DSB多样本标准化实操：背景矩阵配对、独立任务与结果保存

完整操作教程 · 参考固定 · 往返恢复验收

按真实背景批次配对细胞与空液滴ADT矩阵，运行dsb并检查输出名称、有限值与失败原因。

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。

DSB标准化的第一步是配对背景，不能把几份CITE-seq样本按文件名循环就算完成。细胞ADT和空液滴ADT应来自相应实验背景，保留抗体面板、染色、文库和批次信息。先在一批已完成细胞与背景质控的数据上跑通，再扩展到队列。

一个任务对应一套背景关系

样本被混在同一实验文库时，如何使用共同背景需按实验设计确定，不机械地按供者拆开重估。独立染色或不同技术批次也不默认复用同一空液滴矩阵。本文限定有真实空液滴及同型对照的DSBNormalizeProtein路线；无空液滴时官方另有ModelNegativeADTnorm，不能补全零背景来冒充原方法。

输入	本例要求	先处理的问题
cell	蛋白×细胞原始ADT计数	拒绝CLR、log或已缩放矩阵
empty	同背景蛋白×空液滴计数	结合RNA、ADT判断背景，排除残留细胞
蛋白行	名称唯一且两矩阵完全同序	同名不同抗体不能只改名合并
isotypes	真实同型对照的准确名称	没有时另选官方适用设置

保存一批可追溯输入

batch-input.rds保存列表cell、empty、isotypes。矩阵列名分别是本批细胞和背景条形码，二者不重叠。本例明确要求已经准备好的数值普通矩阵，避免代码静默把巨大稀疏矩阵整体转稠密；从稀疏输入转换前先核对容量。粗算双精度载荷可用八乘蛋白数乘两类液滴总数，实际峰值还含副本与模型。

把背景筛选条件、输入校验值及条形码清单放在输入旁，不照搬示例数据的固定阈值。背景过少、混入大量细胞或遗漏主要背景群体都应回到上游检查，不能仅凭标准化输出“看起来更清楚”选择参数。

独立目录运行并立即验收

在已固定版本且安装dsb的R环境执行以下脚本。当前官方说明覆盖dsb 2.0系列的两类路线；这里保留默认算法，不开启可选fast.km。若以后切换快速模式，另记参数并在同一输入核对结果，不把它当无条件等价的资源开关。

```
library(dsb)
x <- readRDS("batch-input.rds")
cell <- x$cell
empty <- x$empty
stopifnot(is.matrix(cell), is.numeric(cell), is.matrix(empty), is.numeric(empty))
for (m in list(cell, empty)) {
  stopifnot(all(dim(m) > 0), all(is.finite(m)), all(m >= 0),
            all(m == floor(m)), !is.null(rownames(m)), !is.null(colnames(m)),
            !anyNA(rownames(m)), !anyNA(colnames(m)),
            all(nzchar(rownames(m))), all(nzchar(colnames(m))),
            !anyDuplicated(rownames(m)), !anyDuplicated(colnames(m)))
}
stopifnot(identical(rownames(cell), rownames(empty)),
          !any(colnames(cell) %in% colnames(empty)),
          length(x$isotypes) > 0, !anyNA(x$isotypes),
          !anyDuplicated(x$isotypes), all(x$isotypes %in% rownames(cell)))
out <- "dsb-batch01"
stopifnot(!file.exists(out), dir.create(out))
```

```
set.seed(101)
normalized <- DSBNormalizeProtein(
  cell_protein_matrix = cell, empty_drop_matrix = empty,
  denoise.counts = TRUE, use.isotype.control = TRUE,
  isotype.control.name.vec = x$isotypes)
stopifnot(is.matrix(normalized), identical(dimnames(normalized), dimnames(cell)),
          all(is.finite(normalized)))
saveRDS(normalized, file.path(out, "normalized.rds"))
write.csv(data.frame(protein=rownames(cell),
                     background_sd=apply(log(empty+10), 1, sd)),
          file.path(out, "background-diagnostic.csv"), row.names=FALSE)
writeLines(capture.output(sessionInfo()), file.path(out, "session.txt"))
saveRDS(list(seed=101, isotypes=x$isotypes, denoise=TRUE),
        file.path(out, "parameters.rds"))
```

完成后检查什么

normalized.rds应保持原细胞矩阵的蛋白和细胞顺序，数值必须有限。DSB值可以为负，不能为方便展示就截成零再用于统计。比较关键蛋白在背景与细胞中的分布、同型对照和预期阴阳群体；维度检查只证明结构，不证明背景选择或生物学分群正确。

若出现非有限值，检查原始计数、背景中的蛋白变化及模型日志；诊断表用伪计数十展示背景对数标准差，不是自动剔除规则。若决定删去无有效染色蛋白，必须在两矩阵同步处理并记下依据。多批次分别成功也不等于消除了染色批次效应。

每个独立任务保留输入映射、参数、完整结果和session信息；失败保留目录，复核后换新目录重跑。先测单任务实际RAM和运行时间，再安排获配范围内的并发，不能从蛋白数少推断任意数量背景液滴都能一次读入。

空液滴和细胞矩阵必须成对

DSB的背景校正依赖空液滴中的蛋白背景分布。cell_protein_matrix和empty_drop_matrix应来自同一实验批次、相同抗体面板和处理流程，行名按蛋白逐项对齐，列名互不重叠。确认空液滴确实没有被当作细胞保留，且两矩阵的数值是原始非负整数计数。若只找到细胞矩阵而没有合适的空液滴，不能复制其他批次背景或将细胞随机打乱来伪造背景。

等位/同型对照抗体名称必须和面板注释完全匹配。检查提供的isotype向量是否在蛋白行名中、是否重复，并记录面板版本。若同型对照不存在或名称不确定，应按DSB官方方法选择适用设置并说明限制，而不是随意挑几个高表达抗体。批次间抗体克隆、试剂批号或面板变化也会影响背景分布，需要按实验批次分别分析或证明可合并。

DSB步骤和输出解释

DSB通常包含空液滴背景校正和细胞内技术噪声去除等步骤。是否使用isotype controls、是否运行denoise counts，应结合实验设置和算法适用条件确定。运行后先核对归一化矩阵的行列名与输入细胞矩阵完全一致、数值有限，并检查每种蛋白的分布和背景诊断。校正值可以为负数，这是标准化尺度上的结果，不应直接截为零或解释为“负蛋白表达”。

background-diagnostic只展示空液滴背景的统计特征，不是自动判定过滤阈值。对每个抗体查看细胞信号与空液滴背景的分离、同型对照表现和已知阳性/阴性群。若所有抗体都没有合理分离，先检查面板、空液滴选择、计数矩阵和实验批次，不能靠调高算力解决信号不足。报告中保留原始计数、DSB矩阵及参数，便于回看标准化前后差异。

分批处理与计算资源

蛋白面板维度通常小于基因表达矩阵，但细胞数和空液滴数量仍会影响对象规模。先在一个代表批次上测量读入、归一化和保存阶段的峰值RSS，再决定是否逐批运行。若逐批，需保持相同面板、空液滴选择原则和参数，同时保留批次ID；不能把不同批次空液滴合并后默认代表共同背景。

资源异常时查看是否将稀疏输入误转成全量dense、是否重复保留多个细胞/背景矩阵副本，以及作业内存限制。输出文件较小不代表运行峰值低。多个R进程同时处理不同批次会重复加载包和矩阵，先测并发再扩展。数据受限项目还需确认空液滴和细胞数据是否可以跨用户共享，以及处理后矩阵的存储和清理策略。

报告与质量边界

方法报告至少写出使用的输入矩阵、空液滴数、面板版本、同型抗体、denoise设置、DSB包版本、细胞数和批次处理方式。展示DSB前后关键抗体分布或对照群的变化，说明采用DSB的依据。结果是相对背景校正的蛋白信号，不能单独当作绝对蛋白浓度或跨平台通用标尺。

如果缺少匹配空液滴或样本规模不满足方法条件，应明确说明限制并评估其他适用方案。不要把没有背景信息的矩阵输入同一流程后，用“成功生成输出”替代方法有效性判断。给出的练习代码只检查矩阵结构和调用路径，不含真实实验数据，也不能证明某个实验批次适合使用DSB。

没有空液滴时怎样决策

首先回看原始实验输出是否保存了未过滤矩阵，且空液滴列是否仍在；检查过滤步骤是否只导出了细胞矩阵。若原始材料可恢复，按同一抗体面板和批次定义空液滴，再结合官方方法确定合适背景集合。若材料已经丢失，不要从细胞矩阵中抽取低信号细胞冒充空液滴；该做法改变背景分布，不能声称完成了DSB背景校正。

如果现有实验条件不适合DSB，应比较可用的蛋白计数标准化方案，并写明为何选择该方案。不同方法得到的尺度和负值含义可能不同，不应把处理后的数值直接拼接。保留各方法的参数与诊断图，邀请对实验设计熟悉的人员核查背景定义。硬件升级不能弥补关键实验对照缺失。

跨批次比较的注意事项

DSB结果主要依据每个实验批次的背景分布进行校正。跨批次比较前核对抗体面板、仪器、试剂批号、空液滴数量和背景变化；如果批次差异明显，先考虑分批校正与后续整合，而不是将所有空液滴和细胞直接拼在一起。选择共同背景或逐批背景会影响结果，应根据实验和研究问题预先设计。

把DSB归一化矩阵用于下游时，保留样本和批次列，避免因矩阵数值形式类似就忘记它与RNA表达数据的尺度不同。差异分析模型应使用适合蛋白标准化值的假设和生物学重复，而不是把单细胞数量当独立重复。说明负值的来源，并在图例中写清展示的是标准化值还是原始计数。

实验批次归档

每个批次归档cell matrix与empty drop matrix的来源及散列、抗体面板清单、同型抗体配置、DSB版本、运行参数、诊断结果和输出对象。若原始数据不能长期保留，至少记录受控存储位置、访问负责人和保留期限。重新运行时应可确认使用的是同一对背景与细胞矩阵，而不是名称相同但来源不同的文件。

提交结果前复核输出细胞数、蛋白数、缺失值和有限性，并抽查若干蛋白的归一化前后分布。若有异常蛋白或实验批次，记录排除理由和对下游样本数的影响。最终结果交付应附方法说明和适用限制，不要用一张标准化热图替代输入质控与背景证据。

先用诊断图识别背景异常

绘制每个蛋白在空液滴和细胞中的原始计数分布，查看同型抗体是否表现出预期背景，以及少数抗体是否存在极高背景或近乎无信号。对异常蛋白检查抗体名称、面板注释、计数单位和批次记录。DSB生成的背景标准差表只是一项诊断，不能由一个统计量自动决定某抗体是否删除。

在同一输入上保存原始分布与校正后分布，并随机抽查细胞ID和蛋白顺序是否保持。若某批次归一化后值异常集中或大量非有限值，检查输入是否含错误尺度、无效空液滴列或常数背景。把诊断图与运行参数归档，后续升级包版本或改变背景定义时可对照变化。

写清楚标准化值的统计用途

DSB标准化矩阵与原始计数的含义不同。下游展示时标注使用标准化蛋白值，不要把负数裁剪后称为原始表达。差异分析应考虑生物学重复并选择适合该尺度的方法；每个细胞不是独立生物学重复。若要合并RNA和蛋白特征，明确两种模态的尺度、缺失情况和整合方法，避免把蛋白标准化结果直接当作RNA表达量。

不同背景处理方法可能改变细胞之间的排序。对关键结论，可对比合理的背景设定或适用的替代标准化方法，重点观察阳性/阴性对照和主要细胞群是否稳定。敏感性分析用于说明方法依赖范围，不应只挑使组间差异最大的设置。若不同方法得出相反结论，应报告这种不确定性并回看实验控制。

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

dsb官方仓库与当前参数 · 来源资料

官方完整CITE-seq流程 · 来源资料

官方多批次与FAQ · 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。