**人类全外显子组上游分析方法（示例）**

对于测序获得的原始数据，首先使用Trimmomatic v0.361进行质量控制，去除接头序列及低质量碱基序列，双端数据的过滤参数设置为："LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36"。获得高质量测序数据后，使用BWA v0.7.182的mem模块将序列比对至人类参考基因组（GRCh38），并通过SAMtools v1.213的sort功能对生成的SAM文件进行排序后，得到BAM文件用于后续分析。

随后进行变异位点分析。首先使用Picard v0.3.3.04的MarkDuplicates功能标记并去除PCR扩增重复，设置参数为："REMOVE\_DUPLICATES=true"。接着，基于GATK4 v4.6.1.05进行SNP和InDel的鉴定。具体流程包括：（1）使用BaseRecalibrator和ApplyBQSR模块进行碱基质量校准（Quality score recalibration, BQSR）；（2）通过HaplotypeCaller在GVCF模式下（"--emit-ref-confidence GVCF"）筛选每个样本中的短变异（Call variants），以便后续多样本合并分析；此外上述两步都使用了人类已知变异位点的vcf文件（来自Ensembl release-110）作为校准和引导变异检测的参考（分别使用参数"--known-sites" 和 "--dbsnp"指定）。（3）在多样本情况下，使用GenomicsDBImport模块合并所有样本GVCF文件，并通过GenotypeGVCFs模块进行联合基因分型；（4）使用GATK的变异质量评分校正（Variant Quality Score Recalibration, VQSR）对已获得的样本变异结果进行筛选。基于HapMap6、Mills7、Omni8、1000G8和dbSNP9等人类高置信度变异数据，通过VariantRecalibrator模块整合DP、QD、MQ、MQRankSum、ReadPosRankSum、FS及SOR等质量指标，分别构建SNP和InDel的训练模型，设置正模型最大高斯数为 6（--max-gaussians 6）。最后以99%真集敏感度阈值（--truth-sensitivity-filter-level 99.0），利用 ApplyVQSR 模块对两类变异筛选，输出最终结果。

使用VEP v113.310对SNP位点的进行注释分析，识别变异发生的区域并预测其对基因功能的潜在影响。注释信息来自Ensembl GRCh38 release-110。

【结果数据统计使用的软件和方法此处未列出，请根据需要自行添加，参考文件见Citation-and-References.md】

1 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

2 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

3 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

4 *Picard toolkit*, <<https://broadinstitute.github.io/picard/>> (2019).

5 McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

6 Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58, doi:10.1038/nature09298 (2010).

7 Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59-65, doi:10.1038/nature09708 (2011).

8 Altshuler, D. M. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-+, doi:10.1038/nature15393 (2015).

9 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311, doi:DOI 10.1093/nar/29.1.308 (2001).

10 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, doi:ARTN 122 10.1186/s13059-016-0974-4 (2016).

英文版

**Raw Data Processing​**  
 Raw sequencing data underwent quality control using Trimmomatic v0.36 to remove adapter sequences and low-quality bases, with paired-end read filtering parameters: "LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36". High-quality reads were aligned to the human reference genome (GRCh38) using BWA-MEM v0.7.18, and SAM files were subsequently sorted with SAMtools v1.21 to generate sorted BAM files.

**Variant Calling Pipeline**

PCR duplicates were marked and removed using the MarkDuplicates function in Picard v3.3.0 with the parameter "REMOVE\_DUPLICATES=true". Then SNP and InDel identification was performed using GATK4 v4.2.6.1 as follows: (1) Base Quality Score Recalibration (BQSR) using the BaseRecalibrator and ApplyBQSR modules; (2) Short variants were called in each sample using HaplotypeCaller in GVCF mode ("--emit-ref-confidence GVCF") for subsequent multi-sample merging analysis. Additionally, both steps utilized a VCF file of known human variants (from Ensembl release-110) as a reference for calibration and guiding variant detection, specified using the parameters "--known-sites" and "--dbsnp," respectively. (3) In the case of multiple samples, the GenomicsDBImport module was used to merge all sample GVCF files, followed by joint genotyping using the GenotypeGVCFs module; (4) The Variant Quality Score Recalibration (VQSR) method from GATK was used to filter the obtained sample variant results. Based on high-confidence human variant data from HapMap, Mills, Omni, 1000G, and dbSNP, the VariantRecalibrator module integrated quality metrics such as DP, QD, MQ, MQRankSum, ReadPosRankSum, FS, and SOR to construct training models for SNPs and InDels separately, with the max number of Gaussians for the positive model set to 6 (--max-gaussians 6). Finally, high-quality variants were obtained using the ApplyVQSR module with the parameter "--truth-sensitivity-filter-level 99.0".

**Variant Annotation**  
 The Variant Effect Predictor (VEP) v113.3 was used for SNP annotation, with the annotation version matching the GRCh38 genome (Ensembl release-110), identifying genomic regions impacted and predicting protein functional impacts.