

Galaxy

Gloria I. Giraldo-Calderón
September 2017



VectorBase

Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

Outline

- Introduction to Galaxy
- RNAseq analysis

Galaxy Overview

- Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research.
 - *Accessibility*: Galaxy enables users without programming experience to easily specify parameters and run tools and workflows.
 - *Reproducibility*: Galaxy captures all information necessary so that any user can repeat and understand a complete computational analysis.
 - *Transparency*: Galaxy enables users to share and publish analyses via the web and create Pages--interactive, web-based documents that describe a complete analysis.
- Galaxy is open source for all organizations. The public Galaxy server makes analysis tools, genomic data, tutorial demonstrations, persistent workspaces, and publication services available to any scientist that has access to the Internet. Local Galaxy servers can be set up by downloading the Galaxy application and customizing it to meet particular needs
- Public server URL is <http://usegalaxy.org/>
- VectorBase's server (for registered VB users) is <https://www.vectorbase.org/galaxy>

Galaxy homepage

The screenshot displays the Galaxy homepage with a dark blue header bar. The header includes the 'Galaxy' logo, navigation links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User', and a 'Using 37%' status indicator. The left sidebar, titled 'Tools', contains a 'Load Data' button and a search bar. Below the search bar is a list of tool categories: Get Data, Send Data, Lift-Over, Text Manipulation, Convert Formats, FASTA manipulation, Filter and Sort, Join, Subtract and Group, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Graph/Display Data, Regional Variation, Multiple regression, Multivariate Analysis, Evolution, Motif Tools, Multiple Alignments, Metagenomic analyses, and Genome Diversity. At the bottom of the sidebar are links for 'NGS TOOLBOX BETA', 'Phenotype Association', and 'NGS: QC and manipulation'. The main content area features a central banner with the text 'Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#).' Below the banner is a grid of logos for 'OSDDLINUX LiveGalaxy', 'OPEN SOURCE DRUG DISCOVERY', and 'iPlant Collaborative'. To the right of the banner is a 'Tweets' section with three tweets from the Galaxy Project (@galaxyproject). The right sidebar, titled 'History', shows 'Unnamed history' with '0 bytes' and a message: 'Your history is empty. Click 'Get Data' on the left pane to start'.

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Using 37%

Tools Load Data

search tools

Get Data
Send Data
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Genome Diversity

NGS TOOLBOX BETA
Phenotype Association
NGS: QC and manipulation

Galaxy is an open source, web-based platform for data intensive biomedical research.
If you are new to Galaxy [start here](#) or consult our [help resources](#).

OSDDLINUX LiveGalaxy

OPEN SOURCE DRUG DISCOVERY

Tweets

Galaxy Project @galaxyproject 4 Feb
MT @UofCr4kids: In Calgary? next generation sequencing, begets need for more bioinformatics. Learn Galaxy THIS WEEK [bit.ly/1bXtjqs](#)

Galaxy Project @galaxyproject 2 Feb
GlobusWorld abstract deadline is Feb 15 Includes new Biosciences/Genomics Program [bit.ly/globusworld2014](#) (and [bit.ly/gxyGlobusGenom...](#))
Expand

Galaxy Project @galaxyproject 31 Jan
The February 2014 Galaxy Update newsletter features new papers, jobs
Tweet to @galaxyproject

History

Unnamed history
0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

PENNSTATE

JOHNS HOPKINS UNIVERSITY

TACC

iPlant Collaborative

Galaxy homepage

List of available tools (actions)

Galaxy

Analyze Data Workflow Shared Data Visualization Help Login or Register

Tools

search tools

Get Data
Lift-Over
Collection Operations
Text Manipulation
Datamash
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Alignments/Sequences
NGS: QC and manipulation
NGS: DeepTools
NGS: Mapping
NGS: RNA Analysis
NGS: SAMtools
NGS: BamTools
NGS: Picard
NGS: VCF Manipulation
NGS: Peak Calling
NGS: Variant Analysis
NGS: RNA Structure
NGS: Du Novo
NGS: Gemini
NGS: Assembly
NGS: Chromosome Conformation
NGS: Motur
Operate on Genomic Intervals
Statistics
Graph/Display Data
Phenotype Association
BEDTools
Genome Diversity
EMBOS
Regional Variation
FASTA manipulation
Multiple Alignment
Metagenomic Analysis
Multiple regression

https://www.tacc.utexas.edu

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#). You can install your own Galaxy by following the [tutorial](#) and choose from thousands of tools from the [Tool Shed](#).

080+
Public Galaxy Servers
and still counting

Tweets by @galaxyproject

Galaxy Project @galaxyproject
Featuring another public Galaxy server [ribogalaxy.ucc.ie](#) "provides on-line tools for analysis and visualization of ribo-seq data"

Galaxy Project Retweeted
Cristel Thomas @crstlthms
Hey! Let us know how we're doing with ImmPort Galaxy! [docs.google.com/forms/d/e/1FAI...](#)

Embed View on Twitter

PENNSTATE
JOHNS HOPKINS UNIVERSITY
OREGON HEALTH & SCIENCE UNIVERSITY
TACC
CYVERSE

The Galaxy Team is a part of the [Center for Comparative Genomics and Bioinformatics](#) at Penn State, the [Department of Biology](#) and at [Johns Hopkins University](#) and the [Computational Biology Program](#) at [Oregon Health & Science University](#).

This instance of Galaxy is utilizing infrastructure generously provided by the [CyVerse](#) at the [Texas Advanced Computing Center](#), with support from the [National Science Foundation](#).

The Galaxy Project is supported in part by [NSF](#) [NHGRI](#), [The Huck Institutes of the Life Sciences](#), [The Institute for CyberScience](#) at Penn State, and [Johns Hopkins University](#).

This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own [local Galaxy instance](#) or run [Galaxy on the cloud](#).

History

search datasets

Unnamed history
(empty)

This history is empty. You can load your own data or get data from an external source

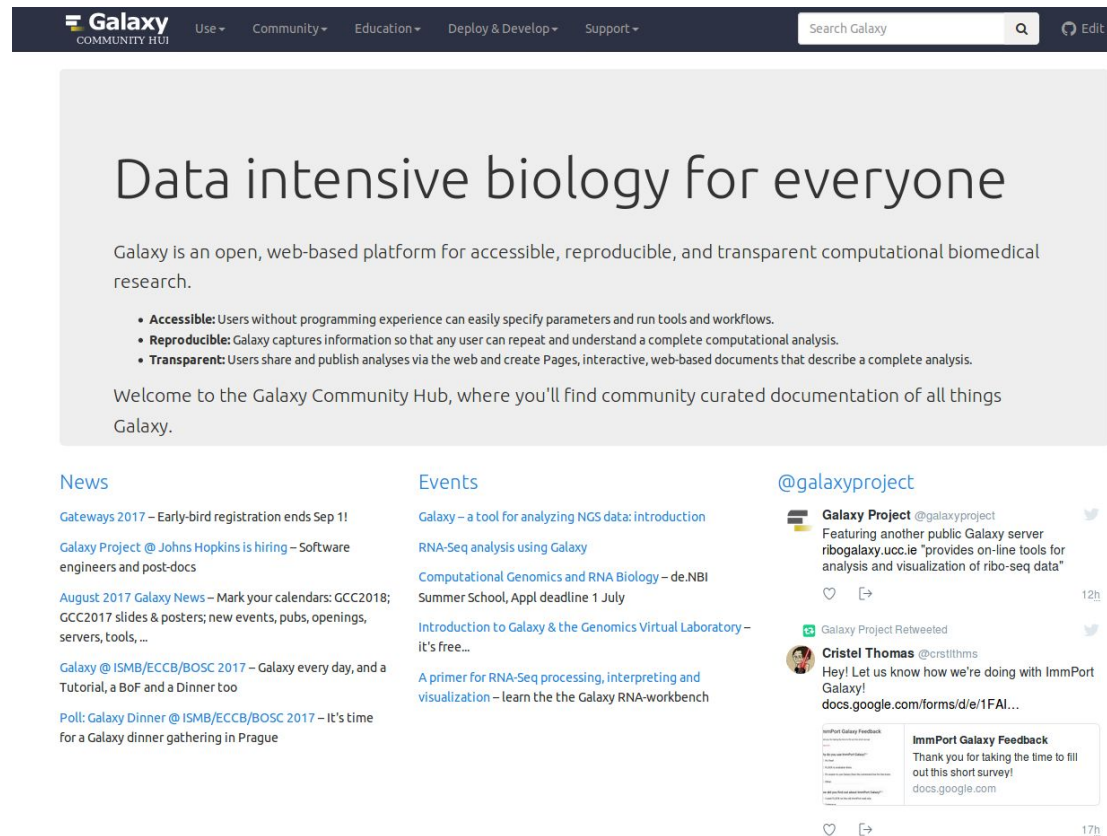
Galaxy version 17.05, commit [1d485399b0a1f88441c4e577f8fb7d6dc265f9618](#)

History of data sets and activity

Main panel - shows options/configuration and displays output

Useful resources: Galaxy Wiki

- Wiki site: <https://galaxyproject.org/>
- Documentation: <https://galaxyproject.org/learn/>



The screenshot shows the Galaxy Community Hub website. The header is dark blue with the Galaxy logo and navigation links: Use, Community, Education, Deploy & Develop, and Support. A search bar and an edit icon are on the right. The main content area has a light gray background with the title "Data intensive biology for everyone". Below the title is a paragraph describing Galaxy as an open, web-based platform for accessible, reproducible, and transparent computational biomedical research. This is followed by three bullet points: Accessible (Users without programming experience can easily specify parameters and run tools and workflows), Reproducible (Galaxy captures information so that any user can repeat and understand a complete computational analysis), and Transparent (Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis). Below this is a welcome message to the Galaxy Community Hub. The bottom section is divided into three columns: News, Events, and a social media section for @galaxyproject. The News column lists several upcoming events and deadlines. The Events column lists various workshops and seminars. The social media section shows a tweet from the Galaxy Project and a tweet from Cristel Thomas, both promoting a survey about the ImmPort Galaxy server.

Galaxy
COMMUNITY HUB

Use • Community • Education • Deploy & Develop • Support •

Search Galaxy 🔍 Edit

Data intensive biology for everyone

Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research.

- **Accessible:** Users without programming experience can easily specify parameters and run tools and workflows.
- **Reproducible:** Galaxy captures information so that any user can repeat and understand a complete computational analysis.
- **Transparent:** Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

Welcome to the Galaxy Community Hub, where you'll find community curated documentation of all things Galaxy.

News

Gateways 2017 – Early-bird registration ends Sep 1!

Galaxy Project @ Johns Hopkins is hiring – Software engineers and post-docs

August 2017 Galaxy News – Mark your calendars: GCC2018; GCC2017 slides & posters; new events, pubs, openings, servers, tools, ...

Galaxy @ ISMB/ECCB/BOSC 2017 – Galaxy every day, and a Tutorial, a BoF and a Dinner too

Poll: Galaxy Dinner @ ISMB/ECCB/BOSC 2017 – It's time for a Galaxy dinner gathering in Prague

Events

Galaxy – a tool for analyzing NGS data: introduction

RNA-Seq analysis using Galaxy

Computational Genomics and RNA Biology – de.NBI Summer School, Appl deadline 1 July

Introduction to Galaxy & the Genomics Virtual Laboratory – it's free...

A primer for RNA-Seq processing, interpreting and visualization – learn the the Galaxy RNA-workbench

@galaxyproject

Galaxy Project @galaxyproject
Featuring another public Galaxy server ribogalaxy.ucc.ie "provides on-line tools for analysis and visualization of ribo-seq data"

Galaxy Project Retweeted

Cristel Thomas @crstlthms
Hey! Let us know how we're doing with ImmPort Galaxy!
docs.google.com/forms/d/e/1FAI...

ImmPort Galaxy Feedback
Thank you for taking the time to fill out this short survey!
[docs.google.com](https://docs.google.com/forms/d/e/1FAI...)

Video guides to using Galaxy

- Available from <http://vimeo.com/galaxyproject>

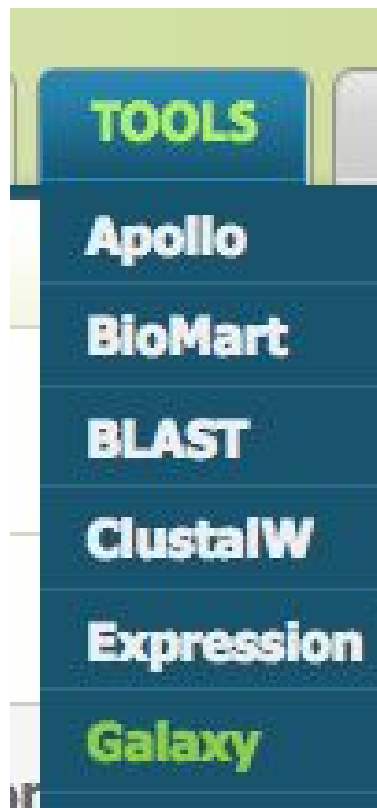
115 videos at time of writing!

- Learning resources (3:45) <http://vimeo.com/75940376>
- Datasets (7:26) <http://vimeo.com/79356949>
- Loading data and understanding datatypes (10:41) <http://vimeo.com/76351539>
- Get data: upload file (7:58) <http://vimeo.com/75938324>
- FASTQ prep (13:42) <http://vimeo.com/76024253>
- Custom genome (1:34) <http://vimeo.com/75918922>

VectorBase Galaxy

- Available to all vector community users
- Hardware: 80 cores, 100GB RAM, many TBs of storage
- If you run NGS analysis once or twice a year, you won't be able to justify a big machine like this on your grants!
- Default 250GB disk space quota per user
- Friendly support!

Let's play with Galaxy on VectorBase



It's in the Tools menu

RNAseq Analysis

Part I - Mapping reads

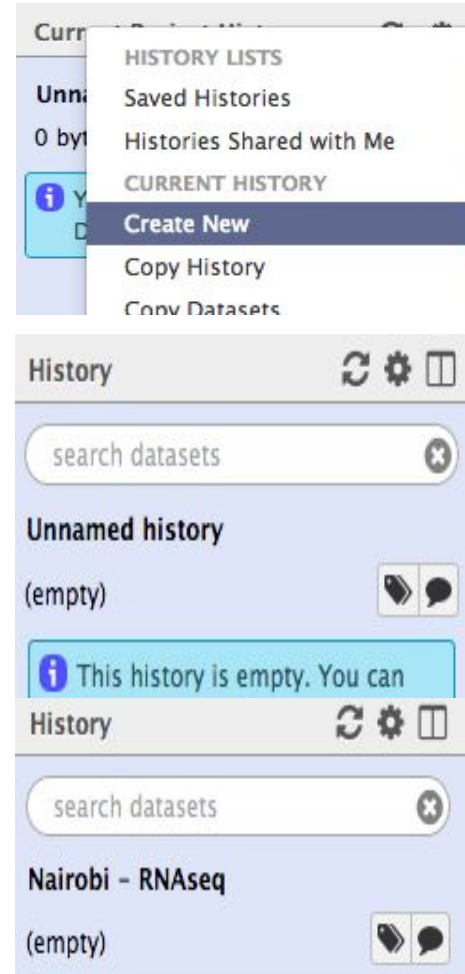
RNAseq analyses in Galaxy

- Allows for QC metrics and filtering of reads (not covered today)
- Alignment to reference genome assembly
- Transcript reconstruction
- Calculate expression values and differential expression analysis

Set up a new history

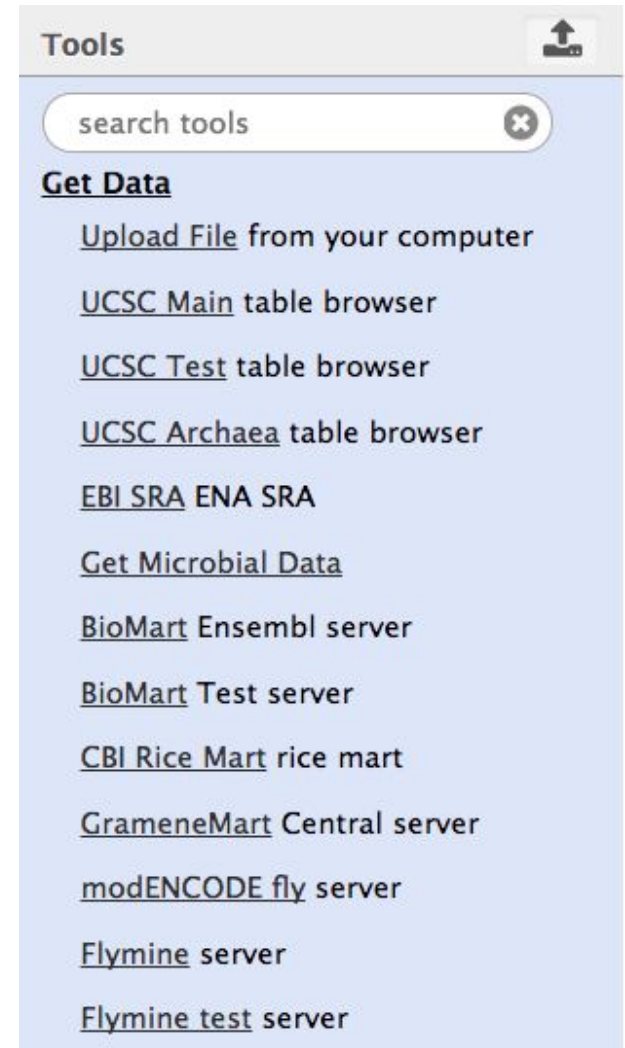
- Create a new history
- Change the name, to something like "my first history"

You will only use it for getting familiar with Galaxy basics



Uploading data

- From local files (not recommended for very large files)
- From HTTP/FTP uploads
- From ENA SRA
- Shared from Galaxy histories



Uploading data from file

This is only recommended for small (MB not GB) files
Download a file from VectorBase **Downloads->Data files** as follows:





- Filter for species *Anopheles gambiae* and look for this file:


Anopheles-gambiae-PEST_BASEFEATURES_AgamP4.7.
gtf.gz



- Download it to your computer
- Galaxy->Get Data->Upload file

Uploading data from file

Regular Composite

Name	Size	Type		Genome	Settings	Status
 Anopheles-gambiae-PEST_BASEFEATURES_AgamP4.7.gtf.gz	1.8 MB	Auto-detect		unspecified (?)		100% 

Type (set all): Auto-detect **Genome (set all):** unspecified (?)

 Choose local file  Paste/Fetch data Pause Reset Start Close

- For local files use the 'Choose local file' option
- For remote files with FTP/HTTP URLs choose 'Paste/Fetch data'

Don't forget to click Start...

Files will be loaded into the current history

Choose the file format or let Galaxy auto-detect

Uploading data from file

- Confirmation of upload job

Galaxy / VectorBase


Analyze DataWorkflowShared Data▼Visualization▼Help▼User▼

Using 87%

Tools

search tools

[Get Data](#)
[Send Data](#)
[Lift-Over](#)
[Text Manipulation](#)
[Filter and Sort](#)
[Join, Subtract and Group](#)
[Convert Formats](#)
[Extract Features](#)
[Fetch Sequences](#)
[Fetch Alignments](#)
[Statistics](#)
[Graph/Display Data](#)
[NGS: RNA Analysis](#)



Welcome to the **VectorBase** Galaxy Server

We are pleased to announce that VectorBase Galaxy has been upgraded to version 16.07. For information on this release, please refer to the [Galaxy release notes](#)

[Galaxy](#) is an open, web-based platform for data intensive biomedical research. The [Galaxy team](#) is a part of [BX](#) at [Penn State](#), and the [Biology](#) and [Mathematics and Computer Science](#) departments at [Emory University](#). The [Galaxy Project](#) is supported in part by [NHGRI](#), [NSF](#), [The Huck Institutes of the Life Sciences](#), [The Institute for CyberScience at Penn State](#), and [Emory University](#).

History

search datasets

demo
1 shown, 1 [deleted](#)
21.47 MB

[2: Anopheles-gambiae-](#)
[PEST BASEFEATURES AgamP4.7.gtf](#)

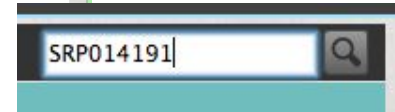
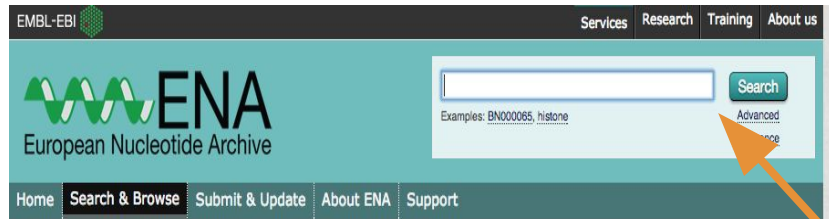
Edit file attributes/format

- Click the filename to expand the file info section
- Note that Galaxy has guessed 'gff' format
- We will change it to 'gtf'

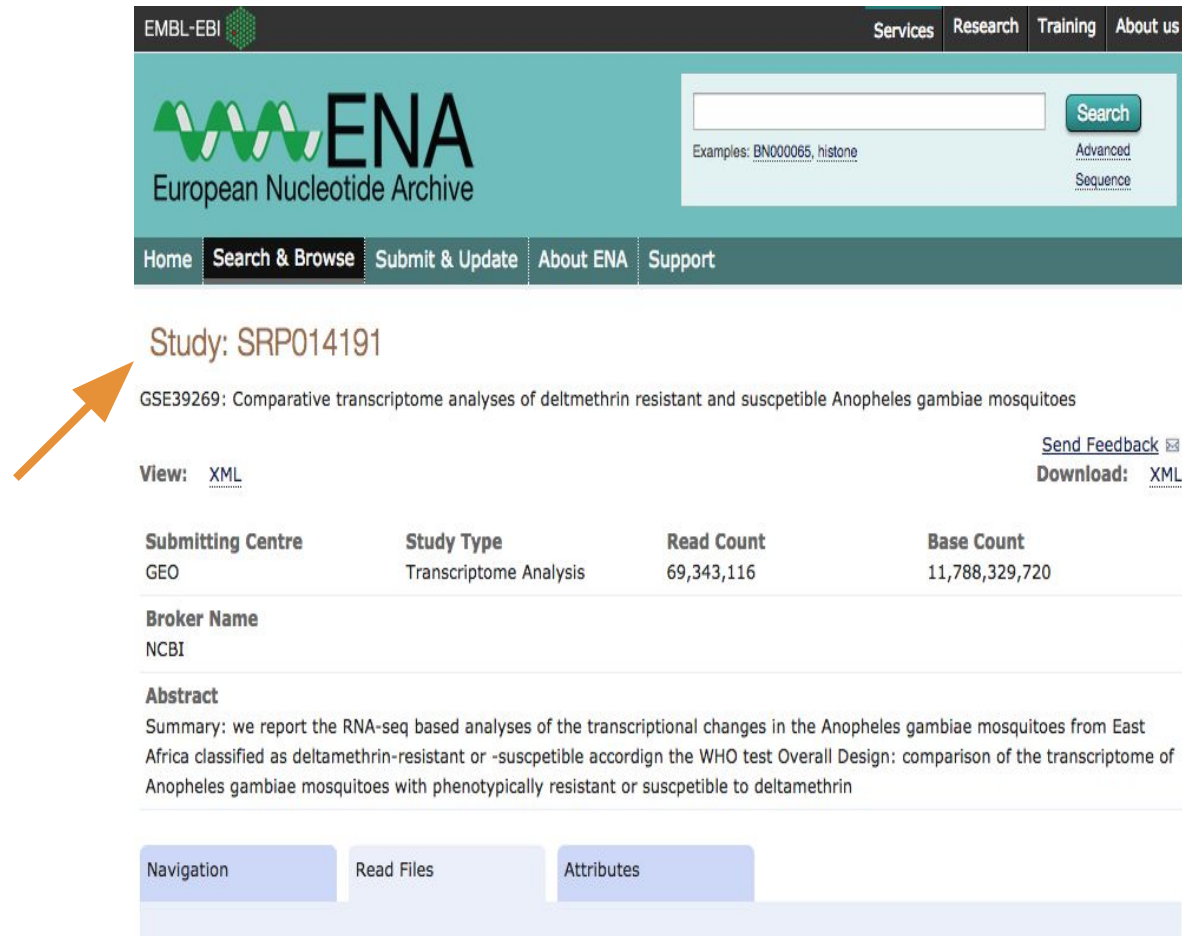
The image shows two panels from the Galaxy web interface. The left panel is the 'Attributes' tab, specifically the 'Datatype' sub-tab. It has a title bar with 'Attributes', 'Convert Format', 'Datatype', and 'Permissions'. The main content area is titled 'Change data type' and contains a 'New Type:' dropdown menu with 'gtf' selected. Below the dropdown is a text box stating: 'This will change the datatype of the existing dataset but not modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.' At the bottom of this section is a 'Save' button. A red arrow labeled '2' points to the 'Datatype' tab. A red arrow labeled '3' points to the 'gtf' dropdown. A red arrow labeled '4' points to the 'Save' button. The right panel is the 'History' tab, showing a list of datasets. The top dataset is 'demo' (21.47 MB). Below it is a dataset named '2: Anopheles-gambiae-PEST_BASEFEATURES AgamP4.7.gtf' with '~200,000 lines' and 'format: gtf, database: ?'. A red arrow labeled '1' points to the edit icon (pencil) next to this dataset name. Below the dataset name, there is a section for '1. Seqname' containing genomic information: '#!genome-build AgamP4', '#!genome-version AgamP4', '#!genome-date 2006-02', '#!genome-build-accession GCA_000005575.1', and '#!genomebuild-last-updated 2015-10'.

Uploading data from SRA

- Galaxy wraps the ENA pages within a frame
- Type the SRA accession you are interested into the search field
- Projects SRPxxxxxx
- Experiments SRXxxxxxx
- Runs SRRxxxxxxx



Uploading data from SRA: SRP014191



EMBL-EBI

Services Research Training About us

ENA
European Nucleotide Archive

Search
Examples: BN000065, histone
Advanced
Sequence

Home Search & Browse Submit & Update About ENA Support

Study: SRP014191

GSE39269: Comparative transcriptome analyses of deltamethrin resistant and susceptible *Anopheles gambiae* mosquitoes

View: [XML](#) [Send Feedback](#) [Download: XML](#)

Submitting Centre	Study Type	Read Count	Base Count
GEO	Transcriptome Analysis	69,343,116	11,788,329,720

Broker Name
NCBI


Abstract
Summary: we report the RNA-seq based analyses of the transcriptional changes in the *Anopheles gambiae* mosquitoes from East Africa classified as deltamethrin-resistant or -susceptible according to the WHO test Overall Design: comparison of the transcriptome of *Anopheles gambiae* mosquitoes with phenotypically resistant or susceptible to deltamethrin

Navigation Read Files Attributes

Uploading data from SRA: SRP014191

EMBL-EBI

ServicesResearchTrainingAbout us



European Nucleotide Archive

Search

Examples: [BN000065](#), [histone](#)

[Advanced](#)
[Sequence](#)

HomeSearch & BrowseSubmit & UpdateAbout ENASupport

NavigationRead FilesAttributes

► Download files

View: [TEXT](#)Download: [TEXT](#)


[Select columns](#)

Showing results 1 - 2 of 2 results

Study accession	Secondary study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)	Submitted files (ftp)	Submitted files (galaxy)	CoL tax ID	CoL scientific name
SRP014191	SRP014191	SAMN01087356	SRS349590	SRX159159	SRR520427	Anopheles gambiae	Illumina Genome Analyzer II	PAIRED	File 1 File 2	File 1 File 2	Not available	Not available	8630925	Anopheles gambiae Giles, 1902
SRP014191	SRP014191	SAMN01087357	SRS349591	SRX159160	SRR520428	Anopheles gambiae	Illumina Genome Analyzer II	PAIRED	File 1 File 2	File 1 File 2	Not available	Not available	8630925	Anopheles gambiae Giles, 1902

Anopheles gambiae mosquitoes with phenotypically resistant or susceptible to deltamethrin

NavigationRead FilesAttributes



Uploading data from SRA



The following job has been successfully added to the queue:

1: EBI SRA: SRP014191 File: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR520/SRR520427/SRR520427_1.fastq.gz



You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.



Project List





search project names and tags



[Advanced Search](#)



<input type="checkbox"/> Project Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated ↑	Status
<input type="checkbox"/> Uploaded Files ▾	1	0 Tags		0 bytes	5 minutes ago	less than a minute ago	current project





Current Project History  

Uploaded Files
0 bytes  

 1: EBI SRA: SRP014191   
File:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR520/SRR520427/SRR520427_1.fastq.gz

Current Project History  

Uploaded Files
0 bytes  

 1: EBI SRA: SRP014191 File:   
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR520/SRR520427/SRR520427_1.fastq.gz

- Galaxy upload files confirmations
- (the same as for the manual uploads)

Uploading data using SRA fastq links

EMBL-EBI

Services Research Training About us

ENA
European Nucleotide Archive

Search

Examples: BN000065, histone

Advanced Sequence

Home Search & Browse Submit & Update About ENA Support

Navigation Read Files Attributes

► Download files

View: [TEXT](#)

[Select columns](#)

Showing results 1 - 2 of 2 results

Download: [TEXT](#)

Study accession	Secondary study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)	Submitted files (ftp)	Submitted files (galaxy)	CoL tax ID	CoL scientific name
SRP014191	SRP014191	SAMN01087356	SRS349590	SRX159159	SRR520427	Anopheles gambiae	Illumina Genome Analyzer II	PAIRED	File 1 File 2	File 1 File 2	Not available	Not available	8630925	Anopheles gambiae Giles, 1902
SRP014191	SRP014191	SAMN01087357	SRS349591	SRX159160	SRR520428	Anopheles gambiae	Illumina Genome Analyzer II	PAIRED	File 1 File 2	File 1 File 2	Not available	Not available	8630925	Anopheles gambiae Giles, 1902

Anopheles gambiae mosquitoes with phenotypically resistant or susceptible to deltamethrin

Navigation Read Files Attributes

Uploading data from SRA url links

The screenshot displays the Galaxy / VectorBase web interface. A modal dialog titled "Download from web or upload from disk" is open in the center. The dialog has two tabs: "Regular" (selected) and "Composite". Below the tabs, a message states: "You added 1 file(s) to the queue. Add more files or click 'Start' to proceed." A table lists the added file:

Name	Size	Type	Genome	Settings	Status
New File	158 b	fastqsanger	unspecified (?)		

Below the table, a text box contains the following URLs:

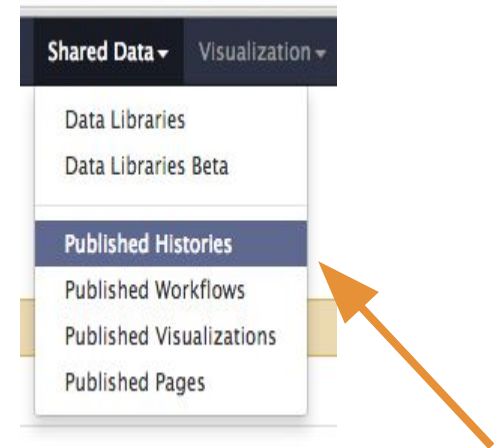
```
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR159/009/SRR1593369/SRR1593369_1.fastq.gz  
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR159/009/SRR1593369/SRR1593369_2.fastq.gz

At the bottom of the dialog, there are two main options: "Choose local file" and "Paste/Fetch data" (which is selected). To the right of these are "Pause", "Reset", "Start" (circled in red), and "Close" buttons. The "Start" button is highlighted with a red circle and the number 4. A red arrow points to the "fastqsanger" dropdown menu in the table, labeled with the number 3. Another red arrow points to the "Paste/Fetch data" button, labeled with the number 2. A red arrow points to the "Upload File from your computer" link in the left sidebar, labeled with the number 1.


```

Copying data from published history

- Data can be shared between users within Galaxy
- Histories are either public (found by search) or shared via URLs or specific user names
- In this exercise we will get our data from a shared history
- Find and click on:
PRJNA170440 Deltamethrin resistance Anopheles gambiae RNA-seq DEMO data v2
 - Click on the "Import history" link, top right
 - Rename your imported history something like "Workshop RNA-Seq analysis"



Copying individual datasets

- You can copy files/datasets from one history to another (but we won't do that today)

Copy any number of history items from one history to another.

Source History:

10: PRJNA170440 Deltamethrin...

All None

- ☐ 1: resistant_1.fastq
- ☐ 2: resistant_2.fastq
- ☐ 3: susceptible_1.fastq
- ☐ 4: susceptible_2.fastq
- ☐ 5: Anopheles-gambiae-PEST_BASEFEATURES_AgamP4.7.gtf
- ☐ 6: Anopheles-gambiae-PEST_CHROMOSOMES_AgamP4.fa

→

Destination History:

2: Nairobi - RNAseq

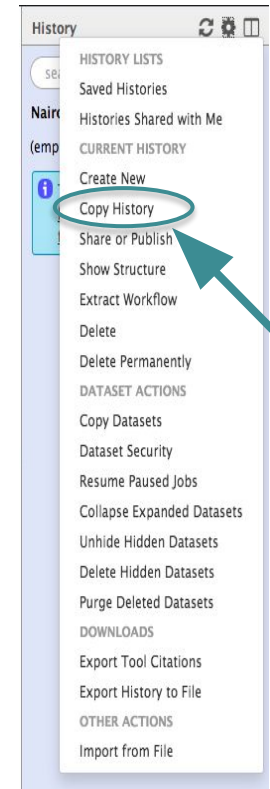
[Choose multiple histories](#)

— OR —

New history named:

Copy History Items

<input type="checkbox"/>	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated ↑	Status
<input type="checkbox"/>	Nairobi - RNAseq	4	0 Tags		14.5 GB	~41 minutes ago	~9 minutes ago	current history



Copying individual datasets

- You can also copy by drag and drop into the current history when in the multiple history column view:

The screenshot shows the Galaxy / VectorBase interface with multiple history columns. The 'Current History' column on the left contains datasets like 'Nairobi - RNAseq' and '4: susceptible 2.fastq'. Other columns show 'PRJNA170440 Deltamethrin resistance' and 'demo 1'. An orange arrow indicates dragging a dataset from 'demo 1' into the 'Current History' column.

The 'History' panel shows a search bar and a list of datasets. The top dataset is 'PRJNA170440 bob testing'. Below it are two highlighted datasets: '37: Cuffdiff on data 3, data 16, and others: transcript FPKM tracking' and '36: Cuffdiff on data 3, data 16, and'. The top-right corner of the panel contains icons for refresh, settings, and a window management icon, which is highlighted by an orange arrow.

Alignments

Alignment & mapping

- We will use the HISAT alignment tool
- Designed to deal with short reads and be splice aware

NGS: Mapping

TopHat for Illumina Find splice junctions using RNA-seq data

TopHat Gapped-read mapper for RNA-seq data

HISAT A fast and sensitive alignment program

Map with BWA - map short reads (< 100 bp) against reference genome

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome

Bowtie2 - map reads against reference genome



Select datasets

- Select the input data format
- Select single/paired-end
- Select reference genome
- default settings...
- Press Execute!

HISAT A fast and sensitive alignment program (Galaxy Version 2.0.3)

Input data format

FASTQ

Single end or paired reads?

Collection of paired reads

Collection of paired reads

Individual paired reads

Individual unpaired reads

Source for the reference genome to align against

Use a built-in genome

Built-in references were created using default options

Select a reference genome

A. gambiae-PEST April. 2016 (VectorBase/AgamP4) (AgamP4)

If your genome of interest is not listed, contact the Galaxy team

Alignment jobs



Successfully invoked workflow RNA-Seq: paired-end FastQ to FPKM [HISAT2].

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

The History pane displays a list of jobs. At the top, there is a search bar labeled 'search datasets'. Below it, the section is titled 'Nairobi - RNAseq' with '7 shown' and a size indicator '14.76 GB'. The jobs are listed in a table-like format with icons for viewing, editing, and deleting each entry.

Job ID	Job Name	View	Edit	Delete
7	HISAT on data 4 and data 3			
6	Anopheles-gambiae- PEST_BASEFEATURES AgamP4.7.qtf			
5	Anopheles-gambiae- PEST_CHROMOSOMES AgamP4.fa			
4	susceptible 2.fastq			
3	susceptible 1.fastq			
2	resistant 2.fastq			
1	resistant 1.fastq			

- Pending/Active jobs are listed on the right-hand side

Alignment jobs

<input type="checkbox"/> <u>Name</u>	<u>Datasets</u>	<u>Tags</u>	<u>Sharing</u>	<u>Size on Disk</u>	<u>Created</u>	<u>Last Updated</u> ↑	<u>Status</u>
<input type="checkbox"/> Nairobi - RNAseq ▾	<div><div>6</div><div>2</div><div>4</div></div>	0 Tags		14.8 GB	~2 hours ago	~1 minute ago	current history

- Back to the Project view summary (User->Saved histories)
- Notice the active jobs in yellow and pending in grey.
- The colours (and numbers) will change as jobs are completed

Running several samples at once

- You can run several files through several tools in "batch" mode using **workflows**
- Your workflows are at the bottom of the Tools list
- System-wide published workflows are in the **Shared Data** menu.

(Note, you can also run multiple files through single tools without using workflows.)

The screenshot displays the Galaxy / VectorBase web interface. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The left sidebar contains a 'Tools' list with various bioinformatics tools, and a 'Workflows' section at the bottom, which is highlighted with a yellow circle and an orange arrow. The main content area shows a 'Welcome to the VectorBase Galaxy Server' message. The right sidebar features a 'History' panel with a search bar and a list of workflows, including '6: Anopheles-gambiae-PEST_CHROMOSOMES_AqamP4.f', '5: Anopheles-gambiae-PEST_BASEFEATURES_AqamP4.3.qtf', '4: susceptible_2.fastq', '3: susceptible_1.fastq', '2: resistant_2.fastq', and '1: resistant_1.fastq'. An orange arrow points to the 'Shared Data' menu in the top navigation bar.

Starting a workflow

- Go to Shared Data->Published Workflows:
- Select and *Import* **"RNA-Seq: paired-end FastQ to FPKM [HISAT2]"** and "start using this workflow"
- Enable "multiple dataset input" with the stacked document icon highlighted yellow, right:
- Select the forward read files (ctrl-click)
- Select the reverse read files (ctrl-click)
- Specify the transcript GTF and reference genome inputs
- Use defaults
- Run workflow...

The screenshot displays the Galaxy / VectorBase web interface. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The main title of the workflow is 'Workflow: RNA-Seq: paired-end FastQ to FPKM [HISAT2]'. A 'Run workflow' button is visible in the top right corner.

The left sidebar contains a 'Tools' section with a search bar and a list of tool categories: Get Data, Send Data, Lift-Over, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, Fetch Sequences, Fetch Alignments, Statistics, Graph/Display Data, NGS: RNA Analysis, Motif Tools, NGS: QC and manipulation, NGS: Mapping, Multiple Alignments, VCF Tools, ND BioApps Tools, NGS: SAM Tools, GATK, Admixture a population structure from large SNP genotype datasets, CCAT Control-based ChIP-seq Analysis Tool, and Workflows. The 'Workflows' section is currently selected, showing 'All workflows'.

The main content area is divided into steps for configuring the workflow:

- History Options:** A section for sending results to a new history, with 'Yes' and 'No' buttons.
- Step 1: Input dataset:** A section for selecting FASTQ files. It shows a list of files: 10: susceptible_2.fastq, 9: susceptible_1.fastq, 8: resistant_2.fastq, and 7: resistant_1.fastq. A yellow circle highlights the 'multiple dataset input' icon (a stacked document icon) next to the file list. Below the list, a note states: 'This is a batch mode input field. Separate jobs will be triggered for each dataset selection.' Batch options are also visible.
- Step 2: Input dataset:** A section for selecting FASTQ files. It shows the same list of files as Step 1. A yellow circle highlights the 'multiple dataset input' icon. A note states: 'This is a batch mode input field. Separate jobs will be triggered for each dataset selection.' Batch options are also visible.
- Step 3: Input dataset:** A section for selecting reference transcripts GTF/GFF3. It shows a dropdown menu with the selected file: 11: Anopheles-gambiae-PEST_BASEFEATURES_AgamP4.7.gtf.
- Step 4: HISAT A fast and sensitive alignment program (Galaxy Version 2.0.3):** A section for selecting the input data format.

The right sidebar shows the 'History' section, which includes a search bar and a list of datasets: demo (6 shown, 6 deleted), 12: Anopheles-gambiae-PEST_CHROMOSOMES_AgamP4.7.gtf, 11: Anopheles-gambiae-PEST_BASEFEATURES_AgamP4.7.gtf, 10: susceptible_2.fastq, 9: susceptible_1.fastq, 8: resistant_2.fastq, and 7: resistant_1.fastq. Each dataset entry has a yellow circle icon and a close button.

Executing a workflow

- Do NOT click/check "Send results to a new history"
"Send results to a new history"
- Then click "Run workflow"

Galaxy / VectorBase | Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 87%

Tools | search tools | Get Data | Send Data | Lift-Over | Text Manipulation | Filter and Sort | Join, Subtract and Group | Convert Formats | Extract Features | Fetch Sequences | Fetch Alignments | Statistics | Graph/Display Data | NGS: RNA Analysis | Motif Tools | NGS: QC and manipulation | NGS: Mapping | Multiple Alignments | VCF Tools | ND BioApps Tools | NGS: SAM Tools | GATK | Admixture a population structure from large SNP genotype datasets | CCAT Control-based CHIP-seq Analysis Tool | Workflows | All workflows

Workflow: RNA-Seq: paired-end FastQ to FPKM [HiSAT2] | Run workflow

History Options

Send results to a new history ☐ Yes ☒ No

Step 1: Input dataset

FASTQ file

10: susceptible_2.fastq
9: susceptible_1.fastq
8: resistant_2.fastq
7: resistant_1.fastq

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Batch options:

Step 2: Input dataset

FASTQ file

10: susceptible_2.fastq
9: susceptible_1.fastq
8: resistant_2.fastq
7: resistant_1.fastq

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Batch options:

Step 3: Input dataset

Reference transcripts GTF/GFF3

11: Anopheles-gambiae-PEST_BASEFEATURES_AgamP4.7.gtf

Step 4: HISAT A fast and sensitive alignment program (Galaxy Version 2.0.3)

Input data format

History | search datasets | demo | 6 shown, 6 deleted | 14.76 GB | 12: Anopheles-gambiae-PEST_CHROMOSOMES_AgamP4.7.gtf | 11: Anopheles-gambiae-PEST_BASEFEATURES_AgamP4.7.gtf | 10: susceptible_2.fastq | 9: susceptible_1.fastq | 8: resistant_2.fastq | 7: resistant_1.fastq

Advantages of workflows

- Less clicking, so less can go wrong
- Workflows (like the one demonstrated here) can be configured to **rename output files** so they are less confusing
- Workflows can also hide unwanted output files (some tools make many files) to reduce clutter and confusion

Add a cuffdiff step

- cuffdiff will report differential expression between samples groups by two or more conditions
- Our example only has one biological replicate* per input file - ideally you would have two or more biological replicates for improved statistical robustness
- *You can start Galaxy jobs to work on files that haven't finished being made yet*
- In the history where the workflow is running, choose the cuffdiff tool and configure as in the next slide...

Tools

search tools

Get Data

Send Data

Lift-Over

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Statistics

Graph/Display Data

NGS: RNA Analysis

[Cuffdiff](#) find significant changes in transcript expression, splicing, and promoter use

[Trinity](#) De novo assembly of RNA-Seq data Using Trinity

[Cufflinks](#) transcript assembly and FPKM (RPKM) estimates for RNA-Seq data

Motif Tools

NGS: QC and manipulation

NGS: Mapping

Multiple Alignments

VCF Tools

ND BioApps Tools

NGS: SAM Tools

GATK

[Admixture](#) a population structure from large SNP genotype datasets

[CCAT](#) Control-based ChIP-seq Analysis Tool

Workflows

- All workflows

Cuffdiff find significant changes in transcript expression, splicing, and promoter use (Galaxy Version 2.2.1.3)

Options

Transcripts

1: Anopheles-gambiae-PEST_BASEFEATURES_AgamP4.7.gtf

A transcript GFF3 or GTF file produced by cufflinks, cuffcompare, or other source.

Omit Tabular Datasets

Yes No

Discard the tabular output.

Generate SQLite

Yes No

Generate a SQLite database for use with cummeRbund.

Input data type

SAM/BAM

CuffNorm supports either CXB (from cuffquant) or SAM/BAM input files. Mixing is not supported. Default: SAM/BAM

Condition

1: Condition

Name

resistant

Replicates

16: resistant_1.fastq.bam
11: susceptible_1.fastq.bam

2: Condition

Name

susceptible

Replicates

16: resistant_1.fastq.bam
11: susceptible_1.fastq.bam

+ Insert Condition

Library normalization method

geometric

Dispersion estimation method

blind

If using only one sample per condition, you must use 'blind.'

History

search datasets

demo workflow

24 shown, 2 hidden

7.5 GB

26: Cufflinks on data

1 and data 16: Skipped

Transcripts

24: Cufflinks on data

1 and data 16:

assembled transcripts

23:

resistant_1.fastq.bam.trans fpkm

22:

resistant_1.fastq.bam.genes fpkm

21: Cufflinks on data

1 and data 11: Skipped

Transcripts

19: Cufflinks on data

1 and data 11:

assembled transcripts

18:

susceptible_1.fastq.bam.trans fpkm

17:

susceptible_1.fastq.bam.genes fpkm

16:

resistant_1.fastq.bam

15: TopHat on data

2, data 4, and data 3:

splice junctions

14: TopHat on data

2, data 4, and data 3:

deletions

13: TopHat on data

2, data 4, and data 3:

insertions

RNAseq Analysis

Part II - Expression values

Alignment jobs

<input type="checkbox"/>	Recife - run ▾	23	0 Tags	42.5 GB	~7 hours ago	~41 minutes ago	current history
--------------------------	----------------	----	------------------------	---------	--------------	-----------------	------------------------

- Check that all the jobs are complete

Expression metrics

Cufflinks

- We used the cufflinks tool to assign expression levels based on the VB-annotated transcripts
- Requires both alignments (BAM) and annotation (GTF) to calculate FPKM expression values for each locus
- FPKM (fragments per kilobase of exon per million fragments mapped)

NCS: RNA-seq

Cuffmerge merge together several Cufflinks assemblies

Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data

Cuffdiff find significant changes in transcript expression, splicing, and promoter use

Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments

Tophat2 Gapped-read mapper for RNA-seq data

Tophat Fusion Post post-processing to identify fusion genes

Tophat for Illumina Find splice junctions using RNA-seq data

Filter Combined Transcripts using tracking file



Expression values - cufflinks

7: Cufflinks Eukaryotic on data 6: gene expression

13,465 lines
format: tabular, database:
Anopheles_gambiae_AgamP3
cufflinks v2.1.1 cufflinks -q --no-update-check -l 300000 -F 0.100000 -j 0.150000 -p 4 -G /rnrrocket/indices/genomes/pathogen_portal_inplace_updates/pathogen_portal_indices/vectorbase/Anopheles_gambiae_AgamP3/Anopheles-gambiae-PEST_BASEFEATURES_AgamP3.7.

1	2	3
tracking_id	class_code	nearest_ref_id
AGAP004680	-	-
AGAP004681	-	-
AGAP004678	-	-
AGAP004682	-	-
AGAP004683	-	-

8: Cufflinks Eukaryotic on data 6: transcript expression

15,315 lines
format: tabular, database:
Anopheles_gambiae_AgamP3
cufflinks v2.1.1 cufflinks -q --no-update-check -l 300000 -F 0.100000 -j 0.150000 -p 4 -G /rnrrocket/indices/genomes/pathogen_portal_inplace_updates/pathogen_portal_indices/vectorbase/Anopheles_gambiae_AgamP3/Anopheles-gambiae-PEST_BASEFEATURES_AgamP3.7.

1	2	3
tracking_id	class_code	nearest_ref_id
AGAP004680-RA	-	-
AGAP004681-RA	-	-
AGAP004678-RA	-	-
AGAP004682-RA	-	-
AGAP004683-RA	-	-

9: Cufflinks Eukaryotic on data 6: assembled transcripts




81,418 lines
format: gtf, database:
Anopheles_gambiae_AgamP3
cufflinks v2.1.1 cufflinks -q --no-update-check -l 300000 -F 0.100000 -j 0.150000 -p 4 -G /rnrrocket/indices/genomes/pathogen_portal_inplace_updates/pathogen_portal_indices/vectorbase/Anopheles_gambiae_AgamP3/Anopheles-gambiae-PEST_BASEFEATURES_AgamP3.7.

display at [vectorbase](#)

1. Seqname	2. Source	3. Feature	4. Start
2L	Cufflinks	transcript	2712
2L	Cufflinks	exon	2712
2L	Cufflinks	transcript	3583
2L	Cufflinks	exon	3583
2L	Cufflinks	exon	3589
2L	Cufflinks	transcript	2037






- Calculate expression values for genes and transcripts
- Reconstruct transcripts from aligned RNAseq data

Viewing results

7: Cufflinks Eukaryotic on   

data 6: gene expression

13,465 lines
format: tabular, database:
Anopheles_gambiae_AgamP3
cufflinks v2.1.1 cufflinks -q --no-
update-check -l 300000 -F 0.100000
-j 0.150000 -p 4 -G
/rnavarocket/indices/genomes/pathoge
n_portal_inplace_updates/pathogen_p
ortal_indices/vectorbase/Anopheles_g
ambiae_AgamP3/Anopheles-gambiae-
PEST_BASEFEATURES_AgamP3.7.

1	2	3
tracking_id	class_code	nearest_ref_id
AGAP004680	-	-
AGAP004681	-	-
AGAP004678	-	-
AGAP004682	-	-
AGAP004683	-	-

View the results as text/table

View details about how this file was generated, e.g. tool, input files and parameters

Expression values - cufflinks

tracking_id	class_code	nearest_ref_id	gene_id	gene_short_name	tss_id	locus	length	coverage	FPKM	FPKM_conf_lo	FPKM_conf_hi	FPKM_status
AGAP004681	-	-	AGAP004681	-	-	2L:358328-359280	-	-	0	0	0	OK
AGAP004679	-	-	AGAP004679	-	-	2L:207893-210460	-	-	13.0137	11.8069	14.2224	OK
AGAP004678	-	-	AGAP004678	-	-	2L:203778-205293	-	-	10.6099	8.01572	10.4124	OK
AGAP004680	-	-	AGAP004680	-	-	2L:271284-271815	-	-	0.54089	0.073170	0.95121	OK
AGAP004682	-	-	AGAP004682	-	-	2L:433502-461627	-	-	18.0336	15.0316	18.3678	OK
AGAP004684	-	-	AGAP004684	-	-	2L:493038-493543	-	-	34.6278	13.5679	20.2638	OK
AGAP004683	-	-	AGAP004683	-	-	2L:485697-488369	-	-	7.38985	6.03467	7.67734	OK
AGAP004685	-	-	AGAP004685	-	-	2L:493578-497632	-	-	7.77028	7.08223	8.45834	OK
AGAP004687	-	-	AGAP004687	-	-	2L:819112-819301	-	-	0	0	0	OK
AGAP004677	-	-	AGAP004677	-	-	2L:157347-186936	-	-	91.6806	86.9678	96.6903	OK

Expression values - cuffdiff

- Test for differential expression between two or more conditions (e.g. resistant and susceptible)
- Requires:
 - tophat alignments of reads to reference genome
 - GTF/GFF3 file of transcript annotations

NGS: RNA-seq

Cuffmerge merge together several Cufflinks assemblies

Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data

Cuffdiff find significant changes in transcript expression, splicing, and promoter use

Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments

Tophat2 Gapped-read mapper for RNA-seq data

Tophat Fusion Post post-processing to identify fusion genes

Tophat for Illumina Find splice junctions using RNA-seq data

Filter Combined Transcripts using tracking file



Expression values - cuffdiff - transcript-level

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
AGAP000002-RA	AGAP000002	-	X:581-16387	Resistant	Susceptible	OK	17.8052	14.1507	-0.331424	-0.730082	0.4928	0.997833	no
AGAP000005-RA	AGAP000005	-	X:32381-38843	Resistant	Susceptible	OK	73.7774	53.4476	-0.465054	-1.04785	0.3582	0.997833	no
AGAP000007-RA	AGAP000007	-	X:83816-88773	Resistant	Susceptible	OK	27.4516	25.6688	-0.0968771	-0.197951	0.8491	0.997833	no
AGAP000008-RA	AGAP000008	-	X:90141-94903	Resistant	Susceptible	OK	88.2536	58.5792	-0.591267	-1.2572	0.25845	0.997833	no
AGAP000009-RA	AGAP000009	-	X:97669-114021	Resistant	Susceptible	OK	12.3586	16.7825	0.441443	0.447786	0.65835	0.997833	no
AGAP000009-RB	AGAP000009	-	X:97669-114021	Resistant	Susceptible	OK	21.8111	12.952	-0.751885	-0.760384	0.4439	0.997833	no
AGAP000009-RC	AGAP000009	-	X:97669-114021	Resistant	Susceptible	OK	18.4501	18.884	0.0335399	0.0380026	0.9681	0.997833	no
AGAP000010-RA	AGAP000010	-	X:120772-123499	Resistant	Susceptible	OK	13.1417	10.8384	-0.277997	-0.651008	0.58325	0.997833	no
AGAP001707-RA	AGAP001707	-	2R:8840694-8843819	Resistant	Susceptible	OK	8.9113	77.5087	3.12065	7.1511	5e-05	0.00985509	yes

Filter and extract significant gene/transcript IDs

- View your "Cuffdiff ... transcript differential expression testing" output (as in previous slide)
- Note that column 14 (significant) contains "yes" or "no"
- Start the tool "Filter and Sort->Filter"
- Select the "Cuffdiff ... transcript differential expression testing" dataset as input
- Filter condition: `c14=='yes'`
- Header lines: 1
- That job should run very quickly
- Now set up a "Text Manipulation->Cut" job to cut columns "c1,c2" from the filtered output of the previous step.
- You should end up with something looking like the screenshot →

1	2
test_id	gene_id
AGAP000047-RA	AGAP000047
AGAP000820-RA	AGAP000820
AGAP001376-RA	AGAP001376
AGAP001707-RA	AGAP001707
AGAP001969-RA	AGAP001969
AGAP002425-RA	AGAP002425
AGAP002442-RA	AGAP002442
AGAP002557-RA	AGAP002557
AGAP003095-RA	AGAP003095
AGAP003247-RA	AGAP003247
AGAP003251-RA	AGAP003251
AGAP003308-RB	AGAP003308
AGAP003691-RA	AGAP003691
AGAP003738-RA	AGAP003738
AGAP003765-RA	AGAP003765
AGAP003841-RA	AGAP003841
AGAP004581-RA	AGAP004581
AGAP004583-RA	AGAP004583
AGAP004794-RA	AGAP004794
AGAP004847-RA	AGAP004847

Find gene annotations in BioMart

So what do these genes do?

1	2
test_id	gene_id
AGAP000047-RA	AGAP000047
AGAP000820-RA	AGAP000820
AGAP001376-RA	AGAP001376
AGAP001707-RA	AGAP001707
AGAP001969-RA	AGAP001969
AGAP002425-RA	AGAP002425
AGAP002442-RA	AGAP002442
AGAP002557-RA	AGAP002557
AGAP003095-RA	AGAP003095
AGAP003247-RA	AGAP003247
AGAP003251-RA	AGAP003251
AGAP003308-RB	AGAP003308
AGAP003691-RA	AGAP003691
AGAP003738-RA	AGAP003738
AGAP003765-RA	AGAP003765
AGAP003841-RA	AGAP003841
AGAP004581-RA	AGAP004581
AGAP004583-RA	AGAP004583
AGAP004794-RA	AGAP004794
AGAP004847-RA	AGAP004847

- Copy-paste the transcript and gene IDs into the Gene->ID list limit filter of BioMart (Anopheles gambiae genes)
- Choose the attributes you want displayed - gene name, gene description, GO terms and InterPro domains are useful.

You could also paste the gene and transcript IDs into the *Anopheles gambiae* **Expression Map** gene search - to see how the differentially expressed genes from this experiment behaved in other experiments.

What next?

You've now seen that running CPU-expensive NGS analysis pipelines in Galaxy is relatively simple for the non-expert.

Genomic variation, ChIP-Seq and many more analyses are possible in Galaxy.

Always seek advice on using the correct tools and parameters.

Always do quality control and sanity checks!

Contact the VectorBase helpdesk if a tool you would like to use is not available, or if you need any other assistance.

How to search for more information or help?

E-mail us at
info@vectorbase.org