Next Generation Sequencing (NGS)

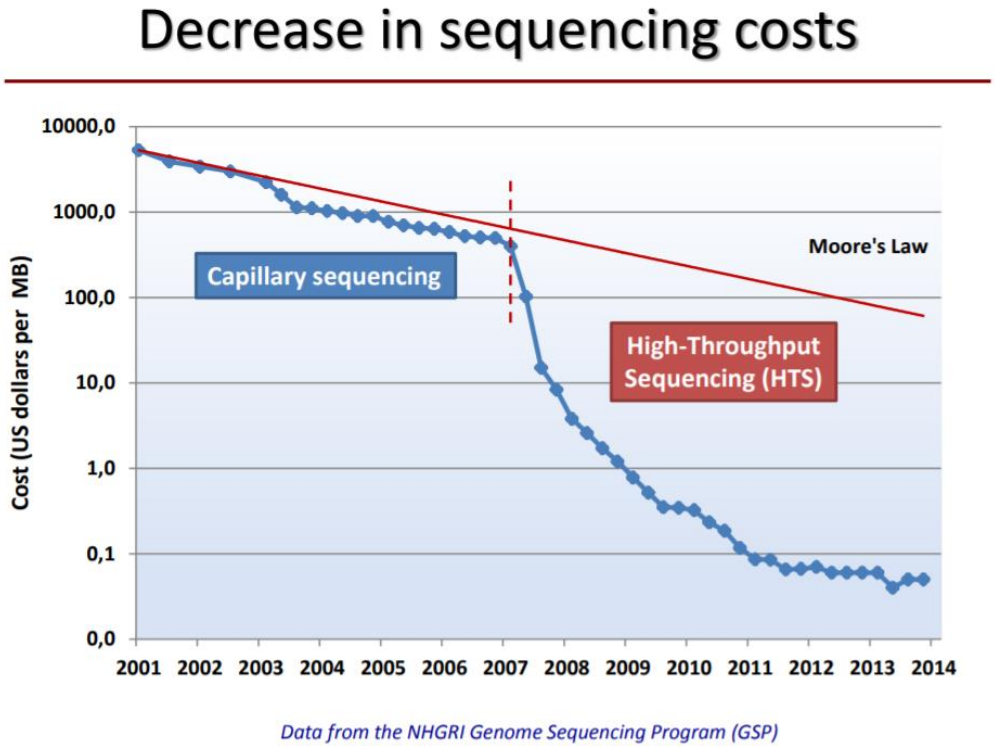
李 余 动
lyd@zjsu.edu.cn

Topics

- **下一代测序原理**
  - Illumina平台
- **NGS测序流程**
  - 建库、测序、数据分析
- **NGS数据质量控制**
  - fastq格式/FastQC
  - Trimmomatic/Fastp

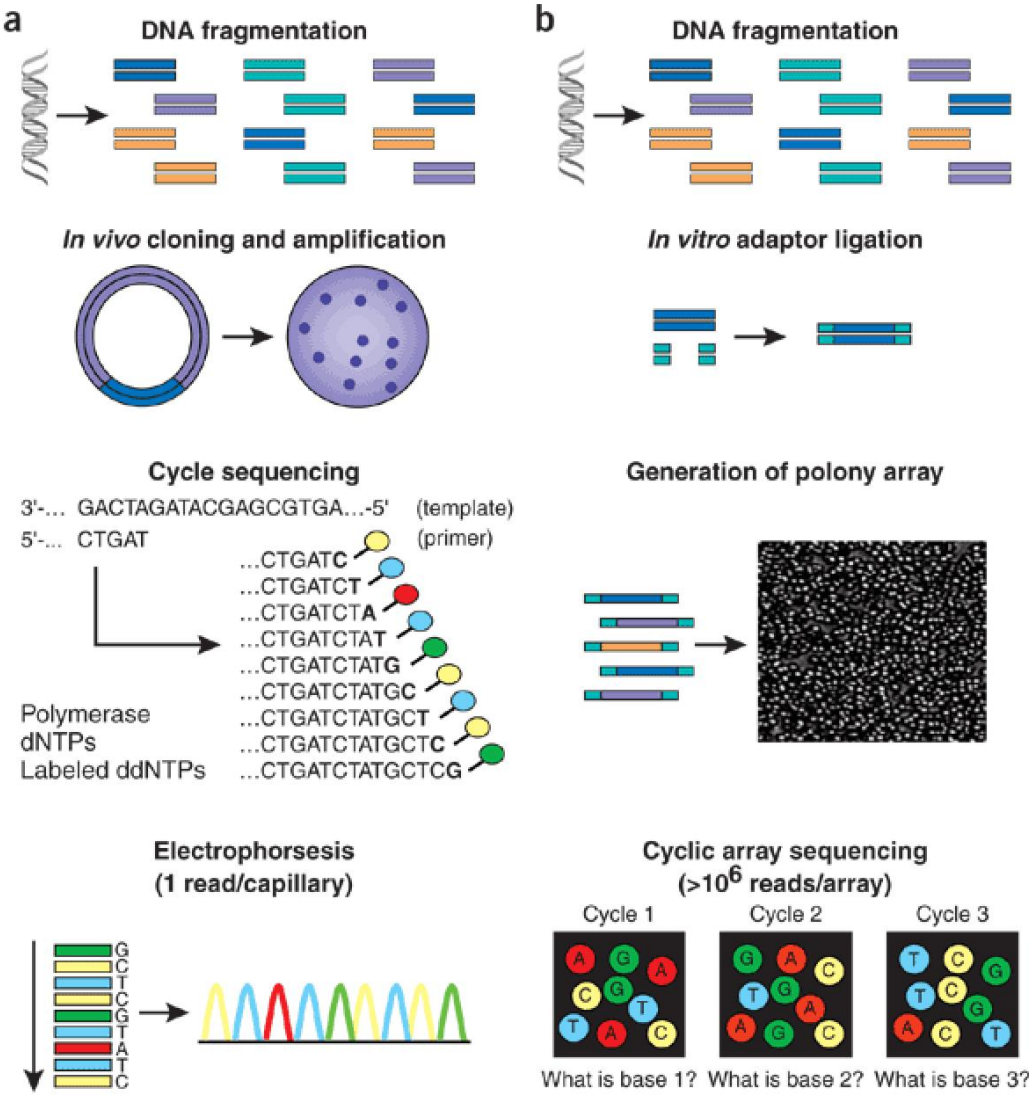
Next Generation Sequencing (NGS)

- 下一代测序技术可以同时对多个DNA片段进行平行测序，又称**高通量测序**(High-Throughput Sequencing, HTS)。



- DNA测序技术快速发展，NGS技术使得基因组测序的通量快速增加，测序成本极大降低，生命科学研究进入组学时代。

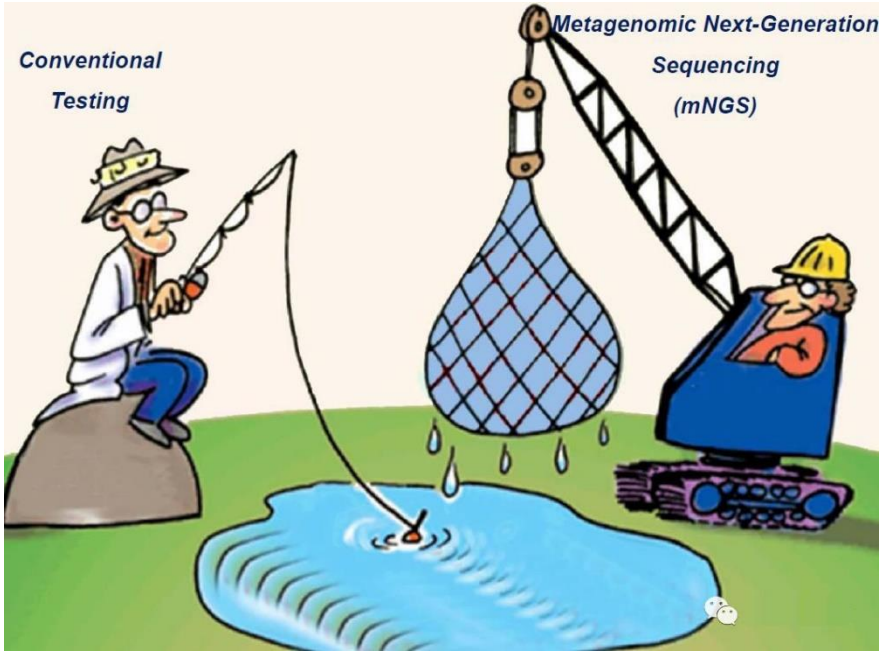
传统Sanger测序与NGS技术原理比较



Sanger sequencing vs. HTS

NGS特点：

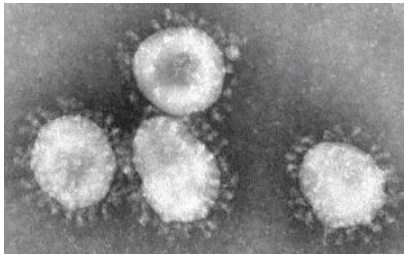
- 测序通量高
- Reads长度短
- 错误率较高

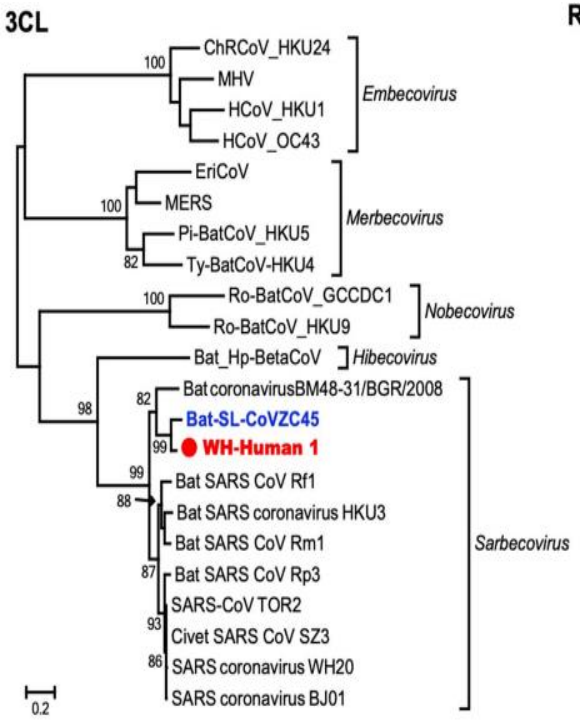


Method	Read length	Advantage	Disadvantage
Sanger sequencing	~ 1000 bp	Long reads. High accurate (>99.9%) Useful for individual gene sequencing applications	Low-throughput Time-consuming Expensive
NGS e.g. HiSeq2000	50 to 300 bp	High-throughput. Useful for genome sequencing applications	Short read length Low accurate (~99%)

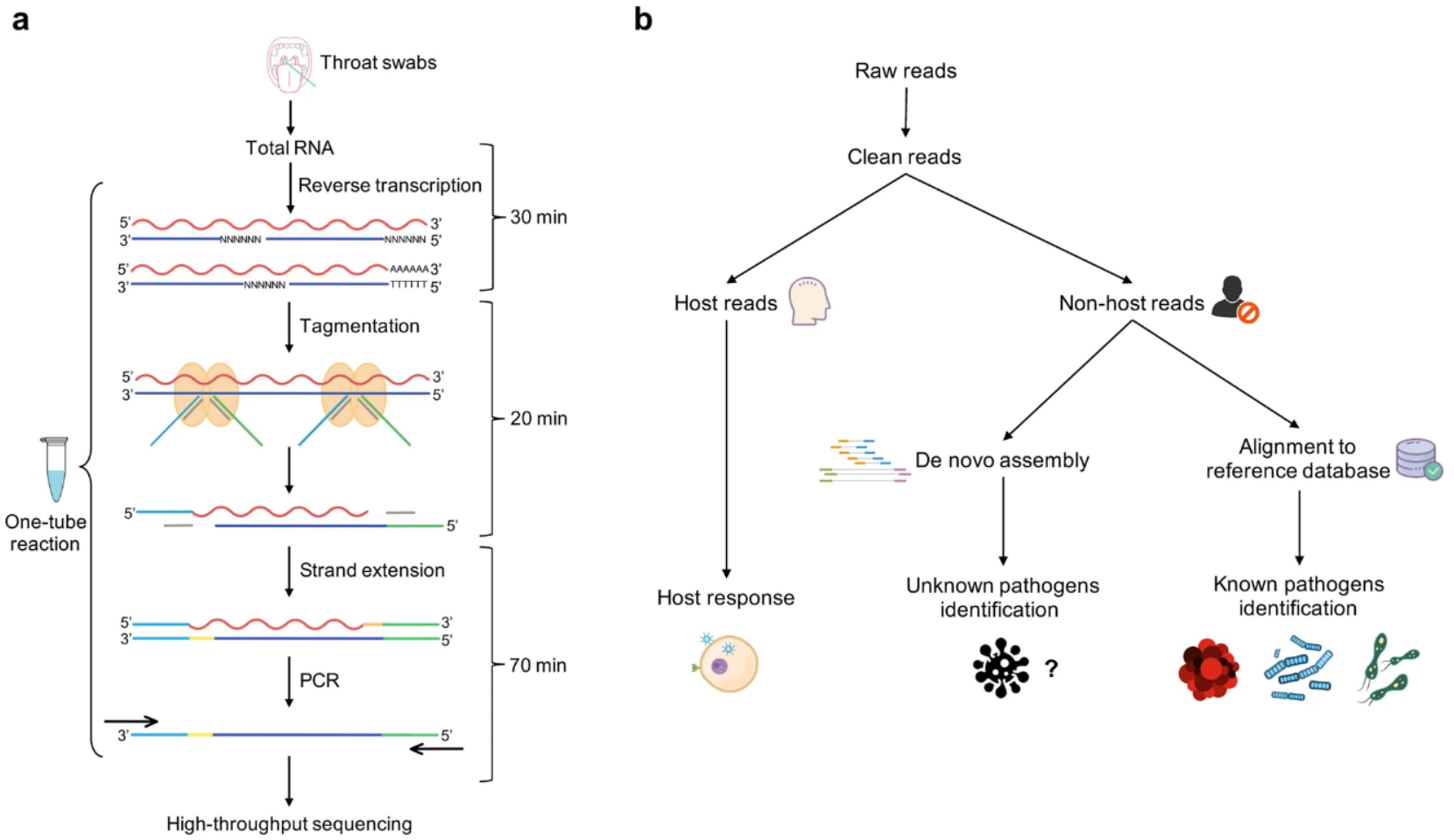
2003冠状病毒与2019新冠病毒

- 2003年中国广东发生SARS疫情，在当时的检测技术条件下，甚至弄明白是什么病原体感染都颇费周折。2002年12月发生疫情，到2003年3月才将病原体聚焦于冠状病毒，2003年4月最终测序成功并确定，经过不懈的努力，科学家鉴定出罪魁祸首为一株冠状病毒，并在广东牲禽市场上所销售的果子狸中发现了基因类似的病毒。
- 2019年底中国武汉暴发的新型冠状病毒肺炎疫情(COVID2019)，中国科学家对患者支气管肺泡灌洗液进行了mNGS测序，鉴定出了一株新型冠状病毒的基因组序列。2020年2月3日，英国Nature杂志在线发表了的研究论文“A new coronavirus associated with human respiratory disease in China”(Wu et al., 2020)。新冠病毒的全基因组序列对新冠疫情的防疫工作具有重大意义。





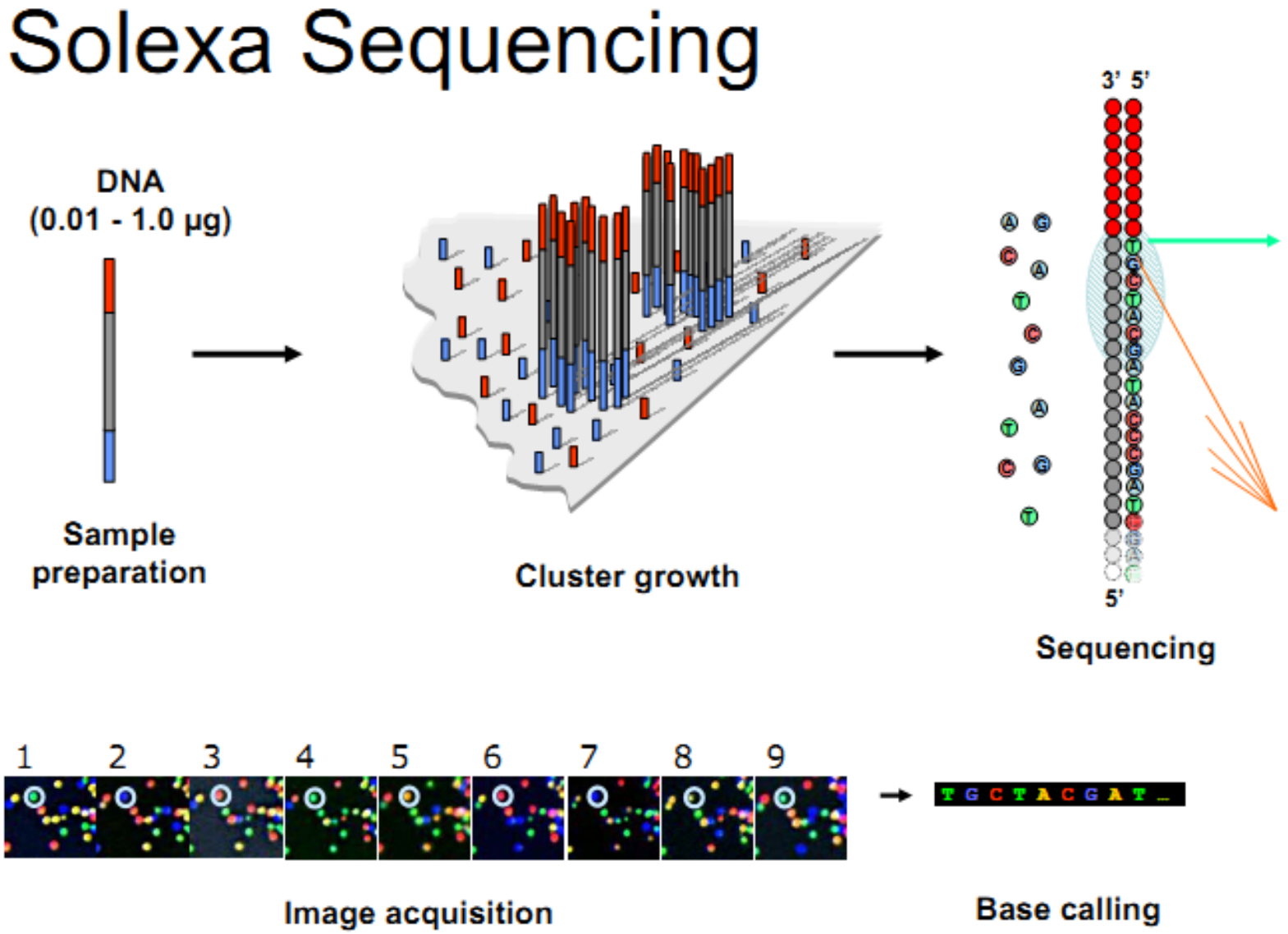
Metagenomic strategy for detecting SARS-CoV-2



下一代测序原理

- 高通量测序技术可以同时对多个DNA片段进行平行测序，主要将打碎后建库的DNA片段锚定在固体介质表面
  - 如Illumina平台通过连接接头的方法将DNA片段锚定在测序通道内表面
  - 通过对每个锚定DNA片段进行桥式PCR
- 检测DNA聚合酶催化荧光标记的dNTP结合到DNA模板时产生的荧光信息。
  - 每加一个碱基进行一次“加上荧光染料 – 洗脱多余染料 – 荧光成像扫描”的循环过程，实现高通量测序。
- 核心：边合成边测序(Sequencing-by-Synthesis, SBS)

Illumina二代测序原理

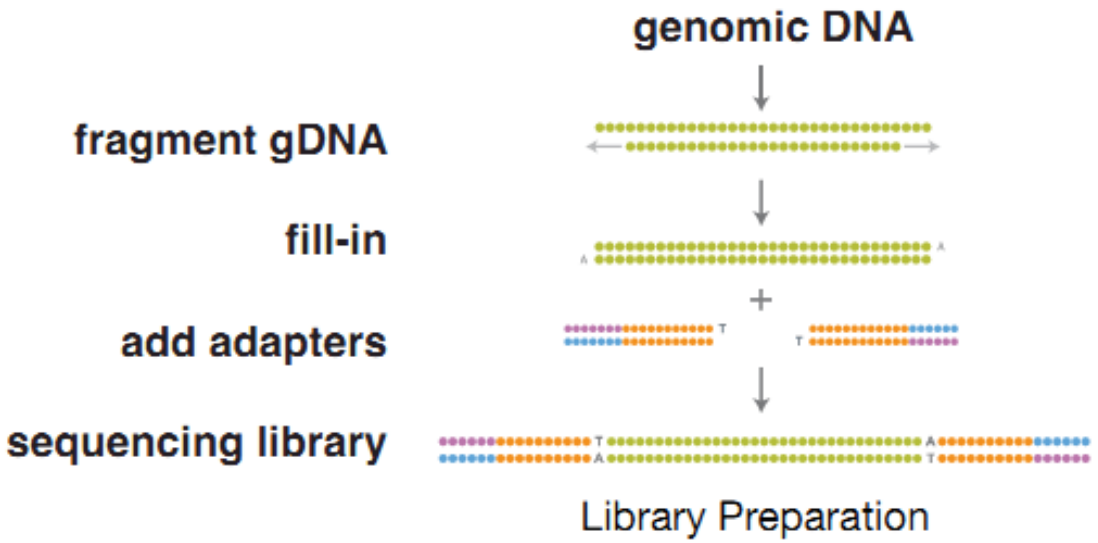


Illumina测序官方视频

https://www.youtube.com/watch?v=fCd6B5HRaZ8

3 steps for genome sequencing

1. Library preparation



文库准备：将待测DNA模板打断成200~800bp的片段，然后在末端补平并加A尾，与特定的测序接头（adapter）连接，纯化后的连接产物构成了测序文库。

双端测序模板(sequencing template)

① 与流动槽（Flow Cell）结合的区域
② Read 1和Read2测序引物结合的区域
③ 插入片段
④ 标签序列区域（Index）

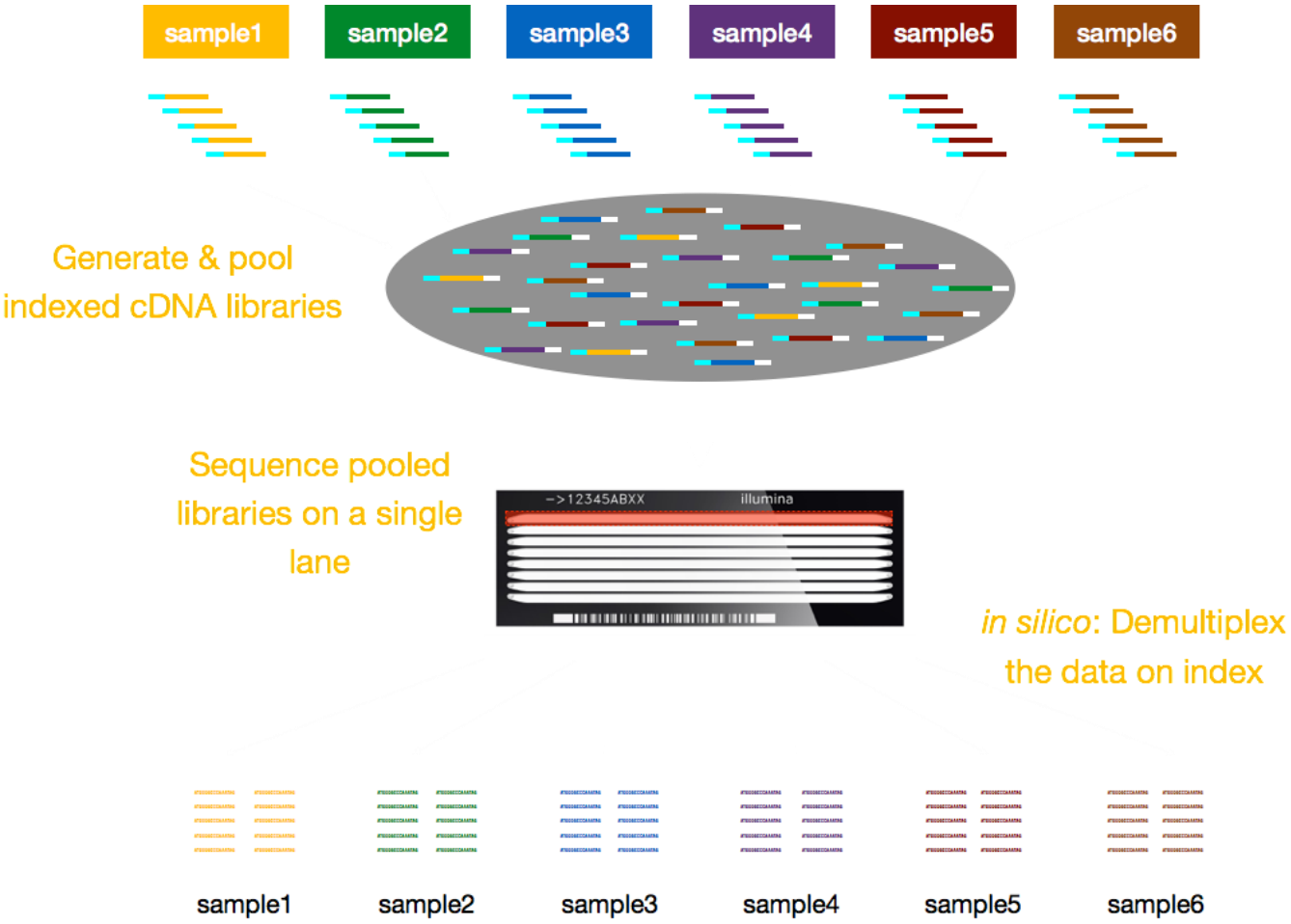
文库制备的目的是在需要测序的DNA片段两端加上能够与测序仪配合的接头序列（Multiplexed, SR, PE）

(1)锚定序列(P5/P7)，用于锚定DNA片段到固体支持物（如玻璃片或磁珠），并在其上进行测序反应；

(2) 通用的测序引物序列(Rd1 SP/Rd2 SP)，用于每个DNA插入片段的测序反应；

(4) 标签序列(Index1/Index2)，用于混合多个样本一起测序时作为区分标签(barcode)。

Multiplexing （混合测序）



3 steps for genome sequencing

- 通过桥式PCR将文库的单片段序列扩增成簇(Cluster)，将信号放大。
- 每个片段在独立的测序单元中进行边合成边测序（SBS）
- 分析每轮测序收集的荧光信号，识别每个模板DNA片段的序列（碱基读取）。

Different sequencing platforms

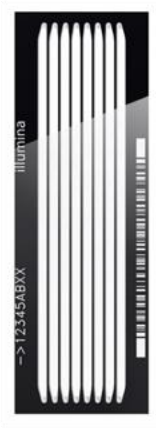
- Differences in platform can alter the length of reads generated, the quality of reads, as well as the total number of reads sequenced per run and the amount of time required to sequence the libraries.

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

Different sequencing platforms

- The different platforms each use a different flow cell, which is a glass surface coated with an arrangement of paired oligos that are complementary to the adapters added to your template molecules.

HiSeq 2500 (2 lane)



HiSeq 3000/4000

NextSeq 500

- Depending on the Illumina platform (MiSeq, HiSeq, NextSeq), the number of lanes per flow cell, and the number of reads that can be obtained per lane varies widely. **You will need to decide on how many reads you would like per sample** (i.e. the sequencing depth) and then based on the platform you choose calculate how many total lanes you will require for your set of samples.

3 steps for genome sequencing

3. Analysis

Align reads to reference genome

▼

Call variants

▼

Filter variants

▼

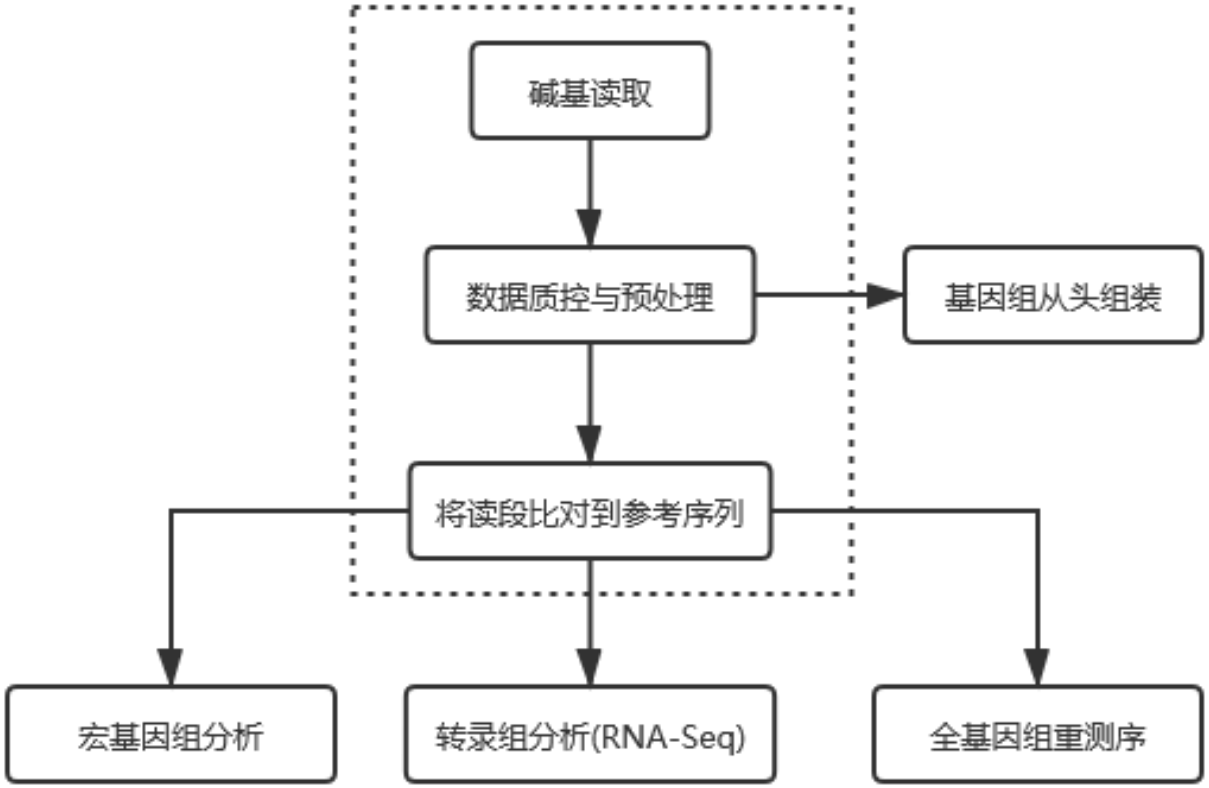
View

Software

Step	Software	Link
QC/preprocessing	FastQC	http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/
	FASTX-Toolkit	http://hannonlab.cshl.edu/fastx_toolkit/
Aligning	Novoalign	http://www.novocraft.com
	BWA	http://bio-bwa.sourceforge.net/
Variant calling	Samtools	http://samtools.sourceforge.net/
	VCFtools	http://vcftools.sourceforge.net/
Variant annotation	SeattleSeq Annotation	http://gvs.gs.washington.edu/SeattleSeqAnnotation/
Data viewing	IGV	http://www.broadinstitute.org/software/igv/
	UCSC Browser	http://genome.ucsc.edu/
Misc	tabix	http://samtools.sourceforge.net/tabix.shtml
	Perl	http://www.perl.org/
	R	http://www.r-project.org/

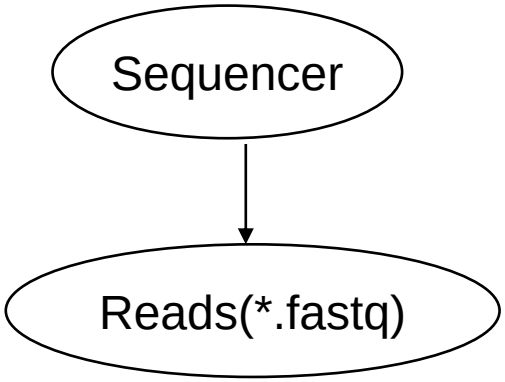
- 下游数据分析基于测序机器的碱基读取序列
  - Illumina测序仪产生的BCL文件通过BCL2FASTQ工具转换为FASTQ文件。
  - PacBio公司的SMRT Link及ONT公司的Guppy等

NGS数据分析的共同步骤



测序数据：Read(读段)

- Read: A short DNA fragment which is read out by sequencer.
  - DNA sequence (碱基)
  - Quality information (测序质量)
- 高通量测序的序列数据一般存储在FASTQ格式文件，文件后缀一般为.fastq, .fq等
  - FastQ is a FASTA file with quality information



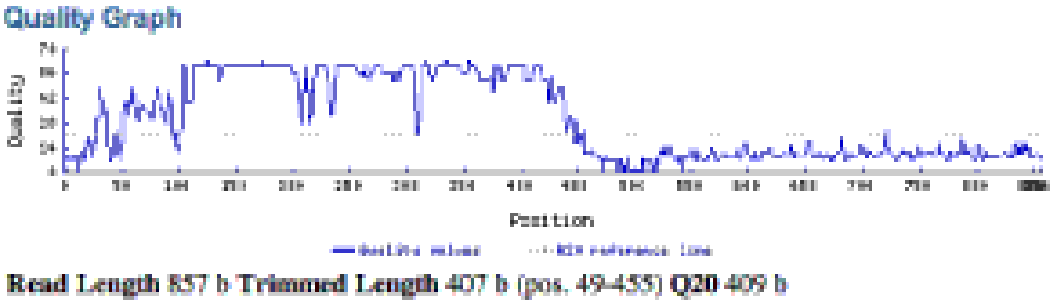
FASTQ文件

- FASTQ格式以每个测序读长(read)为4行，分别为
  - @序列ID, 包含index序列及read1或read2标志；
  - 碱基序列，大写”ACGTN”；
  - +序列ID（可选）；
  - 序列每个碱基的质量分数（字符的ASCII编码表示）。

Index

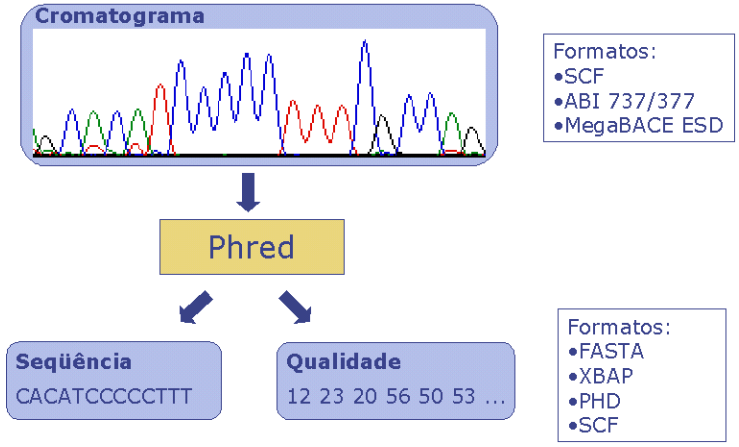
Read1

```
@HISEQ2000:404:C73LWACXX:2:1101:1487:1876 1:N:0:CGATGT
NCCCTCTTGAACTCTCTCTTCAAAGTTCTTTTCAACTTTCCCTTACGGTACTTGTTGACTATCGGTCTCGTGCAGATCGGA
+
#4=DB?:DF?ADCFFGD>BHCEB9F3AAACEFHC>@BBFFFGD@??BF??D9B?FGDFFGDGGB@;AE>ED25;)..;;=;
```



Phred Quality Scores

- 质量值(quality values)表示该碱基的错误率,也称Phred值(Q-score)。
  - Phred is a program that assigns a quality score to each base in a sequence.

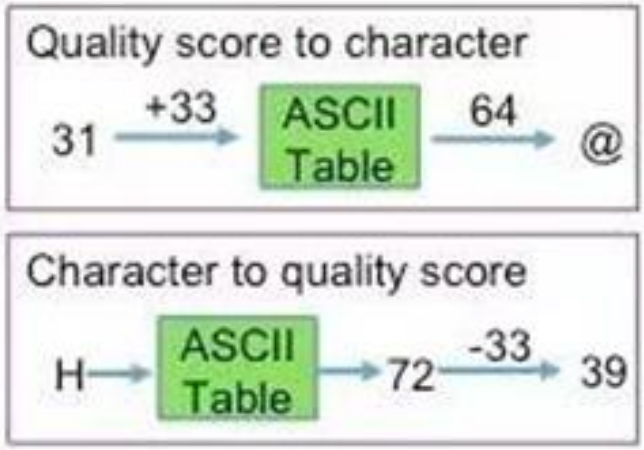


- Phred scores (Q) are logarithmically related to the probability of an error (p):
  - $Q = -10 \log_{10} P$, where Q is the phred score and P is the probability that the base was called incorrectly.
  - 质量值为10时，说明该碱基的错误率为10%；20为1%；30为0.1%等。
  - A score of 20 is generally considered the minimum acceptable score.

p	Q
0.1	10
0.01	20
0.001	30
0.0001	40

Phred质量值编码方式

- To encode the quality scores as a single character, the scores are mapped to the ASCII table
- Sanger测序与Illumina新平台(CASAVA 1.8+)采用ASCII码-33的编码方式(Phred+33)，质量值区间为0-40。

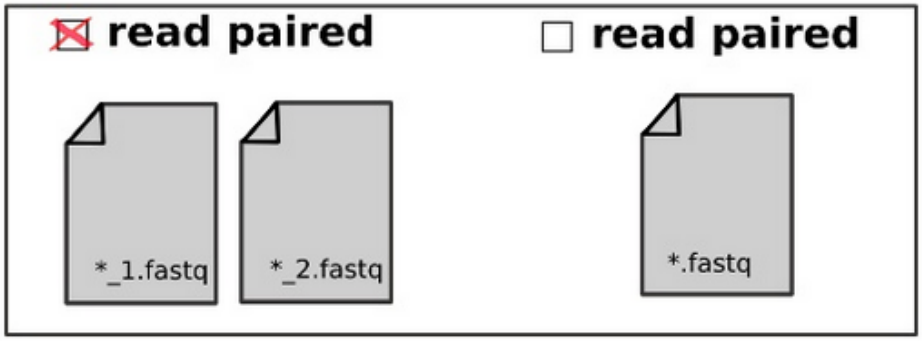


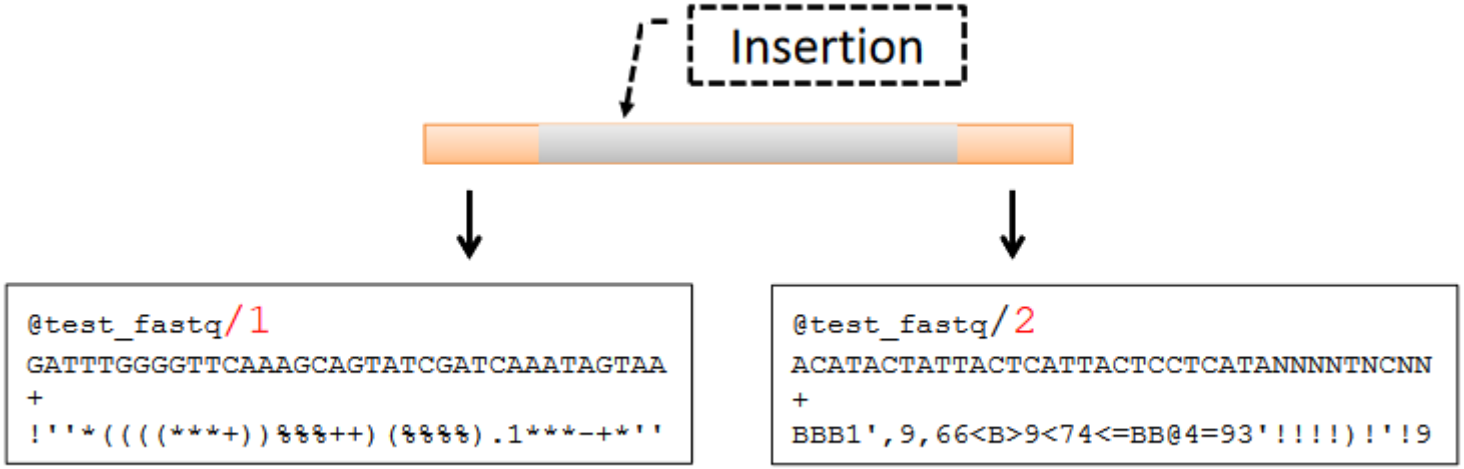
质量值(Q)转换成对应的ASCII码字符

ASCII Code	Symbol	Q-value	ASCII Code	Symbol	Q-value
33	!	0	54	6	21
34	"	1	55	7	22
35	#	2	56	8	23
36	$	3	57	9	24
37	%	4	58	:	25
38	&	5	59	;	26
39	'	6	60	<	27
40	(	7	61	=	28
41	)	8	62	>	29
42	*	9	63	?	30
43	+	10	64	@	31
44	,	11	65	A	32
45	-	12	66	B	33
46	.	13	67	C	34
47	/	14	68	D	35
48	0	15	69	E	36
49	1	16	70	F	37
50	2	17	71	G	38
51	3	18	72	H	39
52	4	19	73	I	40
53	5	20			

测序数据

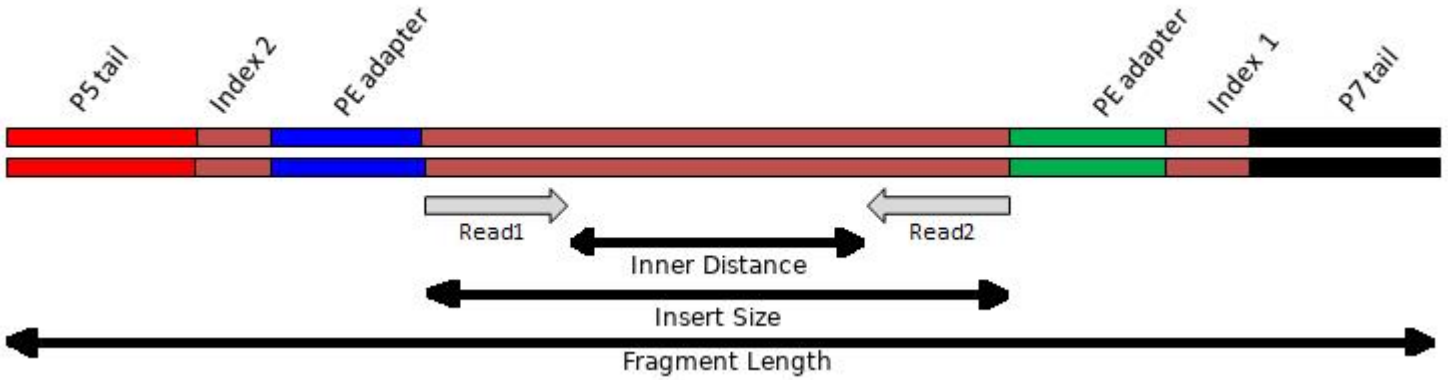
- Illumina高通量测序分双端测序与单端测序
  - Paired-end (PE)
  - Single-end (SE)



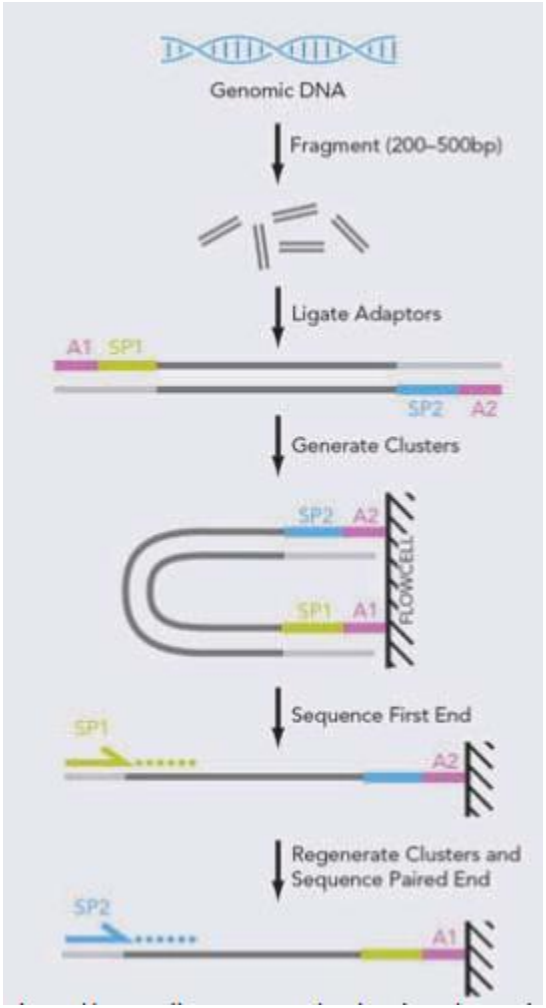


Paired-End Reads vs. Single-End Reads

- More bases from the insert compared to a single-end read - up 2x as many!
- The expected distance between the reads of a pair provide additional constraints around where the read pair can/should align to a genome - the reads from a pair must align within ~300-400 bp of each other.
- The reads must align to the genome with the correct relative orientation. For paired-end reads this is often refered to as: forward-reverse (fr), innies or simply →←



Sequencing Types



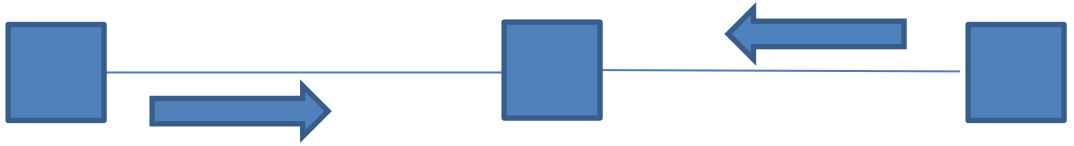
单端测序

Single Read

双端测序

Paired-end read

环化配对测序



Mate-pair read

https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments.html

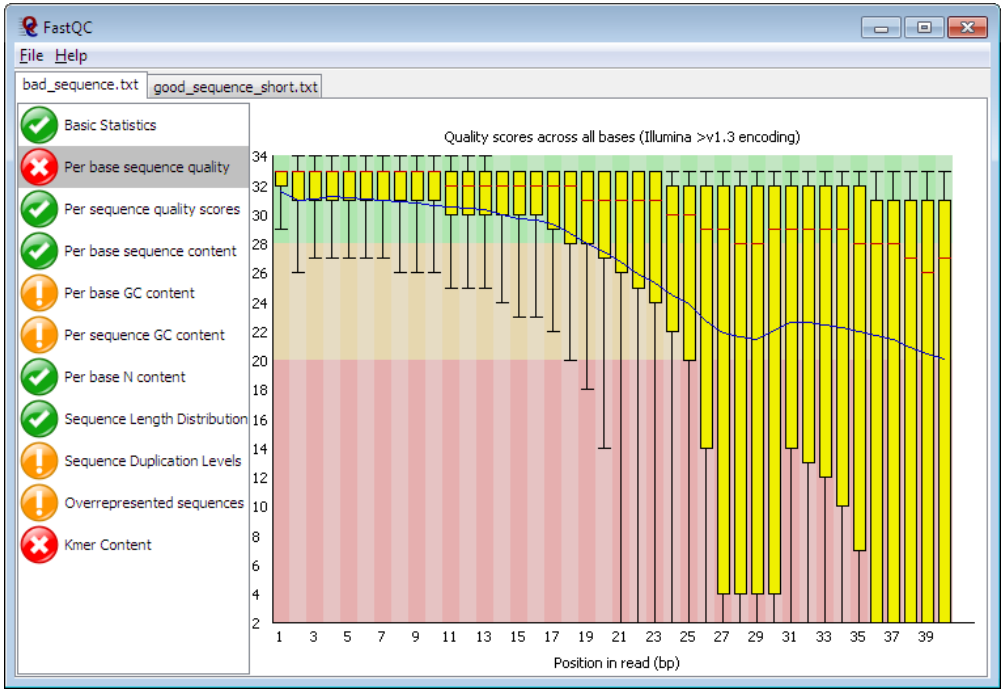
原始测序数据预处理

- 测序数据预处理主要：
  - 剪切(Trimming)去除测序接头序列，Adapter-containing reads will fail to align as the adapter does not match anything in the genome.
  - 过滤(Filtering)去除测序质量比较差的核苷酸序列, Low quality stretches of reads might contain too many sequencing errors and stop the read aligning to the correct location in the genome.

- NGS质控工具：

数据质量评估 - FastQC

- FastQC是目前最常用的NGS数据质量评估软件，可用于统计质量分数、GC含量、测序长度等信息。
  http://www.bioinformatics.babraham.ac.uk/projects/fastqc/



FastQC: Basic Statistics

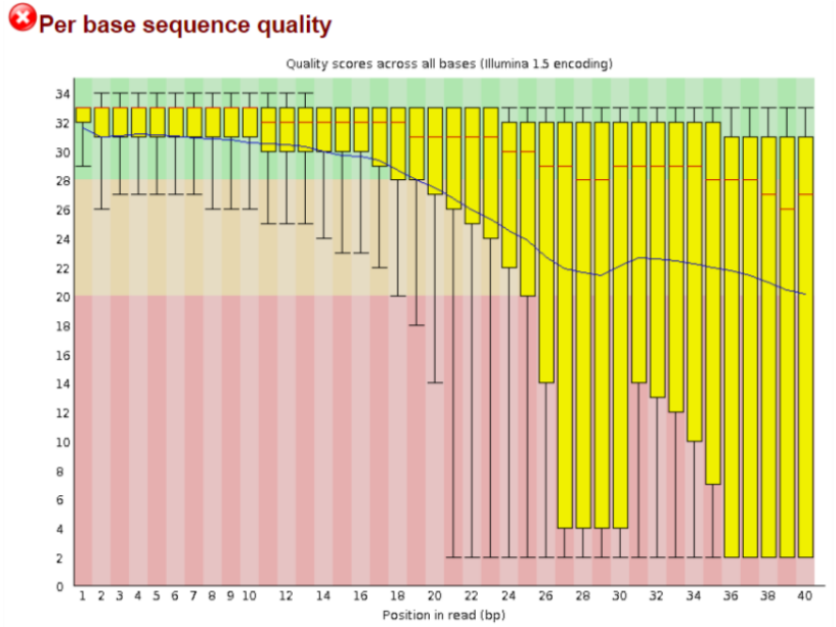
- Basic Statistics给出原始数据的基本信息，包括被分析文件的文件名、文件类型、质量值编码方式(Encoding)、序列总数、标记为低质量的序列数、序列长度和 GC含量。
- Encoding: Phred33 or Phred64 (<1.8)

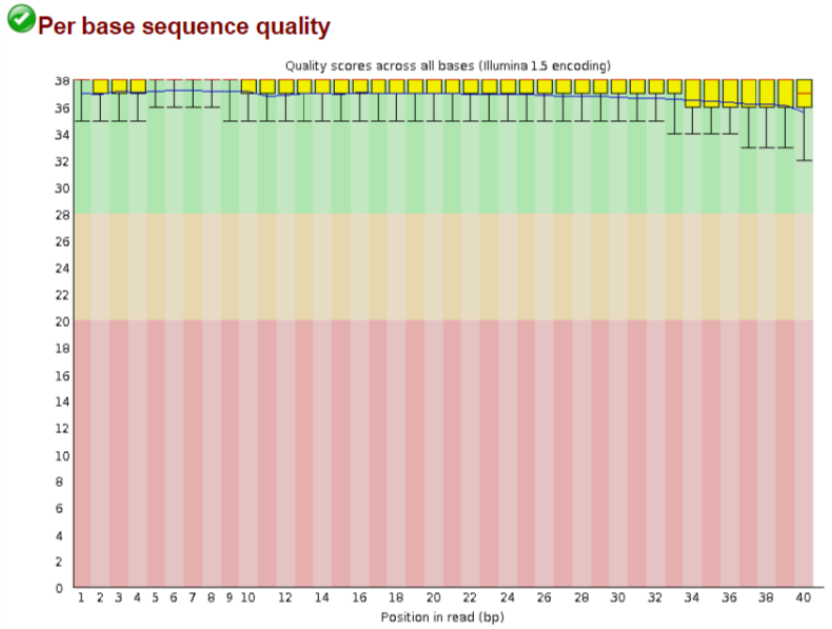
Basic Statistics

Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

FastQC: Per base sequence quality

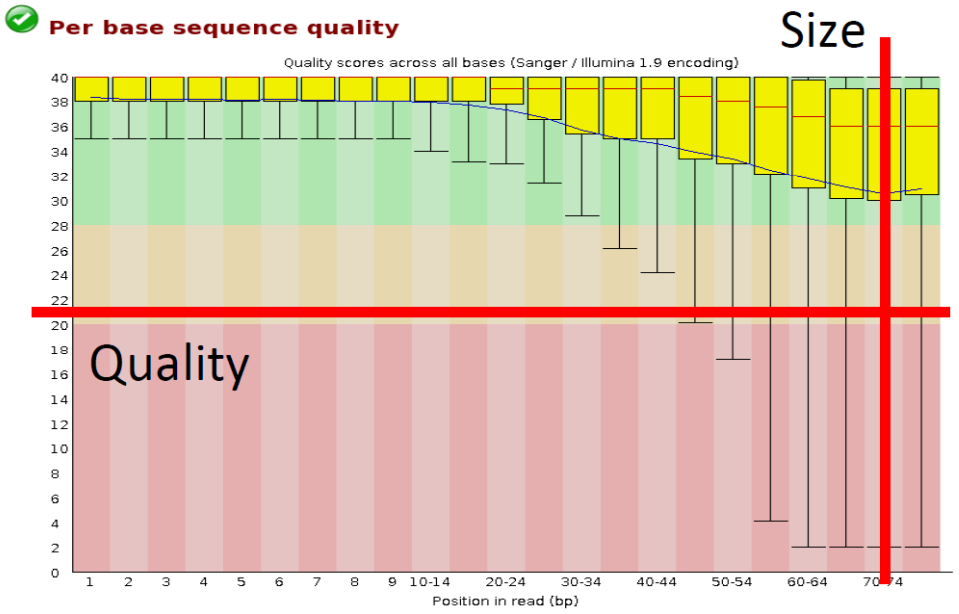
- Overview of the range of quality values across all bases at each position in the FastQ file
- 碱基的质量值越高越好，背景颜色将图分成三部分：碱基质量很好 (绿色)、碱基质量一般(黄色) 以及碱基质量差 (红色)。





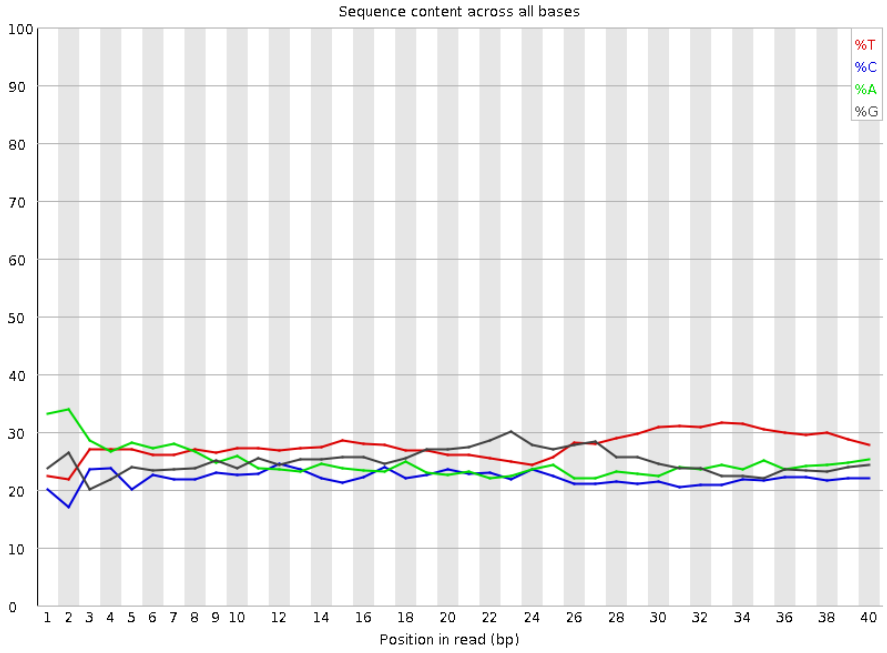
Sequence filtering

- Filtering parameters:
  - Quality filtering
    - Overall mean quality
    - Local mean quality
  - Size filtering
    - Overall sequence size
    - Remaining sequence size after filtering



Per Base Sequence Content (平均碱基组成)

- Per Base Sequence Content显示每个位置上的碱基组成比例
- 一个完全随机的测序文库内每个位置上4种碱基的比例应该大致相同，因此图中的四条线应该相互平行且接近；
- 在序列结尾出现的碱基组成偏离，往往是测序接头的污染造成的。



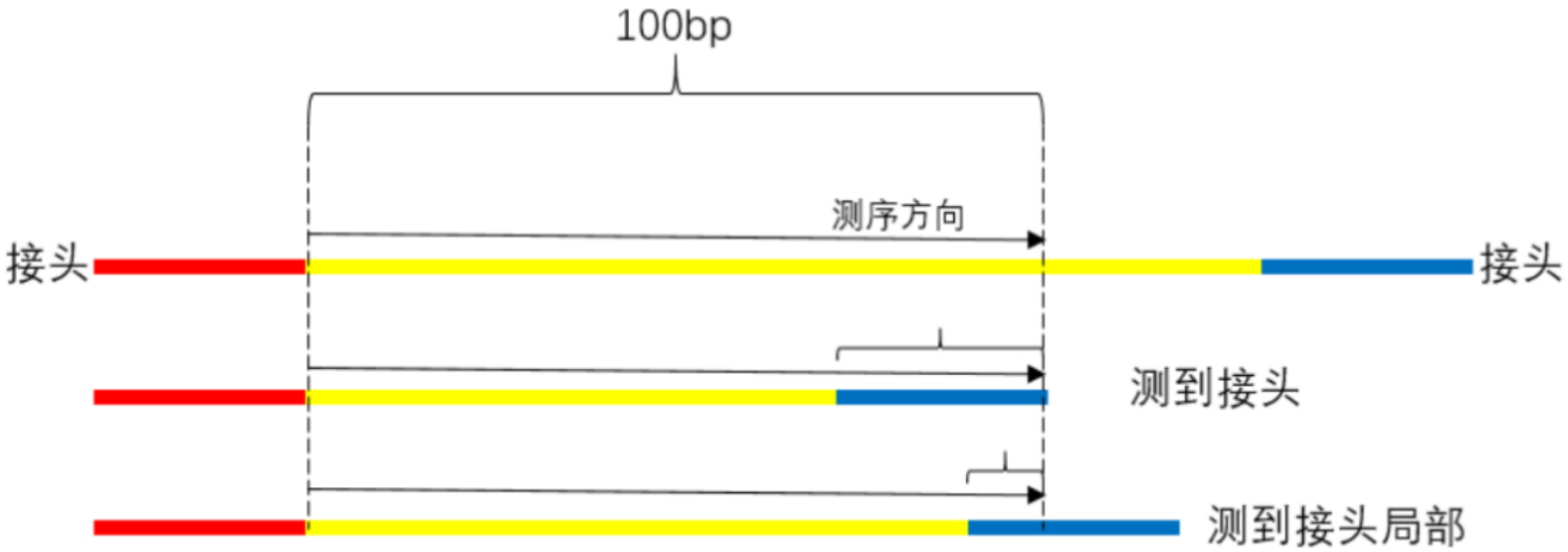
Overrepresented sequences (过度呈现的序列)

- Overrepresented sequences显示同一条read出现次数超过总测序读段数的0.1%的统计情况。
- 对于全基因组鸟枪法测序文库处理，在读段中会含有一小部分的接头序列。如果3'末端的接头含量明显增加，这可能表明基因组DNA在文库处理过程中被过度破碎。

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGA	6276	6.276	TruSeq Adapter, Index 1 (100% over 40bp)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	6274	6.274	Illumina Single End PCR Primer 1 (100% over 40bp)
CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAG	6252	6.252000000000001	Illumina PCR Primer Index 1 (100% over 40bp)
CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAG	6192	6.192	Illumina PCR Primer Index 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	6142	6.142	Illumina Paired End PCR Primer 2 (100% over 40bp)

Trimming adapters

- For whole genome shotgut sequencing library preps we expect to see little adapter content in the reads.
- If there is a significant up-turn in adapter content towards the 3’ end, this may indicate the genomic DNA was over-fragmented during libarary prep.



数据预处理-Trimmomatic

- Trimmomatic可用于过滤与切除低质量序列、接头序列，不需要的污染物序列等
- 网址：http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.38.zip

- 执行命令：

$ java -jar Trimmomatic-0.38/trimmomatic-0.38.jar SE bad_seq.fq bad_output.fq ILLUMINACLIP:testAdapter.txt:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:30

参数说明参见教材11.4.2数据预处理(P138)

数据预处理-fastp

- Fastp是一款可以对NGS数据进行质量控制与序列剪切的工具，实现FastQC+Trimmomatic两款软件的基本功能，并且运行速度非常快。
- 网址： http://opengene.org/fastp
- 执行命令：

```
$ fastp --thread 2 --cut_right  \
-i ./data/ERR1949188_1.fastq.gz \
-I data/ERR1949188_2.fastq.gz \
-o ./qc_reads/fastp/ERR1949188_1.fastq.gz \
-O ./qc_reads/fastp/ERR1949188_2.fastq.gz \
--cut_window_size 4 --cut_mean_quality 20  --length_required 75
```

Trim or not trim?

- Signal/noise -> Preprocessing can remove low-quality "noise", and adapters but the cost is information loss.
  - Some uniformly low-quality reads map uniquely to the genome.
  - Trimming reads to remove lower quality ends can adversely affect alignment, especially if aligning to the genome and the read spans a splice site.
  - **Most aligners can take quality scores into consideration.**
  - Currently, we do not recommend preprocessing reads aside from removing uniformly low quality samples.

- Debate: http://www.ecseq.com/support/ngs/trimming-adapter-sequences-is-it-necessary

作业

- 下载本章所用的测序数据(good/bad_seq.fastq.gz)，练习用FastQC软件分析测序reads的碱基质量及其它质量指标，并用Trimmomatic或Fastp进行质量控制处理。