

The RNA-Seq Pipeline at



Thank you.



Pathogen Portal



Featuring...



RNA Rocket

Align your Illumina fastQ reads against supported genomes, view supported genomes, and estimate gene expression values using an **RNA-Seq Pipeline** running on Galaxy.

Improvements include:

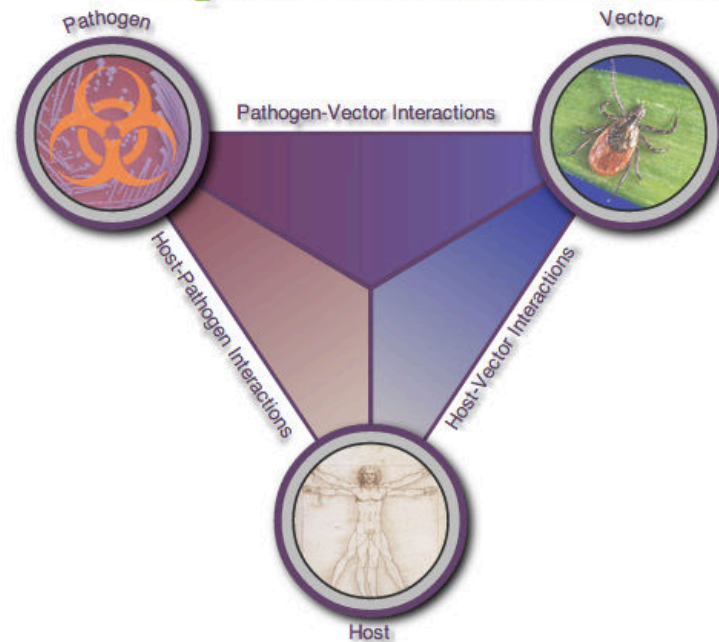
- New user interface
- New reference genomes
- New tools



Pathogen Interaction Gateway

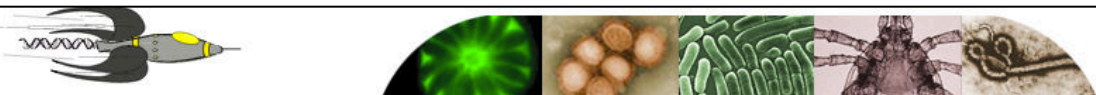
Generate a network graph of **Protein-Protein Interactions**, including **Host-Pathogen Interactions**, from your custom selection of hosts/vectors, bacteria, viruses, and eukaryotic pathogens.

Explore Infectious Disease



RNA-Rocket

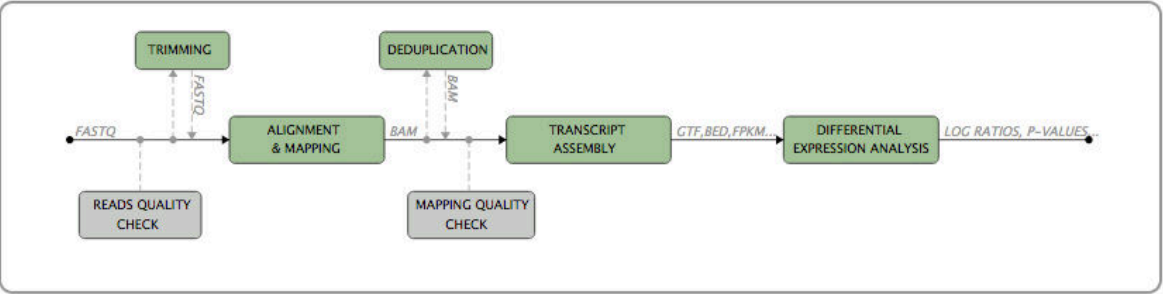
RNA-Rocket

[Login](#) | [Create an Account](#)


GalaxyLaunch PadProject ViewShared DataVisualizationHelpUserUsing 0 bytes

View a [list of supported genomes](#) from [EuPathDB](#), [PATRIC](#), and [VectorBase](#).

Have a question? [Contact the Pathogen Portal Team](#)




Choose an activity below



Uploads

[Upload Files](#)
Upload files for analysis via URL, FTP, or HTTP.



Quality Control

[Login to get started](#)

RNA-Rocket

RNA-Rocket

Galaxy Launch Pad Project View

View a [list of supported genomes](#) from [EuPathDB](#), [PATRIC](#), and [VectorBase](#).

Have a question? [Contact the Pathogen Portal Team](#)

FASTQ

READS QUALITY CHECK

TRIMMING

ALIGNMENT & MAPPING

BAM

Using 0 bytes

- ▶ Free storage 300GB
- ▶ Nightly updates from VectorBase, EuPathDB, PATRIC
- ▶ 6824 bacteria
- ▶ 60 Eukaryotic pathogens
- ▶ 8 model organisms
- ▶ 6 vectors

Choose an activity below

Uploads
[Upload Files](#)
Upload files for analysis via URL, FTP, or HTTP.

Quality Control

[Login to get started](#)

RNA-Seq Pipeline: Initial Release January 2012



Home



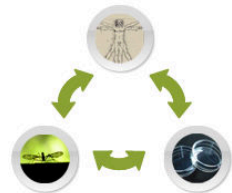
Data



Analyze

About

News and Announcements



Continue Exploring

Host

Pathogen

Environment/Vector

Analyze

Explore and use resources for analyzing host response to infectious disease.

RNA-Seq Pipeline

Map your RNA-Seq Reads to Reference Genomes

-Align your Illumina fastQ reads (**gzipped fastQ files accepted**) against any sequenced genome from EuPathDB, PATRIC, and VectorBase.

-View a list of supported genomes.

Estimate Gene Expression Values

-Obtain BAM files for the resulting alignments and FPKM expression values for annotated genes and novel transcripts.

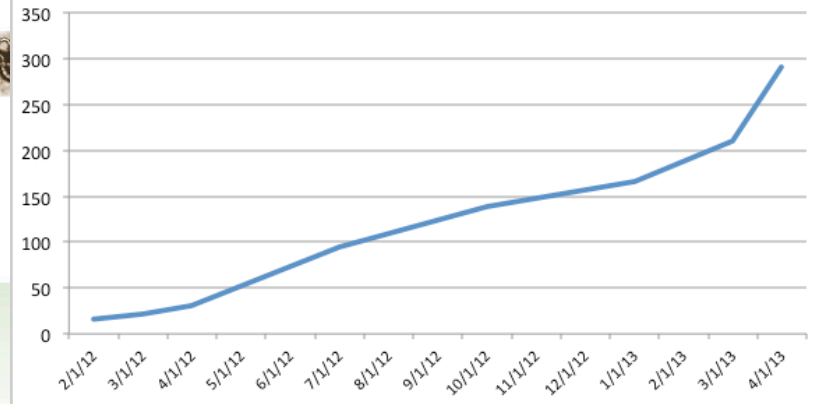
Mouse Model Selection Guide



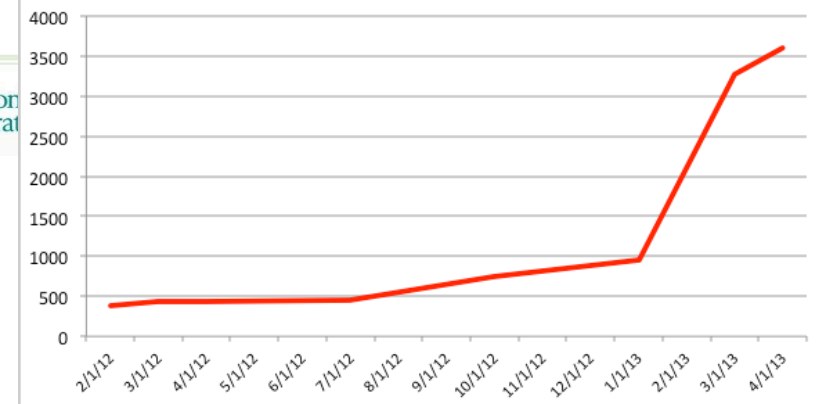
The [Mouse Model Strain Selection Guide](#) was developed by the Pathogen Portal Team in collaboration with The Jackson Laboratory. The Guide lists pathogens along with mouse strains that have been found to be either susceptible or resistant to infection with the pathogen. Links are provided to more information about the mouse strains, including ordering information. Links are also provided to publications documenting susceptibility or resistance of the mouse strains to specific pathogens.



Registered Users



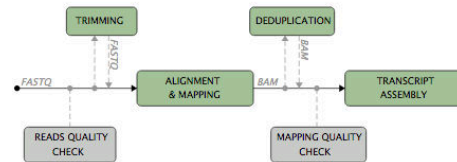
Gb I/O



RNA-Seq Pipeline Features





View a [list of supported genomes](#) from EuPathDB, PATRIC, and VectorBase.

Have a question? [Contact the Pathogen Portal Team](#)



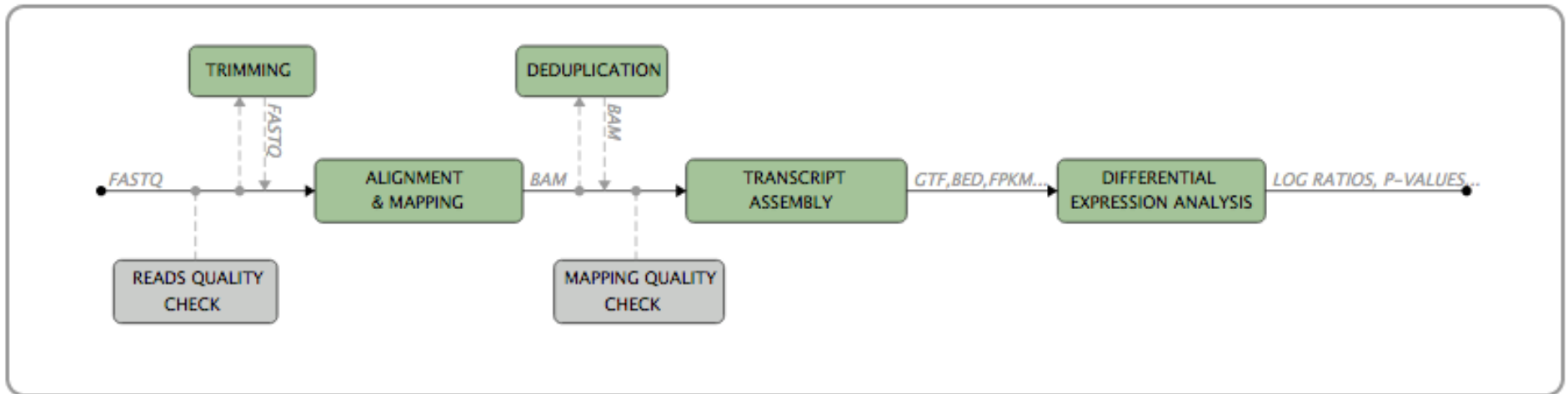
► Guided tour for naïve users

Choose an activity below

-  **Uploads**
Upload Files
Upload files for analysis via URL, FTP, or HTTP.
-  **Quality Control**
Check read quality
Optional: Run FastQC to get a report on the quality of base calls that could affect your read mapping.
Trim Reads
Optional: Run the sickle trimming tool to trim your reads and prepare them for alignment.
Check mapping quality
Optional: Check the number of reads mapped and the alignment quality.
-  **RNA-Seq Analysis**
Align Reads & Assemble Transcripts
Map your reads to the genome and assemble them into transcripts. The alignment step will generate BAM files and the assembly step will generate BED and GTF files.
Differential Expression Analysis
Test RNA-Seq samples to determine if transcripts are differentially expressed.
Create GeneList Summary
Create a GeneList file, for use in PATRIC and other differential expression analysis tools.
-  **Additional Tools**
BED Tools
Use BEDTools to create summary BED files for analysis of genome coverage.
Remove Duplicate Reads
Optional: PCR amplification can lead to bias. For paired-end reads only: if multiple pairs of reads have the exact same coordinates mark all except one as a duplicate and remove.
Alignment Only
To use advanced alignment parameters and/or perform alignment against a non-BRC genome.

Features

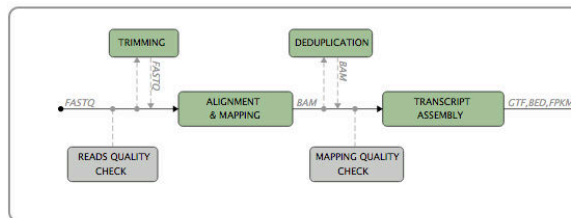
- ▶ Guided tour for naïve users



RNA-Seq Pipeline Features





View a list of supported genomes from [EuPathDB](#), [PATRIC](#), and [VectorBase](#).

Have a question? [Contact the Pathogen Portal Team](#)

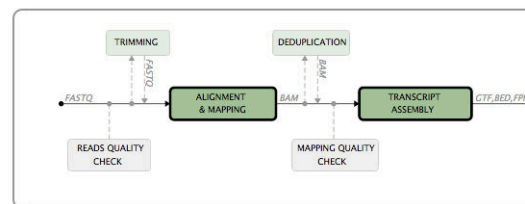


► Guided tour for naïve users

Choose an activity below

-  **Uploads**
Upload Files
Upload files for analysis via URL, FTP, or HTTP.
-  **Quality Control**
Check read quality
Optional: Run FastQC to get a report on the quality of base calls that could affect your read mapping.
Trim Reads
Optional: Run the sickle trimming tool to trim your reads and prepare them for alignment.
Check mapping quality
Optional: Check the number of reads mapped and the alignment quality.
-  **RNA-Seq Analysis**
Align Reads & Assemble Transcripts
Map your reads to the genome and assemble them into transcripts. The alignment step will generate BAM files and the assembly step will generate BED and GTF files.
Differential Expression Analysis
Test RNA-Seq samples to determine if transcripts are differentially expressed.
Create GeneList Summary
Create a GeneList file, for use in PATRIC and other differential expression analysis tools.
-  **Additional Tools**
BED Tools
Use BEDTools to create summary BED files for analysis of genome coverage.
Remove Duplicate Reads
Optional: PCR amplification can lead to bias. For paired-end reads only: if multiple pairs of reads have the exact same coordinates mark all except one as a duplicate and remove.
Alignment Only
To use advanced alignment parameters and/or perform alignment against a non-BRC genome.

RNA-Seq Pipeline Features



► Guided tour for naïve users

Align Reads & Assemble Transcripts

Purpose:

This procedure will map RNA-Seq reads to one of the provided reference genomes and use this mapping to assemble transcripts, map transcripts to existing annotations, and determine the level of expression. Choose the appropriate option for your organism (Eukaryotic/Prokaryotic) and read type (Paired-end/ Single-end) below.

Required Input:

FastQ files

Output:

Read alignments (BAM Files), tab delimited assembly and expression files for known genes, isoforms, and novel transcripts.

Select Analysis Type

- ☐ Eukaryotic Single-End Analysis
- ☐ Prokaryotic Single-End Analysis
- ☐ Eukaryotic Paired-End Analysis
- ☐ Prokaryotic Paired-End Analysis

Select an existing Project or create a new Project to be used during this analysis and populate the Project with the necessary files. Output from this analysis will be saved in the selected Project.

Currently Selected Project: **None Selected**

Target Project:

Select existing project

— OR —

Create project

← Copy

Select and copy files from Uploads or existing project(s) to populate your current Project.

Source Project:

Select source

Imported: CSU Demo

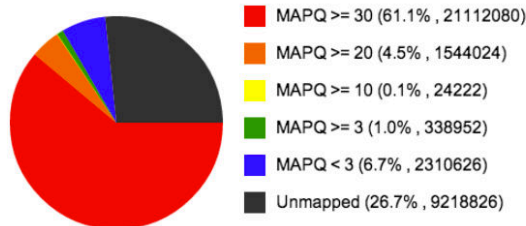
- ☐ Upstream_READ1.fastq
- ☐ Downstream_READ2.fastq
- ☐ Trimmed_Upstream_READ1.fastq
- ☐ Trimmed_Downstream_READ2.fastq
- ☐ BaseQuality_Upstream_READ1.fastq.html
- ☐ BaseQuality_Trimmed_Upstream_READ1.fastq.html
- ☐ Align with Bowtie2 Original
- ☐ Align with Bowtie2 Trimmed
- ☐ SamStat for Align with Bowtie2 on Original
- ☐ SamStat for Align with Bowtie2 Trimmed.html

RNA-Seq Pipeline Features

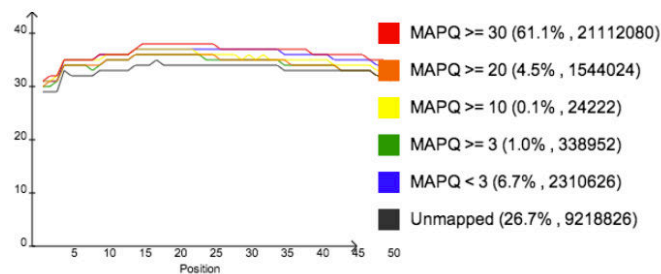
► Quality control tools for read data

Statistics for Align with Bowtie2 on data 1 and data 2: n reads

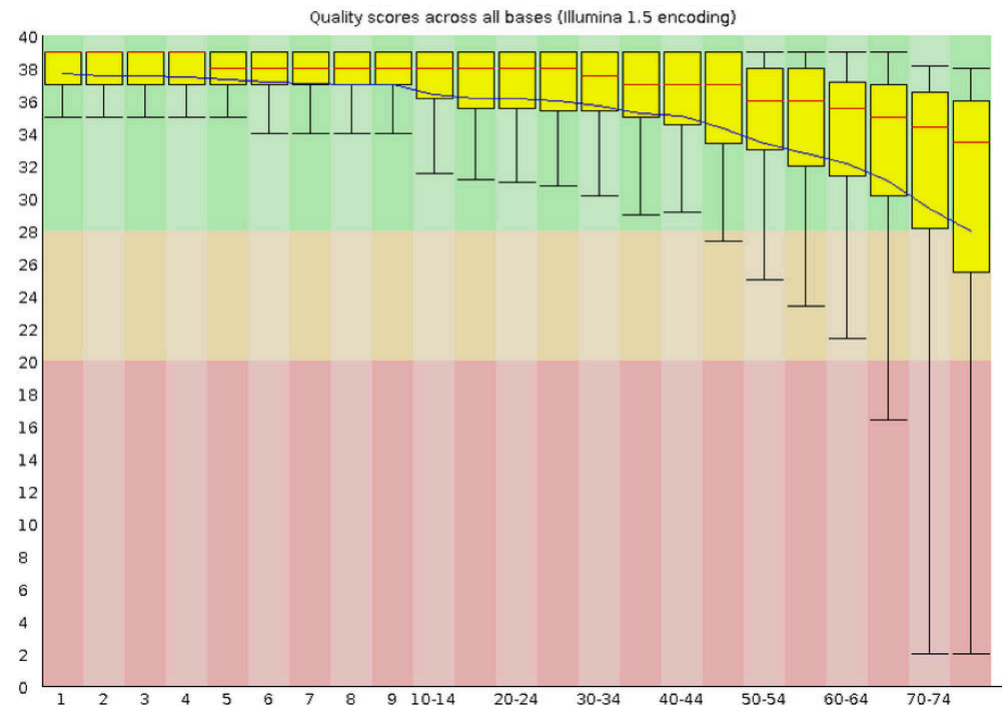
Mapping stats: 73% aligned (25.3M aligned out of 34.5M total)



Mean Base Quality

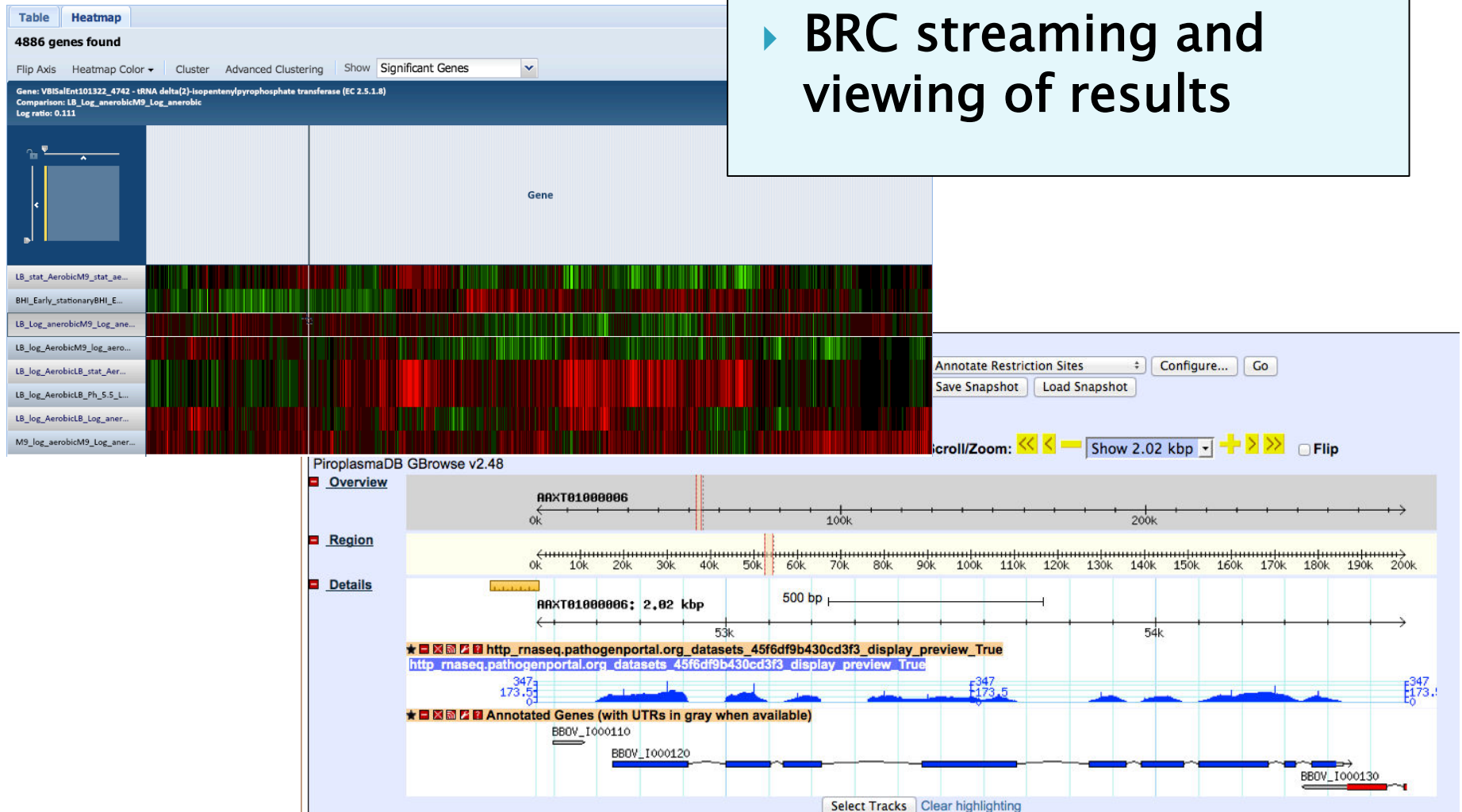


Per base sequence quality



RNA-Seq Pipeline Features

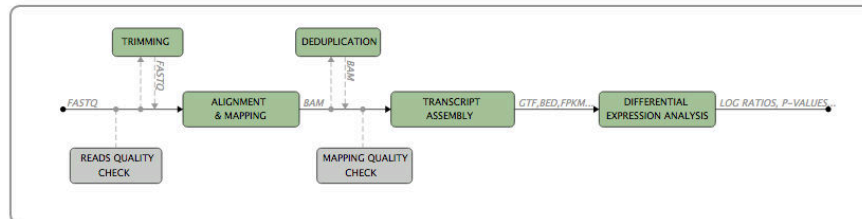
- ▶ BRC streaming and viewing of results



Galaxy Modifications

View a [list of supported genomes](#) from EuPathDB, PATRIC, and VectorBase.

Have a question? [Contact the Pathogen Portal Team](#)



Choose an activity below



Uploads

Upload Files

Upload files for analysis via URL, FTP, or HTTP.



Quality Control

Check read quality

Optional: Run FastQC to get a report on the quality of base calls that could affect

Trim Reads

Optional: Run the sickle trimming tool to trim your reads and prepare them for align

Check mapping quality

Optional: Check the number of reads mapped and the alignment quality.



RNA-Seq Analysis

Align Reads & Assemble Transcripts

Map your reads to the genome and assemble them into transcripts. The alignment step will generate BAM files and the assembly step will generate BED and GTF files.

Differential Expression Analysis

Test RNA-Seq samples to determine if transcripts are differentially expressed.

Create GeneList Summary

Create a GeneList file, for use in PATRIC and other differential expression analysis tools.



Additional Tools

BED Tools

Use BEDTools to create summary BED files for analysis of genome coverage.

Remove Duplicate Reads

Optional: PCR amplification can lead to bias. For paired-end reads only: if multiple pairs of reads have the exact same coordinates mark all except one as a duplicate and remove.

Alignment Only

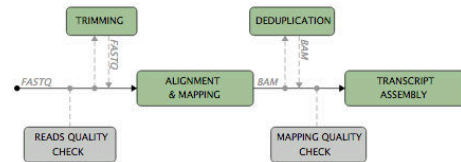
To use advanced alignment parameters and/or perform alignment against a non-BRC genome.

- ▶ Launch Pad
 - D3 library concept diagram

Galaxy Modifications

View a [list of supported genomes](#) from EuPathDB, PATRIC, and VectorBase.

Have a question? [Contact the Pathogen Portal Team](#)



► Launch Pad

- `launch_pad.py` controller
- Tool and workflow lists

Choose an activity below



Uploads

Upload Files

Upload files for analysis via URL, FTP, or HTTP.



Quality Control

Check read quality

Optional: Run FastQC to get a report on the quality of base calls that could affect your read mapping.

Trim Reads

Optional: Run the sickle trimming tool to trim your reads and prepare them for alignment.

Check mapping quality

Optional: Check the number of reads mapped and the alignment quality.



RNA-Seq Analysis

Align Reads & Assemble Transcripts

Map your reads to the genome and assemble them into transcripts. The alignment step will generate BAM files and the assembly step will generate BED and GTF files.

Differential Expression Analysis

Test RNA-Seq samples to determine if transcripts are differentially expressed.

Create GeneList Summary

Create a GeneList file, for use in PATRIC and other differential expression analysis tools.



Additional Tools

BED Tools

Use BEDTools to create summary BED files for analysis of genome coverage.

Remove Duplicate Reads

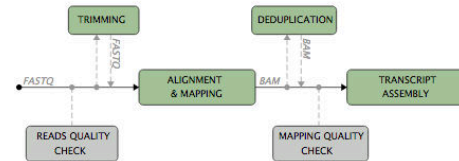
Optional: PCR amplification can lead to bias. For paired-end reads only: if multiple pairs of reads have the exact same coordinates mark all except one as a duplicate and remove.

Alignment Only

To use advanced alignment parameters and/or perform alignment against a non-BRC genome.

Galaxy Modifications

View a [list of supported genomes](#) from EuPathDB, PATRIC, and VectorBase.
Have a question? [Contact the Pathogen Portal Team](#)



► Launch Pad

- launch_pad.py controller
- Tool and workflow lists



Uploads

[Upload Files](#)

Upload files for analysis via URL, FTP, or HTTP.



Quality Control

```
<a href="{h.url_for(controller='launch_pad', action='leave_hanger', launch_type='workflow', retrieval_tag='fastqc')}"><b>Check read quality</b></a>
```



Additional Tools

[BED Tools](#)

Use BEDTools to create summary BED files for analysis of genome coverage.

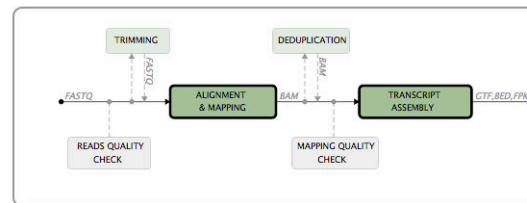
[Remove Duplicate Reads](#)

Optional: PCR amplification can lead to bias. For paired-end reads only: if multiple pairs of reads have the exact same coordinates mark all except one as a duplicate and remove.

[Alignment Only](#)

To use advanced alignment parameters and/or perform alignment against a non-BRC genome.

Galaxy Modifications



Align Reads & Assemble Transcripts

- ▶ Launch configuration
 - Tool/Workflow lists
 - Gateway user
 - API copy

Purpose:

This procedure will map RNA-Seq reads to one of the provided reference genomes and use this mapping to assemble transcripts, map transcripts to existing annotations, and determine the level of expression. Choose the appropriate option for your organism (Eukaryotic/Prokaryotic) and read type (Paired-end/ Single-end) below.

Required Input:

FastQ files

Output:

Read alignments (BAM Files), tab delimited assembly and expression files for known genes, isoforms, and novel transcripts.

Select Analysis Type

- ☐ Eukaryotic Single-End Analysis
- ☐ Prokaryotic Single-End Analysis
- ☐ Eukaryotic Paired-End Analysis
- ☐ Prokaryotic Paired-End Analysis

Select an existing Project or create a new Project to be used during this analysis and populate the Project with the necessary files. Output from this analysis will be saved in the selected Project.

Currently Selected Project: **None Selected**

Target Project:

Select existing project

— OR —

Create project

← Copy

Select and copy files from Uploads or existing project(s) to populate your current Project.

Source Project:

Select source

Imported: CSU Demo

- ☐ Upstream_READ1.fastq
- ☐ Downstream_READ2.fastq
- ☐ Trimmed_Upstream_READ1.fastq
- ☐ Trimmed_Downstream_READ2.fastq
- ☐ BaseQuality_Upstream_READ1.fastq.html
- ☐ BaseQuality_Trimmed_Upstream_READ1.fastq.html
- ☐ Align with Bowtie2 Original
- ☐ Align with Bowtie2 Trimmed
- ☐ SamStat for Align with Bowtie2 on Original
- ☐ SamStat for Align with Bowtie2 Trimmed.html

Galaxy Modifications

Configure Workflow Run for "CSU Demo"

Step 1: Input dataset

Downstream files must be in the same order as

Upstream Read Files

Available

1: Upstream_READ1.fastq
2: Downstream_READ2.fastq
3: Trimmed_Upstream_READ1.fast
4: Trimmed_Downstream_READ2.f

Selected



type to filter, [enter] to select all

Workflow configuration

- Paired-end workflows
- Extjs- run.mako
- Coordinated selection
 - Display type
- AJAX load
- Cufflinks parameters e.g.

Step 2: Input dataset

Downstream files must be in the same order as their corresponding upstream files

Downstream Read Files

Available

1: Upstream_READ1.fastq
2: Downstream_READ2.fastq
3: Trimmed_Upstream_READ1.fast
4: Trimmed_Downstream_READ2.f

Selected



type to filter, [enter] to select all

Galaxy Modifications

Select a reference genome
Rothia dentocariosa M567
If your genome of interest is not listed, contact Pathogen Portal team.

Parameter Settings
Full parameter list
You can use the default settings or set custom values for any of Bowtie's parameters.

Type of alignment ⓘ
End to end

Use Preset options
Yes

Preset option ⓘ
Sensitive

Specify the read group for this file?
No

- ▶ Workflow configuration
 - Paired-end workflows
 - Extjs- run.mako
 - Coordinated selection
 - Display type
 - AJAX load
 - Cufflinks parameters e.g.

Step 4: Cufflinks Prokaryotic (version 2.0.2)

SAM or BAM file of aligned RNA-Seq reads
Output dataset 'output' from step 3

Maximum Intron Length (-I) ⓘ
300000

Minimum Isoform Fraction (-F) ⓘ
0.1

Pre MRNA Fraction (-J) ⓘ
0.15

Overlap Radius ⓘ
50

Perform Quartile Normalization ⓘ
No

Will you select a reference annotation from your history or use a built-in file from Pathogen Portal?
Use provided annotation

Select a reference annotation
Rothia dentocariosa M567
If your annotation of interest is not listed, contact Pathogen Portal team.

Select how to use the provided annotation
Assemble ONLY transcripts matching the annotation

Other Modifications

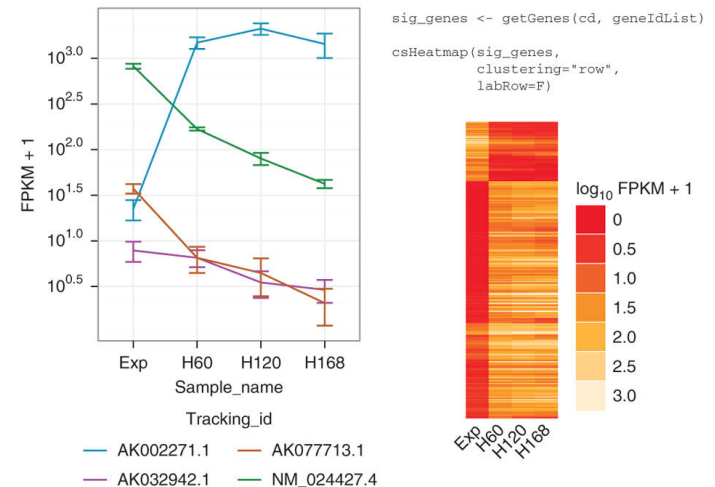
- ▶ Cuffdiff/cuffmerge integration
- ▶ SAM Stat
- ▶ Select2 wait time
- ▶ Power users

Challenges

- ▶ Controller system
- ▶ Workflow vs tool runners
 - Fixed parameters
 - Inputs
- ▶ Advanced parameters

Future Plans

- ▶ Visualization of results
 - CummeRbund
- ▶ Organism specific tools
- ▶ Link outs to BRCs
 - Annotation differences
 - Differential Expression Visualization
 - Alignment results streaming



Acknowledgements

pathogenportal.org

- ▶ Eric Nordberg
- ▶ Dustin Machi
- ▶ Chunhong Mao
- ▶ This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900040C, awarded to BWS Sobral.