RNApip : NGS pipelines made easy

Patrick Schorderet

Patrick.schorderet@molbio.mgh.harvard.edu

Jan 2015

1. Introduction

RNApip is a perl/R package that supports users during the analysis of next generation sequencing (NGS) data as part of the NEAT toolkit (NGS easy analysis toolkit). RNApip, in conjuncture with the NEAT package, provides an easy, rapid and reproducible exploratory data analysis (EDA) tool that allows users to assess RNAseq data in less than 12 hours (based on a 200mio read Highseq run). As such, RNApip manages many of the repetitive, error-prone tasks required for NGS data analysis. It is versatile and easily configurable to meet each user's preferences. RNApip accompanies the user from compressed fastq files (.fastq.gz), usually provided by the sequencing core facility, to bam files using a single command line.

A central feature of RNApip is its ability to perform repetitive tasks on complex sample setups while managing batch submissions and cluster queuing. RNApip can easily be implemented in any institution with limited to no programming experience. The workflow has been designed to efficiently run on a computer cluster using a distributed resource manager such as TORQUE. RNApip has been developed by and for wet-lab scientists as well as bioinformaticiens to ensure user-friendliness, management of complicated experimental setups and reproducibility in the big data era. To start using RNApip, please follow the turotial. This will walk you through the analysis of a small test dataset (provided as part of RNApip) using your own computer cluster. This will also ensure RNApip and its dependencies are correctly installed before submitting large, memory-savvy analysis.

All fastq files from the test data have been subsetted to ca. 15'000 reads. This data comes from an unpublished 50bp single end (SE) sequencing experiment although

RNApip can deal with paired-end (PE) sequencing as well. For more information on the test data provided in this tutorial, please read below.

Although RNApip can be run by scientists with limited to no programming experience, this tutorial require access to a remote server. Users are thereby required to have SSH accessibility with a username and a password. Please refer to your system administrator to obtain such credentials. For more information on how to access a remote server through SSH, please read below.

2. Getting started with UNIX

Notes

In the following tutorial, all unix/perl/R command lines will be bold, italicized and highlighted in blue. Most will be embedded in tables. The command line is the text following, but not including, the dollar sign ($).

[~]$ this is a unix command

This tutorial is intended to run on MacOS/LINUX environments. For MacOS users, we suggest to use the *Terminal* (applications/Terminal) for all following steps. The terminal/shell output will be depicted in black.

Copy pasting the *unix* commands should allow you to follow all steps of this tutorial. Please be aware that *unix* commands are case sensitive, including white spaces.

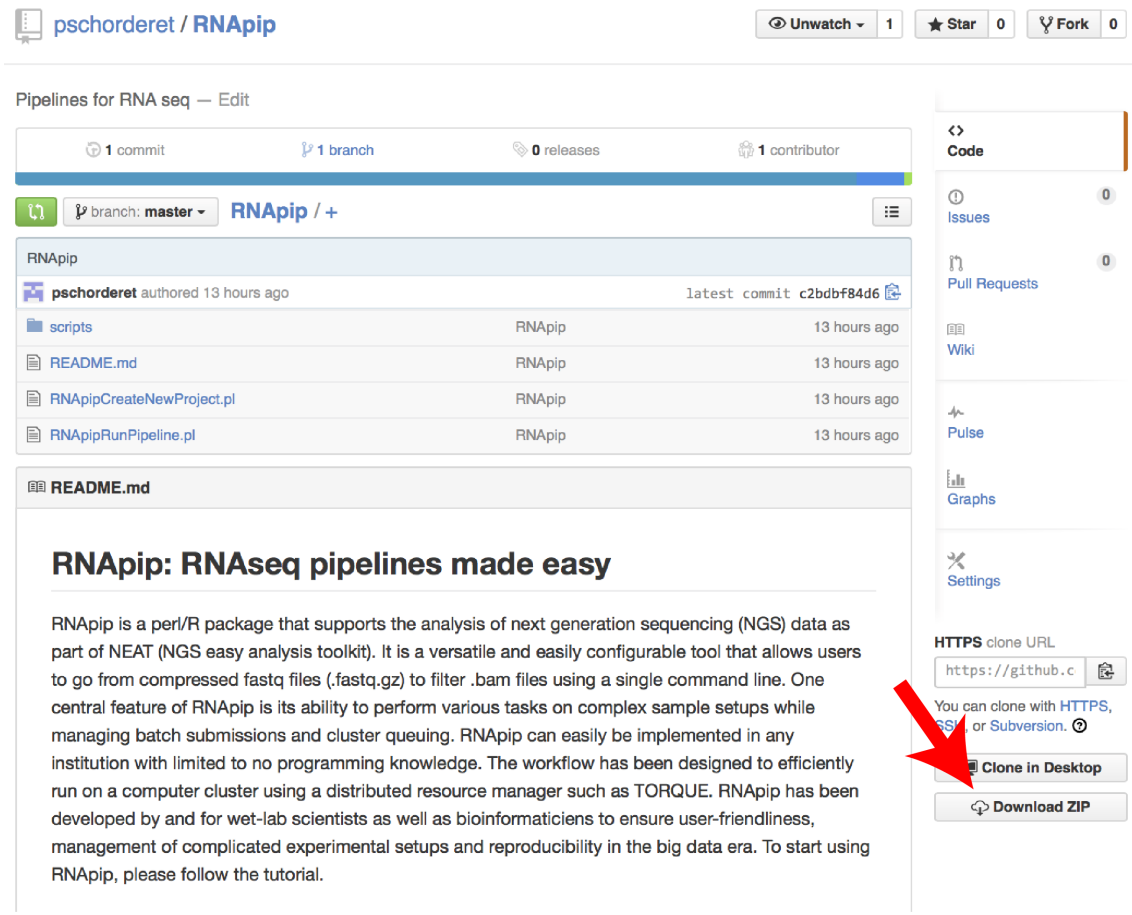
Remote servers and SSH

Secure Shell (SSH) is a cryptographic network protocol for secure data communication. In brief, it is a way for users to access remote computers (and their content) using a secure channel (a tunnel) through an insecure network (the internet). To access your computer cluster, you will need to establish an SSH connection. In analogy to an access card to your building, each user should be provided with an SSH username and password. Finally, the last essential parameter to access your computer cluster is the virtual *address* of the server. However, before accessing the remote server, you will need to copy the RNApip directory from your local computer to the remote server.

3. RNApip

Install RNApip

To install RNApip, download the RNApip repository from GitHub to any directory on your computer cluster.



As an example for the following tutorial, we will suppose the *RNApip/* directory was saved to the user's desktop on a local computer (*~/Desktop*) and that it will be saved to their */<HOME>/* directory on the remote server. In general, home directories can be accessed using the ~ sign. Make sure the folder is named *RNApip* and not *RNApip-master*. Start the transfer to the remote folder by typing the following command:

```
[~Desktop]$ rsync -avz ~/Desktop/RNApip username@serveradress.edu:~/
```

Enter your password and wait for the folders to transfer.

Once it has finished, the RNApip directory should be saved to your remote server. Check this by accessing your remote server. In a *Terminal* window, type the following:

[MY_COMPUTE~]$ ***ssh username@serveradress.edu***
password
[username@setrveradress ~]$

These commands bring you to your <HOME> directory on the remote server. For the following tutorial, we will suppose the name of this directory is */<HOME>/*. If you are not sure of your current working directory, type ***pwd*** to *print* your current *working directory*. Navigate to your */<HOME>/* directory using the ***cd*** and list files using the ***ls*** command. If RNApip was properly copied to your <HOME> directory, you should see the following:

[username@setrveradress ~]$ ***cd ~/RNApip***
[username@setrveradress RNApip]$ ***ls -l***

README.md
RNApipCreateNewProject.pl
RNApipRunPipeline.pl
scripts
Vignette

Note here that *./* is a short cut to your working directory. In this tutorial, because we have set our working directory to *~/RNApip/* by using the command ***cd ~/RNApip/***, *./* will be a short cut for *~/RNApip/* and these can/will be used interchangeably.

4. Test data

Structure and significance of test data

The unpublished test data provided for this tutorial is part of the RNApip folder and can be downloaded on GitHub. If you have followed the previous steps, the compressed fastq files required for the tutorial can be found in the test data folder */RNApip/scripts/testdata/*.

```
[RNApip]:$ ls -l ./scripts/testdata
1153158    PSa29-5_R1.fastq.gz
1158276    PSa29-6_R1.fastq.gz
```

The data consists of a 50 base pair, single end RNAseq experiment on cells under two growing conditions (noDox, Dox). No replicates were generated.

5. Running RNApip PART 1

Creating a new RNApip project

The first step to run RNApip is to create a new RNApip project. Navigate to the RNApip directory and run the *RNApipCreateNewProject.pl* script. We need to add ***perl*** in front of this command to tell the computer it should deal with this file as a perl script. This script requires the user to specify the path where the new RNApip project will be created including the name of the new project. In this example, we will create a project in the RNApip directory named EXAMPLE.

```
[RNApip]:$ perl RNApipCreateNewProject.pl ~/RNApip/EXAMPLE

*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-

Creating a new RNApip project

----------------------------------------------------

New RNApip project has been built as follows:

        ~/RNApip/EXAMPLE
              └─ DataStructure
                    └─ Targets.txt

------------------------------------------------------------------------------------------

IMPORTANT:    Fill in the Targets.txt file before running RNApipRunPipeline.pl
              To modify Targets.txt, copy paste the following command in your terminal:

              nano ~/RNApip/EXAMPLE/DataStructure/Targets.txt

*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-
```

Running the *RNApipCreateNewProject.pl* script should prompt the message above. If not, please troubleshoot before proceeding to the next step.

6. The Targets.txt file structure

Filling in the Targets.txt file

The Targets.txt file found in the *~/EXAMPLE/DataStructure/* directory is the backbone of RNApip. It contains all the information specific to your experiment and your computer cluster, including the names of files, the paths to the reference genomes, the steps to execute, the name of your samples, their relationships, etc. This file is the most important piece of RNApip and users are expected to invest the time to ensure all paths and parameters exist and are correctly set. However, once set, most of these parameters will not be changed on a specific computer cluster (users from a same institute will use the same paths). Therefor, we suggest modifying the *original* Targets.txt template file (see below).

All parameters of the Targets file should be self-explanatory. Here is a brief summary:

My_email	:	If users would like to be notified by emailed when the cluster has finished. This will only work if your computer cluster has activated the emailing feature (please check with system administrator). To ensure servers are not overwhelmed by email services, RNApip is configured in such a way as to notify users only if the pipeline has terminated properly (with no error). Users may change this parameter by modifying the QSUB_header.sh template file found in *./RNApip/scripts/*.
My_project_title	:	This is the name of the folder of your project on the remote server [automatically generated by RNApipCreateNewProject].
Reference_genome	:	The genome your data will be aligned to. Make sure your core faility has this genome reference installed on your cluster and that the extensions of the files are '.fa'.
Path_to_proj_folder	:	Full path to your project folder (without the project name) [automatically generated by RNApipCreateNewProject].
Path_to_RNApip	:	Full path to your RNApip folder. Note that in our example, we have created our project within the RNApip folder itself, but users can freely decide to create a dedicated folder for all of their RNApip projects.
Path_to_orifastq.gz	:	Full path to where your .fastq.gz files are. Usually, your sequencing core facility will let you know where they store these files. Note that all .fastq.gz files can be kept in a single location, they do not need to be copied to your folder.
Path_to_chrLens.dat	:	Path to a .dat file containing chromosome information for your reference genome. Refer to your computer core facility.
Path_to_RefGen.fa	:	Path to folder containing your reference genome files. Refer to your computer core facility.
Paired_end_run	:	"0" for single end sequencing. "1" for paired end sequencing.
Path_to_gtfFile	:	Path to the gtf file corresponding to your genome. Refer to your computer core facility.
Steps_to_execute	:	Users can choose from the following tasks: *unzip*, *qc*, *map* and *filter*. If you do not want to run all of these, simply delete them for the Targets.txt or rename them. Once ran, RNApip will change the value of these from '*unzip*' to '*unzip_DONE*'. Obviously, a certain hierarchy has to be followed, e.g. attempting to filter reads without having previously mapped them (in the same run or in a previous run) will not work. Note that '*qc*' requires Thomas Girke's systemPipeR package; map requires TopHat and

		filtering requires samtools. Refer below for exact requirements.
# Local	:	These parameters are only necessary for users who go on to use the NEAT tookit for metagene analysis, etc on their local computer. If you will not use this package, please disregard (leave as is).
TaxonDatabaseKG	:	Database of preferred feature such as known gene for RNAseq.
TaxonDatabaseDict	:	Idem

Please modify the Targets.txt file to your needs. The paths to the reference genomes should be obtained from your computer core facility (system administrator), as they are the ones usually maintaining these up to date. Moreover, the reference genome files should have a '*.fa*' extension (e.g. mm9.fa). Please check that your core has named these files accordingly as any other extension will lead the pipeline to abort prematurely. To modify the Targets.txt file, we suggest users get accustomed to using a terminal text editor such as ***vi*** or ***nano*** as it will avoid including spaces and special characters.

Fill in your Targets.txt fill using the following command:

[RNApip]:$ ***nano ./EXAMPLE/DataStructure/Targets.txt***

```
#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
#
# Project ID: "EXAMPLE"
#
#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
#
# Remote Server
#
# My_email        =         "your.email@harvard.edu"
#
# My_project_title        =         "EXAMPLE"
# Reference_genome        =         "mm9"
# Path_to_proj_folder     =         "~/RNApip"
# Path_to_RNApip          =         "~/RNApip"
# Path_to_orifastq.gz     =         "~/RNApip/fastq"
# Path_to_chrLens.dat     =         "~/.../reference_files/mm9/chr_lens.dat"
# Path_to_RefGen.fa       =         "~/.../mm9.fa"
# Paired_end_run          =         "0"
#
# Steps_to_execute        =         " unzip + qc + map + filter + peakcalling + cleanbigwig "
#
# Remove_from_bigwig      =         " random, chrM"
#
#
# Local
#
```

```
# TaxonDatabaseKG       =       "TxDb.Mmusculus.UCSC.mm9.knownGene"
# TaxonDatabaseDict     =       "org.Mm.eg.db"
#
#-------------------------------------------------------------------------------------------------------------
#
#
OriFileName   FileName                OriInpName  InpName  Factor  Replicate  FileShort  Experiment  Date
PSa36-1_R1    PSa36-1_noDox_K4me3     PSa37-3_R1  PSa37-3_noDox_Inp  K4me3    1  noDox  1  2015-01-01
PSa36-2_R1    PSa36-2_Dox_K4me3       PSa37-4_R1  PSa37-4_Dox_Inp    K4me3    2  Dox    1  2015-01-01
```

Once all information has been modified, hit ***cmd x*** to save the file. Confirm by hitting the ***y*** and ***enter***. This will save all changes to the file.

To avoid repeating these steps at each new RNApip project creation, we suggest you modify the *original* Targets.txt file that is used as template when creating a new RNApip project. Modify it using the same approach as above:

```
[RNApip]:$ nano ./scripts/NewChIPseqProject/DataStructure/Targets.txt
```

7. Running RNApip PART 2

Run the analysis

Once the Targets.txt file is correctly set up, the *RNApipRunPipeline.pl* script can be run. This script will execute the tasks specified in the *Targets.txt* file. Users can choose to perform the following tasks: *unzip, qc, map* and *filter*. If one or several tasks should not be run, simply discard them from the *Targets.txt* file under *# Steps to execute*.

As does the *RNApipCreateNewProject.pl*, the *RNApipRunPipeline.pl* script requires the user to specify the path to the RNApip project folder (users will obviously feed the same path to both scripts). In our example, the path is *./EXAMPLE*.

```
[RNApip]:$ perl RNApipRunPipeline.pl ~/RNApip/EXAMPLE

##########################################################################
#                                                                        #
#                        RNAseq pipeline v1.0.1 (Jan 2015)               #
#                                                                        #
##########################################################################

My email:               your.email@harvard.edu

expFolder:              EXAMPLE
genome:                 mm9
userFolder:             ~/
path2RNApip:            ~/RNApip
path2expFolder:         ~/RNApip/EXAMPLE
path2fastq.gz:          ~/RNApip/testdata/
Targets:                ~/RNApip/EXAMPLE/DataStructure/Targets.txt
chrlens:                ~/…/reference_files/mm9/chr_lens.dat
refGenome:              ~/…/mm9.fa

Paired end sequencing:      0
Bwa.command.line:           bwa aln -n $ndiff
Remove pcr dupl:            1
Make unique reads:          1
PeakCaller.fdr:             0.01

.....
Performing following tasks:
.....
unzip:                  TRUE
map:                    TRUE
qc:                     TRUE
filter:                 TRUE
peakcalling:            TRUE
cleanbigwig:            TRUE        (remove: random chrM)
.....
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Exiting INITIAL section with no known error

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

This will launch the pipeline and will prompt a summary of the user's parameters. RNApip automatically manages all creations and batch submissions of jobs, dependencies, ordering of files, queuing, etc. If the cluster is using TORQUE, the

processes can be followed by the ***qstat*** command (type ***qstat*** in your terminal). Briefly, *Q* stands for queuing, *R* for running, *E* for exiting and *H* for holding.

From this step on, the user will NOT need to intervene further. The pipeline is completely automated.

Note: This only concerns users following the tutorial using the provided test data. We have experienced corruption issues when fastq.gz files are downloaded from Github, leading RNApip to prematurely terminate. This however does not corrupt any other files or the pipeline. Once the unzip jobs are done, RNApip will stop. Simply re-start it replicating the same command line (***perl RNApipRunPipeline.pl ~/EXAMPLE***). There is no need to change anything else. If the problem persists, please contact us.

Once the pipeline has finished, it will notify the user of its status by email. The first step is to check whether everything ran smoothly. To this end, please open the Targets.txt file and check whether all jobs are marked as *DONE* under the *# Steps to execute* tag. If not, please follow the 'Troubleshooting' section below.

The mock data provided as a test example should take no more than one hour to run, usually a lot less.

8. RNApip workflow

Unzipping and renaming fastq files

Using the *Targets.txt* file, RNApip will unzip fastq.gz files found under '*OriFileName*' /'*OriInpName*' and will rename them to names found under '*FileName*'/'*InpName*'. RNAseq experiments usually do not require inputs, so these entries will be left empty ('-'). RNApip will use the virtual path to the compressed files and will save the unzipped fastq files in the project folder. This minimizes file transfer and ensures all original fastq files can be kept in a central directory.

Quality control (QC)

Quality control of fastq files uses the elegant systemPipeR package developed by Thomas Girke (Girke, 2014).

Mapping and filtering reads

RNApip utilizes TopHat (REFERENCE), a well-established, splice aware and commonly accepted algorithm for RNAseq data. The most common parameters can be changed in the *AdvancedSettings.txt* file (*/EXAMPLE/DataStructure/AdvancedSettings.txt*).

9. Generated data and next steps

Architecture

RNApip will generate many files of which the majority will not be used in further analysis. We should note that the aligned, filter .bam files are stored in */EXAMPLE/Tophat/<sample>/<sample>.bam*. Quality reports (if applicable) are found in */EXAMPLE/QC/*.

RNAmE

Although some users may prefer to take over the analysis from this step, we suggest using RNAmE. RNAmE is part of the NEAT toolkit and has been developed as a downstream module for RNApip. RNAmE accompanies users from an RNApip output to differential gene expression (DEG) analysis in as few as two double clicks. It automatically transfers files from remote server to local computer to create wet-lab scientist readable data including pdf smear graphs, Venn diagrams (overlap of DEG), count tables and RPKM tables. Users interested in such analysis can download the NEAT package from GitHUB and follow the tutorial.

10. Troubleshooting

Output and error files

RNApip is broken down into distinct job sections. For example, all files corresponding to the 'map' section can be found in the scripts folder (*/<HOME>/RNApip/EXAMPLE/scripts/map/*). In each job folder, the qsub directory contains all output (.o.jobID) and error (.e.jobID) files for individual jobs, which makes it easy to troubleshoot any possible errors. Files are named as follows:

```
[RNApip]:$ ls -l ./EXAMPLE/scripts/map/
   1018   map.sh
    759   PSa29-5_noDox_RNA_map.sh
    749   PSa29-6_noDox_RNA_map.sh
    ...
   4096   qsub

[RNApip]:$ ls –l ./EXAMPLE/scripts/map/qsub
      0   Iterate_map.o<jobID>
      0   Iterate_map.e<jobID>
    354   PSa36-1_noDox_K4me3_map.sh.e<jobID>
      0   PSa36-1_noDox_K4me3_map.sh.o<jobID>
```

Most .o.jobID and .e.jobID qsub files should be empty. Exceptions to this are the map.e.jobID and the filter.e.jobID files, which contain terminal outputs. Most of these can be disregarded.

In the scenario where RNApip cannot proceed to all jobs, it will stop. Users can modify the email settings in the QSUB_header.sh to be notified in case of an error (refer to the Advanced settings section below). Users can follow up which jobs induced the stop by looking at the Targets.txt file. The last <job>_DONE is the <job> that induced the premature stop.

Example

As an example, lets suppose RNApip crashed while analyzing the test data. Troubleshoot the error by looking at the Targets.txt file:

[RNApip]:$ ***less*** **./EXAMPLE/DataStructure/Targets.txt**

```
#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
#
# Project ID: "EXAMPLE"
#
#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
#
# Remote Server
#
# My_email        =        "your.email@harvard.edu"
#
# My_project_title        =        "EXAMPLE"
# Reference_genome        =        "mm9"
# Path_to_proj_folder     =        "/<HOME>/RNApip"
# Path_to_RNApip          =        "/<HOME>/RNApip"
# Path_to_orifastq.gz     =        "/<HOME>/RNApip/fastq"
# Path_to_chrLens.dat     =        "/<HOME>/.../reference_files/mm9/chr_lens.dat"
# Path_to_RefGen.fa       =        "/<HOME>/.../"
# Paired_end_run          =        "0"
#
# Steps_to_execute        =        " unzip_DONE+ qc_DONE + map_DONE + filter "
#
# Remove_from_bigwig      =        " random, chrM"
#
#
# Local
#
# TaxonDatabaseKG         =        "TxDb.Mmusculus.UCSC.mm9.knownGene"
# TaxonDatabaseDict       =        "org.Mm.eg.db"
#
#-------------------------------------------------------------------------------------------------------------$
#
#
OriFileName  FileName  OriInpName  InpName  Factor  Replicate  FileShort  Experiment  Date
PSa29-5_R1   PSa29-5_noDox_RNA   -   -   RNA   1   noDox   1   2015-01-01
PSa29-6_R1   PSa29-6_Dox_RNA     -   -   RNA   2   Dox     1   2015-01-01
```

The last <job>_DONE is the map_DONE, which indicates the source of the error. Troubleshoot the origin by looking at the qsub error files corresponding to the 'map' section.

```
[RNApip]:$ ls -l ./scripts/map/qsub

790     PSa29-5_noDox_RNA_map.sh.e<jobID>
  0     PSa29-5_noDox_RNA_map.sh.o<jobID>
```

The .e.<jobID> file is a lot bigger than usual. Use the unix ***less*** command to open it in a read mode. Error messages should be self-explanatory. To exit, hit ***Enter***. In our example, we have the following error message:

```
[RNApip]:$ less ./EXAMPLE/scripts/map/qsub/

[bwt_restore_bwt] fail to open file '/data/ref/mm9/bwa/mm9.fa.bwt'. Abort!
```

RNApip tells you it could not open the '/data/ref/mm9/bwa/mm9.fa.bwt' file. Check the existence of the file in this path. Once the source of the error is determined, modify the Targets.txt file accordingly and move on.

11. Advanced settings

AdvancedSettings.txt

Users can modify advanced settings (map, filter, etc) in the AdvanceSettings.txt file found in the DataStructure directory.

Would users decide to modify the Tophat aligner command (# *Tophat.command.line*), … .

QSUB parameters

The qub header can be modified to meet the requirements of specific clusters, including queuing times, nodes, number of CPUs, etc. If this is of interest, please modify the QSUB_header.sh template file in *./RNApip/scripts/QSUB_header.sh* such as to apply personalized settings to all jobs (this needs to be only once). Please refer to your computer core facility systems administrator for further details.

12. Version information and required packages

Program: bwa (alignment via Burrows-Wheeler transformation)
Version: 0.5.9-r16

Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.18 (r982:295)

R version 3.1.0 (2014-04-10)
Platform: x86_64-redhat-linux-gnu (64-bit)

R packages:

- SPP (spp_1.11)
- systemPipeR (systemPipeR_0.99.0)
- ...

13. Reported bugs

QC

QC does not work on test data due to a corruption in fastq.gz files during transfer. This should not affect QC for your own samples.

14. Funding

This pipeline was developed with funding from the Swiss National Science Foundation.

15. References

- …