目录

一 项目概况 1.1
1.1 项目概况 1.1.1
1.2 项目背景 1.1.2
1.2.1 BSA简介 1.1.2.1
1.2.2 样本信息 1.1.2.2
1.3 项目流程 1.1.3
1.4 分析项目 1.1.4
二 数据过滤、整理及质量评估 1.2
2.1 项目概况 1.2.1
2.2 下机数据统计 1.2.2
2.3 数据质控 1.2.3
2.3.1 碱基质量分布 1.2.3.1
2.3.2 Base Content分布 1.2.3.2
2.3.3 GC Content分布 1.2.3.3
2.3.4 Sequence Base Quality 1.2.3.4
2.4 高质量数据获取 1.2.4
三 对比分析 1.3
3.1 参考基因组整理 1.3.1
3.2 序列对比 1.3.2
3.3 序列对比结果概述 1.3.3
3.4 测序深度及覆盖度概述 1.3.4
四 SNP分析 1.4
4.1 SNP 检测 1.4.1
4.2 SNP 注释 1.4.2
五 InDel分析 1.5
5.1 InDel检测 1.5.1
5.2 InDel注释 1.5.2
六 SNP Index 分析 1.6
6.1 SNP Index计算 1.6.1
6.2 SNP Index分布 1.6.2
6.3 SNP Index差异分析 1.6.3
七 目标性状区域分析 1.7
八 参考文献 1.8

一 项目概况

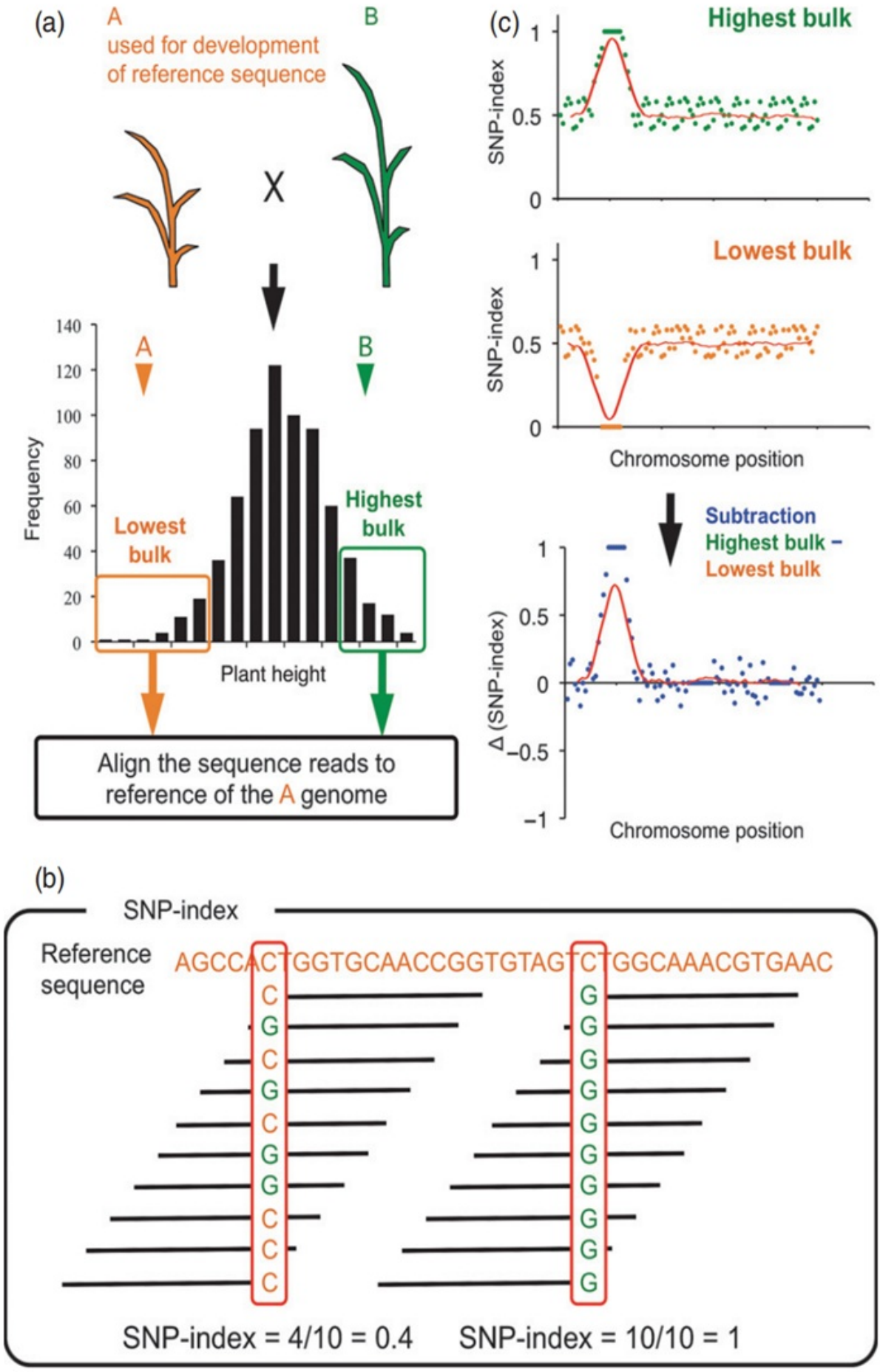
1.1 项目概况

项目编号	**Q202106003**
项目类型	BSA定位分析
物种名称	家蚕
样本形式	DNA
测序平台	Illumina HiSeq
数据量	25G
分析项目	标准生物信息分析
完成日期	2021-04-05
产品经理	贺兴

1.2 项目背景

1.2.1 BSA简介

BSA（Bulked Segregant Analysis），即混合分组分析法，是一种利用极端性状进行功能基因定位的一种方法。具体过程为挑选两组具有极端性状的个体，分别混池后测序，然后计算多态性位点（一般用 SNP位点）的等位基因频率（allele frequency）。等位基因频率有显著差异的多态性位点区域，即是控制目标性状的基因所在的染色体区段。BSA 的基本原理如下图：


(a)
A
used for development
of reference sequence
B
X
A
B
Frequency
Lowest
bulk
Highest
bulk
Plant height
Align the sequence reads to
reference of the A genome
(c)
Highest bulk
SNP-index
Lowest bulk
SNP-index
Chromosome position
Subtraction
Highest bulk −
Lowest bulk
Δ (SNP-index)
Chromosome position
(b)
SNP-index
Reference
sequence
AGCCACTGGTGCAACCGGTGTAGTCTGGCAAACGTGAAC
SNP-index = 4/10 = 0.4
SNP-index = 10/10 = 1

1.2.2 样本信息

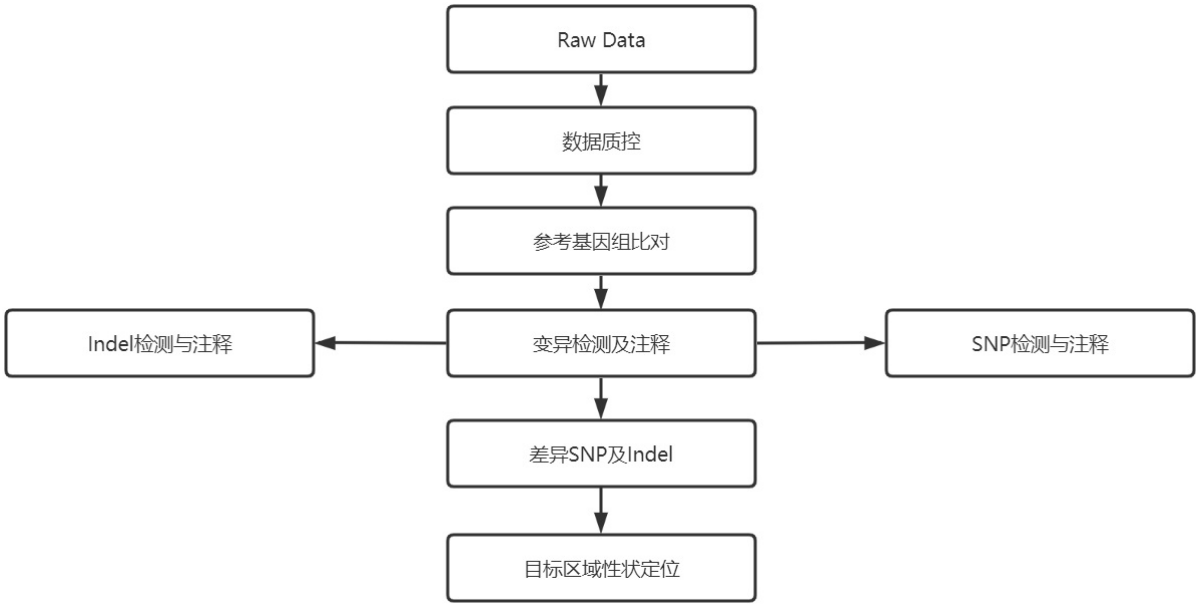
BSA（Bulked Segregant Analysis），即混合分组分析法，是一种利用极端性状进行功能基因定位的一种方法。具体过程是挑选两组具有极端性状的个体，分别混池后测序，然后计算多态性位点（一般用 SNP 位点）的等位基因频率（allele frequency）。等位基因频率有显著差异的多态性位点区域，即是控制目标性状的基因所在的染色体区段。

表1 样本信息

样本类别	个体数量	**Sequencing platform**	**Sequencing Mode**
P50（亲本）	1	lllumina NovaSeq	Paired-end 2*150bp
Spring（亲本）	1	lllumina NovaSeq	Paired-end 2*150bp
WT（混池）	1	lllumina NovaSeq	Paired-end 2*150bp
Mut （混池）	1	lllumina NovaSeq	Paired-end 2*150bp

- Sequencing Platform：测序平台；
- Sequencing Mode： 测序模式。

1.3 项目流程



1.4 分析项目

序号	分析项目	类型	分析
1	数据过滤、整理及质量评估	A	√
2	数据质控	A	√
3	高质量数据获取	A	√
4	参考基因组整理	A	√
5	序列比对	A	√
6	序列比对结果概述	A	√
7	SNP 检测统计	A	√
8	SNP 注释	A	√
9	InDel 检测统计	A	√
10	InDel 注释	A	√
11	SNP index 计算	A	√
12	SNP index 差异分布	A	√
13	目标区域性定位	A	√
14	候选 SNP 及基因筛选	A	√
15	候选区域 InDel 分子标记开发	B	√

- A：标准信息分析
- B：个性化分析

二 数据过滤、整理及质量评估

2.1 项目概况

本项目构建插入片段为 400bp的文库，利用第二代测序技术（Next-Generation Sequencing，NGS），基于 Illumina NovaSeq 测序平台，对这些文库进行双末端（Paired-end，PE）测序。本次共构建了4个文库，文库的基本情况见表2。

表2 测序概述

Sample	Insert Size	Sequencing Platform	Sequencing Mode
P50	400bp	lllumina NovaSeq	Paired-end 2*150bp
Spring	400bp	lllumina NovaSeq	Paired-end 2*150bp
WT	400bp	lllumina NovaSeq	Paired-end 2*150bp
Mut	400bp	lllumina NovaSeq	Paired-end 2*150bp

- Sample：样品名称；
- Insert Size：库插入片段长度；
- Sequencing Platform：测序平台；
- Sequencing Mode：测序模式

2.2 下机数据统计

分别对每个样品的下机数据进行统计，结果见表3。

表3 测序数据统计

Sample	ReadsNum.	Total Bases(bp)	N_rate	GC(%)	Q20(%)	Q30(%)
P50	73,540,094	11,031,014,100	0.0002	39.19	98.17	94.51
Spring	71,963,930	10,794,589,500	0.0002	39.19	98.11	94.39
WT	113,780,226	17,067,033,900	0.0002	39.18	98.23	94.69
Mut	113,744,760	17,061,714,000	0.0002	39.24	98.15	94.49

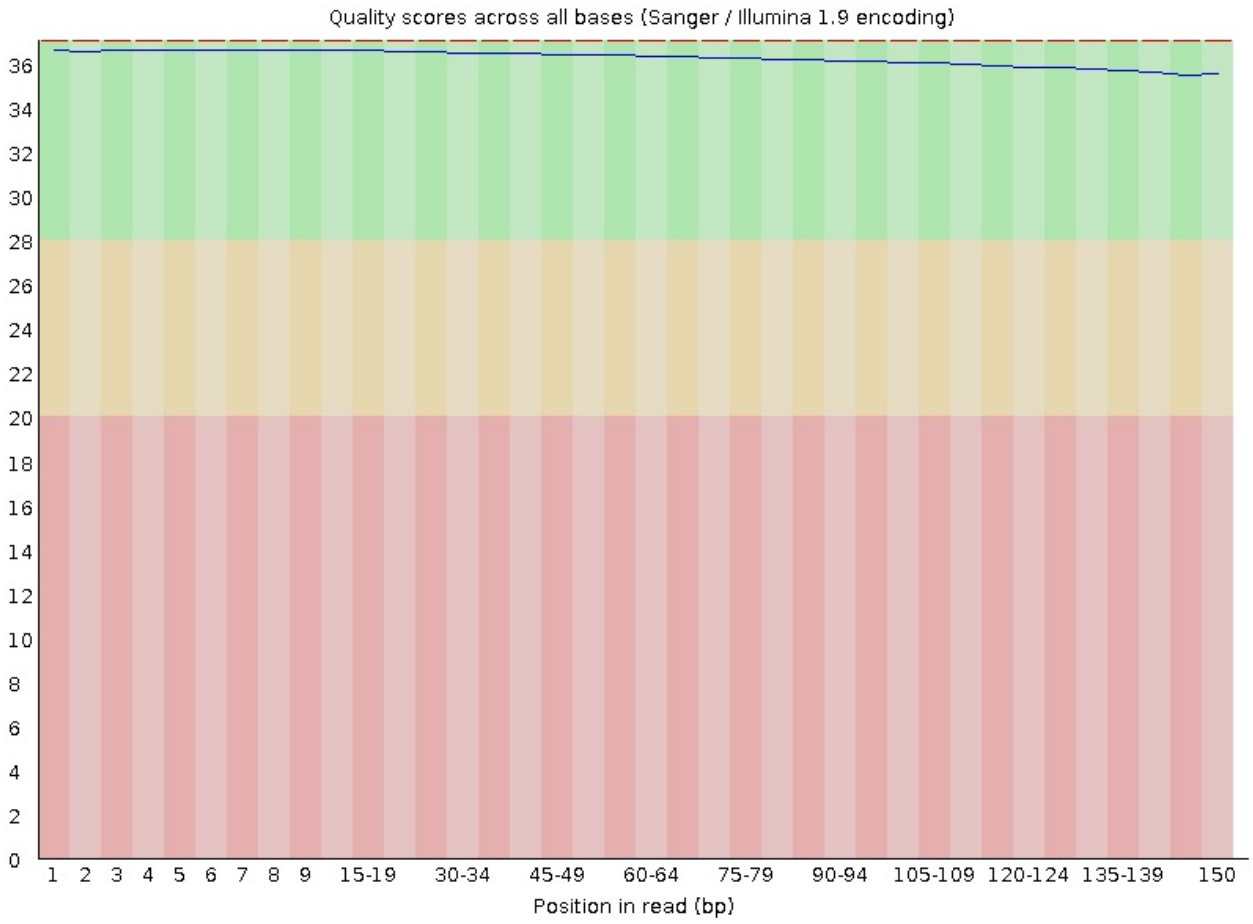
- Sample： 样品名；
- Reads Num.： Reads 总数；
- Total Bases(bp)：碱基总数；
- N_rate(%)： 模糊碱基所占百分比；
- GC(%)： GC 含量；
- Q20(%)： 碱基识别准确率在 99%以上的碱基所占百分比；
- Q30(%)： 碱基识别准确率在 99.9%以上的碱基所占百分比。

2.3 数据质控

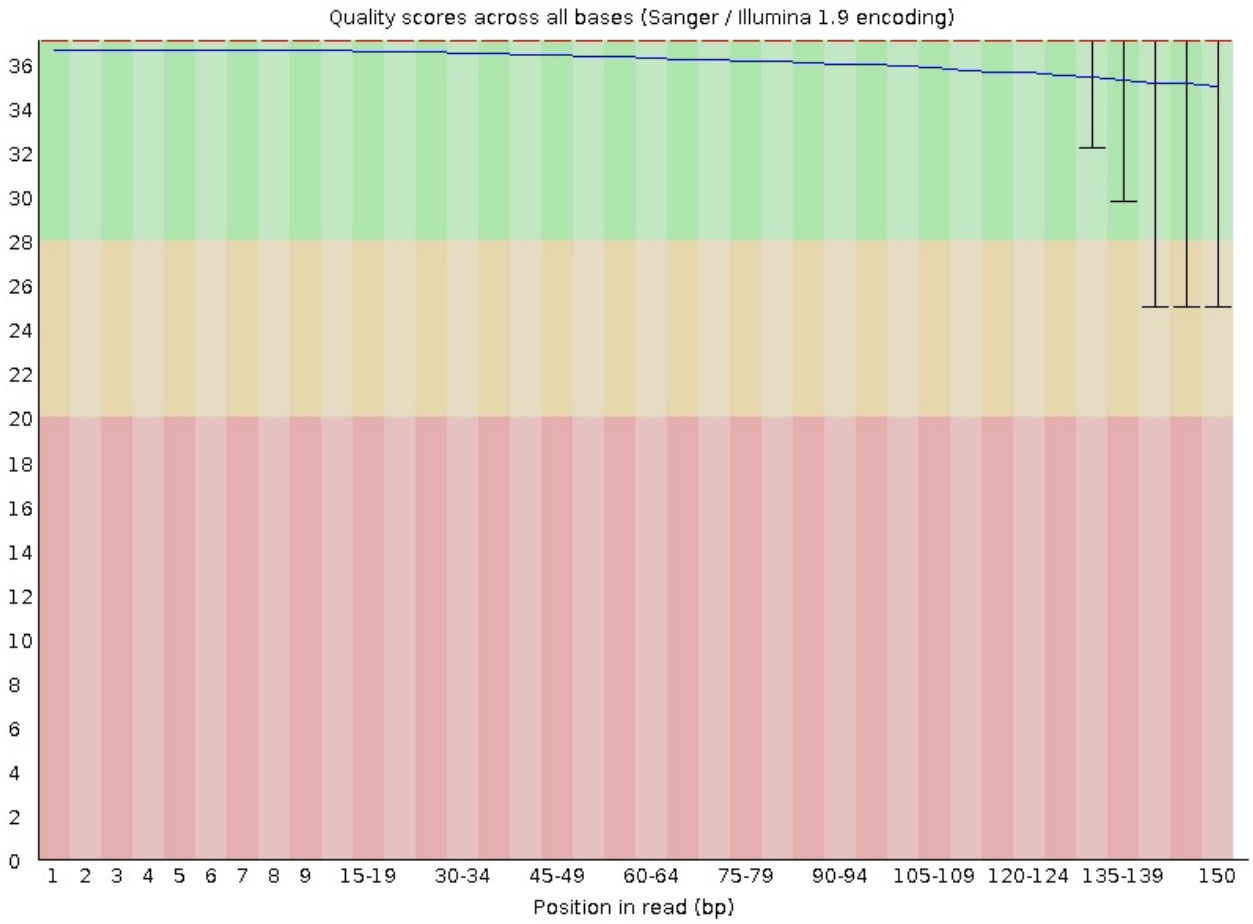
2.3.1 碱基质量分布

采用 FastQC（http://www.bioinformatics.babraham.ac.uk/projects/fastqc）对数据进行质量 控制，具体结果分别见图1-1、图1-2、图1-3 和 图1-4。

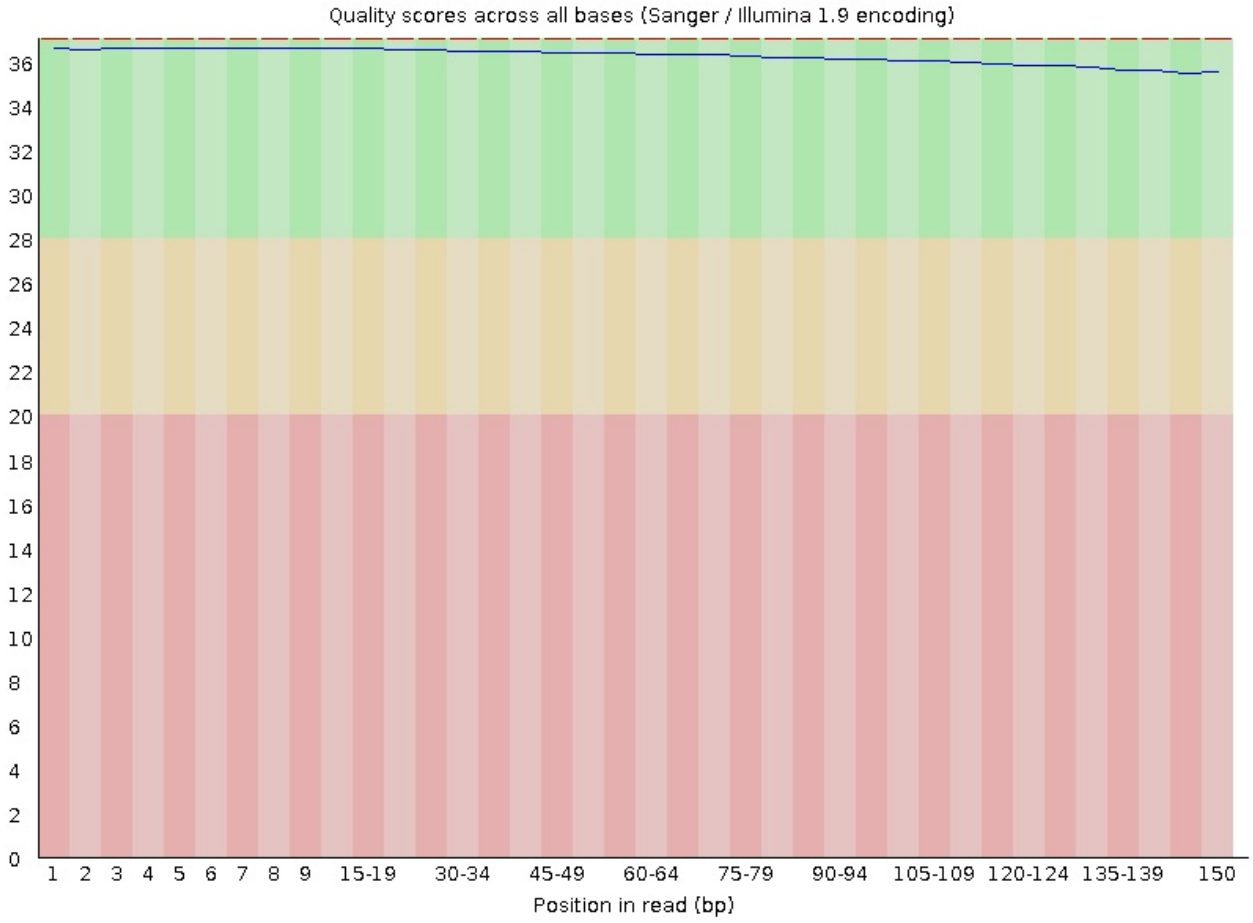
▼ Mut.HQ_R1



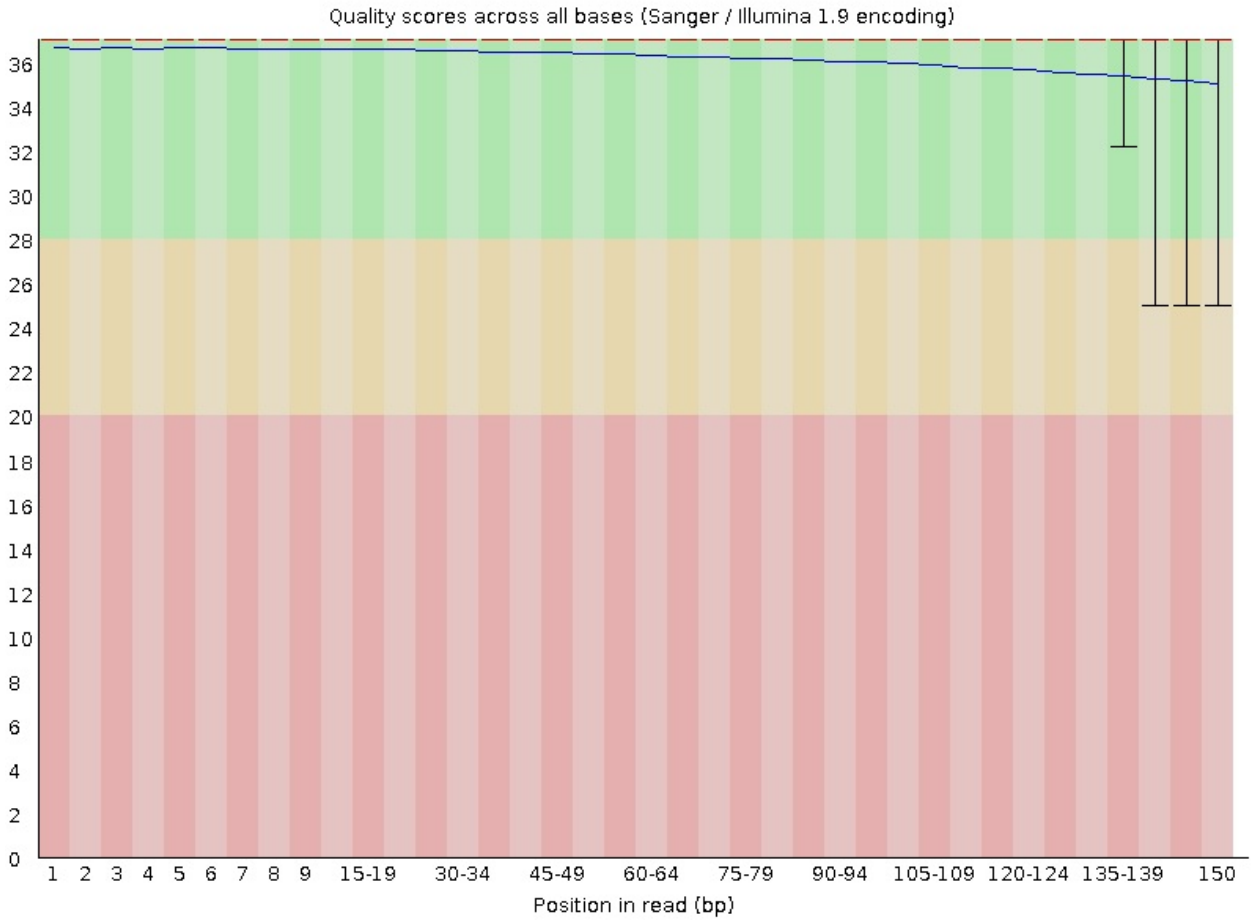
▼ Mut.HQ_R2


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 105-109 120-124 135-139 150
Position in read (bp)

▼ WT.HQ_R1


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 105-109 120-124 135-139 150
Position in read (bp)

▼ WT.HQ_R2

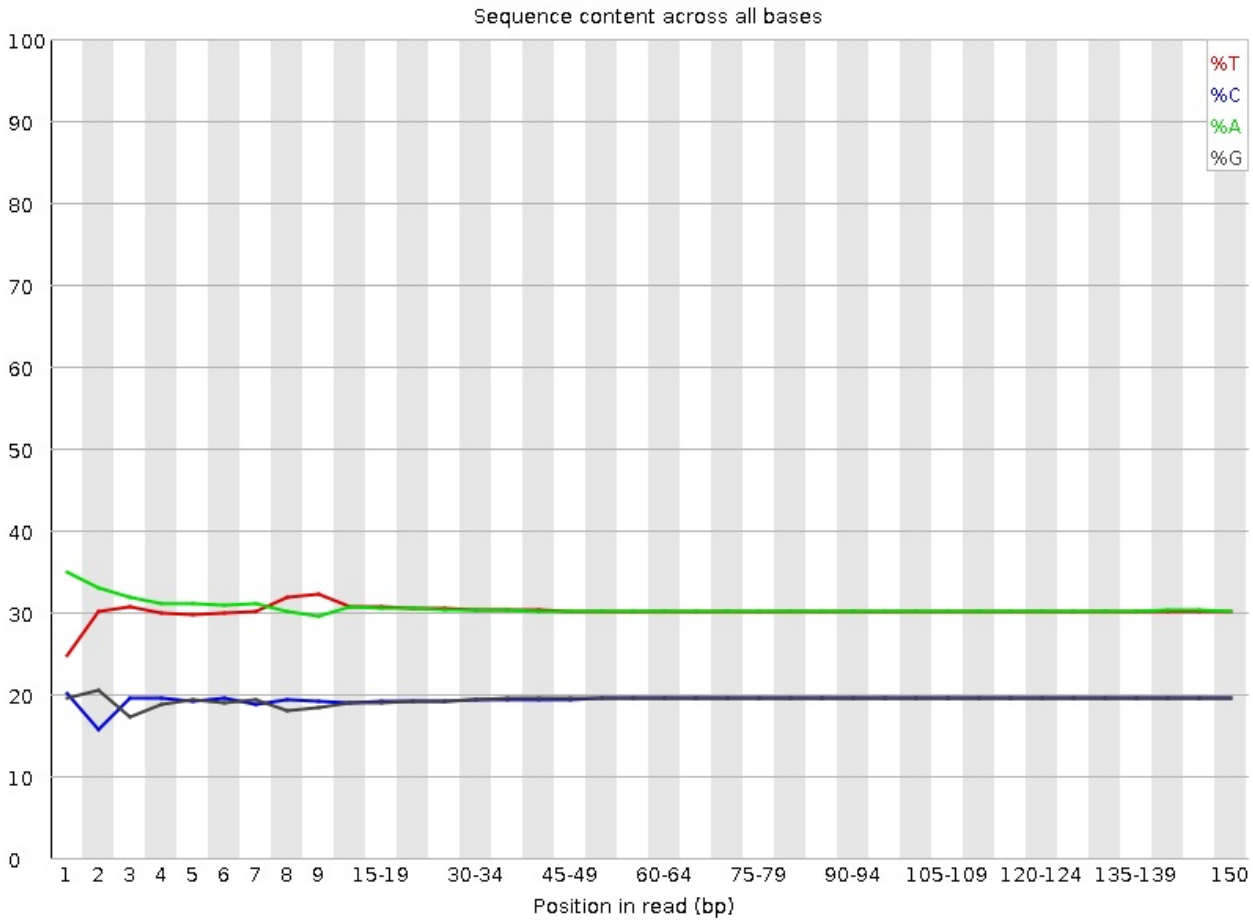


横坐标是 reads 碱基位置（5'->3'），纵坐标是所有 reads 在该位点碱基 Q 值统计。红线代表中位数，蓝 线代表平均数，黄线代表 25%-75%区间，触须是 10%-90%区间。一般而言，reads 的 5'端和 3'端的碱基质 量较低，中间部分的碱基质量较高。从图中可知，本次测序过滤后的数据平均质量非常高。

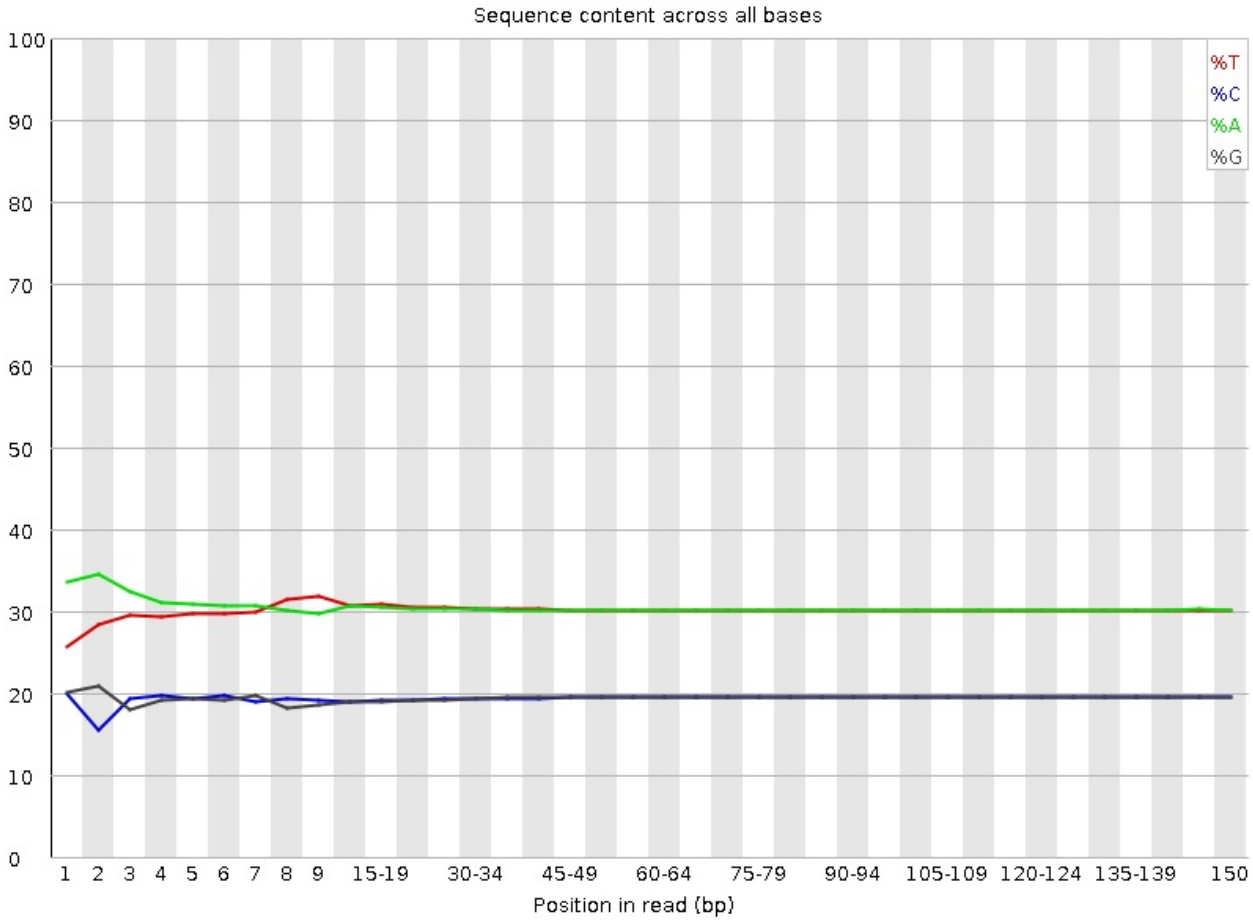
2.3.2 Base content 分布

Base content 分布主要用于检测测序数据是否存在 AT、GC 分离现象。通常，这种现象由建库或测序引入，并影响后续的生物信息分析结果。Base content 分布结果见图2-1、图2-2、图2-3 和 图2-4。结果显示，建库和测序表现出较好的均一度，可以进行后续信息分析。

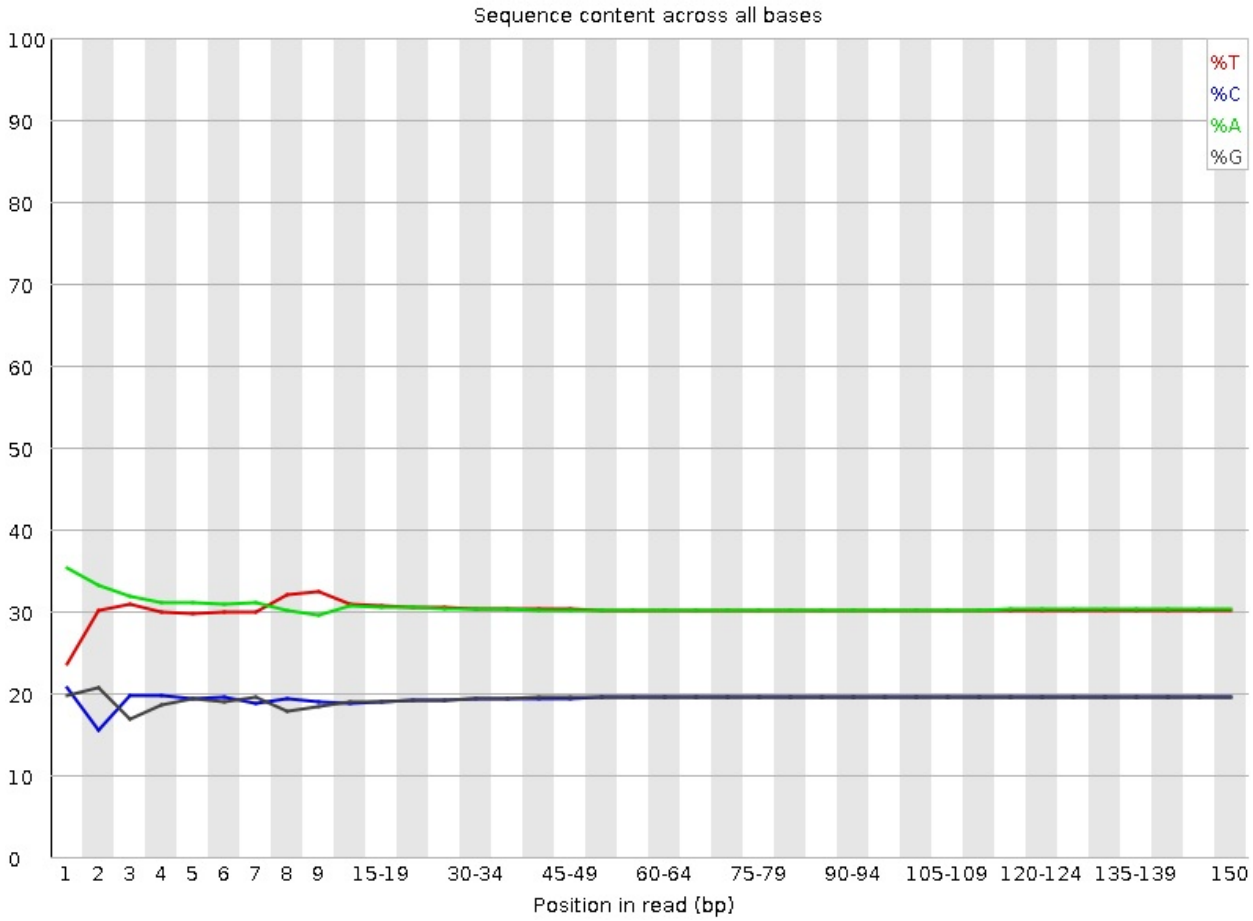
▼ Mut.HQ_R1



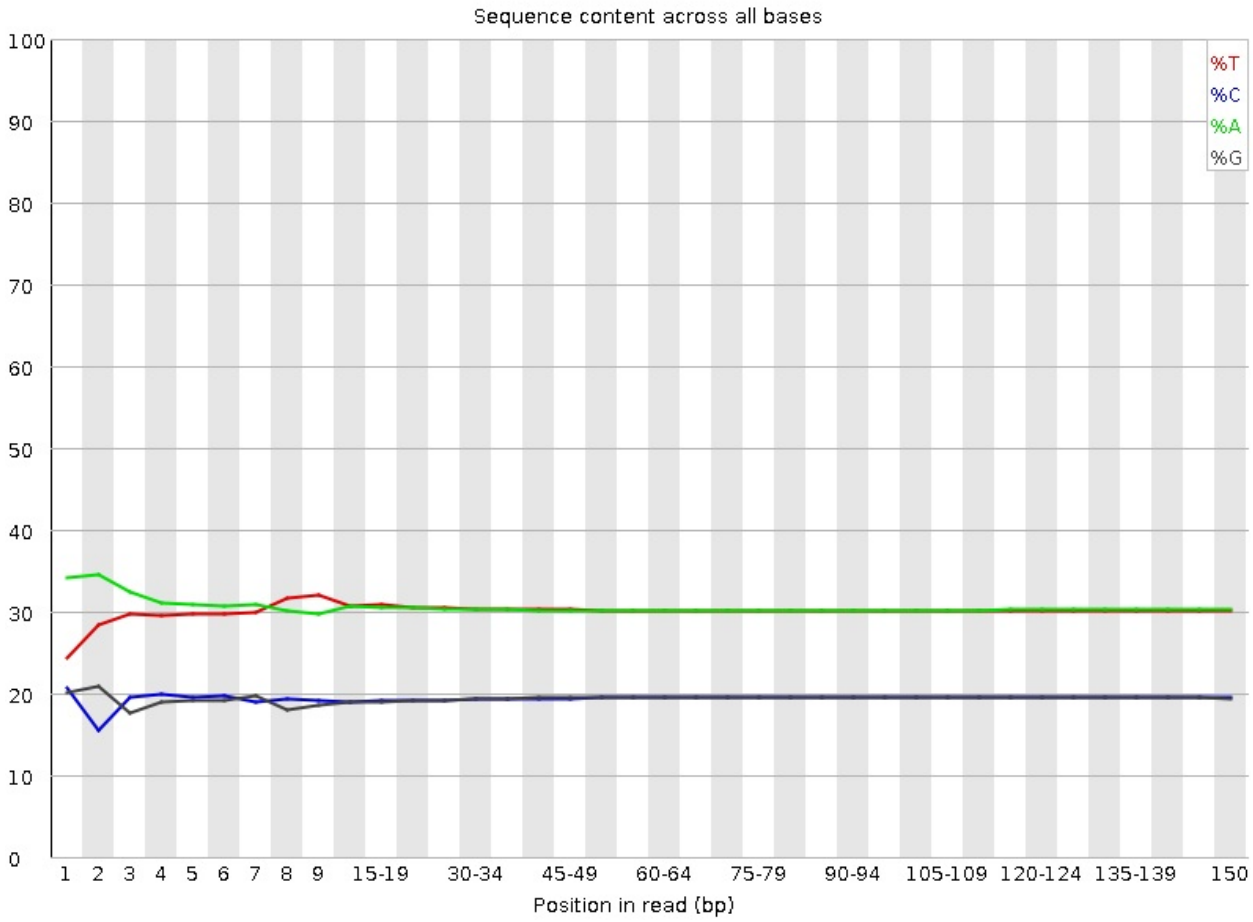
▼ Mut.HQ_R2


Sequence content across all bases
100
90
80
70
60
50
40
30
20
10
0
%T
%C
%A
%G
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 105-109 120-124 135-139 150
Position in read (bp)

▼ WT.HQ_R1


Sequence content across all bases
100
90
80
70
60
50
40
30
20
10
0
%T
%C
%A
%G
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 105-109 120-124 135-139 150
Position in read (bp)

▼ WT.HQ_R2

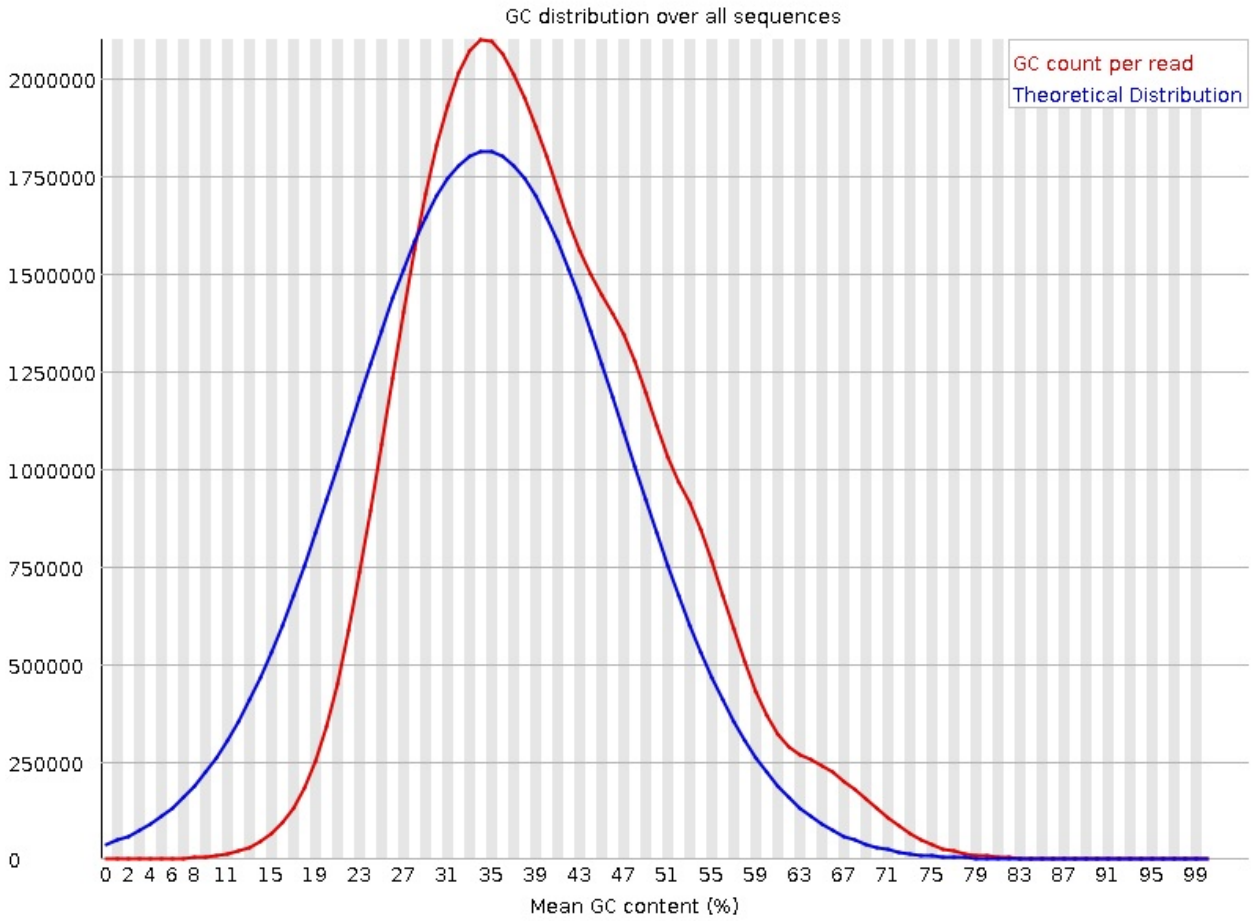


横坐标是 reads 中碱基位置（5’-3’），纵坐标是该位点某碱基所占的比例统计。对基于 Illumina 测序平 台的基因组测序中，基因组片段化过程中前几个位置的核苷酸组成表现出一定的偏好性，而这种偏好性与 测序的物种和实验室环境无关，但会影响基因组测序的均一化程度。除此之外，其余位置四种碱基的出现 频率应该是接近的。因此，好的建库和测序样品的四条线应该平行且接近，当部分位置的碱基出现 Bias 时， 则四条线在某些位置纷乱交织，往往提示有 overrepresented sequence 的污染，当所有位置的碱基比例一致 地表现出 Bias 时，即四条线平行但分开，往往代表文库有 Bias（建库过程或本身特点、或者是测序中的系 统误差）。

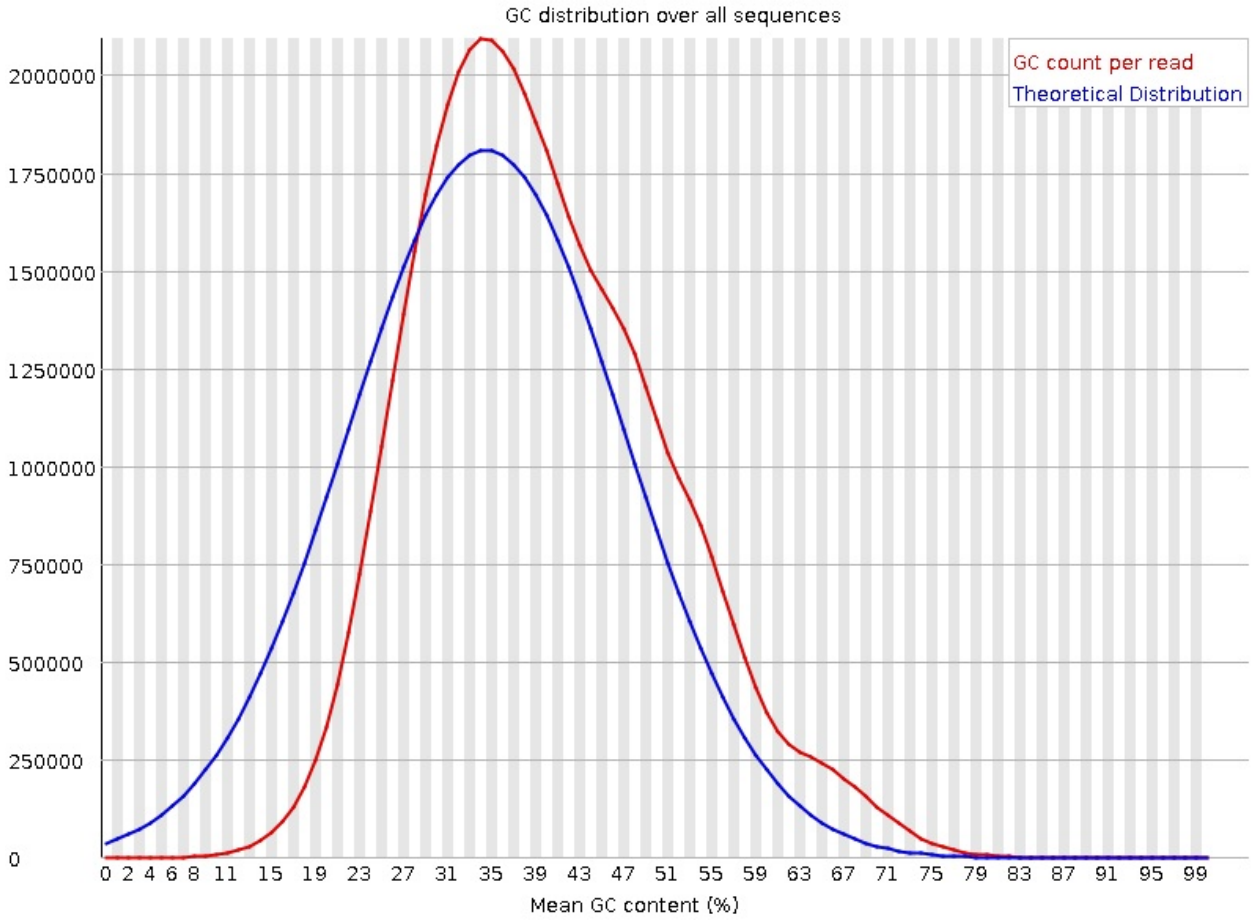
2.3.3 GC Content分布

GC content 分布主要用来检测测序数据的 GC 分布是否正常。曲线形状的偏差往往提示存在不均一的序列组分，这通常是由于文库的内源或外源污染，通过不同的峰值，大致能够判断污染性质和来源。GC content 分布结果见图3-1、图3-2、图3-3 和 图3-4。GC content 分布结果显示理论分布与实际分布一致，可以进行后续分析。

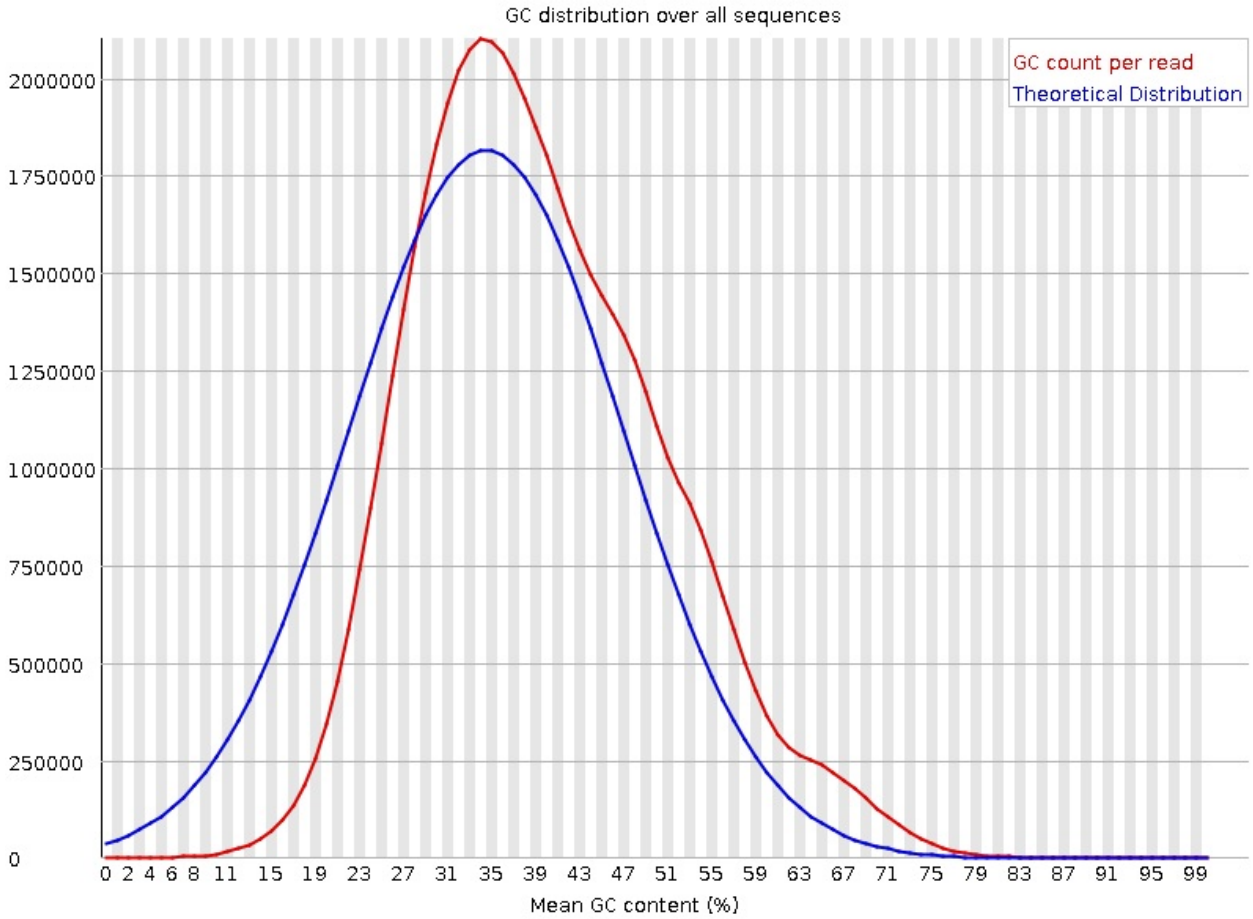
▼ Mut.HQ_R1



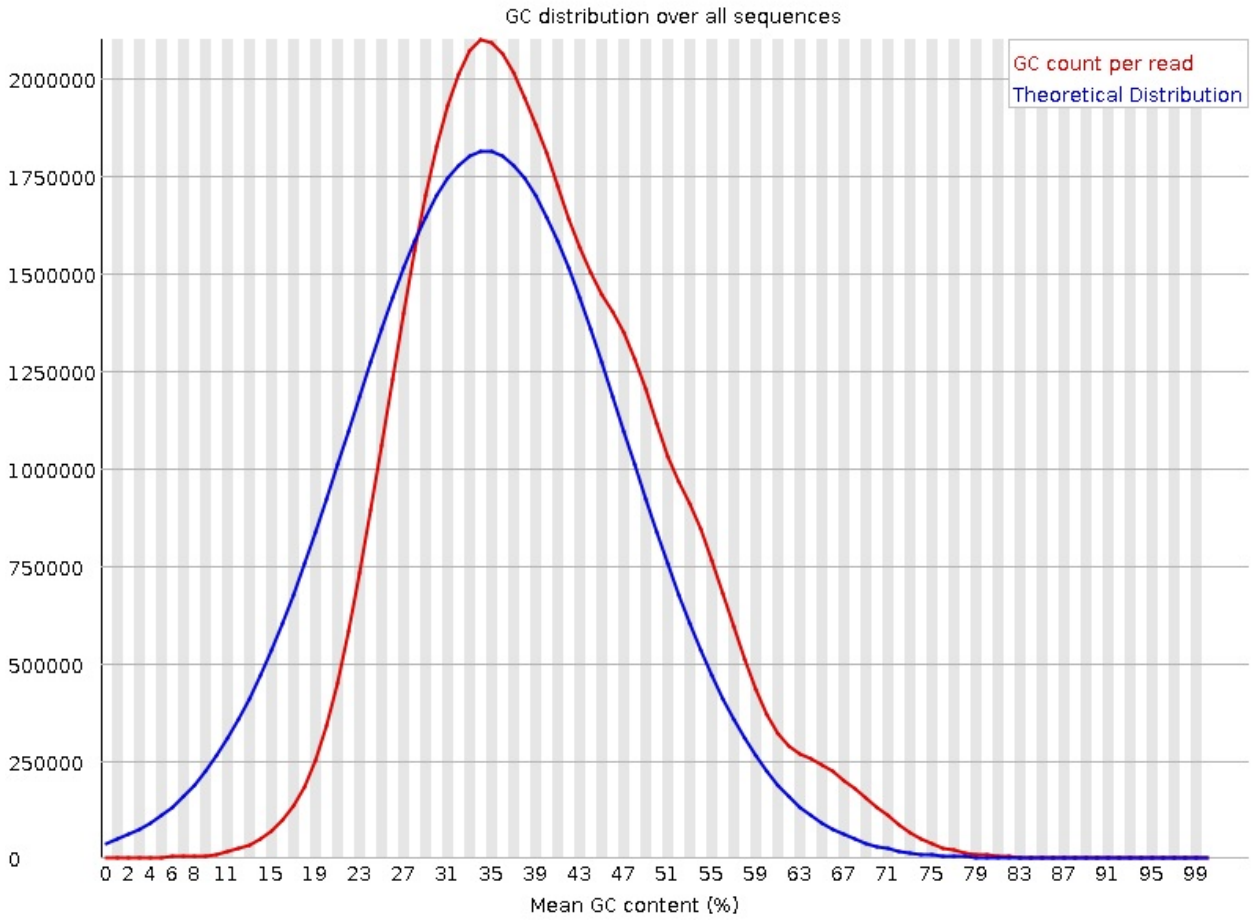
▼ Mut.HQ_R2


GC distribution over all sequences
GC count per read
Theoretical Distribution
2000000
1750000
1500000
1250000
1000000
750000
500000
250000
0
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)

▼ WT.HQ_R1


GC distribution over all sequences
GC count per read
Theoretical Distribution
2000000
1750000
1500000
1250000
1000000
750000
500000
250000
0
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)

▼ WT.HQ_R2

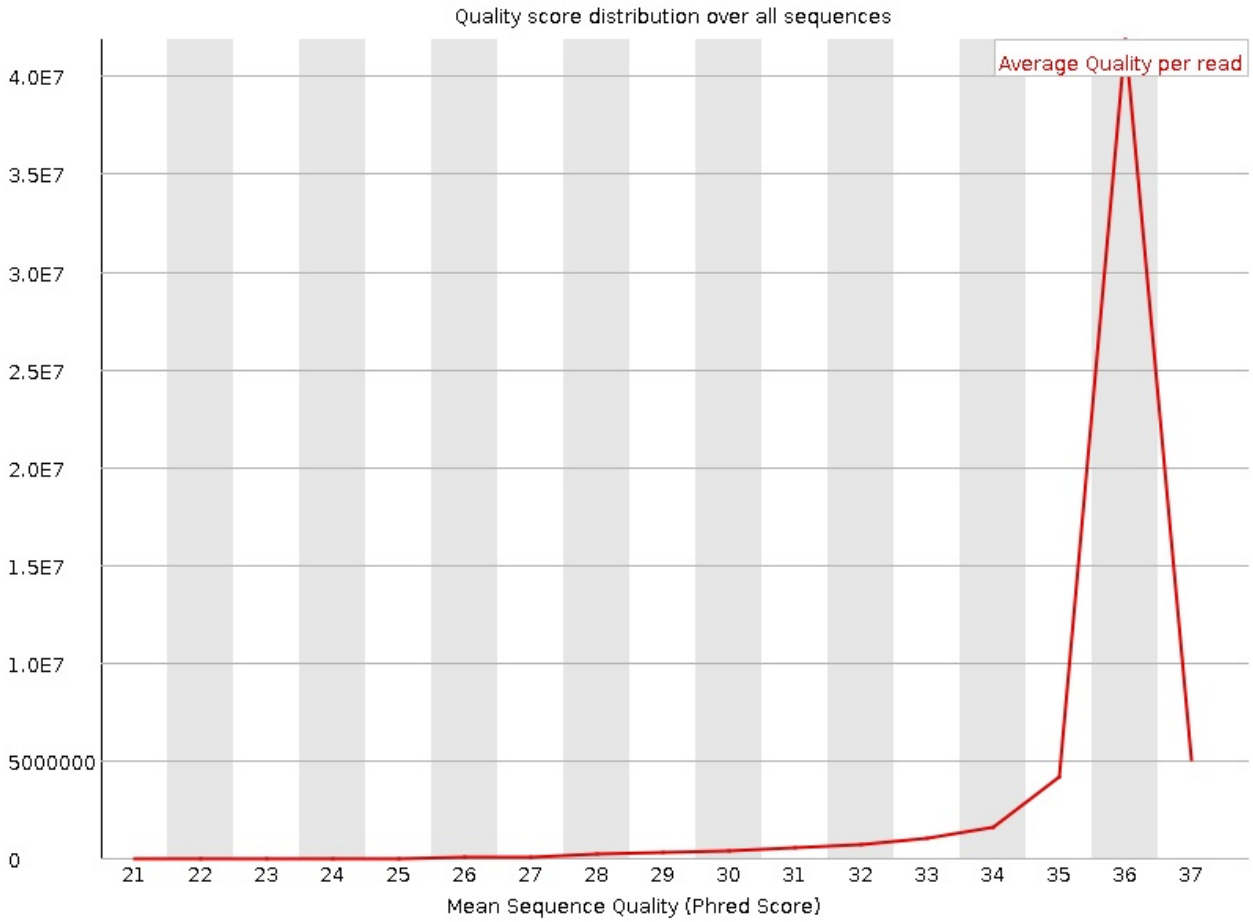


横坐标是 GC 比，纵坐标为对应的 reads 数量。红色表示的是实际分布曲线，蓝色的是理论分布曲线（峰值与物种特性有关）。

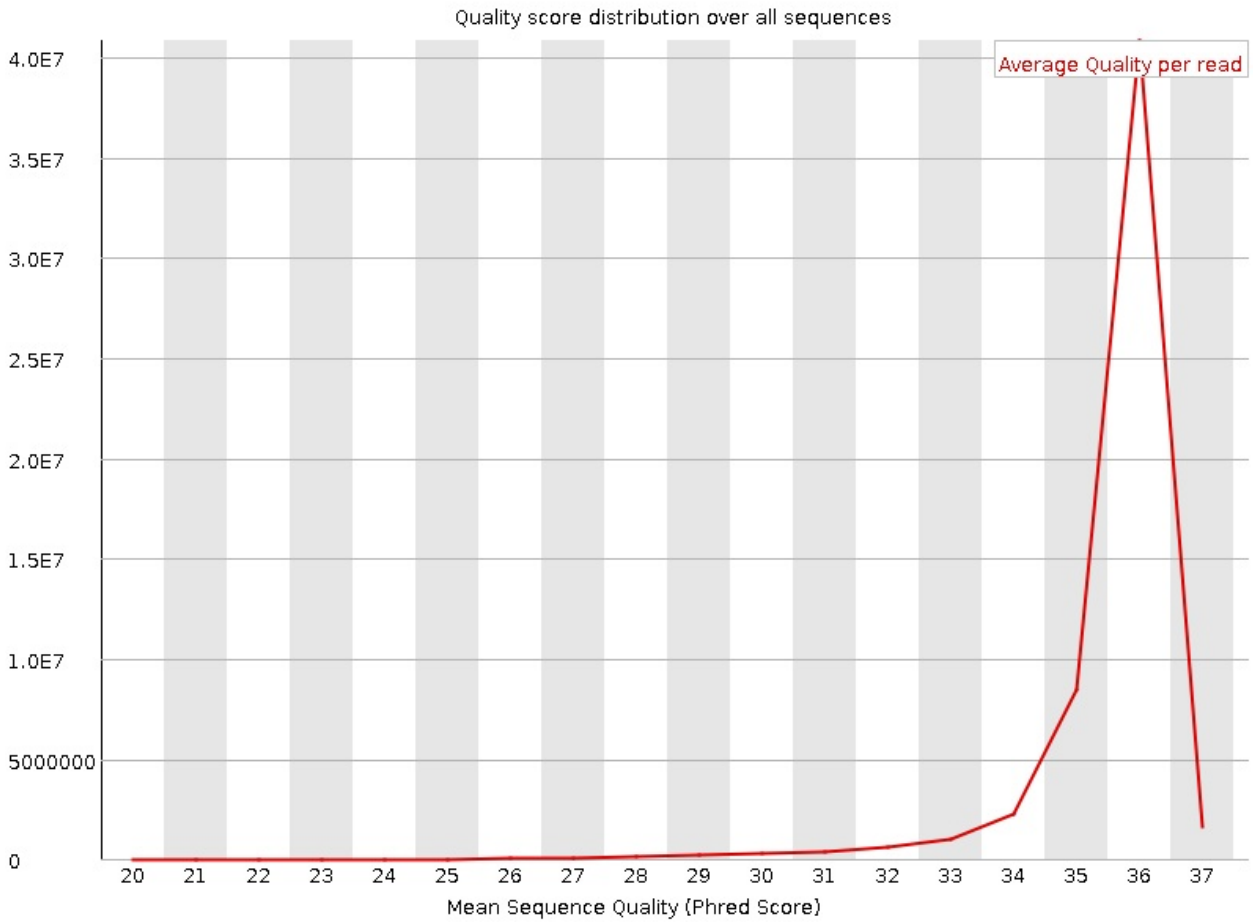
2.3.4 Sequence base quality

Sequence base quality 主要用来检测测序数据的平均质量分布情况。峰尖代表测序质量，峰宽表示测序整体质量分布，较大峰宽或者前拖尾峰表示测序数据的部分数据质量偏低，峰尖对应的值较低表明整体测序结果较差。本次测序的峰值在 Q40 左右，表示测序质量较高。Sequence base quality 结果见图4-1、图4-2、图4-3 和 图4-4。

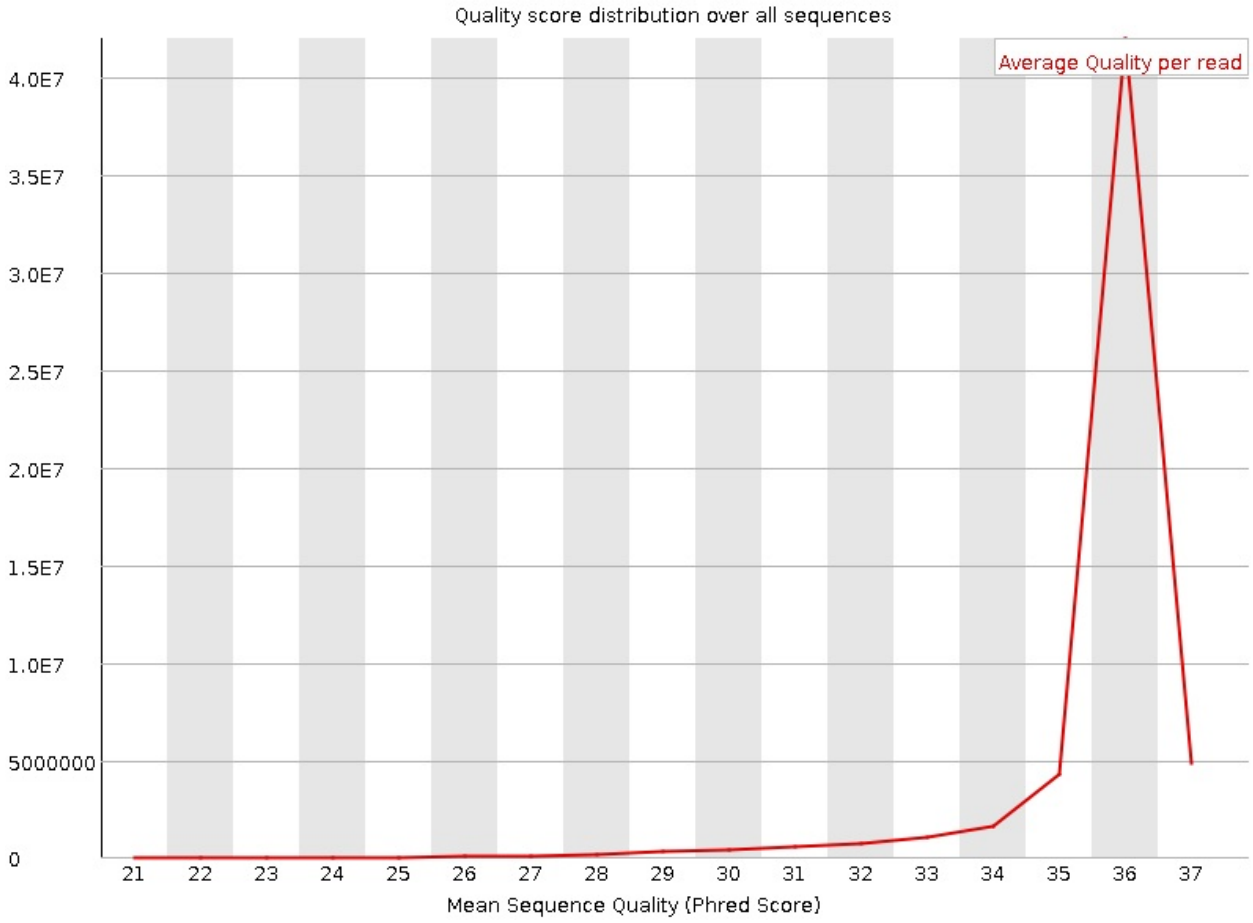
▼ Mut.HQ_R1



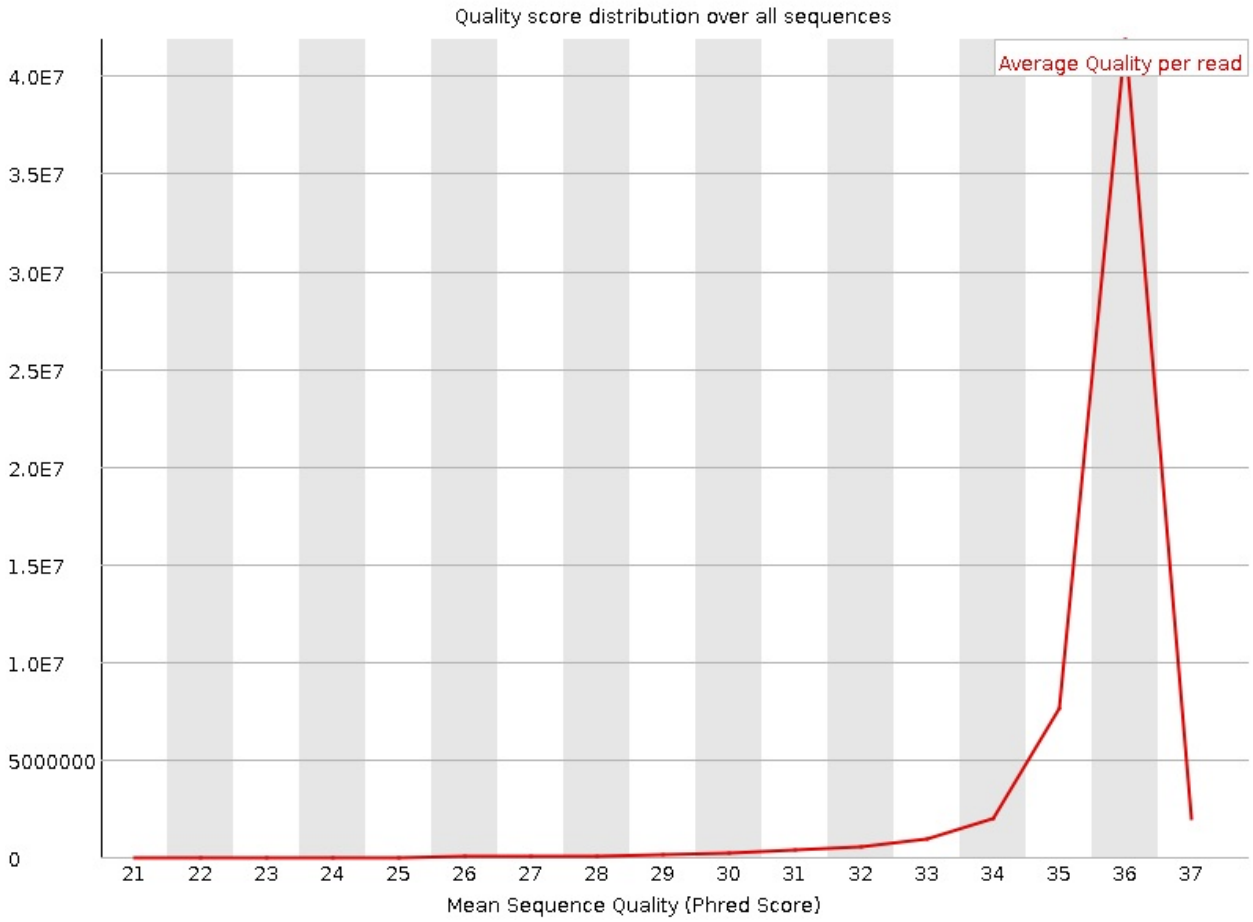
▼ Mut.HQ_R2


Quality score distribution over all sequences
Average Quality per read
4.0E7
3.5E7
3.0E7
2.5E7
2.0E7
1.5E7
1.0E7
5000000
0
20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
Mean Sequence Quality (Phred Score)

▼ WT.HQ_R1


Quality score distribution over all sequences
Average Quality per read
4.0E7
3.5E7
3.0E7
2.5E7
2.0E7
1.5E7
1.0E7
5000000
0
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
Mean Sequence Quality (Phred Score)

▼ WT.HQ_R2



碱基质量分布图，横坐标表示 reads 平均质量，纵坐标为对应平均质量值的 reads 数目。

2.4 高质量数据获取

测序数据包含一些带接头、低质量的 reads，这些序列会对后续的信息分析造成很大的干扰，为了保证后续信息分析质量，需要对下机数据进行进一步过滤，即将原始下机数据（raw data）过滤生成高质量序列（high quality data）。数据过滤的标准主要包括以下几点：

1. 接头污染去除，采用 AdapterRemoval（version 2）（Schubert M 等，2016）去除 3' 端的接头污染；
2. 质量过滤，采用滑动窗口法进行质量过滤，窗口大小设置为 5 bp，步长设置为 1 bp。每一次往前移动一个碱基，取 5 个碱基计算窗口的平均 Q 值，若最后一个碱基的 Q 值 ≤ 2，则仅保留该位置之前的碱基；若窗口的平均 Q 值 ≤ 20，则仅保留该窗口倒数第二个碱基及之前的碱基；
3. 长度过滤，若双末端中任意一条 reads 的长度 ≤ 50 bp，则去除该双末端 reads。数据过滤的基本情况见表4。

表4 数据过滤统计

Sample	HQ Reads (num)	HQ Reads(%)	HQ Data (bp)	HQ Data(%)
P50	73,236,566	99.59	10,978,717,732	99.53
Spring	71,641,752	99.55	10,739,461,034	99.49
WT	113,310,318	99.59	16,985,924,325	99.52
Mut	113,262,146	99.58	16,978,702,179	99.51

- Sample： 样品名；
- HQ Reads(num)： 高质量 reads 数；
- HQ Reads(%)： 高质量 reads 占下机 reads 的百分比；
- HQ Data(bp)： 高质量 reads 碱基数；
- HQ Data(%)： 高质量序列碱基占下机碱基的百分比。

三 对比分析

3.1参考基因组整理

下载参考基因组序列。参考基因组的统计结果如表5。

表5 数据过滤统计

类别	数目
Total_Len	130711987
Total_Seq_Num	28
Total_N_Counts	119362
Total_LowCase_Counts	0
Total_GC_content	0.38

- Total_Len: 基因组长度
- Total_Seq_Num: 基因组染色体数
- Total_N_Counts: 未测通的碱基数
- Total_LowCase_Counts: 重复序列的标志
- Total_GC_content: GC含量

3.2 序列对比

采用 bwa mem 程序将过滤后的高质量数据比对到参考基因组上，比对采用默认参数。采用 Picard 1.107 软件对 sam 文件进行排序并转换为 bam 文件。测序数据的生成过程涉及到文库的扩增和簇的形成，这两个步骤容易产生一些Duplicates（即 PCR Duplicates 和 Optical Duplicates），这些 Duplicates 不能作为变异检测的证据。采用 Picard 软件包中的 “MarkDuplicates” 去除 Duplicates，即如果多个 Paired Reads比对后具有相同的染色体坐标，则仅保留分值最高的 Paired Reads。InDel 附近的 reads 最容易出现 Mapping 错误，为了尽量减少由于 Mapping 错误导致的 SNP，需要对 InDel 附近的 reads 重新进行比对，以提高 SNP Calling 的准确性。

3.3 序列比对结果概述

序列比对结果统计见表6

表6 序列比对结果统计

Sample	Total reads	Mapped reads	Mapping rate	Dup.nums	Dup.rate
Mut	113,262,146	75,744,628	66.88%	8,484,558	7.49%
WT	113,310,318	74,852,064	66.06%	8,420,174	7.43%

- Sample： 样本名称；
- Total reads： 总 reads 数量；
- Mapped reads： 比对至参考基因组上的 reads 数量（包括单端比对和双端比对）；
- Mapping rate： 比对率，比对至参考基因组上的 reads 数量占总 reads 数量的百分比；
- Dup. nums： 重复序列 reads 数量；
- Dup. rate： 重复序列 reads 数量占总 reads 数量的百分比。

3.4 测序深度及覆盖度概述

分别统计每一个样品的平均覆盖深度和覆盖度，结果见表7。

表7 测序概述

Sample	Avg. Depth (x)	Coverage(≥4x)	Coverage(≥10x)	Coverage(≥30x)
Mut	34.64	98.08%	96.90%	65.29%
WT	34.58	98.06%	96.66%	65.21%

- Sample： 样品名称；
- Avg. Depth (×)：平均测序深度，比对至参考基因组的碱基总数除以基因组大小；
- Coverage(≥4x)：参考基因组中至少被 4 条序列覆盖的位点占基因的百分比；
- Coverage(≥10x)：参考基因组中至少被 10 条序列覆盖的位点占基因的百分比。
- Coverage(≥30x)：参考基因组中至少被 30 条序列覆盖的位点占基因的百分比。

四 SNP 分析

4.1 SNP 检测

SNP(单核苷酸多态性)主要是指在基因组水平上由单个核苷酸的变异所引起的 DNA 序 列多态性，包括单个碱基的转换、颠换等。采用 GATK 软件检测所有样本的 SNP，具体操 作步骤如下：

我们利用软件包GATK中的SelectVariants方法进行提取，为了保证 SNP 位点的可靠性，对获得的 SNP 位点进行过滤，过滤标准如下：

- QD < 2.0
- FS > 60.0
- SOR > 3.0
- MQRankSum < -12.5
- ReadPosRankSum < -8.0

> QD：variant的可靠度
>
> FS：Phred的概率分数
>
> SOR: 正义链还是反义链
>
> ReadPosRankSum：用来对比突变位点和原始位点是不是在reads的不同的位置
>
> MQRankSum：支持突变位点的reads和原始位点的reads的Mapping质量

4.2 SNP 注释

采用 snpEff软件对 SNP 位点进行注释。SNP 注释结 果见表8。

表8 SNP 注释结果统计

类型	Number	Precent(%)
DOWNSTEAM	296690	0.167
EXON	3626	0.167
INTERGENIC	1054586	48.462
INTRON	542416	24.926
TRANSCRIPT	202	0.009
SPLICE_SITE_ACCEPTOR	74	0.003
SPLICE_SITE_DONOR	159	0.007
SPLICE_SITE_REGION	3866	0.178
UPSTAREAM	265284	12.191
UTR_3	5991	0.275
UTR_5	3206	0.147

- DOWNSTEAM: 转录终止位点下游 1 kb 的区域
- EXON: 外显子区域；
- INTERGENIC: 基因间区域
- INTRON: 非编码区内含子区域
- TRANSCRIPT: 转录本
- SPLICE_SITE_ACCEPTOR: 剪切位点受体区
- SPLICE_SITE_DONOR: 剪切位点供体区
- SPLICE_SITE_REGION: 剪切位点内部区域
- UPSTAREAM: 转录起始位点上游 1 kb 的区域
- UTR 3': 3' UTR 区域；
- UTR 5': 5' UTR 区域

五 InDel分析

5.1 InDel 检测

InDel是小型的Insertion和Deletion的总称。我们采用GATK软件包中SelectVariants方法提取INDEL。

为了保证InDel结果的可靠性，进一步对InDel位点进行过滤，过滤标准如下：

- QD < 2.0
- FS > 200.0
- SOR > 10.0
- MQRankSum < -12.5
- ReadPosRankSum < -8.0

> QD：variant的可靠度
>
> FS：Phred的概率分数
>
> SOR: 正义链还是反义链
>
> ReadPosRankSum：用来对比突变位点和原始位点是不是在reads的不同的位置
>
> MQRankSum：支持突变位点的reads和原始位点的reads的Mapping质量

5.2 InDel 注释

采用 snpEff软件对 InDel 位点进行注释。InDel 的注 释结果见表9。

表9 InDel 注释结果统计

类型	Number	Precent(%)
DOWNSTEAM	1263790	12.927
EXON	181878	1.86
INTERGENIC	4777321	48.865
INTRON	2312338	23.65266
TRANSCRIPT	201	0.009
SPLICE_SITE_ACCEPTOR	129	0.001
SPLICE_SITE_DONOR	192	0.002
SPLICE_SITE_REGION	19211	0.196
UPSTAREAM	1179366	12.063
UTR_3	23361	0.239
UTR_5	19062	0.195

- DOWNSTEAM: 转录终止位点下游 1 kb 的区域
- EXON: 外显子区域；
- INTERGENIC: 基因间区域
- INTRON: 非编码区内含子区域
- TRANSCRIPT: 转录本
- SPLICE_SITE_ACCEPTOR: 剪切位点受体区
- SPLICE_SITE_DONOR: 剪切位点供体区
- SPLICE_SITE_REGION: 剪切位点内部区域
- UPSTAREAM: 转录起始位点上游 1 kb 的区域
- UTR 3': 3' UTR 区域；
- UTR 5': 5' UTR 区域

六 SNP index 分析

6.1 SNP index 计算

SNP index 的概念最早由日本学者 Takag 等人于 2013 年提出，主要用于筛选与性状基因相关联的候选区域。SNP index 的计算原理见图 5，基本过程描述如下：

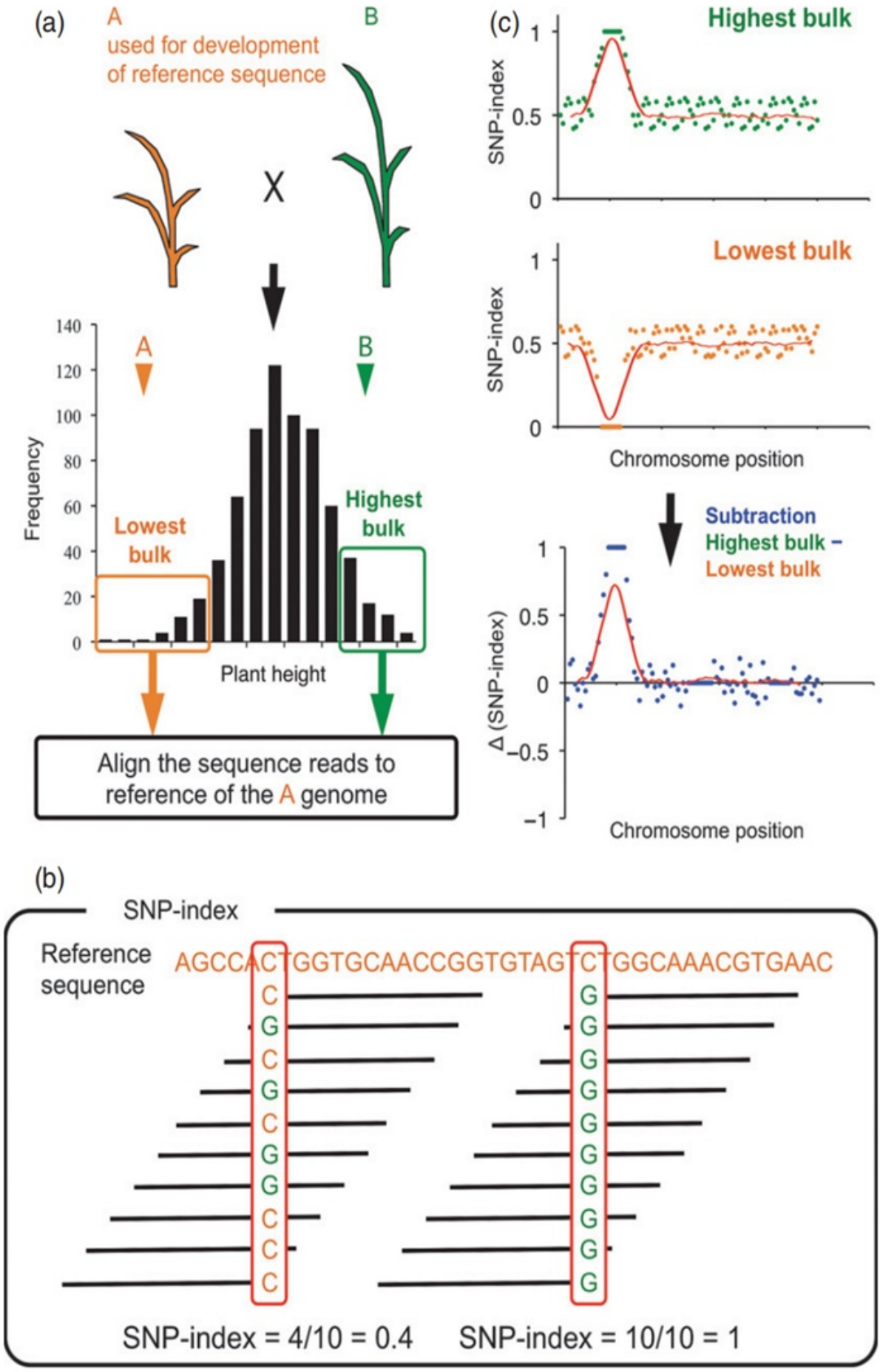

(a)
A
used for development
of reference sequence
B
X
A
B
Frequency
140
120
100
80
60
40
20
0
Lowest
bulk
Highest
bulk
Plant height
Align the sequence reads to
reference of the A genome
(c)
Highest bulk
SNP-index
1
0.5
0
Lowest bulk
SNP-index
Chromosome position
Subtraction
Highest bulk −
Lowest bulk
Δ (SNP-index)
1
0.5
0
−0.5
−1
Chromosome position
(b)
SNP-index
Reference
sequence
AGCCACTGGTGCAACCGGTGTAGTCTGGCAAACGTGAAC
C
G
C
G
C
G
G
C
C
C
G
G
G
G
G
G
G
G
G
G
SNP-index = 4/10 = 0.4
SNP-index = 10/10 = 1

图5 SNP index 原理图

1. 群体构建：具有显著表型差异的亲本 A 和 B 进行杂交，产生 F1 代，F1 代个体进行 自交，产生 F2 代个体；
2. 测序：分别对亲本 A、亲本 B 进行测序；取 F2 代个体具有极端表型的个体分别抽提 总 DNA，等量混合，分别对这两个子代池进行测序；
3. 基因型分析：将亲本 A、亲本 B、子代混池 1 和子代混池 2 比对至参考基因组上，筛选 SNP 多态性标记；
4. SNP Index 计算：假设子代混池 1 在某一个位点上测序深度为 50，其中，与亲本 A 完全相同的碱基测序深度为 30，与亲本 B 完全相同的碱基测序深度为 20，则子代池 1 的 SNP Index 为 0.6 。 分 别 计算每一个多态性位点的SNP index。

6.2 SNP Index 分布

为直观反映子代 SNP-index 在染色体上的分布情况，对 SNP-index 在染色体上的分布进行作图。图中黑点即为以窗口形式展现的 SNPindex 分布，以 1 Mb 大小为窗口， 500 kb 为步距，进行每个窗口中 SNP-index 的计算。以参考基因组的位置为横坐标，SNP index值为纵坐标作图，结果见图6-1 图6-2。

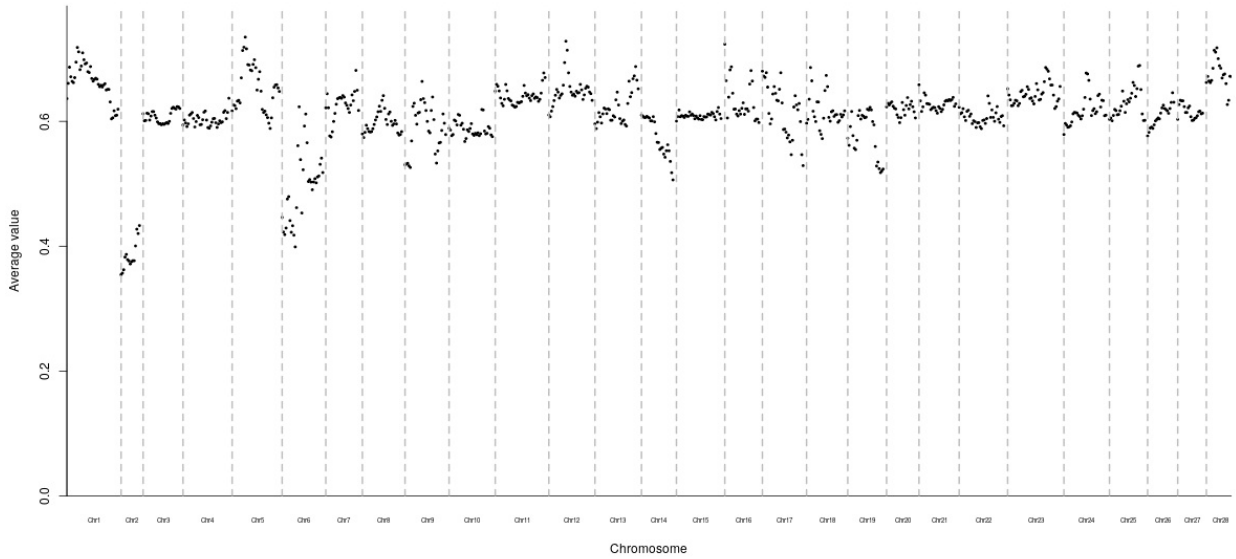


图6-1 子代混池(Mut) SNP index 在染色体上的分布

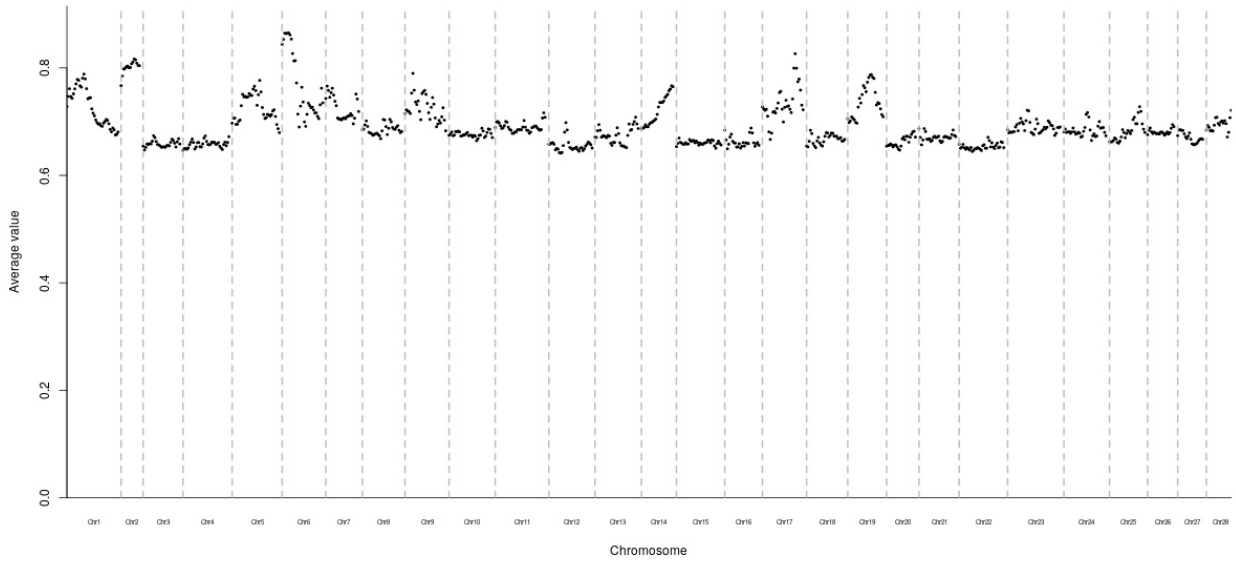


图6-2 子代混池(WT) SNP index 在染色体上的分布

横坐标为各条染色体的名称及其长度，纵坐标代表 SNP index 值。

6.3 SNP index 差异分布

计算两个子代混池的 SNP index 的差值，即得到 △(SNP index) = SNP index(极端性状 A) – SNP index（极端性状 B）。进行 10000 次置换检验，选取 95%（橙色）、99%（绿色） 置信区间作为筛选阈值，图中的黑点/黑色折线即为以窗口形式展现的 △(SNP-index)分布，置信区间 95%以上的窗口作为候选区间，△(SNP index) 在每一条染色体上的分布结果见图7-1 图7-2。

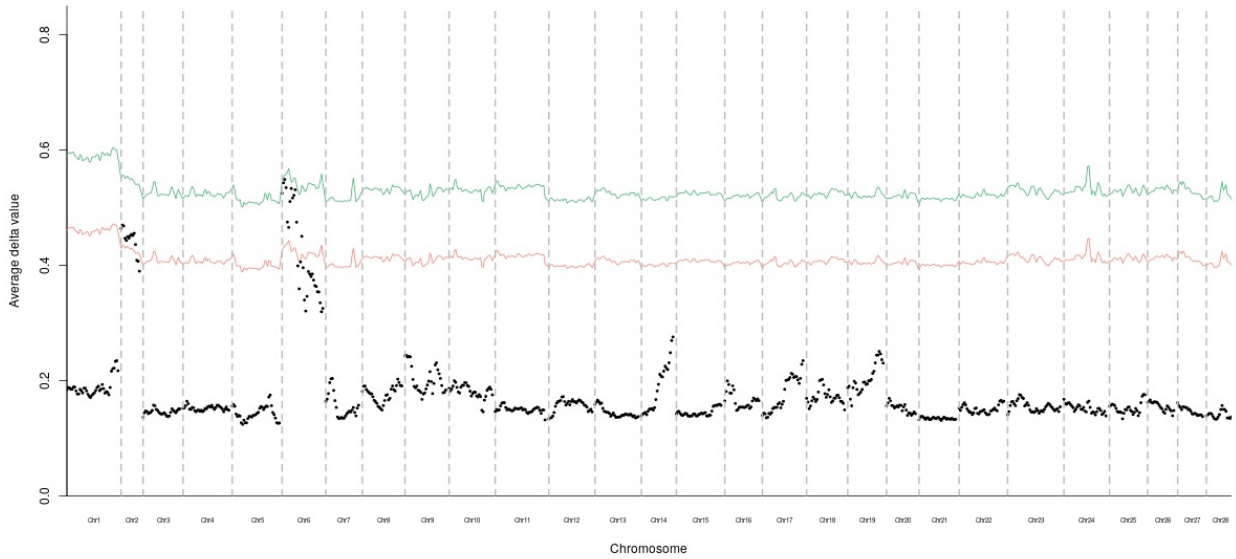


图7-1 △(SNP index) 在全基因组上的分布图

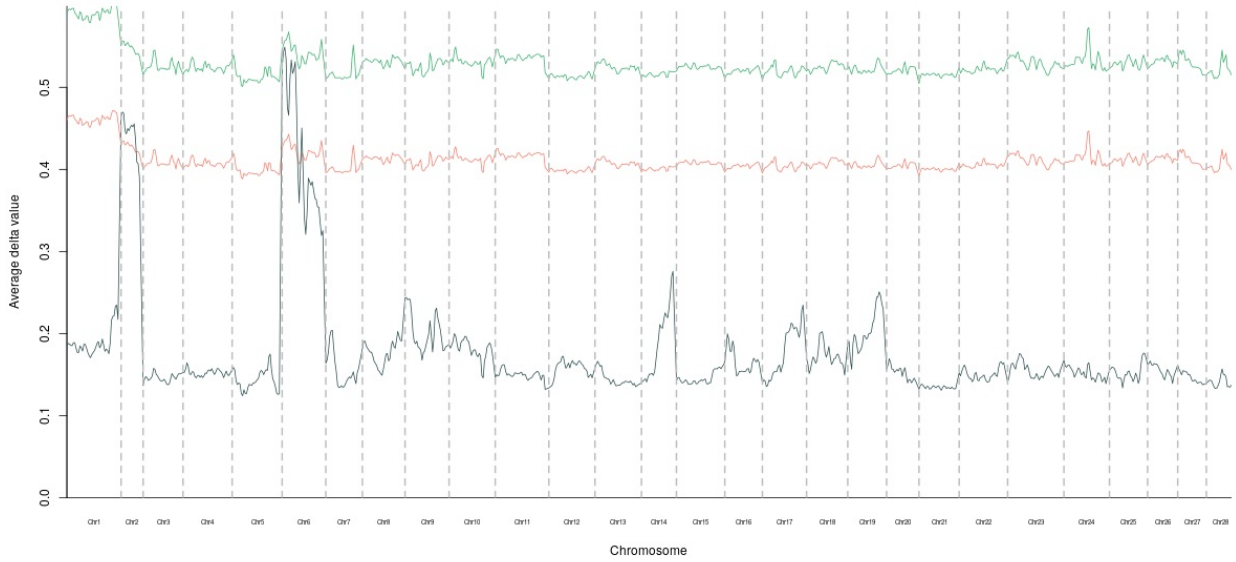


图7-2 △(SNP index) 在全基因组上的分布图

七 目标性状区域分析

△(SNP index)反映了两个混池之间基因型频率的显著差异。在进行测序前，由于对所有混池样本的性状进行选择，选取两个极端性状的混池分别进行测序，即△(SNP index)越接近于1，则该位点的 SNP 与目标性状的连锁强度越高，与目标性状的关联度越高；反之，则该位点的 SNP 与目标性状的连锁强度越低，与目标性状的关联度越低。选取阈值线以上的区域作为性状相关的候选区域，总共筛选到2个区域，总长度为15538441bp。关联区域的结果统计见表10。

表**10** 关联区域统计

Chr	Start	End	Size (bp)
2	595	8395003	8394408
6	263	7144296	7144033

- Chr： 染色体编号；
- Start： 起始位置；
- End： 终止位置；
- Size(bp)：区域大小。

八 参考文献

[1]McKenna A, Hanna M, Banks E. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 2010, 20:1297-303.

[2] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data, Nucleic Acids Research 2010, 38:e164.

[3] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols 2009, 4(7):1073-81.

[4] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. Nature Methods 2010, 7(4):248-249.

[5] Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014, 11(4):361-2.

[6] Hiroki Takagi, Akira Abe, Kentaro Yoshida. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. The Plant Journal 2013, 74: 174–183.

[7] Akira Abe, Shunichi Kosugi, Kentaro Yoshida. Genome sequencing reveals agronomically important loci in rice using MutMap. Nature Biotechnology 2012, 30(2):17.