宏基因组上游任务运行须知

2024年12月30日

> 本文档为宏基因组上游任务 **运行前**，需要项目负责人准备的流程与注意事项。包括任务数据文件准备、任务数据信息表准备、任务运行配置等内容。
>
> - 本文档仅适用于标准宏基因组上游分析任务，**基于Reads的物种注释与功能注释**。有特殊需要的非标准任务请单独联系。
> - 本文档中的路径、文件名等信息仅为示例，实际操作时请根据实际情况进行修改。
> - 请严格按照本文档的要求准备数据文件、信息表、配置文件等内容，以保证任务能够尽快开始运行。

1. 任务数据文件准备

1.1 请将原始数据放置在1块（最好）或多块移动硬盘中，移动硬盘的文件系统请使用 `NTFS`。

> 文件系统类型在Windows系统中可以通过右键点击硬盘，选择"属性"查看。格式化也可以在Windows系统中右键点击硬盘，选择"格式化"进行。苹果系统盘（APFS）部分服务器可能无法识别，建议使用Windows系统格式化。exFAT格式因为没有日志系统，部分情况下可能会出现数据丢失的情况，不建议使用。

1.2 请将原始数据文件按照样本进行分类，每个样本一个文件夹，文件夹名称为样本名。

> 测序文件默认为双端测序文件，reads length为 150bp。reads length等会影响软件运行参数，如果您的数据不符合这个要求，请联系我们进行调整。

1.3 请将每个样本文件夹中的原始双端测序数据文件（压缩的fastq格式）命名为 `样本名_1.clean.fq.gz` 和 `样本名_2.clean.fq.gz`。

> 我们一般使用的是测序公司返回的做过一次质控的数据，因此文件名中包含了 `clean` 字样。如果您的数据没有经过质控，请将 `clean` 替换为 `raw`。

1.4 确保每个样本文件夹中有一个名为 MD5.txt 的文件，文件中包含了该样本文件夹中所有测序文件的MD5值。

> MD5文件的内容格式如下：

```
cat MD5.txt
266b3528cfd65e6a5808649cf90d685a  样本名_1.clean.fq.gz
da273b2c3e84966d55c48d1a47ce7d6d  样本名_2.clean.fq.gz
```

> 原始MD5值一般由测序公司在数据交付时提供，用于验证数据完整性。如果您没有MD5值，在确保文件没有损坏的情况下，在windows系统下可以使用 `certutil -hashfile 文件名 MD5` 命令生成MD5值。在Linux系统下可以使用 `md5sum 文件名` 命令生成MD5值。**文件损坏会导致不可控结果，因此请务必保证数据完整性。**

1.5 完整的数据文件目录结构示例：

```
$tree /data/projectA/00_cleandata/
.
├── SAM001
│   ├── SAM001_1.clean.fq.gz
│   ├── SAM001_2.clean.fq.gz
│   └── MD5.txt
├── CK002
│   ├── CK002_1.clean.fq.gz
│   ├── CK002_2.clean.fq.gz
│   └── MD5.txt
├── 张三
│   ├── 张三_1.clean.fq.gz
│   ├── 张三_2.clean.fq.gz
│   └── MD5.txt
```

2. 任务数据信息表准备

2.1 请准备一个名为 `sampleList.csv` 的样本信息表，表中包含了所有样本的信息。

> 样本信息表的内容格式如下：
>
> - `Server`：运行服务器名称，此列不需要项目人填写
> - `Sample_ID`：样本名，与数据文件夹名称一致，**必填**
> - `Use_Sample`：同 `Sample_ID`，校验用，无特殊情况直接复制 `Sample_ID` 列即可，不可删除
> - `Group_1`：样本所属分组，更多分组可以自行添加列 `Group_2`、`Group_3`等，非必填，无特殊情况直接全填 `y` 即可
> - `Species`：样本物种信息，拉丁名，**必填**
> - `Ref`：去宿主时所用参考基因组信息，**必填**，大部分情况下与 `Species` 列一致，在所选物种没有对应参考基因组时使用近源物种的参考基因组即可；此列为非去除宿主、环境样本、已去除宿主的样本等情况请填写 `none`
> - `Download_link`：`Ref`列使用的参考基因组下载链接，**必填**，请确保链接有效，无链接需自行提供参考基因组文件

```
# sampleList.csv 示例
cat sampleList.csv

Server,Sample_ID,Use_Sample,Group_1,Species,Ref,Download_link
huatuo,S0001C,S0001C,F,Uropsilus gracilis,Uropsilus gracilis,https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/004/024/945/GCA_004
huatuo,S0002C,S0002C,F,Ochotona thibetana,Ochotona princeps,https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/030/435/755/GCF_8304
huatuo,S0003C,S0003C,F,Ochotona thibetana,Ochotona princeps,https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/030/435/755/GCF_8304
huatuo,S0004C,S0004C,F,Ochotona thibetana,Ochotona princeps,https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/030/435/755/GCF_8304
huatuo,SAM001,SAM001,SAM,Homo sapiens,Homo sapiens,https://ftp.ensembl.org/pub/release-113/fasta/homo_sapiens/dna/Homo_sapien
huatuo,CK0002,CK0002,CK,Homo sapiens,Homo sapiens,https://ftp.ensembl.org/pub/release-113/fasta/homo_sapiens/dna/Homo_sapiens
huatuo,张三,张三,ZK,Homo sapiens,Homo sapiens,https://ftp.ensembl.org/pub/release-113/fasta/homo_sapiens/dna/Homo_sapiens.GRCh
huatuo,Water1,Water1,F,none,none,https://ftp.ncbi.nlm.nih.gov/genomes/none.fna.gz
```

2.2 可准备一个名为 `renameList.txt` 的样本名修改对应表，第一列为原样本名，第二列为新样本名(与 `sampleList.csv` 文件总`Sample_ID` 列对应)，无需表头。如不需要改动样本名，第一列与第二列保持一致即可。

> - 部分情况下，样本的原始ID来自不同的采样人员，各自的命名风格也不同，不统一和异常的样本命名方式会为下游的分析带来不必要的麻烦，如：

异常情况	异常示例	可能导致问题
以数字开头	`20240101SDR1`	数字开头会在R分析中被自动加上引号
纯数字组成	`00001`	被识别为数字 `00001` 会自动转化为 `1`
带有特殊符号	`cd-32-41-9`	分符号不识别，`-` 等符号可能会自动转换为 `_`
带有中文	张三	中文字符可能会导致乱码
样本名过长，附带有测序的文库号	`S0001C_FRAS192003827-1a`	部分软件过长ID会被截断导致后续ID不匹配

> - 任务批量处理时按照正则匹配识别样本文件名，因此非常建议将样本名统一为：**固定长度不超过10位的以大写字母开头的格式**
> - 因此我们提供了 `renameList.txt` 以便于在数据处理前将样本ID统一修改，在满足上述1.2-1.5的目录结构中，可以将原有的《文件夹名称、文件名称、MD5文件中文件名称》中的样本ID修改为新的统一的样本ID。当然您也可以手动修改，保证修改后符合上述要求即可。

```
# renameList.txt 示例
cat renameList.txt

S0001C_FRAS192003827-1a S0001C
S0002C  S0002C
S0003C  S0003C
S0004C  S0004C
SAM001  SAM001
CK002   CK0002
张三    ZSCD01
Water1  Water1
```

```
# 修改前示例

├── 张三
│   ├── 张三_1.clean.fq.gz
│   ├── 张三_2.clean.fq.gz
│   └── MD5.txt

cat MD5.txt
266b3528cfd65e6a5808649cf90d685a  张三_1.clean.fq.gz
da273b2c3e84966d55c48d1a47ce7d6d  张三_2.clean.fq.gz

├── ZSCD01
│   ├── ZSCD01_1.clean.fq.gz
│   ├── ZSCD01_2.clean.fq.gz
│   └── MD5.txt

cat MD5.txt
266b3528cfd65e6a5808649cf90d685a  ZSCD01_1.clean.fq.gz
da273b2c3e84966d55c48d1a47ce7d6d  ZSCD01_2.clean.fq.gz
```

3. 任务运行配置准备

> 运行配置文件为 `config.yaml`，请根据实际情况修改配置文件中的参数。

标准宏基因组上游分析的任务使用默认参数不需要修改，如果有特殊需求请联系我们进行调整。

> 如果对于比对数据库的版本有特殊需求，请单独联系我们进行调整。

```
db:
kraken2: kraken2_pluspf_20231009_db
humann_chocophlan: humann3_20231017_db/chocophlan-v31
humann_uniref: humann3_20231017_db/uniref
humann_utility_mapping: humann3_20231017_db/utility_mapping
metaphlan: metaphlan_20231214_db
vfdb: vf_20230915_db
rgi: card_20231214_db
dbcan: dbcan_20231214_db
eggnog: eggnog_20240116_db
checkm: checkm_20150116_db
gtdbtk: gtdbtk_20231214_db
taxonkit: taxonkit_20240108_db
mag_db: download_MAGs_20240322_db
```