

## BIOF3 组学数据分析

# 07 GWAS 入门

导出日期：2026年5月12日

## 07 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.PHEN01.glm.logistic.hybrid
```

### Manhattan plot

```
library(qqman)

results <- read.table("gwas_results.PHEN01.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  $\lambda$ )。  $\lambda > 1.1$  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  $5 \times 10^{-8}$  (Bonferroni 校正  $\sim 1M$  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](#)
- [ggman R 包](#)
- [GWAS Catalog](#)
- [LocusZoom](#)



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  
SHENGXIN F3



BioF3  
SHENGXIN F3



BioF3  
SHENGXIN F3



BioF3  
SHENGXIN F3



BioF3  
SHENGXIN F3