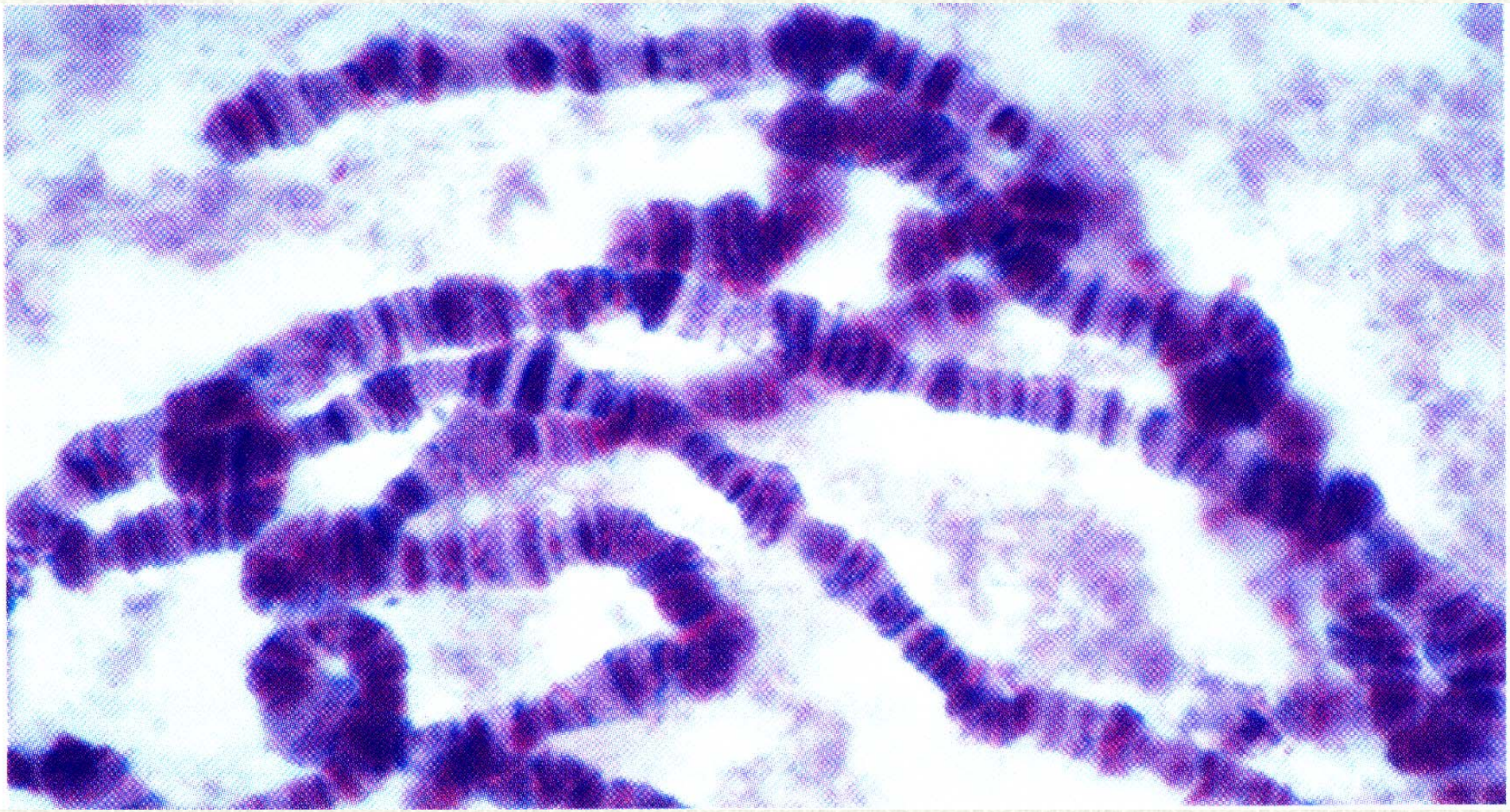
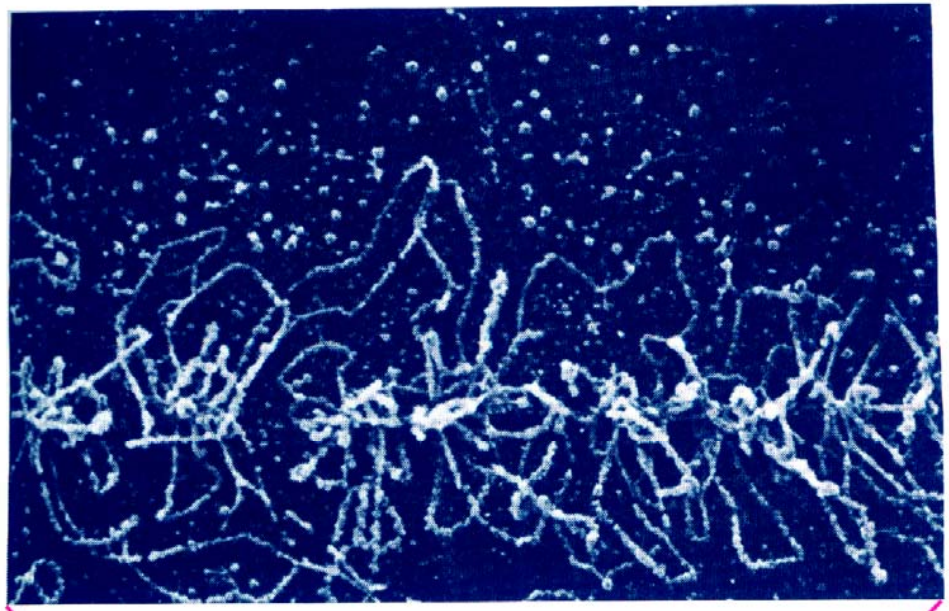
第十讲 基因组与比较基因组学

1940年代第一颗原子弹爆炸，1960年代人类首次登上月球和1990年代提出并已基本完成的人类基因组计划（HGP）是20世纪人类科技发展史上的三大创举。

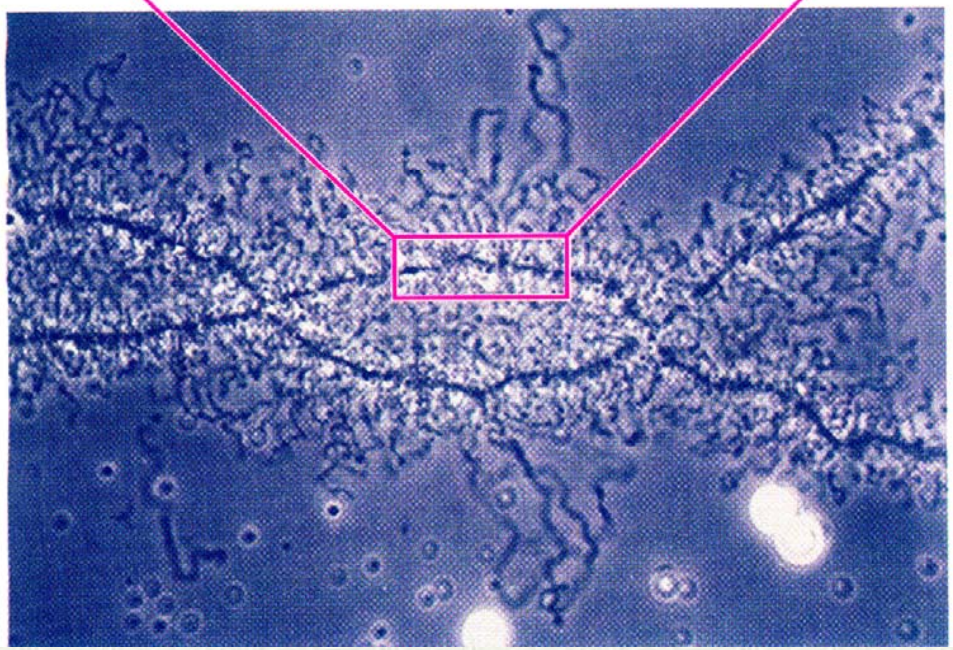


NLPE&PGE

（b）



（a）



NLPE&PGE

10. 1 人类基因组计划

基因组学这一名词是美国人**T.H.Roderick**在**1986**年**7**月造出来的，与一个新的杂志——**Genomics**一道问世，它着眼于研究并解析生物体整个基因组的所有遗传信息。

基因组是生物体内遗传信息的集合，是某个特定物种细胞内全部DNA分子的总和。人类基因组包括23对染色体，单倍体细胞中约有30亿对核苷酸，编码了5-6万个基因，人类基因组中携带了有关人类个体生长发育、生老病死的全部遗传信息。

从整体上看，不同人类个体的基因是相同的，因此，我们说“人类只有一个基因组”，人生来是平等的。当然，不同的人可能拥有不同的等位基因，这一点决定了人与人之间个体上的差异。

10. 1. 1 人类基因组计划的科学意义

到目前为止，已经完成了酵母、线虫、果蝇、拟南芥、人类、小鼠和水稻等7个真核生物基因组以及大肠杆菌等上百个原核生物基因组。

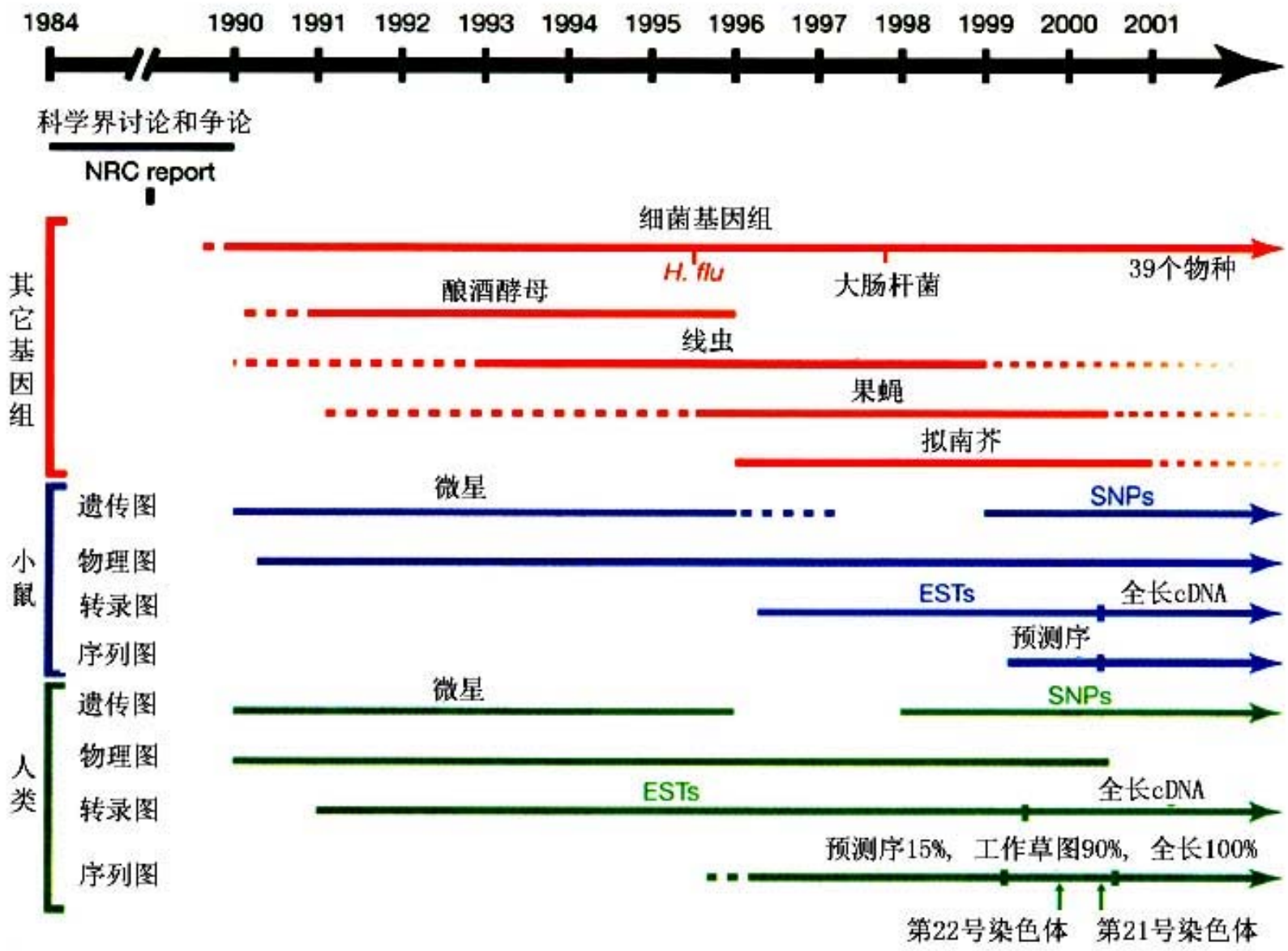
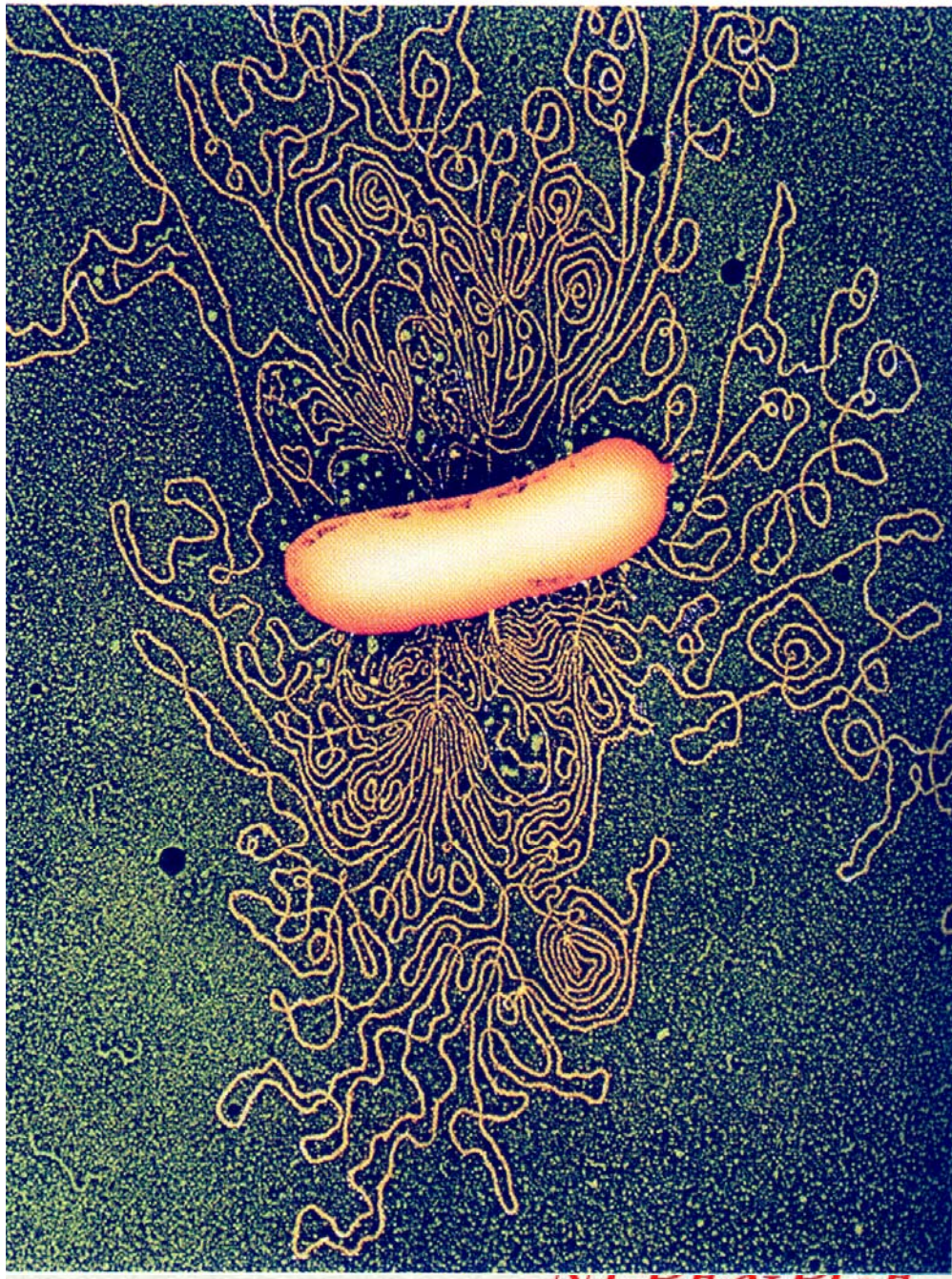


图10-1主要基因组计划到2001年2月为止的进展情况总结

TABLE 17.1 The Genetic Material of Representative Viruses and Bacteria

		Nucleic Acid			
	Organism	*Type*	*SS or DS**	*Length (μm)*	*Overall Size of Viral Head or Bacteria (μm)*
Viruses	ϕX174	DNA	SS	2.0	0.025 × 0.025
	Tobacco mosaic virus	RNA	SS	3.3	0.30 × 0.02
	Lambda phage	DNA	DS	17.0	0.07 × 0.07
	T2 phage	DNA	DS	52.0	0.07 × 0.10
Bacteria	*Haemophilus influenzae*	DNA	DS	832.0	1.00 × 0.30
	Escherichia coli	DNA	DS	1200.0	2.00 × 0.50

*SS = single-stranded; DS = double-stranded.



人类基因组计划的科学意义在于：

（1）确定人类基因组中约3-4万个编码基因的序列及其在基因组中的物理位置，研究基因的产物及其功能。

（2）了解转录和剪接调控元件的结构与位置，从整个基因组结构的宏观水平上理解基因转录与转录后调节。

（3）从整体上了解染色体结构，了解各种不同序列在形成染色体结构、DNA复制、基因转录及表达调控中的影响与作用。

（4）研究空间结构对基因调节的作用。

（5）发现与DNA复制、重组等有关的序列。

（6）研究DNA突变、重排和染色体断裂等，了解疾病的分子机制，为疾病诊断、预防和治疗提供理论依据。

（7）确定人类基因组中转座子、逆转座子和病毒残余序列，研究其周围序列的性质。

（**8**）研究人类个体之间的多态性（**SNP**）情况，用于基因诊断、个体识别、亲子鉴定、组织配型、发育进化等许多医疗、司法和人类学的研究。

人类基因组计划的成果是多方面的，它主要体现在鉴定基因的四张图上。

10. 1. 2 遗传图（Genetic Map）

又称连锁图（**Linkage Map**），是指基因或**DNA**标志在染色体上的相对位置与遗传距离，通常以基因或**DNA**片段在染色体交换过程中的分离频率厘摩（**cM**）来表示。**cM**值越大，两者之间距离越远。

产生配子的减数分裂过程中，亲代同“号”的父源或母源染色体既能相互配对也可能发生片段互换，而父母源染色体等位基因互换导致子代出现**DNA**“重组”的频率与这两个位点之间的距离呈正相关，所以，用两个位点之间的交换或重组频率来表示其“遗传学距离”。

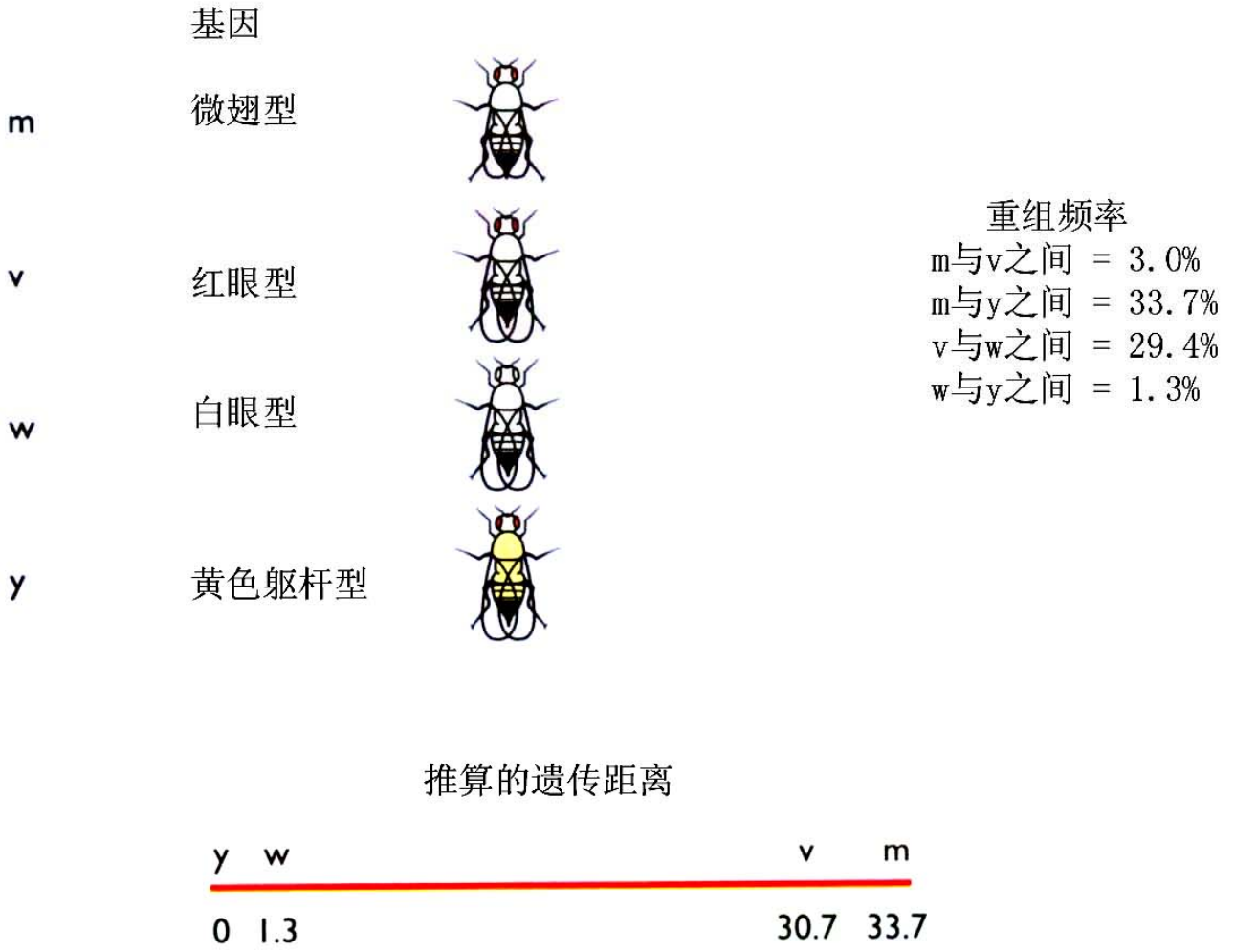


图10-2 遗传距离图的基本数据来自基因的重组。
注：上述4个基因都位于果蝇的X染色体上。

表10-1 酵母遗传分析中最常用的生物化学标签

标　签	表 现 型	筛 选 方 法
ADE2	培养基中需加入腺苷酸	只能在加入腺苷酸的培养基上生长
CAN1	对刀豆氨酸有抗性	能在含有刀豆氨酸的培养基上生长
CUP1	对铜离子有抗性	能在含有铜离子的培养基上生长
CYH1	对环己酰亚胺有抗性	能在含有环己酰亚胺的培养基上生长
LEU2	培养基中需加入亮氨酸	只能在加入亮氨酸的培养基上生长
SUC2	能进行蔗糖发酵	能在以蔗糖作为唯一碳源的培养基上生长
URA3	培养基中需加入尿嘧啶	只能在加入尿嘧啶的培养基上生长

由于不能对人类进行“选择性”婚配，而且人类子代个体数量有限、世代寿命较长，呈共显多态性的蛋白质数量不多，等位基因的数量不多。

DNA技术的建立为人类提供了大量新的遗传标记。

第一代**DNA**遗传标记是**RFLP**（**Restriction Fragment Length Polymorphism**，限制性片段长度多态性）。**DNA**序列上的微小变化，甚至**1**个核苷酸的变化，也能引起限制性内切酶切点的丢失或产生，导致酶切片段长度的变化。

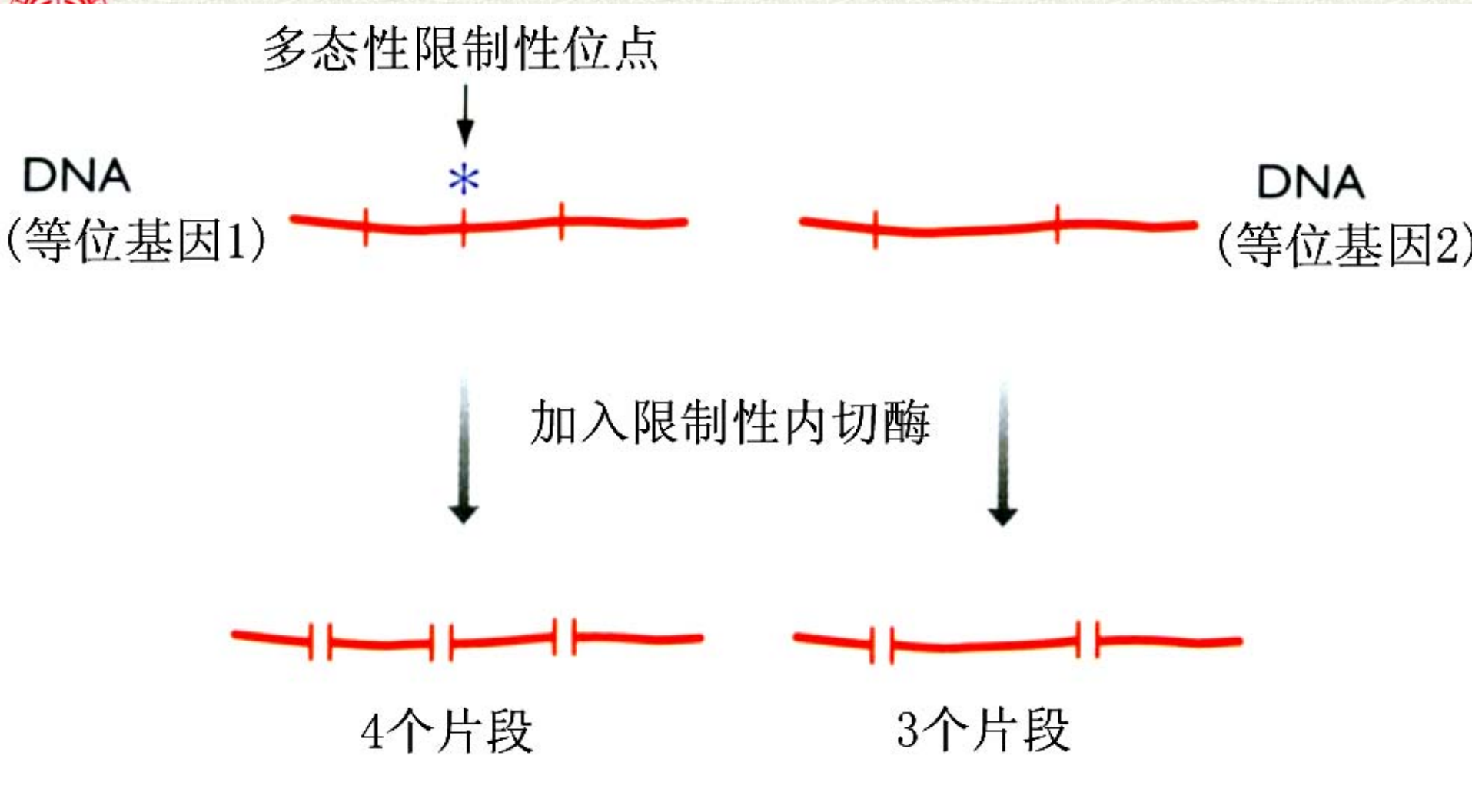


图**10-3** 限制性片段长度多态性（**RFLP**）原理示意图。

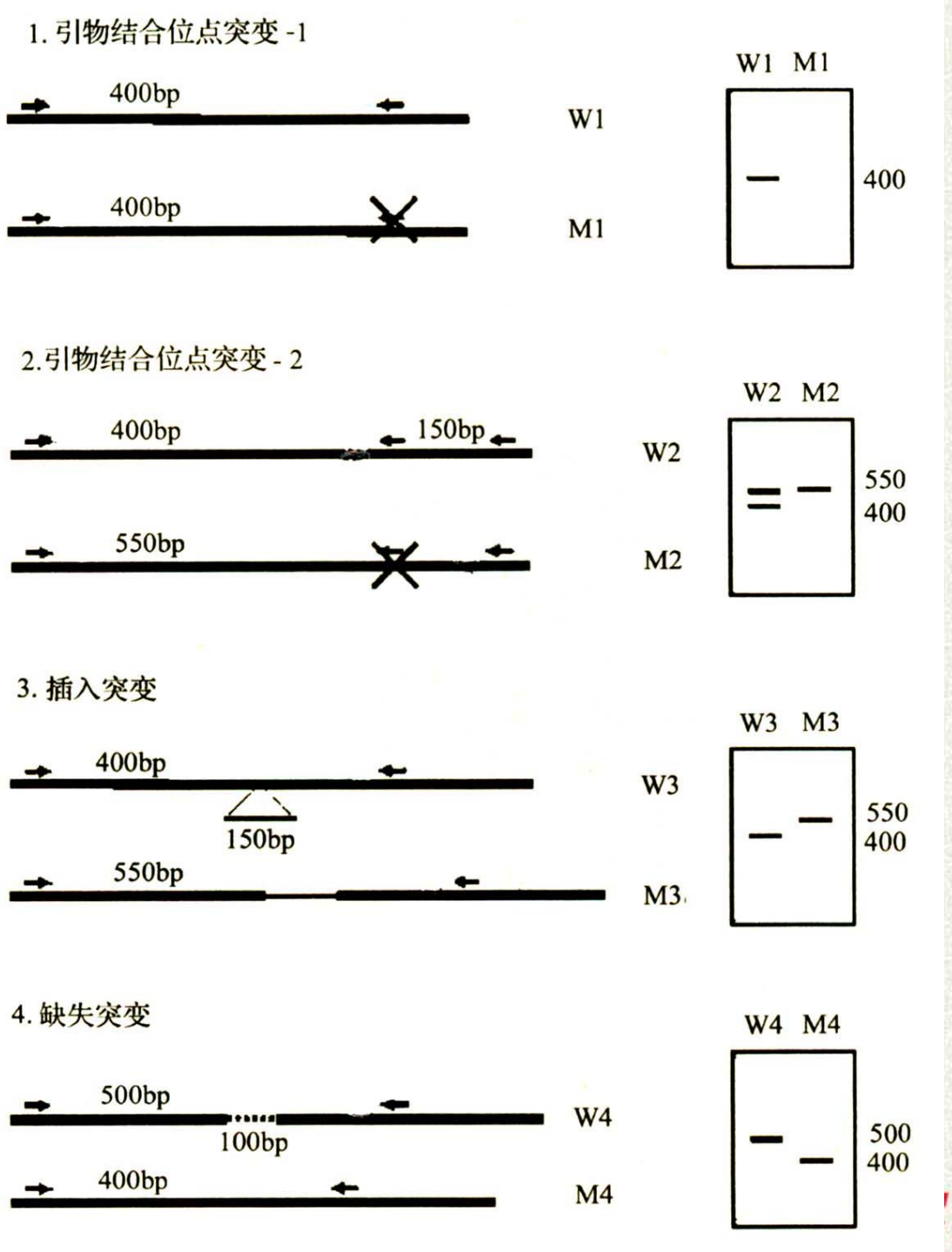
由于核苷酸序列的改变遍及整个基因组，特别是进化中选择压力不是很大的非编码序列之中，**RFLP**的出现频率远远超过了经典的蛋白质多态性。而且，只要选择得当，生物体内出现共显性**RFLP**及**RAPD**分子标记的频率较高。

图**10-4 RFLP**分子标记中的显性与共显多态性分子机制。图中第一种类型是最常见的显性多态性标记，第**2**，**3**，**4**类都是共显性标记。

→：引物

W：野生型（**wildtype**）

M：突变型（**mutant**）



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：

重复单位长度在**15-65**个核苷酸左右的小卫星**DNA**（**minisatellite DNA**）；

重复单位长度在**2-6**个核苷酸之间的微卫星**DNA**（**microsatellite DNA**），后者又称为简短串联重复（**STR**、**STRP**或**SSLP**，**short tandem repeat polymorphism**或者**simple sequence length polymorphism**)。

STRP的优点是“多态性”与“高频率”。由于**(A)n**，**(CA)n**，**(CGG)n**等短重复序列在进化上不受选择，在同一位点上可重复单位数量变化很大，配对时又容易产生“错配”，使这样的位点遍布于整个基因组。

表10-2 人类基因组中的各种主要卫星DNA比较

卫星DNA分类	特　征
卫星DNA： 串联重复的基本单位首尾相接，在基因组中呈不均匀分布，但主要集中于着丝粒、端粒等特定部位，高度或中等重复，分属三个大家族。	
α卫星DNA	中等重复，基本单位长171bp。
小卫星DNA	中等重复，基本单位长15~65bp。
微卫星DNA	中等重复，基本单位长2~8bp

已有5264个STRP为主体的遗传标记“连锁图”，平均分辨率已达600kb，其中第17号染色体上平均每495 kb有一个标记，第9号染色体上平均每767 kb有一个标记，整个基因组中只有三处标记间距大于4Mb。

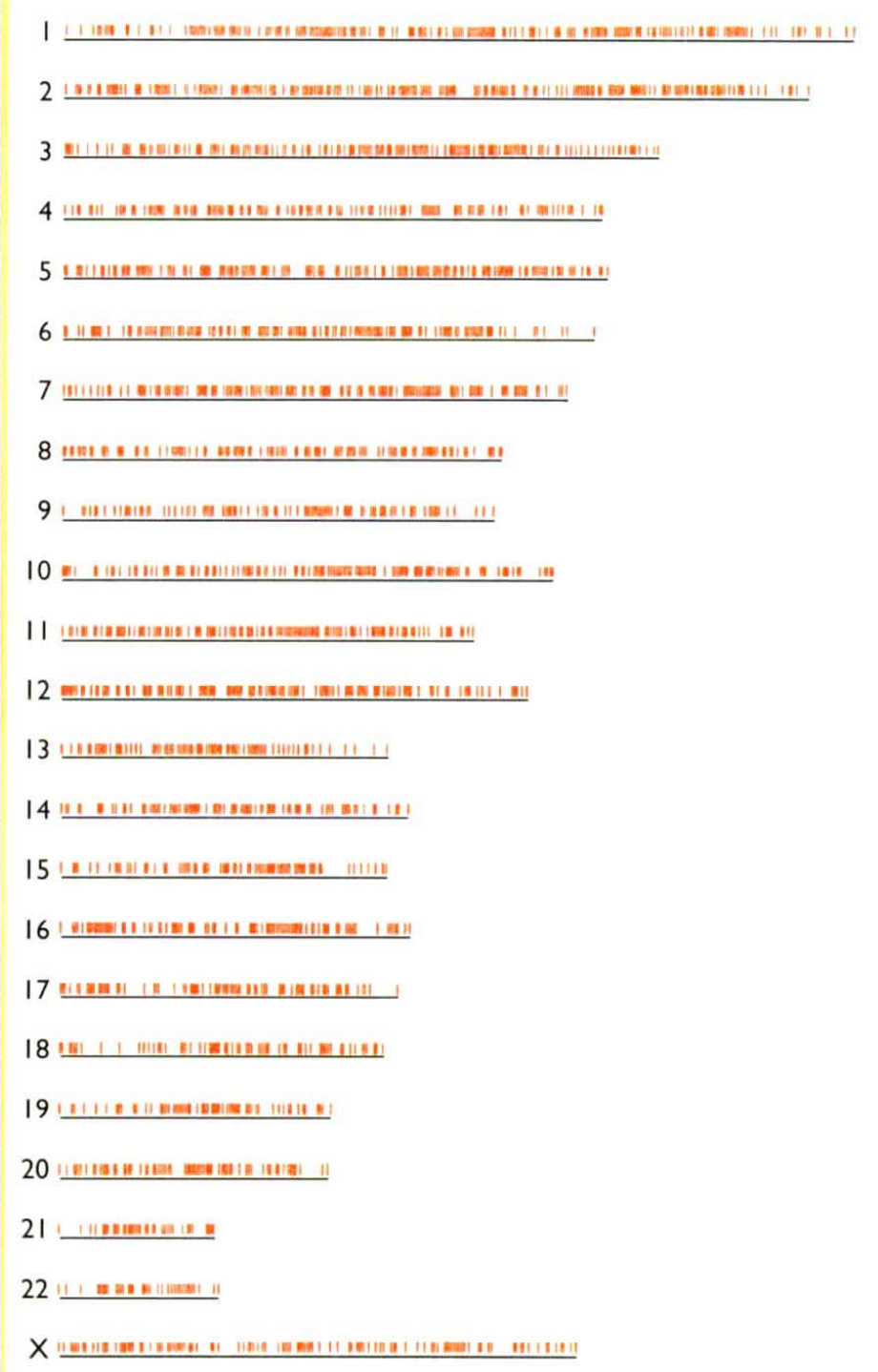


图**10-5** 人类基因组微卫星遗传标记图。

占人类基因组约45%的重复序列来源于转座子复制机制。

序列分析表明，四类转座子产生了这些重复序列，其中前三类转座子以RNA为中间产物进行转座，最后一类则直接以DNA的形式转座。

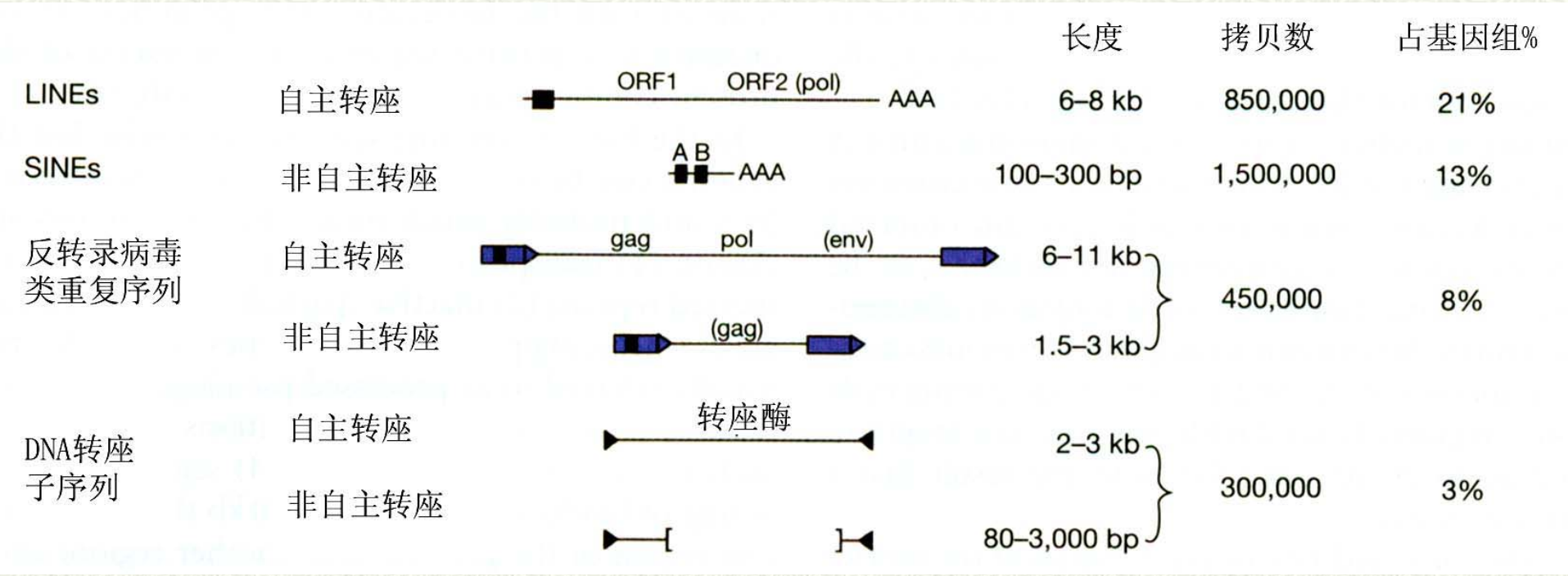


图10-6 存在于人类基因组重复序列中的四类转座子序列分析。

LINEs（long interspersed elements）可能是人类基因组中最古老的重复序列，一般长5-6 kb，含有RNA聚合酶II启动子序列和两个可读框.

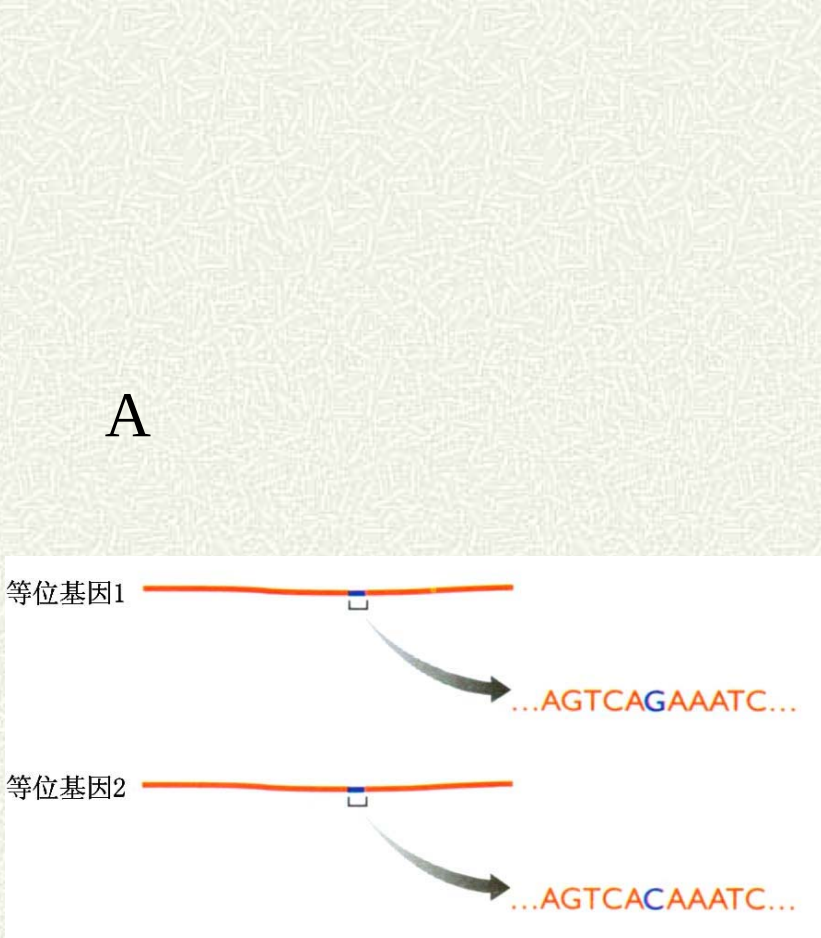
SINEs（short interspersed elements）是非自主转座子，长约100~400 bp，其3’末端与LINEs有同源性，因此能依靠LINEs进行转座。

第三代DNA遗传标记，可能也是最好的遗传标记，是分散于基因组中的单个碱基的差异，即单核苷酸的多态性（SNP），包括单个碱基的缺失、插入和替换。

SNP中大多数为转换，即由一种嘧啶碱基替换另一种嘧啶碱基，或由一种嘌呤碱基替换另一种嘌呤碱基，颠换与转换之比为1：2。

SNP有可能在密度上达到人类基因组“多态”位点数目的极限。估计人类基因组中可能有**300**万个**SNP**位点！

SNP与**RFLP**和**STRP**标记的主要不同之处在于，它不再以**DNA**片段的长度变化作为检测手段，而直接以序列变异作为标记。



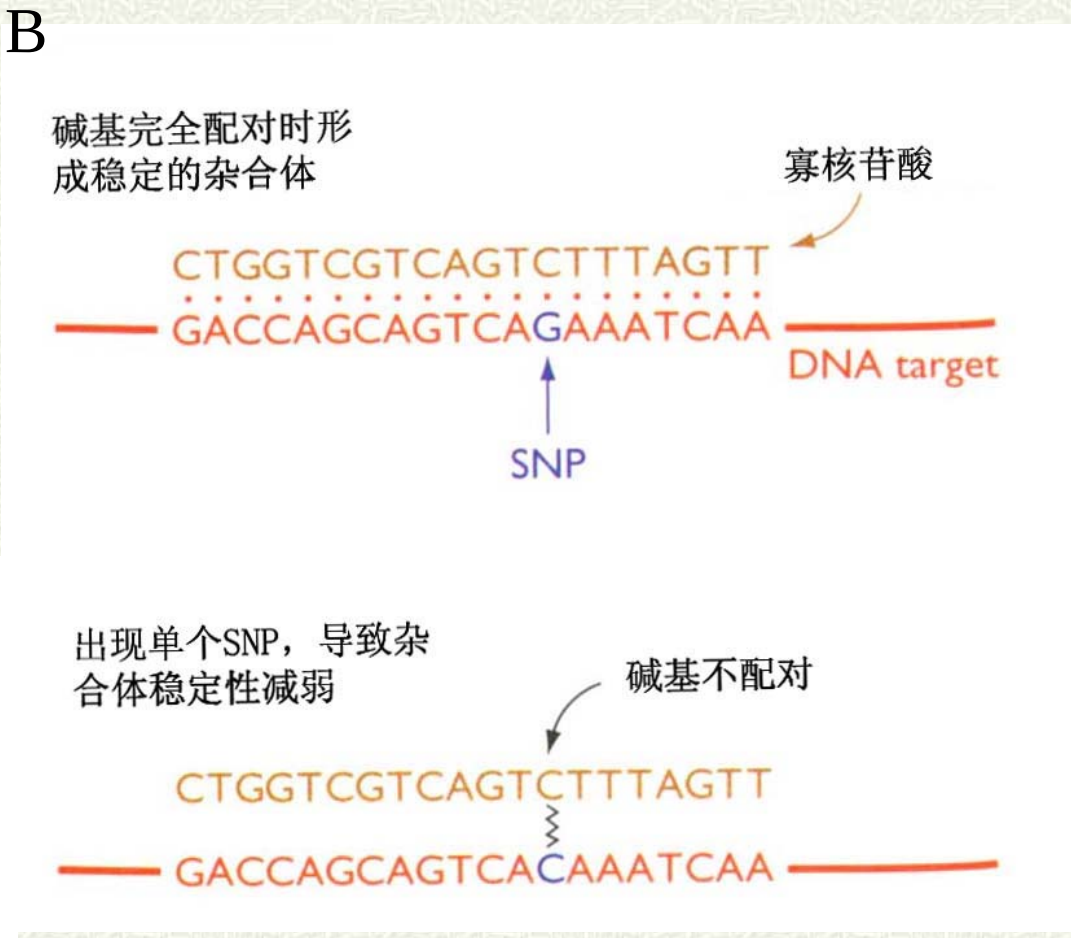


图**10-7** 人类基因组中的**SNP**作为遗传标记的分子机制。**A** 单核苷酸多态性的产生；**B SNP**的产生影响了**DNA**序列间杂交的强度。

NLPE&PGE

“遗传图”的建立为人类疾病相关基因的分离克隆奠定了基础。拥有5000多个遗传学位点，相当于把整个人类基因组划分为5000多个小区，并分别设置了“标牌”。

如果在家系中证实该基因与某个标记不连锁（重组率为50%），表明该基因不在这一标记附近。

如果发现该基因与某个标记有一定程度的“连锁”（重组率小于50%但大于0），表明它可能位于这个标记附近。

如果该基因与某标记间不发生重组（重组率等于0），我们就推测该标记与所研究的疾病基因可能非常接近。

10. 1. 3 物理图（Physical Map）

人类基因组的物理图是指以已知核苷酸序列的DNA片段（序列标签位点，sequence-tagged site，STS）为“路标”，以碱基对（bp，kb，mb）作为基本测量单位（图距）的基因组图。

物理图的主要内容是建立相互重叠连接的“相连DNA片段群”（contigs）。

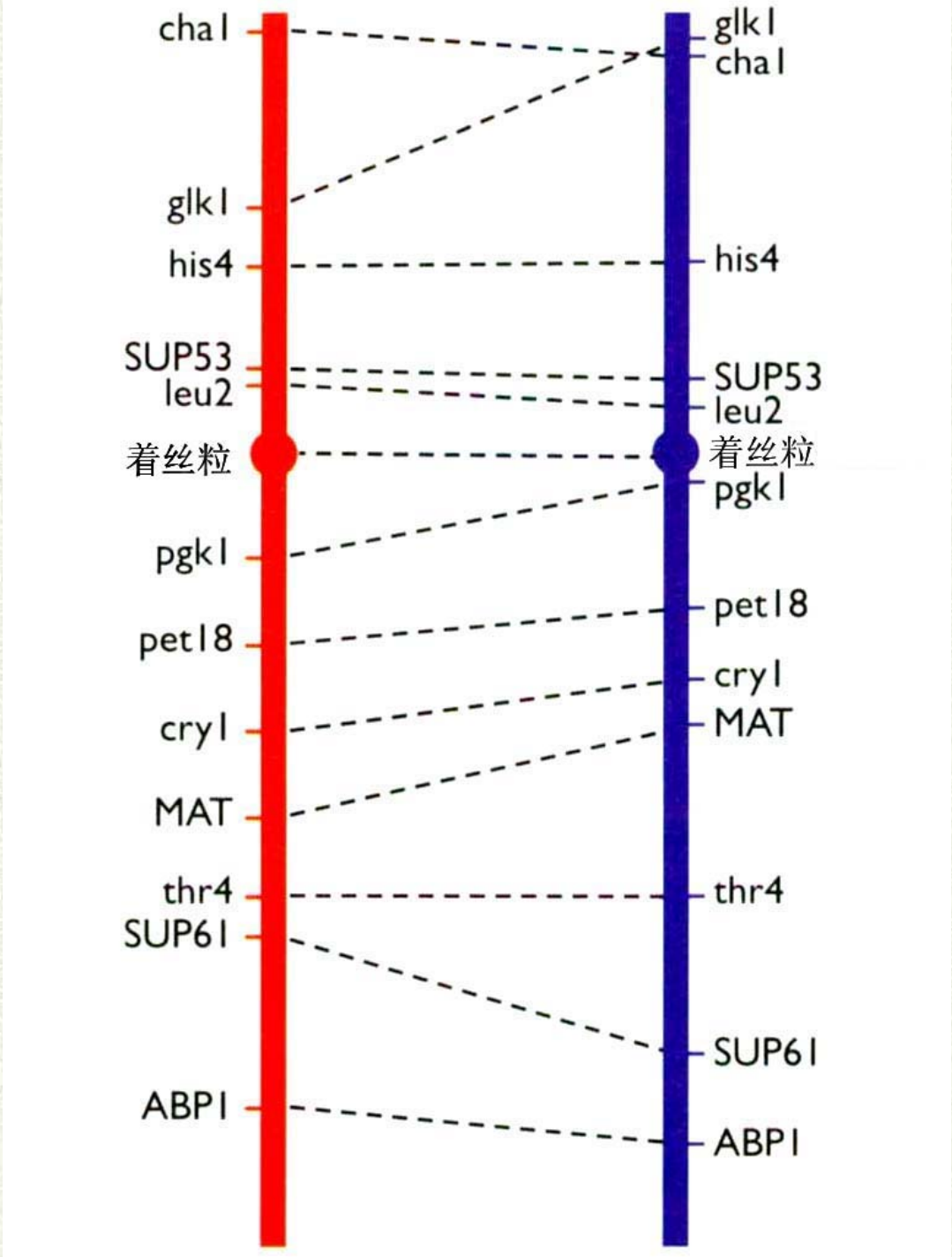


图10-8 酵母第三号染色体遗传图（右）和物理图（左）的比较。

10. 1. 4 转录图（Expression Profiling）

人类的基因转录图（cDNA图），或者基因的cDNA片段图，即表达序列标签图（EST，expressed sequence tag）是人类基因组图的雏型。

在成年个体的每一特定组织中，一般只有10%~20%的结构基因（约1~2万个不同类型的mRNA）表达。

表10-3 不同生物基因组的蛋白质编码能力比较

物　种	基因组总长度	蛋白质种数（个）
尿殖道支原体 ***Mycoplasma genitalium***	**580,073bp**	**467**
肺炎支原体 ***Mycoplasma pneumoniae***	**816,394bp**	**677**
流感嗜血杆菌 ***Haemophilus influenzae***	**1,830,138bp**	**1709**
枯草芽孢杆菌 ***Bacillus subtilis***	**4,214,814bp**	**4100**
大肠杆菌 ***Escherichia coli***	**4,639,221bp**	**4288**

酿酒酵母 ***Saccharomyces cerevisiae***	**13,116,818bp**	**6275**
线 虫 ***Caenorhabditis elegans***	**约97Mp**	**18891**
拟南芥 ***Arabidopsis thaliana***	**115Mp**	**25,498**
果 蝇 ***Drosophila melanogaster***	**116Mp**	**14113**
人 类 ***Homo sapiens***	**3.2×10^9bp**	**约5万左右(?)**

10. 1. 5 人类基因组的序列图（Human Genome Sequence）

人类基因组的核苷酸序列图是分子水平上最高层次、最详尽的物理图。测定总长约1米、由30亿个核苷酸组成的全序列是人类基因组计划的最终目标（图10-10）。

tttccggtatttgggctttaaatccttaattatattatcttg taaaaaaaag ctactcttat aagtaacgtt ttgacccaaa ataaagtaaa gtttcgacat tttgcatata cattaagaaa ctaaataaat atactatgac ccccttcgaa aacatgtcat tcaaaataaa gtacttgtga aaagataaaa ctaaataata taaataatta cctttaaaca gaacaaaatc ttctaaaaca acatttatat tgaaattaag agtaatacat tttagcaata acaaaaaaat tcatgtacaa gatccatgta catataaatgc ctactgatat gtcactttccc caaacgtcacAtta atatctcttc ttctttttt aacatcttaa tcttatttat gattcacaga gaaagaaaaa gagtcaaaat caaaataaca gcttttctcc acataaatcc acatgtgtgt atactggtta ctcgactcta tatatagtcc taaagctaca atgtttctcc atcaaaagta tcaaaagaaa gagaaacaac aaaagcaaat cctataatta taatcacaaaacga

ATGGCGGCCGTTACTTCCTCATGCTCCACCGCGATCTCCGCTTCT
TCCAAAACCCTAGCGAAGCCAGTCGCCGCAAGCTTCGCCCCTAC
TAATCTCTCATTTTCAAAGCTTTCTCCTCAGTCAATCAGGGCTCG
TAGATCCATCACCGTCGGCAGCGCACTAGGCGCCACCAAGGTGT
CGGCTCCTCCCGCCACACATCCCGTTTCGCTCGATTTTGAGACTT
CTGTCTTCAAGAAGGAGAGAGTTAACCTCGCCGGACACGAAGAG

GTTCGGGTTTCTTCTAATTTTTCACTCTACTCTCAGAAATTGACTATTA
CTTTTTATTTTTAAATGAATGATTTTTTTGGTTGATTTGTTGCAG

不同种族、不同个体的基因差异（基因组的多样性）以及“正常”与“疾病”基因的差异，只是同一位点上等位基因的差异，所以，人类基因组全序列来自一个“代表性人类个体”，不属于任何供体。

研究发现，人类基因组与其它动物基因组在染色体水平上有“共线”（即同源）现象。

人类第21号染色体HSA21位点与小鼠第16号染色体MMU16，MMU17和MMU10连锁图中存在着广泛的同源性。

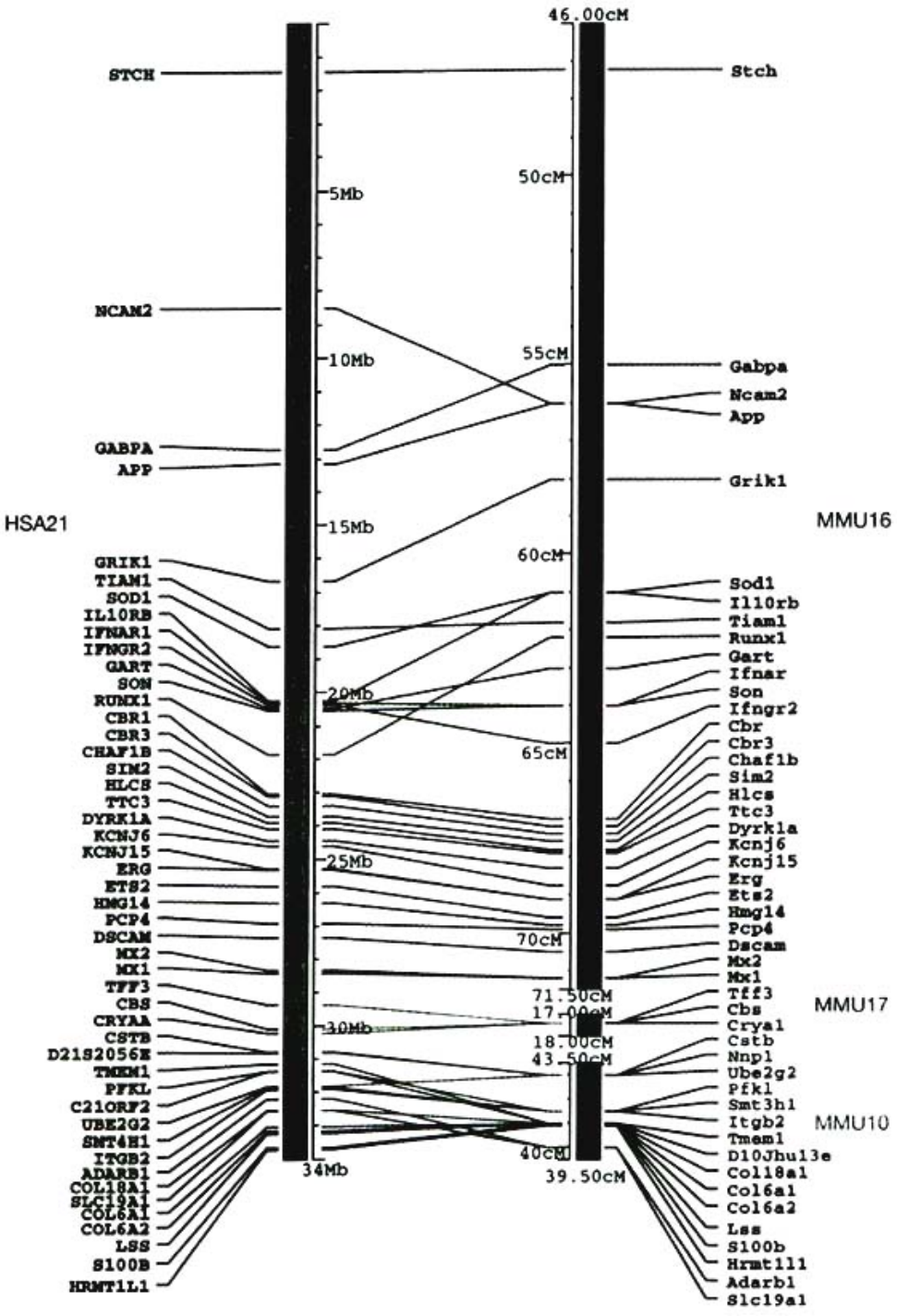


图10-11 人第21号染色体HSA21位点与小鼠16号染色体MMu16，Mmu17和MMu10位点有“同线”性。

10. 2 CLONE-BYCLONE法与鸟枪法序列分析技术的比较

10. 2. 1 基因组DNA大片段文库的构建

酵母人工染色体技术（**yeast artificial chromosome,YAC**）为创制基因组物理图提供了极大的方便。除了**ARS**序列之外，还应加入**CEN**序列，以提高有丝分裂时的稳定性，降低拷贝数。

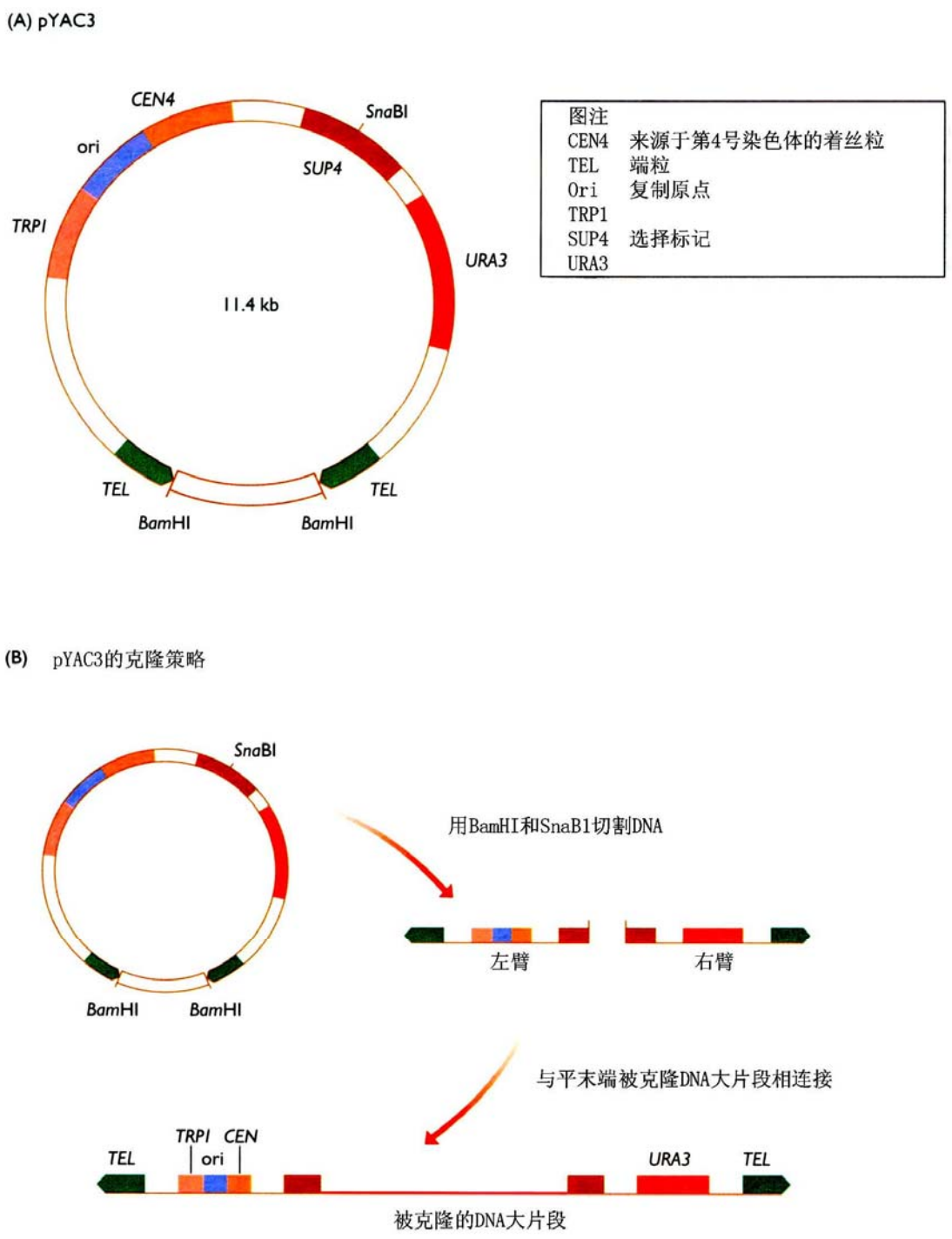


图10-12 酵母人工染色体克隆载体及其克隆策略示意图。

又用细菌的F质粒及其调控基因构建细菌染色体克隆载体，称为BAC（**Bacterial artificial chromosome**），其克隆能力在**125~150 kb**左右。主要包括**oriS**，**repE**（控制**F**质粒复制）和**parA**、**parB**（控制拷贝数）等。

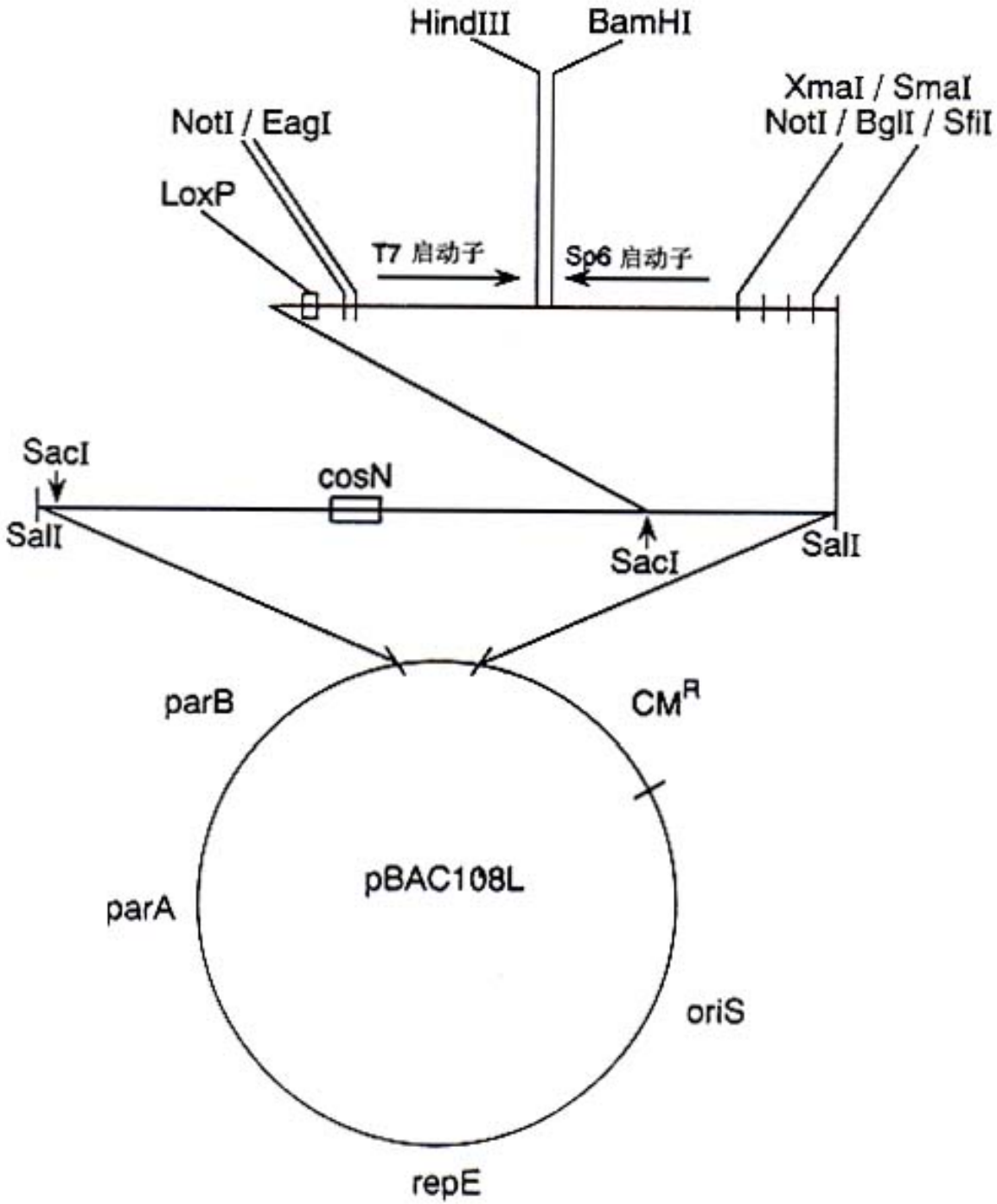
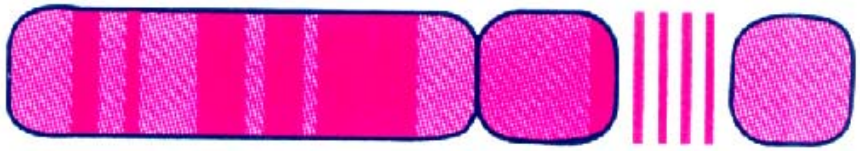


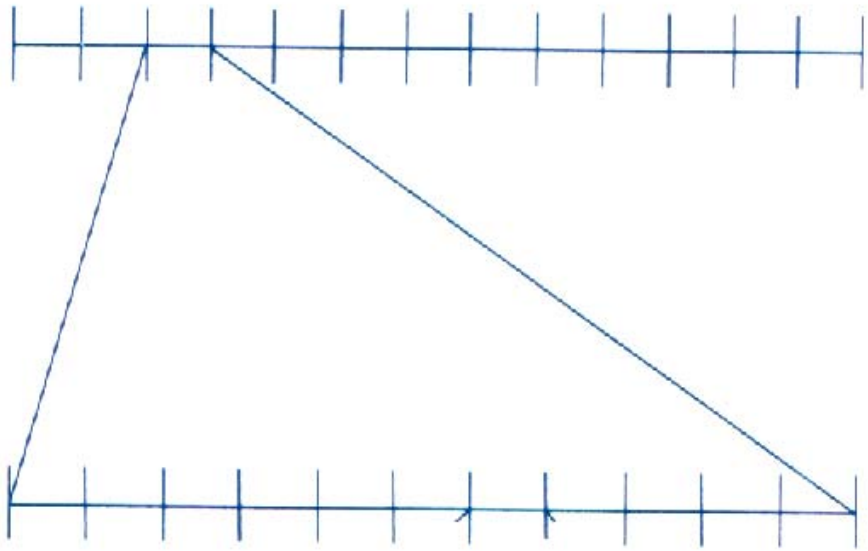
图10-13 细菌人工染色体的构建及其克隆策略。PBAC108L来自细菌的一个小型F质粒，其中oriS，repE控制了质粒的复制起始，parB和parA控制了拷贝数。

10. 2. 2 ＣＬＯＮＥ－ＢＹ－ＣＬＯＮＥ法基因组序列分析技术

(a) CLONE-BY-CLONE METHOD

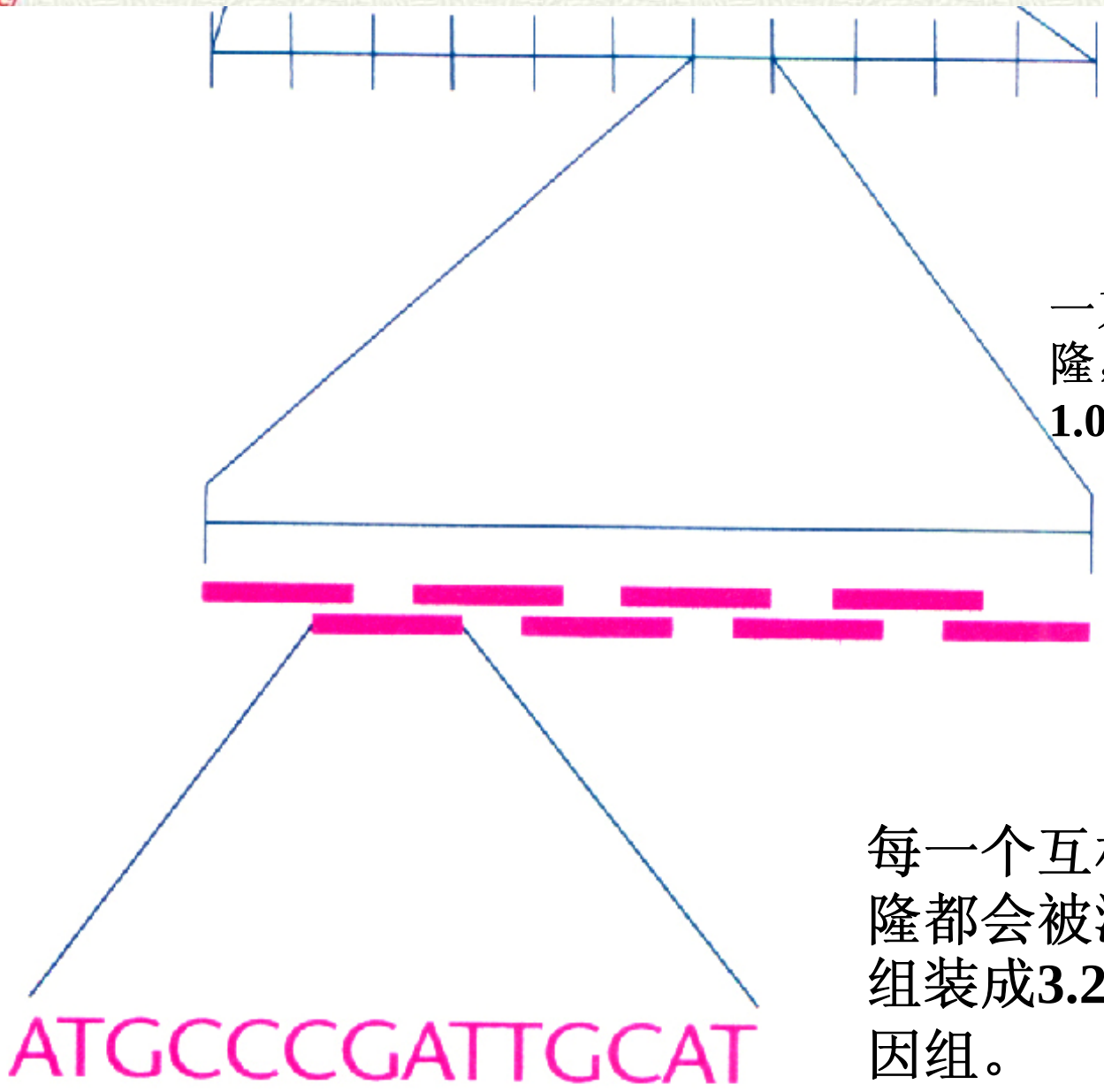


第21号染色体



遗传标记, 如**RFLP**，**STSs**等，在基因组上相距大约**1Mb**有一个标记。主要通过**DNA**重组杂交等技术获取。

用**RFLP**，**STSs**等，画出物理图谱，标出物理距离，物理标签要求密度高于**1/100,000bp**



一系列互相有重叠的克隆，每套都要求长于**0.5-1.0Mb**

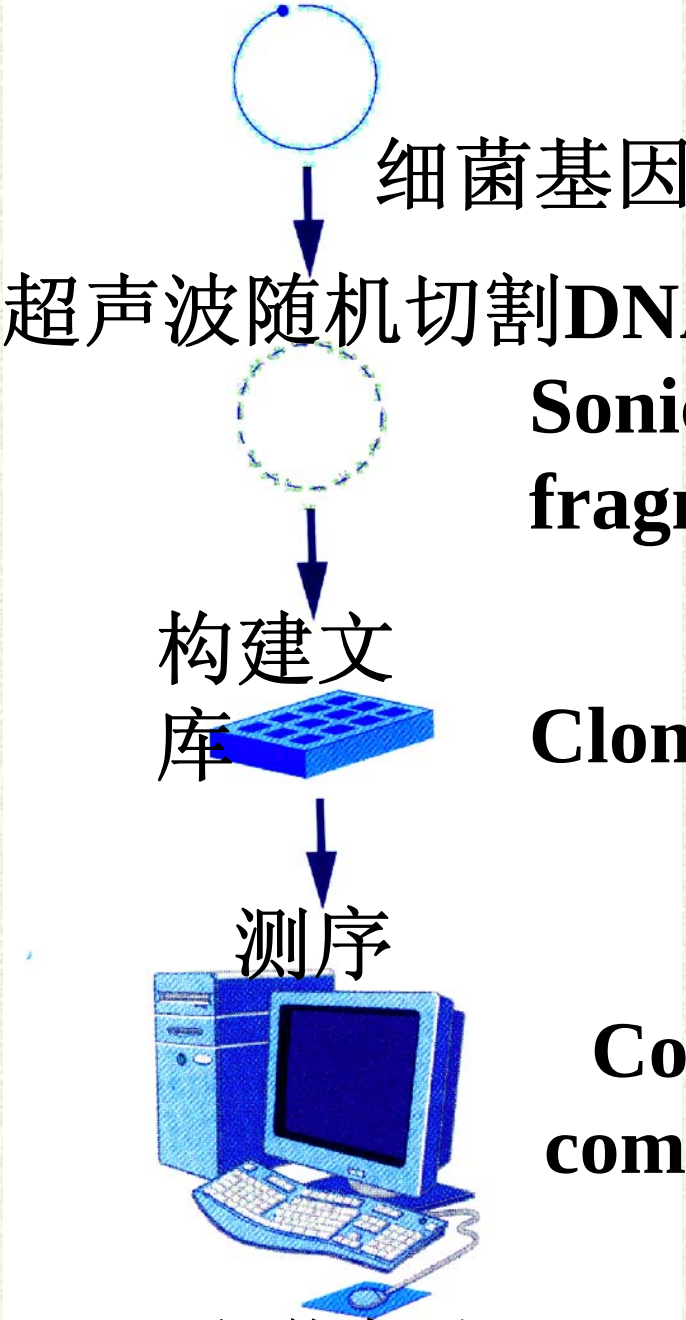
每一个互相重叠的克隆都会被测序，序列组装成**3.2×10^9**人类基因组。

PE&PGE

10．2．3 鸟枪法基因组序列分析技术及其改良

受序列分析技术限制，一次测序的长度不能超过1kb，目前往往采用所谓的全基因组鸟枪法测序技术，随机挑选插入基因组**DNA**的质粒做测序反应，然后用计算机程序进行序列拼接。

鸟枪法测序


细菌基因组
超声波随机切割DNA
Sonicated fragments
构建文库
Cloned library
测序
Computer with compiler software
组装序列

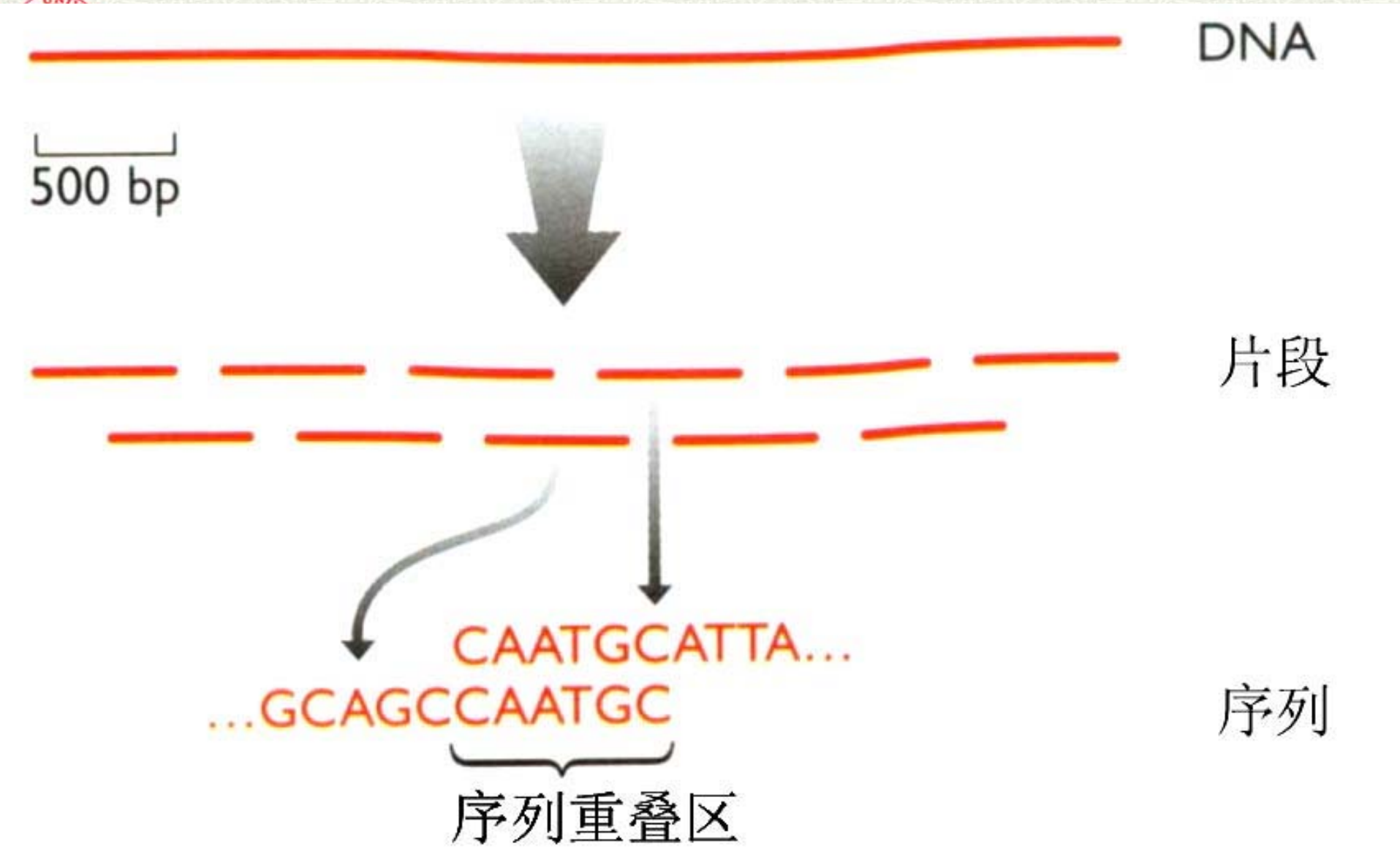


图10-14 基因组DNA的鸟枪法测序原理示意图。

对某基因组文库全部克隆片段进行末端序列测定中未测到的碱基数，即缺口**(gap)**，与已测定的总碱基数相关。随着已测定碱基数的增加，缺口的总碱基数目会按照泊松公式的一个推论（$P=e^{-m}$）迅速减小。

其中**P**为基因组中某个碱基未被测定的概率，**m**为所测定的碱基数与基因组大小相比的倍数。**m**越大**P**值越小。

当 **m=5**（即随机测定的碱基数达到基因组 **5** 倍时），基因组中未测定的碱基数为总碱基数的 **0.67%(e^{-5}=0.0067)**。

对流感嗜血杆菌基因组**(1.83Mb)**来说，可能留有 **128** 个平均长度为 **100bp** 的缺口。

鸟枪法测序的缺点：

随着所测基因组总量增大，所需测序的片段大量增加，各个片段重叠成一个连续体的概率是 $2n^2-2n$

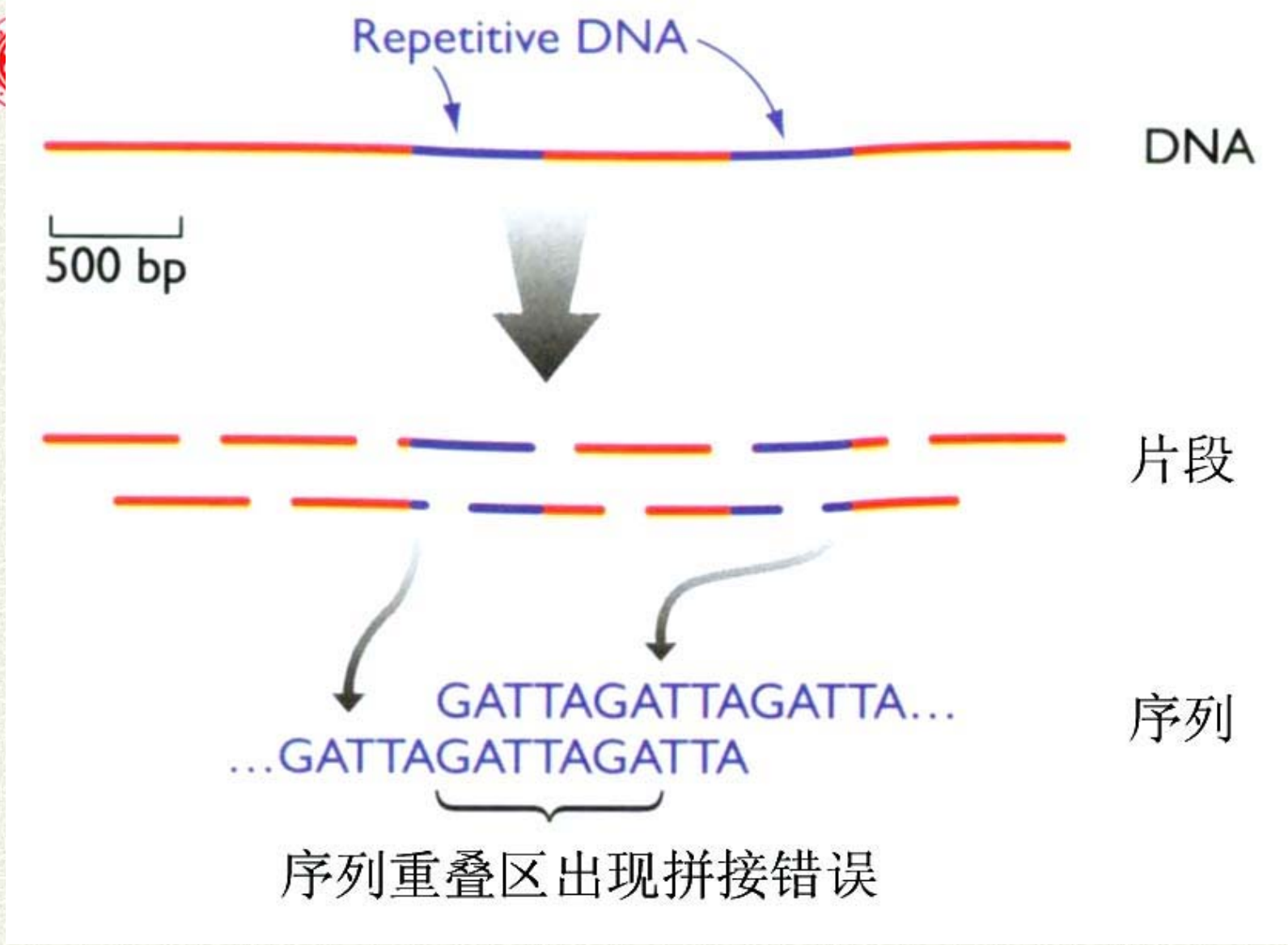


图10-15 鸟枪法测序技术不能鉴别高等真核生物基因组中的重复序列。

对鸟枪法的改进

(1) **Clone contig** 法。首先用稀有内切酶把待测基因组降解为数百 **kb** 以上的片段，再分别测序。

(2) 靶标鸟枪法**(direted shotgun)**。首先根据染色体上已知基因和标记的位置来确定部分 **DNA** 片段的相对位置，再逐步缩小各片段之间的缺口。

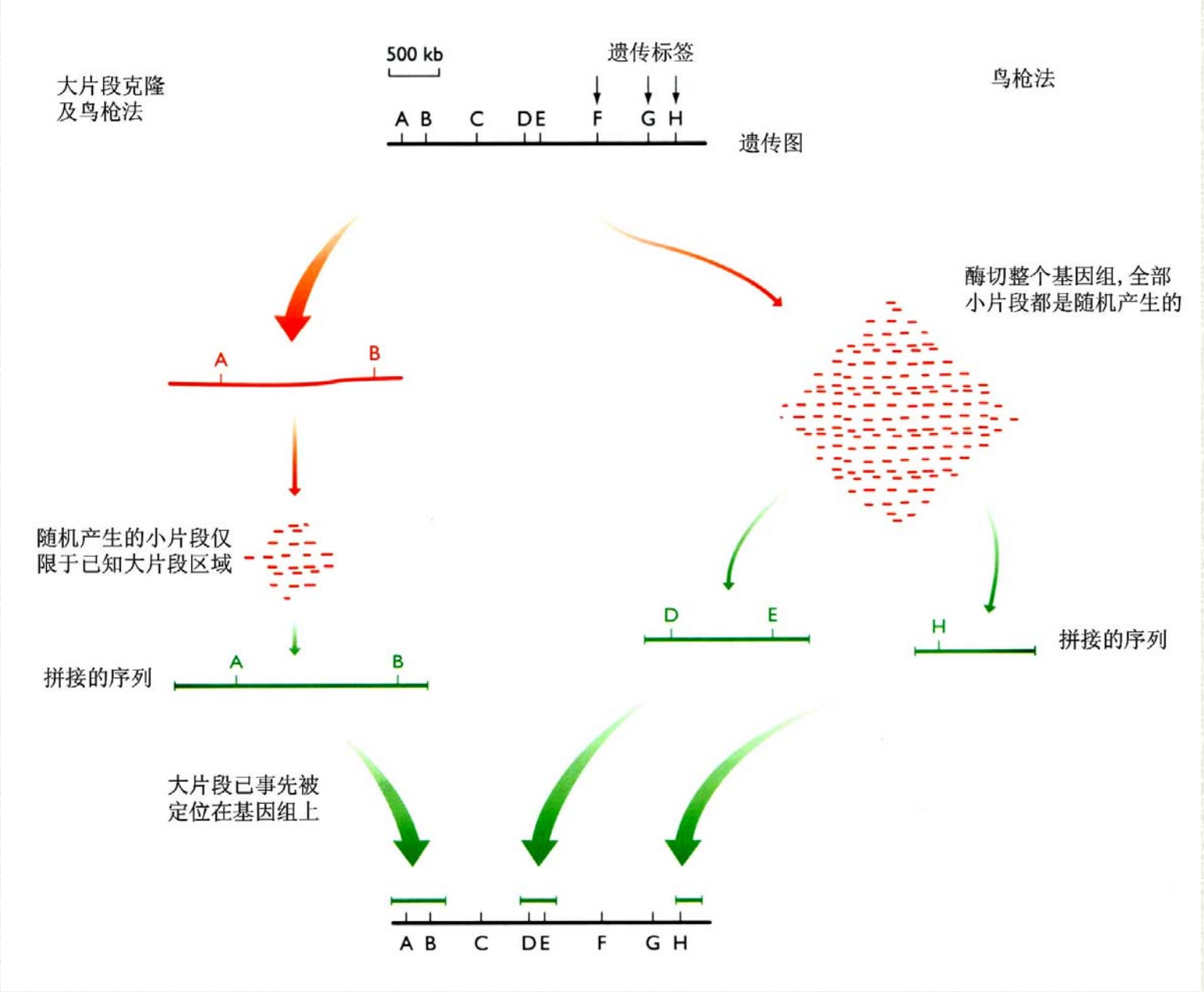


图10-16 改进后的鸟枪测序法原理图。

10. 3 比较基因组学（Comparative genomics）及功能基因组学研究

基因组的序列可被分为三类：

（一）通过比较确知其生理功能的；

（二）在数据库中有相匹配的蛋白质序列，但并不知道其功能的；

（三）在现有数据库中找不到任何相匹配的蛋白质序列的新基因。

10．3．1 通过基因组数据进行全局性分析

低等真核生物如酵母、线虫以及高等植物拟南芥，不但基因组比较小，基因密度比较高，百万碱基对中含有200个或更多的基因，基因组90%以上由常染色质组成。

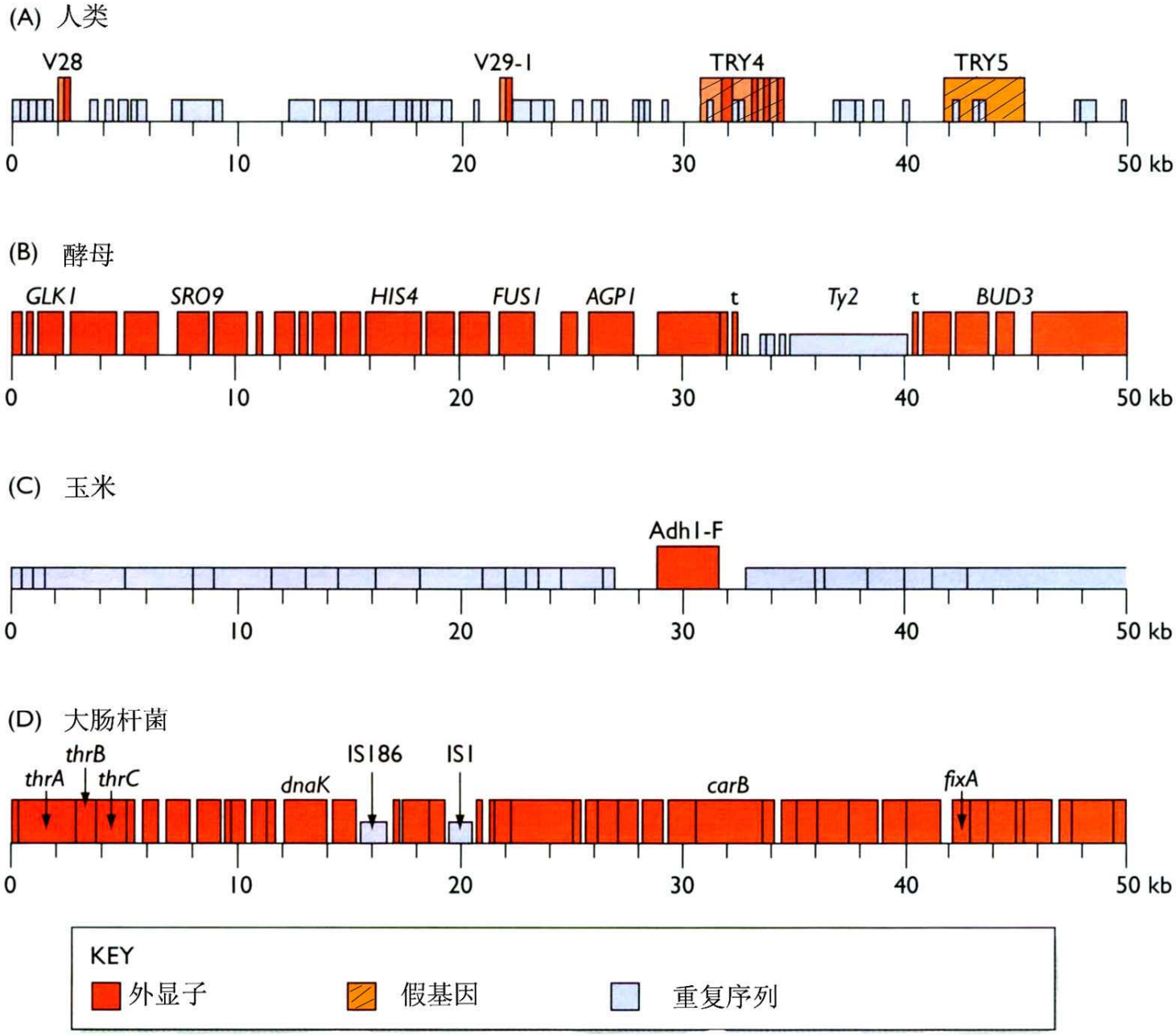

(A) 人类
V28
V29-1
TRY4
TRY5
0
10
20
30
40
50 kb
(B) 酵母
GLK1
SRO9
HIS4
FUS1
AGP1
t
Ty2
t
BUD3
0
10
20
30
40
50 kb
(C) 玉米
Adh1-F
0
10
20
30
40
50 kb
(D) 大肠杆菌
thrA
thrB
thrC
dnaK
IS186
IS1
carB
fixA
0
10
20
30
40
50 kb
KEY
外显子
假基因
重复序列

图10-17 部分典型真核和原核生物基因组成份分析。

A.人β-T细胞受体位点，在50kb大片段中只有一个基因（TRY4，编码胰蛋白酶原），一个假基因（TRY5），两个基因片段（V28，V29-1）和52个存在于全基因组范围内的重复序列，编码功能基因的序列占总序列不到3%。

B. 酵母第Ⅳ号染色体中的一个片段，其中包括**26**个蛋白质编码基因，**2**个**tRNA**基因，**5**个存在于基因组范围内的重复序列，编码功能基因的序列占总序列的**66.4%**，重复序列占**13.5%**（在所有**16**条酵母染色体中，重复序列只有**3.4%**）。在该**50kb**序列中，所有基因都不带内含子，在整个酵母基因组中，一共只有**239**个内含子，而一个人类基因就可能有多达**100**个内含子。

C. 在50kb玉米基因组中，只有1个基因，乙醇脱氢酶I-F基因，其余几乎都是重复序列。在它的5000Mb基因组序列中，50%以上可能是重复序列。

D. 大肠杆菌基因组中，**50kb**序列中可能有**43**个基因（占全序列的**85.9%**），许多基因之间甚至连一点空间都没有（**thrA**和**thrB**之间只隔了一个碱基，**thrC**基因直接位于**thrB**终止子的下游。原核生物基因中没有内含子（少数古细菌基因中可能存在极少的内含子）。原核基因组中没有重复序列，但已发现存在某些插入序列（**IS186**，**IS1**）。

果蝇和人类基因组中异染色质的比例较高，占基因组的**20%~40%**。研究果蝇核**DNA**发现，其**Y**染色体几乎完全异染色质化，第四号染色体也大部分被异染色质化，其**X**染色体在不同家系中变化最大，其异染色质化程度从**30%**到**50%**左右不等。

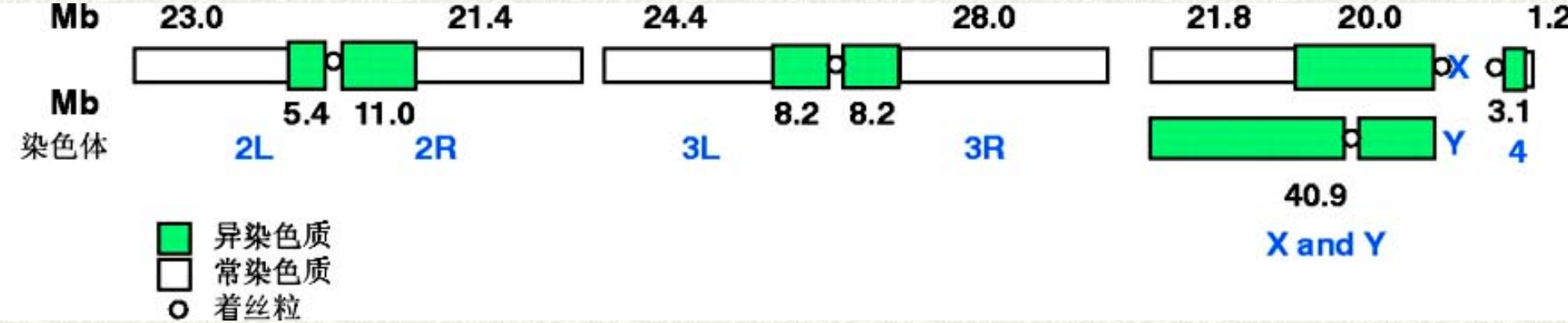


图10-18 果蝇染色体中常染色质和异染色质含量比较。

表10-4 基本完成DNA序列分析的真核生物基因组比较

物　种	完 成 年 份	总 长 度（Mp）	已 完 成 总 长的%	占常染色质%	基 因 数/百万碱基对
酵母	1996	12	93	100	483
线虫	1998	96	99	100	197
果蝇	2000	116	64	97	117
拟南芥	2000	115	92	100	221
人类第21染色体	2000	34	75	100	7
人类第22染色体	1999	34	70	97	16
人类全基因组 Public Sequence	2001	2693	84	90	12
人类全基因组 Celera Sequence	2001	2654	83	99-93	15

大肠杆菌基因组中，尚有38%以上的未知蛋白质。

与物质运转和能量代谢相关的蛋白质含量分别占蛋白质总量的9%左右。

各种功能性酶、细胞结构蛋白、调控蛋白、细胞周期相关因子及参与蛋白质合成、参与重要中间物合成与代谢等过程的蛋白质分别占总蛋白的4%以上。

参与氨基酸合成及代谢，参与DNA合成及代谢的蛋白质也都达到总蛋白的3%左右。

表7-1 大肠杆菌、伤寒杆菌和尿殖道支原体基因组中的基因总量比较与功能分析

	基 因 数		
	大肠杆菌	伤寒杆菌	尿殖道支原体
总读码框数	4288	1727	470
氨基酸合成	131	68	1
辅基等的合成	103	54	5
核苷酸合成	58	53	19
细胞膜合成与装配	237	84	17
能量代谢	243	112	31
其它合成代谢	188	30	6
脂肪代谢	48	25	6
DNA复制、重组和修复	115	87	32
蛋白质结构	9	6	7
调控蛋白	178	64	7
转录	55	27	12
翻译	182	141	101
吸收与运转	427	123	34

分类

表10-5 大肠杆菌所编码的蛋白质分类

功　　能	数 量	占蛋白质总量的%
调控蛋白	178	4.15
细胞结构蛋白	224	5.22
膜蛋白	13	0.3
外源蛋白	87	2.03
物质运转相关蛋白	427	9.95
能量代谢相关蛋白	373	8.70
DNA合成与代谢	115	2.68
转录及RNA合成、代谢与修饰	55	1.28
翻译及蛋白质修饰	182	4.24
细胞周期相关因子	188	4.38
辅基、辅因子及其载体	103	2.40
伴侣蛋白	9	0.21
核苷酸的合成与代谢	58	1.35
氨基酸的合成与代谢	131	3.06
脂肪酸及磷脂的合成与代谢	48	1.12
重要中间物合成与代谢	188	4.38
酶	251	5.85
其它（功能已知蛋白）	26	0.61
未知蛋白	1632	38.06

表10-6人类基因组数据库中蛋白质编码基因的部分重要参数比较

性　　质	平　均　值
外显子（内源性）长度（bp）	145
外显子（内源性）个数	8.8
内含子长度（bp）	3365
3’非翻译区（UTR）	770
5’非翻译区（UTR）	300
开放读码框（ORF）的长度（bp）	1340
核基因长度（kp）	27

人类基因的平均长度为**27kb**左右，

含有**8.8**个长约**145bp**的外显子，

内含子的长度却达到**3365bp**左右，

3’非翻译区（UTR）的平均长度为**770bp**，

5’非翻译区的平均长度为**300bp**，

开放读码框的平均长度只有**1340 bp**，

编码**447**个氨基酸 。

研究发现：

原始生物中单拷贝基因较多，流感嗜血杆菌中单拷贝基因占88.8%，

酵母中占71.4%，

果蝇中占72.5%，

线虫中占55.2%，

拟南芥中只占约35.0%。

表10-7 不同物种中单拷贝基因的数量及占基因总数百分比

物 种	单拷贝基因个数	单拷贝基因占基因总量的%
流感嗜血杆菌	**1587**	**88.8**
酵母	**5105**	**71.4**
果蝇	**10736**	**72.5**
线虫	**14177**	**55.2**
拟南芥	**11601**	**35.0**

Number of Genes in each Arabidopsis Chromosomes

	TAIR	NCBI
Chromosome I	7437	8018
Chromosome II	4753	5152
Chromosome III	5825	6362
Chromosome IV	4574	4848
Chromosome V	6789	7120
Total	29,388	31500

Empirical Analysis of Transcriptional Activity in the *Arabidopsis* Genome.

Science (2003) 302: 842-846.

**Kayoko Yamada,1* Jun Lim,2* Joseph M. Dale,1 Huaming Chen,2,3
Paul Shinn,2,3 Curtis J. Palm,4 Audrey M. Southwick,4
Hank C. Wu,1 Christopher Kim,2,3 Michelle Nguyen,4 Paul Pham,1
Rosa Cheuk,2,3 George Karlin-Newmann,4 Shirley X. Liu,1
Bao Lam,4 Hitomi Sakano,1 Troy Wu,4 Guixia Yu,1
Molly Miranda,4 Hong L. Quach,1 Matthew Tripp,4
Charlie H. Chang,1 Jeong M. Lee,1 Mitsue Toriumi,1
Marie M. H. Chan,1 Carolyn C. Tang,1 Courtney S. Onodera,1
Justine M. Deng,1 Kenji Akiyama,5 YasserAnsar i,2
Takahiro Arakawa,6 Jenny Banh,1 Fumika Banno,1 Leah Bowser,4
Shelise Brooks,3 Piero Carninci,6,7 Qimin Chao,3 Nathan Choy,2
Akiko Enju,5 Andrew D. Goldsmith,1 Mani Gurjal,4
Nancy F. Hansen,4 Yoshihide Hayashizaki,6,7
Chanda Johnson-Hopson,3 Vickie W. Hsuan,1 Kei Iida,5
Meagan Karnes,2 Shehnaz Khan,3 Eric Koesema,2 Junko Ishida,5
Paul X. Jiang,1 Ted Jones,4 Jun Kawai,6,7 Asako Kamiya,5
Cristina Meyers,2 Maiko Nakajima,5 Mari Narusaka,5
Motoaki Seki,5,8 Tetsuya Sakurai,5 Masakazu Satou,5
Racquel Tamse,4 Maria Vaysberg,1 Erika K. Wallender,1
Cecilia Wong,1 Yuki Yamamura,1 Shiaulou Yuan,1
Kazuo Shinozaki,5,8 Ronald W. Davis,4,9 Athanasios Theologis,1*†
Joseph R. Ecker2,3*†**

NLPE&PGE

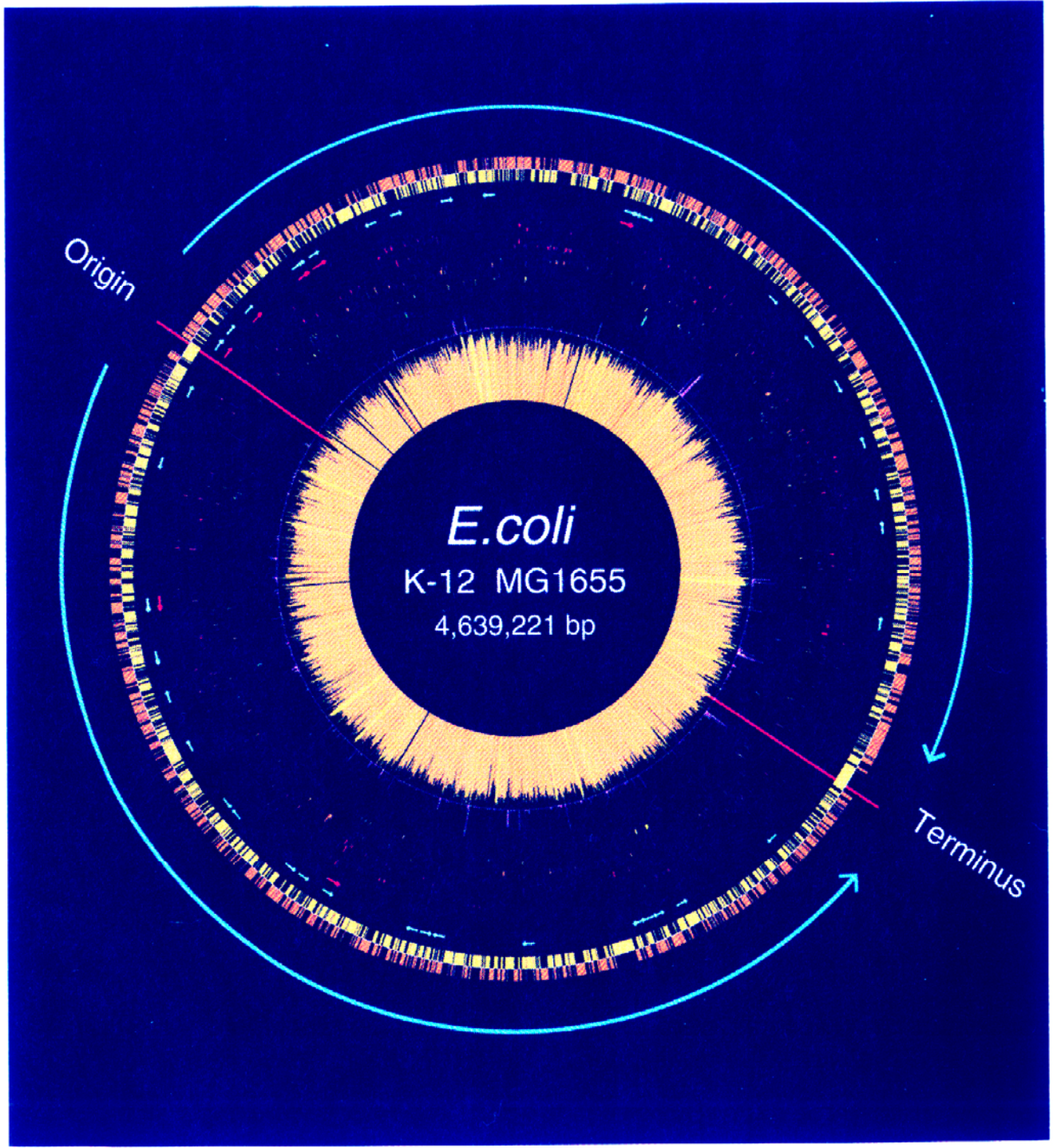
To identify transcription units in the *Arabidopsis* genome, we used custom high-density oligonucleotide arrays that **tile** the entire genome. Our analysis revealed that the *Arabidopsis* genome contains **28,000** genes of which **15,033** (59%) annotated genes lack a fl cDNA clone.

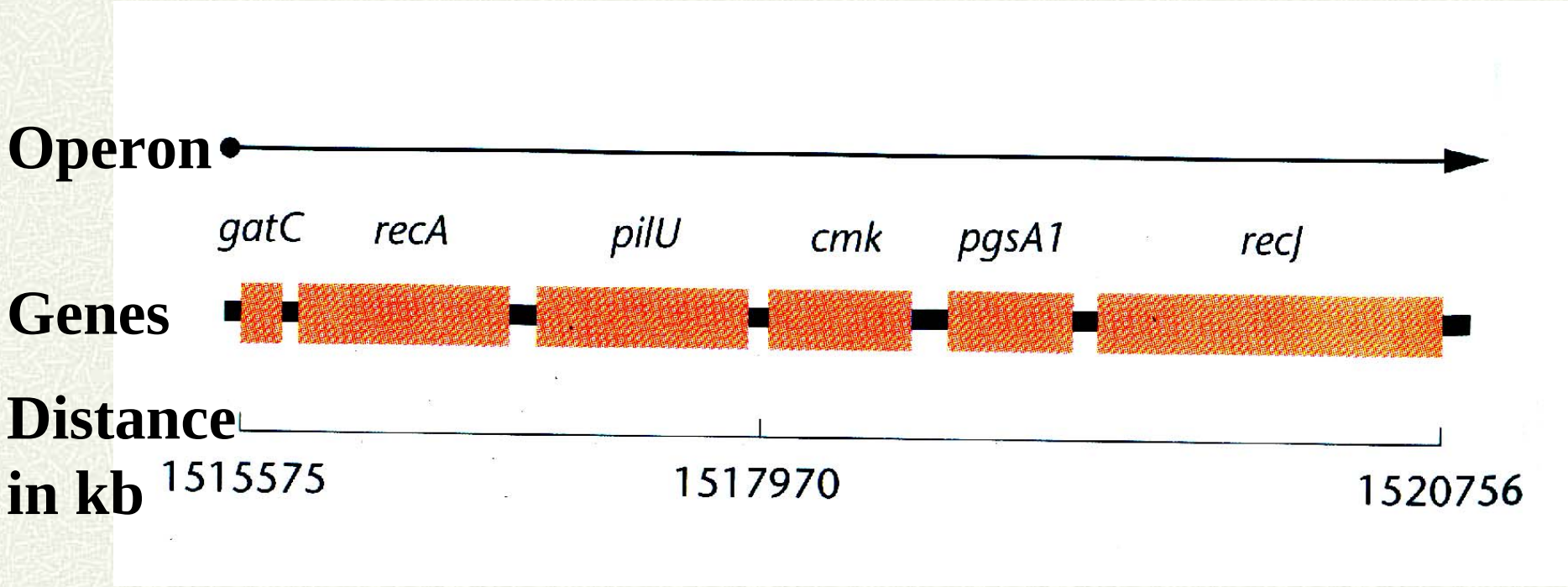
10．3．2通过基因组数据进行比较基因组学研究

尿殖道支原体带有已知最小的基因组，可依此确定能自我复制的细胞必需的一套最少的核心基因。

TABLE 18.3 Genome Size and Gene Number in Selected Prokaryotes

	Genome Size (Mb)	*Number of Genes*
Archaea		
Archaeoglobus fulgidis	2.17	2,493
Methanococcus jannaschii	1.66	1,813
Thermoplasma acidophilum	1.56	1,509
Eubacteria		
Escherichia coli	4.64	4,397
Bacillus subtilis	4.21	4,212
Haemophilus influenzae	1.83	1,791
Aquifex aeolicus	1.55	1,552
Rickettsia prowazekii	1.11	834
Mycoplasma pneumoniae	0.82	710
Mycoplasma genitalium	0.58	503


Origin
E.coli
K-12 MG1655
4,639,221 bp
Terminus


Operon
gatC
recA
pilU
cmk
pgsA1
recJ
Genes
Distance
in kb
1515575
1517970
1520756

在一个操纵子中既有参与蛋白质合成的基因**gatC**，也有参与ＤＮＡ重组的基因**recA**和**recJ**，参与细胞运动的基因**pilU**，参与核苷酸生物合成的基因**cmk**和参与脂肪酸生物合成的基因**pgsA1**.

流感嗜血杆菌的基因组为**1.83Mb**，而尿殖道支原体的基因组只有**0.58Mb**，二者相差**3**倍多，那么，基因组大小影响了基因数目还是基因尺度？

流感嗜血杆菌基因大小平均**900bp**，尿殖道支原体的基因为**1040bp**，基因大小差不多。

流感嗜血杆菌中平均**1042bp** 有**1**个基因，尿殖道支原体中平均**1235bp** 有**1**个基因。

可见基因组尺度减小并不引起基因密度的增加和基因本身尺寸的减小。

二者的差别在于基因数量上，流感嗜血杆菌基因组有**1743**个**ORF**，而尿殖道支原体只有**470**个**ORF**。

表10-8 流感嗜血杆菌和尿殖道支原体各类主要基因比较

分　类	流感嗜血杆菌	尿殖道支原体
总ORF数	1727	470
氨基酸合成	68	1
辅基等的合成	54	5
核苷酸合成	53	29
细胞膜合成与装配	84	27
能量代谢	112	31
糖代谢	30	6
脂肪代谢	25	6
DNA复制、重组和修复	87	42
蛋白质高级结构形成	6	7
调控蛋白	64	7
转录	27	22
翻译	141	101
吸收与转运	123	34

通过对尿殖道支原体与流感嗜血杆菌这两个亲缘关系较远的生物基因组的比较, 选取其共同的基因（共**240**个），再加上一些其他基因，最后组成一套含**256**个基因的最小基因组。

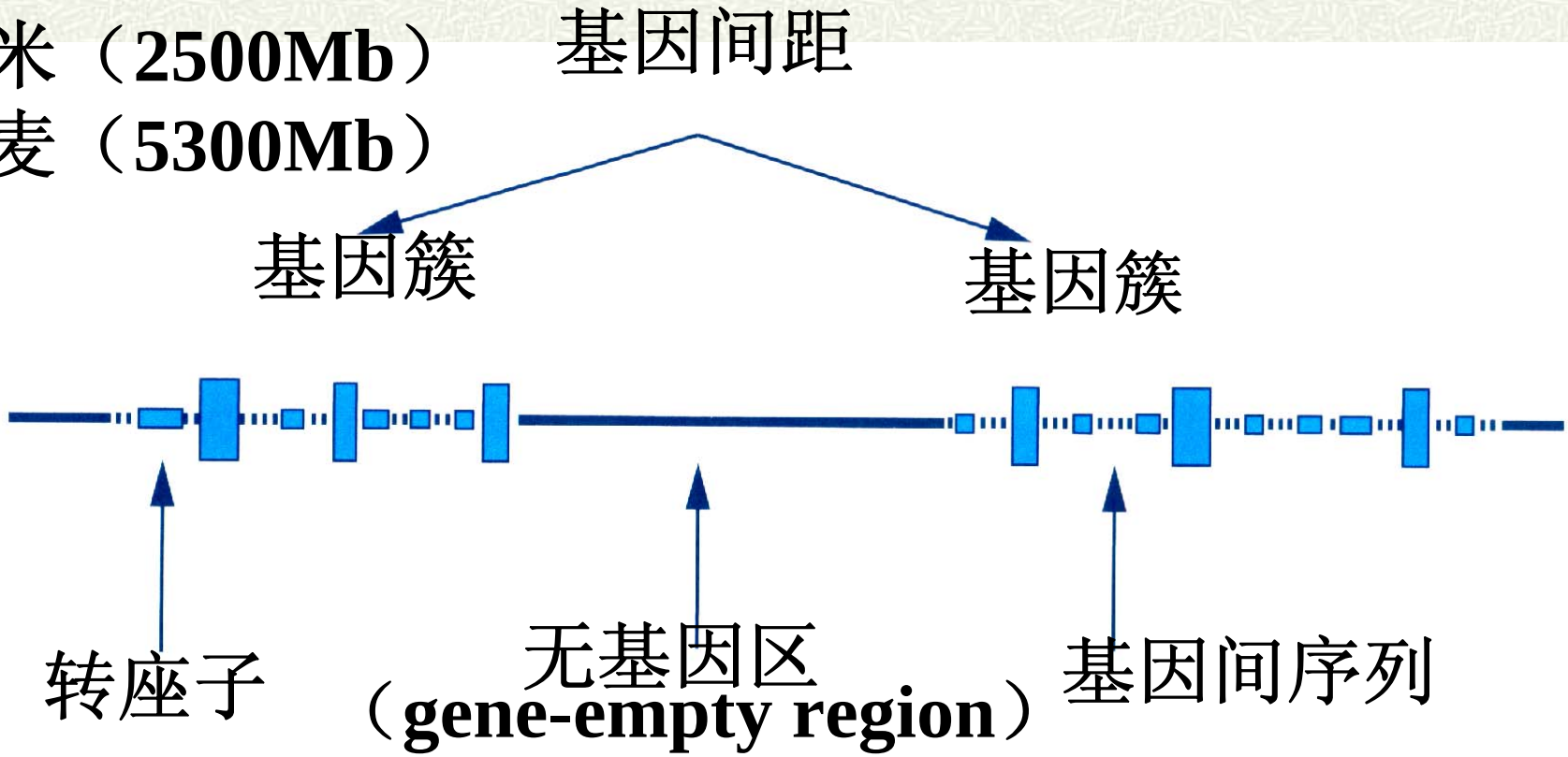
单细胞真核生物酿酒酵母基因组为12，068kb，比单细胞的原核生物和古细菌大一个数量级。

酿酒酵母基因组共有5887个ORF，比原核生物和古细菌要多得多。酿酒酵母的基因密度为1个基因/2kb，小于原核生物流感嗜血杆菌和尿殖道支原体等。

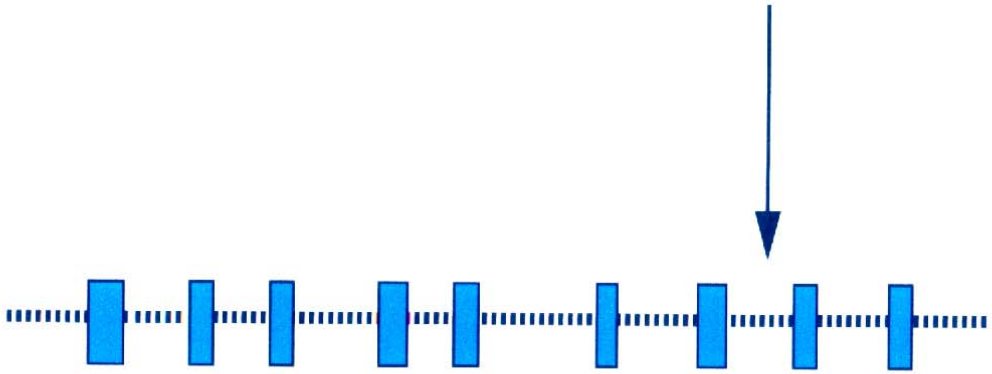
酿酒酵母是最小的真核基因组，裂殖酵母其次，其密度是**1/2.3kb**，

简单多细胞生物线虫的基因密度为**1/30kb**。酿酒酵母只有**4%**的编码基因有内含子，而裂殖酵母则有**40%**编码基因有内含子。

a.水稻（415Mb）
玉米（2500Mb）
大麦（5300Mb）



b.拟南芥
（120Mb）



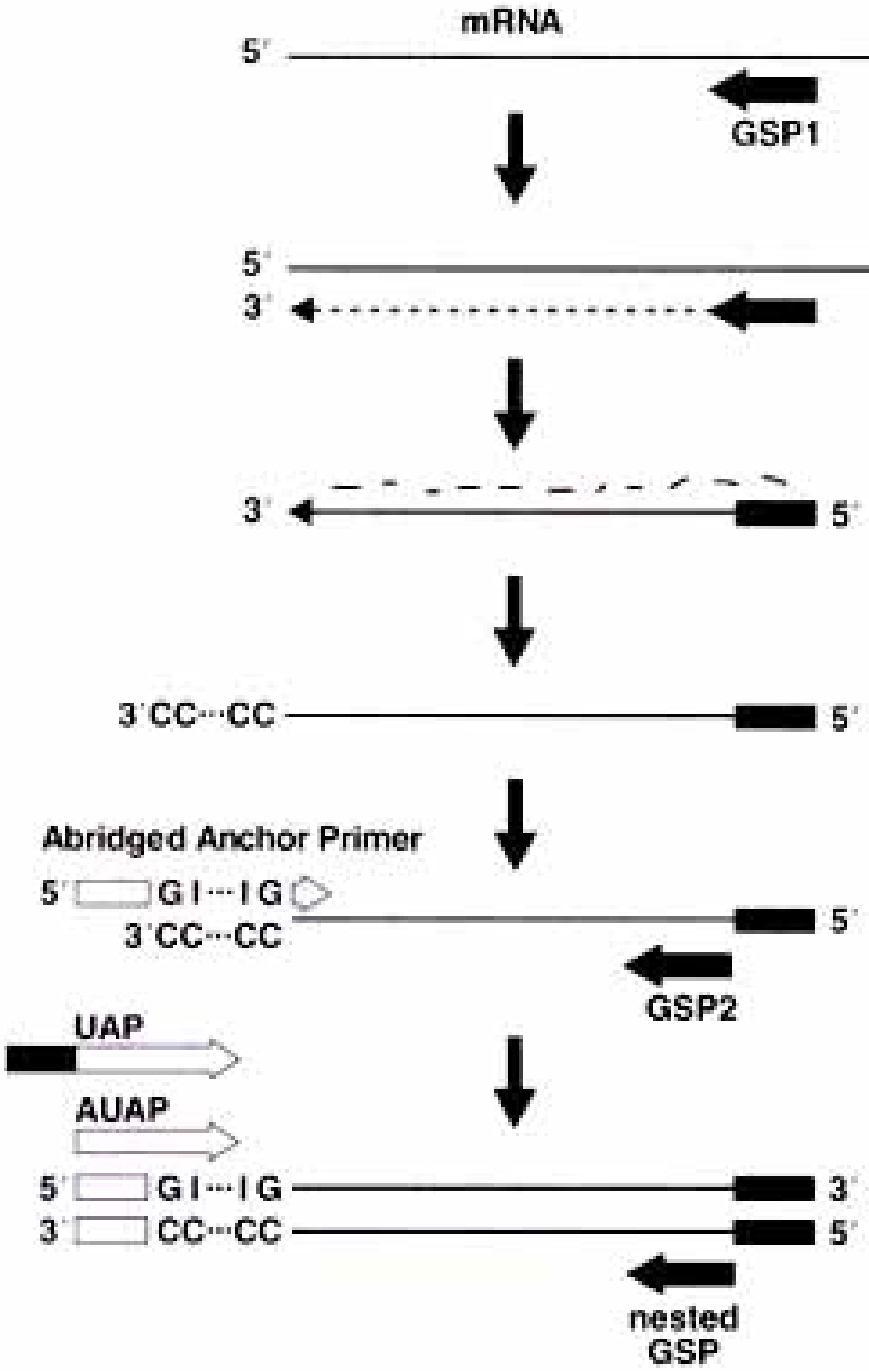
10. 3. 3 功能基因组学研究

整个基因组序列的获得为生物学带来了一种称为功能基因组学的新方法，即在基因组水平上阐明**DNA**序列的功能。

人类和各种模式生物的全长**cDNA**克隆对基因的发现及功能分析都极为有用，因此，获得全长**cDNA**的技术和发现稀有转录物的技术都将被放在高度优先的地位。

cDNA RACE(cDNA末端快速扩增)

是用于从已知cDNA片段扩增全长基因的方法，它根据已知序列设计基因片段内部特异引物，由该片段向外侧进行PCR扩增得到目的序列。用于扩增5'端的方法称为5'RACE，用于扩增3'端的称为3'RACE。



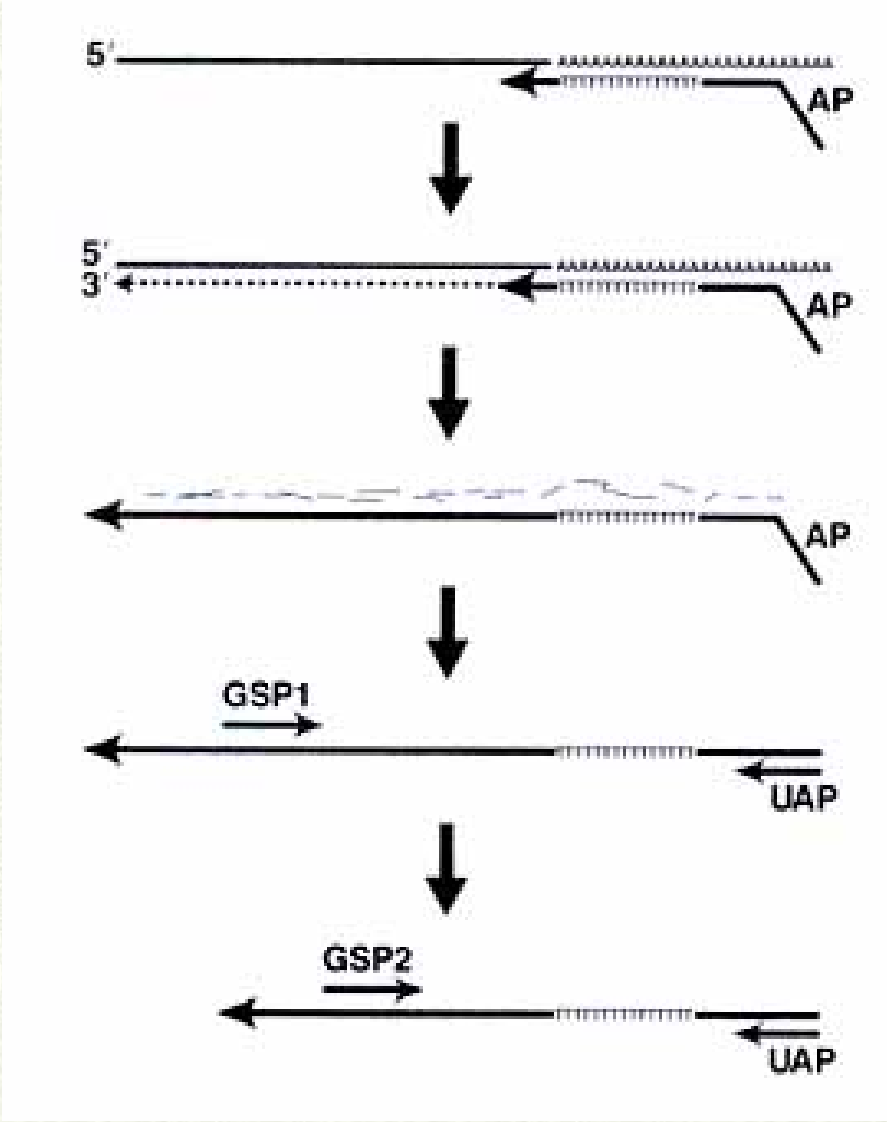
5' RACE

1. 在反转录酶的作用下，以已知基因片段内部。特异性引物启始cDNA第一条链的合成

2. RNase混合物降解模板mRNA，纯化cDNA第一条链。

3. 用末端转移酶在cDNA链3'端加入连续的dCTP

4. 以连有oligo(dG) 的锚定引物和基因片段内部特异的nested引物进行PCR扩增，以期得到目的片段，并可用nest PCR进行检测。



3'RACE

1. 在反转录酶的作用下，以连有可以和polyA配对的oligo(dT)的锚定引物启始cDNA第一条链的合成。

2.用RNaseH 降解模板mRNA。

3. 用通用锚定引物和基因片段内部特异引物进行PCR扩增得到目的3'片段，并可用nest PCR的方法继续进行检测和扩增。