

Zhang-Lab 生信小课堂 第廿四期
Applied Bioinformatics Club (ABC)

和趣求真  秉实生信

(张建伟生物信息学课题组 <https://zhang.hzau.edu.cn>)

结构变异 (SV) 检测工具的选择策略

2025.11.7 二综C102 15:00 欢迎大家交流学习!

主讲人: 路佳雯
2025/11/7

目录

- 01** 背景速览：SV简介与检测挑战
- 02** 工具拆解：主流SV检测器原理与比较
- 03** 四部选择策略：根据场景选择工具的最佳实践
- 04** 总结

结构变异（SV）为何重要？

1.



- 有研究发现，基因组上的SVs比起SNP而言，更能代表人类群体的多样性特征；
- 稀有且相同的一些结构性变异往往和疾病（包括癌症）的发生相互关联甚至还是其直接的致病诱因。

2.

检测核心挑战

技术层面：短读长“断片化”（难跨复杂SV）、长读长高错误率（5-20%，易混淆真实变异与误差）

序列层面：重复序列（如 LCRs、Alu 元件）干扰断点定位

SV 检测工具分类与核心原理

(1) 工具分类表（按检测原理 / 测序技术）：

工具类别	核心原理	典型工具	优势	局限性
短读长工具	读对不一致、拆分读、覆盖度变化	Manta	成本低、通量高，适合大样本筛查	复杂 SV（如嵌套 INV/DUP）检测率低，INS 漏检多
长读长工具	长读长跨断点连续比对信号	Sniffles2、cuteSV、PBSV/SVIM	分辨率高（单碱基级断点），能检测复杂 SV 和 novel INS	对测序深度敏感（5× 以下漏检多），计算资源消耗高
整合工具	合并多工具结果，优化一致性	SURVIVOR	降低假阳性（如 cuteSV+SURVIVOR）	依赖原始工具质量，操作复杂

(2) 核心评价指标

- 准确率 (*Precision*) : $\text{真阳性} / (\text{真阳性} + \text{假阳性})$ —— 避免 “误判不存在的变异”
- 召回率 (*Recall*) : $\text{真阳性} / (\text{真阳性} + \text{假阴性})$ —— 避免 “漏检真实变异”
- *F1* 分数: $2 \times (\text{准确率} \times \text{召回率}) / (\text{准确率} + \text{召回率})$ —— 平衡两者, 文献核心评价指标
- 速度与内存: 长读长工具关键指标 (如 *Sniffles2* 比 *Sniffles1* 快 11.8 倍)
- 功能适配: 是否支持群体分析、嵌合 SV 检测 (如 *Sniffles2* 的 *mosaic* 模式)

核心工具解析: Sniffles

➤ 核心原理 (三大创新)

- 重复序列感知聚类: 动态调整重复区 (如 LCRs) 聚类参数, 解决信号分散
- 覆盖度自适应过滤: 根据全局 / 局部深度调整支持读长阈值 (避免低覆盖漏检)
- 多模式设计: *germline* (生殖系)、群体 (*combine*)、*mosaic* (嵌合)

➤ 性能数据 (GIAB HQ002 样本)

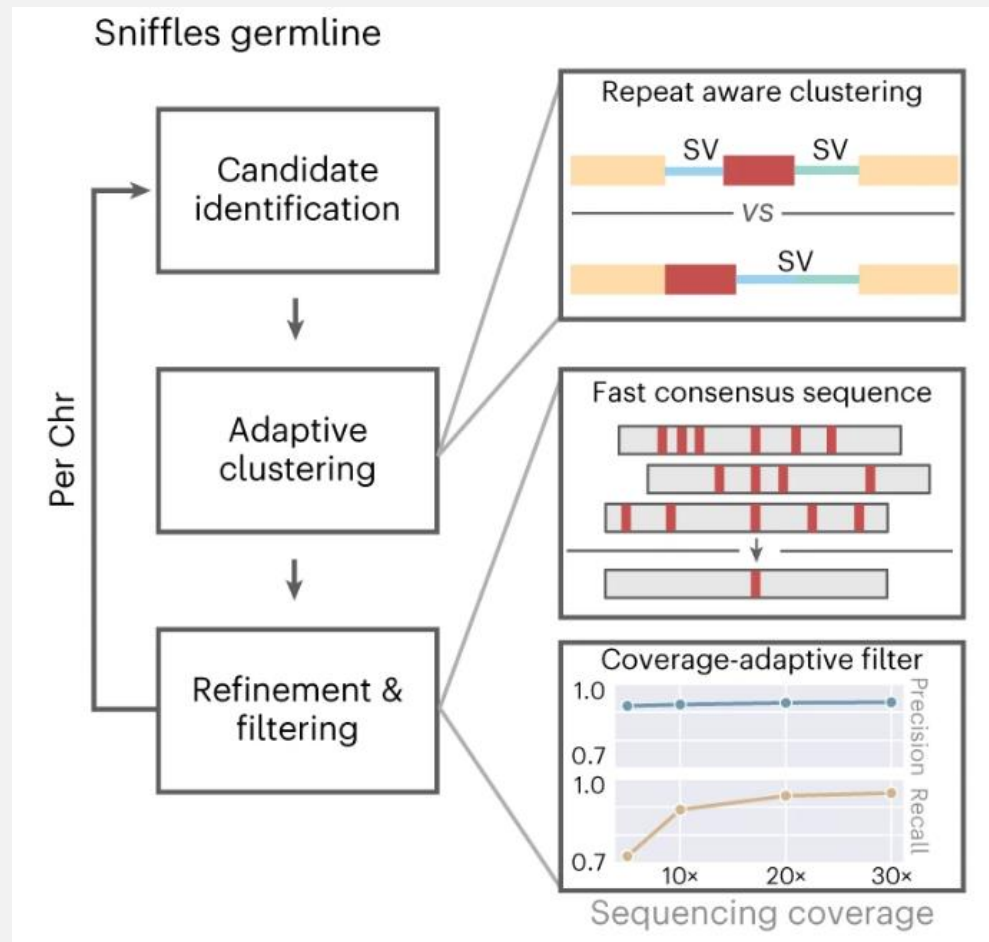
- 速度: 11.8 倍快于 *Sniffles1*, 777 样本合并仅 11h CPU 准确率: 比 *cuteSV*/*PBSV* 高 29% (5-50x ONT/HiFi 数据)
- 嵌合 SV: 5-20% VAF 检测召回率 94.47% (HQ002 spike-in 实验)

➤ 适用场景:

- 孟德尔疾病复杂 SV 检测 (解析 DUP-NML-INV/DUP 结构)
- 群体基因组学 (生成全基因型 VCF, 解决 “n+1” 样本新增问题)
- 嵌合 SV 研究 (MSA 脑区低频率 SV, VAF 5-20%)

➤ 局限性:

- 高度重复区 (99.9% 相似性 *SegDups*) 无法完全解析 (如 DUP-TRP/INV-DUP 的 *Jct1* 断点)
- 计算资源需求较高 (默认 32GB 内存, 大样本需多线程)



核心工具解析: cuteSV

➤ 核心原理 (两大关键)

- 分类型 SV 签名提取: 针对 DEL/INS/INV/DUP/BND 设计专属规则 (如 INS 用 CIGAR + 拆分读验证)

- 两步聚类优化: 先按位置聚类, 再按 SV 长度分亚簇 (解决杂合 SV 信号异质性)

➤ 性能数据 (HG002 样本)

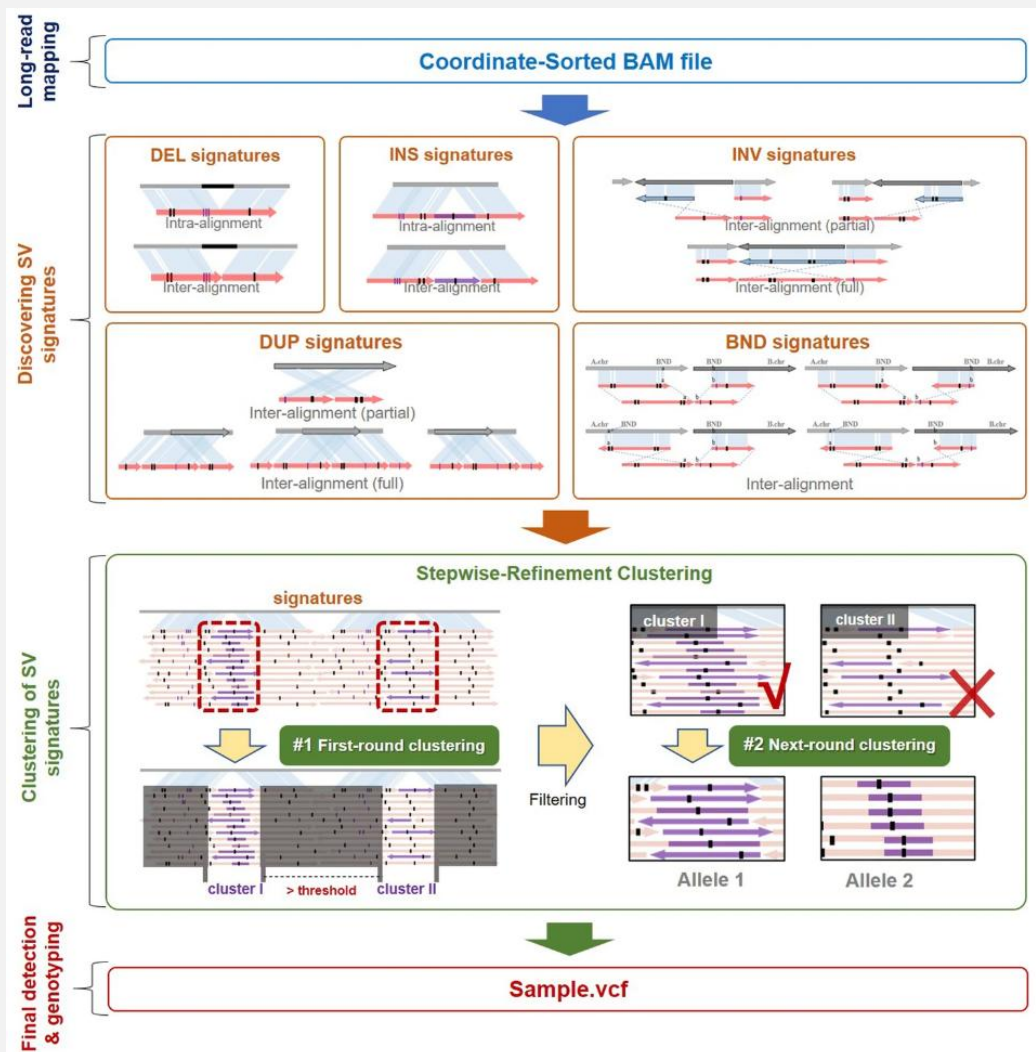
- 多线程效率: 16 线程比 1 线程快 15 倍, 内存仅 0.1-0.4GB (远低于 Sniffles2)
- 低覆盖度表现: 10× ONT 数据 F1=88.85% (比 Sniffles/SVIM 高 7%+)

➤ 适用场景:

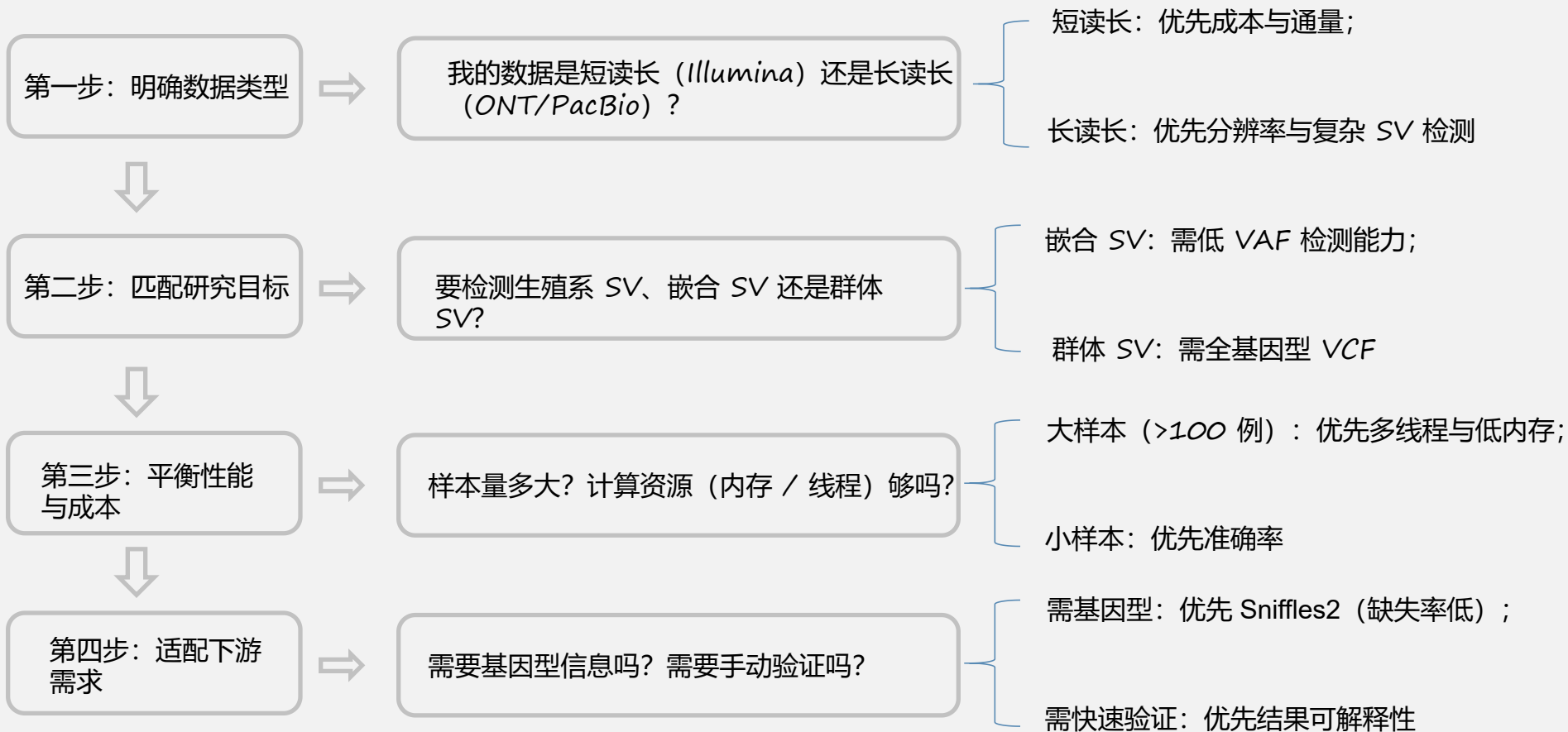
- 大样本低覆盖长读长项目 (如 1000 Genomes 类研究, 计算资源有限)
- 单样本快速检测 (如临床样本初筛, 10× ONT 数据 2 小时出结果)
- 杂合 SV 解析 (同一 loci 多等位基因, 如 108bp+36bp 双 INS)

➤ 局限性:

- 群体分析需依赖 SURVIVOR, 缺失率高 (32.2% vs Sniffles2 的 1.29%)
- 复杂 SV (如 DUP-TRP/INV-DUP) 解析能力弱于 Sniffles2



四步选择策略



！ 最终结果都需要进行验证，如 IGV 可视化

实战案例分析

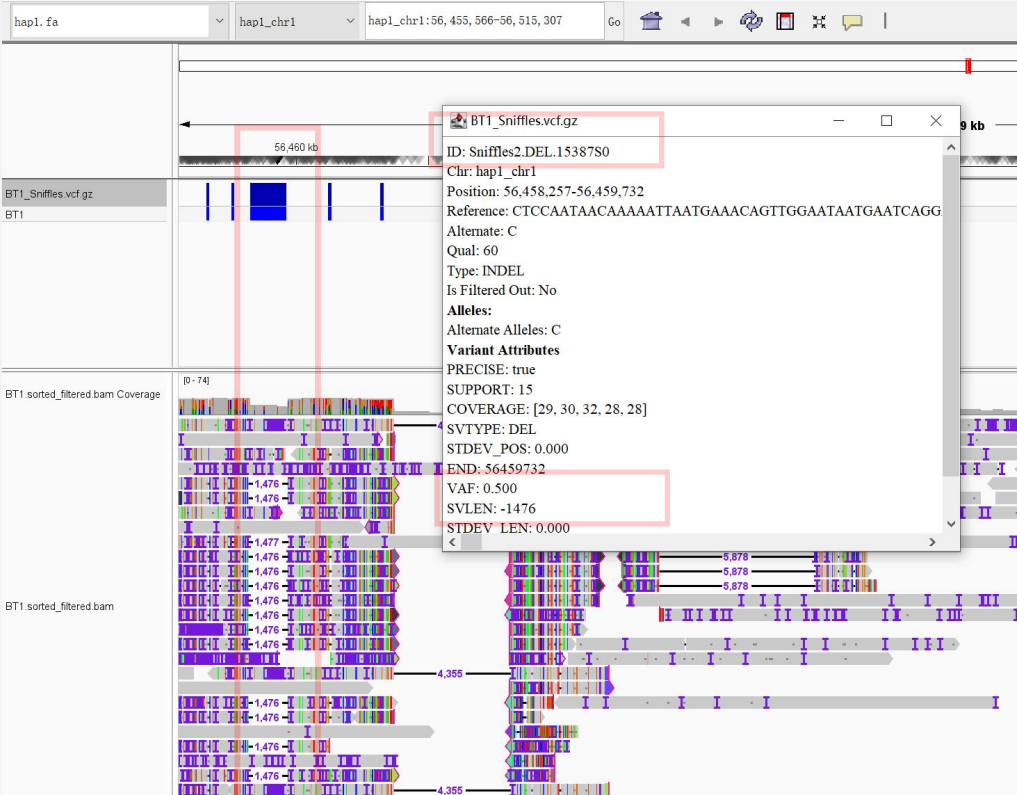
示例：用 Sniffles 进行callSV
代码示例：

```
sniffles --input ${BAM_FILE} \  
--vcf ${OUTPUT_VCF} \  
--reference ${REFERENCE} \  
--threads 16 \  
--sample-id ${SAMPLE_NAME} \  
--minsupport 10 \  
--minsvlen 30 \  
--mapq 20 \  
--min-alignment-length 500 \
```

过滤参数：根据需要
进行严格或宽松

结果与可视化验证：

具体变异类型、事件
长度等可在 IGV 中
可视化验证



常见问题	解决方案	工具 / 方法
不同工具结果不一致（如 Sniffles2 vs cuteSV）	用 SURVIVOR 合并，设置断点匹配阈值（500bp 内），选 “≥2 工具支持” 的 SV	SURVIVOR
复杂 SV（如嵌套 INV+DUP）检测不到	1. 用 Sniffles2 的 --repeat-aware 模式（输入 LCR 注释）； 2. IGV 手动验证长读长跨断点信号	Sniffles2
低覆盖度数据（<10×）漏检多	1. cuteSV 设 --min_support=1-2； 2. Sniffles2 用 mosaic 模式降低阈值； 3. 后续 PCR 验证	cuteSV/Sniffles2
群体分析缺失率高（如 cuteSV）	换用 Sniffles2 的 combine 模式，生成 SNF 文件合并，缺失率会降低	Sniffles2

核心总结：

- 工具无“万能”，只有“适配”：
 - 嵌合 SV / 复杂 SV / 群体基因型：优先 Sniffles2
 - 大样本低覆盖 / 计算资源有限：优先 cuteSV
 - 短读长筛查：Manta+Delly+SURVIVOR
 - 选择逻辑：数据类型→研究目标→性能成本→下游需求（四步走）
- 对于得到的结果都要进行验证

Q&A

汇报人：路佳雯