二代宏基因组流程使用说明

|  |  |  |
| --- | --- | --- |
| 编制人：涂成芳 | 审核人：张阳 | 批准人： |
| 编制日期：20230417 | 审核日期： | 批准日期： |
| 管理部门：质量管理组 | | |
| 发放范围： | | |
| □正本 □ 副本 NO.：v2.1.0 | | |

## 1目的

二代宏基因组流程主要适用于DNA样品建库，包括比对-去宿主-物种注释-基因预测-基因功能注释-多样性分析-差异分析。

本SOP用于规范二代宏基因组分析流程的适用规则。

## 2范围

适用于二代宏基因组在北京/义乌 203

## 3 定义

分析异常：在项目分析中，出现了意外的流程中断。

## 4 职责

信息负责人：如期完成标准分析及个性化分析，分析结果达到质控要求。参考《数据分析标准操作规程》、《信息搜集表评估标准操作规程》

## 5 工作程序

### 5.1信息搜集表准备及说明

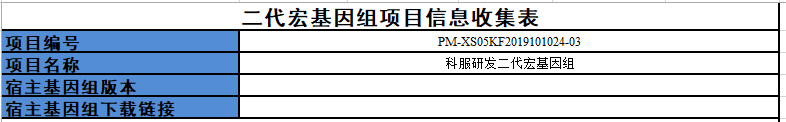
信息搜集表为一个excel文件，分为三个部分，每个部分之间需要一个空行做区分。示例请参考流程目录中的doc/二代宏基因组项目信息收集表.xlsx

#### 5.1.1 信息收集表部分

项目编号：子项目编号

项目名称：任务单名称

宿主基因组下载连接：项管可提供下载连接，分析人员只用将基因组下载，并进行DNA建库（bwa建库步骤即可，其余可不需要）



#### 5.1.2 样品信息表部分

样品名称：对应的下机路径的样品名称，即过滤完成后的样本名；

样品编号：样品对应的文库号；

物种拉丁名：对应样本的物种名，可不填；

样品描述：描述中尽量不要使用特殊符号；

结题报告中样品名称：结题报告中样本对应的名称，结题报告中样品名称列字符长度小于10,以字母开头且只能包含字母、数字、下划线

分组：若需要进行差异分析，则需要提供对应的分组信息，默认使用全部的分组进行分析，若为三个分组，则三个分组同时进行比较。



#### 5.1.3 信搜传到集群

将信息搜集表按照”子项目编号\_info.xlsx”进行命名，并放到 info目录下。该目录下不能有多个信搜。

### 5.2 数据准备

5.2.1 数据准备

数据目录支持两种模式：

1. Analysis/样本/filter/clean/样本\_R1.fq.gz
2. ANNO\*/Cleandata/样本/样本\_R1.fq.gz（只能有一个ANNO\*开头的目录，如果有多个会报错）

如果有以上两种数据格式，优先识别第一种

1. 数据目录下需要有STAT\_Result.xls文件

将目录放到Filter\_Result目录下，并链接到Filter目录下。

5.2.2 信搜准备

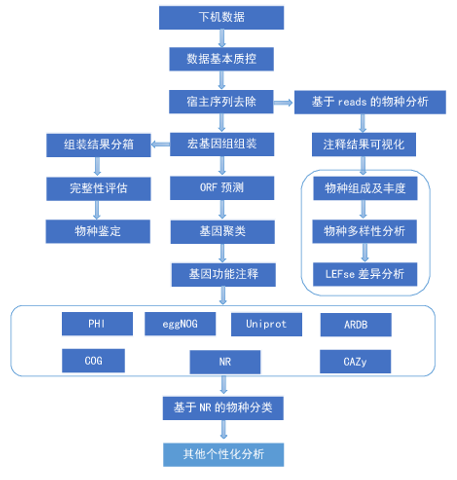
将信搜文件放到分析目录Info目录下，并以\_info.xlsx结尾

5.2.3 参考基因组准备

如果信搜中有参考基因组，则需要提前准备参考基因组。如果该参考基因组已经做过DNA建库，则不需要处理；如果未做过DNA建库，则选择以下两种方案中的一种处理。

1. 联系建库人员，进行参考基因组建库
2. 自己使用Bwa index ref.fa 命令建库

### 5.2 流程分析说明



### 5.3常规流程使用说明

#### 5.3.1 参数说明

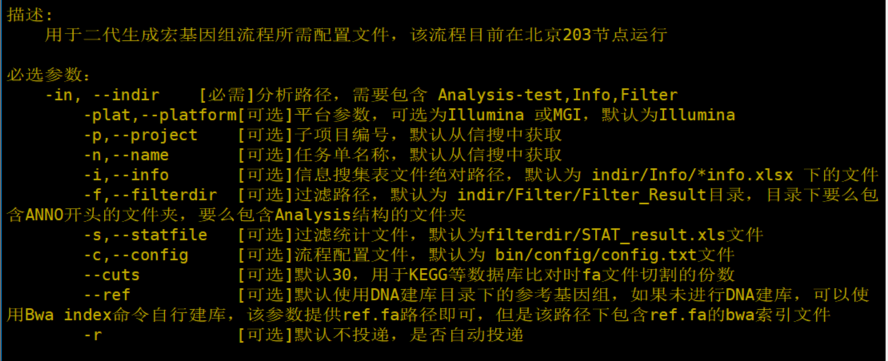
说明：此流程目前在北京203节点，使用chengfangtu账号进行投递。

流程目录为：

$Bin=/annoroad/data1/bioinfo/PROJECT/Commercial/Cooperation/Public/Pipeline/Stable/Mic/NGS\_metagenome/current/

主脚本：$Bin /Pipeline/MetaGenome.py

流程参数说明：



#### -in, --indir [必需]分析路径，需要包含 Analysis-test,Info,Filter

#### -plat,--platform[可选]平台参数，可选为Illumina 或MGI，默认为Illumina

#### -p,--project [可选]子项目编号，默认从信搜中获取

#### -n,--name [可选]任务单名称，默认从信搜中获取

#### -i,--info [可选]信息搜集表文件绝对路径，默认为 indir/Info/\*info.xlsx 下的文件

#### -f,--filterdir [可选]过滤路径，默认为 indir/Filter/Filter\_Result目录，目录下要么包含ANNO开头的文件夹，要么包含Analysis结构的文件夹

#### -s,--statfile [可选]过滤统计文件，默认为filterdir/STAT\_result.xls文件

#### -c,--config [可选]流程配置文件，默认为 bin/config/config.txt文件

#### --cuts [可选]默认30，用于KEGG等数据库比对时fa文件切割的份数

#### --ref [可选]默认使用DNA建库目录下的参考基因组，如果未进行DNA建库，可以使用Bwa index命令自行建库，该参数提供ref.fa路径即可，但是该路径下包含ref.fa的bwa索引文件

#### -r [可选]默认不投递，是否自动投递5.3.2 使用说明

**使用示例1：**

/usr/bin/python3 $Bin /Pipeline/MetaGenome.py –i /annoroad/data1/bioinfo/PROJECT/Commercial/Cooperation/FTP/golden\_data/RD\_All\_Pipeline\_Test\_Example/Microbe/NGS\_Metagenome/PM-XS05KF2022090114-01\_no\_host/ -plat MGI –f /annoroad/data1/bioinfo/PROJECT/Commercial/Cooperation/FTP/golden\_data/RD\_All\_Pipeline\_Test\_Example/Microbe/NGS\_Metagenome/PM-XS05KF2022090114-01\_no\_host/Filter/

Filter\_Result/

`-- ANNO\_PM-XS05KF2022090114-01

|-- Cleandata

`-- STAT\_result.xls

**使用示例2：**

/usr/bin/python3 $Bin /Pipeline/MetaGenome.py –i /annoroad/data1/bioinfo/PROJECT/Commercial/Cooperation/FTP/golden\_data/RD\_All\_Pipeline\_Test\_Example/Microbe/NGS\_Metagenome/PM-XS05KF2022090114-01\_no\_host/ -plat MGI

Filter/

`-- Filter\_Result

|-- ANNO\_PM-XS05KF2022090114-01

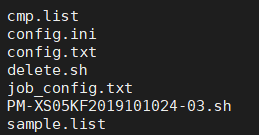
`-- STAT\_result.xls

### 5.4结果说明

在indir/Analysis-test/目录下生成 Analysis和prepare目录，并将项目的配置文件都放到prepare目录下

投递 prepare/项目编号\_qsub.sh即可。

#### 5.4.1 prepare目录



cmp.list：比较组信息；

config.ini：根据信搜表获取的config文件，此文件为样本信息以及流程中用到的项目编号、任务单名称等信息。结题报告中的任务单名称需要修改时，请修改此文件中的Para\_project参数；

config.txt：流程对应的软件以及参数信息；

delete.sh：删除脚本；

job\_config.txt：流程对应投递的job\_config，各模块运行资源配置。流程中如果需要修改投递节点或者内存、线程等资源，请修改此文件中相应模块的参数，并运行主脚本；

PM-XS05KF2019101024-03.sh：流程投递的主脚本，直接sh即可；

sample.list：样本信息。

#### 5.4.2 result目录

|-- annotation -- 预计基因的注释结果，按照数据库创建目录

|-- assemble – 组装结果

|-- Binning – binning结果

|-- Diff – 差异比较组结果

|-- Diversity – 多样性结果

|-- GenePredicion – 基因预测结果

|-- Gene\_Quant – 基因定量结果

|-- Host – 去除宿主后的fq，如果无宿主，则链接clean数据

|-- Kraken – kraken的结果

|-- Kraken\_count – kraken结果合并后的统计结果

|-- ORF – 预测的ORF信息

|-- QC – 进行质控的信息

|-- Report – 结题报告

## 6. 异常问题说明

问题1：5\_0\_diff.sh步骤中断

可能原因：没有差异结果

排查方法:查看.e文件的输出信息，是否没有差异

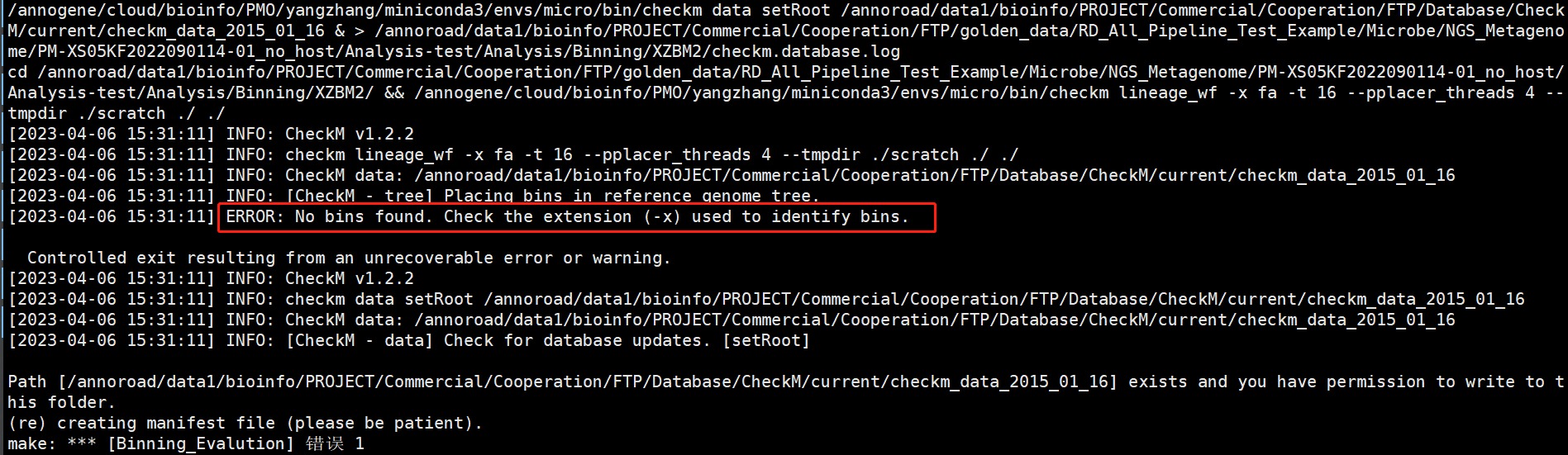
问题2；资源不够

解决方案：如果报错资源不够，可以修改prepare/job\_config.txt文件，然后重新投递 prepare/子项目编号\_qsub.sh

问题3：bin报错

可能原因：没有bin结果

排查方法：查看.e或.o文件查报错信息



## 6 相关文件

### 6.1 软件说明

|  |  |  |  |
| --- | --- | --- | --- |
| **分析内容** | **软件** | **版本** | **参数** |
| 比对分析 | BWA | 0.7.15 | mem算法 |
| 基于Reads注释 | Kraken2 | 2.0.7 | 默认参数 |
| 组装分析 | MEGAHIT | v1.2.9 | 默认参数 |
| 分箱 | MetaBAT2 | 2.12.1 | 默认参数 |
| 分箱结果评估 | checkM | v1.1.3 | "-t 4 -x fa" |
| 基因功能注释 | BLAST | 2.2.28 | "-evalue 1e-5 -num\_alignments 5 -num\_threads 4 -outfmt '6 std stitle'" |
| 基因功能注释 | DIOMAND | v0.9.14.115 | "--outfmt 6 --sensitive --threads 5 --evalue 1e-5" |
| 基因预测 | MetaGeneMark | v.3.38 | "-r -a -d -f G" |
| 非冗余基因 | MMseqs2 | 2-13.45111 | "min-seq-id 0.5 -c 0.8 --cov-mode 1 --threads 8" |
| Alpha/Beta多样性 | QIIME | 2021.4.0 | 默认参数 |
| 差异分析 | LEfSe | 1.1.01 | 默认参数 |

### 6.2 注意事项

### 7 常见问题及解决方法

Q1：支线任务断了咋办？

A：支线任务断了，主线任务会继续运行，而整个流程不受影响。但是查看show\_process的时候，可以看到break的状态。而break的等级是高于run和hold的，因此，需要自行判定程序是否完成。

Q2：主线任务断了咋办？

A：主线断了，那么就断了，需要重新投递。如果主线任务断了，而支线任务没有完成，那么主程序会等候支线任务完成才会实现最终的退出，所以会导致程序一直在运行的假象。这个时候有两个处理办法：

1. 把所有进程都杀掉，然后重新投递

2. 只杀掉主进程（pipeline.py)，并且在log.txt中人为添加支线任务的finish标识（防止再次运行pipeline.py时重新投递），之后重新运行该任务。

Q3：如何监控项目运行状态？

A：运行/annoroad/bioinfo/PMO/liutao/pipeline\_generate/bin/v5/show\_process.py会显示项目的状态。项目运行状态分为running, break, plan, end 四种。其中running表示正在运行，break表示中断，plan表示准备运行，end表示运行完成,hold表示磁盘不够，任务挂起。

Q4：发现任务状态是break，该咋办？

A：当发现任务状态是break的时候，首先需要确定break掉的任务是否是主线任务。

如果是主线任务，如果主程序自然退出(在top或者ps的时候没有发现pipeline.py），则可以重新投递任务；

如果没有自然退出（主程序pipeline.py还在运行），那么可能是有之前的支线任务未完成，可以参照Q2来进行操作；

如果是支线任务，如果查看log.txt发现主线任务全部完成或者正常运行，如果时间允许，可以等待所有任务完成后再投递任务；

如果加急，可以把该步对应的sh文件修改后，手动投递该任务；

如果主线任务也断掉了，那么修改脚本后，重新投递所有任务。

Q5：监控项目的记录文件在哪？

A：程序会在您的home目录下，生成一个记录文件，路径为~/.mission/.pipeline.log，记录了每个项目的分析目录。如果不想显示某个项目，可以对相应的行进行删除或者编辑。

Q6：如果程序断了，咋办？

A：如果程序由于各种因素中断了，仔细检查脚本，如果脚本没错，确定只是中断，那么重新运行一次之前的脚本，默认会把断掉的模块全部重头运行；如果不想将该模块内已经完成的样品重新运行，可以加上-c参数，那么会只运行没有运行成功的样品。

Q7：如何发现配额不够？

A：如果配额不够了，会把所有的任务挂起，使用show\_process查看时，会发现Hold状态；或者使用qstat的时候，会发现hqw，hr，ht等，或者没有任务在运行（因为配额不够，程序会自动不投递任务）

Q8：配额不够了，咋办？

A：第一，找系统组修改配额 ； 第二，修改相应的sh.\*.log文件，加入一行DISK\_QUOTA\*\*G ;之后，程序会自动的release。但需要注意的是，因为之前在程序里设置了较小的配额，所以之后每一步都会被hold。所以需要之后每个log文件都加上 DISK\_QUOTA \*\*G，来每次进行更改；或者杀掉重新来。 或者删除文件来释放空间，这样的话，后面可以不用修改就可以运行。

Q9： 如何杀掉程序？

A:

1. 杀掉所有的子进程 守护进程qsub\_sge.pl，否则的杀掉的任务会重新投递；

2. 杀掉所有的任务 qdel掉

Q10：如何精准的杀掉守护进程？

A：

1. 在重新投递之前，使用ps -f -u name |cat 然后仔细的判别，获得进程ID

2. 查看shell后面的数字，在sh 和log直接的数字，是进程ID，使用kill -9可以杀掉

## 8更新记录

|  |  |  |
| --- | --- | --- |
| **版本号** | **修订内容** | **修订日期** |
| v0.0.1 | 搭建完成 | **20190326** |
| V2.1.0 | 修改比对软件为pbmm2 | **20211021** |
|  |  |  |