**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**

**目标区域测序（疾病）**

**1.实验流程**

**1.1 样品DNA质量检测（Evaluation of DNA quality）**

采用以下两种方法进行DNA质检：

1. 琼脂糖凝胶电泳分析DNA降解程度以及是否有RNA、蛋白质污染；
2. Qubit 3.0对DNA浓度进行精确定量。

其中DNA浓度≥20 ng/µL，总量0.5 µg以上的DNA样品被用来建库。

**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180-280 bp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DNA 3’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 目标区域杂交捕获（Library Hybridization with Exome Array）**

使用Agilent 设计的试剂盒，将接头文库与含生物素标记的探针库进行液相杂交，通过核酸碱基互补的原理，使得探针与目标DNA片段进行结合，再使用链霉亲和素磁珠与该杂交混合液混合，使链霉亲和素磁珠与含生物素的目标片段牢固结合，从而捕获到基因的外显子。经过进一步清洗，去除和磁珠非特异性结合的DNA后，文库中属于外显子区域的DNA得到富集。

**1.9 PCR扩增目标区域DNA文库 (PCR Amplification)**

在50 μL反应体系中应用高保真的聚合酶扩增原始文库，以保证足够的外显子文库总量。PCR扩增循环数控制在10-12之间。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。扩增后的外显子文库经过磁珠纯化即成为可以上机的测序文库。

**1.10 文库库检（Library Quality Assessment）**

文库构建完成后，先使用Qubit 3.0进行初步定量，随后使用NGS3K/Caliper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（3 nM）进行准确定量，以保证文库质量。

**1.11 桥式PCR**

即将捕获后的文库种到Flow Cell芯片上进行扩增的过程。Flow Cell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Flow Cell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。

**1.12 Illumina平台PE150上机测序（sequencing）**

PE150即Pair end 150 bp，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150 bp，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，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**

a.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。

b.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%。

c.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。

d.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≥80%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 H et al.)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SAMtools(Li H et al.)对比对结果进行排序；再用Sambamba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**

**2.2.1 SNP检测结果**

通常，一个人全基因组内会有约3.6~4.4 M个SNP，绝大数（大于95%）的高频（群体中等位基因频率大于5%）的SNP在dbSNP（Sherry S T et al.）中有记录，高频的SNP一般都不是致病的主要突变位点。

在最初的比对结果（BAM文件）的基础上，利用SAMtools识别SNP位点，对其进行统计及注释。统计基因组不同区域上SNP数目，编码区上不同类型SNP数目，转换和颠换的类型分布，SNP数目及基因型分布。利用ANNOVAR（Wang K et al.）软件对SNP进行注释，其中包括dbSNP数据库、千人基因组计划和其他已有的数据库的注释信息，注释内容包括7个部分，分别为优先级信息，基因及区域注释，数据库（频率）注释，保守（有害）性预测，变异位点信息，基因功能及通路注释，基因的组织特异性表达情况。

**2.2.2 InDel检测结果**

InDel（insertion and deletion），即插入和缺失，通常在一个人的全基因组中约有350 kb的InDel，约90%的InDel在dbSNP中有记录。在编码区或剪接位点处发生的InDel都可能会改变蛋白的翻译。发生移码变异，其插入或缺失的碱基串的长度为3的非整数倍，可能导致整个读框的改变；非移码变异即InDel长度为3的整倍数，编码区和剪接位点的读框不发生移码。前者较后者对基因功能影响更大。

在比对结果的基础上，我们利用SAMtools识别InDel，并采用国际惯用的过滤标准对InDel结果进行过滤。利用ANNOVAR（Wang K et al.）软件对InDel进行注释，其中包括dbSNP数据库、千人基因组计划和其他已有的数据库的注释信息，注释内容包括7个部分，分别为优先级信息，基因及区域注释，数据库（频率）注释，保守（有害）性预测，变异位点信息，基因功能及通路注释，基因的组织特异性表达情况。

**2.3 注释**

ANNOVAR（wang k et al.）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。

**参考文献**

[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 [J]. 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, 30(20):2852-9.