疾病 WGS 中文版 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.实验流程

1.1 样品 DNA 质量检测 (DNA Quality Control)

详见样本检测报告。

1.2 DNA 片段化 (DNA Shearing)

将基因组 DNA 经 Covaris 破碎仪随机打断成长度为 350bp 左右的片段。

1.3 末端修复反应 (End Repair)

片段化后的 DNA 存在 5’或 3’端突出，向纯化后的 DNA 片段中加入末端补齐体系，其中 T4 DNA 聚合酶(T4 DNA Polymerase)的外切酶(Exonuclease)活性消化 3’端的单链突出，而聚合酶(Polymerase)活性补齐 5’端的突出；同时磷酸激酶(PNK)在 5’末端加上后续连接反应必需的磷酸基团，经过 Agencourt AMPure XP 磁珠纯化，最终得到 5’端含有磷酸基团的平末端 DNA 短片段文库。

1.4 3’端加“A”尾 (Adenlylate 3’ Ends)

向上述体系中加入 3’末端加“A”缓冲反应体系。在末端修饰完成的双链 DNA3’末端加上单个腺苷酸“A”，防止 DNA 片段之间的平末端自连，还可以与下一步测序接头 5’末端的单个“T”突出互补配对，准确连接，有效降低文库片段之间自身的串联。

1.5 连接测序接头(Adapter Ligation)

向上述反应体系中加入连接缓冲液和双链测序接头，利用 T4 DNA 连接酶将 Illumina 测序接头连接至文库 DNA 两端。

1.6 文库片段筛选(Size Selection)

对于加上接头的文库，应用 Agencourt SPRIselect (Beckman Coulter, USA, Catalog # : 2358413)核酸片段筛选试剂盒在纯化文库的同时，进行片段大小筛选。采用两步法筛选(Double Size Selection)，先用 SPRI 磁珠去掉目标域左侧小片段(Left-side Size Selection)，再去掉位于目标片段区域右侧的大片段(Right-side Size Selection) 最终筛选出片段长度适中的原始文库，用于下一步的 PCR 扩增。经

过纯化后的文库，去掉了体系中过量的测序接头和接头自连产物，避免 PCR 过程的无效扩增，消除对上机测序的影响。

1.7 PCR 扩增 DNA 文库(PCR Amplification)

应用高保真的聚合酶扩增原始文库，以保证足够的文库总量。此外因为只有两端都连有接头的 DNA 片段才能够被扩增，因此该步骤还能够有效富集这部分 DNA。在保证产物足够的前提下，减少因扩增循环数过大而引入的 bias；最终使用 Qubit3.0 精确测定每个文库浓度。

1.8 文库库检(Library Quality Assessment)

文库构建完成后，Agilent 5400 system(AATI)对文库的 insert size 进行检测，insert size 符合预期后，使用 qPCR 方法对文库的有效浓度(1.5nM)进行准确定量，以保证文库质量。

1.9 桥式 PCR(BridgePCR)

库检合格后，根据文库的有效浓度及数据产出在 Illumina Novaseq 平台进行测序。即将捕获后的文库种到 FlowCell 芯片上进行扩增的过程。FlowCell 通道内表面种植有两种不同的 DNA 引物，这两种引物序列与 DNA 文库中两头的接头序列相互补，且以共价键形式连接在 FlowCell 上。具体过程如下:

a. 将 DNA 文库加入到芯片上，由于文库两头的 DNA 序列和芯片上的引物序列互补，产生互补杂交，杂交完后，加入 dNTP 和聚合酶，聚合酶从引物开始，沿着模板，合成一条与原来 DNA 序列互补的 DNA 链;

b. 加入 NaOH 碱溶液，使得 DNA 双链解链，冲走原来那条没有和芯片共价连接的 DNA 链，保留新合成的和芯片共价连接的 DNA 链;

c. 再在液流磁中加入中和液，中和掉碱性溶液，此时 DNA 上的另一端和芯片上的另一个引物发生互补杂交，加入酶和 dNTP，合成一条新的 DNA 链;再次加入碱溶液，使两条 DNA 链分开，再加中和液，DNA 即和芯片上新的引物杂交，加酶和 dNTP，再次从新的引物上合成新链，连续重复这一过程，DNA 链以指数的方式增长。

1.10 Illumina 平台 PE150 上机测序(Sequencing)

PE150 即 Pair-end150bp，高通量测序。在构建的 DNA 小片段文库中，将每条插入片段进行两端测序，每端各测 150bp，具体过程如下:

完成桥式 PCR 之后，将合成的双链变成可以测序的单链；

a. 将芯片上其中一个引物的一个特定基团切断，碱溶液冲洗芯片，使得 DNA 双链解链，且被切断根部的 DNA 链被冲掉， 留下被共价键连接的那条链;

b. 加入中性溶液、测序引物及带荧光标记的 dNTP，四种 dNTP 由四种不同的荧光标记，其 3’末端被叠氮基堵住，再加入聚合酶，使 dNTP 合成到 新的 DNA 链上，由于其 3’末端被叠氮钠基堵住，故每个循环只能延长一个碱基，完成一个循环后将多余的 dNTP、酶等冲掉，置于显微镜下进行激光扫描，根据发出来的荧光判 断新合成的是哪个碱基，通过互补原理可推测模板碱基;

c. 在完成一个循环之后，加入化 学试剂，将叠氮钠基团和荧光基团切掉，使得 3’端羟基暴露出来，加入新的 dNTP 和新的酶，又延长一个碱基，新的碱基延长完成之后，把多余的 dNTP 和酶冲掉，再进行一轮显微激光扫描，再读一轮此碱基，不断重复此循环，就可以读出上百个碱基。

2.生物信息分析

测序结束后对原始序列进行信息分析，通过对数据质量进行评估，判断其是否达到标准， 若符合标准，则对样本进行变异检测，包括 SNP、InDel、CNV、SV，并注释;若不合标准，则需根据实际情况加测或者重新建库。

2.1 数据质量控制

2.1.1 原始序列数据

原始测序数据通过 Illumina 测序平台得到的原始图像数据文件经碱基识别(Base Calling)分析转化为原始测序序列(Sequenced Reads)，即 Raw Data，结果以 FASTQ(简称为 fq)文件格式存储，其中包含测序序列(reads)的序列信息及其对应的 测序质量信息。

2.1.2 测序数据质量评估

2.1.2.1 原始数据过滤

去除带接头(adapter)的 reads 对;去掉单端测序 read 中 N(N 表示无法确定碱基信 息)的比例大于 10%的 reads 对;当单端测序 read 中含有的低质量(低于 5) 碱基数超 过该条 read 长度比例的 50%时，去除此对 reads。

2.1.2.2 检查测序错误率分布

测序错误率是在碱基识别(Base Calling)过程中通过一种判别发生错误概率的模型计算 得到的。它与碱基质量有关，受测序仪本身、测序试剂、样品等多个因素共同影响。测序错误率分布检查用于检测在测序长度范围内，有无异常的碱基位置存在高错误率，一般情况下，每个碱基位置的测序错误率都应该低于1%。

2.1.2.3 检查 GC 含量分布

该检查主要检测有无 AT、GC 分离现象，理论上 A 和 T 碱基及 C 和 G 碱基在每一测序循环上应该分别相等，但在实际测序过程中，会由于 DNA 模板扩增偏差、前几个碱基测序质量较低等原因，导致每个 read 前几个碱基波动较大，属于正常情况。

2.1.2.4 测序数据质量分布

依照测序技术特点，测序片段末端碱基质量一般较前端低。测序数据的质量主要分布在 Q30≥85%以上时，才能保证后续分析正常进行。

2.1.3 测序深度及覆盖度统计

有效测序数据通过 BWA (Li et al., 2018) 比对到参考基因组(GRCh37/hg19/GRCh38)，得到 BAM 格式的最初的比对结果。然后，用 Sambamba (Tarasov et al., 2015)对比 对结果进行排序并标记重复 reads(mark duplicate reads)。最后，利用重复标记后的 比对结果进行覆盖度、深度等的统计。通常，人类样本的测序 reads 能达到 95%以上的比对率;当一个位点的碱基覆盖深度(read depth)达到 10X 以上时，该位点处检测出的 SNP 比较可信。

2.2 变异检测结果

在最初的比对结果(BAM 文件)的基础上，利用 SAMtools(Li et al., 2009)识别 SNP 和 InDel 位点，统计基因组不同区域上 SNP 和 InDel 数目，编码区上不同类型 SNP 和 InDel 数目，转换和颠换的类型分布，SNP 和 InDel 数目及基因型分布。germline SNP 和 InDel 过滤参数如下:QUAL≥20;DV≥4;MQ≥30。

利用 Control-FREEC(Boeva et al., 2012)检测 CNV 的增加和减少，统计不同类型的 CNV 事件数目。利用 Lumpy 软件(Layer et al., 2014)检测 SV，并统计不同类型的 SV 事件数目。

2.3 注释

ANNOVAR (Wang et al., 2010) 是一种高效的软件工具，它能利用最新的信息，对由多个基因组检测出的基因变异进行功能注释。

利用 ANNOVAR 对先前工作中获得的 vcf(variantcallformat)进行注释。

(1) 使用 Refseq(O'Leary et al., 2016)注释变异位点的基因结构，基因类型包括 mRNA、非编码 RNA 等;

(2) 变异位点的基因组特征，对于位于基因组重复区段内的突变需谨慎对待;

(3) 通过 SIFT(Ng et al., 2003)、PolyPhen (Adzhubei et al., 2013)以及 MutationTaster (Reva et al., 2011)等方法全面评估非同义突变对疾病/肿瘤的影响;

(4) 提供了 dbSNP(Sherry et al., 2001)、千人基因组 SNP 数据库(Abecasis et al., 2012)、COSMIC(Tate et al., 2019)已知肿瘤体细胞突变数据库和 esp6500 变异数据库等注释，对变异结果可以进行任何组合的筛选;

(5) 注释内容还包括对突变所在基因进行功能注释，使用的数据库包括 GO(Lee et al., 2004)、KEGG(Kanehisa et al., 2000)、Reactome(Jassal et al., 2020)、Biocarta、 PID(Schaefer et al., 2009)等。

3 参考文献

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)