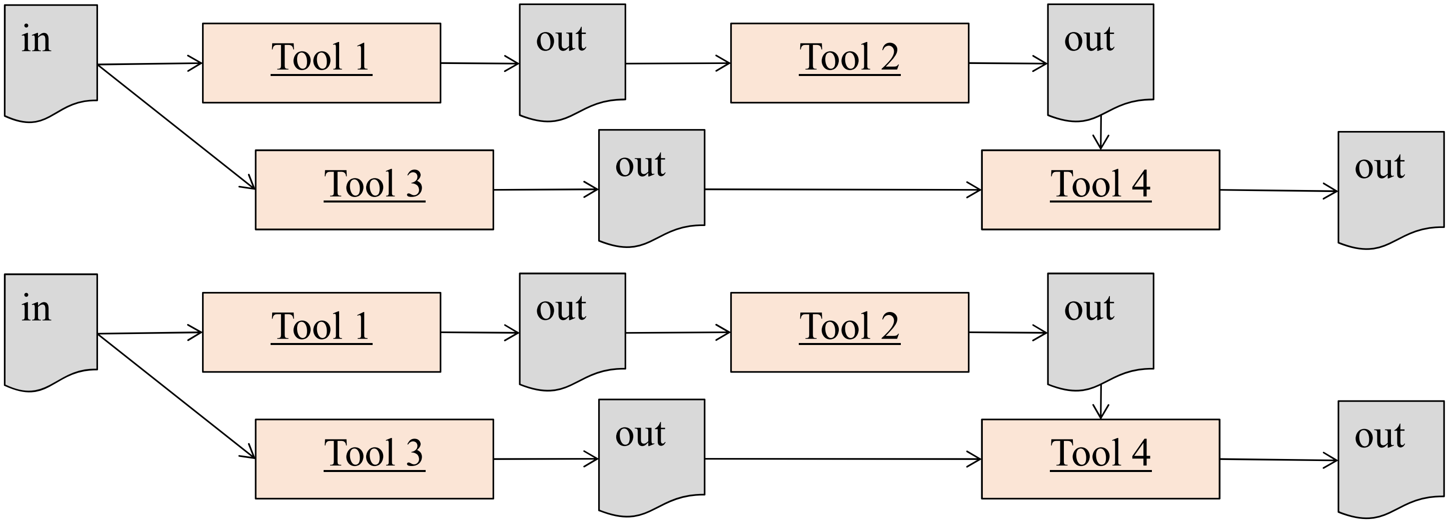
useGalaxy.*

Introduction to Galaxy

What is Galaxy?

- Galaxy is a free web-based scientific analysis platform that can be used by scientists across the world
- Galaxy's tools assist in the analysis of large biological datasets that are common in fields such as genomics or bioinformatics

Motivation of Galaxy



Sequencers throughput require parallel processing of multiple samples

⇒ how do you efficiently monitor all these workflow executions ?

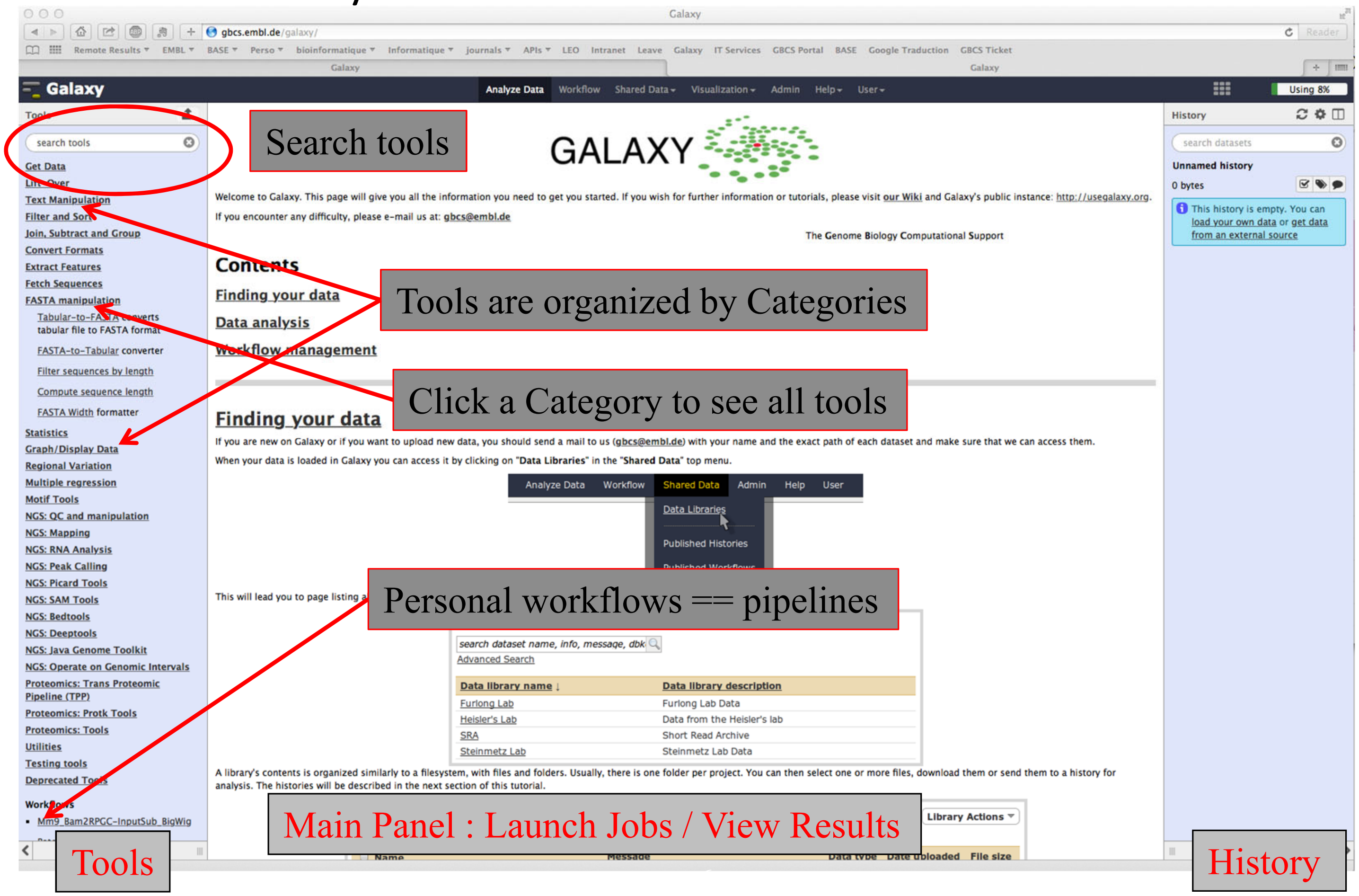
What makes Galaxy special?

- Free
- Available world-wide
- Massive toolbox for data analysis — More than 5500 tools available
- Free training materials — Useful for learning the user interface and several tools
- Helpful community and forums
- Reproducibility — “Workflows” can be saved to show step-by-step uses of tools on data

How to Start

- Go to usegalaxy.org and click the ”Login or Register” tab at the top
- Click on the “Register here” link below the login to begin creating an account
  - Once done filling in the required information, click ”Create”
  - You will be required to verify your account’s email address before being able to use any of the tools or data available
- After creating and verifying your account, you are ready to begin your introduction into Galaxy

Welcome to Galaxy



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

Welcome to Galaxy. This page will give you all the information you need to get you started. If you wish for further information or tutorials, please visit our Wiki and Galaxy's public instance: http://usegalaxy.org.

If you encounter any difficulty, please e-mail us at: gbcs@embl.de

The Genome Biology Computational Support

Contents

Finding your data

Data analysis

Workflow management

Finding your data

If you are new on Galaxy or if you want to upload new data, you should send a mail to us (gbcs@embl.de) with your name and the exact path of each dataset and make sure that we can access them.

When your data is loaded in Galaxy you can access it by clicking on "Data Libraries" in the "Shared Data" top menu.

This will lead you to page listing a

Data library name ↓	Data library description
Furlong Lab	Furlong Lab Data
Heisler's Lab	Data from the Heisler's lab
SRA	Short Read Archive
Steinmetz Lab	Steinmetz Lab Data

A library's contents is organized similarly to a filesystem, with files and folders. Usually, there is one folder per project. You can then select one or more files, download them or send them to a history for analysis. The histories will be described in the next section of this tutorial.

Search tools

Tools are organized by Categories

Click a Category to see all tools

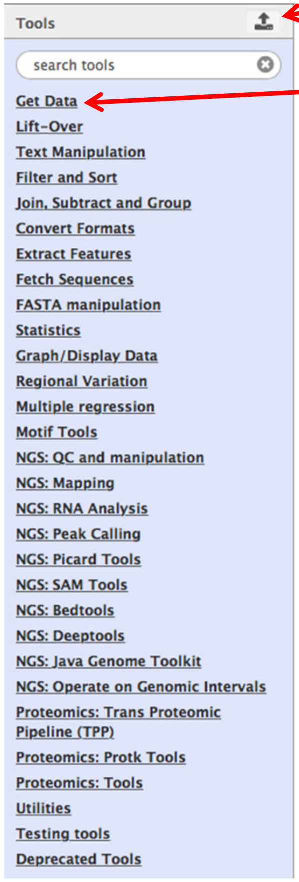
Personal workflows == pipelines

Main Panel : Launch Jobs / View Results

Tools

History

Noticeable Tool Categories



Tools

search tools

- Get Data
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- FASTA manipulation
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Motif Tools
- NGS: QC and manipulation
- NGS: Mapping
- NGS: RNA Analysis
- NGS: Peak Calling
- NGS: Picard Tools
- NGS: SAM Tools
- NGS: Bedtools
- NGS: Deeptools
- NGS: Java Genome Toolkit
- NGS: Operate on Genomic Intervals
- Proteomics: Trans Proteomic Pipeline (TPP)
- Proteomics: Protk Tools
- Proteomics: Tools
- Utilities
- Testing tools
- Deprecated Tools

Upload personal files

Fetch public data from UCSC, ENSEMBL, mouseMINE…

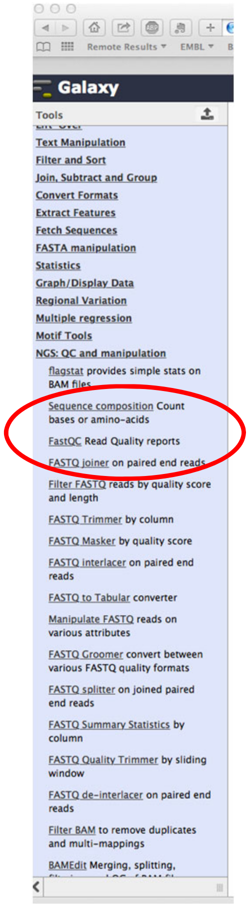
General sequence and text format manipulation Tools

NGS Tools organized by:
- application (QC, mapping, peak calling, RNA, …)
- package name (bedtools, picard)

Proteomics Tools

Utility (local), Test (beta version) and Deprecated (for backward comp.)

Running a Tool is easy



Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

FASTA manipulation

Statistics

Graph/Display Data

Regional Variation

Multiple regression

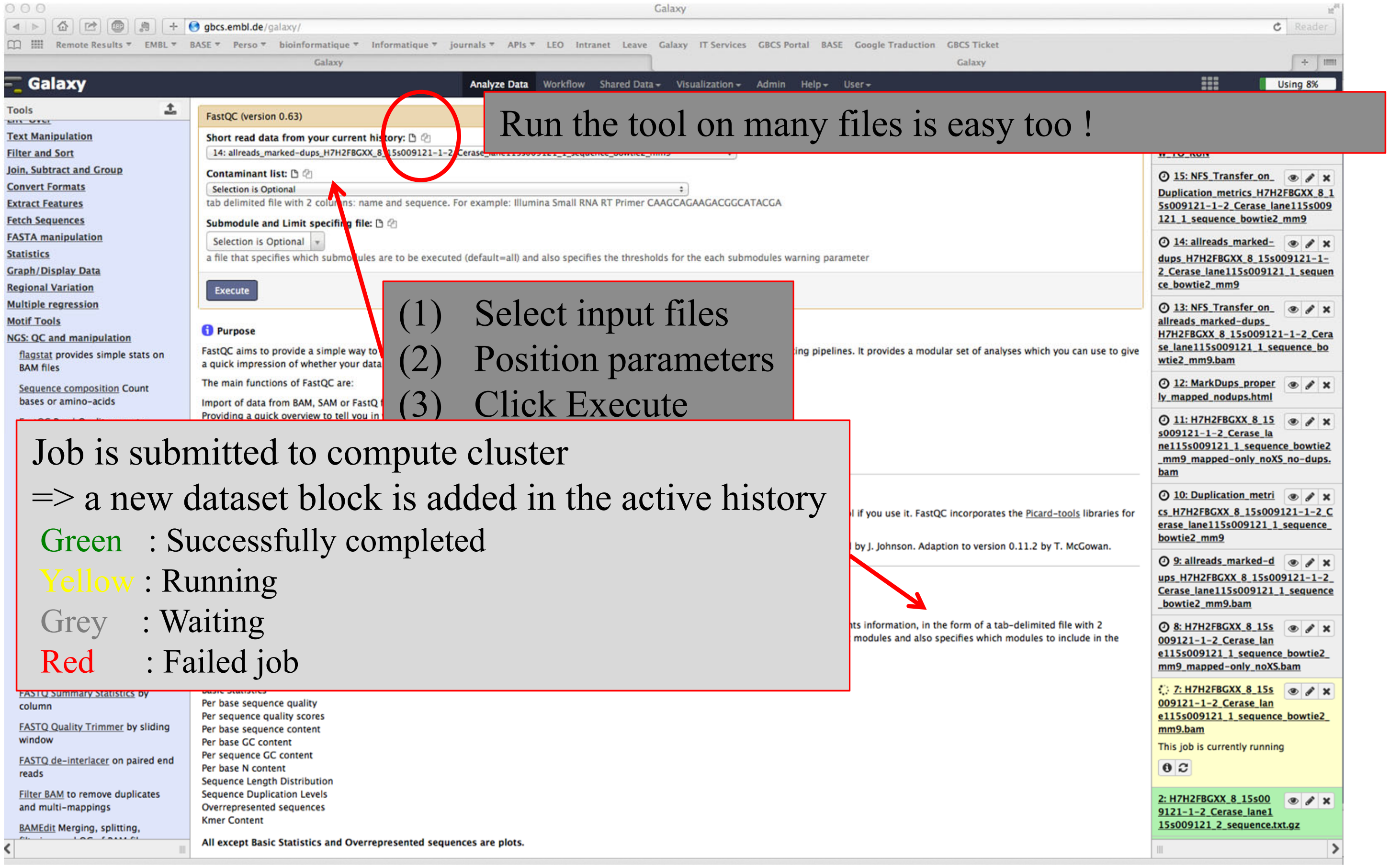
Motif Tools

NGS: QC and manipulation

- flagstat provides simple stats on BAM files
- Sequence composition Count bases or amino-acids
- FastQC Read Quality reports
- FASTQ joiner on paired end reads
- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column
- FASTQ Masker by quality score
- FASTQ interlacer on paired end reads
- FASTQ to Tabular converter
- Manipulate FASTQ reads on various attributes
- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ Summary Statistics by column
- FASTQ Quality Trimmer by sliding window
- FASTQ de-interlacer on paired end reads
- Filter BAM to remove duplicates and multi-mappings
- BAMEdit Merging, splitting,

Click a tool to bring it up in the middle panel eg FastQC

Running a Tool is easy



Run the tool on many files is easy too !

(1) Select input files
(2) Position parameters
(3) Click Execute

Job is submitted to compute cluster
=> a new dataset block is added in the active history
Green : Successfully completed
Yellow : Running
Grey : Waiting
Red : Failed job

flagstat provides simple stats on BAM files

Sequence composition Count bases or amino-acids

FASTQ Summary Statistics by column

FASTQ Quality Trimmer by sliding window

FASTQ de-interlacer on paired end reads

Filter BAM to remove duplicates and multi-mappings

BAMEdit Merging, splitting,

FastQC (version 0.63)

Short read data from your current history:
14: allreads_marked-dups_H7H2FBGXX_8_15s009121-1-2_Cerase_lane115s009121_1_sequence_bowtie2_mm9

Contaminant list:
Selection is Optional
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Submodule and Limit specifing file:
Selection is Optional
a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Purpose

FastQC aims to provide a simple way to … ing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data

The main functions of FastQC are:

Import of data from BAM, SAM or FastQ f
Providing a quick overview to tell you in

l if you use it. FastQC incorporates the Picard-tools libraries for

by J. Johnson. Adaption to version 0.11.2 by T. McGowan.

ts information, in the form of a tab-delimited file with 2 … modules and also specifies which modules to include in the

Per base sequence quality
Per sequence quality scores
Per base sequence content
Per base GC content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Kmer Content

All except Basic Statistics and Overrepresented sequences are plots.

15: NFS_Transfer_on_Duplication_metrics_H7H2FBGXX_8_15s009121-1-2_Cerase_lane115s009121_1_sequence_bowtie2_mm9

14: allreads_marked-dups_H7H2FBGXX_8_15s009121-1-2_Cerase_lane115s009121_1_sequence_bowtie2_mm9

13: NFS_Transfer_on_allreads_marked-dups_H7H2FBGXX_8_15s009121-1-2_Cerase_lane115s009121_1_sequence_bowtie2_mm9.bam

12: MarkDups_properly_mapped_nodups.html

11: H7H2FBGXX_8_15s009121-1-2_Cerase_lane115s009121_1_sequence_bowtie2_mm9_mapped-only_noXS_no-dups.bam

10: Duplication_metrics_H7H2FBGXX_8_15s009121-1-2_Cerase_lane115s009121_1_sequence_bowtie2_mm9

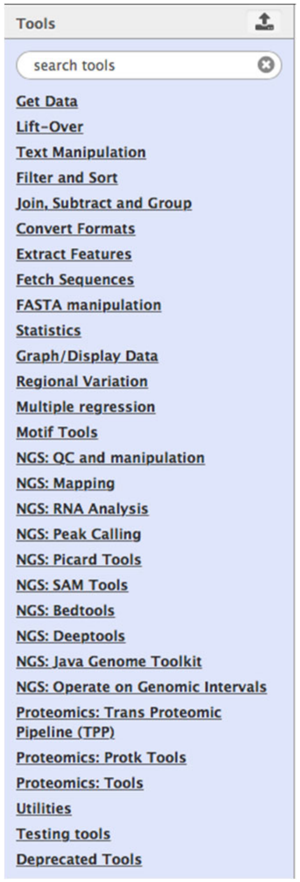
9: allreads_marked-dups_H7H2FBGXX_8_15s009121-1-2_Cerase_lane115s009121_1_sequence_bowtie2_mm9.bam

8: H7H2FBGXX_8_15s009121-1-2_Cerase_lane115s009121_1_sequence_bowtie2_mm9_mapped-only_noXS.bam

7: H7H2FBGXX_8_15s009121-1-2_Cerase_lane115s009121_1_sequence_bowtie2_mm9.bam
This job is currently running

2: H7H2FBGXX_8_15s009121-1-2_Cerase_lane115s009121_2_sequence.txt.gz

Tool summary



More than 350 tools available !

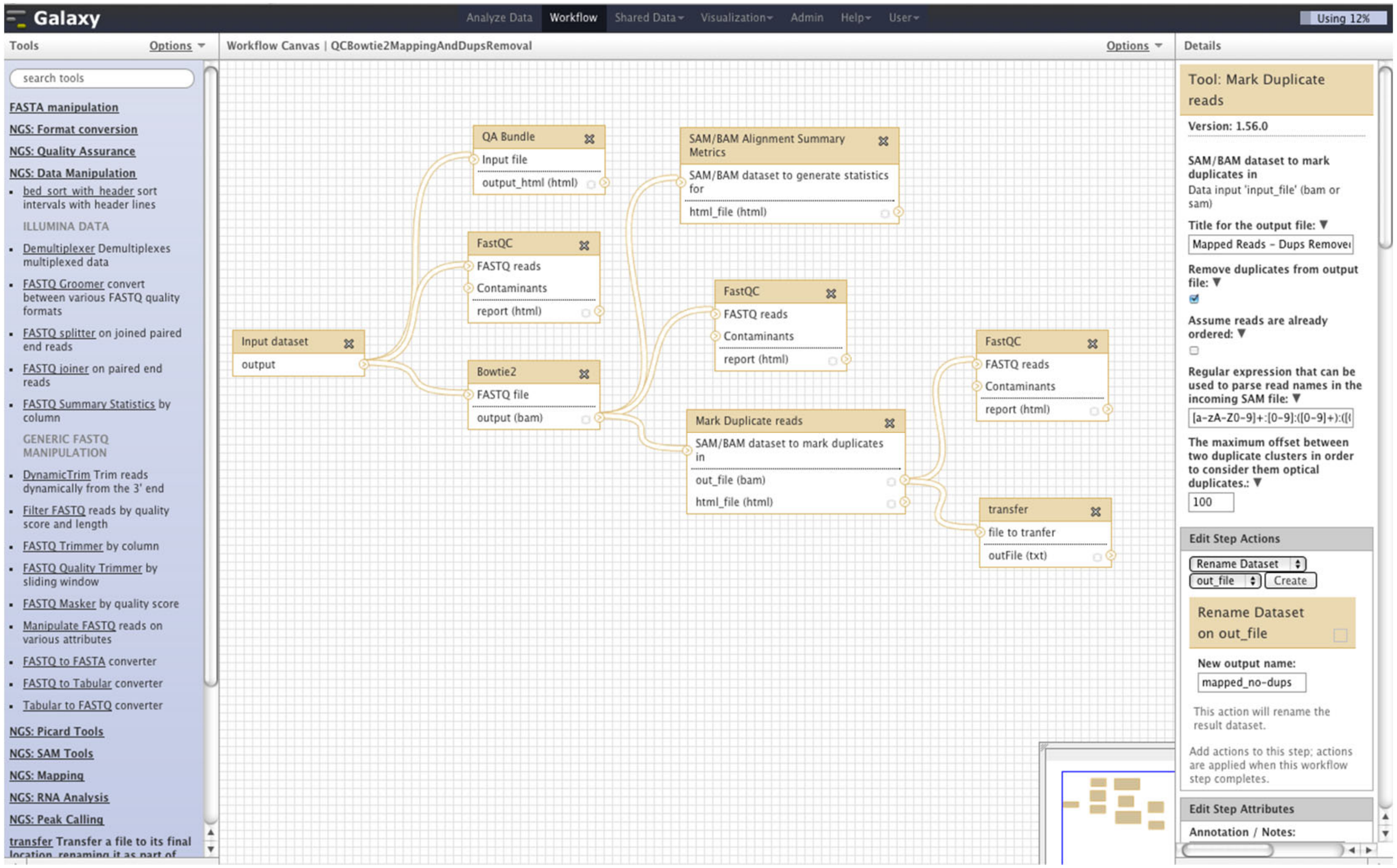
✓ All results can be downloaded or directly transferred to your project folder

✓ Missing Tools can be easily integrated

✓ Easy way to add a GUI to your own script

✓ Parameters for cluster submission can be adjusted for each tool (and even be dynamically computed)

Tools can be assembled into workflows


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 12%
Tools
Options
search tools
FASTA manipulation
NGS: Format conversion
NGS: Quality Assurance
NGS: Data Manipulation
bed_sort_with_header sort intervals with header lines
ILLUMINA DATA
Demultiplexer Demultiplexes multiplexed data
FASTQ Groomer convert between various FASTQ quality formats
FASTQ splitter on joined paired end reads
FASTQ joiner on paired end reads
FASTQ Summary Statistics by column
GENERIC FASTQ MANIPULATION
DynamicTrim Trim reads dynamically from the 3' end
Filter FASTQ reads by quality score and length
FASTQ Trimmer by column
FASTQ Quality Trimmer by sliding window
FASTQ Masker by quality score
Manipulate FASTQ reads on various attributes
FASTQ to FASTA converter
FASTQ to Tabular converter
Tabular to FASTQ converter
NGS: Picard Tools
NGS: SAM Tools
NGS: Mapping
NGS: RNA Analysis
NGS: Peak Calling
transfer Transfer a file to its final location, renaming it as part of
Workflow Canvas | QCBowtie2MappingAndDupsRemoval
Options
Input dataset
output
QA Bundle
Input file
output_html (html)
FastQC
FASTQ reads
Contaminants
report (html)
Bowtie2
FASTQ file
output (bam)
SAM/BAM Alignment Summary Metrics
SAM/BAM dataset to generate statistics for
html_file (html)
FastQC
FASTQ reads
Contaminants
report (html)
Mark Duplicate reads
SAM/BAM dataset to mark duplicates in
out_file (bam)
html_file (html)
FastQC
FASTQ reads
Contaminants
report (html)
transfer
file to tranfer
outFile (txt)
Details

Tool: Mark Duplicate reads

Version: 1.56.0

SAM/BAM dataset to mark duplicates in
Data input 'input_file' (bam or sam)

Title for the output file: ▼
Mapped Reads - Dups Removei

Remove duplicates from output file: ▼
☑

Assume reads are already ordered: ▼
☐

Regular expression that can be used to parse read names in the incoming SAM file: ▼
[a-zA-Z0-9]+:[0-9]:([0-9]+):([(

The maximum offset between two duplicate clusters in order to consider them optical duplicates.: ▼
100

Edit Step Actions

Rename Dataset
out_file Create

Rename Dataset on out_file ☐

New output name:
mapped_no-dups

This action will rename the result dataset.

Add actions to this step; actions are applied when this workflow step completes.

Edit Step Attributes

Annotation / Notes:

Example of Galaxy Usage

- “*RNA–Seq Data Analysis in Galaxy*” is an academic article published on PubMed and SpringerLink
- The purpose of this article is to showcase the tools and steps involved in RNA-Seq analysis meant to find differentially expressed genes with enrichment analysis

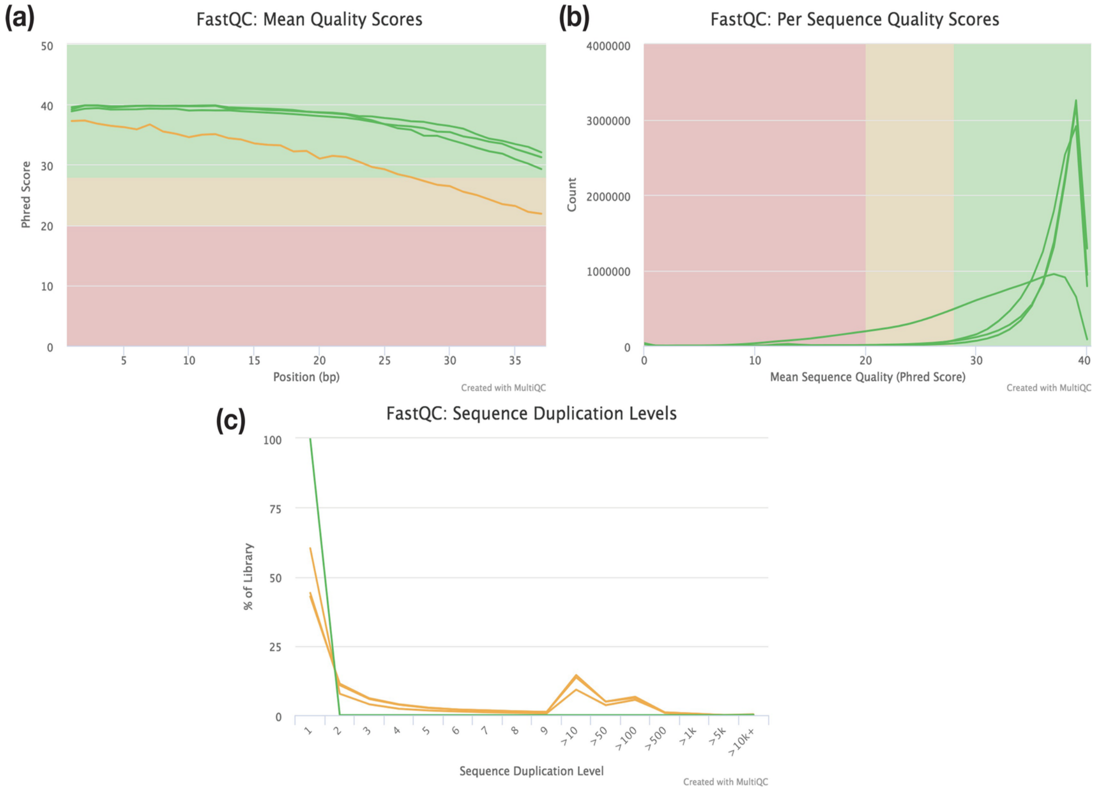
Data Used

- This article used real data for analysis based on a study conducted by Brooks AN, Yang L, Duff MO et al (2011) Conservation of an RNA regulatory map between Drosophila and mammals.
  - This referenced article found genes and pathway that regulated the pasilla gene of *Drosophila melanogaster* and discovered RNAi could deplete the gene
  - They then created RNA-seq libraries for treated and untreated samples of the pasilla gene to gather RNA-seq reads and compare them to monitor the effects of pasilla depletion on gene expression
- Seven of the original datasets were sampled and used for analysis
  - Four untreated samples: GSM461176, GSM461177, GSM461178, GSM461182
  - Three treated samples: GSM461179, GSM461180, and GSM461181

Methods Used: Quality Control

- FASTQ files are taken for the samples and the reads undergo quality control using
  - FastQC: a tool used to perform quality control checks on raw sequence data that is coming from high throughput sequencing pipelines
    - Creates a QC report that flags any issues with the data that the user should be aware of before continuing analysis
  - Cutadapt: a tool used to find and remove any adapter sequences, primers, poly-A tails and other types of unwanted sequences from high-throughput sequencing reads
- Each step up until “Counting” is followed by the use of the MultiQC tool for quality checks
  - MultiQC searches a directory and logs/compiles a report viewable via HTML
    - Useful for summarizing outputs of numerous bioinformatics tools

Methods Used: Quality Control (FASTQC)*

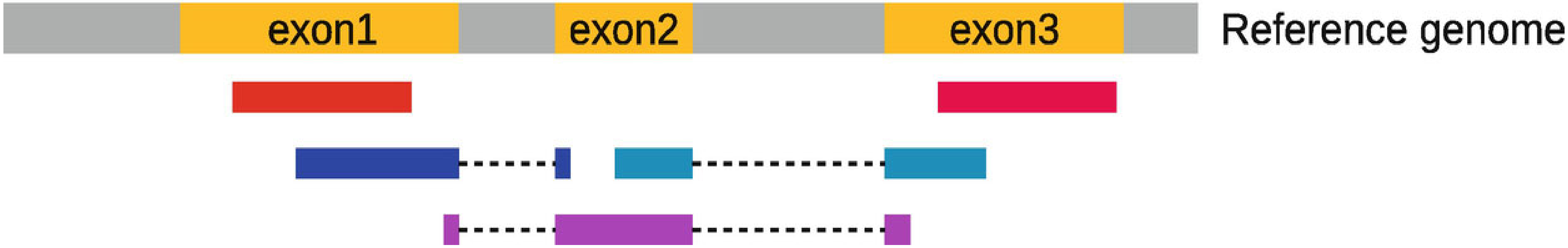


*This is GSM461180 for an example

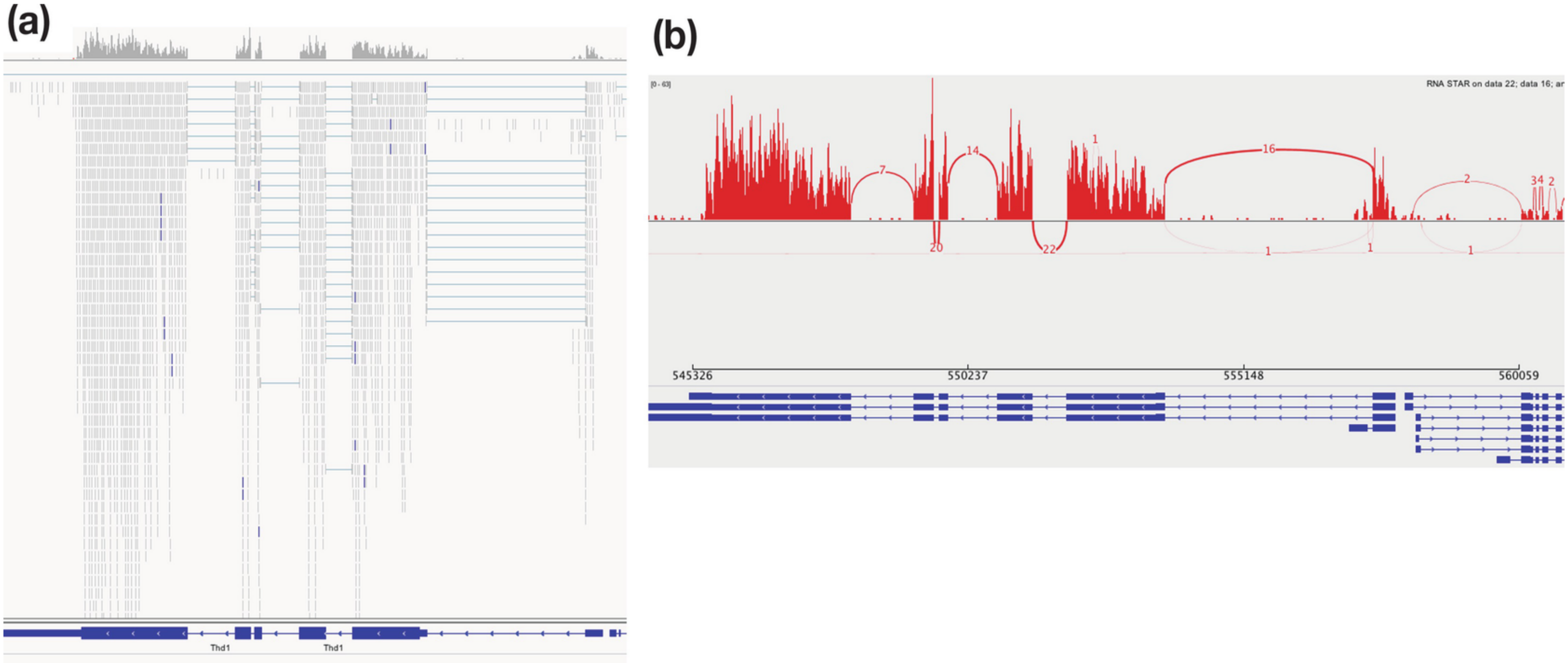
Methods Used: Mapping

- After quality control, the data is mapped to a reference genome of *D. melanogaster* using STAR in order to make sense of the reads collected
  - Used to find where the sequences originated from within the genome and which genes they belong to
  - STAR is used for ultra fast alignment of RNA-seq data
    - Often useful for mapping RNA data from RNA-seq or CLIP experiments
    - The MultiQC revealed that 80% of the reads were able to map exactly once to the referenced genome, anything below 70% could indicate contamination of the sample
- After mapping with STAR, the mapping is checked using Integrative Genomics Viewer (IGV) and some other tools
  - IGV is a visualization tool capable of displaying next-generation sequencing data
  - Here we will see an example of IGV creating a Sashimi plot to analyze splice junctions and reads

Methods Used: Mapping (STAR)



Methods Used: Mapping (IGV)



Methods Used: Counting

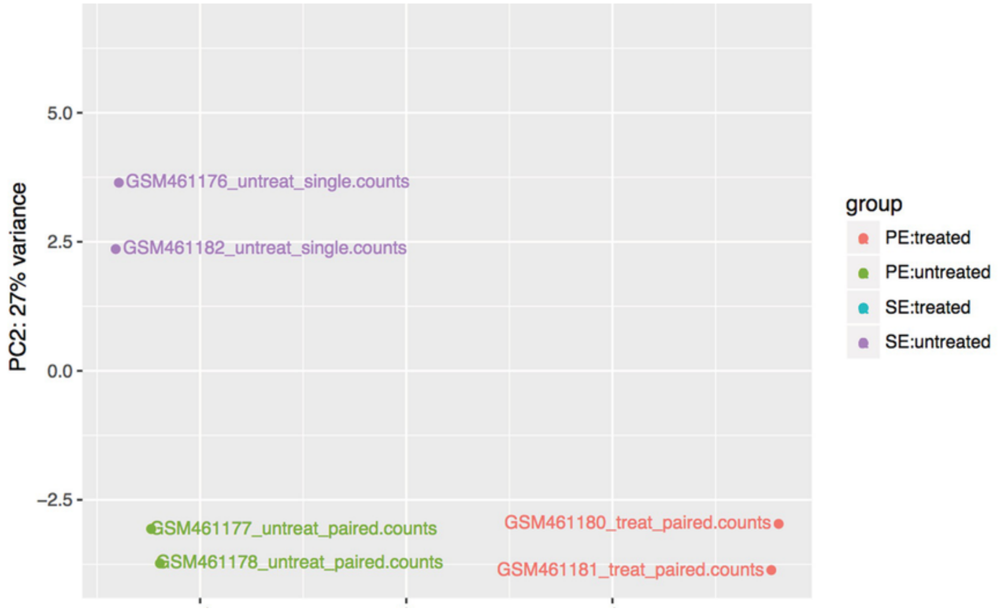
- Once the mapping was checked, the number of reads per annotated gene is counted using the featureCounts tool
  - featureCounts is a tool that is used for quantifying reads generated from RNA or DNA-seq technologies
    - Uses chromosome hashing, feature blocking, and more to read features with high efficiency
    - This is used to compare expressions of genes between different conditions in order to quantify the number of reads per gene/the number of reads mapping to the exons of each gene
    - Outputs include
      - A table with the numbers of reads mapped to each gene
      - A file with the length of each gene

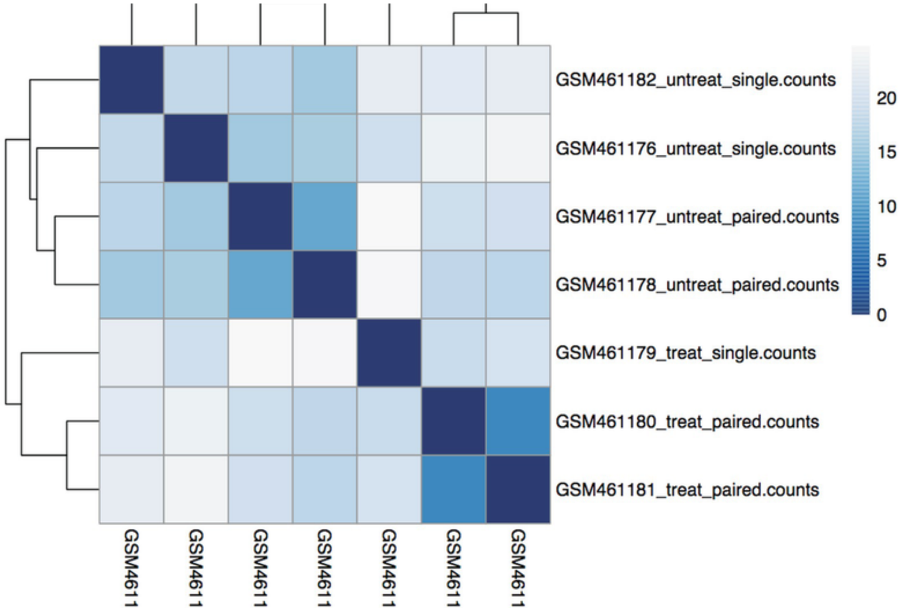
Methods Used: Identification of differentially expressed features; Extraction and Annotation

- DESeq2 extracted the differentially expressed genes and can annotate them.
- The main output of DESeq2 is a summary file that contains certain values relating to each gene
  - Gene identifier
  - Mean normalized counts
  - Fold change in log2
    - Compares fold changes between treated and untreated samples and the values correlate to up- or down-regulation of the gene
  - Standard error estimate for fold change
  - Wald statistic
  - P-values for significance of seen changes
  - All of these outputs can be used to create graphical summaries

Methods Used: Identification of differentially expressed features (DESeq2)

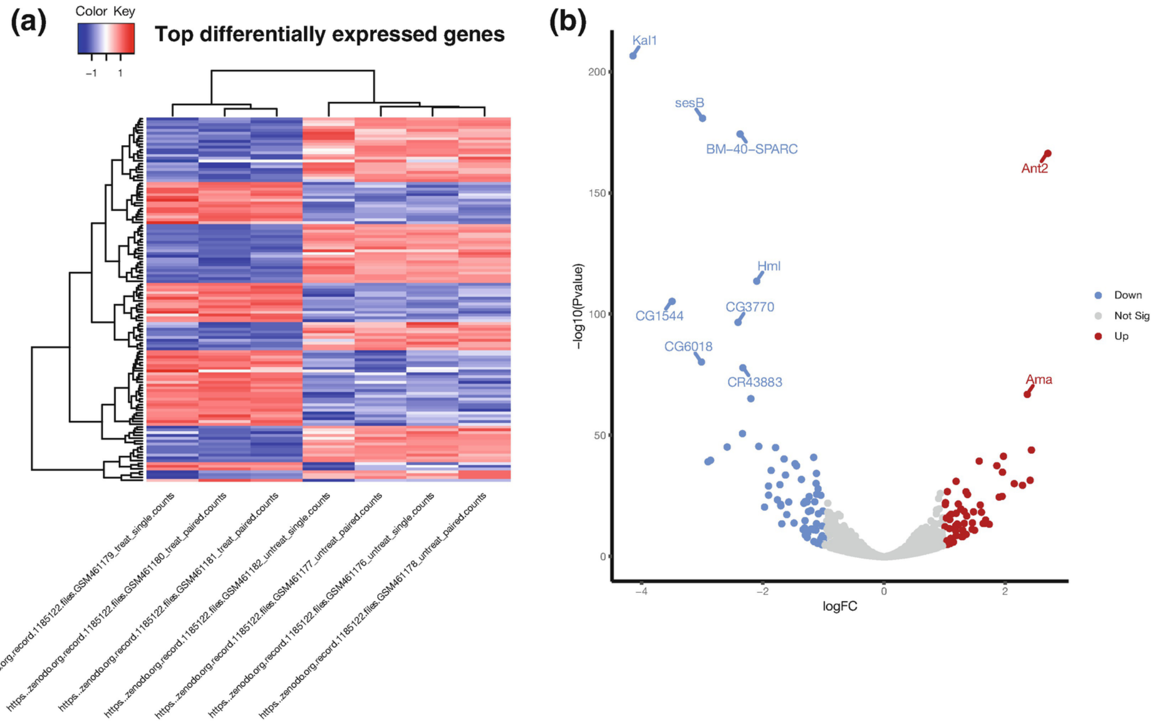
- After checking the counting, DESeq2 was used on these counts to normalize them and extract any differentially expressed genes
  - DESeq2 is a tool designed to normalize, visualize, and implement differential analysis of high-dimensional count data





Methods Used: Visualization (Heatmap2 and Volcano Plot)

- The data was then visualized using Heatmap2 and Volcano Plots to assign colors based on gene expression
  - Red colors indicate significant overexpression
  - Blue colors indicate significant underexpression



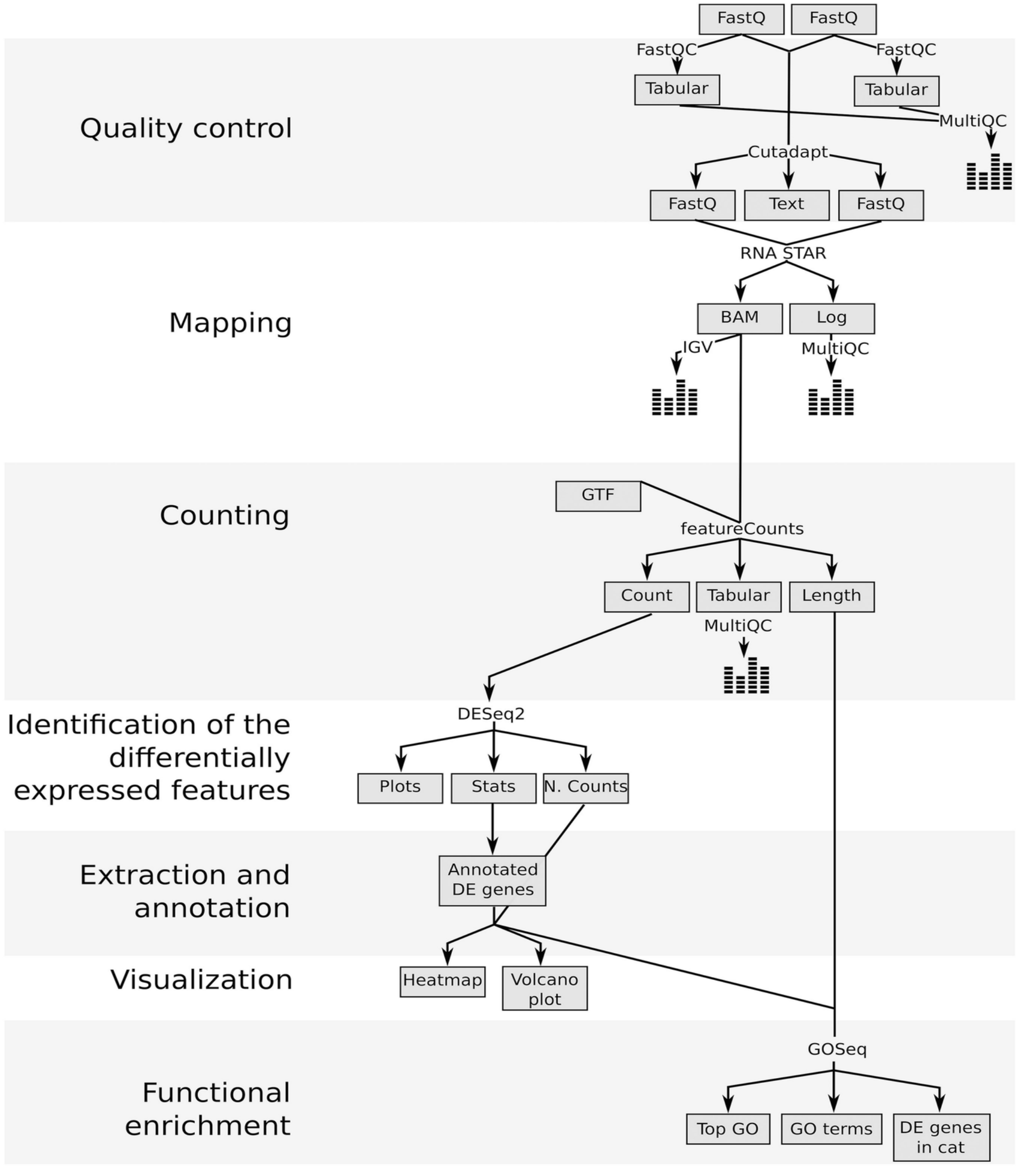
Methods Used: Functional Enrichment Analysis

- Enrichment analysis can now be conducted in order to tell if the differentially expressed genes mentioned before are tied to any genes which play a role in biological function and see how they could be impacted
  - Gene ontology (GO) analysis is popularly used for this analysis as it helps to highlight biological processes within genome expression studies

Methods Used: Functional Enrichment Analysis Cont.

- The tool being used here will be goseq
  - goseq allows GO analysis to be performed on RNA-seq data while taking gene length biases into account
  - The output will be a table with the following columns
    - GO category
    - p-Value for overrepresentation of the term in differentially expressed genes
    - p-Value for underrepresentation of the term in differentially expressed genes
    - Number of differentially expressed genes
    - Number of genes
    - Details about term
    - Ontology with Molecular Function, Cellular Component, and Biological Process
    - p-Value for overrepresentation of the term in differentially expressed genes adjusted for multiple testing with the Benjamini-Hochberg procedure
    - p-Value for underrepresentation of the term in differentially expressed genes adjusted for multiple testing with the Benjamini-Hochberg procedure

Workflow Used*



*This workflow only shows the processing of two datasets

Where Can You Follow This Analysis?

- The article provides detailed steps for doing this analysis yourself within Galaxy by including:
  - Where to access the data
  - How to download the data
  - How to upload the data
  - What tools to use and under what parameters
  - The meaning behind every step and discussion of said findings

References

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coroar, N., Grüning, B., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research, 46*(W1). https://doi.org/10.1093%2Fnar%2Fgky379

Batut, B., van den Beek, M., Doyle, M. A., & Soranzo, N. (2021). RNA-seq data analysis in galaxy. *Methods in Molecular Biology*, *2284*, 367–392. https://doi.org/10.1007/978-1-0716-1307-8_20

References cont.

- https://galaxyproject.github.io/
- https://galaxyproject.org/usegalaxy/
- usegalaxy.org