QIIME 2 DADA2 试跑：按读数、截断参数和线程排查

先看每样本 reads 留存和质量曲线，再确定双端截断长度、重叠区间与线程数，避免盲目调参或占满节点。

QIIME 2 的 DADA2 长时间运行或输出 reads 大量减少时，先不要直接加线程或放宽过滤。先核对样本拆分、引物/接头去除、质量曲线、双端读段顺序和目标扩增子长度。对于 paired-end，正向与反向截断后还必须保留足够重叠区才能合并；缩短 truncation 虽可删除低质量尾部，也可能让两端无法合并。

先确认输入是否满足分析前提

记录 QIIME 2 发行版、DADA2 插件版本、样本数量、每个样本原始 reads 数和输入文件清单。确认 demultiplex 后每个样本的正反向 reads 一一对应，非生物序列已按实验方案移除。先查看代表样本的正反向质量曲线，再根据引物区域、扩增子长度和质量下降位置确定 trim-left 与 trunc-len；不能照抄其他项目的固定截断值。

用质量曲线和重叠长度共同选截断

paired-end 的截断参数会丢弃读段 3' 末端。选择前先估算截断后理论重叠长度：正向长度加反向长度，再减去扩增子长度；还需考虑片段长度变异。QIIME 2 当前文档要求合并时至少保留 12 个核苷酸重叠。ITS 等长度变异大的扩增子不宜机械套用固定截断长度，应核对目标区域与所用方法的约束。

现象	查看统计	优先处理
某一步骤后 reads 急剧减少	每样本 denoising stats 中各阶段留存数	判断是质量过滤、合并还是嵌合体去除造成
正反向大量无法合并	实际截断长度、扩增子长度和重叠区	先验证重叠是否满足要求，再调整质量截断
运行速度慢但 CPU 不忙	线程分配、节点负载、输入/临时目录 I/O	检查存储和并发，而非只增加线程
部分样本几乎无输出	原始 reads、过滤保留率、样本标签	按样本追踪输入与每阶段统计

在线程申请范围内做小步试跑

QIIME 2 的 n_threads=0 表示使用所有可用核心；共享集群里应显式设置不超过作业申请的线程数。先用代表性样本跑通参数，再处理全批次。运行时记录作业峰值内存、CPU 利用率和临时盘空间；提高线程不一定能按比例缩短时间，也可能增加同节点竞争。

```
set -euo pipefail
: "${TRUNC_F:?set forward truncation length}"
: "${TRUNC_R:?set reverse truncation length}"
: "${THREADS:?set threads allocated by scheduler}"
test -r demux-paired.qza
test ! -e dada2-table.qza
qiime dada2 denoise-paired \
  --i-demultiplexed-seqs demux-paired.qza \
  --p-trunc-len-f "$TRUNC_F" --p-trunc-len-r "$TRUNC_R" \
  --p-n-threads "$THREADS" \
  --o-table dada2-table.qza \
  --o-representative-sequences dada2-rep-seqs.qza \
  --o-denoising-stats dada2-stats.qza
```

该命令保留输出对象为 QIIME 2 artifact；实际参数名应以所装发行版的 qiime dada2 denoise-paired --help 为准。正式批次前可将 stats 可视化并逐样本检查，而不是只看总 reads 数。

用留存统计判断是否“跑成功”

DADA2 会输出各样本在过滤、去噪、双端合并和嵌合体去除等阶段的 reads 统计。逐样本查看留存比例，检查是否有样本异常掉到接近零、合并率是否与扩增子长度相符，并结合代表序列和 feature table 验证结果。保留命令、软件版本、质量图、样本清单和 stats 文件。若改变 truncation 或 max expected errors 等过滤参数，应说明其依据并重新评估样本间留存差异；不能只为提高保留率而忽略序列质量和合并条件。

资源申请可依据代表样本实际峰值和样本数估算，但小样本试跑不一定覆盖全批次最复杂的样本。需要扩大规模时，逐步调整并发并观察计算与磁盘 I/O，保留足够输出空间和日志，才能复现每次过滤参数选择。

参考资料

QIIME 2 amplicon 当前版 DADA2 插件说明 · 官方资料

DADA2 官方流程教程与质量/截断说明 · 官方资料

GNU time 资源统计手册 · 官方资料