疾病 WGS 英文版 method

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。

1. Experimental Procedure

1.1.Sample Quality Control

Please refer to QC report for methods of sample quality control.

1.2 DNA fragmentation (DNA Shearing)

The genomic DNA was randomly interrupted by a Covaris fragmentation device into fragments of about 350 bp.

1.3 End repair reaction (End Repair)

The fragmented DNA has a protrusion at the 5' or 3' end, and an end-complement system is added to the purified DNA fragments, in which the exonuclease activity of T4 DNA polymerase is used to digest the single-stranded protrusion at the 3' end, and its polymerase activity is used to complement the protrusion at the 5' end. At the same time phosphate kinase (PNK) adds a phosphate group to the 5' end, which is required for subsequent ligation reactions. After purification with Agencourt AMPure XP magnetic beads, a library of short, flat-ended DNA fragments containing a phosphate group at the 5' end is obtained.

1.4 Addition of poly(A) tail at 3' end (Adenlylate 3’ Ends)

The above system was supplemented with 3' end-addition poly(A) buffer reaction system. A single adenylate "A" is added to the 3' end of the end-modified double-stranded DNA, which is to prevent the flat ends of DNA fragments from self-associating with each other, so that they can be complementarily paired with the single "T" protrusion at the 5' end of the sequencing junction in the next step, and accurately connected, effectively reducing the crosstalk between library fragments themselves.

1.5 Adapter Ligation

Ligation buffer and double-stranded sequencing adapters were added to the above reaction system, T4 DNA Ligase was used to ligate Illumina sequencing adapters to both ends of the library DNA.

1.6 Size Selection

For libraries that have been attached to adapters, fragment size screening should be performed by the Agencourt SPRIselect kit (Beckman Coulter, USA, Catalog # : 2358413) at the same time as the library is purified. Double Size Selection was adopted,specifically, the SPRI beads were used to remove the small fragments on the left side of the target domain, and then the large fragments located on the right side of

the target fragment region were removed, so that the original library with moderate fragment size was selected for the next step of PCR amplification.Excess sequencing adapters and adapter self-linkage products in the system have been removed from the purified library to avoid invalid amplification during the PCR process and to eliminate the impact on on-board sequencing.

1.7 PCR Amplification

The original library is amplified using a high-fidelity polymerase to ensure a sufficient total amount of library. In addition, this step is able to efficiently enrich for DNA fragments with adapters at both ends, which are the only fragments that can be amplified efficiently.With the premise of ensuring sufficient products, the bias introduced due to the excessive number of amplification cycles was reduced.The final concentration of each library was accurately determined using Qubit 3.0.

1.8 Library Quality Assessment

After the library construction was completed, the insert size of the library was detected by Agilent 5400 system (AATI), and after the insert size met the expectation, the effective concentration of the library (1.5 nM) was accurately quantified by qPCR to ensure the quality of the library.

1.9 Bridge PCR

When the library is qualified, the library is then sequenced on the Illumina Novaseq platform according to the validated concentration and data output of the library. Is the process of inoculthe captured library onto the FlowCell chip for amplification.The inner surface of the FlowCell channel has two different DNA primers that complement the adaptor sequences at both ends of the DNA library and are attached to the FlowCell as a covalent bond.The specific steps are described as follows:

a.The DNA library is added to the chip. Since the DNA sequence of the double end and the primer sequence on the chip are complementary, complementary hybridization will be performed. After hybridization, dNTP and polymerase are added. the polymerase starts from the primer, along the template, to synthesize a DNA strand complementary to the original DNA sequence;

b.NaOH solution is added to make the DNA duplex unchain, wash away the DNA chain that is not covalently connected to the chip, and retain the newly synthesized DNA chain that is covalently connected to the chip;

c. The neutralizing liquid is added to the liquid flow magnetic to neutralize the alkaline solution, then the other end of the DNA and the other primer on the chip undergo complementary hybridization, and the enzyme and dNTP are added to synthesize a new DNA chain. The alkali solution is added again to separate the two DNA strands, then

the neutralisation solution is added, followed by the addition of the enzyme and dNTP and the synthesis of the new strand with the new primer. This process is repeated continuously, and the DNA strands grow exponentially.

1.10 Illumina platform for PE150 on-board sequencing

PE150 is Pair-end 150bp, high-throughput sequencing. In the constructed DNA small fragment library, each insert fragment was sequenced at both ends, each end was sequenced at 150bp, as follows.

After completion of bridge PCR, the synthesised double strand is turned into a single strand that can be sequenced;

a. A specific group of one of the primers on the chip is cut and the chip is rinsed with alkali solution, causing the DNA double strand to unravel and the cut DNA strand to be washed away, leaving the strand covalently bonded.
b. Add neutral solution, sequencing primers and fluorescently labelled dNTP, in which the four dNTPs are labelled with four different fluorescent labels and their 3' ends are blocked by azide groups. The dNTP is then synthesised into the new DNA strand by adding polymerase, which can only extend one base per cycle due to the blockage of the 3' end by the sodium azide group. After completing a cycle, the excess dNTP, enzyme, etc. are washed out and placed under a microscope for laser scanning to determine the type of newly synthesised base based on the fluorescence emitted, and the template base can be inferred through the principle of complementarity.
c. c.After completing the cycle, chemicals are added that allow the sodium azide group and the fluorescent group to be cleaved off, exposing the hydroxyl group at the 3' end. New dNTP and enzyme are added, another base is extended, and the excess dNTP and enzyme are flushed out before the round of microscopic laser scanning is performed to read out this round of bases. This cycle is repeated over and over again for hundreds of base reads.

2. Bioinformatic analysis

The raw sequences were analysed for information at the end of sequencing. The data quality was evaluated to determine whether it met the standard. If the data meets the standard, the samples will be tested for variants, including SNP, InDel, CNV, SV, and annotated with the corresponding variant information. If the data do not meet the quality standard, additional testing or library reconstruction is required according to the actual situation.

2.1 Data quality control

2.1.1 Raw sequence data

The raw image data files obtained by the Illumina sequencing platform of the original sequencing data were transformed by base calling analysis into the raw sequenced reads, namely Raw Data.The results are stored in the FASTQ (fq) file format containing the sequence information of Sequenced reads and their corresponding sequencing quality information.

2.1.2 Quality assessment of the sequencing data

2.1.2.1 Raw data filtering

Remove reads pair with adapters; remove reads pairs with more than 10% of N (N indicates the inability to determine the base information);this pair of reads was removed when the single-end sequencing reads containing low quality (below 5) bases exceeded 50% of the length of the reads.

2.1.2.2 Detection of the distribution of the sequencing error rates

The sequencing error rate is calculated from a model that determines the probability of an error during base recognition (Base Calling). It is related to base quality, and is affected by multiple factors such as the sequencer itself, sequencing reagent, samples and more. Sequencing error rate distribution inspection is used to detect high error rates in base positions with or without abnormalities within the sequencing length. Generally, they should be less than 1% per base position

2.1.2.3 Detection of the GC content distribution

The test mainly detect AT, GC separation phenomenon, theoretically A and T base, C and G base on each sequencing cycle should be equal, but in the actual sequencing process, due to the DNA template amplification deviation, the first few base sequencing quality, the first few bases in each read fluctuation is a normal phenomenon.

2.1.2.4 Distribution of the sequencing data quality

According to the characteristics of sequencing technology, the terminal base quality of sequencing fragments is generally lower than that of the front end. Only when the quality of sequencing data is mainly distributed $Q30 \geq 85\%$ can the subsequent analysis be guaranteed normally.

2.1.3 Statistics of the sequencing depth and coverage

Effective sequencing data were aligned by BWA (Li et al., 2018) to the reference genome (GRCh37 / hg19 / GRCh38) to obtain the initial alignment in BAM format. Then, the results were sorted by Sambamba (Tarasov et al., 2015) and the duplicate reads (mark duplicate reads) were labeled.

Finally, the alignment results after repeated labeling are used to calculate the coverage, depth, etc. Generally, the sequencing reads of human samples can exceed 95%; the SNP detected at a site is credible when the base coverage depth (read depth) reaches 10 X.

2.2 Variant detection results

Based on the initial alignment results (BAM files), bcftools (Li et al., 2009) was used to identify SNP and InDel sites, count the number of SNP and InDel in different regions of the genome, the number of different types of SNP and InDel in coding regions, the distribution of transitions and transversions , the number of SNP and InDel and the distribution of genotypes. The germline SNP and InDel filtering parameters are as follows: QUAL ≥ 20; DV ≥ 4; MQ ≥ 30.
The increase and decrease of CNV were detected by Control-FREEC (Boeva et al., 2012), and the number of different types of CNV events was counted. SV was detected using Lumpy software (Layer et al., 2014) and the number of SV of different types was counted.

2.3 Annotations

ANNOVAR (Wang et al., 2010) is an efficient software tool that uses the latest information to provide functional annotation of gene variants detected by multiple genomes. The vcf (variantcallformat) obtained in the previous work was annotated using ANNOVAR
(1) Refseq (O'Leary et al., 2016) was used to annotate the gene structure of the variant site, with gene types including mRNA, non-coding RNA, etc.
(2) The genomic characteristics of the variant sites should be treated with caution for the mutations located within the repetitive segments of the genome.
(3) The effect of the nonsynonymous mutations on the disease / tumor was comprehensively evaluated by SIFT (Ng et al., 2003), PolyPhen (Adzhubei et al., 2013), and MutationTaster (Reva et al., 2011).
(4) Annotations of dbSNP (Sherry et al., 2001), thousand human genome SNP database (Abecasis et al., 2012), known tumor somatic mutation database COSMIC (Tate et al., 2019)and esp6500 variant database are provided, and the variant results can be screened for any combination.
(5) The annotation also includes functional annotation of the genes where the mutation occurred, using databases including GO (Lee et al., 2004), KEGG (Kanehisa et al., 2000), Reactome (Jassal et al., 2020), Biocarta, PID (Schaefer et al., 2009), etc.

3. Reference

Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65. doi:10.1038/nature11632 (1000 Genomes)

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. doi:10.1002/0471142905.hg0720s76 (PolyPhen)

Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012;28(3):423-425. doi:10.1093/bioinformatics/btr670 (Control-FREEC)

Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005-D1012. doi:10.1093/nar/gky1120 (GWAS Catalog)

Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553-1561. doi:10.1101/gr.092619.109 (LRT)

Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 2014;30(17):2503-2505. doi:10.1093/bioinformatics/btu314 (Samblaster)

Frankish A, Diekhans M, Jungreis I, et al. GENCODE 2021.Nucleic Acids Res. 2021;49(D1):D916-D923. doi:10.1093/nar/gkaa1087 (GENCODE)

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25(12):i54-i62. doi:10.1093/bioinformatics/btp190 (SiPhy)

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(Database issue):D514-D517. doi:10.1093/nar/gki033 (OMIM)

Harris MA, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics resource.Nucleic Acids Res. 2004;32(Database issue):D258-D261. doi:10.1093/nar/gkh036 (GO)

Huber CD, Kim BY, Lohmueller KE. Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. PLoS Genet. 2020;16(5):e1008827. Published 2020 May 29. doi:10.1371/journal.pgen.1008827 (GERP)

Jassal B, Matthews L, Viteri G, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020;48(D1):D498-D503. doi:10.1093/nar/gkz1031 (Reactome)

Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30. doi:10.1093/nar/28.1.27 (KEGG PATHWAY)

Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996-1006. doi:10.1101/gr.229102 (UCSC)

Kobayashi Y, Yang S, Nykamp K, Garcia J, Lincoln SE, Topper SE. Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. Genome Med. 2017;9(1):13. Published 2017 Feb 6. doi:10.1186/s13073-017-0403-7 (ExAc)

Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15(6):R84. Published 2014 Jun 26. doi:10.1186/gb-2014-15-6-r84 (Lumpy)

Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform.Bioinformatics. 2010;26(5):589-595. doi:10.1093/bioinformatics/btp698 (BWA_MEM)

Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352 (SAMtools)

Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-3814. doi:10.1093/nar/gkg509 (SIFT)

O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.Nucleic Acids Res. 2016;44(D1):D733-D745. doi:10.1093/nar/gkv1189 (RefSeq)

Pio MG, Siffo S, Scheps KG, et al. Curating the gnomAD database: Report of novel variants in the thyrogobulin gene using in silico bioinformatics algorithms. Mol Cell Endocrinol. 2021; 534:111359. doi: 10.1016/j.mce.2021.111359 (gnomAD)

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20(1):110-121. doi:10.1101/gr.097857.109 (phyloP)

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012;28(3):423-425. doi:10.1093/bioinformatics/btr670 (Control-FREEC)

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome.Nucleic Acids Res. 2019;47(D1):D886-D894. doi:10.1093/nar/gky1016 (CADD)

Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics.Nucleic Acids Res. 2011;39(17):e118. doi:10.1093/nar/gkr407 (MutationAssessor)

Schaefer CF, Anthony K, Krupa S, et al. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009;37(Database issue):D674-D679. doi:10.1093/nar/gkn653 (PID)

Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-311. doi:10.1093/nar/29.1.308 (dbSNP)

Shihab HA, Gough J, Cooper DN, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34(1):57-65. doi:10.1002/humu.22225 (FATHMM)

Steinhaus R, Proft S, Schuelke M, Cooper DN, Schwarz JM, Seelow D. MutationTaster2021. Nucleic Acids Res. 2021;49(W1):W446-W451. doi:10.1093/nar/gkab266 (MutationTaster)

Stenson PD, Mort M, Ball EV, et al. The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting.Hum Genet. 2020;139(10):1197-1207. doi:10.1007/s00439-020-02199-3 (HGMD)

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31(12):2032-2034. doi:10.1093/bioinformatics/btv098 (Sambamba)

Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer.Nucleic Acids Res. 2019;47(D1):D941-D947. doi:10.1093/nar/gky1015 (COSMIC)

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.Nucleic Acids Res. 2010;38(16):e164. doi:10.1093/nar/gkq603 (ANNOVAR)