

rnaQUAST 1.1.0 using Atmosphere

Rationale and background:

Overview of rnaQUAST

rnaQUAST is a tool for evaluating RNA-Seq assemblies using reference genome and gene data database. In addition, rnaQUAST is also capable of estimating gene database coverage by raw reads and *de novo* quality assessment using third-party softwares (STAR, TopHat, GMAP etc.). The detailed manual is available here - <http://spades.bioinf.spbau.ru/rnaquast/release1.1.0/manual.html>

Needs to run rnaQUAST on Atmosphere

1. Atmosphere requirements
 - a. CyVerse username that has an institutional email (e.g. janedoe@email.arizona.edu)
 - b. Atmosphere allocation
 - i. Follow section called "Adding apps and services to your account"
 - ii. Fill in the request sheet with the allocation needed
 1. Include the number of Atmosphere Units (AUs) that you will need
 2. It can be difficult to know how AUs many are needed *a priori*,
 - a. $(\text{number of cores}) \times (\text{real time hours}) \times (\text{days you need to run the instance}) = \text{AUs needed in a month}$
2. Computational Knowledge
 - a. Familiarity with the terminal/shell



Learn about allocations

Learn about CyVerse's allocation policies [here](#).

Protocol: How to use rnaQUAST 1.1.0 on Atmosphere?

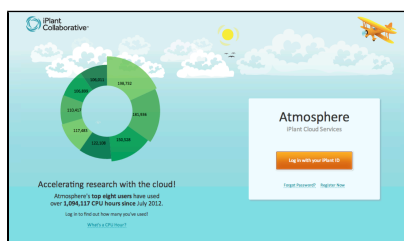
This tutorial will take users through the steps of:

1. Launching the rnaQUAST 1.1.0 Atmosphere image
2. Running rnaQUAST 1.1.0 on a test data

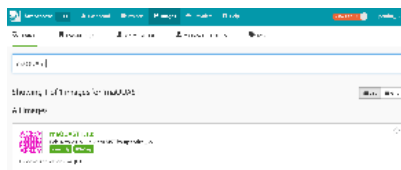
Please work through the tutorial and add your comments on the bottom of this page. Or send comments per email to upendra@cyverse.org. Thank you

Part 1: Connect to an instance of an Atmosphere Image (Virtual Machine)

Step 1. Go to <https://atmo.iplantcollaborative.org> and log in with your CyVerse credentials.



Step 2. Click on the **Launch New Instance** button and search for **rnaQUAST 1.1.0**.



Step 3. Select the image **rnaQUAST 1.1.0** and click **Launch Instance**. It will take 10-15 minutes for the cloud instance to be launched.



Note: Instances can be configured for different amounts of CPU, memory, and storage depending on user needs. This tutorial can be accomplished with the medium instance size, **medium1 (4 CPUs, 8 GB memory, 80 GB root)**

Part 2: Set up a rnaQUAST 1.1.0 run on a test data using the Terminal window

Step 1. Open the **Terminal**. Add the ssh details along with your IP address to connect the instance through the terminal

```
$ ssh <username>@Ipaddress
```

step 2. You will find rnaQUAST v1.1.0 software in "/opt" folder. All the dependencies for running rnaQUAST v1.1.0 are located in "/opt/rnaQUAST-1.1.0"

```
$ cd /opt/rnaQUAST-1.1.0
$ ls rnaQUAST-1.1.0
augustus-3.2.1  download      gmap-2015-12-31          initial.meta.lst.feature
manual.html     other_libs    rnaQUAST.py              seq_bin_3
VERSION
blat            final_model  gms.log                  itr_0.lst
metrics        quast23      rnaQUAST_test_output    STAR-2.5.1b
bowtie2-2.2.7  GeneMarkS-T  GPLv2.txt                LICENSE
ncbi-blast-2.3.0+  README      seq_bin_1                test_data
BUSCO_v1.1b1    general      hmmer-3.1b2-linux-intel-x86_64  Log.out
objects        report      seq_bin_2                tophat-2.1.0.Linux_x86_64
```

Step 3. Before you start using the rnaQUAST 1.1.0, you need to make sure that the following softwares are added to your PATH

```
export PATH=/opt/rnaQUAST-1.1.0/ncbi-blast-2.3.0+/bin/:$PATH
export PATH=/opt/rnaQUAST-1.1.0/blat/:$PATH
export PATH=/opt/rnaQUAST-1.1.0/BUSCO_v1.1b1/:$PATH
export PATH=/opt/rnaQUAST-1.1.0/GeneMarkS-T/:$PATH
export PATH=/opt/rnaQUAST-1.1.0/STAR-2.5.1b/bin/Linux_x86_64_static/:$PATH
export PATH=/opt/rnaQUAST-1.1.0/tophat-2.1.0.Linux_x86_64/:$PATH
export PATH=/opt/rnaQUAST-1.1.0/bowtie2-2.2.7/:$PATH
```

Step 4. The staged example data can be found in folder "rnaQUAST-1.1.0/test_data" within "opt" folder. List its contents with the **ls** command:


```
$ cd /opt/rnaQUAST-1.1.0
$ ls test_data
checkpoint      Saccharomyces_cerevisiae.R64-1-1.75.dna.toplevel.1.bt2
Saccharomyces_cerevisiae.R64-1-1.75.dna.toplevel.fa  spades.311.fasta
idba.fasta      Saccharomyces_cerevisiae.R64-1-1.75.dna.toplevel.2.bt2
Saccharomyces_cerevisiae.R64-1-1.75.dna.toplevel.rev.1.bt2  Trinity.fasta
Paired_ends1.fq  Saccharomyces_cerevisiae.R64-1-1.75.dna.toplevel.3.bt2
Saccharomyces_cerevisiae.R64-1-1.75.dna.toplevel.rev.2.bt2
Paired_ends2.fq  Saccharomyces_cerevisiae.R64-1-1.75.dna.toplevel.4.bt2
Saccharomyces_cerevisiae.R64-1-1.75.gtf
```


- **Trinity.fasta**, **spades.311.fasta** and **idba.fasta** are test assemblies assembled with Trinity, spades and idba respectively
- **Saccharomyces_cerevisiae.R64-1-1.75.gtf** is the test reference annotation file
- **Saccharomyces_cerevisiae.R64-1-1.75.dna.toplevel.fa** is the test reference genome file. The files with extension (.bt2) are bowtie2 indexed reference genome files
- **Paired_end1.fq** and **Paired_end2.fq** are the test paired end fastq read files

Step 5. Set up a rnaQUAST 1.1.0 test run to check that everything is installed correctly and all the software are added to the \$PATH environment al variable

```
$ python rnaQUAST.py --test
```

Basic run:

 It is recommended to redirect the output to home directory because of the disk space restrictions on /opt directory

 By default, the rnaQUAST 1.1.0 uses GMAP. If you want to run the above test using BLAT option.


a. Using rnaQUAST 1.1.0 tool using GMAP (default)

```
$ python rnaQUAST.py --threads 4 --transcripts test_data/idba.fasta
test_data/spades.311.fasta test_data/Trinity.fasta --reference
test_data/Saccharomyces_cerevisiae.R64-1-1.75.dna.toplevel.fa --gene_database
test_data/Saccharomyces_cerevisiae.R64-1-1.75.gtf --output_dir
~/rnaQUAST_test_output_GMAP
```

b. Using rnaQUAST 1.1.0 tool using BLAT

```
$ python rnaQUAST.py --threads 4 --blat --transcripts test_data/idba.fasta
test_data/spades.311.fasta test_data/Trinity.fasta --reference
test_data/Saccharomyces_cerevisiae.R64-1-1.75.dna.toplevel.fa --gene_database
test_data/Saccharomyces_cerevisiae.R64-1-1.75.gtf --output_dir
~/rnaQUAST_test_output_BLAT
```

Read alignment:

 rnaQUAST 1.1.0 is also capable of calculating various statistics using raw reads (e.g. database coverage by reads) using either STAR aligner (or alternatively TopHat aligner +SAM tools)

a. Using rnaQUAST 1.1.0 tool using STAR aligner (default)

```
$ python rnaQUAST.py --threads 2 --transcripts test_data/idba.fasta
test_data/spades.311.fasta test_data/Trinity.fasta --left_reads
test_data/Paired_ends1.fq --right_reads test_data/Paired_ends2.fq --reference
test_data/Saccharomyces_cerevisiae.R64-1-1.75.dna.toplevel.fa --gene_database
test_data/Saccharomyces_cerevisiae.R64-1-1.75.gtf --output_dir
~/rnaQUAST_test_output_STAR
```

b. Using rnaQUAST 1.1.0 tool using TopHat aligner +SAM tools

```
$ python rnaQUAST.py --no_plots --threads 2 --transcripts test_data/idba.fasta
test_data/spades.311.fasta test_data/Trinity.fasta --left_reads
test_data/Paired_ends1.fq --right_reads test_data/Paired_ends2.fq --reference
test_data/Saccharomyces_cerevisiae.R64-1-1.75.dna.toplevel.fa --gene_database
test_data/Saccharomyces_cerevisiae.R64-1-1.75.gtf --output_dir
~/rnaQUAST_test_output_TOPHAT --tophat
```

Software for *de novo* quality assessment:



When reference genome and gene database are unavailable, it is recommend to run [BUSCO](#) and [GeneMarkS-T](#) in rnaQUAST tool.

a. Using rnaQUAST 1.1.0 tool with [BUSCO](#)

```
$ python rnaQUAST.py --threads 4 --transcripts test_data/idba.fasta
test_data/spades.311.fasta test_data/Trinity.fasta --output_dir
~/rnaQUAST_test_output_arthropoda_BUSCO --disable_infer_genes
--disable_infer_transcripts --busco --clade
/opt/rnaQUAST-1.1.0/BUSCO_v1.1b1/arthropoda
```

b. Using rnaQUAST 1.1.0 tool with [GeneMarkS-T](#)

```
$ python rnaQUAST.py --threads 4 --transcripts test_data/idba.fasta
test_data/spades.311.fasta test_data/Trinity.fasta --output_dir
~/rnaQUAST_test_output_GM --gene_mark
```



Outputs

For detailed explanation of the outputs with each of the above runs, please refer to [rnaQUAST 1.1.0 manual](#)