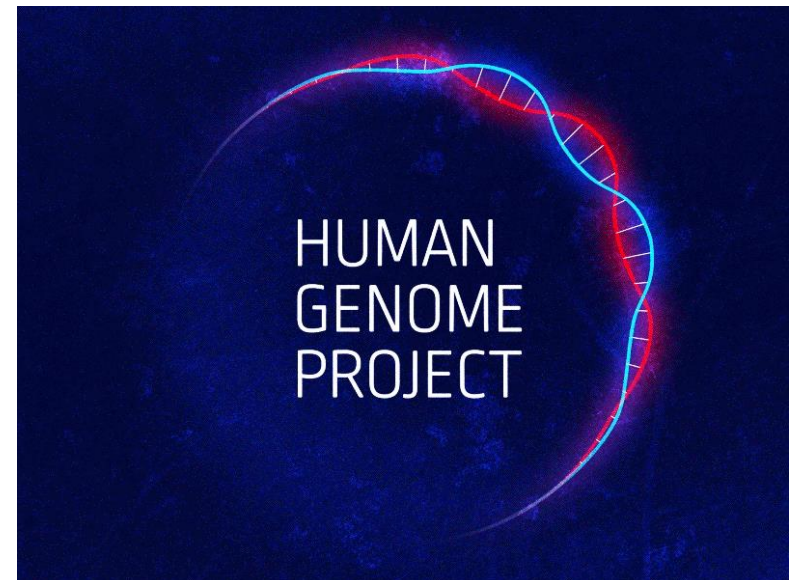


# 第十四章 基因组研究

- 人类基因组计划
- 人类基因组特点
- 模式生物基因组
- 基因组后续计划



# 一、基因组概论

- 1920年，植物学家Hans Winkler将gene和chromosome两词结合，提出了**genome（基因组）**的概念，指的是细胞内全套染色体及其所携带的全部基因。
- 1987年，遗传学家Victor McKusick首次提出了基因组学（**genomics**）的概念，并创办了同名期刊。





# 基因组学：研究基因组的科学

- 基因组学是遗传学的继续和发展，是基于基因组层次和规模的遗传学。
- 基因组学研究最主要的两个理念：
  - 1) 生命是序列的 (Life is of sequence)：遗传信息蕴藏在DNA序列以及不同方式修饰的核苷酸之中；
  - 2) 生命是数字的 (Life is digital)：代代相传的生命指令是数据化的。
- 基因组研究的核心技术：
  - 1) 测序 (Sequencing)：包括DNA、RNA和甲基化组等测序；
  - 2) 信息学 (Bioinformatics)：借助计算机软件进行表型和功能分析。

## 二、人类基因组计划 (*Human Genome Project, HGP*)

1990年，美国国会批准美国的“人类基因组计划” (*Human Genome Project, HGP*) 在10月1日正式启动。其总体规划是准备在15年内（1990—2005）至少投入30亿美元，分析人类的基因组30亿个碱基对。”



James D. Watson  
(1928-)



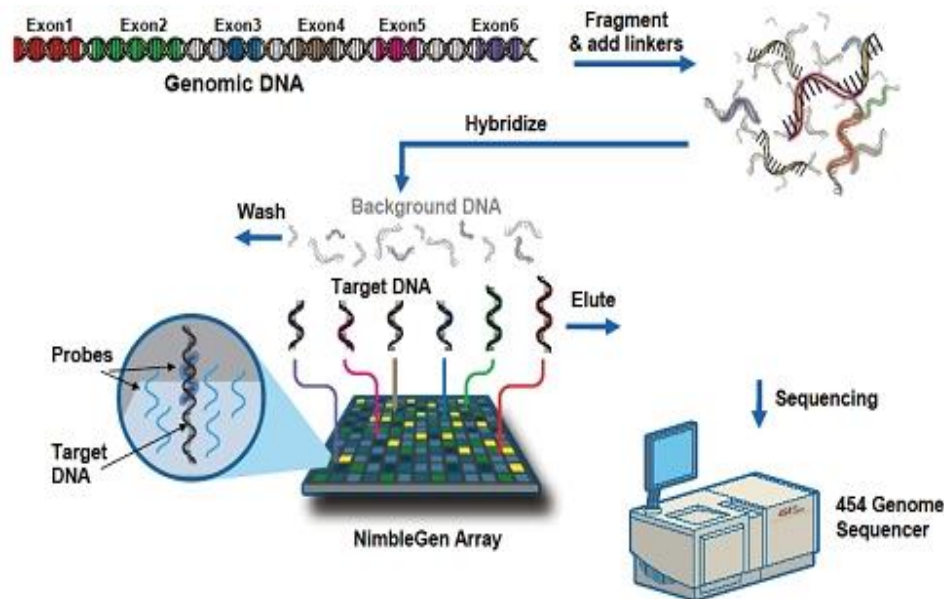
## 国际人类基因组测序协作组

(International Human Genome Sequencing Consortium, **IHGSC**)

- 美国: WASH&MIT等7家研究中心 贡献率为54%
- 英国: SANGER 1家研究中心 贡献率为33%
- 日本: RIKEN等2家研究中心 贡献率为7%
- 法国: GENOSCOPE研究中心 贡献率为2.8%
- 德国: IMB等3家研究中心 贡献率为2.2%
- 中国: 北京华大研究中心、国家南方基因研究中心、  
国家北方基因研究中心3家 贡献率为1%

# (一) DNA测序技术的四个突破:

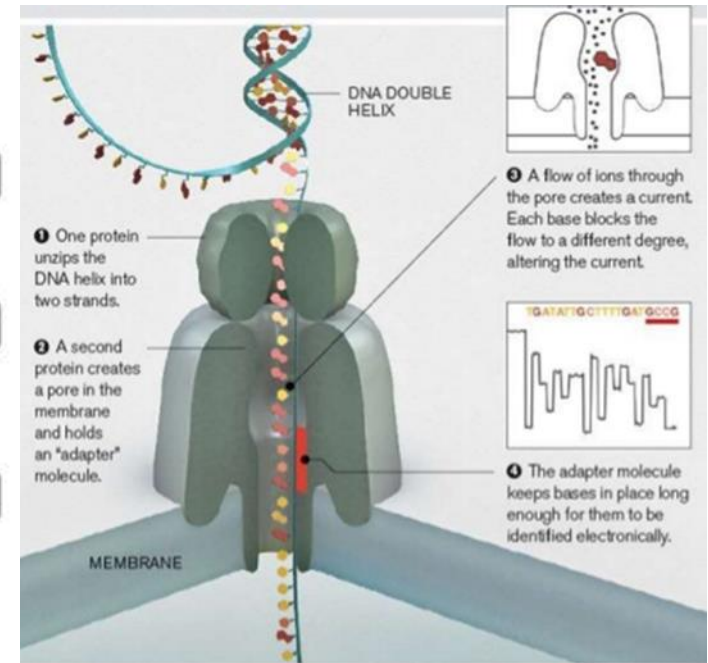
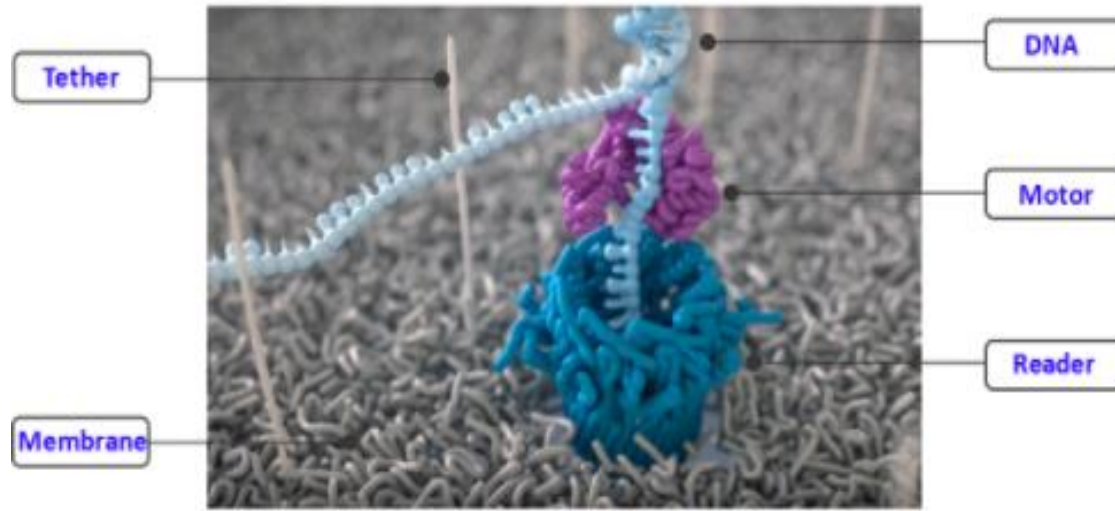
1. 直读法的创立
2. 自动化的开始——平板凝胶电泳法
3. 规模化
4. MPH (Massively Pararell High-throughput, 大规模平行高通量), 又称2<sup>nd</sup> –generation sequencing techniques:



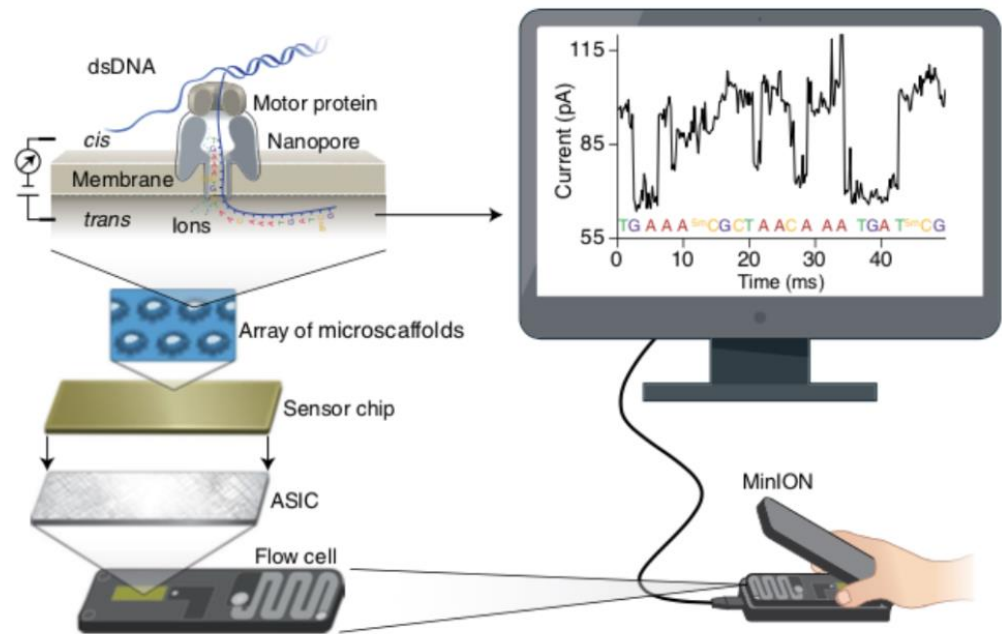
- **High throughput**
- **Short readout: 30-200bp**



## 2. 纳米孔测序：根据ssDNA或RNA模板分子通过纳米孔引起“信号”变化进行实时测序。



**Reader**——经人工改造后的具有天然的蛋白纳米孔的跨膜蛋白；  
**Membrane**——高电阻率的人工多聚物薄膜，膜两侧是离子溶液，在两侧加不同的电位，离子就会在孔中流动，形成电流；  
**Motor**——连接在测序样本接头上的马达蛋白，用于将DNA或RNA分子推入纳米孔中；  
**Tether**——锚定DNA或RNA链，防止其在溶液中飘动，并使其进入纳米孔中。

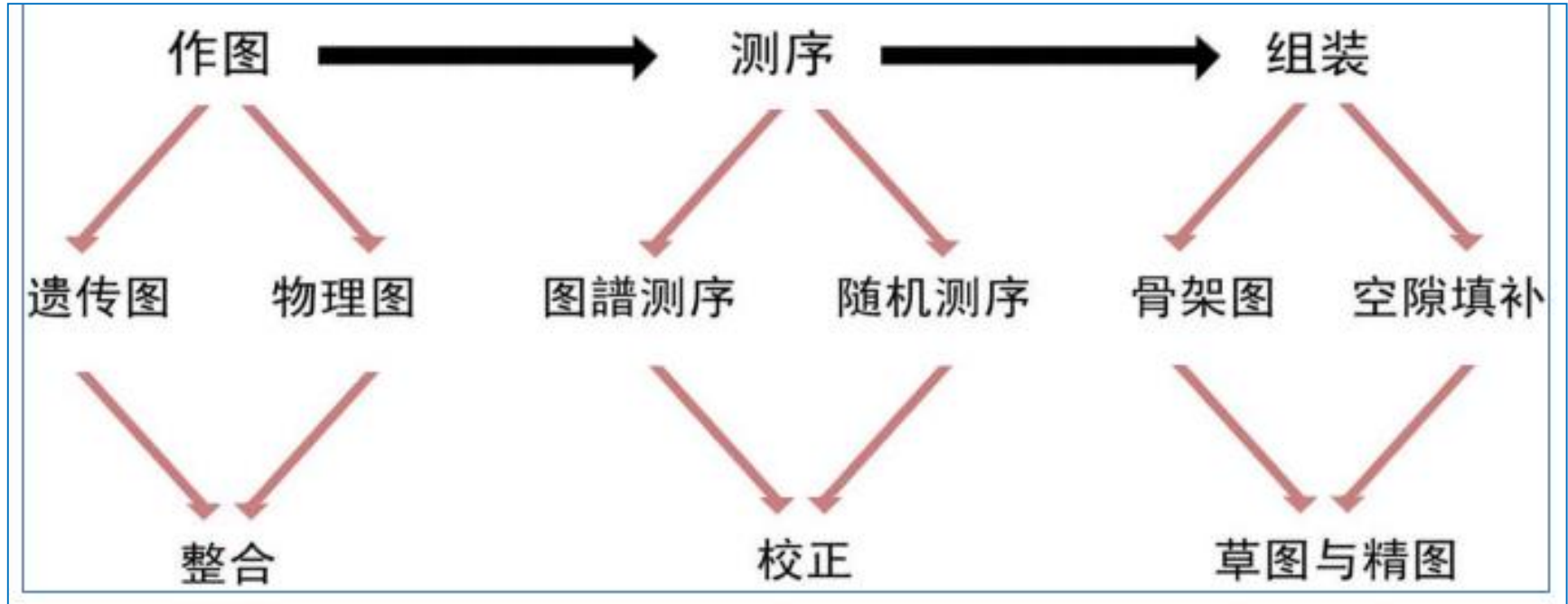


## 主要优势：

- **长读长**：Reads可达Mb；
- **设备成本低**：测序芯片可清洗再生，重复利用；
- **实时获得序列信息**：最快可在1小时内完成测序流程及数据分析，满足动态检测宏基因组需求；
- **便携式测序装置**：重量轻且占用空间小，可以随身携带随时测序；
- **直接测序**：直接测序原始DNA和RNA，不需要进行PCR扩增，保留了原始碱基修饰信息，能够直接读出甲基化的胞嘧啶。

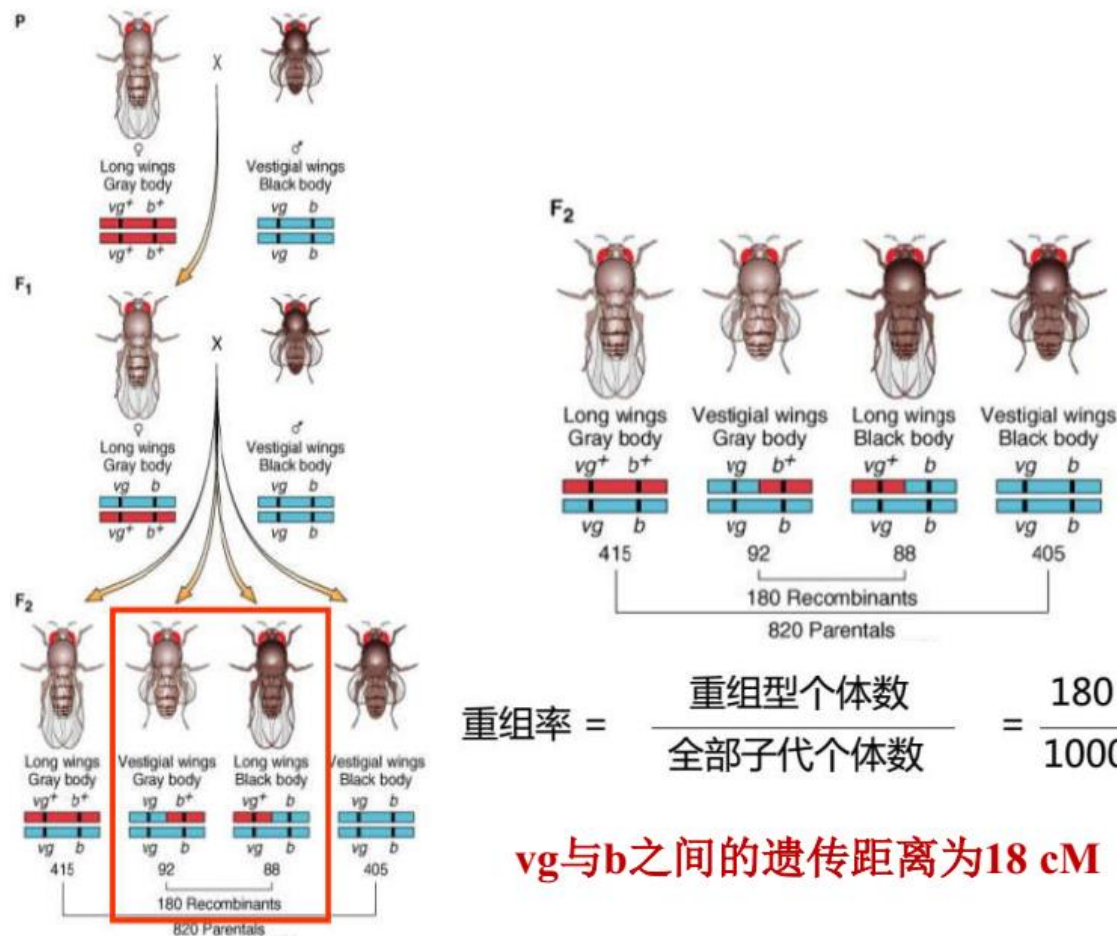


## (二) 基因组测序策略:



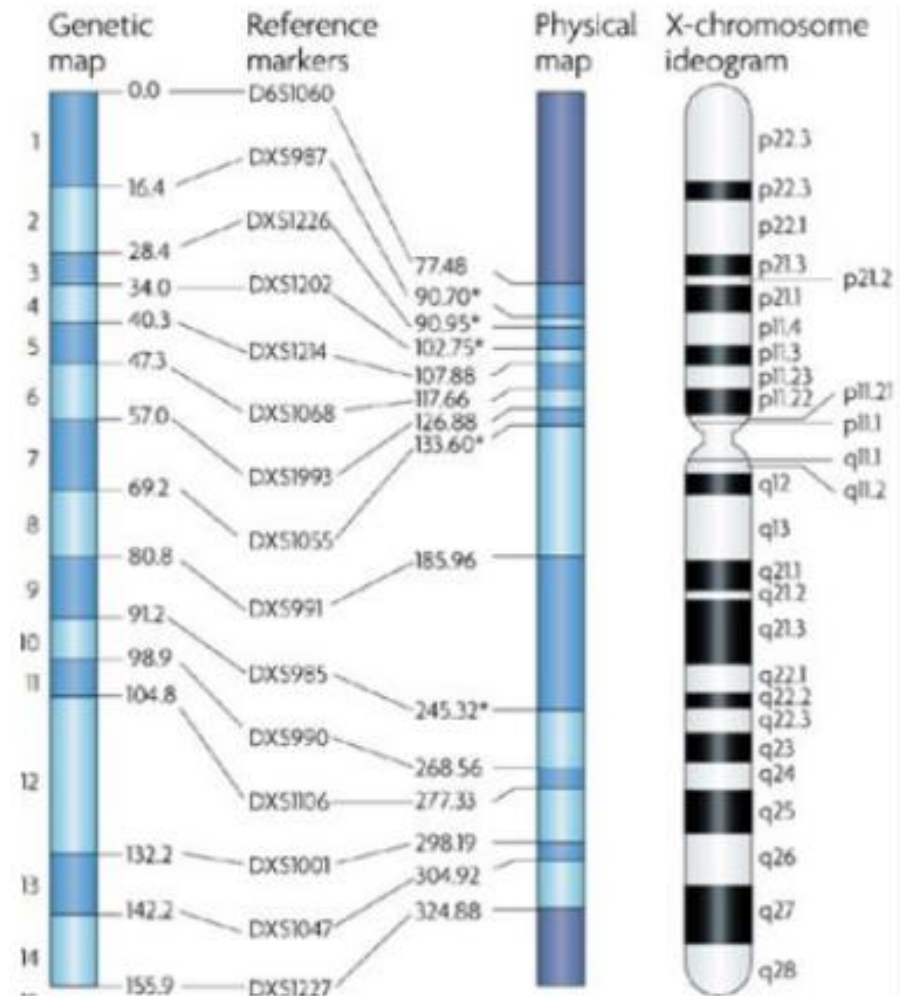
# 1. 遗传图谱 (genetic map)

通过连锁分析，计算遗传标记（或基因）间的交换频率，确定其在染色体上的相对位置与遗传距离，一般用厘摩（cM）表示。



## 2. 物理图谱 (physical map)

- 指采用分子生物学技术直接将DNA分子标记、基因或克隆标定在基因组实际位置的作图方法；
- 反映的是目标DNA分子在染色体上的真实位置，用bp计算标记之间的距离；
- 具体方法包括 荧光原位杂交、限制酶作图、序列标签位点作图以及依靠克隆作图。



Andreas Beyer, et al. 2007.



## 序列标签作图

***STS* (sequence tagged site)** :指染色体上位置固定、核酸序列已知且在基因组中**只有一份拷贝**的**DNA**短片断，一般长**200bp—500bp**。

**STS**要满足**2**个条件：

- ◆ 是一段已知的序列，可据此设计**PCR**引物来检测不同**DNA**片断中是否存在这一序列。
- ◆ **STS**在染色体上必须是独一无二的。如果在基因组中有多个位点出现，作图数据将含混不清。

STS的主要来源：只要是序列已知、单一拷贝（位置唯一）的序列都可开发成为STS。

（1）表达序列标签（expressed sequence tag, EST），EST是通过对互补DNA（complementary DNA, cDNA）进行测序分析得到的短段DNA(300~500bp)。一个EST代表了一个表达基因的部分转录片段。EST既是STS的主要来源，又是获得编码基因信息的重要途径。

（2）遗传标记SSLP。

（3）基因组内随机小片段（序列已知，位置唯一）。



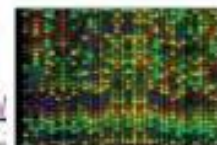
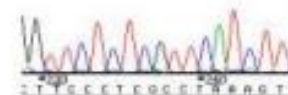
----->  
mRNA  
isolation



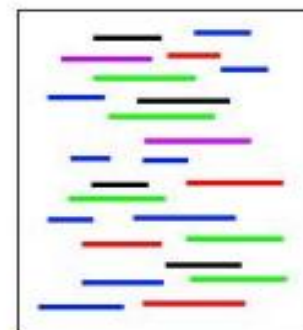
----->  
cDNA  
library  
construction



----->  
sequencing



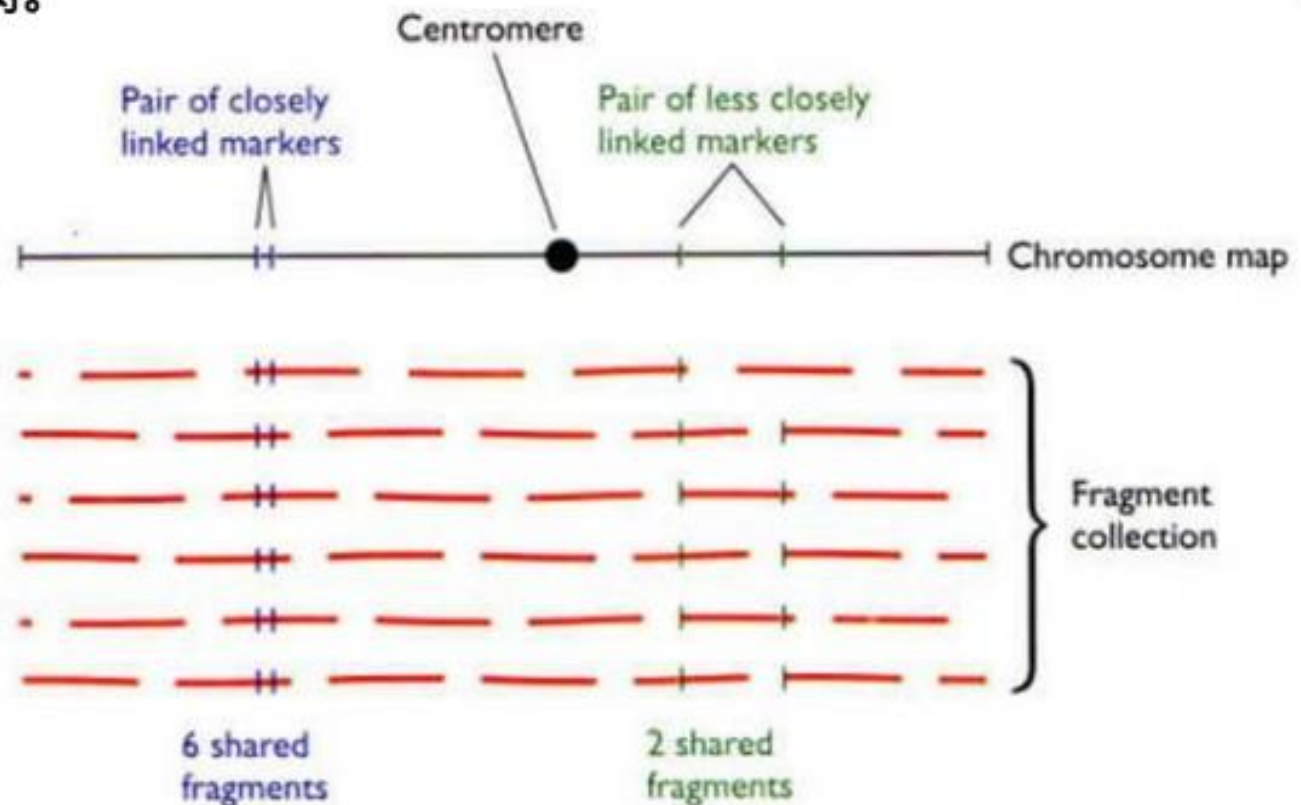
----->  
data  
collection



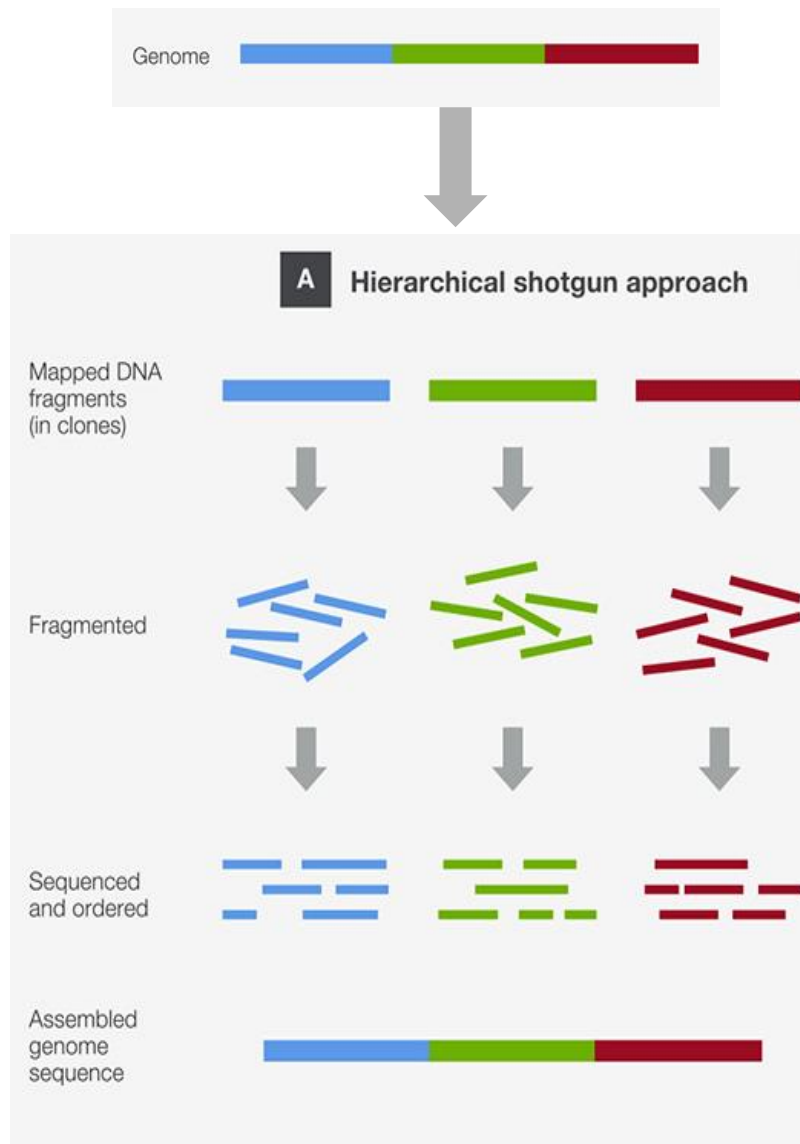
ESTs raw reads  
DNA sequences



- 利用PCR或杂交方法测定序列中的STS，将不同的STS依照它们在染色体上的位置依次排列构建图谱的方法为序列标签位点作图。
- 利用不同STS标记在一条染色体的随机断裂片段中的分离频率来计算STS标记之间的相对距离。





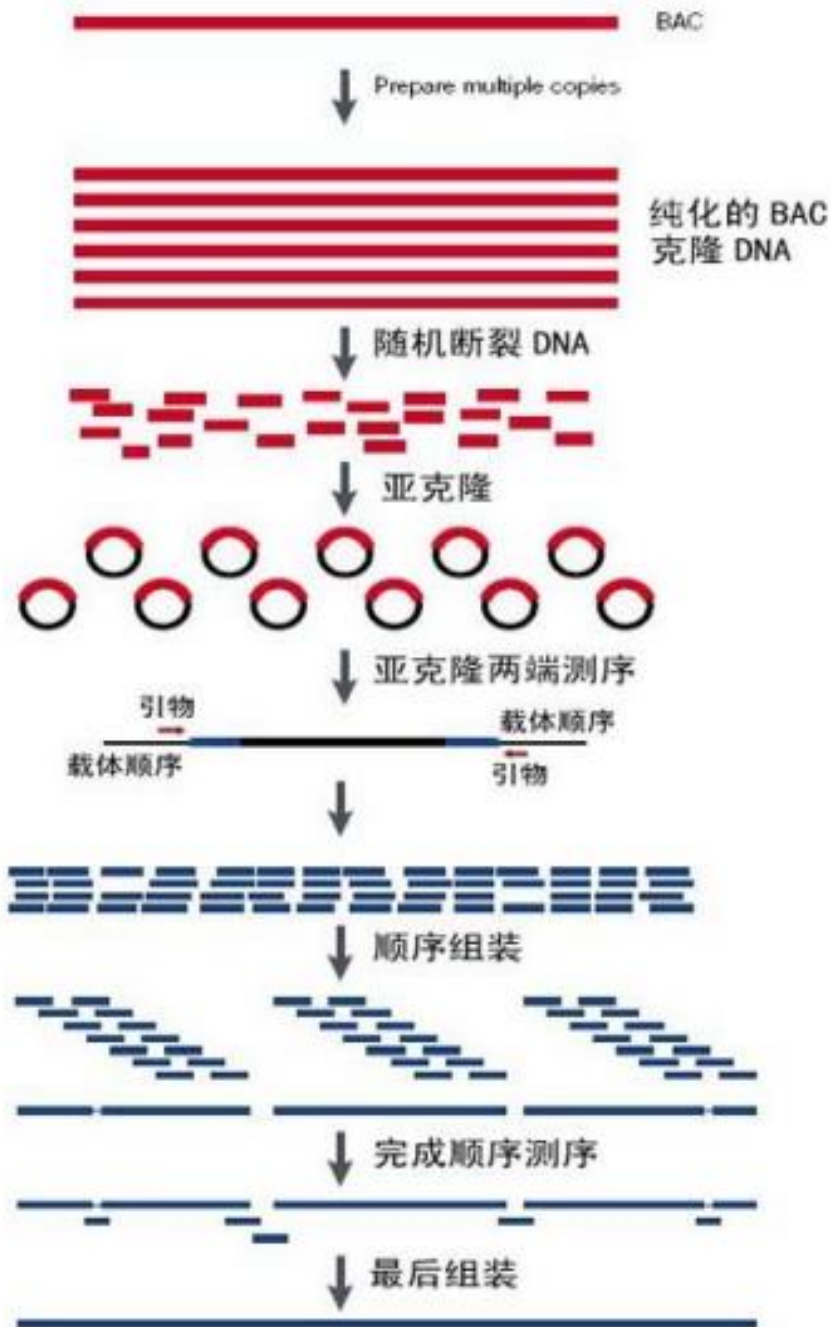


## 逐级克隆的测序策略 Hierarchical Shotgun Approach

- 先制作高密度的遗传和物理图谱;
- 再进行逐个克隆 (clone-by-clone) 测序:

其方法是:

- ① 将基因组DNA分段克隆到BAC载体中, 全基因组共构建包括300, 000个BAC克隆在内的文库;

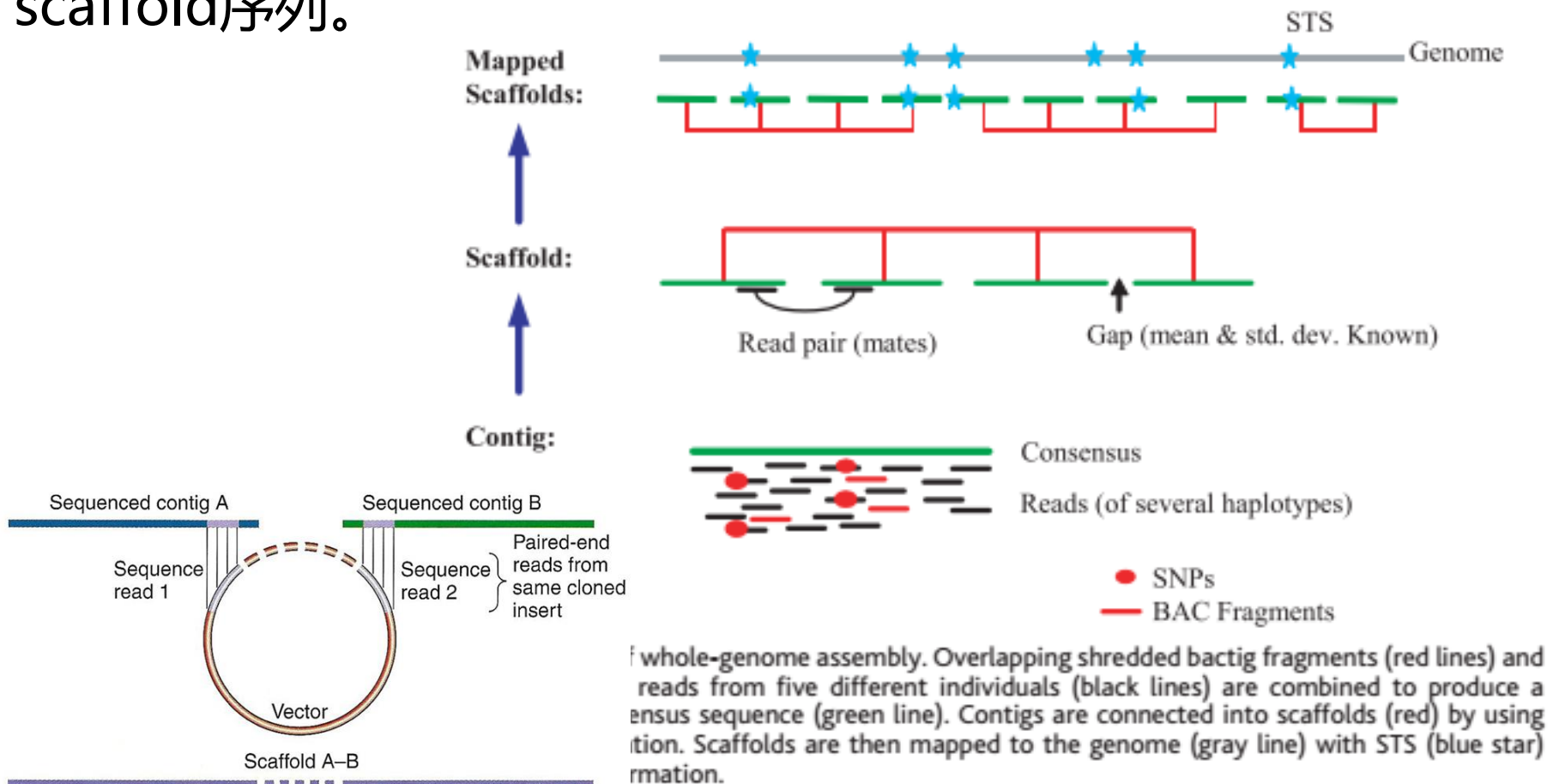


- ② 对基因组BAC文库进行亚克隆，即构建插入片段为2kb的亚克隆群，进行末端测序和拼接，对缺口进行填补；
- ③ 根据不同克隆DNA片段之间的限制性酶切位点图谱或STS图谱重叠顺序构建重叠群（contig）；
- ④ 将全部大片段进行连接，完成全基因组拼装。

**Read:** 一次测序中仪器读取的核苷酸长度。

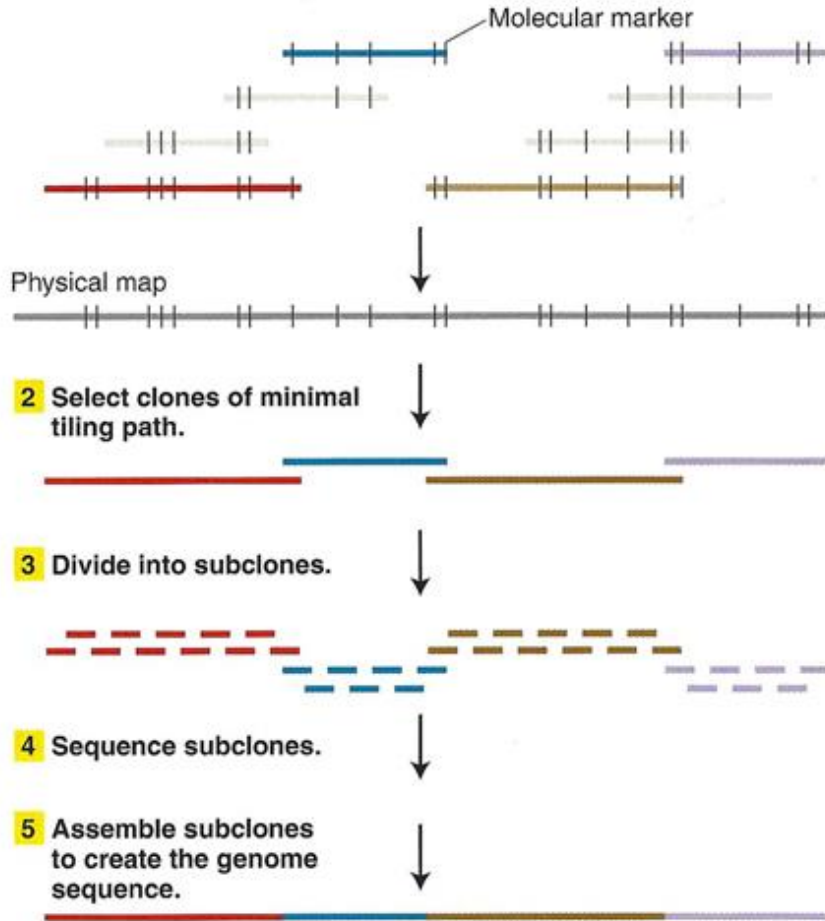
**Contig:** 通过重叠部分将相邻reads组装形成的单元称为 contig。

**Scaffold:** 利用双端测序等其他方法的信息，定位contigs在染色体上的线性排列或相对位置关系，并连接起来形成较长的 scaffold 序列。

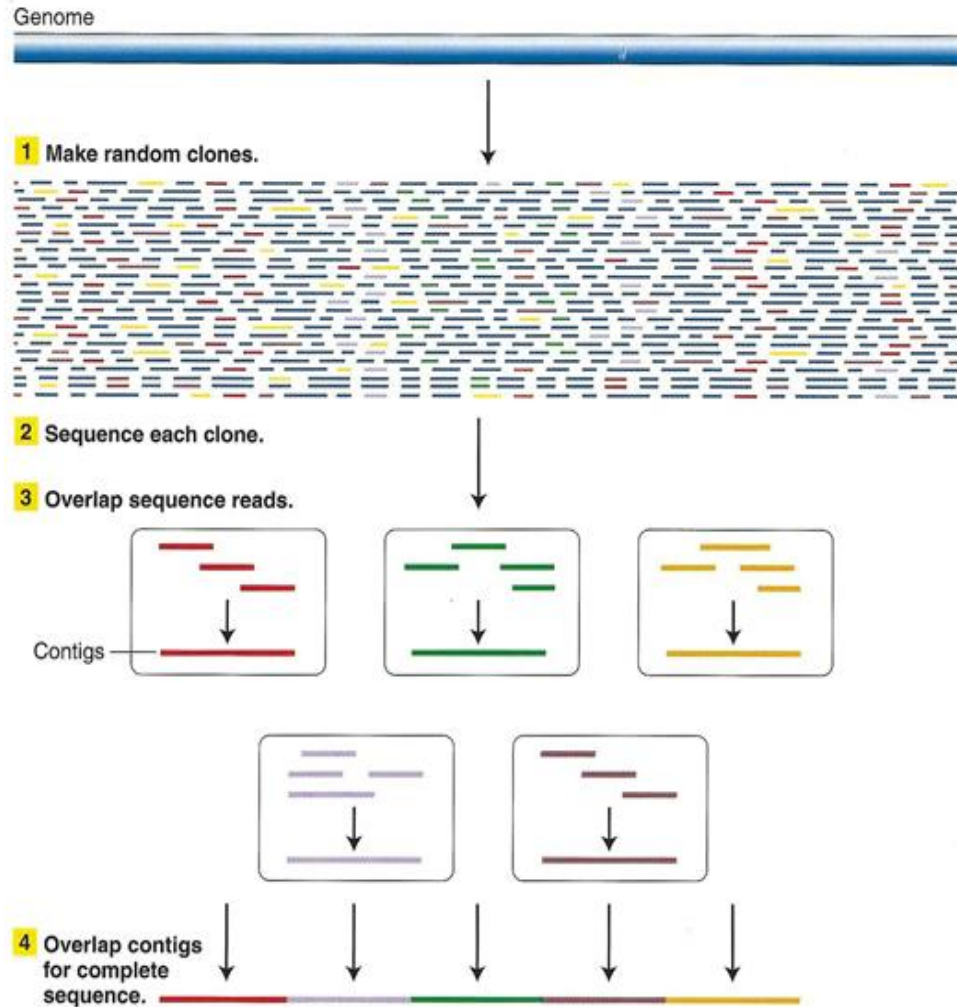


## 逐个克隆测序策略 Hierarchical Shotgun Sequencing

- 1 Order large-insert clones by overlapping fingerprints to create a physical map.



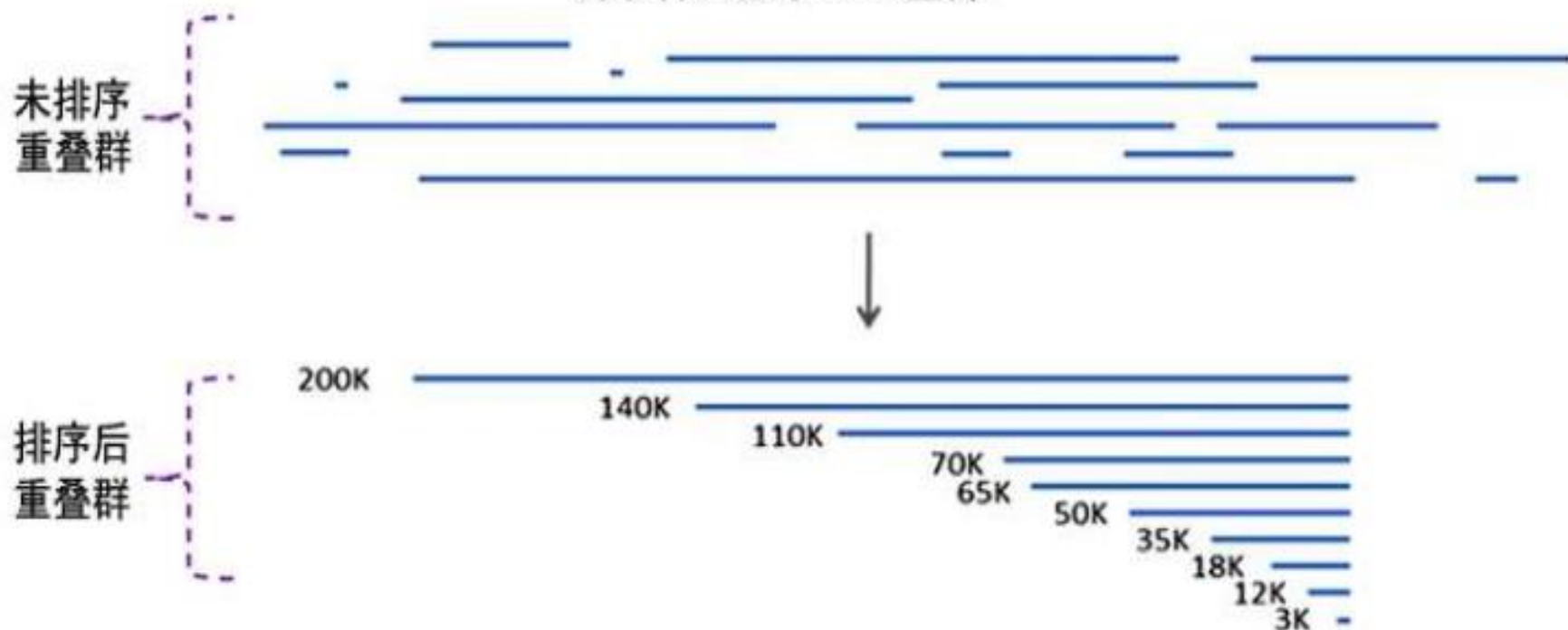
## 全基因组鸟枪法测序策略 Whole-genome Shotgun Sequencing



- 以BAC克隆(100kb)文库为工作单元;
- 图谱依赖性

- 以全基因组(3Gb)为工作单元;
- 图谱非依赖性

**N50 (可信的组装测序序列)**：把contig或scaffold从大到小排序，并对其长度进行累加，当累加长度达到总序列长度一半时，最后一个contig或scaffold长度。Contig大于N50的重叠群可被入选，进入下一步的组装。



叠加后重叠群总长 =  $200K + 140K + 110K + 70K + 65K + 50K + 35K + 18K + 12K + 3K = 703K$

50%重叠群总长 =  $703K \times 50\% = 351.5K$

依次叠加到351.5kb时最后一个重叠群成员：从高到低叠加，到110kb时达到50%总长

$\therefore 200K + 140K + 110K > 351.5K$ ， $\therefore N50 = 110K$ ，110kb重叠群处于最后一个，即为中位N50



- **测序深度 (sequencing depth)** : 实际测序得到的碱基总量与基因组大小的比值, 它是评价测序质量的重要指标。
- 较低的测序深度会导致拼装序列出现大段空缺和较高的错误率。

P: 基因组中某一碱基未被测序的概率;

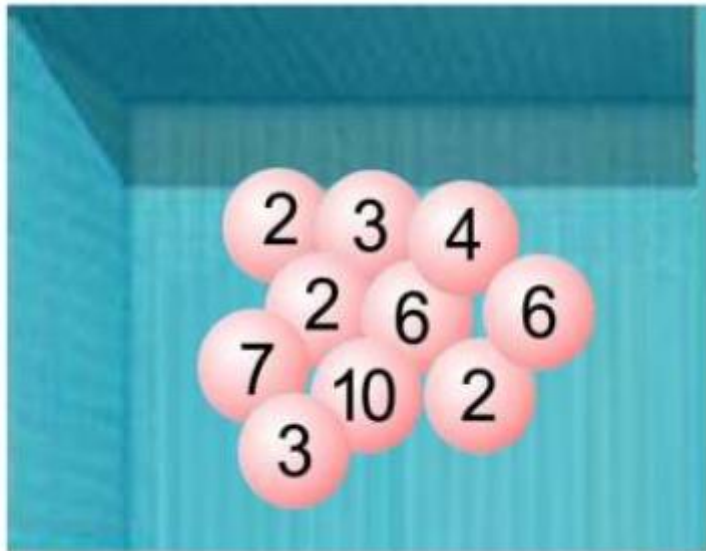
$$P = e^{-m}$$

e: 自然底数;

m: 测序深度;

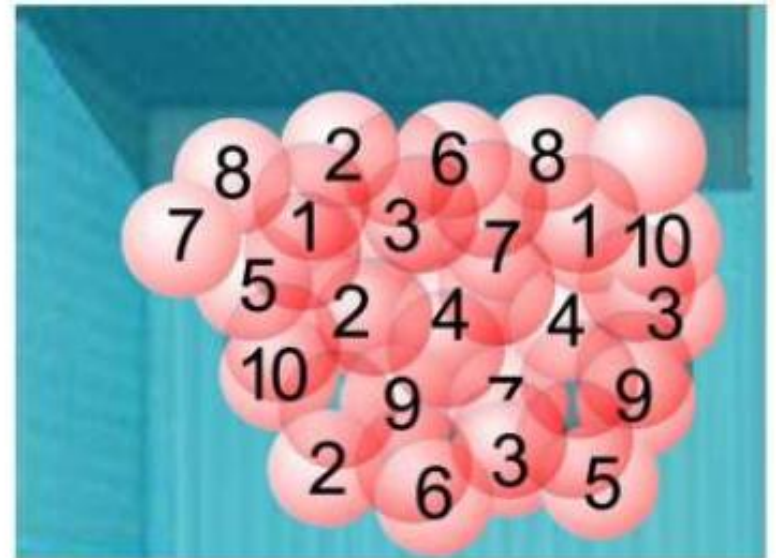
m=1, P=36.8%; m=2, P=13.5%; m=5, P=0.67%

重复取20次



从未拿到的小球占37%

重复取50次



从未拿到的小球占0.67%



## 2000.6 人类基因组草图完成

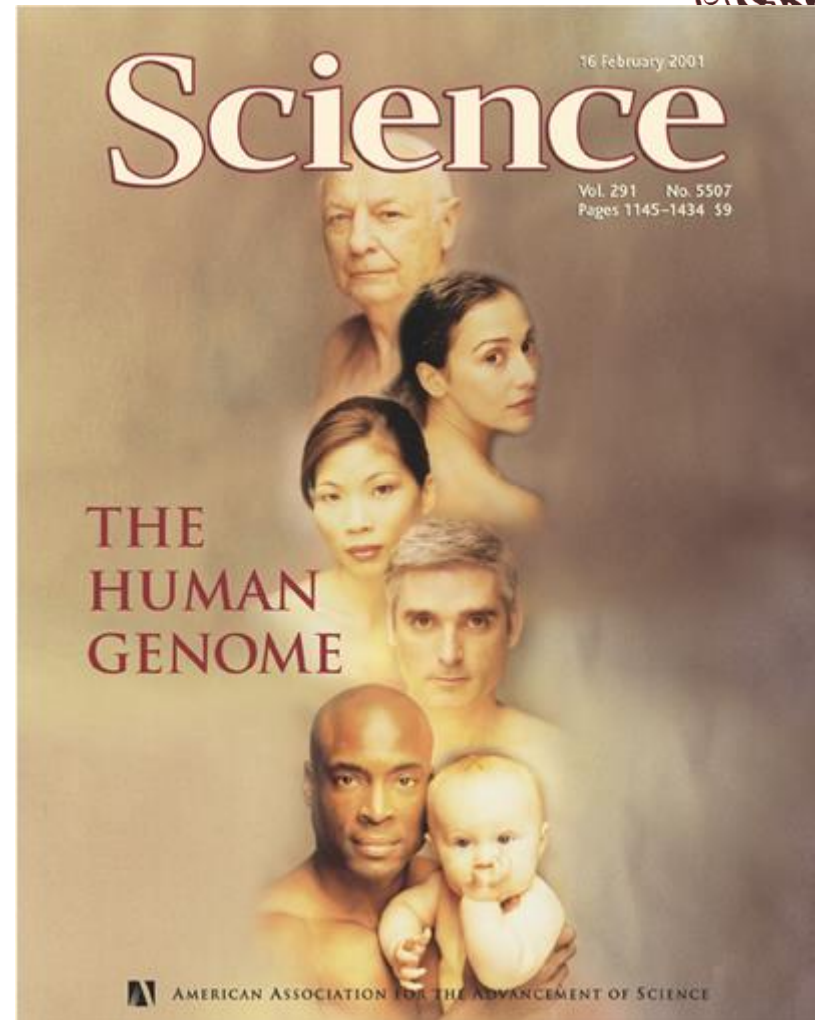
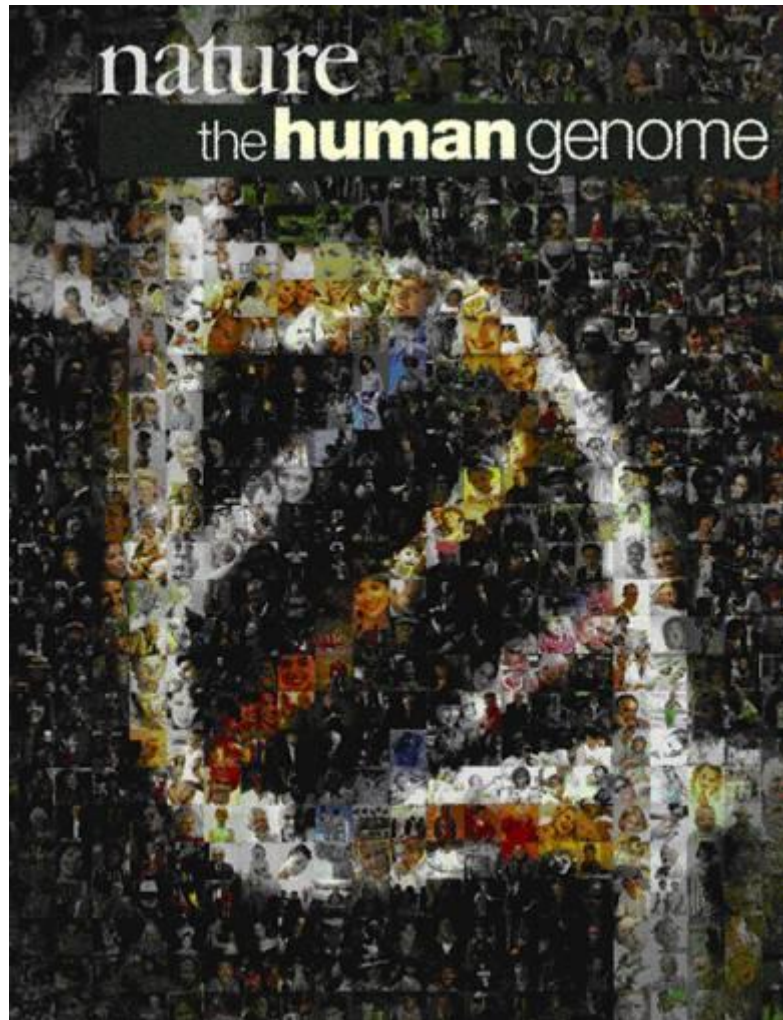
- 2000.6, 美国总统Clinton在白宫宣布人类基因组计划草图（工作框架图）提前完成；
- 2001.2, HUGO与Celera公司分别在《Nature》和《Science》上同时公布了人类基因组草图图谱（覆盖基因组90%以上的序列，错误率<1%）；



## 2003.4 人类基因组精细图完成

- 2003.4, HUGO首席科学家Collins宣布人类基因组精图绘制完成；
- 2004.10, HUGO在《Nature》上公布了人类基因组精图图谱（覆盖基因组99%以上的序列，仅存在341处空隙，错误率<0.001%）。

# The Human Genome Project



**Public HGP**

**Celera Genomics**

**February 2001: Completion of the Draft Human Genome**



## 框 1.26

### 人类基因组计划宣言

——六国政府首脑关于人类基因组序列图完成的联合宣言

(2003 年 4 月 14 日)




我们，美国、英国、日本、法国、德国与中国的政府首脑，骄傲地向全世界宣布：我们六国的科学家已完成了人类生命的分子指南——由 30 亿个碱基对组成的人类基因组 DNA 的关键序列图。


人类“生命天书”全部章节的解读，适逢 DNA 双螺旋结构发表 50 周年。50 年前的这个月，Watson 与 Crick 这一里程碑的发现，使基因研究与生物技术取得了举世瞩目的进展；50 年后的这一天，“国际人类基因组测序协作组”公布了人类基因组序列信息，全世界都可以通过国际互联网从公共数据库中自由分享，免费使用而不受任何限制。

人类基因组是全人类的共同财富和遗产。人类基因组序列图不仅奠定了人类认识自我的基石，推动了生命与医学科学的革命性进展，而且为全人类的健康带来了福音，使我们向着



# 数据查询网址:

**UNIVERSITY OF CALIFORNIA  
SANTA CRUZ** Genomics Institute


**UCSC**


**Genome Browser Gateway**


Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help About Us


**Browse/Select Species**


**POPULAR SPECIES**


  
Human


  
Mouse

  
Rat


  
Zebrafish

  
Fruitfly

  
Worm

  
Yeast

**REPRESENTED SPECIES**



**Find Position**

Human Assembly  
Dec. 2013 (GRCh38/hg38)

Current position: chr1:11,102,837-11,267,747

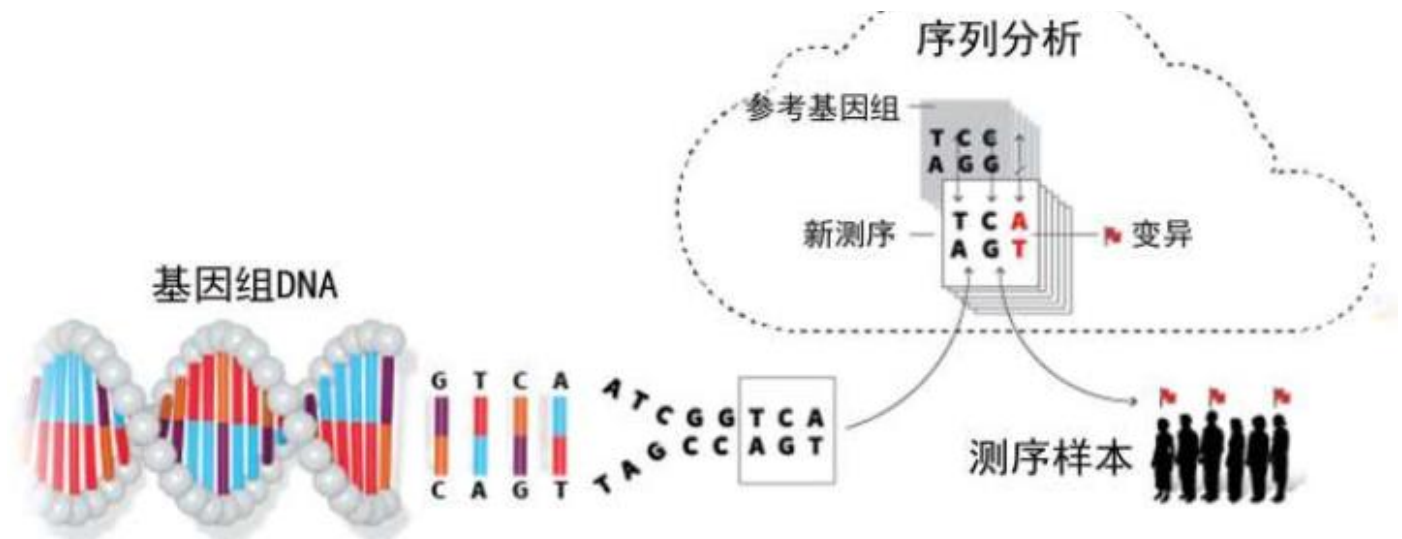
GO

**Human Genome Browser - hg38 assembly**

UCSC Genome Browser assembly ID: hg38  
Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38.p12 (GCA\_000001405.27)  
Assembly date: Dec. 2013 initial release; Dec. 2017 patch release 12  
Assembly accession: [GCA\\_000001405.27](#)  
NCBI Genome ID: 51 (Homo sapiens (human))  
NCBI Assembly ID: 5800238 (GRCh38.p12, GCA\_000001405.27)  
BioProject ID: PRJNA31257

<https://genome-asia.ucsc.edu/cgi-bin/hgGateway?redirect=manual&source=genome.ucsc.edu>

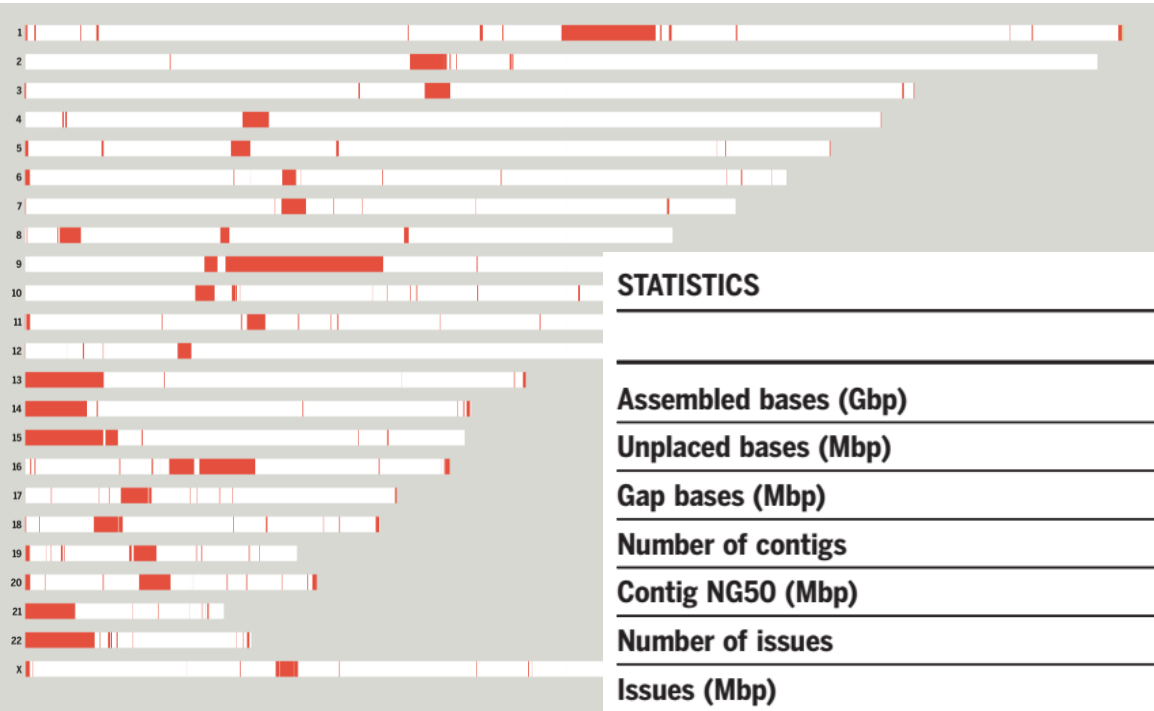
- **参考基因组 (Reference Genome)**是一个物种基因组的标准参考序列，并不代表某个确定的个体基因组。
- 人类参考基因组版本GRCh37来自纽约Buffalo市的13个自愿的匿名个体提供的基因组序列组装而成。人体血液有ABO三种血型，但在参考基因组中只含有一种基因型，即O型等位基因的序列。
- 参考基因组联盟 (Genome Reference Consortium)进行日常维持和改进，2013年12月24日公布了GCCh38.
- 因为参考基因组最接近真实基因组的序列，所以可用来指导新测序获得的人类基因组序列的组装。



# 2022.4 人类基因组完整图发布



## Telomere-to-Telomere (T2T) Consortium



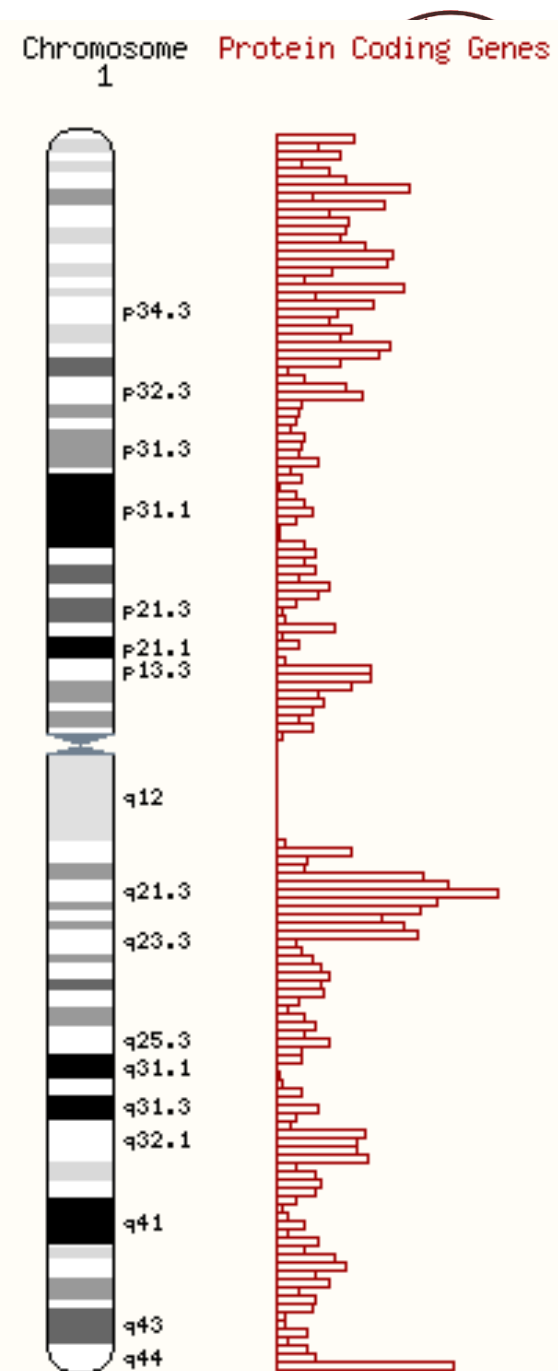
STATISTICS	GRCH38	T2T-CHM13	DIFFERENCE (±%)
Summary			
Assembled bases (Gbp)	2.92	3.05	+4.5
Unplaced bases (Mbp)	11.42	0	-100.0
Gap bases (Mbp)	120.31	0	-100.0
Number of contigs	949	24	-97.5
Contig NG50 (Mbp)	56.41	154.26	+173.5
Number of issues	230	46	-80.0
Issues (Mbp)	230.43	8.18	-96.5
Gene annotation			
Number of genes	60,090	63,494	+5.7
Protein coding	19,890	19,969	+0.4
Number of exclusive genes	263	3,604	
Protein coding	63	140	
Number of transcripts	228,597	233,615	+2.2
Protein coding	84,277	86,245	+2.3
Number of exclusive transcripts	1,708	6,693	
Protein coding	829	2,780	

*The complete sequence of a human genome. Science, 2022*



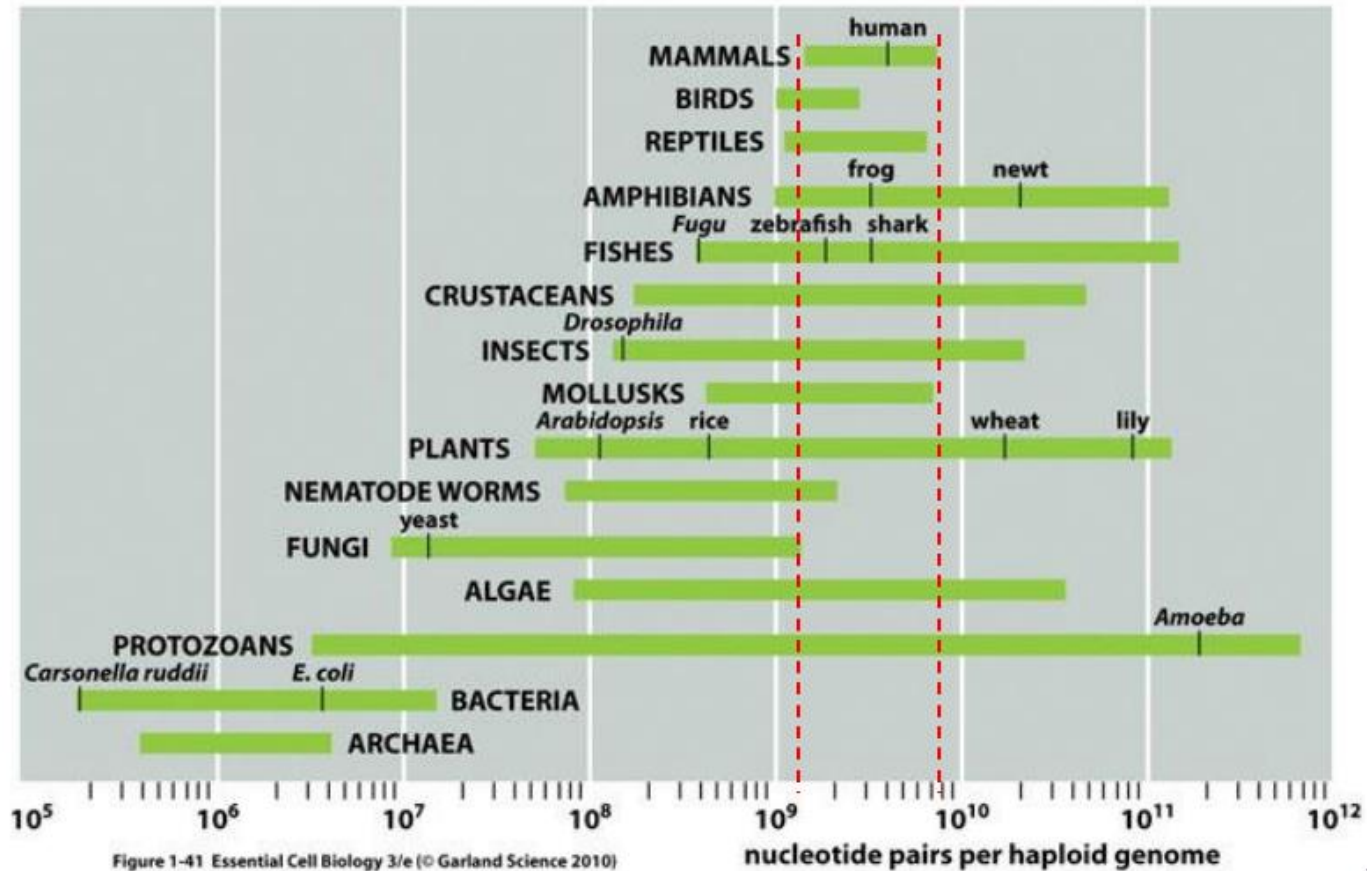
### 三、人类基因组的结构特点

- 核基因组DNA的总长约 $3.2 \times 10^9$ bp;
- 19, 831个蛋白编码基因, 总序列长度约48Mb, 占总基因组序列的1.5%;
- 人类基因平均含有4个外显子, cDNA长度平均为1350bp, 编码450个氨基酸;
- 平均基因密度为5.96个基因/Mb, 但在染色体上的分布并不均匀; 整个基因组的 20% 为基因 “沙漠” 区;
- 罕见重叠基因和多顺反子转录单位;
- 50%以上的区域含重复序列。人类和灵长类基因组最具特征性的重复序列是Alu家族, 在单倍体基因组中重复近100万次。



# 人类基因组组成(1) ---- 基因组大小及GC含量

Genome Size: 是生物体基因组最重要的生物学特点之一。

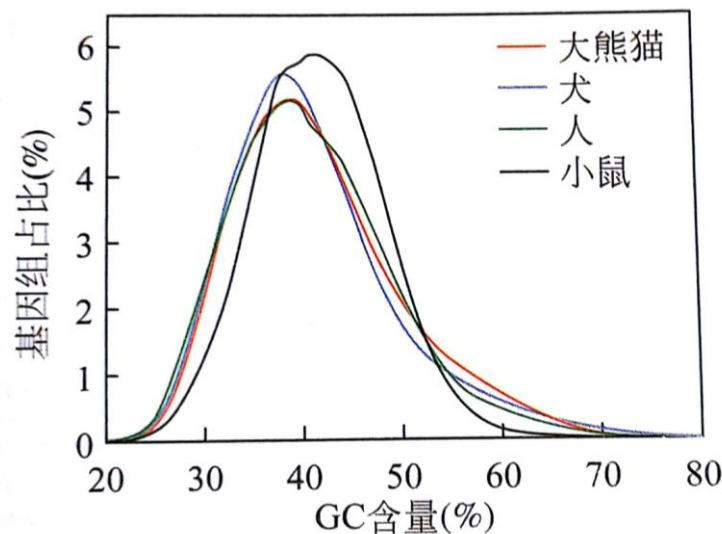


**C值:** 是指一个物种单倍体基因组中DNA的总量。

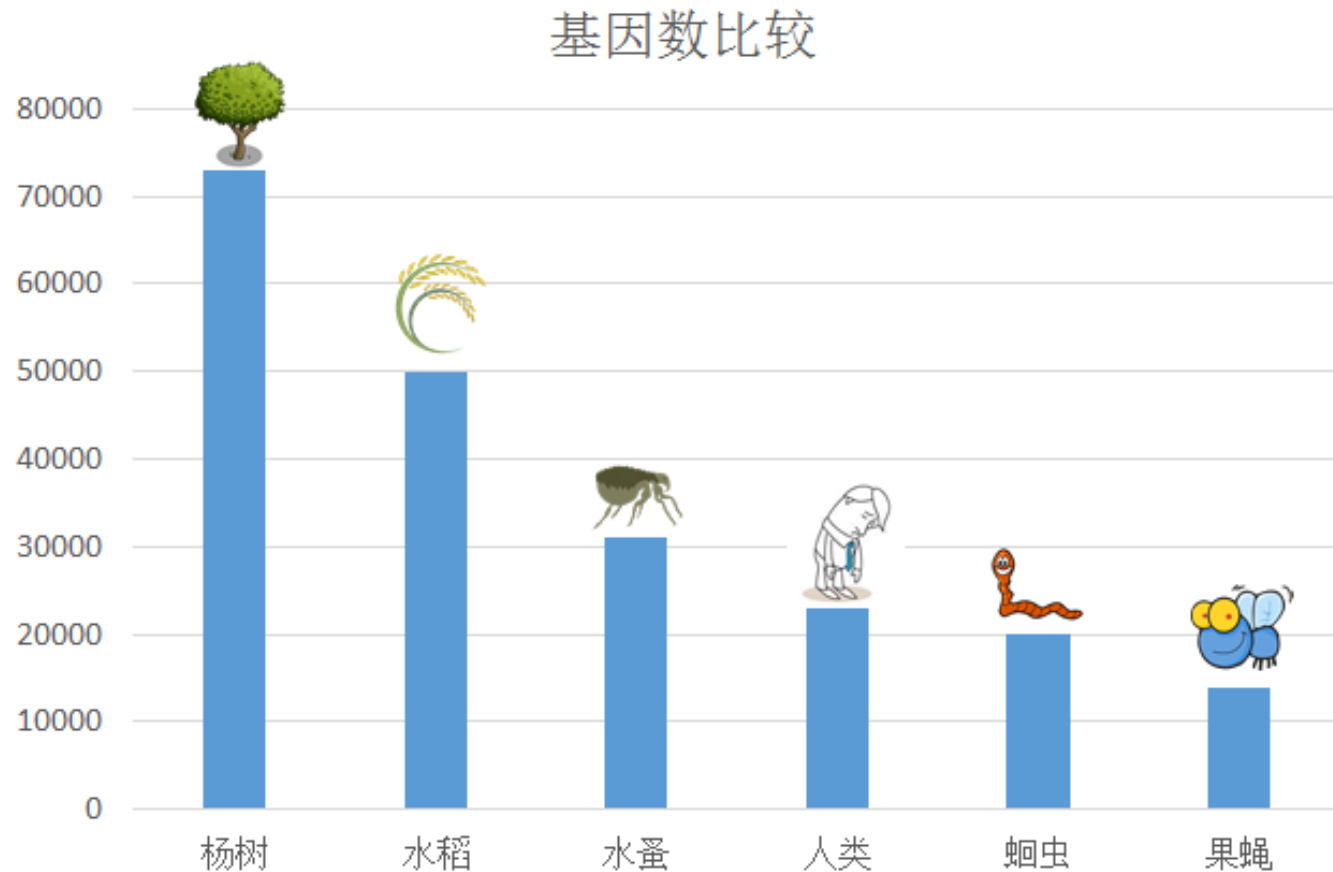
**C值悖论 (C value paradox):** 物种的C值和它的进化复杂性之间无严格对应关系的现象称为C 值悖理，是复杂生物基因组的一个普遍特征。

- GC含量是物种演化的特征之一。不同物种基因组序列之间的GC含量相差很大，近缘物种的GC分布有相似的趋势；
- GC配对有着较高的热稳定性，蛋白编码序列的GC含量较高，而非编码序列的GC含量较低，成为基因注释软件算法的参考因素之一；
- 95%的CpG islands的GC含量为60-70%，长度为0.3-3kb, 与基因组序列甲基化及基因表达有关,是表观基因组研究的重要内容。

恶性疟原虫的GC含量为19.3%，是已知GC含量最低的物种。细菌为67.7%，酿酒酵母为38%，而人为42%左右。

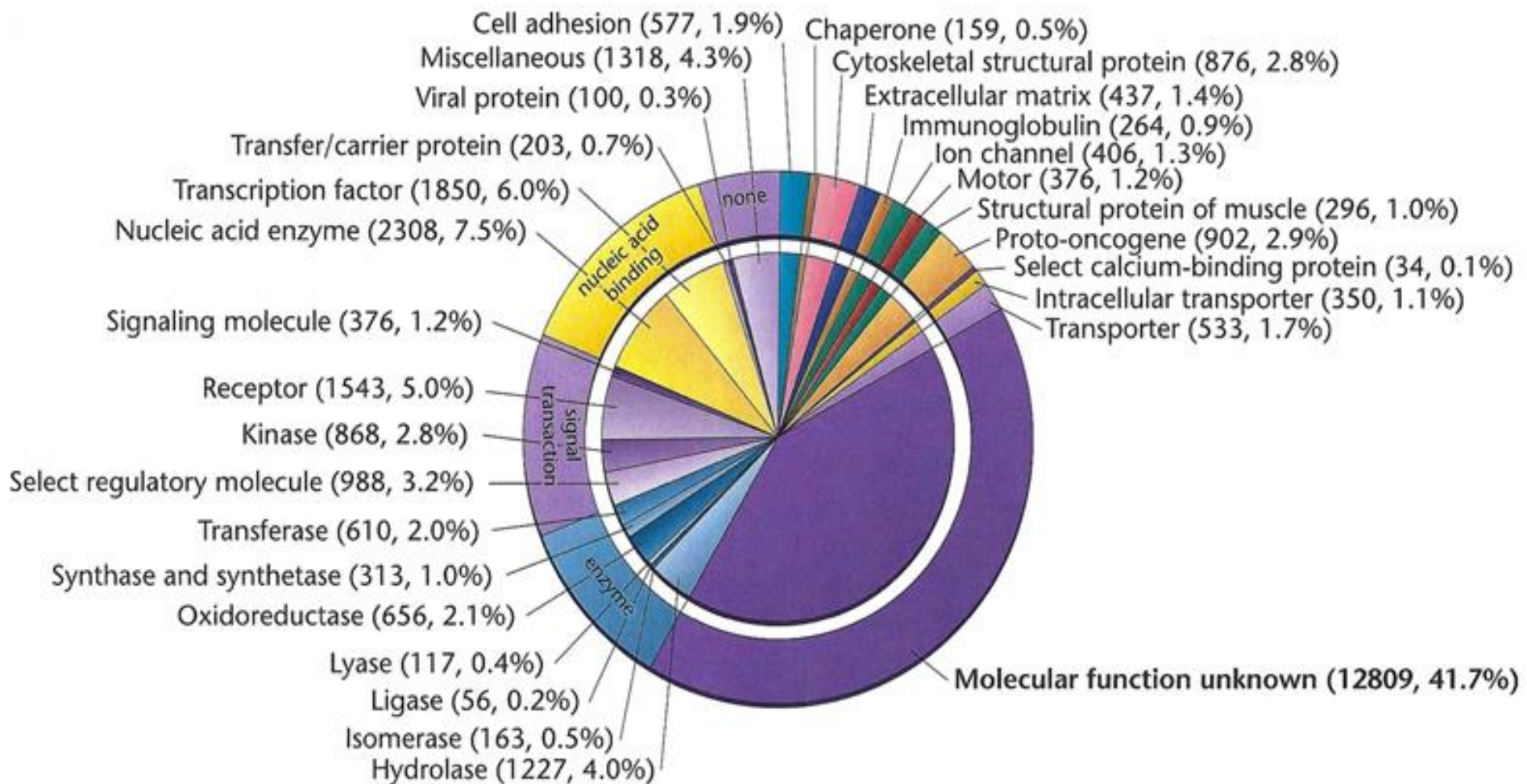


# 人类基因组组成(2) ---- 基因及基因相关序列



**N值：**是指生物体所含有的基因数目。

**N值悖论 (N value paradox)：** 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。

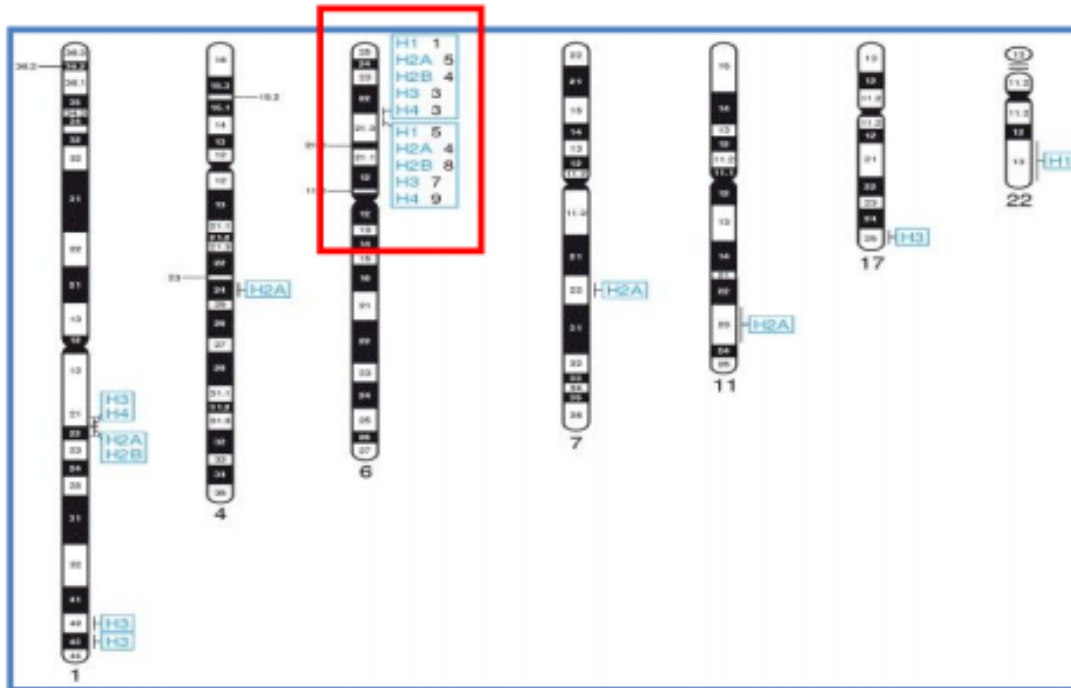


- 基因总长、外显子的大小均与基因产物的大小无关。但表达丰度高的管家基因一般都倾向于含有较小的内含子；



# 基因座与基因簇

- 基因座 (locus) 指基因在染色体上所处的位置。
- 基因簇 (gene cluster): 由一些序列和功能高度一致的基因聚集在染色体的相同位置, 紧密连锁而构成。如人的核糖体RNA基因、组蛋白基因等。



- 人类五种组蛋白的编码基因在基因组中均有多个基因拷贝, 它们分布在7条染色体上, 序列高度保守。
- 但在6号染色体短臂, 组蛋白基因的多个拷贝串联分布, 形成了典型的基因簇结构。



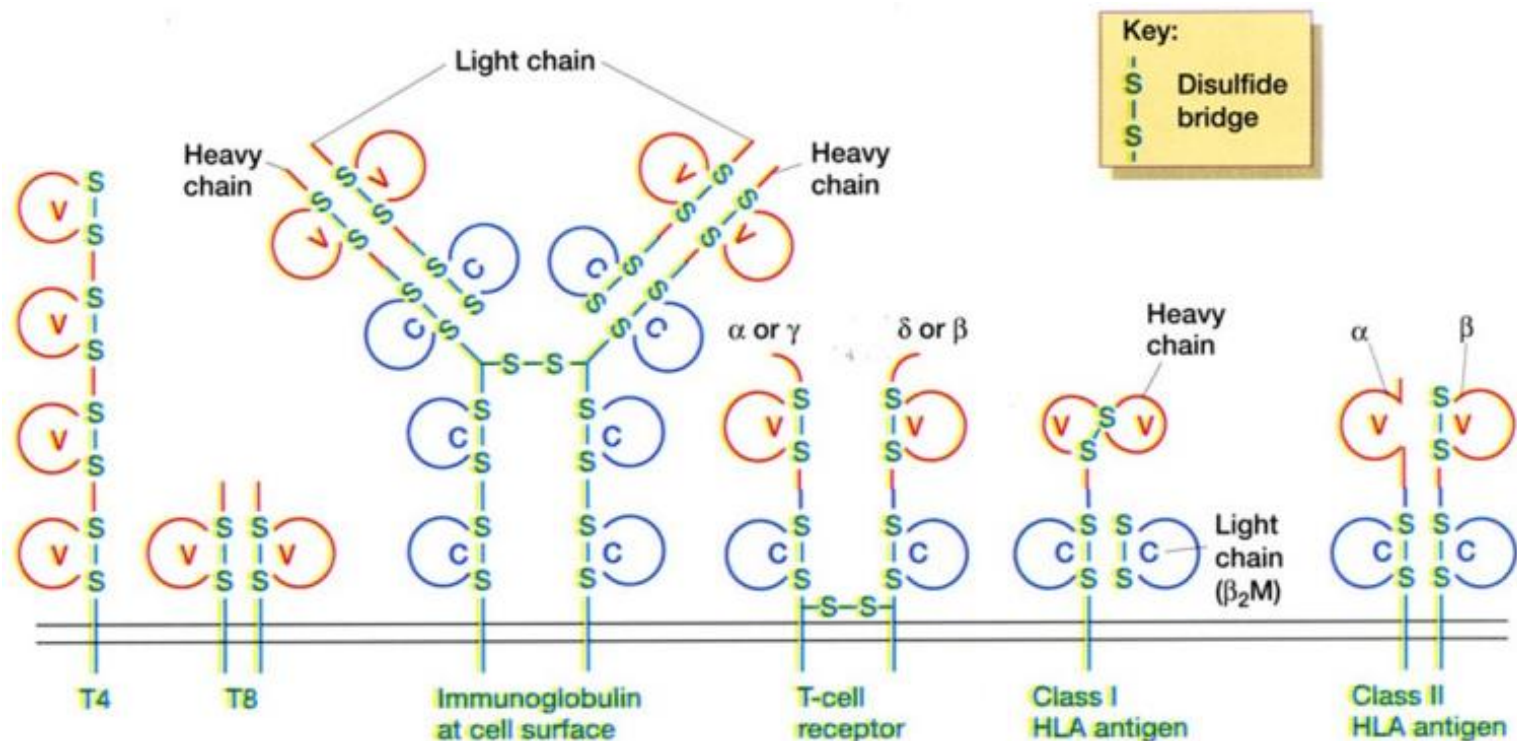
# 基因家族

- **基因家族 (gene family):** 包括多个基因，这些成员的**序列高度同源**，能编码保守的蛋白质结构域或氨基酸基序；  
家族成员在进化上具有共同祖先，**功能相似或相近**；  
基因家族的成员有的以**基因簇**的形式存在，如珠蛋白基因家族；  
有的**散在分布**在多条染色体上，如醛缩酶基因家族。



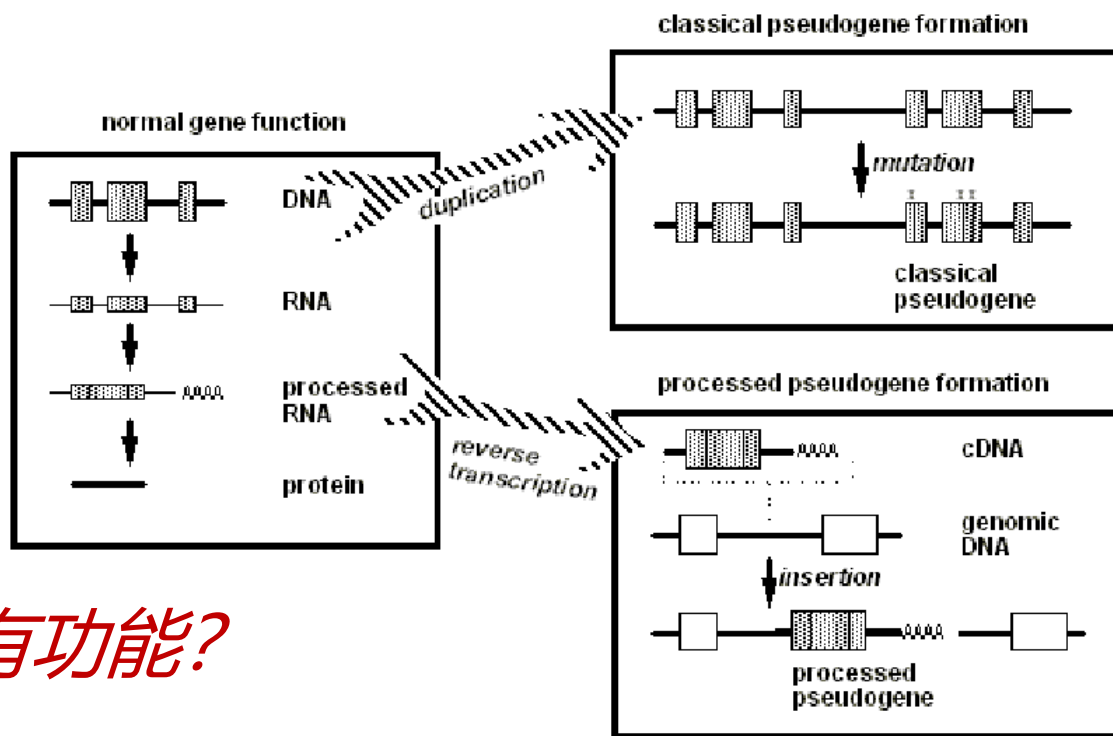
- **基因超家族 (gene superfamily):** 一些基因的**序列同源性低**，基因产物没有保守的蛋白质功能域或氨基酸基序，但**功能相关且具有相同的特征结构**。

如：免疫球蛋白基因超家族，尽管基因序列之间的同源性很低，但基因产物都与免疫应答有关且具有和免疫球蛋白相似的结构特征。这类基因的进化亲缘关系较远。



**假基因 (pseudogene)** 与基因组中有功能的基因具有相似的序列，但失去蛋白质编码功能或不能正常转录表达的DNA序列。

- **常规假基因 (classical/conventional pseudogene)**: 在基因组进化过程中功能基因复制后产生突变的失活产物。
- **加工假基因 (processed pseudogene)**: 功能基因的mRNA转录产物反转录为cDNA后再次插入基因组，形成一个新的基因拷贝，亦称为反转座假基因。基因组中的Alu, LINE1是丰度最高的两种加工假基因。



**假基因有没有功能?**



## 人类基因组组成(3) ---- 基因外DNA

- 人类基因组中基因与基因相关序列仅占约25%，在剩余75%的基因外DNA中，存在大量的非编码重复序列；
- 按重复程度的不同将重复序列分为低度重复序列( $<10$ 拷贝), 中度重复序列( $10-10^3$ 拷贝)和高度重复序列( $> 10^4$ 拷贝)；
- 按重复序列在染色体上的分布可分为串联重复序列和散在重复序列；

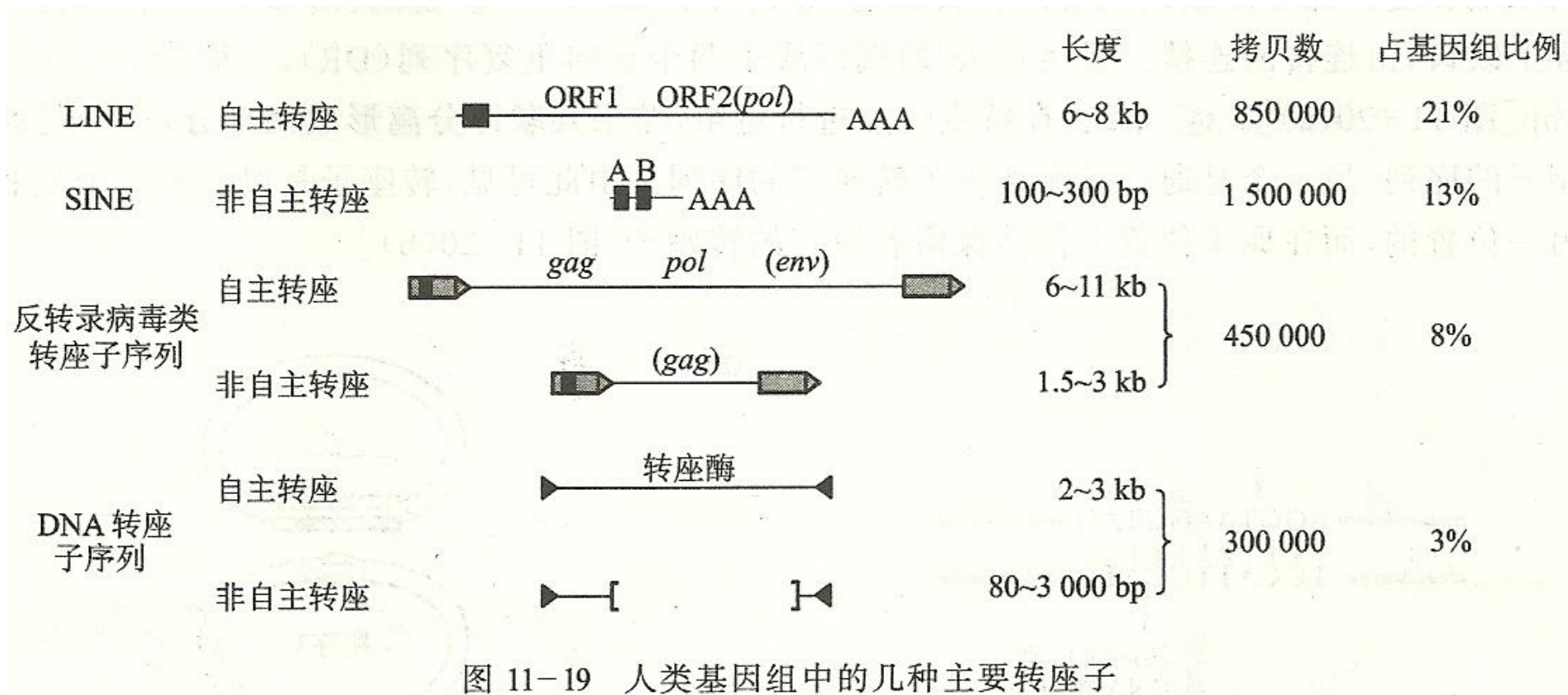


**串联重复序列：**包括**大卫星DNA (megasatellite DNA)**，**卫星DNA (satellite DNA)**，**小卫星DNA (minisatellite DNA)**和**微卫星DNA (microsatellite DNA)**

类别	重复序列大小	主要染色体定位
大卫星 DNA	几 kb	特定染色体上的多个位置
RS447	4.7 kb	4p15, 有 50 ~ 70 拷贝, 8p 远端还有一些
未命名	2.5 kb	4q31 和 19q13, 约 400 拷贝
未命名	3.0 kb	X 染色体上, 约 50 拷贝
卫星 DNA	5 ~ 171 bp	主要着丝粒位置
$\alpha$ (alphoid DNA)	171 bp	全部染色体的着丝粒异染色质区
$\beta$ (Sau3 A 家族)	68 bp	1、9、13、14、15、21、22 和 Y 染色体的着丝粒异染色质区
Satellite 1 (富含 AT)	25 ~ 48 bp	大多数染色体的着丝粒异染色质区, 其他异染色质区
Satellite 2 和 3	5 bp	全部染色体
小卫星 DNA	6 ~ 64 bp	所有染色体的端粒和近端粒位置
端粒家族	6 bp	所有端粒
高度可变家族	9 ~ 64 bp	全部染色体, 常在近端粒区
微卫星 DNA	1 ~ 4 bp	全部染色体上散在分布

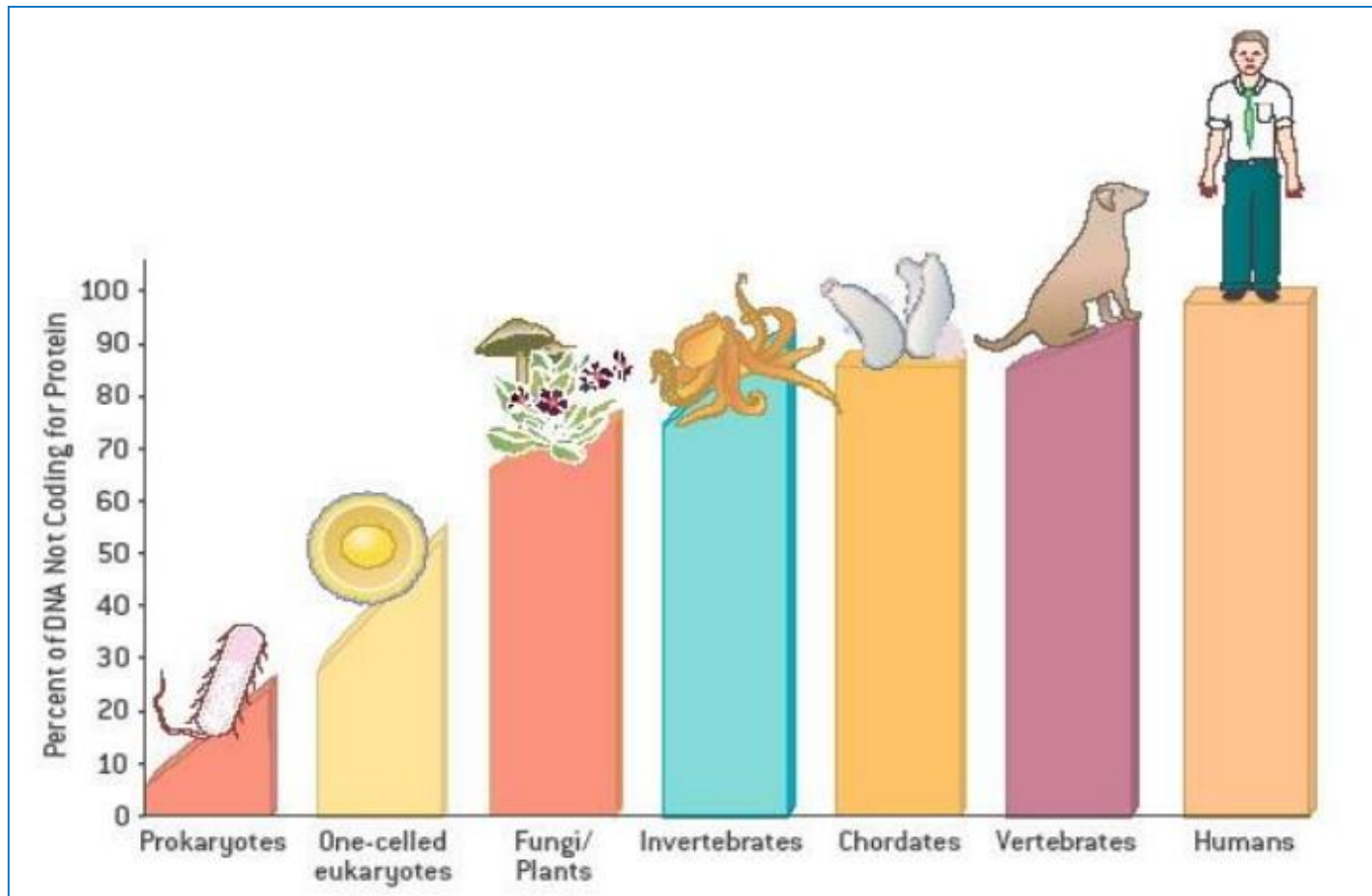


**散在重复序列：**主要包括**反转座元件**和**DNA转座子化石**，前者是真核生物基因组的特有组成部分。



## 人类基因组组成(4) ---- 非编码RNA

非编码序列占人类基因组的98.5%，远高于其他任何一种生物

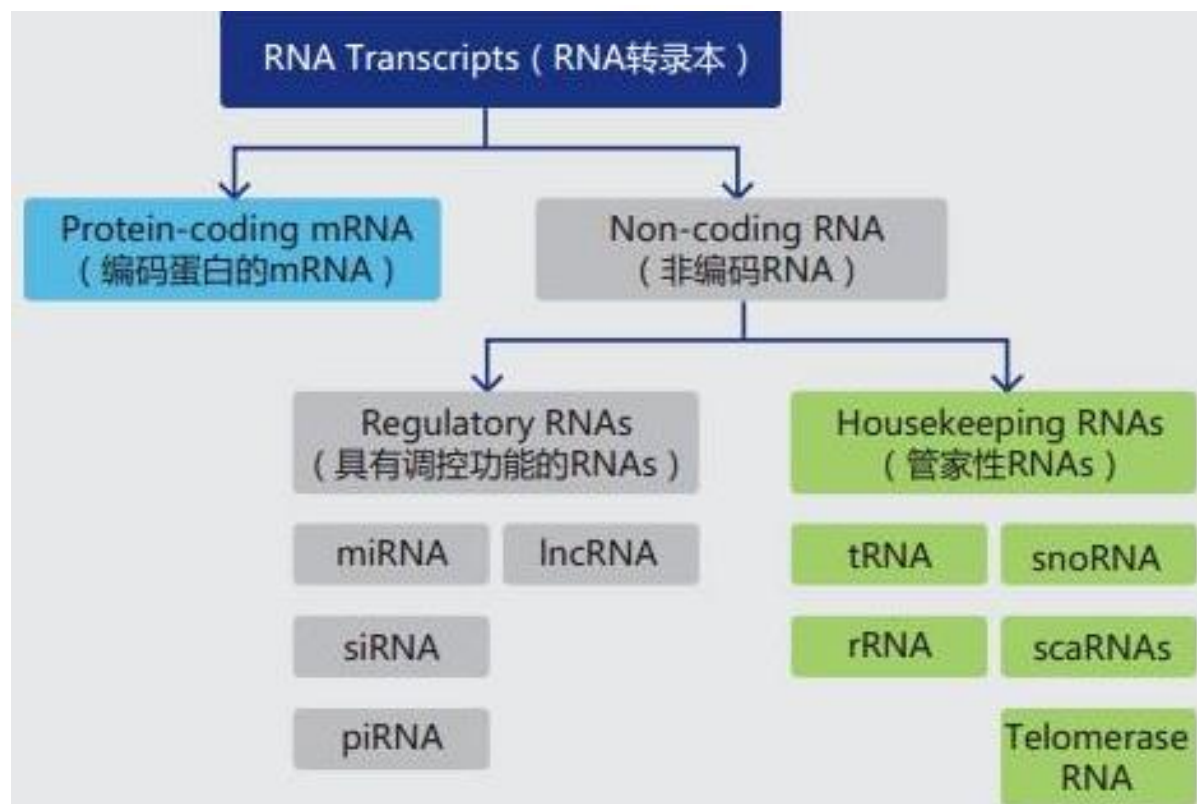


- **非编码RNA (non-coding RNA, ncRNA)** 指不具有蛋白质编码功能的RNA分子。人类基因组中ncRNA基因是功能DNA的重要补充；

表 2.13 非编码 RNA 的分类

长度 (nt)	RNA 种类
$\leq 50$	miRNA、siRNA、piRNA
50~500	rRNA、tRNA 等
$\geq 500$	lncRNA 等非编码 RNA、长的不带 polyA 尾巴的非编码 RNA 等

- ncRNA基因有的位于蛋白编码基因的内部（如内含子），有的位于编码基因的调控序列（如假基因），还有的位于基因间的非编码区。



- rRNA, 参与核糖体组装
- tRNA, 参与氨基酸转运
- snRNA (small nuclear RNA) 小核RNA, 参与内含子剪接
- snoRNA (small nucleolar RNA) 小核仁RNA, 参与rRNA加工
- miRNA (microRNA) 微小RNA; siRNA (small interfering RNA) 小干扰RNA等参与基因表达的转录后调控
- piRNA (PIWI-interacting RNA), 参与转座调控, 精子发生等
- lncRNA (long ncRNA), 参与转录及翻译后调控, 表观遗传修饰等

# 人类基因组的组成

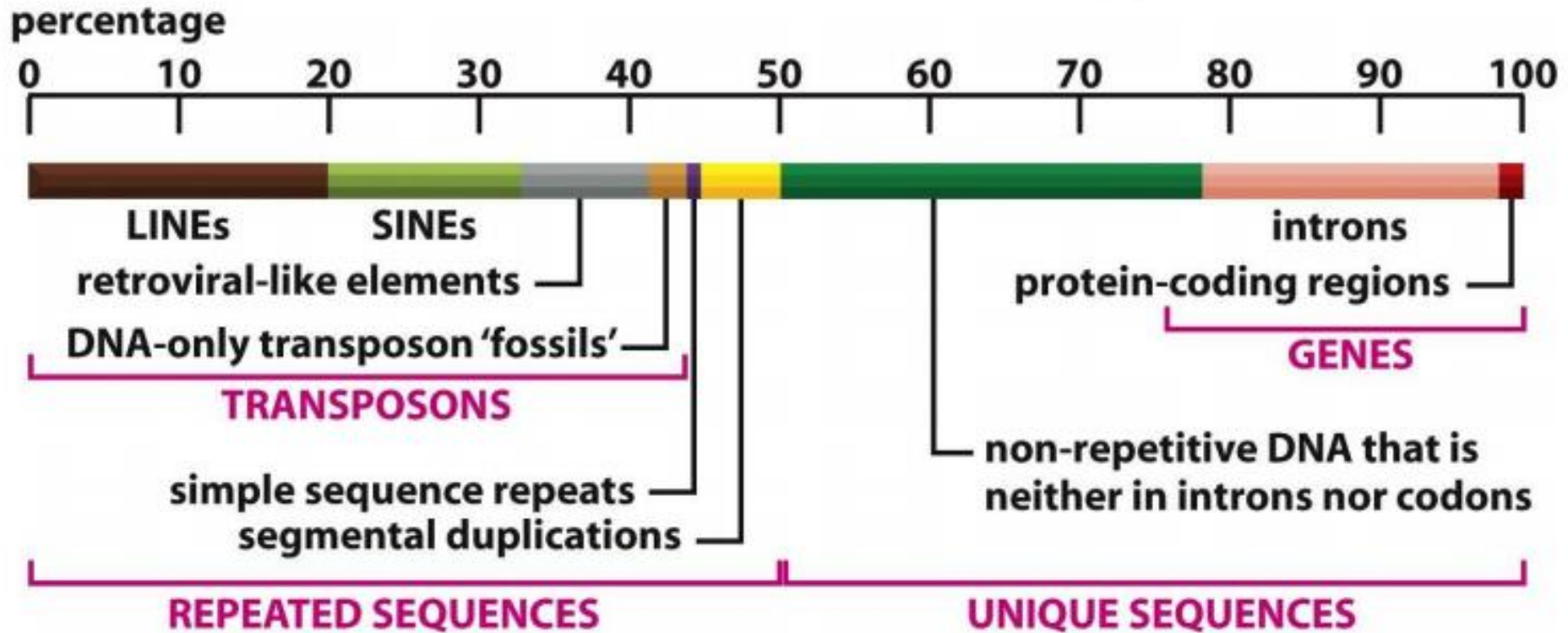


Figure 4-17 Molecular Biology of the Cell 5/e (© Garland Science 2008)