CD Genomics
The Genomics Services Company

Whole Microbial Genome Sequencing

Final Report

Order#: xxxxxxxxxx

Date: xx/xx/xxxx

Contents

1. Summary

Microbes are widely found in nature and are closely related to human production and life. At present, microbes are mainly divided into fungi, actinomycetes, bacteria, spirochetes, rickettsia, chlamydia, mycoplasma and virus. In the late 1990s, the microbial single genome research was initiated. This study began to understand the complete biological function of a single microbial organism based on the analysis of gene structure from the complete nucleotide sequencing of the microorganism. As the microbial genome research has a clear goal, relatively low input, quick results, easy to translate into products and many other advantages, the majority of microbial single genome has been reported or is sequencing, but there are some specific microbial resources in many countries, if strengthened into the genome study, it is possible to get a high starting point soon.

With the significant reduction of sequencing cost and the improvement of sequencing efficiency, genome sequencing has played a great role in promoting microbial single genomics research. Through the whole genome sequencing, the genome database of the species can be constructed, an efficient platform for the further study on the growth, development, evolution and origin of the species can be established, and the DNA sequence information can be provided for subsequent gene mining and functional verification. Genome sequencing can produce a new understanding of microbial single bacteria at the genome level. Subsequent research on the structures and functions of the genomes can provide opportunities and lay the foundation for innovative research and applied research in the fields of medicine, industry, and agriculture.

Bacterial genome sequencing can be divided into two categories which are bacterial genome de novo sequencing and bacterial genome resequencing. The bacterial genome de novo sequencing refers to the sequencing of a bacterial species without any existing sequence information, and the use of bioinformatics analysis to assemble the sequence to obtain the genome of the bacterial sequence. Bacterial genome resequencing is the genome sequencing of different individuals of the reference sequence, and on this basis, individual or group-level differences are analyzed. A large number of single nucleotide polymorphism (SNP), InDel (Insertion and Deletion), structural variation (SV) and other variation information can be detected. Bacterial genome sequencing can predict important genes and proteins to understand their functions and possible regulatory mechanisms that have replaced traditional

methods and become an important tool for studying the genetic mechanisms of bacterial evolution and key functional genes.

In this report, we have completed the whole genome resequencing for one sample, the data analysis includes the following items.

1) Data Statistics. Including the quantity and quality of sequencing data and GC content of the data.
2) Alignment with reference genome. Including the mapping rate, genome coverage and depth of the genome sequencing.
3) Variations detection and annotation. SNP, InDel, SV and CNV.
4) Gene mutation analysis and annotation. Annotation based on KEGG, GO, COG, NR and SwissProt databases.

2. Introduction

2.1Pipeline of Experiment

Illumina HiSeq Sequencing Platform.

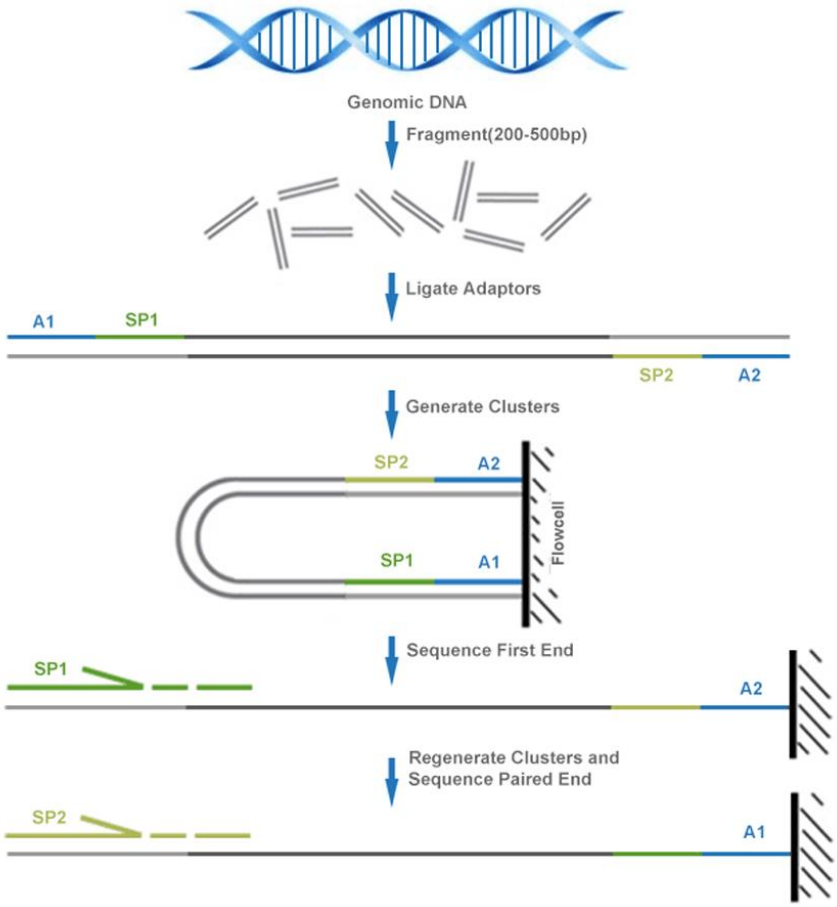

Figure 1. Pipeline of experiment.

After the DNA samples were delivered, we conducted a sample quality test first. Then we used those qualified DNA samples to construct library： Purified DNA samples were sheared into smaller fragments with the desired sizes by Covaris S/E210 or Bioruptor first. Then the overhangs resulting from fragmentation were converted into blunt ends by using T4 DNA polymerase, Klenow Fragment, and T4 Polynucleotide Kinase. After adding an ‘A’ base to the 3' end of the blunt phosphorylated DNA fragments, adapters were then ligated to the ends of the DNA fragments. The desired fragments can be purified through gel-electrophoresis, then selectively enriched and amplified by PCR.

The index tags were introduced into the adapter at the PCR stage and a library quality test was performed. At last, the qualified library would be used for sequencing, and the generated data were used for the downstream bioinformatics analysis.

2.2Pipeline of Bioinformatics Analysis

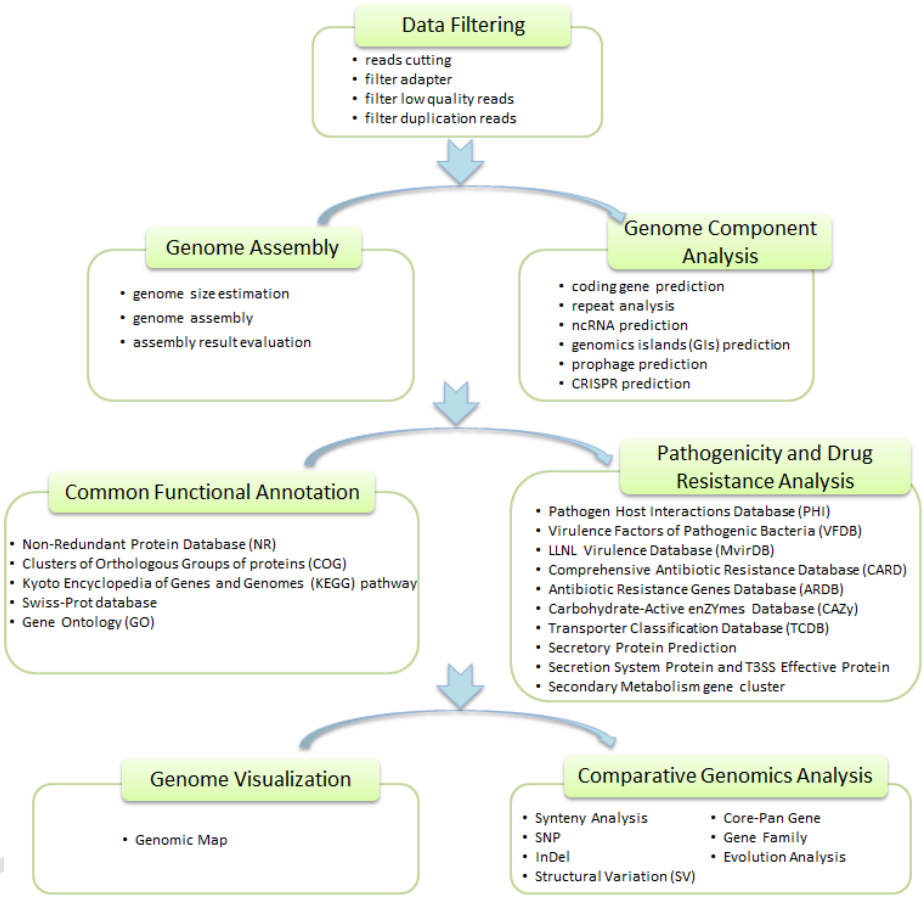

Figure 2. Pipeline of Bioinformatics Analysis.

3. Results

3.1 Data and Statistics

3.1.1 Illumina Data and Statistics

Samples were sequenced using the Illumina HiSeq sequencing platform. Raw data generated on the sequencing platform contained a small percentage of low-quality data, which was then filtered to get high-quality data.The raw data were processed as the following steps:

1. Filter all the reads in the tiles with N's content of greater than or equal to 80%.
3. Filter low quality reads, specifically: normal GC sequences data retained the reads with a base which with a quality greater than 20 greater than 40%, abnormal GC sequences data retained the reads with a base which with a quality greater than 2 greater than 40%.
4. Filter reads N content greater than 10% reads.
5. Filter reads that the adapter sequence can match it greater than or equal to 15bp and mismatch number less than or equal to 3bp.
6. Filter the redundant reads.

The above-mentioned processing methods are simultaneously performed on the operation of read1 and read2. The process will generally remove 10% to 20% of the data (small fragment library data). Due to higher Duplication, more data will be removed from large fragment library data. There is no definite ratio.

* "cycle" represents the base position of all reads equal length in the sequencing. For example, if the sequencing strategy is 150PE, then read1 has 150 cycles and read2 also has 150 cycles.

An example of raw data with FASTQ (fq) is shown in the below Figure 3 and the detailed interpretation is displayed in Table 1.

```
@HWI-7001455:224:HA58FADXX:2:2216:15226:55443 1:N:0:GAAACC
ATATCAGTCCATAGACTTCTTTTAATAATTCCAAGTTCGAAAAAGAATTGATAGCTTCTACCTGTAC
+
CCCFFFFFFHHHHJJJJJJIJJJJJJJIJJJJJJJJJJJJJJJJJJJJJJJJJJIJJJJJIJIIJ
```

Figure 3. A snapshot example of fastq format.

Table 1 Illumina raw data fastq interpretation.

Sequence Identifiers	Information
HWI-ST1276	Instrument – unique identifier of the sequencer
71	run number – Run number on instrument
C1162ACXX	FlowCell ID – ID of flowcell
1	LaneNumber – positive integer
1101	TileNumber – positive integer
1208	X – x coordinate of the spot. Integer which can be negative
2458	Y – y coordinate of the spot. Integer which can be negative
1	ReadNumber - 1 for single reads; 1 or 2 for paired ends
N	whether it is filtered - NB：Y if the read is filtered out, not in the delivered fastq file, N otherwise
0	control number - 0 when none of the control bits are on, otherwise it is an even number
CGATGT	Illumina index sequences

After the sequencing data processing, the detailed statistics of the data of each sample is shown in the following table:

Table 2. Statistics of the data generated in the sequencing.

Sample_ID	Clean_Reads	Clean_Base	Q20(%)	Q30(%)	GC(%)
1a3	8584468	1281820106	97.57	92.78	37.77
WT	7347612	1097278833	97.25	92.00	38.48

Table header comment:

(1) Sample ID: Sample ID

(2) Clean data: amount of clean data after filter

(3) Clean_Base: amount of clean base

(4) Q20 (%): Q20 value of clean data

(5) Q30 (%): Q30 value of clean data

(6) GC (%): GC content of clean data

3.1.2 Quality Distribution

The sequencing quality score of each base is obtained through sequencing, the quality score distribution is plotted as follows:

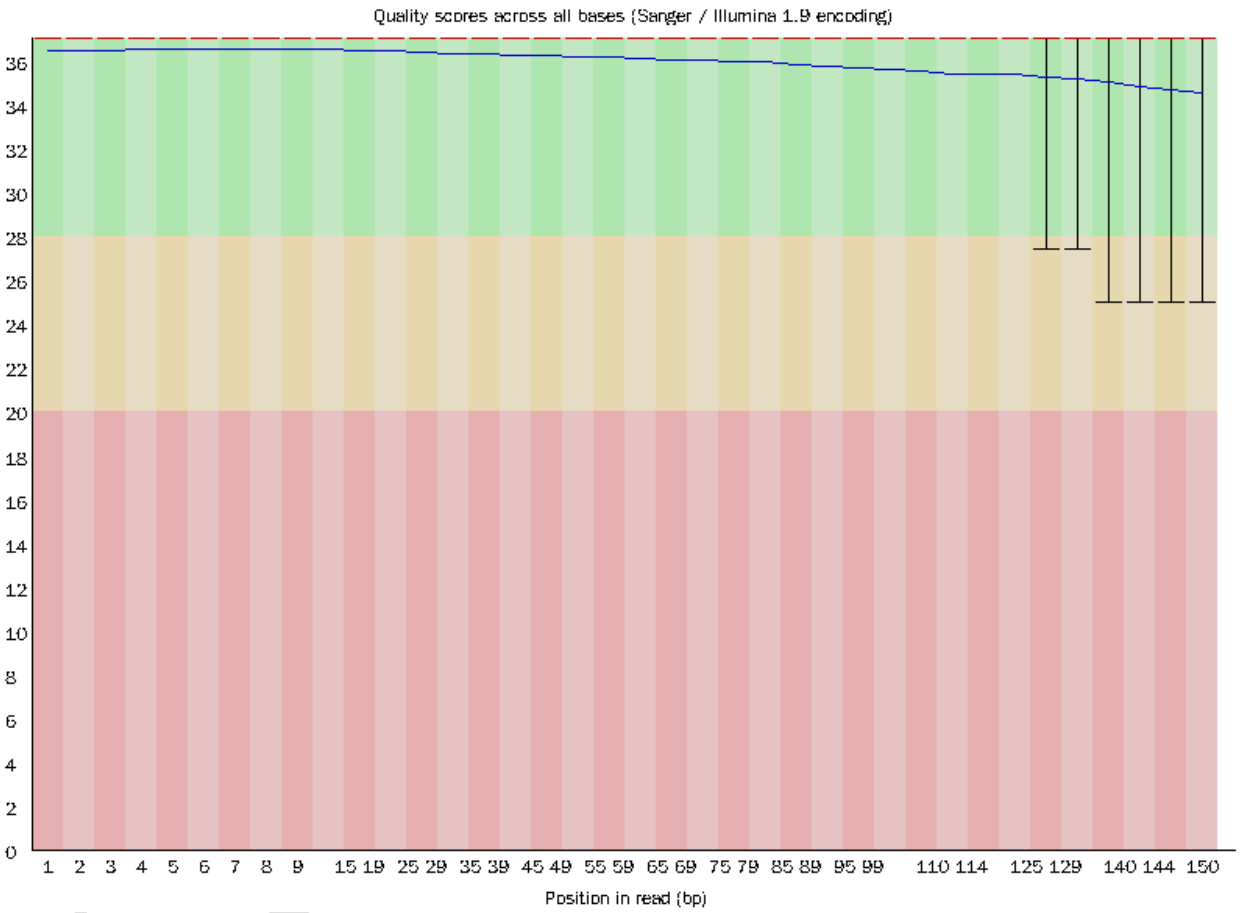

Figure 4. Distribution of base quality. X-coordinate is position of the bases along the reads, Y-coordinate is the quality score at each position along the reads.

3.1.3 Distribution of Base Content

The checking of base content distribution is used to detect the presence or absence of AT, GC separation, which may be caused by sequencing or library construction, and may affect subsequent analysis. The sequenced reads in high-throughput sequencing are randomly interrupted DNA fragments. Since the distribution of the loci on the genome

was similar, the G / C and A / T contents were also approximately uniform. Therefore, according to the large number theorem, in each sequencing cycle, GC, AT content should be equal, and equal to the GC, AT content of the whole genome. Similarly, because of the relationship between the overlapping clusters will lead the fluctuations of the first few bases of the AT, GC, and will be higher than other sequencing sections, and other sections of the GC, AT content is almost equal, uniformly distributed and no separation. The figure is shown below:

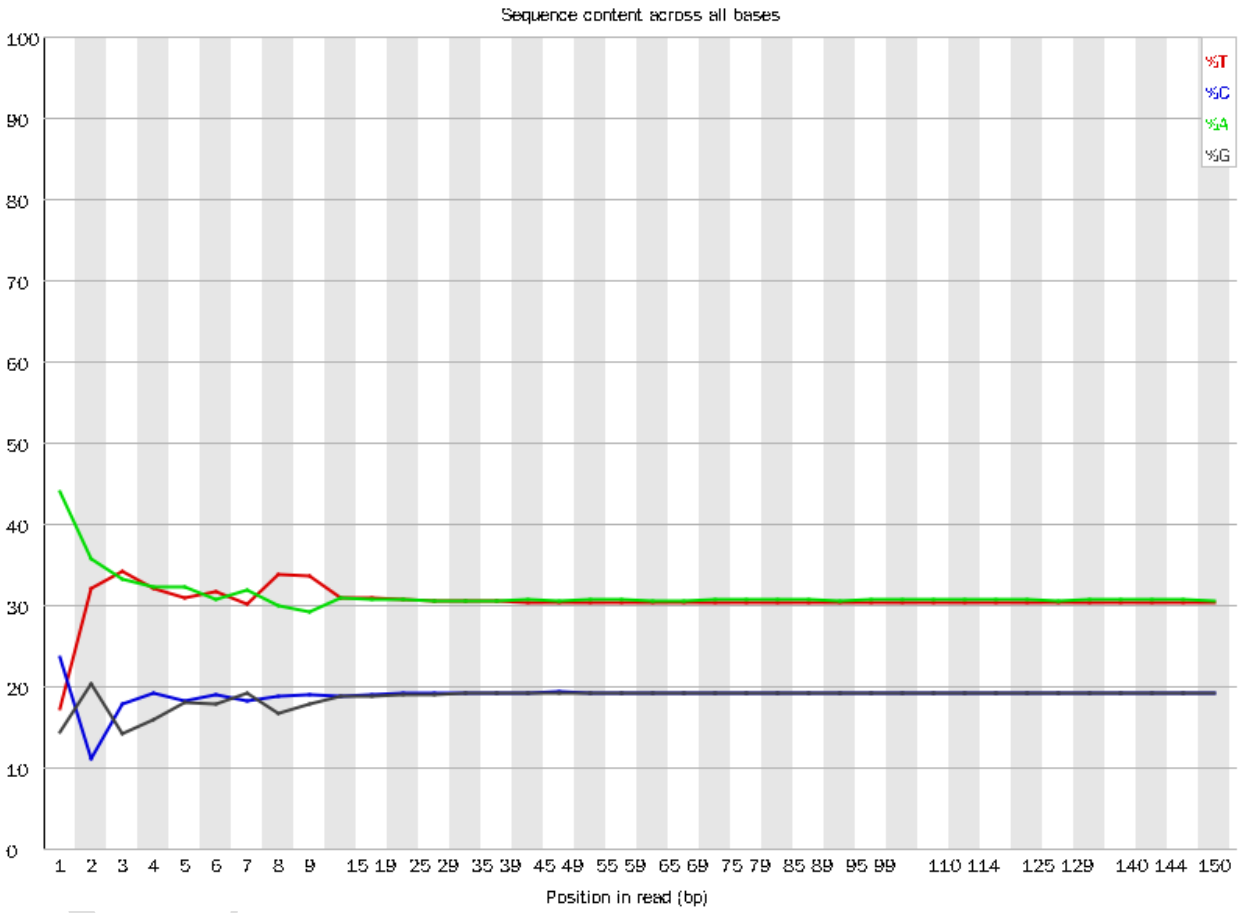

Figure 5. Distribution of base content. X-coordinate represents the position of the bases along the reads. Y-coordinate represents the percentage of each base in the specific position.

3.2 Statistics of Alignment with Reference Genome

The sequencing reads obtained from the re-sequencing need to be mapped to the reference genome for subsequent variation analysis. The bwa [1] software is mainly used for the comparison of short reads obtained from high-throughput sequencing (such as Xten and other sequencing platforms) with reference genomes. The position of clean reads on the reference genome was determined by alignment, and the information such as the depth of sequencing of the samples and the genome coverage were counted and used for variation detection.

3.2.1 Statistics of Alignment Results

The information such as mapping rate are counted in the alignment with reference genome.

https://www.ncbi.nlm.nih.gov/nuccore/U00096

Mapping rate: The percentage of clean reads that can be mapped to the reference genome relative to the total number of clean reads. If the reference genome is properly selected and there is no contamination, the mapping rate of sequencing reads will be higher than 70%. In addition, the level of mapping rate is affected by genetic relationship between the sequencing species and the species with reference genome, the assembly quality and the quality of the sequencing, the more close the species, the more complete the reference genome assembly, the higher quality of sequencing reads, then the higher of mapping rate.

The statistics of alignment is listed in the below table:

Table 3. Statistics of alignment result.

Sample ID	Total_reads	Mapped(%)	Properly_mapped(%)
1a3	8614252	99.88	99.31

WT	7377670	99.78	99.30

Mapped (%): percentage of clean reads that can be mapped to the reference genome relative to the total number of clean reads, achieved through samtools flagstat command.

Properly_mapped(%): the percentage of both paired end reads mapped to the reference genome and the distance conforms to the read length.

3.2.2 Statistics of Sequencing Depth

After mapped to the reference genome by the sequencing reads, the reference genome base coverage can be counted. The percentage of the base number covered by the sequencing reads relative to the base number of whole genome is called the genome coverage; the reads number covered on each base is the coverage depth. Genome coverage can reflect the completeness of the reference genome mutation detection, the more coverage the area, the more mutation sites can be detected. Coverage is mainly affected by the depth of sequencing and the close of the test sample to the species with reference genome. The coverage depth of the genome will affect the accuracy of mutation detection. In addition, if the coverage depth distribution is uniform on the genome, then indicating that the sequencing randomness is good. The average coverage depth and the corresponding genome coverage ratio are displayed in the following table.

Table 4. Statistics of coverage depth and the ratio of coverage.

Sample ID	Ave_depth	Cov_ratio_1X(%)	Cov_ratio_5X(%)	Cov_ratio_10X(%)
1a3	81.212	0.999999	0.999731	0.999416
WT	71.4847	1	0.999789	0.999376

Ave-depth: average of coverage depth. Cov_ratio_(%): The ratio of genome coverage relative to total bases in the reference genome in the specific depth.

3.3 SNP Detection and Annotation

SNP (single nucleotide polymorphism) is mainly referred to the polymorphism caused by single nucleotide variation in genomic level, which can occur in both coding region and non-coding region.

3.3.1 SNP Detection Between Test Sample and Reference

SNP detection is mainly achieved by using the GATK [2] software toolkit. Based on the mapping results, Picard was used to perform Mark Duplicates, GATK was used to

perform Local Realignment, Base Recalibration, etc., to ensure the accuracy of SNP detection, and finally the SNPs set were obtained and filtered by using GATK. The detail process can refer to GRTK Best Practices: https://www.broadinstitute.org/gatk/guide/best-practices.php.

The variants result is shown in vcf format file, an example of vcf file list is shown in the below table.

Table 5. An example of vcf format file.

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	1a3	WT
NC_001133.9	127793	.	G	A	735.5	PASS	AC=2;AF=0.500;AN=4;DP=337;ExcessHet=4.7712;FS=0.000;MLEAC=2;MLEAF=0.500;MQ=58.29;NDA=4;QD=2.63;SOR=0.561	GT:AD:DP:GQ:PL	0/1:170,0:170:99:3390,0,6884	0/1:110,0:110:99:406,0,4461
NC_001133.9	253400	.	C	A	1047.13	PASS	AC=4;AF=1.00;AN=4;DP=281;ExcessHet=3.0103;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=55.70;NDA=2;QD=28.73;SOR=1.089	GT:AD:DP:GQ:PL	1/1:0,131:131:99:5060,393,0	1/1:0,141:141:99:5427,424,0
NC_001133.9	414775	.	G	T	1055.13	PASS	AC=4;AF=1.00;AN=4;DP=270;ExcessHet=3.0103;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=60.00;NDA=2;QD=30.97;SOR=0.740	GT:AD:DP:GQ:PL	1/1:0,113:113:99:4573,340,0	1/1:0,149:149:99:5996,447,0

A detailed vcf file interpretation can be referred to: http://gatkforums.broadinstitute.org/discussion/1268/how-should-i-interpret-vcf-files-produced-by-the-gatk.

SNP types are divided into two types, transition and transversion. The mutations between bases of the same type are called transitions, such as between purine and purine, or

between pyrimidine and pyrimidine, while the mutations between different types of bases are called transversion, such as the variation between purine and pyrimidine. In general, transition is more likely to occur than transversion, so the ratio of transition/transversion (Ti / Tv) is generally greater than 1, and the specific value is related to the species being measured. For a diploid or polyploid species, if a SNP site on the homologous chromosome is the same base, the SNP site is called a homozygous SNP site; if the SNP site on the homologous chromosome contains different types of bases, the SNP site is referred to as a heterzygous SNP site. The more the number of homozygous SNPs, the greater the difference between the sample and the reference genome, the more the number of heterozygous SNPs, the higher the degree of heterozygosity of the sample. The SNP detection results between the sample and the reference genome are shown below.

Table 6.Statistics of the detected SNPs

ID	SNP number	Transition	Transversion	Heterozygous	Homozygous
1a3	141	70	71	17	70
WT	129	65	64	8	67

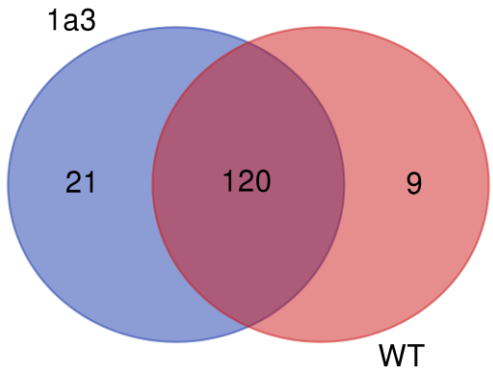

Figure 6. Shared SNP number between samples

3.3.2 The SNPs Detection of a Single Sample

The whole genome SNP mutations can be divided into 6 types. Taking T:A>C:G for example, this type of SNP mutant includes T>C and A>G. Since the sequenced reads can be mapped to both plus and minus strands of the reference genome, when T>C mutation occurs on the plus strand, at the meanwhile, the A>G mutant will occur on the minus strand in the same site, so T>C and A>G are considered as the same type. The SNP mutation type distribution is shown in the below figure.

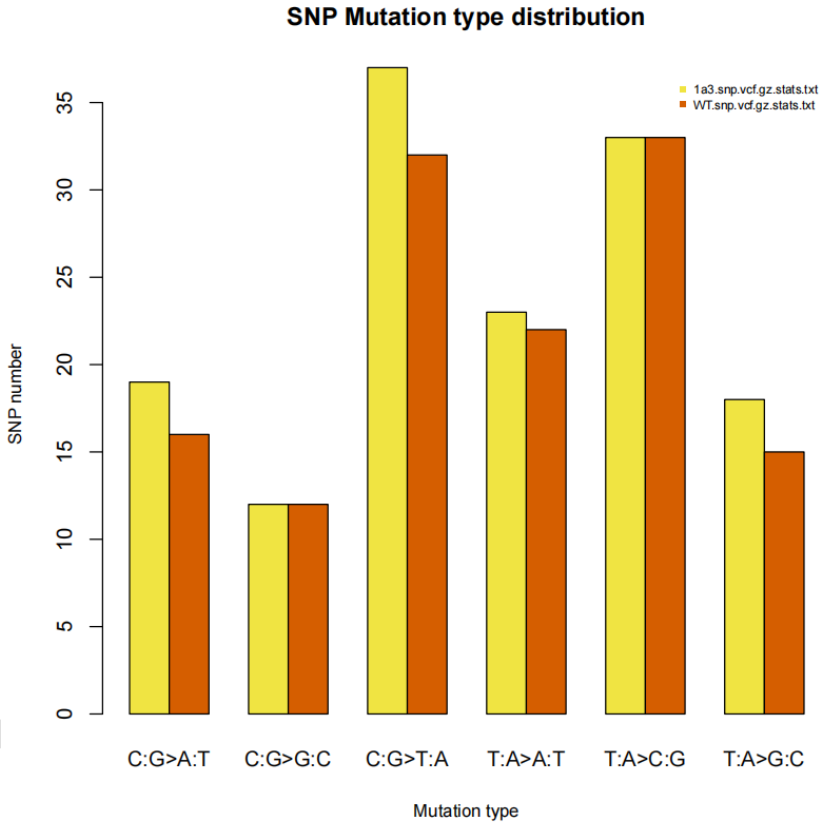

Figure 7. SNP mutation type distribution.

3.3.3 SNP Annotation.

SnpEff[3] is a software for annotating and prediction the effect of variants (SNP, Small InDel) on genes. Based on the position of the mutation site on the reference genome and the gene position information on the reference genome, the effect of the mutation site on the region of the genome (intergenic region, gene region or CDS region), and the effect of mutation (synonymous or non-synonymous mutations, etc.) can be obtained. This

software can use vcf format files as input and output. The output will add the following fields in the INFO column of the vcf file: ANN= Allele | Annotation | Annotation_Impact | Gene_Name | Gene_ID | Feature_Type | Feature_ID | Transcript_BioType | Rank | HGVS.c | HGVS.p | cDNA.pos / cDNA.length | CDS.pos / CDS.length | AA.pos / AA.length | Distance | ERRORS / WARNINGS / INFO. The detailed interpretations of the SnpEff results can refer to: http://snpeff.sourceforge.net/SnpEff_manual.html#output.

The statistics of SNP annotation results is displayed in the following figure:

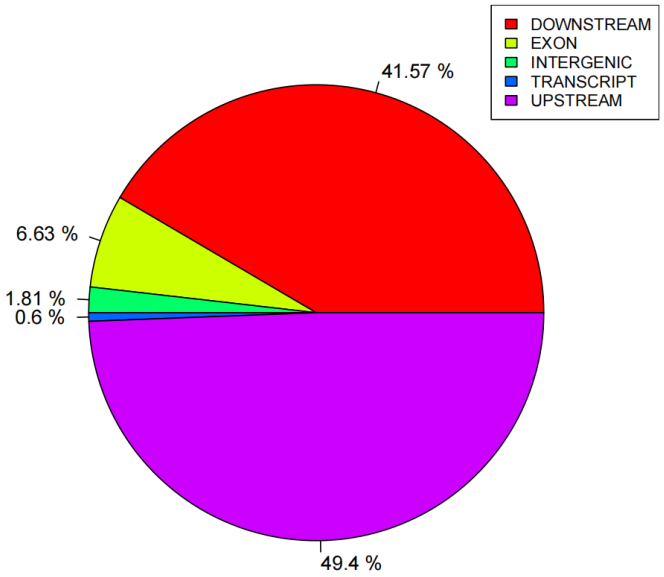

Figure 8. Statistics pie of SNP annotations.

3.4 Small InDel Detection and Annotation

3.4.1 Small InDel Detection Between Genome of Test Sample and Reference Genome

According to the mapping results of the clean reads on the reference genome, the presence or absence of Small InDel between the sample and the reference genome was examined. Again, the GATK was used to detect the insertion and deletion. Generally, the Small InDel mutations are less than SNP mutations, and the InDel in coding regions may cause frameshift mutation, leading to changes in gene function. The statistics of InDel in whole genome and CDS regions is demonstrated in the below table.

Table 7. Statistics of InDel in whole genome and CDS regions

Sample	CDS-Insertion	CDS-Deletion	CDS-Het	CDS-Homo	CDS-Total	Genome-Insertion	Genome-Deletion	Genome-Het	Genome-Homo	Genome-Total
1a3	24	15	8	14	39	219	131	215	68	350
WT	24	14	11	14	38	220	136	220	68	356

Het: Heterozygous Homo: Homozygous

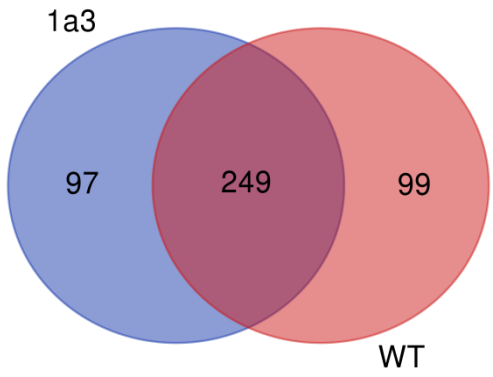

Figure 6. Shared InDel number between samples

The InDel length distribution in the whole genome scale and CDS regions is counted, the graph is displayed in the following figure.

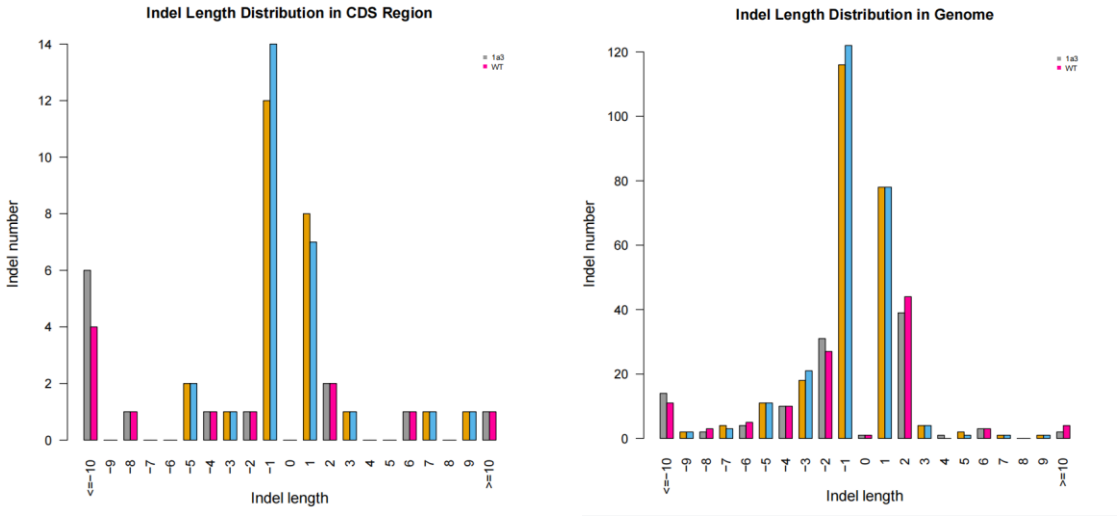

Figure 9. InDel length distribution in both the whole genome scale and CDS regions. The ordinate is the InDel length (10 bp or less), number with greater than 0 represents the insertion and less than 0 represents the deletion. Abscissa showing the corresponding number.

3.4.2 InDels Annotation.

The InDels of each sample is annotated by SnpEff[3]. The statistics of Indels annotation results is displayed in the following figure:

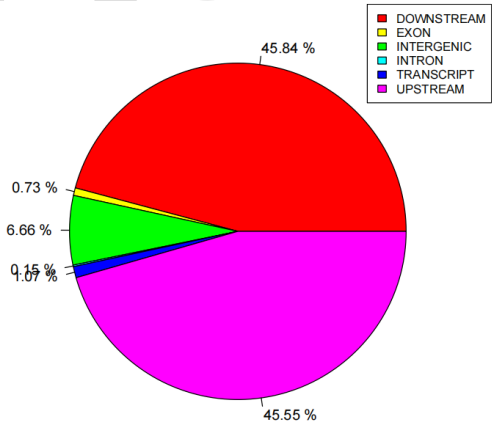

Figure 10. Statistics pie of InDel annotations.

3.5. SV Detection and Annotation

3.5.1 SV Detection.

SV (structural variation) refers to the insertion, deletion, inversion, translocation of large fragments at the genomic level. Structural variations are detected using breakDancer[4]. The breakDancer first obtains the size and variance of the insert of the library based on the alignment result of the sequence and the reference genome. Then, looking for possible structural variations by comparing the result of the aberration between the test sequence and the reference genome (the deviation of the insert, or the alignment orientation does not match). The SV result file contains the header row and the data line as follows:

Table 8. Part of SV result file

#Chr1	Pos1	Orientation1	Chr2	Pos2	Orientation2	Type	Size	Score	num_Reads	num_Reads_lib	Allele_frequency
NC_001133.9	1	20+83-	NC_001133.9	289	20+83-	ITX	-294	33	2	../02.align/1a3_align_sort_rm_dup.bam|2	NA
NC_001133.9	6894	2+0-	NC_001133.9	7254	2+2-	DEL	375	42	2	../02.align/1a3_align_sort_rm_dup.bam|2	2.15
NC_001133.9	8491	2+1-	NC_001133.9	8902	0+2-	DEL	363	49	2	../02.align/1a3_align_sort_rm_dup.bam|2	3.67
NC_001133.9	10349	2+0-	NC_001133.9	10759	1+4-	DEL	375	44	2	../02.align/1a3_align_sort_rm_dup.bam|2	3.62

Note: Columns 1-3 and 4-6 are used to specify the coordinates of the two SV breakpoints. The orientation is a string that records the number of reads mapped to the plus (+) or the minus (-) strand in the anchoring regions.

Column 7 is the type of SV detected: DEL (deletions), INS (insertion), INV (inversion), ITX (intra-chromosomal translocation), CTX (inter-chromosomal translocation), and Unknown.

Column 8 is the size of the SV in bp. It is meaningless for inter-chromosomal translocations.

Column 9 is the confidence score associated with the prediction.

Column 11 can be used to dissect the origin of the supporting read pairs, which is useful in pooled analysis. For example, one may want to give SVs that are supported by more than one libraries higher confidence than those detected in only one library. It can also be used to distinguish somatic events from the germline, i.e., those detected in only the tumor libraries versus those detected in both the tumor and the normal libraries.

Column 12 is currently a placeholder for displaying estimated allele frequency. The allele frequencies estimated in this version are not accurate and should not be trusted.

Using the breakDancer [4] software, based on the alignment of Pair-end reads to the reference genome and the actual Insert Size of the sample, the insertions (INS), deletions (DEL), inversion (INV), intra-chromosomal translocation (ITX), and inter-chromosomal translocation (CTX) are detected. The statistics of the detection of various types of SV is shown in the blow Table:

Table 9. Statistics of various types of SV

Sample ID	SV	INS	DEL	INV	ITX	CTX	UN
1a3	137	7	50	36	32	12	0
WT	103	0	24	28	42	9	0

3.5.2 SV Annotation

According to the positional information on the reference genome of the detected SV variation, the information such as the gene of the reference genome, the location of the CDS (usually in the gff file) can be used to note whether SV mutation occurs in the intergenic region, gene region or CDS region. The three types of structural variation, deletion (DEL), insertion (INS), and inversion (INV) were annotated, the statistical results are shown in the following table:

Table 10. Statistical results of three types of SV annotation

Sample ID	Type	Gene	CDS	Exon	Intron	RNA
1a3	DEL	36	51	21	0	20

1a3	INV	224	237	170	0	165
WT	DEL	7	7	8	0	7
WT	INV	197	254	142	0	137

3.6 CNV Detection

The sequencing depth of the sequenced reads was used to detect CNV by using cnvnator [5], and the distribution of Copy Number Gain and Loss on the reference genome was plotted. Part of the cnvnator result is shown in the following table:

Table 11. Statistical results of two types of CNV annotation

Sample ID	DEL	DUP
1a3	65	58
WT	47	46

Note: Due to an average depth coverage of the bam file about 100×, the bin size for depth detection method is set at 30bp according to cnvnator's publication. CNVs with q value greater than 0.5 or length less than 100bp were filtered out.

4. Software List

Software	URL
BWA	https://github.com/lh3/bwa
TBtools	https://github.com/search?q=TBtools&type=repositories
Picard	https://broadinstitute.github.io/picard/
SnpEff	https://pcingola.github.io/SnpEff/
BreakDancer	https://github.com/genome/breakdancer
CNVnator	https://github.com/abyzovlab/CNVnator

5. Database List

Database	URL
KEGG	https://www.kegg.jp/
GO	https://geneontology.org/
COG	https://www.ncbi.nlm.nih.gov/research/cog/
NR	https://www.ncbi.nlm.nih.gov/protein
SwissProt	https://www.uniprot.org/uniprotkb?query=SwissProt

6. Reference

1. Abuín JM, Pichel JC, Pena TF, Amigo J. BigBWA: approaching the Burrows-Wheeler aligner to Big Data technologies. Bioinformatics. 2015 Dec 15;31(24):4003-5.
2. Chen C, Chen H, Zhang Y, et al. TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. Mol Plant. 2020 Aug 3;13(8):1194-1202.
3. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3., Fly (Austin). 2012 Apr-Jun;6(2):80-92.
4. Fan X, Abbott TE, Larson D, Chen K. BreakDancer: Identification of Genomic Structural Variation from Paired-End Read Mapping. Curr Protoc Bioinformatics. 2014;45:15.6.1-11.

5. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011 Jun;21(6):974-84.