Human Whole Genome Sequencing

Project Demo Report

(Disease)

May 1, 2016

Contents

1 Sample Information

Table 1.1 Sample information

FamilyID	Sample ID	Gender	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.

A total amount of 1.0ug DNA per sample was used as input material for the DNA sample preparations. Sequencing libraries were generated using Truseq Nano DNA HT Sample preparation Kit (Illumina USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. The genomic DNA is randomly fragmented to a size of 350bp by Covaris cracker, then DNA fragments were end polished, A-tailed, and ligated with the full-length adapter for Illumina sequencing with further PCR amplification. At last, PCR products were purified (AMPure XP system) and libraries were analyzed for size distribution by Agilent2100 Bioanalyzer and quantified using real-time PCR.

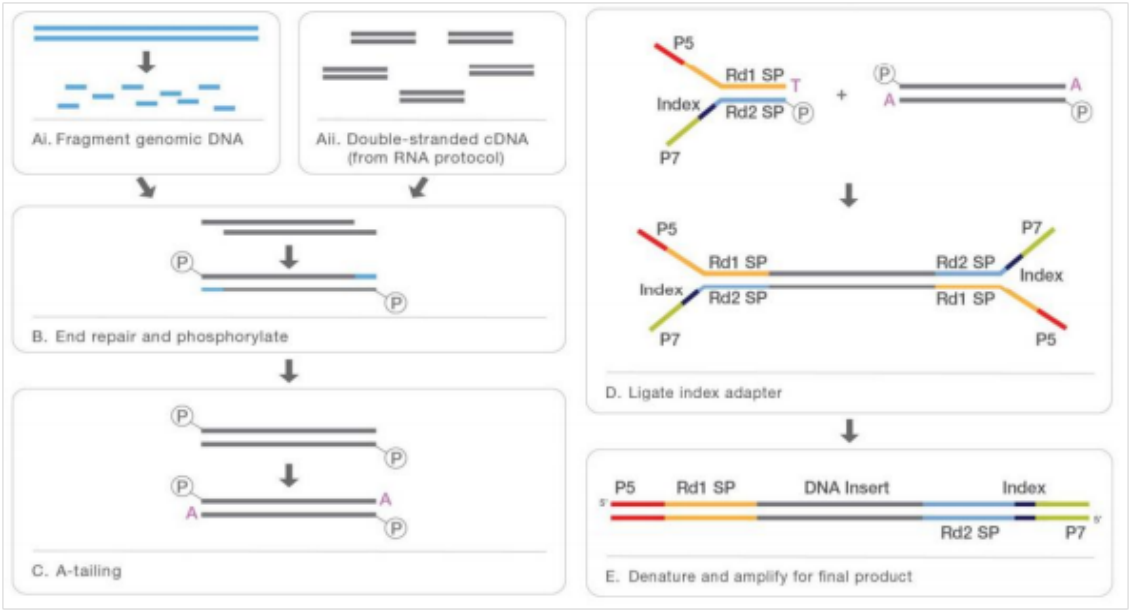

Figure 2.1 Library construction workflow

2.2 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:

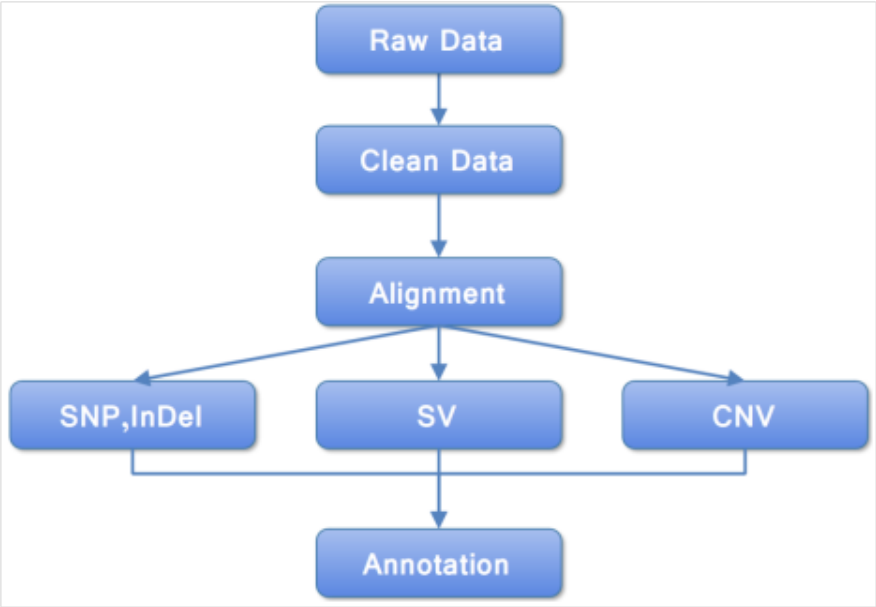

Figure 3.1 Bioinformatics analysis pipeline

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 No.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

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 TruSeqTM DNA Sample Prep Kits) information:

5' Adapter: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

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

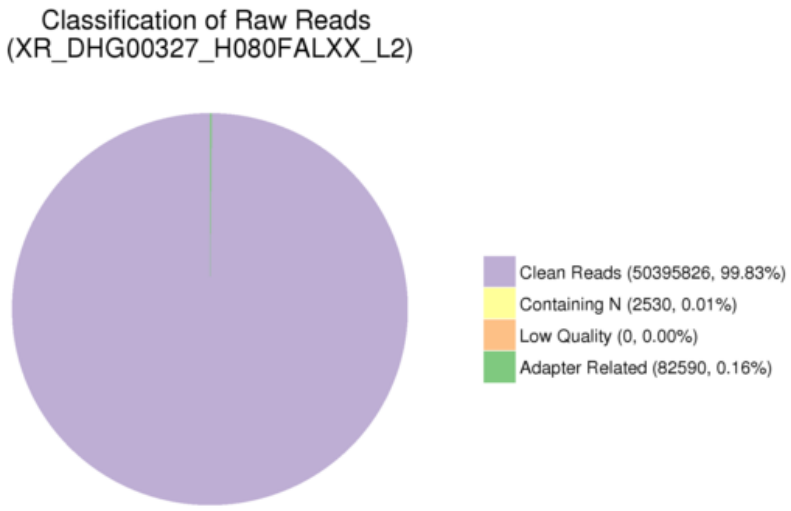

Figure 4.1 Raw data filtration result

Note: Reads were discarded in pairs.
(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.

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 NGS, 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.

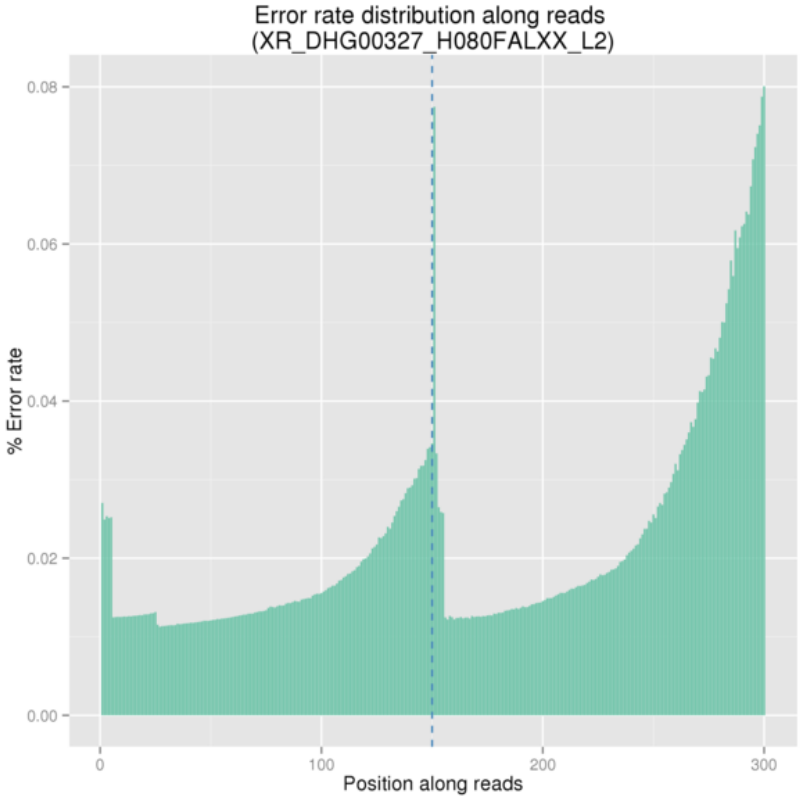

Figure 4.2 Sequencing error rate distribution

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

4.2.3 GC Content Distribution

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

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

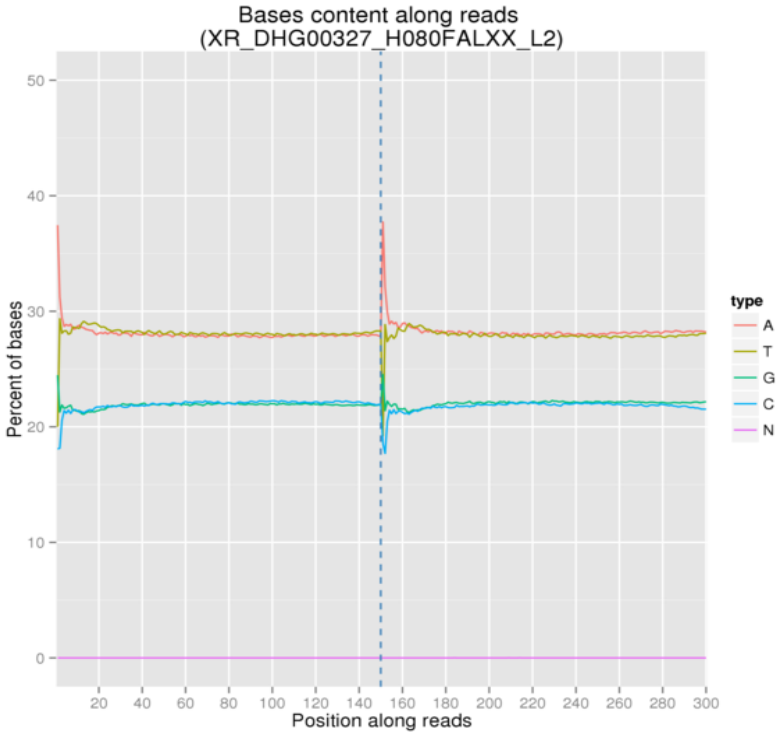

Figure 4.3 GC content distribution

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

4.2.4 Sequencing Quality Distribution

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

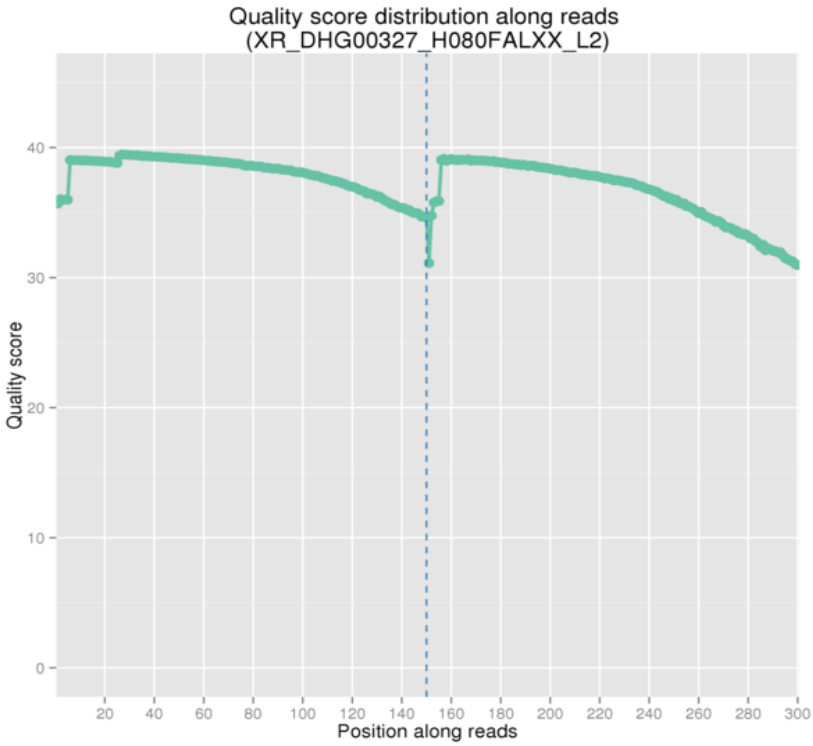

Figure 4.4 Data quality distribution

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

4.2.5 Statistics Summary of Sequencing Quality

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

Table 4.4 Overview of data production quality

Sample name[1]	1st BASE ID[2]	Lane[3]	Raw reads[4]	Rawdata (G)[5]	Raw Depth (X)[6]	Effective (%)[7]	Error (%)[8]	Q20 (%)[9]	Q30 (%)[10]	GC (%)[11]
TEST-1	DHG00321	H04HLALXX_L1	310557209	93.17	32.09	99.10	0.03	95.99	90.32	43.44
TEST-2	DHG00320	H04HLALXX_L2	333468550	100.04	34.62	99.67	0.025	96.03	90.96	43.74
TEST-3	DHG00319	H04HLALXX_L3	278682329	96.05	33.0	99.41	0.03	95.34	89.14	43.94
XR	DHG00327	H04HLALXX_L1-L9	530689772	159.13	50.8	99.63	0.035	95.06	88.50	42.34

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 (UCSC hg19 (Kent W J *et al*.)). The original mapping result in BAM format can be obtained. SAMtools (Li H *et al*.) is used for sorting the BAM file, and

Picard is utilized 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 10× read depth are more robust and authentic.

4.3.1 Sequencing Depth, Coverage Distribution

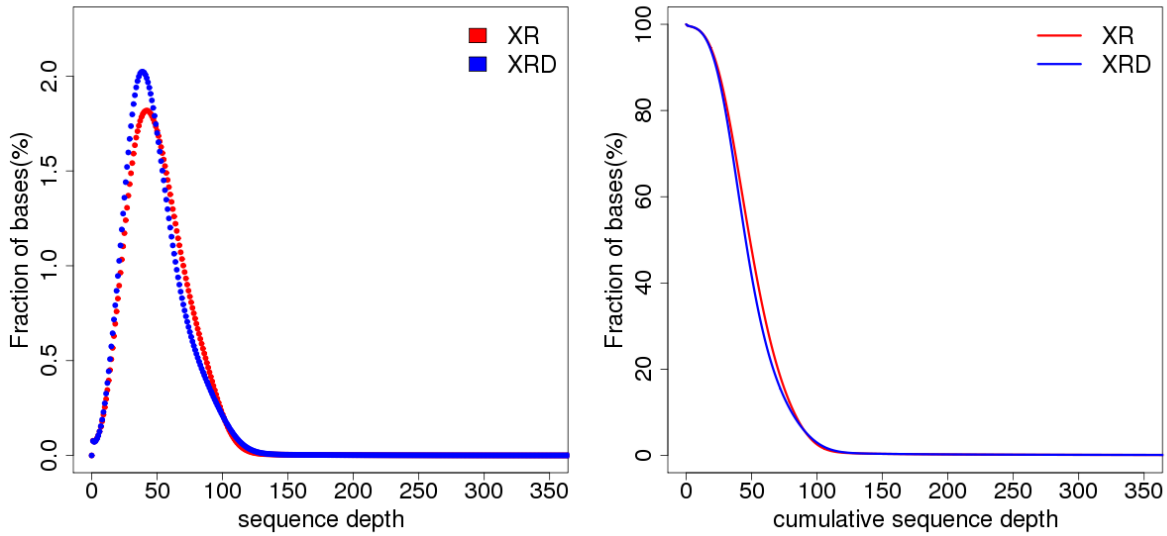

Figure 4.5 Sequencing depth

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

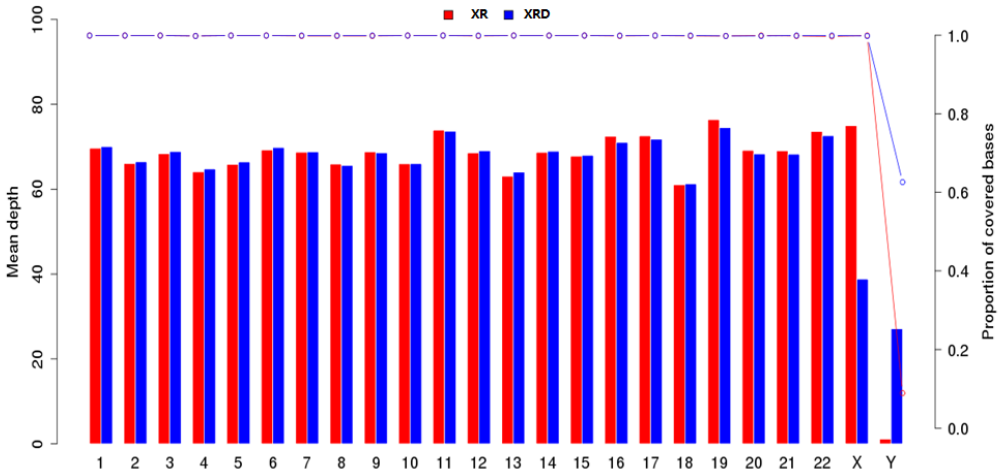

Figure 4.6 The coverage depth (the left coordinate) and coverage rate (the right coordinate) of chromosome

The x-axis is chromosome number; the left y-axis is the average depth of each chromosome (Raw data/length_of_chromosome); the right y-axis is the fraction of covered on each chromosome (The number of bases covered by reads/total number of bases).

4.3.2 Statistics of Coverage

Table 4.5 Mapping rate and coverage

Sample:	XR
Total:[1]	1054972854 (100%)
Duplicate:[2]	113076597 (10.73%)
Mapped:[3]	1054237028 (99.93%)
Properly mapped:[4]	1023490016 (97.02%)
PE mapped:[5]	1053812160 (99.89%)
SE mapped:[6]	849736 (0.08%)
With mate mapped to a different chr:[7]	11157288 (1.06%)
With mate mapped to a different chr ((mapQ>=5)):[8]	8711686 (0.83%)
Average_sequencing_depth:[9]	52.64
Coverage:[10]	99.67%
Coverage_at_least_4x:[11]	99.44%
Coverage_at_least_10x:[12]	98.58%
Coverage_at_least_20x:[13]	93.69%

Note:
(1) Total: The total number of clean reads
(2) Duplicate: The number of duplication 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
(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
(7) With mate mapped to a different chr: The number of mate reads that mapped to the different chromosomes (percentage)
(8) With mate mapped to a different chr (mapQ>=5): The number of mate reads that mapped to the different chromosomes and the MAQ >5
(9) Average_sequencing_depth: The average sequencing depth that mapped to the reference genome
(10) Coverage: The sequence coverage of genome
(11) Coverage_at_least_4X: The percentage of bases with depth >4× in whole genome bases
(12) Coverage_at_least_10X: The percentage of bases with depth >10× in whole genome bases
(13) Coverage_at_least_20X: The percentage of bases with depth >20× in whole genome bases

4.4 Variation Detection Result

4.4.1 SNV Detection Result

Generally, the whole genome of human has about 3.6M SNV. Most (above 95%) SNVs with high frequency (the allele frequency in population is above 5%) have records in dbSNP(Sherry S T *et al.*) .The ration of Ts/Tv can reflect the accuracy of sequencing. Generally, the ratio in genome is about 2.2 and in coding region is about 3.2.
We use GATK to detect SNV, and the statistics of SNVs are as follows:

Table 4.6 The number of SNV in different genomic region

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]
XR	19533	1085383	21665	4660	1793788	8018	123419	18606	19113	2374	222

Note:
(1) Sample: Sample name
(2) exonic: The number of SNV in exonic region
(3) intronic: The number of SNV in intronic region
(4) UTR3: The number of SNV in 3'UTR region
(5) UTR5: The number of SNV in 5'UTR region
(6) intergnic: The number of SNV in intergenic region

(7) ncRNA_exonic: The number of SNV in non-coding RNA exonic region
(8) ncRNA_intronic: The number of SNV in non-coding RNA intronic region
(9) upstream: The number of SNV in the 1kb upstream region of transcription start site
(10) downstream: The number of SNV in the 1kb downstream region of transcription ending site
(11) splicing: The number of SNV in 4bp splicing junction region
(12) ncRNA_splicing: The number of SNV in 4bp splicing junction of non-coding RNA

Table 4.7 The number of of different SNVs types in coding region

Sample[1]	synonymous_SNV[2]	missense_SNV[3]	Stopgain[4]	Stoploss[5]	Unknown[6]
XR	10227	8965	47	12	282

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 cause an amino acid change
(4) stopgain: A nonsynonymous SNV that lead to the immediate creation of stop codon at the variant site
(5) stoploss: A nonsynonymous SNV that lead to the immediate elimination of stop codon at the variant site
(6) unknown: Unknown function (due to various errors in the gene structure definition in the database file)

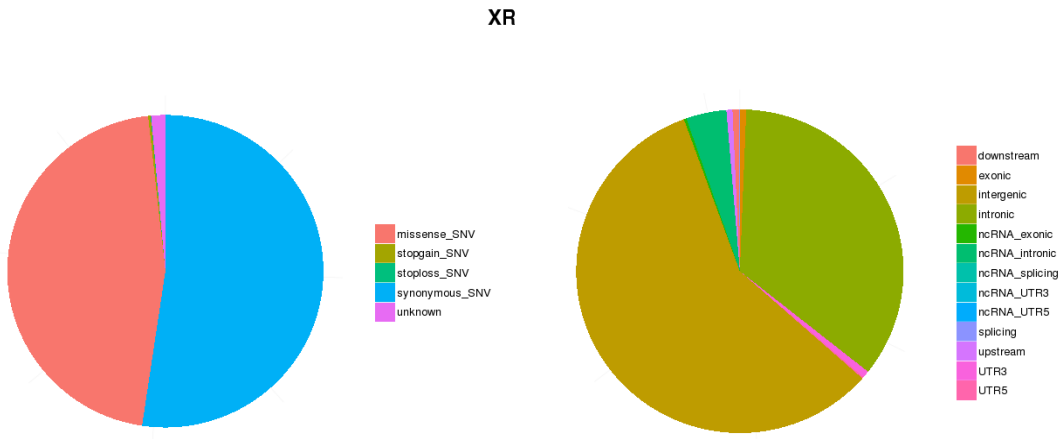

Figure 4.7 The number of of different SNVs types in coding region (left); the number of SNVs in different genomic region (right).

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 means a nonsynonymous SNV that lead to the immediate creation of stop codon at the variant site; stoploss 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, downstream 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 4bp 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 4bp 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, the ratio in genome is about 2.2, in coding region is about 3.2.

Table 4.8 Transition and transversion distribution

Sample[1]	novel_ts[2]	novel_ts/tv[3]	novel_tv[4]	ts[5]	ts/tv[6]	tv[7]
XR	31255	1.992795	15684	2101696	2.112183	995035

Note:
(1) Sample: Sample name
(2) novel_ts: The number of ts SNP that are not in dbSNP
(3) novel_ts/tv: Is calculated as novel ts/ novel tv

(4) novel_tv: The number of tv SNP that are not in dbSNP
(5) transition(ts): Transition
(6) ts/tv: Is calculated as the number of transition/ the number of transversion
(7) transversion(tv): Transversion

Table 4.9 SNV and genotype distribution

Sample[1]	all[2]	genotype.Het[3]	genotype.Hom[4]	novel[5]	novel_proportion[6]
XR	3096781	1838868	1257913	47088	0.015205519

Note:
(1) Sample: Sample name
(2) all: The total number of SNV
(3) genotype.Het: The genotype of heterozygote
(4) genotype.Hom: The genotype of homozygote
(5) novel: SNV not in dbSNP
(6) novel_proportion: Is calculated as novel SNV/total number of SNV

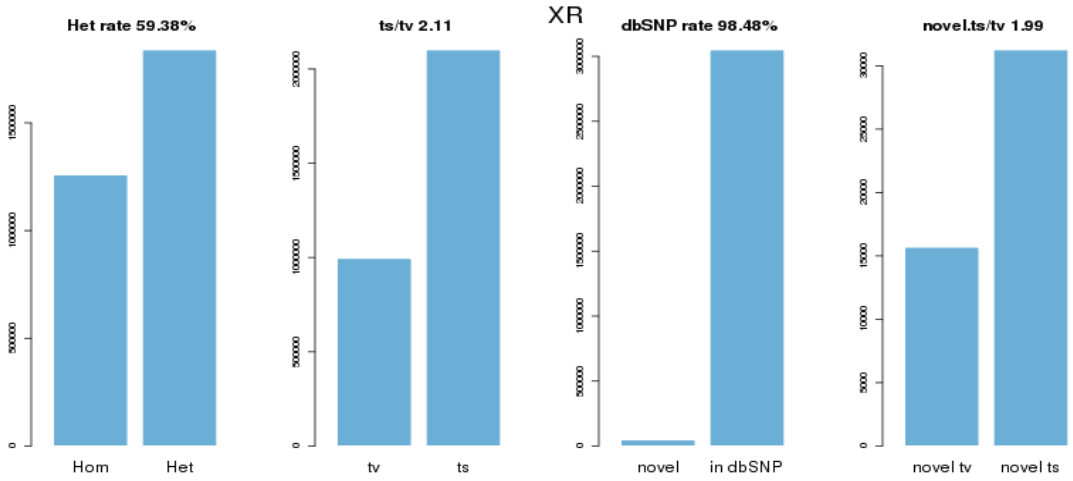

Figure 4.8 Feature of SNV in genome.

In the first figure, Hom is short for homozygous and Het for heterozygous; the y-axis means the number of SNVs with homozygous and heterozygous genotype.
In the second figure, tv means transversion and ts means transition; the y-axis means the number of SNVs.
In the third figure, novel means SNVs not in dbSNP; in dbSNP means SNVs in dbSNP. dbSNP rate is calculated as the number of SNV 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 Detection Result

Generally, the genome of human has about 350K InDel (insertion and deletion, insertion and deletion less than 50bp).
The InDel in coding region or splicing site may change the protein translation. Frameshift mutation, in which the number of inserted or deleted bases is not an integral multiple of three, may lead to the change of the whole reading frame. Compared to non-frameshift mutation, frameshift mutation is more limited by selective pressure.The length distribution of InDel in figure 4.10 shows this phenomenon.
We use GATK to detect InDel, and the obtained InDel result is as follows:

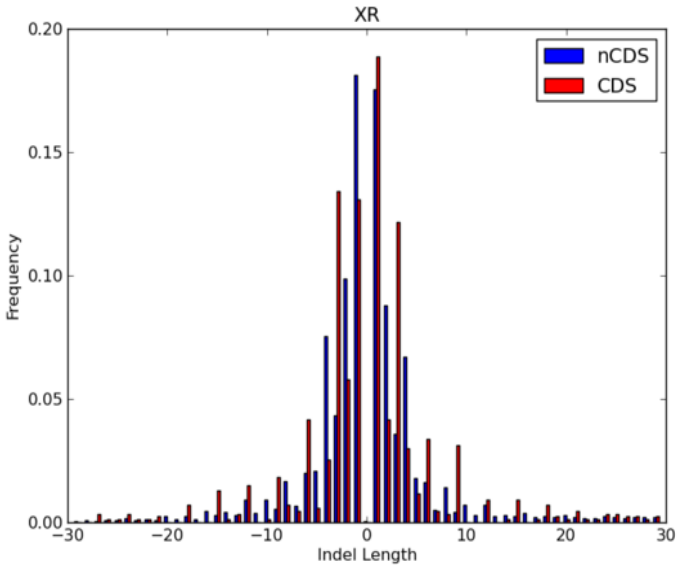

Figure 4.9 InDel length distribution.

The x-axis is indel length and y-axis is frequency. CDS means coding region and splicing site; 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 InDel 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]
XR	584	165935	3958	754	261796	898	18679	3767	3527	352	33

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 4bp splicing junction region
(12) ncRNA_splicing: The number of InDel in 4bp splicing junction of non-coding RNA

Table 4.11 The number of differente types of InDels in coding regions

Sample[1]	frameshift_deletion[2]	frameshift_insertion[3]	nonframeshift_deletion[4]	nonframeshift_insertion[5]	stoploss[6]	stopgain[7]	Unknown[8]
XR	67	56	179	164	0	7	89

Note:
(1) Sample: Sample name
(2) frameshift_deletion: A deletion of one or more nucleotides that cause frameshift changes in protein coding sequence, the deletion length is not multiple of 3
(3) frameshift_insertion: An insertion of one or more nucleotides that cause frameshift changes in protein coding sequence, the insertion length is not multiple of 3
(4) nonframeshift_deletion: Non-frameshift deletion, does not change coding protein frame deletion, the deletion length is multiple of 3
(5) nonframeshift_insertion: Non-frameshift insertion, does not change coding protein frame insertion: the insertion length is multiple of 3
(6) stoploss: Frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate elimination of stop codon at the variant site
(7) stopgain: Frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate creation of stop codon at the variant site
(8) unknown: Unknown function (due to various errors in the gene structure definition in the database file)

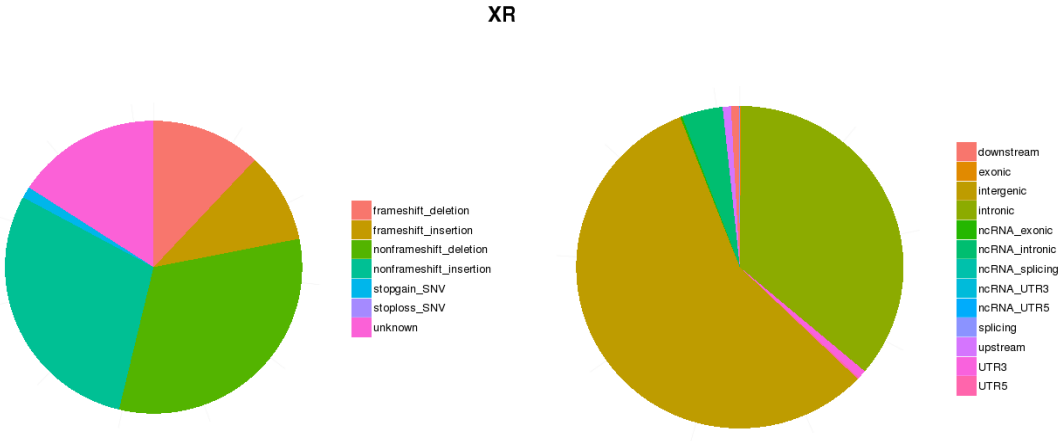

Figure 4.10 The number of of different type InDels in coding regions (left); the number of InDels in different genomic regions (right).

Table 4.12 InDel and genotype distribution

Sample[1]	all[2]	genotype.Het[3]	genotype.Hom[4]	novel[5]	novel_proportion[6]
XR	460257	235404	224853	317575	0.689995

Note:
(1) Sample: Sample name
(2) all: The total number of InDel
(3) genotype.Het: The genotype of heterozygote
(4) genotype.Hom: The genotype of homozygote
(5) novel: InDel not in dbSNP
(6) novel_proportion: Is calculated as novel Indel/total number of Indel

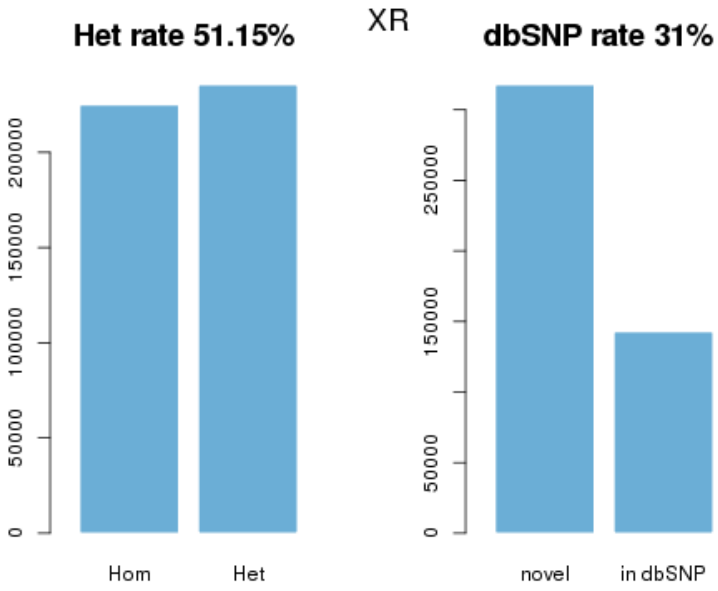

Figure 4.11 Feature of InDels in genome.

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.

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—1st BASE 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 0-1000; 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.

4.4.3 SV Result

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

Table 4.14 SV detection result summary

Sample[1]	varType[2]	total[3]	cds[4]	splicing[5]	utr5[6]	utr3[7]	Intron[8]	upstream[9]	downstream[10]	ncRNA[11]	intergenic[12]	Unknown[13]
XR	Deletion	3239	90	0	3	11	1090	26	30	128	1861	0
	Translocation	602	10	0	3	7	204	3	6	33	336	0
	Inversion	228	78	0	0	0	41	1	0	26	82	0
	Insertion	101	5	0	1	2	35	1	2	2	53	0

Note:
(1) Sample: Sample name
(2) varType: SV Type
(3) total: The total number of SV
(4) cds: The number of SV in CDS region
(5) splicing: The number of SV in splicing junction region
(6) utr5: The number of SV in 5'UTR region
(7) utr3: The number of SV in 3'UTR region
(8) intron: The number of SV in intronic region
(9) upstream: The number of SV in 1Kb region upstream from transcription start site
(10) downstream: The number of SV in 1Kb region downstream from transcription end site
(11) ncRNA: The number of SV in ncRNA region
(12) intergnic: The number of SV in intergenic region
(13) unknown: The number of SV in region with unknown function (due to various errors in the gene structure definition in the database file

Table 4.15 SV annotation result

Chr[1]	Start[2]	End[3]	Ref[4]	Alt[5]	…….	Size[32]	Support[33]	SupportPerID[34]	…….	SVID[38]	SVType[39]
1	869442	870240	0	0	……	855	9	XR,23	…….	1	Deletion
1	869442	869442	0	0	……	855	9	XR,23	…….	1	Breakpoint
1	870240	870240	0	0	……	855	9	XR,23	…….	1	Breakpoint
1	1162752	1162838	0	0	…….	200	7	XR,11	…….	2	Deletion

Note: Annotation information includs four parts: Annotation of Genes and genome regions (1-18), Database annotation (19-22), ENCODE Annotation (23-28), Structuralvariation information (29~39).

The first item are annotation of Genes and genome regions—Genes where the variation locates on may relate to disease. 1st BASE proceeds the annotation of known gene structure and genome regeions related to the varition.

(1) Chr: Chromosome
(2) Start: The start position of variation on chromosome
(3) End: The end position of variation on chromosome
(4) Ref: Base at this position in reference genome
(5) Alt: Base at this position in sequencing data
(6) GeneName: The names of genes related to this variation
(7) Func: Tells whether the variant hit exons or hit intergenic regions, or hit introns, or hit a non-coding RNA genes.
(8) Gene: The IDs of transcript whose function has changed like the value in column 'Func'.
(9) GeneDetail: Description of variations in UTR, splicing, ncRNA, splicing or intergenic.
(10) ExonicFunc: the amino acid changes as a result of the exonic variant.(synonymous_SNV, missense_SNV, stopgain_SNV, stopgloss_SNV or unknown)
(11) AAChange: when 'Func' equals 'exonic'or 'exonic; splicing', this value gives the change of amino acid in each related transcript. For

example, AIM1L:NM_001039775:exon2:c.C2768T:p.P923L, AIM1L is gene name containing this variation; NM_001039775 is ID of transcript; exon2 means the variation is on the second exon of the transcript; c.C2768T means the 2,768 base on cDNA is changed from C to T due to this variation; p.P923L means the 923 amino acid on protein is changed from Pro to Leu due to this variation

(12) Gencode: Gene name in Gencode
(13) cpgIslandExt: The result of prediction of CpG islands
(14) cytoband: Chromosome band
(15) wgRna: snoRNA and miRNA annotation
(16) targetScanS: miRNA target prediction by TargetScan
(17) phastConsElements46way: The conservative region predicted by phastCons basing on the whole genome alignment of vertebrates; 46way means the number of used species
(18) tfbsConsSites: Transcript factor binding site that are conservative in human, mouse and rat; this is acquired from transfac matrix database (v7.0).

The second item is database annotation—1st BASE proceeds the database annotation of the structural variation in order to inquiry if the variation was known or related to some kind of disease.

(19) genomicSuperDups: to detect if the varitation is located in the segmental duplication, most variations called on the segmental duplication result from the fault alignment, which may be false positive, so those variation results should be treated carefully. such as, Score=0.994828;Name=chr19:60000, means the segment of chr19:60000 is similar to the region where the variation locates on, the sequence uniformity is 0.994828
(20) dgvMerged: Annotation from Database of Genomic Variants
(21) gwasCatalog: Tells whether this variation has been identified by published Genome-Wide Association Studies (GWAS), collected in the Catalog of Published Genome-Wide Association Studies at the National Human Genome Research Institute (NHGRI). It lists the diseases related to this variation. "." means this variation has not been reported by published GWAS study.
(22) Repeat: the repeat sequence annotated from RepeatMasker. Such as, Score=180; Name='1385: (CACCC)n (Simple_repeat)', '(CACCC)n' is the name of the repeat, 'Simple' means the type of the repeat. This annotation means the variation is located on the repeat region where it is easy to align falsely. So the variation on the repeat region is not very reliable

The third item is ENCODE Annotation—Functional annotation based on the ENCODE genome region aims to show the functional elements of the genome.

(23) encodeGm12878: Predicted functional elements in the genome of cell line Gm12878. ChromHMM was used to get twenty-five states by integrating ENCODE ChIP-seq, DNase-seq, and FAIRE-seq data; these states were used to segment the genome, and they were then grouped and colored to highlight predicted functional elements. (The relationship between state and functional element is as follow: Tss, TssF——Active Promoter; PromF——Promoter Flanking; PromP——Inactive Promoter; Enh, EnhF——Candidate Strong enhancer; EnhWF, EnhW, DNaseU, DNaseD, FaireW——Candidate Weak enhancer/DNase; CtrcfO, Ctcf——Distal CTCF/Candidate Insulator; Gen5', Elon, ElonW, Gen3', Pol2, H4K20——Transcription associated; Low——Low activity proximal to active states; ReprD, Repr, ReprW——Polycob repressed; Quies, Art——Heterochromatin/Repetitive/Copy Number Variation)
(24) encodeH1hesc: Predicted functional elements in the genome of cell line H1-hESC
(25) encodeHelas3: Predicted functional elements in the genome of cell line HeLa-S3
(26) encodeHepg2: Predicted functional elements in the genome of cell line HepG2
(27) encodeHuvec: Predicted functional elements in the genome of cell line HUVEC
(28) encodeK562: Predicted functional elements in the genome of cell line K562

The fourth item is structuralvariation information——Detailed information of structural variation including variation type, coverage etc.

(29) Orientation1: A string that records the number of reads mapped to the plus (+) or the minus (-) strand in the anchoring regions. For breakpoint 1
(30) Orientation2: A string that records the number of reads mapped to the plus (+) or the minus (-) strand in the anchoring regions. For breakpoint 2
(31) Score: The score of SV
(32) Size: The length of SV
(33) Support: The number of reads supporting this SV
(34) SupportPerID: The number of reads supporting this SV in each sample
(35) TX: For translocation, there are two types-CTX (Inter-chromosomal translocations) and ITX (Intra-chromosomal translocations).
(36) TCHR: For translocation, the chromosome on which the other breakpoint belong.
(37) TSTART: For translocation, the positon of the other breakpoint.
(38) SVID: The ID of SV
(39) SVType: Breakpoint, translocation, deletion, insersion, inversion

4.4.4 CNV Result

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

Table 4.16 CNV detection result

Sample[1]	varType[2]	Total[3]	cds[4]	Splicing[5]	utr5[6]	utr3[7]	Intron[8]	Upstream[9]	Downstream[10]	ncRNA[11]	Intergenic[12]	Unknown[13]
XR	gain	901	139	0	3	5	54	5	7	130	558	0
	loss	2419	62	0	2	6	638	25	20	115	1551	0

Note:
(1) Sample: Sample name
(2) varType: CNV Type
(3) total: The total number of CNV
(4) cds: The number of CNV in CDS region
(5) splicing: The number of CNV in splicing junction region
(6) utr5: The number of CNV in 5'UTR region
(7) utr3: The number of CNV in 3'UTR region
(8) intronic: The number of CNV in intronic region
(9) upstream: The number of CNV in 1Kb region upstream from transcription start site
(10) downstream: The number of CNV in 1Kb region downstream from transcription end site
(11) ncRNA: The number of CNV in ncRNA region
(12) intergnic: The number of CNV in intergenic region
(13) unknown: The number of CNV in region with unknown function (due to various errors in the gene structure definition in the database file)

Table 4.17 CNV annotation result

Chr[1]	Start[2]	End[3]	Ref[4]	Alt[5]	GeneName[6]	Func[7]	Gene[8]	…..	CNVType[32]
1	0	16254500	0	0	CLSTN1,PRKCZ	exonic	NM_000302,NM_000815	.	loss
1	16254500	16254500	0	0	.	intergenic	NONE,NR_046018	.	breakpoint

Note: Annotation information includes four parts: Annotation of Genes and genome regions(1-18), Database annotation(19-22), ENCODE Annotation(23-28), Structural variation information(29~32).

The first item are annotation of Genes and genome regions—Genes where the variation locates on may relate to disease. 1st BASE proceeds the annotation of known gene structure and genome regions related to the variation.

(1) Chr**:** Chromosome
(2) Start: The start position of variation on chromosome
(3) End: The end position of variation on chromosome
(4) Ref: Base at this position in reference genome
(5) Alt: Base at this position in sequencing data
(6) GeneName: The names of genes related to this variation
(7) Func: Tells whether the variant hit exons or hit intergenic regions, or hit introns, or hit a non-coding RNA genes.
(8) Gene: The IDs of transcript whose function has changed like the value in column 'Func'.
(9) GeneDetail: description of variations in UTR, splicing, ncRNA, splicing or intergenic. When the Func column is exonic, ncRNA_exonic, intronic, ncRNA_intronic, upstream, downstream, upstream; downstream, ncRNA_UTR3, ncRNA_UTR5, this colum is blank; When the Func column is intergenic, this column is e.g.dist=1366; dist=22344, representant the distance between two genes aside the CNV.
(10) ExonicFunc: variation type of exonic variant (synonymous_SNV, missense_SNV, stopgain_SNV, stopgloss_SNV or unknown)
(11) AAChange: change of the base and amino acie; if the CNV exists in multiple transcripts, every transcript will be annotated; the format is e.g. OR4F5:NM_001005484:exon1:c.A421G:p.T141A, represents gene name: transcript ID of RefSeq: exonic structure:base variation in cDNA:amino acid change in the protein.
(12) Gencode: Gene name in Gencode
(13) cpgIslandExt: the predicting result of CpG islands, the result is CpG island name, such as CpG:116, 116 refers to the number of CG in the CpG island
(14) cytoband: the chromesome band where the CNV locates (obserbed by Giemas dyeing), if the variation locates in multiple regions, a dash is used for ligation
(15) wgRna: Annotated with microRNA and snoRNA related to the CNV based on miRBase and snoRNABase, show the gene name of microRNA and snoRNA
(16) targetScanS: miRNA target prediction by TargetScan, such as, Score=62;Name=KRAS:miR-181:1, means the score of this microRNA target is 62, located on the 3'UTR of KRAS gene, the microRNA affected by the variation is miR-181:1
(17) phastConsElements46way: The conservative region predicted by phastCons basing on the whole genome alignment of vertebrates; 46way means the number of used species
(18) tfbsConsSites: Transcript factor binding site that are conservative in human, mouse and rat; this is acquired from transfac matrix database (v7.0)

The second item is database annotation—1st BASE proceeds the database annotation of the structural variation in order to inquiry if the variation was known or related to some kind of disease.

(19) genomicSuperDups: to detect if the varitation is located in the segmental duplication, most variations called on the segmental duplication

result from the fault alignment, which may be false positive, so those variation results should be treated carefully. such as, Score=0.994828; Name=chr19:60000, means the segment of chr19:60000 is similar to the region where the variation locates on, the sequence uniformity is 0.994828

(20) dgvMerged: annotation from Database of Genomic Variants to show the reported SV where the variation locate

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

(22) Repeat:the annotation of repeat sequence annotated from RepeatMasker. such as, Score=180; Name="1385: (CACCC)n(Simple_repeat)", (CACCC)n is the name of the repeat, Simple means the type of the repeat. This annotation means the variation is located on the repeat region where it is easy to align falsely. So the variation on the repeat region is not very reliable

The third item is ENCODE Annotation—Functional annotation based on the ENCODE genome region aims to show the functional elements of the genome.

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

(24) encodeH1hesc: Predicted functional elements in the genome of cell line H1-hESC

(25) encodeHelas3: Predicted functional elements in the genome of cell line HeLa-S3

(26) encodeHepg2: Predicted functional elements in the genome of cell line HepG2

(27) encodeHuvec: Predicted functional elements in the genome of cell line HUVEC

(28) encodeK562: Predicted functional elements in the genome of cell line K562

The fourth item is structuralvariation information—Detailed information of structural variation including variation type, coverage etc.

(29) CopyNumber: Copy number

(30) Size: The size of the CNV region

(31) CNVID: The ID of CNV

(32) CNVType: The type of CNV
loss: Reduction of copy number
gain: Increase of copy number
breakpoint: The breakpoint of CNV

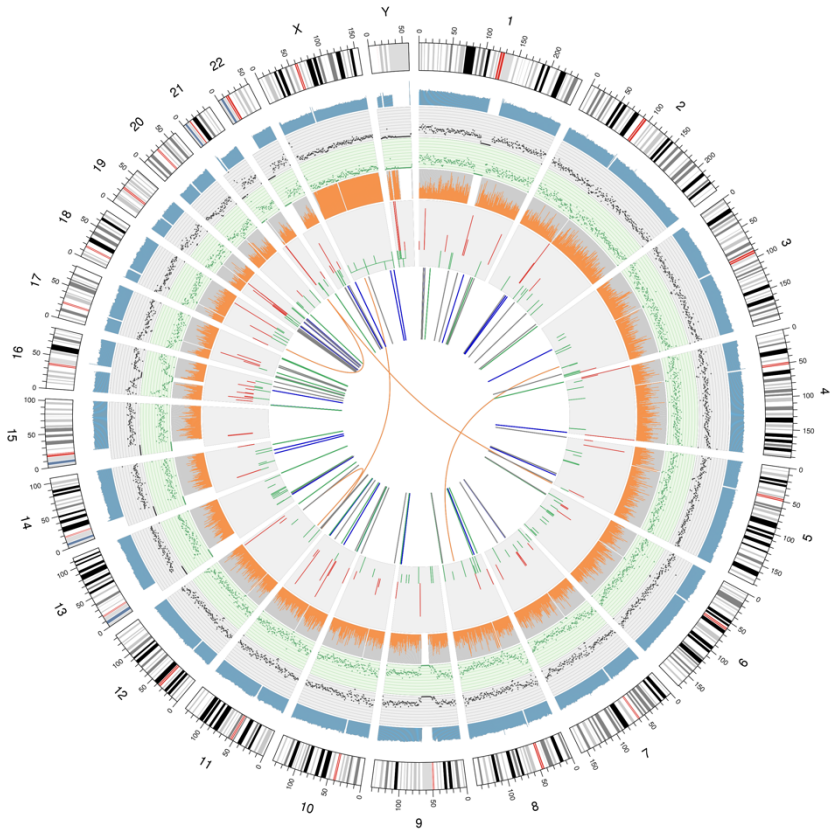

Figure 4.12 Circos

Note: 1st BASE shows Circos only when CNV analysis was carried out. The figure consists seven rings from outer to inner.
(1) The outer concentric ring is chromosomal information.
(2) The second ring represents the read coverage in histogram style. A histogram is the average coverage of a 0.5Mbp region.
(3) The third ring represents indel density in scatter style. A black dot is calculated as indel number in a range of 1Mbp).
(4) The fourth ring represents snp density in scatter style. A green dot is calculated as snp number in a range of 1Mbp).
(5) The fifth ring represents the proportion of homozygous SNP (orange) and heterozygous SNP (grey) in histogram style. A histogram is calculated from a 1Mbp region.
(6) The sixth ring represents the CNV inference. Red means gain, and green means loss.
(7) The most central ring represents the SV inference in exonic and splicing regions. CTX (orange), INS (green), DEL (grey), ITX (pink) and INV (blue).

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.

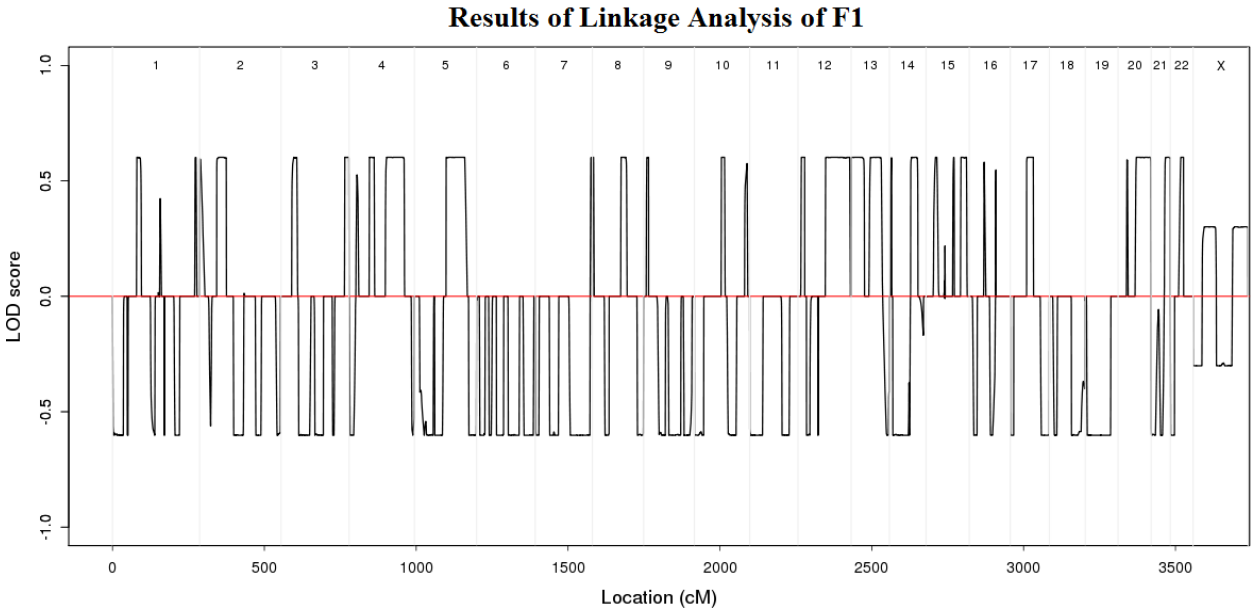

Figure 5.1 Linkage analysis for family F1

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]	GeneName[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]	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: 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. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools [J]. Bioinformatics, 2009, 25(16): 2078-2079. (SAMtools)
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. Nucleic Acids Res. 2003, 1; 31(13):3812-4. (SIFT)
14. Chen K, Wallis J W, McLellan M D, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation [J]. Nature methods, 2009, 6 (9): 677-681. (BreakDancer)
15. Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data [J]. Bioinformatics, 2012, 28(3): 423-425. (Control-FREEC)
16. 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.
17. 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.
18. 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.
19. Teare M D, Barrett J H. Genetic linkage studies[J]. The Lancet, 2005, 366(9490): 1036-1044.
20. 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.
21. Magi, A., L. Tattini, et al. (2014). "H3M2: detection of runs of homozygosity from whole-exome sequencing data." Bioinformatics 30(20): 2852-2859.
22. 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.
23. Low-pass Genomewide Sequencing and Variant Imputation Using Identity-by-descent in an Isolated Human Population.
24. 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 softwares

Analytical	Software	Notes	Version
Quality control	In house		
Alignment	BWA	Map the sequencing reads to the reference genome and the BAM file was obtained	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
SNP/INDEL detection	GATK	Detect and filter SNP, InDel	1.0
	ANNOVAR	Annotate variation site	2015Mar22
SV	breakDancer	Detect SV	1.4.4
	ANNOVAR	Annotate variation site	2015Mar22
CNV	control-FREEC	Detect CNV	V6.7
	ANNOVAR	Annotate variation site	2015Mar22

Appendix B: Verification Method of Sequencing

Verification Method of SNV/Indel

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 make it unable to achieve batch quantity verification. And chances of success is low for the variation in repeated sequence and high GC content sequences.

Technology application: Programme of concentrated and small-scale variation.

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

2. SNaPshot

Technical characteristics: Short period of synthesising 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: Programme of large sample size and the number of variations greater than eight is better. (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/s12881-015-015

3. MassArray

Technical characteristics: High throughput, it only need a few seconds to detecting a reaction-hole which could conduct forty reactions, multiple iPLEX GOLD experiment with 384 samples was able to complete in one chip, 80 bp of primer and with no need for fluorescence labeling, this method is cheaper for the programme with large sample size and multiple variations. However, it has some shortcomings, such as time-consuming for testing experimental conditions especially with confirming the concentration of primer; kit is costly for single-trial.
Technology application: Programme for sample size larger than 500 and number of SNV around 25 fold.
Reference
A large-scale screen for coading 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.

Verification Method of CNV

Droplet Digital PCR
Technical characteristics: high sensitivity makes it can detect mutations with a frequency lower than 0.001%, and high-specificity makes it can detect target sequences with complex backgrounds; absolute quantitative without normal curve, less demanding for sample amount and detecting nucleinic acid in rare sample accurately. Several disadvantages of it are unable to achieve batch quantity verification and include primer with repeated and high GC content sequences, time-consuming for testing experimental conditions, requiring custom-made probe, expensive and long cycle, only a handful of companies undertake those experiments.
Technology application: programme with small sample size and less variatonss.
Reference:
Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and

obesity.Natural Genetics.2015,doi:10.1038/ng.3340

Verification Method of SV

1. PCR amplification combined with Sanger sequencing

Technical characteristics: Good for inversion, translocation and deletion, high accuracy makes it could confirm the position of breakpoint; cheaper for single validation test and short experimental cycle. But it has several disadvantages: low throughput, one breakpoint can be detected at a time, multiple experimental and primer design, poor stability, chances of success is low for the variation in repeated sequence and high GC content sequences, rigorous for initial dose of genomic DNA and purity.

Technology application: Small-scale programme.

Reference:

Precise detection of chromosomal translocation or inversion breakpoints by whole-genome sequencing. Journal of Human Genetics.2014, 1-6

Breakpoint mapping by next generation sequencing reveals causative gene disruption in patients carrying apparently balanced chromosome rearrangements with intellectual deficiency and/or congenital malformations. Journal of Medical Genetics.2013, 50:144-150

2. Fluorescence in situ hybridization(FISH)

Technical characteristics: Simple operation, intuitional results,detecting structure variation in large fragments of chromosome, low false positive and negative rate, easy to detect the position of variation. Several disadvantages of it are as follows: difficult to hybrid completely, lower efficiency particularly in the context of using short cDNA probe, low sensitivity for SV detection in small fragments of chromosome; low resolution, unable to quantify ortho-repeated CNV precisely, low throughput, unable to detect loss of heterozygosity (LOH), long cycle for probe custom, costly, Positive control (centromeric probes) is required when verifying quantity change (non-rearrangement).

Technology application: programme for small sample size and less variation sites.

Reference:

Breakpoint mapping by next generation sequencing reveals causative gene disruption in patients carrying apparently balanced chromosome rearrangements with intellectual deficiency and/or congenital malformations. Journal of Medical Genetics.2013, 50:144-150