

GeneLab

Open Science for Life in Space

RNA-SEQ TUTORIAL USING GENELAB'S ANALYSIS PLATFORM

Version 3.0

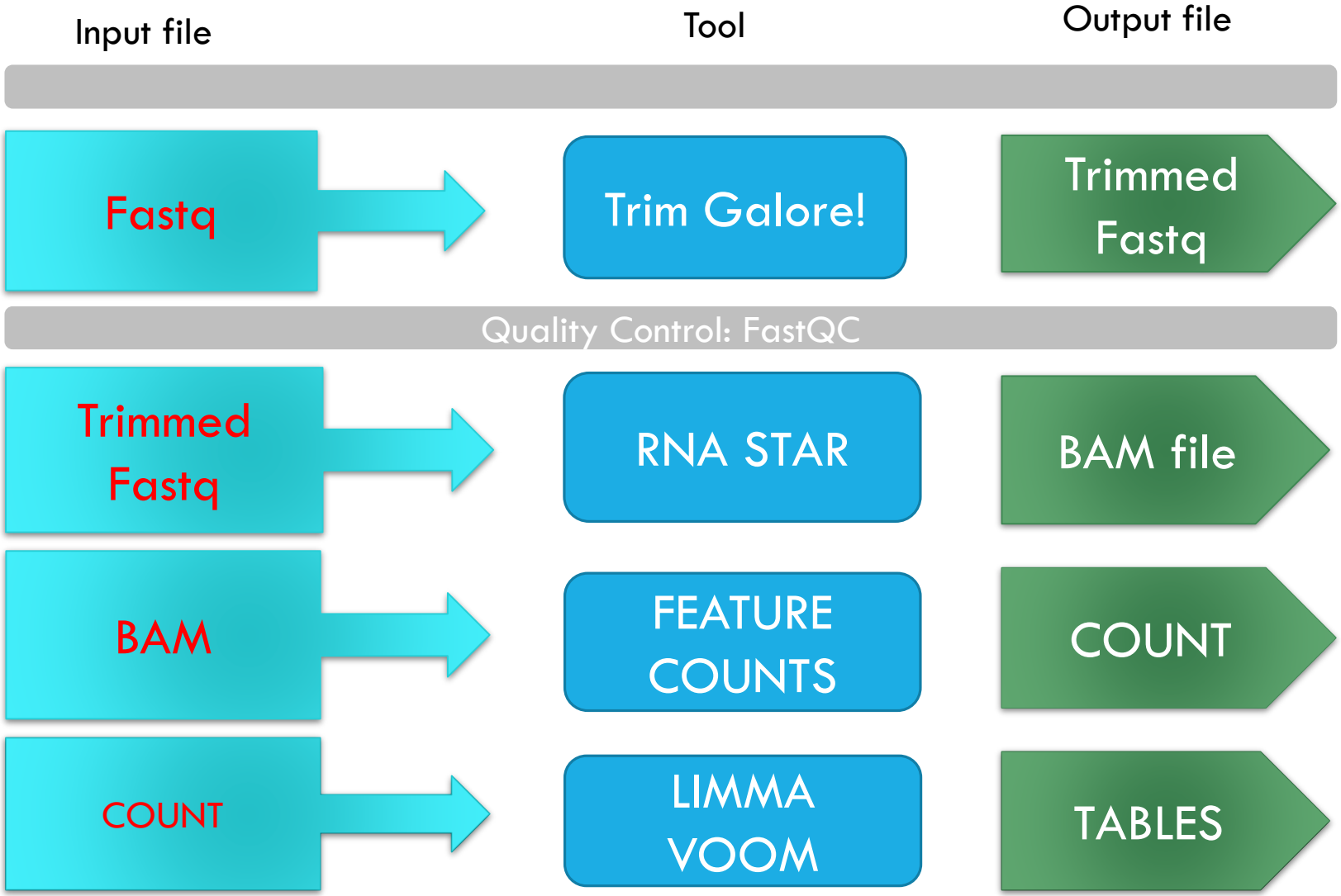
October 1, 2018

PIPELINE

- Quality Control
  - Tools: FastQC
- Trimming
  - Tools : Trim Galore, Trimmomatic, Cutadapt
- Alignment (Mapping)
  - Tools : RNA STAR, Bowtie, HiSAT2
- Count:
  - Tools: Feature Counts, HTSeq, Rsem
- Differential gene expression
  - Tools: Limma-voom, edgeR, DESeq/ DESeq2 (these do Counts too)

The pipeline described in this tutorial was used to generate the GeneLab processed data for RNA-Seq (in red); however, there are other software packages and parameters that can be used in each step of the analysis pipeline that may be better suited for the user's needs.

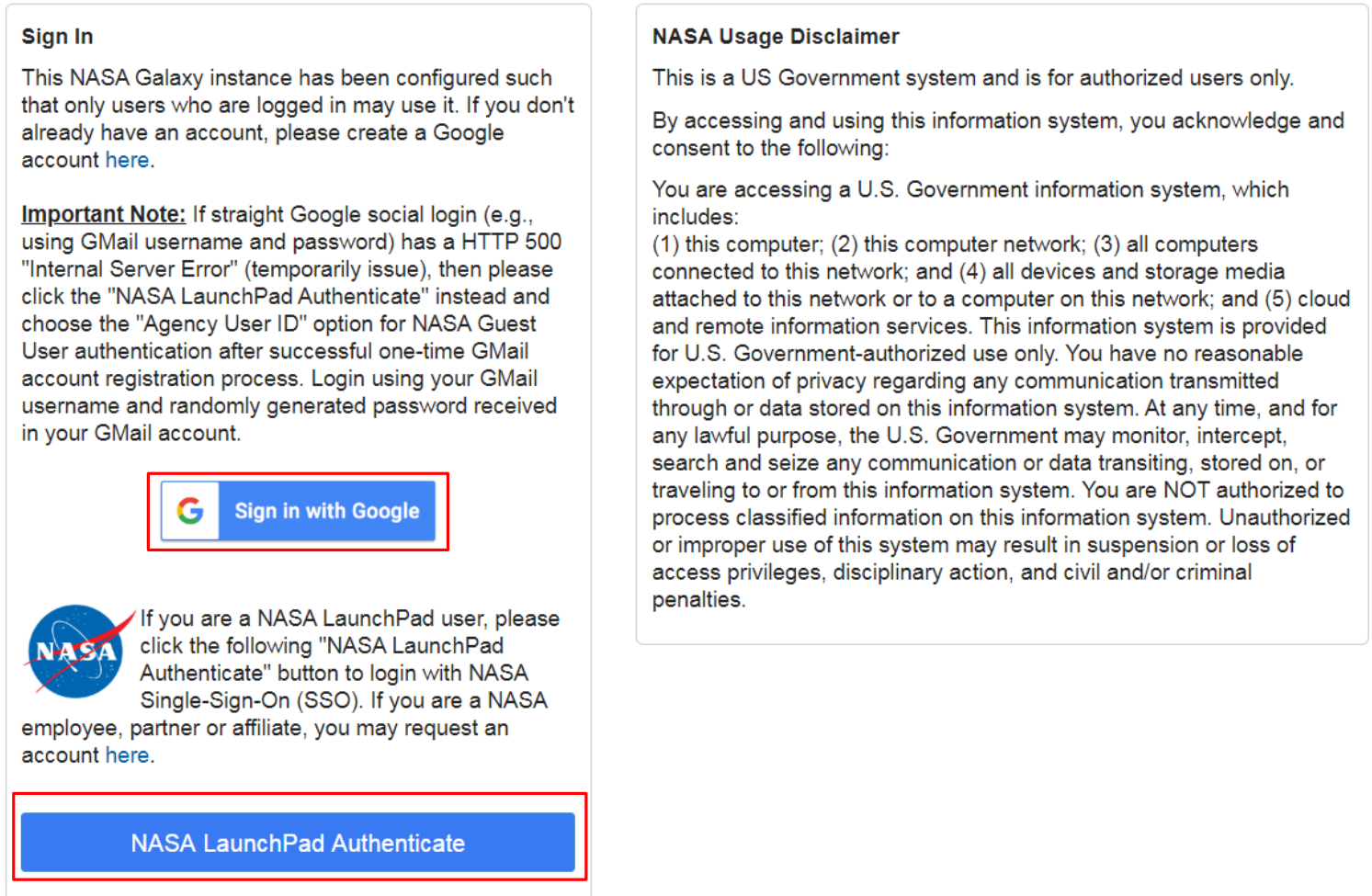
GENELAB RNA-SEQ WORKFLOW



LOGGING INTO THE ANALYSIS PLATFORM

https://genelab-data.ndc.nasa.gov/glxy-sso-login/

Log in using your Google ID or NASA Credentials



Sign In

This NASA Galaxy instance has been configured such that only users who are logged in may use it. If you don't already have an account, please create a Google account here.

Important Note: If straight Google social login (e.g., using GMail username and password) has a HTTP 500 "Internal Server Error" (temporarily issue), then please click the "NASA LaunchPad Authenticate" instead and choose the "Agency User ID" option for NASA Guest User authentication after successful one-time GMail account registration process. Login using your GMail username and randomly generated password received in your GMail account.

Sign in with Google

If you are a NASA LaunchPad user, please click the following "NASA LaunchPad Authenticate" button to login with NASA Single-Sign-On (SSO). If you are a NASA employee, partner or affiliate, you may request an account here.

NASA LaunchPad Authenticate

NASA Usage Disclaimer

This is a US Government system and is for authorized users only.

By accessing and using this information system, you acknowledge and consent to the following:

You are accessing a U.S. Government information system, which includes:
(1) this computer; (2) this computer network; (3) all computers connected to this network; and (4) all devices and storage media attached to this network or to a computer on this network; and (5) cloud and remote information services. This information system is provided for U.S. Government-authorized use only. You have no reasonable expectation of privacy regarding any communication transmitted through or data stored on this information system. At any time, and for any lawful purpose, the U.S. Government may monitor, intercept, search and seize any communication or data transiting, stored on, or traveling to or from this information system. You are NOT authorized to process classified information on this information system. Unauthorized or improper use of this system may result in suspension or loss of access privileges, disciplinary action, and civil and/or criminal penalties.

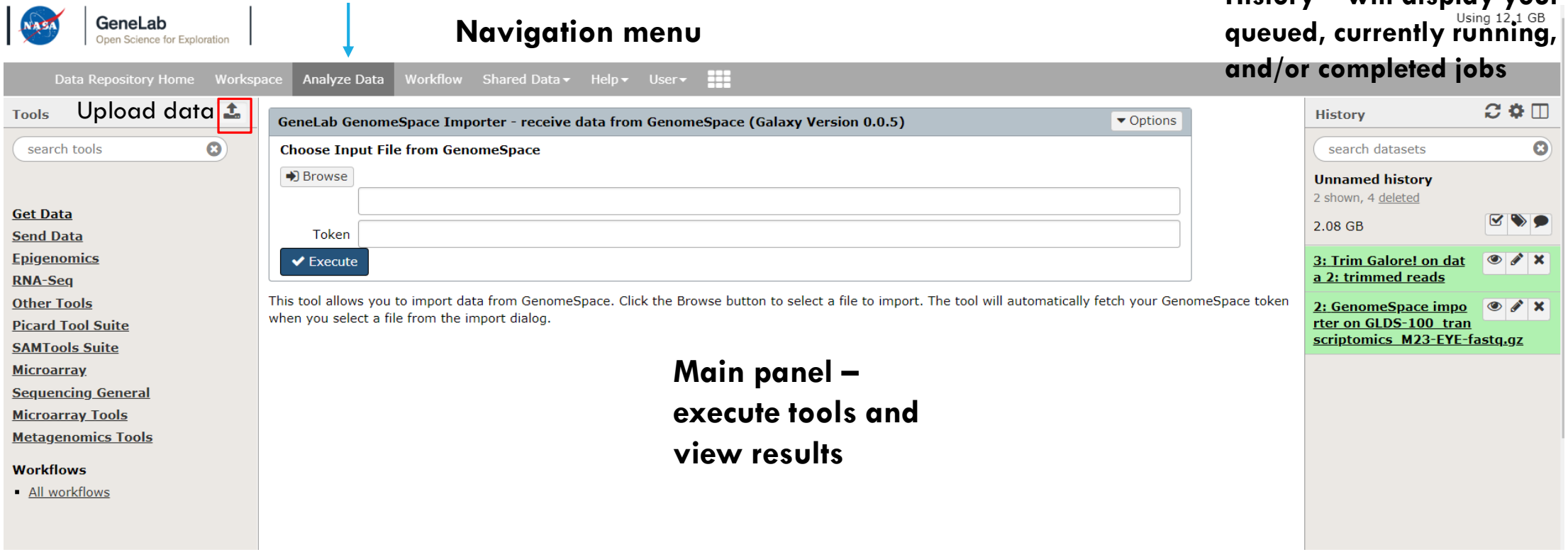
ANALYSIS PLATFORM OVERVIEW

Click Analyze Data to navigate to this user interface

Navigation menu

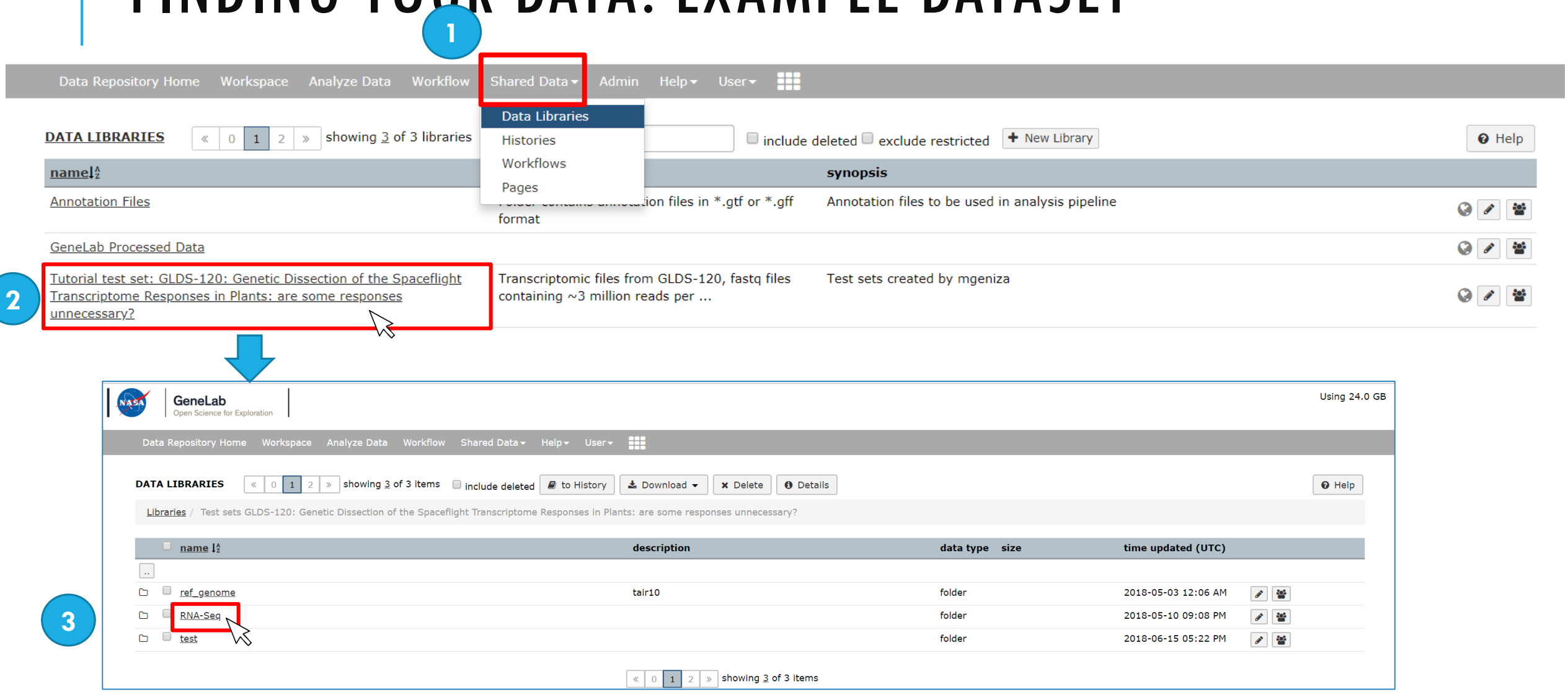
History – will display your queued, currently running, and/or completed jobs

Toolshed - all the available tools



Main panel – execute tools and view results

FINDING YOUR DATA: EXAMPLE DATASET



GeneLab
Open Science for Exploration

Using 24.0 GB

Data Repository Home | Workspace | Analyze Data | Workflow | Shared Data | Help | User

DATA LIBRARIES « 0 1 2 » showing 12 of 12 items include deleted | to History | Download | Delete | Details | Help

Libraries / Test sets GLDS-120: Genetic Dissection of the Spaceflight Transcriptome Responses in Plants: are some responses unnecessary? / RNA-Seq / RawData

name	description	data type	size	time updated (UTC)
..				
GLDS-120_Athaliana_Col0_GC_Root_Light_Day11_Rep1_R1_RNASeq_RawReads.fastq		fastqsanger	716.9 MB	2018-05-10 08:57 PM
GLDS-120_Athaliana_Col0_GC_Root_Light_Day11_Rep1_R2_RNASeq_RawReads.fastq		fastqsanger	716.9 MB	2018-05-10 08:57 PM
GLDS-120_Athaliana_Col0_GC_Root_Light_Day11_Rep2_R1_RNASeq_RawReads.fastq		fastqsanger	717.0 MB	2018-05-10 08:57 PM
GLDS-120_Athaliana_Col0_GC_Root_Light_Day11_Rep2_R2_RNASeq_RawReads.fastq		fastqsanger	717.0 MB	2018-05-10 08:57 PM
GLDS-120_Athaliana_Col0_GC_Root_Light_Day11_Rep3_R1_RNASeq_RawReads.fastq		fastqsanger	745.2 MB	2018-05-10 08:57 PM
GLDS-120_Athaliana_Col0_GC_Root_Light_Day11_Rep3_R2_RNASeq_RawReads.fastq		fastqsanger	745.2 MB	2018-05-10 08:57 PM
GLDS-120_Athaliana_Col0_FLT_Root_Light_Day11_Rep1_R1_RNASeq_RawReads.fastq		fastqsanger	605.8 MB	2018-05-10 08:57 PM
GLDS-120_Athaliana_Col0_FLT_Root_Light_Day11_Rep1_R2_RNASeq_RawReads.fastq		fastqsanger	605.8 MB	2018-05-10 08:57 PM
GLDS-120_Athaliana_Col0_FLT_Root_Light_Day11_Rep2_R1_RNASeq_RawReads.fastq		fastqsanger	628.9 MB	2018-05-10 08:57 PM
GLDS-120_Athaliana_Col0_FLT_Root_Light_Day11_Rep2_R2_RNASeq_RawReads.fastq		fastqsanger	628.9 MB	2018-05-10 08:57 PM
GLDS-120_Athaliana_Col0_FLT_Root_Light_Day11_Rep3_R1_RNASeq_RawReads.fastq		fastqsanger	644.9 MB	2018-05-10 08:57 PM
GLDS-120_Athaliana_Col0_FLT_Root_Light_Day11_Rep3_R2_RNASeq_RawReads.fastq		fastqsanger	644.9 MB	2018-05-10 08:57 PM

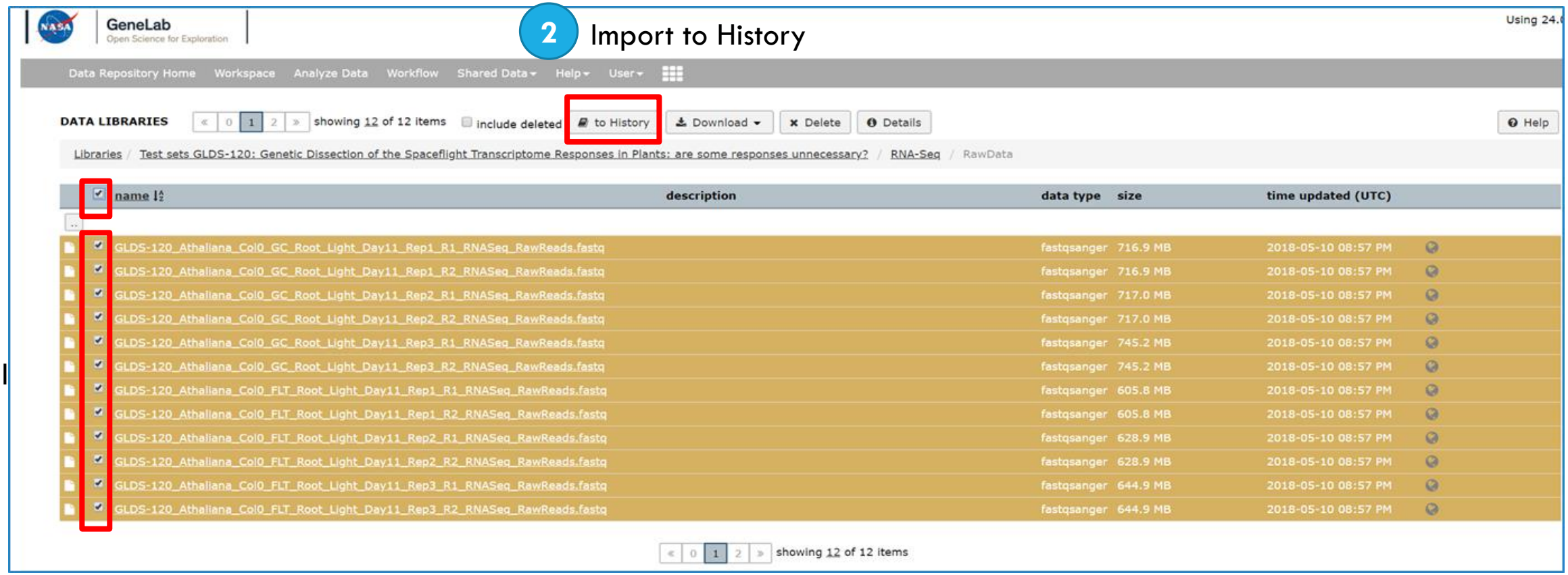
« 0 1 2 » showing 12 of 12 items

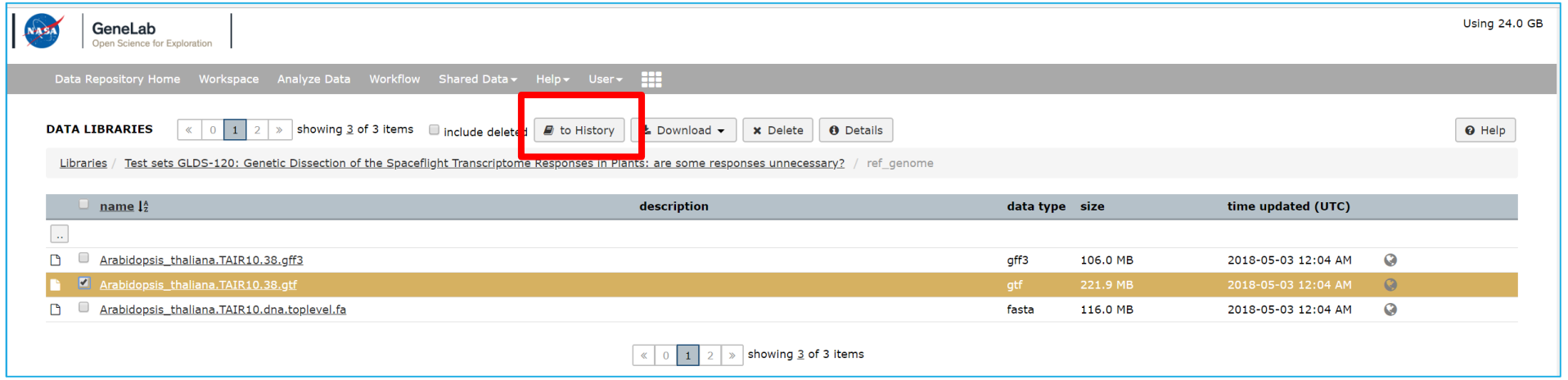
ADDING DATA TO HISTORY

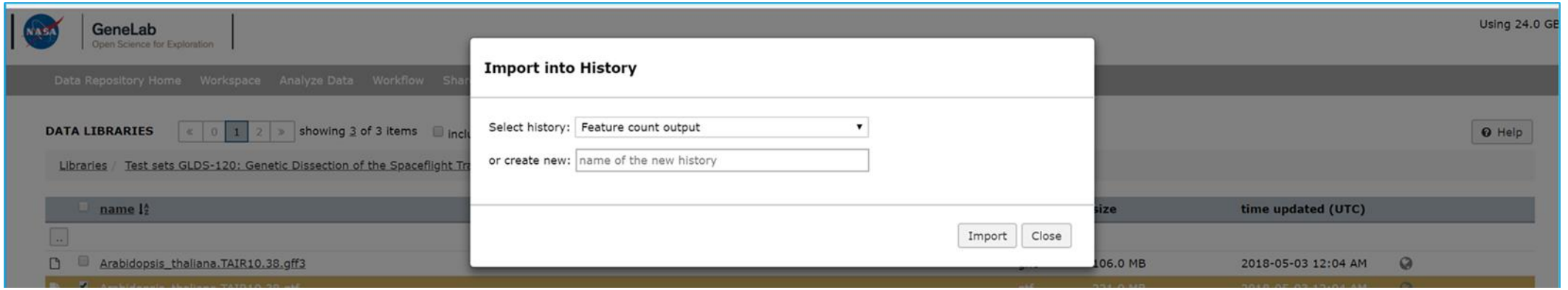
1

Select files – check all

Or individual files





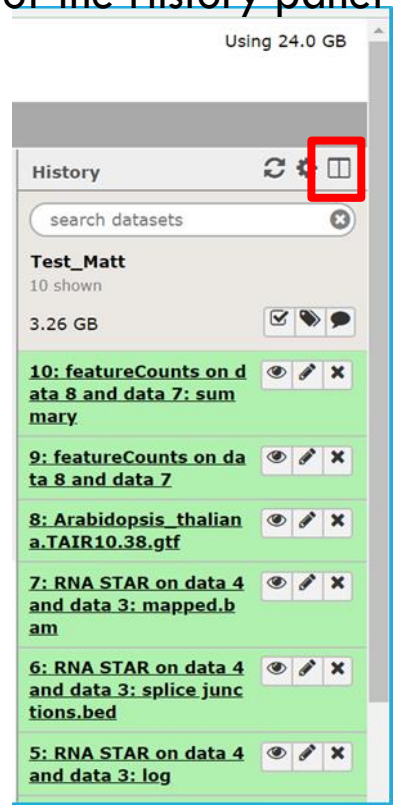


Now you have a CURRENT history containing all RAW files and annotated Reference Genome file (.gtf file)

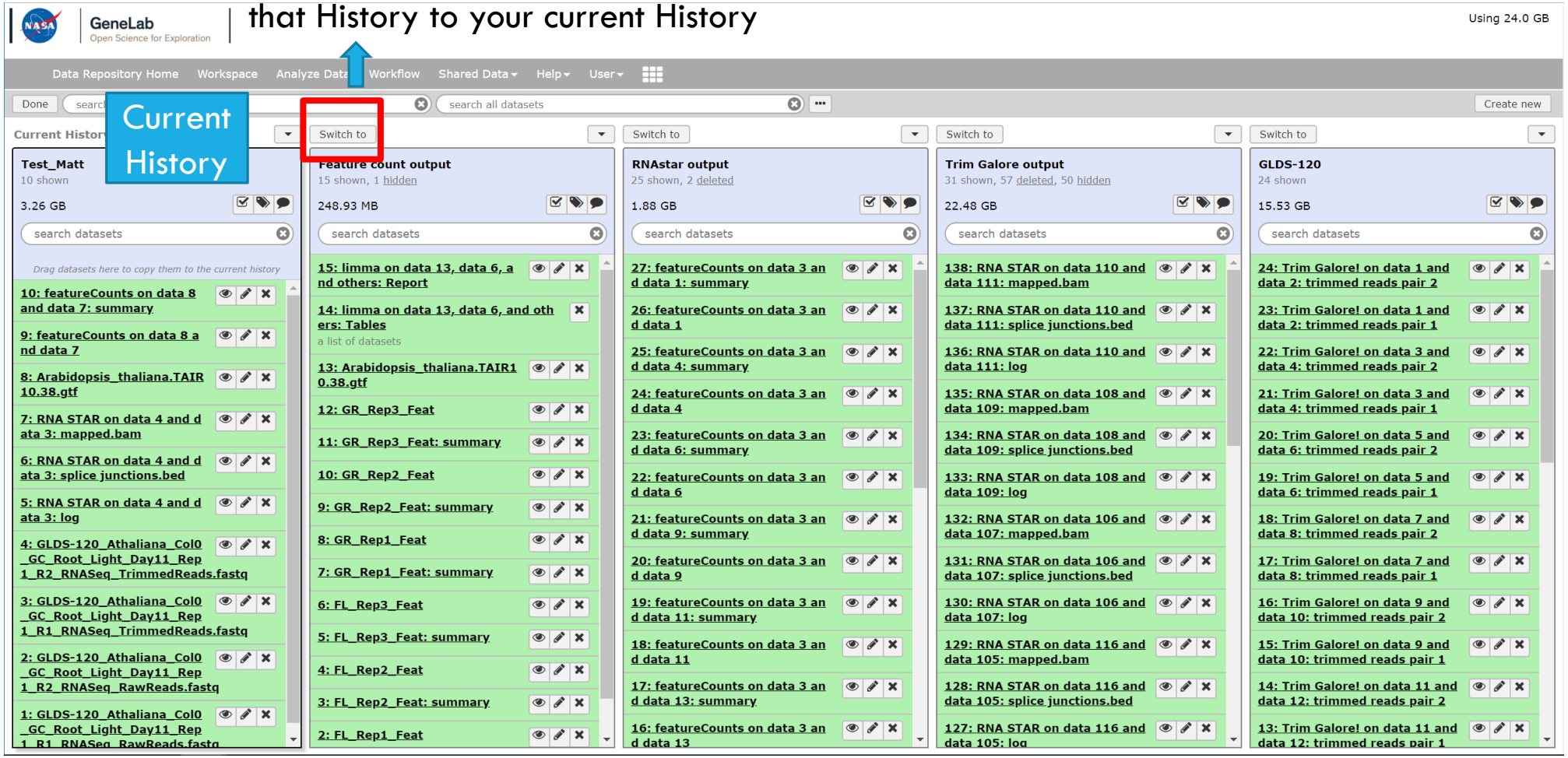
You can choose a pre-existing history from the drop-down menu to place the data from Data Libraries OR you can send them to a new history by typing a title in the ‘create new’ box and clicking the Import button

VIEWING HISTORY

To view your histories click on this icon at the top right corner of the History panel



Click the Switch to button to move that History to your current History

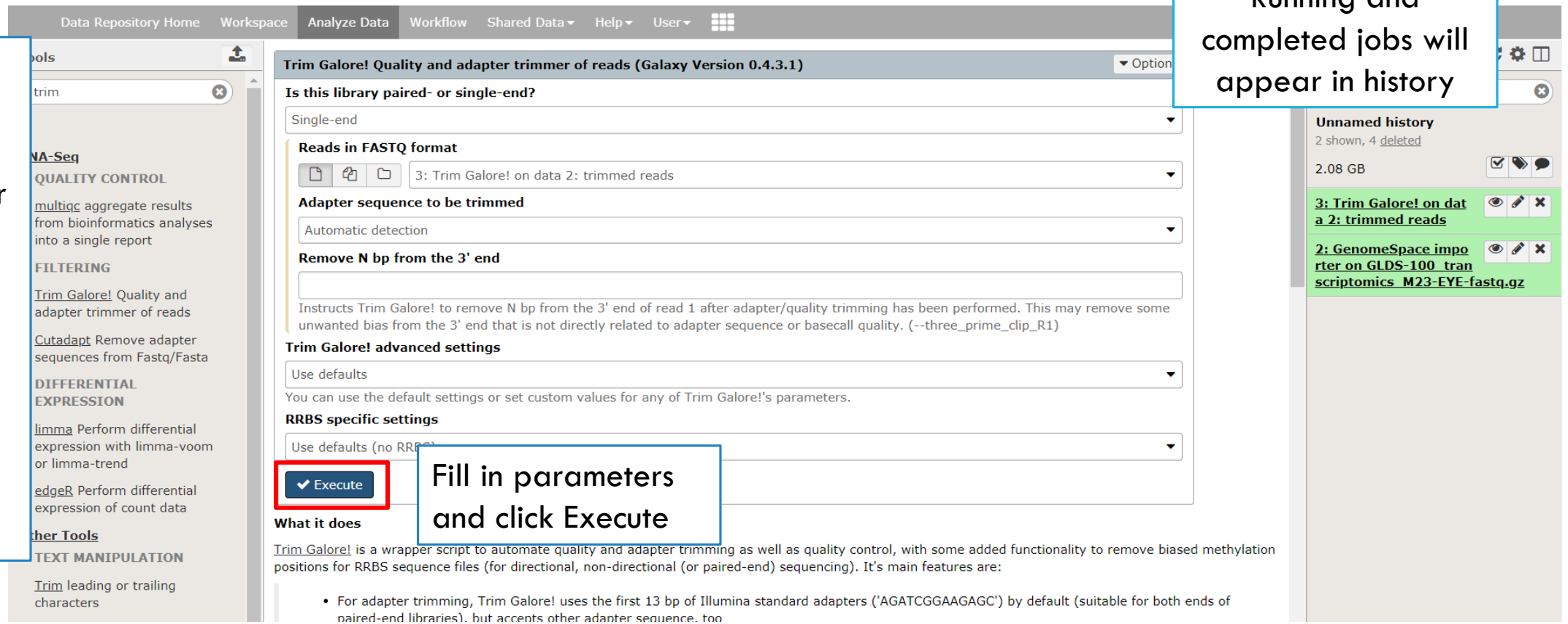


SEARCHING AND RUNNING TOOLS

Search tools by using keywords in the search bar

Or

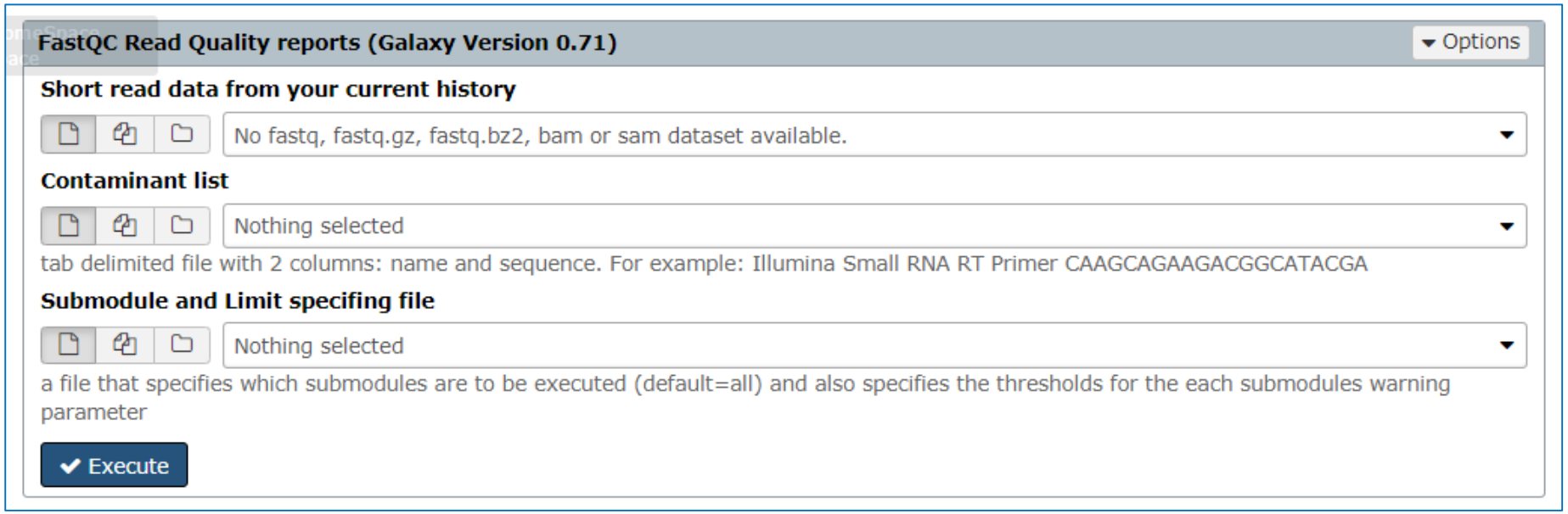
Use the menu and scroll through the various categories



Running and completed jobs will appear in history

Fill in parameters and click Execute

QUALITY CONTROL: CALL FASTQC



Short read data from your current history

➢Select your fastq files or dataset collection

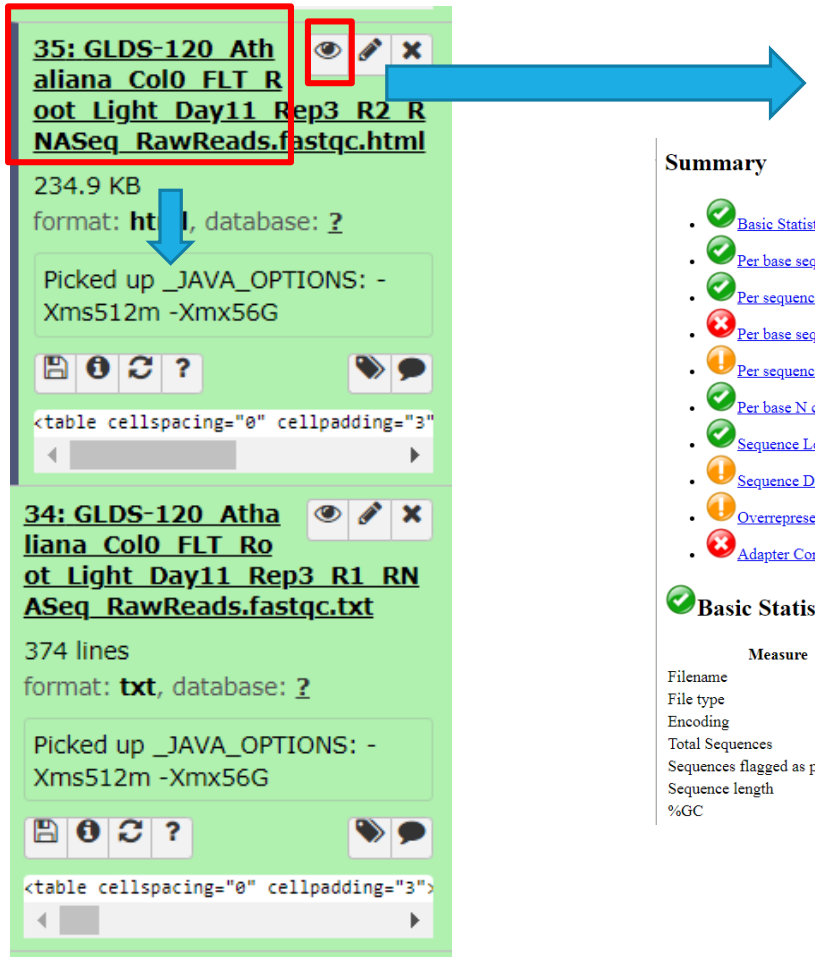
Contaminant list

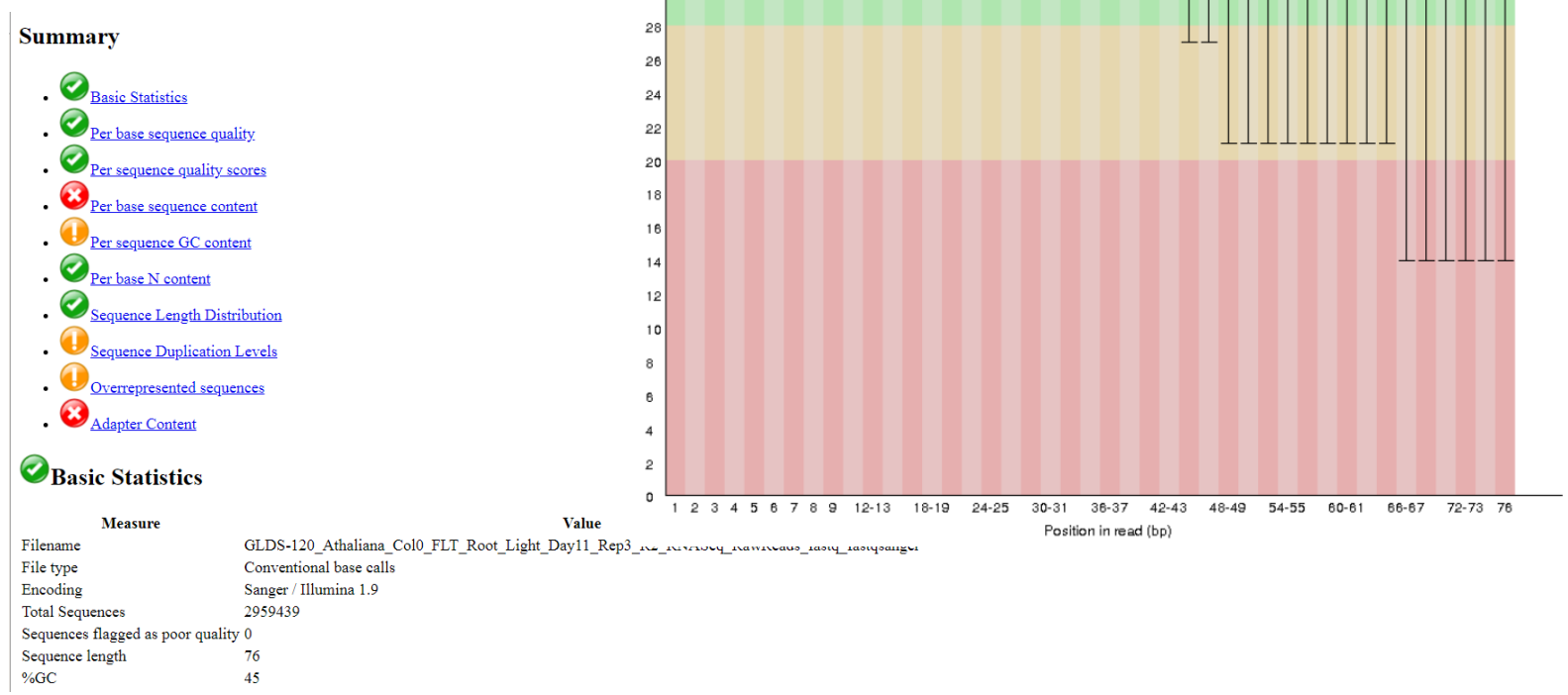
➢Use default

Submodule and Limit specifying file

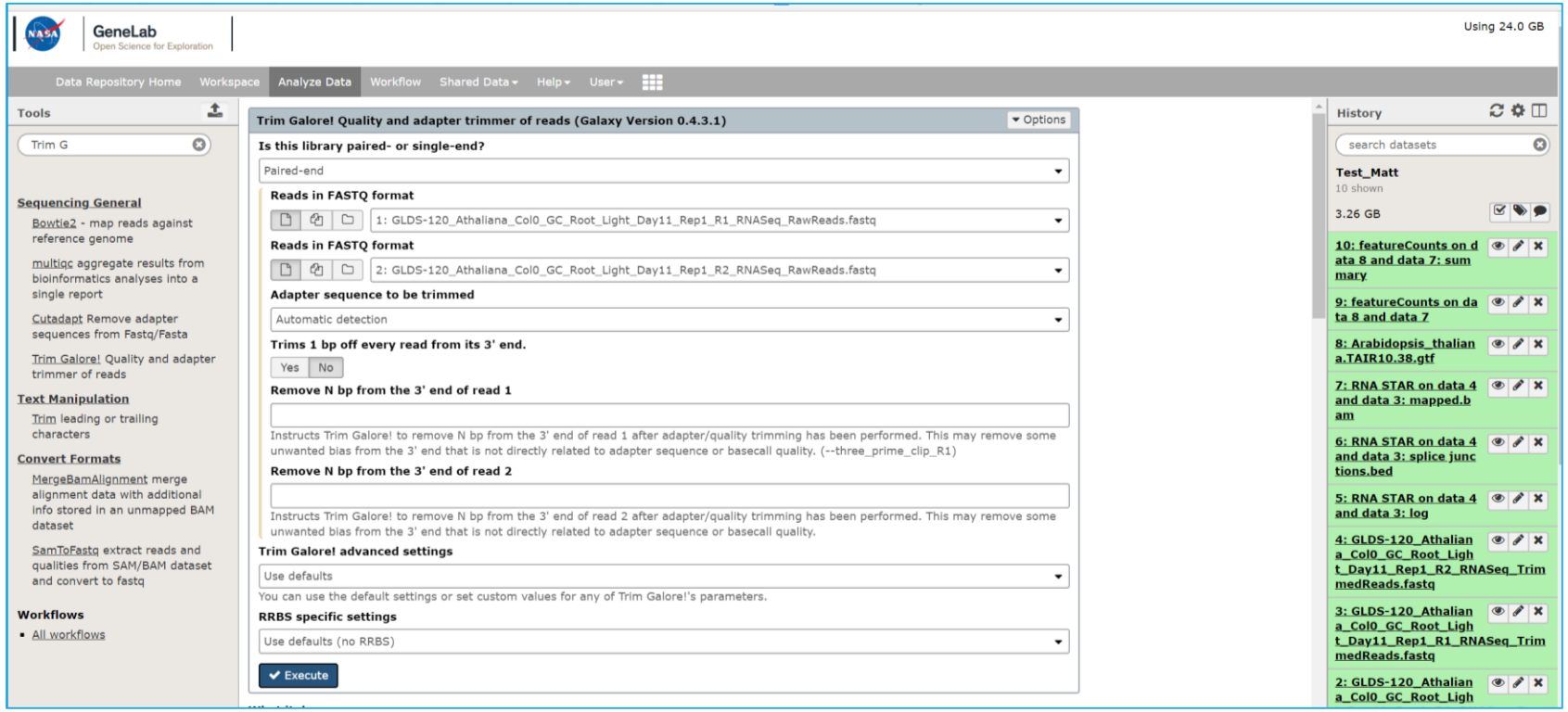
➢ Use default

FASTQC RESULTS





TRIMMING : CALL TRIM GALORE!



Is this library paired- or single-end?

- Paired-end

Reads in FASTQ format

- Add R1 and R2 fastq files of the same sample

Adapter sequence to be trimmed

- Automatic detection

Trims 1 bp off every read from its 3' end.

- False

Remove N bp from the 3' end of read 1

- Not available.

Remove N bp from the 3' end of read 2

- Not available.

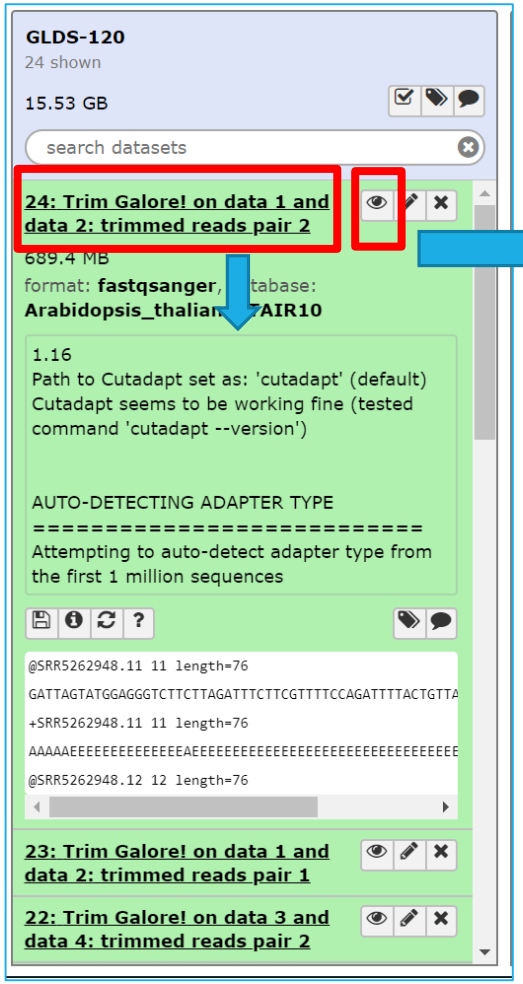
Trim Galore! advanced settings

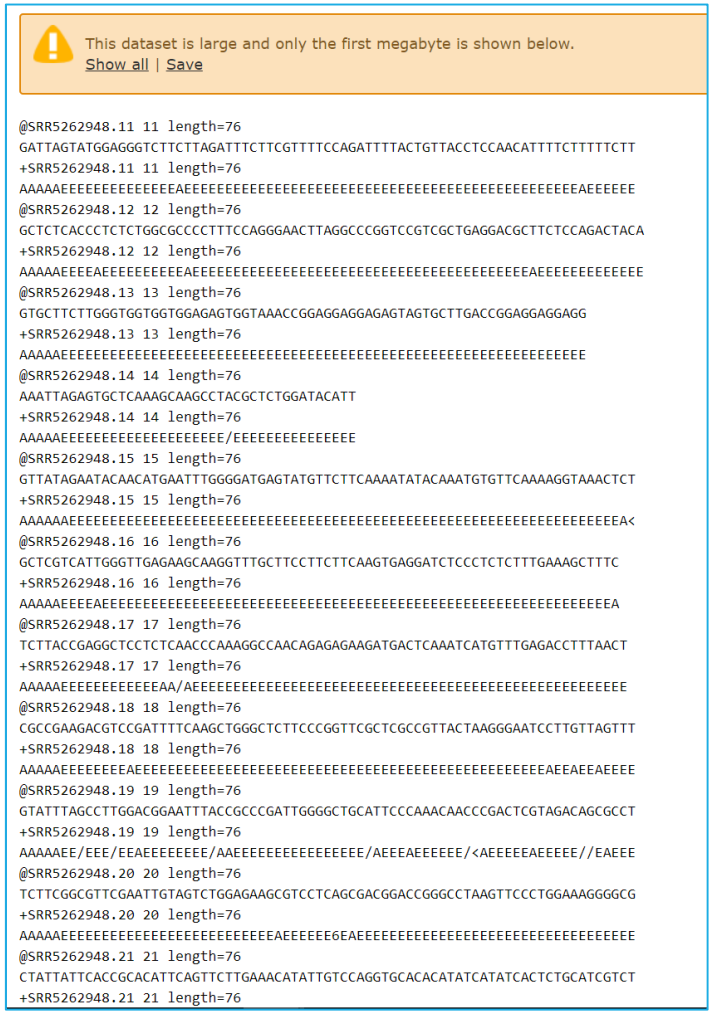
- Use defaults

RRBS specific settings

- Use defaults (no RRBS)

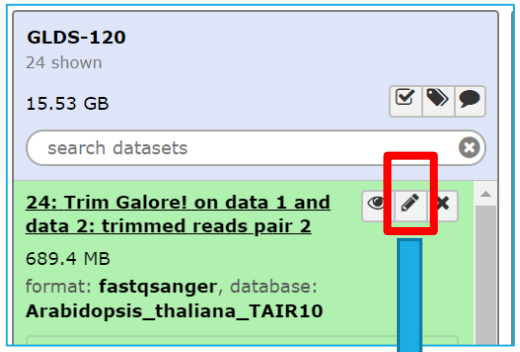
TRIM GALORE! RESULTS





Perform FastQC to check the quality of your trimmed fastq files.

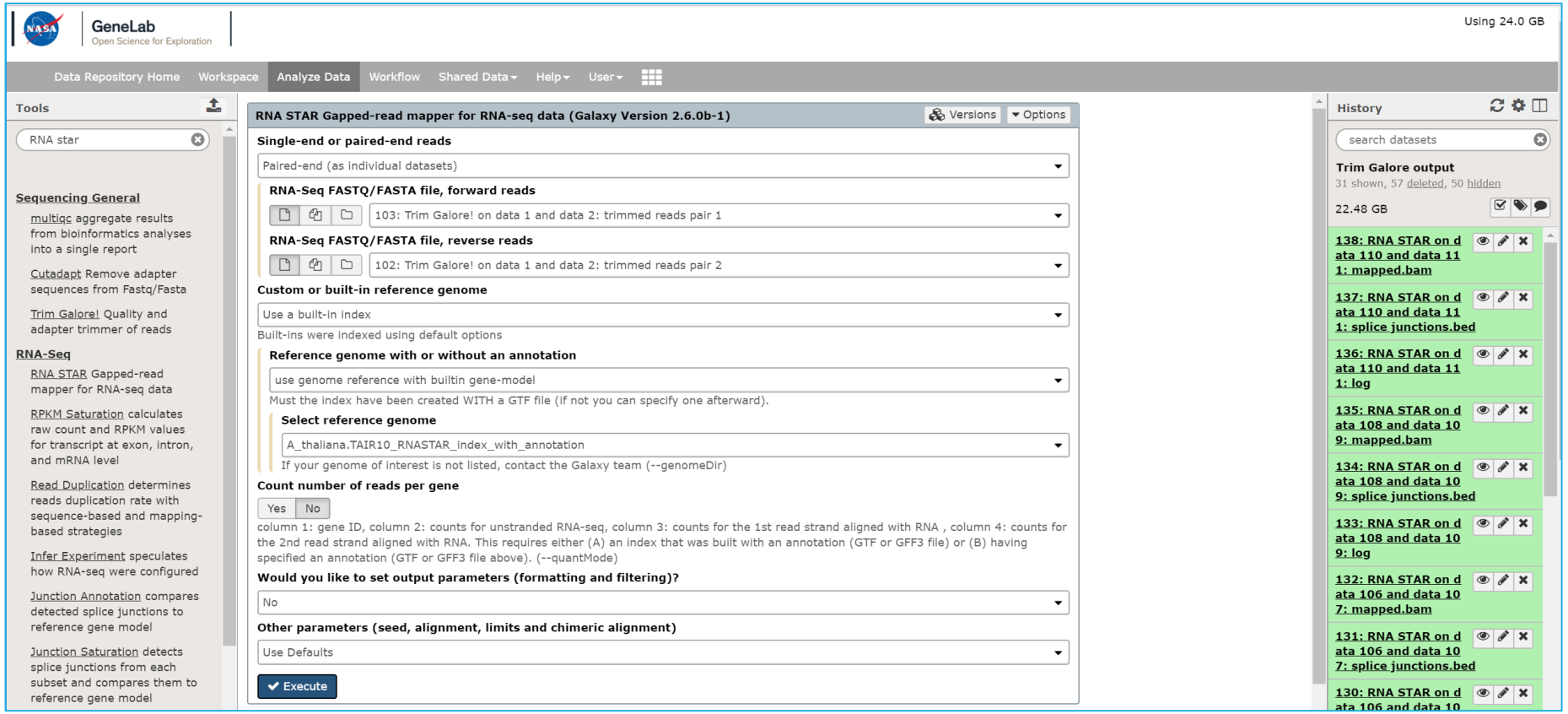
CHANGING FILE NAMES/METADATA



Change file names to reflect Sample info –
Ground_Rep1_Trim
- ❑ **Hit ENTER**

Helps to track Samples

ALIGNMENT : CALL RNA STAR



Single-end or paired-end reads
- ➢Paired-end (as individual datasets)

RNA-Seq FASTQ/FASTA file, forward reads
- ➢Output dataset 'trimmed_reads_pair1' from step 1

RNA-Seq FASTQ/FASTA file, reverse reads
- ➢Output dataset 'trimmed_reads_pair2' from step 1

Custom or built-in reference genome
- ➢Use a built-in index

Reference genome with or without an annotation
- ➢use genome reference with built in gene-model

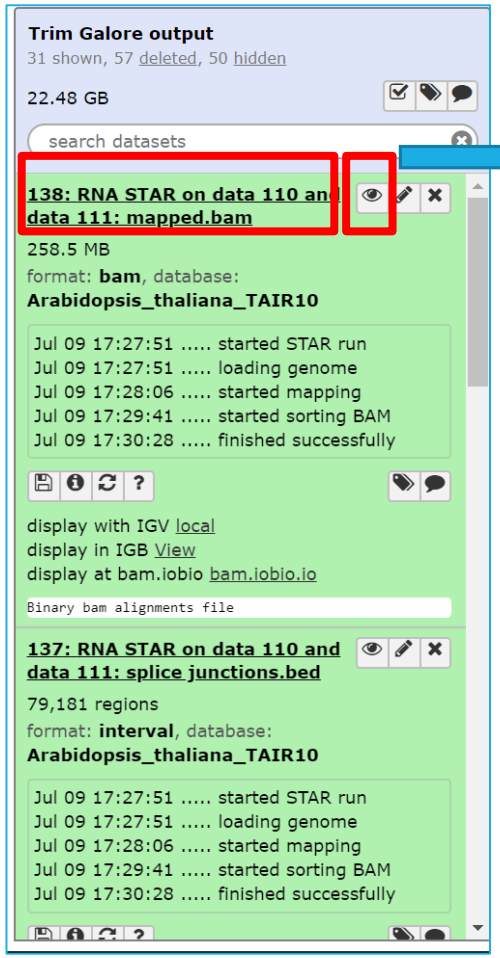
Select reference genome
- ➢Arabidopsis_thaliana.TAIR10

Count number of reads per gene
- ➢False

Would you like to set output parameters (formatting and filtering)?
- ➢No

Other parameters (seed, alignment, limits and chimeric alignment)
- ➢Use Defaults

RNA STAR RESULTS



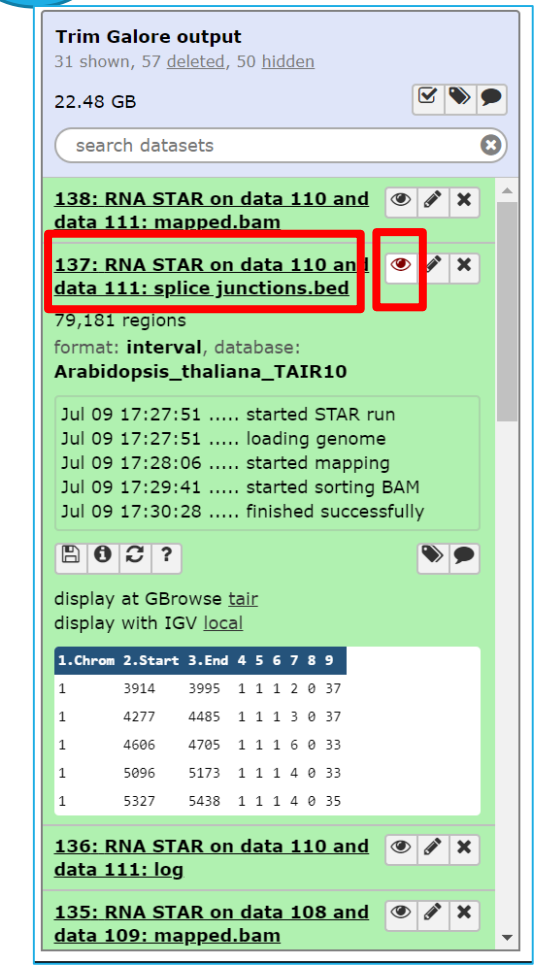
.bam file → Binary equivalent of SAM (Sequence Alignment/Map) file

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ	QUAL
@HD VN:1.4 SO:coordinate										
@SQ SN:1 LN:30427671										
@SQ SN:2 LN:19698289										
@SQ SN:3 LN:23459830										
@SQ SN:4 LN:18585056										
@SQ SN:5 LN:26975502										
@SQ SN:Mt LN:366924										
@SQ SN:Pt LN:154478										
@PG ID:STAR PN:STAR VN:STAR_2.6.0b CL:STAR --runThreadN 8 --genomeDir /opt/galaxy-user/galaxy-17.09-github/tool-data/Arabidopsis_thaliana.TAIR10/rnastar_index2/Arabidopsis_thaliana.TAIR10/dataset_17967_files --genomeLoad NoSharedMemory --readFilesIn /c										
@CO user command line: STAR --runThreadN 8 --genomeLoad NoSharedMemory --genomeDir /opt/galaxy-user/galaxy-17.09-github/tool-data/Arabidopsis_thaliana.TAIR10/rnastar_index2/Arabidopsis_thaliana.TAIR10/dataset_17967_files --readFilesIn /opt2/galaxy/files/0										
SRR5263094.2725762	99	1	3236	255	76M	=	3303	143	CTATTATAATACTAAAAGGGGAATACAAATTTCTACAGAGGATGATATTCAATCCACGGTTCACCCAAACCGATTT	AAAAAEEEEEEEEEAEEEE
SRR5263094.2725762	147	1	3303	255	76M	=	3236	-143	AACCGATTTTATAAAATTTATTATTAAATCTTTTTTAATTGTTAAATTGGTTTAAATCTGAACTCTGTTTACTTAC	EEEEEEEEEEEEEEEEEE
SRR5263094.1922645	99	1	3634	255	75M	=	3686	124	TTATTAGATATACCAAACCAGAGAAAACAAATACATAATCGGAGAAATACAGATTACAGAGAGCGAGAGAGATCG	AAAAAEEEEEEEEEEE6E
SRR5263094.1432922	163	1	3660	255	76M	=	3800	215	ACAAATACATAATCGGAGAAATACAGATTACAGAGAGCGAGAGAGATCGACGGCGAAGCTCTTTACCCGAAAACCA	AAAAAEEEEEEEEEEEEE
SRR5263094.2649309	163	1	3685	255	44M	=	3685	44	GATTACAGAGAGCGAGAGAGATCGACGGCTAAGCTCTTTACCCG	/A/AAEEAEEEAEEE/EE
SRR5263094.2649309	83	1	3685	255	44M	=	3685	-44	GATTACAGAGAGCGAGAGAGATCGACGGCTAAGCTCTTTACCCG	EEEEE/EEEEEE/EEEEE
SRR5263094.1922645	147	1	3686	255	72M	=	3634	-124	ATTACAGAGAGCGAGAGAGATCGACGGCGAAGCTCTTTACCCGGAAACCATTGAAATCGGACGGTTTAGTGA	EEEEEEEEEEEEEEEEEE
SRR5263094.2661221	99	1	3689	255	76M	=	3735	122	ACAGAGAGCGAGAGAGATCGACGGCGAAGCTCTTTACCCGGAAACCATTGAAATCGGACGGTTTAGTGAAAATGGA	AAAAAEEEEEEEEEEEEE
SRR5263094.2661221	147	1	3735	255	76M	=	3689	-122	ATTGAAATCGGACGGTTTAGTGAAAATGGAGGATCAAGTTGGGTTTGGGTTCCGTCCGAACGACGAGGAGCTCGTT	EEEEEEEEEEEEEEEEEE
SRR5263094.1432922	83	1	3800	255	75M	=	3660	-215	AGGAGCTCGTTGGTCACTATCTCCGTAACAAAATCGAAGGAAACACTAGCCGCGACGTTGAAGTAGCCATCAGCG	E6AEEEEEE<EAEEEEE
SRR5263094.454076	163	1	3877	255	37M82N37M	=	4236	644	GTCAACATCTGTAGCTACGATCCTTGGAACTTGCGCTTCCAGTCAAAGTACAAATCGAGAGATGCTATGTGGTA	AAAA6AEEEAA/EEEEEE
SRR5263094.1106877	163	1	3898	255	16M82N60M	=	4051	229	CCTTGGAACTTGCGCTTCCAGTCAAAGTACAAATCGAGAGATGCTATGTGGTACTTCTTCTCTCGTAGAGAAAACA	AAAAAEEEEEEEEEEEEE
SRR5263094.1106877	83	1	4051	255	76M	=	3898	-229	AAACAACAAAGGGAATCGACAGAGCAGGACAACGGTTTCTGGTAAATGGAAGCTTACCGGAGAATCTGTTGAGGTC	EEEEEEEEEEEEEEEEEE
SRR5263094.2253813	163	1	4115	255	56M	=	4115	56	TCTGTTGAGGTCAAGGACCAGTGGGGATTTTGTAGTGAGGGCTTTCGTGGTAAGAT	AAAAAEEEEEEEEEEEEE
SRR5263094.2253813	83	1	4115	255	56M	=	4115	-56	TCTGTTGAGGTCAAGGACCAGTGGGGATTTTGTAGTGAGGGCTTTCGTGGTAAGAT	/EEEEEEEEEAEEEEEEE
SRR5263094.280448	163	1	4146	255	75M	=	4587	617	GTAGTGAGGGCTTTCGTGGTAAGATTGGTCATAAAAGGGTTTTGGTGTTCCTCGATGGAAGATACCCTGACAAAA	AAAAAEEEEEEEEEEEEE
SRR5263094.454076	83	1	4236	255	41M209N35M	=	3877	-644	TTATCCACGAGTTCCACTACGACCTCTTACCAGAACATCAGAGGACATATGTCATCTGCAGACTTGAGTACAAGGG	EEA//<//EAE<E///E/<
SRR5263094.65950	163	1	4238	255	39M209N37M	=	4815	653	ATCCACGAGTTCCACTACGACCTCTTACCAGAACATCAGAGGACATATGTCATCTGCAGACTTGAGTACAAGGGTG	AAAAAEEEEEEEEEEEEE
SRR5263094.1551455	163	1	4268	255	9M209N67M	=	4728	536	GAACATCAGAGGACATATGTCATCTGCAGACTTGAGTACAAGGGTGATGATGCGGACATTCTATCTGCTTATGCAA	AA6AAEEEEEEEEEAEEEE
SRR5263094.77987	99	1	4512	255	44M	=	4512	44	GTACAAGGGTGATGATGCGGACATTCTATCTGCTTATGCAATAG	AAAAAEEEEEEEEEEEEE
SRR5263094.77987	147	1	4512	255	44M	=	4512	-44	GTACAAGGGTGATGATGCGGACATTCTATCTGCTTATGCAATAG	EEEEEEEEEEEEEEEEEE
SRR5263094.583889	99	1	4547	255	59M100N17M	=	4769	298	ATGCAATAGATCCCACTCCCGCTTTTGTCCCCAATATGACTAGTAGTGCAGGTTCTGTGGTCAACCAATCACGTCA	AAAAAEEEEEEEEEEEEE
SRR5263094.949623	99	1	4550	255	56M100N18M	=	4577	203	CAATAGATCCCACTCCCGCTTTTGCCCCCAATATGACTAGTAGTGCAGGTTCTGTGGTCAACCAATCACGTCAA	AAAAAEEEEEEEEEEEEE
SRR5263094.42789	163	1	4573	255	20S33M100N20M2S	=	4573	174	GTATCAACGCAGAGTACGGGGTCCCCAATATGACTATTAGTGCAGGTTCTGTGGTCAACCAATCACGTCAACGTA	AAAAAEEEEEEEEEEEEE
SRR5263094.42789	83	1	4573	255	1S33M100N41M	=	4573	-174	GGTCCCCAATATGACTATTAGTGCAGGTTCTGTGGTCAACCAATCACGTCAACGTAATTCAGGATCTTACAACAC	EEEEEEEEEEEEEEEEEE
SRR5263094.755552	99	1	4575	255	31M100N44M	=	4590	190	CCCCAATATGACTAGTAGTGCAGGTTCTGTGGTCAACCAATCACGTCAACGAAATTCAGGATCTTACAACACTTA	AAAAAEEEEEEEEEAEE
SRR5263094.949623	147	1	4577	255	29M100N47M	=	4550	-203	CCAATATGACTAGTAGTGCAGGTTCTGTGGTCAACCAATCACGTCAACGAAATTCAGGATCTTACAACACTTACTC	EEAEEEEEEAEEEEEEE

https://genome.ucsc.edu/FAQ/FAQformat.html#format1

RNA STAR RESULTS

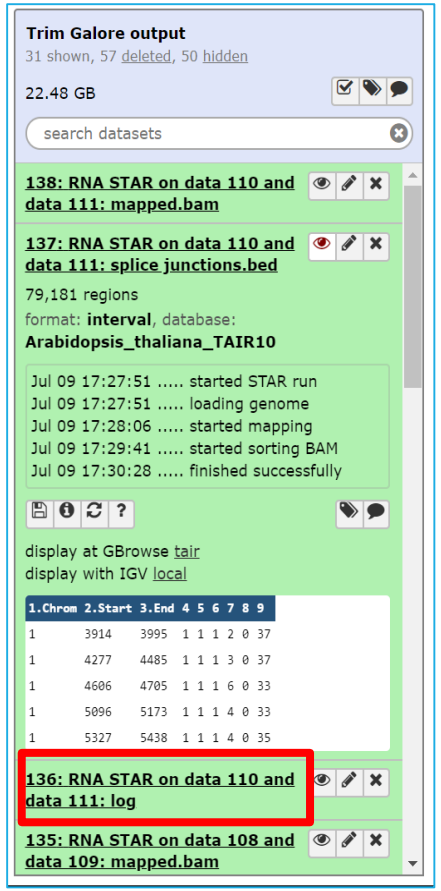
1

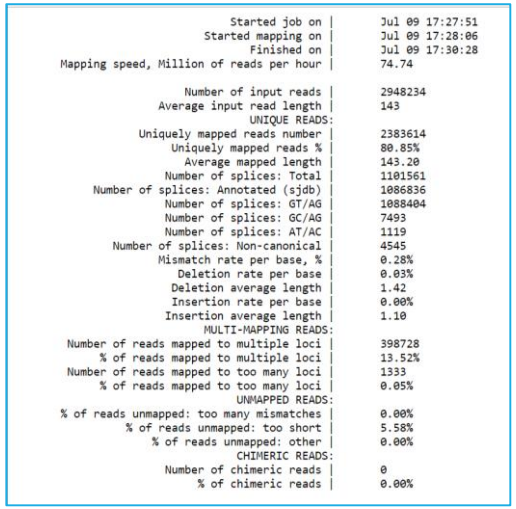


1	2	3	4	5	6	7	8	9
1	3914	3995	1	1	1	2	0	37
1	4277	4485	1	1	1	3	0	37
1	4606	4705	1	1	1	6	0	33
1	5096	5173	1	1	1	4	0	33
1	5327	5438	1	1	1	4	0	35
1	7070	7156	2	2	1	15	0	37
1	7233	7383	2	2	1	24	0	34
1	7451	7563	2	2	1	23	0	36
1	7650	7761	2	2	1	16	0	36
1	7650	8235	2	2	1	1	0	25
1	7836	7941	2	2	1	6	0	24
1	7853	7941	2	2	0	1	0	13
1	7988	8235	2	2	1	12	0	34
1	8326	8416	2	2	1	10	0	32
1	8465	8570	2	2	1	5	0	38
1	8465	8593	2	2	1	3	0	29
1	24656	24751	1	1	1	1	0	19
1	25998	26080	1	1	1	1	0	13
1	26204	26291	1	1	1	1	0	14
1	27714	27802	1	1	1	1	0	21
1	29081	29159	1	1	1	1	0	8
1	30066	30146	1	1	1	1	0	30
1	30312	30409	1	1	1	2	0	35
1	31425	31520	2	2	1	66	0	37
1	31603	31692	2	2	1	57	0	38
1	31814	31932	2	2	1	86	0	37
1	31999	32087	2	2	1	72	0	38
1	32196	32281	2	2	1	59	0	35
1	32348	32430	2	2	1	51	0	30
1	32460	32546	2	2	1	55	0	31
1	32671	33028	2	2	1	38	0	36
1	39055	39135	2	2	1	9	0	36
1	39288	39408	2	2	1	9	0	36
1	39815	40212	2	2	1	6	0	38
1	40330	40472	2	2	1	10	0	36
1	40536	40674	2	2	1	9	0	28
1	42463	50180	2	2	0	1	0	17
1	45560	45645	2	2	1	1	0	18

BED (Browser Extensible Data) format provides a flexible way to define the data lines that are displayed in an annotation track. BED lines have three required fields and nine additional optional fields.

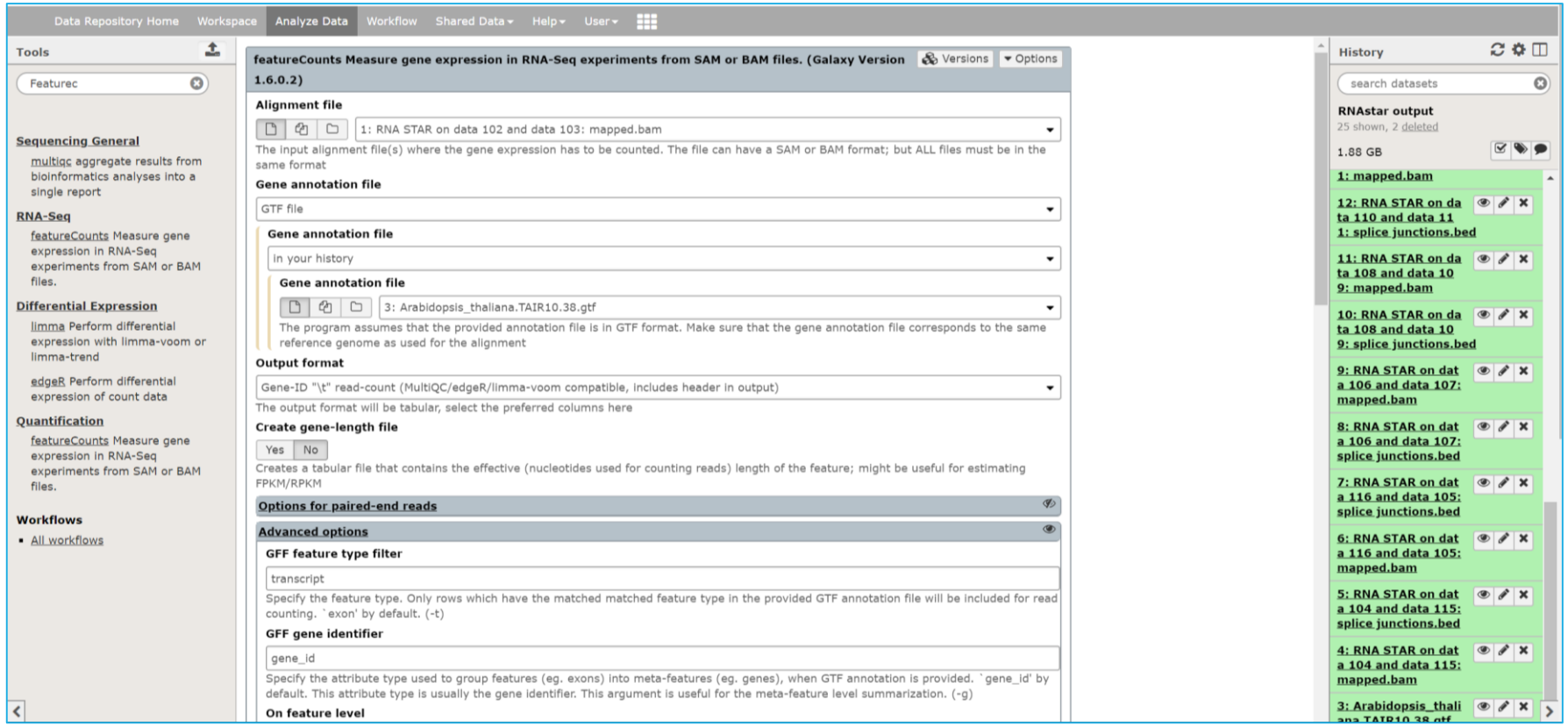
2





Started job on	Jul 09 17:27:51
Started mapping on	Jul 09 17:28:06
Finished on	Jul 09 17:30:28
Mapping speed, Million of reads per hour	74.74
Number of input reads	2948234
Average input read length	143
UNIQUE READS:	
Uniquely mapped reads number	2383614
Uniquely mapped reads %	80.85%
Average mapped length	143.20
Number of splices: Total	1101561
Number of splices: Annotated (sjdb)	1086836
Number of splices: GT/AG	1088404
Number of splices: GC/AG	7493
Number of splices: AT/AC	1119
Number of splices: Non-canonical	4545
Mismatch rate per base, %	0.28%
Deletion rate per base	0.03%
Deletion average length	1.42
Insertion rate per base	0.00%
Insertion average length	1.10
MULTI-MAPPING READS:	
Number of reads mapped to multiple loci	398728
% of reads mapped to multiple loci	13.52%
Number of reads mapped to too many loci	1333
% of reads mapped to too many loci	0.05%
UNMAPPED READS:	
% of reads unmapped: too many mismatches	0.00%
% of reads unmapped: too short	5.58%
% of reads unmapped: other	0.00%
CHIMERIC READS:	
Number of chimeric reads	0
% of chimeric reads	0.00%

COUNT : CALL FEATURE COUNTS



Alignment File
- RNA Star output : mapped.bam file

Gene annotation file
- GTF file

Gene annotation file
- in your history

Gene annotation file
- Arabidopsis_taliana.TAIR10.38.gtf

Output format
- Gene-ID "\t" read-count (MultiQC/edgeR/limma-voom compatible, includes header in output)

Create gene-length file
- False

Options for paired-end reads:
- Default

Advanced options:

GFF feature type filter
- Transcript or gene

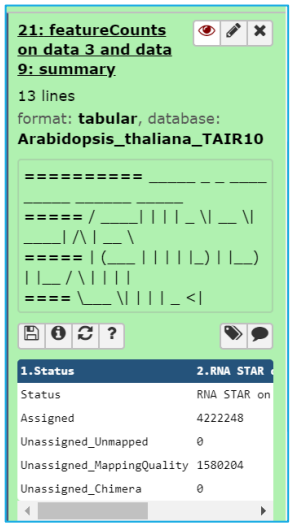
GFF gene identifier
- transcript_id or gene_id

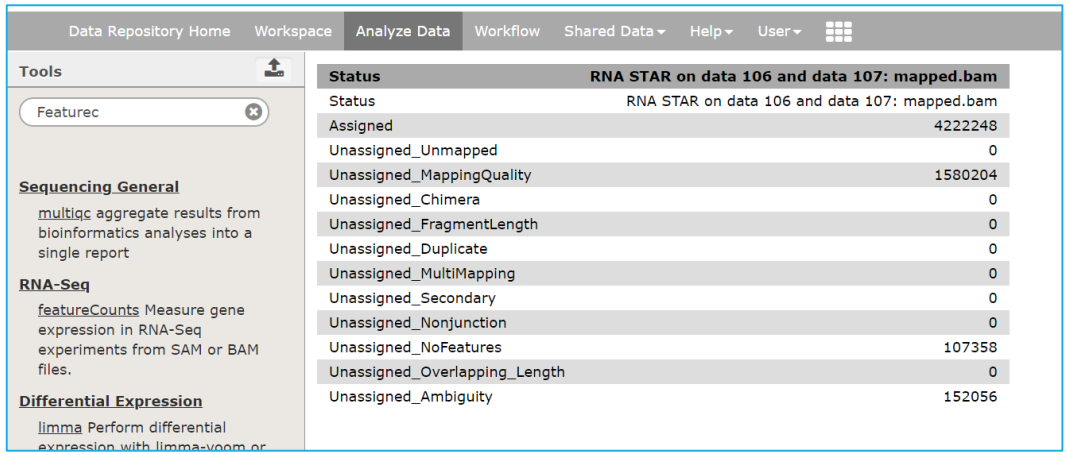
Rest
- Default

Gene Annotation files can be found under Shared Data → Data Libraries → Annotation Files → GTF folder. Find the GTF for your organism and import to history.

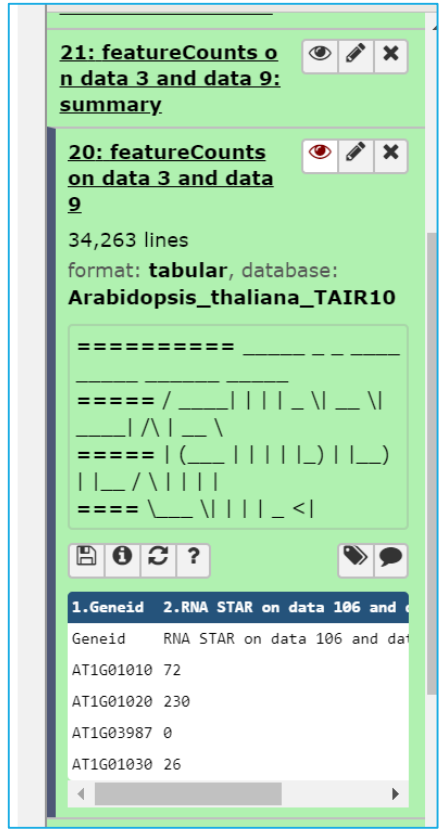
FEATURE COUNTS RESULT

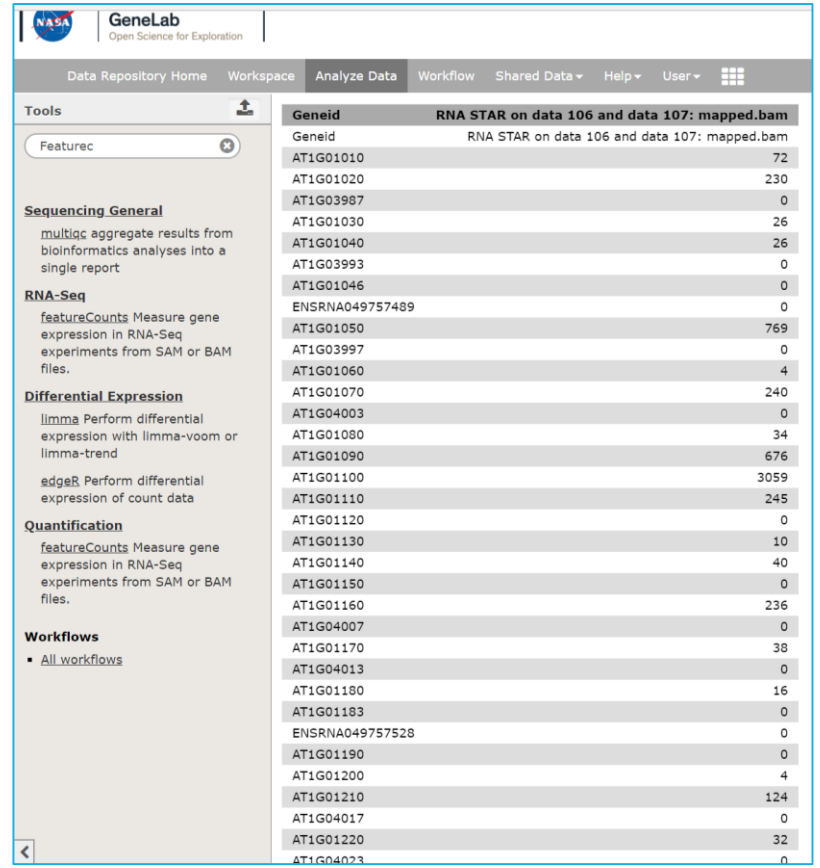
1



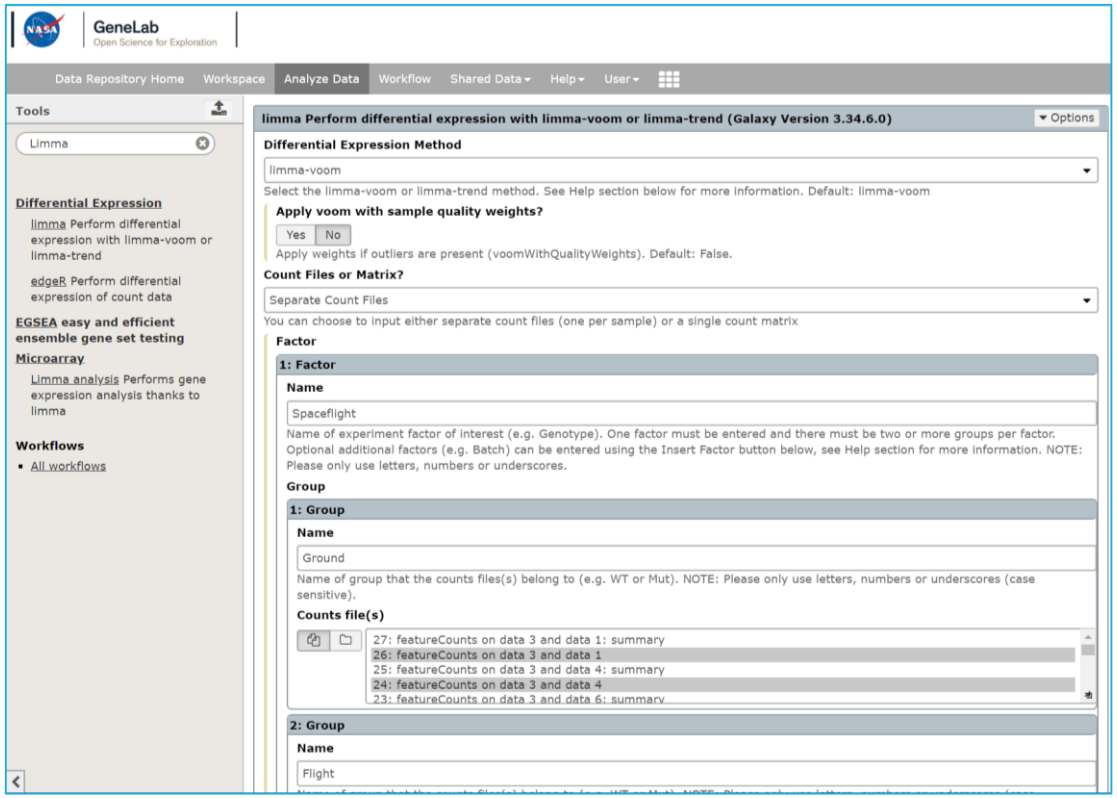


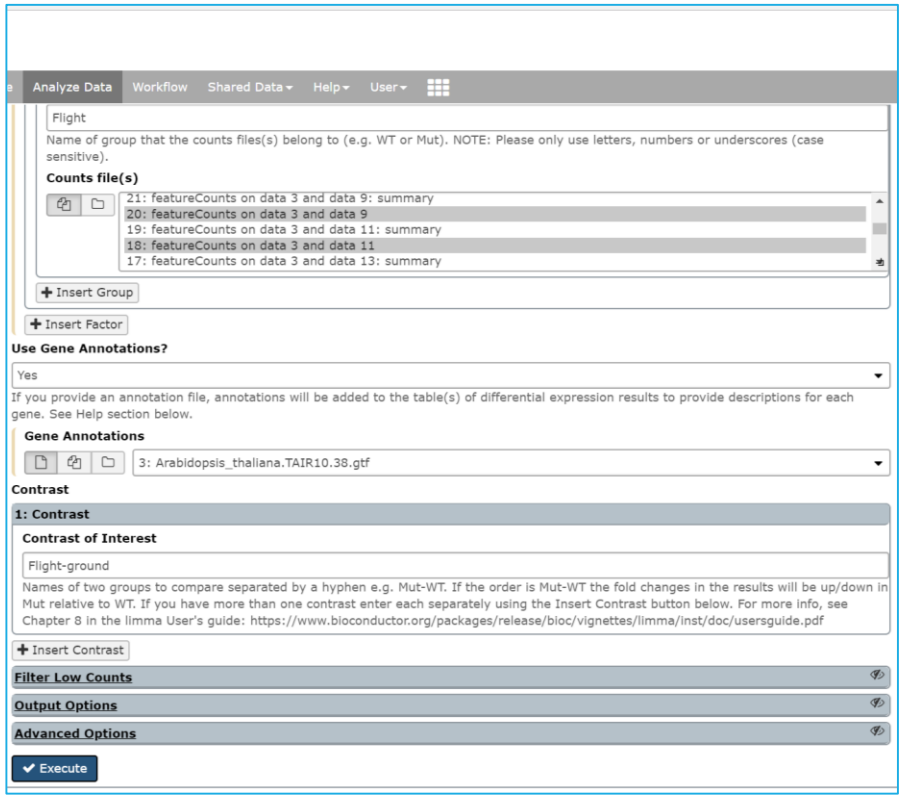
2





DGE: CALL LIMMA





Differential Expression Method
- Limma-voom

Apply voom with sample quality weights
- No

Count files or Matrix
- Separate Count Files

Factor

Name
- Spaceflight (Add your choice)

Group

Name
- Ground (Add your control)
- Choose featureCounts on ground data

Group

Name
- Flight (Add your test)
- Choose featureCounts on Flight data

Use Gene Annotations
- Yes

Gene Annotations
- Arabidopsis_taliana.TAIR10.38.gtf

Contrast

Contrast of Interest
- Flight-Ground (No space, same spelling)

Rest
- Default

LIMMA RESULTS

GeneLab
Open Science for Exploration

Data Repository Home | Workspace | Analyze Data | Workflow | Shared Data | Help | User

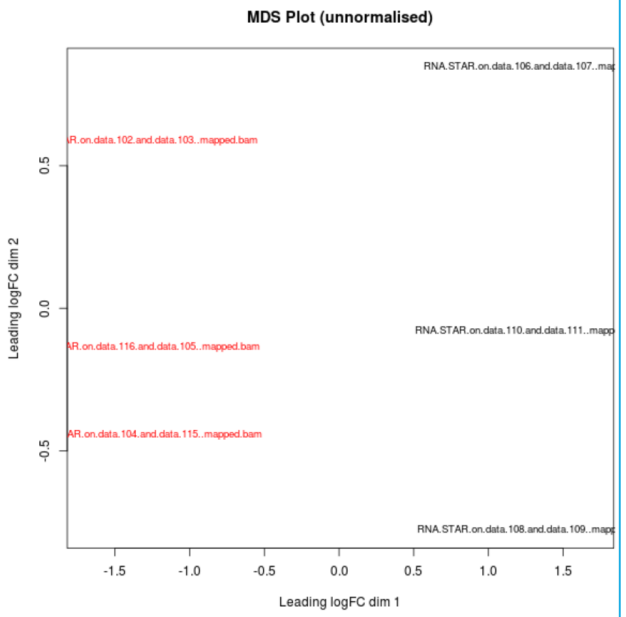
Tools

search tools

Sequencing General
Epigenomics
RNA-Seq
Differential Expression
Get Data
Send Data
Collection Operations
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Statistics
Graph/Display Data
Quantification
Microbiome Analysis
Assembly
Proteomics
Phylogenetics
Annotation
EGSEA easy and efficient ensemble gene set testing
Pathway / Enrichment Analysis
icroarray

Limma Analysis Output:

PDF copies of JPEGS available in 'Plots' section.



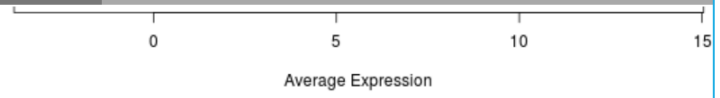
GeneLab
Open Science for Exploration

Data Repository Home | Workspace | Analyze Data | Workflow | Shared Data | Help | User

Tools

search tools

Sequencing General
Epigenomics
RNA-Seq
Differential Expression
Get Data
Send Data
Collection Operations
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Statistics
Graph/Display Data
Quantification
Microbiome Analysis
Assembly



Differential Expression Counts:

	Up	Flat	Down
Flight-Ground	2209	30270	1783

Plots:

MDS Plot (unnormalised).pdf
Voom Plot.pdf
MDS Plot (normalised).pdf
MA Plot_Flight-Ground (.pdf)

Tables:

limma-voom_Flight-Ground.tsv

Alt-click links to download file.

Click floppy disc icon associated history item to download all files.

.tsv files can be viewed in Excel or any spreadsheet program.

History

search datasets

Feature count output
15 shown, 1 hidden

248.93 MB

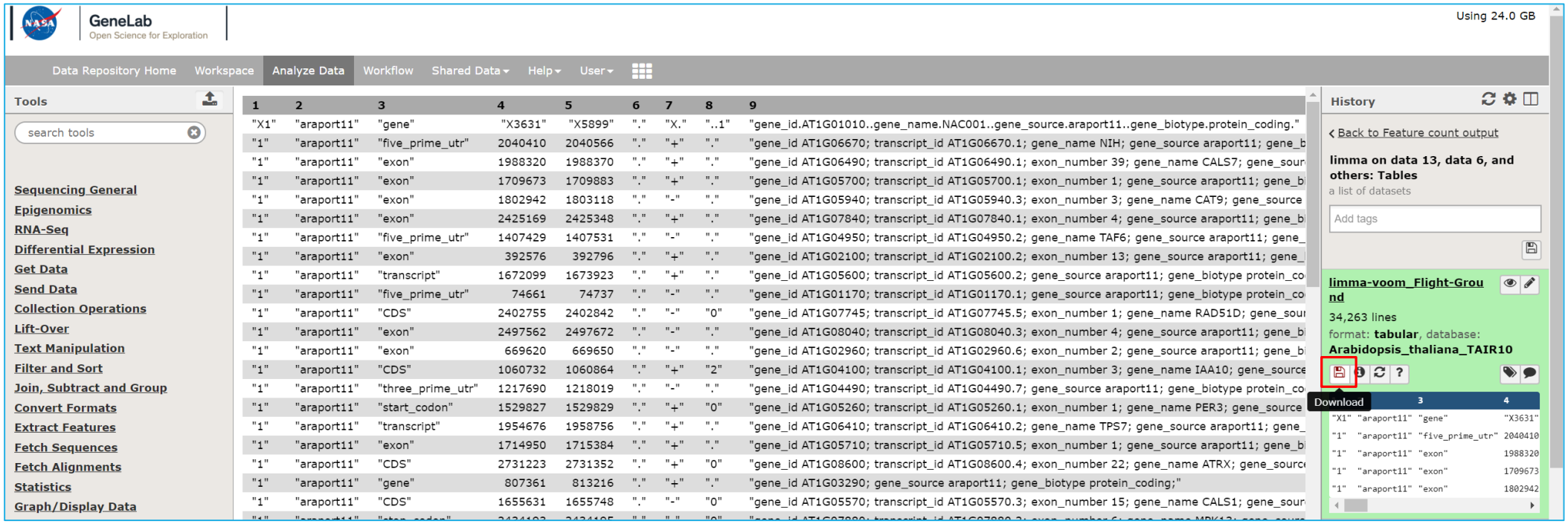
15: limma on data 13, data 6, and others: Report

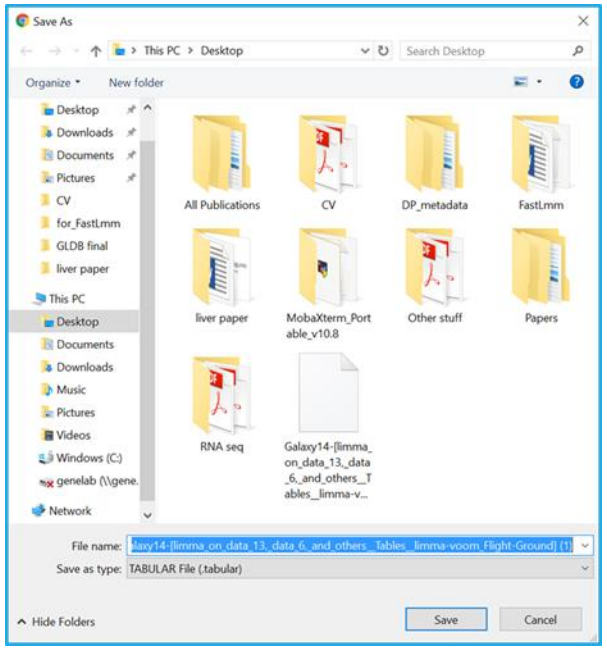
5.0 KB
format: **html**, database: **Arabidopsis_thaliana_TAIR10**

[1] "Extracting counts"
[1] "Generating DGEList object"
[1] "Generating Design"
[1] "Calculating Normalisation Factors"
[1] "Generating Contrasts"
[1] "Generating MDS plot"
[1] "Generating normalised MDS plot"
[1] "Generating DE results"

HTML file

14: limma on data 13, data 6, and others: Tables
a list of 1 datasets





Save file on local computer