转录组自动分析流程
设计使用说明书

冯华东　　Email：ash123fhd@163.com

目录

1. 简介

随着海量转录组测序数据的出现，一个高效、准确、可重复的分析流程对于转录组测序数据的分析变得愈发重要。且随着近年来各种软件的开发优化，软件数量不断增多，导致服务器在软件配置和管理方面出现许多问题。如：各不同分析流程软件互不兼容、服务器环境复杂造成新软件无法安装等，会使研究人员大量时间浪费在管理和配置软件上。因此，本自动化分析转录组测序数据的流程在这样的背景下应运而生。此分析流程适用于所有物种的转录组测序数据的自动分析。流程主要包括三套子流程：一套针对没有参考基因组的物种进行转录组测序数据的分析，两套根据市面上各软件的不同搭配针对有参考基因组的物种进行转录组测序数据的分析。本流程包括基本的处理步骤，如对测序数据的质量控制和预处理、数据比对、转录本的定量和基因的差异分析。此外，针对没有参考基因组的物种，还对分析后的转录本进行功能注释并生成注释报告。此分析流程基于 snakemake 工作流程管理系统和 Docker 容器引擎，安装配置好所有依赖项后即可自动运行。

1.1. 编写目的

本文档通过使用非专门术语的语言，充分地描述该分析系统所具有的功能及基本的使用方法，提供该流程每一个运行的具体过程和有关知识，包括操作方法的细节。使用户通过本手册能够了解该流程的详细用途，并且能够确定在什么情况下，如何使用它。

1.2. 使用对象

本文档的使用对象主要为需要分析转录组测序数据的科研人员。

1.3. 安装

1.3.1. Docker

由于该分析流程是基于 Docker 容器引擎的，因此，需要先在 Linux 的操作系统上安装 Docker 容器引擎。安装步骤如下（以 centos7 为例）：

```
yum install -y yum-utils
yum-config-manager --add-repo https://download.docker.com/linux/centos/docker-ce.repo
yum install docker-ce docker-ce-cli containerd.io docker-buildx-plugin docker-compose-plugin
systemctl start docker
```

通过运行镜像验证 Docker 引擎是否安装成功；

```
docker run hello-world
```

（详细安装步骤可参考 Docker 官方文档：Install Docker Engine on CentOS | Docker Docs）

1.3.2. 工作流程安装

此流程封装好的 Docker 镜像可以在 docker hub 页面找到，运行该流程需要先下载镜像启动容器再使用：

```
docker pull ash123fhd/star-snake:1.0
docker run -itd --name RNA_snakemake ash123fhd/star-snake:1.0 bash
docker exec -it RNA_snakemake bash
```

注意：RNA_snakemake 为容器名称，可以根据用户所需更改。ash123fhd/star-snake:1.0 为镜像名，根据用户不同需求可以下载不同的镜像。

1.4. 输入文件

运行工作流程需要多个输入文件：一个基因组文件（.fa）、一个注释文件 （.gtf） 、压缩测序文件 （.fastq.gz）、工作流配置文件（.yaml）和样本信息表（.csv/.txt）

文件类型	描述
genome.fa	用户提供的基因组序列文件
annotation.gtf	用户提供的基因组注释文件
sequence.fastq.gz	用户提供的压缩测序文件
config.yaml	用于自定义工作流的配置文件
info.csv	样本信息表（有参考基因组）
samples.txt	样本信息表（无参考基因组）

1.4.1. 基因组文件及其注释文件

建议用户从美国国家生物技术信息中心（NCBI）或者专门的物种基因组数据库检索下载所需物种的基因组和注释文件。

注意：如果您使用自定义注释文件，请确保注释文件遵守 gtf 文件格式标准，如果注释文件为 gff 格式文件请转为 gtf 格式文件。

1.4.2. 输入测序文件

这些都是由用户提供的输入文件，仅支持双端测序数据。请确保将.fastq.gz 输入文件全部存放于/home/samples/中。

1.4.3. 样本信息表及配置文件（有参考基因组）

要运行此工作流程，必须提供一个样本信息文件和一个配置文件。这两个文件的模板都可以在/home/configs 中提供。样本信息文件用于指定输入的.fastq.gz 文件的信息，便于后续进行差异基因分析。配置文件用于用户轻松的自定义某些设置。用户可以根据自己的实验数据修改配置目录中的 info.csv 和 config.yaml 文件。

修改样本信息表（/home/configs/info.csv）：

id	condition
id1	CONTROL
id2	CONTROL
id3	CONTROL
id4	TREAT
id5	TREAT
id6	TREAT

id 为测序输入文件 id，condition 为数据处理条件。

自定义配置文件，根据实验条件和数据情况修改 config.yaml 文件。它包含以下变量：

PROJECT: 项目名称

SAMPLES:[测序文件名称]

GENOME:基因组文件所在路径

ANNOTATION:基因组注释文件所在路径

INPUTPATH:输入文件目录所在路径

OUTPUTPATH:输出文件目录所在路径

THREAD: 运行软件时线程数目

INFO:样本信息文件

CONTROL: [对照组条件]

TREAT: [处理组条件]

SPECIES:物种名称

1.4.4. 样本信息文件及配置文件（无参考基因组）

修改样本信息表（/home/configs/samples.txt）：

treatment	id1	OUTPUTPATH/fastp/id1-1.fastq.gz	OUTPUTPATH/fastp/id1-2.fastq.gz
treatment	id2	OUTPUTPATH/fastp/id2-1.fastq.gz	OUTPUTPATH/fastp/id2-2.fastq.gz
treatment	id3	OUTPUTPATH/fastp/id3-1.fastq.gz	OUTPUTPATH/fastp/id3-2.fastq.gz
treatment	id4	OUTPUTPATH/fastp/id4-1.fastq.gz	OUTPUTPATH/fastp/id4-2.fastq.gz

第一列为处理条件，第二列为测序输入文件 id，第三第四列为质控后测序文件的存放位置。

自定义配置文件，根据实验条件和数据情况修改 config.yaml 文件。它包含以下变量：

PROJECT: 项目名称

SAMPLES: [测序文件名称]

INPUTPATH: 输入文件目录所在路径

OUTPUTPATH: 输出文件目录所在路径

THREAD: 运行软件时线程数目

INFO: 样本信息表存放路径

TREATMENT:[测序文件处理方式]

SPECIES: 物种名称

1.5. 输出文件

有参考基因组的流程

输出文件路径	输出文件
OUTPUTPATH/QC	测序数据的质量报告
OUTPUTPATH/fastp	进行质控和过滤后的测序数据及其质量报告

OUTPUTPATH/mapping	读段比对到参考基因组产生的 sam/bam 及对比结果报告等文件
OUTPUTPATH/quantity	转录本组装定量产生的各种文件及最终的表达量及表达矩阵（fpkm、tpm、count）
OUTPUTPATH/DEA	差异分析结果和可视化结果

无参考基因组流程

输出文件路径	输出文件
OUTPUTPATH/QC	测序数据的质量报告
OUTPUTPATH/fastp	进行质控和过滤后的测序数据及其质量报告
OUTPUTPATH/result	分析后的结果存放地址(表达量、差异分析、拼接的转录本、注释结果)
OUTPUTPATH/result/DESeq2	差异分析结果和可视化结果
OUTPUTPATH/result/trinotate_annotation_report.xls	注释结果报告

2. 工作流程配置

此分析流程所包含的三套子流程的 Docker 镜像分别为：ash123fhd/hisat-snake：1.0（有参）、ash123fhd/star-snake：1.0（有参）、ash123fhd/denovo-snake：2.0（无参）。三套子流程镜像均上传至 docker hub 上。所有子流程都允许参数的定制，用户可以设置的自定义工作流的选项显示在配置文件中（.yaml）。

2.1. 各子流程包含的分析软件

有参考基因组分析流程 1（ash123fhd/hisat-snake：1.0）：

软件（工具）	版本	描述
python	3.8.10	python 解释器
snakemake	7.32.2	工作流程管理系统
sratoolkit	3.0.0	ncbi 下载及解压数据的工具
gffread	2.2.1	gff 和 gtf 文件转换工具
fastqc	0.11.9	输出测序数据的质量报告
multiqc	1.15	汇总报告的工具
fastp	0.23.2	去除低质量数据
hisat2	2.2.1	将读段比对到参考基因组
samtools	1.16	处理 bam 文件的工具
stringtie	2.2.1	转录本组装定量工具
R	4.2.1	R 解释器

r-dplyr	1.1.2	处理数据的 R 包
r-yaml	2.3.7	载入 yaml 文件
r-BiocManager	3.16	安装和管理 Bioconductor 软件包
r-DESeq2	1.38.3	差异表达分析
r- ggplot2	3.4.2	用于创建图形的 R 软件包
r-pheatmap	1.0.12	用于绘制热图的 R 软件包

有参考基因组分析流程 2（ash123fhd/star-snake：1.0）：

软件（工具）	版本	描述
python	3.8.10	python 解释器
snakemake	7.32.4	工作流程管理系统
sratoolkit	3.0.0	ncbi 下载及解压数据的工具
gffread	2.2.1	gff 和 gtf 文件转换工具
fastqc	0.11.9	输出测序数据的质量报告
multiqc	1.15	汇总报告的工具
fastp	0.23.2	去除低质量数据
STAR	2.7.10b	将读段比对到参考基因组
samtools	1.16	处理 bam 文件的工具
htseq	2.0.4	基因定量
R	4.2.1	R 解释器
r-dplyr	1.1.3	处理数据的 R 包
r-yaml	2.3.7	载入 yaml 文件
r-BiocManager	3.16	安装和管理 Bioconductor 软件包
r-ggplot2	3.4.3	用于创建图形的 R 软件包
r-pheatmap	1.0.12	用于绘制热图的 R 软件包
r-edgeR	3.40.2	差异表达分析
GenomicFeatures	1.50.4	提取基因组序列

无参考基因组分析流程（ash123fhd/denovo-snake：2.0）

软件（工具）	版本	描述
python	3.8.10	python 解释器
snakemake	7.32.4	工作流程管理系统
sratoolkit	3.0.0	ncbi 下载及解压数据的工具
fastqc	0.11.9	输出测序数据的质量报告
multiqc	1.16	汇总报告的工具
fastp	0.23.2	去除低质量数据
bowtie2	2.4.4	比对软件
Trinity	2.15.1	转录组组装
samtools	1.16	处理 bam 文件的工具
RSEM	1.3.3	转录本定量
Cd-hit	4.8.1	去冗余
diamond	2.1.6	序列比对软件
hmmer	3.3.2	通过 HMMER 工具注释蛋白质结构域

TransDecoder	5.7.0	预测转录本的蛋白编码区域
sqlite	3.41.2	整合数据库数据
blast	2.14.0	序列对比软件
Trinotate	4.0.2	对转录本进行注释
R	4.2.1	R 解释器
r-dplyr	1.1.3	处理数据的 R 包
r-yaml	2.3.7	载入 yaml 文件
r-edgeR	3.40.2	差异表达分析
r-DESeq2	1.38.3	差异表达分析
r-qvalue	2.30.0	处理数据的 R 软件包
r-fastcluster	1.2.3	处理数据的 R 软件包
r-seqLogo	1.64.0	绘制图形的 R 软件包

3. 使用说明

为了方便用户使用该流程，下面将做一个演示运行工作流程。

3.1. 启动容器并进入容器

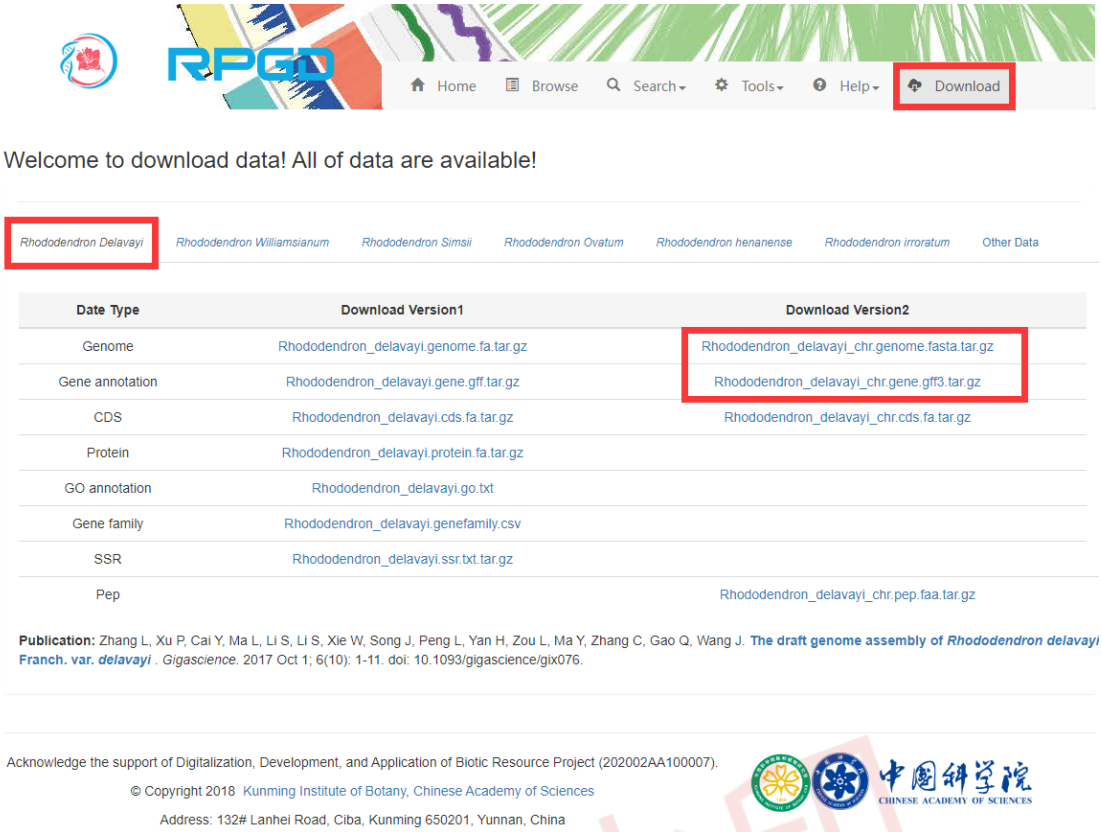
```
docker run -itd --name RNA_snakemake ash123fhd/star-snake:1.0 bash
docker exec -it RNA_snakemake bash
```

注意：用户可以分析数据需要选择有参考基因组还是无参考基因组的分析流程。

3.2. 准备输入文件

3.2.1. 基因组文件及其注释文件（仅限有参考基因组分析流程）

建议用户从美国国家生物技术信息中心（NCBI）或者专门的物种基因组数据库检索下载所需物种的基因组和注释文件。以在杜鹃花基因组数据库（RPGD）下载马缨杜鹃（*R. delavayi*）基因组文件和注释文件进行演示为例：



RPGD: http://bioinfor.kib.ac.cn/RPGD/index.html

使用鼠标右键单击上图中基因组和基因组注释文件获取下载链接下载到已启动的 Docker 容器中并解压缩：

```
cd reference/                    #进入存放参考基因组文件目录
wget http://bioinfor.kib.ac.cn/RPGD/download/Rhododendron_delavayi_chr.genome.fasta.tar.gz
wget http://bioinfor.kib.ac.cn/RPGD/download/Rhododendron_delavayi_chr.gene.gff3.tar.gz
tar -zxvf Rhododendron_delavayi_chr.gene.gff3.tar.gz
tar -zxvf Rhododendron_delavayi_chr.genome.fasta.tar.gz
```

注意：如果您使用自定义注释文件，请确保遵守 gtf 文件格式标准，如果注释文件为 gff 格式文件请转为 gtf 格式文件。

```
gffread -T Rhododendron_delavayi.gene.gff3 -o Rhododendron_delavayi.gtf
```

3.2.2. 测序文件

下载或者将准备好的转录组测序数据放入/home/samples 目录中，示例数据从美国国家生物技术信息中心（NCBI）下载，链接：https://www.ncbi.nlm.nih.gov/bioproject/PRJNA476831

```
root@2dbb47b3d4d1:/home/samples# ls -l
total 15491436
-rw-r--r--. 1 1001 1001 1333462765 Jun 30 07:51 SRR7403459_1.fastq.gz
-rw-r--r--. 1 1001 1001 1328487561 Jun 30 07:51 SRR7403459_2.fastq.gz
-rw-r--r--. 1 1001 1001 1368261934 Jun 30 07:52 SRR7403460_1.fastq.gz
-rw-r--r--. 1 1001 1001 1348426129 Jun 30 07:52 SRR7403460_2.fastq.gz
-rw-r--r--. 1 1001 1001 1297596334 Jun 30 11:37 SRR7403463_1.fastq.gz
-rw-r--r--. 1 1001 1001 1308172681 Jun 30 11:37 SRR7403463_2.fastq.gz
-rw-r--r--. 1 1001 1001 1297332331 Jun 30 07:51 SRR7403464_1.fastq.gz
-rw-r--r--. 1 1001 1001 1311665745 Jun 30 07:51 SRR7403464_2.fastq.gz
-rw-r--r--. 1 1001 1001 1300946403 Jun 30 11:37 SRR7403465_1.fastq.gz
-rw-r--r--. 1 1001 1001 1301318765 Jun 30 11:37 SRR7403465_2.fastq.gz
-rw-r--r--. 1 1001 1001 1328814518 Jun 30 11:38 SRR7403466_1.fastq.gz
-rw-r--r--. 1 1001 1001 1338724361 Jun 30 11:38 SRR7403466_2.fastq.gz
root@2dbb47b3d4d1:/home/samples#
```

3.2.3. 样本信息文件及配置文件

最后，将准备样本信息表（info.csv/samples.txt）和配置文件（config.yaml）。用户可以在配置目录（/home/configs）中修改它们。示例如下：

有参考基因组流程：

样本信息表（/home/configs/info.csv）：

id	condition
SRR7403463	eco
SRR7403465	eco
SRR7403466	eco
SRR7403459	endo
SRR7403460	endo
SRR7403464	endo

配置文件（/home/configs/config.yaml）：

PROJECT: example

SAMPLES:["SRR7403463","SRR7403465","SRR7403466","SRR7403459","SRR7403460","SRR7403464"]

GENOME: /home/reference/genome.fasta

ANNOTATION: /home/reference/ Rhododendron_delavayi.gtf

INPUTPATH: /home/samples

OUTPUTPATH: /home/Rhododendron_1

THREAD: "10"

INFO: /home/configs/info.csv

CONTROL: ["eco"]

TREAT: ["endo"]

SPECIES: Rhododendron delavayi

无参考基因组流程：

样本信息表（/home/configs/samples.txt）：

endo	SRR7403459	/home/Rhododendron_01/fastp/SRR7403459-1.fastq.gz	/home/Rhododendron_01/fastp/SRR7403459-2.fastq.gz
endo	SRR7403460	/home/Rhododendron_01/fastp/SRR7403460-1.fastq.gz	/home/Rhododendron_01/fastp/SRR7403460-2.fastq.gz
eco	SRR7403463	/home/Rhododendron_01/fastp/SRR7403463-1.fastq.gz	/home/Rhododendron_01/fastp/SRR7403463-2.fastq.gz
endo	SRR7403464	/home/Rhododendron_01/fastp/SRR7403464-1.fastq.gz	/home/Rhododendron_01/fastp/SRR7403464-2.fastq.gz
eco	SRR7403465	/home/Rhododendron_01/fastp/SRR7403465-1.fastq.gz	/home/Rhododendron_01/fastp/SRR7403465-2.fastq.gz
eco	SRR7403466	/home/Rhododendron_01/fastp/SRR7403466-1.fastq.gz	/home/Rhododendron_01/fastp/SRR7403466-2.fastq.gz

配置文件（/home/configs/config.yaml）：

PROJECT: example

SAMPLES:["SRR7403459","SRR7403460","SRR7403463","SRR7403464","SRR7403465","SRR7403466"]

INPUTPATH: /home/samples

OUTPUTPATH: /home/Rhododendron_01

THREAD: "40"

INFO: /home/configs/samples.txt

TREATMENT: ["endo","eco"]

SPECIES: Rhododendron delavayi

3.3. 运行工作流程

有参考基因组的分析流程：

```
cd /home/run
snakemake -s run.py -j 6                # -j 参数表示同时运行任务数，用户可根据情况自行修改
#后台运行
nohup snakemake -s run.py -j 6 &
```

注意：运行 run.py 可以直接由测序数据得到基因表达量（FPKM、TPM、count 矩阵）、差异基因分析的结果；运行 exp.py 可以直接从测序数据得到基因表达量而不会进行差异分析；运行 deg.py 前则需要先运行 run.py 或者 exp.py，运行 deg.py 将会根据 config.yaml 配置文件所设置的处理组和对照组对 run.py 或 exp.py 得到的 count 矩阵进行差异分析。

无参考基因组的分析流程：

```
cd /home/run
snakemake -s run.py -j 6                # -j 参数表示同时运行任务数，用户可根据情况自行修改
#后台运行
nohup snakemake -s run.py -j 6 &
```

注意：运行 run.py 可以直接由测序数据得到基因表达量（TPM、count 矩阵）、差异基因分析的结果、预测转录本 ORF 的结果以及注释报告。

3.4. 输出结果

输出结果将存储在自定义配置文件的 OUTPUTPATH 中。