miRDeep2多样本实操：建立输入清单与独立工作目录

完整操作教程 · 参考固定 · 往返恢复验收

用唯一清单、绝对参考路径和新工作目录隔离miRDeep2样本。

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。

miRDeep2多样本批处理应先解决输入类型和工作目录。mapper会生成折叠序列与基因组映射，miRDeep2再利用成熟体和前体参考识别候选；多份任务挤在同一目录，容易混用中间结果。本篇以官方0.1.3发行版接口组织顺序队列，先让一个样本验收，再考虑并发。

输入先处理接头与UMI

本例输入为已经按建库方案剪接头、处理UMI的未压缩FASTQ。小RNA接头进入读段时不能跳过剪切；若交给mapper处理，应使用真实接头的-k参数，不照搬教程序列。参考包括基因组、同物种成熟miRNA、相关物种成熟miRNA和同物种前体，来源版本及物种对应关系全部记录。

文件/工具	要求	验收内容
FASTQ	样本独立、预处理方法固定	读段长度分布及处理统计
Bowtie索引	由对应基因组用Bowtie 1构建	前缀与FASTA匹配
miRNA参考	成熟体、相关物种及前体各自明确	版本、序列ID、物种
工具环境	Perl、Bowtie 1、RNAfold、randfold等	官方测试通过及依赖版本

把样本ID与输入路径分开

准备samples.tsv，无表头、两列TAB、LF换行，样例为S01及/data/S01.trimmed.fastq。以下参考路径都是占位，须换成已验收文件；Bowtie前缀不带索引后缀。mapper的-l 18和-o 2是教学设置，长度规则按研究方法确定，两线程不代表服务器需求。

```
set -euo pipefail
root="$PWD/mirdeep-batch"; mkdir "$root"
genome=/ref/genome.fa
idx=/ref/bowtie1/genome
mature=/ref/mature_this.fa
related=/ref/mature_other.fa
hairpin=/ref/hairpin_this.fa
for f in "$genome" "$mature" "$related" "$hairpin"; do test -s "$f"; done
declare -A seen=(); n=0
while IFS=$'\t' read -r id fq extra || [[ -n "$id$fq$extra" ]]; do
  [[ "$id" =~ ^[A-Za-z0-9][A-Za-z0-9_.-]*$ ]]
  [[ -z "$extra" && "$fq" = /* ]]
  [[ -z "${seen[$id]+yes}" ]]; seen[$id]=1
  test -s "$fq"
  mkdir "$root/$id"
  (
    cd "$root/$id"
    /usr/bin/time -v -o mapper.time mapper.pl "$fq" -e -h -i -j \
      -l 18 -m -p "$idx" -s reads_collapsed.fa \
      -t reads_vs_genome.arf -o 2 -v > mapper.log 2>&1
    test -s reads_collapsed.fa; test -s reads_vs_genome.arf
    /usr/bin/time -v -o mirdeep.time miRDeep2.pl reads_collapsed.fa \
      "$genome" reads_vs_genome.arf "$mature" "$related" "$hairpin" \
      > mirdeep.log 2>&1
    shopt -s nullglob
    csv=(result_*.csv); web=(result_*.html)
    test "${#csv[@]}" -eq 1; test -s "${csv[0]}"
    test "${#web[@]}" -eq 1; test -s "${web[0]}"
  )
  n=$((n+1))
```

```
done < samples.tsv
test "$n" -gt 0
```

先检查折叠与映射，再读候选

运行前留存清单和实际软件版本。mapper输出的折叠FASTA带有丰度信息，应保持其命名规则，不要用普通去重FASTA替换。检查处理后读段数、长度与映射比例是否合理；arf非空只说明存在映射，不能证明输入和物种选择正确。参考序列ID也不能与基因组名称混淆。

每个目录应只有本次生成的一组result时间戳文件，候选表与HTML能互相对应。打开HTML，检查候选结构、成熟体与星链支持及评分背景；若PDF生成失败，先检查相应依赖和日志。这个例子不指定-t物种名，项目需要物种特异标注时按官方支持值设置。

让失败保留在对应样本目录

脚本遇异常即停，已经完成的样本保留原目录，失败样本留日志供修复。重新提交采用新批次根目录，避免时间戳结果累积让验收失效。输入若仍压缩，先在新位置校验并解压；不能把gz文件仅改后缀当作FASTQ。重复样本ID或空清单会拒绝运行。

准备批处理登记表，填样本ID、原始reads数、预处理后reads数、折叠序列数、映射数与结果相对路径。同一份折叠序列可代表多个原始reads，汇总时不能把序列条数当作分子或测序读段总数；样本间采用相同统计口径再比较。

先比较mapper与miRDeep2各段的墙钟时间、最大RSS和盘峰值，再安排多样本并发。Bowtie使用的线程不代表RNA折叠等全部步骤都按相同比例扩展。增加并发前确认独立目录、总内存与共享参考读取；候选miRNA仍需后续证据评估，不因队列全部完成就自动确认新发现。

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

miRDeep2官方教程 · 来源资料

miRDeep2 0.1.3发行版 · 来源资料

mapper输入和线程参数 · 来源资料

miRDeep2输出实现 · 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。