仅供客户在文章写作时参考，分析内容和方法请以结题报告为准，请客户自行承担文章查重等相关风险

Whole genome sequencing

High quality data laid the foundation of satisfied bioinformatics analysis. In order to ensure correct, complete and reliable data, we strictly controlled every link of sequencing by optimized workflow, such as sample testing, library preparation and sequencing.

1.DNA Sample testing

The quality of isolated genomic DNA was verified by using these two methods in combination:

(1) DNA degradation and suspected RNA/Protein contamination were verified by electrophoresis on 1% agarose gels.

(2) The concentration and purity of DNA samples were further quantified precisely by Qubit dsDNA hs assay kit in Qubit®3.0 Flurometer (Life Technologies, CA, USA). A total amount of 0.3μg DNA per sample was required for library generation.

2.Library preparation and sequencing

A total amount of 0.3 μg DNA per sample was used as input material for the DNA library preparations. Sequencing library was generated using Truseq Nano DNA HT Sample Prep Kit (Illumina USA) following manufacturer's recommendations and index codes were added to each sample. Briefly, genomic DNA sample was fragmented by sonication to a size of 350 bp. Then DNA fragments were endpolished, A-tailed, and ligated with the full-length adapter for Illumina sequencing, followed by further PCR amplification. After PCR products were purified (AMPure XP system), libraries were analyzed for size distribution by Agilent 2100 Bioanalyzer and quantified by real-time PCR (3nM). At last, DNA library were sequenced on Illumina for paired-end 150bp reads.

3.Quality Control

The raw image files obtained from illumine platform were processed with Illumina pipeline for base calling and were stored as FastQ format (Raw data), which contain adapter contamination, low-quality nucleotide and undetected nucleotide (N). These sequence artifacts can impose significant influence on downstream processing analysis. Hence quality control, which is listed below, is applied to guarantee the meaningful downstream analysis.

Quality Control:

(1) Filter reads with adapter contamination (>10 nucleotide aligned to the adapter, allowing≤10% mismatches).

(2) Discard a paired reads if more than 10% of bases are uncertain in either one read;

(3) Discard a paired reads if the proportion of low quality (Phred quality <5) bases is over 50% in either one read.

All the downstream bioinformatics analyses are based on the high quality clean data, which can be obtained after these steps. At the same time, QC statistics including total reads number, raw data, raw depth, sequencing error rate, percentage of reads with average quality> Q20（the percent of bases with phred-scaled quality scores greater than 20）, percentage of reads with average quality >Q30（the percent of bases with phred-scaled quality scores greater than 30） and GC content distribution can be calculated.

4.Reads mapping to reference sequence

Valid sequencing data is mapped to the reference genome (GRCh37/hg19) by Burrows-Wheeler Aligner (BWA) software (Li H *et al*.) to get the original mapping result in BAM format. Subsequently, Samtools (Li H *et al*.) and Sambamba are spectively utilized to sort bam files, do duplicate-marking to generate final bam file. If one or one pair read(s) has multiple mapping positions, the strategy adopted by BWA is to select the best one, if there are multi best mapping position, we randomly pick one. Mapping step is very difficult due to mismatches, including true mutation and sequencing error, and duplicates resulted from PCR amplification. These duplicate reads are uninformative and shouldn't be considered as evidence for variants. Picard is employed to mark these duplicates so that we will ignore them in the following analysis.

5.Variant calling

In this step, reads are collected for mutation identification and subsequent analysis. Samtools mpileup and bcftools are used to do variant calling and identify SNP, indels. We employed the reliable user-friendly computational pipeline control-FREEC (Boeva V *et al.*) to discover disruptive genic CNVs in human genetic studies of disease, which might be missed by standard approaches. We provided genome-wide detection of five types of structural variants: inter-chromosomal translocations (CTX), intra-chromosomal translocations (ITX), inversions (INV), deletions (DEL), and insertions (INS). CREST (Wang J *et al*.) was used to identify SVs with standard settings. It mapped the breakpoints of SVs by using the information of soft-clipping reads

and applying an assembly-mapping-searching-assembly alignment procedure consisting of CAP3 and BLAT.

6.Functional Annotation

Functional annotation is very important because the link between genetic variation and disease can be found in this step. ANNOVAR (Wang K *et al*.) is performed to do annotation for VCF (Variant Call Format) file obtained in the previous step. The variant position, variant type, conservative prediction and other information are obtained at this step through a variety of databases, such as dbSNP, 1000 Genome, GnomAD, CADD and HGMD. Since we are interested in exonic variants, gene transcript annotation databases, such as Consensus CDS, RefSeq, Ensemble and UCSC, are also applied for annotation to determine amino acid alternation.

7.Filter

Variants obtained from previous steps are then filtered as follows:

(1) Mutations will be removed if MAF of the mutation > 1% in more than one of the four frequency databases of 1000 genomic data (1000g_all), esp6500siv2_all, gnomAD data (gnomAD_ALL and gnomAD_EAS). A truly pathogenic rare mutation is possible obtained by removing the diversity locus between individuals.

(2) Only SNVs occurring in exons or splice sites (splicing junction 10 bp) are further analyzed since we are interested in amino acid changes.

(3) Then synonymous SNVs which are not relevant to the amino acid alternation predicted by software are discarded; The small fragment non-frameshift mutation in the repeat region (<10bp) are discarded.

(4) Variations are screened according to scores of SIFT, Polyphen, MutationTaster and CADD softwares. The potentially deleterious variations are reserved if the score of more than half of these four softwares support harmfulness of variations. The splicing variations which are no greater than 2bp (±1 ~2bp) from exon are reserved. Mutations that are predicted by dbscSNV to affect splicing are reserved. (for example, the predicted result of a site is 'SIFT=0.0,D', 'Polyphen=0.923,D 0.999,D', 'MutationTaster=1.000,N', 'CADD=.', 'dbscSNV_score=.', so the ratio of softwares which supports harmfulness of variation is 2/4, the variation is reserved. The predicted results of a site is 'SIFT =.', 'Polyphen =.', 'MutationTaster = 1.000, N', 'CADD =.', 'dbscSNV_score = 0.5589, 0.636',

so the predicted software doesn't support harmfulness of variation, but one of the dbsSNV's scores is higher than 0.6, that is, the software predicts that the mutation will affect the splicing and the site is also retained).

8.ACMG

In 2015, the American College of Medical Genetics and Genomics (ACMG) developed guidance for the interpretation of sequence variants, becoming the gold standard for data interpretation (Sue Richards et al, 2015). ACMG developed the classification system of variation， and it is recommended to use a specific standard terminology. The variations are classified into pathogenic, likely pathogenic, uncertain significance, likely benign and benign. There are 28 evidence categories in ACMG variation classification system, and according to the combination form of 28 evidences, to carry out the harmful classification of the variation sites.

9.Analysis of the harmfulness of SV/CNV

Similar to single nucleotide variation (SNVs, single nucleotide variants), many SV/CNVs are normal polymorphisms in the biological genome, and this benign SV/CNV does not lead to pathological changes in the organism. However, it is also found that some malignant SV/CNV is related to neurological disorders, cancer and other diseases.

In order to further filter the benign SV/CNV from the SV/CNV results detected by the software. We used a variety of SV/CNV databases to classify the test results. The classification criteria are as follows:

1. Use the DGV database and its derived series StringentLib, InclusiveLib and DGV.GoldStandard. July2015 to annotate the detected benign SV/CNVs;

2. Use the CNVD database to annotate the detected malignant SV/CNVs;

3. According to the annotation situation, SV/CNVs are divided into four categories: H(high), the SV/CNV is annotated in the malignant database (CNVD database) and is not annotated in the benign database (DGV database); P(Possibly deleterious), the SV/CNV is not annotated neither in the malignant database (CNVD database) nor the benign database (DGV database); M(medium), the SV/CNV is annotated in the malignant database (CNVD database) and the benign database (DGV database); L(low), the SV/CNV is annotated in the benign database (DGV database) and is not annotated in the malignant database (CNVD database). If a variation can be predicted in two

different softwares, it is labeled "*", and if a variation is marked in "genomicSuperDups" or "Repeat", the label is "-".

10.Non-coding region filtering

Non-coding regions play an important role in gene expression regulation. In addition, the variation of non-coding regions will also cause many diseases.

It is essential that screening was conducted based on non-coding region mutation sites. The steps of screening are as follows:

1. According to disease or phenotype, screen non-coding region genetic variation related to disease by Genomiser.

2.Epigenome annotation:(1)Provide relevant organizations (1.Artery_Heart,2.Brain, 3.Digestive_System, 4.Reproductive_System, 5.Endocrine_System, 6.Muscle, 7.Adipose, 8.Skin, 9.Spleen, 10.Lung); (2)Not provide relevant organizations(Artery_Aorta annotation by GTEx database, Aorta（E065）annotation by Roadmap database)

3. Filter screening: (1) frequency screening: allele frequency≤0.01in GnomAD_EAS database; (2) conservative screening: GWAVA and CADD score screening results (GWAVA score of > 0.5, CADD score of ≥10 indicates deleteriousness for the variant). SNP screening strategy: filter out CADD or GWAVA with a score, and indicates harmlessness for the variant; InDel screening: retain the variant of CADD score >10; (3) GTEx database and Epigenomics database (Roadmap and Encode) filtering step (optional): According to providing tissue types, screening mutations can affect the specific gene expression, and annotated as DNA function elements;(4) Screening using the family genetic model (optional, performed by only family samples)

11.Variant filtering under dominant and recessive inheritance models

The mode of inheritance of a monogenic disorder strongly influences both the experimental design and the analytical approach. As we all know, mendelian hereditary disease can be divided into dominant hereditary disease and recessive hereditary disease. In order to screen out possible candidate sites, there are two strategies of variant filtering as below:

11.1 Heterozygote dominant variants

As for dominant genetic disease, the pathogenic mutations are usually heterozygous mutation which come from the father or mother. So heterozygous mutations should be considered first. After filtering the mutation site previously, then reserve loci which patients' autosomal is heterozygous

mutations while normal people have no mutation (sex chromosome reserved mutation loci) as candidate loci.

11.2 Homozygous and compound heterozygote recessive variants

1.After filtering mutation site as previously step, then reserve the loci which patients are homozygous variation, while the loci which normal people are either heterozygous or no mutation in family.

2.The filtering strategy of compound heterozygous model is that reserve the loci which not belong to homozygous site from either patients or normal persons. In addition, the loci must have more than two heterozygous site in patients, besides that the distribution of mutation site which belong to patients was different form normal persons, and it is not a subset of mutation sites constituting normal people.

12. Identification of *de novo* mutations

De novo mutation is the pathogenetic variant which arises for the first time in the offspring of normal parents, and it is a further mechanism that can give rise to an apparently sporadic disorder. Two methods are implemented:

1）Seeking *de novo* mutation by SAMtools (*de novo* SNP/InDel)

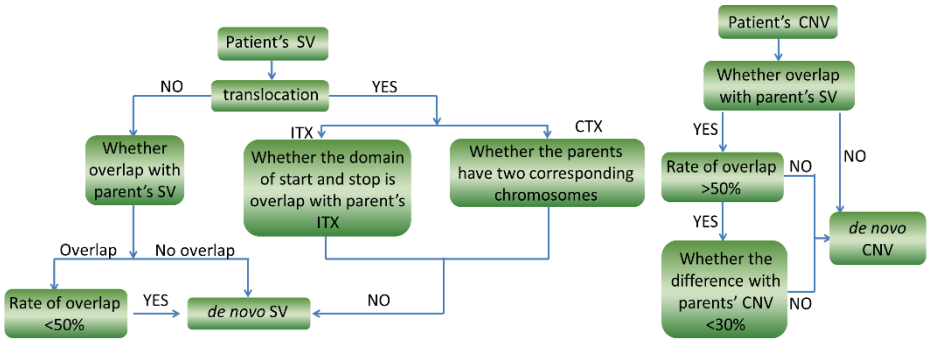
On the basis of agligning to hg19 build of the human genome, variant detection of family trios (patient and both unaffected parents) was performed with SAMtools. Identified candidates for *de novo* mutations were obtain after filtering (the step of "7. filter").

2）Seeking *de novo* mutation by family-based mutation information (*de novo* SNP/InDel)

Seeking *de novo* mutation by family-based mutation information. Identified candidates for *de novo* mutations were obtain after filtering (the step of "7. filter").

The intersection of the above two approaches was also showed in results. *De novo* mutation rate was calculated, and annotation was performed with ANNOVAR (Wang K *et al*.)

De novo SVs/CNVs were identified by the following:



The filtered step of *de novo* SV/CNV

13.Linkage analysis

Linkage analysis can be carried out between a putative disease locus and a single marker locus (two-point linkage) or across a set of markers (multipoint analysis) consisting of a small number of markers or even all markers on a given chromosome.

At present, the most commonly used method is the superior logarithmic score method (LOD) whose LOD value represents the pair value of the probability of two-site linkage and the probability ratio of non-linkage. For monogenic disease, it is definitely linkage when LOD value is greater than 3. It is need to increase the family material further analysis or test priority mutation loci in the candidate area when the LOD values < -2 negative chain or the value between 1 and 2.For complex diseases, LOD has a low threshold.

This linkage analysis using merlin tools and the perl language, combined with the family high throughput sequencing data and the HapMap database of Chinese population (CHB) allele frequency, using the known SNPS as a marker linkage analysis, get the chain candidate area.

14.ROH analysis

The homozygous region which is the homozygous allele in the genome. This interval is due to the fact that the alleles passed from the parents to the offspring are from the same ancestor, that is, the parents have alleles from the same ancestor, and the offspring are passed on as homozygotes. Homozygous localization analysis is commonly used for genetic diseases caused by inbreeding.

15. Overlap based strategy

For sporadic samples, on the basis of the previous step ("7. filter"), the mutation is reserved if this mutation is detected in more than two patients. The number of genes with mutations in multiple affected patients will decrease rapidly by combining data from increasing numbers of patients, resulting in less candidate genes for follow-up.

16. Enrichment analysis of candidate genes

Different genes perform their biological functions by coordinating with each other in organisms. Especially for complex diseases, phenotypic difference could be caused by the mutations of multiple genes. The most important biochemical metabolic pathways and signal transduction pathways involved in mutant genes were determined by significance enrichment analysis.

16.1 Gene Ontology (GO) enrichment analysis

The results of the GO enrichment analysis show as a picture. It includes CC (Cellular Component), BP (Biological Pathway) and MF (Molecular Function).

Gene Ontology (GO) enrichment analysis of candidate genes was implemented by the clusterProfile package. GO terms with corrected P-value less than 0.05 were considered significantly enriched by candidate genes.

16.2 KEGG pathways enrichment analysis

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-through put experimental technologies (http://www.genome.jp/kegg/). We used R software to test the statistical enrichment of candiate genes in KEGG pathways.

17.The analysis of genotype-phenotype correlations

17.1 The annotation from DisGeNeT database

DisGeNET is a discovery platform containing one of the largest publicly available collections of genes and variants associated to human diseases (Piñero et al., 2016; Piñero et al., 2015). Several original metrics are provided to assist the prioritization of genotype－phenotype relationships. We use DisGeNET database to annotate candidate genes.

17.2 The analysis results from Phenolyzer

Phenolyzer stands for Phenotype Based Gene Analyzer, a tool focusing on discovering genes based on user-specific disease/phenotype terms. The selected candidate genes were ranked according to disease association. The figure about association network between genes and disease/phenotype was constructed.

18. PPI network analysis

We mapped interactions between genes harboring validated nonsynonymous SNV by construct PPI

network. We used GeneMania（Warde-Farley D et al.2010-7）to create a PPI network. It includes protein-protein, protein-DNA-genetic interactions, pathways, reactions, gene-protein expression data, protein domains-phenotypic screening profiles.

19. Pharmacogenomics

The most obvious applications of NGS technology to discover novel genetic variation important to drug response include genetic association studies of drug efficacy and drug toxicity in humans. The success of whole-exome sequencing/whole-genome sequencing studies in determining the gene bases for many rare Mendelian diseases suggests that the same approaches may be similarly successful in studying rare drug toxicities. Candidate genes were obtained from the same approaches (eg. Association analysis and Overlap based strategy). We use the PharmGKB and Drugbank databases to annotate variants. According to the related information，the relationship between the mutation and the drug can be analyzed.

20. Association analysis

20.1 Site based association analysis

After applying stringent quality controls, we obtained SNPs for inclusion in GWAS. Genetic association analysis was carried out with Fisher's exact test (*or* Linear and logistic regression models). We used PLINK software(http://pngu.mgh.harvard.edu/~purcell/plink/) (Purcell et al. 2007) to calculate the p value and OR value of each SNP site. Manhattan plots and quantile-quantile plots of the log10 of P-values of the GWAS were generated with custom code in R26.

20.2 Gene based burden analysis

Rare variants may be have large effect to the disease. To enrich variants that were likely to alter the function of the proteins, we defined "qualifying variant" as deleterious, rare exonic and splice site variants with call rate > 90%. Deleterious variants were limited to missense, nonsense, splice sites, frameshift that was not in genomic repeat region, meanwhile, the deleteriousness score was computational predictions by the following methods: SIFT, PolyPhen-2, MutationTaster, CADD and GERP++. Rare variants were filtered using a minor allele frequency (MAF) threshold of 1% in 1000 Genomes and 0.1% in ExAC East Asian.

We then ran an association analysis by coding individuals based on the presence or absence of rare deleterious variants in each sequenced gene and compared the combined frequency of rare variants in each gene between cases and controls with two-tailed fisher's exact test in R statistical package.

The association P values were corrected by several multiple-testing correction methods including Bonferroni correction and Holm-Bonferroni method. False discovery rates were also calculated by Benjamini–Hochberg procedure and Benjamini–Yekutieli procedure.

Reference:

1. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature, 2012, 491(7422): 56-65.（1000G）
2. 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）
3. ExAC：http://exac.broadinstitute.org/ （ExAC）
4. Kent W J, Sugnet C W, Furey T S, et al. The human genome browser at UCSC. Genome research, 2002, 12(6): 996-1006.（UCSC）
5. Krumm N, Sudmant P H, Ko A, et al. Copy number variation detection and genotyping from exome sequence data. [J]. Genome Research, 2012, 22(8):1525-1532.（CoNIFER）
6. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 2009-1, 25(14): 1754-1760.（BWA）
7. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics, 2009-2, 25(16): 2078-2079.（SAMtools）
8. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003,1;31(13):3812-4. （SIFT）
9. Martin K, Witten D M, Preti J, et al. A general framework for estimating the relative pathogenicity of human genetic variants. [J]. Nature Genetics, 2014, 46(3):310-5. （CADD）
10. Picard: http://sourceforge.net/projects/picard/. （Picard）
11. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tartar D, IIBDGC, Cotsapas C, Daly MJ. Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology.PLoS Genetics，2011,7(1): e1001273（DAPPLE）
12. Sherry S T, Ward M H, Kholodov M, et al. dbSNP: the NCBI database of genetic variation [. Nucleic acids research, 2001, 29(1): 308-311.（dbSNP）
13. Schwarz J M, Rodelsperger C, Schuelke M, et al. MutationTaster evaluates disease-causing

potential of sequence alterations. [J]. Nature Methods, 2010, 7(8):575-576. （MutationTaster）

14. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology[J]. Genetics in Medicine Official Journal of the American College of Medical Genetics, 2015, 17(5):405.
15. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research, 2010, 38(16): e164-e164.（ANNOVAR）
16. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations[J]. Nucleic acids research, 2014, 42(D1): D1001-D1006.（gwascatalog）
17. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses[J]. The American Journal of Human Genetics, 2007, 81(3): 559-575.
18. Wang J, Mullighan C G, Easton J, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution[J]. Nature methods, 2011, 8(8): 652-654.（CREST）