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

全基因组测序

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。其中含量在 0.3μg 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 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 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的产物，避免 PCR 过程的无效扩增，消除对上机测序的影响。

1.7PCR 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 Qubit 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，先使用 Qubit 2.0 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。

1.9 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 芯片上具有 8 条通道，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，合成一条与原来的 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。

1.10Illumina 平台 PE150 上机测序（sequencing）

PE150 即 Pair end 150bp，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

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），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

2.2.1SNP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 SNV 数目，编码区上不同类型 SNV 数目，转换和颠换的类型分布，SNV 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。

2.2.2InDel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 350kb 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 InDel，并采用国际惯用的过滤标准对 InDel 结果进行过滤。利 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。

2.2.3CNV检测

CNV 即 copy number variation 的缩写。拷贝数变异指的是基因组上大片段序列拷贝数的增加或者减少，可分为缺失（deletion）和重复(duplication)两种类型，是一种重要的分子机

制。CNV 能够导致孟德尔遗传病与罕见疾病，同时与包括癌症在内的复杂疾病相关，因此，染色体水平的缺失、扩增的研究已经成为疾病研究热点。采用 control-FREEC(Boeva V *et al.*)软件，通过检测样本在一个参考基因组上 reads 的深度分布情况来检测 CNV。利用 ANNOVAR 软件对拷贝数变异进行注释，注释信息包括 4 个部分，分别为基因及区域注释，数据库注释，ENCODE 注释，结构变异信息等。

2.2.4SV检测

SV 是 structural variation 的缩写。结构变异指的是在基因组上一些大的结构性的变异，比如大片段丢失（deletion）、大片段插入（insertion）、大片段重复（duplication）、拷贝数变异（copy number variants）、倒位（inversion）、易位（translocation）。结构变异普遍存在于人类基因组中，一般来说涉及的序列长度在 1kb 到 3Mb 之间，是个人差异和一些疾病易感性的来源。结构变异还可能导致融合基因的发生，一些癌症的发生已经证实和结构变异导致的基因融合事件有关。采用 CREST(Wang J et al.)对 SV 信息进行检测，利用 ANNOVAR 软件对结构变异进行注释，注释信息包括 4 个部分，分别为基因及区域注释、数据库注释、Encode 注释、结构变异信息等。

2.3 高级分析

2.3.1 突变位点筛选

a.去除在千人基因组数据（1000g_all）、ESP6500 数据库（esp6500siv2_all）、gnomAD 数据（gnomAD_ALL 和 gnomAD_EAS）这四个频率数据库中至少有一个频率高于 1%的突变。旨在去除个体间的多样性位点，得到真正可能致病的罕见突变（rare）

b.保留外显子区（exonic）或剪接位点区（splicing，上下 10bp）的变异

c.去除未被软件预测为会影响剪接且处于非高度保守区的同义 SNP 突变；去除处于 Repeat 区的小片段（<10bp）非移码 InDel 突变

d.依据 SIFT，Polyphen，MutationTaster，CADD 这 4 个软件的打分预测情况进行变异位点筛选，要求这 4 个软件中，有分值的软件中至少有一半支持该位点可能有害，该位点被保留；保留与外显子区距离不大于 2bp（±1~2bp）的剪接位点区突变；保留 dbscSNV 预测为会影响剪接的突变。（举例：一个位点的预测结果'SIFT=0.0,D', 'Polyphen=0.923,D 0.999,D', 'MutationTaster=1.000,N', 'CADD=.', 'dbscSNV_score=.', 那么该位点处支持有害的软件比例为 2/4,该位点被保留；一个位点的预测结果为'SIFT=.', 'Polyphen=.', 'MutationTaster=1.000,N', 'CADD=.', 'dbscSNV_score=0.5589,0.636',有害性预测软件不支持认为该位点有害，但 dbsSNV

的分值中有一个高于 0. 6，即软件预测该突变会影响剪接，该位点也被保留）

2.3.2突变位点有害性分类

2015 年美国医学遗传学和基因组学学会(ACMG) 开发了针对序列变异的解读的标准和指南，成为高通量测序后数据解读的金标准(Sue Richards *et al*,2015)。ACMG 开发的变异分类系统并推荐使用特定的标准术语，该系统将变异分为 pathogenic(致病的)、likelypathogenic(可能致病的)、uncertain significance(致病性不明确的)、likely benign(可能良性的)、benign(良性的) 来描述孟德尔疾病致病基因中发现的突变。 ACMG 的变异分类系统中共有 28 个证据类别，根据 28 个证据的组合形式进行变异位点的有害性分类，并对每个有害类别的变异位点数目进行统计。

2.3.3非编码区筛选

非编码区在基因表达调控中发挥着重要作用，另外非编码区的变异也会引起许多疾病的发生。在之前的研究中发现，有相当数量的病人可能存在外显子以外的致病突变。非编码区突变与疾病的密切关系，对非编码区进行研究非常有必要。基于非编码区变异位点进行筛选，具体的筛选过程如下：

1、Genomiser 筛选与疾病相关的非编码区遗传变异；

2、表观基因组学注释：注释相应组织 GTEx 和 Roadmap 数据库结果；

3、过滤筛选：（1）频率筛选：GnomAD_EAS 频率≤0.01 的位点；（2）位点保守型筛选：GWAVA 和 CADD 保守性分值筛选结果（GWAVA＞0.5，CADD≥10 认为有害）。SNP 筛选策略：过滤掉 CADD 或 GWAVA 有分值，且不认为有害的位点；InDel 筛选：保留 CADD 分值>10 的位点；（3）GTEx 数据库及表观数据库（Roadmap 和 Encode）过滤筛选: 筛选变异位点能够影响特定组织基因表达，且注释为 DNA 功能元件; （4）利用家系遗传模式进行筛选。

2.3.4结构变异CNV／SV有害性分析

与单核苷酸变异(SNVs, single nucleotide variatants)类似，很多 CNV／SV 是生物基因组中正常的多态性，这种良性 CNV／SV 不会导致生物体发生病变。但是，还发现有些恶性 CNV／SV 和神经系统障碍、癌症等疾病相关。为了从软件检测到的 CNV／SV 结果中进一步过滤掉良性 CNV／SV，以及保留恶性 CNV／SV，利用多种 CNV／SV 数据库对检测结果进行分类标记。分类标准如下：1、使用数据库 DGV 及其衍生系列 StringentLib，InclusiveLib 和 DGV.GoldStandard. July2015 对检测到的 CNV／SV 进行良性变异注释；2、

使用数据库 CNVD 对检测到的 CNV / SV 进行恶性变异注释；3、根据注释情况将 CNV / SV 分为 4 类，H(high)：恶有良无；P(possibly deleterious)：恶无良无；M(medium)：恶有良有；L(low)：恶无良有。若某变异能在两种不同的软件都能预测到则标为“*”，若某变异在"genomicSuperDups"或"Repeat"列中有标注则标为“-”。

2.3.5显隐性遗传模式筛选

显性遗传模式疾病的致病变异来自父亲或母亲，因此一般是杂合变异，首选考虑候选基因上的杂合变异，隐性遗传模式疾病，其致病变异一般来自父母双方，为纯合或者复合杂合变异。

a.显性遗传模式筛选

在突变位点过滤的基础上，如果某一单基因病在家系中为显性模式遗传，则保留家系中患者常染色体为杂合突变（性染色体保留有突变的位点），且家系中正常人没有突变的位点作为候选位点。

b.隐性遗传模式筛选

隐性遗传致病包括两种情况，基因纯合变异和复合杂合变异。在进行隐性模式筛选的时候，也考虑了 X 连锁的情况。基因纯合变异保留患者为纯合突变，且家系中正常人为杂合突变作为候选位点。复合杂合变异则保留患者和正常人都不为纯合突变的位点，且要求一个基因在患者中至少有两个杂合突变位点，且患者此基因上的突变位点分布不能与任何一个正常人（此基因）的突变位点分布一样，也不能是任何一个正常人（此基因）突变位点的子集。

2.3.6共有突变筛选

在过滤有害性位点的基础上，筛选 2 个以上患者共有的突变基因。

2.3.7新生突变筛选

对于“患者+双亲”的成三/成四家系，目标是：寻找父母中没有，而孩子出现的突变。应用两种方法进行 *de novo* SNP/InDel 筛选，1)基于 SAMtools 软件筛选出 *de novo* 突变；2)基于家系中每个样本的变异结果进行 *de novo* 位点筛查和过滤。其中用到的位点过滤条件是基于多篇参考文献和测序深度选择的(Zaidi S et al.2013，Sanders S J et al.2012)。

a. SAMtools 方法筛选（de novo SNP/InDel）

在比对结果的基础上，基于家系 trio 成员的 jointly analysis，使用 SAMtools *de novo* mutation 分析方法，得到患者中有，患者父母中没有的 *de novo* mutation 位点，再经过突变位点筛选得到最终的候选位点。

b. 基于家系中单个样本方法筛选（*de novo* SNP/InDel）

在得到每个家系成员 SNV，InDel 等突变信息的基础上（基本分析结果基础上），筛选在患者中有，而患者父母没有的突变位点作为最初的新生突变，经过突变位点筛选得到最终的候选突变位点（筛选过程同高级分析第一部分所示）。

d.在 a 和 b 两种方法结果的基础上，取两种方法的交集。

e.计算候选致病新生突变的速率（*de novo* mutation rate），评估这些突变在患病群体中发生的偏好性，验证该疾病的发生与新生突变的发生密切相关。

f.在得到每个家系成员 CNV / SV 突变信息的基础上（基本分析结果基础上），根据患者和父母的 CNV / SV 的差异来筛选患者的 *de novo* CNV / SV，筛选流程如下：

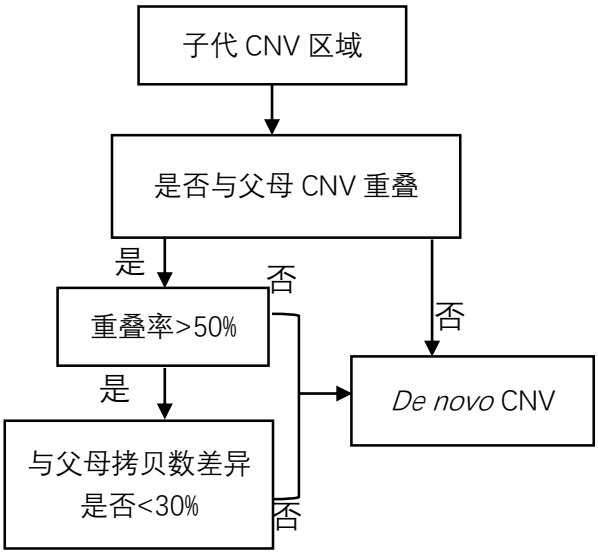

图 1.*de novo* CNV 筛选流程

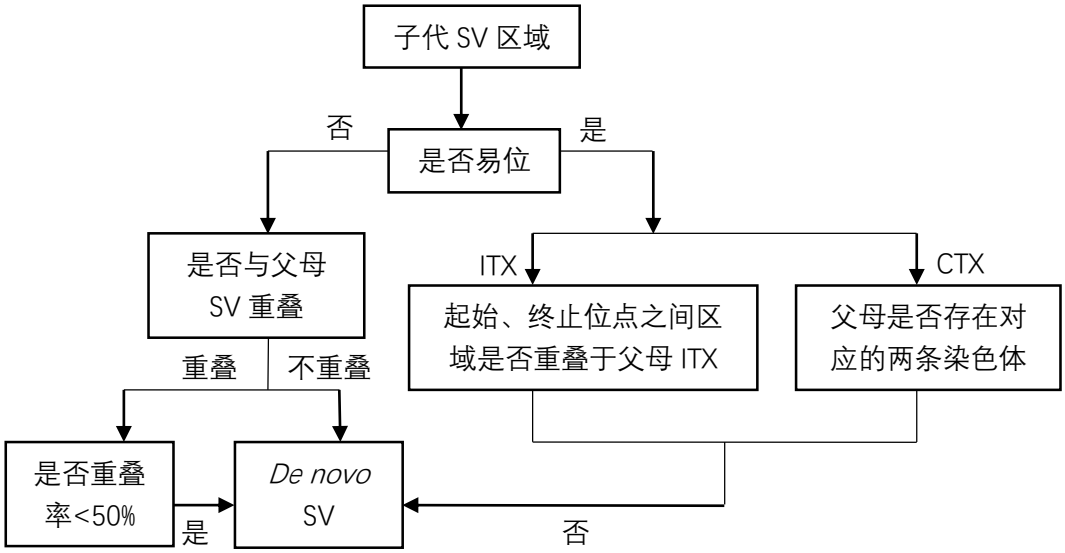

图 2.*de novo* SV 筛选流程

2.3.8连锁分析

以两代或两代以上的家系为材料基础，利用 merlin 工具和 perl 语言，优势对数计分法（LOD），结合家系中高通量测序数据和 HapMap 数据库中中国人群（CHB）的等位基因频率，使用已知的 SNP 作为连锁分析标记，观察标记位点与疾病致病基因位点在家系内是否共分离，并计算出遗传距离及连锁程度，LOD 值代表两位点连锁的概率与不呈连锁的概率比的对数值。对于单基因遗传病，LOD 值大于 3 时肯定连锁，LOD 值<-2 时否定连锁，LOD 值介于 1 与-2 之间的则需要增加家系材料进一步分析或者优先验证这个候选区域内的变异位点。对于复杂疾病，LOD 对应的阈值偏低。

2.3.9ROH分析

纯合子区域即基因组上的纯合等位基因的区域。这个区间的产生是由于父母遗传至子代的等位基因来自于共同的祖先，即父母拥有同一祖先的等位基因，遗传给子代的便是纯合子。一般用于近亲结婚引起的遗传病，利用 H3M2 软件，生成每个家系下样本的 ROH 区域，并对每个样本的 ROH 区域进行注释。

2.3.10候选基因GO／KEGG富集分析

GO 功能富集分析包含 CC（细胞组件）、BP（生物学途径）、MF（分子功能）三种结果。用 clusterProfiler 包进行 GO 富集分析，利用 R 软件的超几何分布法对候选基因进行 KEGG 富集分析。

2.3.11 蛋白功能互作分析

使用 genemania 在线软件 GeneMania（Warde-Farley D et al.2010-7）对候选基因进行蛋白功能互作网络分析，包括 protein-protein, protein-DNA-genetic interactions, pathways, reactions,gene-protein expression data, protein domains-phenotypic screening profiles。

2.3.12 基因-疾病表型关联性分析

2.3.12.1DisGeNet 数据库注释

DisGeNet 数据库是一个专注基因-疾病关联以及突变位点-疾病关联的数据库，该数据库目前版本为 V5.0， 包含 561,119 条基因-疾病关联记录（GDAs）和 135,588 突变-疾病关联记录（VDAs）。使用该数据库中的关联数据，通过疾病获取相关基因以及变异信息，用于后续分析。

2.3.12.2Phenolyzer 分析结果

依据疾病/表型名称，通过精准算法，结合测序结果和多种数据库，对基因进行筛选排序，构建基因-疾病表型之间的关联图。

2.3.12.3 候选基因疾病关联性排序

针对筛选出的候选基因，依据其与疾病的关联性强弱进行排序。

2.3.13关联分析

2.3.13.1位点的（site-based）关联分析

（1）质控

a)样本质控

I)为排除由于人群分层因素导致的假阳性结果，对样本进行 PCA 分析(Principal Component Analysis, PCA) ，画出 PCA 图，分析是否存在人群分层。

II)为排除样本中由于有亲缘关系个体的存在导致的假阳性或者假阴性结果，需要对每个个体的 IBD(identity-by-descent)共享的可能性即亲属系数进行了评估，给出血缘关系系数，排除亲缘关系样本。

III)排除测序质量异常的样本：根据测序质量（Q20，Q30），PE reads mapping 率，覆盖度，有效深度等综合判断。

b)位点质控

I）评价 Hardy-Weinberg Equilibrium，给出相应 P 值；

II）Call rate：保留 case、control 样本中位点分型率都大于 90%的点。

（2）对 SNP位点进行关联分析

利用 PLINK 软件（http://pngu.mgh.harvard.edu/~purcell/plink/）（Purcell et al.2007）对单核苷酸多态性位点进行关联分析，寻找 case 相对 control 存在显著性差异的位点：计算每个 SNP 位点的 p 值和 OR 值，通过关联的显著度（P-value），筛选显著性关联的 SNP 变异。

（3）候选致病突变位点注释

一般认为，致病（causal）的基因和我们找出的显著性关联 SNP 位于同一个单体型区段（LD Block）中，或者简单理解为在染色体上的物理距离较近。根据关联分析结果，对显著关联 SNP 位点所在物理位置上下游一定区域内（如 50kb）的相关基因或者物理距离最近的某几个基因上的所有 SNP 进行注释，注释内容包括变异的位置信息及所在基因的基本信息，1000G 中的频率信息及突变有害性分值等内容。

（4）低频有害突变过滤（仅适用于RVAS）

a) 低频突变位点过滤

过滤千人基因组数据库、Novo-zhonghua 数据库、GnomAD_All 和 GnomAD_EAS 数据库，去除个体间多样性位点，得到真正可能致病的罕见突变（rare）：保留 1000g_All&Novo-zhonghua<0.01 且 GnomAD_All & GnomAD_EAS <0.001。

b) 突变位点有害性过滤

显著性关联（associated）SNP 周边潜在致病（causal）基因上的突变并不一定均是有害的，所以需要进行以下过滤得到致病（causal）变异：

I）保留外显子区（exonic）或剪接位点区（splicing，上下 10 bp）的变异；

II）去除同义突变（不导致氨基酸编码改变的突变），得到对基因表达产物有影响的突变；

III）依据 SIFT、Polyphen、MutationTaster、和 CADD 这 4 个软件，要求这 4 个软件中，至少有一半支持该位点可能有害，该位点才被保留（举例：一个位点的预测结果为 'SIFT=0.07,T', 'Polyphen2-HVAR=0.923,D, Polyphen2-HDIV = 0.999, D', 'MutationTaster=1.000, N','CADD=.'，那么该位点处支持有害的软件比例为 1/3，不到一半，该位点会被丢弃）；

2.3.13.2基因关联分析(Burden analysis)

（1）样本质控

a）为排除由于人群分层因素导致的假阳性结果，对样本进行 PCA 分析（Principal Component Analysis），画出 PCA 图，分析是否存在人群分层。

b）群体中亲缘关系相近的个体将会影响下游关联分析，导致假阳性或假阴性结果的出现。为了检查个体间的亲缘关系，我们对每个个体的 IBD(identity-by-descent)共享的可能性及亲属系数进行了评估，计算血缘关系系数，排除具有亲缘关系的样本。

c）排除测序质量异常样本：根据测序质量（Q20,Q30），PE reads mapping 率、覆盖度，有效深度等综合判断。

（2）位点质控

保留 case、control 样本中位点分型率(call rate)都大于 90%的点。

（3）有害性过滤

a) SNP 过滤：

I. 保留外显子区和剪切位点区（2bp）的变异位点；

II.去除同义突变，stoploss 突变；

III. 保留低频的突变位点 (1000g_All&Novo-zhonghua<0.01&

GnomAD_All&GnomAD_EAS <0.001）；

IV.过滤掉位于 Repeat 或者 genomicSuperdup 区的突变；

V.有害性筛选：保留 SIFT、Polyphen、MutationTaster、CADD 中超过 2 个预测为有害的位点。即这 4 个软件中，至少有一半支持该位点可能有害，该位点才被保留（举例：一个位点的预测结果为'SIFT=0.07, T', 'Polyphen2-HVAR=0.923, D, Polyphen2-HDIV = 0.999, D','MutationTaster=1.000, N', 'CADD=.'，那么该位点处支持有害的软件比例为 1/3，不到一半，该位点会被丢弃）；

VI.保守性过滤：保留 gerp++gt 分值大于 2 的位点。

b) InDel 过滤：

I.保留外显子区和剪切位点区（2bp）的变异位点；

II.去除 nonframeshift，stoploss 突变；

III. 保留低频的突变位点 (1000g_All&Novo-zhonghua<0.01& GnomAD_All&GnomAD_EAS < 0.001）；

IV.过滤掉位于 Repeat 或者 genomic Superdup 区的突变；

（4）Burden 分析

经过突变过滤之后，可以得到与疾病相关联的候选基因及相应有害性突变位点，通过 Burden 或 SKATO 分析方法（Wu et al, 2011），进行以基因为单位的显著性分析。