BIOF3 组学数据分析

05 变异过滤与质量评估

导出日期：2026年5月12日

05 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