Variation Detection Project (GBS, without-reference)

Demo Report

May 1, 2016

Contents

1 Project Background

Species name: XXX;
Number of samples: 2;
Sequencing strategy: Illumina HiSeq PE150;
Analysis content: Sequencing quality control, GBS tag output statistics, sequence alignment, SNP detection and annotation.
About GBS: GBS technology refers to Genotyping By Sequencing, which can be used for development of molecular markers, ultra-high density genetic map construction, population genetic analysis, GWAS and other fields.

2 Experimental Procedures

2.1 DNA Quantification and Qualification

1st BASE utilizes three major QC methods for DNA sample qualification:

(1) Agarose gel electrophoresis analysis for DNA purity and integrity;
(2) NanoDrop® 2000 spectrophotometer measurement for DNA purity by assessing the OD_{260}/OD_{280} ratio;
(3) Qubit® 2.0 flurometer quantitation for accurate measurement of DNA concentration;

Sample DNA, with OD_{260}/OD_{280} ratio of 1.8 to 2.0 and total amount of more than 0.6 μg, was qualified for library construction.

2.2 Library Construction

The genomic DNA of samples was respectively digested using the restriction enzymes, and the obtained fragments were ligated with barcodes, and then they were amplified by PCR. Subsequently, the samples were pooled and selected for the required fragments for library construction. To check the prepared DNA libraries, Qubit® 2.0 fluorometer was firstly used to determine the concentration of the library. After dilution to 1 ng/μl, the Agilent® 2100 bioanalyzer was used to assess the insert size. And finally the quantitative real-time PCR (qPCR) was performed to detect the effective concentration of each library. If the library with appropriate insert size has an effective concentration of more than 2 nM, the constructed libraries are qualified and ready for Illumina® high-throughput sequencing. The experimental procedures of DNA library preparation are shown in **Figure 2.1**.

(1) Restriction enzyme digestion: 0.3~0.6 μg genomic DNA was digested with the restriction enzyme in order to obtain a suitable marker density;
(2) Ligating P1 and P2 adapter: each end of digested fragment was respectively ligated with P1 and P2 adapter (complementarily with digested DNA overhang);
(3) Fragment selection: tags containing both P1 and P2 adapters were amplified through PCR. Then DNA fragments of different samples were pooled, and the desired fragments of DNA were recovered after electrophoresis;

(4) High-throughput sequencing: Cluster preparation, and then sequencing.

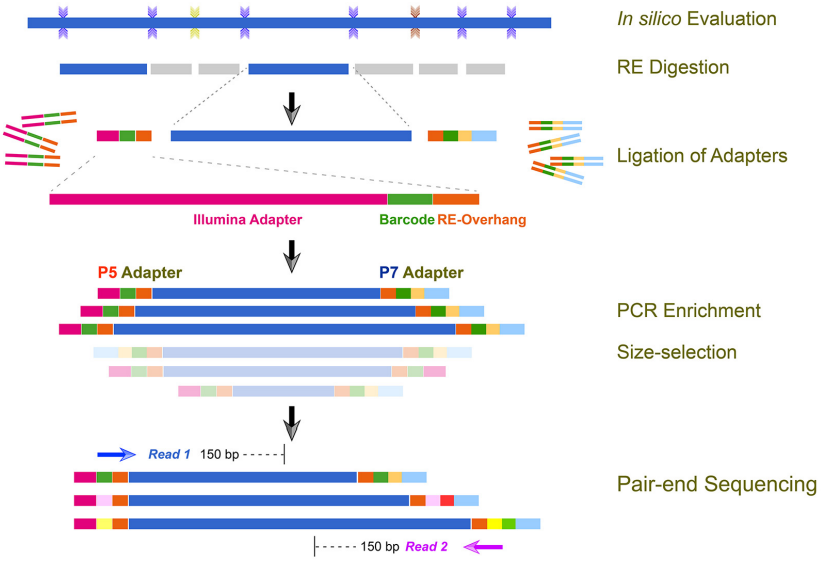

Figure 2.1 Experimental procedures of library preparation in GBS

2.3 High-throughput DNA Sequencing

Pair-end sequencing were performed on Illumina® HiSeq platform, with the read length of 150 bp at each end.

3 Bioinformatics Analysis Procedures

The bioinformatics analysis procedures are as follows:

(1) Quality control of raw sequencing data for clean data filtration;
(2) Mapping clean reads to reference genome;
(3) SNP and InDel detection and annotation according to the reference genome mapping results.

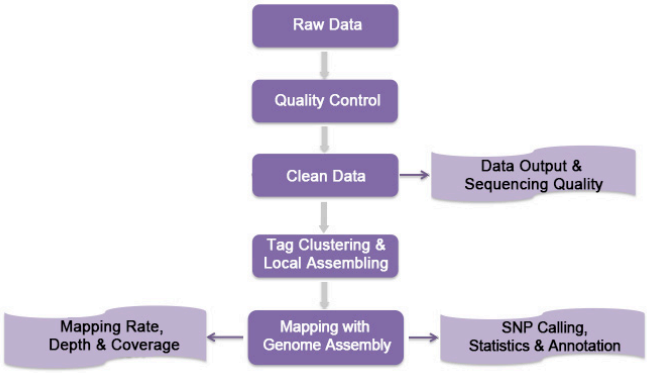

Figure 3.1 Bioinformatics analysis workflow

4 Results of Analyses

4.1 Raw Data

The original sequencing data acquired by high-throughput sequencing platforms (e.g. Illumina HiSeq™ /Miseq™) recorded in image files are firstly transformed to sequence reads by base calling with the CASAVA software. The sequences and corresponding sequencing quality information are stored in a FASTQ file.

Every read in FASTQ format is stored in four lines as follows:

```
@K00124:82:H2MH5BBXX:1:1101:31389:1158 2:N:0:0
TAGCCACATAGAAACCAACAGCCATATAACTGGTAGCTTTAAGCGGCTCACCTTTAGCATCAACAGGCCAC
AACCAACCAGAACGTGAAAAAGCGTCCTGCGTGTAGCGAACTGCGATGGGCATACAGATCGGAAGAGCGTC
GTGTAGGG
+
AAFFFKKKKKKKKFKKKFFKKAAFKKKKKFKKKKFKKA,FKKKKKKKKKAKKFKKKKKKKAKKKKKKFFKK
KKF<FFKKKKKKKKKKKKFKKFKKF7FFFFFKFKKKFKKKKKKKKF<FFKKKKFKKKKKFKFKFKKFK<<
F,A7,AFK
```

Line 1 begins with an '@' character and is followed by Illumina sequence identifiers, and an optional description (such as a FASTA title line).

Line 2 is the sequence of a sequencing read.

Line 3 begins with a '+' character and is optionally followed by Illumina sequence identifier and description.

Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of characters as the bases in the sequence. The per base sequencing quality score could be calculated by the ASCII value of each character in Line 4 minus a constant 33.

Table 4.1 Information of Illumina sequence identifiers

Identifier	Meaning
K00124	Unique instrument name
82	Run ID
H2MH5BBXX	Flowcell ID
1	Flowcell lane number
1101	Tile number within the flowcell lane
31389	'x'-coordinate of the cluster within the tile
1158	'y'-coordinate of the cluster within the tile
2	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
N	Y if the read fails filter (read is bad), N otherwise
0	0 when none of the control bits are on, otherwise it is an even number
ATCACG	Index sequence

4.2 Quality Control of Sequencing Data

4.2.1 Sequencing Quality Distribution

If the sequencing error rate is represented by *e*, and Illumina HiSeq™ /MiSeq™ sequencing quality by Q_{Phred}, the quality score of a base (Phred score) is calculated by the following equation: $Q_{Phred}=-10\log_{10}(e)$. The correspondence relationship between Illunima sequencing quality and Phred score in base calling by Casava version 1.8 is listed as follows:

Table 4.2 Relationship between Illunima sequencing quality and Phred score

Phred Score	Error Rate	Correct Rate	Q-score
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

For next-generation sequencing (NGS), the sequencing platform, chemical reactants, and sample quality can influence sequencing quality and base error rate. Sequencing quality distribution is examined over the full length of all sequences, to detect any sites (base positions) with an unusually low sequencing quality, where incorrect bases may be incorporated at abnormally high levels. For detailed sequencing quality distribution, please refer to Figure 4.2.

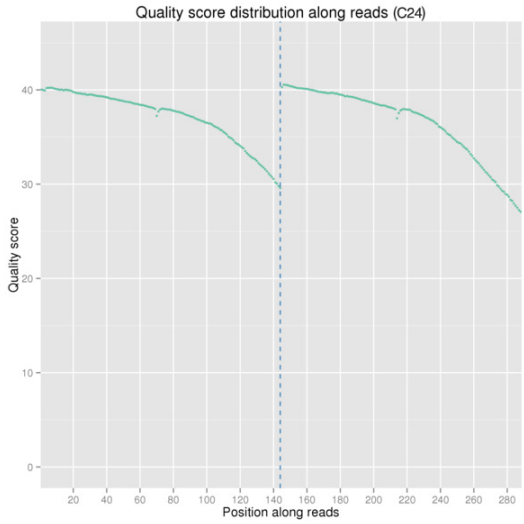

Figure 4.2.1 Distribution of sequencing quality

The x-axis shows the base position within a sequencing read, and the y-axis shows the average phred score of all reads at each position.

(Pair-end sequencing data are plotted together, with the first 150 bp representing read 1 and the following 150 bp for read 2.)

4.2.2 Distribution of Sequencing Errors

Sequencing error rate is related to the base quality of the obtained sequence. The sequencing platform, chemical reactants, and sample quality can all influence sequencing error rate and herein the base quality. For next-generation sequencing (NGS) with sequencing-by-synthesis

strategy, sequencing error rate distribution shows two common features:

(1) Error rate increases with extending of the sequencing reads due to the consumption of chemical reagents, damage of the DNA template by laser irradiation, and possible accumulation of errors during the sequencing cycles. All the Illumina high-throughput sequencing platforms have this feature.

(2) The sequencing error rate is higher for the first several bases than at other positions, which is likely the result of reading errors during the first few cycles after calibration of the optical instruments.

Sequencing error rate distribution is examined over the full length of all sequences, to detect any sites (base positions) with an unusually high error rate, where incorrect bases may be incorporated at abnormally high levels. For detailed sequencing error distribution, please refer to **Figure 4.2.2**.

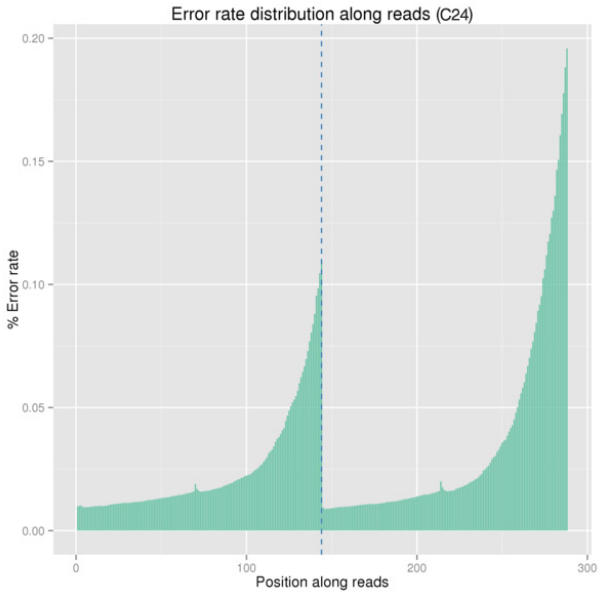

Figure 4.2.2 Distribution of sequencing errors.

The x-axis shows the base position within a sequencing read, and the y-axis shows the average error rate of all reads at each position.

(Pair-end sequencing data are plotted together, with the first 150 bp representing read 1 and the following 150 bp for read 2.).

4.2.3 Sequencing Data Filtration

Raw data obtained from sequencing contains adapter contamination and low-quality reads. These sequencing artifacts may increase the complexity of downstream analyses, and therefore, we utilize quality control steps to remove them. Consequently, all the downstream analyses are based on the clean reads.

The quality control steps are as follows:

(1) Discard the paired reads when either read contains adapter contamination;

(2) Discard the paired reads when uncertain nucleotides (N) constitute more than 10 percent of either read;

(3) Discard the paired reads when low quality nucleotides (base quality less than 5, $Q \leq 5$)

constitute more than 50 percent of either read.

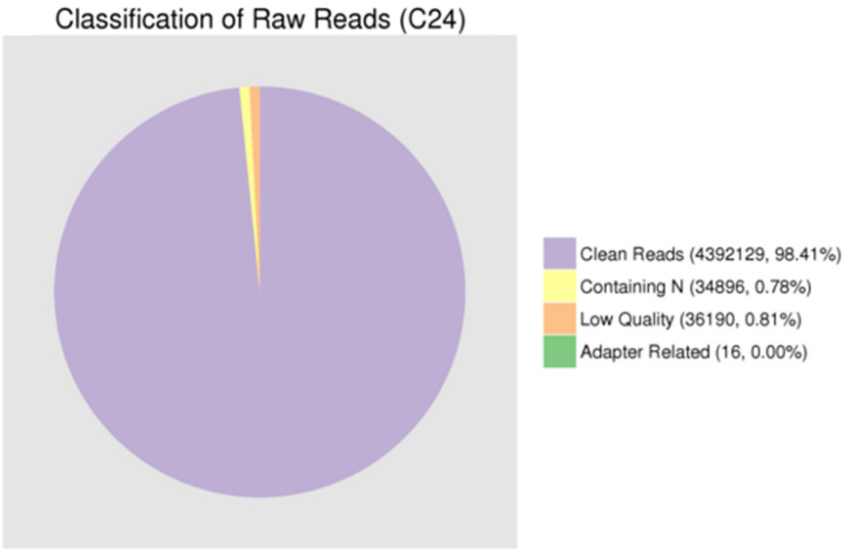

Figure 4.2.3 Classification of the sequenced reads

(1) Adapter related: The proportion of filtered reads containing adapters in total reads. (2) Containing N: The proportion of filtered reads containing more than 10% Ns in total reads. (3) Low quality: The proportion of filtered reads for low quality in total reads. (4) Clean reads: The proportion of clean reads in raw reads.

4.2.4 Statistics of Sequencing Data

Consistent with the Illumina platform sequencing features, for PE data, the error rate should be below 0.1%. The results are shown in **Table 4.3**.

Table 4.3 Statistics of sequencing data

Sample	Raw Base (bp)	Clean Base (bp)	Effective Rate (%)	Error rate (%)	Q20 (%)	Q30 (%)	GC Content (%)
C24	554,557,824	554,446,368	99.98	0.04	94.70	87.52	44.85
C25	495,188,352	495,088,128	99.98	0.03	95.43	89.57	45.10

The details for the sequencing data statistics are as follows:

(1) Sample: Sample name.

(2) Raw Base (bp): The output of raw data calculated by the number and length of sequence (in bp).

(3) Clean Base (bp): The valid data output of sequence (in bp) after filtering low quality reads, calculated by the number and length of sequences in clean data.

(4) Effective Rate (%): The ratio of clean data to raw data.

(5) Error Rate (%): Overall error rate of base.

(6) Q20 and Q30 (%): The percentage of bases with higher Phred score than 20 and 30 in total bases.

(7) GC Content (%): The percentage of G and C in total bases.

4.2.5 Restriction Digestion Statistics

According to the experimental procedures, the reduction representative of the genome was achieved by digestion with sets of restriction enzymes. After the primary restriction digestion, the second set of enzymes were hired to further reduce the genome fragments. The restriction enzyme captured reads, which contains the restriction site of the primary digestion enzyme at each end and without recognition sites of the secondary restriction enzymes, from the filtered clean data were counted to calculate the digestion ratio. The statistics of the enzyme-captured PE reads were shown in **Table 4.4**.

Table 4.4 Statistics of enzyme captured reads

Sample	Total Clean PE Reads	1st Enzyme-captured PE Reads	Completely-digested PE Reads	Capture Ratio (%)	Complete Ratio (%)
C24	1,719,056	1,719,056	1,719,056	100.00	100.00
C25	1,925,161	1,925,161	1,925,161	100.00	100.00

The details for the sequencing data statistics are as follows:

(1) Sample: Sample name.

(2) Total Clean PE Reads: The number of paired reads in clean data.

(3) 1st Enzyme-captured PE Reads: The number of paired reads with restriction site of the primary restriction enzyme at end.

(4) Completely-digested PE Reads: The number of paired reads with the primary restriction site at its end and without the secondary recognition sites on its sequence, which indicates this tag is completely digested.

(5) Capture Ratio (%): The ratio of 1st enzyme captured PE reads to total clean RE reads.

(6) Complete Ratio (%): The percentage of clean PE reads completely digested by both the primary and secondary restriction enzymes.

4.2.6 Statistics of GBS-tag Output

The clean PE reads were clustered into unique tag clusters by Stacks with allowing a maximum 6 bp mismatch. The statistics of tags are shown in **Table 4.5**.

Table 4.5 Statistics of GBS-tags

Sample	Total Clean PE Reads	Total Tag Number
C24	1,719,056	176,737
C25	1,925,161	230,259

The details for the sequencing data statistics are as follows:

(1) Sample: Sample name.

(2) Total Clean PE Reads: The number of paired reads in clean data.

(3) Total Tag Number: The number of clustered unique GBS-tags.

4.2.7 Sequencing Evaluation Summary

Totally 1.050G raw data were sequenced from this run, with 1.049G clean data generated after filtering low-quality data. The raw data production for each sample ranged from 495.188 M to 554.558 M, indicating the sufficient amount of data production. As the Q20 and Q30 reached 94.70% and 87.52%, respectively, the sequencing quality could meet the proper analysis requirements. The GC content of 44.85% to 45.10% are also in the normal distribution range, fulfilling the quality standard.

4.3 Clustering and Local Assembling

4.3.1 Statistics of Clustering and Assembly

The acquired clusters from a parent line were locally assembled according to overlapped sequences. As shown in **Table 4.6**, the similarity in GC content with the sequenced data indicates the representation of the genome.

Table 4.6 Assembling statistics

Sample	Clustered Tag Number	Length (bp)	GC Content (%)
C24	319,269	91,949,472	40.3

(1) Sample: Sample name.

(2) Clustered Tag Number: The number of clustered reads from clean data.

(3) Length (bp): The length of the genome assembly.

(4) GC Content (%):The percentage of G and C in total bases.in the genome assembly.

4.3.2 Evaluation of Clustering and Assembly

The non-duplication reads were mapped to the local assembly sequences, and variations were detected subsequently. The assembly result is illustrated reliable if most heterozygous SNPs and small amount of homozygous SNPs are detected. The mapping and variation detection procedures were conducted as follows:
(1) Parameters for reads aligning with the assembled genome by BWA: 'mem -t 4 -k 32 -M'.
(2) Use 'samtools -bs' to convert sam to bam files, and use 'samtools sort' to sort those bam files.
(3) Use 'samtools rmdup' to remove duplication of the aligned reads (reads aligned to multiple locations).
(4) Use 'samtools mpileup -m 2 -F 0.002 -d 1000' to do SNP calling.
The statistics of assembly mapped reads and detected SNPs are listed in **Table 4.7**.

Table 4.7 Statistics of mapping and SNP calling

Sample	Mapping rate (%)	Average depth	Coverage (%)	Coverage (at least 4X)(%)	SNP number	Het SNP	Het rate (%)
C24	85.03	4.81	90.19	46.04	41,578	34,696	83.45

(1) Sample: Sample name.

(2) Mapping rate (%): The rate of mapped reads to total reads.

(3) Average depth: The average sequencing depth at each site, which was the rate of reads number and genome size.

(4) Coverage (%): The percentage of the assembled genome with more than one read at each site.

(5) Coverage (at least 4X)(%): The percentage of the assembled genome with ⩾4X coverage at each site.

(6) SNP number: The number of SNPs detected with using the assembly sequence as the reference.

(7) Het SNP: The number of heterozygous SNPs.

(8) Het rate (%): The rate of heterozygous SNP in all SNPs.

4.4 Mapping Statistics

The clean reads were mapped to the genome assembly using BWA aligner, with the parameter settings as 'mem -t 4 -k 32 -M'. The aligned reads were further filtered with 'SAMtools rmdup' to remove duplication.

4.4.1 Mapping Statistics

The mapping rates of samples reflect the similarity between each sample and the reference genome. The depth and coverage are indicators of the evenness and homology with the reference genome. The effective sequencing data was aligned with the reference sequence through BWA[1] software (parameters: mem -t 4 -k 32 -M), and the mapping rate and coverage was counted according to the alignment results (see **Table 4.8**). The duplicates were removed by SAMTOOLS[2] (parameters: rmdup).

Table 4.8 Statistics of sequencing depth and coverage

Sample	Mapped reads	Total reads	Mapping rate (%)	Average depth (X)	Coverage (at least 1X)(%)	Coverage (at least 4X)(%)
C24	67,686,930	69,676,436	97.14	25.08	98.77	95.24
C25	33,715,993	34,903,794	96.60	12.77	97.33	94.11

The details for mapping statistics are as follows:

(1) Sample: Sample names.

(2) Mapped reads: The number of clean reads mapped to the reference assembly, including both single-end reads and reads in pairs.

(3) Total reads: Total number of effective reads in clean data.

(4) Mapping rate: The ratio of the reference genome assembly mapped reads to the total sequenced clean reads.

(5) Average depth: The average depth of mapped reads at each site, calculated by the total number of bases in the mapped reads dividing by size of the assembled genome.

(6) Coverage at least 1X: The percentage of the assembled genome with more than one read at each site.

(7) Coverage at least 4X: The percentage of the assembled genome with ≥4X coverage at each site.

4.4.2 Mapping Summary

For the current 91,949,472bp reference genome assembly, the mapping rate of each sample ranges from 94.11% to 95.24%. The average depth on the reference genome (without Ns) is in 12.77X to 25.08X range, while the more than 1X coverage exceeds 97.33%. This result is in the qualified normal range and may serve in the subsequent variation detection and related analyses.

4.5 SNP Detection and Annotation

Single nucleotide polymorphism (SNP) refers to a variation in a single nucleotide which may occur at some specific position in the genome, including transition and transversion of a single nucleotide. We detected the individual SNP variations using SAMTOOLS[2] with the following parameter: 'mpileup -m 2 -F 0.002 -d 1000'.

To reduce the error rate in SNP detection, we filtered the results with the criterion as follows:

(1) The number of support reads for each SNP should be more than 4 and less than 1000;

(2) The mapping quality (MQ) of each SNP should be higher than 20;

4.5.1 Statistics of SNP Variation

SNP calling results are listed in **Table 4.9**.

Table 4.9 Statistics of SNP variation

Sample	Total SNP Number	Hom SNP	Het SNP	Het rate (%)
C25	813,656	746,794	66,862	83.45

(1) Sample: Sample name.

(2) Total SNP Number: The number of SNPs detected.

(3) Hom SNP: The number of homozygous SNPs.

(4) Het SNP: The number of heterozygous SNPs.

(5) Het rate (%): The ratio of heterozygous SNP to all SNPs.

4.5.2 SNP Quality Distribution

To assess the credibility of detected SNPs, we checked the distribution of support reads number, SNP quality, as well as the distance between adjacent SNPs. The results are shown in **Figure 4.5.2**.

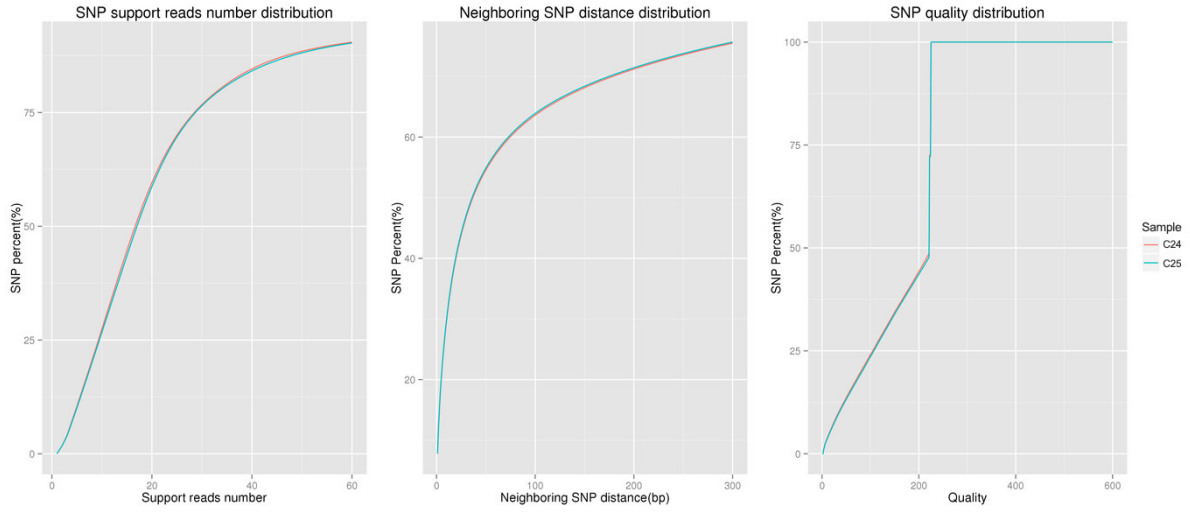

Figure 4.5.2 Cumulative distribution of SNP quality

These figures show the quality distribution of SNPs by, from left to right, the distribution of SNP support reads number, the distribution of distances between adjacent SNPs, and the cumulative distribution of SNP quality.

4.5.3 SNP Mutation Frequency

Take the T:A>C:G mutations as an example, this category includes mutations from T to C and A to G. When T>C mutation appears on either of the double-strand, the A>G mutation will be found in the same position of the other chain. Therefore the T>C and A>G mutations are classified into one category. Accordingly, the whole-genome SNP mutations could be classified

into six categories. The frequency of each type is shown in **Figure 4.5.3**.

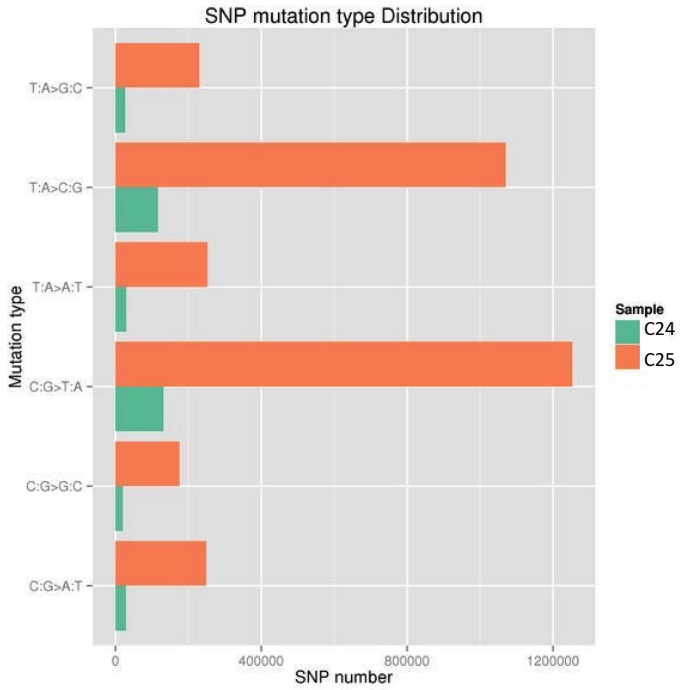

Figure 4.5.3 Frequency of SNP mutations

The x-axis represents the number of the SNPs, and y-axis indicates the mutation types.

4.6 Genotyping

The high-quality SNP makers may serve the subsequent genotyping within populations and accelerate the gene function studies.

Table 4.10 Genotyping Demo with SNP markers

Chromosome	Position	Ref	Sample 1	Sample 2
CM3.5.1_scaffold01596	8839	G	GG	AA
CM3.5.1_scaffold01596	24977	G	GG	CC
CM3.5.1_scaffold01596	25981	A	AA	TT
CM3.5.1_scaffold01596	29104	G	GG	AA

(1) Chromosome: The chromosome id or assembled contig names.

(2) Position: The chromosome location of each SNP variation.

(3) Ref: The sequence of the reference genome at the SNP site.

(4) Sample 1: The haplotype of sample 1 at the SNP site.

(5) Sample 2: The haplotype of sample 2 at the SNP site..

5 References

[1] Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25(14):1754-1760.

[2] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009, 25(16):2078-2079.