Galaxy

Gloria I. Giraldo-Calderón
September 2017

VectorBase
Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

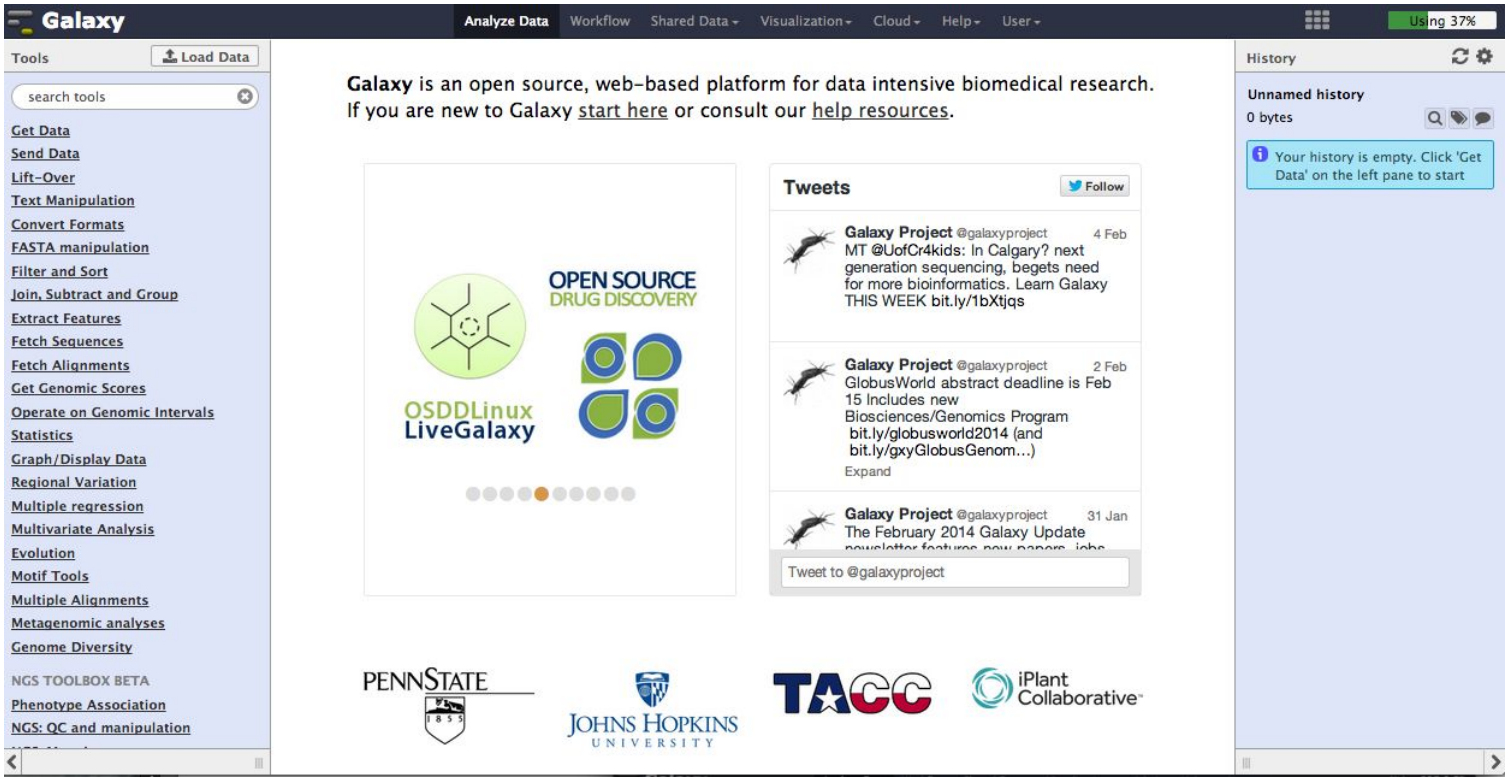
Outline

- Introduction to Galaxy
- RNAseq analysis

Galaxy Overview

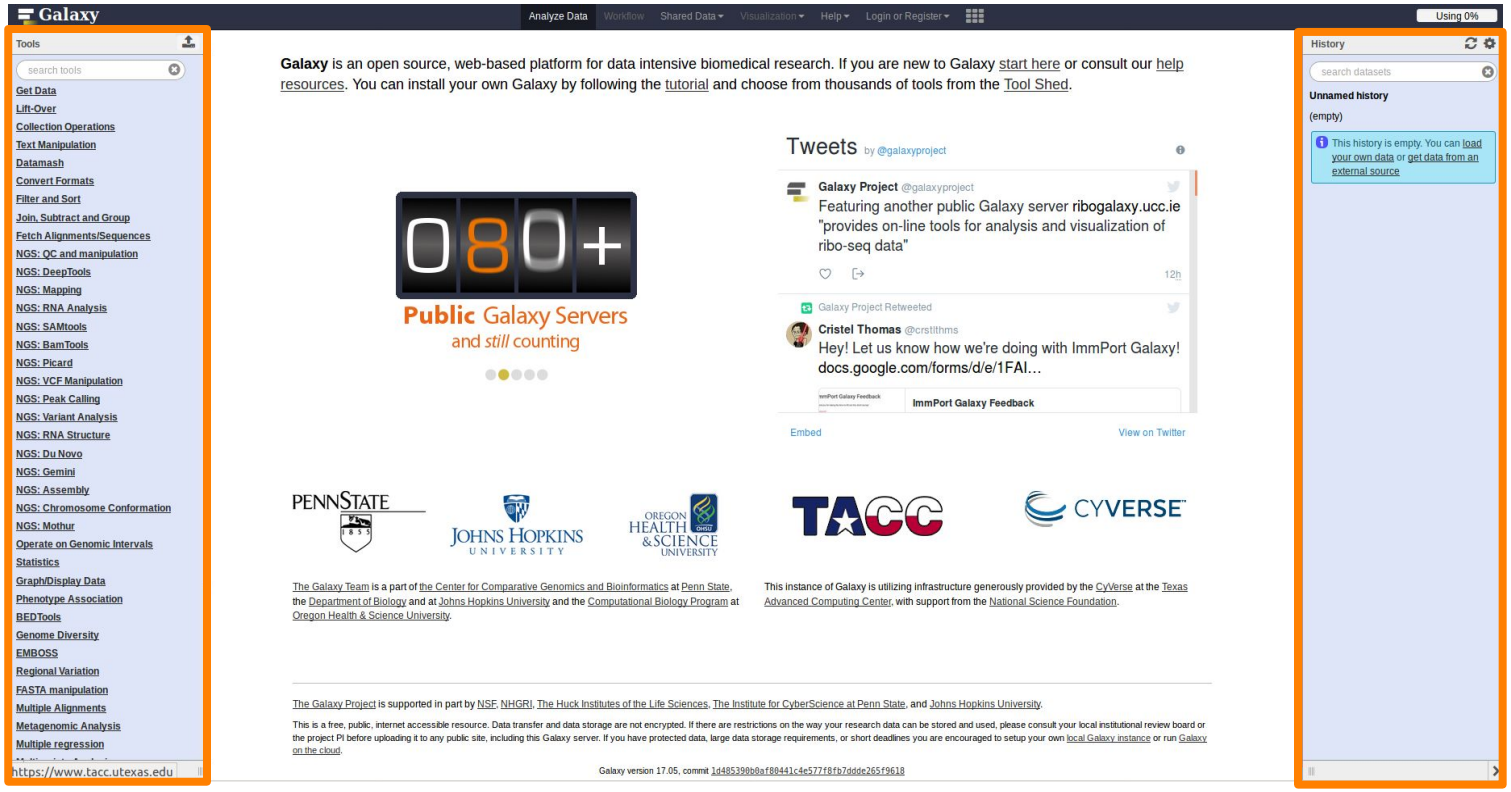
- Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research.
  - *Accessibility*: Galaxy enables users without programming experience to easily specify parameters and run tools and workflows.
  - *Reproducibility*: Galaxy captures all information 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



Galaxy homepage

List of available tools (actions)

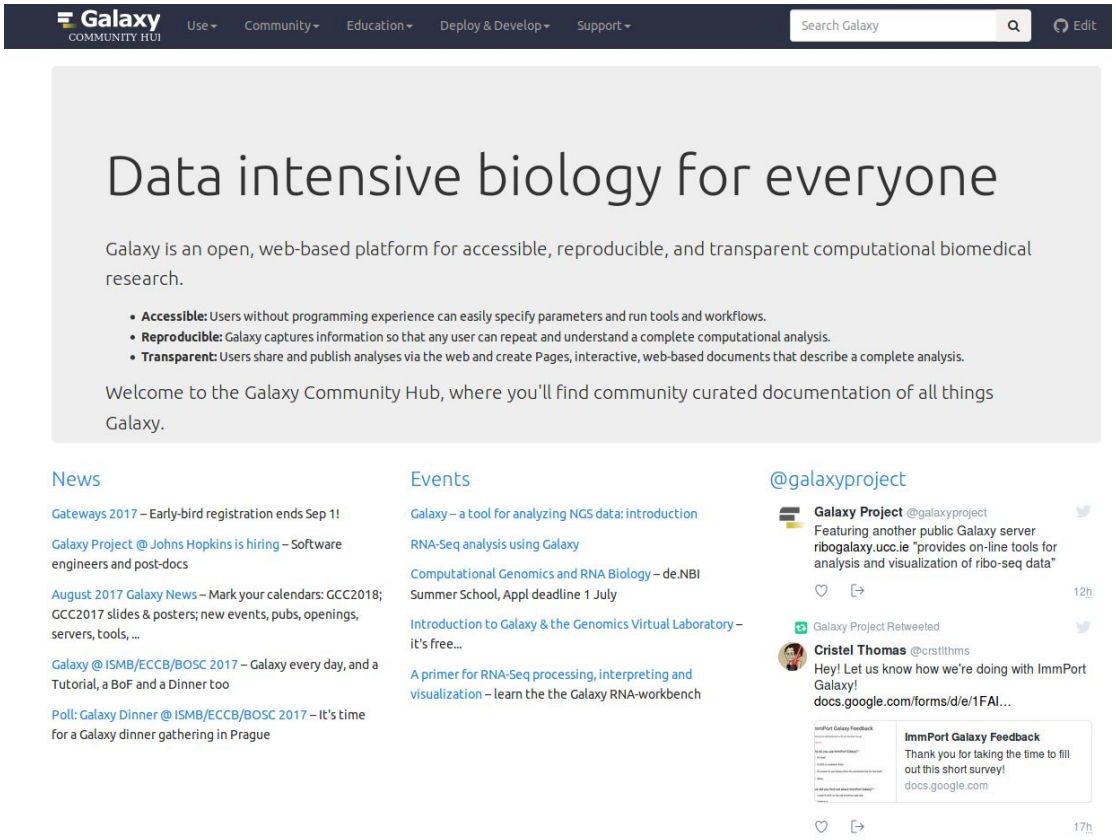


History of data sets and activity

Main panel - shows options/configuration and displays output

Useful resources: Galaxy Wiki

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



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

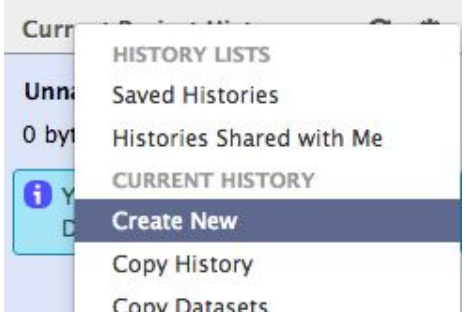
RNAseq Analysis
Part I - Mapping reads

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

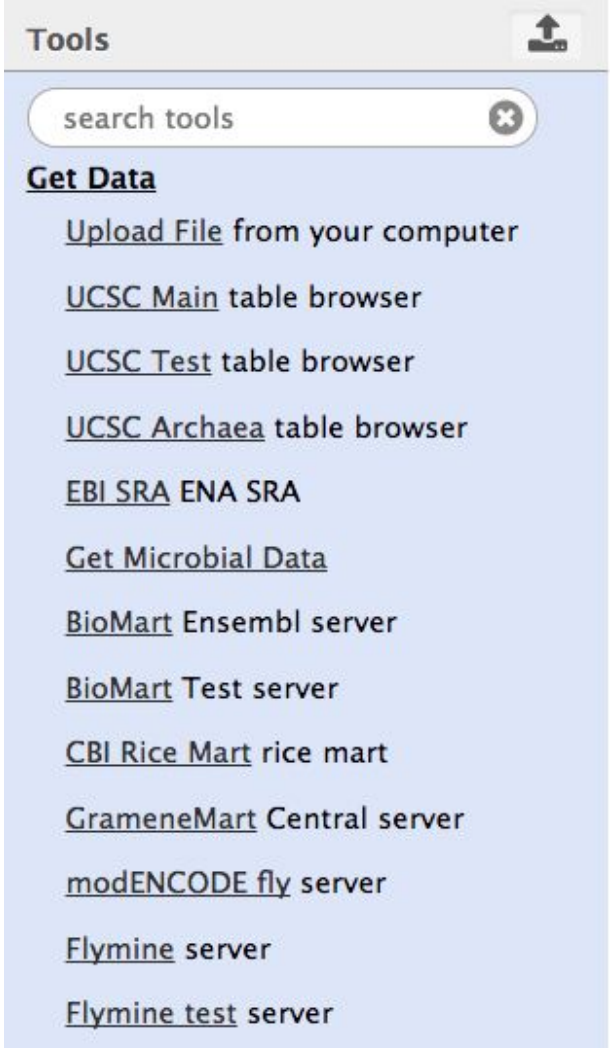


- Change the name, to something like "my first history"

  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

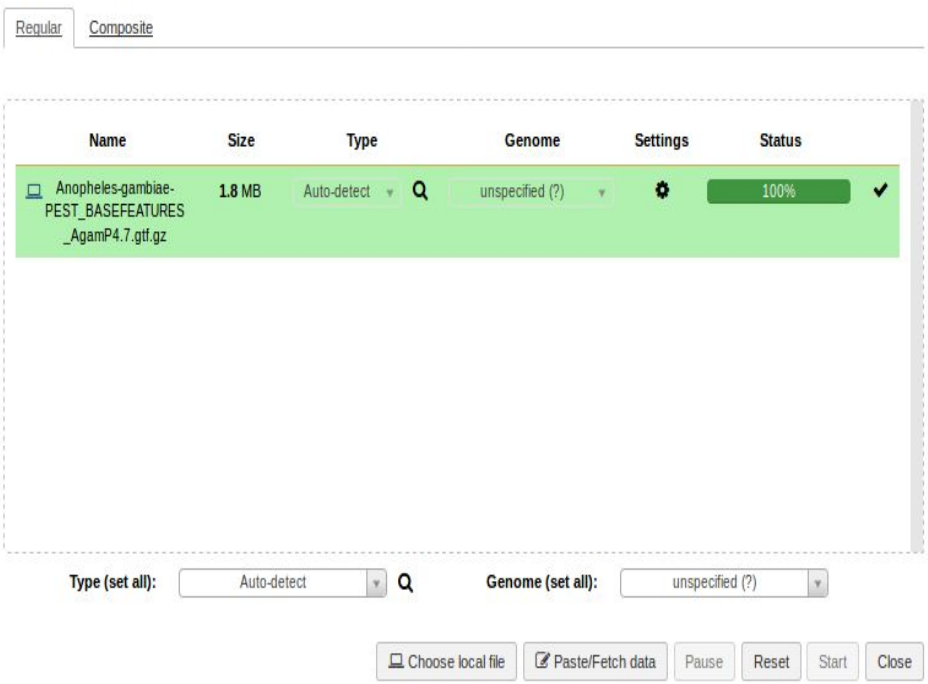
Download a file from VectorBase **Downloads->Data files** as follows:

- 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



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

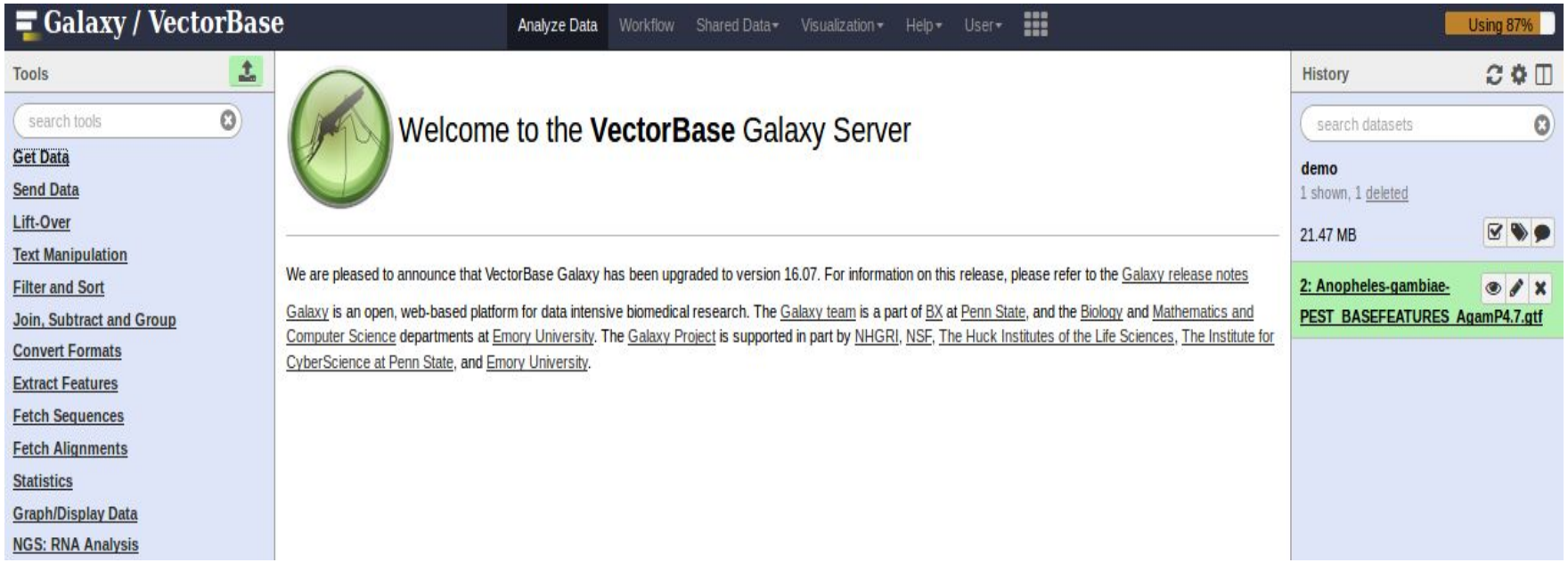
Don't forget to click Start...

Files will be loaded into the current history

Choose the file format or let Galaxy auto-detect

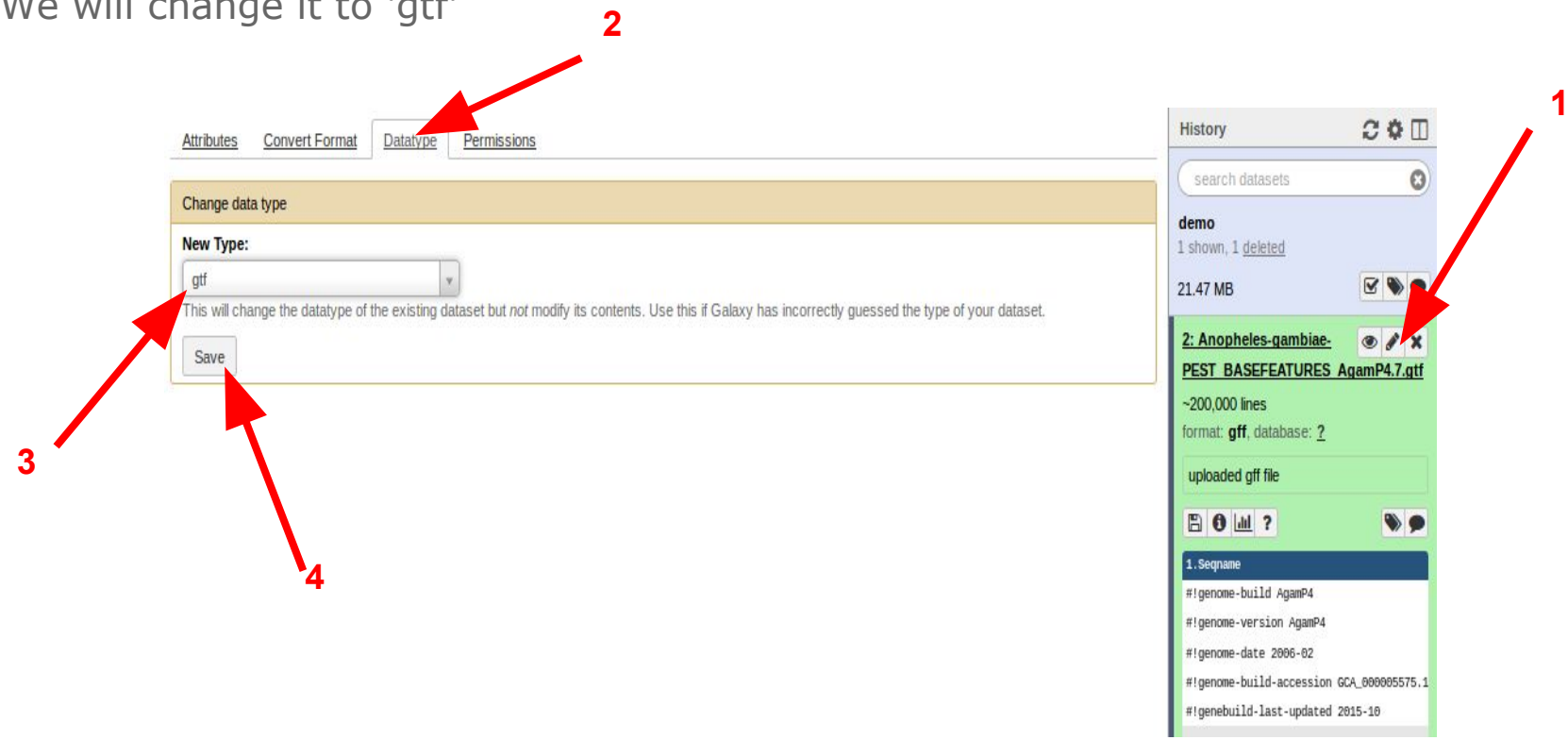
Uploading data from file

- Confirmation of upload job



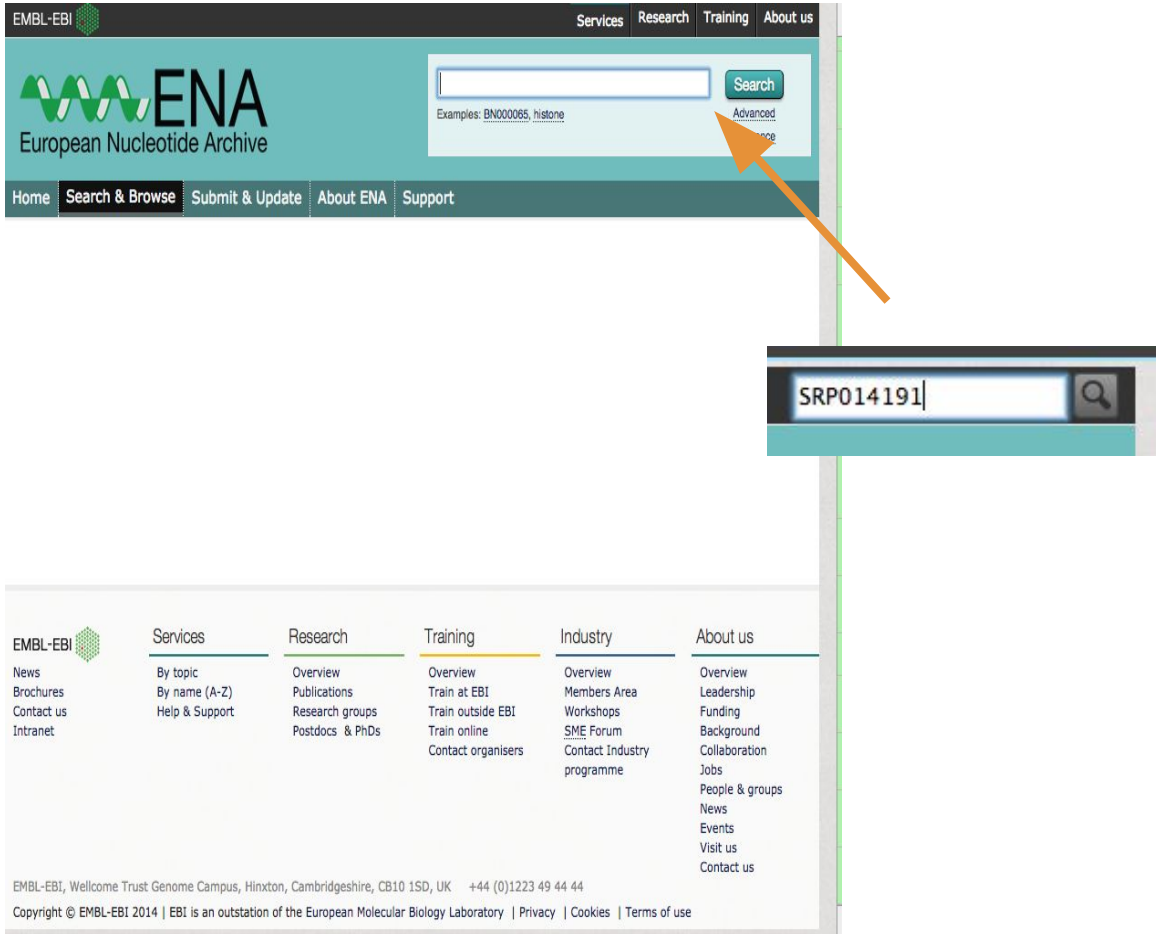
Edit file attributes/format

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

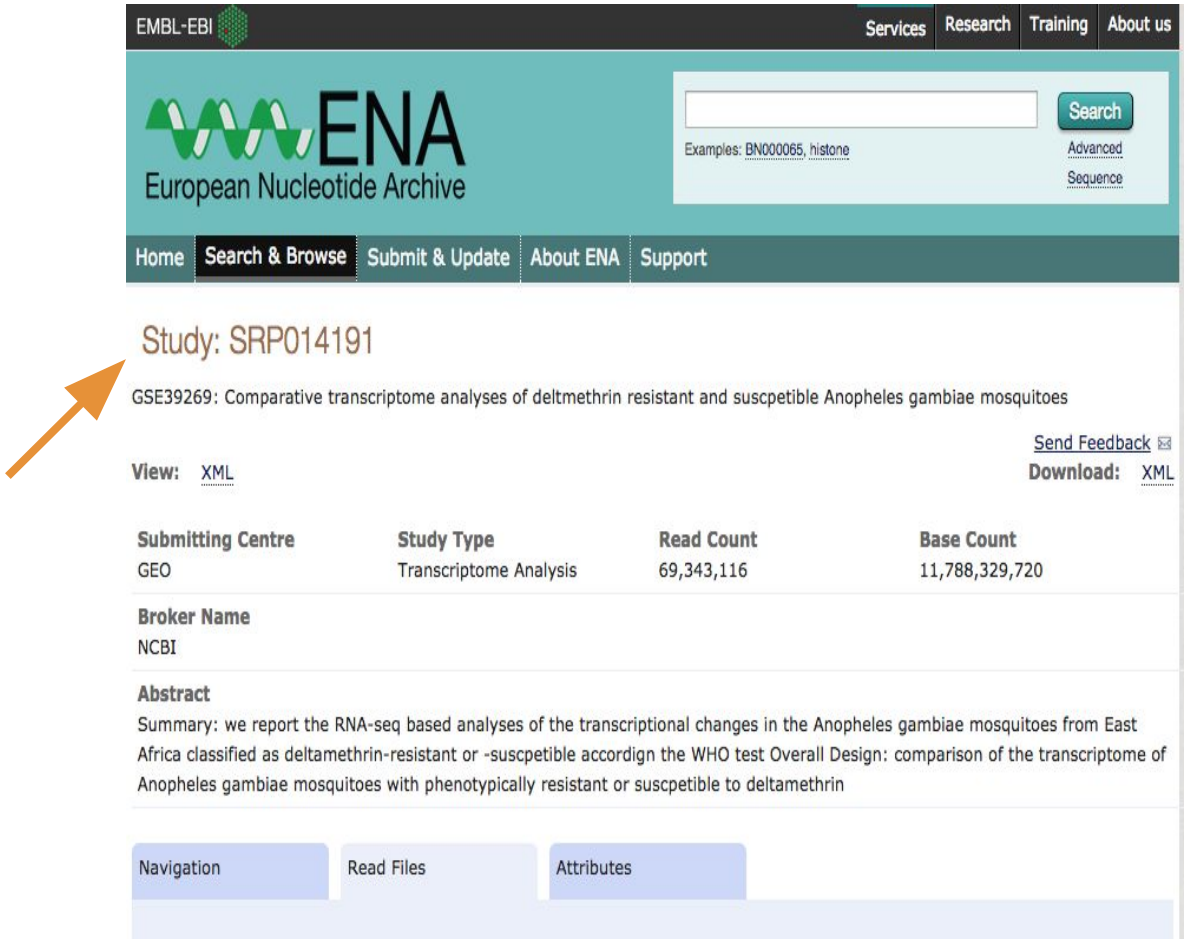


Uploading data from SRA

- 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



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

Navigation | Read Files | Attributes

▶ Download files

View: TEXT

Download: TEXT

Select columns

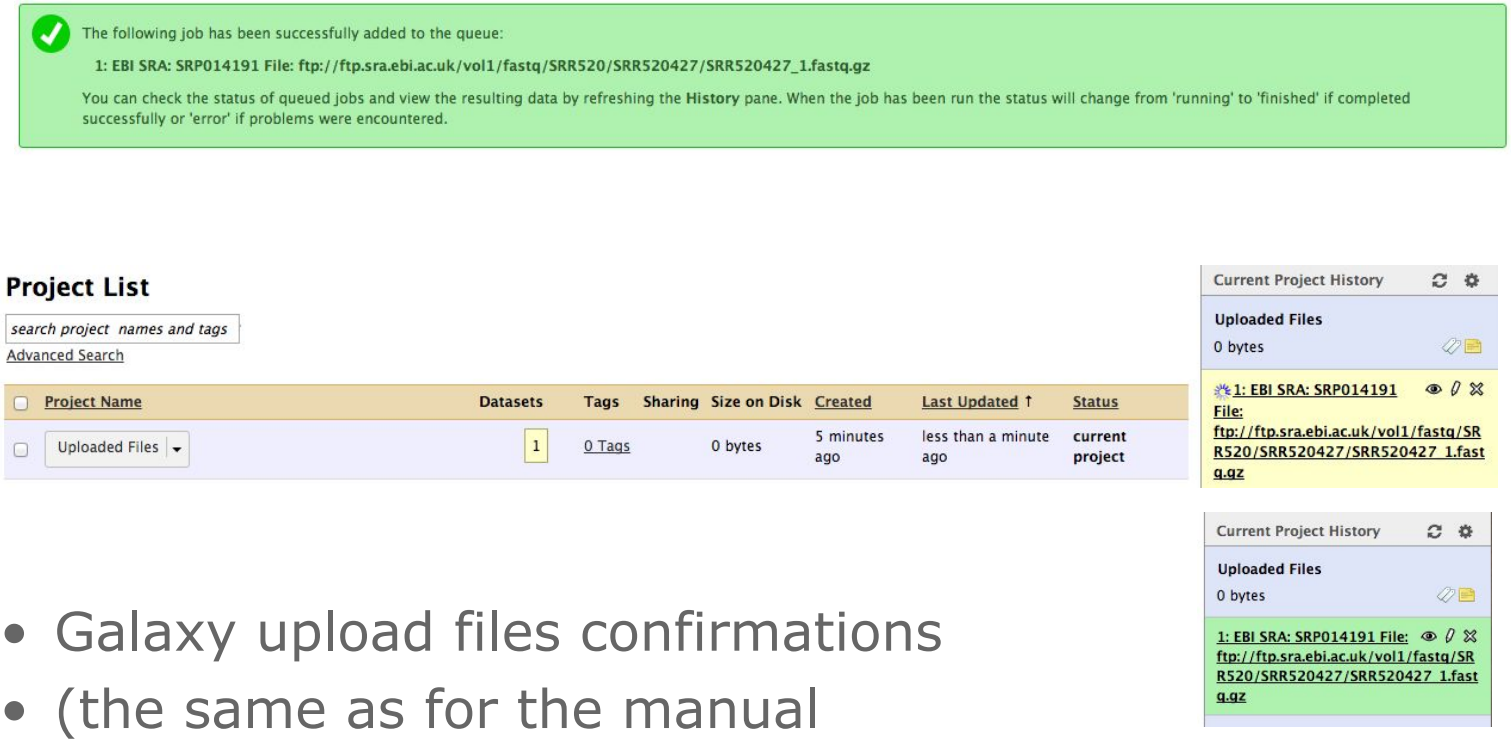
Showing results 1 - 2 of 2 results

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

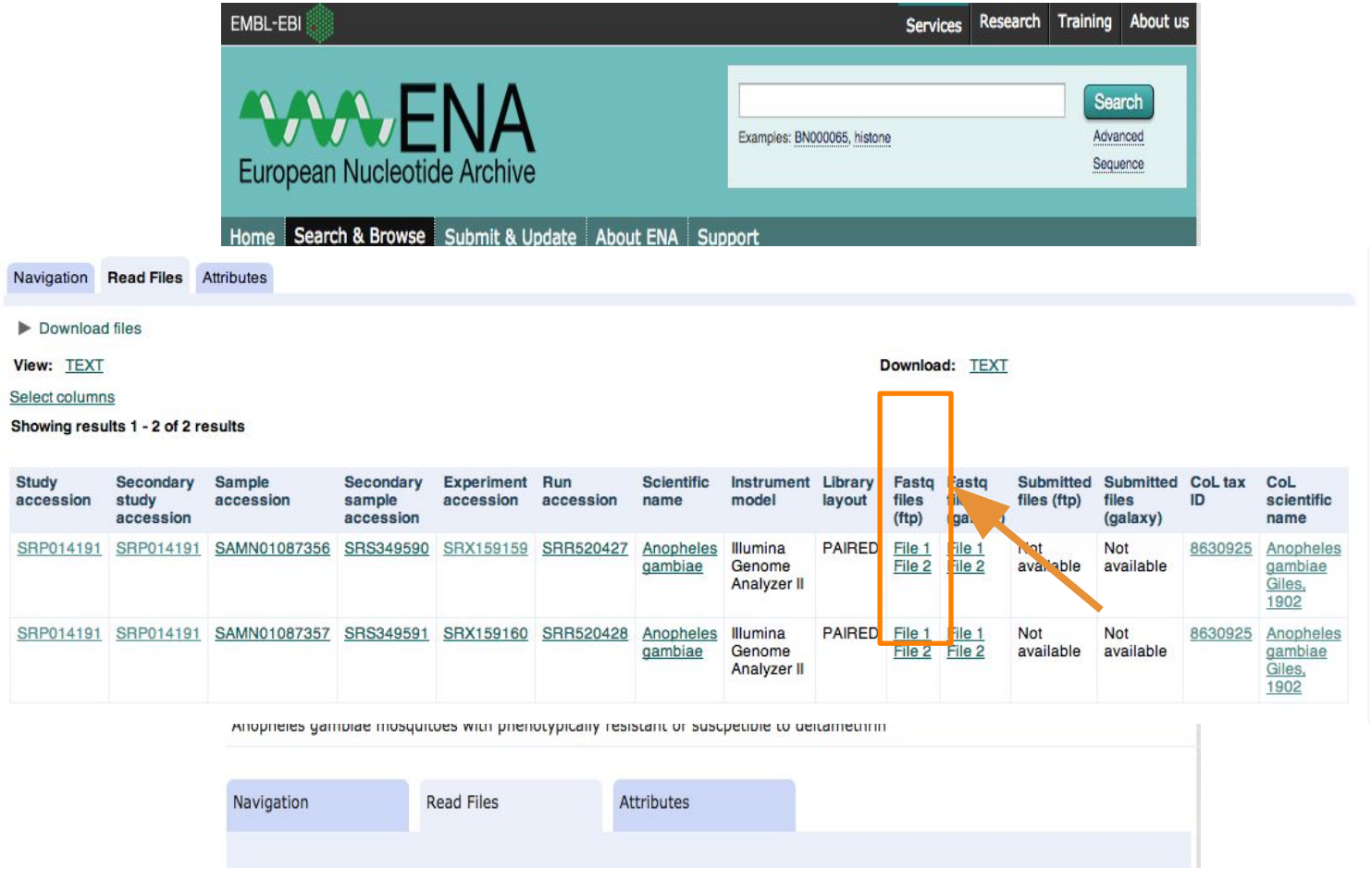
Navigation | Read Files | Attributes

Uploading data from SRA

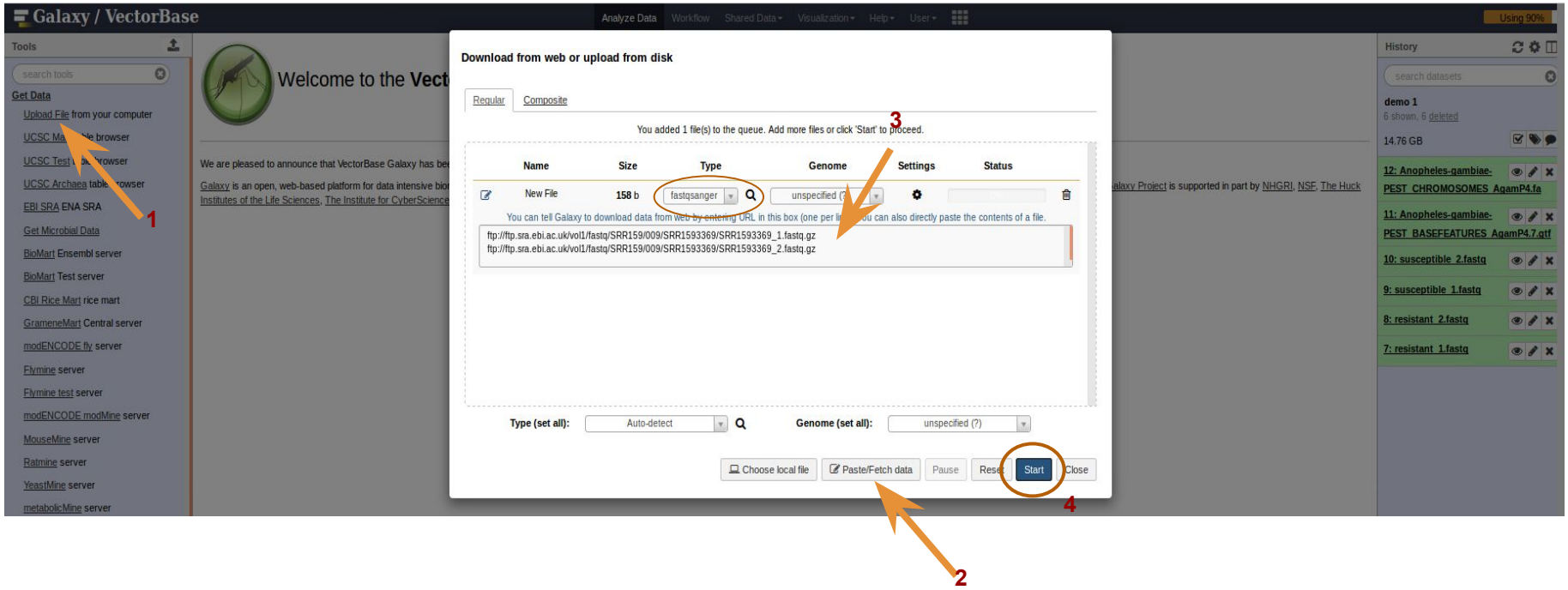


- Galaxy upload files confirmations
- (the same as for the manual uploads)

Uploading data using SRA fastq links



Uploading data from SRA url links

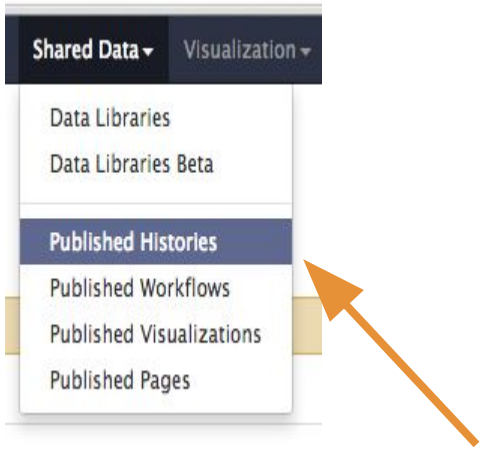


Copying data from published history

- 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

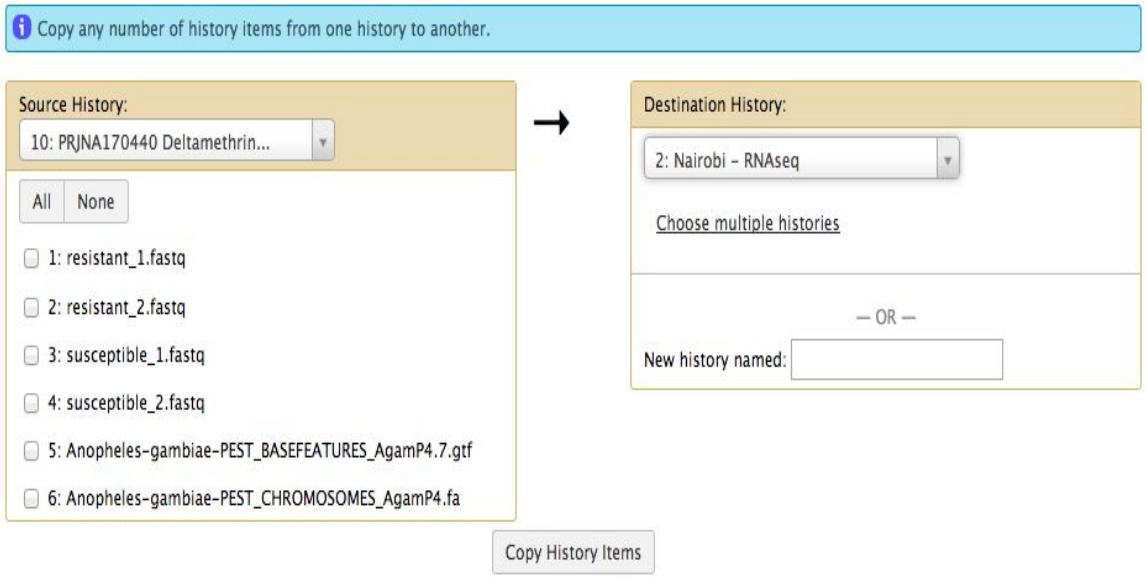
  •Click on the "Import history" link, top right

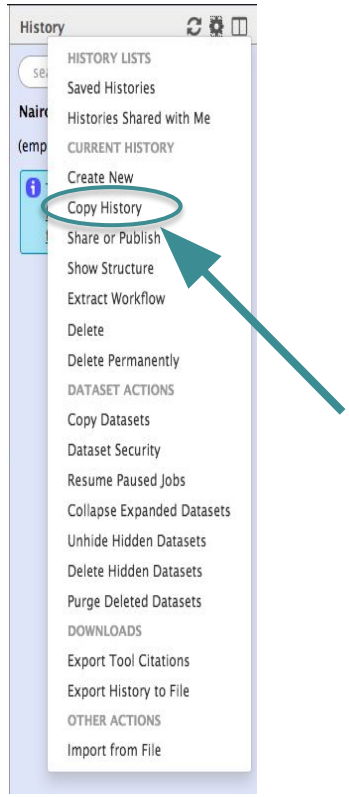
  •Rename your imported history something like "Workshop RNA-Seq analysis"



Copying individual datasets

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

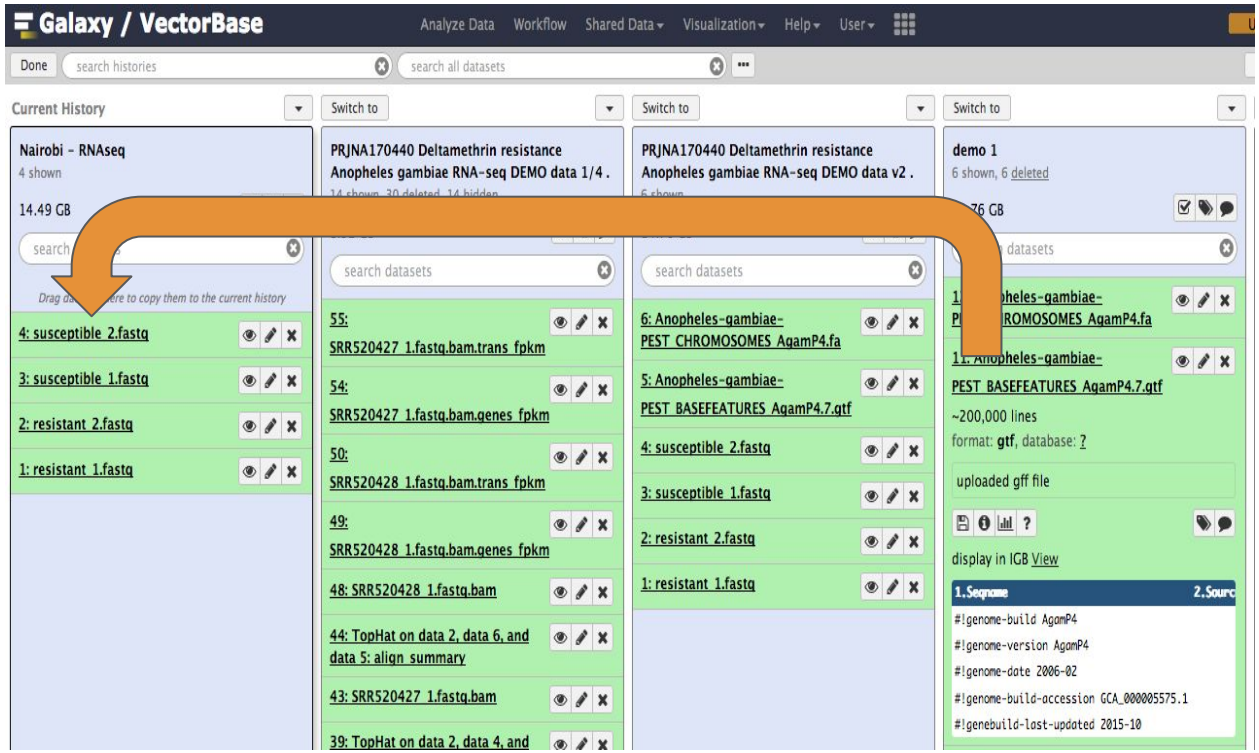


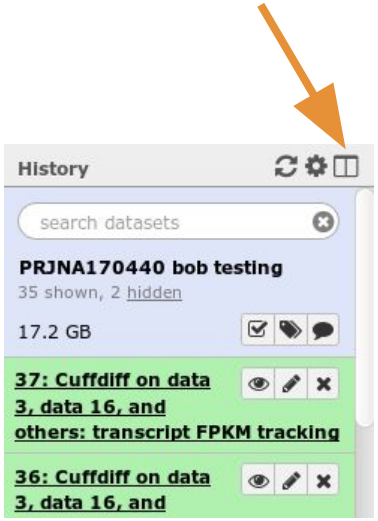


Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated↑	Status
Nairobi - RNAseq	4	0 Tags		14.5 GB	~41 minutes ago	~9 minutes ago	current history

Copying individual datasets

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

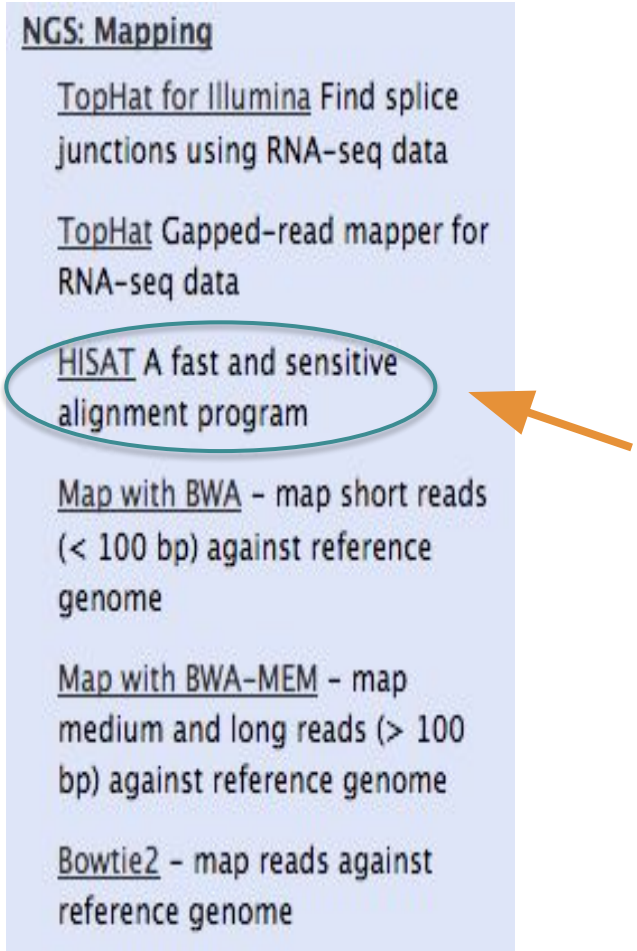




Alignments

Alignment & mapping

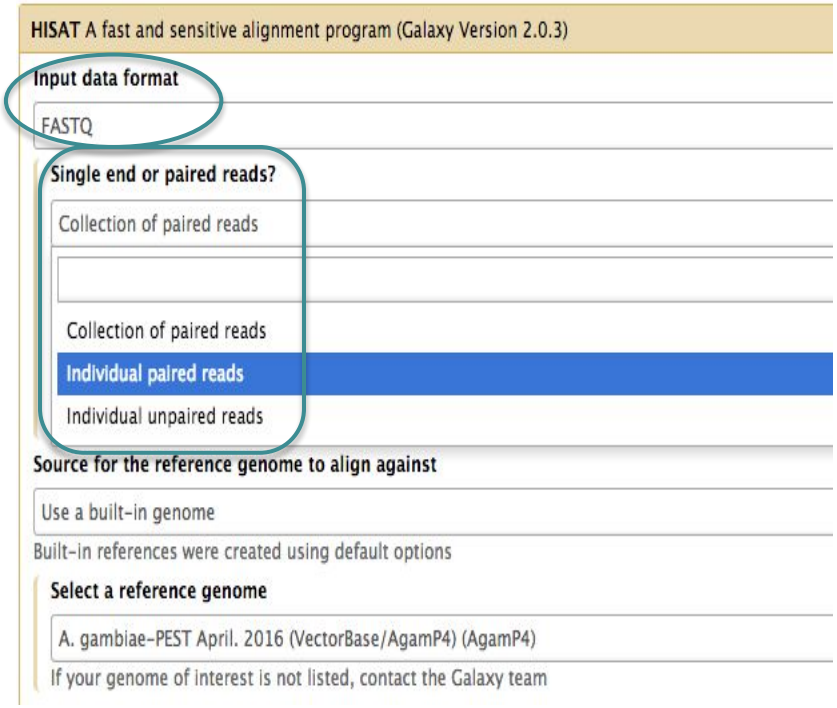
- We will use the HISAT alignment tool
- Designed to deal with short reads and be splice aware



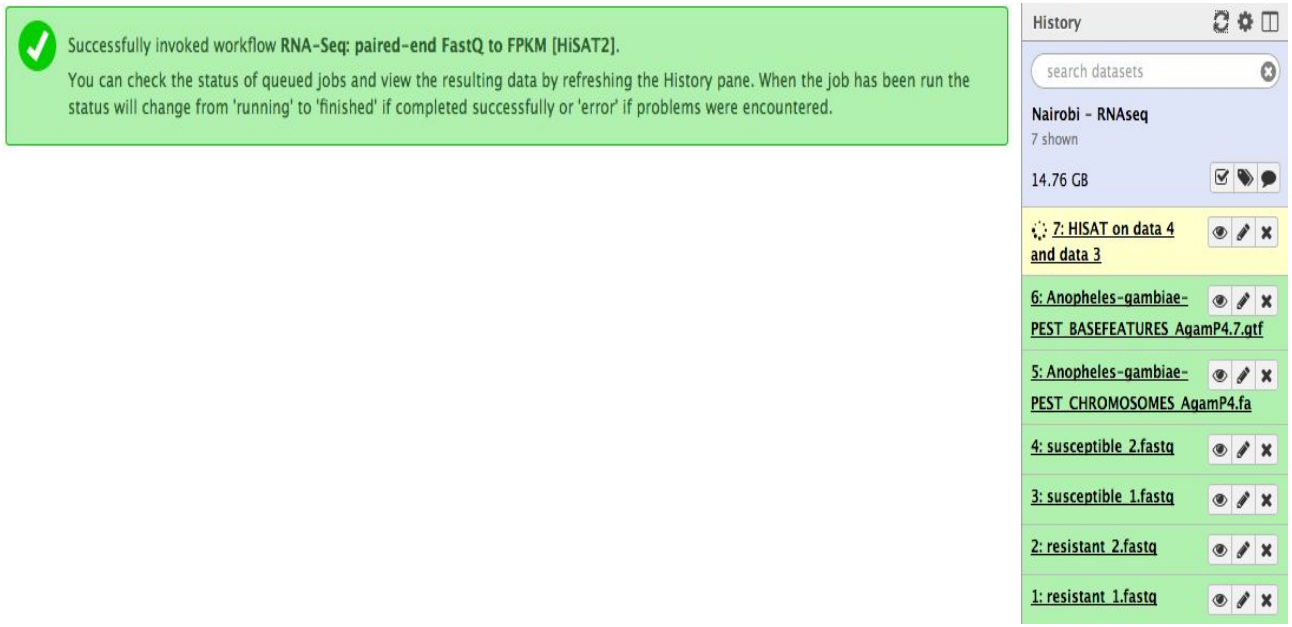
Select datasets

- Select the input data format
- Select single/paired-end
- Select reference genome
- default settings…

- Press Execute!



Alignment jobs



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

Alignment jobs

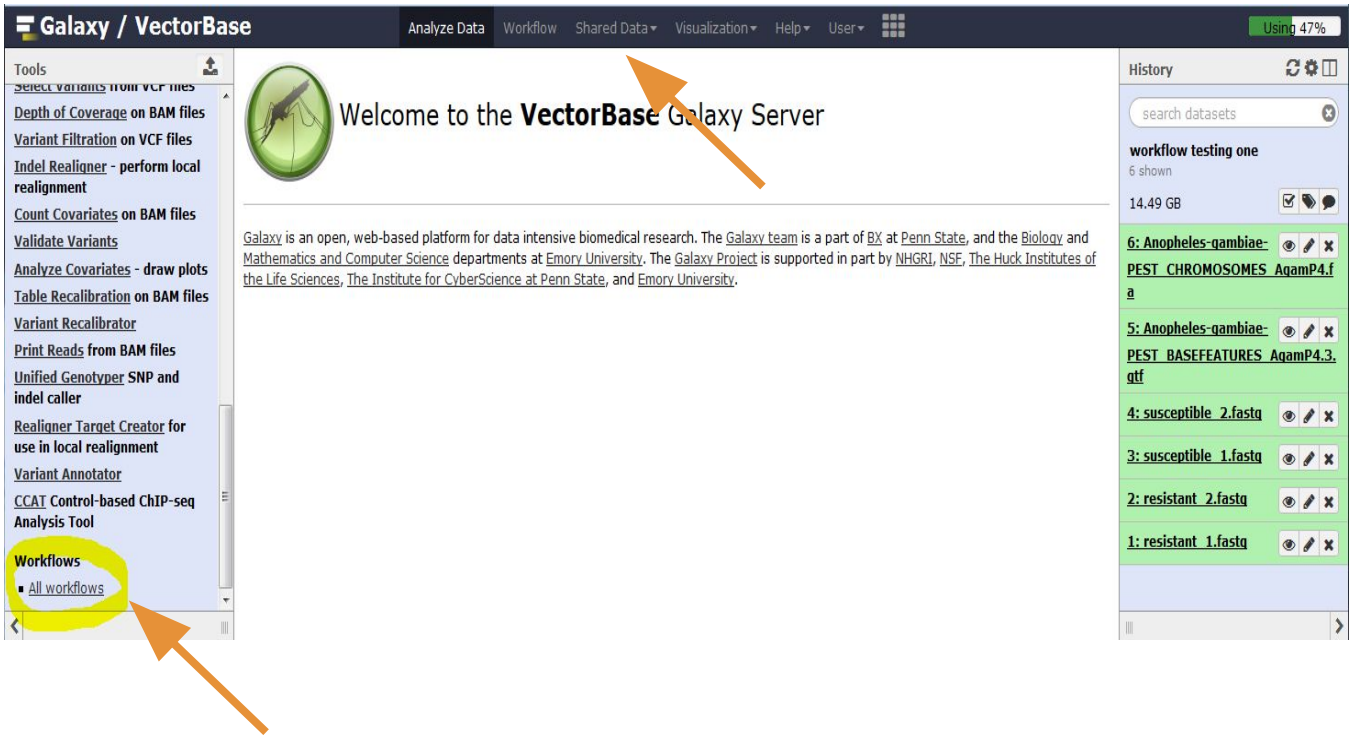
Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated↑	Status
Nairobi - RNAseq	6 2 4	0 Tags		14.8 GB	~2 hours ago	~1 minute ago	current history

- 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

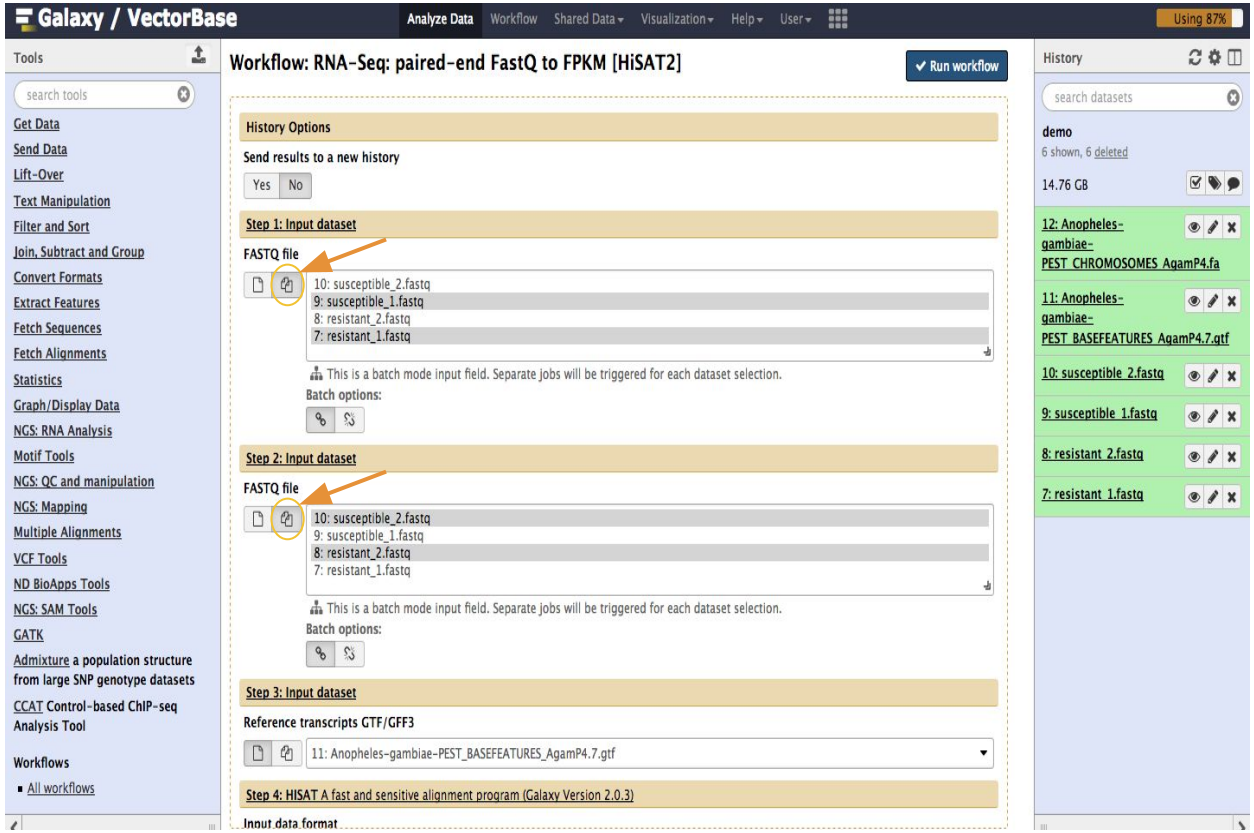
- You can run several files through several tools in "batch" mode using **workflows**
- Your workflows are at the bottom of the Tools list
- System-wide published workflows are in the **Shared Data menu**.

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



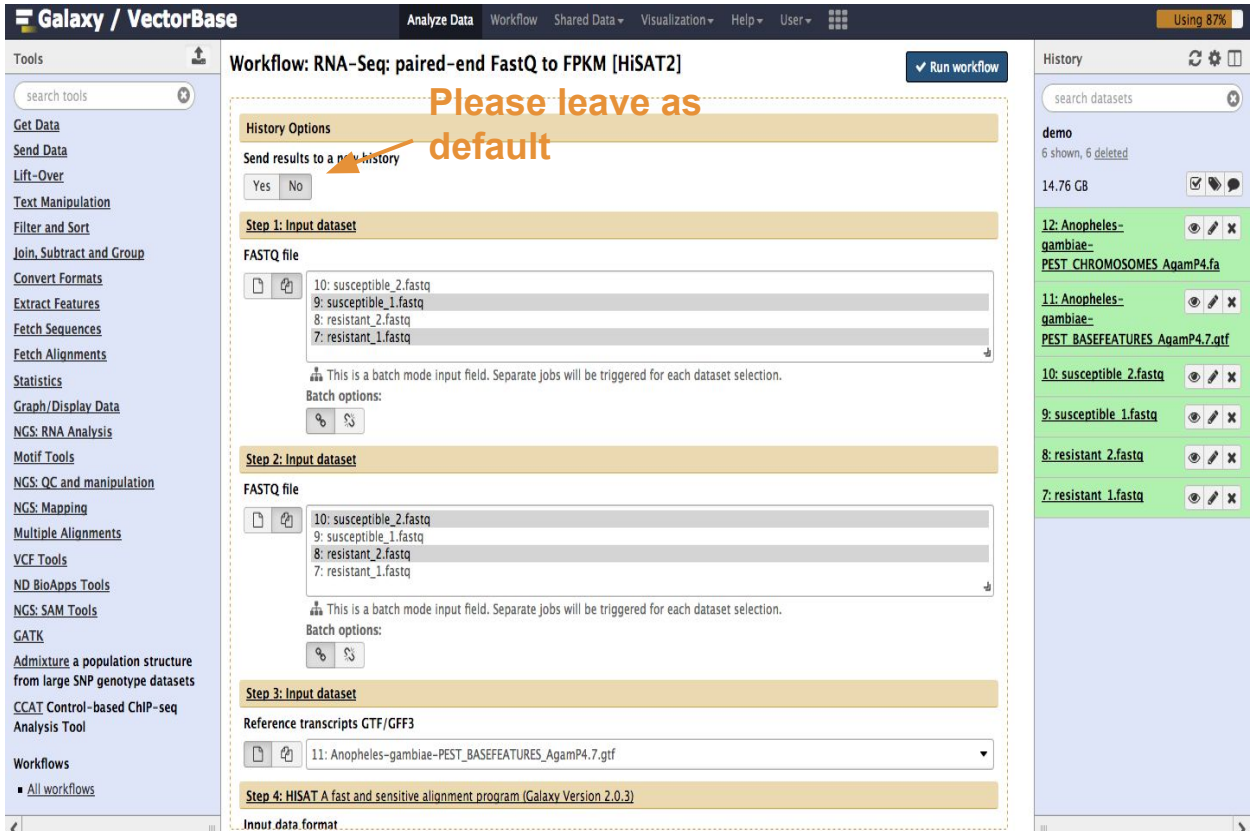
Starting a workflow

- 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...



Executing a workflow

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



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...

Galaxy / VectorBase

Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 86%

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

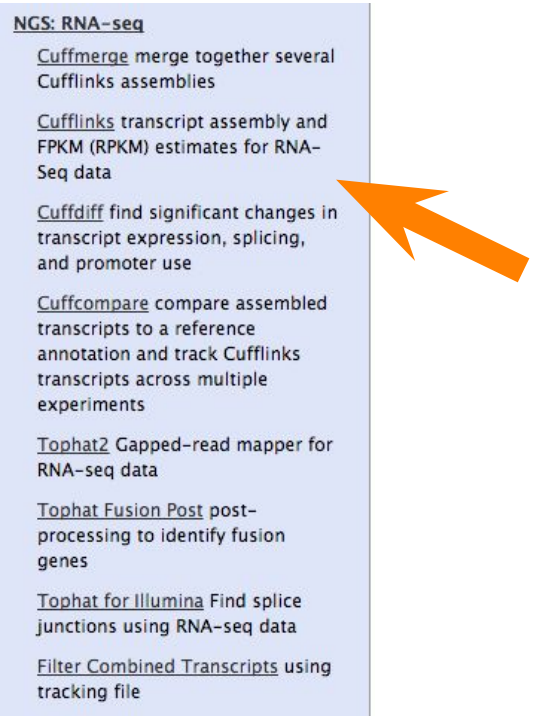
☐	Recife - run	23	0 Tags	42.5 GB	~7 hours ago	~41 minutes ago	**current history**

- 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)



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 -I 300000 -F 0.100000 -j 0.150000 -p 4 -G /rnarocket/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 -I 300000 -F 0.100000 -j 0.150000 -p 4 -G /rnarocket/indices/genomes/pathogen_portal_inplace_updates/pathogen_portal_indices/vectorbase/Anopheles_gambiae_AgamP3/Anopheles-gambiae-PEST_BASEFEATURES_AgamP3.7.

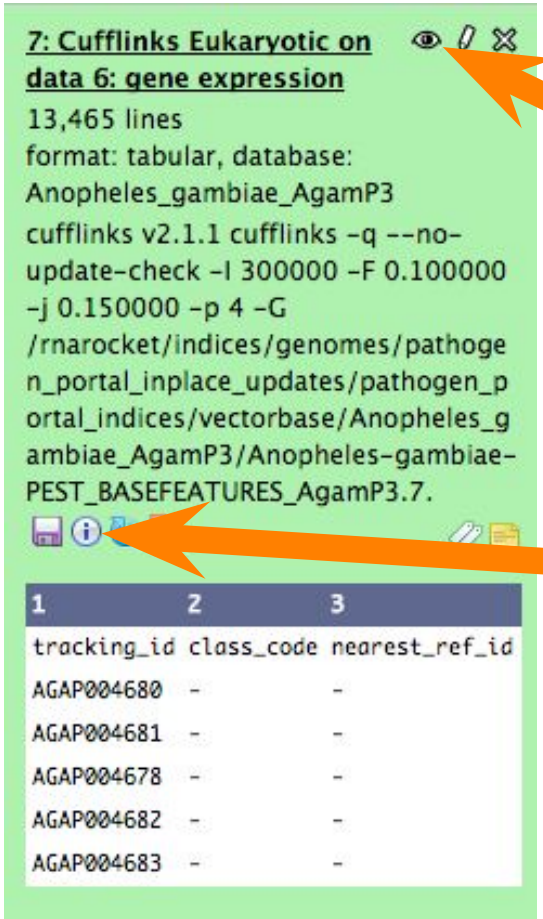
1	2	3
tracking_id	class_code	nearest_ref_i
AGAP004680-RA	-	-
AGAP004681-RA	-	-
AGAP004678-RA	-	-
AGAP004682-RA	-	-
AGAP004683-RA	-	-

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 -I 300000 -F 0.100000 -j 0.150000 -p 4 -G /rnarocket/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.St
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



View the results as text/table

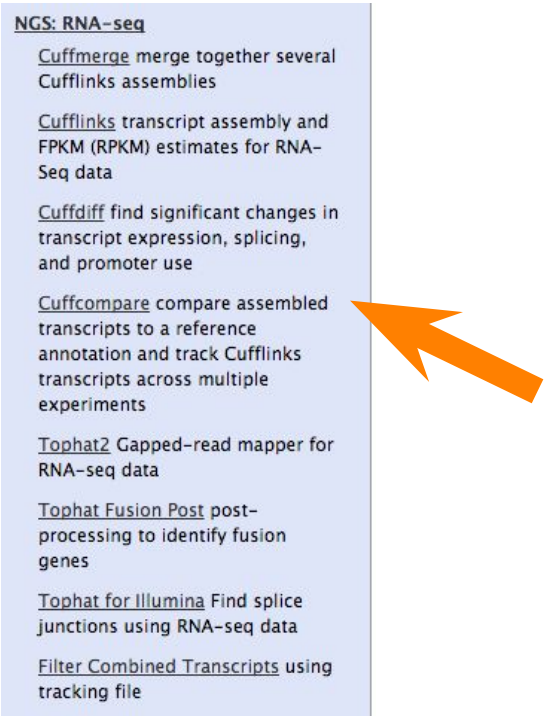
View details about how this file was generated, e.g. tool, input files and parameters

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



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

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

So what do these genes do?

- 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