Reseq Demo Report

01-Aug-2017

1 INTRODUCTION OF WORKFLOW

1.1 Experiment Process

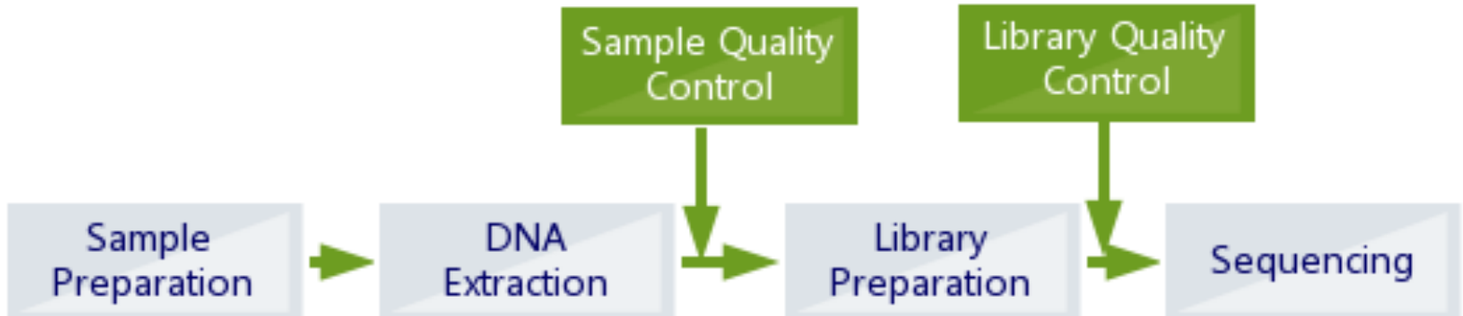

Figure 1-1 Experiment process

After the DNA sample(s) was(were) delivered, we did a sample quality test first. Then we used this(those) qualified DNA sample(s) to construct library： DNA sample is sheared into smaller fragments with a desired size firstly. Then adding an ‘A’ base to the 3' end of the blunt phosphorylated DNA fragments, adapters are ligated to the ends of the DNA fragments. At last, the qualified library would be used for sequencing.

1.2 Pipeline of Bioinformatics Analysis

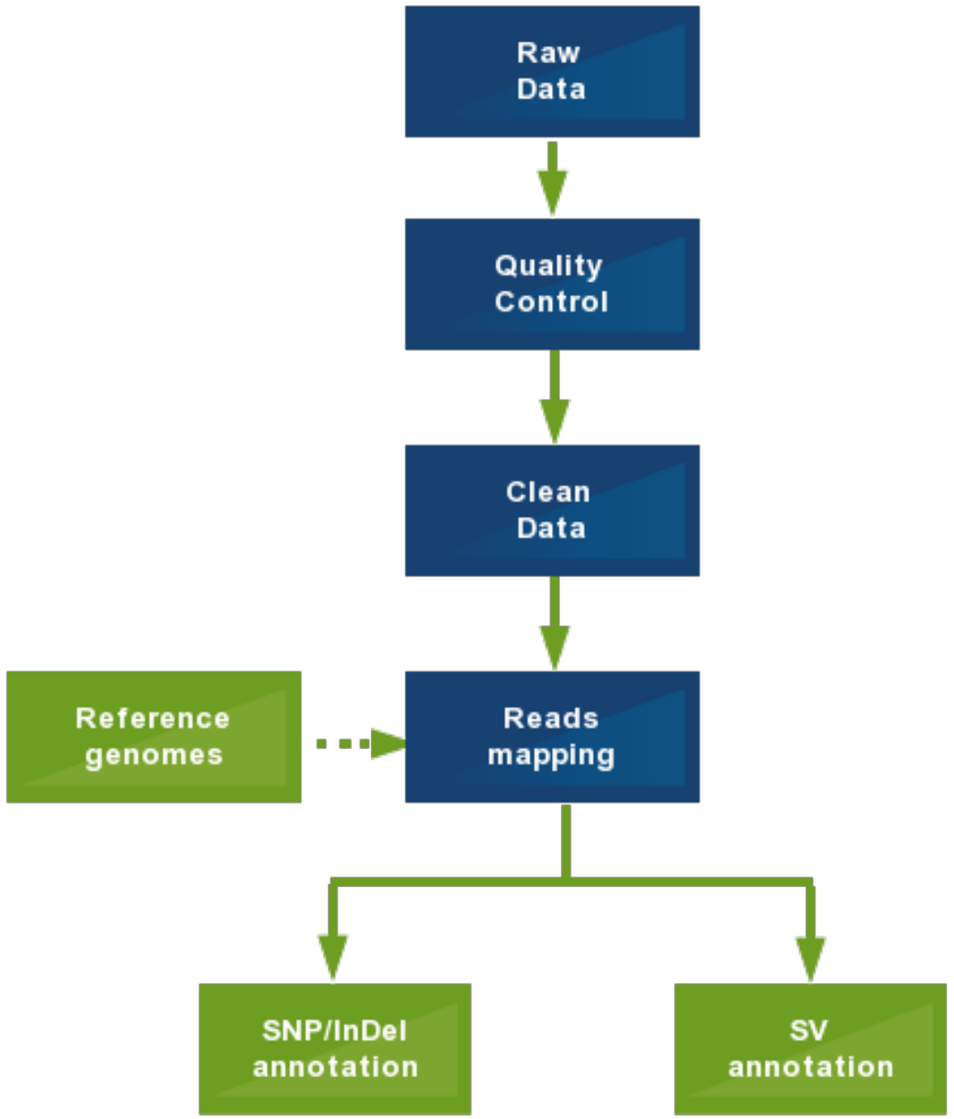

Figure 1-2 Pipeline of Bioinformatics Analysis（based on reads mapping）

Pipeline of Bioinformatics Analysis：

1 Raw data filtering Low quality reads will be filtered and generate Clean Data.

2 Reference genome comparison Accroding to BWA alignment，comparing the high-quality data(Clean Data) with the reference sequence.

3 SNP、InDel、SV Analysis Detecting SNP, InDel and SV of the sample according to the situation of comparison, and commenting the result of SNP,InDel,SV.

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

2 2 ANALYSIS RESULTS

2.1 Sequence Data Statistics

Samples were sequenced by using Hiseq4000 sequencing platform，the following table shows detailed statistics of the sample data：

Table 2-1 Sequence Data Statistics

Sequencing data statistics

Sample_Name	Insert_Size(bp)	Reads_length(bp)	AllRaw_data(Mbp)	Tol_Filtered(%)	Clean_Data(Mbp)	Clean_data_Q20(%)	Clean_data_Q30(%)	Clean_data_GC(%)
BA69.2	350	150:150	1,409.54	9.97	1,269.03	96.26	90.60	46.57
BM70	350	150:150	1,370.87	9.52	1,240.39	96.27	90.58	46.49
BS22.2	350	150:150	1,422.25	10.24	1,276.60	96.21	90.47	43.92
CX12	350	150:150	1,134.54	9.60	1,025.60	96.41	90.91	43.33
TGR2A	350	150:150	2,555.70	10.82	2,279.26	96.37	90.62	66.27

Page 1 of 1 10

View 1 - 5 of 5

Note Insert size：the length of insert fragment；Reads length：the length of reads；ALLRaw data：raw data size；Tol_Filtered(%)：the ratio of filtered reads；Clean data：clean data size；Clean data Q20：Q20 value of clean data； Clean data Q30：Q30 value of clean data；Clean data GC：GC content of clean data；

Distribution of base content and quality of clean data is shown below：

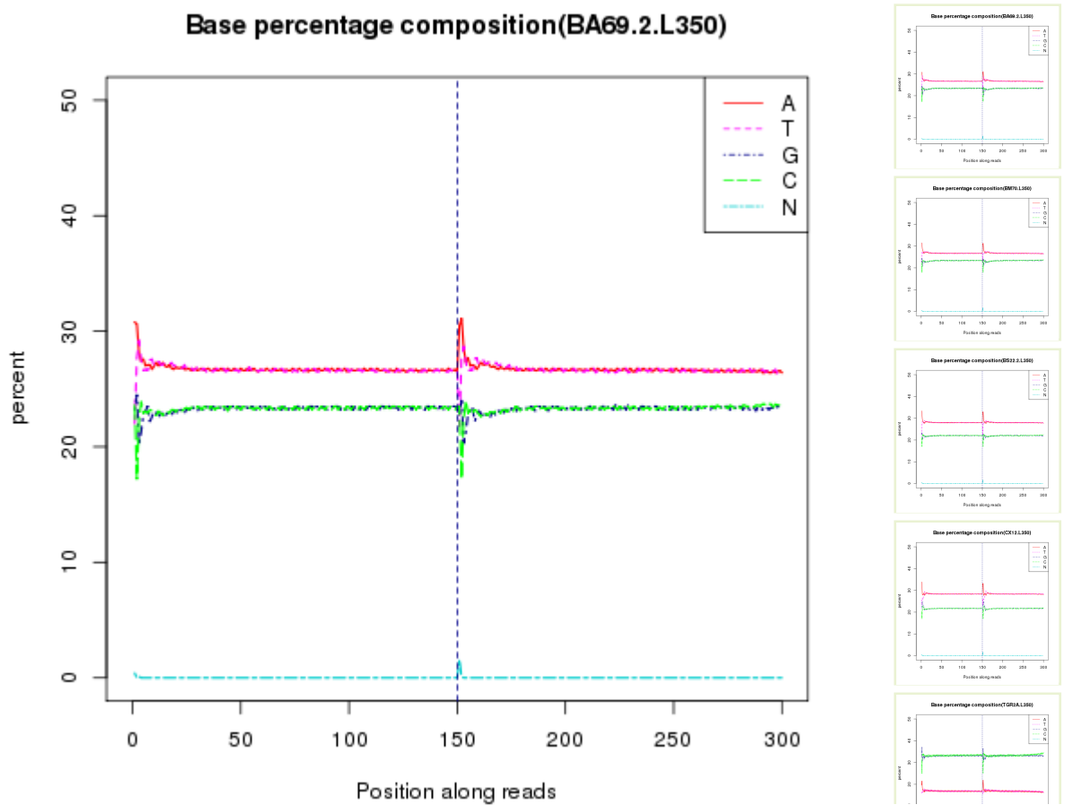

Figure 2-1-1 Distribution of base content

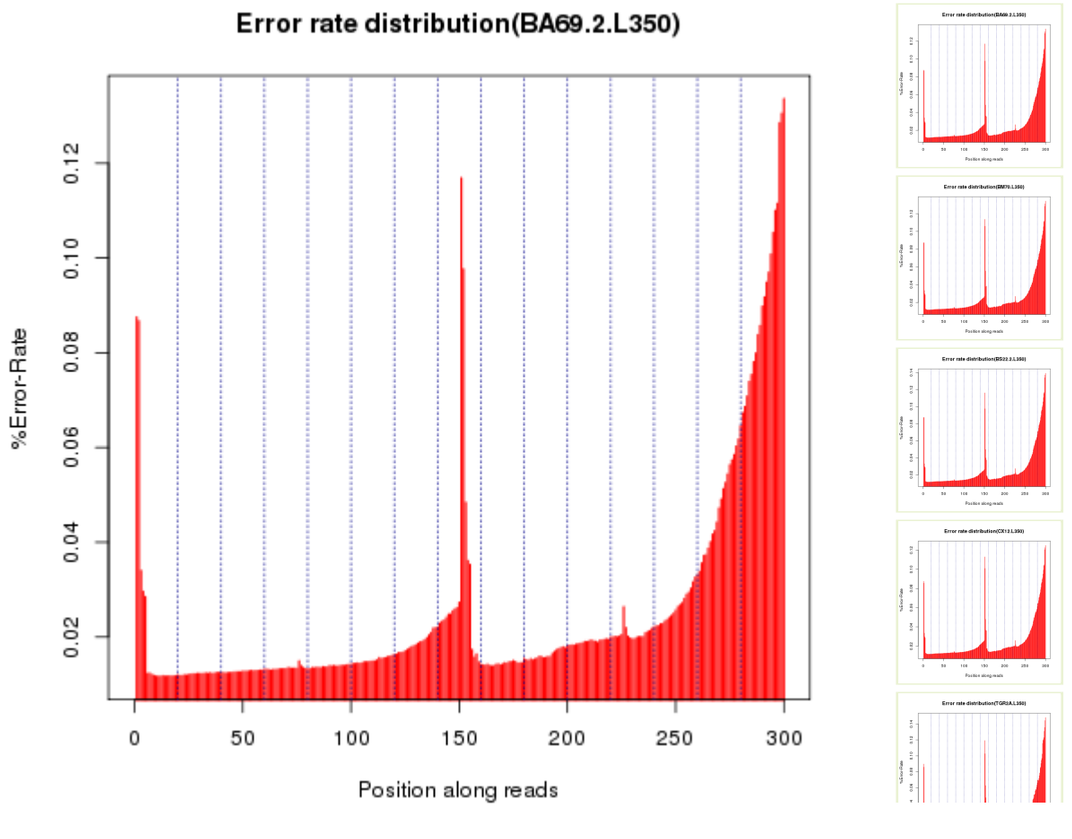

Figure 2-1-2 Distribution of base quality

Note Distribution of base content: X-coordinate is base site of reads. Y-coordinate is base content of every site. Distribution of quality: X-coordinate is base site of reads. Y-coordinate is average error rate of every site.

2.2 Reference genome comparison

The coverage is completed through comparing the sequencing data to reference genome by using SAMTOOLS and BWA, and determining the relatives case between the sample and the reference sequence。Statistics in Table 2-2, Cover depth distribution shown in Figure 2-2.

Table 2-2 Statistics of sequencing depth and coverage

Sequencing depth statistics

ref_name	query_name	avg_depth	coverage>=1X	coverage>=4X	coverage>=10X	coverage>=20X	map_rate	mismatch_rate	total_base	total_map_base	total_mismatch
BMV	BM70	273	92.49	92.34	92.28	92.24	87.21	2.33	1240385100	1081747656	25213494
PAPAO1	TGR2A	334	96.16	96.13	96.11	96.07	92.04	0.86	2279258400	2097801036	18109164
BADSM7	BA69.2	288	92.66	92.41	92.36	92.31	89.95	2.30	1269027300	1141540477	26229473
BS168	BS22.2	285	92.86	92.76	92.72	92.69	93.51	1.34	1276601700	1193739415	16010135
BS	CX12	222	94.81	94.64	94.60	94.56	87.66	1.40	1025604900	899079646	12619454

Page 1 of 1 10

View 1 - 5 of 5

Note There are reference sequence ID, sample ID, alignments rate of sample sequencing data to the reference sequence, average sequencing depth, and the coverage of sequencing depth that more than a certain value.

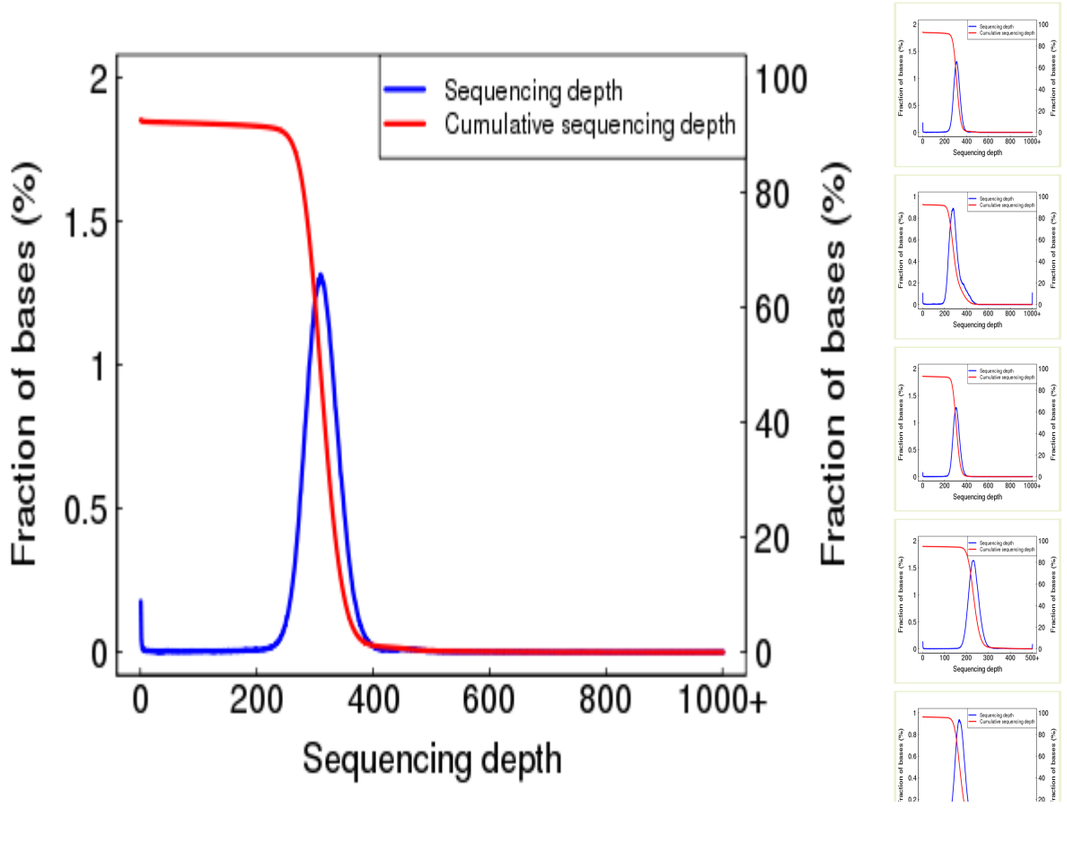

Figure 2-2 Sequencing depth distribution

Note The horizontal represents the sequencing depth of the reference sequence of each sites, Left vertical axis corresponding to Sequencing depth statistics, representing the ratio statistics of different sequencing depth; Right vertical axis corresponding to Cumulative sequencing depth statistics, it represents the accumulated value distribution of the proportion of the sequencing depth, from low to high depth.

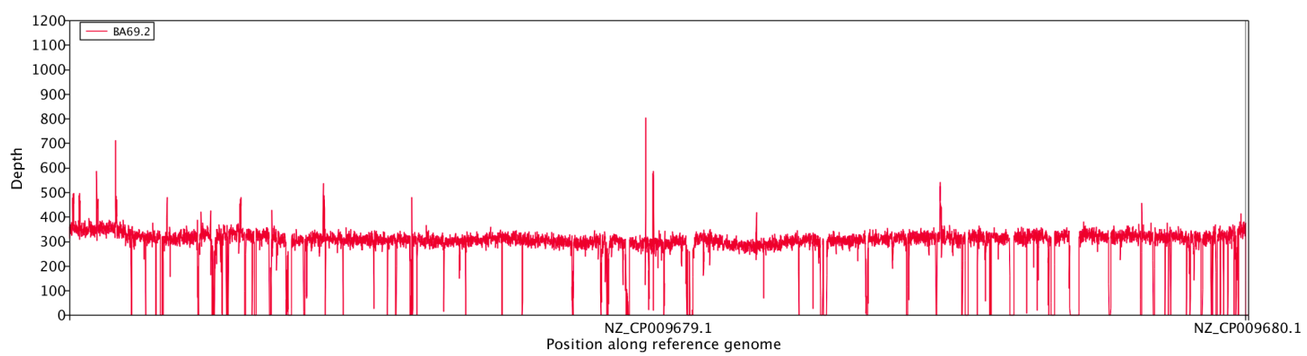

Figure 2-2-1Sequencing depth distribution on BADSM7. The X-coordinate was the chrom or scaffold of the Reference; The Y-coordinate was the average coverage depth of corresponding statistic region(when the genom size was samaller than 10M, the statistic region was 500bp, otherwise was 5Kb).(Click)

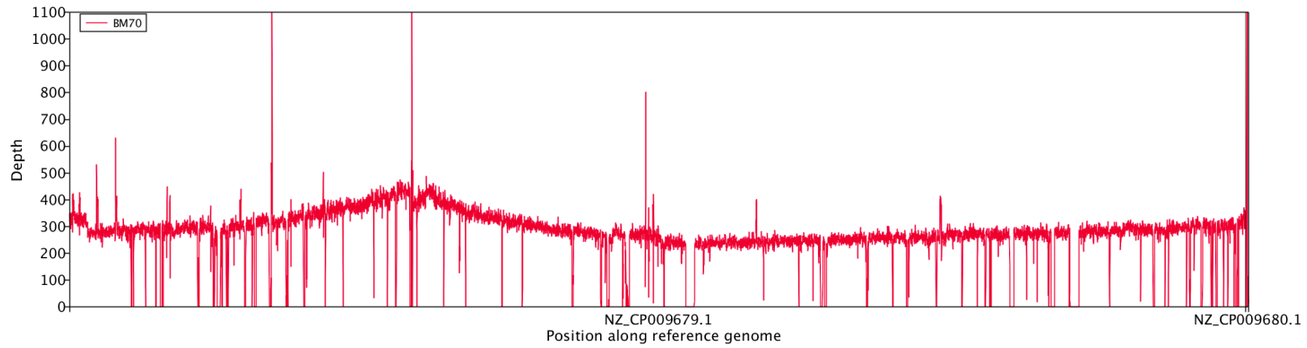

Figure 2-2-2Sequencing depth distribution on BMV. The X-coordinate was the chrom or scaffold of the Reference; The Y-coordinate was the average coverage depth of corresponding statistic region(when the genom size was samaller than 10M, the statistic region was 500bp, otherwise was 5Kb).(Click)

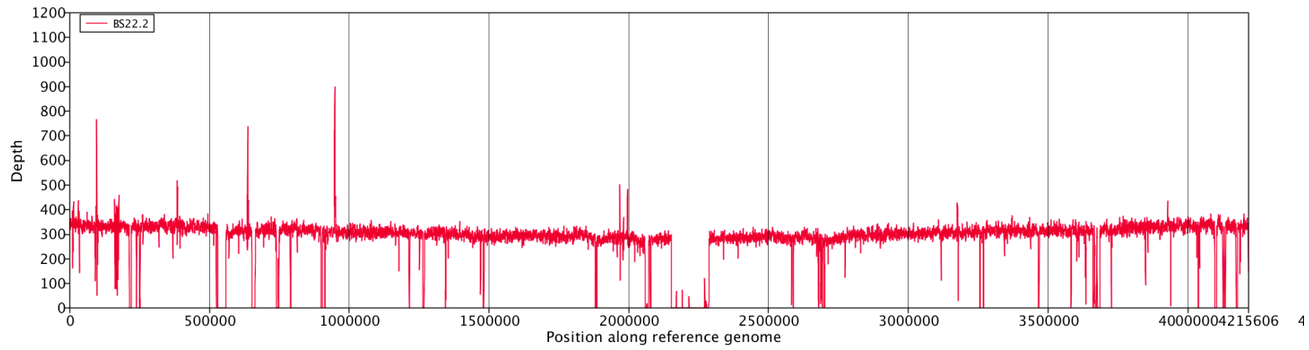

Figure 2-2-3Sequencing depth distribution on BS168. The X-coordinate was the chrom or scaffold of the Reference; The Y-coordinate was the average coverage depth of corresponding statistic region(when the genom size was samaller than 10M, the statistic region was 500bp, otherwise was 5Kb).(Click)

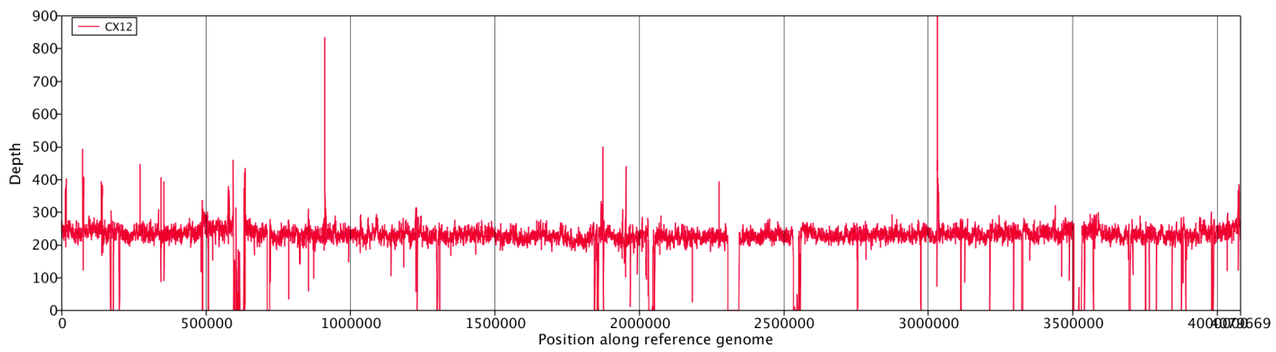

Figure 2-2-4Sequencing depth distribution on BS. The X-coordinate was the chrom or scaffold of the Reference; The Y-coordinate was the average coverage depth of corresponding statistic region(when the genom size was samaller than 10M, the statistic region was 500bp, otherwise was 5Kb).(Click)

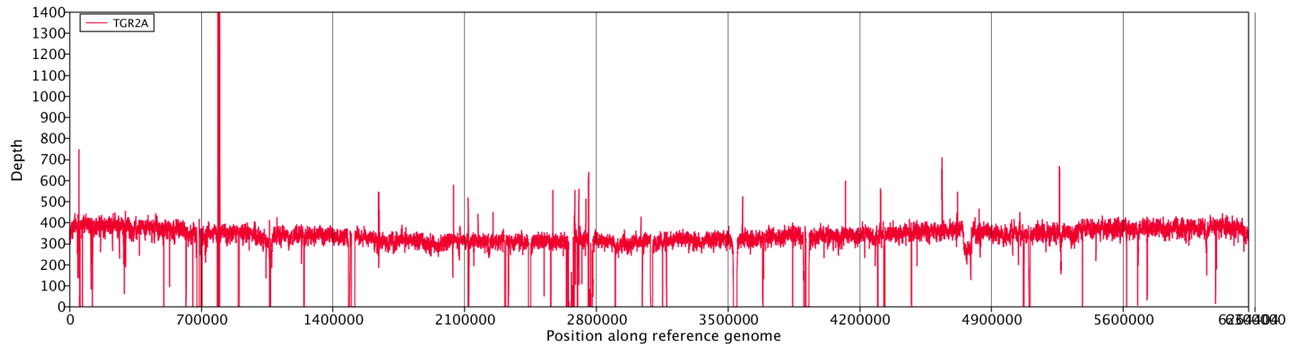

Figure 2-2-5Sequencing depth distribution on PAPAO1. The X-coordinate was the chrom or scaffold of the Reference; The Y-coordinate was the average coverage depth of corresponding statistic region(when the genom size was samaller than 10M, the statistic region was 500bp, otherwise was 5Kb).(Click)

2.3 SNP、InDel Detection and Statistics

2.3.1 SNP Detection and Statistics

SNP（single nucleotide polymorphism）SNP mainly refers to the DNA sequence polymorphisms caused by a single nucleotide mutation variation at the genomic level, including a single base transitions, transversions. We conduct individual SNP detection by using SAMTOOLS, Transitions and Transversions ratio of the whole genome and coding region, hybrid ratio, the number of SNP statistics in Table 2-3.

Table 2-3 SNP Detection results Statistics

SNP Detection and Statistics

Reference	Sample	ts	tv	ts/tv	Het	Hom	HetRate(%)	Total	Density(SNP/Kb)
BS168	BS22.2	26,171	12,081	2.17	148	38,104	0	38,252	9.07
BS	CX12	27,910	12,779	2.18	182	40,507	0	40,689	9.97
BADSM7	BA69.2	48,262	22,831	2.11	317	70,776	0.01	71,093	17.71
BMV	BM70	48,213	22,821	2.11	307	70,727	0.01	71,034	17.69
PAPAO1	TGR2A	21,090	6,481	3.25	124	27,447	0	27,571	4.40

Page 1 of 1 10

View 1 - 5 of 5

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

2.3.2 Indel Detection and Statistics

InDel refers to the insertion and deletion of small fragments of genomic sequences。Small fragment insertion and deletion that length less than 50 bp is detected by using SAMTOOLS, Insert, deletions, hybrid ratio, the number of InDel statistics of the whole genome and coding regions as shown in table 2-4.

Table 2-4 InDel Detection results Statistics

INDEL Detection and Statistics

Reference	Sample	Insertion	Deletion	Het	Hom	HetRate(%)	Total
BS168	BS22.2	129	110	1	238	0	239
BS	CX12	358	373	7	724	0	731
BADSM7	BA69.2	178	200	4	374	0	378
BMV	BM70	307	300	6	601	0	607
PAPAO1	TGR2A	36	41	2	75	0	77

	Page 1 of 1 10	View 1 - 5 of 5

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

2.3.3 SNP、InDel distribution over the genome

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

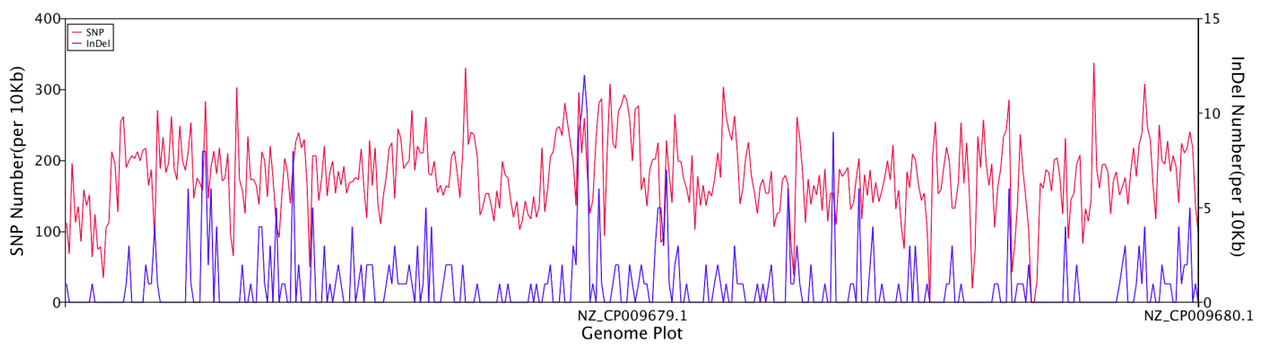

Figure 2-3-1 sample BADSM7_BA69.2_SNP_InDel SNP/InDel distribution over the genome.The horizontal listed chromosome of reference sequence, the vertical axis represents the number of SNP /InDel per 10kb region sequence, left ruler for the SNP, the right side of the scale is InDel.(Click)

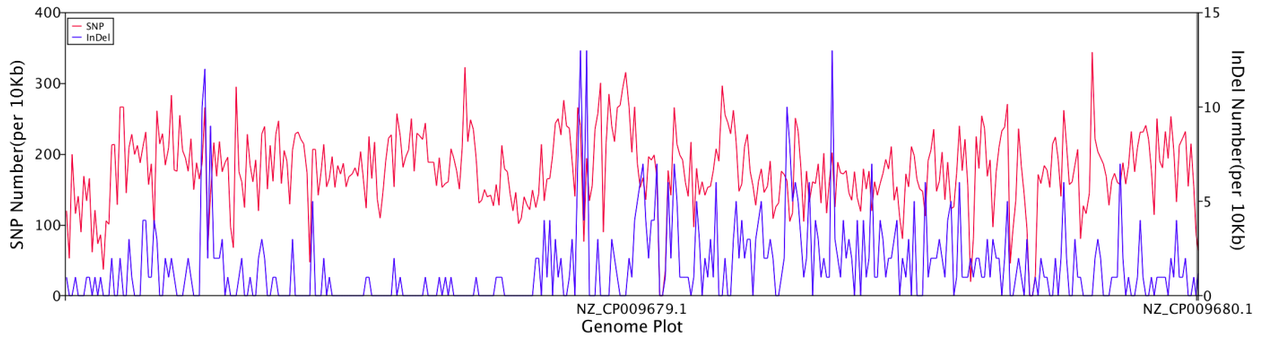

Figure 2-3-2 sample BMV_BM70_SNP_InDel SNP/InDel distribution over the genome.The horizontal listed chromosome of reference sequence, the vertical axis represents the number of SNP /InDel per 10kb region sequence, left ruler for the SNP, the right side of the scale is InDel.(Click)

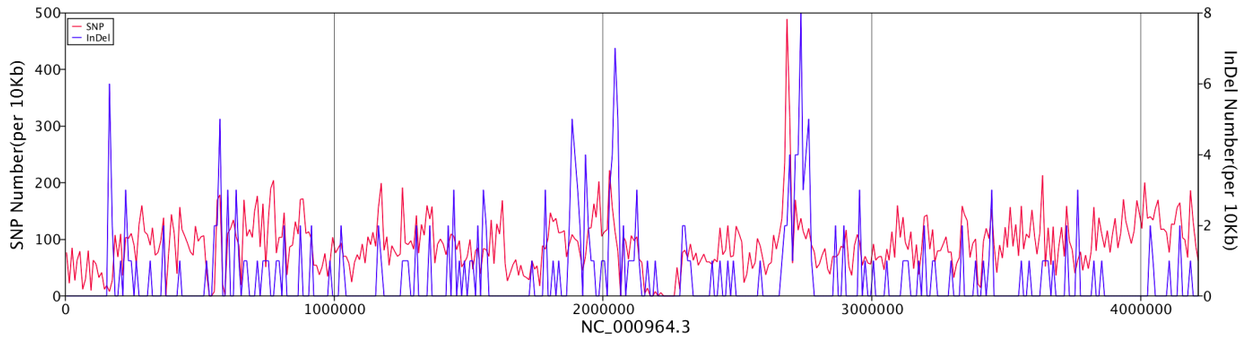

Figure 2-3-3 sample BS168_BS22.2_SNP_InDel SNP/InDel distribution over the genome.The horizontal listed chromosome of reference sequence, the vertical axis represents the number of SNP /InDel per 10kb region sequence, left ruler for the SNP, the right side of the scale is InDel.(Click)

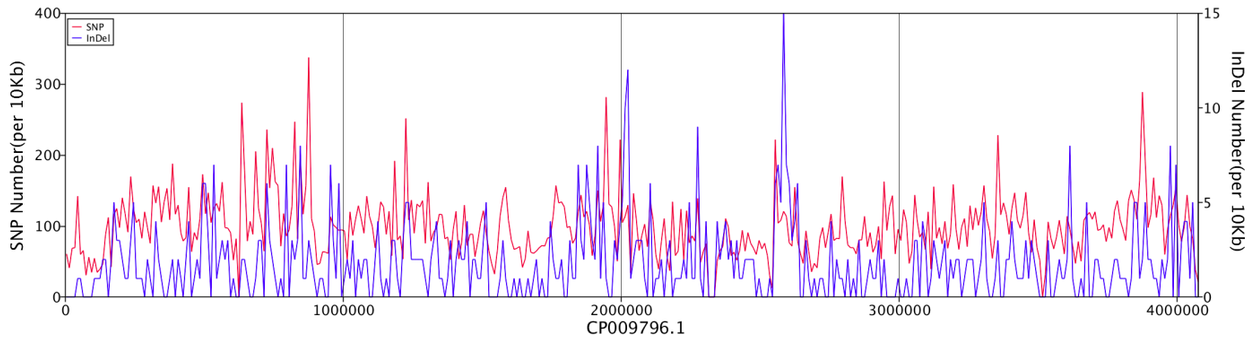

Figure 2-3-4 sample BS_CX12_SNP_InDel SNP/InDel distribution over the genome.The horizontal listed chromosome of reference sequence, the vertical axis represents the number of SNP /InDel per 10kb region sequence, left ruler for the SNP, the right side of the scale is InDel.(Click)

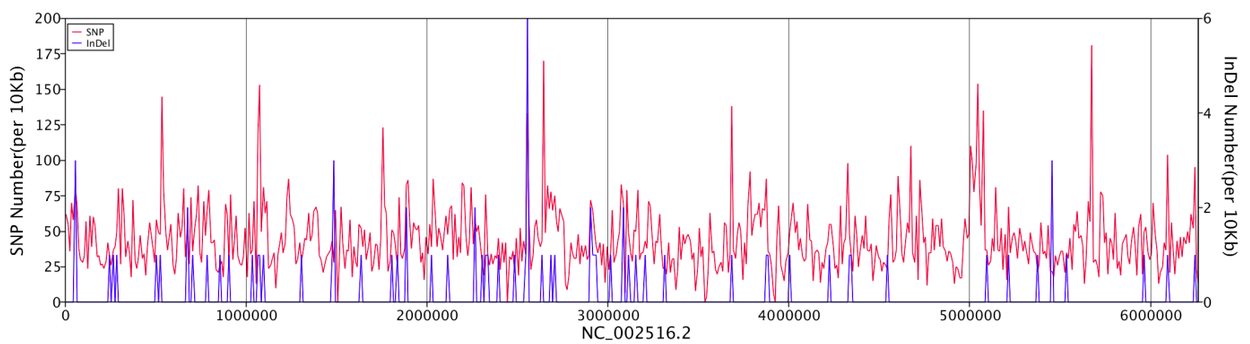

Figure 2-3-5 sample PAPAO1_TGR2A_SNP_InDel SNP/InDel distribution over the genome.The horizontal listed chromosome of reference sequence, the vertical axis represents the number of SNP /InDel per 10kb region sequence, left ruler for the SNP, the right side of the scale is InDel.(Click)

2.4 SNP、InDel Annotation

2.4.1 SNP Annotation

The SNP function can be annotated in accordance with the positional relationship and interaction between the SNP and genetic, the results in the table below:

Table 2-5 SNP Annotation Result Statistics

SNP Annotation Result

Reference	Sample	#Non_syn	#Syn	#Intergenic
BS168	BS22.2	8,425	25,966	3,861
BS	CX12	8,962	27,625	4,102
BADSM7	BA69.2	16,240	48,241	6,612
BMV	BM70	16,262	48,251	6,521
PAPAO1	TGR2A	5,258	18,343	3,970

	Page 1 of 1 10	View 1 - 5 of 5

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

2.4.2 InDel Annotation

InDel Annotation, results in the table below:

Table 2-6 InDel Annotation Result Statistics

INDEL Annotation Result

Reference	Sample	non_shift	shift	Intergenic	Total
BS168	BS22.2	18	21	200	239
BS	CX12	58	76	597	731
BADSM7	BA69.2	43	72	263	378
BMV	BM70	59	83	465	607
PAPAO1	TGR2A	7	4	66	77

	Page 1 of 1 10	View 1 - 5 of 5

Note Non-shift, CDS region don’t cause InDel of frame shift; Shift, CDS region cause InDel of frame shift; Intergenic： mutation is located between the gene district.

2.5 SV Detection and Annotation

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

2.5.1 SV Detection

Structural variation is detected by using BreakDancer, number of SV detected statistics in the follow table, SV length distribution as shown below.

Table 2-7 SV Detection results Statistics

SV Detection Result

Reference	Sample	INS	DEL	INV	ITX	CTX	Unknown	Total
PAPAO1	TGR2A	1	17	0	5	0	0	23
BS168	BS22.2	0	15	0	0	0	0	15
BMV	BM70	0	51	0	1	0	0	52
BADSM7	BA69.2	0	52	0	3	0	0	55
BS	CX12	0	12	0	0	0	0	12

	Page 1 of 1 10	View 1 - 5 of 5

Note There are samples ID, the number of Insert, number of deletions , number of inversions , number of intra-chromosomal translocation, number of inter-chromosomal translocation.

2.5.2 SV Annotation

DEL、INS、INV Annotation results in the table below:

Table 2-8 SV Annotation Result Statistics

SV Annotation Result

Reference	Sample	Gene_region	Intergenic
PAPAO1	TGR2A	15	3
BS168	BS22.2	14	1
BMV	BM70	50	1
BADSM7	BA69.2	49	3
BS	CX12	9	3
	Page 1 of 1 10		View 1 - 5 of 5

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

3 Methods described

1 Data Processing

The original optic data obtained by high-throughput sequencing (Illumina HiSeq platform) were transformed into raw sequenced reads (raw data, or raw reads) by CASAVA base calling and stored in FASTQ (fq) format which contained reads' sequences and corresponding sequencing quality. Quality control were performed and the adapter and low quality sequences were removed. The obtained clean data was used for subsequent analysis.

2 Reads Mapping

Reads were mapped to reference genomes by BWA softwares. The coverage was computed by SAMTOOLS software.

3 SNP/InDel Analysis

SNP (single nucleotide polymorphism) mainly refers to the variation of DNA sequence at the level of the single nucleotide including transition and transversion. InDel refers to the insertion and deletion of small fragments in the genome. SAMTOOLS was used to detect the individual SNP and insertion and deletion of small fragments(<50bp). The position of the SNP/InDel in the functional regions of the genome was also annotated.

4 SV Analysis

SV (structural variation) refers to the insertion, deletion, inversion and translocation of the large segments at 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 are found by BreakDancer software.