Cancer project report

(2014, 11)

Contents

Contents.....2
1 Sample information.....3
2 Experimental procedures.....3
3 Bioinformatics analysis procedures.....5
4 Analysis result.....6
5 References.....32
6 Appendix.....33

1 Sample information

Patient ID	Sample ID	Library ID	Type
Patient1	Sample1	DHG1	N or T
Patient2	Sample2	DHG2	N or T
Patient3	Sample3	DHG3	N or T

Note: Type: (N: normal T: tumor)

2 Experimental procedures

2.1 DNA quantification and qualification

DNA degradation and contamination was monitored on 1% agarose gels.

- DNA purity was checked using the NanoPhotometer® spectrophotometer (IMPLEN, CA, USA)
- DNA concentration was measured using Qubit® DNA Assay Kit in Qubit® 2.0 Flurometer (Life Technologies, CA, USA).
- 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 1.5μg 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 attribute sequences to each sample. Briefly, the DNA sample was fragmented by sonication to a size of 350bp, then DNA fragments were end-polished, 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 Agilent2100 Bioanalyzer and quantified using real-time PCR.

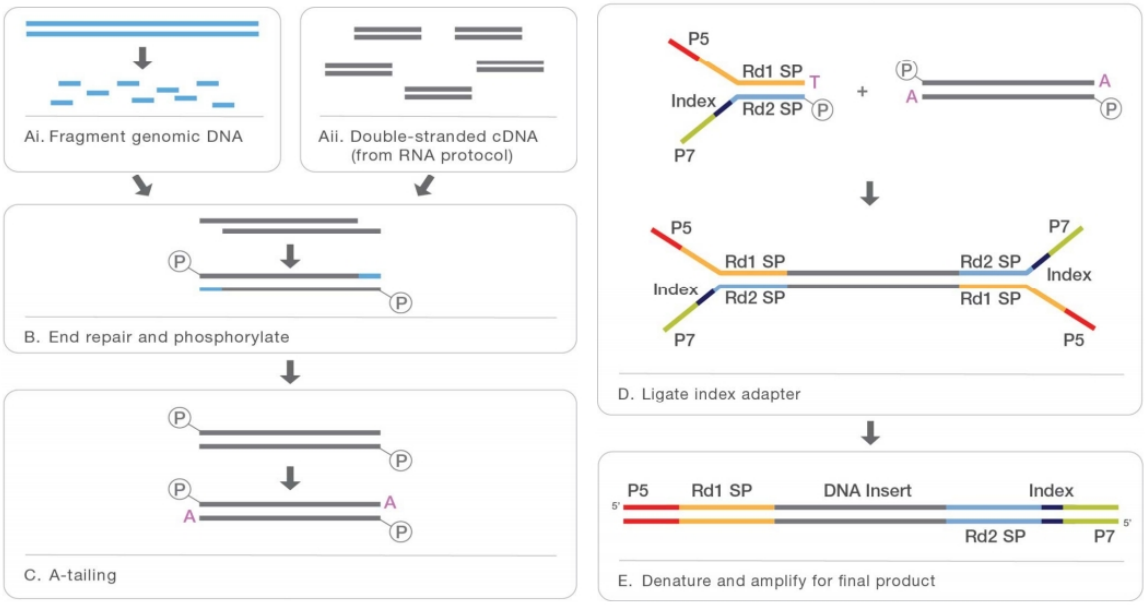


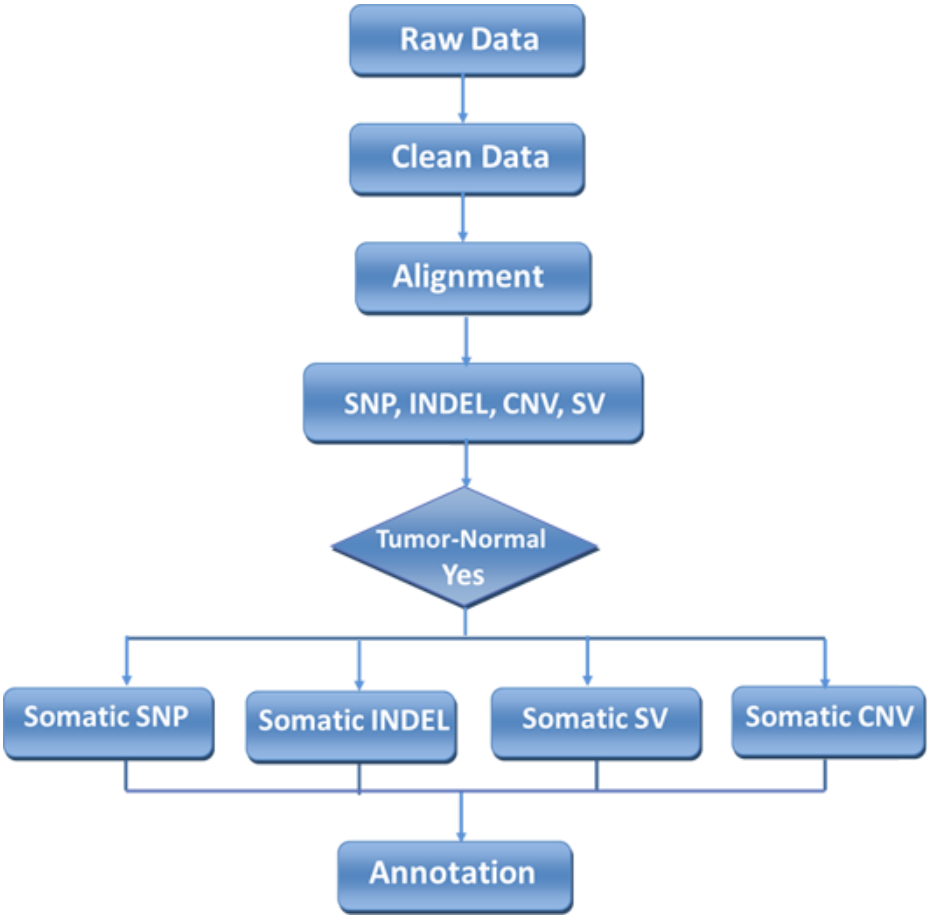
Figure 1 Library construction workflow

2.3 Clustering and sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using HiseqX HD PE Cluster Kit (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illunina Hiseq X Ten platform and paired-end reads were generated.

3 Bioinformatics analysis procedures

If we have reference sequence or reference genome (hg19), bioinformatics analysis as following steps can be applied after obtaining the original sequenced reads.



4 Analysis result

4.1 Raw data

The original raw data obtained from high throughput sequencing platforms (e.g. illumina HiSeq™ XTen//2000/MiSeq) is transformed to sequenced reads by base calling. Raw data is recorded in FASTQ file, 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 and is followed by a sequence identifier and an optional description (like a FASTA title line). Line 2 is the raw sequence reads. Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again. Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of characters as bases in the sequence.

Illumina sequence identifier details:

EAS139	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 for every character at the fourth line minus 33 will be the corresponding sequencing base quality value at the second line. If the sequencing error rate is recorded by “e” and the base quality for illumina HiSeq™ XTen/2000/MiSeq is expressed as Q_{phred}, the equation as below will be obtained:

Equation 1：$Q_{phred} = -10\log_{10}(e)$

The relationship between sequencing error rate (e) and sequencing base quality value(Qphred) is listed as below:

sequencing error rate(e)%	sequencing base quality value(Qphred)	character
5	13	.
1	20	5
0.1	30	?
0.01	40	I

4.2 Sequencing quality control

4.2.1 Sequencing data filtration

Raw data obtained from sequencing contains adapter contamination, low-quality reads. These sequence artifacts can increase the complexity on downstream processing analysis, which means quality control is necessary (listed below). All the downstream analysis will base on clean reads.

The steps of data processing are as follows:

(1) Discard the paired reads when either one read contains adapter contamination.

(2) Discard the paired reads when either one read contains uncertain nucleotides more than 10 percent.

(3) Discard the paired reads when either one read contains low quality nucleotides (base quality less than 5) more than 50 percent.

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

5’ Adapter：

5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3’

3’ Adapter：

5’-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC(6index)ATCTCGTATGCCGTCTTCTGCTTG-3’

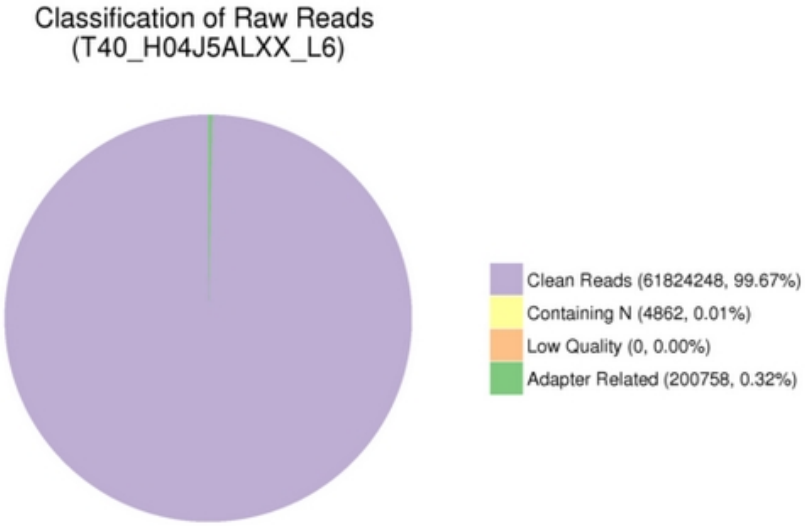


Figure 4.1 Raw data filtration result. (Clean Reads: read pairs passed quality control. Containing N: read pairs with either one read containing uncertain nucleotides more than 10 percent. Low Quality: read pairs with either one read containing low quality nucleotides more than 50 percent. Adapter Related: read pairs with either one read contains adapter contamination).

4.2.2 Sequencing error rate distribution

A Phred score of a base (Phred score, Q_{phred}) is calculated by the equation 1 while the sequencing error rate is obtained from the base calling process. The corresponding relation is listed as below:

Phred score	sequencing error rate	correct sequencing rate	Q-scorce
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

Sequencing platform, chemistry reactant and sample quality all can influence sequencing error rate and base quality. For NGS, sequencing error rate distribution has two features:

(1)Error rate is increasing with sequencing reads extension due to the consumption of chemical reagents. All the illumina high throughput sequencing platforms have this feature.

(2)The front six bases have higher sequencing error rate than others. This length is equal to the random primer which is required in reverse transcription in the process of DNA libraries construction. Hence we can infer that the reason why the front six bases error rate is higher is that the random primer and DNA template are not entirely binding.

Sequencing error rate distribution examination is applied to detect whether there is any abnormal bases with high error rate in the range of sequencing length or not. For example, abnormal bases might present if the middle base sequencing error rate is higher than others. Generally, each bases sequencing error rate should be smaller than 1%.

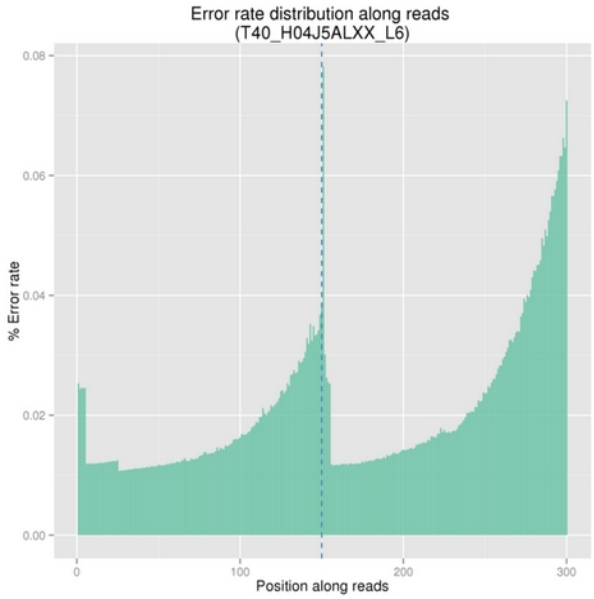


Figure 4.2 Sequencing error rate distribution. The x-axis is the position of base on reads, the y-axis is the average error rate of bases on all reads at this position.

4.2.3 GC content distribution

GC content distribution evaluation is applied to check the potential AT-GC separation phenomenon, which might be produced by sample contamination, sequencing bias or library preparation.

In theory, AT and GC should be equal to each other during every machine cycle, in the meantime their contents should be constant in the whole sequencing procedure. But in practical measurement, due to the primer amplification bias and some other reasons, the first 6 to 7 nucleotides will fluctuate for every read, which is normal and reasonable.

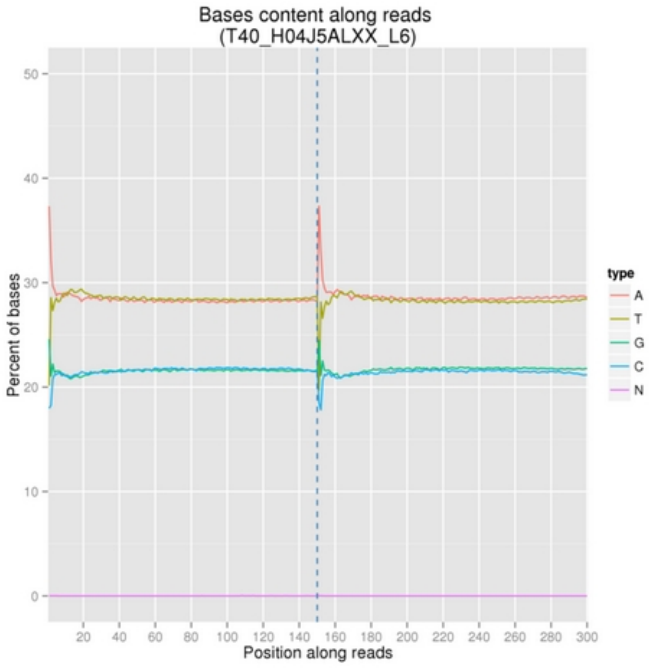


Figure 4.3 GC content distribution. The x-axis is the position of base on reads, the y-axis is single base percentage; each color represents different base type.

4.2.4 Sequencing quality distribution

To ensure downstream analysis, most base quality is required to be greater than Q20. According to sequencing feature, base quality in sequence end is usually lower than that in sequence beginning.

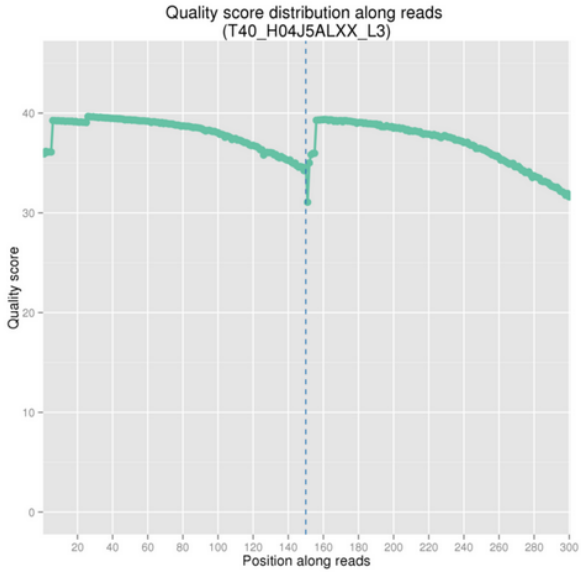


Figure 4.4 Data quality distribution. The x-axis is the position of base on reads, the y-axis is the average quality score of bases on all reads at this position.

4.2.5 Statistics summary of sequencing quality

According to the illumina platform sequencing feature, for PE data, we require that Q30 is above 80%, error rate is below 0.1%.

Table 4-1 Overview of data production quality

Sample[1]	Library[2]	Lane[3]	Raw reads[4]	Raw data(G)[5]	Effective(%)[6]	Error(%)[7]	Q20(%)[8]	Q30(%)[8]	GC(%)[9]
101	DHG00121	H04C3ALXX_L2	395161268	110.40	99.34	0.03;0.05	97.22;97.35	90.03;85.53	43.47%;43.39%
2	DHE00122	H04C3ALXX_L1	406170049	113.42	97.17	0.03;0.03	97.04;97.82	90.54;92.58	43.22;43.15
3	DHE00123	H04C3ALXX_L4	376387852	105.16	97.11	0.04;0.03	96.91;97.67	90.25;92.22	43.52;43.46
4	DHE00124	H04C3ALXX_L5	350735983	97.99	98.22	0.03;0.05	97.88;96.61	90.11;86.77	43.42;43.34

Note:

(1) Sample: sample name

(2) Library: Library name.

(3) Lane: The flowcell ID and lane number of the sequencing on Hiseq machine

(4) Raw reads：the number of sequencing reads pairs; According to the format of FASTQ, four lines will be considered as one unit.

(5) Raw data：the original sequence data

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

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

(8) Q20,Q30：percentage of bases with average quality>Q20 and percentage of bases with average quality>Q30

(9) GC content：Percentage of G and C in the total number of bases

4.3 Sequence alignment

Burrows-Wheeler Aligner (BWA) (Li H et al.) software is utilized to map the paired-end clean reads to the reference genome (UCSC hg19). The original mapping result in BAM format can be obtained. Picard, GATK (DePristo M A et al.) and Samtools (Li H et al.) are then used to do duplicate removal, local realignment, and base quality recalibration etc. Final BAM file can be obtained after these steps. We computed the coverage and depth based on the final BAM file.

Generally, human sample sequencing reads can reach above 95% mapping ratio. SNPs called from sites with more than 10X read depth are more confident than other sites.

4.3.1 Sequencing depth, coverage distribution

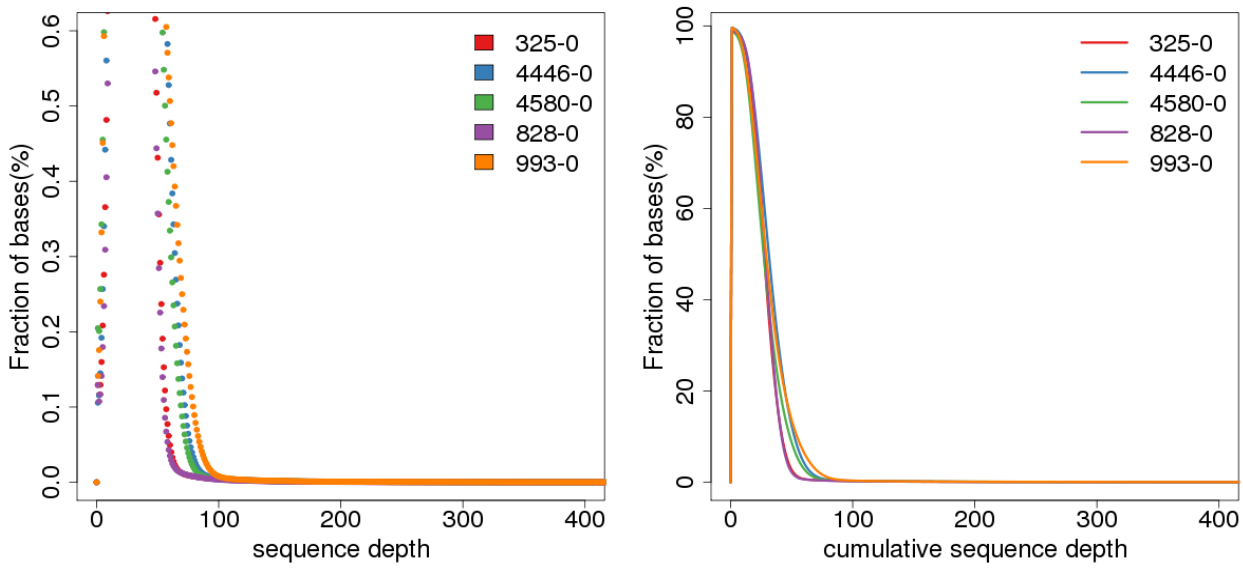


Figure 4.5 Sequencing depth

The left figure is the ratio of bases with different sequencing depth. The x-axis is sequencing depth; the y-axis is the fraction of bases with the given sequencing depth.

The right figure is accumulative base ratio with different depth. The x-axis is sequencing depth, the y-axis is the fraction of bases above the given sequencing depth. For example, the sequencing depth of 0x corresponds to the base ratio of 100%, showing that 100% base's sequencing depth >0X.

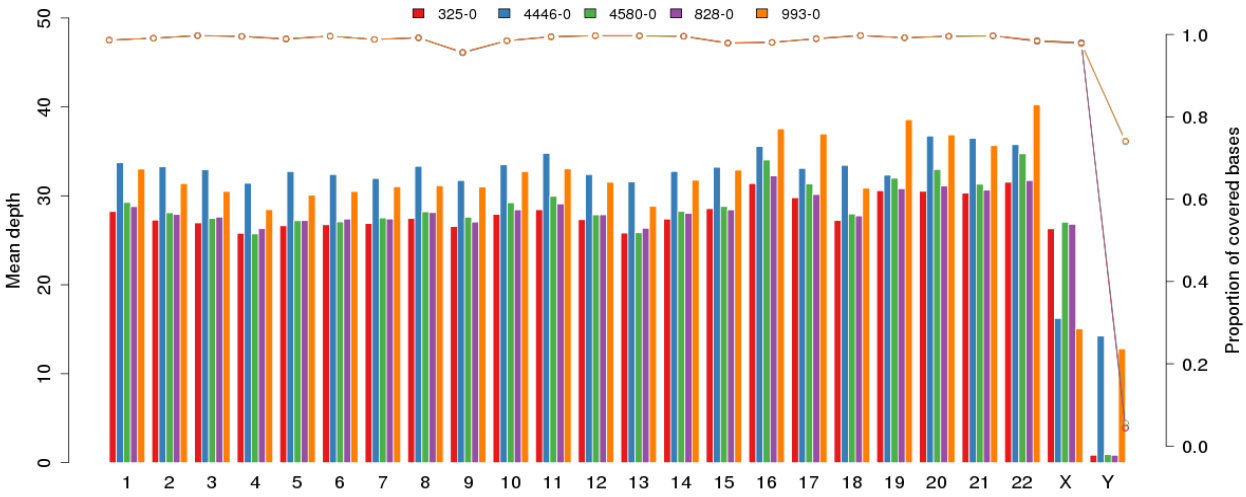


Figure 4.6 The coverage depth (the left coordinate) and coverage rate (the right coordinate) of chromosome

The x-axis is chromosome number; the left y-axis is the average depth of each chromosome; the right y-axis is the fraction of covered on each chromosome.

4.3.2 Statistics of coverage

Table 4-2 Mapping rate and coverage

Sample:[1]	1	2	3	4
Total:[2]	569112086 (100.00%)	657637118 (100.00%)	587520668 (100.00%)	577318526 (100.00%)
Duplicate:[3]	5544498 (0.97%)	8023172 (1.22%)	4641413 (0.79%)	6523699 (1.13%)
Mapped:[4]	568356915 (99.87%)	657175352 (99.93%)	585649459 (99.68%)	576584835 (99.87%)
Properly mapped:[5]	553815332 (97.31%)	628278125 (95.54%)	559519912 (95.23%)	559833922 (96.97%)
PE mapped:[6]	569112086 (100.00%)	656973443 (99.90%)	585371799 (99.63%)	576335044 (99.83%)
SE mapped:[7]	284178 (0.05%)	201909 (0.03%)	277660 (0.05%)	249791 (0.04%)
with mate mapped to a different chr:[8]	4204534 (0.74%)	6807704 (1.04%)	6952981 (1.18%)	5124785 (0.89%)
with mate mapped to a different chr (mapQ>=5):[9]	2982932 (0.52%)	4003436 (0.61%)	4108167 (0.70%)	2566762 (0.44%)
Average_sequencing_depth:[10]	28.63	33.44	29.64	29.22
Coverage:[11]	98.93%	99.66%	98.91%	98.96%
Coverage_at_least_4X:[12]	98.56%	99.29%	98.26%	98.61%
Coverage_at_least_10X:[13]	96.47%	96.82%	94.00%	96.83%

Coverage_at_least_20X:[14]	77.15%	81.06%	71.04%	79.85%

Note：

(1) Sample：sample name

(2) Total: The number of total clean reads

(3) Duplicate: The number of duplication reads

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

(5) Properly mapped: The number of reads that mapped to the reference genome and the direction is right

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

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

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

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

(10) Average_sequencing_depth: The average sequencing depth that mapped to the reference genome

(11) Coverage: The sequence coverage of genome

(12) Coverage_at_least_4X: The percentage of bases with depth >4X in whole genome bases

(13) Coverage_at_least_10X: The percentage of bases with depth >10X in whole genome bases

(14) Coverage_at_least_20X: The percentage of bases with depth >20X in whole genome bases

4.4 Variation detection result

4.4.1 SNP detection result

Generally, the whole genome of human has about 3.6M SNP. Most (above 95%) SNPs with high frequency (the allele frequency in population is above 5%) is collected in dbSNP (Sherry S T et al.) .The ratio of Ts/Tv can reflect the accuracy of sequencing. Generally, this ratio in genome is about 2.2 and in coding region is about 3.2.

The statistics of SNP are as follows:

Table 4-3 The number of SNP in different genomic region

Sample ID[1]	1	2	3	4
CDS[2]	22269	22619	19533	22808
synonymous_SNP[a]	11352	11565	10227	11563
missense_SNP[b]	10429	10562	8965	10692
stopgain_SNP[c]	82	80	47	93
stoploss_SNP[d]	15	15	12	11
Unknown[e]	391	397	282	449
Intronic[3]	1246455	1254792	1085383	1258808
UTR3[4]	24569	24804	21665	24982
UTR5[5]	5544	5618	4660	5592
Intergenic[6]	2148355	2154191	1793788	2167409
ncRNA_exonic[7]	9747	9687	8018	9740
ncRNA_intronic[8]	146348	146410	123419	148839
Upstream[9]	22640	22790	18606	23180
Downstream[10]	22224	22421	19113	22865
Splicing[11]	579	582	2374	598
ncRNA_UTR3[12]	587	590	521	577
ncRNA_UTR5[13]	95	92	76	117
ncRNA_splicing[14]	113	113	222	104
Total[15]	3649525	3664709	3096731	3685619

Note：

(1) Sample：Sample name

(2) CDS：the number of SNP in exonic region

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

b) missense_SNP: a single nucleotide change that cause an amino acid change

c) stopgain_SNP: a nonsynonymous SNP that lead to the immediate creation of stop codon at the variant site

d) stoploss_SNP: a nonsynonymous SNP that lead to the immediate elimination of stop codon at the variant site

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

(3) intronic：the number of SNP in intronic region

(4) UTR3：the number of SNP in 3'UTR region

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

(6) intergenic：the number of SNP in intergenic region

(7) ncRNA_exonic：the number of SNP in non-coding RNA exonic region

(8) ncRNA_intronic：the number of SNP in non-coding RNA intronic region

(9) upstream：the number of SNP in the 1kb upstream region of transcription start site

(10) downstream：the number of SNP in the 1kb downstream region of transcription ending site

(11) splicing：the number of SNP in 4bp splicing junction region

(12) ncRNA_UTR3：the number of SNP in 3'UTR of non-coding RNA

(13) ncRNA_UTR5：the number of SNP in 5'UTR of non-coding RNA

(14) ncRNA_splicing：the number of SNP in 4bp splicing junction of non-coding RNA

(15) Total: the total number of SNP

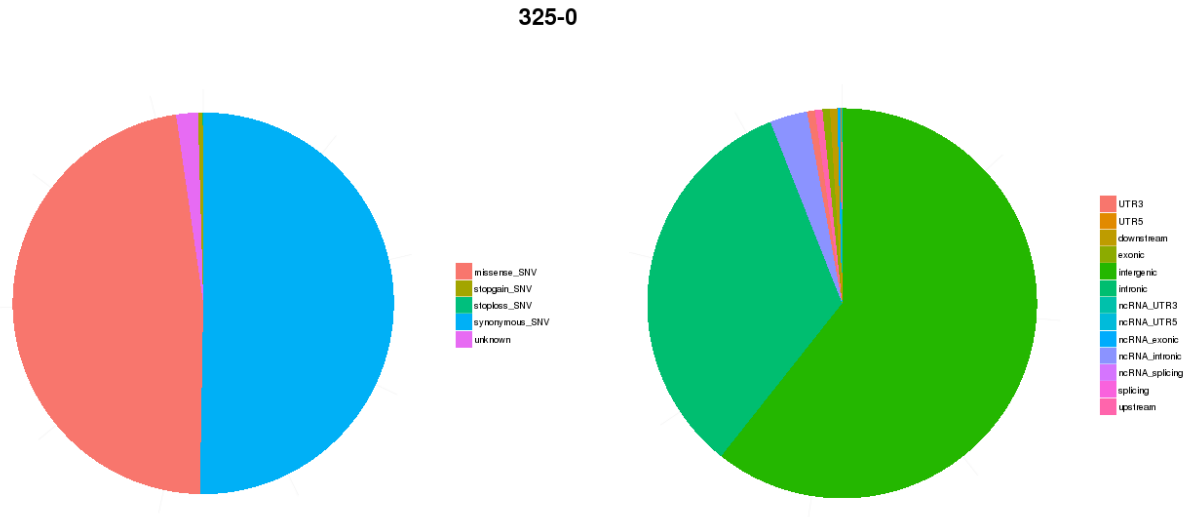


Figure 4.7 The number of SNP of different types in coding region (left), the number of SNPs in different genomic regions (right).

Table 4-4 Feature of SNP in genome.

Sample[1]	1	2	3	4
Total[2]	3649525	3664709	3685619	3636874
Het[3]	2051819	2080735	2090854	2053565
Hom[4]	1597706	1583974	1594765	1583309
Transition[5]	2451247	2465539	2479472	2447806
Transvertion[6]	1198278	1199170	1206147	1189068
ts/tv[7]	2.05	2.06	2.06	2.06
dbSNP percentage[8]	3571583 (97.86%)	3593313 (98.05%)	3612104 (98.01%)	3568498 (98.12%)
Novel[9]	77942	71396	73515	68376
novel ts[10]	45796	42625	43849	41286
novel tv[11]	32146	28771	29666	27090
novel ts/tv[12]	1.42	1.48	1.48	1.52

Note:

(1) Sample：Sample name

(2) Total：the total number of SNP

(3) Het：the genotype of heterozygote

(4) Hom：the genotype of homozygote

(5) transition（ts）：transition

(6) transversion（tv）：transversion

(7) ts/tv: is calculated as the number of transition/ the number of transversion

(8) dbSNP percentage: is calculated as the number of SNP in dbSNP/total number of SNP.

(9) novel: SNP not in dbSNP

(10) novel ts：the number of ts SNP that are not in dbSNP.

(11) novel tv：the number of tv SNP that are not in dbSNP.

(12) novel ts/tv：is calculated as novel ts/ novel tv

4.4.2 INDEL detection result

Generally, the genome of human has about 350K INDEL (insertion and deletion, less than 50bp insertion and deletion).

The INDEL in coding region or splicing site may change the protein translation. Frameshift mutation, in which the number of inserted or deleted bases is not an integral multiple of three, may lead to the change of the whole reading frame. Frameshift mutation is more limited than nonframeshift mutation because of selective pressure. The statistics of INDEL are as follows:

Table 4-5 The number of INDEL in different genomic regions

Sample[1]	1	2	3	4
CDS[2]	712	697	723	694
frameshift_deletion[a]	136	140	143	131
frameshift_insertion[b]	115	102	122	108
nonframeshift_deletion[c]	190	188	198	190
nonframeshift_insertion[d]	169	164	152	164
stopgain[e]	10	9	7	5
stoploss[f]	1	1	1	0
unknown[g]	91	93	100	96
Intronic[3]	243756	220986	223074	214883
UTR3[4]	5571	5147	5191	5013
UTR5[5]	876	857	831	848
Splicing[6]	192	179	178	165
ncRNA_exonic[7]	1321	1250	1211	1189
ncRNA_intronic[8]	27771	25284	25838	24779
ncRNA_UTR3[9]	139	127	134	113

ncRNA_UTR5[10]	23	25	22	23
ncRNA_splicing[11]	29	29	26	23
Upstream[12]	5086	4679	4725	4628
Downstream[13]	4954	4550	4689	4475
Intergenic[14]	383120	350218	353706	341359
Total[15]	673550	614028	620348	598192

Note：

(1) Sample：Sample name

(2) CDS：the number of INDEL in exonic region

a) frameshift_deletion：a deletion of one or more nucleotides that cause frameshift changes in protein coding sequence. the deletion length is not multiple of 3

b) frameshift_insertion：an insertion of one or more nucleotides that cause frameshift changes in protein coding sequence.the insertion length is not multiple of 3.

c) nonframeshift_deletion：non-frameshift deletion, does not change coding protein frame deletion, the deletion length is multiple of 3

d) nonframeshift_insertion：non-frameshift insertion, does not change coding protein frame insertion: the insertion length is multiple of 3

e) stopgain：frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate creation of stop codon at the variant site

f) stoploss：frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate elimination of stop codon at the variant site

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

(3) intronic：the number of INDEL in intronic region

(4) UTR5：the number of INDEL in 5’UTR region

(5) UTR3：the number of INDEL in 3’UTR region

(6) Splicing：the number of INDEL in 10bp splicing junction region

(7) ncRNA_exonic：the number of INDEL in non-coding RNA exonic region

(8) ncRNA_intronic：the number of INDEL in non-coding RNA intronic region

(9) ncRNA_UTR3：the number of INDEL in 3'UTR of non-coding RNA

(10) ncRNA_UTR5：the number of INDEL in 5'UTR of non-coding RNA

(11) ncRNA_splicing：the number of INDEL in 10bp splicing junction of non-coding RNA

(12) upstream：the number of INDEL in the 1kb upstream region of transcription start site

(13) downstream：the number of INDEL in the 1kb downstream region of transcription ending site

(14) intergenic：the number of INDEL in intergenic region

(15) Total：the total number of INDEL

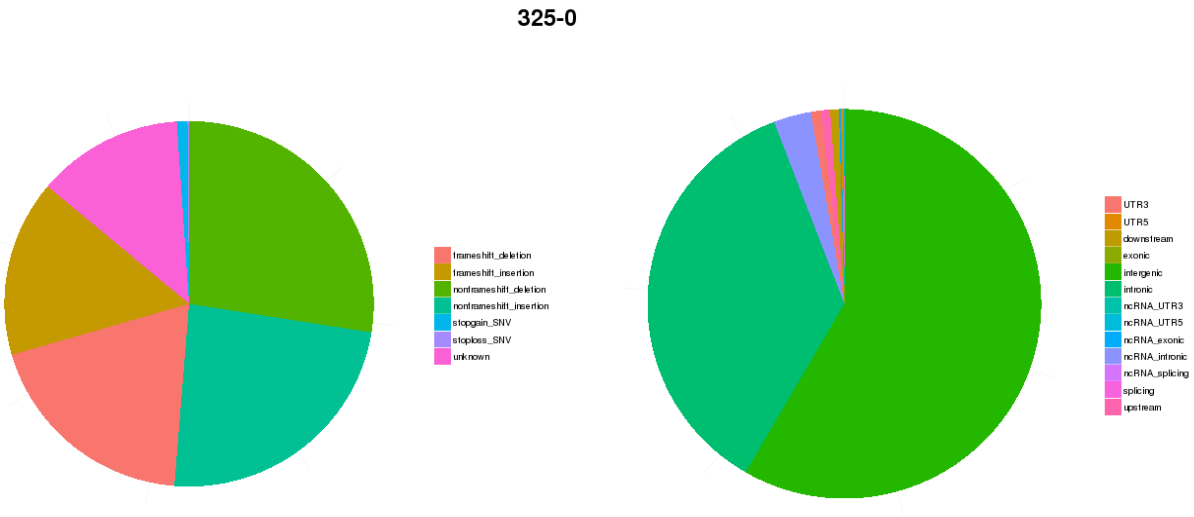


Figure 4.8 The number of different type INDEL in coding region (left), the number of INDEL in different genomic regions right).

Table 4-6 Feature of INDEL in genome

Sample[1]	1	2	3	4	Sample
Total[2]	673550	614028	620348	598192	Total
Het[3]	353002	315630	319841	305740	Het
Hom[4]	320548	298398	300507	292452	Hom
dbSNP percentage[5]	513916 (76.30%)	478656 (77.95%)	483251 (77.90%)	467994 (78.23%)	483062 (77.80%)
Novel[6]	159634	135372	137097	130198	137827

Note:

(1) Sample：Sample name

(2) Total：the total number of INDEL

(3) Het：the genotype of heterozygote

(4) Hom：the genotype of homozygote

(5) dbSNP percentage：is calculated as the number of INDEL in dbSNP/total number of INDEL

(6) novel：INDEL not in dbSNP

4.4.3 SV detection result

SV (structural variation) is the large structural variation in genome, such as deletion, insertion, duplication, copy number variations, inversion and translocation of large fragment. Generally, the sequence length related to SV is between 1kb and 3Mb. SV, which is the source of the individual difference and the disease susceptibility, is widespread in human genome. SV may lead to fusion genes which have been proved to be related to cancer. The statistics of SV are as follows:

Table 4-7 Statistics summary of SV detection result

Sample[1]	varType[2]	Total[3]	Cds[4]	Splicing[5]	UTR5[6]	UTR3[7]	Intron[8]	Upstream[9]	Downstream[10]	ncRNA[11]	Intergenic[12]	Unknown[13]
1	Deletion	2058	57	1	1	3	711	19	19	74	1173	0
	Translocation	330	5	0	1	5	104	4	1	14	196	0
	Insertion	43	3	0	0	1	19	0	0	1	19	0
	Inversion	126	38	0	0	0	30	1	0	13	44	0

Note：each column shows the number of different type of SV in each genomic region.

(1) Sample: sample name
(2) varType：SV Type
(3) total：the total number of SV
(4) CDS：the number of SV in CDS region
(5) splicing：the number of SV in splicing junction region
(6) UTR5：the number of SV in 5'UTR region
(7) UTR3：the number of SV in 3'UTR region
(8) intronic：the number of SV in intronic region
(9) intergenic：the number of SV in intergenic region
(10) upstream：the number of SV in 1Kb region upstream from transcription start site
(11) downstream：the number of SV in 1Kb region downstream from transcription end site
(12) ncRNA: the number of SV in ncRNA region
(13) unknown：the number of SV in region with unknown function (due to various errors in the gene structure definition in the database file)

4.4.4 CNV detection result

CNV (copy number variation) refers to the increase or reduction of copy number of large fragment in the genome and is a very important molecular mechanism. There are two types of CNV: deletion and duplication. Abnormal CNV change can be the cause of many diseases. Thus, it has already been the hotspot of disease study. The CNV detection result is as follows:

Table 4-8 CNV detection result

Sample[1]	varType[2]	Total[3]	Cds[4]	Splicing[5]	utr5[6]	utr3[7]	Intron[8]	Upstream[9]	Downstream[10]	ncRNA[11]	Intergenic[12]	Unknown[13]
1	Deletion	2643	66	0	1	4	700	18	16	94	1744	0
	Duplication	810	127	0	6	3	38	5	10	116	505	0
2	Deletion	2324	67	0	6	5	575	14	14	67	1576	0
	Duplication	2081	620	0	16	7	242	16	18	156	1006	0
3	Deletion	2029	47	0	2	1	525	15	18	65	1356	0
	Duplication	823	200	0	4	5	67	9	9	89	440	0

3	Deletion	1858	55	0	4	3	483	16	18	65	72	1207
	Duplication	3005	1207	0	19	10	307	10	14	187	1251	0

Note: each column shows the number of different type of CNV in each genomic region.

(1) Sample: sample name
(2) varType：CNV type
(3) total：the total number of CNV
(4) CDS：the number of CNV in CDS region
(5) splicing：the number of CNV in splicing junction region
(6) UTR5：the number of CNV in 5'UTR region
(7) UTR3：the number of CNV in 3'UTR region
(8) intronic：the number of CNV in intronic region
(9) upstream：the number of CNV in 1Kb region upstream from transcription start site
(10) downstream：the number of CNV in 1Kb region downstream from transcription end site
(11) ncRNA: the number of CNV in ncRNA region
(12) intergenic: the number of CNV in intergenic region
(13) unknown：the number of CNV in region with unknown function (due to various errors in the gene structure definition in the database file)

4.4.5 Annotation result of variation site

ANNOVAR (Wang K *et al.*) , which contains dbSNP database, the 1000-genome project and other published databases, is utilized to do annotation. Variant position, variant type, conservative prediction and other information can be obtained at this step.

- Refseq and Gencode databases are used to annotate the gene structure of variation site (exons, introns,etc). The gene type, including mRNA, non-coding RNA, smallRNA and microRNA, can also be obtained.
- The genome features of variation site include the CG island, cell cytoband, phastConsElements46way conservative regions, transcription factor binding sites , the annotation of structure and functional elements in the genome of cell line Gm12878
- SIFT, PolyPhen, MutationAssessor and LRT method were used to assess the effect of non-synonymous mutations on tumor or disease.
- The annotation results of dbSNP, 1000 genome, Hapmap, Cosmic（known tumor somatic mutation database）and esp6500 database are also provided.
- GO (biology process、cell component molecular function), KEGG, Reactome, Biocarta, PID are applied to do functional annotation.

Table 4-9 The annotation result of variation site

CHROM[1]	POS[2]	ID[3]	REF[4]	ALT[5]	QUAL[6]	FILTER[7]	GeneName[8]	Func[9]	Gene[10]	GeneDetail[11]	[12-60]
1	13656	.	CAG	C	1467.73	PASS		ncRNA_exonic	DDX11L1	.	.
1	109575	.	C	CGT	99.74	PASS		intergenic	OR4F5(dist=39567),		.

								LOC729737(dist=25198)			
1	895755	.	A	AG	323.75	PASS	upstream	KLHL17	.	.	
1	1209183	.	G	GCGC	417.73	PASS	UTR5	UBE2J2	.	.	
1	1823922	.	ATCC	G	319.77	PASS	intergenic	UBE2J2(dist=2368), SCNN1D(dist=4214)	.	.	

Note:

(1) CHROM: chromosome.

(2) POS: the position of variation on chromosome.

(3) ID: the ID of the variation in dbSNP.

(4) Ref: base at this position in reference genome.

(5) Alt: base at this position in sequencing data.

(6) QAUL: variation quality value.

(7) FILTER: the TAG of filtration, if the site satisfies all the filtration conditions, mark PASS.

(8) GeneName: the names of genes related to this variation

(9) Func: tells whether the variant hit exons or hit intergenic regions, or hit introns, or hit a non-coding RNA genes. (exonic, splicing, UTR5, UTR3, intronic, ncRNA_exonic, ncRNA_intronic, ncRNA_UTR3, ncRNA_UTR5, ncRNA _splicing, upstream, downstream, intergenic).

(10) Gene: the IDs of transcript whose function has changed like the value in column 'Func'.

(11) GeneDetail: description of variations in UTR, splicing, ncRNA, splicing or intergenic.

(12) ExonicFunc: the amino acid changes as a result of the exonic variant.(synonymous_SNP, missense_SNP, stopgain_SNP, stopgloss_SNP or unknown)

(13) AAChange: when 'Func' equals 'exonic'or 'exonic;splicing', this value gives the change of amino acid in each related transcript. For example, AIM1L:NM_001039775:exon2:c.C2768T:p.P923L, AIM1L is gene name containing this variation; NM_001039775 is ID of transcript; exon2 means the variation is on the second exon of the transcript; c.C2768T means the 2,768 base on cDNA is changed from C to T due to this variation; p.P923L means the 923 amino acid on protein is changed from Pro to Leu due to this variation.

(14) Gencode: gene name in Gencode.

(15) cpgIslandExt: the result of prediction of CpG islands.

(16) cytoband: chromosome band.

(17) wgRna: snoRNA and miRNA annotation.

(18) targetScanS: miRNA target prediction by TargetScan.

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

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

(21) genomicSuperDups: repetitive segments in genome. For a region to be included in the track, at least 1 Kb of the total sequence (containing at least 500 bp of non-RepeatMasked sequence) had to align and a sequence identity of at least 90% was required.

(22) dgvMerged: annotation from Database of Genomic Variants.

(23) gwasCatalog: tells whether this variation has been identified by published Genome-Wide Association Studies (GWAS),

collected in the Catalog of Published Genome-Wide Association Studies at the National Human Genome Research Institute (NHGRI). It lists the diseases related to this variation. "." means this variation has not been reported by published GWAS study.

(24) Repeat: annotation of repeats from RepeatMasker program.

(25) encodeGm12878: predicted functional elements in the genome of cell line Gm12878. ChromHMM was used to get twenty-five states by integrating ENCODE ChIP-seq, DNase-seq, and FAIRE-seq data; these states were used to segment the genome, and they were then grouped and colored to highlight predicted functional elements. (The relationship between state and functional element is as follow: Tss, TssF——Active Promoter; PromF——Promoter Flanking; PromP——Inactive Promoter; Enh, EnhF——Candidate Strong enhancer; EnhWF, EnhW, DNaseU, DNaseD, FaireW——Candidate Weak enhancer/DNase; CtrcfO, Ctcf——Distal CTCF/Candidate Insulator; Gen5', Elon, ElonW, Gen3', Pol2, H4K20——Transcription associated; Low——Low activity proximal to active states; ReprD, Repr, ReprW——Polycob repressed; Quies, Art——Heterochromatin/Repetitive/Copy Number Variation)

(26) encodeH1hesc: predicted functional elements in the genome of cell line H1-hESC.

(27) encodeHelas3: predicted functional elements in the genome of cell line HeLa-S3.

(28) encodeHepg2: predicted functional elements in the genome of cell line HepG2.

(29) encodeHuvec: predicted functional elements in the genome of cell line HUVEC.

(30) encodeK562: predicted functional elements in the genome of cell line K562.

(31) snp138: the ID of this variation in dbSNP(version 138)

(32) snp138NonFlagged: tells whether this variation is in snp138NonFlagged. snp138NonFlagged is dbSNP after removing those flagged SNPs (SNPs < 1% minor allele frequency (MAF) (or unknown), mapping only once to reference assembly, flagged in dbSnp as "clinically associated").

(33) 1000g2012apr_eur: gives the alternative allele frequency data of this variation in 1000 Genomes Project (published in April, 2012) for European population. ALL, AMR (admixed American), EUR (European), ASN (), AFR (African) populations

(34) 1000g2012apr_asn: gives the alternative allele frequency data of this variation in 1000 Genomes Project (published in April, 2012) for Asian population.

(35) 1000g2012apr_afr: gives the alternative allele frequency data of this variation in 1000 Genomes Project (published in April, 2012) for African population.

(36) 1000g2012apr_amr: gives the alternative allele frequency data of this variation in 1000 Genomes Project (published in April, 2012) for admixed American population.

(37) 1000g2012apr_all: gives the alternative allele frequency data of this variation in 1000 Genomes Project (published in April, 2012) for ALL population.

(38) hapmapCHB_allele: gives the allele frequency data of this variation in The HapMap Project for Han Chinese in Beijing, China (CHB).

(39) hapmapCHB_genotype: gives the genotype frequency data of this variation in The HapMap Project for Han Chinese in Beijing, China (CHB).

(40) esp6500si_all: gives alternative allele frequency of this variation in all subjects in the NHLBI-ESP project with 6500 exomes.

(41) cosmic68: cosmic database

(42) ljb23_sift: whole-exome SIFT scores with missing values imputed (version 2.3), including raw score, transformed score (0-1, higher values more deleterious, calculated as 1-SIFT) and categorical prediction. The smaller the SIFT raw score is, the more deleterious this variation is.(D: Deleterious (sift<=0.05); T: tolerated (sift>0.05))

(43) ljb23_pp2hvar: whole-exome PolyPhen 2 scores built on HumanVar database (for Mendelian phenotypes) (version 2.3), including raw score and categorical prediction. The larger the raw score is, the more deleterious this variation is.(D: Probably damaging (>=0.909), P: possibly damaging (0.447<=pp2_hvar<=0.909); B: benign (pp2_hvar<=0.446))

(44) ljb23_pp2hdiv: whole-exome PolyPhen 2 scores built on HumanDiv database (for complex phenotypes) (version 2.3), including raw score and categorical prediction. The larger the raw score is, the more deleterious this variation is.(D: Probably damaging (>=0.957), P: possibly damaging (0.453<=pp2_hdiv<=0.956); B: benign (pp2_hdiv<=0.452))

(45) ljb23_mt: whole-exome MutationTaster scores (version 2.3), including raw score, transformed scores (0-1, with higher values more deleterious) and categorical prediction. The larger the transformed score is, the more deleterious this variation is. "A" ("disease_causing_automatic"); "D" ("disease_causing"); "N" ("polymorphism"); "P" ("polymorphism_automatic")

(46) ljb23_lrt: whole-exome LRT scores (version 2.3), including raw score, transformed scores (0-1, with higher values more deleterious) and categorical prediction. (D: Deleterious; N: Neutral; U: Unknown)

(47) ljb23_metalr: whole-exome MetaLR scores, including raw score and categorical prediction. (D: Deleterious; T: Tolerated) The smaller the raw score is, the more deleterious this variation is.

(48) INFO: information about this variation from variation calling software.

(49) FORMAT: Comma-separated list of several tags from variation calling software.

a) GT: Genotype. 0 represents allele same to REF; 1, 2, 3 et. al. represents allele different from REF. 0/0 and 1/1 represent homogeneous genotype. 0/1 represents heterozygous genotype.

b) PL: List of Phred-scaled genotype likelihoods.

c) DP: Number of high-quality bases.

d) DV: Number of high-quality non-reference bases.

(50) OMIM: annotation from OMIM database.

(51) CancerGene: Ocogene and Antioncogene databases

(52) BertVogelstein125: 125 mut-driver genes from BertVogelstein's paper

(53) Predisposition: Susceptibility gene

(54) DriverCNA: Driver mutation

(55) Rearrangement: Gene structure rearrangement

(56) GO_BP, GO_CC, GO_MF: annotation from Gene Ontology

(57) KEGG_PATHWAY: annotation from KEGG.

(58) PID_PATHWAY: annotation from PID (the Pathway Interaction Database).

(59) BIOCARTA_PATHWAY: annotation from BIOCARTA.

(60) REACTOME_PATHWAY: annotation from Reactome Pathway Database.

4.5 Somatic variation detection

Somatic mutation, which will not lead to offspring genetic change, is the mutation in the somatic cell except for germline cell. Somatic mutation, especially driver mutation, plays a crucial role in oncogenesis and tumor development. Cancer drug resistance is also related to the somatic mutation. So somatic mutation is the major target on the study of the cancer genome.

4.5.1 Somatic SNP detection

We use muTect to find Somatic SNP, and use Strelka to find Somatic INDEL. The detection result of Somatic SNP is as follows:

Table 4-10 The number of Somatic SNP in different genomic region

Sample ID[1]	T23	T40	T25	T24
CDS[2]	66	43	424	88
synonymous_SNP[a]	0	0	0	0
missense_SNP[b]	0	0	0	0
stopgain_SNP[c]	3	2	23	3
stoploss_SNP[d]	0	0	0	0
Unknown[e]	1	0	5	0
Intronic[3]	2161	2105	11194	2484
UTR3[4]	31	37	263	32
UTR5[5]	12	5	92	8
Intergenic[6]	2	1	19	1
ncRNA_exonic[7]	14	14	101	18
ncRNA_intronic[8]	298	376	1278	357
Upstream[9]	4	0	5	0
Downstream[10]	0	1	2	0
Splicing[11]	0	0	4	1
ncRNA_UTR3[12]	54	44	298	59
ncRNA_UTR5[13]	43	38	218	46
ncRNA_splicing[14]	3624	4567	17179	4831
Total[15]	3649525	3664709	3096731	3685619

Note：

(1) Sample：Sample name

(2) CDS：the number of Somatic SNP in exonic region

a) synonymous_ SNP: a single nucleotide change that does not cause an amino acid change

b) missense_ SNP: a single nucleotide change that cause an amino acid change
c) stopgain: a nonsynonymous SNP that lead to the immediate creation of stop codon at the variant site
d) stoploss: a nonsynonymous SNP that lead to the immediate elimination of stop codon at the variant site
e) unknown: unknown function (due to various errors in the gene structure definition in the database file)

(3) intronic：the number of Somatic SNP in intronic region
(4) UTR3：the number of Somatic SNP in 3'UTR region
(5) UTR5：the number of Somatic SNP in 5'UTR region
(6) intergenic：the number of Somatic SNP in intergenic region
(7) ncRNA_exonic：the number of Somatic SNP in non-coding RNA exonic region
(8) ncRNA_intronic：the number of Somatic SNP in non-coding RNA intronic region
(9) upstream：the number of Somatic SNP in the 1kb upstream region of transcription start site
(10) downstream：the number of Somatic SNP in the 1kb downstream region of transcription ending site
(11) splicing：the number of Somatic SNP in 10bp splicing junction region
(12) ncRNA_UTR3：the number of Somatic SNP in 3'UTR of non-coding RNA
(13) ncRNA_UTR5：the number of Somatic SNP in 5'UTR of non-coding RNA
(14) ncRNA_splicing：the number of Somatic SNP in 10bp splicing junction of non-coding RNA
(15) Total: the total number of Somatic SNP

We summarize mutation rate and mutation spectrum. Somatic mutation rate shows the frequency of every tumor sample's somatic cell mutation. Through mutation spectrum analysis, we can know the types of mutation (like C:G>T:A) in each tumor sample, which shows whether the sample has any type of mutation preference or not. By means of somatic mutation rate and mutation spectrum analysis, we can have an insight into the mechanism of somatic mutation. For example, maybe there is a positive correlation between mutation rate and age (the longer the time pass by, the more mutations accumulate). The mutation rate of the tumor cell that has the defect on DNA repair mechanisms should also be higher.

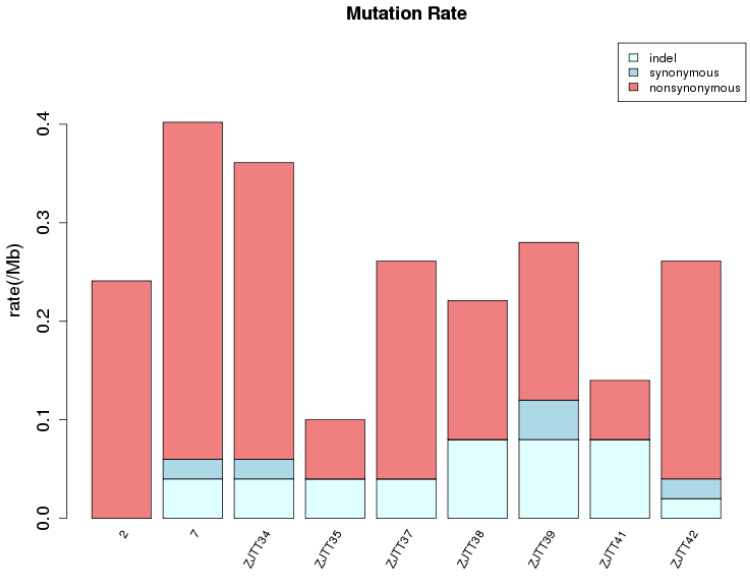


Figure 4.9 Somatic mutation rate figure

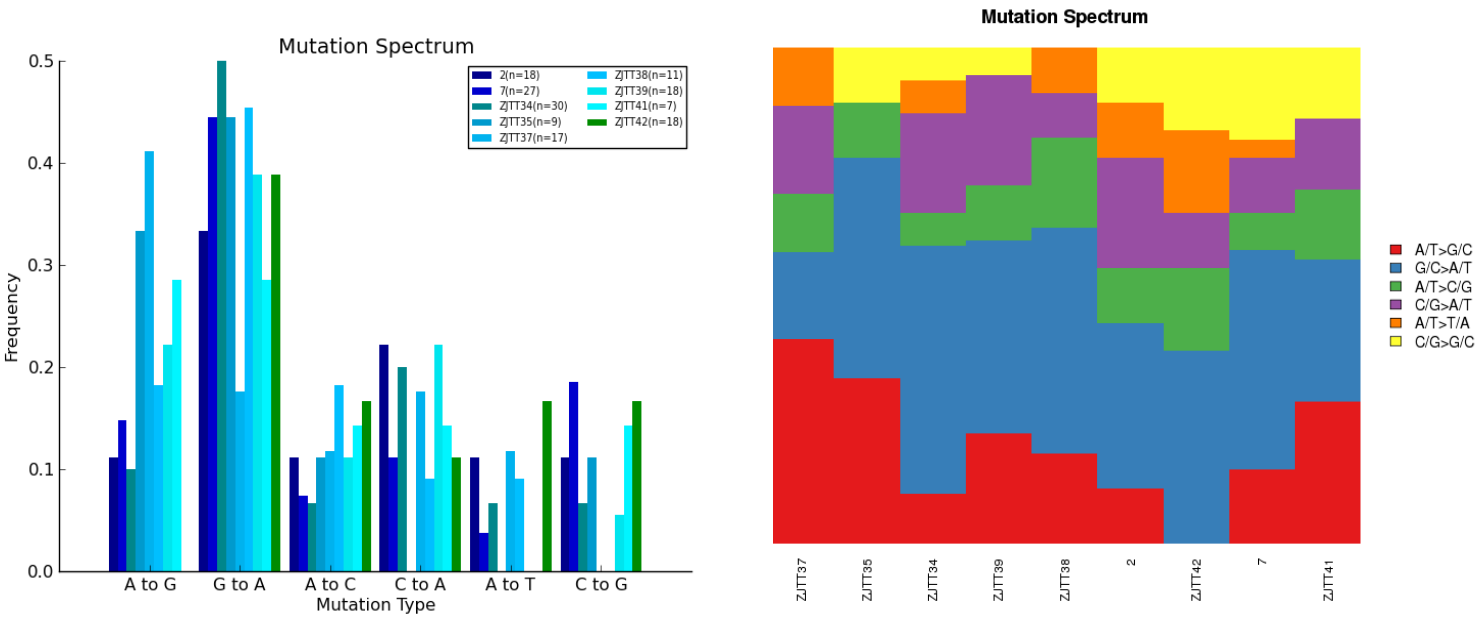


Figure 4.10 Mutation features. The left figure is frequency histogram of different type of mutation. The right figure is frequency spectrum of different type of mutation

4.5.2 Somatic INDEL detection

The detection result of Somatic INDEL is as follows:

Table 4-11 The number of Somatic INDEL in different genomic region

Sample[1]	T40	T25	T24	T22
CDS[2]	0	2	2	1
frameshift_deletion[a]	0	1	1	0
frameshift_insertion[b]	0	0	1	1
nonframeshift_deletion[c]	0	0	0	0
nonframeshift_insertion[d]	0	1	0	0
stopgain[e]	0	0	0	0
stoploss[f]	0	0	0	0
unknown[g]	0	0	0	0
Intronic[3]	1	29	32	30
UTR3[4]	0	0	3	1
UTR5[5]	0	0	0	0
Splicing[6]	0	0	0	0
ncRNA_exonic[7]	0	1	0	0
ncRNA_intronic[8]	0	2	3	2
ncRNA_UTR3[9]	0	0	0	0
ncRNA_UTR5[10]	0	0	0	0
ncRNA_splicing[11]	0	0	0	0
Upstream[12]	0	0	0	1
Downstream[13]	0	1	0	2
Intergenic[14]	27	86	88	85
Total[15]	28	121	128	122

Note：

(1) Sample：Sample name

(2) CDS：the number of somatic INDEL in exonic region

a) frameshift_deletion ：a deletion of one or more nucleotides that cause frameshift changes in protein coding sequence. the deletion length is not multiple of 3

b) frameshift_insertion ：an insertion of one or more nucleotides that cause frameshift changes in protein coding sequence. the insertion length is not multiple of 3.

c) nonframeshift_deletion ：non-frameshift deletion, does not change coding protein frame deletion, the deletion length is multiple of 3

d) nonframeshift_insertion ：non-frameshift insertion, does not change coding protein frame insertion: the insertion length is multiple of 3

e) stopgain ：frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate creation of stop codon at the variant site

f) stoploss ：frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the

immediate elimination of stop codon at the variant site

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

(3) intronic：the number of Somatic INDEL in intronic region

(4) UTR5：the number of Somatic INDEL in 5'UTR region

(5) UTR3：the number of Somatic INDEL in 3'UTR region

(6) splicing：the number of Somatic INDEL in 10bp splicing junction region

(7) ncRNA_exonic: the number of Somatic INDEL in non-coding RNA exonic region

(8) ncRNA_intronic: the number of Somatic INDEL in non-coding RNA intronic region

(9) ncRNA_UTR3: the number of Somatic INDEL in 3'UTR of non-coding RNA

(10) ncRNA_UTR5: the number of Somatic INDEL in 5'UTR of non-coding RNA

(11) ncRNA_splicing: the number of Somatic INDEL in 10bp splicing junction of non-coding RNA

(12) upstream：the number of Somatic INDEL in the 1kb upstream region of transcription start site

(13) downstream：the number of Somatic INDEL in the 1kb downstream region of transcription ending site

(14) intergenic：the number of Somatic INDEL in intergenic region

(15) Total：the total number of Somatic INDEL

4.5.3 Somatic SV detection

The detection result of Somatic SV is as follows:

Table 4-12 Statistics summary of somatic SV

Sample[1]	varType[2]	Total[3]	Cds[4]	Splicing[5]	utr5[6]	utr3[7]	Intron[8]	Upstream[9]	Downstream[10]	ncRNA[11]	Intergenic[12]	Unknown[13]
T24	Insertion	98	4	0	0	0	45	1	4	3	41	0
	Deletion	2315	44	0	2	7	794	17	24	89	1338	0
	Translocation	281	3	0	0	5	79	2	1	7	184	0
	Inversion	140	36	0	0	0	30	2	0	18	54	0
T25	Deletion	2223	44	0	1	6	758	8	24	84	1298	0
	Inversion	133	42	0	0	0	32	2	0	13	44	0
	Translocation	271	3	0	0	4	83	1	0	9	171	0
	Insertion	34	2	0	0	0	16	0	1	1	14	0

Note： Each column shows the number of each type of Somatic SV in the following region

(1) Sample: sample name

(2) varType：Somatic SV Type

(3) total：the total number of total Somatic SV

(4) CDS：the number of Somatic SV in CDS region

(5) splicing：the number of Somatic SV in splicing junction region
(6) UTR5：the number of Somatic SV in 5'UTR region
(7) UTR3：the number of Somatic SV in 3'UTR region
(8) intronic：the number of Somatic SV in intronic region
(9) intergenic：the number of Somatic SV in intergenic region
(10) upstream：the number of Somatic SV in 1Kb region upstream from transcription start site
(11) downstream：the number of Somatic SV in 1Kb region downstream from transcription end site
(12) ncRNA: the number of Somatic SV in ncRNA region
(13) unknown：the number of Somatic SV in region with unknown function (due to various errors in the gene structure definition in the database file)

4.5.3 Somatic CNV detection

The detection result of Somatic CNV is as follows:

Table 4-13 Statistics summary of somatic CNV

Sample[1]	varType[2]	Total[3]	Cds[4]	Splicing[5]	utr5[6]	utr3[7]	Intron[8]	Upstream[9]	Downstream[10]	ncRNA[11]	Intergenic[12]	Unknown[13]
T24	gain	84	37	0	3	0	6	0	0	7	31	0
	loss	46	40	0	0	0	0	0	0	5	1	0
T25	loss	12	10	0	0	0	1	0	0	1	0	0
	gain	53	16	0	0	0	11	2	1	4	19	0
T40	gain	103	47	0	0	1	13	0	1	6	35	0
	loss	10	2	0	0	0	1	0	0	2	5	0
T22	loss	17	14	0	1	0	0	0	0	2	0	0
	gain	62	42	0	0	0	2	0	0	4	14	0

Note：each column shows the number of different type of somatic CNV in each genomic region.

(1) Sample: sample name
(2) varType：Somatic CNV type
(3) total：the total number of Somatic CNV
(4) CDS：the number of Somatic CNV in CDS region
(5) splicing：the number of Somatic CNV in splicing junction region
(6) UTR5：the number of Somatic CNV in 5'UTR region
(7) UTR3：the number of Somatic CNV in 3'UTR region
(8) intronic：the number of Somatic CNV in intronic region
(9) upstream：the number of Somatic CNV in 1Kb region upstream from transcription start site
(10) downstream：the number of Somatic CNV in 1Kb region downstream from transcription end site

(11) ncRNA: the number of Somatic CNV in ncRNA region

(12) intergenic: the number of Somatic CNV in intergenic region

(13) unknown：the number of Somatic CNV in region with unknown function (due to various errors in the gene structure definition in the database file)

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)
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）

6 Appendix

The softwares which were applied in the bioinformatic analysis are listed as below:

Table 6-1 The list of genome-wide analysis software

Analytical content	Software	Comments	Version
Alignment	BWA	Map the sequencing reads to the reference genome and the BAM file was obtained	0.7.8-r455
	Samtools	Sort bam	1
	Picard	Merge the bam file from the same sample and mark the duplicate reads	1.111
	GATK	local realignment and base quality recalibration	v3.1
SNP/INDEL detection	GATK	Detect and filter SNP,INDEL	v3.1
SV detection	Breakdancer	Detect SV	1.4.4
CNV detection	Control-FREEC	Detect CNV	v6.7
Somatic SNP/INDEL detection	MuTect/Strelka	Detect and filter somatic SNP/INDEL	muTect：1.1.4，Strelka：v1.0.13
Somatic SV detection	Breakdancer	Detect somatic SV	1.4.4
Somatic CNV detection	Control-FREEC	Detect Somatic CNV	v6.7
Functional annotation	ANNOVAR	Annotate variation site	2013Aug23