

bulk RNA-seq 实践教程

导出日期：2026年6月27日

bulk RNA-seq 实践教程

bulk RNA-seq 把一份组织或一群细胞打成混合样本测序，得到的是每个基因在样本里的平均表达量。多数实验室的转录组项目仍然围绕它展开：几十个样本、两三个处理组、一篇文章的主线分析。

本专栏打算把 bulk RNA-seq 从"测完序之后怎么办"这一步开始讲，按真实项目常见的顺序走一遍：质控、比对或准比对、定量、差异分析、富集解释。尽量给出能在自己电脑上跑完的数据和脚本，而不是只列工具名。

bulk vs 单细胞 vs 空间：什么时候选哪个

在动手之前先回答一个问题：你的科学问题真的需要 bulk RNA-seq 吗？

问题	推荐技术	理由
两组样本之间整体差异	bulk RNA-seq	便宜、统计成熟、3-5 个样本就够
不同细胞类型的响应不同	单细胞 RNA-seq	bulk 看到的是各类细胞的加权平均，细节被抹掉
表达定位（哪一块组织在做什么）	空间转录组	bulk 和单细胞都没有空间信息
稀有亚群（< 5%）的变化	单细胞	bulk 里 5% 的信号被 95% 稀释
大样本 + 临床/流行病学问题	bulk RNA-seq	TCGA 上千样本量级，单细胞做不起
转录起始位点 / 可变剪接	bulk + 长读长	标准 bulk 也能做，但 Iso-seq / Nanopore 直接

bulk 还是大多数项目的默认起点：便宜、流程成熟、统计学经过 20 年验证、可以做大样本。除非你的问题明显需要细胞分辨率或空间信息，否则先做 bulk。

一个项目大致是什么样子

假设你做了一个时间点比较的实验：对照组三个样本、处理组三个样本、每个样本一次双端测序。下机之后你会拿到 12 个 FASTQ 文件，加起来可能有几十 GB。

从这里到"一张火山图 + 一张 GO 富集结果 + 一句能写进论文的结论"之间，要经过的步骤大致是：

步骤	典型产物	常用工具
原始数据质控	每个样本的 QC 报告	FastQC、MultiQC
比对到参考基因组	每个样本的 BAM	STAR、HISAT2
基因或转录本定量	counts 矩阵 或 TPM 矩阵	featureCounts、Salmon、kallisto
样本层面检查	PCA 图、样本相关性	DESeq2、edgeR、limma
差异表达分析	差异基因表	DESeq2、edgeR、limma-voom
功能富集	GO、KEGG、GSEA 结果	clusterProfiler、fgsea
结果交付	表格、图、方法段	R Markdown、Quarto

"比对 + featureCounts" 和 "Salmon/kallisto 准比对" 是两条常见路线。前者产出 BAM、对可视化和变异分析都友好；后者跳过显式比对，直接估计转录本丰度，速度快、磁盘友好。新项目里如果只关心基因层面的差异表达，用 Salmon 是更经济的选择。

常见工具栈

在 BioF3 的例子里会优先使用下面这套组合，原因是文档、社区和版本都比较稳定：

阶段	工具	运行环境
质控	FastQC、MultiQC	bash
比对	STAR、HISAT2	bash
准比对定量	Salmon、tximport	bash + R
差异分析	DESeq2、edgeR、limma-voom	R
富集分析	clusterProfiler、fgsea	R
可视化	ggplot2、ComplexHeatmap、EnhancedVolcano	R

其他工具（Kallisto、RSEM、StringTie、pheatmap 等）都是合理替代，但这套组合能覆盖大部分教学和小型真实项目。

推荐公开数据集

学这个专栏不一定要用自己的测序数据。下面这几份公开数据都能用来练手，且每份都有配套文献可对照。

数据集	规模	适合	入口
Pasilla (果蝇 RNA-seq)	7 个样本	最小差异分析例子, Bioconductor 官方包 <code>pasilla</code> 里直接带	pasilla
airway (人肺上皮, 4 对照 + 4 处理)	8 个样本	DESeq2 官方教程数据, 做 glucocorticoid 处理	airway
GTEX 子集	可选子集	多组织 / 多样本练习归一化和批次	GTEX
Recount3 项目	任选一个 SRP	预处理好的 counts 矩阵, 跳过比对	Recount3

前两个数据集在 R 里装一下包就能直接用, 不需要下载 FASTQ, 非常适合作为入门。GTEX 和 Recount3 的优点是真实样本量足够做归一化和批次演示。

最小可跑的例子

下面这段代码从安装 DESeq2 到跑出一份差异基因结果一共不到 20 行, 数据就是上面 airway 公共数据集。把它粘到 RStudio 里, 装好包就能跑:

```
# 一次性安装依赖 (如果还没装)
if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")
BiocManager::install(c("DESeq2", "airway"))

library(DESeq2)
library(airway)

# 加载示例数据: 8 个样本 x ~64000 个基因的 counts 矩阵
data(airway)
se <- airway
se$dex <- relevel(se$dex, ref = "untrt")

# 构建 DESeq 对象并拟合模型
dds <- DESeqDataSet(se, design = ~ cell + dex)
dds <- DESeq(dds)

# 提取差异基因 (trt vs untrt)
res <- results(dds, contrast = c("dex", "trt", "untrt"))
summary(res)

# 看一下最显著的几个基因
head(res[order(res$padj), ])
```

运行之后 `summary(res)` 会告诉你上下调基因各多少个, `head(res[order(res$padj), ])` 会列出按校正 p 值排的前几个基因。把这几行读懂, 基本就知道 bulk RNA-seq 差异分析是在做什么了。

真实项目比这段要多两件事: 样本多、要考虑批次和协变量; 定量步骤之前还有比对和质控。这也是后面几个模块要展开讲的内容。

专栏模块规划

模块	主题	状态
01	项目结构、数据入口与 Salmon 定量	已上线
02	DESeq2 差异表达分析	已上线
03	GO / KEGG / GSEA 富集分析	已上线
04	火山图、热图与富集可视化	已上线
05	多时间点与 LRT 检验	已上线
06	批次效应: ComBat / SVA / batch-in-design	已上线
07	DESeq2 / edgeR / limma-voom 三工具对比	已上线
08	可复现分析报告与结果交付	已上线

所有 8 个模块都已上线，前七个模块带一份可以直接跑的 R 脚本。

推荐前置知识

- [编程基础: R、Python、Bash 学习路径](#)
- [R 数据整理与 ggplot2 可视化](#)
- [公共数据库与数据检索](#)

单细胞项目里的质控、归一化和差异分析思路和 bulk 有重叠，已经看过[单细胞实践教程](#)的读者会觉得很多概念上手更快。

常见误区

新手做 bulk RNA-seq 经常掉的几个坑：

误区 1：只做差异分析就交付

差异基因列表 ≠ 完整结果。至少还要补：**PCA / 样本距离的 QC 图、富集分析、关键基因的表达分布图**。只给一份 padj 排序的 csv 给合作者，对方看不出哪些是真信号、哪些是边缘 noise。

误区 2：不看 PCA 直接跑差异

PCA 是“测序质量 + 设计是否合理”的总体检查。**PC1 不分组 = 处理效应弱（或者 design 错了）**；某个样本远离同组其他样本 = 该样本可能有问题。这两件事不解决，跑出来的差异基因没法解释。

误区 3：把 $|\text{LFC}| > 1$ 当成生物学显著的硬标准

$|\log_2\text{FC}| > 1$ 即 fold change > 2 倍。这个阈值在不同项目里意义完全不同：转录因子表达 1.5 倍可能就是大事，组织特异性基因 10 倍变化才有意义。先看自己数据的整体分布，再选阈值。

误区 4：富集结果只看 top 5 通路

GO / KEGG 跑出来几十条通路，只看 top 5 容易被冗余 term（“细胞增殖” / “正向调控细胞增殖” / “细胞周期”基本是同一件事）误导。用 **cnetplot / emapplot** 把通路网络化，看真正的几个独立模块。

误区 5：发表前不锁版本

DESeq2 / clusterProfiler / Salmon 半年一更，包升级有时会改算法默认值。发表后审稿人或读者重跑结果不一致就麻烦。用 `renv` 或 `conda env.yaml` 锁版本，08 章会展开。

参考资源

- [DESeq2 官方教程](#)
- [edgeR 用户指南](#)
- [limma 用户手册](#)
- [RNA-seq workflow \(Bioconductor\)](#)
- [Recount3](#)
- [GTEx Portal](#)



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。