武漢大学 Wuhan University

遗 传 学（第3版）

第10章 基因组学与功能基因组学

1. 真核生物基因组概述
2. 人类基因组计划与基因组学
3. 基因组测序与序列组装
4. 基因组图谱构建与应用
5. 比较基因组学和功能基因组学
6. 基因组学研究新进展及应用

10.1 真核生物基因组概述

与原核生物基因组不同，真核生物基因组的主要特点表现在以下几方面：

①真核生物基因组一般由多条染色体组成。核小体是组成染色体的基本结构单位。核小体上组蛋白上的氨基酸序列能够被进行各种修饰。

②每条染色体的DNA分子具有多个复制起点；基因多为不连续基因。

③DNA编码序列仅占基因组DNA的一小部分，绝大多数为非编码序列。

④一般没有重叠基因。在转录水平上，存在多种转录剪接机制，同一基因转录物在不同的发育阶段和不同的组织细胞中可以翻译成不同的蛋白质。

⑤真核生物基因组中存在着大量的基因家族和奢侈基因。

⑥基因组中存在大量的重复序列。

⑦真核生物基因组的表达调控比较复杂，在个体的不同发育阶段和时期、不同的细胞和组织，存在着不同水平的表达调控机制

10.2 人类基因组计划与基因组学

基因组学（genomics）：

它是研究基因组的组成、结构和功能的学科。

结构基因组学（structural genomics）：

是着重研究基因组的结构并构建高分辨的遗传图、物理图、序列图和转录图以及研究蛋白质组成与结构的学科；

功能基因组学（functional genomics）：

主要是利用结构基因组学研究所得到的各种信息在基因组水平上研究编码序列及非编码序列生物学功能的学科。

10.2.1 人类基因组计划

美国能源部和国立卫生研究院（NIH）合作于1990年正式启动人类基因组计划。主要目标是计划拨款30亿美元，用15年时间完成人类基因组30亿bp全部序列的测定，在2001年完成全部染色体的“工作草图”。

经过参加该项目的1000多名各国科学家的通力合作，人类基因组的工作草图于2000年6月26日胜利绘制完成，该工作草图包含人体90%以上碱基对的位置信息。

2000年6月26日，六国科学家公布人类基因组工作框架图

2001 年2月12日，人类基因组图谱及初步分析结果首次公布

2001 年8月26日，中国提前两年完成1%人类基因组测序任务

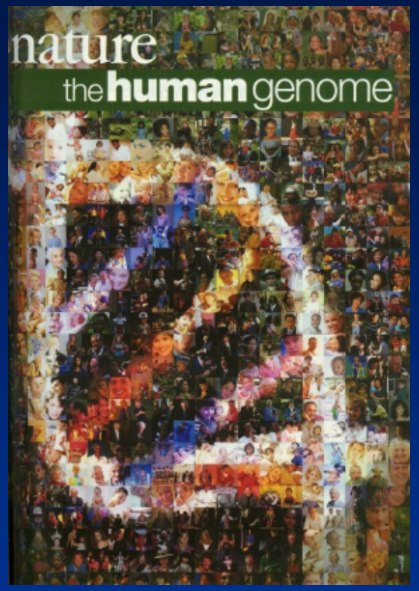
2003年4月15日，六个国家共同宣布人类基因组序列图完成

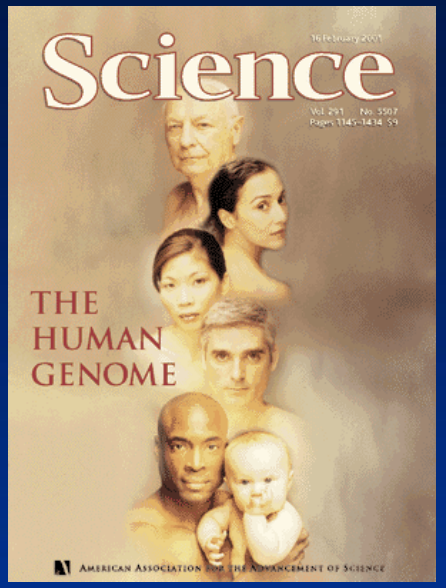
(美、英、德、日、法、中)

已经完成的人类基因组图谱(序列)来自不同人种的5个个体，是一个参考图谱。

2004年10月，国际人类基因组测序联合体在Nature周刊上发表了人类基因组常染色质全序列测定的论文

宣布人类基因组由31.647亿个碱基对组成，拥有编码蛋白质的基因数目大约在2万到2.5万个之间。





February 2001,The HGP consortium publishes its working draft in Nature (15 February), and Celera publishes its draft in Science (16 February).

articles

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

* A list of authors and their affiliations appears in the Supplementary Information

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human enome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

10.2.2 人类基因组的结构特点

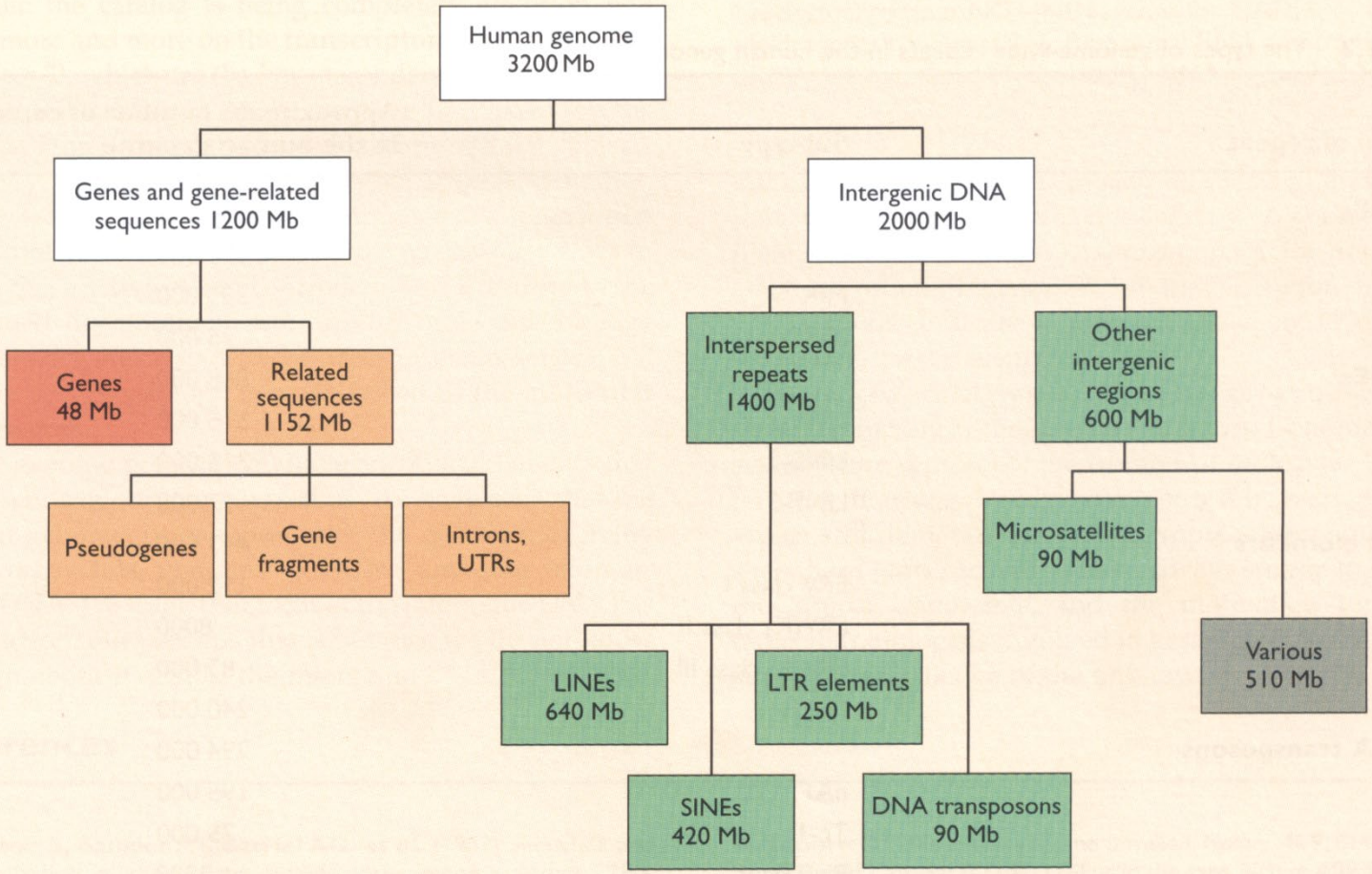


图10-1 人类基因组的组织组成

研究表明：编码基因的序列仅占人类基因组DNA的1%左右，98%以上的序列是非编码序列。基因中内含子占基因组的24%，因此基因组涉及与产生蛋白质有关的序列达到25%

重复序列占50%以上，主要分成5种类型：

①转座子成分：占基因组的45%

均以多拷贝的形式存在于基因组中

②已加工假基因（processed pseudogene）：

约占基因组的0.1%

③简单重复序列，约占基因组的3%

④大片段重复（长10～30 kb）：占基因组约5%。只有少部分在相同的染色体上，多数分布在不同的染色体上

⑤串联重复，主要位于着丝粒和端粒部位。

10.2.3 模式生物基因组研究与宏基因组研究

HGP除了对人类基因组的测序外，还进行包括大肠杆菌、酵母、线虫、果蝇、拟南芥、小鼠等模式生物体基因组的研究计划。对于这些模式生物基因组研究和宏基因组研究分别详见e10-1和e10-2。

10.3 基因组测序与序列组装

10.3.1 基因组测序策略

主要有两种测序策略，一种是自上而下（top-down mapping）或由长到短作图策略；另一种是全基因组鸟枪法（whole-genome shotgun method）作图策略，也称自下而上（Bottom-up approach / mapping）或由短到长作图策略（图10-2）。

(1).Physical mapping （测序策略）

① By the clone contig approach (图10-2)

建立连续重叠克隆系（overlapping clones)/重叠群(Contig),对单个重叠群,采用鸟枪法对其中的克隆逐个进行测序,最后在重叠群内进行拼接,拼接出全长序列。

这种方法也称为 Top-down mapping： “自上而下”

或由长到短作图:

先构建大DNA克隆(100Kb-10000Kb),并把克隆依染色体排列---染色体的克隆图。当把每个克隆测序完成后，就可以拼装出整个染色体的DNA序列。 (Fig10.6)

注：重叠群/叠连群：相互间存在重叠顺序的一组克隆，根据重叠顺序的相对位置将各个克隆首尾连接，构成连续顺序图。

② By the whole-genome shotgun method

(图10-2)

直接将基因组DNA随机切成 2Kb 左右的小片段（BAC克隆）， 随机末端测序，再以基因组的分子标记为起点进行DNA片段拚接，其余过程辅以少量大片段(约10Kb)，计算机分析串连得全序列。

这是一种“自下而上”或由短到长作图

Bottom-up approach/mapping

10.3.2 基因组测序方法与组装

20世纪80年代DNA测序技术主要有：

Sanger等（1977）发明的双脱氧链末端终止法

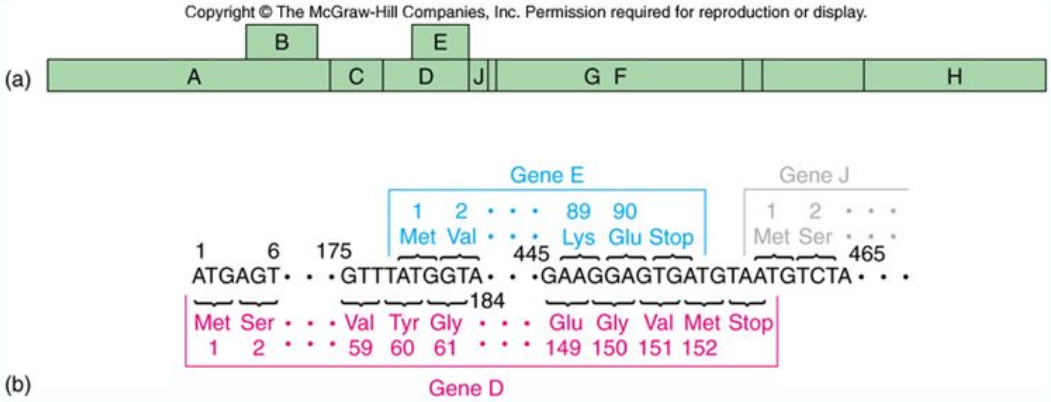
Maxam和 Gilbert（1977）发明的化学降解法

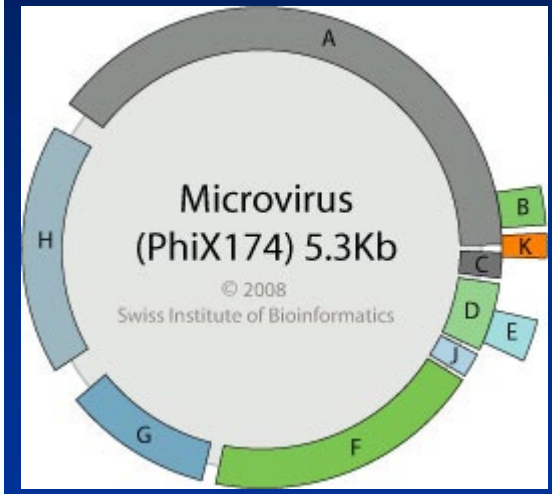
两者原理上差异很大，但都是根据核苷酸在某一固定的点开始，随机在某一个特定的碱基处终止，产生A、T、C、G四组不同长度的一系列核苷酸，然后在尿素变性的PAGE胶上电泳进行检测，从而获得DNA序列。

1977年Sanger实验室采用末端终止法完成的ΦX174全基因组的测定是世界上第一个被测定的基因组。由于基因组测序是大规模的序列测定，要求快速、准确，这就必须要求实现自动化测序。Sanger的末端终止法具有这些特点。因此在后来很快取代了Maxam-Gilbert的化学降解法，而成为DNA序列测定的主流技术。

Phage ϕX174 Genome

- First genome sequenced was a very simple one, phage ϕX174
  - Completed by Sanger in 1977
  - 5375-nucleotides
- Note that some of these phage genes overlap



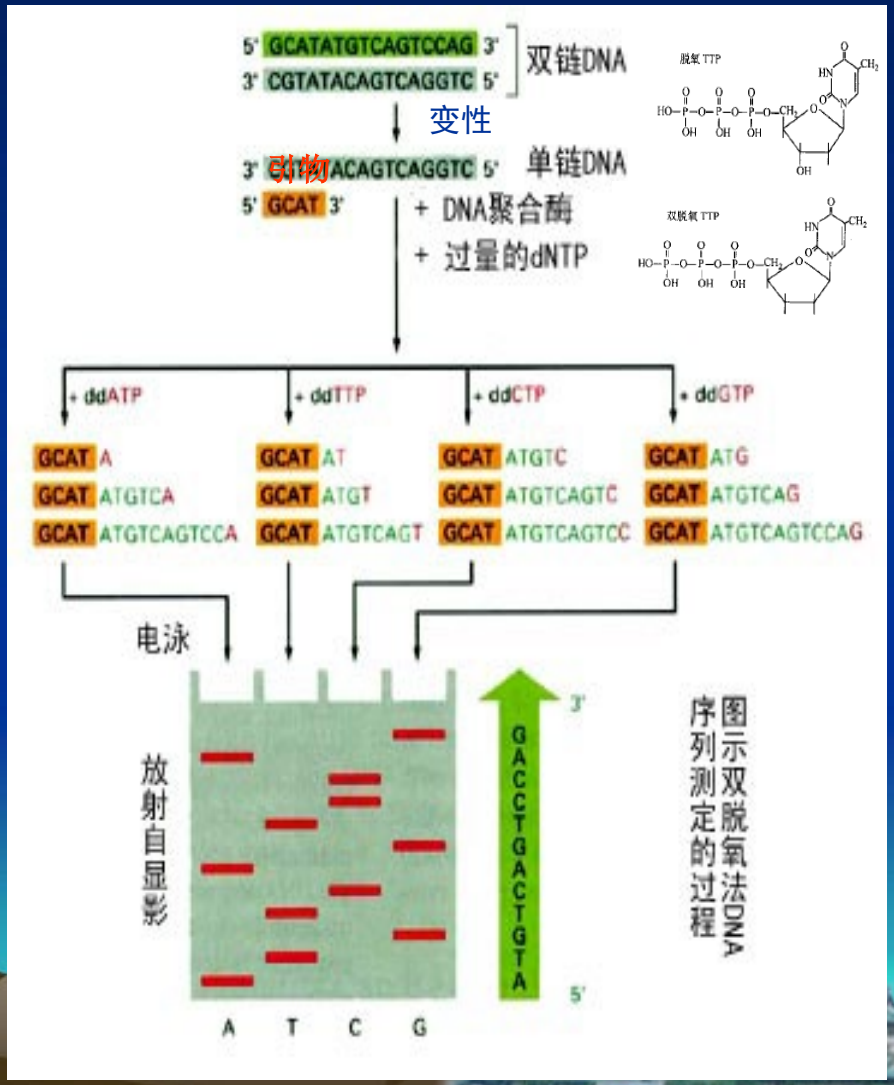


First generation DNA sequencing

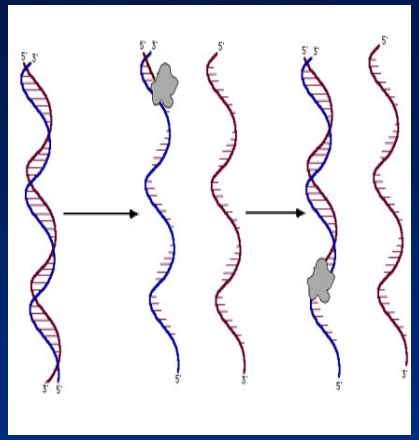
Sanger 等的双脱氧链末端终止法进行测序

该技术：DNA聚合酶+测序胶

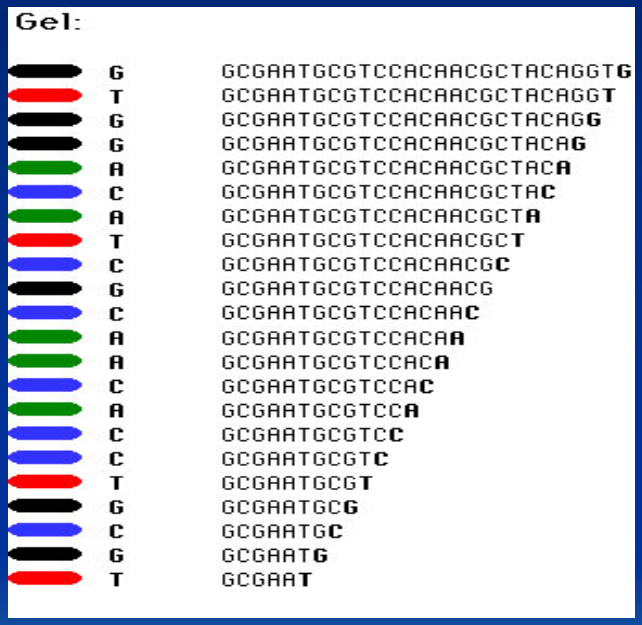
DNA聚合酶是在退火于“模板”DNA的引物的3′端上进行延伸。双脱氧测序法利用了DNA聚合酶能以双脱氧核糖核苷酸为底物的特性。当ddNTP被掺入到引物延伸的引物的3′端时，由于链上3′羟基的缺如，链的延伸就终止于G、A、T或C。在4个测序反应中，每个反应只需各加入4种可能的ddNTP中的其中一种，就将产生如图所示的4个序列阶梯。

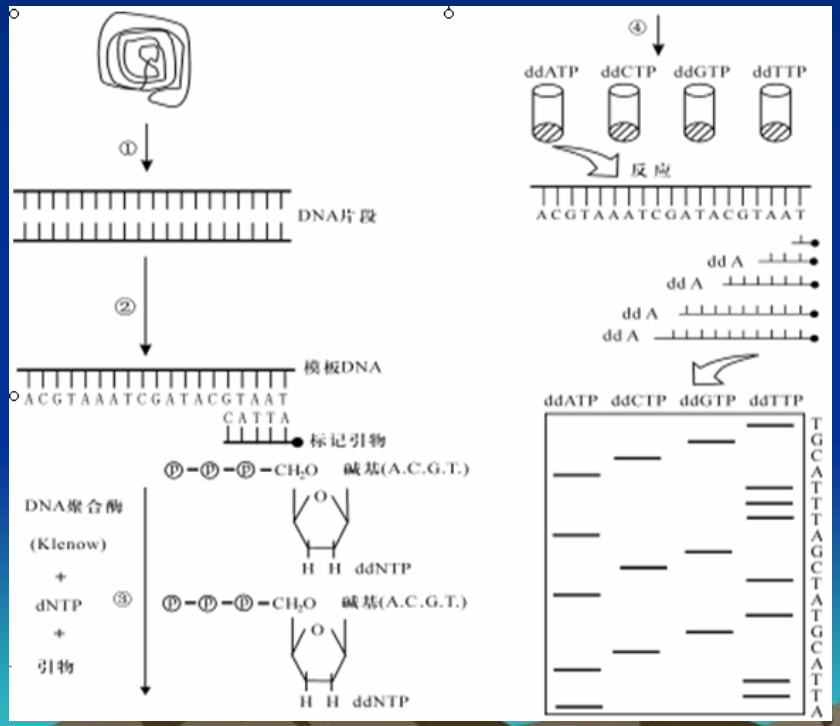


Sanger双脱氧终止法



- 与 PCR反应类似
- 反应体系中包含：模板 DNA, Taq酶, dNTPs ddNTPs和测序引物
- 反应过程： 变性—复性—延伸—终止

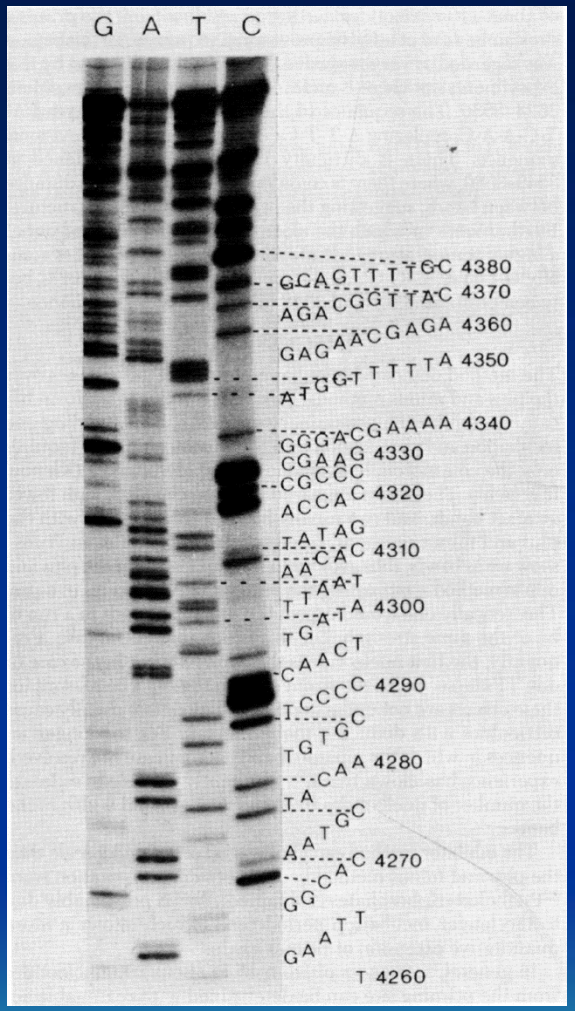




Dideoxynucleotides
（双脱氧核苷酸）

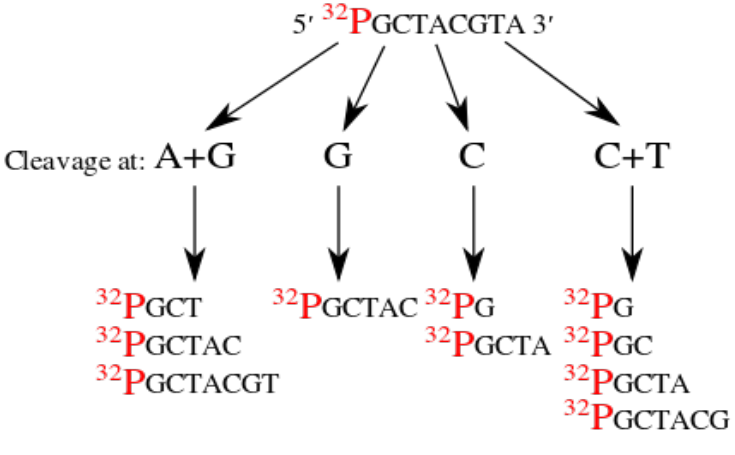
ddNTPs 是反应终止剂可以当作正常碱基参与复制，一旦链入DNA中，其后就不能再继续连接。反应体系中dNTPs的浓度远高ddNTPs (一般3~4 ： 1)。

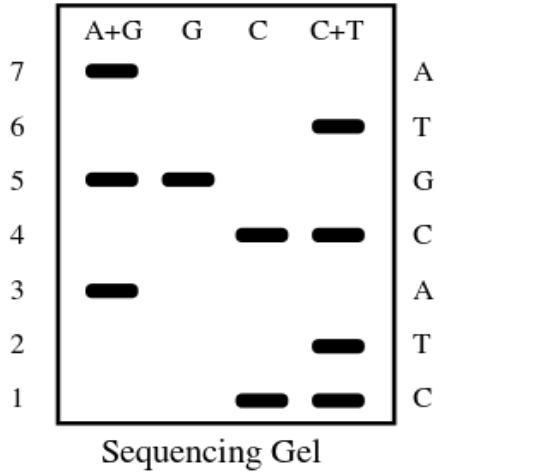
First generation DNA sequencing



Frederick Sanger's enzymatic dideoxy DNA sequencing technique based on the chain-terminating dideoxynucleotide analogues (Sanger et al. 1977)

Allan Maxam and Walter Gilbert's chemical degradation DNA sequencing technique in which terminally labeled DNA fragments were chemically cleaved at specific bases and separated by gel electrophoresis





Proc. Natl. Acad. Sci. USA 74 (1977)

Maxam–Gilbert sequencing1977

Proc. Natl. Acad. Sci. USA 74 (1977)

采用Sanger末端终止法原理的测序技术：荧光毛细管电泳技术是第一代基因组测序的主要技术之一（图10-3）。

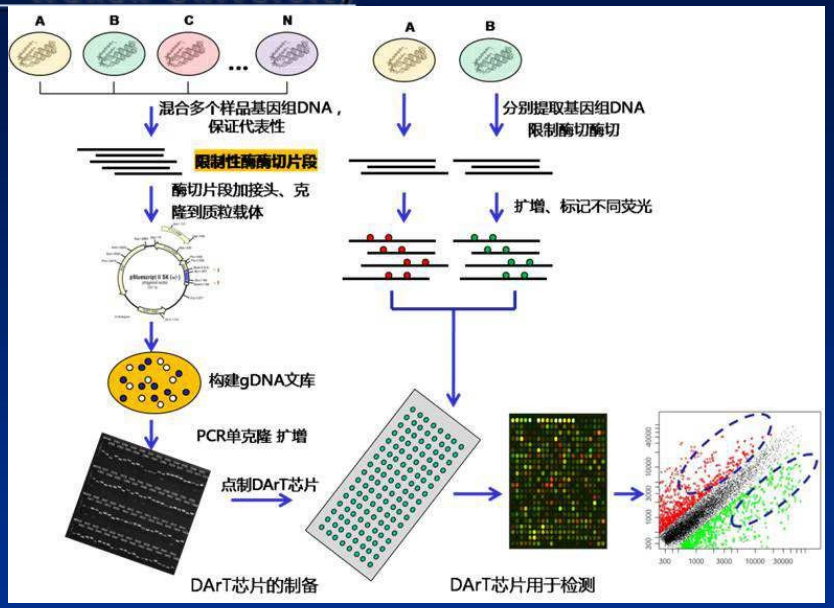
其优点是：

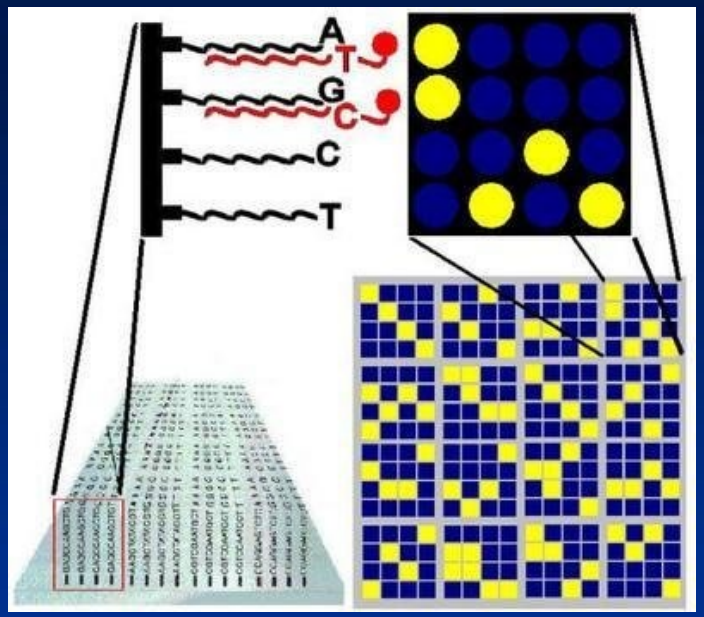
①使用不同荧光分别标记4种ddNTP，加入到一个反应中，聚合链终止反应完成后，可以获得分别带有A，C，T和G的DNA单链。

②自动化荧光测序系统避免了人眼分辨的差错，结果准确可靠，测序质量大大提高。

③测序速度提高，有96个泳道，每次可同时进行96次测序，极大地提高了测序的工作效率，1天可完成近千次反应

DNA芯片杂交测序（sequencing by hybridization）技术





是继荧光毛细管电泳技术后发展起来的一新种技术

DNA芯片（DNA chip）：是指采用类似大规模集成电路的手段将寡核苷酸探针或者DNA、cDNA有规律地排列固定在指甲大小（或更小）的硅片上而形成的测序集成块。在DNA芯片杂交测序技术中，当待测的样品DNA分子与芯片温浴，凡是能杂交的寡核苷酸都会在确定的位置发出杂交信号，然后根据获取的信息将寡核苷酸的序列进行对比组装，拼接成完全的DNA序列。

10.3.3 新一代测序方法

第二代DNA测序技术（second-generation DNA sequencing）也称为新一代DNA测序技术（next-generation DNA sequencing）。主要包括：

（1）Roche (454) GS FLX sequencer

（2）Illumina Solexa（ Illumina genome analyzer ）

（3）Applied Biosystems（Applied Biosystems Genetic Analyzer）/ ABI SOLiD（ SOLiD sequencer ）

第三代DNA测序技术平台

（1） Heliscope

（2）SMRT（Single Molecule Real-time Technology）

（3） Nanopore Sequencing Platform

等多种测序仪器组成的基因组分析实验技术平台

传统Sanger测序法（a）及新一代测序技术（b）流程图

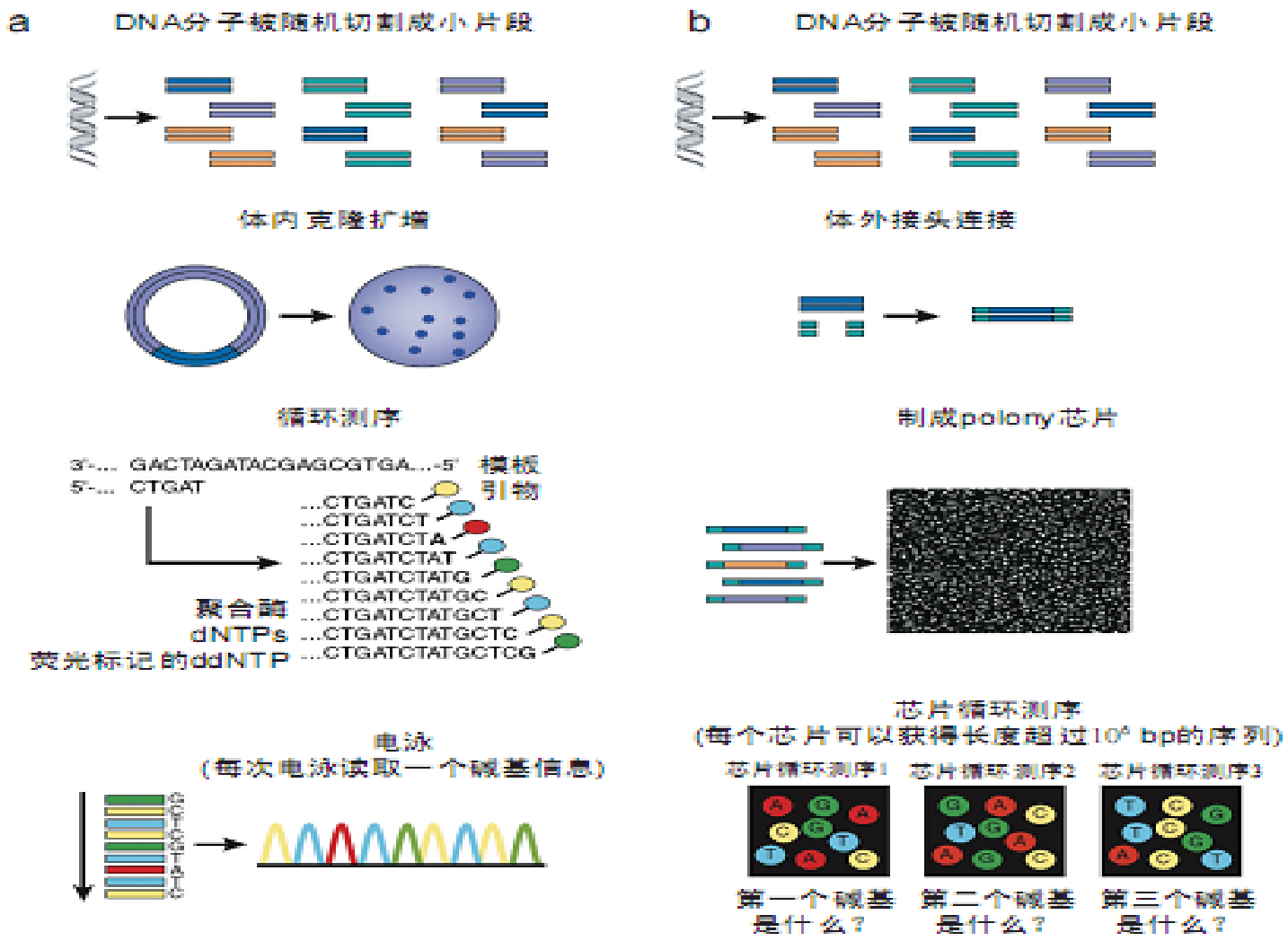


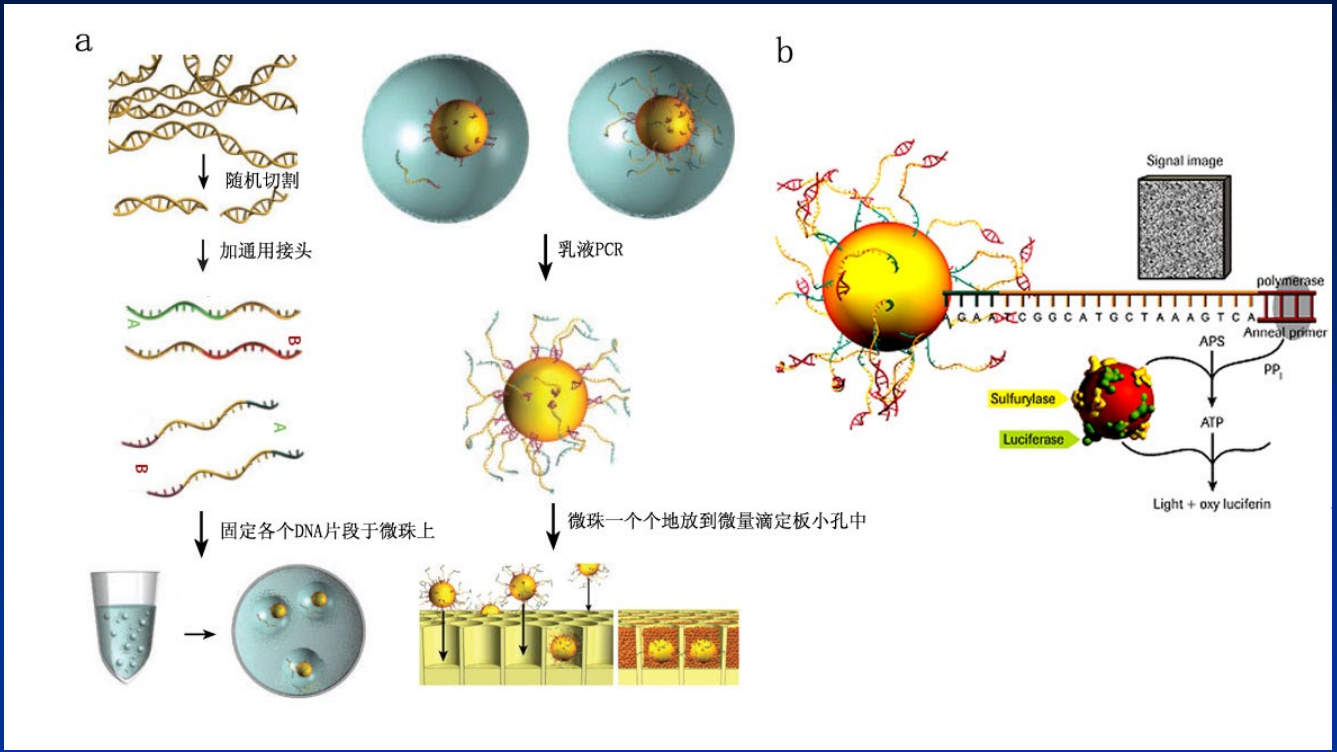
图 1 传统的Sanger测序法及新一代DNA测序技术工作流程图。
（a）高通量鸟枪Sanger测序法。首先基因组DNA被随机切割成小片段分子，接着众多小片段DNA被克隆入质粒载体，随后转化到大肠杆菌中。最后培养大肠杆菌提取质粒，进行测序。每一个测序反应都在只有几微升的反应体系中完成。测序后获得一系列长短不一的末端标记有荧光的片段，最后通过对每一个延伸反应产物末端荧光颜色进行识别来读取DNA序列。（b）鸟枪循环芯片测序法。首先将基因组DNA随机分割成小片段DNA分子，然后在这些小片段

1. Roche (454) GS FLX sequencer

a. 文库制备、模板制备和测序准备过程

b. 焦磷酸测序基本原理

该技术平台最主要的优点是测序读长较长，目前可以准确进行400个以上的碱基序列分析



454的技术平台（ 2005 年）是将一条DNA片段与一个微珠结合，利用乳液PCR（emulsion PCR）技术，每个微珠（磁珠+ DNA模板分子）分别获得几百万DNA片段的拷贝。在微珠进入微滴度扳后，每个DNA碱基（A、C、T、G）在微板上被分别洗出，并边合成边测序。CCD摄像机记录了在焦磷酸测序反应中10万个微球同一时间添加的每一个碱基。

焦磷酸测序技术的原理

1987年Nyren 等发展一种新型的DNA测序技术:

焦磷酸测序技术(Pyrosequencing)

是由四种酶:

- DNA 聚合酶 (DNA polymerase)
- ATP 硫酸化酶(ATP sulfurylase)
- 荧光素酶(luciferase)
- 双磷酸酶(apyrase)

催化的同一反应体系中的酶级联反应

反应底物为:

5′-磷酰硫酸(adenosine 5′-phosphosulfate, APS)

荧光素(luciferin)

反应体系还包括待测序DNA单链和测序引物

武漢大學
Wuhan University

焦磷酸测序技术简介

- introduction 02

武漢大學
Wuhan University

DNA library preparation

武漢大學
Wuhan University

emPCR

武漢大學
Wuhan University

Data Analysis

武漢大學
Wuhan University

Summary

2、Illumina Solexa（ Illumina genome analyzer ）

Solexa技术最早由两位剑桥大学的化学家创立，利用专利核心技术“DNA簇” 和“可逆性末端终结”，达成自动化样本制备及基因组数百万个碱基大规模平行测序。

Illumina公司于2007年收购 Solexa，并推出成熟商业产品 Genome Analyzer (基因组分析系统)。 Genome Analyzer 为新一代革新性技术分子生物学综合技术平台，具有高准确性，高通量，高灵敏度，和低运行成本等突出优势。可以同时完成传统基因组学研究（测序和注释）以及功能基因组学 （基因表达及调控，基因功能，蛋白/核酸相互作用）研究。

Solexa技术的基本原理与测序流程

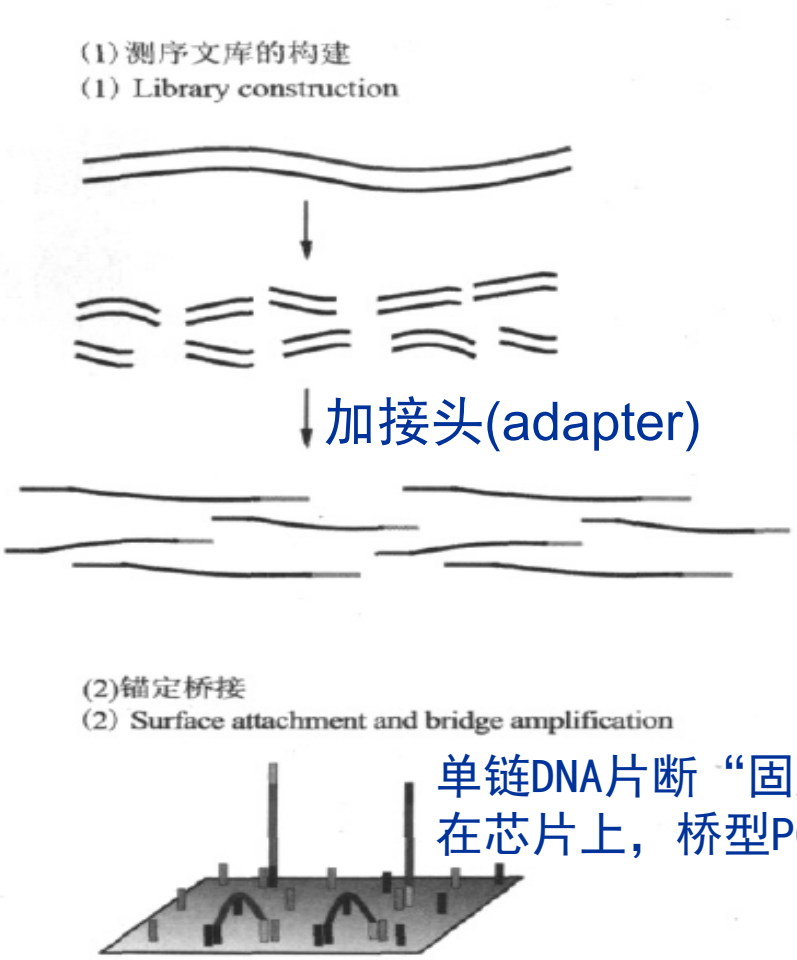


图 3 Illumina genome analyzer (Solexa)测序流程(Mardis, 2008)

注: (1)测序文库的构建. 将基因组 DNA 随机打断后在单链 DNA 两端加上接头; (2)锚定桥接. 每一个带接头的 DNA 片断与测序通道上的接头引物随机结合, 添加未标记的 dNTP 和普通 *Taq* 酶进行固相桥式 PCR 扩增; (3)预扩增. 通过变性和桥式扩增循环在测序通道表面获得数百万条密集成簇的待测 DNA 片断; (4)单碱基延伸测序. 将 4 种被标记的 dNTP、引物和 DNA 聚合酶添加到测序通道内以启动地一个测序循环. 通过激光的激发, 从每个测序通道的测序簇里面产生出对应的荧光, 通过判断捕获的荧光颜色记录待测序簇的第 1 个碱基(http://www.illumina.com/pagesnrn.ilmn?ID=70)

武漢大學
Wuhan University

Solexa 测序原理及流程

技术特色突出表现在

1. 每张芯片有8个通道，每个通道可单独运行一个样品，也可以把多个样品混合在一起检测
2. 一次实验可读取10亿个碱基/芯片
3. 可精确读取重复序列如 AAAAAAAAAAAAAAAAA, TTTTTTTTTTTTT
4. 实验费用低，读取1百万碱基的成本最低至4美元
5. 无需建库，自动化样品制备，简单，成本低
6. 样本使用效率极高，所以对少量样本也可以极灵敏精确地检测 （1 ug DNA即可以进行末端双向测序反应）
7. 可以进行35碱基长度的末端双向测序反应

Solexa技术在应用中有如下特点

1. 不受物种限制，对人，动物，微生物，植物都可进行研究
2. 高灵敏度，精确性，及重复性
3. 无需预先知道模式物种基因组序列，无须合成探针，可直接进行全基因组表达研究
4. 不需要实验假设的支持
5. 可以检测到单拷贝分子的变化
6. 没有传统方法的荧光背景噪音的干扰

3、ABI SOLiD（SOLiD sequencer）

Applied Biosystems SOLiD (Supported Oligo Ligation Detection)（Sequencing by Oligonucleotide Ligation / Detection） sequencer与其他新一代测序仪不同的是，它不是利用DNA 聚合酶在合成过程中读取序列，而是利用DNA连接酶在连接过程读取序列。

SOLiD 测序的基本流程为：

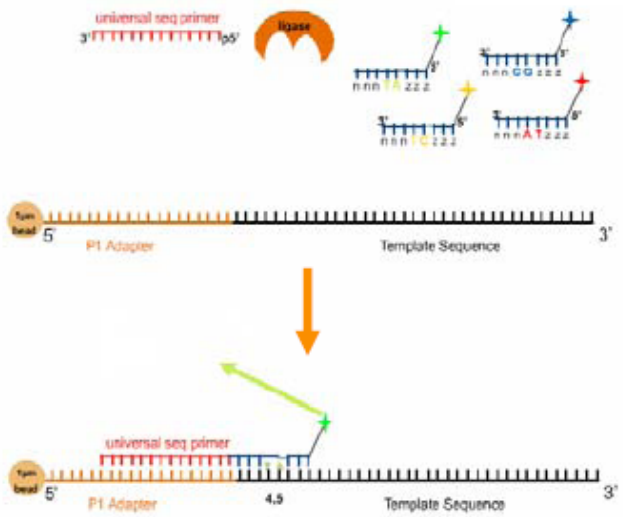
(1) 测序文库：在DNA 碎片两端加上一对接头P1、P2。

(2) 乳滴PCR

(3) 边连接边测序

ABI SOLiD Overview

SOLiD™ Chemistry System 4-color ligation
Ligation reaction



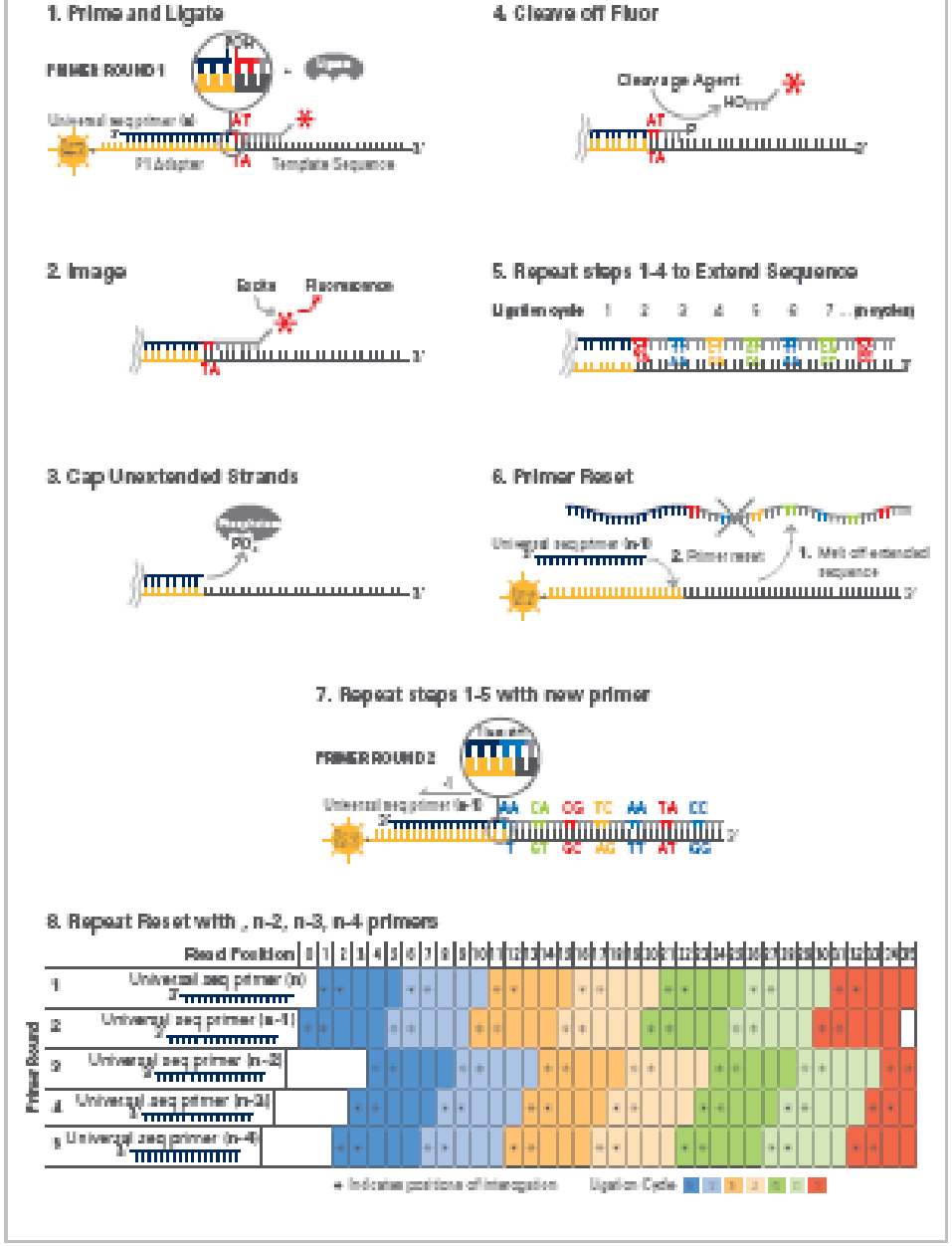


Figure 1: SOLiD™ System – Sequencing by ligation using di-base labeled probes

SOLiD测序原理与步骤

- SOLiD_video_final.wmv

表11 新一代测序技术的应用

应用范围	应用举例
全基因组测序（Complete genome resequencing）	人类个体基因组多态性及突变的全面检测
约化表示测序法（Reduced representation sequencing）	大规模多态性检测
靶向再测序（Targeted genomic resequencing）	靶向多态性及突变检测
末端配对测序（Paired end sequencing）	遗传及获得性结构变异检测
环境基因组测序（Metagenomic sequencing）	传染性及共生菌群检测
转录组测序（Metagenomic sequencing）	定量基因表达及选择性剪切；转录注释；转录SNPs或体细胞突变检测
小RNA测序（Small RNA sequencing）	microRNA表达谱
酸性亚硫酸盐标记DNA测序（Sequencing of bisulfite-treated DNA）	基因组DNA中胞嘧啶甲基化模式的测定
染色质免疫沉淀测序（ChIP-Seq）	全基因组蛋白质与DNA相互作用图谱
核酶片段及测序（Nuclease fragmentation and sequencing）	核小体定位
分子条码（Molecular barcoding）	多个体来源样品的多通路测序

武漢大學 Wuhan University

第三代测序技术

1、Helicos Biosciences

Heliscope

2、Pacific Biosciences

（Single Molecule Real-time Technology）

SMRT

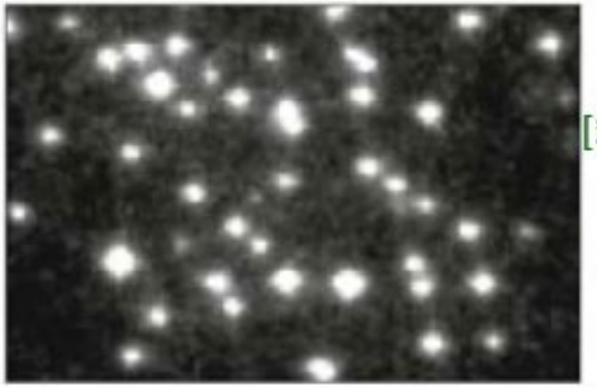
3、 Oxford Nanopore Technologies

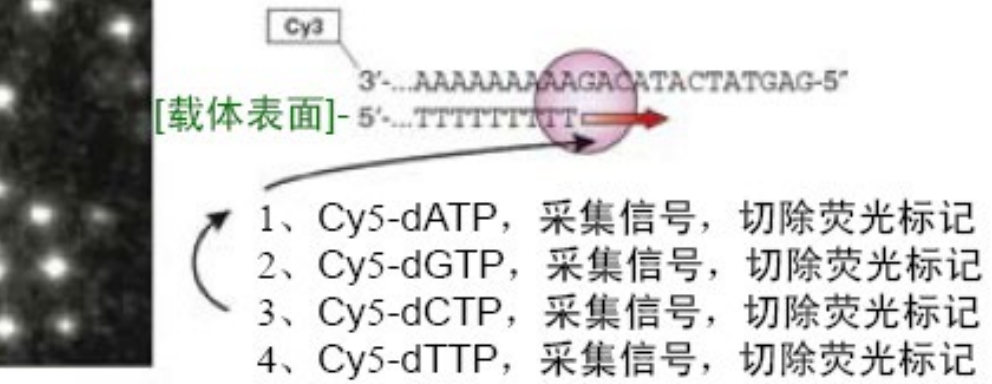
1. HeliScope测序仪

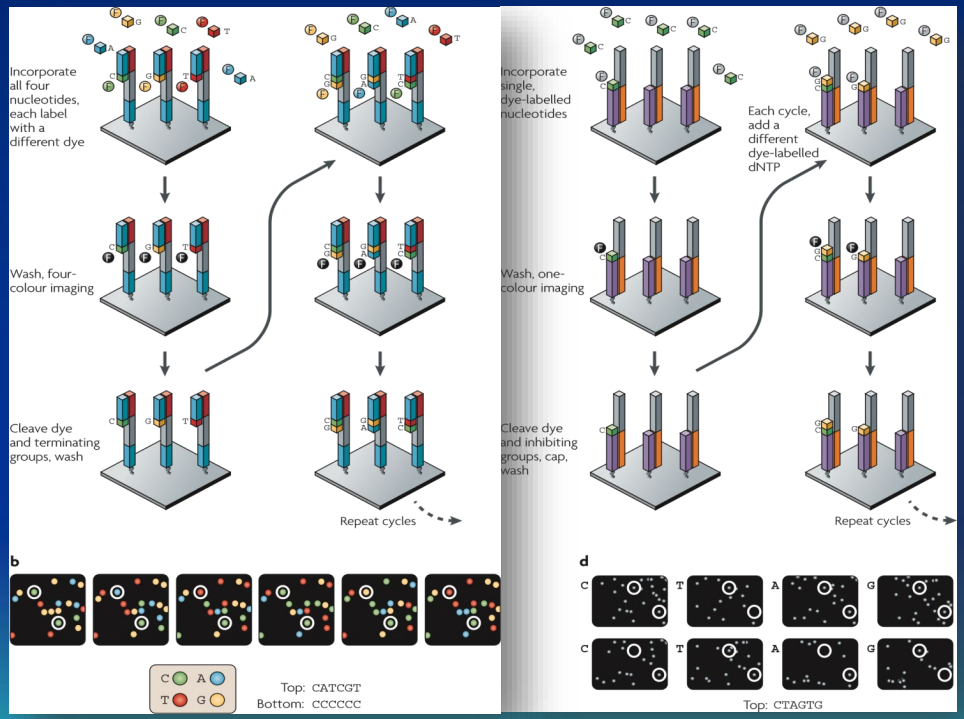
HeliScope测序仪是由Quake团队设计开发的，也是一种循环芯片测序设备。其最大的特点是无需对测序模板进行扩增，它使用了一种高灵敏度的荧光探测仪直接对单链DNA模板进行合成法测序。

首先，将基因组DNA切割成随机的小片段DNA分子，并且在每个片段末端加上poly-A尾。然后通过poly-A尾和固定在芯片上的poly-T杂交，将待测模板固定到芯片上，制成测序芯片。最后借助聚合酶将荧光标记的单核苷酸掺入到引物上。采集荧光信号，切除荧光标记基团，进行下一轮测序反应，如此反复，最终获得完整的序列信息。根据最近的报道，经过数百轮这种单碱基延伸可以获得25bp或更长的测序长度。

HeliScope测序仪的其它特点见下表







HeliScope测序仪：单核酸分子不经扩增直接测序。Poly-A尾被添加到DNA文库片段末端，通过与固定在芯片上的Poly-T互补杂交将模板链固定到芯片，制成测序芯片。模板上标记有Cy3以标识出它们在芯片上的位置。DNA聚合酶将荧光标记的单核苷酸掺入到引物上，采集荧光信号获得序列信息，切除荧光基团，重复上述步骤，完成测序。

Pushkarev D, Neff N F, Quake S R, 2009

HeliScope单分子测序过程

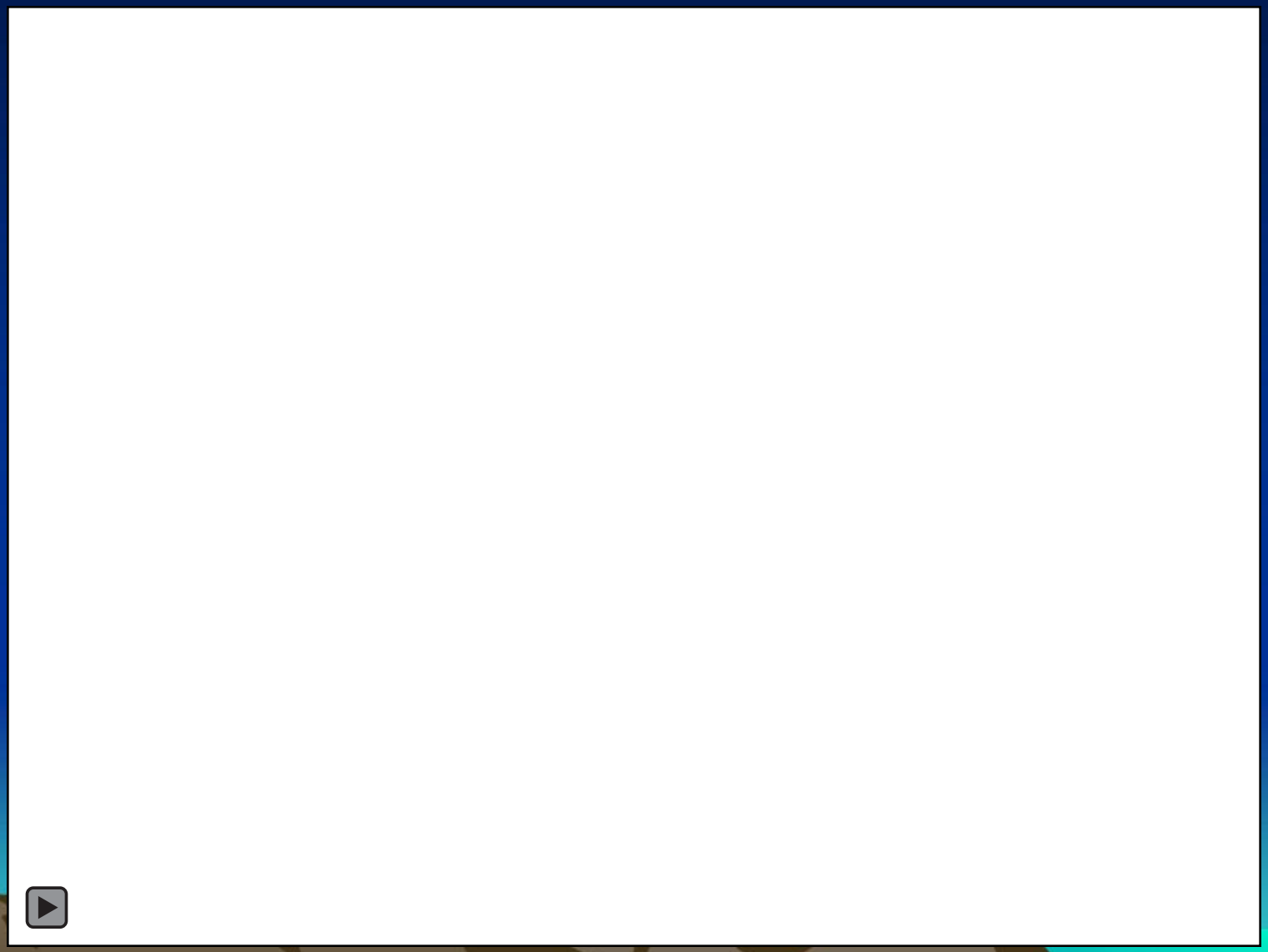


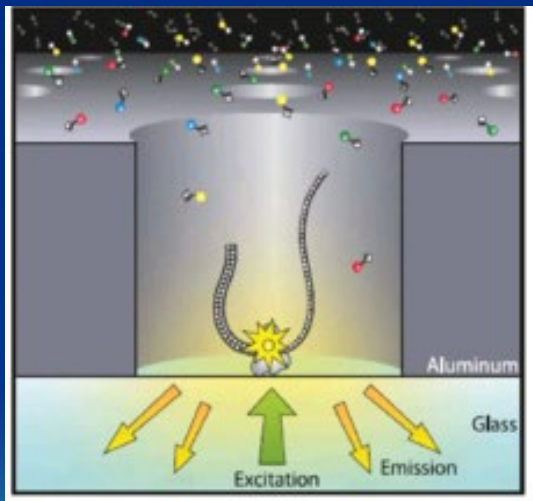
表10 新一代DNA测序平台

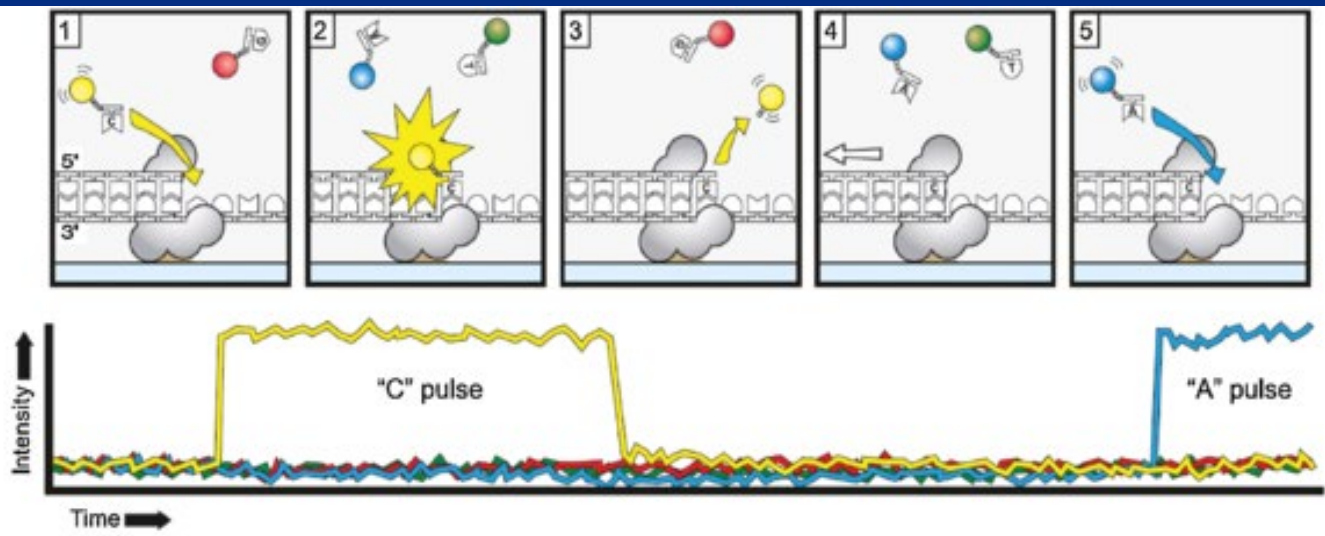
新一代测序仪名称	模板扩增方法	合成测序方法中使用到的酶和试剂	测序费用（美元）/Mb	设备价格（美元/台）	是否末端配对	发生1°错误的突变类型	测序长度
454	微乳液PCR	聚合酶（焦磷酸测序）	~60	500,000	是	插入或缺失突变	250bp
Solexa	桥式PCR	聚合酶（可被切除的终止基团）	~2	430,000	是	替换突变	36bp
SOLiD	微乳液PCR	连接酶（由双碱基编码技术设计的8bp长度探针）	~2	591,000	是	替换突变	35bp
Polonator	微乳液PCR	连接酶（九聚物）	~1	155,000	是	替换突变	13bp
HeliScope	单个分子检测	聚合酶（非同步延长）	~1	1,350,000	是	缺失突变	30bp

注：DNA测序领域的快速发展使得对各类测序方法的价格及读长的评估在很短时间内便失去意义。Roche Applied Science、Illumina及Applied Biosystems公司目前都在不断推出新的产品。表中列出的测序费用只是对使用的反应试剂费用的一个估算。测序长度指的是单链长度。

2、Single Molecule Real-time Technology：SMRT

单分子实时DNA测序





Flusberg B A, Webster D R, Lee J H, et al., 2010

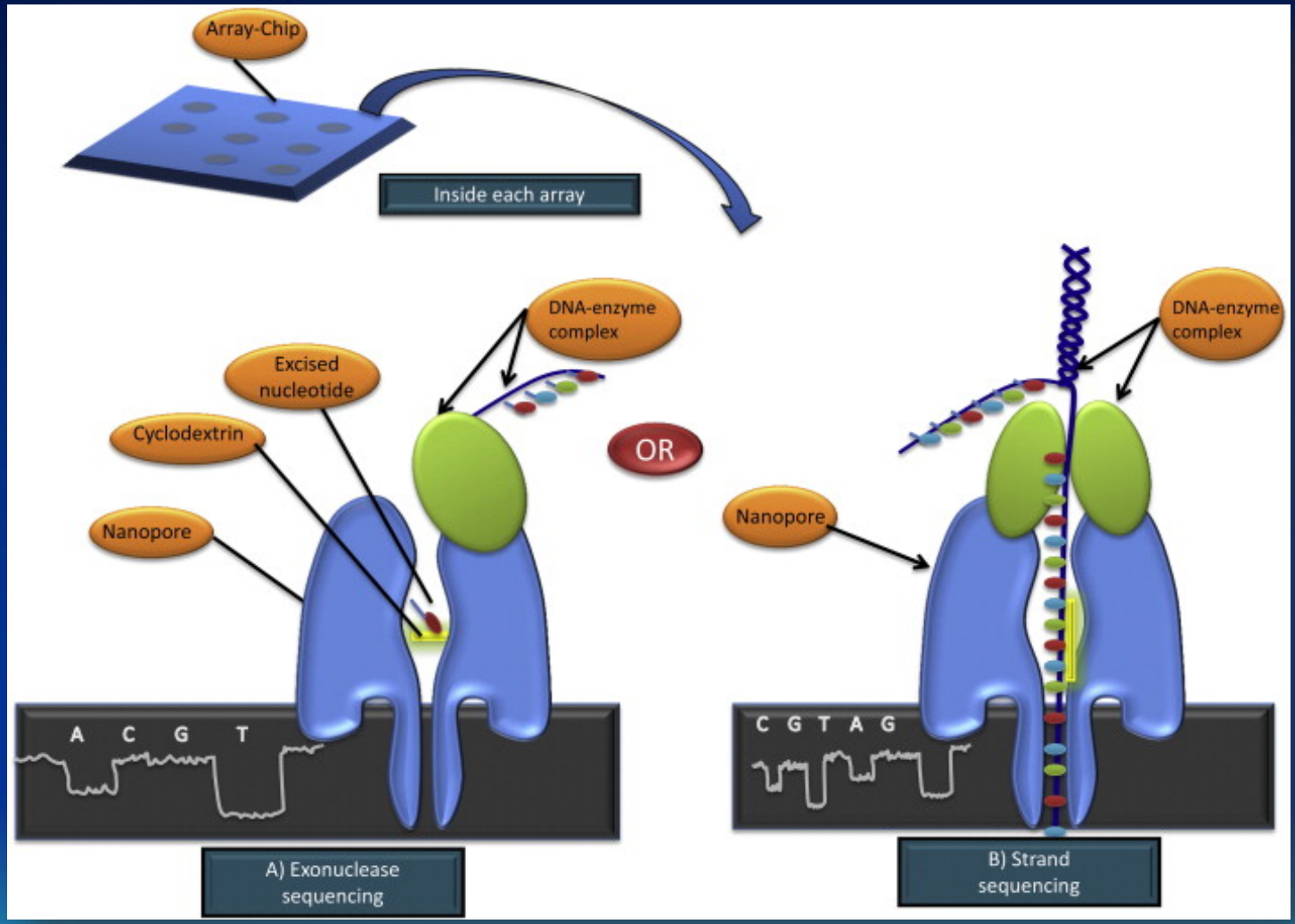
3、Oxford Nanopore Technologies

NANOPORE SEQUENCING PLATFORM

其核心是将在某一面上含有一对电极的特殊的脂质双分子层置于一个微孔之上，该双分子层中含有很多由α溶血素蛋白组成的直径约1.5nm的纳米孔中，结合一个核酸外切酶。当DNA模板进入孔道时，孔道中的核酸外切酶会“抓住”DNA分子，顺序剪切掉穿过纳米孔道的DNA碱基，每一个碱基通过纳米孔时都会产生一个阻断，根据阻断电流的变化就能检测出相应碱基的种类，最终得出DNA分子的序列.

人们使用硅和其他电子材料来制作纳米孔，使用微小的晶体管来控制通过小孔的电流，当4种不同的碱基通过小孔时，它们对电信号影响的方式不同，导致电压上升或者下降，可以依据这种微小的改变来确定通过DNA上碱基的类型。

纳米孔单分子测序技术的一大优势就是仪器构造简单使用成本低廉；因为它不需要对核苷酸进行标记，也不需要复杂的光学探测系统（如激光发射器和CCD信号采集系统等），能直接对RNA分子进行测序。同时由于它是直接检测每一个碱基的特征性电流，因而能对修饰过的碱基进行测序，这一点对于表观遗传学研究具有极高的价值；缺点就是它采用的是水解测序法，不能进行重复测序，因而无法达到一个满意的测序精确度



Wheeler D A, Srinivasan M, Egholm M, et al. 2008

nanopore.mp4

Nanopore Technologies

	测序方法/平台	方法/酶	测序长度	每个循环的数据产出量	每个循环耗时	主要错误来源
第一代测序技术	Sanger/ABI3730 DNA Analyzer	Sanger法/DNA聚合酶	1000 bp	56 Kb		
第二代测序技术	454/GS FLX Titanium Series	焦磷酸测序法/DNA聚合酶	400 bp	400-600 Mb	10 h	插入、缺失
	Solexa/Illumina Genome Analyzer	边合成边测序/DNA聚合酶	2*75 bp	20.5-25 Gb	9.5 d	替换
	SOLiD/SOLiD 3 system	连接酶测序/DNA连接酶	2*50 bp	10-15 Gb	6-7 d	替换
第三代测序技术	Heliscope/Helicos	边合成边测序/DNA聚合酶	30-35 bp	21-28 Gb	8 d	替换
	SMRT	边合成边测序/DNA聚合酶	100000 bp			插入、缺失
	Nanopore Sequencing	电信号测序/核酸外切酶	无限长			缺失
	FRET Sequencing	FRET/DNA聚合酶				缺失
	The Ion Torrent Sequencing	半导体/DNA聚合酶				缺失

新一代测序方法的应用主要有

（1）基因组*de novo* 测序 基因组从头测序（*de novo* sequencing）是在没有参考基因组的情况下，对某种物种进行基因组测序及组装，从而构建出该物种的全基因组组装序列图谱，即完整的物理图谱。

（2）基因组重测序 重测序（Re-sequencing）是对已知基因组序列的物种进行不同个体的基因组测序，并在此基础上对个体或群体进行差异性分析。

（3）简化基因组测序 简化基因组测序（Reduced-representation sequencing）技术是指利用生物信息学方法，设计标记开发方案，筛选特异性长度片段，应用高通量测序技术获得海量标签序列来充分代表目标物种全基因组信息的测序方法。目前简化基因组测序技术主要有：

①利用限制性位点关联的DNA标志物测序（restriction-site associated DNA tags，RAD-seq）技术进行简化基因组测序，通过筛查两侧带有特定核酸内切酶识别位点的DNA短片段（RAD标签）进行遗传变异分析，从而降低基因组测序的复杂度，快速鉴定出成千上万的单核苷酸多态性（SNPs）标记。RAD-seq 技术在模式和非模式生物的基因型-表型关联图谱、系统进化、群体遗传等遗传分析研究领域具有广泛的应用前景

②以高通量测序为基础的SLAF-seq (specific-locus amplified fragment sequencing/ specific length amplified fragment sequencing)技术，以较低的花费实现全基因组范围的标记开发和分型，是目前最热门的标记分型技术之一。利用SLAF-seq对个体或者群体DNA混池进行标记分型，可以构建高密度遗传图谱、进行集群分离分析，从而实现性状相关标记筛选、基因定位的目的等。

（4）转录组测序　转录组是指特定细胞在某一功能状态下所能转录出来的所有RNA的总和。转录组研究是基因功能及结构研究的基础。转录组测序技术是把mRNA、smallRNA和非编码 RNA等用高通量测序技术把它们的序列测出来，全面快速地获取某一物种特定器官或组织在某一状态下的几乎所有转录本，反映出它们的表达水平。

10.4 基因组图谱构建与应用

10.4.1 遗传标记

遗传学中曾将可识别的等位基因称为遗传标记（genetic marker）。现代遗传学将可示踪染色体、染色体片段、基因等传递轨迹的遗传特性也称为遗传标记。

1.形 态 标记： 是指那些能够明确显示遗传多态的外观性状，如株高、粒色等的相对差异

2.细胞学标记： 是指能够明确显示遗传多态的细胞学特征。如染色体结构上和数量上的遗传多态性等。

3.蛋白质标记： 非酶蛋白质和酶蛋白质

在非酶蛋白质中，用得较多的是种子贮藏蛋白；

酶蛋白质主要是同功酶。

4.DNA 标 记： 也称DNA多态性标记、 DNA分子标记，

是 DNA水平上遗传多态性的直接反映。

这里仅介绍遗传的分子标记中的几种DNA分子标记:

RFLP、VNTRs、AFLP、RAPD、SSR、STS和SNP

（1） RFLP标记

(restriction fragment length polymorphism)：

同一物种的亚种、品系或个体间基因组DNA 受到同一种限制性内切酶作用而形成不同的酶切图谱的现象，称为

限制性片段长度多态性（RFLP）。

这是由于基因组DNA某一位点上的变异有可能引起该位点特异性的限制性内切酶识别位点的改变，包括原有位点的消失或出现新的酶切位点，致使酶切片段长度随之发生变化而产生。

对RFLP的检测主要是用Southern杂交的方法进行

基本流程：

组织或细胞→基因组DNA→限制性内切酶消化→琼脂糖凝胶电泳→印迹转移至滤膜→预杂交→加入探针→杂交→洗膜→放射自显影。

RFLP标记的主要特点是：

（1）遍布于整个基因组，低拷贝序列；

（2）无表型效应，不受发育阶段及器官特异性限制；

（3）共显性，可区分纯合子和杂合子；

（4）结果稳定、可靠；

（5）DNA需要量大，检测技术繁杂，难以用于大规模的育种实践中。

用途：RFLP主要用于群体水平和系统发育研究上

进行：个体识别；绘制遗传图谱；目的基因定位；检测群体内或群体间序列差异程度；辅助育种等

（2） VNTRs标记

一般将真核生物基因组长16～100个核苷酸为基本单元的串联重复序列称为小卫星，而将以2～ 6个核苷酸为基本单元的简单串联重复序列，例如（CA）n、（GAG）n、（ GACA）n 等，称为微卫星或简单序列重复（ simple sequence repeats，SSR） 。小卫星和微卫星其多态性来源于重复序列的核苷酸组成和重复的次数不同。一般又将小卫星和微卫星称作可变数目串联重复（ variable number of tandem repeats，VNTRs）。

VNTRs 标记也呈共显性遗传，符合孟德尔遗传传递规律，因此可以用来进行遗传分析和作图（图18－3） 。而且VNTRs在基因组中有广泛的分布，可检出的多态更丰富，出现的频率更高。

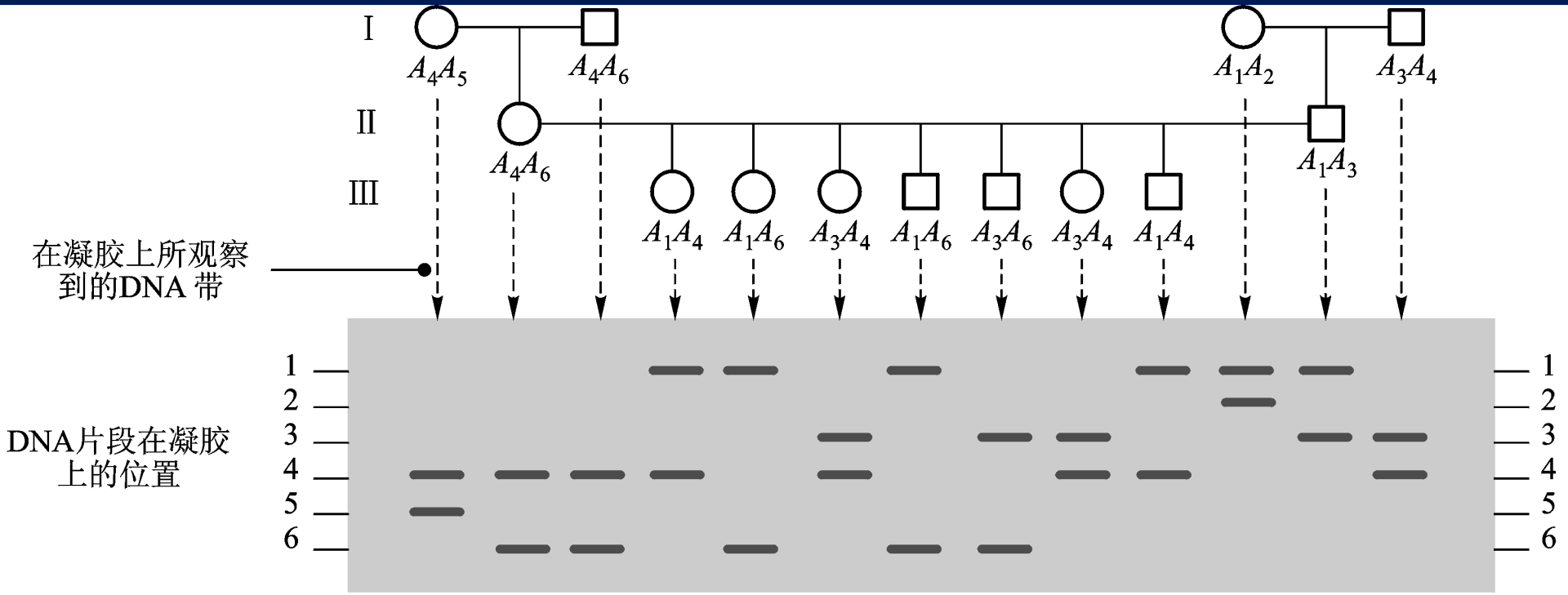


图10－10　人VNTRs的系谱分析（引自Hartl等，2001）

人类系谱分析表明VNTR等位基因的分离。在系谱中出现6个等位基因（ A_1 ～ A_6），但是任何一个个体仅可能有一个等位基因（纯合子）或两个等位基因（杂合子）

（基于PCR的DNA标记：随机引物PCR标记）

（3）RAPD（random amplified polymorphic DNA）标记：

随机扩增多态性DNA，用随机短引物（人工合成的核苷酸）进行DNA的PCR扩增。所扩增的DNA区段是事先未知的，具有随机性和任意性，因此随机引物PCR标记技术可用于对任何未知基因组的研究。（RAPD 原理）

（基于PCR的DNA标记：特异引物PCR标记）

（3）SSR（simple sequence repeats)标记：

简单重复序列多态性，又称微卫星DNA多态性，即由二核苷酸，三核苷酸或四核苷酸串联重复的拷贝数目不等而出现的多态现象。

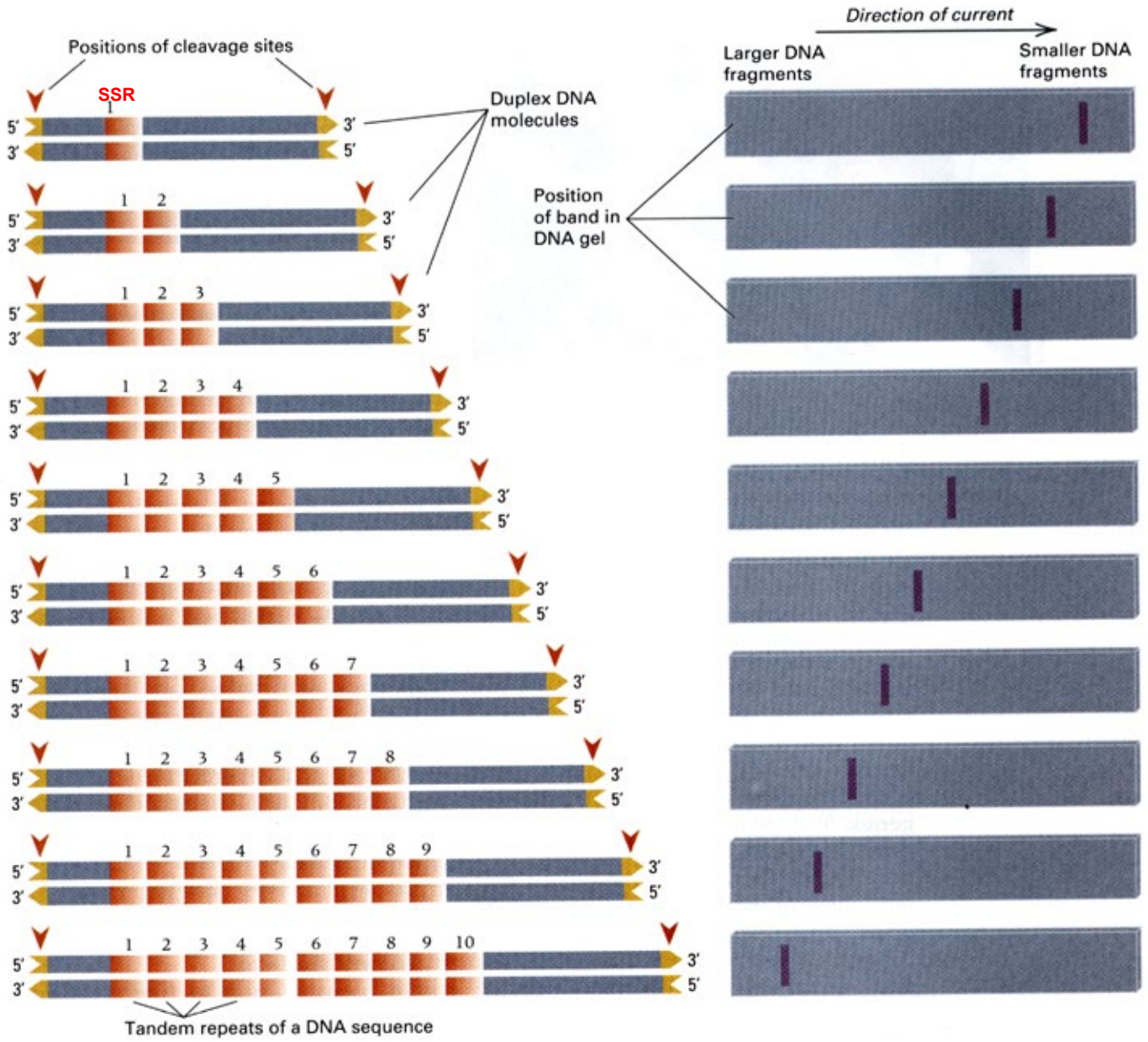


Figure 2.28 In a simple tandem repeat polymorphism (STRP), the alleles in a population differ in the number of copies of a short sequence (typically 2–60 bp) that is repeated in tandem along the DNA molecule. This example shows alleles in which the repeat number varies from 1 to 10. Cleavage at

典型的孟德尔式遗传；共显性（两个亲本的性状在一个个体中同时出现，没有显隐关系）。

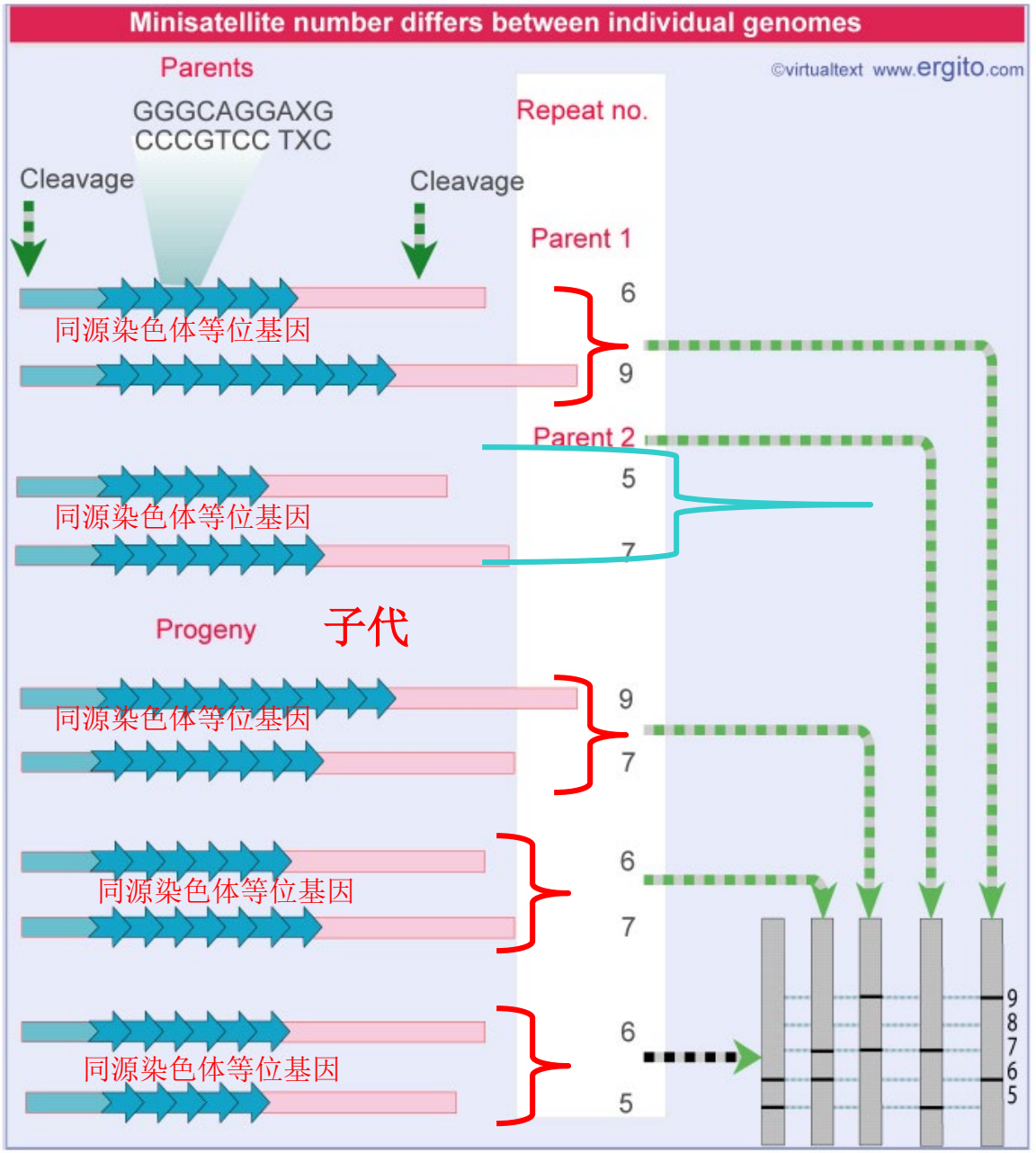


Figure 4.27 Alleles may differ in the number of repeats at a minisatellite locus, so that cleavage on either side generates restriction fragments that differ in length. By using a minisatellite with alleles that differ between parents, the pattern of inheritance can be followed.

SSR技术的优点是：

（1）在基因组中随机分布，检测的多态性频率高；
（2）PCR特异引物，重复性好；
（3）共显性，操作相对简单。

问题是：

（1）SSR需要测序和设计引物，因而需要大量的人力、物力和时间；
（2）另外其种属特异性强，开发所需的费用高昂，因此一些实验室进行了合作，共同开发微卫星引物

显然，利用SSR标记的关键是引物的设计。

对于已经进行基因组全序列测定的生物或遗传学研究比较经典的生物，可以直接从其基因组进行查找和利用有关程序进行引物的设计或利用别人已经发表的引物来进行实验；而对于没有基因组序列信息的生物，就需要进行有关引物的设计工作。

SSR标记应用：

现已证明微卫星DNA 存在于绝大多数真核生物基因组中，因此已广泛应用于遗传图谱构建、品种指纹图谱绘制、品种纯度检测及目标性状分子标记筛选、遗传多样性分析、重要性状基因的定位等

（4）、AFLP技术的原理和应用

荷兰Keygene公司Zabeau等（1992）、Zabeau和Vos等（1993）发展了一种检测DNA多态性的新方法——AFLP（ Amplified fragment length polymorphism ）：

RFLP＋接头＋引物＋PCR→ AFLP

AFLP既是进行酶切位点分析又要对酶切片段进行选择性扩增，因此它既具备RFLP的准确性又具有PCR的高效性，综合了两者的优点，并于1993年获得欧洲专利局专利。

AFLP原理图

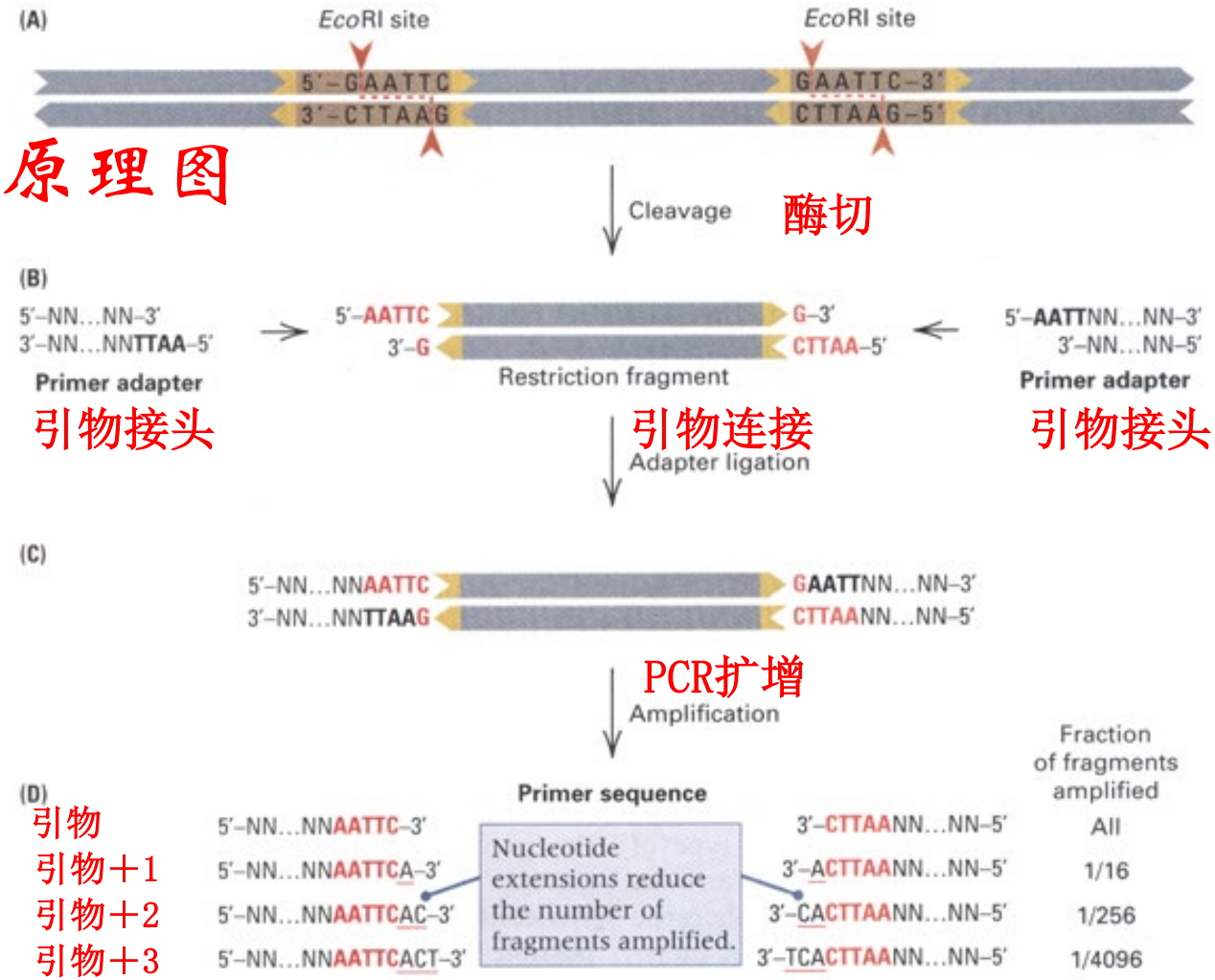
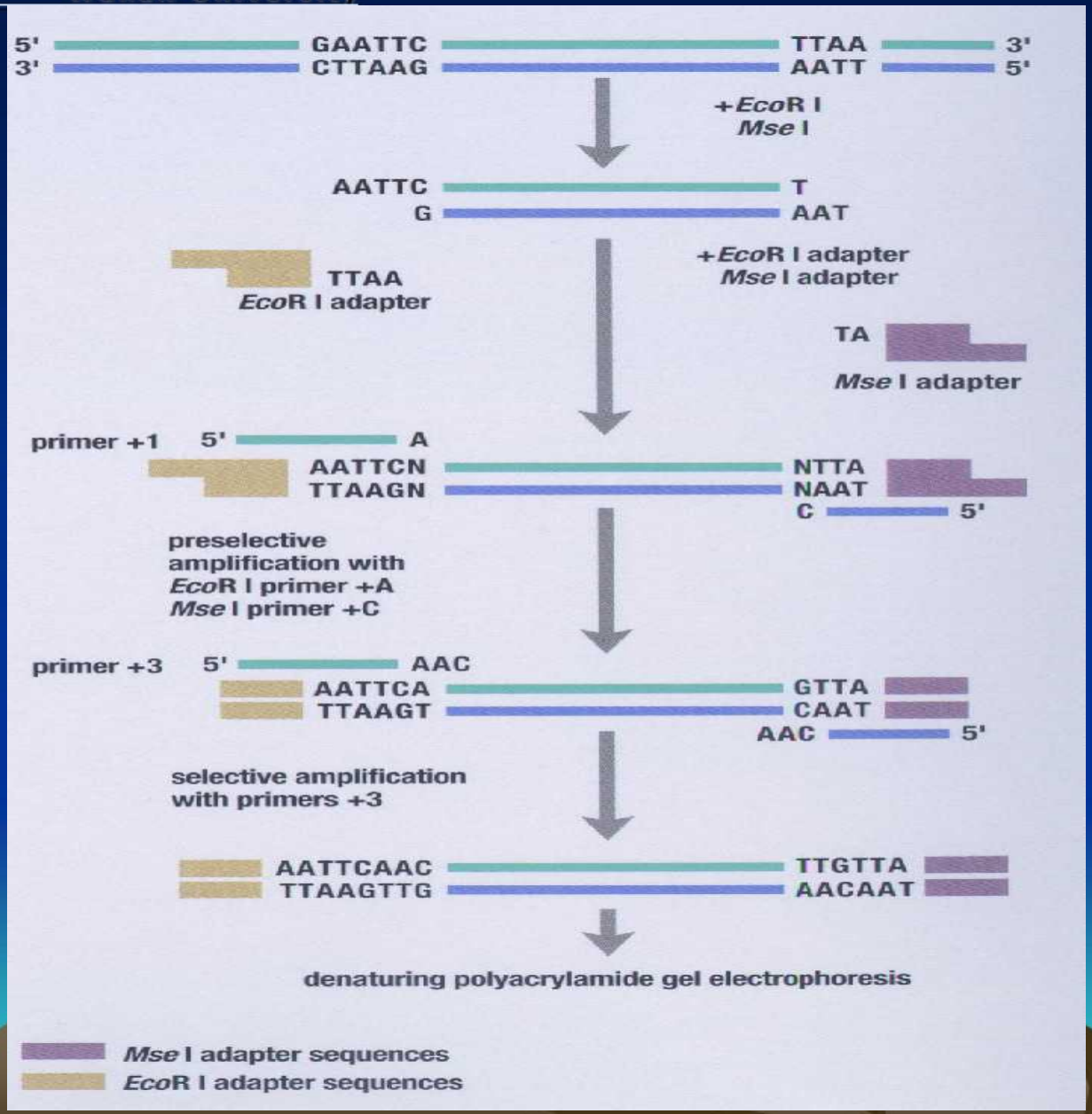


Figure 2.27 An amplified fragment length polymorphism (AFLP). (A and B) Genomic DNA is digested with one or more restriction enzymes (in this case, *Eco*RI). (C) Oligonucleotide adapters are ligated onto the fragments; note that the single-stranded overhang of the adaptors matches those of the genomic DNA fragments. (D) The resulting fragments are subjected to PCR using primers complementary to the adapters. The number of amplified fragments can be adjusted by manipulating the number of nucleotides in the adapters that are also present in the primers.

AFLP：扩增片段长度多态性

通过对基因组DNA酶切片段的选择性扩增来检测DNA酶切片段长度的多态性。

AFLP揭示的DNA多态性是酶切位点和其后的选择性碱基的变异。



为了对扩增片段的大小进行灵活的调节，一般采用两个限制性内切酶。一个是切点多的酶，如具有4bp识别位点的*Mse* I，它产生较小的DNA片段；另一个是切点少的酶，如具有6bp识别位点的*Eco*R I，它产生较大的DNA片段。

AFLP技术的优点：

（1）酶切所需DNA量少（仅需50－300ng），也可以作用于mitDNA；

（2）多态性检测能力高，既能检测酶切位点不同造成的多态性，又能检测随机选择 碱基造成的多态性；

（3）简单高效，一次选择扩增就能比较几十甚至上百个位点；

（4）分辨率高，扩增片段短，片段多态性检出率高；

（5）可靠性与重复性好。AFLP由于是特定引物扩增，退火温度高，因而假阳性低，重复性好。

AFLP集RFLP、PCR和RAPD的优点于一身，并且克服了RFLP和RAPD的缺点。

AFLP技术的不足：

（1）AFLP技术受到专利保护，试剂盒昂贵，成本较高；

（2）对DNA的纯度和内切酶的质量要求较高，技术要求高。现在一般都采用银染和荧光标记的方法来代替放射性同位素标记，因而避免了辐射伤害。

（3）操作繁琐。

AFLP的应用

（1） 分子进化研究：分析同种植物DNA的差异和亲缘关系；鉴定分类关系

（2） 分析生物发育不同时期基因的表达差异情况：cDNA-AFLP；

（3）构建遗传图谱；

（4）遗传多样性分析；

（5）植物雄性不育系与保持系关系的研究等

（5） STS标记:

序列标签位点（ sequence－tagged site，STS）是在染色体上定位的、序列已知的单拷贝DNA短片段，一般长200－500bp。

STS标记的原理：是根据单拷贝的DNA 片段两端的序列，设计一对特异引物，经PCR扩增基因组DNA而产生的一段长度为几百bp的特异序列。由于不同的STS序列在基因组中往往只出现一次，从而能够界定基因组的特异位点。用STS进行物理作图，可通过PCR或杂交途径来完成。STS标记可以作为比较遗传图谱和物理图谱的共同界标，因此在基因组作图上具有非常重要的作用。

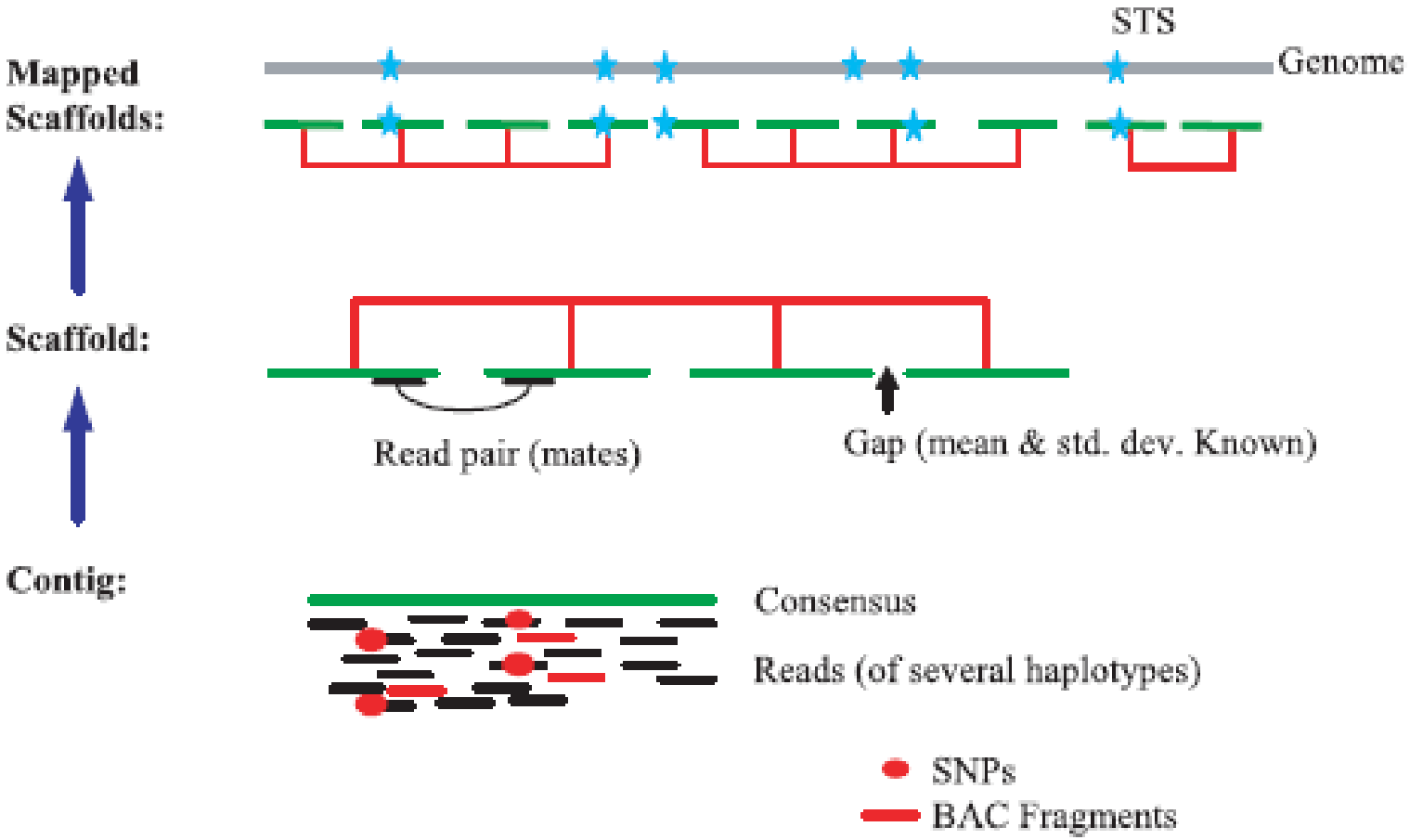


Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

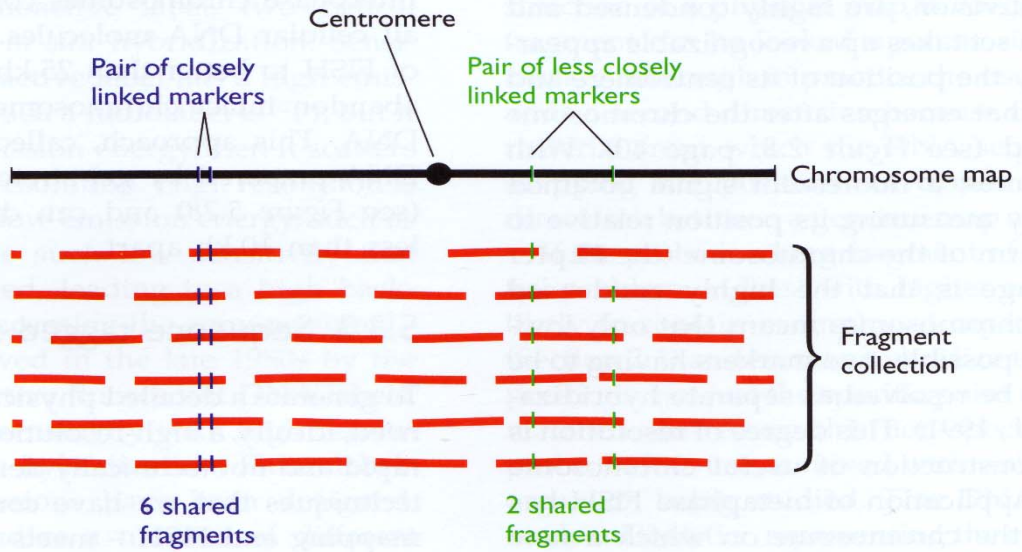


Figure 5.31 A fragment collection suitable for STS mapping.

The fragments span the entire length of a chromosome, with each point on the chromosome present in an average of five fragments. The two blue markers are close together on the chromosome map and there is a high probability that they will be found on the same fragment. The two green markers are more distant from one another and so are less likely to be found on the same fragment.

（单核苷酸多态性的DNA标记）

（6） SNP（single nucleotide polymorphism）

标记:单核苷酸多态性。

同一物种不同个体基因组DNA的等位序列上单个核苷酸存在差别的现象。或者説不同个体基因组DNA序列同一位置上的单个核苷酸的差别。其比较的不是DNA的片段长度，而是相同序列长度里的单个碱基的差别。 (SNP)

SNP是指基因组内特定位置上存在两种不同的碱基，是二等位多态性，其中最少一种在群体中的频率不小于1%；如果出现频率低于1%，则视作点突变。

SNP只涉及单个碱基的变异，这种变异可以由单个碱基的转换（包括C与T互换，在其互补链上则为G与A互换），或颠换（包括C与A、G与T、C与G、A与T互换）引起，也可以由碱基的插入或缺失所致。但是，通常所说的SNP并不包括后两种情况。

根据SNP在基因中的位置，SNP可分为：

基因编码区SNP（coding SNP, cSNP）

基因周边SNP（peripheral SNP, pSNP）

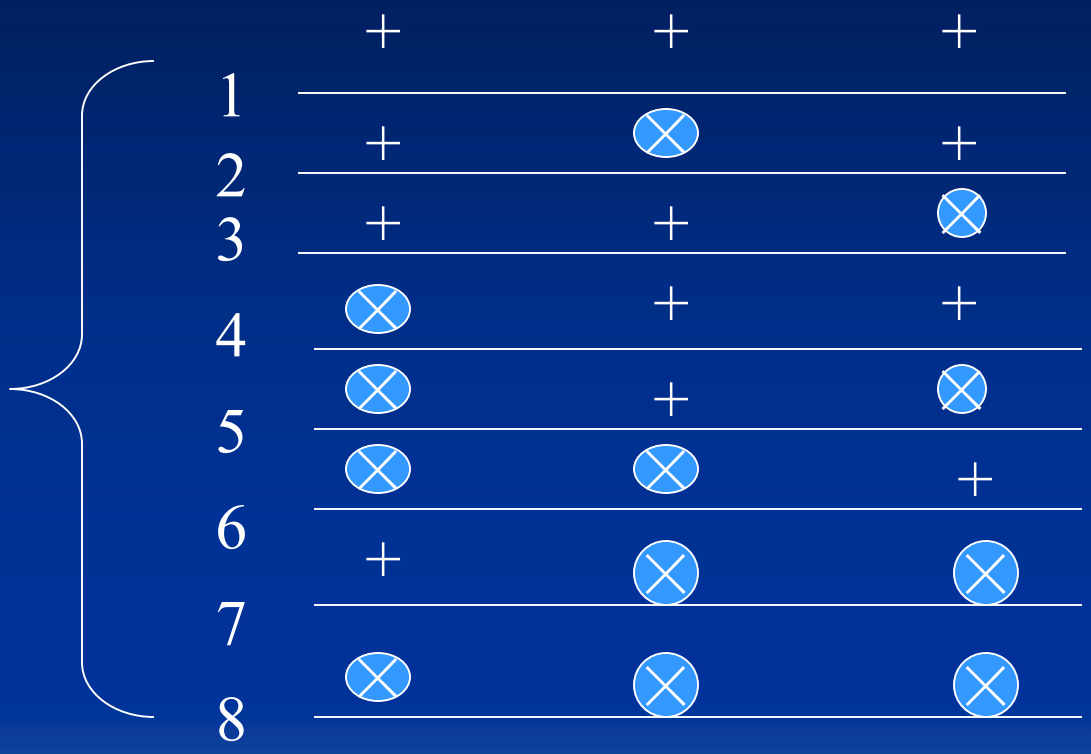
基因间SNP （intronic SNP, iSNP）

尽管单一的SNP所提供的信息量远小于现在常用的分子标记，但SNP的数量极其丰富，并且可以自动化检验，因此其具有广泛的应用前景。

例如：如果假定1/1000的碱基是多态的话，那么人类30亿碱基当中有约300万SNP位点。由此可见，SNP数量比微卫星标记数要高出几个数量级。

例如：3－4个SNP这种相邻的界标构成的单体型（haplotype）就可以有8－16种：

3个SNP可以形成的单倍型8种



Haplotype:又称“单倍型”或“单元型”。一条同源染色体上的等位基因或遗传标记所构成的组合。

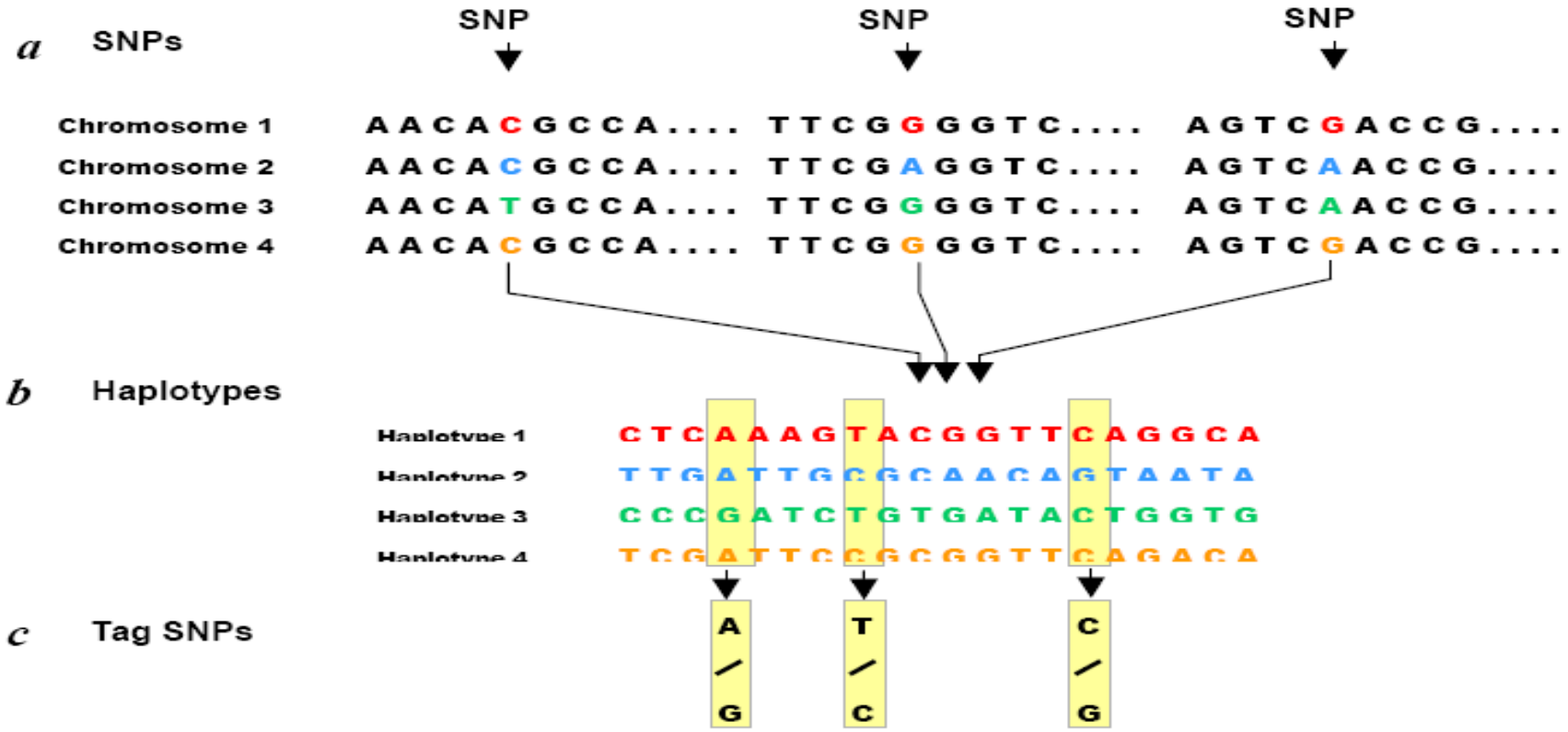


图 1　单体型图示意。a　一段 DNA 上的 SNP 位点。来自不同个体的四条染色体的这一段 DNA 的大部分序列完全相同但有 3 个位点表现出差异。每个 SNP 有两个可能的等位位点（alleles）。b 单体型。一个 Haplotypes 由临近的一系列 SNP 等位位点组成，图中显示从一个 6000 bp DNA 片断上检定出的 20 个 SNPs 的排布类型，包括 a 中的 3 个位点（第二排箭头）。c 标签 SNPs。只要检定分型 20 个 SNPs 中的 3 个，就足以确定这一段序列的单体型。例如，如果测出一个染色体的标签 SNPs 是 A - T – C，则这一序列是单体型 1。（转引自 The International HapMap Consortium. The International HapMap Project, Figure 1. *Nature*, 2003, in press）

10.4.2 遗传图谱（Genetic map）

（1）**经典的遗传学图谱**：主要用来确定生物体的基因在染色体上的排列，只能标明基因之间的相对位置，无法指明基因在染色体上的具体位置，因此无法按图直接克隆分离基因。显然，在人类基因组计划中利用价值不大。

（2）现代遗传学图谱：David Botstein等 于1980年提出，利用RFLP作为标记去构建多态性基因与这些标记连锁关系，进而确定多态性基因的位置。其精髓在于将单纯的表型研究深入到DNA分子的本质上去。当时，由于DNA限制性内切酶和连接酶的应用，RFLP成为一种崭新的DNA多态性标记。

即：表型多样性对应于DNA分子的多样性

将单纯的表现多态性的界标改变为以DNA序列的多态作为作图界标。这些界标包括： RFLP、 VNTR和STR等。

10.4.3 物理图谱

物理图谱（physical map）是指各遗传标记之间或DNA序列两点之间，以物理距离来表示其在DNA分子上的位置而构成的位置图，以实际的碱基对或千碱基对或百万碱基对长度来度量其物理距离。

最早的物理图谱是细胞遗传学图谱，通过原位杂交将基因定位在染色体各区带上。而现在HGP中以YAC 或BAC为载体构建的连续克隆系覆盖人的每条染色体的大片段DNA。以YAC 叠连群或BAC 叠连群作为大尺度物理图谱，同时寻找分布于人类整个基因组的序列标签位点STS。STS是物理作图通用语言，最新制作的物理图包含了52 000个STS位标。以STS为基础的图谱最大的优点是：适合于大规模测序，并很容易在染色体上定位。

10.4.4 人类基因组遗传图谱的构建

（1）经典遗传图谱的构建

首先要对用于作图的基因是否连锁进行判断。在人类遗传学研究中，通常通过适当的统计模型来估算重组率，并采用或然比或似然比（likelihood ratio，r）检验的方法来推断连锁是否存在，即比较假设两个基因间存在连锁（$r<0.5$）的概率与假设没有连锁（$r=0.5$）的概率。该值的对数值称为LOD值（LOD score）。为了确定两对基因之间存在连锁，一般要求或然比大于1 000：1，即LOD＞3；而要否定连锁存在，则要求或然比小于100：1，即LOD＜2。在其他生物遗传图谱的构建中，或然比的概念也用来反映重组值的可靠程度或作为连锁是否真实存在的一种判断尺度（详见第6章）。

一旦若干连锁基因被确定，接下来是估计基因座位的图距，并排列出这些基因的相互远近关系，从而构建出一张基因的连锁图。

由上可知，经典遗传学图谱主要用来确定生物体的基因在染色体上的排列，只能标明基因之间的相对位置，无法指明基因在染色体上的具体位置，因此无法按这种图谱直接分离和克隆基因。因而，在人类基因组计划中利用价值不大。

（2）现代遗传学图谱的构建

现代遗传学图谱的概念是D.Botstein等于1980年提出来的，当时由于DNA限制性内切酶和连接酶的应用，RFLP成为一种崭新的DNA多态性标记。他们利用RFLP作为标记去构建多态性基因与这些标记连锁关系，进而确定多态性基因的位置。

其精髓在于将单纯的表型研究深入到DNA分子的本质上去

建立精细的人类遗传图谱的关键是获得足够多的高度多态性的标记

第一代用作遗传图谱的DNA遗传标记是RFLP；

第二代DNA遗传标记利用了存在于人类基因组中的大量重复序列；

第三代DNA遗传标记是SNP。SNP遗传标记分析完全摒弃了经典的凝胶电泳，代之以最新的DNA芯片技术或高通量测序技术，是目前发现的最好的遗传标记，而且已成为研究基因组多样性和识别、定位疾病相关基因的一种新手段。

10.4.5 植物基因组遗传图谱的构建

植物基因组遗传图谱的构建主要包括以下几个步骤：

（1）遗传标记的选择

（2）亲本的选配

（3）分离群体类型的选择与作图群体的构建

（4）群体大小的确定

（5）连锁图谱制作的统计学方法

（6）DNA标记分离数据的处理

（以上步骤详见第6章：6.3.2　QTL定位的基本步骤PPT）

10.4.6 物理图谱的构建

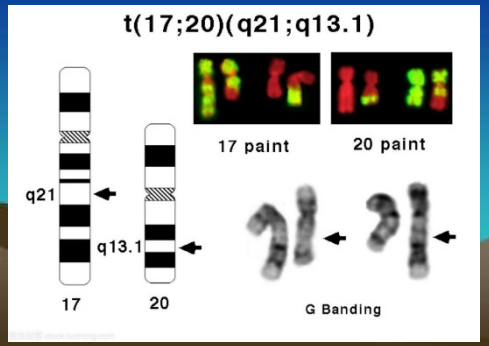
物理图谱有细胞遗传学图谱、限制酶酶切图谱、叠连群图谱及全基因组序列图谱等几种类型。

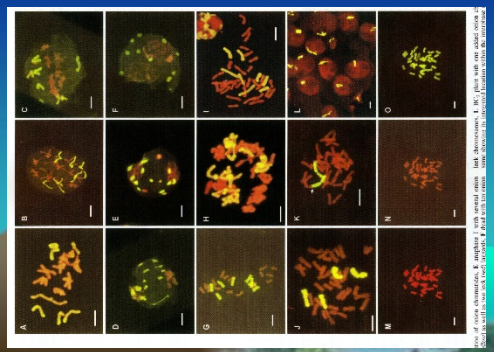
（1）细胞遗传学图谱

细胞遗传学图谱（cytogenetic map）是将基因或DNA片段直观定位于染色体上的物理图谱，因此也称为染色体图谱（chromosome map）。它是将基因或其他被分离出的DNA片段定位在它所在的染色体区域，并且粗略地测出它们之间相距的碱基长度。

细胞遗传学图谱的制作方法：原位杂交技术，将目标基因或特定的DNA片段定位到特定的染色体区带上。（DNA探针+目标区域）

荧光原位杂交技术、多彩色荧光原位杂交技术、基因组原位杂交技术

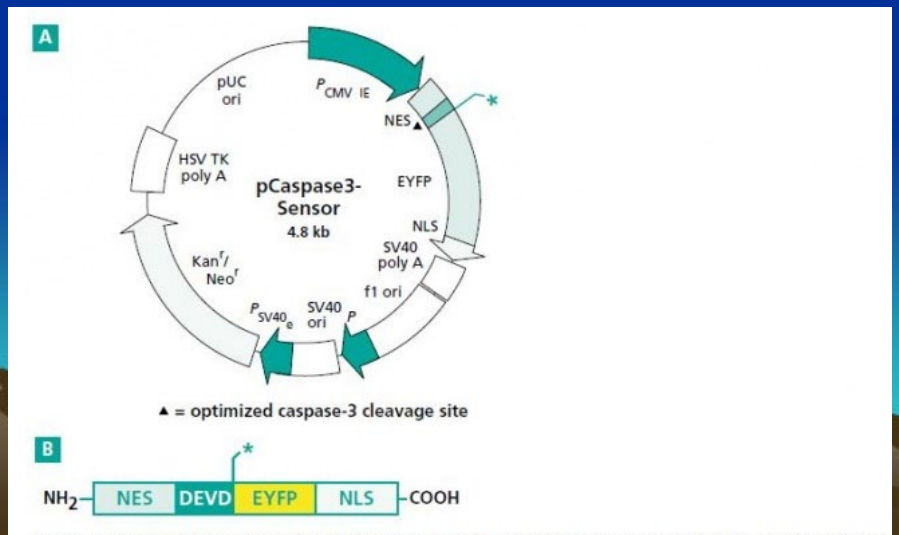


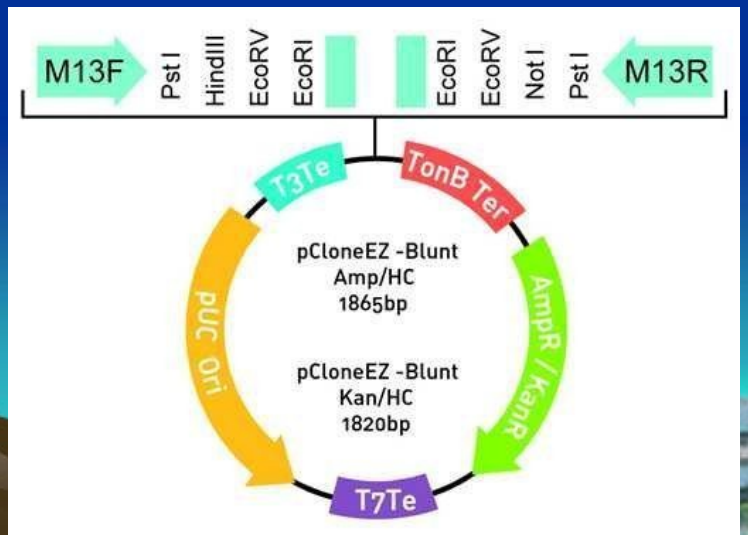


（2）限制酶酶切图谱

DNA限制性内切酶酶切图谱（restriction map）是一种重要的DNA物理图谱，它由一系列位置确定的多种限制性内切酶酶切位点组成，以直线或环状图式表示。限制性内切酶在DNA链上的切口是以特异序列为基础的，核苷酸序列不同的DNA，经酶切后就会产生不同长度的DNA片段，由此而构成独特的酶切图谱。

其主要的步骤是：制备克隆DNA；克隆DNA指纹分析，包括酶切、电泳分离、杂交标记或图像处理；计算机对克隆的排列和DNA片段的排序；填补空隙，包括分离新克隆、PCR验证等。

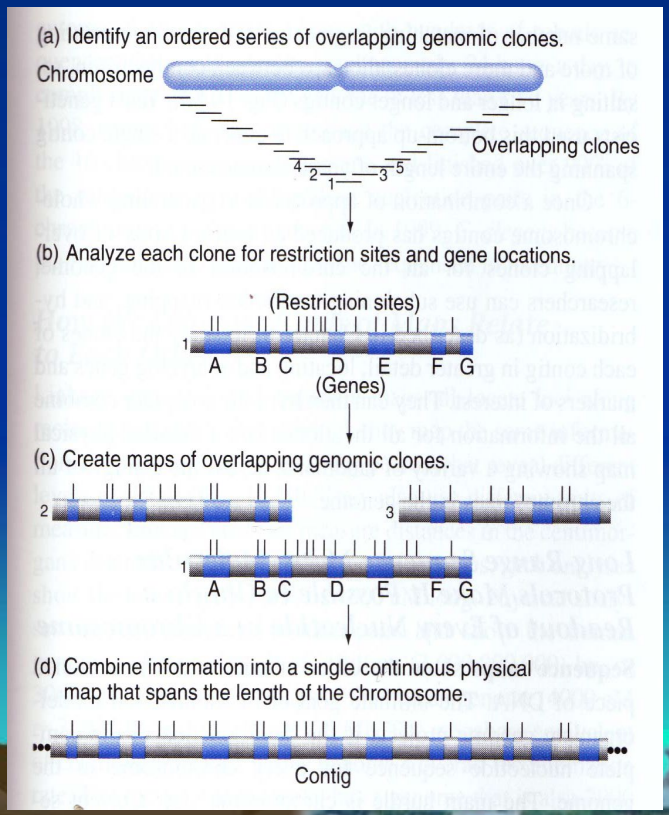




（3）叠连群图谱

一组相互两两头尾拼接的可装配成长片段的DNA序列克隆群称为叠连群。分辨在基因文库克隆中叠连群插入片段的顺序关系，通过相互邻接的两个片段间存在的重叠部分，推断出各叠连群覆盖整个染色体的克隆片段在染色体上的顺序，最后构建出叠连群图谱（contig map）。其具体步骤是将染色体切割成小片段后，克隆并排序，得到由排好序的插入DNA片段克隆组成的叠连群。

克隆叠连群的组建通常采用染色体步移（chromosomal walking）法。首先从基因文库中随机取出一个指定的或随机克隆，然后在文库中寻找与之重叠的第二个克隆。在第二个克隆的基础上再寻找第三个克隆，依次延伸，这便是克隆叠连群图谱构建的基本原理。



（4）基因组DNA大片段文库的构建

基因组DNA大片段文库的构建是建立基因组图的首项工作（详见e10-3）。需要构建的文库包括酵母人工染色体文库、细菌人工染色体文库、P1噬菌体衍生人工染色体文库以及FOSMID文库等。

（5）全基因组序列图谱

最精确的物理图谱是基因组全DNA序列图谱。现在一般采用新一代测序技术进行了全基因组测序，利用叠连群和DNA分子标记例如SSR、STS等从头组装。(100)

（6）物理图与遗传图的关系

两者之间既有联系又有差异(100)

遗传学图、细胞学图和物理图之间的相互关系如图10-12

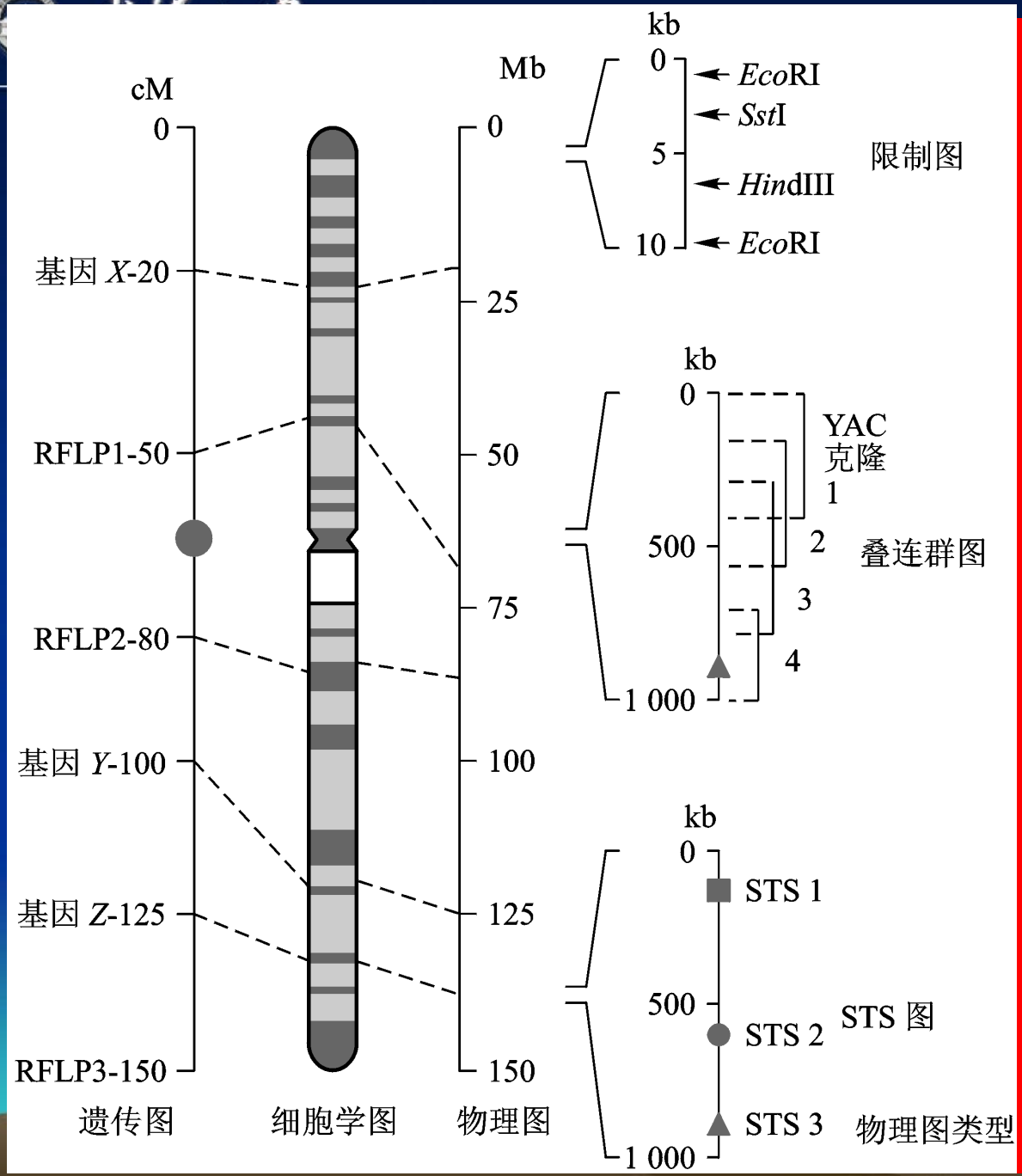


图10-12 三种图谱之间的相互关系

在现代分子标记图谱中，遗传图和物理图所表示的基因（和分子标记位点）在序列上基本相同，但在距离上是不相同的。这是由于遗传图确定的是基因或DNA 标记在染色体上的相对位置与遗传距离。物理图是指以物理距离来表示其在DNA 分子上的位置而构成的位置图。

(7)“位点”的概念：在经典的定义中对于“位点”的概念是基因即遗传位点。而目前基因组研究中通常染色体位点不仅是指基因，还包括客观物组成的染色体上的位点。

在物理图谱中，位点是指一系列的客观物：包括探针位点、限制性内切酶酶切位点、克隆位点、基因位点、中间粒及端粒位点等。

位点——染色体上任何可以区分的染色体位置。

10.4.7 基因组图谱的应用

(1) 寻找新的基因：从基因组序列中进行基因搜寻、分析与基因相关的序列，这是解读整个基因组图谱的基础。

(2) 基因的克隆与分离：根据饱和的基因组图谱，可以对基因进行克隆和分离。

(3) 基因功能的预测：利用基因组图谱还能够对基因功能进行预测，观察基因组作为一个整体如何行使其功能。基因功能的预测可以通过计算机预测和实验确定两种途径进行。

(4) 利用基因组图谱数据进行比较基因组学研究：通过基因组比较作图可以揭示染色体或染色体片段上同线性或共线性的存在，从而对不同物种的基因组结构及基因组进化历程进行精细分析。

(5) 基因定位：借助基因组图谱，可使基因定位在精度、深度、广度等方面有极大的提高。

10.5 比较基因组学和功能基因组学

10.5.1比较基因组学与进化

比较基因组学（comparative genomics）是一门通过运用数理理论和相应计算机程序，对不同物种的基因组进行比较分析来研究基因组大小与基因数量、基因排列顺序、编码序列与非编码序列的长度、数量及特征，以及物种进化关系等生物学问题的科学。最重要也是最能体现比较基因组学学科特点的是不同生物间全基因组的核苷酸序列的整体比较。

目前，通过比较基因组学的研究所获得的重要进展主要表现在（详见e10-4）：①加深了人们对基因组结构和功能的认识；②揭示了近缘物种中存在保守的连锁群或染色体“板块”结构；③进一步理解基因组的进化。同时促进了“系统生物学”（详见e10-5）和“单细胞基因组学”（详见e10-6）的形成和发展。

（1） 基因组比较原位杂交（Comparative genome in situ hybridization）/比较基因组杂交（Comparative genome hybridization,CGH）

将一个物种的基因组总DNA作为探针，与另一物种的染色体进行原位杂交，来检测两基因组同源性。不同种的比较基因组杂交在物种系统进化和亲缘关系研究上有重要作用，它可以直观地看出两基因组间的相似程度。

（Fig 7.6）

（2） 比较作图（Comparative mapping）

比较作图就是利用共同的分子标记（主要是cDNA标记及基因克隆）在相关物种中进行物理或遗传作图，比较这些标记在不同物种基因组中的分布情况，揭示染色体或染色体片段上同线性（synteny）或共线性（collinearity）的存在，，从而对不同物种的基因组结构及基因组进化历程进行精细分析。

例：用一套共同的cDNA克隆为探针对水稻（Oryza）、小麦(Triticum)、大麦(Hordeum)、黑麦（Secale）、燕麦（Avena）、珍珠谷（Pennisetum）、玉米（Zea）、高粱（Sorghum）、甘蔗（Saccharum）等进行比较作图发现，所有这些属（genera）相同的基因在大的染色体片段上以同样的顺序排列。所有这些作物含有一套相同的基因，它们在一些基因组中大量重复（玉米和小麦中更明显），通过易位和其它突变而进行重排，但仍然可识别为具有相同的一套祖先基因。

（Fig9.29 a-1）

（3）过比较基因组学的研究加深人们对基因组结构和基因功能的认识

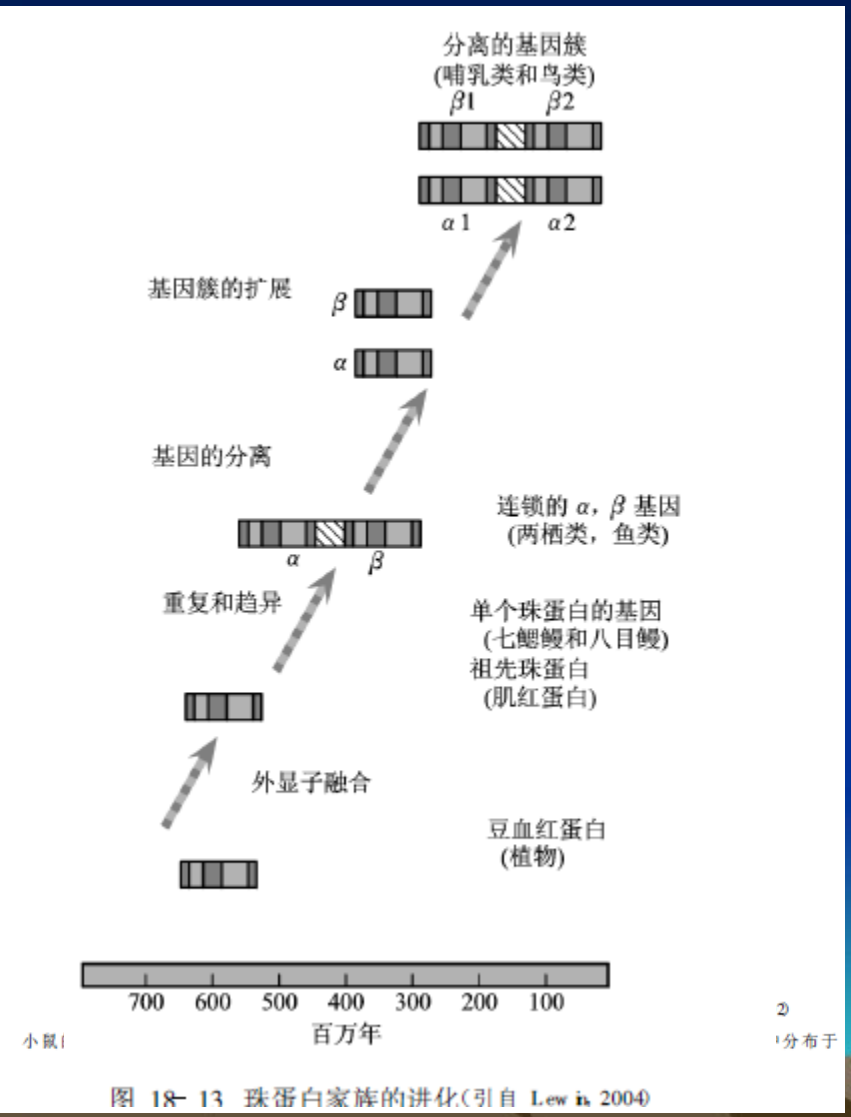


图 18-13 珠蛋白家族的进化(引自 Lewin, 2004)

（2）比较基因组学的研究表明在亲缘关系近的物种中存在保守的连锁群或染色体“板块”结构

（3）基因组的进化

①基因组大小的进化

②基因组的分子进化

10.5.2 功能基因组学

在功能基因组学的研究中通常运用高通量技术（high-throuput techniques），如DNA微阵列（DNA microarrays），反求遗传学技术如基因打靶（gene targeting），转基因（transgene）以及反义mRNA（antisense mRNA）和RNA干扰（RNA interference，RNA i）等技术来系统地分析基因功能及基因间相互作用、基因组的时空表达以及发现和寻找新基因等。

（1） DNA微阵列技术 又称DNA芯片（DNA chips）或基因芯片（gene chips）技术，它通过将对应于不同基因或cDNA的DNA片段或寡聚核苷酸点样于微芯片上形成高密度的矩阵，与荧光标记的总mRNA进行杂交，然后通过激光共聚焦扫描检测并运用计算机软件对杂交信号进行自动化定性定量分析，具有高通量、实时、灵敏、准确等特点。

武漢大學
Wuhan University

血液样本

DNA抽提

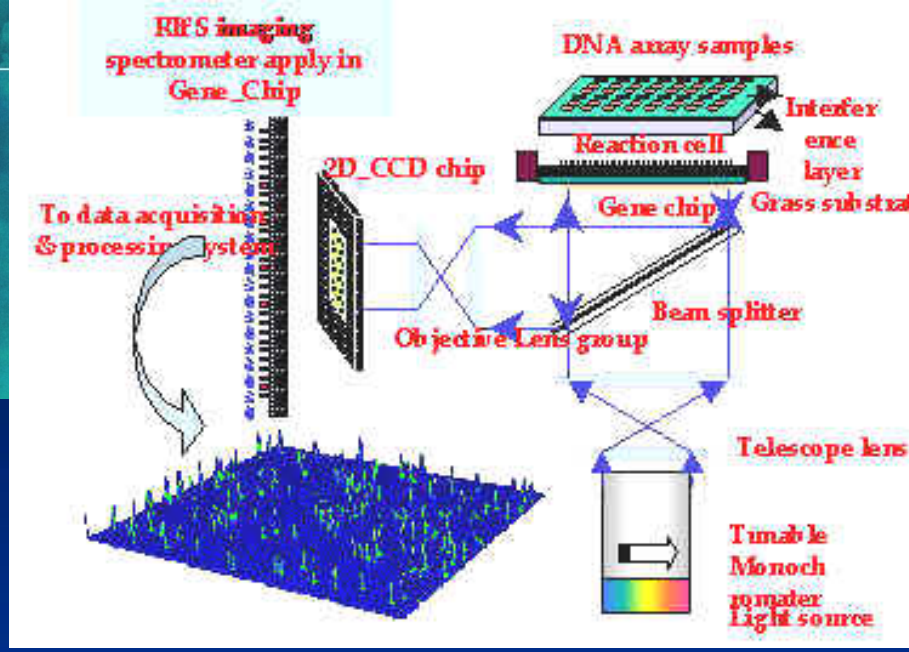
基因组 DNA

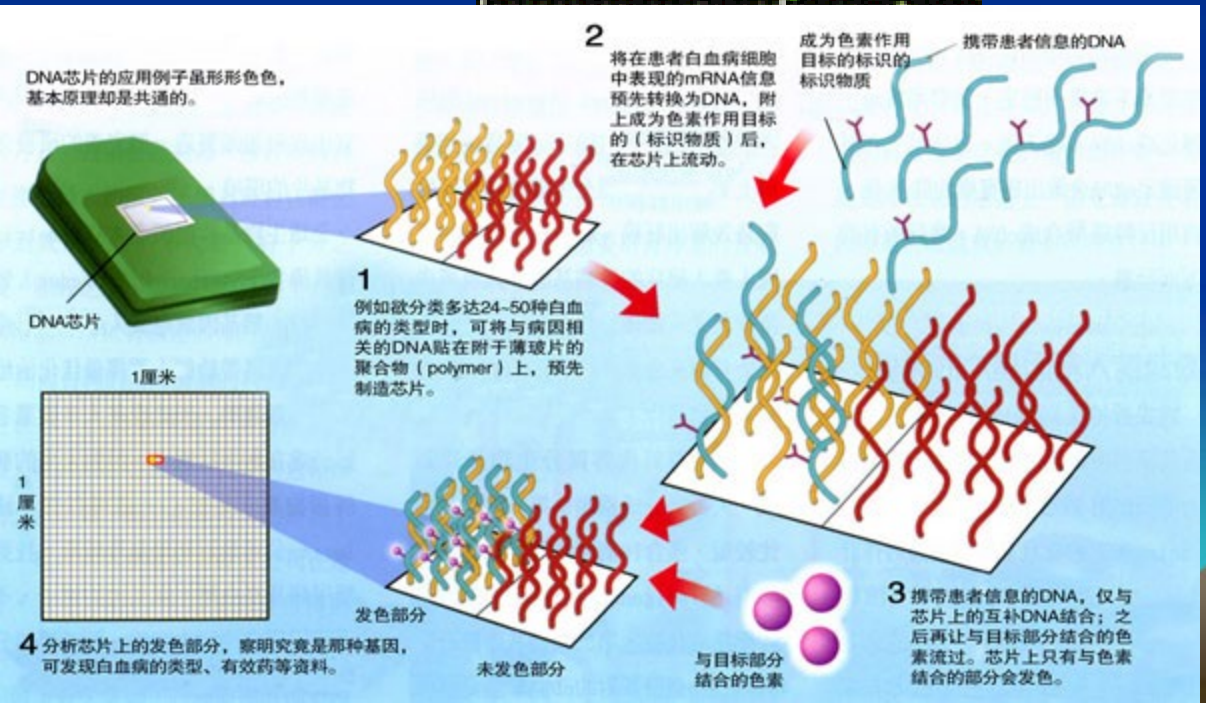
PCR扩增

荧光标记的 PCR 产物

芯片杂交

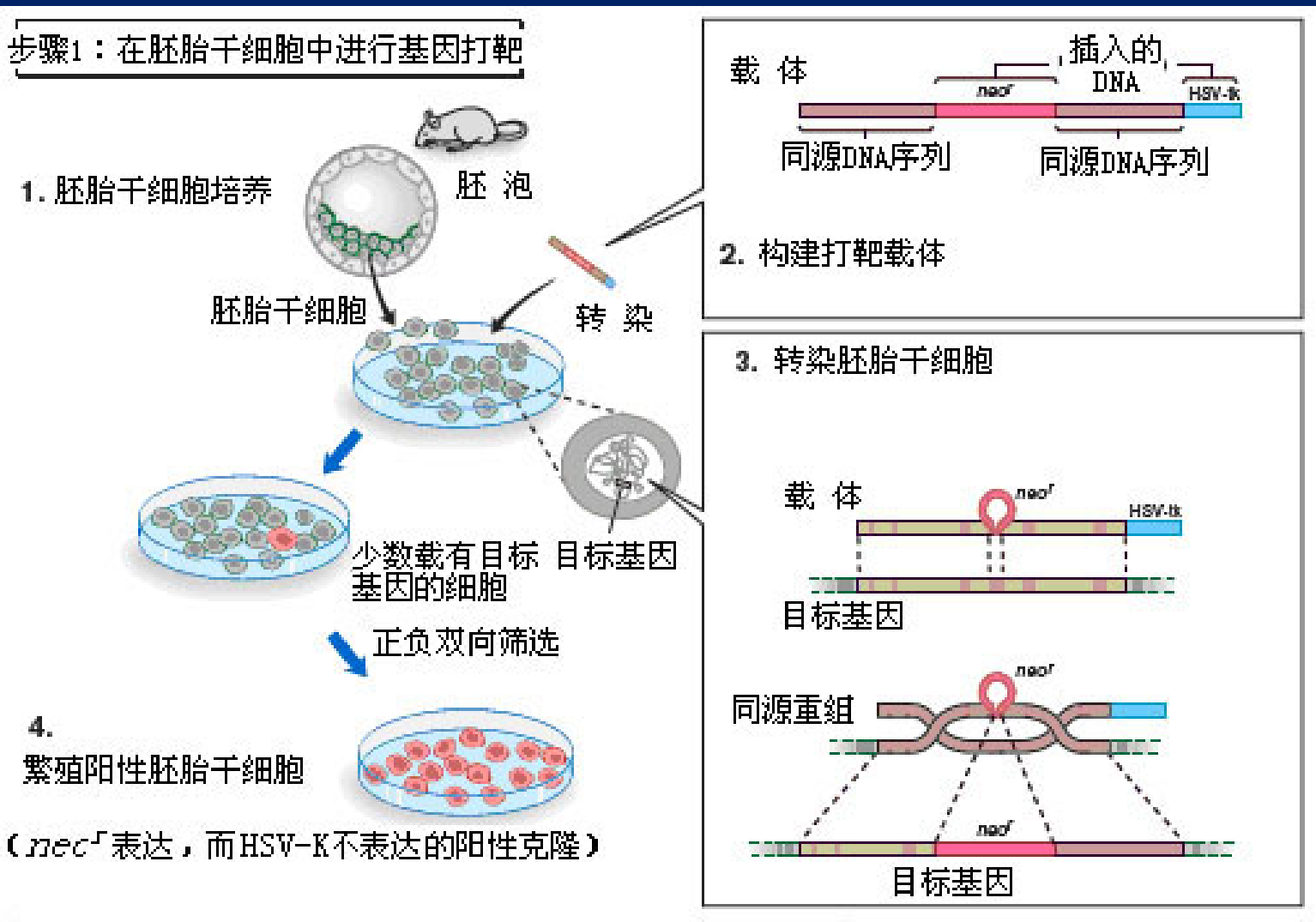
扫描分析





（2）基因打靶（gene targeting）

基因打靶是指通过转染的DNA序列与细胞内同源的基因组序列（靶序列）之间进行同源重组，以改变靶序列来研究其结构和功能或进行基因治疗的技术。



基因打靶技术包括基因敲除（knock-out）和基因敲入（knock-in），前者是用无功能的DNA序列与靶序列重组，破坏原基因组的遗传功能，后者是用有功能的DNA序列与受到破坏的靶序列重组使其恢复遗传功能。基因打靶技术是一种从基因到表型的新研究方法，属于反求遗传学范畴。

步骤2：从基因打靶胚胎干细胞到基因打靶小鼠

5. 向小鼠胚泡注入胚胎干细胞

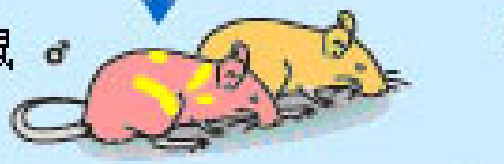
显微注射　　体外繁殖　　胚胎移植

6. 生育、繁殖嵌合体小鼠

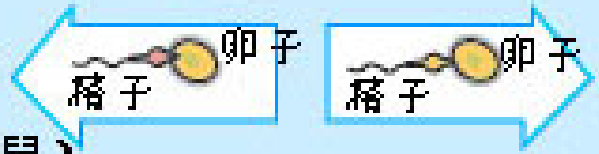
嵌合小鼠和正常小鼠交配
生育含有目的基因的小鼠
及正常小鼠后代

出生的嵌合小鼠

嵌合小鼠 ♂　　♀ 正常小鼠



基因打靶小鼠
（体内目标基因沉默—基因敲处小鼠）



正常小鼠

© The Nobel Committee for Physiology or Medicine　Illustrator: Annika Röhl

（3）反义mRNA技术是通过向细胞导入一段与特定编码mRNA互补的非编码RNA链，使其与该段mRNA特异性结合而定向阻抑靶基因表达的技术。

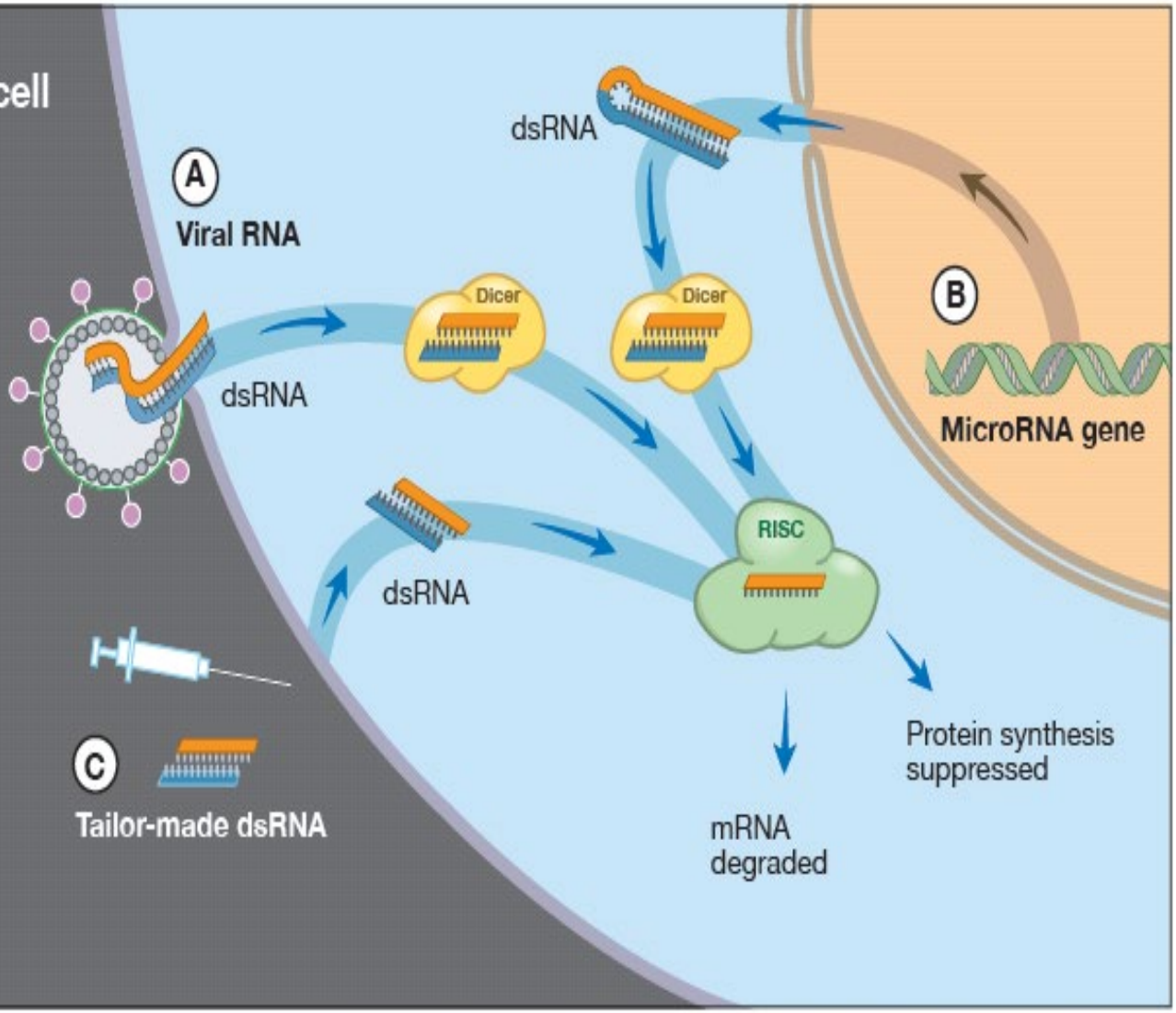
（4） RNA干扰(RNA interference，RNAi)是正常生物体内一些小的双链RNA可以抑制特定基因表达的一种现象。它是指当细胞中导入与内源性mRNA编码区同源的小的双链RNA (double stranded RNA，dsRNA)时，可以通过促使该mRNA降解来高效、特异的阻断体内特定基因的表达，从而导致基因表达沉默的现象，这种现象发生在转录后水平，又称为转录后基因沉默(post-transcriptional gene silencing，PTGS)。这些小的双链RNA称为siRNA （Small /short interfering RNA，siRNA）。

4. Several processes in the cell use RNAi

A. When an RNA virus infects the cell, it injects its genome consisting of double-stranded RNA. RNA interference destroys the viral RNA, preventing the formation of new viruses.

B. Synthesis of many proteins is controlled by genes encoding microRNA. After processing, microRNA prevents the translation of mRNA to protein.

C. In the research laboratory, dsRNA molecules are tailor-made to activate the RISC complex to degrade mRNA for a specific gene.



© The Nobel Committee for Physiology or Medicine Illustration: Annika Röhl

(5) 基因组编辑技术

1）RNA-guided genome editing：

①锌指核酸酶（Zinc finger nucleases，ZFNs)技术

②转录激活样效应因子核酸酶（Transcription activator-like effector nucleases，TALENs)）技术

③成簇的、规律间隔的短回文重复序列-相关蛋白9技术（Clustered regularly interspaced short palindromic repeat-associated proteins 9，CRISPR-Cas9) 技术

2）DNA-guided genome editing using the *Natronobacterium gregoryi Argonaute* (Nat. Biotechnol,韩春雨 2016) (河北科技大学)？？？世界上没有人重复出来

10.5.3 蛋白质组学

蛋白质组学（proteomics）是研究细胞内全部蛋白质的组成、结构与功能的科学。蛋白质组（proteome）是指由一个基因组所表达的全部相应蛋白质。

蛋白质组与基因组相对应，也是一个整体的概念，是基因组表达的全部蛋白质。两者的根本区别在于：一个有机体只有一个确定的基因组，组成该有机体的所有不同细胞基因组都相同；但基因组内各个基因表达的条件和表达的程度则随时间、地点和环境条件的不同而不同，因而它们表达的模式，即表达产物的种类和数量随时间、地点和环境条件也是不同的。所以，蛋白质组是一个动态的概念。

（1）同一细胞的不同时期，其所表达的蛋白质是不同的。同一细胞在不同的生长发育阶段和生长条件下（正常、疾病或外界环境刺激），所表达的蛋白质也是不同的。

（2）再加上由于基因剪接、蛋白质翻译后修饰和蛋白质剪接等，使从基因到蛋白质的遗传信息的表现规律变得更加复杂，1个基因≠ 1个蛋白，一个基因可以表达的蛋白质的数目可能远大于一。如现在测定人类基因总数为20 000～25 000，但是已发现的蛋白质总数就已高达200 000种。

与传统对单一蛋白研究相比，蛋白质组学研究所采用的是高通量和大规模的研究手段。

双向电泳（two-dimensional electrophoresis，2-DE）技术、计算机图像分析与大规模数据处理技术以及质谱（mass spectrometry，MS）技术是蛋白质组学研究的三大基本支撑技术。

蛋白质组学研究的一般方法和技术路线见示意图10-13

蛋白质组研究的技术路线

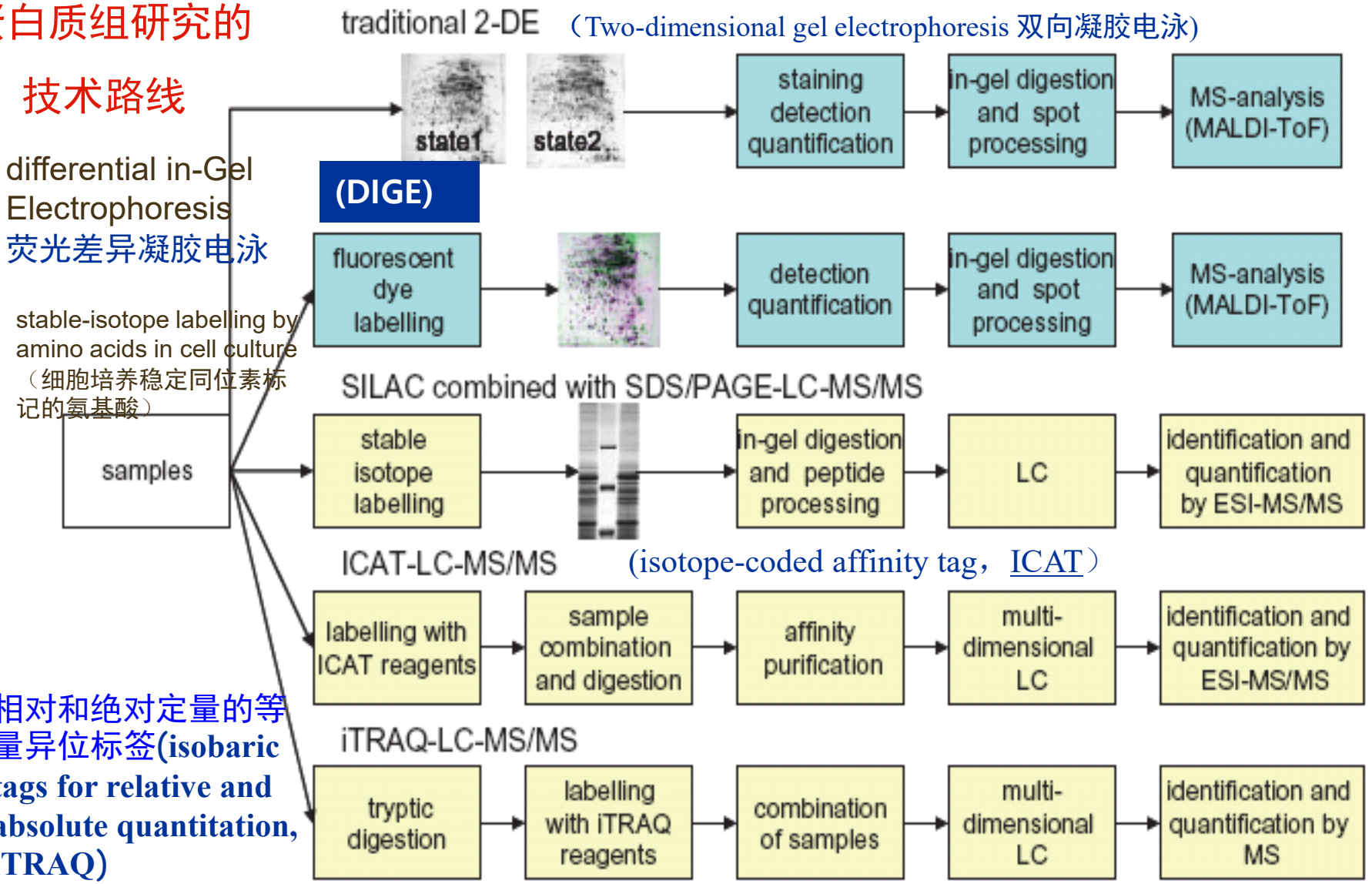


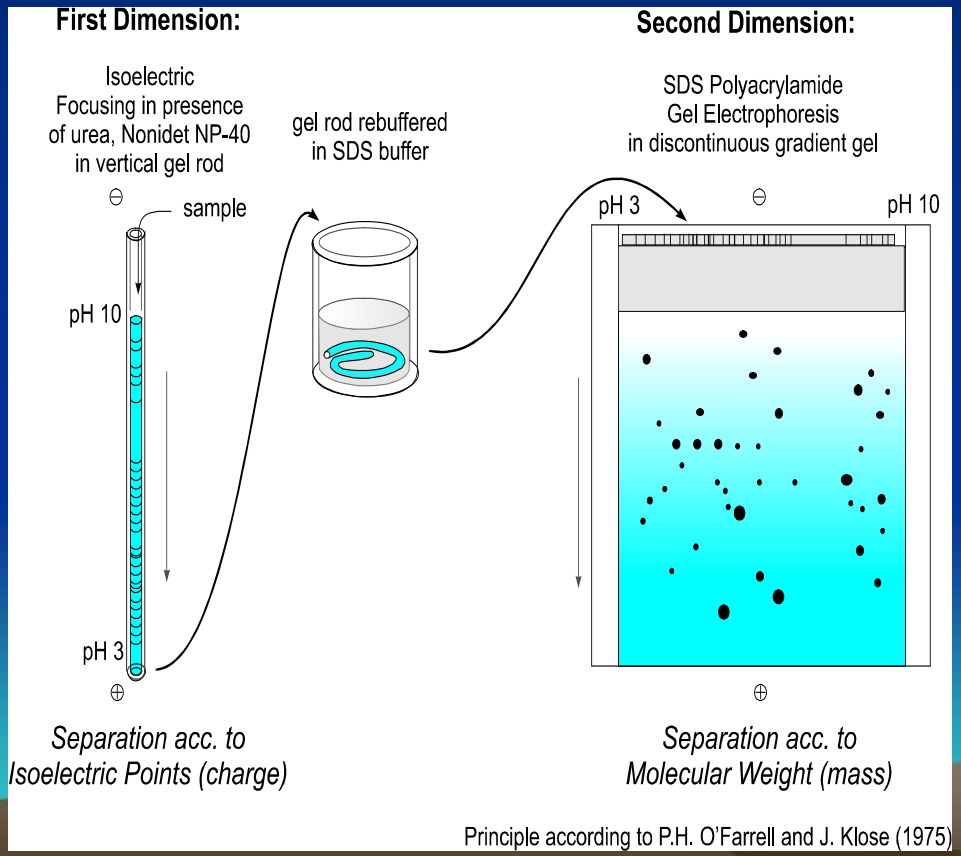
Fig. 1. Schematic presentation of the workflow of gel-based and non-gel-based proteomics approaches.

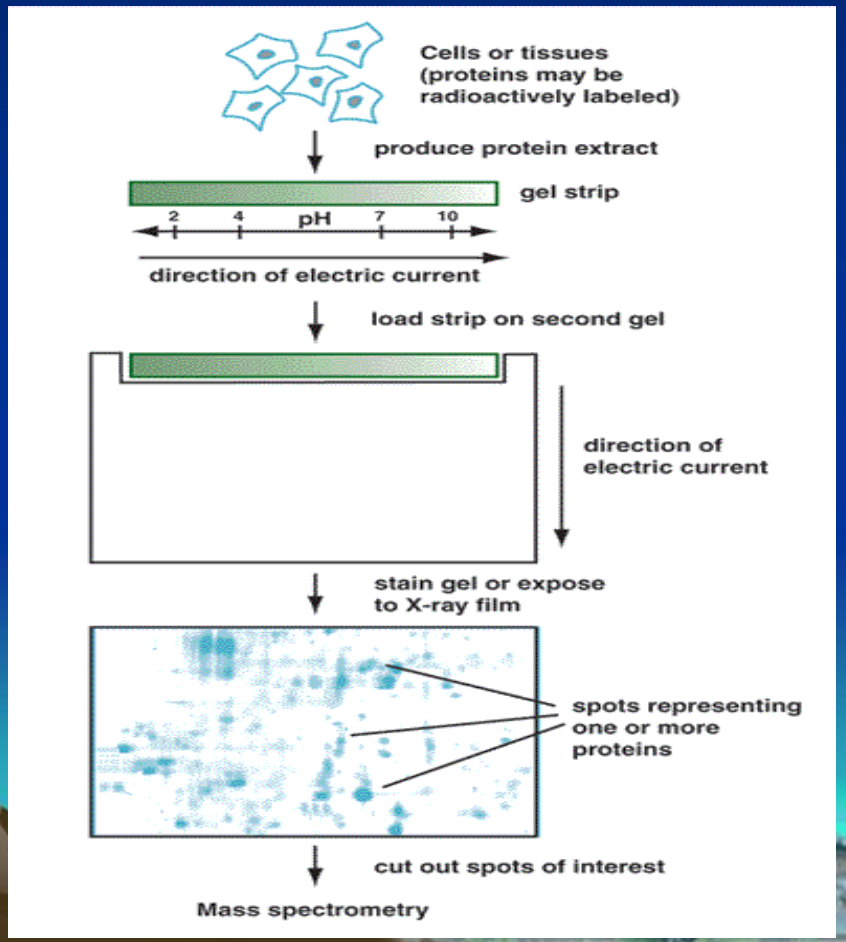
蛋白质组分析-2D Gel

双向凝胶电泳（Two-dimensional gel electrophoresis，2-DGE 或 2-DE ）

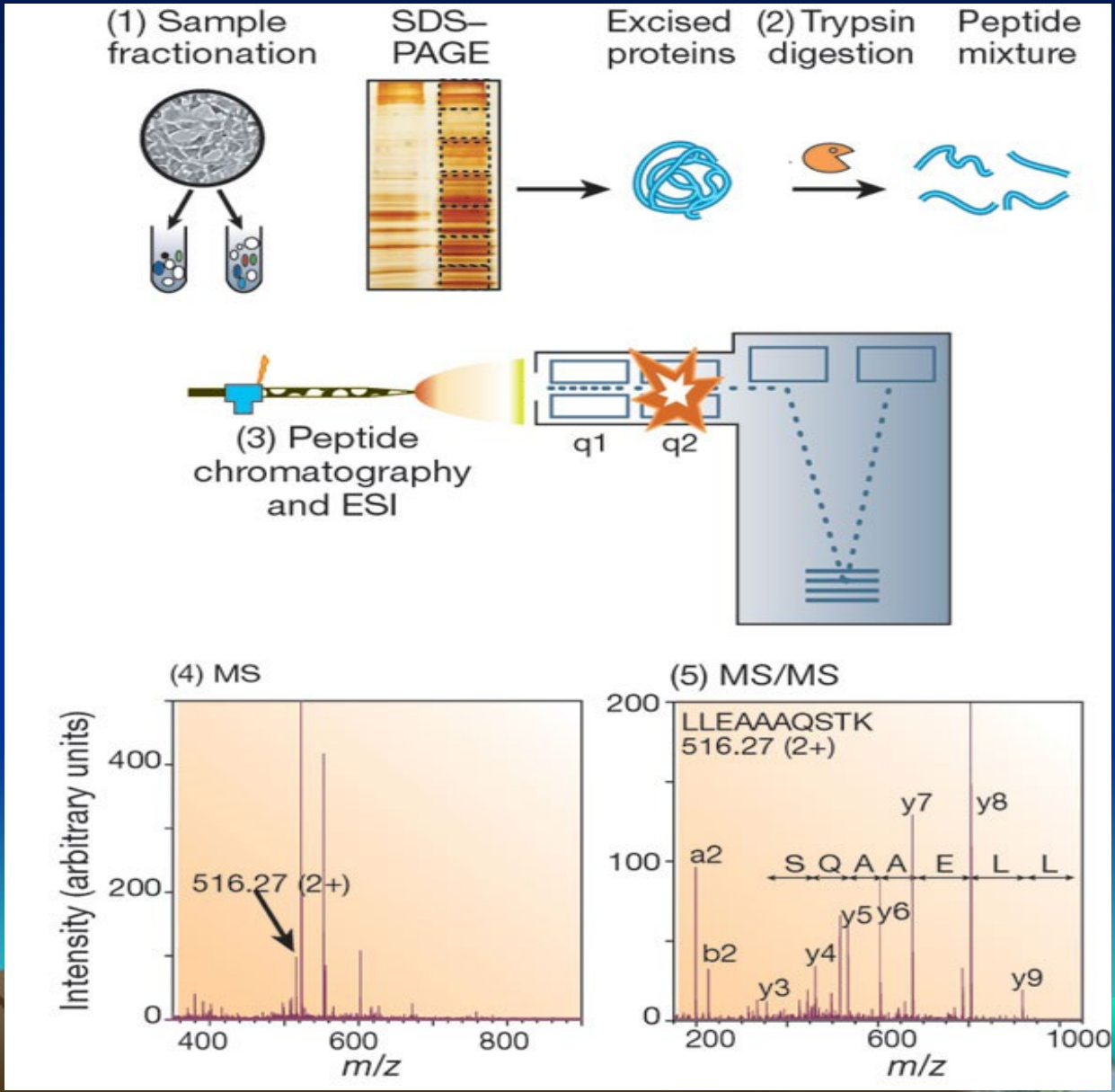
第一向：等电聚焦凝胶电泳（Isoelectrifocusing gel electrophoresis, IEF）

第二向：SDS-PAGE（Sodium dodecyl sulfate-polyacrylamide gel electrophoresis） 十二烷基硫酸钠聚丙烯酰胺凝胶电泳

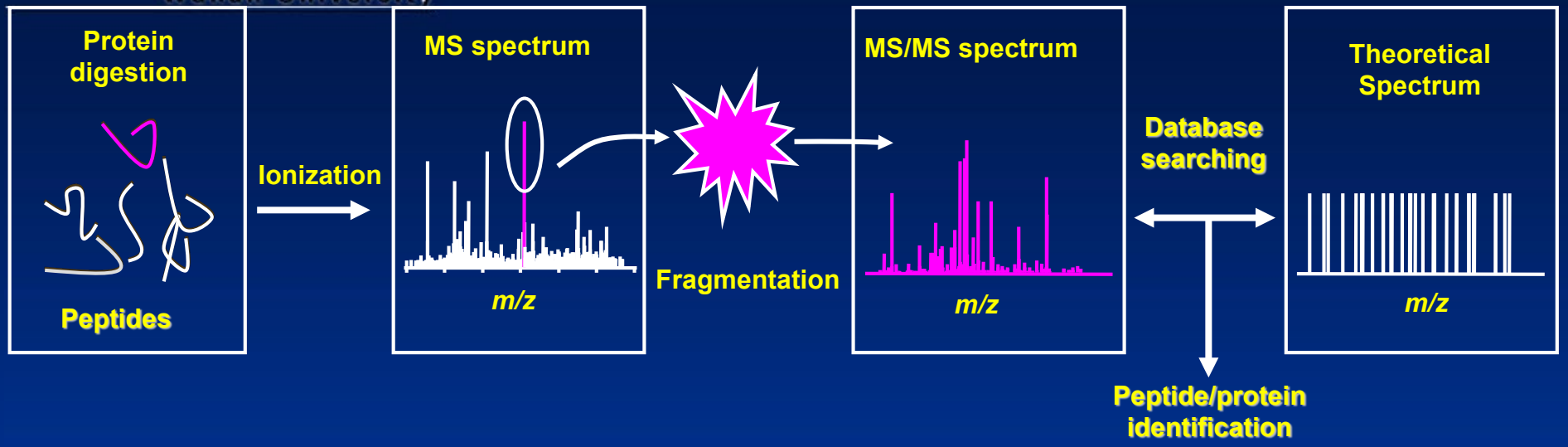




蛋白质组分析-质谱

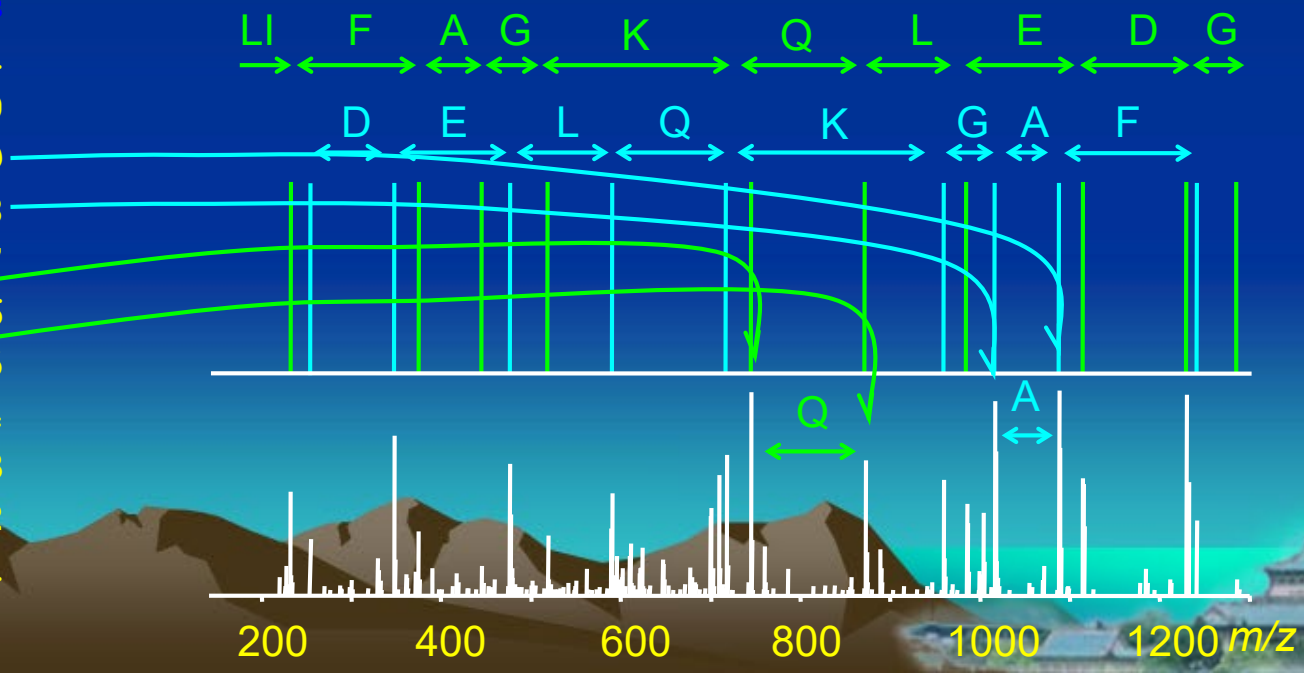

(1) Sample fractionation
SDS–PAGE
Excised proteins
(2) Trypsin digestion
Peptide mixture
(3) Peptide chromatography and ESI
q1
q2
(4) MS
Intensity (arbitrary units)
400
200
0
516.27 (2+)
400
600
800
m/z
(5) MS/MS
200
100
0
LLEAAAQSTK
516.27 (2+)
y7
y8
a2
S Q A A E L L
y5
y6
b2
y4
y3
y9
200
600
1000
m/z

Protein Identification by MS/MS

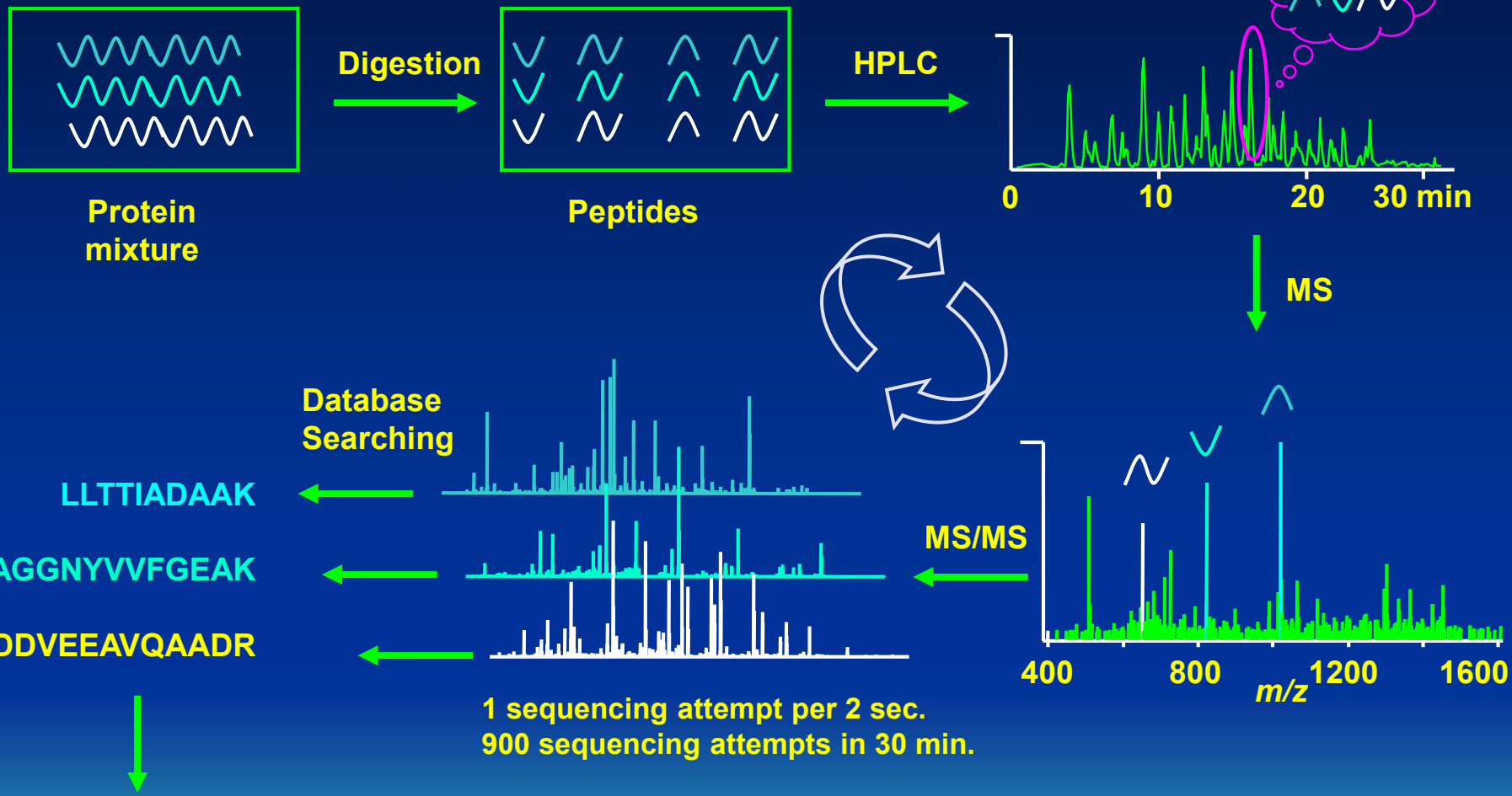


LIFAGKQLEDGR

b ions	y ions
1: L	IFAGKQLEDGR:11
2: LI	FAGKQLEDGR:10
3: LIF	AGKQLEDGR :9
4: LIFA	GKQLEDGR :8
5: LIFAG	KQLEDGR :7
6: LIFAGK	QLEDGR :6
7: LIFAGKQ	LEDGR :5
8: LIFAGKQL	EDGR :4
9: LIFAGKQLE	DGR :3
10: LIFAGKQLED	GR :2
11: LIFAGKQLEDG	R :1



Protein Identification by LC-MS/MS



蛋白质组研究中最重要的一个方面是数据库的建立，包括蛋白质序列数据库、质谱数据库、双向电泳数据库等。著名的蛋白质组数据库，见表10-1。

表 10-1　重要蛋白质组数据库的网址

数据库	网址
NCBI's Entrez proteins	http://www.ncbi.nlm.nih.gov
Swiss-Prot	ftp://ftp.ebi.ac.uk/pub/databases/uniprot/
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq/
TrEMBL	ftp://ftp.ebi.ac.uk/pub/databases/uniprot/
GenPept	ftp://ftp.ncifcrf.gov/pub/genpept/
Ensemble	http://www.ensemble.org/
Hinv-DB ORF	http://www.jbir.aist.go.jp/hinv/index.jsp
UniRef100	ftp://ftp.ebi.ac.uk/pub/databases/uniprot/
PIR(international protein sequence database)	http://www.pir.org/
IPI(international protein index)	ftp://ftp.ebi.ac.uk/pub/databases/IPI/current
SDSPB Protien	http://lifecenter.sgst.cn/main/cn/index.do

10.5.4 代谢组学

代谢组(metabolome)是指某一生物的细胞、组织或器官在某一特定生理时期内所有代谢物的集合。这些代谢产物，包括一系列不同生物化学结构分子，例如肽、糖类、脂质、核酸以及异源物质的催化产物等，也包括一些参与生物体新陈代谢、维持生物体正常生长功能和生长发育的内源性小分子化合物（其相对分子质量Mr≤1 000）。

代谢组学(metabolomics/metabonomics) 是指研究某一生物或细胞，在一特定生理时期内其代谢组的变化，即对所有低相对分子质量代谢产物同时进行定性和定量分析的一门新科学。

代谢组学研究流程包括样品制备、代谢产物分离、检测与鉴定以及数据分析与模型建立三个部分（图10-14），以高通量的检测实验和大规模的计算为特征。

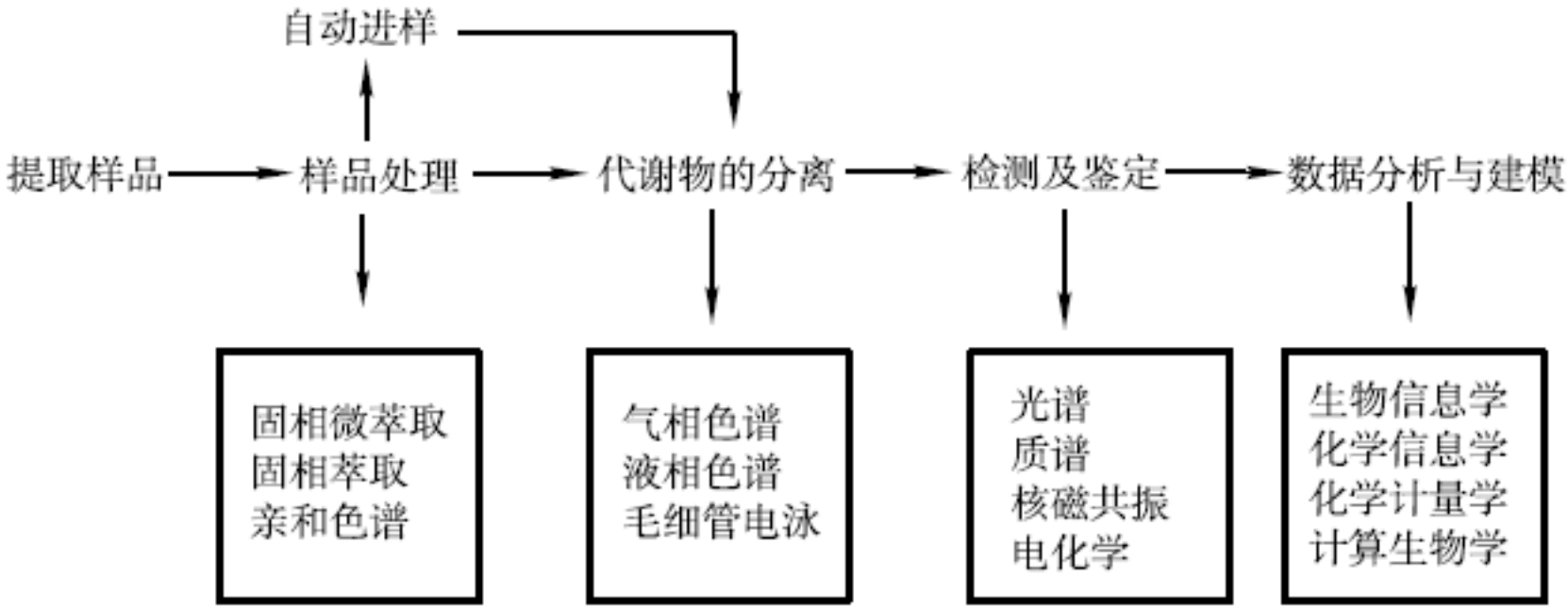


图 10－14　代谢组学各分析流程

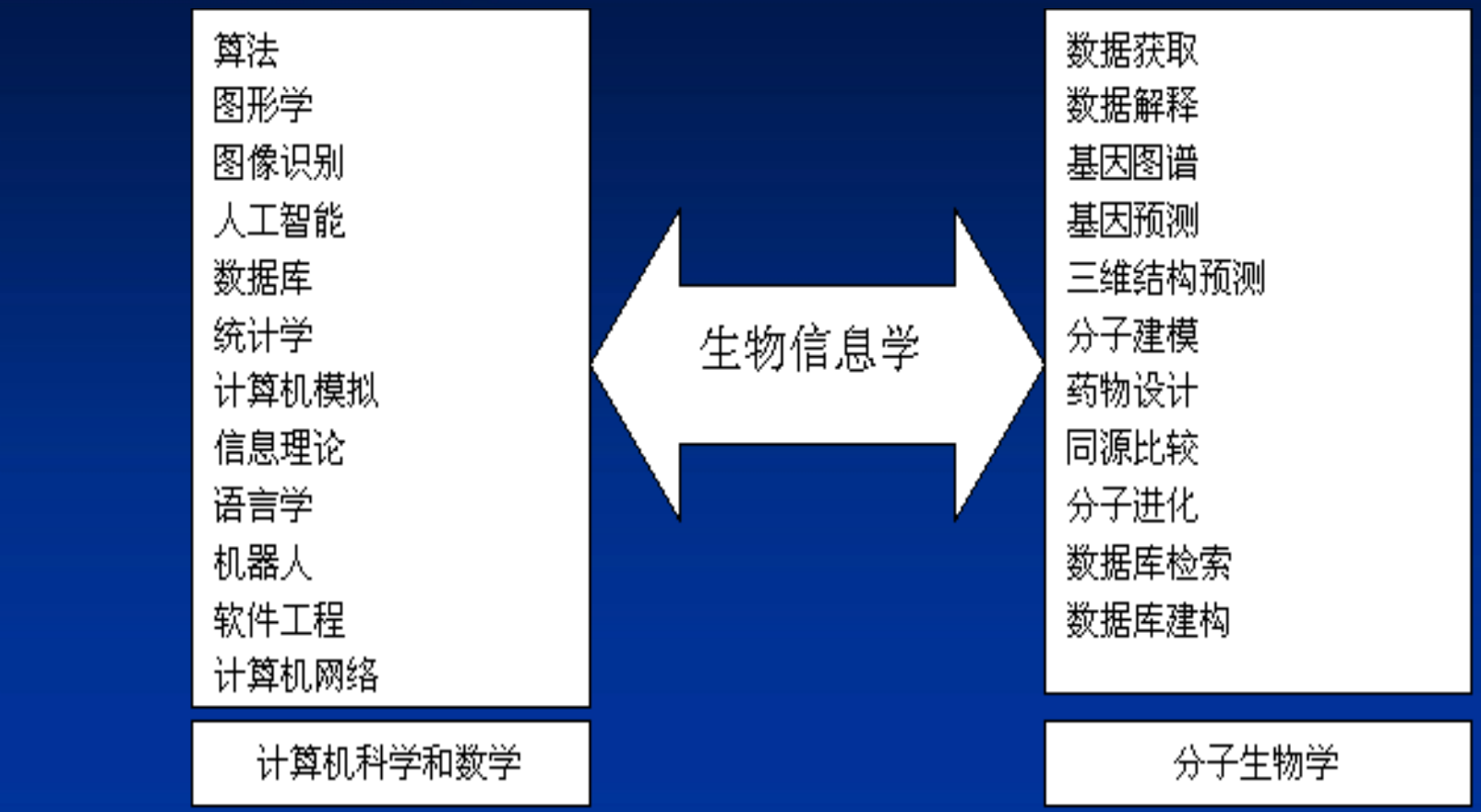
代谢组学技术在临床医学中的应用

代谢组学全流程工作流程

10.5.5 生物信息学与功能基因组学

生物信息学（bioinformatics）是运用计算机技术和信息技术开发新的算法和统计方法，对生物实验数据进行分析，确定数据所含有的生物学意义，并开发新的数据分析工具以实现对各种信息的获取和管理的科学。

生物信息学以计算机为主要工具，开发各种软件，对日益增长的DNA和蛋白质的序列和结构等相关信息和生物学实验数据进行收集、储存、发行、提取、加工、分析和研究，同时建立理论模型，指导实验研究。它由数据库、计算机网络和应用软件三大部分构成，在基因组计划中发挥不可替代的作用。


算法
图形学
图像识别
人工智能
数据库
统计学
计算机模拟
信息理论
语言学
机器人
软件工程
计算机网络
计算机科学和数学
生物信息学
数据获取
数据解释
基因图谱
基因预测
三维结构预测
分子建模
药物设计
同源比较
分子进化
数据库检索
数据库建构
分子生物学

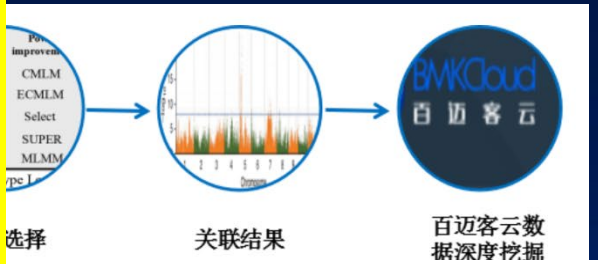
目前生物信息学主要被应用在以下一些方面：大规模基因组测序中的信息处理与分析；基因组相关信息的收集、存取与管理；新基因和新SNP的发现与鉴定，在“国际人类基因组‘单体型图’计划”实施中，进一步确立世界上主要人群基因组的遗传变异图谱；基因识别及编码序列、非编码序列分析；序列比对，遗传密码的起源和生物分子进化；完整基因组的比较研究；大规模基因功能表达谱的分析；生物大分子的结构模拟与药物设计和药物开发；生物信息学分析方法的研究；建立国家生物医学数据库与服务系统等。

10.6.1全基因组关联研究（genome-wide association study，GWAS）

全基因组关联研究（是用来寻找基因变异与表型之间关系的一种遗传学研究方法。进行GWAS的前提条件是：所研究物种的基因组测序工作已完成、基于序列变异-SNP单体型图谱构建和有效的高通量基因分型技术。采用高通量的基因分型技术，对覆盖全基因组的遗传标记进行基因分型，利用全基因组范围内筛选出的高密度遗传标记，例如单核苷酸多态性（SNP）标记或拷贝数变异(CNV）标记，对所研究群体的每个个体进行扫描（一般采用基因芯片的方法），分析扫描得出的分子标记数据与表型性状之间关联的关系。

分析流程

通过全基因组大样本重测序对动植物重要种质资源进行全基因组的基因型鉴定，并与关注的表型数据进行全基因组关联分析（GWAS），进而找到与关注表型相关的SNP位点，定位与性状相关基因，为后续动植物的育种提供科学理论依据。



分析内容

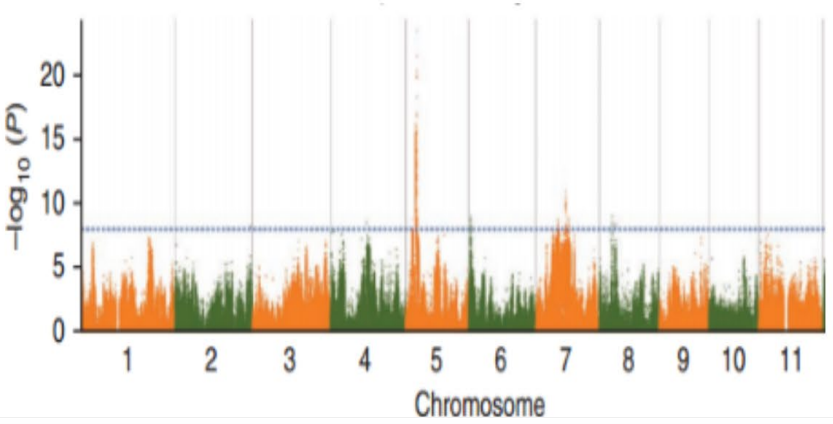
重测序-GWAS

标准分析：

1、测序数据质控
2、与参考基因组比对
3、全基因组变异检测
4、遗传进化分析
5、全基因组关联分析
   1）基因型填充（低于 3X）
   2）SNP 与性状关联分析
   3）Indel 与性状关联分析（10X以上）
   4）候选基因功能注释

定制化分析：

1、显著变异位点的 LD block 分析(确定相关基因后)
2、基因型填充其他方法
3、其他模型类型关联分析



转录组-GWAS

能够在全基因组范围内寻找与重要性状相关联的遗传变异，通过序列和表达量水平双重视角关联分析解析复杂调控网络的的解决方案

有参标准分析：

1、测序数据质控
2、SNP 分析
3、表达量分析和差异基因分析（针对特定时期、特定组织取样）
4、系统发育树构建
5、遗传结构分析
6、关联分析
   1）SNP 关联分析
   2）GEM 关联分析（针对特定时期和特定组织取样）
7、候选基因功能注释和富集分析

无参标准分析（2+3）：

1、三代测序数据质控
2、全长转录本功能注释
3、CDS 预测（针对全长转录本）
4、SSR 分析（针对全长转录本）
5、二代测序数据质控
6、SNP 分析（以全长转录本为参考）
7、表达量分析和差异基因分析（针对特定时期、特定组织取样）
8、系统发育树构建
9、遗传结构分析
10、基因表达分析
11、关联分析
   SNP 关联分析
   GEM 关联分析（针对特定时期和特定组织取样）
12、候选基因功能注释和富集分析

10.6.2 千人基因组计划

千人基因组计划（1 000 Genomes Project）于2008年启动，它是由英国、中国和美国等国的科学家组成的国际协作组，包括Sanger研究所(英国)，深圳华大基因研究院（BGI Shenzhen，中国）以及人类基因组研究所（NHGRI，美国）共同发起的、多家研究机构协作进行的一个国际合作人类基因组测序计划。其目的是将测定选自全世界各地27个群体总计2 500人的全基因组DNA序列，绘制迄今为止最详尽的、最有医学应用价值的人类基因组遗传多态性图谱。

千人基因组计划的全部数据将免费对外开放，其网址：http：//www.1000genomes.org/

10.6.3 1000种动植物基因组计划

中国深圳华大基因研究院于2010年启动“1 000种动植物基因组计划”。该计划在近几年内为1 000种重要动植物进行基因组测序，将为被测序物种的各种研究提供可靠的基础，比如同一物种各种品系的基因组学研究、表观基因组学和转录组学研究等。

华大基因研究院的网址：http://www.genomics.cn/index

本章结束

作业：p252-4, 7, 9

谢谢！

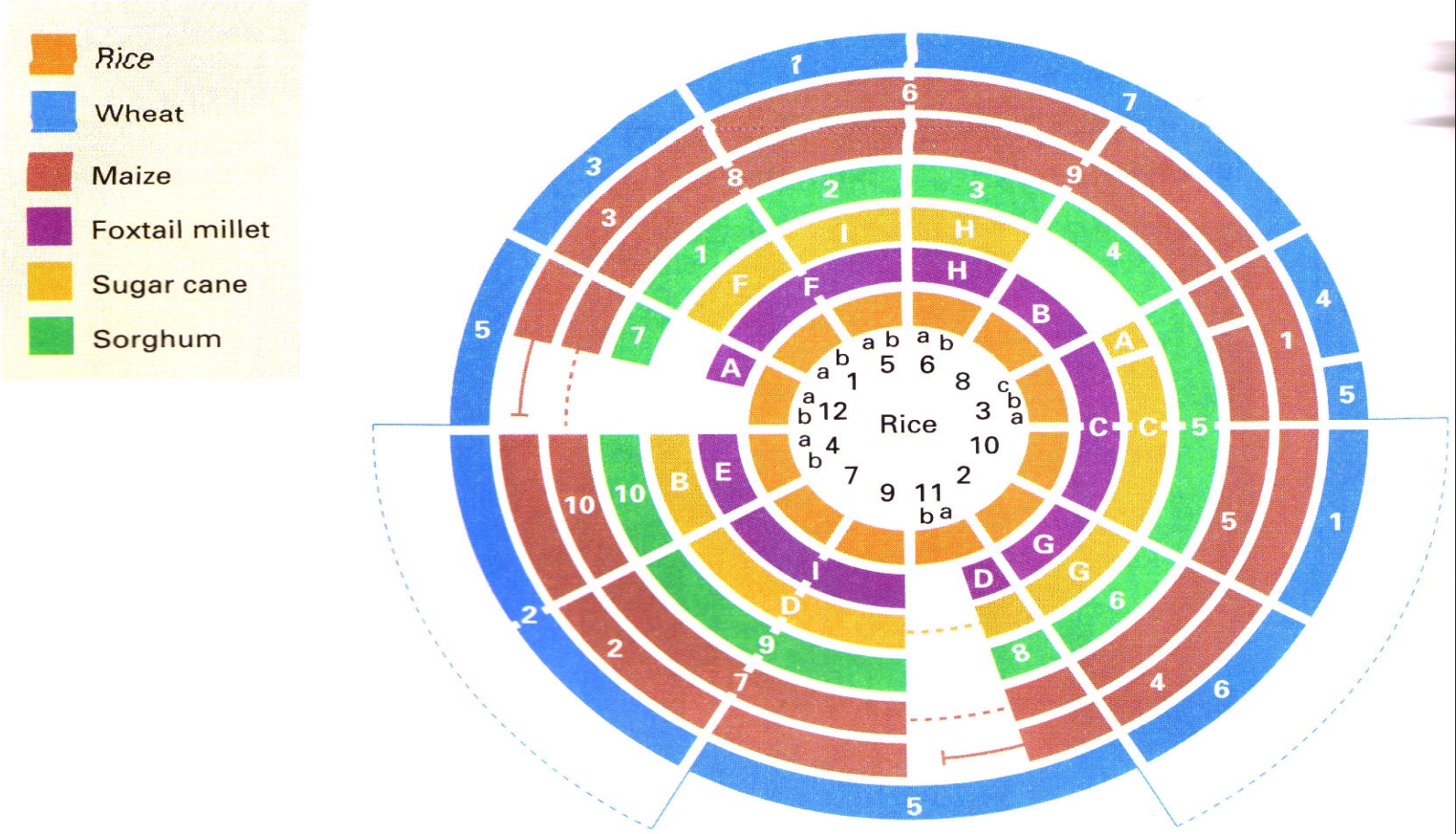


Figure 9.29
Circular arrangement of synteny groups in the cereal grasses makes simultaneous comparisons possible. The thin dashed lines indicate connections between blocks of genes. A number of transpositions of genetic segments are not shown. In some cases, a region within a synteny group is inverted; these inversions are not shown. The circular diagram is for convenience only; there is no indication that the ancestral grass chromosome was actually circular. [Courtesy of Graham Moore. From G. Moore, K. M. Devos, Z. Wang, and M. D. Gale. 1995. *Current Opinion Genet. Devel.* 5: 737.]

图a-1

返回