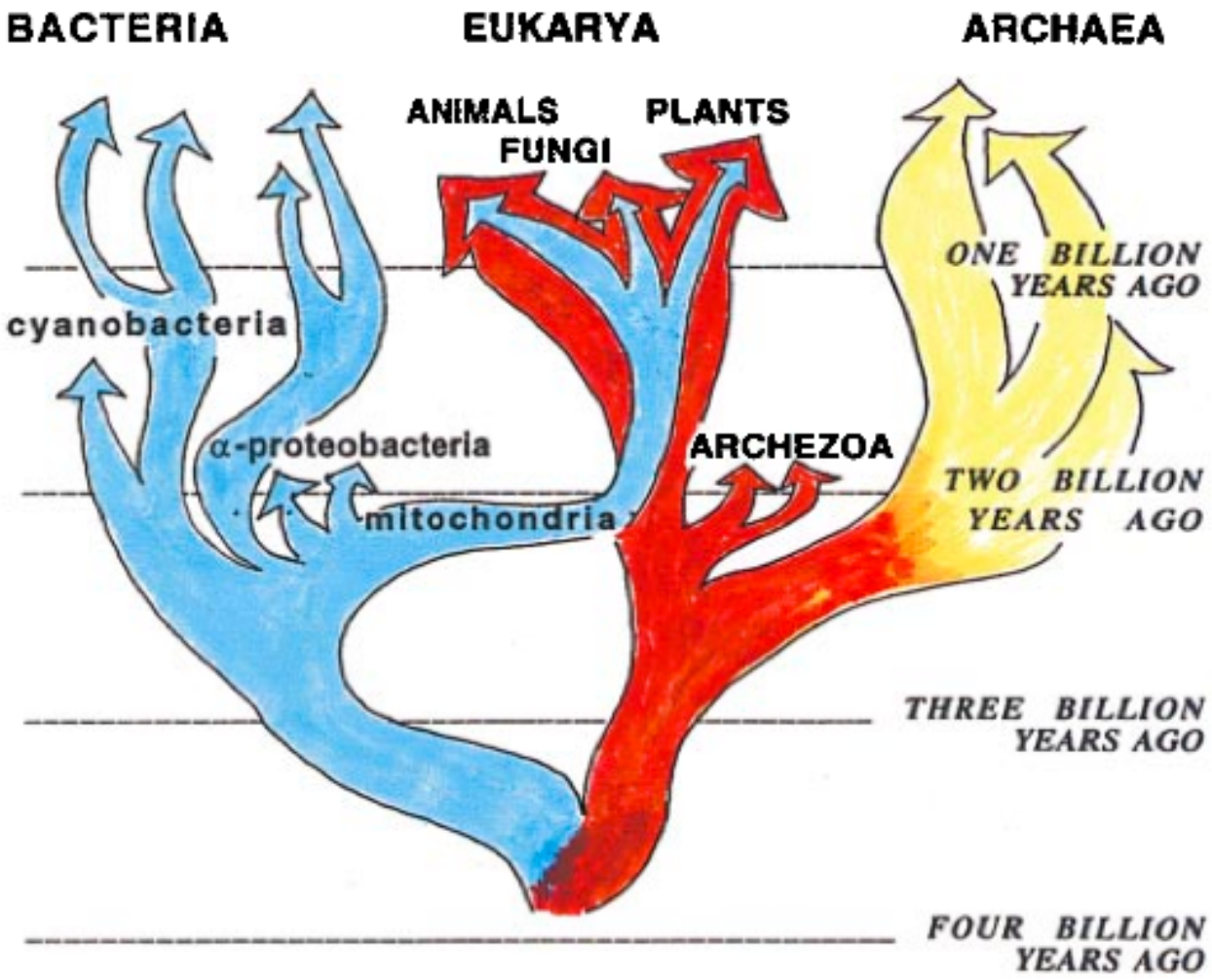
基因组数据分析解读及实例操作

罗奇斌
奇云诺德QY NODE
德国慕尼黑工业大学

为什么要进行新物种的基因组测序？



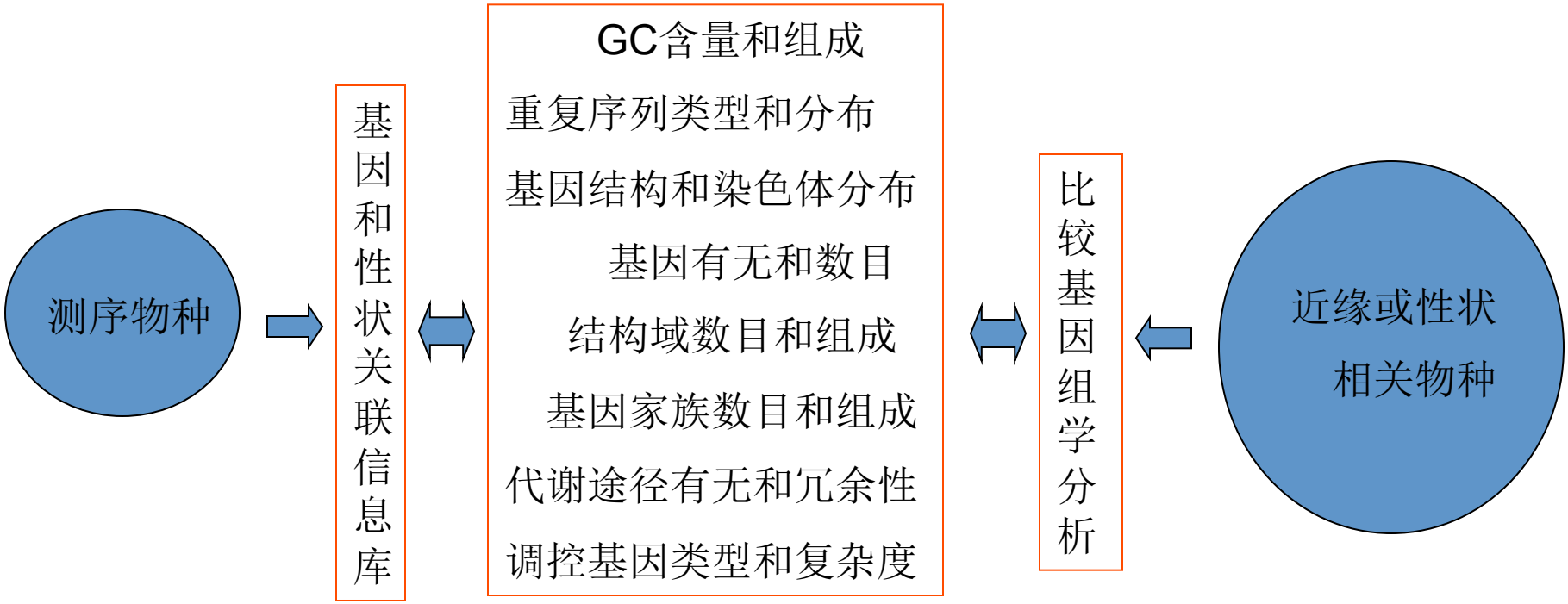
物种重要性及其测序必要性

模式生物；病原性；经济性状；进化地位

1.基因组测序，需利用拼接，基因组注释(RNA，蛋白质和重复序列等)；

2.关键性状的遗传因素 (功能基因组学确定性状决定的基因或基因群)；

3.比较基因组学分析--------进化地位，性状产生和进化的基因组线索等



高通量测序时代全基因组测序发展方向

- 发展针对不同特征的基因组测序策略；

- 经济或病原物种，自成模式生物，基因组数据作为研究基础，如利用功能基因组学手段大规模发现基因-表型关联；
- 针对科学问题，假说优先，如直立人基因组测序计划

Next Generation Sequencing and Assembly

- Platforms and strategies: NGS -- SOLiD, Illumina and Roche; WGS – whole genome shotgun;

454 GS FLX

SOLiD 4.0

GA IIx

Recent *de novo* assemblies

Table 1. Assembly statistics for maize, horse, panda, blue-stain fungus (*G. clavigera*) and *P. syringae* genomes and their cost

	B73 maize	Domestic horse*	Giant panda†	*G. clavigera*†	*P. syringae*†
Genome length	2.3 Gb	2.5-2.7 Gb	2.4-2.5 Gb	32.5 Mb	6.1 Mb
Sequencing technology/ies	Sanger	Sanger	Illumina	Sanger, 454, Illumina	Illumina
Number of contigs	125,325	55,316	198,274	3,361	1,346
Contig N50	40 kb	112 kb	40 kb	32 kb	11 kb
Number of scaffolds	61,161	9,687	81,469	2,322	71
Scaffold N50	76 kb	46 Mb	1.3 Mb	782 kb	317 kb
Estimated sequencing cost	$30 million	$15 million	$0.6 million	$100,000	$4,000

Contiguity statistics are calculated for *contigs and scaffolds 1 kb or longer and †contigs and scaffolds 100 bp or longer.

Second generation sequencers

Roche	454	1	Metagenomics De novo sequencing RNA-seq	Novel genome(s)
illumina	Solexa	3	De novo sequencing RNA-seq, Re-sequencing ChIP-seq， Meth-seq	Both types
Applied Biosystems	SOLiD	5	Re-sequencing ChIP-seq RNA-seq	“known” Genome

通用高性能集群计算环境

- 运算速度快
- 节点多
- 单节点内存大（至少2～4GB/Core）
- 节点内部网络通路无特殊要求

- 应用（NGS数据分析）
  - 全基因组序列比对分析（BWA，SOAP，Bioscope，Bfast，…）
  - 转录组序列比对分析
  - …

专用计算环境

- 大内存服务器
- 128GB/256GB/512GB/1TB

- 用途
  - 大基因组拼接
  - 转录组拼接
  - 基于大数据量的功能分析
  - …

存储资源环境

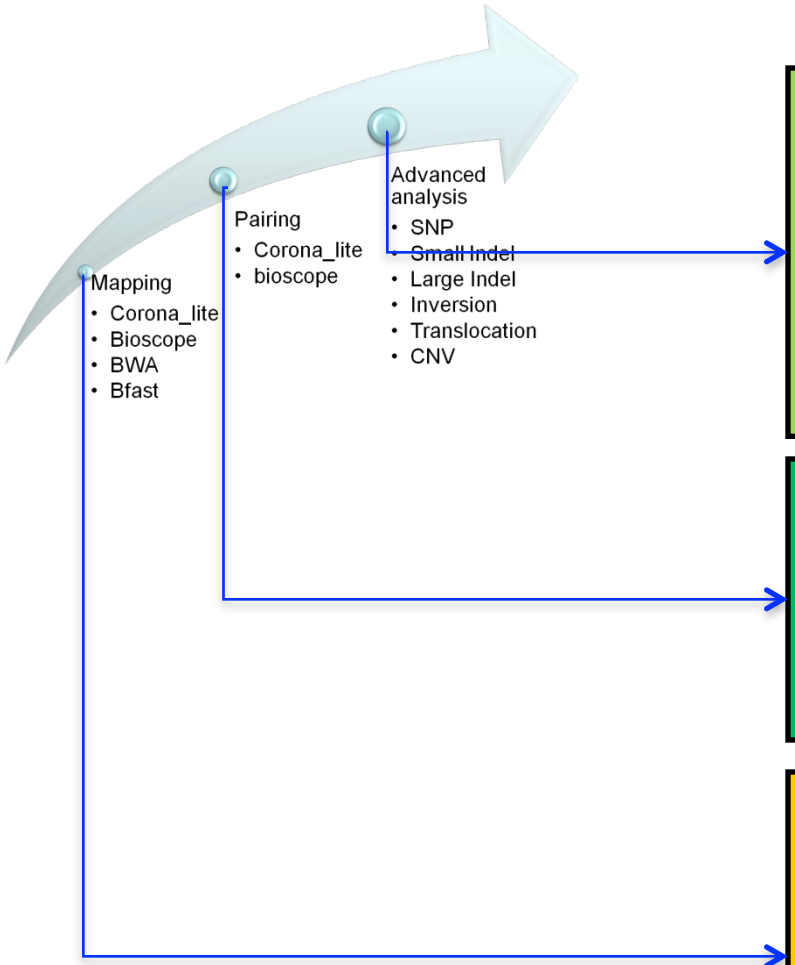
- 分级存储
  - 普通存储 + 高性能存储
  - 高性能存储 + 磁带库存储
  - 普通存储 + 高性能存储 + 磁带库存储
- 基因组所状况
  - 普通存储：700TB（参与小数据计算）
  - 高性能存储：200TB（参与大数据计算）
  - 磁带存储：若干（备份数据）

QY NODE

CPU: 960(core)

Aggregation Capabilities:10. 2TFLOPS

计算资源需求及特点



Advanced analysis：

在Mapping或者pairing 结果上进行计算处理，计算速度取决于分配的资源，有写分析需要一定的内存消耗。

Pairing:

在Mapping结果上进行计算处理，计算资源占用小, 单节点即可完成，内存耗用小，运算速度快。

Mapping:

计算资源占用大, 需要大量夸节点运算，并且单节点内存需求大（如16G），mapping 速度取决于计算资源。

计算资源需求及特点

基因组数据的前处理及组装对计算要求及特点：

❖单节点，多CPU任务，

❖内存消耗完全取决于所测物种基因组的大小及数据的乘数

❖对程序及算法依赖性较强，目前正在开发跨节点算法。

比如：

细菌基因组 （3Mbp），所需内存 500Mb

简单植物如水稻 （500Mbp），所需内存 128Gb

复杂动物如对虾 （1.8Gbp），所需内存 512Gb ~ 1Tb

计算资源需求及特点

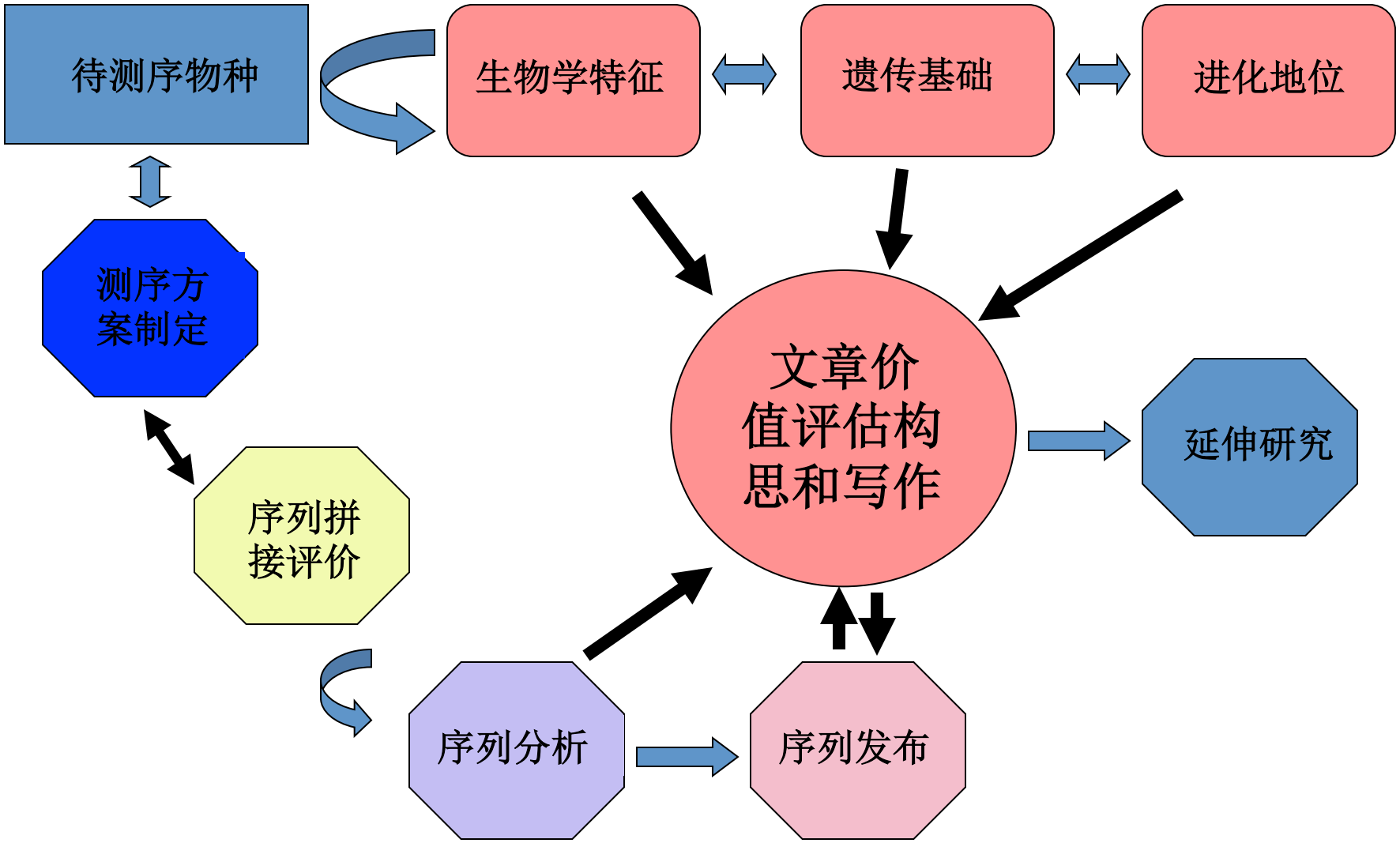
Database

资源取决于数据库用途及用户需求。对于二代数据用户，需要足够的磁盘存储空间，个人用户至少10Gb以上。处理速度取决于是否并行计算。

Software

开发者可以以少量数据做前期开发，需求资源较少，但最终必须要大量数据做稳定测试等，这将大大提高资源要求。

基因组从头测序线路图



基本运算流程的开发和固定

流程开发方式：

广泛调研，集思广益，以完成项目为先期目标(不完美但实用)

基础软件的选择：

最权威(引用)作者，横向比较文章，最经典文章，自行横向比较，公共论坛

固定流程要求：

尽量自动化，傻瓜也能用，包含精炼的输入输出样式；
基础软件版本信息，详细使用说明(使用流程的数据前提)；
结果说明，可能出错的检查点是什么；内置评估脚本或策略说明

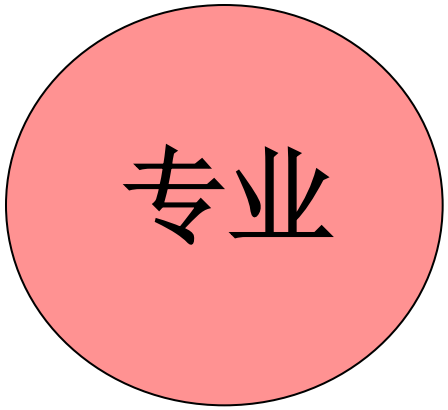
数据分析要求：

与统计检验和生物学问题紧密结合，
一站式解决问题，方便文章写作；
详尽动态可追溯数据库；

相关专业杂志浏览，相关文章或信息公开制度

常用数据库列表及其更新检测制度

常用软件列表及其更新检测制度



Sequencing Glossary

Reads. A collection of clones that over-sample the target genome.

Pair-end reads. Sequence reads derived from both ends of a sequencing-library clone.

Mate-pair reads. Sequence reads derived from both ends of a mate-pair library clone which insert size is usually >1kb.

Insert size. The size of the clone-insert from which a clone-end pair is taken.

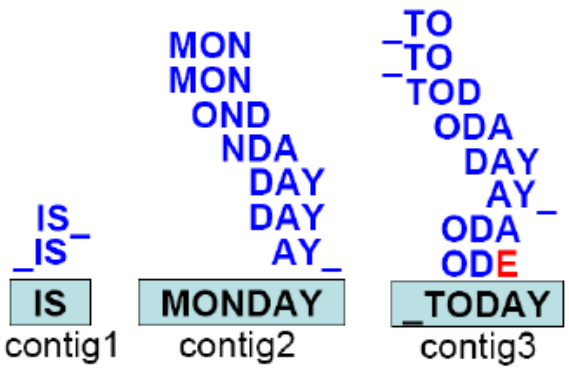
Contig. The result of joining an overlapping collection of sequence reads.

Scaffold. The result of connecting non-overlapping contigs by using pair-end reads.

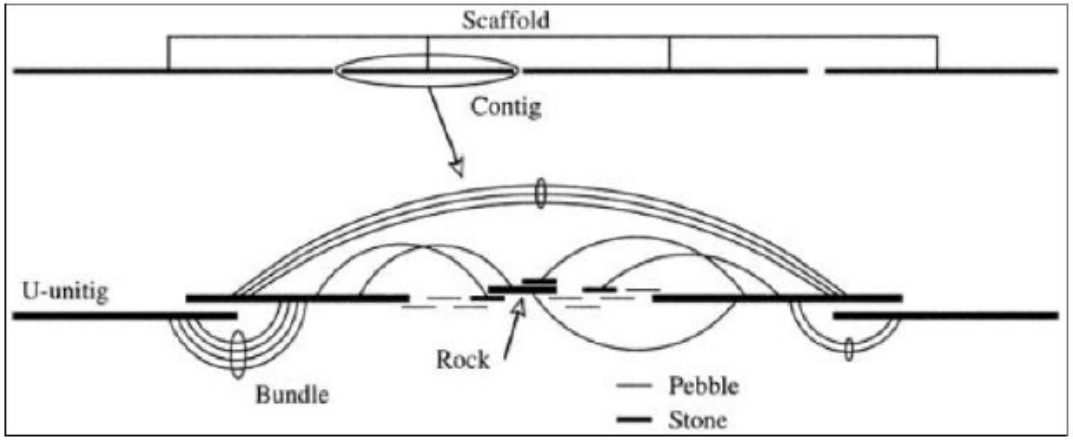
N50 size. As applied to contigs or scaffolds, that size above which 50% of the assembled sequence can be found.

Genome assembly strategy

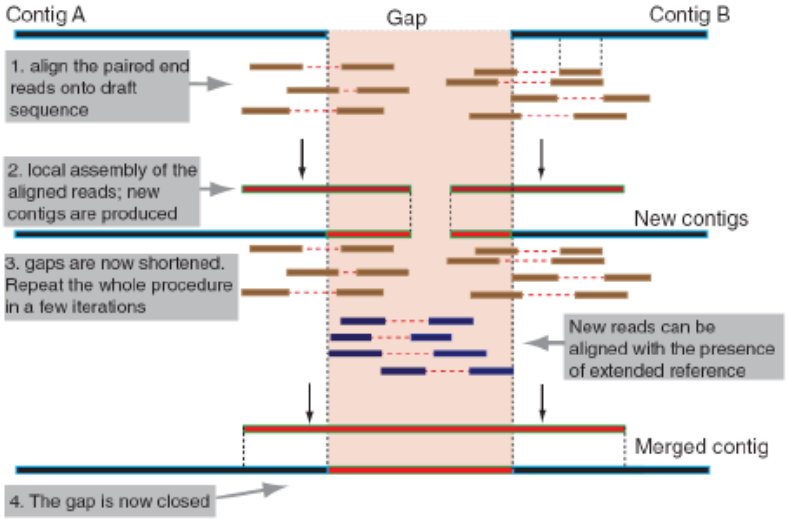
de novo assembly



Contig assembly



Scafffolding



Internal gap closing

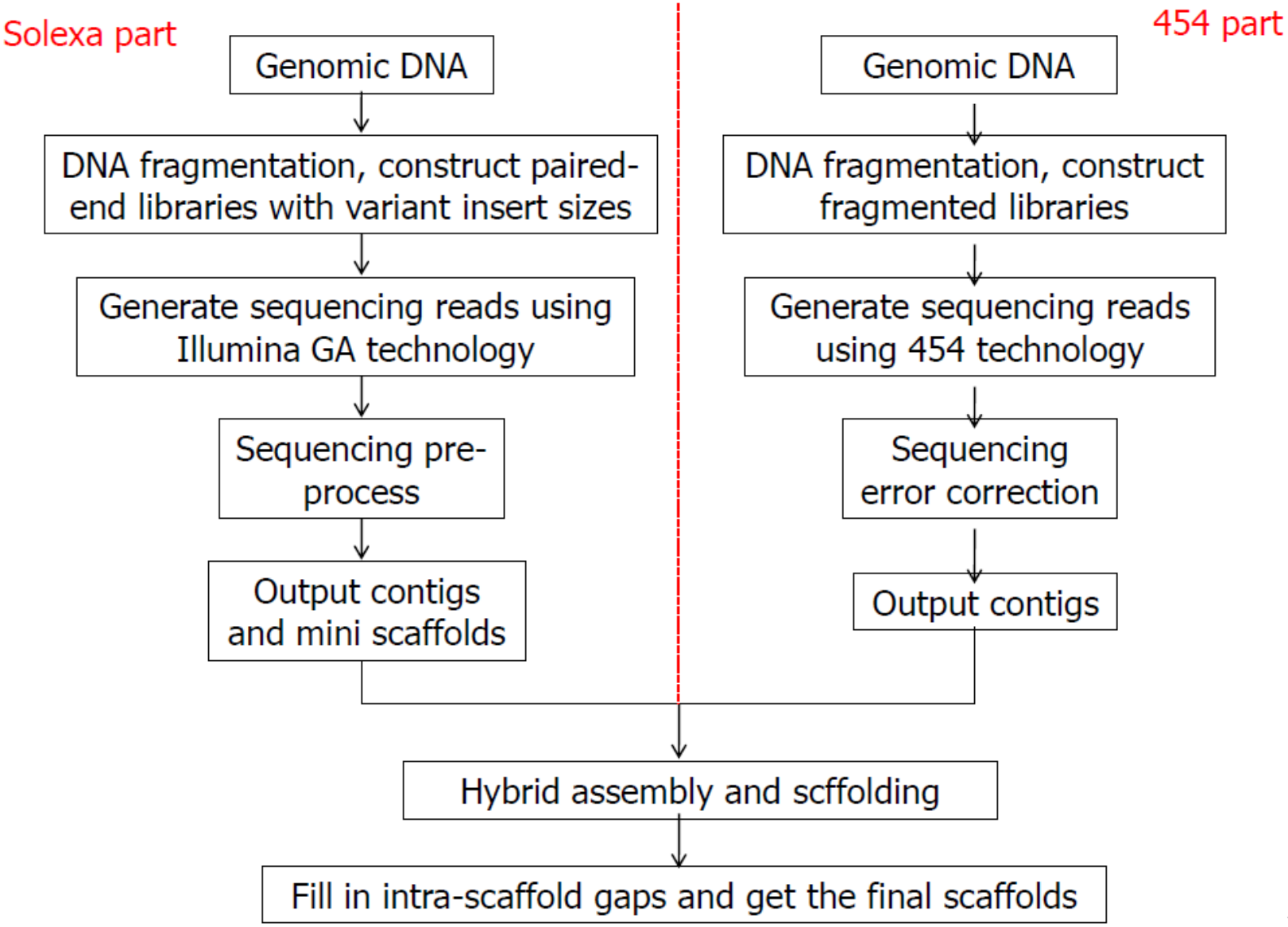
http://genomebiology.com/2010/11/4/R41

Recent whole genome sequencing projects

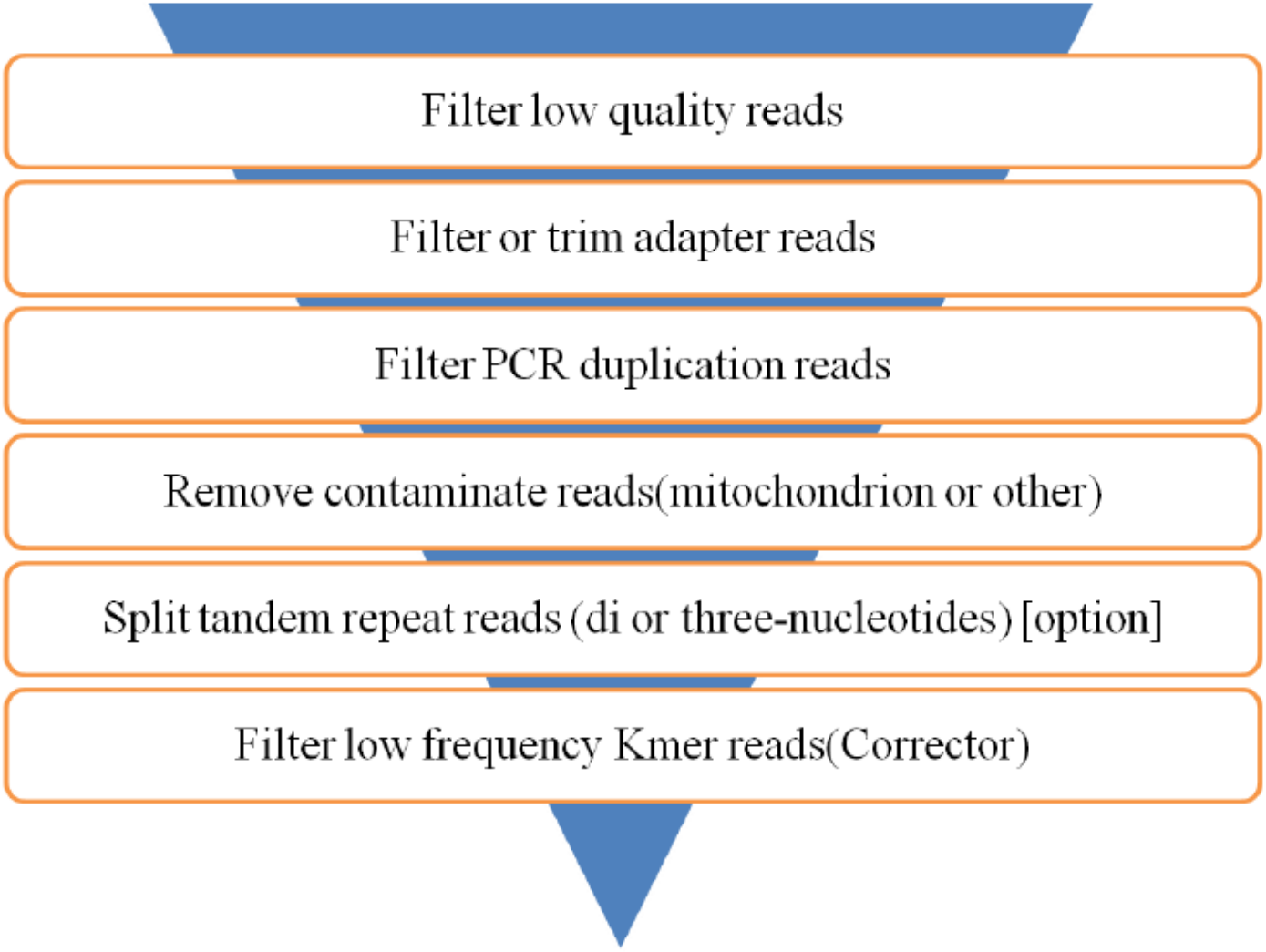
Table. Basic information of Rrecently sequenced genomes.

Organism	Genome size	strategy	Coverage	Contig				Scafffolds			
				#	N50	Max	Total	#	N50	Max	Total
Human	3.0Gb	Solexa	45x	2.76M	1.5Kb	18.8Kb	2.18Gb	NR	NR	NR	NR
Apple	742.3 Mb	Sangr+ 454	4.4x+ 12.5x	122,146	16,171	NR	603.9Mb	1,629	102Kb	NR	598.3
Castor	320Mb	Sanger	4.59x	54,000	21.1kb	190kb	324Mb	25,828	496.5kb	4.7Mb	350.6Mb
Grapevine	500Mb	Sangr+ 454	7x+4.2x	58,611	18.2Kb	238kb	531Mb	2,093	1.33Mb	7.8Mb	421Mb
Panda	2.4Gb	Solexa	74x	200,604	36,728	434,635	2.25Gb	81,496	1.22Mb	6.05Mb	2.30Gb
Straberry	220Mb	454+solexa+solid	24.5x+6.4x+6.4x	16,487	28,072	215,349	202Mb	3,263	1.44Mb	4.1Mb	214Mb
Cacoo	430Mb	454+sanger+solexa	16.7x+ 44x	25,912	19.8kb	190Kb	291.4	4,792	473.8Kb	3415Kb	326.9Mb
Tomato	900Mb	454+sanger+solexa+solid	31x+3.6x +82x+ 140x	110,872	55.7kb	NR	763Mb	3,761	4.45Mb	NR	782Mb
Potato	840Mb	454+solexa+solid	11x+106x +0.2x	111,187	31Kb	NR	683Mb	66,301	387Kb	NR	727Mb

Flowchart of the WGS *de novo* assembly



raw data pre-process



Quality Control

- GERALD Summary.htm

Lane	Lane Yield (kbases)	Clusters (raw)	Clusters (PF)	1st Cycle Int(PF)	% intensity after 20 cycles (PF)	%PF Clusters	% Align (PF)	Alignment Score (PF)	%Error Rate (PF)
1	526305	97464 +/- 4878	87676 +/- 9219	75 +/- 21	86.17 +/- 5.25	89.76 +/- 5.95	99.06 +/- 0.25	102.41 +/- 1.62	1.30 +/- 0.22

Fastq and Quality

Solexa reads of the Fastq format

s_1_1_sequence.txt…

```
@HWI-EAS724_0001:8:32:374:374#0/1
GAGCTGTATATGAATAATAGTTCGTTTTTCATTATCCAAGATGGATCGGTATAAAGTCTGCTAAAATAAAGGTACAACG
+HWI-EAS724_0001:8:32:374:374#0/1
fcfcfggdfggggfggggcgggggggfgggggcgggfWgggggggggfgcggdgcgcggggfacbbb][bgcgggggd
```

s_1_2_sequence.txt …

```
@HWI-EAS724_0001:8:32:374:374#0/2
TACCGTTAATAGCAGTAATATCATAATAGTAATAGCATCATAACGGTAGTCCCATAAAAGTGTGTCAGTAGTAGTAGTA
+HWI-EAS724_0001:8:32:374:374#0/2
ggggfgggggd_adcggggeggfggeggegf`geececdeggggfegcfegggegggfgac[aced`bd__\_c[[Yb
```

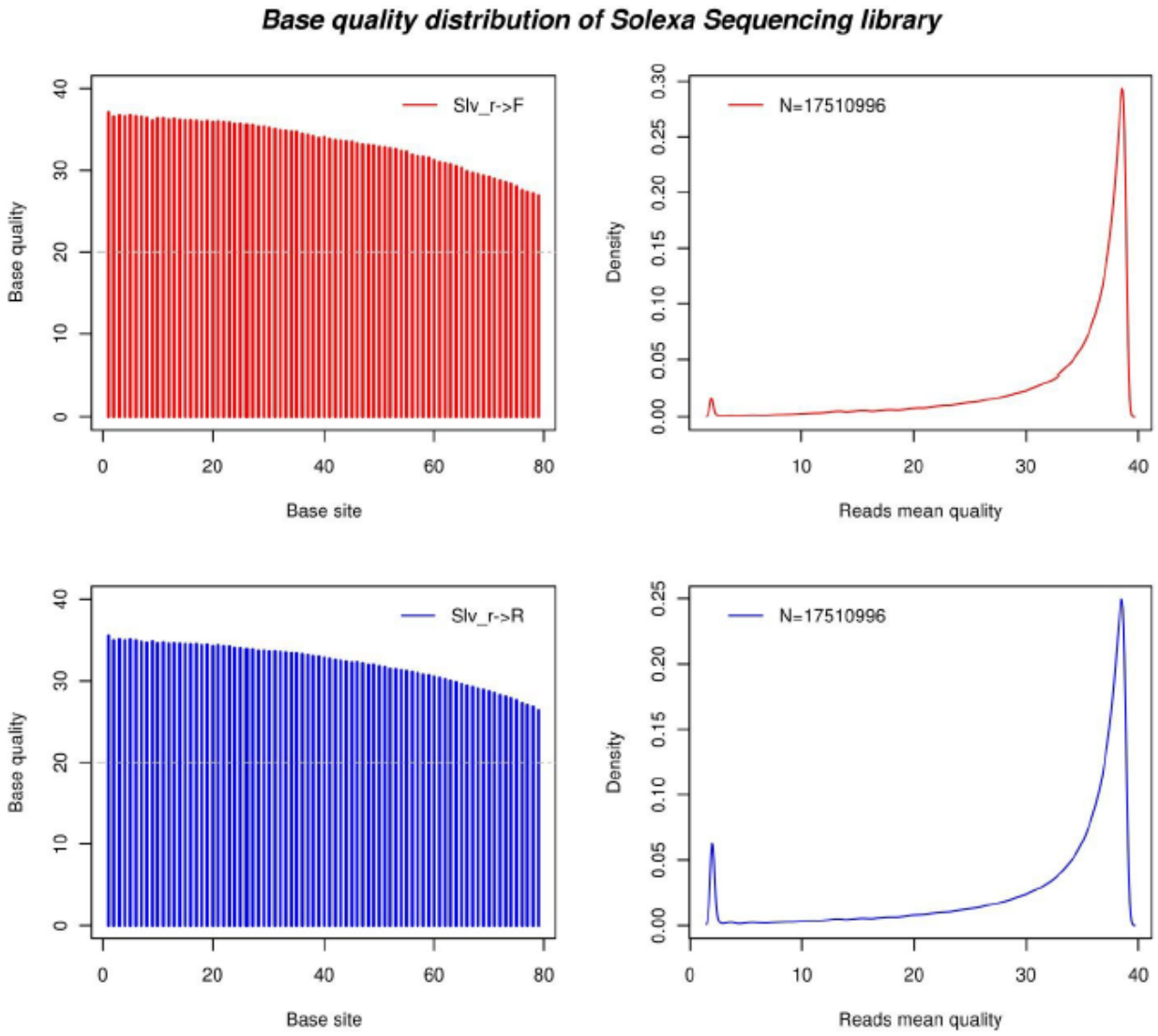
Illumina 1.3 format encodes a Phred quality score from 0 to 40 using ASCII 64 to 104

error probability (p):

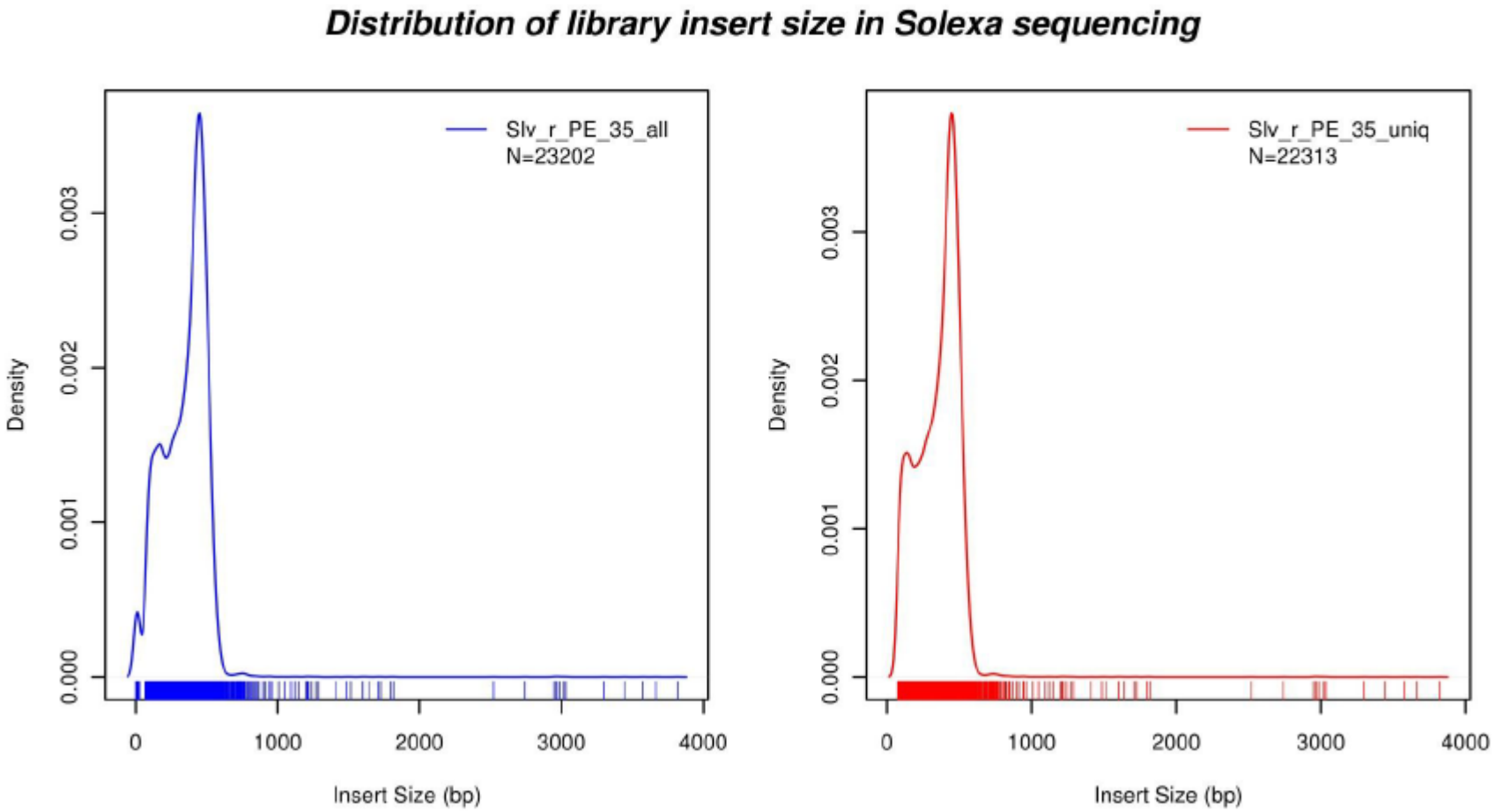
$$Q_{\text{sanger}} = -10 \log_{10} p \qquad Q_{\text{solexa-prior to v.1.3}} = -10 \log_{10} \frac{p}{1-p}$$

for solexa: p = 0.01, Q = 19; p = 0,05, Q = 12.8, p = 0.10, Q = 9.5;
for phred: p = 0.01, Q = 20; p = 0,05, Q = 13, p = 0.10, Q = 10;

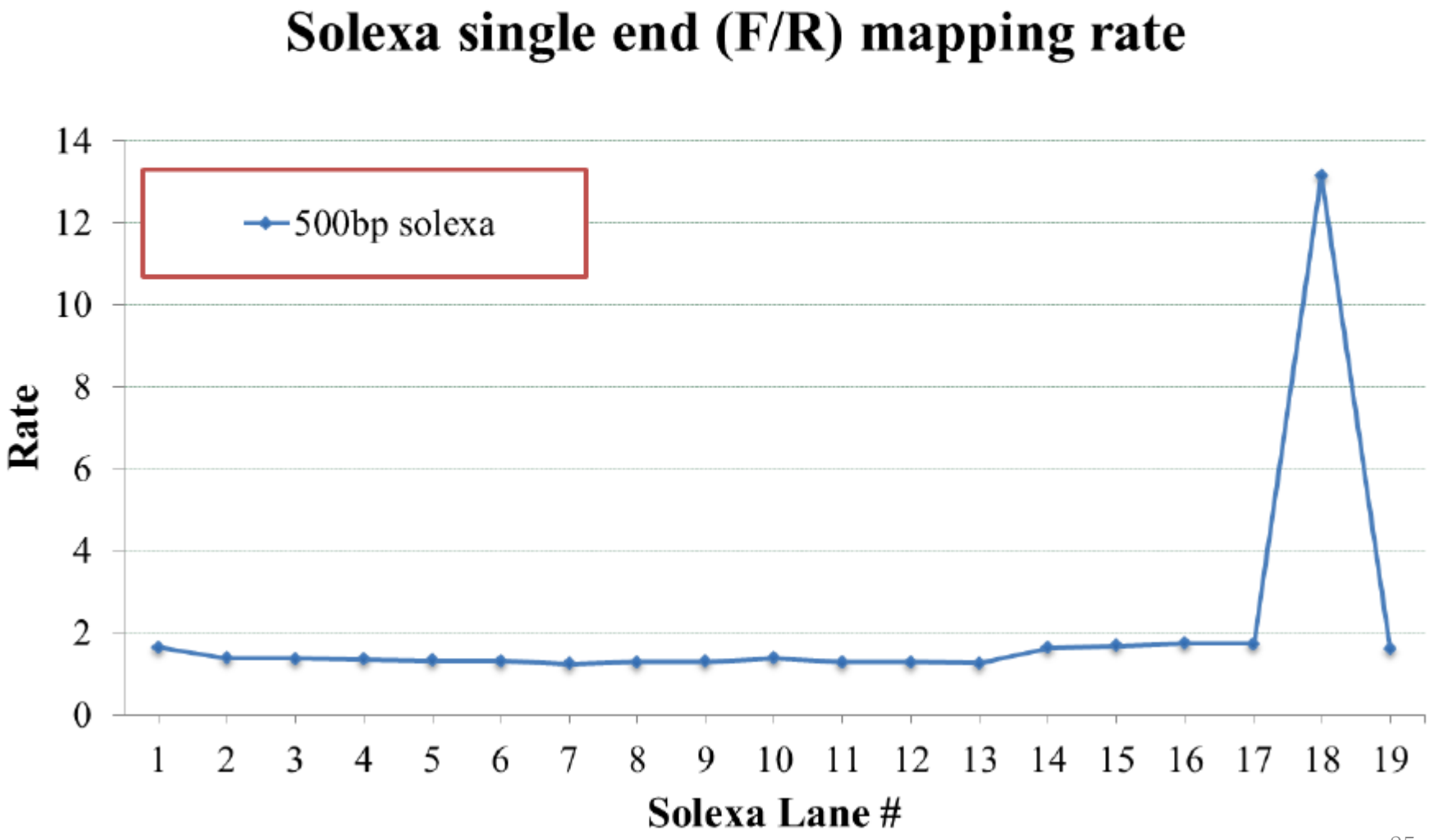
Data assessment I – Read quality distribution



Data assessment II – Library insert size

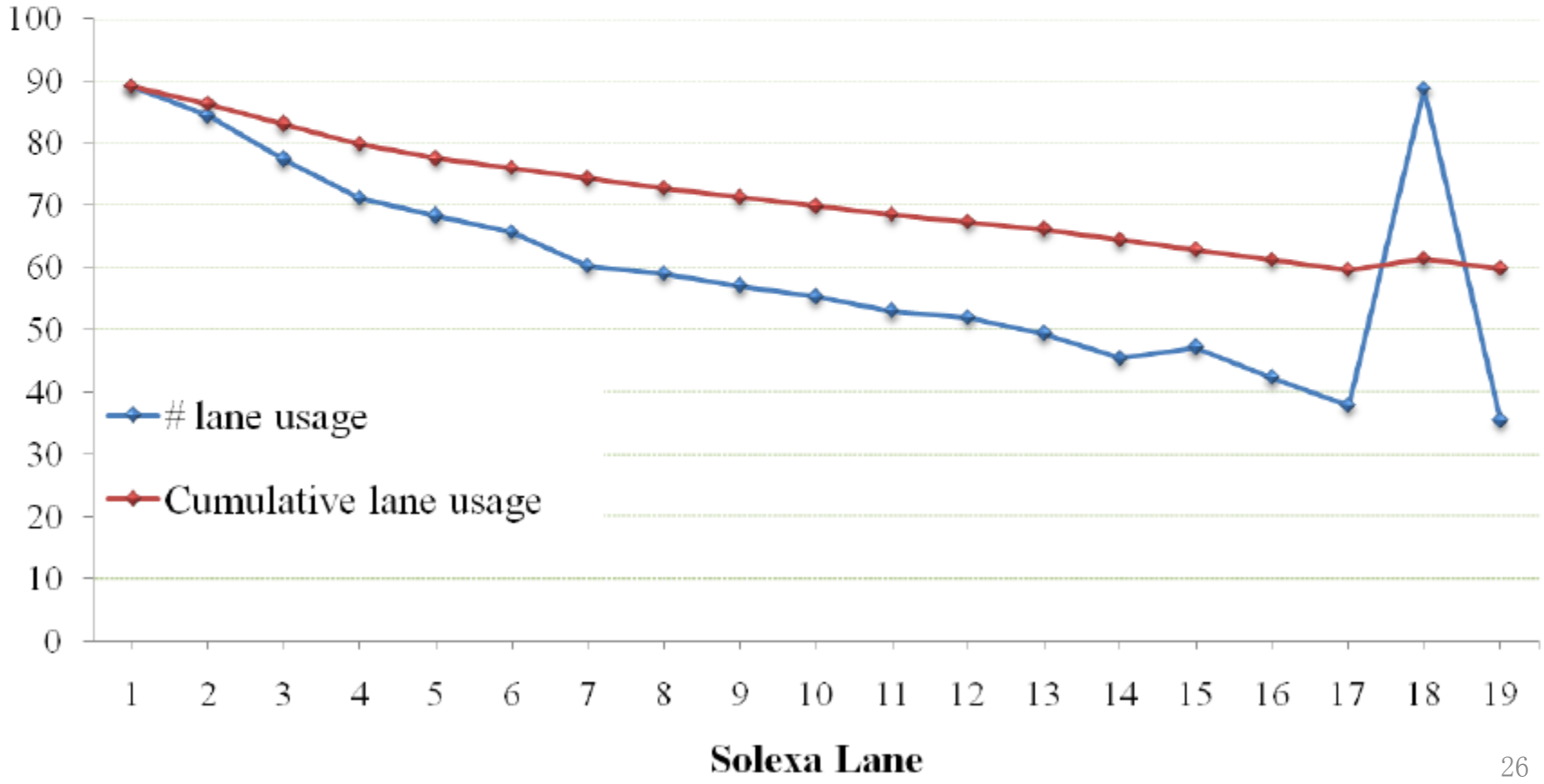


Data assessment III – Mapping Rate

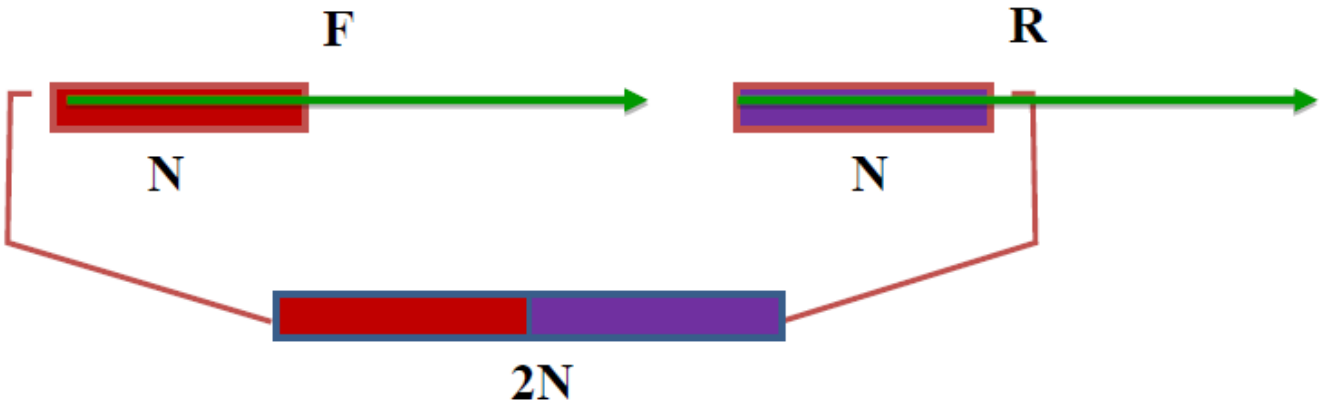


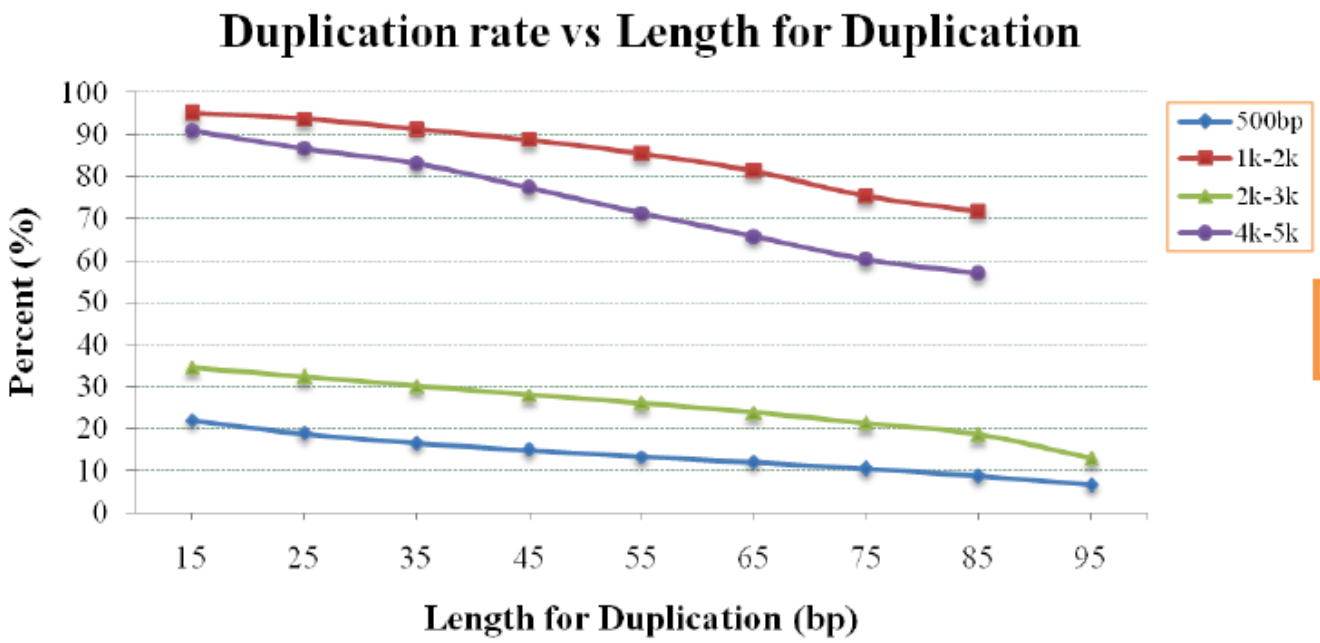
Data assessment IV – Duplication assessment

Solexa Sequencing Data Usage in 500bp Library



Duplicates detection and filter





$Q_{average}$ >20 ?

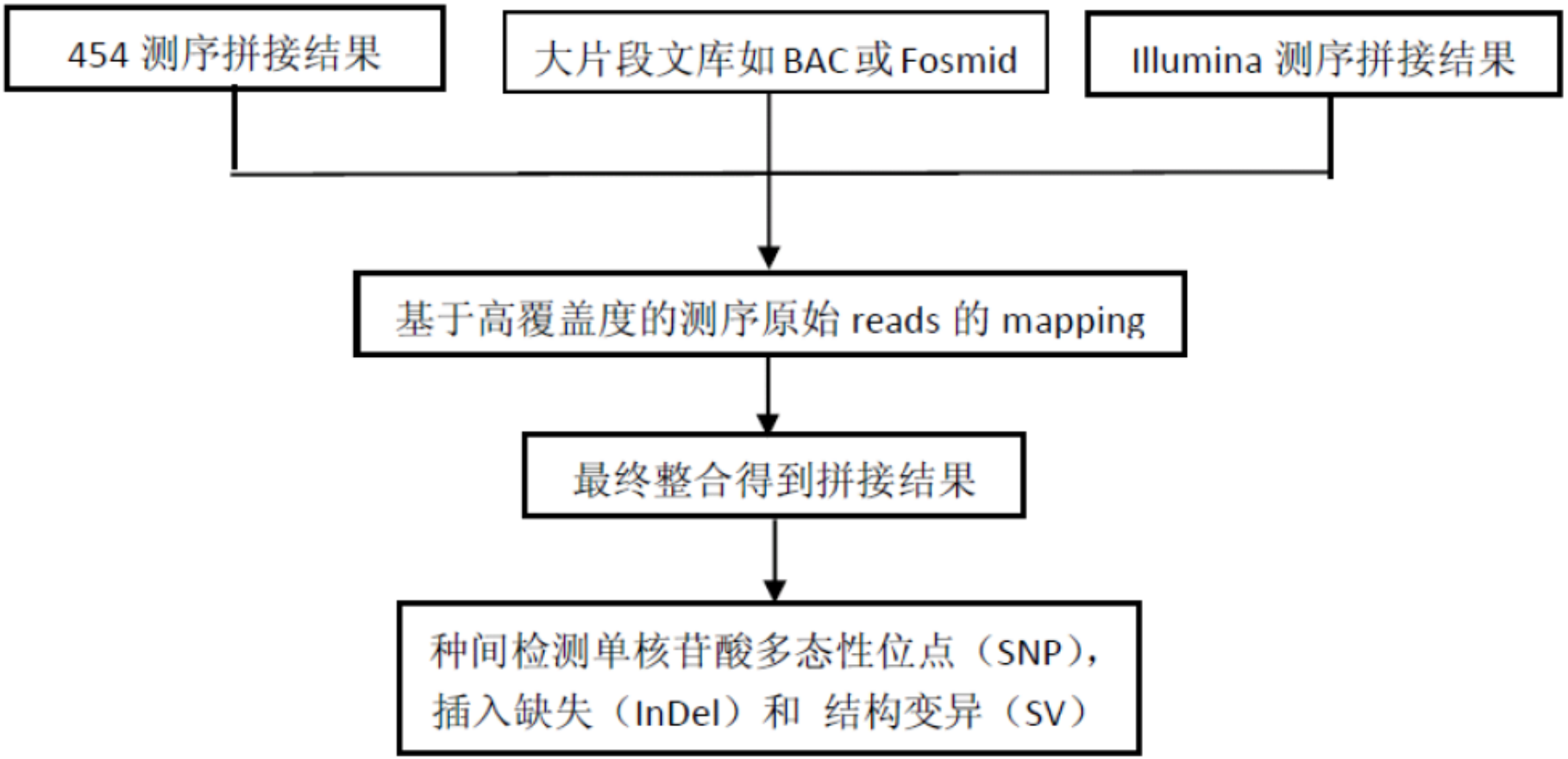
全基因组de novo分析工具

Platform	Correction	Assembly
Solexa	SOAPdenovo	SOAPdenovo Velvet, Abyss
Solid	SAET	Velvet
454	-	newbler

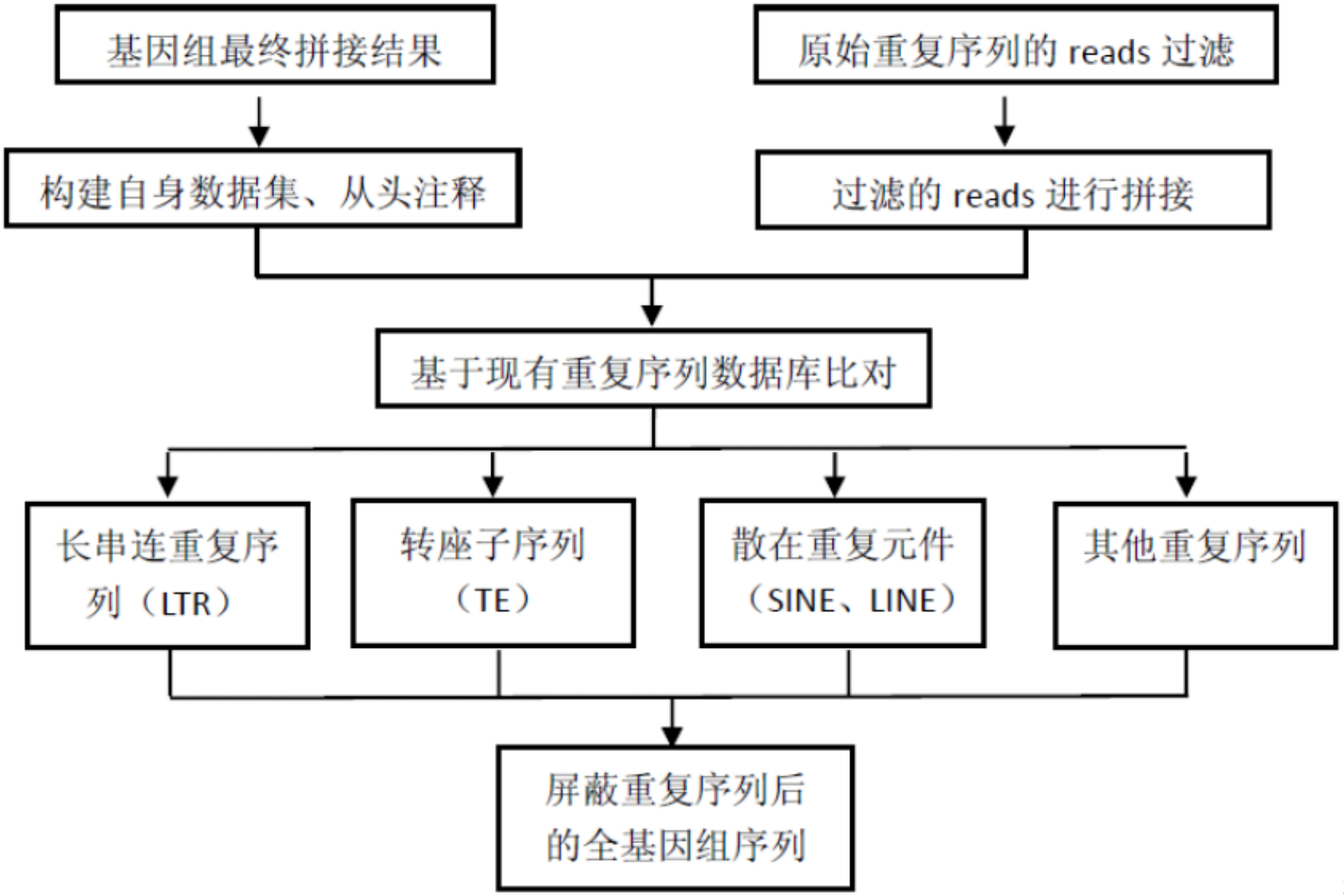
外显子组分析工具

Platform	Alignment	Find Variations
Solexa	bwa, soap	samtools, SOAPsnp
Solid	Bioscope,BFAST	Bioscope,BFAST
454	blast, newbler	newbler

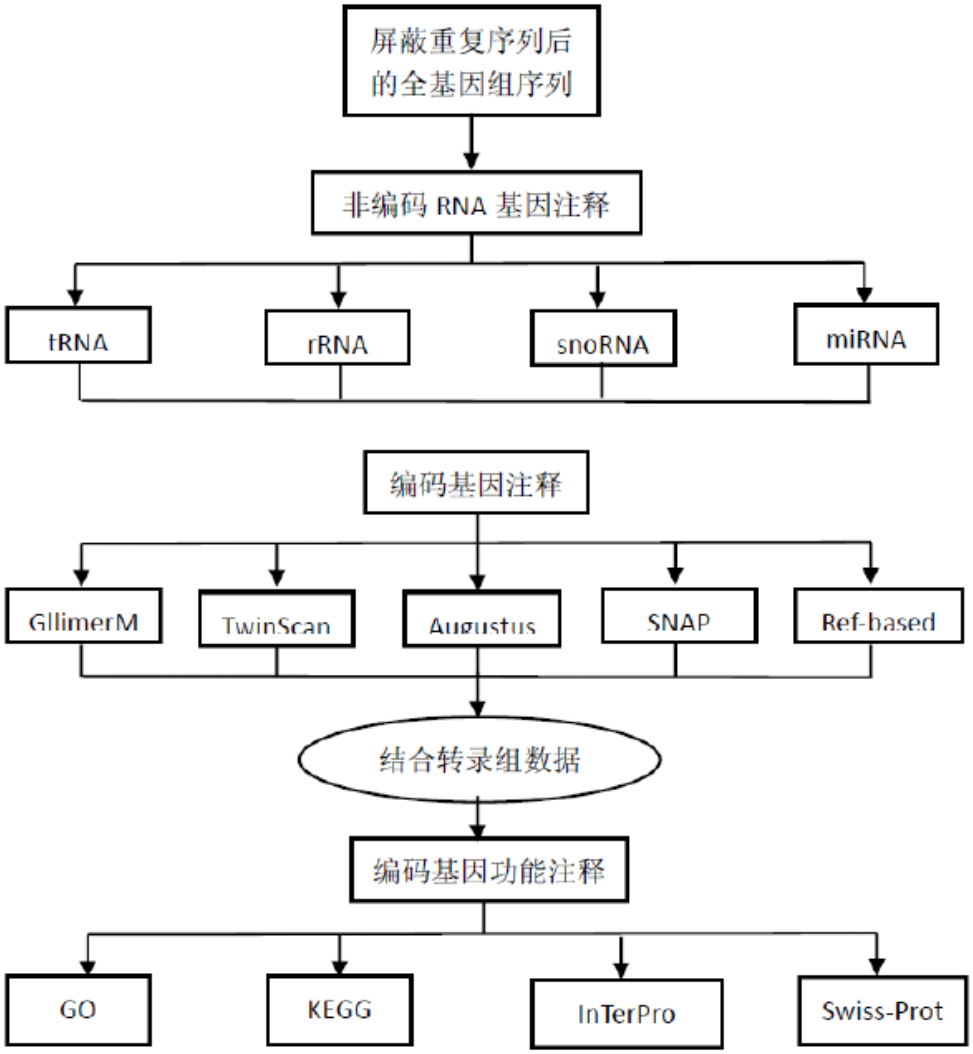
基因组混合拼接验证及结构变异检测流程



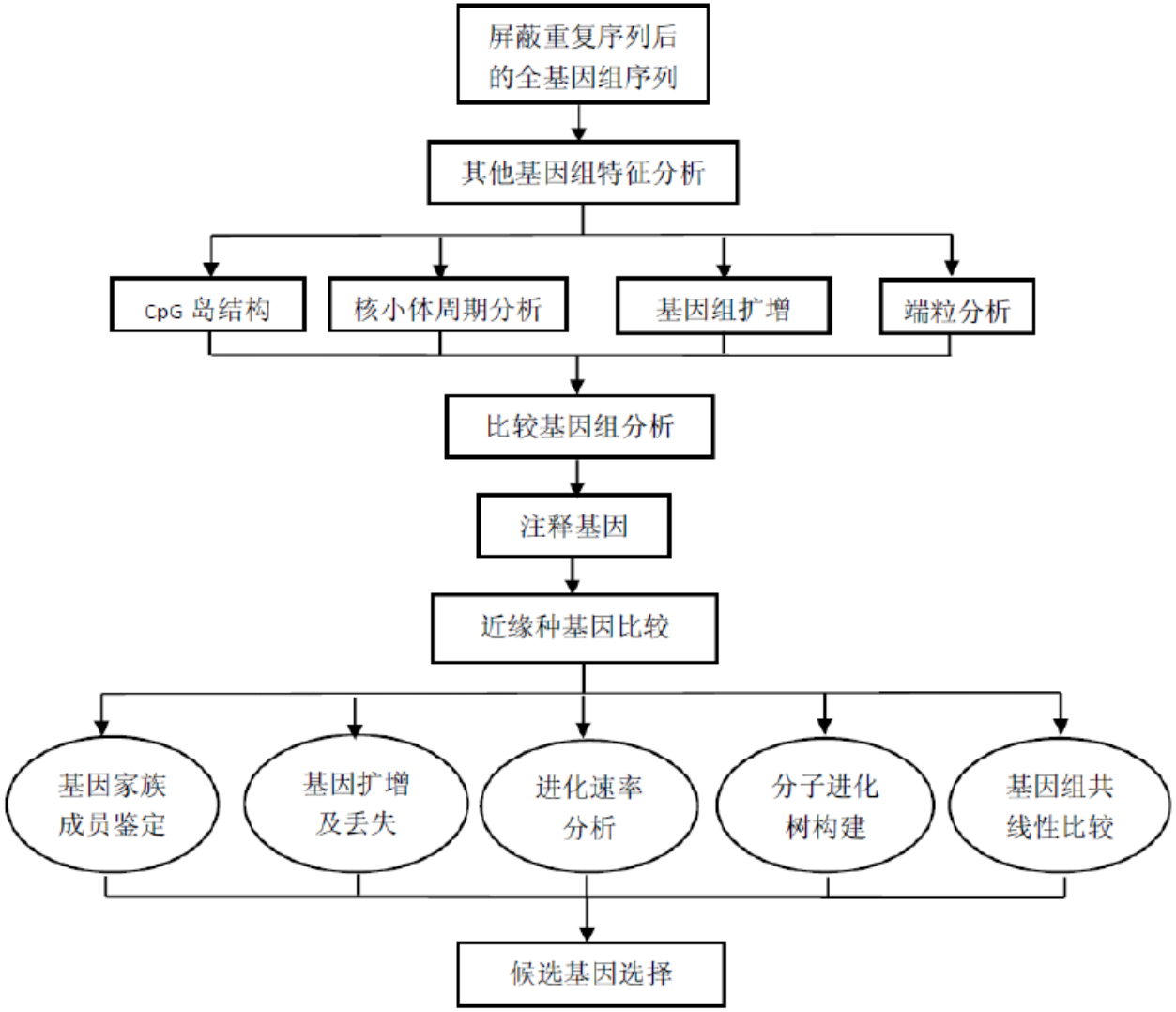
重复序列注释流程


基因组最终拼接结果
原始重复序列的 reads 过滤
构建自身数据集、从头注释
过滤的 reads 进行拼接
基于现有重复序列数据库比对
长串连重复序列（LTR）
转座子序列（TE）
散在重复元件（SINE、LINE）
其他重复序列
屏蔽重复序列后的全基因组序列

基因结构及功能注释技术路线



基因家族进化分析及比较生物学分析技术路线



QY NODE

基因数据解读QQ群
78422213

E-mail: luoqibin@qynode.com