2013

RNA CoMPASS Manual v1.1

Computational Biology Group
Department of Computer Science
University of New Orleans

RNA CoMPASS, a web-based GUI distributed computational pipeline, provides all-in-one functionality including human transcriptome quantification and the typical endogenous RNA-Sequencing analysis along with the investigation of exogenous sequences. RNA CoMPASS is deployable on either a local cluster or a grid environment managed by PBS submission.

Guorong Xu
5/24/2013

RNA CoMPASS Manual

Abstract

Welcome to the *RNA CoMPASS* Manual. Here you will find information on how to install and configure the application. It is a step-by-step, task-oriented guide for configuring *RNA CoMPASS* on your system.

License

Table of Contents

Chapter 1 Introduction

Welcome to the *RNA CoMPASS* Manual.

Here you can find information on how to install and configure RNA CoMPASS. It is a step-by-step, task-oriented guide for configuring *RNA CoMPASS* on your system.

This manual assumes you have a basic understanding of your operating system. Some installation details are covered in Chapter 2: Installation. If you need detailed instructions on using *RNA CoMPASS*, please refer to Chapter 3: Usages.

RNA CoMPASS is hosted by sourceforge and homepage is:

http://RNACoMPASS.sourceforge.net/

Chapter 2 Installation

This chapter provides a quick overview on installing *RNA CoMPASS*.

2.1. Preparing to Install

This section explains RNA CoMPASS's requirements.

3.1.1 System Requirements

Recommended Memory: 8GB+ RAM
Minimum Memory: 4GB RAM
OS: Mac OS X, Linux Ubuntu 11.04

3.1.2 Running Environment

A recent version of the Java Runtime Environment (JRE) is needed prior to using RNA CoMPASS.

For Windows this involves a quick download from the following Oracle site:
http://www.java.com/en/download/manual.jsp

Java is standard on Mac OS X.

2.2 Installing RNA CoMPASS

- Download the zip file *RNACompass.war* to your local disk,
- Copy *RNACompass.war* into the Host appBase,
  For example:
  /apache-tomcat-7.0.14/webapps/
- Copy the folder *properties* to the web Host bin directory and change the access permissions of all files to executable using command chmod,
  For example:
  cd /apache-tomcat-7.0.14/bin/properties/
  chmod 777 *.*
- Create an empty folder under the Host bin directory named by *uploaded*,
  For example:
  mkdir /apache-tomcat-7.0.14/bin/uploaded/
- Launch web application server and access an instance running on localhost,
  For example:
  http://localhost:8080/RNACompass/

2.2 Installing JPPF Grid

A JPPF Grid is made of a workhorse server, to which any numbers of execution nodes are attached. A node is a JPPF software component that is generally installed and running on a separate machine. This is commonly called a master/slave architecture, where the work is distributed by the server to the nodes.

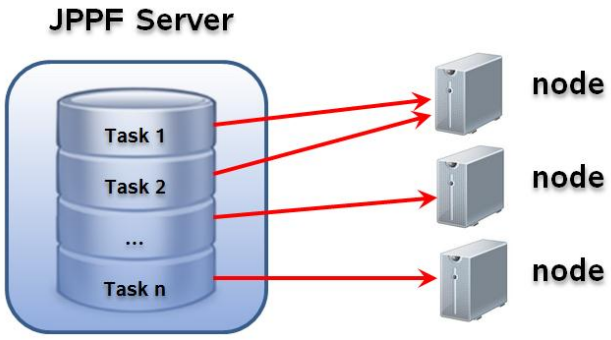

Figure 2.2-1 JPPF framework

2.2.1 To install JPPF server/driver distribution on your workhorse server

- Go to the JPPF homepage to download the JPPF server/driver distribution on your workhorse server,
- Unzip the JPPF-3.0-driver.zip file to your workspace directory,
- Open a terminal and go to the JPPF-3.0-driver folder,
- Type command "ant" to launch the JPPF driver on your server.

Please note: You need to configure the IP address of the workhorse server and the number of node machines in the property file System.properties under the Host bin/properties directory.
For example:

```
#########################################################
### The configuration parameters for JPPF framework ###
#########################################################
ftp_HostName=137.30.125.103
ftp_Port=12222
jppfMachineNum=2
```

Figure 2.2.1-1 Configure host name/IP address

2.2.2 To install JPPF administration and monitoring console on your workhorse server

- Go to the JPPF homepage to download the JPPF administration and monitoring console on your workhorse server,
- Unzip the JPPF-3.0-admin-ui.zip file to your workspace directory,
- Open a terminal and go to the JPPF-3.0-admin-ui folder,
- Type command "ant" to launch the JPPF admin-ui console on your server.

2.2.3 To install JPPF node on your other machines

- Go to the JPPF homepage to download the JPPF node distribution on your other machines,
- Unzip the JPPF-3.0-node.zip file to your workspace directory,
- Open a terminal and go to the JPPF-3.0-node folder,
- Copy the folder ScriptFiles to JPPF-3.0-node and change the access permissions of all files to executable using command chmod,
  For example:
  cd /JPPF-3.0-node/ScriptFiles/
  chmod 777 *.*
- Type command "ant" to launch the JPPF node on your other machines.

Please note: You need to configure the path of bowtie and Novoalign index files, and the path of BLAST database files in the property file Node.properties under the ScriptFiles directory.
For example:

```
bowtieIndexFile=/media/OS/Human_data/hg19/hgChrAll
novoalignIndexFile_human=/media/OS/Human_data/hg19/hgChrAll.ndx
novoalignIndexFile_bacteria=/media/OS/Human_data/hg19/bacteriaPart.ndx
novoalignIndexFile_virus=/media/OS/Human_data/hg19/hgChrAll.ndx
ntDB=/usr/share/nt/nt
humanDB=/home/guest1/Desktop/JPPF/JPPF-2.5.3-node/human_database/HumanRNA
```

Figure 2.2.3-1 Configure the number of JPPF nodes

2.3 Installing a Grid Environment Managed by PBS Submission

RNA CoMPASS allows users to deploy the pipeline on a computer cluster managed by PBS (Portable Batch System) submission. To install on the computer cluster, users need to copy novoalign and blast folders to a specific folder on the cluster.

2.3.1 To install Novoalign on the cluster

- Download novoalign folder under *PBS_Cluster* folder at the RNA CoMPASS website,
- Then build Novoalign index files and save to index folder under novoalign folder,
- Change the access permission of scripts file under scripts folder,
  For example:
  `chmod 777 *.*`

2.3.2 To install BLAST on the cluster

- Download blast folder under *PBS_Cluster* folder at the RNA CoMPASS website,
- Copy NT database to nt folder under blast folder,
- Download human reference sequence file (human.rna.fna.gz) from the FTP site ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/
- Run the command to make blast database files:
  `makeblastdb -taxid 9606 -in human.rna.fna -out HumanRNA -dbtype nucl`
- Save the database files to HumanRNA folder under blast folder,
- Change the access permission of scripts file under scripts folder,
  For example:
  `chmod 777 *.*`

2.3.3 To configure parameters for cluster

- Open the property file System.properties under the Host bin/properties directory.
- Change property useCluster from false to true as Figure 2.3.3-1,
- Configure the cluster host name,
- Configure the cluster user name,
- Configure the cluster user password,
- Configure the Novoalign current folder on the cluster,
- Configure the Novoalign data folder on the cluster,
- Configure the BLAST current folder on the cluster,
- Configure the BLAST data folder on the cluster,

```
#########################################################
### The configuration parameters for PBS Cluster ###
#########################################################
useCluster=true
cluterHostName=ccs1.ccs.tulane.edu
clusterUserName=
clusterPasswd=
clusterMachineNum=2
BlastCurrentDir=/scratch03/gxu2/blast/
NovoalignCurrentDir=/scratch03/gxu2/novoalign/
Cluster_novoalignDataFolder=/scratch03/gxu2/novoalign/data/
Cluster_blastDataFolder=/scratch03/gxu2/blast/data/
```

Figure 2.3.3-1 Configure parameters for the PBS cluster

2.3.4 To modify the path of scripts for cluster

- Open the scripts folder under novoalign folder and blast folder,
- Open the shell script files novalign_single.sh, novoalign_paired.sh and blast.sh,
- Modify the path of scripts folder and executable files as the Figure 2.3.4-1.

```
#!/bin/bash
#PBS -q ccs_short
#PBS -l walltime=06:00:00
#PBS -N novoalign
#PBS -j oe
#PBS -l nodes=1:ppn=4
#PBS -d /scratch03/gxu2/novoalign/scripts
echo "index file=$index_file"
echo "fastq quality=$fastq_quality"
echo "input file=$input_file"
echo "mate file=$mate_file"
echo "output file=$output_file"

/scratch03/gxu2/novoalign/software/novoalign -d $index_file -F $fastq_quality -f $ir
```

Figure 2.3.4-1 Configure shell script files for the cluster

Chapter 3 Usages

This chapter provides a detailed guide of using *RNA CoMPASS*.

3.1. User Management

3.1.1 To create an FTP user account

You can input an FTP user name and password for transferring data to server. Clicking the Check Availability button to check if the user account exists, the new user account will be created if user name does not exist.

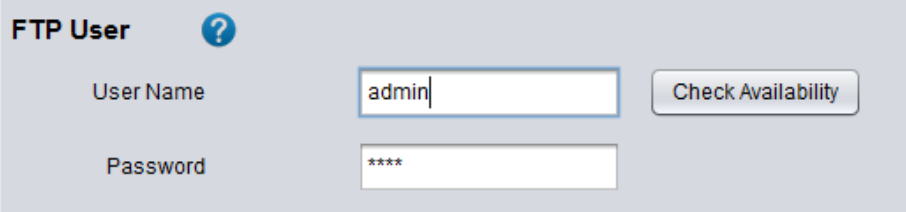

Figure 3.1.1-1 FTP user management

3.2. File Management

You can add/remove the sequence files in FASTQ format and an annotation file to the web application server.

3.3.1 To choose files or choose folder

Users can choose the files or folder using any one of these methods,

- To choose a consecutive group of files in the tab Choose Files, click the first item, hold down the SHIFT key, and then click the last item,
- Or choose nonconsecutive files in the tab Choose Files, hold down CTRL, and then click each item you want to select,
- Or open a folder that contains the files you want to choose in the tab Choose Folder,
- Then press the Upload button to transfer the files to web application server.

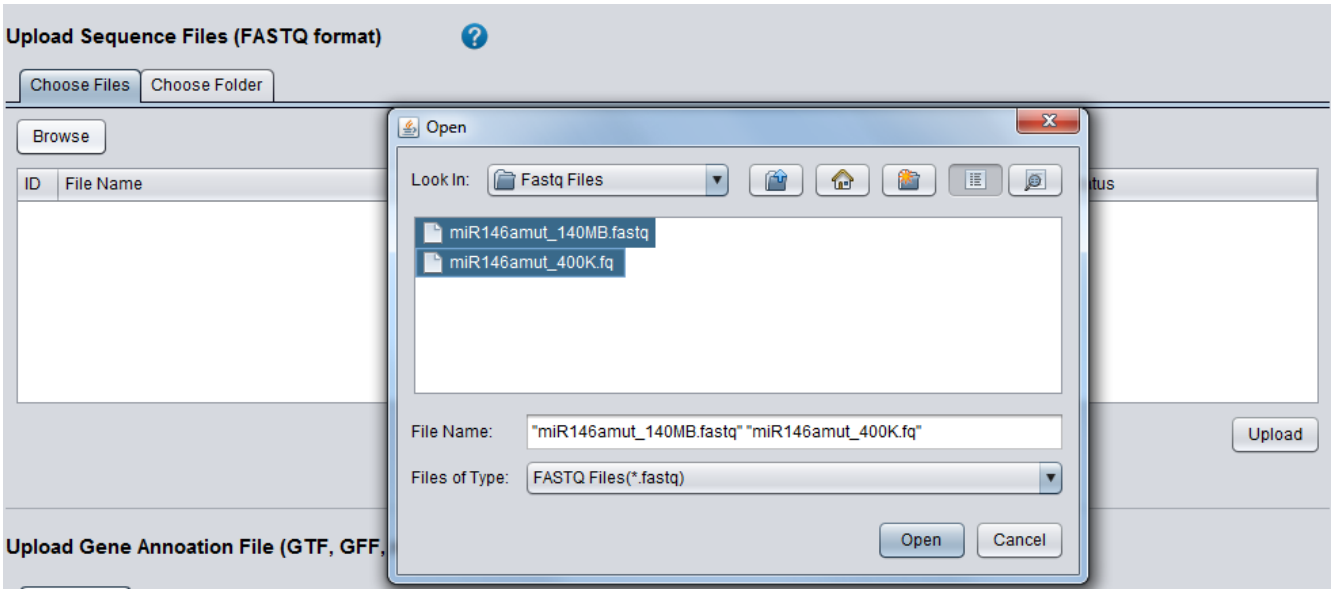

Figure 3.2.1-1 Open directory dialogue and choose files

3.3.2 To remove files

Users can remove files from the web application server.

- To select a consecutive group of files, click the first item, hold down the SHIFT key, and then click the last item,

- Right click to select the Delete to remove the selected files

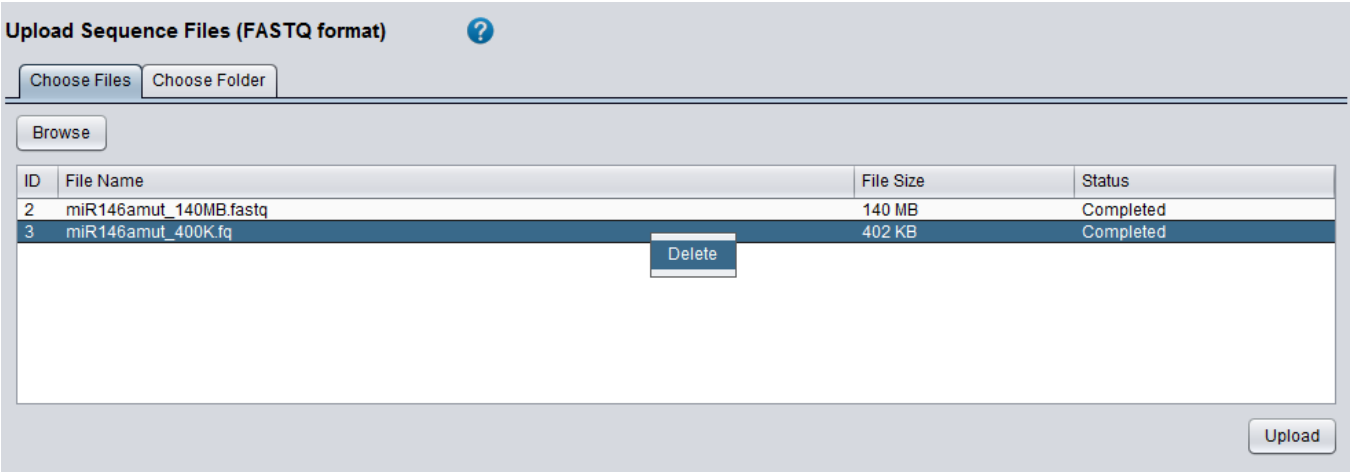

Figure 3.2.2-1 Remove files from the web application server

3.3. Parameters Configurations

You can configure the appropriate parameters for RNA CoMPASS to process.

3.3.1 To check sequence file format and quality scale

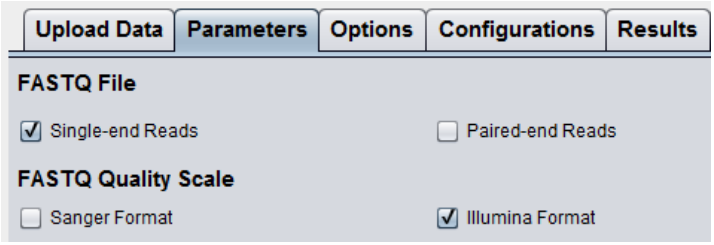

Figure 3.3.1-1 Parameters for sequence file format and quality scale

3.3.2 To configure pipeline parameters

You can configure the appropriate parameters for RNA CoMPASS to process.

For example, when you check the option of Deduplication FASTQ file, RNA CoMPASS will remove the duplicate reads from your FASTQ files, the first 50 bases of reads are used for deduplication by default.

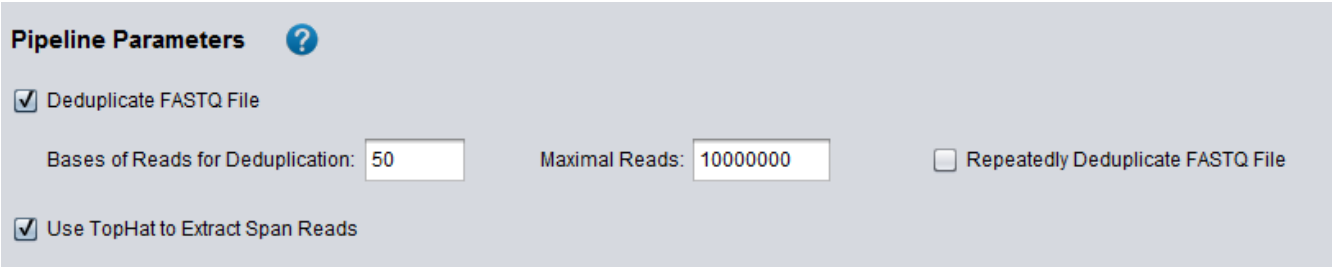

Figure 3.3.2-1 Parameters for sequence file format and quality scale

3.3.3 To group sequence files

You can separate the sequence files into two groups if you enable using edgeR to detect differentially expressed genes and isoforms.

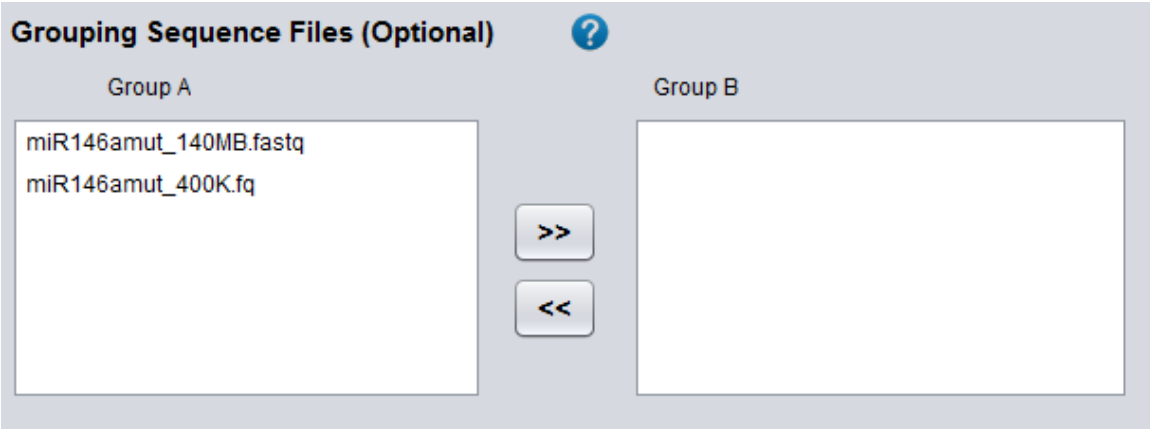

Figure 3.3.3-1 Grouping sequence files

3.3.4 To select pre-built indexes

You can select at least one Novoalign pre-built indexes to align your sequence data by clicking add >> button or remove the indexes from selected indexes list by clicking remove << button.

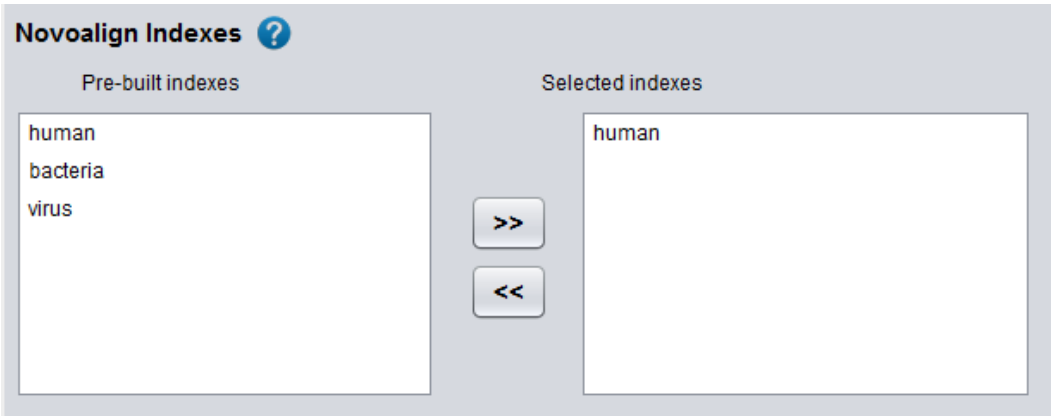

Figure 3.3.3-1 Grouping sequence files

3.3.5 To check options of estimation of transcripts abundance

You can check one of methods to estimate transcripts abundance, One-step iQuant, Iterative iQuant and RAEM procedure. Please note that the options of estimation of transcripts abundance are optional.

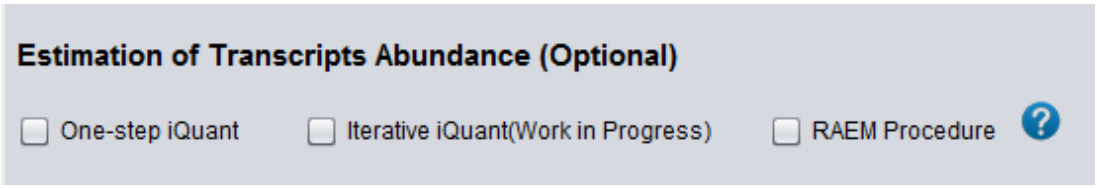

Figure 3.3.5-1 Estimation of transcripts abundance

3.3.6 To check key features for downstream analysis of RNA-seq technology

RNA CoMPASS provides some key features for downstream analysis of RNA-seq technology.

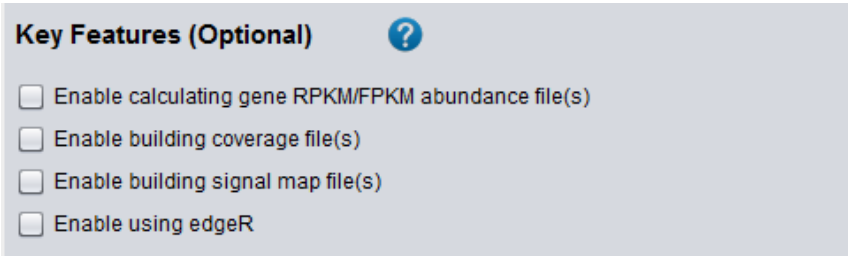

Figure 3.3.6-1 Key features

3.3.7 Customizing the chromosome names

Between the genome annotation file and the RNA-seq data file, the chromosome names are often mismatched due to different databases and/or aligners. To remedy this situation, RNA CoMPASS allows user to customize the relationship map between different chromosome names allowing the system to automatically map the customized chromosome names during calculations.

- Users can define the mapping relationship of chromosome names between different versions.

  For example, by adding add the line:

  gi|89161185|ref|NC_000001.9|NC_000001 chr1

  The string "gi|89161185|ref|NC_000001.9|NC_000001" with "chr1" in the output files will be automatically replaced.

- In the Chromosome Name Mapping table, right click on the desired names, and then left click Delete Selection.

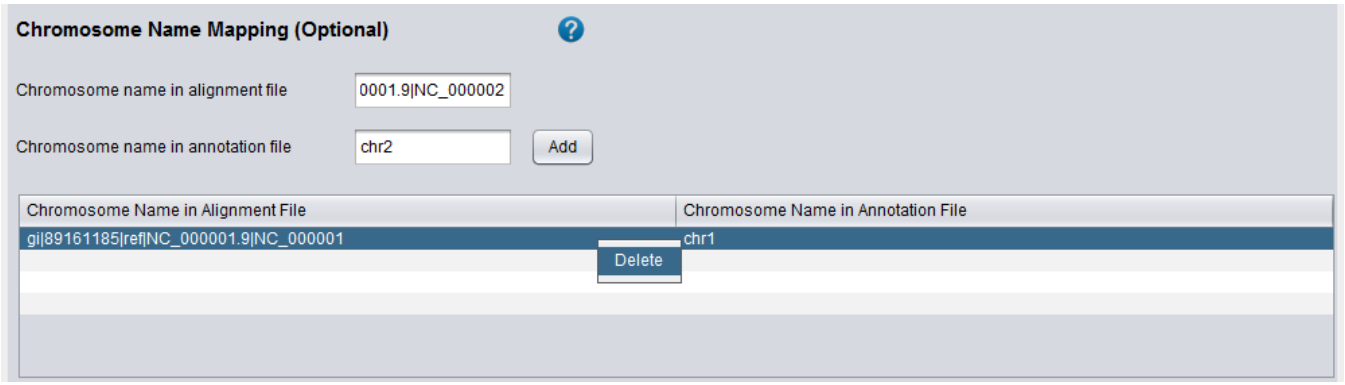

Figure 3.3.7-1 Chromosome name mapping

3.3.8 Customizing the signal map intervals

RNA CoMPASS allows user to customize the signal map intervals to generate the base-wise signal map information that fall between the customized the intervals for peak detection.

- Users can customize the signal map intervals by adding the chromosome name, start position and end position.
- In the Signal Map table, right click on the desired interval, and then left click Delete to remove selected intervals.

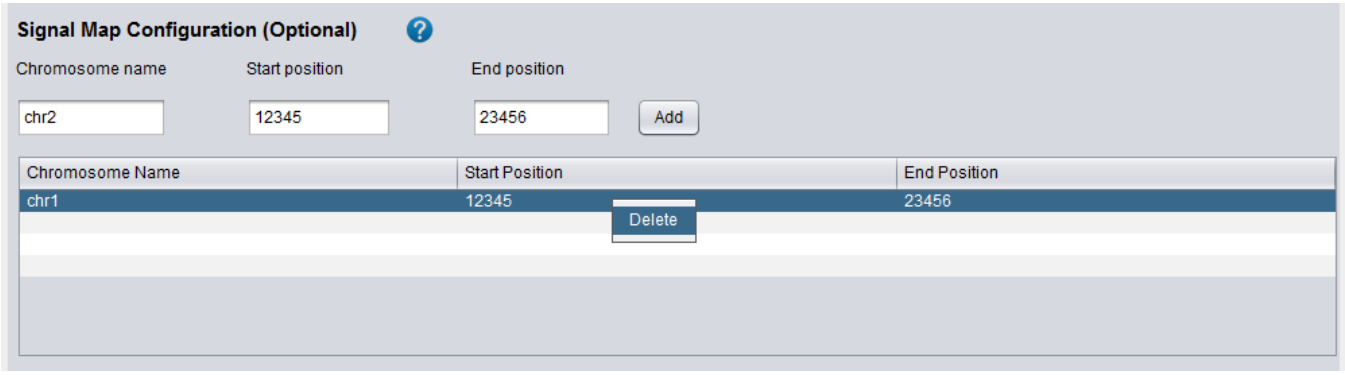

Figure 3.3.8-1 Signal map configuration

3.4. Resulting Files Management

You can download the resulting files from web application server and delete the files from server if the Delete all files from server after downloading check box is checked. You can also view the log file if the View log file button is pressed.

3.4.1 To manage your resulting files

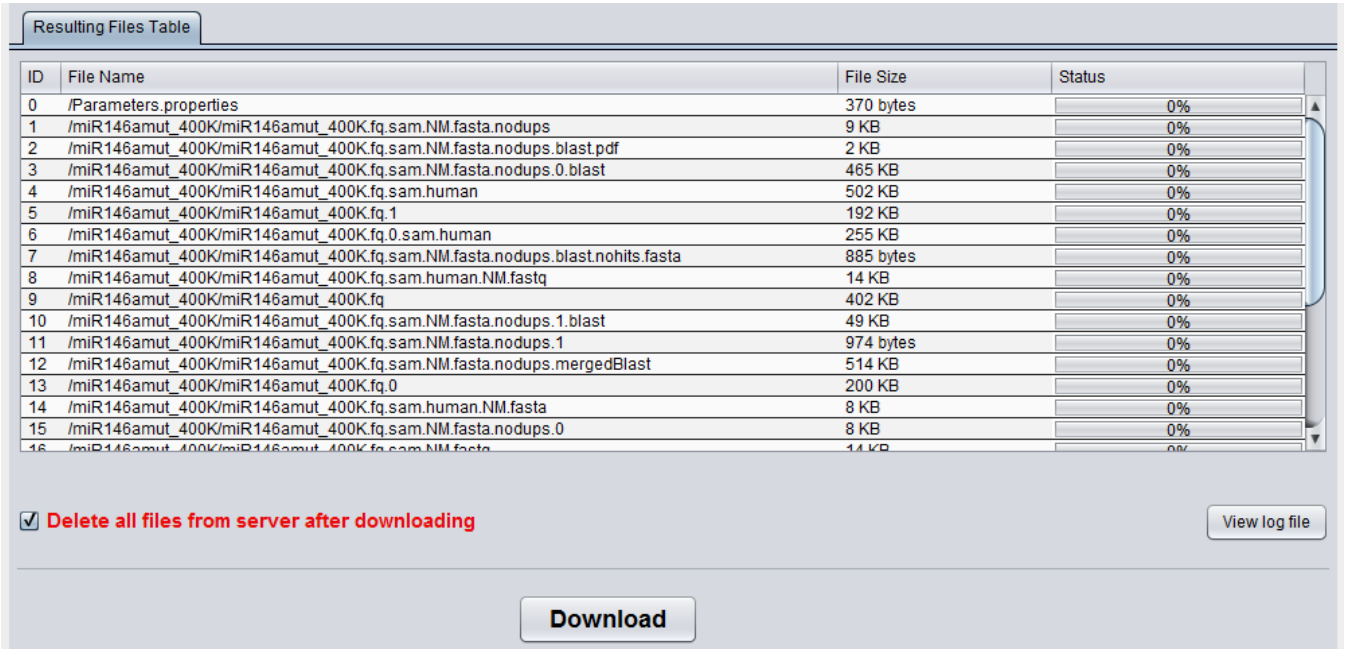

Figure 3.4.1-1 Resulting files table

Chapter 4 Appendix

This chapter provides a detailed installation guide of existing Bioinformatics tools used in RNA CoMPASS.

4.1. Installing Existing Bioinformatics Tools on Computers

4.1.1 To install Novoalign on JPPF node machines

- Download Novoalign executable file at the below website, http://www.novocraft.com/main/downloadpage.php
- Unzip the file and copy novoalign and novoindex executable files to /usr/bin/,
- Build the Novoalign index files using novoindex command and configure the paths in the System.properties file,

```
# After building new novoalign index, pleas add the index n
# then append a new line of the path of novoalign index to
novoalignIndexes_list=human,bacteria,virus,
novoalignIndexFile_human=hChrAll.ndx
novoalignIndexFile_bacteria=hChrAll.ndx
novoalignIndexFile_virus=hChrAll.ndx
```

Figure 4.1.1-1 Configure parameters for RNA CoMPASS pipeline

4.1.2 To install Bowtie on workhorse server

- Download bowtie executable file at the below website, http://bowtie-bio.sourceforge.net/index.shtml
- Unzip the file and copy bowtie executable files to /usr/bin/,
- Download the Bowtie index files and configure the path in the System.properties file,

```
############################################################
### The configuration parameters for RNA CoMPASS pipeline ###
############################################################
giTaxID=/media/Resource/Workspace/human_data/gi_taxid_nucl.dmp
bowtieIndexFile=/media/Resource/Workspace/human_data/hg19/hgChrAll
```

Figure 4.1.2-1 Configure parameters for RNA CoMPASS pipeline

4.1.3 To install TopHat on workhorse server

- Download TopHat executable file at the below website, http://tophat.cbcb.umd.edu/
- Unzip the file and copy tophat executable file to /usr/bin/,

4.1.4 To install SAMtools on workhorse server

- Download SAMtools source files at the below website, http://sourceforge.net/projects/samtools/files/samtools/

- Unzip the file and run make command to build samtools executable file,
- Copy samtools executable file to /usr/bin/,

4.1.5 To install BLAST+ on JPPF node machines

- Download the installation package ncbi-blast*+-src.tar.gz at the below FTP website,
  ftp.ncbi.nlm.nih.gov/blast/executables/blast+
- Unzip the package using the command,
  tar -xzf ncbi-blast*+-src.tar.gz -C /usr/bin
- Open a terminal and type the following command,
  cd /usr/bin/ncbi-*/c++
- Configure and make executable files by using the following commands,
  ./configure --without-debug --with-mt --with-build-root=ReleaseMT;
  cd ReleaseMT/build;
  make all_r;
- Creates links to the following files
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/blastdb_aliastool /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/dustmasker /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/rpstblastn /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/blastdbcheck /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/gene_info_reader /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/segmasker /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/blastdbcmd /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/gumbelparams /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/seqdb_demo /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/blast_formatter /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/srsearch /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/blastn /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/makeblastdb /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/tblastn /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/blastp /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/makembindex /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/tblastx /usr/bin;
  ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/blastx /usr/bin;

```
ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/project_tree_builder /usr/bin;
ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/convert2blastmask /usr/bin;
ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/psiblast /usr/bin;
ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/windowmasker /usr/bin;
ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/datatool /usr/bin;
ln -s /usr/bin/ncbi-blast*/c++/ReleaseMT/bin/rpsblast /usr/bin;
```

4.1.6 To install NT database on JPPF node machines

- Open a terminal and go to,
  `cd /usr/bin/ncbi-blast*/c++/src/app/blast/`
- Then run the command,
  `update_blastdb.pl nt;`
- Unzip all nt files,
  `tar -xzf nt.**.tar.gz`

4.1.7 To install human refseq database on JPPF node machines

- Download human reference sequence file (human.rna.fna.gz) from the FTP site ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/
- Run the command to make blast database files:
  `makeblastdb -taxid 9606 -in human.rna.fna -out HumanRNA -dbtype nucl`

4.1.8 Install gi-taxid_nucl library on workhorse server

- Download gi-taxid_nucl library at the following FTP website,
  ftp://ftp.ncbi.nih.gov/pub/taxonomy/gi_taxid_nucl.zip
- Unzip the zip file to /usr/share using the following command,
  `unzip -d /usr/share gi_taxid_nucl.zip`
- Configure the path in the System.properties file,

```
############################################################
### The configuration parameters for RNA CoMPASS pipeline ###
############################################################
giTaxID=/media/Resource/Workspace/human_data/gi_taxid_nucl.dmp
bowtieIndexFile=/media/Resource/Workspace/human_data/hg19/hgChrAll
```

Figure 4.1.7-1 Configure parameters for RNA CoMPASS pipeline

4.1.9 Install MEGAN on workhorse server

- Download the installation package MEGAN_unix_4_50_6.sh at the website, http://ab.inf.uni-tuebingen.de/software/megan/
- Change the access permission of the shell executable file,
  `chmod +x MEGAN_unix_4_50_6.sh`
- Run the following command to install MEGAN,
  `./MEGAN_unix_4_50_6.sh`