purge_dups运行教程：覆盖度准备、证据检查与资源记录

完整操作教程 · 参考固定 · 往返恢复验收

purge_dups输出的是依据覆盖度与序列相似性识别的候选冗余区域，不能把较小组装直接当作更好结果。本篇限定具有适用PacBio CCS reads的二倍体初始组装，先生成证据、人工检查覆盖阈值，再在独立目录提取序列。多倍体、混样或特殊拷贝结构需要另行评估。

运行边界：示例代码需按安装版本和真实项目输入复核；未提供真实生物数据，语法或小样本检查不代表科学结果和性能验收。

purge_dups输出的是依据覆盖度与序列相似性识别的候选冗余区域，不能把较小组装直接当作更好结果。本篇限定具有适用PacBio CCS reads的二倍体初始组装，先生成证据、人工检查覆盖阈值，再在独立目录提取序列。多倍体、混样或特殊拷贝结构需要另行评估。

把证据生成与删除决策分开

安装并记录purge_dups、minimap2版本。CCS使用map-hifi，其他reads不能只改文件名沿用预设。官方要求大型参考设置minimap2的-I，避免多批索引造成覆盖统计问题；它按目标序列碱基数设索引批大小，不是可用内存上限。先统计本项目参考长度，填写能容纳单批索引的值并评估内存。

文件	生成方式	审核要点
PB.stat / PB.base.cov	reads映射后pbcstat	覆盖分布、低覆盖和不同峰是否合理
cutoffs.auto	calcuts自动推断	不能未经检查直接作为删除依据
self.paf.gz	拆分组装后自比对	与同一组装及版本对应
dups.bed / purged.fa	经核对阈值后生成	保留原组装并检查候选损失

生成一个可复核的证据包

以下为Linux Bash，输入路径是示例。INDEX_BASES由本项目参考长度与预检填写，例如使用minimap2接受的碱基数单位；两线程仅为试跑教学值。reads和组装应已完成基本格式、污染与样本来源检查。命令失败即停止，不将不完整PAF继续交给后续步骤。

```
set -euo pipefail
: "${INDEX_BASES:?[redacted]}"
asm=/data/primary.fa
reads=/data/ccs.fastq.gz
test -s "$asm"
test -s "$reads"
gzip -t "$reads"
mkdir purge-evidence-001
cd purge-evidence-001
sha256sum "$asm" "$reads" > inputs.sha256
minimap2 --version > minimap2.version
/usr/bin/time -v -o map.time minimap2 -x map-hifi -I "$INDEX_BASES" \
  -t 2 "$asm" "$reads" 2> map.log | gzip -c > reads.paf.gz
pbcstat reads.paf.gz > pbcstat.log 2>&1
test -s PB.stat
test -s PB.base.cov
calcuts PB.stat > cutoffs.auto 2> calcuts.log
test -s cutoffs.auto
split_fa "$asm" > assembly.split.fa
test -s assembly.split.fa
/usr/bin/time -v -o self.time minimap2 -x asm5 -DP -I "$INDEX_BASES" \
  -t 2 assembly.split.fa assembly.split.fa 2> self.log | gzip -c > self.paf.gz
```

```
gzip -t reads.paf.gz self.paf.gz
```

检查后才提取候选结果

先画出并阅读覆盖分布，比较自动阈值与组装背景，核对不同峰的来源。把确认后的阈值另存为cutoffs.checked，记录修改理由与版本；无法合理解释覆盖峰时先停止。继续在上述证据目录执行，下例-e只去除末端重复，避免不经审查切除内部片段；这仍不保证所有生物学重复均被正确处理。

```
set -euo pipefail
test -s cutoffs.checked
mkdir result-001
cd result-001
purge_dups -2 -T ../cutoffs.checked -c ../PB.base.cov ../self.paf.gz \
  > dups.bed 2> purge.log
test -f dups.bed
get_seqs -e dups.bed /data/primary.fa > get_seqs.log 2>&1
test -s purged.fa
test -f hap.fa
```

验收包含“没有候选”的分支

dups.bed为空可能表示没有识别到候选，不应一律当失败；hap.fa也可能为空。因此这里区分文件存在与非空。反过来，大量候选也不是自动通过，需要检查其长度、覆盖、坐标范围和原始contig归属。hap.fa是被分出的序列集合，不能直接称为完整另一套单倍型组装。

分别比较原始与purged组装的长度、序列去向、reads支持和完整性指标。重复BUSCO减少时，同时查看缺失与完整BUSCO，防止真实基因被误删；保留审阅表以及所有原始序列。资源记录分reads映射、自比对和提取阶段，除了RSS还测工作盘高水位。先顺序完成一个基因组，确认证据链和结果可解释后，再为不同基因组建立独立任务目录与队列。

核对map.log与self.log中实际建立或读取索引的过程，确认没有因为参考规模超过设置而分批；只看到PAF非空不能证明该前提满足。阈值审核表同时写覆盖峰位置、自动阈值、最终阈值、修改原因和候选长度。调整阈值重新比较时，保留旧result目录，不能覆盖后只留下最好看的组装。提取前重新校验输入SHA，确认参考未被替换。

配套练习包

本PDF对应的ZIP包含本文代码块、主题练习模板或合成数据、README和SHA256校验文件。请从同一教程页面下载练习包；先核对校验值，再在独立目录复现。

参考资料

purge_dups官方工作流与索引限制 · 来源资料

get_seqs输出与-e实现 · 来源资料

purge_dups 官方流程与阈值说明 · 来源资料

操作建议：先在合成小样本上验证流程，再依据项目参考、标签合同、存储策略与获配资源验收真实数据。