Chapter 2. Genome sequencing and assemble technology

Lecture 2.2 Sequencing methods

Huaqin He

PPT slides and Message @ http://jxpt.fafu.edu.cn/meol/homepage/common/

Email: 1156743645@qq.com

OUTLINES

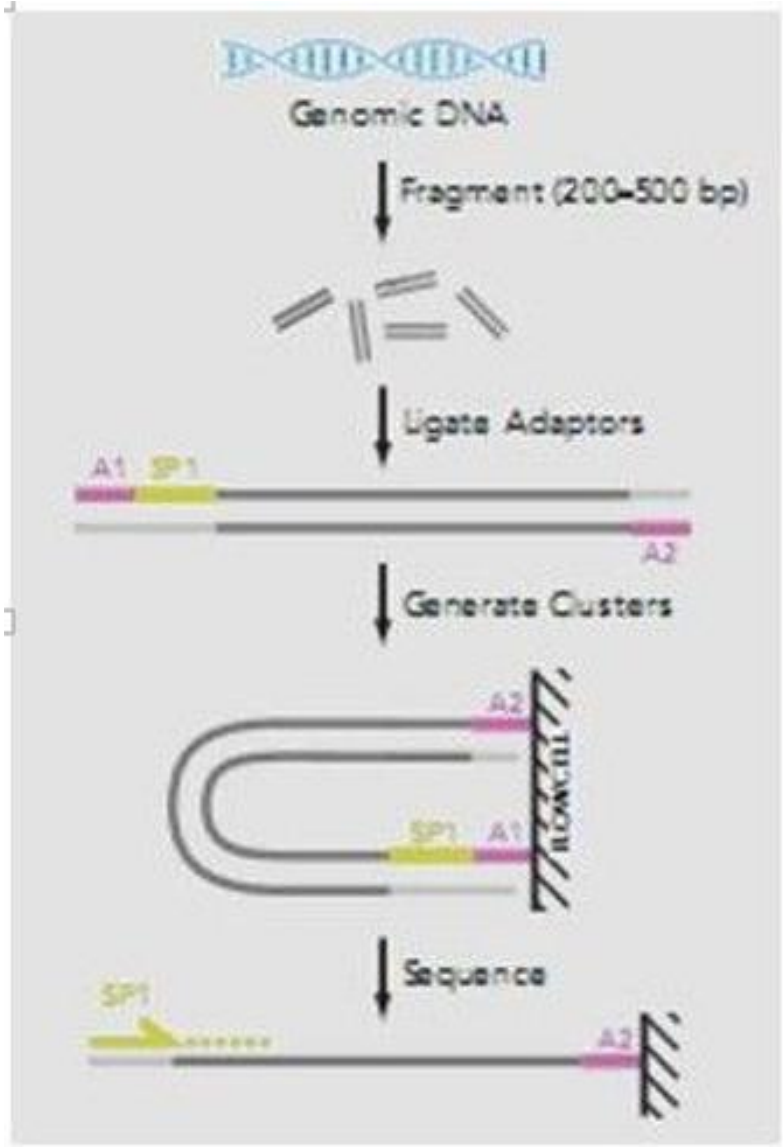
1. Genome sequencing methods 测序方式
2. Read file format 读段记录格式
3. Read quality control and QC 测序质量控制

1. Genome sequencing methods 测序方式

1.1 Single read 单端测序

1.2 Paired-end read 双端测序

1.3 Mate-pair read 配对测序



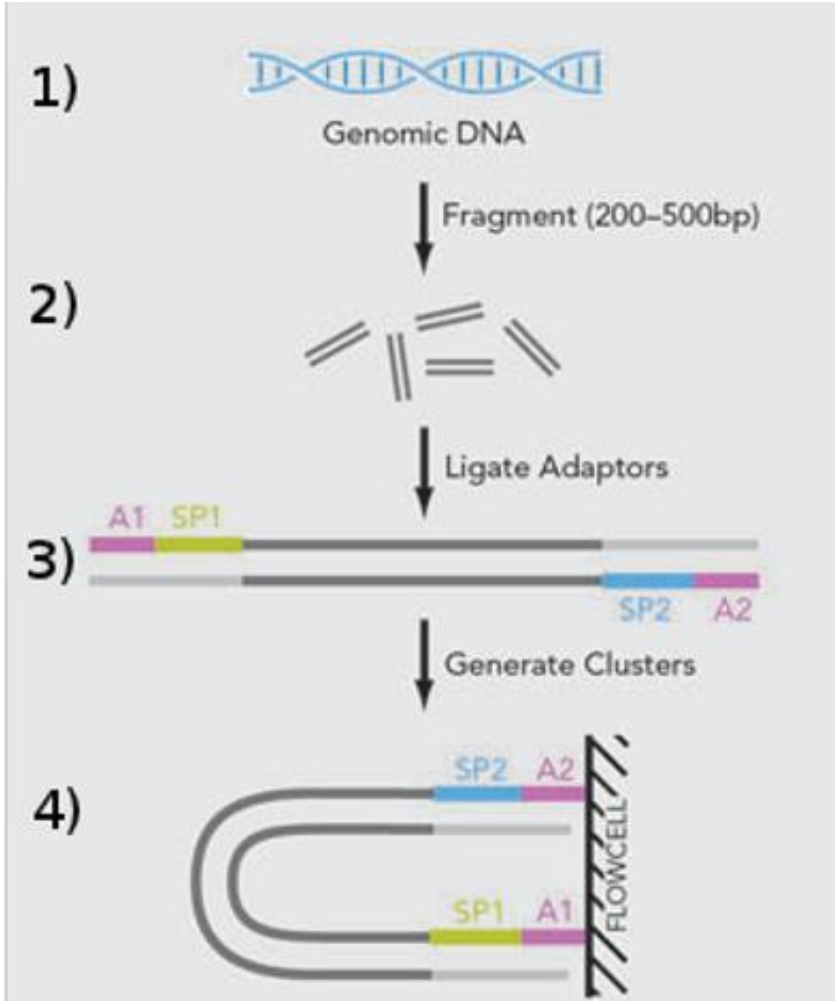
1.1 Single-read library:

❶DNA sample is digested into 200-500bp fragments.

❷Primers and adaptors are ligated to the ends of fragments, respectively.

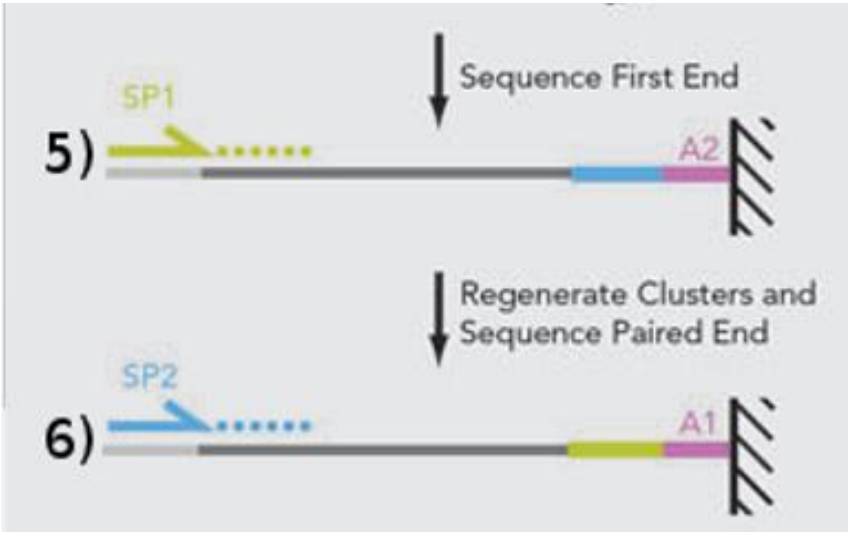
❸DNA fragments are linked to flow cell to generate DNA cluster.

❹Sequencing.



1.2 Paired-end library

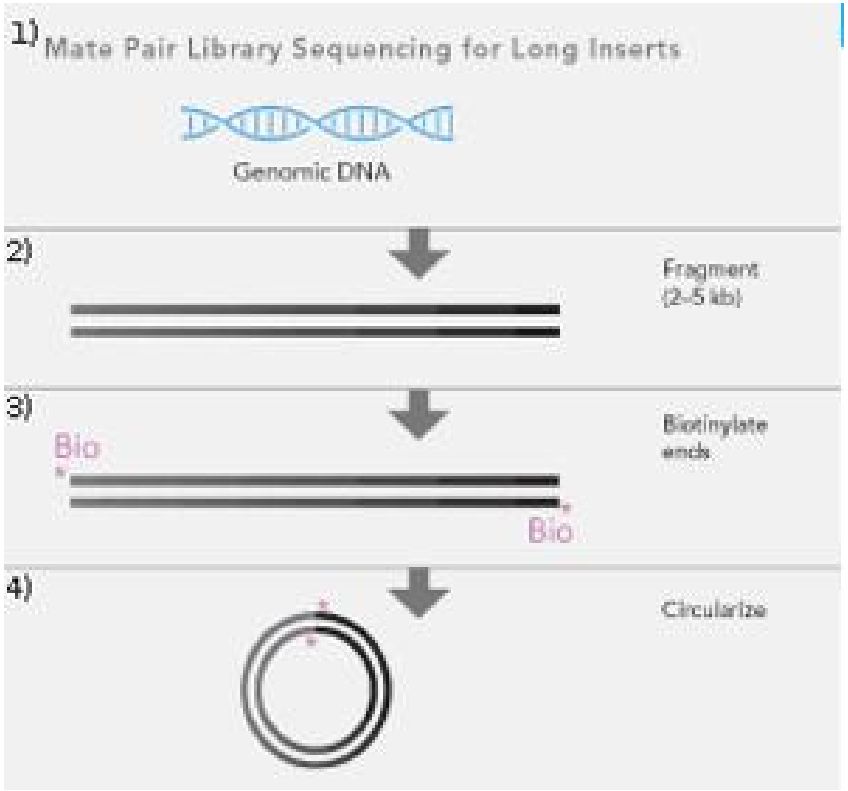
1) Genomic DNA
2) Fragment DNA into 200-500 bp.
3) Add adapter and primers in both ends of the sequence of interest.
4) Generate **clusters** (spots on flow cell of same sequences made by amplification).



1.2 Paired-end library:

5) **Sequencing step** using modified dNTPs and primers for know sequences (SP1 and SP2) you read the reads by light signals.

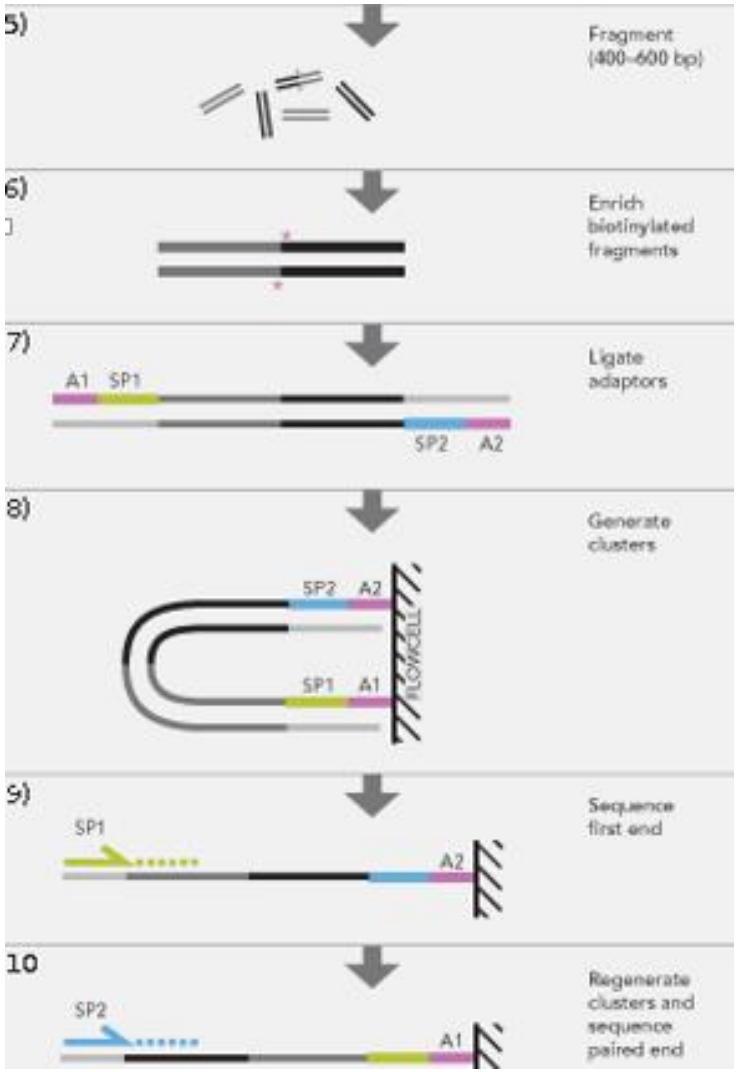
1.3 Mate-pair library:



2) Sequence fragmentation is made in bigger fragments (2-5 kb).

3) A addition of a Biotin in each 5′ ends is done (step 3).

4) The sequence with correct addition of Biotin will circularize and after a wash, the sequencing with non-circularized fragment will be thrown away.

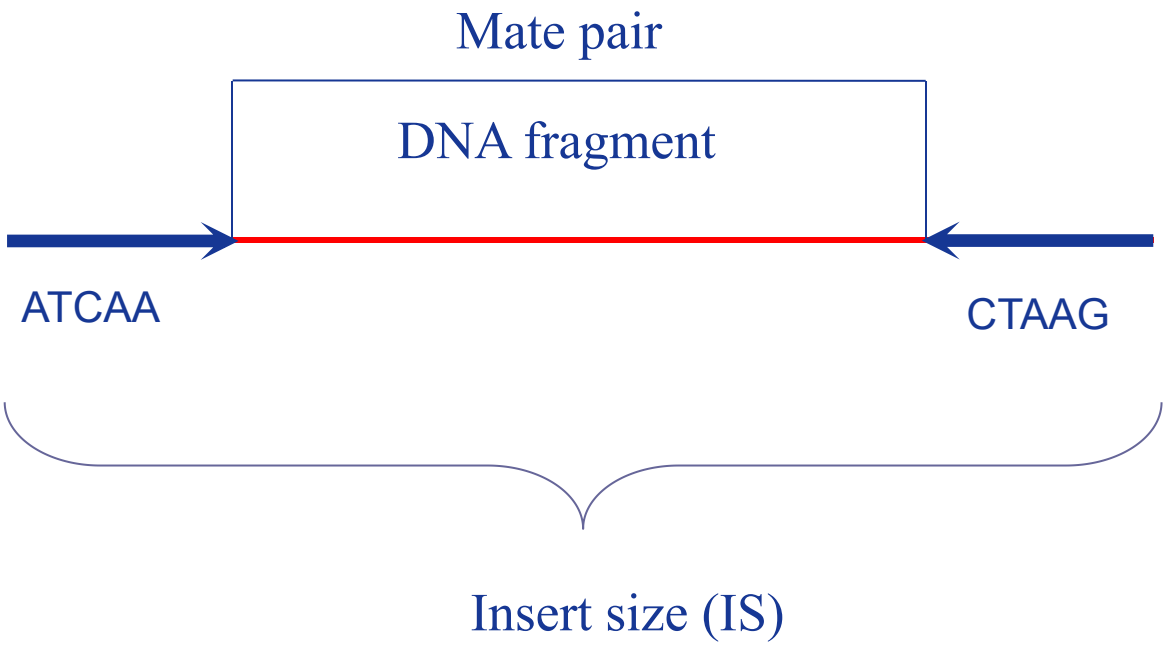


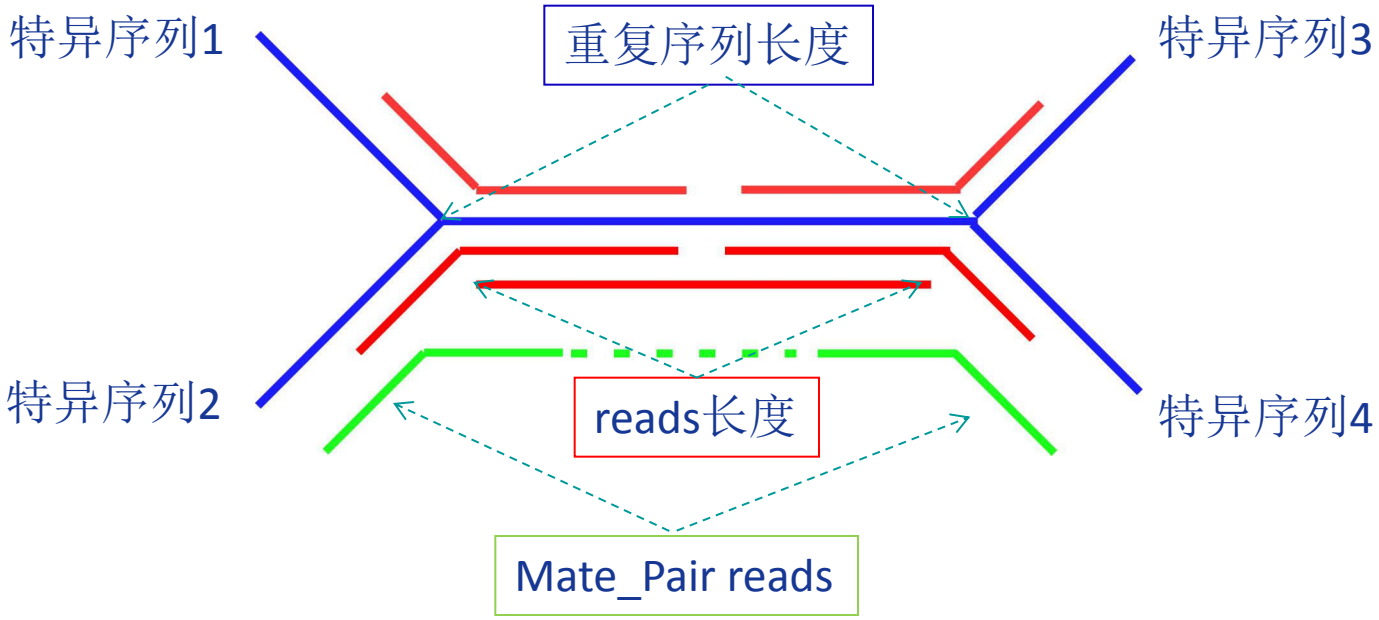
1.3 Mate-pair library

5) The circularized fragments will be cutted with the biotin in the middle and size-selected (400-600 bp).

6) Sequencing is done normally: adapter with primer sequence addition (step 7), the fragments will be spotted and clustered (step 8), and sequencing (step 9 and 10).

1.4 Why different sequencing methods



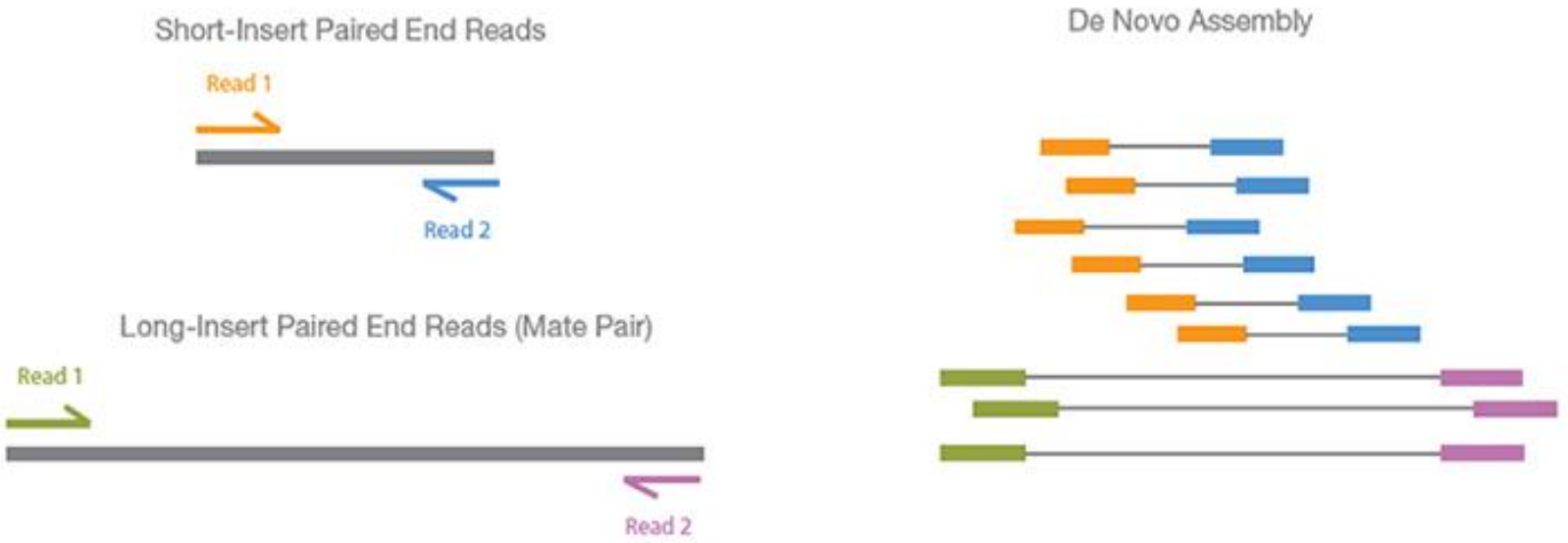


Questions : 重复序列长度超过read的长度，拼接将产生分支，无法延续，从而形成断点.

Solutions : ❶ 选择长read的测序仪器；

❷ 构建大片段Mate_Pair样品库进行测序.

Figure 5. *De Novo* Assembly with Mate Pairs



Using a combination of short and long insert sizes with paired-end sequencing results in maximal coverage of the genome for de novo assembly. Because larger inserts can pair reads across greater distances, they provide a better ability to read through highly repetitive sequences and regions where large structural rearrangements have occurred. Shorter inserts sequenced at higher depths can fill in gaps missed by larger inserts sequenced at lower depths. Thus a diverse library of short and long inserts results in better de novo assembly, leading to fewer gaps, larger contigs, and greater accuracy of the final consensus sequence.

Summary:

❶ 样本建库不同；

❷ Single Read (SR) 只检测待测片段的一端序列信息，PE 或MP检测待测片段的两端序列信息；

❸ 信息具有互补性.

2. Sequencing read file format

Read file format is a text-based format for representing nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows to read by human being and computers.

2.1 Sequencing read file in Fasta format

2.2 Sequencing read file in Fastq format

2.1 Fasta format

❶ The 1st line starts either with a ">" (greater-than) symbol or, less frequently, a ";" (semicolon) and was taken as a comment.

❷ The 2nd line is the actual sequence itself in standard one-letter code. Anything other than a valid code would be ignored.

2.1 Fasta format

❸ It was also common to end the sequence with an "*" (asterisk) character.

❹ Leaving a blank line between the description and the sequence.

2.1 Fasta format--*cases*

```
;LCBO - Prolactin precursor - Bovine
; a sample sequence in FASTA format
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*

>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*

>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

ASF-1.fa

```
>@HWI-ST216_0180:4:1101:1096:2196#GGCTAC/1
TTTTCAGNGAATACTGCAAATCAATAAACTCTTTAG
>@HWI-ST216_0180:4:1101:1158:2236#GGCTAC/1
AAAAGCTCATTTCCTATAGTTAACAGGACATGCCTT
>@HWI-ST216_0180:4:1101:1448:2211#GGCTAC/1
ATTATATAAGATAGCGGCTTTTTCCGTTAGTTTCCT
>@HWI-ST216_0180:4:1101:1331:2227#GGCTAC/1
CACGTTCTCTGTCCCCAATGGTATTTGCATCCCTGT
>@HWI-ST216_0180:4:1101:1376:2237#GGCTAC/1
GCGTCCCTTAGCTGAACTACCCAAACGTACGAATGC
```

ASF-2.fa

```
>@HWI-ST216_0180:4:1101:1096:2196#GGCTAC/2
TCAAGAAAACAACCTGATTATGCTAAGAAAGTAGAG
>@HWI-ST216_0180:4:1101:1158:2236#GGCTAC/2
TACGGTTAATACTTTCCTCTTCGTCTTTTTCTACAC
>@HWI-ST216_0180:4:1101:1448:2211#GGCTAC/2
CAAAACGAATTAAAAAATATGACCGTATTTCTTTTG
>@HWI-ST216_0180:4:1101:1331:2227#GGCTAC/2
GTTCAGATCTTTACAAAGCAATGAAAAAATTCTTCT
>@HWI-ST216_0180:4:1101:1376:2237#GGCTAC/2
AAATTATCTTGTTTCTTTTTGACGTTCTTTGGTACG
```

- Reads are often stored in fasta files
- Separate file for forward and reverse pairs
- header line -- read name/pairing info
- sequence line -- nucleotides

2.2 Fastq format

❶ Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description.

❷ Line 2 is the raw sequence letters.

❸ Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again.

2.2 Fastq format

❹ Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

The character '!' represents the lowest quality while '~' is the highest.

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

ASF-1.fastq

```
@HWI-ST216_0180:4:1101:1096:2196#GGCTAC/1
TTTTCAGNGAATACTGCAAATCAATAAACTCTTTAG
+HWI-ST216_0180:4:1101:1096:2196#GGCTAC/1
ceedb]]B[[]]]][fffff\ddddedeeedf_fbd
@HWI-ST216_0180:4:1101:1158:2236#GGCTAC/1
AAAAGCTCATTTCCTATAGTTAACAGGACATGCCTT
+HWI-ST216_0180:4:1101:1158:2236#GGCTAC/1
ggggggggggggggggegggfffgggggggggfg
@HWI-ST216_0180:4:1101:1448:2211#GGCTAC/1
ATTATATAAGATAGCGGCTTTTTCCGTTAGTTTCCT
```

ASF-2.fastq

```
@HWI-ST216_0180:4:1101:1096:2196#GGCTAC/2
TCAAGAAAACAACCTGATTATGCTAAGAAAGTAGAG
+HWI-ST216_0180:4:1101:1096:2196#GGCTAC/2
c_\cNaZUZIZ^_aZfbf\fdeeeeXc`[]VRYYY\
@HWI-ST216_0180:4:1101:1158:2236#GGCTAC/2
TACGGTTAATACTTTCCTCTTCGTCTTTTTCTACAC
+HWI-ST216_0180:4:1101:1158:2236#GGCTAC/2
gggggegggfggggfgfgggegegggggggggggg
@HWI-ST216_0180:4:1101:1448:2211#GGCTAC/2
CAAAACGAATTAAAAAATATGACCGTATTTCTTTTG
```

- Most reads are stored in fastq
- 4 lines per read

- header line: @SEQUENCE_ID
- sequence line
- line beginning with +
- encoded quality value line

3. Read quality control

Raw reads are always in fastq format. Usually, Quality score (Q-score) of a raw reads is used to measure the confidence in that base calling's identity.

❶ $Q_{\text{-score}} = -10 * \log_{10}P$ (P为碱基识别出错的概率)

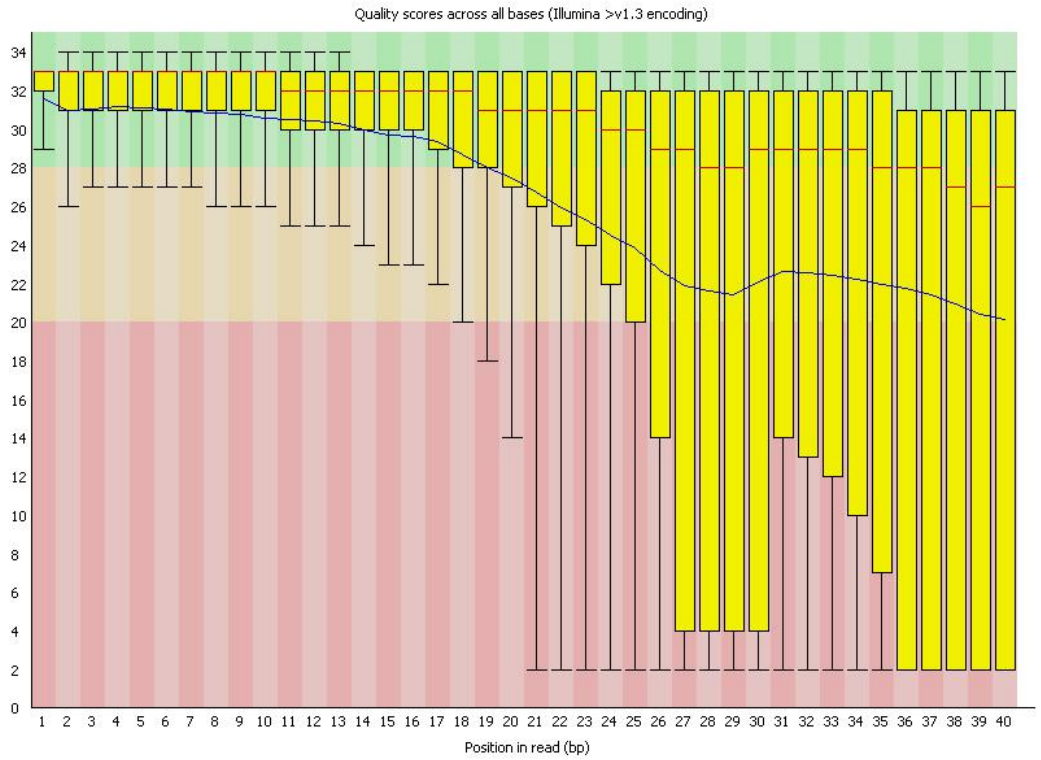
The higher Q-score, the lower P_{error}.

碱基质量值越高表明碱基识别越可靠，碱基测错的可能性越小。

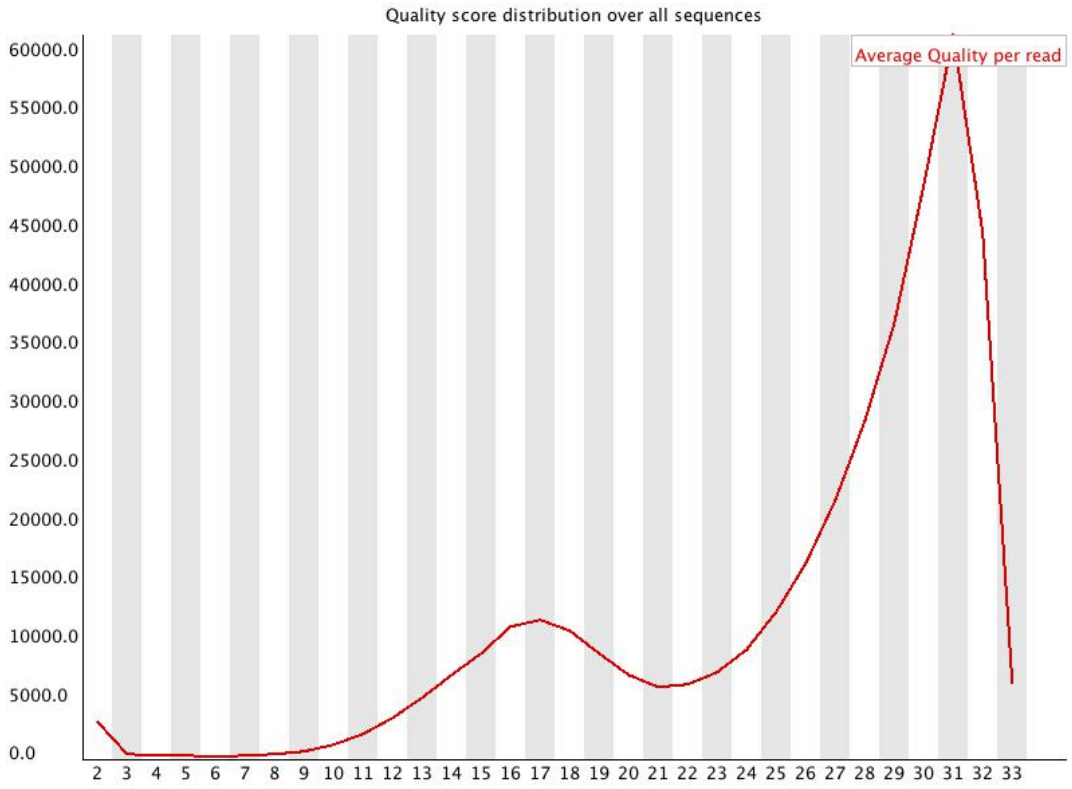
Quality Score	Base Call Accuracy	Incorrect Base Call
Q10	90 %	1/10
Q20	99 %	1/100
Q30	99.9 %	1/1000
Q40	99.99 %	1/10000

❸ Tools could be used to check Q-score for raw reads, such as fastqc. (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to check the Q-score for the raw reads.

fastqc [-o output dir] [--(no)extract] [-f fastq|bam|sam] [-c contaminant file] seqfile1 .. seqfileN



- 红色表示中位数，黄色是25%-75%区间，触须是10%-90%区间，蓝线是平均数。
- 若任一位置的中位数低于25，报“WARN“。
- 若任一位置的中位数低于20，报“FAIL“。



- 横轴为quality，纵轴是reads数目。
- 当峰值小于27（错误率0.2%）时报“WARN”。
- 当峰值小于20（错误率1%）时报"FAIL"。

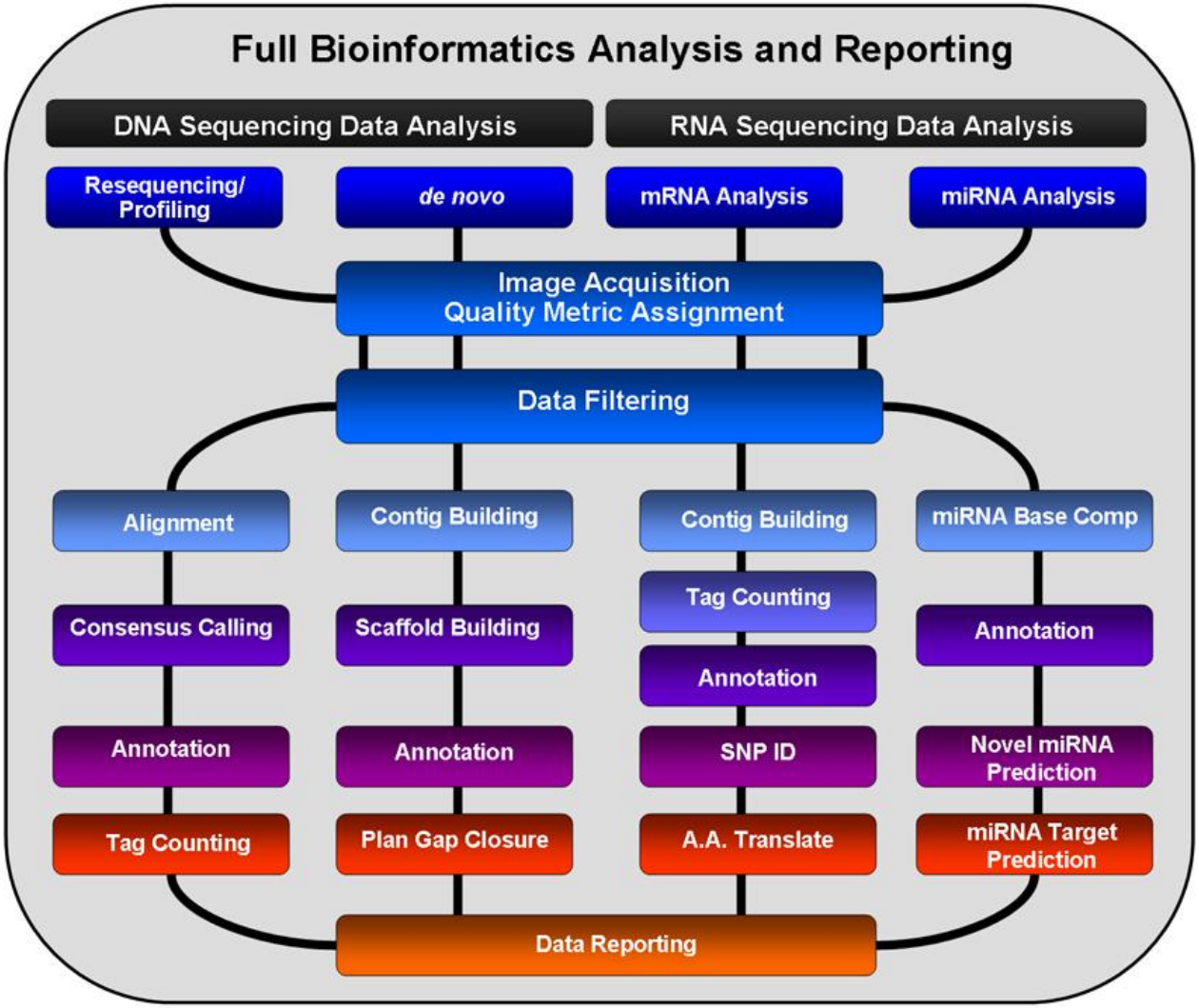
❹ Quality control

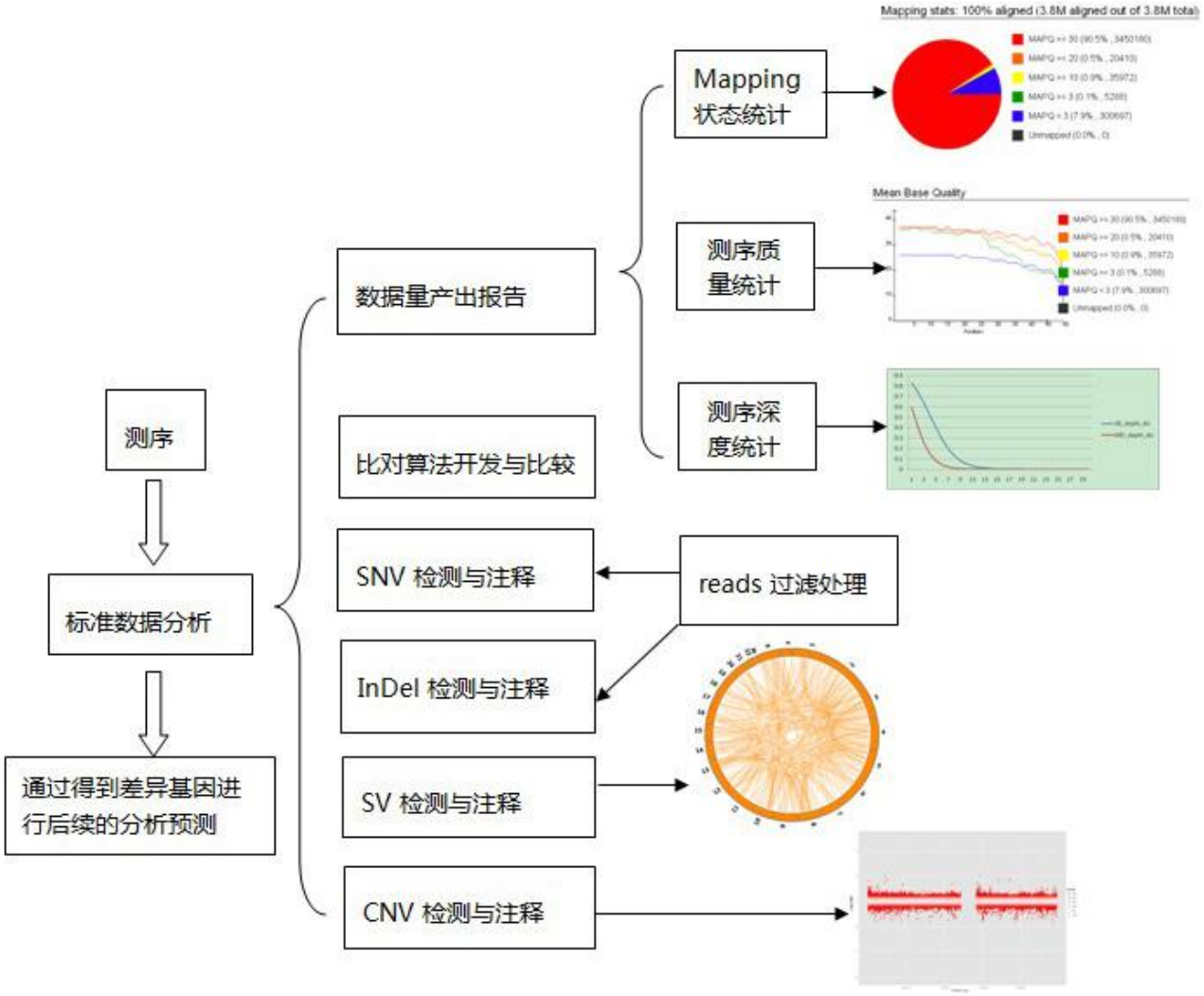
- Delete the PCR primer

（先使用BWA软件进行mapping统计，再使用samtools rmdup 进行去重复）

BWA index → BWA aln → BWA sampe → Samtools view → Samtools sort → Samtools rmdup

- Filter the raw reads with low Q-score.


Full Bioinformatics Analysis and Reporting
DNA Sequencing Data Analysis
RNA Sequencing Data Analysis
Resequencing/ Profiling
de novo
mRNA Analysis
miRNA Analysis
Image Acquisition
Quality Metric Assignment
Data Filtering
Alignment
Contig Building
Contig Building
miRNA Base Comp
Tag Counting
Consensus Calling
Scaffold Building
Annotation
Annotation
Annotation
Annotation
SNP ID
Novel miRNA Prediction
Tag Counting
Plan Gap Closure
A.A. Translate
miRNA Target Prediction
Data Reporting


测序
标准数据分析
通过得到差异基因进行后续的分析预测
数据量产出报告
比对算法开发与比较
SNV 检测与注释
InDel 检测与注释
SV 检测与注释
CNV 检测与注释
Mapping 状态统计
测序质量统计
测序深度统计
reads 过滤处理
Mapping stats: 100% aligned (3.8M aligned out of 3.8M total)
Mean Base Quality

Thanks for your attentions!