Human Whole Exome Sequencing

Project Demo Report

(Disease)

May 1, 2016

Contents

1 Sample Information 1
2 Experimental Procedures 1
2.1 DNA Quantification and Qualification 1
2.2 Library Preparation for Sequencing 1
2.3 Clustering and Sequencing 2
3 Bioinformatics Analysis Procedures 2
4 Analysis Result 2
4.1 Raw Data 2
4.2 Quality Control 3
4.2.1 Sequencing Data Filtration 3
4.2.2 Sequencing Error Rate Distribution 4
4.2.3 Sequencing Quality Distribution 5
4.2.4 Statistics Summary of Sequencing Quality 6
4.3 Sequence Alignment 6
4.3.1 Sequencing Depth, Coverage Distribution 7
4.3.2 Statistics of Coverage 8
4.4 Variation Detection Result 9
4.4.1 SNV Result 9
4.4.2 InDel Result 11
4.4.3 Annotation Result 14
5 Advanced analysis 17
5.1 Variant filtering using known databases 17
5.2 Variant filtering based on disease model 18
5.3 Linkage analysis 18
5.4 Regions of homozygosity (ROH) analysis 19
5.5 *De novo* mutation analysis 20
5.5.1 *De novo* mutation from SAMtools 20
5.5.2 *De novo* mutation from DenovoF 20
5.5.3 Annotation Result 21
References 22
Appendix 23
Appendix A: Software List 23
Appendix B: Verification Method of Sequencing 23
Verification Method of SNV/InDel 23

1 Sample Information

Table 1.1 The sample information

FamilyID	SampleID	SEX	Normal/Patient
F1	test1	F or M	N or P

2 Experimental Procedures

2.1 DNA Quantification and Qualification

Three methods are applied to DNA quantification and qualification: (1) DNA purity was checked using the Nanodrop (OD260/280 ratio); (2) DNA degradation and contamination were monitored on 1% agarose gels; (3) DNA concentration was measured using Qubit. DNA samples with OD260/280 ratio between 1.8~2.0 and concentration above 1.0ug are used to prepare sequencing libraries.

2.2 Library Preparation for Sequencing

Agilent liquid phase hybridization was applied to efficiently enrich whole exons which will be sequenced on Illumina platform. Sequencing libraries and capture were used Agilent SureSelect Human All ExonV6 (Agilent Technologies, CA, USA) with reagents recommended by the instruction manual and following optimized experimental procedures.

Basic experimental procedures: Genomic DNA was randomly fragmented to 180-280bp with Covaris cracker, then DNA fragments were end polished, A-tailed and ligated with the full-length adapter for Illumina sequencing. Fragments with specific indexes were hybridized with biotin labeled probes after pooling, then magnetic beads with streptomycin were used to capture exons. After PCR amplification and quality control, libraries were sequenced (Figure 1).

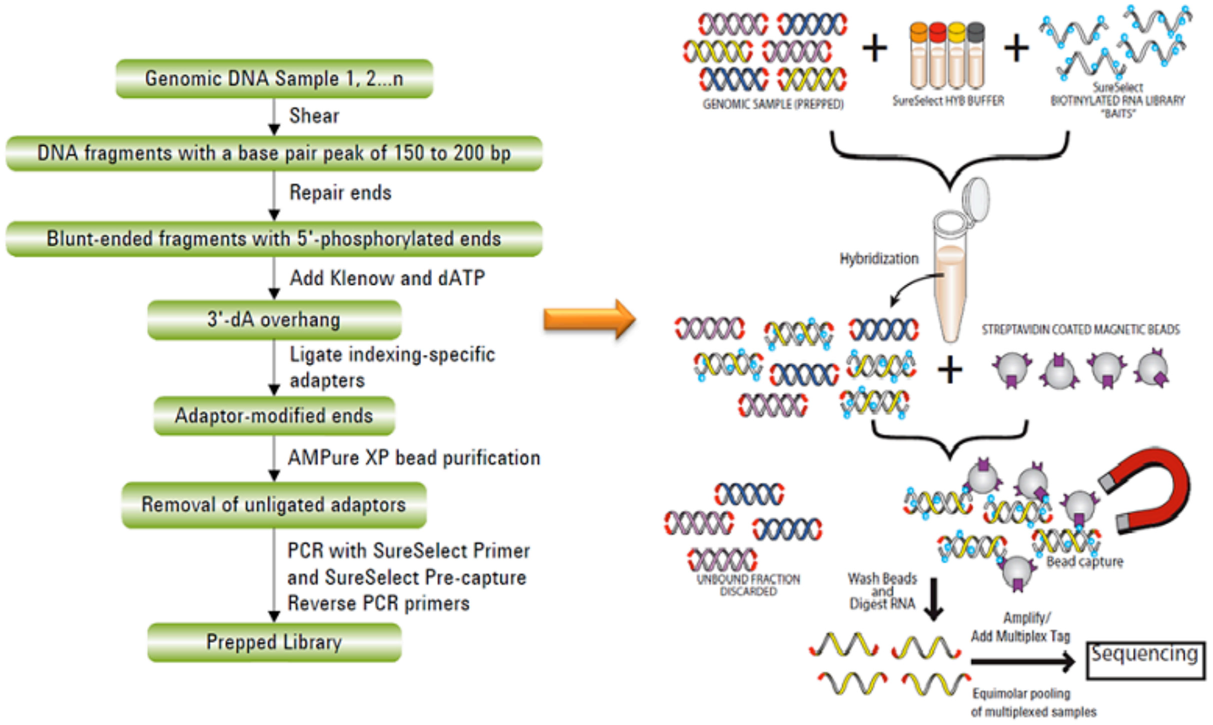
Genomic DNA Sample 1, 2...n
Shear
DNA fragments with a base pair peak of 150 to 200 bp
Repair ends
Blunt-ended fragments with 5'-phosphorylated ends
Add Klenow and dATP
3'-dA overhang
Ligate indexing-specific adapters
Adaptor-modified ends
AMPure XP bead purification
Removal of unligated adaptors
PCR with SureSelect Primer and SureSelect Pre-capture Reverse PCR primers
Prepped Library
GENOMIC SAMPLE (PREPPED)
SureSelect HYB BUFFER
SureSelect BIOTINYLATED RNA LIBRARY "BAITS"
Hybridization
STREPTAVIDIN COATED MAGNETIC BEADS
UNBOUND FRACTION DISCARDED
Bead capture
Wash Beads and Digest RNA
Amplify/ Add Multiplex Tag
Sequencing
Equimolar pooling of multiplexed samples

Figure 2.1 The workflow of library preparation.

2.3 Clustering and Sequencing

If the library qualifies, it will be sequenced on an Illumina platform according to effective concentration and data volume.

3 Bioinformatics Analysis Procedures

By default, the 1000 Genomes (GRCh37 + decoy) human genome reference is used as reference genome. The bioinformatics analysis workflow is as follows

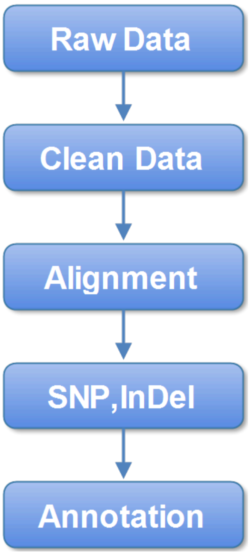

Figure 3.1 Bioinformatics analysis pipeline

4 Analysis Result

4.1 Raw Data

The original raw image data obtained from high throughput sequencing platforms (e.g. Illumina platform) is transformed to sequenced reads by base calling. The sequenced reads are regarded as raw data or raw reads, which is recorded in FASTQ file (fq) containing 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 beginning with a '@' character 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.

Table 4.1 Illumina sequence identifier details

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 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 platform is expressed as $Q_{phred,}$ the equation 1 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 (Q_{phred}) is listed as below (Table 4.2):

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

The higher the quality value, the lower the error rate and the higher the accuracy.

4.2 Quality Control

4.2.1 Sequencing Data Filtration

Raw sequence data contains adapter contamination and low-quality reads. To ensure the quality of bioinformatics analyses, raw data should be filtered to obtain clean reads which will be used in the downstream analyses.

The steps of data processing are as follows:

(1) Discard the read pair if either one read contains adapter contamination.

(2) Discard the read pair if more than 10% of nucleotides are uncertain in either one read.

(3) Discard the read pair if the proportion of low quality nucleotides is over 50% in either one read.

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

5' Adapter:

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

3' Adapter:

5'-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC(6-indexes)ATCTCGTATGCCGTCTTCTGCTTG-3'

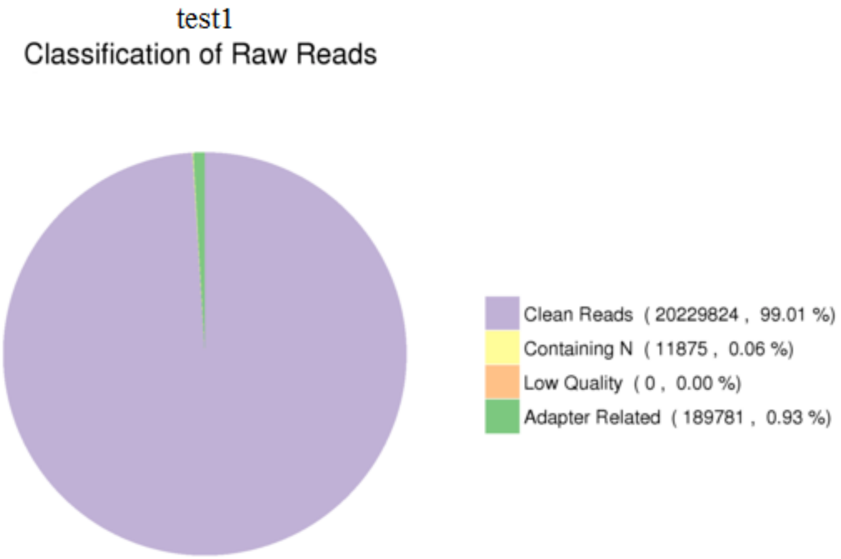

Figure 4.1 The filtration result of raw data

Note:
(1) Containing N: the number of read pairs with either one read containing uncertain nucleotides more than 10%, and the proportion in raw data.
(2) Low Quality: the number of read pairs with either one read containing low quality (below 5) nucleotides more than 50 percent, and the proportion in raw data.
(3) Adapter related: the number of read pairs filtered out with adapter contamination, and the proportion of filtered read pairs in raw data.
(4) Clean reads: the number of read pairs passed quality control and the proportion in raw data.
Note: Reads were discarded in pairs.

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:

Table 4.3 Phred score and corresponding sequencing error rate

Phred score	Sequencing error rate	Sequencing correct rate	Q-score
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 Illumina platforms, sequencing error rate distribution has two features:

(1) Error rate is increasing with sequencing reads extension due to the attenuation of fluorescent signal caused by the incomplete excision of fluorescent mark.

(2) The first several bases have higher sequencing error rate than others. At the beginning of sequencing, the focusing of the sequencer's fluorescence image sensorsensing element is not sensitive enough, thus, the quality of acquired fluorescence image is low.

Sequencing error rate distribution is applied to detect whether there are any abnormal bases with high error rate in reads. For example, abnormal bases might present if the middle base sequencing error rate is higher than others. Generally, the sequencing error rate should be smaller than 1% at each site.

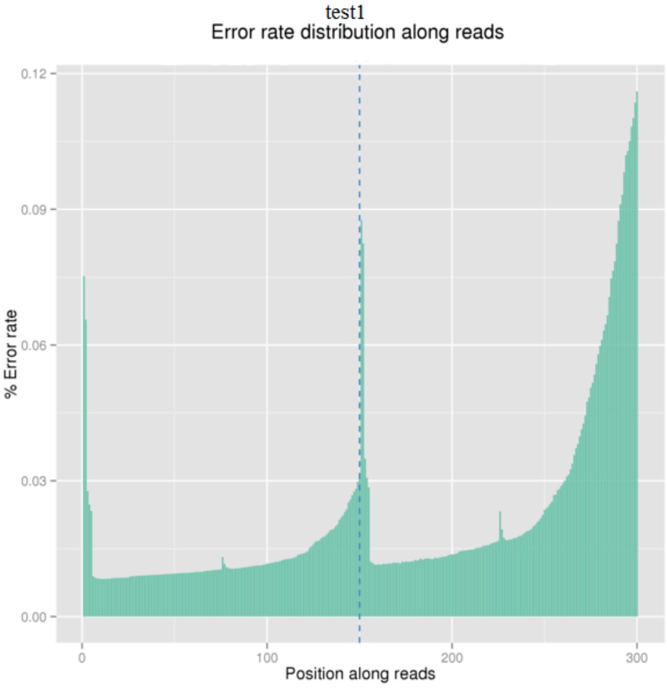

Figure 4.2 The distribution of sequencing error rate

The x-axis is the base position on reads and the y-axis is the average error rate of bases on all reads at this position.

4.2.3 Sequencing Quality Distribution

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

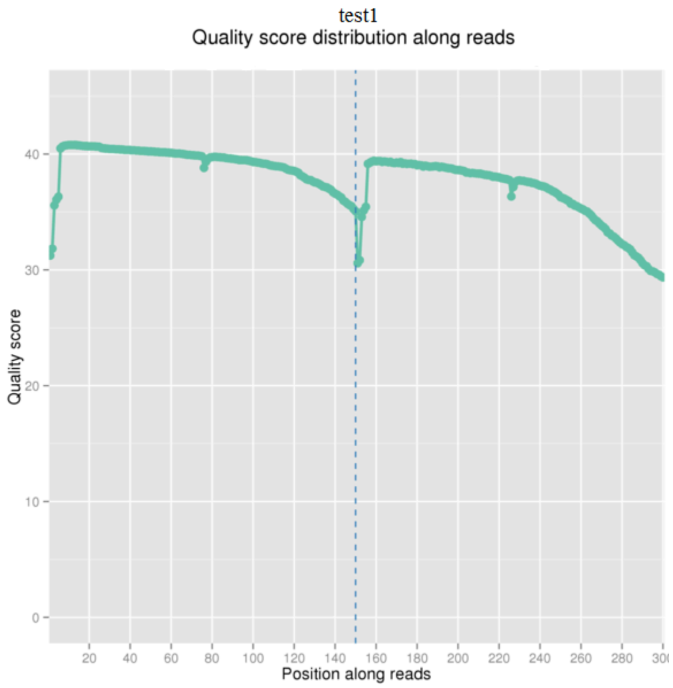

Figure 4.3 The distribution of sequencing quality along reads

The x-axis is the base position on reads and the y-axis is the average quality score (Phred Score) of bases on all reads at this position.

4.2.4 Statistics Summary of Sequencing Quality

According to the Illumina platform sequencing features, for PE data we require the average percentage of Q30 is above 80% and the average error rate is below 0.1%.

Table 4.4 The overview of sequencing quality

Sample name	1st BASE ID	Lane	Raw reads	Raw data (G)	Raw Depth (x)	Effective (%)	Error (%)	Q20 (%)	Q30 (%)	GC (%)
test1	D16000003	C4PU4ANXX_L4	21528191	5.38	53.9	95.59	0.035	95.21	91.04	47.78

Note:
(1) Sample name: Sample name.
(2) 1st BASE ID: 1st BASE ID of the sample.
(3) Lane: The flowcell ID and lane number during the sequencing (FlowcellID_LaneNumber).
(4) Raw reads: The number of sequencing reads pairs; four lines will be considered as one unit according to FASTQ format.
(5) Raw data (G): The original sequence data volume.
(6) Raw depth (x): The original sequence depth.
(7) Effective (%): The percentage of clean reads in all raw reads.
(8) Error (%): The average error rate of all bases.
(9) Q20: The percentage of bases with Phred score ≥20.
(10) Q30: The percentage of bases with Phred score ≥30.
(11) GC: The percentage of G and C in the total 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 (the 1000 Genomes human genome reference is used by default).

The original mapping result in BAM format can be obtained. Then, SAMtools(Li H *et al.*) is used to sort the BAM file, and Picard is used to mark duplicate reads. 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.

4.3.1 Sequencing Depth, Coverage Distribution

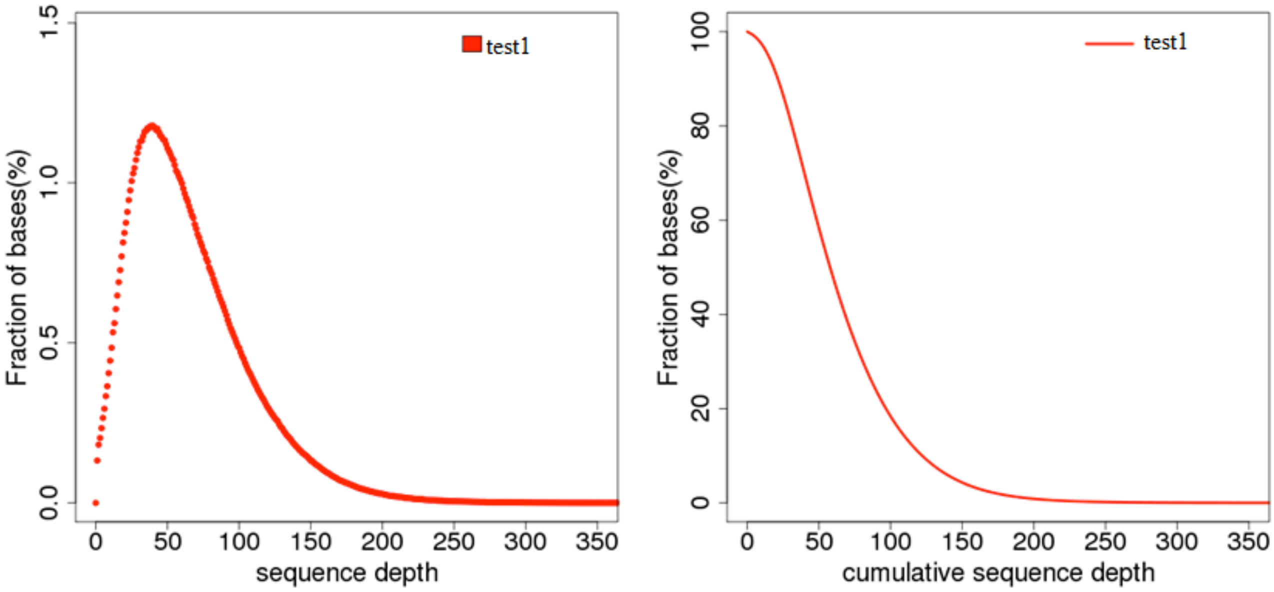

Figure 4.4 The distribution of sequencing depth.

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 curve follows a Poisson distribution around the average read depth. The right figure is the accumulative base ratio with different depth. The x-axis is sequencing depth, and 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 in 100%, shows that all bases are at least sequenced once.

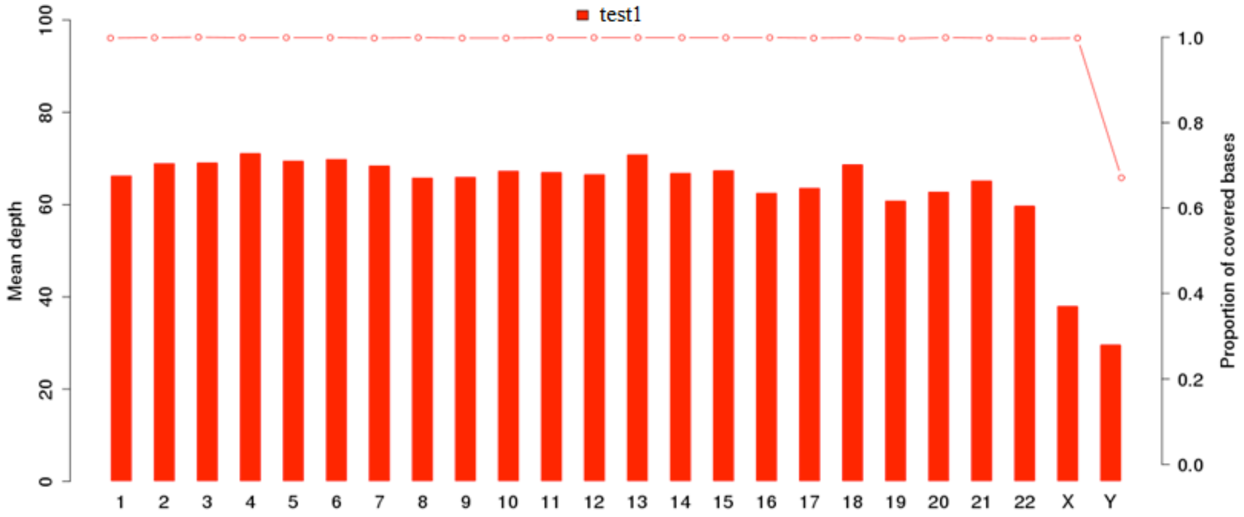

Figure 4.5 The average depth (the left y-axis) and the coverage rate (the right y-axis) of chromosomes. The x-axis is chromosome number; the left y-axis is the average depth of each chromosome (Raw data/length_of_chromosome); the right y-axis is the fraction of covered bases on each chromosome (The number of bases covered/total number of bases).

4.3.2 Statistics of Coverage

Table 4.5 The summary of mapping rate and coverage

Sample:	test1
Total:[1]	34269240 (100%)
Duplicate:[2]	1786151 (5.23%)
Mapped:[3]	34164013 (99.69%)
Properly mapped:[4]	33848206 (98.77%)
PE mapped:[5]	34064960 (99.40%)
SE mapped:[6]	198106 (0.58%)
Initial_bases_on_target:[7]	50390601
Initial_bases_on_or_near_target:[8]	124292823
Total_effective_yield(Mb):[9]	5030.94
Effective_yield_on_target(Mb):[10]	3097.9
Fraction_of_effective_bases_on_target:[11]	61.60%
Fraction_of_effective_bases_on_or_near_target:[12]	84.70%
Average_sequencing_depth_on_target:[13]	61.48
Bases_covered_on_target:[14]	50264948
Coverage_of_target_region:[15]	99.80%
Fraction_of_target_covered_with_at_least_100x:[16]	13.40%
Fraction_of_target_covered_with_at_least_50x:[17]	58.00%
Fraction_of_target_covered_with_at_least_20x:[18]	92.60%
Fraction_of_target_covered_with_at_least_10x:[19]	98.30%
Fraction_of_target_covered_with_at_least_4x:[20]	99.50%

Note:
(1) Total: The total number of clean reads. The below indexes are calculated based on the clean reads.
(2) Duplicate: The number of duplicated reads (percentage: duplicated reads/clead reads).
(3) Mapped: The number of total reads that mapped to the reference genome (percentage).
(4) Properly mapped: The number of reads that mapped to the reference genome and the direction is right (percentage).
(5) PE mapped: The number of pair-end reads that mapped to the reference genome (percentage).
(6) SE mapped: The number of single-end reads that mapped to the reference genome (percentage).
(7) Initial_bases_on_target: Total bases mapped to the target regions(exonic regions we capture).
(8) Initial_bases_near_target: Total based mapped to the flanking regions (The regions nearby target upstream and downstream 200bp).
(9) Total_effective_yield(Mb): Total data size of reads that mapped to the reference genome.
(10) Effective_yield_on_target(Mb): Total data size of reads that mapped to target regions.

(11) Fraction_of_effective_bases_on_target: The percentage of the mapped reads in target regions to the reads in reference genome.
(12) Fraction_of_effective_bases_on_near_target: The percentage of the mapped reads in target regions and flanking regions to the reads in reference genome.
(13) Average_sequencing_depth_on_target: The average sequencing depth that mapped to the reference genome target regions.
(14) Base_covered_on_target: The coverage length of target regions.
(15) Coverage_of_target_region: The percentage of target region coveraged.
(16) Fraction_of_target_covered_with_at_least_100x: The percentage of bases with depth >100x in target regions.
(17) Fraction_of_target_covered_with_at_least_50x: The percentage of bases with depth >50x in target region.
(18) Fraction_of_target_covered_with_at_least_20x: The percentage of bases with depth >20x in target region.
(19) Fraction_of_target_covered_with_at_least_10x: The percentage of bases with depth >10x in target region.
(20) Fraction_of_target_covered_with_at_least_4x: The percentage of bases with depth >4x in target region.

4.4 Variation Detection Result

4.4.1 SNV Result

Generally, the whole human genome has about 3.6~4.4 million SNVs. Most (above 95%) SNVs with high frequency (the allele frequency in population is above 5%) have been recorded in dbSNP (Sherry S T *et al.*). These high-frequency SNVs are generally not pathogenic and have little value for disease research.
We use GATK to detect SNV, and the statistics of SNVs are as follows:

Table 4.6 The number of SNVs in different genomic regions

Sample[1]	Exonic[2]	Intronic[3]	UTR3[4]	UTR5[5]	Intergenic[6]	ncRNA_exonic[7]	ncRNA_intronic[8]	Upstream[9]	Downstream[10]	Splicing[11]	ncRNA_splicing[12]
test1	16635	12418	849	680	1196	698	416	231	86	1571	18

Note:
(1) Sample: Sample name.
(2) exonic: The number of SNVs in exonic regions.
(3) intronic: The number of SNVs in intronic regions.
(4) UTR3: The number of SNVs in 3'UTR regions.
(5) UTR5: The number of SNVs in 5'UTR regions.
(6) intergenic: The number of SNVs in intergenic regions.
(7) ncRNA_exonic: The number of SNVs in non-coding RNA exonic regions.
(8) ncRNA_intronic: The number of SNVs in non-coding RNA intronic regions.
(9) upstream: The number of SNVs in the 1kb upstream regions of transcription start sites.
(10) downstream: The number of SNVs in the 1kb downstream regions of transcription ending sites.
(11) splicing: The number of SNVs in 10bp splicing junction regions.
(12) ncRNA_splicing: The number of SNVs in 10bp splicing junction regions of non-coding RNAs.

Table 4.7 The number of different SNV types in coding regions

Sample[1]	synonymous_SNV[2]	missense_SNV[3]	Stopgain[4]	Stoploss[5]	Unknown[6]
test1	8791	7568	42	9	225

Note:
(1) Sample: Sample name.
(2) synonymous_SNV: A single nucleotide change that does not cause an amino acid change.
(3) missense_SNV: A single nucleotide change that causes an amino acid change.
(4) stopgain: A nonsynonymous SNV that leads to the creation of stop codon at the variant site.
(5) stoploss: A nonsynonymous SNV that leads to the elimination of stop codon at the variant site.
(6) unknown: Unknown function (due to imperfect information in the gene structure definition in the database file).

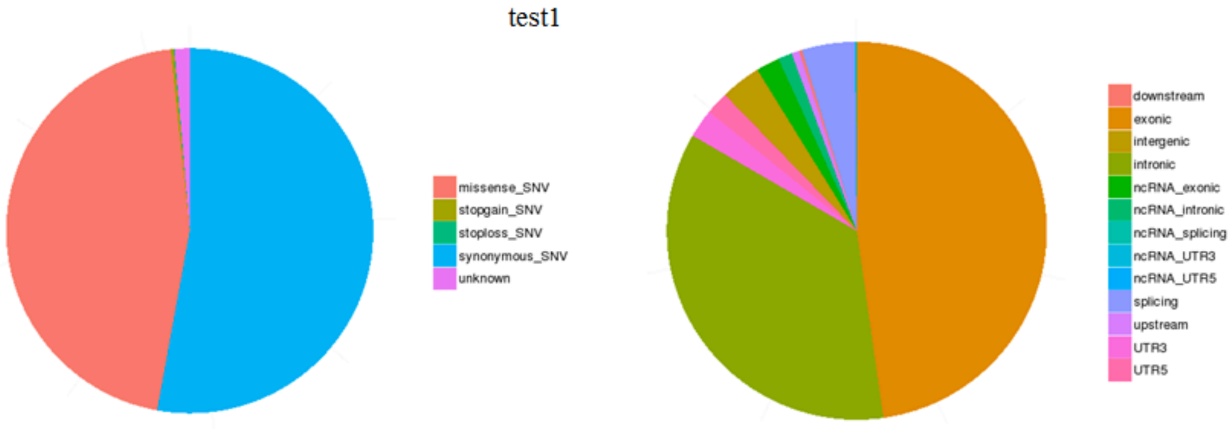

Figure 4.6 The number of different SNV types in coding regions (left) and the number of SNVs in different genomic regions (right).

In the left figure, synonymous_SNV means a single nucleotide change that does not cause an amino acid change; missense_SNV means a single nucleotide change that cause an amino acid change; stopgain_SNV means a nonsynonymous SNV that lead to the immediate creation of stop codon at the variant site; stoploss_SNV means a nonsynonymous SNV that lead to the immediate elimination of stop codon at the variant site; unknown means unknown function (due to various errors in the gene structure definition in the database file).
In the right figure, downsteam means the number of SNV in the 1kb downstream region of transcription ending site; exonic means the number of SNV in exonic region; intergenic means the number of SNV in intergenic region; intronic means the number of SNV in intronic region; ncRNA_exonic means the number of SNV in non-coding RNA exonic region; ncRNA_intronic means the number of SNV in non-coding RNA intronic region; ncRNA_splicing means the number of SNV in 10bp splicing junction of non-coding RNA; ncRNA_UTR3 means the number of SNV in 3'UTR of non-coding RNA; ncRNA_UTR5 means the number of SNV in 5'UTR of non-coding RNA; splicing means the number of SNV in 10bp splicing junction region; upstream means the number of SNV in the 1kb upstream region of transcription start site; UTR3 means the number of SNV in 3'UTR region; UTR5 means the number of SNV in 5'UTR region.

The ratio of Ts/Tv can reflect the accuracy of sequencing. Generally, the ratio is about 2.2 in whole human genome and is about 3.2 in coding regions.

Table 4.8 The distribution of transition and transversion

Sample[1]	novel_ts[2]	novel_ts/tv[3]	novel_tv[4]	ts[5]	ts/tv[6]	tv[7]
test1	362	2.167664671	167	25332	2.678367520	9458

Note:
(1) Sample: Sample name.
(2) novel_ts: The number of ts SNVs that are not in dbSNP
(3) novel_ts/tv: The ratio of novel ts/ novel tv.
(4) novel_tv: The number of tv SNVs that are not in dbSNP.
(5) transition(ts): Transition
(6) ts/tv: The ratio of the number of transition/the number of transversion.
(7) transversion(tv): Transversion

Table 4.9 The distribution of SNVs and genotypes

Sample[1]	all[2]	genotype.Het[3]	genotype.Hom[4]	novel[5]	novel_proportion[6]
test1	34790	20334	14456	529	1.52%

Note:
(1) Sample: Sample name.
(2) all: The total number of SNVs.
(3) genotype.Het: The number of heterozygous genotypes.
(4) genotype.Hom: The number of homozygous genotypes.
(5) novel: The number of SNVs not in dbSNP.
(6) novel_proportion: The ratio of novel SNVs/total number of SNVs.

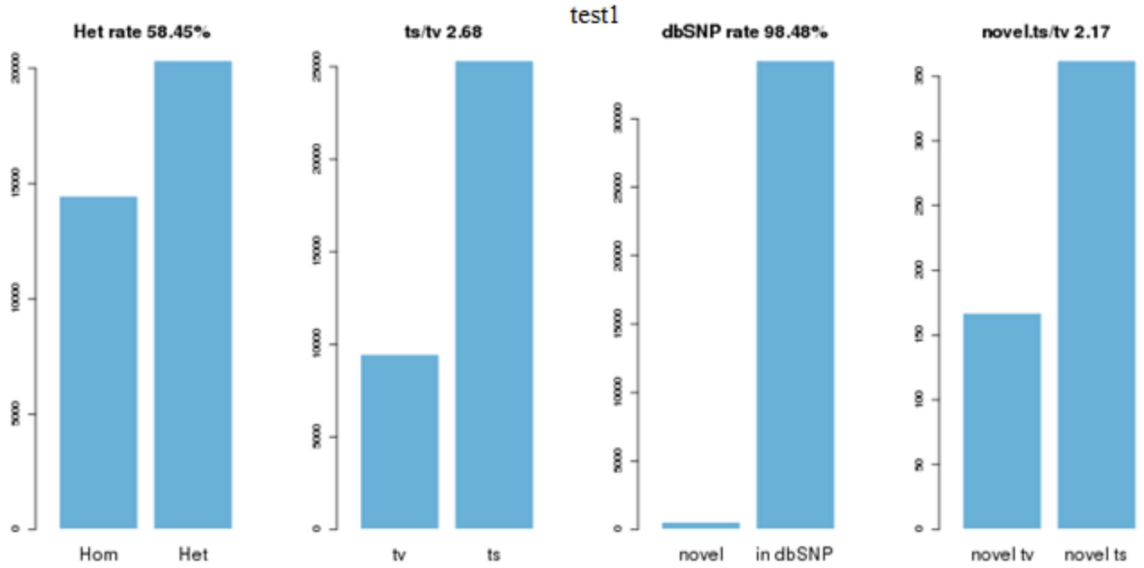

Figure 4.7 SNV features in genome

In the first figure, Hom is short for homozygous and Het for heterozygous; the y-axis means the number of SNVs with homozygous or heterozygous genotype.
In the second figure, tv means transversion and ts means transition; the y-axis means the number of SNVs corresponding to tv or ts.
In the third figure, novel means SNVs not in dbSNP; in dbSNP means SNVs in dbSNP. dbSNP rate is calculated as the number of SNVs in dbSNP/total number of SNVs.
In the fourth figure, novel tv (novel ts) means the number of tv(ts) SNVs that are not in dbSNP.

4.4.2 InDel Result

Generally, the genome of human has about 350K InDels (insertion and deletion). The InDels in coding regions or splicing sites may change 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. Compared to non-frameshift mutation, frameshift mutation is more limited by selective pressure, which is confirmed by figure 4.8 showing the InDel length distribution.
We use GATK to detect InDels, and the obtained result is as follows:

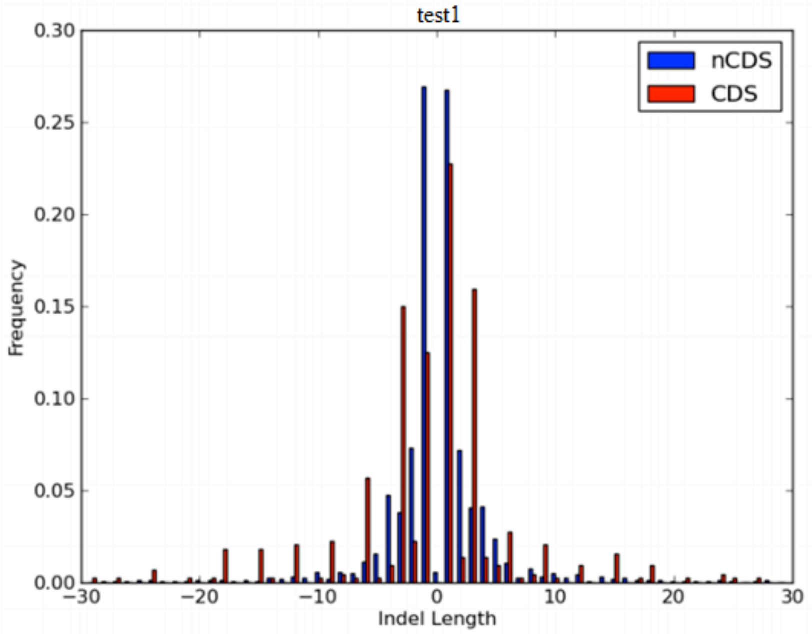

Figure 4.8 The distribution of InDel lengths.

The x-axis is the indel length and y-axis is the corresponding frequency. CDS means coding regions and splicing sites; nCDS means other regions. This figure shows that in CDS, compared to nonframeshift mutation, frameshift mutation is more limited by selective pressure.

Table 4.10 The number of InDels in different genomic regions

Sample[1]	Exonic[2]	Intronic[3]	UTR3[4]	UTR5[5]	Intergenic[6]	ncRNA_exonic[7]	ncRNA_intronic[8]	Upstream[9]	Downstream[10]	Splicing[11]	ncRNA_splicing[12]
test1	351	1326	88	74	93	49	50	15	10	173	3

Note:
(1) Sample: Sample name
(2) exonic: The number of InDel in exonic region
(3) intronic: The number of InDel in intronic region
(4) UTR3: The number of InDel in 3'UTR region
(5) UTR5: The number of InDel in 5'UTR region
(6) intergnic: The number of InDel in intergenic 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) upstream: The number of InDel in the 1kb upstream region of transcription start site
(10) downstream: The number of InDel in the 1kb downstream region of transcription ending site
(11) splicing: The number of InDel in 10bp splicing junction region
(12) ncRNA_splicing: The number of InDel in 10bp splicing junction of non-coding RNA

Table 4.11 The number of different InDel types in coding regions

Sample[1]	frameshift_deletion[2]	frameshift_insertion[3]	nonframeshift_deletion[4]	nonframeshift_insertion[5]	stoploss[6]	stopgain[7]	Unknown[8]
test1	63	45	81	91	0	1	61

Note:
(1) Sample: Sample name
(2) frameshift_deletion: A deletion that causes frameshift changes in protein coding sequence and the deletion length is not multiple of 3.

(3) frameshift_insertion: An insertion that causes frameshift changes in protein coding sequence and the insertion length is not multiple of 3.
(4) nonframeshift_deletion: Non-frameshift deletion, the deletion does not change coding protein frame and the deletion length is multiple of 3.
(5) nonframeshift_insertion: Non-frameshift insertion, the insertion does not change coding protein frame and the insertion length is multiple of 3.
(6) stoploss: An InDel leads to the immediate elimination of stop codon at the variant site
(7) stopgain: An InDel leads to the immediate creation of stop codon at the variant site.
(8) unknown: Unknown function. Due to imperfect information in the gene structure databases, some InDels cannot be annotated.

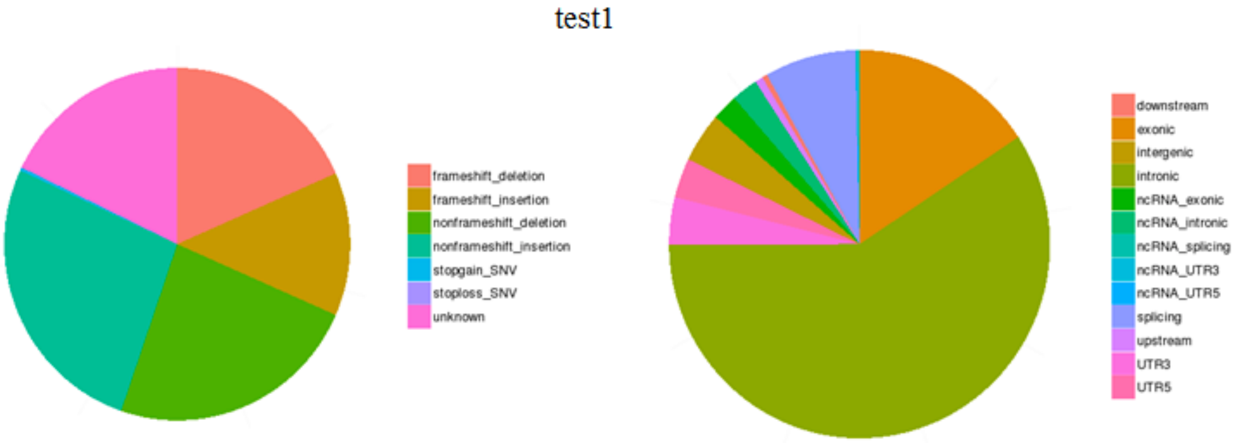

Figure 4.9 The number of different InDel types in coding regions (left), and the number of InDels in different genomic regions (right).

Table 4.12 The distribution of InDels and genotypes

Sample[1]	all[2]	genotype.Het[3]	genotype.Hom[4]	novel[5]	novel_proportion[6]
test1	2223	1045	1178	1321	0.594242015

Note:
(1) Sample: Sample name
(2) all: The total number of InDel
(3) genotype.Het: The number of the heterozygous sites
(4) genotype.Hom: The number of the homozygous sites
(5) novel: InDel not in dbSNP
(6) novel_proportion: The ratio of novel InDel/total number of InDel

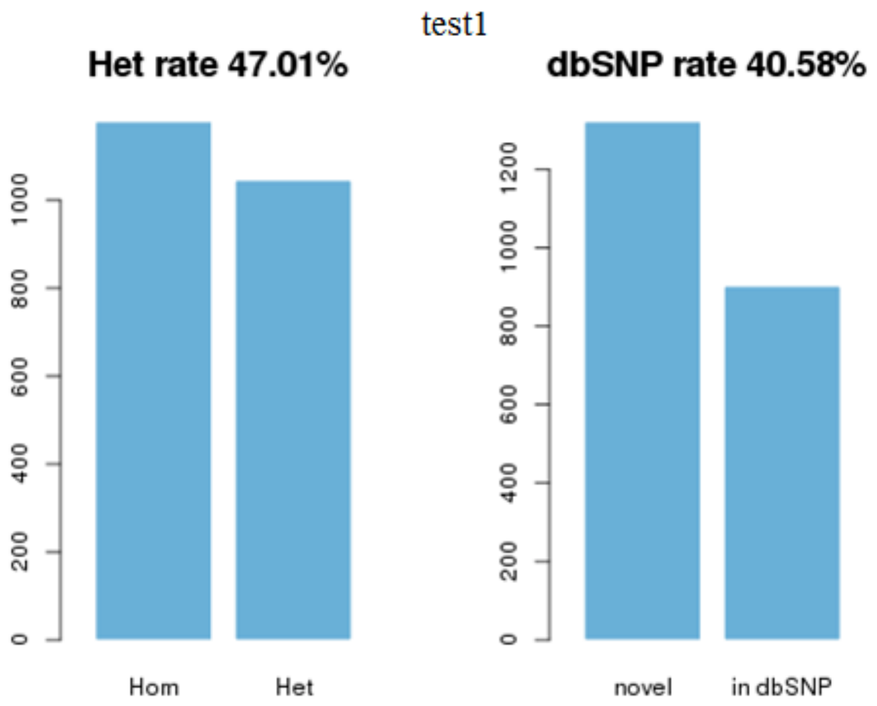

Figure 4.10 The InDel features in genome.

In the left figure, Hom is short for homozygote and Het for heterozygote; the y-axis means the number of InDels with homozygous and heterozygous genotype.
In the right figure, novel means InDels not in dbSNP; in dbSNP means InDels in dbSNP. dbSNP rate is calculated as the number of InDels in dbSNP/total number of InDels.

4.4.3 Annotation Result

We use ANNOVAR (Wang K *et al.*) to annotate SNVs and InDels, which includes annotation information from dbSNP, the 1000 Genomes Project and other published databases. Annotation contains the variation's position, type, conservation prediction, etc. Table 4.13 show annotation details.

Table 4.13 The annotation results

Priority[1]	**CHROM**[2]	**POS**[3]	**ID**[4]	**REF**[5]	**ALT**[6]	**QAUL**[7]	**FILTER**[8]	**GeneName**[9]	**Func**[10]	**Gene**[11]	**GeneDetail**[12]	**ExonicFunc**[13]	**AAChange**[14]	[15-69]
L	**1**	14653	.	C	T	841.77	PASS	WASH7P	ncRNA_exonic	.	.	.	.	…...
H	**1**	16631	.	T	C	541.77	PASS	WASH7P	ncRNA_exonic	.	.	.	.	…...

Note: Annotation information includes six parts: Priority (1), Chromosomal regions and gene structures (2-21), Database annotation (22-39), Functional prediction (40-51), Basic information on the variation (52-58), Gene function and pathway annotation (59-68).

The first part is priority information—1stBASE sets priority levels scientifically according to criterion on prioritizing candidate variants used in published studies. Priority suggests the importance of the variant and serves as some guide.

(1) Priority: The value may be H (high), M (medium) or L (low). High indicates that the following conditions must be met: 1. The variant is not in repeat regions of human genome(i.e. 'genomicSuperDups' and 'Repeat' have no annotation information specified with a dot '.'); 2. Allele frequency of the variant in 1000 Genomes Project is below 0.01; 3. The variant hits exon or splicing regions; 4. At least one of the four functional prediction, i.e. SIFT, Polyphen, MutationTaster and CADD, for this variant is deleterious. Medium indicates that the first three conditions mentioned above must be met. Low indicates the remaining variants.

The second part shows information of chromosomal regions and gene structures related to the variant.

(2) CHROM: Chromosome ID.

(3) POS: The position of the variant on chromosomes. The value refers to the position of the first base in the REF sting.

(4) ID: The rs number of the variant in dbSNP.

(5) REF: Reference base(s).

(6) ALT: Alternate base(s). Comma separated list of alternate non-reference alleles called on at least one of the samples.

(7) QUAL: Quality value for the variant. Phred-scaled quality score for the assertion made in ALT. i.e. $-10log_{10}$ prob(call in ALT is wrong).

(8) FILTER: Filter status, PASS if the position has passed all filters.

(9) GeneName: Names of genes in which this variant is located according to the refGene annotations.

(10) Func: This field tells whether the variant hit exons or hit intergenic regions, or hit introns, or hit a non-coding RNA gene. The value of this field takes the following precedence: exonic = splicing > ncRNA> > UTR5/UTR3 > intronic > upstream/downstream > intergenic. Notes: 1. When a variant hit different genes or transcripts, the variant may fit multiple functional categories, and then the precedence mentioned above is used to decide what function to print out; 2. The "exonic" here refers only to coding exonic portion , but not UTR portion, as there are two keywords (UTR5, UTR3) that are specifically reserved for UTR annotations; 3. If a variant is located in both 5' UTR and 3' UTR region (possibly for two different genes), then the "UTR5,UTR3" will be printed as the output; 4. "splicing" in ANNOVAR is defined as variant that is within 2-bp away from an exon/intron boundary by default, but 1st BASE changed the threshold to be 10-bp; 5. "splicing" in ANNOVAR only refers to the 10bp in the intron that is close to an exon; 6. The term "upstream" and "downstream" is defined as 1-kb away from transcription start site or transcription end site, respectively, taking in account of the strand of the mRNA. If a variant is located in both downstream and upstream region (possibly for 2 different genes), then the "upstream,downstream" will be printed as the output.

(11) Gene: The transcript name(s). If a variant has 'intergenic' in 'Func' field, this field will give the two neighboring transcripts. If a variant hits multiple transcripts with different functional categories, only transcript names in accordance with the value of 'Func' field will be output. For example, rs333970 hits the exonic, splicing, intronic, exonic of the four transcripts of gene *CSF1*, the 'Func' value will be "exonic;splicing" and the 'Gene' value will be "NM_000757, NM_172210, NM_172212" (NM_172211 will be ignored).

(12) GeneDetail: Description of the sequence change in UTR, splicing, ncRNA_splicing or intergenic region. If 'Func' is 'exonic;splicing' or 'splicing', this field gives the sequence change in splicing region(s); for example, NM_172210:exon6:c.1090 +5C>A, NM_172210 is the transcript identifier; exon6:c.1090+5C>A is the sequence change and means that this C>A substitution is at the fifth base downstream from the 6th exon (1090 is the end position of the 6th exon of the cDNA). If 'Func' is 'intergenic', this field gives the distance to the neighboring transcripts, such as 'dist=1366;dist=22344'. If 'Func' is 'UTR*', this field gives the sequence change in UTR; for example, NM_198576:c.*19C>T means that this C>T substitution is at the 19th base downstream from stop codon on NM_198576.

(13) ExonicFunc: This field tells the functional consequences of the variant (possible values include: missense SNV, synonymous SNV, frameshift insertion, frameshift deletion, nonframeshift insertion, nonframeshift deletion, frameshift block substitution, nonframshift block substitution, stopgain, stoploss, unknown).

(14) AAChange: This field tells the amino acid changes as a result of the exonic variant. Only exonic variants have information in this field, i.e. when 'Func' is 'exonic' or 'exonic; splicing', this field gives the amino acid change in each related transcript. For example, AIM1L:NM_001039775:exon2:c.C2768T:p.P923L, AIM1L is gene name; NM_001039775 is the transcript identifier; exon2 means this variant is on the second exon of NM_001039775; c.C2768T is the sequence change and means that this C>T substitution is at the 2,768 position on the cDNA; p.P923L is the amino acid change and means that the 923 amino acid on protein is changed from Pro to Leu due to this variant. Another example, NADK:NM_001198995:exon10:c.1240_1241insAGG:p.G414delinsEG, c.1240_1241insAGG is the sequence change and means that there is a 3bp insertion between position 1,240 and 1,241 on the cDNA; p.G414delinsEG is the amono acid change and means that Gly at the 414th amino acid on protein is changed to Glu-Gly.

(15) Gencode: The transcript name(s) in which this variant is located according to Gencode gene definitions.

(16) cytoband: This field gives the Giemsa-stained chromosomes bands. When a variant spans multiple bands, they will be connected by a dash (for example, 1q21.1-q23.3).

(17) wgRna: Gene names of snoRNAs and microRNAs based on the miRBase Release and snoRNABase.

(18) targetScanS: The targetScanS annotation database offered by UCSC gives conserved mammalian microRNA regulatory target sites for conserved microRNA families in the 3' UTR regions of Refseq Genes, as predicted by TargetScanHuman 5.1. This field tells whether the variant disrupts predicted microRNA binding sites. The output consists of a score and a name. The score of target site ranges from from
1-1000; the smaller the score, the target site is more confident. The name shows the name of microRNA acting on the target. For instance, "Score=62;Name=KRAS:miR-181:1" means that the predicted target site is within the UTR3 region of gene KRAS and that the microRNA named miR-181:1 acts on this target site.

(19) tfbsConsSites: This field tells whether the variant disrupts transcription factor binding sites conserved in the human/mouse/rat alignment and gives the Score and Name annotation for the transcription factor binding sites. The score represents the normalized score. The name represents binding site motif name. For example, Score=765;Name=V$PAX5_02. Users can investigate what transcription factors may recognize this motif using many online resources, for example, MSigDB provides gene list that recognize these motifs, see for example http://www.broadinstitute.org/gsea/msigdb/cards/V$PAX5_02.

(20) genomicSuperDups: This field tells whether the variant hits segmental duplications in reference genome. Variants that are mapped to segmental duplications are most likely sequence alignment errors and should be treated with extreme caution. The 'Score' field in output is the sequence identity ranging from 0 to 1 between two genomic segments. The 'Name' field represents the other "matching" segments in genome. For example, 'Score=0.994828; Name=chr19:60000' means that the fragment at the position of chr19:60000 is homologous to the fragment containing this variant, and the sequence identity is 0.994828. Note, for a region to be included in the segmental duplications, at least 1 Kb of the total sequence (containing at least 500bp of non-RepeatMasked sequence) had to align and a sequence identity of at least 90% was required.

(21) Repeat: This field tells whether the variant hits interspersed repeats and low complexity DNA sequences output by RepeatMasker program, such as SINE, LINE and Simple repeats. For example, 'Score=180;Name=1385:(CACCC) n(Simple_repeat)', 180 is the score of the repeat,(CACCC) n is the name of the repeat, 'Simple_repeat' is type of repeat. Note, variants mapped to repeats are likely to be false and should be treated with extreme caution.

The third item is database annotataion—There are a great number of common polymorphism sites in human population, while many deleterious variants are rare or low-frequency. This part gives the allele frequency and clinical information for each variant.

(22) avsnp144: The rs number of the variant in dbSNP (build 144).

(23) clinvar_20150330: The ClinVar database archives and aggregates information about relationships among variations and human health. An example is 'CLINSIG=probable-non-pathogenic;CLNDBN=not_specified;CLNREVSTAT=single;CLNACC=RCV000116272.2;CLNDSDB=MedGen; CLNDSDBID=CN169374'. CLINSIG refers to Variant Clinical Significance, including unknown, untested, non-pathogenic, probable-non-pathogenic, probable-pathogenic, pathogenic, drug-response, histocompatibility, other; CLINDBN refers to variant disease name; CLNREVSTAT refers to ClinVar Review Status, mult-Classified by multiple submitters, single-Classified by single submitter, not-Classified by submitter, exp-Reviewed by expert panel, prof-Review by professional society; CLINACC refers to Variant Accession and Versions; CLNDSDB refers to variant disease database name; CLNDSDBID refers to variant disease database ID.

(24) gwasCatalog: This field tells whether this variant was previously reported to be associated with diseases or traits in genome-wide association studies. It lists the disease names related to this variation. "." means this variation has not been reported by published GWAS study.

(25) 1000g2015aug_eas: This field gives the allele frequency for the allele in ALT in 1000 Genomes Project (released in August, 2015) in East Asian population.

(26) 1000g2015aug_sas: This field gives the allele frequency for the allele in ALT in 1000 Genomes Project (released in August, 2015) in South Asian population.

(27) 1000g2015aug_eur: This field gives the allele frequency for the allele in ALT in 1000 Genomes Project (released in August, 2015) in European population.

(28) 1000g2015aug_afr: This field gives the allele frequency for the allele in ALT in 1000 Genomes Project (released in August, 2015) in African population.

(29) 1000g2015aug_amr: This field gives the allele frequency for the allele in ALT in 1000 Genomes Project (released in August, 2015) in Admixed American population.

(30) 1000g2015aug_all: This field gives the allele frequency for the allele in ALT in 1000 Genomes Project (released in August, 2015) in ALL population.

(31) esp6500siv2_all: The ESP is a NHLBI funded exome sequencing project aiming to identify genetic variants in exonic regions from over 6000 individuals, including healthy ones as well as subjects with different diseases. This field gives alternative allele frequency for the variant in ESP.

(32) ExAC_ALL: ExAC is short for Exome Aggregation Consortium. The data set spans 60,706 unrelated individuals and should serve as a useful reference set of allele frequencies for severe disease studies. Currently supported population groups include ALL, AFR (African), AMR (Admixed American), EAS (East Asian), FIN (Finnish), NFE (Non-finnish European), OTH (other) and SAS (South Asian). ExAC_ALL gives alternative allele frequency for the variation in ALL ExAC samples.

(33) ExAC_AFR: The alternative allele frequency for the variation in ExAC for African population.

(34) ExAC_AMR: The alternative allele frequency for the variation in ExAC for Admixed American population.

(35) ExAC_EAS: The alternative allele frequency for the variation in ExAC for East Asian population.

(36) ExAC_FIN: The alternative allele frequency for the variation in ExAC for Finnish population.

(37) ExAC_NFE: The alternative allele frequency for the variation in ExAC for Non-Finnish European population.

(38) ExAC_OTH: The alternative allele frequency for the variation in ExAC for other population.

(39) ExAC_SAS: The alternative allele frequency for the variation in ExAC for South Asian population.

The fourth part is functional prediction from multiple tools—these annotations can help to evaluate deleteriousness of a variation. SIFT, Polyphen2, MutationTaster, LRT, MutationAssessor and FATHMM are similar and all predict whether an amino acid substitution affects protein function; only coding variants have these annotations. phyloP, SiPhy, gerp++ and CADD are similar and predict the conservation level of the site; these types of "conservation scores" only consider conservation level at the current base, and they do not care about the actual nucleotide identity, so synonymous and non-synonymous variants at the same site will be scored as the same; these scores are used for finding functionally important sites, so variants that confer increased susceptibility may be scored well.

(40) SIFT: SIFT annotation (dbNSFP version 3.0a). The annotation consists of score and categorical prediction; the scores and predictions are separated by comma. There are two possible predictions: D (Deleterious, score<=0.05)); T (Tolerated, score >0.05).

(41) Polyphen2_HVAR: PolyPhen 2 (dbNSFP version 3.0a) annotation based on HumanVar database. This annotation should be used for diagnostics of Mendelian diseases. The annotation consists of score and categorical prediction. There are three possible predictions: D (Porobably damaging, score>=0.909), P (possibly damaging, 0.447<=score<=0.909), B (benign, score<=0.446).

(42) Polyphen2_HDIV: PolyPhen 2 (dbNSFP version 3.0a) annotation based on HumanDiv database. This annotation should be used when evaluating rare alleles at loci potentially involved in complex phenotypes, dense mapping of regions identified by genome-wide association studies, and analysis of natural selection from sequence data. The annotation consists of score and categorical prediction. There are three possible predictions: D (Porobably damaging, score>=0.957), P (possibly damaging, 0.453<=score<=0.956), B (benign, score<=0.452).

(43) MutationTaster: MutationTaster annotation (dbNSFP version 3.0a). The annotation consists of score and categorical prediction. There are four possible predictions: 'A' (Disease_causing_automatic), 'D (Disease_causing), 'N' (Polymorphism), 'P' (Polymorphism_automatic'). D and N are categorized by only score, while A and P are categorized by score and other information (if nonsynonymous SNV leads to stop-gain, the variation will be predicted an 'A'; if all three genotypes of nonsynonymous SNV has frequency information in HapMap, the variation will be predicted a 'P'). So, both A and D should be considered deleterious.

(44) LRT: LRT annotation (dbNSFP version 3.0a). The annotation consists of score and categorical prediction. There are three possible predictions: D (Deleterious), N (Neutral), U (Unknown).

(45) MutationAssessor: MutationAssessor annotation (dbNSFP version 3.0a). The annotation consists of score and categorical prediction. There are four possible predictions: H (high), M (medium), L (low), N (neutral). H/M means functional and L/N means non-functional.

(46) FATHMM: FATHMM annotation (dbNSFP version 3.0a). The annotation consists of score and categorical prediction. There are two possible predictions: D (Deleterious, score<=-1.5)); T (Tolerated, score >-1.5).

(47) phyloP7way_vertebrate: PhyloP score (dbNSFP version 3.0a) based on the whole genome alignment of 7 vertebrates. Generally the higher the score, the more conserved the site.

(48) **phyloP20way_mammalian:** PhyloP score (dbNSFP version 3.0a) based on the whole genome alignment of 20 mammals.
(49) **SiPhy_29way_logOdds:** SiPhy score (dbNSFP version 3.0a) based on the whole genome alignment of 29 mammals genomes. The larger the score is, the more conserved the site.
(50) **gerp++gt2:** GERP++ scores for all mutations with GERP++>2 in human genome, as this threshold is typically regarded as evolutionarily conserved and potentially functional. Variants with '.' in this field should be considered not conserved. The larger the score is, the more conserved the site.
(51) **CADD:** CADD (Combined Annotation Dependent Depletion) is a score that is based on SVM on multiple other scores. It assigns a score to each possible mutation in the human genome including non-coding and coding variants. In the output, the comma-delimited values are raw scores and phred-scaled scores. "." stands for CADD score <10. For phred-scaled scores, 10 means 10% percentile highest scores, 20 means 1% percentile highest scores, and 30% means 0.1% percentile highest scores. CADD official website suggests 15 as a cutoff; in published studies, 10 or 15 are used as a cutoff.

The fifth item is basic information on the variation—This part shows the detail information of variation, including INFO, genotypes, et al.

(52) **INFO:** Information about this variation from variant calling software. INFO fields are encoded as a semicolon-separated series of short keys with optional values in the format: <key>=<data>[,data].
(53) **FORMAT:** The FORMAT field specifies the data types and order (colon-separated alphanumeric String). This is followed by one field per sample, with the colon-separated data corresponding to the types specified in the FORMAT.

GT: genotype, encoded as allele values separated by either of / or |. The allele values are 0 for the reference allele (what is in the Ori_REF field), 1 for the first allele listed in Ori_ALT, 2 for the second allele list in Ori_ALT and so on. 0/0 and 1/1 represent homozygous. 0/1 represents heterozygous. '.' means that a call cannot be made for a sample at a given locus.

AD: Allelic depths for the ref and alt alleles in the order listed (Allelic depths).
DP: Approximate read depth (reads with MQ=255 or with bad mates are filtered).
GQ: Genotype Quality.
PL: Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification.

(54) **SampleID:** The colon-separated data in this sample corresponding to the types specified in the FORMAT.
(55) **Ori_REF:** The reference allele (what is in the REF field) in VCF file. According to the annotation workflow at 1st BASE (mentioned above), for InDel, the allele in "REF" field in this file may be different from (usually shorter than) the "REF" in VCF file.
(56) **Ori_ALT:** The alternative allele(s) (what is in the ALT field) in VCF file. In this file, the allele in "ALT" field corresponds to one alelle in "Ori_ALT" field; according to the annotation workflow at 1st BASE (mentioned above), for InDel, the allele in "ALT" field may be different from (usually shorter than) the corresponding allele in the "Ori_ALT" field.
(57) **shared_hom:** Number of individuals who have homozygous genotype at this site.

(58) **shared_het:** Number of individuals who have heterozygous genotype at this site.

The sixth item is gene function and pathway annotation—These annotations are for genes containing this variation.

(59) **OMIM:** Annotation from Online Mendelian Inheritance in Man (OMIM).
(60) **GWAS_Pubmed_pValue:** GWAS_Pubmed_pValue: Annotation from the NHGRI-EBI GWAS Catalog. The value is like 'pubmedID(p-value);pubmedID(p-value)'. 'pubmedID' is PubMed ID of publication of the study which reported the association between the variation and disease. 'p-value' is the corresponding p-Value in the publication.
(61) **HGMD_ID_Diseasename:** Annotation from the Human Gene Mutation Database (HGMD®). The value is like 'ID_HGMD(Disease_name);ID_HGMD(Disease_name)'. ID_HGMD is HGMD internal identifier. Disease_name is the name for the disease or condition associated with the mutation.
(62) **HGMD_mutation:** Annotation from the Human Gene Mutation Database (HGMD®). The value is infomation about this variant.
(63-65) **GO_BP, GO_CC, GO_MF:** Annotation from Gene Ontology. BP is Biological Process; CC is cellular component; MF is molecular function.
(66) **KEGG_PATHWAY:** Annotation from KEGG PATHWAY Database.
(67) **PID_PATHWAY:** Annotation from PID (Pathway Interaction Database).
(68) **BIOCARTA_PATHWAY:** Annotation from BioCarta.
(69) **REACTOME_PATHWAY:** Annotation from Reactome Pathway Database.

5 Advanced analysis

5.1 Variant filtering using known databases

We merge VCF files from multiple samples into one and use ANNOVAR to annotate variants. Then, variants are filtered to identify candidate mutations that may be associated with diseases. SNP and InDels are processed separately.
The filtering steps are:

(1) Keep variants with allele frequency < 1% in ALL population from the phase III of the 1000 Genomes Project (i.e. 1000g2015aug_all).

(2) Keep variants that hit exon or splicing regions.

(3) Discard synonymous SNVs.

(4) Keep variants for which at least half of the four functional predictions, i.e. SIFT, Polyphen, MutationTaster and CADD, is deleterious.

Table 5.1 The result of variant filtering

Total	1000G	Function	Synonymous	Deleterious
34790	1013	408	318	192

Notes:
Total: the total number of variants.
1000G: the number of variants survived step (1).
Function: the number of variants survived step (2).
Synonymous: the number of variants survived step (3).
Deleterious: the number of variants survived step (4).

5.2 Variant filtering based on disease model

After variant filtering using known databases, we get candidate mutations that may be associated with diseases. Based on the pedigree of each family, we further perform variant filtering considering disease model for each family.

According to the genetic information form, we can determine the genetic model for the given disease. Then, basing on the disease model, we further screen the variants. For instance, in single gene autosomal dominant model, we selected variants whose genotype is '0/0' for normal individuals but is not '0/0' for patients.

Table 5.2 The statistics of result

Family ID	SNP	INDEL
F1	49	80

Notes: SNP and INDEL mean the number of SNVs and InDels survived the variant filtering steps.

5.3 Linkage analysis

Linkage analysis tests for co-segregation of a chromosomal region and a trait of interest. It relies on using family-based data to detect genetic loci that may harbor disease predisposing genes. Several methods have been proposed to detect linkage: the U scores, the sib pair test, the likelihood ratios, the lod score method. The lod score method is the one most commonly used at present. LOD score, or logarithm of odds score, is a statistical test used in genetic linkage analysis. The LOD score compares the probability of obtaining the test data if the two loci are linked to the probability of obtaining the test data if the two loci are not linked. For Mendelian diseases, the decision thresholds of the test are usually set at -2 and +3, i.e. LOD score > 3 suggests linkage, and LOD score < -2 rejects linkage; for regions with LOD scores between -2 and 3, it is necessary to go on accumulating information.

1st BASE use MERLIN to run a non-parametric linkage analysis based on the called SNVs and SNPs in HapMap 2 (CEU).

Here, we list a simple graphical summary of the linkage results for each family. All chromosomal regions within which LOD scores for all selected SNPs are greater than 0.4 are showed.

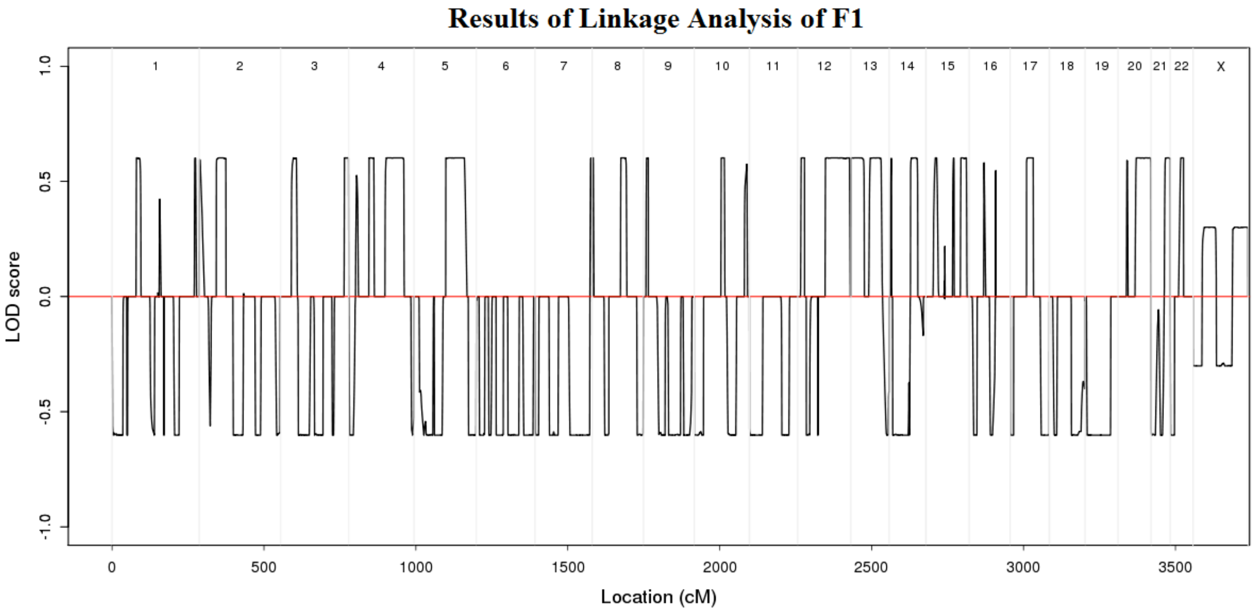

Figure 5.1 Linkage analysis for family F1

Note: The upper x-axis is chromosome number; the lower x-axis is centimorgan (cM); the y-axis is LOD score.

5.4 Regions of homozygosity (ROH) analysis

"Runs of homozygosity" or ROH are regions of the genome where the copies inherited from our parents are identical. This creates a run of homozygous variants, from tens of thousands to millions of letters in length. The two DNA copies are identical because our parents have inherited them from a common ancestor at some point in the past, recently in the case of a cousin marriage, but in fact we all carry ROH, because going far enough back in time we are all related. The distribution of ROH may be important medically. This is because they allow certain variants, called recessive variants to be expressed. Recessive variants only have their effect when present on both copies of an individual's genome, for example in a run of homozygosity. Recessive variants cause many genetic diseases such as cystic fibrosis, phenylketonuria and Tay-Sachs disease.

1st BASE use homozygosity heterogeneous hidden Markov model (H^3M^2) to detect ROH from WES data.

Table 5.3 The result of ROH

Priority[1]	CHROM[2]	POS[3]	ID[4]	REF[5]	ALT[6]	QAUL[7]	FILTER[8]	Gene Name[9]	Func[10]	11-59
Region	6	87963529	88070410	C6orf163;GJB7;SMIM8;ZNF292						
L	6	87963529	rs3857485	G	A	225	PASS	ZNF292	intronic	…...
L	6	87967636	rs6941356	A	G	150	PASS	ZNF292	exonic	
L	6	87969737	rs3734187	C	T	206	PASS	ZNF292	exonic	

L	6	87970301	rs3812132	C	G	217	PASS	ZNF292	exonic

Note:
The lines begin with 'Region' represent a ROH region. These lines give the chomosome ID, the start position of the ROH region, the end position, genes located in the region.
Other lines list SNVs located in the ROH region. Explanations of the header line is same as Table 4.13.

5.5 *De novo* mutation analysis

De novo mutation means an alteration in a gene that is present for the first time in one family member as a result of a mutation in a germ cell (egg or sperm) of one of the parents or in the fertilized egg itself. New mutations have long been known to cause genetic disease, but their true contribution to the disease burden can only now be determined using family-based whole-genome or whole-exome sequencing approaches.

Currently, there are several public softwares developed for *de novo* mutation detection, including SAMtools, GATK, DeNovoGear, FamSeq, and so on. Besides, *de novo* mutation can be obtained by simply screening variants that exist in child while not exist in parents (This method is referred as DenovoF here).

After evaluation of the mentioned methods, 1st BASE decides to provide *de novo* mutation resulted from SAMtools and DenovoF. In addition, Conrad D F *et al*. conclude that the union of these two methods will be the most complete result.

5.5.1 *De novo* mutation from SAMtools

SAMtools jointly analyze the BAM files of child and both parents and output *de novo* mutations. Then, variants are filtered to identify candidate mutations that may be associated with diseases. SNP and InDels are processed separately. The filtering steps are same as Part 5.1.

Table 5.4 The statistics of de novo SNVs from SAMtools

Total	1000G	Function	synonymous	deleterious
71	36	13	8	3

Notes:
Total: the total number of *de novo* SNVs in the family.
1000G: the number of variants survived step (1).
Function: the number of variants survived step (2).
Synonymous: the number of variants survived step (3).
Deleterious: the number of variants survived step (4).

5.5.2 *De novo* mutation from DenovoF

SNVs and InDels in each family member are detected by GATK. Then, *de novo* mutation can be obtained by simply screening variants that exist in child while not exist in parents.

Table 5.5 The statistics of de novo SNVs from DenovoF

Total	1000G	Function	synonymous	deleterious
44420	19825	83	62	12

Notes:
Total: the total number of *de novo* SNVs in the family.
1000G: the number of variants survived step (1).

Function: the number of variants survived step (2).
Synonymous: the number of variants survived step (3).
Deleterious: the number of variants survived step (4).

5.5.3 Annotation Result

We use ANNOVAR (Wang K *et al.*) to annotate *de novo* mutations, which includes annotation information from dbSNP, the 1000 Genomes Project and other published databases. Annotation contains the variation's position, type, conservation prediction, etc.

Table 5.6 The annotation results

Priority[1]	**CHROM[2]**	**POS[3]**	**ID[4]**	**REF[5]**	**ALT[6]**	**QAUL[7]**	**FILTER[8]**	**GeneName[9]**	**Func[10]**	**Gene[11]**	**GeneDetail[12]**	**Exonic Func[13]**	**AAChange[14]**	**[15-69]**
L	**1**	14653	.	C	T	841.77	PASS	WASH7P	ncRNA_exonic	.	.	.	.	…...
H	**1**	16631	.	T	C	541.77	PASS	WASH7P	ncRNA_exonic	.	.	.	.	…...

Note: Explanations of the header line is same as Table 4.13.

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. 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)
6. 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)
7. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes[J]. Nature, 2012, 491(7422):56-65. (1000g)
8. 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)
9. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource[J]. Nucleic acids research, 2004, 32(suppl 1):D258-D261. (GO)
10. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes[J]. Nucleic acids research, 2000, 28(1):27-30. (KEGG PATHWAY)
11. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet, 2013,Chapter 7:Unit7.20. (PolyPhen-2)
12. Augustine K, Frigge M L, Gisli M, et al. Rate of de novo mutations and the importance of father's age to disease risk.[J]. Nature, 2012, 488(7412):471-475.
13. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function[J]. Nucleic Acids Res. 2003, 1; 31(13):3812-4. (SIFT)
14. Georg B, Ehret, Patricia B, Munroe, Kenneth M, Rice, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk[J]. Nature, 2011, 478(7367):103-109.
15. Joshi P K, Esko T, Mattsson H, et al. Directional dominance on stature and cognition in diverse human populations[J]. Nature, 2015, 523(7561):459-462.
16. Keller M C, Simonson M A, Ripke S, et al. Runs of Homozygosity Implicate Autozygosity as a Schizophrenia Risk Factor[J]. Plos Genetics, 2012, 8(4):e1002656.
17. Teare M D, Barrett J H. Genetic linkage studies[J]. The Lancet, 2005, 366(9490): 1036-1044.
18. Abecasis G R, Cherny S S, Cookson W O, et al. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees[J]. Nature genetics, 2002, 30(1): 97-101.
19. Magi, A., L. Tattini, et al. (2014). "H3M2: detection of runs of homozygosity from whole-exome sequencing data." Bioinformatics 30(20): 2852-2859.
20. Kancheva, D., D. Atkinson, et al. (2015). "Novel mutations in genes causing hereditary spastic paraplegia and Charcot-Marie-Tooth neuropathy identified by an optimized protocol for homozygosity mapping based on whole-exome sequencing." Genetics in Medicine.
21. Low-pass Genomewide Sequencing and Variant Imputation Using Identity-by-descent in an Isolated Human Population.
22. Conrad D F, Keebler J E, Depristo M A, et al. Variation in genome-wide mutation rates within and between human families[J]. Nature Genetics, 2011, 43(7).

Appendix

Appendix A: Software List

The list of software during analysis

Analytical content	**Software**	**Notes**	**Version**
Quality control	In house		
Alignment	BWA	Map the sequencing reads to the reference genome	0.7.8-r455
	SAMtools	Sort bam	1.0
	Picard	Merge the bam file from the same sample and mark the duplicate reads	1.111
SNV/INDEL detection and annotation	GATK	Detection and filtering of SNPs and InDels	v3.1
	ANNOVAR	The annotation of the variant sites	2015Mar22

Appendix B: Verification Method of Sequencing

Verification Method of SNV/InDel

1.1 Sanger sequencing

Technical characteristics: Sanger sequencing is the most widely used method verifying the next generation sequencing in disease study, and has several defining features ,such as: high accuracy, shorter experimental period, only five days to get results from designing primer to sequencing. However, low throughput of this method makes it unable to be applied to large-scale studies. It can't accurately detect variations in repeated or high GC content sequences.
Technology application: mutation centralized and small-scale verification.
Reference:
Mutations in HFM1 in recessive primary ovarian insufficiency.The New England Journal of Medicine. 2014,370(10):972-974
Whole-genome sequencing of quartet families with autism spectrum disorder. Nature Medicine. 2015,21:185-191

1.2 SNaPshot

Technical characteristics: Short period of synthesizing primer and probe,speculating whether samples were polluted through checking the form of spectrums,high sensitivity and accuracy (>95%);although SNaPshot increased throughput, rigorous Tm value of primer is required. Besides, pre-experiment is needed to confirm the experimental conditions, time consuming and high demand for sample as with next generation sequencing.

Technology application: Large sample size and the number of variations greater than eight. (Kits are costly if the sample size is small).

Reference:

Whole exome sequencing in an India family links Coats plus syndrome and dextrocardia with a homozygous novel CTC1 and a rare HES7 variation. BMC Medical Genetics. 2015,16:5.DOI 10.1186/s12881-015-0151-8

1.3 MassArray

Technical characteristics: High throughput; fast, only needing a few seconds to detect a reaction-hole which could conduct forty reactions; one chip able to complete multiple iPLEX GOLD experiment with 384 samples; primer is 80bp with no need for fluorescence labeling. This method is cheap for the large sample size and multiple variations. However, it has some shortcomings, such as time-consuming for determining experimental conditions especially with confirming the concentration of primer; kit is costly for single-trial.

Technology application: Sample size larger than 500 and number of SNVs around 25 fold.

Reference:

A large-scale screen for coding variants predisposing to psoriasis.Nature Genetics.2014,46(1):45-50

Rare Variants in FBN1 are associated with severe adolescent idiopathic scoliosis. Human Molecular Genetics. 2014,23(19):5271-5282.