XXXX PROJECT REPORT

Microbe team

xxxx.xx.xx

CATALOG

Ⅰ INTRODUCTION OF WORKFLOW

1 Experiment procedure

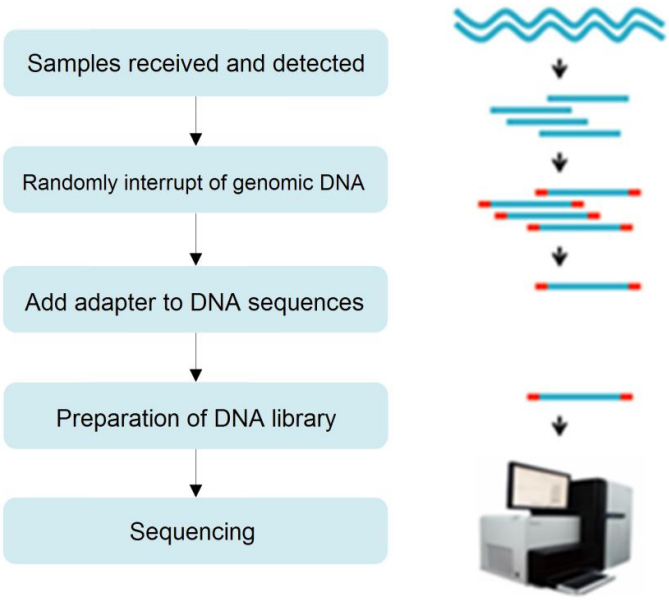

Figure 1-1 Experiment procedure. After the DNA samples were received, the sample for testing would be conducted. Then was the library construction with qualified samples: First, the DNA samples were randomly fragmented to produce DNA fragments of the desired length, followed by the end repair and the adapter ligation. Finally, qualified library were sequenced and clusters were prepared.

2 Bioinformatics analysis pipeline

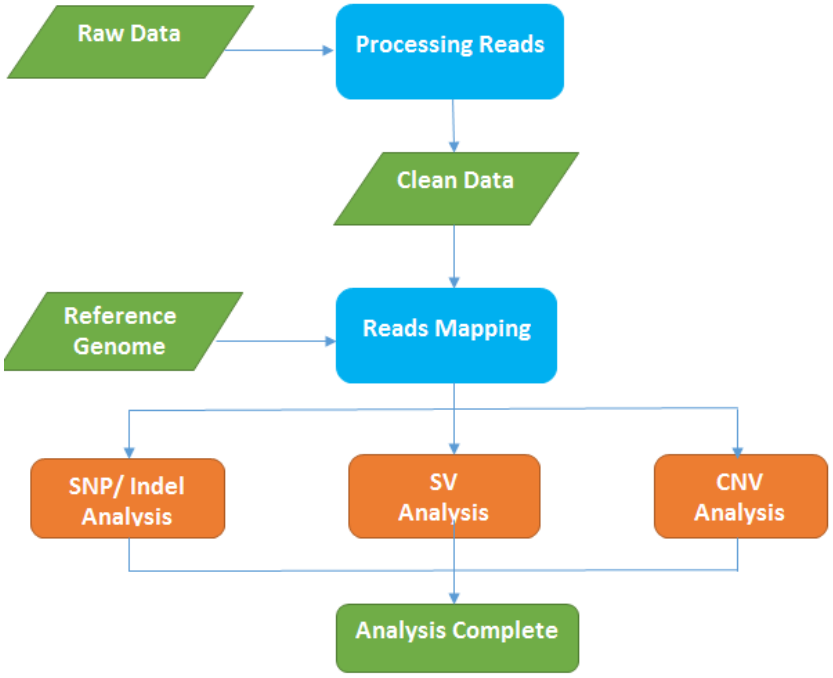

Figure 1-2 Bioinformatics analysis pipeline comparison of mutation detection based on reads First, the original data was filtered to compare the high-quality data(Clean Data) with the reference sequence; Secondly, SNP, InDel and SV of the sample were detected according to the comparison, and the result of SNP, InDel were commented. For the multi-samples, the species evolutionary relationships of the sample could be analyzed through SNP results.

Note: This flow chart included all the analysis of the product and the project-specific analysis contents of this report shall prevail.

II RESULTS

1 Data overview

Samples were sequenced by Hiseq2000 sequencing platform. The following table showed the detailed statistics of the data

Figure 2-1 Statistics of the sequencing data

Sample ID	Insert size (bp)	Reads length (bp)	Raw data (Mb)	Filteredreads (%)	Clean data (Mb)	Clean data Q20(%)	Clean data Q30(%)

Note: Insert size: Insert length, Reads length: Length of sequencing reads, Raw data: The amount of raw data, Filtered reads: the percentage of filtered Reads, Clean data: valid data, Clean data Q20: Q20 value of valid data, Clean data Q30: Q30 value of valid data.

Base content and quality distribution of valid data obtained were shown below

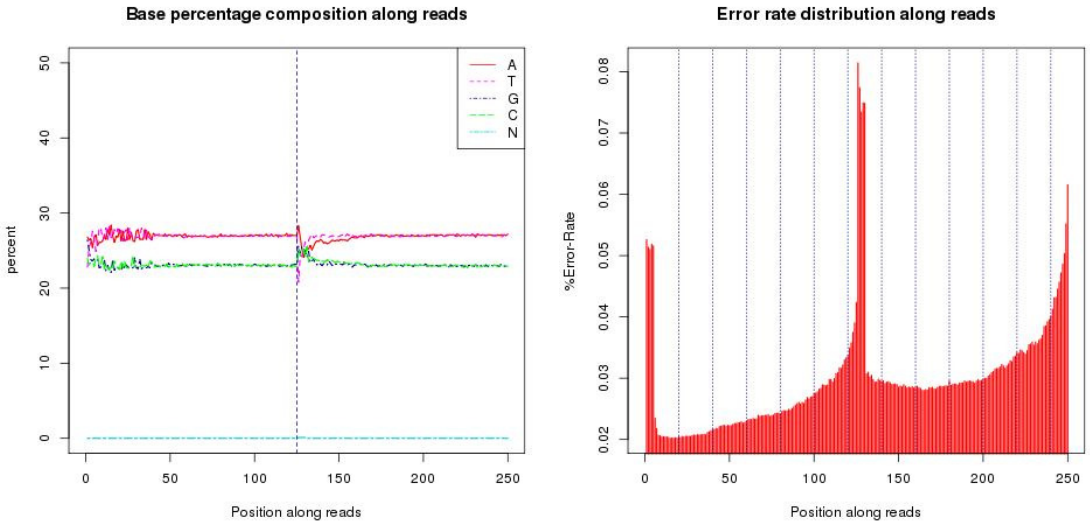

Figure 2-1 Sample, ***bp The base content and quality distribution of library distribution of base content**

The horizontal axis represents the base composition along reads and the vertical axis represents the percentage of each base on a position; Quality distribution: The horizontal axis represents the base composition along reads and the vertical axis represents the quality distribution of the base sequenced.

2 Reference genome comparison

The coverage was completed through comparing the sequencing data ****** to the reference genome by using SAMTOOLS and BWA, determining the relatives case between the sample and the reference sequence. The statistics of sequencing and coverage depth distribution were shown in Table 2-2 and Figure 2-2.

Table 2-2 Statistics of the sequencing depth and coverage

Ref_ID	Sample ID	Mapping rate (%)	Average sequencing depth	Coverage (%) ≥			
				1X	4X	10X	20X

Note: From the left to the right is the reference sequence ID, sample ID, alignment rate of the data to the reference sequence, average sequencing depth, and the coverage of sequencing depth which exceeds a certain value.

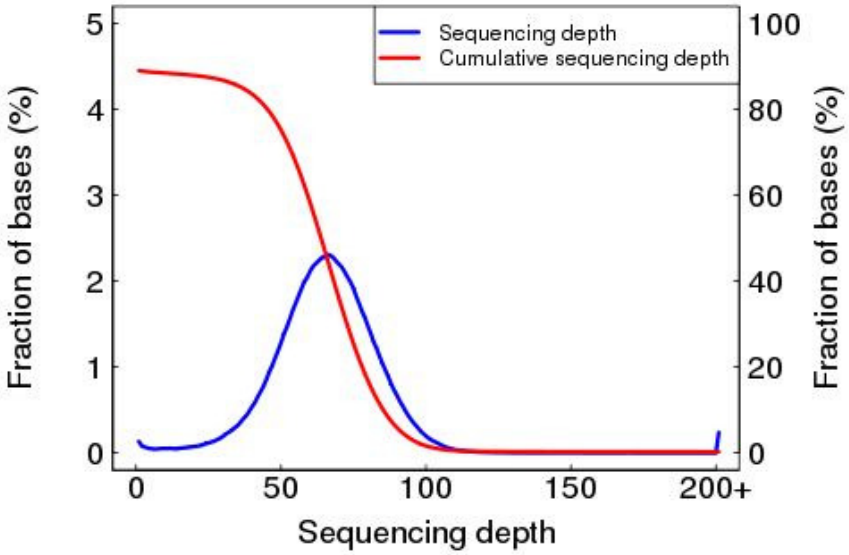

Figure 2-2 sample* Sequencing depth distribution**

The horizontal represents the sequencing depth of the reference sequence of each site. The left vertical axis corresponding to Sequencing depth statistics, represents the ratio statistics of different sequencing depth; The right vertical axis corresponding to Cumulative sequencing depth statistics represents the accumulated value distribution of the proportion of the sequencing depth, from low to high.

3 Statistics of SNP/InDel detection

3.1 Statistics of SNP detection

SNP（single nucleotide polymorphism）SNP mainly refers to the DNA sequence pol ymorphisms caused by a single nucleotide mutation variation at the genomic level, includ ing a single base transition and transversion.

SAMTOOLS was used to conduct individual SNP detection. Transitions and Transversions ratio of the whole genome and coding region, hybrid ratio, and the number of SNP stat istics were shown in Table 2-3.

Table 2-3 Statistics of SNP detection

Sample ID	Ts	Tv	Ts/Tv	Het	Hom	Het rate %	Total

Note: Ts: Transition, Tv: transversion, Ts/Tv: ratio of Transitions to Transversions, Het: hybrid SNP, Hom: Homozygous SNP, Het rate: Het SNP/Total Genome Length.

3.2 Statistics of InDel detection

InDel refers to the insertion and deletion of small fragments of genomic sequences. Small fragment insertion and deletion with a length less than 50 bp was detected by SAMTOOLS. Insert, deletions, hybrid ratio, the number of InDel statistics of the whole genome and coding regions were shown in table2-4.

Table 2-4 Statistics of the InDel detection

Samples ID	Insertion	Deletion	Het	Hom	Het rate (%)	Total

Note: Het: hybrid InDel, Hom: Homozygous InDel, Het rate: Het InDel/total Genome Length.

3.3 SNP/InDel distribution over the genome

The distribution of SNP / InDel of all samples on the reference genome sequence was shown below:

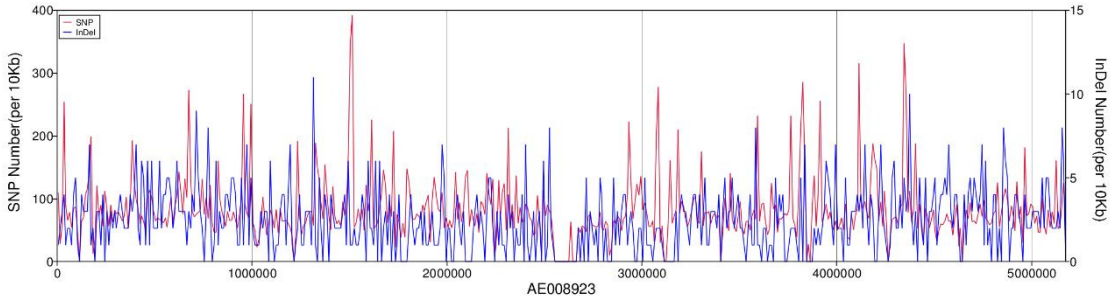

Figure 2-3 sample * SNP/InDel distribution over the genome.** The horizontal axis represents the chromosome of the reference sequence and the vertical axis represents the number of SNP /InDel per 10kb region sequence. The left side of the scale shows the SNP and the right shows the InDel.

4 SNP/InDel annotation

4.1 SNP annotation

Function of SNPs was annotated based on the position on gene structure, and genetic coding transformation affected by these SNPs. The result was shown in the table below:

Table 2-5 Statistics of the SNP annotation

Samples ID	Non-Syn	Syn	Intergenic

Note: Non-Syn, Nonsynonymous mutations in the CDS region; Syn, Synonymous mutations in the CDS region; Intergenic, mutation located between the gene district region.

Table 2-5 Statistics of the SNP annotation

Samples ID	5-UTR	Non-Syn	Syn	3-UTR	Intron	Intergenic

Note: SNP of 5-UTR, 5-UTR region; Non-Syn, Nonsynonymous mutation in the CDS region; Syn, Synonymous mutations in the CDS region; SNP of 3-UTR, 3-UTR region; Intron, number of SNP located in intron regions; Intergenic, mutation located between the gene district region.

4.2 InDel annotation

InDel annotation results were shown in the table below:

Table 2-6 Statistics of the InDel annotation

Samples ID	Non-Shift	Shift	Intergenic

Note: Non-shift, InDel in the CDS region which cause no frame shift; Shift, InDel in the CDS region which cause frame shift; Intergenic: mutation located between the gene district.

Table 2-6 Statistics of the InDel annotation

Samples ID	5-UTR	Non-Shift	Shift	3-UTR	Intron	Intergenic

Note: SNP of 5-UTR, 5-UTR region; Non-shift, InDel in the CDS region which cause no frame shift; Shift, In Del in the CDS region which cause frame shift; 3-UTR, SNP of 3-UTR region; Intron, InDel number located in intron regions; Intergenic, mutation located between the gene district.

5 SV Detection and annotation

SV (Structural Variation) refers to the genomic insertions, deletions, inversions, translocations of large fragments at the genomic level. BreakDancer software was used to detect INS(insertion), DEL(deletion), INV(inversion), ITX(intra-chromosomal translocation) and CTX(inter-chromosomal translocation).

5.1 SV detection

Structural variation was detected by using BreakDancer. The statistics of the SV det ected were shown in the following table. The SV length distribution was also shown below.

Table 2-7 Statistics of the SV detection

Samples ID	INS	DEL	INV	ITX	CTX	Unknown	Total

Note: From the left to the right is: samples ID, the number of Insert, number of deletions, number of inversions, number of intra-chromosomal translocation, number of inter-chromosomal translocation.

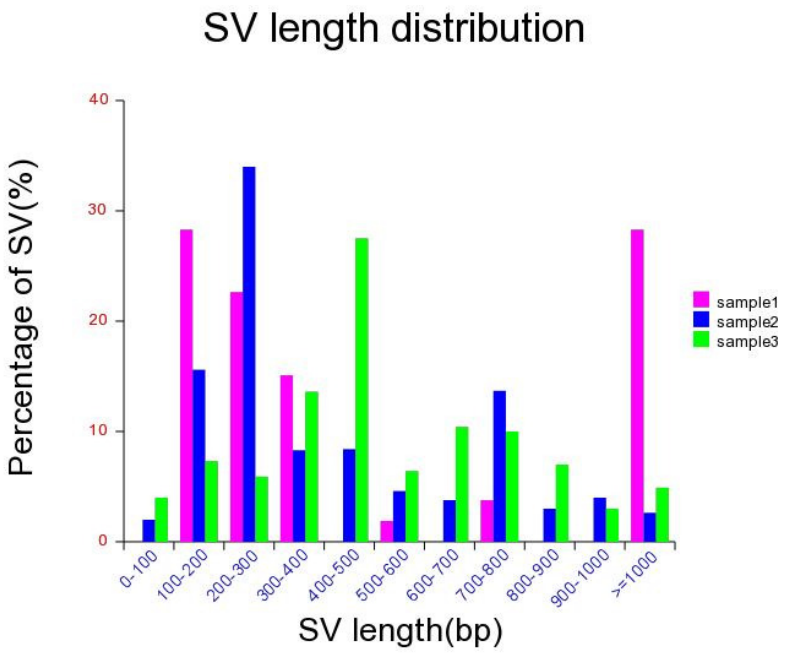

Figure 2-4 SV length distribution The horizontal axis represents the length of the SV and the vertical axis represents the ratio of the SV.

5.2 SV annotation

DEL, INS, INV Annotation, results were shown in the table below:

Table 2-8 Statistics of the SV annotation

Samples ID	Gene region	Intergenic

Note: Gene region: mutation located between the coding gene region. Intergenic: mutation located between the gene region.

6 CNV detection and annotation

CNV Copy number variation is caused by a genomic rearrangement. Usually it refers to the increasing or decreasing of copy number of the large fragments of the genome of which the length is larger than 1 kb, mainly for the submicroscopic level of repeat (DUP) and deletions (DEL).

6.1 CNV detection

Mutation of CNV was detected by using CNVnator. The statistics of the CNV detected were shown in the following table:

Table 2-9 Statistics of the CNV detection

Samples ID	DEL Number	DEL Length	DUP Number	DUP Length

Note: From left to right is: samples ID, number of DEL, total length of DEL, number of DUP, total length of DUP.

6.2 CNV annotation

Annotation results for DEL, DUP of CNV were shown in the table below:

Table 2-9 Statistics of the CNV detection

Samples ID	Gene region DEL	intergenic DEL	Gene region DUP	intergenic DUP

Note: Gene region: mutation located between the coding gene region. Intergenic: mutation located between the gene region.

7 Genome-wide variation map

The reads coverage of the reference sequence for each sample, analytical results of SNP/InDel, reads coverage and distribution of SNP/InDel were shown in the annular FIG

by the Circos which directly compared the reads coverage and variation distribution from the map.

Results were shown below:

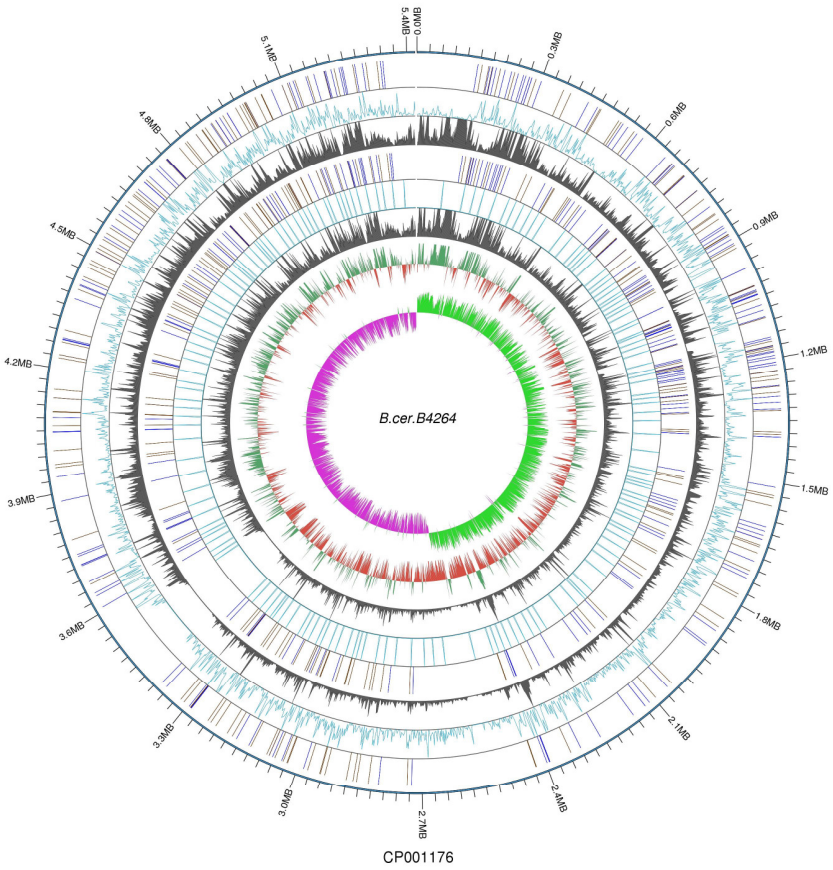

Figure 2-5 Genome-wide variation map based on the reference sequence. The outer ring shows the reference sequence position coordinates. From outside to inside is the distribution of InDel, SNP distribution, reads coverage depth, the genomic GC content of the reference sequence, and the genome GC skew value distribution of reference sequence. For multiple samples, there will be multiple sets of "InDel distribution, SNP distribution, reads coverage depth".

METHODS DESCRIBED

The analysis method was shown below and it should be updated according to actual conditions.

1 Data processing:

The original data obtained by high-throughput sequencing (for example, the platform of Illumina HiSeq TM2000/MiSeq, etc.) was transformed into raw reads (raw data, or raw reads) by CASAVA base calling, and stored in FASTQ (fq) format, containing sequencing information and the corresponding sequencing quality information of the reads.

The sequenced data was filtered and the reads containing adapter and low quality data were removed to obtain the clean data used for subsequent analysis.

2 Reads mapping analysis:

The reads comparison is the basis of the resequencing analysis. The variation information of the sample and the reference was obtained by aligning the sample reads with th e designated reference.

The BWA software was used to map the reads to the reference sequence, counting the coverage of the reference sequence to the reads. The SAMTOOLS software was used o to do the alignment.

3 SNP/InDel analysis:

SNP (single nucleotide polymorphism) mainly refers to the DNA sequence polymorphism caused by the single nucleotide variation at the genome level, including transition, transversion, etc. InDel refers to the insertion and deletion of small fragments of the genome.

SAMTOOLS was used to detect the individual SNP and insertion and deletion of small fragments(<50bp), as well as the analysis of SNP/InDel in the functional regions of the genome.

4 SV analysis:

SV (structural variation) refers to the insertion, deletion, inversion and translocation of the large segments in the genome level.

The insertion (INS), deletion (DEL), inversion (INV), intra-chromosomal translocation (ITX), and inter-chromosomal translocation (CTX) between the reference and the sample were found by BreakDancer software.