BIOF3 组学数据分析

基因组学实践手册

BioF3 基因组学专栏导出版

导出日期：2026年5月12日

扫码关注微信公众号【生信F3】

获取文章完整内容，分享生物信息学最新知识。

BIOF3 组学数据分析

基因组学实践手册目录

BioF3 基因组学专栏导出版

01 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。

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：

```
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

参考资源

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

02 数据类型：WGS vs WES vs Panel

基因组测序有三种主要策略，选择取决于研究目的和预算：

策略	覆盖范围	数据量/样本	适用场景
WGS	全基因组 ~3Gb	30-60x, ~90GB FASTQ	结构变异、非编码区、群体遗传
WES	外显子区 ~60Mb	100-200x, ~6GB FASTQ	肿瘤体细胞突变、遗传病诊断
Panel	几十~几百个基因	500-1000x, ~1GB	临床检测、已知热点突变

WGS vs WES 的分析差异

维度	WGS	WES
变异类型	SNV + Indel + SV + CNV	主要 SNV + Indel
参考区域	全基因组	需要 BED 文件定义 target 区域
覆盖度均匀性	好	受捕获效率影响，边缘区域覆盖低
数据分析工具	GATK / DeepVariant	GATK + Mutect2（肿瘤）
下游重点	群体遗传、GWAS、SV	肿瘤驱动突变、maftools、OncoKB

肿瘤 WES 的特殊性

肿瘤样本通常配对测序（tumor + matched normal），用 Mutect2 做体细胞突变检测。输出的 MAF 文件是 maftools 的标准输入。

```
# Mutect2 典型调用
gatk Mutect2 \
  -R reference.fa \
  -I tumor.bam \
  -I normal.bam \
  -normal normal_sample_name \
  -O somatic.vcf.gz

# VCF -> MAF 转换
vcf2maf.pl --input-vcf somatic.vcf.gz --output-maf somatic.maf \
  --tumor-id tumor --normal-id normal --ref-fasta reference.fa
```

下一步

- 02 比对与变异检测流程
- 03 VCF 注释与可视化
- 04 maftools 肿瘤突变分析

参考资源

- GATK Best Practices
- Mutect2 文档
- vcf2maf

03 比对与变异检测流程

从 FASTQ 到 VCF 的标准流水线。这一章给出完整的 bash 命令序列，适合在服务器上跑。

BWA-MEM 比对

```
# 建索引（一次性）
bwa index reference.fa

# 比对（每个样本）
bwa mem -t 8 -R "@RG\tID:sample1\tSM:sample1\tPL:ILLUMINA" \
  reference.fa sample1_R1.fq.gz sample1_R2.fq.gz \
  | samtools sort -@ 4 -o sample1.sorted.bam

samtools index sample1.sorted.bam
```

`-R` 参数加 read group 信息，GATK 后续步骤必须有。

去重复 + BQSR

```
# 标记 PCR 重复
gatk MarkDuplicates \
  -I sample1.sorted.bam \
  -O sample1.dedup.bam \
  -M sample1.dedup_metrics.txt

# Base Quality Score Recalibration
gatk BaseRecalibrator \
  -R reference.fa \
  -I sample1.dedup.bam \
  --known-sites dbsnp.vcf.gz \
  --known-sites mills_indels.vcf.gz \
  -O sample1.recal_table

gatk ApplyBQSR \
  -R reference.fa \
  -I sample1.dedup.bam \
  --bqsr-recal-file sample1.recal_table \
  -O sample1.recal.bam
```

HaplotypeCaller（胚系变异）

```
# 单样本模式（输出 gVCF）
gatk HaplotypeCaller \
  -R reference.fa \
  -I sample1.recal.bam \
  -O sample1.g.vcf.gz \
  -ERC GVCF

# 多样本联合分型
gatk CombineGVCFs -R reference.fa \
  -V sample1.g.vcf.gz -V sample2.g.vcf.gz \
  -O cohort.g.vcf.gz

gatk GenotypeGVCFs -R reference.fa \
  -V cohort.g.vcf.gz \
  -O cohort.vcf.gz
```

Mutect2（体细胞变异）

```
gatk Mutect2 \
  -R reference.fa \
  -I tumor.recal.bam \
  -I normal.recal.bam \
  -normal normal_sample \
  --germline-resource gnomad.vcf.gz \
  -O somatic_unfiltered.vcf.gz

gatk FilterMutectCalls \
  -R reference.fa \
  -V somatic_unfiltered.vcf.gz \
  -O somatic_filtered.vcf.gz
```

流水线工具

手动跑上面这些命令容易出错。推荐用 nf-core/sarek：

```
nextflow run nf-core/sarek \
  --input samplesheet.csv \
  --genome GRCh38 \
  --tools mutect2,haplotypecaller \
  --outdir results/
```

参考资源

- BWA 手册
- GATK Best Practices
- nf-core/sarek

- DeepVariant

04 VCF 注释与可视化

VCF 文件拿到手之后，第一步是"每个变异落在什么基因、什么区域、有什么功能影响"。本章用 R 的 `VariantAnnotation` 包在内置 chr22 VCF 上演示。

真实示例

配套脚本 `genome03_variant_anno_sci.R` 输出 6 张图：

```
Rscript scripts/genomics/genome03_variant_anno_sci.R
```

每张图看什么

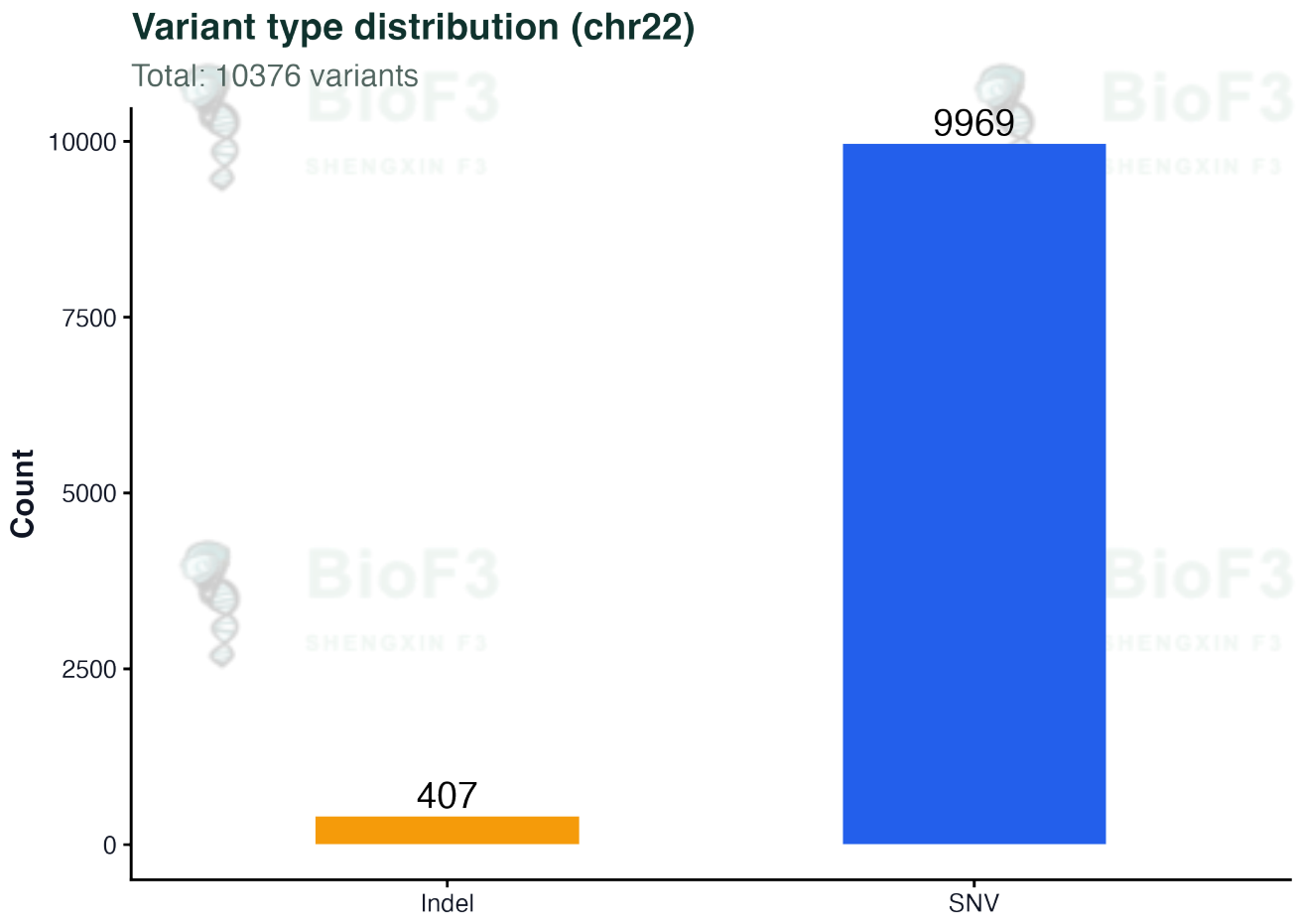

图 1：SNV vs Indel 的数量分布。WGS/WES 里 SNV 通常占 90%+。

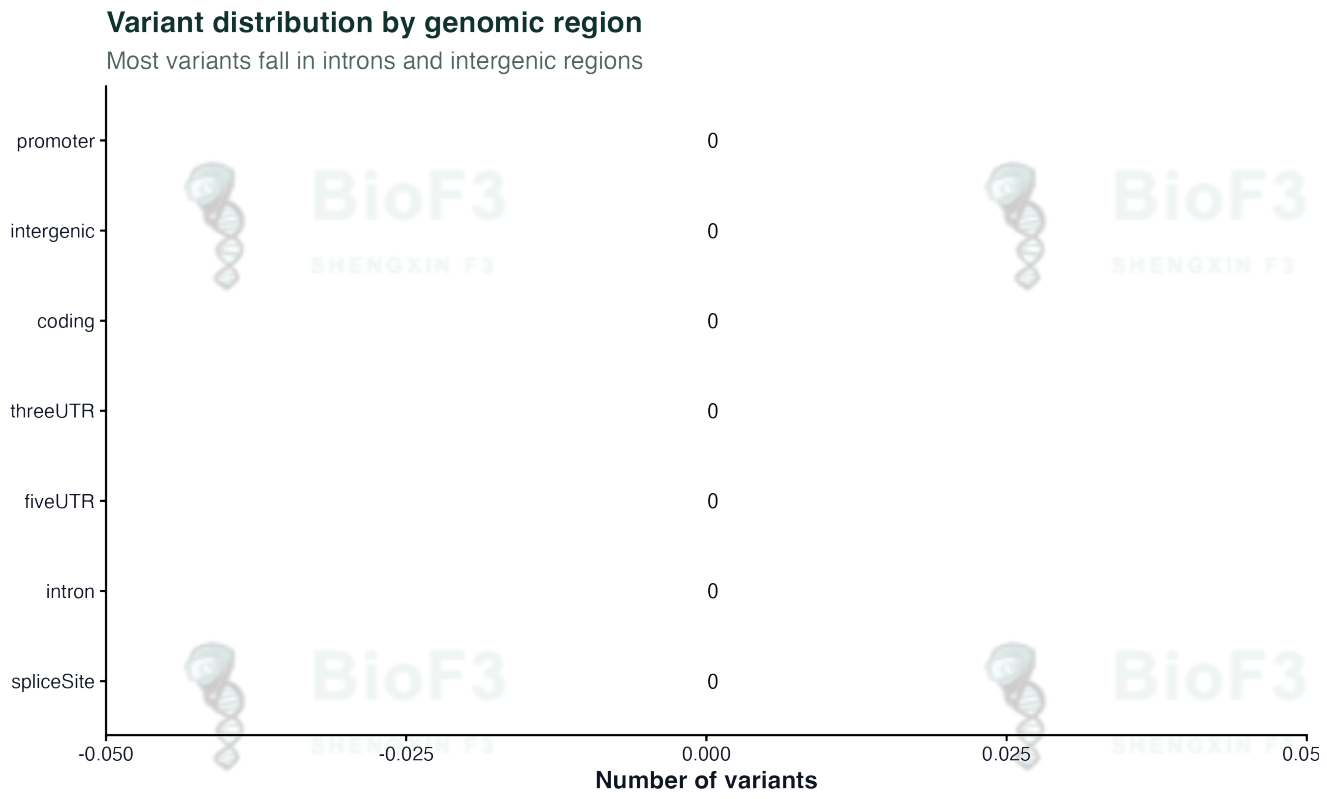

图 2：变异落在哪些基因组区域。大部分在内含子和基因间区（非编码区占基因组 98%+）。

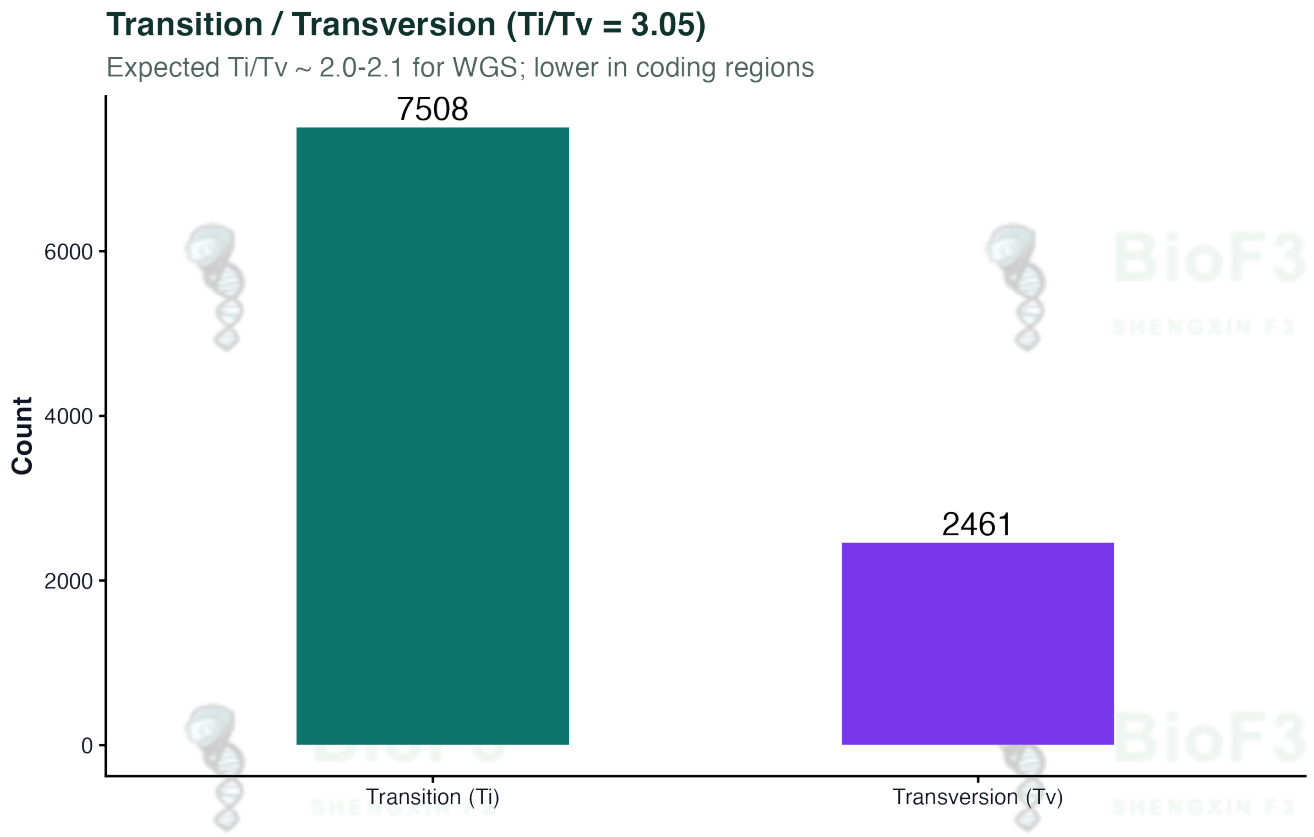

图 3：转换/颠换比。WGS 期望 ~2.0-2.1，WES 编码区 ~3.0。偏低可能说明假阳性多。

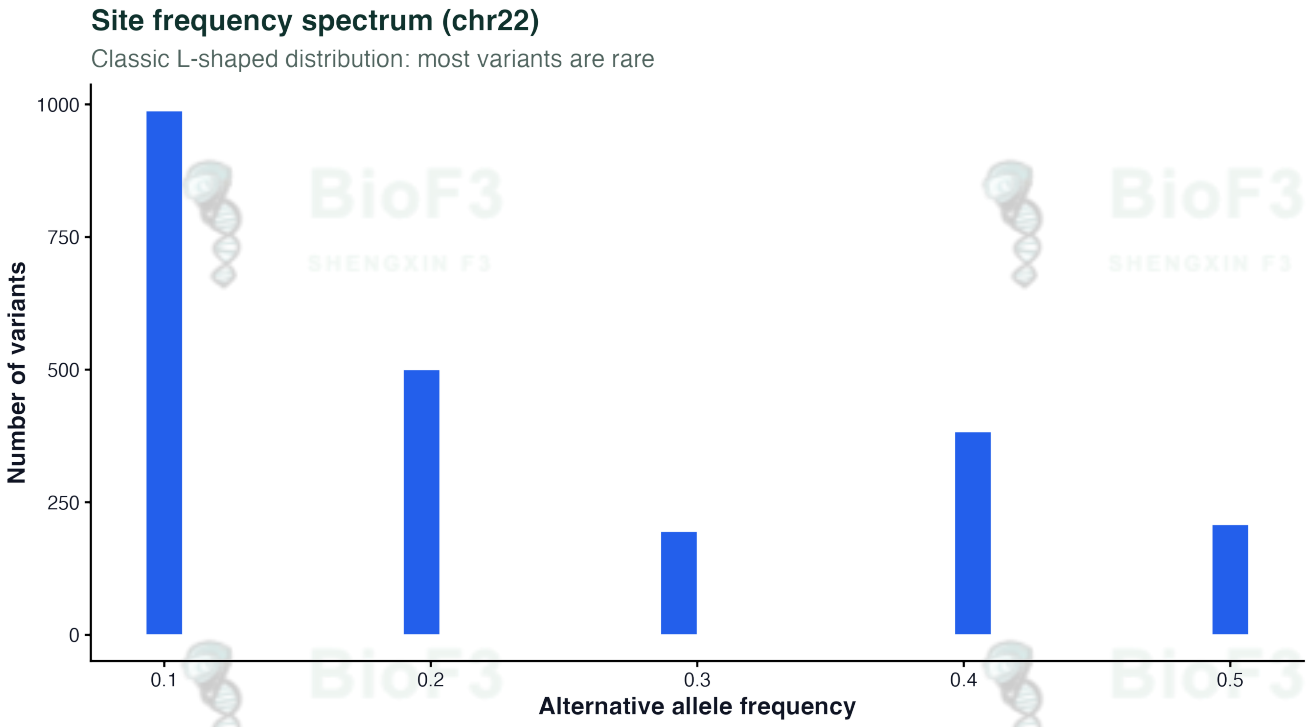

图 4：等位基因频率谱。经典 L 形：大部分变异是稀有的（低频）。

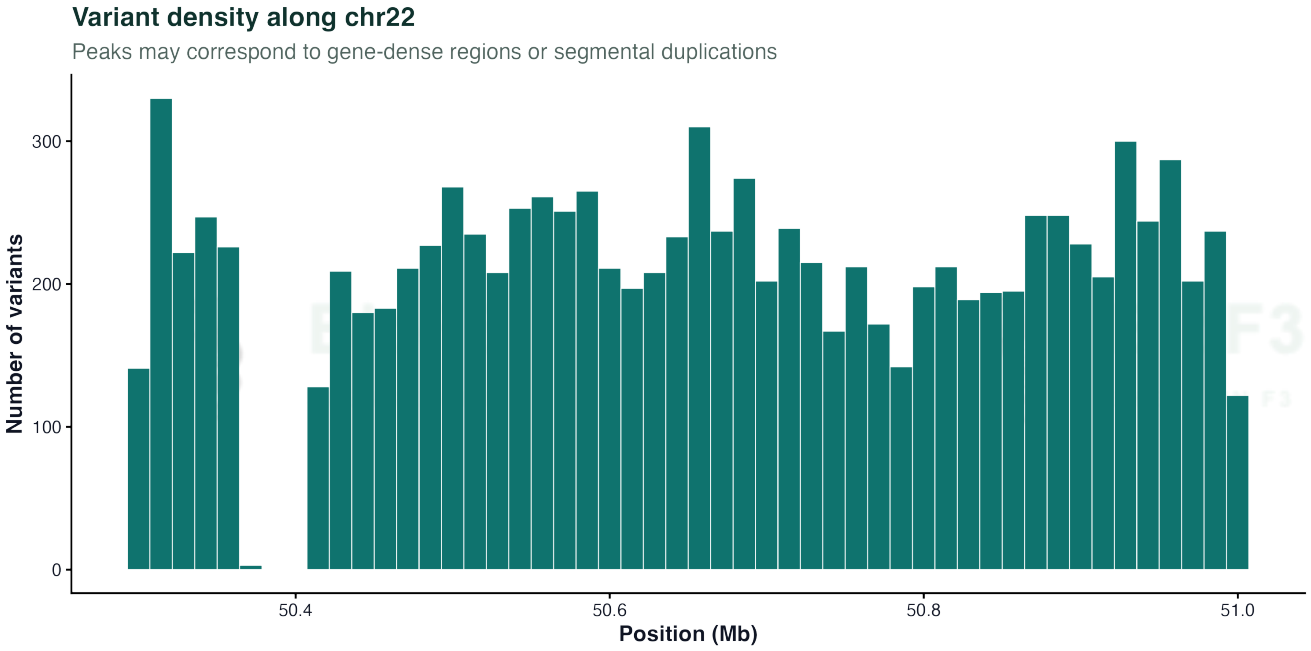

图 5：chr22 上的变异密度分布。某些区域密集可能对应基因密集区或重复序列。

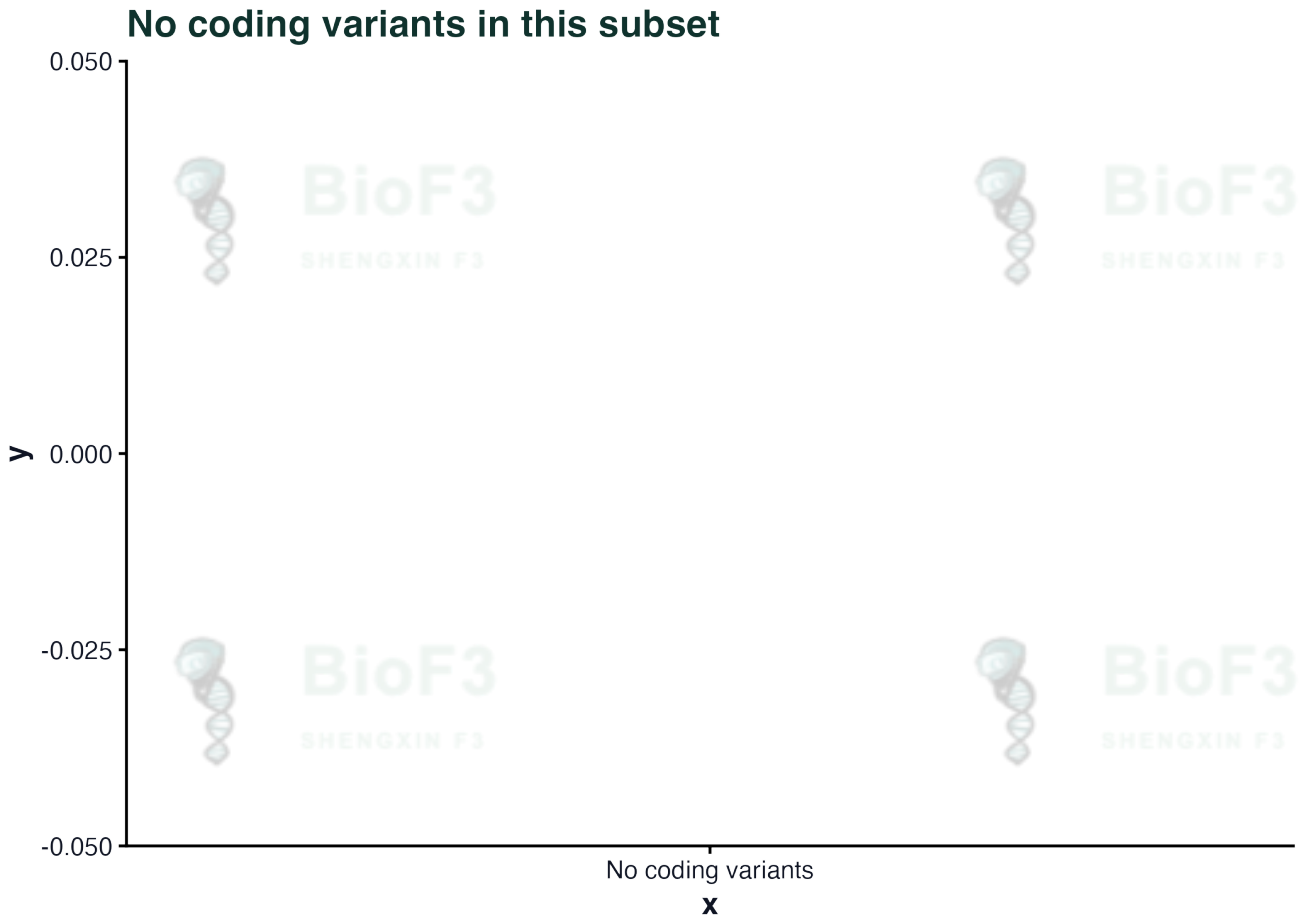

图 6：编码区变异的功能后果（同义/错义/无义）。

下载资源

genome03_variant_anno_sci.R
8 KB

下载 VCF 注释可视化完整脚本 ↗

参考资源

- VariantAnnotation
- Ensembl VEP
- ANNOVAR

05 maftools 肿瘤突变分析

maftools 是肿瘤外显子组（WES）分析里最常用的 R 包。它读入 MAF 格式的体细胞突变文件，一套函数出完整的突变景观图、驱动基因分析、突变签名等。

本章用 maftools 自带的 TCGA LAML（急性髓系白血病，193 个样本）数据演示。

MAF 格式

MAF（Mutation Annotation Format）是 TCGA 定义的标准格式，每行一个突变，关键列：

列	含义
Hugo_Symbol	基因名
Chromosome / Start_Position / End_Position	基因组坐标
Variant_Classification	Missense / Nonsense / Frame_Shift 等
Variant_Type	SNP / INS / DEL
Tumor_Sample_Barcode	样本 ID
Protein_Change	蛋白变化（如 p.R882H）

从 VCF 转 MAF 用 `vcf2maf.pl`（需要 VEP 注释）。

真实示例

配套脚本 `genome04_maftools_sci.R` 输出 6 张图：

```
Rscript scripts/genomics/genome04_maftools_sci.R
```

每张图看什么

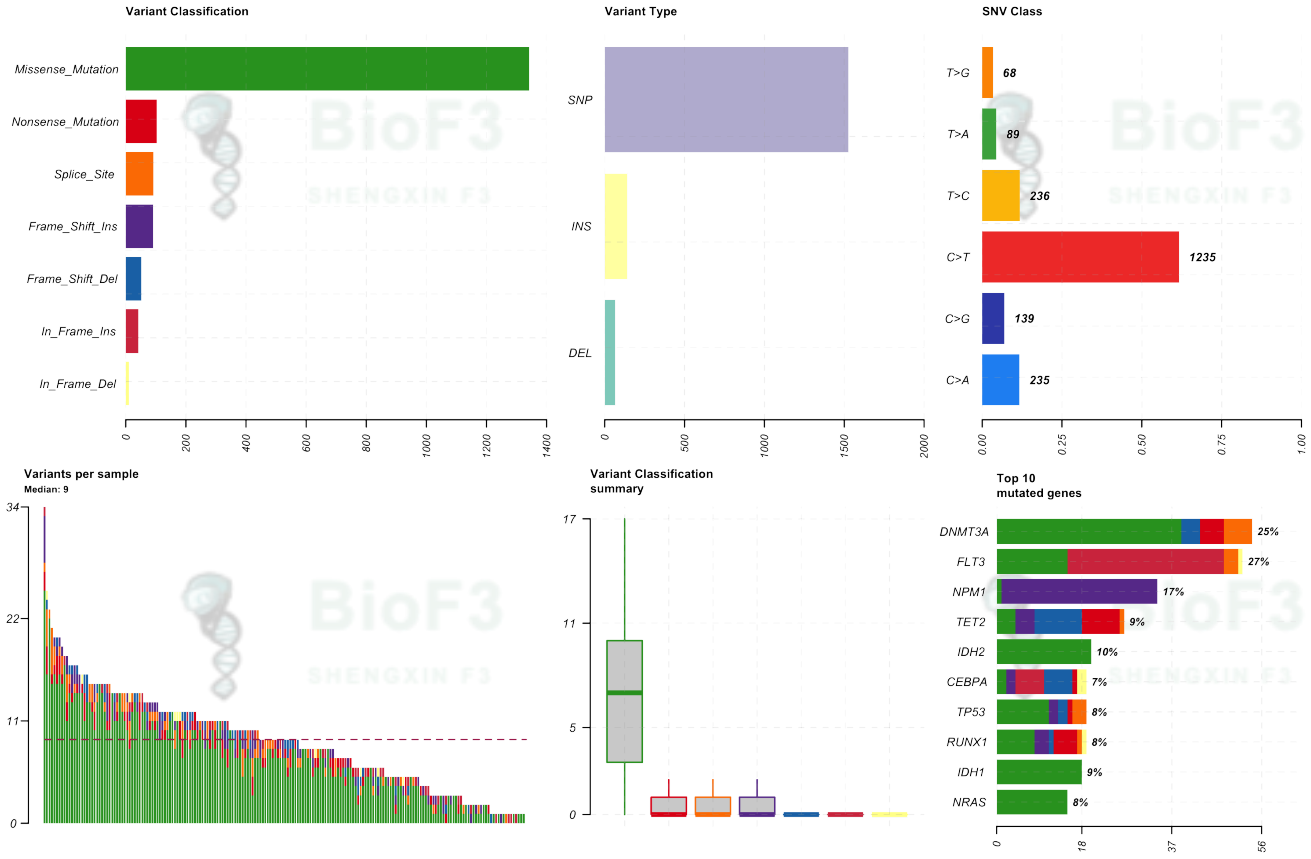

图 1：MAF 总览 dashboard。左上：每个样本的突变数；右上：变异分类分布；左下：变异类型；右下：SNV 碱基替换类型。一张图看完整个队列的突变概况。

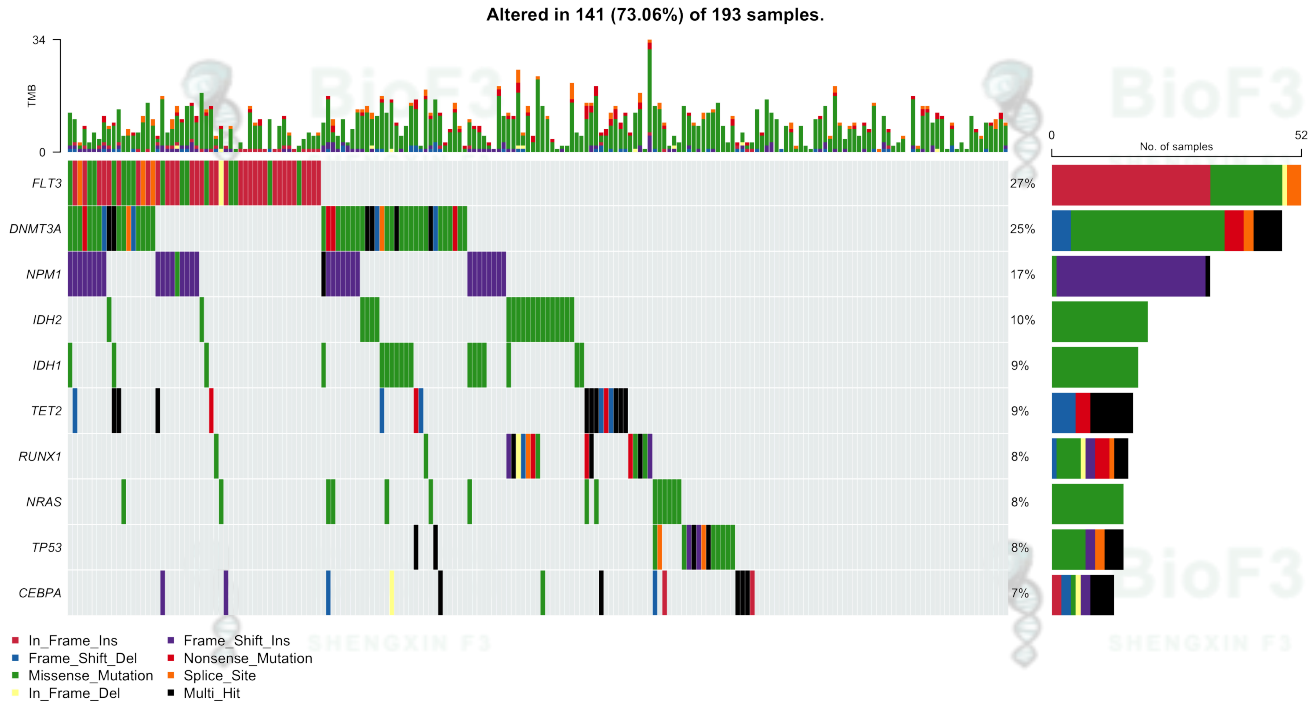

图 2：Oncoplot（突变景观图）。每列一个样本，每行一个基因（按突变频率排序）。颜色代表突变类型。这是肿瘤基因组文章里最核心的一张图 —— 一眼看出哪些基因在队列里反复突变。

LAML 里 FLT3、DNMT3A、NPM1 是 top 3 驱动基因，和文献完全一致。

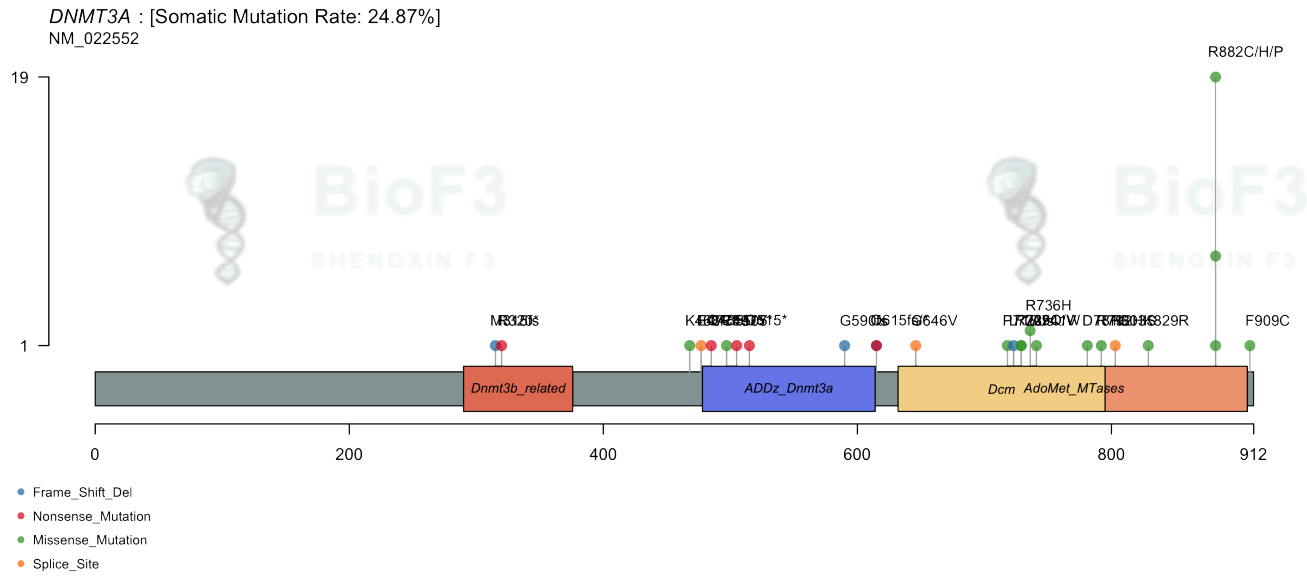

图 3：DNMT3A 的 lollipop 图。横轴是蛋白结构域，每个棒棒糖代表一个突变位点，高度代表该位点在队列里出现的次数。R882 是 DNMT3A 的热点突变，在 AML 里反复出现。

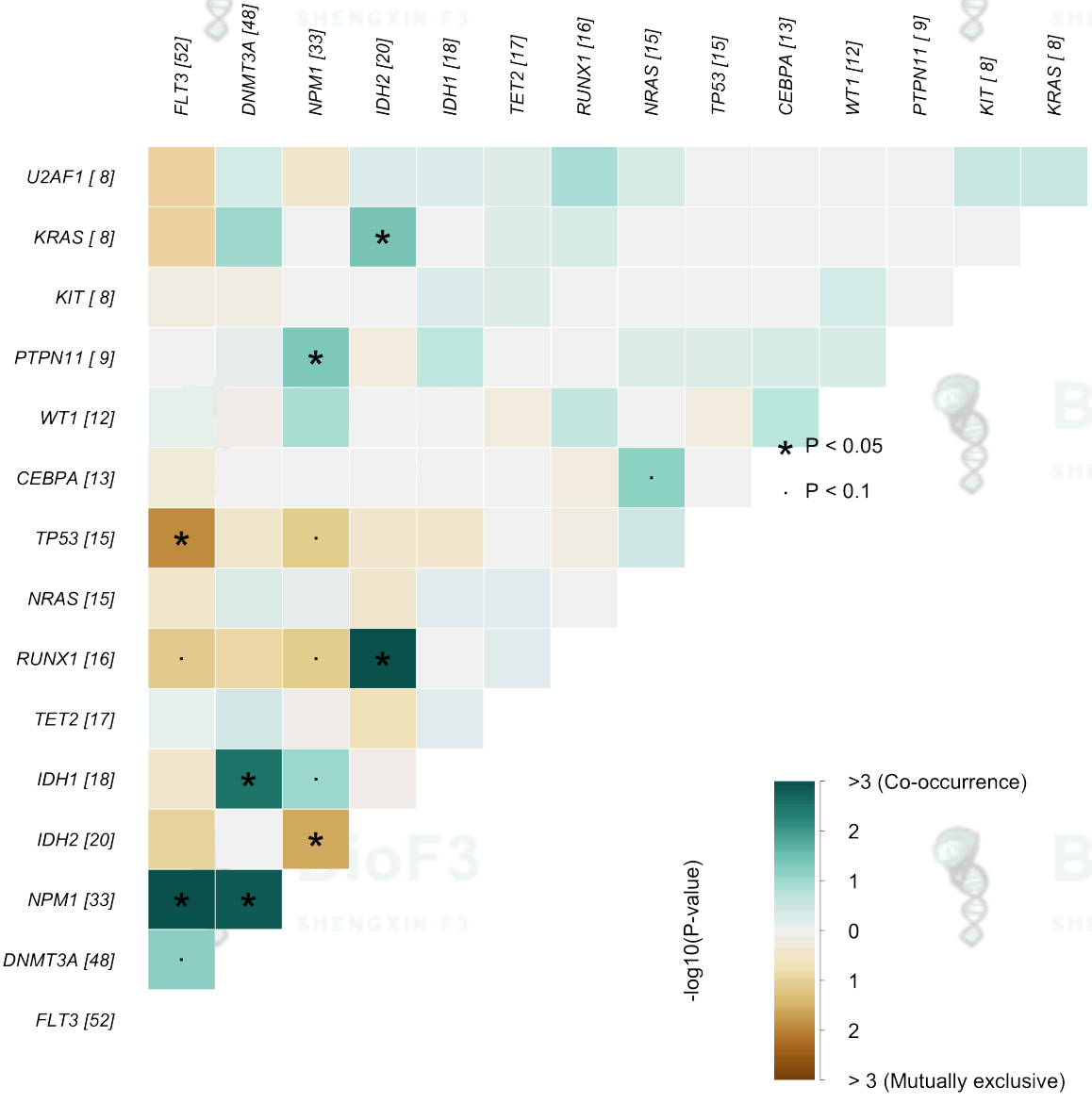

图 4：基因间的共突变 / 互斥关系。绿色 = 共现（co-occurrence），粉色 = 互斥（mutual exclusivity）。互斥的基因对可能在同一条通路上（突变一个就够了）。

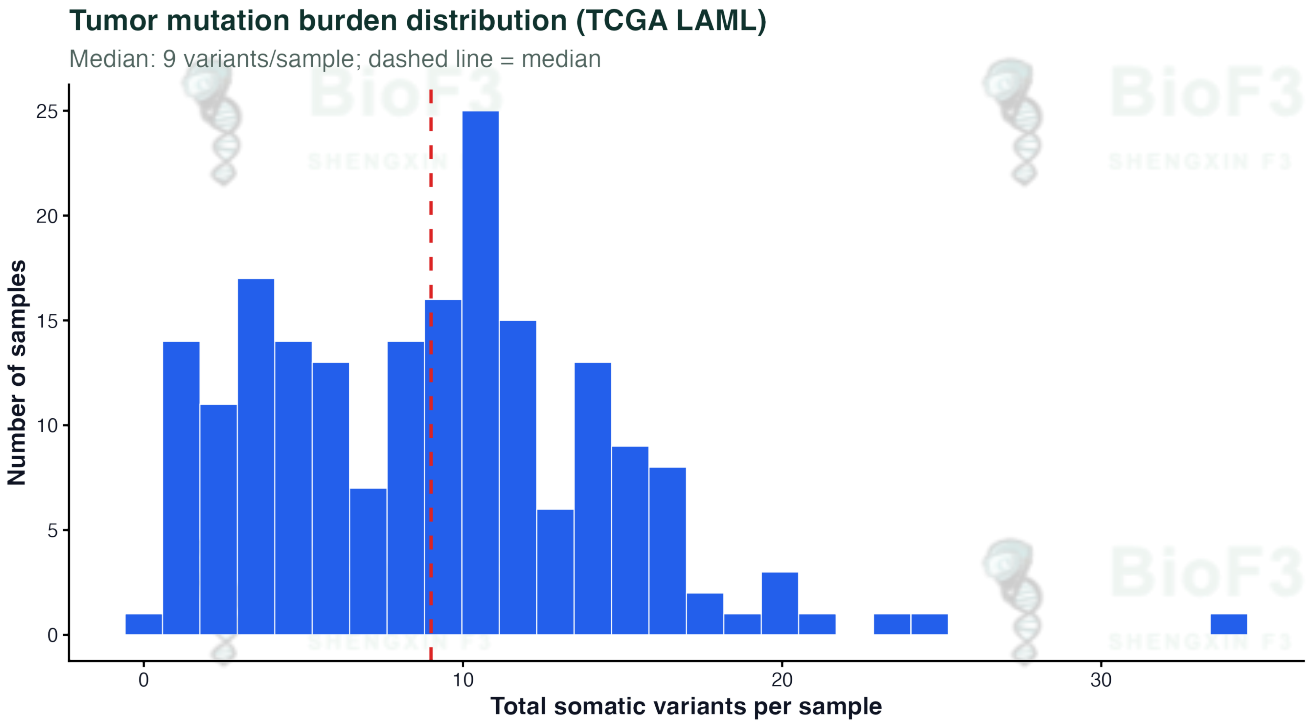

图 5：肿瘤突变负荷（TMB）分布。AML 是低 TMB 肿瘤（中位 ~10 个突变/样本），和黑色素瘤、肺癌（几百个）形成对比。TMB 是免疫治疗响应的预测指标之一。

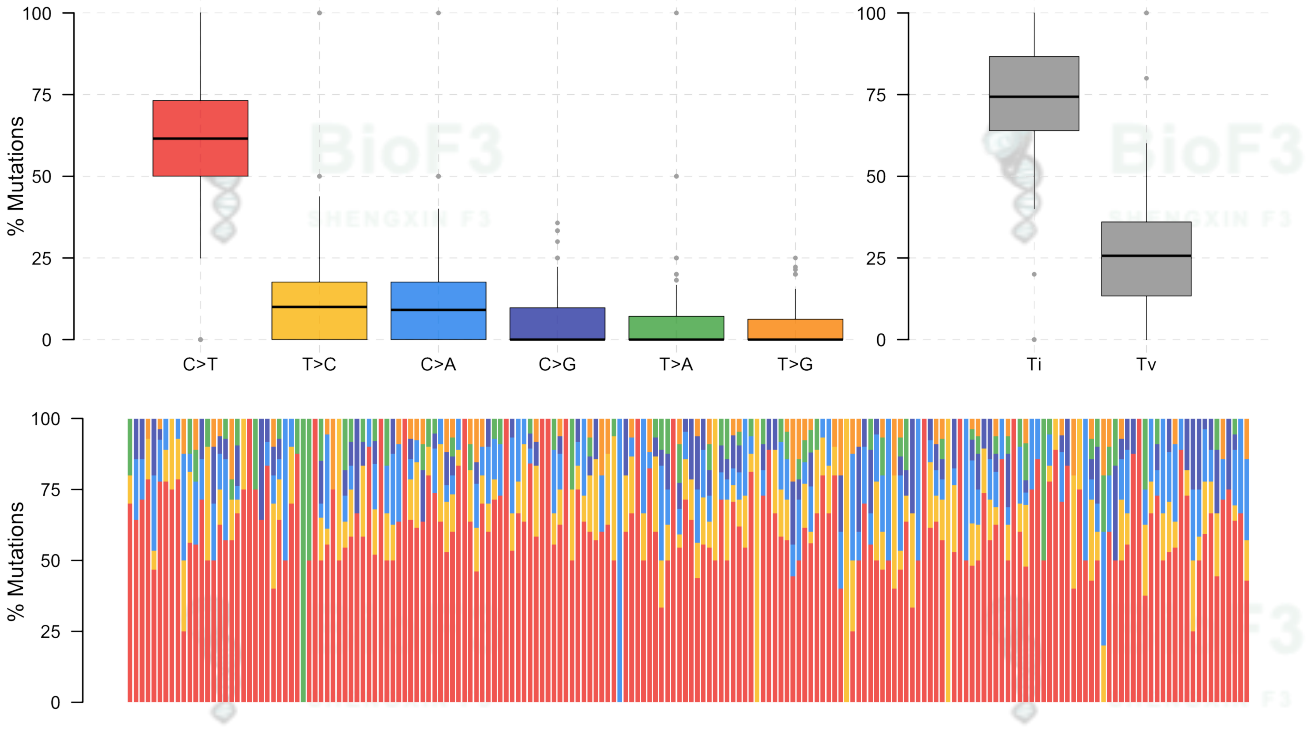

图 6：转换/颠换比（按样本）。Ti/Tv 偏离正常范围可能提示特定的突变过程（如 APOBEC 活性会增加 C>T 转换）。

核心代码

```
library(maftools)

laml <- read.maf(
  maf = system.file("extdata", "tcga_laml.maf.gz", package = "maftools"),
  clinicalData = system.file("extdata", "tcga_laml_annot.tsv", package = "maftools")
)

plotmafSummary(laml, dashboard = TRUE)
oncoplot(laml, top = 10)
lollipopPlot(laml, gene = "DNMT3A", AACol = "Protein_Change")
somaticInteractions(laml, top = 15)
```

套到自己数据上

把 `read.maf()` 的路径换成自己的 MAF 文件即可。如果只有 VCF：

```
# VCF -> MAF（需要 VEP + vcf2maf）
vcf2maf.pl --input-vcf somatic.vcf.gz --output-maf somatic.maf \
  --tumor-id TUMOR --normal-id NORMAL --ref-fasta hg38.fa \
  --vep-path /path/to/vep --vep-data /path/to/vep_cache
```

下载资源

genome04_maftools_sci.R
6 KB

下载 maftools 肿瘤突变分析完整脚本 ↗

参考资源

- maftools 文档
- Mayakonda et al. 2018, maftools 论文
- TCGA MAF 格式规范
- OncoKB 驱动基因注释

06 变异过滤与质量评估

原始 VCF 里有大量假阳性。过滤策略决定了最终结果的可靠性。

VQSR vs 硬过滤

方法	适用	原理
VQSR	样本量 > 30、WGS	用已知变异集训练机器学习模型
硬过滤	样本量少、WES、Panel	手动设阈值（QD、FS、MQ 等）

硬过滤典型参数

```
# SNP 过滤
gatk VariantFiltration -R ref.fa -V raw.vcf.gz \
  --filter-expression "QD < 2.0" --filter-name "LowQD" \
  --filter-expression "FS > 60.0" --filter-name "HighFS" \
  --filter-expression "MQ < 40.0" --filter-name "LowMQ" \
  --filter-expression "MQRankSum < -12.5" --filter-name "LowMQRS" \
  --filter-expression "ReadPosRankSum < -8.0" --filter-name "LowRPRS" \
  -O filtered_snps.vcf.gz

# Indel 过滤
gatk VariantFiltration -R ref.fa -V raw.vcf.gz \
  --filter-expression "QD < 2.0" --filter-name "LowQD" \
  --filter-expression "FS > 200.0" --filter-name "HighFS" \
  --filter-expression "ReadPosRankSum < -20.0" --filter-name "LowRPRS" \
  -O filtered_indels.vcf.gz
```

质量评估指标

- **Ti/Tv ratio**：WGS ~2.0-2.1，WES ~2.8-3.0。偏低 = 假阳性多
- **Het/Hom ratio**：人类 WGS ~1.5-2.0
- **已知变异比例**：和 dbSNP 的重叠率应该 > 95%（WGS）
- **Mendelian error rate**：有 trio 数据时，子代不符合孟德尔遗传的比例应 < 1%

参考资源

- GATK 硬过滤指南
- GATK VQSR 教程

07 群体遗传：PCA / 祖源 / LD

群体遗传分析用大量个体的基因型数据回答"这些人从哪来、彼此什么关系、哪些位点受到选择"。

常用工具

工具	用途
PLINK 2	数据管理、QC、PCA、关联分析
ADMIXTURE	祖源成分估计（K 群体）
vcftools	VCF 统计（Fst、pi、Tajima's D）
EIGENSOFT	PCA + 群体结构

PCA 分析

```
# VCF -> PLINK 格式
plink2 --vcf cohort.vcf.gz --make-bed --out cohort

# LD pruning（去掉高 LD 的 SNP，避免 PCA 被局部 LD 主导）
plink2 --bfile cohort --indep-pairwise 50 5 0.2 --out pruned
plink2 --bfile cohort --extract pruned.prune.in --make-bed --out cohort_pruned

# PCA
plink2 --bfile cohort_pruned --pca 10 --out cohort_pca
```

输出 `cohort_pca.eigenvec` 就是每个样本的 PC1~PC10 坐标，用 R 画散点图。

ADMIXTURE 祖源分析

```
# 跑 K=2 到 K=6
for K in 2 3 4 5 6; do
  admixture --cv cohort_pruned.bed $K | tee log_K${K}.out
done

# 选最优 K：看 CV error 最低的那个
grep "CV error" log_K*.out
```

LD 衰减

```
# 计算 LD（r²）随距离的衰减
plink2 --bfile cohort --ld-window-r2 0 --ld-window 1000 --ld-window-kb 500 \
  --out ld_decay
```

LD 衰减速度反映群体的有效群体大小和重组率。

参考资源

- PLINK 2 文档
- ADMIXTURE
- 1000 Genomes 群体结构教程

08 GWAS 入门

全基因组关联分析（GWAS）找的是"哪些基因组位点和某个表型（疾病、身高、药物响应）有统计学关联"。

基本流程

```
基因型数据（PLINK 格式）+ 表型文件
  → QC（MAF、HWE、缺失率、亲缘关系）
  → 关联检验（线性/逻辑回归，校正 PC）
  → Manhattan plot + QQ plot
  → 显著位点注释
```

PLINK 2 做关联

```
# 二分类表型（case/control）
plink2 --bfile cohort \
  --pheno phenotype.txt \
  --covar covariates.txt \
  --glm \
  --out gwas_results

# 输出 gwas_results.PHENO1.glm.logistic.hybrid
```

Manhattan plot

```
library(qqman)

results <- read.table("gwas_results.PHENO1.glm.logistic.hybrid",
                      header = TRUE)
manhattan(results, chr = "CHROM", bp = "POS", p = "P", snp = "ID",
          suggestiveline = -log10(1e-5), genomewideline = -log10(5e-8))
```

QQ plot

QQ plot 检查 p 值的整体膨胀（genomic inflation factor λ）。λ > 1.1 说明有群体分层或其他混杂没控制好。

```
qq(results$P)
```

常见坑

- **群体分层**：不同祖源的人混在一起会产生假关联。用 PCA 的前几个 PC 做协变量
- **多重检验**：全基因组显著性阈值是 5×10^{-8}（Bonferroni 校正 ~1M 独立检验）
- **LD 结构**：一个显著信号可能对应一整个 LD block 里的几十个 SNP，真正的因果变异需要 fine-mapping

参考资源

- PLINK 2 GWAS 教程
- qqman R 包
- GWAS Catalog
- LocusZoom

09 临床变异解读与报告

从"一份 VCF"到"能给临床医生看的报告"，中间需要变异分级、数据库查询和标准化报告格式。

ACMG 变异分级

美国医学遗传学学会（ACMG）把胚系变异分为 5 级：

级别	含义	行动
Pathogenic	致病	报告 + 临床干预
Likely pathogenic	可能致病	报告
VUS	意义未明	报告但不做临床决策
Likely benign	可能良性	通常不报告
Benign	良性	不报告

常用注释数据库

数据库	内容
ClinVar	变异-疾病关联（NCBI 维护）
gnomAD	人群等位基因频率
OncoKB	肿瘤驱动突变 + 药物靶点
COSMIC	肿瘤体细胞突变数据库
InterVar	自动化 ACMG 分级工具

VEP 注释 + ClinVar

```
vep --input_file variants.vcf \
    --output_file annotated.vcf \
    --cache --dir_cache /path/to/vep_cache \
    --assembly GRCh38 \
    --everything \
    --plugin ClinVar,/path/to/clinvar.vcf.gz
```

报告模板

临床报告通常包含：

1. 患者信息 + 检测方法
2. 阳性发现（Pathogenic / Likely pathogenic 变异）
3. VUS 列表
4. 质量指标（覆盖度、Ti/Tv）

5. 方法学描述
6. 局限性声明

肿瘤报告的特殊性

肿瘤报告额外需要：

- 突变等位基因频率（VAF）
- 肿瘤突变负荷（TMB）
- 微卫星不稳定性（MSI）状态
- 可靶向突变（OncoKB Level 1-4）

参考资源

- ClinVar
- ACMG 2015 指南
- OncoKB
- InterVar
- Franklin（Genoox 自动分级）