第14章 基因组学 P283

§1 基因组学概述

一、概念

- 基因组（genome）：又称染色体组
- 一个物种单倍体的染色体数目，物种全部遗传信息的总和（一个物种的单倍体所包含的DNA（染色体）称为该物种的基因组或染色体组）
- 物种遗传信息的“总词典”
- 控制发育的“总程序”
- 生物进化历史的“总档案”

基因组学(genomics)：

- 研究基因组的结构、功能和进化；
  阐明整个基因组所包含的遗传信息和相互关系。

基因组学（genomics）

- 1986年提出，至今20年，已经发展成为遗传学中最重要的分支学科。

- 对物种的所有基因进行定位、作图、测序和功能分析

1986年首次提出基因组学(Genomics)的概念。基因组计划研究开始于1990年，美国国立卫生研究院(NIH)和能源部(DOE)联合启动了被誉为“人体阿波罗计划”的“人类基因组计划”(human genome project，HGP)

人类基因组计划

- 1990，美国国立卫生研究所和能源部投资＄30亿，启动了人类基因组计划，预计15年时间完成人类基因组全部序列的测定
- 1996，完成标记密度为0.6cM的人类基因组遗传图谱，100kb的物理图谱
- 2000，完成草图
- 2001年2月，公布人类基因组图谱的修订版
- 2002，完成测序工作

- 美国提出人类基因组计划后，英、法、日、前苏联、中等，也相继启动了类似项目
- 2000年6月26日，各国科学家公布了人类基因组工作草图
- 2001年8月26日，人类基因组计划中国部分测序项目汇报及联合验收会在京召开，标志人类基因组“中国卷”通过验收

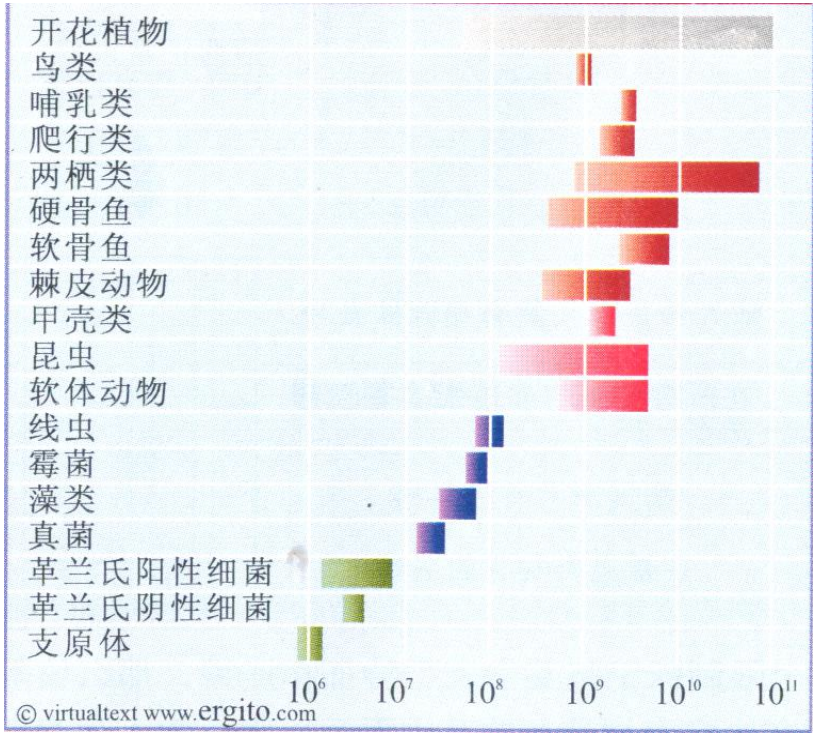
1998年由中国内陆以及台湾地区与日本、美国、法国、韩国、印度等发起，多国共同完成的对水稻基因研究的国际科研工程。1997年9月，水稻基因组测序国际联盟在新加坡举行的植物分子学大会期间成立。1998年2月，中、日、美、英、韩五国代表制定了“国际水稻基因组测序计划”，2002年12月12日，中国科学院、国家科技部、国家发展计划委员会和国家自然基金会联合举行新闻发布会，宣布中国水稻基因组“精细图”已经完成。水稻基因组计划研究包括水稻基因组测序和水稻基因组信息，是继“人类基因组计划”后的又一重大国际合作的基因组研究项目。

几个代表物种的基因组大小

物种	基因组大小/bp
T4噬菌体	2.0×10^{5}
大肠杆菌（*Escherichia coli*）	4.2×10^{6}
酵母（*Sccharomyces cerevisiae*）	1.5×10^{7}
拟南芥（*Arabidopsis thaliana*）	1.0×10^{8}
秀丽小杆线虫（*Caenorhbditis elagans*）	1.0×10^{8}
果蝇（*Drosophila melanogaster*）	1.65×10^{8}
水稻（*Oryza sativa*）	3.89×10^{8}
小白鼠（*Mus musculus*）	3.0×10^{9}
人类（Homo sapiens）	3.3×10^{9}
玉米(*Zea mays*）	5.4×10^{9}
普通小麦（*Triticum aestivum*）	1.6×10^{10}

C值：一个单倍体基因组中DNA的总量。一个特定的种属具有特定的C值

C值悖理：物种的C值和它的进化复杂性之间无严格对应关系的现象



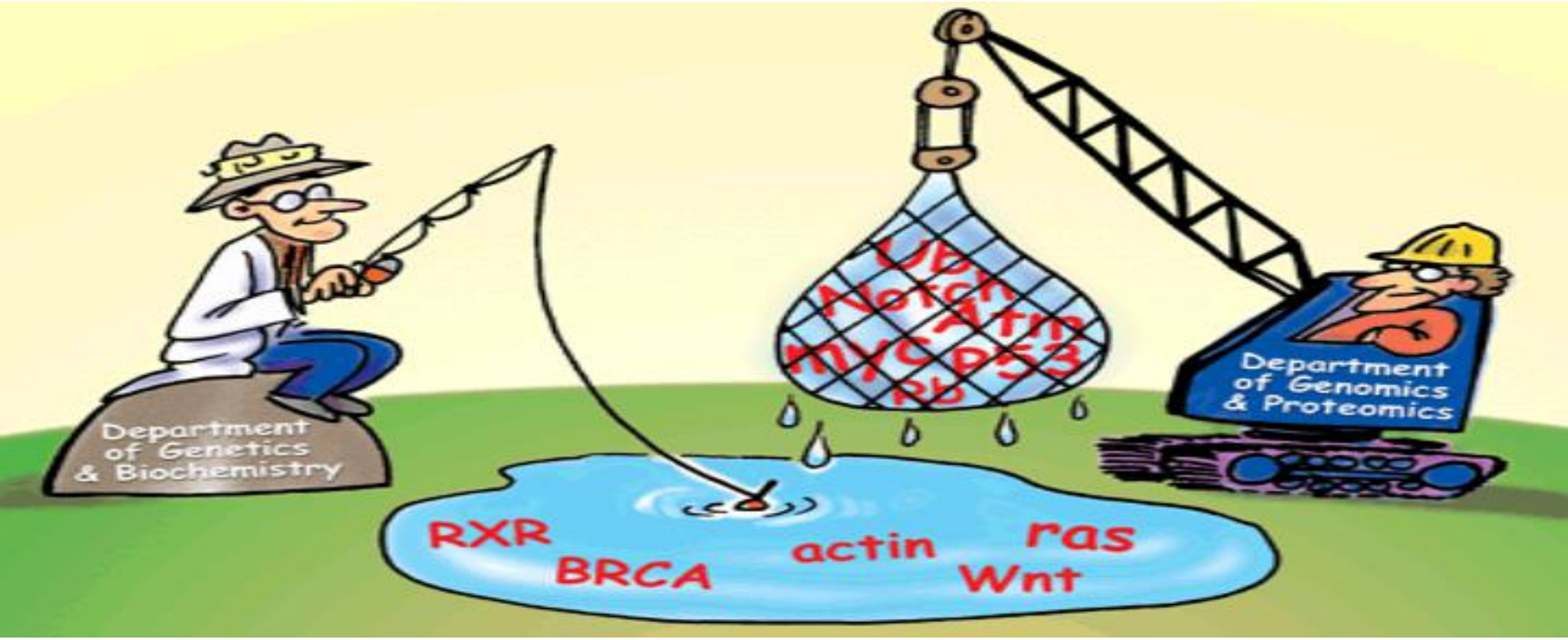
N值：生物体所含有的基因数目

N值悖理：复杂性不同的生物种属所具有的基因数目与其生物结构的复杂性不成比例的现象。 水稻基因数约4万个，人类基因总数约3万个

基因组学(genomics)是遗传学研究进入分子水平后发展起来的一个分支，主要研究生物体全基因组(genome)的分子特征。

基因组学强调的是以**基因组**为单位，而不是以单个基因为单位作为研究对象

Fishing in a More Effective Way!



CREDIT: JOE SUTLIFF
Science, Vol 291: 1221.

基因组学的研究目标：

1. 获得生物体全部基因组序列
2. 注解基因组所含的全部基因
3. 鉴定所有基因的功能及基因间相互作用关系
4. 阐明基因组的复制及进化规律

研究内容：

结构基因组学

1. 构建基因组的遗传图谱；
2. 构建基因组的物理图谱；
3. 测定基因组DNA的全部序列；

功能基因组学

4. 基因组序列诠释

蛋白组学

5.蛋白组学

基因组学的研究内容

结构基因组学：通过基因作图、核苷酸序列分析确定基因组成、进行基因定位的科学

功能基因组学(后基因组学)：利用结构基因组所提供的信息和产物，研究基因组功能表达的一门分支学科。 基因的识别、鉴定和克隆，基因结构与功能及其相互关系，基因表达调控

蛋白质组学：研究细胞内蛋白质组成及其活动规律的新兴学科。 鉴定蛋白质表达、存在方式、结构、功能和相互作用方式等

基因组学的重要组成部分是基因组计划，如人类、水稻基因组计划，大体可分为：

1、构建基因组的遗传图谱

2、构建基因组的物理图谱

3、测定基因组DNA的全部序列

4、构建基因组的转录本图谱

5、分析基因组的功能

- 遗传图谱(genetic map)：采用遗传分析的方法将基因和其它DNA分子标记标定在染色体上构建而成的图谱称为遗传图谱。遗传图距单位为厘摩（cM）。
- 物理图谱（physical map）：采用分子生物学技术，直接确定DNA分子标记、基因或克隆在染色体上的实际位置。通常用标记之间的DNA长度（碱基对）来表示。

经典遗传学

- 在20世纪初，遗传学刚刚诞生的时候，遗传学家的工作主要是鉴别感兴趣的基因，确定这些基因在染色体上的位置。

- 第一个环节：寻找自发突变体，或者利用物理、化学因素诱发突变。

- 第二个环节：通过连锁分析确定新基因与已知基因的相互关系，绘制遗传连锁图。

基因组学的研究内容

- 结构基因组学
- 功能基因组学
- 蛋白质组学

结构基因组学（structural genomics）

- 基因定位
- 基因组作图
- 测定核苷酸序列

功能基因组学（**functional genomics**）

又称后基因组学（postgenomics）

- 基因的识别、鉴定、克隆
- 基因结构、功能及其相互关系
- 基因表达调控的研究

蛋白质组学（**proteomics**）

- 鉴定蛋白质的产生过程、结构、功能和相互作用方式

§2 基因组图谱的构建

- 基因组计划的主要任务是获得全基因组序列
- 但是，现在的测序方法每次只能测800～1000bp
- 大量的测序片段要拼接
- 要知道序列在Chr上的位置才能正确拼接
- 基因组计划的第一个环节：
  构建基因组图谱

在进行大规模序列测定之前，构建基因组图谱，作为序列测定中制定测序方案的依据，以便先重后轻地分析基因，锚定测知的核酸序列在染色体上的位置

在人类基因组计划实施中，首先用了6年时间构建高密度的基因组图谱，然后才进入测序工作

基因组图谱

- 遗传图谱（**genetic map**）
- 物理图谱（**physical map**）

遗传图谱（genetic map）

采用***遗传分析的方法***将基因或其它DNA序列标定在染色体上构建连锁图。

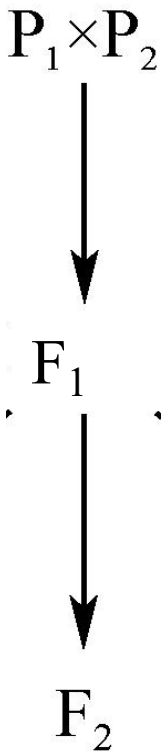
思考题

1. 什么是基因组？基因组学研究内容？
2. 什么是C值悖理、N值悖理？
3. 遗传图谱的概念？

- 为什么要构建遗传图谱？
- 构建方法？

植物基因组遗传图谱的构建

选择亲本
产生构图群体
遗传标记的染色体定位
标记间的连锁分析



一、遗传图谱构建

遗传作图：采用遗传学分析方法将基因或其他DNA顺序标定在染色体上构建连锁图

1、图谱标记

形态标记：主要指可以观察到的一些性状，如种皮颜色、眼色、株高等

细胞学标记：能明确显示遗传多态性的细胞学特征。染色体的结构特征和数量特征是常见的细胞学标记

生化标记：主要是同工酶及种子贮藏蛋白，有时又称蛋白质标记

分子标记：主要指DNA水平上的标记。RFLP，RAPD，SSR，STS，AFLP，CAPS，SNP

遗传标记

- 有可以识别的标记，才能确定目标的方位及彼此之间的相对位置。
- 构建遗传图谱就是寻找基因组不同位置上的特征标记。
- 包括：

  形态标记

  细胞学标记

  生化标记

  DNA分子标记

多态性（polymophism）

所有的标记都必须具有多态性！

- 花色：白色、红色
- 株高：高、矮
- 血型：A、B、O型
- 淀粉：糯、非糯

所有多态性都是基因突变的结果！

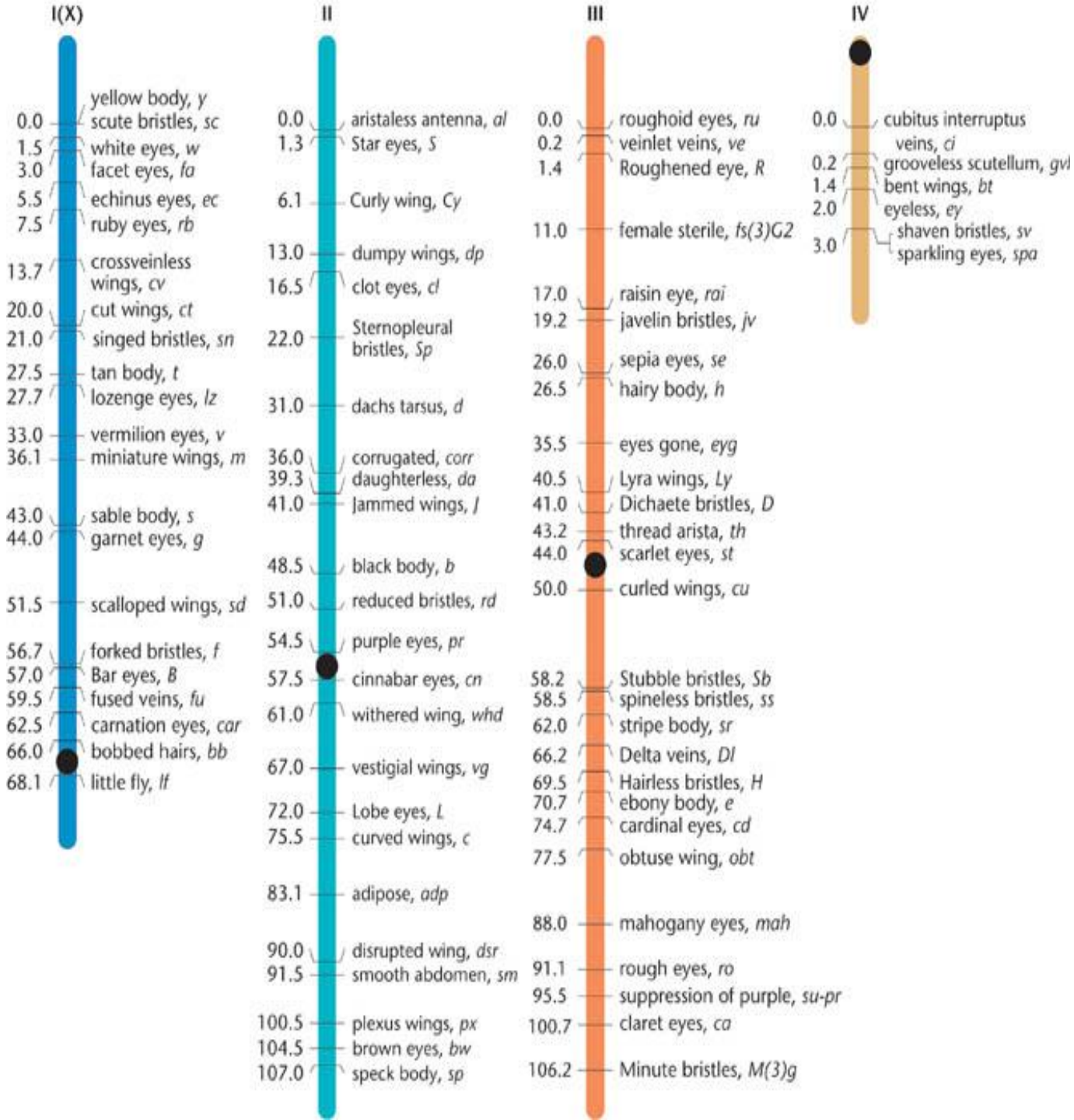
形态标记

- 形态性状：株高、颜色、白化症等
- 又称表型标记

- 数量少
- 很多突变是致死的
- 受环境、生育期等因素的影响

- 最早建立的果蝇连锁图，就是利用控制果蝇眼睛的形状、颜色，躯体的颜色、翅膀的形状等形态性状作为标记，分析它们连锁关系及遗传距离，绘制而成的。

- 控制性状的其实是基因，所以形态标记实质上就是基因标记。


I(X)
0.0 yellow body, y
0.0 scute bristles, sc
1.5 white eyes, w
3.0 facet eyes, fa
5.5 echinus eyes, ec
7.5 ruby eyes, rb
13.7 crossveinless wings, cv
20.0 cut wings, ct
21.0 singed bristles, sn
27.5 tan body, t
27.7 lozenge eyes, lz
33.0 vermilion eyes, v
36.1 miniature wings, m
43.0 sable body, s
44.0 garnet eyes, g
51.5 scalloped wings, sd
56.7 forked bristles, f
57.0 Bar eyes, B
59.5 fused veins, fu
62.5 carnation eyes, car
66.0 bobbed hairs, bb
68.1 little fly, lf
II
0.0 aristaless antenna, al
1.3 Star eyes, S
6.1 Curly wing, Cy
13.0 dumpy wings, dp
16.5 clot eyes, cl
22.0 Sternopleural bristles, Sp
31.0 dachs tarsus, d
36.0 corrugated, corr
39.3 daughterless, da
41.0 Jammed wings, J
48.5 black body, b
51.0 reduced bristles, rd
54.5 purple eyes, pr
57.5 cinnabar eyes, cn
61.0 withered wing, whd
67.0 vestigial wings, vg
72.0 Lobe eyes, L
75.5 curved wings, c
83.1 adipose, adp
90.0 disrupted wing, dsr
91.5 smooth abdomen, sm
100.5 plexus wings, px
104.5 brown eyes, bw
107.0 speck body, sp
III
0.0 roughoid eyes, ru
0.2 veinlet veins, ve
1.4 Roughened eye, R
11.0 female sterile, fs(3)G2
17.0 raisin eye, rai
19.2 javelin bristles, jv
26.0 sepia eyes, se
26.5 hairy body, h
35.5 eyes gone, eyg
40.5 Lyra wings, Ly
41.0 Dichaete bristles, D
43.2 thread arista, th
44.0 scarlet eyes, st
50.0 curled wings, cu
58.2 Stubble bristles, Sb
58.5 spineless bristles, ss
62.0 stripe body, sr
66.2 Delta veins, Dl
69.5 Hairless bristles, H
70.7 ebony body, e
74.7 cardinal eyes, cd
77.5 obtuse wing, obt
88.0 mahogany eyes, mah
91.1 rough eyes, ro
95.5 suppression of purple, su-pr
100.7 claret eyes, ca
106.2 Minute bristles, M(3)g
IV
0.0 cubitus interruptus veins, ci
0.2 grooveless scutellum, gvl
1.4 bent wings, bt
2.0 eyeless, ey
3.0 shaven bristles, sv
sparkling eyes, spa

果蝇连锁图

细胞学标记

- 明确显示遗传多态性的染色体结构特征和数量特征
  - 染色体的核型
  - 染色体的带型
  - 染色体的结构变异
  - 染色体的数目变异
- 优点：不受环境影响
- 缺点：数量少、费力、费时、对生物体的生长发育不利

生化标记

- 又称蛋白质标记
- 就是利用蛋白质的多态性作为遗传标记。

  如：同工酶、贮藏蛋白
- 优点：数量较多，受环境影响小
- 缺点：受发育时间的影响、有组织特异性、只反映基因编码区的信息

DNA分子标记

简称分子标记

以DNA序列的多态性作为遗传标记

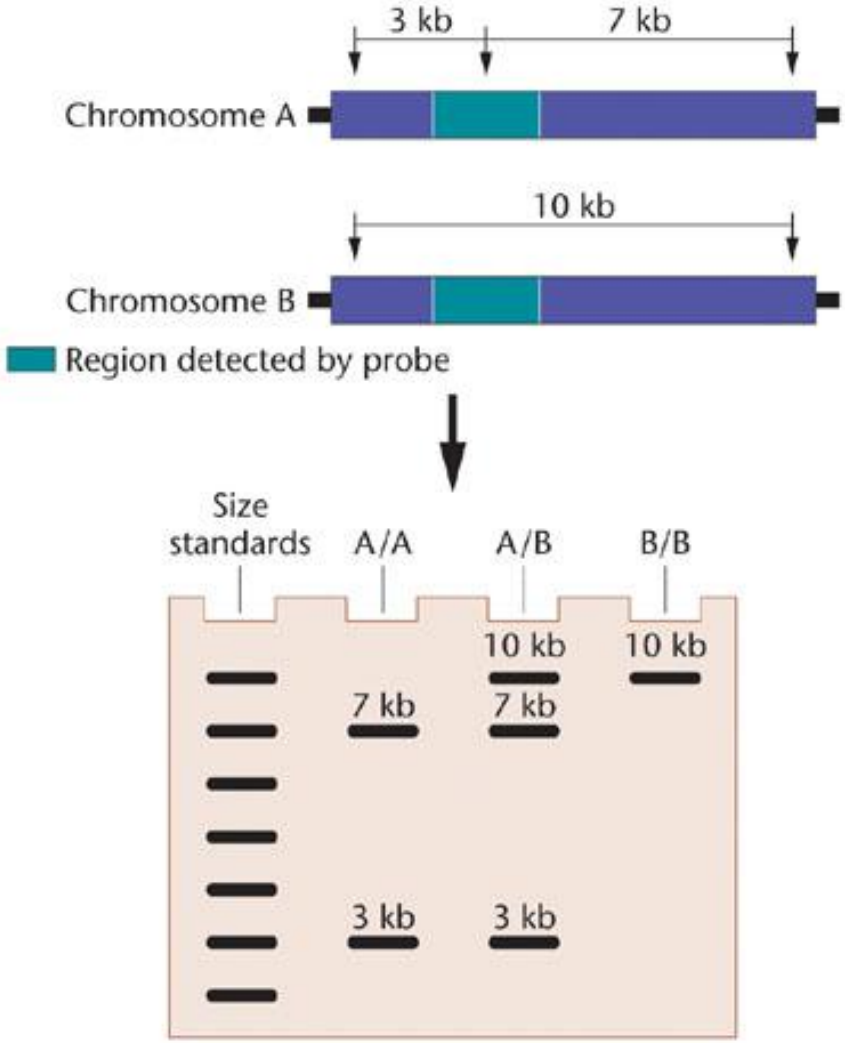
优点：

- 不受时间和环境的限制
- 遍布整个基因组，数量无限
- 不影响性状表达
- 自然存在的变异丰富，多态性好
- 共显性，能鉴别纯合体和杂合体

限制性片段长度多态性
（restriction fragment length polymorphism，RFLP）

- DNA序列能或不能被某一酶酶切，相当于一对等位基因的差异。
- 如有两个DNA分子（一对染色体），一个具有某一种酶的酶切位点，而另一个没有这个位点，酶切后形成的DNA片段长度就有差异，即多态性。
- 可将RFLP作为标记，定位在基因组中某一位置上。
- 人类基因组中有10^5个RFLP位点，每一位点只有两个等位基因。

RFLP分析



Genotypes	Fragment sizes
Homozygous for chromosome A (A/A)	3 kb, 7 kb
Heterozygous (A/B)	3 kb, 7 kb, 10 kb
Homozygous for chromosome B (B/B)	10 kb

微卫星（microsatellite）标记

- 微卫星又称为简单重复序列（simple sequence repeat，SSR）。

- 这种重复序列的重复单位很短，常常只有2个、3个或4个核苷酸

- 如一条染色体TCTGAGAGAGACGC

  另一染色体TCTGAGAGAGAGAGAGAGACGC，就构成了多态性。

2、遗传图谱的构建方法

- 理论基础: 连锁与交换
- 基本方法: 两点测验法和三点测验法

植物遗传图谱的构建

- 选择研究材料（亲本）
- 构建分离群体
- 遗传标记检测
- 标记间的连锁分析

选择亲本

- 要求亲缘关系远，遗传差异大
- 但又不能相差太大以导致引起子代不育。
- 对备选材料进行多态(差异)性检测，综合测定结果，选择有一定量多态性的一对或几对材料作为遗传作图亲本。

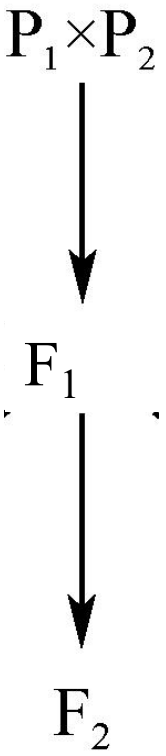
植物基因组遗传图谱的构建

选择亲本

产生构图群体

遗传标记的染色体定位

标记间的连锁分析



构建作图群体

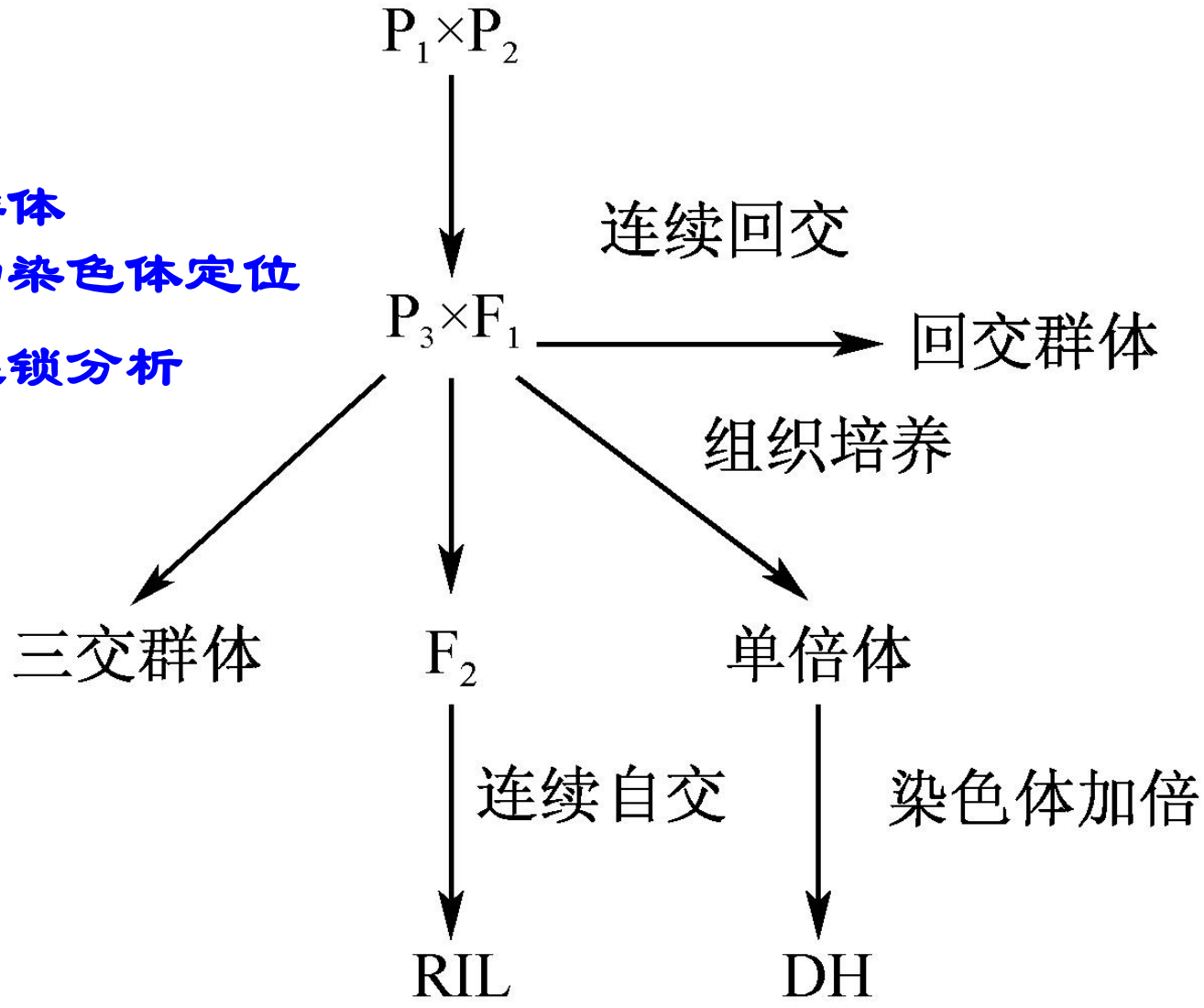
植物基因组遗传图谱的构建

选择亲本

产生构图群体

遗传标记的染色体定位

标记间的连锁分析



构建作图群体

遗传标记的染色体定位

- 利用遗传学方法或其它方法将少数标记锚定在染色体上，作为确定连锁群的参照系。
- 常用的方法：

  单体分析

  三体分析

  代换系分析

  附加系分析

标记间的连锁分析

- 利用在两个亲本间有多态性的标记分析分离群体中所有个体的基因型
- 根据连锁交换的情况，确定标记之间的连锁关系和遗传距离
- 有计算机软件可以应用

水稻遗传图

- 1994年，水稻第一张高密度遗传图谱
  927个座位， 1383个标记
- 1998年，1157个座位，2275个标记
- 2000年，3267个标记
- 高密度的遗传图谱为基因组测序和遗传研究奠定了坚实的基础。

人类基因组遗传图谱的构建

人类的遗传图谱是利用家系分析法，在对8个家系的134个成员的分析中，主要根据5264个STR标记绘制而成的。利用这些家系的资料绘制第1至22号染色体图谱。对于X染色体图谱，还利用了来自另外12个家系，170个成员的资料绘制而成。

将5264个标记定位在2335个座位，据此构建的人类基因组遗传图谱的密度为每个标记599 kb。

人类遗传图谱的构建

- 不可能根据需要选择亲本，设计杂交组合，构建分离群体！
- 只能检测现存家庭连续几代成员的基因型
- 家系分析法
- 资料有限、必须借助于统计学方法

现有的人类遗传图谱

- 1～22号染色体
- 8个家系134个成员
- X染色体，12个家系170个成员
- 5364个SSR标记
- 2335个座位
- 标记间的平均距离599kb

二、物理图谱

由于遗传图谱的分辨率有限、精确性不高，所以还要构建物理图谱

物理图谱的构建

- 用分子生物学方法直接检测DNA标记在染色体上的实际位置绘制成的图谱称为物理图谱。

- 有遗传图谱为什么还要构建物理图谱？

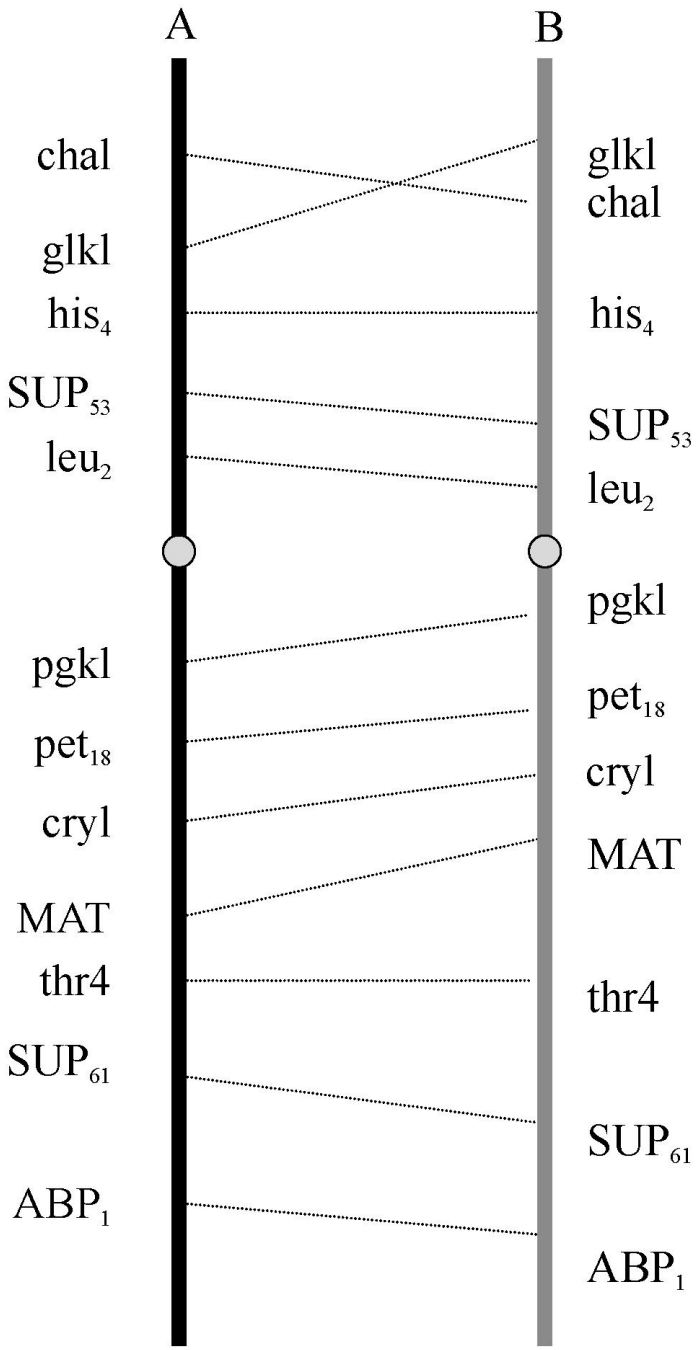
遗传图谱的缺陷

- 分别率有限
- 人类只能研究少数减数分裂事件，不能获得大量子代个体
- 测序要求每个标记的间隔小于100kb
- 实际是599kb

遗传图谱的缺陷

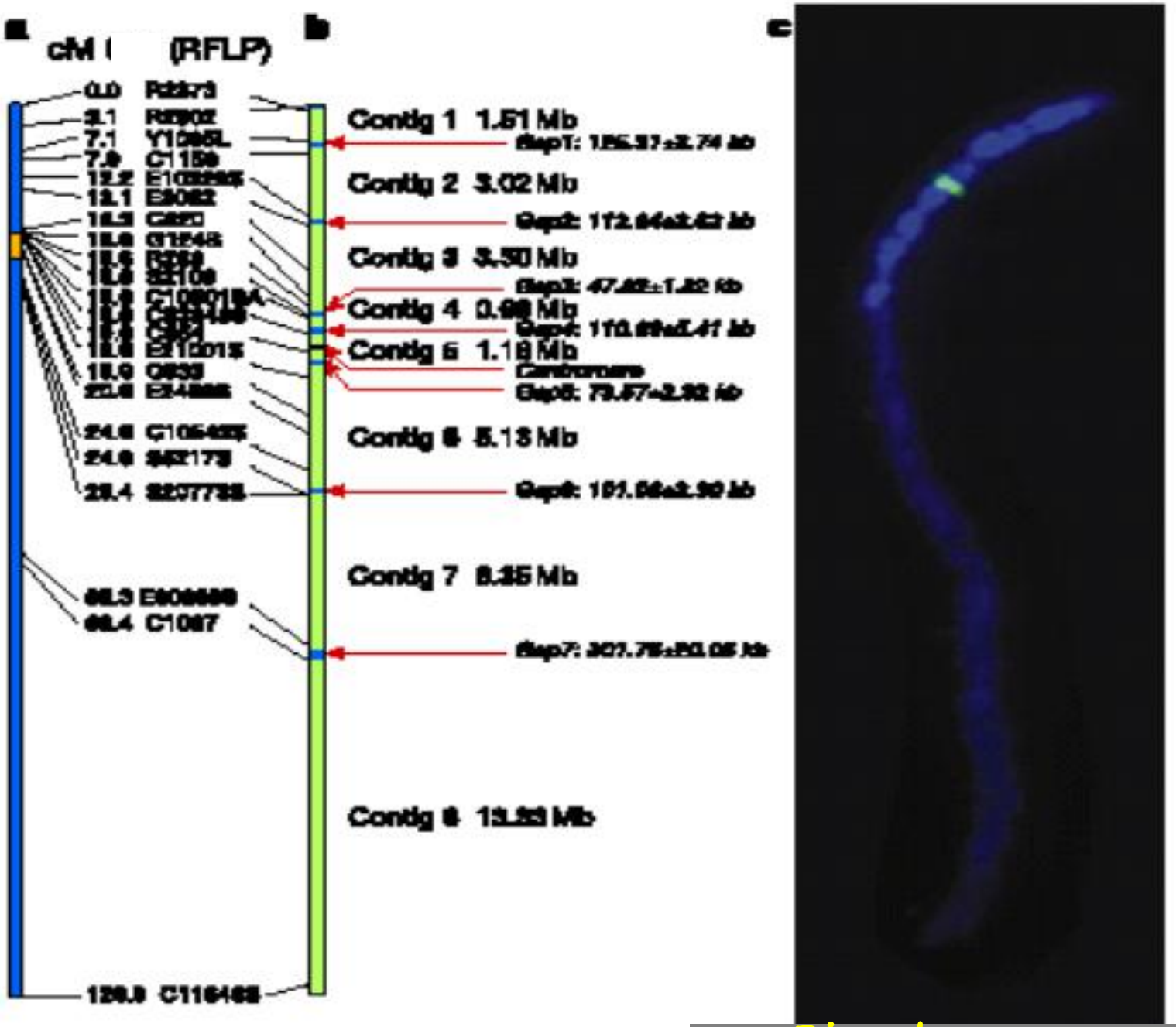
- 精确性不够
- 经典遗传学认为，交换是随机发生的,基因组中有些区域是重组热点
- 倒位、重复等染色体结构变异会限制交换重组

酵母遗传图与物理图比较



A 遗传图

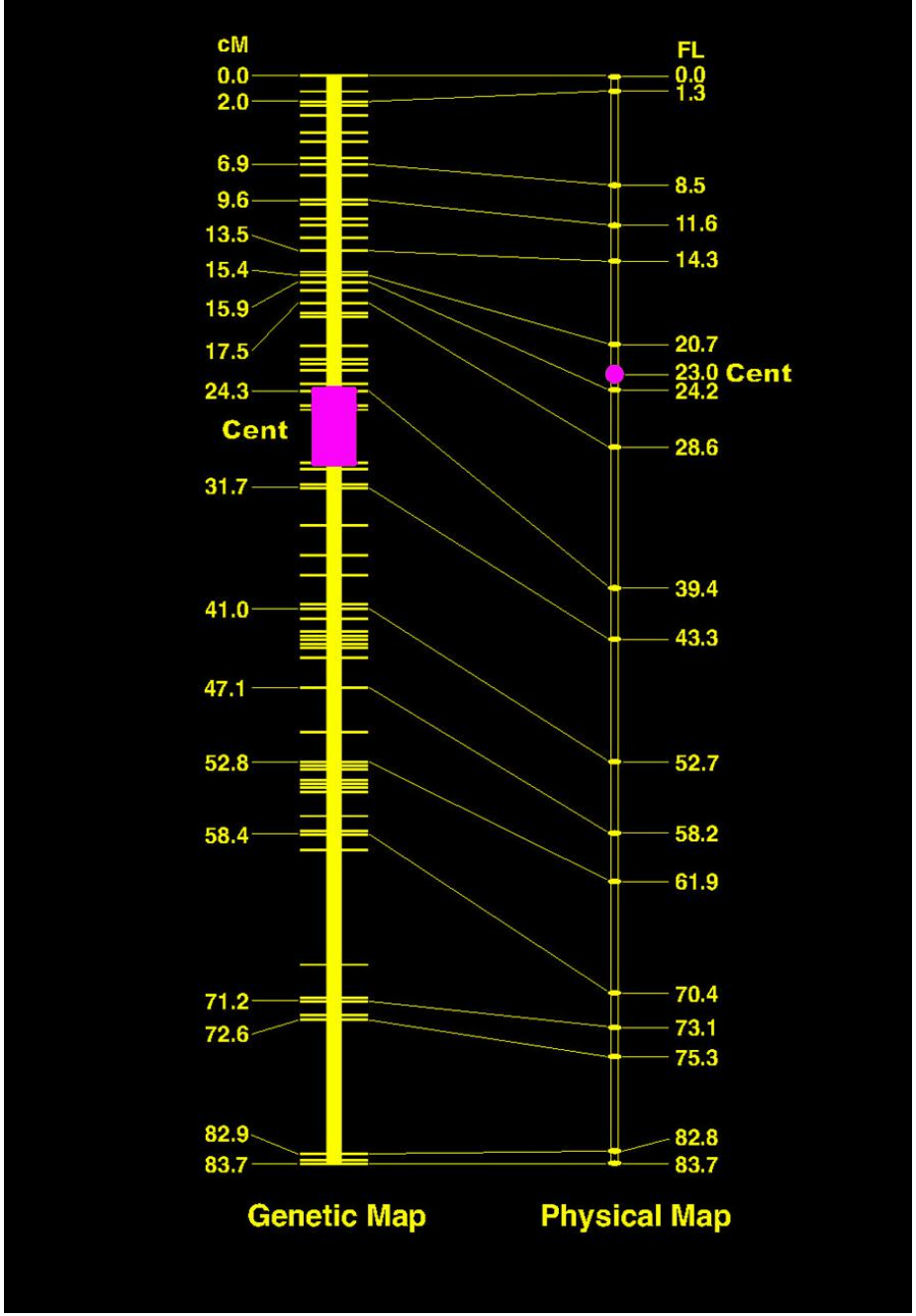
B 物理图



Genetic map

Physical map

Rice chromosome 4.



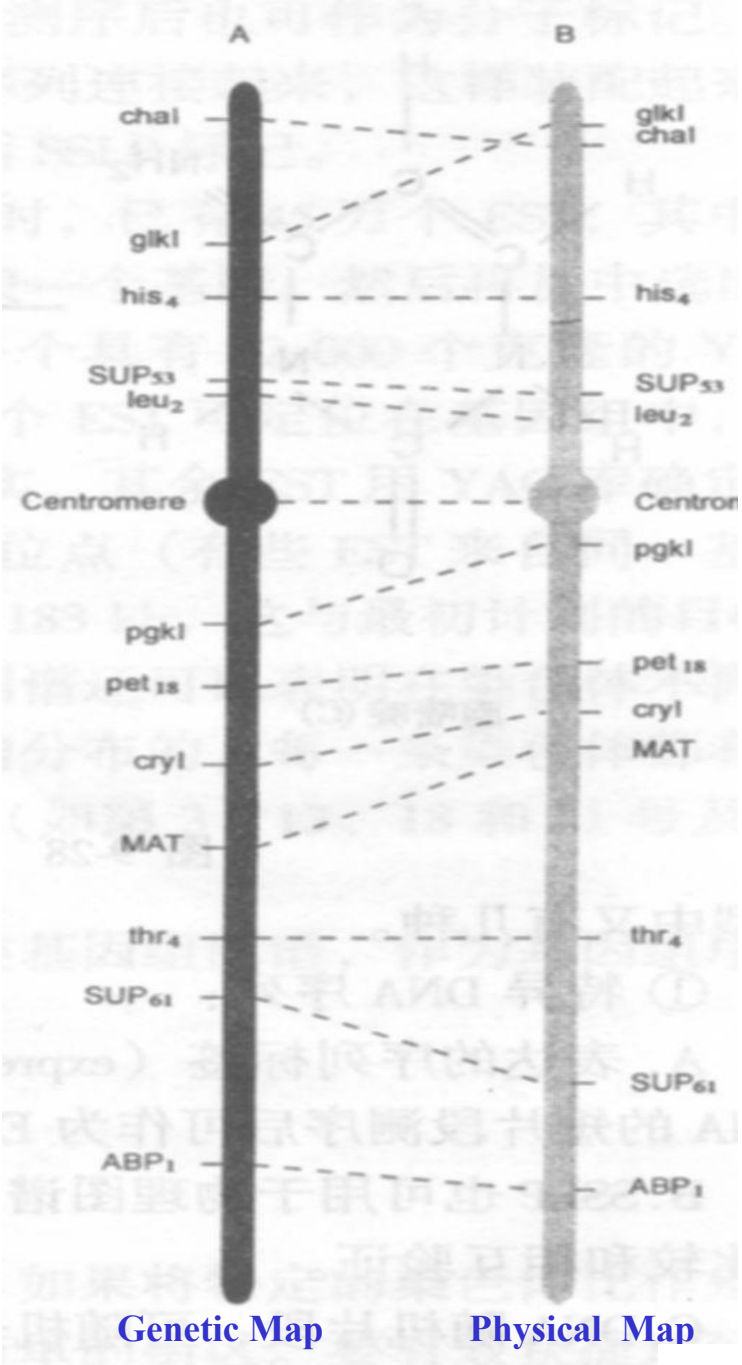


图 9-27 酵母菌第Ⅲ染色体

Chromosome 10 of rice

物理作图的方法

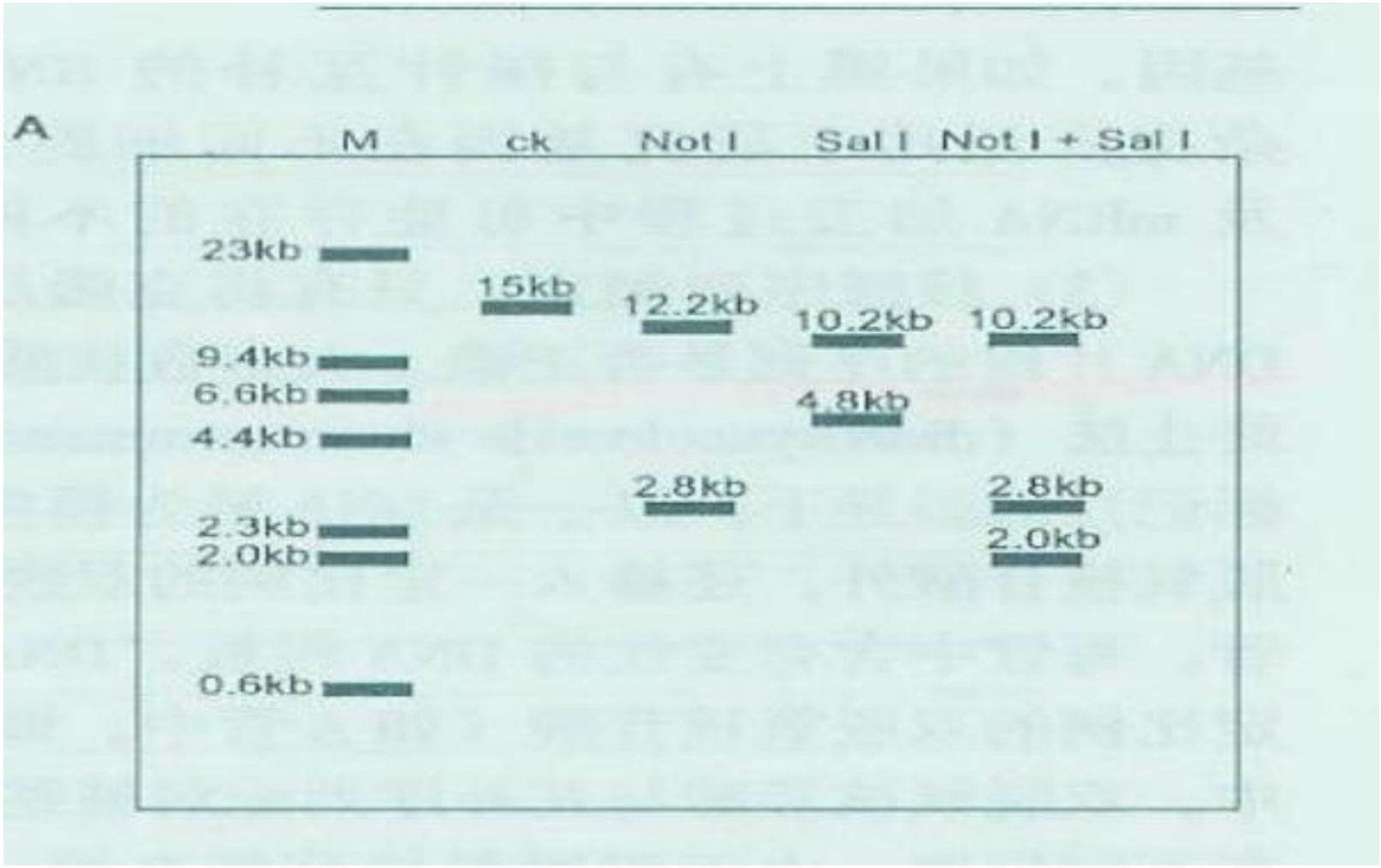
1、限制酶作图

2、依靠克隆的基因组作图

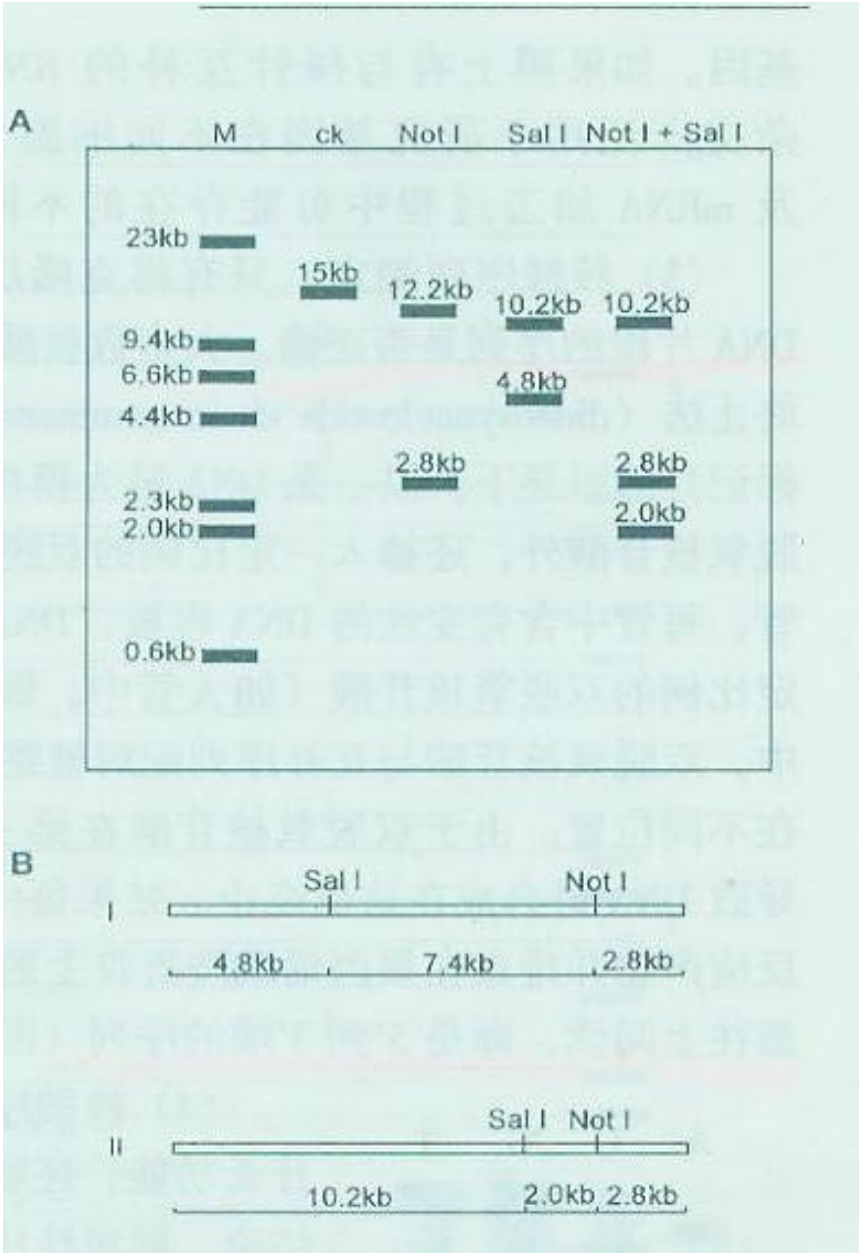
3、荧光原位杂交

4、序列标签位点作图

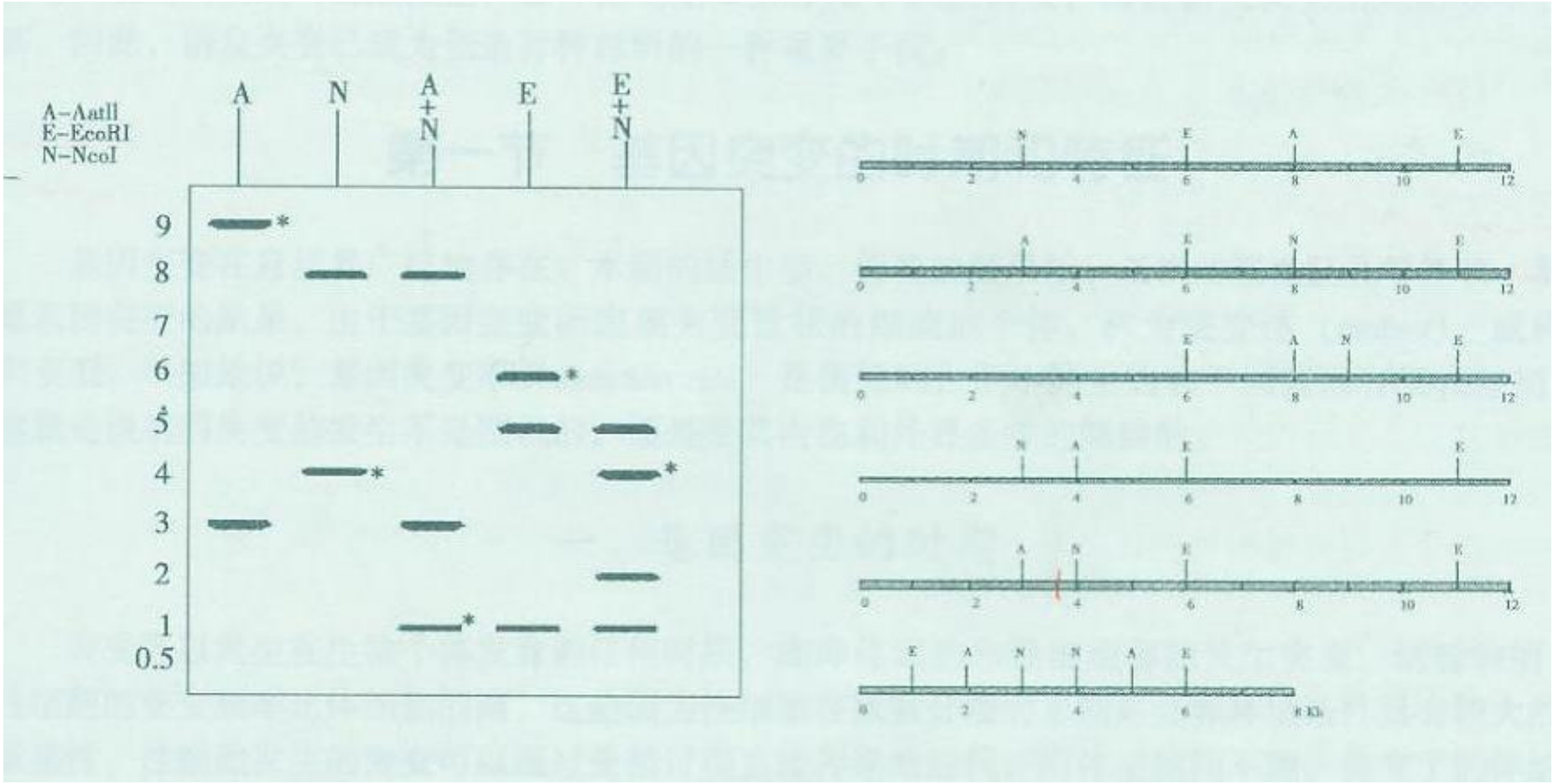
① **限制酶作图：** 比较不同限制酶产生的DNA片段的大小



限制酶作图（**restriction mapping**）



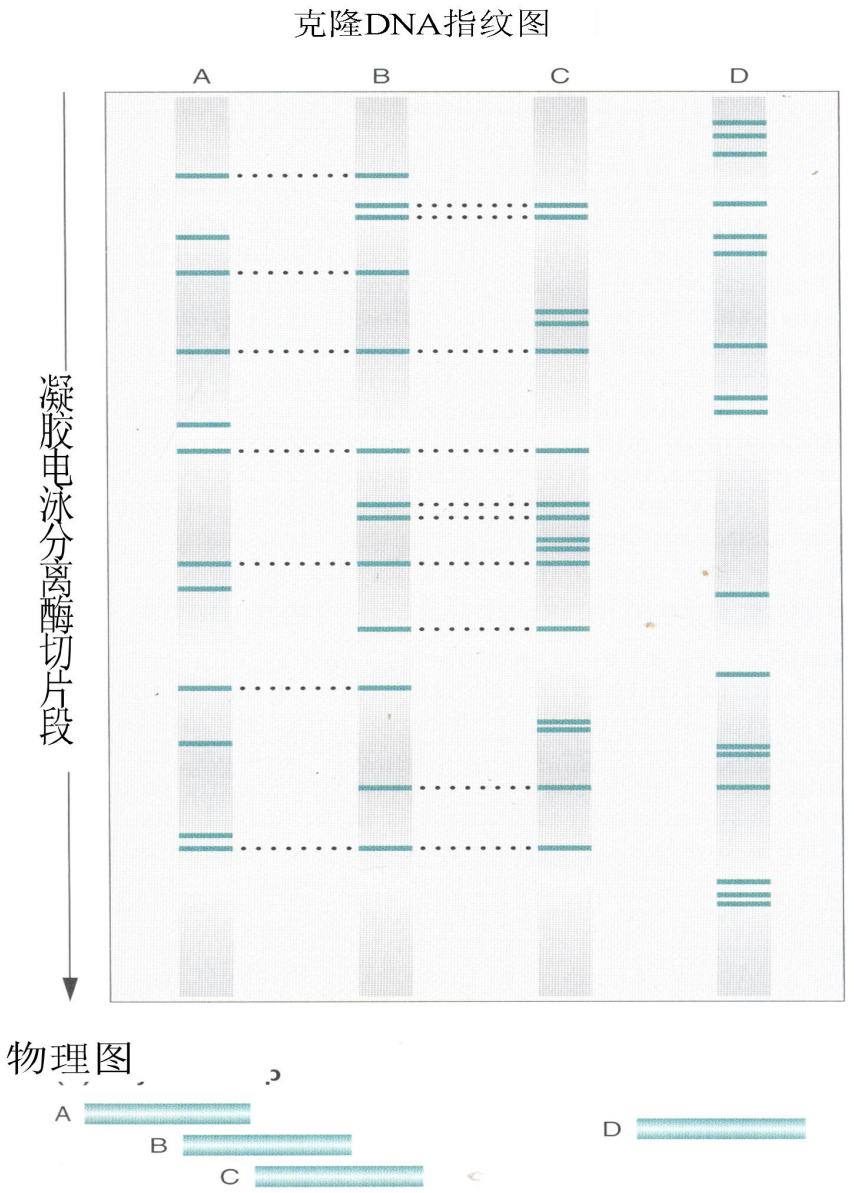
限制酶作图（restriction mapping）



限制酶作图（**restriction mapping**）

② **基于克隆的基因组作图：** 根据克隆的DNA片段之间的重叠顺序构建重叠群 (contig), 绘制物理连锁图。

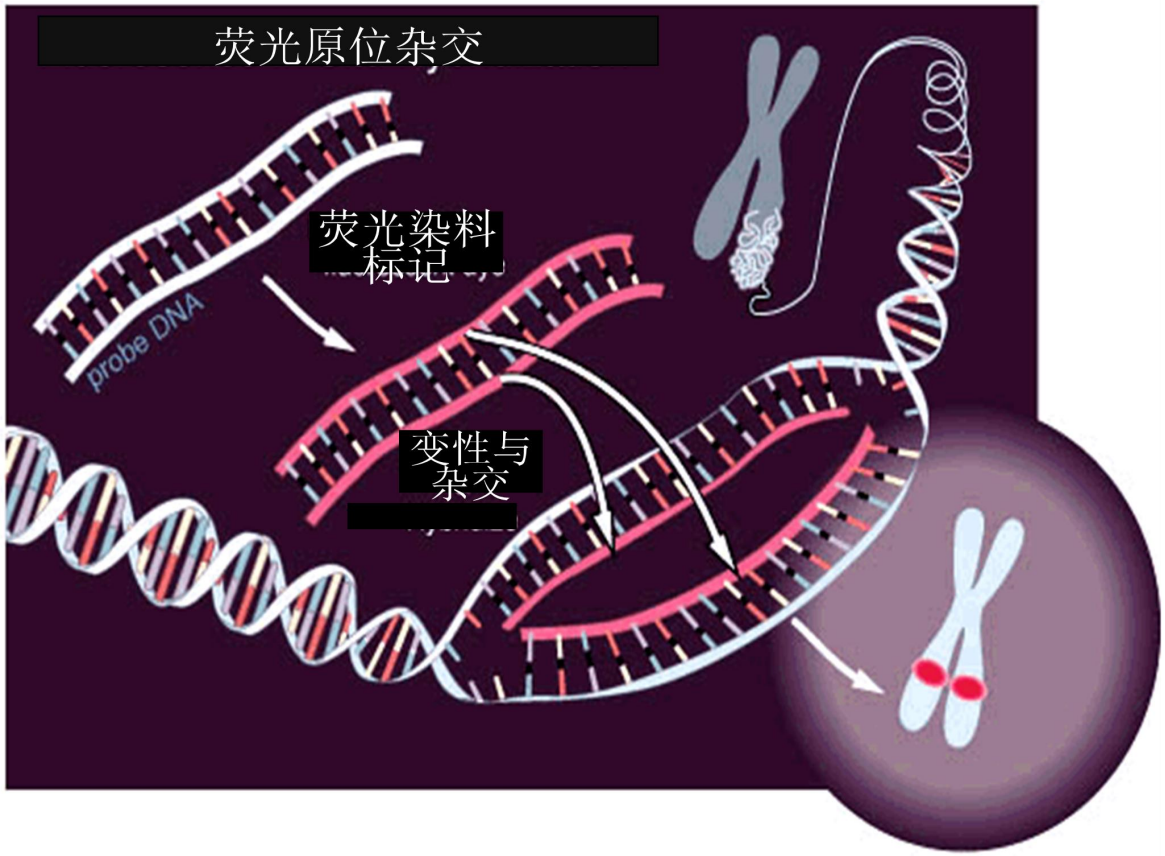
重叠群：相互重叠的DNA片段组成的物理图。克隆重叠群的组建采用染色体步移法



③ 荧光标记原位杂交(FISH)：

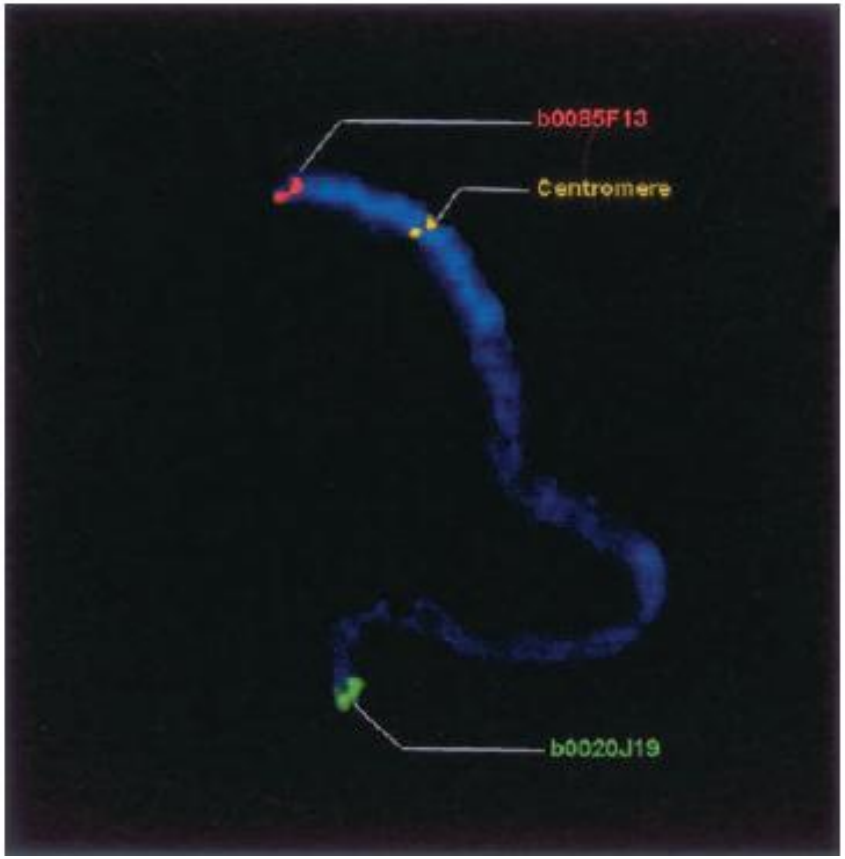
（fluorescent in situ hybridization，FISH）

将荧光标记的探针与染色体杂交确定分子标记所在位置的方法



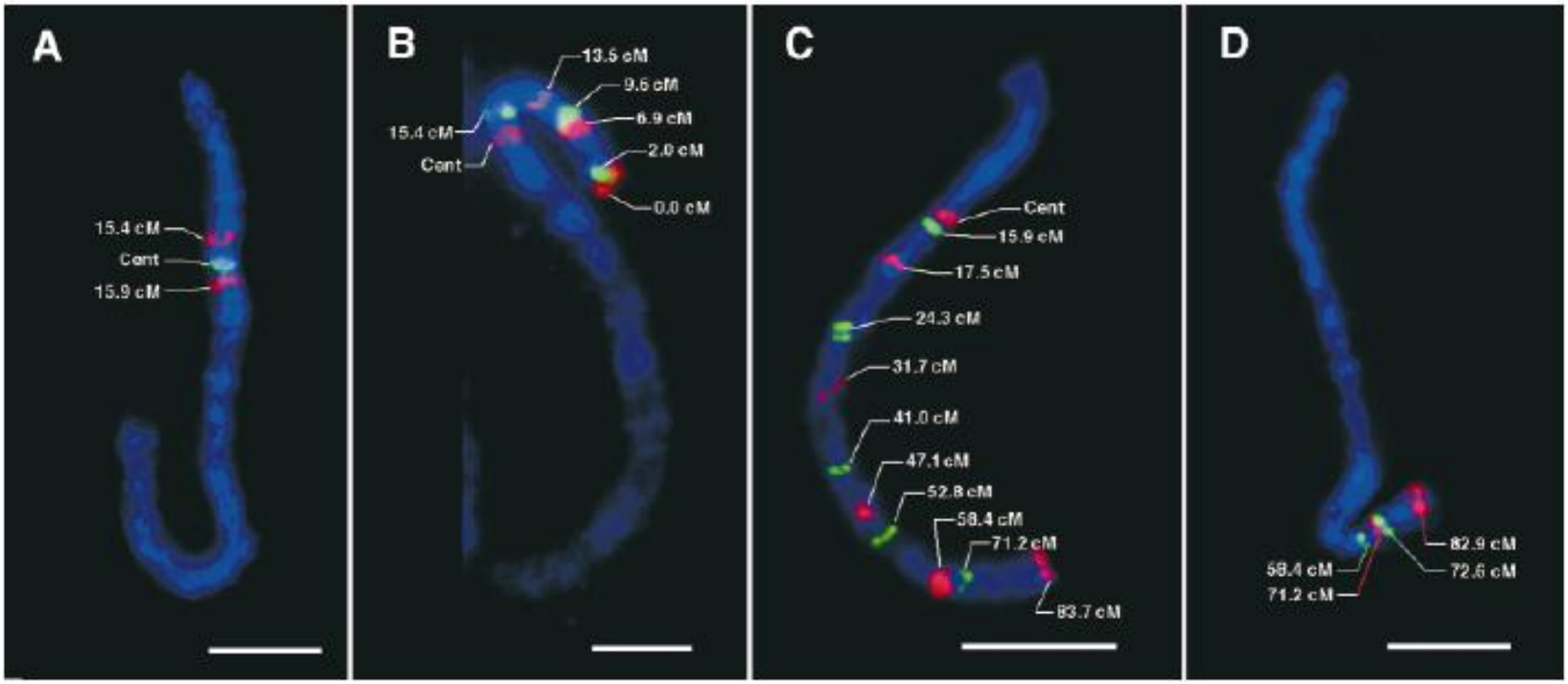
荧光原位杂交

（fluorescent in situ hybridization，FISH）



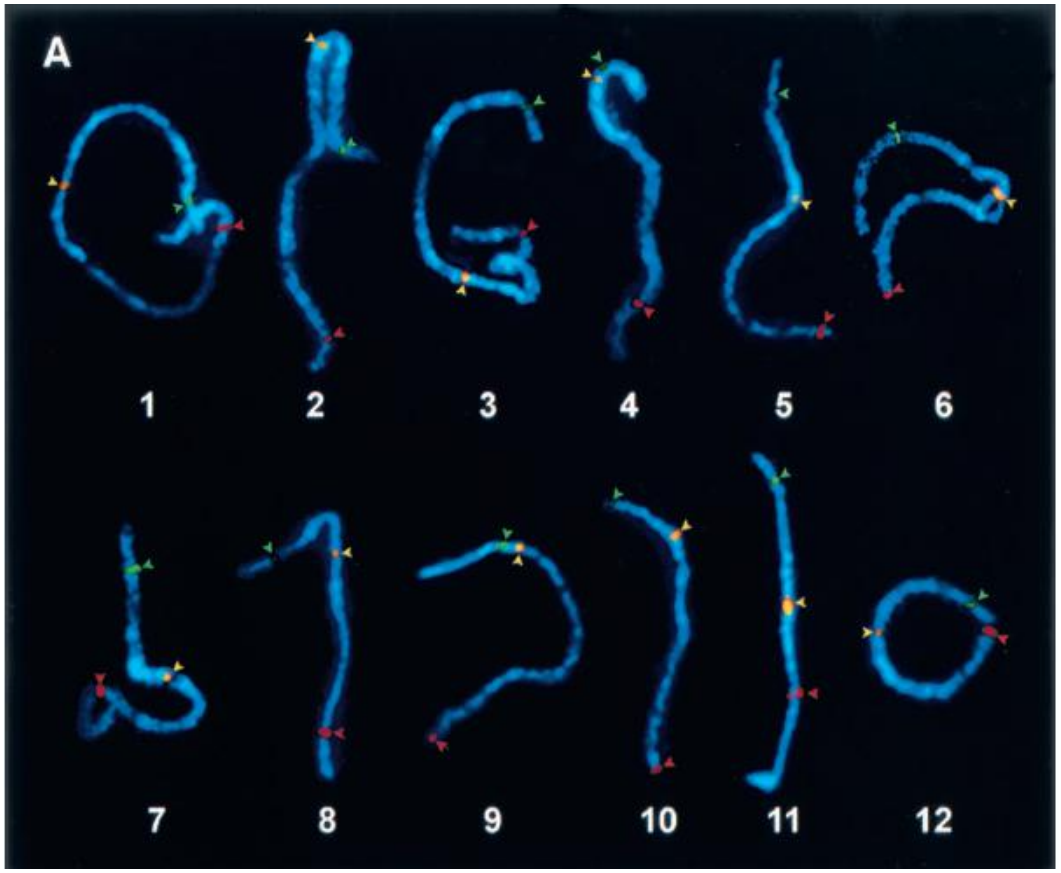
荧光原位杂交

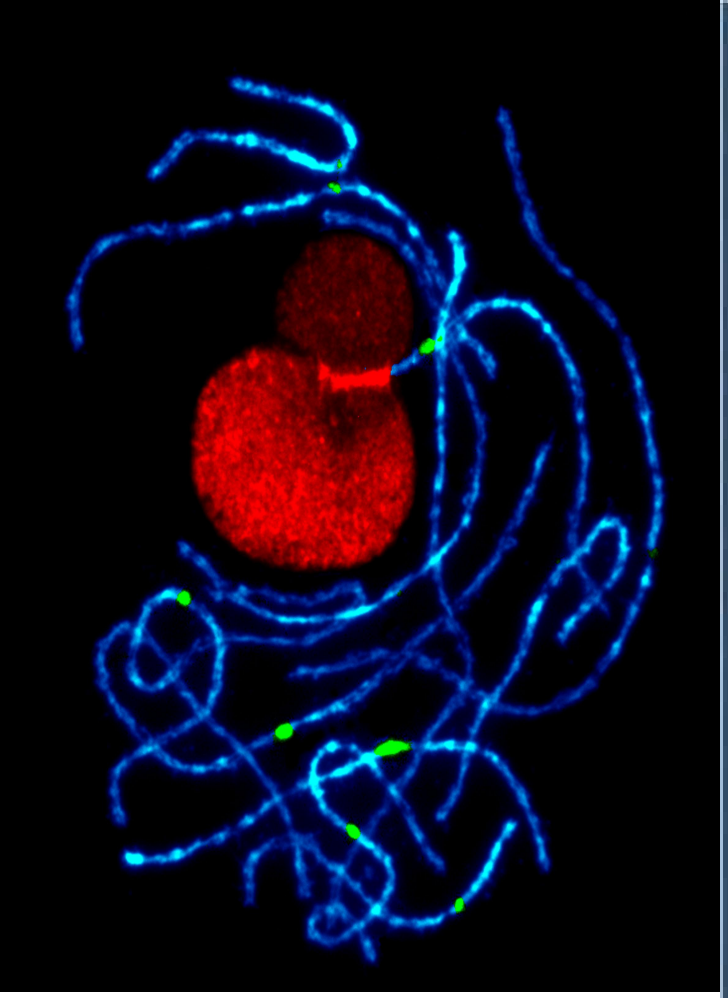
（fluorescent in situ hybridization，FISH）



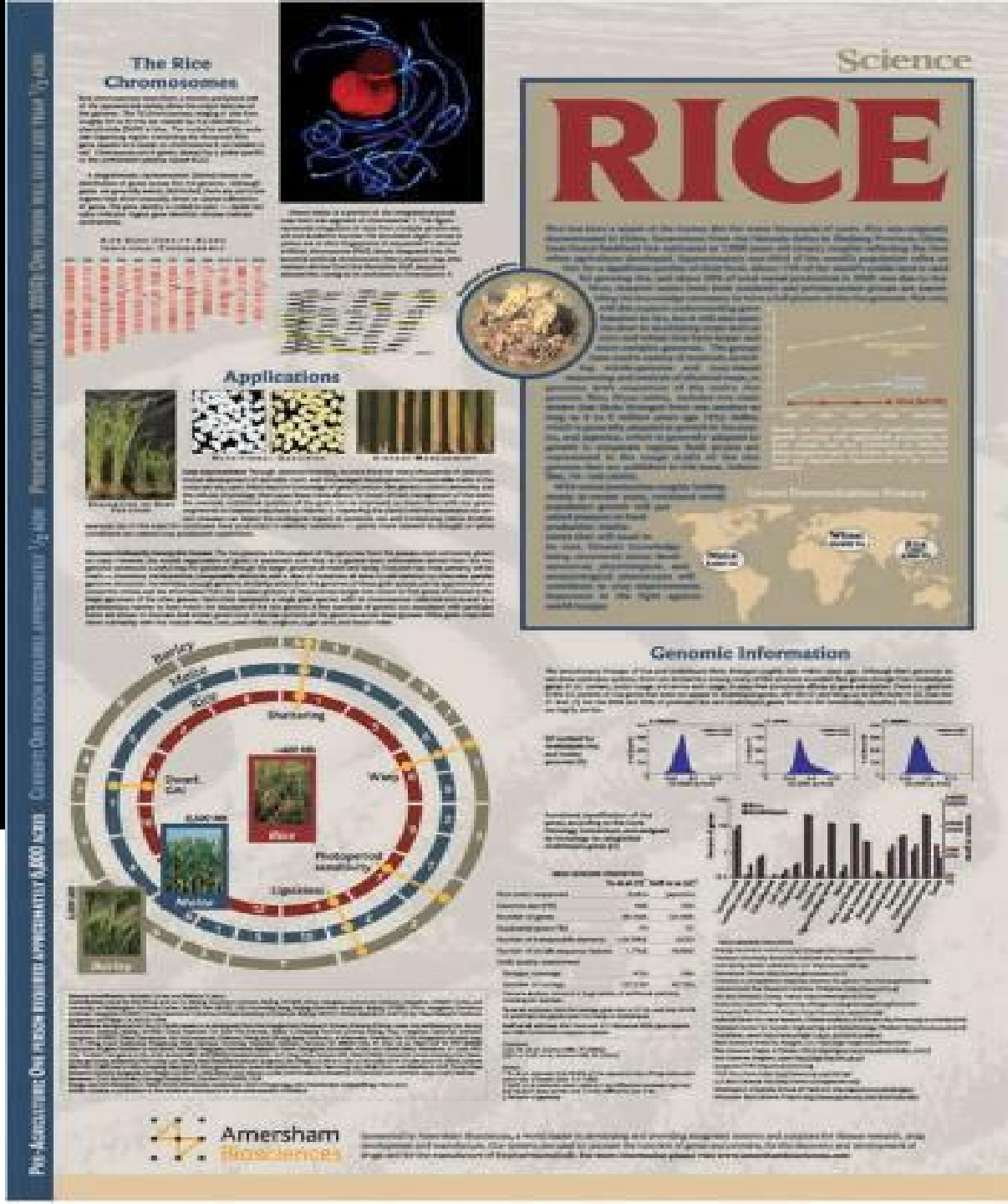
荧光原位杂交

（fluorescent in situ hybridization，FISH）





《*Science*》水稻基因组测序的宣传画



④**序列标签位点**（sequence-tagged site, STS），是已知核苷酸序列的DNA片段，是基因组中任何单拷贝的短DNA序列，长度在100～500bp之间。任何DNA序列，只要知道它在基因组中的位置，都能被用作STS标签。作为基因组中的单拷贝序列，是新一代的遗传标记系统，其数目多，覆盖密度较大，达到平均每1kb一个STS或更密集

人类基因组物理图

- 1987年，RFLP图谱，403个标记，10Mb
- 1994年，5800个标记，0.7Mb
- 1996年，17000多个标记，100kb
- 完全适应全基因组测序的要求

遗传图与物理图的整合

- 有些标记既是遗传标记，又是物理标记

  RFLP标记　SSR标记　某些基因序列

- 借助这些标记可以将遗传图和物理图整合起来

三、基因组图谱的应用

1、基因组序列测定

2、基因定位

3、基因的克隆与分离

4、分子标记辅助选择

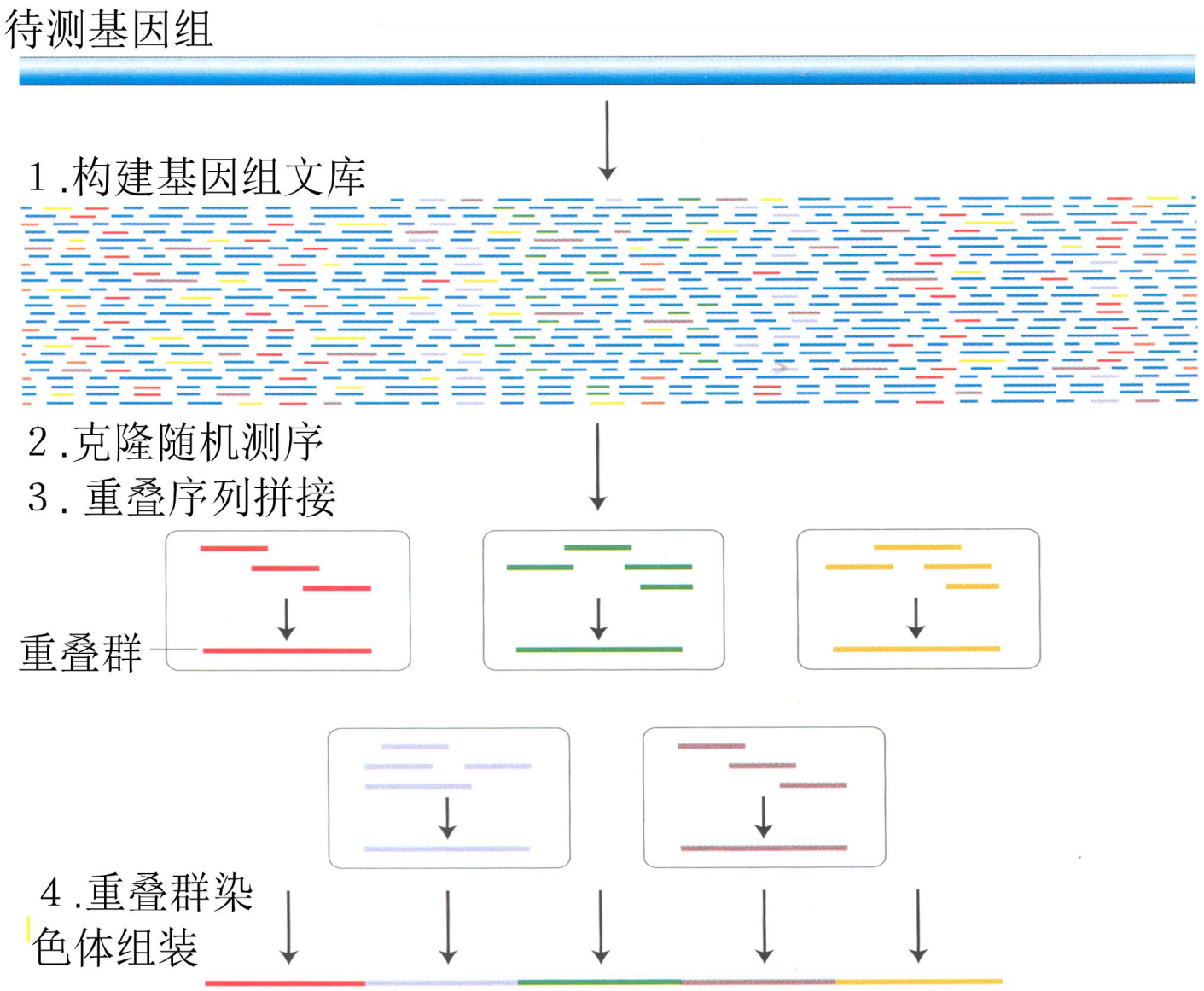
5、比较基因组研究

1、基因组测序策略

- 有了高密度的基因组图谱，就可以开始全基因组测序了
- 测序的技术飞速发展，现在可以全自动化

①鸟枪法测序

用限制酶或超声波处理待测序基因组



鸟枪法的优缺点

- 优点：
  - 不需要高密度的图谱
  - 速度快、简单、成本低
- 缺点：
  - 拼接组装困难，尤其在重复序列多的区域

- 主要用于重复序列少、相对简单的原核生物基因组

第一代测序技术

第一代DNA测序技术用的是1975年由桑格（Sanger）和考尔森（Coulson）开创的链终止法或者是1976-1977年由马克西姆（Maxam）和吉尔伯特（Gilbert）发明的化学法（链降解）.并在1977年，桑格测定了第一个基因组序列，是噬菌体X174的，全长5375个碱基1。自此，人类获得了窥探生命遗传差异本质的能力，并以此为开端步入基因组学时代。研究人员在Sanger法的多年实践之中不断对其进行改进。在2001年，完成的首个人类基因组图谱就是以改进了的Sanger法为其测序基础，Sanger法核心原理是：由于ddNTP的2'和3'都不含羟基，其在DNA的合成过程中不能形成磷酸二酯键，因此可以用来中断DNA合成反应，在4个DNA合成反应体系中分别加入一定比例带有放射性同位素标记的ddNTP（分为：ddATP,ddCTP,ddGTP和ddTTP），通过凝胶电泳和放射自显影后可以根据电泳带的位置确定待测分子的DNA序列

第一代测序技术的主要特点是测序读长可达1000bp，准确性高达99.999%

Dr. Fred Sanger

Frederick Sanger was awarded the prize in both 1958 and 1980. He is the fourth person in the world to have been awarded two Nobel Prizes and the only person to receive both in chemistry.

"dideoxy" sequencing technique (Sanger et al., 1977)

DNA双脱氧链终止法测序

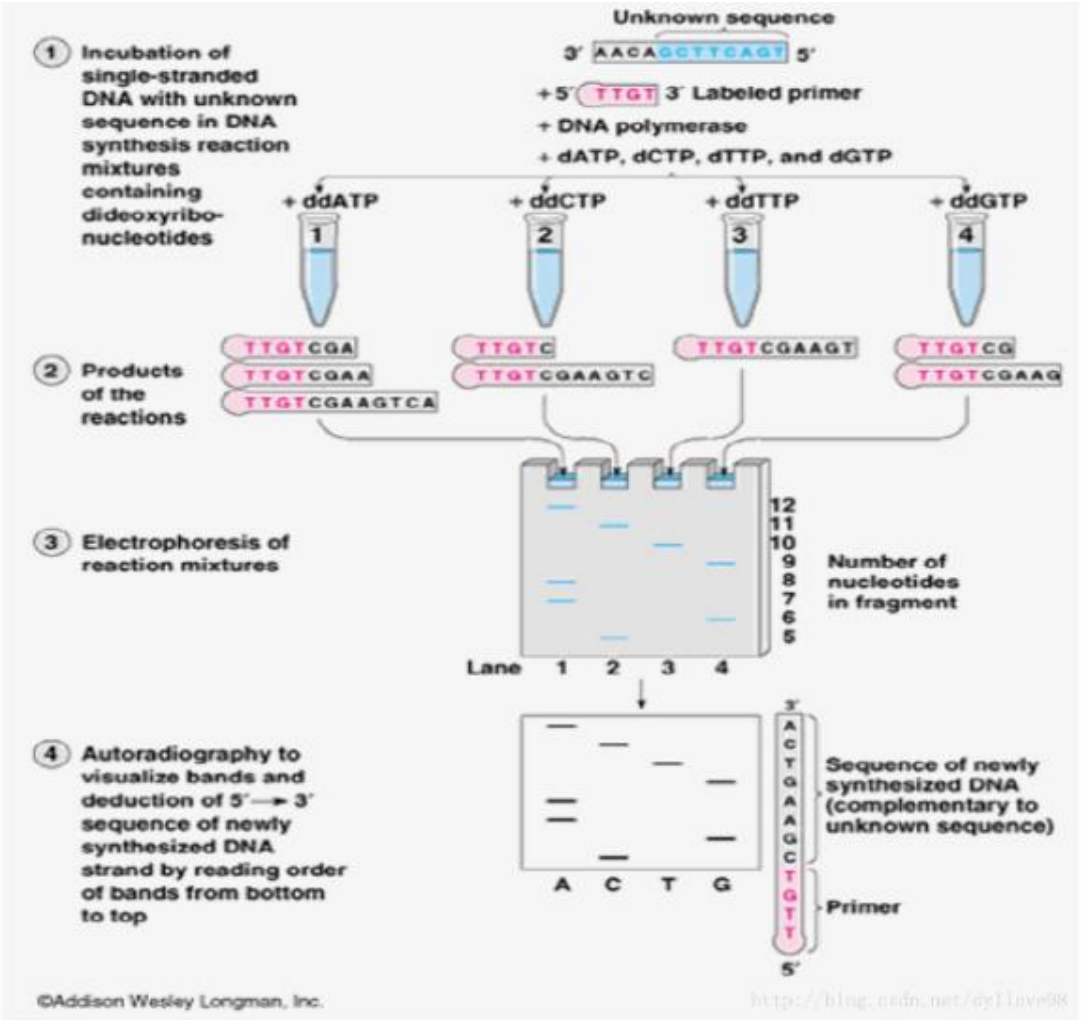


图2：Sanger法测序原理

第二代测序技术

总的说来，第一代测序技术的主要特点是测序读长可达1000bp，准确性高达99.999%，但其测序成本高，通量低等方面的缺点，严重影响了其真正大规模的应用。因而第一代测序技术并不是最理想的测序方法。经过不断的技术开发和改进，以Roche公司的454技术、illumina公司的Solexa，Hiseq技术和ABI公司的Solid技术为标记的第二代测序技术诞生了。第二代测序技术大大降低了测序成本的同时，还大幅提高了测序速度，并且保持了高准确性，以前完成一个人类基因组的测序需要3年时间，而使用二代测序技术则仅仅需要1周，但在序列读长方面比起第一代测序技术则要短很多。表1和图3对第一代和第二代测序技术各自的特点以及测序成本作了一个简单的比较，以下我将对这三种主要的第二代测序技术的主要原理和特点作一个简单的介绍。

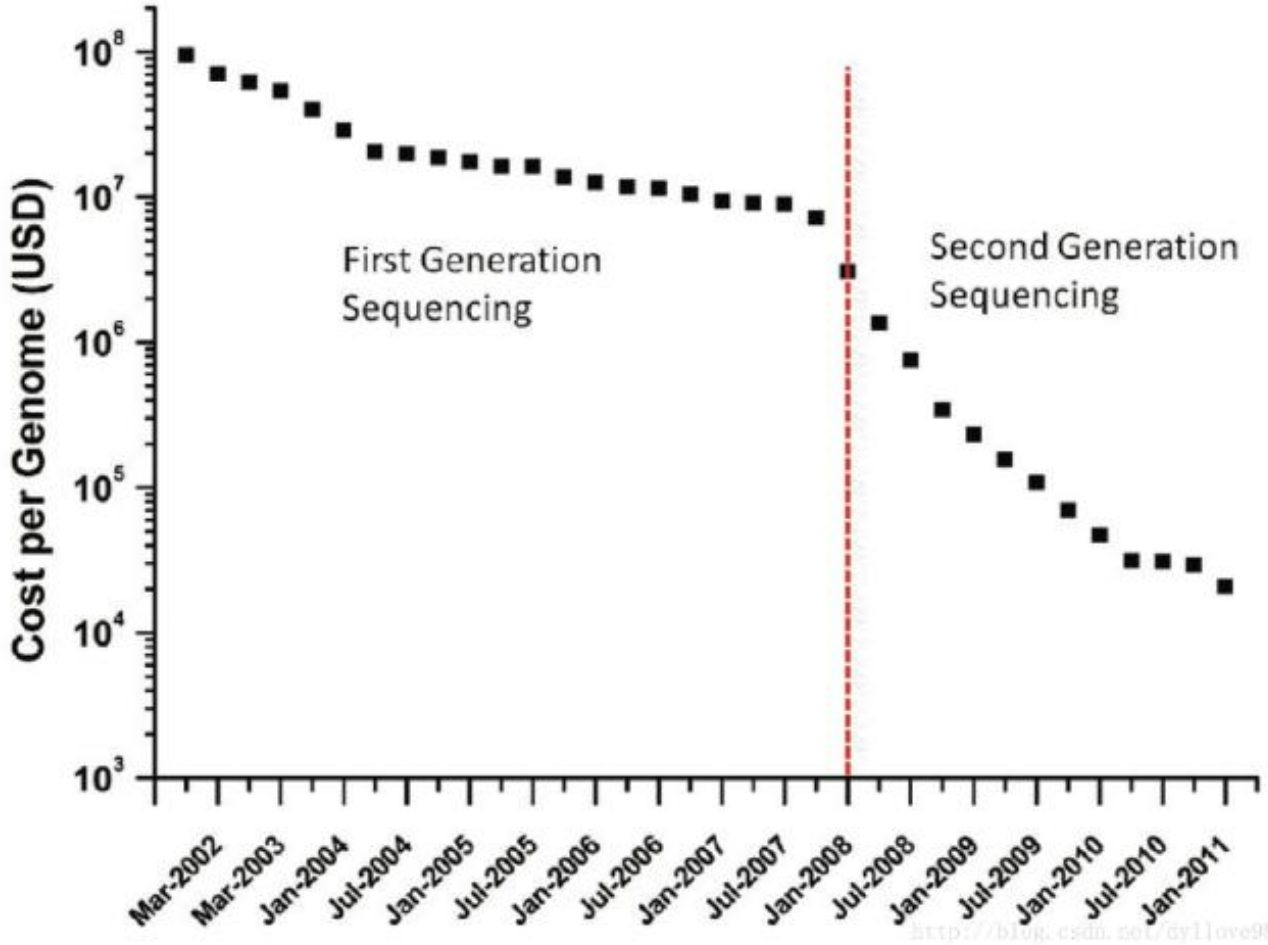

Cost per Genome (USD)
10^8
10^7
10^6
10^5
10^4
10^3
First Generation Sequencing
Second Generation Sequencing
Mar-2002
Mar-2003
Jan-2004
Jul-2004
Jan-2005
Jul-2005
Jan-2006
Jul-2006
Jan-2007
Jul-2007
Jan-2008
Jul-2008
Jan-2009
Jul-2009
Jan-2010
Jul-2010
Jan-2011

图3. 测序成本的变化

1.Illumine

Illumina公司的Solexa和Hiseq应该说是目前全球使用量最大的第二代测序机器，这两个系列的技术核心原理是相同的2,4。这两个系列的机器采用的都是边合成边测序的方法，它的测序过程主要分为以下4步，如图4.

（1）DNA待测文库构建

利用超声波把待测的DNA样本打断成小片段，目前除了组装之外和一些其他的特殊要求之外，主要是打断成200-500bp长的序列片段，并在这些小片段的两端添加上不同的接头，构建出单链DNA文库。

（2）Flowcell

Flowcell是用于吸附流动DNA片段的槽道，当文库建好后，这些文库中的DNA在通过flowcell的时候会随机附着在flowcell表面的channel上。每个Flowcell有8个channel，每个channel的表面都附有很多接头，这些接头能和建库过程中加在DNA片段两端的接头相互配对（这就是为什么flowcell能吸附建库后的DNA的原因），并能支持DNA在其表面进行桥式PCR的扩增。

（3）桥式PCR扩增与变性

桥式PCR以Flowcell表面所固定的接头为模板，进行桥形扩增，如图4.a所示。经过不断的扩增和变性循环，最终每个DNA片段都将在各自的位置上集中成束，每一个束都含有单个DNA模板的很多分拷贝，进行这一过程的目的在于实现将碱基的信号强度放大，以达到测序所需的信号要求。

（4）测序

测序方法采用边合成边测序的方法。向反应体系中同时添加DNA聚合酶、接头引物和带有碱基特异荧光标记的4中dNTP（如同Sanger测序法）。这些dNTP的3'-OH被化学方法所保护，因而每次只能添加一个dNTP。在dNTP被添加到合成链上后，所有未使用的游离dNTP和DNA聚合酶会被洗脱掉。接着，再加入激发荧光所需的缓冲液，用激光激发荧光信号，并有光学设备完成荧光信号的记录，最后利用计算机分析将光学信号转化为测序碱基。这样荧光信号记录完成后，再加入化学试剂淬灭荧光信号并去除dNTP 3'-OH保护基团，以便能进行下一轮的测序反应。Illumina的这种测序技术每次只添加一个dNTP的特点能够很好的地解决同聚物长度的准确测量问题，它的主要测序错误来源是碱基的替换，目前它的测序错误率在1%-1.5%之间，测序周期以人类基因组重测序为例，30x测序深度大约为1周。

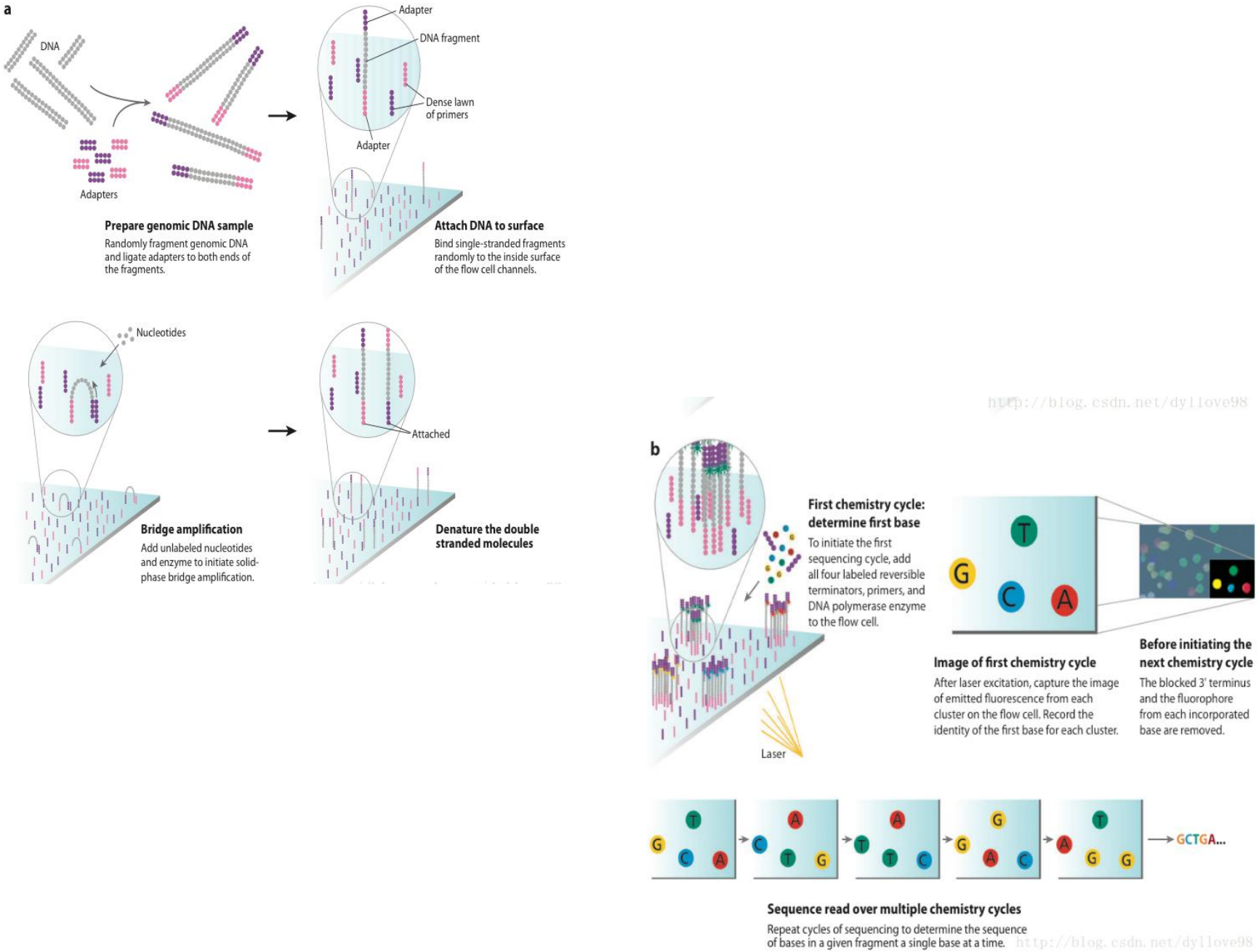


图4. Illumina测序流程

1.Roche 454

Roche 454测序系统是第一个商业化运营二代测序技术的平台。它的主要测序原理是（图5 abc）2：

（1）DNA文库制备

454测序系统的文件构建方式和illumina的不同，它是利用喷雾法将待测DNA打断成300-800bp长的小片段，并在片段两端加上不同的接头，或将待测DNA变性后用杂交引物进行PCR扩增，连接载体，构建单链DNA文库（图5a）。

（2）Emulsion PCR （乳液PCR，其实是一个注水到油的独特过程）

454当然DNA扩增过程也和illumina的截然不同，它将这些单链DNA结合在水油包被的直径约28um的磁珠上，并在其上面孵育、退火。

乳液PCR最大的特点是可以形成数目庞大的独立反应空间以进行DNA扩增。其关键技术是“注水到油”（水包油），基本过程是在PCR反应前，将包含PCR所有反应成分的水溶液注入到高速旋转的矿物油表面，水溶液瞬间形成无数个被矿物油包裹的小水滴。这些小水滴就构成了独立的PCR反应空间。理想状态下，每个小水滴只含一个DNA模板和一个磁珠。

这些被小水滴包被的磁珠表面含有与接头互补的DNA序列，因此这些单链DNA序列能够特异地结合在磁珠上。同时孵育体系中含有PCR反应试剂，所以保证了每个与磁珠结合的小片段都能独立进行PCR扩增，并且扩增产物仍可以结合到磁珠上。当反应完成后，可以破坏孵育体系并将带有DNA的磁珠富集下来。进过扩增，每个小片段都将被扩增约100万倍，从而达到下一步测序所要求的DNA量。

（3）焦磷酸测序

测序前需要先用一种聚合酶和单链结合蛋白处理带有DNA的磁珠，接着将磁珠放在一种PTP平板上。这种平板上特制有许多直径约为44um的小孔，每个小孔仅能容纳一个磁珠，通过这种方法来固定每个磁珠的位置，以便检测接下来的测序反应过程。测序方法采用焦磷酸测序法，将一种比PTP板上小孔直径更小的磁珠放入小孔中，启动测序反应。测序反应以磁珠上大量扩增出的单链DNA为模板，每次反应加入一种dNTP进行合成反应。如果dNTP能与待测序列配对，则会在合成后释放焦磷酸基团。释放的焦磷酸基团会与反应体系中的ATP硫酸化学酶反应生成ATP。生成的ATP和荧光素酶共同氧化使测序反应中的荧光素分子并发出荧光，同时由PTP板另一侧的CCD照相机记录，最后通过计算机进行光信号处理而获得最终的测序结果。由于每一种dNTP在反应中产生的荧光颜色不同，因此可以根据荧光的颜色来判断被测分子的序列。反应结束后，游离的dNTP会在双磷酸酶的作用下降解ATP，从而导致荧光淬灭，以便使测序反应进入下一个循环。由于454测序技术中，每个测序反应都在PTP板上独立的小孔中进行，因而能大大降低相互间的干扰和测序偏差。454技术最大的优势在于其能获得较长的测序读长，当前454技术的平均读长可达400bp，并且454技术和illumina的Solexa和Hiseq技术不同，它最主要的一个缺点是无法准确测量同聚物的长度，如当序列中存在类似于PolyA的情况时，测序反应会一次加入多个T，而所加入的T的个数只能通过荧光强度推测获得，这就有可能导致结果不准确。也正是由于这一原因，454技术会在测序过程中引入插入和缺失的测序错误。

第三代测序技术

测序技术在近两三年中又有新的里程碑。以PacBio公司的SMRT和Oxford Nanopore Technologies纳米孔单分子测序技术，被称之为第三代测序技术。与前两代相比，他们最大的特点就是单分子测序，测序过程无需进行PCR扩增。

其中PacBio SMRT技术其实也应用了边合成边测序的思想5，并以SMRT芯片为测序载体。基本原理是： DNA聚合酶和模板结合,4色荧光标记 4 种碱基（即是dNTP）,在碱基配对阶段,不同碱基的加入,会发出不同光,根据光的波长与峰值可判断进入的碱基类型。同时这个 DNA 聚合酶是实现超长读长的关键之一,读长主要跟酶的活性保持有关,它主要受激光对其造成的损伤所影响。PacBio SMRT技术的一个关键是怎样将反应信号与周围游离碱基的强大荧光背景区别出来。他们利用的是ZMW（零模波导孔）原理：如同微波炉壁上可看到的很多密集小孔。小孔直径有考究,如果直径大于微波波长,能量就会在衍射效应的作用下穿透面板而泄露出来，从而与周围小孔相互干扰。如果孔径小于波长,能量不会辐射到周围，而是保持直线状态（光衍射的原理）,从而可起保护作用。同理,在一个反应管(SMRTCell:单分子实时反应孔)中有许多这样的圆形纳米小孔, 即 ZMW(零模波导孔),外径 100多纳米,比检测激光波长小(数百纳米),激光从底部打上去后不能穿透小孔进入上方溶液区,能量被限制在一个小范围(体积20X 10-21 L)里,正好足够覆盖需要检测的部分,使得信号仅来自这个小反应区域,孔外过多游离核苷酸单体依然留在黑暗中,从而实现将背景降到最低。另外，可以通过检测相邻两个碱基之间的测序时间，来检测一些碱基修饰情况，既如果碱基存在修饰，则通过聚合酶时的速度会减慢，相邻两峰之间的距离增大，可以通过这个来之间检测甲基化等信息（图7）。SMRT技术的测序速度很快，每秒约10个dNTP。但是，同时其测序错误率比较高（这几乎是目前单分子测序技术的通病），达到15%,但好在它的出错是随机的，并不会像第二代测序技术那样存在测序错误的偏向，因而可以通过多次测序来进行有效的纠错。

Oxford Nanopore Technologies公司所开发的纳米单分子测序技术与以往的测序技术皆不同，它是基于电信号而不是光信号的测序技术5。该技术的关键之一是，他们设计了一种特殊的纳米孔，孔内共价结合有分子接头。当DNA碱基通过纳米孔时，它们使电荷发生变化，从而短暂地影响流过纳米孔的电流强度（每种碱基所影响的电流变化幅度是不同的），灵敏的电子设备检测到这些变化从而鉴定所通过的碱基（图8）。

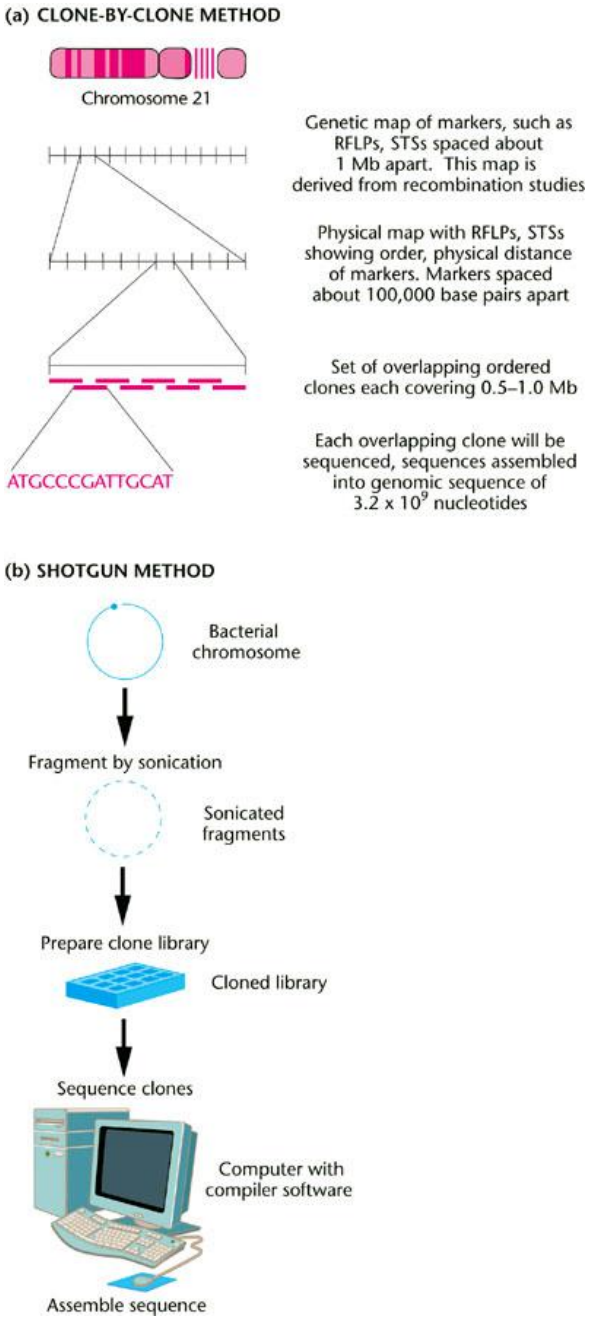
该公司在去年基因组生物学技术进展年会(AGBT)上推出第一款商业化的纳米孔测序仪，引起了科学界的极大关注。纳米孔测序（和其他第三代测序技术）有望解决目前测序平台的不足，纳米孔测序的主要特点是：读长很长，大约在几十kb，甚至100 kb;错误率目前介于1%至4%，且是随机错误，而不是聚集在读取的两端;数据可实时读取;通量很高(30x人类基因组有望在一天内完成);起始DNA在测序过程中不被破坏;以及样品制备简单又便宜。理论上，它也能直接测序RNA。

纳米孔单分子测序计算还有另一大特点，它能够直接读取出甲基化的胞嘧啶，而不必像传统方法那样对基因组进行bisulfite处理。这对于在基因组水平直接研究表观遗传相关现象有极大的帮助。并且改方法的测序准确性可达99.8%，而且一旦发现测序错误也能较容易地进行纠正。但目前似乎还没有应用该技术的相关报道。

2克隆重叠群法（clone contig）

- 将基因组DNA切割长度为0.1Mb－1Mb的大片段，克隆到YAC或BAC载体上
- 然后再进行亚克隆，分别测定单个亚克隆的序列
- 再装配、连接成连续的DNA分子。
- 这是一种自上而下（up to down）的测序策略
- clone-by-clone method

两种基因组测序策略

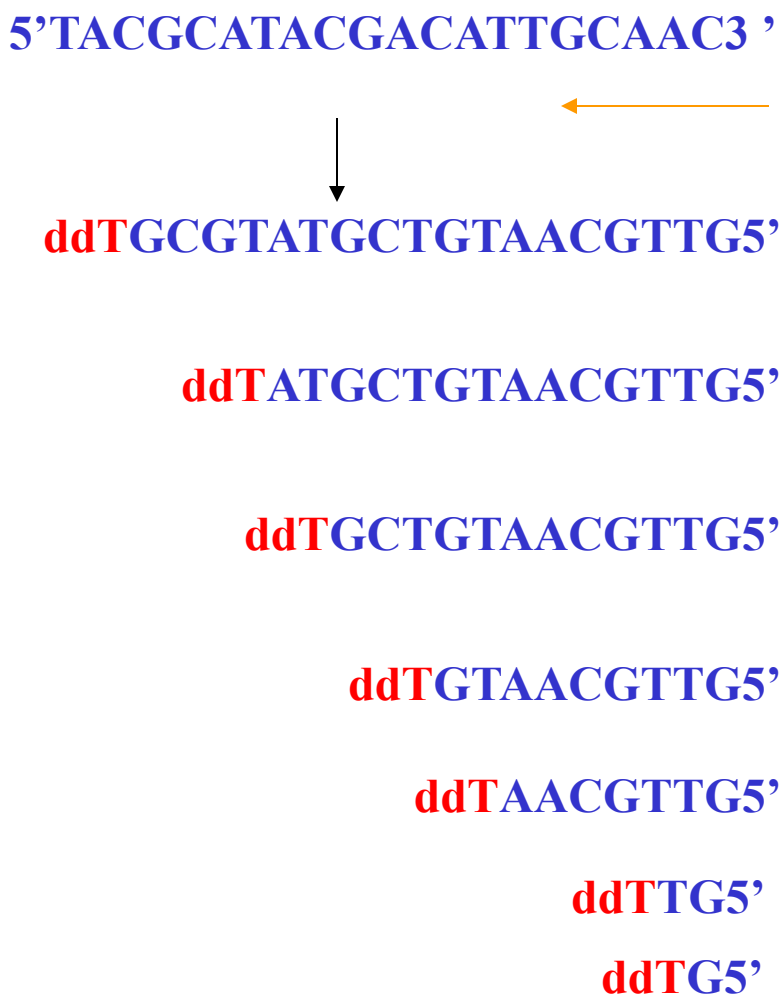

(a) CLONE-BY-CLONE METHOD
Chromosome 21
Genetic map of markers, such as RFLPs, STSs spaced about 1 Mb apart. This map is derived from recombination studies
Physical map with RFLPs, STSs showing order, physical distance of markers. Markers spaced about 100,000 base pairs apart
Set of overlapping ordered clones each covering 0.5–1.0 Mb
ATGCCCGATTGCAT
Each overlapping clone will be sequenced, sequences assembled into genomic sequence of 3.2 x 10^9 nucleotides
(b) SHOTGUN METHOD
Bacterial chromosome
Fragment by sonication
Sonicated fragments
Prepare clone library
Cloned library
Sequence clones
Computer with compiler software
Assemble sequence

序列测定和分析：

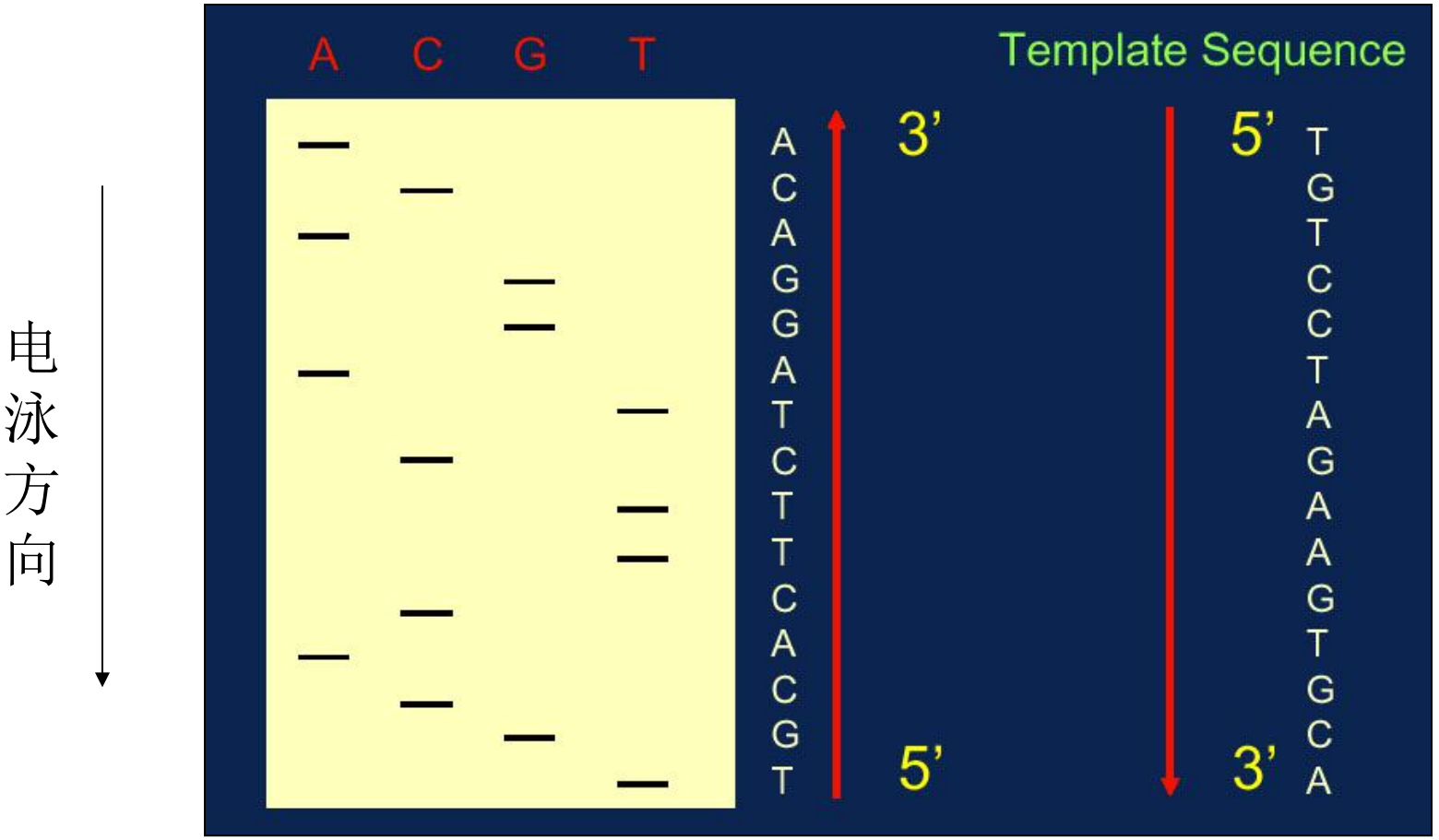
从文库中筛选获得目标克隆后，只有进行测序才能进行生物信息分析和功能预测。

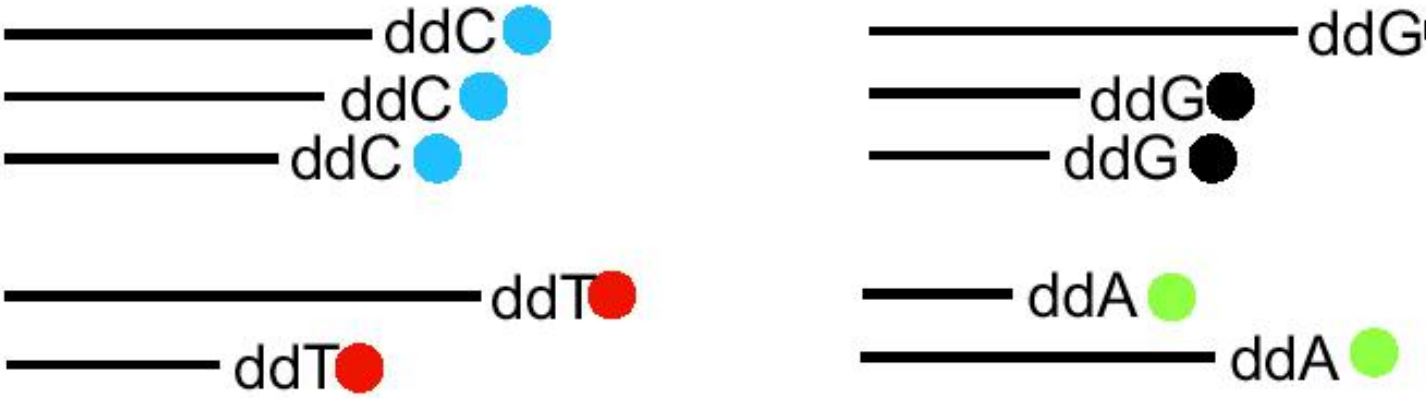
一般采用Sanger(1977)发明的双脱氧核糖核酸终止法(dideoxynucleotide chain termination)测定核酸序列。

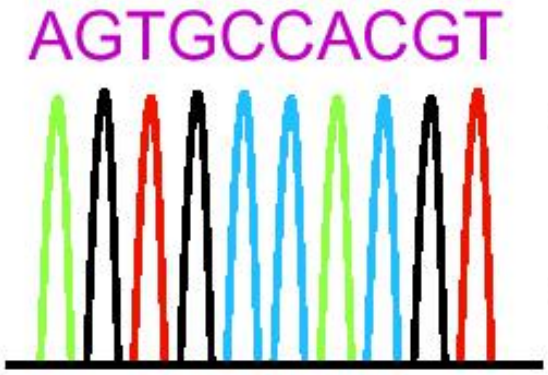
Sanger双脱氧链终止法



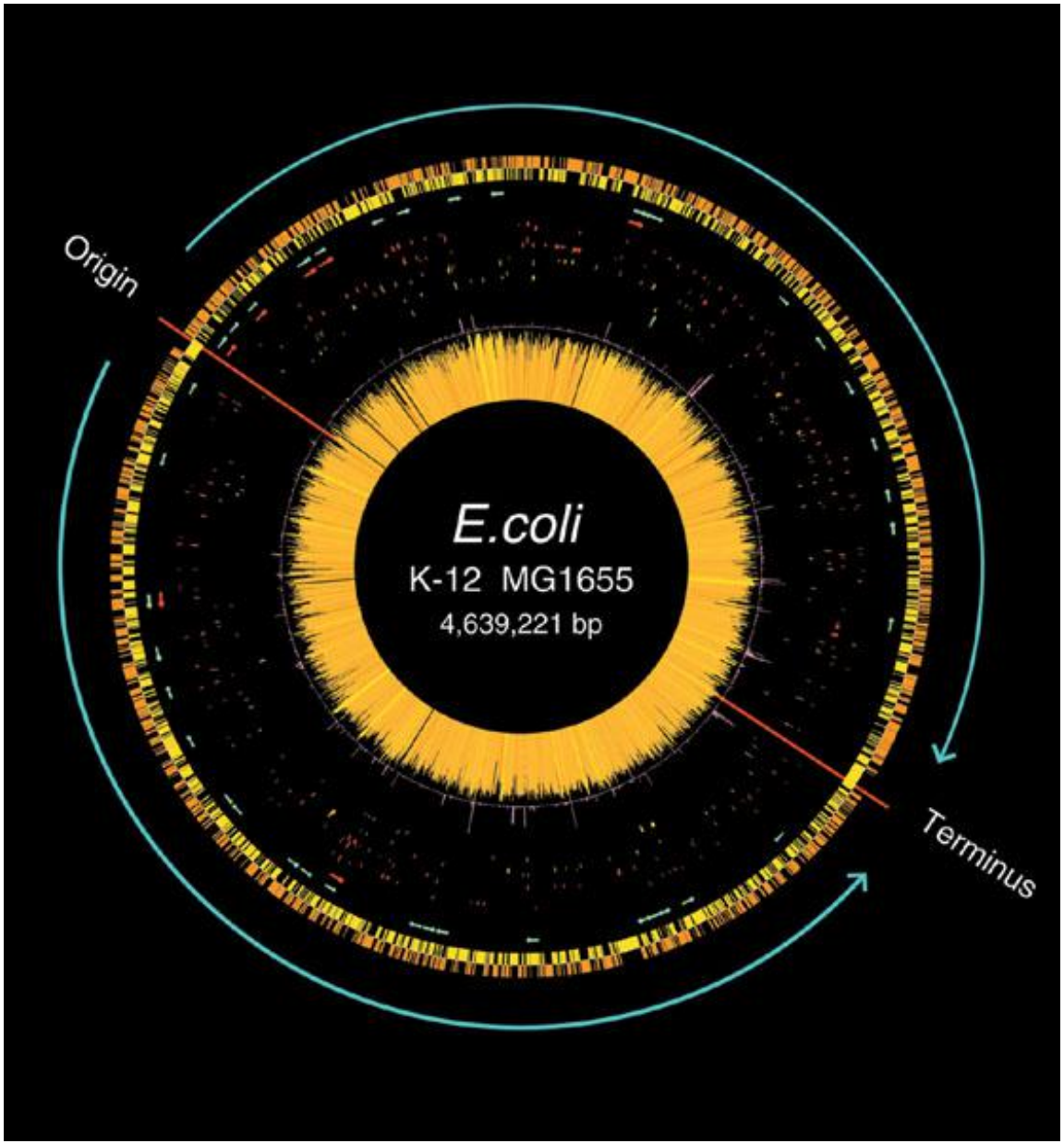
- 待测序的DNA模板
- 引物
- DNA聚合酶
- dNTP
- ddTTP


A C G T
Template Sequence
3’
5’
5’
3’
ACAGGATCTTTCACGT
TGTCCTAGAAAGTGCA
电泳方向

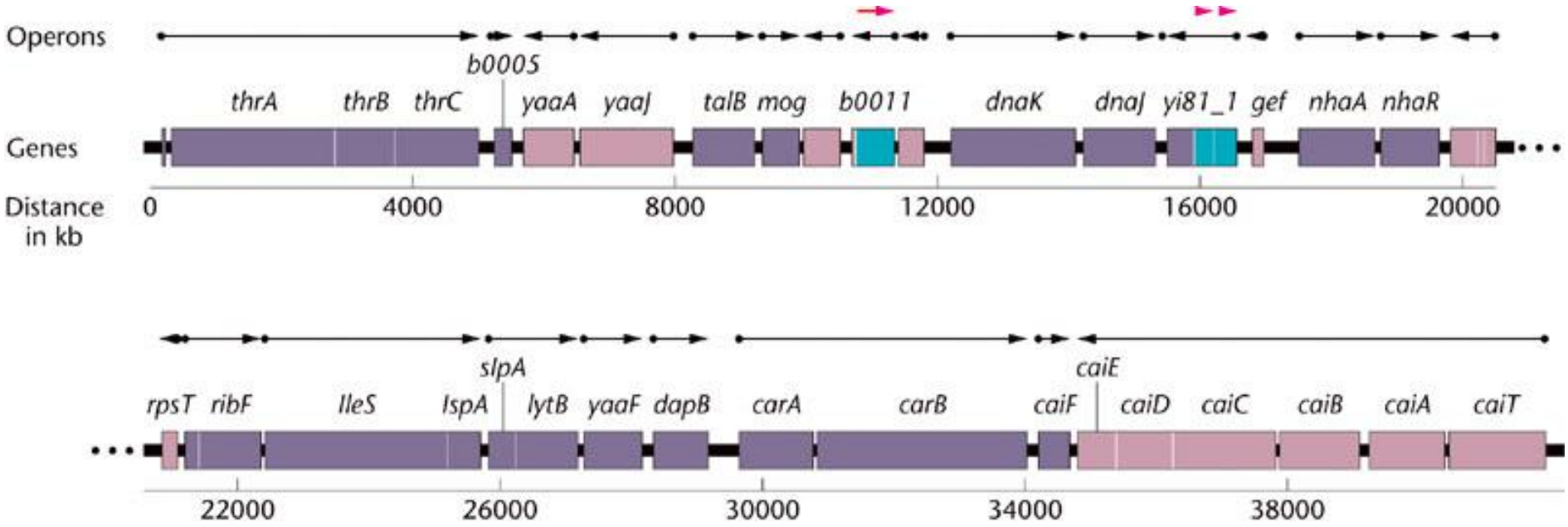




- Use fluorescently-labeled ddNTPs in chain termination reactions;
- Detect by a imaging system



The *E.coli* genome



A portion of the *E.coli* chromosome showing genes and operons.

A dot indicates the promoter for each gene or operon. Arrows and color indicate the direction of transcribtion: dark blue genes are transcribed left to right, light blue are transcribed right to left.Overlapping gene are shown in green.

四、功能基因组学

- 完成基因组测序，仅仅是基因组计划的第一步，更重要的工作在于弄清楚：

  ①基因组序列中所包含的全部遗传信息是什么；

  ②基因组作为一个整体如何行使功能。

- 就是对基因组序列进行诠释的过程，也就是功能基因组学的研究内容。

根据序列分析搜寻基因

- 查找开放阅读框（open reading frame，ORF）
- 开放阅读框都有一个起始密码子，ATG，还要有终止密码子。
- 从ATG开始，然后向下游寻找终止密码子。
- 起始密码子和终止密码子之间的碱基数目要能够被3整除

同源查询

- 利用已经存入数据库的基因序列与待查的基因组序列比对，从中查找可以与之匹配的碱基序列及其比例，用于界定基因。

- 同源查询可以部分弥补ORF扫描的不足。

同源查询的依据

有亲缘关系的物种，基因组可能存在某种程度的相似性：

- 存在某些完全相同的序列；
- ORF的排列相似，如等长的外显子；
- ORF指令的氨基酸序列相似；
- 模拟的多肽链的高级结构相似，等。

基因功能研究

1、计算机预测基因功能

- 依据仍然是同源性比较。同源基因拥有一个共同的祖先基因，它们之间有许多相似的序列。
- 种间同源基因
- 种内同源基因

基因功能研究

2、实验确认基因功能

基因克隆

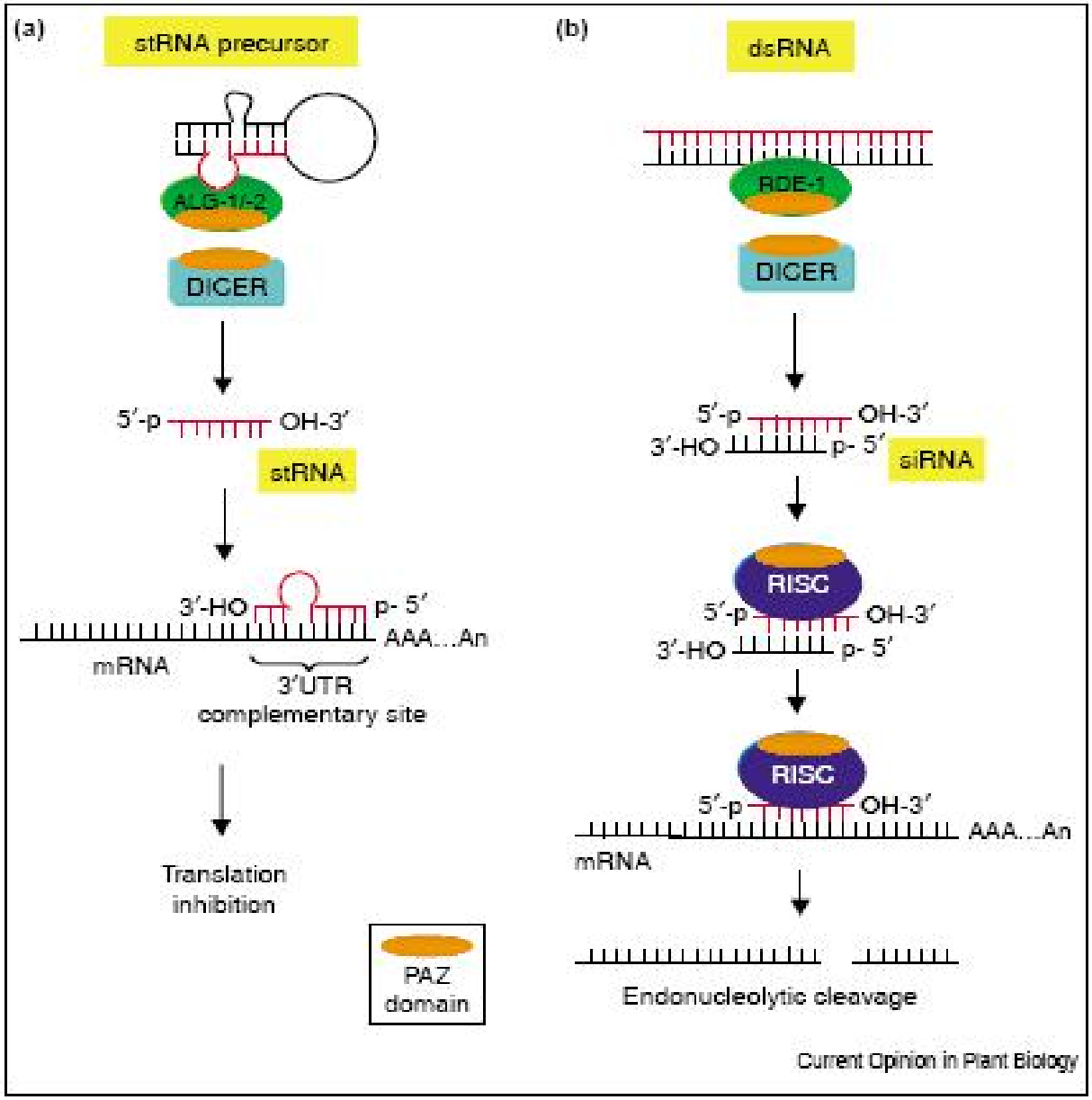
基因敲除（**knock-out**）（CRISPR/Cas9 ）

基因的超表达

反义RNA技术

RNAi

转座子插入突变



反义RNA（antisense RNA）技术

CRISPR (clustered, regularly interspaced, short palindromicrepeats)

是一种来自细菌降解入侵的病毒 DNA 或其他外源 DNA 的免疫机制。在细菌及古细菌中，CRISPR系统共分成3类，其中I类和III类需要多种CRISPR相关蛋白（Cas蛋白）共同发挥作用，而II类系统只需要一种Cas蛋白即可，这为其能够广泛应用提供了便利条。

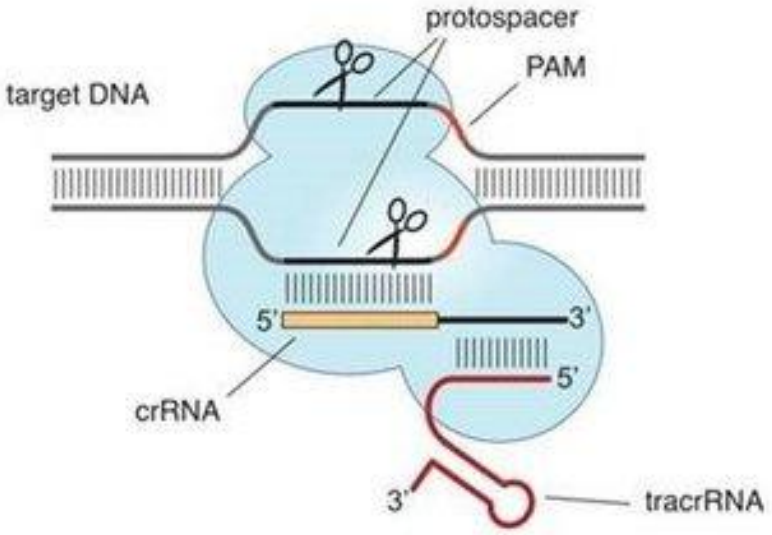
目前，来自*Streptococcus pyogenes* 的CRISPR-Cas9系统应用最为广泛。Cas9 蛋白（含有两个核酸酶结构域，可以分别切割DNA 两条单链。Cas9首先与crRNA及tracrRNA结合成复合物，然后通过PAM序列结合并侵入DNA，形成RNA-DNA复合结构，进而对目的DNA双链进行切割，使DNA双链断裂。

由于PAM序列结构简单（5’-NGG-3’），几乎可以在所有的基因中找 到大量靶点，因此得到广泛的应用。CRISPR-Cas9系统已经成功应用于植物、细菌、酵母、鱼类及哺乳动物细胞，是目前最高效的基因组编辑系统[1]。

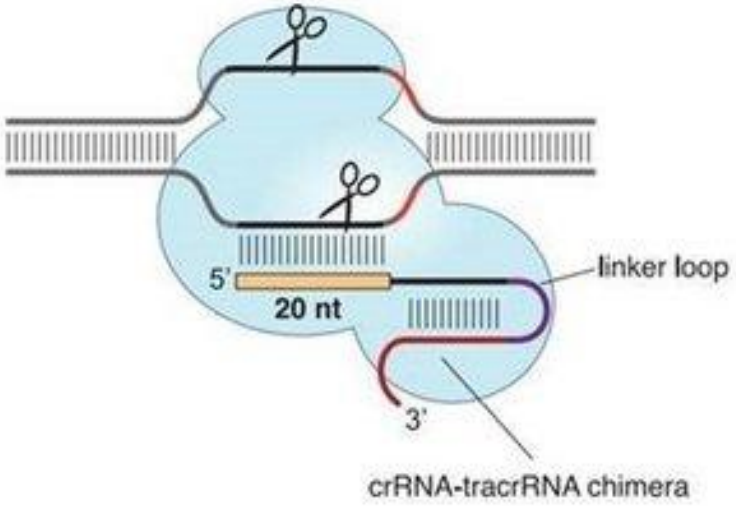
通过基因工程手段对crRNA和tracrRNA进行改造，将其连接在一起得到sgRNA（single guide RNA）。融合的RNA具有与野生型RNA类似的活力，但因为结构得到了简化更方便研究者使用。通过将表达sgRNA的原件与表达Cas9的原件相连接，得到可以同时表达两者的质粒，将其转染细胞，便能够对目的基因进行操作[2,3]。

目前常用的CAS9研究方法是通过普通质粒,质粒构建流程如下：

Cas9 programmed by crRNA:tracrRNA duplex



Cas9 programmed by single chimeric RNA



研究内容：

结构基因组学

1. 构建基因组的遗传图谱；
2. 构建基因组的物理图谱；
3. 测定基因组DNA的全部序列；

功能基因组学

4. 基因组序列诠释

蛋白组学

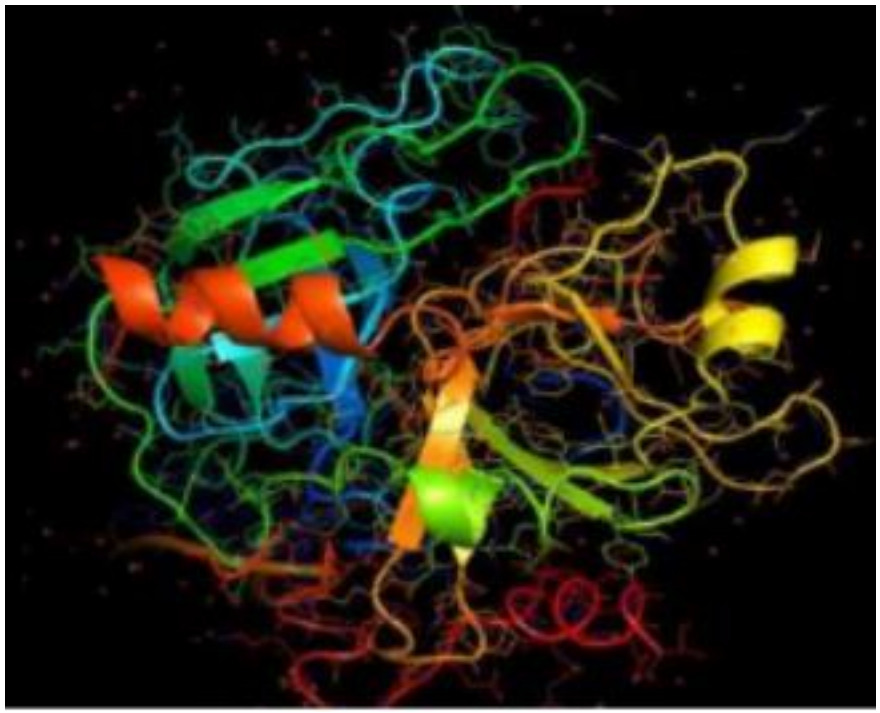
5.蛋白组学

第三节　生物信息学

1、生物信息学(Bioinformatics)

采用计算机技术和信息论方法对蛋白质及其核酸序列等多种生物信息采集、加工、储存、传递、检索、分析和解读，旨在掌握复杂生命现象的形成模式和演化规律的科学

生物信息学的研究材料和结果就是各种各样的生物学数据，其研究工具是计算机，研究方法包括对生物学数据的搜索（收集和筛选）、处理（编辑、整理、管理和显示）及利用（计算、模拟）。主要的研究方向有：序列比对、基因识别、基因重组、蛋白质结构预测、基因表达、蛋白质反应的预测，以及建立进化模型。



生物信息学（**Bioinformatics**）是当今生命科学和自然科学的重大前沿领域之一，同时也将是21世纪自然科学的核心领域之一。其研究重点主要体现在基因组学（Genomics）和蛋白质组学（Proteomics）两方面，具体说就是从核酸和蛋白质序列出发，分析序列中表达的结构功能的生物信息。

2、基因芯片(gene chip),又称DNA微阵列(microarray)

是由大量DNA或寡核苷酸探针密集排列所形成的探针阵列，其基本原理是通过杂交检测信息。

利用基因芯片，可以实现基因信息的大规模检测

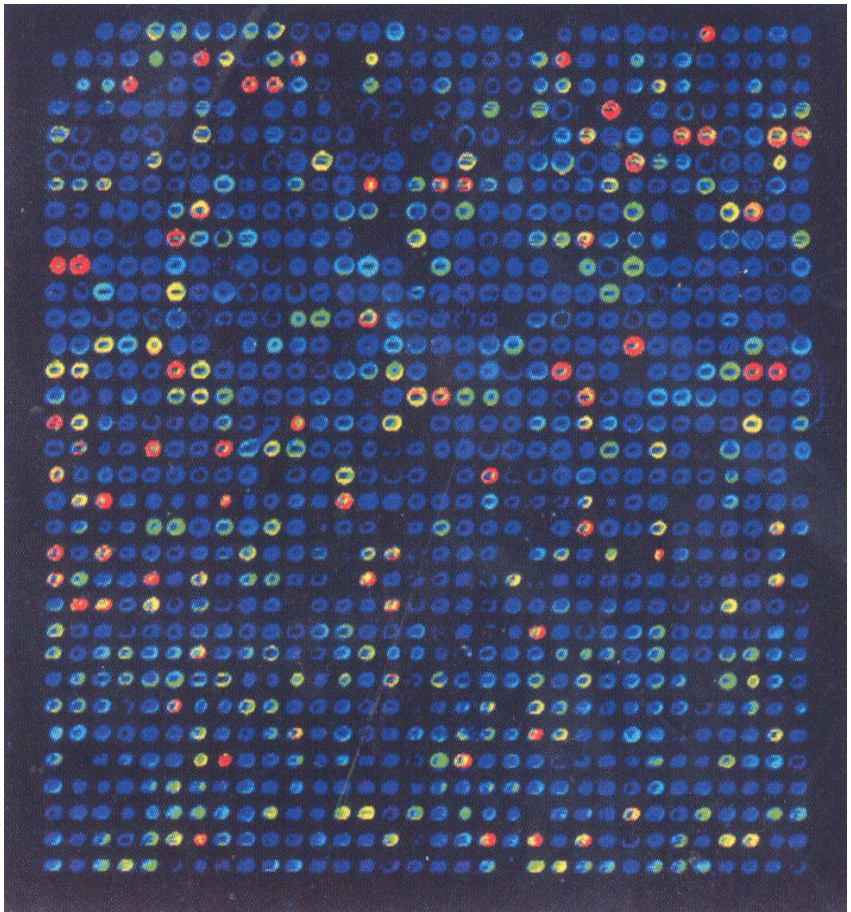
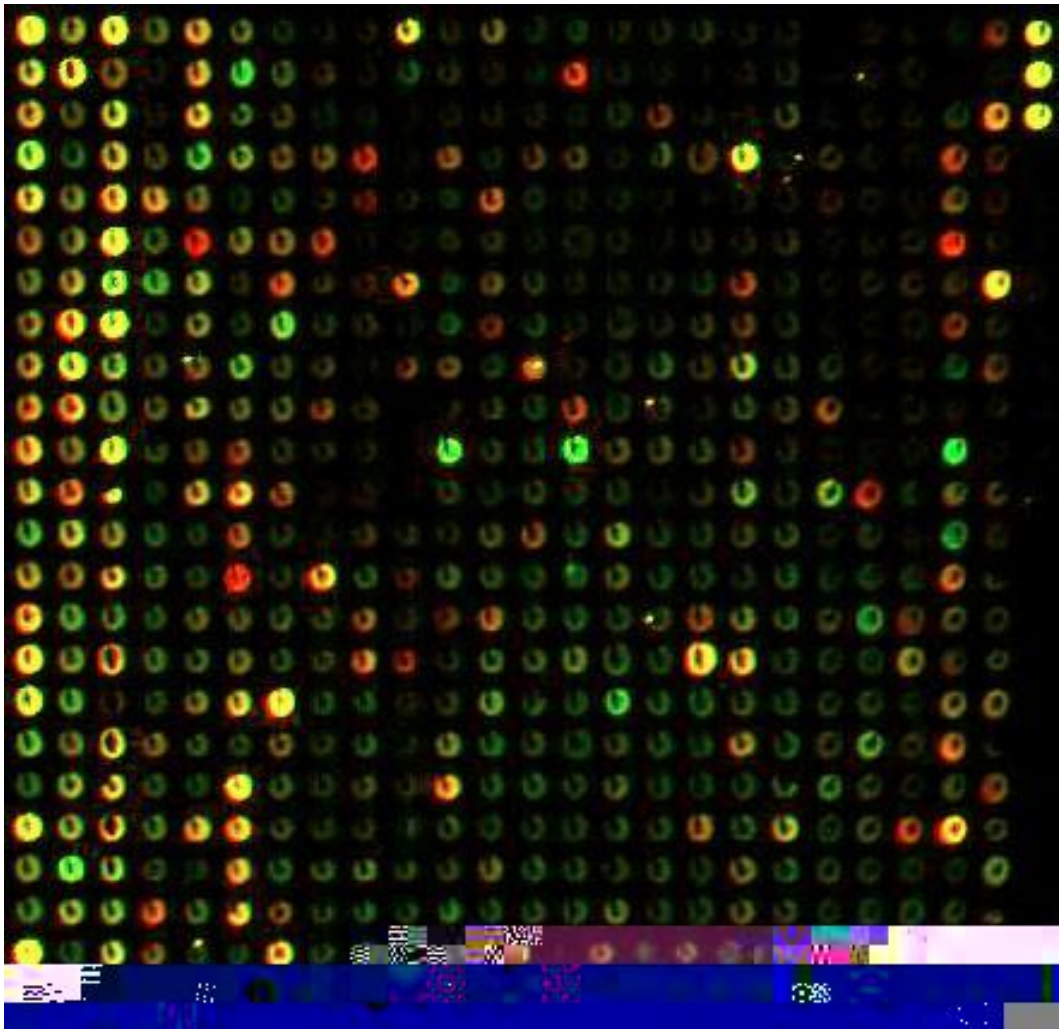


Image from Gene-Chips (Microarray)



利用大量的生物信息资料来了解遗传网络系统、信号传递及相互关系，计算机还可进行一些生物模拟研究。

3、生物信息学的应用

(1) 发现新基因和新的单核苷酸多态性

(2) 分析基因组中非编码蛋白质区域功能

(3) 在基因组水平上研究生物进化

(4) 完整基因组比较研究

第四节　蛋白质组学

1、概念及研究内容

蛋白质组：细胞、器官或组织的蛋白质成分的总称

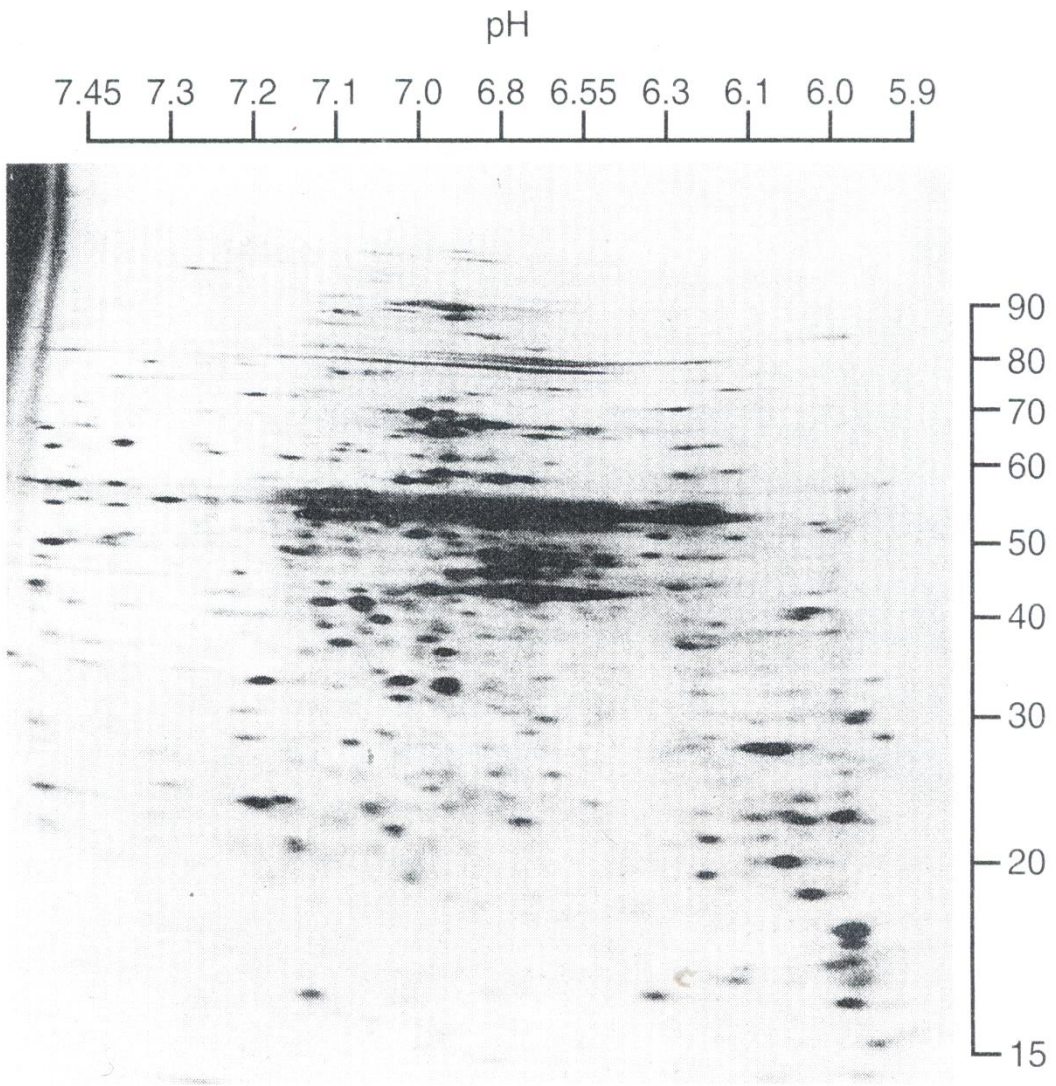
蛋白质组学：研究这些成分在指定的时间或特定的环境条件下的表达

研究内容：蛋白质表达模式的研究，蛋白质组功能模式的研究

2、蛋白质的分离

蛋白质组学研究的第一步就是蛋白质的分离

双相凝胶电泳是蛋白质组研究中的首选分离技术



3、蛋白质的鉴定

鉴定蛋白质组份的性质、结构和功能及其各蛋白质间的相互作用关系，从而最终实现蛋白质组表达模式和功能模式的研究

蛋白质表达模式的鉴定技术：

以质谱为核心的技术

蛋白质微测序

氨基酸组成分析

蛋白质芯片分析

4、蛋白质间的相互作用

研究方法：

酵母双杂交系统

表面等离子共振技术

酵母双杂交系统：酵母双杂交系统是将待研究的两种蛋白质分别克隆（融合）到酵母表达质粒的转录激活因子（如GAL4等）的DNA结合结构域（DNA-BD）和转录激活域（AD）上，构建成融合表达载体，从表达产物分析两种蛋白质相互作用的系统。

酵母双杂交系统可进行两个蛋白互作分析，可用一个已知的蛋白因子（在双杂交系统中称为诱饵蛋白去钓取与其结合的蛋白；也可用进一步验证两个蛋白之间的互作。应用Clontech第三代酵母双杂交系统，

酵母双杂交系统（Yeast two-hybrid system）的建立是基于对真核生物调控转录起始过程的认识。细胞起始基因转录需要有反式转录激活因子的参与。反式转录激活因子，例如酵母转录因子 GAL4在结构上是组件式的（modular），往往由两个或两个以上结构上可以分开，功能上相互独立的结构域（domain）构成，其中有DNA结合功能域（DNA binding domain,DNA-BD）和转录激活结构域（activation domain，DNA-AD）。这两个结合域将它们分开时仍分别具有功能，但不能激活转录，只有当被分开的两者通过适当的途径在空间上较为接近时，才能重新呈现完整的GAL4转录因子活性，并可激活上游激活序列（upstream activating sequence,UAS）的下游启动子，使启动子下游基因得到转录。

双杂交系统的建立得力于对真核生物调控转录起始过程的认识。细胞起始基因转录需要有反式转录激活因子的参与。80年代的工作表明，转录激活因子在结构上是组件式的（modular）， 即这些因子往往由两个或两个以上相互独立的结构域构成，其中有DNA结合结构域（DNA binding domain，简称为BD）和转录激活结构域（activation domain，简称为AD），它们是转录激活因子发挥功能所必需的。单独的BD虽然能和启动子结合，但是不能激活转录。而不同转录激活因子的BD和AD形成的杂合蛋白仍然具有正常的激活转录的功能。如酵母细胞的Gal4蛋白的BD与大肠杆菌的一个酸性激活结构域B42融合得到的杂合蛋白仍然可结合到Gal4结合位点并激活转录。

在酵母双杂交系统中，“诱饵”蛋白X克隆至DNA-BD载体中，表达DNA-BD/X融合蛋白；待测试蛋白Y克隆至AD载体中，表达AD/Y融合蛋白。一旦X与Y蛋白间有相互作用，则DNA-BD和AD也随之被牵拉靠近，恢复行使功能，激活报告重组体中LacZ和HIS3基因的表达。

从1996年酵母菌基因组全序列测定后的4年多时间里：

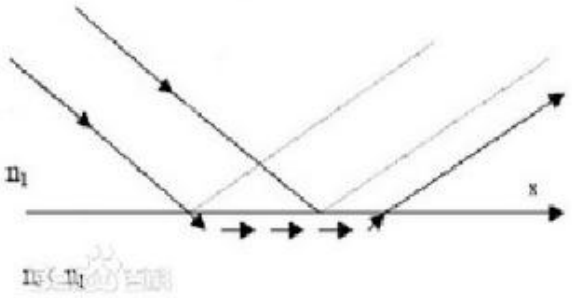
全世界1000多个实验室，5000多名科学家从事酵母菌后基因组学的研究

发表论文7000多篇，鉴定1060个新基因的功能，但仍然还有约1600个阅读框架的功能不清楚

这些结果充分说明后基因组学研究的复杂性

表面等离子共振 (SPR)原理

- 表面等离子共振(Surface Plasmon Resonance, SPR)
- 消逝波：当光从光密介质射入光疏介质，入射角增大到某一角度，使折射角达到90° 时，折射光将完全消失，而只剩下反射光，这种现象叫做全反射。
- 当以波动光学的角度 来研究全反射时，人们发现当入射光到达界面时并不是直接产生反射光，而是先透过光疏介质约一个波长的深度，再沿界面流动约半个波长再返回光密介质。则透过光疏介质的波被称为消逝波。



表面等离子共振 (SPR)原理

- 等离子波：把金属表面的价电子看成是均匀正电荷背景下运动的电子气体，其中正、负带电粒子数目几乎相等，这实际上也是一种等离子体。当金属受电磁干扰时，金属内部的电子密度分布会变得不均匀。因为库仑力的存在，会将部分电子吸引到正电荷过剩的区域，被吸引的电子由于获得动量，故不会在引力与斥力的平衡位置停下而向前运动一段距离，之后电子间存在的斥力会迫使已经聚集起来的电子再次离开该区域。由此会形成一种整个电子系统的集体震荡，而库仑力的存在使得这种集体震荡反复进行，进而形成的震荡称等离子震荡，并以波的形式表现，称为等离子波。

表面等离子共振?

由于金属表面存在大量自由电子，自由电子在入射光场的作用下发生集体振荡，这种集体激发的导体电子的振荡模式被称为表面等离子体 (SP)。在特定条件下，入射光与金属薄膜的振荡电子发生共振，对入射光的吸收显着增强，这种现象被称为「表面等离子共振 (SPR) 」。
表面等离子共振（SPR）是一种物理现象，当入射光以临界角入射到两种不同折射率的介质界面（比如玻璃表面的金或银镀层）时，可引起金属自由电子的共振， 由于电子吸收了光能量，从而使反射光在一定角度内大大减弱。

研究内容：

生物信息学

结构基因组学

1. 构建基因组的遗传图谱；
2. 构建基因组的物理图谱；
3. 测定基因组DNA的全部序列；

功能基因组学

4. 基因组序列诠释

蛋白组学

5.蛋白组学

思考题

1. 常用分子标记有哪几种？
2. 基因组学研究内容与孟德尔遗传定律相关试验研究内容的区别？
3. 植物构建作图群体有哪些主要类型？其优缺点如何？
4. 简述植物基因组遗传图谱的构建方法。

- 作业 P300

1. 什么是基因组？基因组学研究内容？

2. 什么是C值悖理、N值悖理？

3. 遗传图谱与物理图谱的概念，构建遗传图谱与物理图谱的作用？

4. 根据下列凝胶电泳分析结果，构建一个限制性酶图谱，并表明酶切位点及片段的长度，片段总长度为1500bp，电泳分析结果如下：

内切酶	酶切出生 DNA片段长度（bp）		
A	1200	300	
B	550	950	
A+B	250	300	950