BIOF3 组学数据分析

基因组学实践教程

导出日期：2026年6月27日

基因组学实践教程

基因组学研究的是 DNA 序列本身：谁的基因组里有什么变异，这些变异是否落在功能区域，以及在群体里如何分布。即便今天大多数湿实验都在转录组或单细胞这一层，基因组重测序仍然是找到疾病相关变异、做 GWAS、做群体遗传分析的起点。

本专栏打算把"拿到一批测序数据之后怎么走到变异列表"这条主线讲清楚。重点放在重测序和短读长小变异，其他方向（长读长、结构变异、甲基化等）会作为延伸而不是主干。

基因组学项目的三条典型主线

新手容易把"基因组学"当成一回事，其实它是三种很不同的项目类型：

项目类型	核心问题	数据特点	关键工具
胚系变异检测	这个人有什么遗传缺陷	单/几个样本、深度 30-50x	GATK HaplotypeCaller + ClinVar
肿瘤体细胞变异	肿瘤里有什么驱动突变	配对 tumor + normal	Mutect2 + maftools
群体遗传 / GWAS	这个变异在群体里怎么分布	几千-几十万样本	PLINK 2

这三条主线分析流程差很多：胚系关注致病性、肿瘤关注 VAF 和驱动基因、群体关注频率和 LD。**先想清楚自己在做哪一种**，再选工具，比闷头跑流水线效率高。

一个项目大致是什么样子

一个典型的 WES 或 WGS 项目会拿到若干个体的双端 FASTQ 数据。从这里到"一份可以注释、可以过滤、可以做关联分析的 VCF"之间，要走的步骤大致是：

步骤	典型产物	常用工具
质控与接头剪切	清洗后的 FASTQ	FastQC、fastp、Trim Galore
参考基因组比对	sorted.bam + 索引	BWA-MEM、minimap2
重复序列标记与 BQSR	校正后的 BAM	GATK MarkDuplicates、BQSR
变异检测	原始 VCF / gVCF	GATK HaplotypeCaller、DeepVariant
联合基因分型	多样本 VCF	GATK GenotypeGVCFs
变异过滤	高置信变异集	GATK VQSR、硬过滤
变异注释	带功能注释的 VCF	VEP、ANNOVAR、SnpEff
下游分析	关联、群体结构、可视化	PLINK、R

"BWA + GATK" 是短读长小变异的事实标准，文献、社区和官方 best practices 都围绕它展开。如果只做基因型分型而不做科研级变异检测， `bcftools mpileup + bcftools call` 也能跑出可用结果，代码更短。

常见工具栈

这套组合在教学和小到中等规模真实项目里基本够用：

阶段	工具	运行环境
质控	FastQC、fastp、MultiQC	bash
比对	BWA-MEM、minimap2	bash
BAM 处理	samtools、GATK	bash
变异检测	GATK HaplotypeCaller、DeepVariant	bash
变异过滤	GATK VQSR、bcftools	bash
注释	VEP、ANNOVAR、SnpEff	bash
群体与统计	PLINK 2、vcftools、R	bash + R

长读长（PacBio、ONT）和结构变异会用到另一套工具链（minimap2、Sniffles、CuteSV 等），这里暂不展开。

推荐公开数据集

学这个专栏可以从三类数据入手：

数据集	规模	适合	入口
NA12878 / HG001	单样本，广泛验证	流程搭建、variant calling 基准	Genome in a Bottle
1000 Genomes	人群规模	群体结构、LD、祖源分析	1000 Genomes Project
GnomAD	汇总等位基因频率	变异注释、罕见变异过滤参考	gnomAD

NA12878 有 GIAB 发布的"黄金标准" VCF，可以直接用来评估你自己跑出的结果是否合理，是练习 variant calling 流程最方便的参照物。

最小可跑的例子

下面用 Bioconductor 的 `VariantAnnotation` 包做一次最简单的 VCF 读取和浏览。它自带一份 `chr22.vcf.gz` 示例，不需要联网下载：

```r
if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")
BiocManager::install(c("VariantAnnotation", "TxDb.Hsapiens.UCSC.hg19.knownGene"))

library(VariantAnnotation)
library(TxDb.Hsapiens.UCSC.hg19.knownGene)

# 读入包里自带的 chr22 示例 VCF
fl <- system.file("extdata", "chr22.vcf.gz", package = "VariantAnnotation")
vcf <- readVcf(fl, "hg19")
vcf

# 看一下 VCF 结构
header(vcf)
rowRanges(vcf)[1:5]

# 提取基因型信息
geno(vcf)$GT[1:5, 1:5]

# 把变异定位到基因区域
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
loc <- locateVariants(vcf, txdb, CodingVariants())
head(loc)

# 统计每类变异的数量（内含子、外显子、UTR 等）
all_loc <- locateVariants(vcf, txdb, AllVariants())
table(all_loc$LOCATION)
```

`locateVariants` 会把每个 VCF 记录映射到最近的基因和区域类型。`table(all_loc$LOCATION)` 的输出是典型的"变异区域分布"结果——UTR、外显子、内含子、基因间各多少——这也是 variant calling 报告里必写的一栏。

真实重测序项目比这段要多两件事：前半段的 BWA 比对 + GATK 变异检测流水线（要跑几个小时、要有足够磁盘）、后半段的过滤策略和多样本联合分型。但 VCF 拿到手之后的操作思路，和上面这段是一致的。

专栏模块规划

模块	主题	状态
01	数据类型：WGS vs WES vs Panel	已上线
02	BWA / GATK 比对与变异检测	已上线
03	VCF 注释与可视化	已上线
04	maftools 肿瘤突变分析（WES）	已上线
05	变异过滤与质量评估	已上线
06	群体遗传：PCA / 祖源 / LD	已上线
07	GWAS 入门	已上线
08	临床变异解读与报告	已上线

所有 8 个模块已上线。03 和 04 带可跑脚本和真实数据图。

推荐前置知识

- 编程基础：R、Python、Bash 学习路径
- Jupyter 与交互式分析环境
- 公共数据库与数据检索

常见误区

误区 1：以为变异越多越好

WGS 跑出来 5M 变异，**99% 是已知 SNP，对你研究问题毫无意义**。真正有价值的是**罕见变异 + 致病变异**，需要靠 ClinVar / gnomAD 过滤。"我的样本有 5M 个变异" 不是亮点。

误区 2：用 SNV-only 思路做肿瘤分析

肿瘤里 CNV、SV、染色体重排同样关键 — 漏掉这些等于半残。**WGS 项目要补 Manta / GRIDSS（SV）和 ASCAT / FACETS（CNV）**。

误区 3：把胚系变异工具用在肿瘤上

`HaplotypeCaller` 假设变异在 50% 或 100% 的 reads 里（杂合或纯合）。**肿瘤是异质性群体，VAF 可以是 5%、20%、80% 等任意值**，必须用 Mutect2 或类似肿瘤专用 caller。

误区 4：不做 BQSR 直接 call 变异

BQSR 校正测序仪系统偏差，**不做的话假阳性率提高 10-30%**。GATK best practices 里这是必须步骤，不要图省事跳过。

误区 5：群体遗传项目不做 LD pruning 直接 PCA

PCA 会被高 LD 区域（HLA、着丝粒等）主导，主成分变成"局部 LD 模式"而不是真实的群体结构。**PCA 前必须 LD prune（PLINK `--indep-pairwise 50 5 0.2`）**。

参考资源

- GATK Best Practices
- Genome in a Bottle
- 1000 Genomes Project
- gnomAD
- Ensembl VEP
- samtools / bcftools 手册