武汉大学生命科学学院

研究生《高等遗传学》课程英文文献阅读

资料汇编

2016.12

序 言

遗传学是当代生命科学发展最快的学科之一，也是当代生命科学的核心。基于我们教学团队对我院遗传学专业的研究生学位课《高等遗传学》（以前是《分子遗传学》，后来修订教学方案改名为《高等遗传学》）长期的教学实践，体会到当代遗传学的发展日新月异，新的理论、新的技术方法层出不穷，特别是基因组学等组学的迅猛发展，更是令人目不暇接。为了学习这些新的知识、了解学科前沿发展，我们有必要阅读英文著作。阅读英文学术文献，不仅可以提高我们的知识理论水平，也可以提高我们的英文阅读能力，还可以以免出现歧义，忠实理解原文。为此，我们特别收集近年来在生命科学领域，特别是遗传学领域有重要影响的英文学术资料，汇编成册，以供本课程研究生学习阅读，作为《高等遗传学》课程的教学内容之一。

目录

REVIEW

National Science Review
1: 190–204, 2014
doi: 10.1093/nsr/nwu008
Advance access publication 15 May 2014

BIOLOGY & BIOCHEMISTRY

Special Topic: Frontiers in RNA Research

Non-coding RNA: a new frontier in regulatory biology

Xiang-Dong Fu

ABSTRACT

A striking finding in the past decade is the production of numerous non-coding RNAs (ncRNAs) from mammalian genomes. While it is entirely possible that many of those ncRNAs are transcription noises or by-products of RNA processing, increasing evidence suggests that a large fraction of them are functional and provide various regulatory activities in the cell. Thus, functional genomics and proteomics are incomplete without understanding functional ribonomics. As has been long suggested by the 'RNA world' hypothesis, many ncRNAs have the capacity to act like proteins in diverse biochemical processes. The enormous amount of information residing in the primary sequences and secondary structures of ncRNAs makes them particularly suited to function as scaffolds for molecular interactions. In addition, their functions appear to be stringently controlled by default via abundant nucleases when not engaged in specific interactions. This review focuses on the functional properties of regulatory ncRNAs in comparison with proteins and emphasizes both the opportunities and challenges in future ncRNA research.

Keywords: the RNA world, non-coding RNA, biological functions, regulatory mechanisms, experimental approaches, functional genomics

INTRODUCTION

A major surprise since the completion of the human genome and subsequent sequencing of all biological model organisms is the limited number of protein-coding genes, which neither correlates with the complexity of organisms nor accounts for the selection pressure during the evolution of modern organisms [1]. In humans, the protein-coding sequences occupy only ∼1.5% of the genome, and when considering intervening sequences (introns) within protein-coding genes and 5′ and 3′ untranslated regions, this number goes up to only ∼28%. Much of the remaining portion of the human genome used to be considered 'junk' DNA because ∼59% are repeat sequences; however, recent analysis by the Encyclopedia of DNA Elements (ENCODE) project suggests that ∼80% of the genome appears to participate in some sort of biochemical activities that might be functionally important [2]. This suggests a general paradigm for functional DNA elements embedded in the non-coding part of mammalian genomes.

While initial microarray-based results met with skepticism, the ENCODE data generated by the latest deep sequencing technologies demonstrated that at least 70% of the human genome has the capacity to produce transcripts of various sizes, many of which are conserved in animal kingdom [2]. Besides mRNAs already annotated, most other transcripts do not seem to encode for proteins and are generally referred to as non-coding RNAs (ncRNAs) [3]. Although debate continues with respect to the possibility that some of these ncRNAs may still direct synthesis of short peptides, the consensus is that they are largely non-coding, which is supported by the evidence from ribosome profiling [4] and by the large-scale proteomics analysis performed on two ENCODE cell lines [5]. While most of these ncRNAs have yet to be biochemically characterized, we are witnessing functional assignment to an increasing number of ncRNAs, leading to birth of a new discipline in biological research.

Like many emerging disciplines, the ncRNA field has received great attention in recent years from the general research community, and the progress made has been extensively reviewed from the perspective of mechanistic insights [6–8] and/or

Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, University of California, San Diego, La Jolla, CA 92093-0651, USA

E-mail: xdfu@ucsd.edu

Received 9 March 2014; **Revised** 1 April 2014; **Accepted** 2 April 2014

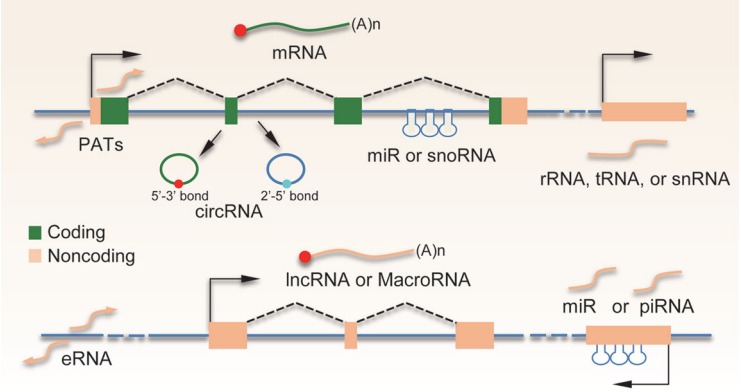


Figure 1. Production of distinct classes of ncRNAs from mammalian genomes. Top: protein-coding (green lines) genes produce divergent PATs at the transcription start site. Certain exonic and intronic sequences have the capacity to generate circRNAs containing either 3′–5′ or 2′–5′ phosphodiester bonds. Many intronic sequences can also encode for miRNAs or snoRNAs. Genes for rRNAs, tRNAs, or a subfraction of snRNAs are transcribed from separate genes. Bottom: similar to protein-coding genes, transcription enhancers also produce divergent transcripts, known as eRNAs. Most of the lncRNA genes contain at least one intron and are transcribed and processed in the same way as protein-coding genes except that they do not have coding potential (yellow line). miRNAs and piRNAs can also be derived from various intergenic regions.

biological functions [9–11]. Instead of enumerating numerous great points that have been made in those reviews, here I highlight the biochemical property of ncRNAs in comparison with proteins to formulate ideas for future research, the uniqueness of ncRNA research, which calls for the great need to develop new experimental approaches, and the potential to exploit ncRNA as a new class of biomarkers or therapeutic targets in biomedical and biotechnological applications.

Housekeeping ncRNAs

Abbreviation	Full name	Function
rRNA	Ribosomal RNA	Translational machinery
tRNA	Transfer RNA	Amino acid carriers
snRNA	Small nuclear RNA	RNA processing
snoRNA	Small nucleolar RNA	RNA modifications
TR	Telomere RNA	Chromosome end synthesis

Regulatory ncRNAs

Abbreviation	Full name	Function
miRNA	MicroRNAs	RNA stability and translation control
endo-siRNA	Endogenous siRNA	RNA degradation
rasiRNA	Repeat-derived RNA	Transcriptional control
piRNA	Piwi-associated RNA	Silencing transposon and mRNA decay
eRNA	Enhancer-derived RNA	Regulation of gene expression
PATs	Promoter-associated RNA	Transcription initiation and pause release
lncRNA	Long non-coding RNA	Imprinting, epigenetics, nuclear structure

Box 1.

ncRNA: OLD AND NEW

ncRNAs may be new to the research community at large, but actually ancient among RNA researchers. Classic ncRNAs that have been intensively studied in the past five decades since the birth of molecular biology include small ncRNAs, such as transfer RNAs (tRNAs) for carrying amino acids, small nucleolus RNAs (snoRNAs) for RNA modifications, and small nuclear RNAs (snRNAs) for RNA splicing, and large ones, such as ribosomal RNAs (rRNAs) for protein synthesis (Box 1 and Fig. 1). These ncRNAs may be considered 'constitutive', because they are abundantly and ubiquitously expressed in all cell types and provide essential functions to the organism. This class may also include the telomere complex-associated guide RNA, which is essential for the end formation and maintenance of chromosomes in normal proliferating cells even though the telomere complex and the ncRNA in it are often compromised in cancer cells [12].

We now have extensive knowledge about 'tiny' ncRNAs, such as microRNAs (miRNAs), endogenous siRNAs (endo-siRNAs), and PIWI-associated small RNAs (piRNAs) that are expressed in animals and plants (Box 1). The biogenesis, targeting, and function of these classes of ncRNAs have been extensively studied and reviewed [13–17] (see also Chen *et al.*, this issue), and are thus not focused in this review. These small RNAs are normally processed from larger RNA precursors, either from their own transcripts or from sequences within specific protein-coding genes (Fig. 1). In contrast to these small RNAs, deep sequencing has identified an increasing number of long intergenic non-coding RNAs (lincRNAs) or simply long non-coding RNAs (lncRNAs), now listed in various databases [18,19], which has received great attention from the research community.

In general, ncRNAs have been classified based on an arbitrary size cut-off of 200 nt to separate small ncRNAs from lncRNAs. However, many ncRNAs may fall into both sides of this cut-off, such as enhancer-associated RNAs (eRNAs), promoter-associated transcripts (PATs), and the more recently emerged circular RNAs (circRNAs) (Box 1; Fig. 1). In fact, these ncRNAs have their own structural features at each end, as eRNAs and PATs have cap, but no poly(A) tail [20], while circRNAs obviously have no ends, which add to structural characteristics of other ncRNAs after processing (e.g. snRNAs with the 5′ tri-methylated cap, miRNAs with the 5′-phosphate, etc.). These features distinguish them from the class of lncRNAs (Box 1), which are transcribed and processed in an identical way to that of protein-coding genes (e.g. capping,

splicing, and polyadenylation, see Fig. 1), and as such, their genes are also associated with characteristic chromatin marks (e.g. H3K4me3 at promoters and H3K36me3 in the gene body), which have been exploited for their prediction, identification, and characterization in mammalian genomes [21].

A common feature of newly identified ncRNAs is their highly regulated expression in different cell types or during development. Our current understanding of their functions, although still quite limited, suggests that these ncRNAs may have diverse regulatory activities (Box 1). Because ncRNAs are either transcribed from specific genomic loci or derived from segments of protein-coding genes, the question is whether all expressed ncRNAs that are detectable by sensitive technologies are functional or some of them may simply reflect transcriptional noises or by-products of RNA processing [22]. A deeper question is whether the process of producing some of those ncRNAs, rather than the final products, is of biological importance because transcription of these ncRNAs is often associated with chromatin remodeling activities. Despite continuous debate on these valid questions, the field has experienced tremendous progress in elucidating the function and mechanism of various ncRNAs, particularly lncRNAs. Thus, for practical reasons, one may first focus on studying ncRNAs that have already some functional evidence, while ignoring many potential 'junk' RNAs, at least for the time being.

FUNCTION OF ncRNA IN COMPARISON WITH PROTEIN

The hypothesis of 'the RNA world' proposes that the development of life, which has to fulfill the requirement of having the ability to carry and replicate its genetic material, may begin with RNA [23,24]. ncRNAs appear to have preserved most, if not all, of their original features and functions in modern organisms that have evolved to adopt more efficient strategies to replicate and express their genetic information along the central dogma from DNA to RNA to protein. As a result of exploring selective advantages of proteins and RNA during evolution, many functions of RNA are passed onto proteins while others are retained. In this regard, it might be informative to compare the function of ncRNAs with proteins to conceptualize ncRNA function and mechanism.

RNA as enzyme

One of the key functions of proteins is to catalyze chemical reactions. Some ncRNAs have long been known to preserve this critical function, known as catalytic RNA, such as the RNAs associated with RNase P required for tRNA processing [25] and auto-catalytic introns [26]. In fact, through *in vitro* selection from random sequences, one may select RNA capable of catalyzing RNA ligation [27] or polymerization [28]. Other ncRNAs preserve their catalytic function only when folded correctly with help of proteins. The best known example is rRNAs in which all key catalytic reactions in reading the coding information in mRNA are provided by the so-called RNA centers [29]. This may also be the case in the spliceosome, which is responsible for intron removal during pre-mRNA splicing and where the catalytic center may form with both RNA and proteins [30]. Therefore, although most catalytic activities of RNA have been passed onto proteins in modern organisms, at least some ncRNAs appear to have kept such function during evolution. Even so, some key functional properties of RNA are maintained in many ribonucleoprotein (RNP) machines. The best known examples are in fact miRNAs and piRNAs in argonaute-containing complexes where these tiny ncRNAs provide targeting information whereas the associated proteins execute the biochemical reactions [31,32]. We thus should not be surprised if many additional ncRNAs are found to make direct contribution to catalysis in the form of RNPs.

RNA as scaffold of molecular interactions

A major function of proteins in the cell is to engage in protein–protein, protein–DNA, and protein–RNA interactions in diverse biochemical reactions. These functions are mediated by specific domains, ∼600 of which have been characterized to date among ∼3000 potential ones [33–35]. In comparison, RNA seems to have similar, if not larger, capacity to perform such molecular interactions through their unique sequence motifs and secondary structures, the latter of which may adapt into different combinations when exposed to different environments or interacting with different proteins. In principle, a specific RNA moiety may interact with DNA or RNA through base-pairing whereas both primary sequences and secondary structures may serve as modules for interactions with specific proteins or protein complexes. For example, specific stem-loop domains in the 7SK RNA are known to interact with distinct protein components [36], and the lncRNA HOTAIR uses its 5′ domain to interact with Polycomb Complex 2 (PRC2) and its 3′ domain to recruit the histone lysine 4 demethylase LSD1, thus coordinating two separate transcription repressor complexes to act on target genes [37]. The ability of a ncRNA to simultaneously engage in interactions with DNA

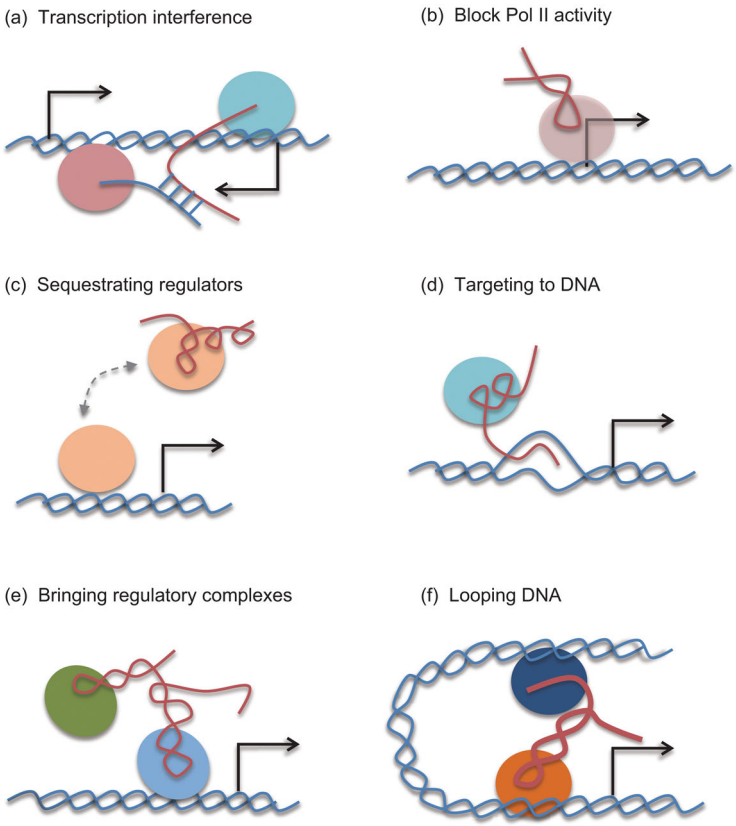


Figure 2. Modes of ncRNA action on genomic DNA in regulated gene expression. lncRNAs are best characterized for their interactions with transcriptional regulators on functional DNA elements. (a) Various antisense transcripts, which appear to be quite widespread in humans and mice [42], may act as ncRNAs to interfere with Pol II elongation [41]. (b) Repeat-derived ncRNAs to block transcription. The prototype ncRNAs in this class are some transcribed Alu sequences, which bind to and interfere with Pol II function at gene promoters [44]. (c) A ncRNA may function as a decoy to compete for a specific transcription factor. The prototype for this mode is PANDA in sequestering the transcription factor NF-YA [45]. (d) A ncRNA may also facilitate the recruitment of a transcription regulator to a specific target site by engaging base-pairing interactions with genomic DNA. The prototype for this mode is the rRNA gene PATs [38]. (e) A ncRNA may bridge protein–protein interactions between transcription regulators to enhance their activities on a common DNA target. The prototype for this mode is the ncRNA HOTAIR in bridging PRC2 and the lysine demethylase LSD1 to mediate gene silencing [37]. (f) A ncRNA may mediate long-distance interactions between promoter and enhancer during transcription activation. Both *cis*-acting eRNAs and lncRNAs have been demonstrated to play such a role [40,46,47,51].

and proteins has been exemplified with the rRNA gene-associated transcripts, which, together with the transcription factor TTF-1, recruit the DNA methyltransferase DNMT3b to CpG islands [38]. These examples illustrate unique advantages of ncRNAs in the regulation of gene expression.

The ncRNA steroid receptor RNA activator is one of the first examples documented to function as a transcription co-activator in gene activation [39], and we now know that many other ncRNAs appear to have such enhancer function [40]. Numerous studies have exposed the mechanisms of regulatory ncRNAs in transcriptional control, including (1) transcription interference by antisense RNA [41,42] (Fig. 2a), (2) direct inhibition of Pol II activity by Alu repeat-derived transcripts [43,44] (Fig. 2b), (3) sequestration of transcriptional regulators [45] (Fig. 2c), (4) guiding transcription regulators to specific regulatory loci through RNA–DNA base-pairing interactions [38] (Fig. 2d), (5) recruitment of additional transcription regulators [37] (Fig. 2e), and (6) mediating long-distance interactions between promoter and enhancer [40,46] (Fig. 2f). Each of these action mechanisms by specific lncRNAs on their target genes has been detailed in multiple recent reviews [6–8,11]. Interestingly, a recent study showed that two lncRNAs (PRNCR1 and PCGEM1) overexpressed in prostate cancer cells interact in a consecutive fashion with the androgen receptor to promote gene expression and cell proliferation in castration-resistant prostate cancer [47]. These and other findings emphasize the involvement of extensive RNA-dependent interactions in transcriptional control.

Cis-acting RNA as regulatory signal

A common property associated with many regulatory ncRNAs is their action in *cis*, meaning that they function at the genomic loci where they are transcribed [40], which is likely due to their rapid turnover once released from the site of synthesis. An analogy may be made in this case with secreted proteins synthesized on endoplasmic reticulum (ER), where the signal peptide guides the protein during translation into the lumen of ER and then removed by peptidase [48]. Some promoter-proximal ncRNAs appear to interfere in *cis* with transcription either through direct interaction with core components of the transcription machinery [49] or through separate RNA-binding proteins (RBPs) [50]. Certain lncRNAs, such as HOTTIP, appear to also act in *cis* because of the difficulty in restoring their functional requirement with exogenous transcripts [51]. However, inactivation of most lncRNAs by RNAi seems to invoke genome-wide responses, implying that those lncRNAs may function in *trans* to module gene expression in multiple locations in the genome [52].

One particular type of ncRNAs that function exclusively near the site of their production is enhancer-transcribed ncRNAs (or eRNAs) [53,54]. Recent studies demonstrated that eRNA production is essential for activating their targeted promoters [20,46,55,56]. As enhancer activities may reflect binding and activity of Pol II, which has been shown to induce chromatin remodeling [57] and promote

DNA looping between enhancer and promoter [56], the question is whether or not the process of such transcriptional activities might be more functionally relevant than the RNA products. A BoxB-λN tethering strategy was first used to demonstrate HOTTIP in coordinating long-range chromatin interactions [51], and a recent study also took this approach to show that eRNA mediates DNA looping between enhancer and promoter [46].

Another class of potential *cis*-acting ncRNAs is PATs. Interestingly, most mammalian genes appear to express divergent transcripts from their promoters, a phenomenon that is not evident in yeast or *Drosophila* [58,59]. Currently, little is known about the function of these ncRNAs transcribed in the opposite direction of the genes. Interestingly, the antisense transcripts tend to lack U1-binding sites whereas the sense transcripts lack the polyadenylation signals [60]. These features might be responsible for the termination of antisense transcription while allowing sense transcription to proceed, as U1 is known to protect the genome by preventing premature transcriptional termination [61]. The sense PATs may also represent aborted transcription products of paused Pol II immediately downstream of mammalian promoters [62]. Interestingly, one such RNA signal has been well studied in HIV-1, where it attracts the HIV tat protein to bind and recruit additional transcription activators, particularly pTEFb, a Pol II CTD kinase, to release paused Pol II into the gene body [63]. A recent study indicates that many cellular genes may employ a similar mechanism through the splicing factor (SRSF2) to facilitate pause release of Pol II from gene promoter into gene body [64], thus suggesting a general role of PATs in providing signals for Pol II to enter productive elongation. It has also been demonstrated that nascent RNA from the gene body near the transcription start site may provide *cis* signals for the Polycomb Complexes to bind [65]. Another important message from these studies is that parts of pre-mRNAs from protein-coding genes may also be considered as a new class of ncRNAs in regulated transcription.

Trans-acting RNA as molecular sink

The molecular sink mechanism is a key strategy for proteins to function in signaling networks in mammalian cells. This concept has also been well documented with many RNA motifs in mRNAs as well as in transcripts from transcribed pseudogenes in mammalian genomes [66,67], again indicating that some parts of mRNAs also function as ncRNAs in nature. These RNA elements have been shown to sequester specific miRNAs to prevent their action on other target mRNAs, but the stoichiometry between competing ncRNAs and target RNAs has to be considered in each case for the physiological relevance of any sequestration effect detected [68]. Some specific lncRNAs have also been shown to sponge miRNA [69] and titrate transcription activators to inhibit cell cycle progression under starvation conditions [70] or in response to DNA damage [45]. Therefore, the entire repertoire of expressed RNAs, whether they are mRNAs or ncRNAs, may participate in diverse RNA–RNA or RNA–protein interaction networks to regulate various cellular activities.

Interestingly, analysis of poly(A−) RNA, which has been largely ignored in the past, revealed many stable ncRNA species, which have been abundantly detected in the oocyte nucleus [71]. One of the general mechanisms for these ncRNAs to remain stable may be that their ends are somehow sealed. Three strategies have been elucidated for stabilization of such ncRNAs. One is to ligate their 5′ and 3′ ends, thus forming circRNAs (see Fig. 1) [72,73]. This likely results from the action of the spliceosome, leading to the ligation of the upstream 3′ splice site to the downstream 5′ splice site of an exon, although the precise mechanism for their production remains to be understood. Interestingly, through characterizing poly(A−) RNAs, another strategy to 'seal' the ends was recently revealed, which is to prevent debranching on some released introns [74]. This type of intron-derived circRNAs is thus sealed by the 2′–5′ phosphodiester bond formed at the branchpoint during pre-mRNA splicing (see Fig. 1). The third strategy to protect the RNA ends is via some stable RNA moieties, such as those found in snoRNAs [75] or the formation of a triple helical structure, such as that characterized at the ends of the stable MALAT-1 RNA [76,77] and some virus-derived ncRNAs [78]. Such RNA structures, either alone or in complex with specific RBPs, protect the RNA from degradation after release from their pre-mRNA precursors.

Functionally, one specific circRNA has been shown to contain an array of binding sites for miRNAs, thus serving as a molecular sink to prevent the miRNAs from interacting with their targets [72,73]. The snoRNA-protected intronic ncRNAs appear to trap a critical RNA binding protein RBFox2, thus titrating its active pool for regulated splicing in the cell [75]. In fact, the classic RNA that serves as a molecular sink is the very abundant 7SK RNA, which has been well characterized to bridge pTEFb to its inhibitor HEXIM1 in the inactive pool of the CTD kinase in the cell [63]. It is unlikely, however, that a molecular sink is the only function associated with various stable ncRNAs. For example, the

intron-derived circRNAs sealed by the 2′–5′ phosphodiester bond appear to play a positive role in transcription of their host genes, although the mechanism has remained elusive [74]. This finding further highlights the functional importance of various sequences in the pre-mRNA of protein-coding genes, as they not only give rise to miRNAs and snoRNAs, but also produce various circRNAs that appear to have both *cis* and *trans* functions.

RNA as ligand

Both small molecules and proteins are well known for their abilities to bind and induce conformational changes of their protein partners, thereby invoking signaling. ncRNAs appear to have a similar role in modulating protein conformation. One such example is a DNA damage-induced ncRNA from the cyclin D1 promoter-proximal region. This ncRNA binds to the RNA binding protein TLS to induce its conformational changes to unmask another domain in the protein for additional protein–protein interactions to take place, eventually leading to transcriptional repression [50].

The miRNA Let-7 appears to also act like a ligand in activating the Toll-like receptor 7, which appears to be a critical event in Let-7-induced neurodegeneration [79]. Small RNAs as ligands have also been exemplified by piRNAs, which, upon incorporating into the PIWI complex, induce conformational changes of the PIWI protein (MIWI in mice) to permit its ubiquitination by a specific E3 ligase [80]. This ncRNA-induced signaling event appears to play a vital role in spermiogenesis by triggering the eventual clearance of the piRNA machinery, a pathway proven to be essential for producing mature sperms in the testis. These findings illustrate that ncRNAs can function as ligands to regulate the conformation of their target proteins to trigger the next set of molecular interactions in some important biological processes. Future structural studies of RNPs may elucidate detailed mechanisms underlying such ncRNA-induced molecular switches.

RNA as organizer of cellular structures

Many ncRNAs are quite large in size and have been referred to as macroRNAs. The best example is the nuclear enriched abundant transcript 1 (NEAT-1). NEAT-1 has two isoforms (the larger one is ∼23 kb in length and the smaller one is 3.7 kb in human, 3.2 kb in mouse), both of which are localized to a specific nuclear domain known as paraspeckles [81,82]. The function of paraspeckles is largely known, although a more recent study suggests an active role of NEAT-1 in facilitating the expression of some antiviral genes [83]. A large number of RBPs have been identified to be part of this nuclear structure, although a few core factors, such as Nono, PSP1, and PSF, appear to be selectively concentrated in this nuclear domain [84]. Many repeat-containing RNAs have been shown to associate with this structure, suggesting that the domain might arise from clustering some specific classes of ncRNAs along with their RBPs [85,86]. The larger NEAT-1 isoform appears to play a critical role in organizing such clusters, as targeted degradation of this ncRNA disrupted the structure [87,88], and ectopic expression of this large, but not small, NEAT-1 isoform was sufficient to induce *de novo* formation of a paraspeckle-like structure around it [89].

The name of paraspeckle is due to the spatial relationship of the domain to another nuclear domain known as speckles [90]. As numerous factors implicated in the splicing reaction have been localized to this structure, it has been a cellular hallmark for the splicing machinery [91]. However, its primary function in pre-mRNA splicing has long been a subject of debate. A popular view is that this domain serves as a storage site for splicing factors; however, increasing evidence points to a more active role of the domain in gene expression via coordinating transcription and splicing reactions at its vicinity, thus suggesting that this nuclear domain may play a larger role in organizing the genome for concerted transcription and post-transcriptional processing events [92,93]. Interestingly, another large lncRNA, known as NEAT-2/MALAT-1 of ∼7.5 kb in size, lies in the heart of individual nuclear speckles. The initial MALAT-1 transcript contains a tRNA-like structure at its 3′ end, which is processed to produce the mature MALAT-1 retained in the nucleus, releasing the tRNA-like small RNA to the cytoplasm [94]. Unlike NEAT-1, mature MALAT-1 does not seem to be responsible for the formation or maintenance of nuclear speckles [95]. However, depletion of this large lncRNA has been shown to affect specific events associated with nuclear speckles, such as SR protein phosphorylation [96], implying that the lncRNA is involved in various protein–protein interactions to facilitate the establishment and dynamics of this non-membrane-bound organelle in the nucleus. Interestingly, NEAT-2/MALAT-1 was originally identified as a nuclear ncRNA that was dramatically elevated in tumor cells [97], which appears to be important for metastasis of lung cancer [98], indicating that this macroRNA may have an active role in cancer initiation and/or progression through its function in regulated gene expression. It is however important to point out that knockout of either NEAT-1 or NEAT-2/MALAT-1 produced no

obvious phenotypic defects, indicating that these ncRNAs are not essential for mouse development [95,99].

Contrary to the nuclear structures associated with active gene expression, other nuclear domains are functionally linked to gene repression, such as the Polycomb body in the nucleus, which contains protein complexes responsible for depositing repressive marks, such as H3K27me3, to chromatin. This domain contains numerous ncRNAs, including Tug 1 [100]. While the precise role of this lncRNA has remained unclear, its association with the Polycomb body may compete with some common gene expression regulators that are partitioned between active and repressive domains in the nucleus, and regulated exchange between these domains appears to be a key event in switching the functional states of many genes [101]. Therefore, specific lncRNAs may provide signals or docking sites for regulatory proteins or protein complexes, thereby contributing to the organization of the human genome in the 3D space of the nucleus. More recently, repeat-derived ncRNAs were suggested to be a key part of nuclear scaffold for maintaining chromosome territories [102]. Together, various nuclear domain-associated lncRNAs may be considered as part of nuclear skeleton in analogy with the cytoskeleton in the cytoplasm.

Secreted RNA as potential hormone

ncRNAs are made in the nucleus either from their own genes or genomic loci or processed from their host genes. As cells have very active machineries to degrade most transcribed RNAs, functional ncRNAs must have evolved some strategies to survive various RNA surveillance mechanisms. As described above, some ncRNAs have specific structures to protect their ends to make them inaccessible to exonucleases while others may gain protection by forming specific RNPs. A fraction of ncRNAs are able to not only survive degradation in the cell, but also make it to the extracellular space. So far, this has been documented for miRNAs, which appear to be assembled into microvesicles for secretion [103]. We are still early in understanding how some miRNAs are imported or assembled into microvesicles for secretion, and how the specificity, if any, might be established in such a process. In any case, the detection of secreted miRNAs in the circulation system seems to provide a unique set of biomarkers for disease diagnosis [104–106]. A more important question is what these secreted miRNAs might do in the circulation system. Do they function as hormones to act in distal organs? Initial studies provide some evidence for such a possibility [107,108]. Remarkably, some exogenous miRNAs from food supply might also have such a role [109], although the finding has remained to be substantiated [110]. Overall, the idea that RNAs can function as hormones has remained as a hypothetic function for secreted miRNAs.

In concluding this section, I wish to make the point that our current knowledge has significantly expanded the function of RNAs as information carriers. They appear to be able to perform a large array of cellular functions that have been ascribed to proteins. Importantly, we are still glimpsing at the tip of iceberg, despite the impression that many working principles have been elucidated with specific ncRNA examples.

STRATEGIES FOR FUNCTIONAL AND MECHANISTIC STUDIES OF ncRNA

Small ncRNAs, particularly miRNAs, are well known for their roles in diverse biological pathways. The existing examples of characterized lncRNAs have also demonstrated their widespread participation in biological functions, ranging from dosage compensation [111,112], cell cycle control [45,113], stem cell maintenance and differentiation [52,114,115], development [116–118], and cancer etiology and progression [47,119,120]. Given their functional resemblance to proteins, essentially all experimental strategies developed to decipher protein functions may be applied to ncRNA research; however, because of their uniqueness as a linear chain of nucleic acids and the ability to fold into multiple secondary and tertiary structures, new approaches are also needed to study their functions and action mechanisms. In this section, I briefly discuss some common and unique approaches developed for ncRNA research (Box 2).

Experimental approaches to defining ncRNA function

As with protein-coding genes, one of the most important experimental approaches to study ncRNAs nowadays is to determine their unique expression patterns associated with a specific biological question under investigation and to conduct loss-of-function studies in a particular biological setting. Using modern genomics strategies, it has become a routine to profile gene expression by RNA-seq in any given biological system [121,122], which may be combined with various affinity methods to detect RNA (both coding and non-coding) at different stages of gene expression [123,124]. The identification of the entire set of expressed lncRNAs would allow comparison under different experimental

Detection of binding events

Abbreviation	Full name	Application	Ref
CLIP	Crosslinking IP	Protein-RNA interactions	142
ChIRP	Chromatin isolation by RNA purification	RNA interaction with genomic DNA	130
CHART	Capture hybridization analysis of RNA targets	RNA interaction with genomic DNA and proteins	128
CLASH	Crosslinking, ligation, sequencing of hybrids	RNA-RNA interactions	33

Measurement of functional consequences

Abbreviation	Full name	Application	Ref
RNA-seq	RNA sequencing	Gene expression profiling	121,122
GRO-seq	Global nuclear Run-On	Detection of nascent RNA	123
Bru-seq	BrU labeling/chasing	Time course analysis of RNA	124
RP	Ribosome profiling	Studying translational control	4,161

Probing RNA secondary structure

Abbreviation	Full name	Application	Ref
SHAPE	Selective 2'-hydroxyl acylation analyzed by primer extension	Chemical probing of RNA secondary structure	132
PARS	Parallel analysis of RNA structure	Enzymatic probing of RNA secondary structure	133
DMS-seq	Dimethyl sulphate-modified RNA for sequencing	Probing unpaired adenine and cytosine in RNA	134,135

Characterization of RNA binding proteins

Abbreviation	Full name	Application	Ref
RNAComplete	RNA complete	RNA binding specificity	144
SeqRS	RNA sequence specificity landscapes	RNA binding affinity	143
IC	Interactome capture	mRNA bound proteins	139
MS2-TRAP	MS2-targged RNA affinity purification	Capture RBPs on RNA	136
RAT	RNA affinity in tandem	Capture RBPs on RNA	137
Csy4 Select	Csy4 selection of RNA-protein complexes	Capture RBPs on RNA	157

Genome editing

Abbreviation	Full name	Application	Ref
TALEN	Transcription activator-like effector nucleases	Targeted gene mutation	156
CRISPR	Clustered, regularly interspaced short palindromic repeats	Targeted gene deletion and insertion	138,157
CRISPRi	CRISPR interference	Interrogate genomic loci to modulate gene expression	158,161

Box 2.

conditions or between different cell types to identify differentially expressed lncRNAs [116,125]. The challenge is to determine on which specific lncRNA(s) to further study. Currently, most studies focus on differentially expressed lncRNAs that are expressed with sufficient abundance. By using siRNA or antisense oligonucleotides (ASO), the latter of which appear to be more efficient in depleting lncRNAs via endogenous RNase H activities [126], one can efficiently deplete specific lncRNAs to evaluate their functional requirement. If resources are available or permit, this loss-of-function approach may be applied genome-wide to obtain a comprehensive set of lncRNAs involved in some defined biological processes, as exemplified on stem cells [52].

The hard part of ncRNA research is to probe for the mechanism and explore new regulatory concepts. The cellular localization of specific ncRNAs may be first determined to obtain an approximation of their functional sites. As mRNAs are known to display remarkable localization patterns in the cell [127], the localization of ncRNAs, particularly lncRNAs, might be informative to their cellular functions. To understand the function of a specific lncRNA, it is often important to identify its protein partners. Furthermore, if the lncRNA under investigation acts in the nucleus to regulate gene expression, one will also need to determine its target genes. To identify protein partners, antibodies are very useful tools for protein research, but for lncRNA, one has to rely on some entirely distinct approaches. One such approach is to use affinity tagged (such as biotin) oligos to capture specific lncRNA followed by deep sequencing of linked DNA and/or by mass spectrometric analysis of associated proteins, a method known as CHART-seq [128], which has been applied to elucidate two-step spreading of Xist ncRNA complexes during X-chromosome inactivation [129]. A related method called ChIRP-seq was developed in parallel to survey lncRNA occupancy on genomic DNA [130]. This technique has been applied to probe the genomic interaction of the 7SK complex on so-called anti-pause enhancers [131].

To efficiently use this approach, it would be helpful to know the exposed RNA regions in the cell by probing RNA structure in living cells [132,133]. Two recent studies reported a more robust method based on dimethyl sulfate modification of exposed adenines and cytosines followed by deep sequencing of RNA containing the modified residues to achieve high-resolution mapping of the RNA secondary structure [134,135]. These new approaches will greatly accelerate the discovery of regulatory events on RNA targets by both ncRNAs and specific RBPs.

Another approach is to epitope tag an lncRNA with an MS2 moiety, thus permitting the capture of the lncRNA-containing RBP with an MS2 fusion protein [136]. An analogous strategy is to use an RNA tag that contains two specific hairpins, thus allowing tandem affinity purification of RNA–protein complexes [137]. This RNA-tagging strategy, however, can be problematic if the lncRNA only acts in *cis* or the overexpressed transcript does not

effectively get assembled into its native RNP complexes. This problem can be addressed by using the latest genome editing technology to tag specific ncRNA genes [138] (see below). Given the nucleic acid nature of lncRNA, future studies may also pursue chemical engineering methods to take advantage of specific sequences or structure moieties to introduce affinity groups for lncRNA localization and affinity purification.

Studying ncRNA from the angle of RBPs

It is conceivable that lncRNA functions are mostly mediated by specific RBPs, and, thus, focusing on specific RBPs of interest may be an effective route to study lncRNA function and mechanism in general. Recent studies indicate that mammalian genomes may express at least 1000 RBPs [139], many of which may not even carry annotated RNA-binding domains [140]. As a matter of fact, we do not know the exact distinction between DNA-binding proteins and RBPs, as they have been traditionally studied based on their interactions with DNA or RNA. As a result, some DNA-binding proteins may also bind RNA and the converse may also be true. For example, two recent studies demonstrated that the PRC2, which is responsible for depositing the repressive H3K27me3 mark on histone, actually has high affinity for RNA [141], explaining its extensive interaction with nascent RNA in the cell [65].

An important point is that the cross-linking immunoprecipitation (CLIP) technology and various variants of the approach have demonstrated effectiveness in identifying protein-associated RNAs and mapping such interactions in the genome [142]. Efficient and high-throughput methods have also been developed to determine the RNA binding specificity of RBPs [143,144], and an increasing number of RBPs have been mapped to mammalian genomes using CLIP technologies. Although most published studies to date have been focused on understanding the function of RBPs in RNA metabolism, such as pre-mRNA splicing, the available mapping data indicate that many RBPs also show extensive interactions with diverse lncRNAs [145]. As the CLIP data accumulate and have been organized in the database [146], one may mine such data to identify proteins mapped to specific lncRNAs under investigation. With candidate RBPs and lncRNAs in hand, loss-of-function studies can then be performed to identify common targets for further mechanistic dissection, as exemplified by the study of p53-regulated gene expression that involves both an lncRNA (lincRNA-p21) and a specific RBP (hnRNP K) [147].

Challenges in structural analysis of RNPs

A common approach in mechanistic studies of proteins or protein complexes is to define specific protein domains engaged in a particular molecular interaction and probe a detailed interaction mechanism in crystal structure. Similar approaches are clearly needed for understanding RNA–protein interactions. The challenge in dissecting RNA domains involved in such an interaction with specific proteins has been showcased with HOTAIR, an lncRNA that interacts with two different chromatin remodeling complexes through distinct RNA segments [37]. However, there is a great uncertainty in dissecting domains with *in vitro* transcribed RNA, as RNA may adopt into distinct secondary structures when made *in vitro* versus produced inside cells where specific RBPs may be assembled onto the RNA during transcription and/or processing, which may take place in a sequential fashion. This may make it difficult to reconstitute RNPs that contain multiple protein components for biochemical studies.

In the protein world, ultimate mechanistic insights are obtained from NMR or crystallography. The structure of the largest RNA machine—the ribosomes in complex with tRNA and mRNA—has been resolved at the atomic levels [148,149], and similarly, structures of miRNAs in argonaute proteins have been determined [150–152]. The structural approach has also been applied to an H/ACA box snoRNP particle [153] and a spliceosome subcomplex [154]. In general, however, it has been quite difficult to obtain crystals of many other RNPs, such as the spliceosome, in part because of insufficient materials one can purify from the cell or the lack of ability to preserve relatively stable structures during the purification process for crystallization. The common practice in protein crystallization is to use recombinant proteins, but in light of various potential problems in assembling RNPs *in vitro*, it will be a major challenge to reconstitute large RBPs for structural studies.

Genome engineering to determine ncRNA function

Similar to investigating protein functions in biology, the decisive information is obtained in many cases by gene targeting, which has recently been applied to a set of lncRNAs [155]. We are at the dawn of applying this genetic approach to ncRNA research, especially in light of the recent development of the powerful TALEN and CRISPR/Cas technologies for genome engineering [138,156]. For instance, the CRISPR technology has been used to tag an

lncRNA in its expression unit in the genome to allow capture of specific RNA–protein complexes assembled *in vivo* [157]. In this elegantly designed strategy, a small RNA hairpin is first inserted in the front of specific ncRNA under investigation in the genome by CRISPR. An inactive version of the Cys4 nuclease is next used to efficiently capture the hairpin as part of RNA hybrid along with associated proteins. The affinity-purified RNP is then released for biochemical analysis by using imidazole to activate the Csy4 nuclease. The CRISPR technology can also be used to selectively remove specific ncRNA sequences embedded in their host genes, such as those transcribed as part of introns, to study their functional requirements. Recently, a catalytic inactive form of Cas9 was exploited to develop the CRISPRi system [158,159], which permits both positive and negative modulation of endogenous genes [160] and real-time imaging dynamic movement of specific genomic loci [161]. It is anticipated that the rapidly evolving CRISPR-based genome editing technologies will find wide applications in studying genomic sequences encoding for both small and large ncRNA in the near future.

ncRNA as an integral part of genomics and proteomics

It has become increasingly evident that ncRNAs provide diverse regulatory functions in the cell, and regulatory RNA networks in general represent a crucial interphase between genomics and proteomics (Fig. 3). Both small and large ncRNAs are subjected to regulation by diverse mechanisms to control their expression, biogenesis, and degradation, all of which have been well documented with miRNAs and piRNAs [15,31]. As many lncRNAs are expressed from their own genes, a battery of transcription factors are likely involved in the regulation of these lncRNAs during development or in different cell types in a similar way to the regulation of protein-coding genes.

Most lncRNAs have been characterized by their functions in the nucleus, and their interactions with various nuclear machineries may thus contribute to their nuclear retention. However, many lncRNAs are also detectable in the cytoplasm and clearly function there, as demonstrated with the BACE1-antisense transcript (BACE1-AS) and an Alu-containing lncRNA in the regulation of mRNA stability [162,163]. Because premature stop codons in mRNA trigger the nonsense-mediated RNA decay (NMD) [164], this raises the question of how various lncRNAs escape such a pathway. One possibility is that lncRNAs are not scanned by ribosome beyond immediate 5′ sequences [4,165], as the translation process is known to activate the NMD pathway [166]. However, the key NMD initiator Upf1 appears to have the capacity to bind mRNAs as well as lncRNAs in a translation-independent manner [167]. At this point, we have little knowledge about whether cytoplasmic lncRNAs are sensitive to NMD, which represents an interesting subject for future studies.

One exciting future research area is to decipher the contribution of lncRNAs to local and long-distance genomic interactions (Fig. 3a,b). Functional studies of eRNAs and certain lncRNAs have exemplified the critical role of ncRNAs in mediating enhancer–promoter interactions [46,56,168]. Recent studies suggest that the Xist complex explores some larger genomic domains to help spread the transcription repressor complex during X-chromosome inactivation [129,169]. This strategy may also be exploited for establishing both active and repressive domains that involve genomic segments separated by long linear distance on the same chromosomes or even from different chromosome, which may in turn contribute to the organization of the genome in the 3D space of the nucleus [170,171] (Fig. 3c). Research along this direction may represent a new frontier of ncRNA cell biology.

The intersection of ncRNA research with gene networks has well been established for miRNAs [172]. It is easily imaginable for numerous RNA-dependent protein–protein and protein–DNA interactions to exist in the cell, but systematic effort has yet to be undertaken to study such RNA-dependent interactions (Fig. 3d). Thus, analysis of gene networks would be incomplete without incorporating regulatory ncRNAs into various biological pathways. Towards this general goal, all classes of ncRNAs and their expression patterns have been organized in an integrated database [173]. Such a systems biology approach will greatly accelerate research on ribonomics and its integration with functional genomics and proteomics.

CONCLUSIONS

ncRNAs have undoubtedly become one of the 'hot' spots in modern biological and biomedical research. The existing data have abundantly demonstrated the connection of ncRNAs to diverse disciplines in biology, and have illuminated regulatory paradigms that have been largely attributed to proteins. As ncRNAs can be efficiently targeted by stable ASO,

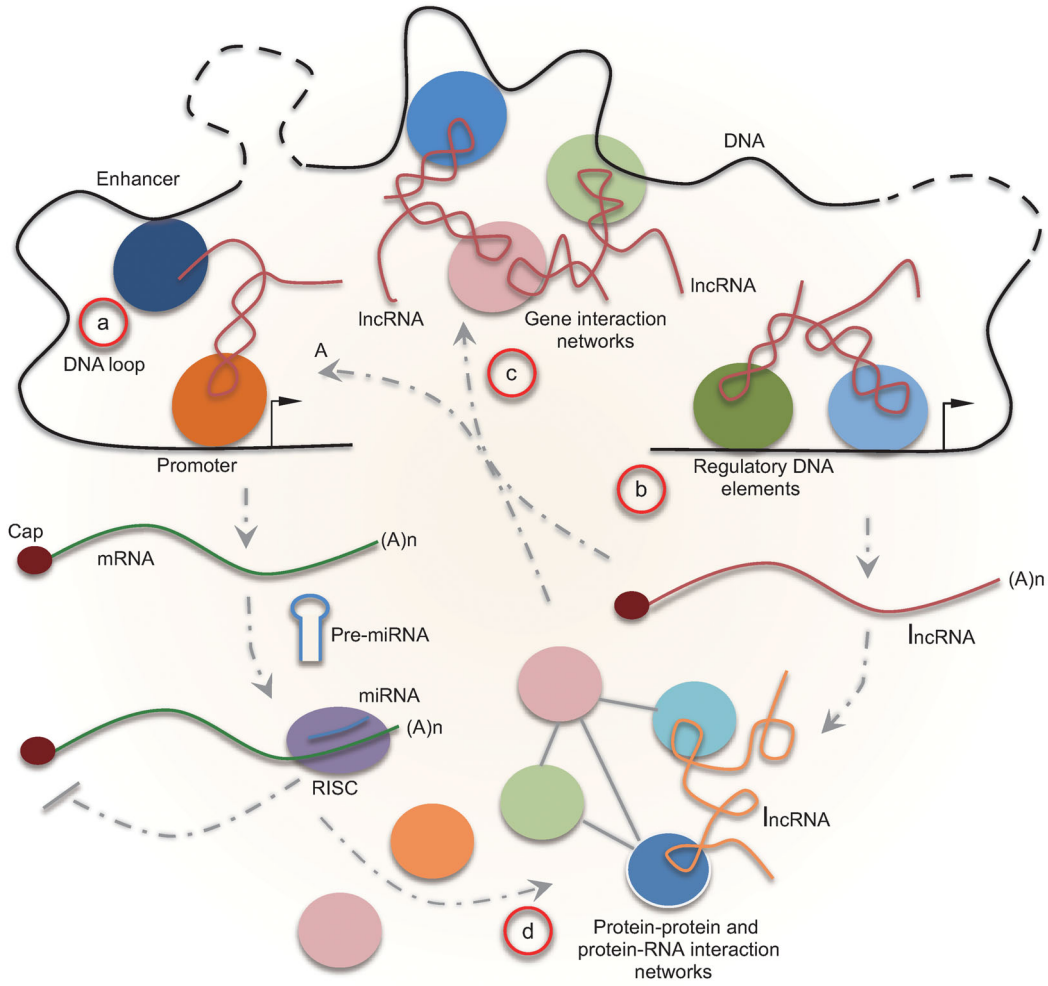


Figure 3. ncRNAs as integrated parts of gene networks. (a) ncRNAs mediate promoter–enhancer interactions to regulate the expression of various protein-coding genes. Protein-coding transcripts are also subjected to regulation by miRNAs to fine tune protein synthesis in the cytoplasm. (b) ncRNA genes produce various regulatory ncRNAs, which then participate in regulated expression of both protein-coding and non-coding genes. (c) ncRNAs may play a critical role in the organization of the genome in the nucleus to coordinate the expression of gene clusters. (d) Regulated gene expression at both the transcriptional and post-transcriptional levels determines the cell type-specific proteome and ncRNAs may also be extensively involved in protein interaction networks, which together contribute to gene networks in the cell.

this approach may be explored as a method to target specific regulatory ncRNAs to understand their biological functions and action mechanisms in basic research and develop novel strategies for disease intervention in clinical applications. The era of ncRNA research has resulted in and benefited from the rapid advance in genomics technologies and informatics approaches that have been developed in recent years. However, we are clearly facing new challenges in dissecting the dark matter in the genome and understanding their mechanisms. Like many breakthroughs made in the history of life science, both opportunities and challenges equalize, which is up to prepared minds to seize the moment in order to make new breakthroughs.

ACKNOWLEDGEMENTS

The author wishes to thank Yang Shi for critical comments, and Xiao Li and Patrick Menzies for proofreading the manuscript. The author is also grateful to reviewers of this article who made numerous insightful suggestions.

FUNDING

Research in the author's research has been supported by grants from National Institutes of Health.

REFERENCES

1. Lander, ES, Linton, LM and Birren, B *et al.* Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860–921.

2. Bernstein, BE, Birney, E and Dunham, I *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**: 57–74.
3. Eddy, SR. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet* 2001; **2**: 919–29.
4. Guttman, M, Russell, P and Ingolia, NT *et al.*, Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 2013; **154**: 240–51.
5. Bánfai, B, Jia, H and Khatun, J *et al.* Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res* 2012; **22**: 1646–57.
6. Wang, X, Song, X and Glass, CK *et al.* The long arm of long noncoding RNAs: roles as sensors regulating gene transcriptional programs. *Cold Spring Harb Perspect Biol* 2011; **3**: a003756.
7. Rinn, JL and Chang, HY. Genome regulation by long noncoding RNAs. *Annu Rev Biochem* 2012; **81**: 145–66.
8. Yang, L, Froberg, JE and Lee, JT. Long noncoding RNAs: fresh perspectives into the RNA world. *Trends Biochem Sci* 2014; **39**: 35–43.
9. Esteller, M. Non-coding RNAs in human disease. *Nat Rev Genet* 2011; **12**: 861–74.
10. Batista, PJ and Chang, HY. Long noncoding RNAs: cellular address codes in development and disease. *Cell* 2013; **152**: 1298–307.
11. Fatica, A and Bozzoni, I. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet* 2014; **15**: 7–21.
12. Nandakumar, J and Cech, TR. Finding the end: recruitment of telomerase to telomeres. *Nat Rev Mol Cell Biol* 2013; **14**: 69–82.
13. Malone, CD and Hannon, GJ. Small RNAs as guardians of the genome. *Cell* 2009; **136**: 656–68.
14. Fabian, MR, Sonenberg, N and Filipowicz, W. Regulation of mRNA translation and stability by microRNAs. *Annu Rev Biochem* 2010; **79**: 351–79.
15. Thomson, T and Lin, H. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annu Rev Cell Dev Biol* 2009; **25**: 355–76.
16. Peng, JC and Lin, H. Beyond transposons: the epigenetic and somatic functions of the Piwi-piRNA mechanism. *Curr Opin Cell Biol* 2013; **25**: 190–4.
17. Pasquinelli, AE. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nat Rev Genet* 2012; **13**: 271–82.
18. Derrien, T, Johnson, R and Bussotti, G *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 2012; **22**: 1775–89.
19. Burge, SW, Daub, J and Eberhardt, R *et al.* Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* 2013; **41**: D226–32.
20. Lam, MT, Cho, H and Lesch, HP *et al.* Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature* 2013; **498**: 511–5.
21. Guttman, M, Amit, I and Garber, M *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 2009; **458**: 223–7.
22. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol* 2007; **14**: 103–5.
23. Joyce, GF. RNA evolution and the origins of life. *Nature* 1989; **338**: 217–24.
24. Gibert, W. The RNA world. *Nature* 1986; **319**: 618.
25. Evans, D, Marquez, SM and Pace, NR. RNase P: interface of the RNA and protein worlds. *Trends Biochem Sci* 2006; **31**: 333–41.
26. Doudna, JA and Cech, TR. The chemical repertoire of natural ribozymes. *Nature* 2002; **418**: 222–8.
27. Wright, MC and Joyce, GF. Continuous in vitro evolution of catalytic function. *Science* 1997; **276**: 614–7.
28. Johnston, WK, Unrau, PJ and Lawrence, MS *et al.* RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science* 2001; **292**: 1319–25.
29. Moore, PB and Steitz, TA. The involvement of RNA in ribosome function. *Nature* 2002; **418**: 229–35.
30. Galej, WP, Oubridge, C and Newman, AJ *et al.* Crystal structure of Prp8 reveals active site cavity of the spliceosome. *Nature* 2013; **493**: 638–43.
31. Czech, B and Hannon, GJ. Small RNA sorting: matchmaking for Argonautes. *Nat Rev Genet* 2011; **12**: 19–31.
32. Helwak, A, Kudla, G and Dudnakova, T *et al.* Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 2013; **153**: 654–65.
33. Marchler-Bauer, A, Zheng, C and Chitsaz, F *et al.* CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res* 2013; **41**: D348–52.
34. Punta, M, Penny, CC and Ruth, YE *et al.* The Pfam protein families database. *Nucleic Acids Res* 2012; **40**: D290–301.
35. Bateman, A, Coggill, P and Finn, RD. DUFs: families in search of function. *Acta Crystallogr F* 2010; **66**: 1148–52.
36. Egloff, S, Van Herreweghe, E and Kiss, T. Regulation of polymerase II transcription by 7SK snRNA: two distinct RNA elements direct P-TEFb and HEXIM1 binding. *Mol Cell Biol* 2006; **26**: 630–42.
37. Tsai, MC, Manor, O and Wan, Y *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 2010; **329**: 689–93.
38. Schmitz, KM, Mayer, C and Postepska, A *et al.* Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Gene Dev* 2010; **24**: 2264–9.
39. Lanz, RB, McKenna, NJ and Onate, SA *et al.* A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. *Cell* 1999; **97**: 17–27.
40. Ørom, UA, Derrien, T and Beringer, M *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* 2010; **143**: 46–58.
41. Hongay, CF, Grisafi, PL and Galitski, T *et al.* Antisense transcription controls cell fate in Saccharomyces cerevisiae. *Cell* 2006; **127**: 735–45.
42. Osato, N, Suzuki, Y and Ikeo, K *et al.* Transcriptional interferences in cis natural antisense transcripts of humans and mice. 2007; *Genetics* **176**: 1299–306.
43. Bevis, M, Wahr, J and Khan, SA *et al.* Bedrock displacements in Greenland manifest ice mass variations, climate cycles and climate change. *Proc Natl Acad Sci USA* 2012; **109**: 11944–8.
44. Mariner, PD, Walters, RD and Espinoza, CA *et al.* Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol Cell* 2008; **29**: 499–509.
45. Hung, T, Wang, Y and Lin, MF *et al.* Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat Genet* 2011; **43**: 621–9.
46. Li, W, Notani, D and Ma, Q *et al.* Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* 2013; **498**: 516–20.
47. Yang, L, Lin, C and Jin, C *et al.* lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. *Nature* 2013; **500**: 598–602.
48. Walter, P and Blobel, G. Mechanism of protein translocation across the endoplasmic reticulum. *Biochem Soc Symp* 1982; **47**: 183–91.
49. Martianov, I, Ramadass, A and Serra Barros, A *et al.* Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* 2007; **445**: 666–70.
50. Wang, X, Arai, S and Song, X *et al.* Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* 2008; **454**: 126–30.
51. Wang, KC, Yang, YW and Liu, B *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 2011; **472**: 120–4.

52. Guttman, M, Donaghey, J and Carey, BW *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 2011; **477**: 295–300.
53. Wang, D, Garcia-Bassets, I and Benner, C *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* 2011; **474**: 390–4.
54. Kim, TK, Hemberg, M and Gray, JM *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* 2010; **465**: 182–7.
55. Melo, CA, Drost, J and Wijchers, PJ *et al.* eRNAs are required for p53-dependent enhancer activity and gene transcription. *Mol Cell* 2013; **49**: 524–35.
56. Lai, F, Orom, UA and Cesaroni, M *et al.* Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* 2013; **494**: 497–501.
57. Sanchez-Elsner, T, Gou, D and Kremmer, E *et al.* Noncoding RNAs of trithorax response elements recruit Drosophila Ash1 to Ultrabithorax. *Science* 2006; **311**: 1118–23.
58. Seila, AC, Calabrese, JM and Levine, SS *et al.* Divergent transcription from active promoters. *Science* 2008; **322**: 1849–51.
59. Core, LJ, Waterfall, JJ and Lis, JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 2008; **322**: 1845–8.
60. Almada, AE, Wu, X and Kriz, AJ *et al.* Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* 2013; **499**: 360–3.
61. Kaida, D, Berg, MG and Younis, I *et al.* U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* 2010; **468**: 664–8.
62. Adelman, K and Lis, JT. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* 2012; **13**: 720–31.
63. Zhou, Q, Li, T and Price, DH. RNA polymerase II elongation control. *Annu Rev Biochem* 2012; **81**: 119–43.
64. Ji, X, Zhou, Y and Pandit, S *et al.* SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. *Cell* 2013; **153**: 855–68.
65. Kaneko, S, Son, J and Shen, SS *et al.* PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat Struct Mol Biol* 2013; **20**: 1258–64.
66. Salmena, L, Poliseno, L and Tay, Y *et al.* A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 2011; **146**: 353–8.
67. Poliseno, L, Salmena, L and Zhang, J *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 2010; **465**: 1033–8.
68. Ulitsky, I and Bartel, DP. lincRNAs: genomics, evolution, and mechanisms. *Cell* 2013; **154**: 26–46.
69. Kallen, AN, Zhou, XB and Xu, J *et al.* The imprinted H19 lncRNA antagonizes let-7 microRNAs. *Mol Cell* 2013; **52**: 101–12.
70. Kino, T, Hurt, DE and Ichijo, T *et al.* Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci Signal* 2010; **3**: ra8.
71. Gardner, EJ, Nizami, ZF and Talbot, CC, Jr *et al.* Stable intronic sequence RNA (sisRNA), a new class of noncoding RNA from the oocyte nucleus of Xenopus tropicalis. *Gene Dev* 2012; **26**: 2550–9.
72. Hansen, TB, Jensen, TI and Clausen, BH *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature* 2013; **495**: 384–8.
73. Memczak, S, Jens, M and Elefsinioti, A *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 2013; **495**: 333–8.
74. Zhang, Y, Zhang, XO and Chen, T *et al.* Circular intronic long noncoding RNAs. *Mol Cell* 2013; **51**: 792–806.
75. Yin, QF, Yang, L and Zhang, Y *et al.* Long noncoding RNAs with snoRNA ends. *Mol Cell* 2012; **48**: 219–30.
76. Brown, JA, Valenstein, ML and Yario, TA *et al.* Formation of triple-helical structures by the 3′-end sequences of MALAT1 and MENβ noncoding RNAs. *Proc Natl Acad Sci USA* 2012; **109**: 19202–7.
77. Wilusz, JE, JnBaptiste, CK and Lu, LY *et al.* A triple helix stabilizes the 3′ ends of long noncoding RNAs that lack poly(A) tails. *Gene Dev* 2012; **26**: 2392–407.
78. Mitton-Fry, RM, DeGregorio, SJ and Wang, J *et al.* Poly(A) tail recognition by a viral RNA element through assembly of a triple helix. *Science* 2010; **330**: 1244–7.
79. Lehmann, SM, Krüger, C and Park, B *et al.* An unconventional role for miRNA: let-7 activates Toll-like receptor 7 and causes neurodegeneration. *Nat Neurosci* 2012; **15**: 827–35.
80. Zhao, S, Gou, LT and Zhang, M *et al.* piRNA-triggered MIWI ubiquitination and removal by APC/C in late spermatogenesis. *Dev Cell* 2013; **24**: 13–25.
81. Bond, CS and Fox, AH. Paraspeckles: nuclear bodies built on long noncoding RNA. *J Cell Biol* 2009; **186**: 637–44.
82. Fox, AH, Lam, YW and Leung, AK *et al.* Paraspeckles: a novel nuclear domain. *Curr Biol* 2002; **12**: 13–25.
83. Imamura, K, Imamachi, N and Akizuki, G *et al.* Long noncoding RNA NEAT1-dependent SFPQ relocation from promoter region to paraspeckle mediates IL8 expression upon immune stimuli. *Mol Cell* 2014; **53**: 393–406.
84. Naganuma, T, Nakagawa, S and Tanigawa, A *et al.* Alternative 3′-end processing of long noncoding RNA initiates construction of nuclear paraspeckles. *EMBO J* 2012; **31**: 4020–34.
85. Prasanth, KV, Prasanth, SG and Xuan, Z *et al.* Regulating gene expression through RNA nuclear retention. *Cell* 2005; **123**: 249–63.
86. Chen, LL and Carmichael, GG. Long noncoding RNAs in mammalian cells: what, where, and why? *Wiley Interdiscip Rev RNA* 2010; **1**: 2–21.
87. Chen, LL and Carmichael, GG. Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. *Mol Cell* 2009; **35**: 467–78.
88. Clemson, CM, Hutchinson, JN and Sara, SA *et al.* An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol Cell* 2009; **33**: 717–26.
89. Mao, YS, Sunwoo, H and Zhang, B *et al.* Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. *Nat Cell Biol* 2011; **13**: 95–101.
90. Fu, XD and Maniatis, T. Factor required for mammalian spliceosome assembly is localized to discrete regions in the nucleus. *Nature* 1990; **343**: 437–41.
91. Spector, DL and Lamond, AI. Nuclear speckles. *Cold Spring Harb Perspect Biol* 2011; **3**: a000646.
92. Dundr, M and Misteli, T. Biogenesis of nuclear bodies. *Cold Spring Harb Perspect Biol* 2010; **2**: a000711.
93. Zhong, XY, Wang, P and Han, J *et al.* SR proteins in vertical integration of gene expression from transcription to RNA processing to translation. *Mol Cell* 2009; **35**: 1–10.
94. Wilusz, JE, Freier, SM and Spector, DL. 3′ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* 2008; **135**: 919–32.
95. Zhang, B, Arun, G and Mao, YS *et al.* The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult. *Cell Rep* 2012; **2**: 111–23.
96. Tripathi, V, Ellis, JD and Shen, Z *et al.* The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell* 2010; **39**: 925–38.

97. Hutchinson, JN, Ensminger, AW and Clemson, CM *et al.* A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* 2007; **8**: 39.
98. Gutschner, T, Hämmerle, M and Eissmann, M *et al.* The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res* 2013; **73**: 1180–9.
99. Nakagawa, S, Naganuma, T and Shioi, G *et al.* Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice. *J Cell Biol* 2011; **193**: 31–9.
100. Khalil, AM, Guttman, M and Huarte, M *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci USA* 2009; **106**: 11667–72.
101. Yang, L, Lin, C and Liu, W *et al.* ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. *Cell* 2011; **147**: 773–88.
102. Hall, LL, Carone, DM and Gomez, AV *et al.* Stable C0T-1 repeat RNA is abundant and is associated with euchromatic interphase chromosomes. *Cell* 2014; **156**: 907–19.
103. Zhang, Y, Liu, D and Chen, X *et al.* Secreted monocytic miR-150 enhances targeted endothelial cell migration. *Mol Cell* 2010; **39**: 133–44.
104. Kosaka, N, Yoshioka, Y and Hagiwara, K *et al.* Trash or Treasure: extracellular microRNAs and cell-to-cell communication. *Front Genet* 2013; **4**: 173.
105. Vickers, KC, Palmisano, BT and Shoucri, BM *et al.* MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins. *Nat Cell Biol* 2011; **13**: 423–33.
106. Gandhi, R, Healy, B and Gholipour, T *et al.* Circulating microRNAs as biomarkers for disease staging in multiple sclerosis. *Ann Neurol* 2013; **73**: 729–40.
107. Png, KJ, Halberg, N and Yoshida, M *et al.* A microRNA regulon that mediates endothelial recruitment and metastasis by cancer cells. *Nature* 2012; **481**: 190–4.
108. Fabbri, M, Paone, A and Calore, F *et al.* MicroRNAs bind to Toll-like receptors to induce prometastatic inflammatory response. *Proc Natl Acad Sci USA* 2012; **109**: E2110–6.
109. Zhang, L, Hou, D and Chen, X *et al.* Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. *Cell Res* 2012; **22**: 107–26.
110. Dickinson, B, Zhang, Y and Petrick, JS *et al.* Lack of detectable oral bioavailability of plant microRNAs after feeding in mice. *Nat Biotechnol* 2013; **31**: 965–7.
111. Brown, CJ, Ballabio, A and Rupert, JL *et al.* A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* 1991; **349**: 38–44.
112. Lee, JT and Bartolomei, MS. X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell* 2013; **152**: 1308–23.
113. Nadal-Ribelles, M, Solé, C and Xu, Z *et al.* Control of Cdc28 CDK1 by a stress-induced lncRNA. *Mol Cell* 2014; **53**: 549–61.
114. Loewer, S, Cabili, MN and Guttman, M *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* 2010; **42**: 1113–7.
115. Dinger, ME, Amaral, PP and Mercer, TR *et al.* Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res* 2008; **18**: 1433–45.
116. Kretz, M, Siprashvili, Z and Chu, C *et al.* Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* 2013; **493**: 231–5.
117. Grote, P, Wittler, L and Hendrix, D *et al.* The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev Cell* 2013; **24**: 206–14.
118. Klattenhoff, CA, Scheuermann, JC and Surface, LE *et al.* Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell* 2013; **152**: 570–83.
119. Gupta, RA, Shah, N and Wang, KC *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 2010; **464**: 1071–6.
120. Yildirim, E, Kirby, JE and Brown, DE *et al.* Xist RNA is a potent suppressor of hematologic cancer in mice. *Cell* 2013; **152**: 727–42.
121. Mortazavi, A, Williams, BA and McCue, K *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008; **5**: 621–8.
122. Zhou, Y, Li, HR and Huang, J *et al.* Multiplex analysis of polyA-linked sequences (MAPS): an RNA-seq strategy to profile poly(A+) RNA. *Methods Mol Biol* 2014; **1125**: 169–78.
123. Core, LJ and Lis, JT. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science* 2008; **319**: 1791–2.
124. Paulsen, MT, Veloso, A and Prasad, J *et al.* Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response. *Proc Natl Acad Sci USA* 2013; **110**: 2240–5.
125. Sheik Mohamed, J, Gaughwin, PM and Lim, B *et al.* Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells. *RNA* 2010; **16**: 324–37.
126. Maamar, H, Cabili, MN and Rinn, J *et al.* linc-HOXA1 is a noncoding RNA that represses Hoxa1 transcription in cis. *Gene Dev* 2013; **27**: 1260–71.
127. Lecuyer, E, Yoshida, H and Parthasarathy, N *et al.* Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* 2007; **131**: 174–87.
128. Simon, MD, Wang, CI and Kharchenko, PV *et al.* The genomic binding sites of a noncoding RNA. *Proc Natl Acad Sci USA* 2011; **108**: 20497–502.
129. Simon, MD, Pinter, SF and Fang, R *et al.* High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature* 2013; **504**: 465–9.
130. Chu, C, Qu, K and Zhong, FL *et al.* Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell* 2011; **44**: 667–78.
131. Liu, W, Ma, Q and Wong, K *et al.* Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell* 2013; **155**: 1581–95.
132. Spitale, RC, Crisalli, P and Flynn, RA *et al.* RNA SHAPE analysis in living cells. *Nat Chem Biol* 2013; **9**: 18–20.
133. Wan, Y, Qu, K and Zhang, QC *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 2014; **505**: 706–9.
134. Ding, Y, Tang, Y and Kwok, CK *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* 2014; **505**: 696–700.
135. Rouskin, S, Zubradt, M and Washietl, S *et al.* Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* 2014; **505**: 701–5.
136. Yoon, JH, Srikantan, S and Gorospe, M. MS2-TRAP (MS2-tagged RNA affinity purification): tagging RNA to identify associated miRNAs. *Methods* 2012; **58**: 81–7.
137. Hogg, JR and Collins, K. RNA-based affinity purification reveals 7SK RNPs with distinct composition and regulation. *RNA* 2007; **13**: 868–80.

138. Mali, P, Esvelt, KM and Church, GM. Cas9 as a versatile tool for engineering biology. *Nat Methods* 2013; **10**: 957–63.
139. Castello, A, Fischer, B and Eichelbaum, K *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* 2012; **149**: 1393–406.
140. Lunde, BM, Moore, C and Varani, G. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol* 2007; **8**: 479–90.
141. Davidovich, C, Zheng, L and Goodrich, KJ *et al.* Promiscuous RNA binding by Polycomb repressive complex 2. *Nat Struct Mol Biol* 2013; **20**: 1250–7.
142. Konig, J, Zarnack, K and Luscombe, NM *et al.* Protein-RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet* 2011; **13**: 77–83.
143. Campbell, ZT, Bhimsaria, D and Valley, CT *et al.* Cooperativity in RNA-protein interactions: global analysis of RNA binding specificity. *Cell Rep* 2012; **1**: 570–81.
144. Ray, D, Kazan, H and Chan, ET *et al.* Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol* 2009; **27**: 667–70.
145. Xiao, R, Tang, P and Yang, B *et al.* Nuclear matrix factor hnRNP U/SAF-A exerts a global control of alternative splicing by regulating U2 snRNP maturation. *Mol Cell* 2012; **45**: 656–68.
146. Li, JH, Liu, S and Zhou, H *et al.* StarBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* 2014; **42**: D92–7.
147. Huarte, M, Guttman, M and Feldser, D *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* 2010; **142**: 409–19.
148. Korostelev, A, Trakhanov, S and Laurberg, M *et al.* Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell* 2006; **126**: 1065–77.
149. Selmer, M, Dunham, CM and Murphy, FV *et al.* Structure of the 70S ribosome complexed with mRNA and tRNA. *Science* 2006; **313**: 1935–42.
150. Wang, Y, Juranek, S and Li, H *et al.* Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. *Nature* 2009; **461**: 754–61.
151. Wang, Y, Juranek, S and Li, H *et al.* Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature* 2008; **456**: 921–6.
152. Wang, Y, Sheng, G and Juranek, S *et al.* Structure of the guide-strand-containing argonaute silencing complex. *Nature* 2008; **456**: 209–13.
153. Li, L and Ye, K. Crystal structure of an H/ACA box ribonucleoprotein particle. *Nature* 2006; **443**: 302–7.
154. Lin, PC and Xu, RM. Structure and assembly of the SF3a splicing factor complex of U2 snRNP. *EMBO J* 2012; **31**: 1579–90.
155. Sauvageau, M, Goff, LA and Lodato, S *et al.* Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* 2013; **2**: e01749.
156. Wei, C, Liu, J and Yu, Z *et al.* TALEN or Cas9—rapid, efficient and specific choices for genome modifications. *J Genet Genomics* 2013; **40**: 281–9.
157. Lee, HY, Haurwitz, RE and Apffel, A *et al.* RNA-protein analysis using a conditional CRISPR nuclease. *Proc Natl Acad Sci USA* 2013; **110**: 5416–21.
158. Qi, LS, Larson, MH and Gilbert, LA *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* 2013; **152**: 1173–83.
159. Larson, MH, Gilbert, LA and Wang, X *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat Protoc* 2013; **8**: 2180–96.
160. Gilbert, LA, Larson, MH and Morsut, L *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* 2013; **154**: 442–51.
161. Chen, B, Gilbert, LA and Cimini, BA *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* 2013; **155**: 1479–91.
162. Faghihi, MA, Modarresi, F and Khalil, AM *et al.* Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat Med* 2008; **14**: 723–30.
163. Gong, C and Maquat, LE. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature* 2011; **470**: 284–8.
164. Popp, MW and Maquat, LE. Organizing principles of mammalian nonsense-mediated mRNA decay. *Annu Rev Genet* 2013; **47**: 139–65.
165. Chew, GL, Pauli, A and Rinn, JL *et al.* Ribosome profiling reveals resemblance between long non-coding RNAs and 5′′ leaders of coding RNAs. *Development* 2013; **140**: 2828–34.
166. Popp, MW and Maquat, LE. The dharma of nonsense-mediated mRNA decay in Mammalian cells. *Mol Cells* 2014; **37**: 1–8.
167. Zund, D, Gruber, AR and Zavolan, M *et al.* Translation-dependent displacement of UPF1 from coding sequences causes its enrichment in 3′ UTRs. *Nat Struct Mol Biol* 2013; **20**: 936–43.
168. Halley, P, Kadakkuzha, BM and Faghihi, MA *et al.* Regulation of the apolipoprotein gene cluster by a long noncoding RNA. *Cell Rep* 2014; **6**: 222–30.
169. Engreitz, JM, Pandya-Jones, A and McDonel, P *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* 2013; **341**: 1237973.
170. Hu, Q, Kwon, YS and Nunez, E *et al.* Enhancing nuclear receptor-induced transcription requires nuclear motor and LSD1-dependent gene networking in interchromatin granules. *Proc Natl Acad Sci USA* 2008; **105**: 19199–204.
171. Nunez, E, Fu, XD and Rosenfeld, MG. Nuclear organization in the 3D space of the nucleus—cause or consequence? *Curr Opin Genet Dev* 2009; **19**: 424–36.
172. Gerstein, MB, Kundaje, A and Hariharan, M *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* 2012; **489**: 91–100.
173. Bu, D, Yu, K and Sun, S *et al.* NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res* 2012; **40**: D210–15.

Cell

Leading Edge
Review

Long Noncoding RNAs: Cellular Address Codes in Development and Disease

Pedro J. Batista[1] and Howard Y. Chang[1,*]
[1]Howard Hughes Medical Institute and Program in Epithelial Biology, Stanford University School of Medicine, Stanford, CA 94305, USA
*Correspondence: howchang@stanford.edu
http://dx.doi.org/10.1016/j.cell.2013.02.012

In biology as in real estate, location is a cardinal organizational principle that dictates the accessibility and flow of informational traffic. An essential question in nuclear organization is the nature of the address code—how objects are placed and later searched for and retrieved. Long noncoding RNAs (lncRNAs) have emerged as key components of the address code, allowing protein complexes, genes, and chromosomes to be trafficked to appropriate locations and subject to proper activation and deactivation. lncRNA-based mechanisms control cell fates during development, and their dysregulation underlies some human disorders caused by chromosomal deletions and translocations.

Introduction

From a single cell to an entire organism, spatial positioning is a key problem in biology. It is well appreciated that robust systems sort and distribute macromolecules, a property essential for the function of cells and tissues (Shevtsov and Dundr, 2011; Wolpert, 2011). A historical example illustrates the general utility of spatial organization. As the Roman Empire expanded and the Romans were faced with the need to construct cities in new lands, they developed a city prototype that included a group of answers to the many practical problems related to the creation and maintenance of a city (Figure 1A). This was a universal plan of simple execution. City walls protected the citizens from attack and delimited the city. At the center stood the forum, where the business and political activities of the city were concentrated. Fountains were placed throughout the city to supply water, and other spaces, such as amphitheaters, temples, and baths, were dedicated to organize daily activities. Thus, a group of structures analogous in function was always present in an organization that follows the original prototype (Grimal and Woloch, 1983).

Just like the Roman city, the nucleus of the eukaryotic cell is a highly organized space (Figure 1B). Evolution gave rise to a "nuclear" prototype that provides answers to the many challenges the cell has to respond to maintain homeostasis and growth, though subject to developmental specialization (Solovei et al., 2009). Chromosomes are not randomly organized in the nucleus, and during interphase, each chromosome occupies a discrete territory (reviewed in Cremer and Cremer, 2010). Furthermore, whereas the densely compacted heterochromatin is localized at the nuclear envelope, euchromatin localizes to the interior regions of the nucleus. Gene expression is also localized and occurs mostly at nuclear center. In addition, active genes that are coregulated are often found forming clusters. During development, individual loci such as immunoglobulin or *Hox* genes are known to change position within the nucleus according to their transcriptional status (reviewed in Misteli, 2007). Large portions of the genome are partitioned into topological domains of chromatin interaction ranging from hundreds of kilobases to megabases (the resolution of current methods), within which the genes tend to be more coregulated (Dixon et al., 2012; Nora et al., 2012). The complex task of gene expression—ensuring the proper timing, space, and rate of expression—involves noncoding regions of the genome, chromatin modifications, and the arrangement of chromosomes and nuclear domains. Here, we review the evidence that lncRNAs are a rich source of molecular addresses in the eukaryotic nucleus.

Biogenesis and Characteristics

Efforts over the last decade revealed that a large fraction of the noncoding genome is transcribed. Extensive annotation of lncRNA has been performed in multiple model organisms (reviewed in Rinn and Chang, 2012), and there is now evidence that, whereas 2% of the genome encodes for proteins (IHGSC, 2004), primary transcripts cover 75% of the human genome, with processed transcripts covering 62.1% of the genome (Djebali et al., 2012). In this Review, we focus on a particular class of noncoding transcripts known as long noncoding RNAs (lncRNAs) and the roles that they play in nuclear organization.

lncRNAs are currently defined as transcripts of greater than 200 nucleotides without evident protein coding function (Rinn and Chang, 2012). It is important to note that lncRNA is a broad definition that encompasses different classes of RNA transcripts, including enhancer RNAs, small nucleolar RNA (snoRNA) hosts, intergenic transcripts, and transcripts overlapping other transcripts in either sense or antisense orientation. lncRNAs predominantly localize to the nucleus and have, on average, a lower level of expression than protein coding genes, although details vary for different classes (Djebali et al., 2012; Ravasi et al., 2006). Multiple studies have shown that lncRNA expression is more cell type specific than protein-coding genes (Cabili et al., 2011; Djebali et al., 2012; Ravasi et al., 2006). At the DNA and

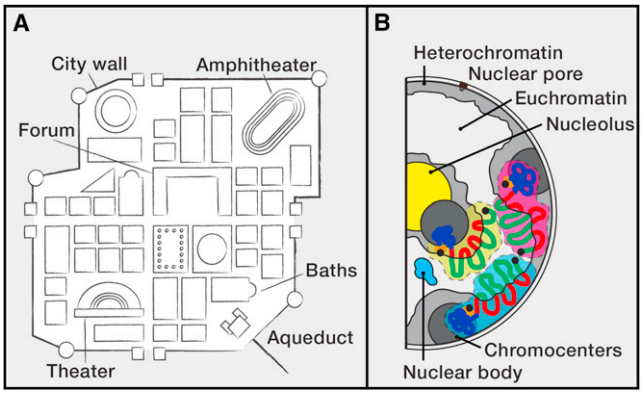


Figure 1. Comparison between a Roman City and the Cell Nucleus Reveals the Importance of Spatial Organization

(A) Depiction of the basic features of a Roman city. City walls delimit the city, with gates at the two main roads that intersect at the center of the city. The Forum was the business and political center of the city, and many buildings provided specific functions that were essential for city life.

(B) Schematic representation of the typical nuclear organization during interphase. Each chromosome occupies a discrete territory. Euchromatin localizes to the interior regions of the nucleus, and the densely compacted heterochromatin localizes near the nuclear envelope. Many specialized functions are executed in distinct regions in the nucleus, known as nuclear bodies. One example is the nucleolus, where ribosomes are assembled. Adapted from Solovei et al., 2009.

chromatin level, lncRNA loci are similar to mRNA loci, but lncRNAs show a bias for having just one intron and a trend for less-efficient cotranscriptional splicing (Derrien et al., 2012; Tilgner et al., 2012). Although lncRNAs are under lower selective pressure than protein-coding genes, sequence analysis shows that lncRNAs are under higher selective pressures than ancestral repeat sequences, which are considered to be under neutral selection. Interestingly, the promoters of lncRNAs are the region of the lncRNA gene under higher selective pressure, displaying levels of selection comparable to the promoters of protein-coding genes (Derrien et al., 2012; Guttman et al., 2009; Marques and Ponting, 2009; Ørom et al., 2010; Ponjavic et al., 2007). This analysis has also revealed a high number of correlated positions between lncRNA in sequence alignments, an observation that fits the hypothesis that lncRNAs are under selective pressure to maintain a functional RNA structure (Derrien et al., 2012). Comparison between mammalian and zebrafish lncRNAs revealed that short stretches of conserved sequence are functionally important and that location and structure of lncRNAs can be conserved, even in the absence of strong sequence conservation. The ability to induce a loss-of-function phenotype by blocking the short conserved motif in addition to the ability to rescue loss of function of two lncRNAs with the addition of human and mouse lncRNAs (Ulitsky et al., 2011) demonstrates that these "in silico" observations are of biological significance.

Sequence analysis of lncRNAs, focusing on presence and size of open reading frames as well as codon conservation frequency, has been used to exclude protein coding potential. Ribosome profiling, a method that enumerates transcripts associated with ribosomes, had detected many lncRNAs, but it was unclear whether these lncRNAs are just being scanned similarly to 5′ untranslated regions or actually are productively engaged in translation (Ingolia et al., 2011). Comparison of RNA sequencing (RNA-seq) data to tandem mass spectrometry data for two cell lines suggests that ~92% of the annotated lncRNAs do not yield detectable peptides in these cell lines (Bánfai et al., 2012; Derrien et al., 2012). Although the differences between these two studies may stem from measuring two different endpoints, they suggest that lncRNAs have low translational potential even when ribosomes attempt to decode them. Current annotations suggest that the actual number of lncRNAs exceeds that of protein coding genes (Derrien et al., 2012).

The repertoire of roles performed by lncRNAs is growing, as there is now evidence that lncRNAs participate in multiple networks regulating gene expression and function. Several characteristics of lncRNAs make them the ideal system to provide the nucleus with a system of molecular addresses. lncRNAs, unlike proteins, can function both in *cis*, at the site of transcription, or in *trans*. An RNA-based address code may be deployed more rapidly and economically than a system that relies only on proteins. lncRNAs do not need to be translated and do not require transport between the cytoplasm and the nucleus. lncRNAs can also interact with multiple proteins, enabling scaffolding functions and combinatorial control (Wang and Chang, 2011). As such, the act of transcription can rapidly create an anchor that will lead to the formation, or remodeling, of nuclear domains through the recruitment or sequestration of proteins already present in the nuclear compartment. Using lncRNAs allows cells to create addresses that are regional-, locus- or even allele-specific (Lee, 2009). At the regional level, lncRNAs can influence the formation of nuclear domains and the transcriptional status of an entire chromosome, and they can participate in the interaction of two different chromosomal regions. At a more fine-grained level, lncRNAs can control the chromatin state and activity of a chromosomal locus or specific gene. We explore each of these concepts below with recently published examples.

Locus Control of Gene Regulation

Cells can use noncoding RNAs to modulate gene expression by changing the accessibility of gene promoters. These mechanisms can be used to fine-tune gene expression in response to environmental conditions or to silence a gene as part of a developmental program.

First, the act of noncoding RNA (ncRNA) transcription itself can be purposed for regulatory function. For example, transcription through a regulatory sequence, such as a promoter, can block its function, a mechanism termed transcriptional interference (Figure 2A) first identified in yeast (Martens et al., 2004). In such instances, the lncRNA promoter is finely tuned to receive appropriate inputs to exert regulatory function; the lncRNA product is typically a faithful biomarker of transcriptional interference in action but is not required for its success. In conditions that limit vegetative growth, diploid *S. cerevisiae* cells enter sporulation, a differentiation program that results in the formation of haploid daughter cells. Entry into meiosis has catastrophic consequences in haploid cells and is therefore inhibited via a transcriptional interference mechanism. A transcription factor in haploid cells activates the expression of IRT1(SUT643),

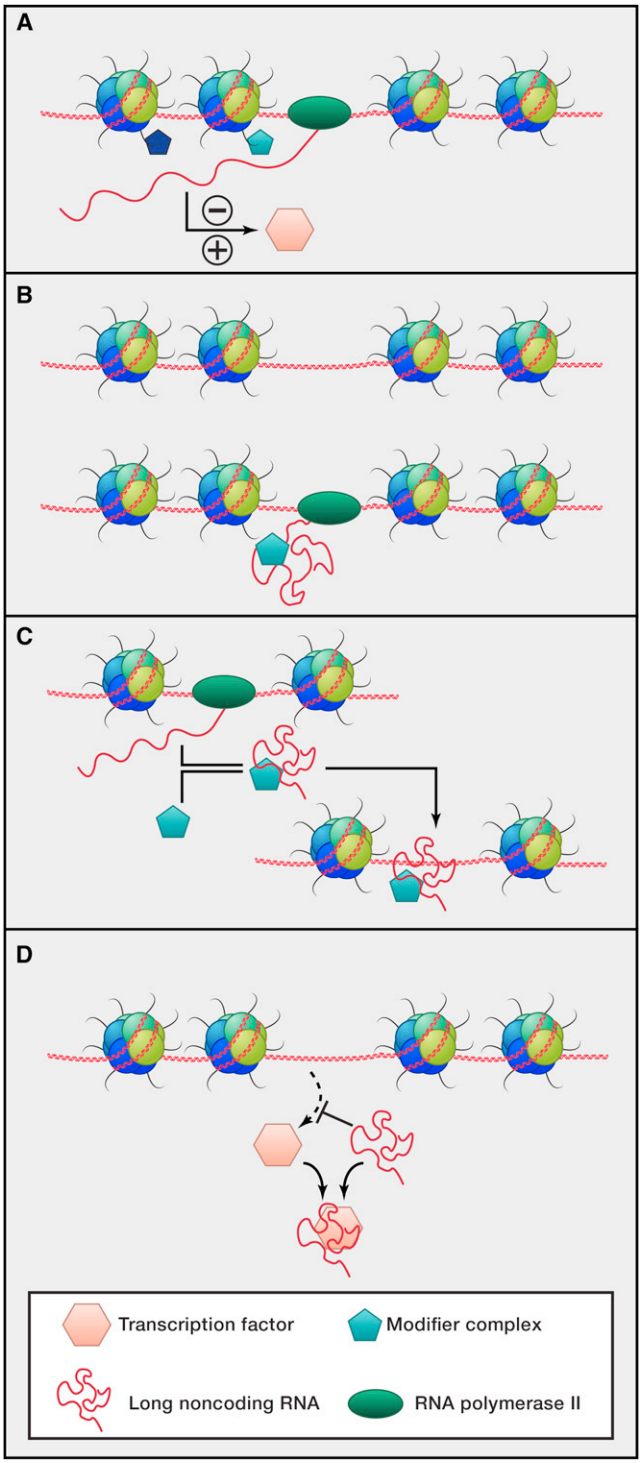


Figure 2. Functional Modules of lncRNAs in the Nucleus

(A) The act of transcription at noncoding regions can modulate gene expression through the recruitment of chromatin modifiers to the site of transcription. These complexes can create a local chromatin environment that facilitates or blocks the binding of other regulators.

(B) lncRNAs can function in *cis*, recruiting protein complexes to their site of transcription and thus creating a locus-specific address. Cells can use this mechanism to repress or activate gene expression.

(C) lncRNAs can function in *trans* and recruit protein complexes to chromatin loci away from their site of transcription.

(D) lncRNAs can bind and sequester transcription factors away from their target chromosomal regions.

a noncoding RNA that overlaps the promoter of *IME1*, the master regulator of sporulation. Transcription of IRT1 establishes a repressive chromatin state at the *IME1* promoter through the recruitment of histone methyltransferase Set2 and the histone deacetylase Set3 (van Werven et al., 2012). The use of noncoding transcription to control chromatin modification is a widespread strategy. The Set3 histone deacetylase has also been implicated in the modulation of gene induction kinetics during changes of carbon source. Transcription of ncRNAs that overlap the regulated genes leads to the establishment of H3K4me2, which recruits Set3 and leads to the deacetylation of the gene promoter. Deacetylation of the promoter results in delayed or reduced induction of the regulated genes. This mechanism is also involved in the inhibition of cryptic promoters (Kim et al., 2012). Expression of GAL10-ncRNA, driven by Reb1, leads to deacetylation across the GAL1-10 promoter, facilitating glucose repression of GAL1-10 (Houseley et al., 2008).

In mammalian imprinting, the noncoding RNA Air (also known as Airn) is expressed from the paternal chromosome and is involved in silencing the paternal alleles of multiple genes. The promoter of one of these genes, *Igf2r*, overlaps with the *Air* transcriptional unit and is silenced by transcriptional interference (Latos et al., 2012).

Transcriptional interference can also be used to activate gene expression by inhibiting the action of repressor elements, functioning as an antisilencing mechanism. In *Drosophila* embryogenesis, transcription through Polycomb response elements (PRE) alters the function of these elements, blocking the establishment of repressive chromatin (Schmitt et al., 2005).

Second, lncRNAs can silence or activate gene expression in *cis*, acting on neighboring genes of the lncRNA locus. Some of the first studied examples of lncRNA function involve dosage compensation and genomic imprinting, whereby lncRNAs provide allele-specific gene regulation to differentially control two copies of the same gene within one cell (see the Review by Lee and Bartolomei on page 1308 of this issue; Lee and Bartolomei, 2013) (Figure 2B). Several such lncRNAs are now recognized to interact with and recruit histone modification complexes, including Xist (recruits PRC2 for H3K27me3 and RYBP-PRC1 for H2A ubiquitylation) and Kcnq1ot1 (recruits G9a for H3K9me3 and PRC2) (Pandey et al., 2008; Tavares et al., 2012; Zhao et al., 2010). The Air lncRNA (the transcription of which inhibits *Igfr2*) targets G9a and H3K9me3 to silence more distantly located genes on the paternal chromosome (Nagano et al., 2008); hence, one lncRNA gene can employ multiple mechanisms to regulate nearby and distantly located genes. In genome-wide studies, numerous lncRNAs have now been found to interact with chromatin modification complexes (Guil et al., 2012; Guttman et al., 2011; Khalil et al., 2009; Zhao et al., 2010). In the plant *A. thaliana*, two cold-inducible lncRNAs, COOLAIR and COLDAIR, are embedded antisense or intronic to the flowering control locus gene *FLC*, and they help to recruit PRC2 to stably silence *FLC* in a cold-dependent manner, a key

mechanism to ensure the proper flowering time after winter termed "vernalization" (reviewed in Ietswaart et al., 2012). In an analogous fashion, DNA damage induces a lncRNA from the promoter of cyclin D1 gene (*CCND1*); this lncRNA binds to TLS protein to allosterically inhibit histone acetyltransferase in *cis*, which suppresses *CCND1* transcription (Wang et al., 2008).

DNA methylation can occur as a long-term silencing mechanism downstream of repressive histone modifications, and lncRNAs may also guide DNA methylation in addition to histone modification. The ribosomal DNA (rDNA) loci are tandemly repeated in the genome, with some copies being transcriptionally active, whereas others are silenced by DNA methylation and histone modifications. Each ribosomal DNA transcribes rRNA separated by intergenic spacers (IGSs) as a polycistronic unit, and IGSs can be processed to 150–250 nt fragments termed "promoter RNAs (pRNAs)" (reviewed in Bierhoff et al., 2010). pRNA serves as a platform to recruit the de novo cytosine methylase DNMT3 and the NoRC complex containing poly-ADP ribose polymerase-1 (PARP-1) to promote silencing of rDNA (Guetg et al., 2012; Mayer et al., 2006). Notably, a stretch of 20 nt in pRNA binds the rDNA promoter, forming a RNA:DNA:DNA triplex (Schmitz et al., 2010). This triplex structure is proposed to recruit DNMT3 and also serves as the specific recognition mechanism between lncRNA and genomic DNA—a model that likely applies to other lncRNA-DNA interactions (Martianov et al., 2007).

A distinct family of lncRNAs serves to activate gene expression. Many active enhancer elements transcribe lncRNAs, termed "eRNAs" (De Santa et al., 2010; Kim et al., 2010), and several lncRNAs are required to activate gene expression, which are termed "enhancer-like RNAs" (Ørom et al., 2010). Evf is a *cis*-acting lncRNA that is required for the activation of *Dlx5/6* genes and generation of GABAergic interneurons in vivo (Bond et al., 2009). A key mechanism of lncRNA specificity in *cis* is the higher-order chromosomal configuration (Wang et al., 2011). The noncoding RNA HOTTIP is expressed from the 5′ end tip of the *HoxA* locus and drives histone H3 lysine 4 trimethylation and gene transcription of *HoxA* distal genes through the recruitment of the WDR5/MLL complex (Wang et al., 2011). Endogenous HOTTIP is brought to its target genes by chromosomal looping, and ectopic HOTTIP only activates transcription when it is artificially tethered to the reporter gene (Wang et al., 2011). The MLL complex is also recruited to the *Hox* locus by the noncoding RNA Mistral, located between *Hoxa6* and *Hoxa7*. Mistral directly interacts with MLL1, leading to changes at the chromatin level that activate *Hoxa6* and *Hoxa7* (Bertani et al., 2011). Hence, lncRNA interaction with MLL/Trx complexes and likely additional proteins will define their function in enforcing active chromatin states and gene activation.

Third, lncRNAs can control chromatin states at distantly located genes (i.e., in *trans*) for both gene silencing and activation (Figure 2C). These lncRNAs bind to some of the same effector chromatin modification complexes but target them to genomic loci genome-wide. For instance, human HOTAIR lncRNA binds to PRC2 and LSD1 complexes and couples H3K27 methylation and H3K4 demethylation activity to hundreds of sites genome-wide (Chu et al., 2011; Tsai et al., 2010). HOTAIR is located in the *HOXC* locus and is regulated in an anatomic position-specific fashion. Linc-p21 is induced by p53 during DNA damage and recruits hnRNPK via physical interaction to mediate p53-mediated gene repression (Huarte et al., 2010). Linc-p21 also has a recently recognized role in translational control (Yoon et al., 2012). In contrast, PANDA, another lncRNA induced by p53, acts as a decoy by binding to the transcription factor NF-YA and preventing NF-YA from activating genes encoding cell death proteins (Hung et al., 2011) (Figure 2D). lncRNA-mediated activation can also occur in *trans*. Jpx, an X-linked lncRNA that activates Xist expression, is important for X chromosome inactivation in female cells, and Jpx deletion can be rescued by Jpx supplied in *trans* (Tian et al., 2010).

Nuclear Domains

The concept of lncRNA recruitment of factors to genes may be more properly considered a two-way street, with genes being moved into specific cytotopic locations by lncRNAs. One type of molecular address can be found in the formation of nuclear domains. These are regions of the nucleus where specific functions are performed. Unlike cellular organelles, these domains are not membrane delimited. They are instead characterized by the components that form them. These domains are believed to form through molecular interactions between its components. Once a stable interaction is found, the components remain associated. These domains are often formed around the sites of transcription of RNA components, which function as molecular anchors (reviewed in Dundr and Misteli, 2010). The noncoding RNA NEAT1, an essential component of the Paraspeckle, is a well-characterized example of how noncoding RNAs can function as structural components of nuclear bodies. Upon transcription of NEAT1, diffusible components of this domain nucleate at the site of NEAT1 accumulation, leading to the formation of the Paraspeckle (Figure 3A) (Chen and Carmichael, 2009; Clemson et al., 2009; Mao et al., 2011; Sasaki et al., 2009; Shevtsov and Dundr, 2011; Sunwoo et al., 2009).

Nuclear domains can be dynamically regulated in an RNA-dependent fashion. In response to serum stimulation, the demethylase KDM4C is recruited to the promoters of genes controlled by the cell-cycle-specific transcription factor E2F, where it demethylates Polycomb protein Pc2. Whereas methylated Pc2 interacts with the noncoding RNA TUG1, a component of Polycomb bodies, unmethylated Pc2 interacts with the noncoding RNA MALAT1/NEAT2, a component of interchromatin granules. Therefore, changes in the methylation status of Pc2 lead to the relocation of growth control genes from an environment that inhibits gene expression, the Polycomb body, to a domain that is permissive of gene expression, the interchromatin granule (Figure 3B). Interestingly, the reading ability of Pc2 is modulated by the noncoding RNA that it is interacting with. When bound to TUG1, Pc2 reads H4R3me^{2s} and H3K27me^{2}, whereas it reads H2AK5ac and H2AK13ac when interacting with MALAT1/NEAT2 (Yang et al., 2011). These interplays control the growth-factor-dependent expression of cell-cycle genes in vitro, but it came as a surprise that mouse knockouts of either NEAT1 or MALAT1/NEAT2 had no little overt phenotype (Eissmann et al., 2012; Nakagawa et al., 2012; Nakagawa et al., 2011; Zhang et al., 2012). Clearly, the question of redundancy or compensation in vivo needs to be addressed in the future.

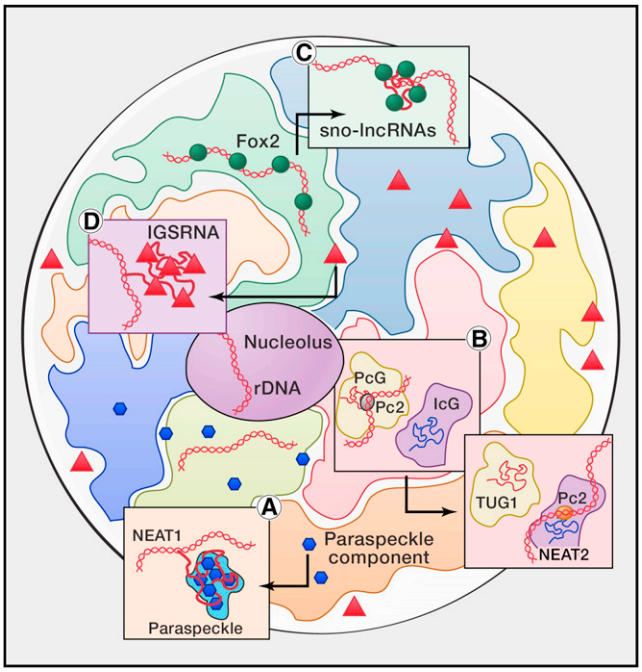


Figure 3. Schematic Representation of the Cell Nucleus, Showing the Nucleolus and Chromosomal Territories

(A) Protein components of the Paraspeckle diffused throughout the nucleoplasm aggregate upon the transcription of NEAT1, forming the Paraspeckle nuclear domain.

(B) Pc2 differentially binds MALAT1/NEAT2 or TUG1 depending on methylation status. Methylated Pc2 interacts with TUG1, bringing associated growth control genes to a repressive environment, the polycomb body (PcG). Unmethylated Pc2 interacts with MALAT1/NEAT2 at the interchromatin granule (ICG), where gene expression is permitted.

(C) Expression of lncRNAs with snoRNA ends from the Prader-Willi syndrome locus functions as a sink for the FOX2 protein, leading to redistribution of this splicing factor in this nuclear region.

(D) In response to cellular stress, transcription of specific IGSRNAs leads to the retention of targeted proteins at the nucleolus. Different types of stress lead to the retention of different proteins through the expression of specific noncoding RNAs.

Unusual processing mechanisms may explain the localization activity of certain lncRNAs. An imprinted region in chromosome 15 (15q11-q13) that had been implicated in Prader-Willi syndrome (PWS) hosts multiple intron-derived lncRNAs with small nucleolar RNAs at their ends—so called "sno-lncRNAs." It is probable that the presence of structured snoRNAs at the ends of lncRNAs stabilizes these molecules, which have no 5′ cap or polyA tail. These RNAs are retained in the nucleus and localize to, or remain near, their sites of transcription. Knockdown of sno-lncRNAs has little effect on the expression of nearby genes, suggesting that it does not affect gene expression in *cis*. Instead, these sno-lncRNAs seem to create a "domain" where the splicing factor Fox2 is enriched. These sno-lncRNAs contain multiple binding sites for Fox2, and altering the level of sno-lncRNAs led to a redistribution of Fox2 in the nucleus and changes in mRNA splicing patterns. Hence, the sno-lncRNAs appear to function as Fox2 sinks, participating in the regulation of splicing in specific subnuclear domains (Yin et al., 2012) (Figure 3C). Similarly, formation of a blunt-ended triplex RNA structure at the 3′ end of MALAT1/NEAT2 lncRNA, which lacks a polyA tail, stabilizes the lncRNA and presumably limits its export to the cytoplasm (Brown et al., 2012; Wilusz et al., 2012). Viral nuclear lncRNAs have also adapted this strategy and hide their 3′ polyA tails in a triplex RNA structure to prevent decay (Mitton-Fry et al., 2010; Tycowski et al., 2012).

Gene Control through Sequestration

In contrast to the model of nuclear domains that concentrate and thereby facilitate molecular interactions, spatial control can also separate reactants until the moment is right. For example, certain environmental stresses trigger the retention of select proteins in the nucleolus away from their normal site of action. The retention at the nucleolus requires a signal sequence and the expression of specific noncoding RNAs expressed from the large intergenic spacer (IGS) of the rDNA repeats. IGS ncRNAs turn out to gate the responses to cellular stress. Unique IGS ncRNAs are transcriptionally induced by specific stressors, functioning as baits for proteins with specific signal sequences. Interfering with a specific IGSRNA does not affect the function of other IGSRNAs (Audas et al., 2012) (Figure 3D).

In *S. pombe*, both mRNAs and lncRNAs function together to form heterochromatin and sequester genes in the control of meiosis. During vegetative growth, the expression of meiotic genes is repressed through selective elimination of meiotic mRNAs. Meiotic genes contain within their transcripts a region known as determinant of selective removal (DSR) that determines their degradation. This sequence is recognized by Mmi1, which promotes both mRNA degradation (Harigaya et al., 2006) as well as formation of facultative heterochromatic islands (Zofall et al., 2012). Hence, aberrant nascent mRNAs can function in an lncRNA-like fashion to tether the formation for heterochromatin. Furthermore, during vegetative growth, Mei2p, an RNA-binding protein that is crucial for entry in meiosis, is kept in an inactive form. When cells commit to the meiosis expression program, Mei2p accumulates in its active form and sequesters Mmi1 to a structure known as Mei2 dot, where Mmi1 function is inhibited. The Mei2 dot forms at the *sme2* locus at the site of transcription of two noncoding RNAs, meiRNA-S and meiRNA-L, which are necessary for the formation of the Mei2 dot structure and, therefore, entry in meiosis (Yamamoto, 2010).

Higher-Order Chromosomal Interactions

An intriguing possibility is that lncRNAs can regulate the three-dimensional structure of the chromosomes by facilitating the interaction of specific chromosomal loci. The act of transcription itself can influence gene expression and genome organization by promoting chromatin modifications, by recruiting gene active regions to common transcription factories, or by exposing the DNA strands to enzymatic activity. Hence, the presence of multiple lncRNA genes in a region may help chromosomal loci adopt distinct conformation with transcriptional activation. For example, in the *Hox* loci, collinear expression of *Hox* mRNA genes and *Hox* lncRNAs along the chromosome is associated with the progressive recruitment of those chromosomal segments into a tightly interacting domain that is distinct from the transcriptionally silent portion of the loci (Noordermeer

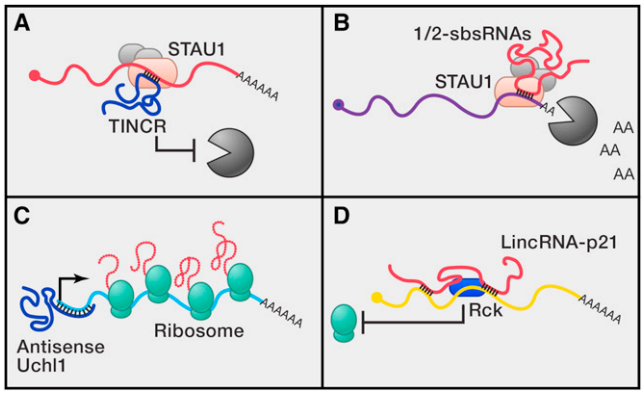


Figure 4. lncRNAs Regulate Gene Expression in the Cytoplasm
(A) The lncRNA TINCR interacts with STAU1 and target mRNAs containing the TINCR box motif, promoting their stability.
(B) lncRNAs of the 1/2-sbsRNAs class hybridize with 3′-UTR-containing Alu elements and promote the degradation of these target mRNAs.
(C) Under stress conditions, the lncRNA antisense to *Uchl1* moves from the nucleus to the cytoplasm and binds the 5′ end of the *Uchl1* mRNA to promote its translation under stress conditions.
(D) lincRNA-p21 interacts with and targets RcK to mRNAs, resulting in translation inhibition.

et al., 2011; Wang et al., 2011). A similar phenomenon was first appreciated in the β-globin locus, and intergenic transcripts from its locus control regions (Ashe et al., 1997). Transcription-coupled looping is likely to be related to the fact that the Mediator complex that links transcription factors to basal transcription machinery promotes long-range enhancer-promoter interactions (Kagey et al., 2010). A similar transcription-directed mechanism has also been proposed to guide DNA recombination of lymphocyte receptor genes over megabases (Verma-Gaur et al., 2012). The lncRNA transcripts are useful readouts of the chromosomal configuration but are not necessarily required for the chromosomal interactions.

lncRNAs can also regulate chromosome structure through direct mechanisms. High-throughput chromosomal conformation assays revealed that the active and inactive X chromosomes adopt quite distinct conformations. The inactive X (Xi) is coated by the Xist lncRNA, which is required for choosing the inactive X chromosome. Importantly, conditional knockout of Xist has demonstrated that the folding of inactive X requires the Xist RNA. After Xist deletion, the Xi chromosome adopts a conformation that is more similar to that of the active X chromosome (Xa) without reactivation of Xi gene expression. Hence, Xist appears to regulate X chromosome structure through mechanisms other than the relocation of active genes to transcriptional factories (Splinter et al., 2011). One intriguing clue is that conditional Xist deletion also led to loss of PRC2 and H3K27me3 marks. The conformations of the two X chromosomes appear to be regulated by distinct mechanisms because PRC2 is dispensable for the topological domains of Xa (Nora et al., 2012). Whether one or several Xa-expressed lncRNA controls Xa conformation remains to be seen.

lncRNAs can also regulate the interaction between chromosomes, a concept that is exemplified by *S. pombe* meiosis. In order for chromosomes to properly segregate in meiosis and prevent aneuploidy, homologous chromosomes must interact and generate stable associations. The *sme2* locus plays a key role in the mutual identification of homologous chromosomes during meiosis, in addition to its role in the mitosis/meiosis switch discussed above. The meiRNA-L transcript accumulates at the *sme2* locus and is necessary for the robust chromosomal pairing (Ding et al., 2012). These studies suggest that noncoding RNAs can be components of a *cis*-acting pairing factor that allows homologous chromosomes to identify each other.

Cytoplasmic Functions

The ultimate function of mRNAs is to be translated, and like other steps of gene expression, multiple layers of posttranscriptional regulation exist in the cytoplasm (Figure 4). lncRNAs can also "identify" mRNAs in the cytoplasm and modulate their life cycle. Recent works demonstrated that lncRNAs impact both the mRNA half-life and translation of mRNAs. The lncRNA TINCR (terminal differentiation-induced ncRNA) is induced during epidermal differentiation and is required for normal induction of key mediators of epidermal differentiation. TINCR localizes to the cytoplasm, where it interacts with Staufen 1 protein (STAU1) to promote the stability of mRNAs containing the TINCR box motif (Kretz et al., 2013) (Figure 4A). Hence, the TINCR mechanism is the diametric opposite of posttranscriptional silencing by small regulatory RNAs like siRNA or miRNAs. STAU1 can also be programmed by other lncRNAs to facilitate mRNA degradation. The half-STAU1-binding site RNAs (1/2-sbsRNAs) contain Alu elements that bind to Alu elements in the 3′UTR of actively transcribed target genes, generating a STAU1-binding site. These mRNAs are therefore identified as STAU1-mediated messenger RNA decay (SMD) targets (Gong and Maquat, 2011) (Figure 4B). In addition, a recently identified class of lncRNA impacts gene expression by promoting translation of targets mRNAs. Expression of antisense *Uchl1* RNA leads to an increase in Uchl1 protein level without any change at the mRNA level. Antisense *Uchl1* lncRNA is composed by a region that overlaps with the first 73 nucleotides of *Uchl1* and two embedded repetitive sequences, one of which (SINEB2) is required for the ability of the lncRNA to induce protein translation. Under stress conditions in which cap-dependent translation is inhibited, antisense *Uchl1* lncRNA, previously enriched in the nucleus, moves into the cytoplasm and hybridizes with *Uchl1* mRNA to enable cap-independent translation of *Uchl1*. In other words, the lncRNA acts like a mobile internal ribosomal entry element to promote selective translation. Other SINEB2-containing antisense lncRNAs may function in a similar way (Carrieri et al., 2012) (Figure 4C). Conversely, lincRNA-p21 can inhibit the translation of target mRNAs. In the absence of HuR, lincRNA-p21 is stable and interacts with the mRNAs CTNNB1 and JUNB and translational repressor Rck, repressing the translation of the targeted mRNAs (Yoon et al., 2012) (Figure 4D). These emerging examples illustrate that lncRNAs can provide a rich palette of regulatory capacities in the cytoplasm.

Human Diseases

Considering the wide range of roles that lncRNAs play in cellular networks, it is not surprising that noncoding RNAs have been implicated in disease. Genome-wide association studies have

revealed that only 7% of disease or trait-associated single-nucleotide polymorphisms (SNPs) reside in protein-coding exons, whereas 43% of trait-/disease-associated SNP are found outside of protein-coding genes (Hindorff et al., 2009). In addition to the example of sno-lncRNAs in Prader-Willi syndrome discussed above, several recent discoveries of lncRNAs in Mendelian disorders illustrate the emerging recognition of lncRNAs in human diseases.

Facioscapulohumeral muscular dystrophy (FSHD) is the third most common myopathy and is predominantly caused by a contraction in copy number of the D4Z4 repeats mapping to 4q35. The D4Z4 repeat is the target of several chromatin modifications, including H3K9me3 and H3K27me3, which are reduced in FSHD patients. Cabianca et al. found that a long array of D4Z4 repeats recruit Polycomb complexes to promote the formation of a repressive chromatin state that inhibits the expression of genes at 4q35. Loss of D4Z4 repeats results in derepression of DBE-T, a novel lncRNA that functions in *cis* and localizes to the FSHD locus. DBE-T recruits ASH1L (a component of MLL/TrX complex), leading to improper establishment of active chromatin and expression of genes from 4q35 (Cabianca et al., 2012). Hence, DBE-T is a lncRNA that functions as a locus control element by promoting active chromatin domain, and FSHD results from lncRNA "promoter mutations" that perturb DBE-T regulation.

HELLP syndrome (*h*emolysis, *e*levated *l*iver enzymes, *l*ow *p*latelets) is a recessively inherited life-threatening pregnancy complication. Linkage analysis narrowed the HELLP locus to a gene desert between *C12orf48* and *IGF1* on 12q23.2, where a single 205 kb capped and polyadenylated lncRNA is transcribed (van Dijk et al., 2012). Knockdown of this lncRNA revealed a role in the transition from G2 to mitosis and trophoblast cell invasion, although the precise mechanism is still unclear. Notably, morpholino oligonucleotides complementary to the mutation site in HELLP lncRNA boosted lncRNA level and reversed the gene expression and cell invasion defects.

Similarly, deletions in a coding-gene desert at 16q24.1 lead to *a*lveolar *c*apillary *d*ysplasia with *m*isalignment of *p*ulmonary *v*eins (ACD/MPV) (Szafranski et al., 2013). This region contains a distant enhancer of *FOXF1*, a key regulator of lung development. This enhancer element interacts with *FOXF1* in human pulmonary microvascular endothelial cells, but not in lymphoblasts, suggesting that FOXF1 expression in the lung endothelium is regulated at the chromatin structure levels. In addition to transcription-factor-binding sites, the focal deletion includes two lncRNA expressed specifically in the lung. An intriguing possibility is that the expression of these lncRNAs, which happens specifically in the lung, contributes to the establishment of a chromatin loop that brings the enhancer in close proximity to *FOXF1*.

Chromosomal translocations lead to inheritable structural and genetic changes and, as such, are relevant causes of genetic disease. One way that chromosomal translocations can lead to disease is through disruption of the higher-order chromatin organization and the *cis*-regulatory landscape. Recently, two different translocations have been identified in brachydactyly type E (BDE) that implicate lncRNA dysregulation (Maass et al., 2012). These translocations affect a regulatory region that interacts in *cis* with *PTHLH* and in *trans* with *SOX9*. Interestingly, this region is home to a lncRNA whose expression is important for the proper expression of *PTHLH* and *SOX9*. Depletion of this lncRNA (*DA125942*) resulted in downregulation of *PTHLH* and *SOX9*. The lncRNA interacts with both loci, and the occupancy is reduced in chromatin originated from BDE patients. This study demonstrates how lncRNAs and chromatin higher-order organization collaborate in the regulation of gene expression.

Recognition of the roles of lncRNAs in human disease has unveiled new diagnostic and therapeutic opportunities. lncRNAs are expressed in a more tissue-specific fashion than mRNA genes, a pattern that has been found to hold true in pathologic states such as cancer (Brunner et al., 2012). lncRNA measurements could hence trace cancer metastases or circulating cancer cells to their origins. In addition, a strong connection between lncRNAs and cancer has been clearly established, as many lncRNAs are dysregulated in human cancers. The lncRNA HOTAIR in overexpressed in breast, colon, pancreas, and liver cancers, and overexpression of HOTAIR has been shown to drive breast cancer metastasis in vivo (Gupta et al., 2010; Gutschner and Diederichs, 2012). lncRNAs appear to be more structured and stable than mRNA transcripts, which facilitate their detection as free nucleic acids in body fluid such as urine and blood—knowledge already put to good use in clinically approved tests for prostate cancer (Fradet et al., 2004; Shappell, 2008; Tinzl et al., 2004). Aberrant lncRNAs can be knocked down in vivo using oligonucleotide "drugs" (Modarresi et al., 2012; Wheeler et al., 2012), which should spur advance in lncRNA genetics and therapeutics.

Conclusions

lncRNAs are well poised to be molecular address codes, particularly in the nucleus. On the one hand, transcription of lncRNAs is often exquisitely regulated, reflecting the particular developmental stage and external environment that the cell has experienced. On the other, the capacity of lncRNAs to function as guides, scaffolds, and decoys endows them with enormous regulatory potential in gene expression and for spatial control within the cell. These outstanding properties of long RNAs have already been leveraged to make designer RNA scaffolds for synthetic cell circuits (Delebecque et al., 2011). Many questions remain to be addressed in this rapidly expanding field. First, the in vivo function of most lncRNAs has not been determined. An extensive catalog of lncRNAs has recently been described available for several model organisms (Nam and Bartel, 2012; Pauli et al., 2012; Ulitsky et al., 2011), opening the door of a wide array of powerful techniques to be used in the in vivo study of lncRNAs that will complement the study of human lncRNAs. In addition, detailed knowledge of structure-function relationship in lncRNAs is still lacking, which prohibits the de novo prediction of lncRNA domains and functions that we take for granted in protein-coding transcripts. New technologies to deconvolute RNA structure and function (Martin et al., 2012; Wan et al., 2012), probe RNA-chromatin interactions (Chu et al., 2011; Simon et al., 2011), and track RNA movement in real time (Paige et al., 2011) will be crucial for understanding lncRNAs and realizing their therapeutic potential.

ACKNOWLEDGMENTS

We thank members of the Chang lab for discussion and apologize to colleagues whose works are not discussed due to space limitation. We acknowledge support from NIH and California Institute for Regenerative Medicine (H.Y.C.). P.J.B. is the Kenneth G. and Elaine A. Langone Fellow of the Damon Runyon Cancer Research Foundation. H.Y.C. is an Early Career Scientist of the Howard Hughes Medical Institute. H.Y.C. is on the Scientific Advisory Board of RaNA Therapeutics, which works on long noncoding RNAs.

REFERENCES

Ashe, H.L., Monks, J., Wijgerde, M., Fraser, P., and Proudfoot, N.J. (1997). Intergenic transcription and transinduction of the human beta-globin locus. Genes Dev. *11*, 2494–2509.

Audas, T.E., Jacob, M.D., and Lee, S. (2012). Immobilization of proteins in the nucleolus by ribosomal intergenic spacer noncoding RNA. Mol. Cell *45*, 147–157.

Bánfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W.E., Jr., Kundaje, A., Gunawardena, H.P., Yu, Y., Xie, L., et al. (2012). Long noncoding RNAs are rarely translated in two human cell lines. Genome Res. *22*, 1646–1657.

Bertani, S., Sauer, S., Bolotin, E., and Sauer, F. (2011). The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. Mol. Cell *43*, 1040–1046.

Bierhoff, H., Schmitz, K., Maass, F., Ye, J., and Grummt, I. (2010). Noncoding transcripts in sense and antisense orientation regulate the epigenetic state of ribosomal RNA genes. Cold Spring Harb. Symp. Quant. Biol. *75*, 357–364.

Bond, A.M., Vangompel, M.J., Sametsky, E.A., Clark, M.F., Savage, J.C., Disterhoft, J.F., and Kohtz, J.D. (2009). Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. Nat. Neurosci. *12*, 1020–1027.

Brown, J.A., Valenstein, M.L., Yario, T.A., Tycowski, K.T., and Steitz, J.A. (2012). Formation of triple-helical structures by the 3′-end sequences of MALAT1 and MENβ noncoding RNAs. Proc. Natl. Acad. Sci. USA *109*, 19202–19207.

Brunner, A.L., Beck, A.H., Edris, B., Sweeney, R.T., Zhu, S.X., Li, R., Montgomery, K., Varma, S., Gilks, T., Guo, X., et al. (2012). Transcriptional profiling of lncRNAs and novel transcribed regions across a diverse panel of archived human cancers. Genome Biol. *13*, R75.

Cabianca, D.S., Casa, V., Bodega, B., Xynos, A., Ginelli, E., Tanaka, Y., and Gabellini, D. (2012). A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. Cell *149*, 819–831.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer, I., Collavin, L., Santoro, C., et al. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature *491*, 454–457.

Chen, L.L., and Carmichael, G.G. (2009). Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. Mol. Cell *35*, 467–478.

Chu, C., Qu, K., Zhong, F.L., Artandi, S.E., and Chang, H.Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol. Cell *44*, 667–678.

Clemson, C.M., Hutchinson, J.N., Sara, S.A., Ensminger, A.W., Fox, A.H., Chess, A., and Lawrence, J.B. (2009). An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol. Cell *33*, 717–726.

Cremer, T., and Cremer, M. (2010). Chromosome territories. Cold Spring Harb. Perspect. Biol. *2*, a003889.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. *8*, e1000384.

Delebecque, C.J., Lindner, A.B., Silver, P.A., and Aldaye, F.A. (2011). Organization of intracellular reactions with rationally designed RNA assemblies. Science *333*, 470–474.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Ding, D.Q., Okamasa, K., Yamane, M., Tsutsumi, C., Haraguchi, T., Yamamoto, M., and Hiraoka, Y. (2012). Meiosis-specific noncoding RNA mediates robust pairing of homologous chromosomes in meiosis. Science *336*, 732–736.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Dundr, M., and Misteli, T. (2010). Biogenesis of nuclear bodies. Cold Spring Harb. Perspect. Biol. *2*, a000711.

Eissmann, M., Gutschner, T., Hämmerle, M., Günther, S., Caudron-Herger, M., Groß, M., Schirmacher, P., Rippe, K., Braun, T., Zörnig, M., and Diederichs, S. (2012). Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development. RNA Biol. *9*, 1076–1087.

Fradet, Y., Saad, F., Aprikian, A., Dessureault, J., Elhilali, M., Trudel, C., Mâsse, B., Piché, L., and Chypre, C. (2004). uPM3, a new molecular urine test for the detection of prostate cancer. Urology *64*, 311–315, discussion 315–316.

Gong, C., and Maquat, L.E. (2011). lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. Nature *470*, 284–288.

Grimal, P., and Woloch, M. (1983). Roman cities = Les villes romaines (Madison, Wis.: University of Wisconsin Press).

Guetg, C., Scheifele, F., Rosenthal, F., Hottiger, M.O., and Santoro, R. (2012). Inheritance of silent rDNA chromatin is mediated by PARP1 via noncoding RNA. Mol. Cell *45*, 790–800.

Guil, S., Soler, M., Portela, A., Carrère, J., Fonalleras, E., Gómez, A., Villanueva, A., and Esteller, M. (2012). Intronic RNAs mediate EZH2 regulation of epigenetic targets. Nat. Struct. Mol. Biol. *19*, 664–670.

Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.C., Hung, T., Argani, P., Rinn, J.L., et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature *464*, 1071–1076.

Gutschner, T., and Diederichs, S. (2012). The hallmarks of cancer: a long non-coding RNA point of view. RNA Biol. *9*, 703–719.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300.

Harigaya, Y., Tanaka, H., Yamanaka, S., Tanaka, K., Watanabe, Y., Tsutsumi, C., Chikashige, Y., Hiraoka, Y., Yamashita, A., and Yamamoto, M. (2006). Selective elimination of messenger RNA prevents an incidence of untimely meiosis. Nature *442*, 45–50.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.

Houseley, J., Rubbi, L., Grunstein, M., Tollervey, D., and Vogelauer, M. (2008). A ncRNA modulates histone modification and mRNA induction in the yeast GAL gene cluster. Mol. Cell *32*, 685–695.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M., et al. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell *142*, 409–419.

Hung, T., Wang, Y., Lin, M.F., Koegel, A.K., Kotake, Y., Grant, G.D., Horlings, H.M., Shah, N., Umbricht, C., Wang, P., et al. (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat. Genet. *43*, 621–629.

Ietswaart, R., Wu, Z., and Dean, C. (2012). Flowering time control: another window to the connection between antisense RNA and chromatin. Trends Genet. *28*, 445–453.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802.

IHGSC (International Human Genome Sequencing Consortium). (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc. Natl. Acad. Sci. USA *106*, 11667–11672.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.

Kim, T., Xu, Z., Clauder-Münster, S., Steinmetz, L.M., and Buratowski, S. (2012). Set3 HDAC mediates effects of overlapping noncoding transcription on gene induction kinetics. Cell *150*, 1158–1169.

Kretz, M., Siprashvili, Z., Chu, C., Webster, D.E., Zehnder, A., Qu, K., Lee, C.S., Flockhart, R.J., Groff, A.F., Chow, J., et al. (2013). Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature *493*, 231–235. Published online December 2, 2012. http://dx.doi.org/10.1038/nature11661.

Latos, P.A., Pauler, F.M., Koerner, M.V., Şenergin, H.B., Hudson, Q.J., Stocsits, R.R., Allhoff, W., Stricker, S.H., Klement, R.M., Warczok, K.E., et al. (2012). Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science *338*, 1469–1472.

Lee, J.T. (2009). Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. Genes Dev. *23*, 1831–1842.

Lee, J.T., and Bartolomei, M.S. (2013). X-inactivation, imprinting, and long noncoding RNAs in health and disease. Cell *152*, this issue, 1308–1323.

Maass, P.G., Rump, A., Schulz, H., Stricker, S., Schulze, L., Platzer, K., Aydin, A., Tinschert, S., Goldring, M.B., Luft, F.C., and Bähring, S. (2012). A misplaced lncRNA causes brachydactyly in humans. J. Clin. Invest. *122*, 3990–4002.

Mao, Y.S., Sunwoo, H., Zhang, B., and Spector, D.L. (2011). Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. Nat. Cell Biol. *13*, 95–101.

Marques, A.C., and Ponting, C.P. (2009). Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol. *10*, R124.

Martens, J.A., Laprade, L., and Winston, F. (2004). Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. Nature *429*, 571–574.

Martianov, I., Ramadass, A., Serra Barros, A., Chow, N., and Akoulitchev, A. (2007). Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. Nature *445*, 666–670.

Martin, L., Meier, M., Lyons, S.M., Sit, R.V., Marzluff, W.F., Quake, S.R., and Chang, H.Y. (2012). Systematic reconstruction of RNA functional motifs with high-throughput microfluidics. Nat. Methods *9*, 1192–1194.

Mayer, C., Schmitz, K.M., Li, J., Grummt, I., and Santoro, R. (2006). Intergenic transcripts regulate the epigenetic state of rRNA genes. Mol. Cell *22*, 351–361.

Misteli, T. (2007). Beyond the sequence: cellular organization of genome function. Cell *128*, 787–800.

Mitton-Fry, R.M., DeGregorio, S.J., Wang, J., Steitz, T.A., and Steitz, J.A. (2010). Poly(A) tail recognition by a viral RNA element through assembly of a triple helix. Science *330*, 1244–1247.

Modarresi, F., Faghihi, M.A., Lopez-Toledano, M.A., Fatemi, R.P., Magistri, M., Brothers, S.P., van der Brug, M.P., and Wahlestedt, C. (2012). Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. Nat. Biotechnol. *30*, 453–459.

Nagano, T., Mitchell, J.A., Sanz, L.A., Pauler, F.M., Ferguson-Smith, A.C., Feil, R., and Fraser, P. (2008). The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science *322*, 1717–1720.

Nakagawa, S., Naganuma, T., Shioi, G., and Hirose, T. (2011). Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice. J. Cell Biol. *193*, 31–39.

Nakagawa, S., Ip, J.Y., Shioi, G., Tripathi, V., Zong, X., Hirose, T., and Prasanth, K.V. (2012). Malat1 is not an essential component of nuclear speckles in mice. RNA *18*, 1487–1499.

Nam, J.W., and Bartel, D.P. (2012). Long noncoding RNAs in C. elegans. Genome Res. *22*, 2529–2540.

Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., De Laat, W., and Duboule, D. (2011). The dynamic architecture of Hox gene clusters. Science *334*, 222–225.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Ørom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., et al. (2010). Long noncoding RNAs with enhancer-like function in human cells. Cell *143*, 46–58.

Paige, J.S., Wu, K.Y., and Jaffrey, S.R. (2011). RNA mimics of green fluorescent protein. Science *333*, 642–646.

Pandey, R.R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-Dinardo, D., and Kanduri, C. (2008). Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol. Cell *32*, 232–246.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., and Schier, A.F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. *22*, 577–591.

Ponjavic, J., Ponting, C.P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. *17*, 556–565.

Ravasi, T., Suzuki, H., Pang, K.C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M.C., Gongora, M.M., et al. (2006). Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res. *16*, 11–19.

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.

Sasaki, Y.T., Ideue, T., Sano, M., Mituyama, T., and Hirose, T. (2009). MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles. Proc. Natl. Acad. Sci. USA *106*, 2525–2530.

Schmitt, S., Prestel, M., and Paro, R. (2005). Intergenic transcription through a polycomb group response element counteracts silencing. Genes Dev. *19*, 697–708.

Schmitz, K.M., Mayer, C., Postepska, A., and Grummt, I. (2010). Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. Genes Dev. *24*, 2264–2269.

Shappell, S.B. (2008). Clinical utility of prostate carcinoma molecular diagnostic tests. Rev. Urol. *10*, 44–69.

Shevtsov, S.P., and Dundr, M. (2011). Nucleation of nuclear bodies by RNA. Nat. Cell Biol. *13*, 167–173.

Simon, M.D., Wang, C.I., Kharchenko, P.V., West, J.A., Chapman, B.A., Alekseyenko, A.A., Borowsky, M.L., Kuroda, M.I., and Kingston, R.E. (2011). The genomic binding sites of a noncoding RNA. Proc. Natl. Acad. Sci. USA *108*, 20497–20502.

Solovei, I., Kreysing, M., Lanctôt, C., Kösem, S., Peichl, L., Cremer, T., Guck, J., and Joffe, B. (2009). Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. Cell *137*, 356–368.

Splinter, E., de Wit, E., Nora, E.P., Klous, P., van de Werken, H.J., Zhu, Y., Kaaij, L.J., van Ijcken, W., Gribnau, J., Heard, E., and de Laat, W. (2011). The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes Dev. *25*, 1371–1383.

Sunwoo, H., Dinger, M.E., Wilusz, J.E., Amaral, P.P., Mattick, J.S., and Spector, D.L. (2009). MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. Genome Res. *19*, 347–359.

Szafranski, P., Dharmadhikari, A.V., Brosens, E., Gurha, P., Kolodziejska, K.E., Ou, Z., Dittwald, P., Majewski, T., Mohan, K.N., Chen, B., et al. (2013). Small non-coding differentially-methylated copy-number variants, including lncRNA genes, cause a lethal lung developmental disorder. Genome Res. *23*, 23–33. Published online October 3, 2012. http://dx.doi.org/10.1101/gr.141887.112.

Tavares, L., Dimitrova, E., Oxley, D., Webster, J., Poot, R., Demmers, J., Bezstarosti, K., Taylor, S., Ura, H., Koide, H., et al. (2012). RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3. Cell *148*, 664–678.

Tian, D., Sun, S., and Lee, J.T. (2010). The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. Cell *143*, 390–403.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res. *22*, 1616–1625.

Tinzl, M., Marberger, M., Horvath, S., and Chypre, C. (2004). DD3PCA3 RNA analysis in urine—a new perspective for detecting prostate cancer. Eur. Urol. *46*, 182–186, discussion 187.

Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. Science *329*, 689–693.

Tycowski, K.T., Shu, M.D., Borah, S., Shi, M., and Steitz, J.A. (2012). Conservation of a triple-helix-forming RNA stability element in noncoding and genomic RNAs of diverse viruses. Cell Rep. *2*, 26–32.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537–1550.

van Dijk, M., Thulluru, H.K., Mulders, J., Michel, O.J., Poutsma, A., Windhorst, S., Kleiverda, G., Sie, D., Lachmeijer, A.M., and Oudejans, C.B. (2012). HELLP babies link a novel lincRNA to the trophoblast cell cycle. J. Clin. Invest. *122*, 4003–4011.

van Werven, F.J., Neuert, G., Hendrick, N., Lardenois, A., Buratowski, S., van Oudenaarden, A., Primig, M., and Amon, A. (2012). Transcription of two long noncoding RNAs mediates mating-type control of gametogenesis in budding yeast. Cell *150*, 1170–1181.

Verma-Gaur, J., Torkamani, A., Schaffer, L., Head, S.R., Schork, N.J., and Feeney, A.J. (2012). Noncoding transcription within the Igh distal V(H) region at PAIR elements affects the 3D structure of the Igh locus in pro-B cells. Proc. Natl. Acad. Sci. USA *109*, 17004–17009.

Wan, Y., Qu, K., Ouyang, Z., Kertesz, M., Li, J., Tibshirani, R., Makino, D.L., Nutter, R.C., Segal, E., and Chang, H.Y. (2012). Genome-wide measurement of RNA folding energies. Mol. Cell *48*, 169–181.

Wang, K.C., and Chang, H.Y. (2011). Molecular mechanisms of long noncoding RNAs. Mol. Cell *43*, 904–914.

Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M.G., Glass, C.K., and Kurokawa, R. (2008). Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. Nature *454*, 126–130.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., et al. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.

Wheeler, T.M., Leger, A.J., Pandey, S.K., MacLeod, A.R., Nakamori, M., Cheng, S.H., Wentworth, B.M., Bennett, C.F., and Thornton, C.A. (2012). Targeting nuclear RNA for in vivo correction of myotonic dystrophy. Nature *488*, 111–115.

Wilusz, J.E., Jnbaptiste, C.K., Lu, L.Y., Kuhn, C.D., Joshua-Tor, L., and Sharp, P.A. (2012). A triple helix stabilizes the 3′ ends of long noncoding RNAs that lack poly(A) tails. Genes Dev. *26*, 2392–2407.

Wolpert, L. (2011). Positional information and patterning revisited. J. Theor. Biol. *269*, 359–365.

Yamamoto, M. (2010). The selective elimination of messenger RNA underlies the mitosis-meiosis switch in fission yeast. Proc. Jpn. Acad., Ser. B, Phys. Biol. Sci. *86*, 788–797.

Yang, L., Lin, C., Liu, W., Zhang, J., Ohgi, K.A., Grinstein, J.D., Dorrestein, P.C., and Rosenfeld, M.G. (2011). ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. Cell *147*, 773–788.

Yin, Q.F., Yang, L., Zhang, Y., Xiang, J.F., Wu, Y.W., Carmichael, G.G., and Chen, L.L. (2012). Long noncoding RNAs with snoRNA ends. Mol. Cell *48*, 219–230.

Yoon, J.H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J.L., De, S., Huarte, M., Zhan, M., Becker, K.G., and Gorospe, M. (2012). LincRNA-p21 suppresses target mRNA translation. Mol. Cell *47*, 648–655.

Zhang, B., Arun, G., Mao, Y.S., Lazar, Z., Hung, G., Bhattacharjee, G., Xiao, X., Booth, C.J., Wu, J., Zhang, C., and Spector, D.L. (2012). The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult. Cell Rep. *2*, 111–123.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol. Cell *40*, 939–953.

Zofall, M., Yamanaka, S., Reyes-Turcu, F.E., Zhang, K., Rubin, C., and Grewal, S.I. (2012). RNA elimination machinery targeting meiotic mRNAs promotes facultative heterochromatin formation. Science *335*, 96–100.

CellPress

Cell Stem Cell
Review

Long Noncoding RNAs in Cell-Fate Programming and Reprogramming

Ryan A. Flynn[1] and Howard Y. Chang[1,*]
[1]Howard Hughes Medical Institute and Program in Epithelial Biology, Stanford University School of Medicine, Stanford, CA, 94305, USA
*Correspondence: howchang@stanford.edu
http://dx.doi.org/10.1016/j.stem.2014.05.014

In recent years, long noncoding RNAs (lncRNAs) have emerged as an important class of regulators of gene expression. lncRNAs exhibit several distinctive features that confer unique regulatory functions, including exquisite cell- and tissue-specific expression and the capacity to transduce higher-order spatial information. Here we review evidence showing that lncRNAs exert critical functions in adult tissue stem cells, including skin, brain, and muscle, as well as in developmental patterning and pluripotency. We highlight new approaches for ascribing lncRNA functions and discuss mammalian dosage compensation as a classic example of an lncRNA network coupled to stem cell differentiation.

Introduction

Efforts to understand how tissues are patterned during development and maintained by stem cells throughout life have traditionally focused on the protein-coding genome. Over the past decade, however, our understanding of the noncoding genome and its impact on cell fate has dramatically expanded. Contrary to previous notions of genome organization and function, the identification of thousands of long and short noncoding RNAs (ncRNAs) has revealed that much of the genome is in fact transcribed. Long noncoding RNAs (lncRNAs) are operationally defined as transcripts of greater than 200 nucleotides that function by means other than coding for proteins; lncRNAs are typically transcribed by RNA polymerase II and are frequently spliced and polyadenylated (reviewed by Rinn and Chang, 2012). As a class, lncRNAs tend to be expressed at lower levels and are predominantly localized in the nucleus, in contrast to messenger RNAs, which are abundant and enriched in the cytoplasm (Derrien et al., 2012). Notwithstanding these generalizations, lncRNAs exhibit a wide range of expression levels and distinct cytotopic localizations, reflecting a large and diverse class of regulators (reviewed by Batista and Chang, 2013). Several well-studied examples of lncRNAs suggest that they can operate through distinct modes, including as signals, scaffolds for protein-protein interactions, molecular decoys, and guides to target elements in the genome or transcriptome (Wang and Chang, 2011). The discovery of novel lncRNAs has historically outpaced their functional annotation; however, efforts to more specifically ascribe function to either previously identified or novel lncRNAs have increased in recent years. Stem cells offer an attractive system for studying lncRNA function since previous findings have suggested that lncRNA expression is more cell-type-specific than mRNA expression (Cabili et al., 2011), leading to the possibility that lncRNAs may be key regulators of cell fate.

Here we review recent developments that illuminate the roles of lncRNAs in stem cell biology. We explore efforts to characterize the functions of lncRNAs in the development and patterning of several somatic tissues, including skin, brain, and musculature. Additionally, we examine how lncRNAs contribute to the pluripotent state and can be used to assess reprogramming status.

LncRNAs in Adult Tissue Stem Cells

Skin: An Ideal Model

Studying the biology of tissues at the molecular level necessitates robust model systems. While there are few systems that are suitable for detailed molecular characterization, well-developed human models exist for the skin based on ex vivo tissue regeneration that can also be grafted in vivo (Sen et al., 2010; Truong et al., 2006). Such models provide cellular material for molecular and biochemical studies that would be otherwise inaccessible and offer a system for testing the function of lncRNAs. Surveying the pattern of gene expression during epidermal differentiation, Khavari and colleagues discovered two key lncRNAs, ANCR and TINCR, that are expressed in epidermal stem cells and their terminally differentiated progeny, respectively (Kretz et al., 2012, 2013) (Figure 1). Antidifferentiation noncoding RNA (ANCR) provides a prime example of an lncRNA that controls the differentiation state of a somatic stem cell (Kretz et al., 2012). Specifically, ANCR depletion results in ectopic differentiation of epidermal stem cells, implying that ANCR's role is to suppress the differentiation pathway in the epidermis and maintain the stem cell compartment.

While ANCR appears to inhibit differentiation, a different lncRNA termed terminal differentiation-induced noncoding RNA (TINCR) promotes epidermal differentiation (Kretz et al., 2013). TINCR is kept at very low levels in epidermal stem cells, but it is dramatically induced upon differentiation. Mechanistic studies of TINCR revealed that TINCR is a cytoplasmic lncRNA that interacts with the RNA-binding protein (RBP) STAU1 and converts STAU1 into an mRNA stability factor (Figure 1). Together, TINCR and STAU1 bind to and functionally stabilize mRNAs that encode structural and regulatory proteins critical for terminally differentiated keratinocytes. Additionally, TINCR expression is downregulated in human squamous cell carcinoma, providing evidence that lncRNAs can functionally regulate healthy and disease tissues.

The development of two techniques made these insights possible: (1) RNA interactome analysis (RIA), which allows the retrieval and unbiased discovery of RNAs interacting with an lncRNA of interest, and (2) protein microarray hybridization, which allows rapid discovery of direct RBP partners of an

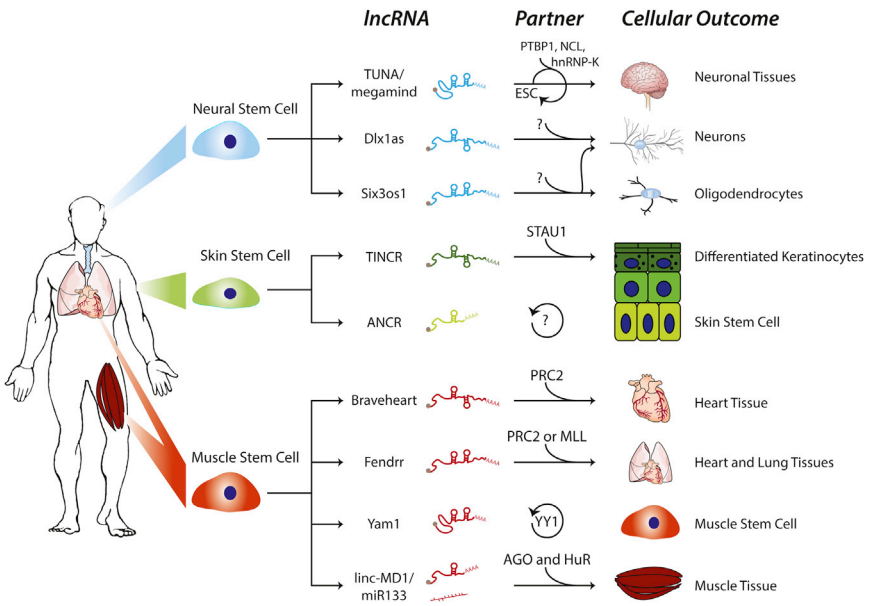


Figure 1. lncRNAs Control Differentiation and Self-Renewal

Several lncRNAs that regulate specific somatic tissue stem cell renewal or differentiation and their protein partners are depicted. Some lncRNAs maintain the stem cell state, while others promote a differentiation program. Their functions are often facilitated by protein partners that impart the ability to activate or repress gene expression or posttranscriptionally regulate other RNAs.

lncRNA (Kretz et al., 2013). Moreover, both ANCR and TINCR were identified from large-scale expression profiling studies, suggesting that many additional lncRNAs may be identified and characterized using this system. Indeed, the differentiation of the skin is a multistep and highly regulated process that could benefit from the diverse set of lncRNAs hiding in the genome. The development of techniques such as RIA and the implementation of protein microarrays facilitated the functional characterization of TINCR but are applicable to uncovering mechanisms of other lncRNAs. Within the skin, the regulated and sequential expression of lncRNAs is clearly essential for their function; thus, understanding what controls the spatiotemporal expression of lncRNAs, such as ANCR and TINCR, should be the focus of future studies.

Regulation in the Brain

Transcription and alternative splicing in the brain appear to be the most complex among all organs (Mehler and Mattick, 2007; Mercer et al., 2008). An early example of lncRNAs controlling neural cell fates involves the Evf2 lncRNA and the *Dlx5/6* genomic locus (Bond et al., 2009). *Evf2* is transcribed antisense to *Dlx6*, which encodes a transcription factor, and is located immediately downstream of the *Dlx5* genomic locus. The act of transcribing *Evf2* can control the levels of Dlx6 in *cis*, and after disengaging the polymerase, Evf2 acts in *trans* to modulate the methylation of the *Dlx5/6* enhancer and transcription of *Dlx5*. Therefore, by regulating the cellular levels of the Dlx5 and Dlx6 transcription factors, Efv2 controls GABAergic interneuron activity (Berghoff et al., 2013; Bond et al., 2009). A different study characterizing another lncRNA important for neural differentiation found that an enhancer region of the gene encoding the Neurogenin 1 transcription factor was transcribed and produced an lncRNA that positively regulated Neurogenin 1 expression (Onoguchi et al., 2012). These few examples begin to build the case that lncRNAs play an important role in neural biology.

The starting point of many lncRNA studies is unbiased gene expression analysis, which can reveal novel lncRNAs and their expression pattern in a developmental context. Recent large-scale efforts have employed next generation sequencing ("-seq") technologies, from RNA-seq to chromatin immunoprecipitation (ChIP-seq), to identify transcripts and define their genomic positions (reviewed by Rinn and Chang, 2012). In the mouse brain, Lim and colleagues isolated three separate regions, subventricular zone (SVZ), olfactory bulb (OB), and the dentate gyrus (DG), and subjected these samples to short-read RNA-seq and ChIP-seq (Ramos et al., 2013). Over 3,600 novel lncRNAs were identified, and clustering of the lncRNAs and mRNAs by their expression patterns revealed that the lncRNAs were more tissue specific than mRNAs, consistent with previous reports (Cabili et al., 2011). Application of CaptureSeq, a technique that circumvents some drawbacks of short-read sequencing (Mercer et al., 2012), to further characterize the transcriptome of adult SVC tissue doubled the number (to ~7,000) of novel lncRNAs identified. To functionally validate the cataloging effort, two lncRNAs were identified by selecting loci marked by H3K4me3, which is associated with expressed genes, in NPC-SVC cells. This search identified Six3os and Dlx1as for further testing. Notably, Six3os has been previously reported to control retinal development (Rapicavoli et al., 2011). To characterize the neural role of Six3os and Dlx1as, SVZ neural progenitor cells were challenged in a 7-day differentiation assay with short hairpin RNAs targeting the two lncRNAs or control shRNAs. Depletion of Six3os lncRNA leads to fewer Tuj1 (neuron marker)- and OLIG2 (oligodentrocyte marker)-positive cells, whereas depletion of Dlx1as specifically affected the number of Tuj1-positive cells (Figure 1). While the molecular mechanisms of these lncRNAs were not explored, Six3os has been shown to physically interact with Ezh2, a component of the Polycomb repressive complex 2 (PRC2), to repress specific genes in retinal cells (Rapicavoli et al., 2011). These examples illustrate that mapping spatiotemporal patterns of lncRNAs can highlight functional transcripts. Larger-scale validation efforts will be required to fully realize the extent of lncRNA regulation in the different regions of the brain.

A complementary approach identifies potential lncRNA regulators based on their loss-of-function phenotypes in large-scale depletion studies (Guttman et al., 2011; Lin et al., 2014). Rana and colleagues targeted 1,280 mouse lncRNAs and identified 20 lncRNAs that were required for the maintenance of mouse embryonic stem cell (mESC) pluripotency. One lncRNA, named TUNA, was previously identified as *megamind* in zebrafish.

TUNA/Megamind depletion in zebrafish led to altered neurodevelopment and impaired locomotor response (Ulitsky et al., 2011; Lin et al., 2014). TUNA is highly conserved in human and fish, is required for the maintenance of pluripotency, and is also expressed in the brain, spinal cord, and eyes in adult tissues. Indeed, TUNA expression was increased when mESCs differentiated toward the neural lineages, and TUNA depletion inhibited neural differentiation of ESCs (Figure 1). Purifying proteins that associate with in vitro-transcribed TUNA identified hnRNP-K, Nucleolin (NCL), and PTBP1 as interaction partners. Importantly, depletion of several of these proteins phenocopied TUNA depletion (Lin et al., 2014). An important caveat to consider is that while the candidate approach characterized TUNA, Six3os, and Dlx1as lncRNAs as successful validation of genome-wide screens, such approaches leave the function of thousands of other transcripts, many of which may play important roles, unaddressed.

Many lncRNAs have been implicated in the regulation of chromatin states (Rinn and Chang, 2012), but direct evidence for their association has only recently been possible through the development chromatin isolation by RNA purification (ChIRP; and others methods discussed below) (Chu et al., 2011). ChIRP uses DNA capture probes to retrieve a specific lncRNA with its associated genomic DNA targets, and together with deep sequencing can generate a genome-wide map of lncRNA-chromatin interactions. Careful optimization of in vivo crosslinking, both of the chemical crosslinking agent and duration, and selection of proper oligonucleotide probes are important to obtain reliable measurement. This process often includes multiple but distinct DNA capture probe sets, probes targeting irrelevant RNAs as negative controls, and positive control regions to assay during pilot experiments (Chu et al., 2011). Successful implementation of ChIRP has revealed the lncRNA TUNA occupies promoter regions of *Nanog*, *Sox2*, and *Fgf4*, genes that are important for pluripotency and neural lineage commitment (Lin et al., 2014). Together with its protein partners and its chromatin localization, TUNA may regulate gene expression at both the transcriptional and post-transcriptional level. Thus, TUNA represents an lncRNA that is important for at least two cell states (ESC pluripotency and neural differentiation) and probably operates through multiple molecular mechanisms. This example highlights the concept that a single lncRNA can, under different cellular context and protein partners, function to control multiple molecular pathways.

lncRNAs and Muscle

LncRNAs also control development of mesodermal tissues and have similarly benefited from large-scale sequencing efforts to identify functionally important transcripts. One example of a heart-specific lncRNA named Braveheart was first functionally characterized as a key factor involved in cardiac lineage commitment because its depletion resulted in a severe reduction in the number of spontaneous beating cardiomyocytes formed during embryoid body differentiation (Klattenhoff et al., 2013) (Figure 1). Further characterization of Braveheart found that it interacts with Suz12, a subunit of PRC2, and acts in *trans* to regulate heart-specific differentiation genes such as *MesP1*. The regulation of master drivers of cardiac differentiation, such as *MesP1* by Braveheart, offers new tools toward the goal of achieving highly efficient and reproducible in vitro reprograming (Burridge et al., 2012). Producing cardiomyocytes from induced pluripotency stem cells (iPSCs) or directly from other differentiated cell types may benefit from engineering specific lncRNA expression during in vitro production.

While small interfering RNA (siRNA) knockdown of lncRNAs (used in most of the discussed work) often provides a great deal of insight into function, off-target effects and incomplete depletion must always been considered. As with protein-coding genes, knockout (KO) strategies offer potential remedies to these siRNA-related issues, but the specific strategy employed is critical (discussed below: *Developmental Patterning by lncRNAs*). Utilizing this concept, Herrmann and colleagues inserted a premature polyadenylation (polyA) signal into the lncRNA *Fendrr*'s locus to promote depletion of the full-length Fendrr RNA (Grote et al., 2013). Initial characterization of Fendrr found it expressed in the caudal end of the lateral plate mesoderm (LPM), which develops into the structures like the heart and body wall. Fendrr KO resulted in embryonic lethality at embryonic day 13.75, abdominal wall defects, and pooling of blood in the right atrium. By partnering with both activating (mixed-lineage Leukemia [MLL], WDR5) and silencing (PRC2) chromatin complexes, Fendrr was proposed to modulate the epigenetic landscape during development (Figure 1). More recently ChIRP was used to show that Fendrr physically associates with the promoters of FoxF1 and Pitx2 mRNAs, two genes repressed by Fendrr (Grote and Herrmann, 2013; Grote et al., 2013). Fendrr therefore represents a dual-function lncRNA that may control both positive and negative chromatin modifying complexes to guide development.

Long RNAs Controlling Small RNAs

The differentiation of a myoblast progenitor cell (MB) to a fully differentiated muscle cell is a highly regulated process that relies on Ying Yang 1 (YY1), a multifunctioning transcription factor (Deng et al., 2010; Lu et al., 2012). Examination of YY1's chromatin binding pattern in MBs revealed that it bound the promoter of many ncRNA loci, and these target noncoding genes were named YY1-associated muscle lncRNAs (Yam) (Lu et al., 2013). Characterization of one of these lncRNAs, Yam1, identified it as a key regulator of myogenesis, as it was able to repress key muscle differentiation genes including myogenin, Tnni2, and α-actin, (Figure 1). Furthermore, Yam1 increased levels of microRNA-715 (miR-715), which targets Wnt7b, a protein that normally promotes muscle differentiation (Lu et al., 2013). Yam1 thus provides evidence that in muscle lncRNAs can modulate the levels of both mRNAs and other ncRNAs, such as miRNAs, providing additional network control to cells.

The regulation of miRNA networks reveals an additional mechanism through which lncRNAs exert control. Recently, multiple lncRNAs have been shown to act as competing endogenous RNAs (ceRNAs), where the lncRNAs are proposed to bind to and compete miRNAs away from cognate mRNA targets (Tay et al., 2014). Pseudogene lncRNAs are prime candidates for the ceRNA mechanism because they may share multiple miRNA binding sites, allowing more effective competition with cognate mRNAs. The ceRNA hypothesis requires that ceRNAs are expressed highly enough and have sufficient numbers of miRNA binding sites to substantially affect the pool of cellular miRNAs. Recent work exploring the dynamics of miRNA-regulated gene repression has shown that it is highly susceptible to thresholds. In certain contexts, small concentration changes of miRNA-mRNA or

miRNA-ceRNA pairs can substantially modulate the gene expression network (Bosia et al., 2013; Mukherji et al., 2011). Moreover, one example of a ceRNA, linc-MD1, has been previously show to regulate muscle differentiation through its ability to sponge miR-133 and miR-135 away from the mRNAs MAML1 and MEF2C (Cesana et al., 2011). These two mRNAs are important transcriptional activators of the muscle differentiation program. Linc-MD1 itself contains an miR-133b, which represses muscle differentiation when processed. Recent molecular characterization of this network revealed the RBP HuR bound to linc-MD1 and the levels of linc-MD1 positively correlated with HuR protein abundance (Legnini et al., 2014) (Figure 1). HuR controlled the fate of linc-MD1, as cellular depletion of HuR favored the processing of linc-MD1 into miR-133b, tipping the balance in favor of the miRNA over the ceRNA. HuR has known roles in myogenesis and its interaction with linc-MD1 fine-tunes the levels of miRNAs important in the muscle differentiation program. Together, these studies explore lncRNA functions in muscle tissue and help to expand the possible modes of lncRNA functions within the already complex system of miRNA-mediated gene regulation.

Developmental Patterning by lncRNAs

lncRNAs also orchestrate the patterning of cells into tissues and organs during development. HOTAIR lncRNA was one of the first characterized lncRNAs that acts at distance (in *trans*) to modulate Hox gene expression (Rinn et al., 2007). HOTAIR is a repressive lncRNA and serves a scaffold between two distinct chromatin modification complexes (Rinn et al., 2007; Tsai et al., 2010). Other Hox-encoded lncRNAs such as HOTTIP, Mistral, and HOTAIRm1 were shown to regulate different members of HoxA genes (Bertani et al., 2011; Wang et al., 2011; Zhang et al., 2009). For example, HOTTIP is expressed in distal anatomic structures and activates the expression of *HOXA9-HOXA13* genes to promote distal limb development (Wang et al., 2011). Characterization of these lncRNAs has often occurred through overexpression or siRNA knockdown studies. While these strategies often yield relevant results, transcriptional modulation is often not complete, especially using siRNA (or even short hairpin RNA), necessitating alternative methods.

Recently there have been a number of studies utilizing gene KO to understand lncRNA biology (Grote et al., 2013; Li et al., 2013; Sauvageau et al., 2013). At least three KO strategies have been reported: (1) insertion of a polyA signal near the transcription start sites; (2) insertion of a reporter gene under the control of the endogenous promoter; and (3) complete deletion of the lncRNA locus. The latter is the most dramatic and may, in addition to removing the lncRNA exons/intron structure, remove unknown regulatory elements. Insertion of a reporter gene has the advantage of being able to monitor expression of the lncRNA throughout development; however, depending on which sequences are replaced, it may also carry similar drawbacks as the deletion strategy. Finally, insertion of a polyA signal near the transcription start sites likely has the least off-target effects; however, background expression from the lncRNA locus could still result from not removing downstream sequences, cryptic start sites, or inefficient polyA tailing and cleavage.

Elucidating lncRNA Tissue Patterning by KO Models

Recent efforts have begun to utilize full KO strategies to characterize additional lncRNAs including Hox encoded candidates. The developmental functions of mouse Hotair were investigated by full lncRNA locus deletion in the mouse (Li et al., 2013). Loss of Hotair resulted in aberrant patterning of the skeletal system during development, as was evident in abnormalities in the wrist and spine, including a switch of vertebral segment identity called homeotic transformation. Further, genome-wide characterization of the Hotair KO mouse confirmed that murine Hotair acted similarly to human HOTAIR, namely as a *trans*-acting lncRNA controlling histone modification at specific genomic loci (Li et al., 2013). More recently, in an effort to dramatically expand the number of lncRNA KOs, Rinn and colleagues used the reporter gene approach to generate 18 separate lncRNA knockout mice (Sauvageau et al., 2013). By replacing lncRNA exonic regions with a LacZ construct, both KO and tagging was achieved. Three of the 18 lncRNAs (Fendrr, Peril, and Mdgt) showed variable penetrance and lethality. The Mdgt and Pint KO lead to abnormally low body weight and slower growth. The detailed characterization of the lncRNA Brn1b revealed its role in cortical development; specifically, this lncRNA was important for the embryonic patterning in certain areas of projection neurons. By creating a large number of lncRNA KO mice and characterizing many of their functions in vivo, this study helped to solidify the functional importance of lncRNAs. While thousands of lncRNAs remain to be genetically tested, new and more facile genome-editing tools should speed future characterization (Mali et al., 2013).

Sauvageu et al. also generated a new Fendrr KO mouse (Sauvageau et al., 2013). Under these conditions, Fendrr was expressed much more widely than previously observed and most highly in the developing lung. Fendrr KO resulted in perinatal lethality, as Fendrr −/− embryos either failed to initiate breathing or stop breathing within 5 hr of birth, neither of which was observed in wild-type pups. While the most striking phenotype of this KO was pulmonary, heart septal defects were also apparent even though their LacZ construct did not stain the heart for expression. This discrepancy is an important example of the possible phenotypic difference achieved by differential KO strategies such as reporter construct replacement or early polyA termination (Grote et al., 2013; Sauvageau et al., 2013). Specifically, addition of the polyA sites resulted in minimal disruption of the endogenous Fendrr locus, but extremely low levels of Fendrr were still detectable (Grote and Herrmann, 2013). On the other hand, the LacZ construct replaced ~20 kb of the genome, resulting in a complete lack of Fendrr transcripts; however, this large replacement may have removed other functional elements from the genome responsible for regulating other genes. Therefore, while both approaches confirmed loss of the lncRNA transcript, additional investigation is necessary and careful consideration of the cellular outcomes from any particular targeting strategy must be included in the experimental design.

Single-Cell Analysis of lncRNA Function

Most transcript-profiling experiments of lncRNAs have employed bulk measurements, reporting results from an average of thousands or millions of cells. Recent work at the single-cell level has revealed how much heterogeneity exists even within a "clonal" population of cells (Buganim et al., 2012; Shalek et al., 2013). Thus, it follows that examination of the noncoding genome and its function at the single-cell level could also reveal

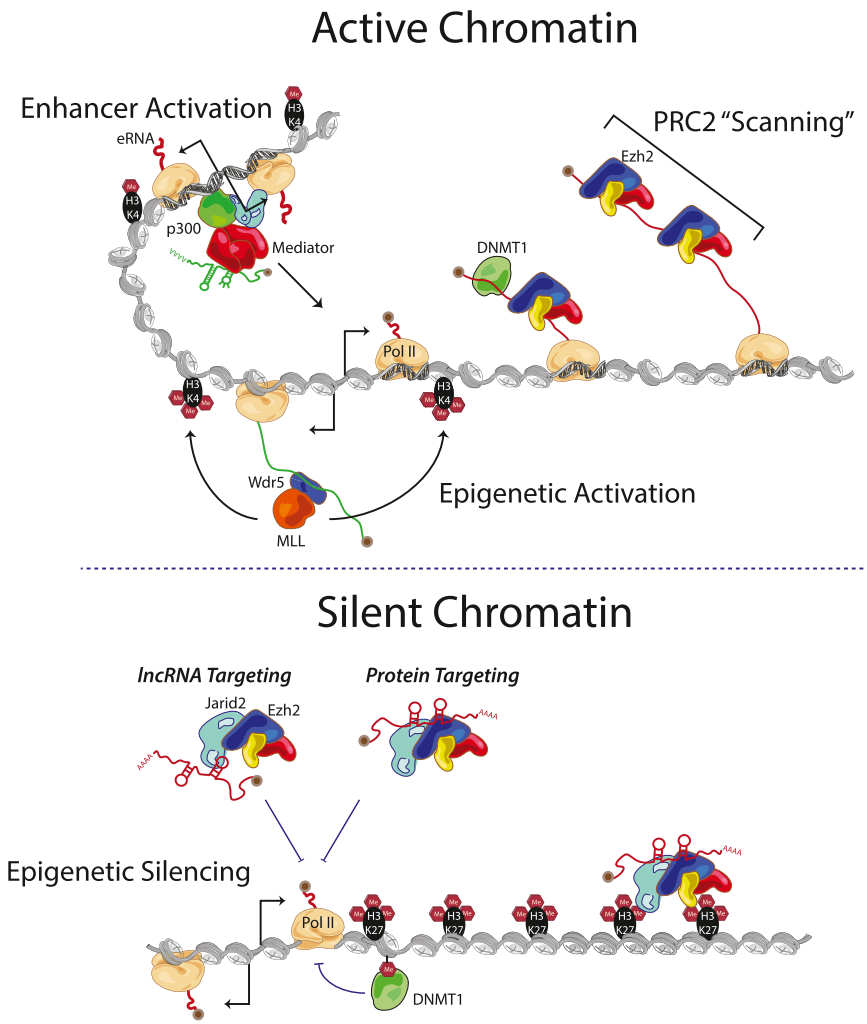


Figure 2. lncRNAs Program Active and Silent Chromatin States

Top: in ESCs active chromatin is achieved and maintained through multiple mechanisms. *cis*-acting lncRNAs can recruit the MLL/WDR5 complex to deposit H3K4me3 at promoters. Enhancer regions can transcribe enhancer RNAs (eRNAs); some enhancer-like RNAs bring Mediator to promoters to contribute to gene activation. Additionally, through interactions with the nascent transcribed RNA, canonical silencing factors such as PRC2 and DNMT1 are titrated away from active chromatin. Bottom: chromatin also employs many lncRNA-based mechanisms to stay silent. Ezh2 and JARID2 (subunits of PRC2) may bind lncRNAs to facilitate specific chromatin targeting or to enhance PRC2 complex assembly and stability. Additionally, when nascent RNA production is low, DNMT1 can interact with the chromatin and act to silence through DNA methylation.

novel modes of action. Additionally, while some studies have successfully elucidated the role lncRNAs present at a low copy number (Wang et al., 2011), the accuracy of such reports remains challenging when working with bulk populations.

Recent characterization of an lncRNA, named lincHOXA1, located in the 3′ end of the HoxA cluster by Raj and colleagues, brought to light the importance of carefully examining, at the single-cell level, the function of lowly expressed lncRNAs (Maamar et al., 2013). Initial analysis, at the bulk cell level, ascribed a positive correlation to the expression of lincHOXA1 and a nearby mRNA HOXA1. Surprisingly, however, single-cell analysis revealed an anticorrelation, and specifically a switch-like relationship was observed such that if a cell had above ten copies of lincHOXA1, *HOXA1* was repressed. Knockdown studies used both siRNA and antisense oligonucleotides (ASO, via RNase H-mediated cleavage of the target RNA). The two depletion methods differ in their capacity to reduce lincHOXA1 levels on the chromatin versus total levels, with siRNA treatment unable to efficiently lower chromatin-associated transcripts. Functionally, lincHOXA1 was found to partner with purine-rich element-binding protein B (PURB) and exert transcriptional silencing of HOXA1. Importantly, this study highlights two key and common methodological decision points: context of cellular measurements and RNA knockdown strategies. In this case, bulk measurements would have masked the anticorrelated relationship between lincHOXA1 and HOXA1, which could have led to key misinterpretations. Additionally, use of siRNAs, which was effective in reducing total cellular levels of lincHOXA1, was not efficient at depleting the functional lincHOXA1 transcripts. Future work examining the molecular roles of both coding and noncoding transcripts should choose carefully the methods and context in which experiments are performed. As single-cell analysis and ASO technology become more robust and widely adopted, it is likely that many unknown features of known lncRNAs may be revealed.

LncRNAs Regulation of Pluripotency

The richness of the lncRNA regulatory landscape is perhaps best exemplified in ESCs, where the noncoding transcriptome has been under intense study. The expansive number of genomic data sets, both RNA- and chromatin-based, now available in ESCs provides a rich database to characterize lncRNA function. Recent progress in understanding lncRNA control of pluripotency and dosage-compensation mechanisms have revealed intimate connections between lncRNAs and chromatin state (Figure 2). Some of the most studied lncRNA binding proteins belong to chromatin modification complexes, including PRC2 and MLL, which act to suppress and activate, respectively, transcription through methylation of histone protein.

Transition between Cell States

Characterization of the transcriptome of ESCs has revealed many lncRNAs that participate in the regulation of the pluripotent state (Guttman et al., 2011, 2009; Lin et al., 2014; Ng et al., 2012; Sheik Mohamed et al., 2010). Through a comprehensive "perturb-and-measure" strategy, Guttman et al. showed that dozens

of lncRNAs are required for the setting the gene expression patterns of mouse ESCs or the first step of differentiation toward different germ layers (Guttman et al., 2011). A subset of these lncRNAs bound one or more chromatin modification complexes, including readers, writers, or erasers of repressive histone modifications.

In contrast, the "regulator of reprogramming" lncRNA (lincRNA-RoR) was identified as an important factor for the reprogramming process as its depletion or overexpression leads to a lower or higher efficiency of reprogramming fibroblasts to iPSCs, respectively (Loewer et al., 2010). However only recently was the molecular mechanism investigated (Wang et al., 2013). Pull-down experiments with lincRNA-RoR specifically isolated miR-145-5p, 181a-5p, and 99b-3p, as well as the miR-targeting protein Argonaute2 (Ago2). These miRs have been previously shown to regulate core pluripotency factors such as *Pou5f1*, *Sox2*, and *NANOG*, suggesting that lincRNA-RoR might act as a ceRNA. Indeed, functional assays revealed that lincRNA-RoR regulated the mature form of miR-145, characteristic of a ceRNA. Loss of lncRNA-RoR caused human ESCs (hESCs) to differentiate toward mesoderm and ectoderm, while overexpression conferred a differentiation defect. Additionally, in the context of cancer, a rapidly proliferative state similar to ESCs, lincRNA-RoR was recently shown to act in a regulatory loop suppressing the expression of the tumor suppressor p53 (Zhang et al., 2013). Together, this characterization of lincRNA-RoR further advances the idea that each lncRNA may control many pathways in different cellular contexts including tumor growth and core pluripotency gene network utilizing a ceRNA mechanism.

Activation of the Epigenome with lncRNAs

To date, the vast majority of lncRNAs have annotated functions in repressive complexes, with only a few examples of activating or enhancing lncRNAs (Wang et al., 2011; Zhang et al., 2009). HOTTIP, named due to its location at the distal "tip" of the HOXA gene cluster, enforces an active chromatin state by recruiting the WDR5 subunit of the MLL complex (Wang et al., 2011) (Figure 2). The *HOTTIP* locus comes into spatial proximity with its target genes, and all the while the expression level of HOTTIP remains near one copy per cell (Wang et al., 2011). The low copy number of HOTTIP ensures that HOTTIP acts precisely in *cis* on target genes defined by proximity in three-dimensional nuclear space but not broadly on other genes. More recently, biochemical characterization of the interaction between WDR5 and HOTTIP revealed a specific RNA-binding pocket of WDR5 and that RNA binding could stabilize chromatin-associated WDR5 (Yang et al., 2014). This finding suggested that in vivo, not only the localization, but also the half-life of WDR5 could be modulated by HOTTIP. Given that WDR5/MLL acts at many genomic loci, RNA immunoprecipitation-seq (RIP-seq) was used to identify over 1,400 WDR5 interacting RNAs, including many coding and noncoding RNAs. An lncRNA-binding pocket on WDR5 was discovered, and a specific mutation of the RNA-binding pocket selectively abrogated RNA binding but no other functions of the WDR5-MLL complex (Yang et al., 2014). This selective WDR5 mutant revealed that RNA binding is important for the temporal stability of the active chromatin mark H3K4me3 over time and maintenance of ESC pluripotency. These studies suggest a generalizable mechanism for functional MLL/WDR5-RNA interaction. Specifically, HOTTIP acts in *cis* and is expressed at far too low levels per cell to globally modulate the MLL/WDR5 chromatin localization. The RIP-seq of WDR5 in mESCs (which do not express HOTTIP) revealed that more than one thousand cellular RNAs could interact with and may modulate the chromatin modification complex. Because WDR5 targets over 10,000 genomic sites (Ang et al., 2011), whether the three-dimensional organization of the genome facilitates lncRNA coregulation of the mESC self-renewal program remains to be addressed in future studies.

Epigenetic Repression through lncRNA-PRC2 Interactions

Unlike activating chromatin complexes, chromatin-modifying complexes that repress transcription have been more extensively studied in the context of lncRNA interactions, resulting in a richer set of known interactions. The focus of many of these studies has been the PRC2 complex, responsible for depositing H3K27me3, which plays roles pluripotency, differentiation, XCI, and diseases such as cancer (Margueron and Reinberg, 2011). An initial survey of the RNA-interactome of Ezh2 yielded more than 9,000 target RNAs using RIP-seq in mESCs (Zhao et al., 2010). Recently, two studies have revisited this observation to further clarify the interplay between RNA and PRC2 (Davidovich et al., 2013; Kaneko et al., 2013) (Figure 2). Biochemical interaction and photoactivated RNA-crosslinking experiments suggest that Ezh2 can interact with numerous RNAs, including the 5′ end of nascent RNAs that are actively transcribed. The apparently specific interactions of PRC2 with several lncRNAs in lysate and in vivo are not recapitulated in vitro by the core PRC2 complex alone. The promiscuous RNA binding of Ezh2 may be modulated by additional proteins, such as Jarid2 and others, to facilitate higher degrees of specificity in vivo (Davidovich et al., 2013; see below). Moreover, Ezh2 may scan the genome surveying the transcriptional status of its targets. Actively transcribed regions may continually push Ezh2 away via their elongating mRNAs, while silent regions or those stably bound by lncRNAs (generated in *trans*) can be silenced. This proposed mechanism reinforces the status quo of gene transcription and silencing and is consistent with the known genetic role of Polycomb group proteins in chromatin state maintenance.

A similar RNA surveillance mechanism is also employed by the DNA methylase DNMT1 that interacts with many cellular transcripts, including the nonpolyadenylated extracoding *CEBPA* (*ecCEBPA*) lncRNA. The *ecCEBPA* lncRNA adopts a characteristic stem-loop structure critical for interaction with DNMT1 and, when transcribed, acts to shield the *CEBPA* locus from DNA methylation (Di Ruscio et al., 2013) (Figure 2). These two examples provide evidence that cells employ RNAs to modulate the deposition of repressive epigenetic marks in a genome-wide manner. Nonetheless, recognition of the potentially broad interactions between RNA and PRC2 highlights the need for high-quality in vivo controls and validation of RNA-protein interactions. Methodological choice is critical as each assay type has its own strengths and weaknesses, which will impact the results obtained and conclusions drawn.

While PCR2 operates in a wide range of cell types, certain subunits, such as JARID2, are specifically expressed and partner with PCR2 in ESCs and certain dividing cells, including cancer cells (Pasini et al., 2010; Peng et al., 2009; Shen et al., 2009). These initial studies established JARID2's capacity to regulate

the stability of the PRC2 complex as well as its enzymatic activity (Figure 2). Further expanding the cellular functions of JARID2, in vitro RNA-binding assays and in vivo PAR-CLIP suggest that JARID2 directly interacts with cellular RNAs (Kaneko et al., 2014). JARID2 and Ezh2 reproducibly crosslinked to 106 and 165 lncRNAs, respectively, and 53 lncRNAs were commonly bound. The MEG3 lncRNA was bound by both subunits of PRC2; however, the RNA-binding region (RBR) of JARID2 provided the largest contribution of MEG3 binding to PRC2. Additionally, cellular levels of MEG3 contribute to PRC2's chromatin association, as low expression of MEG3 resulted in the loss of PRC2 subunits from specific loci leading to derepression of the nearby genes. Finally, the in vitro interaction between JARID2 and Ezh2 was facilitated by HOTAIR and MEG3, and Ezh2's chromatin association was shown to be partially dependent on JARID2's RNA-binding domain. Thus, JARID2, an ESC-specific subunit of PRC2, appears to modulate the localization of PRC2, and thus the chromatin state, in an RNA-dependent manner. While this study offers an additional layer of regulation with respect to the Polycomb complex, little is known about the other RNA targets of JARID2, which may significantly contribute to its cellular function. Additionally, studies to rigorously interrogate the enzymatic properties of the PRC2 complex inside cells with and without its RNA partners will be very informative.

An lncRNA Network to Control Dosage Compensation

Dosage compensation of genes encoded on the X chromosome is accomplished by divergent strategies in different species; however, the use of lncRNAs is a common feature. In *Drosophila*, dosage compensation is achieved by precisely upgregulating the X chromosome in males by 2-fold (Lucchesi et al., 2005). A desire to understand how dosage compensation operates fueled the development of ChIRP and CHART, genomic tools that map the chromatin association of lncRNAs (Chu et al., 2011; Simon et al., 2011). Initially, ChIRP and CHART were applied to the *Drosophila* roX2 lncRNA, which provided evidence that roX2 co-occupies genomic loci with the known dosage compensation protein factors on the X chromosome. Importantly, they proved that mapping the genomic locations of lncRNAs can generate novel hypotheses for functions of lncRNAs. While studies in *Drosophila* and other model systems have provided key insights into mechanisms of dosage compensation, we will focus on recent investigations conducted in mammalian cells.

Xist Spreading

In mammals, the strategy for dosage compensation is reverse from *Drosophila*: female cells selectively repress one entire chromosome by upregulating the repressive lncRNA Xist (Lee, 2012). Xist is transcribed from the X-inactivation center (XIC) and is responsible for physically coating and silencing the X chromosome targeted for the Barr body (the Inactive X, X_i). Another lncRNA, Tsix, is transcribed from the active X chromosome (X_a) and enforces silencing of *Xist* (Lee, 2012). These two lncRNAs, together with others described below, form a complex RNA-protein regulatory network that controls X chromosome dosage compensation in mammals.

Traditional techniques such as immunofluorescence (IF) and RNA fluorescence in situ hybridization (FISH) have been widely applied to study X chromosome inactivation (XCI) and have arrived at a consensus mechanism: elevated Xist expression from the future X_i leads to a cloud-like coating of Xist on X_i and finally epigenetic silencing and chromatin compaction. While informative, IF and RNA-FISH studies had resolution limitations, and as was true for the roX2 RNA, specifically mapping the genomic locations of Xist held the promise of answering mechanistic features of its function. Recently, application of CHART and the development and application of RAP (a method with similar principles as ChIRP and CHART) to the Xist lncRNA defined its precise chromatin association (Engreitz et al., 2013; Simon et al., 2013). Together, the studies revealed that the initiation of Xist spreading occurs from the *Xist* locus to distinct sites across the X chromosome that are not directly adjacent to its locus. These regions are highly accessible by DNaseI footprinting and contain many genes that are actively transcribed prior to silencing. Once Xist is deposited on these early sites, it proceeds to spread and coat the rest of the chromosome to fully silence all but a few genes that escape XCI. It is proposed that the initial deposition process is mediated through higher-order chromatin architecture (Engreitz et al., 2013); however, experimental design differences between the two studies described above make it difficult to directly compare the chromatin conformation results measured. While further investigation is clearly needed to solidify and refine these results, using high-resolution genomic tools (ChIRP, CHART, or RAP) can provide critical insight into lncRNA-controlled systems previously hidden from view.

Mechanisms of Xist Regulation

Intense study of the Xist regulatory network has uncovered many novel lncRNAs in and around the XIC, often illuminating novel mechanistic concepts for how lncRNAs function. Within the lncRNA network that controls Xist, Tsix and Jpx oppose each other's function by repressing or activating, respectively, the transcription of Xist (Lee, 2000; Tian et al., 2010). Recently, additional characterization of the Jpx pathway revealed an unexpected interplay between the lncRNA Jpx and CCCTC-binding factor (CTCF), a major DNA-binding protein involved in higher-order chromosomal folding and interactions (Sun et al., 2013). During female mESC differentiation, CTCF is lost from the *Xist* locus, therefore allowing allele-specific *Xist* upregulation. Molecular characterization of this regulatory loop revealed that CTCF directly binds Jpx and this interaction can titrate CTCF from its DNA targets. Within the conceptual framework of dosage compensation, this puts Jpx and CTCF as central players in the balance between activation and silencing of the X chromosome. Cellular levels of Jpx, as partially determined by the number of X chromosomes, would control the ability of CTCF to bind and inhibit transcription at the *Xist* locus only under conditions when XCI is required. Another recent study more globally characterized the RNA-binding capacity of CTCF and found a multitude of RNA targets, including Wrap53, an lncRNA that controls the induction of the tumor suppressor p53 upon DNA damage (Saldaña-Meyer et al., 2014). Interestingly, biochemical characterization of CTCF's protein domains revealed that the RBR and RNA promoted multimerization of CTCF.

While Xist is modulated by CTCF localization and the spatiotemporal deposition of Xist has been initially defined through CHART and RAP, how Xist interacts with protein effectors of XCI remains poorly understood. The repeat A (RepA) domain of

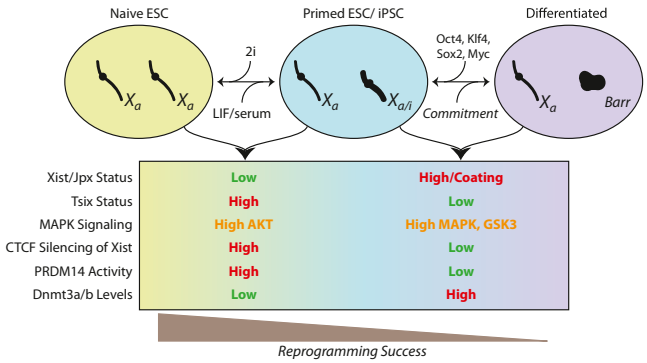


Figure 3. lncRNAs Mark ESC State and Reprogramming Success
X chromosome inactivation (XCI) is a key step in the commitment of ESCs to differentiated cell types. The network on lncRNAs, signaling pathways, and protein effectors that control XCI are depicted. These features can distinguish the stemness of different ESC states and iPSC quality.

Xist has been reported to mediate the interaction with the PRC2 complex (Zhao et al., 2008). Recent characterization of the JARID2 subunit of PRC2 also implicates it in functionally interacting with Xist (da Rocha et al., 2014). Specifically, the authors observed JARID2 and other PCR2 subunits co-occupying genomic regions on the X_i, and a requirement for JARID2 for the deposition of H3K27me3. Further, Xist deletion experiments defined the RepB and RepF regions within the RNA as responsible for JARID2 targeting to the X_i. Interestingly, this function was not depended on its previously identified RBR (Kaneko et al., 2014), suggesting that JARID2 is a multifunctioning RNA-binding protein that mediates the association of PRC2 to the X_i through Xist. These examples suggest that within the context of XCI, as well as during other critical cellular decisions, lncRNAs (such as Xist) can act to modulate chromosome architecture and chromatin modification patterns.

XCI as a Marker of Reprogramming

The ability to transform differentiated cells back into pluripotent cells holds tremendous possibilities for regenerative medicine, but many hurdles still remain before this technology is fully matured (Sánchez Alvarado and Yamanaka, 2014). Because biallelic X activation is a key epigenetic marker of pluripotency, the status of Xist and Xist-mediated gene silencing (or lack thereof) can be exploited to phenotype ESCs and iPSCs (Figure 3). Careful analysis of human iPSCs derived from female cells revealed that many carried an X_i, failing to undergo X chromosome reactivation (XCR), and are epigenetically dynamic, suggesting that the derivation of hiPSCs may not result in pristinely pluripotent cells as desired (Tchieu et al., 2010). A subsequent study used X-inactivation markers to segregate populations of hiPSCs and found that female-derived iPSCs are likely to be less stable in culture than male-derived cells (Anguera et al., 2012). Indeed, erosion of dosage compensation has been observed in female hiPSCs over time in culture, significantly impacting the potential use of these cells for modeling X-linked disease (Mekhoubad et al., 2012). More recent work characterized XCR in the context of iPSC reprogramming and found PRDM14, involved in the ESC pluripotency network, controls Xist silencing (Payer et al., 2013). With the help of Tsix, PRDM14 represses Xist activators (Rnf12 and Jpx) and the *Xist* locus itself by recruiting PRC2, placing PRDM14 expression as a marker for XCR. Work from Heard and colleagues also explored how Xist status can directly regulate ESC differentiation, notably within the framework of the primed/metastable and ground/naive states, with the latter representing a more primordial state of pluripotency. Schultz et al. reported that an X-linked inhibitor of MAPK signaling couples the status of X chromosomes to ESC differentiation. In the ground state where both X chromosomes are active, MAPK is inhibited concomitantly with other molecular changes that block ESC differentiation (Schulz et al., 2014). Upon X chromosome inactivation in the primed state, the relief of MAPK inhibition leads to high MAPK signaling and the capacity to proceed with differentiation. Therefore the characteristic expression of Xist and X-silencing genes provides new ways to evaluate the efficiency and ultimately control of reprograming during iPSC generation. Combining traditional pluripotency markers with new markers like X-inactivation will be critical to achieve the standardization and consistency necessary for clinical application of iPSC technologies.

Lessons and Future Prospects

While the myriad examples to date highlight the functions of a small fraction of known lncRNAs, they illustrate the principle that lncRNAs are intimately involved in the specification, self-renewal, differentiation, and patterning of stem cells and their differentiated progenies. It is reasonable to anticipate that similar principles will be uncovered in many additional organ systems and cell types. A frequently asked question is "Why RNA"? lncRNAs exhibit exquisite cell-type- and organ-specific expression patterns, in fact, to a greater extent than mRNAs. Evolution has probably taken advantage of this fertile soil of cell-type- and state-specific transcription to evolve regulatory functions. Thus, one area of future investigation should focus on the regulation of lncRNA expression—what exactly makes them different and endows them with such state-specific expression? A second challenge for the field is the need to predict the functions of lncRNAs from primary sequence. Finally, understanding how the structure of lncRNAs guides their function remains largely unexplored. As has been true for protein biochemistry, understanding the physical conformations lncRNAs adopt inside cells will undoubtedly uncover novel functional domains and structural elements responsible for their cellular activities.

ACKNOWLEDGMENTS

We apologize to colleagues whose work was not discussed due to space constraints. We thank M. Wernig, E. Heard, R.C. Spitale, A.J. Rubin, and members of the H.Y.C. lab for comments. This work was supported by NIH Medical Scientist Training Program (R.A.F.) and NIH (RO1-CA118750, RO1-ES023168) and California Institute for Regenerative Medicine (H.Y.C.). H.Y.C. is an Early Career Scientist of the Howard Hughes Medical Institute.

REFERENCES

Ang, Y.-S., Tsai, S.-Y., Lee, D.-F., Monk, J., Su, J., Ratnakumar, K., Ding, J., Ge, Y., Darr, H., Chang, B., et al. (2011). Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. Cell *145*, 183–197.

Anguera, M.C., Sadreyev, R., Zhang, Z., Szanto, A., Payer, B., Sheridan, S.D., Kwok, S., Haggarty, S.J., Sur, M., Alvarez, J., et al. (2012). Molecular signatures of human induced pluripotent stem cells highlight sex differences and cancer genes. Cell Stem Cell *11*, 75–90.

Batista, P.J., and Chang, H.Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. Cell *152*, 1298–1307.

Berghoff, E.G., Clark, M.F., Chen, S., Cajigas, I., Leib, D.E., and Kohtz, J.D. (2013). Evf2 (Dlx6as) lncRNA regulates ultraconserved enhancer methylation and the differential transcriptional control of adjacent genes. Development *140*, 4407–4416.

Bertani, S., Sauer, S., Bolotin, E., and Sauer, F. (2011). The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. Mol. Cell *43*, 1040–1046.

Bond, A.M., Vangompel, M.J.W., Sametsky, E.A., Clark, M.F., Savage, J.C., Disterhoft, J.F., and Kohtz, J.D. (2009). Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. Nat. Neurosci. *12*, 1020–1027.

Bosia, C., Pagnani, A., and Zecchina, R. (2013). Modelling competing endogenous RNA networks. PLoS ONE *8*, e66609.

Buganim, Y., Faddah, D.A., Cheng, A.W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S.L., van Oudenaarden, A., and Jaenisch, R. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell *150*, 1209–1222.

Burridge, P.W., Keller, G., Gold, J.D., and Wu, J.C. (2012). Production of de novo cardiomyocytes: human pluripotent stem cell differentiation and direct reprogramming. Cell Stem Cell *10*, 16–28.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell *147*, 358–369.

Chu, C., Qu, K., Zhong, F.L., Artandi, S.E., and Chang, H.Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol. Cell *44*, 667–678.

da Rocha, S.T., Boeva, V., Escamilla-Del-Arenal, M., Ancelin, K., Granier, C., Matias, N.R., Sanulli, S., Chow, J., Schulz, E., Picard, C., et al. (2014). Jarid2 is implicated in the initial Xist-induced targeting of PRC2 to the inactive X chromosome. Mol. Cell *53*, 301–316.

Davidovich, C., Zheng, L., Goodrich, K.J., and Cech, T.R. (2013). Promiscuous RNA binding by Polycomb repressive complex 2. Nat. Struct. Mol. Biol. *20*, 1250–1257.

Deng, Z., Cao, P., Wan, M.M., and Sui, G. (2010). Yin Yang 1: a multifaceted protein beyond a transcription factor. Transcription *1*, 81–84.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Di Ruscio, A., Ebralidze, A.K., Benoukraf, T., Amabile, G., Goff, L.A., Terragni, J., Figueroa, M.E., De Figueiredo Pontes, L.L., Alberich-Jorda, M., Zhang, P., et al. (2013). DNMT1-interacting RNAs block gene-specific DNA methylation. Nature *503*, 371–376.

Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., et al. (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science *341*, 1237973.

Grote, P., and Herrmann, B.G. (2013). The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. RNA Biol. Published online August 22, 2013. http://dx.doi.org/10.4161/rna.26165.

Grote, P., Wittler, L., Hendrix, D., Koch, F., Währisch, S., Beisaw, A., Macura, K., Bläss, G., Kellis, M., Werber, M., and Herrmann, B.G. (2013). The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Dev. Cell *24*, 206–214.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300.

Kaneko, S., Son, J., Shen, S.S., Reinberg, D., and Bonasio, R. (2013). PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. Nat. Struct. Mol. Biol. *20*, 1258–1264.

Kaneko, S., Bonasio, R., Saldaña-Meyer, R., Yoshida, T., Son, J., Nishino, K., Umezawa, A., and Reinberg, D. (2014). Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. Mol. Cell *53*, 290–300.

Klattenhoff, C.A., Scheuermann, J.C., Surface, L.E., Bradley, R.K., Fields, P.A., Steinhauser, M.L., Ding, H., Butty, V.L., Torrey, L., Haas, S., et al. (2013). Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. Cell *152*, 570–583.

Kretz, M., Webster, D.E., Flockhart, R.J., Lee, C.S., Zehnder, A., Lopez-Pajares, V., Qu, K., Zheng, G.X.Y., Chow, J., Kim, G.E., et al. (2012). Suppression of progenitor differentiation requires the long noncoding RNA ANCR. Genes Dev. *26*, 338–343.

Kretz, M., Siprashvili, Z., Chu, C., Webster, D.E., Zehnder, A., Qu, K., Lee, C.S., Flockhart, R.J., Groff, A.F., Chow, J., et al. (2013). Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature *493*, 231–235.

Lee, J.T. (2000). Disruption of imprinted X inactivation by parent-of-origin effects at Tsix. Cell *103*, 17–27.

Lee, J.T. (2012). Epigenetic regulation by long noncoding RNAs. Science *338*, 1435–1439.

Legnini, I., Morlando, M., Mangiavacchi, A., Fatica, A., and Bozzoni, I. (2014). A feedforward regulatory loop between HuR and the long noncoding RNA linc-MD1 controls early phases of myogenesis. Mol. Cell *53*, 506–514.

Li, L., Liu, B., Wapinski, O.L., Tsai, M.-C., Qu, K., Zhang, J., Carlson, J.C., Lin, M., Fang, F., Gupta, R.A., et al. (2013). Targeted disruption of Hotair leads to homeotic transformation and gene derepression. Cell Rep. *5*, 3–12.

Lin, N., Chang, K.-Y., Li, Z., Gates, K., Rana, Z.A., Dang, J., Zhang, D., Han, T., Yang, C.-S., Cunningham, T.J., et al. (2014). An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. Mol. Cell *53*, 1005–1019.

Loewer, S., Cabili, M.N., Guttman, M., Loh, Y.-H., Thomas, K., Park, I.H., Garber, M., Curran, M., Onder, T., Agarwal, S., et al. (2010). Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat. Genet. *42*, 1113–1117.

Lu, L., Zhou, L., Chen, E.Z., Sun, K., Jiang, P., Wang, L., Su, X., Sun, H., and Wang, H. (2012). A novel YY1-miR-1 regulatory circuit in skeletal myogenesis revealed by genome-wide prediction of YY1-miRNA network. PLoS ONE *7*, e27596.

Lu, L., Sun, K., Chen, X., Zhao, Y., Wang, L., Zhou, L., Sun, H., and Wang, H. (2013). Genome-wide survey by ChIP-seq reveals YY1 regulation of lincRNAs in skeletal myogenesis. EMBO J. *32*, 2575–2588.

Lucchesi, J.C., Kelly, W.G., and Panning, B. (2005). Chromatin remodeling in dosage compensation. Annu. Rev. Genet. *39*, 615–651.

Maamar, H., Cabili, M.N., Rinn, J., and Raj, A. (2013). linc-HOXA1 is a noncoding RNA that represses Hoxa1 transcription in cis. Genes Dev. *27*, 1260–1271.

Mali, P., Esvelt, K.M., and Church, G.M. (2013). Cas9 as a versatile tool for engineering biology. Nat. Methods *10*, 957–963.

Margueron, R., and Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. Nature *469*, 343–349.

Mehler, M.F., and Mattick, J.S. (2007). Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease. Physiol. Rev. *87*, 799–823.

Mekhoubad, S., Bock, C., de Boer, A.S., Kiskinis, E., Meissner, A., and Eggan, K. (2012). Erosion of dosage compensation impacts human iPSC disease modeling. Cell Stem Cell *10*, 595–609.

Mercer, T.R., Dinger, M.E., Mariani, J., Kosik, K.S., Mehler, M.F., and Mattick, J.S. (2008). Noncoding RNAs in Long-Term Memory Formation. Neuroscientist *14*, 434–445.

Mercer, T.R., Gerhardt, D.J., Dinger, M.E., Crawford, J., Trapnell, C., Jeddeloh, J.A., Mattick, J.S., and Rinn, J.L. (2012). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat. Biotechnol. *30*, 99–104.

Mukherji, S., Ebert, M.S., Zheng, G.X.Y., Tsang, J.S., Sharp, P.A., and van Oudenaarden, A. (2011). MicroRNAs can generate thresholds in target gene expression. Nat. Genet. *43*, 854–859.

Ng, S.-Y., Johnson, R., and Stanton, L.W. (2012). Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. EMBO J. *31*, 522–533.

Onoguchi, M., Hirabayashi, Y., Koseki, H., and Gotoh, Y. (2012). A noncoding RNA regulates the neurogenin1 gene locus during mouse neocortical development. Proc. Natl. Acad. Sci. USA *109*, 16939–16944.

Pasini, D., Cloos, P.A.C., Walfridsson, J., Olsson, L., Bukowski, J.-P., Johansen, J.V., Bak, M., Tommerup, N., Rappsilber, J., and Helin, K. (2010). JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. Nature *464*, 306–310.

Payer, B., Rosenberg, M., Yamaji, M., Yabuta, Y., Koyanagi-Aoi, M., Hayashi, K., Yamanaka, S., Saitou, M., and Lee, J.T. (2013). Tsix RNA and the germline factor, PRDM14, link X reactivation and stem cell reprogramming. Mol. Cell *52*, 805–818.

Peng, J.C., Valouev, A., Swigut, T., Zhang, J., Zhao, Y., Sidow, A., and Wysocka, J. (2009). Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. Cell *139*, 1290–1302.

Ramos, A.D., Diaz, A., Nellore, A., Delgado, R.N., Park, K.-Y., Gonzales-Roybal, G., Oldham, M.C., Song, J.S., and Lim, D.A. (2013). Integration of genome-wide approaches identifies lncRNAs of adult neural stem cells and their progeny in vivo. Cell Stem Cell *12*, 616–628.

Rapicavoli, N.A., Poth, E.M., Zhu, H., and Blackshaw, S. (2011). The long noncoding RNA Six3OS acts in trans to regulate retinal development by modulating Six3 activity. Neural Dev. *6*, 32.

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., and Chang, H.Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311–1323.

Saldaña-Meyer, R., González-Buendía, E., Guerrero, G., Narendra, V., Bonasio, R., Recillas-Targa, F., and Reinberg, D. (2014). CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53. Genes Dev. *28*, 723–734.

Sánchez Alvarado, A., and Yamanaka, S. (2014). Rethinking differentiation: stem cells, regeneration, and plasticity. Cell *157*, 110–119.

Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., Spence, M., et al. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. Elife *2*, e01749.

Schulz, E.G., Meisig, J., Nakamura, T., Okamoto, I., Sieber, A., Picard, C., Borensztein, M., Saitou, M., Blüthgen, N., and Heard, E. (2014). The two active X chromosomes in female ESCs block exit from the pluripotent state by modulating the ESC signaling network. Cell Stem Cell *14*, 203–216.

Sen, G.L., Reuter, J.A., Webster, D.E., Zhu, L., and Khavari, P.A. (2010). DNMT1 maintains progenitor function in self-renewing somatic tissue. Nature *463*, 563–567.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236–240.

Sheik Mohamed, J., Gaughwin, P.M., Lim, B., Robson, P., and Lipovich, L. (2010). Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells. RNA *16*, 324–337.

Shen, X., Kim, W., Fujiwara, Y., Simon, M.D., Liu, Y., Mysliwiec, M.R., Yuan, G.-C., Lee, Y., and Orkin, S.H. (2009). Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells. Cell *139*, 1303–1314.

Simon, M.D., Wang, C.I., Kharchenko, P.V., West, J.A., Chapman, B.A., Alekseyenko, A.A., Borowsky, M.L., Kuroda, M.I., and Kingston, R.E. (2011). The genomic binding sites of a noncoding RNA. Proc. Natl. Acad. Sci. USA *108*, 20497–20502.

Simon, M.D., Pinter, S.F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S.K., Kesner, B.A., Maier, V.K., Kingston, R.E., and Lee, J.T. (2013). High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. Nature *504*, 465–469.

Sun, S., Del Rosario, B.C., Szanto, A., Ogawa, Y., Jeon, Y., and Lee, J.T. (2013). Jpx RNA activates Xist by evicting CTCF. Cell *153*, 1537–1551.

Tay, Y., Rinn, J., and Pandolfi, P.P. (2014). The multilayered complexity of ceRNA crosstalk and competition. Nature *505*, 344–352.

Tchieu, J., Kuoy, E., Chin, M.H., Trinh, H., Patterson, M., Sherman, S.P., Aimiuwu, O., Lindgren, A., Hakimian, S., Zack, J.A., et al. (2010). Female human iPSCs retain an inactive X chromosome. Cell Stem Cell *7*, 329–342.

Tian, D., Sun, S., and Lee, J.T. (2010). The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. Cell *143*, 390–403.

Truong, A.B., Kretz, M., Ridky, T.W., Kimmel, R., and Khavari, P.A. (2006). p63 regulates proliferation and differentiation of developmentally mature keratinocytes. Genes Dev. *20*, 3185–3197.

Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. Science *329*, 689–693.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537–1550.

Wang, K.C., and Chang, H.Y. (2011). Molecular mechanisms of long noncoding RNAs. Mol. Cell *43*, 904–914.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., et al. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.

Wang, Y., Xu, Z., Jiang, J., Xu, C., Kang, J., Xiao, L., Wu, M., Xiong, J., Guo, X., and Liu, H. (2013). Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. Dev. Cell *25*, 69–80.

Yang, Y.W., Flynn, R.A., Chen, Y., Qu, K., Wan, B., Wang, K.C., Lei, M., and Chang, H.Y. (2014). Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. Elife *3*, e02046.

Zhang, X., Lian, Z., Padden, C., Gerstein, M.B., Rozowsky, J., Snyder, M., Gingeras, T.R., Kapranov, P., Weissman, S.M., and Newburger, P.E. (2009). A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. Blood *113*, 2526–2534.

Zhang, A., Zhou, N., Huang, J., Liu, Q., Fukuda, K., Ma, D., Lu, Z., Bai, C., Watabe, K., and Mo, Y.-Y. (2013). The human long non-coding RNA-RoR is a p53 repressor in response to DNA damage. Cell Res. *23*, 340–350.

Zhao, J., Sun, B.K., Erwin, J.A., Song, J.J., and Lee, J.T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science *322*, 750–756.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol. Cell *40*, 939–953.

A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity

Martin Jinek,[1,2]* Krzysztof Chylinski,[3,4]* Ines Fonfara,[4] Michael Hauer,[2]† Jennifer A. Doudna,[1,2,5,6]‡ Emmanuelle Charpentier[4]‡

Clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems provide bacteria and archaea with adaptive immunity against viruses and plasmids by using CRISPR RNAs (crRNAs) to guide the silencing of invading nucleic acids. We show here that in a subset of these systems, the mature crRNA that is base-paired to trans-activating crRNA (tracrRNA) forms a two-RNA structure that directs the CRISPR-associated protein Cas9 to introduce double-stranded (ds) breaks in target DNA. At sites complementary to the crRNA-guide sequence, the Cas9 HNH nuclease domain cleaves the complementary strand, whereas the Cas9 RuvC-like domain cleaves the noncomplementary strand. The dual-tracrRNA:crRNA, when engineered as a single RNA chimera, also directs sequence-specific Cas9 dsDNA cleavage. Our study reveals a family of endonucleases that use dual-RNAs for site-specific DNA cleavage and highlights the potential to exploit the system for RNA-programmable genome editing.

Bacteria and archaea have evolved RNA-mediated adaptive defense systems called clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) that protect organisms from invading viruses and plasmids (*1*–*3*). These defense systems rely on small RNAs for sequence-specific detection and silencing of foreign nucleic acids. CRISPR/Cas systems are composed of *cas* genes organized in operon(s) and CRISPR array(s) consisting of genome-targeting sequences (called spacers) interspersed with identical repeats (*1*–*3*). CRISPR/Cas-mediated immunity occurs in three steps. In the adaptive phase, bacteria and archaea harboring one or more CRISPR loci respond to viral or plasmid challenge by integrating short fragments of foreign sequence (protospacers) into the host chromosome at the proximal end of the CRISPR array (*1*–*3*). In the expression and interference phases, transcription of the repeat-spacer element into precursor CRISPR RNA (pre-crRNA) molecules followed by enzymatic cleavage yields the short crRNAs that can pair with complementary protospacer sequences of invading viral or plasmid targets (*4*–*11*). Target recognition by crRNAs directs the silencing of the foreign sequences by means of Cas proteins that function in complex with the crRNAs (*10*, *12*–*20*).

There are three types of CRISPR/Cas systems (*21*–*23*). The type I and III systems share some overarching features: specialized Cas endonucleases process the pre-crRNAs, and once mature, each crRNA assembles into a large multi-Cas protein complex capable of recognizing and cleaving nucleic acids complementary to the crRNA. In contrast, type II systems process pre-crRNAs by a different mechanism in which a trans-activating crRNA (tracrRNA) complementary to the repeat sequences in pre-crRNA triggers processing by the double-stranded (ds) RNA-specific ribonuclease RNase III in the presence of the Cas9 (formerly Csn1) protein (fig. S1) (*4*, *24*). Cas9 is thought to be the sole protein responsible for crRNA-guided silencing of foreign DNA (*25*–*27*).

We show here that in type II systems, Cas9 proteins constitute a family of enzymes that require a base-paired structure formed between the activating tracrRNA and the targeting crRNA to cleave target dsDNA. Site-specific cleavage occurs at locations determined by both base-pairing complementarity between the crRNA and the target protospacer DNA and a short motif [referred to as the protospacer adjacent motif (PAM)] juxtaposed to the complementary region in the target DNA. Our study further demonstrates that the Cas9 endonuclease family can be programmed with single RNA molecules to cleave specific DNA sites, thereby raising the exciting possibility of developing a simple and versatile RNA-directed system to generate dsDNA breaks for genome targeting and editing.

Cas9 is a DNA endonuclease guided by two RNAs. Cas9, the hallmark protein of type II systems, has been hypothesized to be involved in both crRNA maturation and crRNA-guided DNA interference (fig. S1) (*4*, *25*–*27*). Cas9 is involved in crRNA maturation (*4*), but its direct participation in target DNA destruction has not been investigated. To test whether and how Cas9 might be capable of target DNA cleavage, we used an overexpression system to purify Cas9 protein derived from the pathogen *Streptococcus pyogenes* (fig. S2, see supplementary materials and methods) and tested its ability to cleave a plasmid DNA or an oligonucleotide duplex bearing a protospacer sequence complementary to a mature crRNA, and a bona fide PAM. We found that mature crRNA alone was incapable of directing Cas9-catalyzed plasmid DNA cleavage (Fig. 1A and fig. S3A). However, addition of tracrRNA, which can pair with the repeat sequence of crRNA and is essential to crRNA maturation in this system, triggered Cas9 to cleave plasmid DNA (Fig. 1A and fig. S3A). The cleavage reaction required both magnesium and the presence of a crRNA sequence complementary to the DNA; a crRNA capable of tracrRNA base pairing but containing a noncognate target DNA-binding sequence did not support Cas9-catalyzed plasmid cleavage (Fig. 1A; fig. S3A, compare crRNA-sp2 to crRNA-sp1; and fig. S4A). We obtained similar results with a short linear dsDNA substrate (Fig. 1B and fig. S3, B and C). Thus, the trans-activating tracrRNA is a small noncoding RNA with two critical functions: triggering pre-crRNA processing by the enzyme RNase III (*4*) and subsequently activating crRNA-guided DNA cleavage by Cas9.

Cleavage of both plasmid and short linear dsDNA by tracrRNA:crRNA-guided Cas9 is site-specific (Fig. 1, C to E, and fig. S5, A and B). Plasmid DNA cleavage produced blunt ends at a position three base pairs upstream of the PAM sequence (Fig. 1, C and E, and fig. S5, A and C) (*26*). Similarly, within short dsDNA duplexes, the DNA strand that is complementary to the target-binding sequence in the crRNA (the complementary strand) is cleaved at a site three base pairs upstream of the PAM (Fig. 1, D and E, and fig. S5, B and C). The noncomplementary DNA strand is cleaved at one or more sites within three to eight base pairs upstream of the PAM. Further investigation revealed that the noncomplementary strand is first cleaved endonucleolytically and subsequently trimmed by a 3′-5′ exonuclease activity (fig. S4B). The cleavage rates by Cas9 under single-turnover conditions ranged from 0.3 to 1 min^{-1}, comparable to those of restriction endonucleases (fig. S6A), whereas incubation of wild-type (WT) Cas9-tracrRNA:crRNA complex with a fivefold molar excess of substrate DNA provided evidence that the dual-RNA–guided Cas9 is a multiple-turnover enzyme (fig. S6B). In

[1]Howard Hughes Medical Institute (HHMI), University of California, Berkeley, CA 94720, USA. [2]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA. [3]Max F. Perutz Laboratories (MFPL), University of Vienna, A-1030 Vienna, Austria. [4]The Laboratory for Molecular Infection Medicine Sweden, Umeå Centre for Microbial Research, Department of Molecular Biology, Umeå University, S-90187 Umeå, Sweden. [5]Department of Chemistry, University of California, Berkeley, CA 94720, USA. [6]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

*These authors contributed equally to this work.
†Present address: Friedrich Miescher Institute for Biomedical Research, 4058 Basel, Switzerland.
‡To whom correspondence should be addressed. E-mail: doudna@berkeley.edu (J.A.D.); emmanuelle.charpentier@mims.umu.se (E.C.)

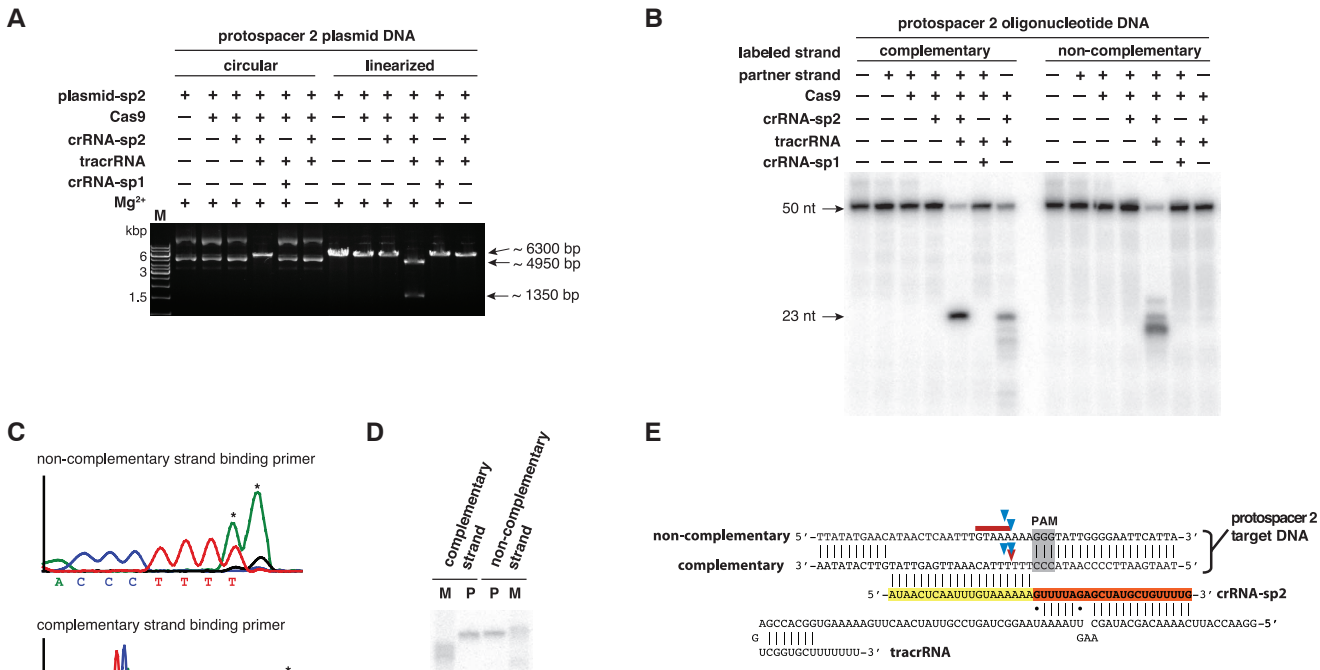


Fig. 1. Cas9 is a DNA endonuclease guided by two RNA molecules. (**A**) Cas9 was programmed with a 42-nucleotide crRNA-sp2 (crRNA containing a spacer 2 sequence) in the presence or absence of 75-nucleotide tracrRNA. The complex was added to circular or XhoI-linearized plasmid DNA bearing a sequence complementary to spacer 2 and a functional PAM. crRNA-sp1, specificity control; M, DNA marker; kbp, kilo–base pair. See fig. S3A. (**B**) Cas9 was programmed with crRNA-sp2 and tracrRNA (nucleotides 4 to 89). The complex was incubated with double- or single-stranded DNAs harboring a sequence complementary to spacer 2 and a functional PAM (*4*). The complementary or noncomplementary strands of the DNA were 5′-radiolabeled and annealed with a nonlabeled partner strand. nt, nucleotides. See fig. S3, B and C. (**C**) Sequencing analysis of cleavage products from Fig. 1A. Termination of primer extension in the sequencing reaction indicates the position of the cleavage site. The 3′ terminal A overhang (asterisks) is an artifact of the sequencing reaction. See fig. S5, A and C. (**D**) The cleavage products from Fig. 1B were analyzed alongside 5′ end-labeled size markers derived from the complementary and noncomplementary strands of the target DNA duplex. M, marker; P, cleavage product. See fig. S5, B and C. (**E**) Schematic representation of tracrRNA, crRNA-sp2, and protospacer 2 DNA sequences. Regions of crRNA complementarity to tracrRNA (orange) and the protospacer DNA (yellow) are represented. The PAM sequence is shown in gray; cleavage sites mapped in (C) and (D) are represented by blue arrows (C), a red arrow [(D), complementary strand], and a red line [(D), noncomplementary strand].

contrast to the CRISPR type I Cascade complex (*18*), Cas9 cleaves both linearized and supercoiled plasmids (Figs. 1A and 2A). Therefore, an invading plasmid can, in principle, be cleaved multiple times by Cas9 proteins programmed with different crRNAs.

Each Cas9 nuclease domain cleaves one DNA strand. Cas9 contains domains homologous to both HNH and RuvC endonucleases (Fig. 2A and fig. S7) (*21*–*23*, *27*, *28*). We designed and purified Cas9 variants containing inactivating point mutations in the catalytic residues of either the HNH or RuvC-like domains (Fig. 2A and fig. S7) (*23*, *27*). Incubation of these variant Cas9 proteins with native plasmid DNA showed that dual-RNA–guided mutant Cas9 proteins yielded nicked open circular plasmids, whereas the WT Cas9 protein-tracrRNA:crRNA complex produced a linear DNA product (Figs. 1A and 2A and figs. S3A and S8A). This result indicates that the Cas9 HNH and RuvC-like domains each cleave one plasmid DNA strand. To determine which strand of the target DNA is cleaved by each Cas9 catalytic domain, we incubated the mutant Cas9-tracrRNA:crRNA complexes with short dsDNA substrates in which either the complementary or noncomplementary strand was radiolabeled at its 5′ end. The resulting cleavage products indicated that the Cas9 HNH domain cleaves the complementary DNA strand, whereas the Cas9 RuvC-like domain cleaves the noncomplementary DNA strand (Fig. 2B and fig. S8B).

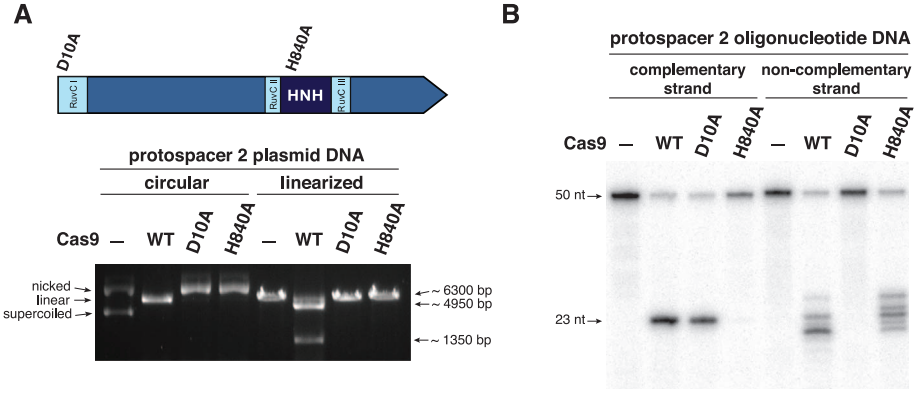


Fig. 2. Cas9 uses two nuclease domains to cleave the two strands in the target DNA. (**A**) (Top) Schematic representation of Cas9 domain structure showing the positions of domain mutations. D10A, Asp^{10}→Ala^{10}; H840A; His^{840}→Ala^{840}. (Bottom) Complexes of WT or nuclease mutant Cas9 proteins with tracrRNA: crRNA-sp2 were assayed for endonuclease activity as in Fig. 1A. (**B**) Complexes of WT Cas9 or nuclease domain mutants with tracrRNA and crRNA-sp2 were tested for activity as in Fig. 1B.

Dual-RNA requirements for target DNA binding and cleavage. tracrRNA might be required for target DNA binding and/or to stimulate the nuclease activity of Cas9 downstream of target recognition. To distinguish between these possibilities, we used an electrophoretic mobility shift assay to monitor target DNA binding by catalytically inactive Cas9 in the presence or absence of crRNA and/or tracrRNA. Addition of tracrRNA substantially enhanced target DNA binding by Cas9, whereas we observed little specific DNA binding with Cas9 alone or Cas9-crRNA (fig. S9). This indicates that tracrRNA is required for target DNA recognition, possibly by properly orienting the crRNA for interaction with the complementary strand of target DNA. The predicted tracrRNA:crRNA secondary structure includes base pairing between the 22 nucleotides at the 3′ terminus of the crRNA and a segment near the 5′ end of the mature tracrRNA (Fig. 1E). This interaction creates a structure in which the 5′-terminal 20 nucleotides of the crRNA, which vary in sequence in different crRNAs, are available for target DNA binding. The bulk of the tracrRNA downstream of the crRNA base-pairing region is free to form additional RNA structure(s) and/or to interact with Cas9 or the target DNA site. To determine whether the entire length of the tracrRNA is necessary for site-specific Cas9-catalyzed DNA cleavage, we tested Cas9-tracrRNA:crRNA complexes reconstituted using full-length mature (42-nucleotide) crRNA and various truncated forms of tracrRNA lacking sequences at their 5′ or 3′ ends. These complexes were tested for cleavage using a short target dsDNA. A substantially truncated version of the tracrRNA retaining nucleotides 23 to 48 of the native sequence was capable of supporting robust dual-RNA–guided Cas9-catalyzed DNA cleavage (Fig. 3, A and C, and fig. S10, A and B). Truncation of the crRNA from either end showed that Cas9-catalyzed cleavage in the presence of tracrRNA could be triggered with crRNAs missing the 3′-terminal 10 nucleotides (Fig. 3, B and C). In contrast, a 10-nucleotide deletion from the 5′ end of crRNA abolished DNA cleavage by Cas9 (Fig. 3B). We also analyzed Cas9 orthologs from various bacterial species for their ability to support *S. pyogenes* tracrRNA:crRNA-guided DNA cleavage. In contrast to closely related *S. pyogenes* Cas9 orthologs, more distantly related orthologs were not functional in the cleavage reaction (fig. S11). Similarly, *S. pyogenes* Cas9 guided by tracrRNA:crRNA duplexes originating from more distant systems was unable to cleave DNA efficiently (fig. S11). Species specificity of dual-RNA–guided cleavage of DNA indicates coevolution of Cas9, tracrRNA, and the crRNA repeat, as well as the existence of a still unknown structure and/or sequence in the dual-RNA that is critical for the formation of the ternary complex with specific Cas9 orthologs.

To investigate the protospacer sequence requirements for type II CRISPR/Cas immunity in bacterial cells, we analyzed a series of protospacer-containing plasmid DNAs harboring single-nucleotide mutations for their maintenance following transformation in *S. pyogenes* and their ability to be cleaved by Cas9 in vitro. In contrast to point mutations introduced at the 5′ end of the protospacer, mutations in the region close to the PAM and the Cas9 cleavage sites were not tolerated in vivo and resulted in decreased plasmid cleavage efficiency in vitro (Fig. 3D). Our results are in agreement with a previous report of protospacer escape mutants selected in the type II CRISPR system from *S. thermophilus* in vivo (*27*, *29*). Furthermore, the plasmid maintenance and cleavage results hint at the existence of a "seed" region located at the 3′ end of the protospacer sequence that is crucial for the interaction with crRNA and subsequent cleavage by Cas9. In support of this notion, Cas9 enhanced complementary DNA strand hybridization to the crRNA; this enhancement was the strongest in the 3′-terminal region of the crRNA targeting sequence (fig. S12). Corroborating this finding, a contiguous stretch of at least 13 base pairs between the crRNA and the target DNA site proximal to the PAM is required for efficient target cleavage, whereas up to six contiguous mismatches in the 5′-terminal region of the protospacer are tolerated (Fig. 3E). These findings are reminiscent of the previously observed seed-sequence requirements for target nucleic acid recognition in Argonaute proteins

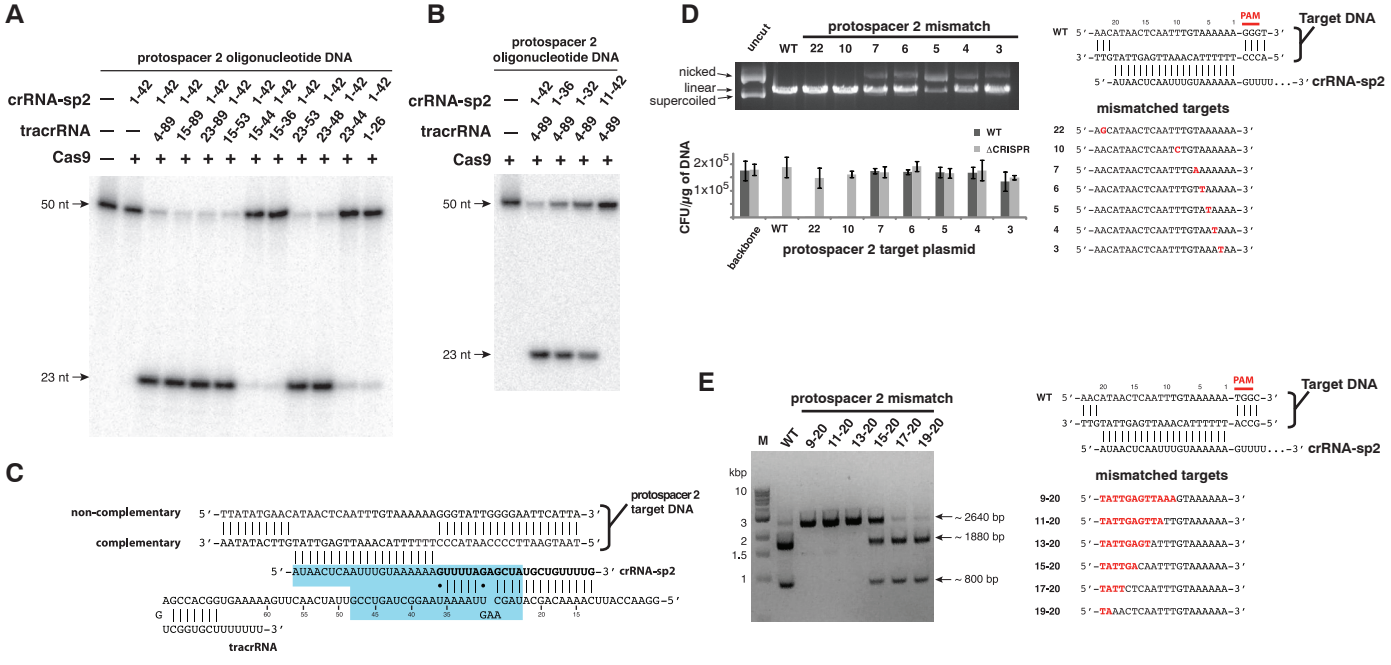


Fig. 3. Cas9-catalyzed cleavage of target DNA requires an activating domain in tracrRNA and is governed by a seed sequence in the crRNA. (**A**) Cas9-tracrRNA:crRNA complexes were reconstituted using 42-nucleotide crRNA-sp2 and truncated tracrRNA constructs and were assayed for cleavage activity as in Fig. 1B. (**B**) Cas9 programmed with full-length tracrRNA and crRNA-sp2 truncations was assayed for activity as in (A). (**C**) Minimal regions of tracrRNA and crRNA capable of guiding Cas9-mediated DNA cleavage (blue shaded region). (**D**) Plasmids containing WT or mutant protospacer 2 sequences with indicated point mutations (right) were cleaved in vitro by programmed Cas9 as in Fig. 1A (top-left) and used for transformation assays of WT or pre-crRNA–deficient *S. pyogenes* (bottom-left). The transformation efficiency was calculated as colony-forming units (CFU) per microgram of plasmid DNA. Error bars represent SDs for three biological replicates. (**E**) Plasmids containing WT and mutant protospacer 2 inserts with varying extent of crRNA-target DNA mismatches (right) were cleaved in vitro by programmed Cas9 (left). The cleavage reactions were further digested with XmnI. The 1880- and 800-bp fragments are Cas9-generated cleavage products. M, DNA marker.

(*30*, *31*) and the Cascade and Csy CRISPR complexes (*13*, *14*).

A short sequence motif dictates R-loop formation. In multiple CRISPR/Cas systems, recognition of self versus nonself has been shown to involve a short sequence motif that is preserved in the foreign genome, referred to as the PAM (*27*, *29*, *32*–*34*). PAM motifs are only a few base pairs in length, and their precise sequence and position vary according to the CRISPR/Cas system type (*32*). In the *S. pyogenes* type II system, the PAM conforms to an NGG consensus sequence, containing two G:C base pairs that occur one base pair downstream of the crRNA binding sequence, within the target DNA (*4*). Transformation assays demonstrated that the GG motif is essential for protospacer plasmid DNA elimination by CRISPR/Cas in bacterial cells (fig. S13A), consistent with previous observations in *S. thermophilus* (*27*). The motif is also essential for in vitro protospacer plasmid cleavage by tracrRNA:crRNA-guided Cas9 (fig. S13B). To determine the role of the PAM in target DNA cleavage by the Cas9-tracrRNA:crRNA complex, we tested a series of dsDNA duplexes containing mutations in the PAM sequence on the complementary or noncomplementary strands, or both (Fig. 4A). Cleavage assays using these substrates showed that Cas9-catalyzed DNA cleavage was particularly sensitive to mutations in the PAM sequence on the noncomplementary strand of the DNA, in contrast to complementary strand PAM recognition by type I CRISPR/Cas systems (*18*, *34*). Cleavage of target single-stranded DNAs was unaffected by mutations of the PAM motif. This observation suggests that the PAM motif is required only in the context of target dsDNA and may thus be required to license duplex unwinding, strand invasion, and the formation of an R-loop structure. When we used a different crRNA-target DNA pair (crRNA-sp4 and protospacer 4 DNA), selected due to the presence of a canonical PAM not present in the protospacer 2 target DNA, we found that both G nucleotides of the PAM were required for efficient Cas9-catalyzed DNA cleavage (Fig. 4B and fig. S13C). To determine whether the PAM plays a direct role in recruiting the Cas9-tracrRNA:crRNA complex to the correct target DNA site, we analyzed binding affinities of the complex for target DNA sequences by native gel mobility shift assays (Fig. 4C). Mutation of either G in the PAM sequence substantially reduced the affinity of Cas9-tracrRNA:crRNA for the target DNA. This finding argues for specific recognition of the PAM sequence by Cas9 as a prerequisite for target DNA binding and possibly strand separation to allow strand invasion and R-loop formation, which would be analogous to the PAM sequence recognition by CasA/Cse1 implicated in a type I CRISPR/Cas system (*34*).

Cas9 can be programmed with a single chimeric RNA. Examination of the likely secondary structure of the tracrRNA:crRNA duplex (Figs. 1E and 3C) suggested the possibility that the features required for site-specific Cas9-catalyzed DNA cleavage could be captured in a single chimeric RNA. Although the tracrRNA:crRNA

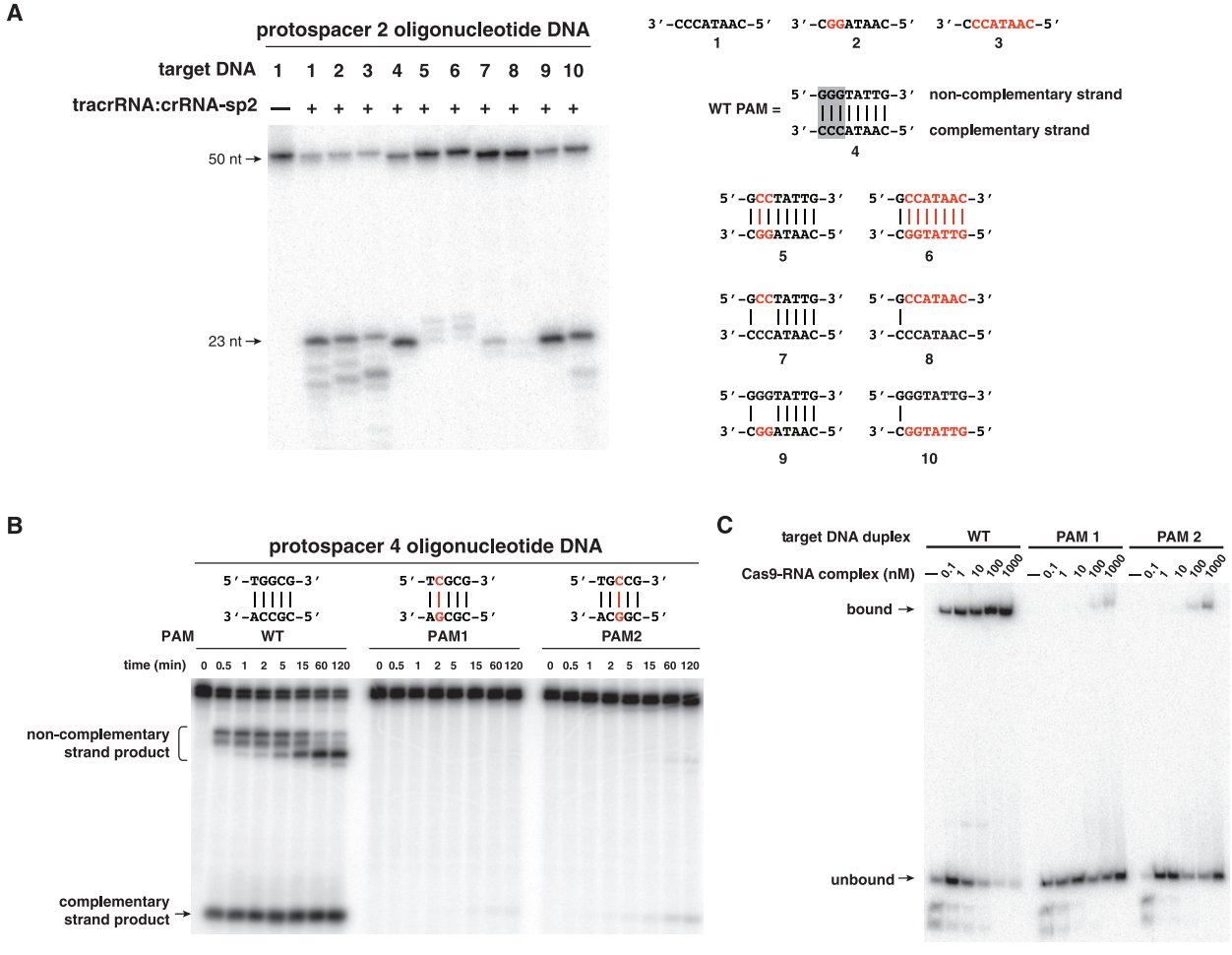


Fig. 4. A PAM is required to license target DNA cleavage by the Cas9-tracrRNA:crRNA complex. (**A**) Dual RNA-programmed Cas9 was tested for activity as in Fig. 1B. WT and mutant PAM sequences in target DNAs are indicated (right). (**B**) Protospacer 4 target DNA duplexes (labeled at both 5′ ends) containing WT and mutant PAM motifs were incubated with Cas9 programmed with tracrRNA:crRNA-sp4 (nucleotides 23 to 89). At the indicated time points (in minutes), aliquots of the cleavage reaction were taken and analyzed as in Fig. 1B. (**C**) Electrophoretic mobility shift assays were performed using RNA-programmed Cas9 (D10A/H840A) and protospacer 4 target DNA duplexes [same as in (B)] containing WT and mutated PAM motifs. The Cas9 (D10A/H840A)–RNA complex was titrated from 100 pM to 1 μM.

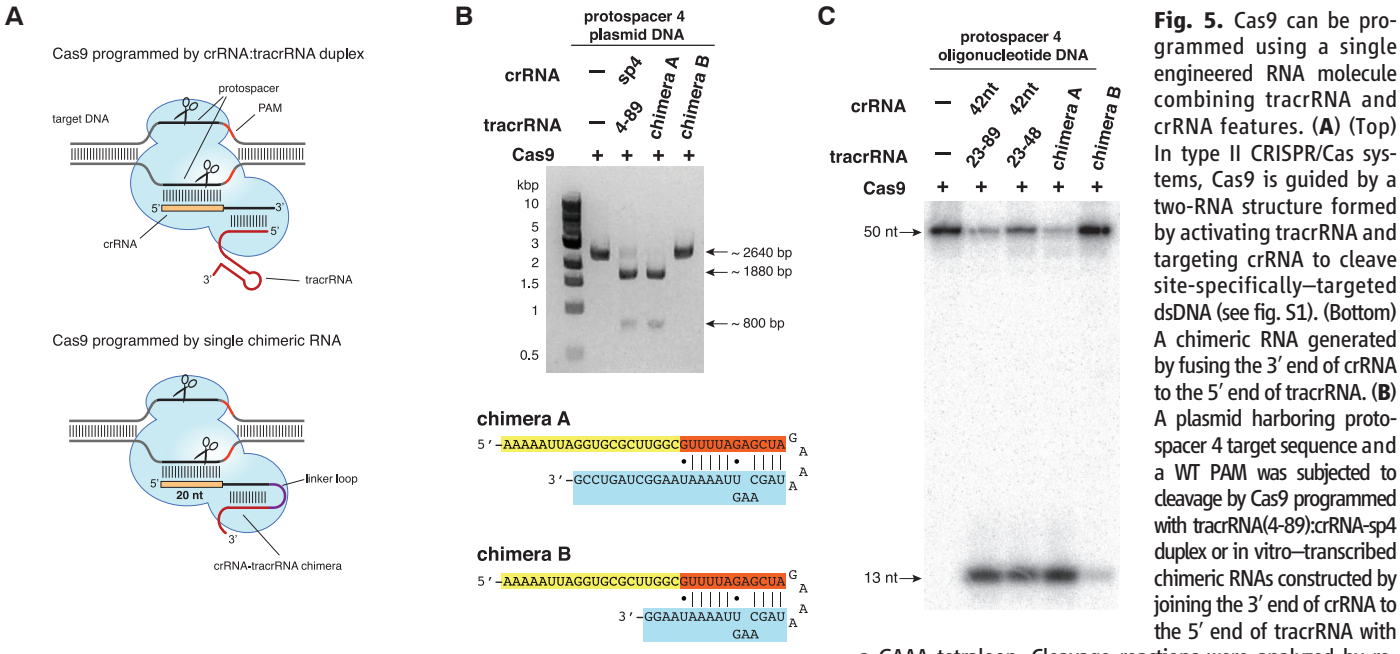


Fig. 5. Cas9 can be programmed using a single engineered RNA molecule combining tracrRNA and crRNA features. (**A**) (Top) In type II CRISPR/Cas systems, Cas9 is guided by a two-RNA structure formed by activating tracrRNA and targeting crRNA to cleave site-specifically–targeted dsDNA (see fig. S1). (Bottom) A chimeric RNA generated by fusing the 3′ end of crRNA to the 5′ end of tracrRNA. (**B**) A plasmid harboring protospacer 4 target sequence and a WT PAM was subjected to cleavage by Cas9 programmed with tracrRNA(4-89):crRNA-sp4 duplex or in vitro–transcribed chimeric RNAs constructed by joining the 3′ end of crRNA to the 5′ end of tracrRNA with a GAAA tetraloop. Cleavage reactions were analyzed by restriction mapping with XmnI. Sequences of chimeric RNAs A and B are shown with DNA-targeting (yellow), crRNA repeat-derived sequences (orange), and tracrRNA-derived (light blue) sequences. (**C**) Protospacer 4 DNA duplex cleavage reactions were performed as in Fig. 1B. (**D**) Five chimeric RNAs designed to target the GFP gene were used to program Cas9 to cleave a GFP gene–containing plasmid. Plasmid cleavage reactions were performed as in Fig. 3E, except that the plasmid DNA was restriction mapped with AvrII after Cas9 cleavage.

target-selection mechanism works efficiently in nature, the possibility of a single RNA-guided Cas9 is appealing due to its potential utility for programmed DNA cleavage and genome editing (Fig. 5A). We designed two versions of a chimeric RNA containing a target recognition sequence at the 5′ end followed by a hairpin structure retaining the base-pairing interactions that occur between the tracrRNA and the crRNA (Fig. 5B). This single transcript effectively fuses the 3′ end of crRNA to the 5′ end of tracrRNA, thereby mimicking the dual-RNA structure required to guide site-specific DNA cleavage by Cas9. In cleavage assays using plasmid DNA, we observed that the longer chimeric RNA was able to guide Cas9-catalyzed DNA cleavage in a manner similar to that observed for the truncated tracrRNA:crRNA duplex (Fig. 5B and fig. S14, A and C). The shorter chimeric RNA did not work efficiently in this assay, confirming that nucleotides that are 5 to 12 positions beyond the tracrRNA:crRNA base-pairing interaction are important for efficient Cas9 binding and/or target recognition. We obtained similar results in cleavage assays using short dsDNA as a substrate, further indicating that the position of the cleavage site in target DNA is identical to that observed using the dual tracrRNA:crRNA as a guide (Fig. 5C and fig. S14, B and C). Finally, to establish whether the design of chimeric RNA might be universally applicable, we engineered five different chimeric guide RNAs to target a portion of the gene encoding the green-fluorescent protein (GFP) (fig. S15, A to C) and tested their efficacy against a plasmid carrying the GFP coding sequence in vitro. In all five cases, Cas9 programmed with these chimeric RNAs efficiently cleaved the plasmid at the correct target site (Fig. 5D and fig. S15D), indicating that rational design of chimeric RNAs is robust and could, in principle, enable targeting of any DNA sequence of interest with few constraints beyond the presence of a GG dinucleotide adjacent to the targeted sequence.

Conclusions. We identify a DNA interference mechanism involving a dual-RNA structure that directs a Cas9 endonuclease to introduce site-specific double-stranded breaks in target DNA. The tracrRNA:crRNA-guided Cas9 protein makes use of distinct endonuclease domains (HNH and RuvC-like domains) to cleave the two strands in the target DNA. Target recognition by Cas9 requires both a seed sequence in the crRNA and a GG dinucleotide-containing PAM sequence adjacent to the crRNA-binding region in the DNA target. We further show that the Cas9 endonuclease can be programmed with guide RNA engineered as a single transcript to target and cleave any dsDNA sequence of interest. The system is efficient, versatile, and programmable by changing the DNA target-binding sequence in the guide chimeric RNA. Zinc-finger nucleases and transcription-activator–like effector nucleases have attracted considerable interest as artificial enzymes engineered to manipulate genomes (*35*–*38*). We propose an alternative methodology based on RNA-programmed Cas9 that could offer considerable potential for gene-targeting and genome-editing applications.

References and Notes

1. B. Wiedenheft, S. H. Sternberg, J. A. Doudna, *Nature* **482**, 331 (2012).
2. D. Bhaya, M. Davison, R. Barrangou, *Annu. Rev. Genet.* **45**, 273 (2011).
3. M. P. Terns, R. M. Terns, *Curr. Opin. Microbiol.* **14**, 321 (2011).
4. E. Deltcheva *et al.*, *Nature* **471**, 602 (2011).
5. J. Carte, R. Wang, H. Li, R. M. Terns, M. P. Terns, *Genes Dev.* **22**, 3489 (2008).
6. R. E. Haurwitz, M. Jinek, B. Wiedenheft, K. Zhou, J. A. Doudna, *Science* **329**, 1355 (2010).
7. R. Wang, G. Preamplume, M. P. Terns, R. M. Terns, H. Li, *Structure* **19**, 257 (2011).
8. E. M. Gesner, M. J. Schellenberg, E. L. Garside, M. M. George, A. M. Macmillan, *Nat. Struct. Mol. Biol.* **18**, 688 (2011).
9. A. Hatoum-Aslan, I. Maniv, L. A. Marraffini, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 21218 (2011).
10. S. J. J. Brouns *et al.*, *Science* **321**, 960 (2008).
11. D. G. Sashital, M. Jinek, J. A. Doudna, *Nat. Struct. Mol. Biol.* **18**, 680 (2011).
12. N. G. Lintner *et al.*, *J. Biol. Chem.* **286**, 21643 (2011).

13. E. Semenova *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **108**, 10098 (2011).
14. B. Wiedenheft *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **108**, 10092 (2011).
15. B. Wiedenheft *et al*., *Nature* **477**, 486 (2011).
16. C. R. Hale *et al*., *Cell* **139**, 945 (2009).
17. J. A. L. Howard, S. Delmas, I. Ivančić-Baće, E. L. Bolt, *Biochem. J.* **439**, 85 (2011).
18. E. R. Westra *et al*., *Mol. Cell* **46**, 595 (2012).
19. C. R. Hale *et al*., *Mol. Cell* **45**, 292 (2012).
20. J. Zhang *et al*., *Mol. Cell* **45**, 303 (2012).
21. K. S. Makarova *et al*., *Nat. Rev. Microbiol.* **9**, 467 (2011).
22. K. S. Makarova, N. V. Grishin, S. A. Shabalina, Y. I. Wolf, E. V. Koonin, *Biol. Direct* **1**, 7 (2006).
23. K. S. Makarova, L. Aravind, Y. I. Wolf, E. V. Koonin, *Biol. Direct* **6**, 38 (2011).
24. S. Gottesman, *Nature* **471**, 588 (2011).
25. R. Barrangou *et al*., *Science* **315**, 1709 (2007).
26. J. E. Garneau *et al*., *Nature* **468**, 67 (2010).
27. R. Sapranauskas *et al*., *Nucleic Acids Res.* **39**, 9275 (2011).
28. G. K. Taylor, D. F. Heiter, S. Pietrokovski, B. L. Stoddard, *Nucleic Acids Res.* **39**, 9705 (2011).
29. H. Deveau *et al*., *J. Bacteriol.* **190**, 1390 (2008).
30. B. P. Lewis, C. B. Burge, D. P. Bartel, *Cell* **120**, 15 (2005).
31. G. Hutvagner, M. J. Simard, *Nat. Rev. Mol. Cell Biol.* **9**, 22 (2008).
32. F. J. M. Mojica, C. Díez-Villaseñor, J. García-Martínez, C. Almendros, *Microbiology* **155**, 733 (2009).
33. L. A. Marraffini, E. J. Sontheimer, *Nature* **463**, 568 (2010).
34. D. G. Sashital, B. Wiedenheft, J. A. Doudna, *Mol. Cell* **46**, 606 (2012).
35. M. Christian *et al*., *Genetics* **186**, 757 (2010).
36. J. C. Miller *et al*., *Nat. Biotechnol.* **29**, 143 (2011).
37. F. D. Urnov, E. J. Rebar, M. C. Holmes, H. S. Zhang, P. D. Gregory, *Nat. Rev. Genet.* **11**, 636 (2010).
38. D. Carroll, *Gene Ther.* **15**, 1463 (2008).

Acknowledgments: We thank K. Zhou, A. M. Smith, R. Haurwitz and S. Sternberg for excellent technical assistance; members of the Doudna and Charpentier laboratories and J. Cate for comments on the manuscript; and B. Meyer and T.-W. Lo (Univ. of California, Berkeley/HHMI) for providing the GFP plasmid. This work was funded by the HHMI (M.J. and J.A.D.), the Austrian Science Fund (grant W1207-B09; K.C. and E.C.), the Univ. of Vienna (K.C.), the Swedish Research Council (grants K2010-57X-21436-01-3 and 621-2011-5752-LiMS; E.C.), the Kempe Foundation (E.C.), and Umeå University (K.C. and E.C.). J.A.D. is an Investigator and M.J. is a Research Specialist of the HHMI. K.C. is a fellow of the Austrian Doctoral Program in RNA Biology and is cosupervised by R. Schroeder. We thank A. Witte, U. Bläsi, and R. Schroeder for helpful discussions, financial support to K.C., and for hosting K.C. in their laboratories at MFPL. M.J., K.C., J.A.D., and E.C. have filed a related patent.

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1225829/DC1
Materials and Methods
Figs. S1 to S15
Tables S1 to S3
References (*39*–*47*)

8 June 2012; accepted 20 June 2012
Published online 28 June 2012;
10.1126/science.1225829

LETTER

doi:10.1038/nature13769

Programmable RNA recognition and cleavage by CRISPR/Cas9

Mitchell R. O'Connell[1], Benjamin L. Oakes[1], Samuel H. Sternberg[2], Alexandra East-Seletsky[1], Matias Kaplan[3]† & Jennifer A. Doudna[1,2,3,4]

The CRISPR-associated protein Cas9 is an RNA-guided DNA endonuclease that uses RNA–DNA complementarity to identify target sites for sequence-specific double-stranded DNA (dsDNA) cleavage[1–5]. In its native context, Cas9 acts on DNA substrates exclusively because both binding and catalysis require recognition of a short DNA sequence, known as the protospacer adjacent motif (PAM), next to and on the strand opposite the twenty-nucleotide target site in dsDNA[4–7]. Cas9 has proven to be a versatile tool for genome engineering and gene regulation in a large range of prokaryotic and eukaryotic cell types, and in whole organisms[8], but it has been thought to be incapable of targeting RNA[5]. Here we show that Cas9 binds with high affinity to single-stranded RNA (ssRNA) targets matching the Cas9-associated guide RNA sequence when the PAM is presented in *trans* as a separate DNA oligonucleotide. Furthermore, PAM-presenting oligonucleotides (PAMmers) stimulate site-specific endonucleolytic cleavage of ssRNA targets, similar to PAM-mediated stimulation of Cas9-catalysed DNA cleavage[7]. Using specially designed PAMmers, Cas9 can be specifically directed to bind or cut RNA targets while avoiding corresponding DNA sequences, and we demonstrate that this strategy enables the isolation of a specific endogenous messenger RNA from cells. These results reveal a fundamental connection between PAM binding and substrate selection by Cas9, and highlight the utility of Cas9 for programmable transcript recognition without the need for tags.

CRISPR–Cas immune systems must discriminate between self and non-self to avoid an autoimmune response[9]. In type I and II systems, foreign DNA targets that contain adjacent PAM sequences are targeted for degradation, whereas potential targets in CRISPR loci of the host do not contain PAMs and are avoided by RNA-guided interference complexes[3,5,6,10]. Single-molecule and bulk biochemical experiments showed that PAMs act both to recruit Cas9–guide-RNA (Cas9–gRNA) complexes to potential target sites and to trigger nuclease domain activation[7]. Cas9 from *Streptococcus pyogenes* recognizes a 5′-NGG-3′ PAM on the non-target (displaced) DNA strand[4,6], suggesting that PAM recognition may stimulate catalysis through allosteric regulation. Moreover, the HNH nuclease domain of Cas9, which mediates target-strand cleavage[4,5], is homologous to other HNH domains that cleave RNA substrates[11,12]. Based on the observations that single-stranded DNA (ssDNA) targets can be activated for cleavage by a separate PAMmer[7], and that similar HNH domains can cleave RNA, we wondered whether a similar strategy would enable Cas9 to cleave ssRNA targets in a programmable fashion (Fig. 1a).

Using *S. pyogenes* Cas9 and dual-guide RNAs (Methods), we performed *in vitro* cleavage experiments using a panel of RNA and DNA targets

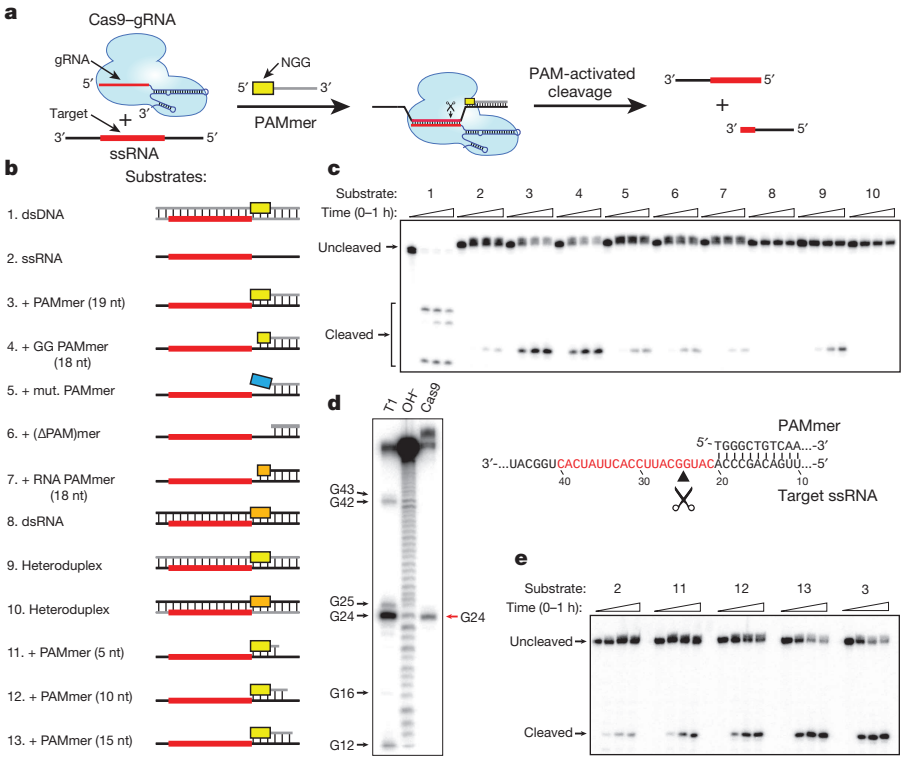


Figure 1 | RNA-guided Cas9 cleaves ssRNA targets in the presence of a short PAM-presenting DNA oligonucleotide (PAMmer). **a**, Schematic depicting the approach used to target ssRNA for programmable, sequence-specific cleavage. **b**, The panel of nucleic acid substrates examined in this study. Substrate elements are coloured as follows: DNA, grey; RNA, black; guide-RNA target sequence, red; DNA PAM, yellow; mutated DNA PAM, blue; RNA PAM, orange. The 18-nucleotide 'GG PAMmer' contains only a GG dinucleotide PAM sequence. nt, nucleotide. **c**, Representative cleavage assay for 5′-radiolabelled nucleic acid substrates using Cas9–gRNA, numbered as in **b**. **d**, Cas9–gRNA cleavage site mapping assay for substrate 3. T1 and OH^- denote RNase T1 and hydrolysis ladders, respectively; the sequence of the target ssRNA is shown at right. Sites of G cleavage by RNase T1 are shown at left. Site of Cas9 cleavage (G24) shown at right. **e**, Representative ssRNA cleavage assay in the presence of PAMmers of increasing length, numbered as in **b**.

[1]Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. [2]Department of Chemistry, University of California, Berkeley, California 94720, USA. [3]Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA. [4]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. †Present address: Department of Agricultural and Biological Engineering, University of Florida, Gainesville, Florida 32611, USA.

(Fig. 1b and Extended Data Table 1). Deoxyribonucleotide-comprised PAMmers specifically activated Cas9 to cleave ssRNA (Fig. 1c), an effect that required a 5′-NGG-3′ or 5′-GG-3′ PAM. RNA cleavage was not observed using ribonucleotide-based PAMmers, suggesting that Cas9 may recognize the local helical geometry and/or deoxyribose moieties within the PAM. Consistent with this hypothesis, dsRNA targets were not cleavable and RNA–DNA heteroduplexes could only be cleaved when the non-target strand was composed of deoxyribonucleotides. Notably, we found that Cas9 cleaved the ssRNA target strand between positions 4 and 5 of the base-paired gRNA–target-RNA hybrid (Fig. 1d), in contrast to the cleavage between positions 3 and 4 observed for dsDNA[3–5]. This is probably due to subtle differences in substrate positioning. However, we did observe a significant reduction in the pseudo-first-order cleavage rate constant of PAMmer-activated ssRNA as compared to ssDNA[7] (Extended Data Fig. 1).

We hypothesized that PAMmer nuclease activation would depend on the stability of the hybridized PAMmer–ssRNA duplex and tested this by varying PAMmer length. As expected, ssRNA cleavage was lost when the predicted melting temperature for the duplex decreased below the temperature used in our experiments (Fig. 1e). In addition, large molar excesses of di- or tri-deoxyribonucleotides in solution were poor activators of Cas9 cleavage (Extended Data Fig. 2). Collectively, these data demonstrate that hybrid substrate structures composed of ssRNA and deoxyribonucleotide-based PAMmers that anneal upstream of the RNA target sequence can be cleaved efficiently by RNA-guided Cas9.

We investigated the binding affinity of catalytically inactive dCas9 (Cas9 (D10A;H840A))–gRNA for ssRNA targets with and without PAMmers using a gel mobility shift assay. Notably, whereas our previous results showed that ssDNA and PAMmer-activated ssDNA targets are bound with indistinguishable affinity[7], PAMmer-activated ssRNA targets were bound >500-fold tighter than ssRNA alone (Fig. 2a, b). A recent crystal structure of Cas9 bound to a ssDNA target revealed deoxyribose-specific van der Waals interactions between the protein and the DNA backbone[13], suggesting that energetic penalties associated with ssRNA binding must be attenuated by favourable compensatory binding interactions with the provided PAM. The equilibrium dissociation constant measured for a PAMmer–ssRNA substrate was within fivefold of that for dsDNA (Fig. 2b), and this high-affinity interaction again required a cognate deoxyribonucleotide-comprised 5′-GG-3′ PAM (Fig. 2a). Tight binding also scaled with PAMmer length (Fig. 2c), consistent with the cleavage data presented above.

It is known that Cas9 possesses an intrinsic affinity for RNA, but sequence specificity of the interaction had not been explored[5]. Thus, to verify the programmable nature of PAMmer-mediated ssRNA cleavage by Cas9–gRNA, we prepared three distinct guide RNAs (λ2, λ3 and λ4; each targeting 20-nucleotide sequences within λ2, λ3 and λ4 RNAs, respectively) and showed that their corresponding ssRNA targets could be efficiently cleaved using complementary PAMmers without any detectable cross-reactivity (Fig. 3a). This result indicates that complementary RNA–RNA base pairing is critical in these reactions. Notably however, dCas9 programmed with the λ2 guide RNA bound all three PAMmer–ssRNA substrates with similar affinity (Fig. 3b). This observation suggests that high-affinity binding in this case may not require correct base pairing between the guide RNA and the ssRNA target, particularly given the compensatory role of the PAMmer.

During dsDNA targeting by Cas9–gRNA, duplex melting proceeds directionally from the PAM and strictly requires the formation of complementary RNA–DNA base pairs to offset the energetic costs associated with dsDNA unwinding[7]. We therefore wondered whether binding specificity for ssRNA substrates would be recovered using PAMmers containing 5′-extensions that create a partially double-stranded target region requiring unwinding (Fig. 3c). We found that use of a 5′-extended PAMmer enabled dCas9 bearing the λ2 guide sequence to bind sequence-selectively to the λ2 PAMmer–ssRNA target. The λ3 and λ4 PAMmer–ssRNA targets were not recognized (Fig. 3d and Extended Data Fig. 3),

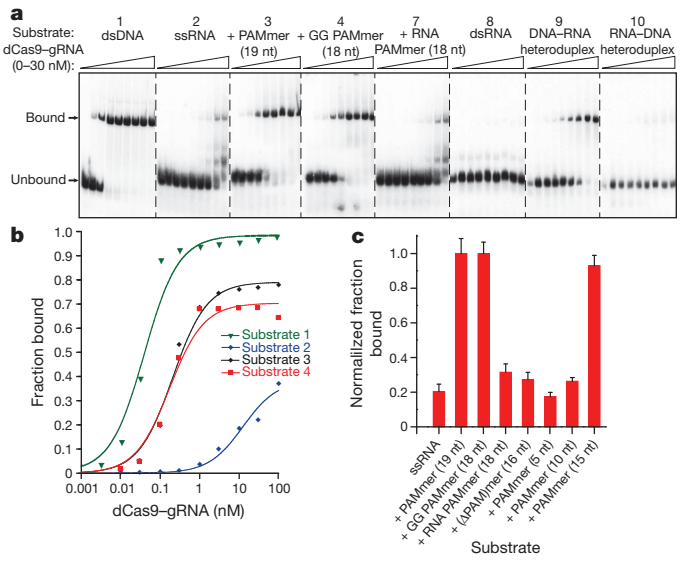


Figure 2 | dCas9–gRNA binds ssRNA targets with high affinity in the presence of PAMmers. **a**, Representative electrophoretic mobility shift assay for binding reactions with dCas9–gRNA and a panel of 5′-radiolabelled nucleic acid substrates, numbered as in Fig. 1b. **b**, Quantified binding data for substrates 1–4 from **a** fitted with standard binding isotherms. Measured dissociation constants from three independent experiments (mean ± s.d.) were 0.036 ± 0.003 nM (substrate 1), >100 nM (substrate 2), 0.20 ± 0.09 nM (substrate 3) and 0.18 ± 0.07 nM (substrate 4). **c**, Relative binding data for 1 nM dCas9–gRNA and 5′-radiolabelled ssRNA with a panel of different PAMmers. The data are normalized to the amount of binding observed at 1 nM dCas9–gRNA with a 19-nucleotide (nt) PAMmer; error bars represent the standard deviation from three independent experiments.

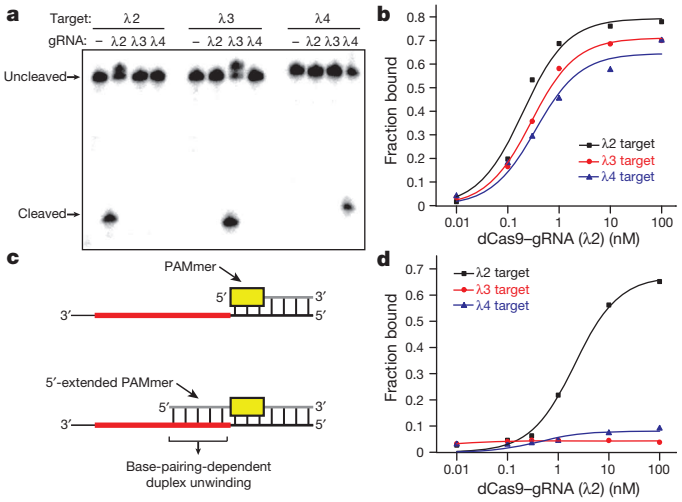


Figure 3 | 5′-extended PAMmers are required for specific target ssRNA binding. **a**, Cas9 programmed with either λ2-, λ3- or λ4-targeting gRNAs exhibits sequence-specific cleavage of 5′-radiolabelled λ2, λ3 and λ4 target ssRNAs, respectively, in the presence of cognate PAMmers. **b**, dCas9 programmed with a λ2-targeting gRNA exhibits similar binding affinity to λ2, λ3 and λ4 target ssRNAs in the presence of cognate PAMmers. Dissociation constants from three independent experiments (mean ± s.d.) were 0.20 ± 0.09 nM (λ2), 0.33 ± 0.14 nM (λ3) and 0.53 ± 0.21 nM (λ4). **c**, Schematic depicting the approach used to restore gRNA-mediated ssRNA binding specificity, which involves 5′-extensions to the PAMmer that cover part or all of the target sequence. **d**, dCas9 programmed with a λ2-targeting gRNA specifically binds the λ2 ssRNA but not λ3 and λ4 ssRNAs in the presence of complete 5′-extended PAMmers. Dissociation constants from three independent experiments (mean ± s.d.) were 3.3 ± 1.2 nM (λ2) and >100 nM (λ3 and λ4).

although we did observe a tenfold reduction in overall ssRNA substrate binding affinity. By systematically varying the length of the 5′ extension, we found that PAMmers containing 2–8 additional nucleotides upstream of the 5′-NGG-3′ offer an optimal compromise between gains in binding specificity and concomitant losses in binding affinity and cleavage efficiency (Extended Data Fig. 4).

Next we investigated whether nuclease activation by PAMmers requires base pairing between the 5′-NGG-3′ and corresponding nucleotides on the ssRNA. Prior studies have shown that DNA substrates containing a cognate PAM that is mismatched with the corresponding nucleotides on the target strand are cleaved as efficiently as a fully base-paired PAM[4]. This could enable targeting of RNA while precluding binding or cleavage of corresponding genomic DNA sites lacking PAMs (Fig. 4a). To test this possibility, we first demonstrated that Cas9–gRNA cleaves PAMmer–ssRNA substrates regardless of whether or not the PAM is base paired (Fig. 4b, c). When Cas9–RNA was incubated with both a PAMmer–ssRNA substrate and the corresponding dsDNA template containing a cognate PAM, both targets were cleaved. In contrast, when a dsDNA target lacking a PAM was incubated together with a PAMmer–ssRNA substrate bearing a mismatched 5′-NGG-3′ PAM, Cas9–gRNA selectively targeted the ssRNA for cleavage (Fig. 4c). The same result was obtained using a mismatched PAMmer with a 5′ extension (Fig. 4c), demonstrating that this general strategy enables the specific targeting of RNA transcripts while effectively eliminating any targeting of their corresponding dsDNA template loci.

We next explored whether Cas9-mediated RNA targeting could be applied in tagless transcript isolation from HeLa cells (Fig. 4d). The immobilization of Cas9 on a solid-phase resin is described in Methods (see also Extended Data Fig. 5). As a proof of concept, we first isolated *GAPDH* mRNA from HeLa total RNA using biotinylated dCas9, gRNAs and PAMmers (Extended Data Table 2) that target four non-PAM-adjacent sequences within exons 5–7 (Fig. 4e). We observed a substantial enrichment of *GAPDH* mRNA relative to control β-actin mRNA by northern blot analysis, but saw no enrichment using a non-targeting gRNA or dCas9 alone (Fig. 4f).

We then used this approach to isolate endogenous *GAPDH* transcripts from HeLa cell lysate under physiological conditions. In initial experiments, we found that Cas9–gRNA captured two *GAPDH*-specific RNA fragments rather than the full-length mRNA (Fig. 4g). Based on the sizes of these bands, we hypothesized that RNA–DNA heteroduplexes formed between the mRNA and PAMmer were cleaved by cellular RNase H. Previous studies have shown that modified DNA oligonucleotides can abrogate RNase H activity[14], and therefore we investigated whether Cas9 would tolerate chemical modifications to the PAMmer. We found that a wide range of modifications (locked nucleic acids, 2′-OMe and 2′-F ribose moieties) still enabled PAMmer-mediated nuclease

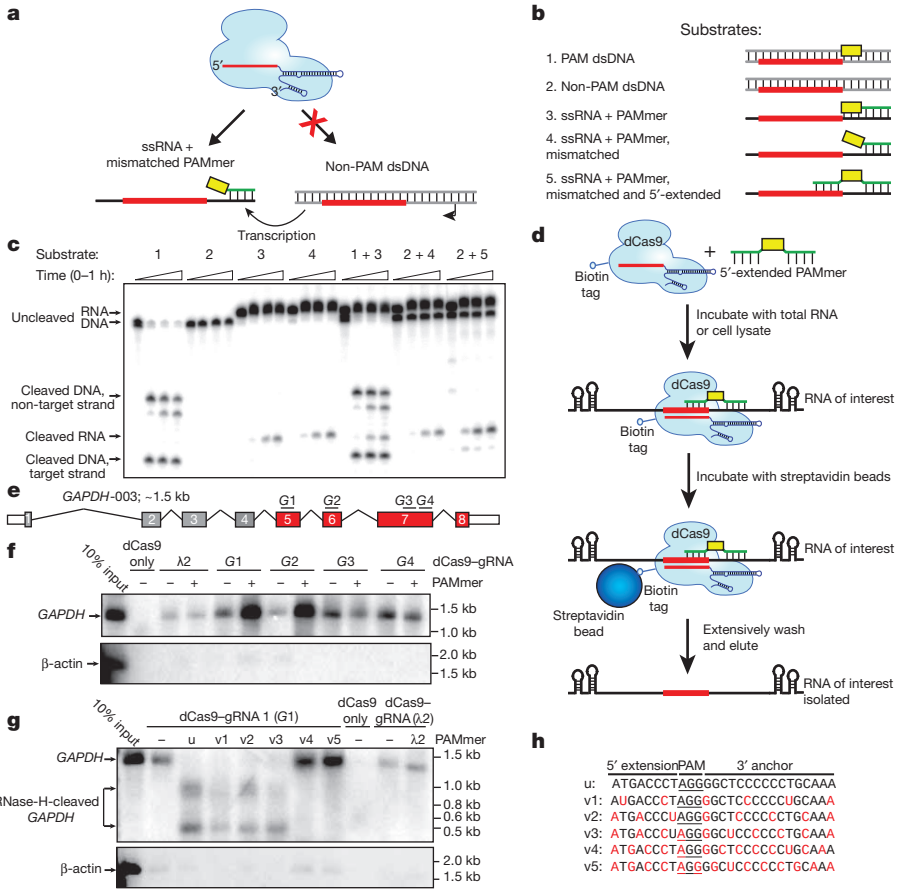


Figure 4 | RNA-guided Cas9 can target non-PAM sites on ssRNA and isolate *GAPDH* mRNA from HeLa cells in a tagless manner. **a**, Schematic of the approach designed to avoid cleavage of template DNA by targeting non-PAM sites in the ssRNA target. **b**, The panel of nucleic acid substrates tested in **c**. **c**, Cas9–gRNA cleaves ssRNA targets with equal efficiency when the 5′-NGG-3′ of the PAMmer is mismatched with the ssRNA. This strategy enables selective cleavage of ssRNA in the presence of non-PAM target dsDNA. **d**, Schematic of the dCas9 RNA pull-down experiment. **e**, *GAPDH* mRNA transcript isoform 3 (*GAPDH*-003) shown schematically, with exons common to all *GAPDH* protein-coding transcripts in red and gRNA/PAMmer targets *G1*–*G4* indicated. kb, kilobase pairs. **f**, Northern blot showing that gRNAs and corresponding 5′-extended PAMmers enable tagless isolation of *GAPDH* mRNA from HeLa total RNA; β-actin mRNA is shown as a control. **g**, Northern blot showing tagless isolation of *GAPDH* mRNA from HeLa cell lysate with varying 2′-OMe-modified PAMmers. RNase H cleavage is abrogated with v4 and v5 PAMmers; β-actin mRNA is shown as a control. u, unmodified PAMmer (*G1*). v1-v5, increasingly 2′-OMe-modified PAMmers (*G1*), see **g** for PAMmer sequences. **h**, Sequences of unmodified and modified *GAPDH* PAMmers used in **g**; 2′-OMe-modified nucleotides are shown in red.

activation (Extended Data Fig. 6). Furthermore, by varying the pattern of 2′-OMe modifications in the PAMmer, we could completely eliminate RNase-H-mediated cleavage during the pull-down and successfully isolate intact *GAPDH* mRNA (Fig. 4g, h). Notably, we consistently observed specific isolation of *GAPDH* mRNA in the absence of any PAMmer, albeit with lower efficiency, suggesting that Cas9–gRNA can bind to *GAPDH* mRNA through direct RNA–RNA hybridization (Fig. 4f, g and Extended Data Fig. 7). These experiments demonstrate that RNA-guided Cas9 can be used to purify endogenous untagged RNA transcripts. In contrast to current oligonucleotide-mediated RNA-capture methods, this approach works well under physiological salt conditions and does not require crosslinking or large sets of biotinylated probes[15–17].

Here we have demonstrated the ability to re-direct the dsDNA targeting capability of CRISPR/Cas9 for RNA-guided ssRNA binding and/or cleavage (which we now denote RCas9, an RNA-targeting Cas9). Programmable RNA recognition and cleavage has the potential to transform the study of RNA function, much as site-specific DNA targeting is changing the landscape of genetic and genomic research[8] (Extended Data Fig. 8). Although certain engineered proteins such as PPR proteins and Pumilio/FBF (PUF) repeats show promise as platforms for sequence-specific RNA targeting[18–22], these strategies require re-designing the protein for every new RNA sequence of interest. While RNA interference has proven useful for manipulating gene regulation in certain organisms[23], there has been a strong motivation to develop orthogonal nucleic-acid-based RNA recognition systems, such as the CRISPR/Cas Type III-B Cmr complex[24–28] and the atypical Cas9 from *Francisella novicida*[29,30]. In contrast to these systems, the molecular basis for RNA recognition by RCas9 is now clear and requires only the design and synthesis of a matching gRNA and complementary PAMmer. The ability to recognize endogenous RNAs within complex mixtures with high affinity and in a programmable manner paves the way for direct transcript detection, analysis and manipulation without the need for genetically encoded affinity tags.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 April; accepted 14 August 2014.
Published online 28 September 2014.

1. Wiedenheft, B., Sternberg, S. H. & Doudna, J. A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482,** 331–338 (2012).
2. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315,** 1709–1712 (2007).
3. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468,** 67–71 (2010).
4. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821 (2012).
5. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl Acad. Sci. USA* **109,** E2579–E2586 (2012).
6. Mojica, F. J. M., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiol* **155,** 733–740 (2009).
7. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–67 (2014).
8. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nature Methods* **10,** 957–963 (2013).
9. Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463,** 568–571 (2010).
10. Sashital, D. G., Wiedenheft, B. & Doudna, J. A. Mechanism of foreign DNA selection in a bacterial adaptive immune system. *Mol. Cell* **46,** 606–615 (2012).
11. Pommer, A. J. *et al.* Mechanism and cleavage specificity of the H-N-H endonuclease colicin E9. *J. Mol. Biol.* **314,** 735–749 (2001).
12. Hsia, K. C. *et al.* DNA binding and degradation by the HNH protein ColE7. *Structure* **12,** 205–214 (2004).
13. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156,** 935–949 (2014).
14. Wu, H. J., Lima, W. F. & Crooke, S. T. Properties of cloned and expressed human RNase H1. *J. Biol. Chem.* **274,** 28270–28278 (1999).
15. Engreitz, J. M. *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341** (2013).
16. Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. & Chang, H. Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA–chromatin interactions. *Mol. Cell* **44,** 667–678 (2011).
17. Simon, M. D. *et al.* The genomic binding sites of a noncoding RNA. *Proc. Natl Acad. Sci. USA* **108,** 20497–20502 (2011).
18. Mackay, J. P., Font, J. & Segal, D. J. The prospects for designer single-stranded RNA-binding proteins. *Nature Struct. Mol. Biol.* **18,** 256–261 (2011).
19. Filipovska, A. & Rackham, O. Designer RNA-binding proteins: new tools for manipulating the transcriptome. *RNA Biol.* **8,** 978–983 (2011).
20. Wang, Y., Wang, Z. & Tanaka Hall, T. M. Engineered proteins with Pumilio/fem-3 mRNA binding factor scaffold to manipulate RNA metabolism. *FEBS J.* **280,** 3755–3767 (2013).
21. Yin, P. *et al.* Structural basis for the modular recognition of single-stranded RNA by PPR proteins. *Nature* **504,** 168–171 (2013).
22. Yagi, Y., Nakamura, T. & Small, I. The potential for manipulating RNA with pentatricopeptide repeat proteins. *Plant J.* **78,** 772–782 (2014).
23. Kim, D. H. & Rossi, J. J. RNAi mechanisms and applications. *Biotechniques* **44,** 613–616 (2008).
24. Hale, C. R. *et al.* RNA-guided RNA cleavage by a CRISPR RNA–Cas protein complex. *Cell* **139,** 945–956 (2009).
25. Hale, C. R. *et al.* Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol. Cell* **45,** 292–302 (2012).
26. Staals, R. H. J. *et al.* Structure and activity of the RNA-targeting type III-B CRISPR-Cas complex of *Thermus thermophilus*. *Mol. Cell* **52,** 135–145 (2013).
27. Spilman, M. *et al.* Structure of an RNA silencing complex of the CRISPR-Cas immune system. *Mol. Cell* **52,** 146–152 (2013).
28. Terns, R. M. & Terns, M. P. CRISPR-based technologies: prokaryotic defense weapons repurposed. *Trends Genet.* **30,** 111–118 (2014).
29. Sampson, T. R., Saroj, S. D., Llewellyn, A. C., Tzeng, Y. L. & Weiss, D. S. A. CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature* **497,** 254–257 (2013).
30. Sampson, T. R. & Weiss, D. S. Exploiting CRISPR/Cas systems for biotechnology. *Bioessays* **36,** 34–38 (2014).

Acknowledgements We thank B. Staahl and K. Zhou for technical assistance, A. Iavarone for assistance with mass spectrometry measurements, Integrated DNA Technologies for the synthesis of DNA and RNA oligonucleotides, and members of the Doudna laboratory and J. Cate for discussions and critical reading of the manuscript. S.H.S. acknowledges support from the National Science Foundation and National Defense Science & Engineering Graduate Research Fellowship programs. A.E.-S. and B.L.O. acknowledge support from NIH NRSA trainee grants. Funding was provided by the NIH-funded Center for RNA Systems Biology (P50GM102706-03). J.A.D. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions M.R.O. and S.H.S. conceived the project. M.R.O., B.L.O., S.H.S., A.E.-S. and M.K. conducted experiments. All authors discussed the data, and M.R.O., S.H.S., B.L.O. and J.A.D. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.A.D. (doudna@berkeley.edu).

METHODS

Cas9 and nucleic acid preparation. Wild-type Cas9 and catalytically inactive dCas9 (Cas9(D10A;H840A)) from *S. pyogenes* were purified as previously described[4]. Forty-two-nucleotide crRNAs were either ordered synthetically (Integrated DNA Technologies) or transcribed *in vitro* with T7 polymerase using single-stranded DNA templates, as described[31]. Using the previously described numbering scheme[4], tracrRNA was transcribed *in vitro* and contained nucleotides 15–87. Single-guide RNAs (sgRNAs) targeting λ-RNAs were transcribed *in vitro* from linearized plasmids and contain full-length crRNA and tracrRNA connected via a GAAA tetraloop insertion. *GAPDH* mRNA-targeting sgRNAs were transcribed *in vitro* from dsDNA PCR products based on an optimized sgRNA design[32]. Target ssRNAs (55–56 nucleotides) were transcribed *in vitro* using single-stranded DNA templates. Sequences of all nucleic acid substrates used in this study can be found in Extended Data Tables 1 and 2.

All RNAs were purified using 10–15% denaturing polyacrylamide gel electrophoresis (PAGE). Duplexes of crRNA and tracrRNA were prepared by mixing equimolar concentrations of each RNA in hybridization buffer (20 mM Tris-HCl, pH 7.5, 100 mM KCl, 5 mM $MgCl_2$), heating to 95 °C for 30 s and slow cooling. Fully double-stranded DNA/RNA substrates (substrates 1, 8–10 in Fig. 1 and substrates 1 and 2 in Fig. 4) were prepared by mixing equimolar concentrations of each nucleic acid strand in hybridization buffer, heating to 95 °C for 30 s, and slow cooling. RNA, DNA and chemically modified PAMmers were synthesized commercially (Intergrated DNA Technologies). DNA and RNA substrates were 5′-radiolabelled using [γ-^{32}P]ATP (PerkinElmer) and T4 polynucleotide kinase (New England Biolabs). Double-stranded DNA and dsRNA substrates (Figs 1c and 4c) were 5′-radiolabelled on both strands, whereas only the target ssRNA was 5′-radiolabelled in other experiments.

Cleavage assays. Cas9–gRNA complexes were reconstituted before cleavage experiments by incubating Cas9 and the crRNA–tracrRNA duplex for 10 min at 37 °C in reaction buffer (20 mM Tris-HCl, pH 7.5, 75 mM KCl, 5 mM $MgCl_2$, 1 mM dithiothreitol (DTT), 5% glycerol). Cleavage reactions were conducted at 37 °C and contained ~1 nM 5′-radiolabelled target substrate, 100 nM Cas9–RNA, and 100 nM PAMmer, where indicated. Aliquots were removed at each time point and quenched by the addition of RNA gel-loading buffer (95% deionized formamide, 0.025% (w/v) bromophenol blue, 0.025% (w/v) xylene cyanol, 50 mM EDTA (pH 8.0), 0.025% (w/v) SDS). Samples were boiled for 10 min at 95 °C before being resolved by 12% denaturing PAGE. Reaction products were visualized by phosphorimaging and quantified with ImageQuant (GE Healthcare).

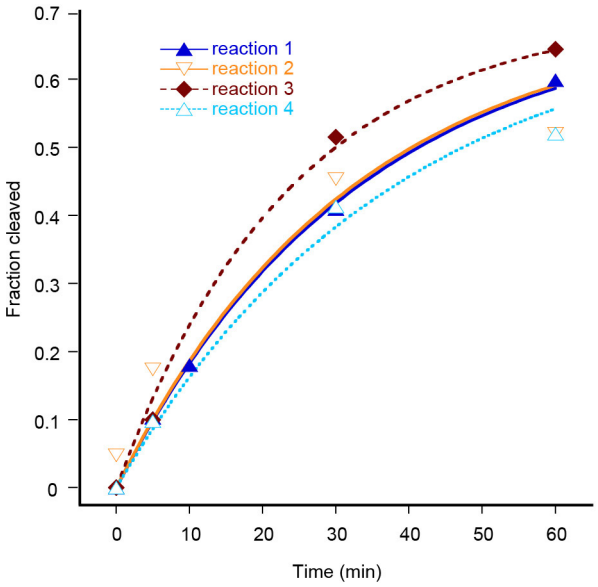
RNA cleavage site mapping. A hydrolysis ladder (OH^-) was obtained by incubating ~25 nM 5′-radiolabelled λ2 target ssRNA in hydrolysis buffer (25 mM CAPS (*N*-cyclohexyl-3-aminopropanesulphonic acid), pH 10.0, 0.25 mM EDTA) at 95 °C for 10 min, before quenching on ice. An RNase T1 ladder was obtained by incubating ~25 nM 5′-radiolabelled λ2 target ssRNA with 1 U RNase T1 (New England Biolabs) for 5 min at 37 °C in RNase T1 buffer (20 mM sodium citrate, pH 5.0, 1 mM EDTA, 2 M urea, 0.1 mg ml^{-1} yeast transfer RNA). The reaction was quenched by phenol/chloroform extraction before adding RNA gel-loading buffer. All products were resolved by 15% denaturing PAGE.

Electrophoretic mobility shift assays. In order to avoid dissociation of the Cas9–gRNA complex at low concentrations during target ssRNA binding experiments, binding reactions contained a constant excess of dCas9 (300 nM), increasing concentrations of sgRNA, and 0.1–1 nM of target ssRNA. The reaction buffer was supplemented with 10 μg ml^{-1} heparin in order to avoid non-specific association of *apo*-dCas9 with target substrates[7]. Reactions were incubated at 37 °C for 45 min before being resolved by 8% native PAGE at 4 °C (0.5 × TBE buffer with 5 mM $MgCl_2$). RNA and DNA were visualized by phosphorimaging, quantified with ImageQuant (GE Healthcare), and analysed with Kaleidagraph (Synergy Software).

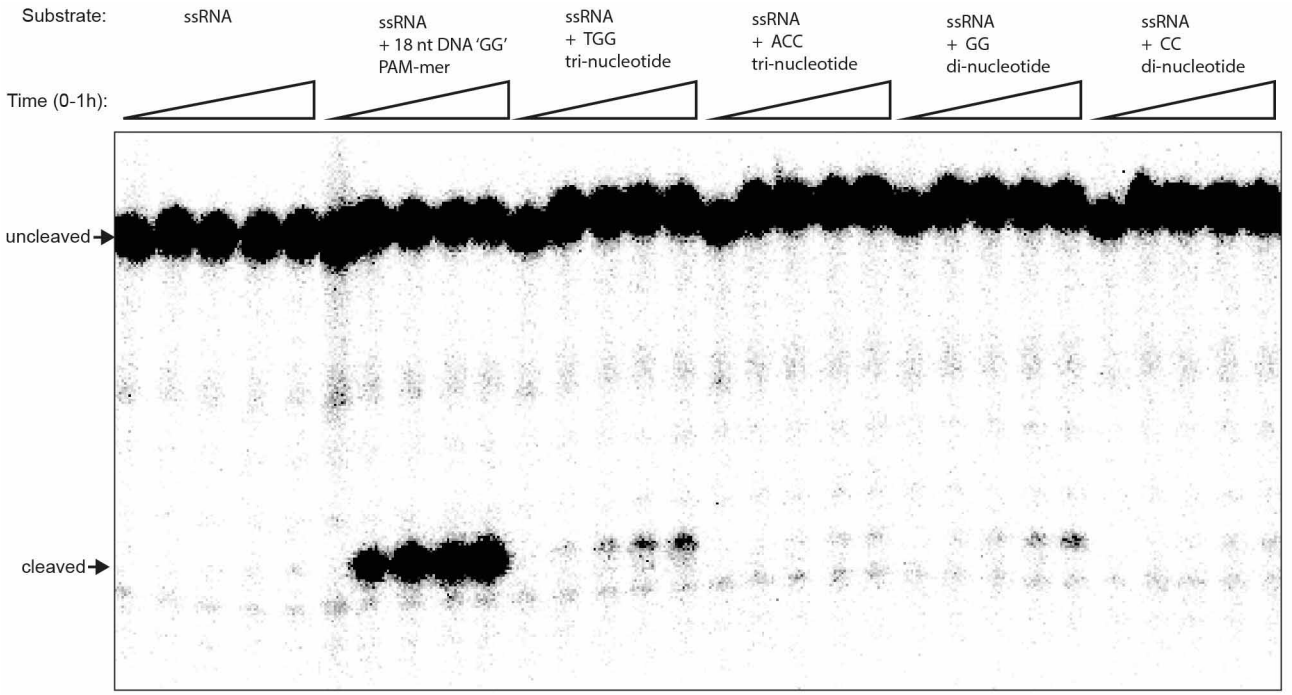
Cas9 biotin labelling. To ensure specific labelling at a single residue on Cas9, two naturally occurring cysteine residues were mutated to serine (C80S and C574S) and a cysteine point mutant was introduced at residue Met 1. To attach the biotin moiety, 10 μM wild-type Cas9 or dCas9 was reacted with a 50-fold molar excess of EZ-Link Maleimide-PEG2-Biotin (Thermo Scientific) at 25 °C for 2 h. The reaction was quenched by the addition of 10 mM DTT, and unreacted Maleimide-PEG2-Biotin was removed using a Bio-Gel P-6 column (Bio-Rad). Labelling was verified using a streptavidin bead binding assay, where 8.5 pmol of biotinylated Cas9 or non-biotinylated Cas9 was mixed with either 25 μl streptavidin-agarose (Pierce Avidin Agarose; Thermo Scientific) or 25 μl streptavidin magnetic beads (Dynabeads MyOne Streptavidin C1; Life Technologies). Samples were incubated in Cas9 reaction buffer at room temperature for 30 min, followed by three washes with Cas9 reaction buffer and elution in boiling SDS–PAGE loading buffer. Elutions were analysed using SDS–PAGE. Cas9 M1C biotinylation was also confirmed using mass spectroscopy performed in the QB3/Chemistry Mass Spectrometry Facility at UC Berkeley. Samples of intact Cas9 proteins were analysed using an Agilent 1200 liquid chromatograph equipped with a Viva C8 (100 mm × 1.0 mm, 5 μm particles, Restek) analytical column and connected in-line with an LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific). Mass spectra were recorded in the positive ion mode. Mass spectral deconvolution was performed using ProMass software (Novatia).

***GAPDH* mRNA pull-down.** HeLa-S3 cell lysates were prepared as previously described[33]. Total RNA was isolated from HeLa-S3 cells using Trizol reagent according to the manufacturer's instructions (Life Technologies). Cas9–sgRNA complexes were reconstituted before pull-down experiments by incubating a twofold molar excess of Cas9 with sgRNA for 10 min at 37 °C in reaction buffer. HeLa total RNA (40 μg) or HeLa lysate (~5×10^6 cells) was added to reaction buffer with 40 U RNasin (Promega), PAMmer (5 μM) and the biotin-dCas9 (50 nM)–sgRNA (25 nM) in a total volume of 100 μl and incubated at 37 °C for 1 h. This mixture was then added to 25 μl magnetic streptavidin beads (Dynabeads MyOne Streptavidin C1; Life Technologies) pre-equilibrated in reaction buffer and agitated at 4 °C for 2 h. Beads were then washed six times with 300 μl wash buffer (20 mM Tris-HCl, pH 7.5, 150 mM NaCl, 5 mM $MgCl_2$, 0.1% Triton X-100, 5% glycerol, 1 mM DTT, 10 μg ml^{-1} heparin). Immobilized RNA was eluted by heating beads at 70 °C in the presence of DEPC-treated water and a phenol/chloroform mixture. Eluates were then treated with an equal volume of glyoxal loading dye (Life Technologies) and heated at 50 °C for 1 h before separation via 1% BPTE agarose gel (30 mM Bis-Tris, 10 mM PIPES, 10 mM EDTA, pH 6.5). Northern blot transfers were carried out as previously described[34]. Following transfer, membranes were crosslinked using UV radiation and incubated in pre-hybridization buffer (UltraHYB Ultrasensitive Hybridization Buffer; Life Technologies) for 1 h at 46 °C before hybridization. Radioactive northern probes were synthesized using random priming of *GAPDH* and β-actin partial cDNAs (for cDNA primers, see Extended Data Table 2) in the presence of [α-^{32}P]dATP (PerkinElmer), using a Prime-It II Random Primer Labelling kit (Agilent Technologies). Hybridization was carried out for 3 h in pre-hybridization buffer at 46 °C followed by two washes with 2 × SSC (300 mM NaCl, 30 mM trisodium citrate, pH 7, 0.5% (w/v) SDS) for 15 min at 46 °C. Membranes were imaged using a phosphorscreen.

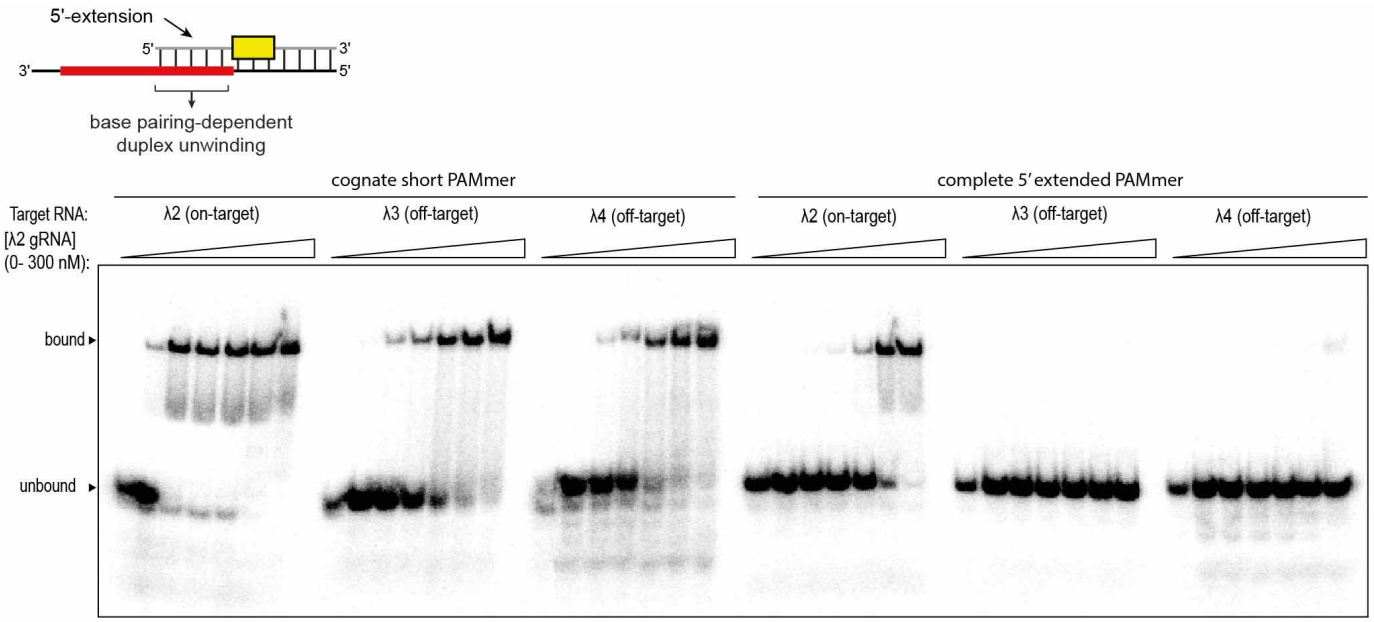
31. Sternberg, S. H., Haurwitz, R. E. & Doudna, J. A. Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA* **18,** 661–672 (2012).
32. Chen, B. *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155,** 1479–1491 (2013).
33. Lee, H. Y. *et al.* RNA-protein analysis using a conditional CRISPR nuclease. *Proc. Natl Acad. Sci. USA* **110,** 5416–5421 (2013).
34. Chomczynski, P. One-hour downward alkaline capillary transfer for blotting of DNA and RNA. *Anal. Biochem.* **201,** 134–139 (1992).



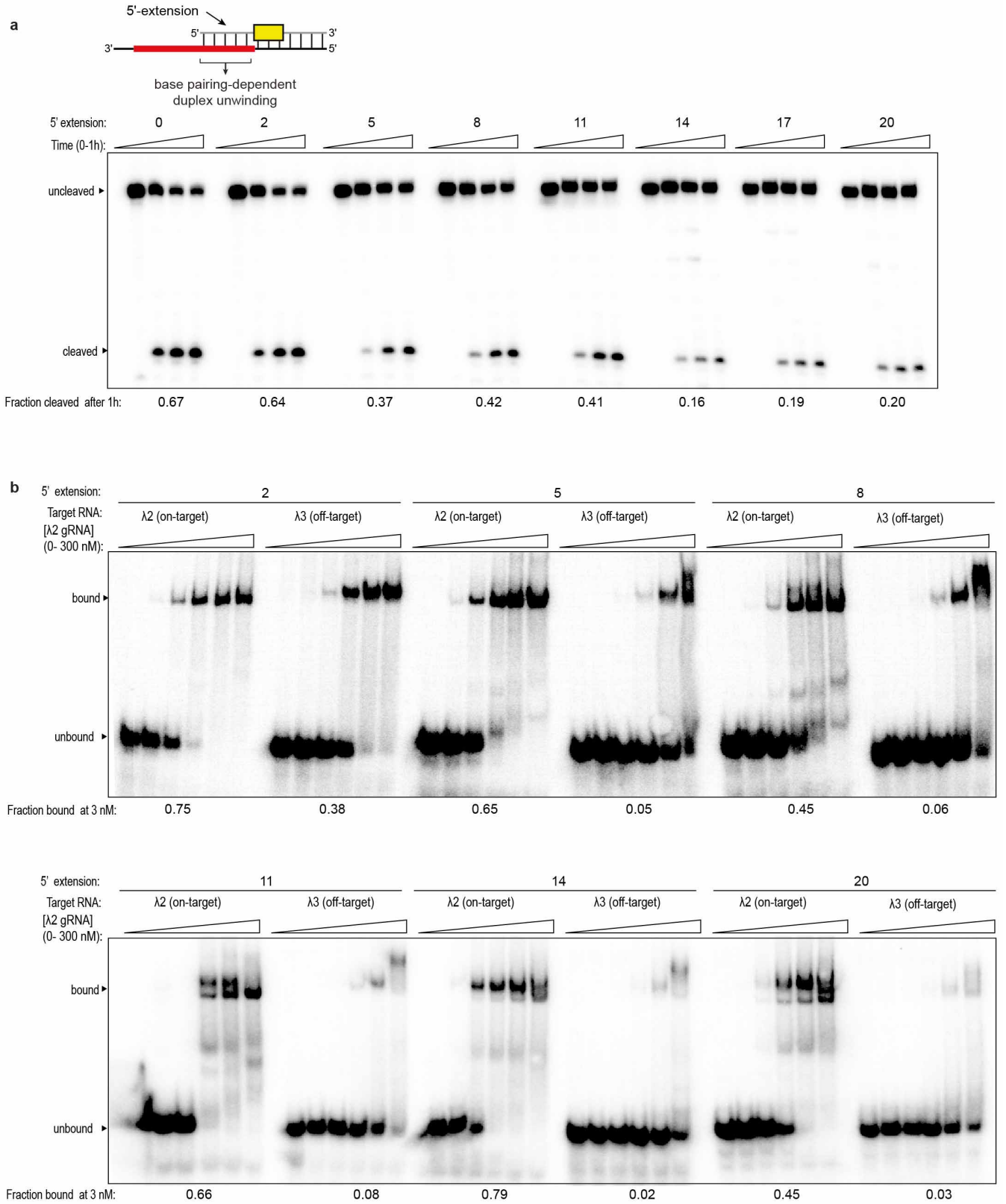
Extended Data Figure 1 | Quantified data for cleavage of ssRNA by Cas9–gRNA in the presence of a 19-nucleotide PAMmer. Cleavage assays were conducted as described in the Methods, and the quantified data were fitted with single-exponential decays. Results from four independent experiments yielded an average apparent pseudo-first-order cleavage rate constant (mean ± s.d.) of $0.032 \pm 0.007\ min^{-1}$. This is slower than the rate constant determined previously for ssDNA in the presence of the same 19-nucleotide PAMmer ($7.3 \pm 3.2\ min^{-1}$)[7].



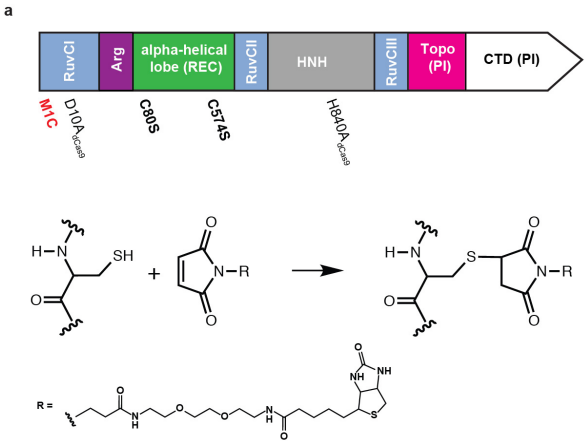
Extended Data Figure 2 | RNA cleavage is marginally stimulated by di- and tri-deoxyribonucleotide PAMmers. Cleavage reactions contained ~1 nM 5′-radiolabelled target ssRNA and no PAMmer (left), 100 nM 18-nt PAMmer (second from left), or 1 mM of the indicated di- or tri-nucleotide (remaining lanes). Reaction products were resolved by 12% denaturing polyacrylamide gel electrophoresis (PAGE) and visualized by phosphorimaging.



Extended Data Figure 3 | Representative binding experiment demonstrating guide-specific ssRNA binding with 5′-extended PAMmers. Gel shift assays were conducted as described in the Methods. Binding reactions contained Cas9 programmed with λ2 gRNA and either λ2 (on-target), λ3 (off-target) or λ4 (off-target) ssRNA in the presence of short cognate PAMmers or cognate PAMmers with complete 5′-extensions, as indicated. The presence of a cognate 5′-extended PAM-mer abrogates off-target binding. Three independent experiments were conducted to produce the data shown in Fig. 3b, d.



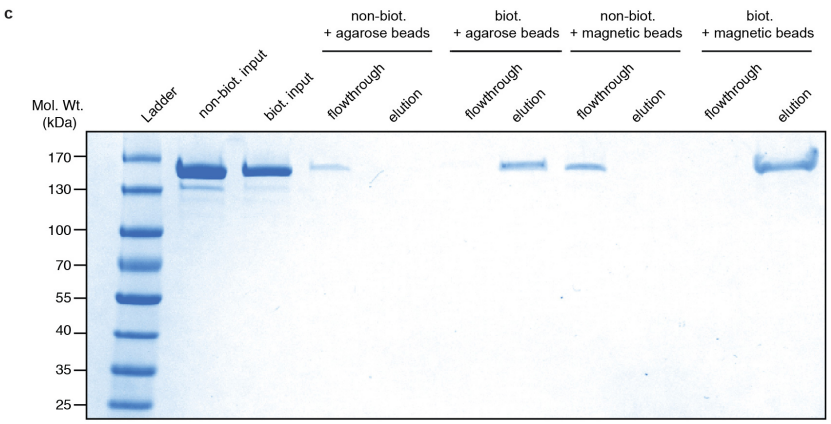
Extended Data Figure 4 | Exploration of RNA cleavage efficiencies and binding specificity using PAMmers with variable 5′-extensions. **a**, Cleavage assays were conducted as described in Methods. Reactions contained Cas9 programmed with λ2 gRNA and λ2 ssRNA targets in the presence of PAMmers with 5′-extensions of variable length. The ssRNA cleavage efficiency decreases as the PAMmer extends further into the target region, as indicated by the fraction of RNA cleaved after 1 h. **b**, Binding assays were conducted as described in the Methods, using mostly the same panel of 5′-extended PAMmers as in **a**. Binding reactions contained Cas9 programmed with λ2 gRNA and either λ2 (on-target) or λ3 (off-target) ssRNA in the presence of cognate PAMmers with 5′-extensions of variable length. The binding specificity increases as the PAMmer extends further into the target region, as indicated by the fraction of λ3 (off-target) ssRNA bound at 3 nM Cas9–gRNA. PAMmers with 5′ extensions also cause a slight reduction in the relative binding affinity of λ2 (on-target) ssRNA.

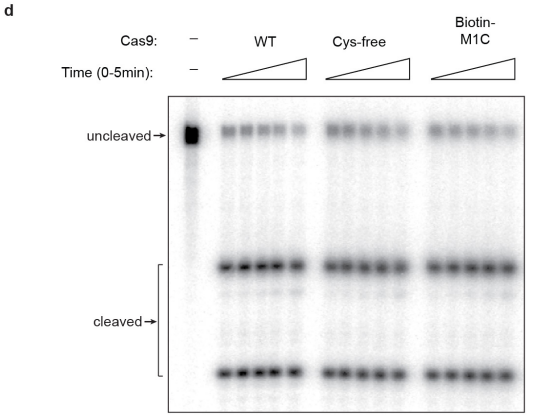


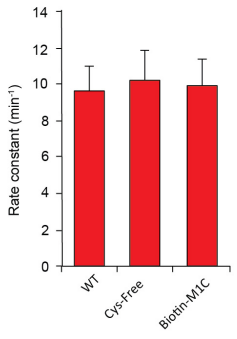
b

Mass Spectroscopy Data

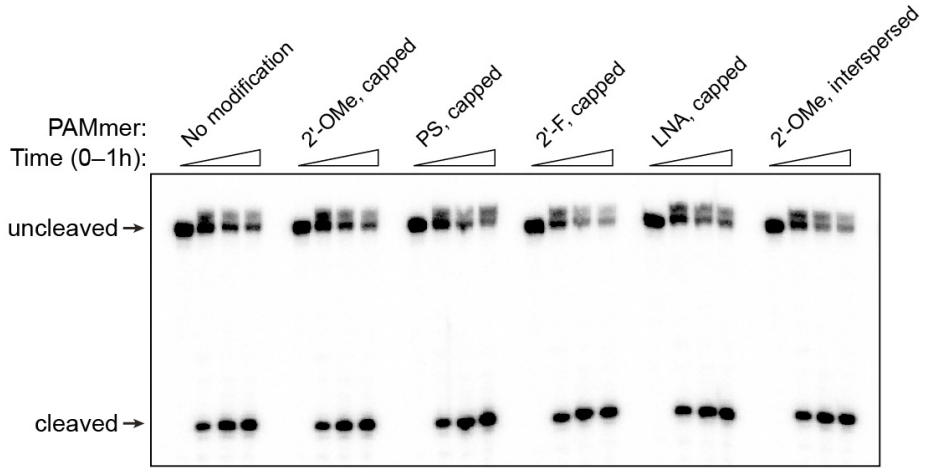
Cas9 Mutant	Predicted Mass (Da)	Unlabeled Mass (Da)	Labeled Mass (Da)	ΔLabel (Da)
Cys-Free	158,681.5	158,681	158,681	0
M1C	158,653.5	158,653	159,179	527



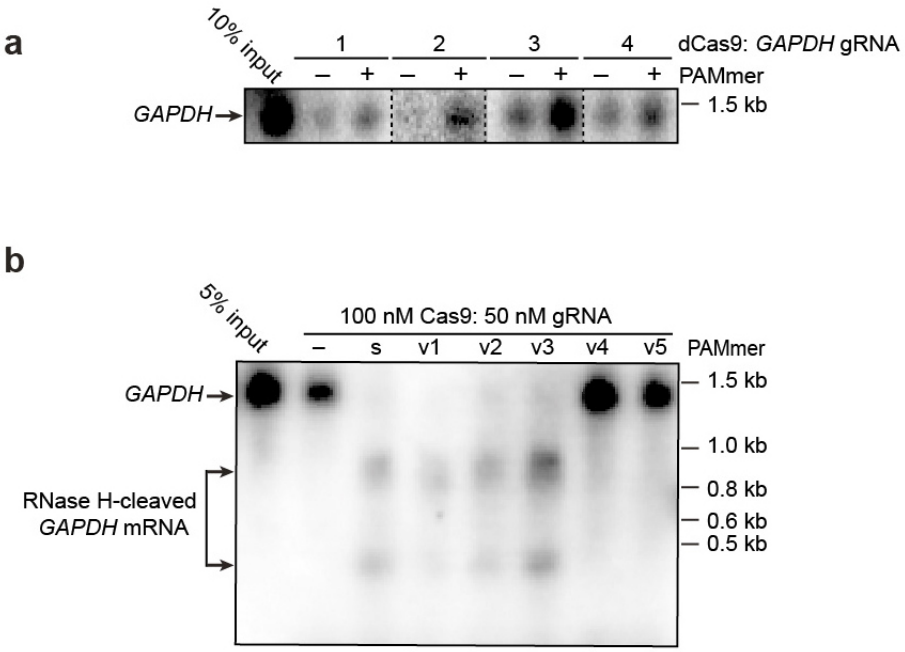




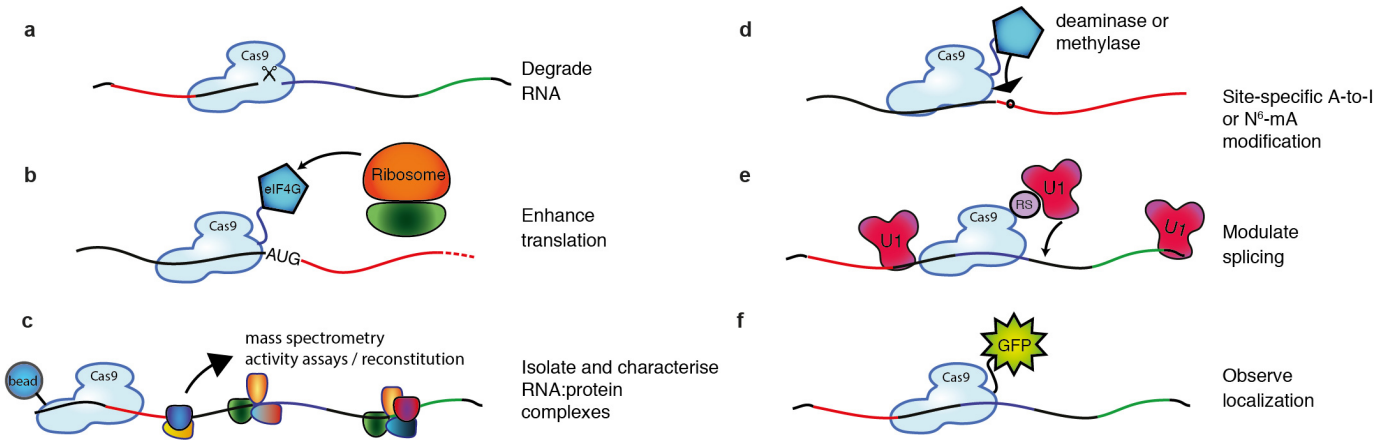
Extended Data Figure 5 | Site-specific biotin labelling of Cas9. **a**, In order to introduce a single biotin moiety on Cas9, the solvent accessible, non-conserved amino-terminal methionine was mutated to a cysteine (M1C; red text) and the naturally occurring cysteine residues were mutated to serine (C80S and C574S; bold text). This enabled cysteine-specific labelling with EZ-link Maleimide-PEG2-biotin through an irreversible reaction between the reduced sulphydryl group of the cysteine and the maleimide group present on the biotin label. Mutations of dCas9 are also indicated in the domain schematic. **b**, Mass spectrometry analysis of the Cas9 biotin-labelling reaction confirmed that successful biotin labelling only occurs when the M1C mutation is present in the Cys-free background (C80S;C574S). The mass of the Maleimide-PEG2-biotin reagent is 525.6 Da. **c**, Streptavidin bead binding assay with biotinylated (biot.) or non-biotinylated (non-biot.) Cas9 and streptavidin agarose or streptavidin magnetic beads. Cas9 only remains specifically bound to the beads after biotin labelling. **d**, Cleavage assays were conducted as described in the Methods and resolved by denaturing PAGE. Reactions contained 100 nM Cas9 programmed with λ2 gRNA and ~1 nM 5′-radiolabelled λ2 dsDNA target. **e**, Quantified cleavage data from triplicate experiments were fitted with single-exponential decays to calculate the apparent pseudo-first-order cleavage rate constants (average ± standard deviation). Both Cys-free and biotin-labelled Cas9(M1C) retain wild-type activity.



Extended Data Figure 6 | RNA-guided Cas9 can utilize chemically modified PAMmers. Nineteen-nucleotide PAMmer derivatives containing various chemical modifications on the 5′ and 3′ ends (capped) or interspersed throughout the strand still activate Cas9 for cleavage of ssRNA targets. These types of modification are often used to increase the *in vivo* half-life of short oligonucleotides by preventing exo- and endonuclease-mediated degradation. Cleavage assays were conducted as described in the Methods. PS, phosphorothioate bonds; LNA, locked nucleic acid.



Extended Data Figure 7 | Cas9 programmed with *GAPDH*-specific gRNAs can pull down *GAPDH* mRNA in the absence of PAMmers. a, Northern blot showing that, in some cases, Cas9–gRNA is able to pull down detectable amounts of *GAPDH* mRNA from total RNA without requiring a PAMmer. **b**, Northern blot showing that Cas9–gRNA *G*1 is also able to pull down quantitative amounts of *GAPDH* mRNA from HeLa cell lysate without requiring a PAMmer. s, standard; v1-5, increasingly 2′-OMe-modified PAMmers. See Fig. 4g for PAMmer sequences.



Extended Data Figure 8 | Potential applications of RCas9 for untagged transcript analysis, detection and manipulation. **a**, Catalytically active RCas9 could be used to target and cleave RNA, particularly those for which RNA-interference-mediated repression/degradation is not possible. **b**, Tethering the eukaryotic initiation factor eIF4G to a catalytically inactive dRCas9 targeted to the 5′ untranslated region of an mRNA could drive translation. **c**, dRCas9 tethered to beads could be used to specifically isolate RNA or native RNA–protein complexes of interest from cells for downstream analysis or assays including identification of bound-protein complexes, probing of RNA structure under native protein-bound conditions, and enrichment of rare transcripts for sequencing analysis. **d**, dRCas9 tethered to RNA deaminase or N6-mA methylase domains could direct site-specific A-to-I editing or methylation of RNA, respectively. **e**, dRCas9 fused to a U1 recruitment domain (arginine- and serine-rich (RS) domain) could be programmed to recognize a splicing enhancer site and thereby promote the inclusion of a targeted exon. **f**, dRCas9 tethered to a fluorescent protein such as GFP could be used to observe RNA localization and transport in living cells. Adapted from Mackay *et al.*[18]

Extended Data Table 1 | λ-Oligonucleotide sequences

Description	Sequence *	Used in
Oligo for T7 promoter, in vitro transcription	5′-**TAATACGACTCACTATA**-3′	NA
λ2-targeting crRNA	5′-GUGAUAAGUGGAAUGCCAUGGUUUUAGAGCUAUGCUGUUUUG-3′	Fig. 1c-e, 3a, 4c-d, ED1-2, 4a
λ3-targeting crRNA	5′-CUGGUGAACUUCCGAUAGUGGUUUUAGAGCUAUGCUGUUUUG-3′	Fig. 3a
λ4-targeting crRNA	5′-CAGATATAGCCTGGTGGTTCGUUUUAGAGCUAUGCUGUUUUG-3′	Fig. 3a
ssDNA T7 template†, tracrRNA	5′-AAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATGCTGT**CCTA TAGTGAGTCGTATTA**-3′	NA
tracrRNA (nt 15-87)	5′-GGACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUU-3′	Fig. 1c-e, 3a, 4c-d, ED1-2,4a
λ2-targeting sgRNA T7 template‡	5′-**TAATACGACTCACTATAGG**TGATAAGTGGAATGCCATGGTTTTAGAGCTATGCTGTTTTGGAAACAAAACAGCATA GCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′	NA
λ2-targeting sgRNA	5′-**GG**UGAUAAGUGGAAUGCCAUGGUUUUAGAGCUAUGCUGUUUUGGAAACAAAACAGCAUAGCAAGUUAAAAUAAGGC UAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUU-3′	Fig 2, 3b,d, ED3, 4b
λ2 target dsDNA duplex	5′-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATGTGGGCTGTCAAAATTGAGC-3′ 3′-CTCACCTTCCTACGGTCACTATTCACCTTACGGTACACCCGACAGTTTTAACTCG-5′	Fig. 1c, 2a, 4c
λ2 ssDNA target strand (used to make heteroduplex DNA:RNA)	3′-CTCACCTTCCTACGGTCACTATTCACCTTACGGTACACCCGACAGTTTTAACTCG-5′	Fig. 1c, 2a
λ2 ssDNA non-target strand (used to make heteroduplex DNA:RNA)	5′-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATGTGGGCTGTCAAAATTGAGC-3′	Fig 1c, 2a, 3d, ED3
λ2 ssRNA target strand T7 template	5′-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATGTGGGCTGTCAAAATTGAGC**CTATAGTGAGTCGTATTA**-3′	NA
λ2 ssRNA target strand	3′-CUCACCUUCCUACGGUCACUAUUCACCUUACGGUACACCCGACAGUUUUAACUCGG-5′	Fig. 1c-e, 2, 3, 4, ED1-4
λ2 ssRNA non-target strand T7 template	5′-GCTCAATTTTGACAGCCCACATGGCATTCCACTTATCACTGGCATCCTTCCACTCC**CTATAGTGAGTCGTATTA**-3′	NA
λ2 ssRNA non-target strand (used to make dsRNA)	5′-GGAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATGTGGGCTGTCAAAATTGAGC-3′	Fig. 1c, 2a
19 nt λ2 DNA PAMmer	5′-TGGGCTGTCAAAATTGAGC-3′	Fig. 1c-e, 2, 3a-b, ED 1-4
18 nt λ2 "GG" DNA PAMmer	5′-GGGCTGTCAAAATTGAGC-3′	Fig. 1c, 2
19 nt λ2 DNA mutated PAMmer	5′-ACCGCTGTCAAAATTGAGC-3′	Fig. 1c, 2c
16 nt λ2 DNA "PAM-less" PAMmer	5′-GCTGTCAAAATTGAGC-3′	Fig. 1c, 2c
18 nt λ2 RNA PAMmer	5′-GGGCUGUCAAAAUUGAGC-3′	Fig. 1c, 2a
5 nt λ2 DNA PAMmer	5′-TGGGC-3′	Fig. 1e, 2c
10 nt λ2 DNA PAMmer	5′-TGGGCTGTCA-3′	Fig. 1e, 2c
15 nt λ2 DNA PAMmer	5′-TGGGCTGTCAAAATT-3′	Fig. 1e, 2c
λ3 ssRNA target strand T7 template	5′-AACGTGCTGCGGCTGGCTGGTGAACTTCCGATAGTGCGGGTGTTGAATGATTTCC**TATAGTGAGTCGTATTA**-3′	NA
λ3 ssRNA target strand	3′-UUGCACGACGCCGACCGACCACUUGAAGGCUAUCACGCCCACAACUUACUAAAGG-5′	Fig. 3a,b,d, ED3, 4b
λ4 ssRNA target strand T7 template	5′-TCACAACAATGAGTGGCAGATATAGCCTGGTGGTTCAGGCGGCGCATTTTTATTGC**CTATAGTGAGTCGTATTA**-3′	NA
λ4 ssRNA target strand	3′-AGUGUUGUUACUCACCGUCUAUAUCGGACCACCAAGUCCGCCGCGUAAAAAUAACGG-5′	Fig. 3a,b,d, ED3
λ3 ssDNA non-target strand	5′-AACGTGCTGCGGCTGGCTGGTGAACTTCCGATAGTGCGGGTGTTGAATGATTTCC-3′	Fig. 3d, ED3
λ4 ssDNA non-target strand	5′-TCACAACAATGAGTGGCAGATATAGCCTGGTGGTTCAGGCGGCGCATTTTTATTG-3′	Fig. 3d, ED3
19 nt λ3 DNA PAMmer	5′-CGGGTGTTGAATGATTTCC-3′	Fig. 3a,b,d, ED3, 4
19 nt λ4 DNA PAMmer	5′-AGGCGGCGCATTTTTATTG-3′	Fig. 3a,b,d, ED3
21 nt λ2 5′-extended DNA PAMmer	5′-TGTGGGCTGTCAAAATTGAGC-3′	Fig. 4c, ED 4a,b
21 nt λ3 5′-extended DNA PAMmer	5′-TGCGGGTGTTGAATGATTTCC-3′	ED 4b
24 nt λ2 5′-extended DNA PAMmer	5′-CCATGTGGGCTGTCAAAATTGAGC-3′	ED 4a,b
24 nt λ3 5′-extended DNA PAMmer	5′-TAGTGCGGGTGTTGAATGATTTCC-3′	ED 4b
27 nt λ2 5′-extended DNA PAMmer	5′-ATGCCATGTGGGCTGTCAAAATTGAGC-3′	Fig. 4f,g, ED 4a,b
27 nt λ3 5′-extended DNA PAMmer	5′-CGATAGTGCGGGTGTTGAATGATTTCC-3′	ED 4b
30 nt λ2 5′-extended DNA PAMmer	5′-GGAATGCCATGTGGGCTGTCAAAATTGAGC-3′	ED 4a,b
30 nt λ3 5′-extended DNA PAMmer	5′-TTCCGATAGTGCGGGTGTTGAATGATTTCC-3′	ED 4b
33 nt λ2 5′-extended DNA PAMmer	5′-AGTGGAATGCCATGTGGGCTGTCAAAATTGAGC-3′	ED 4a,4b
33 nt λ3 5′-extended DNA PAMmer	5′-AACTTCCGATAGTGCGGGTGTTGAATGATTTCC-3′	ED 4b
36 nt λ2 5′-extended DNA PAMmer	5′-ATAAGTGGAATGCCATGTGGGCTGTCAAAATTGAGC-3′	ED 4a
39 nt λ2 5′-extended DNA PAMmer	5′-GTGATAAGTGGAATGCCATGTGGGCTGTCAAAATTGAGC-3′	ED 4a,4b
39 nt λ3 5′-extended DNA PAMmer	5′-CTGGTGAACTTCCGATAGTGCGGGTGTTGAATGATTTCC-3′	Fig. 4b
non-PAM λ2 dsDNA	5′-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATG**ACG**GCTGTCAAAATTGAGC-3′ 3′-CTCACCTTCCTACGGTCACTATTCACCTTACGGTACTGGCGACAGTTTTAACTCG-5′	Fig. 4c
non-PAM λ2 ssRNA target strand T7 template	5′-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATGACCGCTGTCAAAATTGAG**CCTATAGTGAGTCGTATTA**-3′	NA
non-PAM λ2 ssRNA target strand	3′-CUCACCUUCCUACGGUCACUAUUCACCUUACGGUACTGGCGACAGUUUUAACUCGG-5′	Fig. 4c
λ2 2′OMe capped PAMmer§	5′-*UGGGCUGUCAAAAUUGAG*C-3′	ED 6
λ2 PS capped PAMmer§	5′-T*GGGCTGTCAAAATTGAG*C-3′	ED 6
λ2 2′F capped PAMmer§	5′-*UGGGCUGUCAAAAUUGAG*C-3′	ED 6
λ2 LNA capped PAMmer§	5′-*TGGGCTGTCAAAATTGAG*C-3′	ED 6
λ2 19 nt 2′OMe interspersed PAMmer§	5′-*UGGG*CUGUC*AAAAUU*GAG*C-3′	ED 6

* Guide crRNA sequences and complementary DNA target strand sequences are shown in red. PAM sites (5′-NGG-3′) are highlighted in yellow on the non-target strand when adjacent to the target sequence or in the PAMmer.
† The T7 promoter is indicated in bold (or reverse complement of), as well as 5′ G or GG included in the ssRNA product by T7 polymerase.
‡ sgRNA template obtained from pIDT, subsequently linearized by AflII for run-off transcription.
§ Positions of modifications depicted with asterisks preceding each modified nucleotide in each case (except for PS linkages which are depicted between bases).
PS, phosphorothioate bond; NA, not applicable; LNA, locked nucleic acid.

Extended Data Table 2 | Oligonucleotides used in the *GAPDH* mRNA pull-down experiment

Description	Sequence *	Used in
GAPDH-targeting sgRNA 1 T7 template†	5′-**TAATACGACTCACTATA**GGGGCAGAGATGATGACCCTGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAA GGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′	NA
GAPDH-targeting sgRNA 1	5′-GGGGCAGAGATGATGACCCTGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCAAC TTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′	Fig. 4f,g, ED 7
GAPDH-targeting sgRNA 2 T7 template†	5′-**TAATACGACTCACTATAGG**CCAAAGTTGTCATGGATGACGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAAT AAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′	NA
GAPDH-targeting sgRNA 2	5′-**GG**CCAAAGTTGTCATGGATGACGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCA ACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′	Fig. 4f, ED 7
GAPDH-targeting sgRNA 3 T7 template†	5′-**TAATACGACTCACTATAGG**CCAAAGTTGTCATGGATGACGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAAT AAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′	NA
GAPDH-targeting sgRNA 3	5′-**GG**CCAAAGTTGTCATGGATGACGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCA ACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′	Fig. 4f, ED 7
GAPDH-targeting sgRNA 4 T7 template†	5′-**GG**ATGTCATCATATTTGGCAGGGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCA ACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′	NA
GAPDH-targeting sgRNA 4	5′-**TAATACGACTCACTATAGG**ATGTCATCATATTTGGCAGGGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAAT AAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′	Fig. 4f, ED 7
GAPDH PAMmer 1	5′-ATGACCCT**TGG**GGCTCCCCCCTGCAAA-3′	Fig. 4f,g, ED 7
GAPDH PAMmer 2	5′-TGGATGAC**CGG**GGCCAGGGGTGCTAAG-3′	Fig. 4f, ED 7
GAPDH PAMmer 3	5′-TTGGCAGG**TGG**TTCTAGACGGCAGGTC-3′	Fig. 4f, ED 7
GAPDH PAMmer 4	5′-CCCCAGCG**TGG**AAGGTGGAGGAGTGGG-3′	Fig. 4f, ED 7
GAPDH PAMmer 1 2′OMe v1‡	5′-A*UGACC*CTAGG*GGCTC*CCCCC*UGCAA*A-3′	Fig. 4g, ED 7
GAPDH PAMmer 1 2′OMe v2‡	5′-*ATG*ACC*CU*AGG*GGC*UCC*CCC*CTG*CAA*A-3′	Fig. 4g, ED 7
GAPDH PAMmer 1 2′OMe v3‡	5′-*ATG*ACCC*UAGG*GGCT*CCCC*CCTG*CAA*A-3′	Fig. 4g, ED 7
GAPDH PAMmer 1 2′OMe v4‡	5′-*AT*GA*CC*CT*AGG*GG*CT*CC*CC*CC*UG*CA*AA-3′	Fig. 4g, ED 7
GAPDH PAMmer 1 2′OMe v5‡	5′-*AT*GA*CC*CT*AG*GG*GC*TC*CC*CC*CU*GC*AA*A-3′	Fig. 4g, ED 7
GAPDH cDNA primer Fwd	5′-CTCACTGTTCTCTCCCTCCGC-3′	Fig. 4g,f, ED7
GAPDH cDNA primer Rev	5′-AGGGGTCTACATGGCAACTG-3′	Fig. 4g,f, ED7
β-actin cDNA primer Fwd	5′-AGAAAATCTGGCACCACACC-3′	Fig. 4g,f, ED7
β-actin cDNA primer Rev	5′-GGAGTACTTGCGCTCAGGAG-3′	Fig. 4g,f, ED7

* Guide crRNA sequences and complementary DNA target strand sequences are shown in red. PAM sites (5′-NGG-3′) are highlighted in yellow on the non-target strand when adjacent to the target sequence or in the PAMmer.
† The T7 promoter is indicated in bold (or reverse complement of), as well as 5′ G or GG included in the ssRNA product by T7 polymerase. sgRNAs for *GAPDH* were designed according to Chen *et al.*[32]
‡ Positions of 2′-OMe modifications depicted with asterisks preceding each modified nucleotide.
NA, not applicable.

Cancer Letters 365 (2015) 141–148

Contents lists available at ScienceDirect

Cancer Letters

journal homepage: www.elsevier.com/locate/canlet

Mini-review

Circular RNA: A new star of noncoding RNAs

Shibin Qu [1], Xisheng Yang [1], Xiaolei Li [1], Jianlin Wang, Yuan Gao, Runze Shang, Wei Sun, Kefeng Dou, Haimin Li *

Department of Hepatobiliary Surgery, Xijing Hospital, The Fourth Military Medical University, Xi'an, 710032, China

ARTICLE INFO

Article history:
Received 30 April 2015
Received in revised form 1 June 2015
Accepted 1 June 2015

Keywords:
Circular RNA
Alternative circularization
MicroRNA sponge
Gene expression regulation
Biomarker

ABSTRACT

Circular RNAs (circRNAs) are a novel type of RNA that, unlike linear RNAs, form a covalently closed continuous loop and are highly represented in the eukaryotic transcriptome. Recent studies have discovered thousands of endogenous circRNAs in mammalian cells. CircRNAs are largely generated from exonic or intronic sequences, and reverse complementary sequences or RNA-binding proteins (RBPs) are necessary for circRNA biogenesis. The majority of circRNAs are conserved across species, are stable and resistant to RNase R, and often exhibit tissue/developmental-stage-specific expression. Recent research has revealed that circRNAs can function as microRNA (miRNA) sponges, regulators of splicing and transcription, and modifiers of parental gene expression. Emerging evidence indicates that circRNAs might play important roles in atherosclerotic vascular disease risk, neurological disorders, prion diseases and cancer; exhibit aberrant expression in colorectal cancer (CRC) and pancreatic ductal adenocarcinoma (PDAC); and serve as diagnostic or predictive biomarkers of some diseases. Similar to miRNAs and long noncoding RNAs (lncRNAs), circRNAs are becoming a new research hotspot in the field of RNA and could be widely involved in the processes of life. Herein, we review the formation and properties of circRNAs, their functions, and their potential significance in disease.

Introduction

Circular RNAs (circRNAs) were recently discovered as a special novel type of endogenous noncoding RNA and represent a recent research hotspot in the field of RNA. Unlike linear RNAs that are terminated with 5′ caps and 3′ tails, circRNAs form covalently closed loop structures with neither 5′–3′ polarities nor polyadenylated tails [1].

CircRNA was first found in RNA viruses as early as the 1970s [2]. Unfortunately, only a handful of such circRNAs were serendipitously discovered over the past 30 years [3–9]. Such molecules were typically considered to be molecular flukes or products of aberrant RNA splicing due to their low levels of expression. However, with the development of RNA deep sequencing technology and bioinformatics, recent work has revealed that large numbers of circRNAs are endogenous, abundant, conserved and stable in mammalian cells [10–16]. Furthermore, several researchers have confirmed that reversed complementary sequences including inverted repeated Alu pairs (IRAlus) and exon skipping are essential to circRNA formation [17–25]. Moreover, RNA-binding proteins (RBPs) also regulate circRNA formation [23,26].

Specifically, subsequent reports revealed that circRNAs could function as microRNA (miRNA) sponges, regulate alternative splicing, and modulate the expression of parental genes [13,14,16,23,27]. More importantly, it is becoming evident that circRNAs may be involved in atherosclerotic vascular disease risk, neurological disorders, prion diseases and cancer [28–30]; are aberrantly expressed in colorectal cancer (CRC) [31] and pancreatic ductal adenocarcinoma (PDAC) (S.B.Q., unpublished observations). CircRNAs were described as potential disease biomarkers in human saliva and as biomarkers for aging and gastric cancer (GC) [32–34]. Taken together, these findings indicate that circRNAs have great potential to perform special regulating roles in biological development and disease initiation and progression, become new clinical diagnostic and prognostic markers, and provide new insights into the treatment of diseases.

In this review, we briefly delineate the diversity of circRNAs and discuss the highlights of the biogenesis of circRNAs, their characteristics, their potential functions and their relationships with the disease.

Diversity of circRNAs

CircRNAs are expressed at low levels and were originally thought to be by-products of spliceosome-mediated splicing errors [35] or intermediates that escaped from intron lariat debranching [36,37]. Thus, circRNAs received little attention and were thought to be

* Corresponding author. Tel.: +86 29 84771098; fax: +86 29 84771098.
E-mail address: lihaim@fmmu.edu.cn (H. Li).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.canlet.2015.06.003
0304-3835/

Table 1
Overview of human circRNAs identified recently.

Sample	Special treatment	Detection method	Number of circRNAs	References
Cell line (HeLa)	Pol II CLIP	RNA-seq	15 EIciRNAs (most abundant)	[16]
39 ENCODE data sets	rRNA depletion	RNA-seq	7112 predicted circRNAs (circRNA fraction ≥10%)	[15]
Cell line (H9)	poly(A) RNA depletion rRNA depletion RNase R	RNA-seq	103 ciRNAs (at least 2-fold enrichment)	[14]
Cell line (Hs68)	rRNA depletion RNase R	RNA-seq	25,166 predicted circRNAs (high-confidence)	[12]
15 Cell lines (including cancer and non-cancer cell lines from public ENCODE RNA-seq data)	poly(A) RNA depletion rRNA depletion	RNA-seq	46,866 predicted circRNAs (at an FDR of 0.025)	[11]
4 Cell lines ($CD19^+$ leukocytes, HEK293, $CD34^+$ leukocytes, neutrophils)	rRNA depletion	RNA-seq	1950 predicted circRNAs (at least two independent reads)	[13]
5 Cell lines ($CD19^+$ leukocytes, HeLa, H9, $CD34^+$ leukocytes, neutrophils)	rRNA depletion	RNA-seq	2748 predicted circRNAs	[10]

Special treatments were conducted after total RNAs were extracted from the samples. Then, circRNAs were identified via RNA-seq.
circRNAs: circular RNAs; EIciRNAs: exon–intron circRNAs; ciRNAs: circular intronic RNAs; rRNA: ribosomal RNA; RNA-seq: RNA-sequencing; Pol II CLIP: RNA polymerase II crosslinking and immunoprecipitation; FDR: false discovery rate; RNase R: ribonuclease R.

unlikely to play critical roles in biological processes. Until 2010, few circRNAs had been discovered, and research into circRNA biogenesis was minimal. However, with the development of high-throughput sequencing technology and computational analysis, thousands of circRNAs across species from Archaea to humans have been discovered [10–16,38]. The expression of some circRNAs is >10-fold higher than those of their canonical linear transcripts of the same genes [12]. The recently identified human circRNAs are depicted in Table 1.

Biogenesis of circRNAs

Recent studies have revealed that the biogenesis of circRNAs via backsplicing is different from the canonical splicing of linear RNAs [18]. Furthermore, several recent advances in our understanding of circRNA biogenesis, particularly regarding its regulation and the competition between backsplicing and canonical splicing, have been made [1]. For example, Jeck et al. put forward two models of circRNA formation [12]. Model 1 is termed 'lariat-driven circularization' or 'exon skipping' (Fig. 1a), and model 2 is termed 'intron-pairing-driven circularization' or 'direct backsplicing' (Fig. 1b). Notably, Kelly and colleagues also found that exon circularization is widespread and correlated with exon skipping in human umbilical vein endothelial cells (HUVECs) treated with tumor necrosis factor α (TNFα) or tumor growth factor β (TGFβ) [22]. Although some evidence has indicated that intron-pairing-driven circularization might occur more frequently than lariat-driven circularization [39], accumulated evidence has verified the model of intron-pairing-driven circularization and suggested that reverse complementary sequences, including IRAlus, are important for circRNA biogenesis [17–21,23–25]. Shortly thereafter, Zhang and others discovered a new type of circRNA in human cells that is derived from introns and was termed circular intronic RNAs (ciRNAs). ciRNA biogenesis depends on a consensus motif containing a 7-nt GU-rich element near the 5′ splice site and an 11-nt C-rich element near the branchpoint site [14] (Fig. 1c). Very recently, Li et al. also found exons that are circularized with introns 'retained' between the exons. These authors termed them exon–intron circRNAs or EIciRNAs and found that they could be overexpressed with their flanking complementary sequences [16]. However, the mechanism of EIciRNA formation remains unknown. These mechanisms add considerably to the regulatory complexity of the human transcriptome.

Additionally, researchers have identified the *muscleblind* protein (MBL), which can bind to circMbl flanking introns to provoke the formation of circRNAs that act as RBPs to bridge two flanking introns close together [23]. Similarly, researchers reported an additional mode of circRNA biogenesis in which interactions between RBPs form a bridge between the flanking introns, which causes the splice donor and splice acceptor to close to promote circRNA biogenesis [40] (Fig. 1d). Surprisingly, Conn and others have recently found that RBP Quaking (QKI) regulates the formation of circRNAs [26]. In contrast, Ivanov and others noted that the RNA-editing enzyme ADAR1 can bind to double-stranded RNA to antagonize circRNA biogenesis by melting the stem structure [20]. Therefore, RBPs may serve as activators or inhibitors of the formation of circRNAs in some conditions.

Remarkably, Zhang et al. first proposed a model of alternative circularization that is similar to alternative splicing [18] (Fig. 2). These authors found that competition in RNA pairing by complementary sequences (either repetitive or nonrepetitive) across or within individual flanking introns could significantly affect splicing selection and exon circularization. Complementary sequences within individual flanking introns can be sufficient to promote liner mRNA generation. Conversely, complementary sequences across flanking introns can benefit exon circularization. The competition between reverse complementary sequences can result in multiple circRNA transcripts being processed from a single gene (Fig. 2). However, alternative circularization can be species-specific due to the different distributions of complementary sequences across species. The existence of complementary sequences is necessary but not sufficient for exon circularization [18]. This model suggests that the mechanism of alternative circularization is very complicated and is also possibly regulated by other factors, such as RBPs [1].

Properties of circRNAs

According to recent research, there are several noteworthy properties of circRNAs that are produced by backsplicing. Firstly, these circRNAs have covalently closed loop structures with neither 5′–3′ polarity nor a polyadenylated tail, which makes them much more stable than liner RNA and insusceptible to degradation by RNA exonuclease or RNase R [41]. For example, researchers identified >400 circRNAs in human cell-free saliva (CFS) from healthy individuals. These data represent experimental validation of circRNAs in any type of extracellular body fluid [33]. Secondly, there is a great diversity of circRNAs [40]. In some cases, the abundance of circular molecules exceeds those of the corresponding linear mRNAs by >10-fold [12]. Thirdly, circRNAs are largely composed of exons, which primarily reside in the cytoplasm and possibly have miRNA response elements (MREs) [11–13]. Moreover, circRNAs harbor significant reductions in polymorphisms at predicted miRNA target sites [42]. Some circRNAs come from introns or exons with introns that are 'retained' between exons and are primarily located in the nucleus in eukaryotes and may regulate gene expression [14,16].

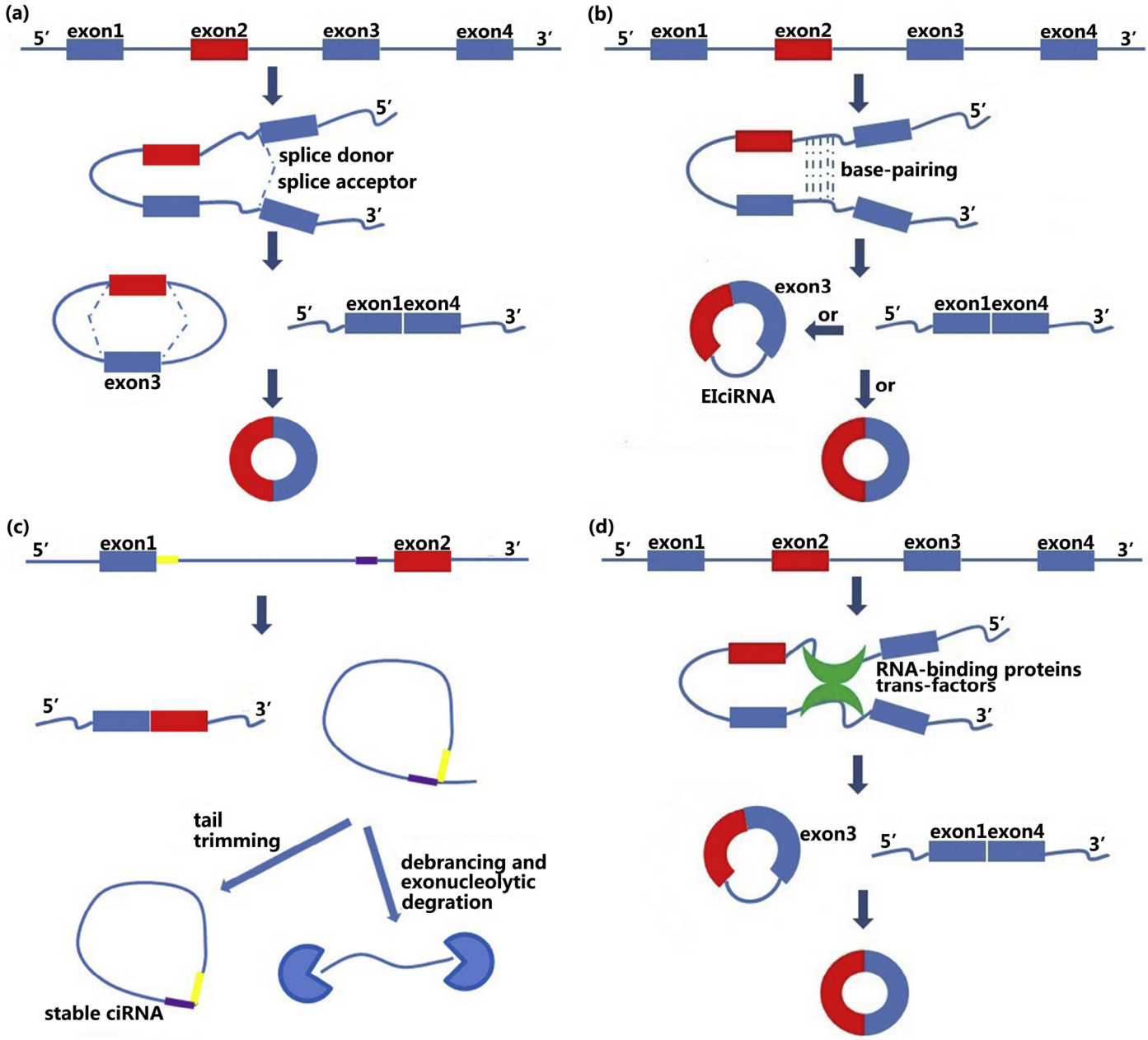


Fig. 1. Models of circRNA biogenesis. (a) Lariat-driven circularization. That splice donor in 3′ end of exon 1 covalently splices to splice acceptor in 5′ end of exon 4 forms a lariat via exon skipping. Then the introns are removed via splicesome. CircRNA finally is formed [12]. (b) Intron-pairing-driven circularization. Intron 1 and intron 3 are formed circular structure via base-pairing. Then introns are removed or retained to form circRNA or EIciRNA [12,16]. (c) Circular intronic RNA. The lariat intron is generated from the splicing reaction. GU-rich sequences near the 5′ splice site (yellow box) and C-rich sequences near branch point (purple box) are minimally sufficient for an intron to escape debranching and degradation. 3′ 'tail' downstream from the branch point is trimmed to result in a stable ciRNA [14]. (d) RBP or trans-factor driven circularization. RBPs or trans-factors (green) can bridge two flanking introns close together. Then the introns are removed to form circRNA [26,40]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fourthly, circRNAs often exhibit tissue/developmental-stage-specific expression [11,13,16]. For example, hsa_circRNA_2149 has been detected in $CD19^{+}$ leukocytes but not $CD34^{+}$ leukocytes, neutrophils or HEK293 cells. Some nematode circRNAs seem to be expressed in oocytes but absent in 1- or 2-cell embryos according to sequencing data [13]. Fifthly, the vast majority of circRNAs are endogenous noncoding RNAs, and only a small portion of exogenous circRNAs, such as Hepatitis δ (HDV) and engineered circRNAs with internal ribosome entry sites (IRESs), are translated [5,43,44]. Finally, circRNAs are evolutionarily conserved between different species [11,12,45]; however, some ciRNAs are much less evolutionarily conserved [14]. Taken together, these properties indicate that circRNAs have the potential to play important roles in transcription and post-transcription and to become ideal biomarkers in the diagnosis of diseases.

CircRNA function

CircRNAs function as competing endogenous RNAs or miRNA sponges

The competitive endogenous RNAs (ceRNAs) contain shared MREs, such as mRNAs, pseudogenes and long noncoding RNAs (lncRNAs), and can compete for miRNA binding [46]. Thus, the presence or absence of ceRNAs influences the activities of miRNAs regarding the

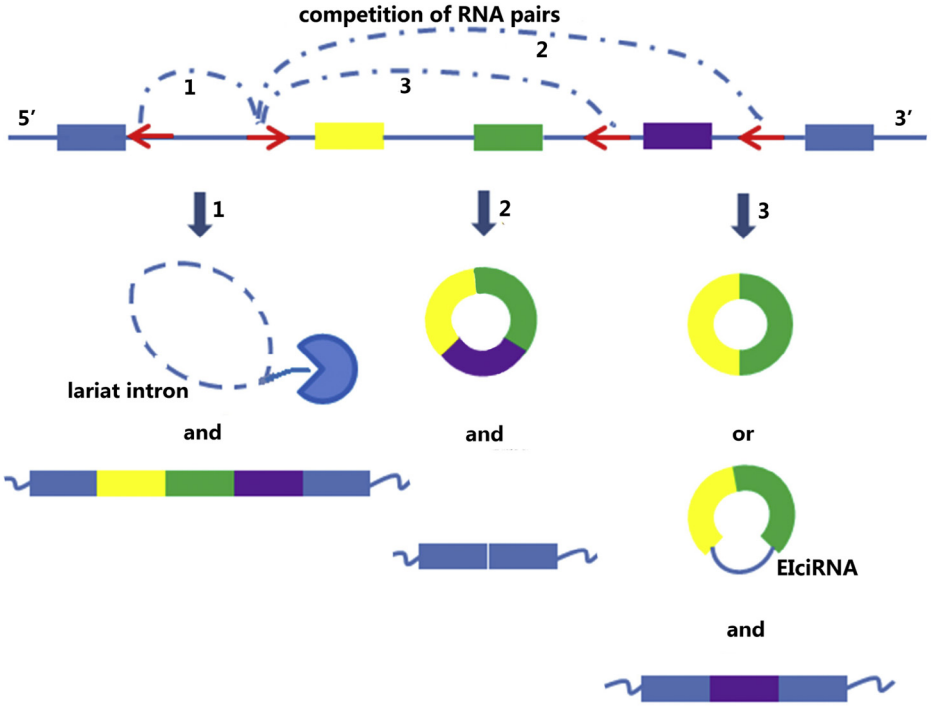


Fig. 2. Possible models of alternative circularization. Multiple circRNAs from either exons or introns, such as ciRNA, circRNA and EIciRNA, can be generated from a single gene locus via the competition of RNA pairing across different introns (blue arcs). (1) Lariat intron without key RNA elements containing a 7-nt GU-rich element near the 5′ splice site and an 11-nt C-rich element near the branchpoint site fail to escape from debranching and exonucleolytic degradation [14]. (2) CircRNA is formed via the competition of RNA pairing across flanking introns [1]. (3) Complementary sequences formed a circular structure via the competition of RNA pairing. CircRNA is formed via removing the intron. If the intron is retained with unknown mechanisms, EIciRNA will be processed [16,18]. Red arrows, complementary sequences. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

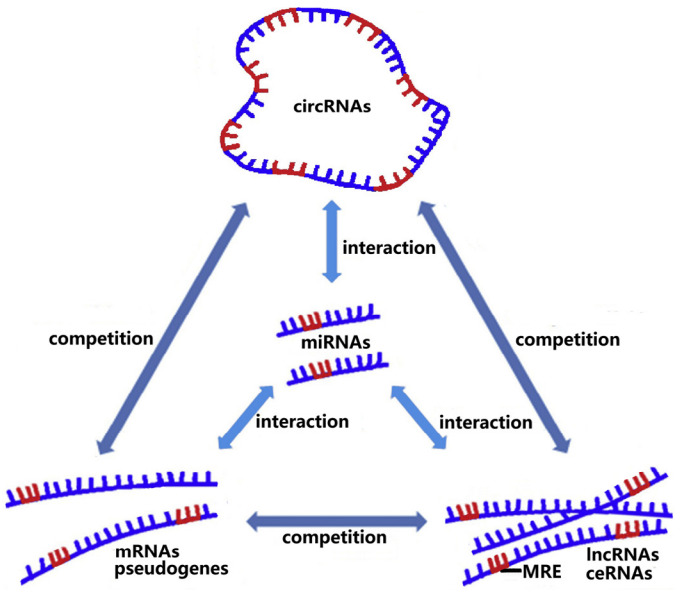


Fig. 3. Network of ceRNAs. circRNAs, mRNAs, pseudogenes and lncRNAs act as ceRNAs to regulate their respective roles, which contain shared MREs to compete for microRNA binding. They maintain dynamic balance to regulate cellular homeostasis. If homeostasis of ceRNAs is dysregulated, the diseases may occur.

regulation of gene expression. Recently accumulated evidence indicates that circRNAs can function as miRNA sponges or potent ceRNA molecules [13,27,30,47] (Fig. 3) and can be depleted of polymorphisms at microRNA binding sites [41]. For example, the exonic circRNAs *ciRS-7/CDR1as* (for circular RNA sponge for miR-7 or *CDR1* antisense) and *Sry* have been shown to bind miRNAs without being degraded, which makes them excellent candidates for ceRNA activity [39]. Hansen et al. discovered that the cerebellar degeneration-related protein 1 (CDR1) gene can translate a natural circular antisense transcript termed antisense to the cerebellar degeneration-related protein 1 transcript (CDR1as). CDR1as can interact with miRNAs and be cleaved by miR-671 [48]. The miR-671 binding site exhibits near-perfect complementarity and little variation across species [27]. Subsequent research revealed that CDR1as contains over 70 conserved seed matches for miR-7 and is densely bound by Argonaute proteins (i.e., the proteins that bind to miRNAs). Notably, the limits in the complementarity of the seed matches protect CDR1as from degradation from the bound miR-7 [27]. The silencing of CDR1as or the overexpression of miR-671 decreases the expression of published miR-7 target genes [13,27], such as SNCA, EGFR and IRS2 [49–51]. In comparison, CDR1as overexpression prevents the downregulation of miR-7 targets [27]. Moreover, CDR1as is expressed at higher levels in nervous tissue. The overexpression of CDR1as in zebrafish embryos, which lack the *cdr1* locus, substantially reduces midbrain size and mimics the phenotype of miR-7 loss-of-function, which causes morphological defects in the midbrain [13]. Similarly, Murine Sex-determining region Y (*Sry*) is the gene responsible for mammalian sex determination and can produce a testis-specific circular transcript [7]. This single-exon circRNA has 16 binding sites for miR-138 and can be co-precipitated with Argonaute 2 (AGO2) in HEK293 cells that are co-transfected with the circRNA *Sry* expression vector and pJEBB-138. These data indicate that the circular *Sry* RNA likely also acts as a miR-138 sponge [27].

However, some analyses of the large set of exonic circRNAs identified by CircleSeq suggest that very few circRNAs in mammalian

cells contain more than ten binding sites for an individual miRNA. Furthermore, many exonic circRNAs only contain smaller numbers of putative miRNA binding sites [39]. Analogously, Guo et al. also found that few circRNAs exhibit properties expected of miRNA sponges [15]. Fortunately, Li et al. detected that cir-ITCH spans several exons of the E3 ubiquitin (Ub) protein ligase (ITCH) and acts as a sponge of miR-7, miR-17 and miR-214 [30]. Therefore, whether circular miRNA sponges are a general phenomenon and how networks of circRNAs, miRNAs and ceRNAs maintain balance to regulate cellular homeostasis remain to be clarified.

CircRNAs regulate alternative splicing or transcription

Previous studies have suggested that circRNAs are involved in the regulation of alternative splicing or transcription. For example, Ashwal-Fluss et al. discovered that circMbl is generated by the second exon of the splicing factor MBL, which competes with canonical pre-mRNA splicing. circMbl flanking introns and circMbl itself have conserved MBL binding sites that are strongly and specifically bound by MBL. The modulation of MBL levels significantly affects circMbl formation, and this effect depends on MBL binding sites in the flanking intronic sequences [23]. These findings suggest that general splicing factors, such as MBL, may have effects on alternative splicing that modulate the balance between circRNA biogenesis and canonical splicing. Moreover, Chao et al. noticed that the mouse formin (*Fmn*) gene can produce circRNA via backsplicing. Notably, this circRNA that contains the translation start site functions as an 'mRNA trap' and leaves a noncoding linear transcript and thereby reduces the expression level of the *Fmn* protein [52]. Moreover, Jeck and Sharpless uncovered that many of single-exon circRNAs contain a translation start site in human fibroblasts [39]. These discoveries indicate that circRNAs could act as mRNA traps by sequestering the translation start site to regulate protein expression.

CircRNAs regulate the expression of parental gene

Recent advances have revealed that circRNAs could regulate the expression of parental genes (Fig. 4). For instance, Zhang and colleagues discovered that the formation of ciRNAs depends on the key flanking RNA elements that might be essential for the intron lariat to escape from debranching. These ciRNAs have little enrichment for microRNA target sites, indicating that they are functionally distinct [53]. Detailed studies have demonstrated that some ciRNAs are abundant in the nucleus and interact with the polymerase II (Pol II) machinery and modulate host transcription activity in a cis-acting manner [14]. Subsequently, researchers also reported a special class of circRNAs termed ElciRNAs that are associated with RNA Pol II in human cells. ElciRNAs, such as circEIF3J and circPAIP2, are predominantly localized to the nucleus, interact with U1 small nuclear ribonucleoproteins (snRNPs) and enhance the transcription of their parental genes in a cis-acting manner [16]. Similarly, Li and others found that both cir-ITCH and the 3′-untranslated region (UTR) of ITCH share some miRNA binding sites. Further study indicated that the interactions of cir-ITCH with miR-7, miR-17, and miR-214 might increase the level of ITCH [30]. We speculate that exon-only circRNAs may fulfill regulatory functions in the cytoplasm, whereas intronic circRNAs, such as ciRNAs and ElciRNAs, seem to be efficient for transcriptional regulation in the nucleus.

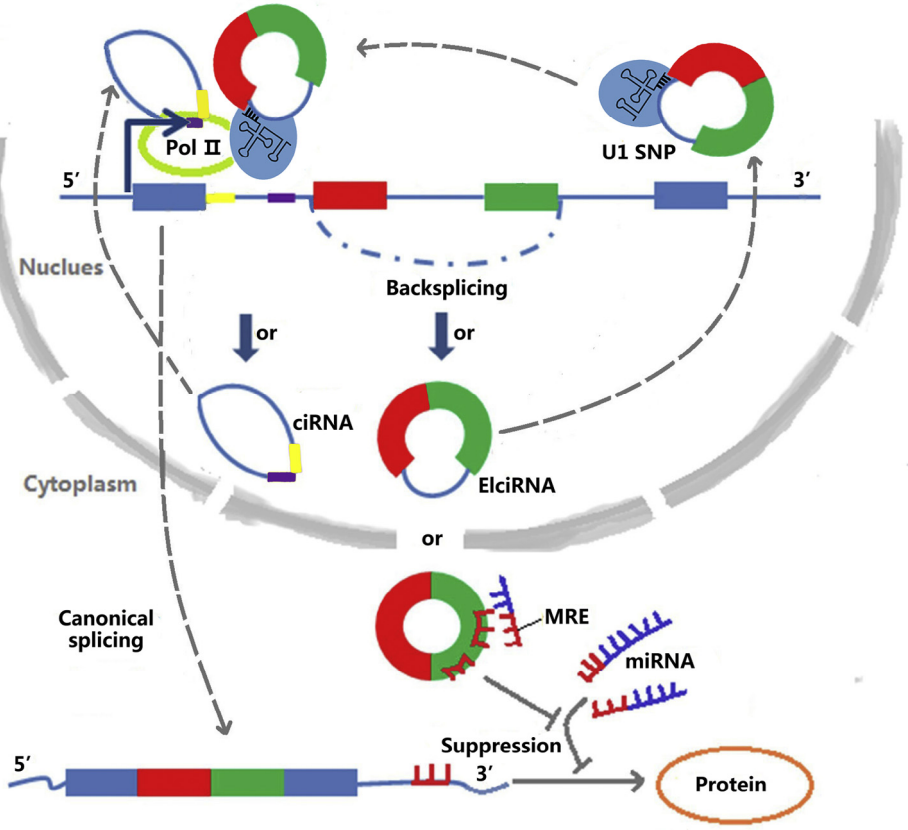


Fig. 4. Three models of circRNA regulating the expression of parental gene. ciRNAs are produced from lariat introns that escape debranching. The stable ciRNA binds to elongating RNA Pol II and promotes transcription [14]. ElciRNA binds to U1 snRNP through specific RNA–RNA interaction between U1 snRNA and ElciRNA, and then the ElciRNA–U1 snRNP complexes might interact with RNA Pol II transcription complex to promote host gene expression [16]. CircRNA shares some miRNA binding sites with 3′-UTR of the transcript from their parental gene. CircRNA acts as miRNA sponge and increases the translations of the transcript from its parental gene [30].

Other possible functions of circRNAs

Few circRNAs can be translated. Researchers reported that engineered circRNAs that were inserted an IRES in upstream of the start codons of a protein could be translated in vitro [42] or in vivo [54]. Similarly, Perriman and Ares reported that an engineered circular mRNA containing a simple green fluorescent protein (GFP) open reading frame can direct GFP expression in *Escherichia coli* [44]. Interestingly, thus far, only one naturally occurring circRNA is known to encode a single protein in eukaryotic cells, i.e., HDV, which is a subviral satellite virus of the hepatitis B virus (HBV) [5]. The encapsulation of HDV with HBV virions results in the production of a single viral protein that is specific to pathogenicity, but the principle of translation is noncanonical and probably associated with specific viral agents [55,56]. However, to date, there is no evidence that suggests that naturally occurring endogenous circRNAs undergo translation [10,12]. Additionally, researchers have reported putative additional plausible roles of circRNAs [57,58]. For example, circRNAs could function as RBP sponges, e.g., the strong and direct interaction between MBL protein and circMbl [23] or function in the assembly of RBP factories or their allosteric regulators. CircRNAs could also directly target mRNAs by partial base pairing. Some circRNAs even serve as templates for translation, as indicated by findings that synthetic circRNAs can be efficiently translated. These findings demonstrate that further studies are necessary to clarify the other potential functions of circRNAs.

CircRNAs in disease

Recent works have suggested that circRNAs may play important roles in the initiation and development of disease could potentially become new biomarkers for these processes. For instance, the expression of ciRS-7/CDR1as but not CDR1 is induced by stable overexpression of the prion protein (PrPC) in HEK293 cells [59]. Therefore, PrPC could possibly be involved in the regulation of CDR1as. It would be interesting to unveil the function of CDR1as in prion disease [29].

CircMbl and its flanking intron sequences can be combined with MBL. Alterations in MBL levels strongly affect circMbl biosynthesis. circRNA production competes with canonical *mbl* pre-mRNA splicing [23]. MBL can regulate *mbl* pre-mRNA splicing efficiency between *mbl* mRNA and circMbl. Moreover, circMbl can sponge out the excess MBL protein by binding to it. However, MBL functional deficiency is known to cause a severe degenerative disease called myotonic dystrophy. Hence, we speculate that circMbl could be involved in myotonic dystrophy initiation and progression.

It is clear that miRNAs have been shown to be involved in nearly all aspects of cellular functions [60] and play critical roles in disease initiation and progression [61,62]. Given that circRNAs interact with miRNAs to regulate their target genes, circRNAs could possibly be involved in diseases correlated with miRNAs. For example, it is evident that CDR1as is highly expressed in the brain and has over 60 binding sites for miR-7 [13,27,48,63]. It is important to note that miR-7 is implicated in numerous pathways and diseases, including its function as a direct regulator of α-synuclein and ubiquitin protein ligase A (UBE2A). CDR1as has been implicated in Parkinson's disease, Alzheimer's disease and brain development [13,27,29,64]. Simultaneously, because miR-7 has been characterized as having both oncogenic and tumor-suppressive properties [65–68], the CDR1as/miR-7 axis is likely involved in cancer initiation and progression [33]. Remarkably, Li and others have shown that cir-ITCH expression is typically downregulated in esophageal squamous cell carcinoma (ESCC) compared to paired adjacent tissue. Cir-ITCH may have an antitumor function in ESCC that acts through interactions with miRNAs such as miR-7, miR-17, and miR-214 and an increase in the level of ITCH, which facilitates ubiquitin-mediated Dvl2 degradation and decreases the expression of the oncogene *c-myc*. This process therefore inhibits canonical Wnt signaling [30]. Moreover, researchers have found circRNAs are globally reduced in CRC tissues via analyses of RNA-sequencing data from 12 matched normal colon mucosa and tumor tissues [31]. We have also identified that the circRNA expression signatures of PDAC are dysregulated via microarray platform (S.B.Q., unpublished observations). The microarray profile has been deposited in GEO with accession number GSE69362 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69362). These findings indicate that dysregulated circRNAs may be involved in the progression of CRC and PDAC.

Finally, circRNAs have also been described as a class of aging biomarkers in *Drosophila* [32] and as putative disease biomarkers in human saliva [25,33]. Burd and colleagues discovered that cANRIL (circular antisense non-coding RNA in the *INK4* locus, cANRIL) is an antisense transcript from the *INK4A-ARF* locus. Single nucleotide polymorphisms (SNPs) on chromosome 9p21.3 near the *INK4/ARF* (*CDKN2a/b*) locus within the atherosclerotic vascular disease (ASVD) risk interval may regulate *ANRIL* splicing and cANRIL production. Intriguingly, cANRIL expression correlates with *INK4/ARF* transcription and ASVD risk [28]. Moreover, researchers have also discovered that hsa_circ_002059 is downregulated in gastric cancer and could represent a potential novel biomarker for the diagnosis of GC [34]. These findings suggest that circRNAs may be involved in ESCC, CRC, PDAC and GC initiation and progression. As research into circRNAs proliferates, circRNAs may also be found to play roles in other tumors.

Conclusion

CircRNAs were previously largely thought to arise from errors in RNA splicing. However, with the advancements in high-throughput sequencing technologies and bioinformatics progression, the biogenesis and function of circRNAs that have been hidden in the multifarious ncRNAs have drawn the attention of many scientists. More importantly, the study of circRNAs has gradually become one of the most noticeable areas in the field of RNA biology [69]. A circRNA database has been constructed (http://www.circbase.org/) [70]. This database will facilitate further research on circRNAs. In this review, we have described natural circRNAs as an abundant, stable, diverse and conserved class of RNA molecules. Reverse complementary sequences and RBPs play profound roles in circRNA biogenesis, but very little is known about the degradation and localization of most circRNAs. CircRNAs can act as competing endogenous RNAs to bind to miRNAs or regulate transcription or affect parental gene expression, and it seems that other functions will be revealed. Moreover, some circRNAs may be involved in differentiation or disease, especially in cancer. A database of disease–circRNA association in Circ2Traits has also been constructed (http://gyanxet-beta.com/circdb/) [71]. This database enriches knowledge base of potential association of circular RNAs with cancer in humans. CircRNAs are associated with cancer-related miRNAs and some circRNA–miRNA axes may be involved in cancer-related pathways. Hsa_circ_002059 is first found to be significantly downregulated in GC, and may be a potential novel and stable biomarker for the diagnosis of GC [34]. Cir-ITCH expression is typically downregulated in ESCC, and may have an inhibitory effect on ESCC by suppressing the Wnt/β-catenin pathway [30]. Additionally, circRNAs exhibit aberrant expression in CRC [31] and PDAC via high-throughput screening. Although there are just few studies of circRNAs in cancer, studies of circRNAs in cancer are on the way. The prospect of research and applications about circRNAs in cancer is promising. Therefore, we could potentially construct engineered circRNAs as molecular tools or therapies. Engineered circRNAs could be effective either for sequestering miRNAs and other RNAs

or RBPs or for releasing these stored molecules via cleavage of the circRNA.

Taken together, the functions and related mechanisms of circRNAs may be rather diverse. CircRNAs may affect life processes, serve as diagnostic or predictive biomarkers of disease and also provide new potential therapeutic targets. Nevertheless, compared with coding RNA and miRNA and lncRNA, there are still significant gaps in our current understanding of circRNAs. The circRNA world is full of treasure. CircRNAs have provided new insights into the "dark matter" of the human genome. The latent roles of circRNAs in the diagnosis and treatment of diseases could be massive. Recent advances have primarily focused on the mechanisms of circRNA biogenesis. The biological and molecular mechanisms of circRNAs in the development of diverse diseases are not yet fully understood. With the development of technology and research, additional circRNAs will be identified. Moreover, further studies will reveal the functions of the vast majority of circRNAs in terms of both physiological and pathological processes.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (No. 81772061) and Health Public Welfare Industry Special Scientific Research Projects of China (No. 201202007). Thanks for anonymous reviewers' constructive comments and suggestions.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] L.L. Chen, L. Yang, Regulation of circRNA biogenesis, RNA Biol. 12 (2015) 381–388.
[2] H.L. Sanger, G. Klotz, D. Riesner, H.J. Gross, A.K. Kleinschmidt, Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures, Proc. Natl. Acad. Sci. U.S.A. 73 (1976) 3852–3856.
[3] M.T. Hsu, M. Coca-Prados, Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells, Nature 280 (1979) 339–340.
[4] A.C. Arnberg, G.J. Van Ommen, L.A. Grivell, E.F. Van Bruggen, P. Borst, Some yeast mitochondrial RNAs are circular, Cell 19 (1980) 313–319.
[5] A. Kos, R. Dijkema, A.C. Arnberg, P.H. van der Meide, H. Schellekens, The hepatitis delta (delta) virus possesses a circular RNA, Nature 323 (1986) 558–560.
[6] J.M. Nigro, K.R. Cho, E.R. Fearon, S.E. Kern, J.M. Ruppert, J.D. Oliner, et al., Scrambled exons, Cell 64 (1991) 607–613.
[7] B. Capel, A. Swain, S. Nicolis, A. Hacker, M. Walter, P. Koopman, et al., Circular transcripts of the testis-determining gene Sry in adult mouse testis, Cell 73 (1993) 1019–1030.
[8] P.G. Zaphiropoulos, Circular RNAs from transcripts of the rat cytochrome P450 2C24 gene: correlation with exon skipping, Proc. Natl. Acad. Sci. U.S.A. 93 (1996) 6536–6541.
[9] Z. Pasman, M.D. Been, M.A. Garcia-Blanco, Exon circularization in mammalian nuclear extracts, RNA 2 (1996) 603–610.
[10] J. Salzman, C. Gawad, P.L. Wang, N. Lacayo, P.O. Brown, Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types, PLoS ONE 7 (2012) e30733.
[11] J. Salzman, R.E. Chen, M.N. Olsen, P.L. Wang, P.O. Brown, Cell-type specific features of circular RNA expression, PLoS Genet. 9 (2013) e1003777.
[12] W.R. Jeck, J.A. Sorrentino, K. Wang, M.K. Slevin, C.E. Burd, J. Liu, et al., Circular RNAs are abundant, conserved, and associated with ALU repeats, RNA 19 (2013) 141–157.
[13] S. Memczak, M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, et al., Circular RNAs are a large class of animal RNAs with regulatory potency, Nature 495 (2013) 333–338.
[14] Y. Zhang, X.O. Zhang, T. Chen, J.F. Xiang, Q.F. Yin, Y.H. Xing, et al., Circular intronic long noncoding RNAs, Mol. Cell 51 (2013) 792–806.
[15] J.U. Guo, V. Agarwal, H. Guo, D.P. Bartel, Expanded identification and characterization of mammalian circular RNAs, Genome Biol. 15 (2014) 409.
[16] Z. Li, C. Huang, C. Bao, L. Chen, M. Lin, X. Wang, et al., Exon-intron circular RNAs regulate transcription in the nucleus, Nat. Struct. Mol. Biol. 22 (2015) 256–264.
[17] R.A. Dubin, M.A. Kazmi, H. Ostrer, Inverted repeats are necessary for circularization of the mouse testis Sry transcript, Gene 167 (1995) 245–248.
[18] X.O. Zhang, H.B. Wang, Y. Zhang, X. Lu, L.L. Chen, L. Yang, Complementary sequence-mediated exon circularization, Cell 159 (2014) 134–147.
[19] D. Liang, J.E. Wilusz, Short intronic repeat sequences facilitate circular RNA production, Genes Dev. 28 (2014) 2233–2247.
[20] A. Ivanov, S. Memczak, E. Wyler, F. Torti, H.T. Porath, M.R. Orejuela, et al., Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals, Cell Rep. 10 (2015) 170–177.
[21] S. Starke, I. Jost, O. Rossbach, T. Schneider, S. Schreiner, L.H. Hung, et al., Exon circularization requires canonical splice signals, Cell Rep. 10 (2015) 103–111.
[22] S. Kelly, C. Greenman, P.R. Cook, A. Papantonis, Exon skipping is correlated with exon circularization, J. Mol. Biol. 15 (2015) 112–116.
[23] R. Ashwal-Fluss, M. Meyer, N.R. Pamudurti, A. Ivanov, O. Bartok, M. Hanan, et al., circRNA biogenesis competes with pre-mRNA splicing, Mol. Cell 56 (2014) 55–66.
[24] Q. Vicens, E. Westhof, Biogenesis of circular RNAs, Cell 159 (2014) 13–14.
[25] S. Petkovic, S. Müller, RNA circularization strategies in vivo and in vitro, Nucleic Acids Res. 43 (2015) 2454–2465.
[26] S.J. Conn, K.A. Pillman, J. Toubia, V.M. Conn, M. Salmanidis, C.A. Phillips, et al., The RNA binding protein quaking regulates formation of circRNAs, Cell 160 (2015) 1125–1134.
[27] T.B. Hansen, T.I. Jensen, B.H. Clausen, J.B. Bramsen, B. Finsen, C.K. Damgaard, et al., Natural RNA circles function as efficient microRNA sponges, Nature 495 (2013) 384–388.
[28] C.E. Burd, W.R. Jeck, Y. Liu, H.K. Sanoff, Z. Wang, N.E. Sharpless, Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk, PLoS Genet. 6 (2010) e1001233.
[29] T.B. Hansen, J. Kjems, C.K. Damgaard, Circular RNA and miR-7 in cancer, Cancer Res. 73 (2013) 5609–5612.
[30] F. Li, L. Zhang, W. Li, J. Deng, J. Zheng, M. An, et al., Circular RNA ITCH has inhibitory effect on ESCC by suppressing the Wnt/β-catenin pathway, Oncotarget 6 (2015) 6001–6013.
[31] A. Bachmayr-Heyda, A.T. Reiner, K. Auer, N. Sukhbaatar, S. Aust, T. Bachleitner-Hofmann, et al., Correlation of circular RNA abundance with proliferation-exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human tissues, Sci. Rep. 5 (2015) 8057.
[32] J.O. Westholm, P. Miura, S. Olson, S. Shenker, B. Joseph, P. Sanfilippo, et al., Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation, Cell Rep. 9 (2014) 1966–1980.
[33] J.H. Bahn, Q. Zhang, F. Li, T.M. Chan, X. Lin, Y. Kim, et al., The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human saliva, Clin. Chem. 61 (2015) 221–230.
[34] P. Li, S. Chen, H. Chen, X. Mo, T. Li, Y. Shao, et al., Using circular RNA as a novel type of biomarker in the screening of gastric cancer, Clin. Chim. Acta 444 (2015) 132–136.
[35] C. Cocquerelle, B. Mascrez, D. Hétuin, B. Bailleul, Mis-splicing yields circular RNA molecules, FASEB J. 7 (1993) 155–160.
[36] C.C. Kopczynski, M.A. Muskavitch, Introns excised from the Delta primary transcript are localized near sites of Delta transcription, J. Cell Biol. 119 (1992) 503–512.
[37] L. Qian, M.N. Vu, M. Carter, M.F. Wilkinson, A spliced intron accumulates as a lariat in the nucleus of T cells, Nucleic Acids Res. 20 (1992) 5345–5350.
[38] M. Danan, S. Schwartz, S. Edelheit, R. Sorek, Transcriptome-wide discovery of circular RNAs in Archaea, Nucleic Acids Res. 40 (2012) 3131–3142.
[39] W.R. Jeck, N.E. Sharpless, Detecting and characterizing circular RNAs, Nat. Biotechnol. 32 (2014) 453–461.
[40] E. Lasda, R. Parker, Circular RNAs: diversity of form and function, RNA 20 (2014) 1829–1842.
[41] H. Suzuki, T. Tsukahara, A view of pre-mRNA splicing from RNase R resistant RNAs, Int. J. Mol. Sci. 15 (2014) 9331–9342.
[42] L.F. Thomas, P. Sætrom, Circular RNAs are depleted of polymorphisms at microRNA binding sites, Bioinformatics 30 (2014) 2243–2246.
[43] C.Y. Chen, P. Sarnow, Initiation of protein synthesis by the eukaryotic translational apparatus on circular RNAs, Science 268 (1995) 415–417.
[44] R. Perriman, M. Ares Jr., Circular mRNA can direct translation of extremely long repeating-sequence proteins in vivo, RNA 4 (1998) 1047–1054.
[45] P.L. Wang, Y. Bao, M.C. Yee, S.P. Barrett, G.J. Hogan, M.N. Olsen, et al., Circular RNA is expressed across the eukaryotic tree of life, PLoS ONE 9 (2014) e90859.
[46] X. Shi, M. Sun, H. Liu, Y. Yao, Y. Song, Long non-coding RNAs: a new frontier in the study of human diseases, Cancer Lett. 339 (2013) 159–166.
[47] R. Taulli, C. Loretelli, P.P. Pandolfi, From pseudo-ceRNAs to circ-ceRNAs: a tale of cross-talk and competition, Nat. Struct. Mol. Biol. 20 (2013) 541–543.
[48] T.B. Hansen, E.D. Wiklund, J.B. Bramsen, S.B. Villadsen, A.L. Statham, et al., miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA, EMBO J. 30 (2011) 4414–4422.
[49] E. Junn, K.W. Lee, B.S. Jeong, T.W. Chan, J.Y. Im, M.M. Mouradian, Repression of alpha-synuclein expression and toxicity by microRNA-7, Proc. Natl. Acad. Sci. U.S.A. 106 (2009) 13052–13057.
[50] B. Kefas, J. Godlewski, L. Comeau, Y. Li, R. Abounader, M. Hawkinson, et al., microRNA-7 inhibits the epidermal growth factor receptor and the Akt pathway and is down-regulated in glioblastoma, Cancer Res. 68 (2008) 3566–3572.
[51] L. Jiang, X. Liu, Z. Chen, Y. Jin, C.E. Heidbreder, A. Kolokythas, et al., MicroRNA-7 targets IGF1R (insulin-like growth factor 1 receptor) in tongue squamous cell carcinoma cells, Biochem. J. 432 (2010) 199–205.
[52] C.W. Chao, D.C. Chan, A. Kuo, P. Leder, The mouse formin (*Fmn*) gene: abundant circular RNA transcripts and gene-targeted deletion analysis, Mol. Med. 4 (1998) 614–628.

[53] Y. Zhang, L. Yang, L.L. Chen, Life without A tail: new formats of long noncoding RNAs, Int. J. Biochem. Cell Biol. 54 (2014) 338–349.
[54] Y. Wang, Z. Wang, Efficient backsplicing produces translatable circular mRNAs, RNA 21 (2015) 172–179.
[55] Z. Abbas, R. Afzal, Life cycle and pathogenesis of hepatitis D virus: a review, World J. Hepatol. 5 (2013) 666–675.
[56] C. Alves, C. Branco, C. Cunha, Hepatitis delta virus: a peculiar virus, Adv. Virol. 2013 (2013) 560105.
[57] M.W. Hentze, T. Preiss, Circular RNAs: splicing's enigma variations, EMBO J. 32 (2013) 923–925.
[58] J.E. Wilusz, P.A. Sharp, Molecular biology. A circuitous route to noncoding RNA, Science 340 (2013) 440–441.
[59] J. Satoh, T. Yamamura, Gene expression profile following stable expression of the cellular prion protein, Cell. Mol. Neurobiol. 24 (2004) 793–814.
[60] B. Humphries, C. Yang, The microRNA-200 family: small molecules with novel roles in cancer development, progression and therapy, Oncotarget 6 (2015) 6472–6498.
[61] R. Garzon, G.A. Calin, C.M. Croce, MicroRNAs in cancer, Annu. Rev. Med. 60 (2009) 167–179.
[62] A. Esquela-Kerscher, F.J. Slack, Oncomirs – microRNAs with a role in cancer, Nat. Rev. Cancer 6 (2006) 259–269.
[63] E.J. Dropcho, Y.T. Chen, J.B. Posner, L.J. Old, Cloning of a brain protein identified by autoantibodies from a patient with paraneoplastic cerebellar degeneration, Proc. Natl. Acad. Sci. U.S.A. 84 (1987) 4552–4556.
[64] W.J. Lukiw, Circular RNA (circRNA) in Alzheimer's disease (AD), Front. Genet. 4 (2013) 307.
[65] K.F. Meza-Sosa, E.I. Pérez-García, N. Camacho-Concha, O. López-Gutiérrez, G. Pedraza-Alva, L. Pérez-Martínez, MiR-7 promotes epithelial cell transformation by targeting the tumor suppressor KLF4, PLoS ONE 9 (2014) e103987.
[66] T. Suto, T. Yokobori, R. Yajima, H. Morita, T. Fujii, S. Yamaguchi, et al., MicroRNA-7 expression in colorectal cancer is associated with poor prognosis and regulates cetuximab sensitivity via EGFR regulation, Carcinogenesis 36 (2015) 338–345.
[67] Z. Hao, J. Yang, C. Wang, Y. Li, Y. Zhang, X. Dong, et al., MicroRNA-7 inhibits metastasis and invasion through targeting focal adhesion kinase in cervical cancer, Int. J. Clin. Exp. Med. 8 (2015) 480–487.
[68] F.C. Kalinowski, R.A. Brown, C. Ganda, K.M. Giles, M.R. Epis, J. Horsham, et al., microRNA-7: a tumor suppressor miRNA with therapeutic potential, Int. J. Biochem. Cell Biol. 54 (2014) 312–317.
[69] X.Y. Fang, H.F. Pan, R.X. Leng, D.Q. Ye, Long noncoding RNAs: novel insights into gastric cancer, Cancer Lett. 356 (2015) 357–366.
[70] P. Glažar, P. Papavasileiou, N. Rajewsky, circBase: a database for circular RNAs, RNA 20 (2014) 1666–1670.
[71] S. Ghosal, S. Das, R. Sen, P. Basak, J. Chakrabarti, Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits, Front. Genet. 4 (2013) 283.

NATURE|Vol 447|24 May 2007|doi:10.1038/nature05913

Perceptions of epigenetics

Adrian Bird[1]

Geneticists study the gene; however, for epigeneticists, there is no obvious 'epigene'. Nevertheless, during the past year, more than 2,500 articles, numerous scientific meetings and a new journal were devoted to the subject of epigenetics. It encompasses some of the most exciting contemporary biology and is portrayed by the popular press as a revolutionary new science — an antidote to the idea that we are hard-wired by our genes. So what is epigenetics?

There has always been a place in biology for words that have different meanings for different people. Epigenetics is an extreme case, because it has several meanings with independent roots. To Conrad Waddington, it was the study of epigenesis: that is, how genotypes give rise to phenotypes during development[1]. By contrast, Arthur Riggs and colleagues defined epigenetics as "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence"[2]: in other words, inheritance, but not as we know it. These definitions differ markedly, although they are often conflated as though they refer to a single phenomenon. Waddington's term encompasses the activity of all developmental biologists who study how gene activity during development causes the phenotype to emerge, but it suffers from the disadvantage that developmental biologists themselves rarely, if ever, use this word to describe their field. In this sense, the usage is obsolete. The definition put forward by Riggs and colleagues tells us what epigenetics is not (inheritance of mutational changes), leaving open what kinds of mechanism are at work. In this article, I give examples of how epigenetic phenomena are studied and interpreted, and I propose a revised definition that embodies contemporary usage of the word.

The molecular basis of heritable epigenetics has been studied in a variety of organisms. The DNA methylation system and the Polycomb/Trithorax systems come closest to the ideal, because alterations in these systems are often inherited by subsequent generations of cells and sometimes organisms (Box 1). A classic case of what Robin Holliday named epimutation[3] is the peloric variant of toadflax (*Linaria*) flowers (Fig. 1), first described by Linnaeus. In this variant, heritable silencing of the gene *Lcyc*, which controls flower symmetry, is due not to a conventional mutation (that is, a mutation in the nucleotide sequence) but to the stable transmission of DNA methylation at this locus from generation to generation[4]. Although most variants arising in laboratory plants are due to conventional mutations rather than epimutations of this kind, examples of transgenerational epigenetics are now well documented in plants (see page 418) and fungi. In animals, however, the transmission of epigenetic traits between organismal generations has, so far, been detectable only by using highly sensitive genetic assays[5]. The mouse agouti locus (also known as nonagouti), which affects coat colour, is the best-studied example, being affected by the extent of DNA methylation at an upstream transposon. Genetically identical parents whose agouti genes are in different epigenetic states tend to produce offspring with different coat colours, although the effect is variable.

Despite the paucity of data from animal studies, this type of epigenetics has caught the general imagination because, in principle, it is stable but potentially affected by the environment. The possibility that acquired 'marks' can be passed from parents to children has a deliciously lamarckian flavour that has proved difficult to resist as a potential antidote to genetic determinism. A recent BBC television science programme hailed the advent of epigenetics as a profound shift in our understanding of inheritance (http://www.bbc.co.uk/sn/tvradio/programmes/horizon/ghostgenes.shtml). It summarized the implications of the emergent science as follows: "At the heart of this new field is a simple but contentious idea — that genes have a 'memory'. That the lives of your grandparents — the air they breathed, the food they ate, even the things they saw — can directly affect you, decades later, despite your never experiencing these things yourself." Is there any evidence for these heady claims, and how reliable is it? The answer to the first part of the question is yes.

Genes learning by experience?

Several studies have reported evidence that links the environment or ageing to long-lasting epigenetic effects on phenotype. One study

Box 1 | Epigenetic paradigms

There are two classic epigenetic systems: the Polycomb and Trithorax (Polycomb/Trithorax) systems, and DNA methylation. The Polycomb and Trithorax groups of proteins, which are named after mutants of the fruitfly *Drosophila melanogaster*, work to maintain repressed or active transcription states, respectively, of developmentally important genes. In the absence of these systems, the genes that specify the different segments of the fruitfly are initially expressed correctly, but this pattern cannot be maintained. It can be inferred from this that the Polycomb/Trithorax systems stably 'memorize' gene-expression patterns that have been set up by other cellular mechanisms. There is evidence that Polycomb-imposed silencing can even be transmitted between fruitfly generations at low frequency[18]. Biochemical studies have enabled the identification of components of the two key Polycomb-system protein complexes and have established a close link with modification of the lysine residue at position 27 of histone H3. The mechanism by which silencing is transmitted between cell generations remains obscure.

In the case of DNA methylation, biochemical information preceded genetic understanding of the system. The methylated sequence in vertebrates is CG, which is paired with the same sequence on the opposite DNA strand. This symmetry means that sites are transiently methylated on only one of the two DNA strands (that is, hemimethylated) after DNA replication. CG methylation patterns are copied between cell generations by the DNA methyltransferase DNMT1, which 'completes' hemimethylated but not unmethylated sites. In plants and fungi, the base 5-methylcytosine is also present in non-symmetrical DNA sequences, so the mechanism of copying is less obvious. DNA methylation is associated with stable gene silencing (for example, on the inactive X chromosome), either through interference with transcription-factor binding or through the recruitment of repressors that specifically bind sites containing methylated CG.

[1]Wellcome Trust Centre for Cell Biology, Edinburgh University, The King's Buildings, Edinburgh EH9 3JR, UK.

examined monozygotic (that is, identical) twins, whom, perhaps oddly, epigeneticists often use to exemplify their system at work. To many, twins epitomize the awesome power of genetics to determine human form and function regardless of environment. Indeed, 'concordance' of a particular characteristic in monozygotic and dizygotic twins is one of the most reliable ways of assessing its genetic basis. What has attracted the attention of epigeneticists, however, is that monozygotic twins do not always show the same disease susceptibility, raising the possibility that epigenetic differences that arise during ageing are at work[6]. Accordingly, it has been reported that young twins have similar amounts of DNA methylation, whereas older twins differ considerably in the amounts and patterns of this modification[7]. Might these non-genetic age-dependent differences in gene marking give rise to the divergent disease predispositions seen in some twins? At present, this is unclear, and a recent study emphasizes the need for further basic work on twins. The largest high-resolution analysis of human DNA methylation patterns so far found that 873 genes on 3 chromosomes showed no significant variation in DNA methylation between individuals in their mid-20s and those in their mid-60s[8]. The remarkable uniformity of DNA methylation among unrelated individuals of disparate ages does not square easily with the large divergence reported in twins of the same age.

Another high-profile study has raised the possibility that a mother's behaviour can affect the chemistry of DNA in her offspring. Quality of early maternal care has long been acknowledged to have long-term repercussions during the lifetime of an individual. A potential mechanism for this effect was deduced from a study reporting that maternal nurturing in rats alters DNA methylation at the gene encoding the glucocorticoid receptor[9]. The authors suggest that in the absence of appropriate nurturing, there is less methylation of this gene in the hippocampus, resulting in overexpression of the receptor in later life. The implication is that the glucocorticoid-mediated stress-response pathway is epigenetically fixed at the level of gene transcription. In addition, transgenerational effects of environmental insults have been reported in mammals: for example, the exposure of embryonic rats to the anti-androgenic compound vinclozolin led to a decrease in spermatogenesis not only in the treated animals but also in males of several subsequent generations[10]. Altered DNA methylation was again suggested as a potential mediator of this effect, although, during development, mammalian embryos pass through a profoundly hypomethylated state, which might be expected to jeopardize the heritability of such marks. Despite uncertainties about the mechanism(s) at work, these studies have raised the profile of epigenetics as a potential mechanistic explanation for the long-term impact of the environment on physiology and behaviour (see page 433). Time will tell whether that potential is realized.

Epigenetics and inheritance

Should heritability be mandatory in a contemporary view of epigenetics? The requirement that epigenetic characters should be transmissible through mitosis or meiosis has the virtue of clarity but can be a liability. To explain why, it is necessary to introduce a third, somewhat informal, 'definition' of epigenetics that has crept into widespread use. This incarnation of epigenetics encompasses the biology of chromatin, including the complex language of chromatin marks (see page 407), the transcriptional effects of RNA interference (see page 399) and, for good measure, the effects of the higher-order structure of chromosomes and the nucleus (see page 413). The attraction of this usage is that it brackets together some of the most exciting contemporary work in biology. Its drawback is that it does not sit easily with the prevailing textbook definitions. One reason for this is that many chromatin marks are short-lived. For example, phosphorylation of the variant histone H2AX (also known as H2AFX) after a double-strand break[11] would qualify as an epigenetic mark under the emerging definition, but it is too transient to qualify as a heritable epigenetic mark (Fig. 2). Histone modifications associated with transcription are also ambiguous with respect to heritability. On the one hand, DNA methylation affects histone acetylation and histone methylation, so these modifications can be viewed as heritably epigenetic, albeit indirectly[12]. On the other hand, these histone marks can also result from events that seem to involve neither DNA methylation nor Polycomb group proteins, and the marks are not necessarily transmissible between generations. Therefore, a single histone modification could, in principle, be rated as either epigenetic or not epigenetic according to the heritability credentials of its origin. Such a complicated classification system would have limited utility.

The issue of replicative accuracy is also relevant when considering heritability. DNA synthesis is spectacularly accurate, making only

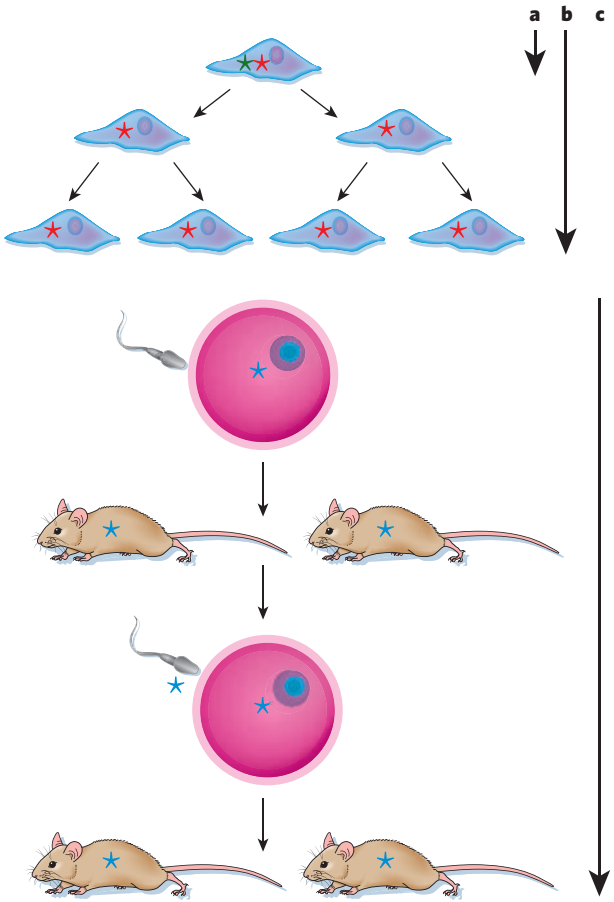


Figure 2 | Persistence of epigenetic marks. Alterations that last less than one cell cycle (green asterisk, **a**) do not qualify as epigenetic under the definition that strictly requires heritability, whereas non-mutational changes that are transmitted from one cell to its daughters (red asterisk, **b**) or between generations of an organism (blue asterisk, **c**) do qualify.

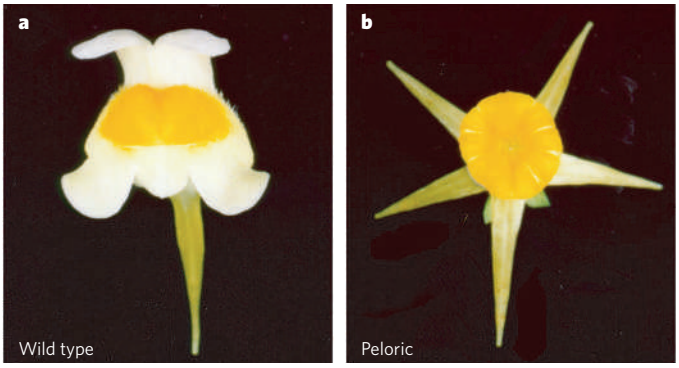


Figure 1 | Frontal view of a wild-type toadflax flower and a peloric epimutant. a, The wild-type flower is dorsoventrally asymmetrical. **b**, By contrast, the peloric flower is radially symmetrical with all petals resembling the ventral petal of the wild-type flower. (Image reprinted, with permission, from ref. 4.)

1 'unforced' error for every 10^7–10^8 bases copied[13]. But DNA methylation has an apparent accuracy of ~96%, which is ~1 error for every 25 methylated sites copied[14]. Because of this error rate, cloning from a single cell quickly results in a population of cells in which DNA methylation patterns are diverse[15]. Methylated domains are more stably maintained, even though the detailed location of methylated sites varies within them. But even the peloric variant of toadflax, which is an otherwise perfect example of heritable epigenetics in action, shows considerable instability as the plant grows. So how accurately transmitted should an epigenetic mark be? Variation due to faulty copying is compounded by current evidence that all histone modifications, as well as DNA methylation itself, can be abruptly removed during development, thereby preventing the persistence of these modifications in a heritable epigenetic sense (see page 425). The restrictiveness of the heritable view of epigenetics is perhaps best illustrated by considering the brain. A growing idea is that functional states of neurons, which can be stable for many years, involve epigenetic phenomena[16], but these states will not be transmitted to daughter cells because almost all neurons never divide.

Refining a definition

Given that there are several existing definitions of epigenetics, it might be felt that another is the last thing we need. Conversely, there might be a place for a view of epigenetics that keeps the sense of the prevailing usages but avoids the constraints imposed by stringently requiring heritability. The following could be a unifying definition of epigenetic events: the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states. This definition is inclusive of chromosomal marks, because transient modifications associated with both DNA repair or cell-cycle phases and stable changes maintained across multiple cell generations qualify. It focuses on chromosomes and genes, implicitly excluding potential three-dimensional architectural templating of membrane systems and prions, except when these impinge on chromosome function. Also included is the exciting possibility that epigenetic processes are buffers of genetic variation, pending an epigenetic (or mutational) change of state that leads an identical combination of genes to produce a different developmental outcome[17].

An implicit feature of this proposed definition is that it portrays epigenetic marks as responsive, not proactive. In other words, epigenetic systems of this kind would not, under normal circumstances, initiate a change of state at a particular locus but would register a change already imposed by other events. Such events could be, for example, the collision of DNA with ionizing radiation or a developmental switch in gene expression. It could be argued that the responsive nature of epigenetic processes is a unifying feature, because classic epigenetic systems such as the DNA methylation system and the Polycomb/Trithorax systems seem to respond to previous switches in gene activity in this way. Therefore, their sophisticated feature is the ability, in the 'darkness' of the nucleus, to sense and mark changes in the chromosomal status. For example, transcriptional activation through sequence-specific DNA-binding proteins brings in histone acetyltransferases, which then epigenetically adapt the promoter region for transcription (for histone acetyl groups, although ephemeral, would now be epigenetic). Similarly, elongating polymerases carry enzymes that restrain the spurious transcriptional initiation that might arise within the temporarily disrupted chromatin of an active gene. Without such epigenetic mechanisms, hard-won changes in genetic programming could be dissipated and lost; transient disruptions of chromosomal organization might go uncompensated; and DNA damage might escape repair. ■

1. Waddington, C. H. *The Strategy of the Genes* (Allen & Unwin, London, 1957).
2. Russo, V. E. A., Martienssen, R. A. & Riggs, A. D. (eds) *Epigenetic Mechanisms of Gene Regulation* (Cold Spring Harbor Laboratory Press, Woodbury, 1996).
3. Jeggo, P. A. & Holliday, R. Azacytidine-induced reactivation of a DNA repair gene in Chinese hamster ovary cells. *Mol. Cell. Biol.* **6,** 2944–2949 (1986).
4. Cubas, P., Vincent, C. & Coen, E. An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* **401,** 157–161 (1999).
5. Chong, S. & Whitelaw, E. Epigenetic germline inheritance. *Curr. Opin. Genet. Dev.* **14,** 692–696 (2004).
6. Wong, A. H., Gottesman, I. I. & Petronis, A. Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Hum. Mol. Genet.* **14,** R11–R18 (2005).
7. Fraga, M. F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl Acad. Sci. USA* **102,** 10604–10609 (2005).
8. Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genet.* **38,** 1378–1385 (2006).
9. Weaver, I. C. *et al.* Epigenetic programming by maternal behavior. *Nature Neurosci.* **7,** 847–854 (2004).
10. Anway, M. D., Cupp, A. S., Uzumcu, M. & Skinner, M. K. Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science* **308,** 1466–1469 (2005).
11. Rogakou, E. P., Boon, C., Redon, C. & Bonner, W. M. Megabase chromatin domains involved in DNA double-strand breaks *in vivo*. *J. Cell Biol.* **146,** 905–916 (1999).
12. Klose, R. J. & Bird, A. P. Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.* **31,** 89–97 (2006).
13. Kunkel, T. A. DNA replication fidelity. *J. Biol. Chem.* **279,** 16895–16898 (2004).
14. Laird, C. D. *et al.* Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. *Proc. Natl Acad. Sci. USA* **101,** 204–209 (2004).
15. Silva, A. J., Ward, K. & White, R. Mosaic methylation in clonal tissue. *Dev. Biol.* **156,** 391–398 (1993).
16. Hong, E. J., West, A. E. & Greenberg, M. E. Transcriptional control of cognitive development. *Curr. Opin. Neurobiol.* **15,** 21–28 (2005).
17. Sollars, V. *et al.* Evidence for an epigenetic mechanism by which Hsp90 acts as a capacitor for morphological evolution. *Nature Genet.* **33,** 70–74 (2003).
18. Cavalli, G. & Paro, R. The *Drosophila Fab-7* chromosomal element conveys epigenetic inheritance during mitosis and meiosis. *Cell* **93,** 505–518 (1998).

Acknowledgements I thank the Wellcome Trust for research support.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence should be addressed to the author (apbird@staffmail.ed.ac.uk).

NATURE|Vol 447|24 May 2007|doi:10.1038/nature05917

Epigenetic inheritance in plants

Ian R. Henderson[1] & Steven E. Jacobsen[1]

The function of plant genomes depends on chromatin marks such as the methylation of DNA and the post-translational modification of histones. Techniques for studying model plants such as *Arabidopsis thaliana* have enabled researchers to begin to uncover the pathways that establish and maintain chromatin modifications, and genomic studies are allowing the mapping of modifications such as DNA methylation on a genome-wide scale. Small RNAs seem to be important in determining the distribution of chromatin modifications, and RNA might also underlie the complex epigenetic interactions that occur between homologous sequences. Plants use these epigenetic silencing mechanisms extensively to control development and parent-of-origin imprinted gene expression.

Eukaryotic genomes are covalently modified with a diverse set of chromatin marks, which are present on both the DNA and the associated histones (see page 407). Although these changes do not alter the primary DNA sequence, they are frequently heritable through cell division, sometimes for multiple generations, and can thus often be classified as epigenetic marks. These conserved epigenetic marks have been found to influence many aspects of gene expression and chromosome biology, and they have characteristic genomic distributions.

The size of eukaryotic genomes varies extensively and does not correlate with gene number[1]. This is often because of the presence of large amounts of non-gene sequences, which can include pseudogenes, transposable elements, integrated viruses and simple repeats[1]. At the chromosomal level, genomes are organized into euchromatin, which is gene-rich, and heterochromatin, which is repeat-rich[2]. Heterochromatin is defined by three main properties: greater compaction than other genomic regions during interphase, lower accessibility than other regions to transcription and recombination machinery, and the formation of structured nucleosome arrays[2] (see page 399). The defining characteristics of heterochromatin depend on epigenetic information, including post-translational modification of histones and methylation of cytosine bases in DNA[2,3]. The silencing of transposable-element sequences within heterochromatin is probably a genome-defence strategy. However, heterochromatin can also have important roles during chromosomal segregation[4], and transposons and epigenetic silencing have been shown to both modulate gene expression and contribute to *cis*-regulatory sequences[5,6]. Plant systems have been a rich source for the study of epigenetic inheritance, and examples of important discoveries include transposable elements[7], paramutation[8], small interfering RNAs (siRNAs)[9] and RNA-directed DNA methylation[10].

Genomic resources for studying the model plant *Arabidopsis thaliana* have begun to provide insight into the epigenetic 'landscape' of this organism[11,12]. *A. thaliana* has a compact ~130-megabase (Mb) genome, although it contains considerable amounts of heterochromatin, which is repeat-rich and largely located in the centromeric and pericentromeric regions[13,14] (Fig. 1). High-resolution mapping of cytosine methylation by using whole-genome microarrays has confirmed previous reports, showing that this modification co-localizes with repeat sequences and with the centromeric regions[11,12,15]. Fewer than 5% of expressed genes were shown to have methylated promoters, although about one-third of genes were methylated in their open reading frame[11,12]. The significance of methylation in the body of a gene is not fully understood, but such methylation was found to correlate with genes that are both highly transcribed and constitutively expressed[11,12]. By contrast, genes with methylated promoters had lower expression levels and frequently had tissue-specific expression patterns[11,12]. This distribution of cytosine methylation is in contrast to that observed in mammalian genomes, which are often densely methylated but have hypomethylated CG islands in gene promoters[3]. It will be important to describe the 'methylome' of other repeat-rich plant genomes, such as those of the grasses, to test the generality of the patterns observed in *A. thaliana*. Here, we review the emerging and prominent role of RNA in epigenetic inheritance in plants and how such mechanisms are used to control development.

Mediating silencing with RNA

A central question in understanding the epigenetic regulation of genomes is how sequences are recognized or avoided as targets for silencing. There is an increasing appreciation that siRNAs, which are generated by the RNA interference (RNAi) pathway, can provide sequence specificity to guide epigenetic modifications in a diverse range of eukaryotes. Well-studied examples include transcriptional silencing in yeast[16] (see page 399), cytosine methylation in plants[10,17] and genome rearrangements in ciliates[18]. RNA-directed DNA methylation was discovered in tobacco, in which genomic sequences homologous to infectious RNA viroids were found to become cytosine methylated[10]. Subsequently, the expression of double-stranded RNA (dsRNA) in plants was shown to generate siRNAs and cause dense cytosine methylation of homologous DNA in all sequence contexts[19]. This is reflected by the high coincidence of endogenous siRNA clusters with methylated sequences and repeats in *A. thaliana*[11,12,15,20].

All known *de novo* DNA methylation in *A. thaliana* is carried out by DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2), which is a homologue of the mammalian DNA methyltransferase 3 (DNMT3) enzymes[21–24] (Fig. 2b). DRM2 can be targeted to a sequence by siRNAs generated from the expression of either direct or inverted repeats[23,24]. Plants encode multiple homologues of the RNAi-machinery components, some of which are specialized for function in RNA-directed DNA methylation[25,26]. The endoribonuclease DICER-LIKE 3 (DCL3) generates 24-nucleotide siRNAs, which are loaded into the PAZ- and PIWI-domain-containing protein ARGONAUTE 4 (AGO4)[26–31] (Fig. 2a). These AGO4-associated siRNAs are proposed to guide the cytosine-methyltransferase activity of DRM2 (refs 26–31). The mechanism by which siRNAs target epigenetic modifications is poorly understood and could involve either DNA–RNA or RNA–RNA hybridization events. Interestingly, epigenetic modifications guided by AGO4 in

[1]Department of Molecular, Cell and Developmental Biology, Howard Hughes Medical Institute, University of California, Los Angeles, California 90095, USA.

A. thaliana have been shown to depend partly on the RNaseH ('slicer') catalytic activity of AGO4 (ref. 30). This could be taken as support for RNA–RNA hybridization having an important role in the targeting of epigenetic modifications.

The accumulation of siRNAs associated with RNA-directed DNA methylation in *A. thaliana* often depends on RNA-DEPENDENT RNA POLYMERASE 2 (RDR2) and the plant-specific protein NUCLEAR RNA POLYMERASE IV A (also known as NUCLEAR RNA POLYMERASE D 1A; NRPD1A), which are involved in a putative amplification pathway[26,32–35] (Fig. 2a). Together, RDR2 and NRPD1A might generate dsRNA substrates for DCL3 to process into siRNAs, although how these proteins are recruited to target loci is unknown. Several loci also show dependence on AGO4 and DRM2 for siRNA accumulation, suggesting that there might be a feedback loop between transcriptional silencing and siRNA generation[24,26].

NRPD1A functions in a complex with NRPD2. A variant of this NRPD complex, which contains NRPD1B instead of NRPD1A, is also required for RNA-directed DNA methylation but participates less frequently in siRNA accumulation[33,35] (Fig. 2a). One possible function for the NRPD1B-containing complex is to generate a target transcript that can hybridize with siRNA-loaded AGO4-containing complexes. Indeed, AGO4 has been observed to bind directly to NRPD1B[28]. The SWI–SNF-family chromatin-remodelling protein DEFECTIVE IN RNA-DIRECTED DNA METHYLATION 1 (DRD1) is also required for RNA-directed DNA methylation and could function to facilitate access of DRM2 to target DNA[27,36]. Recently, several proteins in the RNA-directed DNA-methylation pathway have been found to localize to distinct nuclear bodies, including the Cajal body, which is a centre for the processing and modification of many non-coding RNAs[28,29]. Localization to these bodies might be required for the efficient loading of AGO4-containing complexes with siRNA before these complexes travel to the nucleoplasm and, together with DRM2, direct RNA-directed DNA methylation.

Plants show extensive methylation of cytosine bases in the CG, CNG (where N denotes any nucleotide) and CHH (where H denotes A, C or T) sequence contexts[37]. By contrast, most cytosine methylation in mammals is found in the CG sequence context[3,38]. CG methylation is maintained by the homologous proteins METHYLTRANSFERASE 1 (MET1) and DNMT1 in plants and mammals, respectively[39,40] (Fig. 2b). DNMT1 has a catalytic preference for hemimethylated substrates, providing an attractive model for the efficient maintenance of CG methylation after DNA replication and during cell division[38]. Most non-CG methylation in plants is maintained redundantly by DRM2 and the plant-specific protein CHROMOMETHYLASE 3 (CMT3)[23,37] (Fig. 2b); however, some loci show residual non-CG methylation in *drm1 drm2 cmt3* triple mutants, which might be maintained by MET1 (ref. 25). Non-CG methylation differs from CG methylation, because it seems to require an active maintenance signal after DNA replication. At some loci, siRNAs seem to provide this signal, acting through DRM2 activity: for example, at the *MEA-ISR* locus (*MEDEA INTERSTITIAL SUBTELOMERIC REPEATS* locus, an array of seven tandem repeats located downstream of the *MEDEA* gene), the repeats lose all non-CG methylation in *drm2* mutants and in several RNAi-pathway mutants such as *ago4* and *rdr2* (refs 23, 37). By contrast, other loci — for example, the SINE-class retrotransposon *AtSN1* — completely lose non-CG methylation only in *drm1 drm2 cmt3* triple mutants. At *AtSN1*, CMT3 contributes to the maintenance of both CNG methylation and asymmetrical (CHH) methylation. The activity of CMT3 largely depends on the main methyltransferase for H3K9 (the lysine residue at position 9 of histone H3) — SU(VAR)3-9 HOMOLOGUE 4 (SUVH4; also known as KRYPTONITE) — showing that histone methylation is also an important signal for the maintenance of non-CG methylation[41,42]. At present, the factors that determine the relative importance of the RNAi pathway and histone methylation for the maintenance of non-CG methylation at different loci remain unclear.

Communication of silent information

Epigenetically silent expression states can show remarkable stability throughout mitosis and meiosis, although they can retain the ability to revert to an active state[2]. This gives rise to the concept of the epigenetic allele (epiallele), which is defined as an allele that shows a heritable difference in expression as a consequence of epigenetic modifications and not changes in DNA sequence. For example, hypermethylated (silent) epialleles of *SUPERMAN* (which is involved in floral development) known as *clark kent* are stable during many generations of inbreeding, but they can revert to an unmethylated (active) state at a frequency of ~3% per generation[43]. Another notable characteristic of certain epialleles is their ability to influence other homologous sequences both *in cis* and

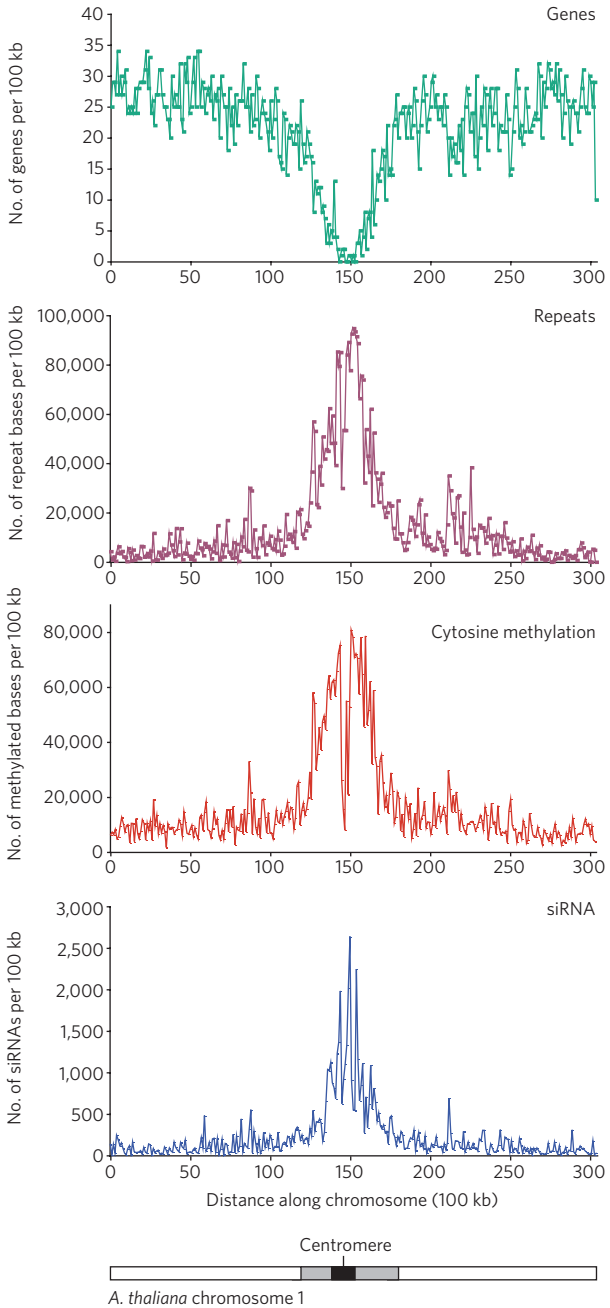


Figure 1 | The epigenetic 'landscape' of *A. thaliana*. The relative abundance of genes (number of annotated genes[11]), repeats (repeat bases per 100 kb; ref. 11), cytosine methylation (methylated bases per 100 kb; ref. 11) and siRNAs (cloned siRNAs per 100 kb; ref. 20) is shown for the length of *A. thaliana* chromosome 1, which is ~30 Mb. Numbers on the *x* axis represent 100-kb windows along the chromosome. A diagram of chromosome 1 is also shown, with white bars indicating euchromatic arms, grey bars indicating pericentromeric heterochromatin and the black bar indicating the centromeric core. (Figure courtesy of X. Zhang, University of California, Los Angeles.)

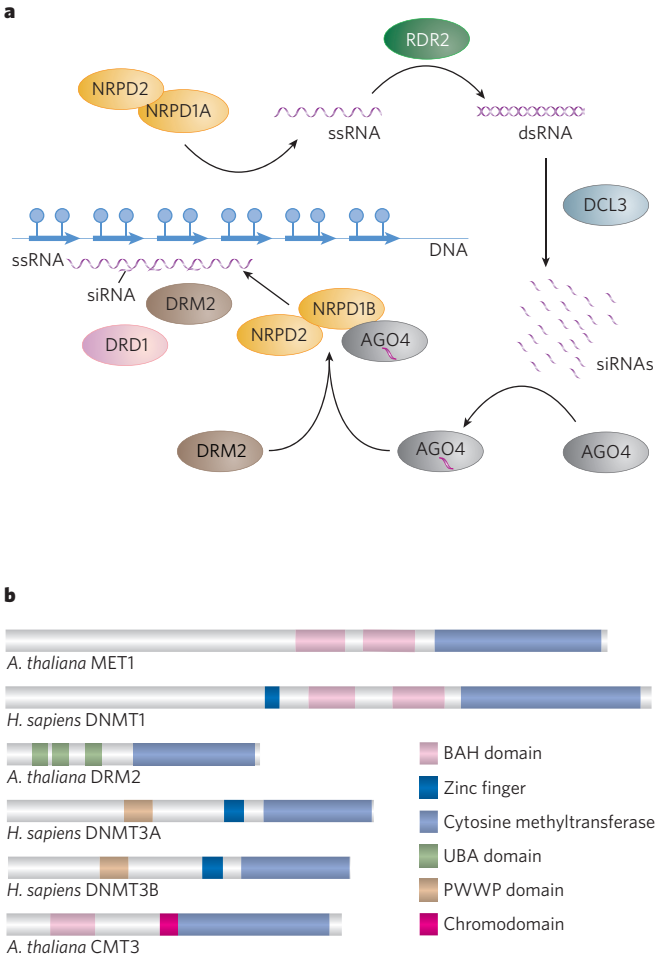


Figure 2 | RNA-directed DNA methylation. a, Putative pathway for RNA-directed DNA methylation in *A. thaliana*. Target loci (in this case tandemly repeated sequences; coloured arrows) recruit an RNA polymerase IV complex consisting of NRPD1A and NRPD2 through an unknown mechanism, and this results in the generation of a single-stranded RNA (ssRNA) species. This ssRNA is converted to double-stranded RNA (dsRNA) by the RNA-dependent RNA polymerase RDR2. The dsRNA is then processed into 24-nucleotide siRNAs by DCL3. The siRNAs are subsequently loaded into the PAZ- and PIWI-domain-containing protein AGO4, which associates with another form of the RNA polymerase IV complex, NRPD1B–NRPD2. AGO4 that is 'programmed' with siRNAs can then locate homologous genomic sequences and guide the protein DRM2, which has *de novo* cytosine methyltransferase activity. Targeting of DRM2 to DNA sequences also involves the SWI–SNF-family chromatin-remodelling protein DRD1. The NRPD1B–NRPD2 complex might generate a target transcript (ssRNA) to which the AGO4-associated siRNAs can hybridize. Given that siRNAs homologous to some loci are absent in *drm2* mutants and *ago4* mutants, it is possible that DNA methylation (blue circles) also stimulates siRNA generation and reinforces silencing. **b**, DNA methyltransferase structure and function. Plant and mammalian genomes encode homologous cytosine methyltransferases, of which there are three classes in plants and two in mammals. *A. thaliana* MET1 and *Homo sapiens* (human) DNMT1 both function to maintain CG methylation after DNA replication, through a preference for hemimethylated substrates, and both have amino-terminal bromo-adjacent homology (BAH) domains of unknown function. *De novo* DNA methylation is carried out by the homologous proteins DRM2 (in *A. thaliana*) and DNMT3A and DNMT3B (both in *H. sapiens*). Despite their homology, these proteins have distinct N-terminal domains, and the catalytic motifs present in the cytosine methyltransferase domain are ordered differently in DRM2 and the DNMT3 proteins. Plants also have another class of methyltransferase, which is not found in mammals. CMT3 functions together with DRM2 to maintain non-CG methylation. PWWP, Pro-Trp-Trp-Pro motif; UBA, ubiquitin associated.

in trans[2]. One example is paramutation, which was discovered in plants and is defined as allelic interactions that cause a meiotically heritable change in the expression of one of the alleles[8]. *Trans*-phenomena similar to paramutation have also been described in mammals, including at a chimaeric version of the mouse *Rasgrf1* (Ras protein-specific guanine-nucleotide-releasing factor 1) locus that contained the imprinting control region from the insulin-like growth factor 2 receptor gene[44].

One of the best-studied paramutation systems is the maize (*Zea mays*) locus *b1*, which encodes a transcription factor that is required for accumulation of the pigment anthocyanin[8]. The paramutagenic epiallele *B′*, which causes light pigmentation, arises spontaneously at a low frequency from its paramutable parent allele *B-I*, which causes dark pigmentation[45]. *B′* epialleles convert *B-I* alleles to *B′* epialleles when heterozygous with 100% penetrance, and the newly created paramutated *B′* epialleles can pass on their silent state in subsequent crosses[45] (Fig. 3). *B′* epialleles are transcribed at one-twentieth to one-tenth the rate of *B-I* alleles but have identical gene sequences[45,46]. Fine-structure recombination mapping of alleles resulting from a cross between individuals with paramutagenic alleles and those with neutral alleles (which cannot participate in paramutation) enabled the sequences required for paramutation to be defined; these sequences are present as an array of 7 tandem 853-base repeats, which is located ~100 kilobases (kb) upstream of *b1* (refs 45, 46). The sequences are present as a single copy in neutral alleles. Recombinant alleles with three repeats show partial paramutational ability, whereas alleles with seven repeats are fully active in paramutation[45,46]. These repeats were also shown to have a closed chromatin structure and more cytosine methylation in *B′* epialleles than in *B-I* alleles[46]. However, for *B′*, cytosine methylation was found to be established after the silent state, so it is unlikely to be the cause[46]. There are several models of *trans*-communication between alleles, including physical pairing of alleles and transmission of an RNA signal. A model for paramutagenic interactions being mediated by siRNA is supported by the finding that a genetic suppressor of paramutation, *mediator of paramutation1* (*mop1*), encodes the maize orthologue of the RNA-dependent RNA polymerase RDR2 (refs 47, 48). So far, siRNAs homologous to the tandem repeats upstream of *B′* have not been reported, although such repeats are commonly associated with small RNAs[20,49]. The *mop1* gene is also required for silencing transgenes and *Mutator*-like transposons, indicating that RNA-dependent RNA polymerases and siRNAs have a role in heterochromatic silencing in monocotyledonous plants[50]. The detailed relationships between siRNAs, chromatin structure at the repeats upstream of *B′*, and the ability to transfer epigenetic states will be intriguing to determine.

The *A. thaliana* gene *FWA* has similarities to maize *b1* in that it has tandem repeats upstream that, when methylated, cause heritable silencing of expression[51]. Stably hypomethylated *fwa-1* epialleles have been found to be generated spontaneously and in *met1* mutant backgrounds[39,40,51], causing overexpression of the transcription factor FWA and a dominant late-flowering phenotype[51]. In contrast to *B′* epialleles, methylated and unmethylated *fwa* epialleles are not influenced by the presence of one another in heterozygotes[23,49,51]. However, introduction of unmethylated transgenic copies of *FWA* by *Agrobacterium tumefaciens*-mediated transformation leads to efficient *de novo* silencing of the incoming transgene, in a process that depends on both DRM2 and the RNA-directed DNA-methylation RNAi pathway[22,23] (Fig. 3). Intriguingly, an unmethylated *FWA* transgene obtained after transformation into a *drm2* mutant does not become remethylated after outcrossing to wild-type *A. thaliana*[22,23]. This finding suggests that, during the transformation process, there is a 'surveillance' window when the incoming *FWA* transgene is competent to be silenced. *A. tumefaciens* targets the female gametophyte (which is haploid) during transformation, but introduction of *FWA* into *DRM2/drm2* heterozygotes revealed that the silencing window must be present after fertilization[49]. Structure–function analysis of an *FWA* transgene showed that the upstream tandem repeats are necessary and sufficient for transformation-dependent silencing and were also found to produce homologous siRNA[49]. Interestingly, the efficiency by which an incoming

FWA transgene is silenced can be influenced by the methylation state of endogenous *FWA*[49]. Whereas introduction of an *FWA* transgene into a background in which the endogenous *FWA* gene is methylated leads to extremely efficient silencing of the transgene, transformation into the *fwa-1* background, which contains an unmethylated endogenous gene, leads to inefficient methylation and silencing of the *FWA* transgene[49] (Fig. 3). Furthermore, an introduced transgene can occasionally cause silencing of the unmethylated *fwa-1* endogenous gene[49]. These results reveal extensive communication between the transgenic and endogenous *FWA* gene copies during transformation, and this communication depends on the DNA methylation state of the endogenous gene. Surprisingly, these differences between *fwa-1* epialleles are not accounted for by siRNA production, because the repeat-derived siRNAs accumulate equally in plants with wild-type *FWA* and those with *fwa-1* (ref. 49). Hence, recruitment of siRNA machinery to a locus is not always sufficient for RNA-directed DNA methylation and probably also requires modifications of chromatin.

Maintenance of silencing at *FWA* depends mainly on CG methylation, because *met1* alleles generate hypomethylated *fwa-1* epialleles at a high frequency[39,40]. Although the tandem repeats upstream of *FWA* are also methylated at non-CG sequences, loss of this methylation in *drm1 drm2 cmt3* triple mutants does not cause reactivation and late flowering[37]. Genome-wide analysis of cytosine methylation and transcription in *drm1 drm2 cmt3* triple mutants has identified genes with methylated promoters, the expression of which depends strongly on DRM- and CMT3-mediated non-CG methylation[11]. These methylated genes might be responsible for the developmental phenotypes of *drm1 drm2 cmt3* triple mutants, which include misshapen leaves and reduced stature[27,37].

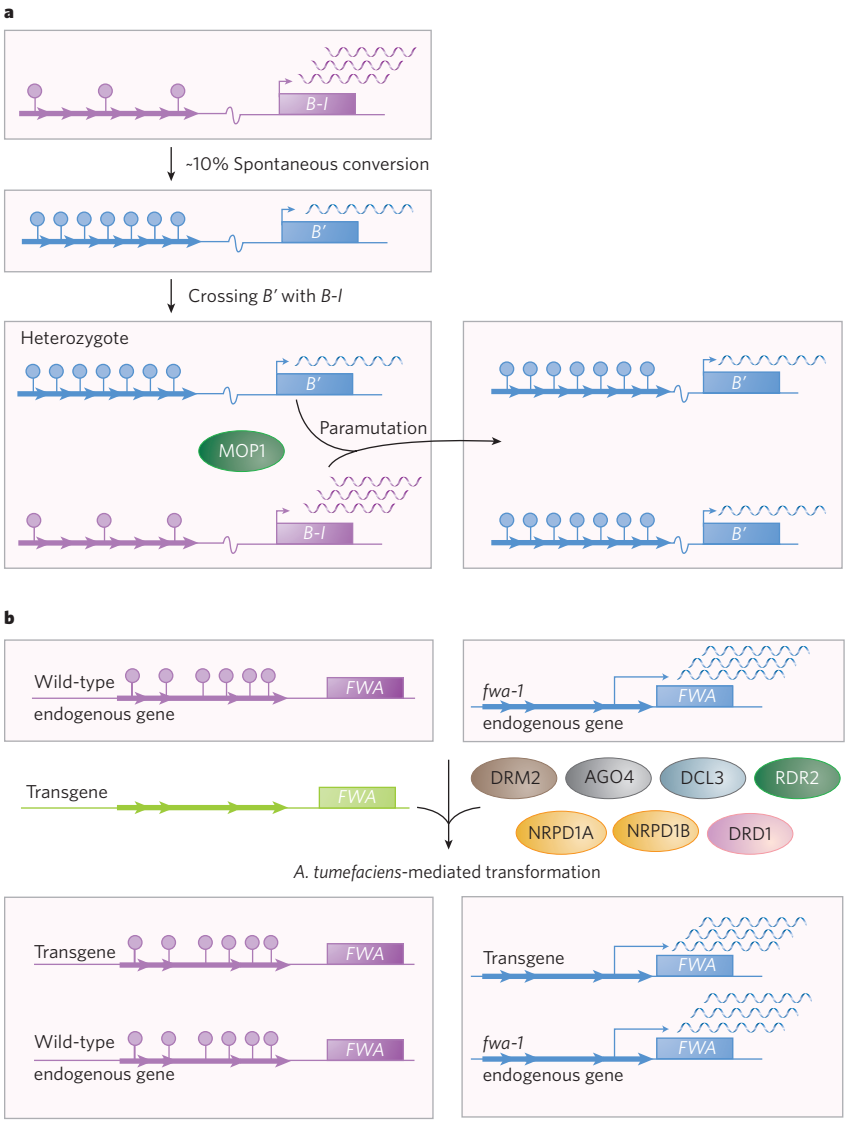


Figure 3 | *Trans*-epiallele interactions at *b1* and *FWA*. **a**, Paramutation at the *b1* locus in maize. The *B-I* allele (pink) of the *b1* gene in maize has an upstream tandem-repeat region (coloured arrows) and spontaneously gives rise to silenced *B′* epialleles (blue) at a low frequency. *B′* epialleles are more heavily methylated at cytosine bases in the repeat region and are less frequently transcribed. When the *B′* epiallele is brought together with a new copy of *B-I* by crossing of maize plants, the *B-I* allele is paramutated to a silenced *B′* state with 100% penetrance. *Trans*-communication between epialleles requires MOP1, the maize homologue of *A. thaliana* RDR2, suggesting that siRNA-mediated silencing might be involved in the conversion of *B-I* to *B′*. **b**, *De novo* silencing of *FWA* transgenes in wild-type and *fwa-1 A. thaliana*. The *FWA* gene in wild-type *A. thaliana* (pink) is methylated at cytosine bases in a pair of tandem repeats in its promoter, silencing its expression. Mutations that decrease DNA methylation give rise to hypomethylated *fwa-1* epialleles (blue), which overexpress the transcription factor FWA, thereby causing late flowering. Introduction of an unmethylated *FWA* transgene (green) by *A. tumefaciens*-mediated transformation of wild-type plants results in efficient methylation and silencing of the incoming transgene. This process depends on DRM2, AGO4, DCL3, RDR2, NRPD1A, NRPD1B and DRD1. By contrast, transformation of an *fwa-1* background results in inefficient silencing of the transgene, indicating that the methylation state of endogenous *FWA* is important for transgene silencing.

In contrast to the independently segregating epialleles that arise in *met1* mutants (as a result of the stable loss of CG methylation)[39,40,51], backcrossing *drm1 drm2 cmt3* triple mutants to wild-type plants or reintroducing either *DRM2* or *CMT3* by transformation immediately rescues these morphological phenotypes[27]. This finding suggests that non-CG methylation can be more easily re-established, possibly allowing flexible regulation of genes. However, it is unclear how commonly this type of regulation is used, because few examples of DNA-methylation-regulated plant genes have been described.

Silencing through time and development

The life cycles of plants differ from those of animals in that the products of meiosis undergo mitotic proliferation to form multicellular gametophytes (that is, the embryo sac and the pollen in flowering plants). The embryo sac (female) contains an egg cell, which is haploid, and this is fertilized by a sperm nucleus, which is also haploid, to form a diploid embryo. A second sperm nucleus fertilizes the central cell, which is diploid, to form triploid endosperm, an extra-embryonic tissue that has a supportive role during embryogenesis. The central cell and the endosperm show parent-of-origin-dependent monoallelic expression, or imprinting, which is important for proper seed development[52]. For example, in *A. thaliana*, the tandem repeats of maternal *FWA* alleles are specifically demethylated in the central cell and the endosperm, leading to expression of *FWA* in these tissues[53]. Demethylation and activation of *FWA* depend on maternal expression of the gene encoding the DNA glycosylase–lyase DEMETER (DME), which can directly excise the base 5-methylcytosine[54–56]. Because the endosperm is a terminally differentiating extra-embryonic tissue, this mechanism does not necessitate remethylation of *FWA*[53]. This is in contrast to mammals, in which demethylation of imprinted genes occurs in primordial germ cells (the cells that ultimately generate the germ line) and is followed by germline-specific remethylation and silencing (see page 425). Other imprinted genes such as *MEA* and *FERTILIZATION-INDEPENDENT SEED 2* also have cytosine-methylated regions in their promoters that are associated with maternally restricted expression[55,57]. However, only for *FWA* has it been shown that differential methylation of particular sequences is required for the regulation of imprinting[53,58].

Cytosine demethylation is also likely to have an important role in the control of silencing in situations other than gametophytic generation and imprinting. *DME* belongs to a small *A. thaliana* gene family that includes the somatically expressed gene *REPRESSOR OF SILENCING 1* (*ROS1*)[54,59]. Mutations in *ROS1* have been shown to increase RNA-directed DNA methylation, and ROS1 has been shown to function as a cytosine demethylase[56,59,60]. Together, these exciting discoveries have defined a long-sought cytosine demethylation pathway, and they raise many interesting questions. For example, to what extent are genomic methylation patterns balanced by the targeting of *de novo* DNA methyltransferases and DNA glycosylases? Furthermore, there are indications of a similar mechanism for cytosine demethylation in vertebrates[61,62].

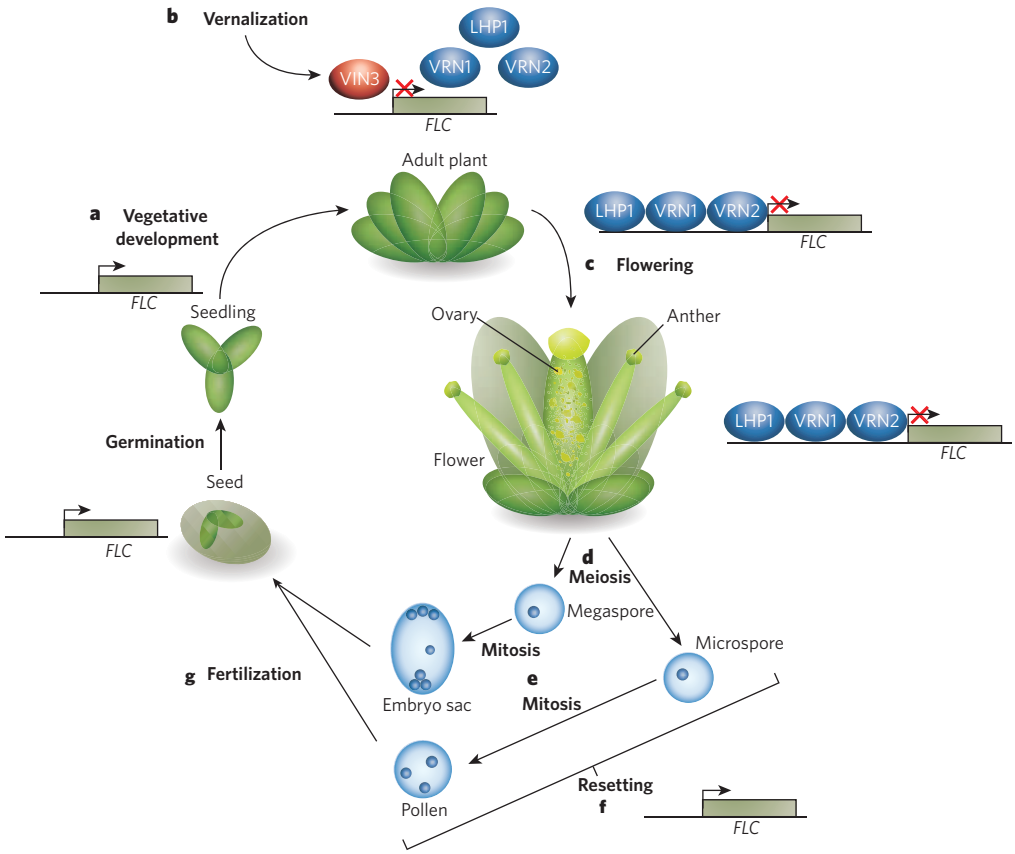


Figure 4 | PcG-protein-mediated silencing throughout the *A. thaliana* life cycle. The activation state of the PcG protein target *FLC* is illustrated throughout the plant life cycle. **a**, *FLC* is transcriptionally active in seeds and seedlings, preventing the plant from flowering and prolonging vegetative development. **b**, Exposure to a long period of cold (that is, vernalization) results in the expression of VIN3 (red), which initiates repression of *FLC* transcription, and the binding of the PcG protein VRN2, as well as VRN1 and LHP1 (blue). In this process, chromatin at *FLC* is epigenetically modified by the trimethylation of H3K27. **c**, After warmer temperatures return, *FLC* repression is maintained, allowing flowering to be induced by other cues. **d**, During flower development, the anthers and ovaries are sites of meiotic differentiation, giving rise to haploid cells known as microspores and megaspores, respectively. **e**, These meiotic products undergo mitotic proliferation to form the multicellular embryo sac and pollen gametophytes. **f**, PcG-protein-mediated repression at *FLC* is removed during an undefined resetting process. **g**, Then, the pollen contributes sperm nuclei to the embryo sac, and these fertilize the haploid egg cell and diploid central cell (not shown), forming the embryo and endosperm (respectively) in a new seed, in which *FLC* is re-expressed.

Other examples of imprinted genes are maize *fertilization-independent endosperm1* (*fie1*) and *fie2*, which show monoallelic expression from the maternal allele during endosperm development. This is reflected by the promoters of the silent paternal alleles having differentially methylated regions (DMRs)[63,64]. Analysis of DMR methylation of *fie* alleles in sperm, egg and central cells showed interesting differences in the mechanism for imprinting *fie1* and *fie2* (ref. 64). The DMR of *fie1* is heavily methylated in all three cell types, but the maternal alleles in the central cell (which contribute to the endosperm) become specifically demethylated, resembling the imprinting mechanism described for *A. thaliana FWA*[64]. By contrast, the DMR of *fie2* is unmethylated in all gametes, although the paternal allele becomes methylated *de novo* in the endosperm. Furthermore, the *fie2* DMR also showed extensive non-CG methylation, which is consistent with a DRM2-type-mediated RNA-directed DNA methylation process[64]. A further instance of potential gene regulation by *de novo* DNA methylation is provided by the *Brassica rapa SP11* locus, which encodes a pollen self-incompatibility determinant[65]. The *B. rapa* self-incompatibility phenotype is controlled by dominance relationships between *S*-haplotypes, and recessive *SP11* alleles were found to be specifically methylated *de novo* and silenced in the anther tapetal tissues[65]. It will be interesting to determine the prevalence of such instances of tissue-specific gene regulation by DNA methylation.

In addition to the gametophytic tissues being an important location for the establishment of imprinted gene expression, they also maintain pre-existing patterns of cytosine methylation. Evidence that silencing is important during gametophytic generation is provided by null *met1* alleles in *A. thaliana*, which produce hypomethylated epialleles even when the individual is heterozygous for the null allele[40]. This is caused by loss of cytosine methylation in the gametophytes of *met1* mutants, a loss that is greater when *met1* is inherited through the female gametophyte than the male[40]. This difference is probably accounted for by the female gametophyte (that is, the embryo sac) undergoing one more postmeiotic round of DNA replication before fertilization than the male gametophyte (that is, the pollen)[40].

A different epigenetic system used to developmentally silence genes during plant life cycles involves Polycomb group (PcG) proteins[66]. A conserved complex known as Polycomb repressive complex 2 (PRC2) functions to maintain patterns of gene repression in both plants and animals, using H3K27 methylation[66] (see page 425). However, in plants, there are several PRC2 complexes, with overlapping subunit compositions, specialized for distinct developmental roles[66]. For example, the PcG proteins have an important role in the regulation of imprinted gene expression. *A. thaliana MEA*, which is a homologue of *Drosophila melanogaster Enhancer of zeste*, shows maternally imprinted expression[67]. An important component of *MEA* imprinting is repression of the paternal *MEA* allele in the endosperm, and this process has been found to involve *MEA* autoregulation, using H3K27 trimethylation[55,68,69]. Interestingly, the mammalian PcG protein EED (embryonic ectoderm development) has also been shown to have an important role in the control of imprinted gene expression[70].

Another well-understood example of PcG-protein-mediated regulation in plants involves silencing of the floral-repressor gene *FLOWERING LOCUS C* (*FLC*) during the vernalization response in *A. thaliana*[71–73] (Fig. 4). Expression of *FLC*, which encodes a MADS-box-containing transcription factor, delays flowering and can be silenced by exposure of the plant to long periods of cold (that is, vernalization)[71–73]. In nature, this cold treatment occurs in winter and leads to flowering in favourable spring conditions. After the cold signal has been removed, *FLC* silencing is stable[71–73]. Mutations in the *VERNALIZATION 2* (*VRN2*) gene, which encodes a homologue of the *D. melanogaster* PcG protein Suppressor of zeste 12, cause late flowering after vernalization as a result of high levels of *FLC* expression[72]. Interestingly, *vrn2* mutants can silence *FLC* expression during the cold but fail to maintain this repression after the cold signal has been removed[72]. VRN2 is also required for acquisition of H3K27 dimethylation and trimethylation at *FLC* during vernalization, consistent with the known functions of PRC2 in maintaining patterns of gene repression[71,73,74].

The mechanism by which the vernalization-specific PcG-protein complex is recruited to *FLC* is not well understood but is known to require the PHD-finger-domain-containing protein VERNALIZATION INSENSITIVE 3 (VIN3)[73]. Because *VIN3* expression is induced after cold treatment, this protein might be a component of the signalling pathway that recruits PcG-protein-mediated repression to *FLC*[73] (Fig. 4). Recently, the *A. thaliana* homologue of *D. melanogaster* Heterochromatin protein 1 (HP1) — LIKE HETEROCHROMATIN PROTEIN 1 (LHP1; also known as TFL2) — was found to be required for the maintenance of *FLC* silencing after vernalization[75,76]. LHP1 becomes associated with the silenced *FLC* locus, a process that depends on an intronic sequence element[76]. The role of LHP1 in the repression of PcG-protein-regulated genes differs markedly from the main role of animal HP1 in heterochromatic silencing (see page 399). The DNA-binding protein VRN1 is also required for the maintenance of *FLC* silencing and associates with mitotic chromosomes[75,77]. Interestingly, VRN1 is absent from meiotic chromosomes of developing pollen[75]. One speculation is that this absence is associated with the resetting of *FLC* expression, which leads to a requirement for vernalization, at the start of each generation. Indeed, all PcG-protein-mediated silencing might be reset at some point during meiosis or gametogenesis, through an unknown mechanism (Fig. 4).

Conclusions

Plants continue to be excellent systems for the study of epigenetics, and their silencing mechanisms have marked similarities to those of mammals. An advantage of using plants is that they are tolerant of genome stresses, such as large losses of DNA methylation and changes in chromosome number. The elegant genetic tools available for organisms such as maize and *A. thaliana* are facilitating the dissection of epigenetic control. Recent advances such as the development of whole-genome microarrays and high-throughput sequencing are allowing the generation of large-scale data sets for epigenetic modifications and small RNAs that are extending our view to a genome-wide scale. Together, these approaches should enable major advances in our understanding of epigenetics to be made using plant systems: for example, how specific chromatin modifications are established and maintained, how they influence one another, and the extent to which they are used throughout the genome. This work should provide important insight for fields as diverse as cancer biology, development and evolution. ■

1. Gregory, T. R. The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. *Ann. Bot. (Lond.)* **95,** 133–146 (2005).
2. Hall, I. M. & Grewal, S. I. in *RNAi: A Guide to Gene Silencing* (ed. Hannon, G. J.) 205–232 (Cold Spring Harbor Laboratory Press, Woodbury, 2003).
3. Bernstein, B. E., Meissner, A. & Lander, E. S. The epigenome. *Cell* **128,** 669–681 (2006).
4. Bernard, P. *et al.* Requirement of heterochromatin for cohesion at centromeres. *Science* **294,** 2539–2542 (2001).
5. Bejerano, G. *et al.* A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441,** 87–90 (2006).
6. Liu, J., He, Y., Amasino, R. & Chen, X. siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in *Arabidopsis*. *Genes Dev.* **18,** 2873–2878 (2004).
7. Comfort, N. C. From controlling elements to transposons: Barbara McClintock and the Nobel Prize. *Trends Biochem. Sci.* **26,** 454–457 (2001).
8. Chandler, V. L. & Stam, M. Chromatin conversations: mechanisms and implications of paramutation. *Nature Rev. Genet.* **5,** 532–544 (2004).
9. Hamilton, A. J. & Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286,** 950–952 (1999).
10. Wassenegger, M., Heimes, S., Riedel, L. & Sanger, H. L. RNA-directed *de novo* methylation of genomic sequences in plants. *Cell* **76,** 567–576 (1994).
11. Zhang, X. *et al.* Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell* **126,** 1189–1201 (2006).
12. Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T. & Henikoff, S. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genet.* **39,** 61–69 (2007).
13. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
14. Fransz, P. F. *et al.* High-resolution physical mapping in *Arabidopsis thaliana* and tomato by fluorescence *in situ* hybridization to extended DNA fibres. *Plant J.* **9,** 421–430 (1996).
15. Lippman, Z. *et al.* Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430,** 471–476 (2004).
16. Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297,** 1833–1837 (2002).
17. Aufsatz, W., Mette, M. F., van der Winden, J., Matzke, A. J. & Matzke, M. RNA-directed DNA methylation in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **99** (suppl. 4), 16499–16506 (2002).

18. Mochizuki, K., Fine, N. A., Fujisawa, T. & Gorovsky, M. A. Analysis of a *piwi*-related gene implicates small RNAs in genome rearrangement in *Tetrahymena*. *Cell* **110,** 689–699 (2002).
19. Matzke, M., Matzke, A. J. & Kooter, J. M. RNA: guiding gene silencing. *Science* **293,** 1080–1083 (2001).
20. Lu, C. *et al.* Elucidation of the small RNA component of the transcriptome. *Science* **309,** 1567–1569 (2005).
21. Cao, X. *et al.* Role of the DRM and CMT3 methyltransferases in RNA-directed DNA methylation. *Curr. Biol.* **13,** 2212–2217 (2003).
22. Cao, X. & Jacobsen, S. E. Role of the *Arabidopsis* DRM methyltransferases in *de novo* DNA methylation and gene silencing. *Curr. Biol.* **12,** 1138–1144 (2002).
23. Chan, S. W. *et al.* RNA silencing genes control *de novo* DNA methylation. *Science* **303,** 1336 (2004).
24. Zilberman, D. *et al.* Role of *Arabidopsis ARGONAUTE4* in RNA-directed DNA methylation triggered by inverted repeats. *Curr. Biol.* **14,** 1214–1220 (2004).
25. Henderson, I. R. *et al.* Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nature Genet.* **38,** 721–725 (2006).
26. Xie, Z. *et al.* Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol.* **2,** e104 (2004).
27. Chan, S. W. *et al.* RNAi, DRD1, and histone methylation actively target developmentally important non-CG DNA methylation in *Arabidopsis*. *PLoS Genet.* **2,** e83 (2006).
28. Li, C. F. *et al.* An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in *Arabidopsis thaliana*. *Cell* **126,** 93–106 (2006).
29. Pontes, O. *et al.* The *Arabidopsis* chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell* **126,** 79–92 (2006).
30. Qi, Y. *et al.* Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature* **443,** 1008–1012 (2006).
31. Zilberman, D., Cao, X. & Jacobsen, S. E. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299,** 716–719 (2003).
32. Herr, A. J., Jensen, M. B., Dalmay, T. & Baulcombe, D. C. RNA polymerase IV directs silencing of endogenous DNA. *Science* **308,** 118–120 (2005).
33. Kanno, T. *et al.* Atypical RNA polymerase subunits required for RNA-directed DNA methylation. *Nature Genet.* **37,** 761–765 (2005).
34. Onodera, Y. *et al.* Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120,** 613–622 (2005).
35. Pontier, D. *et al.* Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in *Arabidopsis*. *Genes Dev.* **19,** 2030–2040 (2005).
36. Kanno, T. *et al.* Involvement of putative SNF2 chromatin remodeling protein DRD1 in RNA-directed DNA methylation. *Curr. Biol.* **14,** 801–805 (2004).
37. Cao, X. & Jacobsen, S. E. Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes. *Proc. Natl Acad. Sci. USA* **99** (suppl. 4), 16491–16498 (2002).
38. Goll, M. G. & Bestor, T. H. Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.* **74,** 481–514 (2005).
39. Kankel, M. W. *et al. Arabidopsis* MET1 cytosine methyltransferase mutants. *Genetics* **163,** 1109–1122 (2003).
40. Saze, H., Mittelsten Scheid, O. & Paszkowski, J. Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis. *Nature Genet.* **34,** 65–69 (2003).
41. Jackson, J. P., Lindroth, A. M., Cao, X. & Jacobsen, S. E. Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature* **416,** 556–560 (2002).
42. Malagnac, F., Bartee, L. & Bender, J. An *Arabidopsis* SET domain protein required for maintenance but not establishment of DNA methylation. *EMBO J.* **21,** 6842–6852 (2002).
43. Jacobsen, S. E. & Meyerowitz, E. M. Hypermethylated *SUPERMAN* epigenetic alleles in *Arabidopsis*. *Science* **277,** 1100–1103 (1997).
44. Herman, H. *et al. Trans* allele methylation and paramutation-like effects in mice. *Nature Genet.* **34,** 199–202 (2003).
45. Stam, M. *et al.* The regulatory regions required for *B'* paramutation and expression are located far upstream of the maize *b1* transcribed sequences. *Genetics* **162,** 917–930 (2002).
46. Stam, M., Belele, C., Dorweiler, J. E. & Chandler, V. L. Differential chromatin structure within a tandem array 100 kb upstream of the maize *b1* locus is associated with paramutation. *Genes Dev.* **16,** 1906–1918 (2002).
47. Alleman, M. *et al.* An RNA-dependent RNA polymerase is required for paramutation in maize. *Nature* **442,** 295–298 (2006).
48. Woodhouse, M. R., Freeling, M. & Lisch, D. Initiation, establishment, and maintenance of heritable *MuDR* transposon silencing in maize are mediated by distinct factors. *PLoS Biol.* **4,** e339 (2006).
49. Chan, S. W.-L., Zhang, X., Bernatavichute, Y. V. & Jacobsen, S. E. Two-step recruitment of RNA-directed DNA methylation to tandem repeats. *PLoS Biol.* **4,** e363 (2006).
50. Lisch, D., Carey, C. C., Dorweiler, J. E. & Chandler, V. L. A mutation that prevents paramutation in maize also reverses *Mutator* transposon methylation and silencing. *Proc. Natl Acad. Sci. USA* **99,** 6130–6135 (2002).
51. Soppe, W. J. *et al.* The late flowering phenotype of *fwa* mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. *Mol. Cell* **6,** 791–802 (2000).
52. Gehring, M., Choi, Y. & Fischer, R. L. Imprinting and seed development. *Plant Cell* **16,** S203–S213 (2004).
53. Kinoshita, T. *et al.* One-way control of *FWA* imprinting in *Arabidopsis* endosperm by DNA methylation. *Science* **303,** 521–523 (2004).
54. Choi, Y. *et al.* DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in *Arabidopsis*. *Cell* **110,** 33–42 (2002).
55. Gehring, M. *et al.* DEMETER DNA glycosylase establishes *MEDEA* polycomb gene self-imprinting by allele-specific demethylation. *Cell* **124,** 495–506 (2006).
56. Morales-Ruiz, T. *et al.* DEMETER and REPRESSOR OF SILENCING 1 encode 5-methylcytosine DNA glycosylases. *Proc. Natl Acad. Sci.USA* **103,** 6853–6858 (2006).
57. Jullien, P. E., Kinoshita, T., Ohad, N. & Berger, F. Maintenance of DNA methylation during the *Arabidopsis* life cycle is essential for parental imprinting. *Plant Cell* **18,** 1360–1372 (2006).
58. Kinoshita, Y. *et al.* Control of *FWA* gene silencing in *Arabidopsis thaliana* by SINE-related direct repeats. *Plant J.* **49,** 38–45 (2007).
59. Gong, Z. *et al.* ROS1, a repressor of transcriptional gene silencing in *Arabidopsis*, encodes a DNA glycosylase/lyase. *Cell* **111,** 803–814 (2002).
60. Agius, F., Kapoor, A. & Zhu, J. K. Role of the *Arabidopsis* DNA glycosylase/lyase ROS1 in active DNA demethylation. *Proc. Natl Acad. Sci. USA* **103,** 11796–11801 (2006).
61. Barreto, G. *et al.* Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation. *Nature* **445,** 671–675 (2007).
62. Jost, J. P., Siegmann, M., Sun, L. & Leung, R. Mechanisms of DNA demethylation in chicken embryos. Purification and properties of a 5-methylcytosine-DNA glycosylase. *J. Biol. Chem.* **270,** 9734–9739 (1995).
63. Danilevskaya, O. N. *et al.* Duplicated *fie* genes in maize: expression pattern and imprinting suggest distinct functions. *Plant Cell* **15,** 425–438 (2003).
64. Gutierrez-Marcos, J. F. *et al.* Epigenetic asymmetry of imprinted genes in plant gametes. *Nature Genet.* **38,** 876–878 (2006).
65. Shiba, H. *et al.* Dominance relationships between self-incompatibility alleles controlled by DNA methylation. *Nature Genet.* **38,** 297–299 (2006).
66. Kohler, C. & Grossniklaus, U. Epigenetic inheritance of expression states in plant development: the role of Polycomb group proteins. *Curr. Opin. Cell Biol.* **14,** 773–779 (2002).
67. Kinoshita, T., Yadegari, R., Harada, J. J., Goldberg, R. B. & Fischer, R. L. Imprinting of the *MEDEA* polycomb gene in the *Arabidopsis* endosperm. *Plant Cell* **11,** 1945–1952 (1999).
68. Baroux, C., Gagliardini, V., Page, D. R. & Grossniklaus, U. Dynamic regulatory interactions of Polycomb group genes: *MEDEA* autoregulation is required for imprinted gene expression in *Arabidopsis*. *Genes Dev.* **20,** 1081–1086 (2006).
69. Jullien, P. E., Katz, A., Oliva, M., Ohad, N. & Berger, F. Polycomb group complexes self-regulate imprinting of the Polycomb group gene *MEDEA* in *Arabidopsis*. *Curr. Biol.* **16,** 486–492 (2006).
70. Mager, J., Montgomery, N. D., de Villena, F. P. & Magnuson, T. Genome imprinting regulated by the mouse Polycomb group protein Eed. *Nature Genet.* **33,** 502–507 (2003).
71. Bastow, R. *et al.* Vernalization requires epigenetic silencing of *FLC* by histone methylation. *Nature* **427,** 164–167 (2004).
72. Gendall, A. R., Levy, Y. Y., Wilson, A. & Dean, C. The *VERNALIZATION 2* gene mediates the epigenetic regulation of vernalization in *Arabidopsis*. *Cell* **107,** 525–535 (2001).
73. Sung, S. & Amasino, R. M. Vernalization in *Arabidopsis thaliana* is mediated by the PHD finger protein VIN3. *Nature* **427,** 159–164 (2004).
74. Sung, S., Schmitz, R. J. & Amasino, R. M. A PHD finger protein involved in both the vernalization and photoperiod pathways in *Arabidopsis*. *Genes Dev.* **20,** 3244–3248 (2006).
75. Mylne, J. S. *et al.* LHP1, the *Arabidopsis* homologue of HETEROCHROMATIN PROTEIN1, is required for epigenetic silencing of *FLC*. *Proc. Natl Acad. Sci. USA* **103,** 5012–5017 (2006).
76. Sung, S. *et al.* Epigenetic maintenance of the vernalized state in *Arabidopsis thaliana* requires LIKE HETEROCHROMATIN PROTEIN 1. *Nature Genet.* **38,** 706–710 (2006).
77. Levy, Y. Y., Mesnage, S., Mylne, J. S., Gendall, A. R. & Dean, C. Multiple roles of *Arabidopsis* VRN1 in vernalization and flowering time control. *Science* **297,** 243–246 (2002).

Acknowledgements We thank S. Chan, C. Fei Li, K. Niakan, M. Ong and all members of the Jacobsen laboratory for useful comments and discussion. We apologize to colleagues whose research we did not have space to discuss. I.R.H. was supported by a long-term fellowship from the European Molecular Biology Organization, a Special Fellow grant from The Leukemia & Lymphoma Society, and a grant from the National Institutes of Health. S.E.J is an investigator of the Howard Hughes Medical Institute.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to S.E.J. (jacobsen@ucla.edu).

Stability and flexibility of epigenetic gene regulation in mammalian development

Wolf Reik[1]

During development, cells start in a pluripotent state, from which they can differentiate into many cell types, and progressively develop a narrower potential. Their gene-expression programmes become more defined, restricted and, potentially, 'locked in'. Pluripotent stem cells express genes that encode a set of core transcription factors, while genes that are required later in development are repressed by histone marks, which confer short-term, and therefore flexible, epigenetic silencing. By contrast, the methylation of DNA confers long-term epigenetic silencing of particular sequences — transposons, imprinted genes and pluripotency-associated genes — in somatic cells. Long-term silencing can be reprogrammed by demethylation of DNA, and this process might involve DNA repair. It is not known whether any of the epigenetic marks has a primary role in determining cell and lineage commitment during development.

Development is, by definition, epigenetic. Differences in the programmes of gene expression that result in the development of different organs and tissues occur without changes to the sequence of our DNA (with one or two exceptions). There is nothing mysterious in this concept; subsets of the ~30,000 genes in our genome are active in different tissues and organs, depending on their regulation by different sets or combinations of transcription factors. This implies that if we were to take all of the transcription factors that activate genes in a liver cell and transfer them to a brain cell (while inactivating all brain-specific transcription factors), then the brain cell would turn into a liver cell.

A recent study provides tantalizing insight into this concept of epigenetic control of development. Takahashi and Yamanaka identified four transcriptional regulators that when expressed in fibroblasts, resulted in these cells being reprogrammed to become embryonic stem (ES)-like cells[1]. Extending this concept a little further, in somatic-cell nuclear transfer, the nucleus of a somatic cell from an adult individual is transplanted into an oocyte from which the nucleus has been removed, resulting in reprogramming of the adult nucleus and therefore successful development of the cloned animal.

Cloning, however, is inefficient, because most (if not all) cloned animals have epigenetic defects, particularly in DNA methylation. Therefore, our lack of understanding of how epigenetic marks are reprogrammed is a key obstacle to cloning[2]. Similarly, the reprogramming of fibroblasts to become ES-like cells is a rare event *in vitro*, and epigenetic defects such as lack of demethylation of the *Oct4* (also known as *Pou5f1*) promoter, affecting expression of the encoded transcription factor, have been noted in these ES-like cells[1].

These observations highlight that, in addition to transcription factors, changes in gene expression during development are accompanied or caused by epigenetic modifications[2–7], such as methylation of DNA at CpG sequences (in vertebrates[4,5]), modification of histone tails[6] and the presence of non-nucleosomal chromatin-associated proteins[7]. Therefore, as development and differentiation proceed, differentiated cells accumulate epigenetic marks that differ from those of pluripotent cells, and differentiated cells of different lineages also accumulate different marks.

In this review, I focus on the role of epigenetic regulation in development, particularly comparing the short-term flexibility of certain epigenetic marks (which can be removed before a cell divides or within very few cell divisions) with the long-term stability and heritability of other marks (which can be maintained for many divisions) (Fig. 1). During the early stages of development, genes that are required later in development are transiently held in a repressed state by histone modifications, which are highly flexible and easily reversed when expression of these genes is needed. During differentiation, genes that are crucial for pluripotency are silenced by histone modifications, as well as by DNA methylation. Some of these genes are also silent in mature germ cells, meaning that epigenetic marks probably need to be reversed rapidly after fertilization to allow re-expression of pluripotency-associated genes in the next generation. By contrast, long-term silencing of transposons and imprinted genes — which is based on DNA methylation — needs to be stably maintained from the gametes into the early embryo and the adult organism. Methylation of imprinted genes can only be erased in primordial germ cells (PGCs), the cells that ultimately give rise to the germ line. Probably because there is a requirement for both removing epigenetic marks and retaining epigenetic marks between generations, epigenetic information can sometimes be inherited across multiple generations. In this review, I address how the fascinating interplay between transcription factors and epigenetic factors is beginning to provide an explanation for how pluripotency and development are regulated.

Flexibility for developmental gene regulation

In this section, three issues are addressed. First, are differentiation-specific genes held in an epigenetically silenced manner in pluripotent cell types, in order to be activated later? And is the removal of epigenetic marks from these genes needed for their activation? Second, are pluripotency-associated genes epigenetically inactivated in differentiated cell types? This inactivation could, in principle, be irreversible, because somatic cell types are not required to give rise to pluripotent cells. One exception is the germ line, where reactivation of pluripotency-associated genes is needed at the initial stages of development; however, later, the silencing of these genes is essential for the differentiation of mature germ cells. And therefore, third, is the removal of 'permanent' silencing marks from the gametic genomes after fertilization crucial to activate essential genes, such as pluripotency-associated genes, early in development?

[1]Laboratory of Developmental Genetics and Imprinting, The Babraham Institute, Cambridge CB22 3AT, UK.

There is recent evidence for the first type of epigenetic regulation: that is, the temporary inactivation of differentiation-specific genes in pluripotent cell types (Fig. 2a). Genes that are required during development and differentiation — for example, those in the homeobox (*Hox*), distal-less homeobox (*Dlx*), paired box (*Pax*) and sine-oculis-related homeobox (*Six*) gene families — are held repressed in pluripotent ES cells by the Polycomb group (PcG)-protein repressive system in mice and humans. This system marks the histones associated with these genes by inducing methylation of the lysine residue at position 27 of histone H3 (H3K27)[8–10]. ES cells that lack EED (embryonic ectoderm development), a component of the PcG-protein repressive complex (PRC), have partly derepressed developmental genes and are prone to spontaneous differentiation[8,10]. Interestingly, some developmental genes are present within 'bivalent' chromatin regions, which contain both inactivating marks (methylated H3K27) and activating marks (methylated H3K4)[9,11]. This could indicate that after the repressive marks have been removed (when expression of the components of PRCs are downregulated during differentiation), these genes are automatically poised for transcriptional activation through the H3K4 methylation mark. It is important to note that epigenetic silencing by PRCs might be mitotically heritable (through an unknown mechanism)[7], but these marks could presumably be rapidly removed by enzymatic demethylation of H3K27 (by an unidentified demethylase)[12]. The H3K27 methylation mark occurs mostly outside the context of DNA methylation. In contrast to the terminal silencing achieved by DNA methylation (discussed later), developmental genes that are silenced by PRCs in pluripotent tissues require repressive marks to be rapidly and flexibly removed when differentiation begins. Strikingly, in cancer cells, the genes targeted by PRCs often become DNA methylated, which might result in a more permanent locking-in of a 'pluripotent' state in cancer stem cells[13].

The second type of epigenetic regulation to be considered is whether pluripotency-associated genes are epigenetically inactivated in differentiated cell types. Several genes that are required for early development or for germ-cell development only — for example, those that encode pluripotency-sustaining transcription factors (such as OCT4 and NANOG) — are known to be expressed by ES cells but silenced on the differentiation of these cells, with a defined kinetics of acquiring repressive histone modifications and DNA methylation[14] (Fig. 2b). Silencing by both histone modifications and DNA methylation in somatic tissues seems to be typical of this group of genes and of those that encode cancer–testis antigens, which are expressed during spermatogenesis[15]. It is probable that this permanent type of epigenetic silencing safeguards against accidental expression of these genes in differentiated cells, because that might lead to dedifferentiation and, perhaps, to a predisposition to

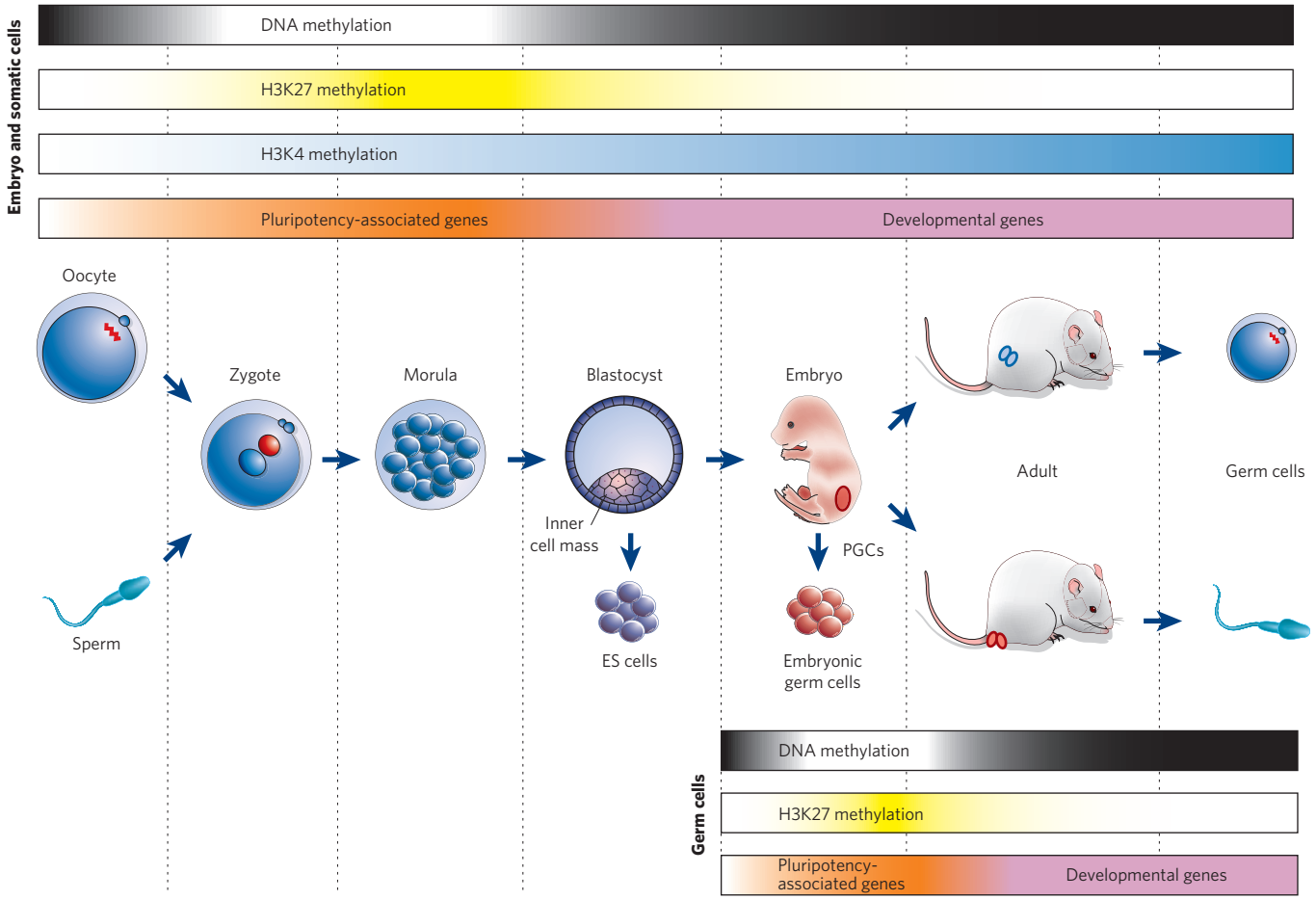


Figure 1 | Epigenetic gene regulation during mammalian development. Key developmental events are shown together with global epigenetic modifications and gene-expression patterns. Very early in development, DNA methylation is erased. In addition, pluripotency-associated genes begin to be expressed, and developmental genes are repressed by the PcG protein system and H3K27 methylation. During the differentiation of pluripotent cells such as ES cells, pluripotency-associated genes are repressed, potentially permanently, as a result of DNA methylation. At the same time, developmental genes begin to be expressed, and there is an increase in H3K4 methylation. During the early development of PGCs, DNA methylation and repressive histone modifications (such as H3K9 methylation) are also erased. Pluripotency-associated genes are re-expressed during a time window that allows embryonic germ cells to be derived in culture. Imprinted genes are demethylated during this period, and developmental genes are expressed afterwards. Flexible histone marks such as H3K27 methylation enable developmental genes to be silenced for a short time in pluripotent cells. By contrast, DNA methylation enables the stable silencing of imprinted genes, transposons and some pluripotency-associated genes.

cancer[16]. Consequently, these genes are difficult to reactivate in cloned embryos because of inefficient reprogramming of repressive marks, particularly of DNA methylation[17].

Special epigenetic regulation needs to occur in PGCs developing in the early post-implantation embryo[18]. Because these cells emerge from cell types in the egg cylinder that are already on the way to lineage commitment and differentiation, the somatic gene-expression programme needs to be suppressed. One of the key regulators of this process is BLIMP1 (B-lymphocyte-induced maturation protein 1), which associates with the arginine methyltransferase PRMT5. PRMT5 might partly repress *Hox*-family genes and other somatic genes in PGCs[19] (Fig. 2c). Pluripotency-associated genes and genes that have later roles in germ-cell development can also be repressed by DNA methylation (Fig. 2b). So genes such as *Mvh* (also known as *Ddx4*), *Dazl* (deleted in azoospermia-like) and *Sycp3* (synaptonemal complex protein 3) are methylated in early PGCs and begin to be expressed after the erasure of DNA methylation[20], which occurs between embryonic day (E) 8.0 and E12.5 in PGCs. Interestingly, pluripotency-associated genes such as *Nanog* also begin to be reactivated at these stages, but it is not known whether this involves demethylation of DNA. PGCs at these stages have similar properties to pluripotent cells, including the ability to form embryonic germ cells in culture[21]. These studies are important because they are the first to show that in some developmental situations, removal of epigenetic marks (H3K27 methylation in the ES-cell study, and DNA methylation in the PGC study) could be crucial for the activation of developmental genes. Whether DNA methylation in PGCs is erased by an active or a passive mechanism is unclear (discussed later). The promoters of the genes that undergo 'developmental' demethylation (for example, *Mvh*, *Dazl* and *Sycp3*) contain CpG islands, as do the differentially methylated regions (DMRs) of imprinted genes, which also undergo demethylation at these stages of PGC development. I am not aware of any reports of demethylation of CpG islands during development other than in PGCs or in the zygote and pre-implantation embryo (discussed later). Methylation of CpG islands might only be removable under exceptional circumstances.

Some key pluripotency-associated genes (such as *Oct4* and *Nanog*) are epigenetically inactivated at later stages of gametogenesis and in the mature gametes, including by DNA methylation. Therefore, after

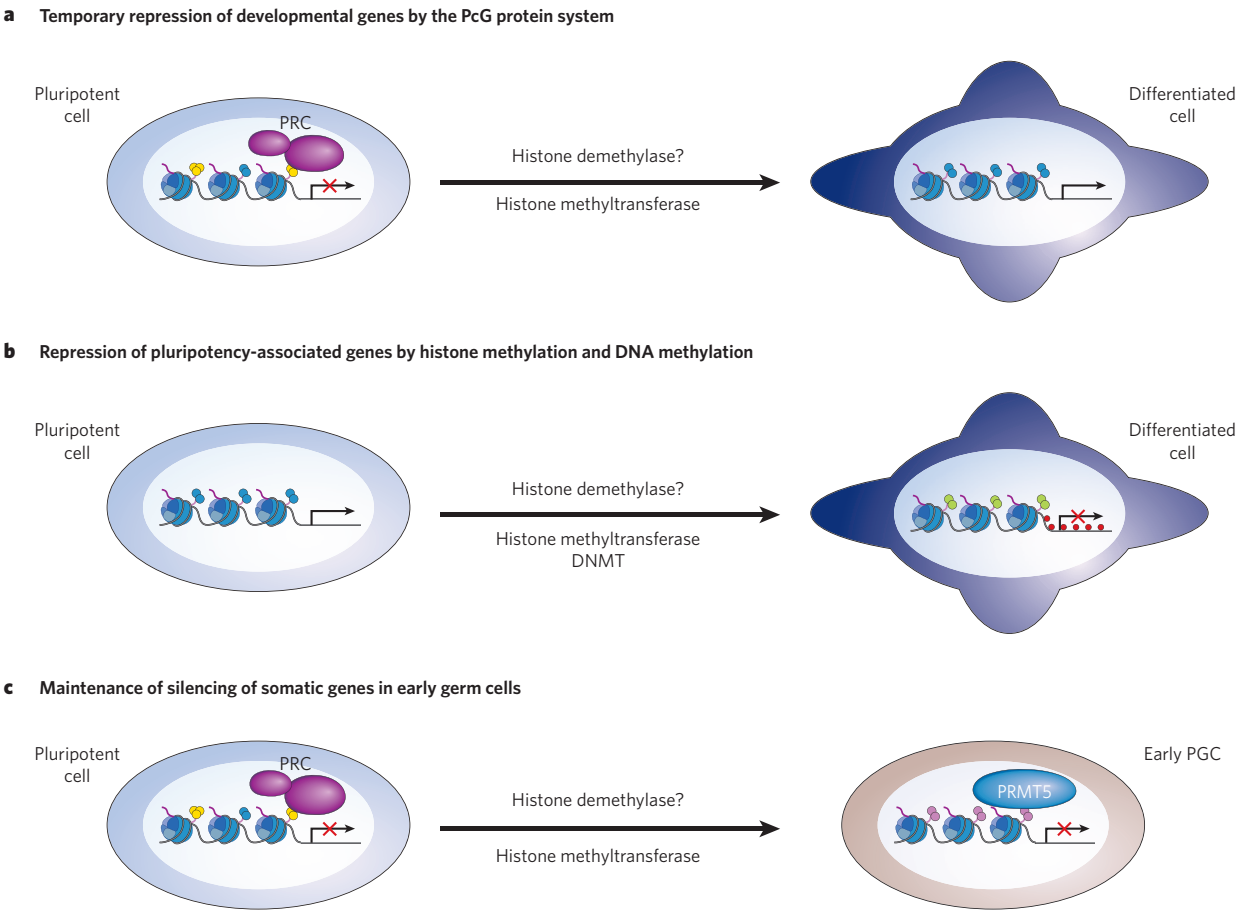


Figure 2 | Epigenetic regulation of pluripotency-associated genes and developmental genes during the differentiation of somatic cells and germ cells. The expression or repression of pluripotency-associated genes and developmental genes is indicated, and the associated modifications of the histone tails and/or DNA are represented by different colours. **a**, In pluripotent cells, the repression of genes that are needed later in development is flexible and can involve the PcG-protein repressive system. Silent developmental genes can be marked by both H3K27 methylation (yellow) and H3K4 methylation (blue), possibly allowing rapid gene activation after loss of repression by PcG-protein-containing repressive complexes (PRCs). Whether the loss of H3K27 methylation involves a histone demethylase is unknown. Further increases in H3K4 methylation might be required for proper developmental gene expression. **b**, Pluripotency-associated genes are stably silenced during differentiation, through histone methylation and DNA methylation. For example, genes such as *Oct4* and *Nanog* are silenced during ES-cell differentiation, and this process can involve both histone methylation (such as methylation of H3K9 mediated by G9A; also known as EHMT2) (green) and DNA methylation (red). Whether a histone demethylase is required for the removal of H3K4 methylation is unknown. **c**, For germ-cell development, the repression of somatic genes needs to be maintained in early germ cells, and this process might involve histone arginine methylation (pink). *Hox*-family genes and other developmental genes remain silent in early germ cells; some of this silencing might require histone arginine methylation brought about by PRMT5.

fertilization, the repressive epigenetic marks might need to be removed for transcriptional activation of these genes and correct early lineage development to take place (discussed later).

Stability for transposon silencing and imprinting

In contrast to developmental genes, which need to be epigenetically regulated with flexibility, transposons (if possible) need to be silenced completely and stably (at least from the perspective of the host) to prevent them from moving around in the genome and potentially causing mutations[22]. Therefore, many transposon families are both methylated themselves and marked by repressive histone modifications (such as H3K9 methylation), and these marks are important for the heritable silencing of transposons. Some transposon families (such as intracisternal A particles; IAPs) are also resistant to the erasure of DNA methylation in the zygote and in PGCs, possibly resulting in epigenetic inheritance across generations (discussed later).

Imprinted genes are a class of mammalian genes with possible mechanistic relationships to transposons[23], in that CpG islands in their promoters become methylated and in that silencing relies on long-term epigenetic stability. In imprinted genes (and transposons), DNA methylation is introduced during either oogenesis or spermatogenesis, by the *de novo* methyltransferase DNA methyltransferase 3A (DNMT3A) and its cofactor DNMT3-like (DNMT3L)[24,25] (Fig. 3a). How particular imprinted genes are selected for *de novo* methylation during oogenesis or spermatogenesis is not understood, although this targeting could involve pre-existing histone marks[26]. After fertilization, the methylation of imprinted-gene DMRs is maintained by DNMT1o (the oocyte form of DNMT1) for one division cycle during very early pre-implantation development[27] and then by DNMT1s (the somatic form of DNMT1) in embryonic and adult tissues[28].

Imprinted genes can be directly silenced by methylation of DMRs (which often contain CpG islands) that overlap the promoter. More frequently, however, imprinted genes occur in clusters, and there is usually a single DMR that is methylated in the germ line and is responsible for regulating gene silencing in the rest of the cluster. So far, there are two distinct models for how, after fertilization, imprinted genes are silenced through the action of nearby unmethylated DMRs. First, the DMR overlaps the promoter of a long, non-coding, unspliced, nuclear RNA[29,30]. The presence of the unmethylated and expressed copy of the non-coding RNA results in the silencing of linked genes, a process that involves repressive histone modifications[31,32]. It is unclear how the presence of the non-coding RNA leads to gene silencing *in cis*. In one model, repressive complexes (for example, PRCs) might be targeted during transcription[33]. Alternatively, the RNA might 'coat' the region to be inactivated, similarly to how *Xist* RNA (inactive X-specific transcripts) coats the inactive X chromosome[31,34]. This might establish a physical structure from which RNA polymerase II (Pol II) is excluded, resulting in transcriptional silencing[35] (Fig. 3b). In one case of silencing mediated by an imprinted non-coding RNA, the developmental kinetics of inactivation are markedly similar to those of imprinted X-chromosome inactivation. Both

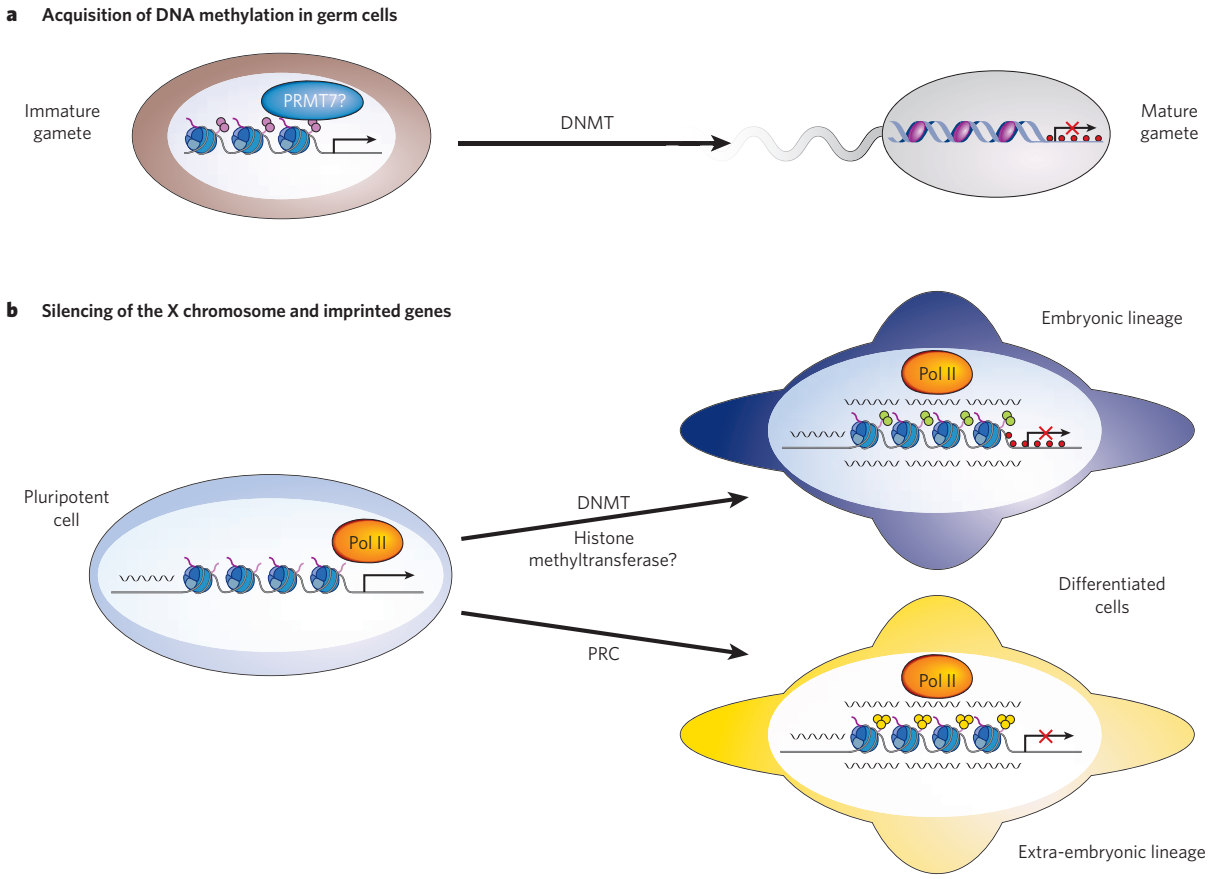


Figure 3 | Developmental regulation of imprinting and X-chromosome inactivation. a, During germ-cell development, selected imprinted genes and transposons become methylated. This process depends on *de novo* methyltransferases such as DNMT3A and its cofactor DNMT3L. It is possible that the targeting of DNA methylation requires arginine methylation of histones, carried out by PRMT7. Mature male germ cells have chromatin that is largely based on non-histone proteins known as protamines (dark pink); this alters the packaging of the DNA. **b**, Expression of non-coding RNAs (wavy black line) *in cis* can result in the silencing of adjacent genes as a consequence of the physical exclusion of Pol II and the acquisition of histone modifications and/or DNA methylation, depending on the embryonic lineage. DNA methylation stabilizes gene silencing in embryonic tissues but is less important in extra-embryonic tissues, where PRC-mediated silencing might predominate. This mechanism of postzygotic gene silencing occurs in X-chromosome inactivation and in some forms of autosomal gene imprinting. H3K9 methylation is shown in green, H3K27 methylation in yellow, histone arginine methylation in pink and DNA methylation in red.

non-coding RNAs (*Kcnq1ot1* and *Xist*) begin to be expressed from the paternal allele in the two-cell embryo, and gene silencing *in cis* and the acquisition of histone modifications follow during the next few cleavage divisions and are largely complete by the blastocyst stage[34] (Fig. 3b).

The second model of how imprinted genes are silenced involves an epigenetically regulated chromatin insulator. In this model, tissue-specific enhancers are located on one side of the DMR overlapping with the insulator, whereas the silenced genes are on the other side[36]. Silencing occurs when the DMR is unmethylated and binds chromatin-organizing proteins such as CTCF (CCCTC-binding factor), resulting in a higher-order chromatin structure that prevents interactions between remote enhancers and promoters[37].

X-chromosome inactivation is another example of a relatively stable epigenetic silencing event; in this case, large regions of a whole chromosome are involved. In mice, imprinted X-chromosome inactivation is probably largely initiated by expression of *Xist* from the paternal chromosome at the two-cell stage[38]. (The nature of the imprinting leading to paternal expression is still unknown, but it is unlikely to be DNA methylation.) Imprinted X-chromosome inactivation is then stable (even in the absence of DNA methylation[39]) in the extra-embryonic tissues. Although the PcG protein system (which confers H3K27 methylation marks) has some influence on gene silencing, these modifications do not seem to confer heritable silencing[40]. Random X-chromosome inactivation is initiated in the epiblast after reprogramming of imprinted inactivation[41,42]. This reprogramming might be initiated by the silencing of *Xist* expression, and if this is the case, it is possible that the mitotic 'memory' for inactivation simply resides in the expression of *Xist*. The subsequent upregulation of *Xist* expression during the differentiation of epiblast cells is again followed by coating, gene silencing and acquisition of histone marks[43]. However, in contrast to imprinted X-chromosome inactivation, CpG islands in inactivated genes on the X chromosome become methylated and, although it has not been tested genetically, this might constitute long-term memory for inactivation during embryonic and adult life[43] (Fig. 3b). It is important to note that this methylation of CpG islands seems to be a dead end in that it does not need to be reprogrammed during the normal life cycle. (In the germ line, the inactivated X chromosome does not become methylated.)

Breaking stability by epigenetic reprogramming

DNA-methylation patterns that have been acquired during development are stable in somatic cells and during adult life. DNA-methylation patterns are somatically heritable essentially through the action of DNMT1, the maintenance methyltransferase[44]. At most CpG sites, the error rate of maintaining methylation (~1% per division) is low in relation to the number of cell divisions that are needed to produce a mammalian organism (44 for humans). Indeed, methylation of CpG islands is never erased during normal development. By contrast, methylation of CpG islands in imprinted-gene DMRs needs to be erased in the germ line so that gender-specific methylation can be imposed subsequently, during germ-cell development. This erasure takes place in a defined period — from E10.5 to E12.5 in PGCs — in all imprinted genes that have been tested[45,46], and it could occur by active demethylation of DNA by an unknown mechanism, possibly involving DNA repair (discussed later). This mechanism for erasure might also underlie the demethylation and activation of non-imprinted genes such as *Mvh*, *Dazl* and *Sycp3,* which takes place at about the same stage[20] (Fig. 4a).

Epigenetic reprogramming in PGCs entails widespread loss of DNA methylation, as well as H3K9 methylation[47]. In addition to the erasure of genomic imprints, this epigenetic reprogramming might also help to return PGCs to a pluripotent state (because at these stages of PGC development, pluripotent embryonic germ cells can be established in culture), through the reactivation of genes such as *Nanog*. Not all genomic methylation is lost, however, at these stages; some transposons such as IAPs remain fairly highly methylated[48]. Later in oogenesis and spermatogenesis, *de novo* methylation occurs not only sex-specifically in imprinted genes but also in transposons and in single-copy gene sequences. For example, the *Nanog* promoter becomes highly methylated in mature sperm[49].

Distinct genome-wide reprogramming events also occur immediately after fertilization and during early pre-implantation development (Fig. 4b). Many sequences in the paternal genome become suddenly demethylated shortly after fertilization[50–53]. This demethylation occurs after the removal of protamines (basic proteins that are associated with DNA in sperm) and the acquisition of histones by the paternal genome during the long G1 phase, before DNA replication. Methylation can be observed by staining cells with an immunofluorescently labelled antibody specific for 5-methylcytosine. Judged by the substantial loss of immunofluorescence signal, together with the considerable loss of methylation of *Line1* elements as determined by bisulphite sequencing[48], the paternal genome loses a significant amount of methylation, although more precise measurements and more information about which sequences are affected and unaffected would be valuable. Sequences that are known not to be affected include IAPs and paternally methylated DMRs in imprinted genes (Fig. 4c). A recent study provides intriguing insight into a protein that might protect the genome from demethylation. The protein stella (also known as DPPA3) binds to DNA and was originally identified because expression of the encoding gene is upregulated during early PGC development. Stella is present in large amounts in oocytes and, after fertilization, translocates to both pronuclei. Deletion of the gene from the oocyte (and therefore removal of the protein from the zygote) results in early pre-implantation lethality of embryos, as well as loss of methylation of the following sequences: the maternally methylated genes *Peg1* (also known as *Mest*), *Peg5* (also known as *Nnat*) and *Peg10*; the paternally methylated genes *H19* and *Rasgrf1* (Ras protein-specific guanine-nucleotide-releasing factor 1); and IAPs[54]. So stella might, either directly or indirectly, protect specific sequences from demethylation in the zygote, but it is unknown how other sequences are protected (Fig. 4c).

The mechanism of active demethylation in the zygote is still unknown. However, the DNA deaminases AID and APOBEC1 have been shown *in vitro* to deaminate 5-methylcytosine in DNA to thymine[55]; this results in T•G mismatches, which can be repaired by the base-excision repair pathway. Interestingly, *Aid* and *Apobec1* are located in a cluster of genes with *Stella*, growth differentiation factor 3 (*Gdf3*) and *Nanog*. *Stella*, *Gdf3* and *Nanog* are all expressed in pluripotent tissues, and *Gdf3* and *Nanog* have important roles in conferring stem-cell identity on ES cells. Indeed, *Aid* and *Apobec1* are also expressed by oocytes, stem cells and germ cells[55], and recent work shows that *in vivo* targeting of AID to the methylated *H19* DMR in the zygote results in efficient and substantial demethylation of this region (C. F. Chan, H. Morgan, F. Santos, D. Lucifero, S. Petersen-Mahrt, W. Dean and W.R., unpublished observations). Although it is unclear whether AID and/or APOBEC1 are responsible for the demethylation of the paternal genome in the zygote, the evidence suggests that base-excision or mismatch repair might have a role in this process. I think that this suggestion is supported by the recent identification of a DNA glycosylase–lyase — DEMETER — that preferentially excises 5-methylcytosine from DNA in *Arabidopsis thaliana*[56,57]. DEMETER is required for the demethylation and activation of the imprinted gene *MEDEA* (see page 418). Another DNA-damage-responsive gene, the mouse gene *Gadd45* (growth arrest and DNA-damage-inducible 45), might also have a role in demethylation[58].

Although there have been suggestions that the methyl group could be directly removed from DNA by hydrolytic attack or by oxidation, these mechanisms have not been substantiated[2]. The relative flexibility of histone methylation might be brought about by the attachment of the methyl group through a carbon–nitrogen bond, together with the existence of enzymes that can directly remove the methyl group, leaving the rest of the histone molecule intact[12]. By contrast, the current evidence suggests that methyl groups attach through a carbon–carbon bond to the cytosine base and therefore might not be able to be directly removed, so demethylation inevitably has to proceed by pathways that involve base-excision or mismatch repair[55–57].

This active demethylation of the paternal genome is followed by passive demethylation of both maternal and paternal genomes, presumably brought about by exclusion of DNMT1o (the main form of DNMT1 present in the oocyte) from the nuclei of pre-implantation embryos[27]. Although DNMT1s can maintain the methylation of imprinted-gene DMRs during this period, total genome methylation decreases, reaching an overall low at the blastocyst stage. The purpose of active and passive demethylation during early embryogenesis is unknown. Demethylation of the paternal genome has been proposed to account for the paucity of paternal imprints[59] or to be a consequence of DNA-repair processes that are potentially involved in the protamine-to-histone transition[53]. General demethylation during this period could also have a role in returning the gametic genomes to pluripotency. For example, early expression of genes such as *Oct4* or *Nanog* is required for the establishment and maintenance of the inner-cell-mass lineage in the blastocyst[60]. Because the *Nanog* and *Oct4* promoters are methylated in sperm, and because methylation of these promoters is repressive, they need to be demethylated for proper expression to occur (Fig. 4b).

Epigenetic spillover across generations

Many of the epigenetic marks that are inherited and acquired by germ cells are therefore erased in PGCs and in early embryos, making way for new generations to develop and grow into adults purely on the basis of their genetic make-up. However, it also seems that epigenetic information can spill over to the next generation. The ability of somatic cells in the offspring to inherit the methylation of imprinted genes from parental germ cells is a mechanistic example of this (Fig. 4c). Another important example of spillover is inheritance of the epigenetic states conferred on some genes by adjacent insertion of IAPs. This can alter the expression of the endogenous genes; however, more importantly, the epigenetic state of the IAP (that is, methylated or unmethylated) regulates the expression of the nearby gene[61]. Because IAPs seem generally resistant to reprogramming during PGC and pre-implantation development, the state of expression of the genes that are regulated by IAP insertion can be inherited across several generations. It is interesting to note that there is an example of epigenetic inheritance being maternally transmitted but not paternally transmitted (the agouti viable yellow epiallele in mice), and the methylation of the IAP in the sperm is, unusually, erased in the zygote in this case[62]. So epigenetic inheritance is 'broken' by erasure of methylation of the paternal genome after fertilization.

There are other possible spillovers across generations. In *Caenorhabditis elegans*, the X chromosomes are epigenetically marked (by histone modifications) during gametogenesis, and some of these marks are maintained for several cell divisions in the new embryo (for an unknown reason)[63]. In mammalian embryos, some of the histone modifications acquired during the silencing of X-linked genes in spermatogenesis might be carried over into the zygote, leading to early silencing of some genes on the paternal X chromosome without the action of *Xist*[64].

One other area that is unique to mammalian biology deserves consideration with regard to epigenetic spillovers from the previous generation. At present, we have no understanding of how molecular decisions are taken to set up the first two cell lineages in the embryo: the trophectoderm and the inner cell mass[65]. However, a recent study suggests that differential histone arginine methylation of individual blastomeres, as early as the four-cell stage, could be one of the earliest marks for this lineage commitment[66]. There is much work to be done in this area, but it is an exciting possibility that the spillover of epigenetic marks from the gametes of parents might be responsible for setting up some of the earliest developmental decisions in the newly developing embryo.

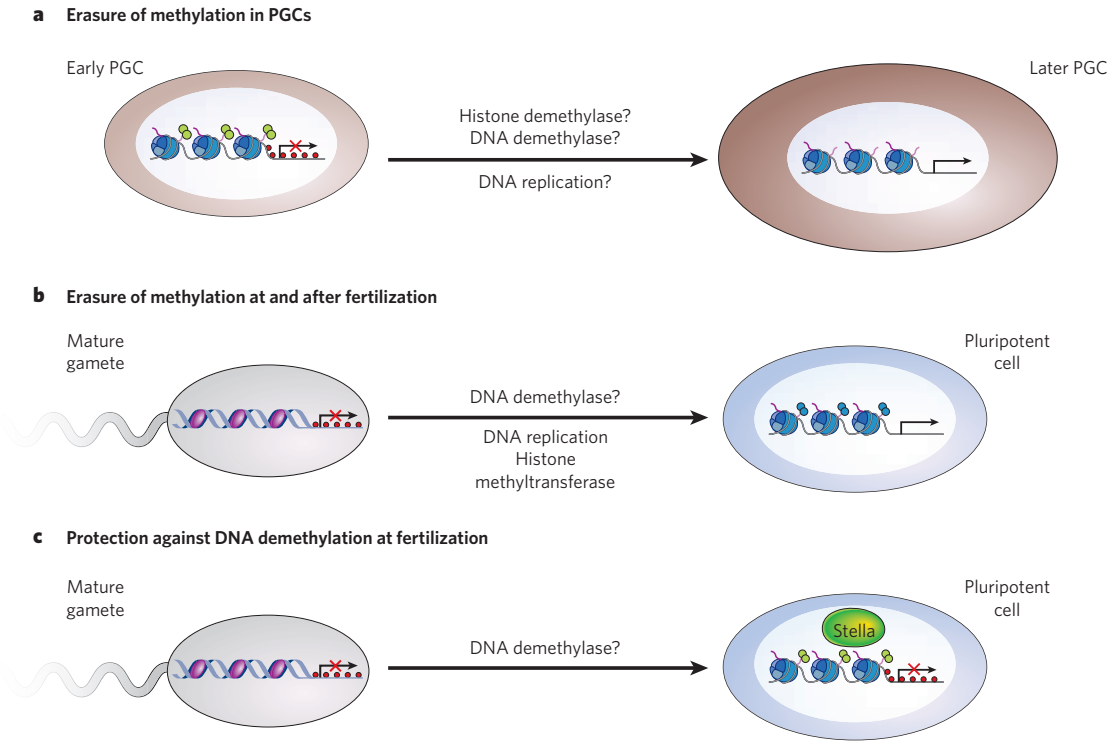


Figure 4 | Reprogramming of epigenetic marks in the germ line and the early embryo. a, During the development of PGCs, methylation of CpG islands in imprinted genes and other genes can be erased. This is a rapid process and might involve a demethylase or might occur by DNA replication without methylation being maintained. **b**, Many gene sequences that are methylated in mature gametes become demethylated in the early embryo. Some of this demethylation occurs in the absence of DNA replication and is therefore likely to be mediated by a demethylase. Demethylation might be important for the expression of pluripotency-associated genes. Active histone marks are also likely to be important for the expression of pluripotency-associated genes. **c**, In mature gametes, some DNA sequences that are methylated are protected from demethylation at or after fertilization. These sequences include imprinted genes and some transposons. The protein stella has recently been implicated in protection against demethylation at fertilization. H3K9 methylation is shown in green, H3K4 methylation in blue and DNA methylation in red.

Conclusions and outlook

Development might be a one-way street because of the somatic inheritance of epigenetic marks. Whether there is a linear relationship between acquisition of epigenetic marks and developmental progression is doubtful; some key restrictions in developmental potential that are brought about by epigenetic regulation might occur very early in development. Judging from somatic-cell nuclear-transfer experiments, it is far from clear whether more-differentiated cells have more epigenetic marks or have marks that are more difficult for the oocyte to reprogramme[67].

Natural epigenetic reprogramming might be needed to ensure that development can start afresh in every new generation. Although various mechanisms for the rapid erasure of histone modifications have recently been identified, the mechanism of DNA demethylation still needs to be determined. Recent work on the erasure of DNA methylation from imprinted plant genes shows that base-excision repair has an important role, and it is possible that this is also the case in mammals. Because of the generally accurate heritability of DNA methylation and because of its chemical stability, erasure of DNA methylation might only be possible either by replicating DNA in the absence of DNMT1 or by breaking DNA.

It is fascinating to see that both transcription-factor interactions and epigenetic programming and reprogramming seem to be needed to maintain pluripotency in early embryos and ES cells. Indeed, experimental reprogramming of differentiated nuclei without using somatic-cell nuclear transfer or cell fusion has been achieved recently, using a mix of pluripotency factors[1]. It could be expected that forcing the expression of pluripotency transcription-factor networks would also activate epigenetic reprogramming factors, but whether this occurs is unclear. Perhaps combinations of transcription factors and epigenetic reprogramming factors are needed for more complete reprogramming of somatic cells to a pluripotent state, and this would be of great fundamental scientific and medical interest.

In the animal kingdom, some epigenetic systems, such as imprinting, have evolved only in mammals. Many of the basic molecular building blocks for epigenetics, such as the enzymes for DNA methylation and histone modifications, are highly conserved in vertebrates, but the regulation of epigenetic modifiers might evolve more rapidly together with specific developmental strategies. Therefore, evolutionary epigenetics and epigenomics will have an important role in discovering links between developmental adaptations and epigenetic regulators.

There is probably a conflict between the requirement for erasing epigenetic marks between generations and the requirement for maintaining others, such as those in imprinted genes and in some transposons. This conflict most probably underlies the observation that some epigenetic marks are not erased between generations, thereby leading to multigenerational influences on inheritance and phenotype (see page 396). Epigenetic inheritance across generations is relatively common in plants, but it is still unclear how widespread this phenomenon is in mammals or whether it has any role in shaping evolution[61].

An exciting question for future work is whether segregation of epigenetic marks in early development has any primary role in determining cell and lineage commitment. For example, the mechanism by which the first two cell lineages are allocated in mammalian pre-implantation embryos, although a matter of hot debate, is not really understood. An epigenetic hypothesis might allow us to take a fresh look at a long-standing fundamental problem in developmental biology. ■

1. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).
2. Morgan, H. D., Santos, F., Green, K., Dean, W. & Reik, W. Epigenetic reprogramming in mammals. *Hum. Mol. Genet.* **14**, R47–R58 (2005).
3. Allis, C. D., Jenuwein, T. & Reinberg, D. (eds) *Epigenetics* (Cold Spring Harbor Laboratory Press, Woodbury, 2007).
4. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
5. Li, E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Rev. Genet.* **3**, 662–673 (2002).
6. Turner, B. M. Defining an epigenetic code. *Nature Cell Biol.* **9**, 2–6 (2007).
7. Ringrose, L. & Paro, R. Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu. Rev. Genet.* **38**, 413–443 (2004).
8. Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441,** 349–353 (2006).
9. Szutorisz, H. *et al.* Formation of an active tissue-specific chromatin domain initiated by epigenetic marking at the embryonic stem cell stage. *Mol. Cell. Biol.* **25**, 1804–1820 (2005).
10. Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nature Cell Biol.* **8**, 532–538 (2006).
11. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).
12. Klose, R. J., Kallin, E. M. & Zhang, Y. JmjC-domain-containing proteins and histone demethylation. *Nature Rev. Genet.* **7**, 715–727 (2006).
13. Ohm, J. E. *et al.* A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nature Genet.* **39**, 237–242 (2007).
14. Feldman, N. Y. *et al.* G9a-mediated irreversible epigenetic inactivation of *Oct-3/4* during early embryogenesis. *Nature Cell Biol.* **8**, 188–194 (2006).
15. Simpson, A. J., Caballero, O. L., Jungbluth, A., Chen, Y. T. & Old, L. J. Cancer/testis antigens, gametogenesis and cancer. *Nature Rev. Cancer* **5**, 615–625 (2005).
16. Hochedlinger, K., Yamada, Y., Beard, C. & Jaenisch, R. Ectopic expression of Oct-4 blocks progenitor-cell differentiation and causes dysplasia in epithelial tissues. *Cell* **121,** 465–477 (2005).
17. Boiani, M., Eckardt, S., Scholer, H. R. & McLaughlin, K. J. Oct4 distribution and level in mouse clones: consequences for pluripotency. *Genes Dev.* **16**, 1209–1219 (2002).
18. Surani, M. A., Hayashi, K. & Hajkova, P. Genetic and epigenetic regulators of pluripotency. *Cell* **128**, 747–762 (2007).
19. Ancelin, K. *et al.* Blimp1 associates with Prmt5 and directs histone arginine methylation in mouse germ cells. *Nature Cell Biol.* **8**, 623–630 (2006).
20. Maatouk, D. M. *et al.* DNA methylation is a primary mechanism for silencing postmigratory primordial germ cell genes in both germ cell and somatic cell lineages. *Development* **133**, 3411–3418 (2006).
21. Surani, A. & Reik, W. in *Epigenetics* (eds Allis, C. D., Jenuwein, T. & Reinberg, D.) 315–327 (Cold Spring Harbor Laboratory Press, Woodbury, 2007).
22. Bourc'his, D. & Bestor, T. H. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431,** 96–99 (2004).
23. Barlow, D. P. Methylation and imprinting: from host defense to gene regulation? *Science* **260**, 309–310 (1993).
24. Bourc'his, D., Xu, G. L., Lin, C. S., Bollman, B. & Bestor, T. H. Dnmt3L and the establishment of maternal genomic imprints. *Science* **294**, 2536–2539 (2001).
25. Kaneda, M. *et al.* Essential role for *de novo* DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature* **429**, 900–903 (2004).
26. Jelinic, P., Stehle, J. C. & Shaw, P. The testis-specific factor CTCFL cooperates with the protein methyltransferase PRMT7 in *H19* imprinting control region methylation. *PLoS Biol.* [online] **4**, e355 (2006) (doi:10.1371/journal.pbio.0040355).
27. Howell, C. Y. *et al.* Genomic imprinting disrupted by a maternal effect mutation in the *Dnmt1* gene. *Cell* **104**, 829–838 (2001).
28. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366**, 362–365 (1993).
29. Sleutels, F., Zwart, R. & Barlow, D. P. The non-coding *Air* RNA is required for silencing autosomal imprinted genes. *Nature* **415**, 810–813 (2002)
30. Mancini-Dinardo, D., Steele, S. J., Levorse, J. M., Ingram, R. S. & Tilghman, S. M. Elongation of the *Kcnq1ot1* transcript is required for genomic imprinting of neighboring genes. *Genes Dev.* **20**, 1268–1282 (2006).
31. Lewis, A. *et al.* Imprinting on distal chromosome 7 in the placenta involves repressive histone methylation independent of DNA methylation. *Nature Genet.* **36**, 1291–1295 (2004).
32. Umlauf, D. *et al.* Imprinting along the *Kcnq1* domain on mouse chromosome 7 involves repressive histone methylation and recruitment of Polycomb group complexes. *Nature Genet.* **36**, 1296–1300 (2004).
33. Kanduri, C., Thakur, N. & Pandey, R. R. The length of the transcript encoded from the *Kcnq1ot1* antisense promoter determines the degree of silencing. *EMBO J.* **25**, 2096–2106 (2006).
34. Lewis, A. *et al.* Epigenetic dynamics of the *Kcnq1* imprinted domain in the early embryo. *Development* **133**, 4203–4210 (2006).
35. Chaumeil, J., Le Baccon, P., Wutz, A. & Heard, E. A novel role for *Xist* RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev.* **20**, 2223–2227 (2006).
36. Verona, R. I., Mann, M. R. & Bartolomei, M. S. Genomic imprinting: intricacies of epigenetic regulation in clusters. *Annu. Rev. Cell Dev. Biol.***19**, 237–259 (2003).
37. Kurukuti, S. *et al.* CTCF binding at the *H19* imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to *Igf2*. *Proc. Natl Acad. Sci. USA* **103**, 10684–10689 (2006).
38. Okamoto, I. *et al.* Evidence for *de novo* imprinted X-chromosome inactivation independent of meiotic inactivation in mice. *Nature* **438**, 369–373 (2005).
39. Sado, T. *et al.* X inactivation in the mouse embryo deficient for *Dnmt1*: distinct effect of hypomethylation on imprinted and random X inactivation. *Dev. Biol.* **225**, 294–303 (2000).
40. Kohlmaier, A. *et al.* A chromosomal memory triggered by *Xist* regulates histone methylation in X inactivation. *PLoS Biol.* [online] **2,** e171 (2004) (doi:10.1371/journal.pbio.0020171).
41. Mak, W. *et al.* Reactivation of the paternal X chromosome in early mouse embryos. *Science* **303**, 666–669 (2004).
42. Okamoto, I., Otte, A. P., Allis, C. D., Reinberg, D. & Heard, E. Epigenetic dynamics of imprinted X inactivation during early mouse development. *Science* **303**, 644–649 (2004).
43. Heard, E. & Disteche, C. M. Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes Dev.* **20**, 1848–1867 (2006).
44. Goll, M. G. & Bestor, T. H. Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.* **74**, 481–514 (2005).
45. Hajkova, P. *et al.* Epigenetic reprogramming in mouse primordial germ cells. *Mech. Dev.* **117**, 15–23 (2002).
46. Lee, J. *et al.* Erasing genomic imprinting memory in mouse clone embryos produced from day 11.5 primordial germ cells. *Development* **129**, 1807–1817 (2002).
47. Seki, Y. *et al.* Extensive and orderly reprogramming of genome-wide chromatin modifications associated with specification and early development of germ cells in mice. *Dev. Biol.* **278**, 440–458 (2005).

48. Lane, N. *et al*. Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *Genesis* **35**, 88–93 (2003).
49. Imamura, M. *et al*. Transcriptional repression and DNA hypermethylation of a small set of ES cell marker genes in male germline stem cells. *BMC Dev. Biol.* [online] **6**, 34 (2006) (doi:10.1186/1471-213X-6-34).
50. Oswald, J. *et al*. Active demethylation of the paternal genome in the mouse zygote. *Curr. Biol.* **10**, 475–478 (2000).
51. Mayer, W., Niveleau, A., Walter, J., Fundele, R. & Haaf, T. Demethylation of the zygotic paternal genome. *Nature* **403**, 501–502 (2000).
52. Dean, W. *et al*. Conservation of methylation reprogramming in mammalian development: aberrant reprogramming in cloned embryos. *Proc. Natl Acad. Sci. USA* **98**, 13734–13738 (2001).
53. Santos, F., Hendrich, B., Reik, W. & Dean, W. Dynamic reprogramming of DNA methylation in the early mouse embryo. *Dev. Biol.* **241**, 172–182 (2002).
54. Nakamura, T. *et al*. PGC7/Stella protects against DNA demethylation in early embryogenesis. *Nature Cell Biol.* **9**, 64–71 (2006).
55. Morgan, H. D., Dean, W., Coker, H. A., Reik, W. & Petersen-Mahrt, S. K. Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. *J. Biol. Chem.* **279**, 52353–52360 (2004).
56. Gehring, M. *et al*. DEMETER DNA glycosylase establishes *MEDEA* Polycomb gene self-imprinting by allele-specific demethylation. *Cell* **124**, 495–506 (2006).
57. Morales-Ruiz, T. *et al. DEMETER* and *REPRESSOR OF SILENCING 1* encode 5-methylcytosine DNA glycosylases. *Proc. Natl Acad. Sci. USA* **103**, 6853–6858 (2006).
58. Barreto, G. *et al.* Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation. *Nature* **445**, 671–675 (2007).
59. Reik, W. & Walter, J. Evolution of imprinting mechanisms: the battle of the sexes begins in the zygote. *Nature Genet.* **27**, 255–256 (2001).
60. Smith, A. G. Embryo-derived stem cells: of mice and men. *Annu. Rev. Cell Dev. Biol.* **17**, 435–462 (2002).
61. Whitelaw, N. C. & Whitelaw, E. How lifetimes shape epigenotype within and across generations. *Hum. Mol. Genet.* **15**, R131–R137 (2006).
62. Blewitt, M. E., Vickaryous, N. K., Paldi, A., Koseki, H. & Whitelaw, E. Dynamic reprogramming of DNA methylation at an epigenetically sensitive allele in mice. *PLoS Genet.* [online] **2**, e49 (2006) (doi:10.1371/journal.pgen.0020049).
63. Bean, C. J., Schaner, C. E. & Kelly, W. G. Meiotic pairing and imprinted X chromatin assembly in *Caenorhabditis elegans*. *Nature Genet.* **36**, 100–105 (2004).
64. Namekawa, S. H. *et al*. Postmeiotic sex chromatin in the male germline of mice. *Curr. Biol.* **16**, 660–667 (2006).
65. Rossant, J. Lineage development and polar asymmetries in the peri-implantation mouse blastocyst. *Semin. Cell Dev. Biol.* **15**, 573–581 (2004).
66. Torres-Padilla, M.E., Parfitt, D.E., Kouzarides, T. & Zernicka-Goetz, M. Histone arginine methylation regulates pluripotency in the early mouse embryo. *Nature* **445**, 214–218 (2007).
67. Yang, X. *et al.* Nuclear reprogramming of cloned embryos and its implications for therapeutic cloning. *Nature Genet.* **39**, 295–302 (2007).

Acknowledgements I thank all my colleagues, past and present, for their contributions to the work and ideas described in this paper, especially W. Dean, F. Santos, A. Lewis, and G. Smits. Funding from the Biotechnology and Biological Sciences Research Council, the Medical Research Council, the European Union Epigenome Network of Excellence, CellCentric and the Department of Trade & Industry is gratefully acknowledged.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares competing financial interests: details accompany the paper at www.nature.com/nature. Correspondence should be addressed to the author (wolf.reik@bbsrc.ac.uk).

HISTONE VARIANTS MEET THEIR MATCH

Kavitha Sarma and Danny Reinberg

Abstract | A fascinating aspect of how chromatin structure impacts on gene expression and cellular identity is the transmission of information from mother to daughter cells, independently of the primary DNA sequence. This epigenetic information seems to be contained within the covalent modifications of histone polypeptides and the distinctive characteristics of variant histone subspecies. There are specific deposition pathways for some histone variants, which provide invaluable mechanistic insights into processes whereby the major histones are exchanged for their more specialized counterparts.

ORPHAN GENE
A protein-coding region that bears little or no homology to genes in distant species.

Howard Hughes Medical Institute, Division of Nucleic Acids Enzymology, Department of Biochemistry, Robert Wood Johnson Medical School, Piscataway, New Jersey 08854, USA. Correspondence to D.R. e-mail: reinbedf@umdnj.edu
doi:10.1038/nrm1567

The nucleosomes form the basic repeating units of chromatin in eukaryotes. The composition of the individual nucleosomes is fundamentally similar and consists of an octameric core of four types of histones — H2A, H2B, H3 and H4 — around which 147 bp of DNA is wrapped. Each octamer contains two copies of each histone. Modulation of the fundamental nucleosome units contributes to the dynamic structural characteristics of chromatin, which are heritable and impact on transcription and, therefore, cellular identity. This entails various post-translational modifications of the histone proteins and also the incorporation of variant histone subspecies. These variant or 'replacement' histones were discovered on the basis of the small — and sometimes even large — differences in their amino-acid sequence relative to the major histone species (FIG. 1).

The nucleus is characterized by distinct chromatin domains. The dynamics, maintenance and post-translational modifications in these domains have sparked intensive interest in the field of chromatin biology, and recent discoveries have helped elucidate their structural and functional regulation (BOX 1). Some of these specialized domains in chromatin are enriched for the specific histone variants, which operate with other factors to ensure the proper functioning of these domains. In this review, we highlight the roles of these variant histones, their modes of deposition by specific chaperones and also how these might relate to other known histone exchangers.

Histone deposition and exchange

The expression of the major histones is tightly regulated during the cell cycle, and the histones are deposited onto DNA in a process that is strictly coupled to DNA replication. However, histone variants are expressed from a set of genes known as ORPHAN GENES, which are not subject to this stringent regulation. These genes are expressed throughout the cell cycle and their products are deposited during, as well as after, the completion of S phase. These variants have evolved particular characteristics that impact on the transcriptional capacity of the nucleosomal regions they inhabit, some of which are described briefly in this review (for a comprehensive review, see REF. 1). Chromatin is further compacted by the incorporation of the linker histone H1, which has been reported to have eight isoforms in higher eukaryotes. This topic has been extensively covered in a forthcoming review by Kimmins and Sassone-Corsi[2], and we will therefore not address histone H1 or its variants.

Although previous studies have indicated low levels of histone exchange in the absence of transcription or replication[3], the first direct visual evidence was obtained by using cells that expressed green fluorescent protein (GFP)-tagged histones in conjunction with photobleaching of a small area of the nucleus. The recovery of fluorescence in these 'bleached' areas was scored for the level of histone mobility — in other words, histone exchange. These analyses indicated that histones are not readily replaced. In fact, histones H3 and H4 were

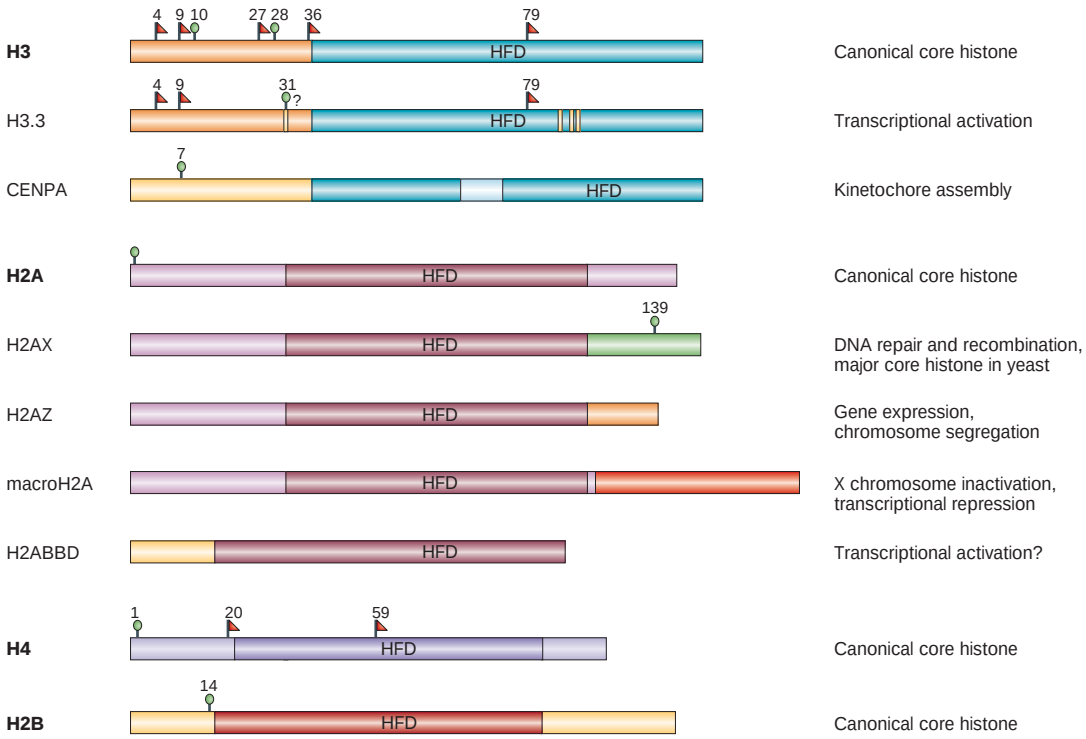


Figure 1 | **Canonical core histones and their variants.** The major core histones contain a conserved histone-fold domain (HFD). In addition, they contain N- and C-terminal tails that harbour sites for various post-translational modifications. For simplicity, only well-established sites for lysine methylation (red flags) and serine phosphorylation (green circles) are shown (other types of modifications, such as ubiquitylation, are not shown). In the histone H3.3 variant, the residues that differ from the major histone H3 (also known as H3.1) are highlighted in yellow. Three of these residues are contained in the globular domain and one resides in the N terminus. This N-terminal residue (Ser31) has been speculated to be a potential site for phosphorylation on H3.3. The centromeric histone CENPA has a unique N terminus, which does not resemble other core histones. Two sites of phosphorylation have been identified in this region, of which Ser7 phosphorylation has been shown to be essential for completion of cytokinesis. The region in the globular domain that is required for targeting CENPA to the centromere is highlighted in light blue. Histone H2A variants differ significantly from the major core H2A in their C terminus. The C terminus of H2AX harbours a conserved serine residue (Ser139), the phosphorylation of which is an early event in response to DNA double-strand breaks. A short region in the C terminus of H2AZ is essential for viability in *Drosophila melanogaster*. MacroH2A has an extended C-terminal macro domain, the function of which is unknown. Finally, the H2ABBD is the smallest of the H2A variants and contains a distinct N terminus, which lacks all of the conserved modification sites that are present in H2A. The C terminus is also truncated and lacks the docking domain that is found in other H2A species. The histones H4 and H2B are also shown, including their known methylation and phosphorylation sites. The proposed functions of the variants are listed.

found to show almost no recovery after photobleaching, whereas histones H2A and H2B showed slightly higher levels of exchange[4]. However, the linker histone H1 showed recovery within a few minutes, which indicated that it has a high rate of diffusion[5].

The popular model for nucleosome deposition onto DNA is that it is coupled to DNA replication and occurs in a stepwise manner, initiated by the deposition of the H3–H4 tetramer followed by the deposition of the H2A–H2B dimers[6,7]. The addition of post-translational modifications could occur before or after deposition. However, for DNA-replication-dependent deposition of the tetramer, the acetylation of lysines 5 and 12 of the H4 tail are thought to be necessary[8,9]. The significance of acetylation at these residues is debatable as they are not required for chromatin assembly in budding yeast, and deletion of H3 and H4 N-terminal tails does not compromise the chromatin-assembly factor-1 (CAF1)-dependent assembly of these histones onto DNA *in vitro*[10,11]. Nonetheless, Kadonaga and co-workers isolated a DNA-replication-dependent histone-deposition complex — known as replication-coupling assembly factor (RCAF) — which included the anti-silencing factor-1 (ASF1; see below) and CAF1, as well as the histones H3 and H4 with appropriate acetylated residues[12].

Whether other histone modifications that are important in defining 'chromatin domains' are established before or subsequent to histone deposition is not clear. However, owing to the disparate expression pattern of enzymes that incorporate stable histone marks during the cell cycle, such as lysine methylation, some of these modifications are likely to be incorporated during mitosis and, therefore, post-replication[13,14]. Other complexes that contain HISTONE LYSINE METHYLTRANSFERASE activity seem to be expressed in normal cells during S phase, and so a

HISTONE LYSINE METHYLTRANSFERASE
An enzyme that catalyses the transfer of methyl groups onto the ε-amino residue of lysines in histones.

Box 1 | Definition of chromatin domains

Historically, chromatin domains have been broadly classified into two forms, euchromatin and heterochromatin.

Euchromatin
This is the region of chromatin that is decondensed and is thought to represent loci that are transcriptionally active. Genes in this region replicate early, the chromatin contains hyperacetylated histones and stains poorly in the nucleus. Active genes in this region are enriched for methylation at Lys4 of histone H3 (H3-K4), H3-K36 and H3-K79.

Heterochromatin
This is highly compacted chromatin with regions of silenced DNA. It replicates late, contains hypoacetylated histones and high levels of DNA methylation. Heterochromatin is further classified into pericentric or constitutive heterochromatin and facultative heterochromatin.

Pericentric or constitutive heterochromatin
This is the region that is juxtaposed to centromeres on the chromosome and contains large blocks of ALPHA SATELLITE REPEATS in humans (known as major satellite repeats in mice). This region contains H3 tri-methyl K9 and mono-methyl K27, and H4 tri-methyl K20 (REFS 84–86). As the name suggests, it is irreversibly silenced and remains so throughout the cycles of cell division.

Facultative heterochromatin
This type of heterochromatin has the ability or 'faculty' to become transcriptionally active again. A classic example of facultative heterochromatin is the inactive X chromosome in mammals, which is characterized by the presence of H3 tri-methyl K27 and di-methyl K9, and H4 mono-methyl K20 (REFS 87–90). Silenced euchromatic genes contain methylated H3-K9, H3 tri-methyl K27 and H4 mono-methyl K20 (REF. 85).

methylation mark at a specific histone residue (for example, at Lys27 of H3; H3-K27) is probably incorporated during DNA replication[15]. Whether this modification is incorporated into the histone before deposition onto DNA or immediately after deposition is at present unknown.

The correct incorporation of histones onto DNA requires the assistance of additional factors. This function is fulfilled by HISTONE CHAPERONES. Chaperones are thought to function in coordination with CHROMATIN-REMODELLING FACTORS to mediate the accurate positioning of nucleosomes on a DNA template. Several histone chaperones have been identified and characterized both biochemically and genetically (for details, see REF. 16). Recently, some chaperones were shown to function in the deposition of specific histone variants. Although histones H2B and H4 were once thought to be invariable, recent studies have identified two testis-specific H2B variants in humans, the functions of which have yet to be determined[17,18]. However, no H4 variant has been reported so far.

Next, we discuss the different histone H3 and H2A variants and their modes of deposition, in particular H3.3 and H2AZ, as these have been most well studied.

H3 variants

Four different isoforms of histone H3 have been reported: H3.1, H3.2 and H3.3, which are similar, and CENPA, the centromeric histone H3, which shows a wide variability in amino-acid composition between species and even within the same species when compared with the other H3 isoforms (FIG. 1; BOX 2). Of these H3 variants, H3.3 and CENPA have been studied the most intensively and have been found to carry out distinct functions.

H3.3. Recent studies showed that the variant histone H3.3 is present at transcriptionally active loci[19]. Deposition of the GFP-tagged variant H3.3 was observed during DNA-replication-coupled (RC) processes as well as in a DNA-replication-independent (RI) manner. During the RC phase, GFP–H3.3 was found throughout the genome; however, during RI assembly, GFP–H3.3 was found predominantly at rDNA arrays, which indicates incorporation at sites of active transcription. On the other hand, cells transfected with the GFP-tagged core histone H3 (H3.1) and exposed to the S-phase inhibitor aphidicolin were inhibited for GFP–H3.1 deposition onto DNA. Moreover, artificial expression of H3.1 outside S phase did not result in its incorporation into chromatin, which showed that deposition of H3.1 is tightly coupled to DNA replication. Importantly, even though H3.1 and H3.3 differ by only four amino acids (FIG. 1), these three residues in the globular domain are crucial for their distinctive deposition during the cell cycle. The H3.1-specific residues apparently impede its assembly outside S phase. Mutation of any of the H3.1-specific residues to the corresponding residue in H3.3 led to both RC and partial RI deposition.

Further clues regarding the function of H3.3 were derived from the identification of its post-translational modifications. This variant is enriched for the presence of 'marks' that reflect transcriptional competence, such as di- and tri-methylation of Lys4, acetylation at Lys9, Lys18 and Lys23, and methylation at K79. Of note, although H3.3 is present at lower levels in dividing cells, on terminal differentiation, the level of H3.3 increases significantly and contributes to more than half of the total amount of H3 protein in the cell[20]. This confirms again that H3.3 is deposited at all stages of the cell cycle, whereas H3.1 incorporation is restricted to the S phase. So, in differentiated cells, the increased level of H3.3 does not correlate with the amount of transcriptional activity.

Insight into the mechanisms by which histone H3.1 and H3.3 are deposited onto DNA has come from biochemical studies. The purification of H3.1- and H3.3-containing complexes from stable cell lines that contain tagged H3-species revealed that these variants associate with different chromatin-assembly complexes *in vivo*[21]. The H3.1 complex was associated with CAF1, whereas the H3.3-containing complex was associated with the histone chaperone HIRA. Interestingly, both complexes contained the HISTONE ACETYLTRANSFERASE HAT1, which supported previous evidence that histones are transiently acetylated before deposition onto DNA. This also indicated that the complexes represented the pre-deposited forms of histones. Furthermore, both complexes included histone H4 and ASF1, a histone chaperone that was initially shown to promote deposition of histones in a DNA-replication-dependent manner *in vitro*[12]. However, recent *in vivo* studies suggest that deletion of ASF1 in

ALPHA SATELLITE REPEAT
Large highly repetitive stretches of (A+T)-rich DNA sequences in the human genome that are usually untranscribed.

HISTONE CHAPERONE
A protein that escorts histones to DNA for deposition.

CHROMATIN-REMODELLING FACTOR
A protein that alters the dynamic organization of nucleosomes to help in the activation or repression of gene expression.

HISTONE ACETYLTRANSFERASE
An enzyme that catalyses the addition of an acetyl group to specific lysine residues in histones.

yeast leads to the formation of more compact chromatin, indicating that ASF1 functions in the disassembly rather than the assembly of chromatin[22]. Whether the H4 histone in the H3.1 complex contained the acetylation pattern that is important for DNA-replication-dependent deposition was not analysed.

CAF1 is one of the most well-studied chaperones and consists of three subunits: p150, p60 and p48. CAF1 interacts with the proliferating cell nuclear antigen (PCNA), a multifunctional protein complex that partakes in various functions such as DNA replication, repair, cell-cycle regulation and chromatin assembly. During the process of DNA synthesis, it functions as a sliding clamp around the replication fork and stimulates the processivity of DNA polymerases δ and ε[23]. PCNA is thought to recruit CAF1 for replication-dependent assembly of nucleosomes[24,25]. CAF1 was now found to specifically interact with H3.1 *in vitro* and to deposit H3.1 onto DNA in a DNA-repair-synthesis-dependent manner. In the case of HIRA, its role as a chaperone was first uncovered in *Xenopus laevis* extracts, which on HIRA depletion were able to support RC but not RI histone deposition. This deficiency was restored when HIRA and H3–H4 tetramers were added[26]. HIRA deposits histone H3.3 independently of DNA synthesis; whether it does so during DNA synthesis as well, is not yet clear.

Importantly, the isolated complexes contained the tagged H3 isoform and histone H4, but were devoid of H2A and H2B. Most importantly, these complexes, which were isolated on the basis of the tag that is present on the H3 isoforms, were also devoid of endogenous H3, despite the fact that endogenous H3 was present in the extracts at levels much greater than those for the tagged H3 isoforms. This latter finding was interpreted as indicating that each complex contained one dimer of H3–H4. Importantly, mononucleosomes that were isolated from these stable cell lines contained endogenous and tagged versions of either H3.1 or the H3.3 variant, but not both isoform types, which suggested that the nucleosomes that formed through these pathways are homogeneous in their H3 composition.

These observations raise at least two important questions. First, under the assay conditions, are the H3–H4 histone polypeptides deposited as tetramers, as previous studies that were carried out *in vivo* have shown, or as dimers? Second, if dimers are deposited, how is the second copy of the H3–H4 dimer (for example, the endogenous one) brought to the tagged H3–H4-dimer–DNA complex? A possibility is that the ASF1 chaperone, which is common to both H3.1- and H3.3-containing complexes, deposits the 'respective' endogenous H3 isoform together with H4. Previous observations support this possibility by demonstrating that ASF1 exists in a RCAF complex, which includes H3 and H4 and that also synergizes with CAF1 (REFS 12,27). These studies show that RCAF or CAF1 alone assemble chromatin inefficiently, but that together, their deposition abilities are stimulated. ASF1 might function in localizing an H3–H4 dimer and subsequently deposit the second H3–H4 dimer. However, as ASF1 was found in both the H3.1 and H3.3 complexes, the important question that remains is how the specificity for the respective H3 isoform (H3.1 or H3.3) is attained?

A more general question that remains is whether there are different pathways for nucleosome deposition during DNA replication and during DNA-repair synthesis. The authors scored for *de novo* nucleosome assembly on naked DNA either in the absence of DNA synthesis (HIRA/H3.3 competent) or in the presence of DNA-repair synthesis (CAF1/H3.1 competent)[21]. Do these complexes have similar roles on DNA templates that are assembled with histones? Another important consideration is that even though the process of DNA synthesis during DNA repair is similar to DNA replication, these processes are mechanistically different. During DNA replication, the DNA strands are separated and the nucleosomes are segregated. This is not the case in DNA-repair synthesis, which encompasses just a 20–30 nucleotide stretch. The nucleosome(s) in this process are probably altered or relocated to facilitate access to the DNA-repair machinery. But even if some nucleosomes are evicted, the integrity of others close to the site of repair need not be jeopardized. When coupled to this nucleosome 'alteration' at the site of DNA damage, the H3.3 and H3.1 complexes might then mediate histone deposition or exchange. With regard to DNA replication, the role of the H3.1 complex might be clarified following its isolation during the S phase of the cell cycle.

Whether the H3–H4 tetramer is displaced during DNA replication or half of the tetramer is displaced, leaving behind the other H3–H4 dimer, is a topic that warrants more investigation. It bears on whether H3–H4 is then deposited or replaced in the form of a tetramer or a dimer during this process. The issue of dimer versus tetramer deposition is an important one in the context of epigenetics. How is the information that is contained within the mononucleosomes, with respect to histone isoform type and histone modifications, which reflect active and inactive chromatin regions, retained when histones are deposited during DNA replication? If, during the process of DNA replication, an H3–H4 dimer remains behind on the template DNA and another must be re-deposited, given the results of Tagami *et al.* — who showed the homogeneity of histone H3.3 or H3.1 isoforms within nucleosomes — the remaining dimer might determine the isoform that is brought in[21]. Modifications that are contained on the remaining dimer might be copied to the new dimer, by as-yet-unknown means. On the other hand, if it is the tetramer that is displaced during DNA replication, how then is the original information safeguarded and restored? The mechanism of nucleosome segregation during DNA replication needs to be revisited, as nucleosome deposition and histone exchange during DNA synthesis (or DNA-repair synthesis), transcription (see below) and DNA replication seem to be more different than was first thought.

CENPA. The centromeric histone CENPA was found to be a histone-H3 variant on the basis of its tendency to co-purify with the other core histones[28]. Sequence

Box 2 | Histone exchange — why bother?

Several post-translational modifications have been identified both at the N-terminal tails and the globular domains of histones. These include acetylation, phosphorylation, methylation, polyADP ribosylation and monoubiquitylation[91]. Of these 'marks', methylation has been shown to occur on both lysine and arginine residues. The function of these methyl marks on histones has been extensively investigated recently[92–94]. As several methyl marks have been implicated in the regulation of gene expression, it follows that the addition and removal of these modifications must also be controlled. This stems from the need to return the gene to its original state before the stimulatory or repressive signal. This can be done in two ways. The first is an enzymatic reaction that catalyses the removal of the methyl group. The second is the removal of the entire histone molecule. Unlike acetylation and phosphorylation, both of which have been shown to be dynamic marks that are subject to reversal by deacetylases and phosphatases, respectively, lysine methylation has been found to be relatively stable; so far, no enzymes have been found that 'demethylate' modified lysine residues that are involved in repression (H3-K9, H3-K27 and H4-K20). However, recently, the H3-K4 methyl mark, which is involved in activation, has been shown to be demethylated by the enzyme LSD1 (previously known as p110 or BHC110)[97]. On the other hand, arginine methylation was recently shown to be reversible. This is the consequence of a deimination reaction that converts both unmethylated and mono-methylated but not di-methylated arginines of histone H3 and H4 to citrulline. The enzyme peptidyl arginine deiminase-4 (PADI4 or PAD4) catalyses this reaction, and this in turn antagonizes the activity of the histone arginine methyltransferase CARM1 (REFS 95,96). Whether deimination by PAD4 occurs *in vivo* as a secondary step after the removal of a single methyl moiety from di-methylated arginines remains to be tested.

A histone-exchange reaction would have a dual function *in vivo*. First, it would remove all epigenetic marks on histones and facilitate reprogramming of the gene in question. Second, it would allow for the incorporation of replacement histones that have evolved to carry out diverse functions in cells. The removal of stable epigenetic marks poses a paradox, as such marks are presumably transmitted to the daughter cell. However, it is interesting to note that the lysines that are reported as methylated in H3.1 are completely conserved in variant H3.3, which indicates that H3.3 could be subject to the same modifications. This is also the case for H2A, H2AX and H2AZ. Most of the methylated lysines are conserved, except for one in H2AZ in which there is a shift in position by one amino acid. This means that the integrity of the epigenetic programme need not be perturbed on histone exchange with variant species.

analysis revealed that it shared a similar C-terminal histone-fold domain with H3, but varied extensively in its N-terminal region (FIG. 1). CENPA is localized exclusively to centromeres, but when overexpressed, CENPA spreads along the chromosome arms[29]. Although not much is known about how CENPA is targeted to centromeres, it is known to carry out an essential function(s), as a homozygous knockout of this gene in mice results in lethality[30]. Domain-swap analysis between H3.1 and CENPA has shown that the highly conserved histone-fold domain, but not the variant N terminus, is essential for targeting to the centromeres[29]. This is similar to the findings with H3.3 in that the amino acids within the conserved histone-fold motif have an important role in its localization to chromatin domains.

Interestingly, in fission yeast, the histone chaperones Caf1 and Hira have also been shown to be associated with centromeric chromatin, and deletion of both gives rise to an altered centromeric structure[31]. The localization of Cse4 (the yeast centromeric H3, which is homologous to CENPA) was also affected. In the absence of Caf1 and Hira, although Cse4 was found to localize to centromeric DNA, it was also distributed in non-centromeric regions. This indicates that even though these histone chaperones are not essential for the deposition of Cse4 to centromeric regions, they might be required for imparting specificity to its localization. Whether the same Hira- or Caf1-containing complexes that function in restricting the spreading of centromeric CENPA also function in chromatin assembly during DNA synthesis and repair (see above) remains to be determined, as at least three different complexes that contain Hira and Caf1 were separated during glycerol-gradient sedimentation[21].

Recently, Wieland and colleagues showed that the yeast centromeric histone Cse4 could functionally complement human CENPA. In this study, the phenotype that is induced by depletion of CENPA by RNA interference (RNAi) could be complemented by the ectopic expression of the yeast Cse4 (REF. 32). Structural studies have provided new insights into the mechanism of targeting CENPA to centromeres. CENPA that is complexed with H4 forms a more rigid subnucleosomal structure compared with the H3–H4 tetramer, and this results in reduced solvent accessibility for histone H4 (REF. 33). It is also interesting to note that the globular region of CENPA that varies from H3 (consisting of loop 1 and the α2 helix) is conserved across species, and this could account for the complementation of human CENPA by yeast Cse4. In addition, when these regions were replaced with the corresponding H3 regions, targeting to the centromere was abrogated. This is reminiscent of the situation whereby amino-acid substitutions in the H3.1 to the corresponding amino acids in H3.3 in the globular domain confer RI deposition[19].

H2A variants

Four H2A variants have been reported so far — H2AX, H2AZ, macroH2A and H2A-bar-body-deficient (H2ABBD). These variants function in diverse cellular pathways, some of which are discussed below.

H2AZ. H2AZ (for which the yeast homologue is Htz1) is the most studied H2A variant with respect to function.

In mammals, it is encoded by an essential gene, as homozygous knockout of the gene in mice resulted in embryos that failed to develop beyond gastrulation[34]. The C-terminus of H2AZ contains a short region that is essential for development beyond the larval stages in *Drosophila melanogaster*[35]. In *Tetrahymena thermophila*, H2AZ was found exclusively in the transcriptionally active MACRONUCLEUS and was expressed in the silent MICRONUCLEUS only during CONJUGATION before gene activation[36]. This gave the first indication that H2AZ might be involved in the activation of gene expression. Studies in yeast indicated that Htz1, the yeast orthologue of H2AZ, localized within actively transcribed regions, particularly those that flank heterochromatin that is associated with the Sir silencing complex; specifically, telomeric chromatin, MATING-TYPE LOCUS and rDNA[37]. The Sir silencing complex comprises the Sir2–4 proteins. Sir2 is an NAD^+-dependent histone deacetylase, which preferentially targets H4-K16 *in vivo*, and Sir3 and Sir4 in this complex bind to the H3 and H4 N-terminal tails of deacetylated nucleosomes. The spread of heterochromatin that is mediated by the Sir proteins is thought to be through the deacetylation of H4-K16 by Sir2, followed by the binding of Sir3 and Sir4 to these hypoacetylated regions[38]. As the silenced regions spread in the absence of Htz1 *in vivo*, Hzt1 was postulated to be a participant in the maintenance or establishment of the boundary between heterochromatic and euchromatic regions.

The crystal structure of the H2AZ-containing nucleosome suggests that two H2AZ molecules are preferred over one copy of H2AZ and H2A[39]. Once again, homogeneity of a variant form in the nucleosome is preferred. Previously, H2AZ was shown to localize to pericentric heterochromatin in early mouse development[40]. Depletion of H2AZ by targeted disruption or RNAi resulted in early mouse embryonic lethality[34]. More recently, knockdown of H2AZ levels by RNAi in mammalian cell lines led to chromosomal missegregation and disruption of the normal distribution of the heterochromatin-specific protein HP1α[41]. This indicates that, in higher eukaryotes, H2AZ is also involved in confining HP1α to specific regions and, therefore, in the maintenance of facultative heterochromatin (see BOX 1). Taken together, the reason for chromosome missegregation could be attributed to incomplete chromatin condensation or heterochromatin formation in mitosis. As previous studies have shown a putative role for H2AZ in transcriptional activation or prevention of the spread of repressive chromatin, the pertinent question now arises: how does H2AZ find its way to these euchromatic regions?

Three independent studies revealed the existence of a novel complex that is required for H2AZ deposition in yeast[42–44]. The complex contains, among other polypeptides, Swr1 (a member of the ATP-dependent SWI/SNF family of chromatin-remodelling factors), H2AZ and H2B. The complex was also found to contain Bdf1, a protein containing BROMODOMAINS that interacts with acetylated histone H4. Bdf1 also associates with the yeast TFIID complex and shows similarity to the C-terminal region of human TAF1 (REF. 45).

The H2AZ deposition complex was discovered using divergent approaches and while pursuing diverse goals. Wu and co-workers initiated studies to investigate the role of a lesser-known member of the SWI/SNF family, Swr1. Through purification of a tagged Swr1, they isolated a Swr1 complex that contained H2AZ[42]. Greenblatt and colleagues used a genetic screen to isolate synthetic mutations that function in conjunction with mutants in chromatin-modifying factors and RNA polymerase II transcription elongation, followed by a proteomic approach. This led to the characterization of a complex that contained Swr1 and Bdf1, among other polypeptides. The presence of H2AZ in this complex was discovered when it co-immunoprecipitated with Bdf1 (REF. 43). Rine's group specifically looked for complexes that contain H2AZ and that might facilitate its deposition — they discovered a similar complex, which they called Swr1-Com[44]. The presence of Bdf1 in the Swr1 complex was postulated as a possible candidate to facilitate Swr1/H2AZ localization to active chromatin regions.

The role of the Swr1 complex in H2AZ exchange was directly broached by biochemical studies from the Wu laboratory. The Swr1 complex contains several polypeptides, some of which are also present in the INO80 complex, as well as some polypeptides that are unique to each. Both the Swr1 and INO80 members of the SWI/SNF family differ from the other family members in having a split ATPase domain[46]. When Wu's laboratory discovered that the Swr1 complex contained H2AZ, they used tagged H2AZ to purify two H2AZ-containing complexes that also contained H2B; the Swr1 complex, thereby confirming the integrity of the original association, and a second complex that contained the nucleosome-assembly protein-1 (Nap1) chaperone. The Nap1-containing complex, designated NAP-Z, contains at least five other polypeptides that are yet to be described. In the presence of ATP, the Swr1 complex transferred H2AZ–H2B dimers that were free or those that were associated with Nap1 to immobilized nucleosomes. As free dimers were also exchanged, the chaperone characteristics of Nap1 remain unresolved; Nap1 might function as a source of dimers without escorting them to the chromatin. Nonetheless, these findings clearly established the precedent for a nucleosome-remodelling complex, such as Swr1, to function also in ATP-dependent histone exchange.

All three studies demonstrate the dependency of H2AZ on Swr1 for its deposition onto chromatin *in vivo*. The presence of H2AZ at previously identified regions was significantly reduced in both an H2AZ-deletion mutant strain and, correspondingly, in a Swr1-deletion strain, as evidenced by CHROMATIN IMMUNOPRECIPITATION (ChIP) analyses[42–44]. This substantiates the requirement of Swr1 in H2AZ deposition. Regions that are deficient in H2AZ show significant overlap with reduced gene expression using whole-genome microarrays. However, there were some surprising findings. H2AZ localized to the polyadenylation and 3′ mRNA cleavage sites at higher levels than was seen at the promoters for the *ADH1*, *PMA1* and *GAL1* genes when probed in one study[43]. Also, H2AZ was more prevalent at the *GAL1*

MACRONUCLEUS
The larger of the two nuclei in the unicellular ciliate *Tetrahymena thermophila*. This is the somatic nucleus and is transcriptionally active.

MICRONUCLEUS
The smaller 'germline' nucleus in *Tetrahymena thermophila*, which is transcriptionally silent.

CONJUGATION
A process of sexual reproduction that occurs in some unicellular organisms and that involves the exchange of genetic material between two cells through a so-called sex pilus.

MATING-TYPE LOCUS
The genomic region in yeast that determines the mating type or 'sex' of the haploid yeast cell.

BROMODOMAIN
An evolutionarily conserved domain that has been shown to bind to acetylated residues.

CHROMATIN IMMUNOPRECIPITATION
(ChIP). A technique by which direct or indirect protein–DNA interactions in chromatin can be studied using antibodies against specific chromosomal proteins.

promoter when repressed. H2AZ was found not only throughout the region near the telomere, as expected, but also overlapped with silenced telomeric regions of chromosome V[43]. So, H2AZ probably does not function alone in establishing the heterochromatic–euchromatic boundary in this region.

So, what does this mean in the context of yeast and higher eukaryotes? The genome in yeast is less complex than in humans, and the finding that H2AZ in yeast localized at large distances from the telomeres strongly indicates a role other than that of impeding the spread of heterochromatin. The genome in higher eukaryotes is more complex as it contains blocks of 'junk' DNA. So, the equivalent Swr1 complex might function in the exchange of H2A for H2AZ to allow a nucleosome environment that would favour the action of chromatin-remodelling factors to facilitate transcription. It has been shown that the incorporation of H2AZ into chromatin stabilizes the octamer within the nucleosome, but impedes oligomerization (and therefore condensation) of chromatin fibres[47,48]. Nakatani and co-workers (personal communication) have isolated a similar Swr1 complex from human cells that contains substoichiometric amounts of FACT ('facilitates chromatin transcription'), a complex that disassembles or reassembles chromatin during RNA polymerase II transit[49,50]. This, in conjunction with the findings from H2AZ knockdown by RNAi, portrays a more global role for the Swr1 homologue in human cells.

H2AX. H2AX is a histone variant in higher eukaryotes, which, although absent in nematodes, is the 'normal' histone H2A in budding yeast[51]. As pointed out in an excellent review by Malik and Henikoff [1], the copy number of the gene seems to correlate directly with the extent of homologous recombination in the organism. For example, yeast, which has high levels of homologous recombination, only has H2AX and not H2A, and nematodes that seem to have little homologous recombination lack H2AX altogether. Similarly, in humans and flies, the copy number of the gene that encodes H2AX is low, which correlates with the low levels of homologous recombination. On the other hand, *T. thermophila* has high levels of homologous recombination and a large number of genes that encode H2AX (approximately similar to the number of genes encoding H2A)[1].

The yeast H2A and the higher eukaryote H2AX histones contain an extension at the C-terminus, which includes the conserved amino-acid sequence SQ(E/D)ϕ (where ϕ denotes a hydrophobic residue). Ser139 in this unique C-terminal region is phosphorylated in response to DNA double-strand breaks (DSBs) and seems to be an early step in the response to DNA damage[52]. Whereas the main kinase that phosphorylates Ser139 is thought to be ATM (ataxia-telangiectasia mutated), the DNA-dependent protein kinase (DNA-PK) has a redundant function in this event[53]. In *D. melanogaster*, the function of H2AX has been taken over by a chimeric molecule that contains the H2AZ globular domain coupled to the C-terminal H2AX tail[54]. The means by which H2AX is targeted to DSBs is unknown at present, but it is thought that H2AX is randomly deposited and that it is phosphorylated around DSBs (see below). The deletion of H2AX in mice, although not lethal, caused a reduction in the number of irradiation-induced foci (IRIF) and resulted in genomic instability and male infertility[55]. It has recently been shown that although DNA-repair factors are recruited to sites of DNA damage in H2AX-deficient cells, their retention is transient and they fail to form IRIF. This was also the case when H2AX-deficient cells were stably transformed with an H2AX form carrying a mutation at its phosphorylation site[56]. So, phosphorylation of H2AX seems to be essential for the formation of efficient repair foci in cells. A function for H2AX that is independent of phosphorylation was observed in male meiosis. H2AX is required for the condensation of the mouse X and Y chromosome pair and for their maintenance in a silenced state during meiosis[57].

Apoptotic DNA damage was found to promote the phosphorylation of another histone, H2B, at Ser14. This phosphorylation has a broad distribution pattern in nuclei[58]. In a recent study, H2B Ser14 was found to be phosphorylated in the absence of H2AX, but its localization at DSBs was compromised in the absence of H2AX Ser139 phosphorylation[59]. As the N terminus of H2B is required for chromosome condensation[60], one possibility to explain this observation is the existence of an interplay between the modifications in the H2AX C-terminus and the H2B N-terminus. Compaction would have two consequences. First, chromosomes would be unable to separate until the DNA-repair process was complete, and second, the concentration of repair factors around the lesion would increase to promote efficient repair. The latter would be possible if H2AX is distributed at regular intervals in the genome to monitor the integrity of chromosomes, thereby functioning as the 'histone guardian of the genome'[61]. Whether phosphorylation of H2AX and H2B can occur together on the same nucleosome remains to be tested.

MacroH2A. MacroH2A is a vertebrate-specific variant, which has two distinct domains — the N-terminus, which is similar to H2A, and a large C-terminus, which has no similarity to other histones[62]. MacroH2A is enriched on the inactive X (Xi) chromosome in mammalian female cells[63]. Although this variant is a hallmark of X INACTIVATION, its presence is not essential for maintenance of the inactivated state. Its deposition occurs after localization of the inactive-X-specific transcript, *Xist*, on the Xi[64]. In the absence of the *Xist* transcript, macroH2A cannot localize to the Xi. This suggests that an RNA molecule might be involved in promoting histone exchange. MacroH2A in undifferentiated embryonic stem (ES) cells (that is, before X inactivation) is concentrated at the centrosomes of the nucleus, where it is tethered by microtubules[65]. At the onset of differentiation, macroH2A shows reorganization with enrichment on the inactive X chromosome. MacroH2A has a general role in silencing, as evidenced by findings that the C-terminal 'macro' domain inhibits the binding of transcription factors and that the N-terminal H2A domain

X INACTIVATION
The process whereby one of the two copies of the X chromosome in female mammals is silenced to compensate for the presence of a single copy in males.

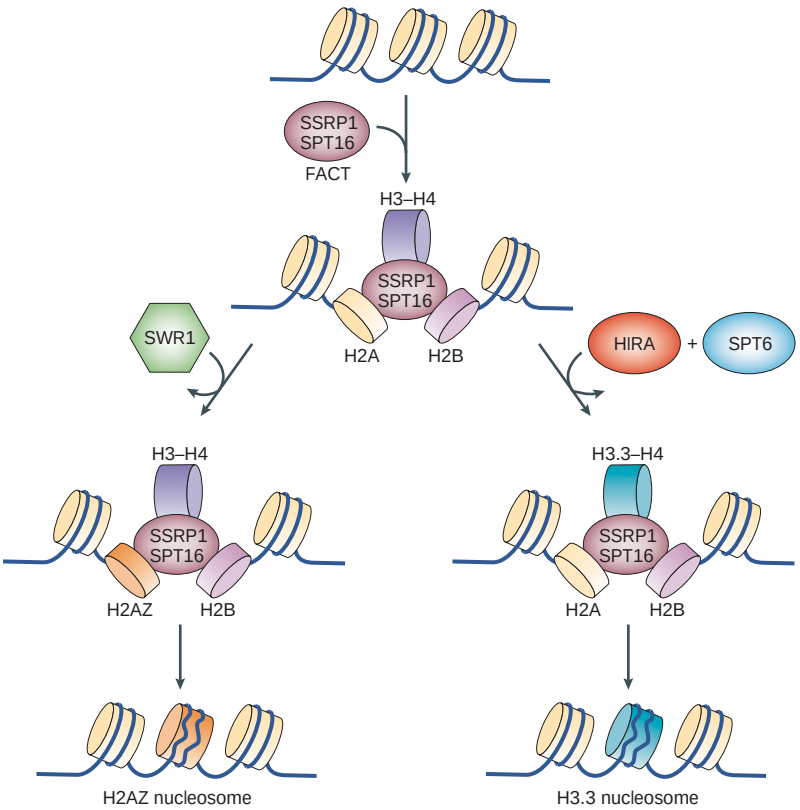


Figure 2 | **Synergism between SWR1, HIRA and FACT.** The major histones can be replaced by their variants to allow for a more transcriptionally competent chromatin state. Here, we show a model for the synergy between the H2AZ and H3.3 exchange complexes with FACT ('facilitates chromatin transcription'), which disassembles and reassembles chromatin during transcription. In one situation, displacement of an H2A–H2B dimer by the SPT16 subunit of FACT could allow exchange of the displaced H2A with H2AZ by SWR1 (a member of the ATP-dependent SWI/SNF family of chromatin-remodelling factors), which leads to an altered nucleosome that is homogeneous in its composition of H2AZ (see main text). In a second situation, the SSRP1 subunit of FACT could coordinate with the elongation factor SPT6 and the histone chaperone HIRA to replace H3 with H3.3. Both of these events would result in the formation of chromatin that is more amenable to transcription, either on the basis of the intrinsic structure of the variant nucleosome or by the presence of post-translational modifications on the variant histones.

interferes with the activity of nucleosome-remodelling factors[66]. The C-terminal of macroH2A contains a LEUCINE-ZIPPER MOTIF that has been implicated in protein dimerization. Such dimerization in macroH2A-containing nucleosomes might facilitate inter-nucleosome interactions, thereby promoting the compaction of large chromatin domains.

H2ABBD. H2ABBD is the most recently isolated H2A variant and little is known about it, with the following exceptions. It is excluded from the Xi chromosome in mammalian cells and colocalizes with H4 that is acetylated at Lys12, which is indicative of a euchromatic function[67]. Stability and structural studies on nucleosomes that have been reconstituted with this variant led to the conclusion that nucleosomes are more 'open' or less stable than conventional H2A-containing nucleosomes[68,69]. Interestingly, photobleaching studies[68] showed that the mobility of H2ABBD in the nucleus was found to be faster than that of H2A.

LEUCINE-ZIPPER MOTIF
A leucine-rich protein domain that mediates interactions with other proteins with a similar domain.

Chaperones and exchange factors

Two histone-variant-specific exchange activities have been identified so far — the HIRA and the Swr1 complexes, which catalyse the replacement of H3.3 and H2AZ, respectively. Apart from these, several ATP-dependent chromatin-remodelling factors have been shown to catalyse the displacement of H2A–H2B dimers[70]. Of the remodelling factors tested, SWI/SNF, among others, efficiently catalyses this displacement. The loss of an H2A–H2B dimer agrees with genetic studies, which showed that the depletion of H2A–H2B *in vivo* alleviates the requirement of SWI/SNF at a subset of promoters[71]. Whether this ability of SWI/SNF to displace dimers is an early step in an exchange reaction remains to be determined.

An important factor in facilitating transcription through chromatin is FACT, which removes one copy of the H2A–H2B dimer in a transcription-coupled manner[72,73]. FACT consists of two subunits, SSRP1 and SPT16 (REF. 50). Through SPT16, FACT binds to nucleosomes, but not H3–H4 tetramers, with the SPT16 subunit making contacts with the H2A–H2B dimer[49]. Interestingly, SSRP1 binds to H3–H4 tetramers but not as part of intact nucleosomes. One interpretation of these observations is that SSRP1 helps to stabilize the H3–H4 tetramers, and assists in promoting the reassembly of nucleosomes after the transit of RNA polymerase II. FACT dislocates one dimer, leaving behind a hexasome, and this makes chromatin more accessible to the transit of RNA polymerase II. So, with one H2A–H2B dimer displaced from the nucleosome, FACT can then function synergistically with SWR1, or another dimer-exchange factor, to allow the incorporation of an H2AZ–H2B dimer. A homogeneous population of H2AZ-containing nucleosomes could arise if the removal of one dimer destabilizes the nucleosome and promotes the removal of a second H2A–H2B dimer. As the crystal structure of H2AZ argues against the presence of a mixed population of H2AZ and H2A in the same nucleosomes (as discussed above), whether FACT binds preferentially to the H2A–H2B dimer relative to H2AZ–H2B remains to be investigated. In such a case, a hypothetical functional interaction between FACT and SWR1 could result in the complete displacement of H2A from the nucleosomes with replacement by H2AZ (FIG. 2).

Several histone-binding proteins have been shown to function together with the elongating form of RNA polymerase II during active transcription. One such molecule is Spt6. The role of Spt6 as an elongation factor is evident by its ability to increase the rate of transcription by RNA polymerase II on naked DNA template[74]. It was also shown that Spt6 interacts weakly with both RNA polymerase II and the elongation factor DSIF (DRB-sensitivity-inducing factor), which is comprised of two subunits, Spt4 and Spt5 (REFS 75,76). Spt6 colocalizes with the phosphorylated, elongating form of RNA polymerase II on *D. melanogaster* polytene chromosomes[77,78] as well as with FACT. The function of Spt6 as a chaperone comes from yeast studies, which

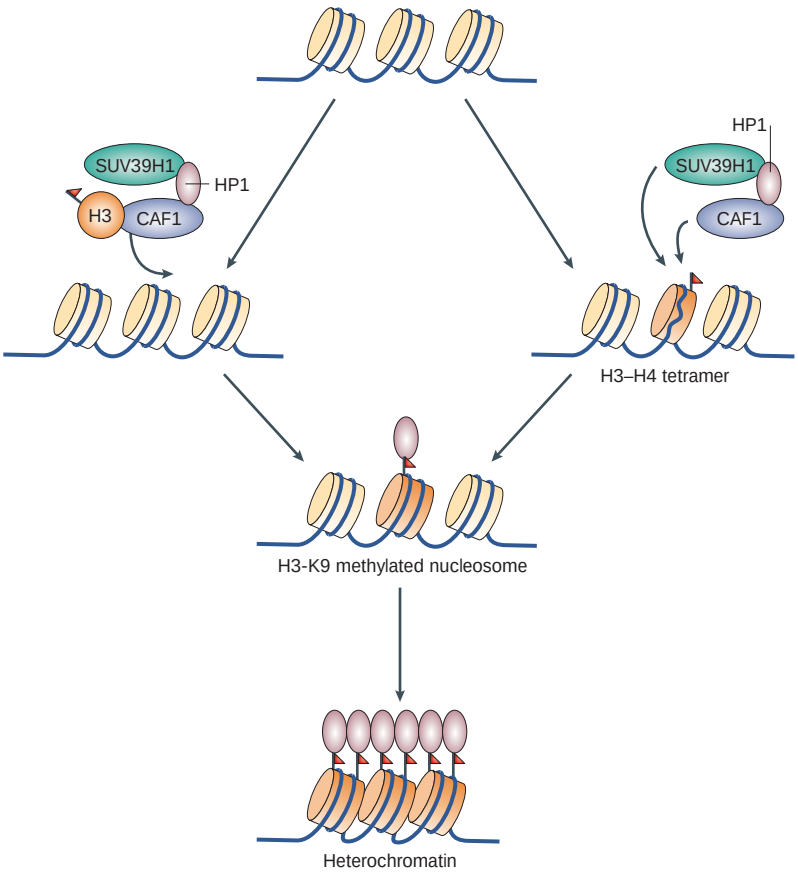


Figure 3 | **Mechanism of repression by histone-exchange complexes and histone methyltransferases.** The chromatin-assembly factor-1 (CAF1) has been shown to interact with the heterochromatin-binding protein HP1, which, in turn, has been associated with the H3-K9 (histone H3 methylated at Lys9) methyltransferase SUV39H1. The methylation of histone H3 that is associated with CAF1 could occur before (left) or after (right) the tetramers have been deposited onto DNA, but before the incorporation of H2A–H2B dimers. This would be followed by the transfer of HP1 from the CAF1 complex to the methylated residue. The stimulation of tri-methylation of H4-K20 by the SUV420H1 and SUV4202 enzymes, which depend on the activity of SUV39H1 and the presence of HP1, would then occur. The establishment of both these marks would lead to the formation and propagation of pericentric heterochromatin.

showed that Spt6 interacts genetically with the histone H3 globular domain. Further biochemical characterization showed that Spt6 binds preferentially to the yeast histones H3 and H4 (REF. 79), and *in vitro* experiments showed that Spt6 transfers H3–H4 onto DNA, albeit rather inefficiently. As Spt6 interacts with Spt5 *in vivo*, the presence of the DSIF elongation factor might be required for increasing the deposition potential of Spt6. In addition, the Spt16 subunit of FACT genetically interacts with Spt6 (REF. 80). This communication could result in a mechanism whereby destabilization of the nucleosome by FACT could facilitate the exchange of an H3–H4 tetramer by Spt6. The exchange function of Spt6 in coordination with a variant-specific exchanger such as HIRA could facilitate the replacement of H3.1 by H3.3 during transcription (FIG. 2).

Exchange of methylated histones by CAF1

The deposition of H3.3 and H2AZ seems to contribute to the establishment of euchromatin, whereas the only histone variant that is implicated in the formation of repressive chromatin is macroH2A, but this is a unique case (as discussed above). This leads us to question the mechanism whereby methylated histones are deposited onto chromatin in the formation of more generalized repressive domains. Of particular interest is the methylation of H3 at Lys9 (H3-K9), as it is by far the most studied histone modification with respect to function and has been clearly shown to repress transcription through the recruitment of HP1. Interestingly, most enzymes that modify H3 at Lys9 seem to be efficient in catalysing this reaction on octamers but not nucleosomes. With the exception of ESET (ERG-associated protein with a SET domain), other H3-K9 methyltransferases such as SUV39H1 and G9A modify Lys9 on core histone substrates, and their activity is inhibited when histones are presented in the form of nucleosomes *in vitro*. This being the case, how do methylated H3-K9 histones find their way to chromatin? CAF1 as well as SUV39H1 bind to HP1 *in vivo*, thereby establishing their association indirectly[81,82]. The H3–H4 dimers or tetramers complexed with CAF1 and HP1 could be methylated by SUV39H1 and deposited onto chromatin. HP1 could then be transferred to the methylated H3-K9 residue, leading to the formation and propagation of repressive chromatin (FIG. 3). Recent studies have confirmed the presence of a replication-specific pool of HP1α that localizes to the boundaries of pericentric heterochromatin in a CAF1-dependent manner. Although the CAF1–HP1α complex is distinct from the H3.1 complex (see above) owing to the absence of the histone H3 and H4 polypeptides, it can assemble newly synthesized cytosolic histones into chromatin as efficiently as the H3.1 complex[83]. Although SUV39H1 cannot methylate nucleosomes *in vitro*, the possibility that tetramers that are assembled on DNA are a favourable substrate cannot be excluded. In this aspect, the dimer versus tetramer deposition of H3–H4 becomes an important issue. If methylated dimers are added to DNA, are the second dimers that are added also methylated? If not, are nucleosomes that are 'hemi-methylated' sufficient to mediate repression? Future efforts to address the scope of modifications within the same nucleosome are needed.

Concluding remarks

Two histone-variant-specific exchange complexes have been described recently — HIRA and the Swr1 complex, which deposit histones H3.3 and H2AZ onto chromatin, respectively. Recent and exciting advances in studies of histone variants, their function and specific exchange complexes have enriched our understanding of the regulation of gene expression. Further studies are required to establish a link between complexes that disrupt nucleosome structure and those that swap histones on remodelled chromatin. In addition, the discovery of new chaperones that are involved in the assembly of the other variants into chromatin will give us a fuller appreciation of how diverse, and yet prescribed, this process is.

1. Malik, H. S. & Henikoff, S. Phylogenomics of the nucleosome. *Nature Struct. Biol.* **10**, 882–891 (2003).
An outstanding, comprehensive review on the evolution and functions of variant histones.
2. Kimmins, S. S. & Sassone-Corsi, P. Chromatin remodelling and epigenetic features of germ cells. *Nature* (in the press).
3. Jackson, V. *In vivo* studies on the dynamics of histone–DNA interaction: evidence for nucleosome dissolution during replication and transcription and a low level of dissolution independent of both. *Biochemistry* **29**, 719–731 (1990).
4. Kimura, H. & Cook, P. R. Kinetics of core histones in living human cells: little exchange of H3 and H4 and some rapid exchange of H2B. *J. Cell Biol.* **153**, 1341–1353 (2001).
5. Lever, M. A., Th'ng, J. P., Sun, X. & Hendzel, M. J. Rapid exchange of histone H1.1 on chromatin in living human cells. *Nature* **408**, 873–876 (2000).
6. Wilhelm, F. X., Wilhelm, M. L., Erard, M. & Duane, M. P. Reconstitution of chromatin: assembly of the nucleosome. *Nucleic Acids Res.* **5**, 505–521 (1978).
7. Smith, S. & Stillman, B. Stepwise assembly of chromatin during DNA replication *in vitro*. *EMBO J.* **10**, 971–980 (1991).
8. Sobel, R. E., Cook, R. G., Perry, C. A., Annunziato, A. T. & Allis, C. D. Conservation of deposition-related acetylation sites in newly synthesized histones H3 and H4. *Proc. Natl Acad. Sci. USA* **92**, 1237–1241 (1995).
9. Mello, J. A. & Almouzni, G. The ins and outs of nucleosome assembly. *Curr. Opin. Genet. Dev.* **11**, 136–141 (2001).
10. Ma, X. J., Wu, J., Altheim, B. A., Schultz, M. C. & Grunstein, M. Deposition-related sites K5/K12 in histone H4 are not required for nucleosome deposition in yeast. *Proc. Natl Acad. Sci. USA* **95**, 6693–6698 (1998).
11. Shibahara, K., Verreault, A. & Stillman, B. The N-terminal domains of histones H3 and H4 are not necessary for chromatin assembly factor-1-mediated nucleosome assembly onto replicated DNA *in vitro*. *Proc. Natl Acad. Sci. USA* **97**, 7766–7771 (2000).
12. Tyler, J. K. *et al.* The RCAF complex mediates chromatin assembly during DNA replication and repair. *Nature* **402**, 555–560 (1999).
13. Nishioka, K. *et al.* PR-Set7 is a nucleosome-specific methyltransferase that modifies lysine 20 of histone H4 and is associated with silent chromatin. *Mol. Cell* **9**, 1201–1213 (2002).
14. Rice, J. C. *et al.* Mitotic-specific methylation of histone H4 Lys 20 follows increased PR-Set7 expression and its localization to mitotic chromosomes. *Genes Dev.* **16**, 2225–2230 (2002).
15. Bracken, A. P. *et al.* EZH2 is downstream of the pRB–E2F pathway, essential for proliferation and amplified in cancer. *EMBO J.* **22**, 5323–5335 (2003).
16. Loyola, A. & Almouzni, G. Histone chaperones, a supporting role in the limelight. *Biochim. Biophys. Acta* **1677**, 3–11 (2004).
17. Zalensky, A. O. *et al.* Human testis/sperm-specific histone H2B (hTSH2B). Molecular cloning and characterization. *J. Biol. Chem.* **277**, 43474–43480 (2002).
18. Churikov, D. *et al.* Novel human testis-specific histone H2B encoded by the interrupted gene on the X chromosome. *Genomics* **84**, 745–756 (2004).
19. Ahmad, K. & Henikoff, S. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Mol. Cell* **9**, 1191–1200 (2002).
Shows the direct visualization of RI incorporation of H3.3 into chromatin.
20. McKittrick, E., Gafken, P. R., Ahmad, K. & Henikoff, S. Histone H3.3 is enriched in covalent modifications associated with active chromatin. *Proc. Natl Acad. Sci. USA* **101**, 1525–1530 (2004).
21. Tagami, H., Ray-Gallet, D., Almouzni, G. & Nakatani, Y. Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis. *Cell* **116**, 51–61 (2004).
Describes the existence of unique pre-deposition complexes that mediate the incorporation of histones H3.1 and H3.3.
22. Adkins, M. W. & Tyler, J. K. The histone chaperone Asf1p mediates global chromatin disassembly *in vivo*. *J. Biol. Chem.* **279**, 52069–52074 (2004).
23. Maga, G. & Hubscher, U. Proliferating cell nuclear antigen (PCNA): a dancer with many partners. *J. Cell Sci.* **116**, 3051–3060 (2003).
24. Moggs, J. G. *et al.* A CAF-1–PCNA-mediated chromatin assembly pathway triggered by sensing DNA damage. *Mol. Cell. Biol.* **20**, 1206–1218 (2000).
25. Shibahara, K. & Stillman, B. Replication-dependent marking of DNA by PCNA facilitates CAF-1-coupled inheritance of chromatin. *Cell* **96**, 575–585 (1999).
26. Ray-Gallet, D. *et al.* HIRA is critical for a nucleosome assembly pathway independent of DNA synthesis. *Mol. Cell* **9**, 1091–1100 (2002).
27. Mello, J. A. *et al.* Human Asf1 and CAF-1 interact and synergize in a repair-coupled nucleosome assembly pathway. *EMBO Rep.* **3**, 329–334 (2002).
28. Palmer, D. K., O'Day, K., Wener, M. H., Andrews, B. S. & Margolis, R. L. A 17-kD centromere protein (CENP-A) copurifies with nucleosome core particles and with histones. *J. Cell Biol.* **104**, 805–815 (1987).
29. Sullivan, K. F., Hechenberger, M. & Masri, K. Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere. *J. Cell Biol.* **127**, 581–592 (1994).
30. Howman, E. V. *et al.* Early disruption of centromeric chromatin organization in centromere protein A (Cenpa) null mice. *Proc. Natl Acad. Sci. USA* **97**, 1148–1153 (2000).
31. Sharp, J. A., Franco, A. A., Osley, M. A. & Kaufman, P. D. Chromatin assembly factor I and Hir proteins contribute to building functional kinetochores in *S. cerevisiae*. *Genes Dev.* **16**, 85–100 (2002).
32. Wieland, G., Orthaus, S., Ohndorf, S., Diekmann, S. & Hemmerich, P. Functional complementation of human centromere protein A (CENP-A) by Cse4p from *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **24**, 6620–6630 (2004).
33. Black, B. E. *et al.* Structural determinants for generating centromeric chromatin. *Nature* **430**, 578–582 (2004).
34. Faast, R. *et al.* Histone variant H2A.Z is required for early mammalian development. *Curr. Biol.* **11**, 1183–1187 (2001).
35. Clarkson, M. J., Wells, J. R., Gibson, F., Saint, R. & Tremethick, D. J. Regions of variant histone His2AvD required for *Drosophila* development. *Nature* **399**, 694–697 (1999).
36. Allis, C. D. *et al.* hv1 is an evolutionarily conserved H2A variant that is preferentially associated with active genes. *J. Biol. Chem.* **261**, 1941–1948 (1986).
37. Meneghini, M. D., Wu, M. & Madhani, H. D. Conserved histone variant H2A.Z protects euchromatin from the ectopic spread of silent heterochromatin. *Cell* **112**, 725–736 (2003).
38. Rusche, L. N., Kirchmaier, A. L. & Rine, J. The establishment, inheritance, and function of silenced chromatin in *Saccharomyces cerevisiae*. *Annu. Rev. Biochem.* **72**, 481–516 (2003).
39. Suto, R. K., Clarkson, M. J., Tremethick, D. J. & Luger, K. Crystal structure of a nucleosome core particle containing the variant histone H2A.Z. *Nature Struct. Biol.* **7**, 1121–1124 (2000).
40. Rangasamy, D., Berven, L., Ridgway, P. & Tremethick, D. J. Pericentric heterochromatin becomes enriched with H2A.Z during early mammalian development. *EMBO J.* **22**, 1599–1607 (2003).
41. Rangasamy, D., Greaves, I. & Tremethick, D. J. RNA interference demonstrates a novel role for H2A.Z in chromosome segregation. *Nature Struct. Mol. Biol.* **11**, 650–655 (2004).
42. Mizuguchi, G. *et al.* ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* **303**, 343–348 (2004).
Describes the purification of the Swr1 chromatin-remodelling complex that resulted in the recognition of its role in the exchange of H2AZ into chromatin.
43. Krogan, N. J. *et al.* A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1. *Mol. Cell* **12**, 1565–1576 (2003).
Reports the identification of the H2AZ exchange complex Swr1 by a genetic screen that was designed to identify novel genes involved in transcription elongation.
44. Kobor, M. S. *et al.* A protein complex containing the conserved Swi2/Snf2-related ATPase Swr1p deposits histone variant H2A.Z into euchromatin. *PLoS Biol.* **2**, E131 (2004).
Describes the identification of the Swr1 histone-exchange complex by the biochemical analysis of the interaction partners of H2AZ that facilitate its incorporation into chromatin.
45. Matangkasombut, O., Buratowski, R. M., Swilling, N. W. & Buratowski, S. Bromodomain factor 1 corresponds to a missing piece of yeast TFIID. *Genes Dev.* **14**, 951–962 (2000).
46. Ebbert, R., Birkmann, A. & Schuller, H. J. The product of the SNF2/SWI2 paralogue INO80 of *Saccharomyces cerevisiae* required for efficient expression of various yeast structural genes is part of a high-molecular-weight protein complex. *Mol. Microbiol.* **32**, 741–751 (1999).
47. Park, Y. J., Dyer, P. N., Tremethick, D. J. & Luger, K. A new fluorescence resonance energy transfer approach demonstrates that the histone variant H2AZ stabilizes the histone octamer within the nucleosome. *J. Biol. Chem.* **279**, 24274–24282 (2004).
48. Fan, J. Y., Gordon, F., Luger, K., Hansen, J. C. & Tremethick, D. J. The essential histone variant H2A.Z regulates the equilibrium between different chromatin conformational states. *Nature Struct. Biol.* **9**, 172–176 (2002).
49. Belotserkovskaya, R. *et al.* FACT facilitates transcription-dependent nucleosome alteration. *Science* **301**, 1090–1093 (2003).
50. Orphanides, G., LeRoy, G., Chang, C. H., Luse, D. S. & Reinberg, D. FACT, a factor that facilitates transcript elongation through nucleosomes. *Cell* **92**, 105–116 (1998).
51. Downs, J. A., Lowndes, N. F. & Jackson, S. P. A role for *Saccharomyces cerevisiae* histone H2A in DNA repair. *Nature* **408**, 1001–1004 (2000).
52. Rogakou, E. P., Pilch, D. R., Orr, A. H., Ivanova, V. S. & Bonner, W. M. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J. Biol. Chem.* **273**, 5858–5868 (1998).
53. Stiff, T. *et al.* ATM and DNA-PK function redundantly to phosphorylate H2AX after exposure to ionizing radiation. *Cancer Res.* **64**, 2390–2396 (2004).
54. Madigan, J. P., Chotkowski, H. L. & Glaser, R. L. DNA double-strand break-induced phosphorylation of *Drosophila* histone variant H2Av helps prevent radiation-induced apoptosis. *Nucleic Acids Res.* **30**, 3698–3705 (2002).
55. Celeste, A. *et al.* Genomic instability in mice lacking histone H2AX. *Science* **296**, 922–927 (2002).
56. Celeste, A. *et al.* Histone H2AX phosphorylation is dispensable for the initial recognition of DNA breaks. *Nature Cell Biol.* **5**, 675–679 (2003).
57. Fernandez-Capetillo, O. *et al.* H2AX is required for chromatin remodeling and inactivation of sex chromosomes in male mouse meiosis. *Dev. Cell* **4**, 497–508 (2003).
58. Cheung, W. L. *et al.* Apoptotic phosphorylation of histone H2B is mediated by mammalian sterile twenty kinase. *Cell* **113**, 507–517 (2003).
59. Fernandez-Capetillo, O., Allis, C. D. & Nussenzweig, A. Phosphorylation of histone H2B at DNA double-strand breaks. *J. Exp. Med.* **199**, 1671–1677 (2004).
60. de la Barre, A. E., Angelov, D., Molla, A. & Dimitrov, S. The N-terminus of histone H2B, but not that of histone H3 or its phosphorylation, is essential for chromosome condensation. *EMBO J.* **20**, 6383–6393 (2001).
61. Fernandez-Capetillo, O., Lee, A., Nussenzweig, M. & Nussenzweig, A. H2AX: the histone guardian of the genome. *DNA Repair (Amst.)* **3**, 959–967 (2004).
62. Pehrson, J. R. & Fried, V. A. MacroH2A, a core histone containing a large nonhistone region. *Science* **257**, 1398–1400 (1992).
63. Costanzi, C. & Pehrson, J. R. Histone macroH2A1 is concentrated in the inactive X chromosome of female mammals. *Nature* **393**, 599–601 (1998).
64. Mermoud, J. E., Costanzi, C., Pehrson, J. R. & Brockdorff, N. Histone macroH2A1.2 relocates to the inactive X chromosome after initiation and propagation of X-inactivation. *J. Cell Biol.* **147**, 1399–1408 (1999).
65. Rasmussen, T. P., Mastrangelo, M. A., Eden, A., Pehrson, J. R. & Jaenisch, R. Dynamic relocalization of histone MacroH2A1 from centrosomes to inactive X chromosomes during X inactivation. *J. Cell Biol.* **150**, 1189–1198 (2000).
66. Angelov, D. *et al.* The histone variant macroH2A interferes with transcription factor binding and SWI/SNF nucleosome remodeling. *Mol. Cell* **11**, 1033–1041 (2003).
67. Chadwick, B. P., Valley, C. M. & Willard, H. F. Histone variant macroH2A contains two distinct macrochromatin domains capable of directing macroH2A to the inactive X chromosome. *Nucleic Acids Res.* **29**, 2699–2705 (2001).
68. Gautier, T. *et al.* Histone variant H2ABbd confers lower stability to the nucleosome. *EMBO Rep.* **5**, 715–720 (2004).
69. Bao, Y. *et al.* Nucleosomes containing the histone variant H2A.Bbd organize only 118 base pairs of DNA. *EMBO J.* **23**, 3314–3324 (2004).
70. Bruno, M. *et al.* Histone H2A/H2B dimer exchange by ATP-dependent chromatin remodeling activities. *Mol. Cell* **12**, 1599–1606 (2003).
71. Hirschhorn, J. N., Brown, S. A., Clark, C. D. & Winston, F. Evidence that SNF2/SWI2 and SNF5 activate transcription in yeast by altering chromatin structure. *Genes Dev.* **6**, 2288–2298 (1992).
72. Mason, P. B. & Struhl, K. The FACT complex travels with elongating RNA polymerase II and is important for the fidelity of transcriptional initiation *in vivo*. *Mol. Cell. Biol.* **23**, 8323–8333 (2003).
73. Saunders, A. *et al.* Tracking FACT and the RNA polymerase II elongation complex through chromatin *in vivo*. *Science* **301**, 1094–1096 (2003).
74. Endoh, M. *et al.* Human Spt6 stimulates transcription elongation by RNA polymerase II *in vitro*. *Mol. Cell. Biol.* **24**, 3324–3336 (2004).

75. Wada, T. *et al.* DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev.* **12**, 343–356 (1998).
76. Hartzog, G. A., Wada, T., Handa, H. & Winston, F. Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in *Saccharomyces cerevisiae*. *Genes Dev.* **12**, 357–369 (1998).
77. Kaplan, C. D., Morris, J. R., Wu, C. & Winston, F. Spt5 and spt6 are associated with active transcription and have characteristics of general elongation factors in *D. melanogaster*. *Genes Dev.* **14**, 2623–2634 (2000).
78. Andrulis, E. D., Guzman, E., Doring, P., Werner, J. & Lis, J. T. High-resolution localization of *Drosophila* Spt5 and Spt6 at heat shock genes *in vivo*: roles in promoter proximal pausing and transcription elongation. *Genes Dev.* **14**, 2635–2649 (2000).
79. Bortvin, A. & Winston, F. Evidence that Spt6p controls chromatin structure by a direct interaction with histones. *Science* **272**, 1473–1476 (1996).
80. Kaplan, C. D., Laprade, L. & Winston, F. Transcription elongation factors repress transcription initiation from cryptic sites. *Science* **301**, 1096–1099 (2003).
81. Murzina, N., Verreault, A., Laue, E. & Stillman, B. Heterochromatin dynamics in mouse cells: interaction between chromatin assembly factor 1 and HP1 proteins. *Mol. Cell* **4**, 529–540 (1999).
82. Yamamoto, K. & Sonoda, M. Self-interaction of heterochromatin protein 1 is required for direct binding to histone methyltransferase, SUV39H1. *Biochem. Biophys. Res. Commun.* **301**, 287–292 (2003).
83. Quivy, J. P. *et al.* A CAF-1 dependent pool of HP1 during heterochromatin duplication. *EMBO J.* **23**, 3516–3526 (2004).
84. Rice, J. C. *et al.* Histone methyltransferases direct different degrees of methylation to define distinct chromatin domains. *Mol. Cell* **12**, 1591–1598 (2003).
85. Peters, A. H. *et al.* Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. *Mol. Cell* **12**, 1577–1589 (2003).
86. Schotta, G. *et al.* A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. *Genes Dev.* **18**, 1251–1262 (2004).
87. Heard, E. *et al.* Methylation of histone H3 at Lys-9 is an early mark on the X chromosome during X inactivation. *Cell* **107**, 727–738 (2001).
88. Silva, J. *et al.* Establishment of histone h3 methylation on the inactive X chromosome requires transient recruitment of Eed–Enx1 polycomb group complexes. *Dev. Cell* **4**, 481–495 (2003).
89. Plath, K. *et al.* Role of histone H3 lysine 27 methylation in X inactivation. *Science* **300**, 131–135 (2003).
90. Kohlmaier, A. *et al.* A chromosomal memory triggered by Xist regulates histone methylation in X inactivation. *PLoS Biol.* **2**, E171 (2004).
91. Vaquero, A., Loyola, A. & Reinberg, D. The constantly changing face of chromatin. *Sci. Aging Knowledge Environ.* **2003**, RE4 (2003).
92. Zhang, Y. & Reinberg, D. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev.* **15**, 2343–2360 (2001).
93. Turner, B. M. Cellular memory and the histone code. *Cell* **111**, 285–291 (2002).
94. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
95. Cuthbert, G. L. *et al.* Histone deimination antagonizes arginine methylation. *Cell* **118**, 545–553 (2004).
96. Wang, H. *et al.* mAM facilitates conversion by ESET of dimethyl to trimethyl lysine 9 of histone H3 to cause transcriptional repression. *Mol. Cell* **12**, 475–487 (2003).
97. Shi, Y. *et al.* Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell* **119**, 941–953 (2004).

Acknowledgements

We thank Ken Marians, Jerry Hurwitz, Steven Henikoff, Yoshihiro Nakatani and Geneviève Almouzni for helpful discussions. We would also like to thank Lynne Vales for critical reading of the manuscript and members of the Reinberg laboratory for discussions. We apologize to colleagues whose work we have not cited owing to space limitations. This work was supported by grants from the National Institutes of Health and the Howard Hughes Medical Institute to D.R.

Competing interests statement

The authors declare no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

Swiss-Prot: http://www.expasy.ch

Bdf1 | CAF1 | CENPA | Cse4 | H2ABBD | H2AX | H2AZ | H3.1 | H3.2 | H3.3 | HIRA | macroH2A | PCNA | Swr1 | Spt6

FURTHER INFORMATION

Kavitha Sarma's web page: http://www2.umdnj.edu/ngeweb/HTML/sarma.html

Danny Reinberg's laboratory: http://www2.umdnj.edu/ngeweb

Access to this interactive links box is free online.

Cryo-EM Study of the Chromatin Fiber Reveals a Double Helix Twisted by Tetranucleosomal Units

Feng Song,[1,2]* Ping Chen,[1]* Dapeng Sun,[1,2] Mingzhu Wang,[1] Liping Dong,[1,2] Dan Liang,[1,2] Rui-Ming Xu,[1] Ping Zhu,[1]† Guohong Li[1]†

The hierarchical packaging of eukaryotic chromatin plays a central role in transcriptional regulation and other DNA-related biological processes. Here, we report the 11-angstrom–resolution cryogenic electron microscopy (cryo-EM) structures of 30-nanometer chromatin fibers reconstituted in the presence of linker histone H1 and with different nucleosome repeat lengths. The structures show a histone H1-dependent left-handed twist of the repeating tetranucleosomal structural units, within which the four nucleosomes zigzag back and forth with a straight linker DNA. The asymmetric binding and the location of histone H1 in chromatin play a role in the formation of the 30-nanometer fiber. Our results provide mechanistic insights into how nucleosomes compact into higher-order chromatin fibers.

Understanding the structure of chromatin is key to illuminating the functions of chromatin dynamics in epigenetic regulation of gene expression. The structure of the native 30-nm chromatin fiber in nuclei or isolated from nuclei is a regular helix of nucleosomes with a diameter of about 30 nm and a packing density of about 6 to 7 nucleosomes per 11 nm (*1*–*7*). Nucleosomes can be arranged either linearly in a one-start solenoid-type helix with bent linker DNA or zigzag back and forth in a two-start stack of nucleosomes connected by a relatively straight DNA linker (*7*–*9*). The latter class can be further divided into the helical ribbon model and the twisted crossed-linker model by the different orientation angles between the linker DNA and fiber axes (*8*–*10*). The manner by which nucleosome core particles (NCPs) interact with each other in a beads-on-a-string nucleosomal array to form a condensed 30-nm chromatin fiber remains unresolved (*11*). The irregular native chromatin fiber cannot readily form a homogeneous structure suitable for high-resolution structural analyses. The problem has been partially addressed by reconstituting chromatin fibers in vitro on regular tandem repeats of unique nucleosome-positioning DNA sequences with purified histone proteins (*12*, *13*). Fibers reconstituted in the presence of histone H5 with different nucleosome repeat lengths (NRLs) showed similar structures and could fit the one-start interdigitated solenoid structure (*14*). Chromatin fiber with two-start zigzag conformation was also observed on long reconstituted nucleosome arrays (*15*). The x-ray crystal structure of a tetranucleosome with a 20–base pair (bp) linker DNA and without linker histones reveals two stacks of nucleosomes connected by straight linker DNA (*16*); however, a tetranucleosomal array is too short to form a solenoid structure. Linker histones, which are present at close to one molecule per nucleosome in the majority of eukaryotic organisms, have been considered to be essential for 30-nm chromatin fiber formation (*17*–*20*), but their precise location and exact roles in the organization of the higher-order structure still remain to be determined.

Cryo-EM Reconstruction of 30-nm Chromatin Fibers

We have determined, at about 11 Å resolution, a three-dimensional (3D) cryogenic electron microscopy (cryo-EM) structure of 30-nm chromatin fibers reconstituted in vitro on the 12 tandem repeats of 187-bp and 177-bp (12 × 187 bp and 12 × 177 bp) 601 DNA sequence with recombinant *Xenopus laevis* canonical histones lacking post-translational modifications (Fig. 1). The reconstituted 30-nm fibers are in a compact form in the presence of histone H1 with stoichiometry about one H1 per nucleosome under low-salt conditions (fig. S1). The reconstituted chromatin fibers were fixed by 0.2% glutaraldehyde before cryo-EM analysis and displayed as homogeneously compacted particles in the representative field views of the cryo-EM images (Fig. 1A). About 31,000 particles of 12 × 187 bp and 25,000 particles of 12 × 177 bp 30-nm fibers were visually screened and subjected to 2D classification and 3D reconstruction, beginning with an initial model of a featureless Gaussian blob (fig. S2A). The Euler angle distribution indicates that the particles used in our 3D reconstruction have no preferred orientation (fig. S2B). Some selected unsupervised 2D classification averages (Fig. 1A, right) agree well with the raw particles, as indicated by the white box in the micrograph (Fig. 1A, left). The 3D cryo-EM map defines the spatial location of all individual nucleosomes and the path of linker DNA in the 30-nm fiber (Fig. 1, B and C, and movie S1). The overall structure of the dodecanucleosomal 30-nm fiber comprises three tetranucleosomal structural units, which are twisted against each other with linker DNA extended straight to form a two-start helix (Fig. 1, B and D). This disposition can also be deduced from the particularly oriented reference-free average, in which obvious densities connect two adjacent nucleosomes (Fig. 1A, right, white dashed brackets indicated). The four nucleosomes within the structural unit zigzag back and forth to form two stacks of two nucleosome cores (Fig. 1D), which is consistent with the previous observations that deoxyribonuclease I digestion of nuclei produces dinucleosomal periodicity patterns (*21*). According to the previously proposed zigzag two-start helix model, the diameter of the chromatin fiber could be increased accordingly with the length of DNA crossing back and forth between the two-start helix (*8*–*10*). To analyze how the variations in NRLs affect the overall structure of the resulting fibers, we compared the 3D cryo-EM structures of 30-nm fibers reconstituted with two different NRLs, that is, 12 × 187 bp and 12 × 177 bp 601 DNA sequences (Fig. 1C and fig. S2). The two 30-nm fibers with 177-bp NRL and 187-bp NRL display a very similar overall stacking mode of nucleosomes with the connected linker DNAs extended and straight. An increase of 10 bp of NRL does not affect the overall structure and organization of the reconstituted chromatin fiber, but the increase does change the fiber dimension (diameter × height) from about 27.2 × 28.7 nm for 177-bp repeats to about 29.9 × 27.0 nm for 187-bp repeats (Fig. 1C), which is consistent with a basic zigzag two-start helix model. Overall, our cryo-EM structure shows that the 30-nm chromatin fiber follows a path that is basically compatible with a zigzag two-start helix (Fig. 1D), although the fine details of the structure are distinct from the originally proposed model.

Tetranucleosomal Unit with a Two-Start Zigzag Conformation

Within the tetranucleosomal structural unit of the 30-nm fiber, two stacks of two nucleosome cores are connected by straight linker DNA (Fig. 2A and fig. S3C). The two nucleosomes in each stack directly contact head to head through their octamer surfaces. For the structural unit with a 187-bp NRL, the two stacked nucleosome cores are separated center to center by 53.6 Å with each superhelical axis and dyad axis angled at 11.8° and 16.5°, respectively (Fig. 2A, left). In the interface between the cores, a bulk density is present at the junction of the adjacent H2A-H2B dimer, indicating a strong interaction between the H2B-helix α1/αC and the adjacent H2A-helix α2 (Fig. 2B). As described in the x-ray structure of the tetranucleosome, this strong interacting interface does not allow the internucleosomal interaction between the positive N terminus of histone H4

[1]National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China. [2]University of Chinese Academy of Sciences, Beijing 100049, China.

*These authors contributed equally to this work.
†Corresponding author. E-mail: liguohong@ibp.ac.cn (G.L.); zhup@ibp.ac.cn (P.Z.)

(amino acids 16 to 26) and the acidic patch of the H2A-H2B dimer observed in the nucleosome core-particle structure (*16*). The H2B-α3/αC internucleosomal four-helix bundle might be involved in the compaction and stabilization of chromatin fiber (*22*), which was not observed in our cryo-EM map of reconstituted chromatin fiber, presumably because of the interaction of DNA involved in chromatin reconstitutions.

The two dinucleosomal stacks are separated by 196 Å center to center and are rotated 54.5° left-handed with respect to their stack axes (the axis passing through the centers of the two nucleosomes in each stack) for the 187-bp NRL (Fig. 2A, right). We docked the x-ray atomic structure of the 167-bp NRL in the absence of linker histone (*16*) into the EM density map of our tetranucleosomal structural units with 187- and 177-bp NRLs, respectively (Fig. 2B and fig. S3C). The structures for each stack of two nucleosomes fit very well with our cryo-EM map in both cases. However, the distance and rotation between the two stacks are quite different, which may be caused by the presence of histone H1 or the different NRLs used. The two stacks are separated by 146.1 Å for the tetranucleosome with the 167-bp NRL used for x-ray investigation, compared with 167 Å for the 177-bp NRL (fig. S3C) and 196 Å for the 187-bp NRL (Fig. 2A) in this study, confirming that the fiber diameters change accordingly to the lengths of the NRLs. In addition, the rotation between the two stacks in the x-ray structure is left-handed by 71.3°, compared with 63.7° for the 177-bp NRL and 54.5° for the 187-bp NRL in our cryo-EM structure (Fig. 2 and fig. S3). If we only consider the intrinsic property of the difference in NRLs, each 10-bp increment in linker DNA should increase the length of DNA by 3.4 nm and change the twist between adjacent nucleosomes by −17 based on the DNA twist of 10.5 bp per turn, suggesting that the presence of H1 may impose additional effects on the specific distance and the rotation of the two stacks in our cryo-EM structure (Fig. 3A). The defined location of H1 in each nucleosome, as indicated in Fig. 3B, reveals that the direct interaction of H1 with both the dyad and the entering and exiting DNA in the NCPs may alter the angle of the entry-exit DNA and constrain the linker length and rotation angle between the stacks.

Twist and Interactions Between Tetranucleosomal Units

For the dodecanucleosomal 30-nm chromatin fiber, we define three structural units as unit 1 for nucleosomes N1 to N4, unit 2 for N5 to N8, and unit 3 for N9 to N12. The rotation angles and the separation distances between units 1 and 2 and between units 2 and 3 are slightly different, which may be a result of the end effect of the dodecanucleosomal chromatin fiber. In the cryo-EM structure of the 30-nm fiber with the 187-bp NRL, units 1 and 2 are related by a 48.9° rotation around the fiber axis and a 72.2 Å translation along the axis, whereas a 52.9° rotation and a 68.8 Å translation were observed for units 2 and 3. Here, the fiber axis is defined as the axis bisecting the angle between the two stack axes and orthogonally intersecting the axis connecting the centers of the two stacks in each structural unit. The internucleosomal interactions between the structural units are different from that observed within the tetranucleosomal units. The internucleosomal interface between the structural units displays relatively strong density where the N terminus of H4 meets the adjacent H2A-H2B dimer (Fig. 3C), which indicates that the internucleosomal interactions between the positively charged residues of the H4 N-terminal tail (residues 16 to 23) and the acidic patch of the H2A-H2B dimer may account for the twist between the tetranucleosomal units. The interactions within this region were not observed in the previous x-ray tetranucleosome structure (*16*), but these interactions have been reported to be crucial for the formation of the 30-nm chromatin fibers (*15*, *23*). We generated a series of single mutants in the H4 N-terminal tail, including H4 Lys^{16}→Ala^{16} (K16A), H4R17A (R, Arg), H4R19A, H4K20A, and H4R23A, to examine the functions of these residues in chromatin folding in the presence of histone H1 (fig. S4). Sedimentation velocity in conjunction with van Holde–Weischet analysis was used to identify the structural changes in the chromatin fibers. The 30-nm chromatin fibers containing H4K16A, H4R17A, H4R19A, or H4K20A showed very similar distributions of sedimentation coefficients in comparison with those of the wild-type fibers with a S_{ave} of 46.6S ± 1.7S (where S_{ave} is defined as the sedimentation coefficient at the boundary fraction equal to 50%). However, the chromatin

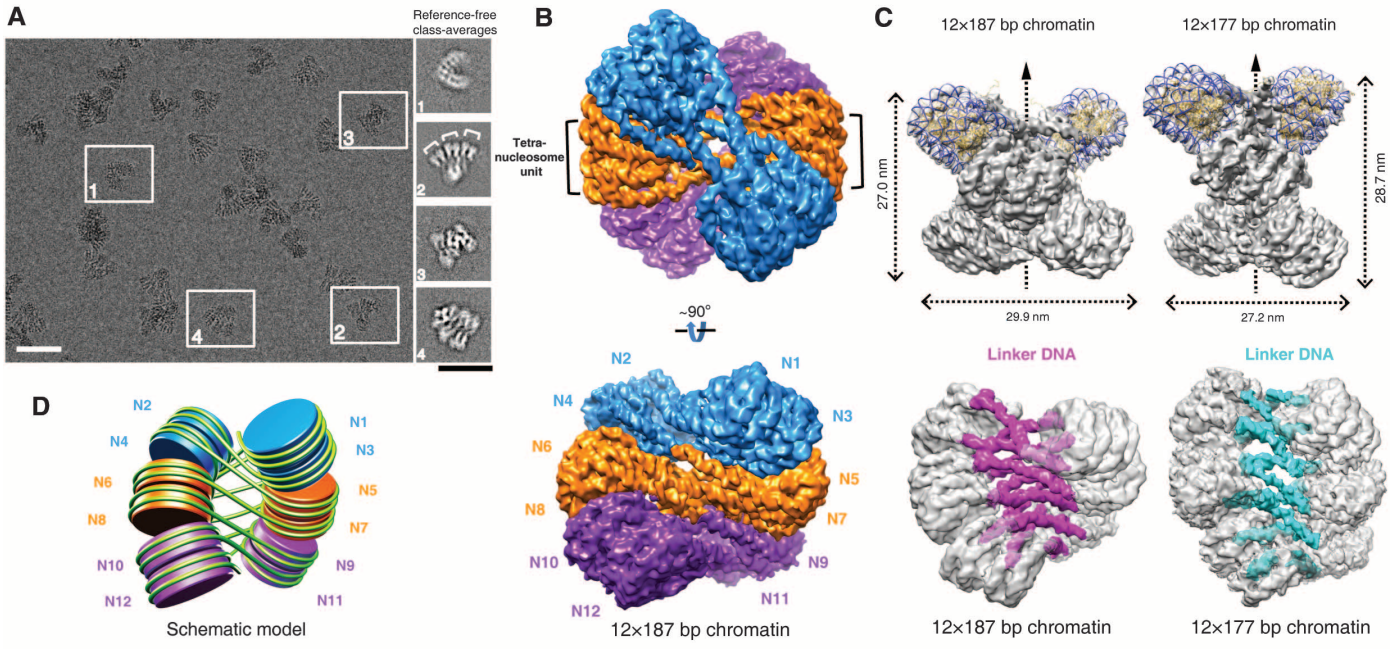


Fig. 1. Cryo-EM reconstruction of 30-nm chromatin fibers. (**A**) A representative cryo-EM micrograph of 30-nm chromatin fibers reconstituted on 12 × 187 bp DNA. Scale bars indicate 50 nm. Four selected unsupervised classification generated averages are shown in the right images, which are in good agreement with the raw particles indicated by white boxes in the micrograph. (**B**) The overall 3D cryo-EM map of the 30-nm chromatin fibers reconstituted on 12 × 187 bp DNA with the three tetranucleosomal structural units highlighted by different colors and viewed from two angles. (**C**) A comparison of the overall structure of 30-nm chromatin fibers reconstituted on 12 × 187 bp and 12 × 177 bp DNA, viewed from two angles with the fiber dimensions directly labeled and their straight linker DNA highlighted. (**D**) A schematic representation of the cryo-EM structure of 30-nm chromatin fibers as shown in (B).

fibers containing H4R23A shifted the sedimentation coefficient distribution to a S_{ave} of 50.8*S* (fig. S4, B and C), which indicated that the chromatin containing H4R23A could still fold into a compact structure but that the mutant H4R23A might alter the chromatin folding mode. Cryo-EM images and the corresponding reference-free class averages of the H4R23A 30-nm fibers showed that the particles appeared much more frequently as two parallel stacks of nucleosomes than the wild-type fibers and with more blurred averages (Fig. 3, D and E). The results suggest that the mutant H4R23A changes the internucleosomal interactions between the structural units and makes it difficult for the tetranucleosomal units to twist stably against each other to form a helical structure.

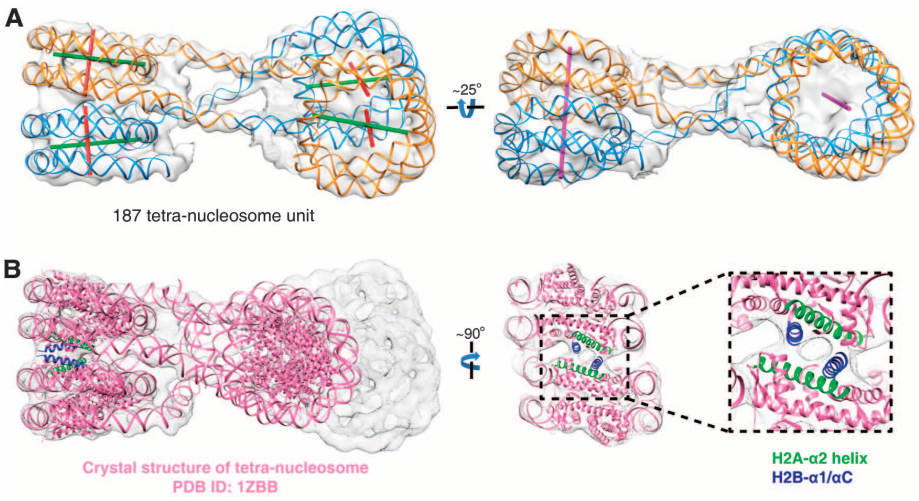


Fig. 2. The structure of a tetranucleosomal unit with the 187-bp NRL. (**A**) The segmented density map for the tetranucleosomal unit in the 30-nm chromatin fibers reconstituted on 12 × 187 bp DNA, shown with the atomic structure of DNA from a docked mononucleosome crystal structure (PDB 1AOI) and modeled presumptive linker DNA and viewed from two angles. Different axes are highlighted by colors, including nucleosome core dyad (green), nucleosome superhelix axes (red), and stack axes (pink). (**B**) A comparison of the 3D cryo-EM map (gray) with the x-ray structure (PDB 1ZBB, pink) of the tetranucleosome (*16*). The strong density where the adjacent H2A-H2B dimer meets is magnified and highlighted in the interface between the nucleosome cores within each stack.

The Asymmetric Location and Self-Association of Histone H1 in Chromatin Fiber

The incorporation of histone H1 plays a fundamental role in determining the higher-order structure of the 30-nm chromatin fiber. As shown in Fig. 3A and figs. S1 and S5A, histone H1 in the 30-nm fibers exhibit a proper 1:1 stoichiometric association with the nucleosome cores. The well-defined locations of the 12 histone H1 molecules in the dodecanucleosomal 30-nm fiber can be visualized in our cryo-EM structure (Fig. 3A, fig. S5A, and movie S2). In each nucleosome core, histone H1 directly interacts with both the dyad and the entering and exiting linker DNA (Fig. 3B), which determines the trajectory of the entry or exit linker DNA segments in nucleosomes and stabilizes the fiber. This interaction mode of histone H1 with the nucleosome core has been proposed previously (*20*, *24*–*26*). In addition, an

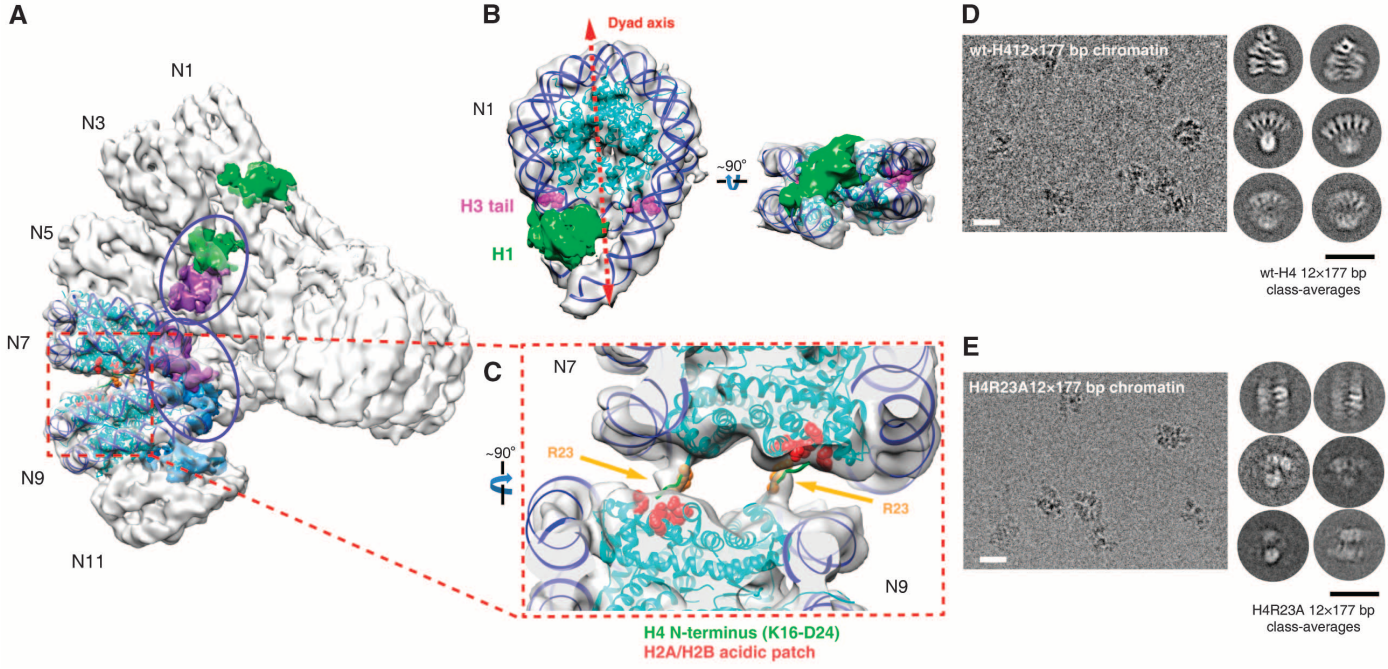


Fig. 3. The interactions between the tetranucleosomal units within the 30-nm fibers. (**A**) The tetranucleosomal units are twisted against each other to form 30-nm fibers. The locations of H1 are highlighted by colors. (**B**) The asymmetric location of H1 in the nucleosome core as viewed from two angles. The mononucleosome N1 segmented from the 12 × 187 bp map (A) with linker DNA density at both entry and exit regions is fitted with an atomic structure of the mononucleosome with an arbitrary length of linker DNA extracted from the tetranucleosome structure (PDB 1ZBB). The presumptive histone H1 is highlighted in green. The H3 tails in the x-ray structure are colored in magenta to locate their relative positions to H1. (**C**) Detail of the strong density where the N terminus of H4 meets the adjacent H2A-H2B dimer. (**D** and **E**) A comparison of the cryo-EM images of the particles and their related unsupervised classification generated averages for 30-nm chromatin fibers reconstituted with wild-type (wt)–H4 (D) and H4R23A (E), respectively. A subset of the wt-H4 12 × 177 bp data with the same particle number, 2664, as that of the H4R23A 12 × 177 data set were randomly picked from the entire data set and subjected to an independent unsupervised 2D classification for comparison. Scale bars, 30 nm.

apparent off-axis location of the globular domain of histone H1 not only in a mononucleosome (Fig. 3B) but also in a tetranucleosomal unit (Fig. 3A, fig. S5A, and movie S2) can be observed in our cryo-EM structure, and this domain localization plays a critical role in the formation of a twist between tetranucleosomal units. The asymmetric location of H1 in mononucleosomes results in the discrimination between different sides of a nucleosome, similar to a coin; we define the "head" as the side of nucleosome with the small portion of H1 and the "tail" as the side with the large portion of H1. The head-to-head nucleosomal interaction in each stack within each tetranucleosomal unit only permits the tail-to-tail nucleosomal interactions between tetranucleosomal units for further stacking and twisting of the chromatin fiber (Fig. 3A). The self-association of H1 had been reported to play an important role in the organization and stabilization of the 30-nm chromatin fiber (*27*). Our cryo-EM structure also revealed a previously unknown arrangement of histone H1 molecules in the reconstituted 30-nm chromatin fiber, in which only the large portions of H1 (most likely its global domain) on the tail side of nucleosomes can interact directly with each other and impart an additional twist between each structural unit (Fig. 3A and fig. S5A). The specific asymmetric binding and location of histone H1 in both mononucleosome and tetranucleosomal units determine the formation of the double-helical 30-nm fiber with a spiral twist of tetranucleosomal units. To further explore the interactions between histone H1 and nucleosomal DNAs, we extracted the densities of individual mononucleosomes with the presumptive H1 from the reconstructed map of the dodecanucleosomal chromatin fiber and averaged them on a 3D level (for details, see materials and methods). Guided by a rigid-body fitting, we found that the crystal structure of *Gallus gallus* histone H5 globular domain (gH5) [Protein Data Bank (PDB) code 1HST] can be docked well into a region, which is a part of the presumptive H1 density, in the averaged H1-containing mononucleosomal EM density map (fig. S5B). This region, most likely the globular domain of human histone H1.4 used in this study, interacts with the nucleosomal entry, exit, and dyad DNAs in a three-contact mode (fig. S5B, red dots), in agreement with the computational analysis of gH5 (*28*) and the mapping of histone H1.5-nucleosome interactions (*24*). Extra densities that cannot be assigned to the globular domain of H1 are presumably contributed by the N- and C-termini of H1 (fig. S5B, blue lines and black stars indicated) and/or partially contributed by the N-termini of histone H3 (Fig. 3B and fig. S5B, magenta tail), which was previously shown to contribute to chromatin folding (*29*).

Structural Model for Chromatin Fibers

Our 3D cryo-EM structure of the 30-nm chromatin fiber allows construction of a fine structural model for 30-nm fibers in the presence of histone H1 (Fig. 4). To evaluate the possible end effects of the relatively short 12 tandem repeats of nucleosomes, we reconstituted the 30-nm fibers with 24 repeats of 177-bp (24 × 177 bp) 601 DNA template and acquired the 3D cryo-EM structure at ~25 Å resolution (Fig. 4A, fig. S6, and movie S3). Two copies of 12 × 177 bp chromatin density map can be docked into the map of 24 × 177 bp chromatin without further modification (Fig. 4B). In addition, the interactions between the intra- and interunits in the 24 × 177 bp chromatin are well maintained compared with the 12 × 177 bp chromatin model (fig. S6E). We then built the model of 30-nm chromatin by directly stacking the cryo-EM structure of dodecanucleosomal 30-nm fibers on top of each other to form a continuous fiber. The resulting direct model reveals that the chromatin fiber is exhibited as a left-handed double helical structure twisted by tetranucleosomal units (Fig. 4C). The 30-nm fiber with the 187-bp NRL contains 7.07 tetranucleosomal units (28.3 nucleosomes) per turn in a period of 49.9 nm, yielding a nucleosome packing density of about 6.2 nucleosomes per 11 nm; this density is very similar to the measured values for mass per unit length of the chromatin fiber (*4*, *30*). To examine the effects of the previously indicated slight difference in rotation and separation between different units in the 30-nm fiber of dodecanucleosomes in the model construction, we also built two other models with a repeat stacking of units 1 and 2 or units 2 and 3, respectively (fig. S7, A and B). These two alternative models show a high resilience to the double-helical twist model formed from the direct stacking of dodecanucleosomes, except for slight changes in the packing density of nucleosomes that were identified as 6.1 nucleosomes per 11 nm for the model built by units 1 and 2 and 6.4 nucleosomes per 11 nm for the model built by units 2 and 3 (Fig. 4C and fig. S7).

Conclusions

The existence of 30-nm fibers in nuclei still remains to be elucidated, whereas the structure of chromatin fiber must be variable in vivo because of its highly dynamic and heterogeneous property with intrinsic compositions, such as different NRLs, linker histones, histone and DNA modifications, and histone variants. Although our 3D cryo-EM structures for the reconstituted 30-nm chromatin fiber show basically a two-start zigzag configuration for nucleosome arrangement, other forms of chromatin structures may exist in different conditions, for example, the one-start solenoid structure in the presence of H5 and magnesium with longer NRLs as suggested previously (*14*, *31*). Nevertheless, the formation of the double-helical structure of the reconstituted canonical 30-nm fibers using the 177-bp and 187-bp NRLs in the presence of H1 under a low-salt condition is basically driven by their intrinsic biophysical and biochemical properties; thus, the fundamental principles may also be applicable in vivo. Histone modifications and histone variants may also play important roles in the regulation of higher-order chromatin structure via modulating the internucleosomal surface interactions between tetranucleosomal units.

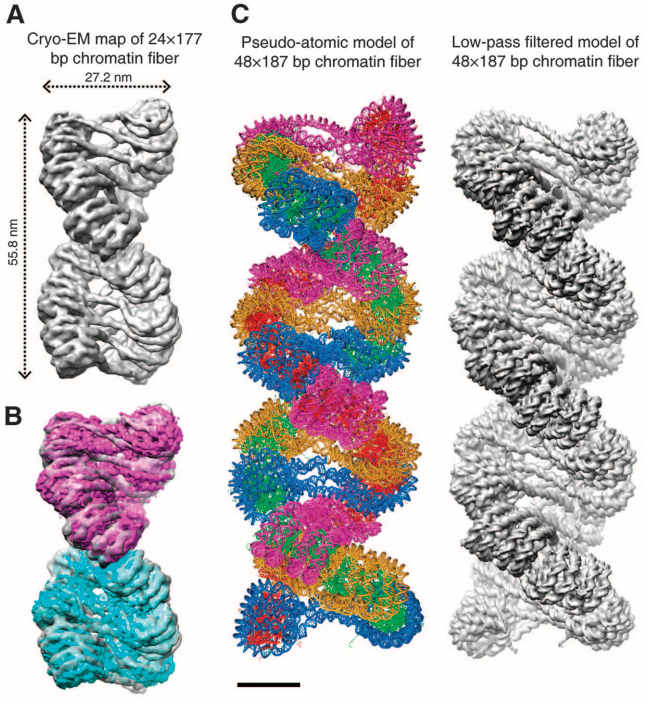


Fig. 4. The 30-nm chromatin fiber model. (A) The overall 3D cryo-EM map of the 30-nm chromatin fiber reconstituted on 24 × 177 bp 601 DNA, with the length and diameter of the fiber indicated. (**B**) The structure of 24 × 177 bp 30-nm fibers is docked by two copies of the cryo-EM structure of 12 × 177 bp 30-nm fibers. The fitting was optimized by the reported correlation value in UCSF Chimera. (**C**) A pseudo-atomic model (left, structure of H1 is not included) and its corresponding density map low-pass filtered to 11 Å (right), built by directly stacking the cryo-EM structure of the dodecanucleosomal 30-nm fiber with 187-bp NRLs on top of each other to form a continuous fiber. Scale bar, 11 nm.

References and Notes

1. R. Ghirlando, G. Felsenfeld, *J. Mol. Biol.* **376**, 1417–1425 (2008).
2. E. C. Pearson, P. J. Butler, J. O. Thomas, *EMBO J.* **2**, 1367–1372 (1983).
3. V. Graziano, S. E. Gerchman, D. K. Schneider, V. Ramakrishnan, *Nature* **368**, 351–354 (1994).
4. S. E. Gerchman, V. Ramakrishnan, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7802–7806 (1987).

5. J. Widom, A. Klug, *Cell* **43**, 207–213 (1985).
6. J. P. Langmore, J. R. Paulson, *J. Cell Biol.* **96**, 1120–1131 (1983).
7. J. T. Finch, A. Klug, *Proc. Natl. Acad. Sci. U.S.A.* **73**, 1897–1901 (1976).
8. C. L. Woodcock, L. L. Frado, J. B. Rattner, *J. Cell Biol.* **99**, 42–52 (1984).
9. S. P. Williams *et al*., *Biophys. J.* **49**, 233–248 (1986).
10. M. F. Smith, B. D. Athey, S. P. Williams, J. P. Langmore, *J. Cell Biol.* **110**, 245–254 (1990).
11. H. G. Davies, J. V. Small, *Nature* **217**, 1122–1125 (1968).
12. L. M. Carruthers, C. Tse, K. P. Walker 3rd, J. C. Hansen, *Methods Enzymol.* **304**, 19–35 (1999).
13. P. T. Lowary, J. Widom, *J. Mol. Biol.* **276**, 19–42 (1998).
14. P. J. Robinson, L. Fairall, V. A. Huynh, D. Rhodes, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 6506–6511 (2006).
15. B. Dorigo *et al*., *Science* **306**, 1571–1573 (2004).
16. T. Schalch, S. Duda, D. F. Sargent, T. J. Richmond, *Nature* **436**, 138–141 (2005).
17. F. Thoma, T. Koller, A. Klug, *J. Cell Biol.* **83**, 403–427 (1979).
18. J. O. Thomas, *Curr. Opin. Cell Biol.* **11**, 312–317 (1999).
19. D. L. Bates, J. O. Thomas, *Nucleic Acids Res.* **9**, 5883–5894 (1981).
20. J. Allan, P. G. Hartman, C. Crane-Robinson, F. X. Aviles, *Nature* **288**, 675–679 (1980).
21. D. Z. Staynov, *Bioessays* **30**, 1003–1009 (2008).
22. T. D. Frouws, H. G. Patterton, B. T. Sewell, *Biophys. J.* **96**, 3363–3371 (2009).
23. K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, T. J. Richmond, *Nature* **389**, 251–260 (1997).
24. S. H. Syed *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **107**, 9620–9625 (2010).
25. D. Z. Staynov, C. Crane-Robinson, *EMBO J.* **7**, 3685–3691 (1988).
26. B. R. Zhou *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19390–19395 (2013).
27. G. J. Carter, K. van Holde, *Biochemistry* **37**, 12477–12488 (1998).
28. L. Fan, V. A. Roberts, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 8384–8389 (2006).
29. S. H. Leuba, C. Bustamante, K. van Holde, J. Zlatanova, *Biophys. J.* **74**, 2830–2839 (1998).
30. J. Bednar *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14173–14178 (1998).
31. A. Routh, S. Sandin, D. Rhodes, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 8872–8877 (2008).

Acknowledgments: This work was supported by grants from the National Basic Research Program of China (2010CB912400 to P.Z., 2011CB966300 to G.L., and 2009CB825500 to R.M.X. and P.Z.); the National Natural Science Foundation of China (91219202 to G.L., 31230018 to P.Z., 91019007 to G.L., 21261130090 to P.Z., and 31000566 to P.C.); Strategic Priority Research Program (XDA01010304 to G.L. and XDB08010100 to P.Z. and R.M.X.) and Key Research Program (KJZD-EW-L05 to P.Z., G.L., and R.M.X.) from the Chinese Academy of Sciences; and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, to P.C. All EM data were collected and processed at the Center for Bio-imaging, Institute of Biophysics, Chinese Academy of Sciences. We thank G. Ji and X. Huang for their technical help and support with electron microscopy and L. Ling for technical help and support with the data processing in the High Performance Computing Service Station. We are also indebted to the colleagues whose work could not be cited because of the limitation of space. The cryo-EM maps for the 12 × 177 bp, 12 × 187 bp and 24 × 177 bp chromatin fibers were deposited into the Electron Microscopy Data Bank with the accession codes EMD-2600, EMD-2601 and EMD-2602, respectively. The authors declare no conflicts of interest.

Supplementary Materials

www.sciencemag.org/content/344/6182/376/suppl/DC1
Materials and Methods
Figs. S1 to S8
References (*32*–*41*)
Movies S1 to S3

28 January 2014; accepted 18 March 2014
10.1126/science.1251413

Transcription and RNA interference in the formation of heterochromatin

Shiv I. S. Grewal[1] & Sarah C. R. Elgin[2]

Transcription in heterochromatin seems to be an oxymoron — surely the 'silenced' form of chromatin should not be transcribed. But there have been frequent reports of low-level transcription in heterochromatic regions, and several hundred genes are found in these regions in *Drosophila*. Most strikingly, recent investigations implicate RNA interference mechanisms in targeting and maintaining heterochromatin, and these mechanisms are inherently dependent on transcription. Silencing of chromatin might involve *trans*-acting sources of the crucial small RNAs that carry out RNA interference, but in some cases, transcription of the region to be silenced seems to be required — an apparent contradiction.

Chromatin fibres, which make up chromosomes, are composed of nucleosome arrays, with each nucleosome consisting of an octamer of core histones associated with double-stranded DNA. Great variety in chromatin biochemistry is achieved by a complex system of accessory proteins, which modify, bind and reorganize histone complexes to generate different functional regions in eukaryotic chromosomes. Chromatin can be considered to have two main types of domain: euchromatin, which is gene-rich; and heterochromatin, which is gene-poor. These domains have different patterns of histone modification, are associated with different modes of nucleosome packaging[1] and therefore, presumably, have differences in higher-order packaging[2,3] and nuclear organization (see page 413).

Heterochromatin was initially defined as the portion of the genome that retains deep staining with DNA-specific dyes as the dividing cell returns to interphase from metaphase. Subsequent investigation showed that heterochromatin has a constellation of properties (Box 1). A link between heterochromatin formation and gene silencing has been inferred from the loss of most gene activity on the inactive X chromosome, which is visibly condensed in female mammals, and from the loss of gene expression, correlated with condensed packaging, in position-effect variegation (PEV) in *Drosophila* and other organisms. PEV occurs when a gene that is normally euchromatic is juxtaposed with heterochromatin, through rearrangement or transposition; the resultant variegating phenotype indicates that the gene has been silenced in a proportion of the cells in which it is normally active[1]. Reporter genes that show PEV (packaged in heterochromatin) have a more uniform nucleosome array and, perhaps as a consequence, suffer a loss of 5′ nuclease-hypersensitive sites (that is, regions that are presumed to be nucleosome free and are generally associated with regulatory sequences present in active or readily induced genes)[2,4]. Loss of nuclease-hypersensitive sites depends on Heterochromatin protein 1 (HP1; also known as Suppressor of variegation 205, SU(VAR)205)[5]. Studies in fission yeast, *Schizosaccharomyces pombe*, have shown that HP1-family proteins mediate recruitment and/or spreading of chromatin-modifying factors, such as the multi-enzyme complex SHREC (SNF2- and histone deacetylase (HDAC)-containing repressor complex). Such complexes presumably facilitate the nucleosome modification and positioning needed to organize the higher-order chromatin structures that are essential for diverse heterochromatin functions, including silencing of transcription, suppression of recombination, long-range chromatin interactions and maintenance of genomic integrity[1,3,6].

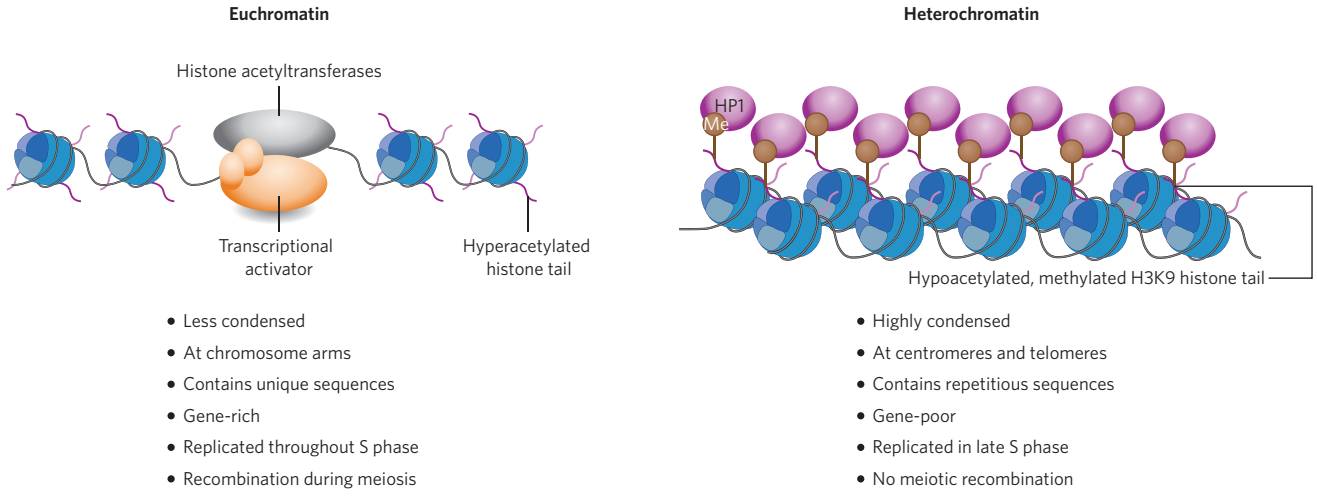
An important characteristic of heterochromatin is the ability of this form of packaging to spread, as evidenced by the occurrence of PEV in *Drosophila* and as shown in *S. pombe*[7] (discussed later). After heterochromatin has been established, it can be stably maintained through mitosis, as shown by the patchy coat of the tortoiseshell (calico) cat, in which coat colour depends on which X chromosome is inactivated. The general properties of euchromatin are antagonistic to those of heterochromatin (Box 1), although it is anticipated that, in euchromatin, there is much more variation in the modification state of the histones and in the arrangement of the nucleosome array, both of which depend on the transcriptional state of a given gene (see page 407). Indeed, greater expression of a *Drosophila* gene (embedded in heterochromatin) that confers variegation has been reported in response to introducing increased amounts of a transcription factor (Gal4), suggesting that there is constant competition in establishing these alternative states[8]. Furthermore, despite the clear distinctions between heterochromatin and euchromatin, low-level transcription has often been found to occur in heterochromatic regions, and these regions contain several hundred genes in *Drosophila*[9]. Resolving these apparent contradictions will provide new insight into how genomes function. In this review article, we focus on recent findings about how heterochromatin formation is targeted and maintained in specific regions of the genome, examining the potential role of transcription associated with the RNA interference (RNAi) system. We draw mainly on results from studies of fungi and animals; interesting results from plants are reported on page 418.

Heterochromatin assembly in *Drosophila*

A key tool for investigating heterochromatin has been the ability to screen for suppressors of PEV (*Su(var)*): that is, mutations elsewhere in the genome that result in loss of silencing at a variegating locus. About 15 such loci have been characterized in *Drosophila melanogaster*, and many more candidates have been identified[10]. The *Su(var)* genes typically encode either proteins that participate directly in the structure of heterochromatin or enzymes that control changes in the modification of histones. A transition between euchromatin and heterochromatin (as might occur in PEV) can roughly be viewed as a series of reactions in which the histone modifications and the proteins associated with the

[1]Laboratory of Biochemistry and Molecular Biology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [2]Washington University, Department of Biology CB1137, St. Louis, Missouri 63130, USA.

Box 1 | Properties of euchromatic and heterochromatic regions



- Less condensed
- At chromosome arms
- Contains unique sequences
- Gene-rich
- Replicated throughout S phase
- Recombination during meiosis

- Highly condensed
- At centromeres and telomeres
- Contains repetitious sequences
- Gene-poor
- Replicated in late S phase
- No meiotic recombination

Trying to define heterochromatin is like trying to define life itself: a cluster of important properties can be specified, but there are exceptions in every instance. For example, centromeres are usually associated with blocks of flanking heterochromatin. However, the inner centromere of *Drosophila* chromosomes is associated with blocks of nucleosomes that contain CENP-A (also known as CID), a variant of histone H3, interspersed with blocks of nucleosomes that contain H3 with a different modification pattern[78]. The elements at the telomeres of *Drosophila* chromosomes are non-LTR (non-long terminal repeat) retrotransposons, which are transcribed[79]. By contrast, the proximal telomere-associated sequences show more of the properties of heterochromatin. The characteristics listed in the figure are most consistently observed in pericentromeric heterochromatin: that is, in the regions that flank the centromeres of many eukaryotic chromosomes. These regions are rich in remnants of transposable elements. It should be noted that little is known about either the stoichiometry of HP1 or the folding of the chromatin fibre in heterochromatin; the figure is meant to convey only the association of HP1 and the condensation of the chromatin fibre.

active state are removed, and then the histone modifications and proteins associated with the inactive state are added (Fig. 1). A sequential set of reactions is required: for example, the lysine residue at position 9 of histone H3 (H3K9) cannot be methylated until it is deacetylated; the binding of SUV4-20 to heterochromatic loci occurs through interaction with HP1 and requires the activity of SU(VAR)3-9 (ref. 11). This sequential requirement undoubtedly contributes to the relative stability of the alternative packaging states. Although the heterochromatic state can be inherited through mitotic and even meiotic cell divisions, a given site can switch from a repressed to an active chromatin state and vice versa at a low frequency. PEV cannot be scored this way in single-celled organisms such as yeast, but this switching can be observed in the phenotype of sectors of a growing colony (Fig. 2).

In *Drosophila*, a small group of proteins are considered likely structural components of pericentromeric heterochromatin because of the observed dose response: whereas having one copy of the encoding gene results in loss of silencing, having three copies results in increased silencing, presumably due to mass action[12]. This set of proteins includes the following: HP1, the first chromodomain protein to be identified[13]; HP2 (also known as SU(VAR)2-HP2), a large protein with no conserved structural motifs[14]; SU(VAR)3-7, a zinc-finger protein[15]; and SU(VAR)3-9, an H3K9-specific histone methyltransferase[16]. However, a well-defined complex of these heterochromatin-associated proteins has not been isolated (despite HP1 interacting with all of these proteins and other proteins[17]), suggesting that an organized protein assembly is present only on the chromatin fibre.

HP1 is a small protein (206 amino acids in *D. melanogaster*) with two conserved domains, an amino-terminal chromodomain and a carboxy-terminal chromoshadow domain, separated by a hinge region (Fig. 3). The chromodomain, found in many chromosomal proteins, folds to create a binding site for the N-terminal tails of histones. HP1 dimerizes through the chromoshadow domain, forming a peptide-binding surface. HP1 interacts stably with SU(VAR)3-9 through the chromoshadow and hinge domains and with di- or trimethylated H3K9 (H3K9me2 or H3K9me3) through the chromodomain[1]. By interacting with both histone-modifying enzyme and modified histone, HP1 provides a foundation for a self-assembly and spreading mechanism, which has been anticipated from studies of PEV (Fig. 3) (see ref. 18 for a review of possible spreading mechanisms). This core assembly seems to be conserved across animals and fungi[19,20] (Table 1). It should be noted that, in many organisms, there are several homologues of HP1 and multiple H3K9 methyltransferases, suggesting the possibility of alternative protein assemblies[19]. However, in *Drosophila*, only HP1A (referred to as HP1 in this review) seems to be associated with known heterochromatic regions, and the ability of other homologues to mimic HP1 in establishing heterochromatin packaging remains to be determined.

The role of RNAi in *S. pombe*

Genetic and biochemical studies using *S. pombe* as a model system have provided great insight into the mechanisms of heterochromatin assembly. Many of the factors involved in heterochromatin formation in *Drosophila* and mammals are conserved in *S. pombe*[19,20]. In particular, the protein Clr4 (cryptic loci regulator 4) — which is the *S. pombe* homologue of *Drosophila* SU(VAR)3-9 and is present in an E3–ubiquitin ligase complex that contains cullin 4 (also known as Pcu4) — has been shown to methylate H3K9 specifically[21–26]. Methylated H3K9 functions as a binding site for recruitment of chromodomain-containing proteins — including chromodomain protein 1 (Chp1), Chp2 and Swi6 (the last of which is a homologue of *Drosophila* HP1) — to heterochromatic loci[22,27–29]. Heterochromatin-associated factors, including methylated H3K9 and Swi6, were found to map to extended chromosomal regions that are coated with heterochromatin complexes at centromeres, telomeres and the mating-type locus[30]. Interestingly, all three of these heterochromatic regions have a common feature — each contains *dg* and *dh* repeat elements, which are preferential targets of heterochromatin formation[7,31–33]. Recent investigations into mechanisms by which these repeats might trigger heterochromatin formation led to the surprising discovery that the RNAi system is involved in the nucleation and assembly of heterochromatin[7,33].

RNAi was first described as a post-transcriptional silencing mechanism in which double-stranded RNA triggers the destruction of cognate RNAs[34]. Subsequent studies have implicated RNAi-associated

mechanisms in diverse cellular functions. In *S. pombe*, mutations in genes encoding factors that are involved in RNAi — such as dicer (Dcr1; an enzyme that cleaves double-stranded RNA), argonaute (Ago1; a PAZ- and PIWI-domain-containing protein that can bind small RNAs) and RNA-directed RNA polymerase 1 (Rdp1) — result in defects in heterochromatin assembly, as shown by loss of silencing at reporter loci[7,33]. An RNAi-induced transcriptional silencing complex (RITS), which contains both a chromatin-associated protein and an RNAi-associated protein, has been identified[35]. RITS contains Chp1 (a chromodomain protein), Ago1 and a protein of unknown function, Tas3 (RITS subunit 3). In addition, RITS also contains small interfering RNAs (siRNAs) derived from the *dg* and *dh* repeats present at the different heterochromatic loci[30,35]. Genome-mapping analyses have shown that Rdp1 and components of RITS are distributed throughout heterochromatic regions in a pattern that is almost identical to the distribution of Swi6 and of H3K9 methylation[30]. Stable binding of RITS to chromatin depends, at least in part, on the binding of the chromodomain of Chp1 to methylated H3K9 (ref. 36). Deletion of *clr4*, or a mutation in the chromodomain-encoding region of *chp1*, results in delocalization of RITS from heterochromatic loci. Interestingly, there are concurrent defects in the processing of *dg* and *dh* repeat transcripts into siRNAs[30,36], suggesting that siRNAs are produced in a heterochromatic environment.

RITS also recruits an RNA-directed RNA polymerase complex (RDRC) that contains Rdp1; this polymerase activity is essential for siRNA production and heterochromatin assembly[37,38]. The generation of siRNAs also requires an RNaseH-like RNA-cleavage activity (referred to as slicer activity) known to be associated with argonaute-family proteins, such as Ago1, found in RITS. Mutations in conserved Ago1 residues that abolish this activity severely affect the processing of *dg* and *dh* repeat transcripts and result in defects in heterochromatin assembly[39,40]. The slicer function of Ago1 has been suggested to be important for the spreading of heterochromatin[39]. It is also possible that siRNAs generated by Ago1-mediated processing of transcripts have a direct structural role in the assembly of higher-order structures that, in addition to mediating silencing, facilitates the local spreading of heterochromatin. These mechanisms, however, cannot by themselves account for the spreading of heterochromatin across large regions, because this requires the HP1-family protein Swi6, which functions as a platform for recruiting the chromatin-modifying effectors (that is, proteins or complexes) involved in heterochromatin assembly[1,7].

These findings suggest that RNAi-mediated heterochromatin assembly in *S. pombe* might occur through a self-reinforcing loop[36,37]. In this model, siRNAs (possibly generated elsewhere) and/or DNA-binding proteins mediate the initial targeting of heterochromatin-associated factors, resulting in the establishment of H3K9 methylation. The presence of methylated H3K9 and associated silencing factors, in turn, allows stable binding of RITS across heterochromatic regions (Fig. 4). RITS presumably functions as a core for the binding of other RNAi-associated factors, such as RDRC, that are essential for the processing of any *dg* and *dh* repeat transcripts. The siRNA-guided cleavage of nascent repeat transcripts by Ago1 (a component of RITS) is thought to be an important step in producing additional siRNAs. It is possible that cleaved transcripts are preferential targets for Rdp1. Rdp1 generates double-stranded RNAs, which are necessary for the generation of siRNAs by Dcr1. Those siRNAs produced *in cis* can feed back to target more heterochromatin complexes but might also have other functions (discussed in the next section).

The exact mechanism by which siRNAs target histone modifications is unclear. The binding of RITS to heterochromatic regions requires *dg* and *dh* siRNAs to be part of the complex. It has been suggested that RITS, tethered to nascent transcripts by siRNAs, might mediate the recruitment of histone methyltransferases such as Clr4 (ref. 35) or that siRNAs directly facilitate the recruitment of chromatin-modifying effectors, such as the Clr4-containing complex, to heterochromatic repeats[7,23]. It is certainly possible that siRNAs target heterochromatin by base-pairing with nascent transcripts[41]; subunits of RITS and RDRC can be crosslinked to transcripts of non-coding centromeric repeats[38]. However, it is unknown whether this binding simply reflects the roles of these factors in processing repeat transcripts or whether it indicates an additional function in recruiting heterochromatin proteins. Recently, artificial tethering of RITS to nascent transcripts has been shown to induce

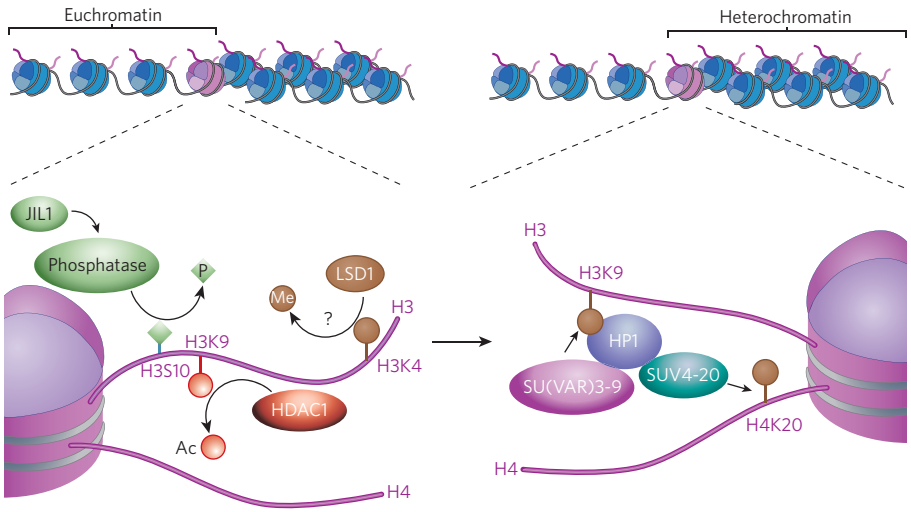


Figure 1 | Changes in histone modification implicated in the switch from a euchromatic to a heterochromatic state in *Drosophila*. Active genes are frequently marked by H3K4me3 (see page 407); this modification is presumably removed by LSD1 (which has not yet been characterized in *Drosophila*). H3K9 is normally acetylated in euchromatin, and this modification must be removed by a histone deacetylase, typically HDAC1. Phosphorylation of H3S10 can interfere with the methylation of H3K9; its dephosphorylation might involve a phosphatase targeted through the carboxy terminus of the protein kinase JIL1 (ref. 10). These transitions set the stage for acquisition of the modifications that are associated with silencing: these include the methylation of H3K9 by SU(VAR)3-9 or another histone methyltransferase, the binding of HP1, and the subsequent methylation of H4K20 by SUV4-20 (an enzyme that is recruited by HP1). Other silencing marks such as methylation of H3K27 by E(Z) (enhancer of zeste; not shown) seem to be relevant in some regions, although this mark is more prominently used by the Polycomb system. Supporting data come from genetic identification of modifiers of PEV, as well as biochemical characterization of the activities of such modifiers and tests of protein–protein interactions[10]. (Figure adapted, with permission, from ref. 10.)

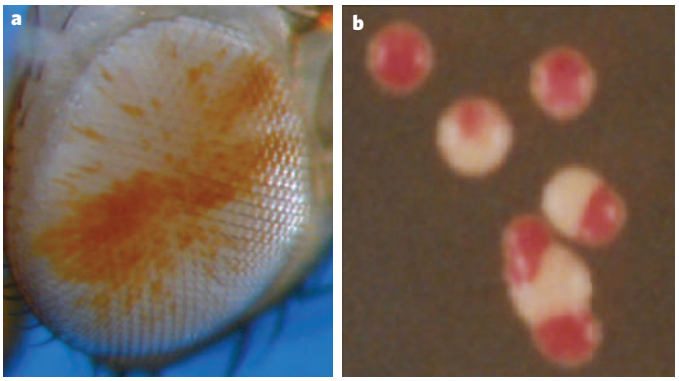


Figure 2 | Variegating phenotypes. Although alternative chromatin packaging states (that is, euchromatin and heterochromatin) can be inherited, they switch at a low frequency. This results in a variegating phenotype in a clonal population of cells. **a**, The image shows a *Drosophila* eye. The *white* gene, expression of which results in a red eye, is active in some eye facets but silenced in others. (Image courtesy of E. Gracheva, Washington University in St. Louis, Missouri.) **b**, The image shows colonies of the fission yeast, *S. pombe*, each of which has differently coloured sectors as a result of variegated expression of the *ade6* gene inserted in a heterochromatic region. (Image courtesy of K. Noma, National Cancer Institute, National Institutes of Health, Bethesda, Maryland.)

local heterochromatin assembly[42]. However, this process requires Dcr1, presumably for the production of siRNAs. Therefore, in addition to the targeting of RITS, other siRNA-dependent steps are required for stable RNAi-mediated heterochromatin nucleation. The emerging view is that, through associations with components of heterochromatin, the RNAi machinery — tethered to specific loci — helps to process transcripts generated from these loci into siRNAs, thereby effectively causing post-transcriptional silencing *in cis*[30,36]. The siRNAs produced in this process also facilitate further targeting of heterochromatin modifications, such as H3K9 methylation. H3K9 methylation enables HP1-family proteins such as Swi6 to localize across heterochromatic regions; these proteins, in turn, facilitate the localization of effectors (such as SHREC) with diverse cellular activities. The HDAC and ATPase (SNF2-like) activities of SHREC are crucial for the proper positioning of nucleosomes to achieve transcriptional silencing[3]. But how can transcription that generates siRNAs occur in a silenced region?

Transcription of heterochromatic repeats

From the results described in the previous section, it can be argued that heterochromatic repeats need to be transcribed to generate the siRNAs that target heterochromatin formation — a circular process. In support of this idea, recent studies have shown that heterochromatic repeats present in the *S. pombe* genome are transcribed by RNA polymerase II (Pol II)[30,43] and that mutations in Pol II impair RNAi-mediated heterochromatin assembly[43,44]. However, an apparent paradox arises in that heterochromatin, in general, is thought to be relatively inaccessible to factors involved in various aspects of DNA metabolism, including the transcriptional machinery[1]. How does Pol II gain access to sequences that are packaged as heterochromatin? Because heterochromatic silencing is thought to be plastic and can be overcome by an increased concentration of transcription factors[8], it can be argued that the promoters driving the transcription of repeats, unlike the promoters of euchromatic genes, have evolved to be somewhat impervious to heterochromatic repression. Indeed, one strand of centromeric repeats in *S. pombe* is always transcribed at a low level[33] but is silenced post-transcriptionally by RNAi-mediated processing of transcripts[33,36].

The transcription of repeats might be facilitated by a specialized mechanism(s) that modulates heterochromatin to provide access for factors involved in different chromosomal processes. In *S. pombe*, Swi6 (*Drosophila* HP1) is thought to function as an 'oscillator' of heterochromatin transcription by directing recruitment of both silencing and antisilencing factors[20]. In addition to factors (such as SHREC) that repress Pol-II-mediated transcription[3], Swi6 also recruits the JmjC-domain-containing protein Epe1 (ref. 40), which was identified in a screen for factors that negatively regulate heterochromatic silencing[45]. Epe1 facilitates Pol-II-mediated transcription of repeats specifically in the context of heterochromatin. It does not seem to have an obligatory role in transcription *per se*, because it is dispensable when heterochromatin is disrupted[40]. The mechanism by which Epe1 counteracts heterochromatic silencing is unknown. Because several JmjC-domain-containing proteins have been shown to catalyse histone demethylation[46], it is possible that Epe1 affects heterochromatin stability through the removal of repressive methylation of lysine residues. However, no such activity has been detected for Epe1 (ref. 47). Epe1 could modulate chromatin by an as yet undefined mechanism. Additional factors targeted to heterochromatic loci by Swi6 or by other mechanisms are also probably important for the transcription of heterochromatin.

In addition to heterochromatin assembly, the transcription of repeats embedded in heterochromatic regions probably has other biological implications. It has been suggested that the transcription of heterochromatic repeats is necessary for continuous production of siRNAs that prime the RNA-induced silencing complex (RISC)-like complexes required to neutralize future invasions by similar sequences[20]. The role of RNAi in destroying viral or transposable element transcripts is conserved in other species, including in *Tetrahymena thermophila* and *Drosophila*[48,49]. Heterochromatin-bound RNAi-associated factors might be components of a memory mechanism that selectively generates a reservoir of siRNAs directed against parasitic DNA elements[20]. It should be noted that, in *S. pombe*, RNAi machinery that is targeted to specific elements can spread to surrounding sequences, including nearby genes, by a process that depends on H3K9 methylation and Swi6 (ref. 36). This might also enable the RNAi machinery to exert heritable control over the expression of sequences located adjacent to repeats.

Silencing of repetitious sequences in metazoans

Although heterochromatin composed of repetitious DNA has become an essential part of the eukaryotic chromosome, maintaining the repetitious sequences in a stable, silent form (repressing both transposition and recombination) is clearly a challenge and a necessity. After it has been initiated, the packaging of heterochromatin occurs in a self-reinforcing manner, through multiple feedback loops[50]. The RNAi machinery seems to be able to detect and respond to repetitious DNA in a variety of ways. But, in metazoans, to what extent might RNAi components be used to target silent regions initially or to maintain these regions? And to what extent might silencing of repetitious DNA depend on transcription *in cis*? The system described in the previous section is unlikely to be universally applicable, because many metazoans, including *Drosophila* and mammals, seem to lack a canonical RNA-dependent RNA polymerase[19].

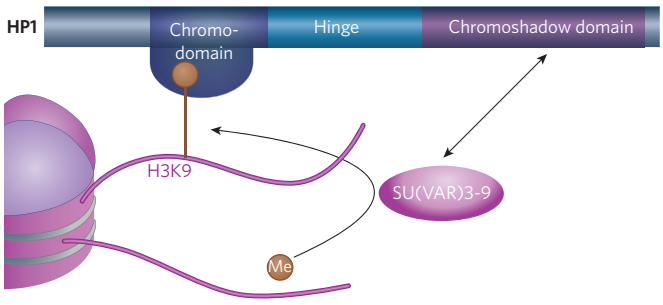


Figure 3 | HP1 and its interactions. HP1 interacts with H3K9me2 and H3K9me3 through its chromodomain, and with SU(VAR)3-9 through its chromoshadow domain. By interacting with both the modified histone and the enzyme responsible for the histone modification, HP1 provides a foundation for heterochromatin spreading and epigenetic inheritance. (Figure adapted, with permission, from ref. 10.)

Table 1 | Factors implicated in heterochromatin formation

Component	*S. pombe*	*Neurospora*	*Drosophila*	Mouse	*Arabidopsis*
Repetitious DNA	Yes	Yes	Yes	Yes	Yes
DNA methylation	No	Yes	No*	Yes	Yes
H3K9 methylation	Yes	Yes	Yes	Yes	Yes
HP1	Yes	Yes	Yes	Yes	No*
Small RNAs	Yes	No*	Yes	Yes	Yes
Pol II	Yes	ND	ND	ND	ND
RDR	Yes	No*	No	No	Yes

Yes indicates that the factor has been implicated to have a role in heterochromatin formation in the given organism. No indicates that the factor is not present in the organism. No* indicates that the organism has the factor but that it seems not to have a role in heterochromatin formation. ND means that the organism has the factor but whether it has a role in heterochromatin formation is unknown. *Arabidopsis*, *Arabidopsis thaliana*; *Neurospora*, *Neurospora crassa*; RDR, RNA-dependent RNA polymerase. (Table adapted from ref. 19.)

Post-transcriptional gene silencing mediated by RNAi, either the degradation of mRNA or a block in its translation, is known to occur in all metazoans that have been examined so far. The first suggestion of RNAi-based transcriptional gene silencing in *Drosophila* came from work showing a loss of expression when multiple copies of a transgene are present[51]. Subsequent analysis showed suppression of PEV (that is, a loss of silencing; as monitored through tandem arrays of *mini–white* and through *white* transgenes in heterochromatin) as a result of mutations in factors involved in the RNAi pathway[51–53]. The loss of silencing is associated with decreased levels of H3K9me2 (ref. 52). Similarly to the RNAi system in other organisms (notably plants; see page 418), the system in *D. melanogaster* might have originated as an antiviral defence mechanism[54,55]. About one-third of the genome is considered to be heterochromatic, and much of that DNA consists of remnants of transposable elements, both DNA transposons and retroviruses. The *Drosophila* genome encodes five PAZ- and PIWI-domain-containing proteins — PIWI, aubergine (AUB), AGO1, AGO2 and AGO3 — which are thought to bind small RNAs. PIWI is required for the self-renewal of germline stem cells, apparently having a key role in silencing retrotransposons and blocking their mobilization in the germ line[56]. Both PIWI and AUB are found associated with siRNAs of 24–29 nucleotides that are derived from repetitious sequences in the germ line[49,57]. *In vitro*, PIWI has RNA-cleavage activity[57], and it has been suggested that germline siRNAs might be generated by a unique processing mechanism that depends on cleavage of long single-stranded transcripts rather than double-stranded RNA[49]. How this silencing activity might influence heterochromatin formation in somatic cells (if at all) is unclear at present.

The effects of mutations in *AGO2* are clearly seen in early *Drosophila* embryos as defects in chromosome condensation, nuclear kinesis and spindle assembly, all potentially correlated with defects in the formation of centric heterochromatin[58]. Similar defects are observed when heterochromatin fails to form in *S. pombe* and other species[59–61]. In *Drosophila* with mutations in the genes that encode SU(VAR)3-9, HP1 or DCR-2, cells have disorganized nucleoli, as well as disorganized centric heterochromatin. In these circumstances, there is a substantial increase in extrachromosomal repetitious DNA in mutant tissues[62]. Similarly, mutations in the genes encoding the RNAi machinery in *S. pombe* also result in defects in maintaining chromosome integrity, including high rates of recombination at genes that encode ribosomal RNA[30]. Therefore, although repetitious DNA now contributes to essential chromosome structures, it is crucial to maintain this DNA specifically in a heterochromatic form, and genetic analysis indicates that the RNAi system has a role in this process. In the absence of any recognizable RNA-directed RNA polymerase activity, this might be accomplished by targeting heterochromatin formation to specific sites, either through DNA–protein interactions or through an RNAi-based recognition system, followed by spreading of the heterochromatin modifications and structure. Similarly to *S. pombe*, the spreading of heterochromatin in *D. melanogaster* (as monitored by PEV) depends on HP1 and SU(VAR)3-9.

Targeting heterochromatin formation

Although much of our discussion focuses on RNAi-based mechanisms, it is important to note that heterochromatin proteins can be recruited to specific sites (known as silencers) by DNA-binding factors. For example, in addition to the RNAi-mediated targeting of heterochromatin to a *dg*- and *dh*-like repeat element located in the silent mating-type region of *S. pombe*, the DNA-binding proteins Atf1 (activating transcription factor 1) and Pcr1, which belong to the ATF/CREB (cyclic-AMP-responsive-element-binding protein) family, have been shown to cooperate with components of SHREC to nucleate heterochromatin assembly independently in this region[6,63,64]. Similarly, redundant mechanisms of heterochromatin nucleation also operate at telomeres in *S. pombe*, where the TRF (TTAGGG repeat factor)-family DNA-binding protein Taz1, in conjunction with Ccq1 (coiled-coil quantitatively enriched protein 1), functions in parallel to the RNAi machinery to nucleate heterochromatin[3,32]. Regardless of the nucleation mechanism, heterochromatin targeted to specific sites can spread, and it provides a sequence-independent platform for cellular effectors with appropriate activities (such as SHREC, the RNAi machinery and cohesin) to be recruited across large regions[20].

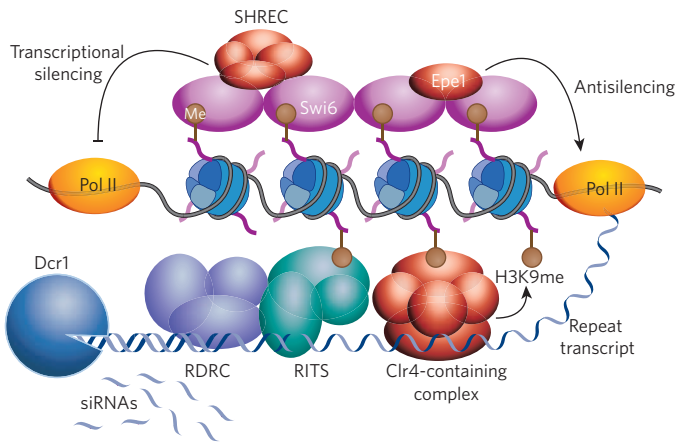


Figure 4 | Model showing RNAi-mediated heterochromatin assembly and silencing in *S. pombe*. Centromeric repeat (*dg* and *dh*) transcripts produced by Pol II are processed by the RNAi machinery, including the complexes RITS and RDRC (which interact with each other and localize across heterochromatic regions). The slicer activity of Ago1 (a component of RITS) and the RNA-directed RNA polymerase activity of Rdp1 (a component of RDRC) are required for processing the repeat transcripts into siRNAs. The siRNA-guided cleavage of nascent transcripts by Ago1 might make these transcripts preferential substrates for Rdp1 to generate double-stranded RNA, which in turn is processed into siRNAs by Dcr1. The targeting of histone-modifying effectors, including the Clr4-containing complex, is thought to be mediated by siRNAs. This process most probably involves the base-pairing of siRNAs with nascent transcripts, but the precise mechanism remains undefined. siRNAs produced by heterochromatin-bound RNAi 'factories' might also prime the assembly of RISC-like complexes capable of mounting a classic RNAi response. Methylation of H3K9 by Clr4 is necessary for the stable association of RITS with heterochromatic loci, apparently through binding to the chromodomain of Chp1. This methylation event also recruits Swi6, which, together with other factors, mediates the spreading of various effectors, such as SHREC. SHREC might facilitate the proper positioning of nucleosomes to organize the higher-order chromatin structure that is essential for the diverse functions of heterochromatin, including transcriptional gene silencing. Swi6 also recruits an antisilencing protein, Epe1, that modulates heterochromatin to facilitate the transcription of repeat elements, in addition to other functions. A dynamic balance between silencing and antisilencing activities determines the expression state of a locus within a heterochromatic domain.

In several cases documented in mammalian cells, HP1 can be targeted to specific promoters by interaction with DNA-binding complexes and seems to contribute to silencing at these loci (see ref. 65 for an example). However, in these cases, a different histone methyltransferase seems to be responsible for the accompanying H3K9 methylation, and spreading is generally not observed. These findings suggest that the interactions of HP1 with both the modified histone (that is, H3K9me3) and the modifying enzyme (usually SU(VAR)3-9) are crucial for heterochromatin spreading (Fig. 3).

Repetitious DNA is a hallmark of heterochromatin. In the case of satellite DNA (simple sequence tandem repeats), it can be suggested that a specific DNA-binding protein recognizes the satellite DNA sequence, thereby triggering heterochromatin assembly. In *Drosophila*, the protein D1, which when mutated results in a loss of silencing, preferentially binds satellite III DNA, which is (A+T)-rich[66,67]. Similarly, the heterochromatin-associated protein DDP1 (dodeca-satellite-binding protein 1; also known as DP1) binds to a conserved dodeca-satellite DNA sequence; however, this protein, which has 15 tandemly organized KH domains, also binds strongly to single-stranded nucleic acids with this sequence, including RNA[68]. Recent work has shown that DDP1, which also causes a loss of silencing when mutated, has a crucial role in the deposition of HP1 and methylated H3K9 at centromeric heterochromatin[68]. Given the ability of DDP1 (and its mammalian homologues, the vigilins) to bind RNA, it is possible that RNA mediates this interaction. Except for the blocks of satellite DNA, the repetitious sequences present in the *Drosophila* genome (which are mainly remnants of transposable elements and DNA transposons) are diverse. Consequently, a recognition process based on RNA (rather than specific protein binding) seems to be most parsimonious, and this suggestion has been supported by studies on the fourth chromosome of *D. melanogaster*.

The small fourth chromosome of *D. melanogaster* is considered to be entirely heterochromatic by the criteria described in Box 1, but it has 88 genes in the distal 1.2 megabases. Mapping with a *white* reporter transgene showed the presence of interspersed heterochromatic regions (inducing a variegating phenotype) (Fig. 2a) and euchromatic regions (allowing expression that results in a full red eye). Detailed examination of the region around the *Host cell factor* (*Hcf*) gene resulted in the *1360* element, which consists of remnants of a DNA transposon, being identified as a potential site for heterochromatin initiation: *D. melanogaster* with reporters lying within 10 kilobases of a *1360* element showed a variegating phenotype, indicating heterochromatin packaging and silencing, whereas *D. melanogaster* with reporters farther away from a *1360* element showed a red eye, indicating euchromatin packaging and full expression[69]. A direct test — using a *P* transposon carrying one copy of *1360*, upstream of a *white* reporter — demonstrated that *1360* contributes to silencing, because silencing of the reporter is largely lost when the adjacent *1360* is deleted. However, stable heterochromatin (resulting in a variegating phenotype) is only observed when that *P* element is located in a region close to the centromere, indicating a requirement for a high density of repeats locally and/or proximity to the pericentromeric heterochromatin, where HP1 is most abundant. Genetic analysis indicates that this silencing depends not only on HP1 and SU(VAR)3-9 but also on RNAi-pathway components, notably AUB[53]. Whether transcription occurs at the target element *1360* is unknown. Small RNA products have been recovered from *1360* and from ~40 other transposable elements in *Drosophila*[70]. It is probable that other transposable elements, in addition to *1360*, are targets for heterochromatin formation. However, it seems unlikely that all transposable-element remnants are targets, given the mapping results obtained on chromosome *4* with the *white* reporter[69]. The crucial characteristics of targets are unknown but could include the presence of start sites for transcription[53]. Many *1360* remnants contain a sequence known to function as a promoter at the multi-copy *Su(Ste)* locus, resulting in the generation of inverse transcripts that are used in the suppression of the multi-copy *Stellate* gene[71]. These results suggest that remnants of transposable elements could be targeted for silencing by a mechanism using a small RNA and that transcription of some of these elements might be involved.

Concluding remarks

Eukaryotes that tolerate large amounts of repetitious sequences in their genomes generally have both the RNAi machinery and the enzymes and structural proteins required to generate a heterochromatin structure based on H3K9 methylation. Whereas some features of the RNAi system (such as RNA-dependent RNA polymerase) and some features of the heterochromatin structure (such as DNA methylation) are used in only a subset of metazoans, this key shift in histone modification from euchromatin to heterochromatin seems to be universal (Table 1; Fig. 1). The RNAi system *per se* can limit gene expression through post-transcriptional gene silencing and can therefore eliminate some sources of damage from invading repetitious elements. However, by itself, it cannot generate the compact chromatin structures that are required to maintain chromosome integrity and chromosome function in mitosis. Hence, the suggestion that post-transcriptional gene silencing is sufficient to explain the silencing of repetitious elements seems unlikely. So, taking into account our new knowledge (described here) of the delicate balance between the need for expression and the need for silencing, an attractive model remains one in which the RNAi machinery has a key role by generating small RNAs involved in specifically targeting chromatin components (including HP1 and H3K9 methyltransferase) to silence repetitious DNA.

Although an assembly of heterochromatin structure based on binding of HP1 proteins to methylated H3K9 provides a foundation for spreading, the molecular mechanisms by which heterochromatin exerts long-range repressive effects are not fully understood. The oligomerization of chromatin-bound HP1 through the chromoshadow domains might mediate condensation. However, recent evidence suggests that HP1 binding is dynamic[72,73]. An alternative emerging view is that HP1-family proteins facilitate recruitment of regulatory proteins (effectors) that are involved in silencing and other chromosomal processes[20]. Indeed, as described earlier, HP1-family proteins mediate preferential binding of SHREC, which has HDAC activity[3]. The deacetylation of histones, which is a universal property of heterochromatic regions, might result in a lower affinity of transcription factors for target loci or could be crucial for higher-order packaging of nucleosomes, both of which would contribute to silencing. The HP1 and H3K9-methylation system might use several routes to minimize H3K9 acetylation, a key characteristic of the active state.

Evidence from different systems suggests that once triggered, a repressive chromatin structure can be sustained for many cell generations. In *S. pombe*, heterochromatin structures established by RNAi and/or DNA-binding factors are inherited *in cis* for many generations in a manner dependent on Swi6 and histone-modifying activities[7,74]. Moreover, a recent study in *Caenorhabditis elegans*, an organism known to silence repetitious DNA by using RNAi and chromatin-associated factors[75,76], showed that a single exposure to RNAi resulted in dominant silencing of a reporter gene in ~30% of the progeny for many generations[77]. A screen for mutations that affected the maintenance of silencing identified four essential genes: *hda-4* (which encodes a histone deacetylase), *K03D10.3* (which encodes a histone acetyltransferase), *isw-1* (which encodes a homologue of the chromatin-remodelling protein ISW1) and *mrg-1* (which encodes a chromodomain protein). Coupled with the observation that trichostatin A (a histone deacetylase inhibitor) relieves silencing, the results imply that maintenance of silencing is a consequence of heterochromatin formation, heritable even in the absence of the initial RNAi stimulus[77]. Although much remains to be learned about the mechanisms involved, it is clear that proper interplay of the RNAi and heterochromatin systems is crucial for the maintenance and function of our genomes. ■

Note added in proof: Two recent publications have shed light on the production of small repeat-associated RNAs in the germ line of *Drosophila*. PIWI and AUB are found associated with RNAs that are mainly antisense to transposons, whereas AGO3 is found associated with RNAs arising mainly from the sense strand. Complementary relationships between these sense and antisense RNA populations suggest that the

slicer activities of the three proteins work together to produce significant amounts of small RNA from endogenous transcripts[80,81]. Such small RNAs, which are maternally inherited, might promote both transcriptional and post-transcriptional silencing of repetitious DNA.

The demethylation of H3K4 has been suggested to be crucial for heterochromatin formation (Fig. 1), and this has now been shown in *Drosophila*[82].

In addition, a histone H2B ubiquitylation ligase complex (HULC) that facilitates Pol-II-mediated transcription of repeat elements in *S. pombe* has been identified[83]. HULC ubiquitylates H2BK119, and this, in addition to promoting euchromatic gene expression, contributes to the transcription of heterochromatic repeats.

1. Grewal, S. I. & Elgin, S. C. Heterochromatin: new possibilities for the inheritance of structure. *Curr. Opin. Genet. Dev.* **12,** 178–187 (2002).
2. Sun, F. L., Cuaycong, M. H. & Elgin, S. C. Long-range nucleosome ordering is associated with gene silencing in *Drosophila melanogaster* pericentric heterochromatin. *Mol. Cell. Biol.* **21,** 2867–2879 (2001).
3. Sugiyama, T. *et al.* SHREC, an effector complex for heterochromatic transcriptional silencing. *Cell* **128,** 491–504 (2007).
4. Wallrath, L. L. & Elgin, S. C. Position effect variegation in *Drosophila* is associated with an altered chromatin structure. *Genes Dev.* **9,** 1263–1277 (1995).
5. Cryderman, D. E., Tang, H., Bell, C., Gilmour, D. S. & Wallrath, L. L. Heterochromatic silencing of *Drosophila* heat shock genes acts at the level of promoter potentiation. *Nucleic Acids Res.* **27,** 3364–3370 (1999).
6. Yamada, T., Fischle, W., Sugiyama, T., Allis, C. D. & Grewal, S. I. The nucleation and maintenance of heterochromatin by a histone deacetylase in fission yeast. *Mol. Cell* **20,** 173–185 (2005).
7. Hall, I. M. *et al.* Establishment and maintenance of a heterochromatin domain. *Science* **297,** 2232–2237 (2002).
8. Ahmad, K. & Henikoff, S. Modulation of a transcription factor counteracts heterochromatic gene silencing in *Drosophila*. *Cell* **104,** 839–847 (2001).
9. Yasuhara, J. C. & Wakimoto, B. T. Oxymoron no more: the expanding world of heterochromatic genes. *Trends Genet.* **22,** 330–338 (2006).
10. Elgin, S. C. R. & Reuter, G. in *Epigenetics* (eds Allis, C. D., Jenuwein, T. & Reinberg, R.) 81–100 (Cold Spring Harbor Laboratory Press, Woodbury, 2007).
11. Schotta, G. *et al.* A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. *Genes Dev.* **18,** 1251–1262 (2004).
12. Locke, J., Kotarski, M. A. & Tartof, K. D. Dosage-dependent modifiers of position effect variegation in *Drosophila* and a mass action model that explains their effect. *Genetics* **120,** 181–198 (1988).
13. Eissenberg, J. C. *et al.* Mutation in a heterochromatin-specific chromosomal protein is associated with suppression of position-effect variegation in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **87,** 9923–9927 (1990).
14. Shaffer, C. D. *et al.* Heterochromatin protein 2 (HP2), a partner of HP1 in *Drosophila* heterochromatin. *Proc. Natl Acad. Sci. USA* **99,** 14332–14337 (2002).
15. Reuter, G. *et al.* Dependence of position-effect variegation in *Drosophila* on dose of a gene encoding an unusual zinc-finger protein. *Nature* **344,** 219–223 (1990).
16. Tschiersch, B. *et al.* The protein encoded by the *Drosophila* position-effect variegation suppressor gene *Su(var)3-9* combines domains of antagonistic regulators of homeotic gene complexes. *EMBO J.* **13,** 3822–3831 (1994).
17. Greil, F., de Wit, E., Bussemaker, H. J. & van Steensel, B. HP1 controls genomic targeting of four novel heterochromatin proteins in *Drosophila*. *EMBO J.* **26,** 741–751 (2007).
18. Talbert, P. B. & Henikoff, S. Spreading of silent chromatin: inaction at a distance. *Nature Rev. Genet.* **7,** 793–803 (2006).
19. Huisinga, K. L., Brower-Toland, B. & Elgin, S. C. The contradictory definitions of heterochromatin: transcription and silencing. *Chromosoma* **115,** 110–122 (2006).
20. Grewal, S. I. & Jia, S. Heterochromatin revisited. *Nature Rev. Genet.* **8,** 35–46 (2007).
21. Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406,** 593–599 (2000).
22. Nakayama, J., Rice, J. C., Strahl, B. D., Allis, C. D. & Grewal, S. I. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292,** 110–113 (2001).
23. Jia, S., Kobayashi, R. & Grewal, S. I. Ubiquitin ligase component Cul4 associates with Clr4 histone methyltransferase to assemble heterochromatin. *Nature Cell Biol.* **7,** 1007–1013 (2005).
24. Hong, E. J. E., Villen, J., Gerace, E. L., Gygi, S. & Moazed, D. A cullin E3 ubiquitin ligase complex associates with Rik1 and the Clr4 histone H3-K9 methyltransferase and is required for RNAi-mediated heterochromatin formation. *RNA Biol.* **2,** 106–111 (2005).
25. Horn, P. J., Bastie, J. N. & Peterson, C. L. A Rik1-associated, cullin-dependent E3 ubiquitin ligase is essential for heterochromatin formation. *Genes Dev.* **19,** 1705–1714 (2005).
26. Thon, G. *et al.* The Clr7 and Clr8 directionality factors and the Pcu4 cullin mediate heterochromatin formation in the fission yeast *Schizosaccharomyces pombe*. *Genetics* **171,** 1583–1595 (2005).
27. Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410,** 120–124 (2001).
28. Partridge, J. F., Scott, K. S., Bannister, A. J., Kouzarides, T. & Allshire, R. C. *cis*-acting DNA from fission yeast centromeres mediates histone H3 methylation and recruitment of silencing factors and cohesin to an ectopic site. *Curr. Biol.* **12,** 1652–1660 (2002).
29. Sadaie, M., Iida, T., Urano, T. & Nakayama, J. A chromodomain protein, Chp1, is required for the establishment of heterochromatin in fission yeast. *EMBO J.* **23,** 3825–3835 (2004).
30. Cam, H. P. *et al.* Comprehensive analysis of heterochromatin- and RNAi-mediated epigenetic control of the fission yeast genome. *Nature Genet.* **37,** 809–819 (2005).
31. Grewal, S. I. & Klar, A. J. A recombinationally repressed region between *mat2* and *mat3* loci shares homology to centromeric repeats and regulates directionality of mating-type switching in fission yeast. *Genetics* **146,** 1221–1238 (1997).
32. Kanoh, J., Sadaie, M., Urano, T. & Ishikawa, F. Telomere binding protein Taz1 establishes Swi6 heterochromatin independently of RNAi at telomeres. *Curr. Biol.* **15,** 1808–1819 (2005).
33. Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297,** 1833–1837 (2002).
34. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391,** 806–811 (1998).
35. Verdel, A. *et al.* RNAi-mediated targeting of heterochromatin by the RITS complex. *Science* **303,** 672–676 (2004).
36. Noma, K. *et al.* RITS acts *in cis* to promote RNA interference-mediated transcriptional and post-transcriptional silencing. *Nature Genet.* **36,** 1174–1180 (2004).
37. Sugiyama, T., Cam, H., Verdel, A., Moazed, D. & Grewal, S. I. RNA-dependent RNA polymerase is an essential component of a self-enforcing loop coupling heterochromatin assembly to siRNA production. *Proc. Natl Acad. Sci. USA* **102,** 152–157 (2005).
38. Motamedi, M. R. *et al.* Two RNAi complexes, RITS and RDRC, physically interact and localize to noncoding centromeric RNAs. *Cell* **119,** 789–802 (2004).
39. Irvine, D. V. *et al.* Argonaute slicing is required for heterochromatic silencing and spreading. *Science* **313,** 1134–1137 (2006).
40. Zofall, M. & Grewal, S. I. RNAi-mediated heterochromatin assembly in fission yeast. *Cold Spring Harb. Symp. Quant. Biol.* **71,** 487–496 (2006).
41. Grewal, S. I. & Moazed, D. Heterochromatin and epigenetic control of gene expression. *Science* **301,** 798–802 (2003).
42. Buhler, M., Verdel, A. & Moazed, D. Tethering RITS to a nascent transcript initiates RNAi- and heterochromatin-dependent gene silencing. *Cell* **125,** 873–886 (2006).
43. Djupedal, I. *et al.* RNA Pol II subunit Rpb7 promotes centromeric transcription and RNAi-directed chromatin silencing. *Genes Dev.* **19,** 2301–2306 (2005).
44. Kato, H. *et al.* RNA polymerase II is required for RNAi-dependent heterochromatin assembly. *Science* **309,** 467–469 (2005).
45. Ayoub, N. *et al.* A novel jmjC domain protein modulates heterochromatization in fission yeast. *Mol. Cell. Biol.* **23,** 4356–4370 (2003).
46. Klose, R. J., Kallin, E. M. & Zhang, Y. JmjC-domain-containing proteins and histone demethylation. *Nature Rev. Genet.* **7,** 715–727 (2006).
47. Tsukada, Y. *et al.* Histone demethylation by a family of JmjC domain-containing proteins. *Nature* **439,** 811–816 (2006).
48. Mochizuki, K. & Gorovsky, M. A. Small RNAs in genome rearrangement in *Tetrahymena*. *Curr. Opin. Genet. Dev.* **14,** 181–187 (2004).
49. Vagin, V. V. *et al.* A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313,** 320–324 (2006).
50. Richards, E. J. & Elgin, S. C. Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. *Cell* **108,** 489–500 (2002).
51. Pal-Bhadra, M., Bhadra, U. & Birchler, J. A. RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in *Drosophila*. *Mol. Cell* **9,** 315–327 (2002).
52. Pal-Bhadra, M. *et al.* Heterochromatic silencing and HP1 localization in *Drosophila* are dependent on the RNAi machinery. *Science* **303,** 669–672 (2004).
53. Haynes, K. A., Caudy, A. A., Collins, L. & Elgin, S. C. Element 1360 and RNAi components contribute to HP1-dependent silencing of a pericentric reporter. *Curr. Biol.* **16,** 2222–2227 (2006).
54. van Rij, R. P. *et al.* The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in *Drosophila melanogaster*. *Genes Dev.* **20,** 2985–2995 (2006).
55. Wang, X. H. *et al.* RNA interference directs innate immunity against viruses in adult *Drosophila*. *Science* **312,** 452–454 (2006).
56. Kalmykova, A. I., Klenov, M. S. & Gvozdev, V. A. Argonaute protein PIWI controls mobilization of retrotransposons in the *Drosophila* male germline. *Nucleic Acids Res.* **33,** 2052–2059 (2005).
57. Saito, K. *et al.* Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev.* **20,** 2214–2222 (2006).
58. Deshpande, G., Calhoun, G. & Schedl, P. *Drosophila* argonaute-2 is required early in embryogenesis for the assembly of centric/centromeric heterochromatin, nuclear division, nuclear migration, and germ-cell formation. *Genes Dev.* **19,** 1680–1685 (2005).
59. Allshire, R. C., Nimmo, E. R., Ekwall, K., Javerzat, J. P. & Cranston, G. Mutations derepressing silent centromeric domains in fission yeast disrupt chromosome segregation. *Genes Dev.* **9,** 218–233 (1995).
60. Fukagawa, T. *et al.* Dicer is essential for formation of the heterochromatin structure in vertebrate cells. *Nature Cell Biol.* **6,** 784–791 (2004).
61. Hall, I. M., Noma, K. & Grewal, S. I. RNA interference machinery regulates chromosome dynamics during mitosis and meiosis in fission yeast. *Proc. Natl Acad. Sc.i USA* **100,** 193–198 (2003).
62. Peng, J. C. & Karpen, G. H. H3K9 methylation and RNA interference regulate nucleolar organization and repeated DNA stability. *Nature Cell Biol.* **9,** 25–35 (2007).
63. Jia, S., Noma, K. & Grewal, S. I. RNAi-independent heterochromatin nucleation by the stress-activated ATF/CREB family proteins. *Science* **304,** 1971–1976 (2004).
64. Kim, H. S., Choi, E. S., Shin, J. A., Jang, Y. K. & Park, S. D. Regulation of Swi6/HP1-dependent heterochromatin assembly by cooperation of components of the mitogen-activated protein kinase pathway and a histone deacetylase Clr6. *J. Biol. Chem.* **279,** 42850–42859 (2004).
65. Schultz, D., Ayyanathan, K., Negorev, D., Maul, G. & Rauscher, F. R. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev.* **16,** 1855–1869 (2002).
66. Aulner, N. *et al.* The AT-hook protein D1 is essential for *Drosophila melanogaster* development and is implicated in position-effect variegation. *Mol. Cell. Biol.* **22,** 1218–1232 (2002).
67. Blattes, R. *et al.* Displacement of D1, HP1 and topoisomerase II from satellite heterochromatin by a specific polyamide. *EMBO J.* **25,** 2397–2408 (2006).

68. Huertas, D., Cortes, A., Casanova, J. & Azorin, F. *Drosophila* DDP1, a multi-KH-domain protein, contributes to centromeric silencing and chromosome segregation. *Curr. Biol.* **14,** 1611–1620 (2004).
69. Sun, F. L. *et al. cis*-Acting determinants of heterochromatin formation on *Drosophila melanogaster* chromosome four. *Mol. Cell. Biol.* **24,** 8210–8220 (2004).
70. Aravin, A. A. *et al.* The small RNA profile during *Drosophila melanogaster* development. *Dev. Cell* **5,** 337–350 (2003).
71. Aravin, A. A. *et al.* Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Curr. Biol.* **11,** 1017–1027 (2001).
72. Cheutin, T. *et al.* Maintenance of stable heterochromatin domains by dynamic HP1 binding. *Science* **299,** 721–725 (2003).
73. Festenstein, R. *et al.* Modulation of heterochromatin protein 1 dynamics in primary mammalian cells. *Science* **299,** 719–721 (2003).
74. Nakayama, J., Klar, A. J. & Grewal, S. I. A chromodomain protein, Swi6, performs imprinting functions in fission yeast during mitosis and meiosis. *Cell* **101,** 307–317 (2000).
75. Grishok, A., Sinskey, J. L. & Sharp, P. A. Transcriptional silencing of a transgene by RNAi in the soma of *C. elegans*. *Genes Dev.* **19,** 683–696 (2005).
76. Robert, V. J., Sijen, T., van Wolfswinkel, J. & Plasterk, R. H. Chromatin and RNAi factors protect the *C. elegans* germline against repetitive sequences. *Genes Dev.* **19,** 782–787 (2005).
77. Vastenhouw, N. L. *et al.* Gene expression: long-term gene silencing by RNAi. *Nature* **442,** 882 (2006).
78. Sullivan, B. A. & Karpen, G. H. Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. *Nature Struct. Mol. Biol.* **11,** 1076–1083 (2004).
79. George, J. A. & Pardue, M. L. The promoter of the heterochromatic *Drosophila* telomeric retrotransposon, HeT-A, is active when moved into euchromatic locations. *Genetics* **163,** 625–635 (2003).
80. Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128,** 1089–1103 (2007).
81. Gunawardane, L.S. *et al.* A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science* **315,** 1587–1590 (2007).
82. Rudolph, T. *et al.* Heterochromatin formation in *Drosophila* is initiated through active removal of H3K4 methylation by the LSD1 homolog SU(VAR)3-3. *Mol. Cell* **26,** 103–115 (2007).
83. Zofall, M. & Grewal, S. I. HULC, a histone H2B ubiquitinating complex, modulates heterochromatin independent of histone H3 lysine 4 methylation in fission yeast. *J. Biol. Chem.* **282,** 14065–14072 (2007).

Acknowledgements We thank members of our laboratories for critical reading of the manuscript, and G. Farkas for the design of Figs 1 and 3. Our work is supported by grants from the National Institutes of Health (S.C.R.E.) and National Institutes of Health intramural support (S.I.S.G.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to S.C.R.E. (selgin@biology.wustl.edu) or S.I.S.G. (grewals@mail.nih.gov).

LETTER

doi:10.1038/nature10508

ATM controls meiotic double-strand-break formation

Julian Lange[1], Jing Pan[1]†, Francesca Cole[2], Michael P. Thelen[3], Maria Jasin[2] & Scott Keeney[1,4]

In many organisms, developmentally programmed double-strand breaks (DSBs) formed by the SPO11 transesterase initiate meiotic recombination, which promotes pairing and segregation of homologous chromosomes[1]. Because every chromosome must receive a minimum number of DSBs, attention has focused on factors that support DSB formation[2]. However, improperly repaired DSBs can cause meiotic arrest or mutation[3,4]; thus, having too many DSBs is probably as deleterious as having too few. Only a small fraction of SPO11 protein ever makes a DSB in yeast or mouse[5] and SPO11 and its accessory factors remain abundant long after most DSB formation ceases[1], implying the existence of mechanisms that restrain SPO11 activity to limit DSB numbers. Here we report that the number of meiotic DSBs in mouse is controlled by ATM, a kinase activated by DNA damage to trigger checkpoint signalling and promote DSB repair. Levels of SPO11–oligonucleotide complexes, by-products of meiotic DSB formation, are elevated at least tenfold in spermatocytes lacking ATM. Moreover, *Atm* mutation renders SPO11–oligonucleotide levels sensitive to genetic manipulations that modulate SPO11 protein levels. We propose that ATM restrains SPO11 via a negative feedback loop in which kinase activation by DSBs suppresses further DSB formation. Our findings explain previously puzzling phenotypes of *Atm*-null mice and provide a molecular basis for the gonadal dysgenesis observed in ataxia telangiectasia, the human syndrome caused by ATM deficiency.

SPO11 creates DSBs via a covalent protein–DNA intermediate that is endonucleolytically cleaved to release SPO11 attached to a short oligonucleotide, freeing DSB ends for further processing and recombination[5] (Fig. 1a). SPO11–oligonucleotide complexes are a quantitative by-product of DSB formation that can be exploited to study DSB number and distribution[5–7] (Supplementary Fig. 1). We examined SPO11–oligonucleotide complexes by SPO11 immunoprecipitation and 3′-end labelling of whole-testis extracts from $Atm^{-/-}$ mutant mice, which have multiple catastrophic meiotic defects, including chromosome synapsis failure and apoptosis[8–12]. The $Atm^{-/-}$ phenotype resembles that of mutants lacking DSB repair factors such as DMC1, indicating that absence of ATM causes meiotic recombination defects. Although $Spo11^{-/-}$ mutation is epistatic to $Atm^{-/-}$ (refs 11, 12), the functional relationship between ATM and SPO11 is complex, as meiotic defects of $Atm^{-/-}$ mice are substantially rescued by reducing *Spo11* gene dosage[13,14] (discussed later).

Unexpectedly, we found that adult $Atm^{-/-}$ testes exhibited an approximately tenfold elevation in steady-state levels of SPO11–oligonucleotide complexes relative to wild-type littermates (Fig. 1b) (11.3 ± 4.5-fold, mean and standard deviation, $n = 7$ littermate pairs). This finding contrasts with $Dmc1^{-/-}$ testes, which showed a ~50% reduction in SPO11–oligonucleotide complexes (0.51 ± 0.06-fold relative to wild type, $n = 5$) (Fig. 1c), as previously shown[5,7]. The mutants share similar arrest points in prophase I, as determined by molecular and histological data[12]; thus, increased SPO11–oligonucleotide complexes in $Atm^{-/-}$ spermatocytes are not an indirect consequence of arrest or of an increased fraction of meiocytes harbouring such complexes.

In $Atm^{-/-}$ testes, levels of free SPO11 (that is, not bound to an oligonucleotide) were much lower than in wild type (Fig. 1b). This is not because a large fraction of SPO11 has been consumed in covalent complexes with DNA—which alters its electrophoretic mobility—as free SPO11 was not restored to wild-type levels by nuclease treatment (Fig. 1d). Instead, because *Spo11* transcript levels in wild type are highest in later stages of meiotic prophase[15–18], after the arrest point of $Atm^{-/-}$ cells, reduced free SPO11 is attributable to the lack of later meiotic cell types, consistent with the reduced free SPO11 also found in

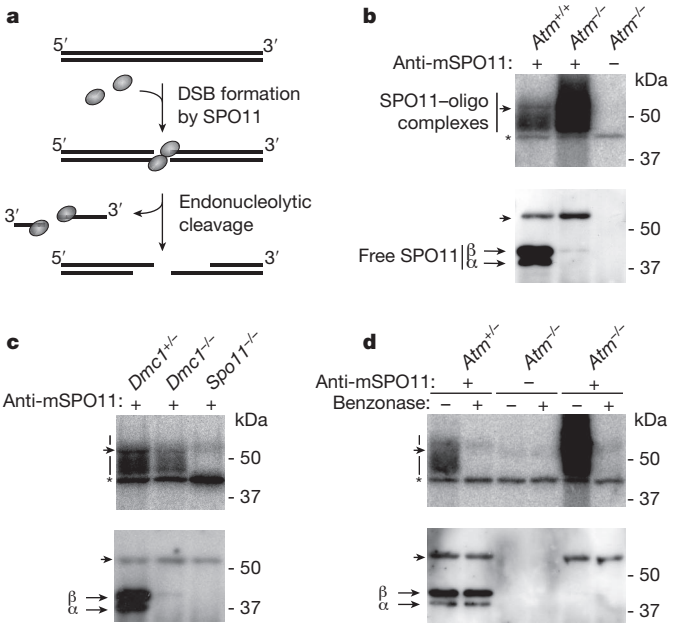


Figure 1 | SPO11 activity and expression in the absence of ATM. **a**, SPO11 attacks the DNA phosphodiester backbone, forming a covalent intermediate with the 5′ strand termini of the DSB. Endonucleolytic cleavage removes SPO11 covalently attached to an oligonucleotide. **b**, **c**, Steady-state levels of SPO11–oligonucleotide (SPO11–oligo) complexes are elevated in $Atm^{-/-}$ testes (**b**), but are decreased in $Dmc1^{-/-}$ testes (**c**). Anti-mSPO11, anti-mouse SPO11 antibody. SPO11 immunoprecipitates from extracts of whole adult testes were treated with terminal transferase and [α-^{32}P] dCTP, resolved by SDS–PAGE, and transferred to a membrane. Representative experiments using littermates of the indicated genotypes are shown. Top, autoradiograph. Bottom, anti-SPO11 western blot detection. Vertical lines, extent of SPO11-specific signals; α and β, major SPO11 isoforms; asterisk, non-specific terminal transferase labelling; arrowheads, migration position of immunoglobulin heavy chain. **d**, Treatment of labelled SPO11 immunoprecipitates with benzonase does not detectably alter levels of free SPO11, but this sequence non-specific nuclease efficiently removes the 3′-end label (compare lanes ± benzonase), and was previously shown to completely remove DNA covalently bound to yeast Spo11 (ref. 1).

[1]Molecular Biology Program, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10065, USA. [2]Developmental Biology Program, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10065, USA. [3]Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California 94550, USA. [4]Howard Hughes Medical Institute, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10065, USA. †Present address: Cell Biology Department, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA.

Dmc1$^{-/-}$ cells (Fig. 1c). As expected, the residual SPO11 protein in *Atm*$^{-/-}$, like *Dmc1*$^{-/-}$, testes was mostly SPO11β (Fig. 1b, c). SPO11α and SPO11β are major protein isoforms encoded by developmentally regulated splice variants; SPO11β is expressed earlier and is sufficient for nearly normal DSB levels[5,15,17–20].

Elevated SPO11–oligonucleotide complexes can be explained by an increased number of meiotic DSBs and/or a longer lifespan of complexes. To distinguish between these possibilities, we examined the initial appearance and persistence of SPO11–oligonucleotide complexes in juvenile mice, in which the first suite of spermatogenic cells proceeds through meiosis in a semi-synchronous fashion[21]. First, we assayed SPO11–oligonucleotide complexes in whole-testis extracts from wild-type C57BL/6J mice at postnatal days (d)7 to 24 (Fig. 2a). SPO11–oligonucleotide complexes first appeared between d9 and d10, when most cells of the initial cohort had entered leptonema. SPO11–oligonucleotide complexes persisted or increased slightly until d15, when the first cohort had progressed into pachynema. Levels rose still further from d16 to d18, coincident with the second cohort of spermatogenic cells reaching leptonema[21]. Thus, SPO11–oligonucleotide complexes appear at the same time as cell types that experience the majority of meiotic DSBs. Consistent with findings in mutants (see earlier), only trace amounts of free SPO11 protein were seen when SPO11–oligonucleotide complexes first appeared, with SPO11β the predominant isoform at these times (Fig. 2a). Importantly, SPO11–oligonucleotide complex levels did not decline between the first and second spermatogenic cohorts. We infer that the lifespan of the complexes is long relative to the duration of prophase, and that an increased lifespan is not a likely explanation for the large increase in steady-state SPO11–oligonucleotides in adult *Atm*$^{-/-}$ testes.

In support of this interpretation, we found that SPO11–oligonucleotide complexes were undetectable in *Atm*$^{-/-}$ testes at d7 (data not shown) but were already elevated 3.3-fold compared with a wild-type littermate when they first appeared, increasing to 8.4-fold over wild type by d12 (Fig. 2b). Because *Atm*$^{-/-}$ juveniles showed higher SPO11–oligonucleotide levels as soon as the first leptotene cells appeared, we conclude that most, if not all, of the increase reflects a greater number of meiotic DSBs occurring during prophase I.

Meiotic defects of mice lacking ATM are substantially suppressed by reducing *Spo11* gene dosage: *Spo11*$^{+/-}$ *Atm*$^{-/-}$ spermatocytes pair and recombine their autosomes and progress through meiotic prophase to metaphase I, where they arrest due to a failure in sex chromosome pairing and recombination[13,14]. The reason for this puzzling rescue was unknown, but our current findings suggest an explanation: the majority of meiotic defects in *Atm*-null spermatocytes are caused by grossly elevated DSB levels, which are lowered by *Spo11* heterozygosity (which reduces SPO11 protein levels by half in adult and juvenile testes (ref. 17 and our unpublished data)). Indeed, we found SPO11–oligonucleotide complexes in *Spo11*$^{+/-}$ *Atm*$^{-/-}$ mice to be substantially reduced compared with *Atm*$^{-/-}$ littermates (Fig. 3a). The remaining increase in SPO11–oligonucleotide complexes in *Spo11*$^{+/-}$ *Atm*$^{-/-}$ mutants compared with wild type (range of 4.5- to 7.8-fold, $n = 2$) is not simply a consequence of metaphase arrest, because SPO11–oligonucleotide complexes were not elevated in mice that exhibit a similar arrest point due to absence of MLH1, a protein involved late in recombination[22] (Fig. 3a). The fact that DSBs are still elevated in *Spo11*$^{+/-}$ *Atm*$^{-/-}$ spermatocytes relative to wild type may account for some or all of the remaining defects in this mutant, including axis interruptions at sites of ongoing recombination and persistent unrepaired DSBs late in prophase I (ref. 14).

Our findings indicate that the absence of ATM renders the extent of DSB formation sensitive to SPO11 expression levels. Therefore, we reasoned that increasing SPO11 expression should further elevate DSB formation in ATM-deficient cells. To test this prediction, we used a previously described transgene (*Xmr-Spo11β*$_B$) that expresses the SPO11β isoform[18]. Indeed, there was substantial further elevation of SPO11–oligonucleotide complex levels (20.9 ± 1.5-fold over wild-type littermates, $n = 3$) upon introduction of this transgene in an *Atm*-null background with intact endogenous *Spo11* (Fig. 3b). By contrast, the transgene resulted in only a modest increase in SPO11–oligonucleotide complexes in an ATM-proficient background (1.1 ± 0.05-fold, $n = 3$) (Fig. 3b).

SPO11–oligonucleotide complexes from *Atm*-null testes were consistently shifted to a higher electrophoretic mobility compared to wild type or other mutants (Figs 1, 2b and 3). To examine the distribution of

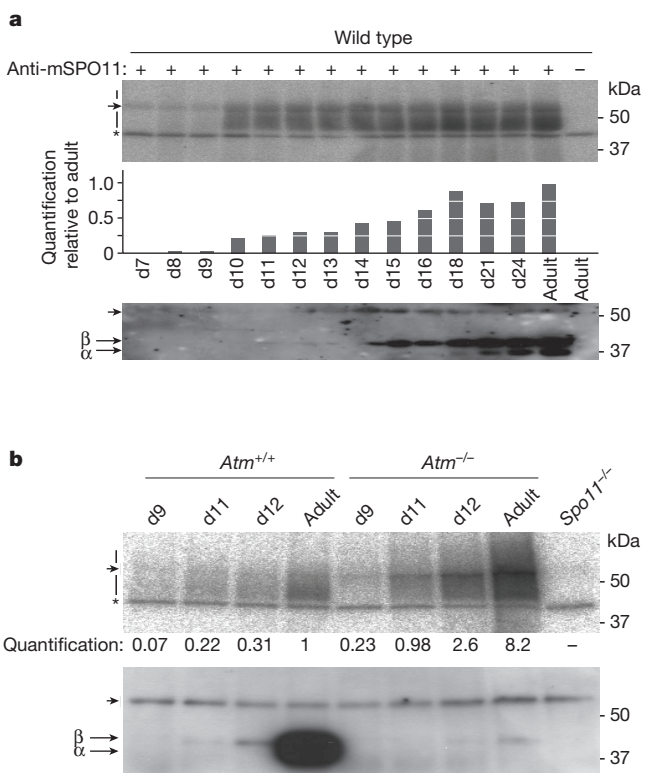


Figure 2 | SPO11–oligonucleotide complexes from juvenile mice. **a**, SPO11–oligonucleotide complexes from testes of wild-type mice from d7–d24. Top, autoradiograph. Middle, quantification. Bottom, anti-SPO11 western detection. **b**, SPO11–oligonucleotides are elevated in testes from juvenile *Atm*$_{-/-}$ mice. Top, autoradiograph. Bottom, anti-SPO11 western detection.

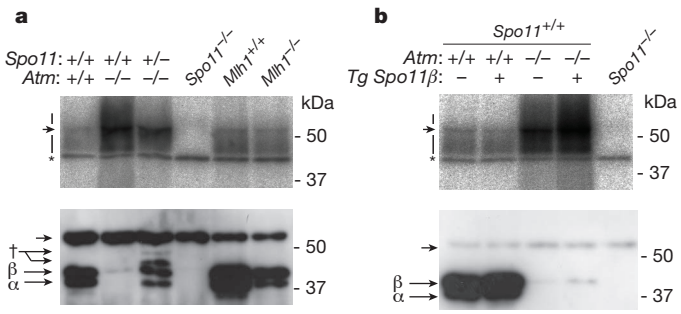


Figure 3 | *Spo11* gene dosage modulates SPO11–oligonucleotide complex levels in *Atm*-deficient spermatocytes. **a**, SPO11–oligonucleotide complexes are reduced in *Spo11*$^{+/-}$ *Atm*$^{-/-}$ testes relative to *Atm*$^{-/-}$, but are more abundant than in wild type or in an *Mlh1*$^{-/-}$ mutant, which similarly arrests at metaphase. Consistent with further meiotic progression than *Atm*$^{-/-}$, both SPO11 isoforms (α and β) are expressed in *Spo11*$^{+/-}$ *Atm*$^{-/-}$ testes, although at reduced levels due to *Spo11* heterozygosity. Dagger, lower-mobility polypeptides probably originating from the *Spo11* knockout allele[18]. **b**, SPO11–oligonucleotide complexes are further elevated by SPO11β expression from the *Xmr-Spo11β*$_B$ transgene (*Tg Spo11β*)[18] in *Atm*$^{-/-}$ spermatocytes. Introducing this transgene into an otherwise wild-type background only modestly increased SPO11–oligonucleotide complex levels.

oligonucleotide lengths, labelled complexes were protease-digested and the resulting oligonucleotides were electrophoresed on a high-resolution gel (Fig. 4a). As previously shown[5], SPO11–oligonucleotides from wild type have a bimodal length distribution with prominent subpopulations at apparent sizes of ~15–27 and ~31–35 nucleotides. *Atm*$^{-/-}$ mice showed a different pattern with or without the *Spo11* transgene: oligonucleotides in the shorter size range were less abundant relative to the ~31–35 nucleotide class and longer oligonucleotides appeared, including an abundant class of ~40–70 nucleotides and a subpopulation that ranged to >300 nucleotides. *Spo11*$^{+/-}$ *Atm*$^{-/-}$ mice showed an intermediate pattern, with more pronounced enrichment of the ~31–35 nucleotide class relative to both smaller and longer oligonucleotides. These results indicate that ATM influences an early step in nucleolytic processing of meiotic DSBs, as has been proposed in yeast[23]. In principle, altered oligonucleotide sizes could reflect changes in preferred positions of the endonucleolytic cleavage that releases the SPO11–oligonucleotide complex, effects on 3′→5′ exonucleolytic digestion of SPO11–oligonucleotides after they are formed, or occurrence of SPO11-induced DSBs at adjacent positions on the same DNA duplex (M. Neale, personal communication). Resection defects and adjacent DSBs (which conventional cytology would be unable to resolve) are both possible explanations for why SPO11–oligonucleotide complexes in *Atm*$^{-/-}$ spermatocytes show a greater increase than RAD51 focus numbers[14].

Our results reveal an essential but previously unsuspected function for ATM in controlling the number of SPO11-generated DSBs. We suggest that activation of ATM by DSBs triggers a negative feedback loop that leads to inhibition of further DSB formation (Fig. 4b) via phosphorylation of SPO11 or its accessory proteins, several of which are known to be phosphorylated in budding yeast (for example, ref. 24) and are conserved in mammals[2]. ATM is activated in the vicinity of DSBs, as judged by SPO11- and ATM-dependent appearance of γH2AX (phosphorylated histone variant H2AX) on chromosomes at leptonema[12,13,25]. Thus, we envision that the negative feedback loop operates at least in part at a local level, perhaps discouraging additional DSBs from forming close to where a DSB has already formed. Such a mechanism could minimize instances where both sister chromatids are cut in the same region, and could also promote more even spacing of DSBs along chromosomes. These studies provide a new molecular framework for understanding the gonadal phenotypes of patients with ataxia telangiectasia[26], which is caused by ATM deficiency[27].

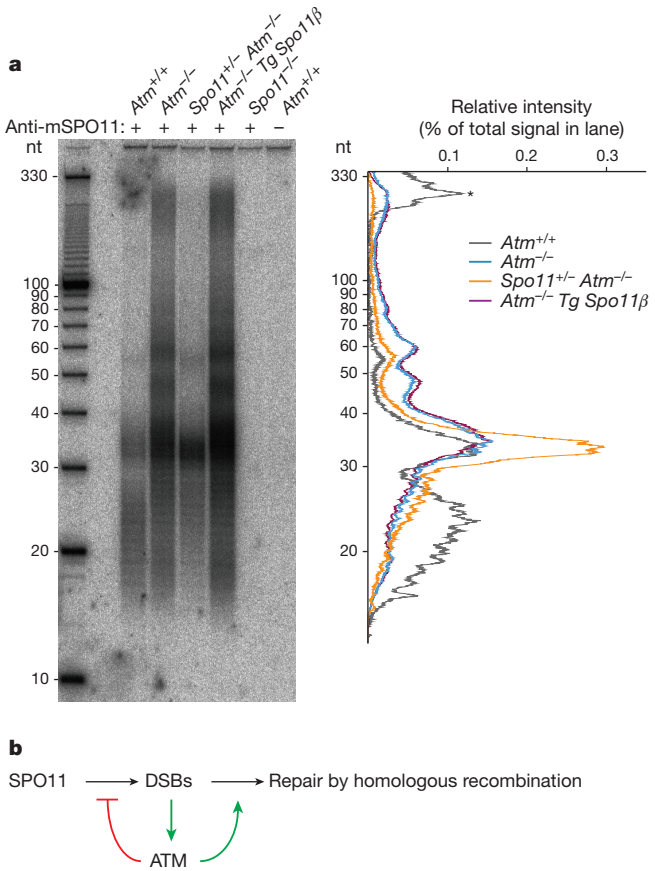


Figure 4 | Roles of ATM in DSB formation and processing. **a**, SPO11–oligonucleotide length distribution is altered in *Atm*$^{-/-}$ spermatocytes. End-labelled SPO11–oligonucleotide complexes were treated with protease to digest the bound protein before electrophoresis on denaturing PAGE. Left, autoradiograph. Right, background-subtracted lane traces normalized to total signal within each lane. Asterisk, autoradiograph background. Each lane contains SPO11–oligonucleotides from the equivalent of different numbers of mice in order to better compare sizes: *Atm*$^{+/+}$, 15 mice; *Atm*$^{-/-}$, 2 mice; *Spo11*$^{+/-}$ *Atm*$^{-/-}$, 4 mice; *Atm*$^{-/-}$ plus transgene, 2 mice; *Spo11*$^{-/-}$, 2 mice; mock, 15 wild-type mice. nt, nucleotides. **b**, Negative feedback loop by which ATM regulates meiotic DSB levels. DSBs generated by SPO11 activate the ATM kinase, inhibiting further DSB formation. ATM may also have roles in repair of DSBs by homologous recombination; for example, by promoting DSB end resection.

METHODS SUMMARY

Mouse mutant alleles and the *Spo11β* transgene were previously described[10,18,28–30]. Experimental animals were compared with controls from the same litter. Experiments conformed to regulatory standards and were approved by the MSKCC Institutional Animal Care and Use Committee. For measurement of SPO11–oligonucleotide complexes, both testes from each mouse were used per experiment, that is, littermate comparisons were made on a per-testis basis (Supplementary Fig. 1). Testis extract preparation, immunoprecipitation and western blot analysis were performed essentially as described[7]. Radiolabelled species were quantified with Fuji phosphor screens and ImageGuage software. The anti-mouse SPO11 monoclonal antibody was produced from hybridoma cell line 180 (M.P.T., unpublished data). The size distribution of SPO11–oligonucleotides was determined essentially as described[5] after radiolabelling with [α-^{32}P] cordycepin. Benzonase treatment of SPO11–oligonucleotide complexes followed manufacturer's instructions (Novagen).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 June; accepted 25 August 2011.
Published online 16 October 2011.

1. Keeney, S. in *Recombination and Meiosis: Crossing-Over and Disjunction* Vol. 2 (ed. Lankenau, D. H.) 81–123 (Springer, 2007).
2. Cole, F., Keeney, S. & Jasin, M. Evolutionary conservation of meiotic DSB proteins: more than just Spo11. *Genes Dev.* **24,** 1201–1207 (2010).
3. Sasaki, M., Lange, J. & Keeney, S. Genome destabilization by homologous recombination in the germ line. *Nature Rev. Mol. Cell Biol.* **11,** 182–195 (2010).
4. Hochwagen, A. & Amon, A. Checking your breaks: surveillance mechanisms of meiotic recombination. *Curr. Biol.* **16,** R217–R228 (2006).
5. Neale, M. J., Pan, J. & Keeney, S. Endonucleolytic processing of covalent protein-linked DNA double-strand breaks. *Nature* **436,** 1053–1057 (2005).
6. Pan, J. *et al.* A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* **144,** 719–731 (2011).
7. Daniel, K. *et al.* Meiotic homologue alignment and its quality surveillance are controlled by mouse HORMAD1. *Nature Cell Biol.* **13,** 599–610 (2011).
8. Barlow, C. *et al.* Atm-deficient mice: a paradigm of ataxia telangiectasia. *Cell* **86,** 159–171 (1996).
9. Xu, Y. *et al.* Targeted disruption of ATM leads to growth retardation, chromosomal fragmentation during meiosis, immune defects, and thymic lymphoma. *Genes Dev.* **10,** 2411–2422 (1996).
10. Barlow, C. *et al.* Atm deficiency results in severe meiotic disruption as early as leptonema of prophase I. *Development* **125,** 4007–4017 (1998).
11. Di Giacomo, M. *et al.* Distinct DNA-damage-dependent and -independent responses drive the loss of oocytes in recombination-defective mouse mutants. *Proc. Natl Acad. Sci. USA* **102,** 737–742 (2005).
12. Barchi, M. *et al.* Surveillance of different recombination defects in mouse spermatocytes yields distinct responses despite elimination at an identical developmental stage. *Mol. Cell. Biol.* **25,** 7203–7215 (2005).
13. Bellani, M. A., Romanienko, P. J., Cairatti, D. A. & Camerini-Otero, R. D. SPO11 is required for sex-body formation, and Spo11 heterozygosity rescues the prophase arrest of *Atm*$^{-/-}$ spermatocytes. *J. Cell Sci.* **118,** 3233–3245 (2005).

14. Barchi, M. *et al.* ATM promotes the obligate XY crossover and both crossover control and chromosome axis integrity on autosomes. *PLoS Genet.* **4,** e1000076 (2008).
15. Keeney, S. *et al.* A mouse homolog of the *Saccharomyces cerevisiae* meiotic recombination DNA transesterase Spo11p. *Genomics* **61,** 170–182 (1999).
16. Shannon, M., Richardson, L., Christian, A., Handel, M. A. & Thelen, M. P. Differential gene expression of mammalian *SPO11/TOP6A* homologs during meiosis. *FEBS Lett.* **462,** 329–334 (1999).
17. Bellani, M. A., Boateng, K. A., McLeod, D. & Camerini-Otero, R. D. The expression profile of the major mouse SPO11 isoforms indicates that SPO11β introduces double strand breaks and suggests that SPO11α has an additional role in prophase in both spermatocytes and oocytes. *Mol. Cell. Biol.* **30,** 4391–4403 (2010).
18. Kauppi, L. *et al.* Distinct properties of the XY pseudoautosomal region crucial for male meiosis. *Science* **331,** 916–920 (2011).
19. Romanienko, P. J. & Camerini-Otero, R. D. Cloning, characterization, and localization of mouse and human *SPO11*. *Genomics* **61,** 156–169 (1999).
20. Romanienko, P. J. & Camerini-Otero, R. D. The mouse *Spo11* gene is required for meiotic chromosome synapsis. *Mol. Cell* **6,** 975–987 (2000).
21. Bellve, A. R. *et al.* Spermatogenic cells of the prepuberal mouse. Isolation and morphological characterization. *J. Cell Biol.* **74,** 68–85 (1977).
22. Eaker, S., Cobb, J., Pyle, A. & Handel, M. A. Meiotic prophase abnormalities and metaphase cell death in MLH1-deficient mouse spermatocytes: insights into regulation of spermatogenic progress. *Dev. Biol.* **249,** 85–95 (2002).
23. Terasawa, M., Ogawa, T., Tsukamoto, Y. & Ogawa, H. Sae2p phosphorylation is crucial for cooperation with Mre11p for resection of DNA double-strand break ends during meiotic recombination in *Saccharomyces cerevisiae*. *Genes Genet. Syst.* **83,** 209–217 (2008).
24. Sasanuma, H. *et al.* Cdc7-dependent phosphorylation of Mer2 facilitates initiation of yeast meiotic recombination. *Genes Dev.* **22,** 398–410 (2008).
25. Mahadevaiah, S. K. *et al.* Recombinational DNA double-strand breaks in mice precede synapsis. *Nature Genet.* **27,** 271–276 (2001).
26. Sedgwick, R. P. & Boder, E. in *Handbook of Clinical Neurology* Vol. 16 (ed. de Jong, J. M. B. V.) 347–423 (Elsevier, 1991).
27. Savitsky, K. *et al.* A single ataxia telangiectasia gene with a product similar to PI-3 kinase. *Science* **268,** 1749–1753 (1995).
28. Edelmann, W. *et al.* Meiotic pachytene arrest in MLH1-deficient mice. *Cell* **85,** 1125–1134 (1996).
29. Pittman, D. L. *et al.* Meiotic prophase arrest with failure of chromosome synapsis in mice deficient for *Dmc1*, a germline-specific RecA homolog. *Mol. Cell* **1,** 697–705 (1998).
30. Baudat, F., Manova, K., Yuen, J. P., Jasin, M. & Keeney, S. Chromosome synapsis defects and sexually dimorphic meiotic progression in mice lacking Spo11. *Mol. Cell* **6,** 989–998 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Neale for discussions, R. Cha and K. McKim for sharing data before publication, and M. Hwang for assistance in monoclonal antibody development. This work was supported by NIH grants HD040916 and HD053855 (to M.J. and S.K.) and GM058673 (to S.K.). J.P. was supported in part by a Leukemia and Lymphoma Society Fellowship and F.C. by a Ruth L. Kirschstein NRSA (F32 HD51392). S.K. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions J.L., J.P. and F.C. performed experiments. M.P.T. generated the anti-SPO11 monoclonal hybridoma line. J.L., M.J., and S.K. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.K. (s-keeney@ski.mskcc.org) or M.J. (m-jasin@ski.mskcc.org).

METHODS

Mouse mutant alleles and the *Spo11β* transgene were previously described[10,18,28–30]. Experiments conformed to regulatory standards and were approved by the MSKCC Institutional Animal Care and Use Committee. For measurement of SPO11–oligonucleotide complexes, both testes from each mouse were used per experiment, that is, littermate comparisons were made on a per-testis basis (Supplementary Fig. 1). The anti-mouse SPO11 monoclonal antibody was produced from hybridoma cell line 180 (M.P.T., unpublished data).

Testis extract preparation, immunoprecipitation and western blot analysis were performed essentially as described[7]. Testes were decapsulated, then lysed in 800 µl lysis buffer (1% Triton X-100, 400 mM NaCl, 25 mM HEPES-NaOH at pH 7.4, 5 mM EDTA). Lysates were centrifuged at 100,000 r.p.m. (355,040*g*) for 25 min in a TLA100.2 rotor. Supernatants were incubated with anti-mouse SPO11 antibody 180 (5 µg per pair of testes) at 4 °C for 1 h, followed by addition of 30–40 µl protein-A–agarose beads (Roche) and incubation for another 3 h. Beads were washed three times with IP buffer (1% Triton X-100, 150 mM NaCl, 15 mM Tris-HCl at pH 8.0). Immunoprecipitates were eluted with Laemmli sample buffer and diluted six- to sevenfold in IP buffer. Eluates were incubated with additional anti-mouse SPO11 antibody 180 at 4 °C for 1 h, followed by addition of 30–40 µl protein-A–agarose beads and incubation at 4 °C overnight. Beads were washed three times with IP buffer and twice with buffer NEB4 (New England BioLabs). SPO11–oligonucleotide complexes were radiolabelled at 37 °C for 1 h using terminal deoxynucleotidyl transferase (Fermentas) and [α-^{32}P] dCTP. Beads were washed three times with IP buffer, boiled in Laemmli sample buffer and fractionated on 8% SDS–PAGE. Complexes were transferred to a PVDF membrane by semi-dry transfer (Bio-Rad). Radiolabelled species were detected and quantified with Fuji phosphor screens and ImageGuage software. For western blot analysis, membranes were probed with anti-mouse SPO11 antibody 180 (1:2,000 in PBS containing 0.1% Tween 20 and 5% non-fat dry milk), then horseradish-peroxidase-conjugated protein A (Abcam; 1:10,000 in PBS containing 0.1% Tween 20 and 5% non-fat dry milk), and detected using the ECL+ reagent (GE Healthcare). The size distribution of SPO11–oligonucleotides was determined by radiolabelling with [α-^{32}P] cordycepin then protease digestion followed by denaturing PAGE. Benzonase treatment of SPO11–oligonucleotide complexes was performed as per manufacturer's instructions (Novagen).

Identification of *MIR390a* precursor processing-defective mutants in Arabidopsis by direct genome sequencing

Josh T. Cuperus[a,b], Taiowa A. Montgomery[a,b], Noah Fahlgren[a,b], Russell T. Burke[b], Tiffany Townsend[b], Christopher M. Sullivan[b,c], and James C. Carrington[b,c,1]

[a]Molecular and Cellular Biology Program; [b]Department of Botany and Plant Pathology; and [c]Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR 97331

Contributed by James C. Carrington, November 16, 2009 (sent for review November 2, 2009)

Transacting siRNA (tasiRNA) biogenesis in Arabidopsis is initiated by microRNA (miRNA) –guided cleavage of primary transcripts. In the case of *TAS3* tasiRNA formation, ARGONAUTE7 (AGO7)–miR390 complexes interact with primary transcripts at two sites, resulting in recruitment of RNA-DEPENDENT RNA POLYMERASE6 for dsRNA biosynthesis. An extensive screen for Arabidopsis mutants with specific defects in *TAS3* tasiRNA biogenesis or function was done. This yielded numerous *ago7* mutants, one *dcl4* mutant, and two mutants that accumulated low levels of miR390. A direct genome sequencing-based approach to both map and rapidly identify one of the latter mutant alleles was developed. This revealed a G-to-A point mutation (*mir390a-1*) that was calculated to stabilize a relatively nonpaired region near the base of the *MIR390a* foldback, resulting in misprocessing of the miR390/miR390* duplex and subsequent reduced *TAS3* tasiRNA levels. Directed substitutions, as well as analysis of variation at paralogous miR390-generating loci (*MIR390a* and *MIR390b*), indicated that base pair properties and nucleotide identity within a region 4–6 bases below the miR390/miR390* duplex region contributed to the efficiency and accuracy of precursor processing.

high-throughput sequencing | miRNA | trans-acting siRNA

Small RNAs, including microRNA (miRNA), several classes of endogenous small interfering RNA (siRNA), and Piwi-associated RNA (piRNA), direct silencing activities that shape transcriptomes and proteomes of eukaryotic organisms. miRNAs arise from transcripts containing self-complementary foldback structures that are initially processed to form 21–22nt miRNA/miRNA* duplexes. In animals, primary transcripts with miRNA foldbacks (pri-miRNA) are processed first by the Microprocessor complex, which contains the RNase III-type protein Drosha and its cofactor Pasha (also known as DGCR8 in humans), then by Dicer, with partners that include the dsRNA-binding domain protein Loquacious (1). Plants orchestrate both pri-miRNA and pre-miRNA processing with the same (or very similar) complex, which includes the RNase-III like enzyme DICER-LIKE1 (DCL1) as the catalytic component (2–5). DCL1 interacts with the dsRNA binding protein HYPONASTIC LEAVES1 (HYL1) and the zinc-finger protein SERRATE (SE), both of which promote efficient and accurate miRNA biogenesis (2, 6–8).

The transacting siRNA (tasiRNA) class represents a specialized type of amplification-dependent siRNA (9, 10). Primary tasiRNA-generating transcripts are first processed by miRNA-guided cleavage (11, 12). Either the 3′ (*TAS1*, *TAS2* and *TAS4* families) or 5′ (*TAS3* family) cleavage product is stabilized and converted to dsRNA by RNA-DEPENDENT RNA POLYMERASE6 (RDR6) (9, 10, 12, 13). Phased, 21-nt siRNAs are generated in register with the miRNA-guided cleavage site through sequential processing by DCL4. Routing of *TAS3* precursor RNA requires two miRNA-guided events, both of which involve AGO7-miR390 complexes (14, 15). Interaction of AGO7-miR390 at a 3′ proximal target site results in primary transcript cleavage, and sets the register for phased siRNA generation. The 3′ cleavage function of AGO7-miR390 is generic, as any of several heterologous miRNA working through AGO1 can substitute for AGO7-miR390 (15). A second miR390 target site at a 5′-proximal position in the processed precursor interacts with AGO7-miR390 in a noncleavage mode (14, 16).

Here, we identify several mutants with defects in *TAS3* tasiRNA biogenesis, including those with defects in the *MIR390a*-derived foldback, revealing a key role for structures near the base of the foldback for efficient and accurate miR390 processing.

Results and Discussion

Screen for *TAS3*-Based syn-tasiRNA–Deficient Mutants. *TAS3a*-based synthetic (syn)-tasiRNAs with complementarity to the *PDS* mRNA provide a visual readout for tasiRNA activity in transgenic Arabidopsis (15). The *35S:TAS3aPDS-1* construct yields tandem syn-tasiRNAs from the 5′ D7[+] and 5′ D8[+] positions in place of siRNA2141 and siRNA2142, also known as tasi-ARFs (11, 12). These repress mRNAs encoding several AUXIN RESPONSE FACTORS, including *ARF3* and *ARF4*, regulation of which is essential for proper developmental timing and lateral organ development (15, 17, 18) (Fig. 1*A*). In wild-type (Col-0) plants expressing *35S:TAS3aPDS-1*, photobleaching emanates from the midrib and major veins, with the phenotype most prominent when viewed from the adaxial side of leaves (Fig. 1*B*) (15). Syn-tasiRNA accumulation and photobleaching are suppressed in plants containing loss-of-function *rdr6-15*, *dcl4-2* and *zip-1* (AGO7-defective) mutations (15).

A screen for mutants with *TAS3* tasiRNA specific defects was done using the syn-tasiRNA line. Besides loss of photobleaching, mutants with *TAS3*-specific defects were predicted to have 1) low or no syn-tasiRNA and endogenous tasi-ARF (siRNA2142), 2) normal levels of *TAS1* tasiRNA (siR255), 3) normal levels of miRNA, such as miR171, that do not function in the *TAS3* pathway, and 4) an accelerated vegetative phase change (AVPC) phenotype, which is associated with loss of *TAS3* tasiRNA (15, 17–20). *TAS3* pathway-specific mutants were not expected to have severe developmental defects, as would be expected for general loss-of-miRNA function mutants (21, 22). The AVPC phenotype is characterized by downward-curled rosette leaves, giving the appearance of a narrow leaf phenotype, and early development of abaxial trichomes (19). In all, 200 pools of seedlings from the M2 generation were screened. A total of 355

Author contributions: J.T.C., T.A.M., and J.C.C. designed research; J.T.C., T.A.M., N.F., R.T.B., and T.T. performed research; N.F. and C.M.S. analyzed data; and J.T.C. and J.C.C. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: carrington@cgrb.oregonstate.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0913203107/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0913203107

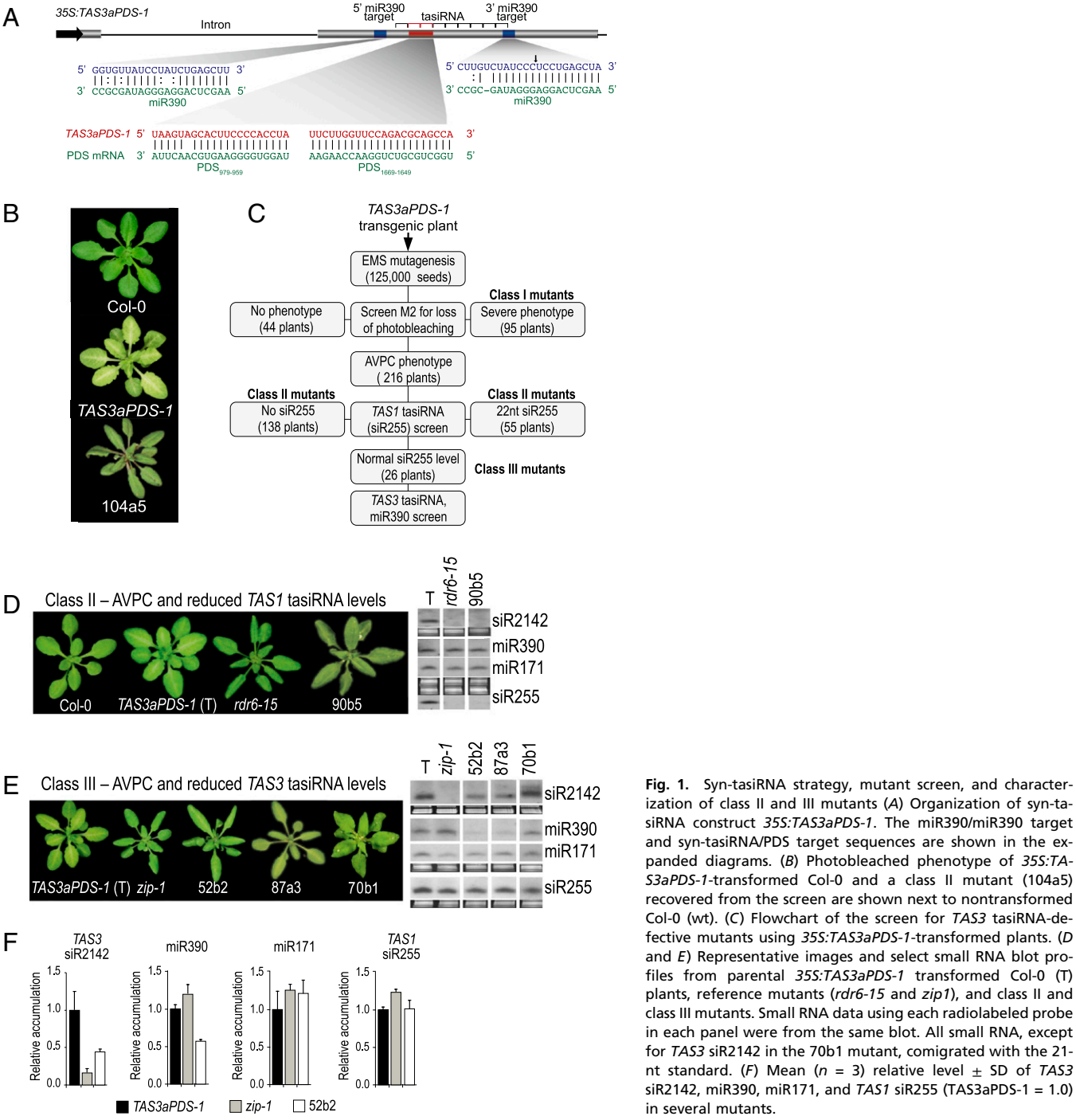


Fig. 1. Syn-tasiRNA strategy, mutant screen, and characterization of class II and III mutants (*A*) Organization of syn-tasiRNA construct *35S:TAS3aPDS-1*. The miR390/miR390 target and syn-tasiRNA/PDS target sequences are shown in the expanded diagrams. (*B*) Photobleached phenotype of *35S:TAS3aPDS-1*-transformed Col-0 and a class II mutant (104a5) recovered from the screen are shown next to nontransformed Col-0 (wt). (*C*) Flowchart of the screen for *TAS3* tasiRNA-defective mutants using *35S:TAS3aPDS-1*-transformed plants. (*D* and *E*) Representative images and select small RNA blot profiles from parental *35S:TAS3aPDS-1* transformed Col-0 (T) plants, reference mutants (*rdr6-15* and *zip1*), and class II and class III mutants. Small RNA data using each radiolabeled probe in each panel were from the same blot. All small RNA, except for *TAS3* siR2142 in the 70b1 mutant, comigrated with the 21-nt standard. (*F*) Mean (n = 3) relative level ± SD of *TAS3* siR2142, miR390, miR171, and *TAS1* siR255 (TAS3aPDS-1 = 1.0) in several mutants.

hygromycin-resistant (transgene-containing) individuals with a reduced-photobleaching phenotype were recovered (Fig. 1*C*).

Of the mutants, 95 had relatively strong or severe vegetative defects, and 44 had no visible morphological defects. Most of the severely affected plants (class I) were dwarfed, had serrated leaves, and resembled known miRNA-debilitated mutants, such as those with *hen1, hyl1, se,* or strong hypomorphic *ago1* alleles (2, 8, 23) (Fig. S1). Among seven class I mutants analyzed, each had reduced levels of miR171 and *TAS1* siR255, indicating that they were generally deficient in miRNA accumulation or activity (Fig. S1). Class I mutants were not analyzed further.

As exemplified by mutant 104a5, 216 mutants had an AVPC phenotype (Fig. 1 *B* and *C*). Approximately 64% (138) or 25% (55) of these mutants lacked the *TAS1* tasiRNA siR255 altogether, or produced siR255-related small RNA that migrated during electrophoresis as a 22-nt RNA, respectively, indicating that they possessed general (*TAS3* and *TAS1*) tasiRNA defects (Fig. 1 *C* and *D*). These 193 plants with general tasiRNA defects comprised class II mutants. Complementation analyses of a random sampling of 15 mutants, including 90b5 (Fig. 1*D*), revealed that the complete loss-of-siR255 subgroup within class II was dominated by *rdr6* (10) and *sgs3* (5) mutants (Table S1). Fourteen mutants produced 21-to-22 nt, size-shifted *TAS1* siR255, and 14 of 14 of these possessed *dcl4* defects based on complementation tests (Table S1). Loss of *DCL4* is known to

result in 22-nt size-shifted tasiRNA, because of the surrogate activity of DCL2 (24–26).

Only 12% (26) of plants possessed an AVPC phenotype, normal levels of 21-nt siR255, and normal levels of miR171 (Fig. 1*C*). Nearly all of these, which were designated as class III mutants, possessed low or undetectable levels of *TAS3* siR2142. Among the class III mutants, complementation analysis revealed 23 independent *ago7* mutants, 14 of which were subjected to *ago7* allele sequencing. Most of the *ago7* alleles contained substitutions affecting the PIWI domain, whereas single mutants with mid-domain or N-terminal domain substitutions were identified (Table S1). A *TAS3*-specific *dcl4* mutant (70b1), in which *TAS3* siR2142-related small RNA, but not *TAS1* siR255, was shifted to 22 nt, was recovered, although minor reductions of both *TAS1* and *TAS3* tasiRNA were noted (Fig. 1*E* and Fig. S2). The 70b1 *dcl4* allele contained a nonconserved Gly-to-Arg substitution affecting a region between the PAZ domain and first RNaseIII domain (Table S1).

Two recessive mutants, 52b2 and 87a3, could not be assigned to any of the complementation groups tested through crosses to *zip1*, *rdr6-15*, *sgs3-11*, and *dcl4-2*. These mutants had similar, moderate AVPC phenotypes (Fig. 1*E*). The 52b2 mutant accumulated significantly reduced levels siR2142 (44.9% compared with Col-0; $P < 0.0028$), but normal levels of miR171 and *TAS1* siR255 (Fig. 1 *E* and *F*). Interestingly, both 52b2 and 87a3 had low levels of miR390 (Fig. 1*E*), with quantitative blot assays revealing a significant difference ($P < 0.0001$) between 52b2 and Col-0 plants (Fig. 1*F*).

Identification of *mir390a-1* by Pooled Genome Sequencing. In principle, direct genome sequencing of a mutant genome using high-throughput sequencing (HTS) technology can identify sites of mutation. However, each EMS-mutagenized genome can possess hundreds or thousands of changes in addition to the causal mutation. We developed a strategy for direct sequencing of a bulk segregant population of genomes for identification of the causative 52b2 mutation. A segregating F2 population from a cross between 52b2 (Col-0 background) and the polymorphic accession Ler was prepared, and 93 homozygous plants with both AVPC and low photobleaching phenotypes were identified. DNA from the 93 plants was pooled and subjected to high-throughput sequencing, which provided 221,000,000 36-base reads.

A pipeline, Mapping and Assembly with Short Sequences (MASS; Fig. 2*A*), was devised to map and assemble sequence data. Approximately 143,000 SNPs (27) were used to identify and quantify Col-0- and Ler-specific reads from repeat-filtered sequences. The ratios of summed Col-0 SNPs/summed Ler SNPs were calculated in 100,000 base windows (20,000 base scroll) across the Arabidopsis genome. A major peak of enriched Col-0 SNPs was identified on chromosome II (Fig. 2*B*). In addition, several minor peaks of Col-0-enriched SNPs were identified around pericentromeric regions. The basis for these minor peaks was not determined conclusively, although the peaks may reflect miscalled SNPs that do not exist in Ler. A 1.52-Mb region encompassing the major Col-0–enriched peak was assembled with the program Mapping and Assembly with Quality (MAQ) (28) using all high-quality sequencing reads, revealing five

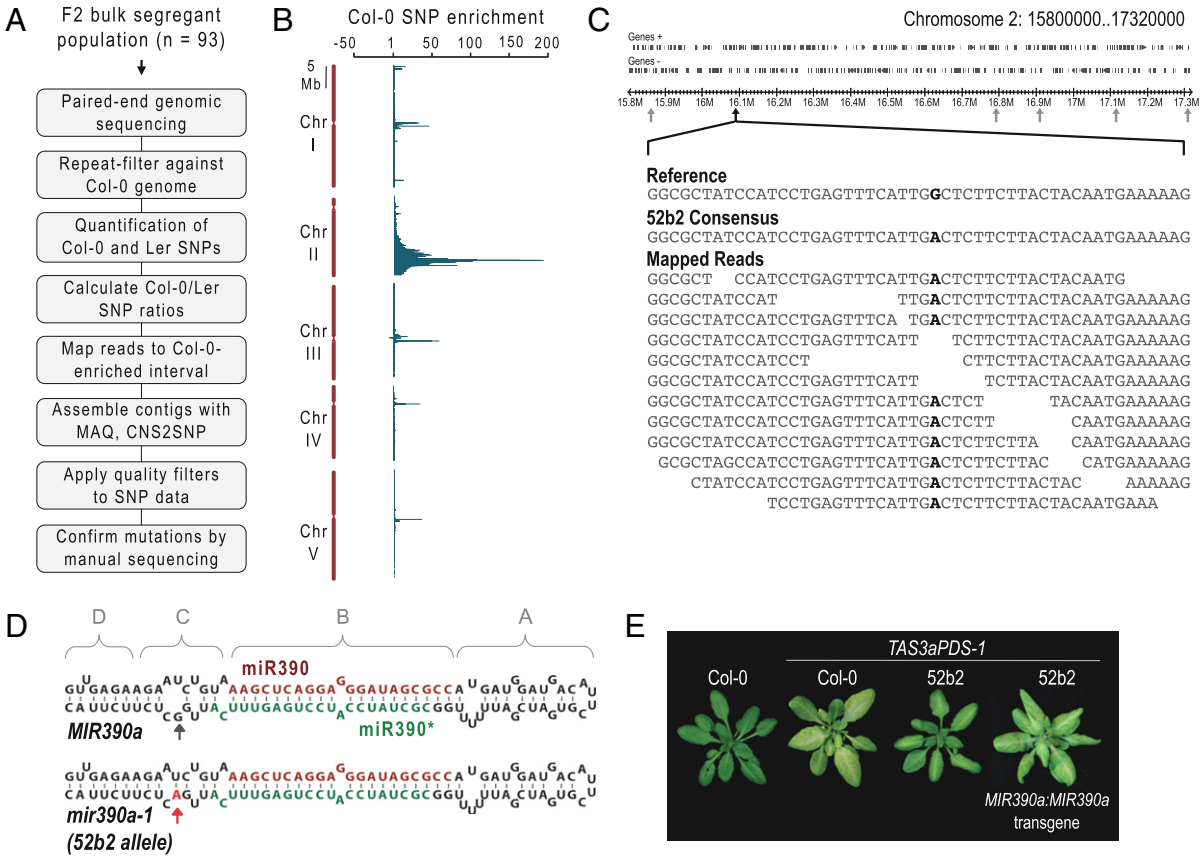


Fig. 2. High-throughput sequencing of the 52b2 mutant genome and identification of the causal mutation (*A*) Flowchart of sequence-based mapping and mutation identification using a bulk segregant population. (*B*) Scrolling window plot of ratios (Col-0/Ler) of total SNPs detected in the bulk segregant sequence dataset. (*C*) A 152-Mb interval spanning the major Col-0–enriched region of chromosome 2 is illustrated. Each nucleotide position that deviates from the reference genome position is indicated by an arrow. The complete or partial sequences of mapped reads from a 50-base segment (chromosome 2 16069100–16069149) from the *MIR390a* locus is shown in the expanded portion. (*D*) Restoration of photobleaching phenotype in 52b2 mutant plants by transformation with a wild-type *MIR390a* transgene. (*E*). Foldback sequence and predicted structure from wild-type *MIR390a* and mutant 52b2 *mir390a-1* alleles. The position corresponding to the mutation is indicated by arrows. For comparative purposes, four foldback domains were assigned, as indicated by the brackets.

G-to-A or C-to-T changes that were consistent with EMS-induced mutation (Table S2). One mutation affected the sequence at the base of the foldback from *MIR390a* (Fig. 2 *C* and *D*). Portions of the 87a3 mutant genome were sequenced manually across the loci corresponding to the G-to-A or C-to-T positions in 52b2, revealing exactly the same *mir390* mutation but wild-type Col-0 sequences at each of the other positions. Thus, 52b2 and 87a3 were independent mutants containing the same *mir390a* allele (*mir390a-1*). A genomic fragment containing wild-type *MIR390a* was introduced into 52b2 mutant plants. This restored photobleaching to the 52b2 mutant line and partially suppressed the leaf curling phenotype (Fig. 2*E*), confirming that the *mir390a-1* mutation was causal.

The sequencing-based approach that identified the *mir390a-1* mutation should be broadly applicable to identification of other markerless (e.g., EMS-induced) mutations. The major benefit of the approach is the simultaneous mapping and sequencing at a genome-wide level. The ability to score all known polymorphisms in individuals from the mapping population affords tremendous marker density, and the MASS pipeline provides a straightforward route to identification of a small number of candidate genes within a relatively small interval of 1–2 Mb. Similar high-throughput sequencing-based approaches for identification of casual mutations were presented recently (29–31).

The *mir390a-1* mutation affects position 94 (G94-to-A94 substitution) from the 5′ end of the predicted foldback. Using both mFOLD and RNAFold (32, 33), G94 in the wild-type sequence was predicted to be nonpaired, or to base pair with U12 with low probability, in the "C region" of the foldback below the miR390/miR390* segment (Fig. 2*D*). In *mir390a-1*, A94 was predicted to base pair with high probability to U12 (Fig. 2*D*). Due to the distance of the mutation away from the miR390/miR390* segment and the variability of this position between *MIR390a* and *MIR390b* (discussed below), this position may have been overlooked in a directed mutagenesis approach to identify precursor processing determinants.

Defective Processing of the *mir390a-1* Foldback. *MIR390a*, *mir390a-1*, and the paralogous *MIR390b* loci specify the identical miR390 sequence, but the foldbacks differ in sequence and predicted base-pair structure. The C region from the *MIR390b* foldback contains more predicted base-paired positions at and adjacent to C112, which occupies the spatially equivalent position as G94 in *MIR390a* (Fig. 3 *A* and *B*). The effects of the *mir390a-1* A94 mutation, as well as the differences in the C region between *MIR390a* and *MIR390b* foldbacks, on miR390 biogenesis and *TAS3* tasiRNA formation, were tested in a transient expression assay using *Nicotiana benthamiana* plants (34).

35S:MIR390a, 35S:mir390a-1, and *35S:MIR390b* were expressed individually to analyze miR390 biogenesis and accumulation, or coexpressed with *35S:TAS3aPDS-2* (syn-tasiRNA) and *35S:HA-AGO7* to test for *TAS3* tasiRNA initiation activity. Compared with *35S:MIR390a*, *35S:mir390a-1* yielded miR390 at 28.3% ($P < 3.02 \times 10^{-5}$) or 28.6% ($P < 0.002$) when expressed individually or with the other *TAS3* tasiRNA components, respectively (Fig. 3*A*), which was consistent with the low levels of *TAS3* tasiRNA detected in the 52b2 mutant plants (Fig. 1 *E* and *F*). Interestingly, *35S:MIR390b* also yielded low levels of miR390 when expressed individually (17.0%, $P < 9.11 \times 10^{-6}$) or with *TAS3* tasiRNA components (19.8%, $P < 0.0019$)(Fig. 3*A*). In addition, the functional amounts of miR390, as reflected by the levels of *TAS3*-based syn-tasiRNA, were significantly lower using *35S:mir390a-1* (21.9%, $P < 0.0072$) and *35S:MIR390b* (33.5%, $P < 0.013$), compared with using *35S:MIR390a* (Fig. 3*A*). These data suggest that processing of the *mir390a-1* and *MIR390b* foldbacks occurs inefficiently.

To analyze processing accuracy of *MIR390a*, *mir390a-1,* and *MIR390b* foldbacks, small RNA libraries from triplicate samples were subjected to high-throughput sequencing analysis after transient expression in *N. benthamiana*. Reads were first normalized based on library size and spike-in standards (35). Reads from within 29-nt windows, centered around the middle of the annotated miR390 or miR390* sequences, were analyzed for size, 5′ position, and 3′ position. The information content of each dataset was used to calculate Shannon's entropy (*H*) (36, 37), providing measures of small RNA uniformity or processing accuracy at both ends of each sequence (Fig. 3 *B* and *C*; *SI Methods*). *MIR390a* yielded predominantly 21-nt, canonical miR390 with highly uniform 5′ and 3′ ends, and moderately heterogeneous 20–21 nt miR390* sequences

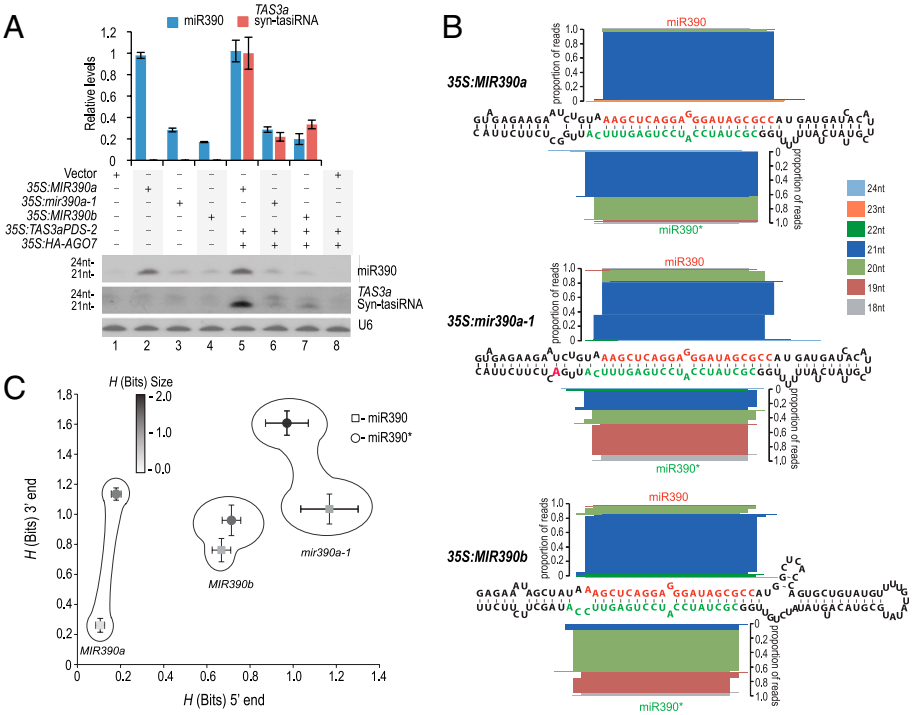


Fig. 3. Foldback processing and *TAS3* tasiRNA-initiation activity of *mir390a-1* in transient assays. (*A*) Accumulation of miR390 and *TAS3* syn-tasiRNA in *N. benthamiana* transient assays. One of three independent replicates is shown. Mean ($n = 3$) relative miR390 (blue) and *TAS3a* (red) syn-tasiRNA levels ± SD (lane 2 and lane 5 = 1.0) are shown. Syn-tasiRNA levels were measured only in assays containing *35S:TAS3aPDS-2* (lanes 5–8). U6 RNA is shown as a loading control. (*B*) Analysis of miR390 and miR390* sequences after transient expression of *MIR390a*, *mir390a-1,* and *MIR390b*. Proportions of reads containing specific sequences are plotted as stacked bars based on size (color coded), 5′ position and 3′ position, with end positions aligned to the respective sequences shown in the foldbacks. miR390 and related sequences are plotted upward, and miR390* sequences are plotted downward. (*C*) Shannon's entropy (*H*) for 5′ end (*x* axis), 3′ end (*y* axis) and size (gray scale) of small RNA populations shown in (*B*). High *H* values reflect high information content, which correlates with variability.

with uniform 5′ ends but with 3′ ends from two major positions (Fig. 3*B*, Fig. S3 and Table S3). In contrast, *35S:mir390a*-1 yielded 5′-, 3′-, and size-heterogeneous miR390 and miR390* sequences, with only 45.6% ± 22.6% of miR390-related sequences containing accurately processed 5′ and 3′ ends (Fig. 3*B* and Fig. S3). This was reflected in high *H* values for each *35S:mir390a*–derived miR390 and miR390* parameter (Fig. 3*C*). *MIR390b* yielded sequences with intermediate processing accuracy. Both ends of miR390, and the 5′ end of miR390*, exhibited more heterogeneity than the comparable ends of sequences from *MIR390a* (Fig. 3 *B* and *C*). Combined with the syn-tasiRNA biogenesis data (Fig. 3*A*), these experimental findings indicate that the *mir390a-1* mutation affects both processing accuracy and efficiency, resulting in low levels of functional miR390. The findings also indicate that *MIR390b* possesses the properties of a low-efficiency mutant allele. Natural variation affecting foldback structure and miRNA biogenesis has been shown previously in plants and animals (38, 39).

Mutational Analysis of the *MIR390a* Foldback. The G-to-A substitution in the *mir390a-1* mutant could conceivably debilitate processing because of a change in foldback base pairing, loss of a base determinant, or both. Computational analysis of predicted foldback variants suggested that the *mir390a-1* structure possessed a higher probability of base pairing between U12 and A94, compared with the probability of pairing between U12 and G94 in the wild-type foldback (Fig. 4*A*). This was reflected in a lower calculated entropy at both positions in the *mir390a-1* foldback (Fig. 4*A*) (36, 37). The *MIR390b* predicted foldback, with even more extensive base pairing, yielded lower calculated positional entropies at nearly all bases in region C (Fig. 4*A*) (36, 37). Seven *35S:MIR390a* mutants with substitutions at either position 94 and/or position 12 were constructed (Table S4). Including *mir390a-1*, the series resulted in foldbacks containing all possible single-base substitutions at both positions, and two combinations of dual-base substitutions (Fig. 4*A*). In addition, the sequences comprising *MIR390a* region C were substituted for the approximate equivalent sequences from *MIR390b*. Predicted foldback structures, positional entropies and miR390 biogenesis levels in a transient assay were determined.

Each substitution at position 94 (*mir390a-1, mir390a-94U,* and *mir390a-94C*) resulted in significantly ($P < 0.003$) lower miR390 levels compared with wild-type *MIR390a*, although the *mir390a-94U* and *mir390a-94C* defects were only modest (Fig. 4*B*). Unlike *mir390a-1*, *mir390a-94U,* and *mir390a-94C* mutations were not predicted to base pair with U12 (Fig. 4*A*). These data generally reinforce a role for G94, as either a single base-determinant or a high-entropy, weak–base-pair partner with U12. Among the position 12 substitutions, *mir390-12C* was significantly ($P < 8.8\ 10^{-9}$) debilitated for miR390 biogenesis and was predicted to form a low-entropy base pair with G94. *mir390a-12A* retained both A12 and G94 in a predicted nonpaired configuration and yielded wild-type levels of miR390 (Fig. 4). These position 12 mutants lend support to the idea that a nonpaired or weakly paired G94 contributes to miR390 biogenesis. In contrast, *mir390a-12G* was predicted to adopt a fold involving low-entropy, highly base-paired 12G and G94 positions, but led to wild-type levels of miR390 (Fig. 4). However, the *mir390a-12G* local stem structure was predicted to include novel, high-entropy asymmetric bulges that differed from the comparable positions from *MIR390a* (Fig. 4*A*).

Among the double mutants, *mir390a-12C94A* contained the A94 mutation from *mir390a-1* and a base-pair–disrupting change at position 12 (Fig. 4*A*). This mutant was highly debilitated for miR390 biogenesis, indicating that the *mir390a-1* defect (A94) was not due solely to the increased base pair configuration between positions 12 and 94 (Fig. 4). Interestingly, the *mir390a-12G94U* mutant foldback, which contained the G and U positions from wild-type *MIR390a* reversed, yielded nearly wild-type levels of miR390 (Fig. 4).

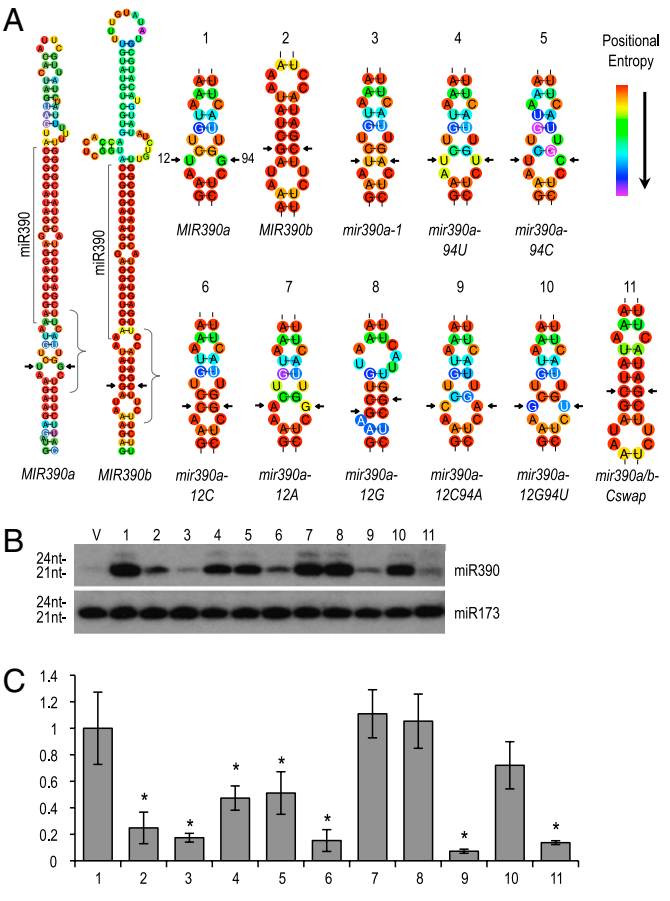


Fig. 4. Directed mutational analysis of *MIR390a* foldback (*A*) Predicted base pair structure of *MIR390a* and *MIR390b* foldbacks (*Left*). Enlarged region corresponds to the bracketed region for 11 mutant or variant foldbacks (*Right*). Positions 12 (U in *MIR390a*) and 84 (G in *MIR390a*) are indicated by arrows. Positional entropy values range from 0 (red) to 1.6 (purple) for all *MIR390a*-based foldbacks and 1.8 (purple) for *MIR390b*. Folding and Shannon's entropy values reflect the probability of variant base-pair states and were calculated using RNAfold. (*B*) Blot assays for miR390 derived from expression of *35S:MIR390a* (lane 1), *35S:MIR390b* (lane 2), *35S:mir390a-1* (lane 3), and each directed mutant construct (lanes 4–11) are shown, along with a negative control sample expressing empty vector (V). One of six independent replicates is shown. U6 RNA is shown as a loading control. (*C*) Mean relative miR390 levels ± SD (*35S:MIR390a* = 1.0).

Substitution of the base of the *MIR390a* stem with that from *MIR390b* led to significant ($P < 6.73\ 10^{-6}$) debilitation of miR390 formation (Fig. 4 *B* and *C*). The mutant foldback region C was predicted to contain the same low-entropy, highly base-paired configuration as predicted for *MIR390b* foldback (Fig. 4*A*). However, the functional significance of the *MIR390b* locus remains unclear.

The relatively high diversity of sequences, sizes, and secondary structures of plant *MIRNA* foldbacks (40) means that processing determinants are not particularly obvious. Based on in vitro processing assays with *MIR167b* foldbacks, DCL1 is sufficient to catalyze ATP-dependent pri- and premiRNA transcript processing, although only a minority of such products possess accurate 5′ and 3′ ends (6). The dsRNA binding motifs of DCL1 may provide a basal function for foldback recognition. However, inclusion of both SE and HYL1 in these reactions increases the rate and accuracy of processing (6, 41). This may indicate that SE and HYL1 function as accessory factors that position DCL1 accurately on substrates through interaction with one or more foldback structural features. We propose that the inaccurate and inefficient processing of the *mir390a-1* foldback is due to loss of interaction with key factors

promoting miRNA biogenesis. In particular, it is attractive to consider G94 in a flexible, high-entropy context as a recognition determinant for HYL1 and/or SE. Both SE and HYL1 promote miR390 accumulation in vivo (23, 42). Importantly, the effects of the *mir390a-1* mutation on foldback processing in transient assays are very similar to the effects of *MIR167b* foldback processing in the absence of SE and HYL1 in vitro (6). It seems unlikely, however, that foldback position G94 is the sole determinant for such interactions, as there is high sequence and structural diversity at this position among foldbacks from conserved *MIRNA* families. By analogy with the Drosha-Pasha/DGCR8 complex interacting with the base of animal foldbacks (43), features defining the junction between the base of the stem and the nonpaired region outside of the stem may also interact with the DCL1-HYL1-SE complex for positioning of the first set of cuts at the proximal end of the miRNA/miRNA* duplex. Indeed, Mateos et al. (44), Song et al. (45), and Werner et al. (46) revealed a key role for a single-stranded/base-paired stem junction ~15 nucleotides from the miRNA/miRNA* duplex in accurate processing of many Arabidopsis miRNAs.

Methods

References for *rdr6-15, dcl4-2, sgs3-11, hen1-1, hyl1-2, se-2, hst-15, dcl1-7, and zip-1* alleles were described (12). Detailed descriptions, protocols, and references for transgenic plant materials, RNA blot assays, transient expression assays in *N. benthamiana*, Arabidopsis mutagenesis and genetic screen, sequencing and analysis of small RNA populations, and the MASS pipeline are provided in *SI Methods*.

ACKNOWLEDGMENTS. We thank Sarah Dvorak for assistance with the genetics screen and technical assistance. We also thank Detlef Weigel, Korbinian Schneeberger, and Richard Clark for helpful discussion, genomic sequencing test data sets, and SNP data; and we thank Nina Fedoroff, Mike Axtell, Detlef Weigel, and Javier Palatnik for sharing data and manuscripts prior to publication. This work was supported by grants from the National Science Foundation (MCB-0618433), National Institutes of Health (AI43288) and United States Department of Agriculture–National Research Initiative (2006-35301-17420).

1. Carthew RW, Sontheimer EJ (2009) Origins and mechanisms of miRNAs and siRNAs. *Cell* 136:642–655.
2. Kurihara Y, Takashi Y, Watanabe Y (2006) The interaction between DCL1 and HYL1 is important for efficient and precise processing of pri-miRNA in plant microRNA biogenesis. *RNA* 12:206–212.
3. Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP (2002) MicroRNAs in plants. *Genes Dev* 16:1616–1626.
4. Golden TA, et al. (2002) SHORT INTEGUMENTS1/SUSPENSOR1/CARPEL FACTORY, a Dicer homolog, is a maternal effect gene required for embryo development in Arabidopsis. *Plant Physiol* 130:808–822.
5. Park W, Li J, Song R, Messing J, Chen X (2002) CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in Arabidopsis thaliana. *Curr Biol* 12:1484–1495.
6. Dong Z, Han MH, Fedoroff N (2008) The RNA-binding proteins HYL1 and SE promote accurate in vitro processing of pri-miRNA by DCL1. *Proc Natl Acad Sci USA* 105:9970–9975.
7. Hiraguri A, et al. (2005) Specific interactions between Dicer-like proteins and HYL1/DRB-family dsRNA-binding proteins in Arabidopsis thaliana. *Plant Mol Biol* 57:173–188.
8. Yang L, Liu Z, Lu F, Dong A, Huang H (2006) SERRATE is a novel nuclear regulator in primary microRNA processing in Arabidopsis. *Plant J* 47:841–850.
9. Vazquez F, et al. (2004) Endogenous trans-acting siRNAs regulate the accumulation of *Arabidopsis* mRNAs. *Mol Cell* 16:69–79.
10. Peragine A, Yoshikawa M, Wu G, Albrecht HL, Poethig RS (2004) SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in *Arabidopsis*.. *Genes Dev* 18:2368–2379.
11. Yoshikawa M, Peragine A, Park MY, Poethig RS (2005) A pathway for the biogenesis of trans-acting siRNAs in *Arabidopsis*. *Genes Dev* 19:2164–2175.
12. Allen E, Xie Z, Gustafson AM, Carrington JC (2005) microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell* 121:207–221.
13. Rajagopalan R, Vaucheret H, Trejo J, Bartel DP (2006) A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev* 20:3407–3425.
14. Axtell MJ, Jan C, Rajagopalan R, Bartel DP (2006) A two-hit trigger for siRNA biogenesis in plants. *Cell* 127:565–577.
15. Montgomery TA, et al. (2008) Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. *Cell* 133:128–141.
16. Howell MD, et al. (2007) Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. *Plant Cell* 19:926–942.
17. Fahlgren N, et al. (2006) Regulation of AUXIN RESPONSE FACTOR3 by TAS3 ta-siRNA affects developmental timing and patterning in Arabidopsis. *Curr Biol* 16:939–944.
18. Garcia D, Collier SA, Byrne ME, Martienssen RA (2006) Specification of leaf polarity in Arabidopsis via the trans-acting siRNA pathway. *Curr Biol* 16:933–938.
19. Poethig RS (2003) Phase change and the regulation of developmental timing in plants. *Science* 301:334–336.
20. Adenot X, et al. (2006) DRB4-dependent TAS3 trans-acting siRNAs control leaf morphology through AGO7. *Curr Biol* 16:927–932.
21. Clarke JH, Tack D, Findlay K, Van Montagu M, Van Lijsebettens M (1999) The SERRATE locus controls the formation of the early juvenile leaves and phase length in Arabidopsis. *Plant J* 20:493–501.
22. Jacobsen SE, Running MP, Meyerowitz EM (1999) Disruption of an RNA helicase/RNAse III gene in Arabidopsis causes unregulated cell division in floral meristems. *Development* 126:5231–5243.
23. Montgomery TA, et al. (2008) AGO1-miR173 complex initiates phased siRNA formation in plants. *Proc Natl Acad Sci USA* 105:20055–20062.
24. Dunoyer P, Himber C, Voinnet O (2005) DICER-LIKE 4 is required for RNA interference and produces the 21-nucleotide small interfering RNA component of the plant cell-to-cell silencing signal. *Nat Genet* 37:1356–1360.
25. Gasciolli V, Mallory AC, Bartel DP, Vaucheret H (2005) Partially redundant functions of *Arabidopsis* DICER-like enzymes and a role for DCL4 in producing trans-acting siRNAs. *Curr Biol* 15:1494–1500.
26. Xie Z, Allen E, Wilken A, Carrington JC (2005) DICER-LIKE 4 functions in trans-acting small interfering RNA biogenesis and vegetative phase change in *Arabidopsis thaliana*.. *Proc Natl Acad Sci USA* 102:12984–12989.
27. Clark RM, et al. (2007) Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science* 317:338–342.
28. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18:1851–1858.
29. Sarin S, Prabhu S, O'Meara MM, Pe'er I, Hobert O (2008) Caenorhabditis elegans mutant allele identification by whole-genome sequencing. *Nat Methods* 5:865–867.
30. Blumenstiel JP, et al. (2009) Identification of EMS-induced mutations in Drosophila melanogaster by whole-genome sequencing. *Genetics* 182:25–32.
31. Schneeberger K, et al. (2009) SHOREmap: Simultaneous mapping and mutation identification by deep sequencing. *Nat Methods* 6:550–551.
32. Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* 9:133–148.
33. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31:3406–3415.
34. Llave C, Xie Z, Kasschau KD, Carrington JC (2002) Cleavage of Scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA. *Science* 297:2053–2056.
35. Fahlgren N, et al. (2009) Computational and analytical framework for small RNA profiling by high-throughput sequencing. *RNA* 15:992–1002.
36. Shannon CE (1948) A mathematical theory of communication. *Bell Sys Tech J* 27: 379–423623–656.
37. Schneider TD (1997) Information content of individual genetic sequences. *J Theor Biol* 189:427–441.
38. de Meaux J, Hu JY, Tartler U, Goebel U (2008) Structurally different alleles of the ath-MIR824 microRNA precursor are maintained at high frequency in Arabidopsis thaliana. *Proc Natl Acad Sci USA* 105:8994–8999.
39. Sun G, et al. (2009) SNPs in human miRNA genes affect biogenesis and function. *RNA* 15:1640–1651.
40. Axtell MJ (2008) Evolution of microRNAs and their targets: Are all microRNAs biologically relevant? *Biochim Biophys Acta* 1779:725–734.
41. Han MH, Goud S, Song L, Fedoroff N (2004) The Arabidopsis double-stranded RNA-binding protein HYL1 plays a role in microRNA-mediated gene regulation. *Proc Natl Acad Sci USA* 101:1093–1098.
42. Chitwood DH, et al. (2009) Pattern formation via small RNA mobility. *Genes Dev* 23: 549–554.
43. Kim VN, Han J, Siomi MC (2009) Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* 10:126–139.
44. Mateos JL, Bologna NG, Chorostecki U, Palatnik J (2010) Identification of structural determinants for microRNA processing in plants by random mutagenesis of MIR172a precursor. *Curr Biol*, in press.
45. Song L, Axtell MJ, Fedoroff NV (2010) RNA secondary structural determinants of miRNA precursor processing in Arabidopsis. *Curr Biol*, in press.
46. Werner S, Wollman H, Weigel D (2010) Sequence determinants for accurate processing of miR172a in *Arabidopsis thaliana*. *Curr Biol*, in press.

Supporting Information

Cuperus et al. 10.1073/pnas.0913203107

SI Methods

Construction of Transgenes. Transgene sequences were PCR-amplified from genomic DNA. Constructs yielding syn-tasiRNA and miRNA were generated by site-overlap extension as previously described (1) and introduced into the vectors pMDC32, pGWB2, or pGWB1 (2, 3). Modified *MIR390* constructs were designed as described (4). These were introduced into pENTR/D-TOPO (Invitrogen) and subsequently recombined with LR clonase (Invitrogen) into pMDC32 (2).

Conventional Sequencing. Sequencing using the Sanger method was done using the National Center for Biotechnology Information (NCBI) primer design tool (available at http://www.ncbi.nlm.nih.gov/tools/primer-blast/). For sequencing EMS-induced mutations in Arabidopsis genomic DNA, two independent PCR products were gel purified and sequenced along with fragments from a nonmutagenized parental control.

RNA-Blot Assays. RNA was isolated using TRIzol reagent (Invitrogen). Two chloroform extractions were done and RNA was precipitated in an equal volume of isopropanol for 20 min. RNA blot assays were performed as described (5). Briefly 5, 10, or 20 μg of total normalized RNA was resolved by denaturing PAGE. RNA was transferred to positively charged nitrocellulose membranes. DNA or LNA probes were end-labeled using [^{32}]ATP and Optikinase (USB). Probes were hybridized to RNA on membranes at 38–50° C. Quantification of small RNA blot hybridization intensities was done using an Instant Imager (Packard Bioscience) and normalized relative to the parental *TAS3aPDS-1* transformed control or appropriate transient assay control.

Genomic Sequencing. A bulk segregant population (F2) from a cross between the 52b2 mutant and the polymorphic parent Ler was generated. Homozygous mutant plants (92 individuals) were inferred based on phenotype, and DNA was isolated from each and pooled. Sequence analysis was done with an Illumina Genome Analyzer I (GA I). Seven lanes of a paired-end flow cell were used. The Illumina genomic DNA sample preparation protocol was followed with modifications. After the PCR amplification step, additional gel purification was done to remove adapter–adapter product. Sequencing and base calling were done according to the manufacturer's recommendations, resulting in 221.5 million independent reads.

MASS. Reads from the bulk segregant population were mapped to the Arabidopsis (Col-0, TAIR8) genome using Cache Assisted Hash Search using XOR logic (CASHX) (6), resulting in ~12× average coverage for perfect-match reads (1.6 GB). Using 143,508 available SNPs (6), a database of 71-bp sequences centered on each SNP (Col-0 vs. Ler) was created. When 71mers overlapped, they were joined into one larger database entry. Illumina 1G reads were aligned to entries in the database using CASHX. Reads that hit Col-0 or Ler SNPs were summed in 100,000-bp windows, using a 20,000-bp scroll, and ratios were calculated. These ratios were plotted using R and visualized (Fig. 2*B*). Illumina reads that aligned with up to two mismatches to ChrII:15800000–17320000 were parsed using Short Oligonucleotide Analysis Package (SOAP) (7). Using the MAQ program easyrun (8), 967,616 sequences (with their Illumina-based quality scores) that mapped with two mismatches or less to the 1.5-MB interval were assembled. An A-to-G difference at genome coordinate ChrII:16766679 was detected, but this was due to a bona fide difference between the reference and initially mutagenized genome. Four of the G-to-A mutations were sequenced using the Sanger method and confirmed as post-EMS specific in the 52b2 mutant.

The MASS package contains scripts to run CASHX, SOAP and MAQ, and is available for download (http://jcclab.science.oregonstate.edu/MASS). In addition to the MASS mapping and alignment tools, the MASS package contains the entire pipeline used to identify the *mir390a-1* mutation. It includes programs for creating plots of SNP enrichment, alignment with MAQ, and filtering of SNPs. MASS is designed to take any indicated read length and to create an appropriate database of sequences centered on a SNP nucleotide, forcing each read to align across the SNP site. The MASS pipeline filters the SNP data set (cns.snp) from the MAQ output. Using Illumina quality scores, data are filtered based on the following criteria: consensus base is a true base; a phred-like quality score of 43; a minimum read depth of 5; a maximum read depth of 50; and no second-best base call. The phred-like quality score is based on Illumina quality scores. In part, these filtering values are based on ~12× coverage; quality scores and read depth may be adjusted based on coverage, read length, and quality of reads.

Small RNA Sequencing from Transient Expression Assays. Small RNA amplicons were prepared in triplicate as described (9). Four synthetic oligoribonucleotides (Std2, Std3, and Std6 [see ref. 9] and Std11 [pUGUCCGACACGAUGCAGAUCC]) were added to 40 μg total RNA per sample before amplicon preparation at four concentrations (Std11, 0.0001 pmol; Std6, 0.001 pmol; Std3, 0.01 pmol; Std2, 0.1 pmol). In addition, samples were barcoded using four variants of the standard 5′ adaptor (5′GUUCAGAGUUCUACAGUCCGACGAUAAC3′ [barcode A], 5′GUUCAGAGUUCUACAGUCCGACGAUCCC3′ [barcode C], 5′GUUCAGAGUUCUACAGUCCGACGAUGGC3′ [barcode G], and 5′GUUCAGAGUUCUACAGUCCGACGAUUUC3′ [barcode U], barcoded sequence underlined) and multiplexed. Sequencing by synthesis was done with an Illumina Genome Analyzer I (GAI). Multiplexed amplicons (four samples, 2.5 pmol total) were added per lane. Reads were computationally parsed based on detection of the 5′ barcode (AAC, CCC, GGC, or TTC) and the first six nucleotides of the 3′ adapter (CTGTAG). Read proportions were based on total reads (18–24 nts) that matched perfectly within a 29-base window surrounding the annotated miR390 or miR390* sequences from *MIR390a* and *MIR390b*. Control samples to measure the low levels of endogenous miR390-related sequences in *N. benthamiana* leaves were prepared after transient expression of *35S:GUS* (Table S4). Sequence and size information content from miR390-related sequences recovered after the transient assays was analyzed using Shannon's entropy formula (10). Calculations were done independently for 5′ end, 3′ end and size.

Mutagenesis of *MIR390a*. Six oligos were used to create *MIR390a* substitution constructs. Briefly, 5′ and 3′ fragments that partially overlapped the *MIR390a* foldbacks were amplified with KOD polymerase (Novagen) using a plasmid containing *35S:MIR390a* as template. The *MIR390a* genomic primers flanked the miR390 sequence by 250 nucleotides on each side:

MIR390a F [caccTATAGGGGGGAAAAAAGGTAG]
MIR390a R [GAGACTAAAGATGAGATCTA]
MIR390b F 5′ [caccTTCCAAAATATGTAATATGGGGA]

MIR390b R 5′ [CTAACAAACTGCTTAGATGTGTGAA].

Sequences CACC on forward primers are not genomic, but were added for cloning into pENTR/D-TOPO (Invitrogen). Forward primers were combined with RI primers, while reverse flanking primers were combined with F2 primers (Table S3). Fragments were gel purified, then mixed with loop containing fragments, or overlapping oligos containing loop sequence, in a second round of PCR. Round 2 PCR fragments were gel purified.

Entropy Calculations. Shannon's entropy formula (10) was used to quantify the diversity of reads mapping to the miRNA or miRNA* from small RNA amplicons. Similarly, Shannon's entropy formula is used by RNAfold as a measurement of the diversity of base pair probability of each position in RNA secondary structures. Shannon's entropy (*H*) will vary from zero, where only one specific miRNA position, size, or type of base pairing is possible, up to $\log_2(x)$ where all positions, size, or type of base pairing occur at the same frequency.

RNA Folding. Computational analysis of foldback sequences shown in Fig. 4*A* was done using RNAfold (11) with the following options: -p -T 22 -d2. Shannon's entropy (10) at each position was calculated using the program RNApdist.pl by summing entropy values for all base pairing probabilities, as well as the probability of not base pairing (1 − sum of base-pair probabilities). Color coding of entropy values was done using the program relplot.pl. All programs are part of the Vienna package (12).

1. Ho SN, Hunt HD, Horton RM, Pullen JK, Pease LR (1989) Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene* 77:51–59.
2. Curtis MD, Grossniklaus U (2003) A gateway cloning vector set for high-throughput functional analysis of genes in planta. *Plant Physiol* 133:462–469.
3. Nakagawa T, et al. (2007) Development of series of gateway binary vectors, pGWBs, for realizing efficient construction of fusion genes for plant transformation. *J Biosci Bioeng* 104:34–41.
4. Montgomery TA, et al. (2008) Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. *Cell* 133:128–141.
5. Montgomery TA, et al. (2008) AGO1-miR173 complex initiates phased siRNA formation in plants. *Proc Natl Acad Sci USA* 105:20055–20062.
6. Clark RM, et al. (2007) Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science* 317:338–342.
7. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: Short Oligonucleotide Alignment Program. *Bioinformatics* 24:713–714.
8. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18:1851–1858.
9. Fahlgren N, et al. (2009) Computational and analytical framework for small RNA profiling by high-throughput sequencing. *RNA* 15:992–1002.
10. Shannon CE (1948) A mathematical theory of communication. *Bell Sys Tech J* 27: 379–623423–656.
11. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31:3406–3415.
12. Bonhoeffer S, McCaskill JS, Stadler PF, Schuster P (1993) RNA multi-structure landscapes. A study based on temperature dependent partition functions. *Eur Biophys J* 22:13–24.

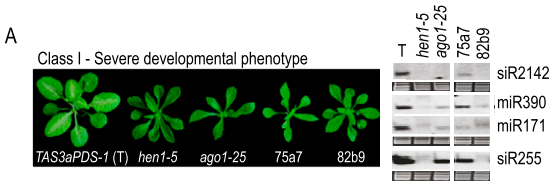


Fig. S1. Characterization of class I mutants (*A*) Representative images and select small RNA blot profiles from parental *35S:TAS3aPDS-1* transformed Col-0 (T) plants, reference mutants (*hen1-5* and *ago1-25*), and class I mutants. Small RNA data using each radiolabeled probe in each panel were from the same blot.

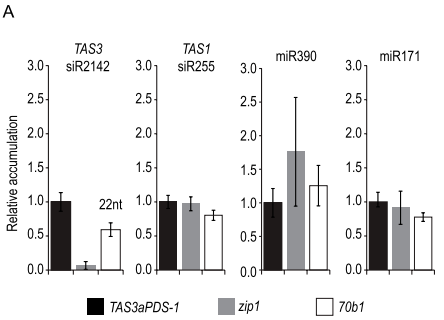


Fig. S2. Quantification of small RNA accumulation in the *dcl4* mutant 70b1. (*A*) Mean (n = 3) relative level ± SD of *TAS3* siR2142, *TAS1* siR255, miR171, and miR390 (TAS3aPDS-1 = 1.0).

A

***35S:MIR390a* miR390**

	Small RNA Size Class																				
	18			19			20			21			22			23			24		
	Replicate			Replicate			Replicate			Replicate			Replicate			Replicate			Replicate		
Offset	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
-4	-	0.158	-	-	-	-	-	-	-	-	-	0.228	-	-	-	-	-	-	-	-	-
-3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.047	-	-
-1	-	-	-	-	-	-	-	-	-	0.093	0.315	0.456	0.233	-		0.047	-	-	-	-	-
0	0.093	-	-	-	-	-	2.33	2.681	4.556	96.32	96.06	94.31	-	-	0.456	-	-	-	-	-	-
1	-	-	-	-	-	-	0.093	0.158	-	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	0.14	-	-	-	-	-	-	-	-	0.047	-	-	-	-	-	-	-	-
3	-	0.158	-	-	-	-	-	-	-	-	0.158	-	-	0.158	-	-	-	-	0.28	-	-
4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.28	0.047	-

35S:MIR390a* miR390

	Small RNA Size Class																				
	18			19			20			21			22			23			24		
	Replicate			Replicate			Replicate			Replicate			Replicate			Replicate			Replicate		
Offset	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
-4	-	-	-	-	-	-	-	-	-	-	-	-	-	0.109	-	-	-	-	-	0.218	-
-3	-	-	-	-	0.109	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.218	-
-2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0	0.352	0.327	0.656	3.612	3.162	0.984	26.87	36.1	33.77	66.26	57.36	62.95	0.264	-	-	-	0.109		-	-	-
1	0.264	0.436	-	1.145	1.418	0.328	0.705	0.218	0.328	0.088	-	-	-	-	-	-	-	-	-	-	-
2	0.088	0.218	0.984	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.176	-	-
3	0.088	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-	-	-	-	-	0.088	-	-	-	-	-	-	-	-

B

***35S:mir390a-1* miR390**

	Small RNA Size Class																				
	18			19			20			21			22			23			24		
	Replicate			Replicate			Replicate			Replicate			Replicate			Replicate			Replicate		
Offset	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
-4	-	-	-	-	-	-	0.484	-	0.423	1.937	0.196	2.114	-	-	-	-	-	-	1	-	-
-3	-	-	-	-	-	0.423	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-2	-	-	-	-	0.196	-	-	-	-	-	-	-	0.242	-	-	-	-	-	-	-	-
-1	-	-	-	-	-	-	-	-	-	57.38	21.22	24.74	-	0.098	0.423	-	-	-	-	-	-
0	-	0.196	-	-	-	0.423	17.68	13.56	13.53	20.1	63.26	53.49	-	-	-	-	-	-	-	-	-
1	-	-	-	-	0.196			0.295	1.48		0.098	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.196	0.211
3	-	0.196	-	-	-	-	-	-	-	-	-	-	-	-	0.634	-	-	-	0.969	0.196	1.691
4	-	-	-	-	-	-	-	-	-	-	-	0.423	-	-	-	-	-	-	0.969	0.098	-

35S:mir390a-1* miR390

	Small RNA Size Class																				
	18			19			20			21			22			23			24		
	Replicate			Replicate			Replicate			Replicate			Replicate			Replicate			Replicate		
Offset	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
-4	-	-	-	-	-	-	-	-	-	-	2		-	-	-	-	-	1	-	-	-
-3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	-
-2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-1	-	-	-	-	-	-	-		0.153	-	-	-	-	-	-	-	-	-	-	1	-
0	-	0.321	-	0.726	0.963	1.529	9.144	14.77	14.83	7.257	31.78	29.82	-	-	-	-	-	-	-	-	-
1	12.05	7.223	5.046	59.51	32.74	32.87	3.483	5.618	7.034	4.499	2.729	4.893	2.032	2.889	3.364	-	-	-	-	-	-
2	0.29	0.161	0.153	-	-	-	-	0.161	0.153	-	-	-	-	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	0.726	-	-	-	-	-	0.145	-	-	-	-	-	-	-	-	-	-	-	-	-	-

C

***35S:MIR390b* miR390**

	Small RNA Size Class																				
	18			19			20			21			22			23			24		
	Replicate			Replicate			Replicate			Replicate			Replicate			Replicate			Replicate		
Offset	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
-4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-1	-	-	-	-	-	-	-	1.6	1.613	1.14	7.2	2.581	-	4.8	0.323	-	-	-	-	-	-
0	-	-	-	0.285	-	0.968	5.128	5.6	8.387	86.61	76.8	75.81	0.57	-	-	-	-	-	-	0.8	-
1	-	-	-	0.285	-	1.935	2.849	3.2	5.161	3.134	-	1.935	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-	-	-	-	0.968	-	-	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-	0.323	-	-	-	-	-	-	-	-	-	-	-	-

35S:MIR390b* miR390

	Small RNA Size Class																				
	18			19			20			21			22			23			24		
	Replicate			Replicate			Replicate			Replicate			Replicate			Replicate			Replicate		
Offset	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
-4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-3	-	-	-	-	-	-	-	-	-	-	-	0.253	-	-	-	-	-	-	-	-	-
-2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-1	-	-	-	0.345	-	-	-	-	-	-	1.382	0.253	-	-	-	-	-	-	-	-	-
0	1.034	-	0.253	7.241	9.677	5.808	66.55	43.78	62.63	6.552	7.834	8.586	-	-	-	-	-	-	-	-	-
1	-	5.991	2.525	16.9	27.19	16.92	0.345	1.843	1.768	-	-	-	-	-	-	-	-	-	-	-	-
2	1.034	1.843	0.758	-	-	-	-	0.461	0.253	-	-	-	-	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Fig. S3. Percentage of small RNA reads mapping within a 29-nt window centered on miR390 or miR390* from transient expression assay (*A–D*) Percentage of small RNA reads mapping to miR390 or miR390* from a transient expression assay. Offset refers to the 5′ position of reads, where 0 is the 5′ position of miR390 or miR390*. A negative offset refers to positions 5′ upstream of miR390 or miR390*, whereas positive offset refers to positions 3′ downstream of miR390 or miR390*. Reads were mapped against appropriate, infiltrated *MIR390* foldback, or, in the case of *35S:GUS*, reads were mapped to both *MIR390a* and *MIR390b*.

Table S1. Position of ago7 and dcl4 mutations

Allele	Genomic coordinate	AGI gene coordinate	Transcript coordinate	Reference base	Mutant base	Reference amino acid	Mutant amino acid	Independent alleles	Domain
Ago7-2	1:26107506	1325	778	C	T	Q	stop	1	Near N-term
Ago7-3	1:26105772	3059	2429	G	A	S	N	1	Piwi
Ago7-4	1:26105371	3460	2831	G	A	R	K	2	Piwi
Ago7-5	1:26105954	2877	2248	G	A	A	T	4	Piwi
ago7-6	1:26105896	2935	2306	C	T	S	F	1	Piwi
ago7-7	1:26106454	2377	1748	G	A	G	E	1	Mid
ago7-8	1:26106199	2632	2003	C	T	S	L	1	Piwi
ago7-9	1:26106703	2128	-	G	A	Splice junction	1	Piwi	
ago7-10	1:26105750	3081	2452	C	T	L	F	1	Piwi
ago7-11	1:26105735	3096	2467	C	T	Q	Stop	1	Piwi
ago7-12	1:26105882	2949	2320	C	T	Q	Stop	1	Piwi
ago7-13	1:26106197	2634	2005	G	A	E	K	1	Piwi
dcl4-7	5:6862527	6742	3382	G	A	G	R	1	Between PAZ and RNase III
dcl4-8	5:6863775	5492	2546	G	A	W	Stop	1	Between dsRBD and PAZ
dcl4-9	5:6862527	5429	2483	G	A	G	D	1	Between dsRBD and PAZ

Table S2. Polymorphisms in the 1.52-Mb mapping interval in 52b2 and 87a3 mutants

Locus	Feature	Reference base	52b2	87a3	MAQ quality score	Read depth	Both strands
2:15841833	Intergenic	G	A	G	75	16	Y
2:16069126	MIR390a	G	A	A	54	9	Y
2:16766679	Intergenic	A*	G	G	51	8	Y
2:16883030	Transposon	C	T	C	45	6	Y
2:17096007	AT2G40950	G	A	G	51	8	Y
2:17302163	AT2G41470	G	A	G	63	12	Y

*Reference base refers to TAIR8 sequence data. Sequence data revealed a G at position 2:16766679 in the parental line, 52b2, and 87a3, indicating a difference between the TAIR8 reference sequence and our Col-0 isolate.

Table S3. Small RNA reads mapping within a 29-nt window centered on miR390 or miR390* from transient expression assays

Sample and reads	Replicates (barcode)		
35S:MIR390a - miR390 reads	2146 (C)	634 (A)	439 (U)
35S:MIR390a - miR390* reads	1135 (C)	917 (A)	305 (U)
35S:mir390a-1 - miR390 reads	413 (G)	1018 (C)	473 (A)
35S:mir390a-1 - miR390* reads	689 (G)	623 (C)	654 (A)
35S:MIR390b - miR390 reads	351 (U)	125 (G)	310 (C)
35S:MIR390b - miR390* reads	290 (U)	217 (G)	396 (C)
35S:GUS - miR390 reads	14 (A)	46 (U)	36 (G)
35S:GUS - miR390* reads	11 (A)	30 (U)	28 (G)

Table S4. Primers used to design MIR390-based constructs

Primer description	Sequence*
MIR390a R primer I	GGCGCTATCCCTCCTGAGCTTTACAGATTCTTCTCTACTTTG
MIR390a F primer II	GGCGCTATCCATCCTGAGTTTCATTGGCTCTTCTTACTACAATG
MIR390a F primer III	AAGCTCAGGAGGGATAGCGCCATGATGATCACATTCGTTATC
MIR390a R primer IV	TGAAACTCAGGATGGATAGCGCCAAAAAATAGATAACGAATGTGATC
mir390a R primer I 12C	GGCGCTATCCCTCCTGAGCTTTACAG**G**TTCTTCTCTACTTTG
mir390a R primer I 12G	GGCGCTATCCCTCCTGAGCTTTACAG**C**TTCTTCTCTACTTTG
mir390a R primer I 12A	GGCGCTATCCCTCCTGAGCTTTACAG**T**TTCTTCTCTACTTTG
mir390a F primer II 96A	GGCGCTATCCATCCTGAGTTTCATTG**A**CTCTTCTTACTACAATG
mir390a F primer II 96C	GGCGCTATCCATCCTGAGTTTCATTG**C**CTCTTCTTACTACAATG
mir390a F primer II 96U	GGCGCTATCCATCCTGAGTTTCATTG**T**CTCTTCTTACTACAATG
mir390a/b-Cswap R primer I	GGCGCTATCCCTCCTGAGCTTT**ATAGCTAT**TTCTCTACTTTG
mir390a/b-Cswap F primer II	GGCGCTATCCATCCTGAGTTTC**ATAGCTTC**TTCTTACTACAATG
MIR390a_loop_F	AAGCTCAGGAGGGATAGCGCCATGATGATCACATTCGTTATCTATTTTTTGGCGCTATCCATCCTGAGTTTCA
MIR390a_loop_R	TGAAACTCAGGATGGATAGCGCCAAAAAATAGATAACGAATGTGATCATCATGGCGCTATCCCTCCTGAGCTT
MIR390b F primer III	AGAATCTGTAAAGCTCAGGAGGGATAGCGCCATGGCTCACCAGTGCTGATG
MIR390b R primer IIII	AGAGCCTTTGGAACTCAGGATGGATAGCGCCAACAGATATACATGTA

*Boldface nucleotides represent nucleotide substitutions in MIR390a constructs.

The Sequence of the Human Genome

J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1] Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1] Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1] Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1] Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1] George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5] Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9] Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1] Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1] Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1] Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1] Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1] Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1] Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1] Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1] Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1] Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12] Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1] Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1] Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1] Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1] Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1] Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1] Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1] Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1] Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1] Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1] Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1] Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1] Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1] Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1] Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1] Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1] Sandra Windsor,[1] Emily Winn-Deen,[1] Keriellen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1] Josep F. Abril,[14] Roderic Guigó,[14] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1] Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1] Anushya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1] Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1] Louis Blick,[1] Marcelo Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1] Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1] Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1] Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1] Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1] Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1] Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1] Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1] Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1] Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

A 2.91-billion base pair (bp) consensus sequence of the euchromatic portion of the human genome was generated by the whole-genome shotgun sequencing method. The 14.8-billion bp DNA sequence was generated over 9 months from 27,271,853 high-quality sequence reads (5.11-fold coverage of the genome) from both ends of plasmid clones made from the DNA of five individuals. Two assembly strategies—a whole-genome assembly and a regional chromosome assembly—were used, each combining sequence data from Celera and the publicly funded genome effort. The public data were shredded into 550-bp segments to create a 2.9-fold coverage of those genome regions that had been sequenced, without including biases inherent in the cloning and assembly procedure used by the publicly funded group. This brought the effective coverage in the assemblies to eightfold, reducing the number and size of gaps in the final assembly over what would be obtained with 5.11-fold coverage. The two assembly strategies yielded very similar results that largely agree with independent mapping data. The assemblies effectively cover the euchromatic regions of the human chromosomes. More than 90% of the genome is in scaffold assemblies of 100,000 bp or more, and 25% of the genome is in scaffolds of 10 million bp or larger. Analysis of the genome sequence revealed 26,588 protein-encoding transcripts for which there was strong corroborating evidence and an additional ~12,000 computationally derived genes with mouse matches or other weak supporting evidence. Although gene-dense clusters are obvious, almost half the genes are dispersed in low G+C sequence separated by large tracts of apparently noncoding sequence. Only 1.1% of the genome is spanned by exons, whereas 24% is in introns, with 75% of the genome being intergenic DNA. Duplications of segmental blocks, ranging in size up to chromosomal lengths, are abundant throughout the genome and reveal a complex evolutionary history. Comparative genomic analysis indicates vertebrate expansions of genes associated with neuronal function, with tissue-specific developmental regulation, and with the hemostasis and immune systems. DNA sequence comparisons between the consensus sequence and publicly funded genome data provided locations of 2.1 million single-nucleotide polymorphisms (SNPs). A random pair of human haploid genomes differed at a rate of 1 bp per 1250 on average, but there was marked heterogeneity in the level of polymorphism across the genome. Less than 1% of all SNPs resulted in variation in proteins, but the task of determining which SNPs have functional consequences remains an open challenge.

Decoding of the DNA that constitutes the human genome has been widely anticipated for the contribution it will make toward understanding human evolution, the causation of disease, and the interplay between the environment and heredity in defining the human condition. A project with the goal of determining the complete nucleotide sequence of the human genome was first formally proposed in 1985 (*1*). In subsequent years, the idea met with mixed reactions in the scientific community (*2*). However, in 1990, the Human Genome Project (HGP) was officially initiated in the United States under the direction of the National Institutes of Health and the U.S. Department of Energy with a 15-year, $3 billion plan for completing the genome sequence. In 1998 we announced our intention to build a unique genome-sequencing facility, to determine the sequence of the human genome over a 3-year period. Here we report the penultimate milestone along the path toward that goal, a nearly complete sequence of the euchromatic portion of the human genome. The sequencing was performed by a whole-genome random shotgun method with subsequent assembly of the sequenced segments.

The modern history of DNA sequencing began in 1977, when Sanger reported his method for determining the order of nucleotides of DNA using chain-terminating nucleotide analogs (*3*). In the same year, the first human gene was isolated and sequenced (*4*). In 1986, Hood and co-workers (*5*) described an improvement in the Sanger sequencing method that included attaching fluorescent dyes to the nucleotides, which permitted them to be sequentially read by a computer. The first automated DNA sequencer, developed by Applied Biosystems in California in 1987, was shown to be successful when the sequences of two genes were obtained with this new technology (*6*). From early sequencing of human genomic regions (*7*), it became clear that cDNA sequences (which are reverse-transcribed from RNA) would be essential to annotate and validate gene predictions in the human genome. These studies were the basis in part for the development of the expressed sequence tag (EST) method of gene identification (*8*), which is a random selection, very high throughput sequencing approach to characterize cDNA libraries. The EST method led to the rapid discovery and mapping of human genes (*9*). The increasing numbers of human EST sequences necessitated the development of new computer algorithms to analyze large amounts of sequence data, and in 1993 at The Institute for Genomic Research (TIGR), an algorithm was developed that permitted assembly and analysis of hundreds of thousands of ESTs. This algorithm permitted characterization and annotation of human genes on the basis of 30,000 EST assemblies (*10*).

The complete 49-kbp bacteriophage lambda genome sequence was determined by a shotgun restriction digest method in 1982 (*11*). When considering methods for sequencing the smallpox virus genome in 1991 (*12*), a whole-genome shotgun sequencing method was discussed and subsequently rejected owing to the lack of appropriate software tools for genome assembly. However, in 1994, when a microbial genome-sequencing project was contemplated at TIGR, a whole-genome shotgun sequencing approach was considered possible with the TIGR EST assembly algorithm. In 1995, the 1.8-Mbp *Haemophilus influenzae* genome was completed by a whole-genome shotgun sequencing method (*13*). The experience with several subsequent genome-sequencing efforts established the broad applicability of this approach (*14*, *15*).

A key feature of the sequencing approach used for these megabase-size and larger genomes was the use of paired-end sequences (also called mate pairs), derived from subclone libraries with distinct insert sizes and cloning characteristics. Paired-end sequences are sequences 500 to 600 bp in length from both ends of double-stranded DNA clones of prescribed lengths. The success of using end sequences from long segments (18 to 20 kbp) of DNA cloned into bacteriophage lambda in assembly of the microbial genomes led to the suggestion (*16*) of an approach to simulta-

[1]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [2]GenetixXpress, 78 Pacific Road, Palm Beach, Sydney 2108, Australia. [3]Berkeley *Drosophila* Genome Project, University of California, Berkeley, CA 94720, USA. [4]Department of Biology, Penn State University, 208 Mueller Lab, University Park, PA 16802, USA. [5]Department of Genetics, Case Western Reserve University School of Medicine, BRB-630, 10900 Euclid Avenue, Cleveland, OH 44106, USA. [6]Johns Hopkins University School of Medicine, Johns Hopkins Hospital, 600 North Wolfe Street, Blalock 1007, Baltimore, MD 21287–4922, USA. [7]Rockefeller University, 1230 York Avenue, New York, NY 10021–6399, USA. [8]New England BioLabs, 32 Tozer Road, Beverly, MA 01915, USA. [9]Division of Biology, 147-75, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA. [10]Yale University School of Medicine, 333 Cedar Street, P.O. Box 208000, New Haven, CT 06520–8000, USA. [11]Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA. [12]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [13]Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, 52900 Israel. [14]Grup de Recerca en Informàtica Mèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, 08003-Barcelona, Catalonia, Spain.

*To whom correspondence should be addressed. E-mail: humangenome@celera.com

neously map and sequence the human genome by means of end sequences from 150-kbp bacterial artificial chromosomes (BACs) (*17*, *18*). The end sequences spanned by known distances provide long-range continuity across the genome. A modification of the BAC end-sequencing (BES) method was applied successfully to complete chromosome 2 from the *Arabidopsis thaliana* genome (*19*).

In 1997, Weber and Myers (*20*) proposed whole-genome shotgun sequencing of the human genome. Their proposal was not well received (*21*). However, by early 1998, as less than 5% of the genome had been sequenced, it was clear that the rate of progress in human genome sequencing worldwide was very slow (*22*), and the prospects for finishing the genome by the 2005 goal were uncertain.

In early 1998, PE Biosystems (now Applied Biosystems) developed an automated, high-throughput capillary DNA sequencer, subsequently called the ABI PRISM 3700 DNA Analyzer. Discussions between PE Biosystems and TIGR scientists resulted in a plan to undertake the sequencing of the human genome with the 3700 DNA Analyzer and the whole-genome shotgun sequencing techniques developed at TIGR (*23*). Many of the principles of operation of a genome-sequencing facility were established in the TIGR facility (*24*). However, the facility envisioned for Celera would have a capacity roughly 50 times that of TIGR, and thus new developments were required for sample preparation and tracking and for whole-genome assembly. Some argued that the required 150-fold scale-up from the *H. influenzae* genome to the human genome with its complex repeat sequences was not feasible (*25*). The *Drosophila melanogaster* genome was thus chosen as a test case for whole-genome assembly on a large and complex eukaryotic genome. In collaboration with Gerald Rubin and the Berkeley *Drosophila* Genome Project, the nucleotide sequence of the 120-Mbp euchromatic portion of the *Drosophila* genome was determined over a 1-year period (*26*–*28*). The *Drosophila* genome-sequencing effort resulted in two key findings: (i) that the assembly algorithms could generate chromosome assemblies with highly accurate order and orientation with substantially less than 10-fold coverage, and (ii) that undertaking multiple interim assemblies in place of one comprehensive final assembly was not of value.

These findings, together with the dramatic changes in the public genome effort subsequent to the formation of Celera (*29*), led to a modified whole-genome shotgun sequencing approach to the human genome. We initially proposed to do 10-fold sequence coverage of the genome over a 3-year period and to make interim assembled sequence data available quarterly. The modifications included a plan to perform random shotgun sequencing to ~5-fold coverage and to use the unordered and unoriented BAC sequence fragments and subassemblies published in GenBank by the publicly funded genome effort (*30*) to accelerate the project. We also abandoned the quarterly announcements in the absence of interim assemblies to report.

Although this strategy provided a reasonable result very early that was consistent with a whole-genome shotgun assembly with eight-fold coverage, the human genome sequence is not as finished as the *Drosophila* genome was with an effective 13-fold coverage. However, it became clear that even with this reduced coverage strategy, Celera could generate an accurately ordered and oriented scaffold sequence of the human genome in less than 1 year. Human genome sequencing was initiated 8 September 1999 and completed 17 June 2000. The first assembly was completed 25 June 2000, and the assembly reported here was completed 1 October 2000. Here we describe the whole-genome random shotgun sequencing effort applied to the human genome. We developed two different assembly approaches for assembling the ~3 billion bp that make up the 23 pairs of chromosomes of the *Homo sapiens* genome. Any GenBank-derived data were shredded to remove potential bias to the final sequence from chimeric clones, foreign DNA contamination, or misassembled contigs. Insofar as a correctly and accurately assembled genome sequence with faithful order and orientation of contigs is essential for an accurate analysis of the human genetic code, we have devoted a considerable portion of this manuscript to the documentation of the quality of our reconstruction of the genome. We also describe our preliminary analysis of the human genetic code on the basis of computational methods. Figure 1 (see fold-out chart associated with this issue; files for each chromosome can be found in Web fig. 1 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1) provides a graphical overview of the genome and the features encoded in it. The detailed manual curation and interpretation of the genome are just beginning.

To aid the reader in locating specific analytical sections, we have divided the paper into seven broad sections. A summary of the major results appears at the beginning of each section.

1 Sources of DNA and Sequencing Methods

Summary. This section discusses the rationale and ethical rules governing donor selection to ensure ethnic and gender diversity along with the methodologies for DNA extraction and library construction. The plasmid library construction is the first critical step in shotgun sequencing. If the DNA libraries are not uniform in size, nonchimeric, and do not randomly represent the genome, then the subsequent steps cannot accurately reconstruct the genome sequence. We used automated high-throughput DNA sequencing and the computational infrastructure to enable efficient tracking of enormous amounts of sequence information (27.3 million sequence reads; 14.9 billion bp of sequence). Sequencing and tracking from both ends of plasmid clones from 2-, 10-, and 50-kbp libraries were essential to the computational reconstruction of the genome. Our evidence indicates that the accurate pairing rate of end sequences was greater than 98%.

Various policies of the United States and the World Medical Association, specifically the Declaration of Helsinki, offer recommendations for conducting experiments with human subjects. We convened an Institutional Review Board (IRB) (*31*) that helped us establish the protocol for obtaining and using human DNA and the informed consent process used to enroll research volunteers for the DNA-sequencing studies reported here. We adopted several steps and procedures to protect the privacy rights and confidentiality of the research subjects (donors). These included a two-stage consent process, a secure random alphanumeric coding system for specimens and records, circumscribed contact with the subjects by researchers, and options for off-site contact of donors. In addition, Celera applied for and received a Certificate of Confidentiality from the Department of Health and Human Services. This Certificate authorized Celera to protect the privacy of the individuals who volunteered to be donors as provided in Section 301(d) of the Public Health Service Act 42 U.S.C. 241(d).

Celera and the IRB believed that the initial version of a completed human genome should be a composite derived from multiple donors of diverse ethnic backgrounds Prospective donors were asked, on a voluntary basis, to self-designate an ethnogeographic category (e.g., African-American, Chinese, Hispanic, Caucasian, etc.). We enrolled 21 donors (*32*).

Three basic items of information from each donor were recorded and linked by confidential code to the donated sample: age, sex, and self-designated ethnogeographic group. From females, ~130 ml of whole, heparinized blood was collected. From males, ~130 ml of whole, heparinized blood was

collected, as well as five specimens of semen, collected over a 6-week period. Permanent lymphoblastoid cell lines were created by Epstein-Barr virus immortalization. DNA from five subjects was selected for genomic DNA sequencing: two males and three females—one African-American, one Asian-Chinese, one Hispanic-Mexican, and two Caucasians (see Web fig. 2 on *Science* Online at www.sciencemag.org/cgi/content/291/5507/1304/DC1). The decision of whose DNA to sequence was based on a complex mix of factors, including the goal of achieving diversity as well as technical issues such as the quality of the DNA libraries and availability of immortalized cell lines.

1.1 Library construction and sequencing

Central to the whole-genome shotgun sequencing process is preparation of high-quality plasmid libraries in a variety of insert sizes so that pairs of sequence reads (mates) are obtained, one read from both ends of each plasmid insert. High-quality libraries have an equal representation of all parts of the genome, a small number of clones without inserts, and no contamination from such sources as the mitochondrial genome and *Escherichia coli* genomic DNA. DNA from each donor was used to construct plasmid libraries in one or more of three size classes: 2 kbp, 10 kbp, and 50 kbp (Table 1) (*33*).

In designing the DNA-sequencing process, we focused on developing a simple system that could be implemented in a robust and reproducible manner and monitored effectively (Fig. 2) (*34*).

Current sequencing protocols are based on the dideoxy sequencing method (*35*), which typically yields only 500 to 750 bp of sequence per reaction. This limitation on read length has made monumental gains in throughput a prerequisite for the analysis of large eukaryotic genomes. We accomplished this at the Celera facility, which occupies about 30,000 square feet of laboratory space and produces sequence data continuously at a rate of 175,000 total reads per day. The DNA-sequencing facility is supported by a high-performance computational facility (*36*).

The process for DNA sequencing was modular by design and automated. Intermodule sample backlogs allowed four principal modules to operate independently: (i) library transformation, plating, and colony picking; (ii) DNA template preparation; (iii) dideoxy sequencing reaction set-up and purification; and (iv) sequence determination with the ABI PRISM 3700 DNA Analyzer. Because the inputs and outputs of each module have been carefully matched and sample backlogs are continuously managed, sequencing has proceeded without a single day's interruption since the initiation of the *Drosophila* project in May 1999. The ABI 3700 is a fully automated capillary array sequencer and as such can be operated with a minimal amount of hands-on time, currently estimated at about 15 min per day. The capillary system also facilitates correct associations of sequencing traces with samples through the elimination of manual sample loading and lane-tracking errors associated with slab gels. About 65 production staff were hired and trained, and were rotated on a regular basis through the four production modules. A central laboratory information management system (LIMS) tracked all sample plates by unique bar code identifiers. The facility was supported by a quality control team that performed raw material and in-process testing and a quality assurance group with responsibilities including document control, validation, and auditing of the facility. Critical to the success of the scale-up was the validation of all software and instrumentation before implementation, and production-scale testing of any process changes.

1.2 Trace processing

An automated trace-processing pipeline has been developed to process each sequence file (*37*). After quality and vector trimming, the average trimmed sequence length was 543 bp, and the sequencing accuracy was exponentially distributed with a mean of 99.5% and with less than 1 in 1000 reads being less than 98% accurate (*26*). Each trimmed sequence was screened for matches to contaminants including sequences of vector alone, *E. coli* genomic DNA, and human mitochondrial DNA. The entire read for any sequence with a significant match to a contaminant was discarded. A total of 713 reads matched *E. coli* genomic DNA and 2114 reads matched the human mitochondrial genome.

1.3 Quality assessment and control

The importance of the base-pair level accuracy of the sequence data increases as the size and repetitive nature of the genome to be sequenced increases. Each sequence read must be placed uniquely in the ge-

Table 1. Celera-generated data input into assembly.

	Individual	Number of reads for different insert libraries				Total number of base pairs
		2 kbp	10 kbp	50 kbp	Total	
No. of sequencing reads	A	0	0	2,767,357	2,767,357	1,502,674,851
	B	11,736,757	7,467,755	66,930	19,271,442	10,464,393,006
	C	853,819	881,290	0	1,735,109	942,164,187
	D	952,523	1,046,815	0	1,999,338	1,085,640,534
	F	0	1,498,607	0	1,498,607	813,743,601
	Total	13,543,099	10,894,467	2,834,287	27,271,853	14,808,616,179
Fold sequence coverage (2.9-Gb genome)	A	0	0	0.52	0.52	
	B	2.20	1.40	0.01	3.61	
	C	0.16	1.17	0	0.32	
	D	0.18	0.20	0	0.37	
	F	0	0.28	0	0.28	
	Total	2.54	2.04	0.53	5.11	
Fold clone coverage	A	0	0	18.39	18.39	
	B	2.96	11.26	0.44	14.67	
	C	0.22	1.33	0	1.54	
	D	0.24	1.58	0	1.82	
	F	0	2.26	0	2.26	
	Total	3.42	16.43	18.84	38.68	
Insert size* (mean)	Average	1,951 bp	10,800 bp	50,715 bp		
Insert size* (SD)	Average	6.10%	8.10%	14.90%		
% Mates†	Average	74.50	80.80	75.60		

*Insert size and SD are calculated from assembly of mates on contigs. †% Mates is based on laboratory tracking of sequencing runs.

nome, and even a modest error rate can reduce the effectiveness of assembly. In addition, maintaining the validity of mate-pair information is absolutely critical for the algorithms described below. Procedural controls were established for maintaining the validity of sequence mate-pairs as sequencing reactions proceeded through the process, including strict rules built into the LIMS. The accuracy of sequence data produced by the Celera process was validated in the course of the *Drosophila* genome project (*26*). By collecting data for the entire human genome in a single facility, we were able to ensure uniform quality standards and the cost advantages associated with automation, an economy of scale, and process consistency.

2 Genome Assembly Strategy and Characterization

Summary. We describe in this section the two approaches that we used to assemble the genome. One method involves the computational combination of all sequence reads with shredded data from GenBank to generate an independent, nonbiased view of the genome. The second approach involves clustering all of the fragments to a region or chromosome on the basis of mapping information. The clustered data were then shredded and subjected to computational assembly. Both approaches provided essentially the same reconstruction of assembled DNA sequence with proper order and orientation. The second method provided slightly greater sequence coverage (fewer gaps) and was the principal sequence used for the analysis phase. In addition, we document the completeness and correctness of this assembly process

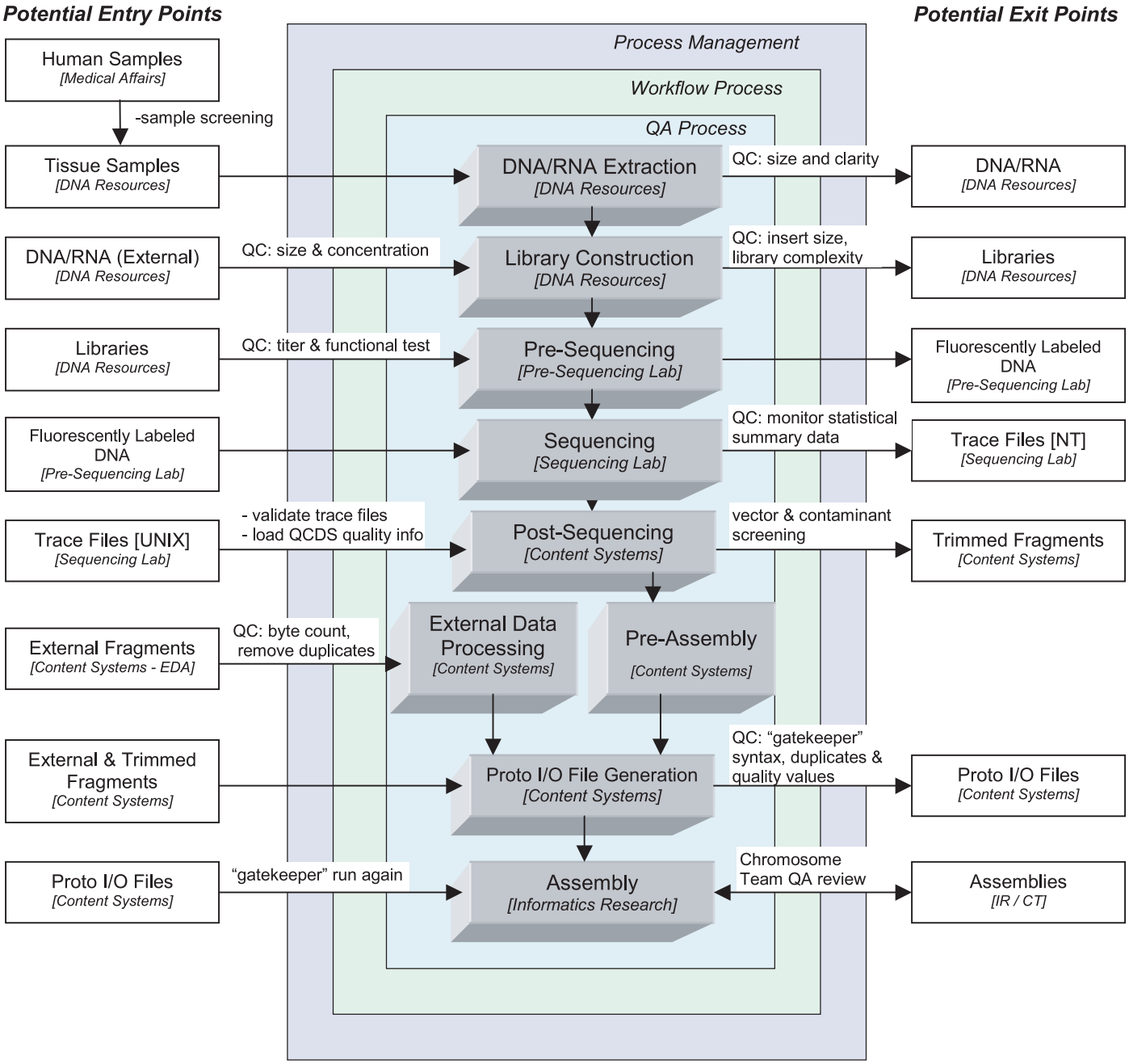


Fig. 2. Flow diagram for sequencing pipeline. Samples are received, selected, and processed in compliance with standard operating procedures, with a focus on quality within and across departments. Each process has defined inputs and outputs with the capability to exchange samples and data with both internal and external entities according to defined quality guidelines. Manufacturing pipeline processes, products, quality control measures, and responsible parties are indicated and are described further in the text.

and provide a comparison to the public genome sequence, which was reconstructed largely by an independent BAC-by-BAC approach. Our assemblies effectively covered the euchromatic regions of the human chromosomes. More than 90% of the genome was in scaffold assemblies of 100,000 bp or greater, and 25% of the genome was in scaffolds of 10 million bp or larger.

Shotgun sequence assembly is a classic example of an inverse problem: given a set of reads randomly sampled from a target sequence, reconstruct the order and the position of those reads in the target. Genome assembly algorithms developed for *Drosophila* have now been extended to assemble the ~25-fold larger human genome. Celera assemblies consist of a set of contigs that are ordered and oriented into scaffolds that are then mapped to chromosomal locations by using known markers. The contigs consist of a collection of overlapping sequence reads that provide a consensus reconstruction for a contiguous interval of the genome. Mate pairs are a central component of the assembly strategy. They are used to produce scaffolds in which the size of gaps between consecutive contigs is known with reasonable precision. This is accomplished by observing that a pair of reads, one of which is in one contig, and the other of which is in another, implies an orientation and distance between the two contigs (Fig. 3). Finally, our assemblies did not incorporate all reads into the final set of reported scaffolds. This set of unincorporated reads is termed "chaff," and typically consisted of reads from within highly repetitive regions, data from other organisms introduced through various routes as found in many genome projects, and data of poor quality or with untrimmed vector.

2.1 Assembly data sets

We used two independent sets of data for our assemblies. The first was a random shotgun data set of 27.27 million reads of average length 543 bp produced at Celera. This consisted largely of mate-pair reads from 16 libraries constructed from DNA samples taken from five different donors. Libraries with insert sizes of 2, 10, and 50 kbp were used. By looking at how mate pairs from a library were positioned in known sequenced stretches of the genome, we were able to characterize the range of insert sizes in each library and determine a mean and standard deviation. Table 1 details the number of reads, sequencing coverage, and clone coverage achieved by the data set. The clone coverage is the coverage of the genome in cloned DNA, considering the entire insert of each clone that has sequence from both ends. The clone coverage provides a measure of the amount of physical DNA coverage of the genome. Assuming a genome size of 2.9 Gbp, the Celera trimmed sequences gave a 5.1× coverage of the genome, and clone coverage was 3.42×, 16.40×, and 18.84× for the 2-, 10-, and 50-kbp libraries, respectively, for a total of 38.7× clone coverage.

The second data set was from the publicly funded Human Genome Project (PFP) and is primarily derived from BAC clones (*30*). The BAC data input to the assemblies came from a download of GenBank on 1 September 2000 (Table 2) totaling 4443.3 Mbp of sequence. The data for each BAC is deposited at one of four levels of completion. Phase 0 data are a set of generally unassembled sequencing reads from a very light shotgun of the BAC, typically less than 1×. Phase 1 data are unordered assemblies of contigs, which we call BAC contigs or bactigs. Phase 2 data are ordered assemblies of bactigs. Phase 3 data are complete BAC sequences. In the past 2 years the PFP has focused on a product of lower quality and completeness, but on a faster time-course, by concentrating on the production of Phase 1 data from a 3× to 4× light-shotgun of each BAC clone.

We screened the bactig sequences for contaminants by using the BLAST algorithm against three data sets: (i) vector sequences in Univec core (*38*), filtered for a 25-bp match at 98% sequence identity at the ends of the sequence and a 30-bp match internal to the sequence; (ii) the nonhuman portion of the High Throughput Genomic (HTG) Seqences division of GenBank (*39*), filtered at 200 bp at 98%; and (iii) the nonredundant nucleotide sequences from GenBank without primate and human virus entries, filtered at 200 bp at 98%. Whenever 25 bp or more of vector was found within 50 bp of the end of a contig, the tip up to the matching vector was excised. Under these criteria we removed 2.6 Mbp of possible contaminant and vector from the Phase 3 data, 61.0 Mbp from the Phase 1 and 2 data, and 16.1 Mbp from the Phase 0 data (Table 2). This left us with a total of 4363.7 Mbp of PFP sequence data 20% finished, 75% rough-draft (Phase 1 and 2), and 5% single sequencing reads (Phase 0). An additional 104,018 BAC end-sequence mate pairs were also downloaded and included in the data sets for both assembly processes (*18*).

2.2 Assembly strategies

Two different approaches to assembly were pursued. The first was a whole-genome assembly process that used Celera data and the PFP data in the form of additional synthetic shotgun data, and the second was a compartmentalized assembly process that first partitioned the Celera and PFP data into sets localized to large chromosomal segments and then performed ab initio shotgun assembly on each set. Figure 4 gives a schematic of the overall process flow.

For the whole-genome assembly, the PFP data was first disassembled or "shredded" into a synthetic shotgun data set of 550-bp reads that form a perfect 2× covering of the bactigs. This resulted in 16.05 million "faux" reads that were sufficient to cover the genome 2.96× because of redundancy in the BAC data set, without incorporating the biases inherent in the PFP assembly process. The combined data set of 43.32 million reads (8×), and all associated mate-pair information, were then subjected to our whole-genome assembly algorithm to produce a reconstruction of the genome. Neither the location of a BAC in the genome nor its assembly of bactigs was used in this process. Bactigs were shredded into reads because we found strong evidence that 2.13% of them were misassembled (*40*). Furthermore, BAC location

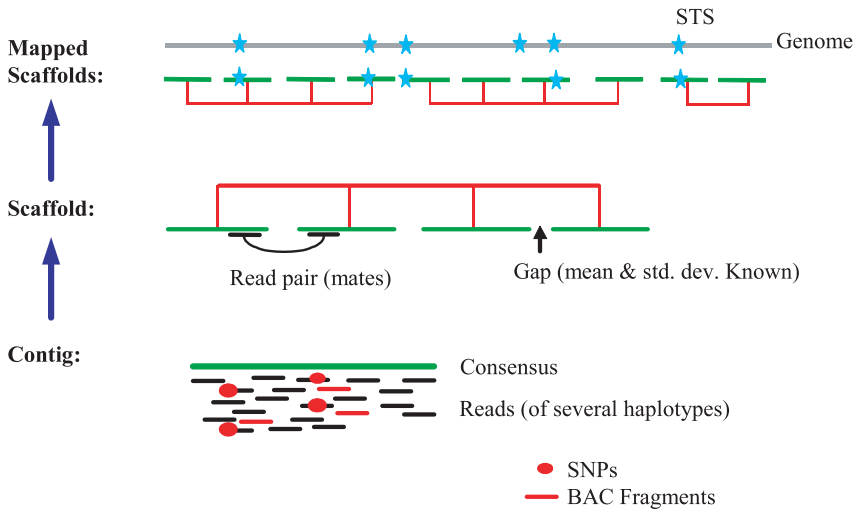


Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

information was ignored because some BACs were not correctly placed on the PFP physical map and because we found strong evidence that at least 2.2% of the BACs contained sequence data that were not part of the given BAC (*41*), possibly as a result of sample-tracking errors (see below). In short, we performed a true, ab initio whole-genome assembly in which we took the expedient of deriving additional sequence coverage, but not mate pairs, assembled bactigs, or genome locality, from some externally generated data.

Table 2. GenBank data input into assembly.

Center	Statistics	Completion phase sequence		
		0	1 and 2	3
Whitehead Institute/ MIT Center for Genome Research, USA	Number of accession records	2,825	6,533	363
	Number of contigs	243,786	138,023	363
	Total base pairs	194,490,158	1,083,848,245	48,829,358
	Total vector masked (bp)	1,553,597	875,618	2,202
	Total contaminant masked (bp)	13,654,482	4,417,055	98,028
	Average contig length (bp)	798	7,853	134,516
Washington University, USA	Number of accession records	19	3,232	1,300
	Number of contigs	2,127	61,812	1,300
	Total base pairs	1,195,732	561,171,788	164,214,395
	Total vector masked (bp)	21,604	270,942	8,287
	Total contaminant masked (bp)	22,469	1,476,141	469,487
	Average contig length (bp)	562	9,079	126,319
Baylor College of Medicine, USA	Number of accession records	0	1,626	363
	Number of contigs	0	44,861	363
	Total base pairs	0	265,547,066	49,017,104
	Total vector masked (bp)	0	218,769	4,960
	Total contaminant masked (bp)	0	1,784,700	485,137
	Average contig length (bp)	0	5,919	135,033
Production Sequencing Facility, DOE Joint Genome Institute, USA	Number of accession records	135	2,043	754
	Number of contigs	7,052	34,938	754
	Total base pairs	8,680,214	294,249,631	60,975,328
	Total vector masked (bp)	22,644	162,651	7,274
	Total contaminant masked (bp)	665,818	4,642,372	118,387
	Average contig length (bp)	1,231	8,422	80,867
The Institute of Physical and Chemical Research (RIKEN), Japan	Number of accession records	0	1,149	300
	Number of contigs	0	25,772	300
	Total base pairs	0	182,812,275	20,093,926
	Total vector masked (bp)	0	203,792	2,371
	Total contaminant masked (bp)	0	308,426	27,781
	Average contig length (bp)	0	7,093	66,978
Sanger Centre, UK	Number of accession records	0	4,538	2,599
	Number of contigs	0	74,324	2,599
	Total base pairs	0	689,059,692	246,118,000
	Total vector masked (bp)	0	427,326	25,054
	Total contaminant masked (bp)	0	2,066,305	374,561
	Average contig length (bp)	0	9,271	94,697
Others*	Number of accession records	42	1,894	3,458
	Number of contigs	5,978	29,898	3,458
	Total base pairs	5,564,879	283,358,877	246,474,157
	Total vector masked (bp)	57,448	279,477	32,136
	Total contaminant masked (bp)	575,366	1,616,665	1,791,849
	Average contig length (bp)	931	9,478	71,277
All centers combined†	Number of accession records	3,021	21,015	9,137
	Number of contigs	258,943	409,628	9,137
	Total base pairs	209,930,983	3,360,047,574	835,722,268
	Total vector masked (bp)	1,655,293	2,438,575	82,284
	Total contaminant masked (bp)	14,918,135	16,311,664	3,365,230
	Average contig length (bp)	811	8,203	91,466

*Other centers contributing at least 0.1% of the sequence include: Chinese National Human Genome Center; Genomanalyse Gesellschaft fuer Biotechnologische Forschung mbH; Genome Therapeutics Corporation; GENOSCOPE; Chinese Academy of Sciences; Institute of Molecular Biotechnology; Keio University School of Medicine; Lawrence Livermore National Laboratory; Cold Spring Harbor Laboratory; Los Alamos National Laboratory; Max-Planck Institut fuer Molekulare, Genetik; Japan Science and Technology Corporation; Stanford University; The Institute for Genomic Research; The Institute of Physical and Chemical Research, Gene Bank; The University of Oklahoma; University of Texas Southwestern Medical Center, University of Washington. †The 4,405,700,825 bases contributed by all centers were shredded into faux reads resulting in 2.96× coverage of the genome.

In the compartmentalized shotgun assembly (CSA), Celera and PFP data were partitioned into the largest possible chromosomal segments or "components" that could be determined with confidence, and then shotgun assembly was applied to each partitioned subset wherein the bactig data were again shredded into faux reads to ensure an independent ab initio assembly of the component. By subsetting the data in this way, the overall computational effort was reduced and the effect of interchromosomal duplications was ameliorated. This also resulted in a reconstruction of the genome that was relatively independent of the whole-genome assembly results so that the two assemblies could be compared for consistency. The quality of the partitioning into components was crucial so that different genome regions were not mixed together. We constructed components from (i) the longest scaffolds of the sequence from each BAC and (ii) assembled scaffolds of data unique to Celera's data set. The BAC assemblies were obtained by a combining assembler that used the bactigs and the 5× Celera data mapped to those bactigs as input. This effort was undertaken as an interim step solely because the more accurate and complete the scaffold for a given sequence stretch, the more accurately one can tile these scaffolds into contiguous components on the basis of sequence overlap and mate-pair information. We further visually inspected and curated the scaffold tiling of the components to further increase its accuracy. For the final CSA assembly, all but the partitioning was ignored, and an independent, ab initio reconstruction of the sequence in each component was obtained by applying our whole-genome assembly algorithm to the partitioned, relevant Celera data and the shredded, faux reads of the partitioned, relevant bactig data.

2.3 Whole-genome assembly

The algorithms used for whole-genome assembly (WGA) of the human genome were enhancements to those used to produce the sequence of the *Drosophila* genome reported in detail in (*28*).

The WGA assembler consists of a pipeline composed of five principal stages: Screener, Overlapper, Unitigger, Scaffolder, and Repeat Resolver, respectively. The Screener finds and marks all microsatellite repeats with less than a 6-bp element, and screens out all known interspersed repeat elements, including Alu, Line, and ribosomal DNA. Marked regions get searched for overlaps, whereas screened regions do not get searched, but can be part of an overlap that involves unscreened matching segments.

The Overlapper compares every read against every other read in search of complete end-to-end overlaps of at least 40 bp and with no more than 6% differences in the match. Because all data are scrupulously vector-trimmed, the Overlapper can insist on complete overlap matches. Computing the set of all overlaps took roughly 10,000 CPU hours with a suite of four-processor Alpha SMPs with 4 gigabytes of RAM. This took 4 to 5 days in elapsed time with 40 such machines operating in parallel.

Every overlap computed above is statistically a 1-in-10^{17} event and thus not a coincidental event. What makes assembly combinatorially difficult is that while many overlaps are actually sampled from overlapping regions of the genome, and thus imply that the sequence reads should be assembled together, even more overlaps are actually from two distinct copies of a low-copy repeated element not screened above, thus constituting an error if put together. We call the former "true overlaps" and the latter "repeat-induced overlaps." The assembler must avoid choosing repeat-induced overlaps, especially early in the process.

We achieve this objective in the Unitigger. We first find all assemblies of reads that appear to be uncontested with respect to all other reads. We call the contigs formed from these subassemblies unitigs (for uniquely assembled contigs). Formally, these unitigs are the uncontested interval subgraphs of the graph of all overlaps (*42*). Unfortunately, although empirically many of these assemblies are correct (and thus involve only true overlaps), some are in fact collections of reads from several copies of a repetitive element that have been overcollapsed into a single subassembly. However, the overcollapsed unitigs are easily identified because their average coverage depth is too high to be consistent with the overall level of sequence coverage. We developed a simple statistical discriminator that gives the logarithm of the odds ratio that a unitig is composed of unique DNA or of a repeat consisting of two or more copies. The discriminator, set to a sufficiently stringent threshold, identifies a subset of the unitigs that we are certain are correct. In addition, a second, less stringent threshold identifies a subset of remaining unitigs very likely to be correctly assembled, of which we select those that will consistently scaffold (see below), and thus are again almost certain to be correct. We call the union of these two sets U-unitigs. Empirically, we found from a 6× simulated shotgun of human chromosome 22 that we get U-unitigs covering 98% of the stretches of unique DNA that are >2 kbp long. We are further able to identify the boundary of the start of a repetitive element at the ends of a U-unitig and leverage this so that U-unitigs span more than 93% of all singly interspersed Alu elements and other 100-to 400-bp repetitive segments.

The result of running the Unitigger was thus a set of correctly assembled subcontigs covering an estimated 73.6% of the human genome. The Scaffolder then proceeded to use mate-pair information to link these together into scaffolds. When there are two or more mate pairs that imply that a given pair of U-unitigs are at a certain distance and orientation with respect to each other, the probability of this being wrong is again roughly 1 in 10^{10}, assuming that mate pairs are false less than 2% of the time. Thus, one can with high confidence link together all U-unitigs that are linked by at least two 2- or 10-kbp mate pairs producing intermediate-sized scaffolds that are then recursively linked together by confirming 50-kbp mate pairs and BAC end sequences. This process yielded scaffolds that are on the order of megabase pairs in size with gaps between their contigs that generally correspond to repetitive elements and occasionally to small sequencing gaps. These scaffolds reconstruct the majority of the unique sequence within a genome.

For the *Drosophila* assembly, we engaged in a three-stage repeat resolution strategy where each stage was progressively more aggressive and thus more likely to make a mistake. For the human assembly, we continued to use the first "Rocks" substage where all unitigs with a good, but not definitive, discriminator score are placed in a scaffold gap. This was done with the condition that two or more mate pairs with one of their reads already in the scaffold unambiguously place the unitig in the given gap. We estimate the probability of inserting a unitig into an incorrect gap with this strategy to be less than 10^{-7} based on a probabilistic analysis.

We revised the ensuing "Stones" substage of the human assembly, making it more like the mechanism suggested in our earlier work (*43*). For each gap, every read R that is placed in the gap by virtue of its mated pair M being in a contig of the scaffold and implying R's placement is collected. Celera's mate-pairing information is correct more than 99% of the time. Thus, almost every, but not all, of the reads in the set belong in the gap, and when a read does not belong it rarely agrees with the remainder of the reads. Therefore, we simply assemble this set of reads within the gap, eliminating any reads that conflict with the assembly. This operation proved much more reliable than the one it replaced for the *Drosophila* assembly; in the assembly of a simulated shotgun data set of human chromo-

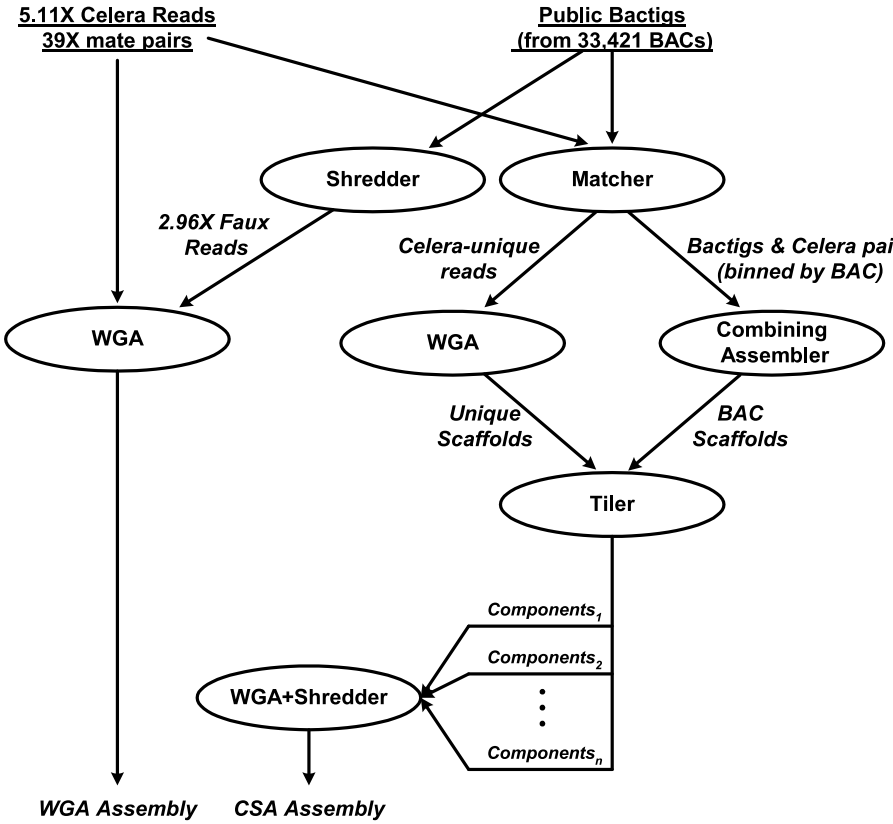


Fig. 4. Architecture of Celera's two-pronged assembly strategy. Each oval denotes a computation process performing the function indicated by its label, with the labels on arcs between ovals describing the nature of the objects produced and/or consumed by a process. This figure summarizes the discussion in the text that defines the terms and phrases used.

some 22, all stones were placed correctly.

The final method of resolving gaps is to fill them with assembled BAC data that cover the gap. We call this external gap "walking." We did not include the very aggressive "Pebbles" substage described in our *Drosophila* work, which made enough mistakes so as to produce repeat reconstructions for long interspersed elements whose quality was only 99.62% correct. We decided that for the human genome it was philosophically better not to introduce a step that was certain to produce less than 99.99% accuracy. The cost was a somewhat larger number of gaps of somewhat larger size.

At the final stage of the assembly process, and also at several intermediate points, a consensus sequence of every contig is produced. Our algorithm is driven by the principle of maximum parsimony, with quality-value–weighted measures for evaluating each base. The net effect is a Bayesian estimate of the correct base to report at each position. Consensus generation uses Celera data whenever it is present. In the event that no Celera data cover a given region, the BAC data sequence is used.

A key element of achieving a WGA of the human genome was to parallelize the Overlapper and the central consensus sequence–constructing subroutines. In addition, memory was a real issue—a straightforward application of the software we had built for *Drosophila* would have required a computer with a 600-gigabyte RAM. By making the Overlapper and Unitigger incremental, we were able to achieve the same computation with a maximum of instantaneous usage of 28 gigabytes of RAM. Moreover, the incremental nature of the first three stages allowed us to continually update the state of this part of the computation as data were delivered and then perform a 7-day run to complete Scaffolding and Repeat Resolution whenever desired. For our assembly operations, the total compute infrastructure consists of 10 four-processor SMPs with 4 gigabytes of memory per cluster (Compaq's ES40, Regatta) and a 16-processor NUMA machine with 64 gigabytes of memory (Compaq's GS160, Wildfire). The total compute for a run of the assembler was roughly 20,000 CPU hours.

The assembly of Celera's data, together with the shredded bactig data, produced a set of scaffolds totaling 2.848 Gbp in span and consisting of 2.586 Gbp of sequence. The chaff, or set of reads not incorporated in the assembly, numbered 11.27 million (26%), which is consistent with our experience for *Drosophila*. More than 84% of the genome was covered by scaffolds >100 kbp long, and these averaged 91% sequence and 9% gaps with a total of 2.297 Gbp of sequence. There were a total of 93,857 gaps among the 1637 scaffolds >100 kbp. The average scaffold size was 1.5 Mbp, the average contig size was 24.06 kbp, and the average gap size was 2.43 kbp, where the distribution of each was essentially exponential. More than 50% of all gaps were less than 500 bp long, >62% of all gaps were less than 1 kbp long, and no gap was >100 kbp long. Similarly, more than 65% of the sequence is in contigs >30 kbp, more than 31% is in contigs >100 kbp, and the largest contig was 1.22 Mbp long. Table 3 gives detailed summary statistics for the structure of this assembly with a direct comparison to the compartmentalized shotgun assembly.

2.4 Compartmentalized shotgun assembly

In addition to the WGA approach, we pursued a localized assembly approach that was intended to subdivide the genome into segments, each of which could be shotgun assembled individually. We expected that this would help in resolution of large interchromosomal duplications and improve the statistics for calculating U-unitigs. The compartmentalized assembly process involved clustering Celera reads and bactigs into large, multiple megabase regions of the genome, and then running the WGA assembler on the Celera data and shredded, faux reads obtained from the bactig data.

The first phase of the CSA strategy was to separate Celera reads into those that matched the BAC contigs for a particular PFP BAC entry, and those that did not match any public data. Such matches must be guaranteed to

Table 3. Scaffold statistics for whole-genome and compartmentalized shotgun assemblies.

	Scaffold size				
	All	>30 kbp	>100 kbp	>500 kbp	>1000 kbp
			Compartmentalized shotgun assembly		
No. of bp in scaffolds (including intrascaffold gaps)	2,905,568,203	2,748,892,430	2,700,489,906	2,489,357,260	2,248,689,128
No. of bp in contigs	2,653,979,733	2,524,251,302	2,491,538,372	2,320,648,201	2,106,521,902
No. of scaffolds	53,591	2,845	1,935	1,060	721
No. of contigs	170,033	112,207	107,199	93,138	82,009
No. of gaps	116,442	109,362	105,264	92,078	81,288
No. of gaps ≤1 kbp	72,091	69,175	67,289	59,915	53,354
Average scaffold size (bp)	54,217	966,219	1,395,602	2,348,450	3,118,848
Average contig size (bp)	15,609	22,496	23,242	24,916	25,686
Average intrascaffold gap size (bp)	2,161	2,054	1,985	1,832	1,749
Largest contig (bp)	1,988,321	1,988,321	1,988,321	1,988,321	1,988,321
% of total contigs	100	95	94	87	79
			Whole-genome assembly		
No. of bp in scaffolds (including intrascaffold gaps)	2,847,890,390	2,574,792,618	2,525,334,447	2,328,535,466	2,140,943,032
No. of bp in contigs	2,586,634,108	2,334,343,339	2,297,678,935	2,143,002,184	1,983,305,432
No. of scaffolds	118,968	2,507	1,637	818	554
No. of contigs	221,036	99,189	95,494	84,641	76,285
No. of gaps	102,068	96,682	93,857	83,823	75,731
No. of gaps ≤1 kbp	62,356	60,343	59,156	54,079	49,592
Average scaffold size (bp)	23,938	1,027,041	1,542,660	2,846,620	3,864,518
Average contig size (bp)	11,702	23,534	24,061	25,319	25,999
Average intrascaffold gap size (bp)	2,560	2,487	2,426	2,213	2,082
Largest contig (bp)	1,224,073	1,224,073	1,224,073	1,224,073	1,224,073
% of total contigs	100	90	89	83	77

properly place a Celera read, so all reads were first masked against a library of common repetitive elements, and only matches of at least 40 bp to unmasked portions of the read constituted a hit. Of Celera's 27.27 million reads, 20.76 million matched a bactig and another 0.62 million reads, which did not have any matches, were nonetheless identified as belonging in the region of the bactig's BAC because their mate matched the bactig. Of the remaining reads, 2.92 million were completely screened out and so could not be matched, but the other 2.97 million reads had unmasked sequence totaling 1.189 Gbp that were not found in the GenBank data set. Because the Celera data are 5.11× redundant, we estimate that 240 Mbp of unique Celera sequence is not in the GenBank data set.

In the next step of the CSA process, a combining assembler took the relevant 5× Celera reads and bactigs for a BAC entry, and produced an assembly of the combined data for that locale. These high-quality sequence reconstructions were a transient result whose utility was simply to provide more reliable information for the purposes of their tiling into sets of overlapping and adjacent scaffold sequences in the next step. In outline, the combining assembler first examines the set of matching Celera reads to determine if there are excessive pileups indicative of unscreened repetitive elements. Wherever these occur, reads in the repeat region whose mates have not been mapped to consistent positions are removed. Then all sets of mate pairs that consistently imply the same relative position of two bactigs are bundled into a link and weighted according to the number of mates in the bundle. A "greedy" strategy then attempts to order the bactigs by selecting bundles of mate-pairs in order of their weight. A selected mate-pair bundle can tie together two formative scaffolds. It is incorporated to form a single scaffold only if it is consistent with the majority of links between contigs of the scaffold. Once scaffolding is complete, gaps are filled by the "Stones" strategy described above for the WGA assembler.

The GenBank data for the Phase 1 and 2 BACs consisted of an average of 19.8 bactigs per BAC of average size 8099 bp. Application of the combining assembler resulted in individual Celera BAC assemblies being put together into an average of 1.83 scaffolds (median of 1 scaffold) consisting of an average of 8.57 contigs of average size 18,973 bp. In addition to defining order and orientation of the sequence fragments, there were 57% fewer gaps in the combined result. For Phase 0 data, the average GenBank entry consisted of 91.52 reads of average length 784 bp. Application of the combining assembler resulted in an average of 54.8 scaffolds consisting of an average of 58.1 contigs of average size 873 bp. Basically, some small amount of assembly took place, but not enough Celera data were matched to truly assemble the 0.5× to 1× data set represented by the typical Phase 0 BACs. The combining assembler was also applied to the Phase 3 BACs for SNP identification, confirmation of assembly, and localization of the Celera reads. The phase 0 data suggest that a combined whole-genome shotgun data set and 1× light-shotgun of BACs will not yield good assembly of BAC regions; at least 3× light-shotgun of each BAC is needed.

The 5.89 million Celera fragments not matching the GenBank data were assembled with our whole-genome assembler. The assembly resulted in a set of scaffolds totaling 442 Mbp in span and consisting of 326 Mbp of sequence. More than 20% of the scaffolds were >5 kbp long, and these averaged 63% sequence and 27% gaps with a total of 302 Mbp of sequence. All scaffolds >5 kbp were forwarded along with all scaffolds produced by the combining assembler to the subsequent tiling phase.

At this stage, we typically had one or two scaffolds for every BAC region constituting at least 95% of the relevant sequence, and a collection of disjoint Celera-unique scaffolds. The next step in developing the genome components was to determine the order and overlap tiling of these BAC and Celera-unique scaffolds across the genome. For this, we used Celera's 50-kbp mate-pairs information, and BAC-end pairs (*18*) and sequence tagged site (STS) markers (*44*) to provide long-range guidance and chromosome separation. Given the relatively manageable number of scaffolds, we chose not to produce this tiling in a fully automated manner, but to compute an initial tiling with a good heuristic and then use human curators to resolve discrepancies or missed join opportunities. To this end, we developed a graphical user interface that displayed the graph of tiling overlaps and the evidence for each. A human curator could then explore the implication of mapped STS data, dot-plots of sequence overlap, and a visual display of the mate-pair evidence supporting a given choice. The result of this process was a collection of "components," where each component was a tiled set of BAC and Celera-unique scaffolds that had been curator-approved. The process resulted in 3845 components with an estimated span of 2.922 Gbp.

In order to generate the final CSA, we assembled each component with the WGA algorithm. As was done in the WGA process, the bactig data were shredded into a synthetic 2× shotgun data set in order to give the assembler the freedom to independently assemble the data. By using faux reads rather than bactigs, the assembly algorithm could correct errors in the assembly of bactigs and remove chimeric content in a PFP data entry. Chimeric or contaminating sequence (from another part of the genome) would not be incorporated into the reassembly of the component because it did not belong there. In effect, the previous steps in the CSA process served only to bring together Celera fragments and PFP data relevant to a large contiguous segment of the genome, wherein we applied the assembler used for WGA to produce an ab initio assembly of the region.

WGA assembly of the components resulted in a set of scaffolds totaling 2.906 Gbp in span and consisting of 2.654 Gbp of sequence. The chaff, or set of reads not incorporated into the assembly, numbered 6.17 million, or 22%. More than 90.0% of the genome was covered by scaffolds spanning >100 kbp long, and these averaged 92.2% sequence and 7.8% gaps with a total of 2.492 Gbp of sequence. There were a total of 105,264 gaps among the 107,199 contigs that belong to the 1940 scaffolds spanning >100 kbp. The average scaffold size was 1.4 Mbp, the average contig size was 23.24 kbp, and the average gap size was 2.0 kbp where each distribution of sizes was exponential. As such, averages tend to be underrepresentative of the majority of the data. Figure 5 shows a histogram of the bases in scaffolds of various size ranges. Consider also that more than 49% of all gaps were <500 bp long, more than 62% of all gaps were <1 kbp, and all gaps are <100 kbp long. Similarly, more than 73% of the sequence is in contigs > 30 kbp, more than 49% is in contigs >100 kbp, and the largest contig was 1.99 Mbp long. Table 3 provides summary statistics for the structure of this assembly with a direct comparison to the WGA assembly.

2.5 Comparison of the WGA and CSA scaffolds

Having obtained two assemblies of the human genome via independent computational processes (WGA and CSA), we compared scaffolds from the two assemblies as another means of investigating their completeness, consistency, and contiguity. From each assembly, a set of reference scaffolds containing at least 1000 fragments (Celera sequencing reads or bactig shreds) was obtained; this amounted to 2218 WGA scaffolds and 1717 CSA scaffolds, for a total of 2.087 Gbp and 2.474 Gbp. The sequence of each reference scaffold was compared to the sequence of all scaffolds from the other assembly with which it shared at least 20 fragments or at least 20% of the fragments of the smaller scaffold. For each such comparison, all matches of at least 200 bp with at most 2% mismatch were tabulated.

From this tabulation, we estimated the amount of unique sequence in each assembly in two ways. The first was to determine the number of bases of each assembly that were

not covered by a matching segment in the other assembly. Some 82.5 Mbp of the WGA (3.95%) was not covered by the CSA, whereas 204.5 Mbp (8.26%) of the CSA was not covered by the WGA. This estimate did not require any consistency of the assemblies or any uniqueness of the matching segments. Thus, another analysis was conducted in which matches of less than 1 kbp between a pair of scaffolds were excluded unless they were confirmed by other matches having a consistent order and orientation. This gives some measure of consistent coverage: 1.982 Gbp (95.00%) of the WGA is covered by the CSA, and 2.169 Gbp (87.69%) of the CSA is covered by the WGA by this more stringent measure.

The comparison of WGA to CSA also permitted evaluation of scaffolds for structural inconsistencies. We looked for instances in which a large section of a scaffold from one assembly matched only one scaffold from the other assembly, but failed to match over the full length of the overlap implied by the matching segments. An initial set of candidates was identified automatically, and then each candidate was inspected by hand. From this process, we identified 31 instances in which the assemblies appear to disagree in a nonlocal fashion. These cases are being further evaluated to determine which assembly is in error and why.

In addition, we evaluated local inconsistencies of order or orientation. The following results exclude cases in which one contig in one assembly corresponds to more than one overlapping contig in the other assembly (as long as the order and orientation of the latter agrees with the positions they match in the former). Most of these small rearrangements involved segments on the order of hundreds of base pairs and rarely >1 kbp. We found a total of 295 kbp (0.012%) in the CSA assemblies that were locally inconsistent with the WGA assemblies, whereas 2.108 Mbp (0.11%) in the WGA assembly were inconsistent with the CSA assembly.

The CSA assembly was a few percentage points better in terms of coverage and slightly more consistent than the WGA, because it was in effect performing a few thousand shotgun assemblies of megabase-sized problems, whereas the WGA is performing a shotgun assembly of a gigabase-sized problem. When one considers the increase of two-and-a-half orders of magnitude in problem size, the information loss between the two is remarkably small. Because CSA was logistically easier to deliver and the better of the two results available at the time when downstream analyses needed to be begun, all subsequent analysis was performed on this assembly.

2.6 Mapping scaffolds to the genome

The final step in assembling the genome was to order and orient the scaffolds on the chromosomes. We first grouped scaffolds together on the basis of their order in the components from CSA. These grouped scaffolds were reordered by examining residual mate-pairing data between the scaffolds. We next mapped the scaffold groups onto the chromosome using physical mapping data. This step depends on having reliable high-resolution map information such that each scaffold will overlap multiple markers. There are two genome-wide types of map information available: high-density STS maps and fingerprint maps of BAC clones developed at Washington University (*45*). Among the genome-wide STS maps, GeneMap99 (GM99) has the most markers and therefore was most useful for mapping scaffolds. The two different mapping approaches are complementary to one another. The fingerprint maps should have better local order because they were built by comparison of overlapping BAC clones. On the other hand, GM99 should have a more reliable long-range order, because the framework markers were derived from well-validated genetic maps. Both types of maps were used as a reference for human curation of the components that were the input to the regional assembly, but they did not determine the order of sequences produced by the assembler.

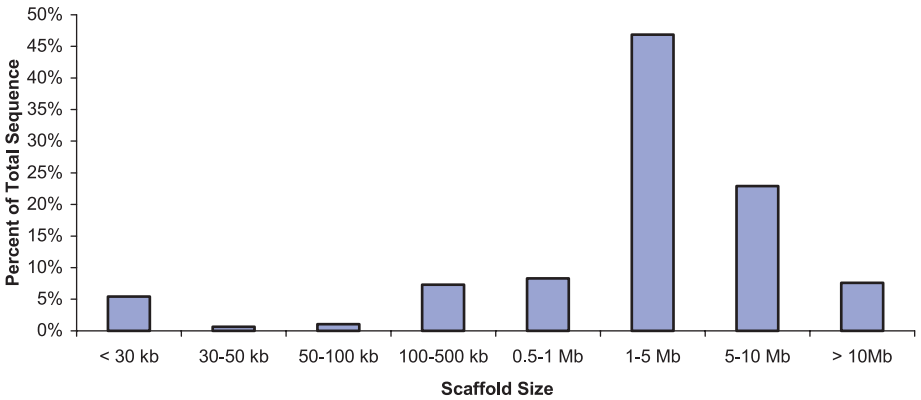


Fig. 5. Distribution of scaffold sizes of the CSA. For each range of scaffold sizes, the percent of total sequence is indicated.

In order to determine the effectiveness of the fingerprint maps and GM99 for mapping scaffolds, we first examined the reliability of these maps by comparison with large scaffolds. Only 1% of the STS markers on the 10 largest scaffolds (those >9 Mbp) were mapped on a different chromosome on GM99. Two percent of the STS markers disagreed in position by more than five framework bins. However, for the fingerprint maps, a 2% chromosome discrepancy was observed, and on average 23.8% of BAC locations in the scaffold sequence disagreed with fingerprint map placement by more than five BACs. When further examining the source of discrepancy, it was found that most of the discrepancy came from 4 of the 10 scaffolds, indicating this there is variation in the quality of either the map or the scaffolds. All four scaffolds were assembled, as well as the other six, as judged by clone coverage analysis, and showed the same low discrepancy rate to GM99, and thus we concluded that the fingerprint map global order in these cases was not reliable. Smaller scaffolds had a higher discordance rate with GM99 (4.21% of STSs were discordant by more than five framework bins), but a lower discordance rate with the fingerprint maps (11% of BACs disagreed with fingerprint maps by more than five BACs). This observation agrees with the clone coverage analysis (*46*) that Celera scaffold construction was better supported by long-range mate pairs in larger scaffolds than in small scaffolds.

We created two orderings of Celera scaffolds on the basis of the markers (BAC or STS) on these maps. Where the order of scaffolds agreed between GM99 and the WashU BAC map, we had a high degree of confidence that that order was correct; these scaffolds were termed "anchor scaffolds." Only scaffolds with a low overall discrepancy rate with both maps were considered anchor scaffolds. Scaffolds in GM99 bins were allowed to permute in their order to match WashU ordering, provided they did not violate their framework orders. Orientation of individual scaffolds was determined by the presence of multiple mapped markers with consistent order. Scaffolds with only one marker have insufficient information to assign orientation. We found 70.1% of the genome in anchored scaffolds, more than 99% of which are also oriented (Table 4). Because GM99 is of lower resolution than the WashU map, a number of scaffolds without STS matches could be ordered relative to the anchored scaffolds because they included sequence from the same or adjacent BACs on the WashU map. On the other hand, because of occasional WashU global ordering discrepancies, a number of scaffolds determined to be "unmappable" on the WashU map could be ordered relative to the anchored scaffolds

with GM99. These scaffolds were termed "ordered scaffolds." We found that 13.9% of the assembly could be ordered by these additional methods, and thus 84.0% of the genome was ordered unambiguously.

Next, all scaffolds that could be placed, but not ordered, between anchors were assigned to the interval between the anchored scaffolds and were deemed to be "bounded" between them. For example, small scaffolds having STS hits from the same GeneMap bin or hitting the same BAC cannot be ordered relative to each other, but can be assigned a placement boundary relative to other anchored or ordered scaffolds. The remaining scaffolds either had no localization information, conflicting information, or could only be assigned to a generic chromosome location. Using the above approaches, ~98% of the genome was anchored, ordered, or bounded.

Finally, we assigned a location for each scaffold placed on the chromosome by spreading out the scaffolds per chromosome. We assumed that the remaining unmapped scaffolds, constituting 2% of the genome, were distributed evenly across the genome. By dividing the sum of unmapped scaffold lengths with the sum of the number of mapped scaffolds, we arrived at an estimate of interscaffold gap of 1483 bp. This gap was used to separate all the scaffolds on each chromosome and to assign an offset in the chromosome.

During the scaffold-mapping effort, we encountered many problems that resulted in additional quality assessment and validation analysis. At least 978 (3% of 33,173) BACs were believed to have sequence data from more than one location in the genome (*47*). This is consistent with the bactig chimerism analysis reported above in the Assembly Strategies section. These BACs could not be assigned to unique positions within the CSA assembly and thus could not be used for ordering scaffolds. Likewise, it was not always possible to assign STSs to unique locations in the assembly because of genome duplications, repetitive elements, and pseudogenes.

Because of the time required for an exhaustive search for a perfect overlap, CSA generated 21,607 intrascaffold gaps where the mate-pair data suggested that the contigs should overlap, but no overlap was found. These gaps were defined as a fixed 50 bp in length and make up 18.6% of the total 116,442 gaps in the CSA assembly.

We chose not to use the order of exons implied in cDNA or EST data as a way of ordering scaffolds. The rationale for not using this data was that doing so would have biased certain regions of the assembly by rearranging scaffolds to fit the transcript data and made validation of both the assembly and gene definition processes more difficult.

2.7 Assembly and validation analysis

We analyzed the assembly of the genome from the perspectives of completeness (amount of coverage of the genome) and correctness (the structural accuracy of the order and orientation and the consensus sequence of the assembly).

Completeness. Completeness is defined as the percentage of the euchromatic sequence represented in the assembly. This cannot be known with absolute certainty until the euchromatin sequence has been completed. However, it is possible to estimate completeness on the basis of (i) the estimated sizes of intrascaffold gaps; (ii) coverage of the two published chromosomes, 21 and 22 (*48*, *49*); and (iii) analysis of the percentage of an independent set of random sequences (STS markers) contained in the assembly. The whole-genome libraries contain heterochromatic sequence and, although no attempt has been made to assemble it, there may be instances of unique sequence embedded in regions of heterochromatin as were observed in *Drosophila* (*50*, *51*).

The sequences of human chromosomes 21 and 22 have been completed to high quality and published (*48*, *49*). Although this sequence served as input to the assembler, the finished sequence was shredded into a shotgun data set so that the assembler had the opportunity to assemble it differently from the original sequence in the case of structural polymorphisms or assembly errors in the BAC data. In particular, the assembler must be able to resolve repetitive elements at the scale of components (generally multimegabase in size), and so this comparison reveals the level to which the assembler resolves repeats. In certain areas, the assembly structure differs from the published versions of chromosomes 21 and 22 (see below). The consequence of the flexibility to assemble "finished" sequence differently on the basis of Celera data resulted in an assembly with more segments than the chromosome 21 and 22 sequences. We examined the reasons why there are more gaps in the Celera sequence than in chromosomes 21 and 22 and expect that they may be typical of gaps in other regions of the genome. In the Celera assembly, there are 25 scaffolds, each containing at least 10 kb of sequence, that collectively span 94.3% of chromosome 21. Sixty-two scaffolds span 95.7% of chromosome 22. The total length of the gaps remaining in the Celera assembly for these two chromosomes is 3.4 Mbp. These gap sequences were analyzed by RepeatMasker and by searching against the entire genome assembly (*52*). About 50% of the gap sequence consisted of common repetitive elements identified by RepeatMasker; more than half of the remainder was lower copy number repeat elements.

A more global way of assessing completeness is to measure the content of an independent set of sequence data in the assembly. We compared 48,938 STS markers from Genemap99 (*51*) to the scaffolds. Because these markers were not used in the assembly processes, they provided a truly independent measure of completeness. ePCR (*53*) and BLAST (*54*) were used to locate STSs on the assembled genome. We found 44,524 (91%) of the STSs in the mapped genome. An additional 2648 markers (5.4%) were found by searching the unassembled data or "chaff." We identified 1283 STS markers (2.6%) not found in either Celera sequence or BAC data as of September 2000, raising the possibility that these markers may not be of human origin. If that were the case, the Celera assembled sequence would represent 93.4% of the human genome and the unassembled data 5.5%, for a total of 98.9% coverage. Similarly, we compared CSA against 36,678 TNG radiation hybrid markers (*55a*) using the same method. We found that 32,371 markers (88%) were located in the mapped CSA scaffolds, with 2055 markers (5.6%) found in the remainder. This gave a 94% coverage of the genome through another genome-wide survey.

Correctness. Correctness is defined as the structural and sequence accuracy of the assembly. Because the source sequences for the Celera data and the GenBank data are from different individuals, we could not directly compare the consensus sequence of the as-

Table 4. Summary of scaffold mapping. Scaffolds were mapped to the genome with different levels of confidence (anchored scaffolds have the highest confidence; unmapped scaffolds have the lowest). Anchored scaffolds were consistently ordered by the WashU BAC map and GM99. Ordered scaffolds were consistently ordered by at least one of the following: the WashU BAC map, GM99, or component tiling path. Bounded scaffolds had order conflicts between at least two of the external maps, but their placements were adjacent to a neighboring anchored or ordered scaffold. Unmapped scaffolds had, at most, a chromosome assignment. The scaffold subcategories are given below each category.

Mapped scaffold category	Number	Length (bp)	% Total length
Anchored	1,526	1,860,676,676	70
Oriented	1,246	1,852,088,645	70
Unoriented	280	8,588,031	0.3
Ordered	2,001	369,235,857	14
Oriented	839	329,633,166	12
Unoriented	1,162	39,602,691	2
Bounded	38,241	368,753,463	14
Oriented	7,453	274,536,424	10
Unoriented	30,788	94,217,039	4
Unmapped	11,823	55,313,737	2
Known chromosome	281	2,505,844	0.1
Unknown chromosome	11,542	52,807,893	2

sembly against other finished sequence for determining sequencing accuracy at the nucleotide level, although this has been done for identifying polymorphisms as described in Section 6. The accuracy of the consensus sequence is at least 99.96% on the basis of a statistical estimate derived from the quality values of the underlying reads.

The structural consistency of the assembly can be measured by mate-pair analysis. In a correct assembly, every mated pair of sequencing reads should be located on the consensus sequence with the correct separation and orientation between the pairs. A pair is termed "valid" when the reads are in the correct orientation and the distance between them is within the mean ± 3 standard deviations of the distribution of insert sizes of the library from which the pair was sampled. A pair is termed "misoriented" when the reads are not correctly oriented, and is termed "misseparated" when the distance between the reads is not in the correct range but the reads are correctly oriented. The mean ± the standard deviation of each library used by the assembler was determined as described above. To validate these, we examined all reads mapped to the finished sequence of chromosome 21 (*48*) and determined how many incorrect mate pairs there were as a result of laboratory tracking errors and chimerism (two different segments of the genome cloned into the same plasmid), and how tight the distribution of insert sizes was for those that were correct (Table 5). The standard deviations for all Celera libraries were quite small, less than 15% of the insert length, with the exception of a few 50-kbp libraries. The 2- and 10-kbp libraries contained less than 2% invalid mate pairs, whereas the 50-kbp libraries were somewhat higher (~10%). Thus, although the mate-pair information was not perfect, its accuracy was such that measuring valid, misoriented, and misseparated pairs with respect to a given assembly was deemed to be a reliable instrument for validation purposes, especially when several mate pairs confirm or deny an ordering.

The clone coverage of the genome was 39×, meaning that any given base pair was, on average, contained in 39 clones or, equivalently, spanned by 39 mate-paired reads. Areas of low clone coverage or areas with a high proportion of invalid mate pairs would indicate potential assembly problems. We computed the coverage of each base in the assembly by valid mate pairs (Table 6). In summary, for scaffolds >30 kbp in length, less than 1% of the Celera assembly was in regions of less than 3× clone coverage. Thus, more than 99% of the assembly, including order and orientation, is strongly supported by this measure alone.

We examined the locations and number of all misoriented and misseparated mates. In addition to doing this analysis on the CSA assembly (as of 1 October 2000), we also performed a study of the PFP assembly as of 5 September 2000 (*30*, *55b*). In this latter case, Celera mate pairs had to be mapped to the PFP assembly. To avoid mapping errors due to high-fidelity repeats, the only pairs mapped were those for which both reads matched at only one location with less than 6% differences. A threshold was set such that sets of five or more simultaneously invalid mate pairs indicated a potential breakpoint, where the construction of the two assemblies differed. The graphic comparison of the CSA chromosome 21 assembly with the published sequence (Fig. 6A) serves as a validation of this methodology. Blue tick marks in the panels indicate breakpoints. There were a similar (small) number of breakpoints on both chromosome sequences. The exception was 12 sets of scaffolds in the Celera assembly (a total of 3% of the chromosome length in 212 single-contig scaffolds) that were mapped to the wrong positions because they were too small to be mapped reliably. Figures 6 and 7 and Table 6 illustrate the mate-pair differences and breakpoints between the two assemblies. There was a higher percentage of misoriented and misseparated mate pairs in the large-insert libraries (50 kbp and BAC ends) than in the small-insert libraries in both assemblies (Table 6). The large-insert libraries are more likely to identify discrepancies simply because they span a larger segment of the genome. The graphic comparison between the two assemblies for chromosome 8 (Fig. 6, B and C) shows that there are many

Table 5. Mate-pair validation. Celera fragment sequences were mapped to the published sequence of chromosome 21. Each mate pair uniquely mapped was evaluated for correct orientation and placement (number of mate pairs tested). If the two mates had incorrect relative orientation or placement, they were considered invalid (number of invalid mate pairs).

		Chromosome 21						Genome		
Library type	Library no.	Mean insert size (bp)	SD (bp)	SD/ mean (%)	No. of mate pairs tested	No. of invalid mate pairs	% invalid	Mean insert size (bp)	SD (bp)	SD/ mean (%)
2 kbp	1	2,081	106	5.1	3,642	38	1.0	2,082	90	4.3
	2	1,913	152	7.9	28,029	413	1.5	1,923	118	6.1
	3	2,166	175	8.1	4,405	57	1.3	2,162	158	7.3
10 kbp	4	11,385	851	7.5	4,319	80	1.9	11,370	696	6.1
	5	14,523	1,875	12.9	7,355	156	2.1	14,142	1,402	9.9
	6	9,635	1,035	10.7	5,573	109	2.0	9,606	934	9.7
	7	10,223	928	9.1	34,079	399	1.2	10,190	777	7.6
50 kbp	8	64,888	2,747	4.2	16	1	6.3	65,500	5,504	8.4
	9	53,410	5,834	10.9	914	170	18.6	53,311	5,546	10.4
	10	52,034	7,312	14.1	5,871	569	9.7	51,498	6,588	12.8
	11	52,282	7,454	14.3	2,629	213	8.1	52,282	7,454	14.3
	12	46,616	7,378	15.8	2,153	215	10.0	45,418	9,068	20.0
	13	55,788	10,099	18.1	2,244	249	11.1	53,062	10,893	20.5
	14	39,894	5,019	12.6	199	7	3.5	36,838	9,988	27.1
BES	15	48,931	9,813	20.1	144	10	6.9	47,845	4,774	10.0
	16	48,130	4,232	8.8	195	14	7.2	47,924	4,581	9.6
	17	106,027	27,778	26.2	330	16	4.8	152,000	26,600	17.5
	18	160,575	54,973	34.2	155	8	5.2	161,750	27,000	16.7
	19	164,155	19,453	11.9	642	44	6.9	176,500	19,500	11.05
Sum					102,894	2,768 (mean = 2.7)	2.7			

more breakpoints for the PFP assembly than for the Celera assembly. Figure 7 shows the breakpoint map (blue tick marks) for both assemblies of each chromosome in a side-by-side fashion. The order and orientation of Celera's assembly shows substantially fewer breakpoints except on the two finished chromosomes. Figure 7 also depicts large gaps (>10 kbp) in both assemblies as red tick marks. In the CSA assembly, the size of all gaps have been estimated on the basis of the mate-pair data. Breakpoints can be caused by structural polymorphisms, because the two assemblies were derived from different human genomes. They also reflect the unfinished nature of both genome assemblies.

3 Gene Prediction and Annotation

Summary. To enumerate the gene inventory, we developed an integrated, evidence-based approach named Otto. The evidence used to increase the likelihood of identifying genes includes regions conserved between the mouse and human genomes, similarity to ESTs or other mRNA-derived data, or similarity to other proteins. A comparison of Otto (combined Otto-RefSeq and Otto homology) with Genscan, a standard gene-prediction algorithm, showed greater sensitivity (0.78 versus 0.50) and specificity (0.93 versus 0.63) of Otto in the ability to define gene structure. Otto-predicted genes were complemented with a set of genes from three gene-prediction programs that exhibited weaker, but still significant, evidence that they may be expressed. Conservative criteria, requiring at least two lines of evidence, were used to define a set of 26,383 genes with good confidence that were used for more detailed analysis presented in the subsequent sections. Extensive manual curation to establish precise characterization of gene structure will be necessary to improve the results from this initial computational approach.

3.1 Automated gene annotation

A gene is a locus of cotranscribed exons. A single gene may give rise to multiple transcripts, and thus multiple distinct proteins with multiple functions, by means of alternative splicing and alternative transcription initiation and termination sites. Our cells are able to discern within the billions of base pairs of the genomic DNA the signals for initiating transcription and for splicing together exons separated by a few or hundreds of thousands of base pairs. The first step in characterizing the genome is to define the structure of each gene and each transcription unit.

The number of protein-coding genes in mammals has been controversial from the outset. Initial estimates based on reassociation data placed it between 30,000 to 40,000, whereas later estimates from the brain were >100,000 (*56*). More recent data from both the corporate and public sectors, based on extrapolations from EST, CpG island, and transcript density–based extrapolations, have not reduced this variance. The highest recent number of 142,634 genes emanates from a report from Incyte Pharmaceuticals, and is based on a combination of EST data and the association of ESTs with CpG islands (*57*). In stark contrast are three quite different, and much lower estimates: one of ~35,000 genes derived with genome-wide EST data and sampling procedures in conjunction with chromosome 22 data (*58*); another of 28,000 to 34,000 genes derived with a comparative methodology involving sequence conservation between humans and the puffer fish *Tetraodon nigroviridis* (*59*); and a figure of 35,000 genes, which was derived simply by extrapolating from the density of 770 known and predicted genes in the 67 Mbp of chromosomes 21 and 22, to the approximately 3-Gbp euchromatic genome.

The problem of computational identification of transcriptional units in genomic DNA sequence can be divided into two phases. The first is to partition the sequence into segments that are likely to correspond to individual genes. This is not trivial and is a weakness of most de novo gene-finding algorithms. It is also critical to determining the number of genes in the human gene inventory. The second challenge is to construct a gene model that reflects the probable structure of the transcript(s) encoded in the region. This can be done with reasonable accuracy when a full-length cDNA has been sequenced or a highly homologous protein sequence is known. De novo gene prediction, although less accurate, is the only way to find genes that are not represented by homologous proteins or ESTs. The following section describes the methods we have developed to address these problems for the prediction of protein-coding genes.

We have developed a rule-based expert system, called Otto, to identify and characterize genes in the human genome (*60*). Otto attempts to simulate in software the process that a human annotator uses to identify a gene and refine its structure. In the process of annotating a region of the genome, a human curator examines the evidence provided by the computational pipeline (described below) and examines how various types of evidence relate to one another. A curator puts different levels of confidence in different types of evidence and looks for certain patterns of evidence to support gene annotation. For example, a curator may examine homology to a number of ESTs and evaluate whether or not they can be connected into a longer, virtual mRNA. The curator would also evaluate the strength of the similarity and the contiguity of the match, in essence asking whether any ESTs cross splice-junctions and whether the edges of putative exons have consensus splice sites. This kind of manual annotation process was used to annotate the *Drosophila* genome.

The Otto system can promote observed evidence to a gene annotation in one of two ways. First, if the evidence includes a high-quality match to the sequence of a known gene [here defined as a human gene represented in a curated subset of the RefSeq database (*61*)], then Otto can promote this to a gene annotation. In the second method, Otto evaluates a broad spectrum of evidence and determines if this evidence is adequate to support promotion to a gene annotation. These processes are described below.

Initially, gene boundaries are predicted on the basis of examination of sets of overlapping protein and EST matches generated by a computational pipeline (*62*). This pipeline searches the scaffold sequences against protein, EST, and genome-sequence databases to define regions of sequence similarity and runs three de novo gene-prediction programs.

To identify likely gene boundaries, regions of the genome were partitioned by Otto on the basis of sequence matches identified by BLAST. Each of the database sequences matched in the region under analysis was compared by an algorithm that takes into account both coordinates of the matching sequence, as well as the sequence type (e.g., protein, EST, and so forth). The results were used to group the matches into bins of related sequences that may define a gene and identify

Table 6. Genome-wide mate pair analysis of compartmentalized shotgun (CSA) and PFP assemblies.*

Genome library	CSA			PFP		
	% valid	% mis-oriented	% mis-separated†	% valid	% mis-oriented	% mis-separated†
2 kbp	98.5	0.6	1.0	95.7	2.0	2.3
10 kbp	96.7	1.0	2.3	81.9	9.6	8.6
50 kbp	93.9	4.5	1.5	64.2	22.3	13.5
BES	94.1	2.1	3.8	62.0	19.3	18.8
Mean	97.4	1.0	1.6	87.3	6.8	5.9

*Data for individual chromosomes can be found in Web fig. 3 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1. †Mates are misseparated if their distance is >3 SD from the mean library size.

gene boundaries. During this process, multiple hits to the same region were collapsed to a coherent set of data by tracking the coverage of a region. For example, if a group of bases was represented by multiple overlapping ESTs, the union of these regions matched by the set of ESTs on the scaffold was marked as being supported by EST evidence. This resulted in a series of "gene bins," each of which was believed to contain a single gene. One weakness of this initial implementation of the algorithm was in predicting gene boundaries in regions of tandemly duplicated genes. Gene clusters frequently resulted in homologous neighboring genes being joined together, resulting in an annotation that artificially concatenated these gene models.

Next, known genes (those with exact matches of a full-length cDNA sequence to the genome) were identified, and the region corresponding to the cDNA was annotated as a predicted transcript. A subset of the curated human gene set RefSeq from the National Center for Biotechnology Information (NCBI) was included as a data set searched in the computational pipeline. If a RefSeq transcript matched the genome assembly for at least 50% of its length at >92% identity, then the SIM4 (*63*) alignment of the RefSeq transcript to the region of the genome under analysis was promoted to the status of an Otto annotation. Because the genome sequence has gaps and sequence errors such as frameshifts, it was not always possible to predict a transcript that agrees precisely with the experimentally determined cDNA sequence. A total of 6538 genes in our inventory were identified and transcripts predicted in this way.

Regions that have a substantial amount of sequence similarity, but do not match known genes, were analyzed by that part of the Otto system that uses the sequence similarity information to predict a transcript. Here, Otto

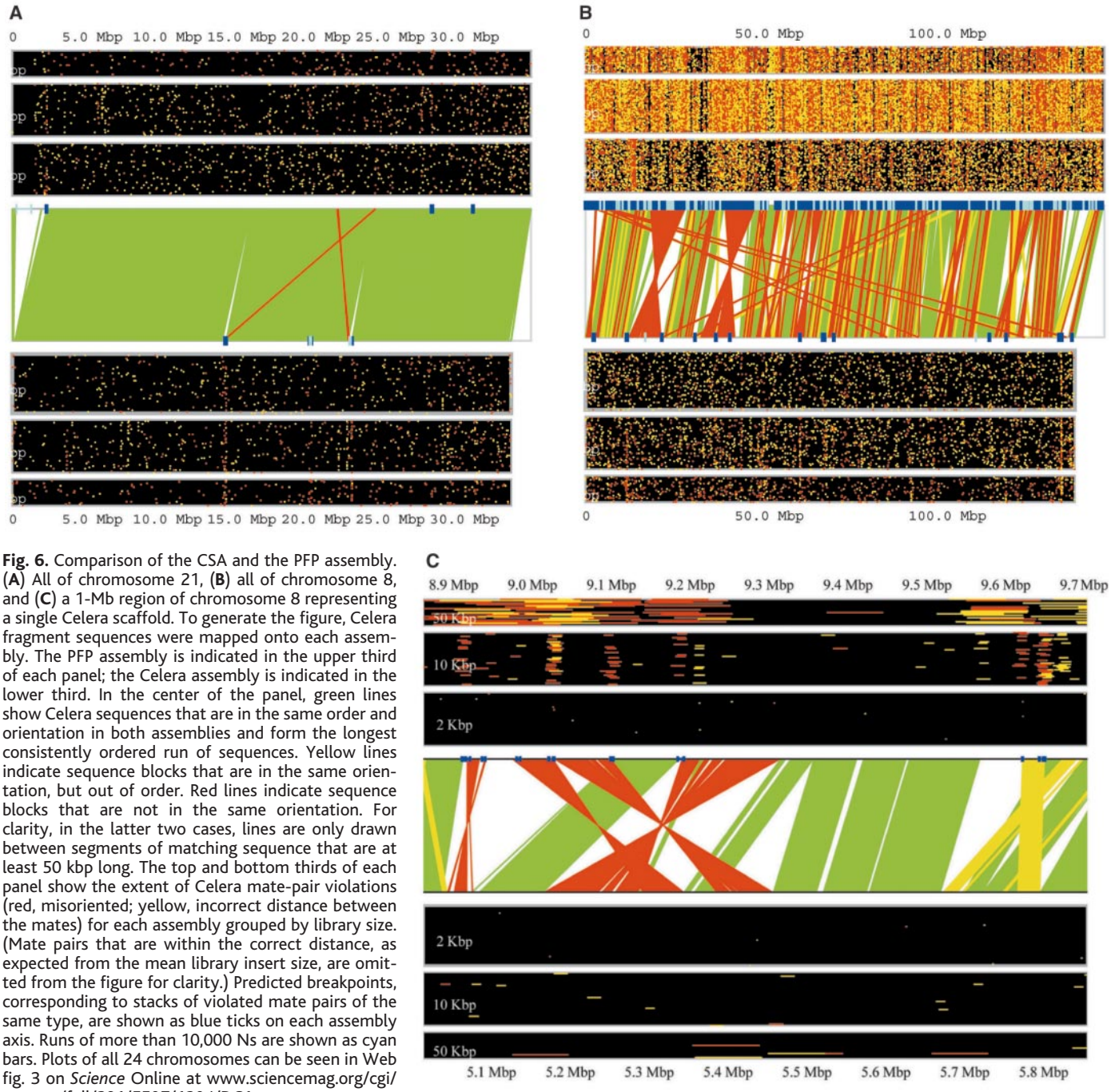


Fig. 6. Comparison of the CSA and the PFP assembly. (**A**) All of chromosome 21, (**B**) all of chromosome 8, and (**C**) a 1-Mb region of chromosome 8 representing a single Celera scaffold. To generate the figure, Celera fragment sequences were mapped onto each assembly. The PFP assembly is indicated in the upper third of each panel; the Celera assembly is indicated in the lower third. In the center of the panel, green lines show Celera sequences that are in the same order and orientation in both assemblies and form the longest consistently ordered run of sequences. Yellow lines indicate sequence blocks that are in the same orientation, but out of order. Red lines indicate sequence blocks that are not in the same orientation. For clarity, in the latter two cases, lines are only drawn between segments of matching sequence that are at least 50 kbp long. The top and bottom thirds of each panel show the extent of Celera mate-pair violations (red, misoriented; yellow, incorrect distance between the mates) for each assembly grouped by library size. (Mate pairs that are within the correct distance, as expected from the mean library insert size, are omitted from the figure for clarity.) Predicted breakpoints, corresponding to stacks of violated mate pairs of the same type, are shown as blue ticks on each assembly axis. Runs of more than 10,000 Ns are shown as cyan bars. Plots of all 24 chromosomes can be seen in Web fig. 3 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1.

evaluates evidence generated by the computational pipeline, corresponding to conservation between mouse and human genomic DNA, similarity to human transcripts (ESTs and cDNAs), similarity to rodent transcripts (ESTs and cDNAs), and similarity of the translation of human genomic DNA to known proteins to predict potential genes in the human genome. The sequence from the region of genomic DNA contained in a gene bin was extracted, and the subsequences supported by any homology evidence were marked (plus 100

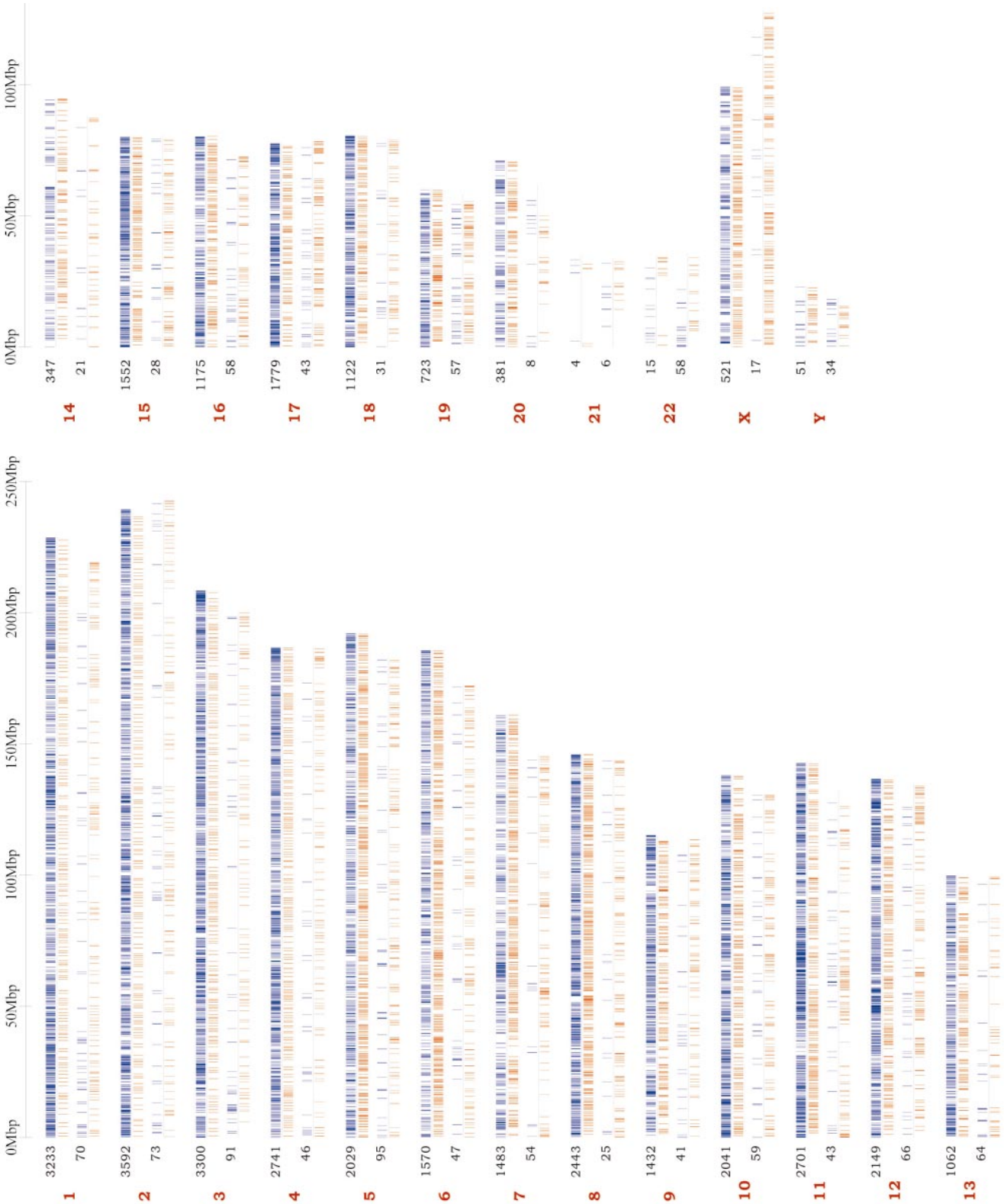


Fig. 7. Schematic view of the distribution of breakpoints and large gaps on all chromosomes. For each chromosome, the upper pair of lines represent the PFP assembly, and the lower pair of lines represent Celera's assembly. Blue tick marks represent breakpoints, whereas red tick marks represent a gap of larger than 10,000 bp. The number of breakpoints per chromosome is indicated in black, and the chromosome numbers in red.

bases flanking these regions). The other bases in the region, those not covered by any homology evidence, were replaced by N's. This sequence segment, with high confidence regions represented by the consensus genomic sequence and the remainder represented by N's, was then evaluated by Genscan to see if a consistent gene model could be generated. This procedure simplified the gene-prediction task by first establishing the boundary for the gene (not a strength of most gene-finding algorithms), and by eliminating regions with no supporting evidence. If Genscan returned a plausible gene model, it was further evaluated before being promoted to an "Otto" annotation. The final Genscan predictions were often quite different from the prediction that Genscan returned on the same region of native genomic sequence. A weakness of using Genscan to refine the gene model is the loss of valid, small exons from the final annotation.

The next step in defining gene structures based on sequence similarity was to compare each predicted transcript with the homology-based evidence that was used in previous steps to evaluate the depth of evidence for each exon in the prediction. Internal exons were considered to be supported if they were covered by homology evidence to within ±10 bases of their edges. For first and last exons, the internal edge was required to be within 10 bases, but the external edge was allowed greater latitude to allow for 5′ and 3′ untranslated regions (UTRs). To be retained, a prediction for a multi-exon gene must have evidence such that the total number of "hits," as defined above, divided by the number of exons in the prediction must be >0.66 or must correspond to a RefSeq sequence. A single-exon gene must be covered by at least three supporting hits (±10 bases on each side), and these must cover the complete predicted open reading frame. For a single-exon gene, we also required that the Genscan prediction include both a start and a stop codon. Gene models that did not meet these criteria were disregarded, and those that passed were promoted to Otto predictions. Homology-based Otto predictions do not contain 3′ and 5′ untranslated sequence. Although three de novo gene-finding programs [GRAIL, Genscan, and FgenesH (*63*)] were run as part of the computational analysis, the results of these programs were not directly used in making the Otto predictions. Otto predicted 11,226 additional genes by means of sequence similarity.

Table 7. Sensitivity and specificity of Otto and Genscan. Sensitivity and specificity were calculated by first aligning the prediction to the published RefSeq transcript, tallying the number (*N*) of uniquely aligned RefSeq bases. Sensitivity is the ratio of *N* to the length of the published RefSeq transcript. Specificity is the ratio of *N* to the length of the prediction. All differences are significant (Tukey HSD; $P < 0.001$).

Method	Sensitivity	Specificity
Otto (RefSeq only)*	0.939	0.973
Otto (homology)†	0.604	0.884
Genscan	0.501	0.633

*Refers to those annotations produced by Otto using only the Sim4-polished RefSeq alignment rather than an evidence-based Genscan prediction. †Refers to those annotations produced by supplying all available evidence to Genscan.

3.2 Otto validation

To validate the Otto homology-based process and the method that Otto uses to define the structures of known genes, we compared transcripts predicted by Otto with their corresponding (and presumably correct) transcript from a set of 4512 RefSeq transcripts for which there was a unique SIM4 alignment (Table 7). In order to evaluate the relative performance of Otto and Genscan, we made three comparisons. The first involved a determination of the accuracy of gene models predicted by Otto with only homology data other than the corresponding RefSeq sequence (Otto homology in Table 7). We measured the sensitivity (correctly predicted bases divided by the total length of the cDNA) and specificity (correctly predicted bases divided by the sum of the correctly and incorrectly predicted bases). Second, we examined the sensitivity and specificity of the Otto predictions that were made solely with the RefSeq sequence, which is the process that Otto uses to annotate known genes (Otto-RefSeq). And third, we determined the accuracy of the Genscan predictions corresponding to these RefSeq sequences. As expected, the alignment method (Otto-RefSeq) was the most accurate, and Otto-homology performed better than Genscan by both criteria. Thus, 6.1% of true RefSeq nucleotides were not represented in the Otto-refseq annotations and 2.7% of the nucleotides in the Otto-RefSeq transcripts were not contained in the original RefSeq transcripts. The discrepancies could come from legitimate differences between the Celera assembly and the RefSeq transcript due to polymorphisms, incomplete or incorrect data in the Celera assembly, errors introduced by Sim4 during the alignment process, or the presence of alternatively spliced forms in the data set used for the comparisons.

Because Otto uses an evidence-based approach to reconstruct genes, the absence of experimental evidence for intervening exons may inadvertantly result in a set of exons that cannot be spliced together to give rise to a transcript. In such cases, Otto may "split genes" when in fact all the evidence should be combined into a single transcript. We also examined the tendency of these methods to incorrectly split gene predictions. These trends are shown in Fig. 8. Both RefSeq and homology-based predictions by Otto split known genes into fewer segments than Genscan alone.

3.3 Gene number

Recognizing that the Otto system is quite conservative, we used a different gene-prediction strategy in regions where the homology evidence was less strong. Here the results of de novo gene predictions were used. For these genes, we insisted that a predicted transcript have at least two of the following types of evidence to be included in the gene set for further analysis: protein, human EST, rodent EST, or mouse genome fragment matches. This final class of predicted genes is a subset of the predictions made by the three gene-finding programs that were used in the computational pipeline. For these, there was not sufficient sequence similarity information for Otto to attempt to predict a gene structure. The three de novo gene-finding programs resulted in about 155,695 predictions, of which ~76,410 were nonredundant (nonoverlapping with one another). Of these, 57,935 did not overlap known genes or predictions made by Otto. Only 21,350 of the gene predictions that did not overlap Otto predictions were partially supported by at least one type of sequence similarity evidence, and 8619 were partially supported by two types of evidence (Table 8).

The sum of this number (21,350) and the number of Otto annotations (17,764), 39,114, is near the upper limit for the human gene complement. As seen in Table 8, if the requirement for other supporting evidence is made more stringent, this number drops rapidly so that demanding two types of evidence reduces the total gene number to 26,383 and demanding three types reduces it to ~23,000. Requiring that a prediction be supported by all four categories of evidence is too stringent because it would eliminate genes that encode novel proteins (members of currently undescribed protein families). No correction for pseudogenes has been made at this point in the analysis.

In a further attempt to identify genes that were not found by the autoannotation process or any of the de novo gene finders, we examined regions outside of gene predictions that were similar to the EST sequence, and where the EST matched the genomic sequence across a splice junction. After correcting for potential 3′ UTRs of predicted genes, about 2500 such regions remained. Addition of a requirement for at least one of the following evidence types—homology to mouse genomic sequence fragments, rodent ESTs, or cDNAs—or similarity to a known protein reduced this number to 1010. Adding this to the numbers from the previous paragraph would give us estimates of about 40,000, 27,000, and 24,000 potential genes in the human genome, depending on the stringency of evidence considered. Table 8 illustrates the number of genes and presents the degree of

confidence based on the supporting evidence. Transcripts encoded by a set of 26,383 genes were assembled for further analysis. This set includes the 6538 genes predicted by Otto on the basis of matches to known genes, 11,226 transcripts predicted by Otto based on homology evidence, and 8619 from the subset of transcripts from de novo gene-prediction programs that have two types of supporting evidence. The 26,383 genes are illustrated along chromosome diagrams in Fig. 1. These are a very preliminary set of annotations and are subject to all the limitations of an automated process. Considerable refinement is still necessary to improve the accuracy of these transcript predictions. All the predictions and descriptions of genes and the associated evidence that we present are the product of completely computational processes, not expert curation. We have attempted to enumerate the genes in the human genome in such a way that we have different levels of confidence based on the amount of supporting evidence: known genes, genes with good protein or EST homology evidence, and de novo gene predictions confirmed by modest homology evidence.

3.4 Features of human gene transcripts

We estimate the average span for a "typical" gene in the human DNA sequence to be about 27,894 bases. This is based on the average span covered by RefSeq transcripts, used because it represents our highest confidence set.

The set of transcripts promoted to gene annotations varies in a number of ways. As can be seen from Table 8 and Fig. 9, transcripts predicted by Otto tend to be longer, having on average about 7.8 exons, whereas those promoted from gene-prediction programs average about 3.7 exons. The largest number of exons that we have identified in a transcript is 234 in the titin mRNA. Table 8 compares the amounts of evidence that support the Otto and other predicted transcripts. For example, one can see that a typical Otto transcript has 6.99 of its 7.81 exons supported by protein homology evidence. As would be expected, the Otto transcripts generally have more support than do transcripts predicted by the de novo methods.

4 Genome Structure

Summary. This section describes several of the noncoding attributes of the assembled genome sequence and their correlations with the predicted gene set. These include an analysis of G+C content and gene density in the context of cytogenetic maps of the genome, an enumerative analysis of CpG islands, and a brief description of the genome-wide repetitive elements.

4.1 Cytogenetic maps

Perhaps the most obvious, and certainly the most visible, element of the structure of the genome is the banding pattern produced by Giemsa stain. Chromosomal banding studies have revealed that about 17% to 20% of the human chromosome complement consists of C-bands, or constitutive heterochromatin (*64*). Much of this heterochromatin is highly polymorphic and consists of different families of alpha satellite DNAs with various higher order repeat structures (*65*). Many chromosomes have complex inter- and intrachromosomal duplications present in pericentromeric regions (*66*). About 5% of the sequence reads were identified as alpha satellite sequences; these were not included in the assembly.

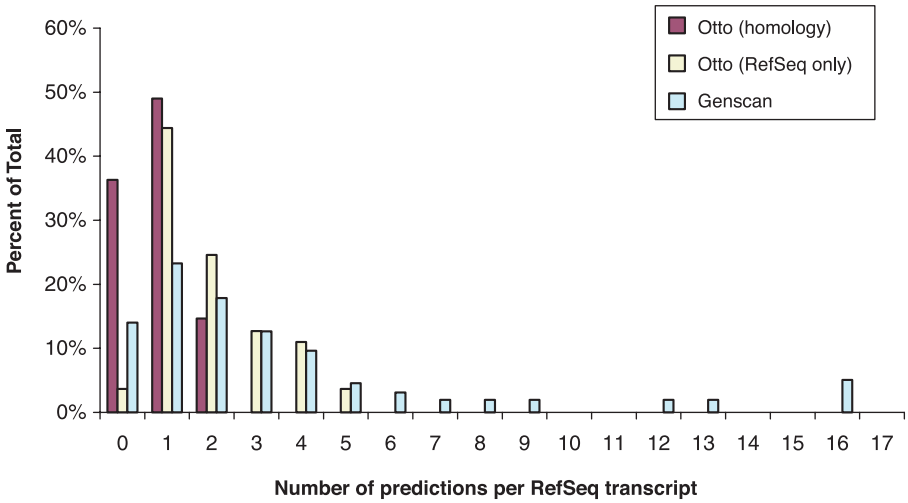


Fig. 8. Analysis of split genes resulting from different annotation methods. A set of 4512 Sim4-based alignments of RefSeq transcripts to the genomic assembly were chosen (see the text for criteria), and the numbers of overlapping Genscan, Otto (RefSeq only) annotations based solely on Sim4-polished RefSeq alignments, and Otto (homology) annotations (annotations produced by supplying all available evidence to Genscan) were tallied. These data show the degree to which multiple Genscan predictions and/or Otto annotations were associated with a single RefSeq transcript. The zero class for the Otto-homology predictions shown here indicates that the Otto-homology calls were made without recourse to the RefSeq transcript, and thus no Otto call was made because of insufficient evidence.

Table 8. Numbers of exons and transcripts supported by various types of evidence for Otto and de novo gene prediction methods. Highlighted cells indicate the gene sets analyzed in this paper (boldface, set of genes selected for protein analysis; italic, total set of accepted de novo predictions).

		Total	Types of evidence				No. of lines of evidence*			
			Mouse	Rodent	Protein	Human	≥1	≥2	≥3	≥4
Otto	Number of transcripts	17,969	17,065	14,881	15,477	16,374	**17,968**†	17,501	15,877	12,451
	Number of exons	141,218	111,174	89,569	108,431	118,869	140,710	127,955	99,574	59,804
De novo	Number of transcripts	58,032	14,463	5,094	8,043	9,220	*21,350*	**8,619**	4,947	1,904
	Number of exons	319,935	48,594	19,344	26,264	40,104	79,148	31,130	17,508	6,520
No. of exons per transcript	Otto	7.84	5.77	6.01	6.99	7.24	7.81	7.19	6.00	4.28
	De novo	5.53	3.17	3.80	3.27	4.36	3.7	3.56	3.42	3.16

*Four kinds of evidence (conservation in 3× mouse genomic DNA, similarity to human EST or cDNA, similarity to rodent EST or cDNA, and similarity to known proteins) were considered to support gene predictions from the different methods. The use of evidence is quite liberal, requiring only a partial match to a single exon of predicted transcript. †This number includes alternative splice forms of the 17,764 genes mentioned elsewhere in the text.

Examination of pericentromeric regions is ongoing.

The remaining ~80% of the genome, the euchromatic component, is divisible into G-, R-, and T-bands (*67*). These cytogenetic bands have been presumed to differ in their nucleotide composition and gene density, although we have been unable to determine precise band boundaries at the molecular level. T-bands are the most G+C- and gene-rich, and G-bands are G+C-poor (*68*). Bernardi has also offered a description of the euchromatin at the molecular level as long stretches of DNA of differing base composition, termed isochores (denoted L, H1, H2, and H3), which are >300 kbp in length (*69*). Bernardi defined the L (light) isochores as G+C-poor (<43%), whereas the H (heavy) isochores fall into three G+C-rich classes representing 24, 8, and 5% of the genome. Gene concentration has been claimed to be very low in the L isochores and 20-fold more enriched in the H2 and H3 isochores (*70*). By examining contiguous 50-kbp windows of G+C content across the assembly, we found that regions of G+C content >48% (H3 isochores) averaged 273.9 kbp in length, those with G+C content between 43 and 48% (H1+H2 isochores) averaged 202.8 kbp in length, and the average span of regions with <43% (L isochores) was 1078.6 kbp. The correlation between G+C content and gene density was also examined in 50-kbp windows along the assembled sequence (Table 9 and Figs. 10 and 11). We found that the density of genes was greater in regions of high G+C than in regions of low G+C content, as expected. However, the correlation between G+C content and gene density was not as skewed as previously predicted (*69*). A higher proportion of genes were located in the G+C-poor regions than had been expected.

Chromosomes 17, 19, and 22, which have a disproportionate number of H3-containing bands, had the highest gene density (Table 10). Conversely, of the chromosomes that we found to have the lowest gene density, X, 4, 18, 13, and Y, also have the fewest H3 bands. Chromosome 15, which also has few H3 bands, did not have a particularly low gene density in our analysis. In addition, chromosome 8, which we found to have a low gene density, does not appear to be unusual in its H3 banding.

How valid is Ohno's postulate (*71*) that mammalian genomes consist of oases of genes in otherwise essentially empty deserts? It appears that the human genome does indeed contain deserts, or large, gene-poor regions. If we define a desert as a region >500 kbp without a gene, then we see that 605 Mbp, or about 20% of the genome, is in deserts. These are not uniformly distributed over the various chromosomes. Gene-rich chromosomes 17, 19, and 22 have only about 12% of their collective 171 Mbp in deserts, whereas gene-poor chromosomes 4, 13, 18, and X have 27.5% of their 492 Mbp in deserts (Table 11). The apparent lack of predicted genes in these regions does not necessarily imply that they are devoid of biological function.

4.2 Linkage map

Linkage maps provide the basis for genetic analysis and are widely used in the study of the inheritance of traits and in the positional cloning of genes. The distance metric, centimorgans (cM), is based on the recombination rate between homologous chromosomes during meiosis. In general, the rate of recombination in females is greater than that in males, and this degree of map expansion is not uniform across the genome (*72*). One of the opportunities enabled by a nearly complete genome sequence is to produce the ultimate physical map, and to fully analyze its correspondence with two other maps that have been widely used in genome and genetic analysis: the linkage map and the cytogenetic map. This would close the loop between the mapping and sequencing phases of the genome project.

We mapped the location of the markers that constitute the Genethon linkage map to the genome. The rate of recombination, expressed as cM per Mbp, was calculated for 3-Mbp windows as shown in Table 12. Higher rates of recombination in the telomeric region of the chromosomes have been previously documented (*73*). From this mapping result, there is a difference of 4.99 between lowest rates and highest rates and the largest difference of 4.4 between males and females (4.99 to 0.47 on chromosome 16). This indicates that the variability in recombination rates among regions of the genome exceeds the differences in recombination rates between males and females. The human genome has recombination hotspots, where recombination rates vary fivefold or more over a space of 1 kbp, so the picture one gets of the magnitude of variability in recombination rate will depend on the size of the window

Table 9. Characteristics of G+C in isochores.

Isochore	G+C (%)	Fraction of genome		Fraction of genes	
		Predicted*	Observed	Predicted*	Observed
H3	>48	5	9.5	37	24.8
H1/H2	43–48	25	21.2	32	26.6
L	<43	67	69.2	31	48.5

*The predictions were based on Bernardi's definitions (*70*) of the isochore structure of the human genome.

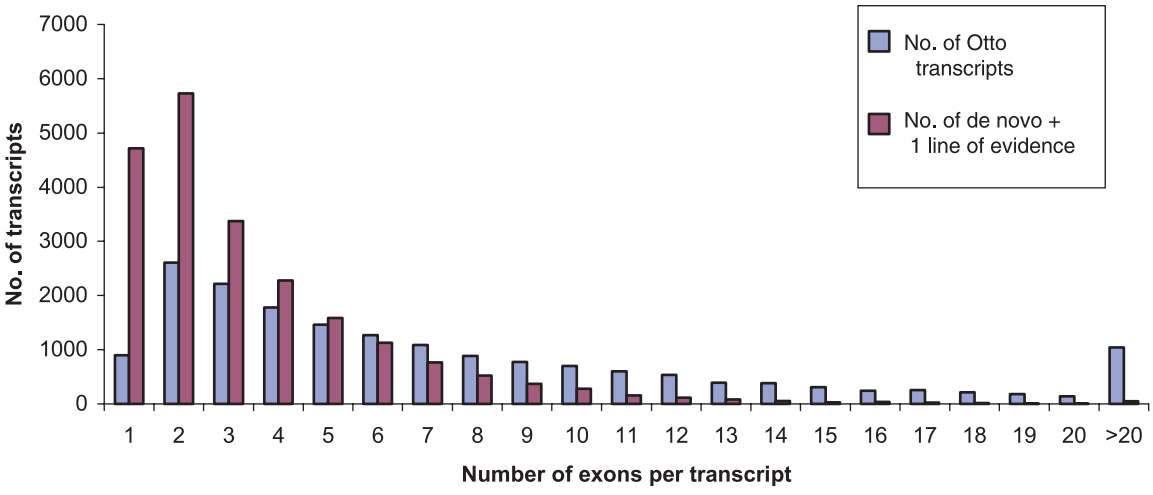


Fig. 9. Comparison of the number of exons per transcript between the 17,968 Otto transcripts and 21,350 de novo transcript predictions with at least one line of evidence that do not overlap with an Otto prediction. Both sets have the highest number of transcripts in the two-exon category, but the de novo gene predictions are skewed much more toward smaller transcripts. In the Otto set, 19.7% of the transcripts have one or two exons, and 5.7% have more than 20. In the de novo set, 49.3% of the transcripts have one or two exons, and 0.2% have more than 20.

examined. Unfortunately, too few meiotic crossovers have occurred in Centre d'Étude du Polymorphism Humain (CEPH) and other reference families to provide a resolution any finer than about 3 Mbp. The next challenge will be to determine a sequence basis of recombination at the chromosomal level. An accurate predictor for the rate for variation in recombination rates between any pair of markers would be extremely useful in designing markers to narrow a region of linkage, such as in positional cloning projects.

4.3 Correlation between CpG islands and genes

CpG islands are stretches of unmethylated DNA with a higher frequency of CpG dinucleotides when compared with the entire genome (*74*). CpG islands are believed to preferentially occur at the transcriptional start of genes, and it has been observed that most housekeeping genes have CpG islands at the 5′ end of the transcript (*75*, *76*). In addition, experimental evidence indicates that CpG island methylation is correlated with gene inactivation (*77*) and has been shown to be important during gene imprinting (*78*) and tissue-specific gene expression (*79*)

Experimental methods have been used that resulted in an estimate of 30,000 to 45,000 CpG islands in the human genome (*74*, *80*) and an estimate of 499 CpG islands on human chromosome 22 (*81*). Larsen *et al.* (*76*) and Gardiner-Garden and Frommer (*75*) used a computational method to identify CpG islands and defined them as regions of DNA of >200 bp that have a G+C content of >50% and a ratio of observed versus expected frequency of CG dinucleotide ≥0.6.

It is difficult to make a direct comparison of experimental definitions of CpG islands with computational definitions because computational methods do not consider the methylation state of cytosine and experimental methods do not directly select regions of high G+C content. However, we can determine the correlation of CpG island with gene starts, given a set of annotated genomic transcripts and the whole genome sequence. We have analyzed the publicly available annotation of chromosome 22, as well as using the entire human genome in our assembly and the computationally annotated genes. A variation of the CpG island computation was compared with Larsen *et al*. (*76*). The main differences are that we use a sliding window of 200 bp, consecutive windows are merged only if they overlap, and we recompute the CpG value upon merging, thus rejecting any potential island if it scores less than the threshold.

To compute various CpG statistics, we used two different thresholds of CG dinucleotide likelihood ratio. Besides using the original threshold of 0.6 (method 1), we used a higher threshold of CG dinucleotide likelihood ratio of 0.8 (method 2), which results in the number of CpG islands on chromosome 22 close to the number of annotated genes on this chromosome. The main results are summarized in Table 13. CpG islands computed with method 1 predicted only 2.6% of the CSA sequence as CpG, but 40% of the gene starts (start codons) are contained inside a CpG island. This is comparable to ratios reported by others (*82*). The last two rows of the table show the observed and expected average distance, respectively, of the closest CpG island from the first exon. The observed average closest CpG islands are smaller than the corresponding expected distances, confirming an association between CpG island and the first exon.

We also looked at the distribution of CpG island nucleotides among various sequence classes such as intergenic regions, introns, exons, and first exons. We computed the likelihood score for each sequence class as the ratio of the observed fraction of CpG island nucleotides in that sequence class and the expected fraction of CpG island nucleotides in that sequence class. The result of applying method 1 on CSA were scores of 0.89 for intergenic region, 1.2 for intron, 5.86 for exon, and 13.2 for first exon. The same trend was also found for chromosome 22 and after the application of a higher threshold (method 2) on both data sets. In sum, genome-wide analysis has extended earlier analysis and suggests a strong correlation between CpG islands and first coding exons.

4.4 Genome-wide repetitive elements

The proportion of the genome covered by various classes of repetitive DNA is presented in Table 14. We observed about 35% of the genome in these repeat classes, very similar to values reported previously (*83*). Repetitive sequence may be underrepresented in the Celera assembly as a result of incomplete repeat resolution, as discussed above. About 8% of the scaffold length is in gaps, and we expect that much of this is repetitive sequence. Chromosome 19 has the highest repeat density (57%), as well as the highest gene density (Table 10). Of interest, among the different classes of repeat elements, we observe a clear association of Alu elements and gene density, which was not observed between LINEs and gene density.

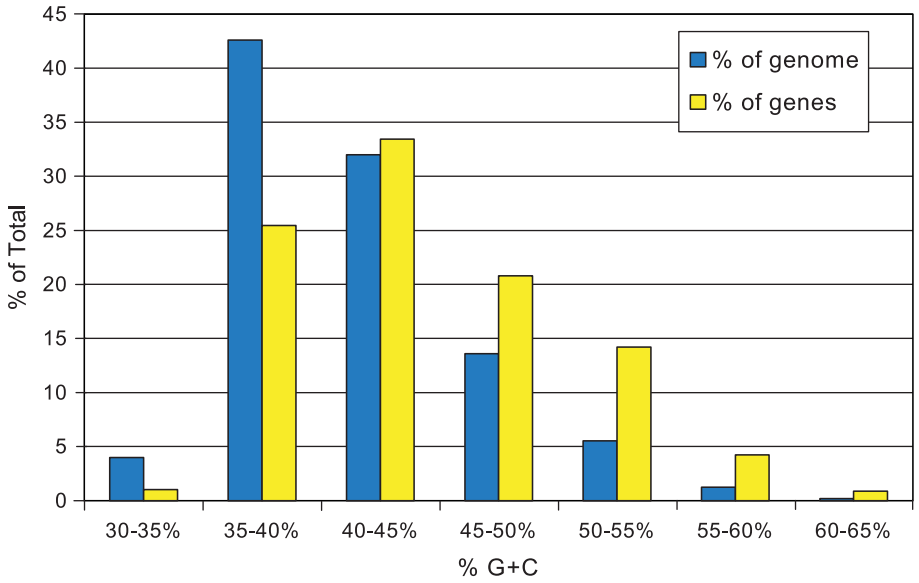


Fig. 10. Relation between G+C content and gene density. The blue bars show the percent of the genome (in 50-kbp windows) with the indicated G+C content. The percent of the total number of genes associated with each G+C bin is represented by the yellow bars. The graph shows that about 5% of the genome has a G+C content of between 50 and 55%, but that this portion contains nearly 15% of the genes.

5 Genome Evolution

Summary. The dynamic nature of genome evolution can be captured at several levels. These include gene duplications mediated by RNA intermediates (retrotransposition) and segmental genomic duplications. In this section, we document the genome-wide occurrence of retrotransposition events generating functional (intronless paralogs) or inactive genes (pseudogenes). Genes involved in translational processes and nuclear regulation account for nearly 50% of all intronless paralogs and processed pseudogenes detected in our survey. We have also cataloged the extent of segmental genomic duplication and provide evidence for 1077 duplicated blocks covering 3522 distinct genes.

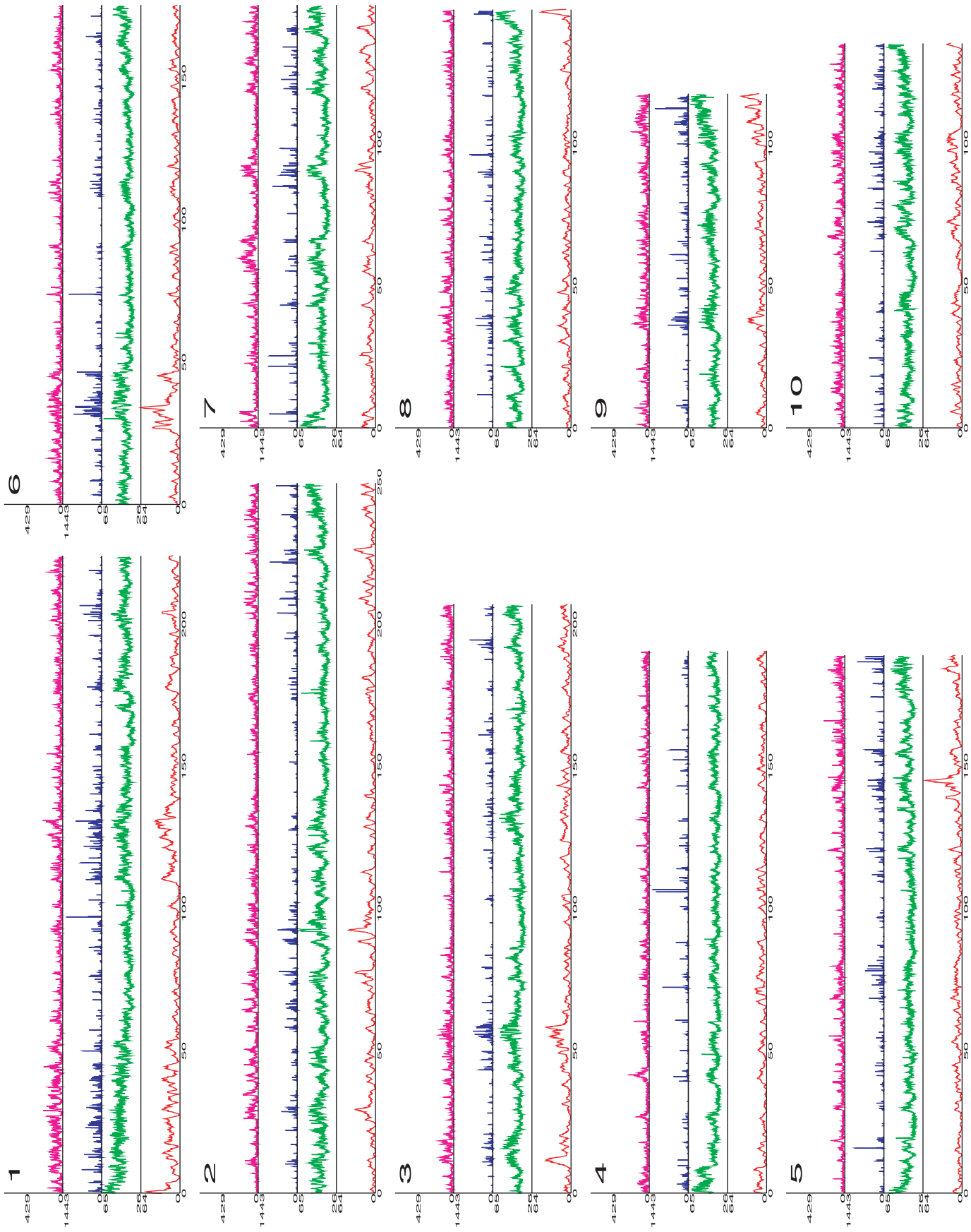


Fig. 11. Genome structural features.

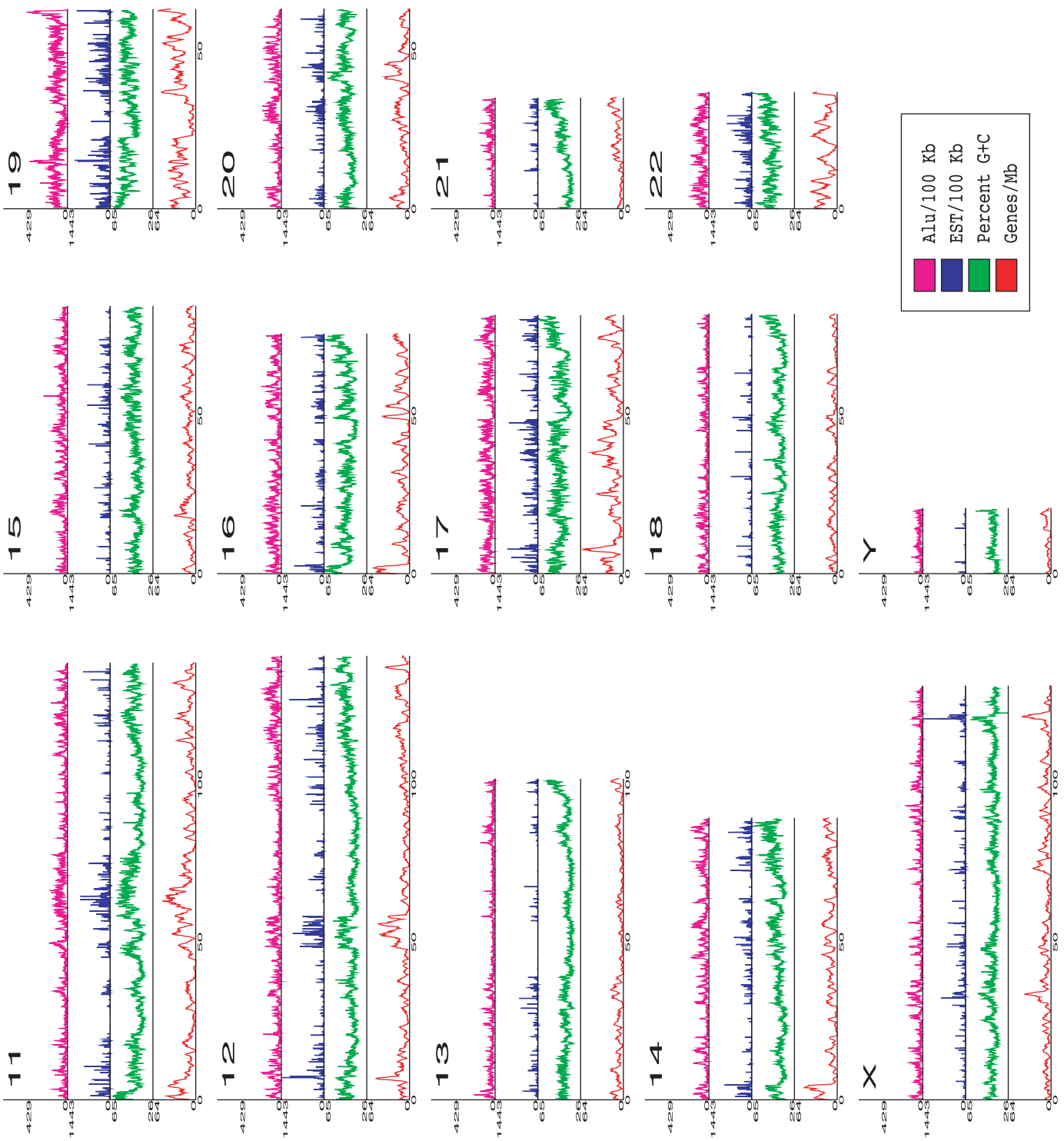


Fig. 11 (continued). Relation among gene density (orange), G+C content (green), EST density (blue), and Alu density (pink) along the lengths of each of the chromosomes. Gene density was calculated in 1-Mbp windows. The percent of G+C nucleotides was calculated in 100-kbp windows. The number of ESTs and Alu elements is shown per 100-kbp window.

5.1 Retrotransposition in the human genome

Retrotransposition of processed mRNA transcripts into the genome results in functional genes, called intronless paralogs, or inactivated genes (pseudogenes). A paralog refers to a gene that appears in more than one copy in a given organism as a result of a duplication event. The existence of both intron-containing and intronless forms of genes encoding functionally similar or identical proteins has been previously described (*84*, *85*). Cataloging these evolutionary events on the genomic landscape is of value in understanding the functional consequences of such gene-duplication events in cellular biology. Identification of conserved intronless paralogs in the mouse or other mammalian genomes should provide the basis for capturing the evolutionary chronology of these transposition events and provide insights into gene loss and accretion in the mammalian radiation.

A set of proteins corresponding to all 901

Table 10. Features of the chromosomes. De novo/any refers to the union of de novo predictions that do not overlap Otto predictions and have at least one other type of supporting evidence; de novo/2x refers to the union of de novo predictions that do not overlap Otto predictions and have at least two types of evidence. Deserts are regions of sequence with no annotated genes.

	Sequence coverage (CS assembly)						Base composition			Gene prediction*					Gene density (genes/Mbp)						
Chr.	Size (Mbp)	No. of scaffolds	Largest scaffold (Mbp)	No. of scaffolds >500 kbp	Sequence covered by scaffolds >500 kbp	% of total sequence in scaffolds >500 kbp	% repeat	% GC	No of CpG islands	Otto	De novo/ any	De novo/ 2×	Total (Otto + de novo/ any)	Total (Otto + de novo/ any)	Sequence in deserts >500/ kbp	Sequence in deserts >1 Mbp	Otto	De novo/ any	De novo/ 2×	Otto + de novo/ any	Otto + de novo/ 2×
1	220	2,549	11	82	192	88	37	42	2,335	1,743	1,710	710	3,453	2,453	29	6	8	8	3	16	11
2	240	3,263	13	78	217	91	36	40	1,703	1,183	1,771	633	2,954	1,816	55	19	5	7	2	12	7
3	200	3,532	7	78	173	87	37	40	1,271	1,013	1,414	598	2,427	1,611	50	12	5	7	3	12	8
4	186	2,180	10	70	169	91	37	38	1,081	696	1,165	449	1,861	1,145	55	18	4	6	2	10	6
5	182	3,231	11	63	163	89	37	40	1,302	892	1,244	474	2,136	1,366	46	15	5	7	2	11	7
6	172	1,713	13	58	160	93	37	40	1,384	943	1,314	524	2,257	1,467	38	9	6	7	3	13	8
7	146	1,326	14	53	130	89	38	40	1,406	759	1,072	460	1,831	1,219	26	12	5	7	3	12	8
8	146	1,772	11	54	135	92	36	40	948	583	977	357	1,560	940	33	6	4	7	2	11	6
9	113	1,616	8	40	101	89	38	41	1,315	689	848	329	1,537	1,018	22	9	6	7	3	13	8
10	130	2,005	9	55	116	89	36	42	1,087	685	968	342	1,653	1,027	21	8	5	7	2	12	7
11	132	2,814	9	44	116	88	39	42	1,461	1,051	1,134	535	2,185	1,586	27	9	8	8	4	16	12
12	134	2,614	8	51	117	87	38	41	1,131	925	936	417	1,861	1,342	24	9	7	7	3	14	10
13	99	1,038	13	34	91	91	36	38	644	341	691	241	1,032	582	31	16	4	7	2	10	5
14	87	576	11	16	83	95	40	41	913	583	700	290	1,283	873	34	20	7	8	3	14	10
15	80	1,747	8	31	70	87	37	42	722	558	640	246	1,198	804	8	1	7	8	3	15	10
16	75	1,520	8	27	62	82	40	44	1,533	748	673	247	1,421	995	13	3	10	9	3	19	12
17	78	1,683	6	40	61	78	39	45	1,489	897	648	313	1,545	1,210	15	6	12	8	4	19	15
18	79	1,333	13	18	72	92	36	40	510	283	543	189	826	472	21	10	4	7	2	10	6
19	58	2,282	3	31	38	67	57	49	2,804	1,141	534	268	1,675	1,409	3	0	20	9	4	29	23
20	61	580	14	17	58	94	41	44	997	517	469	180	986	697	7	1	8	7	3	16	11
21	33	358	10	6	32	96	38	41	519	184	265	102	449	286	15	9	6	8	3	13	8
22	36	333	11	12	32	88	44	48	1,173	494	341	147	835	641	3	0	14	9	4	23	17
X	128	1,346	4	91	93	73	46	39	726	605	860	387	1,465	992	29	8	5	6	3	11	7
Y	19	638	2	10	12	65	50	39	65	55	155	49	210	104	4	2	3	8	2	11	5
U*	75	11,542	1						479	196	278	132	474	328							
Total	2907	53,591		1,059	2,490				28,519	17,764	21,350	8,619	39,114	26,383	606	208					
Avg.	116	2,144	9	44	104	87	40	41	1,160	714	812	333	1,526	1,047	25	9	7	7	3	14	9

*Chromosomal assignment unknown.

Otto-predicted, single-exon genes were subjected to BLAST analysis against the proteins encoded by the remaining multiexon predicted transcripts. Using homology criteria of 70% sequence identity over 90% of the length, we identified 298 instances of single- to multi-exon correspondence. Of these 298 sequences, 97 were represented in the GenBank data set of experimentally validated full-length genes at the stringency specified and were verified by manual inspection.

We believe that these 97 cases may represent intronless paralogs (see Web table 1 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1) of known genes. Most of these are flanked by direct repeat sequences, although the precise nature of these repeats remains to be determined. All of the cases for which we have high confidence contain polyadenylated [poly(A)] tails characteristic of retrotransposition.

Recent publications describing the phenomenon of functional intronless paralogs speculate that retrotransposition may serve as a mechanism used to escape X-chromosomal inactivation (*84*, *86*). We do not find a bias toward X chromosome origination of these retrotransposed genes; rather, the results show a random chromosome distribution of both the intron-containing and corresponding intronless paralogs. We also have found several cases of retrotransposition from a single source chromosome to multiple target chromosomes. Interesting examples include the retrotransposition of a five exon–containing ribosomal protein L21 gene on chromosome 13 onto chromosomes 1, 3, 4, 7, 10, and 14, respectively. The size of the source genes can also show variability. The largest example is the 31-exon diacylglycerol kinase zeta gene on chromosome 11 that has an intronless paralog on chromosome 13. Regardless of route, retrotransposition with subsequent gene changes in coding or noncoding regions that lead to different functions or expression patterns, represents a key route to providing an enhanced functional repertoire in mammals (*87*).

Our preliminary set of retrotransposed intronless paralogs contains a clear overrepresentation of genes involved in translational processes (40% ribosomal proteins and 10% translation elongation factors) and nuclear regulation (HMG nonhistone proteins, 4%), as well as metabolic and regulatory enzymes. EST matches specific to a subset of intronless paralogs suggest expression of these intronless paralogs. Differences in the upstream regulatory sequences between the source genes and their intronless paralogs could account for differences in tissue-specific gene expression. Defining which, if any, of these processed genes are functionally expressed and translated will require further elucidation and experimental validation.

5.2 Pseudogenes

A pseudogene is a nonfunctional copy that is very similar to a normal gene but that has been altered slightly so that it is not expressed. We developed a method for the preliminary analysis of processed pseudogenes in the human genome as a starting point in elucidating the ongoing evolutionary forces

Table 11. Genome overview.

Size of the genome (including gaps)	2.91 Gbp
Size of the genome (excluding gaps)	2.66 Gbp
Longest contig	1.99 Mbp
Longest scaffold	14.4 Mbp
Percent of A+T in the genome	54
Percent of G+C in the genome	38
Percent of undetermined bases in the genome	9
Most GC-rich 50 kb	Chr. 2 (66%)
Least GC-rich 50 kb	Chr. X (25%)
Percent of genome classified as repeats	35
Number of annotated genes	26,383
Percent of annotated genes with unknown function	42
Number of genes (hypothetical and annotated)	39,114
Percent of hypothetical and annotated genes with unknown function	59
Gene with the most exons	Titin (234 exons)
Average gene size	27 kbp
Most gene-rich chromosome	Chr. 19 (23 genes/Mb)
Least gene-rich chromosomes	Chr. 13 (5 genes/Mb), Chr. Y (5 genes/Mb)
Total size of gene deserts (>500 kb with no annotated genes)	605 Mbp
Percent of base pairs spanned by genes	25.5 to 37.8*
Percent of base pairs spanned by exons	1.1 to 1.4*
Percent of base pairs spanned by introns	24.4 to 36.4*
Percent of base pairs in intergenic DNA	74.5 to 63.6*
Chromosome with highest proportion of DNA in annotated exons	Chr. 19 (9.33)
Chromosome with lowest proportion of DNA in annotated exons	Chr. Y (0.36)
Longest intergenic region (between annotated + hypothetical genes)	Chr. 13 (3,038,416 bp)
Rate of SNP variation	1/1250 bp

*In these ranges, the percentages correspond to the annotated gene set (26, 383 genes) and the hypothetical + annotated gene set (39,114 genes), respectively.

Table 12. Rate of recombination per physical distance (cM/Mb) across the genome. Genethon markers were placed on CSA-mapped assemblies, and then relative physical distances and rates were calculated in 3-Mb windows for each chromosome. NA, not applicable.

Chrom.	Male			Sex-average			Female		
	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.
1	2.60	1.12	0.23	2.81	1.42	0.52	3.39	1.76	0.68
2	2.23	0.78	0.33	2.65	1.12	0.54	3.17	1.40	0.61
3	2.55	0.86	0.23	2.40	1.07	0.42	2.71	1.30	0.33
4	1.66	0.67	0.15	2.06	1.04	0.60	2.50	1.40	0.77
5	2.00	0.67	0.18	1.87	1.08	0.42	2.26	1.43	0.62
6	1.97	0.71	0.28	2.57	1.12	0.37	3.47	1.67	0.64
7	2.34	1.16	0.48	1.67	1.17	0.47	2.27	1.21	0.34
8	1.83	0.73	0.14	2.40	1.05	0.46	3.44	1.36	0.43
9	2.01	0.99	0.53	1.95	1.32	0.77	2.63	1.66	0.82
10	3.73	1.03	0.22	3.05	1.29	0.66	2.84	1.51	0.76
11	1.43	0.72	0.31	2.13	0.99	0.47	3.10	1.32	0.49
12	4.12	0.76	0.26	3.35	1.16	0.49	2.93	1.55	0.59
13	1.60	0.75	0.01	1.87	0.95	0.17	2.49	1.19	0.32
14	3.15	0.98	0.18	2.65	1.30	0.62	3.14	1.63	0.75
15	2.28	0.94	0.34	2.31	1.22	0.42	2.53	1.56	0.54
16	1.83	1.00	0.47	2.70	1.55	0.63	4.99	2.32	1.12
17	3.87	0.87	0.00	3.54	1.35	0.54	4.19	1.83	0.94
18	3.12	1.37	0.86	3.75	1.66	0.43	4.35	2.24	0.72
19	3.02	0.97	0.10	2.57	1.41	0.49	2.89	1.75	0.87
20	3.64	0.89	0.00	2.79	1.50	0.83	3.31	2.15	1.34
21	3.23	1.26	0.69	2.37	1.62	1.08	2.58	1.90	1.18
22	1.25	1.10	0.84	1.88	1.41	1.08	3.73	2.08	0.93
X	NA	NA	NA	NA	NA	NA	3.12	1.64	0.72
Y	NA	NA	NA	NA	NA	NA	NA	NA	NA
Genome	4.12	0.88	0.00	3.75	1.22	0.17	4.99	1.55	0.32

that account for gene inactivation. The general structural characteristics of these processed pseudogenes include the complete lack of intervening sequences found in the functional counterparts, a poly(A) tract at the 3′ end, and direct repeats flanking the pseudogene sequence. Processed pseudogenes occur as a result of retrotransposition, whereas unprocessed pseudogenes arise from segmental genome duplication.

We searched the complete set of Otto-predicted transcripts against the genomic sequence by means of BLAST. Genomic regions corresponding to all Otto-predicted transcripts were excluded from this analysis. We identified 2909 regions matching with greater than 70% identity over at least 70% of the length of the transcripts that likely represent processed pseudogenes. This number is probably an underestimate because specific methods to search for pseudogenes were not used.

We looked for correlations between structural elements and the propensity for retrotransposition in the human genome. GC content and transcript length were compared between the genes with processed pseudogenes (1177 source genes) versus the remainder of the predicted gene set. Transcripts that give rise to processed pseudogenes have shorter average transcript length (1027 bp versus 1594 bp for the Otto set) as compared with genes for which no pseudogene was detected. The overall GC content did not show any significant difference, contrary to a recent report (*88*). There is a clear trend in gene families that are present as processed pseudogenes. These include ribosomal proteins (67%), lamin receptors (10%), translation elongation factor alpha (5%), and HMG–non-histone proteins (2%). The increased occurrence of retrotransposition (both intronless paralogs and processed pseudogenes) among genes involved in translation and nuclear regulation may reflect an increased transcriptional activity of these genes.

5.3 Gene duplication in the human genome

Building on a previously published procedure (*27*), we developed a graph-theoretic algorithm, called Lek, for grouping the predicted human protein set into protein families (*89*). The complete clusters that result from the Lek clustering provide one basis for comparing the role of whole-genome or chromosomal duplication in protein family expansion as opposed to other means, such as tandem duplication. Because each complete cluster represents a closed and certain island of homology, and because Lek is capable of simultaneously clustering protein complements of several organisms, the number of proteins contributed by each organism to a complete cluster can be predicted with confidence depending on the quality of the annotation of each genome. The variance of each organism's contribution to each cluster can then be calculated, allowing an assessment of the relative importance of large-scale duplication versus smaller-scale, organism-specific expansion and contraction of protein families, presumably as a result of natural selection operating on individual protein families within an organism. As can be seen in Fig. 12, the large variance in the relative numbers of human as compared with *D. melanogaster* and *Caenorhabditis elegans* proteins in complete clusters may be explained by multiple events of relative expansions in gene families in each of the three animal genomes. Such expansions would give rise to the distribution that shows a peak at 1:1 in the ratio for human-worm or human-fly clusters with the slope spread covering both human and fly/worm predominance, as we observed (Fig. 12). Furthermore, there are nearly as many clusters where worm and fly proteins predominate despite the larger numbers of proteins in the human. At face value, this analysis suggests that natural selection acting on individual protein families has been a major force driving the expansion of at least some elements of the human protein set. However, in our analysis, the difference between an ancient whole-genome duplication followed by loss, versus piecemeal duplication, cannot be easily distinguished. In order to differentiate these scenarios, more extended analyses were performed.

Table 13. Characteristics of CpG islands identified in chromosome 22 (34-Mbp sequence length) and the whole genome (2.9-Gbp sequence length) by means of two different methods. Method 1 uses a CG likelihood ratio of ≥0.6. Method 2 uses a CG likelihood ratio of ≥0.8.

	Chromosome 22		Whole genome (CS assembly)	
	Method 1	Method 2	Method 1	Method 2
Number of CpG islands detected	5,211	522	195,706	26,876
Average length of island (bp)	390	535	395	497
Percent of sequence predicted as CpG	5.9	0.8	2.6	0.4
Percent of first exons that overlap a CpG island	44	25	42	22
Percent of first exons with first position of exon contained inside a CpG island	37	22	40	21
Average distance between first exon and closest CpG island (bp)	1,013	10,486	2,182	17,021
Expected distance between first exon and closest CpG island (bp)	3,262	32,567	7,164	55,811

Table 14. Distribution of repetitive DNA in the compartmentalized shotgun assembly sequence.

Repetitive elements	Megabases in assembled sequences	Percent of assembly	Previously predicted (%) (*83*)
Alu	288	9.9	10.0
Mammalian interspersed repeat (MIR)	66	2.3	1.7
Medium reiteration (MER)	50	1.7	1.6
Long terminal repeat (LTR)	155	5.3	5.6
Long interspersed nucleotide element (LINE)	466	16.1	16.7
Total	1025	35.3	35.6

5.4 Large-scale duplications

Using two independent methods, we searched for large-scale duplications in the human genome. First, we describe a protein family–based method that identified highly conserved blocks of duplication. We then describe our comprehensive method for identifying all interchromosomal block duplications. The latter method identified a large number of duplicated chromosomal segments covering parts of all 24 chromosomes.

The first of the methods is based on the idea of searching for blocks of highly conserved homologous proteins that occur in more than one location on the genome. For this comparison, two genes were considered equivalent if their protein products were de-

termined to be in the same family and the same complete Lek cluster (essentially paralogous genes) (*89*). Initially, each chromosome was represented as a string of genes ordered by the start codons for predicted genes along the chromosome. We considered the two strands as a single string, because local inversions are relatively common events relative to large-scale duplications. Each gene was indexed according to the protein family and Lek complete cluster (*89*). All pairs of indexed gene strings were then aligned in both the forward and reverse directions with the Smith-Waterman algorithm (*90*). A match between two proteins of the same Lek complete cluster was given a score of 10 and a mismatch −10, with gap open and extend penalties of −4 and −1. With these parameters, 19 conserved interchromosomal blocks of duplication were observed, all of which were also detected and expanded by the comprehensive method described below. The detection of only a relatively small number of block duplications was a consequence of using an intrinsically conservative method grounded in the conservative constraints of the complete Lek clusters.

In the second, more comprehensive approach, we aligned all chromosomes directly with one another using an algorithm based on the MUMmer system (*91*). This alignment method uses a suffix tree data structure and a linear-time algorithm to align long sequences very rapidly; for example, two chromosomes of 100 Mbp can be aligned in less than 20 min (on a Compaq Alpha computer) with 4 gigabytes of memory. This procedure was used recently to identify numerous large-scale segmental duplications among the five chromosomes of *A. thaliana* (*92*); in that organism, the method revealed that 60% of the genome (66 Mbp) is covered by 24 very large duplicated segments. For *Arabidopsis*, a DNA-based alignment was sufficient to reveal the segmental duplications between chromosomes; in the human genome, DNA alignments at the whole-chromosome level are insufficiently sensitive. Therefore, a modified procedure was developed and applied, as follows. First, all 26,588 proteins (9,675,713 million amino acids) were concatenated end-to-end in order as they occur along each of the 24 chromosomes, irrespective of strand location. The concatenated protein set was then aligned against each chromosome by the MUMmer algorithm. The resulting matches were clustered to extract all sets of three or more protein matches that occur in close proximity on two different chromosomes (*93*); these represent the candidate segmental duplications. A series of filters were developed and applied to remove likely false-positives from this set; for example, small blocks that were spread across many proteins were removed. To refine the filtering methods, a shuffled protein set was first created by taking the 26,588 proteins, randomizing their order, and then partitioning them into 24 shuffled chromosomes, each containing the same number of proteins as the true genome. This shuffled protein set has the identical composition to the real genome; in particular, every protein and every domain appears the same number of times. The complete algorithm was then applied to both the real and the shuffled data, with the results on the shuffled data being used to estimate the false-positive rate. The algorithm after filtering yielded 10,310 gene pairs in 1077 duplicated blocks containing 3522 distinct genes; tandemly duplicated expansions in many of the blocks explain the excess of gene pairs to distinct genes. In the shuffled data, by contrast, only 370 gene pairs were found, giving a false-positive estimate of 3.6%. The most likely explanation for the 1077 block duplications is ancient segmental duplications. In many cases, the order of the proteins has been shuffled, although proximity is preserved. Out of the 1077 blocks, 159 contain only three genes, 137 contain four genes, and 781 contain five or more genes.

To illustrate the extent of the detected duplications, Fig. 13 shows all 1077 block duplications indexed to each chromosome in 24 panels in which only duplications mapped to the indexed chromosome are displayed. The figure makes it clear that the duplications are ubiquitous in the genome. One feature that it displays is many relatively small chromosomal stretches, with one-to-many duplication relationships that are graphically striking. One such example captured by the analysis is the well-documented olfactory receptor (OR) family, which is scattered in blocks throughout the genome and which has been analyzed for genome-deployment reconstructions at several evolutionary stages (*94*). The figure also illustrates that some chromosomes, such as chromosome 2, contain many more detected large-scale duplications than others. Indeed, one of the largest duplicated segments is a large block of 33 proteins on chromosome 2, spread among eight smaller blocks in 2p, that aligns to a paralogous set on chromosome 14, with one rearrangement (see chromosomes 2 and 14 panels in Fig. 13). The proteins are not contiguous but span a region containing 97 proteins on chromosome 2 and 332 proteins on chromosome 14. The likelihood of observing this many duplicated proteins by chance, even over a span of this length, is 2.3×10^{-68} (*93*). This duplicated set spans 20 Mbp on chromosome 2 and 63 Mbp on chromosome 14, over 70% of the latter chromosome. Chromosome 2 also contains a block duplication that is nearly as large, which is shared by chromosome arm 2q and chromosome 12. This duplication incorporates two of the four known Hox gene clusters, but considerably expands the extent of the duplications proximally and distally on the pair of chromosome arms. This breadth of duplication is also seen on the two chromosomes carrying the other two Hox clusters.

An additional large duplication, between chromosomes 18 and 20, serves as a good example to illustrate some of the features common to many of the other observed large duplications (Fig. 13, inset). This duplication contains 64 detected ordered intrachromosomal pairs of homologous genes. After discounting a 40-Mb stretch of chromosome 18 free of matches to chromosome 20, which is likely to represent a large insert (between the gene assignments "Krup rel" and "collagen rel" on chromosome 18 in Fig. 13), the full duplication segment covers 36 Mb on chromosome 18 and 28 Mb on chromosome 20.

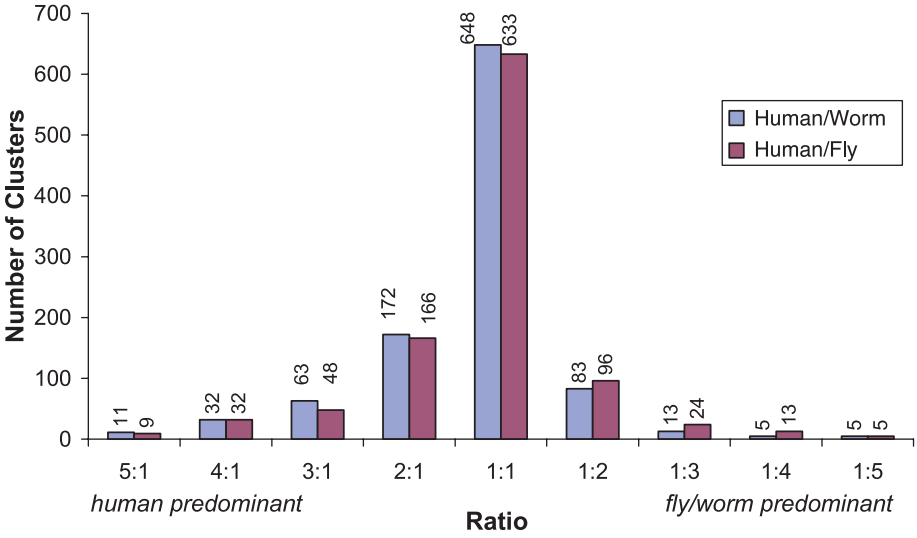


Fig. 12. Gene duplication in complete protein clusters. The predicted protein sets of human, worm, and fly were subjected to Lek clustering (*27*). The numbers of clusters with varying ratios (whole number) of human versus worm and human versus fly proteins per cluster were plotted.

By this measure, the duplication segment spans nearly half of each chromosome's net length. The most likely scenario is that the whole span of this region was duplicated as a single very large block, followed by shuffling owing to smaller scale rearrangements. As such, at least four subsequent rearrangements would need to be invoked to explain the relative insertions and inversions seen in the duplicated segment interval. The 64 protein pairs in this alignment occur among 217 protein assignments on chromosome 18, and among 322 protein assignments on chromosome 20, for a density of involved proteins of 20 to 30%. This is consistent with an ancient large-scale duplication followed by subsequent gene loss on one or both chromosomes. Loss of just one member of a gene pair subsequent to the duplication would result in a failure to score a gene pair in the block; less than 50% gene loss on the chromosomes would lead to the duplication density observed here. As an independent verification of the significance of the alignments detected, it can be seen that a substantial number of the pairs of aligning proteins in this duplication, including some of those annotated (Fig. 13), are those populating small Lek complete clusters (see above). This indicates that they are members of very small families of paralogs; their relative scarcity within the genome validates the uniqueness and robust nature of their alignments.

Two additional qualitative features were observed among many of the large-scale duplications. First, several proteins with disease associations, with OMIM (Online Mendelian Inheritance in Man) assignments, are members of duplicated segments (see web table 2 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1). We have also observed a few instances where paralogs on both duplicated segments are associated with similar disease conditions. Notable among these genes are proteins involved in hemostasis (coagulation factors) that are associated with bleeding disorders, transcriptional regulators like the homeobox proteins associated with developmental disorders, and potassium channels associated with cardiovascular conduction abnormalities. For each of these disease genes, closer study of the paralogous genes in the duplicated segment may reveal new insights into disease causation, with further investigation needed to determine whether they might be involved in the same or similar genetic diseases. Second, although there is a conserved number of proteins and coding exons predicted for specific large duplicated spans within the chromosome 18 to 20 alignment, the genomic DNA of chromosome 18 in these specific spans is in some cases more than 10-fold longer than the corresponding chromosome 20 DNA. This selective accretion of noncoding DNA (or conversely, loss of noncoding DNA) on one of a pair of duplicated chromosome regions was observed in many compared regions. Hypotheses to explain which mechanisms foster these processes must be tested.

Evaluation of the alignment results gives some perspective on dating of the duplications. As noted above, large-scale ancient segmental duplication in fact best explains many of the blocks detected by this genome-wide analysis. The regions of human chromosomes involved in the large-scale duplications expanded upon above (chromosomes 2 to 14, 2 to 12, and 18 to 20) are each syntenic to a distinct mouse chromosomal region. The corresponding mouse chromosomal regions are much more similar in sequence conservation, and even in order, to their human synteny partners than the human duplication regions are to each other. Further, the corresponding mouse chromosomal regions each bear a significant proportion of genes orthologous to the human genes on which the human duplication assignments were made. On the basis of these factors, the corresponding mouse chromosomal spans, at coarse resolution, appear to be products of the same large-scale duplications observed in humans. Although further detailed analysis must be carried out once a more complete genome is assembled for mouse, the underlying large duplications appear to predate the two species' divergence. This dates the duplications, at the latest, before divergence of the primate and rodent lineages. This date can be further refined upon examination of the synteny between human chromosomes and those of chicken, pufferfish (*Fugu rubripes*), or zebrafish (*95*). The only substantial syntenic stretches mapped in these species corresponding to both pairs of human duplications are restricted to the Hox cluster regions. When the synteny of these regions (or others) to human chromosomes is extended with further mapping, the ages of the nearly chromosome-length duplications seen in humans are likely to be dated to the root of vertebrate divergence.

The MUMmer-based results demonstrate large block duplications that range in size from a few genes to segments covering most of a chromosome. The extent of segmental duplications raises the question of whether an ancient whole-genome duplication event is the underlying explanation for the numerous duplicated regions (*96*). The duplications have undergone many deletions and subsequent rearrangements; these events make it difficult to distinguish between a whole-genome duplication and multiple smaller events. Further analysis, focused especially on comparing the estimated ages of all the block duplications, derived partially from interspecies genome comparisons, will be necessary to determine which of these two hypotheses is more likely. Comparisons of genomes of different vertebrates, and even cross-phyla genome comparisons, will allow for the deconvolution of duplications to eventually reveal the stagewise history of our genome, and with it a history of the emergence of many of the key functions that distinguish us from other living things.

6 A Genome-Wide Examination of Sequence Variations

Summary. Computational methods were used to identify single-nucleotide polymorphisms (SNPs) by comparison of the Celera sequence to other SNP resources. The SNP rate between two chromosomes was ~1 per 1200 to 1500 bp. SNPs are distributed nonrandomly throughout the genome. Only a very small proportion of all SNPs (<1%) potentially impact protein function based on the functional analysis of SNPs that affect the predicted coding regions. This results in an estimate that only thousands, not millions, of genetic variations may contribute to the structural diversity of human proteins.

Having a complete genome sequence enables researchers to achieve a dramatic acceleration in the rate of gene discovery, but only through analysis of sequence variation in DNA can we discover the genetic basis for variation in health among human beings. Whole-genome shotgun sequencing is a particularly effective method for detecting sequence variation in tandem with whole-genome assembly. In addition, we compared the distribution and attributes of SNPs ascertained by three other methods: (i) alignment of the Celera consensus sequence to the PFP assembly, (ii) overlap of high-quality reads of genomic sequence (referred to as "Kwok"; 1,120,195 SNPs) (*97*), and (iii) reduced representation shotgun sequencing (referred to as "TSC"; 632,640 SNPs) (*98*). These data were consistent in showing an overall nucleotide diversity of $\sim 8 \times 10^{-4}$, marked heterogeneity across the genome in SNP density, and an overwhelming preponderance of noncoding variation that produces no change in expressed proteins.

6.1 SNPs found by aligning the Celera consensus to the PFP assembly

Ideally, methods of SNP discovery make full use of sequence depth and quality at every site, and quantitatively control the rate of false-positive and false-negative calls with an explicit sampling model (*99*). Comparison of consensus sequences in the absence of these details necessitated a more ad hoc approach (quality scores could not readily be obtained for the PFP assembly). First, all sequence differences between the two consensus sequences were identified; these were then filtered to reduce the contribution of sequencing errors and misassembly. As a measure of the effectiveness of the filtering step, we monitored the ratio of transition and transversion substitutions, because a 2:1 ratio has been well documented as typical in mammalian evolution (*100*) and in human SNPs

(*101*, *102*). The filtering steps consisted of removing variants where the quality score in the Celera consensus was less than 30 and where the density of variants was greater than 5 in 400 bp. These filters resulted in shifting the transition-to-transversion ratio from 1.57:1 to 1.89:1. When applied to 2.3 Gbp of alignments between the Celera and PFP consensus sequences, these filters resulted in identification of 2,104,820 putative SNPs from a total of 2,778,474 substitution differences. Overlaps between this set of SNPs and those found by other methods are described below.

6.2 Comparisons to public SNP databases

Additional SNPs, including 2,536,021 from dbSNP (www.ncbi.nlm.nih.gov/SNP) and 13,150 from HGMD (Human Gene Mutation Database, from the University of Wales, UK), were mapped on the Celera consensus sequence by a sequence similarity search with the program PowerBlast (*103*). The two largest data sets in dbSNP are the Kwok and TSC sets, with 47% and 25% of the dbSNP records. Low-quality alignments with partial coverage of the dbSNP sequence and alignments that had less than 98% sequence identity between the Celera sequence and the dbSNP flanking sequence were eliminated. dbSNP sequences mapping to multiple locations on the Celera genome were discarded. A total of 2,336,935 dbSNP variants were mapped to 1,223,038 unique locations on the Celera sequence, implying considerable redundancy in dbSNP. SNPs in the TSC set mapped to 585,811 unique genomic locations, and SNPs in the Kwok set mapped to 438,032 unique locations. The combined unique SNPs counts used in this analysis, including Celera-PFP, TSC, and Kwok, is 2,737,668. Table 15 shows that a substantial fraction of SNPs identified by one of these methods was also found by another method. The very high overlap (36.2%) between the Kwok and Celera-PFP SNPs may be due in part to the use by Kwok of sequences that went into the PFP assembly. The unusually low overlap (16.4%) between the Kwok and TSC sets is due to their being the smallest two sets. In addition, 24.5% of the Celera-PFP SNPs overlap with SNPs derived from the Celera genome sequences (*46*). SNP validation in population samples is an expensive and laborious process, so confirmation on multiple data sets may provide an efficient initial validation "in silico" (by computational analysis).

One means of assessing whether the three sets of SNPs provide the same picture of human variation is to tally the frequencies of the six possible base changes in each set of SNPs (Table 16). Previous measures of nucleotide diversity were mostly derived from small-scale analysis on candidate genes (*101*), and our analysis with all three data sets validates the previous observations at the whole-genome scale. There is remarkable homogeneity between the SNPs found in the Kwok set, the TSC set, and in our whole-genome shotgun (*46*) in this substitution pattern. Compared with the rest of the data sets, Celera-PFP deviates slightly from the 2:1 transition-to-transversion ratio observed in the other SNP sets. This result is not unexpected, because some fraction of the computationally identified SNPs in the Celera-PFP comparison may in fact be sequence errors. A 2:1 transition:transversion ratio for the bona fide SNPs would be obtained if one assumed that 15% of the sequence differences in the Celera-PFP set were a result of (presumably random) sequence errors.

6.3 Estimation of nucleotide diversity from ascertained SNPs

The number of SNPs identified varied widely across chromosomes. In order to normalize these values to the chromosome size and sequence coverage, we used π, the standard statistic for nucleotide diversity (*104*). Nucleotide diversity is a measure of per-site heterozygosity, quantifying the probability that a pair of chromosomes drawn from the population will differ at a nucleotide site. In order to calculate nucleotide diversity for each chromosome, we need to know the number of nucleotide sites that were surveyed for variation, and in methods like reduced respresentation sequencing, we need to know the sequence quality and the depth of coverage at each site. These data are not readily available, so we could not estimate nucleotide diversity from the TSC effort. Estimation of nucleotide diversity from high-quality sequence overlaps should be possible, but again, more information is needed on the details of all the alignments.

Estimation of nucleotide diversity from a shotgun assembly entails calculating for each column of the multialignment, the probability that two or more distinct alleles are present, and the probability of detecting a SNP if in fact the alleles have different sequence (i.e., the probability of correct sequence calls). The greater the depth of coverage and the higher the sequence quality, the higher is the chance of successfully detecting a SNP (*105*). Even after correcting for variation in coverage, the nucleotide diversity appeared to vary across autosomes. The significance of this heterogeneity was tested by analysis of variance, with estimates of π for 100-kbp windows to estimate variability within chromosomes (for the Celera-PFP comparison, $F = 29.73$, $P < 0.0001$).

Average diversity for the autosomes estimated from the Celera-PFP comparison was 8.94×10^{-4}. Nucleotide diversity on the X chromosome was 6.54×10^{-4}. The X is expected to be less variable than autosomes, because for every four copies of autosomes in the population, there are only three X chromosomes, and this smaller effective population size means that random drift will more rapidly remove variation from the X (*106*).

Having ascertained nucleotide variation genome-wide, it appears that previous estimates of nucleotide diversity in humans based on samples of genes were reasonably accurate (*101*, *102*, *106*, *107*). Genome-wide, our estimate of nucleotide diversity was 8.98×10^{-4} for the Celera-PFP alignment, and a published estimate averaged over 10 densely resequenced human genes was 8.00×10^{-4} (*108*).

6.4 Variation in nucleotide diversity across the human genome

Such an apparently high degree of variability among chromosomes in SNP density raises the question of whether there is heterogeneity at a finer scale within chromo-

Table 15. Overlap of SNPs from genome-wide SNP databases. Table entries are SNP counts for each pair of data sets. Numbers in parentheses are the fraction of overlap, calculated as the count of overlapping SNPs divided by the number of SNPs in the smaller of the two databases compared. Total SNP counts for the databases are: Celera-PFP, 2,104,820; TSC, 585,811; and Kwok 438,032. Only unique SNPs in the TSC and Kwok data sets were included.

	TSC	Kwok
Celera-PFP	188,694 (0.322)	158,532 (0.362)
TSC		72,024 (0.164)

Table 16. Summary of nucleotide changes in different SNP data sets.

SNP data set	A/G (%)	C/T (%)	A/C (%)	A/T (%)	C/G (%)	T/G (%)	Transition: transversion
Celera-PFP	30.7	30.7	10.3	8.6	9.2	10.3	1.59:1
Kwok*	33.7	33.8	8.5	7.0	8.6	8.4	2.07:1
TSC†	33.3	33.4	8.8	7.3	8.6	8.6	1.99:1

*November 2000 release of the NCBI database dbSNP (www.nci.nlm.nih.gov/SNP/) with the method defined as Overlap SnpDetectionWithPolyBayes. The submitter of the data is Pui-Yan Kwok from Washington University. †November 2000 release of NCBI dbSNP (www.ncbi.nlm.nih.gov/SNP/) with the methods defined as TSC-Sanger, TSC-WICGR, and TSC-WUGSC. The submitter of the data is Lincoln Stein from Cold Spring Harbor Laboratory.

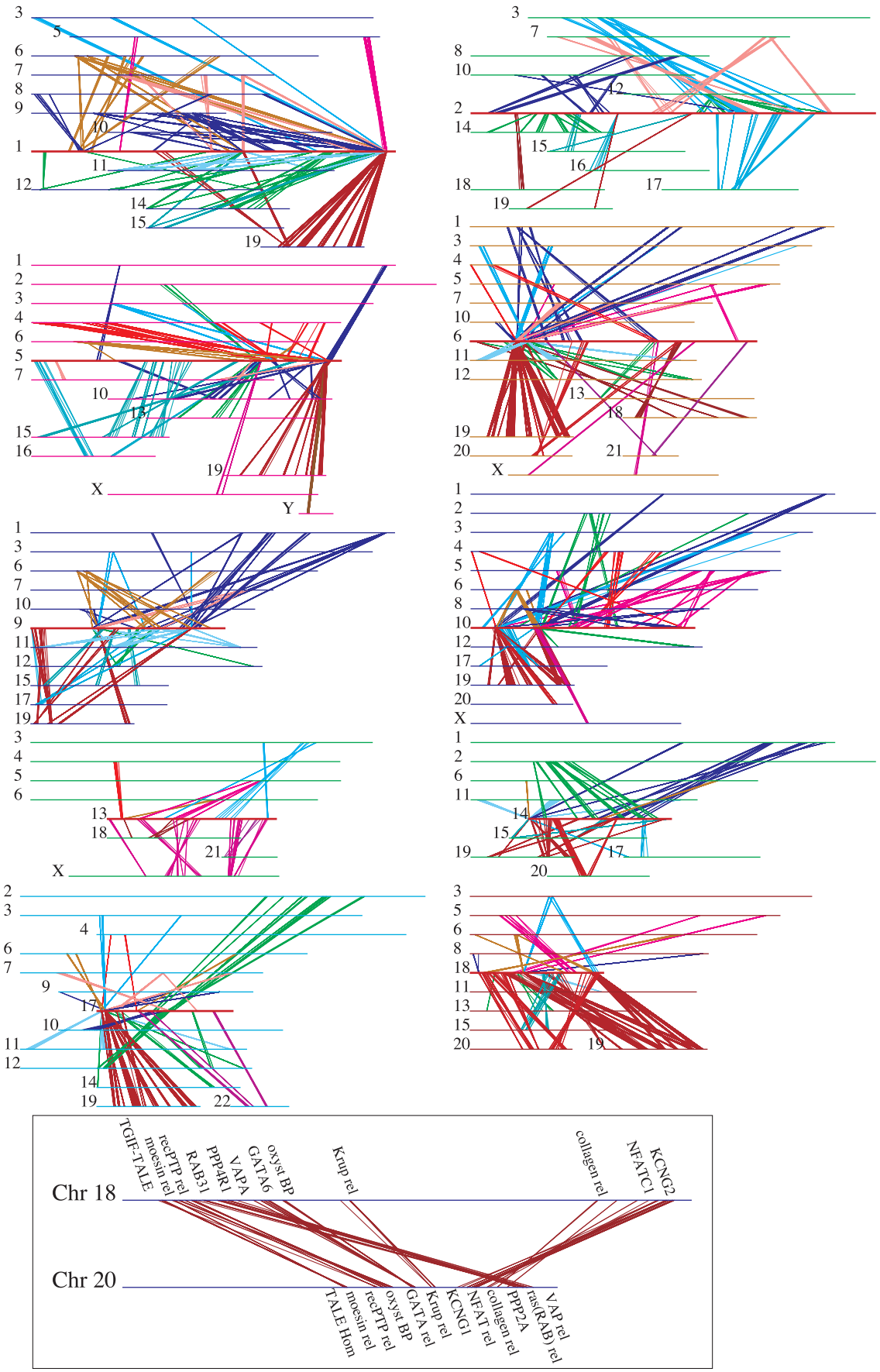
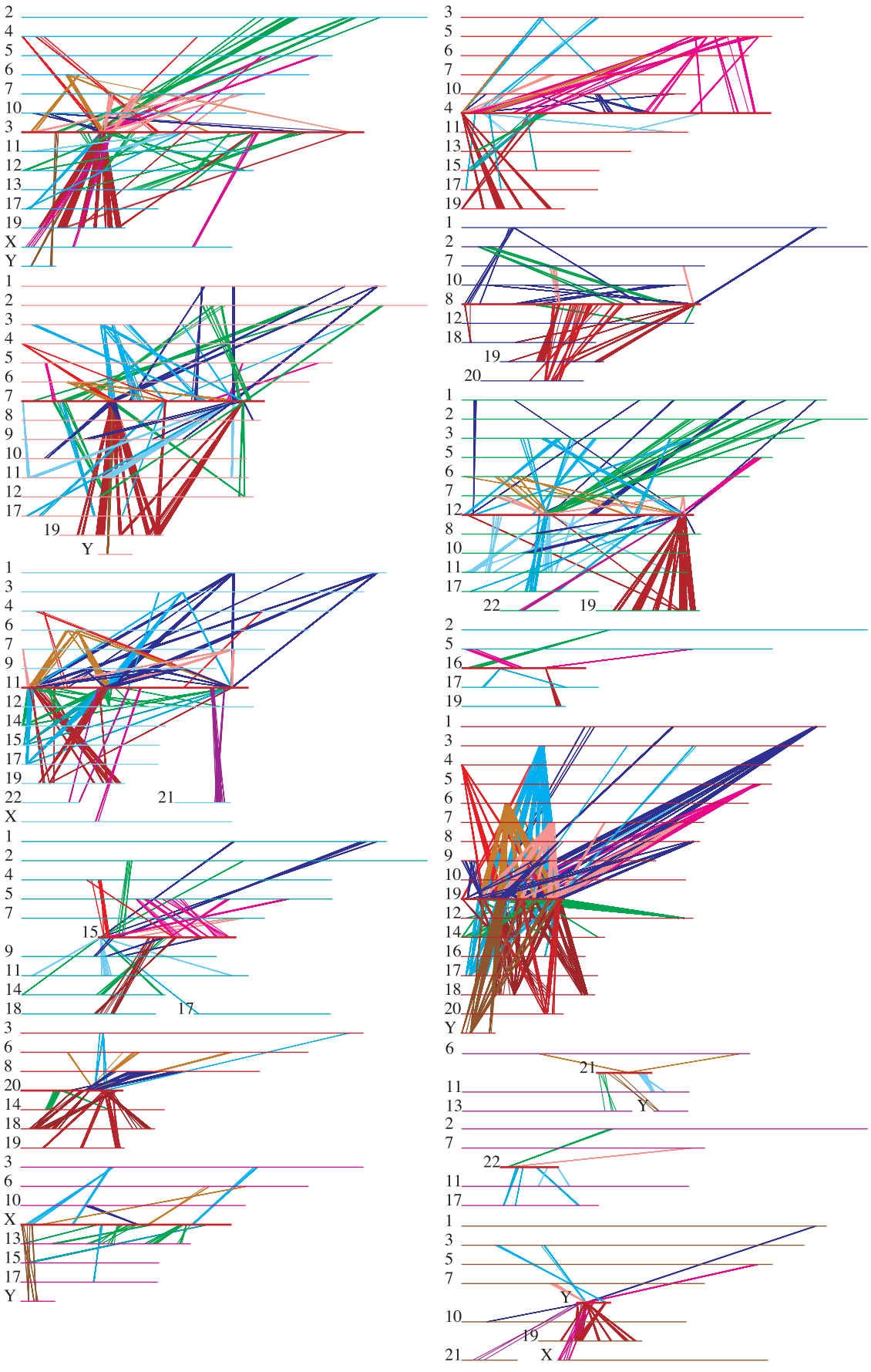


Fig. 13. Segmental duplications between chromosomes in the human genome. The 24 panels show the 1077 duplicated blocks of genes, containing 10,310 pairs of genes in total. Each line represents a pair of homologous genes belonging to a block; all blocks contain at least three genes on each of the chromosomes where they appear. Each panel shows all the duplications between a single chromosome and other chromosomes with shared blocks. The chromosome at the center of each panel is shown as a thick red line for emphasis. Other chromosomes are displayed from top to bottom within each panel ordered by chromosome number. The inset (bottom, center right) shows a close-up of one duplication between chromosomes 18 and 20, expanded to display the gene names of 12 of the 64 gene pairs shown.



somes, and whether this heterogeneity is greater than expected by chance. If SNPs occur by random and independent mutations, then it would seem that there ought to be a Poisson distribution of numbers of SNPs in fragments of arbitrary constant size. The observed dispersion in the distribution of SNPs in 100-kbp fragments was far greater than predicted from a Poisson distribution (Fig. 14). However, this simplistic model ignores the different recombination rates and population histories that exist in different regions of the genome. Population genetics theory holds that we can account for this variation with a mathematical formulation called the neutral coalescent (*109*). Applying well-tested algorithms for simulating the neutral coalescent with recombination (*110*), and using an effective population size of 10,000 and a per-base recombination rate equal to the mutation rate (*111*), we generated a distribution of numbers of SNPs by this model as well (*112*). The observed distribution of SNPs has a much larger variance than either the Poisson model or the coalescent model, and the difference is highly significant. This implies that there is significant variability across the genome in SNP density, an observation that begs an explanation.

Several attributes of the DNA sequence may affect the local density of SNPs, including the rate at which DNA polymerase makes errors and the efficacy of mismatch repair. One key factor that is likely to be associated with SNP density is the G+C content, in part because methylated cytosines in CpG dinucleotides tend to undergo deamination to form thymine, accounting for a nearly 10-fold increase in the mutation rate of CpGs over other dinucleotides. We tallied the GC content and nucleotide diversities in 100-kbp windows across the entire genome and found that the correlation between them was positive ($r = 0.21$) and highly significant ($P < 0.0001$), but G+C content accounted for only a small part of the variation.

6.5 SNPs by genomic class

To test homogeneity of SNP densities across functional classes, we partitioned sites into intergenic (defined as >5 kbp from any predicted transcription unit), 5′-UTR, exonic (missense and silent), intronic, and 3′-UTR for 10,239 known genes, derived from the NCBI RefSeq database and all human genes predicted from the Celera Otto annotation. In coding regions, SNPs were categorized as either silent, for those that do not change amino acid sequence, or missense, for those that change the protein product. The ratio of missense to silent coding SNPs in Celera-PFP, TSC, and Kwok sets (1.12, 0.91, and 0.78, respectively) shows a markedly reduced frequency of missense variants compared with the neutral expectation, consistent with the elimination by natural selection of a fraction of the deleterious amino acid changes (*112*). These ratios are comparable to the missense-to-silent ratios of 0.88 and 1.17 found by Cargill *et al*. (*101*) and by Halushka *et al*. (*102*). Similar results were observed in SNPs derived from Celera shotgun sequences (*46*).

It is striking how small is the fraction of SNPs that lead to potentially dysfunctional alterations in proteins. In the 10,239 RefSeq genes, missense SNPs were only about 0.12, 0.14, and 0.17% of the total SNP counts in Celera-PFP, TSC, and Kwok SNPs, respectively. Nonconservative protein changes constitute an even smaller fraction of missense SNPs (47, 41, and 40% in Celera-PFP, Kwok, and TSC). Intergenic regions have been virtually unstudied (*113*), and we note that 75% of the SNPs we identified were intergenic (Table 17). The SNP rate was highest in introns and lowest in exons. The SNP rate was lower in intergenic regions than in introns, providing one of the first discriminators between these two classes of DNA. These SNP rates were confirmed in the Celera SNPs, which also exhibited a lower rate in exons than in introns, and in extragenic regions than in introns (*46*). Many of these intergenic SNPs will provide valuable information in the form of markers for linkage and association studies, and some fraction is likely to have a regulatory function as well.

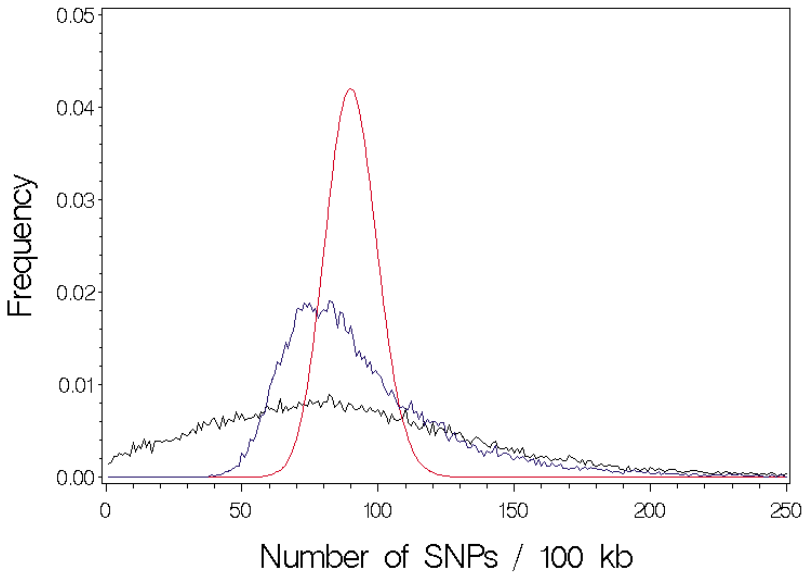


Fig. 14. SNP density in each 100-kbp interval as determined with Celera-PFP SNPs. The color codes are as follows: black, Celera-PFP SNP density; blue, coalescent model; and red, Poisson distribution. The figure shows that the distribution of SNPs along the genome is nonrandom and is not entirely accounted for by a coalescent model of regional history.

7 An Overview of the Predicted Protein-Coding Genes in the Human Genome

Summary. This section provides an initial computational analysis of the predicted protein set with the aim of cataloging prominent differences and similarities when the human genome is compared with other fully sequenced eukaryotic genomes. Over 40% of the predicted protein set in humans cannot be ascribed a molecular function by methods that assign proteins to known families. A protein domain–based analysis provides a detailed catalog of the prominent differences in the human genome when compared with the fly and worm genomes. Prominent among these are domain expansions in proteins involved in developmental regulation and in cellular processes such as neuronal function, hemostasis, acquired immune response, and cytoskeletal complexity. The final enumeration of protein families and details of protein structure will rely on additional experimental work and comprehensive manual curation.

A preliminary analysis of the predicted human protein-coding genes was conducted. Two methods were used to analyze and classify the molecular functions of 26,588 predicted proteins that represent 26,383 gene predictions with at least two lines of evidence as described above. The first method was based on an analysis at the level of protein families, with both the publicly available Pfam database (*114*, *115*) and Celera's Panther Classification (CPC) (Fig. 15) (*116*). The second method was based on an analysis at the level of protein domains, with both the Pfam and SMART databases (*115*, *117*).

The results presented here are preliminary and are subject to several limitations.

Both the gene predictions and functional assignments have been made by using computational tools, although the statistical models in Panther, Pfam, and SMART have been built, annotated, and reviewed by expert biologists. In the set of computationally predicted genes, we expect both false-positive predictions (some of these may in fact be inactive pseudogenes) and false-negative predictions (some human genes will not be computationally predicted). We also expect errors in delimiting the boundaries of exons and genes. Similarly, in the automatic functional assignments, we also expect both false-positive and false-negative predictions. The functional assignment protocol focuses on protein families that tend to be found across several organisms, or on families of known human genes. Therefore, we do not assign a function to many genes that are not in large families, even if the function is known. Unless otherwise specified, all enumeration of the genes in any given family or functional category was taken from the set of 26,588 predicted proteins, which were assigned functions by using statistical score cutoffs defined for models in Panther, Pfam, and SMART.

For this initial examination of the predicted human protein set, three broad questions were asked: (i) What are the likely molecular functions of the predicted gene products, and how are these proteins categorized with current classification methods? (ii) What are the core functions that appear to be common across the animals? (iii) How does the human protein complement differ from that of other sequenced eukaryotes?

7.1 Molecular functions of predicted human proteins

Figure 15 shows an overview of the putative molecular functions of the predicted 26,588 human proteins that have at least two lines of supporting evidence. About 41% (12,809) of the gene products could not be classified from this initial analysis and are termed proteins with unknown functions. Because our automatic classification methods treat only relatively large protein families, there are a number of "unclassified" sequences that do, in fact, have a known or predicted function. For the 60% of the protein set that have automatic functional predictions, the specific protein functions have been placed into broad classes. We focus here on molecular function (rather than higher order cellular processes) in order to classify as many proteins as possible. These functional predictions are based on similarity to sequences of known function.

In our analysis of the 12,731 additional low-confidence predicted genes (those with only one piece of supporting evidence), only 636 (5%) of these additional putative genes were assigned molecular functions by the automated methods. One-third of these 636 predicted genes represented endogenous retroviral proteins, further suggesting that the majority of these unknown-function genes are not real genes. Given that most of these additional 12,095 genes appear to be unique among the genomes sequenced to date, many may simply represent false-positive gene predictions.

The most common molecular functions are the transcription factors and those involved in nucleic acid metabolism (nucleic acid enzyme). Other functions that are highly represented in the human genome are the receptors, kinases, and hydrolases. Not surprisingly, most of the hydrolases are proteases. There are also many proteins that are members of proto-oncogene families, as well as families of "select regulatory molecules": (i) proteins involved in specific steps of signal transduction such as heterotrimeric GTP-binding proteins (G proteins) and cell cycle regulators, and (ii) proteins that modulate the activity of kinases, G proteins, and phosphatases.

Table 17. Distribution of SNPs in classes of genomic regions.

Genomic region class	Size of region examined (Mb)	Celera-PFP SNP density (SNP/Mb)
Intergenic	2185	707
Gene (intron + exon)	646	917
Intron	615	921
First intron	164	808
Exon	31	529
First exon	10	592

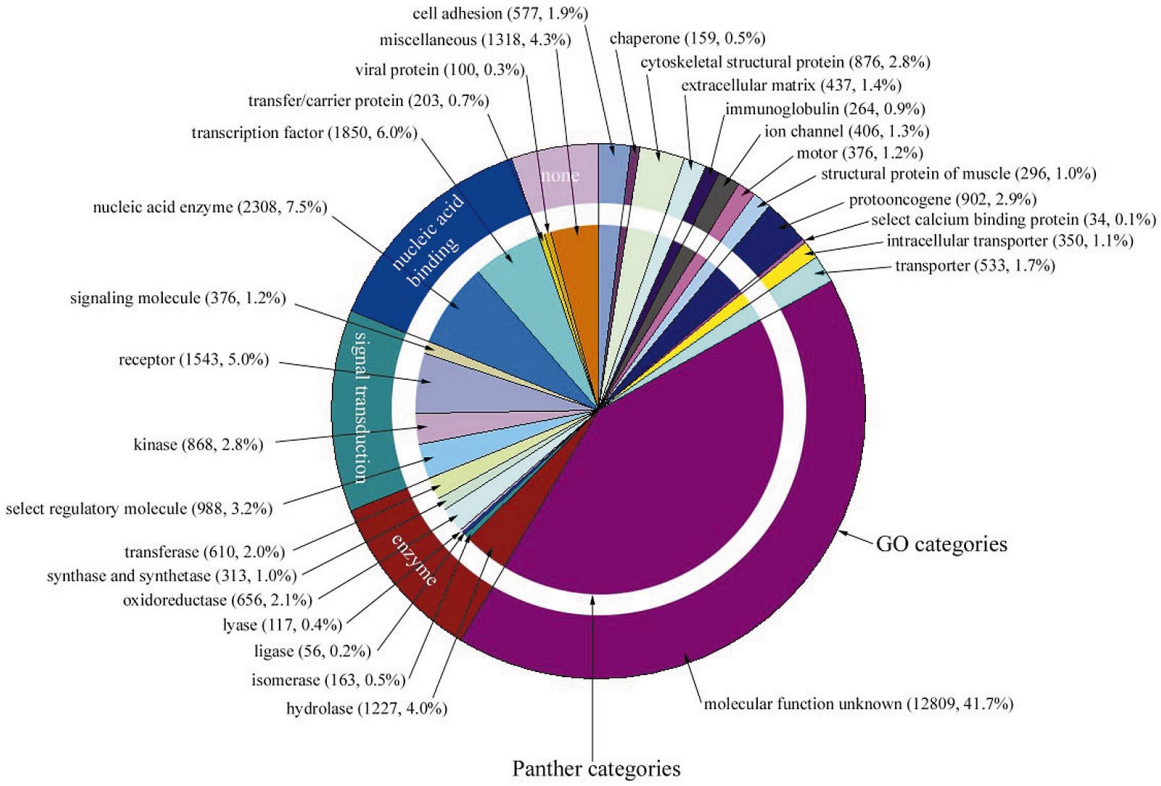


Fig. 15. Distribution of the molecular functions of 26,383 human genes. Each slice lists the numbers and percentages (in parentheses) of human gene functions assigned to a given category of molecular function. The outer circle shows the assignment to molecular function categories in the Gene Ontology (GO) (*179*), and the inner circle shows the assignment to Celera's Panther molecular function categories (*116*).

7.2 Evolutionary conservation of core processes

Because of the various "model organism" genome-sequencing projects that have already been completed, reasonable comparative information is available for beginning the analysis of the evolution of the human genome. The genomes of *S. cerevisiae* ("bakers' yeast") (*118*) and two diverse invertebrates, *C. elegans* (a nematode worm) (*119*) and *D. melanogaster* (fly) (*26*), as well as the first plant genome, *A. thaliana*, recently completed (*92*), provide a diverse background for genome comparisons.

We enumerated the "strict orthologs" conserved between human and fly, and between human and worm (Fig. 16) to address the question, What are the core functions that appear to be common across the animals? The concept of orthology is important because if two genes are orthologs, they can be traced by descent to the common ancestor of the two organisms (an "evolutionarily conserved protein set"), and therefore are likely to perform similar conserved functions in the different organisms. It is critical in this analysis to separate orthologs (a gene that appears in two organisms by descent from a common ancestor) from paralogs (a gene that appears in more than one copy in a given organism by a duplication event) because paralogs may subsequently diverge in function. Following the yeast-worm ortholog comparison in (*120*), we identified two different cases for each pairwise comparison (human-fly and human-worm). The first case was a pair of genes, one from each organism, for which there was no other close homolog in either organism. These are straightforwardly identified as orthologous, because there are no additional members of the families that complicate separating orthologs from paralogs. The second case is a family of genes with more than one member in either or both of the organisms being compared. Chervitz *et al.* (*120*) deal with this case by analyzing a phylogenetic tree that described the relationships between all of the sequences in both organisms, and then looked for pairs of genes that were nearest neighbors in the tree. If the nearest-neighbor pairs were from different organisms, those genes were presumed to be orthologs. We note that these nearest neighbors can often be confidently identified from pairwise sequence comparison without having to examine a phylogenetic tree (see legend to Fig. 16). If the nearest neighbors are not from different organisms, there has been a paralogous expansion in one or both organisms after the speciation event (and/or a gene loss by one organism). When this one-to-one correspondence is lost, defining an ortholog becomes ambiguous. For our initial computational overview of the predicted human protein set, we could not answer this question for every predicted protein. Therefore, we consider only "strict orthologs," i.e., the proteins with unambiguous one-to-one relationships (Fig. 16). By these criteria, there are 2758 strict human-fly orthologs, 2031 human-worm (1523 in common between these sets). We define the evolutionarily conserved set as those 1523 human proteins that have strict orthologs in both *D. melanogaster* and *C. elegans.*

The distribution of the functions of the conserved protein set is shown in Fig. 16. Comparison with Fig. 15 shows that, not surprisingly, the set of conserved proteins is not distributed among molecular functions in the same way as the whole human protein set. Compared with the whole human set (Fig. 15), there are several categories that are overrepresented in the conserved set by a factor of ~2 or more. The first category is nucleic acid enzymes, primarily the transcriptional machinery (notably DNA/RNA methyltransferases, DNA/RNA polymerases, helicases, DNA ligases, DNA- and RNA-processing factors, nucleases, and ribosomal proteins). The basic transcriptional and translational machinery is well known to have been conserved over evolution, from bacteria through to the most complex eukaryotes. Many ribonucleoproteins involved in RNA splicing also appear to be conserved among the animals. Other enzyme types are also overrepresented (transferases, oxidoreductases, ligases, lyases, and isomerases). Many of these en-

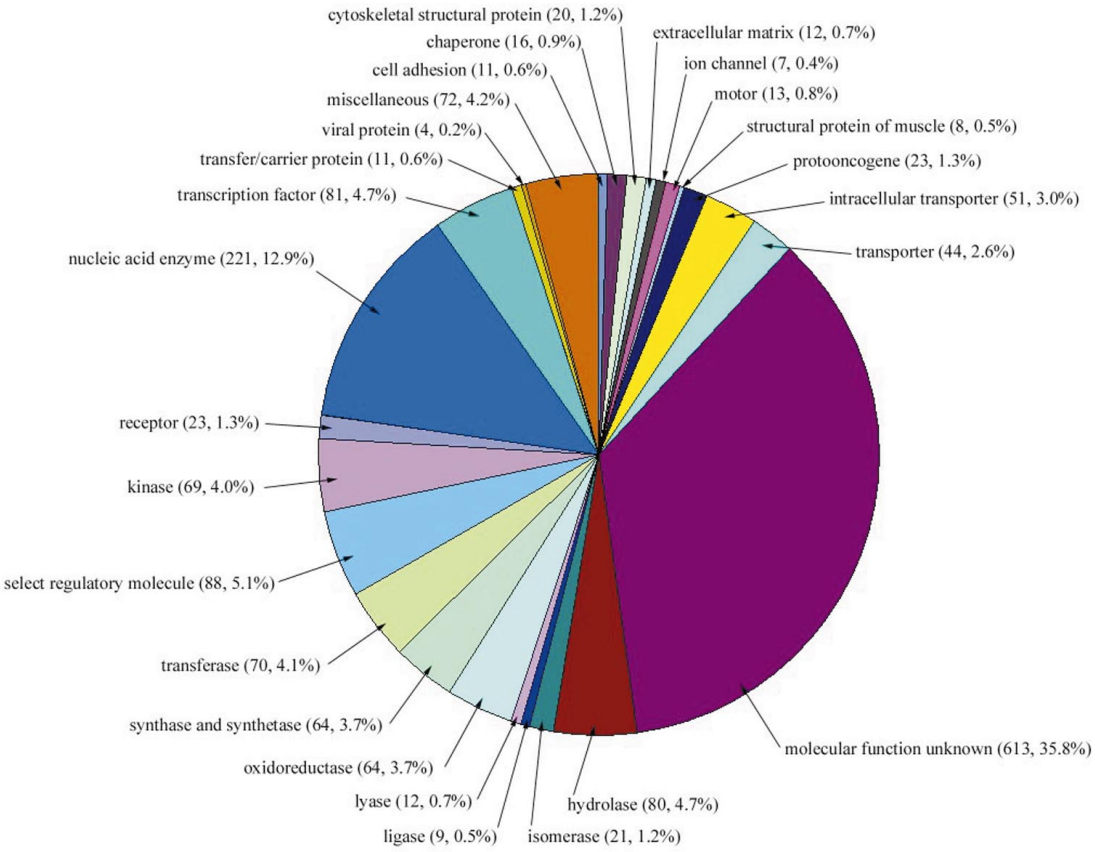


Fig. 16. Functions of putative orthologs across vertebrate and invertebrate genomes. Each slice lists the number and percentages (in parentheses) of "strict orthologs" between the human, fly, and worm genomes involved in a given category of molecular function. "Strict orthologs" are defined here as bi-directional BLAST best hits (*180*) such that each orthologous pair (i) has a BLASTP *P*-value of $\leq 10^{-10}$ (*120*), and (ii) has a more significant BLASTP score than any paralogs in either organism, i.e., there has likely been no duplication subsequent to speciation that might make the orthology ambiguous. This measure is quite strict and is a lower bound on the number of orthologs. By these criteria, there are 2758 strict human-fly orthologs, and 2031 human-worm orthologs (1523 in common between these sets).

zymes are involved in intermediary metabolism. The only exception is the hydrolase category, which is not significantly overrepresented in the shared protein set. Proteases form the largest part of this category, and several large protease families have expanded in each of these three organisms after their divergence. The category of select regulatory molecules is also overrepresented in the conserved set. The major conserved families are small guanosine triphosphatases (GTPases) (especially the Ras-related superfamily, including ADP ribosylation factor) and cell cycle regulators (particularly the cullin family, cyclin C family, and several cell division protein kinases). The last two significantly overrepresented categories are protein transport and trafficking, and chaperones. The most conserved groups in these categories are proteins involved in coated vesicle-mediated transport, and chaperones involved in protein folding and heat-shock response [particularly the DNAJ family, and heat-shock protein 60 (HSP60), HSP70, and HSP90 families]. These observations provide only a conservative estimate of the protein families in the context of specific cellular processes that were likely derived from the last common ancestor of the human, fly, and worm. As stated before, this analysis does not provide a complete estimate of conservation across the three animal genomes, as paralogous duplication makes the determination of true orthologs difficult within the members of conserved protein families.

7.3 Differences between the human genome and other sequenced eukaryotic genomes

To explore the molecular building blocks of the vertebrate taxon, we have compared the human genome with the other sequenced eukaryotic genomes at three levels: molecular functions, protein families, and protein domains.

Molecular differences can be correlated with phenotypic differences to begin to reveal the developmental and cellular processes that are unique to the vertebrates. Tables 18 and 19 display a comparison among all sequenced eukaryotic genomes, over selected protein/domain families (defined by sequence similarity, e.g., the serine-threonine protein kinases) and superfamilies (defined by shared molecular function, which may include several sequence-related families, e.g., the cytokines). In these tables we have focused on (super) families that are either very large or that differ significantly in humans compared with the other sequenced eukaryote genomes. We have found that the most prominent human expansions are in proteins involved in (i) acquired immune functions; (ii) neural development, structure, and functions; (iii) intercellular and intracellular signaling pathways in development and homeostasis; (iv) hemostasis; and (v) apoptosis.

Acquired immunity. One of the most striking differences between the human genome and the *Drosophila* or *C. elegans* genome is the appearance of genes involved in acquired immunity (Tables 18 and 19). This is expected, because the acquired immune response is a defense system that only occurs in vertebrates. We observe 22 class I and 22 class II major histocompatibility complex (MHC) antigen genes and 114 other immunoglobulin genes in the human genome. In addition, there are 59 genes in the cognate immunoglobulin receptor family. At the domain level, this is exemplified by an expansion and recruitment of the ancient immunoglobulin fold to constitute molecules such as MHC, and of the integrin fold to form several of the cell adhesion molecules that mediate interactions between immune effector cells and the extracellular matrix. Vertebrate-specific proteins include the paracrine immune regulators family of secreted 4-alpha helical bundle proteins, namely the cytokines and chemokines. Some of the cytoplasmic signal transduction components associated with cytokine receptor signal transduction are also features that are poorly represented in the fly and worm. These include protein domains found in the signal transducer and activator of transcription (STATs), the suppressors of cytokine signaling (SOCS), and protein inhibitors of activated STATs (PIAS). In contrast, many of the animal-specific protein domains that play a role in innate immune response, such as the Toll receptors, do not appear to be significantly expanded in the human genome.

Neural development, structure, and function. In the human genome, as compared with the worm and fly genomes, there is a marked increase in the number of members of protein families that are involved in neural development. Examples include neurotrophic factors such as ependymin, nerve growth factor, and signaling molecules such as semaphorins, as well as the number of proteins involved directly in neural structure and function such as myelin proteins, voltage-gated ion channels, and synaptic proteins such as synaptotagmin. These observations correlate well with the known phenotypic differences between the nervous systems of these taxa, notably (i) the increase in the number and connectivity of neurons; (ii) the increase in number of distinct neural cell types (as many as a thousand or more in human compared with a few hundred in fly and worm) (*121*); (iii) the increased length of individual axons; and (iv) the significant increase in glial cell number, especially the appearance of myelinating glial cells, which are electrically inert supporting cells differentiated from the same stem cells as neurons. A number of prominent protein expansions are involved in the processes of neural development. Of the extracellular domains that mediate cell adhesion, the connexin domain–containing proteins (*122*) exist only in humans. These proteins, which are not present in the *Drosophila* or *C. elegans* genomes, appear to provide the constitutive subunits of intercellular channels and the structural basis for electrical coupling. Pathway finding by axons and neuronal network formation is mediated through a subset of ephrins and their cognate receptor tyrosine kinases that act as positional labels to establish topographical projections (*123*). The probable biological role for the semaphorins (22 in human compared with 6 in the fly and 2 in the worm) and their receptors (neuropilins and plexins) is that of axonal guidance molecules (*124*). Signaling molecules such as neurotrophic factors and some cytokines have been shown to regulate neuronal cell survival, proliferation, and axon guidance (*125*). Notch receptors and ligands play important roles in glial cell fate determination and gliogenesis (*126*).

Other human expanded gene families play key roles directly in neural structure and function. One example is synaptotagmin (expanded more than twofold in humans relative to the invertebrates), originally found to regulate synaptic transmission by serving as a Ca^{2+} sensor (or receptor) during synaptic vesicle fusion and release (*127*). Of interest is the increased co-occurrence in humans of PDZ and the SH3 domains in neuronal-specific adaptor molecules; examples include proteins that likely modulate channel activity at synaptic junctions (*128*). We also noted expansions in several ion-channel families (Table 19), including the EAG subfamily (related to cyclic nucleotide gated channels), the voltage-gated calcium/sodium channel family, the inward-rectifier potassium channel family, and the voltage-gated potassium channel, alpha subunit family. Voltage-gated sodium and potassium channels are involved in the generation of action potentials in neurons. Together with voltage-gated calcium channels, they also play a key role in coupling action potentials to neurotransmitter release, in the development of neurites, and in short-term memory. The recent observation of a calcium-regulated association between sodium channels and synaptotagmin may have consequences for the establishment and regulation of neuronal excitability (*129*).

Myelin basic protein and myelin-associated glycoprotein are major classes of protein components in both the central and peripheral nervous system of vertebrates. Myelin P0 is a major component of peripheral myelin, and myelin proteolipid and myelin oligodendrocyte glycopotein are found in the central nervous system. Mutations in any of these

Table 18. Domain-based comparative analysis of proteins in *H. sapiens* (H), *D. melanogaster* (F), *C. elegans* (W), *S. cerevisiae* (Y), and *A. thaliana* (A). The predicted protein set of each of the above eukaryotic organisms was analyzed with Pfam version 5.5 using E value cutoffs of 0.001. The number of proteins containing the specified Pfam domains as well as the total number of domains (in parentheses) are shown in each column. Domains were categorized into cellular processes for presentation. Some domains (i.e., SH2) are listed in more than one cellular process. Results of the Pfam analysis may differ from results obtained based on human curation of protein families, owing to the limitations of large-scale automatic classifications. Representative examples of domains with reduced counts owing to the stringent E value cutoff used for this analysis are marked with a double asterisk (**). Examples include short divergent and predominantly alpha-helical domains, and certain classes of cysteine-rich zinc finger proteins.

Accession number	Domain name	Domain description	H	F	W	Y	A
		Developmental and homeostatic regulators					
PF02039	Adrenomedullin	Adrenomedullin	1	0	0	0	0
PF00212	ANP	Atrial natriuretic peptide	2	0	0	0	0
PF00028	Cadherin	Cadherin domain	100 (550)	14 (157)	16 (66)	0	0
PF00214	Calc_CGRP_IAPP	Calcitonin/CGRP/IAPP family	3	0	0	0	0
PF01110	CNTF	Ciliary neurotrophic factor	1	0	0	0	0
PF01093	Clusterin	Clusterin	3	0	0	0	0
PF00029	Connexin	Connexin	14 (16)	0	0	0	0
PF00976	ACTH_domain	Corticotropin ACTH domain	1	0	0	0	0
PF00473	CRF	Corticotropin-releasing factor family	2	1	0	0	0
PF00007	Cys_knot	Cystine-knot domain	10 (11)	2	0	0	0
PF00778	DIX	Dix domain	5	2	4	0	0
PF00322	Endothelin	Endothelin family	3	0	0	0	0
PF00812	Ephrin	Ephrin	7 (8)	2	4	0	0
PF01404	EPh_lbd	Ephrin receptor ligand binding domain	12	2	1	0	0
PF00167	FGF	Fibroblast growth factor	23	1	1	0	0
PF01534	Frizzled	Frizzled/Smoothened family membrane region	9	7	3	0	0
PF00236	Hormone6	Glycoprotein hormones	1	0	0	0	0
PF01153	Glypican	Glypican	14	2	1	0	0
PF01271	Granin	Grainin (chromogranin or secretogranin)	3	0	0	0	0
PF02058	Guanylin	Guanylin precursor	1	0	0	0	0
PF00049	Insulin	Insulin/IGF/Relaxin family	7	4	0	0	0
PF00219	IGFBP	Insulin-like growth factor binding proteins	10	0	0	0	0
PF02024	Leptin	Leptin	1	0	0	0	0
PF00193	Xlink	LINK (hyaluron binding)	13 (23)	0	1	0	0
PF00243	NGF	Nerve growth factor family	3	0	0	0	0
PF02158	Neuregulin	Neuregulin family	4	0	0	0	0
PF00184	Hormone5	Neurohypophysial hormones	1	0	0	0	0
PF02070	NMU	Neuromedin U	1	0	0	0	0
PF00066	Notch	Notch (DSL) domain	3 (5)	2 (4)	2 (6)	0	0
PF00865	Osteopontin	Osteopontin	1	0	0	0	0
PF00159	Hormone3	Pancreatic hormone peptides	3	0	0	0	0
PF01279	Parathyroid	Parathyroid hormone family	2	0	0	0	0
PF00123	Hormone2	Peptide hormone	5 (9)	0	0	0	0
PF00341	PDGF	Platelet-derived growth factor (PDGF)	5	1	0	0	0
PF01403	Sema	Sema domain	27 (29)	8 (10)	3 (4)	0	0
PF01033	Somatomedin_B	Somatomedin B domain	5 (8)	3	0	0	0
PF00103	Hormone	Somatotropin	1	0	0	0	0
PF02208	Sorb	Sorbin homologous domain	2	0	0	0	0
PF02404	SCF	Stem cell factor	2	0	0	0	0
PF01034	Syndecan	Syndecan domain	3	1	1	0	0
PF00020	TNFR_c6	TNFR/NGFR cysteine-rich region	17 (31)	1	0	0	0
PF00019	TGF-β	Transforming growth factor β-like domain	27 (28)	6	4	0	0
PF01099	Uteroglobin	Uteroglobin family	3	0	0	0	0
PF01160	Opiods_neuropep	Vertebrate endogenous opioids neuropeptide	3	0	0	0	0
PF00110	Wnt	Wnt family of developmental signaling proteins	18	7 (10)	5	0	0
		Hemostasis					
PF01821	ANATO	Anaphylotoxin-like domain	6 (14)	0	0	0	0
PF00386	C1q	C1q domain	24	0	0	0	0
PF00200	Disintegrin	Disintegrin	18	2	3	0	0
PF00754	F5_F8_type_C	F5/8 type C domain	15 (20)	5 (6)	2	0	0
PF01410	COLFI	Fibrillar collagen C-terminal domain	10	0	0	0	0
PF00039	Fn1	Fibronectin type I domain	5 (18)	0	0	0	0
PF00040	Fn2	Fibronectin type II domain	11 (16)	0	0	0	0
PF00051	Kringle	Kringle domain	15 (24)	2	2	0	0
PF01823	MACPF	MAC/Perforin domain	6	0	0	0	0
PF00354	Pentaxin	Pentaxin family	9	0	0	0	0
PF00277	SAA_proteins	Serum amyloid A protein	4	0	0	0	0
PF00084	Sushi	Sushi domain (SCR repeat)	53 (191)	11 (42)	8 (45)	0	0
PF02210	TSPN	Thrombospondin N-terminal–like domains	14	1	0	0	0
PF01108	Tissue_fac	Tissue factor	1	0	0	0	0
PF00868	Transglutamin_N	Transglutaminase family	6	1	0	0	0
PF00927	Transglutamin_C	Transglutaminase family	8	1	0	0	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00594	Gla	Vitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain	11	0	0	0	0
		Immune response					
PF00711	Defensin_beta	Beta defensin	1	0	0	0	0
PF00748	Calpain_inhib	Calpain inhibitor repeat	3 (9)	0	0	0	0
PF00666	Cathelicidins	Cathelicidins	2	0	0	0	0
PF00129	MHC_I	Class I histocompatibility antigen, domains alpha 1 and 2	18 (20)	0	0	0	0
PF00993	MHC_II_alpha**	Class II histocompatibility antigen, alpha domain	5 (6)	0	0	0	0
PF00969	MHC_II_beta**	Class II histocompatibility antigen, beta domain	7	0	0	0	0
PF00879	Defensin_propep	Defensin propeptide	3	0	0	0	0
PF01109	GM_CSF	Granulocyte-macrophage colony-stimulating factor	1	0	0	0	0
PF00047	Ig	Immunoglobulin domain	381 (930)	125 (291)	67 (323)	0	0
PF00143	Interferon	Interferon alpha/beta domain	7 (9)	0	0	0	0
PF00714	IFN-gamma	Interferon gamma	1	0	0	0	0
PF00726	IL10	Interleukin-10	1	0	0	0	0
PF02372	IL15	Interleukin-15	1	0	0	0	0
PF00715	IL2	Interleukin-2	1	0	0	0	0
PF00727	IL4	Interleukin-4	1	0	0	0	0
PF02025	IL5	Interleukin-5	1	0	0	0	0
PF01415	IL7	Interleukin-7/9 family	1	0	0	0	0
PF00340	IL1	Interleukin-1	7	0	0	0	0
PF02394	IL1_propep	Interleukin-1 propeptide	1	0	0	0	0
PF02059	IL3	Interleukin-3	1	0	0	0	0
PF00489	IL6	Interleukin-6/G-CSF/MGF family	2	0	0	0	0
PF01291	LIF_OSM	Leukemia inhibitory factor (LIF)/oncostatin (OSM) family	2	0	0	0	0
PF00323	Defensins	Mammalian defensin	2	0	0	0	0
PF01091	PTN_MK	PTN/MK heparin-binding protein	2	0	0	0	0
PF00277	SAA_proteins	Serum amyloid A protein	4	0	0	0	0
PF00048	IL8	Small cytokines (intecrine/chemokine), interleukin-8 like	32	0	0	0	0
PF01582	TIR	TIR domain	18	8	2	0	131 (143)
PF00229	TNF	TNF (tumor necrosis factor) family	12	0	0	0	0
PF00088	Trefoil	Trefoil (P-type) domain	5 (6)	0	2	0	0
		PI-PY-rho GTPase signaling					
PF00779	BTK	BTK motif	5	1	0	0	0
PF00168	C2	C2 domain	73 (101)	32 (44)	24 (35)	6 (9)	66 (90)
PF00609	DAGKa	Diacylglycerol kinase accessory domain (presumed)	9	4	7	0	6
PF00781	DAGKc	Diacylglycerol kinase catalytic domain (presumed)	10	8	8	2	11 (12)
PF00610	DEP	Domain found in Dishevelled, Egl-10, and Pleckstrin (DEP)	12 (13)	4	10	5	2
PF01363	FYVE	FYVE zinc finger	28 (30)	14	15	5	15
PF00996	GDI	GDP dissociation inhibitor	6	2	1	1	3
PF00503	G-alpha	G-protein alpha subunit	27 (30)	10	20 (23)	2	5
PF00631	G-gamma	G-protein gamma like domains	16	5	5	1	0
PF00616	RasGAP	GTPase-activator protein for Ras-like GTPase	11	5	8	3	0
PF00618	RasGEFN	Guanine nucleotide exchange factor for Ras-like GTPases; N-terminal motif	9	2	3	5	0
PF00625	Guanylate_kin	Guanylate kinase	12	8	7	1	4
PF02189	ITAM	Immunoreceptor tyrosine-based activation motif	3	0	0	0	0
PF00169	PH	PH domain	193 (212)	72 (78)	65 (68)	24	23
PF00130	DAG_PE-bind	Phorbol esters/diacylglycerol binding domain (C1 domain)	45 (56)	25 (31)	26 (40)	1 (2)	4
PF00388	PI-PLC-X	Phosphatidylinositol-specific phospholipase C, X domain	12	3	7	1	8
PF00387	PI-PLC-Y	Phosphatidylinositol-specific phospholipase C, Y domain	11	2	7	1	8
PF00640	PID	Phosphotyrosine interaction domain (PTB/PID)	24 (27)	13	11 (12)	0	0
PF02192	PI3K_p85B	PI3-kinase family, p85-binding domain	2	1	1	0	0
PF00794	PI3K_rbd	PI3-kinase family, ras-binding domain	6	3	1	0	0
PF01412	ArfGAP	Putative GTP-ase activating protein for Arf	16	9	8	6	15
PF02196	RBD	Raf-like Ras-binding domain	6 (7)	4	1	0	0
PF02145	Rap_GAP	Rap/ran-GAP	5	4	2	0	0
PF00788	RA	Ras association (RalGDS/AF-6) domain	18 (19)	7 (9)	6	1	0
PF00071	Ras	Ras family	126	56 (57)	51	23	78
PF00617	RasGEF	RasGEF domain	21	8	7	5	0
PF00615	RGS	Regulator of G protein signaling domain	27	6 (7)	12 (13)	1	0
PF02197	RIIa	Regulatory subunit of type II PKA R-subunit	4	1	2	1	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00620	RhoGAP	RhoGAP domain	59	19	20	9	8
PF00621	RhoGEF	RhoGEF domain	46	23 (24)	18 (19)	3	0
PF00536	SAM	SAM domain (Sterile alpha motif)	29 (31)	15	8	3	6
PF01369	Sec7	Sec7 domain	13	5	5	5	9
PF00017	SH2	Src homology 2 (SH2) domain	87 (95)	33 (39)	44 (48)	1	3
PF00018	SH3	Src homology 3 (SH3) domain	143 (182)	55 (75)	46 (61)	23 (27)	4
PF01017	STAT	STAT protein	7	1	1 (2)	0	0
PF00790	VHS	VHS domain	4	2	4	4	8
PF00568	WH1	WH1 domain	7	2	2 (3)	1	0
		Domains involved in apoptosis					
PF00452	Bcl-2	Bcl-2	9	2	1	0	0
PF02180	BH4	Bcl-2 homology region 4	3	0	1	0	0
PF00619	CARD	Caspase recruitment domain	16	0	2	0	0
PF00531	Death	Death domain	16	5	7	0	0
PF01335	DED	Death effector domain	4 (5)	0	0	0	0
PF02179	BAG	Domain present in Hsp70 regulators	5 (8)	3	2	1	5
PF00656	ICE_p20	ICE-like protease (caspase) p20 domain	11	7	3	0	0
PF00653	BIR	Inhibitor of Apoptosis domain	8 (14)	5 (9)	2 (3)	1 (2)	0
		Cytoskeletal					
PF00022	Actin	Actin	61 (64)	15 (16)	12	9 (11)	24
PF00191	Annexin	Annexin	16 (55)	4 (16)	4 (11)	0	6 (16)
PF00402	Calponin	Calponin family	13 (22)	3	7 (19)	0	0
PF00373	Band_41	FERM domain (Band 4.1 family)	29 (30)	17 (19)	11 (14)	0	0
PF00880	Nebulin_repeat	Nebulin repeat	4 (148)	1 (2)	1	0	0
PF00681	Plectin_repeat	Plectin repeat	2 (11)	0	0	0	0
PF00435	Spectrin	Spectrin repeat	31 (195)	13 (171)	10 (93)	0	0
PF00418	Tubulin-binding	Tau and MAP proteins, tubulin-binding	4 (12)	1 (4)	2 (8)	0	0
PF00992	Troponin	Troponin	4	6	8	0	0
PF02209	VHP	Villin headpiece domain	5	2	2	0	5
PF01044	Vinculin	Vinculin family	4	2	1	0	0
		ECM adhesion					
PF01391	Collagen	Collagen triple helix repeat (20 copies)	65 (279)	10 (46)	174 (384)	0	0
PF01413	C4	C-terminal tandem repeated domain in type 4 procollagen	6 (11)	2 (4)	3 (6)	0	0
PF00431	CUB	CUB domain	47 (69)	9 (47)	43 (67)	0	0
PF00008	EGF	EGF-like domain	108 (420)	45 (186)	54 (157)	0	1
PF00147	Fibrinogen_C	Fibrinogen beta and gamma chains, C-terminal globular domain	26	10 (11)	6	0	0
PF00041	Fn3	Fibronectin type III domain	106 (545)	42 (168)	34 (156)	0	1
PF00757	Furin-like	Furin-like cysteine rich region	5	2	1	0	0
PF00357	Integrin_A	Integrin alpha cytoplasmic region	3	1	2	0	0
PF00362	Integrin_B	Integrins, beta chain	8	2	2	0	0
PF00052	Laminin_B	Laminin B (Domain IV)	8 (12)	4 (7)	6 (10)	0	0
PF00053	Laminin_EGF	Laminin EGF-like (Domains III and V)	24 (126)	9 (62)	11 (65)	0	0
PF00054	Laminin_G	Laminin G domain	30 (57)	18 (42)	14 (26)	0	0
PF00055	Laminin_Nterm	Laminin N-terminal (Domain VI)	10	6	4	0	0
PF00059	Lectin_c	Lectin C-type domain	47 (76)	23 (24)	91 (132)	0	0
PF01463	LRRCT	Leucine rich repeat C-terminal domain	69 (81)	23 (30)	7 (9)	0	0
PF01462	LRRNT	Leucine rich repeat N-terminal domain	40 (44)	7 (13)	3 (6)	0	0
PF00057	Ldl_recept_a	Low-density lipoprotein receptor domain class A	35 (127)	33 (152)	27 (113)	0	0
PF00058	Ldl_recept_b	Low-density lipoprotein receptor repeat class B	15 (96)	9 (56)	7 (22)	0	0
PF00530	SRCR	Scavenger receptor cysteine-rich domain	11 (46)	4 (8)	1 (2)	0	0
PF00084	Sushi	Sushi domain (SCR repeat)	53 (191)	11 (42)	8 (45)	0	0
PF00090	Tsp_1	Thrombospondin type 1 domain	41 (66)	11 (23)	18 (47)	0	0
PF00092	Vwa	von Willebrand factor type A domain	34 (58)	0	17 (19)	0	1
PF00093	Vwc	von Willebrand factor type C domain	19 (28)	6 (11)	2 (5)	0	0
PF00094	Vwd	von Willebrand factor type D domain	15 (35)	3 (7)	9	0	0
		Protein interaction domains					
PF00244	14-3-3	14-3-3 proteins	20	3	3	2	15
PF00023	Ank	Ank repeat	145 (404)	72 (269)	75 (223)	12 (20)	66 (111)
PF00514	Armadillo_seg	Armadillo/beta-catenin-like repeats	22 (56)	11 (38)	3 (11)	2 (10)	25 (67)
PF00168	C2	C2 domain	73 (101)	32 (44)	24 (35)	6 (9)	66 (90)
PF00027	cNMP_binding	Cyclic nucleotide-binding domain	26 (31)	21 (33)	15 (20)	2 (3)	22
PF01556	DnaJ_C	DnaJ C terminal region	12	9	5	3	19
PF00226	DnaJ	DnaJ domain	44	34	33	20	93
PF00036	Efhand**	EF hand	83 (151)	64 (117)	41 (86)	4 (11)	120 (328)
PF00611	FCH	Fes/CIP4 homology domain	9	3	2	4	0
PF01846	FF	FF domain	4 (11)	4 (10)	3 (16)	2 (5)	4 (8)
PF00498	FHA	FHA domain	13	15	7	13 (14)	17

myelin proteins result in severe demyelination, which is a pathological condition in which the myelin is lost and the nerve conduction is severely impaired (*130*). Humans have at least 10 genes belonging to four different families involved in myelin production (five myelin P0, three myelin proteolipid, myelin basic protein, and myelin-oligodendrocyte glycoprotein, or MOG), and possibly more-remotely related members of the MOG family. Flies have only a single myelin proteolipid, and worms have none at all.

Intercellular and intracellular signaling pathways in development and homeostasis. Many protein families that have expanded in humans relative to the invertebrates are involved in signaling processes, particularly in response to development and differentiation

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00254	FKBP	FKBP-type peptidyl-prolyl cis-trans isomerases	15 (20)	7 (8)	7 (13)	4	24 (29)
PF01590	GAF	GAF domain	7 (8)	2 (4)	1	0	10
PF01344	Kelch	Kelch motif	54 (157)	12 (48)	13 (41)	3	102 (178)
PF00560	LRR**	Leucine Rich Repeat	25 (30)	24 (30)	7 (11)	1	15 (16)
PF00917	MATH	MATH domain	11	5	88 (161)	1	61 (74)
PF00989	PAS	PAS domain	18 (19)	9 (10)	6	1	13 (18)
PF00595	PDZ	PDZ domain (Also known as DHR or GLGF)	96 (154)	60 (87)	46 (66)	2	5
PF00169	PH	PH domain	193 (212)	72 (78)	65 (68)	24	23
PF01535	PPR**	PPR repeat	5	3 (4)	0	1	474 (2485)
PF00536	SAM	SAM domain (Sterile alpha motif)	29 (31)	15	8	3	6
PF01369	Sec7	Sec7 domain	13	5	5	5	9
PF00017	SH2	Src homology 2 (SH2) domain	87 (95)	33 (39)	44 (48)	1	3
PF00018	SH3	Src homology 3 (SH3) domain	143 (182)	55 (75)	46 (61)	23 (27)	4
PF01740	STAS	STAS domain	5	1	6	2	13
PF00515	TPR**	TPR domain	72 (131)	39 (101)	28 (54)	16 (31)	65 (124)
PF00400	WD40**	WD40 domain	136 (305)	98 (226)	72 (153)	56 (121)	167 (344)
PF00397	WW	WW domain	32 (53)	24 (39)	16 (24)	5 (8)	11 (15)
PF00569	ZZ	ZZ-Zinc finger present in dystrophin, CBP/p300	10 (11)	13	10	2	10
		Nuclear interaction domains					
PF01754	Zf-A20	A20-like zinc finger	2 (8)	2	2	0	8
PF01388	ARID	ARID DNA binding domain	11	6	4	2	7
PF01426	BAH	BAH domain	8 (10)	7 (8)	4 (5)	5	21 (25)
PF00643	Zf-B_box**	B-box zinc finger	32 (35)	1	2	0	0
PF00533	BRCT	BRCA1 C Terminus (BRCT) domain	17 (28)	10 (18)	23 (35)	10 (16)	12 (16)
PF00439	Bromodomain	Bromodomain	37 (48)	16 (22)	18 (26)	10 (15)	28
PF00651	BTB	BTB/POZ domain	97 (98)	62 (64)	86 (91)	1 (2)	30 (31)
PF00145	DNA_methylase	C-5 cytosine-specific DNA methylase	3 (4)	1	0	0	13 (15)
PF00385	Chromo	chromo' (CHRromatin Organization MOdifier) domain	24 (27)	14 (15)	17 (18)	1 (2)	12
PF00125	Histone	Core histone H2A/H2B/H3/H4	75 (81)	5	71 (73)	8	48
PF00134	Cyclin	Cyclin	19	10	10	11	35
PF00270	DEAD	DEAD/DEAH box helicase	63 (66)	48 (50)	55 (57)	50 (52)	84 (87)
PF01529	Zf-DHHC	DHHC zinc finger domain	15	20	16	7	22
PF00646	F-box**	F-box domain	16	15	309 (324)	9	165 (167)
PF00250	Fork_head	Fork head domain	35 (36)	20 (21)	15	4	0
PF00320	GATA	GATA zinc finger	11 (17)	5(6)	8 (10)	9	26
PF01585	G-patch	G-patch domain	18	16	13	4	14 (15)
PF00010	HLH**	Helix-loop-helix DNA-binding domain	60 (61)	44	24	4	39
PF00850	Hist_deacetyl	Histone deacetylase family	12	5 (6)	8 (10)	5	10
PF00046	Homeobox	Homeobox domain	160 (178)	100 (103)	82 (84)	6	66
PF01833	TIG	IPT/TIG domain	29 (53)	11 (13)	5 (7)	2	1
PF02373	JmjC	JmjC domain	10	4	6	4	7
PF02375	JmjN	JmjN domain	7	4	2	3	7
PF00013	KH-domain	KH domain	28 (67)	14 (32)	17 (46)	4 (14)	27 (61)
PF01352	KRAB	KRAB box	204 (243)	0	0	0	0
PF00104	Hormone_rec	Ligand-binding domain of nuclear hormone receptor	47	17	142 (147)	0	0
PF00412	LIM	LIM domain containing proteins	62 (129)	33 (83)	33 (79)	4 (7)	10 (16)
PF00917	MATH	MATH domain	11	5	88 (161)	1	61 (74)
PF00249	Myb_DNA-binding	Myb-like DNA-binding domain	32 (43)	18 (24)	17 (24)	15 (20)	243 (401)
PF02344	Myc-LZ	Myc leucine zipper domain	1	0	0	0	0
PF01753	Zf-MYND	MYND finger	14	14	9	1	7
PF00628	PHD	PHD-finger	68 (86)	40 (53)	32 (44)	14 (15)	96 (105)
PF00157	Pou	Pou domain—N-terminal to homeobox domain	15	5	4	0	0
PF02257	RFX_DNA_binding	RFX DNA-binding domain	7	2	1	1	0
PF00076	Rrm	RNA recognition motif (a.k.a. RRM, RBD, or RNP domain)	224 (324)	127 (199)	94 (145)	43 (73)	232 (369)
PF02037	SAP	SAP domain	15	8	5	5	6 (7)
PF00622	SPRY	SPRY domain	44 (51)	10 (12)	5 (7)	3	6
PF01852	START	START domain	10	2	6	0	23
PF00907	T-box	T-box	17 (19)	8	22	0	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF02135	Zf-TAZ	TAZ finger	2 (3)	1 (2)	6 (7)	0	10 (15)
PF01285	TEA	TEA domain	4	1	1	1	0
PF02176	Zf-TRAF	TRAF-type zinc finger	6 (9)	1 (3)	1	0	2
PF00352	TBP	Transcription factor TFIID (or TATA-binding protein, TBP)	2 (4)	4 (8)	2 (4)	1 (2)	2 (4)
PF00567	TUDOR	TUDOR domain	9 (24)	9 (19)	4 (5)	0	2
PF00642	Zf-CCCH	Zinc finger C-x8-C-x5-C-x3-H type (and similar)	17 (22)	6 (8)	22 (42)	3 (5)	31 (46)
PF00096	Zf-C2H2**	ZInc finger, C2H2 type	564 (4500)	234 (771)	68 (155)	34 (56)	21 (24)
PF00097	Zf-C3HC4	Zinc finger, C3HC4 type (RING finger)	135 (137)	57	88 (89)	18	298 (304)
PF00098	Zf-CCHC	Zinc knuckle	9 (17)	6 (10)	17 (33)	7 (13)	68 (91)

(Tables 18 and 19). They include secreted hormones and growth factors, receptors, intracellular signaling molecules, and transcription factors.

Developmental signaling molecules that are enriched in the human genome include growth factors such as wnt, transforming growth factor–β (TGF-β), fibroblast growth factor (FGF), nerve growth factor, platelet derived growth factor (PDGF), and ephrins. These growth factors affect tissue differentiation and a wide range of cellular processes involving actin-cytoskeletal and nuclear regulation. The corresponding receptors of these developmental ligands are also expanded in humans. For example, our analysis suggests at least 8 human ephrin genes (2 in the fly, 4 in the worm) and 12 ephrin receptors (2 in the fly, 1 in the worm). In the wnt signaling pathway, we find 18 wnt family genes (6 in the fly, 5 in the worm) and 12 frizzled receptors (6 in the fly, 5 in the worm). The Groucho family of transcriptional corepressors downstream in the wnt pathway are even more markedly expanded, with 13 predicted members in humans (2 in the fly, 1 in the worm).

Extracellular adhesion molecules involved in signaling are expanded in the human genome (Tables 18 and 19). The interactions of several of these adhesion domains with extracellular matrix proteoglycans play a critical role in host defense, morphogenesis, and tissue repair (*131*). Consistent with the well-defined role of heparan sulfate proteoglycans in modulating these interactions (*132*), we observe an expansion of the heparin sulfate sulfotransferases in the human genome relative to worm and fly. These sulfotransferases modulate tissue differentiation (*133*). A similar expansion in humans is noted in structural proteins that constitute the actin-cytoskeletal architecture. Compared with the fly and worm, we observe an explosive expansion of the nebulin (35 domains per protein on average), aggrecan (12 domains per protein on average), and plectin (5 domains per protein on average) repeats in humans. These repeats are present in proteins involved in modulating the actin-cytoskeleton with predominant expression in neuronal, muscle, and vascular tissues.

Comparison across the five sequenced eukaryotic organisms revealed several expanded protein families and domains involved in cytoplasmic signal transduction (Table 18). In particular, signal transduction pathways playing roles in developmental regulation and acquired immunity were substantially enriched. There is a factor of 2 or greater expansion in humans in the Ras superfamily GTPases and the GTPase activator and GTP exchange factors associated with them. Although there are about the same number of tyrosine kinases in the human and *C. elegans* genomes, in humans there is an increase in the SH2, PTB, and ITAM domains involved in phosphotyrosine signal transduction. Further, there is a twofold expansion of phosphodiesterases in the human genome compared with either the worm or fly genomes.

The downstream effectors of the intracellular signaling molecules include the transcription factors that transduce developmental fates. Significant expansions are noted in the ligand-binding nuclear hormone receptor class of transcription factors compared with the fly genome, although not to the extent observed in the worm (Tables 18 and 19). Perhaps the most striking expansion in humans is in the C2H2 zinc finger transcription factors. Pfam detects a total of 4500 C2H2 zinc finger domains in 564 human proteins, compared with 771 in 234 fly proteins. This means that there has been a dramatic expansion not only in the number of C2H2 transcription factors, but also in the number of these DNA-binding motifs per transcription factor (8 on average in humans, 3.3 on average in the fly, and 2.3 on average in the worm). Furthermore, many of these transcription factors contain either the KRAB or SCAN domains, which are not found in the fly or worm genomes. These domains are involved in the oligomerization of transcription factors and increase the combinatorial partnering of these factors. In general, most of the transcription factor domains are shared between the three animal genomes, but the reassortment of these domains results in organism-specific transcription factor families. The domain combinations found in the human, fly, and worm include the BTB with C2H2 in the fly and humans, and homeodomains alone or in combination with Pou and LIM domains in all of the animal genomes. In plants, however, a different set of transcription factors are expanded, namely, the myb family, and a unique set that includes VP1 and AP2 domain–containing proteins (*134*). The yeast genome has a paucity of transcription factors compared with the multicellular eukaryotes, and its repertoire is limited to the expansion of the yeast-specific C6 transcription factor family involved in metabolic regulation.

While we have illustrated expansions in a subset of signal transduction molecules in the human genome compared with the other eukaryotic genomes, it should be noted that most of the protein domains are highly conserved. An interesting observation is that worms and humans have approximately the same number of both tyrosine kinases and serine/threonine kinases (Table 19). It is important to note, however, that these are merely counts of the catalytic domain; the proteins that contain these domains also display a wide repertoire of interaction domains with significant combinatorial diversity.

Hemostasis. Hemostasis is regulated primarily by plasma proteases of the coagulation pathway and by the interactions that occur between the vascular endothelium and platelets. Consistent with known anatomical and physiological differences between vertebrates and invertebrates, extracellular adhesion domains that constitute proteins integral to hemostasis are expanded in the human relative to the fly and worm (Tables 18 and 19). We note the evolution of domains such as FIMAC, FN1, FN2, and C1q that mediate surface interactions between hematopoeitic cells and the vascular matrix. In addition, there has been extensive recruitment of more-ancient animal-specific domains such as VWA, VWC, VWD, kringle, and FN3 into multidomain proteins that are involved in hemostatic regulation. Although we do not find a large expansion in the total number of serine proteases, this enzymatic domain has been specifically recruited into several of these multidomain proteins for proteolytic regulation in the vascular compartment. These are represented in plasma proteins that belong to the kinin and complement pathways. There is a

significant expansion in two families of matrix metalloproteases: ADAM (a disintegrin and metalloprotease) and MMPs (matrix metalloproteases) (Table 19). Proteolysis of extracellular matrix (ECM) proteins is critical for tissue development and for tissue degradation in diseases such as cancer, arthritis, Alzheimer's disease, and a variety of inflammatory conditions (*135*, *136*). ADAMs are a family of integral membrane proteins with a pivotal role in fibrinogenolysis and modulating interactions between hematopoietic components and the vascular matrix components. These proteins have been shown to cleave matrix proteins, and even signaling molecules: ADAM-17 converts tumor necrosis factor–α, and ADAM-10 has been implicated in the Notch signaling pathway (*135*). We have identified 19 members of the matrix metalloprotease family, and a total of 51 members of the ADAM and ADAM-TS families.

Apoptosis. Evolutionary conservation of some of the apoptotic pathway components across eukarya is consistent with its central role in developmental regulation and as a response to pathogens and stress signals. The signal transduction pathways involved in programmed cell death, or apoptosis, are mediated by interactions between well-characterized domains that include extracellular domains, adaptor (protein-protein interaction) domains, and those found in effector and regulatory enzymes (*137*). We enumerated the protein counts of central adaptor and effector enzyme domains that are found only in the apoptotic pathways to provide an estimate of divergence across eukarya and relative expansion in the human genome when compared with the fly and worm (Table 18). Adaptor domains found in proteins restricted only to apoptotic regulation such as the DED domains are vertebrate-specific, whereas others like BIR, CARD, and Bcl2 are represented in the fly and worm (although the number of Bcl2 family members in humans is significantly expanded). Although plants and yeast lack the caspases, caspase-like molecules, namely the para- and meta-caspases, have been reported in these organisms (*138*). Compared with other animal genomes, the human genome shows an expansion in the adaptor and effector domain–containing proteins involved in apoptosis, as well as in the proteases involved in the cascade such as the caspase and calpain families.

Expansions of other protein families. *Metabolic enzymes.* There are fewer cytochrome P450 genes in humans than in either the fly or worm. Lipoxygenases (six in humans), on the other hand, appear to be specific to the vertebrates and plants, whereas the lipoxygenase-activating proteins (four in humans) may be vertebrate-specific. Lipoxygenases are involved in arachidonic acid metabolism, and they and their activators have been implicated in diverse human pathology ranging from allergic responses to cancers. One of the most surprising human expansions, however, is in the number of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) genes (46 in humans, 3 in the fly, and 4 in the worm). There is, however, evidence for many retrotransposed GAPDH pseudogenes (*139*), which may account for this apparent expansion. However, it is interesting that GAPDH, long known as a conserved enzyme involved in basic metabolism found across all phyla from bacteria to humans, has recently been shown to have other functions. It has a second cat-

Table 19. Number of proteins assigned to selected Panther families or subfamilies in *H. sapiens* (H), *D. melanogaster* (F), *C. elegans* (W), *S. cerevisiae* (Y), and *A. thaliana* (A).

Panther family/subfamily*	H	F	W	Y	A
Neural structure, function, development					
Ependymin	1	0	0	0	0
Ion channels					
Acetylcholine receptor	17	12	56	0	0
Amiloride-sensitive/degenerin	11	24	27	0	0
CNG/EAG	22	9	9	0	30
IRK	16	3	3	0	0
ITP/ryanodine	10	2	4	0	0
Neurotransmitter-gated	61	51	59	0	19
P2X purinoceptor	10	0	0	0	0
TASK	12	12	48	1	5
Transient receptor	15	3	3	1	0
Voltage-gated Ca^{2+} alpha	22	4	8	2	2
Voltage-gated Ca^{2+} alpha-2	10	3	2	0	0
Voltage-gated Ca^{2+} beta	5	2	2	0	0
Voltage-gated Ca^{2+} gamma	1	0	0	0	0
Voltage-gated K^{+} alpha	33	5	11	0	0
Voltage-gated KQT	6	2	3	0	0
Voltage-gated Na^{+}	11	4	4	9	1
Myelin basic protein	1	0	0	0	0
Myelin PO	5	0	0	0	0
Myelin proteolipid	3	1	0	0	0
Myelin-oligodendrocyte glycoprotein	1	0	0	0	0
Neuropilin	2	0	0	0	0
Plexin	9	2	0	0	0
Semaphorin	22	6	2	0	0
Synaptotagmin	10	3	3	0	0
Immune response					
Defensin	3	0	0	0	0
Cytokine†	86	14	1	0	0
GCSF	1	0	0	0	0
GMCSF	1	0	0	0	0
Intercrine alpha	15	0	0	0	0
Intercrine beta	5	0	0	0	0
Inteferon	8	0	0	0	0
Interleukin	26	1	1	0	0
Leukemia inhibitory factor	1	0	0	0	0
MCSF	1	0	0	0	0
Peptidoglycan recognition protein	2	13	0	0	0
Pre-B cell enhancing factor	1	0	0	0	0
Small inducible cytokine A	14	0	0	0	0
Sl cytokine	2	0	0	0	0
TNF	9	0	0	0	0
Cytokine receptor†	62	1	0	0	0
Bradykinin/C-C chemokine receptor	7	0	0	0	0
Fl cytokine receptor	2	0	0	0	0
Interferon receptor	3	0	0	0	0
Interleukin receptor	32	0	0	0	0
Leukocyte tyrosine kinase receptor	3	0	0	0	0
MCSF receptor	1	0	0	0	0
TNF receptor	3	0	0	0	0
Immunoglobulin receptor†	59	0	0	0	0
T-cell receptor alpha chain	16	0	0	0	0
T-cell receptor beta chain	15	0	0	0	0
T-cell receptor gamma chain	1	0	0	0	0
T-cell receptor delta chain	1	0	0	0	0
Immunoglobulin FC receptor	8	0	0	0	0
Killer cell receptor	16	0	0	0	0
Polymeric-immunoglobulin receptor	4	0	0	0	0

alytic activity, as a uracil DNA glycosylase (*140*) and functions as a cell cycle regulator (*141*) and has even been implicated in apoptosis (*142*).

Translation. Another striking set of human expansions has occurred in certain families involved in the translational machinery. We identified 28 different ribosomal subunits that each have at least 10 copies in the genome; on average, for all ribosomal proteins there is about an 8- to 10-fold expansion in the number of genes relative to either the worm or fly. Retrotransposed pseudogenes may account for many of these expansions [see the discussion above and (*143*)]. Recent evidence suggests that a number of ribosomal proteins have secondary functions independent of their involvement in protein biosynthesis; for example, L13a and the related L7 subunits (36 copies in humans) have been shown to induce apoptosis (*144*).

There is also a four- to fivefold expansion in the elongation factor 1-alpha family (eEF1A; 56 human genes). Many of these expansions likely represent intronless paralogs that have presumably arisen from retrotransposition, and again there is evidence that many of these may be pseudogenes (*145*). However, a second form (eEF1A2) of this factor has been identied with tissue-specific expression in skeletal muscle and a complementary expression pattern to the ubiquitously expressed eEF1A (*146*).

Ribonucleoproteins. Alternative splicing results in multiple transcripts from a single gene, and can therefore generate additional diversity in an organism's protein complement. We have identified 269 genes for ribonucleoproteins. This represents over 2.5 times the number of ribonucleoprotein genes in the worm, two times that of the fly, and about the same as the 265 identified in the *Arabidopsis* genome. Whether the diversity of ribonucleoprotein genes in humans contributes to gene regulation at either the splicing or translational level is unknown.

Posttranslational modifications. In this set of processes, the most prominent expansion is the transglutaminases, calcium-dependent enzymes that catalyze the cross-linking of proteins in cellular processes such as hemostasis and apoptosis (*147*). The vitamin K–dependent gamma carboxylase gene product acts on the GLA domain (missing in the fly and worm) found in coagulation factors, osteocalcin, and matrix GLA protein (*148*). Tyrosylprotein sulfotransferases participate in the posttranslational modification of proteins involved in inflammation and hemostasis, including coagulation factors and chemokine receptors (*149*). Although there is no significant numerical increase in the counts for domains involved in nuclear protein modification, there are a number of domain arrangements in the predicted human proteins that are not found in the other currently sequenced genomes. These include the tandem association of two histone deacetylase domains in HD6 with a ubiquitin finger domain, a feature lacking in the fly genome. An additional example is the co-occurrence of important nuclear regulatory enzyme PARP (poly-ADP ribosyl transferase) domain fused to protein-interaction domains—BRCT and VWA in humans.

Concluding remarks. There are several possible explanations for the differences in phenotypic complexity observed in humans when compared to the fly and worm. Some of these relate to the prominent differences in the immune system, hemostasis, neuronal, vascular, and cytoskeletal complexity. The finding that the human genome contains fewer genes than previously predicted might be compensated for by combinatorial diversity generated at the levels of protein architecture, transcriptional and translational control, posttranslational modification of proteins, or posttranscriptional regulation. Extensive domain shuffling to increase or alter combinatorial diversity can provide an exponential

Table 19 (*Continued*)

Panther family/subfamily*	H	F	W	Y	A
MHC class I	22	0	0	0	0
MHC class II	20	0	0	0	0
Other immunoglobulin†	114	0	0	0	0
Toll receptor–related	10	6	0	0	0
Developmental and homeostatic regulators					
Signaling molecules†					
Calcitonin	3	0	0	0	0
Ephrin	8	2	4	0	0
FGF	24	1	1	0	0
Glucagon	4	0	0	0	0
Glycoprotein hormone beta chain	2	0	0	0	0
Insulin	1	0	0	0	0
Insulin-like hormone	3	0	0	0	0
Nerve growth factor	3	0	0	0	0
Neuregulin/heregulin	6	0	0	0	0
neuropeptide Y	4	0	0	0	0
PDGF	1	1	0	0	0
Relaxin	3	0	0	0	0
Stannocalcin	2	0	0	0	0
Thymopoeitin	2	0	1	0	0
Thyomosin beta	4	2	0	0	0
TGF-β	29	6	4	0	0
VEGF	4	0	0	0	0
Wnt	18	6	5	0	0
Receptors†					
Ephrin receptor	12	2	1	0	0
FGF receptor	4	4	0	0	0
Frizzled receptor	12	6	5	0	0
Parathyroid hormone receptor	2	0	0	0	0
VEGF receptor	5	0	0	0	0
BDNF/NT-3 nerve growth factor receptor	4	0	0	0	0
Kinases and phosphatases					
Dual-specificity protein phosphatase	29	8	10	4	11
S/T and dual-specificity protein kinase†	395	198	315	114	1102
S/T protein phosphatase	15	19	51	13	29
Y protein kinase†	106	47	100	5	16
Y protein phosphatase	56	22	95	5	6
Signal transduction					
ARF family	55	29	27	12	45
Cyclic nucleotide phosphodiesterase	25	8	6	1	0
G protein-coupled receptors†‡	616	146	284	0	1
G-protein alpha	27	10	22	2	5
G-protein beta	5	3	2	1	1
G-protein gamma	13	2	2	0	0
Ras superfamily	141	64	62	26	86
G-protein modulators†					
ARF GTPase-activating	20	8	9	5	15
Neurofibromin	7	2	0	2	0
Ras GTPase-activating	9	3	8	1	0
Tuberin	7	3	2	0	0
Vav proto-oncogene family	35	15	13	3	0

increase in the ability to mediate protein-protein interactions without dramatically increasing the absolute size of the protein complement (*150*). Evolution of apparently new (from the perspective of sequence analysis) protein domains and increasing regulatory complexity by domain accretion both quantitatively and qualitatively (recruitment of novel domains with preexisting ones) are two features that we observe in humans. Perhaps the best illustration of this trend is the C2H2 zinc finger–containing transcription factors, where we see expansion in the number of domains per protein, together with vertebrate-specific domains such as KRAB and SCAN. Recent reports on the prominent use of internal ribosomal entry sites in the human genome to regulate translation of specific classes of proteins suggests that this is an area that needs further research to identify the full extent of this process in the human genome (*151*). At the posttranslational level, although we provide examples of expansions of some protein families involved in these modifications, further experimental evidence is required to evaluate whether this is correlated with increased complexity in protein processing. Posttranscriptional processing and the extent of isoform generation in the human remain to be cataloged in their entirety. Given the conserved nature of the spliceosomal machinery, further analysis will be required to dissect regulation at this level.

8 Conclusions

8.1 The whole-genome sequencing approach versus BAC by BAC

Experience in applying the whole-genome shotgun sequencing approach to a diverse group of organisms with a wide range of genome sizes and repeat content allows us to assess its strengths and weaknesses. With the success of the method for a large number of microbial genomes, *Drosophila*, and now the human, there can be no doubt concerning the utility of this method. The large number of microbial genomes that have been sequenced by this method (*15*, *80*, *152*) demonstrate that megabase-sized genomes can be sequenced efficiently without any input other that the de novo mate-paired sequences. With more complex genomes like those of *Drosophila* or human, map information, in the form of well-ordered markers, has been critical for long-range ordering of scaffolds. For joining scaffolds into chromosomes, the quality of the map (in terms of the order of the markers) is more important than the number of markers per se. Although this mapping could have been performed concurrently with sequencing, the prior existence of mapping data was beneficial. During the sequencing of the *A. thaliana* genome, sequencing of individual BAC clones permitted extension of the se-

Table 19 (*Continued*)

Panther family/subfamily*	H	F	W	Y	A
Transcription factors/chromatin organization					
C2H2 zinc finger–containing†	607	232	79	28	8
COE	7	1	1	0	0
CREB	7	1	2	0	0
ETS-related	25	8	10	0	0
Forkhead-related	34	19	15	4	0
FOS	8	2	1	0	0
Groucho	13	2	1	0	0
Histone H1	5	0	1	0	0
Histone H2A	24	1	17	3	13
Histone H2B	21	1	17	2	12
Histone H3	28	2	24	2	16
Histone H4	9	1	16	1	8
Homeotic†	168	104	74	4	78
ABD-B	5	0	0	0	0
Bithoraxoid	1	8	1	0	0
Iroquois class	7	3	1	0	0
Distal-less	5	2	1	0	0
Engrailed	2	2	1	0	0
LIM-containing	17	8	3	0	0
MEIS/KNOX class	9	4	4	2	26
NK-3/NK-2 class	9	4	5	0	0
Paired box	38	28	23	0	2
Six	5	3	4	0	0
Leucine zipper	6	0	0	0	0
Nuclear hormone receptor†	59	25	183	1	4
Pou-related	15	5	4	1	0
Runt-related	3	4	2	0	0
ECM adhesion					
Cadherin	113	17	16	0	0
Claudin	20	0	0	0	0
Complement receptor-related	22	8	6	0	0
Connexin	14	0	0	0	0
Galectin	12	5	22	0	0
Glypican	13	2	1	0	0
ICAM	6	0	0	0	0
Integrin alpha	24	7	4	0	1
Integrin beta	9	2	2	0	0
LDL receptor family	26	19	20	0	2
Proteoglycans	22	9	7	0	5
Apoptosis					
Bcl-2	12	1	0	0	0
Calpain	22	4	11	1	3
Calpain inhibitor	4	0	0	0	1
Caspase	13	7	3	0	0
Hemostasis					
ADAM/ADAMTS	51	9	12	0	0
Fibronectin	3	0	0	0	0
Globin	10	2	3	0	3
Matrix metalloprotease	19	2	7	0	3
Serum amyloid A	4	0	0	0	0
Serum amyloid P (subfamily of Pentaxin)	2	0	0	0	0
Serum paraoxonase/arylesterase	4	0	3	0	0
Serum albumin	4	0	0	0	0
Transglutaminase	10	1	0	0	0
Other enzymes					
Cytochrome p450	60	89	83	3	256
GAPDH	46	3	4	3	8
Heparan sulfotransferase	11	4	2	0	0
Splicing and translation					
EF-1alpha	56	13	10	6	13
Ribonucleoproteins†	269	135	104	60	265
Ribosomal proteins†	812	111	80	117	256

*The table lists Panther families or subfamilies relevant to the text that either (i) are not specifically represented by Pfam (Table 18) or (ii) differ in counts from the corresponding Pfam models. †This class represents a number of different families in the same Panther molecular function subcategory. ‡This count includes only rhodopsin-class, secretin-class, and metabotropic glutamate-class GPCRs.

quence well into centromeric regions and allowed high-quality resolution of complex repeat regions. Likewise, in *Drosophila*, the BAC physical map was most useful in regions near the highly repetitive centromeres and telomeres. WGA has been found to deliver excellent-quality reconstructions of the unique regions of the genome. As the genome size, and more importantly the repetitive content, increases, the WGA approach delivers less of the repetitive sequence.

The cost and overall efficiency of clone-by-clone approaches makes them difficult to justify as a stand-alone strategy for future large-scale genome-sequencing projects. Specific applications of BAC-based or other clone mapping and sequencing strategies to resolve ambiguities in sequence assembly that cannot be efficiently resolved with computational approaches alone are clearly worth exploring. Hybrid approaches to whole-genome sequencing will only work if there is sufficient coverage in both the whole-genome shotgun phase and the BAC clone sequencing phase. Our experience with human genome assembly suggests that this will require at least 3× coverage of both whole-genome and BAC shotgun sequence data.

8.2 The low gene number in humans

We have sequenced and assembled ~95% of the euchromatic sequence of *H. sapiens* and used a new automated gene prediction method to produce a preliminary catalog of the human genes. This has provided a major surprise: We have found far fewer genes (26,000 to 38,000) than the earlier molecular predictions (50,000 to over 140,000). Whatever the reasons for this current disparity, only detailed annotation, comparative genomics (particularly using the *Mus musculus* genome), and careful molecular dissection of complex phenotypes will clarify this critical issue of the basic "parts list" of our genome. Certainly, the analysis is still incomplete and considerable refinement will occur in the years to come as the precise structure of each transcription unit is evaluated. A good place to start is to determine why the gene estimates derived from EST data are so discordant with our predictions. It is likely that the following contribute to an inflated gene number derived from ESTs: the variable lengths of 3′- and 5′-untranslated leaders and trailers; the little-understood vagaries of RNA processing that often leave intronic regions in an unspliced condition; the finding that nearly 40% of human genes are alternatively spliced (*153*); and finally, the unsolved technical problems in EST library construction where contamination from heterogeneous nuclear RNA and genomic DNA are not uncommon. Of course, it is possible that there are genes that remain unpredicted owing to the absence of EST or protein data to support them, although our use of mouse genome data for predicting genes should limit this number. As was true at the beginning of genome sequencing, ultimately it will be necessary to measure mRNA in specific cell types to demonstrate the presence of a gene.

J. B. S. Haldane speculated in 1937 that a population of organisms might have to pay a price for the number of genes it can possibly carry. He theorized that when the number of genes becomes too large, each zygote carries so many new deleterious mutations that the population simply cannot maintain itself. On the basis of this premise, and on the basis of available mutation rates and x-ray–induced mutations at specific loci, Muller, in 1967 (*154*), calculated that the mammalian genome would contain a maximum of not much more than 30,000 genes (*155*). An estimate of 30,000 gene loci for humans was also arrived at by Crow and Kimura (*156*). Muller's estimate for *D. melanogaster* was 10,000 genes, compared to 13,000 derived by annotation of the fly genome (*26*, *27*). These arguments for the theoretical maximum gene number were based on simplified ideas of genetic load—that all genes have a certain low rate of mutation to a deleterious state. However, it is clear that many mouse, fly, worm, and yeast knockout mutations lead to almost no discernible phenotypic perturbations.

The modest number of human genes means that we must look elsewhere for the mechanisms that generate the complexities inherent in human development and the sophisticated signaling systems that maintain homeostasis. There are a large number of ways in which the functions of individual genes and gene products are regulated. The degree of "openness" of chromatin structure and hence transcriptional activity is regulated by protein complexes that involve histone and DNA enzymatic modifications. We enumerate many of the proteins that are likely involved in nuclear regulation in Table 19. The location, timing, and quantity of transcription are intimately linked to nuclear signal transduction events as well as by the tissue-specific expression of many of these proteins. Equally important are regulatory DNA elements that include insulators, repeats, and endogenous viruses (*157*); methylation of CpG islands in imprinting (*158*); and promoter-enhancer and intronic regions that modulate transcription. The spliceosomal machinery consists of multisubunit proteins (Table 19) as well as structural and catalytic RNA elements (*159*) that regulate transcript structure through alternative start and termination sites and splicing. Hence, there is a need to study different classes of RNA molecules (*160*) such as small nucleolar RNAs, antisense riboregulator RNA, RNA involved in X-dosage compensation, and other structural RNAs to appreciate their precise role in regulating gene expression. The phenomenon of RNA editing in which coding changes occur directly at the level of mRNA is of clinical and biological relevance (*161*). Finally, examples of translational control include internal ribosomal entry sites that are found in proteins involved in cell cycle regulation and apoptosis (*162*). At the protein level, minor alterations in the nature of protein-protein interactions, protein modifications, and localization can have dramatic effects on cellular physiology (*163*). This dynamic system therefore has many ways to modulate activity, which suggests that definition of complex systems by analysis of single genes is unlikely to be entirely successful.

In situ studies have shown that the human genome is asymmetrically populated with G+C content, CpG islands, and genes (*68*). However, the genes are not distributed quite as unequally as had been predicted (Table 9) (*69*). The most G+C-rich fraction of the genome, H3 isochores, constitute more of the genome than previously thought (about 9%), and are the most gene-dense fraction, but contain only 25% of the genes, rather than the predicted ~40%. The low G+C L isochores make up 65% of the genome, and 48% of the genes. This inhomogeneity, the net result of millions of years of mammalian gene duplication, has been described as the "desertification" of the vertebrate genome (*71*). Why are there clustered regions of high and low gene density, and are these accidents of history or driven by selection and evolution? If these deserts are dispensable, it ought to be possible to find mammalian genomes that are far smaller in size than the human genome. Indeed, many species of bats have genome sizes that are much smaller than that of humans; for example, *Miniopterus*, a species of Italian bat, has a genome size that is only 50% that of humans (*164*). Similarly, *Muntiacus*, a species of Asian barking deer, has a genome size that is ~70% that of humans.

8.3 Human DNA sequence variation and its distribution across the genome

This is the first eukaryotic genome in which a nearly uniform ascertainment of polymorphism has been completed. Although we have identified and mapped more than 3 million SNPs, this by no means implies that the task of finding and cataloging SNPs is complete. These represent only a fraction of the SNPs present in the human population as a whole. Nevertheless, this first glimpse at genome-wide variation has revealed strong inhomogeneities in the distribution of SNPs across the genome. Polymorphism in DNA carries with it a snapshot of the past operation of population genetic forces, including mutation, migration, selection, and genetic drift. The availability of a dense array of SNPs will allow questions related to each of these factors to be addressed on a genome-wide basis. SNP studies can establish the range of haplo-

types present in subjects of different ethnogeographic origins, providing insights into population history and migration patterns. Although such studies have suggested that modern human lineages derive from Africa, many important questions regarding human origins remain unanswered, and more analyses using detailed SNP maps will be needed to settle these controversies. In addition to providing evidence for population expansions, migration, and admixture, SNPs can serve as markers for the extent of evolutionary constraint acting on particular genes. The correlation between patterns of interspecies and interspecies genetic variation may prove to be especially informative to identify sites of reduced genetic diversity that may mark loci where sequence variations are not tolerated.

The remarkable heterogeneity in SNP density implies that there are a variety of forces acting on polymorphism—sparse regions may have lower SNP density because the mutation rate is lower, because most of those regions have a lower fraction of mutations that are tolerated, or because recent strong selection in favor of a newly arisen allele "swept" the linked variation out of the population (*165*). The effect of random genetic drift also varies widely across the genome. The nonrecombining portion of the Y chromosome faces the strongest pressure from random drift because there are roughly one-quarter as many Y chromosomes in the population as there are autosomal chromosomes, and the level of polymorphism on the Y is correspondingly less. Similarly, the X chromosome has a smaller effective population size than the autosomes, and its nucleotide diversity is also reduced. But even across a single autosome, the effective population size can vary because the density of deleterious mutations may vary. Regions of high density of deleterious mutations will see a greater rate of elimination by selection, and the effective population size will be smaller (*166*). As a result, the density of even completely neutral SNPs will be lower in such regions. There is a large literature on the association between SNP density and local recombination rates in *Drosophila*, and it remains an important task to assess the strength of this association in the human genome, because of its impact on the design of local SNP densities for disease-association studies. It also remains an important task to validate SNPs on a genomic scale in order to assess the degree of heterogeneity among geographic and ethnic populations.

8.4 Genome complexity

We will soon be in a position to move away from the cataloging of individual components of the system, and beyond the simplistic notions of "this binds to that, which then docks on this, and then the complex moves there. . . ." (*167*) to the exciting area of network perturbations, nonlinear responses and thresholds, and their pivotal role in human diseases.

The enumeration of other "parts lists" reveals that in organisms with complex nervous systems, neither gene number, neuron number, nor number of cell types correlates in any meaningful manner with even simplistic measures of structural or behavioral complexity. Nor would they be expected to; this is the realm of nonlinearities and epigenesis (*168*). The 520 million neurons of the common octopus exceed the neuronal number in the brain of a mouse by an order of magnitude. It is apparent from a comparison of genomic data on the mouse and human, and from comparative mammalian neuroanatomy (*169*), that the morphological and behavioral diversity found in mammals is underpinned by a similar gene repertoire and similar neuroanatomies. For example, when one compares a pygmy marmoset (which is only 4 inches tall and weighs about 6 ounces) to a chimpanzee, the brain volume of this minute primate is found to be only about 1.5 cm^3, two orders of magnitude less than that of a chimp and three orders less than that of humans. Yet the neuroanatomies of all three brains are strikingly similar, and the behavioral characteristics of the pygmy marmoset are little different from those of chimpanzees. Between humans and chimpanzees, the gene number, gene structures and functions, chromosomal and genomic organizations, and cell types and neuroanatomies are almost indistinguishable, yet the developmental modifications that predisposed human lineages to cortical expansion and development of the larynx, giving rise to language, culminated in a massive singularity that by even the simplest of criteria made humans more complex in a behavioral sense.

Simple examination of the number of neurons, cell types, or genes or of the genome size does not alone account for the differences in complexity that we observe. Rather, it is the interactions within and among these sets that result in such great variation. In addition, it is possible that there are "special cases" of regulatory gene networks that have a disproportionate effect on the overall system. We have presented several examples of "regulatory genes" that are significantly increased in the human genome compared with the fly and worm. These include extracellular ligands and their cognate receptors (e.g., wnt, frizzled, TGF-β, ephrins, and connexins), as well as nuclear regulators (e.g., the KRAB and homeodomain transcription factor families), where a few proteins control broad developmental processes. The answers to these "complexities" perhaps lie in these expanded gene families and differences in the regulatory control of ancient genes, proteins, pathways, and cells.

8.5 Beyond single components

While few would disagree with the intuitive conclusion that Einstein's brain was more complex than that of *Drosophila*, closer comparisons such as whether the set of predicted human proteins is more complex than the protein set of *Drosophila*, and if so, to what degree, are not straightforward, since protein, protein domain, or protein-protein interaction measures do not capture context-dependent interactions that underpin the dynamics underlying phenotype.

Currently, there are more than 30 different mathematical descriptions of complexity (*170*). However, we have yet to understand the mathematical dependency relating the number of genes with organism complexity. One pragmatic approach to the analysis of biological systems, which are composed of nonidentical elements (proteins, protein complexes, interacting cell types, and interacting neuronal populations), is through graph theory (*171*). The elements of the system can be represented by the vertices of complex topographies, with the edges representing the interactions between them. Examination of large networks reveals that they can self-organize, but more important, they can be particularly robust. This robustness is not due to redundancy, but is a property of inhomogeneously wired networks. The error tolerance of such networks comes with a price; they are vulnerable to the selection or removal of a few nodes that contribute disproportionately to network stability. Gene knockouts provide an illustration. Some knockouts may have minor effects, whereas others have catastrophic effects on the system. In the case of vimentin, a supposedly critical component of the cytoplasmic intermediate filament network of mammals, the knockout of the gene in mice reveals them to be reproductively normal, with no obvious phenotypic effects (*172*), and yet the usually conspicuous vimentin network is completely absent. On the other hand, ~30% of knockouts in *Drosophila* and mice correspond to critical nodes whose reduction in gene product, or total elimination, causes the network to crash most of the time, although even in some of these cases, phenotypic normalcy ensues, given the appropriate genetic background. Thus, there are no "good" genes or "bad" genes, but only networks that exist at various levels and at different connectivities, and at different states of sensitivity to perturbation. Sophisticated mathematical analysis needs to be constantly evaluated against hard biological data sets that specifically address network dynamics. Nowhere is this more critical than in attempts to come to grips with "complexity," particularly because deconvoluting and correcting complex networks that have undergone perturbation, and have resulted in human diseases, is the greatest significant challenge now facing us.

It has been predicted for the last 15 years that complete sequencing of the human ge-

nome would open up new strategies for human biological research and would have a major impact on medicine, and through medicine and public health, on society. Effects on biomedical research are already being felt. This assembly of the human genome sequence is but a first, hesitant step on a long and exciting journey toward understanding the role of the genome in human biology. It has been possible only because of innovations in instrumentation and software that have allowed automation of almost every step of the process from DNA preparation to annotation. The next steps are clear: We must define the complexity that ensues when this relatively modest set of about 30,000 genes is expressed. The sequence provides the framework upon which all the genetics, biochemistry, physiology, and ultimately phenotype depend. It provides the boundaries for scientific inquiry. The sequence is only the first level of understanding of the genome. All genes and their control elements must be identified; their functions, in concert as well as in isolation, defined; their sequence variation worldwide described; and the relation between genome variation and specific phenotypic characteristics determined. Now we know what we have to explain.

Another paramount challenge awaits: public discussion of this information and its potential for improvement of personal health. Many diverse sources of data have shown that any two individuals are more than 99.9% identical in sequence, which means that all the glorious differences among individuals in our species that can be attributed to genes falls in a mere 0.1% of the sequence. There are two fallacies to be avoided: determinism, the idea that all characteristics of the person are "hard-wired" by the genome; and reductionism, the view that with complete knowledge of the human genome sequence, it is only a matter of time before our understanding of gene functions and interactions will provide a complete causal description of human variability. The real challenge of human biology, beyond the task of finding out how genes orchestrate the construction and maintenance of the miraculous mechanism of our bodies, will lie ahead as we seek to explain how our minds have come to organize thoughts sufficiently well to investigate our own existence.

References and Notes

1. R. L. Sinsheimer, *Genomics* **5**, 954 (1989); U.S. Department of Energy, Office of Health and Environmental Research, *Sequencing the Human Genome: Summary Report of the Santa Fe Workshop*, Santa Fe, NM, 3 to 4 March 1986 (Los Alamos National Laboratory, Los Alamos, NM, 1986).
2. R. Cook-Deegan, *The Gene Wars: Science, Politics, and the Human Genome* (Norton, New York, 1996).
3. F. Sanger *et al.*, *Nature* **265**, 687 (1977).
4. P. H. Seeburg *et al.*, *Trans. Assoc. Am. Physicians* **90**, 109 (1977).
5. E. C. Strauss, J. A. Kobori, G. Siu, L. E. Hood, *Anal. Biochem.* **154**, 353 (1986).
6. J. Gocayne *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 8296 (1987).
7. A. Martin-Gallardo *et al.*, *DNA Sequence* **3**, 237 (1992); W. R. McCombie *et al.*, *Nature Genet.* **1**, 348 (1992); M. A. Jensen *et al.*, *DNA Sequence* **1**, 233 (1991).
8. M. D. Adams *et al.*, *Science* **252**, 1651 (1991).
9. M. D. Adams *et al.*, *Nature* **355**, 632 (1992); M. D. Adams, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* **4**, 256 (1993); M. D. Adams, M. B. Soares, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* **4**, 373 (1993); M. H. Polymeropoulos *et al.*, *Nature Genet.* **4**, 381 (1993); M. Marra *et al.*, *Nature Genet.* **21**, 191 (1999).
10. M. D. Adams *et al.*, *Nature* **377**, 3 (1995); O. White *et al.*, *Nucleic Acids Res.* **21**, 3829 (1993).
11. F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, *J. Mol. Biol.* **162**, 729 (1982).
12. B. W. J. Mahy, J. J. Esposito, J. C. Venter, *Am. Soc. Microbiol. News* **57**, 577 (1991).
13. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995).
14. C. M. Fraser *et al.*, *Science* **270**, 397 (1995).
15. C. J. Bult *et al.*, *Science* **273**, 1058 (1996); J. F. Tomb *et al.*, *Nature* **388**, 539 (1997); H. P. Klenk *et al.*, *Nature* **390**, 364 (1997).
16. J. C. Venter, H. O. Smith, L. Hood, *Nature* **381**, 364 (1996).
17. H. Schmitt *et al.*, *Genomics* **33**, 9 (1996).
18. S. Zhao *et al.*, *Genomics* **63**, 321 (2000).
19. X. Lin *et al.*, *Nature* **402**, 761 (1999).
20. J. L. Weber, E. W. Myers, *Genome Res.* **7**, 401 (1997).
21. P. Green, *Genome Res.* **7**, 410 (1997).
22. E. Pennisi, *Science* **280**, 1185 (1998).
23. J. C. Venter *et al.*, *Science* **280**, 1540 (1998).
24. M. D. Adams *et al.*, *Nature* **368**, 474 (1994).
25. E. Marshall, E. Pennisi, *Science* **280**, 994 (1998).
26. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
27. G. M. Rubin *et al.*, *Science* **287**, 2204 (2000).
28. E. W. Myers *et al.*, *Science* **287**, 2196 (2000).
29. F. S. Collins *et al.*, *Science* **282**, 682 (1998).
30. International Human Genome Sequencing Consortium (2001), *Nature* **409**, 860 (2001).
31. Institutional review board: P. Calabresi (chairman), H. P. Freeman, C. McCarthy, A. L. Caplan, G. D. Rogell, J. Karp, M. K. Evans, B. Margus, C. L. Carter, R. A. Millman, S. Broder.
32. Eligibility criteria for participation in the study were as follows: prospective donors had to be 21 years of age or older, not pregnant, and capable of giving an informed consent. Donors were asked to self-define their ethnic backgrounds. Standard blood bank screens (screening for HIV, hepatitis viruses, and so forth) were performed on all samples at the clinical laboratory prior to DNA extraction in the Celera laboratory. All samples that tested positive for transmissible viruses were ineligible and were discarded. Karyotype analysis was performed on peripheral blood lymphocytes from all samples selected for sequencing; all were normal. A two-staged consent process for prospective donors was employed. The first stage of the consent process provided information about the genome project, procedures, and risks and benefits of participating. The second stage of the consent process involved answering follow-up questions and signing consent forms, and was conducted about 48 hours after the first.
33. DNA was isolated from blood (*173*) or sperm. For sperm, a washed pellet (100 μl) was lysed in a suspension (1 ml) containing 0.1 M NaCl, 10 mM tris-Cl–20 mM EDTA (pH 8), 1% SDS, 1 mg proteinase K, and 10 mM dithiothreitol for 1 hour at 37°C. The lysate was extracted with aqueous phenol and with phenol/chloroform. The DNA was ethanol precipitated and dissolved in 1 ml TE buffer. To make genomic libraries, DNA was randomly sheared, end-polished with consecutive BAL31 nuclease and T4 DNA polymerase treatments, and size-selected by electrophoresis on 1% low-melting-point agarose. After ligation to Bst XI adapters (Invitrogen, catalog no. N408-18), DNA was purified by three rounds of gel electrophoresis to remove excess adapters, and the fragments, now with 3′-CACA overhangs, were inserted into Bst XI-linearized plasmid vector with 3′-TGTG overhangs. Libraries with three different average sizes of inserts were constructed: 2, 10, and 50 kbp. The 2-kbp fragments were cloned in a high-copy pUC18 derivative. The 10- and 50-kbp fragments were cloned in a medium-copy pBR322 derivative. The 2- and 10-kbp libraries yielded uniform-sized large colonies on plating. However, the 50-kbp libraries produced many small colonies and inserts were unstable. To remedy this, the 50-kbp libraries were digested with Bgl II, which does not cleave the vector, but generally cleaved several times within the 50-kbp insert. A 1264-bp Bam HI kanamycin resistance cassette (purified from pUCK4; Amersham Pharmacia, catalog no. 27-4958-01) was added and ligation was carried out at 37°C in the continual presence of Bgl II. As Bgl II–Bgl II ligations occurred, they were continually cleaved, whereas Bam HI–Bgl II ligations were not cleaved. A high yield of internally deleted circular library molecules was obtained in which the residual insert ends were separated by the kanamycin cassette DNA. The internally deleted libraries, when plated on agar containing ampicillin (50 μg/ml), carbenicillin (50 μg/ml), and kanamycin (15 μg/ml), produced relatively uniform large colonies. The resulting clones could be prepared for sequencing using the same procedures as clones from the 10-kbp libraries.
34. Transformed cells were plated on agar diffusion plates prepared with a fresh top layer containing no antibiotic poured on top of a previously set bottom layer containing excess antibiotic, to achieve the correct final concentration. This method of plating permitted the cells to develop antibiotic resistance before being exposed to antibiotic without the potential clone bias that can be introduced through liquid outgrowth protocols. After colonies had grown, QBot (Genetix, UK) automated colony-picking robots were used to pick colonies meeting stringent size and shape criteria and to inoculate 384-well microtiter plates containing liquid growth medium. Liquid cultures were incubated overnight, with shaking, and were scored for growth before passing to template preparation. Template DNA was extracted from liquid bacterial culture using a procedure based upon the alkaline lysis miniprep method (*173*) adapted for high throughput processing in 384-well microtiter plates. Bacterial cells were lysed; cell debris was removed by centrifugation; and plasmid DNA was recovered by isopropanol precipitation and resuspended in 10 mM tris-HCl buffer. Reagent dispensing operations were accomplished using Titertek MAP 8 liquid dispensing systems. Plate-to-plate liquid transfers were performed using Tomtec Quadra 384 Model 320 pipetting robots. All plates were tracked throughout processing by unique plate barcodes. Mated sequencing reads from opposite ends of each clone insert were obtained by preparing two 384-well cycle sequencing reaction plates from each plate of plasmid template DNA using ABI-PRISM BigDye Terminator chemistry (Applied Biosystems) and standard M13 forward and reverse primers. Sequencing reactions were prepared using the Tomtec Quadra 384-320 pipetting robot. Parent-child plate relationships and, by extension, forward-reverse sequence mate pairs were established by automated plate barcode reading by the onboard barcode reader and were recorded by direct LIMS communication. Sequencing reaction products were purified by alcohol precipitation and were dried, sealed, and stored at 4°C in the dark until needed for sequencing, at which time the reaction products were resuspended in deionized formamide and sealed immediately to prevent degradation. All sequence data were generated using a single sequencing platform, the ABI PRISM 3700 DNA Analyzer. Sample sheets were created at load time using a Java-based application that facilitates barcode scanning of the sequencing plate barcode, retrieves sample information from the central LIMS, and reserves unique trace identifiers. The application permitted a single sample sheet file in the linking directory and deleted previously created sample sheet files immediately upon scanning of a

sample plate barcode, thus enhancing sample sheet-to-plate associations.
35. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977); J. M. Prober *et al.*, *Science* **238**, 336 (1987).
36. Celera's computing environment is based on Compaq Computer Corporation's Alpha system technology running the Tru64 Unix operating system. Celera uses these Alphas as Data Servers and as nodes in a Virtual Compute Farm, all of which are connected to a fully switched network operating at Fast Ethernet speed (for the VCF) and gigabit Ethernet speed (for data servers). Load balancing and scheduling software manages the submission and execution of jobs, based on central processing unit (CPU) speed, memory requirements, and priority. The Virtual Compute Farm is composed of 440 Alpha CPUs, which includes model EV6 running at a clock speed of 400 MHz and EV67 running at 667 MHz. Available memory on these systems ranges from 2 GB to 8 GB. The VCF is used to manage trace file processing, and annotation. Genome assembly was performed on a GS 160 running 16 EV67s (667 MHz) and 64 GB of memory, and 10 ES40s running 4 EV6s (500 MHz) and 32 GB of memory. A total of 100 terabytes of physical disk storage was included in a Storage Area Network that was available to systems across the environment. To ensure high availability, file and database servers were configured as 4-node Alpha TruClusters, so that services would fail over in the event of hardware or software failure. Data availability was further enhanced by using hardware- and software-based disk mirroring (RAID-0), disk striping (RAID-1), and disk striping with parity (RAID-5).
37. Trace processing generates quality values for base calls by means of Paracel's TraceTuner, trims sequence reads according to quality values, trims vector and adapter sequence from high-quality reads, and screens sequences for contaminants. Similar in design and algorithm to the phred program (*174*), TraceTuner reports quality values that reflect the log-odds score of each base being correct. Read quality was evaluated in 50-bp windows, each read being trimmed to include only those consecutive 50-bp segments with a minimum mean accuracy of 97%. End windows (both ends of the trace) of 1, 5, 10, 25, and 50 bases were trimmed to a minimum mean accuracy of 98%. Every read was further checked for vector and contaminant matches of 50 bp or more, and if found, the read was removed from consideration. Finally, any match to the 5′ vector splice junction in the initial part of a read was removed.
38. National Center for Biotechnology Information (NCBI); available at www.ncbi.nlm.nih.gov/.
39. NCBI; available at www.ncbi.nlm.nih.gov/HTGS/.
40. All bactigs over 3 kbp were examined for coverage by Celera mate pairs. An interval of a bactig was deemed an assembly error where there were no mate pairs spanning the interval and at least two reads that should have their mate on the other side of the interval but did not. In other words, there was no mate pair evidence supporting a join in the breakpoint interval and at least two mate pairs contradicting the join. By this criterion, we detected and broke apart bactigs at 13,037 locations, or equivalently, we found 2.13% of the bactigs to be misassembled.
41. We considered a BAC entry to be chimeric if, by the Lander-Waterman statistic (*175*), the odds were 0.99 or more that the assembly we produced was inconsistent with the sequence coming from a single source. By this criterion, 714 or 2.2% of BAC entries were deemed chimeric.
42. G. Myers, S. Selznick, Z. Zhang, W. Miller, *J. Comput. Biol.* **3**, 563 (1996).
43. E. W. Myers, J. L. Weber, in *Computational Methods in Genome Research*, S. Suhai, Ed. (Plenum, New York, 1996), pp. 73–89.
44. P. Deloukas *et al.*, *Science* **282**, 744 (1998).
45. M. A. Marra *et al.*, *Genome Res.* **7**, 1072 (1997).
46. J. Zhang *et al.*, data not shown.
47. Shredded bactigs were located on long CSA scaffolds (>500 kbp) and the distribution of these fragments on the scaffolds was analyzed. If the spread of these fragments was greater than four times the reported BAC length, the BAC was considered to be chimeric. In addition, if >20% of bactigs of a given BAC were found on a different scaffolds that were not adjacent in map position, then the BAC was also considered as chimeric. The total chimeric BACs divided by the number of BACs used for CSA gave the minimal estimate of chimerism rate.
48. M. Hattori *et al.*, *Nature* **405**, 311 (2000).
49. I. Dunham *et al.*, *Nature* **402**, 489 (1999).
50. A. B. Carvalho, B. P. Lazzaro, A. G. Clark, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 13239 (2000).
51. The International RH Mapping Consortium, available at www.ncbi.nlm.nih.gov/genemap99/.
52. See http://ftp.genome.washington.edu/RM/RepeatMasker.html.
53. G. D. Schuler, *Trends Biotechnol.* **16**, 456 (1998).
54. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
55a. M. Olivier *et al.*, *Science* **291**, 1298 (2001).
55b. See http://genome.ucsc.edu/.
56. N. Chaudhari, W. E. Hahn, *Science* **220**, 924 (1983); R. J. Milner, J. G. Sutcliffe, *Nucleic Acids Res.* **11**, 5497 (1983).
57. D. Dickson, *Nature* **401**, 311 (1999).
58. B. Ewing, P. Green, *Nature Genet.* **25**, 232 (2000).
59. H. Roest Crollius *et al.*, *Nature Genet.* **25**, 235 (2000).
60. M. Yandell, in preparation.
61. K. D. Pruitt, K. S. Katz, H. Sicotte, D. R. Maglott, *Trends Genet.* **16**, 44 (2000).
62. Scaffolds containing greater than 10 kbp of sequence were analyzed for features of biological importance through a series of computational steps, and the results were stored in a relational database. For scaffolds greater than one megabase, the sequence was cut into single megabase pieces before computational analysis. All sequence was masked for complex repeats using Repeatmasker (*52*) before gene finding or homology-based analysis. The computational pipeline required ~7 hours of CPU time per megabase, including repeat masking, or a total compute time of about 20,000 CPU hours. Protein searches were performed against the nonredundant protein database available at the NCBI. Nucleotide searches were performed against human, mouse, and rat Celera Gene Indices (assemblies of cDNA and EST sequences), mouse genomic DNA reads generated at Celera (3×), the Ensembl gene database available at the European Bioinformatics Institute (EBI), human and rodent (mouse and rat) EST data sets parsed from the dbEST database (NCBI), and a curated subset of the RefSeq experimental mRNA database (NCBI). Initial searches were performed on repeat-masked sequence with BLAST 2.0 (*54*) optimized for the Compaq Alpha computeserver and an effective database size of 3×10^9 for BLASTN searches and 1×10^9 for BLASTX searches. Additional processing of each query-subject pair was performed to improve the alignments. All protein BLAST results having an expectation score of $<1 \times 10^{-4}$, human nucleotide BLAST results having an expectation score of $<1 \times 10^{-8}$ with >94% identity, and rodent nucleotide BLAST results having an expectation score of $<1 \times 10^8$ with >80% identity were then examined on the basis of their high-scoring pair (HSP) coordinates on the scaffold to remove redundant hits, retaining hits that supported possible alternative splicing. For BLASTX searches, analysis was performed separately for selected model organisms (yeast, mouse, human, *C. elegans*, and *D. melanogaster*) so as not to exclude HSPs from these organisms that support the same gene structure. Sequences producing BLAST hits judged to be informative, nonredundant, and sufficiently similar to the scaffold sequence were then realigned to the genomic sequence with Sim4 for ESTs, and with Lap for proteins. Because both of these algorithms take splicing into account, the resulting alignments usually give a better representation of intron-exon boundaries than standard BLAST analyses and thus facilitate further annotation (both machine and human). In addition to the homology-based analysis described above, three ab initio gene prediction programs were used (*63*).
63. E. C. Uberbacher, Y. Xu, R. J. Mural, *Methods Enzymol.* **266**, 259 (1996); C. Burge, S. Karlin, *J. Mol. Biol.* **268**, 78 (1997); R. J. Mural, *Methods Enzymol.* **303**, 77 (1999); A. A. Salamov, V. V. Solovyev, *Genome Res.* **10**, 516 (2000); Floreal *et al.*, *Genome Res.* **8**, 967 (1998).
64. G. L. Miklos, B. John, *Am. J. Hum. Genet.* **31**, 264 (1979); U. Francke, *Cytogenet. Cell Genet.* **65**, 206 (1994).
65. P. E. Warburton, H. F. Willard, in *Human Genome Evolution*, M. S. Jackson, T. Strachan, G. Dover, Eds. (BIOS Scientific, Oxford, 1996), pp. 121–145.
66. J. E. Horvath, S. Schwartz, E. E. Eichler, *Genome Res.* **10**, 839 (2000).
67. W. A. Bickmore, A. T. Sumner, *Trends Genet.* **5**, 144 (1989).
68. G. P. Holmquist, *Am. J. Hum. Genet.* **51**, 17 (1992).
69. G. Bernardi, *Gene* **241**, 3 (2000).
70. S. Zoubak, O. Clay, G. Bernardi, *Gene* **174**, 95 (1996).
71. S. Ohno, *Trends Genet.* **1**, 160 (1985).
72. K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, J. L. Weber, *Am. J. Hum. Genet.* **63**, 861 (1998).
73. M. J. McEachern, A. Krauskopf, E. H. Blackburn, *Annu. Rev. Genet.* **34**, 331 (2000).
74. A. Bird, *Trends Genet.* **3**, 342 (1987).
75. M. Gardiner-Garden, M. Frommer, *J. Mol. Biol.* **196**, 261 (1987).
76. F. Larsen, G. Gundersen, R. Lopez, H. Prydz, *Genomics* **13**, 1095 (1992).
77. S. H. Cross, A. Bird, *Curr. Opin. Genet. Dev.* **5**, 309 (1995).
78. J. Peters, *Genome Biol.* **1**, reviews1028.1 (2000) (http://genomebiology.com/2000/1/5/reviews/1028).
79. C. Grunau, W. Hindermann, A. Rosenthal, *Hum. Mol. Genet.* **9**, 2651 (2000).
80. F. Antequera, A. Bird, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11995 (1993).
81. S. H. Cross *et al.*, *Mamm. Genome* **11**, 373 (2000).
82. D. Slavov *et al.*, *Gene* **247**, 215 (2000).
83. A. F. Smit, A. D. Riggs, *Nucleic Acids Res.* **23**, 98 (1995).
84. D. J. Elliott *et al.*, *Hum. Mol. Genet.* **9**, 2117 (2000).
85. A. V. Makeyev, A. N. Chkheidze, S. A. Lievhaber, *J. Biol. Chem.* **274**, 24849 (1999).
86. Y. Pan, W. K. Decker, A. H. H. M. Huq, W. J. Craigen, *Genomics* **59**, 282 (1999).
87. P. Nouvel, *Genetica* **93**, 191 (1994).
88. I. Goncalves, L. Duret, D. Mouchiroud, *Genome Res.* **10**, 672 (2000).
89. Lek first compares all proteins in the proteome to one another. Next, the resulting BLAST reports are parsed, and a graph is created wherein each protein constitutes a node; any hit between two proteins with an expectation beneath a user-specified threshold constitutes an edge. Lek then uses this graph to compute a similarity between each protein pair *ij* in the context of the graph as a whole by simply dividing the number of BLAST hits shared in common between the two proteins by the total number of proteins hit by *i* and *j*. This simple metric has several interesting properties. First, because the similarity metric takes into account both the similarity and the differences between the two sequences at the level of BLAST hits, the metric respects the multidomain nature of protein space. Two multidomain proteins, for instance, each containing domains A and B, will have a greater pairwise similarity to each other than either one will have to a protein containing only A or B domains, so long as A-B–containing multidomain proteins are less frequent in the proteome than are single-domain proteins containing A or B domains. A second interesting property of this similarity metric is that it can be used to produce a similarity matrix for the proteome as a whole without having to first produce a multiple alignment for each protein family, an error-prone and very time-consuming process. Finally, the metric does not require that either sequence have significant homology to the other in order to have a defined similarity to each other, only that they

share at least one significant BLAST hit in common. This is an especially interesting property of the metric, because it allows the rapid recovery of protein families from the proteome for which no multiple alignment is possible, thus providing a computational basis for the extension of protein homology searches beyond those of current HMM- and profile-based search methods. Once the whole-proteome similarity matrix has been calculated, Lek first partitions the proteome into single-linkage clusters (*27*) on the basis of one or more shared BLAST hits between two sequences. Next, these single-linkage clusters are further partitioned into subclusters, each member of which shares a user-specified pairwise similarity with the other members of the cluster, as described above. For the purposes of this publication, we have focused on the analysis of single-linkage clusters and what we have termed "complete clusters," e.g., those subclusters for which every member has a similarity metric of 1 to every other member of the subcluster. We believe that the single-linkage and complete clusters are of special interest, in part, because they allow us to estimate and to compare sizes of core protein sets in a rigorous manner. The rationale for this is as follows: if one imagines for a moment a perfect clustering algorithm capable of perfectly partitioning one or more perfectly annotated protein sets into protein families, it is reasonable to assume that the number of clusters will always be greater than, or equal to, the number of single-linkage clusters, because single-linkage clustering is a maximally agglomerative clustering method. Thus, if there exists a single protein in the predicted protein set containing domains A and B, then it will be clustered by single linkage together with all single-domain proteins containing domains A or B. Likewise, for a predicted protein set containing a single multidomain protein, the number of real clusters must always be less than or equal to the number of complete clusters, because it is impossible to place a unique multidomain protein into a complete cluster. Thus, the single-linkage and complete clusters plus singletons should comprise a lower and upper bound of sizes of core protein sets, respectively, allowing us to compare the relative size and complexity of different organisms' predicted protein set.

90. T. F. Smith, M. S. Waterman, *J. Mol. Biol.* **147**, 195 (1981).
91. A. L. Delcher *et al.*, *Nucleic Acids Res.* **27**, 2369 (1999).
92. *Arabidopsis* Genome Initiative, *Nature* **408**, 796 (2000).

93. The probability that a contiguous set of proteins is the result of a segmental duplication can be estimated approximately as follows. Given that protein A and B occur on one chromosome, and that A′ and B′ (paralogs of A and B) also exist in the genome, the probability that B′ occurs immediately after A′ is 1/*N*, where *N* is the number of proteins in the set (for this analysis, $N = 26{,}588$). Allowing for B′ to occur as any of the next J-1 proteins [leaving a gap between A′ and B′ increases the probability to $(J - 1)/N$; allowing B′A′ or A′B′ gives a probability of $2(J - 1)/N$]. Considering three genes ABC, the probability of observing A′B′C′ elsewhere in the genome, given that the paralogs exist, is $1/N^2$. Three proteins can occur across a spread of five positions in six ways; more generally, we compute the number of ways that *K* proteins can be spread across *J* positions by counting all possible arrangements of $K - 2$ proteins in the $J - 2$ positions between the first and last protein. Allowing for a spread to vary from *K* positions (no gaps) to *J* gives

$$L = \sum_{X=K-2}^{J-2} \binom{X}{K-2}$$

arrangements. Thus, the probability of chance occurrence is L/N^{K-1}. Allowing for both sets of genes (e.g., ABC and A′B′C′) to be spread across *J* positions increases this to L^2/N^{K-1}. The duplicated segment might be rearranged by the operations of reversal or translocation; allowing for *M* such rearrangements gives us a probability $P = L^2M/N^{K-1}$. For example, the probability of observing a duplicated set of three genes in two different locations, where the three genes occur across a spread of five positions in both locations, is $36/N^2$; the expected number of such matched sets in the predicted protein set is approximately $(N)36/N^2 = 36/N$, a value $\ll 1$. Therefore, any such duplications of three genes are unlikely to result from random rearrangements of the genome. If any of the genes occur in more than two copies, the probability that the apparent duplication has occurred by chance increases. The algorithm for selecting candidate duplications only generates matched protein sets with $P \ll 1$.

94. B. J. Trask *et al.*, *Hum. Mol. Genet.* **7**, 13 (1998); D. Sharon *et al.*, *Genomics* **61**, 24 (1999).
95. W. B. Barbazuk *et al.*, *Genome Res.* **10**, 1351 (2000); A. McLysaght, A. J. Enright, L. Skrabanek, K. H. Wolfe, *Yeast* **17**, 22 (2000); D. W. Burt *et al.*, *Nature* **402**, 411 (1999).
96. Reviewed in L. Skrabanek, K. H. Wolfe, *Curr. Opin. Genet. Dev.* **8**, 694 (1998).
97. P. Taillon-Miller, Z. Gu, Q. Li, L. Hillier, P. Y. Kwok, *Genome Res.* **8**, 748 (1998); P. Taillon-Miller, E. E. Piernot, P. Y. Kwok, *Genome Res.* **9**, 499 (1999).
98. D. Altshuler *et al.*, *Nature* **407**, 513 (2000).
99. G. T. Marth *et al.*, *Nature Genet.* **23**, 452 (1999).
100. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997).
101. M. Cargill *et al.*, *Nature Genet.* **22**, 231 (1999).
102. M. K. Halushka *et al.*, *Nature Genet.* **22**, 239 (1999).
103. J. Zhang, T. L. Madden, *Genome Res.* **7**, 649 (1997).
104. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).

105. From the observed coverage of the sequences at each site for each individual, we calculated the probability that a SNP would be detected at the site if it were present. For each level of coverage, there is a binomial sampling of the two homologs for each individual, and a heterozygous site could only be ascertained if both homologs are present, or if two alleles from different individuals are present. With coverage *x* from a given individual, both homologs are present in the assembly with probability $1 - (1/2)x^{-1}$. Even if both homologs are present, the probability that a SNP is detected is <1 because a fraction of sites failed the quality criteria. Integrating over coverage levels, the binomial sampling, and the quality distribution, we derived an expected number of sites in the genome that were ascertained for polymorphism for each individual. The nucleotide diversity was then the observed number of variable sites divided by the expected number of sites ascertained.

106. M. W. Nachman, V. L. Bauer, S. L. Crowell, C. F. Aquadro, *Genetics* **150**, 1133 (1998).
107. D. A. Nickerson *et al.*, *Nature Genet.* **19**, 233 (1998); D. A. Nickerson *et al.*, *Genomic Res.* **10**, 1532 (2000); L. Jorde *et al.*, *Am. J. Hum. Genet.* **66**, 979 (2000); D. G. Wang *et al.*, *Science* **280**, 1077 (1998).
108. M. Przeworski, R. R. Hudson, A. Di Rienzo, *Trends Genet.* **16**, 296 (2000).
109. S. Tavare, *Theor. Popul. Biol.* **26**, 119 (1984).
110. R. R. Hudson, in *Oxford Surveys in Evolutionary Biology*, D. J. Futuyma, J. D. Antonovics, Eds. (Oxford Univ. Press, Oxford, 1990), vol. 7, pp. 1–44.
111. A. G. Clark *et al.*, *Am. J. Hum. Genet.* **63**, 595 (1998).
112. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
113. H. Kaessmann, F. Heissig, A. von Haeseler, S. Paabo, *Nature Genet.* **22**, 78 (1999).
114. E. L. Sonnhammer, S. R. Eddy, R. Durbin, *Proteins* **28**, 405 (1997).
115. A. Bateman *et al.*, *Nucleic Acids Res.* **28**, 263 (2000).

116. Brief description of the methods used to build the Panther classification. First, the June 2000 release of the GenBank NR protein database (excluding sequences annotated as fragments or mutants) was partitioned into clusters using BLASTP. For the clustering, a seed sequence was randomly chosen, and the cluster was defined as all sequences matching the seed to statistical significance (E-value $< 10^{-5}$) and "globally" alignable (the length of the match region must be >70% and <130% of the length of the seed). If the cluster had more than five members, and at least one from a multicellular eukaryote, the cluster was extended. For the extension step, a hidden Markov Model (HMM) was trained for the cluster, using the SAM software package, version 2. The HMM was then scored against GenBank NR (excluding mutants but including fragments for this step), and all sequences scoring better than a specific (NLL-NULL) score were added to the cluster. The HMM was then retrained (with fixed model length) and all sequences in the cluster were aligned to the HMM to produce a multiple sequence alignment. This alignment was assessed by a number of quality measures. If the alignment failed the quality check, the initial cluster was rebuilt around the seed using a more restrictive E-value, followed by extension, alignment, and reassessment. This process was repeated until the alignment quality was good. The multiple alignment and "general" (i.e., describing the entire cluster, or "family") HMM (*176*) were then used as input into the BETE program (*177*). BETE calculates a phylogenetic tree for the sequences in the alignment. Functional information about the sequences in each cluster were parsed from SwissProt (*178*) and GenBank records. "Tree-attribute viewer" software was used by biologist curators to correlate the phylogenetic tree with protein function. Subfamilies were manually defined on the basis of shared function across subtrees, and were named accordingly. HMMs were then built for each subfamily, using information from both the subfamily and family (K. Sjölander, in preparation). Families were also manually named according to the functions contained within them. Finally, all of the families and subfamilies were classified into categories and subcategories based on their molecular functions. The categorization was done by manual review of the family and subfamily names, by examining SwissProt and GenBank records, and by review of the literature as well as resources on the World Wide Web. The current version (2.0) of the Panther molecular function schema has four levels: category, subcategory, family, and subfamily. Protein sequences for whole eukaryotic genomes (for the predicted human proteins and annotated proteins for fly, worm, yeast, and *Arabidopsis*) were scored against the Panther library of family and subfamily HMMs. If the score was significant (the NLL-NULL score cutoff depends on the protein family), the protein was assigned to the family or subfamily function with the most significant score.

117. C. P. Ponting, J. Schultz, F. Milpetz, P. Bork, *Nucleic Acids Res.* **27**, 229 (1999).
118. A. Goffeau *et al.*, *Science* **274**, 546, 563 (1996).
119. *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
120. S. A. Chervitz *et al.*, *Science* **282**, 2022 (1998).
121. E. R. Kandel, J. H. Schwartz, T. Jessell, *Principles of Neural Science* (McGraw-Hill, New York, ed. 4, 2000).
122. D. A. Goodenough, J. A. Goliger, D. L. Paul, *Annu. Rev. Biochem.* **65**, 475 (1996).
123. D. G. Wilkinson, *Int. Rev. Cytol.* **196**, 177 (2000).
124. F. Nakamura, R. G. Kalb, S. M. Strittmatter, *J. Neurobiol.* **44**, 219 (2000).
125. P. J. Horner, F. H. Gage, *Nature* **407**, 963 (2000); P. Casaccia-Bonnefil, C. Gu, M. V. Chao, *Adv. Exp. Med. Biol.* **468**, 275 (1999).
126. S. Wang, B. A. Barres, *Neuron* **27**, 197 (2000).
127. M. Geppert, T. C. Sudhof, *Annu. Rev. Neurosci.* **21**, 75 (1998); J. T. Littleton, H. J. Bellen, *Trends Neurosci.* **18**, 177 (1995).
128. A. Maximov, T. C. Sudhof, I. Bezprozvanny, *J. Biol. Chem.* **274**, 24453 (1999).
129. B. Sampo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 3666 (2000).
130. G. Lemke, *Glia* **7**, 263 (1993).
131. M. Bernfield *et al.*, *Annu. Rev. Biochem.* **68**, 729 (1999).
132. N. Perrimon, M. Bernfield, *Nature* **404**, 725 (2000).
133. U. Lindahl, M. Kusche-Gullberg, L. Kjellen, *J. Biol. Chem.* **273**, 24979 (1998).
134. J. L. Riechmann *et al.*, *Science* **290**, 2105 (2000).
135. T. L. Hurskainen, S. Hirohata, M. F. Seldin, S. S. Apte, *J. Biol. Chem.* **274**, 25555 (1999).

136. R. A. Black, J. M. White, *Curr. Opin. Cell Biol.* **10**, 654 (1998).
137. L. Aravind, V. M. Dixit, E. V. Koonin, *Trends Biochem. Sci.* **24**, 47 (1999).
138. A. G. Uren *et al.*, *Mol. Cell* **6**, 961 (2000).
139. P. Garcia-Meunier, M. Etienne-Julan, P. Fort, M. Piechaczyk, F. Bonhomme, *Mamm. Genome* **4**, 695 (1993).
140. K. Meyer-Siegler *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 8460 (1991).
141. N. R. Mansur, K. Meyer-Siegler, J. C. Wurzer, M. A. Sirover, *Nucleic Acids Res.* **21**, 993 (1993).
142. N. A. Tatton, *Exp. Neurol.* **166**, 29 (2000).
143. N. Kenmochi *et al.*, *Genome Res.* **8**, 509 (1998).
144. F. W. Chen, Y. A. Ioannou, *Int. Rev. Immunol.* **18**, 429 (1999).
145. H. O. Madsen, K. Poulsen, O. Dahl, B. F. Clark, J. P. Hjorth, *Nucleic Acids Res.* **18**, 1513 (1990).
146. D. M. Chambers, J. Peters, C. M. Abbott, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4463 (1998); A. Khalyfa, B. M. Carlson, J. A. Carlson, E. Wang, *Dev. Dyn.* **216**, 267 (1999).
147. D. Aeschlimann, V. Thomazy, *Connect. Tissue Res.* **41**, 1 (2000).
148. P. Munroe *et al.*, *Nature Genet.* **21**, 142 (1999); S. M. Wu, W. F. Cheung, D. Frazier, D. W. Stafford, *Science* **254**, 1634 (1991); B. Furie *et al.*, *Blood* **93**, 1798 (1999).
149. J. W. Kehoe, C. R. Bertozzi, *Chem. Biol.* **7**, R57 (2000).
150. T. Pawson, P. Nash, *Genes Dev.* **14**, 1027 (2000).
151. A. W. van der Velden, A. A. Thomas, *Int. J. Biochem. Cell Biol.* **31**, 87 (1999).
152. C. M. Fraser *et al.*, *Science* **281**, 375 (1998); H. Tettelin *et al.*, *Science* **287**, 1809 (2000).
153. D. Brett *et al.*, *FEBS Lett.* **474**, 83 (2000).
154. H. J. Muller, H. Kern, *Z. Naturforsch. B* **22**, 1330 (1967).
155. H. J. Muller, in *Heritage from Mendel*, R. A. Brink, Ed. (Univ. of Wisconsin Press, Madison, WI, 1967), p. 419.
156. J. F. Crow, M. Kimura, *Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).
157. K. Kobayashi *et al.*, *Nature* **394**, 388 (1998).
158. A. P. Feinberg, *Curr. Top. Microbiol. Immunol.* **249**, 87 (2000).
159. C. A. Collins, C. Guthrie, *Nature Struct. Biol.* **7**, 850 (2000).
160. S. R. Eddy, *Curr. Opin. Genet. Dev.* **9**, 695 (1999).
161. Q. Wang, J. Khillan, P. Gadue, K. Nishikura, *Science* **290**, 1765 (2000).
162. M. Holcik, N. Sonenberg, R. G. Korneluk, *Trends Genet.* **16**, 469 (2000).
163. T. A. McKinsey, C. L. Zhang, J. Lu, E. N. Olson, *Nature* **408**, 106 (2000).
164. E. Capanna, M. G. M. Romanini, *Caryologia* **24**, 471 (1971).
165. J. Maynard Smith, *J. Theor. Biol.* **128**, 247 (1987).
166. D. Charlesworth, B. Charlesworth, M. T. Morgan, *Genetics* **141**, 1619 (1995).
167. J. E. Bailey, *Nature Biotechnol.* **17**, 616 (1999).
168. R. Maleszka, H. G. de Couet, G. L. Miklos, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 3731 (1998).
169. G. L. Miklos, *J. Neurobiol.* **24**, 842 (1993).
170. J. P. Crutchfield, K. Young, *Phys. Rev. Lett.* **63**, 105 (1989); M. Gell-Mann, S. Lloyd, *Complexity* **2**, 44 (1996).
171. A. L. Barabasi, R. Albert, *Science* **286**, 509 (1999).
172. E. Colucci-Guyon *et al.*, *Cell* **79**, 679 (1994).
173. J. Sambrook, E. F. Fritch, T. Maniatis, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 2, 1989).
174. B. Ewing, P. Green, *Genome Res.* **8**, 186 (1998); B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res.* **8**, 175 (1998).
175. E. S. Lander, M. S. Waterman, *Genomics* **2**, 231 (1988).
176. A. Krogh, K. Sjölander, *J. Mol. Biol.* **235**, 1501 (1994).
177. K. Sjölander, *Proc. Int. Soc. Mol. Biol.* **6**, 165 (1998).
178. A. Bairoch, R. Apweiler, *Nucleic Acids Res.* **28**, 45 (2000).
179. GO, available at www.geneontology.org/.
180. R. L. Tatusov, M. Y. Galperin, D. A. Natale, E. V. Koonin, *Nucleic Acids Res.* **28**, 33 (2000).

181. We thank E. Eichler and J. L. Goldstein for many helpful discussions and critical reading of the manuscript, and A. Caplan for advice and encouragement. We also thank T. Hein, D. Lucas, G. Edwards, and the Celera IT staff for outstanding computational support. The cost of this project was underwritten by the Celera Genomics Group of the Applera Corporation. We thank the Board of Directors of Applera Corporation: J. F. Abely Jr. (retired), R. H. Ayers, J.-L. Bélingard, R. H. Hayes, A. J. Levine, T. E. Martin, C. W. Slayman, O. R. Smith, G. C. St. Laurent Jr., and J. R. Tobin for their vision, enthusiasm, and unwavering support and T. L. White for leadership and advice. Data availability: The genome sequence and additional supporting information are available to academic scientists at the Web site (www.celera.com). Instructions for obtaining a DVD of the genome sequence can be obtained through the Web site. For commercial scientists wishing to verify the results presented here, the genome data are available upon signing a Material Transfer Agreement, which can also be found on the Web site.

5 December 2000; accepted 19 January 2001

ERRATUM

post date 8 June 2001

Gene prediction*				
Otto	De novo/ any	De novo/ 2×	Total (Otto + de novo/ any)	Total (Otto + de novo/ 2×)
1,743	1,710	710	3,453	2,453
1,183	1,771	633	2,954	1,816
1,013	1,414	598	2,427	1,611
696	1,165	449	1,861	1,145
892	1,244	474	2,136	1,366
943	1,314	524	2,257	1,467
759	1,072	460	1,831	1,219
583	977	357	1,560	940
689	848	329	1,537	1,018
685	968	342	1,653	1,027
1,051	1,134	535	2,185	1,586
925	936	417	1,861	1,342
341	691	241	1,032	582
583	700	290	1,283	873
558	640	246	1,198	804
748	673	247	1,421	995
897	648	313	1,545	1,210
283	543	189	826	472
1,141	534	268	1,675	1,409
517	469	180	986	697
184	265	102	449	286
494	341	147	835	641
605	860	387	1,465	992
55	155	49	210	104
196	278	132	474	328
17,764	21,350	8,619	39,114	26,383
714	812	333	1,526	1,047

REPORTS: "The sequence of the human genome" by J. C. Venter *et al.* (16 Feb. 2001, p. 1304). In Table 10, the last column under the heading "Gene prediction" should have read "Total (Otto + de novo/2×)." This section of the table with the corrected column heading is shown here. The asterisk indicates that the chromosomal assignment is unknown.

In the References and Notes section, the authors for reference 176 should have read "A. Krogh *et al.*"; the journal name in reference 177 should have been "*Proc. Intell. Syst. Mol. Biol.*"; and in note 181, the acknowledgement list should have included after G. Edwards the names L. Foster, D. Bhandari, P. Davies, T. Safford, and J. Schira.