用文件清单连接OpenMS处理步骤与结果核验

完整操作教程 · 参考固定 · 往返恢复验收

OpenMS批处理先把每份文件和每个步骤对应起来。输入转换、峰提取、鉴定和定量是不同任务，不能把某个工具退出零当成完整蛋白组工作流完成。本篇固定OpenMS3.6.0，对适用的高分辨率profile mzML做

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。

OpenMS批处理先把每份文件和每个步骤对应起来。输入转换、峰提取、鉴定和定量是不同任务，不能把某个工具退出零当成完整蛋白组工作流完成。本篇固定OpenMS3.6.0，对适用的高分辨率profile mzML做PeakPickerHiRes，再用FileInfo读取前后文件，建立可复查的小批次。

输入必须适合这个步骤

files.tsv无表头，每行文件ID与mzML绝对路径，TAB分隔。示例S1、S2各对应一份经过转换验收的profile数据，保留仪器、转换软件、MS级别与centroid/profile标识。厂商RAW转mzML属于上游任务，已centroid的数据不能无说明再次进行profile峰提取。

环节	本例内容	验收
输入	适用高分辨率profile mzML	格式、谱图数量及元信息
参数	固定PeakPickerHiRes INI	记录算法和MS级别
步骤依赖	输入检查→峰提取→输出检查	前步失败即停止
结果	centroid mzML及FileInfo	不等于鉴定/定量结果
批次	ID与输出目录一一对应	缺文件不能静默跳过

生成参数，顺序连接三个步骤

先在目标环境运行PeakPickerHiRes -write_ini picker.ini，按项目验证的算法设置检查并保存该INI。下面保存为run_openms.py。-threads 2为演示的TOPP线程设置，不承诺每个工具或每个阶段都能使用两个核；processOption等内存方式不在脚本中暗改。

```
import re,subprocess
from pathlib import Path
rows=[r.split("\t") for r in Path("files.tsv").read_text().splitlines()]
assert rows and all(len(r)==2 for r in rows)
ids=[r[0] for r in rows];paths=[Path(r[1]) for r in rows]
assert len(set(ids))==len(ids) and all(re.fullmatch(r"[A-Za-z0-9_-]+",x) for x in ids)
assert all(p.is_absolute() and p.is_file() and p.stat().st_size>0 for p in paths)
assert len({p.resolve() for p in paths})==len(paths)
params=Path("picker.ini").resolve();assert params.is_file()
root=Path("openms-batch-001");root.mkdir()
for sample,path in rows:
    d=root/sample;d.mkdir();picked=d/"centroid.mzML"
    with (d/"run.log").open("w") as log:
        subprocess.run(["FileInfo","-in",path,"-out",str(d/"input.info.txt"),"-m"],
                       stdout=log,stderr=subprocess.STDOUT,check=True)
        subprocess.run(["/usr/bin/time","-v","-o",str(d/"resources.txt"),
            "PeakPickerHiRes","-ini",str(params),"-in",path,"-out",str(picked),"-
threads","2"],
            stdout=log,stderr=subprocess.STDOUT,check=True)
        assert picked.is_file() and picked.stat().st_size>0
        subprocess.run(["FileInfo","-in",str(picked),"-out",str(d/"output.info.txt"),"-
m","-c"],
                       stdout=log,stderr=subprocess.STDOUT,check=True)
    assert (d/"output.info.txt").stat().st_size>0
    (d/"FILES_CHECKED").write_text("inspect spectrum types, counts and signals before
accepting\n")
```

比较前后元数据和实际峰形

读取两个FileInfo报告，核对谱图/MS级别、保留时间范围、前体信息及处理记录；按本次处理范围确认谱图对应，不能只比输出字节数。选择代表谱图查看峰位置、强度和分辨率，判断centroid结果是否合理。FileInfo的格式检查不能替代仪器和算法适用性评估。

FILES_CHECKED只记录命令链和最低文件检查通过。每个预期ID均应有两份报告及输出，缺一个就标未完成；输入没有谱图、错误XML或损坏索引应回到转换与文件完整性。参数文件若与版本不匹配，先解决INI结构和算法差异，不通过force绕过检查。

后续步骤另有自己的输入合同

需要鉴定时，再确定检索引擎、蛋白数据库、酶切、修饰及FDR流程；需要定量时，再确定实验设计、对齐、缺失与汇总方法。不能将centroid mzML命名为“蛋白定量结果”，也不把批量峰提取成功说成候选蛋白已经可靠检出。

保存原mzML、转换记录、INI、工具版本与输出校验值。若调整信噪阈值、MS级别或处理方式，应在新目录重跑并比较代表谱图；不要只对失败样本更改方法后混进统一批次，保留每次尝试的来源和差异。

按步骤测量CPU与工作盘

分别记录文件读取、峰提取、输出检查的完整时间、RAM与存储；压缩mzML解码后内存可能明显不同，不能按压缩体积乘固定倍数报价。PeakPickerHiRes可有不同内存处理方式，若要切换，先按该版本支持与结果一致性验证，再做资源比较。

先用一份代表文件和一份较大的文件测量，再试少量并发。多个工具同时读取原mzML并写新文件，可能先受共享盘限制；每任务线程与外层任务数共同计入额度。保留原始、转换、处理和待交接副本的共存量，确认科学验收后再执行项目保留策略。

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

OpenMS3.6.0 PeakPickerHiRes参数与适用输入 · 来源资料
OpenMS3.6.0 FileInfo输入与检查参数 · 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。