Human Whole Genome Sequencing Methods

1. Experimental Procedure

1.1.Sample Quality Control

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

1.2 Library Construction, Quality Control and Sequencing

The genomic DNA sample was fragmented into short fragments. These DNA fragments were then end-polished, A-tailed, and ligated with full-length adapters for Illumina sequencing before further size selection. PCR amplification was then conducted unless specified as PCR-free. Purification was then conducted through the AMPure XP system (Beverly). The resulting library was assessed on the Agilent Fragment Analyzer System (Agilent) and quantified to 1.5nM through Qubit (Thermo Fisher Scientific) and qPCR.

The qualified libraries were pooled and sequenced on Illumina platforms, according to the effective library concentration and data amount required.

2. Bioinformatics Analysis

2.1 Raw data

The original fluorescence image files obtained from the Illumina platform were transformed into short reads (raw data) by base-calling and these short reads were recorded in the FASTQ format, which contains sequences and the corresponding sequencing quality information.

2.2 Data Quality Control

It was the sequence artifacts, including reads containing adapter contamination, low-quality nucleotides, and unrecognizable nucleotide (N), that undoubtedly set the barrier for the subsequent reliable bioinformatics analysis. Therefore, quality control is an essential step to mitigate these obstacles and could be applied to guarantee meaningful downstream analysis.

The steps of data processing were as follows:

(1) Discard a pair of reads if either one of them contains adapter contamination (>10 nucleotides aligned to the adapter, allowing ≤ 10% mismatches);

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

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

Total reads number, raw data, error rate, and percentage of reads with Q30 (the percent of bases with Phred-scaled quality scores greater than 30) were calculated and summarized. After which, filtered reads were used as clean data for subsequent analysis.

2.3 Sequence Alignment

Clean data were mapped to the reference genome (b37/hg19/hg38) by Burrows-Wheeler Aligner (BWA) software (Li et al., 2018) to generate BAM files. Subsequently, Sambamba (Tarasov et al., 2015) was used to sort BAM files according to chromosome position. Picard tools (The Quick Start is available online: https://broadinstitute.github.io/picard/) were then utilized to merge BAM files and mark duplicate reads.

2.4 Variant Detection

2.4.1 Germline Mutation Detection

GATK (DePristo et al., 2011) HaplotypeCaller was used to call germline SNP and InDel, while the use of GATK VariantFiltration module was carried to filter germline SNP and InDel. The filter parameters of SNP and InDel are shown as follows:

SNP: QD < 2.0 || FS > 60.0 || MQ < 40.0 || HaplotypeScore > 13.0 || MappingQualityRankSum < -12.5 || ReadPosRankSum < -8.0

InDel: QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0

Control-FREEC (Boeva et al., 2012) and DELLY (Rausch et al., 2012) were used to call germline CNV and SV, respectively. The parameter with window = 2000 and step = 1000 was set in the config file for CNV calling.

2.4.2 Somatic Mutation Detection (Only for normal-tumor paired samples)

Somatic mutation detection is commonly applied to normal and tumor-paired samples. The Somatic SNV was detected by MuTect (Cibulskis et al., 2013), while the somatic InDel was identified by Strelka (Saunders et al., 2012). Softwares for Somatic CNV and SV detection are the same as that for germline CNV and SV detection.

2.5 Annotation

ANNOVAR (Wang et al., 2010) was used to perform variant annotation. Annotation contents refer to protein-coding changes, genomic regions affected by the variants, allele frequency, deleterious prediction, etc. The main databases used were as follows:

Genes and regions annotation

RefSeq (O'Leary et al., 2016) and Gencode (Frankish et al., 2021) databases were used to find genomic regions affected by variants and possible changes in the protein. We annotated the features of the genomic regions affected by the variants, such as cytoband, small RNA, conserved mammalian microRNA regulatory target sites, conservative regions of vertebrates, transcription factor binding sites, repeats, etc.

Databases with frequency annotation

The established databases, such as 1000 Genomes (1000 Genomes Project Consortium) (Abecasis et al., 2012), Exome Aggregation Consortium (ExAC) (Kobayashi et al., 2017), Genome Aggregation Database (gnomAD) (Pio et al., 2021) and exome sequencing project (ESP), were all used to find alternative allele frequencies in the populations that were reported. There are a great number of common polymorphism sites in the human population, while many deleterious variants are rare or of low frequency.

Databases and scores with conservative and deleterious annotation

SIFT (Ng et al., 2003), PolyPhen (Adzhubei et al., 2013), MutationAssessor (Reva et al., 2011), LRT (Chun et al., 2009), and CADD (Rentzsch et al., 2019) scores were used to predict the deleterious mutations. GERP++ (Huber et al., 2020) scores were used to evaluate the conservation of mutations. SIFT, Polyphen2, MutationTaster (Steinhaus et al., 2021), LRT, MutationAssessor, and FATHMM (Shihab et al., 2013) were all used to predict whether an amino acid substitution affected protein function. SiPhy (Garber et al., 2009), phyloP (Pollard et al., 2010), GERP++, and CADD were all used to predict the conservative level of the site. It should be noted that the conservation scores only consider the conservative level at the current site, but not the one involved in the nucleotide identity. Therefore, synonymous and non-synonymous variants at the same site will have the same scores. These scores are used for finding functionally important sites, which means that variants that confer increased susceptibility would score well.

Databases with cancer and disease-related annotation

dbSNP (Sherry et al., 2001), COSMIC (Tate et al., 2019), OMIM (Hamosh et al., 2005), GWAS Catalog (Buniello et al., 2019), and HGMD (Stenson et al., 2020) were used to find reported information of the variants.

Databases with functional and pathway annotation

Gene Ontology (Lee et al., 2004), KEGG (Kanehisa et al., 2000), Reactome (Jassal et al., 2020), and PID (Schaefer et al., 2009) databases were applied to provide functional or pathway annotation.

3. References

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)

Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213-219. doi:10.1038/ nbt.2514 (muTect)

DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491-498. doi:10.1038/ng.806 (GATK)

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)

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)

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)

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28(18):i333-i339. doi:10.1093/ bioinformatics/bts378 (DELLY)

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)

Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small- variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28(14):1811-1817. doi:10.1093/bioinformatics/bts271 (Strelka)

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)