中国科学院北京基因组研究所（国家生物信息中心）
BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

国家基因组科学数据中心
National Genomics Data Center

第3章 基因组图谱与测序

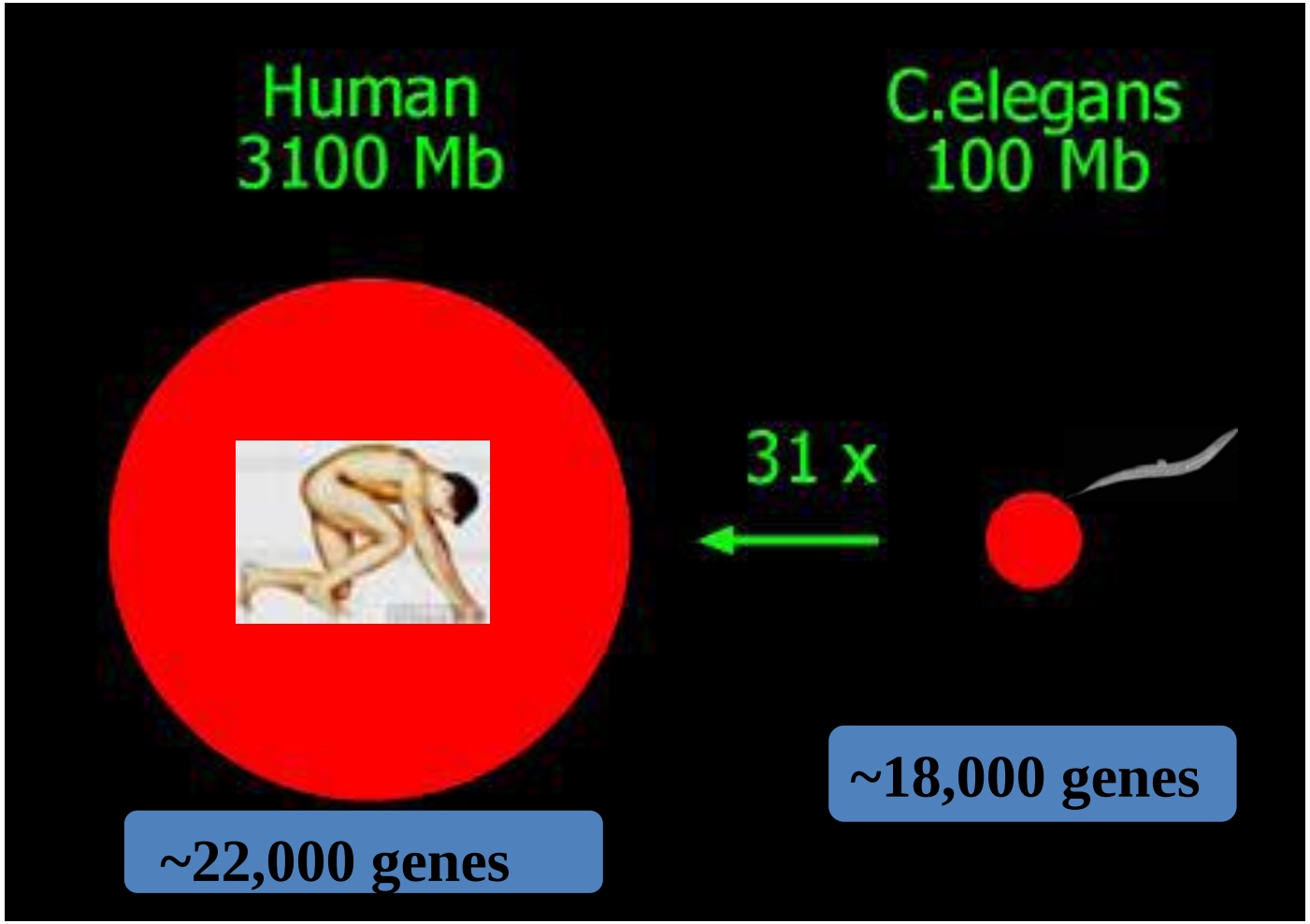
章张 博士

内容提纲

- 基因组遗传图谱
- 基因组物理图谱
- DNA测序技术
- 基因组测序策略
- 单细胞测序
- 发展趋势

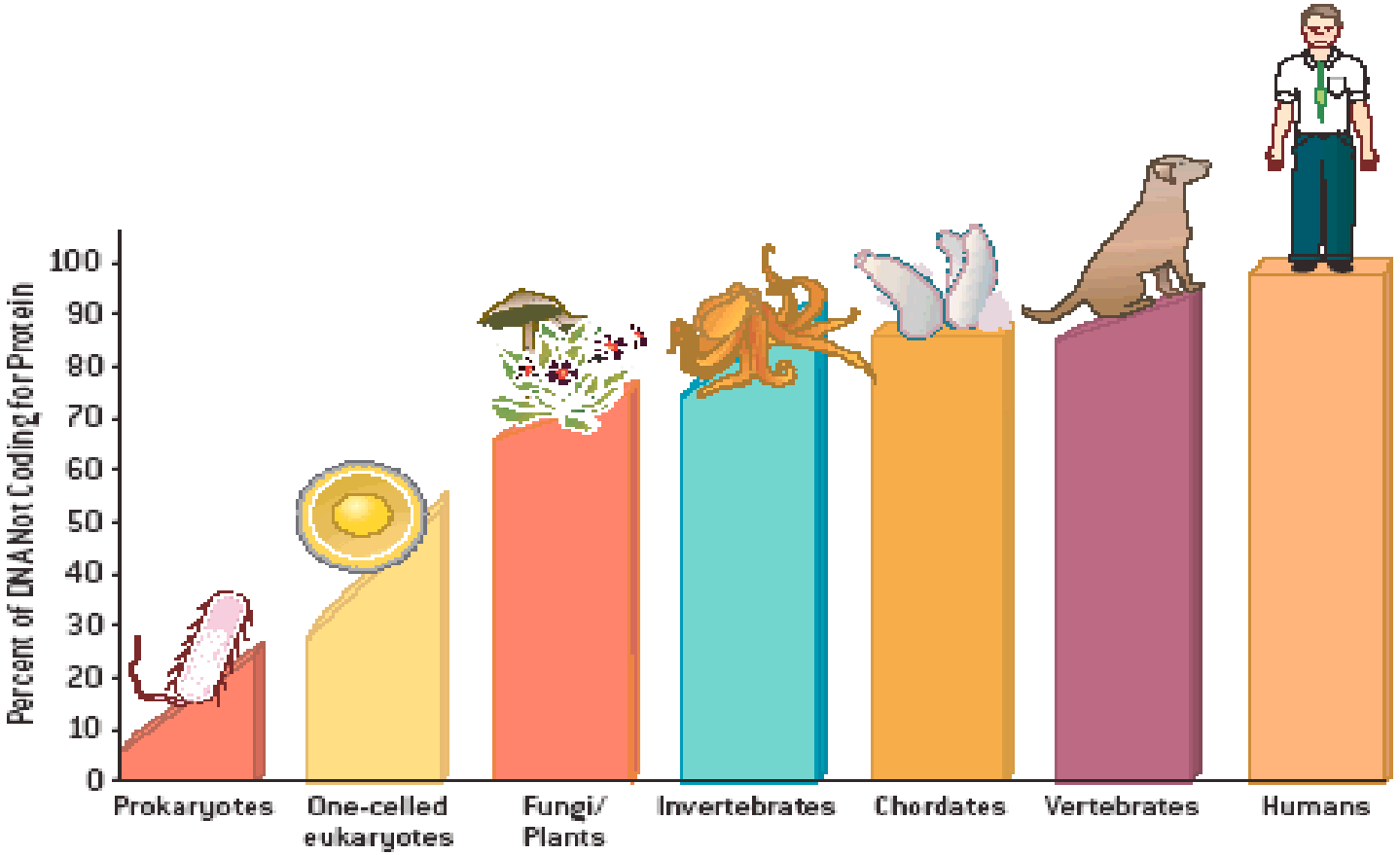
工欲善其事，必先利其器。——孔子《论语》

为什么要对不同的物种测序?



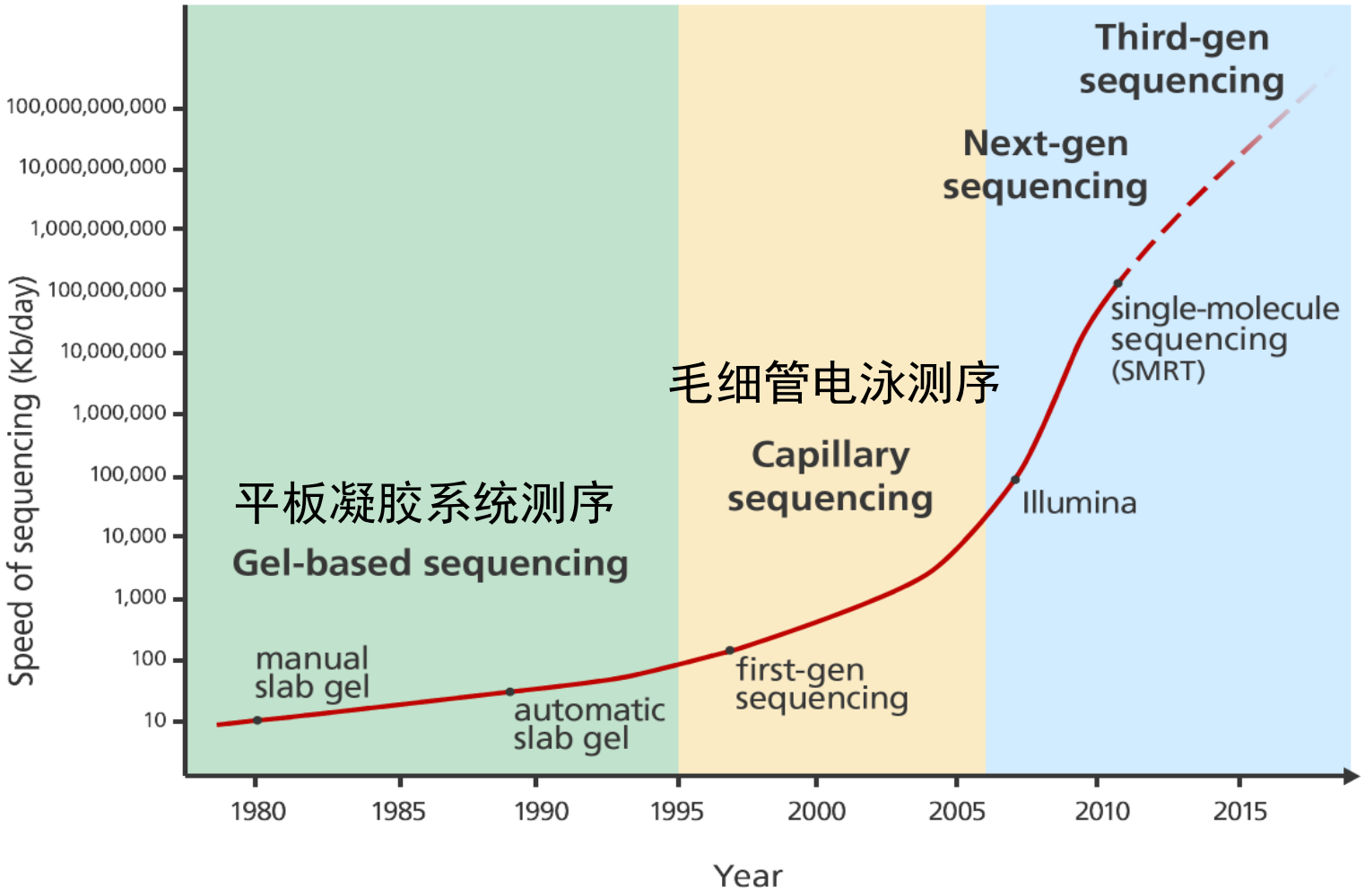
不同物种的基因组主要差别在哪里?

非编码序列



NONPROTEIN-CODING SEQUENCES make up only a small fraction of the DNA of prokaryotes. Among eukaryotes, as their complexity increases, generally so, too, does the proportion of their DNA that does not code for protein. The noncoding sequences have been considered junk, but perhaps it actually helps to explain organisms' complexity.

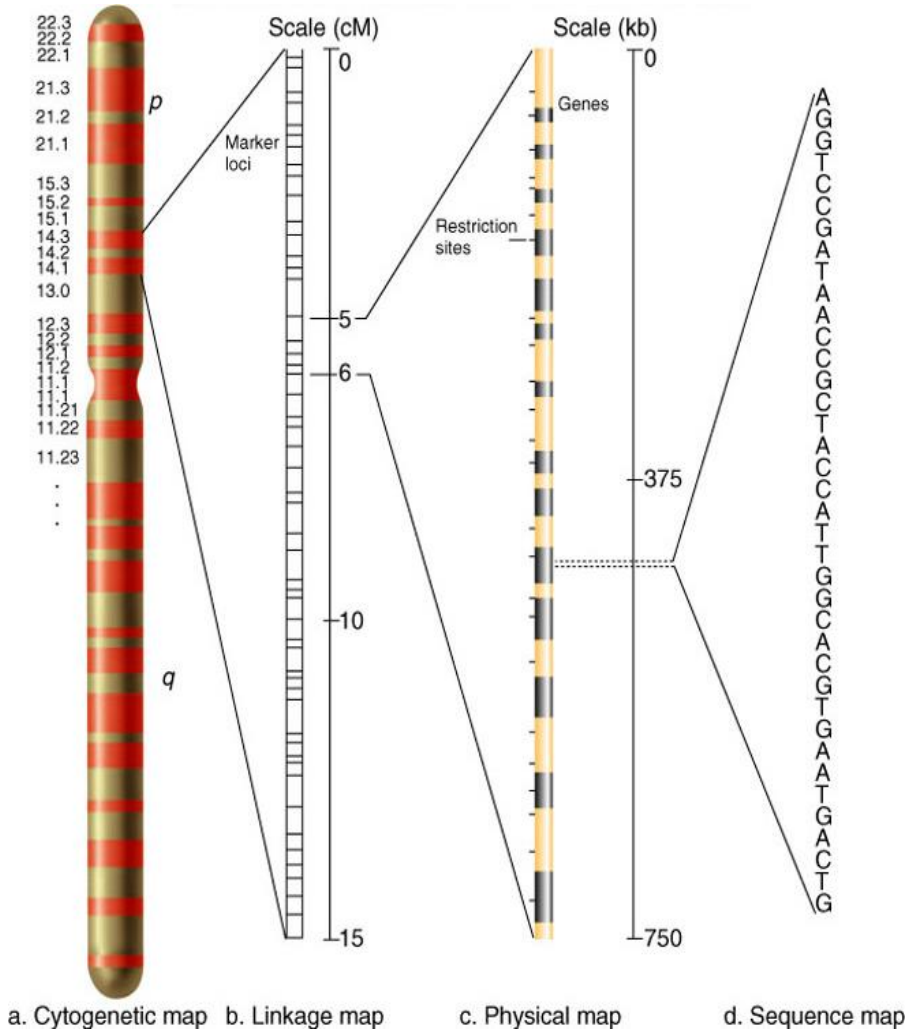
DNA测序技术发展



摘自Pareek CS等，2011

人类基因组计划主要成果：四张图

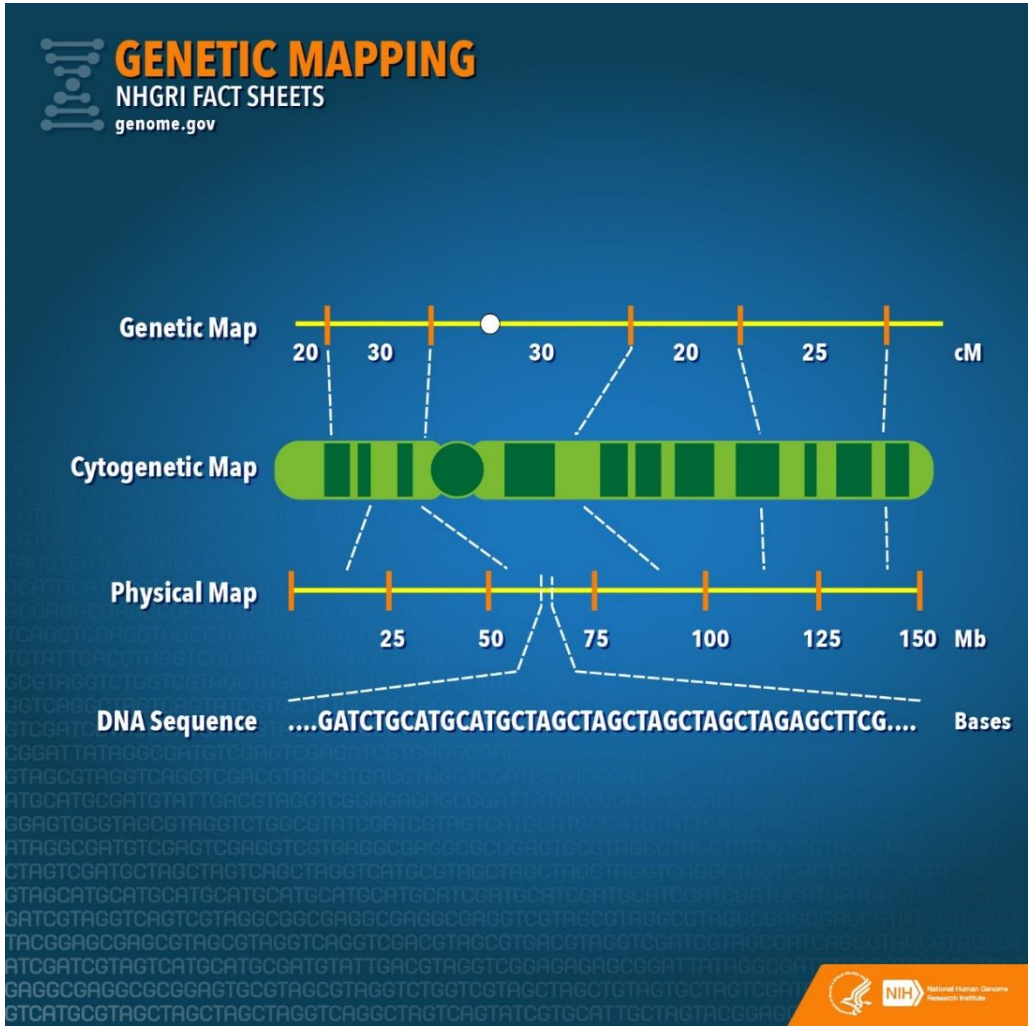
- 遗传图谱：又称为连锁图谱（linkage map），指基因或DNA标志在染色体上的相对位置与遗传距离
- 物理图谱：以定位的DNA标记序列如STS作为路标，以DNA实际长度即bp、kb、Mb为图距的基因组图谱
- 转录图谱：利用EST作为标记所构建的分子遗传图谱
- 序列图谱：通过基因组测序得到的，以A、T、G、C为标记单位的基因组DNA序列



遗传图谱和物理图谱

- 遗传图谱 (Genetic Map)：某一物种的染色体图谱（连锁图谱），显示**所知的基因和/或遗传标记**的**相对位置**。标记间距离（遗传图距）用减数分裂中的交换频率来表示，单位为**厘摩**（centi-Morgan, cM），每单位厘摩定义为1%交换率。

- 物理图谱 (Physical Map)：是DNA中一些**可识别的标记**(如限制性酶切位点、基因等)在DNA上的**物理位置**，图距是**物理长度**单位，如染色体的带区、核苷酸对的数量等。



https://www.genome.gov/genetics-glossary/Genetic-Map

遗传图谱

遗传图谱

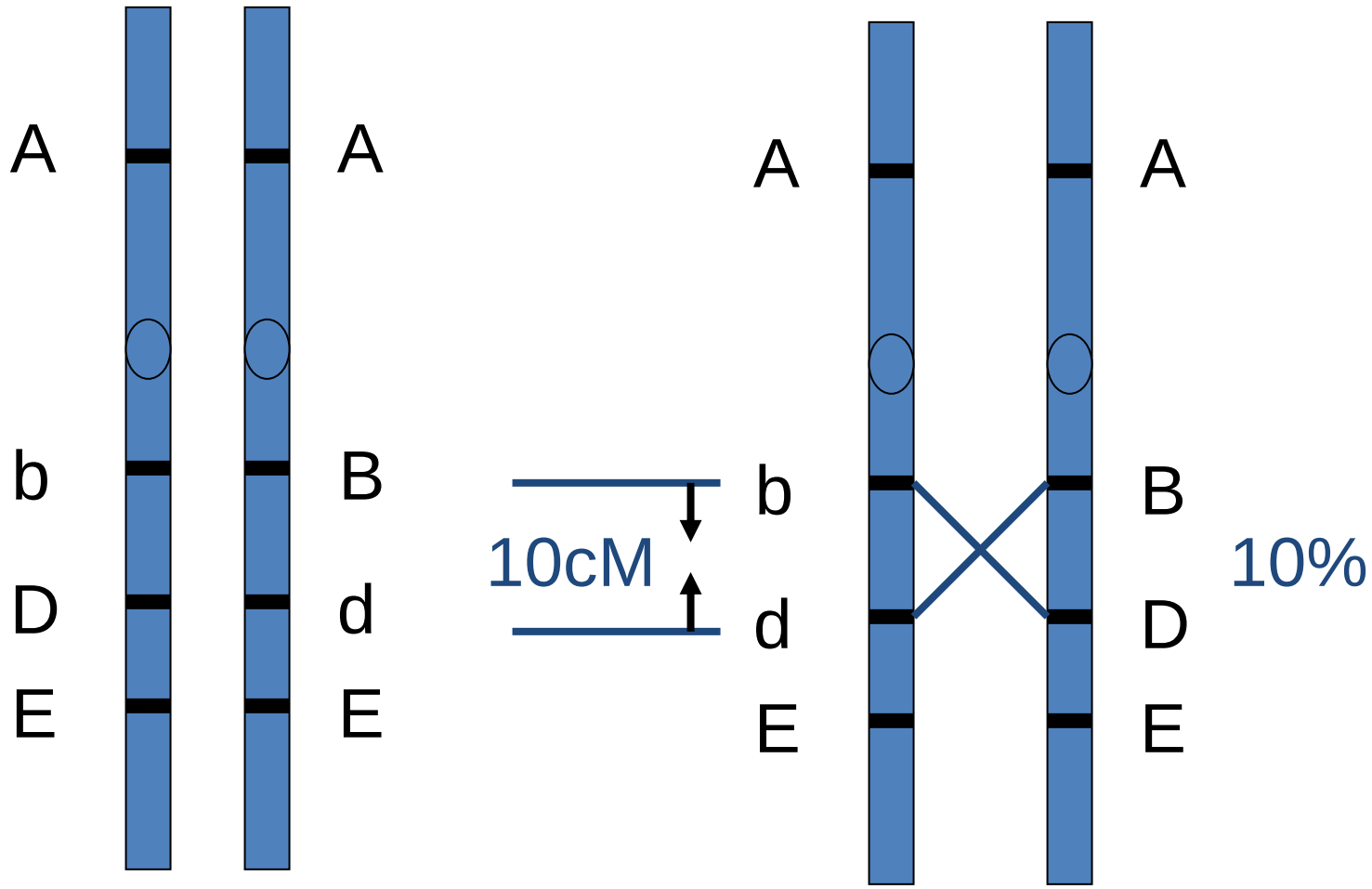
遗传图谱（genetic map）或连锁图谱（linkage map）

- 以在某个遗传位点上具有多个等位基因的遗传标记作为“路标”，以遗传学上的距离即两个遗传位点之间进行交换、重组的百分率cM作为“图距”，反映基因遗传效应的基因组图
- 根据重组频率来确定突变点之间的距离
- 通过测量基因组DNA位点间的重组来绘制

遗传图谱

- 遗传图谱是应用遗传学技术构建能显示基因以及其它序列特征在基因组上位置的图
- 方法是以多态的遗传标记作为界标，计算细胞减数分裂过程中遗传标记之间发生重组的频率，来确定两个遗传标记在染色体上的相对位置
- 遗传标记之间的相对距离，即图距以厘摩（cM，厘摩尔根，centi-Morgan）为单位
- 当两个遗传标记之间的重组值为1%时，图距即为1cM

遗传距离



两对等位基因之间重组互换的频率即遗传距离

遗传图谱构建

遗传图谱构建中需要可以鉴别的标记（marker），分为基因标记和DNA标记

- **基因标记**
  - 利用表型性状基因作标记，根据连锁交换原理来分析基因之间的连锁关系和遗传距离，绘制连锁图谱
- **DNA标记**
  - 限制片断长度多态（RFLP，Restriction Fragment Length Polymorphism）、扩增片段长度多态（AFLP，Amplified Fragment Length Polymorphism）、 随机扩增多态性DNA（RAPD，Random Amplified Polymorphic DNA）
  - 微卫星序列标记：短串连重复序列（STR，Short Tandem Repeat）、简单重复序列（SSR，Simple Sequence Repeat）
  - 单核苷酸多态（SNP，Single Nucleotide Polymorphism）

遗传图谱标记

标记类型	方法	数量	特点
基因标记	表型性状基因	有限	遗传距离较大
DNA标记 第一代	限制片断长度多态（RFLP）及其AFLP、RAPD	~10^5	多态程度较低 利用价值受限
DNA标记 第二代	微卫星序列（STR、SSR）	> 10^4	高度多态
DNA标记 第三代	单核苷酸多态（SNP）	3×10^6	一般为二态 单体型分析

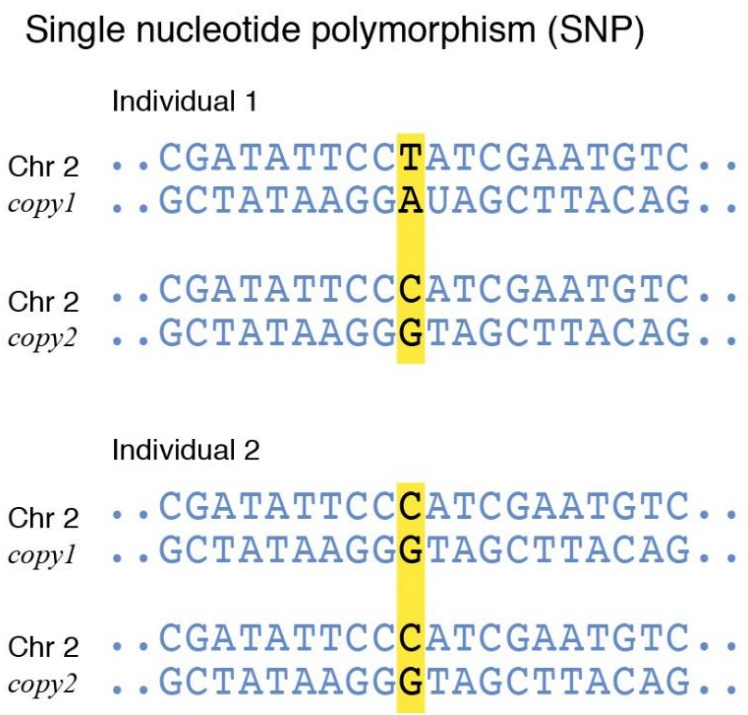
以DNA序列的多态性作为遗传标记

- 多态性（Polymorphism）：DNA序列上平均每几百个碱基会出现变异（variation），并按照孟德尔遗传规律由亲代传给子代，且在不同个体间表现出不同，因而被称为多态性

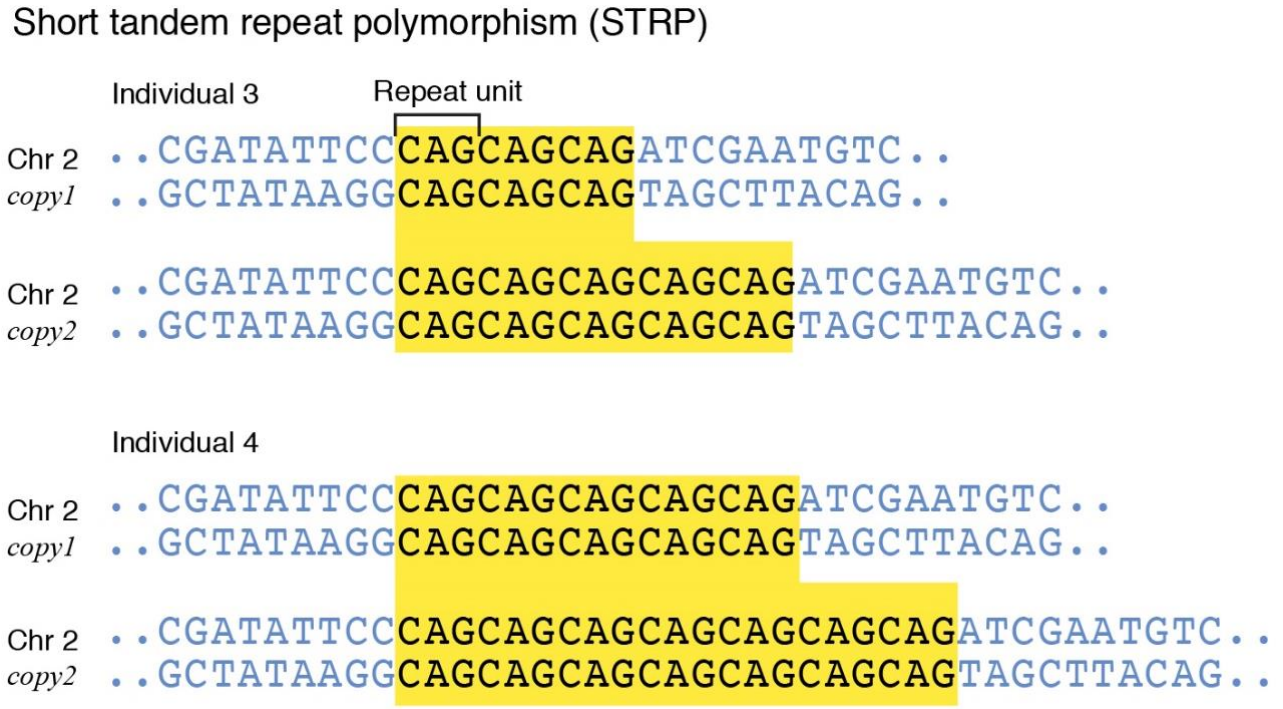
- 优点：
  - 不受时间和环境的限制
  - 遍布整个基因组，数量无限
  - 不影响性状表达
  - 自然存在的变异丰富，多态性好
  - 共显性，能鉴别纯合体和杂合体

遗传多态性（Polymorphism）

单核苷酸多态SNP



短串联重复序列多态STR



单个碱基的缺失、插入和替换
人类基因组中可能有300万个SNP位点

遗传图谱的局限性

- 分辨率有限
  - 人类或其它高等生物子代数量有限，可供研究的减数分裂事件偏少，限制了连锁分析
  - 人类基因组测序要求每100kb有一个标记，1996年发表的人类遗传图谱分辨率只达到600kb一个标记

- 精确度较低
  - 染色体上存在重组热点，影响邻近区段的遗传图谱准确性
  - 由于交换热点的存在，使某一区段的交换频率远高于其它区段，无法绘制精确均衡的遗传图谱

物理图谱

物理图谱—序列标签位点（STS）作图

- 序列标签位点（STS，Sequence Tagged Site）
  - 基因组中长度在200～500bp左右的特异性序列片段
  - 基因组中的单拷贝序列，只出现一次
  - 基因组中的位置和序列已知
  - 用于构建最为详尽的大基因组物理图谱的主流技术

- 物理图谱：以STS为“路标”有序地定位到基因组上，“路标”之间的物理距离（或图距）用碱基对（base pair；bp，Kb, Mb)表示

A **sequence-tagged site** (or **STS**) is a short DNA segment (200~500 bp) that has **a single occurrence in a genome and whose exact location and base sequence are known**.

https://en.wikipedia.org/wiki/Sequence-tagged_site

物理图谱绘制

- 物理图谱绘制需要筛选大量STS标记，制备大容量载体（cosmid、YAC、BAC）并进行大量复杂和繁琐的分析。

- 1995年，第一张以STS为物理标记的物理图谱问世，它包括了94%的基因组和1500多个标记位点，平均间距为200Kb（分辨率），把人类庞大基因组分成具有界标的1500个小区域。

- 人类基因组物理图的问世是基因组计划中的一个重要里程碑，被遗传学家誉为20世纪的"生命（生物学）周期表"。

遗传图谱和物理图谱的区别

Map	Genetic Map	Physical Map
距离方式	相对距离	物理距离
测量原理	遗传连锁重组分析	分子生物学技术
距离单位	cM	bp
分辨度	低	高
精确度	低	高
孟德尔遗传信息	依赖	不依赖

https://www.genome.gov/genetics-glossary/Genetic-Map

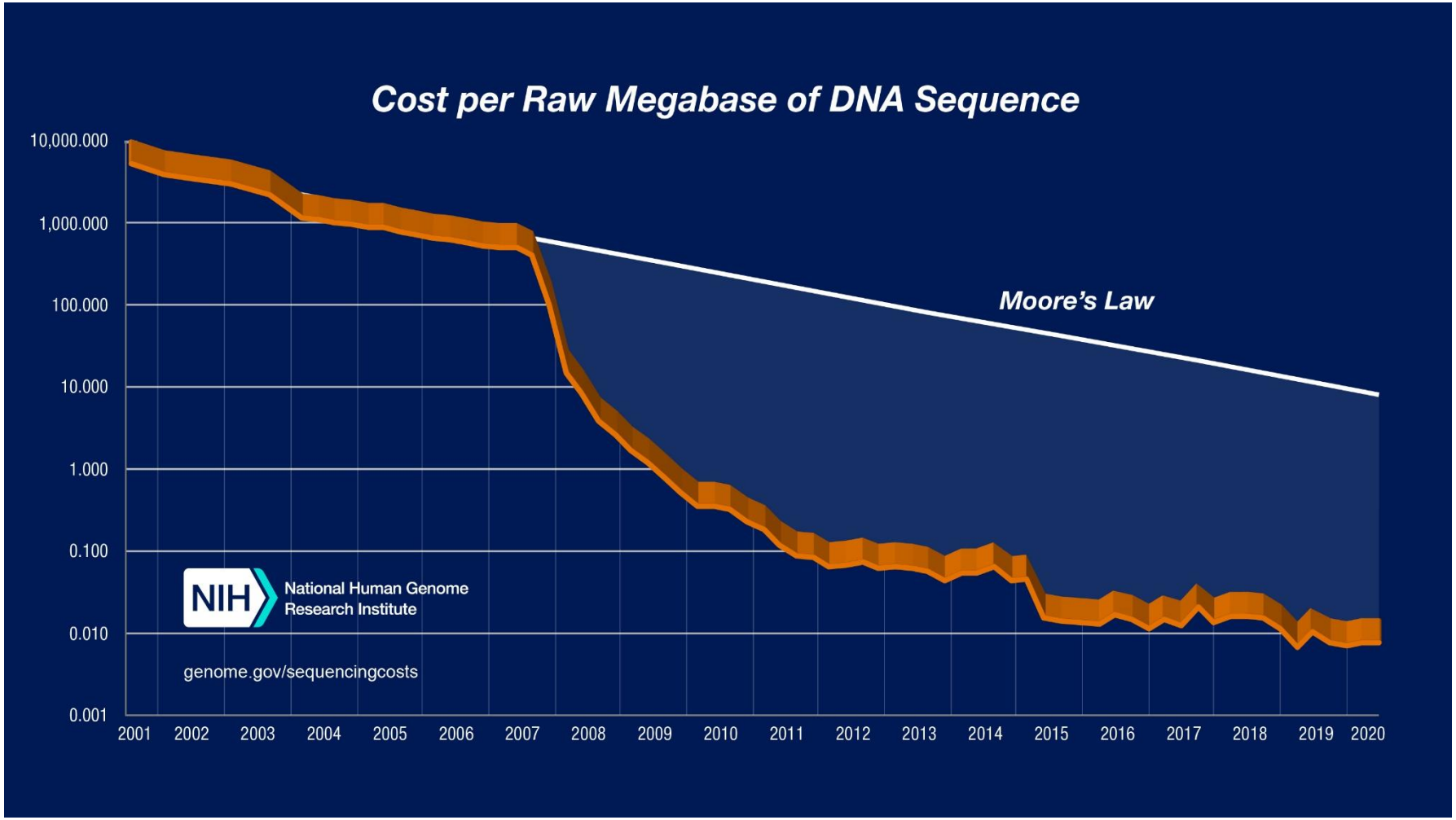
Human Genome Project Results

Area	Goal	Achieved	Date
Genetic Map	2- to 5-cM resolution map (600 - 1,500 markers)	1-cM resolution map (3,000 markers)	September 1994
Physical Map	30,000 STSs	52,000 STSs	October 1998
DNA Sequence	95% of gene-containing part of human sequence finished to 99.99% accuracy	99% of gene-containing part of human sequence finished to 99.99% accuracy	April 2003
Sequence Variation	100,000 mapped human SNPs	3.7 million mapped human SNPs	February 2003

https://www.genome.gov/human-genome-project/results

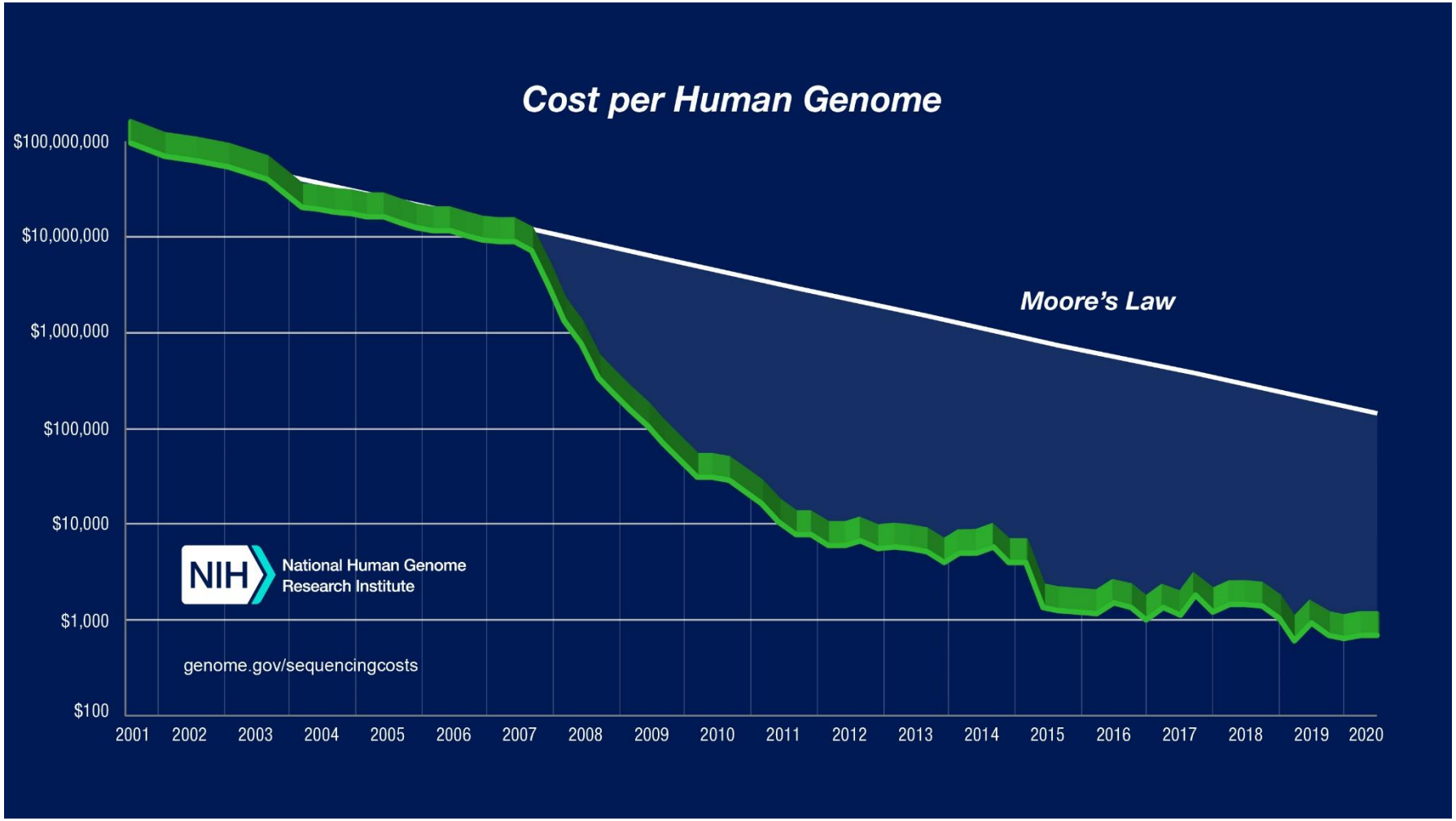
基因组测序技术

测序成本



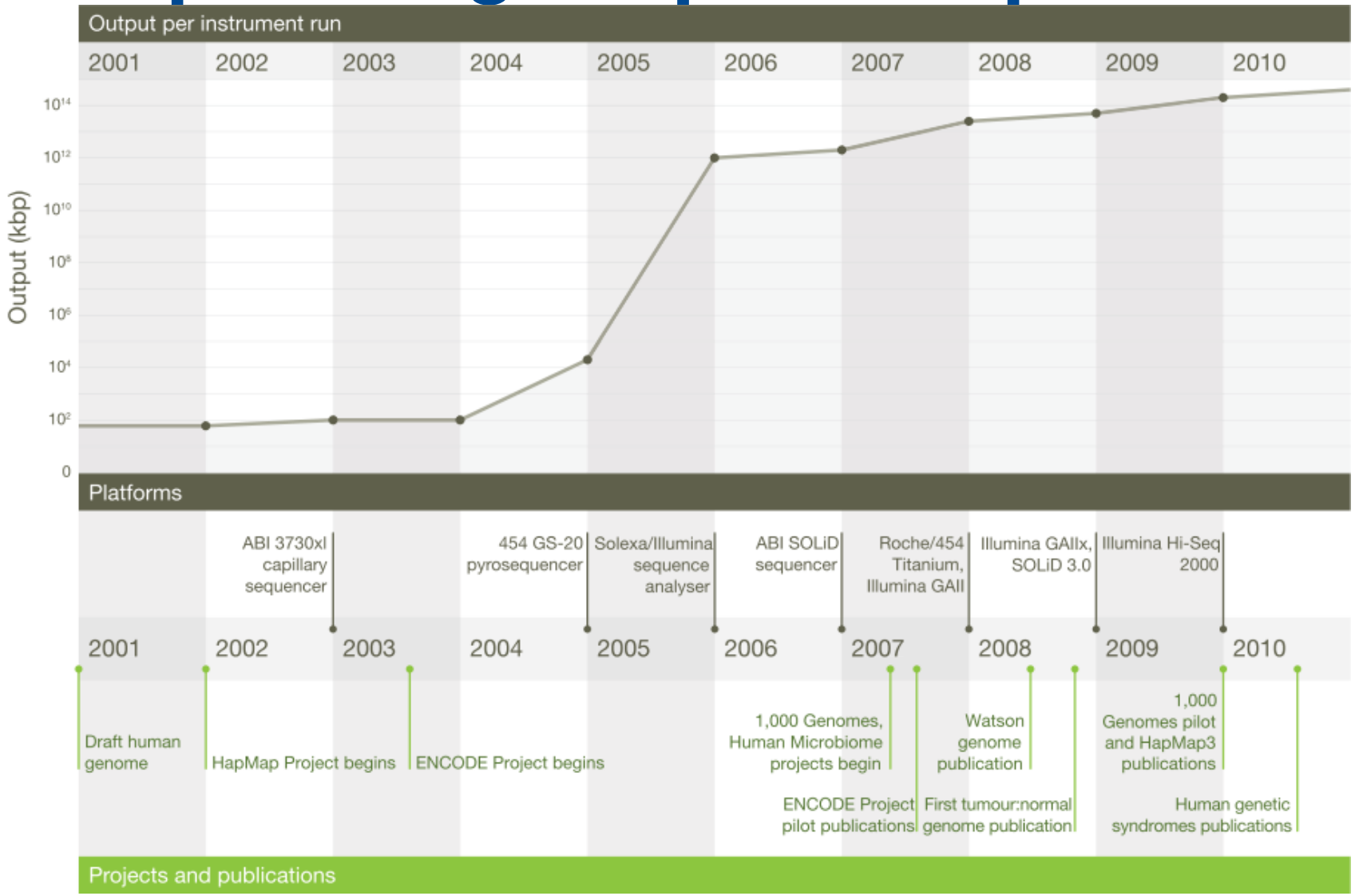
www.genome.gov/sequencingcostsdata

人类基因组测序成本



www.genome.gov/sequencingcostsdata

Sequencing output and platform



测序技术发展

- 第一代测序技术：又称Sanger测序法，由Sanger等人在1975年开创， 利用**双脱氧核苷酸**会终止PCR的原理，测序读长可达 1000 bp，但成本高、通量低。

- 第二代测序技术：又称NGS技术，是对第一代测序技术的划时代变革的核心。现有的技术平台主要包括HiSeq系列等，其所运用的测序原理为**循环微阵列法**，特点是通量高、读长短。

- 第三代测序技术：是指**单分子测序技术**，DNA测序时，不需要经过PCR扩增，实现了对每一条DNA分子的单独测序，也叫从头测序技术，即单分子实时DNA测序，特点是读长长、错误率高、成本高。

测序关键技术演变趋势

测序 关键技术	第一代	第二代	第三代
	Sanger法	循环阵列	单分子实时
分子杂交	√	√	√
酶介导生化反应	√	√	√
荧光标记底物	√	√	√
PCR扩增	√	√	
电泳/流控池	√	√	
微纳加工		√	√
荧光信号	√	√	√
单分子电信号			√
微量液流控			√

测序技术的核心组成逐渐从生物学技术向物理学技术倾斜

Massive sequencing platforms

Methods of DNA Sequencing		Platform
Short-read sequencing	Sequencing by ligation	ABI SOLiD (Thermo Fisher)
		Complete Genomics (BGI)
	Sequencing by synthesis	Illumina
		GeneReader (Qiagen)
		454 Roche
		Ion Torrent
Single-molecule real-time long-read sequencing		Pacific Biosciences
		Oxford Nanopore

https://www.mdpi.com/1424-8220/17/3/588/htm

Sequencers: generating big data

454

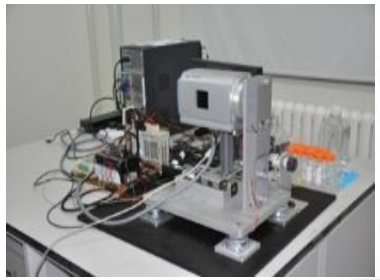
SOLiD4.0

SOLiD 5500xl

SOLEXA GAII

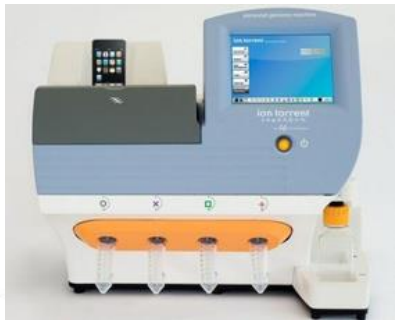
HiSeq 2000

PacBio

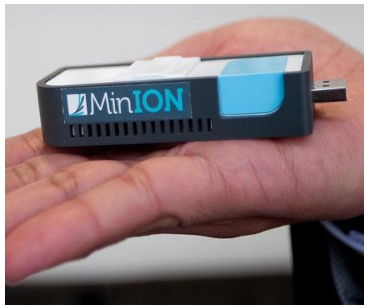


BIGIS-4

MiSeq



Ion Torrent



Nanopore's MinION

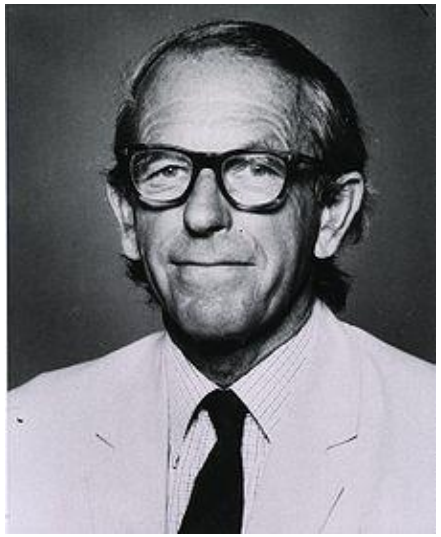
测序仪性能比较

Platform	Read length per run (bp)	No. reads per run	Time	Cost per 10^6 bases	Raw error rate (%)	Cost ($ approx.)	Chemistry
First-generation							
Sanger/Life Technologies/84 kb	800	1	2 h	2400	0.3	95,000	Dideoxy terminator
Second-generation							
454 GS FLX+/Roche/0.7 Gb	700	1×10^6	24/48 h	10	1	500,000	Pyrosequencing
HiSeq/Illumina/1500 Gb	2x150	5×10^9	27/240 h	0.1	0.8	750,000	Reversible terminators
MiSeq/Illumina/15 Gb	2x300	3×10^8	27 h	0.13	0.8	125,000	Reversible terminators
SOLiD/Life Technologies/120 Gb	50	1×10^9	14 days	0.13	0.01	350,000	Ligation
Third-generation							
SMRT/Pac Bio/1 Gb	>10,000	1×10^6	1–2 h	2	12.9	750,000	Real-time SMS
Nanopore/Oxford Nanopore Technologies/1 Gb	>5000	6×10^4	48/72 h	<1	34	1000	Real-time SMS

https://www.intechopen.com/books/next-generation-sequencing-advances-applications-and-challenges/next-generation-sequencing-an-overview-of-the-history-tools-and-omic-applications

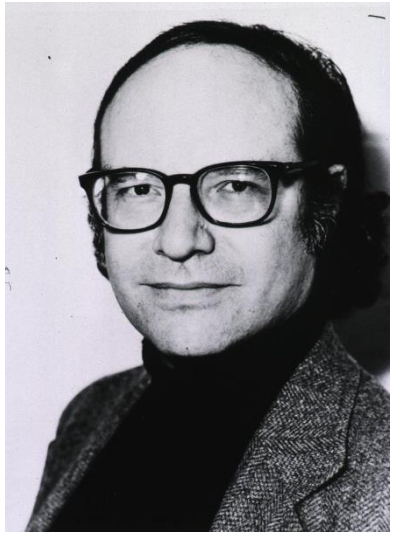
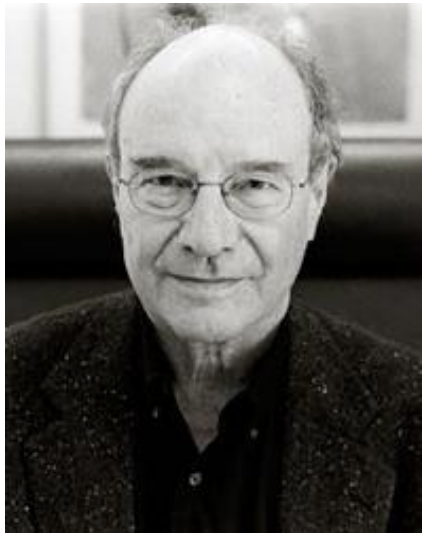
DNA测序技术的发展

- 第一代测序技术，自1977年开始

Frederick Sanger (1918-2013)

“双脱氧末端终止法” （Dideoxy termination method），即 “桑格法”

Walter Gilbert (1932-)

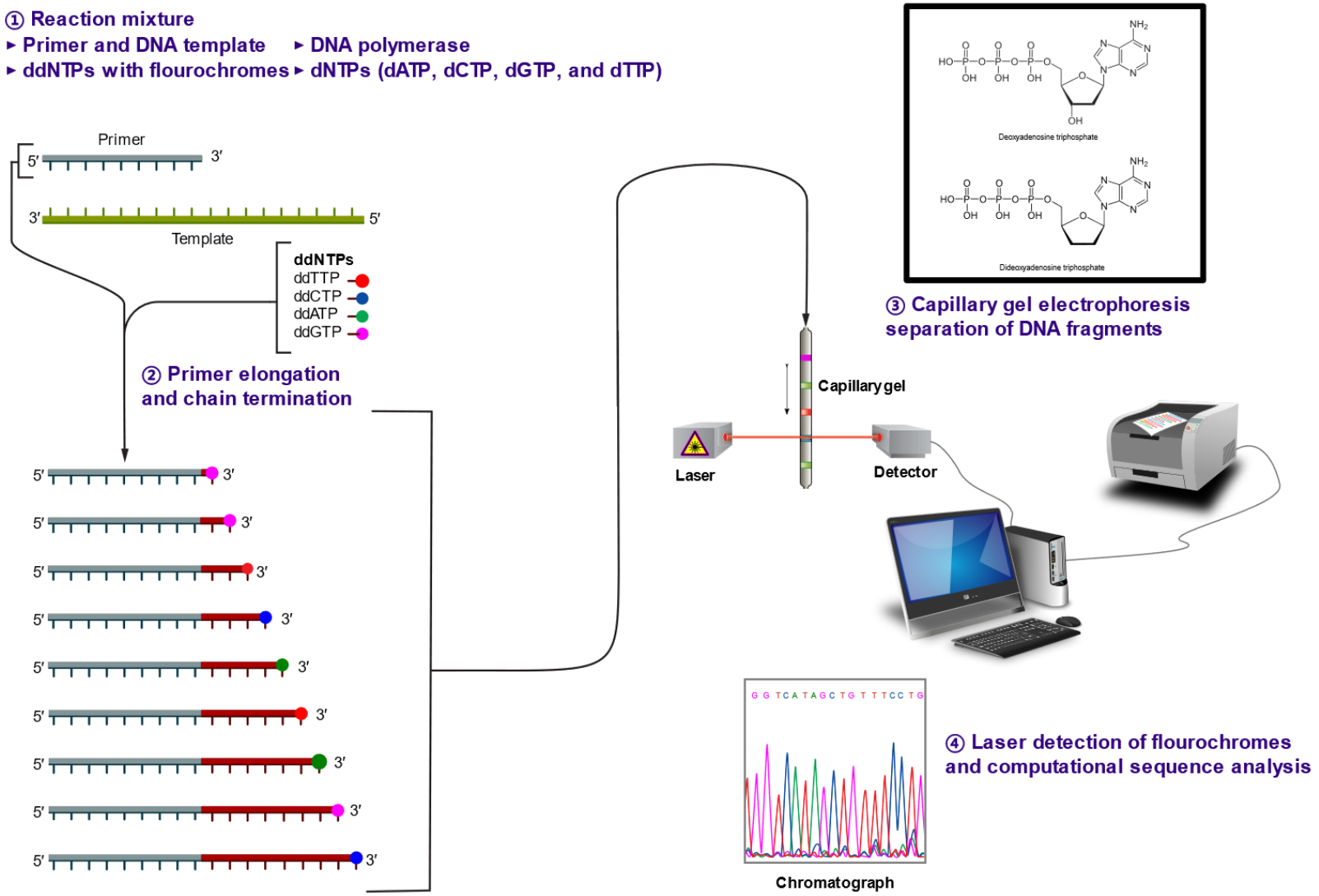
1977年发明DNA测序新方法； “RNA世界”一词的提出者（1986），创造了外显子exon与内含子intron两个名词

1980 Nobel Prize in Chemistry

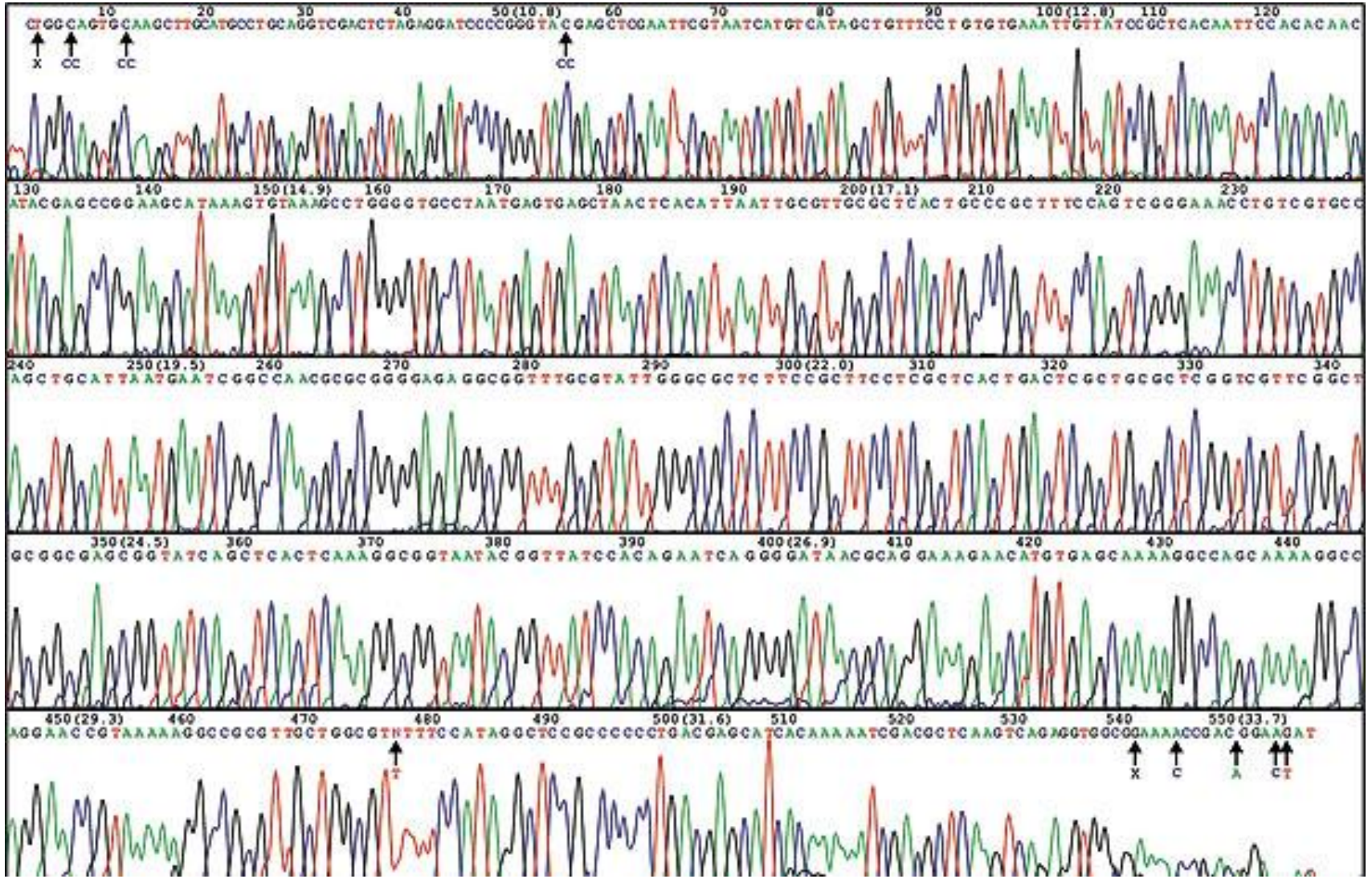
双脱氧末端终止

双脱氧核苷三磷酸（dideoxy-ribonucleoside triphosphate，ddNTP）

ddNTP缺少链延伸所需要的3'羟基基团，加入到合成产物末端后无法进一步延伸从而终止反应。



测序结果峰图



核苷酸名称和代码

代码	英文含义	中文含义
A	Adenine	腺嘌呤
G	Guanine	鸟嘌呤
C	Cytosine	胞嘧啶
T/U	Thymine/Uracil	胸腺嘧啶/胞嘧啶*
R（A or G）	purine	嘌呤
Y（C or T）	pYrimidine	嘧啶
M（A or C）	aMino	氨基（腺嘌呤或胞嘧啶）
K（G or T）	Ketone	酮基（鸟嘌呤或胸腺嘧啶）
S（G or C）	Strong interaction	三对（强）氢键（鸟嘌呤或胞嘧啶）
W（A or T）	Weak interaction	两对（弱）氢键（腺嘌呤或胸腺嘧啶）
H（A or C or T）	Not-G （H after G）	非鸟嘌呤
B（C or G or T）	Not-A （B after A）	非腺嘌呤
V（A or C or G）	Not-T （V after U）*	非胸腺嘧啶
D（A or G or T）	Not-C （D after C）	非胞嘧啶
N（A or C or G or T）	aNy	任意碱基

第一代测序仪

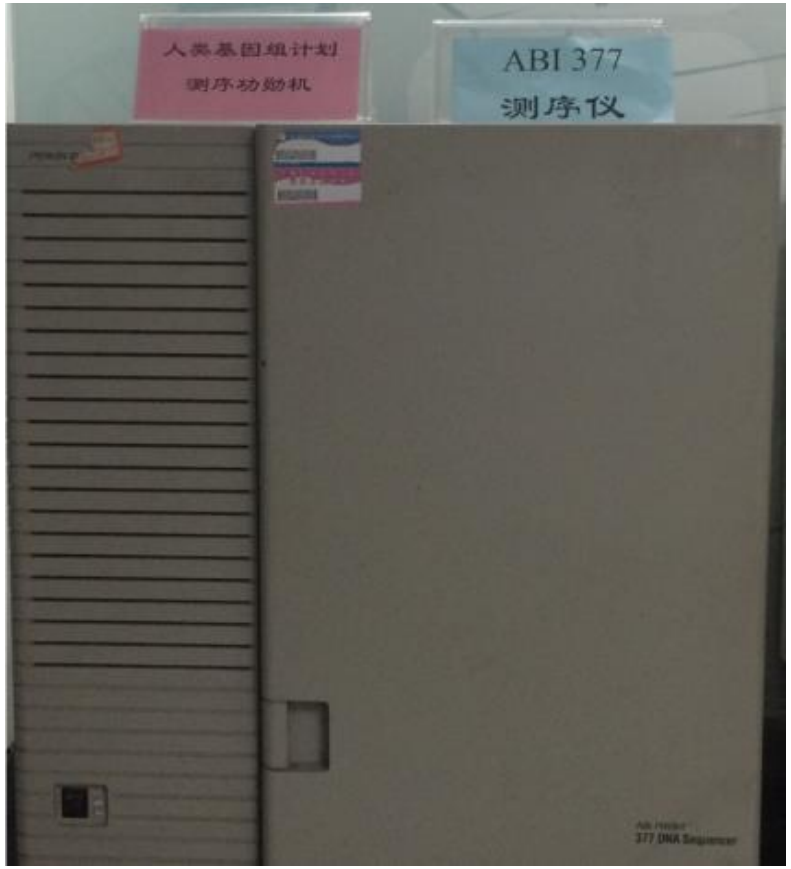
ABI377

MEGABACE 1000

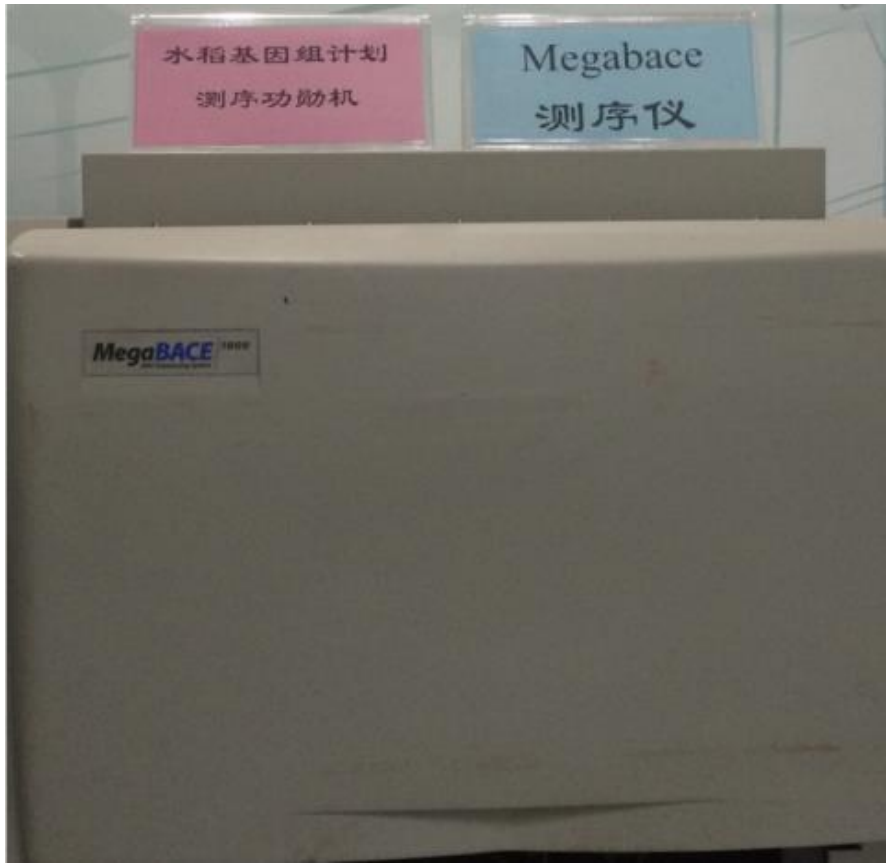
ABI3730XL

- 1985年，美国系统生物学家Leroy Hood（1938-）实验室改进了Sanger法测序
- 1986年，美国应用生物系统公司（Applied Biosystems Inc.，ABI）采用Hood实验室技术开发了第一台自动化DNA测序仪ABI 370

人类基因组计划和水稻基因组计划测序功勋



国际人类基因组计划
1%任务



超级杂交水稻基因组计划
籼稻 9311

DNA测序技术的发展

第二代测序技术：1990~2010

- Roche 454 焦磷酸测序法
- Illumina Solexa 聚合酶测序法
- ABI SOLiD 连接酶测序法

454 GS FLX

GA IIx

SOLiD 4.0

第二代测序仪：Solexa GA-II （2006年）

10万 reads per lane

Read 长度 36 bp

每个read有2到3个碱基错误

Solexa成立于1998年，
创始人是剑桥大学

12bp – 25bp – 30bp

2006年Illumina 6.5亿美元
收购Solexa

DNA测序技术的发展

- 第三代测序技术，2010-

单分子
非光学显微镜成像
纳米孔技术

Helicos BioSciences Corporation

Single molecule real time sequencing (SMRT)

PACIFIC BIOSCIENCES

Single-molecule DNA sequencing technology

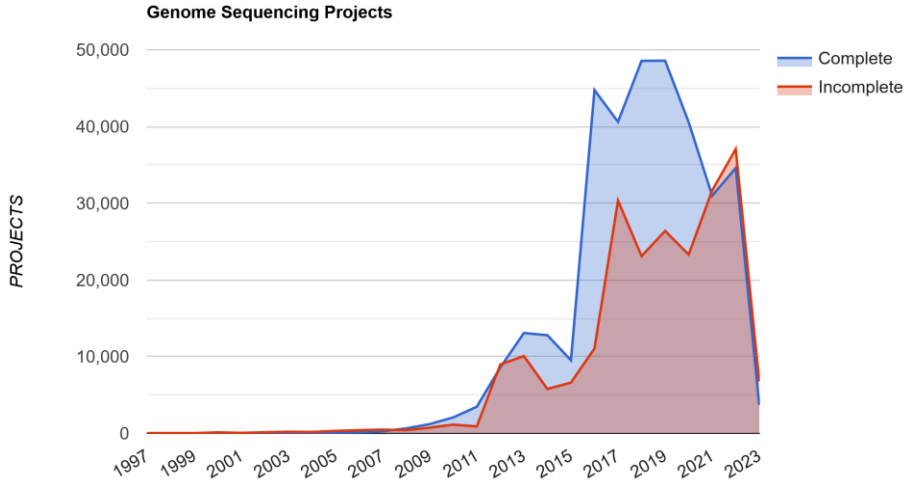
Oxford NANOPORE Technologies

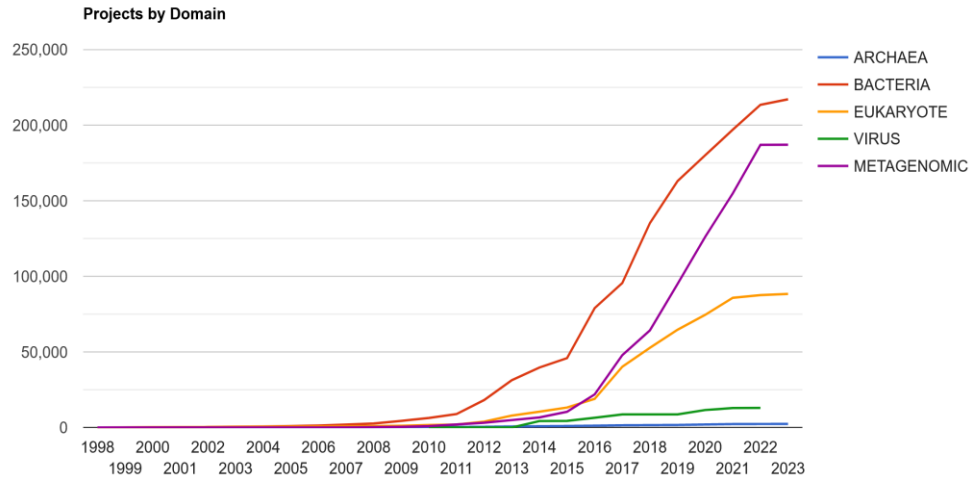
Single molecule nanopore sequencing

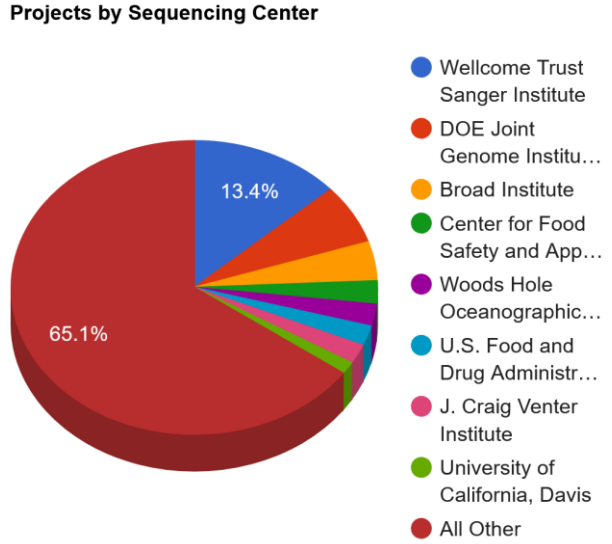
DNA测序速度

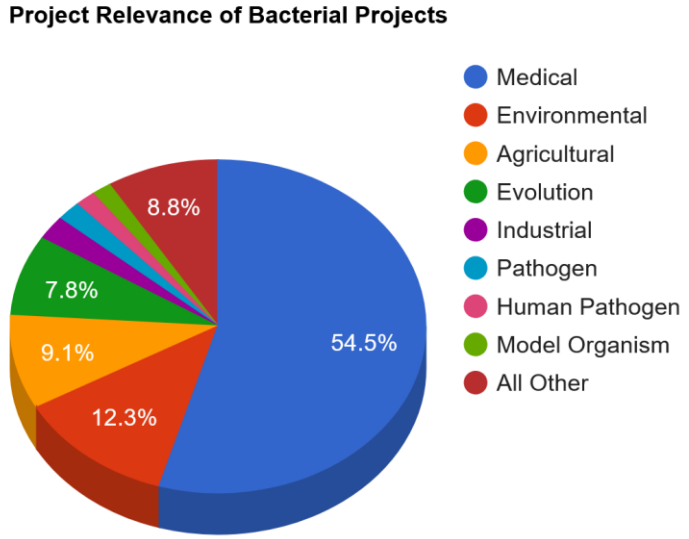
- 1985年：如果每人每年测定10,000 bp
  - 需要 300,000人/年或30亿美元
  - 3,000人工作100年或30,000人工作10年
- 1995年：3,000人工作十年
- 2000年：300人工作十年
- 2003年：毛细管电泳
- 2007年：NGS测序仪
- 2013年：每人每天一个人类基因组

Genome Statistics



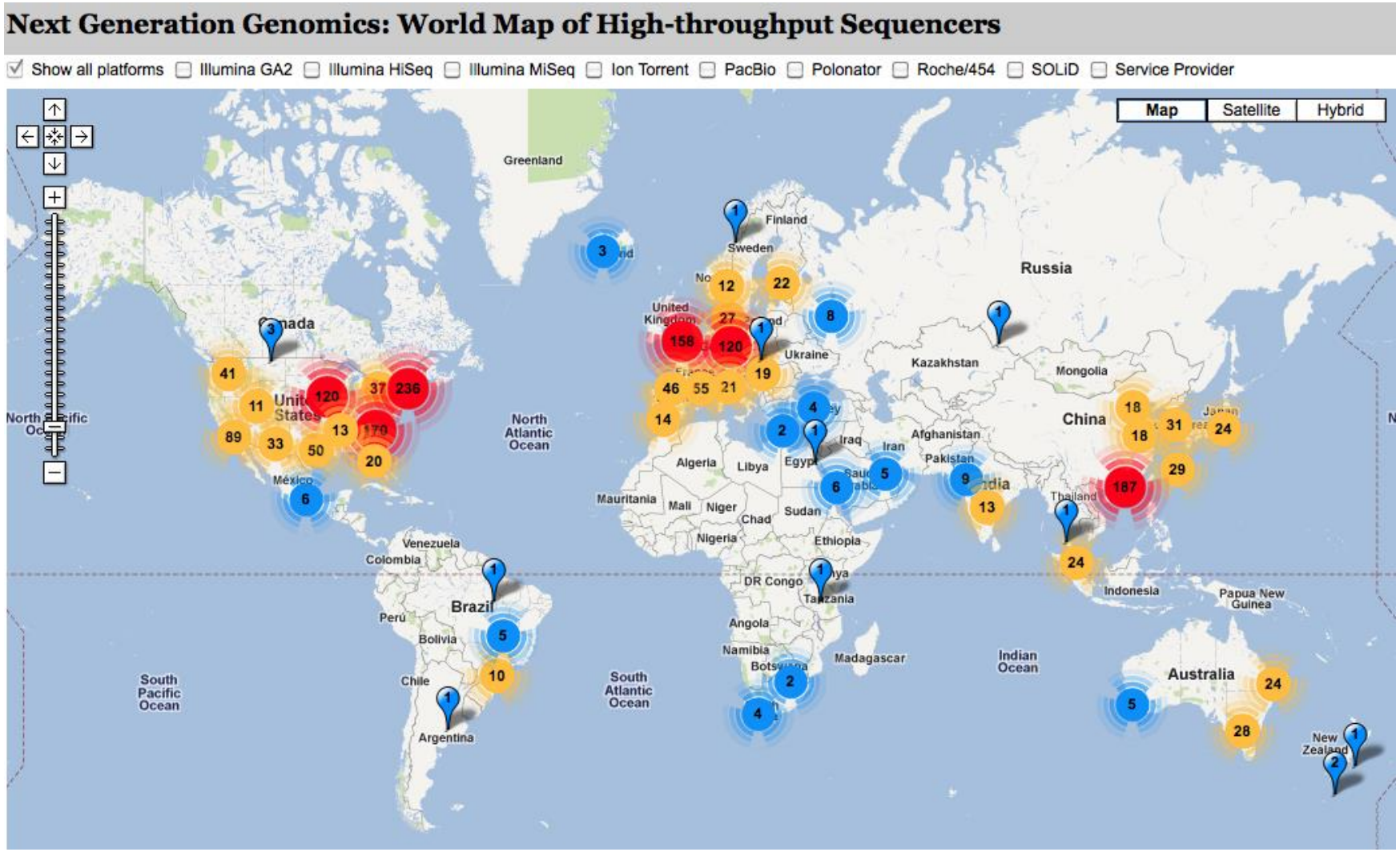






https://gold.jgi.doe.gov/statistics

高通量测序仪的世界分布



http://pathogenomics.bham.ac.uk/hts (unavailable; not updated)

Illumina HiSeq系列测序仪（低通量系列）

实验室级别的，外显子组测序、小RNA测序、转录组测序等

	 MiniSeq	 MiSeq*	 MiSeq Dx†	 MiSeq FGx‡	 NextSeq*
Output Range	1.8-7.5 Gb	0.3-15 Gb	0.3-15 Gb	0.3-15 Gb	20-120 Gb
Run Time	4-24 hr	5-55 hr	4-55 hr	4-55 hr	11-29 hr
Reads per Run	8-25 million	1-25 million	1-25 million	1-25 million (for research)	130-400 million
Maximum Read Length	2 x 150 bp	2 x 300 bp	2 x 300 bp	2 x 300 bp	2 x 150 bp
Samples per Run§	1-96	1-96	N/A	N/A	12-36
Relative Price per Sample§	Higher Cost	Mid Cost	Mid Cost	Mid Cost	Lower Cost
Relative Instrument Price§	Lower Cost	Mid Cost	Mid Cost	Mid Cost	Higher Cost
Downloads	Spec Sheet	Spec Sheet	Spec Sheet	Spec Sheet	Spec Sheet
System Overview	MiniSeq Overview >	MiSeq Overview >	MiSeq Dx Overview >	MiSeq FGx Overview >	NextSeq Overview >

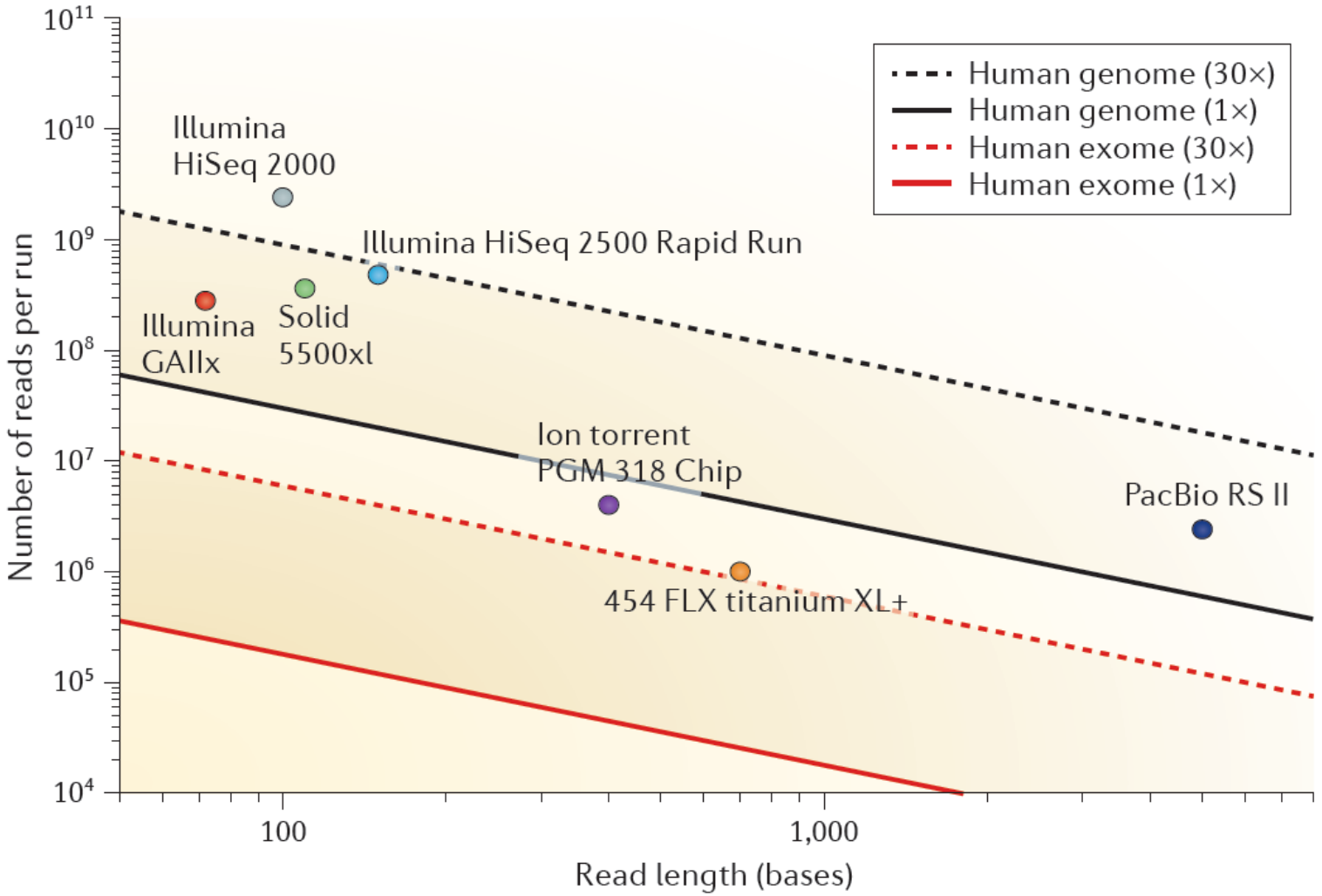
https://www.illumina.com/systems/sequencing-platforms.html

Illumina HiSeq 系列测序仪（高通量系列）

	 NextSeq*	 HiSeq 4000*	 NovaSeq 5000††*	 NovaSeq 6000††*
Output Range	20-120 Gb	125-1500 Gb	167-2000 Gb	167-6000 Gb
Run Time	11-29 hr	<1-3.5 days	TBA	19-40 hr
Reads per Run	130-400 million	2.5-5 billion	1.4-6.6 billion	1.4-20 billion
Maximum Read Length	2 x 150 bp	2 x 150 bp	2 x 150 bp	2 x 150 bp
Samples per Run†	1	6-12	4-16	4-48
Relative Price per Sample†	Lower Cost	Lower Cost	Lower Cost	Lower Cost
Relative Instrument Price†	Higher Cost	Higher Cost	Higher Cost	Higher Cost
Downloads	Spec Sheet	Spec Sheet	Spec Sheet	Spec Sheet
System Overview	NextSeq Overview ›	HiSeq Overview ›	NovaSeq Overview ›	NovaSeq Overview ›

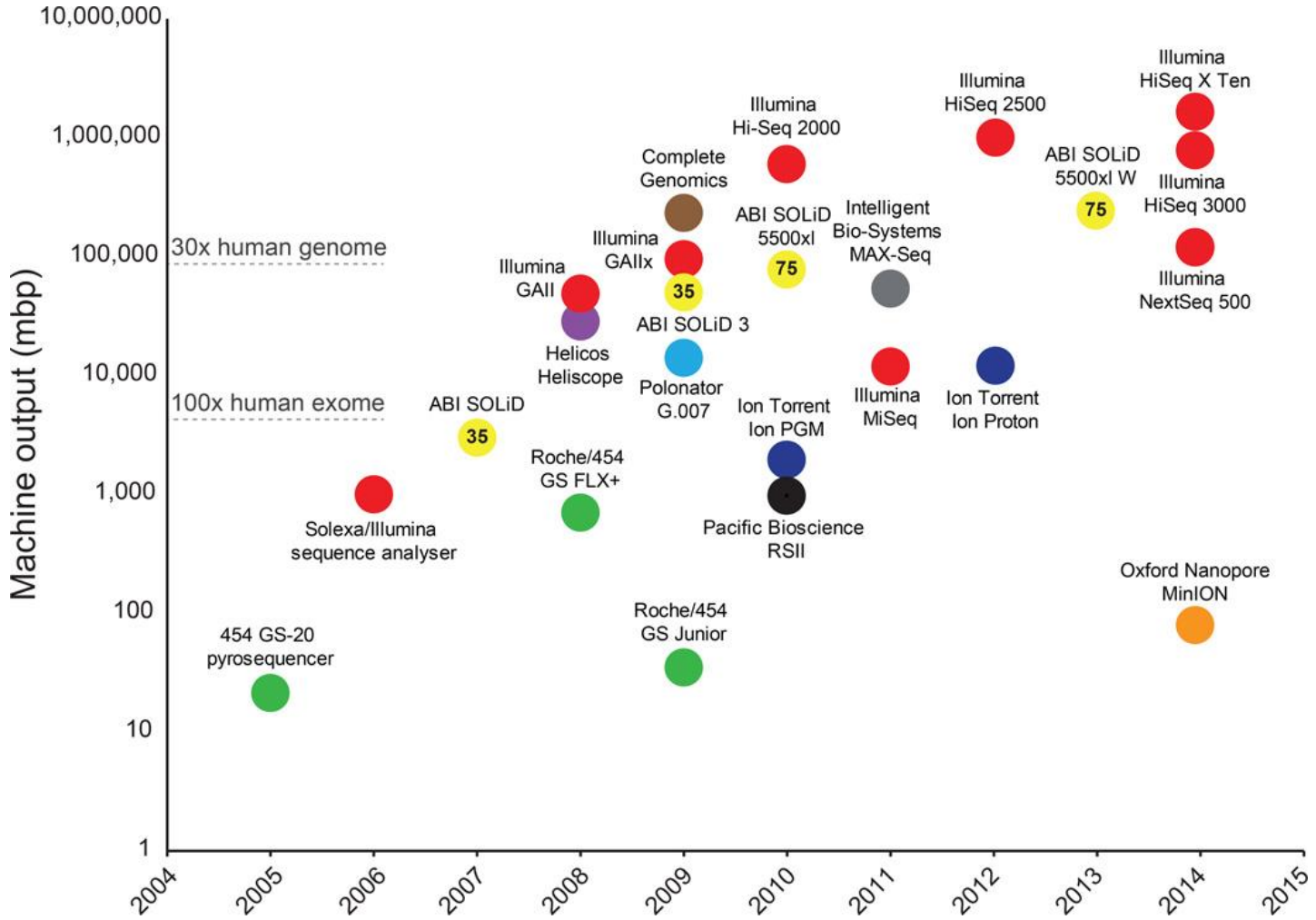
https://www.illumina.com/systems/sequencing-platforms.html

测序通量与读长



Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* 2014

Timeline and comparison of HTS instruments



High-Throughput Sequencing Technologies. *Molecular Cell* 2015. doi: 10.1016/j.molcel.2015.05.004

核酸测序技术发展

- 多代共存：个体化基因组研究、测定所有物种、metagenomics、pangenomics、多层次数据的获取、niche market等决定
- 多原理共存：缺乏“完美”技术、专利的限制和新原理的不断出现
- 高低通量共存：科研需求、应用需求、市场需求
- 长读长与短读长共存：读长、质量、通量和价格等无法同时满足

测序类型

测序类型	研究对象	目的
DNA-seq	基因组	测定生物样本的基因组序列
RNA-seq	转录组	测定打断后的转录本片段序列
Ribo-seq	翻译组	测定与核糖体结合的转录本
WGBS	表观组	测定DNA/RNA的5mC甲基化修饰
ChIP-seq	表观组	测定与蛋白质作用的DNA片段
MeRIP-seq	表观组	测定RNA甲基化修饰m6A
ATAC-seq	表观组	测定染色质开放区域
Hi-C	表观组	测定染色质高级结构

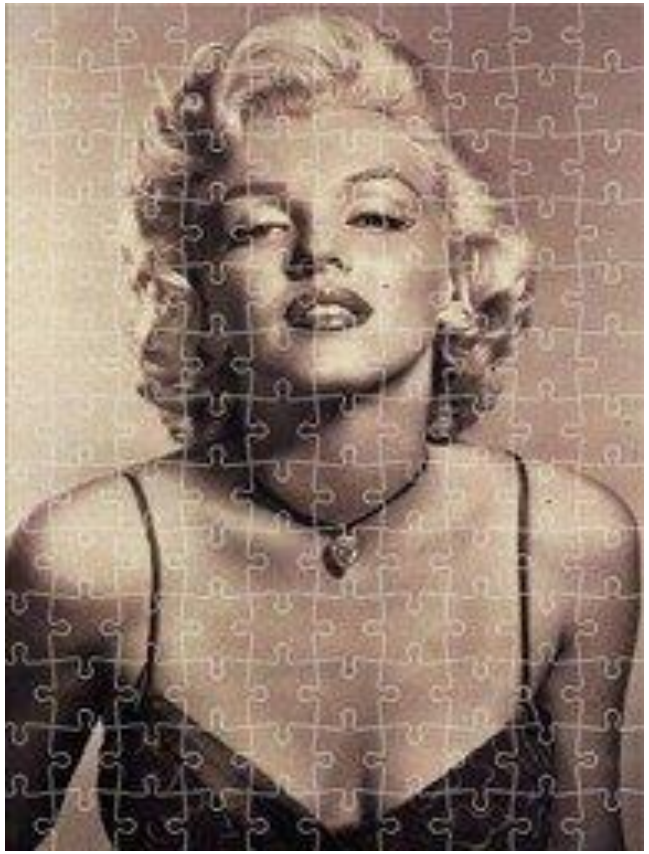
基因组测序里程碑事件

年份	事件	简介
2001	人类基因组草图	国际人类基因组测序联盟和Celera Genomics公司发表人类基因组草图
2004	宏基因组学诞生	利用测序技术所产生的测序数据无需纯化培养而重建微生物群落
2005	新一代测序技术	高通量、大规模并行测序技术登上历史舞台
2007	染色质免疫沉淀测序技术	全基因组层面探索蛋白质与DNA的相互作用
2008	个体基因组测序	利用新一代测序技术测定人类个体基因组序列（一名非裔和一名亚裔）
2008	癌症基因组测序	发布首个癌症（急性髓系白血病）样本的全基因组序列
2009	转录组测序	利用新一代测序技术进行全转录组高通量测序
2009	长读长测序	以单分子实时技术和纳米孔单分子技术为代表的第三代测序技术出现
2009	全外显子组测序技术	全外显子组测序被首次应用于单基因疾病研究
2009	染色质构象捕获技术	基于高通量测序方法在全基因组范围内研究染色质空间构象的新技术
2009	单细胞测序技术	首次在单个细胞水平进行全转录组测序分析
2010	古DNA测序技术	尼安德特人基因组序列的绘制标志着古基因组学研究的重要转折
2014	泛基因组	泛基因组图谱可以更全面的捕获物种基因集及其遗传变异信息
2017	基因组进入白金时代	长短读长测序、光学图谱、染色质构象捕获等多种技术相结合实现高质量基因组序列组装
2020	染色体完整测序	首次完成人类X染色体端粒到端粒无间隙完整测序组装
2022	基因组完整测序	发布第一个无间隙完整组装的人类基因组序列T2T-CHM13

https://www.nature.com/collections/genomic-sequencing-milestones

基因组测序策略

基因组测序 – 拼图游戏



基因组测序的两种策略

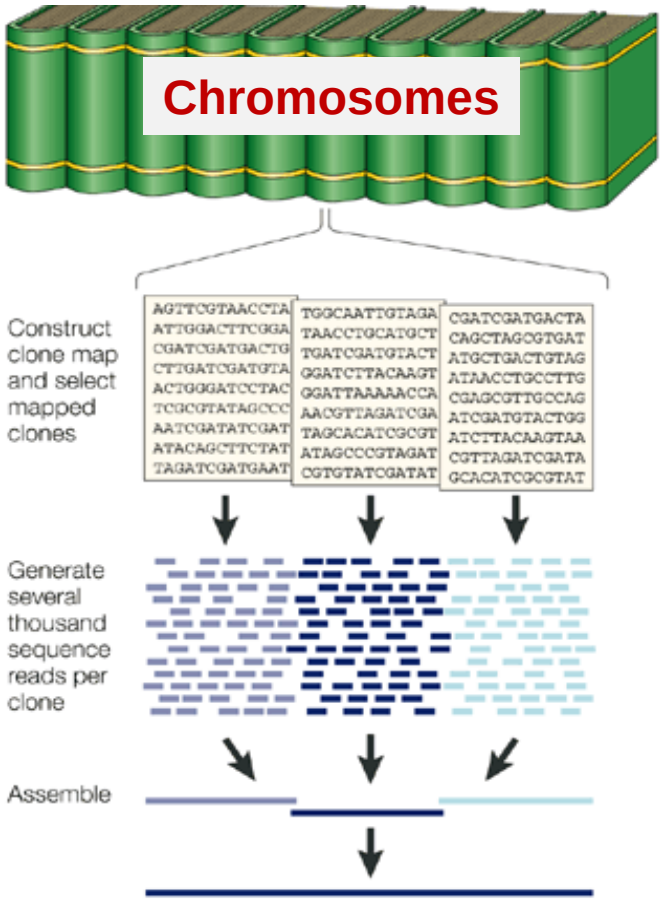
- **逐步克隆法**：按照大分子DNA克隆绘制的物理图谱分别在单个DNA克隆内部进行测序与组装，然后将彼此相连的大分子克隆按次序搭建支架，最后以分子标记为向导将搭建好的支架锚定在基因组整合图上。

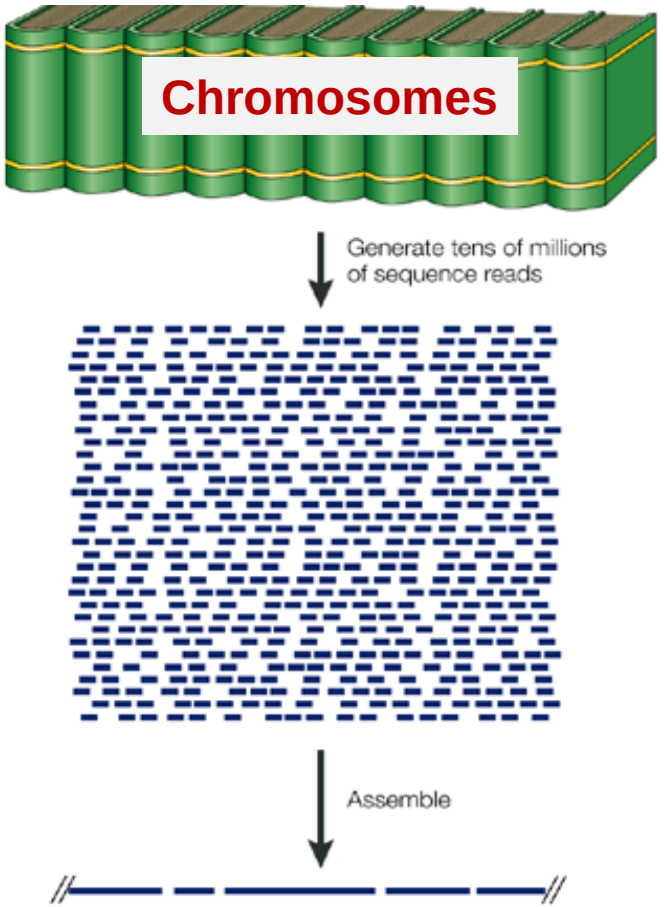
- **全基因组霰xiàn弹法**（鸟枪法）：将整个基因组DNA打断成小片段后将其克隆到载体中，然后随机挑取克隆进行测序，以获得的序列构建重叠群，进一步搭建序列支架，最后以分子标记为向导将序列支架锚定到基因组整合图上。

基因组测序的两种策略

逐步克隆法
BAC-to-BAC (clone-by-clone) sequencing

全基因组霰弹法
Whole genome shot-gun sequencing





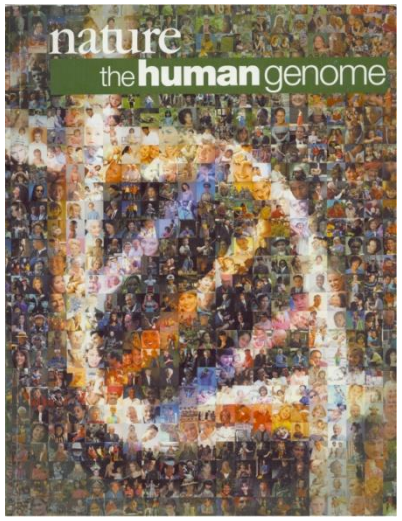
BAC文库
BAC libraries

Contig拼接限制在每个BAC里面，最后再拼接BAC序列

直接在全基因组范围内拼接

两种测序策略的应用

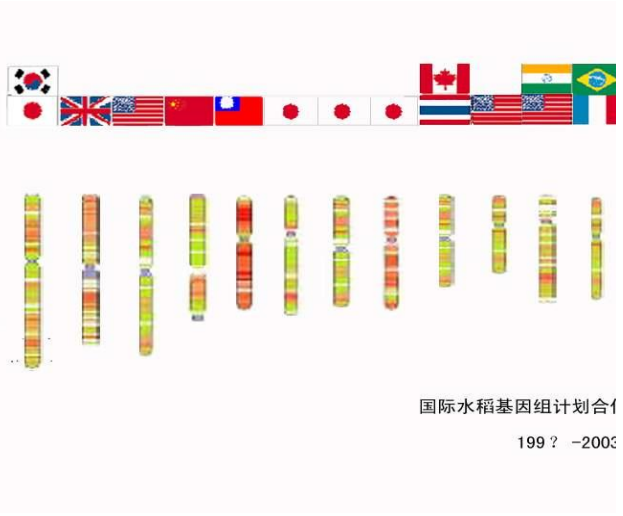
BAC-to-BAC



the sequencing of the human genome is likely to be the only large sequencing project carried to completion by the methods described in this issue.

Maynard V. Olson, The maps: Clone by clone by clone , Nature 409, 816 - 818 (2001)

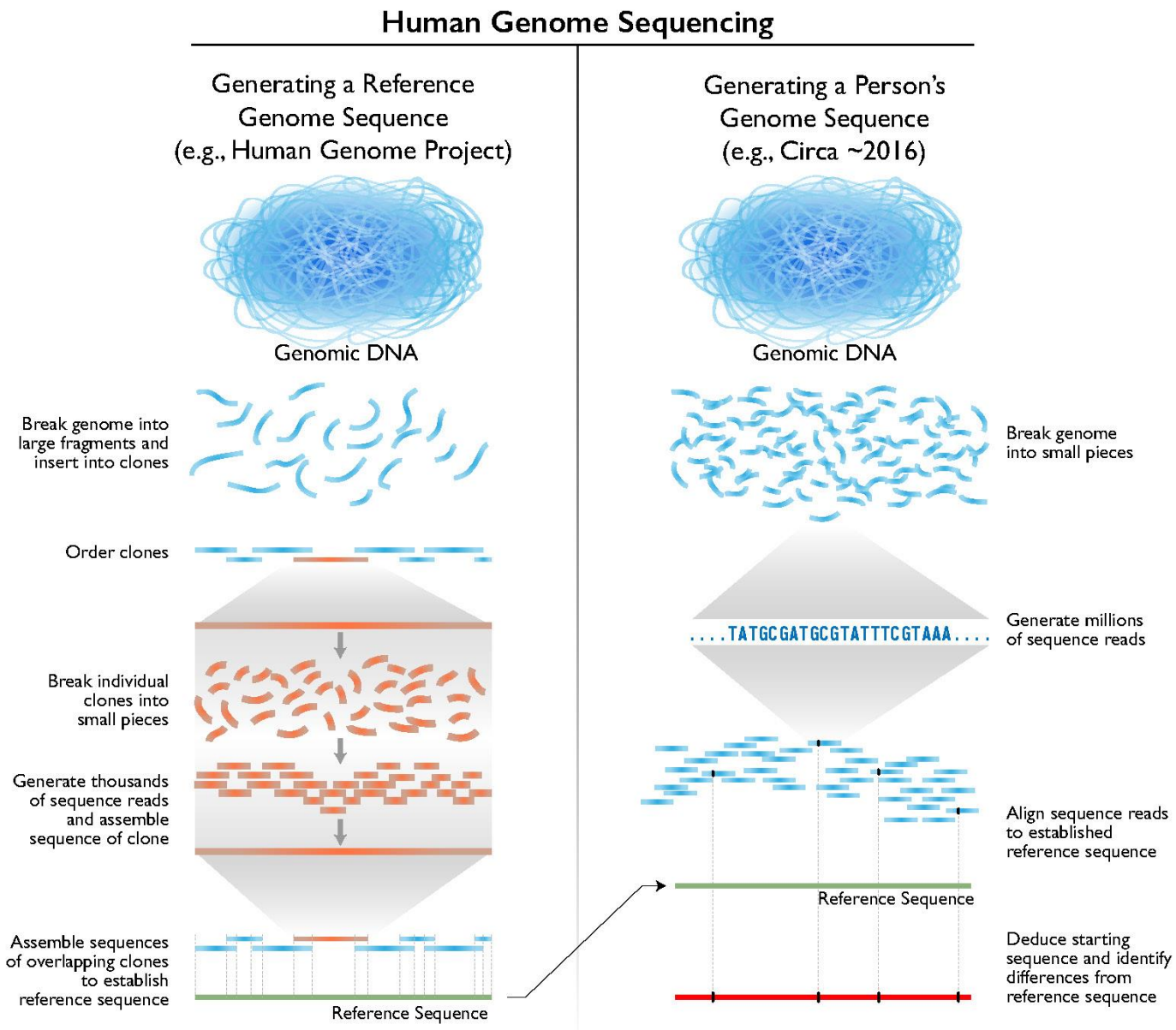
Whole Genome Shotgun



两种基因组测序策略的比较

因素	全基因组霰弹法	逐步克隆法
遗传背景	不需要	需要（需构建精确的物理图谱）
速度	快	慢
费用	低	高
计算机性能	高 (以全基因组为单位进行拼接)	低 (以BAC为单位进行拼接)
适用范围	工作框架图	精细图
代表测序物种	果蝇、水稻	人、线虫

人类基因组测序方法

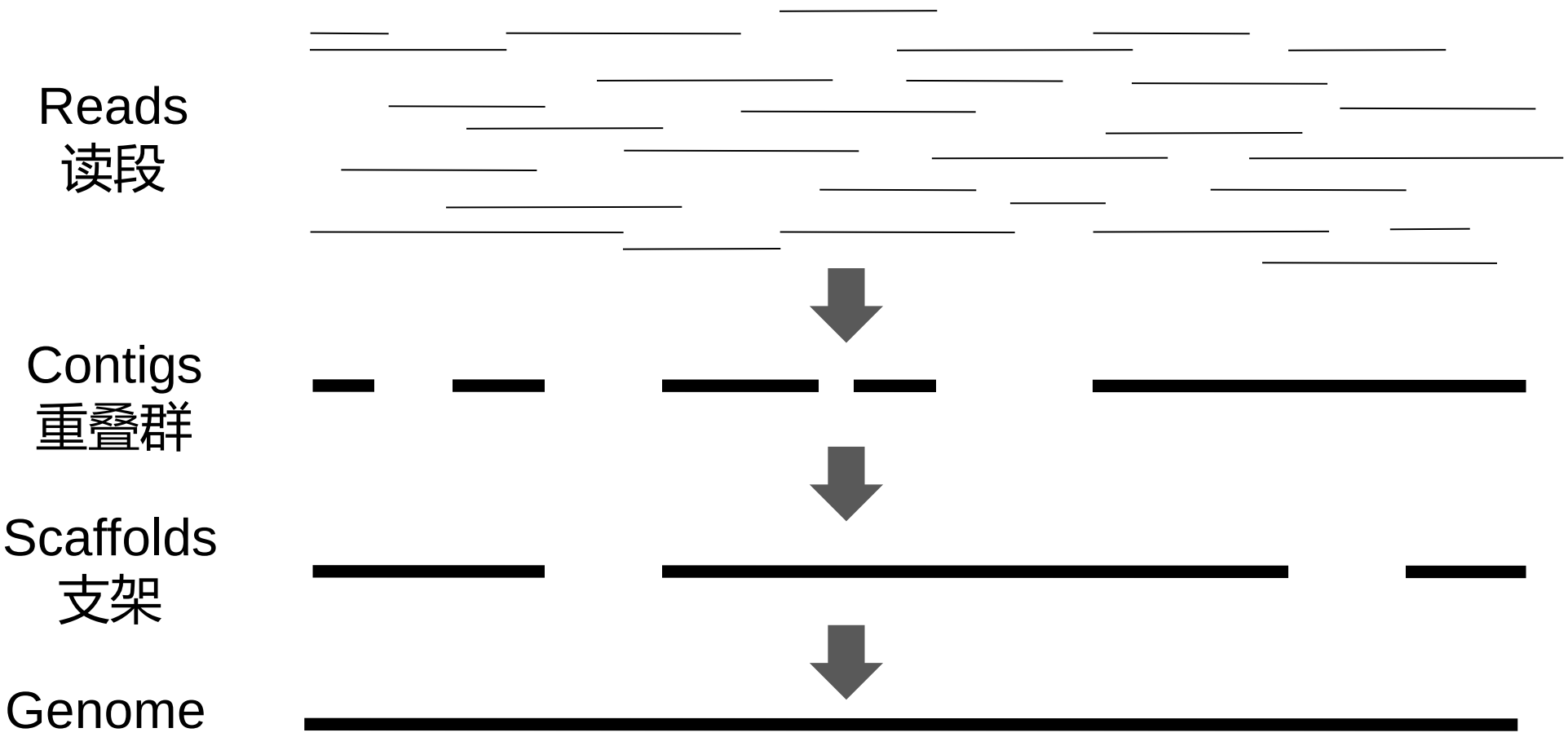


https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

基因组测序与拼接的三种基本方式

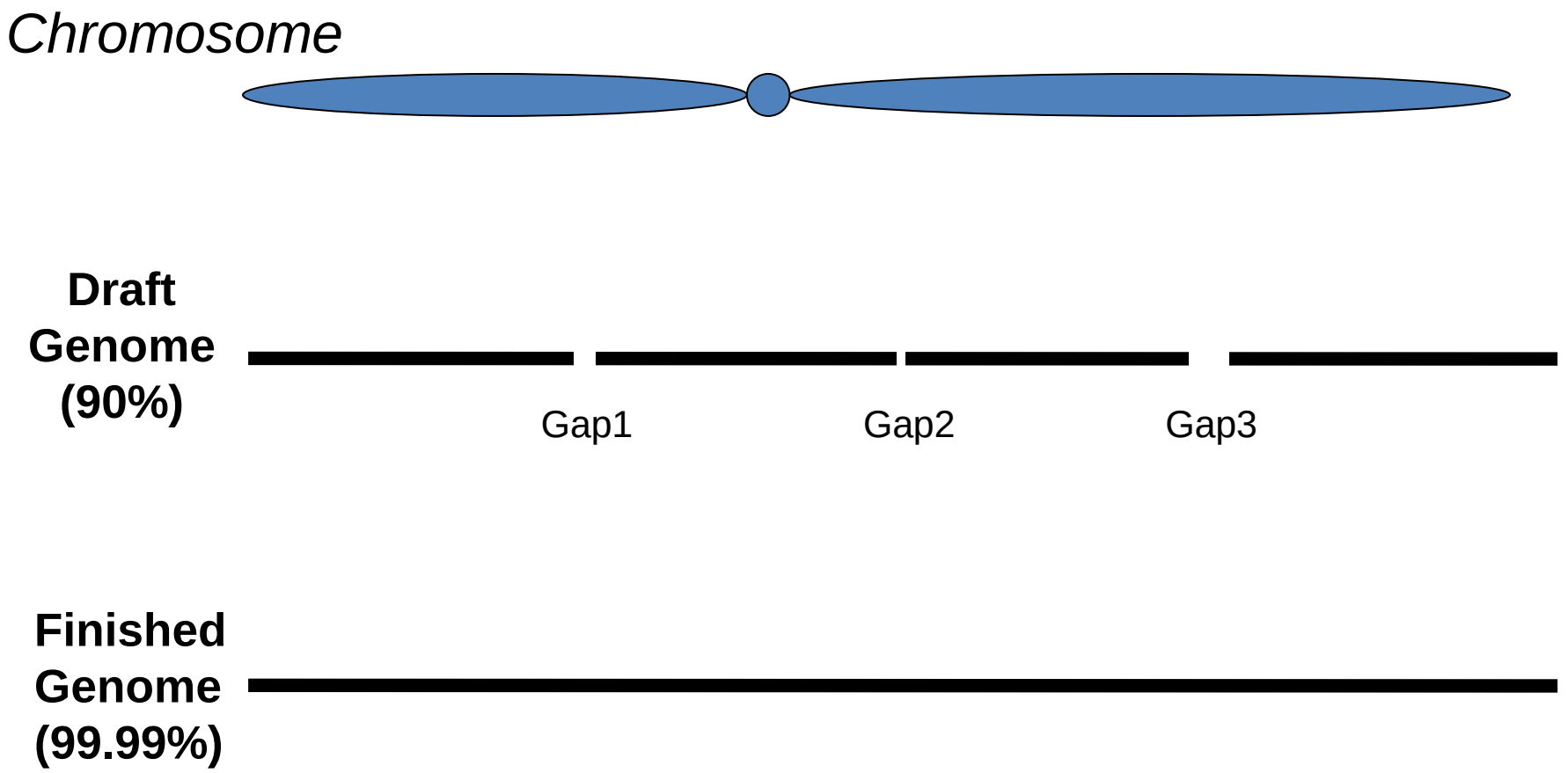
- ***de novo* Genome Sequencing and Assembly**
  - 测序完全一个新的模式物种的基因组
  - 大部分模式物种都是用BAC by BAC的精细测序、拼装获得
  - 质量最高，但人力、物力、财力花费巨大
- **Reference-guided Assembly**
  - 测序相近的物种、亚种，并用已测序的模式基因组作为参考进行指导拼接，以节省基因组拼接的资源，降低误差
  - 测序数据mapping到模式生物上，推测可能的基因组序列
- **Re-sequencing Mapping**
  - 以病人、突变体、近等基因系（一组遗传背景相同或相近，而某个特定性状或其遗传基础有差异的一组品系）、生态型（同一物种在不同生态环境下长期生活的不同类群）等为研究对象
  - 以鉴定变异位点、SNP、群体关联分析为目的
  - 大部分是用二代高通量测序获得

基因组拼接组装原理

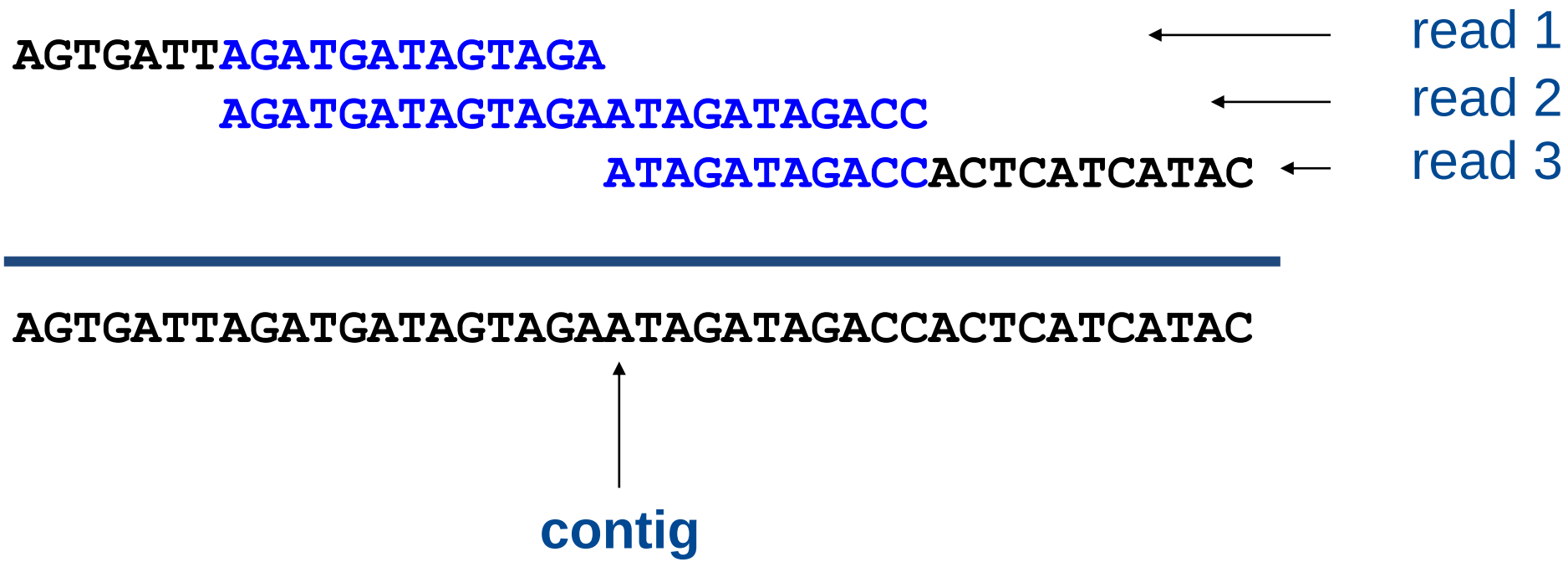


A **contig** (from **contiguous**) is a set of overlapping DNA segments that together represent a consensus region of DNA. **Scaffolds** consist of overlapping contigs separated by gaps.

工作草图与完成图/精细图

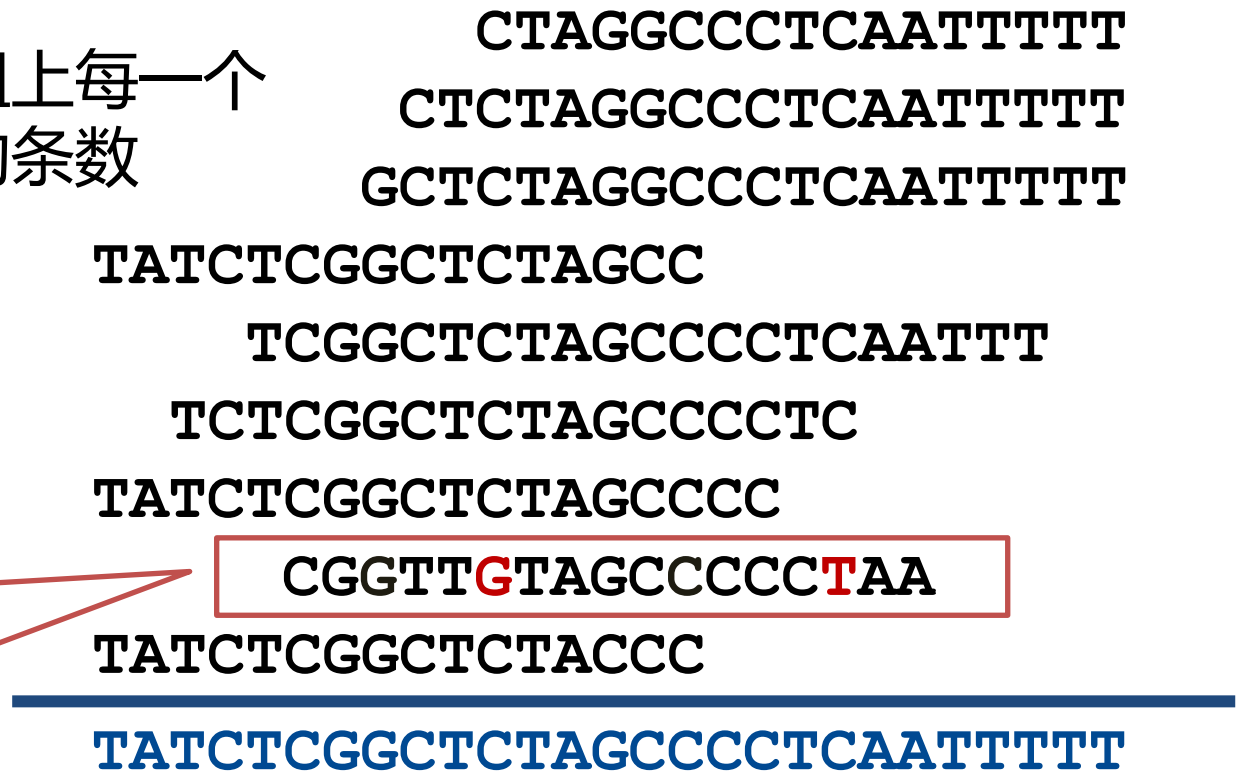


从测序读段到重叠群


AGTGATTAGATGATAGTAGA
AGATGATAGTAGAATAGATAGACC
ATAGATAGACCACTCATCATAC
read 1
read 2
read 3
AGTGATTAGATGATAGTAGAATAGATAGACCACTCATCATAC
contig

测序拼接原理

测序覆盖度：基因组上每一个位置平均覆盖读段的条数



166个碱基

测序错误或杂合子或mapping错误？

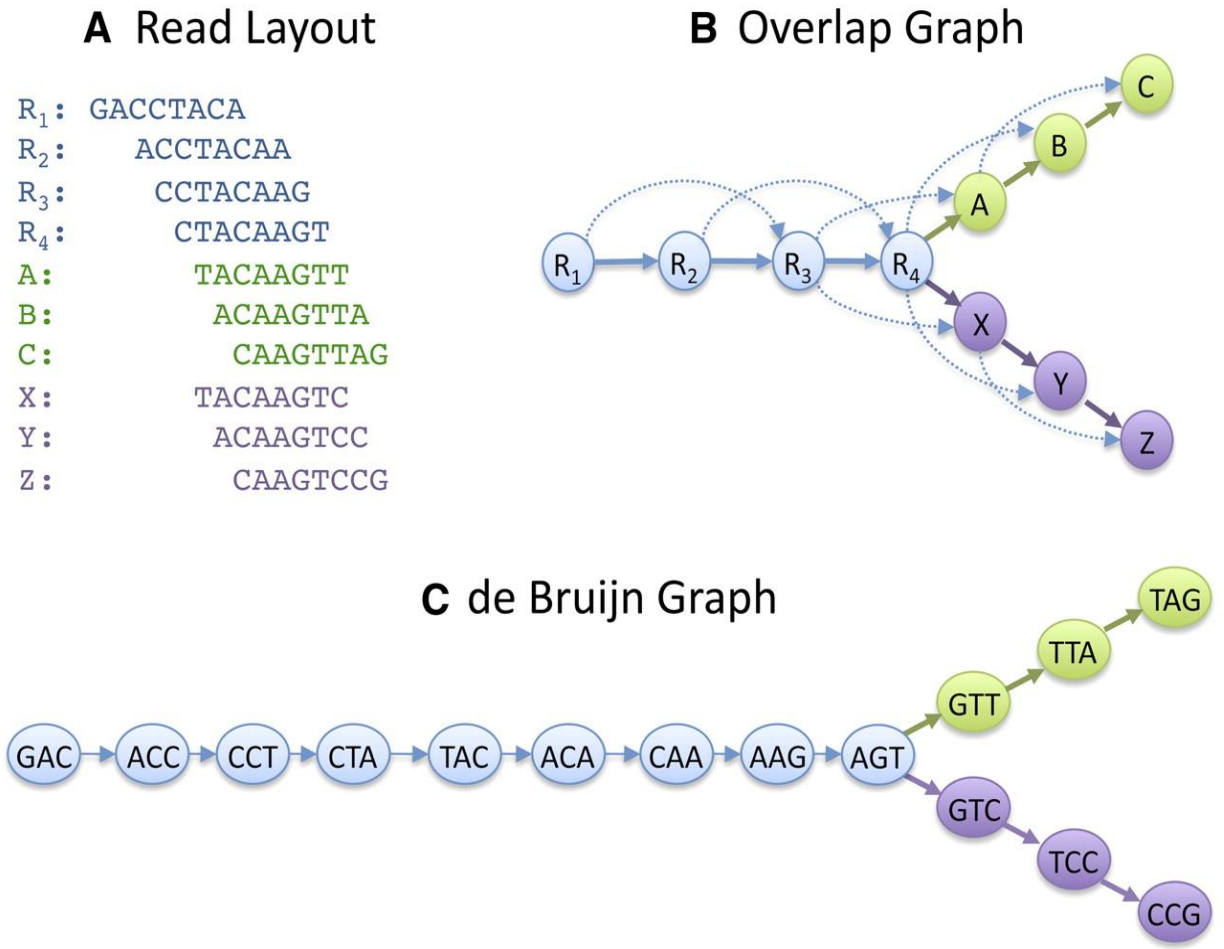
28个碱基

覆盖度C = *nl* / *L*

覆盖度C = 166 / 28 = 6x

NGS测序拼接的三种常用算法

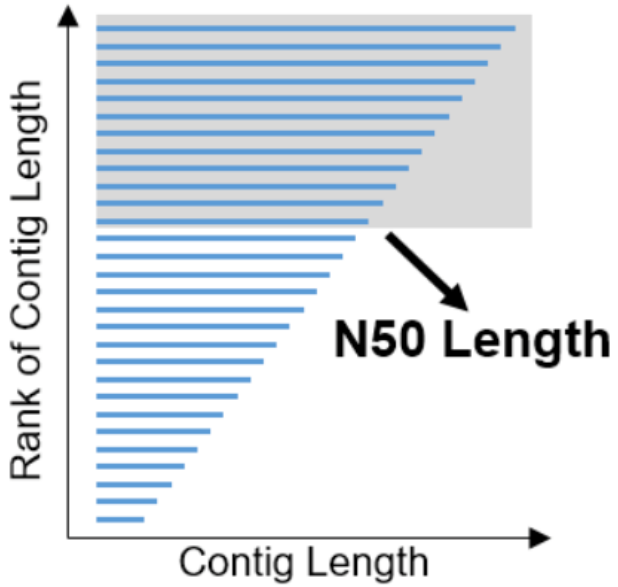
- Overlap Graph assemblers
  - Newbler
  - Edena
- de Bruijn Graph (DBG) methods
  - Velvet
  - ABySS
  - Euler
  - SOAPdenovo
- Greedy graph algorithms
  - SSAKE
  - VCAKE



Assembly of large genomes using second-generation sequencing. *Genome Research* 2010
https://genome.cshlp.org/content/20/9/1165.full

基因组拼接质量评估

- 基本统计: contigs个数, contigs总长, contigs平均和中值长度, 最长contig的大小（越多长contigs 、越少短contigs，效果越好）
- N50是拼接质量评估的重要参数：将contigs按其长度从大到小排列，然后从大到小相加，当恰好加到该基因组长度的50%，即为Contig N50，数值越大说明组装的质量越好（N90类似）
- Scaffold和chromosome的拼接涉及很多人为因素，参考性不大



For example, 9 contigs with the lengths 2,3,4,5,6,7,8,9,and 10; their sum is 54, half of the sum is 27. Thus,
N50=8, since 10 + 9 + 8 = 27

N50 statistic defines assembly quality in terms of contiguity. Given a set of contigs, the N50 is defined as the sequence length of the shortest contig at 50% of the total genome length.

二代拼接软件与工具

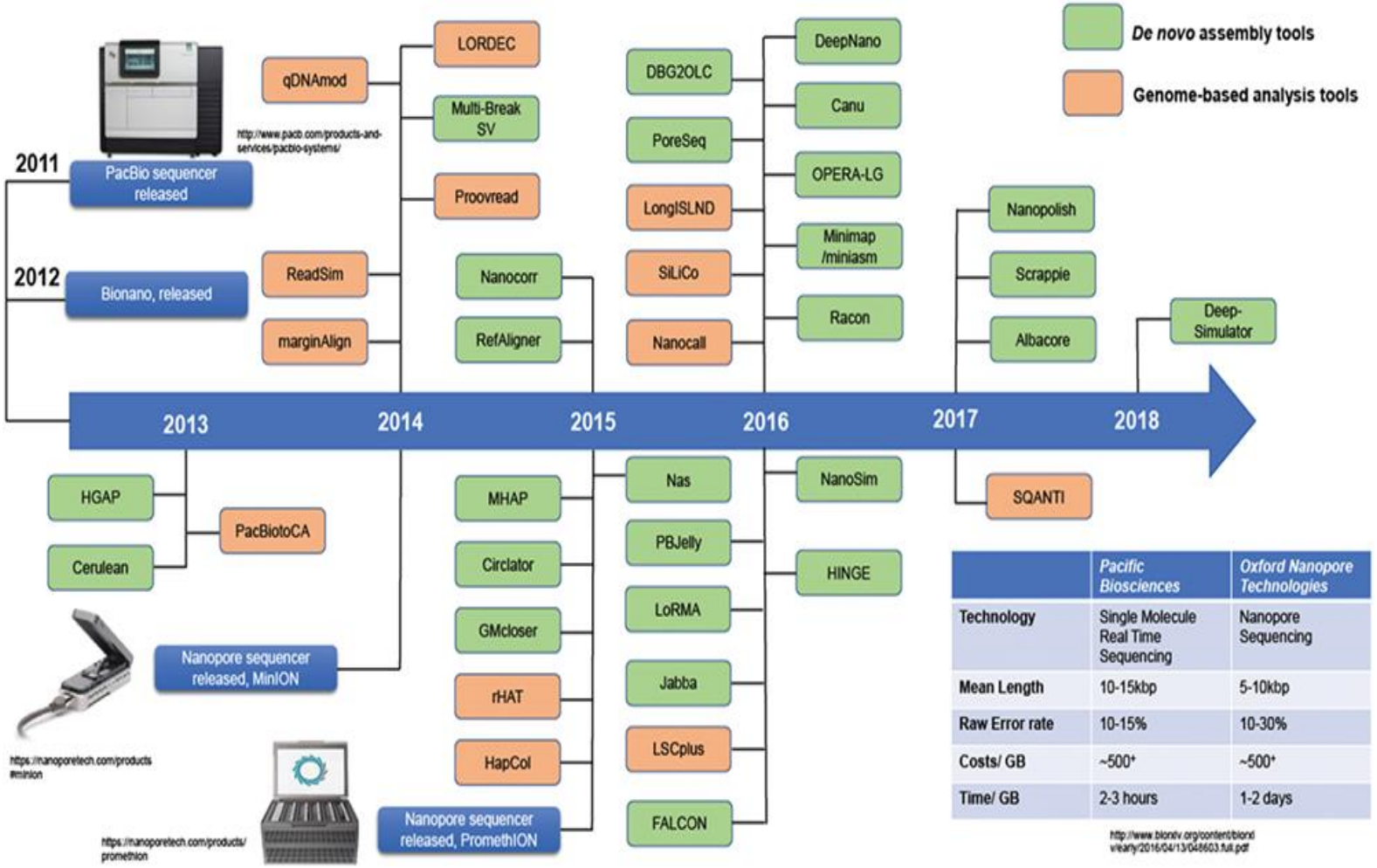
Table 1. De novo assemblies of second-generation sequencing projects

Organism/genome size	Assembler/status[a]	Input sequence: Type	Pair size	Avgerage read (bp)	No. of reads	Read coverage[b]	Pair coverage[c]	Assembly: Contigs No.	Contigs N50	Contigs Max	Contigs Total	Scaffolds No.	Scaffolds N50	Scaffolds Max	Scaffolds Total	Notes
Human (*H. sapiens*)/3.0 Gb	ABYSS published 2009	GA	210 bp	35–46	3.5 B	45×	120×	2.76 M	1.5 kb	18.8 kb	2.18 Gb	NR	NR	NR	NR	
Grapevine (*V. vinifera*)/500 Mb	Myriad published 2007	Sanger	2–10 kb	579	5.95 M	6.9×	21×									
		Sanger	40 kb	460	144 k	0.13×	4.4×	58,611	18.2 kb	238 kb	531 Mb[d]	2093	1.33 Mb	7.8 Mb	421 Mb[d]	
		Sanger	120 kb	369	68 k	0.02×	4.2×									
		454	None	169	12.5 M	4.2×	—									
Cucumber (*C. sativus*)/367 Mb	RePS2 published 2009	Sanger	2–6 kb	439	2.08 M	3.35×	9.9×									
		Sanger	40 kb	496	339 K	0.46×	16.7×	62,412	19,807	NR	226 Mb	47,837	1.15 Mb	NR	244 Mb	
		Sanger	140 kb	551	33.2 k	0.04×	5.6×									
		GA	200 bp	42	282 M	32.5×	76.8×	NR	2.6 kb	NR	204 Mb	NR	19 kb	NR	238 Mb	Assembly of only Sanger reads
		GA	400 bp	44	173 M	20.6×	94.4×									
		GA	2 kb	53	105 M	15.3×	286×	NR	12.5 kb	NR	190 Mb	NR	172 kb	NR	200 Mb	Assembly of only GA reads
Panda (*A. melanoleura*)/2.4 Gb	SOAPdenovo published 2010	GA	150	45	1.31 B	24.5×	43.3×									
		GA	500	67	917 M	25.5×	90.2×	200,604	36,728	434,635	2.25 Gb	81,469[e]	1.22 Mb	6.05 Mb	2.30 Gb	
		GA	2 kb	71	397 M	11.8×	192×									
		GA	5 kb	38	505 M	8.0×	533×									
		GA	10 kb	35	254 M	3.7×	571×									
Strawberry (*F. vesca*)/220 Mb	CABOG and Velvet announced	454	None	209	7.73 M	7.3×	—									
		454	None	368	787 M	13.2×	—	16,487	28,072	215,349	202 Mb	3263	1.44 Mb	4.1 Mb	214 Mb	
		454	2.5 kb	193	2.39 M	2.1×	6.9×									
		454	20 kb	236	1.58 M	1.7×	20×									
		GA	None	76	36 M	12.4×	—									
		SOLiD	2 kb	25	1.30 M	0.14×	6.4×									
Turkey (*M. gallopavo*)/1.1 Gb	CABOG announced	454	3 kb	180	6 M	1×	8×									
		454	20 kb	195	2 M	0.3×	18×	128,271	12,594	90 kb	931 Mb	26,917	1.5 Mb	9 Mb	NR	
		454	None	366	13 M	4×	—									
		GA	180 bp	74	200 M	13×	16×									
		GA	None	74	200 M	13×	—									

Results from de novo assembly of genomes by second-generation sequencing platforms. Summary of inputs and assembly results of recent genome assemblies using SGS reads.
[a]Status indicates when the assembly was published; "announced" assemblies have been described publicly but not yet published.
[b]The number of estimated genome size units contained in the sum of read lengths.
[c]The same value for the sum of lengths of fragments from which paired reads were sequenced.
[d]Contig total greater than scaffold total is largely attributable to "single haplotype contigs."
[e]Number of scaffolds includes single-contig scaffolds. There were 5201 multicontig scaffolds.
GA, Illumina Genome Analyzer; SOLiD, Applied Biosystems SOLiD System.

Assembly of large genomes using second-generation sequencing. *Genome Research* 2010
https://genome.cshlp.org/content/20/9/1165.full

三代拼接软件与工具



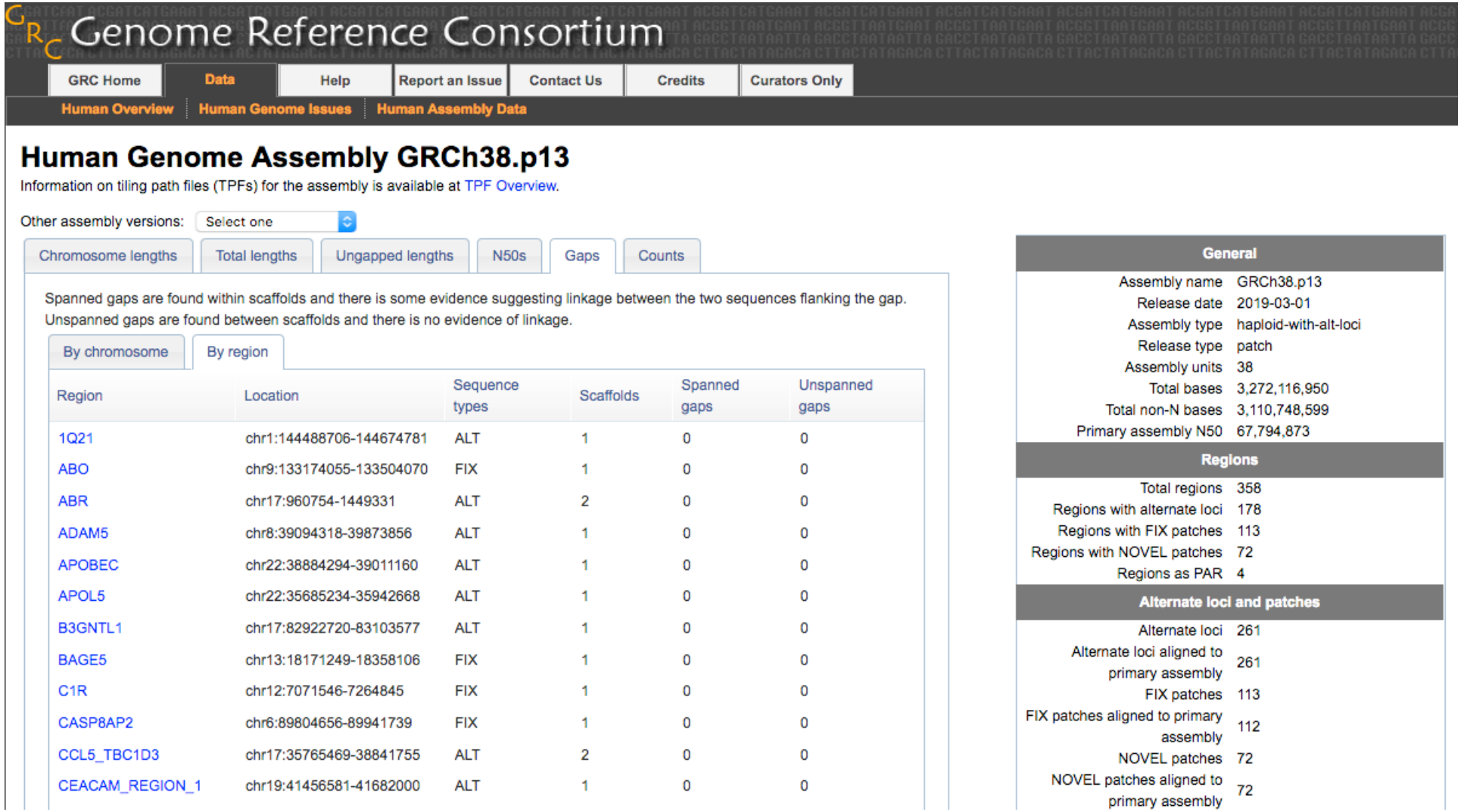
The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing
Briefings in Functional Genomics 2019 https://doi.org/10.1093/bfgp/ely037

人类基因组N50比较

	炎黄YH2.0	华夏HX1	北方汉族NH1.0	国际人类参考基因组 GRCh38
Population	Southern Chinese	Southern Chinese	Northern Han Chinese	European
Sequencing methods	HiSeq (fosmid)	PacBio + BioNano	PacBio + Bionano + 10×Genomics	Sanger (BAC + fosmid)
Assembly software	SOAPdenovo	FALCON	CANU + Supernova	NA
Scaffold N50 (Mb)	20.52	21.98	46.63	67.79
Contig N50 (Mb)	0.02	8.33	3.6	56.41
No. of scaffolds	125,643	5367	5574	735
No. of gaps	235,514	10,901	8484	999
Assembly size (bp)	2,911,235,363	2,934,084,193	2,892,287,479	3,209,286,105

Whole Genome Analyses of Chinese Population and De Novo Assembly of A Northern Han Genome.
Genomics, Proteomics & Bioinformatics 2019 https://doi.org/10.1016/j.gpb.2019.07.002

国际人类参考基因组



Genome Reference Consortium

GRC Home | Data | Help | Report an Issue | Contact Us | Credits | Curators Only

Human Overview | Human Genome Issues | Human Assembly Data

Human Genome Assembly GRCh38.p13

Information on tiling path files (TPFs) for the assembly is available at TPF Overview.

Other assembly versions: Select one

Chromosome lengths | Total lengths | Ungapped lengths | N50s | Gaps | Counts

Spanned gaps are found within scaffolds and there is some evidence suggesting linkage between the two sequences flanking the gap.
Unspanned gaps are found between scaffolds and there is no evidence of linkage.

By chromosome | By region

Region	Location	Sequence types	Scaffolds	Spanned gaps	Unspanned gaps
1Q21	chr1:144488706-144674781	ALT	1	0	0
ABO	chr9:133174055-133504070	FIX	1	0	0
ABR	chr17:960754-1449331	ALT	2	0	0
ADAM5	chr8:39094318-39873856	ALT	1	0	0
APOBEC	chr22:38884294-39011160	ALT	1	0	0
APOL5	chr22:35685234-35942668	ALT	1	0	0
B3GNTL1	chr17:82922720-83103577	ALT	1	0	0
BAGE5	chr13:18171249-18358106	FIX	1	0	0
C1R	chr12:7071546-7264845	FIX	1	0	0
CASP8AP2	chr6:89804656-89941739	FIX	1	0	0
CCL5_TBC1D3	chr17:35765469-38841755	ALT	2	0	0
CEACAM_REGION_1	chr19:41456581-41682000	ALT	1	0	0

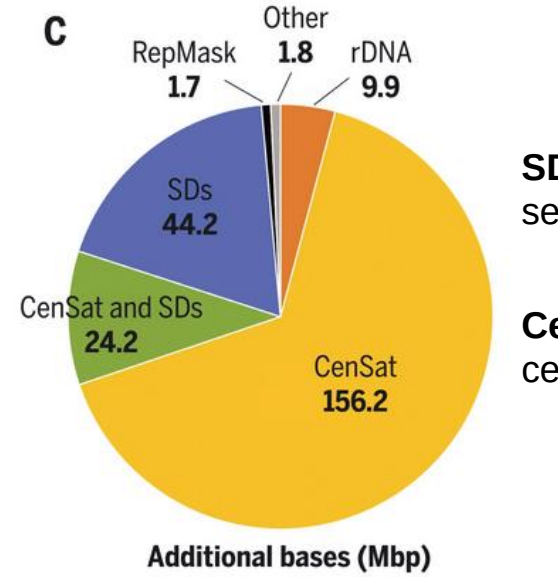
General	
Assembly name	GRCh38.p13
Release date	2019-03-01
Assembly type	haploid-with-alt-loci
Release type	patch
Assembly units	38
Total bases	3,272,116,950
Total non-N bases	3,110,748,599
Primary assembly N50	67,794,873
Regions	
Total regions	358
Regions with alternate loci	178
Regions with FIX patches	113
Regions with NOVEL patches	72
Regions as PAR	4
Alternate loci and patches	
Alternate loci	261
Alternate loci aligned to primary assembly	261
FIX patches	113
FIX patches aligned to primary assembly	112
NOVEL patches	72
NOVEL patches aligned to primary assembly	72

https://www.ncbi.nlm.nih.gov/grc/human/data

人类基因组完整图

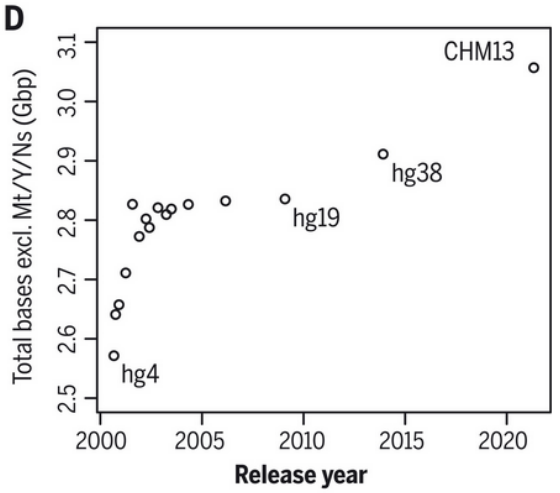
STATISTICS	GRCH38	T2T-CHM13
Assembled bases (Gbp)	2.92	3.05
Unplaced bases (Mbp)	11.42	0
Gap bases (Mbp)	120.31	0
Number of contigs	949	24
Contig NG50 (Mbp)	56.41	154.26
Segmental duplication bases (Mbp)	151.71	201.93
Number of segmental duplications	24097	41528
Percentage of repeats (%)	51.89	53.94
Repeat bases (Mbp)	1,516.37	1,647.81

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



SD
segmental duplication

CenSat
centromeric satellite



中国汉族人高质量参考基因组——T2T-YAO

首例汉族人二倍体高质量参考基因组（44+XY）：T2T-YAO（唐尧）

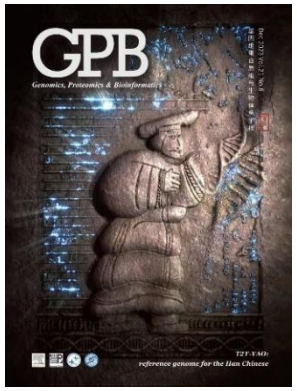
样本采集—代表性

样品采自明初大槐树移民出发地
——尧都临汾，洪洞县

测序拼接—精确性

优于2022年美国NIH发布的参考基因组CHM13，质量世界第一

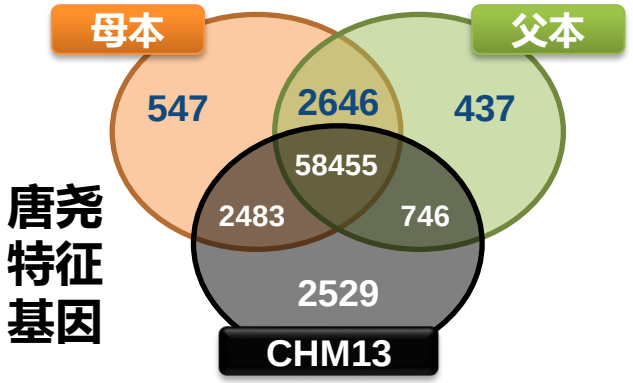
	YAO-hp	CHM13
染色体核型	22+XY	22+X
完整性	99.6%	99.0%
间隙数	0	0
质量值（QV）	74.69	73.94
错误间隔（Mb）	29.5	24.8
估计错误数	104	123



基因序列—差异性

发现众多中国人差异基因
从基因组特征回答“何以中国人”

- 差异序列：~330Mb
- 差异基因：>3000个
- 差异结构：数万个



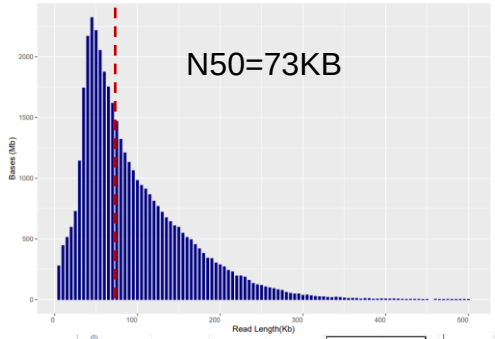
T2T-YAO高质量组装

- PacBio HiFi测序深度 92×
- 超长ONT测序深度 336×

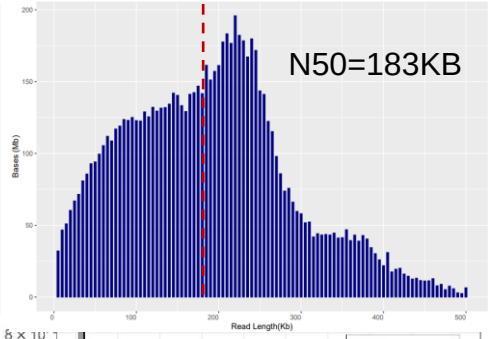
  N50=153KB

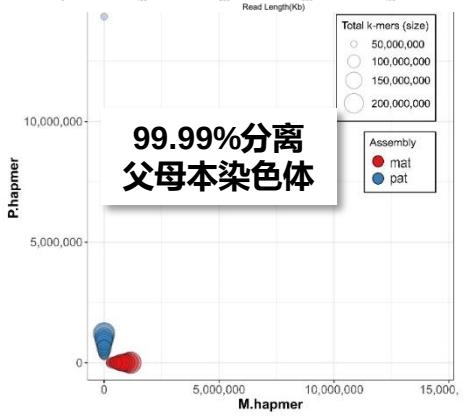
  >250kb reads占总测序量的 12%

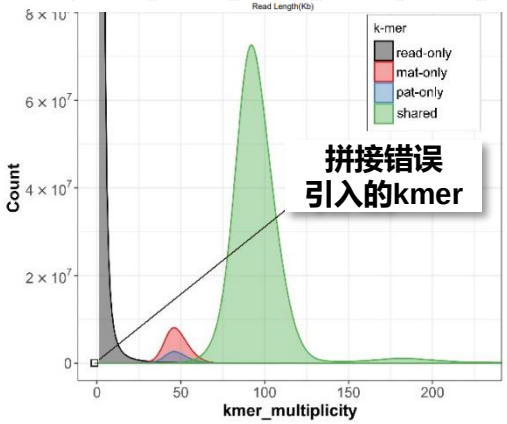
ONT ultra long kit



优化建库







	YAO		HG002[1]		CN1[2]	
	Mat	Pat	Mat	Pat	Mat	Pat
Chr	22+X	22+Y	22+X	22+Y	22+X	22+Y
Assembled bases (Gbp)	3.02	2.92	3.06	2.96	3.04	2.94
Gaps	0	0	109	117	0	0
Gap bases (Mbp)	0	0	16.4	25.3	0	0
N50(Mbp)	155.06	145.07	62.88	81.56	INA	INA
QV	**70.49**	**72.28**	**59.30**	**58.60**	**60.11**	**59.37**
Estimated errors	**269**	**173**	**3620**	**4070**	**2962**	**3400**
Completeness %(haploid)	99.65	99.59	99.70	99.00	INA	INA
False duplications %	0.30	0.20	0.49	0.37	INA	INA
Haplotype switch errors %	0.019	0.011	0.028	0.018	INA	INA
Collapsed bases (Mbp)						
with common repeats	4.71	4.96	18.53	17.64	INA	INA
without common repeats	0.56	0.35	7.61	7.81	INA	INA
rRNA copy number	79	149	INA	INA	132	117

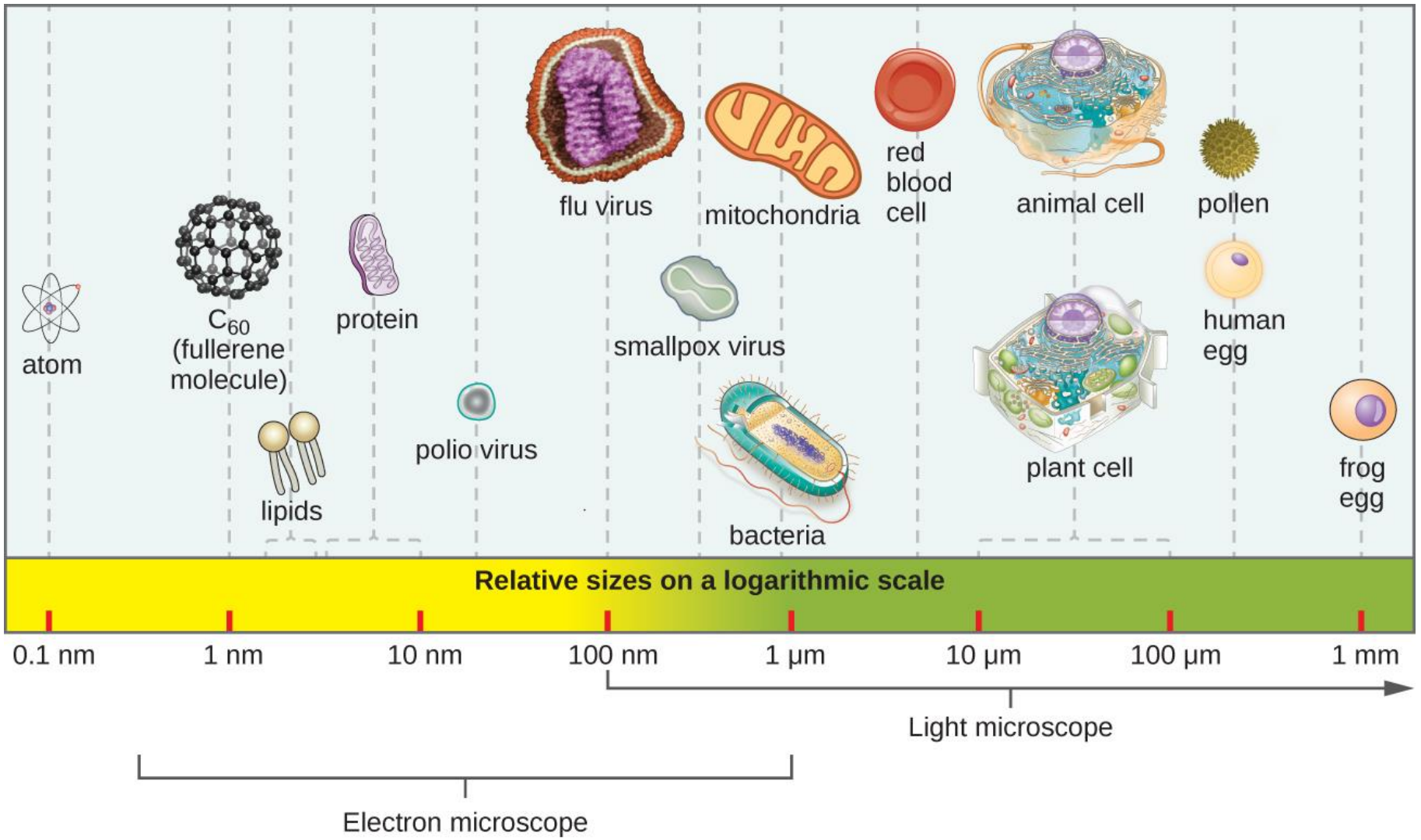
[1] Jarvis et al. ***Nature*** 2022; [2] Yang et al. *Cell Research* 2003

单细胞测序

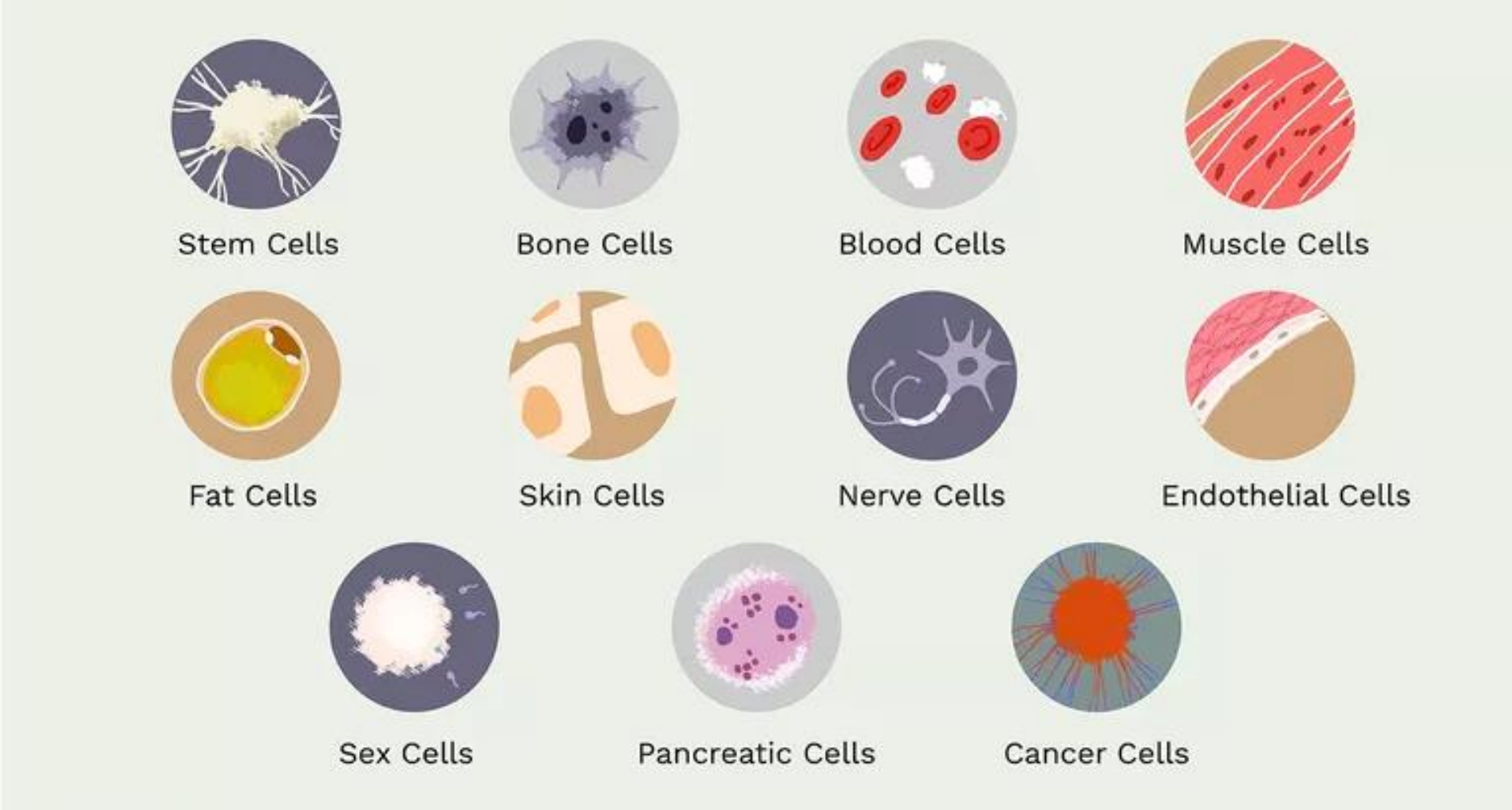
单细胞测序技术


2017年《Nature》封面
被MIT列为10大突破技术
10
BREAKTHROUGH
TECHNOLOGIES

How big are our cells?



Human Cell Types



https://www.thoughtco.com/types-of-cells-in-the-body-373388

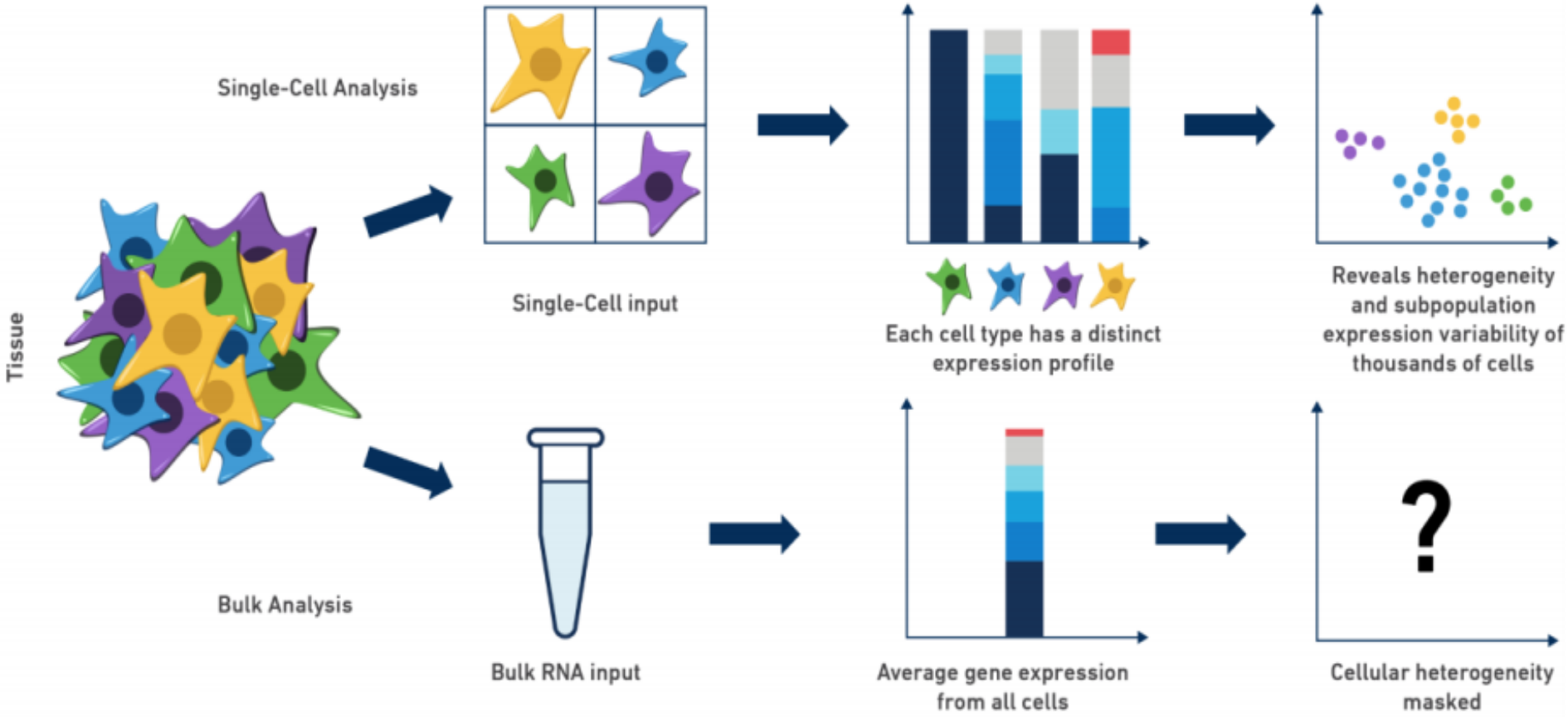
群体细胞测序 vs. 单细胞测序

Bulk Sequencing

Single-Cell Sequencing

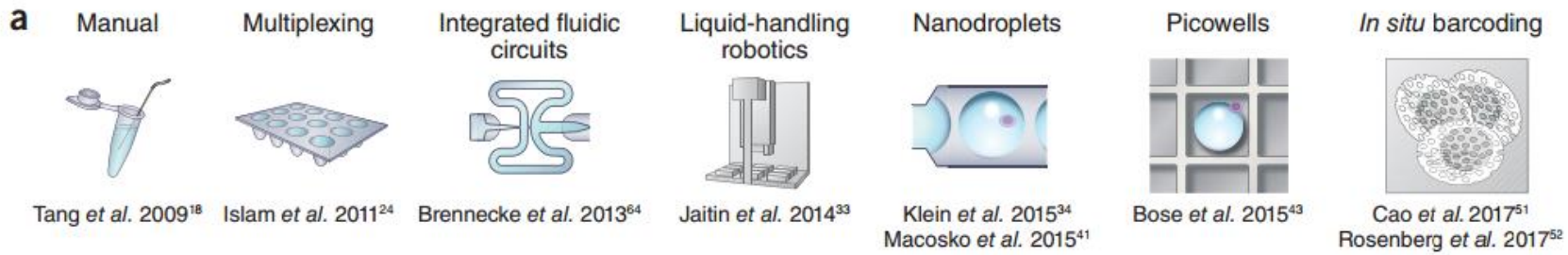
细胞异质性、多样性以及单细胞内分子间互作关系

群体细胞测序 vs. 单细胞测序



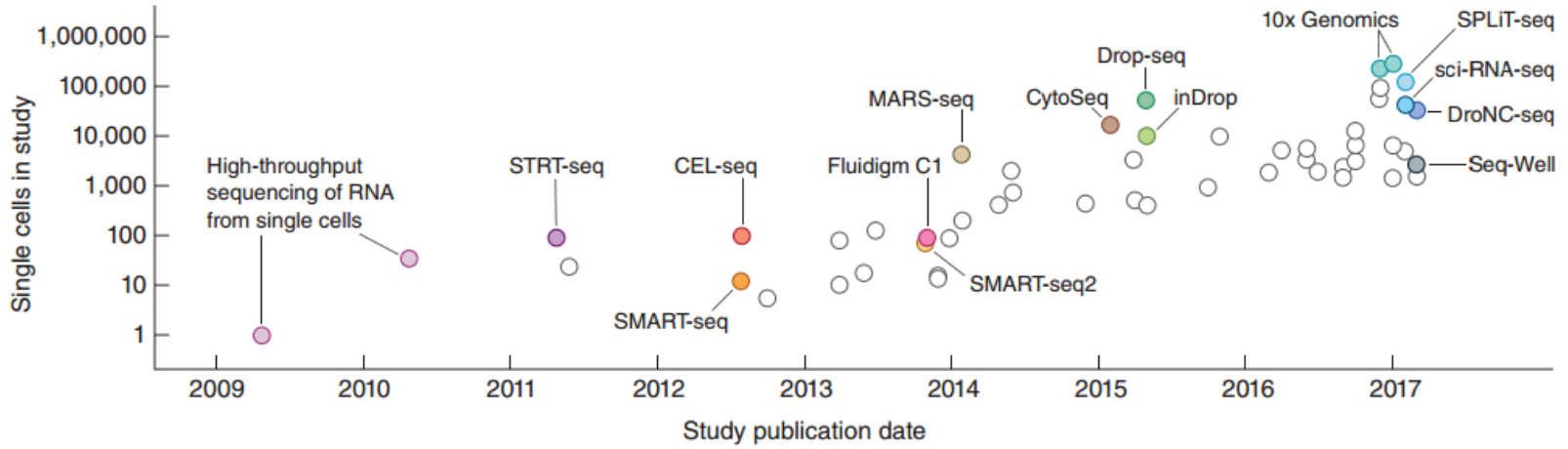
https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

单细胞捕获



通量	单细胞捕获方法	技术平台	细胞选择	细胞分析	样品体积	细胞数量	细胞完整性
低	显微操作法	显微镜/操作仪	有偏向	显微镜图像/荧光信号	微升μl	10^1	★★★★★
低	口吸管技术	显微镜/毛细管	有偏向	显微镜图像/荧光信号	微升μl	10^1	★★★★★
低	激光显微切割法	激光显微切割仪	有偏向	显微镜图像/荧光信号	固态	10^2	★★
高	流体细胞分选	流式细胞仪	有偏向	散射光/荧光信号	纳升nl	10^2	★★★
高	微孔芯片技术	芯片	无偏向	可结合成像系统	纳升(nl-pl)	10^2-10^3	★★★★
高	微流体技术	微流控	无偏向	无	纳升(nl-pl)	10^3	★★★★

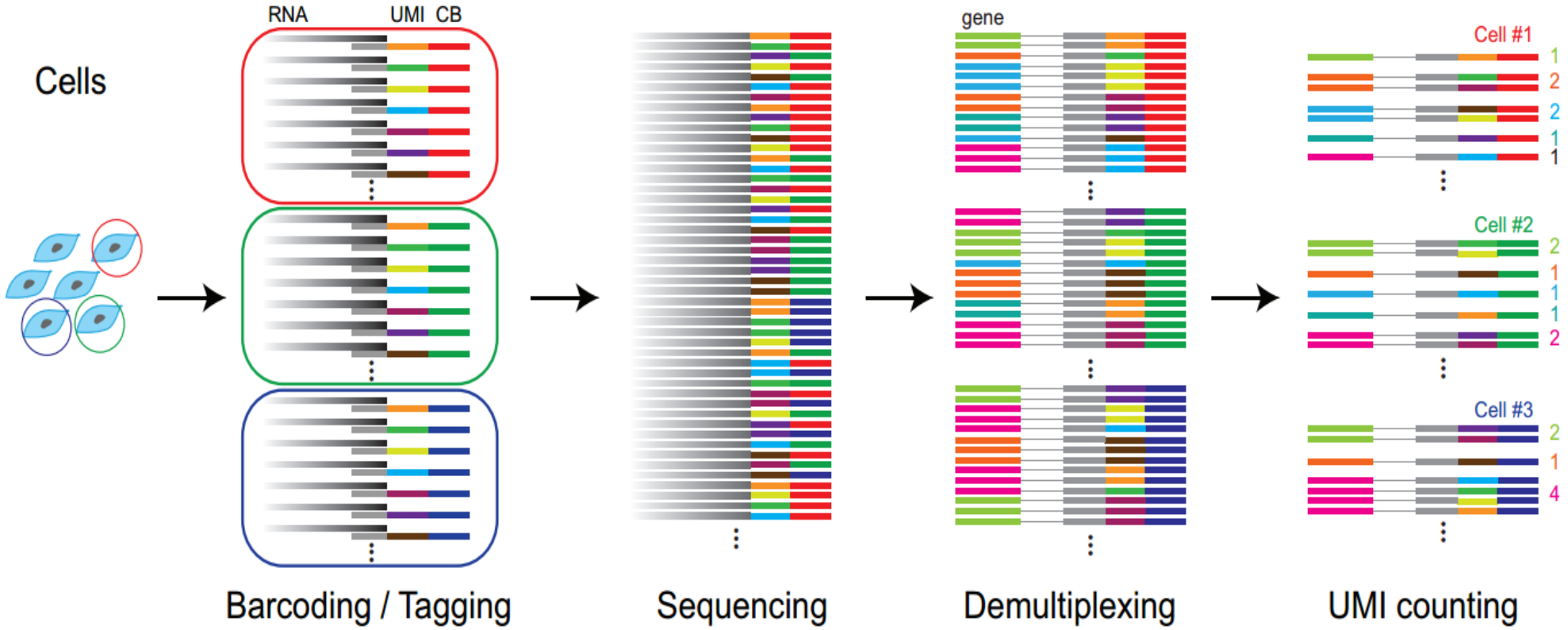
单细胞转录组测序



Protocol example	*C1 (SMARTer)*	*Smart-seq2*	*MATQ-seq*	*MARS-seq*	*CEL-seq*	*Drop-seq*	*InDrop*	*Chromium*	*SEQ-well*	*SPLIT-seq*
Transcript data	Full length	Full length	Full length	3′-end counting	3′-end counting	3′-end counting	3′-end counting	3′-end counting	3′-end counting	3′-end counting
Platform	Microfluidics	Plate-based	Plate-based	Plate-based	Plate-based	Droplet	Droplet	Droplet	Nanowell array	Plate-based
Throughput (number of cells)	10^2–10^3	10^2–10^3	10^2–10^3	10^2–10^3	10^2–10^3	10^3–10^4	10^3–10^4	10^3–10^4	10^3–10^4	10^3–10^5
Typical read depth (per cell)	10^6	10^6	10^6	10^4–10^5	10^4–10^5	10^4–10^5	10^4–10^5	10^4–10^5	10^4–10^5	10^4
Reaction volume	Nanoliter	Microliter	Microliter	Microliter	Nanoliter	Nanoliter	Nanoliter	Nanoliter	Nanoliter	Microliter
Reference	[63]	[57]	[39]	[10]	[64]	[45]	[46]	[47]	[101]	[38]

Exponential scaling of single-cell RNA-seq in the last decade. *Nature Protocols* 2018
A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* 2017

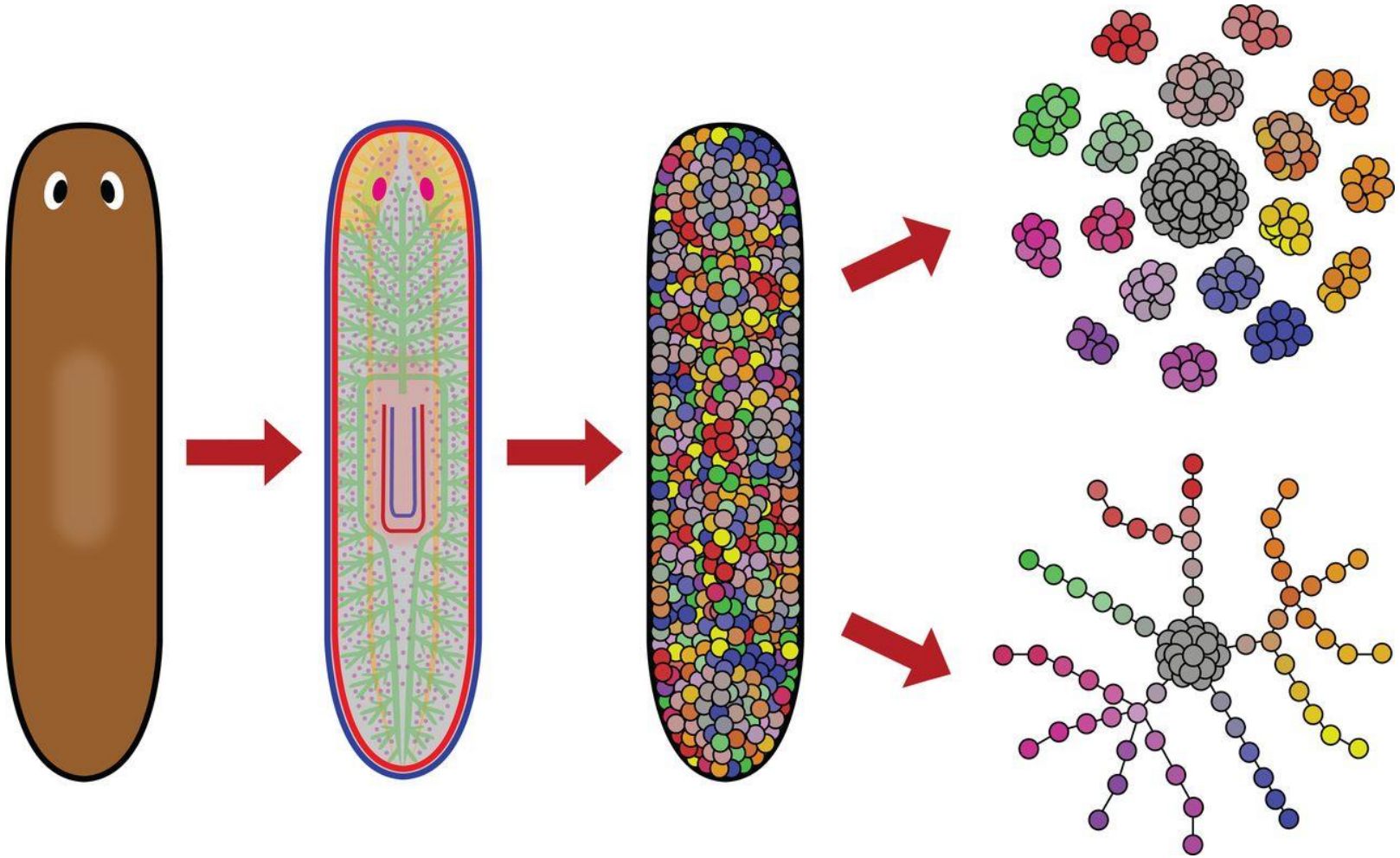
Barcode区分细胞, UMI区分转录本



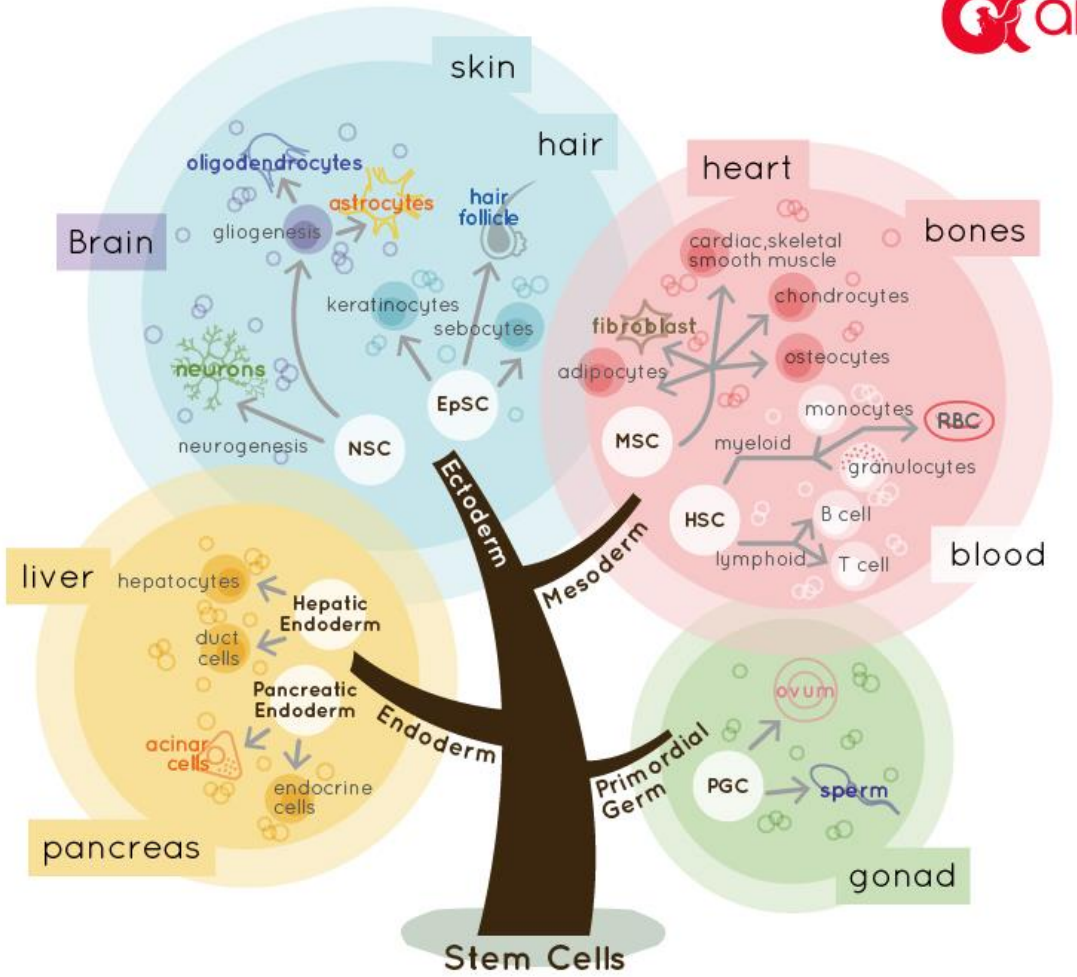
UMI: Unique Molecular Identifier
CB: Cell Barcode

Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular Cell* 2018

A lineage tree by single-cell transcriptomics



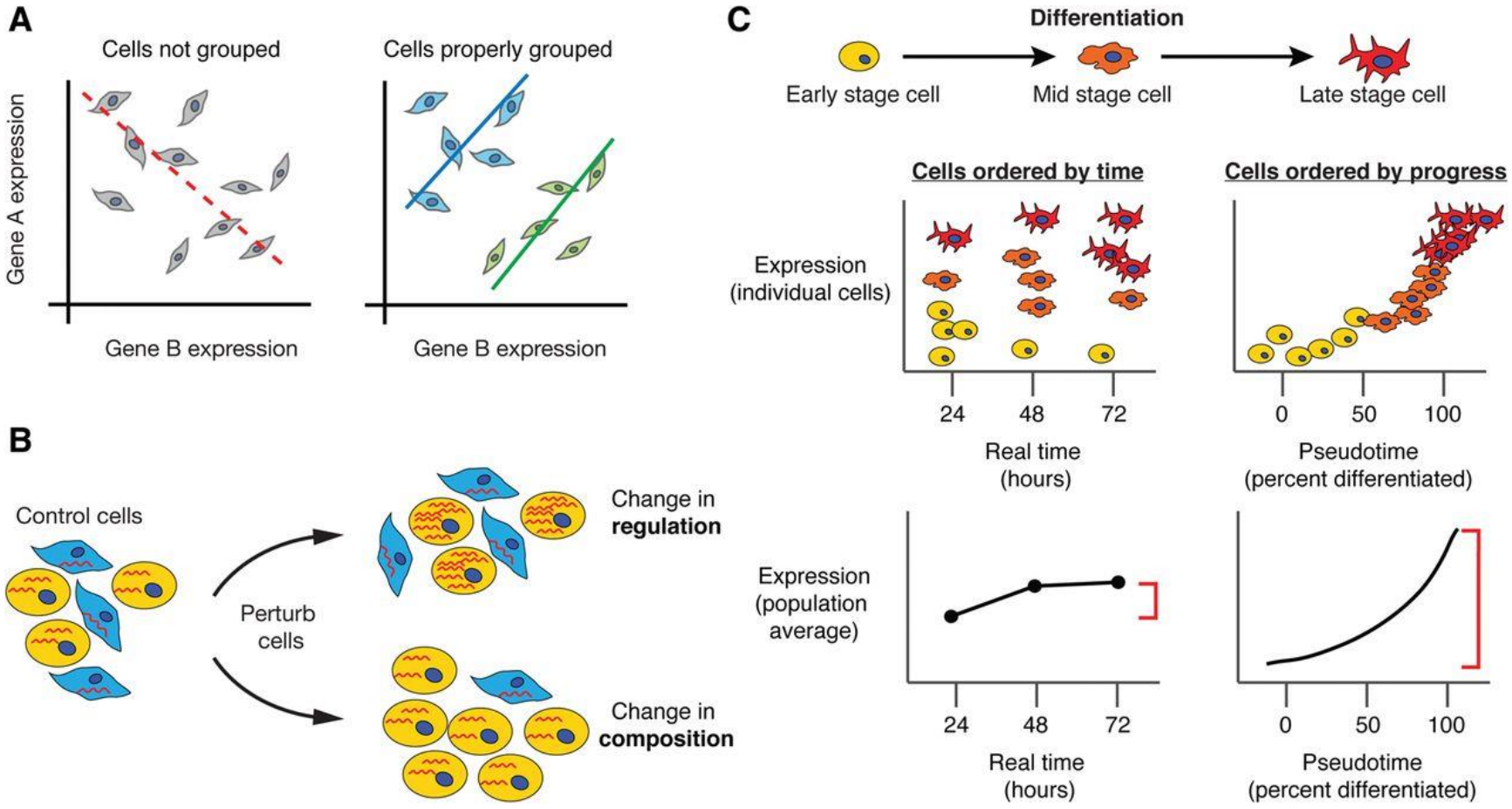
Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* 2018



The tree of stem cell differentiation

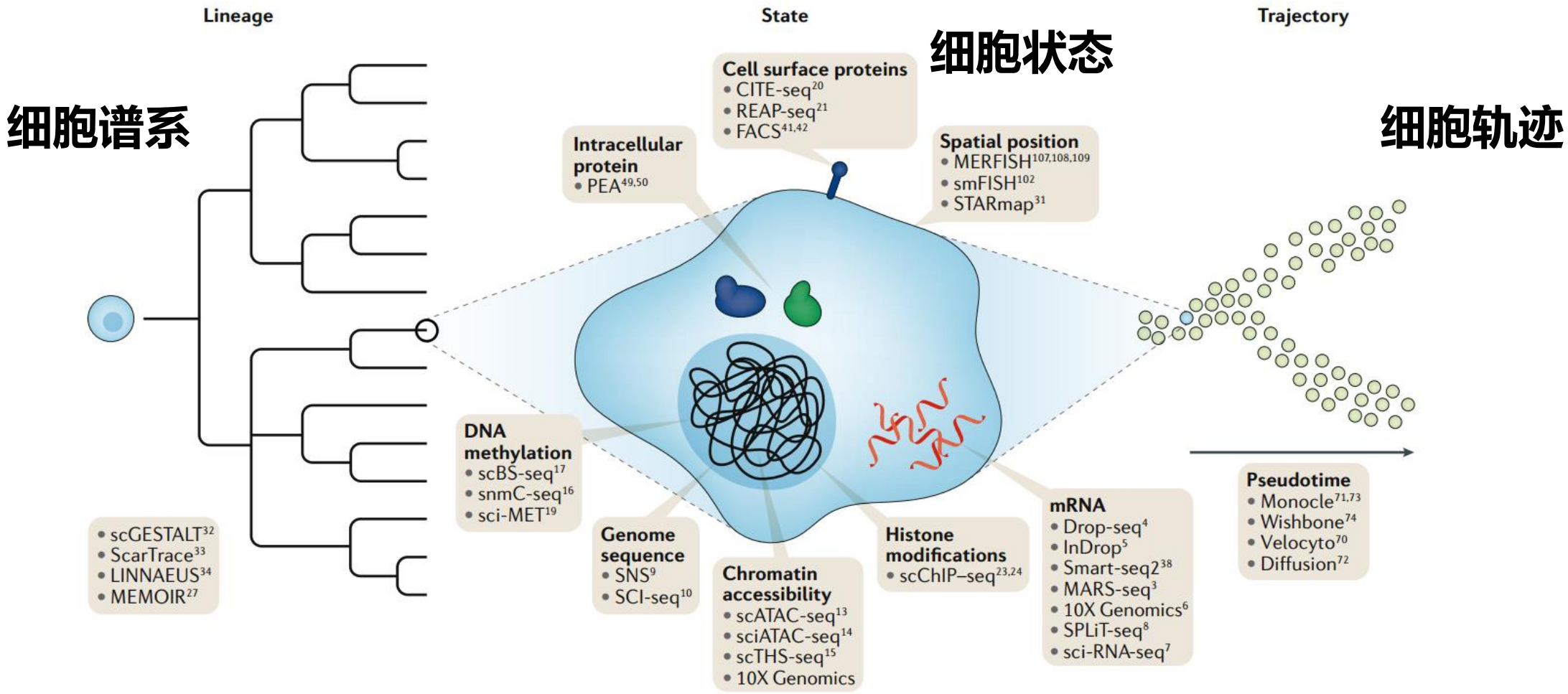
Renew Differentiate Specialize

Cell types and states



Defining cell types and states with single-cell genomics. *Genome Research* 2015

Single-cell Genomics



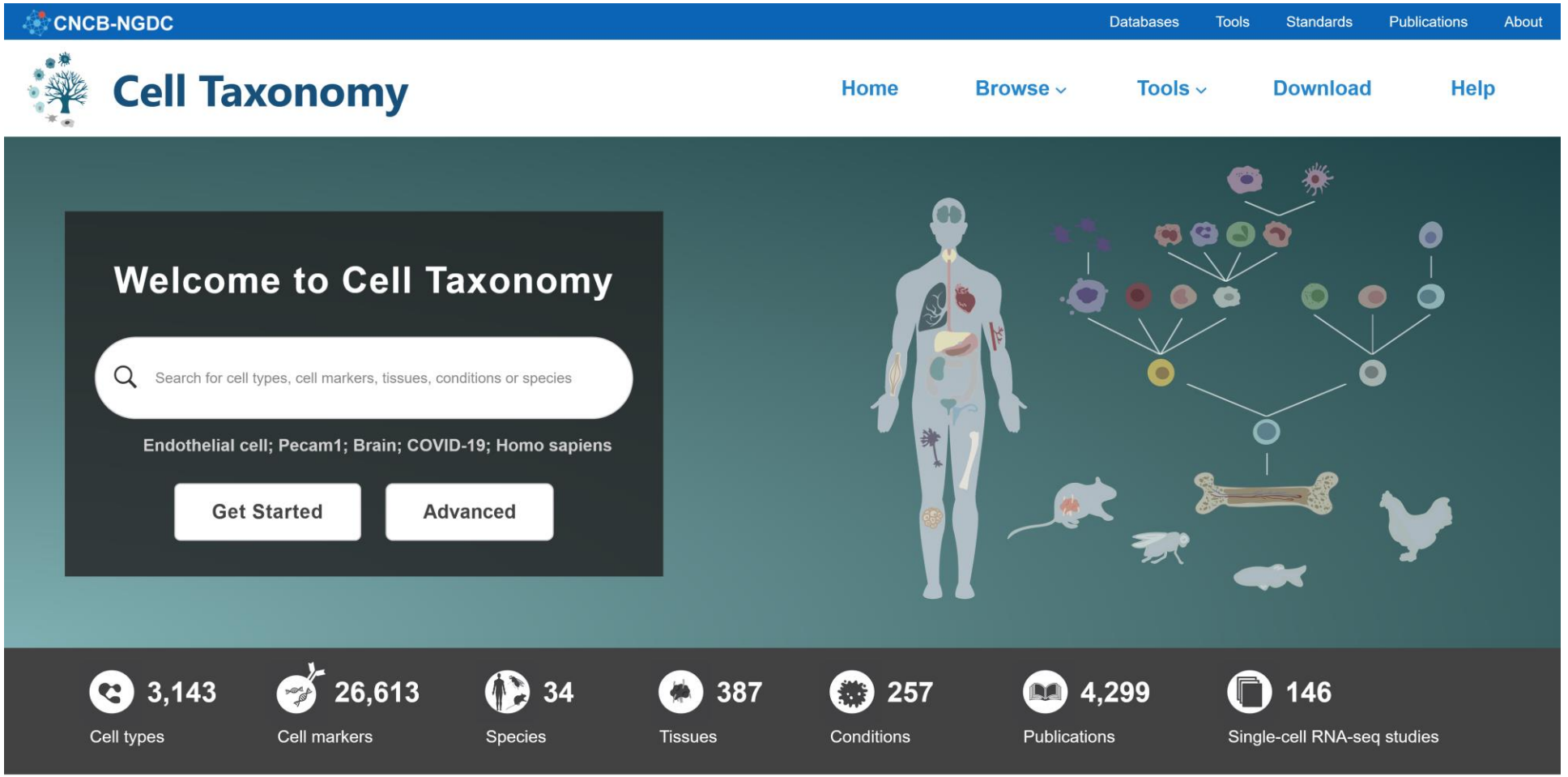
Integrative single-cell analysis. *Nature Reviews Genetics* 2019

人类细胞图谱计划 (Human Cell Atlas)

人体～37万亿个细胞

https://www.humancellatlas.org

Cell Taxonomy



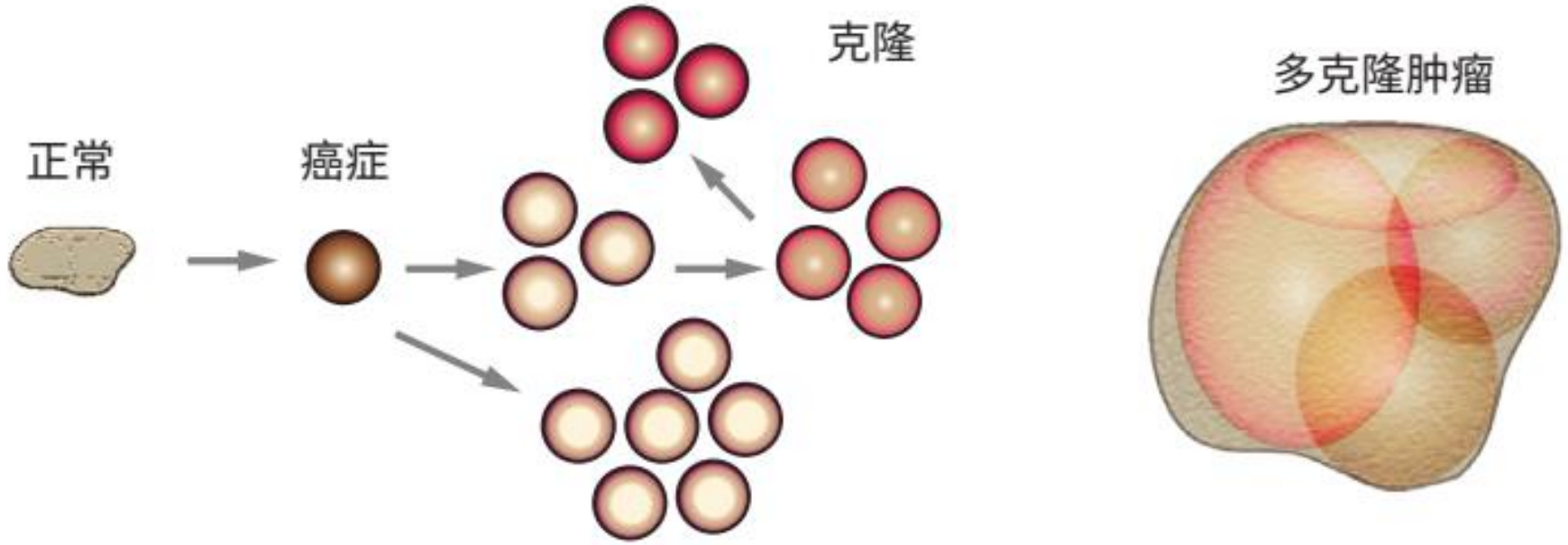
Cell Taxonomy: a curated repository of cell types with multifaceted characterization. *Nucleic Acids Res* 2023

https://ngdc.cncb.ac.cn/celltaxonomy/

单细胞测序技术的应用

Single-cell Cancer Genomics

肿瘤内异质性：体细胞突变的进行性累积导致异质性多克隆肿瘤，晚期肿瘤可能含有大量独特的亚克隆，其中不同的克隆可能对治疗反应不同。



Cancer genomics: one cell at a time. *Genome Biol* 2014

Single-cell Metagenomics

单细胞宏基因组学可帮助我们了解每种微生物对其周围环境的作用

Analysis of single-cell genome sequences of bacteria and archaea *Emerging Topics in Life Sciences* 2017

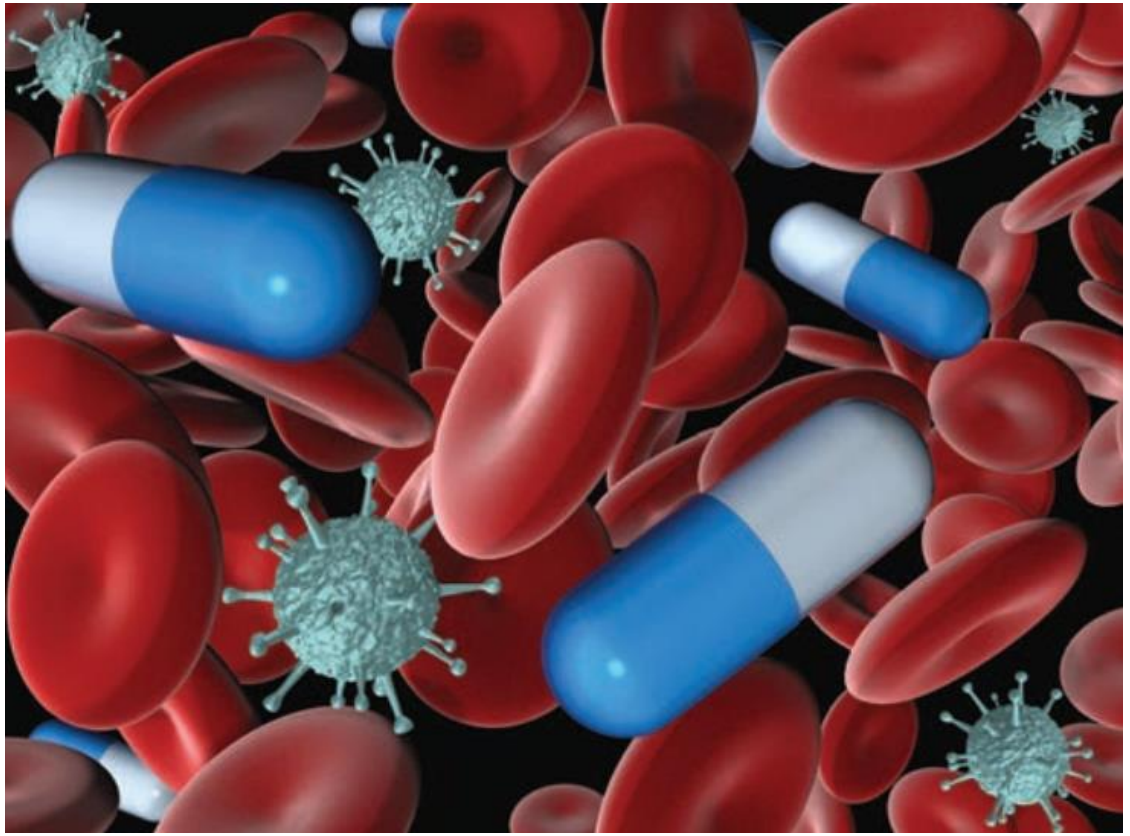
Single-cell Neurobiology

“单细胞RNA测序是一项令人兴奋的新技术，能够分析单个细胞的转录组，非常适用于应对中枢神经系统固有的复杂性和动态特征。”

——Ofengeim 等，2017 年

Single-Cell RNA Sequencing: Unraveling the Brain One Cell at a Time. *Trends Mol Med* 2017

Single-cell aided drug discovery



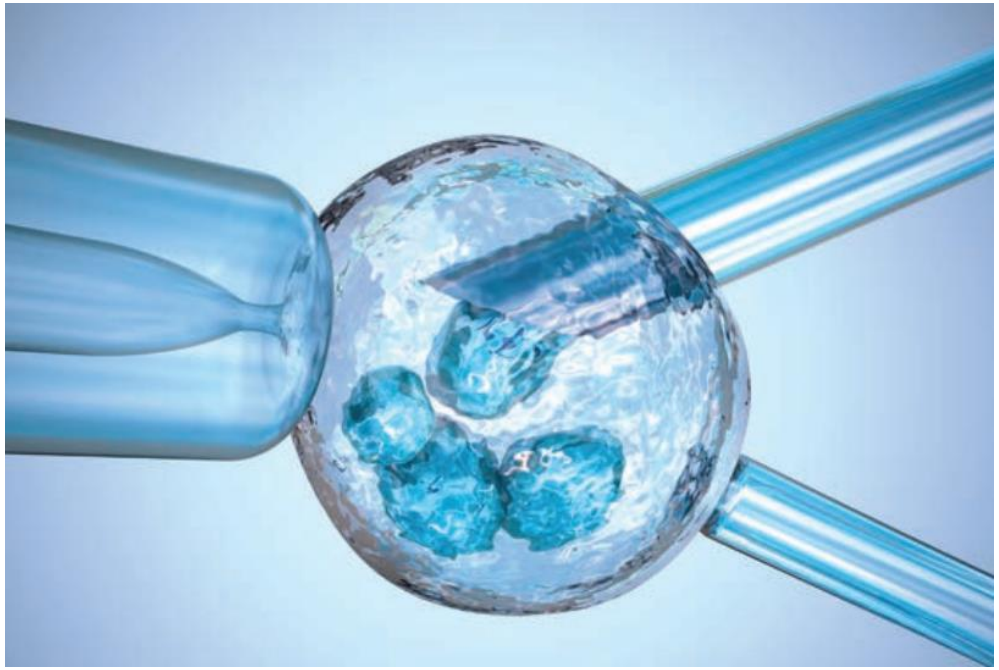
单细胞基因组可在药物研究和设计中带来更高效率

Single-cell analysis tools for drug discovery and development. *Nat Rev Drug Discov* 2016

产前诊断

“单细胞测序能够基因组、外显子组、转录组或表观基因组进行全局测序，进而对早期胚胎进行全面分析。”

——Zhu et.al，2017 年



单细胞测序技术可在植入前筛查胚胎是否有染色体异常，可加速研究、改善胚胎状态的早期检测，并确保能够移植健康的体外受精胚胎。

Next-generation molecular diagnosis: single-cell sequencing from bench to bedside. *Cell Mol Life Sci* 2017

发展趋势

长读长测序 (long-read sequencing)

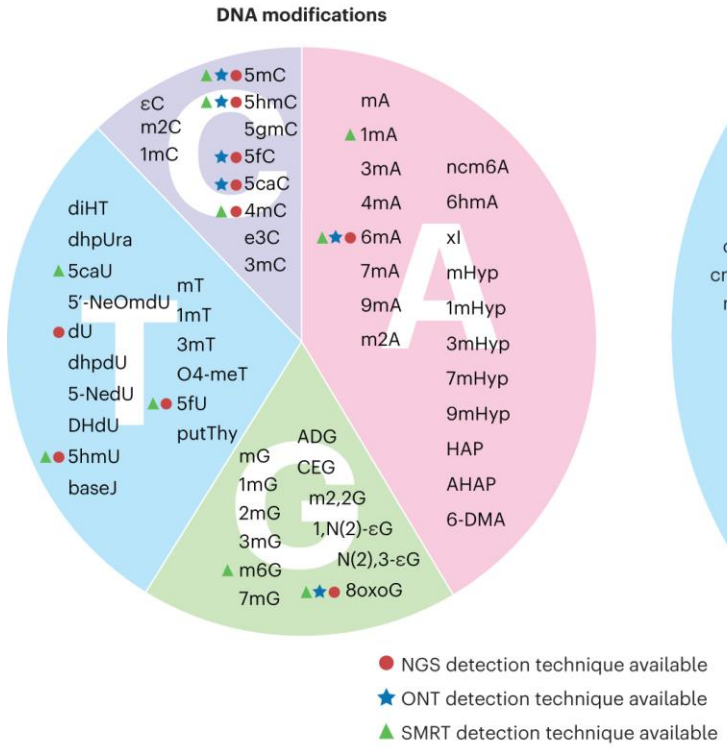
nature › nature methods › focus

Focus | 12 January 2023

Method of the Year 2022: Long-read sequencing

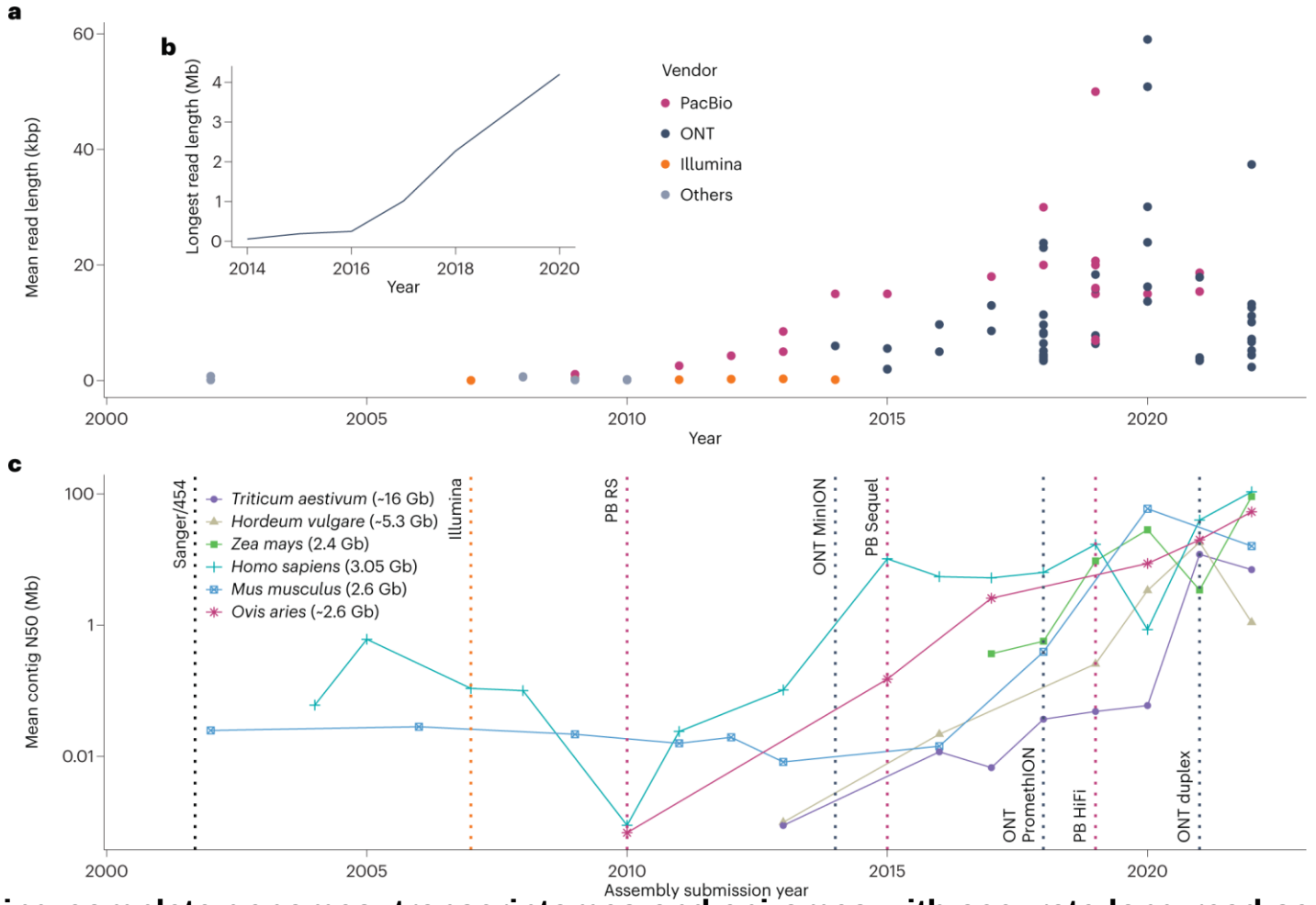
Long-read sequencing is our choice for Method of the Year 2022. This class of technologies is enabling a more complete reading and understanding of genomes, epigenomes and transcriptomes. The special collection also includes Methods to Watch – our predictions of techniques that will become more impactful in the near future.

As long-read sequencing technologies continue to advance, the possibility of obtaining maps of DNA and RNA modifications at single-molecule resolution has become a reality.



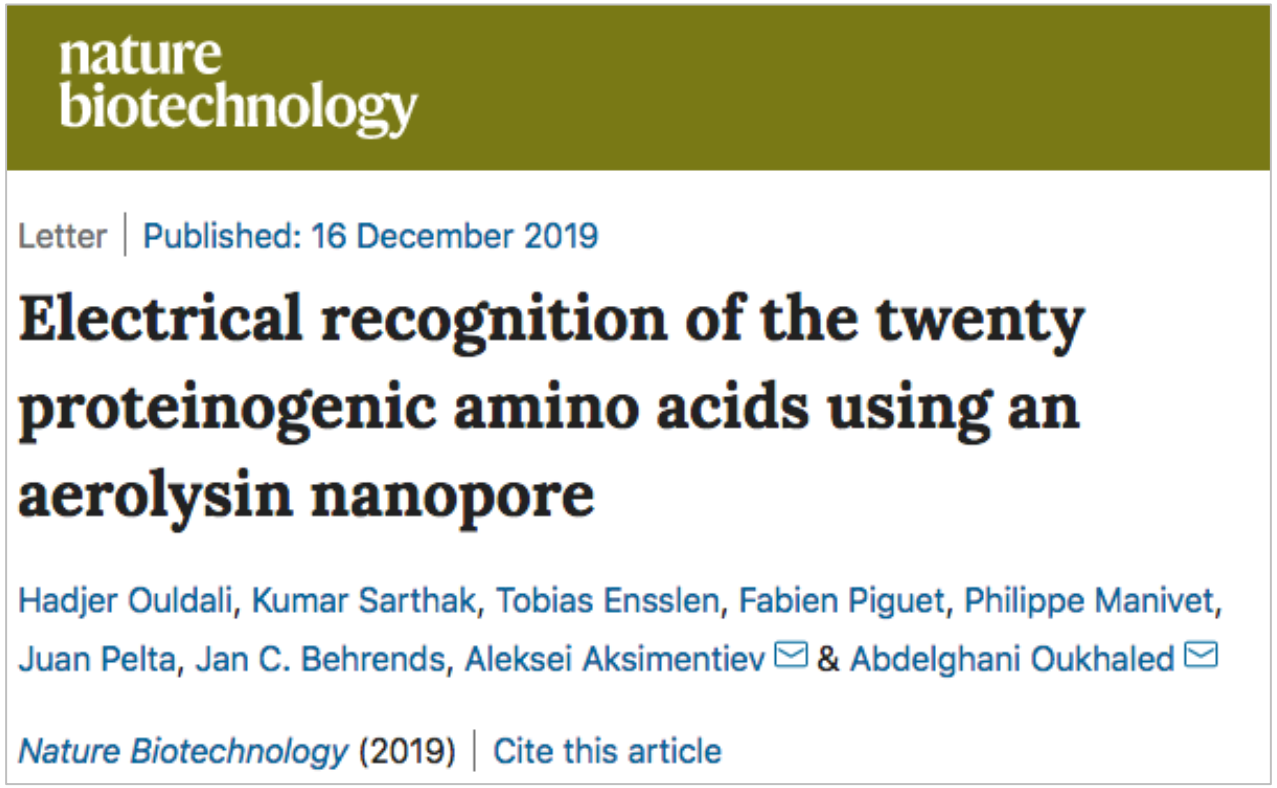
https://www.nature.com/articles/s41592-022-01724-8

Improved long-read sequencing



Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing.
Nature Methods 2023 https://www.nature.com/articles/s41592-022-01716-8

纳米孔蛋白质测序



nature biotechnology

Letter | Published: 16 December 2019

Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore

Hadjer Ouldali, Kumar Sarthak, Tobias Ensslen, Fabien Piguet, Philippe Manivet, Juan Pelta, Jan C. Behrends, Aleksei Aksimentiev ✉ & Abdelghani Oukhaled ✉

Nature Biotechnology (2019) | Cite this article

https://doi.org/10.1038/s41587-019-0345-2

基于纳米孔技术的蛋白质组学测序系统，能够对气溶素纳米孔中所有20种天然蛋白质氨基酸中的15种进行识别

RNA直测（Direct RNA Sequencing）

Unlike traditional RNA-Seq techniques, long-read nanopore RNA sequencing allows ***accurate quantification and complete, full-length characterisation*** *of native RNA or cDNA without fragmentation or amplification – streamlining analysis and* ***removing potential sources of bias****. Direct RNA sequencing also enables the* ***identification of base modifications*** *alongside nucleotide sequence.*

nature methods

Article | Published: 18 November 2019

Nanopore native RNA sequencing of a human poly(A) transcriptome

Rachael E. Workman, Alison D. Tang, Paul S. Tang, Miten Jain, John R. Tyson, Roham Razaghi, Philip C. Zuzarte, Timothy Gilpatrick, Alexander Payne, Joshua

nature communications

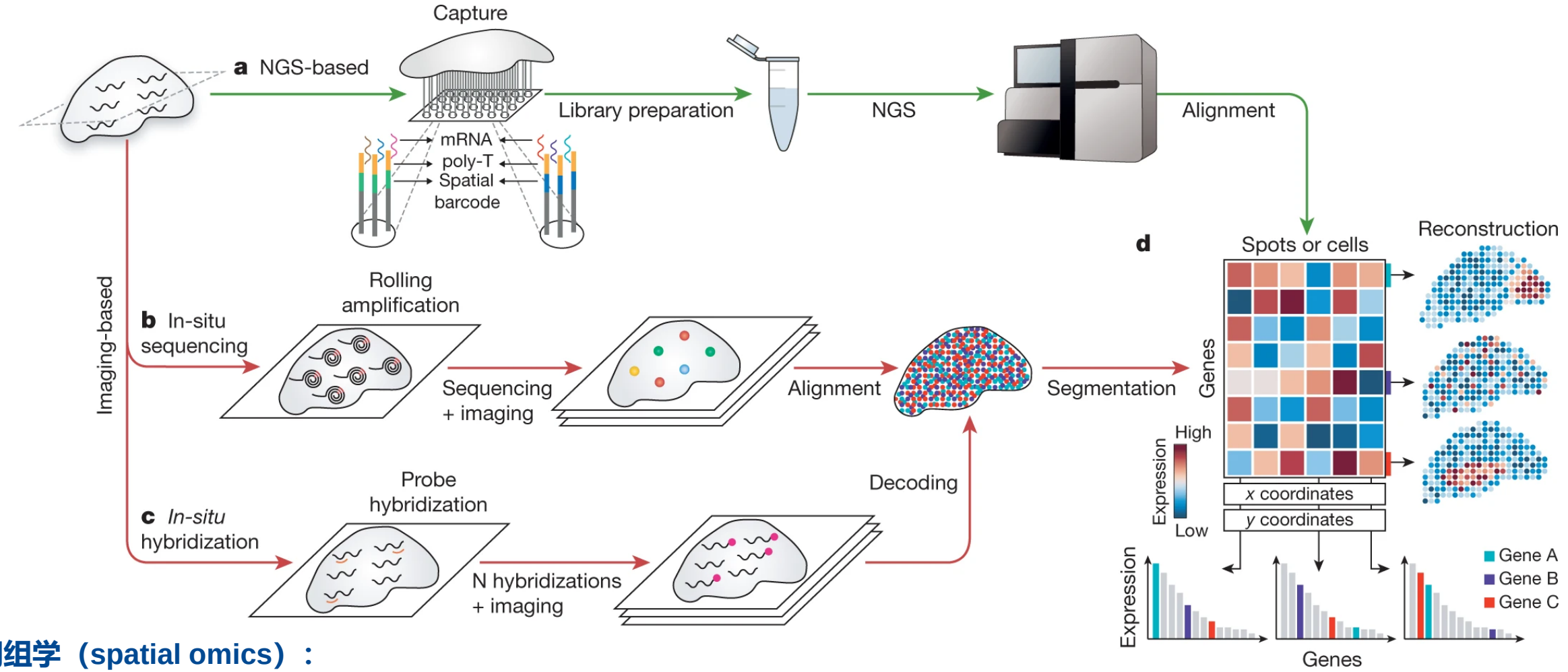
Article | Open Access | Published: 31 July 2019

A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes

Charlotte Soneson, Yao Yao, Anna Bratus-Neuenschwander, Andrea Patrignani, Mark D. Robinson & Shobbir Hussain

Nature Communications **10**, Article number: 3359 (2019) | Cite this article

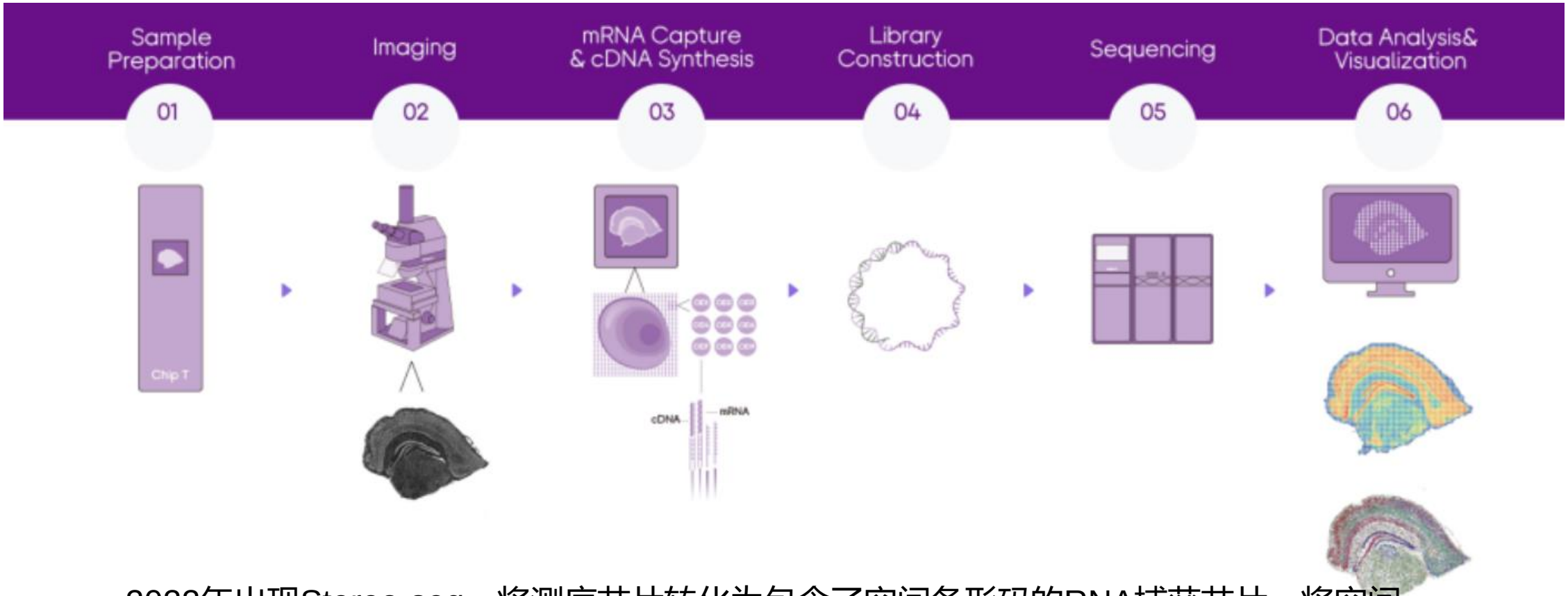
空间转录组技术



空间组学（spatial omics）：
保留细胞或组织的原位空间位置信息，与基因表达信息结合

Exploring tissue architecture using spatial transcriptomics. *Nature* 2021
https://www.nature.com/articles/s41592-022-01409-2

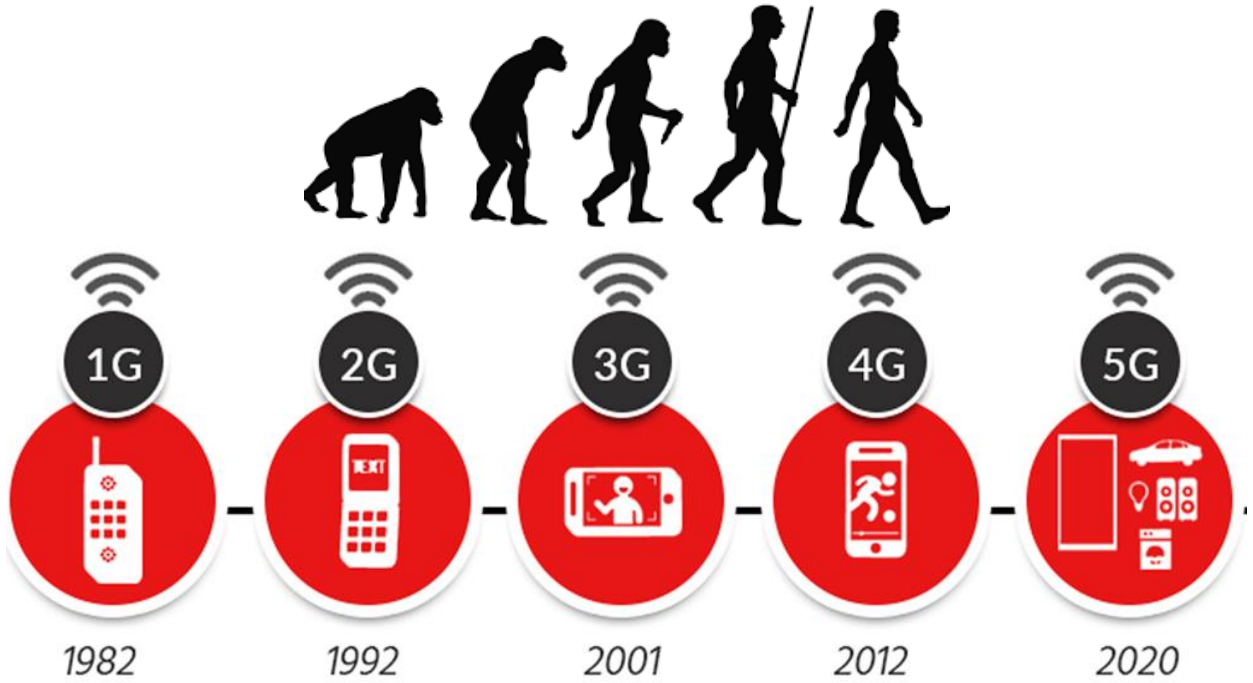
空间增强分辨率组学测序Stereo-seq



2022年出现Stereo-seq，将测序芯片转化为包含了空间条形码的RNA捕获芯片，将空间组学测序技术的分辨率提高到亚细胞水平（500nm），且包含厘米级别的捕获区域

https://www.bgi.com/global/service/spatial-transcriptome-stereo-seq

测序技术发展日新月异



测序平台选择

物种、目的、读长、质量、通量、测序深度、经费、时间

Work smarter – be imaginative and what seems impossible today can be the standard tomorrow

Summary

- 遗传图谱与物理图谱的定义和区别
- DNA测序技术发展：一代、二代、三代，各自特点与不足
- 两种测序策略：逐步克隆法、全基因组霰弹法，各自优缺点
- 基因组测序与拼接的三种方式：*de novo*、reference-guided、re-sequencing
- 基因组拼接：read、contig、scaffold
- 拼接质量评估：N50、contig个数、gap数等
- 单细胞测序技术、长读长测序技术

Further Reading

- **Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing**. *Nature Methods* 2023
- **Fast and accurate long-read assembly with wtdbg2.** *Nature Methods* 2020
- **Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads**. *Nature Communications* 2019
- **Integrative single-cell analysis**. *Nature Reviews Genetics* 2019
- **Sequencing depth and coverage: key considerations in genomic analyses**. *Nature Reviews Genetics* 2014
- **DNA sequencing technologies: 2006–2016.** *Nature Protocols* 2017
- **Single-cell genome sequencing: current state of the science**. *Nature Reviews Genetics* 2016
- **Sequence assembly demystified.** *Nature Reviews Genetics* 2013
- **RNA-Seq a revolutionary tool for transcriptomics**. *Nature Reviews Genetics* 2009