Variation Detection Project (GBS)

Demo Report

May 1, 2016

Contents

1 Project Background

Species name: XXX;
Number of samples: 30;
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.

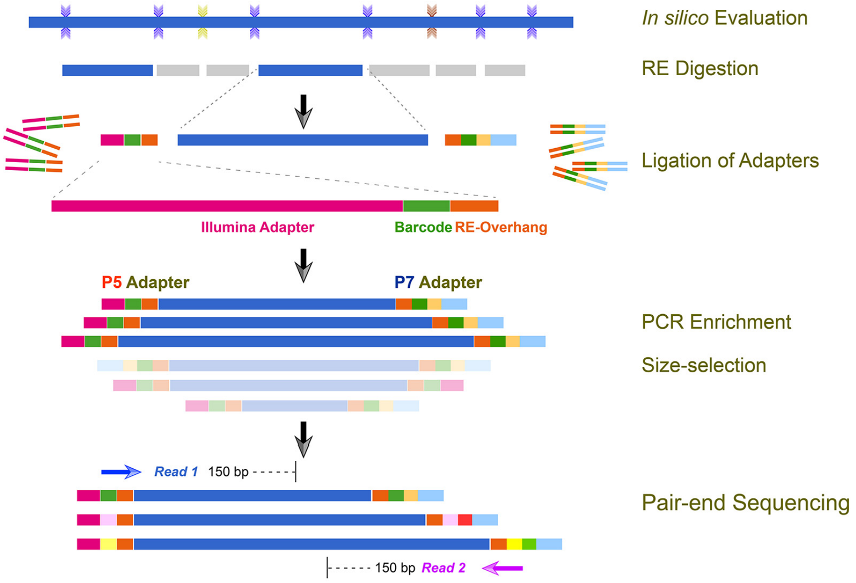

Figure 2.1 Experimental procedures of library preparation in GBS

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.

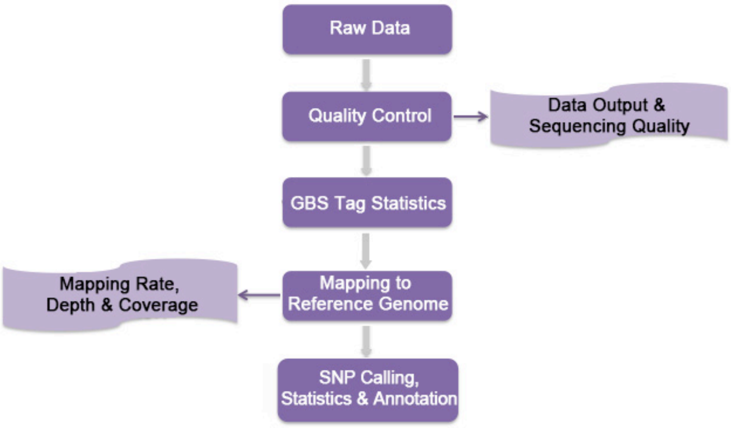
Raw Data
Quality Control
Data Output & Sequencing Quality
GBS Tag Statistics
Mapping Rate, Depth & Coverage
Mapping to Reference Genome
SNP Calling, Statistics & Annotation

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^{TM} /Miseq^{TM}) 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
TAGCCACATAGAAACCAACAGCCATATAACTGGTAGCTTTAAGCGGCTCACCTTTAGCATCAACAGGCCACAACCAA
CCAGAACGTGAAAAAGCGTCCTGCGTGTAGCGAACTGCGATGGGCATACAGATCGGAAGAGCGTCGTGTAGGG
+
AAFFFKKKKKKKKFKKKFFKKAAFKKKKKFKKKKFKKA,FKKKKKKKKKAKKFKKKKKKKAKKKKKKFFKKKKF<FF
KKKKKKKKKKKKKFKKFKKF7FFFFFKFKKKFKKKKKKKKF<FFKKKKFKKKKKFKFKFKKKFK<<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^{TM} /MiSeq^{TM} sequencing quality by Q_{Phred}, the quality score of a base (Phred score) is calculated by the following equation: Q_{Phred}=-10log$_{10}$(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.

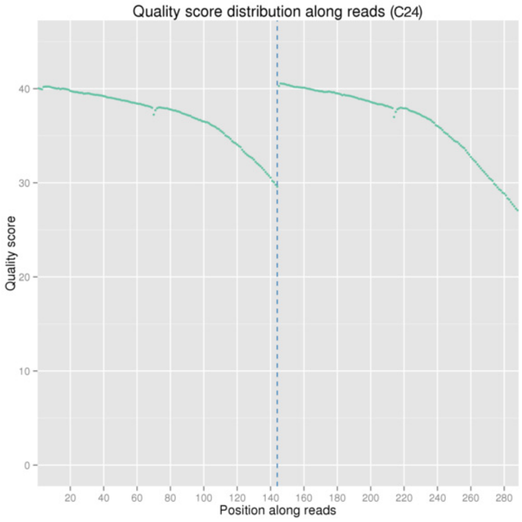

Figure 4.2.1 Distribution of sequencing quality

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**.

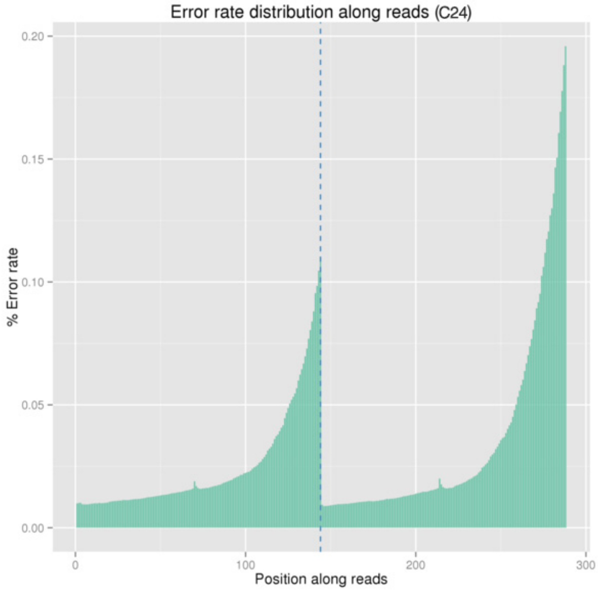

Figure 4.2.2 Distribution of sequencing errors.

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.

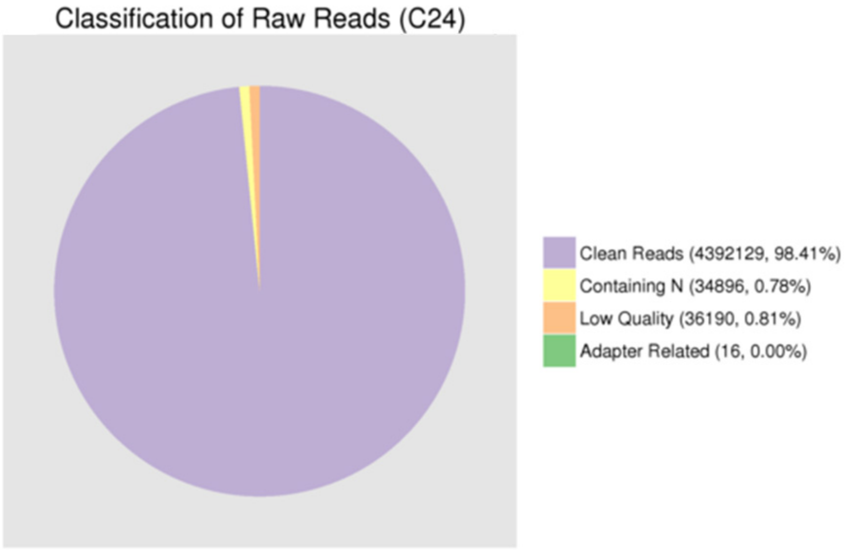

Figure 4.2.3 Classification of the sequenced reads

(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 (%)
C1	1,062,248,978	1,045,326,702	98.41	0.04	94.73	90.44	38.97
C2	796,074,776	782,980,254	98.36	0.04	94.20	89.57	39.28
C3	833,538,594	820,714,202	98.46	0.04	94.38	89.84	39.09
C4	773,327,450	757,772,960	97.99	0.04	93.85	88.93	38.31
C5	708,937,026	697,284,784	98.36	0.04	94.38	89.82	39.03
C6	775,992,574	762,224,036	98.23	0.04	93.55	88.35	39.07
C7	703,972,822	691,687,738	98.25	0.04	93.42	88.21	38.45
C8	793,588,628	779,681,098	98.25	0.04	93.76	88.66	38.53
C9	782,836,502	767,277,252	98.01	0.04	93.66	88.57	38.54
C10	776,462,386	763,480,914	98.33	0.04	94.31	89.62	39.14
C11	784,464,660	770,068,516	98.16	0.04	93.94	89.06	38.82
C12	709,337,580	696,468,206	98.19	0.04	92.91	87.27	38.96
C13	786,265,606	772,883,342	98.30	0.04	94.26	89.66	38.86
C14	858,298,210	844,453,036	98.39	0.04	94.27	89.56	39.30

C15	654,968,384	643,349,462	98.23	0.04	94.53	90.13	38.20
C16	719,231,716	706,093,640	98.17	0.04	93.44	88.16	38.70
C17	469,463,092	433,131,202	92.26	0.05	92.13	85.00	38.22
C18	707,965,748	694,681,540	98.12	0.04	92.56	86.49	39.09
C19	614,542,894	603,029,882	98.13	0.04	93.65	88.59	38.47
C20	733,854,198	722,237,656	98.42	0.04	93.62	88.49	38.98
C21	843,626,462	828,022,230	98.15	0.04	93.27	87.90	38.87
C22	968,627,394	953,875,202	98.48	0.04	94.84	90.56	39.21
C23	822,338,314	809,721,458	98.47	0.04	94.54	90.09	38.89
C24	843,473,428	828,735,992	98.25	0.04	93.02	87.35	39.08
C25	857,744,146	843,913,014	98.39	0.04	94.67	90.26	39.05
C26	634,215,736	623,181,818	98.26	0.04	93.34	87.88	39.02
C27	788,233,152	775,443,508	98.38	0.04	94.25	89.58	39.22
C28	820,986,474	807,637,768	98.37	0.04	93.61	88.40	39.11
C29	814,791,096	800,727,438	98.27	0.04	94.66	90.22	38.96
C30	875,069,118	860,226,962	98.30	0.04	93.77	88.75	39.10

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 Sequencing Evaluation Summary

Totally 23.314G raw data were sequenced from this run, with 22.886G clean data generated after filtering low-quality data. The raw data production for each sample ranged from 469.463 M to 1,062.249 M, indicating the sufficient amount of data production. As the Q20 and Q30 reached 92.13% and 85.0%, respectively, the sequencing quality could meet the proper analysis requirements. The GC content of 38.2% to 39.3% are also in the normal distribution range, fulfilling the quality standard.

4.3 Mapping Statistics

4.3.1 Statistics of Reference Genome

Reference genome is downloaded from: ftp://ftp.ensemblgenomes.org/pub/release-84/plants/fasta/xxxxxx/dna. The statistics of reference genome are listed in **Table 4.4.**

Table 4.4 The statistics of reference genome

Seq number	Total length	GC content (%)	Gap rate (%)	N50 length	N90 length
xx	xxxxxx	34.80	15.78	61,165,649	48,614,681

(1) Seq number: the total number of the assembled genomic sequences.

(2) Total length: the total length of the assembled genomic sequence.

(3) GC content: the GC content of the reference genome.

(4) Gap rate: the proportion of unknown sequence (N) in the reference genome assembly.

(5) N50 length: the length of scaffold N50, of which 50% of the sequence is higher than this level.

(6) N90 length: the length of scaffold N90, of which 90% of sequence is higher than this level.

4.3.2 Mapping Statistics with Reference Genome and Tag Summary

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 4.5**). The duplicates were removed by SAMTOOLS[2] (parameters: rmdup).

Table 4.5 The statistics of mapping rate and coverage

Sample	Mapped reads	Total reads	Tag number	Mapping rate (%)	Average depth(X)	Coverage at least 1X (%)	Coverage at least 4X (%)
C1	8,675,498	8,646,433	329,960	99.66	9.64	10.24	3.83
C2	6,484,774	6,464,509	301,278	99.69	8.38	8.82	3.47
C3	6,780,840	6,759,264	301,357	99.68	7.14	10.77	3.53
C4	6,242,128	6,219,402	293,065	99.64	6.15	11.45	3.45
C5	5,771,530	5,753,858	290,539	99.69	6.13	10.67	3.40
C6	6,271,786	6,248,377	300,320	99.63	7.90	9.05	3.48
C7	5,672,200	5,651,608	260,131	99.64	5.48	11.62	3.10
C8	6,404,712	6,383,214	301,756	99.66	6.42	11.28	3.54
C9	6,302,508	6,278,396	289,099	99.62	6.24	11.41	3.40
C10	6,276,966	6,253,523	298,940	99.63	6.49	10.96	3.53
C11	6,345,996	6,321,522	299,763	99.61	6.62	10.85	3.52
C12	5,719,538	5,700,007	291,840	99.66	6.78	9.57	3.38
C13	6,392,336	6,369,841	303,734	99.65	7.74	9.39	3.51
C14	6,969,916	6,945,095	309,264	99.64	7.06	11.19	3.64
C15	5,316,936	5,297,484	258,641	99.63	5.69	10.52	3.05
C16	5,829,800	5,808,438	293,380	99.63	7.02	9.43	3.41
C17	3,570,248	3,561,944	217,285	99.77	4.35	9.22	2.54
C18	5,715,346	5,696,620	279,165	99.67	6.09	10.60	3.27
C19	4,961,648	4,945,184	248,663	99.67	4.67	11.89	2.96
C20	5,942,966	5,925,941	292,667	99.71	5.88	11.42	3.44
C21	6,833,170	6,812,337	306,250	99.70	5.68	13.54	3.67
C22	7,731,682	7,713,019	321,577	99.76	6.56	13.31	3.84

C23	6,698,404	6,681,459	308,347	99.75	6.32	12.00	3.64
C24	6,817,806	6,798,623	304,170	99.72	7.38	10.50	3.55
C25	6,986,538	6,969,974	312,063	99.76	8.32	9.58	3.61
C26	5,086,244	5,071,744	261,704	99.71	4.56	12.48	3.12
C27	6,383,846	6,367,769	305,420	99.75	7.05	10.31	3.56
C28	6,632,952	6,615,597	301,445	99.74	5.79	12.90	3.58
C29	6,613,722	6,596,308	306,208	99.74	6.74	11.13	3.58
C30	7,082,308	7,063,752	310,371	99.74	7.05	11.39	3.64

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) Tag number: Total number of unique tags (enzyme cutting fragment).

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

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

(6) 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.

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

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

4.3.3 Mapping Summary

For the current xxxx bp reference genome, the mapping rate of each sample ranges from 99.61% to 99.77%. The average depth on the reference genome (without Ns) is in 4.35X to 9.64X range, while the more than 1X coverage exceeds 8.82%. This result is in the qualified normal range and may serve in the subsequent variation detection and related analyses.

4.4 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;

4.4.1 Statistics of SNP Detection and Annotation

ANNOVAR[3] is a widely used software in variation annotation with multiple capabilities, including gene-based annotation, region-based annotation, filter-based annotation as well as other functionalities. 1st BASE use ANNOVAR to do annotation of detected SNPs. The results are listed in **Table 4.6**.

Table 4.6 Statistics of SNP detection and annotation

Sample	Upstream	Exonic				Intronic	Splicing
		Stop gain	Stop loss	Synonymous	Non-synonymous		

C1	1.558	0	0	1.288	468	78.196	3
C2	1,369	0	0	1,202	433	72,552	2
C3	1,382	0	0	1,159	437	72,946	2
C4	1,229	0	0	1,093	408	68,563	4
C5	1,359	0	0	1,121	433	69,598	1
C6	1,390	2	0	1,183	399	70,384	2
C7	1,147	2	0	1,029	383	61,889	4
C8	1,330	0	0	1,178	420	72,485	2
C9	1,217	1	0	1,108	394	67,560	4
C10	1,323	0	0	1,182	417	69,416	2
C11	1,267	0	0	1,133	433	69,228	4
C12	1,367	0	0	1,178	437	71,633	5
C13	1,390	0	0	1,160	426	72,177	2
C14	1,333	0	0	1,242	463	73,403	2
C15	1,093	0	0	972	373	60,065	4
C16	1,322	1	0	1,121	414	68,998	1
C17	972	0	0	886	324	53,406	3
C18	1,216	0	0	1,137	407	65,741	4
C19	1,139	0	0	1,068	380	62,422	3
C20	1,412	0	0	1,227	445	72,776	2
C21	1,468	1	0	1,301	468	77,861	4
C22	1,617	2	0	1,374	488	82,593	3
C23	1,444	1	0	1,250	495	76,537	3
C24	1,383	0	0	1,269	449	75,509	3
C25	1,473	1	0	1,230	448	76,201	2
C26	1,298	1	0	1,135	394	67,266	1
C27	1,470	1	0	1,235	441	76,441	4
C28	1,531	1	0	1,262	467	76,538	4
C29	1,435	1	0	1,268	463	76,311	2

Sample	Downstream	Upstream/ Downstream	Intergenic	ts	tv	ts/tv	Het rate(‰)	Total
C1	1,883	36	110,971	136,706	57,697	2	0.05	194,403
C2	1,747	25	101,424	125,819	52,935	2	0.05	178,754
C3	1,677	34	101,892	126,429	53,100	2	0.05	179,529
C4	1,646	46	96,334	118,611	50,712	2	0.04	169,323
C5	1,640	21	96,732	120,372	50,533	2	0.04	170,905
C6	1,679	29	100,454	123,161	52,361	2	0.05	175,522
C7	1,440	26	87,491	107,904	45,507	2	0.04	153,411
C8	1,679	32	101,733	125,377	53,482	2	0.05	178,859
C9	1,532	42	94,635	116,950	49,543	2	0.04	166,493
C10	1,757	56	98,886	121,661	51,378	2	0.04	173,039
C11	1,714	52	98,636	121,090	51,377	2	0.04	172,467
C12	1,765	34	98,449	123,000	51,868	2	0.05	174,868
C13	1,675	36	103,063	126,384	53,545	2	0.05	179,929
C14	1,808	53	102,641	127,210	53,735	2	0.04	180,945
C15	1,376	30	85,000	104,431	44,482	2	0.03	148,913
C16	1,622	34	96,699	119,489	50,723	2	0.04	170,212
C17	1,273	19	71,920	90,628	38,175	2	0.03	128,803
C18	1,641	51	90,213	113,175	47,235	2	0.03	160,410

C19	1,469	31	83,335	105,590	44,257	2	0.04	149,847
C20	1,806	37	97,833	123,710	51,828	2	0.05	175,538
C21	1,945	35	104,011	131,920	55,174	2	0.05	187,094
C22	2,055	37	109,328	139,311	58,186	2	0.05	197,497
C23	1,931	28	104,034	130,911	54,812	2	0.05	185,723
C24	1,890	37	102,432	129,075	53,897	2	0.05	182,972
C25	1,871	37	104,453	130,806	54,910	2	0.05	185,716
C26	1,580	35	88,320	113,090	46,940	2	0.04	160,030
C27	1,878	37	103,941	130,660	54,788	2	0.05	185,448
C28	1,903	46	102,008	129,743	54,017	2	0.05	183,760
C29	1,838	36	103,518	130,301	54,571	2	0.05	184,872
C30	1,951	44	104,996	133,530	55,751	2	0.05	189,281

The details for SNP detection and annotation statistics are as follows:

(1) Sample: Sample name;

(2) Upstream: SNPs located within 1 kb upstream (away from transcription start site) of the gene.

(3) Exonic: SNPs located in exonic region; Non-synonymous: single nucleotide mutation with changing amino acid sequence; Stop gain/loss: a nonsynonymous SNP that leads to the introduction/removal of stop codon at the variant site; Synonymous: single nucleotide mutation without changing amino acid sequence;

(4) Intronic: SNPs located in intronic region;

(5) Splicing: SNPs located in the splicing site (2 bp range of the intron/exon boundary).

(6) Downstream: SNPs located within 1 kb downstream (away from transcription termination site) of the gene region.

(7) Upstream/Downstream: SNPs located within the < 2 kb intergenic region, which is in 1 kb downstream or upstream of the genes.

(8) Intergenic: SNPs located within the > 2 kb intergenic region.

(9) ts: Transitions, a point mutation that changes a purine nucleotide to another purine (A ↔ G) or a pyrimidine nucleotide to another pyrimidine (C ↔ T). Approximately two out of three SNPs are transitions.

(10) tv: Transversions, the substitution of a (two ring) purine for a (one ring) pyrimidine or vice versa.

(11) ts/tv: The ratio of transitions to transversions.

(12) Het rate: Genome-wide heterozygous rate, calculated by the ratio of heterozygous SNPs to the total number of genome bases.

(13) Total: The total number of SNPs.

4.4.2 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 4.4.2**.

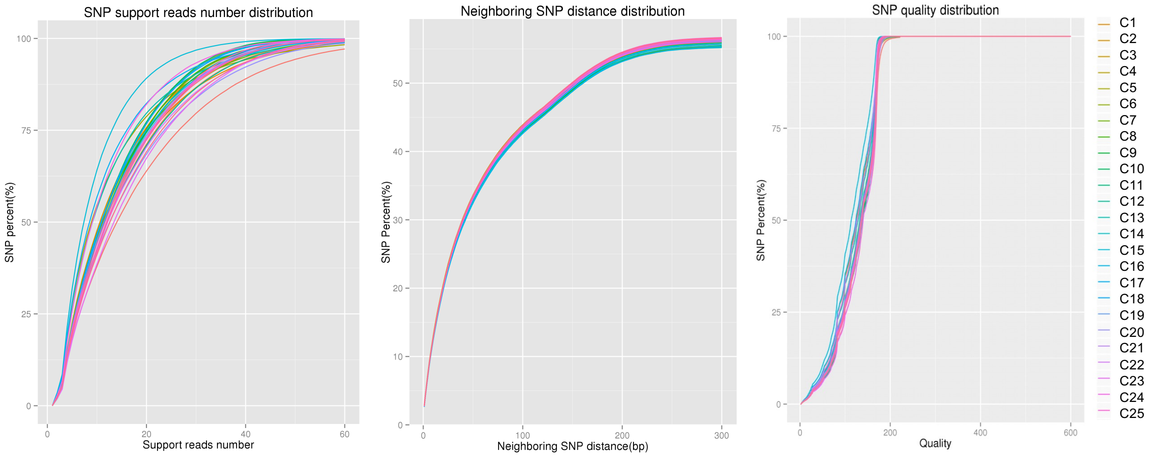

Figure 4.4.2 Cumulative distribution of SNP quality

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.

4.4.3 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 4.4.3**.

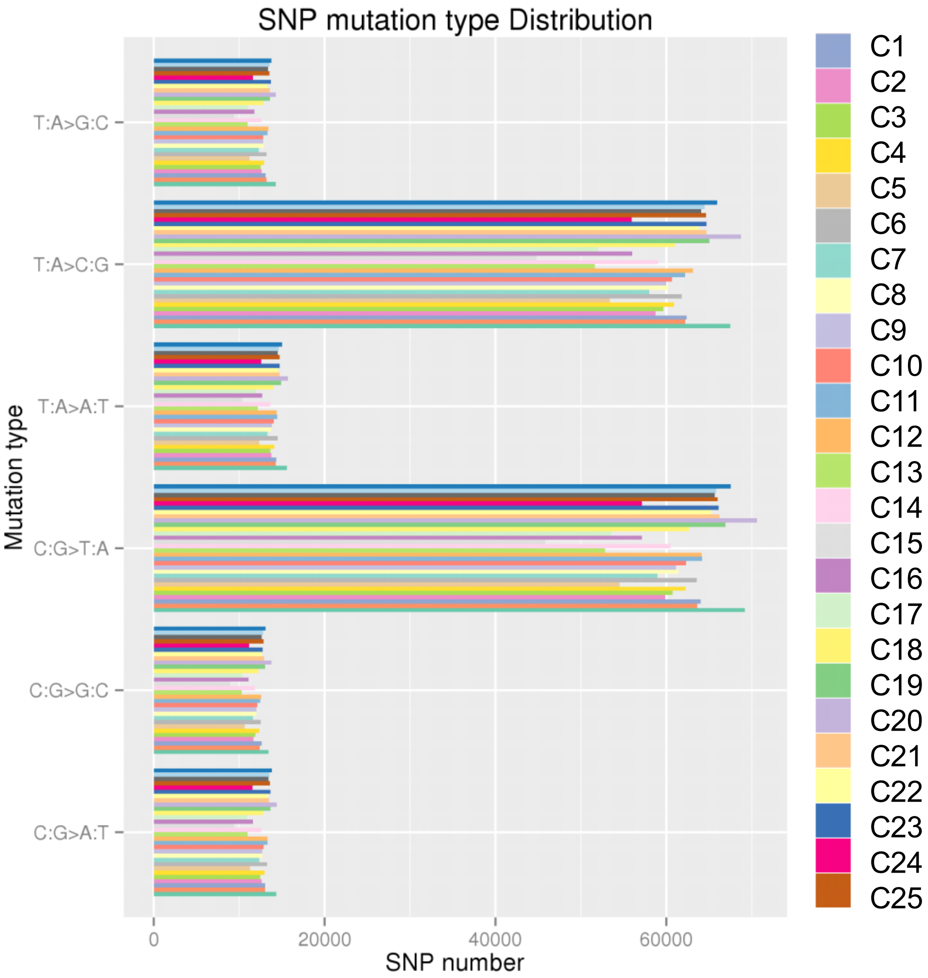

Figure 4.4.3 Frequency of SNP mutations

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

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.

[3] Wang K, Li M, Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 2010, 38(16):e164.

[4] Krzywinski M, Schein J, Birol İ, et al. Circos: an information aesthetic for comparative genomics[J]. Genome research, 2009, 19(9): 1639-1645.