

RUM 2 (RNA-Seq Unified Mapper)

Mike DeLaurentis

University of Pennsylvania

August 15, 2012



Agenda

The RUM Pipeline

Enhancements in RUM 2

- Installation

- Command-line interface

- Job status

Demo

Web resources

- User guide

- Downloads

- Issues

Future direction



About me

- Software engineering background
- Working at Penn since January
- Experience in a variety of languages (Perl, Java, Clojure (lisp), Python, Ruby)

Agenda

The RUM Pipeline

Enhancements in RUM 2

- Installation

- Command-line interface

- Job status

Demo

Web resources

- User guide

- Downloads

- Issues

Future direction



About RUM

- “RUM is an alignment, junction calling, and feature quantification pipeline specifically designed for Illumina RNA-Seq data”
- Written by Gregory Grant (ggrant@grant.org)
- Runs on Linux / UNIX / Mac
- Can distribute work across multiple machines
- Written in Perl 5



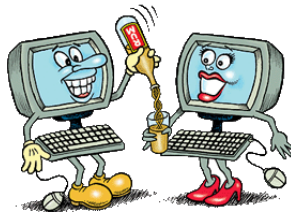
The RUM Pipeline

- Align all reads against genome using Bowtie
- Align all reads against transcriptome using Bowtie
- Merge genome and transcriptome alignments and identify unmapped reads
- Align unmapped reads against genome using BLAT
- Merge Bowtie and Blat alignments
- Produce some output files based on merged results



RUM output files

- Files of unique and non-unique alignments
- All alignments in SAM format
- Coverage plots
- Feature quantifications
- Junction calls
- List of novel inferred internal exons





Phases

Preprocessing Split the input into N chunks.

Processing Run Bowtie and Blat. Each chunk can be processed independently of the others.

Postprocessing Merge alignments for chunks together. Produce coverage files, quantifications, junction files, find novel internal exons.



Work distribution



- RUM can run alignment phase in parallel
- Splits input reads into N chunks
- Alignment for each chunk is independent of other chunks
- Results are merged together for coverage, quantification, junction calling



Agenda

The RUM Pipeline

Enhancements in RUM 2

- Installation

- Command-line interface

- Job status

Demo

Web resources

- User guide

- Downloads

- Issues

Future direction



RUM 2 Features

- Standard installation process
- New command-line interface
- Get status of running job
- Restart a job where it left off
- More reliable kill command
- Run a chunk or postprocessing by itself
- Relocatable indexes
- SAM file is closer to conforming to standard



Installation

- Uses standard Perl Makefile.PL
- Should be familiar to system administrators
- Download tarball from
<https://github.com/PGFI/rum/downloads>
- Run `perl Makefile.PL`
- Install indexes using `bin/rum_indexes`
- That's it!



Command-line interface

Usage is `rum_runner ACTION [OPTIONS]` where action is one of:

- `align` - Run an alignment
- `status` - Check the status of a job
- `stop` - Stop a job (can be restarted later)
- `kill` - Stop a job and clean it up (to restart from scratch)
- `clean` - Remove output files for a job
- `help` - Get help
- `version` - Show version number



rum_runner align

Use `rum_runner align` to run an alignment:

```
rum_runner align \  
  --output dir \  
  --index ~/rum_indexes/hg19 \  
  --name   TestJob \  
  --chunks 25 \  
  ~/samples/forward.fq ~/samples/reverse.fq
```



Job status

- Use `rum_runner status` to check on the status of a running job.

Processing in 25 chunks

```

XXXXXXXXXXXXXXXXXXXXXXXXXXXX Run bowtie on genome
XXXXXXXXXXXXXXXXXXXXXXXXXXXX Parse genome Bowtie output
X XXX  XX XX XXXX  XXXXX  Run bowtie on transcriptome
X XXX  XX XX XXXX  XXXXX  Parse transcriptome Bowtie output
X XXX  XX XX XXXX  XXXXX  Merge unique mappers together
X XXX  XX XX XXXX  XXXXX  Merge non-unique mappers together
X XXX  XX XX XXXX  XXXXX  Make unmapped reads file for blat
X XXX  XX XX XXXX  XXXXX  Run blat on unmapped reads
X XXX  XX XX XXXX  XXXXX  Run mdust on unmapped reads
X XXX  XX XX XXXX  XXXXX  Parse blat output
X XXX  XX XX XXXX  XXXXX  Merge bowtie and blat results
X XX   XX XX  XXX   XXXX   Clean up RUM files
X XX   XX XX  XXX   XXXX   Produce RUM_Unique
X XX   XX XX  XXX   XXXX   Sort RUM_Unique by location
X X    XX XX  XXX   XXXX   Sort cleaned non-unique mappers by ID
X X    XX XX  XXX   XXXX   Remove duplicates from NU
X X    XX XX  XXX   XXXX   Create SAM file
X X    XX XX  XXX   XXXX   Create non-unique stats
X X    XX XX  XXX   XXXX   Sort RUM_NU
X X    XX XX  XXX   XXXX   Generate quant
...

```



Job status

Postprocessing

```
-----  
Merge RUM_NU files  
Make non-unique coverage  
Merge RUM_Unique files  
Compute mapping statistics  
Make unique coverage  
Finish mapping stats  
Merge SAM headers  
Concatenate SAM files  
Merge quant  
make_junctions  
Sort junctions (all, bed) by location  
Sort junctions (all, rum) by location  
Sort junctions (high-quality, bed) by location  
Get inferred internal exons  
Quantify novel exons
```

All the chunk error log files are empty. That's good.
Main error log file is empty. That's good.

RUM is running (job ids 815718, 815720).



Recovering from errors



SUCCESS

Well, you can always try a second time...

\o/ MotivatedPhotos.com

- Sometimes bad things happen
- RUM 2 allows easier recovery from infrastructure failures
- Running “`rum_runner align`” again will restart a job from where it left off
- Can save *a lot* of time when recovering from infrastructure failure
- Determines state of job by looking at which output files exist



Killing a job

- To stop a job and remove all of its output:

```
rum_runner kill -o dir
```

- Useful if you've run a job with incorrect parameters and need to start over

Agenda

The RUM Pipeline

Enhancements in RUM 2

- Installation

- Command-line interface

- Job status

Demo

Web resources

- User guide

- Downloads

- Issues

Future direction



Agenda

The RUM Pipeline

Enhancements in RUM 2

Installation

Command-line interface

Job status

Demo

Web resources

User guide

Downloads

Issues

Future direction



Main Github page

`https://github.com/PGFI/rum`

The screenshot shows the GitHub repository page for PGFI/rum. At the top, there's a navigation bar with the GitHub logo, a search bar, and links for Explore, Gist, Blog, and Help. Below this, the repository name 'PGFI / rum' is displayed with icons for Pull Request, Unwatch, Unstar, 7 forks, and 2 branches. The 'Code' tab is selected, showing options to clone the repository (Clone in Mac, ZIP, HTTP, SSH, Git Read-Only) and a link to the repository URL. Below the cloning options, there's a section for the latest commit to the master branch, showing a commit by mdelaurentis 5 days ago with the commit hash 4ffdc4730f. At the bottom, there's a table of files with columns for name, age, message, and history.

| name | age | message | history |
|------|--------------|--|---------|
| bin | 12 days ago | Only archive log dir in parent process [mdelaurentis] | |
| conf | 2 months ago | Try to handle failure on a cluster better [mdelaurentis] | |



User guide

`https://github.com/PGFI/rum/wiki`

The screenshot shows the GitHub repository page for PGFI/rum. The 'Wiki' tab is selected and circled in red. The page displays the repository name, navigation links (Pull Request, Unwatch, Unstar, Fork), and a table of repository statistics (Code, Network, Pull Requests, Issues, Wiki, Graphs, Admin). Below the statistics, there is a section for the latest commit to the master branch, showing the commit message and the author's name. At the bottom, there is a table of repository files.

github • Search... Explore Gist Blog Help mdelaurentis

PGFI / rum Pull Request Unwatch Unstar 7 Fork 2

| Code | Network | Pull Requests 0 | Issues 16 | Wiki | Graphs | Admin |
|------|---------|-----------------|-----------|------|--------|-------|
|------|---------|-----------------|-----------|------|--------|-------|

RNA-Seq Unified Mapper — Read more
<http://cbil.upenn.edu/RUM>

Clone in Mac ZIP HTTP SSH Git Read-Only `https://github.com/PGFI/rum.git` Read+Write access

branch: master Files Commits Branches 11 Tags 25 Downloads 3

Latest commit to the master branch

Add pod

mdelaurentis authored 5 days ago commit 4ffdc4730f


rum /

| name | age | message | history |
|------|--------------|--|---------|
| bin | 12 days ago | Only archive log dir in parent process [mdelaurentis] | |
| conf | 2 months ago | Try to handle failure on a cluster better [mdelaurentis] | |







User guide


`https://github.com/PGFI/rum/wiki`





github 

Search...

Explore Gist Blog Help

 mdelaurentis   

 **PGFI / rum**

 Pull Request  Unwatch  Unstar 7  Fork 2

Code

Network

Pull Requests 0

Issues 16

Wiki

Graphs

Admin

Home

Pages

Wiki History

Git Access

Home

[New Page](#)[Edit Page](#)[Page History](#)

Welcome, gentle user. These pages describe the RNA-Seq Unified Mapper (RUM). We recommend reading them roughly in the order listed here.

You can find the latest version of RUM at our [downloads page](#). If you find any issues or have any feature requests, please enter them into our [issues list](#). We would really appreciate your feedback.

User guide

1. [About RUM](#)
2. [Installing RUM](#)
3. [Installing indexes](#)
4. [Creating indexes](#) (optional)
5. [Running RUM](#)
6. [RUM output files](#)
7. [Running RUM on a cluster \(optional\)](#)





Downloads

`https://github.com/PGFI/rum/downloads`

Search... Explore Gist Blog Help

rum Pull Request Unwatch Unstar 7 Fork 2

| | | | | | | |
|------|---------|-----------------|-----------|------|--------|-------|
| Code | Network | Pull Requests 0 | Issues 16 | Wiki | Graphs | Admin |
|------|---------|-----------------|-----------|------|--------|-------|

RNA-Seq Unified Mapper — [Read more](#)
<http://cbil.upenn.edu/RUM>

Clone in Mac ZIP HTTP SSH Git Read-Only Read+Write access

branch: master Files Commits Branches 11 Tags 25 **Downloads 3**

Latest commit to the master branch

Add pod

mdelaurentis authored 5 days ago commit 4ffdc4730f

rum /

| name | age | message | history |
|------|--------------|--|---------|
| bin | 12 days ago | Only archive log dir in parent process [mdelaurentis] | |
| conf | 2 months ago | Try to handle failure on a cluster better [mdelaurentis] | |



Downloads

`https://github.com/PGFI/rum/downloads`

Search... Explore Gist Blog Help

rum Pull Request Unwatch Unstar 7 Fork 2

Code Network Pull Requests 0 Issues 16 Wiki Graphs Admin

Files Commits Branches 11 Tags 25 **Downloads 3**

Download as zip Download as tar.gz

Upload a new file

Choose a new file **Short Description**

Start Upload

Download Packages **Manage Downloads**

RUM-Pipeline-v2.0.2_01.tar.gz — RUM Pipeline 2.0.2_01
4.4MB · Uploaded 11 days ago 10 downloads







RUM-Pipeline-v2.0.2.tar.gz — RUM Pipeline 2.0.2
4.4MB · Uploaded 20 days ago 3 downloads


RUM-Pipeline-v2.0.1.tar.gz — RUM Pipeline 2.0.1
4.4MB · Uploaded 2 months ago 8 downloads



Issues




`https://github.com/PGFI/rum/issues`


  Search... [Explore](#) [Gist](#) [Blog](#) [Help](#)  [mdelaurentis](#)   


 [PGFI / rum](#) [Pull Request](#) [Unwatch](#) [Unstar](#) [7](#) [Fork](#) [2](#)

[Code](#) [Network](#) [Pull Requests](#) [0](#) [Issues](#) [16](#) [Wiki](#) [Graphs](#) [Admin](#)



[RNA-Seq Unified Mapper](#) — [Read more](#)
<http://cbil.upenn.edu/RUM>

 [Clone in Mac](#)  [ZIP](#) [HTTP](#) [SSH](#) [Git Read-Only](#)  [Read+Write access](#)



 [branch: master](#) [Files](#) [Commits](#) [Branches](#) [11](#) [Tags](#) [25](#) [Downloads](#) [3](#)

 Latest commit to the **master** branch

[Add pod](#)

 [mdelaurentis](#) authored 5 days ago  [commit 4ffdc4730f](#)


[rum](#) /

| name | age | message | history |
|--|--------------|--|---------|
|  bin | 12 days ago | Only archive log dir in parent process [mdelaurentis] | |
|  conf | 2 months ago | Try to handle failure on a cluster better [mdelaurentis] | |








Issues





`https://github.com/PGFI/rum/issues`



Explore Gist Blog Help


 mdelaurentis   

 PGFI / rum

 Pull Request  Unwatch  Unstar 7  Fork 2

Code Network Pull Requests 0 **Issues 16** Wiki Graphs Admin

Browse Issues Milestones


Search: Issues & Milestones...  **New Issue**

Everyone's Issues 81


Assigned to you 47


Created by you 64


Mentioning you 0


No milestone selected 

Labels


 bug 11

 sam 3

 usability 8




 possible regression 0




No active filters. Use the sidebar to filter issues.





Keyboard shortcuts available 





16 open issues **81 closed issues** Submitted Updated Comments





Reopen Label Assignee Milestone 1 2 3 4 Next »


  #97 **Use of uninitialized value \$_ in length at /usr/local/share/perl/5.10.1/RUM/JobReport.pm line 110.**
by dipu 11 days ago  2 comments

  #96 **Unrecognized option --variable-read-lengths**
by dipu 11 days ago  2 comments

  #95 **Make "rum_runner status" indicate if job is actually running**  

  #94 **Changes to user guide**
by mdelaurentis 17 days ago  1 comment 

  #93 **Not coloring junctions properly**  

 1 comment

Agenda

The RUM Pipeline

Enhancements in RUM 2

Installation

Command-line interface

Job status

Demo

Web resources

User guide

Downloads

Issues

Future direction



Possible future enhancements

- Use original read names throughout the pipeline
- Performance improvements