BIOF3 组学数据分析

02 比对与变异检测流程

导出日期：2026年5月12日

02 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

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。