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

---中国区 重测序业务线

Target Region sequencing（Disease）

1 Experimental Procedure

1.1 Evaluation of DNA quality

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

(1) DNA degradation and contamination were monitored on 1% agarose gels;

(2) DNA concentration was measured by Qubit® DNA Assay Kit in Qubit® 3.0 Flurometer (Invitrogen, USA).

1.2 Library Preparation

To get the target gene regions, we designed probes on the website of Agilent about XX genes according the design description. Briefly, fragmentation was carried out by hydrodynamic shearing system (Covaris, Massachusetts, USA) to generate 180-280 bp fragments. Extracted DNA was amplified by ligation-mediated PCR (LM-PCR), purified, and hybridized to the probe for enrichment, and non-hybridized fragments were washed out. Both non-captured and captured LM-PCR products were subjected to real-time PCR to estimate the magnitude of enrichment. Each captured library was then loaded on Illumina platform, and we performed high-throughput sequencing for each captured library independently to ensure that each sample met the desired average fold coverage.

1.3 Clustering & Sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using Illumina PE Cluster Kit (Illumina, USA) according to the manufacturer’s instructions. After cluster generation, the DNA libraries were sequenced on Illumina platform and 150 bp paired-end reads were generated.

2 Bioinformatics Analysis Pipeline

2.1 Data Quality Control

2.1.1 Raw data

The original fluorescence image files obtained from Illumina platform are transformed to short reads (Raw data) by base calling and these short reads are recorded in FASTQ format, which contains sequence information and corresponding sequencing quality information.

2.1.2 Evaluation of data (Data quality control)

Sequence artifacts, including reads containing adapter contamination, low-quality nucleotides and unrecognizable nucleotide (N), undoubtedly set the barrier for the subsequent reliable bioinformatics analysis. Hence quality control is an essential step and applied to guarantee the meaningful downstream analysis.

The steps of data processing were as follows:

(1) Discard a paired reads if either one read contains adapter contamination (>10 nucleotides 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 were based on the high quality clean data, which were retained after these steps. At the same time, QC statistics including total reads number, raw data, raw depth, sequencing error rate and percentage of reads with Q30 (the percent of bases with phred-scaled quality scores greater than 30) were calculated and summarized.

2.2 Reads Mapping to Reference Sequence

Valid sequencing data is mapped to the reference genome (GRCh37/hg19/GRCh38) by BurrowsWheeler 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 are 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. Sambamba is employed to mark these duplicates so that we will ignore them in the following analysis.

2.3 Variant detection

SAMtools (Wysoker A, et al.) mpileup and bcftools were used to do variant calling and identify SNP, InDels.

2.4 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.

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] Artem T, Vilella A J, Edwin C, et al. Sambamba: fast processing of NGS alignment formats[J]. Bioinformatics, 2015(12):2032-2034.

[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] Consortium G O. 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] Georg B, Ehret, Patricia B, Munroe, Kenneth M, Rice, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk[J]. Nature, 2011, 478(7367):103-109.

[15] Joshi P K, Esko T, Mattsson H, et al. Directional dominance on stature and cognition in diverse human populations[J]. Nature, 2015, 523(7561): 459-462.

[16] 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).

[17] Muona M, Berkovic S F, Dibbens L M, et al. A recurrent de novo mutation in KCNC1 causes progressive myoclonus epilepsy.[J]. Nature Genetics, 2014, 47.

[18] Zaidi S, Choi M, Wakimoto H, et al. De novo mutations in histone-modifying genes in congenital heart disease[J]. Nature, 2013, 498(7453): 220-223.

[19] Sanders S J, Murtha M T, Gupta A R, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism[J]. Nature, 2012, 485(7397): 237-241.

[20] 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.

[21] Magi A,Tattini L,Palombo F, et al. H3M2: detection of runs of homozygosity from whole-exome sequencing data[J]. Bioinformatics, 2014, Oct 15;30(20):2852-9.