FANSe2 序列比对算法极大提升测序数据分析精度并通过翻译组测序建立蛋白质组研究新策略

张弓

暨南大学 生命与健康工程研究院

地基不牢，再漂亮的建筑也会倒塌

2009年6月27日，上海闵行区一幢13层在建商品楼发生倒塌事故

一片玻璃都没碎，质量可见一斑，但地基不稳，一切都白搭

测序的高楼大厦

高级分析：
SNP
甲基化
mRNA profiling
miRNA profiling
Ribosomal footprinting
…

基础分析：序列比对(mapping)

建库与测序实验

序列比对(mapping)是绝大多数测序应用的基础分析，其精度至关重要

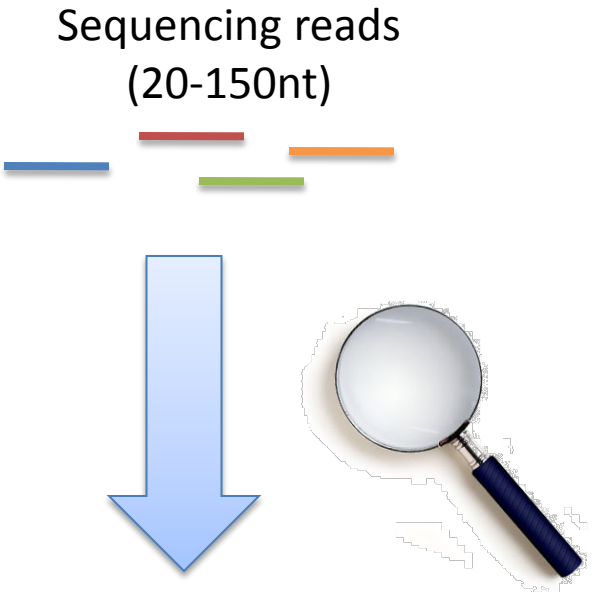
如果这一步的数据不可靠或不完全，其上所有的分析都是错误的。

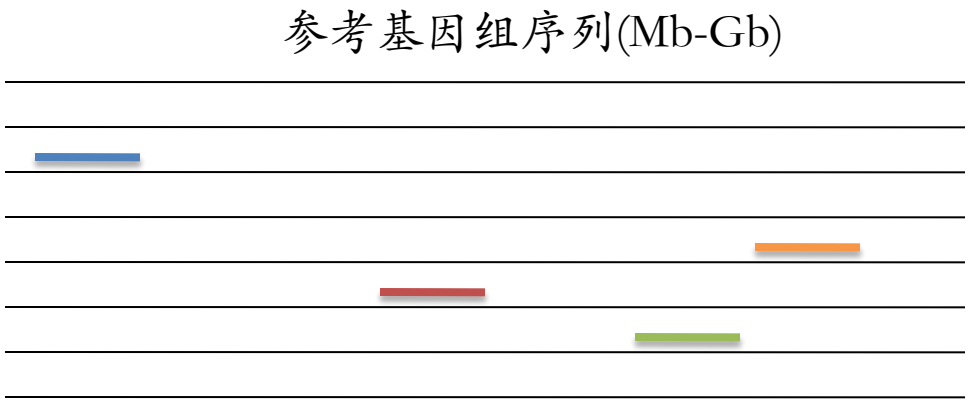
Mapping

将测序产生的数百万个短读序列(reads)向参考基因组上比对，定位其位置。

要求:

- 准确度高
- 足够快
- 可接受的内存耗用





精确的算法实在太慢

- 最精确的动态规划序列比对算法: Smith-Waterman algorithm
- 不甚精确但比较快的启发式算法: BLAST

尽管如此，基因组序列太长，而且我们需要比对几百万甚至上亿reads。

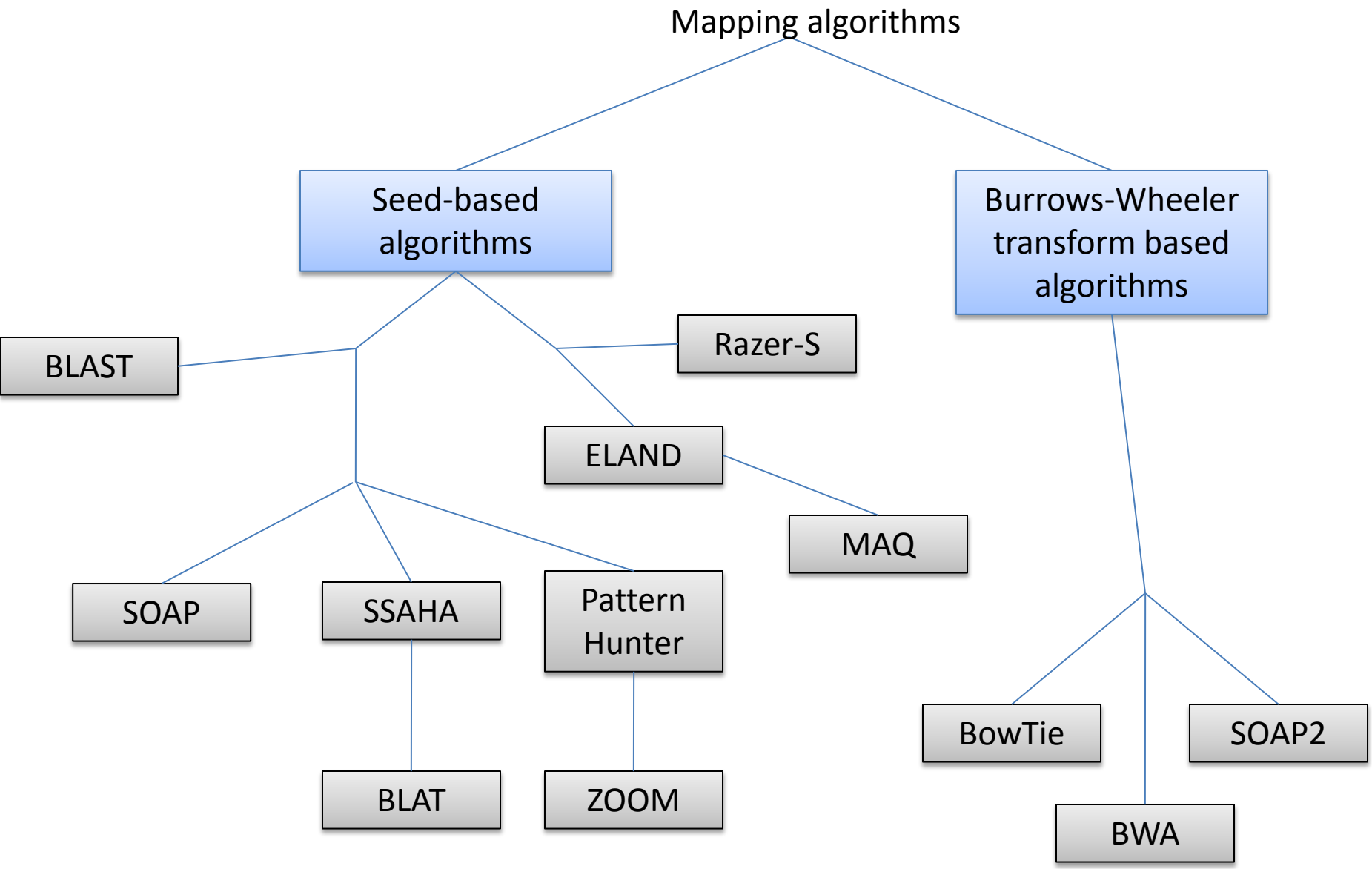
"A large, expensive computer grid might map the reads from this experiment in a few days using traditional alignment algorithms such as BLAST or BLAT."

- Nature Biotechnology (2009) 27:5, 455

World's fastest supercomputer…

… is not here.

Mapping 算法的进化树



我们该相信谁?

不同的mapping算法对同一个数据集可以给出大相径庭的结果。

我们该信谁?

目前主流的做法是——

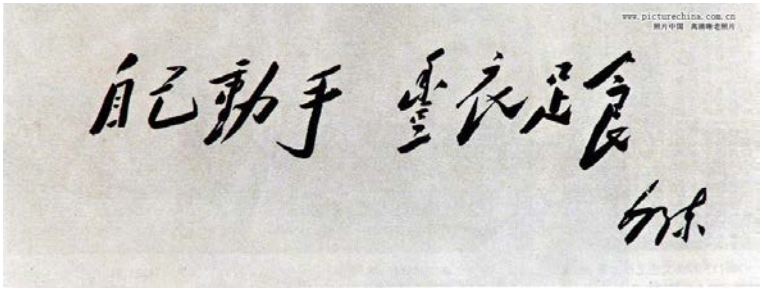
看哪个顺眼就用哪个,

或者看大牛用什么咱也用什么……

	Time (s)	% mapped
BFAST	43,775	32.1
BLAT*	68,758	24.3
Bowtie	2,270	13.1
BWA	7,682	16
MAQ	8,607	28.7

PLoS ONE (2009) 4:11

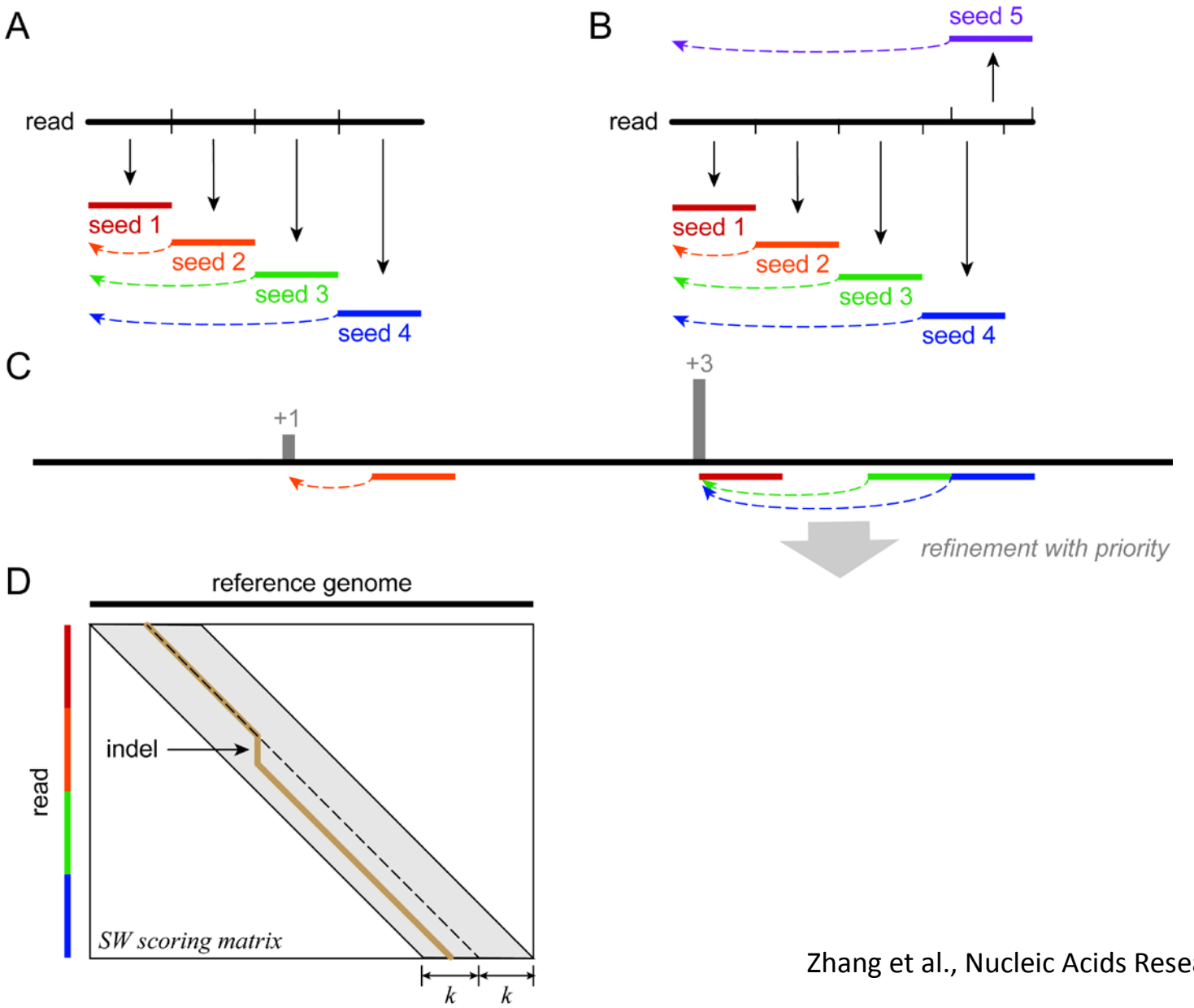
毛主席教导我们:



算法设计目标

- 高精度！不要漏掉可以mapping的read！准确率需高达99.9%以上
- 保持合理的速度
- 稳健性好
- 适用面广，适应多种测序平台与测序应用

FANSe 算法的基本原理

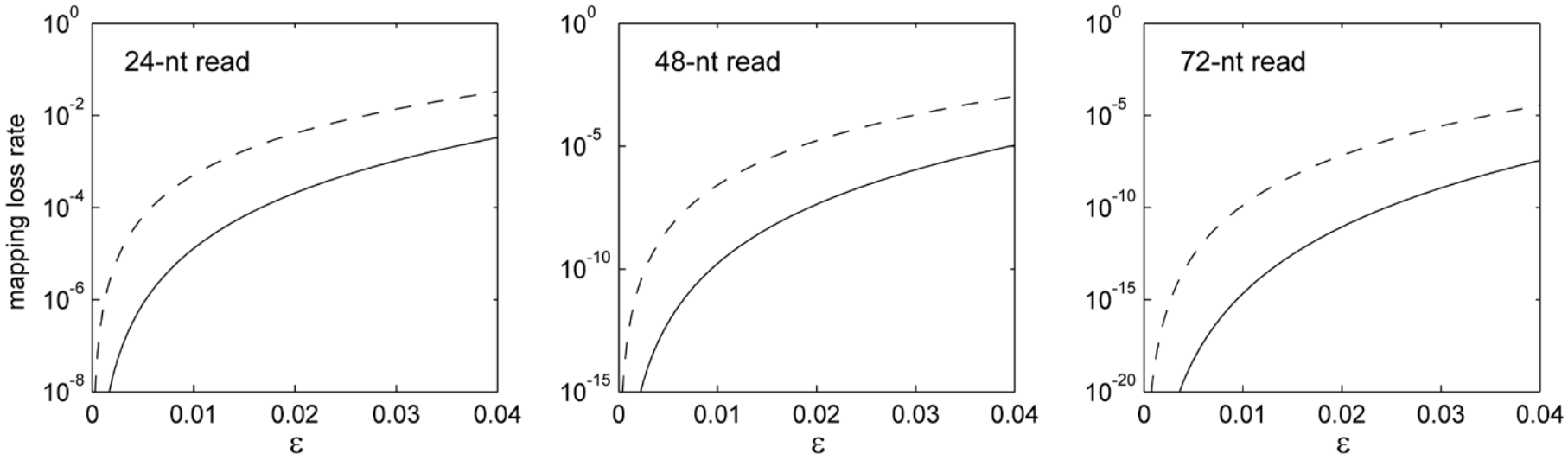


Zhang et al., Nucleic Acids Research, 2012

准确度

- FANSe 准确度极高，几乎不漏掉可以mapping的reads
- Mapping错误率是完全可以预估的，这是第一种可以预估错误率的算法，可以帮助用户选择合适的参数。

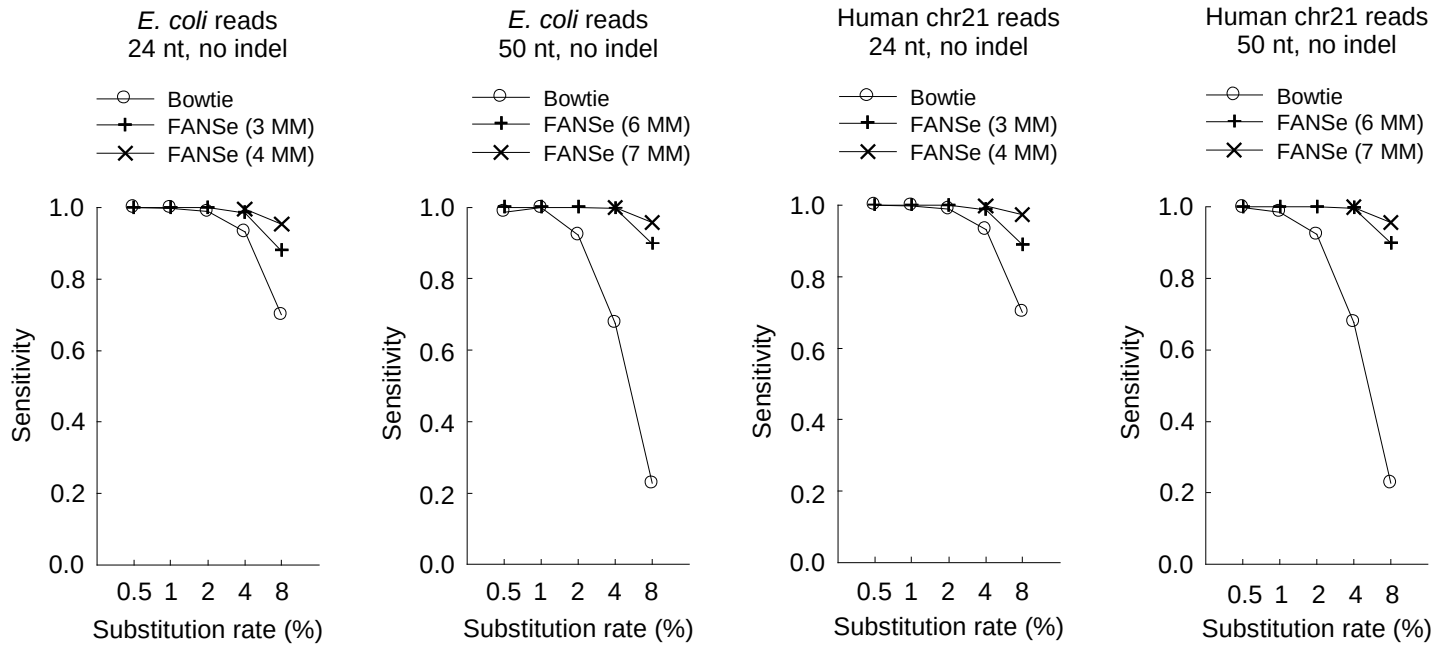
Mapping错误率：



Zhang et al., Nucleic Acids Research, 2012

FANSe的准确性远超其他算法（模拟测试）

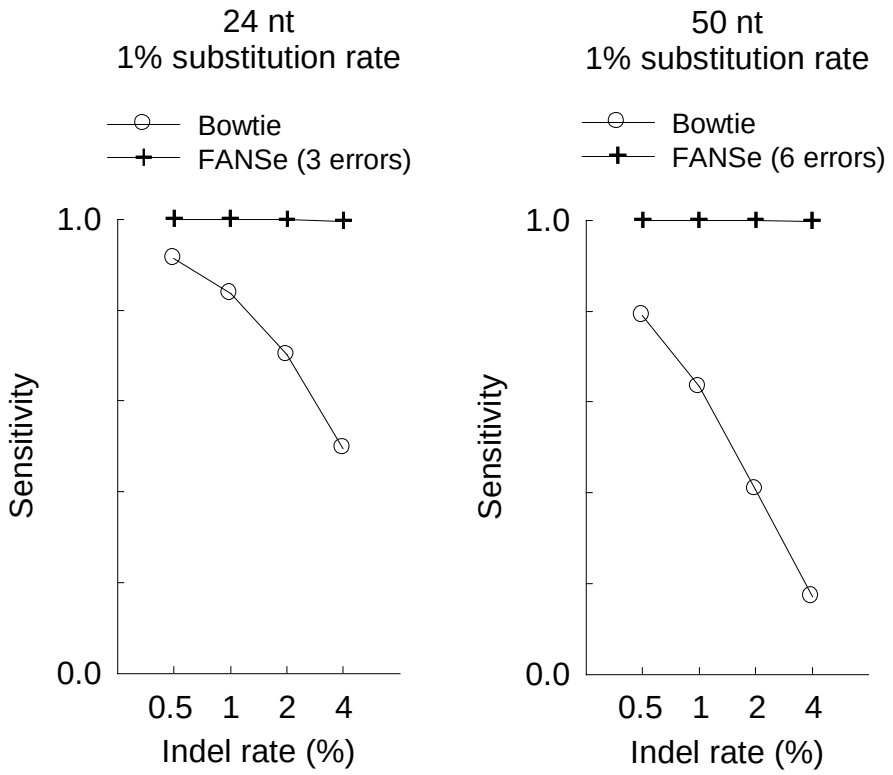
- 在高达4%的测序错误率下，FANSe 依然保持 >99.97% 的灵敏度
- 提高错配容限，可以进一步提高敏感性，同时并不降低准确度——reads仍然会被定位到最优位置上!
- 在高错误率下，其他算法完全不可用。FANSe甚至能对付高达8%的错误率



Zhang et al., Nucleic Acids Research, 2012

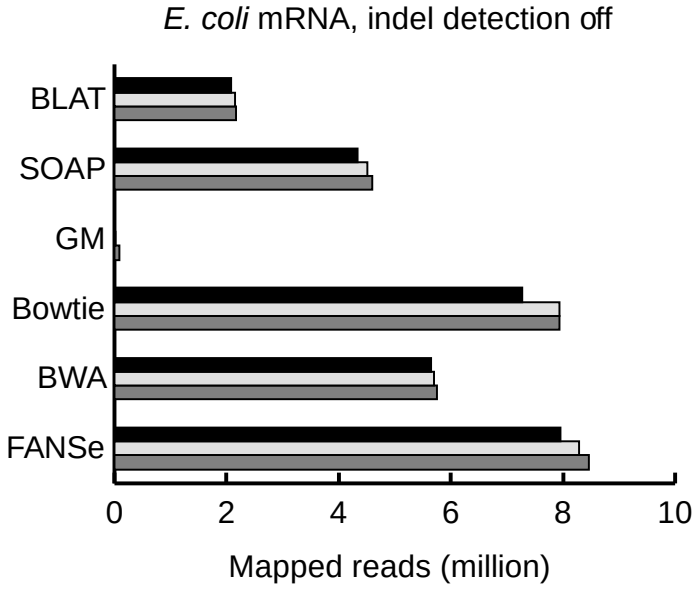
FANSe的准确性远超其他算法（模拟测试）

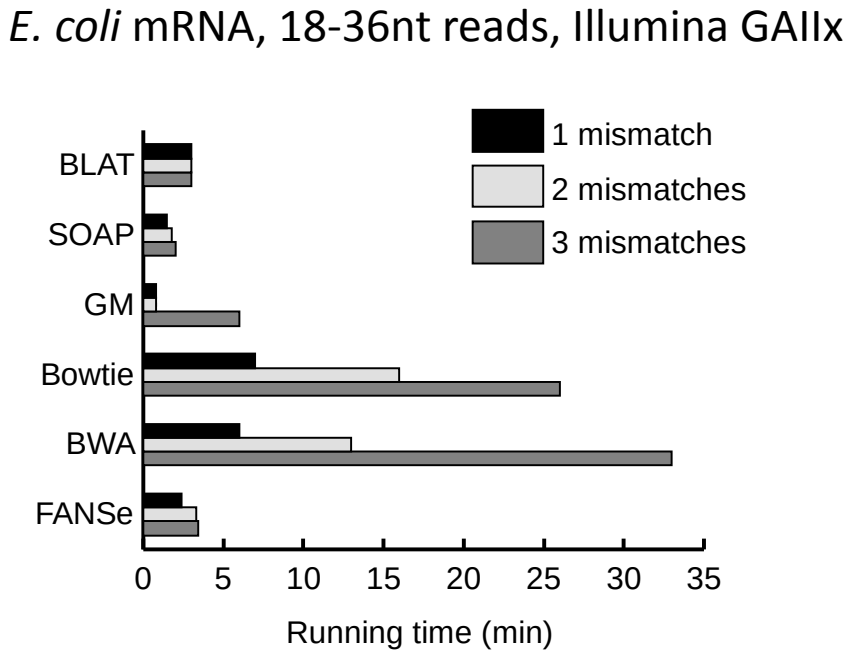
- FANSe 可提供对indel的完美检测。在indel存在的情况下，其他算法几乎不可用，而FANSe的表现近乎完美。

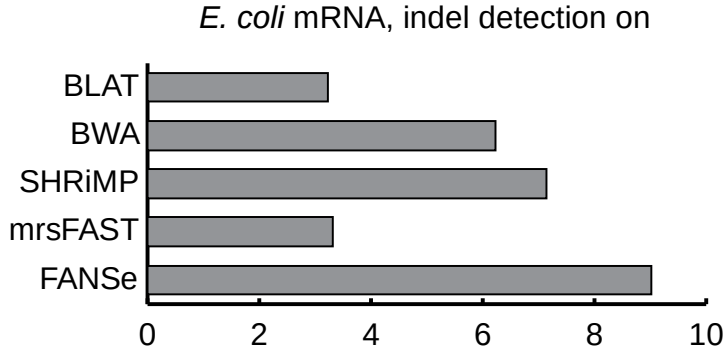


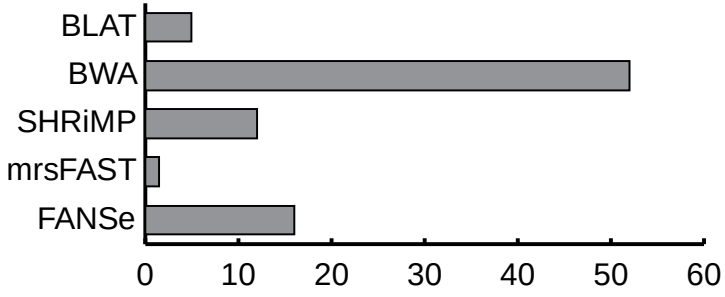
Zhang et al., Nucleic Acids Research, 2012

FANSe的准确性远超其他算法（实战测试）





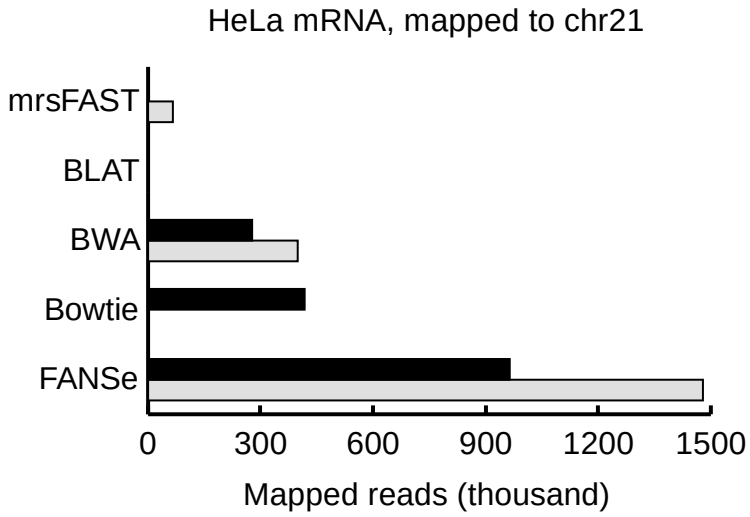


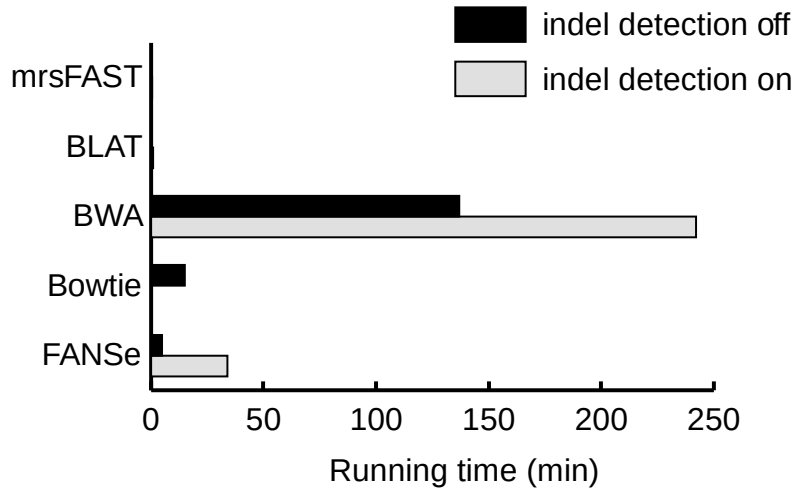


Zhang et al., Nucleic Acids Research, 2012

FANSe的准确性远超其他算法（实战测试）

HeLa mRNA, 18-36nt reads, Illumina GAIIx



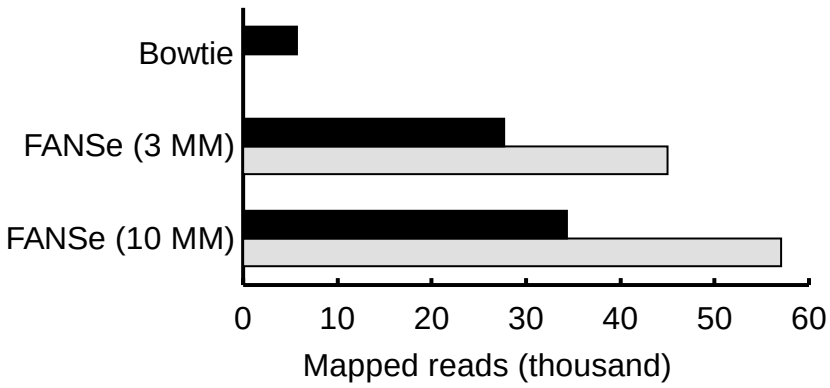


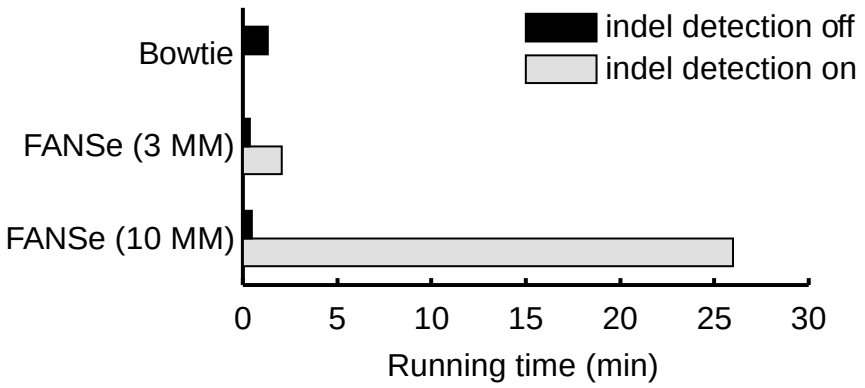
Zhang et al., Nucleic Acids Research, 2012

FANSe同样能对付454平台输出的长reads

- FANSe允许设定任意数量的错配容限，因此更能有效处理长的reads

E. coli genomic DNA, 140-300nt reads, 454 GS FLX





Zhang et al., Nucleic Acids Research, 2012

FANSe将对大规模测序技术带来变革

- FANSe 最精确
  - 可以忽略不计的mapping错误率
  - 迄今为止唯一一种可以预估mapping错误率的算法，极为稳健

- FANSe 适用面最广
  - 适用于除SOLiD以外的所有测序平台（一代、二代、三代测序平台全兼容）
  - 适用于绝大多数RNA/DNA测序应用，并具有极佳的可扩展性
  - 对长短reads全面兼容
  - 对参考基因组完全没有要求
  - 完全兼容masked genome

- FANSe 完全不怕高错误率和 indel
  - 对低质量高错误率平台（如ion torrent）和高indel率平台（如Helicos）是非常大的福音

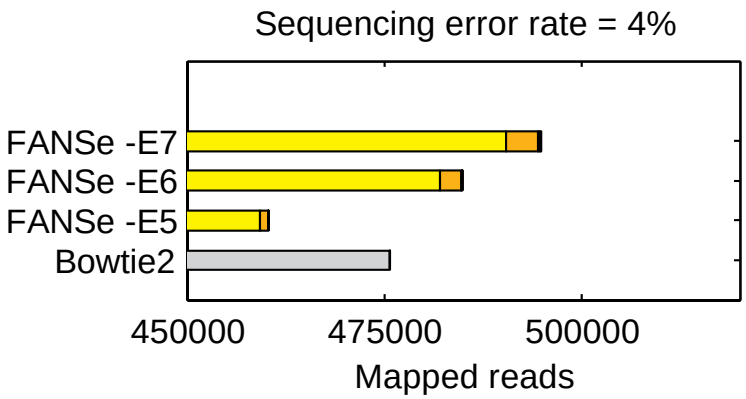
Zhang et al., Nucleic Acids Research, 2012

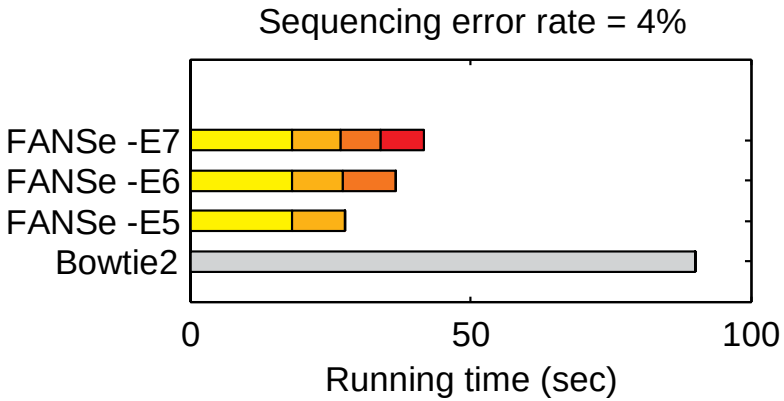
FANSe 的弱项：速度

- FANSe 以准确性为主要设计诉求，速度较慢
  - 不支持多核并行运算
  - 10M reads 向人基因组mapping需要4天

FANSe2: 提高速度，仍然保持准确性

Simulated 500000 75-nt reads mapped to human Chr1
FANSe1 输出结果相同，但需要约15分钟运行时间
FANSe2 快约30倍





FANSe2 强大的并行计算能力

FANSe2 是世界上第一种分布式mapping算法：

- 可在Windows下运行，利用剩余计算能力，不干扰日常工作
- 多机分布式并行运算，任务包形式
- 运算时网络通讯量极少
- 配置简易，无需集群

Illumina HiSeq-2000

608M reads

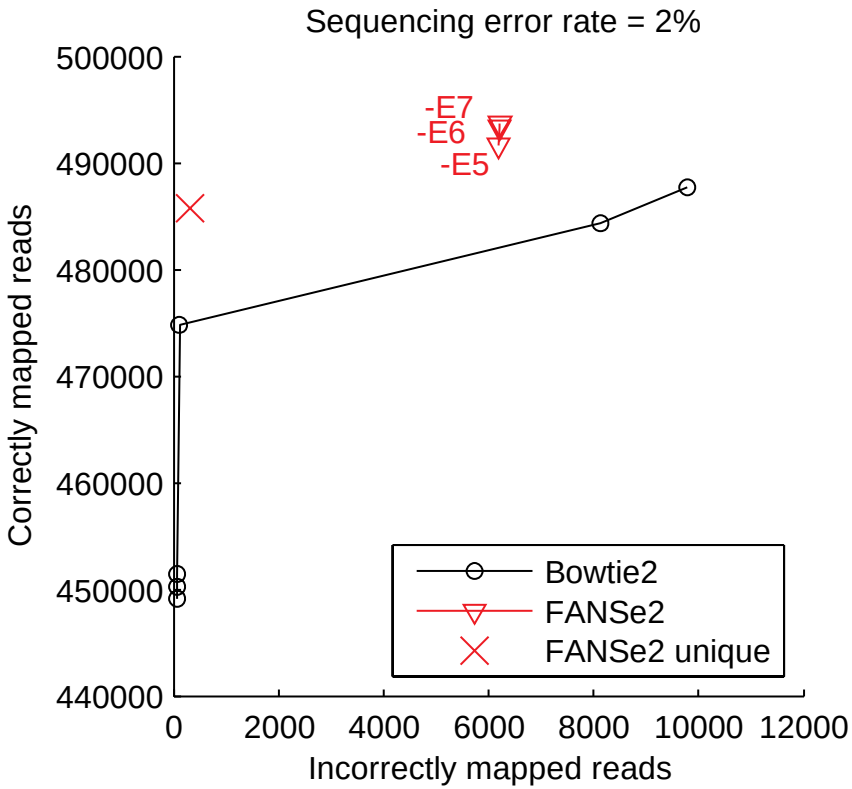
向人基因组比对：

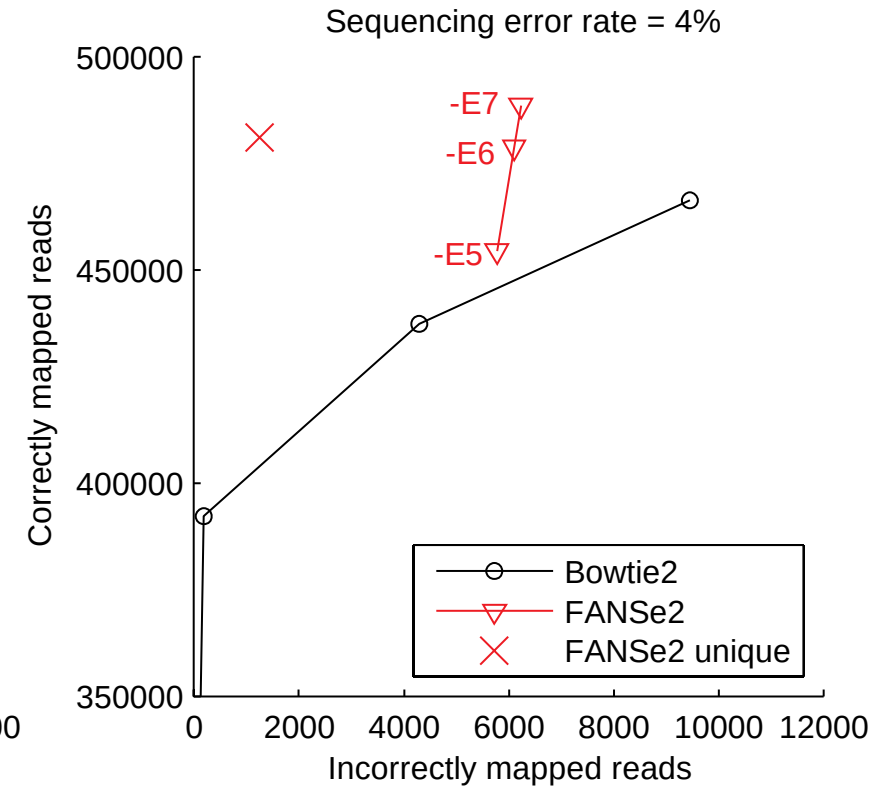
- 单机(4000元组装机)：8小时
- 实验室7台机器并行：1.5小时

Bowtie2: 2天

FANSe2: 提高速度，仍然保持准确性

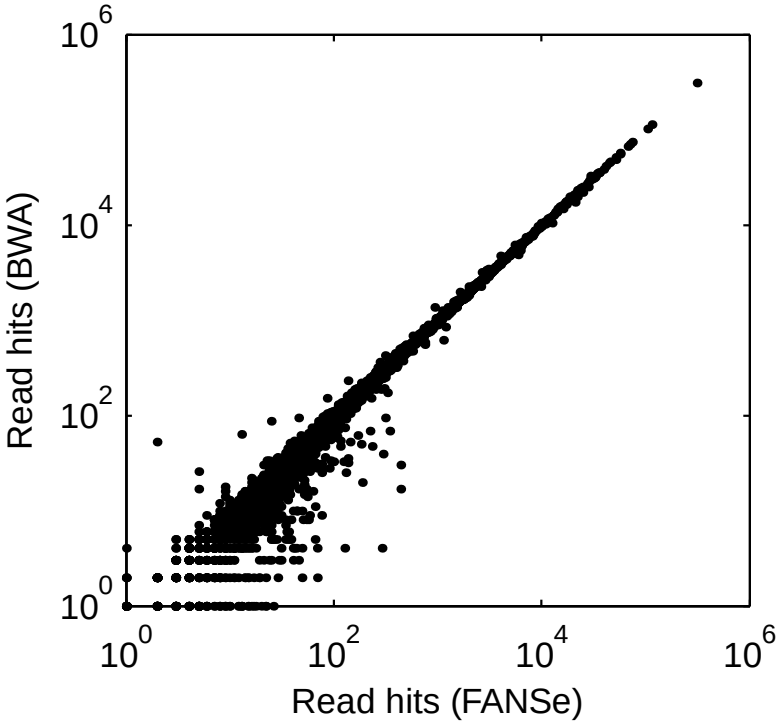
Bowtie2 以丢弃大量reads为代价来保证正确性
FANSe2 稳健性极好，任何情况下都显著优于Bowtie2
且mapping正确性与参数设置关系不大

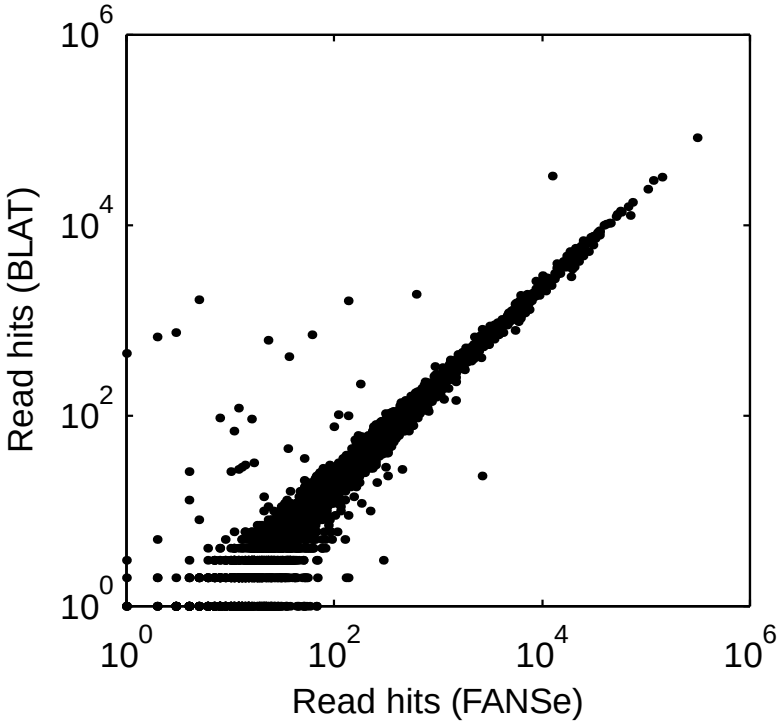




其他mapping算法将会误导RNA定量

E. coli transcriptome quantification



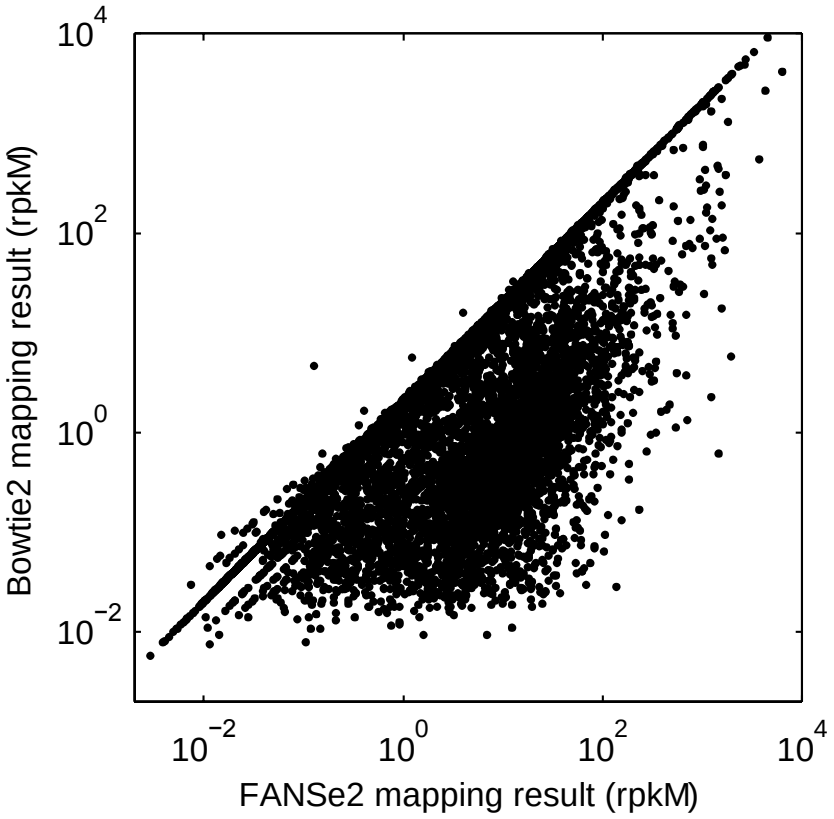


传统算法对低丰度mRNA的定量失准，很可能会误导得出不符合事实的结论！

其他mapping算法将会丢失大量重要信息

真核系统情况尤甚

A549 cell mRNA reads mapped to RefSeq human RNA sequences



一些丰度很高的mRNA不能被Bowtie2鉴定到！

Gene	Function
HNRNPC	Heterogeneous nuclear ribonucleoprotein C
LMNA	Lamin A/C
MORF4L2	Mortality factor 4 like 2
EIF3CL	Eukaryotic translation initiation factor 3, subunit C-like
CNBP	CCHC-type zinc finger, nucleic acid binding protein
RPL13	Ribosomal protein L13

FANSe2 稳定地鉴定到这些mRNA

RT-qPCR结果证明这些mRNA确实大量存在

FANSe/FANSe2 给测序高级分析打下坚实的基础

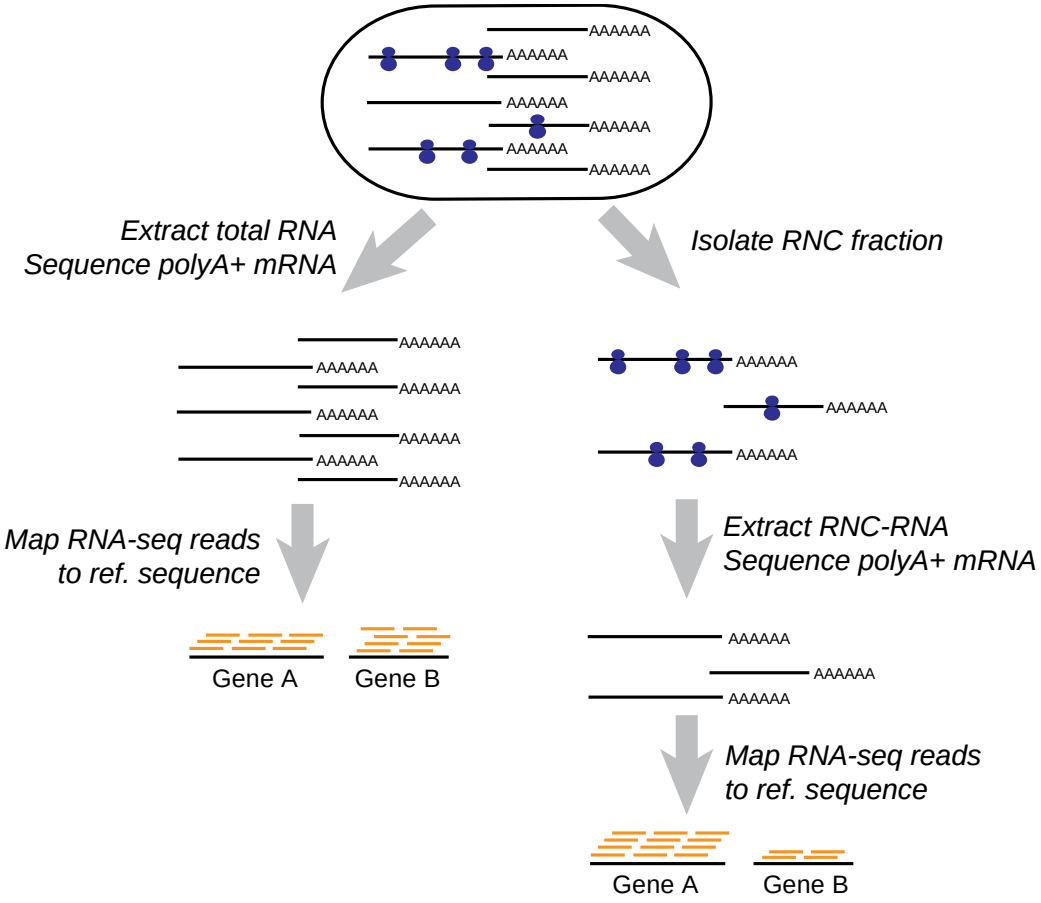
蛋白质组学研究新策略：翻译组测序

转录组测序测的是mRNA

翻译组测序测的是正在翻译的mRNA

翻译组测序可得到哪些蛋白质正在被合成！

难点：正在翻译的mRNA的提取

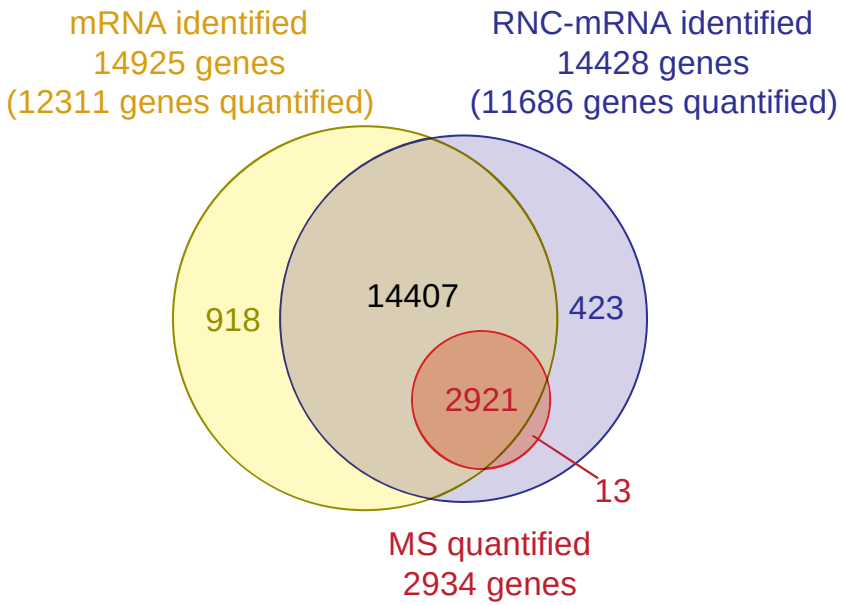


使用 FANSe/FANSe2 进行 mapping

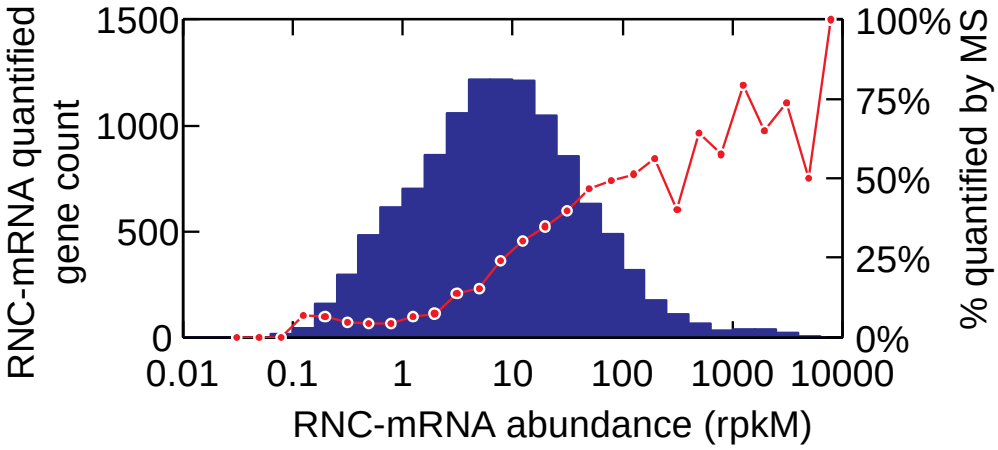
A549肺癌细胞系的翻译组测序结果

翻译组测序灵敏度高

基本上有mRNA就会被翻译

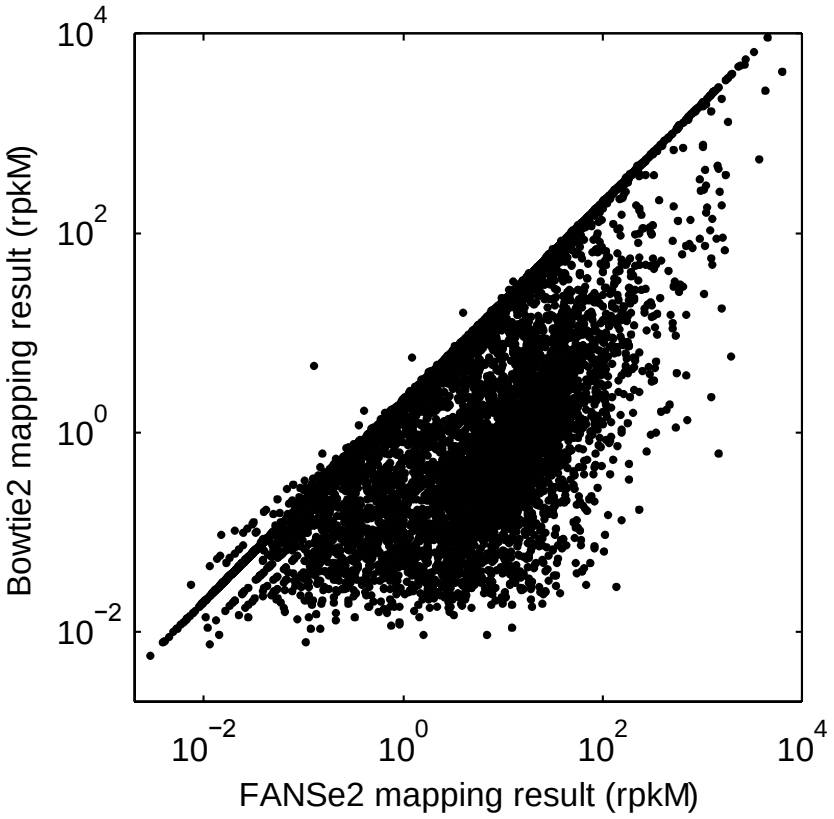


低丰度mRNA的蛋白质质谱鉴定率非常低，但高丰度mRNA的蛋白质质谱鉴定率也只是刚过一半而已。



其他mapping算法将会丢失大量重要信息

A549 cell mRNA reads mapped to RefSeq human RNA sequences



一些丰度很高的mRNA不能被Bowtie2鉴定到！

Gene	Function
HNRNPC	Heterogeneous nuclear ribonucleoprotein C
LMNA	Lamin A/C
MORF4L2	Mortality factor 4 like 2
EIF3CL	Eukaryotic translation initiation factor 3, subunit C-like
CNBP	CCHC-type zinc finger, nucleic acid binding protein
RPL13	Ribosomal protein L13

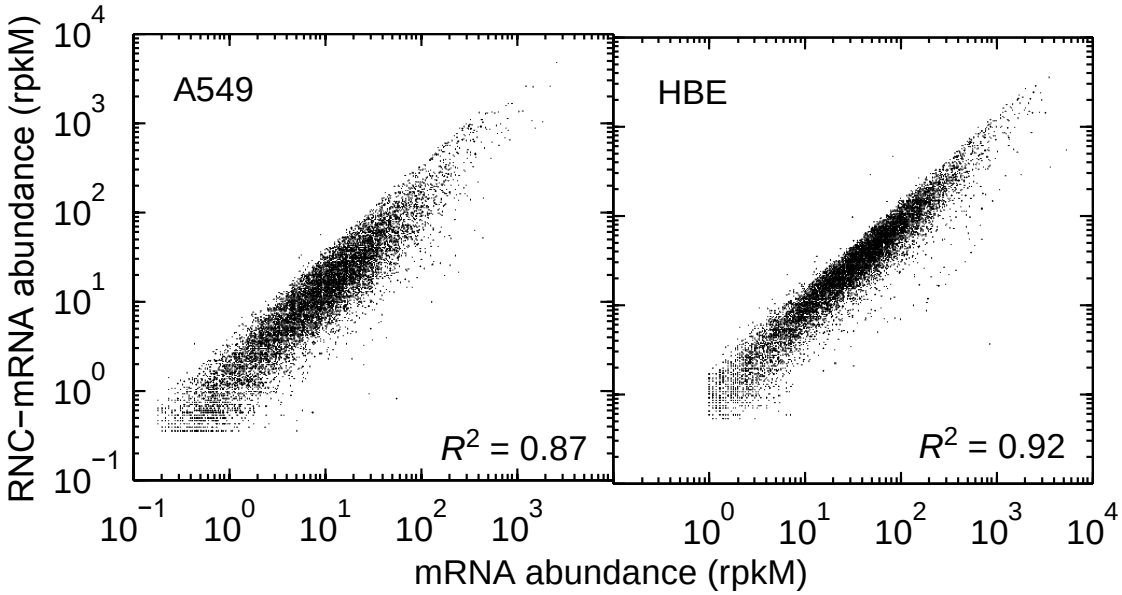
FANSe2 稳定地鉴定到这些mRNA

RT-qPCR结果证明这些mRNA确实大量存在

RNC-mRNA与mRNA的量正相关

RNC-mRNA与mRNA的量正相关——转录量与翻译量基本成正比

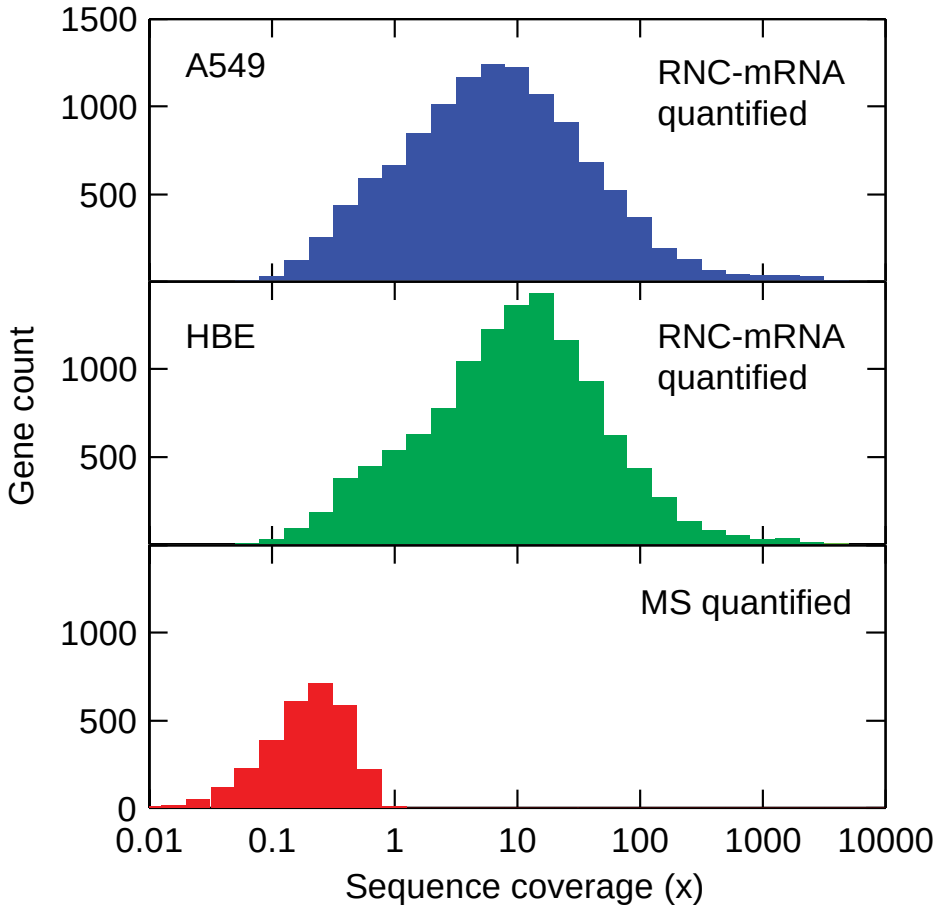
但散布可达一个数量级以上——翻译调控是普遍现象



翻译组测序置信度高

翻译组测序的序列覆盖度很高：

- 9940 genes (85.1%) > 1x
- 5031 genes (43.1%) > 10x



翻译组测序置信度高，并能鉴定新蛋白

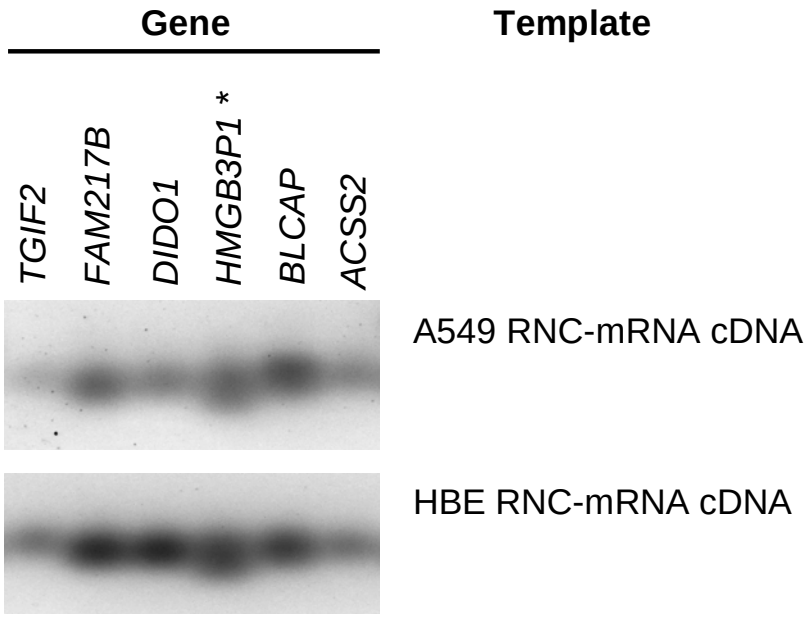
随机选取6个质谱未鉴定到，但翻译组鉴定到的蛋白质，从RNC-mRNA组分中进行RT-PCR验证，均能验证。

翻译量很低的基因(2 rpkM) 都可被RT-PCR验证。

我们还发现了新蛋白：

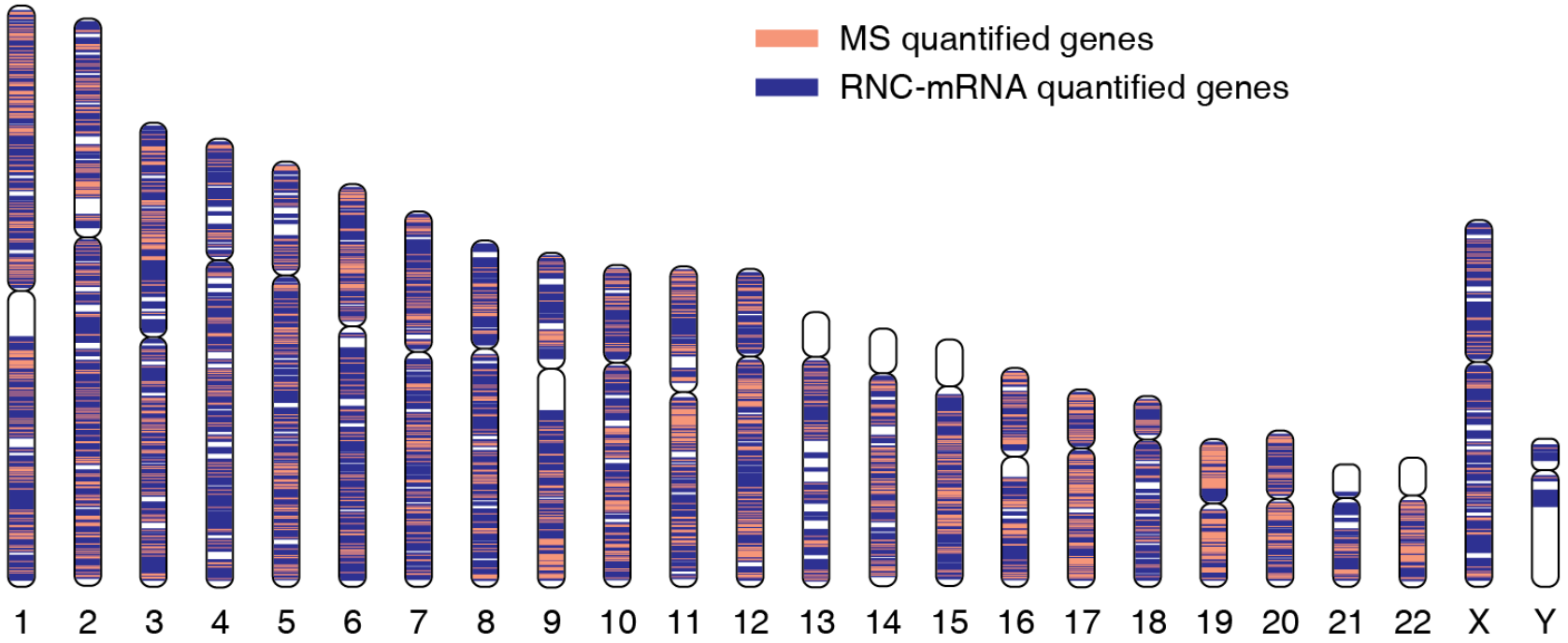
HMGB3P1 (RefSeq: NR_002165)

此前人们认为它是个假基因(pseudogene)，但我们可以证明它正在翻译



翻译组测序极大地加强了染色体上蛋白质的注释

翻译组测序的高灵敏度和高置信度，极大地加强了染色体上各种基因的注释。



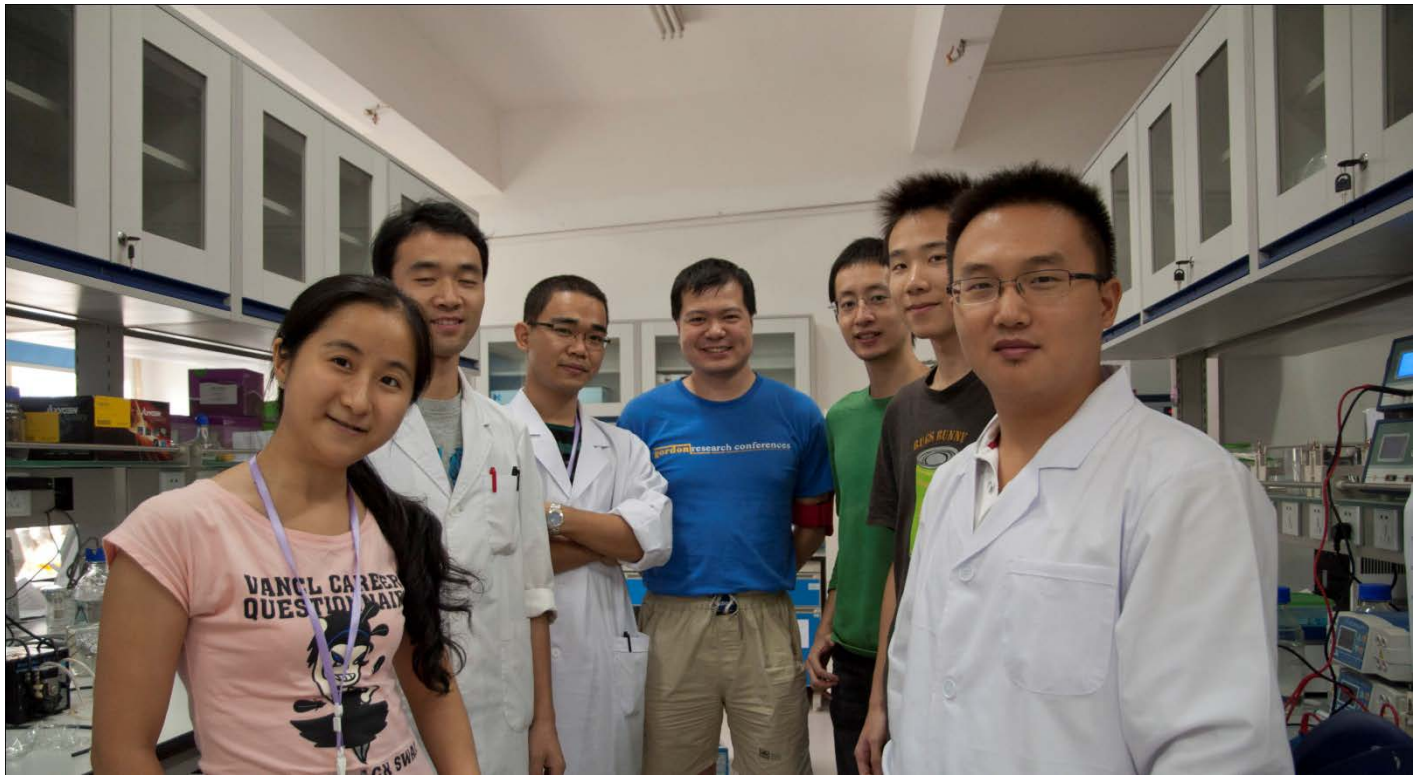
致谢

- 细胞生物学
  - 王通 博士
  - 崔毅峙
  - 郭嘉慧
- 蛋白质质谱
  - 何庆瑜 教授
  - 王贵宾
  - 银兴峰
- 生物信息学与统计学
  - 肖传乐
  - 金静洁

暨南大学生命与健康工程研究院

暨南大学生命与健康工程研究院

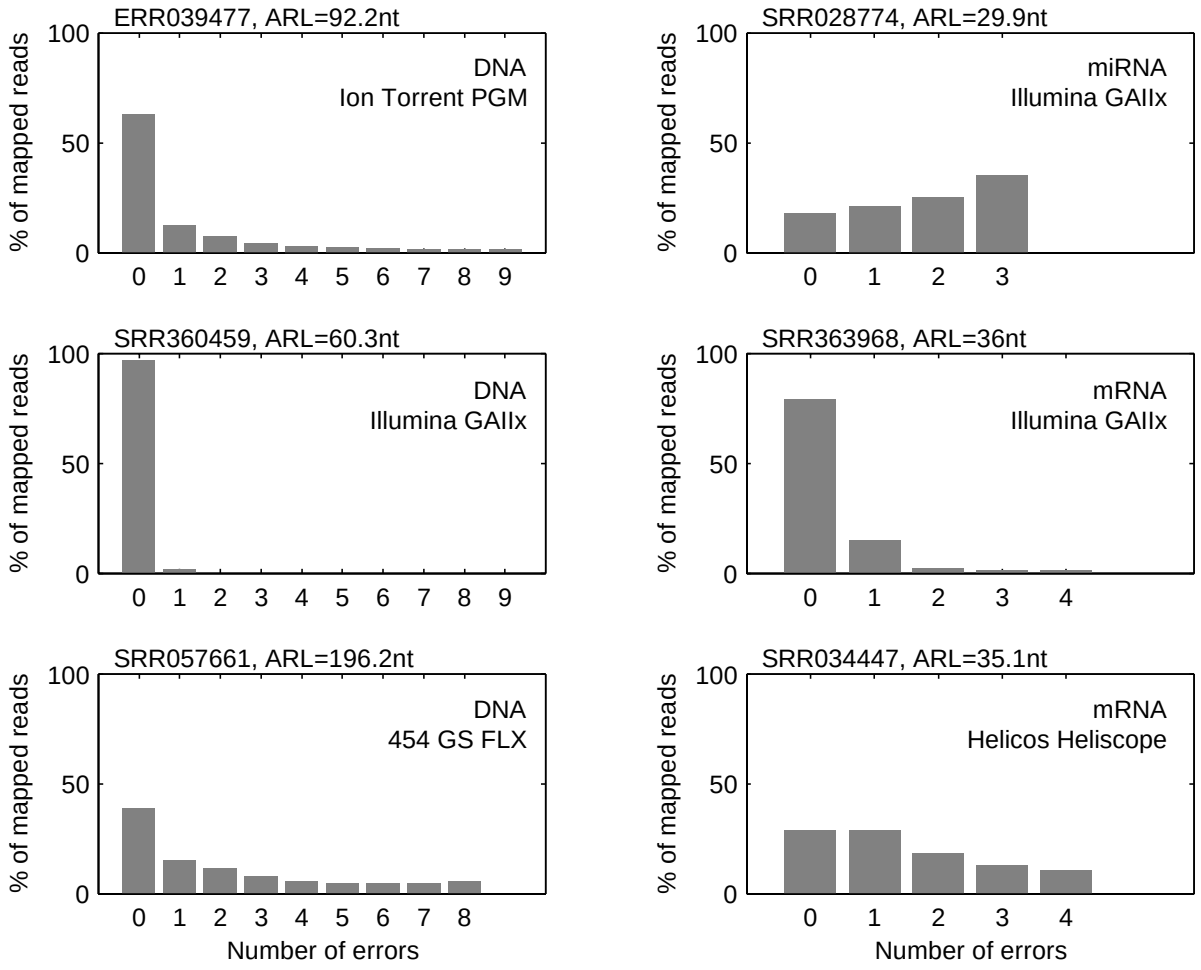
920 翻译组学实验室

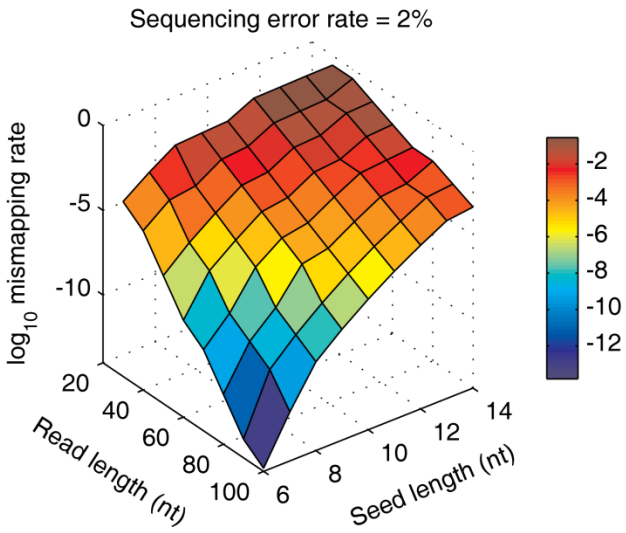


Danke für Ihre Aufmerksamkeit

感谢各位的关注

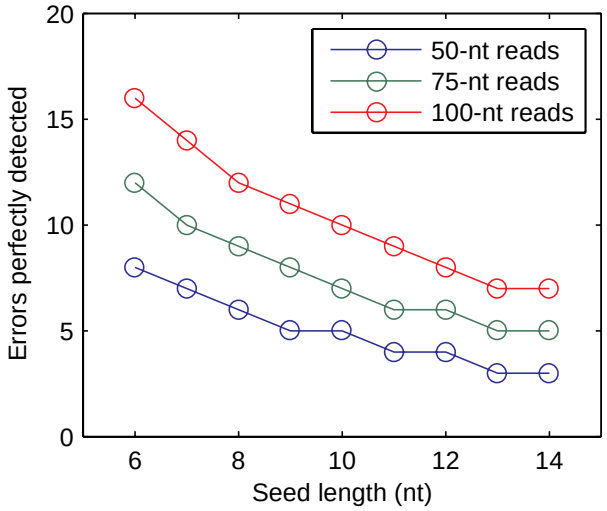
实际状况中绝大部分reads都只含有少量错配





增大Seed length，错误率仍在可控范围内，但速度呈指数上升

长seed未能检测到的reads，继续缩短seed，从而保证速度又不损失精度



即便是更长的seed，对那些含有少量错配的reads仍能完美检测

所以绝大部分reads都可以在长seed阶段被mapping到