souporcell拆样实操：核对基因型相关输入并管理并行步骤

完整操作教程 · 参考固定 · 往返恢复验收

为单个混样文库核对BAM、条形码、参考与预期供者数，按阶段运行souporcell并验收未分配及双细胞。

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。

souporcell根据遗传差异拆分混样单细胞，不能仅凭增加CPU就辨认没有足够遗传差异的供者。先确认混样设计、物种、文库及预期供者数，再检查输入。它得到的遗传簇也不自动等于已经知道姓名或实验编号的供者。

把一份文库作为一个任务

本文限定兼容流程中的Cell Ranger scRNA-seq BAM、对应filtered barcodes及基因组FASTA，采用当前pipeline的默认重比对路线。BAM应保留所需CB和UMI标签，条形码后缀与BAM一致；不要把不同混样文库的同名barcode直接合起来。

输入	核对内容	失败时先查
BAM	来源、可读性、CB/UMI标签、文库	中间处理是否丢标签
barcodes	本批过滤细胞，一行一个，唯一	是否误用raw或另一批细胞
参考	物种、组装、序列及变异坐标	不要混用不同参考的常见变异
k与倍性	实验预期和方法适用性	不是为凑出漂亮簇数随意选择
运行环境	pipeline版本、依赖及路径	缺工具不等于RAM不足

先做只读输入与环境检查

记录pipeline提交、BAM与参考校验值，保留安装版帮助。当前主线pipeline声明3.0.0；旧镜像不一定具有相同参数。下面先验证一个文库，PIPELINE填写兼容环境内的实际脚本，THREADS与DONORS填写正整数且在实验和平台允许范围内。

```
set -euo pipefail
: "${PIPELINE:?[redacted]pipeline[redacted]}"
: "${BAM:?[redacted]BAM}"
: "${BARCODES:?[redacted]}"
: "${FASTA:?[redacted]FASTA}"
: "${THREADS:?[redacted]}"
: "${DONORS:?[redacted]}"
: "${OUT:?[redacted]}"
for value in "$THREADS" "$DONORS"; do
  case "$value" in ""|*[!0-9]*) exit 2;; esac
  test "$value" -ge 1 || exit 2
done
test -r "$PIPELINE" && test -s "$BAM" && test -s "$BARCODES" || exit 2
test -s "$FASTA" && test ! -e "$OUT" && test ! -L "$OUT" || exit 2
python "$PIPELINE" -h > "$OUT.help.txt" 2>&1
samtools quickcheck "$BAM"
sha256sum "$BAM" "$BARCODES" "$FASTA" > "$OUT.inputs.sha256"
/usr/bin/time -v -o "$OUT.resources.txt" \
  python "$PIPELINE" -i "$BAM" -b "$BARCODES" -f "$FASTA" \
  -t "$THREADS" -o "$OUT" -k "$DONORS" \
  > "$OUT.stdout.log" 2> "$OUT.stderr.log"
test -s "$OUT/clusters.tsv" && test -s "$OUT/cluster_genotypes.vcf" || exit 1
```

日志与资源文件位于OUT旁，父目录需可写。quickcheck只是BAM基础结构检查，不证明标签、样本身份或完整内容正确。条形码数量太少等pipeline守卫触发时，先判断样例与输入是否适合，不直接使用ignore绕过。

按阶段解释资源与错误

读取、重比对、变异计数、聚类及双细胞识别的资源不同，查清最后成功步骤与第一条实质错误。先单任务记录作业RAM、CPU和盘高水位，再按实际重叠安排并发；不套用README某个人类测试数据的固定RAM或磁盘倍数。

skip_remap不是任意数据都可用的加速开关。若研究确需常见变异或已知基因型路线，按该版本要求核对参考、位点和文件格式，单独建运行并重新验证。改变k、过滤或最大位点数也会改变分析条件，不拿它们偷偷换取资源下降。

验收分配而非只验进程

检查clusters.tsv的实际列、所有输入细胞覆盖、singlet、doublet及未分配情况，保留概率与质量信息。核对cluster基因型和ambient RNA等相应产物，按项目判断变异证据和簇稳定性。细胞被分配并不说明该遗传簇已映射到真实供者。

跨文库的cluster0可能对应不同人，真实身份需已知基因型或其他独立设计信息匹配。失败后保留原目录与日志，确认旧任务已结束，再按当前支持的恢复规则或新目录重跑；不手工拼接部分分配表当完整结果。最后逐文库汇总状态，全部必要文库通过后再做队列统计。

每批状态表至少记录文库、输入条形码数、供者数、结束码、结果位置和复核结论。进程成功而缺少必要产物仍属失败；结果行数异常时先定位缺失细胞，不能只按输出文件存在判断通过。

确定输入和实验问题是否适合拆样

souporcell面向混样单细胞表达数据，利用表达读段中的遗传变异信号推断遗传簇。若供者之间遗传差异不足、覆盖太低、文库细胞数偏少或表达读段不包含可用位点，增加线程或RAM不能创造区分信号。运行前列出混样文库数量、每份文库的预期供者数、物种、样本是否同一人重复采集及是否有外部基因型可用。

检查BAM是否来自相应单细胞建库流程、barcode标签是否对应所用barcodes文件、参考基因组与比对版本是否匹配。filtered barcode列表应与本次文库一致，不可混用raw barcode或另一批次文件。可以先检查BAM头部、条形码样例和CB/UB标签存在情况；samtools quickcheck只能发现部分文件截断/结构问题，不能验证科学来源和标签正确性。

参数k应来自实验设计中的预期混样供者数量，而不是多跑几个k后只挑簇最漂亮的结果。若确实需要敏感性分析，预先说明候选范围和判读指标，保留所有结果。cluster编号只是算法标签，不等于供者姓名；只有通过独立基因型或实验信息匹配后，才可映射到真实身份。

跑前环境、磁盘与输出目录检查

souporcell及依赖可由不同镜像或安装方式提供，CLI参数也可能随版本变化。先记录容器或环境版本、pipeline脚本散列并保存-h输出；逐项核对当前脚本对输入文件、输出目录和参考索引的要求。不要因为旧文章中某个参数存在就假设本机版本接受它。线程数必须不超过作业额度，临时目录应位于有足够空间且权限正确的文件系统。

输入数据和中间重比对文件可能远大于最终结果。开始前估算可用磁盘并设定工作目录，避免把临时文件写进系统盘或共享结果目录。输出目录不存在、父目录可写且路径不含意外符号；对重要文件记录SHA256。正式跑之前可先做小规模的文件完整性和参数检查，但小规模成功只说明工具入口正常，不能替代真实文库的完整性评估。

执行命令时同时保存标准输出、标准错误、资源统计和结束码。脚本进程退出为零只是第一道检查；若作业被调度器超时终止或日志末尾有被中断迹象，要标记失败。重跑前确认旧进程已经结束，使用新的尝试目录，避免覆盖部分输出导致后续误判。

逐个结果文件核对并解释不确定性

完成后检查每个主要输出文件是否存在且非空，并按当前版本文档确认列名和含义。clusters.tsv应与输入细胞条形码覆盖一致；核对 singlet、doublet、unassigned 等分类比例是否符合项目预期，但不要仅凭比例高低判定成功。若大量细胞未分配，可能涉及测序深度、表达变异位点、过滤、供者数设置或数据结构，应回看每个阶段日志。

查看供者基因型聚类证据、变异位点支持、环境RNA相关结果以及doublet判定。不同供者之间可能遗传接近，遗传簇边界也会受测序覆盖影响。最好将推断结果与已有基因型或实验混样信息比较；若无独立真值，结论应保留为算法推断的簇，而不是已经确认身份。

最后按文库输出一张状态表，记录输入细胞数、被分配细胞数、未分配和双细胞数量、输出文件、软件版本、资源峰值和人工复核结论。不同文库的cluster0不可直接视为同一供者。只有每份文库通过文件、细胞覆盖和科学复核后，才进入跨文库分析；任何失败文库需单独重跑和说明。

异常结果先按证据链排查

如果输出几乎只有一个遗传簇，先检查预期供者数与实际混样设计、每个供者覆盖和可检测变异，而不是立即把k调大。若簇数多于预期，可能是供者数设置、批次结构、双细胞或低质量细胞造成，需要结合基因型支持核实。没有足够证据时，将结果标记为不确定，比强行合并簇更可靠。

若细胞覆盖不完整，先比对输入barcode清单、BAM中的CB标签和输出细胞列表，检查barcode后缀是否统一、是否错误使用另一文库的文件。对输出多出的细胞逐项追溯其来源；不应通过截断输出表或仅保留交集来让数量看上去一致。所有过滤规则都应在重新运行前写入方法记录，并保留未过滤计数。

保护样本身份和大文件的可追溯性

BAM、参考和基因型文件可能包含敏感或受限数据。只在项目授权的存储和计算环境中处理，限制输出目录访问范围，交付时避免把原始BAM或不必要的个体基因型发送到普通共享渠道。即使聚类编号不含姓名，也可能与其他表格关联，因此对外共享应遵循课题组的数据管理要求。

迁移大型输入到服务器时，使用校验值确认传输完整，记录原始文件位置和权限。输出目录按文库和尝试编号隔离，避免不同分析者覆盖。清理临时中间文件前先核对最终产物、日志和可复现材料是否齐全，并遵循数据保留策略；不能只为释放磁盘而删掉唯一的失败证据或重建所需文件。

文库间汇总必须另有映射证据

把每个文库的聚类结果先保存在各自命名空间中，例如library_id加cluster_label作为联合键，避免不同文库都有cluster0时合并冲突。若课题组有已知供者基因型，应在批准的数据处理方案下用独立信息映射到算法簇；映射表单独保存并限制访问。没有该映射时，报告应写“推断遗传簇”，不可把编号改成供者姓名。

跨文库比较时，检查每份文库的供者数、未分配比例、doublet比例及支持位点覆盖。若同一供者在不同批次出现的簇质量差别很大，可能是测序深度或批次导致，不意味着生物学差异。合并分析前定义最低质量标准，未通过的文库应标为需复查，不能为了统一样本数而直接删除后不作说明。

耗时较长时如何安全续作

默认先确认pipeline是否支持阶段续作，以及它如何判断中间文件有效。未经文档支持，不要从目录里挑几个中间文件拼回流程；损坏或旧版本中间文件可能被误用。若工具没有可靠resume方式，重新运行可在条件允许时放入新目录，并保留旧日志供追查。较大任务可先确认本项目最长作业时限和磁盘策略，避免接近时限才发现必须从头开始。

作业被系统终止时，区分超时、内存、磁盘、节点重启和用户主动停止。调度器日志和系统事件往往比最后一行stdout更能解释原因。重新提交前检查磁盘剩余、临时目录和线程限制，并以唯一attempt编号启动。若失败原因与资源有关，先用相同版本和代表输入测量一个阶段，再调整资源；若是文件格式或路径错误，增加资源不会解决问题。

归档前由第二人按文库核对输入散列、有效细胞数、必需结果文件和身份映射依据。未核对的输出保持待复核状态；在报告和共享表格中明确区分工具推断、人工审核和外部基因型确认三类证据。

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

souporcell官方输入、阶段与输出 · 来源资料

souporcell当前pipeline参数和守卫 · 来源资料

samtools quickcheck范围 · 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。