# Human Whole Exome Sequencing (Cancer)

# Primary Analysis Demo Report

**May 1, 2016**

# Contents

# 1 Sample Information

**Table 1. 1 Sample information**

| PatientID | SampleID | LibraryID | Type |
|---|---|---|---|
| P001 | P001_T | DHE01682 | T |
| P001 | P001_N | DHE01681 | N |
| P002 | P002_N | DHE01684 | N |
| P002 | P002_T | DHE01683 | 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.

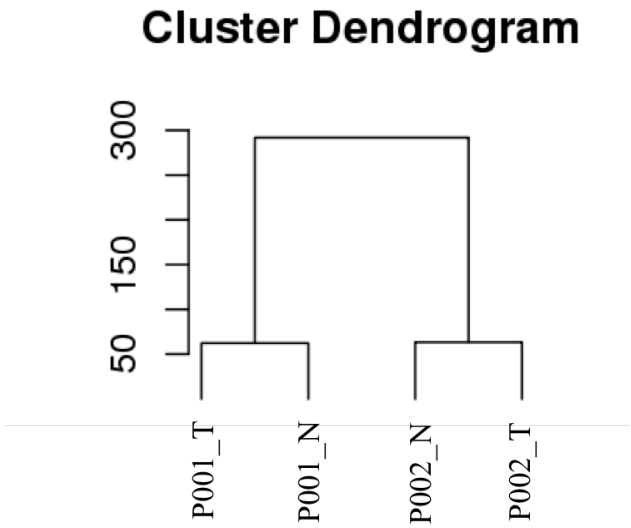



**Figure 1.1 Cluster analysis among the samples**

# 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 genomic DNA per sample was used as input material for the DNA sample preparation. Sequencing libraries were generated using Agilent SureSelect Human All Exon kit (Agilent Technologies, CA, USA) following manufacturer's recommendations and index codes were added to each sample. Briefly, fragmentation was carried out by hydrodynamic shearing system (Covaris, Massachusetts, USA) to generate 180-280bp fragments. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities and enzymes were removed. After adenylation of 3' ends of DNA fragments, adapter oligonucleotides were ligated. DNA fragments with ligated adapter molecules on both ends were selectively enriched in a PCR reaction. After PCR reaction, library hybridize with Liquid phase with biotin labeled probe, then use magnetic beads with streptomycin to capture the exons of genes. Captured libraries were enriched in a PCR reaction to add index tags to prepare for hybridization. Products were purified using AMPure XP system (Beckman Coulter, Beverly, USA) and quantified using the Agilent high sensitivity DNA assay on the Agilent Bioanalyzer 2100 system.

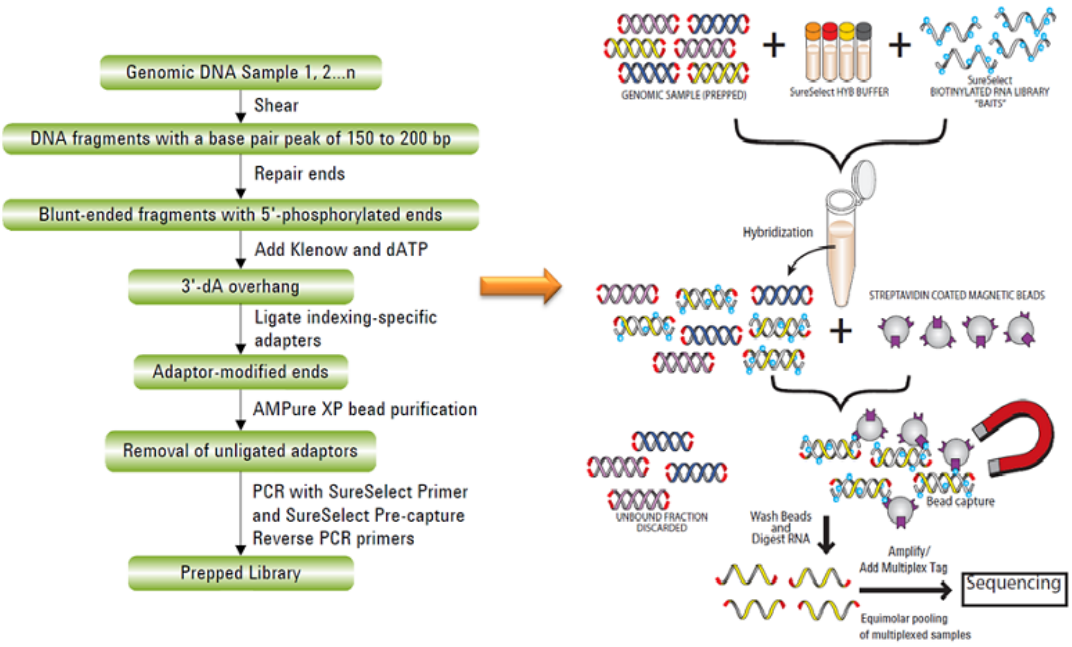



**Figure 2.1 Library construction workflow**

### 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.

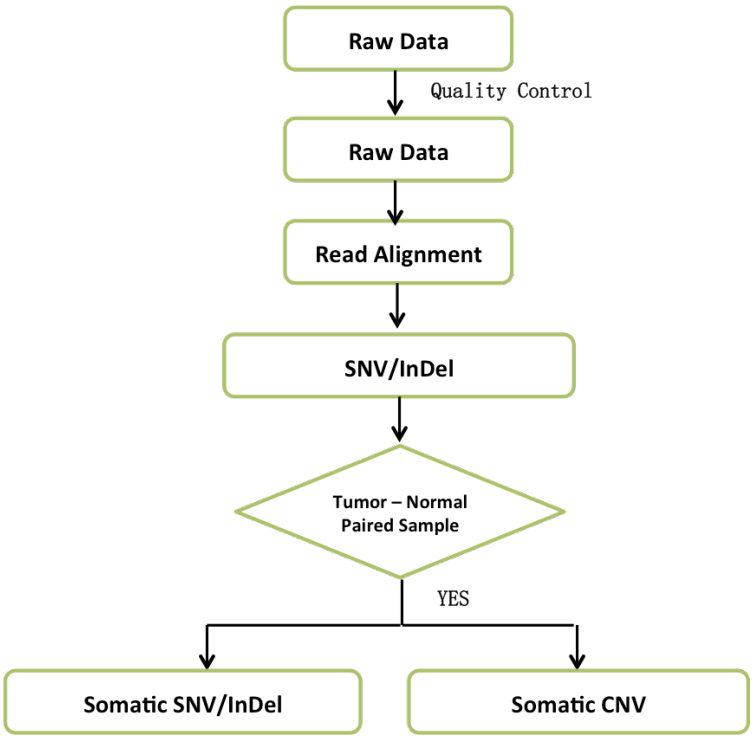



**Figure 3.1 Bioinformatics analysis pipeline**

# 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
+
@@CFFFDEHHHHHFIJJJ@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 contain the same number of characters as bases in line 2.

Illumina sequence identifier details:

**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’

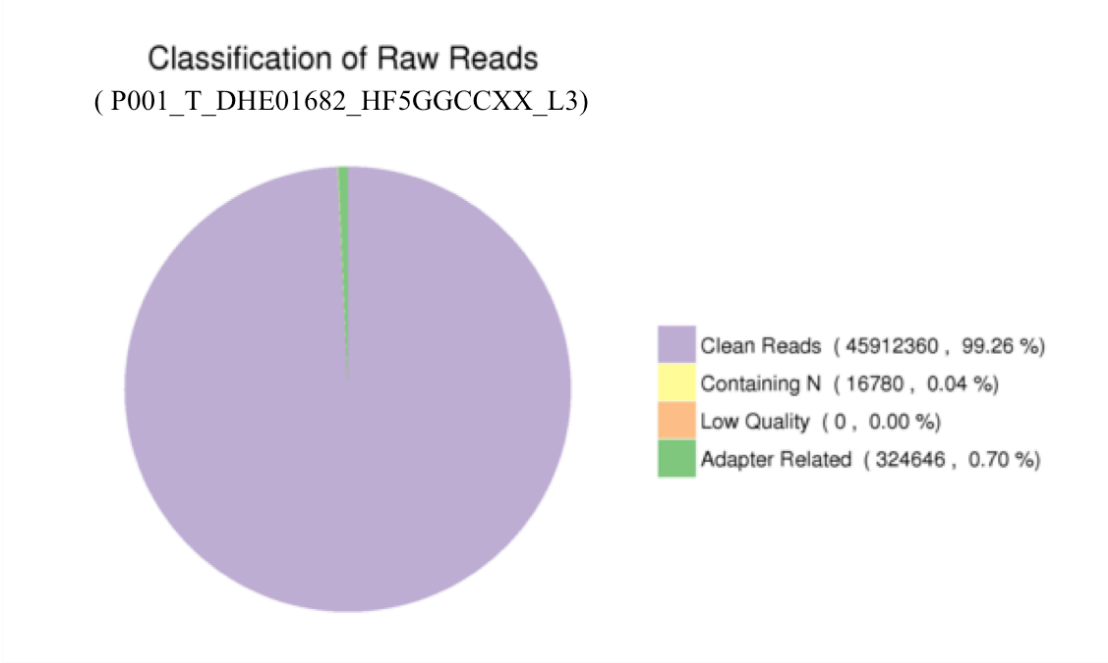



**Figure 4.1 Raw data filtration result**

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.

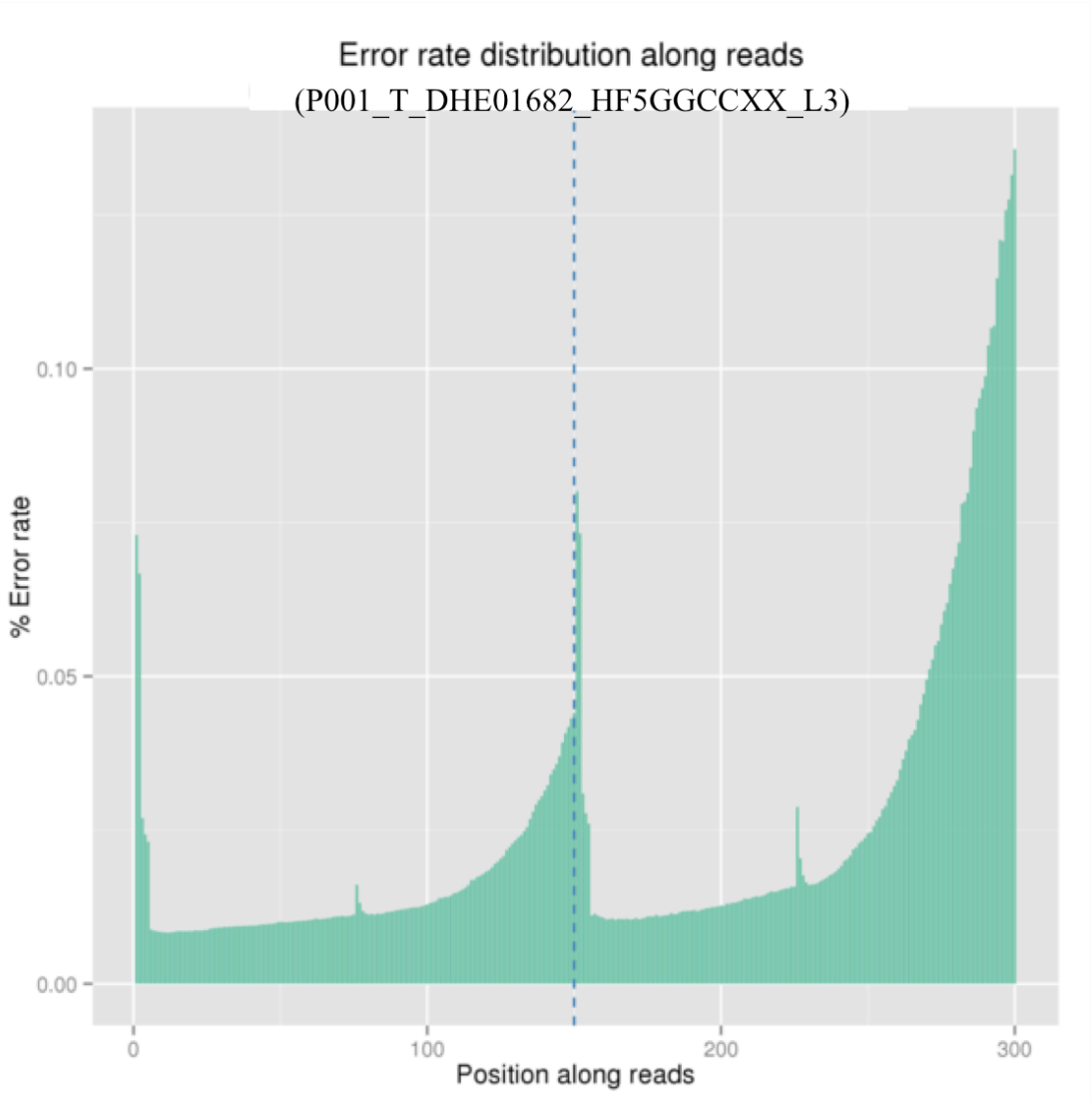



**Figure 4.2 Sequencing error rate distribution**

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.

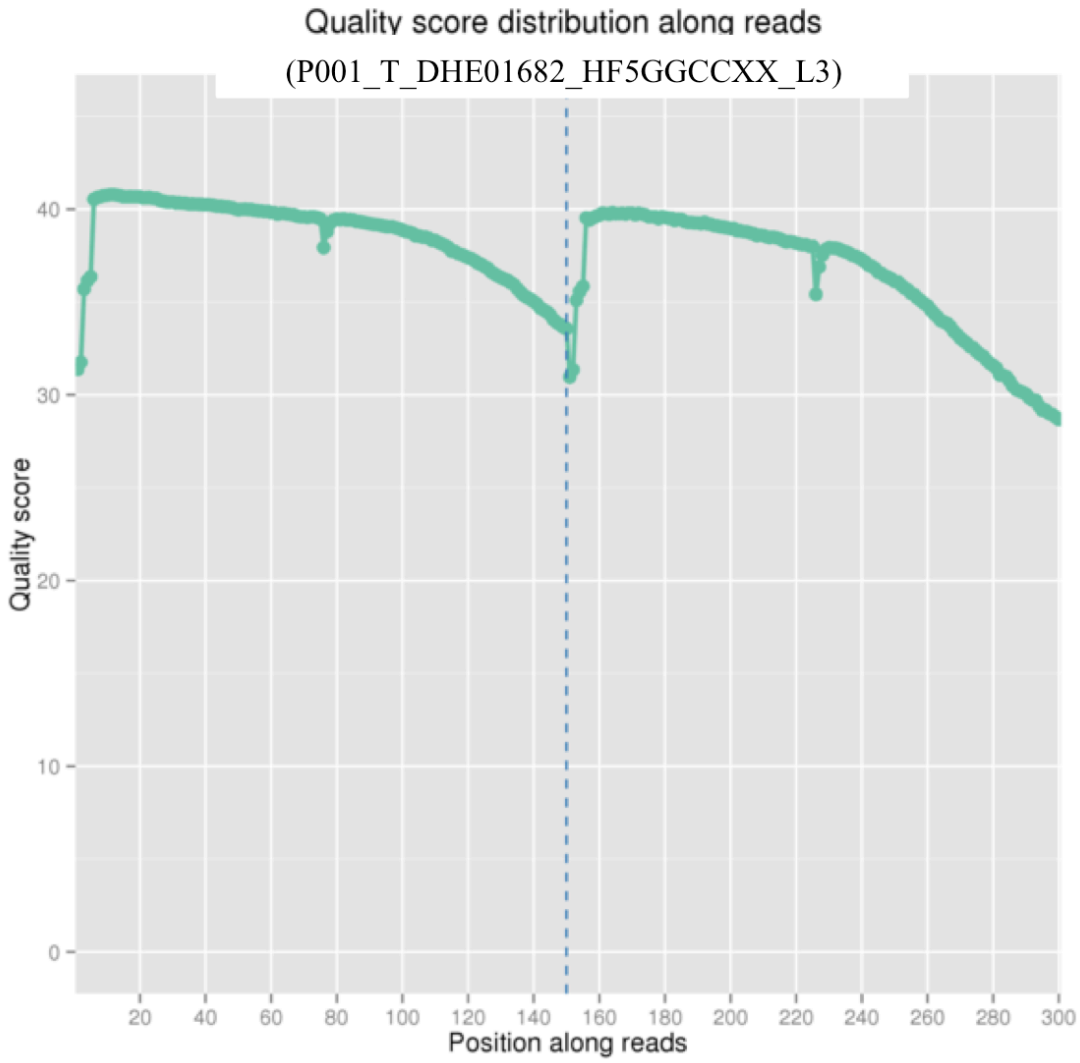



**Figure 4.3 Sequencing quality distribution**

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 (%) |
|---|---|---|---|---|---|---|---|---|---|
| P001_T | DHE01682 | HF5GGCCXX_L3 | 46253786 | 13.88 | 99.26 | 0.03 | 94.77 | 87.75 | 47.64 |
| P001_N | DHE01681 | HF5GGCCXX_L3 | 22787618 | 6.84 | 99.32 | 0.03 | 94.62 | 87.41 | 48.2 |
| P002_N | DHE01684 | HF5GGCCXX_L3 | 24416970 | 7.33 | 99.28 | 0.03 | 94.75 | 87.67 | 48.51 |
| P002_T | DHE01683 | HF5GGCCXX_L3 | 48315582 | 14.49 | 99.18 | 0.03 | 94.76 | 87.68 | 48.73 |

(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

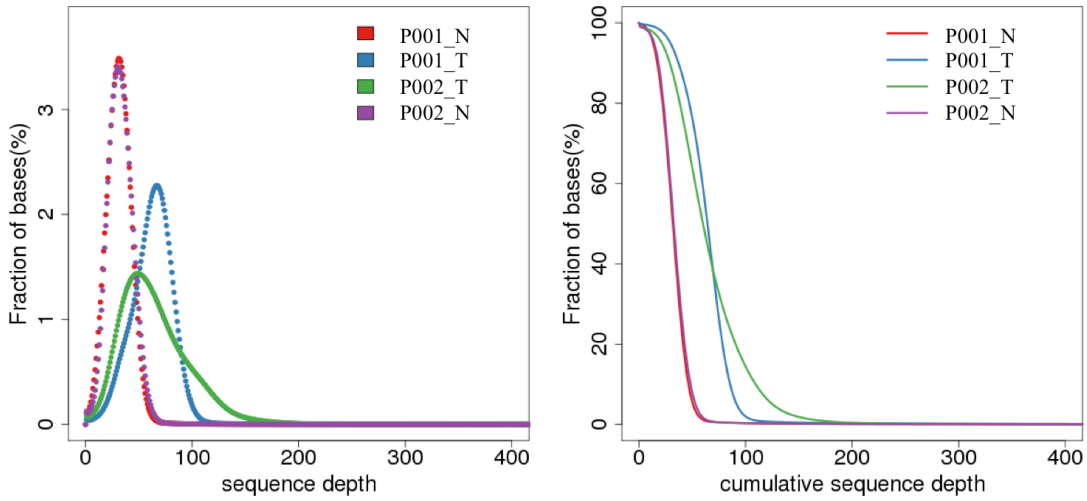
(9) Q30: the percent of bases with phred-scaled quality scores greater than 30

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

## 4.3 Sequence Alignment

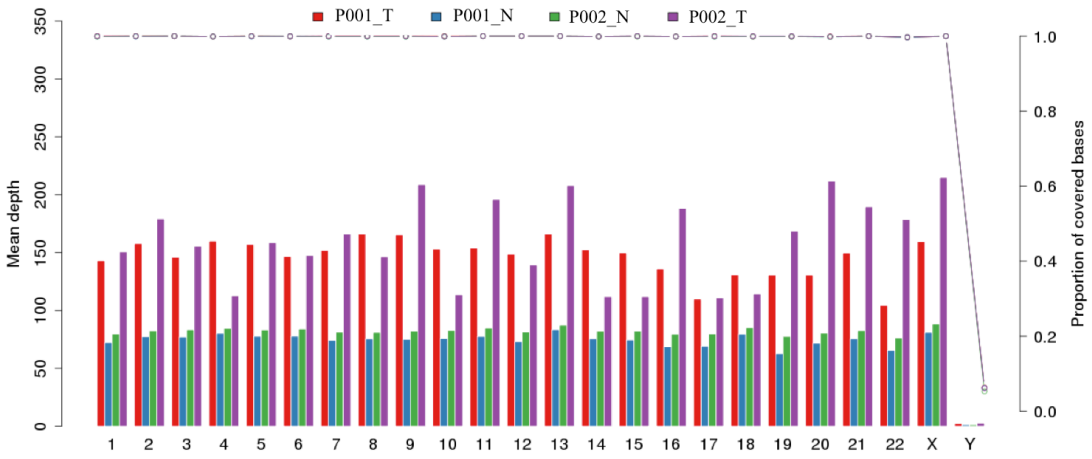
Burrows-Wheeler Aligner (BWA) was utilized to map the paired-end clean reads to the human reference genome (b37 + decoy). After sorting with samtools and marking duplicates with Picard, the results of read alignment was finally stored in the format of BAM. We then compute the coverage and depth based 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_T | P001_N | P002_N | P002_T |
|---|---|---|---|---|
| Total | 91824720 (100%) | 45264602 (100%) | 48480168 (100%) | 95838666 (100%) |
| Duplicate | 8893470 (9.70%) | 3437757 (7.61%) | 3701392 (7.65%) | 8122425 (8.49%) |
| Mapped | 91670368 (99.83%) | 45199549 (99.86%) | 48401584 (99.84%) | 95697216 (99.85%) |
| Properly mapped | 91269090 (99.39%) | 44991474 (99.40%) | 48194996 (99.41%) | 95250734 (99.39%) |
| PE mapped | 91614762 (99.77%) | 45172664 (99.80%) | 48373636 (99.78%) | 95643238 (99.80%) |
| SE mapped | 111212 (0.12%) | 53770 (0.12%) | 55896 (0.12%) | 107956 (0.11%) |
| With mate mapped to a different chr | 237210 (0.26%) | 127680 (0.28%) | 123960 (0.26%) | 272236 (0.28%) |
| With mate mapped to a different chr ((mapQ>=5)) | 161059 (0.18%) | 90610 (0.20%) | 84424 (0.17%) | 188391 (0.20%) |
| Initial_bases_on_target | 50390601 | 50390601 | 50390601 | 50390601 |
| Initial_bases_near_target | 73902222 | 73902222 | 73902222 | 73902222 |
| Initial_bases_on_or_near_target | 124292823 | 124292823 | 124292823 | 124292823 |
| Total_effective_reads | 91873630 | 45293251 | 48498634 | 95905511 |
| Total_effective_yield(Mb) | 12288.35 | 6199.99 | 6637.31 | 13001.76 |
| Effective_sequences_on_target(Mb) | 7509.46 | 3737.82 | 4120.08 | 8050.07 |
| Effective_sequences_near_target(Mb) | 3070.22 | 1536.77 | 1666.11 | 3211.86 |
| Effective_sequences_on_or_near_target(Mb) | 10579.69 | 5274.6 | 5786.19 | 11261.93 |
| Fraction_of_effective_bases_on_target | 0.611 | 0.603 | 0.621 | 0.619 |
| Fraction_of_effective_bases_on_or_near_target | 0.861 | 0.851 | 0.872 | 0.866 |
| Average_sequencing_depth_on_target | 149.03 | 74.18 | 81.76 | 159.75 |
| Average_sequencing_depth_near_target | 41.54 | 20.79 | 22.54 | 43.46 |
| Mismatch_rate_in_target_region | 0.0071 | 0.0071 | 0.0069 | 0.007 |
| Mismatch_rate_in_all_effective_sequence | 0.0059 | 0.0059 | 0.0057 | 0.0058 |
| Base_covered_on_target | 50282747 | 50264854 | 50273990 | 50287426 |
| Coverage_of_target_region | 0.998 | 0.998 | 0.998 | 0.998 |
| Base_covered_near_target | 73196952 | 71920339 | 72013230 | 73031550 |
| Coverage_of_flanking_region | 0.99 | 0.973 | 0.974 | 0.988 |
| Fraction_of_target_covered_with_at_least_20x | 0.988 | 0.954 | 0.965 | 0.988 |

| | | | | |
|---|---|---|---|---|
| Fraction_of_target_covered_with_at_least_10x | 0.995 | 0.988 | 0.99 | 0.995 |
| Fraction_of_target_covered_with_at_least_4x | 0.997 | 0.995 | 0.996 | 0.997 |
| Fraction_of_flanking_region_covered_with_at_least_20x | 0.596 | 0.379 | 0.406 | 0.6 |
| Fraction_of_flanking_region_covered_with_at_least_10x | 0.786 | 0.609 | 0.628 | 0.784 |
| Fraction_of_flanking_region_covered_with_at_least_4x | 0.937 | 0.843 | 0.852 | 0.93 |

(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) Initial_bases_on_target: the number of bases in the target region

(11) Initial_bases_near_target: the number of bases in the flanking (200bp upstream and downstream) region of the target

(12) Initial_bases_on_or_near_target: the number of bases in the target region or flanking region of the target

(13) Total_effective_reads: the number of effective reads mapped to the reference genome

(14) Total_effective_yield (Mb) : the data size of the effective reads mapped to the reference genome (MB as a unit)

(15) Effective_sequences_on_target (Mb) : the data size of the reads mapped to the target region

(16) Effective_sequences_near_target (Mb) : the data size of the reads mapped to the flanking region of the target

(17) Effective_sequences_on_or_near_target (Mb) : the data size of the reads total reads that mapped to mapped to the target region or flanking region

(18) Fraction_of_effective_bases_on_target: the percentage of bases mapped to the target region in all bases mapped to the reference genome (Effective_sequences_on_target/Total_effective_yield)

(19) Fraction_of_effective_bases_on_or_near_target: the percentage of bases mapped to the target or flanking region in all bases mapped to the reference genome

(20) Average_sequencing_depth_on_target: the average sequencing depth in the target region (Effective_sequences_on_target * 1 million/Initial_bases_on_target)

(21) Average_sequencing_depth_near_target: the average sequencing depth in the flanking region of the target

(22) Mismatch_rate_in_target_region: the percentage of mismatches in the target region

(23) Mismatch_rate_in_all_effective_sequence: the percentage of mismatches in the reference genome

(24) Base_covered_on_target: the number of the bases covered in the target region

(25) Coverage_of_target_region: the percentage of target region (Base_covered_on_target/Initial_bases_on_target)

(26) Base_covered_near_target: the number of the bases covered in the flanking region

(27) Coverage_of_flanking_region: the number of bases coverage in the flanking region

(28) Fraction_of_target_covered_with_at_least_20x: the percentage of bases with depth > 20X in the target region

(29) Fraction_of_target_covered_with_at_least_10x: the percentage of bases with depth > 10X in the target region

(30) Fraction_of_target_covered_with_at_least_4x: the percentage of bases with depth > 4X in the target region

(31) Fraction_of_flanking_region_covered_with_at_least_20x: the percentage of bases with depth > 20X in the flanking region

(32) Fraction_of_flanking_region_covered_with_at_least_10x: the percentage of bases with depth > 10X in the flanking region

(33) Fraction_of_flanking_region_covered_with_at_least_4x: the percentage of bases with depth > 4X in the flanking region

## 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 | P001_N | P002_N | P002_T |
|---|---|---|---|---|
| CDS | 23552 | 23527 | 23492 | 22935 |
| synonymous_SNP | 11769 | 11754 | 11750 | 11365 |
| missense_SNP | 11179 | 11173 | 11161 | 11000 |
| stopgain | 102 | 103 | 82 | 87 |
| stoploss | 15 | 15 | 9 | 10 |
| unknown | 508 | 503 | 512 | 496 |
| intronic | 151736 | 104825 | 102717 | 149710 |
| UTR3 | 6144 | 4978 | 5006 | 6230 |
| UTR5 | 3168 | 2760 | 2810 | 3193 |
| splicing | 535 | 532 | 552 | 541 |
| ncRNA_exonic | 3820 | 3292 | 3170 | 3805 |
| ncRNA_intronic | 16261 | 9599 | 9281 | 15982 |
| ncRNA_UTR3 | 0 | 0 | 0 | 0 |
| ncRNA_UTR5 | 0 | 0 | 0 | 0 |
| ncRNA_splicing | 42 | 41 | 37 | 36 |

| | | | | |
|---|---|---|---|---|
| upstream | 5135 | 3605 | 3533 | 5345 |
| downstream | 2880 | 1768 | 1685 | 3058 |
| intergenic | 121679 | 61455 | 55663 | 123419 |
| Total | 335274 | 216599 | 208156 | 334552 |

(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 termination site

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

(20) Total: the total number of SNPs

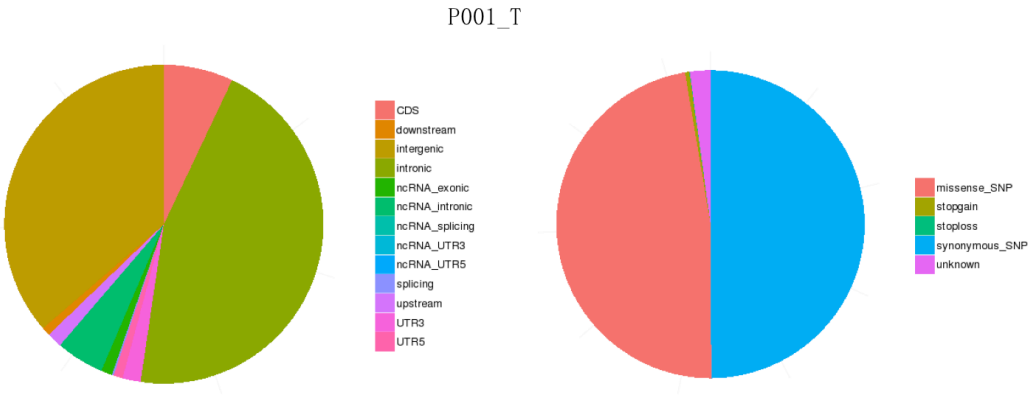



**Figure 4.6 Number of SNPs in various genomic regions (left pie plot); number of different types of SNPs in coding region (right pie plot)**

**Table 4.6 Feature of SNPs**

| Sample | P001_T | P001_N | P002_N | P002_T |
|---|---|---|---|---|
| Total | 335274 | 216599 | 208156 | 334552 |
| Het | 119168 | 87582 | 87740 | 114144 |
| Hom | 216106 | 129017 | 120416 | 220408 |
| transition | 230367 | 148808 | 142979 | 229664 |
| transvertion | 104907 | 67791 | 65177 | 104888 |
| ts/tv | 2.2 | 2.2 | 2.19 | 2.19 |
| dbSNP | 312415 | 200440 | 192427 | 309648 |
| percentage | (93.18%) | (92.54%) | (92.44%) | (92.56%) |
| novel | 22859 | 16159 | 15729 | 24904 |
| novel ts | 14591 | 10275 | 10044 | 15805 |
| novel tv | 8268 | 5884 | 5685 | 9099 |
| novel ts/tv | 1.76 | 1.75 | 1.77 | 1.74 |

(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 divided by the number of transversions

(8) dbSNP percentage: the number of SNPs that have been reported in dbSNP database divided 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 divided 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 | P001_N | P002_N | P002_T |
|---|---|---|---|---|
| CDS | 769 | 760 | 768 | 755 |
| frameshift_deletion | 137 | 133 | 141 | 140 |
| frameshift_insertion | 126 | 132 | 114 | 115 |
| nonframeshift_deletion | 222 | 216 | 228 | 221 |
| nonframeshift_insertion | 179 | 175 | 185 | 177 |
| stopgain | 7 | 9 | 10 | 11 |
| stoploss | 1 | 1 | 1 | 1 |
| unknown | 105 | 102 | 99 | 101 |
| intronic | 24121 | 16275 | 16172 | 23374 |
| UTR3 | 1028 | 794 | 730 | 995 |
| UTR5 | 535 | 455 | 457 | 552 |
| splicing | 212 | 182 | 188 | 199 |
| ncRNA_exonic | 444 | 365 | 333 | 419 |
| ncRNA_intronic | 2553 | 1503 | 1455 | 2398 |
| ncRNA_UTR3 | 0 | 0 | 0 | 0 |
| ncRNA_UTR5 | 0 | 0 | 0 | 0 |
| ncRNA_splicing | 7 | 8 | 10 | 8 |
| upstream | 862 | 588 | 597 | 858 |
| downstream | 409 | 263 | 247 | 430 |
| intergenic | 18004 | 9271 | 8558 | 17715 |
| Total | 48990 | 30496 | 29547 | 47 |

(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

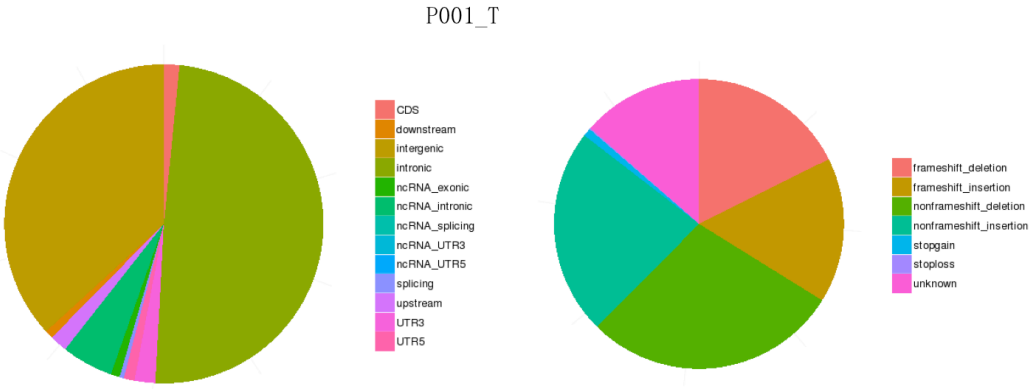



**Figure 4.7 Number of InDels in various genomic regions (left); number of different types of InDels in coding region (right)**

**Table 4.8 Feature of InDels in genome**

| Sample | P001_T | P001_N | P002_N | P002_T |
|---|---|---|---|---|
| Total | 48990 | 30496 | 29547 | 47757 |
| Het | 15849 | 10276 | 10432 | 14533 |
| Hom | 33141 | 20220 | 19115 | 33224 |
| dbSNP | 42795 | 26845 | 25885 | 41434 |
| percentage | (87.35%) | (88.03%) | (87.61%) | (86.76%) |
| novel | 6195 | 3651 | 3662 | 6323 |

(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 Variant Annotation

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 4.9 Annotation result of detected variants (only a part is shown here)**

| CHROM | POS | ID | REF | ALT | QAUL | FILTER | GeneName | Func | Gene | … |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 15211 | rs11586607 | T | G | 66.28 | PASS | WASH7P | ncRNA_intronic | NR_024540 | … |
| 1 | 15274 | rs62636497 | A | T | 66.28 | PASS | WASH7P | ncRNA_intronic | NR_024540 | … |

***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 4.10 Number of somatic SNPs in different genomic regions**

| Sample | P001_T | P002_T |
|---|---|---|
| CDS | 147 | 219 |
| synonymous_SNP | 43 | 60 |
| missense_SNP | 93 | 145 |
| stopgain | 8 | 10 |
| stoploss | 0 | 0 |
| unknown | 3 | 4 |
| intronic | 166 | 226 |
| UTR3 | 6 | 16 |
| UTR5 | 12 | 15 |
| splicing | 8 | 2 |
| ncRNA_exonic | 13 | 12 |
| ncRNA_intronic | 15 | 40 |
| ncRNA_UTR3 | 0 | 0 |
| ncRNA_UTR5 | 0 | 0 |
| ncRNA_splicing | 1 | 1 |
| upstream | 6 | 6 |
| downstream | 3 | 4 |
| intergenic | 81 | 93 |

| | | |
|---|---|---|
| Total | 458 | 635 |

(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 4.11 Number of somatic InDels in different genomic regions**

| Sample | P001_T | P002_T |
|---|---|---|
| CDS | 2 | 5 |
| frameshift_deletion | 0 | 1 |
| frameshift_insertion | 1 | 4 |
| nonframeshift_deletion | 1 | 0 |
| nonframeshift_insertion | 0 | 0 |
| stopgain | 0 | 0 |
| stoploss | 0 | 0 |
| unknown | 0 | 0 |
| intronic | 7 | 8 |
| UTR3 | 0 | 0 |
| UTR5 | 0 | 0 |
| splicing | 0 | 0 |
| ncRNA_exonic | 0 | 0 |
| ncRNA_intronic | 0 | 0 |
| ncRNA_UTR3 | 0 | 0 |

| | | |
|---|---|---|
| ncRNA_UTR5 | 0 | 0 |
| ncRNA_splicing | 0 | 0 |
| upstream | 0 | 1 |
| downstream | 0 | 0 |
| intergenic | 0 | 1 |
| Total | 9 | 15 |

(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 CNV Detection Result

Statistics of detected somatic CNVs are listed below:

**Table 4.12 Statistics of detected somatic CNVs**

| Sample | gain_count | gain_size | loss_count | loss_size | total_count | total_size |
|---|---|---|---|---|---|---|
| P002_T | 198 | 888508381 | 198 | 888508381 | 396 | 1777016762 |
| P001_T | 57 | 164292707 | 57 | 164292707 | 114 | 328585414 |

(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

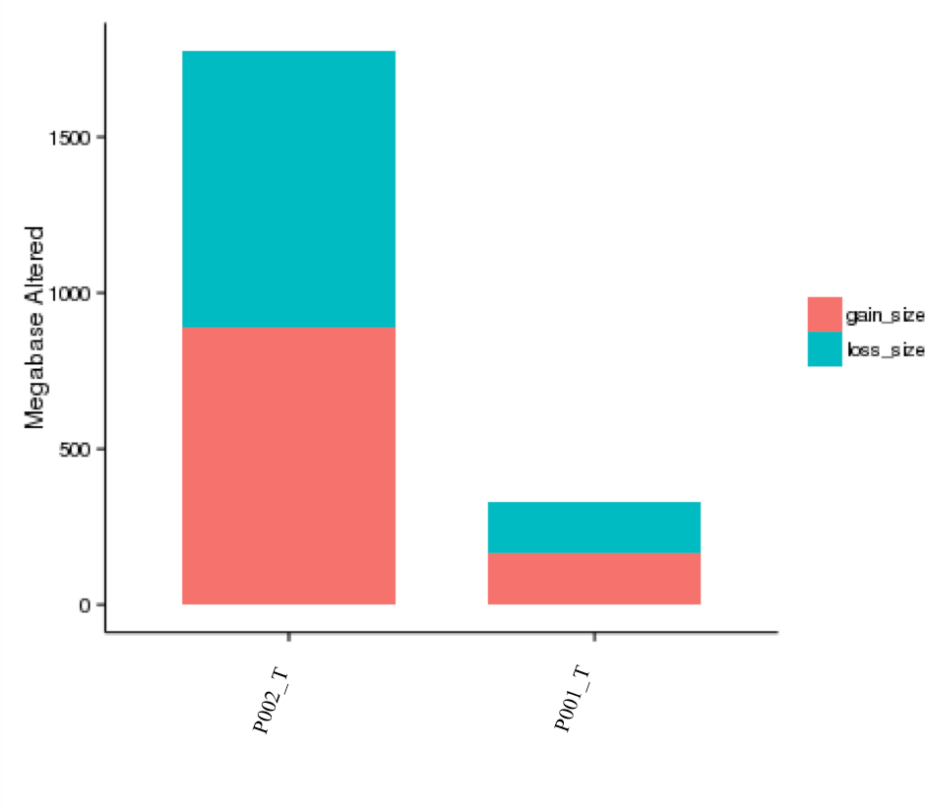



**Figure 4.8 The size of genomic regions affected by somatic CNVs in each sample**

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] 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 4.13 Softwares used in the analysis**

| Analysis | Software | Comment | 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 |
| Somatic SNP/InDel | MuTect/Strelka | Detect and filter Somatic SNPs/InDels | muTect:1.1.4; Strelka:v1.0.13 |
| Somatic CNV | Control-FREEC | Detect somatic CNVs | v6.7 |
| Annotation | ANNOVAR | Annotate variants | 2015Mar22 |