**RSNAP使用手册**

目 录

[一、 软件简介 1](#_Toc137138642)

[二、 软件运行环境 1](#_Toc137138643)

[（一） 操作系统 1](#_Toc137138644)

[（二） 所需依赖 1](#_Toc137138645)

[（三） BLAST 3](#_Toc137138646)

[三、 使用说明 4](#_Toc137138647)

[（一） 基本用法 4](#_Toc137138648)

[（二） 参数配置文件详解 4](#_Toc137138649)

[（三） 使用案例 7](#_Toc137138650)

[（四） 结果解释 11](#_Toc137138651)

## 软件简介

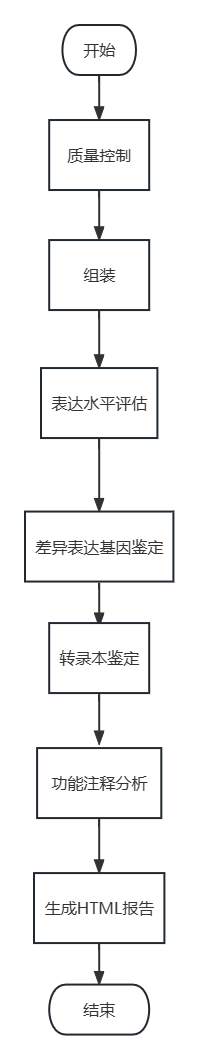


图 1 RSNAP总体框架流程图

RNA-seq无参分析是指在没有参考基因组序列或注释信息的情况下，对RNA-seq数据进行转录本的组装、定量、鉴定和注释等分析。这种分析对于那些基因组信息缺乏或不完善的物种，如许多植物、动物和微生物，具有重要的意义，可以发现新的基因或转录本，揭示其生物学功能和进化关系。然而，RNA-seq无参分析涉及多个步骤和软件，需要考虑数据质量、测序深度、组装算法、表达量估计、差异表达分析、功能注释分析等多方面的因素。每个步骤都可能影响最终的分析结果和解释，因此需要选择合适的方法和参数，以及进行有效的评估和验证。目前已有一些工具或流程可以实现部分或全部的RNA-seq无参分析，如Trinity、ExpressAnalyst等，但它们仍然存在一些不足或局限性，如运行速度慢、内存占用大、功能不完善或不灵活等。为了解决这些问题，我们开发了这个工具，它基于Python语言编写，使用多个开源软件和数据库，可以根据用户输入的参数配置文件，完成从原始测序数据到功能注释结果的一系列分析，并生成一个html格式的报告。该工具具有易用性、灵活性和可扩展性等特点，可以为RNA-seq无参分析提供一个方便快捷的解决方案，有助于那些缺乏完整基因组注释信息的物种的转录组学研究。

## 软件运行环境

### 操作系统

支持主流Unix/Linux发行版本。

### 所需依赖

1. **Python（版本大于3.7）**

os：使用操作系统相关功能的便携方式，比如文件和目录操作、进程管理和环境变量。

sys：提供了一些与Python解释器和它的环境相关的变量和函数，比如标准输入输出、命令行参数和退出状态。

subprocess：创建和管理子进程的方式，可以灵活地控制它们的输入输出和错误流。

zipfile：读写ZIP文件的工具，支持压缩和解压缩。

logging：记录日志信息的功能，可以灵活地配置日志级别、格式、输出目标等。

glob：根据通配符匹配文件名的方法，比如\*.txt或a?.py。

#gseapy：进行基因集富集分析（GSEA）和富集图（Enrichment Map）的方法，可以从多个数据库中获取基因集。

#jinja2：创建动态HTML模板的方法，支持变量、控制结构、过滤器等。

#click：创建命令行界面（CLI）的方法，支持参数、选项、子命令等。

#rpy2：在Python中调用R语言的方法，支持R对象、函数、包等。

#numpy：处理多维数组和矩阵的方法，支持高效的数学运算和随机数生成等。（版本大于等于1.9.0）

#scipy：进行科学计算的方法，支持线性代数、优化、统计、信号处理等。

#matplotlib：进行数据可视化的方法，支持绘制折线图、柱状图、散点图等。（版本大于等于1.4.3）

#pandas：处理结构化数据的方法，支持数据读取、清洗、分析、聚合等。（版本大于等于0.16）

#requests：发送HTTP请求的方法，支持GET、POST、PUT等方法和各种参数。

注：上述python包中，不带#号的为python内置模块，带#号的需要安装。

1. **R（版本大于3.7）**

还需安装如下R包：

DESeq2：一种进行差异表达分析的方法，可以处理RNA测序数据，使用负二项分布模型和Wald检验。

limma：一种进行线性模型分析的方法，可以处理基因芯片和RNA测序数据，使用经验贝叶斯方法和t检验。

edgeR：一种进行差异表达分析的方法，可以处理RNA测序数据，使用负二项分布模型和精确检验。（版本大于等于3.34.0）

pheatmap：一种绘制热图的方法，可以显示数据的聚类和相关性。

data.table：一种处理大型数据表的方法，支持快速的子集、排序、分组和连接操作。

1. **依赖程序**

还需安装如下程序：

Pigz：一个并行实现的gzip，可以利用多个处理器和多核心来压缩数据12。

blast：一个用于比较生物序列的工具，可以找到相似或同源的序列。

fastqc：一个用于检查原始测序数据质量的工具，可以生成各种质量指标的图表和报告。

trimmomatic：一个用于去除测序数据中的低质量或接头序列的工具，可以提高后续分析的准确性。

gzip：一个用于压缩和解压缩文件的工具，可以节省存储空间和传输时间。

Trinity：一个用于从RNA测序数据中组装转录本的工具，可以重建基因结构和表达量。

Bowtie：一个用于将短读比对到参考基因组的工具，可以快速且准确地找到比对位置。

Bowtie2：一个改进版的Bowtie，可以比对更长或更不规则的短读，也可以处理配对端测序数据。

RSEM：一个用于估计基因和转录本表达量的工具，可以从RNA测序数据中计算相对丰度。

jellyfish：一个用于统计k-mer频率的工具，可以从DNA或RNA测序数据中生成k-mer分布。

注：jellyfish应该为Trinity程序自带，但在测试中发现某些版本中执行Trinity时会报错不存在jellyfish。

## 使用说明

### 基本用法

python RSNAP.py (--config config.py)

--config 参数指定参数配置文件地址（默认为./config.py）

### 参数配置文件详解

参数配置文件为.py格式文件，所以遵循python代码格式规范。为字典和列表的嵌套结构。

* 用户输入参数（user\_args）

1. samples

# 测序样本文件绝对路径（双末端）

示例为：

"samples": { # 左 #右

"S1": ["./test/Con-1\_1.fq.gz", "./test/Con-1\_2.fq.gz"],

"S2": ["./test/Con-2\_1.fq.gz", "./test/Con-2\_2.fq.gz"],

"S3": ["./test/Sam-1\_1.fq.gz", "./test/Sam-1\_2.fq.gz"],

"S4": ["./test/Sam-2\_1.fq.gz", "./test/Sam-2\_2.fq.gz"]

},

# 单末端文件示例：

# "samples": {

# "S1": ["./test/Con-1\_1.fq.gz"],

# "S2": ["./test/Con-2\_1.fq.gz"],

# "S3": ["./test/Sam-1\_1.fq.gz"],

# "S4": ["./test/Sam-2\_1.fq.gz"]

# },

1. groups

# 样本分组信息，其中S1和group1为示例可以更改为其他的名称，但要与上下的参数对应

"groups": {

"group1": ["S1", "S2"],

"group2": ["S3", "S4"]

},

1. compare

# 样本比较信息 对照组在前 ，实验组在后。如果多组比较则在后面新建["group3", "group4"]等内容。

"compare": [["group1", "group2"],],

1. layout

# 测序类型（单双末端） paired或者single

"layout": "paired",

1. format

# 测序文件类型 fastq或者fasta

"format": "fastq",

1. output\_dir

# 结果目录（默认为./RSNAP\_out/）

"output\_dir": "./RSNAP\_out/",

1. query

在uniport上的检索关键字，用于获得uniport结果文件。

# 检索关键词，例如"(taxonomy\_id:4751) AND (reviewed:true)"

详细请查看手册：https://www.uniprot.org/help/query-fields

"query": "(taxonomy\_id:4751) AND (reviewed:true)",

1. addition\_fields

检索的返回条目，如果有额外需要可以添加

# '检索该物种信息的额外条目(列)，用逗号分割，默认为:"accession", "xref\_geneid","gene\_names", "organism\_id", "organism\_name", "cc\_pathway", "go\_id", "go"')

# https://www.uniprot.org/help/return\_fields

# 如果不需要则为空字符''

"addition\_fields": ''

* 其他参数

下面是子程序的一些非必要参数，采用默认的参数，如有需要可以更改。

1. # Quality\_Control默认参数

qc\_args = {

"script": "./Quality\_Control.py", # 使用的脚本，后续可以更新拓展

"not\_cut": False, # 是否执行裁剪测序数据，默认为是(False表示不不切割，如过为True则为切割)

"threads": "10", # 线程数，默认为10

"threshold": 0.02, # 进行裁剪长度判断时的阈值，默认为0.01

"software": "trimmomatic", # 进行裁剪的程序，默认为trimmomatic

}

1. # Assemble\_trinity默认参数

at\_args = {

"script": "./Assemble\_trinity.py", # 使用的脚本，后续可以更新拓展

"result\_file": "assemblies.fa", # 组装结果文件名

"summary\_file": "summary.txt", # 组装摘要结果文件名

"max\_memory": "60", # 最大内存占用数（G）（默认为6G）

"cpu": "4", # cup核数，默认为4

}

1. # Expression\_evaluation\_rsem默认参数

eer\_args = {

"script": "./Expression\_evaluation\_rsem.py", # 使用的脚本，后续可以更新拓展

"threads": "10", # 线程数，默认为8

}

1. # Differential\_expression\_gene\_identification默认参数

degi\_args = {

"script": "./Differential\_expression\_gene\_identification.py", # 使用的脚本，后续可以更新拓展

"software": "edgeR", # 差异分析R包,默认为'edgeR'。支持("edgeR", "DESeq2")

"foldchange": "1", # 差异分析log2foldchange筛选阈值，默认为1

"padj": "0.05", # 异分析矫正p值padj筛选阈值，默认为0.05

}

1. # Transcript\_identification默认参数

ti\_args = {

"script": "./Transcript\_identification.py", # 使用的脚本，后续可以更新拓展

"cpu": "10", # CPU核数(默认为1)

"evalue": 10 \*\* -5, # evalue值(默认为10e-5)

}

1. # Feature\_annotation\_analysis默认参数

faa\_args = {

"script": "./Feature\_annotation\_analysi.py", # 使用的脚本，后续可以更新拓展

"all\_out": "all\_anno.tsv", # 所有可鉴别转录本的各类注释信息结果文件名，默认为all\_anno.tsv

"deg\_out": "deg\_anno.tsv", # 差异表达基因的各类注释信息结果文件名，默认为deg\_anno.tsv

"enrichment\_out": "enrichment\_out", # 差异表达基因的各类注释信息的富集分析结果文件目录名,默认为enrichment\_out（即在output目录下的enrichment\_out子目录）

}

注：参数赋值尽量避免使用中文字符，否则可能导致程序出现异常错误。

### 使用案例

1. **数据介绍**

本次的数据来源于NCBI BioProject数据库，ID为PRJNA350822。为3对3的样本，其中实验组为使用NaCl处理（3g），对照组为做对照处理（S0g）。（https://www.ncbi.nlm.nih.gov/bioproject/PRJNA350822）

1. **参数设置和命令示例**
2. **参数设置**

本案例中的配置参数模块的内容如下，其含义为：

在用户输入参数中，样本绝对路径为/home/zhanggaochuan/Dr.Wei/2019-AP-RNA-Seq-NaCl/combined/目录下的六个样本文件；样本分组信息为group1为对照组，group2为实验组；测序类型为双末端；测序文件类型为fastq；结果目录为当前目录下的RSNAP\_out目录；Uniport检索关键词为"(taxonomy\_id:4751) AND (reviewed:true)"检索物种id为4751（真菌），并且经过人工审查。

在Quality\_Control参数中，进行裁剪数据；线程数目为10；进行裁剪长度判断时的阈值为0.02；进行裁剪的程序为trimmomatic。

在Assemble\_trinity参数中，组装结果文件名为assemblies.fa；组装摘要结果文件名为summary.txt；最大内存占用数（G）为60；cup核数为4。

在Expression\_evaluation\_rsem参数中，线程数为10。

在Differential\_expression\_gene\_identification参数中，差异分析R包为edgeR；差异分析log2foldchange筛选阈值为1；异分析矫正p值padj筛选阈值为0.05。

在Transcript\_identification参数中，CPU核数为10；evalue筛选阈值为10e-5。

在Feature\_annotation\_analysis参数中，所有可鉴别转录本的各类注释信息结果文件名为all\_anno.tsv；差异表达基因的各类注释信息结果文件名为deg\_anno.tsv；差异表达基因的各类注释信息的富集分析结果文件目录名为enrichment\_out。

config.py文件具体设置如下：

user\_args = {

# 样本绝对路径（双末端）

"samples": { # 左 #右

"S4": ["/home/zhanggaochuan/Dr.Wei/2019-AP-RNA-Seq-NaCl/combined/3g-L-1\_combined\_R1.fastq.gz", "/home/zhanggaochuan/Dr.Wei/2019-AP-RNA-Seq-NaCl/combined/3g-L-1\_combined\_R2.fastq.gz"],

"S5": ["/home/zhanggaochuan/Dr.Wei/2019-AP-RNA-Seq-NaCl/combined/3g-L-2\_combined\_R1.fastq.gz", "/home/zhanggaochuan/Dr.Wei/2019-AP-RNA-Seq-NaCl/combined/3g-L-2\_combined\_R2.fastq.gz"],

"S6": ["/home/zhanggaochuan/Dr.Wei/2019-AP-RNA-Seq-NaCl/combined/3g-L-3\_combined\_R1.fastq.gz", "/home/zhanggaochuan/Dr.Wei/2019-AP-RNA-Seq-NaCl/combined/3g-L-3\_combined\_R2.fastq.gz"],

"S1": ["/home/zhanggaochuan/Dr.Wei/2019-AP-RNA-Seq-NaCl/combined/S0g-L-1\_combined\_R1.fastq.gz", "/home/zhanggaochuan/Dr.Wei/2019-AP-RNA-Seq-NaCl/combined/S0g-L-1\_combined\_R2.fastq.gz"],

"S2": ["/home/zhanggaochuan/Dr.Wei/2019-AP-RNA-Seq-NaCl/combined/S0g-L-2\_combined\_R1.fastq.gz", "/home/zhanggaochuan/Dr.Wei/2019-AP-RNA-Seq-NaCl/combined/S0g-L-2\_combined\_R2.fastq.gz"],

"S3": ["/home/zhanggaochuan/Dr.Wei/2019-AP-RNA-Seq-NaCl/combined/S0g-L-3\_combined\_R1.fastq.gz", "/home/zhanggaochuan/Dr.Wei/2019-AP-RNA-Seq-NaCl/combined/S0g-L-3\_combined\_R2.fastq.gz"],

},

# 样本分组信息,对照组放前面，实验组放在后面

"groups": {

"group1": ["S1", "S2", "S3"],

"group2": ["S4", "S5", "S6"]

},

},

# 样本比较信息 对照组在前 ，实验组在后

"compare": [["group1", "group2"], ["group3", "group4"], ],

# 测序类型（单双末端） paired或者single

"layout": "paired",

# 测序文件类型 fastq或者fasta

"format": "fastq",

# 结果目录（默认为./RSNAP\_out/）

"output\_dir": "./RSNAP\_out\_1\_6/",

# 检索关键词，例如"(taxonomy\_id:4751) AND (reviewed:true)"\n详细请查看手册：https://www.uniprot.org/help/query-fields

"query": "(taxonomy\_id:4751) AND (reviewed:true)",

# '检索该物种信息的额外条目(列)，用逗号分割，默认为:"accession", "xref\_geneid","gene\_names", "organism\_id", "organism\_name", "cc\_pathway", "go\_id", "go"')

# https://www.uniprot.org/help/return\_fields

# 如果不需要则为空字符''

"addition\_fields": ''

}

# Quality\_Control默认参数

qc\_args = {

"script": "./Quality\_Control.py", # 使用的脚本，后续可以更新拓展

"not\_cut": False, # 是否执行裁剪测序数据，默认为是(False表示不不切割，如过为True则为切割)

"threads": "10", # 线程数，默认为10

"threshold": 0.02, # 进行裁剪长度判断时的阈值，默认为0.01

"software": "trimmomatic", # 进行裁剪的程序，默认为trimmomatic

}

# Assemble\_trinity默认参数

at\_args = {

"script": "./Assemble\_trinity.py", # 使用的脚本，后续可以更新拓展

"result\_file": "assemblies.fa", # 组装结果文件名

"summary\_file": "summary.txt", # 组装摘要结果文件名

"max\_memory": "60", # 最大内存占用数（G）（默认为6G）

"cpu": "4", # cup核数，默认为4

}

# Expression\_evaluation\_rsem默认参数

eer\_args = {

"script": "./Expression\_evaluation\_rsem.py", # 使用的脚本，后续可以更新拓展

"threads": "10", # 线程数，默认为8

}

# Differential\_expression\_gene\_identification默认参数

degi\_args = {

"script": "./Differential\_expression\_gene\_identification.py", # 使用的脚本，后续可以更新拓展

"software": "edgeR", # 差异分析R包,默认为'edgeR'。支持("edgeR", "DESeq2")

"foldchange": "1", # 差异分析log2foldchange筛选阈值，默认为1

"padj": "0.05", # 异分析矫正p值padj筛选阈值，默认为0.05

}

# Transcript\_identification默认参数

ti\_args = {

"script": "./Transcript\_identification.py", # 使用的脚本，后续可以更新拓展

"cpu": "10", # CPU核数(默认为1)

"evalue": 10 \*\* -5, # evalue值(默认为10e-5)

}

# Feature\_annotation\_analysis默认参数

faa\_args = {

"script": "./Feature\_annotation\_analysi.py", # 使用的脚本，后续可以更新拓展

"all\_out": "all\_anno.tsv", # 所有可鉴别转录本的各类注释信息结果文件名，默认为all\_anno.tsv

"deg\_out": "deg\_anno.tsv", # 差异表达基因的各类注释信息结果文件名，默认为deg\_anno.tsv

"enrichment\_out": "enrichment\_out", # 差异表达基因的各类注释信息的富集分析结果文件目录名,默认为enrichment\_out（即在output目录下的enrichment\_out子目录）

}

1. **命令行实现**

* 若在本地目录

在运行程序时，首先确保所有文件具有可执行权限：

chmod 755 ./\*

执行RSNAP主程序，由于config文件默认为./ config.py，故不用指定。因为程序运行需要很长时间。故让其后台运行，并且将日志输出到mylog.log文件：

nohup ./RSNAP.py > mylog.log &

* 放入bin目录

也可将所有程序及其模板放入/urs/bin目录

sudo cp ./\* /usr/bin/

然后执行RSNAP主程序，由于config文件与主程序不在同一目录，故需要指定绝对路径。因为程序运行需要很长时间。故让其后台运行，并且将日志输出到mylog.log文件：

nohup ./RSNAP.py –config /home/ubuntu/RSNAP/config.py > mylog.log &

### 结果解释

在./RSNAP\_out/目录下，结果目录如下表格 1 结果文档和目录结构表：

表格 1 结果文档和目录结构表

|  |  |
| --- | --- |
| **结果文档和目录结构** | **描述** |
| --APIGMT.log | RSNAP日志文件 |
| --all\_data.csv | 结果总表 |
| --report.html | 运行结果报告文件 |
| --qc\_fastqc\_output  |--quality\_control.log  |--\*\_fastqc.zip  |--\*\_summary.txt  |--\*\_fastqc.html  |--\*\_ trim\_unpaired.fq.gz  |--\*\_trim.fq.gz | 质量控制结果目录  质量控制日志  Fastqc结果文件  质控摘要文件  Fastqc报告文件  裁剪后的非配对文件  裁剪后的结果文件 |
| --assemble\_trinity\_output  |--assemble\_trinity.log  |--summary.txt  |--assemblies.fa  |--其他文件及其目录 | 组装结果目录  组装日志  组装摘要文件  组装结果文件  Trinity运行结果 |
| --expression\_rsem\_output  |--expression\_evaluation\_rsem.log  |--genes.expected\_count.matrix  |--其他文件及其目录 | 表达水平评估结果目录  表达水平评估日志  基因计数矩阵  RESM运行结果 |
| --DEG\_identification\_output  |--differential\_expression\_gene\_identification.log  |--scatter.pdf  |--plotMDS2.pdf  |--volcano.pdf  |--clustering\_tree\_plot.pdf  |--sigDEGs\_pheatmap.pdf  |--sigDEG.tsv  |--edgeR\_allGenes.out  |--edgeR\_upDEGs\_FDR0.05\_logFC1.out  |--edgeR\_downDEGs\_FDR0.05\_logFC1.out  |--edgeR\_sigDEGs\_FDR0.05\_logFC1.out  |--edgeR\_sigDEGs\_FDR0.05.out | 差异表达基因鉴别结果目录  差异表达基因鉴别日志  scatter map 散点图  多维缩放（MDS）聚类图  火山图  层次聚类树状图  差异表达基因聚类热图  差异表达基因摘要结果  所有转录本差异分析结果  上调基因差异分析结果  下调基因差异分析结果  按FDR和logFC筛选的差异基因差异分析结果  按FDR筛选的差异基因差异分析结果 |
| --transcript\_indentification\_output  |--differential\_expression\_gene\_identification.log  |--blastdb  |--all\_identified\_transcripts.tsv  |--blastx.stat  |--blastx.summary  |--taxonomy\_id\_4751\_and\_reviewed\_true.fasta  |--taxonomy\_id\_4751\_and\_reviewed\_true.fasta.gz  |--taxonomy\_id\_4751\_and\_reviewed\_true.outfmt7  |--taxonomy\_id\_4751\_and\_reviewed\_true.summary  |--UniprotKB\_records.txt | 转录本鉴别结果目录  转录本鉴别日志  Blast数据库文件夹  所有鉴别到的转录本  Blastx结果统计信息  Blastx结果总结  Uniport下载的测序文件  Uniport下载的测序文件(压缩)  Blastx结果文件  Uniport下载的注释信息文件  Uniport检索关键字日志 |
| --feature\_annotation\_analysis\_output  |--feature\_annotation\_analysis.log  |--all\_anno.tsv  |--deg\_anno.tsv  |--gene2anno\_Gene Ontology (GO).tsv  |--gene2anno\_Pathway.tsv  |--enrichment\_out  |--up\_pathway\_dotplot.pdf  |--up\_pathway\_barplot.pdf  |--up\_gene\_ontology\_go\_dotplot.pdf  |--up\_gene\_ontology\_go\_barplot.pdf  |--down\_pathway\_dotplot.pdf  |-- down\_pathway\_barplot.pdf  |-- down\_gene\_ontology\_go\_dotplot.pdf  |-- down\_gene\_ontology\_go\_barplot.pdf | 功能注释分析结果目录  功能注释分析日志  所有鉴定到的转录本的注释信息  差异基因的注释信息  GO富集分析结果表  KEGG通路富集分析结果表  富集分析可视化结果目录  上调基因KEGG通路富集分析气泡图  上调基因KEGG通路富集分析柱状图  上调基因GO富集分析气泡图  上调基因GO富集分析柱状图  下调基因KEGG通路富集分析气泡图  下调基因KEGG通路富集分析柱状图  下调基因GO富集分析气泡图  下调基因GO富集分析柱状图 |