Plant and Animal Re-sequencing (WGS) with Advanced Analysis

Demo Report

May 1, 2016

Contents

1 Project Background

Species name: XXXX;
Number of samples: 12;
Sequencing strategy: Illumina HiSeq PE150; 42.5 G per sample;
Standard analysis of content: Sequencing quality control; SNP calling, annotation and statistics; InDel calling, annotation and statistics; SV calling, annotation and statistics; CNV calling, annotation and statistics.

2 Experimental Procedures

2.1 DNA 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 1.5 μg, was qualified for library construction.

2.2 Library Construction

The genomic DNA of each sample was randomly sheared into short fragments of about 350 bp, respectively. The obtained fragments were subjected to library construction using the Illumina TruSeq Library Construction Kit, with strictly following the instructions. Briefly, as followed by end repairing, dA-tailing and further ligation with Illumina adapter, the required fragments (in 300-500 bp size) with both P5 and P7 sequences were PCR selected and amplified. After gel electrophoresis and subsequent purification, the required fragments were obtained for library construction. The experimental procedures of DNA library preparation are shown in **Figure 2.1**.

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.

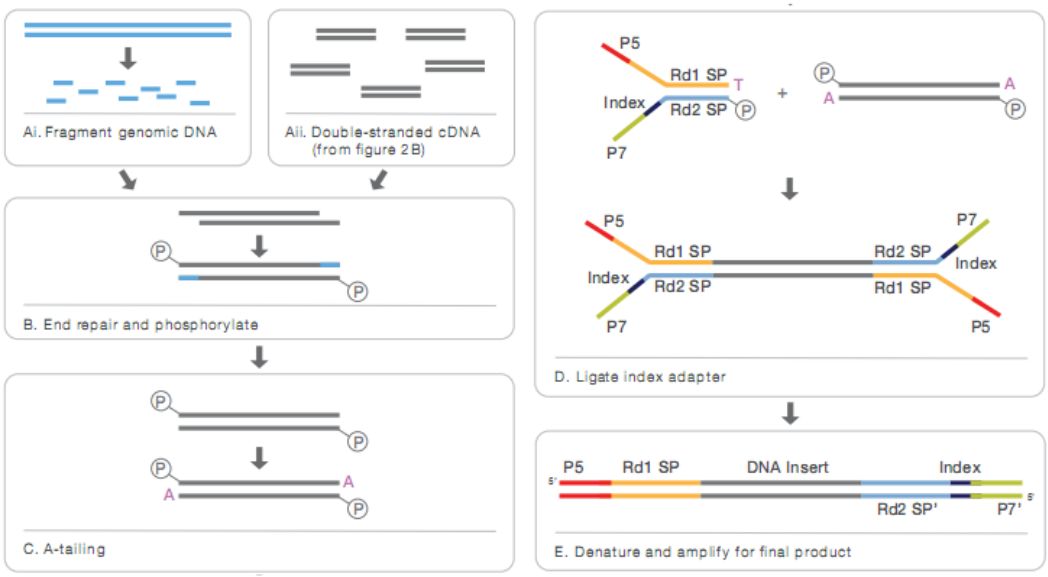

Figure 2.1 Experimental procedures of library preparation

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, InDel, SV and CNV detection and annotation according to the reference genome mapping results.

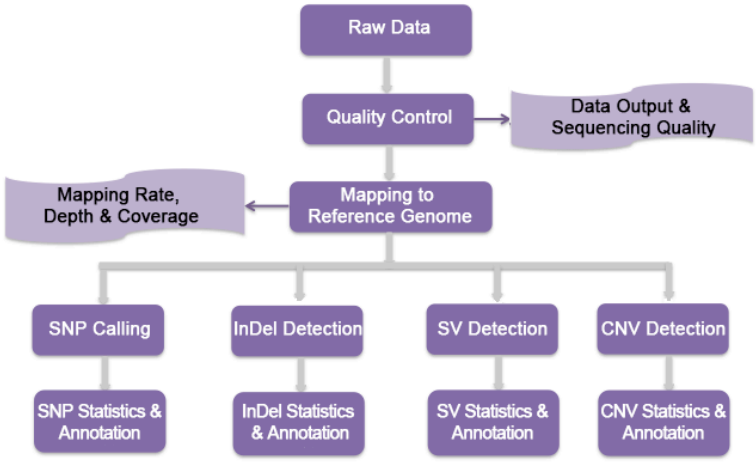

Figure 3.1 Bioinformatics analysis workflow

4 Results of Analysis

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:

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:0 ATCACG
TAGCCACATAGAAACCAACAGCCATATAACTGGTAGCTTTAAGCGGCTCACCTTTAGCATCAACAGGCCAC
AACCAACCAGAACGTGAAAAAGCGTCCTGCGTGTAGCGAACTGCGATGGGCATACAGATCGGAAGAGCGTC
GTGTAGGG
+
AAFFFKKKKKKKKFKKKFFKKAAFKKKKKFKKKKFKKA,FKKKKKKKKKAKKFKKKKKKKAKKKKKKFFKK
KKF<FFKKKKKKKKKKKKKFKKFKKF7FFFFFKFKKKFKKKKKKKKF<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 3.1 Information of Illumina sequence identifiers

Identifer	Meaning
EAS139	Unique instrument name
136	Run ID
FC706VJ	Flowcell ID
2	Flowcell lane number
2104	Tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
Y	Y if the read fails filter (read is bad), N otherwise
18	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**.1.

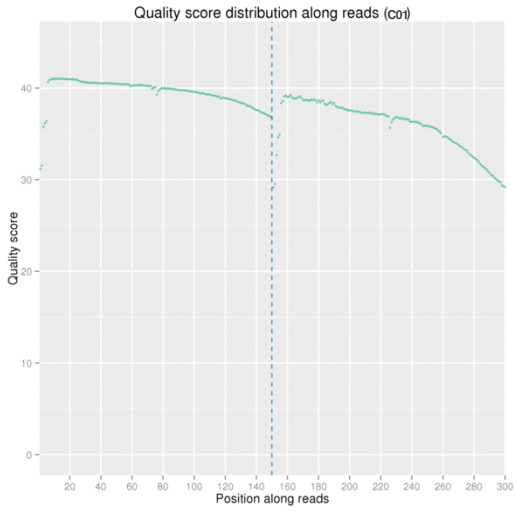

Figure 4.2.1 Distribution of sequencing quality

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**.

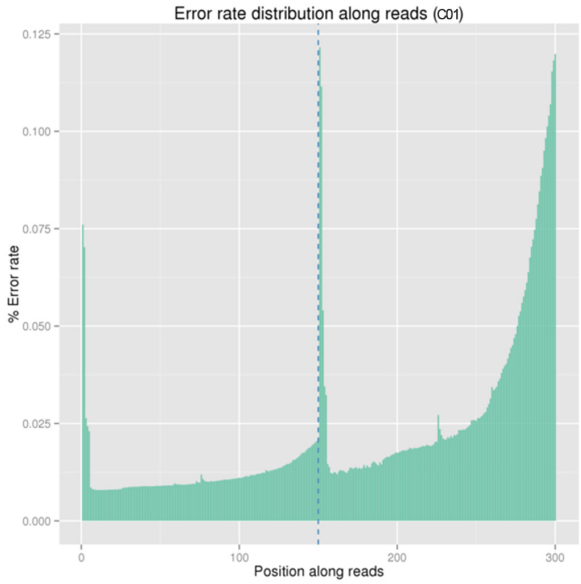

Figure 4.2.2 Distribution of sequencing errors.

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 ≤ 5) constitute more than 50 percent of either read.

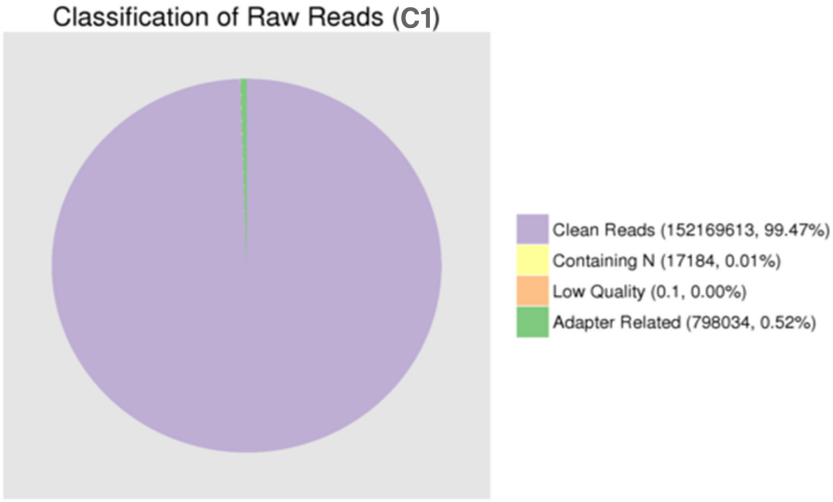

Figure 4.2.3 Classification of the sequenced reads

(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.

4.3 Statistics of Sequencing Data

Consistent with the Illumina platform sequencing features, for PE data, the error rate required to 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 (%)
C01	58,254,060,000	55,718,217,300	95.65	0.03	96.91	92.86	36.71
C02	48,999,575,100	48,813,596,700	99.62	0.03	95.50	90.56	36.77
C03	42,688,887,600	42,356,871,000	99.22	0.04	94.59	88.82	37.28
C04	62,424,593,400	59,089,362,000	94.66	0.03	96.22	91.64	37.47
C05	47,131,144,200	46,938,246,300	99.59	0.03	95.25	90.13	37.50
C06	47,767,277,700	47,665,563,900	99.79	0.03	96.47	91.89	37.55
C07	45,895,449,300	45,650,883,900	99.47	0.04	94.18	88.19	37.90
C08	46,121,013,900	46,029,951,300	99.80	0.03	96.68	92.34	37.91
C09	43,359,441,600	43,264,053,900	99.78	0.03	96.56	92.02	38.09
C10	42,028,694,700	41,925,456,600	99.75	0.03	96.25	91.51	38.34
C11	47,653,513,800	47,560,274,400	99.80	0.03	96.79	92.54	38.42
C12	42,083,946,900	41,940,697,500	99.66	0.03	96.59	92.12	39.67

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.

4.4 Mapping Statistics

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

Reference genome is downloaded from: ftp://ftp.ensemblgenomes.org/pub/release-83/plants/fasta/solanum_tuberosum/dna. The statistics of reference genome are listed in **Table 4.4.1.**

Table 4.4.1 The statistics of reference genome

Seq number	Total length	GC content (%)	Gap rate (%)	N50 length	N90 length
13	810,654,046	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.

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.

Table 4.4.2 The statistics of mapping rate and coverage

Sample	Mapped reads	Total reads	Mapping rate (%)	Average depth(X)	Coverage at least 1X(%)	Coverage at least 4X(%)
C01	358963886	371454782	96.64	63.82	97.67	96.19
C02	313264430	325423978	96.26	56.30	97.55	95.88
C03	269298889	282379140	95.37	52.30	96.49	94.66
C04	383297010	393929080	97.30	66.99	97.09	95.53
C05	298295535	312921642	95.33	50.67	94.31	91.05
C06	309280937	317770426	97.33	59.10	95.09	92.69
C07	291648824	304339226	95.83	56.78	97.38	95.34
C08	290309631	306866342	94.60	54.43	95.51	91.45

C09	282745767	288427026	98.03	55.05	95.53	93.11
C10	266152902	279503044	95.22	50.99	96.85	93.47
C11	308804824	317068496	97.39	59.35	95.75	92.88
C12	270368070	279604650	96.70	50.63	94.47	90.07

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.

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;

5.1 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 uses ANNOVAR to do annotation of detected SNPs. The results are listed in **Table 5.1**.

Table 5.1 Statistics of SNP detection and annotation

Sample	Upstream	Stop gain	Stop loss	Exonic		Intronic	Splicing
				Synonymous	Non-synonymous		
C01	495,196	5,942	1,159	212,165	243,302	976,484	1,366
C02	449,298	5,071	1,083	204,176	227,448	946,815	1,291
C03	502,581	4,621	1,144	166,024	196,634	845,589	1,180
C04	473,883	4,439	1,059	155,586	186,695	788,493	1,104
C05	823,097	8,412	1,789	265,197	322,883	1,397,185	2,077
C06	591,293	6,231	1,429	190,074	239,214	998,613	1,618
C07	379,454	5,268	1,111	210,742	233,049	892,337	1,271
C08	819,989	9,861	1,854	417,257	442,099	1,860,450	2,319

C09	464,553	4,632	1,064	156,792	189,223	798,370	1,155
C10	196,579	3,912	746	194,924	198,457	649,753	837
C11	498,165	6,540	1,441	213,380	256,585	1,011,843	1,678
C12	740,229	7,814	1,724	268,088	321,158	1,359,855	1,997

(Continued next page)

Sample	Downstream	Upstream/ Downstream	Intergenic	ts	tv	ts/tv	Het rate(%)	Total
C01	487,176	39,698	6,787,578	5,712,918	3,537,148	1.615	11.620	9,250,066
C02	454,217	37,467	6,061,645	5,194,566	3,193,945	1.626	9.891	8,388,511
C03	462,082	39,151	6,418,618	5,361,623	3,276,001	1.636	9.489	8,637,624
C04	436,733	35,440	6,239,168	5,169,518	3,153,082	1.639	9.287	8,322,600
C05	753,628	63,059	11,162,737	9,158,984	5,641,080	1.623	14.827	14,800,06
C06	549,472	44,837	8,438,969	6,830,062	4,231,688	1.614	11.161	11,061,75
C07	403,742	32,176	5,494,006	4,770,026	2,883,130	1.654	9.097	7,653,156
C08	836,234	68,233	10,606,328	9,235,684	5,828,940	1.584	17.902	15,064,62
C09	435,034	35,855	6,224,982	5,155,470	3,156,190	1.633	8.434	8,311,660
C10	242,420	17,512	2,987,053	2,745,576	1,746,617	1.571	5.083	4,492,193
C11	505,818	40,266	7,446,697	6,208,373	3,774,040	1.645	11.629	9,982,413
C12	705,933	57,674	10,253,648	8,537,256	5,180,864	1.647	15.328	13,718,12

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.

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 5.2**.

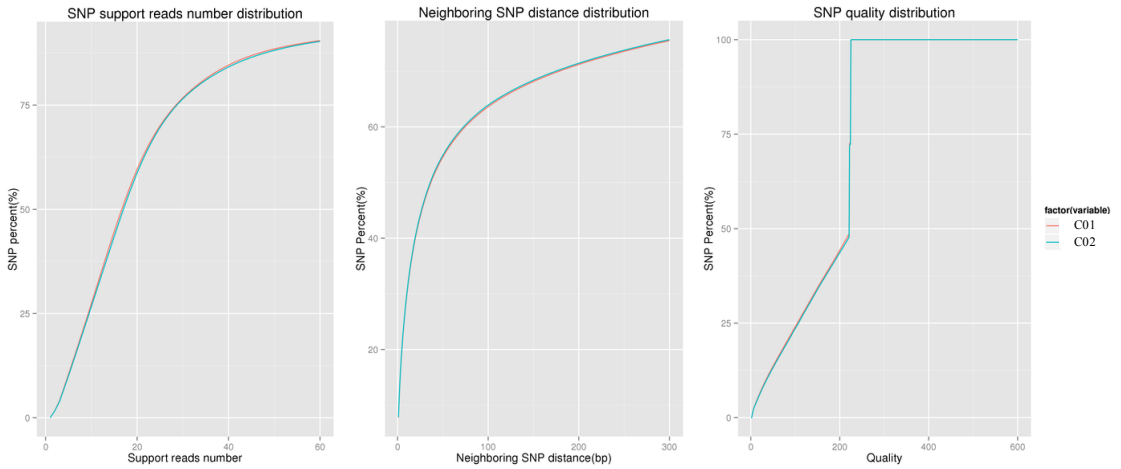

Figure 5.2 Cumulative distribution of SNP quality

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.

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 5.3**.

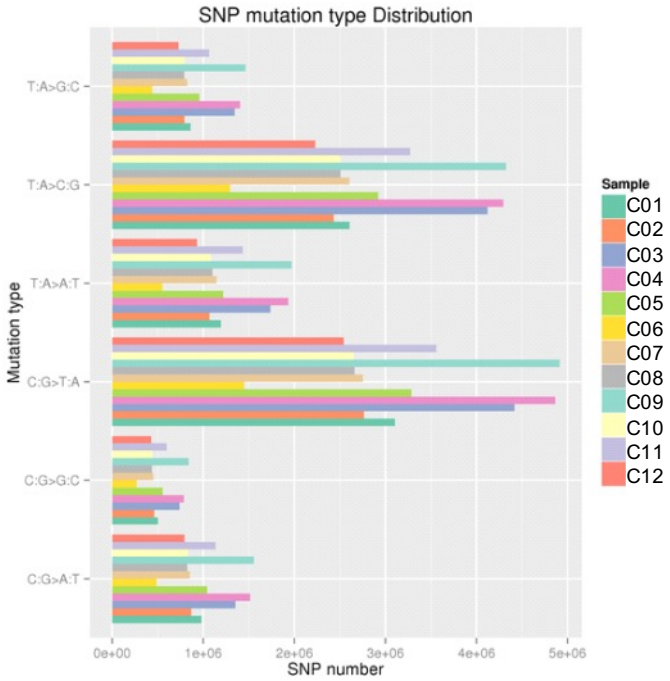

Figure 5.3 Frequency of SNP mutations

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

6 InDel Detection and Annotation

InDel refers to the insertion or deletion of ≤ 50 bp sequences in the DNA. 1st BASE uses SAMTOOLS (mpileup -m 2 -F 0.002 -d 1000) to detect InDels and followed by annotation using ANNOVAR.

6.1 Statistics of InDel Detection and Annotation

Table 6.1 Statistics of InDel detection and annotation

Sample	Upstream	Stop gain	Stop loss	Frameshift deletion	Exonic Frameshift insertion	Non-frameshift insertion	Non-frameshift insertion
C01	103,895	303	105	4,416	3,334	3,463	2,948
C02	109,365	307	121	4,417	3,267	3,793	3,215
C03	86,059	223	88	3,740	2,731	3,151	2,828
C04	82,751	248	94	3,636	2,646	3,030	2,709
C05	140,288	402	130	6,128	4,324	5,180	4,497
C06	97,194	308	112	4,537	3,308	3,710	3,288
C07	100,284	273	120	4,416	3,187	3,717	3,014
C08	162,692	520	173	7,664	5,289	7,028	5,274
C09	77,258	220	88	3,707	2,656	3,032	2,730
C10	103,152	333	124	4,878	3,705	4,142	3,328
C11	87,694	299	98	4,454	3,251	3,580	2,963
C12	112,032	403	127	5,811	4,224	5,175	4,558

Sample	Intronic	Splicing	Downstream	Upstream/ Downstream	Intergenic	Insertion	Deletion	Het rate(%)	Total
C01	150,304	389	93,664	8,770	915,037	592,017	694,622	1,395	1,286,639
C02	160,958	428	100,371	9,470	905,869	595,607	705,995	1.341	1,301,602
C03	126,924	340	79,514	7,501	727,386	482,658	557,841	0.895	1,040,499
C04	119,783	343	76,115	6,772	725,283	474,452	548,972	0.886	1,023,424
C05	205,889	581	128,473	12,148	1,163,568	775,531	896,107	1.427	1,671,638
C06	144,131	406	90,169	8,396	870,304	569,905	655,979	1.024	1,225,884
C07	151,274	421	92,867	8,787	844,485	554,549	658,314	1.272	1,212,863
C08	275,864	737	156,940	14,653	1,291,518	878,928	1,049,45	2.057	1,928,378
C09	119,294	342	73,336	6,684	678,988	447,465	520,880	0.764	968,345
C10	169,085	471	101,711	9,639	893,359	600,399	693,546	1.452	1,293,945
C11	141,846	388	84,846	7,648	782,843	512,579	607,342	1.102	1,119,921
C12	193,616	540	112,636	10,374	939,728	643,099	746,148	1.293	1,389,247

The details of InDel annotation statistics are as follows:

(1) Sample: Sample names.

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

(3) Exonic: InDels located in exonic region; Stop gain/loss: InDel that leads to the introduction/removal of stop codon at the variant site; Frameshift deletion/insertion: InDel mutation changing the open reading frame with deletion or insertion; Non-Frameshift deletion/insertion: InDel mutation without changing the open reading frame with deletion or insertion sequences of 3 or multiple of 3 bases;

(4) Intronic: InDel located in intronic region;

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

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

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

(8) Intergenic: InDel located within the > 2 Kb intergenic region.

(9) Het rate: InDel heterozygous rate, calculated by the ratio of InDels to the total number of genome bases.

(10) Total: The total number of InDels.

6.2 Length Distribution of CDS-located InDels

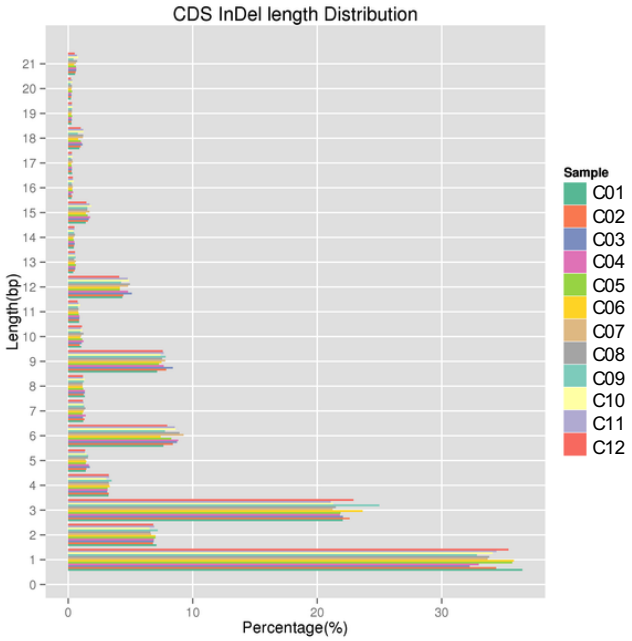

Figure 6.2 Length distribution of CDS-located InDels

The x-axis represents the proportion of the InDels with a certain length, and y-axis indicates the length of the InDels.

7 SV Detection and Annotation

Structural variants (SVs) are genomic variation with muations of relatively larger size (>50 bp), including deletions, duplications, insertions, inversions and translocations. BreakDancer[4] software were used to detect insertion (INS), deletion (DEL), inversion (INV), intra-chromosomal translocation (ITX) and inter-chromosomal translocation (CTX) mutations, based on the reference genome mapping results and the detected insert size. The detected SVs were filtered by removing those with less than 2 supporting PE reads, the INS, DEL and INV were

further annotated by ANNOVAR.

7.1 Statistics of SV Detection and Annotation

Table 7.1 Statistics of SV detection and annotation

Sample	Upstream	Exonic	Downstream	Intronic	Upstream/ Downstream	Intergenic	Splicing
C01	715	1,898	629	1,055	53	7,608	12
C02	603	1,514	540	1,012	55	6,356	8
C03	488	1,561	476	828	42	5,437	12
C04	662	1,916	576	1,033	58	7,124	15
C05	753	1,596	675	1,188	66	7,348	14
C06	684	1,888	638	1,167	50	7,887	23
C07	479	1,390	462	904	43	5,597	6
C08	456	1,232	499	1,247	46	4,817	18
C09	492	1,555	462	929	49	6,018	10
C10	368	1,015	347	810	30	4,034	10
C11	522	1,649	539	1,052	42	6,536	8
C12	448	1,313	439	1,043	48	4,955	21

Sample	INS	DEL	INV	ITX	CTX	Total
C01	1,357	10,606	7	1,562	8,423	21,955
C02	1,050	9,034	4	1,241	6,158	17,487
C03	1,131	7,708	5	1,527	5,539	15,910
C04	1,530	9,845	9	1,700	7,784	20,868
C05	921	10,712	7	1,016	7,606	20,262
C06	1,291	11,037	9	1,456	7,617	21,410
C07	942	7,935	4	1,279	4,818	14,978
C08	757	7,555	3	804	5,545	14,664
C09	1,177	8,334	4	1,476	5,704	16,695
C10	647	5,964	3	825	3,264	10,703
C11	1,106	9,233	9	1,257	6,142	17,747
C12	862	7,398	7	951	5,102	14,320

The details of SV detection statistics are as follows:

(1) Sample: Sample names.

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

(3) Exonic: SVs located in exonic region.

(4) Intronic: SVs located in intronic region.

(5) Downstream: SVs located within 1 Kb downstream (away from transcription termination site) of the gene region.

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

(7) Intergenic: SVs located within the > 2 Kb intergenic region.

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

(9) INS: Insersion.

(10) DEL: Deletion.

(11) INV: Inversion.

(12) ITX: Intra-chromosomal translocations.

(13) CTX: Inter-chromosomal translocations.

(14) Total: The total number of SVs.

7.2 Length Distribution of SVs

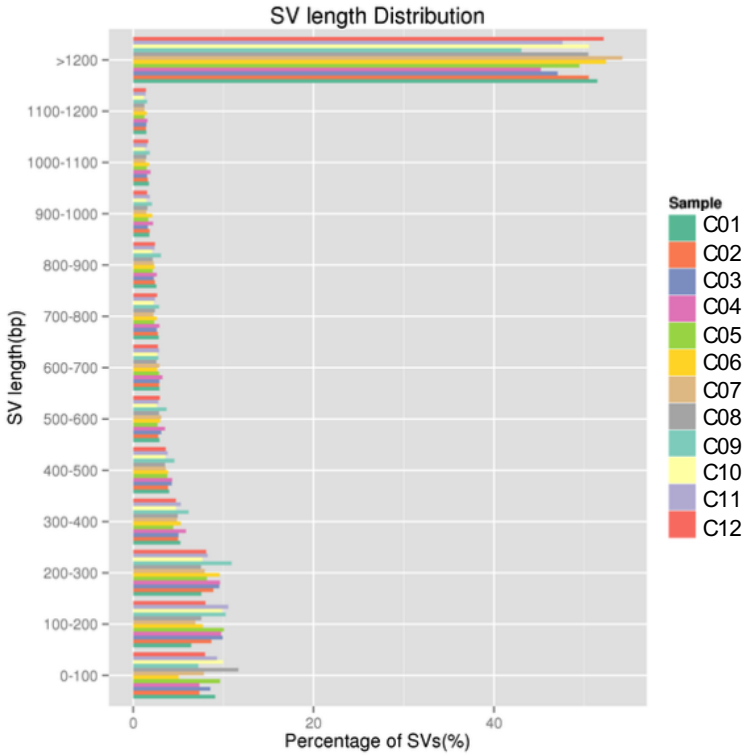

Figure 7.2 Length distribution of SVs

The x-axis represents samples, and the y-axis indicates the proportion of each type of SVs. Note, the length of DNA insert in library construction impacts the SVs detection greatly..

8 CNV Detection and Annotation

Copy-number variations (CNVs) are alterations of the genome that results in the cell having an abnormal or, for certain genes, a normal variation in the number of copies of one or more sections of the DNA. CNVs correspond to relatively large regions of the genome that have been deleted (fewer than the normal number) or duplicated (more than the normal number) on certain chromosomes. Based on the reads depth of the reference genome, CNVnator[5] (-call 100) were used to detect CNVs of potential deletions and duplications. The detected CNVs were further annotated by ANNOVAR.

Table 8 Statistics of CNV detection and annotation

Sample	Upstream	Exonic	Intronic	Downstream	Upstream/ Downstream	Intergenic
C01	1,216	5,164	551	1,035	103	18,527
C02	1,281	5,133	524	1,063	94	17,890
C03	1,389	4,035	594	1,147	117	18,451
C04	1,440	3,937	687	1,174	133	19,797
C05	1,108	7,060	357	887	106	15,610
C06	1,325	5,477	520	1,105	128	18,262
C07	1,075	6,023	389	935	102	15,525
C08	698	9,512	127	554	57	10,859
C09	1,311	4,600	476	1,031	132	17,201
C10	741	8,928	141	627	73	11,569
C11	1,071	6,354	288	822	106	15,088
C12	915	8,531	135	663	80	11,819

Sample	Duplication number	Deletion number	Duplication length (bp)	Deletion length (bp)	Total
C01	5,761	20,836	72,408,600	236,679,900	26,597
C02	5,793	20,193	69,397,000	233,863,500	25,986
C03	3,692	22,043	29,636,200	232,695,400	25,735
C04	3,479	23,692	26,965,400	230,513,900	27,171
C05	7,831	17,297	137,833,600	266,139,800	25,128
C06	5,972	20,846	76,608,200	255,683,200	26,818
C07	6,888	17,162	121,112,500	232,793,800	24,050
C08	10,424	11,383	341,210,100	205,827,100	21,807
C09	4,607	20,145	52,636,800	240,624,300	24,752
C10	9,817	12,262	284,813,300	208,857,400	22,079
C11	7,591	16,138	157,138,700	232,292,600	23,729
C12	9,494	12,649	259,029,400	210,366,500	22,143

The details of CNV detection and annotation are as follows:

(1) Sample: Sample names.

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

(3) Exonic: CNVs located in exonic region.

(4) Intronic: CNVs located in intronic region.

(5) Downstream: CNVs located within 1 Kb downstream (away from transcription termination site) of the gene region.

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

(7) Intergenic: CNVs located within the > 2 Kb intergenic region.

(8) Duplication: CNVs with increased copy number.

(9) Deletion: CNVs with decreased copy number.

(10) Duplication length (bp): The total length of CNV duplication.

(11) Deletion length (bp): The total length of CNV deletion.

(12) Total: The total number of CNVs.

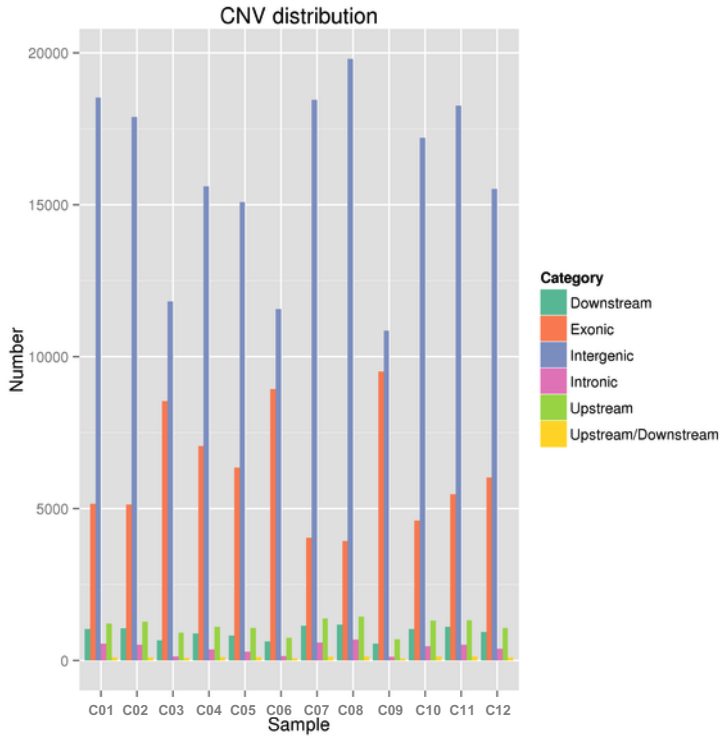

Figure 8.1 Distribution of CNVs on the genome

9 Visualization of Variation

Genome-wide Structrual Variation Visualization

For proper visualization of the structural variations on the whole-genome, we present them according to mutation types:

(1) for SNP/InDel type, the density distribution is drawn;

(2) for SV/CNV type, the location and size are drawn;

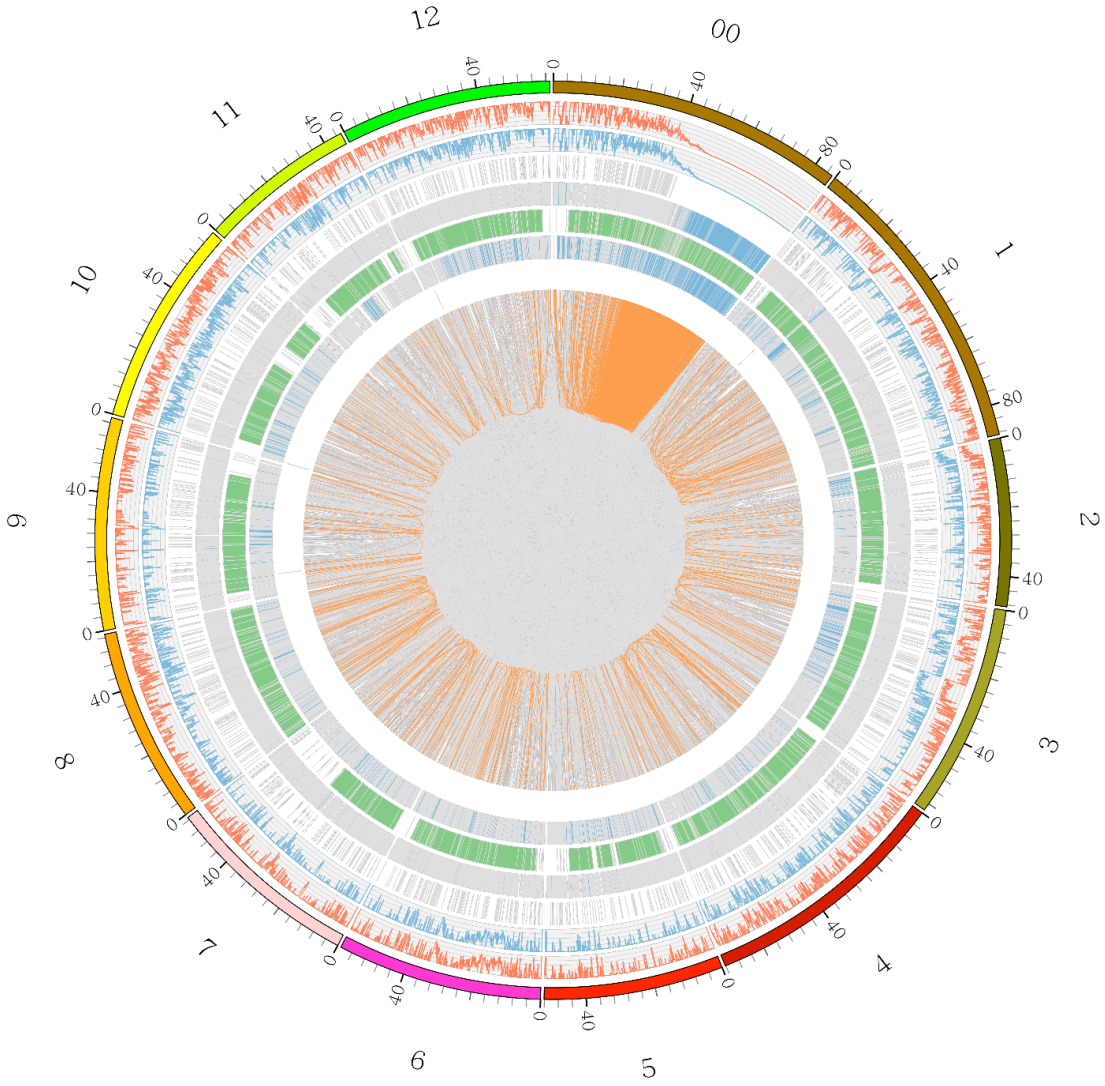

Figure 9 Genome-wide variation distribution

From outer to inner, chromosome, SNP, InDel, CNV duplication, CNV deletion, SV insertion, SV deletion, SV invertion, SV ITX, SV CTX.

10 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.