Human Whole Genome Sequencing (Cancer)

Primary Analysis Demo Report

May 1, 2016

Contents

1 Sample Information 1
2 Experimental Procedure 2
2.1 DNA Quantification & Qualification 2
2.2 Library Preparation for Sequencing 2
2.3 Clustering & Sequencing 3
3 Bioinformatics Analysis Pipeline 4
4 Analysis Result 4
4.1 Raw data 4
4.2 Quality Control 6
4.2.1 Sequencing Data Filtration 6
4.2.2 Sequencing Error Rate Examination 7
4.2.3 Sequencing Quality Distribution 8
4.2.4 Statistics of Sequencing Quality 9
4.3 Sequence Alignment 10
4.3.1 Sequencing Depth & Coverage Distribution 11
4.3.2 Statistics of Mapping, Coverage & Depth 12
4.4 Variant Detection 13
4.4.1 SNP Detection Result 13
4.4.2 InDel Detection Result 15
4.4.3 SV Detection Result 17
4.4.4 CNV Detection Result 19
4.4.5 Variant Annotation Result 20
4.5 Somatic Mutation Detection 23
4.5.1 Somatic SNP Detection Result 24
4.5.2 Somatic InDel Detection Result 25
4.5.3 Somatic SV Detection Result 26
4.5.4 Somatic CNV Detection Result 27
5 References 28
6 Appendix 30

1 Sample Information

Table 1.1 Sample information

PatientID	SampleID	LibraryID	Type
P001	P001_N	DHG04561	N
P001	P001_T	DHG04554	T
P002	P002_N	DHG04557	N
P002	P002_T	DHG04550	T

Type: sample type (N: normal; T: tumor; U: unknown)

We performed hierarchical cluster analysis among the samples based on the SNP genotype information. Only SNPs called on chromosome 1 were considered. This analysis was performed by using R function “hclust” with the agglomeration method “ward”. The result can help us determine whether two paired samples were from the same patient.

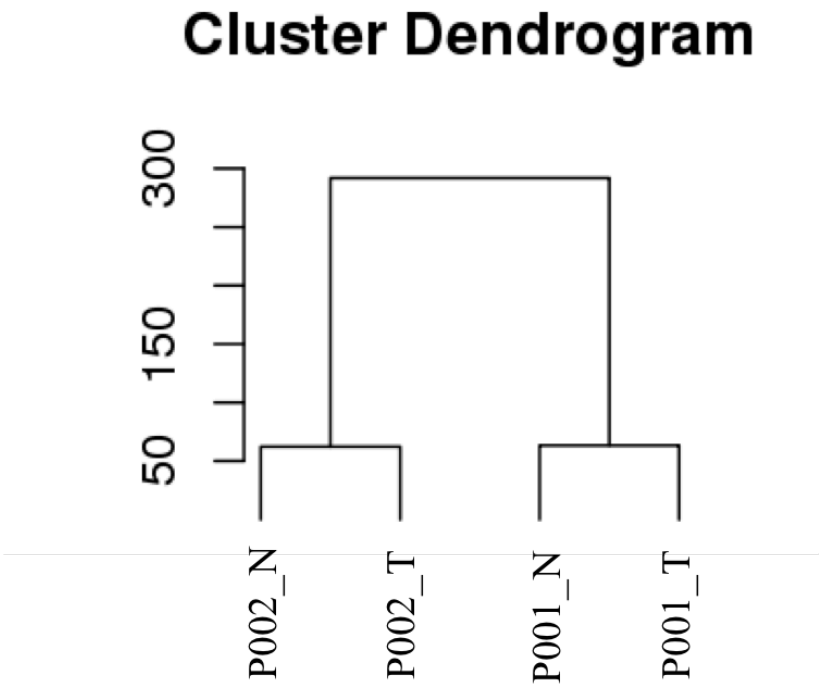

Figure 1.1 Cluster analysis among the samples

2 Experimental Procedure

2.1 DNA Quantification & Qualification

1) DNA degradation and contamination were monitored on 1% agarose gels.

2) DNA purity was checked using the NanoPhotometer® spectrophotometer (IMPLEN, CA, USA).

3) DNA concentration was measured using Qubit® DNA Assay Kit in Qubit® 2.0 Flurometer (Life Technologies, CA, USA).

4) Fragment distribution of DNA library was measured using the DNA Nano 6000 Assay Kit of Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA).

2.2 Library Preparation for Sequencing

A total amount of 1.0μg DNA per sample was used as input material for the DNA sample preparations. Sequencing libraries were generated using Truseq Nano DNA HT Sample Preparation Kit (Illumina USA) following manufacturer's recommendations and index codes were added to each sample. Briefly, the DNA sample was fragmented by sonication to a size of 350bp, and then DNA fragments were endpolished, A-tailed, and ligated with the full-length adapter for Illumina sequencing with further PCR amplification. At last, PCR products were purified (AMPure XP system) and libraries were analyzed for size distribution by Agilent 2100 Bioanalyzer and quantified using real-time PCR.

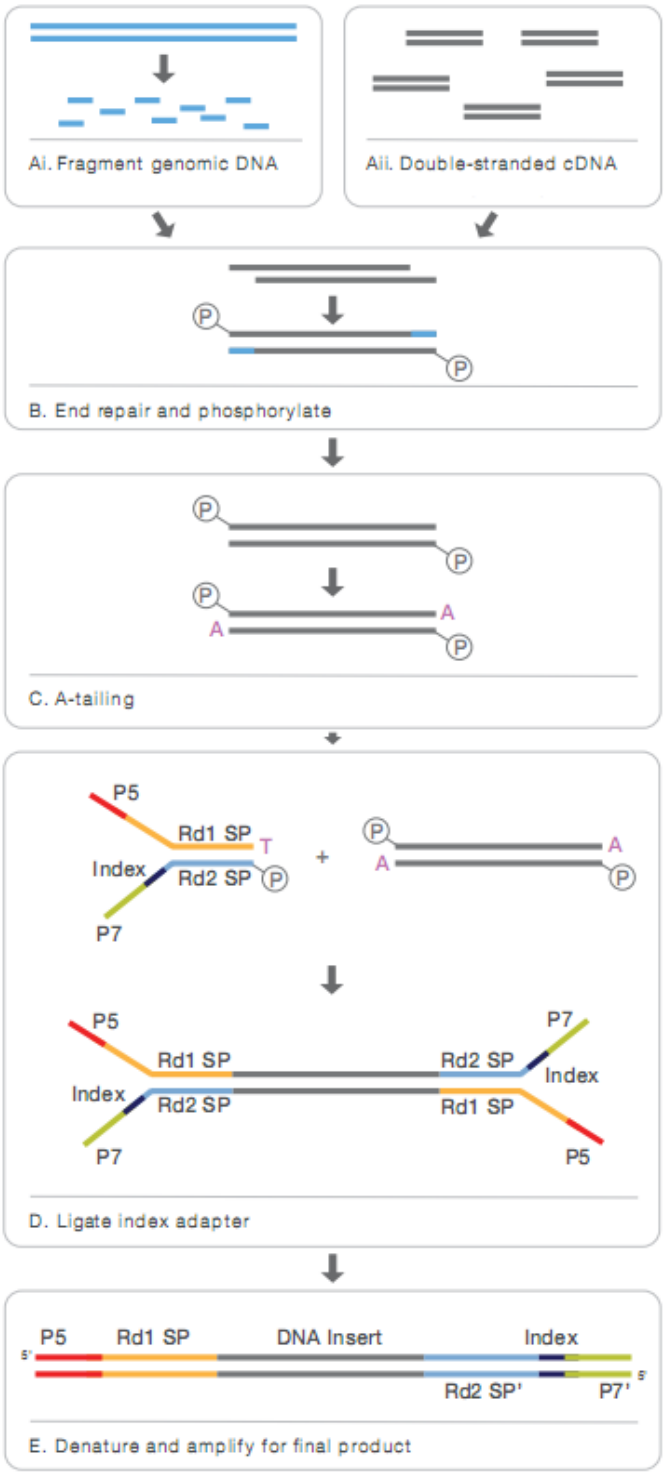

Figure 2.1 Library construction workflow

2.3 Clustering & Sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v4-cBot-HS (Illumina, San Diego,

USA) according to the manufacturer's instructions. After cluster generation, the libraries were sequenced on an Illumina sequencing platform.

3 Bioinformatics Analysis Pipeline

The flowchart below depicts the bioinformatics analysis pipeline we used. Somatic analyses are performed only when tumor-normal paired samples are provided.

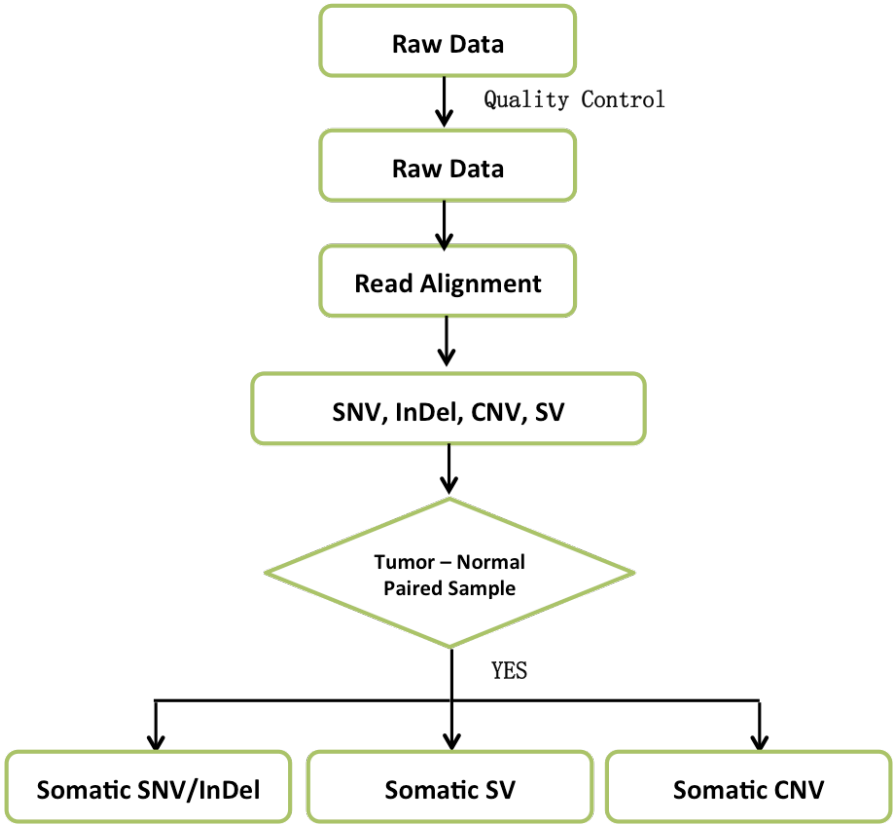

Figure 3.1 Bioinformatics analysis pipeline

4 Analysis Result

4.1 Raw data

The original fluorescence images obtained from high throughput sequencing platforms are transformed to short reads by base calling. These short reads (Raw data) are recorded in FASTQ format, which contains sequence information (reads) and corresponding sequencing quality information.

Every read in FASTQ format is stored in four lines as follows:

@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GCTCTTTGCCCTTCTCGTCGAAAATTGTCTCCTCATTCGAAACTTCTCTGT
+
@@CFFFDEHHHHFIJJJ@FHGIIIEHIIJBHHHIJJEGIIJJIGHIGHCCF

Line 1 begins with a '@' character which is followed by a sequence identifier and an optional description. Line 2 shows the sequenced bases. Line 3 begins with a '+' character and is optionally followed by the same sequence identifier. Line 4 encodes the sequencing quality for each base in line 2, and contains the same number of characters as bases in line 2.

Table 4.1 Illumina sequence identifier

EAS139	The unique instrument name
136	Run ID
FC706VJ	Flowcell ID
2	Flowcell lane
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

The ASCII value of every character at the fourth line minus 33 equals to the phred-scaled quality value of the corresponding sequenced base in the second line. The relationship between sequencing error rate (e) and base quality value (Qphred) can be expressed by the following equation:

$$Qphred = -10log10(e)$$

The table below shows examples of corresponding values among sequencing error rate (e), base quality value (Qphred) and charater.

Table 4.2 Sequencing error rate and corresponding base quality value

Sequencing error rate	Sequencing quality value	Corresponding character
5%	13	.
1%	20	5
0.1%	30	?
0.01%	40	I

4.2 Quality Control

4.2.1 Sequencing Data Filtration

Raw sequencing data may contain adapter contaminated and low-quality reads. These sequence artifacts may increase the complexity of downstream analyses, which means that quality control is an essential step. All the downstream analyses will be based on clean reads that pass quality control.

We performed quality control according to the following procedure:

1) Discard a read pair if either one read contains adapter contamination;
2) Discard a read pair if more than 10% of bases are uncertain in either one read;
3) Discard a read pair if the proportion of low quality bases is over 50% in either one read.

DNA-Seq Adapter (Adapter, Oligonucleotide sequences for TruSeq™ DNA Sample Prep Kits) information:

5’ Adapter:
5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3’
3’ Adapter:
5’-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC(6-nt index)ATCTCGTATGCCGTCTTCTGCTTG-3’

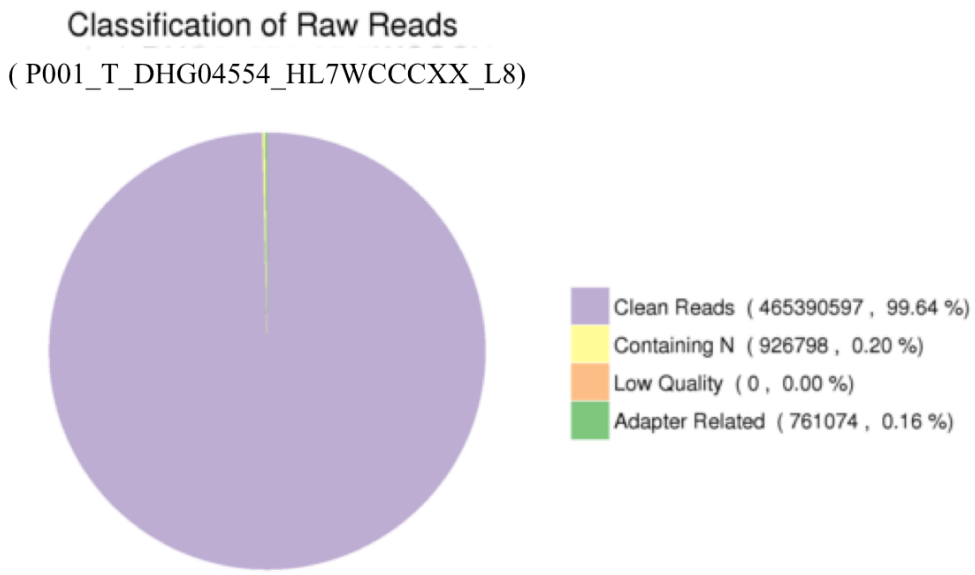

Figure 4.1 Raw data filtration result

Clean Reads: read pairs that passed quality control; Containing N: read pairs in either one read of which more than 10% of bases are uncertain; Low Quality: read pairs in either one read of which the proportion of low quality bases is over 50%; Adapter Related: read pairs that contain adapter contamination in either one read.

4.2.2 Sequencing Error Rate Examination

Sequencing error rate and base quality can be affected by various factors such as sequencing platform, chemical reagent and sample quality. Due to the consumption of chemical reagents, error rate is increasing with read extension, which is a common feature of Illumina high throughput sequencing platforms.

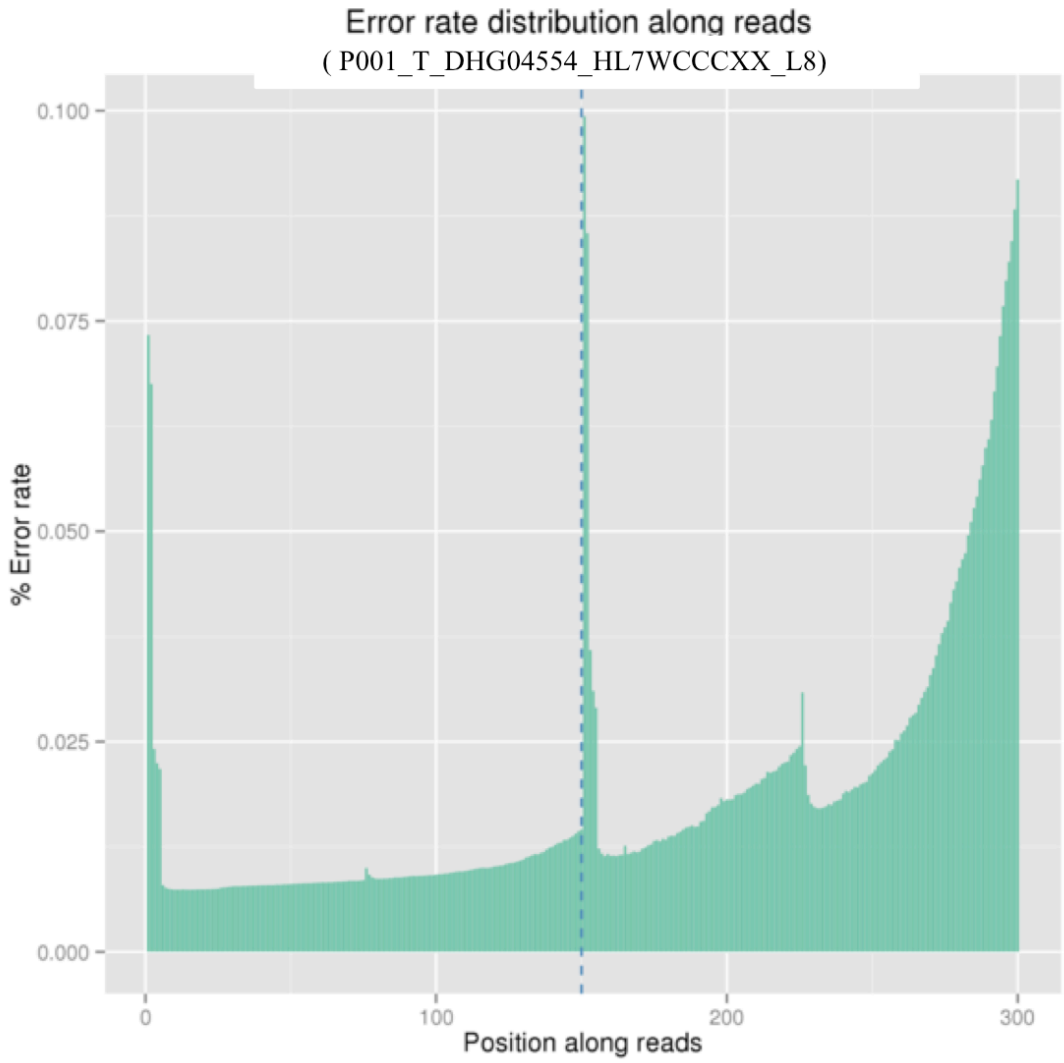

Figure 4.2 Sequencing error rate distribution

The x-axis represents position in reads, and the y-axis represents the average error rate of bases of all reads at a position.

4.2.3 Sequencing Quality Distribution

The phred-scaled quality scores of most bases should be greater than 20, which are required by downstream analyses. It is common to see that base quality decreases along reads, which is an inherent characteristic of next generation sequencing.

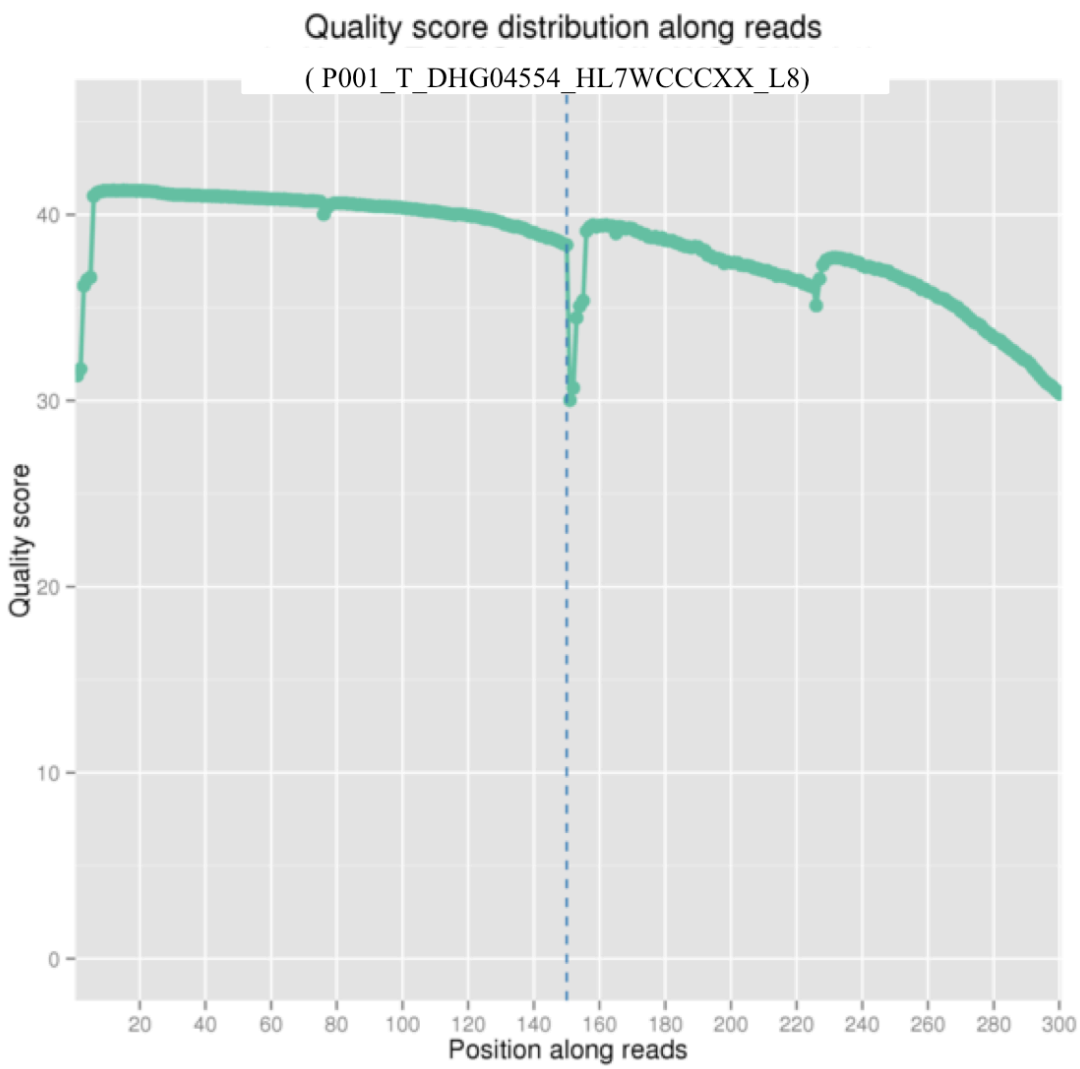

Figure 4.3 Sequencing quality distribution

The x-axis is position in reads, and the y-axis is the average quality score of bases of all reads at a position.

4.2.4 Statistics of Sequencing Quality

According to the sequencing feature of Illumina platforms, for paired-end sequencing data we require that Q30 (the percent of bases with phred-scaled quality scores greater than 30) should be above 80%.

Table 4.3 Overview of data production quality

Sample name	Library	Flowcell/Lane	Raw reads	Raw data (G)	Effective (%)	Error (%)	Q20 (%)	Q30 (%)	GC (%)
P006_-T	DHG04554	HL7WCCCXX_L8	467078469	195.73	99.64	0.03	95.12	90.16	42
P006_-T	DHG04554	HL7WCCCXX_L4	79673815		99.81	0.03	95.01	89.86	41.92
P006_-T	DHG04554	HL7WCCCXX_L5	105667602		99.76	0.03	94.82	89.58	41.96
P006_N	DHG04561	HL7WCCCXX_L4	34450121	98.25	99.73	0.03	95.15	90.07	42.59
P006_N	DHG04561	HL7WCCCXX_L5	46210349		99.68	0.03	94.94	89.78	42.62
P006_N	DHG04561	HL7WCCCXX_L6	57143979		99.7	0.04	94.55	89.02	42.59
P006_N	DHG04561	HL7WCCCXX_L7	57908735		99.68	0.04	94.62	89.18	42.61
P006_N	DHG04561	HL7WCCCXX_L2	66045053		99.75	0.04	94.6	89.16	42.75
P006_N	DHG04561	HL7WCCCXX_L3	65725438		99.74	0.04	94.48	88.91	42.75
P002_T	DHG04550	HJFK3CCXX_L1	437992437	201.85	99.69	0.03	96.59	92.44	43.88
P002_T	DHG04550	HJFK3CCXX_L6	92458231		99.71	0.03	96.77	92.87	43.82
P002_T	DHG04550	HJFK3CCXX_L5	58438773		99.71	0.03	96.62	92.57	43.78
P002_T	DHG04550	HJFK3CCXX_L4	27738594		99.73	0.03	96.58	92.5	43.8
P002_T	DHG04550	HJFK3CCXX_L2	56199932		99.73	0.03	96.41	92.17	43.94
P002_N	DHG04557	HL7WCCCXX_L5	48132835	101.87	99.71	0.03	94.92	89.69	43
P002_N	DHG04557	HL7WCCCXX_L4	35897994		99.76	0.03	95.12	89.98	42.97
P002_N	DHG04557	HL7WCCCXX_L3	69026068		99.76	0.04	94.45	88.8	43.14
P002_N	DHG04557	HL7WCCCXX_L7	57955740		99.71	0.04	94.62	89.12	42.99
P002_N	DHG04557	HL7WCCCXX_L2	69067301		99.77	0.04	94.58	89.05	43.14
P002_N	DHG04557	HL7WCCCXX_L6	59492715		99.73	0.04	94.53	88.95	42.97

(1) Sample name: sample name

(2) Library: library name

(3) Flowcell/Lane: the flowcell ID and lane number

(4) Raw reads: the number of sequencing reads pairs; four lines would be considered as one unit according to the format of FASTQ

(5) Raw data: the original sequencing data

(6) Effective: the percentage of clean reads in all raw reads

(7) Error: the average error rate of all bases on read1 and read2; the error rate of a base is obtained from equation 1

(8) Q20: the percent of bases with phred-scaled quality scores greater than 20

(9) Q30: the percent of bases with phred-scaled quality scores greater than 30

(10) GC content: the percentage of G and C in the all bases

4.3 Sequence Alignment

Burrows-Wheeler Aligner (BWA) was utilized to map the paired-end clean reads to the human reference genome (b37 + decoy). After sorting with samtools and marking

duplicates with Picard, the results of read alignment was finally stored in the format of BAM. We then compute the coverage and depth based on the final BAM file.

4.3.1 Sequencing Depth & Coverage Distribution

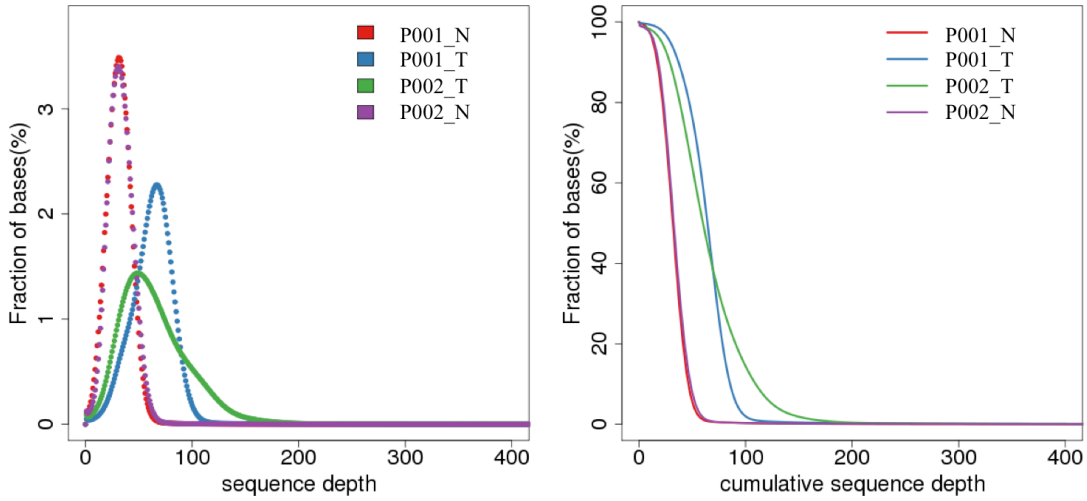

Figure 4.4 Sequencing depth

The left figure shows sequencing depth distribution of all bases in each sample. The x-axis is sequencing depth, and the y-axis is the fraction of bases with the given sequencing depth. The right figure shows accumulative equencing depth distribution of all bases in each sample. The x-axis is accumulative sequencing depth, and the y-axis is the fraction of bases above the given sequencing depth.

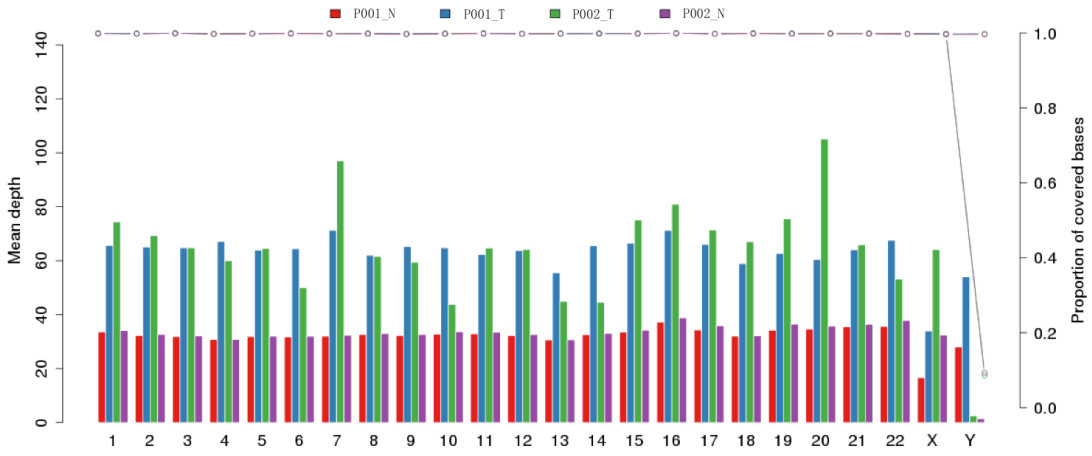

Figure 4.5 Average sequencing depth (bar plot) and coverage (dot-line plot) in each chromosome

The x-axis represents chromosome; the left y-axis is the average depth; the right y-axis is the coverage (proportion of covered bases).

4.3.2 Statistics of Mapping, Coverage & Depth

Table 4.4 Statistics of mapping, coverage and depth in each sample

Sample	P001_N	P001_T	P002_T	P002_N
Total	653091374 (100%)	1300651880 (100%)	1341612930 (100%)	677400658 (100%)
Duplicate	80035220 (12.29%)	198084520 (15.26%)	162995191 (12.16%)	80333844 (11.89%)
Mapped	651322827 (99.73%)	1297739397 (99.78%)	1340355508 (99.91%)	675496625 (99.72%)
Properly mapped	628585682 (96.25%)	1271434068 (97.75%)	1310590942 (97.69%)	651844748 (96.23%)
PE mapped	649766678 (99.49%)	1295120450 (99.57%)	1339559188 (99.85%)	673849392 (99.48%)
SE mapped	3112298 (0.48%)	5237894 (0.40%)	1592640 (0.12%)	3294466 (0.49%)
With mate mapped to a different chr	8470924 (1.30%)	6669962 (0.51%)	7954516 (0.59%)	4834750 (0.71%)
With mate mapped to a different chr ((mapQ>=5))	6810995 (1.04%)	4274555 (0.33%)	5130264 (0.38%)	3174167 (0.47%)
Average_sequencing_depth	32.76	64.69	66.86	33.9
Coverage	0.9974	0.9979	0.9906	0.9903
Coverage_at_least_4X	0.9938	0.9966	0.9883	0.987
Coverage_at_least_10X	0.9744	0.9936	0.9823	0.9745
Coverage_at_least_20X	0.8476	0.9817	0.9564	0.8694

(1) Sample: sample name

(2) Total: the total number of clean reads

(3) Duplicate: the number of duplication reads (percentage)

(4) Mapped: the number of reads that are mapped to the reference genome (percentage)

(5) Properly mapped: the number of reads with themselves and mate reads mapped, and within the expected insert size (percentage)

(6) PE mapped: the number of pair-end reads that are mapped to the reference genome (percentage)

(7) SE mapped: the number of single-end reads that mapped to the reference genome (percentage)

(8) With mate mapped to a different chr: the number of reads with mate reads mapped to different chromosomes (percentage)

(9) With mate mapped to a different chr (mapQ >= 5): the number of reads with mate reads mapped to different chromosomes and the MAQ > 5

(10) Average_sequencing_depth: the average sequencing depth in the whole genome

(11) Coverage: the coverage in the whole genome

(12) Coverage_at_least_4X: the coverage in the whole genome when only bases with depth > 4X are considered

(13) Coverage_at_least_10X: the coverage in the whole genome when only bases with depth > 10X are considered

(14) Coverage_at_least_20X: the coverage in the whole genome when only bases with depth > 20X are considered

4.4 Variant Detection

4.4.1 SNP Detection Result

Single nucleotide polymorphisms (SNPs), also known as single nucleotide variants (SNVs), constitute the largest class of genome variants in genome. A typical whole genome of human has about 3.6 million SNPs. Statistics of detected SNPs are shown below.

Table 4.5 The number of SNPs in various genomic regions

Sample	P001_T	P002_T	P001_N	P002_N
CDS	22807	22386	22715	22641
synonymous_SNP	11583	11264	11557	11376
missense_SNP	10655	10558	10594	10678
stopgain	80	81	79	85
stoploss	15	13	16	12
unknown	474	470	469	490
intronic	1257472	1245042	1245943	1263543
UTR3	24877	25075	24712	25430
UTR5	5529	5451	5494	5511
splicing	562	568	566	566
ncRNA_exonic	11662	11732	11609	11826
ncRNA_intronic	194303	192764	192233	195216
ncRNA_UTR3	728	675	725	705
ncRNA_UTR5	137	155	135	158
ncRNA_splicing	133	116	131	117
upstream	23445	23609	23158	23635
downstream	23555	22888	23333	23174
intergenic	2102788	2102408	2081683	2127289
Total	3667998	3652869	3632437	3699811

(1) Sample: sample name

(2) CDS: the number of SNPs in coding region

(3) synonymous_SNP: a single nucleotide change that does not cause an amino acid change

(4) missense_SNP: a single nucleotide change that causes an amino acid change

(5) stopgain: a nonsynonymous SNP that leads to the immediate creation of stop codon at the variant site

(6) stoploss: a nonsynonymous SNP that leads to the immediate elimination of stop codon at the variant site

(7) unknown: unknown function (due to various errors in the gene structure definition in the database file

(8) intronic: the number of SNPs in intronic region

(9) UTR3: the number of SNPs in 3'UTR region

(10) UTR5: the number of SNPs in 5'UTR region

(11) splicing: the number of SNPs within 4bp away from an exon/intron boundary

(12) ncRNA_exonic: the number of SNPs in exonic region of non-coding RNAs

(13) ncRNA_intronic: the number of SNPs in intronic region of non-coding RNAs

(14) ncRNA_UTR3: the number of SNPs in 3'UTR of non-coding RNAs

(15) ncRNA_UTR5: the number of SNPs in 5'UTR of non-coding RNAs

(16) ncRNA_splicing: the number of SNPs within 4bp away from an exon/intron boundary of non-coding RNAs

(17) upstream: the number of SNPs within 1kb away from the transcription start site

(18) downstream: the number of SNPs within the 1kb away from the transcription ending site

(19) intergenic: the number of SNPs in intergenic region

(20) otal: the total number of SNPs

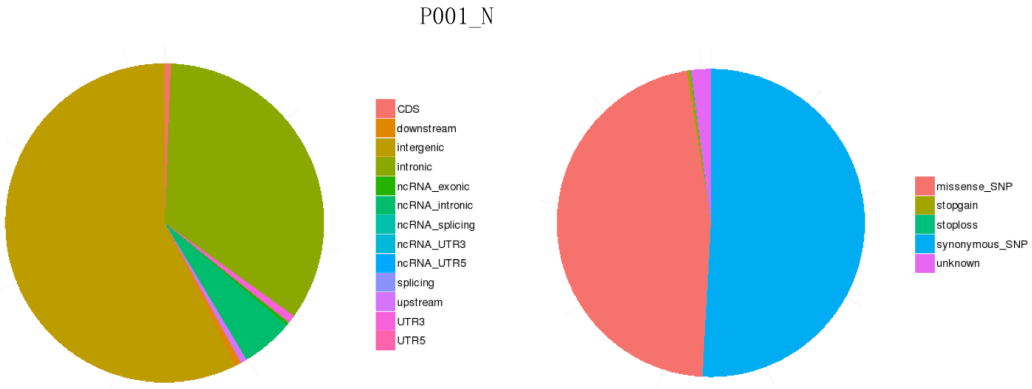

Figure 4.6 Number of SNPs in various genomic regions (left pie plot); number of different types of SNPs in coding region (right pie plot)

Table 4.6 Feature of SNPs

Sample	P001_T	P002_T	P001_N	P002_N
Total	3667998	3652869	3632437	3699811
Het	2189360	2207953	2155883	2268559
Hom	1478638	1444916	1476554	1431252
transition	2464174	2455188	2444243	2488992
transvertion	1203824	1197681	1188194	1210819
ts/tv	2.05	2.05	2.06	2.06
dbSNP percentage	3601228 (98.18%)	3587924 (98.22%)	3573549 (98.38%)	3641366 (98.42%)
novel	66770	64945	58888	58445
novel ts	38366	38054	34436	34433
novel tv	28404	26891	24452	24012
novel ts/tv	1.35	1.42	1.41	1.43

(1) Sample: sample name

(2) Total: the total number of SNPs

(3) Het: the number of heterozygotes

(4) Hom: the number of homozygotes

(5) transition (ts) : the number of transitions

(6) transversion (tv) : the number of transversions

(7) ts/tv: the number of transitions devided by the number of transversions

(8) dbSNP percentage: the number of SNPs that have been reported in dbSNP database devided by the total number of called SNPs

(9) novel: the number of SNPs not reported in dbSNP

(10) novel ts: the number of transition SNPs that have not been reported in dbSNP

(11) novel tv: the number of transversion SNPs that have not been reported in dbSNP

(12) novel ts/tv: novel ts devided by novel tv

4.4.2 InDel Detection Result

Small insertions and deletions (InDels) that are less than 50bp in length constitute another class of genomic variants in human genome. A typical human genome may contain about 350,000 InDels.

The InDels occurred in coding region or splicing sites may cause changes in transcripts and proteins. If the number of inserted or deleted nucleotides is not three or multiples of three, the whole reading frame would be altered. The statistics of InDels called in the samples are listed below:

Table 4.7 Number of InDel in various genomic regions

Sample	P001_T	P002_T	P002_N	P001_N
CDS	756	704	725	738
frameshift_deletion	137	139	145	131
frameshift_insertion	107	99	105	103
nonframeshift_deletion	222	191	192	220
nonframeshift_insertion	173	166	171	174
stopgain	4	7	10	3
stoploss	2	0	0	2
unknown	110	102	102	105
intronic	284536	267149	271860	260469
UTR3	6330	6019	6025	5934
UTR5	913	918	936	862
splicing	239	224	227	211
ncRNA_exonic	1642	1553	1615	1558
ncRNA_intronic	41711	38577	39089	38192
ncRNA_UTR3	184	175	187	168
ncRNA_UTR5	30	29	34	29
ncRNA_splicing	43	43	46	41
upstream	5715	5312	5617	5293
downstream	5912	5570	5689	5452
intergenic	426493	403019	406204	391116
Total	774504	729292	738254	710063

(1) Sample: sample name

(2) CDS: the number of InDels in coding region

(3) frameshift_deletion: a deletion of one or more nucleotides that cause frameshift changes in protein coding sequence

(4) frameshift_insertion: an insertion of one or more nucleotides that cause frameshift changes in protein coding sequence

(5) nonframeshift_deletion: a deletion that does not cause frameshift changes

(6) nonframeshift_insertion: an insertion that does not cause frameshift changes

(7) stopgain: an insertion or a deletion that leads to the immediate creation of stop codon at the variant site

(8) stoploss: an insertion or a deletion that leads to the immediate elimination of stop codon at the variant site

(9) unknown: unknown function (due to various errors in the gene structure definition in the database file)

(10) intronic: the number of InDels in intronic region

(11) UTR3: the number of InDels in 3'UTR region

(12) UTR5: the number of InDels in 5'UTR region

(13) splicing: the number of InDels within 4bp away from an exon/intron boundary

(14) ncRNA_exonic: the number of InDels in exonic region of non-coding RNAs

(15) ncRNA_intronic: the number of InDels in intronic region of non-coding RNAs

(16) ncRNA_UTR3: the number of InDels in 3'UTR of non-coding RNAs

(17) ncRNA_UTR5: the number of InDels in 5'UTR of non-coding RNAs

(18) ncRNA_splicing: the number of InDels within 4bp away from an exon/intron boundary of non-coding RNAs

(19) upstream: the number of InDels within 1kb away from transcription start site

(20) downstream: the number of InDels iwithin 1kb away from transcription ending site

(21) intergenic: the number of InDels in intergenic region

(22) Total: the total number of InDels

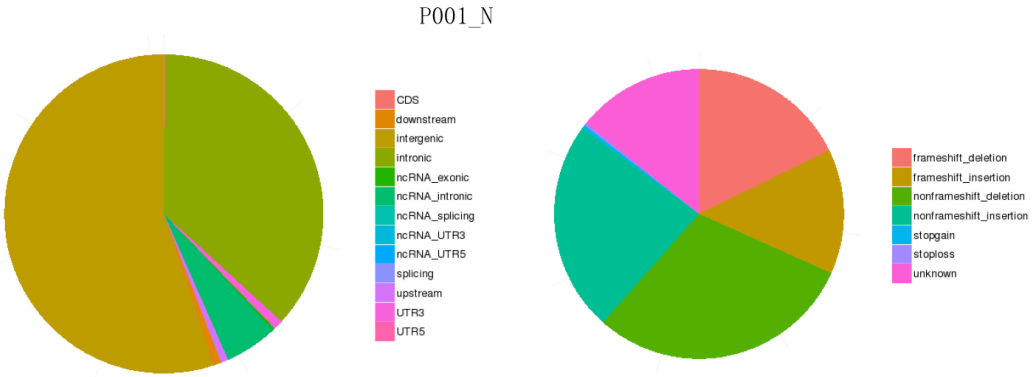

Figure 4.7 Number of InDels in various genomic regions (left); number of different types of InDels in coding region (right)

Table 4.8 Feature of InDels

Sample	P001_T	P002_N	P002_T	P001_N
Total	774504	729292	738254	710063
Het	449926	427314	424816	398887
Hom	324578	301978	313438	311176
dbSNP	561712	541134	536660	527318
percentage	(72.53%)	(74.20%)	(72.69%)	(74.26%)
novel	212792	188158	201594	182745

(1) Sample: sample name

(2) Total: the total number of InDels

(3) Het: the number of heterozygotes

(4) Hom: the number of homozygotes

(5) dbSNP percentage: the number of InDels that have been reported in dbSNP database devided by the total number of called InDels

(6) novel: the number of InDels that have not been reported in dbSNP

4.4.3 SV Detection Result

Structural variants (SVs) are genomic variants with relatively large size (>50 bp), including deletions, duplications, insertions, inversions and translocations. SVs may form the underlying genetic basis of individual differences, and have potential effect on disease and cancer susceptibility. The statistics of SV are shown below:

Table 4.9 SV detection result

Sample	Inversion	Insertion	Translocation	Deletion
P001_N	118	236	293	2285
P001_T	222	515	506	3329
P002_T	262	474	607	3625
P002_N	144	238	307	2312

(1) Sample: sample name

(2) Inversion: the number of inversions

(3) Insertion: the number of insertions

(4) Translocation: the number of translocations

(5) Deletion: the number of deletions

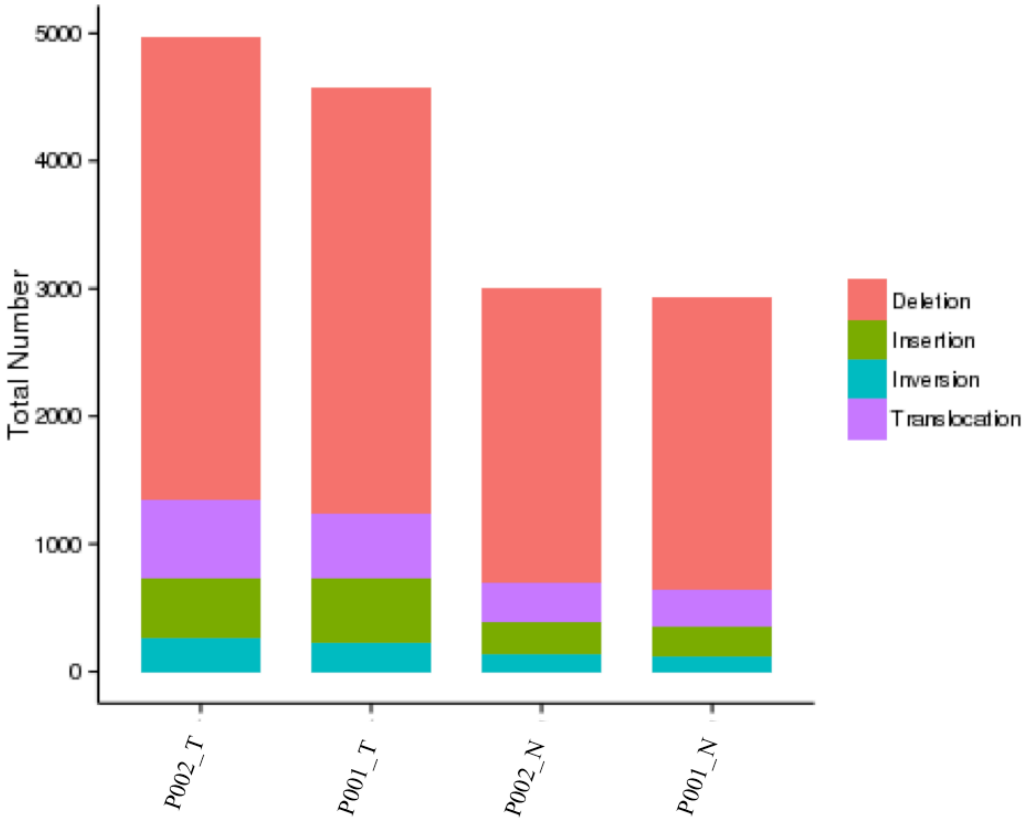

Figure 4.8 Number of different types of SV in each sample

The x-axis represents samples, and the y-axis is the number of each type of SV

4.4.4 CNV Detection Result

Copy number variants (CNVs) are genomic variants that lead to variation in copy number of relatively large fragment (longer than 50 bp) among individuals. There are two types of CNVs, i.e. gains and losses of copies. CNVs may form the underlying genetic basis of individual differences and cancer. The statistics of detected CNVs are listed below:

Table 4.10 CNV detection result

Sample	gain_count	gain_size	loss_count	loss_size	total_count	total_size
P001_N	144	3093144	78	153053078	222	156146222
P001_T	128	7190128	103	153205103	231	160395231
P002_T	422	340715422	138	447437138	560	788152560
P002_N	148	4417148	54	1026054	202	5443202

(1) Sample: sample name

(2) gain_count: the number of gains

(3) gain_size: the total size of gains

(4) loss_count: the number of losses

(5) loss_size: the total size of losses

(6) total_count: the total number of CNVs

(7) total_size: the total size of CNVs

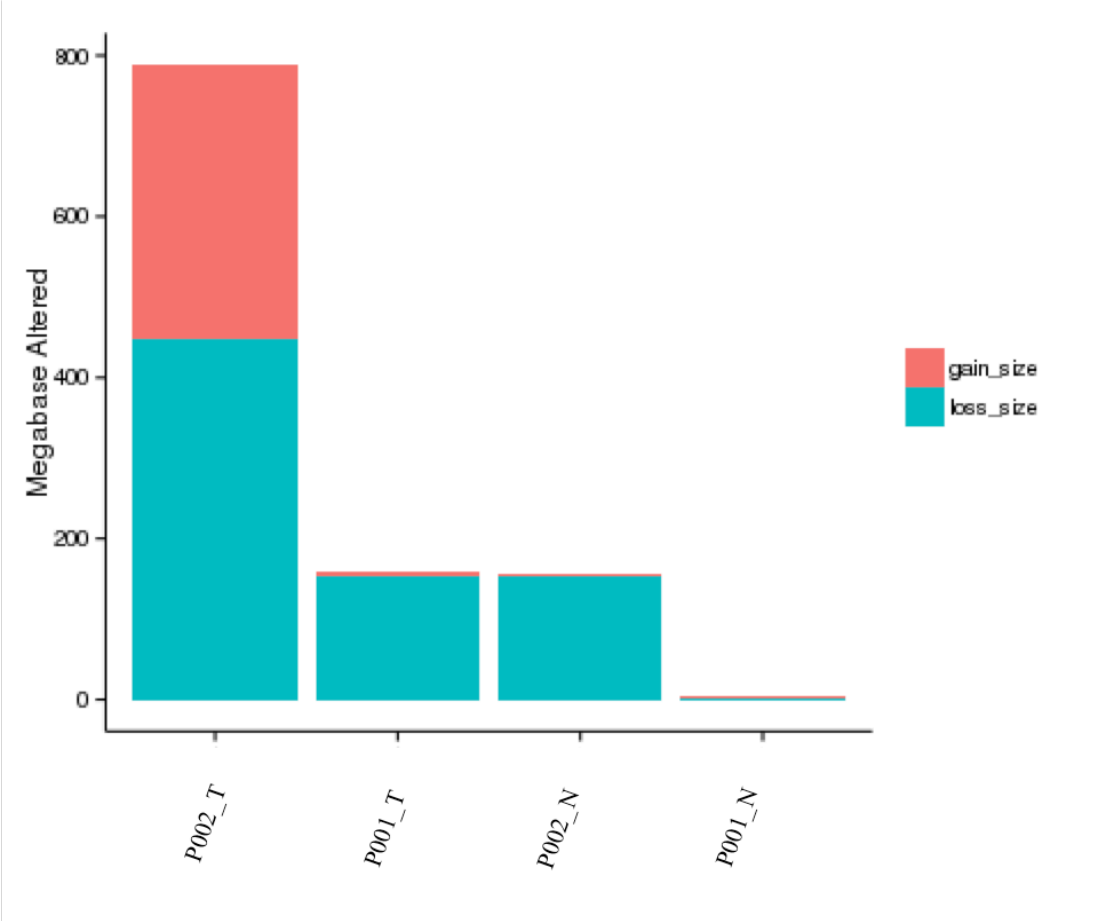

Figure 4.9 The size of genomic regions affected by CNVs in each sample

The x-axis represents samples, and the y-axis is the total size of genomic regions affected by gains or losses (Mb)

4.4.5 Variant Annotation Result

Following genomic variant detection, we performed annotation of variants with the tool ANNOVAR (Wang K et al.) in multiple aspects, including protein coding changes, affected genomic regions, allele frequency reported by some big projects, deleteriousness prediction, etc. The main databases used are as follows:

- RefSeq and Gencode databases were used to find genomic regions affected by the variant and possible changes in protein.
- We annotated the features of the genomic regions affected by the variants, such as cytoband, small RNA, conserved mammalian microRNA regulatory target sites, conservative regions of vertebrates, transcription factor binding sites, repeats, etc.
- SIFT, PolyPhen, MutationAssessor, LRT and CADD scores were used to predict the deleteriousness of mutations. GERP++ scores were used to access the conservation of mutations.

- Alternative allele frequencies in populations reported by big sequencing projects were provided, including 1000 Human Genome, Exome Aggregation Consortium (ExAC) and exome sequencing project (ESP).
- Databases dbSNP, COSMIC, OMIM, GWAS Catalog and HGMD were used to find reported information of the variant, such as top SNPs in GWAS and cancer/disease associations.
- Databases including Gene Otology, KEGG, Reactome, Biocarta and PID were applied to provide functional or pathway annotation.

Table 4.11 Annotation result of detected variants (only a part is shown here)

CHROM	POS	ID	REF	ALT	QAUL	FILTER	GeneName	Func	Gene	…
1	17538	rs200046632	C	A	288.77	PASS	WASH7P	ncRNA_intronic	NR_024540	…
1	51479	rs116400033	T	A	1112.77	PASS	.	intergenic	NR_026820,NM_001005484	…

Part One: Basic information of the variant and its affected genomic elements

(1) CHROM: chromosome ID

(2) POS: the position of the variant on chromosomes

(3) ID: the identifier of the variant in dbSNP database

(4) Ref: reference allele

(5) Alt: alternative allele

(6) QAUL: quality value for the variant

(7) FILTER: filter status; PASS if this variant has passed all filter thresholds

(8) GeneName: the name(s) of the gene(s) affected by the variant according to RefSeq gene annotation

(9) Func: functional category overlapped by the variant

(10) Gene: the name(s) of the transcript(s) affected by the variant

(11) GeneDetail: details of sequence changes as a result of the variant

(12) ExonicFunc: functional consequences of the variant in exonic region

(13) AAChange: amino acid changes as a result of the variant

(14) Gencode: the name(s) of the transcript(s) affected by the variant according to GENCODE gene annotation.

(15) cytoband: chromosome bands overlapped by the variant

(16) wgRna: snoRNAs and microRNAs overlapped by the variant

(17) targetScanS: conserved mammalian microRNA regulatory target sites affected by the variant for conserved microRNA families in the 3' UTR regions of Refseq genes

(18) phastConsElements46way: the conservative region predicted by phastCons basing on the whole genome alignment of vertebrates; 46way means the number of used species

(19) tfbsConsSites: transcript factor binding sites that are conservative in human, mouse and rat; this is acquired from transfac matrix database (v7.0)

(20) genomicSuperDups: this field tells whether the variant hit segmental duplications.

(21) dgvMerged: annotation from Database of Genomic Variants

(22) Repeat: this field tells whether the variant hit interspersed repeats and low complexity DNA sequences output by RepeatMasker program, such as SINE, LINE and Simple repeats

Part Two: Variant information deposited in public databases

(23) gwasCatalog: this field provides information about whether the variant has been reported in previous GWAS studies and what disease the variant may be associated with

(24) avsnp144: this field tells whether the variant has been already reported in dbSNP database and provide the corresponding 'rs' identifiers if exists

(25) cosmic70: this field tells whether the variant has been reported in Catalogue Of Somatic Mutations In Cancer (COSMIC) database

(26) clinvar_20150629: this field tells whether the variant has been reported in ClinVar database

Part Three: Alternative allele frequency reported by famous sequencing projects

(27) 1000g2015aug_eas: alternative allele frequency of the mutation in East Asian population reported by 1000 Human Genome Project

(28) 1000g2015aug_sas: alternative allele frequency of the mutation in South Asian population reported by 1000 Human Genome Project

(29) 1000g2015aug_eur: alternative allele frequency of the mutation in European population reported by 1000 Genome Project

(30) 1000g2015aug_afr: alternative allele frequency of the mutation in African population reported by 1000 Human Genome Project

(31) 1000g2015aug_amr: alternative allele frequency of the mutation in admixed American population in 1000 Human Genome Project

(32) 1000g2015aug_all: alternative allele frequency of the mutation in all populations reported by 1000 Human Genome Project

(33) esp6500siv2_all: this field gives alternative allele frequency of the mutation reported by the exome sequencing project (ESP)

(34) ExAC_ALL: this field provides alternative allele frequency of the mutation in all populations reported by the Exome Aggregation Consortium (ExAC)

(35) ExAC_AFR: this field provides alternative allele frequency of the mutation in African population reported by ExAC

(36) ExAC_AMR: this field provides alternative allele frequency of the mutation in Admixed American population reported by ExAC

(37) ExAC_EAS: this field provides alternative allele frequency of the mutation in East Asian population reported by ExAC

(38) ExAC_FIN: this field provides alternative allele frequency of the mutation in Finnish population reported by ExAC

(39) ExAC_NFE: this field provides alternative allele frequency of the mutation in Non-finnish population reported by ExAC

(40) ExAC_OTH: this field provides alternative allele frequency of the mutation in other population reported by ExAC

(41) ExAC_SAS: this field provides alternative allele frequency of the mutation in South Asian population reported by ExAC

Part Four：Deleteriousness prediction of the variant

(42) SIFT: deleteriousness prediction of the variant with SIFT score (dbNSFPv3.0a)

(43) Polyphen2_HVAR: deleteriousness prediction of the variant with Polyphen2 HVAR score (dbNSFPv3.0a)

(44) Polyphen2_HDIV: deleteriousness prediction of the variant with Polyphen2 HDIV score (dbNSFPv3.0a)

(45) MutationTaster: deleteriousness prediction of the variant with MutationTaster score (dbNSFPv3.0a)

(46) LRT: deleteriousness prediction of the variant with LRT score (dbNSFPv3.0a)

(47) gerp++gt2: conservation evaluation of the variant with GERP++ score

(48) CADD: deleteriousness prediction of the variant with CADD score

Part Five: Supplementary Information about the variant including genotype, related disease and pathway

(49) INFO: information about the variant from variation calling software

(50) FORMAT: comma-separated list of several tags from variation calling software

GT: genotype

AD: allelic depth

DP: read depth at this position

GQ: genotype Quality

PL: list of Phred-scaled genotype likelihoods

(51) Sample ID: comma-separated genotype information of the variant; The data type and order are specified in the "FORMAT" field

(52) Ori_REF: reference allele

(53) Ori_ALT: alternative allele

(54) shared_hom: whether the mutation is homozygous (1) or heterozygous (0)

(55) shared_het: whether the mutation is homozygous (0) or heterozygous (1)

(56) OMIM: annotation from Online Mendelian Inheritance in Man (OMIM)

(57) GWAS_Pubmed_pValue: p value of the variant reported by GWAS studies in Pubmed

(58) HGMD_ID_Diseasename: annotation from Human Gene Mutation Database (HGMD); HGMD is a comprehensive data on published human inherited disease mutations

(59) GO_BP: gene otology term annotation; BP: Biological process

(60) GO_CC, gene otology term annotation; CC: Cellular component

(61) GO_MF: gene otology term annotation; MF: Molecular function

(62) KEGG_PATHWAY: KEGG pathway annotation

(63) PID_PATHWAY: PID database annotation

(64) BIOCARTA_PATHWAY: BIOCARTA database annotation

(65) REACTOME_PATHWAY: REACTOME database annotation

4.5 Somatic Mutation Detection

Somatic mutations refer to genomic variants that have been accumulated in somatic cells. Some of somatic mutations, namely driver mutations, play a crucial role in tumor initiation and progression. Through analyzing sequencing data from tumor-normal paired samples, we detected somatic mutations that have been accumulated in tumor cells.

4.5.1 Somatic SNP Detection Result

We used the tool muTect to detect somatic SNPs, and the tool Strelka to detect somatic InDels. The statistics of detected somatic SNPs in the tumor samples are listed below:

Table 4.12 Number of somatic SNPs in different genomic regions

Sample	P001_T	P002_T
CDS	87	102
synonymous_SNP	25	37
missense_SNP	58	61
stopgain	3	2
stoploss	0	0
unknown	1	2
intronic	3870	3260
UTR3	68	67
UTR5	16	9
splicing	7	5
ncRNA_exonic	37	36
ncRNA_intronic	809	611
ncRNA_UTR3	2	1
ncRNA_UTR5	1	0
ncRNA_splicing	3	2
upstream	86	75
downstream	79	64
intergenic	8670	6079
Total	13735	1031

(1) Sample: sample name

(2) CDS: the number of somatic SNPs in coding region

(3) synonymous_SNP: a single nucleotide change that does not cause an amino acid change

(4) missense_SNP: a single nucleotide change that causes an amino acid change

(5) stopgain: a nonsynonymous SNP that leads to the immediate creation of stop codon at the variant site

(6) stoploss: a nonsynonymous SNP that leads to the immediate elimination of stop codon at the variant site

(7) unknown: unknown function (due to various errors in the gene structure definition in the database file

(8) intronic: the number of somatic SNPs in intronic region

(9) UTR3: the number of somatic SNPs in 3’UTR region

(10) UTR5: the number of somatic SNPs in 5’UTR region

(11) splicing: the number of somatic SNPs within 4bp away from an exon/intron boundar

(12) ncRNA_exonic: the number of somatic SNPs in exonic region of non-coding RNA

(13) ncRNA_intronic: the number of somatic SNPs in intronic region of non-coding RNAs

(14) ncRNA_UTR3: the number of somatic SNPs in 3'UTR of non-coding RNAs

(15) ncRNA_UTR5: the number of somatic SNPs in 5'UTR of non-coding RNAs

(16) ncRNA_splicing: the number of somatic SNPs within 4bp away from an exon/intron boundary of non-coding RNAs

(17) upstream: the number of somatic SNPs within 1kb away from the transcription start site

(18) downstream: the number of somatic SNPs within the 1kb away from the transcription ending site

(19) intergenic: the number of somatic SNPs in intergenic region

(20) Total: the total number of somatic SNPs

4.5.2 Somatic InDel Detection Result

Table 4.13 Number of somatic InDels in different genomic regions

Sample	P001_T	P002_T
CDS	0	1
frameshift_deletion	0	0
frameshift_insertion	0	0
nonframeshift_deletion	0	1
nonframeshift_insertion	0	0
stopgain	0	0
stoploss	0	0
unknown	0	0
intronic	30	94
UTR3	9	4
UTR5	0	1
splicing	0	0
ncRNA_exonic	0	1
ncRNA_intronic	0	15
ncRNA_UTR3	0	0
ncRNA_UTR5	0	0
ncRNA_splicing	0	0
upstream	0	1
downstream	0	3
intergenic	11	144
Total	50	264

(1) Sample: sample name

(2) CDS: the number of somatic InDels in coding region

(3) frameshift_deletion: a deletion of one or more nucleotides that cause frameshift changes in protein coding sequence

(4) frameshift_insertion: an insertion of one or more nucleotides that cause frameshift changes in protein coding sequence

(5) nonframeshift_deletion: a deletion that does not cause frameshift changes

(6) nonframeshift_insertion: an insertion that does not cause frameshift changes

(7) stopgain: an insertion or a deletion that leads to the immediate creation of stop codon at the variant site

(8) stoploss: an insertion or a deletion that leads to the immediate elimination of stop codon at the variant site

(9) unknown: unknown function (due to various errors in the gene structure definition in the database file)

(10) intronic: the number of somatic InDels in intronic region

(11) UTR3: the number of somatic InDels in 3'UTR region

(12) UTR5: the number of somatic InDels in 5'UTR region

(13) splicing: the number of somatic InDels within 4bp away from an exon/intron boundary

(14) ncRNA_exonic: the number of somatic InDels in exonic region of non-coding RNAs

(15) ncRNA_intronic: the number of somatic InDels in intronic region of non-coding RNAs

(16) ncRNA_UTR3: the number of somatic InDels in 3'UTR of non-coding RNAs

(17) ncRNA_UTR5: the number of somatic InDels in 5'UTR of non-coding RNAs

(18) ncRNA_splicing: the number of somatic InDels within 4bp away from an exon/intron boundary of non-coding RNAs

(19) upstream: the number of somatic InDels within 1kb away from transcription start site

(20) downstream: the number of somatic InDels iwithin 1kb away from transcription termination site

(21) intergenic: the number of somatic InDels in intergenic region

(22) Total: the total number of somatic InDels

4.5.3 Somatic SV Detection Result

Statistics of detected somatic SVs are listed below:

Table 4.14 Statistics of detected somatic SVs

Sample	Inversion	Insertion	Translocation	Deletion
P001_T	121	219	277	2488
P002_T	156	183	298	2486

(1) Sample: sample name

(2) Inversion: the number of inversions

(3) Insertion: the number of insertions

(4) Translocation: the number of translocations

(5) Deletion: the number of deletions

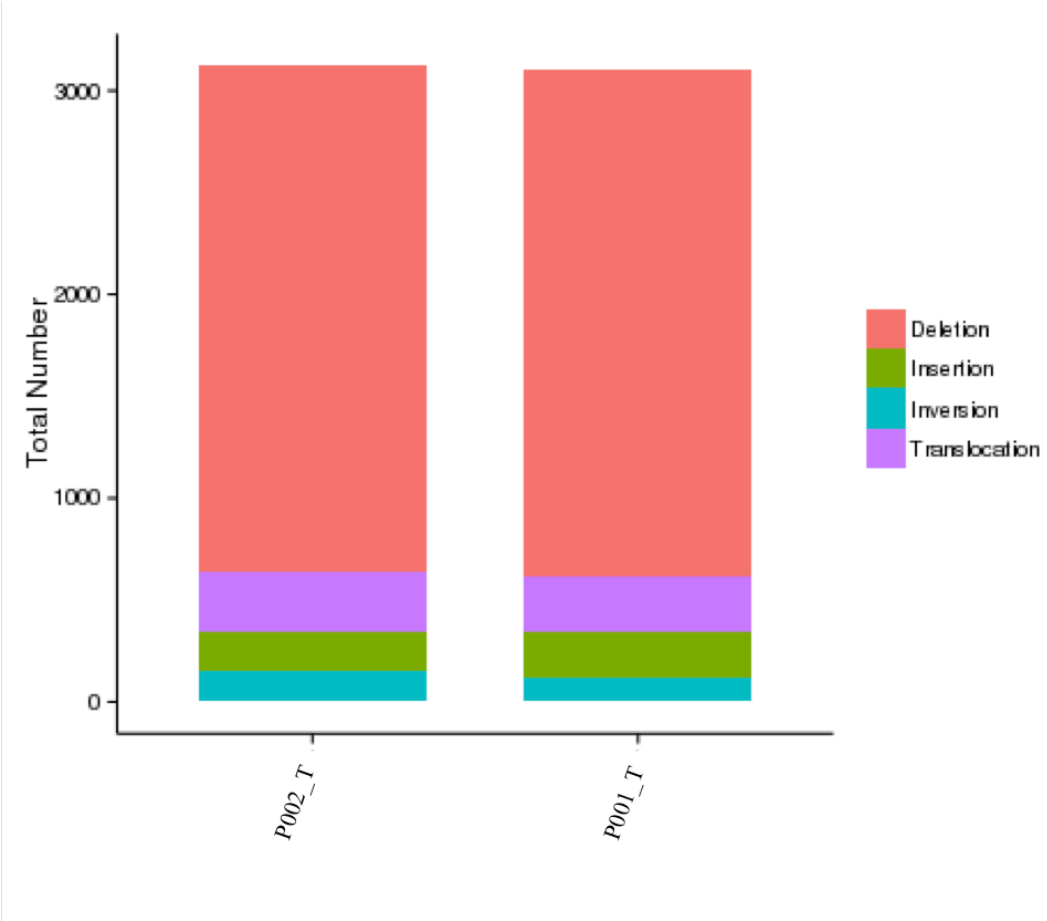

Figure 4.10 Number of different types of somatic SV in each sample

The x-axis is sample, and the y-axis is the number of each type of somatic SV

4.5.4 Somatic CNV Detection Result

Statistics of detected somatic CNVs are listed below:

Table 4.15 Statistics of detected somatic CNVs

Sample	gain_count	gain_size	loss_count	loss_size	total_count	total_size
P001_T	20	160911358	16	140923678	36	301835036
P002_T	324	342898893	83	481484502	407	824383395

(1) Sample: sample name

(2) gain_count: the number of gains

(3) gain_size: the total size of gains

(4) loss_count: the number of losses

(5) loss_size: the total size of losses

(6) total_count: the total number of CNVs

(7) total_size: the total size of CNVs

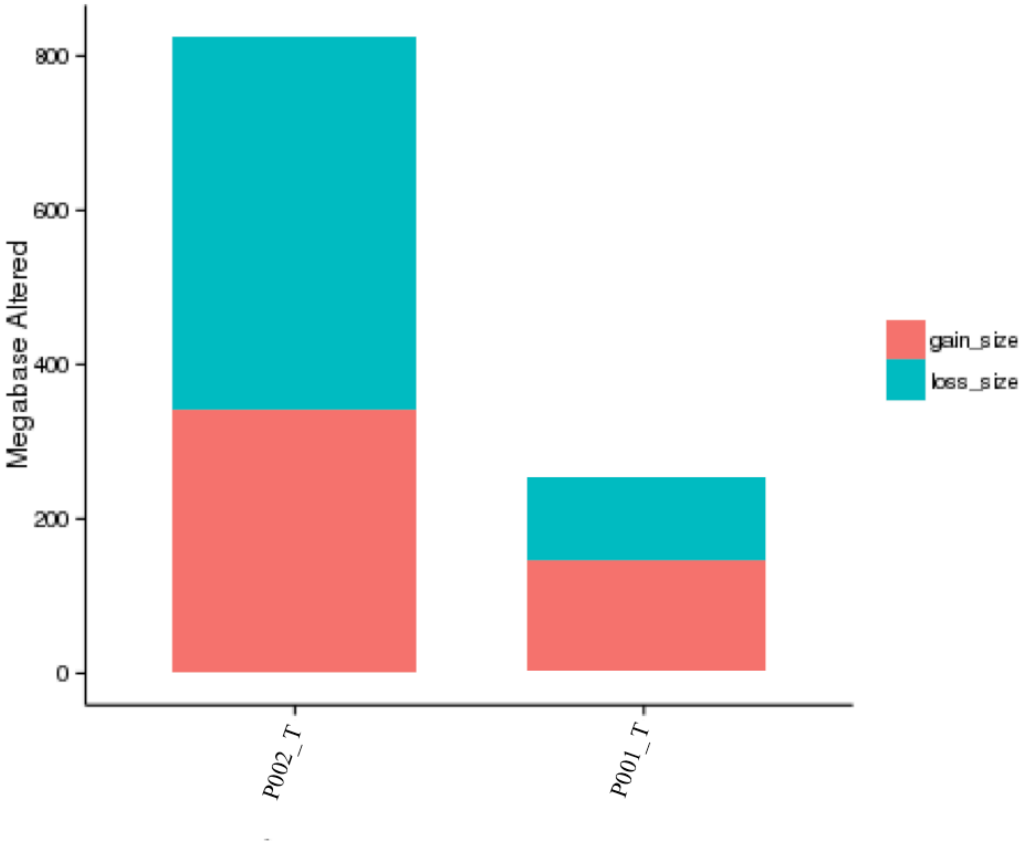

Figure 4.11 The size of genomic regions affected by somatic CNVs in each sample

The x-axis represents samples, and the y-axis is the total size of genomic regions affected by gains or losses (Mb)

5 References

[1] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform[J]. Bioinformatics, 2009, 25(14): 1754-1760. (BWA_MEM)

[2] Kent W J, Sugnet C W, Furey T S, et al. The human genome browser at UCSC[J]. Genome research, 2002, 12(6): 996-1006. (UCSC)

[3] Picard: http://sourceforge.net/projects/picard/ . (Picard)

[4] DePristo M A, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data[J]. Nature genetics, 2011, 43(5): 491-498. (GATK)

[5] Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools[J]. Bioinformatics, 2009, 25(16): 2078-2079. (Samtools)

[6] Sherry S T, Ward M H, Kholodov M, et al. dbSNP: the NCBI database of genetic variation[J]. Nucleic acids research, 2001, 29(1): 308-311. (dbSNP)

[7] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from

high-throughput sequencing data[J]. Nucleic acids research, 2010, 38(16): e164-e164. (ANNOVAR)
[8] 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes[J]. Nature, 2012, 491(7422): 56-65. (1000g)
[9] Hamosh A, Scott A F, Amberger J S, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders[J]. Nucleic acids research, 2005, 33(suppl 1): D514-D517. (OMIM)
[10] Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource[J]. Nucleic acids research, 2004, 32(suppl 1): D258-D261. (GO)
[11] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes[J]. Nucleic acids research, 2000, 28(1): 27-30. (KEGG PATHWAY)
[12] Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnology, 2013.doi:10.1038/nbt.2514.(muTect)
[13] Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK: Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 2012, 28(14):1811-1817. (Strelka)
[14] Chen K, Wallis J W, McLellan M D, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation[J]. Nature methods, 2009,6(9): 677-681. (BreakDancer)
[15] Wang et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nature methods, 2011 Jun 12;8(8):652-4. (Crest)
[16] Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. Control-FREEC: a tool for assessing copy number and allelic content using next generation sequencing data. Bioinformatics. Bioinformatics, 2012, 28(3):423-5. PubMed PMID: 22155870. (Control-FREEC)

6 Appendix

The used softwares in the analysis are listed below:

Table 6.1 Softwares used in the analysis

Analysis	Software	Comments	Version
Alignment	BWA	Map the sequencing reads to the reference genome, and output the alignment file in the bam format	0.7.8-r455
	Samtools	Sort the bam file	1
	Picard	Merge all bam files from the same sample and mark the duplicated reads	1.111
SNP/InDel	GATK	Detect and filter SNPs/InDels	v3.1
SV	breakdancer	Detect SVs	1.4.4
CNV	control-FREEC	Detect CNVs	v6.7
Somatic SNP/InDel	MuTect/Strelka	Detect and filter Somatic SNPs/InDels	muTect:1.1.4; Strelka:v1.0.13
Somatic SV	Breakdancer	Detect somatic SVs	1.4.4
Somatic CNV	Control-FREEC	Detect somatic CNVs	v6.7
Annotation	ANNOVAR	Annotate variants	2015Mar22