DNA 测序技术的发展

三代测序的原理与应用

王强

October 11, 2017

南京大学生命科学学院

Outline

一代: Sanger 法电泳测序

二代: (短读长的) 高通量测序

三代: (长读长的) 单分子测序

总结

一代: Sanger 法电泳测序

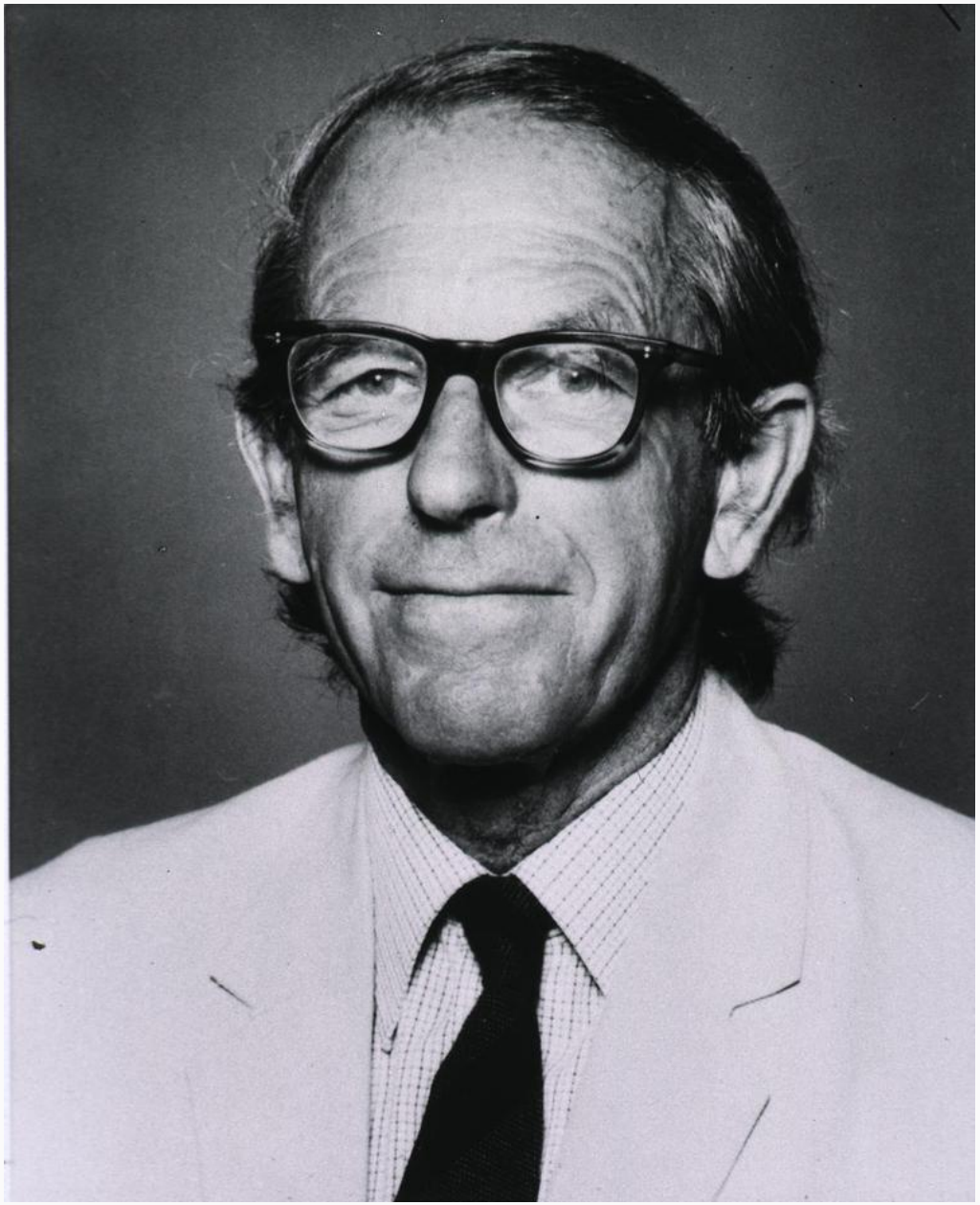


Figure 1. Frederick Sanger

- 1955 年, 第一个蛋白质测序, 胰岛素
- 1958 年, 诺贝尔化学奖
- 1975 年, 双脱氧法, ddNTP
- 1977 年, 第一个基因组, ϕ–X174 噬菌体
- 1980 年, 再度获得诺贝尔化学奖

人类基因组计划

- 1985 年, 美国能源部正式提出人类基因组测序.
- 1990 年, 正式启动人类基因组测序.

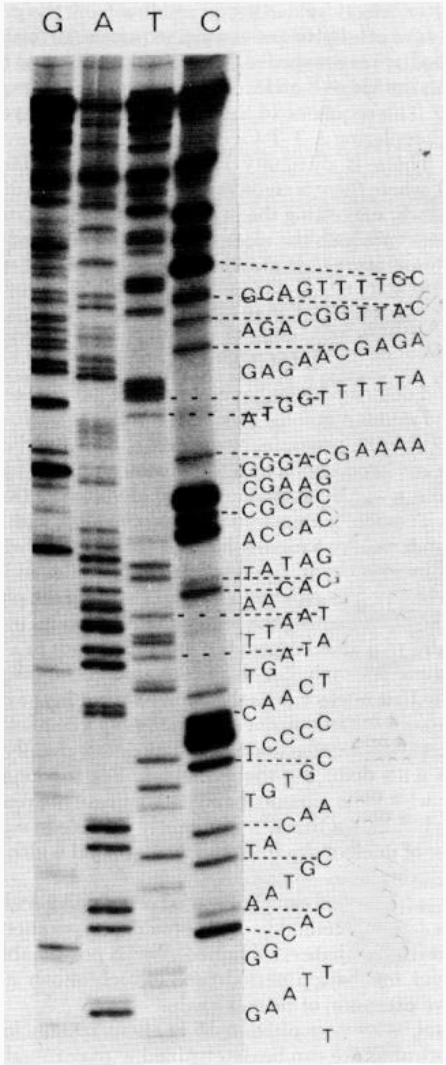


Figure 2. 聚丙烯酰胺凝胶电泳

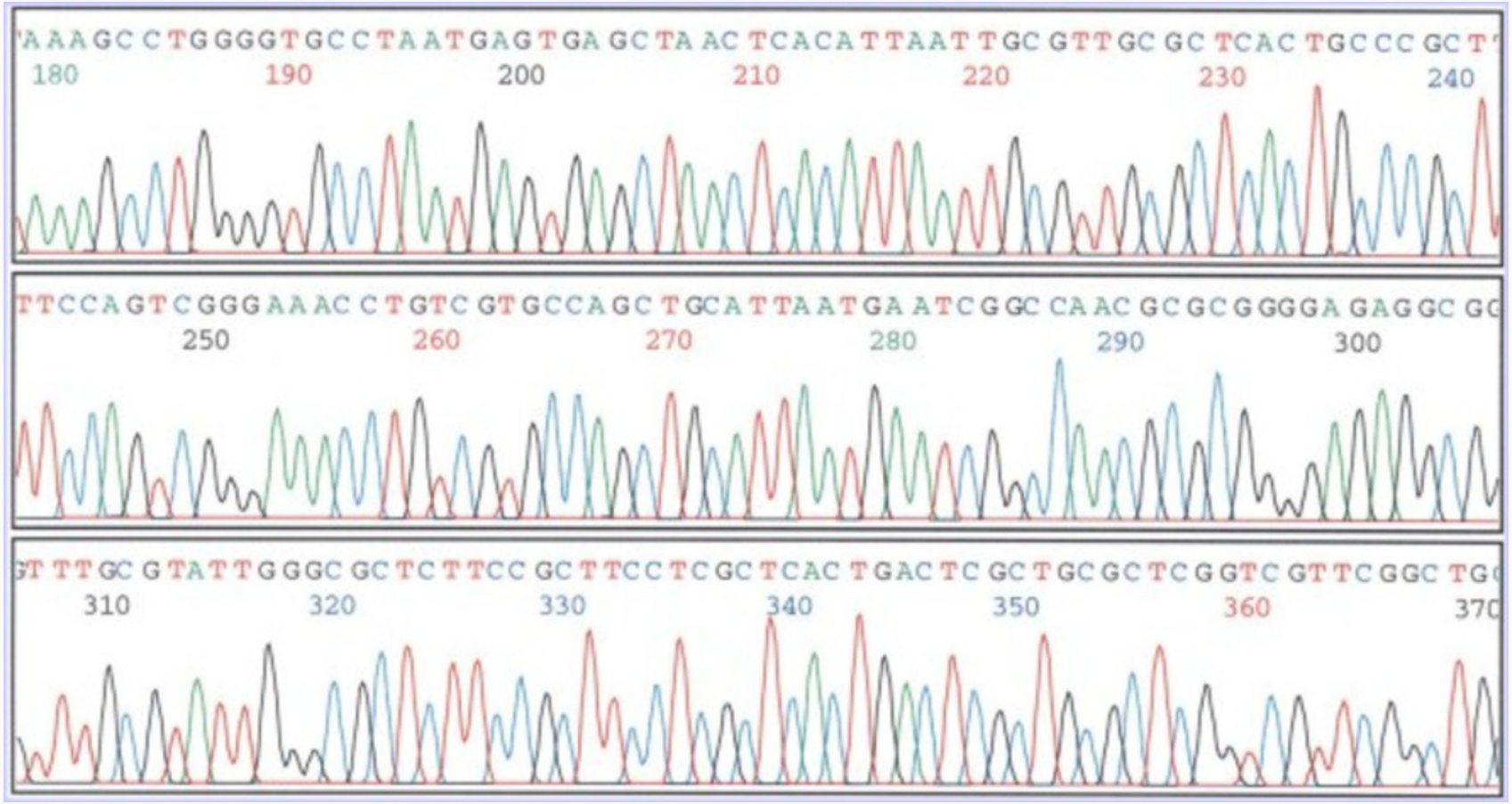


Figure 3. 毛细管电泳图谱

Figure 4. ABI 3730XL

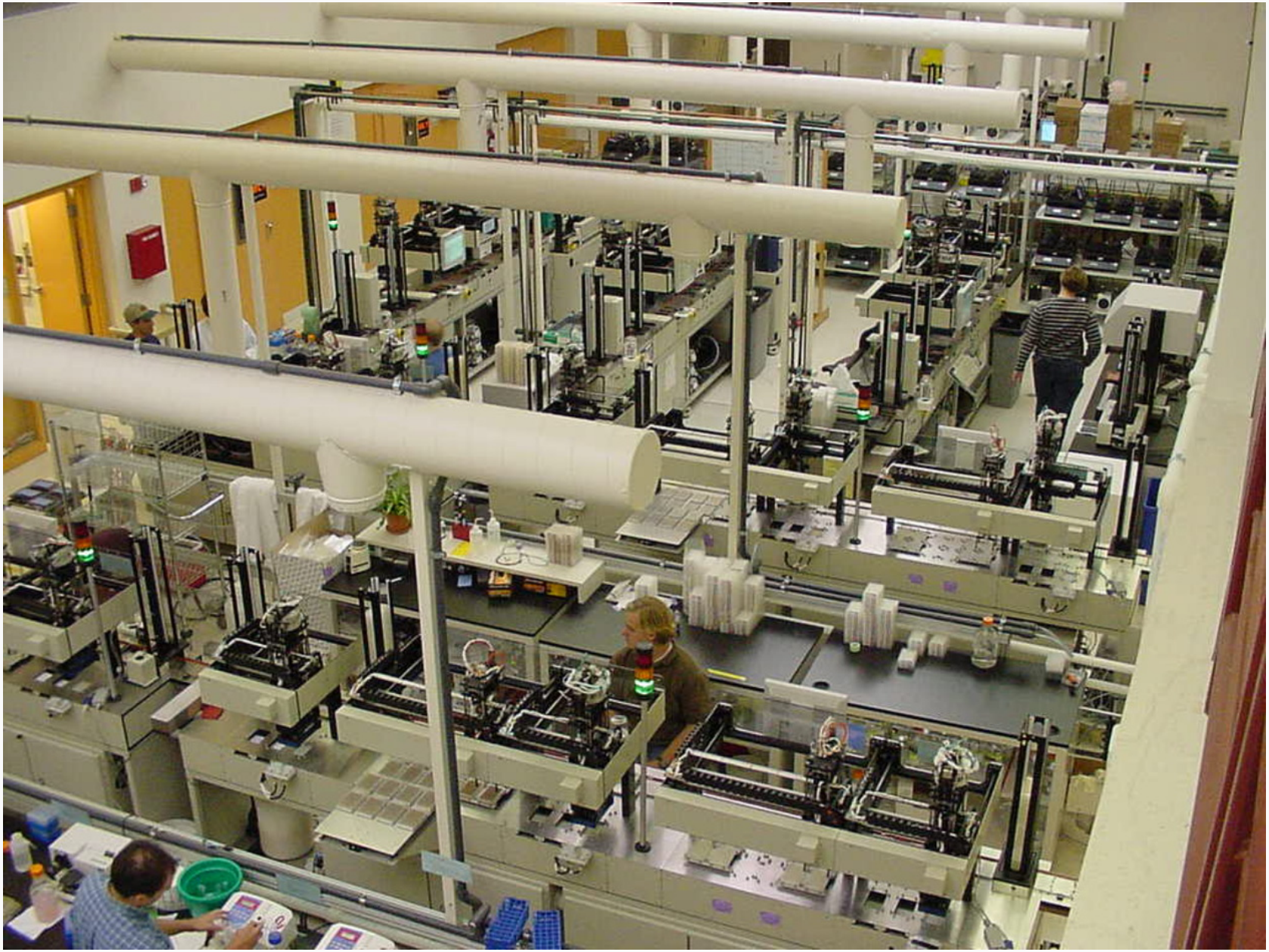


Figure 5. 模板与测序室

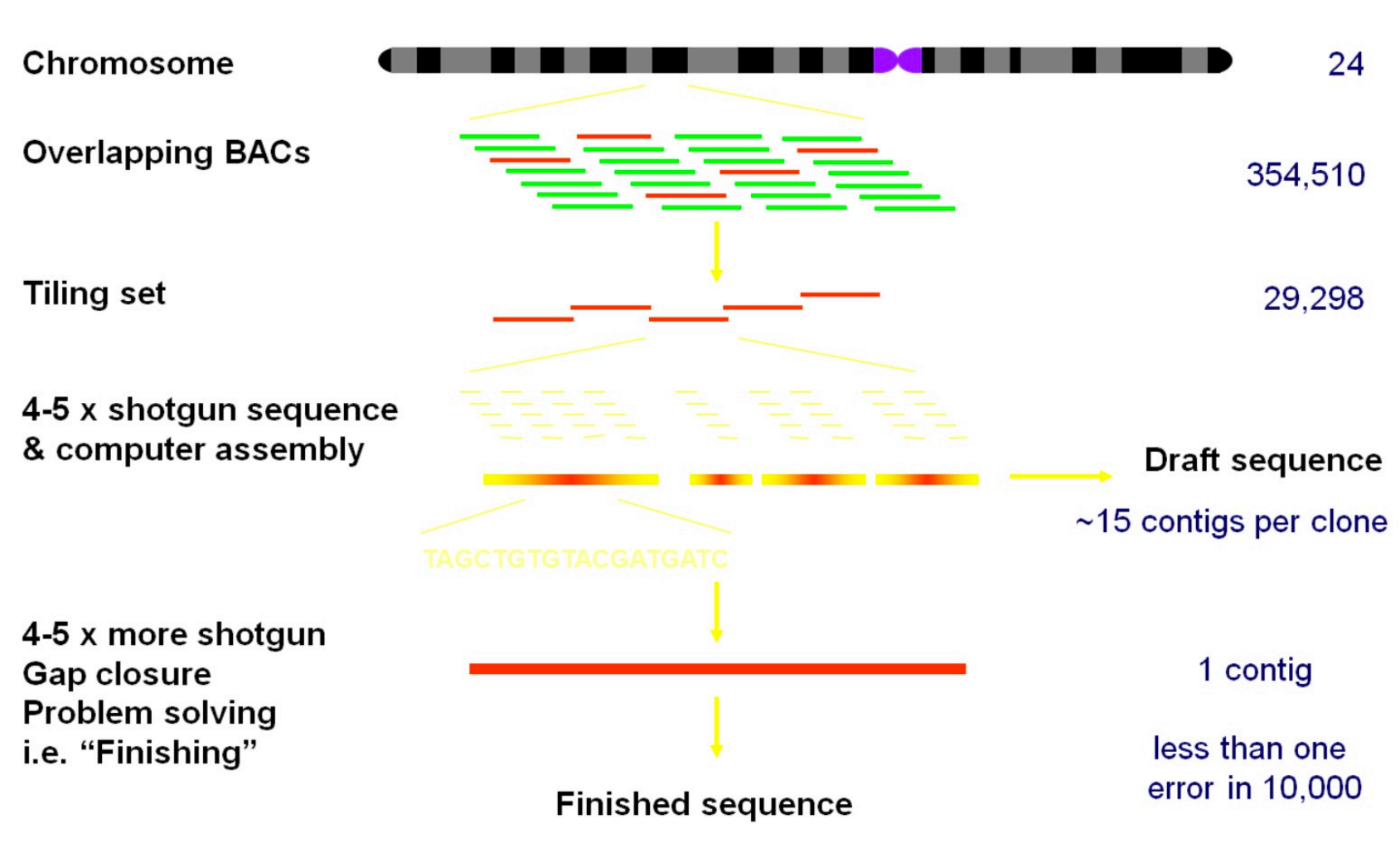


Figure 6. HGP sequencing strategy

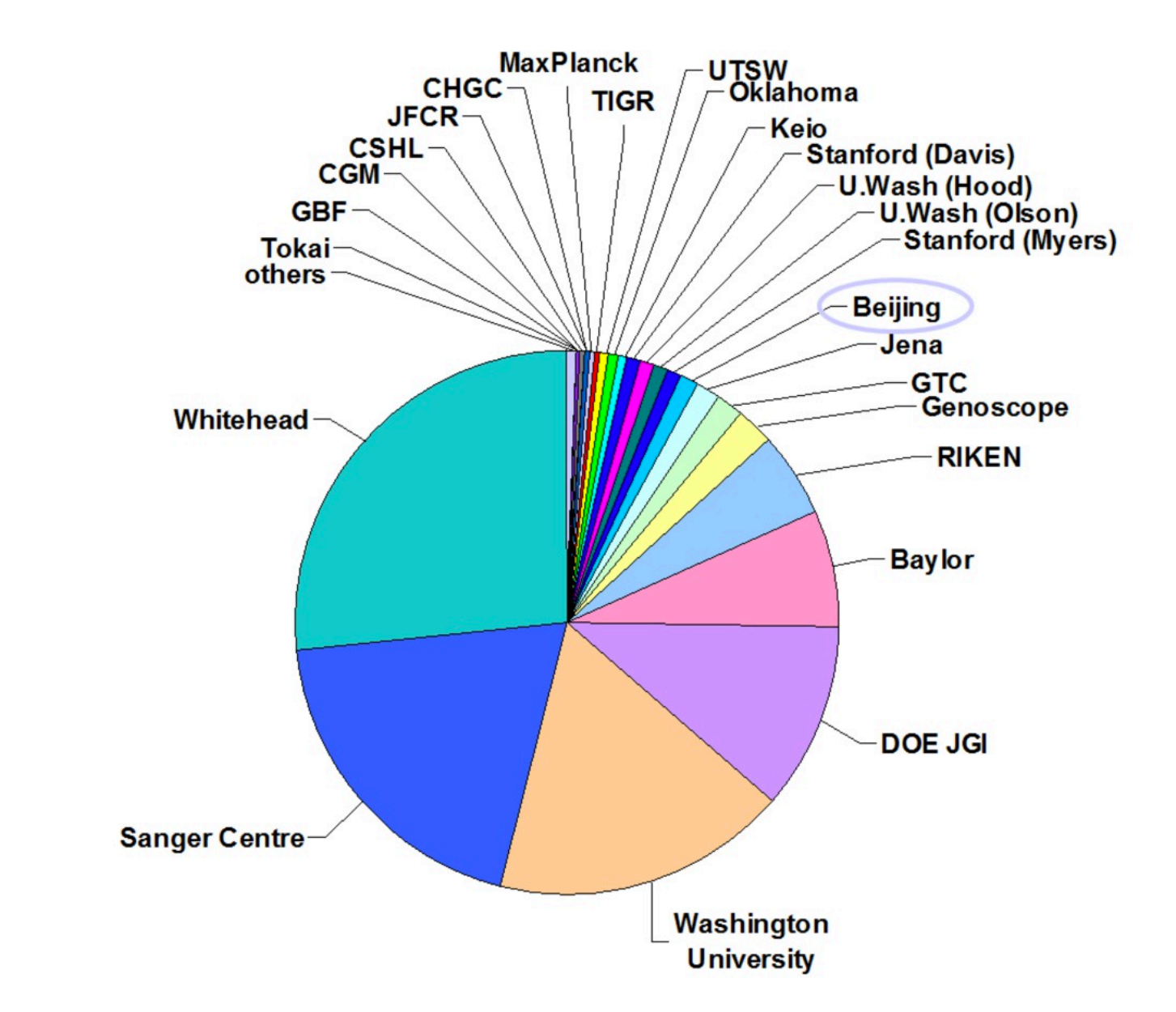

MaxPlanck
UTSW
CHGC
Oklahoma
TIGR
JFCR
Keio
CSHL
Stanford (Davis)
CGM
U.Wash (Hood)
GBF
U.Wash (Olson)
Tokai
Stanford (Myers)
others
Beijing
Jena
GTC
Genoscope
Whitehead
RIKEN
Baylor
DOE JGI
Sanger Centre
Washington University

Figure 7. 贡献

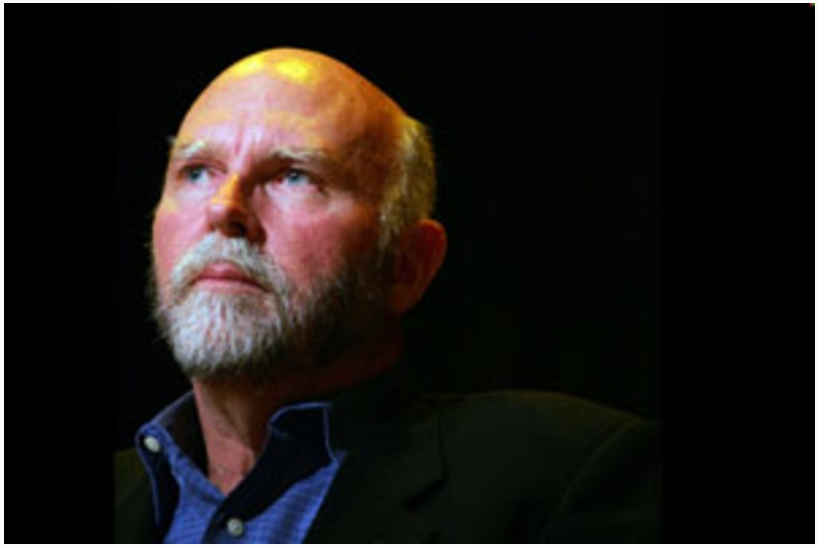


Figure 8. J. Craig Venter (文特尔)

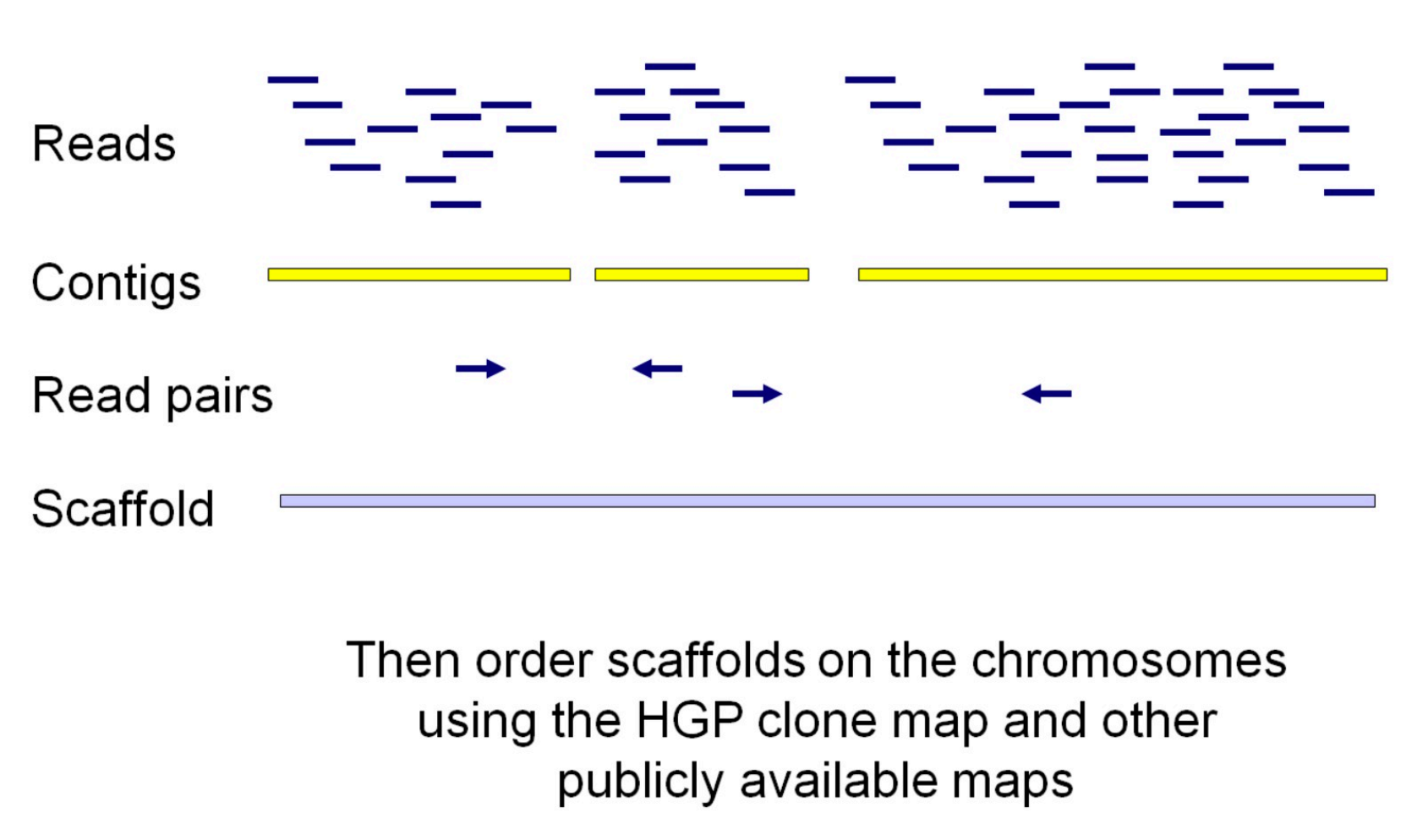
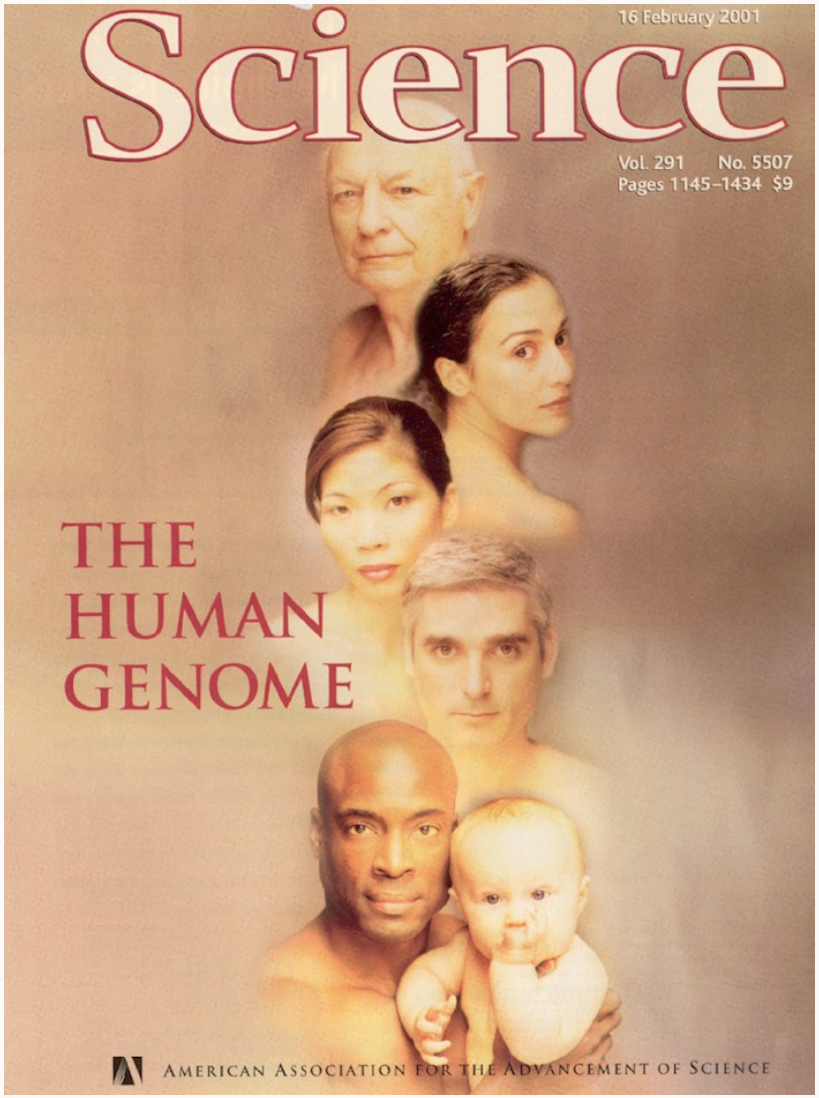


Figure 9. Celera assembly strategy

Figure 10. Celera stocks

(a) Nature



(b) Science

Figure 11. 基因组草图, 2001

二代: (短读长的) 高通量测序

Personal Genomes

- Craig Venter
- James Watson
- Stephen Quake
- George Church
- Marjolein Kriek
- Hermann Hauser
- Han Chinese
- Seong-Jin Kim
- Korean AK1
- Yoruban African NA18507
- 14 others sequenced by Complete Genomics
- Unknown number sequenced by Knome
- 6 genomes sequenced at high depth by the 1000 Genomes Project
- 180 genomes sequenced at low coverage by the 1000 Genomes Project
- Two acute myeloid leukemia patients

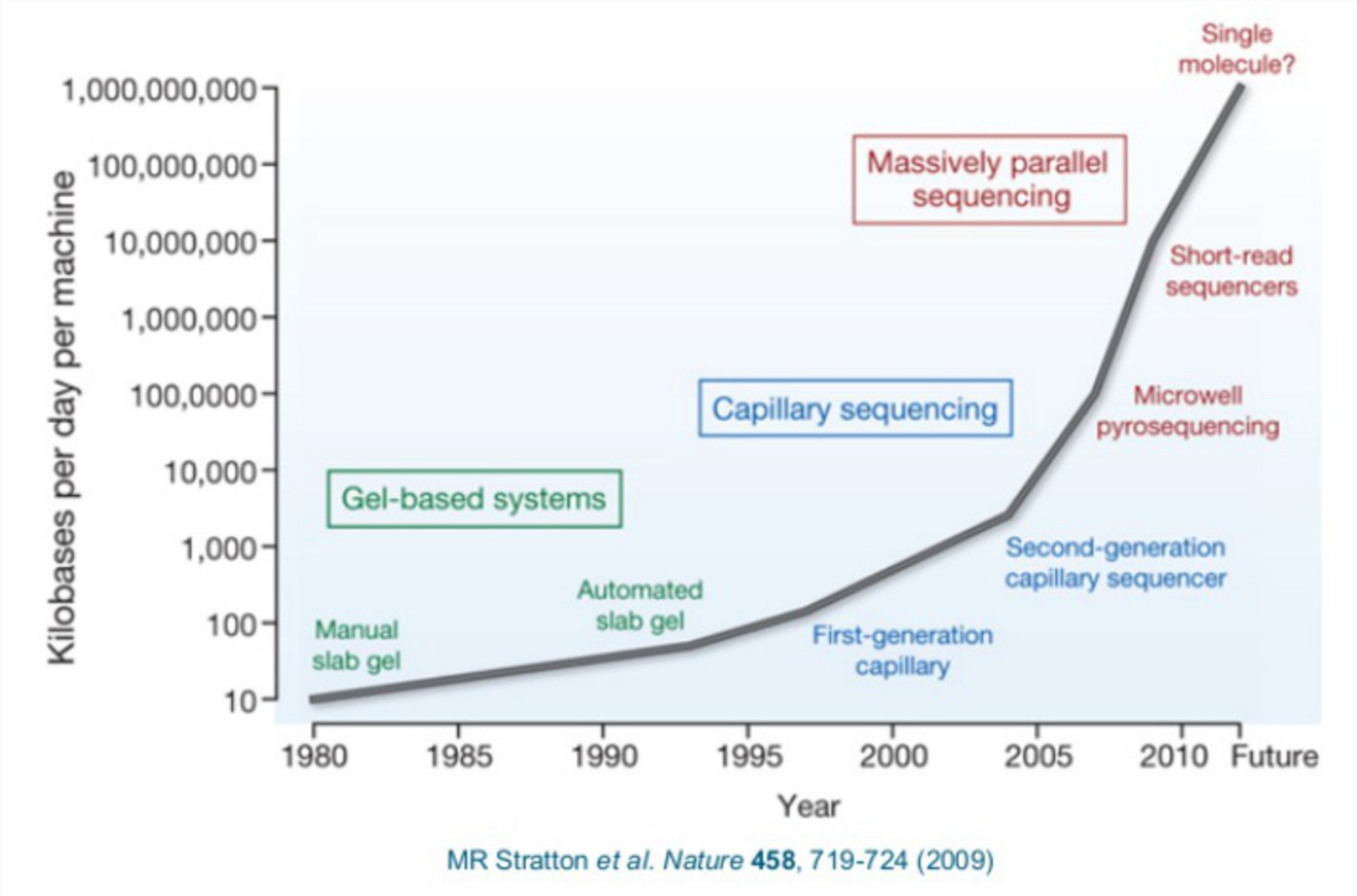


Figure 12. 测序能力的增长

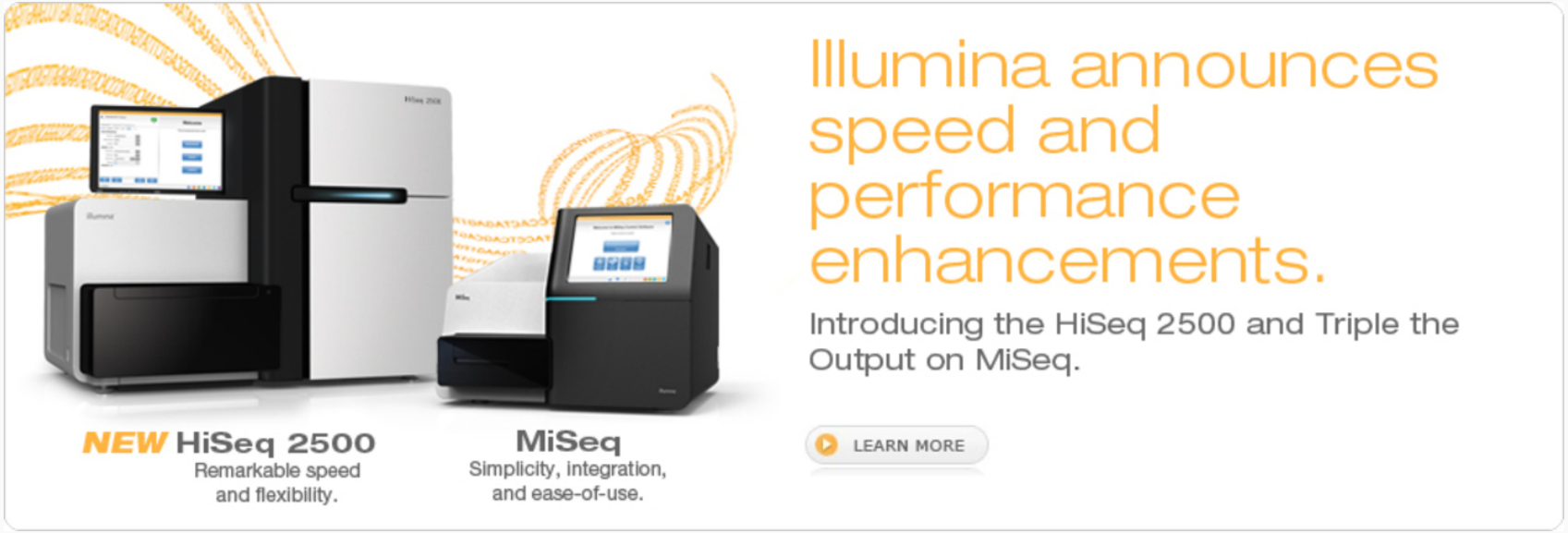


Figure 13. Illumina HiSeq

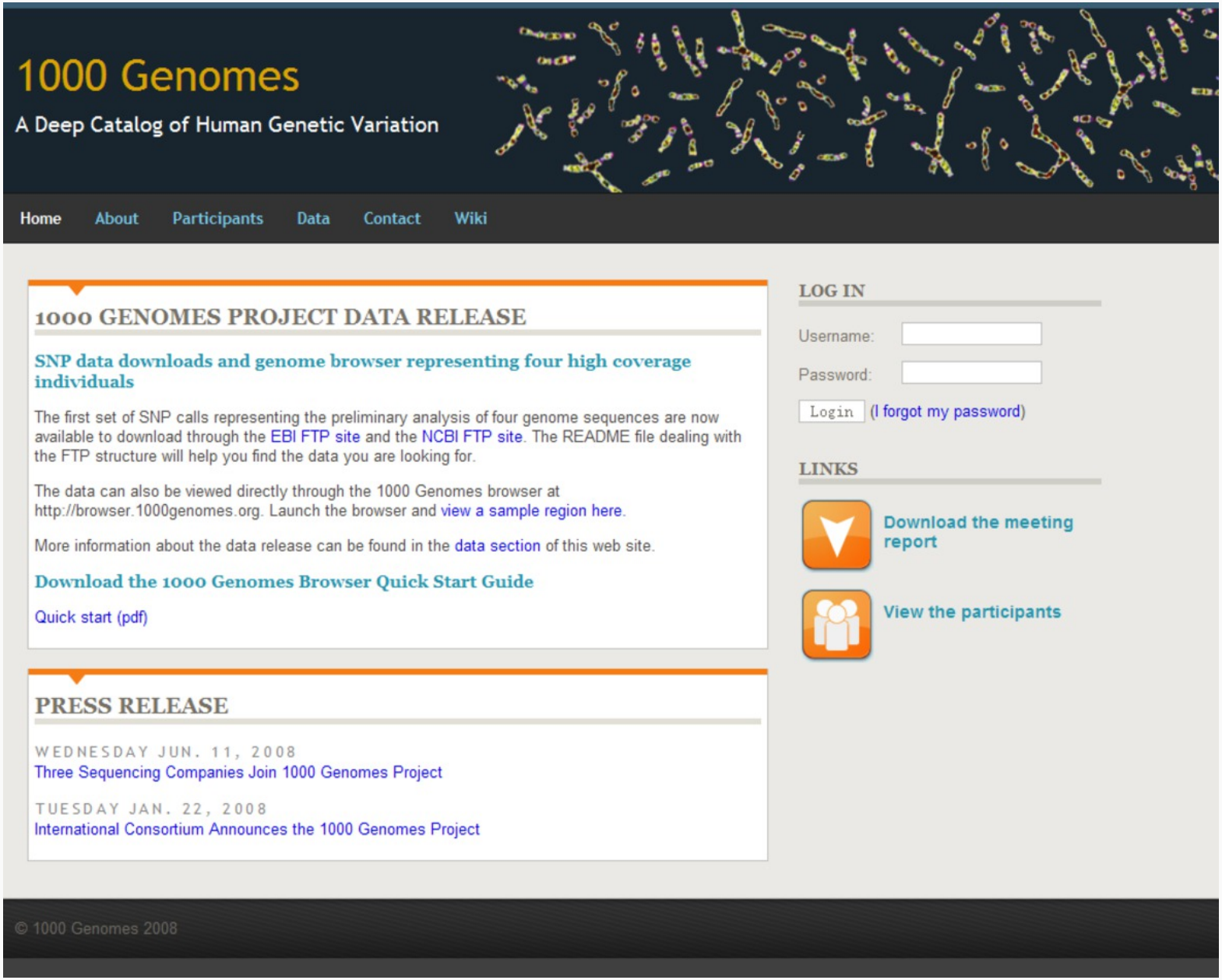


Figure 14. 千个基因组计划

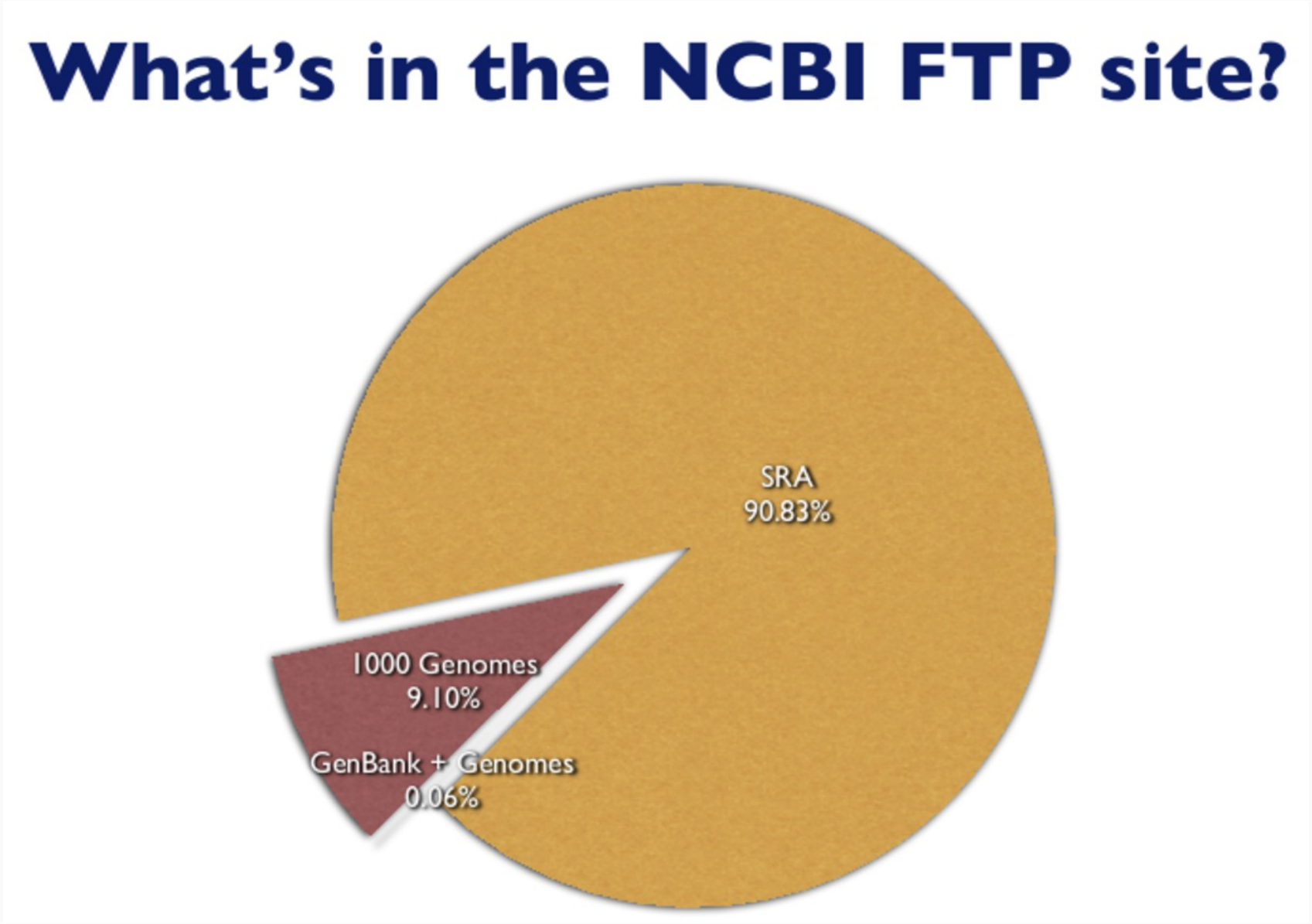


Figure 15. 爆发性增长的数据量

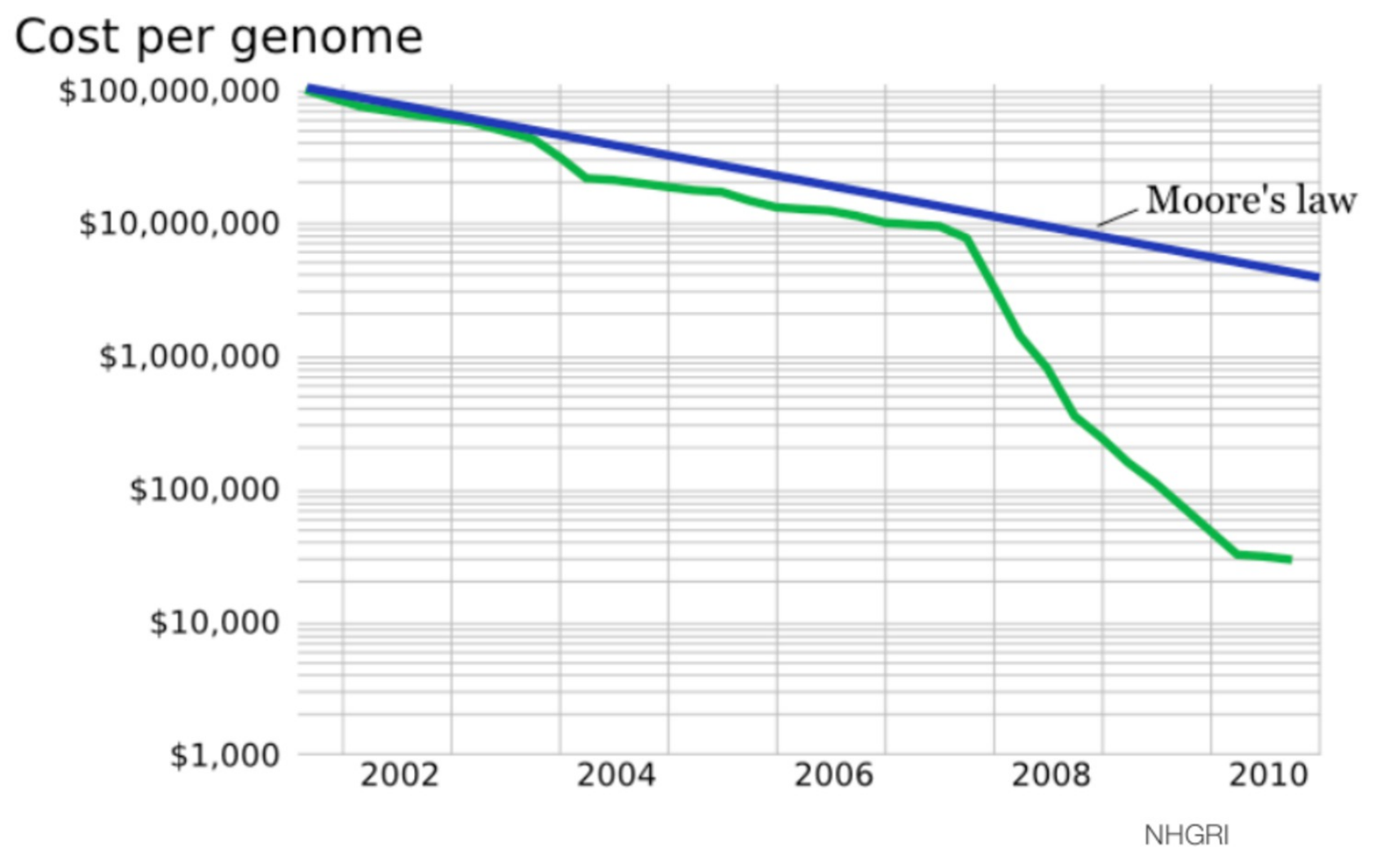


Figure 16. 比摩尔定律更快

Figure 17. HiSeq X

100 GB ≈ 33X Human genome ≈ 10,000 CNY

别再升级通量了，测序仪卖不出去了.

CEO，侬脑子哇特了？

其它二代方法

- Ion Torrent
- 454
- SOLiD
- Complete genomics

在最近与彭博社的一次采访中, 奥巴马透露了自己想要进入科技风投界的想法.

> 我与硅谷以及风投的交流, 极大地满足了我对科学和组织的兴趣... 你可以只花几千美元, 而不是十万美元, 就可以把个人的基因绘制出来. 你可以有能力辨识自己的(基因)倾向, 去生产对你这个个体而言最有效的药物. 这只是可以让我坐下来, 与别人谈几个小时的科技创业的例子之一.

三代: (长读长的) 单分子测序

二代的缺点

- 建库过程中扩增带来的偏性
- 高 GC 区域的覆盖度与准确性
- 读长短
  - 转座子 (transposons and retrotransposons)
  - 片段重复 (tandom or segment duplications)
  - 一般不超过 5 kbp, 但就是二代跨越不了的障碍

我们在甘蓝 (*Brassica oleracea*)[1] 基因组草图中发现了一个约 2 kbp 的片段, 重复数超过了 1000 次.

在其它完全基于二代测序的基因组里, 重复片段的数量都远大于用基于 BAC 的方法测序的物种.

基本都是测序与拼装中的错误.

[1] Liu, S. *et al.* The Brassica Oleracea Genome Reveals the Asymmetrical Evolution of Polyploid Genomes. *Nature Communications* **5** (2014)

二代的对策

- Pair end (short jump)
- Mate pair (long jump)
- 10X Genomics

三代的优势

读长 长!

三代的原理

几个名词

- PacBio: Pacific Biosciences
- SMRT: Single Molecule Real Time Sequencing
- ZMW: Zero-mode waveguide, 20 zl (zeptoliters, 10^{-21})

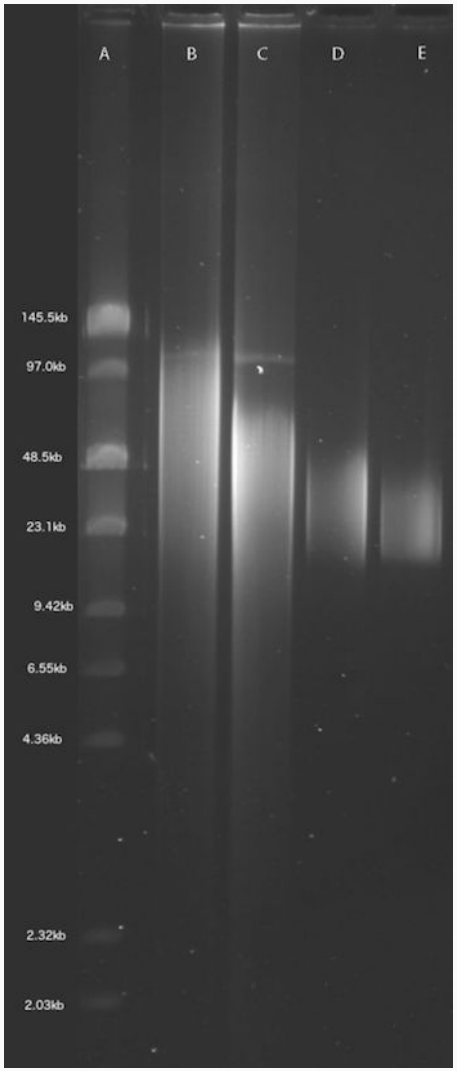


Figure 18. 建库, 脉冲场电泳[2]

[2] Chakraborty, M. *et al.* Contiguous and Accurate *de Novo* Assembly of Metazoan Genomes with Modest Long Read Coverage. *Nucleic Acids Research*, gkw654 (2016)

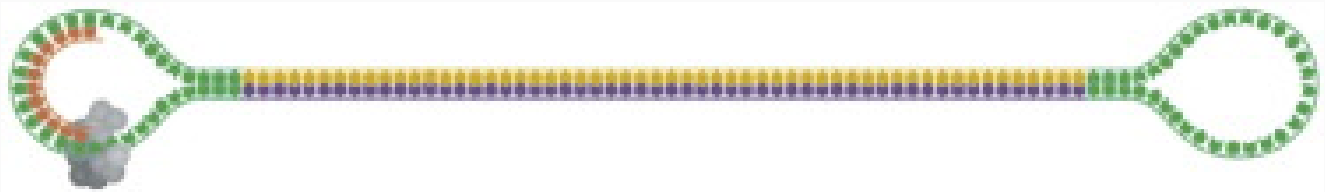


Figure 19. 模板, SMRTbell[3]

[3] Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics* **13,** 278–289 (2015)

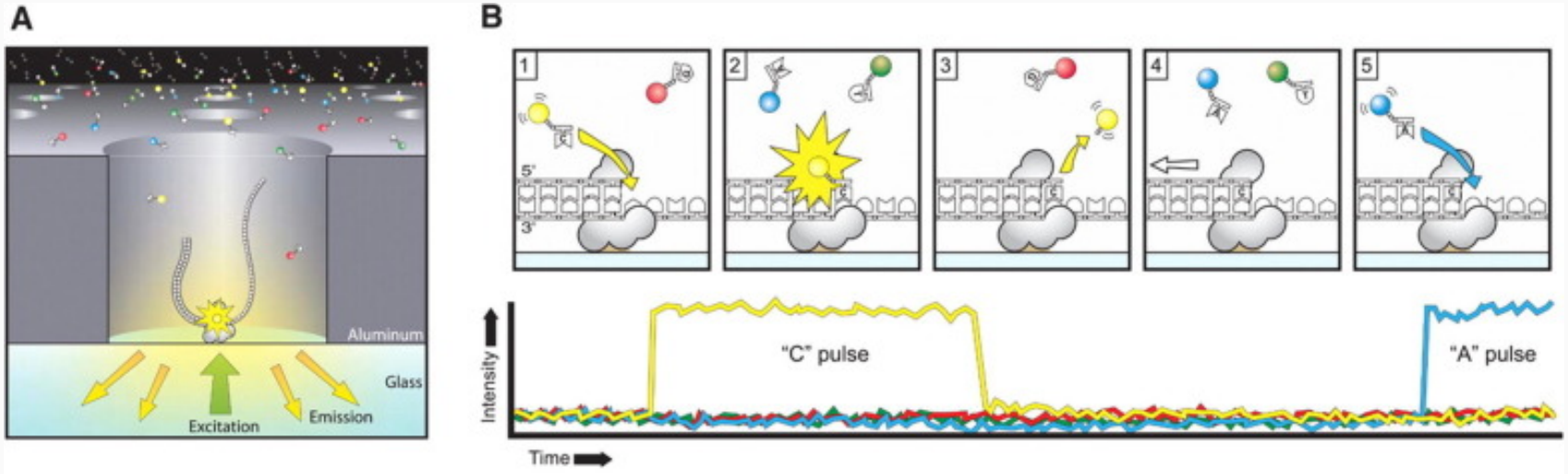


Figure 20. 单分子荧光测序

Figure 21. 一个 SMRT cell

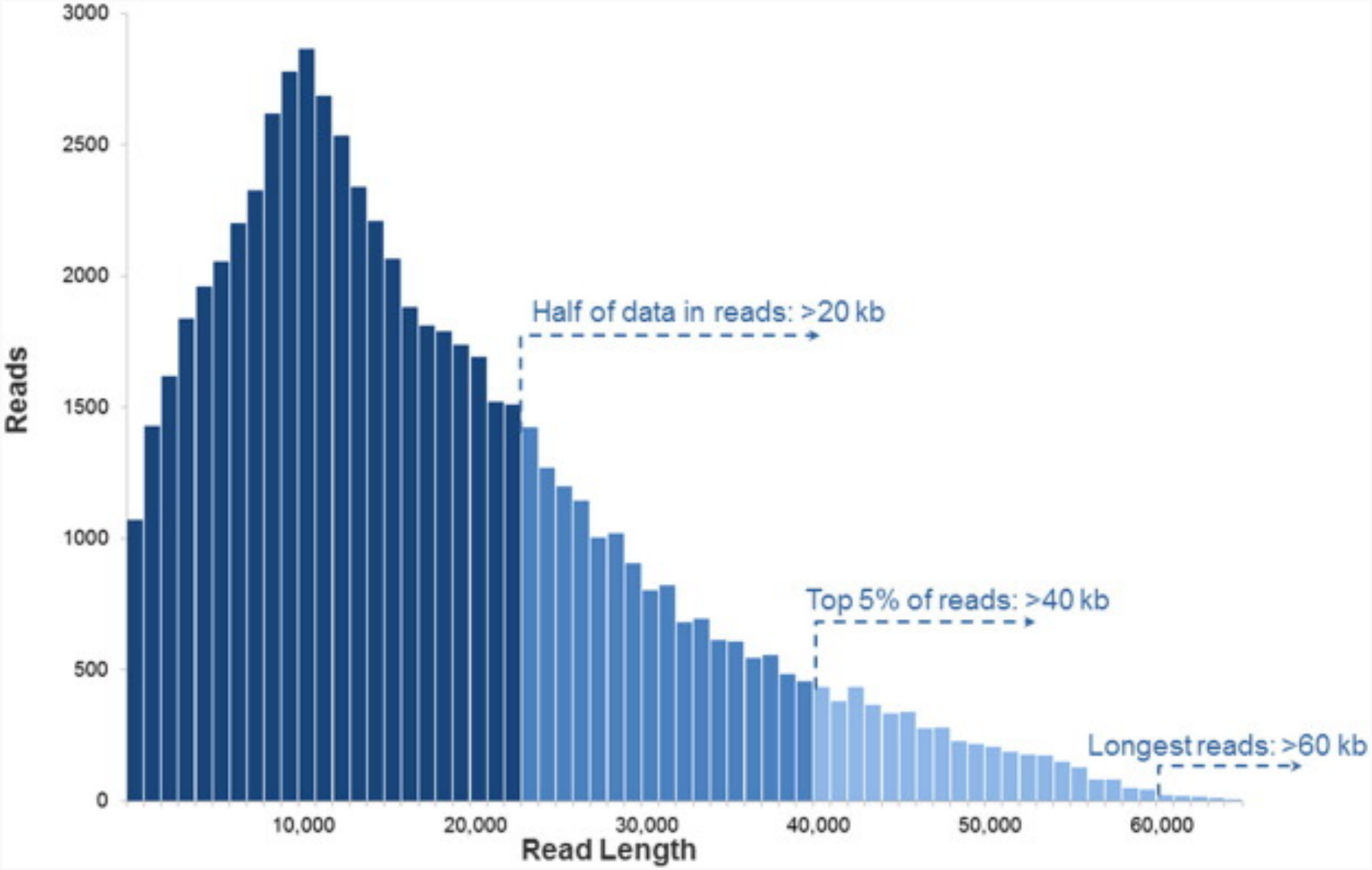


Figure 22. 读长分布

RS II 与 Sequel 对比

	RS II (P6-C4)	Sequel
运行时间	240 min	240 min
输出量	0.5-1 Gb	5-10 Gb
每日输出量	2 Gb	20 Gb
平均读长	10-15 kb	10-15 kb
单程准确率	**~86%**	**~86%**
30X 准确率	>99.999%	>99.999%
Reads 数	50k	500k
平台价格	$700k	$350k
运行成本	$400	$850

三代应用: 哺乳动物基因组

大猩猩基因组[4]

	Susie3	gorGor4	panTro4
Assembly size	3,080,414,926	3,063,362,754	3,323,267,922
Total coverage	74.8X	101.1X	6X
Technology	PacBio	Sanger/Illumina	Sanger
Contig N50	9,558,608	52,934	50,656
#contigs	16,073	170,105	183,860

[4] Gordon, D. *et al.* Long-Read Sequence Assembly of the Gorilla Genome. *Science* **352,** aae0344–aae0344 (2016)

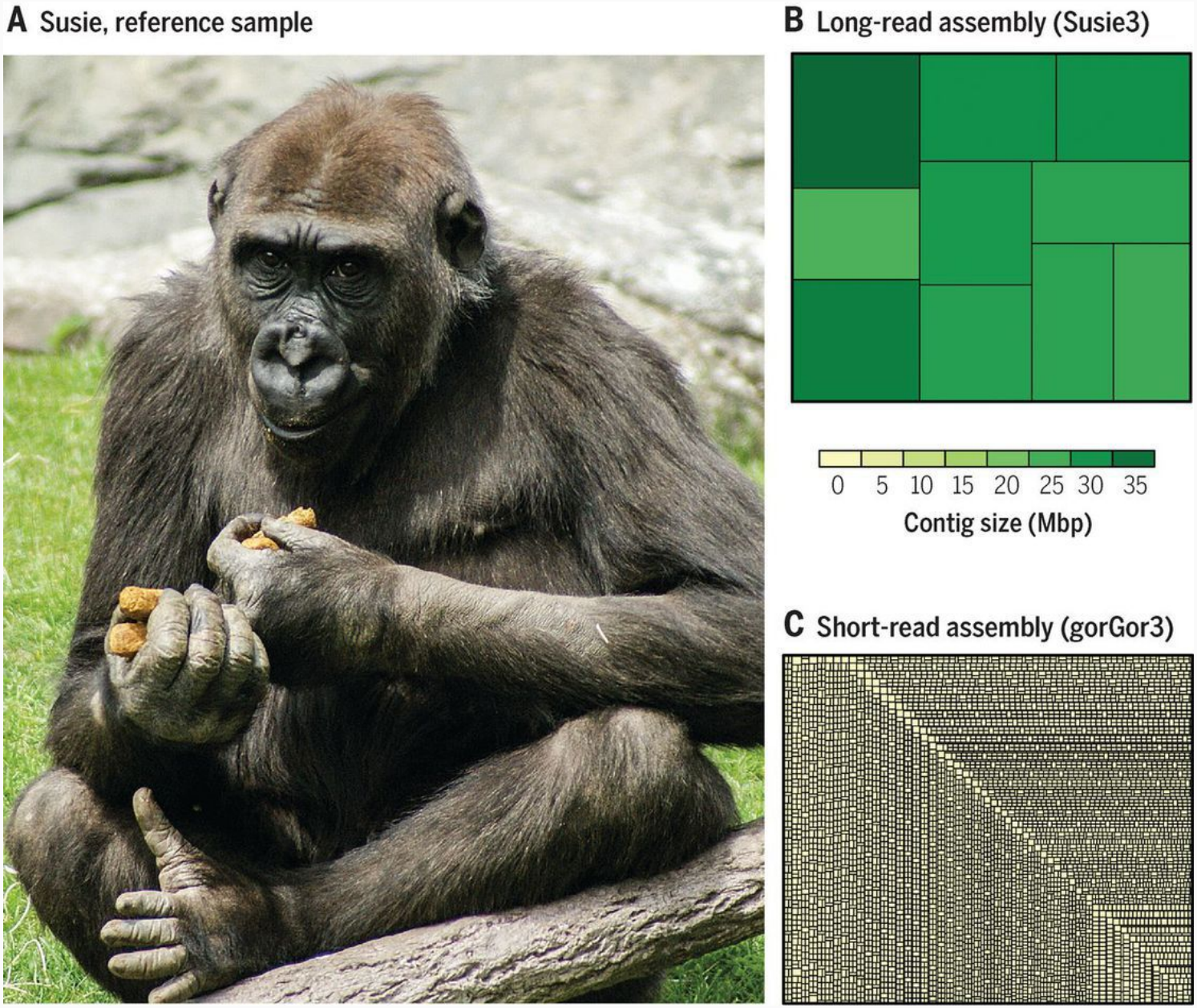


Figure 23. Susie

三代应用: 全长转录组

玉米转录组[5], 6 个组织.

二代测序只能准确地确定剪接点 (splice junctions), 对完整的转录本, 只能靠算法 (猜).

对于三代, 一个转录本就是一个单分子.

---

[5] Wang, B. *et al.* Unveiling the Complexity of the Maize Transcriptome by Single-Molecule Long-Read Sequencing. *Nature Communications* **7,** 11708 (2016)

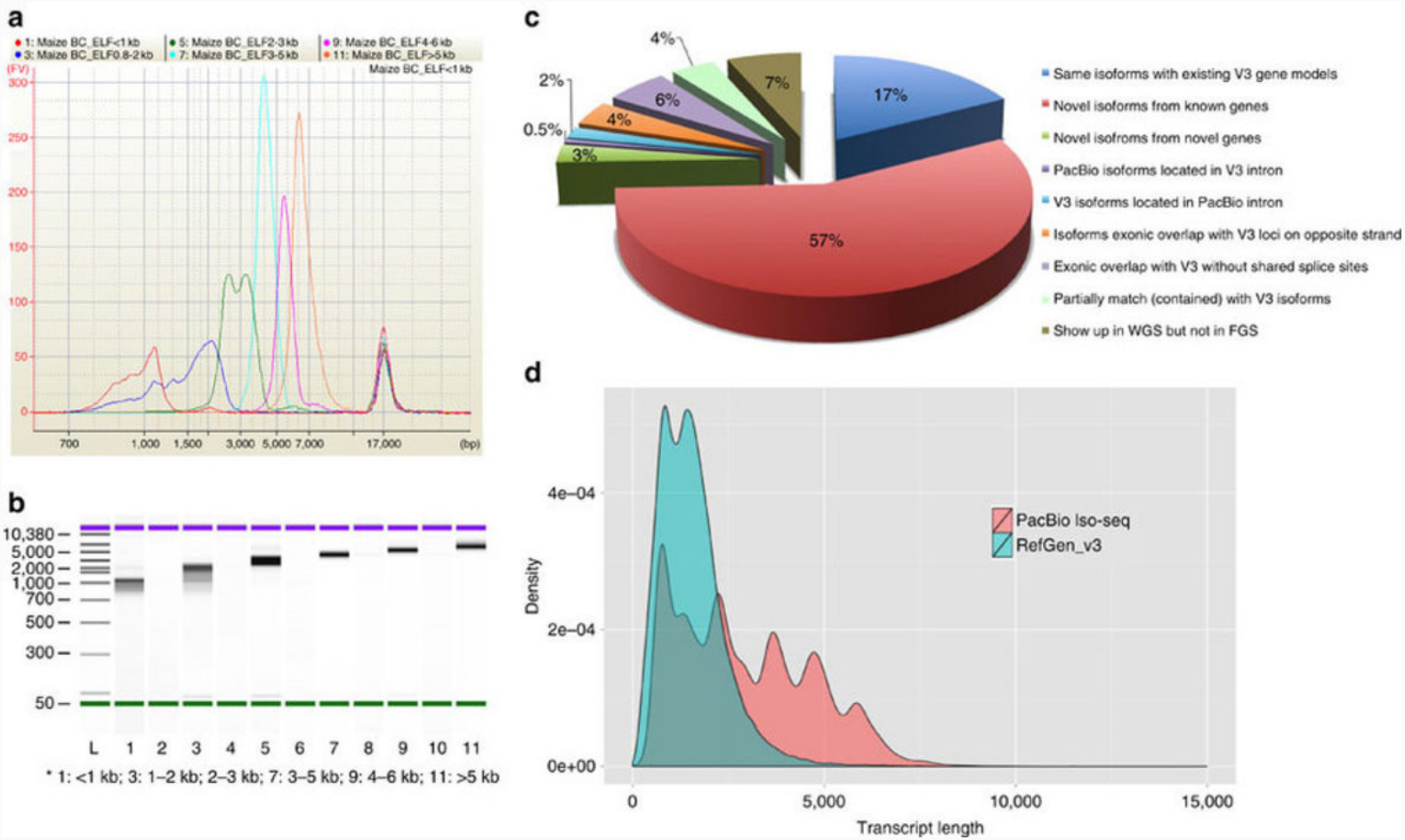


Figure 24. 玉米全长转录组

适用范围

	Sequel	原因
人类全基因组	Ok/Good	贵; 利于鉴定结构变异及组装
小基因组	Good	长读长, 只需要较低的通量
靶向测序	Good	长读长, 只需要较低的通量
转录组	Poor/Good	贵; 可得到全长的转录本
宏基因组	Poor/Ok	贵; 利于 de novo 组装
外显子组	Poor	贵; 长读长对外显子没有用处
表达谱	Poor	贵
ChIP-Seq	Poor	贵

信息学的问题

二代所用的算法和策略很多没法在三代上使用, 只能重新开发.

SMRT Analysis Software 包括了大量自有和第三方程序:

- 编程语言
  - C/C++
  - Bash
  - Java, Scala
  - Mono 3 (C#, VB.net)
  - Perl, Python
- 平台: Tomcat, MySQL
- 文件格式 HDF5 $\rightarrow$ BAM
- Celera Assembler, GMAP, HMMER, SAMtools 等

商业化服务

- http://www.pacb.com/products-and-services/service-providers/
- http://allseq.com/providers/

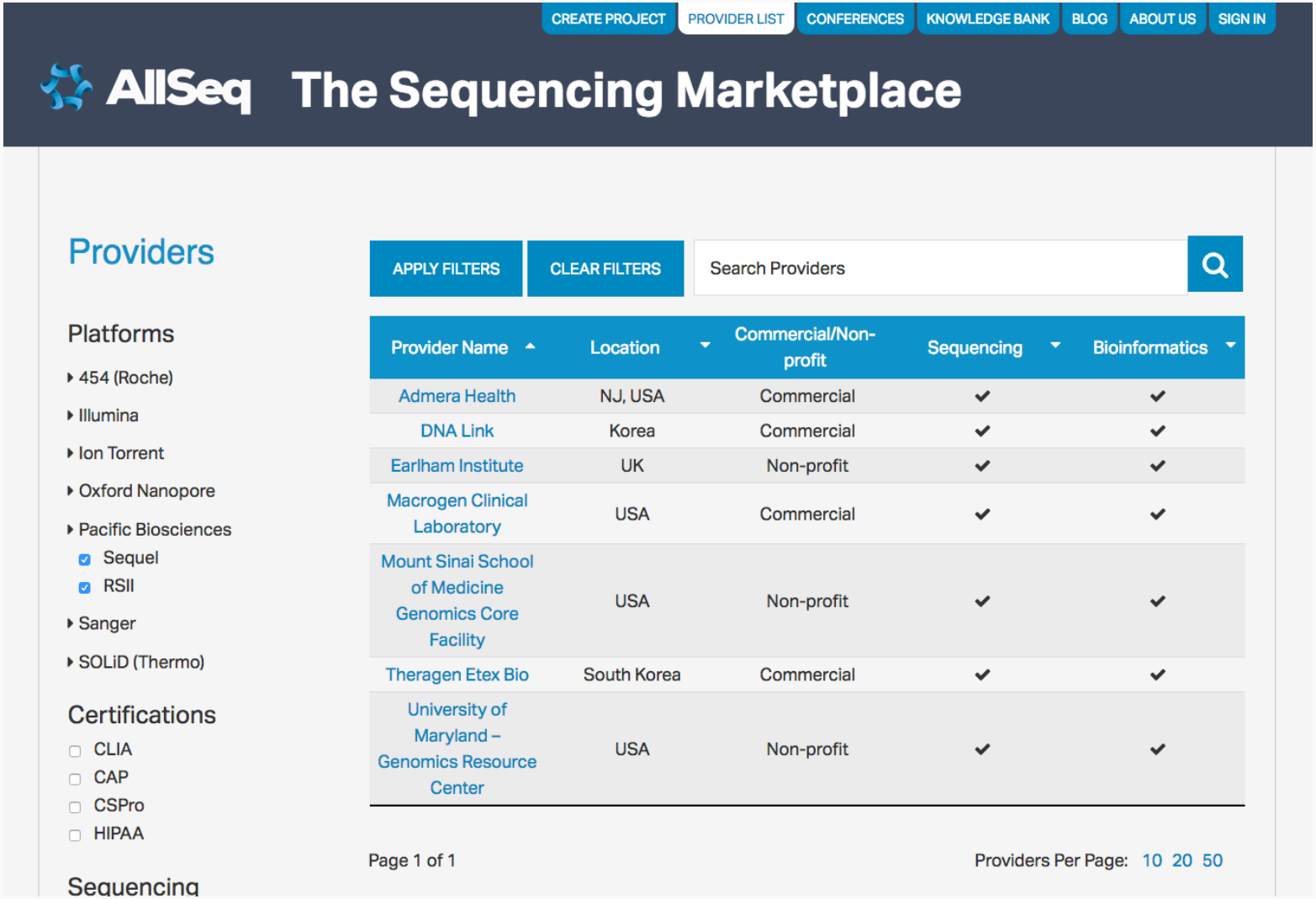


Figure 25. 商业化服务列表

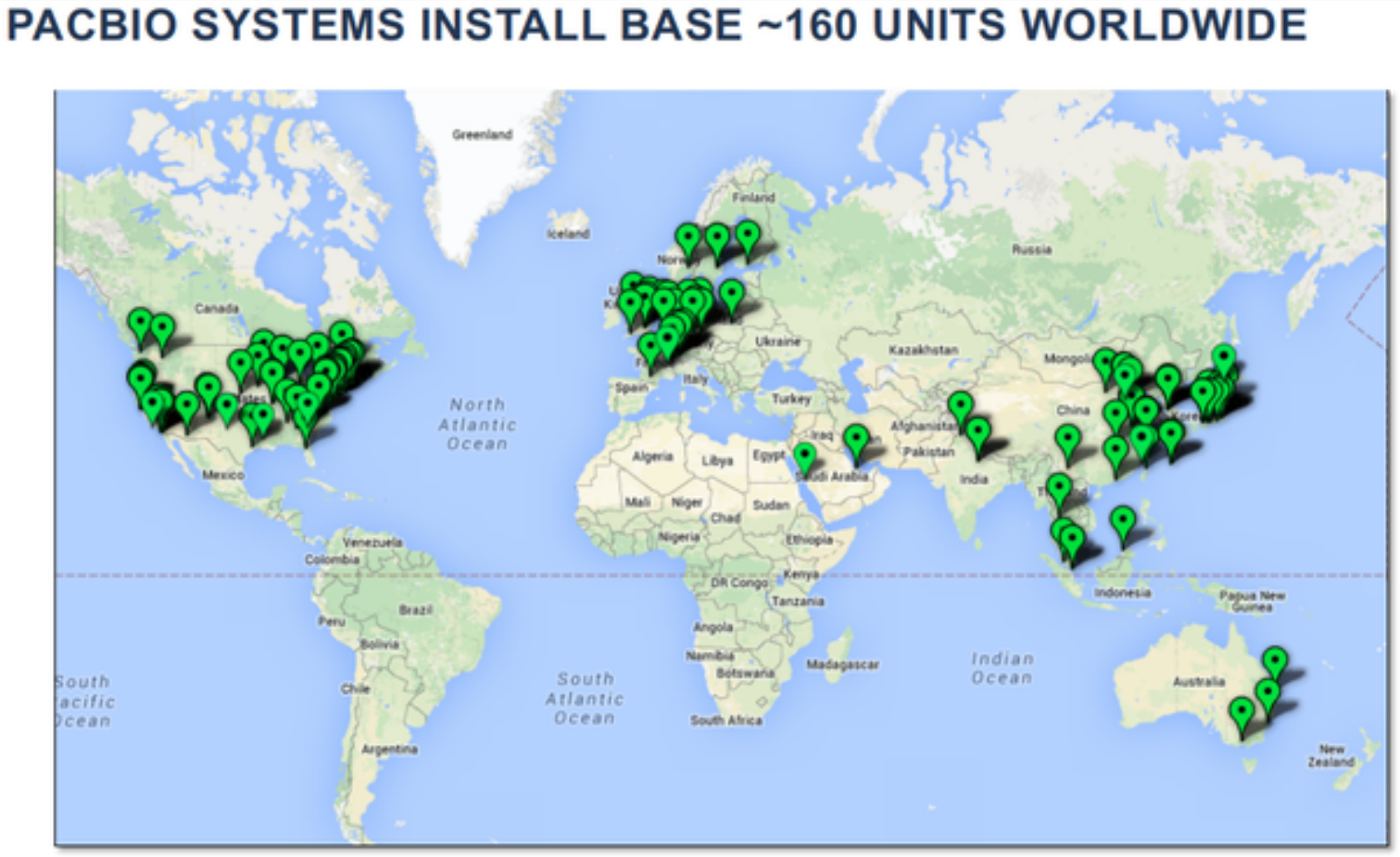


Figure 26. 2015 年全球装机约 160 台

武汉菲沙基因信息有限公司联合美国 Pacific Biosciences 公司在武汉东湖综合保税区共建高通量测序中心, 引进了不少于 6 台 Sequel 三代测序系统, 将建成大规模的三代测序中心.

1 GB ≈ 5,000 CNY

其它三代方法

- Helicos, 读长过短, 已经破产
- Oxford Nanopore

Figure 27. 纳米孔测序

Figure 28. MinION

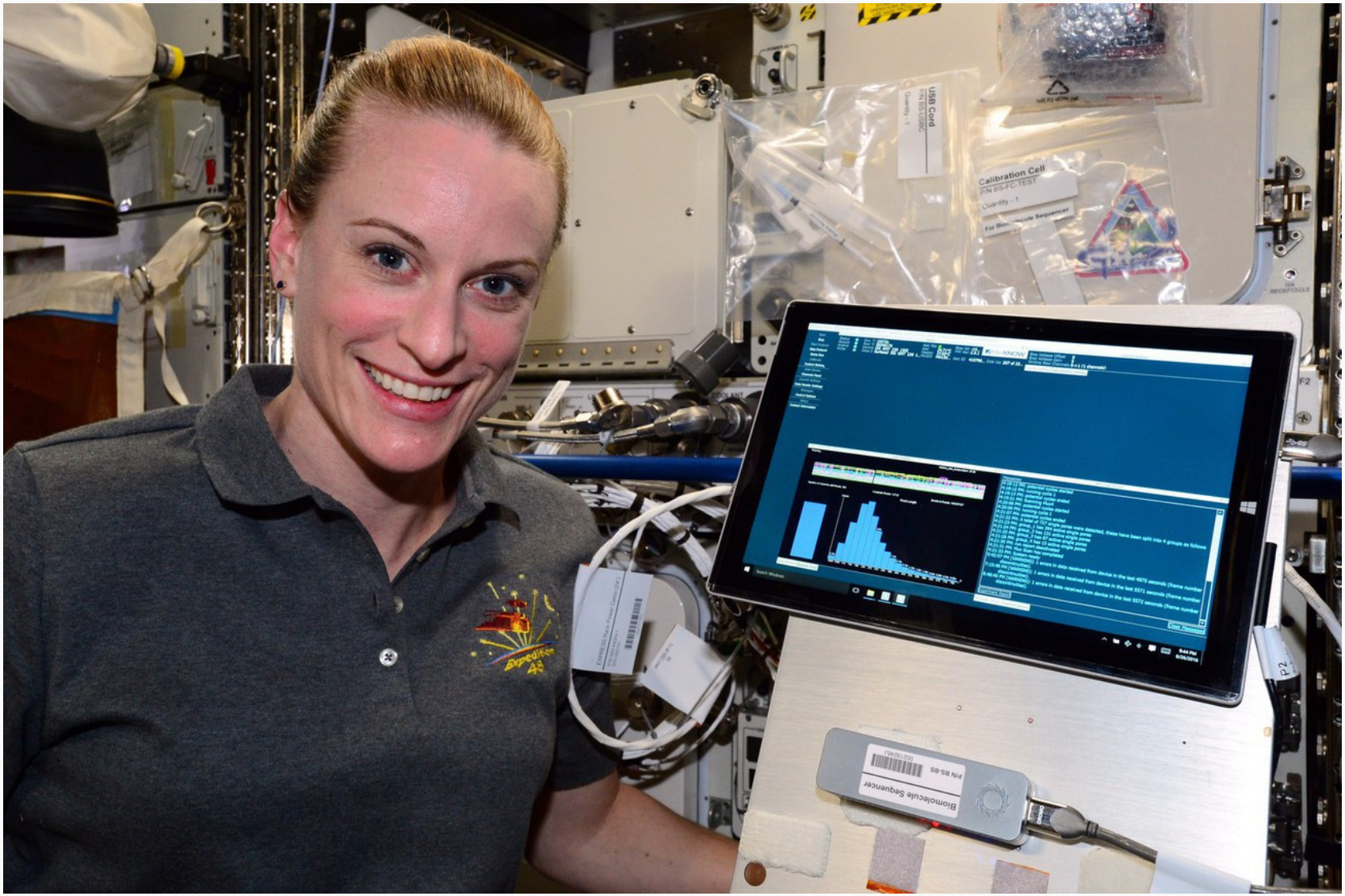


Figure 29. 第一次在太空中测序 DNA

总结

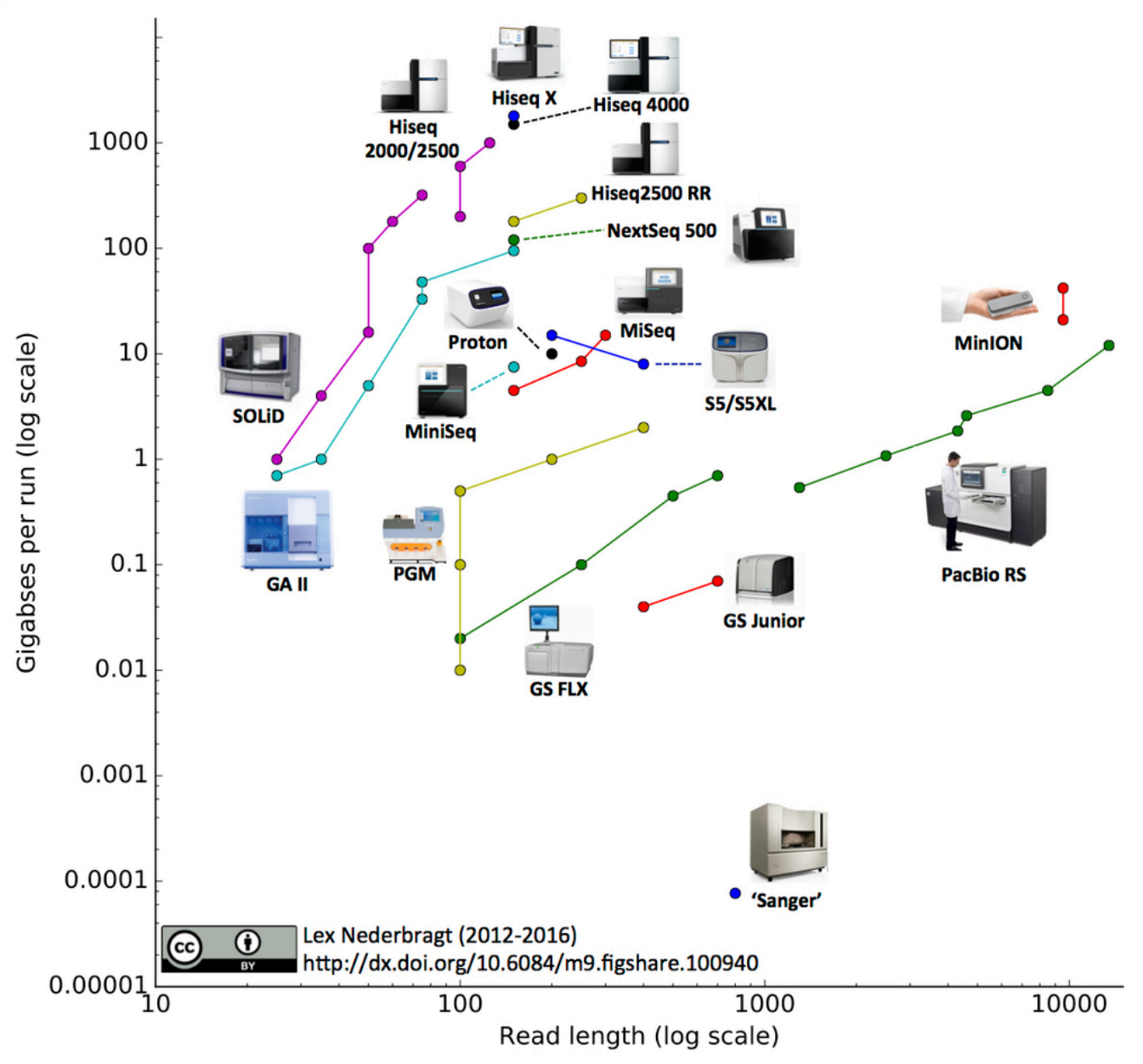
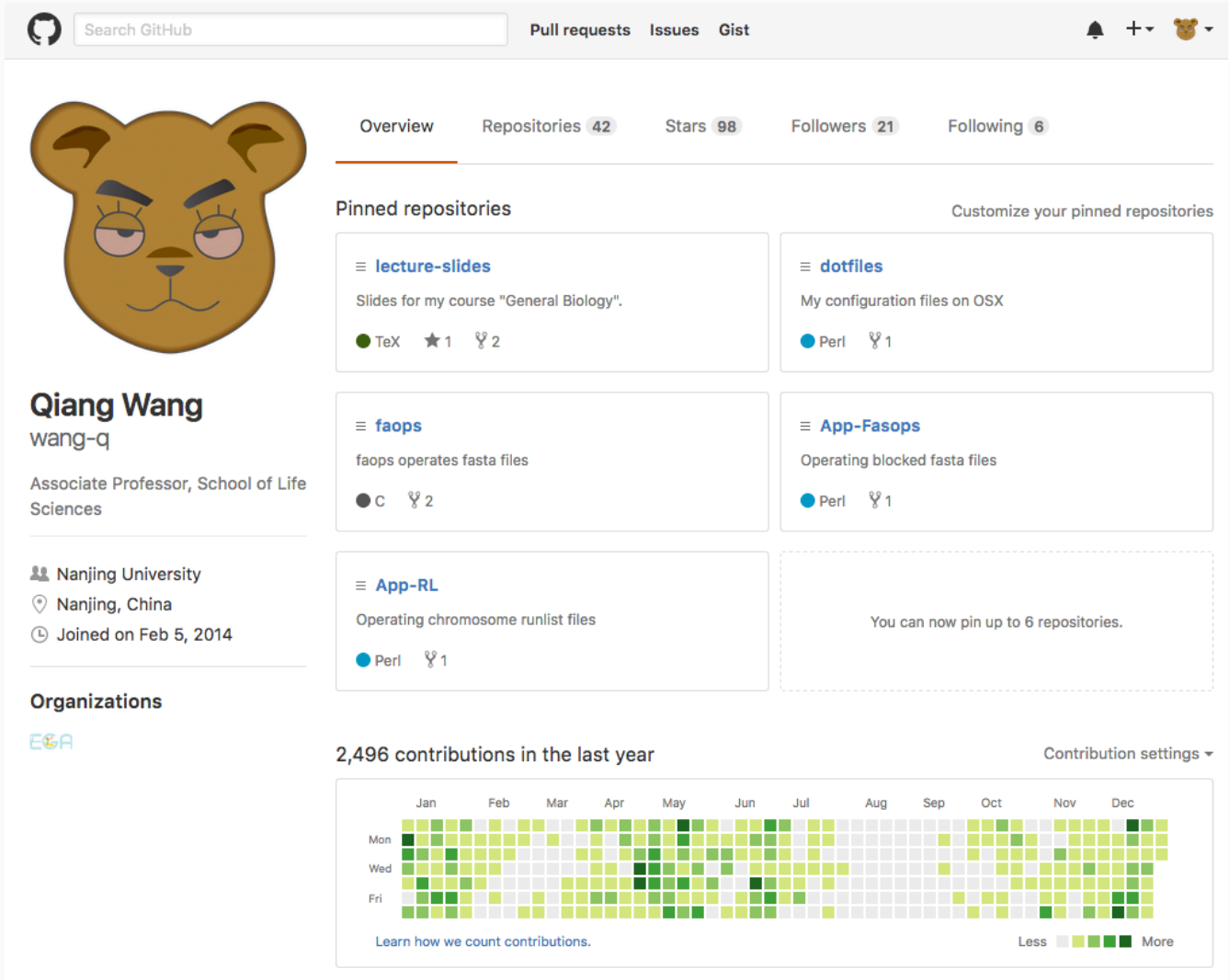


Figure 30. 高通量测序的发展



https://github.com/wang-q/lecture-slides/blob/master/slides/pacbio.slides.pdf