流程优化组重构sRNA_v1.0流程使用说明

`sRNA_v1.0重构流程脚本:/ifs/TJPROJ3/RAD/Pipeline_result/smallRNA/BioMain/smallrna_pipeline.py`
`sRNA_v1.0重构流程说明文档存放：/ifs/TJPROJ3/RAD/Pipeline_result/smallRNA/BioDoc`
`sRNA_v1.0重构流程测试目录：/ifs/TJPROJ3/RAD/Pipeline_result/smallRNA/BioDemo/zma1`

如果对本流程有任何疑问，请联系诺禾致源流程优化组
联系人：王云凯
邮箱： wangyunkai@novogene.com

重构流程更新说明

- sRNA_v1.0流程基于转录调控业务线sRNA_v2.3流程进行重构
- 目前重构流程植物有参与无参均已经测通，动物有参无参预计2018/05/19前测通

- 重构流程采用SGE调度系统，通过sjm进行任务投递，减少任务人为投递时间，更改为更为合理的串并行

```
如果使用者环境变量中没有配置sjm环境变量，任务运行前可以在本地执行
export LD_LIBRARY_PATH=/PUBLIC/software/public/System/boost_1_55_0/lib:$LD_LIBRARY_PATH`
然后通过/PUBLIC/software/public/System/sjm-1.2.0/bin/sjm  *.job进行任务投递
```

- sRNA_v1.0重构流程较sRNA_v2.3流程目录结构变化较大

```
目录结构如下：
注：为了让项目执行人员熟记项目分析内容，项目分析的编号和内容固定，例如：
    5.repeat中5和repeat进行绑定，如无参项目不进行repeat分析，NAT，gene分析，则没有5.repeat等目录，
    直接生成8.novel目录，但是释放结果目录和原来一致，编号随分析内容递增。

    修改raw_data目录结构，原始数据分样本存储
    QC_results目录从QC目录中移到项目分析下
    所有分析目录下面将不在含有abbr对应的子目录，分析内容直接存在对应模块分析下，使项目分析架构更清晰，例如：
    abbr 物种缩写为zma，原目录结构如下：
    3.known/zma/{expression_analyses，pdfs_zma.known，ref，zma.known}
    修改为：
    3.known/{expression_analyses，known_miRNAs，known_miRNAs_expressed_pdfs，miRbase}
    其他目录架构，修改类似

目录结构：

└── raw_data
│   ├── sampleA
│   └── sampleB
├── 1.QC
│   ├── sampleA
│   └── sampleB
├── 2.map
│   ├── sampleA
│   └── sampleB
├── P101SC17060525-05-Zea_mays_QC_results
│   ├── results
│   └── src
├── 3.known
│   ├── expression_analyses
│   ├── known_miRNAs
│   ├── known_miRNAs_expressed_pdfs
│   └── miRbase
├── 4.ncRNA
```

```
│   ├── input
│   └── output
├── 5.repeat
│   └── repeat_result
├── 6.NAT
│   ├── input
│   └── output
├── 7.gene
│   ├── input
│   └── output
├── 8.novel
│   ├── expression_analyses
│   ├── miRbase
│   ├── novel_miRNAs
│   ├── novel_miRNAs_expressed_pdfs
│   └── predict_novel_miRNA
├── 9.TAS
│   ├── input
│   ├── known_TAS
│   ├── output
│   ├── phase.out
│   └── phase.predict
├── 10.Category
├── 11.edit_family
│   ├── edit_analy
│   └── family_analy
├── 12.target
├── 13.diff
│   └── diffAnalysisResult
├── Blast
│   └── temp
├── 14.enrich
│   └── sampleAvssampleB
├── P101SC17060525-05-Zea_mays_sRNA_result
│   ├── P101SC17060525-05-Zea_mays_report
│   └── P101SC17060525-05-Zea_mays_results
├── log
│   ├── *.job
```

- 不在生成libraryID和P101SC17060525-05-Zea_mays_parameter.txt文件，流程根据mapfile文件信息， 自行判断运行样本，根据流程参数自行check和读取项目信息
- 重新编写result和report的脚本，是result和report分析能为一个单纯的分析模块，代码不在依赖主流程进行生成，便于后期维护
- 添加GO功能富集pvalue的check，当项目分析过程中GO功能富集结果为空时，可以通过查看对应比较组合的功能富集.e文件，将会提醒： NO values meet the filter criteria !!!

有参流程脚本配置

主要功能：

```
    1.重写sRNA分析流程，不在包含generate*.sh等嵌套脚本，直接在对应的分析目录下面生成分析内容脚本，便于熟悉分析内容与流程维护。
    2.根据项目要求准备run.sh，运行run.sh生成*job文件，通过sjm进行任务投递。
    3.--fq_dir 参数删除，准备线下项目数据和诺禾下机数据一样，进行mapfile文件配置
      --ownername --yunying目前没有使用，项目运行时可以不进行填写
    4.添加参数：
      --new_job  用于生成job及项目日志目录，默认为年.月.日.时.job，建议传递参数
      --sched    sjm投递的参数，默认为根据qselect提取用户可用队列构造,为了防止跟命令行参数冲突，
                 本参数传递时必须加引号例如："-V -cwd -S /bin/bash -q rad.q "
      --mdspedb  存储已准备好配置文件的模式物种数据库路径


参数选择：
--project            项目编号-B*-4-物种拉丁名(*代表分期数)
--contract           项目编号_合同名称（以下划线分隔)
                     例：P101SC16120904_北京友谊医院3个小鼠SmallRNA测序及分析软件RNAseq-Mus-ZhD的开发
--English            生成英文结题报告，填此参数，“y”，中文报告不填此参数；
--ownername          信息分析的中文名字 [目前可以不用填写]
--yunying            运营的中文名字      [目前可以不用填写]
--org                物种，只能是植物（refplant）或是动物（refanimal）
--mapfile            mapfile文件
--sample             样本名，以逗号隔开，e.g.TR1,TR2,TS1,TS2
--group              样本分组方式，组内用冒号隔开，组间用逗号隔开
                     e.g. TR1,TR2,TS1,TS2,TR1:TR2,TS1:TS2【选填】
```

```
--groupname          组名，以逗号隔开,对应分的样本组 e.g.TR1,TR2,TS1,TS2,TR,TS 【选填】
--compare            样本组比较方式，处理:对照，组内用冒号隔开，组间用逗号隔开
                     e.g."2:1,1:3,2:3"【选填】
                     如果比较组合过多，可以将信息搜集表中的compare组合整理成com.txt文件，格式如下，然后填此文件即可：
                    |组合1 |FmE |M_E |
                    |组合2 |FmL |M_L |
                    |组合3 |FmP |M_P |
                    |组合4| F_A |M_A |
--venn_cluster       venn画图方式，适合2~4 组比较；同一张venn图内的比较组用下划线隔开，
                     不同的venn图间用逗号隔开，默认不画
                     e.g."2:1_1:3_2:3,1:3_2:3"
                     注：只有一组compare时，次参数不填，默认做单样本间venn图【选填】；如果比较组合过多，
                     可将信息搜集表中的venn组合整理成一个如下的venn.txt文本,此参数填此venn.txt即可：
                     |组合比较Venn 1| 组合1 |组合2 |组合3
                     |组合比较Venn 2 |组合2 |组合3 |组合4 |组合5
--mdspe              e.g. hsa_GRCh38.Ensemble.87 （拉丁名缩写_基因组版本信息）
                     注意如果使用此参数，数据准备文件格式需要规范化，可参考已准备好的文件路
                     径：/BJPROJ/RNA/reference_data/sRNA/hsa/GRCh38.Ensembl.84，
                     填写此参数后则--refer --go --geneAnn --rRNA --tRNA --snRNA --snoRNA
                     --repdir --exon --intron --gtf --utr3 参数均不需再填写。

--refer              基因组参考文件
--abbr               物种缩写，3个小写字母，由老师提供，可以是多个，以逗号分隔，相应list
                     在/ifs/TJPROJ3/RAD/wangyunkai/sRNA/BioDB/miRBase21/organisms.txt，
                     如果老师要分析miRBase中所有已知miRNA,则--abbr
                     参数填 all ，所有动物填animal ，所有植物填plant
--mode               预测novel模式，value only be1,2,3; 1 for animal; 2 for monocot; 3 for dicots
--go                 go 文件
--geneAnn            gene注释文件
--chrNum             需要画密度图的染色体数
--rRNA               rRNA.fa，提前准备
--tRNA               tRNA.fa，提前准备
--snRNA              snRNA.fa，提前准备
--snoRNA             snoRNA.fa，提前准备
--spe                物种的全名，如没有，选择相近物种，repeat分析时用到
                     list在/ifs/TJPROJ3/RAD/wangyunkai/sRNA/BioModule/5repeat/RepeatMasker_Species.txt
--repdir             repeat预测结果目录，提前准备
--exon               exon.fa
--intron             intron.fa
--gtf                gtf 文件
--kegg               KEGG 缩写，list在/ifs/TJPROJ3/RAD/wangyunkai/sRNA/BioDB/KEGG/kobas2.0-20140801/abbr_list.txt
                     优先级: "species itself> related species "
--utr3               3UTR.fa 动物需要（转录本3'UTR）【动物必填】
--NAT                NAT物种缩写，list在/ifs/TJPROJ3/RAD/wangyunkai/sRNA/BioDB/PlantNATsDB/spe.list
                     如果没有 用"other"代替【植物必填】
--gene               transcript.fa， (转录本序列)【植物必填】
--common             如果分析内容里有"4.样品间的公共序列和特异序列的分析（>;=2个样本)"，
                     填写y，如果没有填写n，单个样本填n（新增参数），默认是n
--exosome            如果是外泌体项目填写‘y’，其他项目不填
--type               根据物种类别在以下三个中选择一个3utr_human、3utr_fly或者3utr_worm"【动物必填】
--new_job  用于生成job及项目日志目录，默认为年.月.日.时.job，建议传递参数
--sched    sjm投递的参数，默认为根据qselect提取用户可用队列构造,为了防止跟命令行参数冲突，
                     本参数传递时必须加引号例如："-V -cwd -S /bin/bash -q rad.q "
--mdspedb  存储已准备好配置文件的模式物种数据库路径
```

无参流程脚本配置

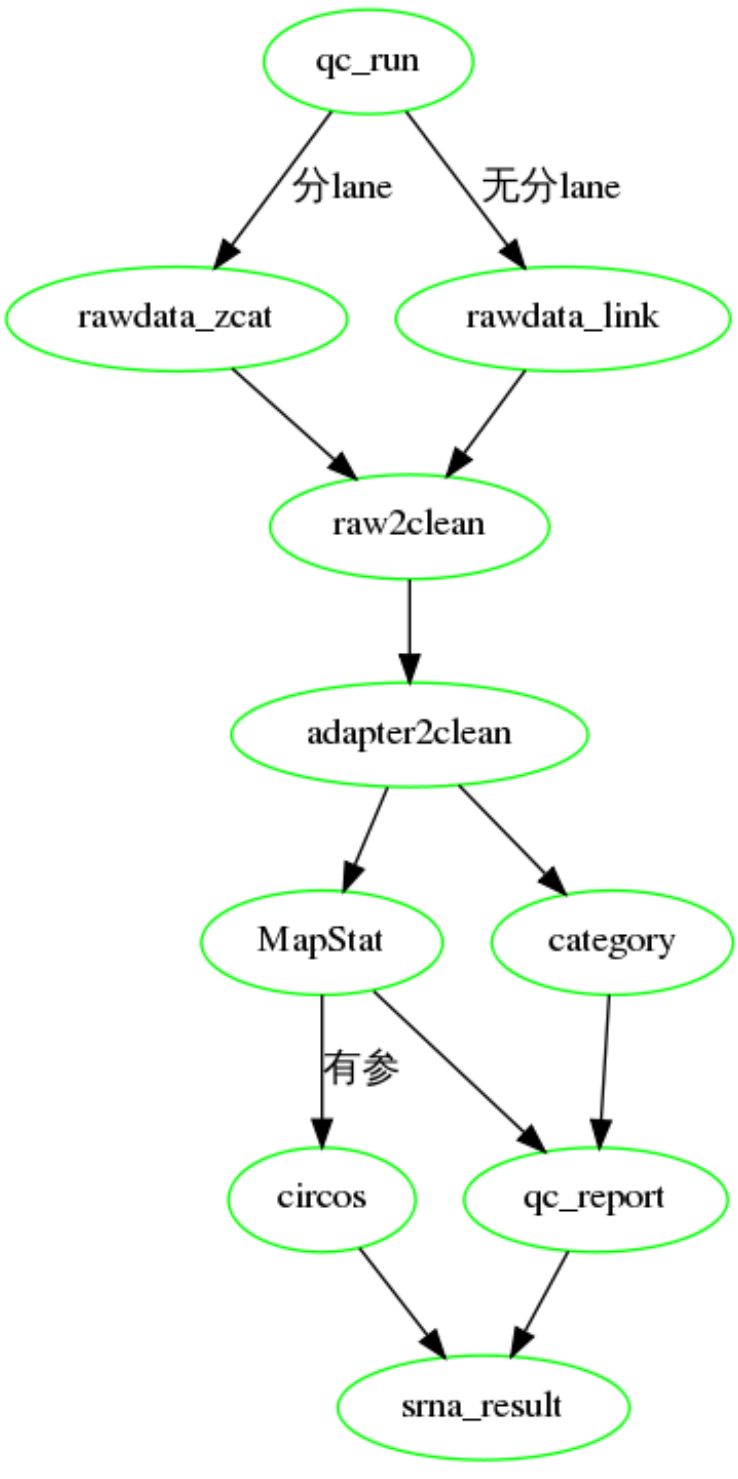
参数选择：

```
--project            项目编号-B*-4-物种拉丁名(*代表分期数)
--English            生成英文结题报告，填此参数，“y”，中文报告不填此参数；
--contract           项目编号_合同名称（以下划线分隔）
                     例：P101SC16120904_北京友谊医院3个小鼠SmallRNA测序及分析软件RNAseq-Mus-ZhD的开发
--code_number        结题编号，例：P101SC16120904-B1(B后面的数字对应分期)
--org                物种，只能是植物（norefplant）或是动物（norefanimal）
--ownername          信息分析的中文名字
--yunying            运营的中文名字
--mapfile            mapfile文件,三列，下机路径\t文库号\t样本名
--sample             样本名，以逗号隔开，e.g.TR1,TR2,TS1,TS2
 --group               样本分组方式，组内用冒号隔开，组间用逗号隔开
                     e.g. TR1,TR2,TS1,TS2,TR1:TR2,TS1:TS2【选填】
--groupname          组名，以逗号隔开,对应分的样本组 e.g.TR1,TR2,TS1,TS2,TR,TS 【选填】
--compare            样本组比较方式，处理:对照，组内用冒号隔开，组间用逗号隔开
                     e.g."2:1,1:3,2:3"【选填】
                     如果比较组合过多，可以将信息搜集表中的compare组合整理成com.txt文件，格式如下，然后填此文件即可：
```

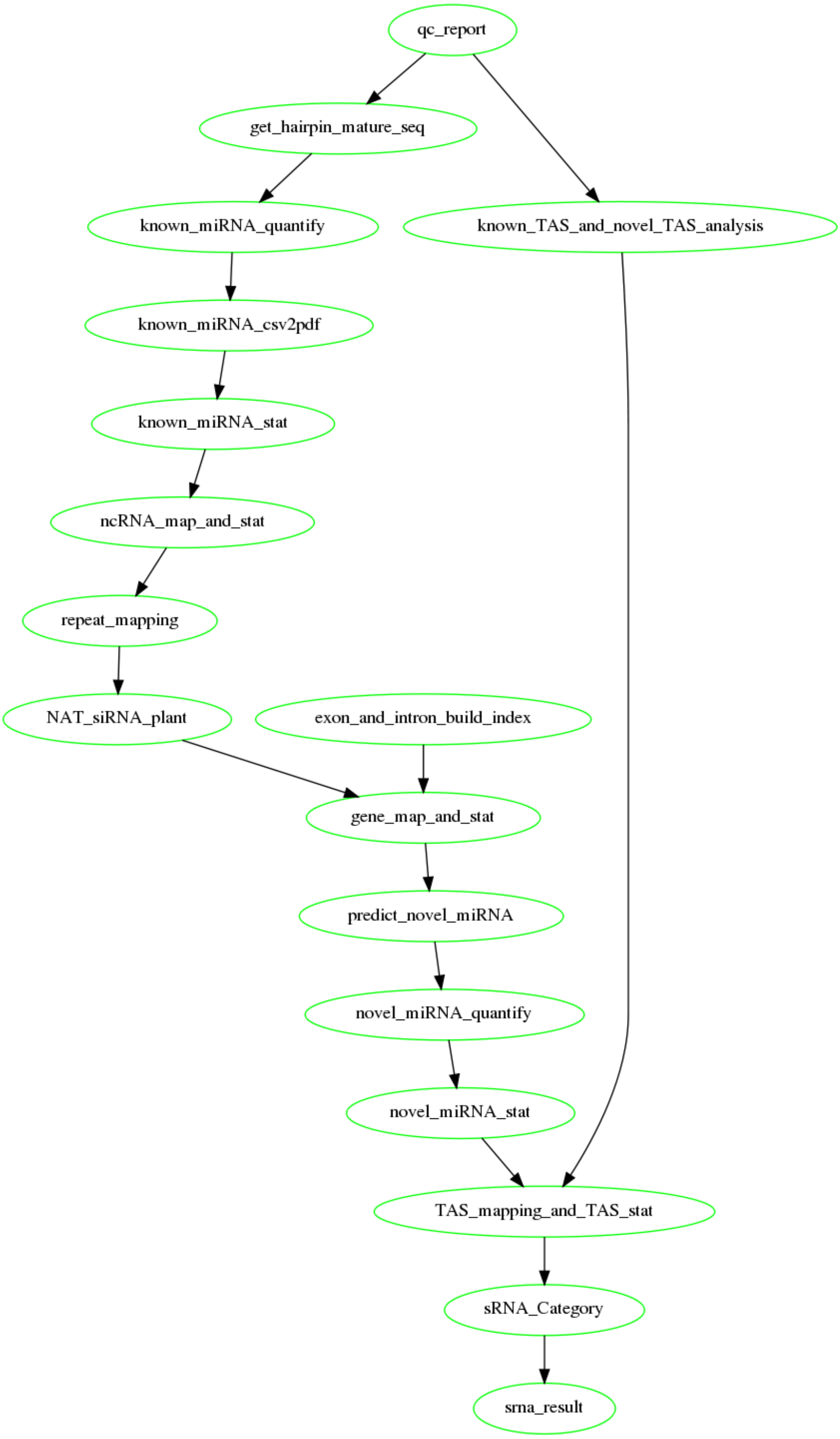
```
                    |组合1 |FmE |M_E |
                    |组合2 |FmL |M_L |
                    |组合3 |FmP |M_P |
                    |组合4| F_A |M_A |
--venn_cluster      venn画图方式，适合2~4 组比较; 同一张venn图内的比较组用下划线隔开，
                    不同的venn图间用逗号隔开，默认不画
                    e.g."2:1_1:3_2:3,1:3_2:3"
                    注：只有一组compare时，次参数不填，默认做单样本间venn图【选填】；如果比较组合过多，
                    可将信息搜集表中的venn组合整理成一个如下的venn.txt文本,此参数填此venn.txt即可：
                    |组合比较Venn 1| 组合1 |组合2 |组合3
                    |组合比较Venn 2 |组合2 |组合3 |组合4 |组合5
--refer             unigene.fasta
--abbr              物种缩写，3个小写字母，由老师提供，可以是多个，以逗号分隔，相应list
                    在/ifs/TJPROJ3/RAD/wangyunkai/sRNA/BioDB/miRBase21/organisms.txt，
                    如果老师要分析miRBase中所有已知miRNA,则--abbr
                    参数填 all , 所有动物填animal , 所有植物填plant
--mode              预测novel模式, value only be1,2,3; 1 for animal; 2 for monocot; 3 for dicots
--go                go 文件
--geneAnn           gene注释文件
--gff3              包含CDS信息的gff3文件【动物必填】
--ko                ko 文件
--length            gene长度信息文件
--gene              gene.fa， 植物需要【植物必填】，unigene.fa
--common            如果分析内容里有"4.样品间的公共序列和特异序列的分析（>;=2个样本)"，
                    填写y，如果没有填写n，单个样本填n（新增参数），默认是n
--new_job   用于生成job及项目日志目录，默认为年.月.日.时.job，建议传递参数
--sched     sjm投递的参数，默认为根据qselect提取用户可用队列构造,为了防止跟命令行参数冲突，
                本参数传递时必须加引号例如："-V -cwd -S /bin/bash -q rad.q "
--mdspedb   存储已准备好配置文件的模式物种数据库路径
```

项目流程图

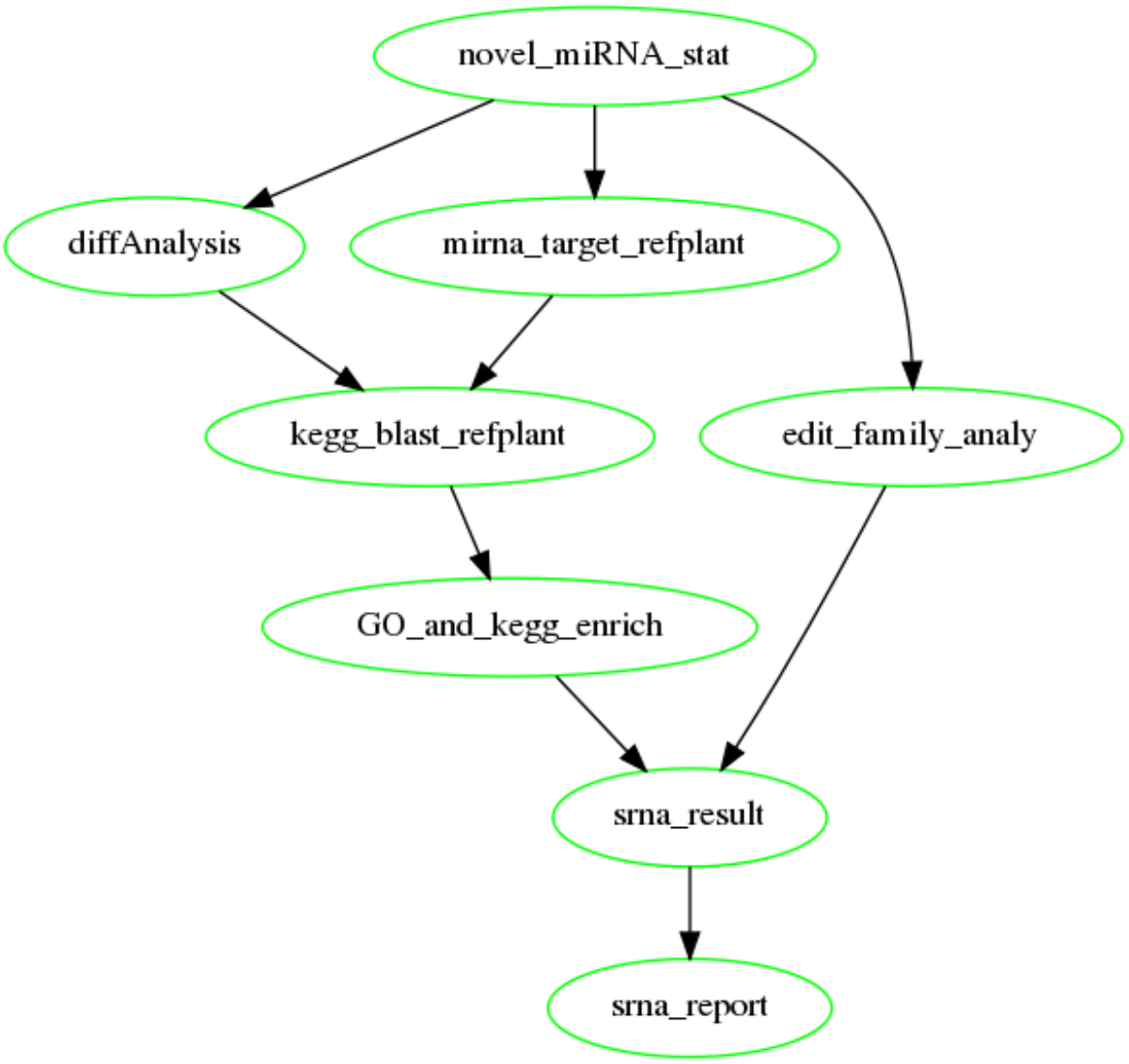
质控和比对分析



sRNA分类筛选


qc_report
get_hairpin_mature_seq
known_miRNA_quantify
known_TAS_and_novel_TAS_analysis
known_miRNA_csv2pdf
known_miRNA_stat
ncRNA_map_and_stat
repeat_mapping
NAT_siRNA_plant
exon_and_intron_build_index
gene_map_and_stat
predict_novel_miRNA
novel_miRNA_quantify
novel_miRNA_stat
TAS_mapping_and_TAS_stat
sRNA_Category
srna_result

miRNA分析



执行方式

```
sh run.sh 生成*job文件
export LD_LIBRARY_PATH=/PUBLIC/software/public/System/boost_1_55_0/lib:$LD_LIBRARY_PATH
/PUBLIC/software/public/System/sjm-1.2.0/bin/sjm *job 进行任务投递

PS：当物种为植物时，当运行到靶基因预测的时候，需要将projpath/12.target/*transcript.fa和mature.fa下载下来，
在http://plantgrn.noble.org/psRNATarget/analysis?function=3 下手动输入文件，
进行在线靶基因的预测，大约花费10min，将结果文件下载下来后上传到projpath/12.target/路径下，重新投递*.job.staus即可
例如：/PUBLIC/software/public/System/sjm-1.2.0/bin/sjm *job.staus

注：初步测试，流程优化组重构流程整体运行时间，以玉米为例（不考虑项目执行人投递任务中间隔时间）：
    原sRNA_v2.3流程运行时间为60小时
    流程优化组重构流程项目分析时间大约为40小时
    固现流程提速在33%左右
```