fastANI比较任务：基因组清单、成对调度与矩阵核验

先定义fastANI有向任务规模，检查五列结果和缺失项，再分批比较。

fastANI的大量比较先要定义方向：A作为query与B作为reference，和反向任务不完全相同。只计算无序组合再当作所有方向的结果，可能丢掉查询片段分母和方向差异。本篇用两份明确清单运行查询对参考比较，按五列结果验收，不把缺失项填成零。

先写出任务规模与清单

输入为适用的微生物基因组组装，质量和比较范围已论证，不将短contig或原始reads当完整基因组使用。每份清单一行一个绝对FASTA路径，规范路径且无重复，样本ID与路径有独立映射。query有Q份、reference有R份时，计划的有向任务数为Q乘R；同一清单全比较则有N平方项。

对象	本例约定	验收要点
方向	query对reference	保留原方向，不默认对称
输出	query、reference、ANI、匹配片段、query总片段	按查询分母计算比对比例
缺失	不满足报告条件等原因	不是ANI为零或自动100
线程	示例4线程	完整时间与内存另做测量

先建立可核对的查询批次

下例假定两份清单已完成路径、文件、重复和研究范围检查。保存软件帮助与输入清单，参数及碎片长度等默认值按安装版记录；不要为了得到完整矩阵随意降低报告条件。工作目录采用新编号，文件未生成与有效零行输出分开检查。

```
#!/usr/bin/env bash
set -euo pipefail
queries=/data/query-paths.txt
references=/data/reference-paths.txt
test -s "$queries"
test -s "$references"
mkdir fastani-pilot-001
root="$PWD/fastani-pilot-001"
cp "$queries" "$root/queries.txt"
cp "$references" "$root/references.txt"
fastANI -h > "$root/help.txt" 2>&1
/usr/bin/time -v -o "$root/resources.txt" \
  fastANI --ql "$queries" --rl "$references" -t 4 \
  -o "$root/ani.tsv" > "$root/stdout.txt" 2> "$root/run.log"
test -f "$root/ani.tsv"
python3 - "$root" <<'PY'
import math,sys
from pathlib import Path
root=Path(sys.argv[1])
def listing(path):
    rows=path.read_text().splitlines()
    assert rows and all(rows) and len(rows)==len(set(rows))
    return set(rows)
queries=listing(root/'queries.txt');refs=listing(root/'references.txt');seen=set()
with (root/'ani.tsv').open() as f:
    for line in f:
        x=line.rstrip('\n').split('\t');assert len(x)==5
        q,r=x[:2];assert q in queries and r in refs and (q,r) not in seen
        value=float(x[2]);matched,total=map(int,x[3:])
        assert math.isfinite(value) and 0<=value<=100 and 0<=matched<=total and total>0
        seen.add((q,r))
print('reported, not reported',len(seen),len(queries)*len(refs)-len(seen))
PY
```

代码检查输出方向、ID、数值和重复记录，未报告计数不解释具体原因。官方说明较低相似度的组合可能不报告ANI，短或低质量输入及匹配比例条件也要结合日志核对。不要以补零后的热图宣称远缘关系已精确量化。

矩阵展示保留方法含义

匹配片段数除以query总片段数是相对于查询组装的比例，不能直接换成reference覆盖率或所有碱基精确比对比例。反向ANI可略有差异；若使用matrix模式，官方采用的对称显示规则需一并记录，不能拿它替代原始有向五列表做所有分析。

只有一部分组合需要比较时，明确查询集合与参考集合能减少无关任务。分批可以按官方支持方式拆参考清单，保持完整query和参数一致，合并前核分片互不重叠且完全覆盖原参考；不能重复运行同一对后无条件平均，也不能把失败分片当未报告组合。

按组合数与数据质量测资源

代表任务记录Q、R、总组装长度、完整时间、RAM及输出，不能只按单个FASTA大小推算全比较。先测单任务线程，再测小并发；参考索引、共享盘读取及内存会影响收益。先完成组装质量和方法检查，再讨论真实吞吐缺口。

归档两份原始清单、样本路径映射、命令、帮助、日志和五列原表。清单路径变更或增加样本后，重新计算计划组合并核查已完成集合；最终报告分别列失败、未报告和已报告结果，避免把不完整队列误当完整距离矩阵。

上面按组合数计算未报告数量，没有在RAM中展开N平方的完整列表；大队列宜保留原始稀疏结果，再按研究需要生成展示矩阵。

参考资料

FastANI官方输入输出及方向说明 https://github.com/ParBLiSS/FastANI

FastANI官方线程及参数实现 https://github.com/ParBLiSS/FastANI/blob/master/src/map/include/parseCmdArgs.hpp

FastANI原始研究 https://doi.org/10.1038/s41467-018-07641-9