多物种共线性任务编排：同源输入、坐标和成对任务验收

完整操作教程 · 参考固定 · 往返恢复验收

多物种共线性先要解决标识和坐标一致，增加并发不能补救错误输入。本篇以MCScanX的两个物种任务为单位：整理四列坐标和BLAST格式同源结果，通过检查后运行，再登记每一对是否产生可解释的共线性块。多物种总任务量按物种对规划。

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。

多物种共线性先要解决标识和坐标一致，增加并发不能补救错误输入。本篇以MCScanX的两个物种任务为单位：整理四列坐标和BLAST格式同源结果，通过检查后运行，再登记每一对是否产生可解释的共线性块。多物种总任务量按物种对规划。

按程序实际读取格式准备输入

当前官方源码读取前缀.gff与前缀.blast，结果为前缀.collinearity。这里的.gff是简化四列表，不是直接传入九列GFF3；README部分旧文字写bed或synteny，不能据此猜实际文件名。蛋白质同源搜索与注释应使用同一套代表转录本，先保存基因ID映射。

输入项目	本例约定	常见错误
pair.gff	染色体、基因ID、起点、终点，TAB分隔	把原始GFF3直接改扩展名
物种前缀	AA_chr1与BB_chr1	两个物种均叫chr1
pair.blast	标准BLAST m8的12列	查询/目标ID与坐标表不一致
pair.collinearity	本次物种对结果	只看文件存在，不看块和锚点

源码用染色体名称前两个字符识别物种，因此两物种必须使用不同的两字符前缀。当前读取逻辑把第三列作为位置参与排序，不是用起终点自动计算中点；坐标应来自统一协议。下面的检查采用非负整数、32位范围和两个物种的明确约束，不替代对注释来源的生物学审核。

先检查标识，再交给MCScanX

将以下保存为check_pair.py，与pair.gff、pair.blast放在同一项目目录。检查器允许物种内同源命中，但要求至少一个跨物种命中；全部是物种内比对时，不能把空的跨物种分析误当正常资源试跑。

```
import math
genes={}
with open("pair.gff") as f:
    for line in f:
        fields=line.rstrip("\n").split("\t")
        assert len(fields)==4, "■■■■■■■■■"
        chrom,gene,left,right=fields
        assert len(chrom)>2 and gene and gene not in genes
        left,right=int(left),int(right)
        assert 0<=left<=right<2**31
        genes[gene]=chrom[:2]
assert genes and len(set(genes.values()))==2
cross=0
with open("pair.blast") as f:
    for line in f:
        fields=line.split()
        assert len(fields)==12, "■■■■■■■■12■"
        q,s=fields[:2]
        assert q in genes and s in genes, "■■ID■■■"
        e=float(fields[10]);assert math.isfinite(e) and e>=0
        cross+=genes[q]!=genes[s]
assert cross>0, "■■■■■■■■"
print("genes",len(genes),"cross_species_hits",cross)
```

```
set -euo pipefail
python3 check_pair.py > pair.input-check.txt
mkdir mcscan-AA-BB-001
cp pair.gff pair.blast mcscan-AA-BB-001/
cd mcscan-AA-BB-001
sha256sum pair.gff pair.blast > inputs.sha256
/usr/bin/time -v -o run.time MCScanX pair -a -b 2 > run.log 2>&1
test -s pair.collinearity
```

验收要区分空块与进程失败

-a仅计算成对共线性块，-b 2限制跨物种块，适用于本例；不要把它理解成只启动两个线程。读取collinearity中的实际块记录和锚点数，只有注释或表头不代表发现了共线性。没有块时依次检查ID交集、物种前缀、同源命中分布、注释版本与参数，再判断是否为真实生物学差异。

同源搜索的过滤协议会影响锚点，不为得到更多连线反复放宽阈值后只保留最好看的结果。记录过滤前后命中数和方法理由，绘图使用同一份经过验收的块文件。必要时抽查一组锚点在原注释中的位置和方向，确认没有把转录本ID与基因ID混合。

扩为多物种任务清单

N个物种的无向两两比较共有N×(N−1)/2对。任务表至少写物种对、两个蛋白集和注释的SHA、搜索设置、输出前缀、退出状态、实际块数及检查结论。每对建立新目录，先测同源搜索和MCScanX各自耗时、RSS与磁盘；它们不是一个资源阶段。确认单对正确后再增加外层并发，并保留无块及失败任务，不能在汇总时静默丢弃。

例如AA_chr1、AA_gene001、100、900与BB_chr1、BB_gene007、500、1200可作为两条坐标记录，而同源表必须仍使用AA_gene001和BB_gene007。若蛋白标题带额外说明，只统一取ID部分并保存对应表；不能随意截断版本号，使多个原始ID合并成一个。

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

MCScanX官方说明 · 来源资料

MCScanX实际坐标与物种读取 · 来源资料

MCScanX命令及输出实现 · 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。