While we are waiting:

- Connect to Tufts network (on campus or VPN)
- Chrome browser https://galaxy.cluster.tufts.edu
- Login with Tufts credentials
- Let me know if you have trouble logging in

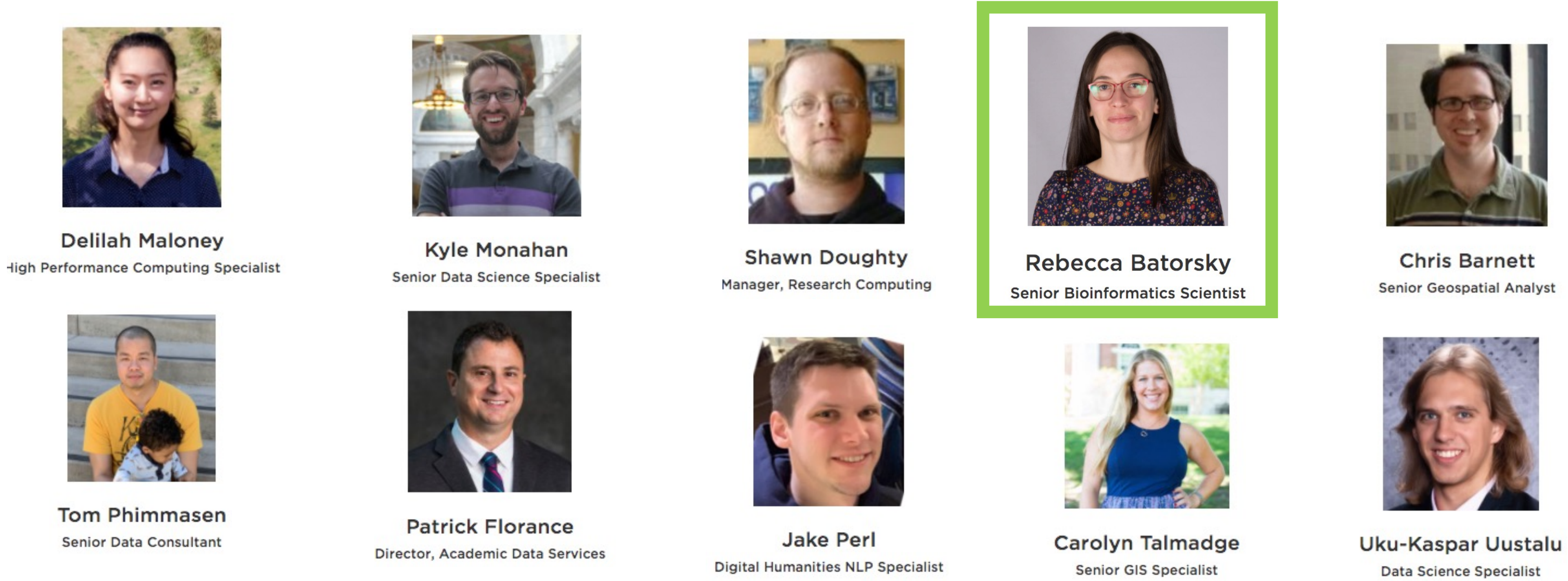
Intro to Next Generation Sequencing Data Analysis with Galaxy

Rebecca Batorsky

Pr Bioinformatics Scientist

Nov 2021

Research Technology Team



Delilah Maloney
High Performance Computing Specialist

Kyle Monahan
Senior Data Science Specialist

Shawn Doughty
Manager, Research Computing

Rebecca Batorsky
Senior Bioinformatics Scientist

Chris Barnett
Senior Geospatial Analyst

Tom Phimmasen
Senior Data Consultant

Patrick Florance
Director, Academic Data Services

Jake Perl
Digital Humanities NLP Specialist

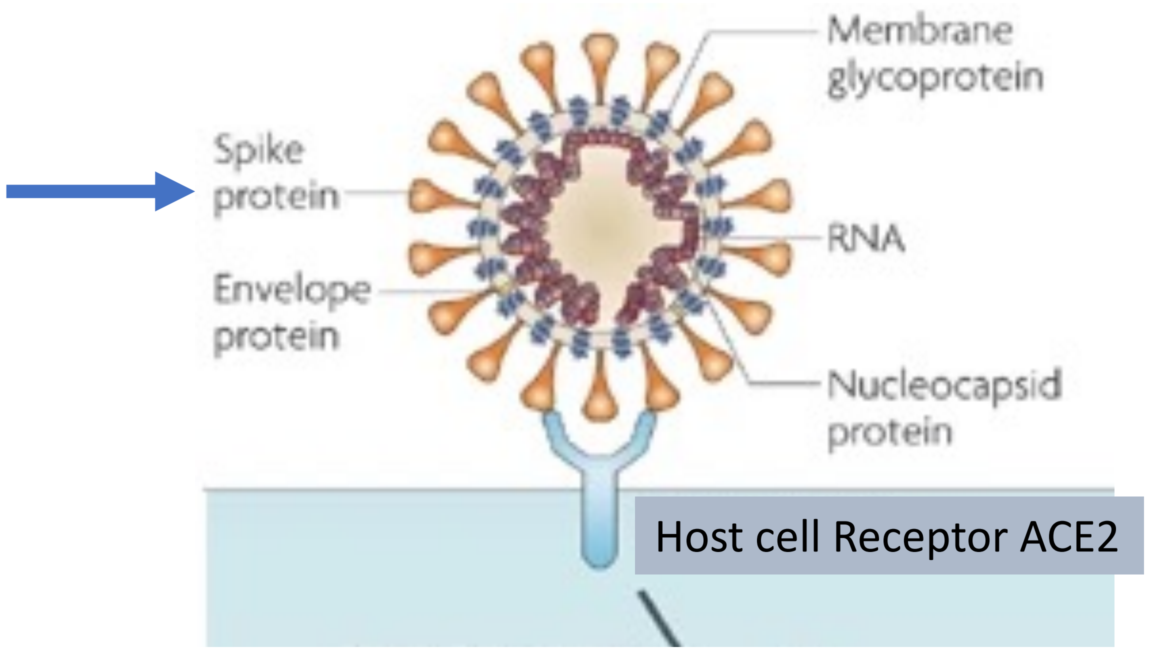
Carolyn Talmadge
Senior GIS Specialist

Uku-Kaspar Uustalu
Data Science Specialist

- ✓ Consultation on Projects and Grants
- ✓ High Performance Compute Cluster
- ✓ Workshops

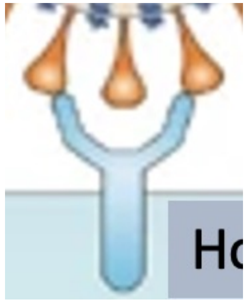
https://it.tufts.edu/research-technology

SARS-CoV2 Spike Protein

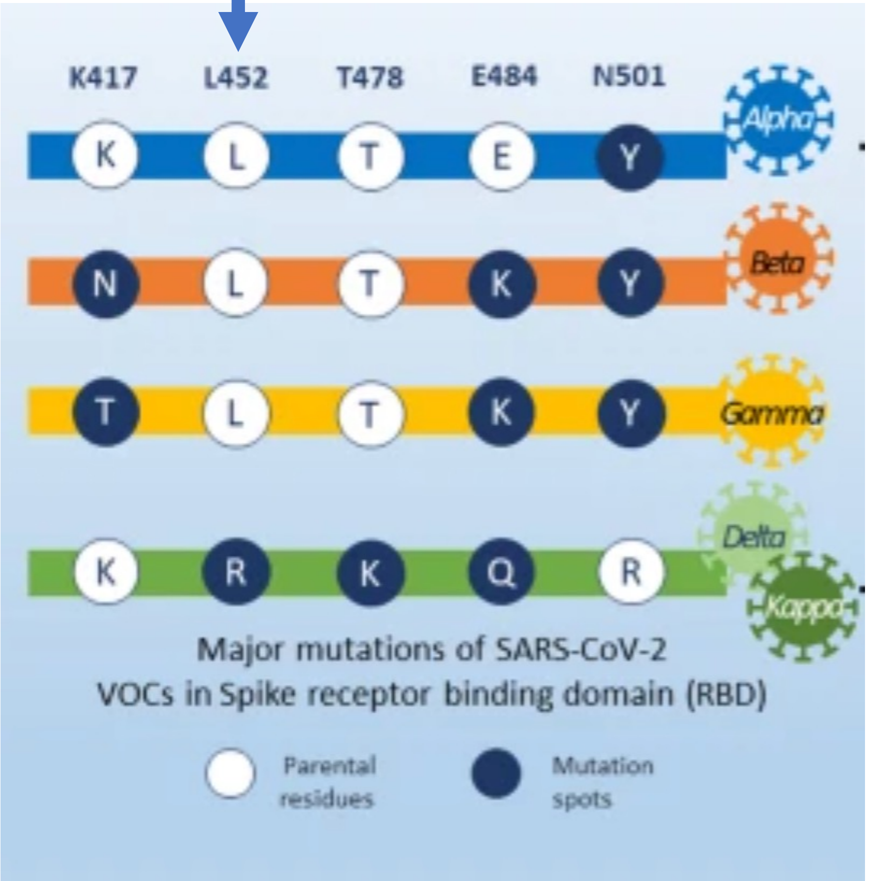


https://www.nature.com/articles/nrmicro2090

SARS-CoV2 Spike Protein VOCs



Spike protein



	K417	L452	T478	E484	N501
Alpha	K	L	T	E	Y
Beta	N	L	T	K	Y
Gamma	T	L	T	K	Y
Delta Kappa	K	R	K	Q	R

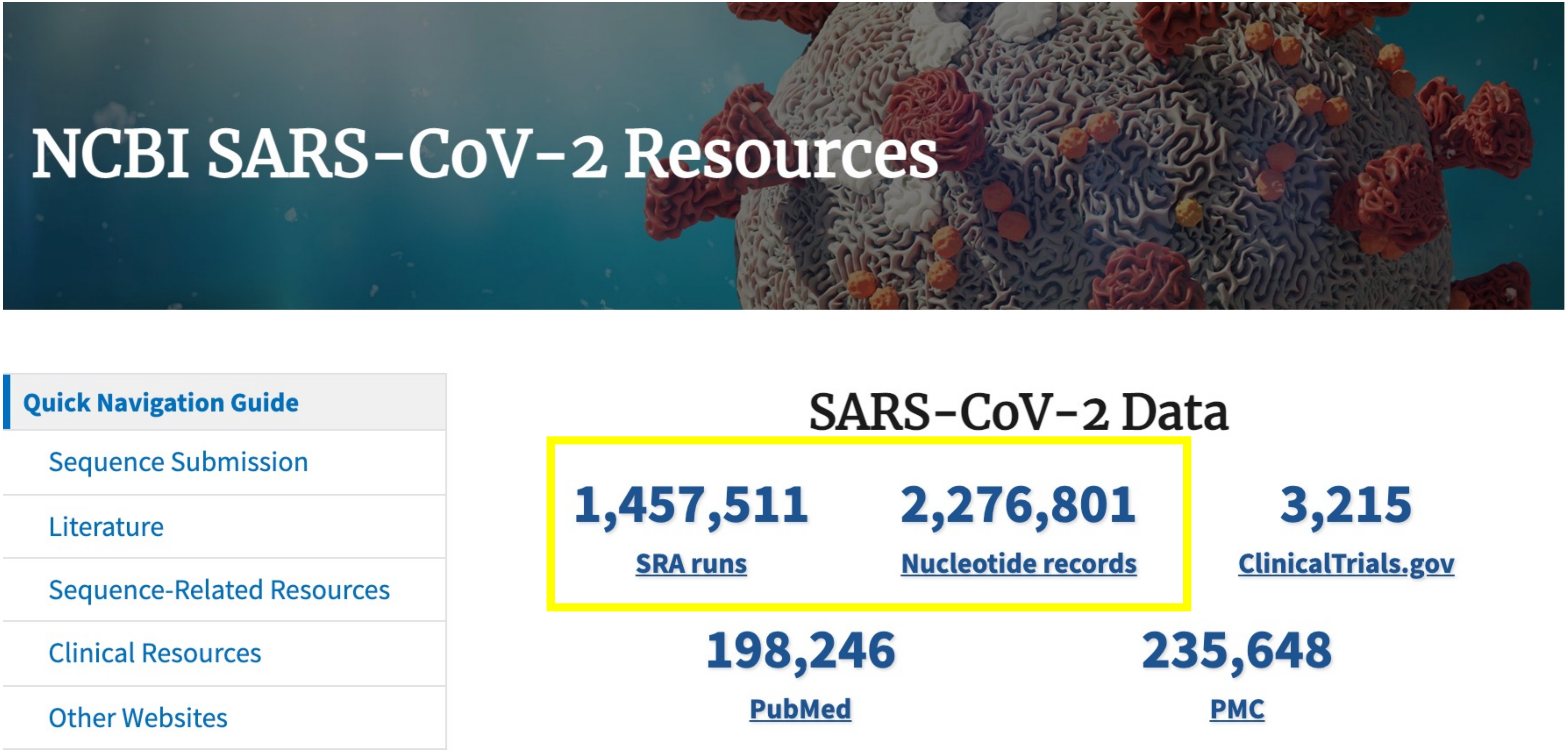
Major mutations of SARS-CoV-2 VOCs in Spike receptor binding domain (RBD)

○ Parental residues ● Mutation spots

Increase

- ACE2 affinity
- Transmissibility

https://ccforum.biomedcentral.com/articles/10.1186/s13054-021-03662-x



NCBI SARS-CoV-2 Resources

Quick Navigation Guide

- Sequence Submission
- Literature
- Sequence-Related Resources
- Clinical Resources
- Other Websites

SARS-CoV-2 Data

1,457,511
SRA runs

2,276,801
Nucleotide records

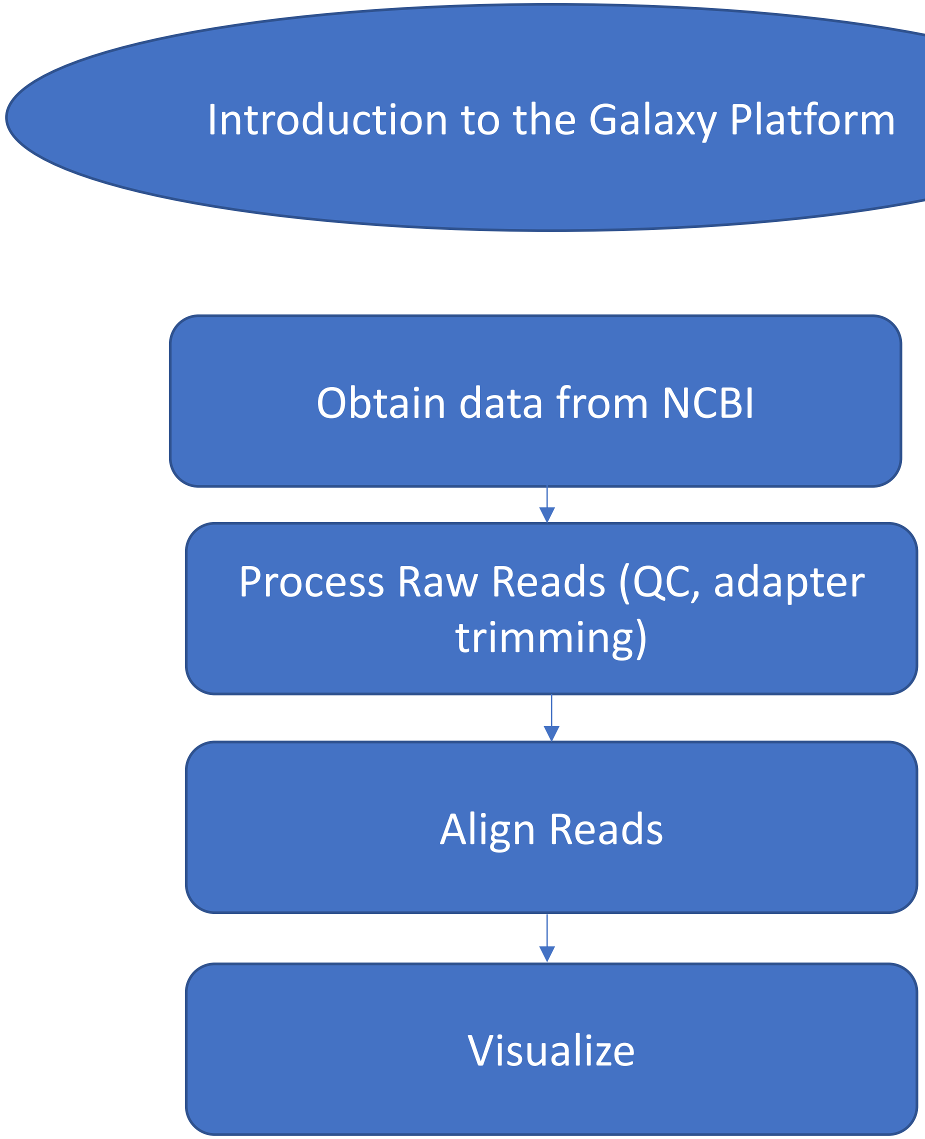
3,215
ClinicalTrials.gov

198,246
PubMed

235,648
PMC

https://www.ncbi.nlm.nih.gov/sars-cov-2/

Outline



Introduction to the Galaxy Platform

Obtain data from NCBI

Process Raw Reads (QC, adapter trimming)

Align Reads

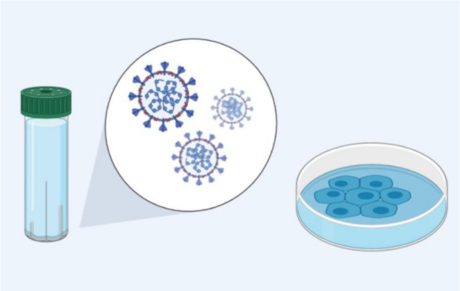
Visualize

SARS-Cov-2 Alpha variant reference sequence
SARS-Cov-2 Delta variant NGS sample

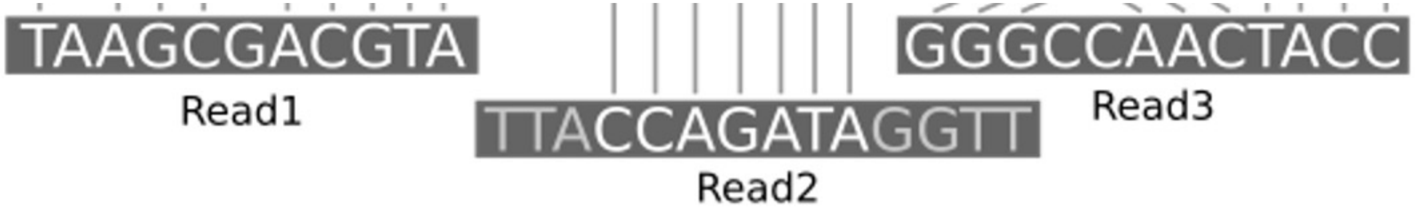
Verify delta variant mutations relative to ancestral sequence

Viral Genome Next Generation Sequencing (NGS)

- Specimen Collected

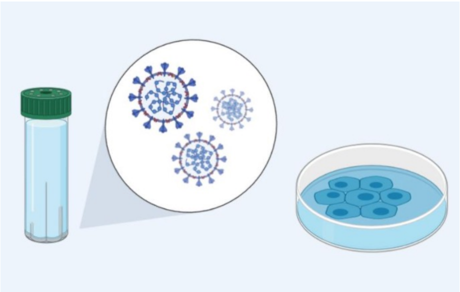


- NGS reads

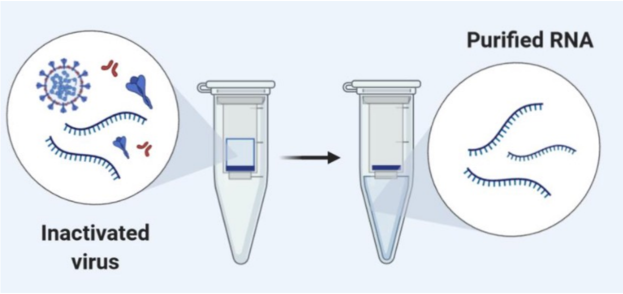


Viral Genome Next Generation Sequencing (NGS)

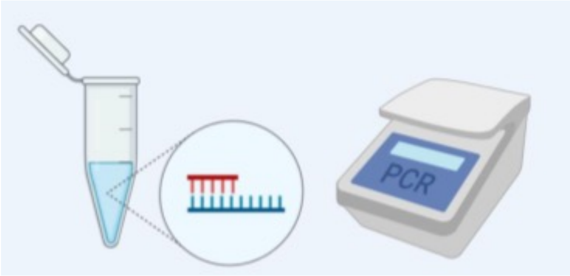
- Specimen Collected



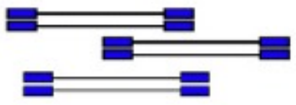
- RNA extraction



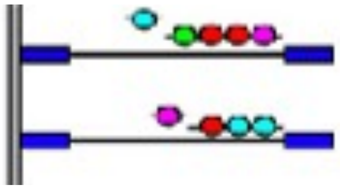
- cDNA synthesis (using virus-specific primers)
- Amplification



- NGS library prep

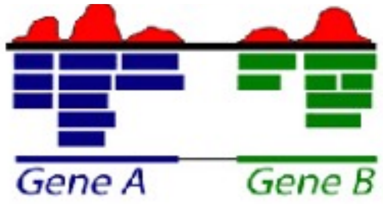


- NGS sequencing

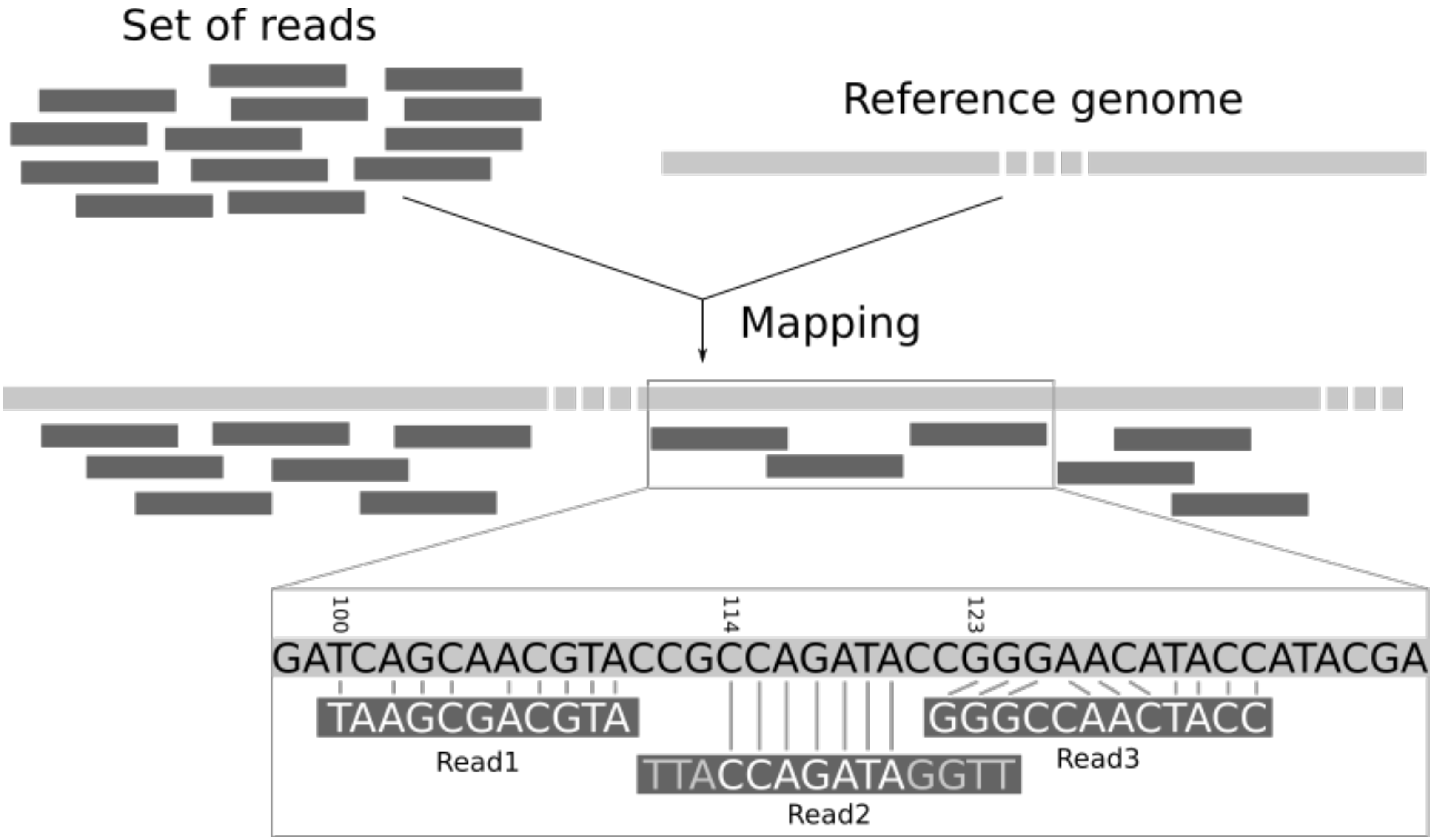


flowcell

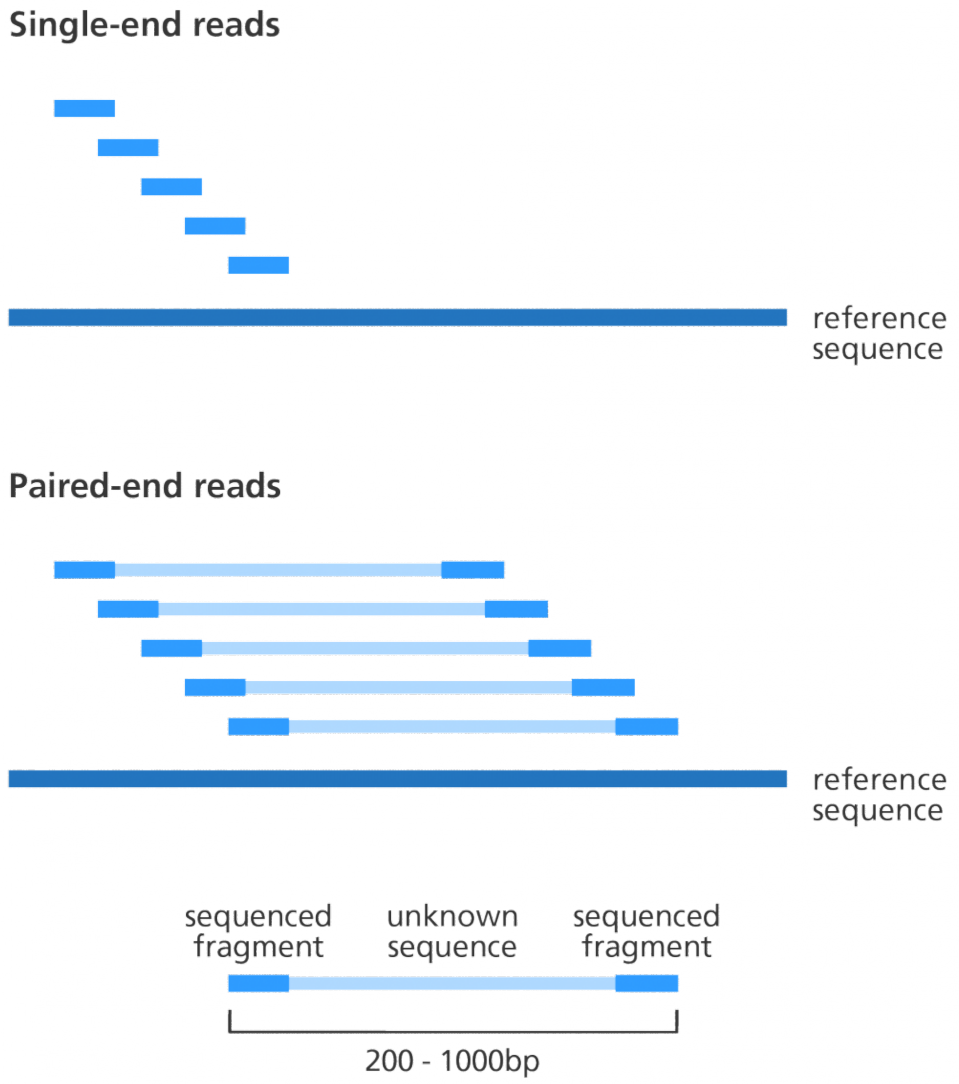
- Alignment



Short Read Alignment



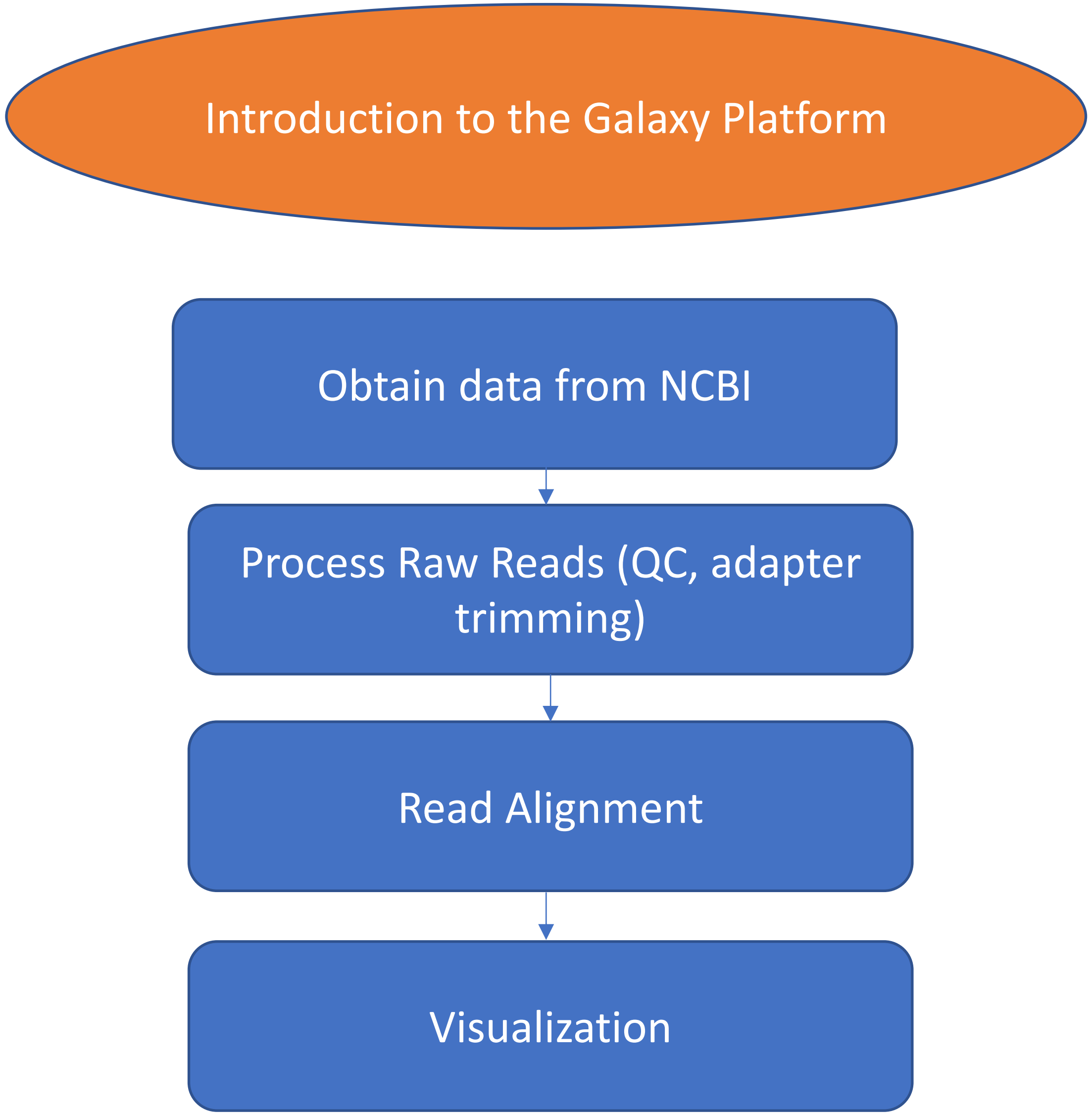
Paired end vs Single end reads



cDNA fragment

https://www.biostars.org/p/267167/

Outline



Introduction to the Galaxy Platform

Obtain data from NCBI

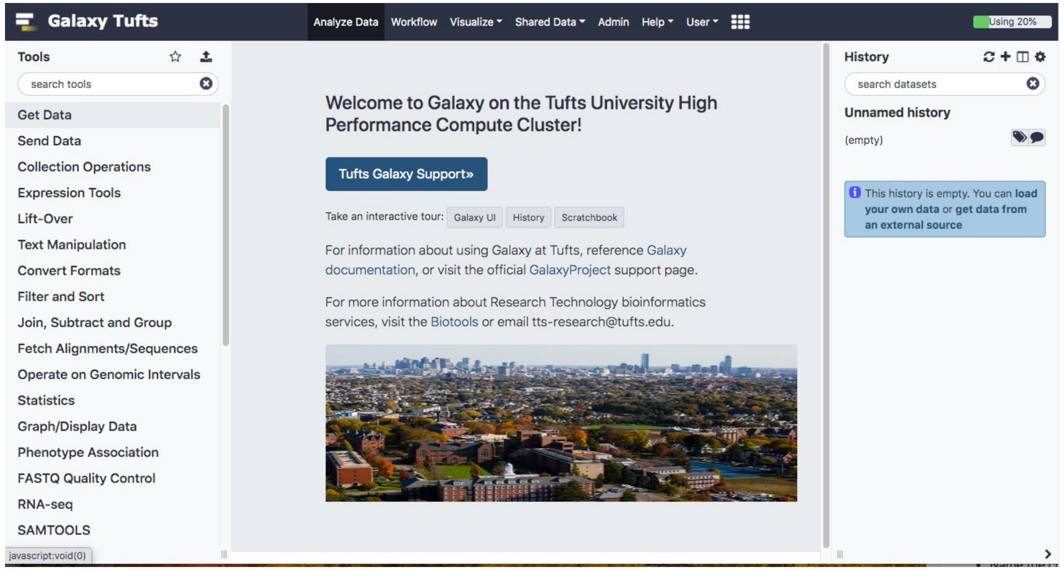
Process Raw Reads (QC, adapter trimming)

Read Alignment

Visualization

Log into Galaxy

- Connect to Tufts Network, either on campus or via VPN
- Visit https://galaxy.cluster.tufts.edu
- Log in with you Tufts credentials



Welcome to Galaxy on the Tufts University High Performance Compute Cluster!

Tufts Galaxy Support»

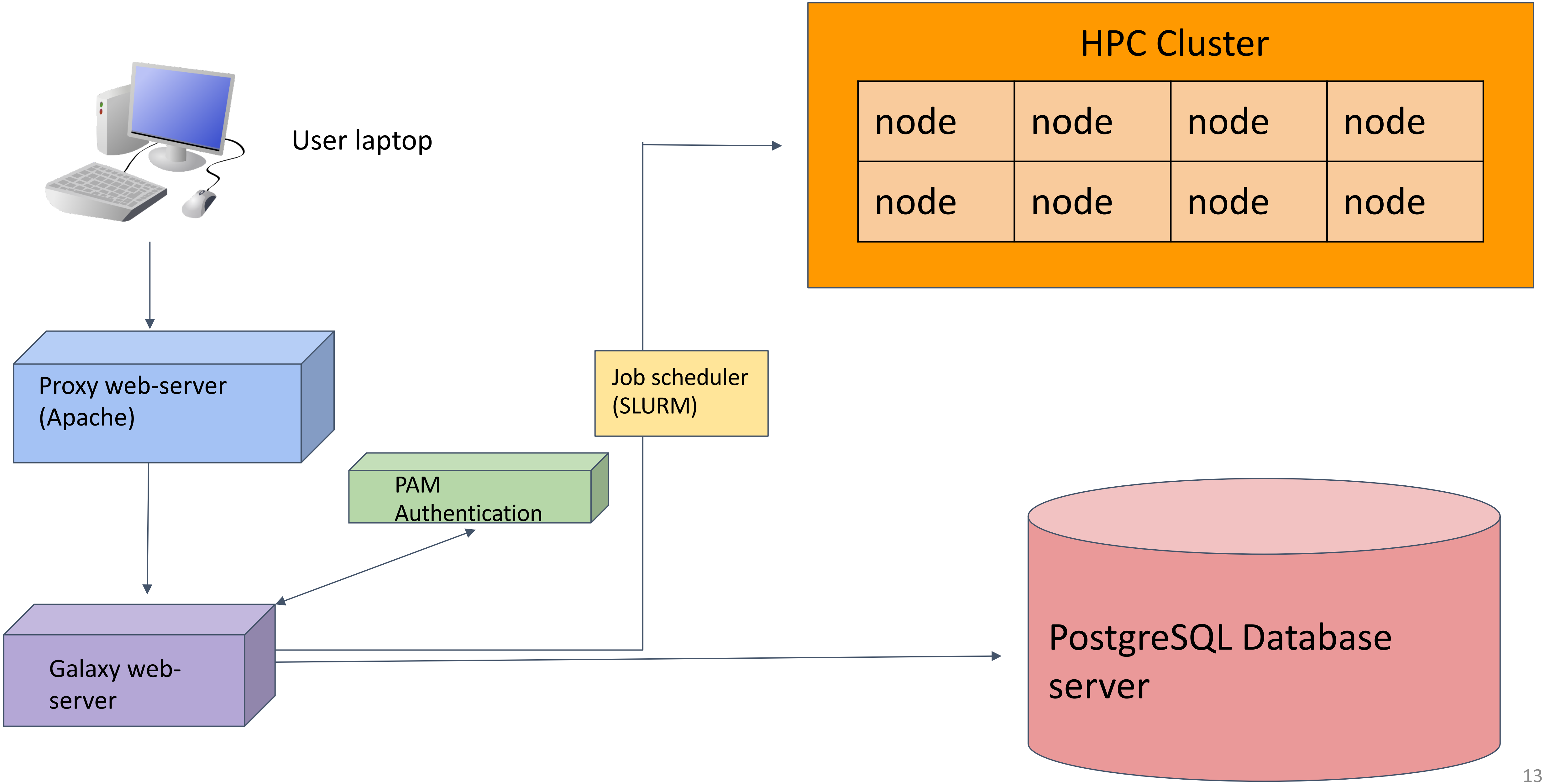
Take an interactive tour: Galaxy UI History Scratchbook

For information about using Galaxy at Tufts, reference Galaxy documentation, or visit the official GalaxyProject support page.

For more information about Research Technology bioinformatics services, visit the Biotools or email tts-research@tufts.edu.

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

Galaxy on the Tufts High Performance Compute (HPC) Cluster



User laptop

HPC Cluster

node	node	node	node
node	node	node	node

Proxy web-server (Apache)

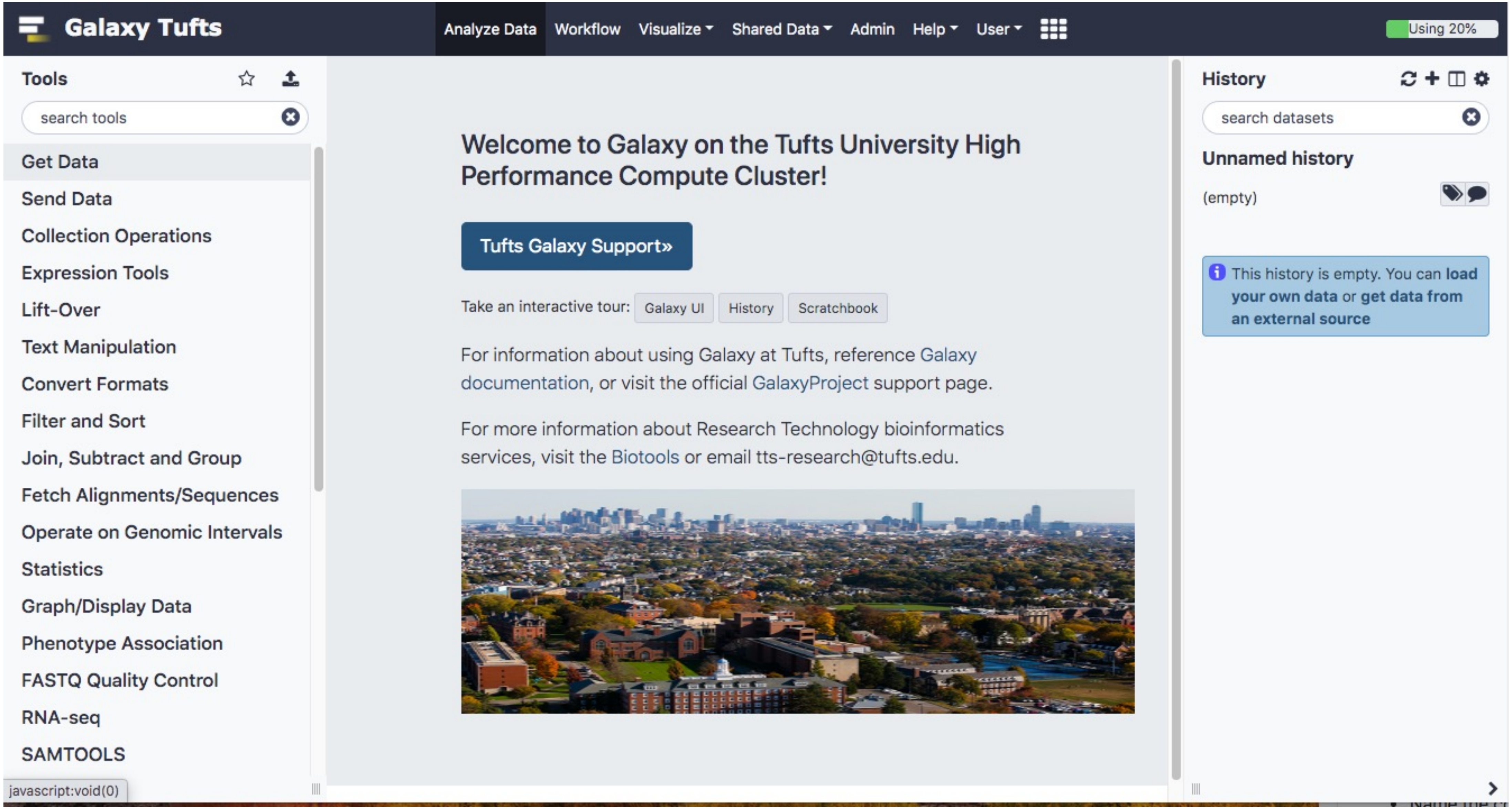
Job scheduler (SLURM)

PAM Authentication

Galaxy web-server

PostgreSQL Database server

User Interface



Welcome to Galaxy on the Tufts University High Performance Compute Cluster!

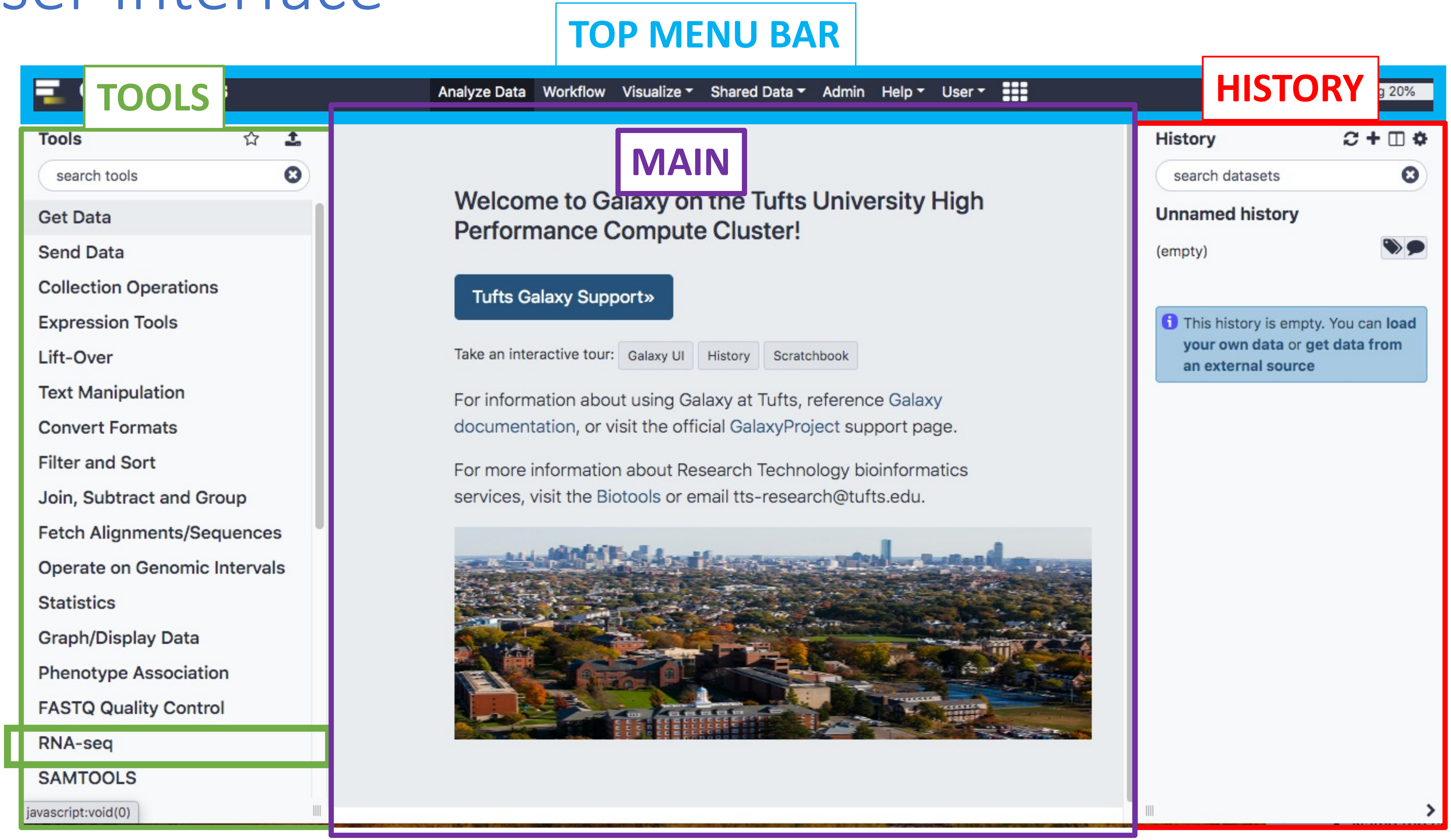
Take an interactive tour: Galaxy UI History Scratchbook

For information about using Galaxy at Tufts, reference Galaxy documentation, or visit the official GalaxyProject support page.

For more information about Research Technology bioinformatics services, visit the Biotools or email tts-research@tufts.edu.

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

User Interface



Tools

search tools

- Get Data
- Send Data
- Collection Operations
- Expression Tools
- Lift-Over
- Text Manipulation
- Convert Formats
- Filter and Sort
- Join, Subtract and Group
- Fetch Alignments/Sequences
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Phenotype Association
- FASTQ Quality Control
- RNA-seq
- SAMTOOLS

javascript:void(0)

Welcome to Galaxy on the Tufts University High Performance Compute Cluster!

Tufts Galaxy Support»

Take an interactive tour: Galaxy UI | History | Scratchbook

For information about using Galaxy at Tufts, reference Galaxy documentation, or visit the official GalaxyProject support page.

For more information about Research Technology bioinformatics services, visit the Biotools or email tts-research@tufts.edu.

History

search datasets

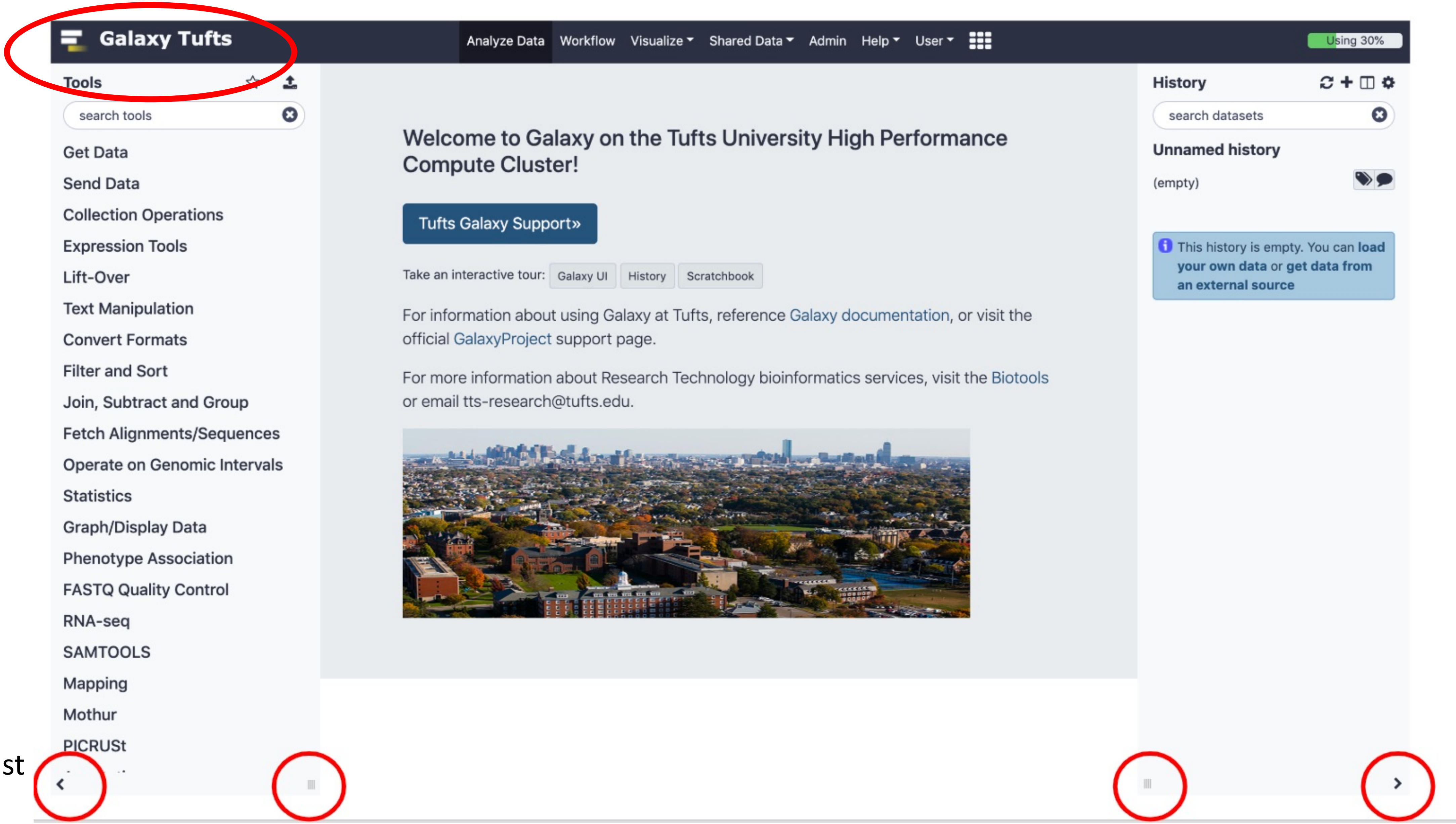
Unnamed history

(empty)

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

Galaxy User Interface

To return to home screen



Welcome to Galaxy on the Tufts University High Performance Compute Cluster!

Tufts Galaxy Support»

Take an interactive tour: Galaxy UI History Scratchbook

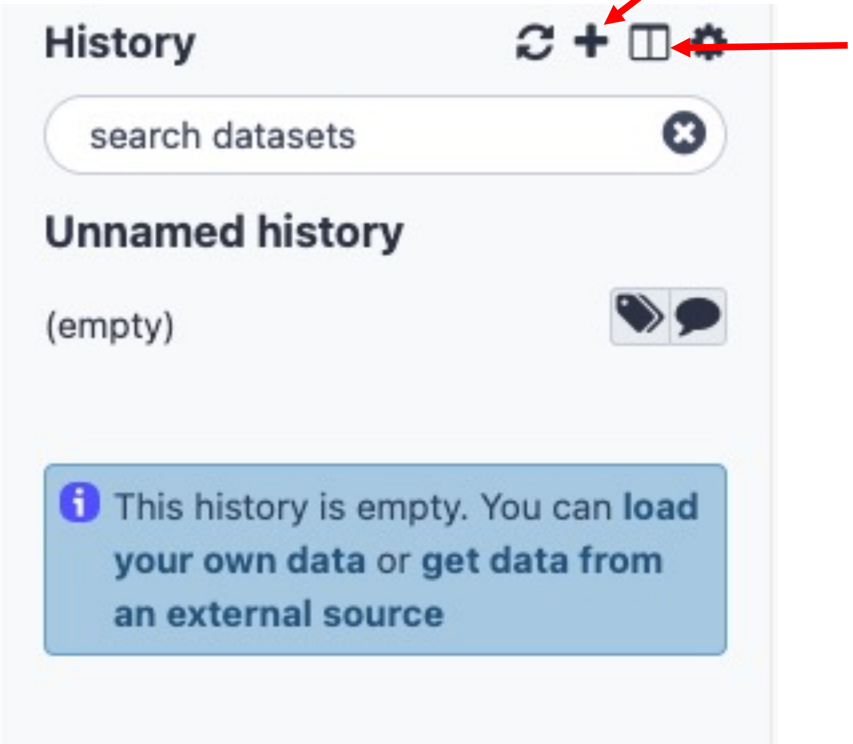
For information about using Galaxy at Tufts, reference Galaxy documentation, or visit the official GalaxyProject support page.

For more information about Research Technology bioinformatics services, visit the Biotools or email tts-research@tufts.edu.

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

Minimize/Adjust toolbars

History



History

search datasets

Unnamed history

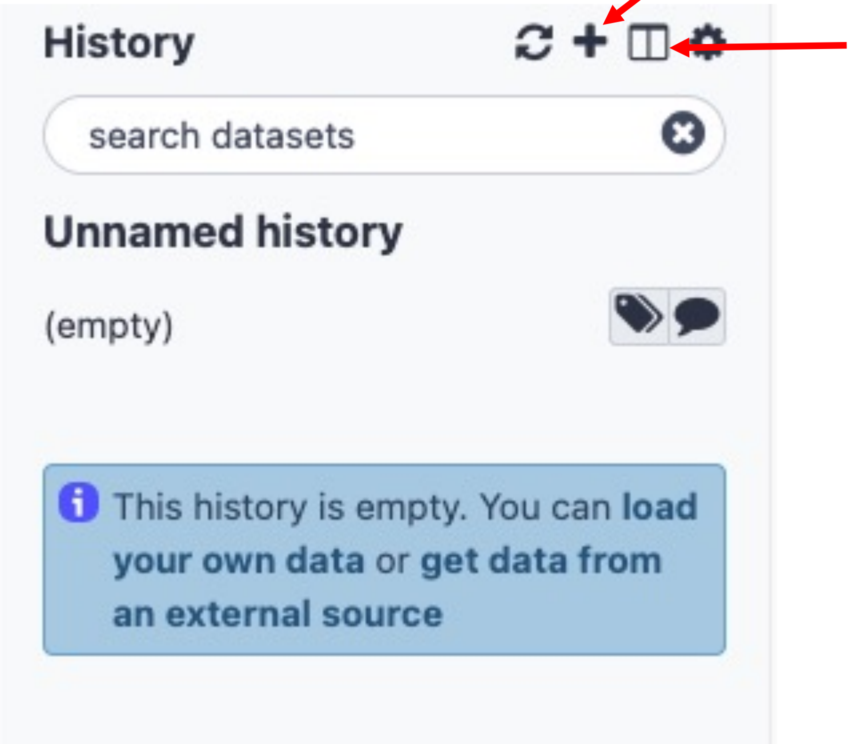
(empty)

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

Create New History

View all Histories

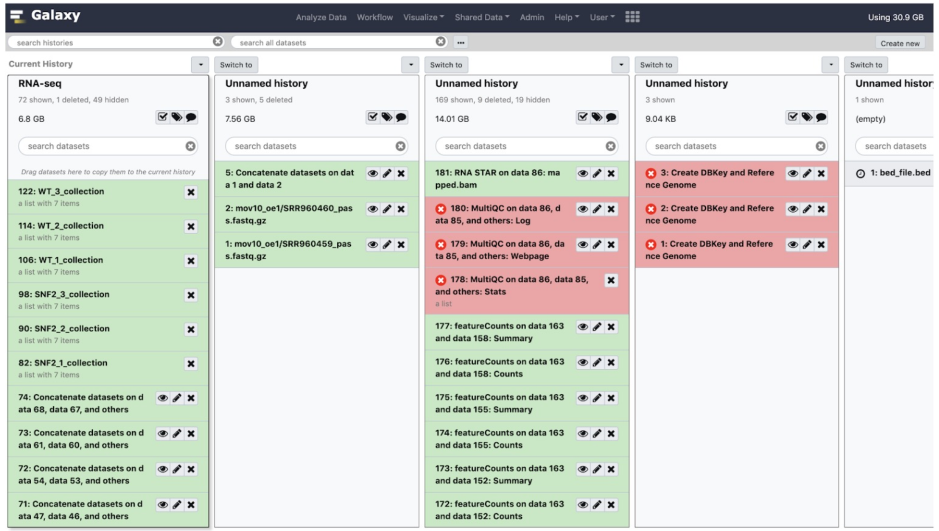
History



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

Create New History

View all Histories



RNA-seq

72 shown, 1 deleted, 49 hidden

6.8 GB

Drag datasets here to copy them to the current history

- 122: WT_3_collection
  a list with 7 items
- 114: WT_2_collection
  a list with 7 items
- 106: WT_1_collection
  a list with 7 items
- 98: SNF2_3_collection
  a list with 7 items
- 90: SNF2_2_collection
  a list with 7 items
- 82: SNF2_1_collection
  a list with 7 items
- 74: Concatenate datasets on data 68, data 67, and others
- 73: Concatenate datasets on data 61, data 60, and others
- 72: Concatenate datasets on data 54, data 53, and others
- 71: Concatenate datasets on data 47, data 46, and others

Unnamed history

3 shown, 5 deleted

7.56 GB

- 5: Concatenate datasets on data 1 and data 2
- 2: mov10_oe1/SRR960460_pass.fastq.gz
- 1: mov10_oe1/SRR960459_pass.fastq.gz

Unnamed history

169 shown, 9 deleted, 19 hidden

14.01 GB

- 181: RNA STAR on data 86: mapped.bam
- 180: MultiQC on data 86, data 85, and others: Log
- 179: MultiQC on data 86, data 85, and others: Webpage
- 178: MultiQC on data 86, data 85, and others: Stats
  a list
- 177: featureCounts on data 163 and data 158: Summary
- 176: featureCounts on data 163 and data 158: Counts
- 175: featureCounts on data 163 and data 155: Summary
- 174: featureCounts on data 163 and data 155: Counts
- 173: featureCounts on data 163 and data 152: Summary
- 172: featureCounts on data 163 and data 152: Counts

Unnamed history

3 shown

9.04 KB

- 3: Create DBKey and Reference Genome
- 2: Create DBKey and Reference Genome
- 1: Create DBKey and Reference Genome

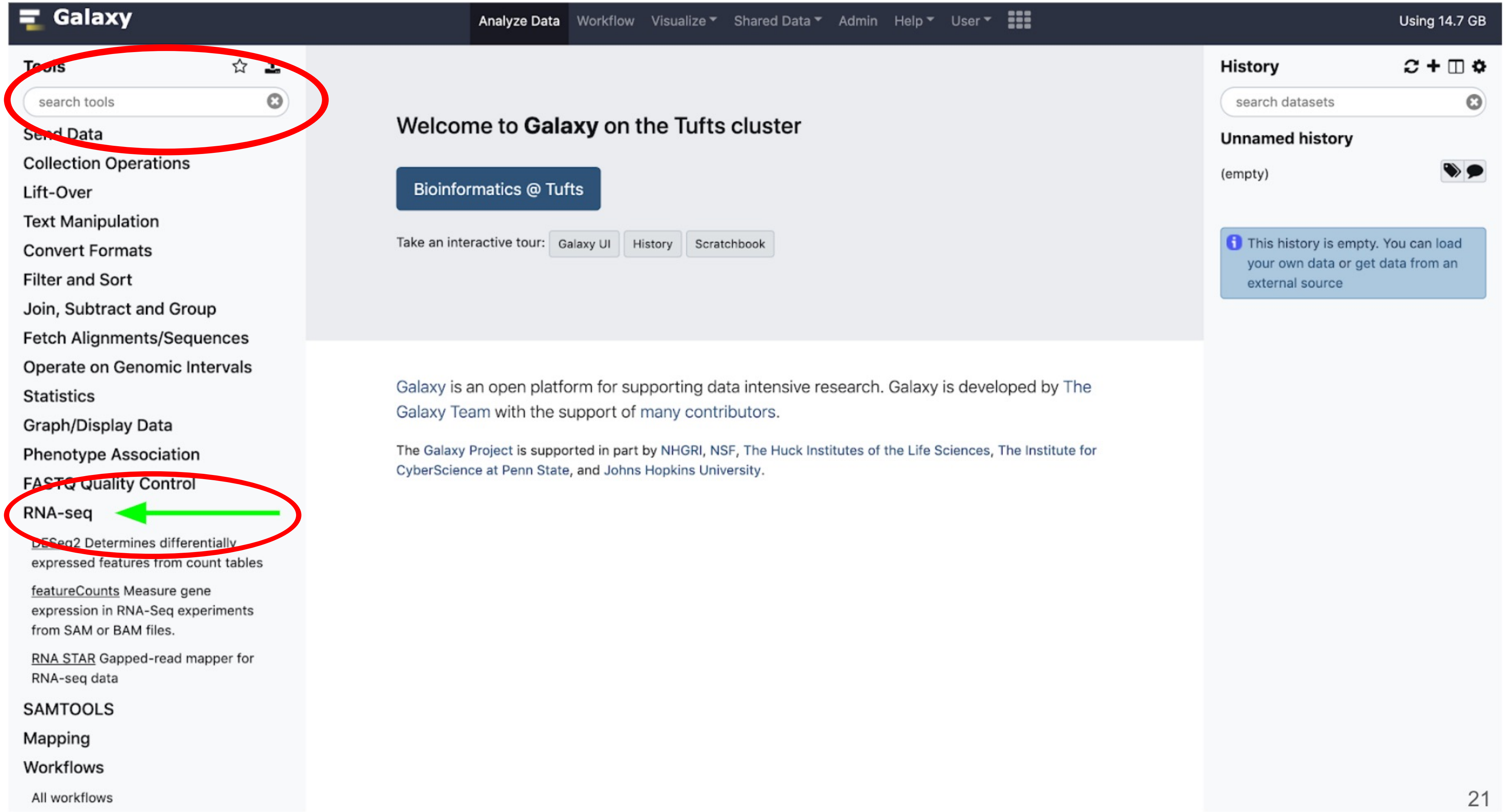
Unnamed histor

1 shown

(empty)

- 1: bed_file.bed

Tools



Send Data
Collection Operations
Lift-Over
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Alignments/Sequences
Operate on Genomic Intervals
Statistics
Graph/Display Data
Phenotype Association
FASTQ Quality Control
RNA-seq

DESeq2 Determines differentially expressed features from count tables

featureCounts Measure gene expression in RNA-Seq experiments from SAM or BAM files.

RNA STAR Gapped-read mapper for RNA-seq data

SAMTOOLS
Mapping
Workflows
All workflows

Welcome to **Galaxy** on the Tufts cluster

Bioinformatics @ Tufts

Take an interactive tour: Galaxy UI | History | Scratchbook

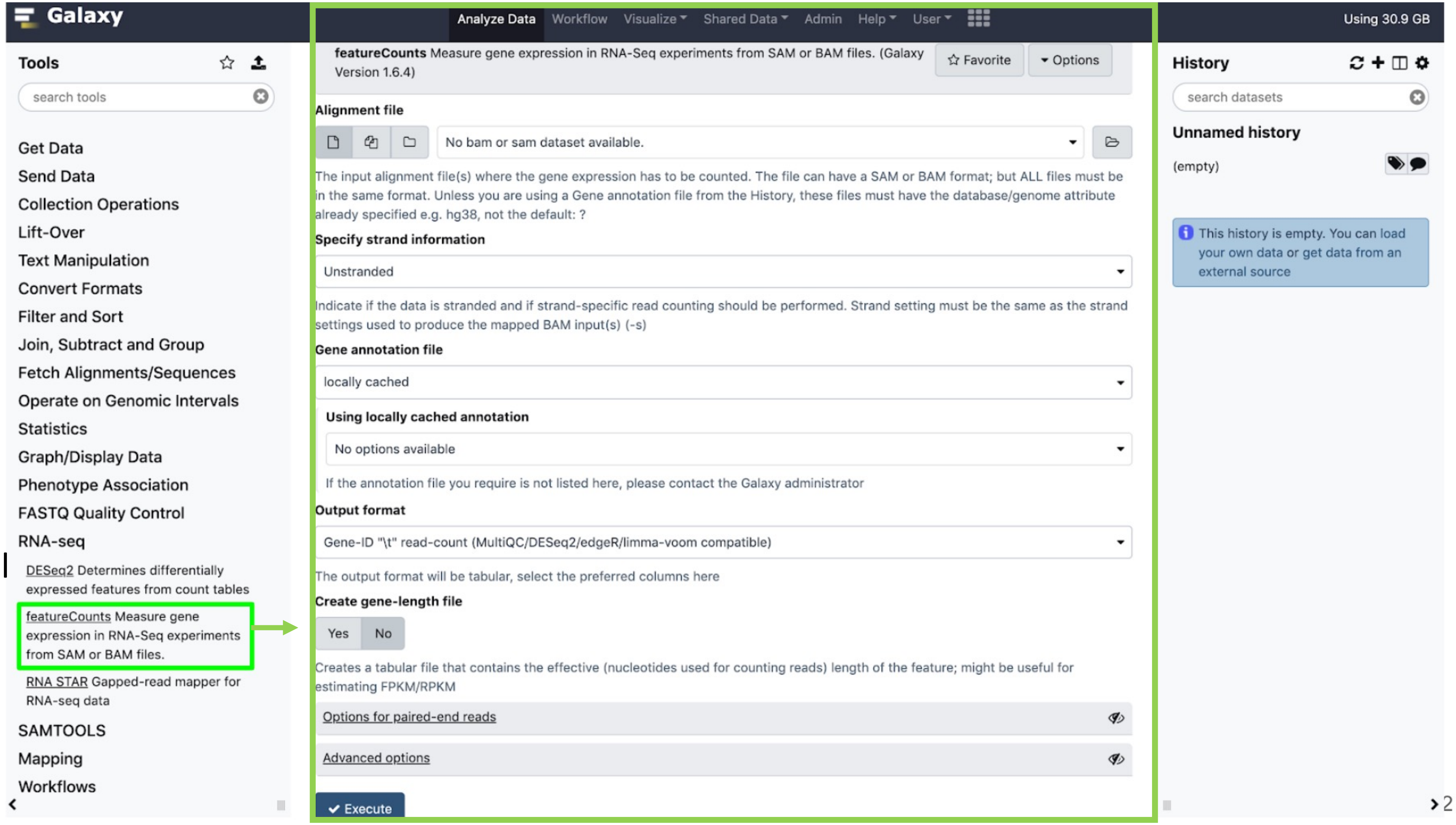
Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors.

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 Johns Hopkins University.

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

Tools

Click on the name of the tool to open it in the main panel



featureCounts Measure gene expression in RNA-Seq experiments from SAM or BAM files. (Galaxy Version 1.6.4)

Alignment file

No bam or sam dataset available.

The input alignment file(s) where the gene expression has to be counted. The file can have a SAM or BAM format; but ALL files must be in the same format. Unless you are using a Gene annotation file from the History, these files must have the database/genome attribute already specified e.g. hg38, not the default: ?

Specify strand information

Unstranded

Indicate if the data is stranded and if strand-specific read counting should be performed. Strand setting must be the same as the strand settings used to produce the mapped BAM input(s) (-s)

Gene annotation file

locally cached

Using locally cached annotation

No options available

If the annotation file you require is not listed here, please contact the Galaxy administrator

Output format

Gene-ID "\t" read-count (MultiQC/DESeq2/edgeR/limma-voom compatible)

The output format will be tabular, select the preferred columns here

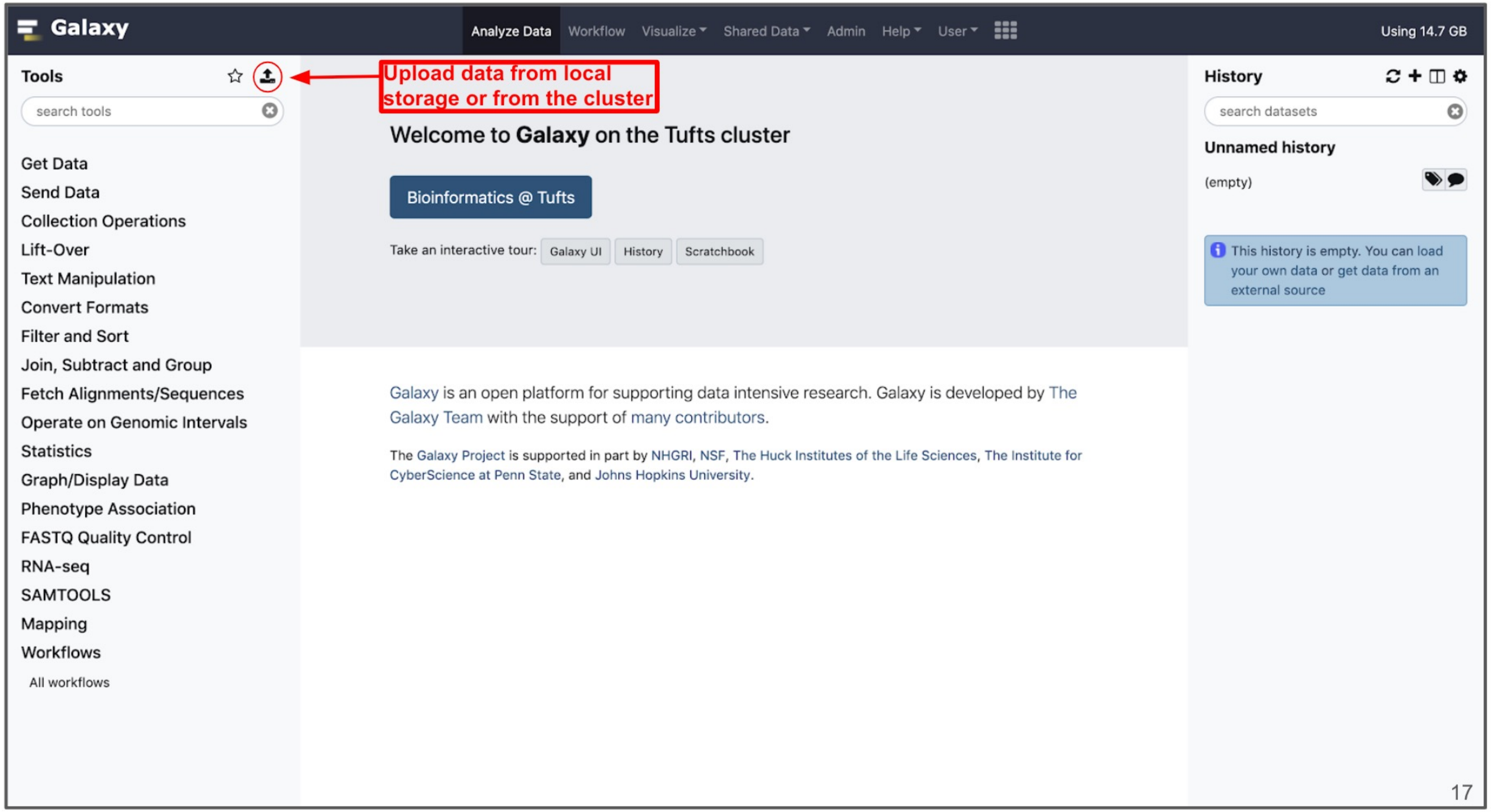
Create gene-length file

Yes No

Creates a tabular file that contains the effective (nucleotides used for counting reads) length of the feature; might be useful for estimating FPKM/RPKM

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

Importing data



Upload data from local storage or from the cluster

Welcome to **Galaxy** on the Tufts cluster

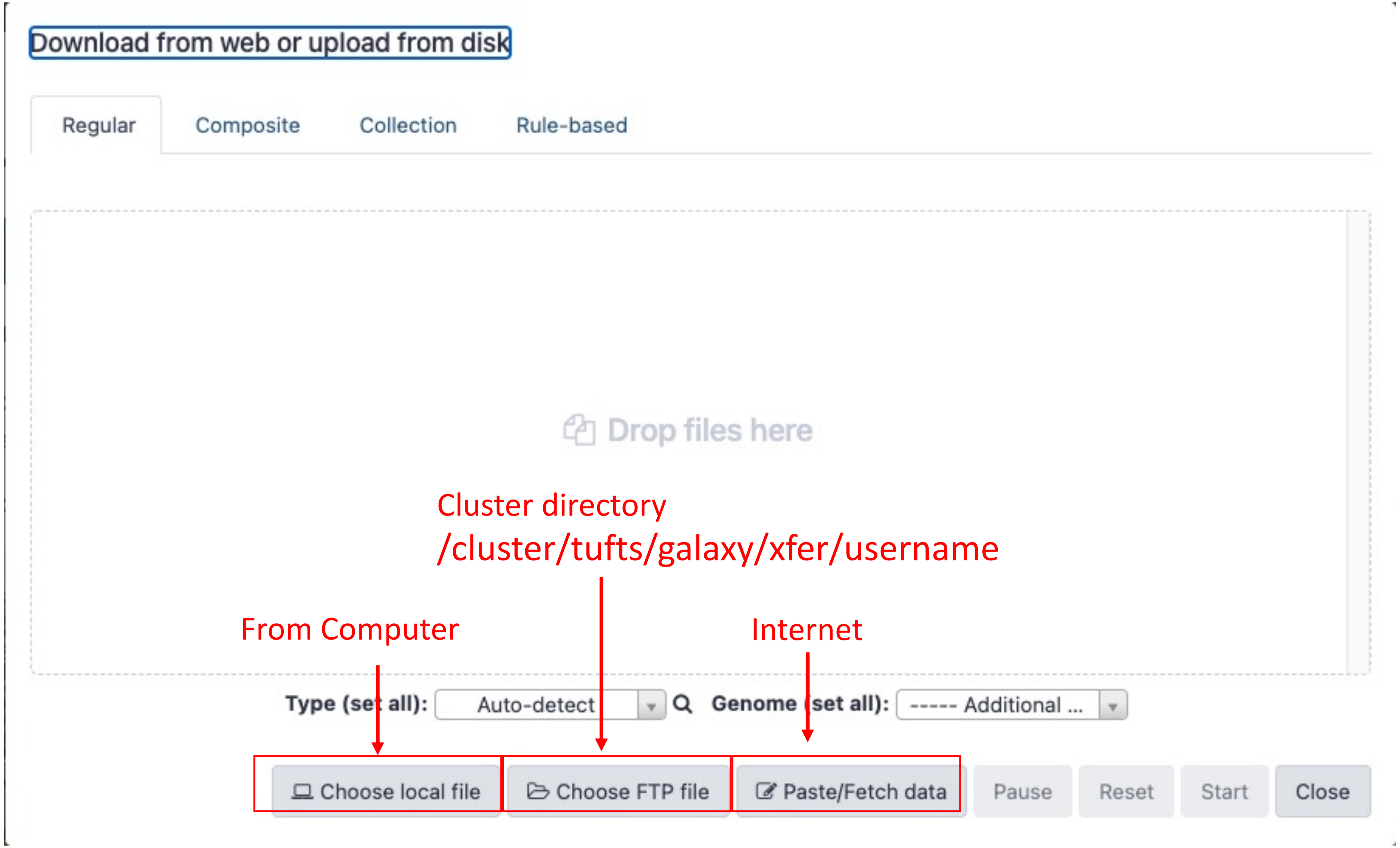
Take an interactive tour: Galaxy UI History Scratchbook

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

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors.

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 Johns Hopkins University.

Importing data



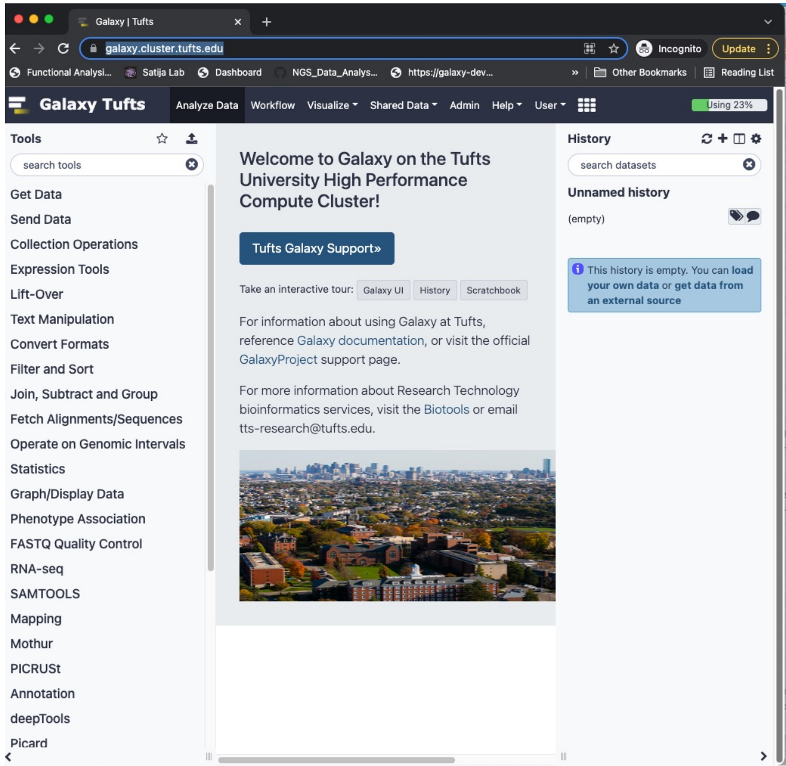
From Computer

Cluster directory
/cluster/tufts/galaxy/xfer/username

Internet

Log into Galaxy and open course website

- Connect to Tufts Network, either on campus or via VPN
- Visit https://galaxy.cluster.tufts.edu
- Log in with you Tufts credentials
- Visit course website https://rbatorsky.github.io/intro-to-galaxy-ngs-sarscov2



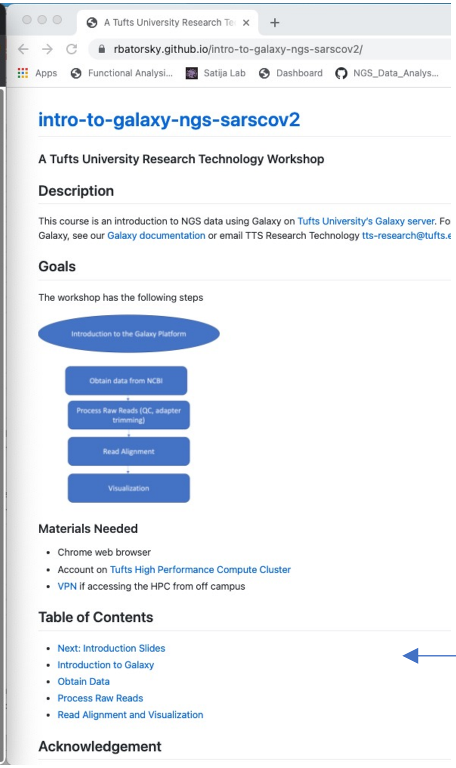
Welcome to Galaxy on the Tufts University High Performance Compute Cluster!

Take an interactive tour: Galaxy UI History Scratchbook

For information about using Galaxy at Tufts, reference Galaxy documentation, or visit the official GalaxyProject support page.

For more information about Research Technology bioinformatics services, visit the Biotools or email tts-research@tufts.edu.

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


A Tufts University Research Te
rbatorsky.github.io/intro-to-galaxy-ngs-sarscov2/
Apps
Functional Analysi...
Satija Lab
Dashboard
NGS_Data_Analys...
Introduction to the Galaxy Platform
Obtain data from NCBI
Process Raw Reads (QC, adapter trimming)
Read Alignment
Visualization

intro-to-galaxy-ngs-sarscov2

A Tufts University Research Technology Workshop

Description

This course is an introduction to NGS data using Galaxy on Tufts University's Galaxy server. Fo
Galaxy, see our Galaxy documentation or email TTS Research Technology tts-research@tufts.e

Goals

The workshop has the following steps

Materials Needed

- Chrome web browser
- Account on Tufts High Performance Compute Cluster
- VPN if accessing the HPC from off campus

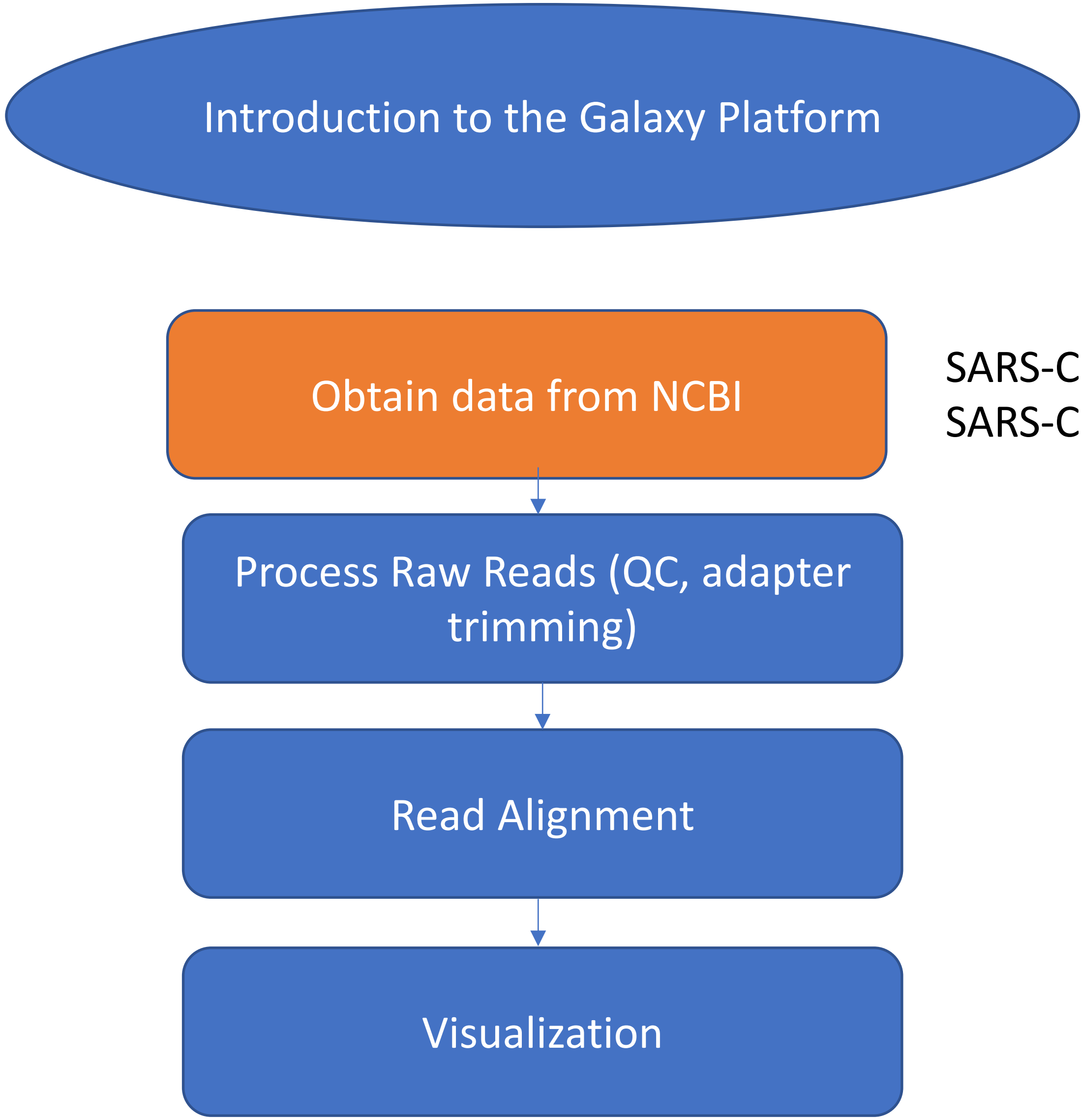
Table of Contents

- Next: Introduction Slides
- Introduction to Galaxy
- Obtain Data
- Process Raw Reads
- Read Alignment and Visualization

Acknowledgement

Navigate to **Obtain_Data** section

Outline



Introduction to the Galaxy Platform

Obtain data from NCBI

Process Raw Reads (QC, adapter trimming)

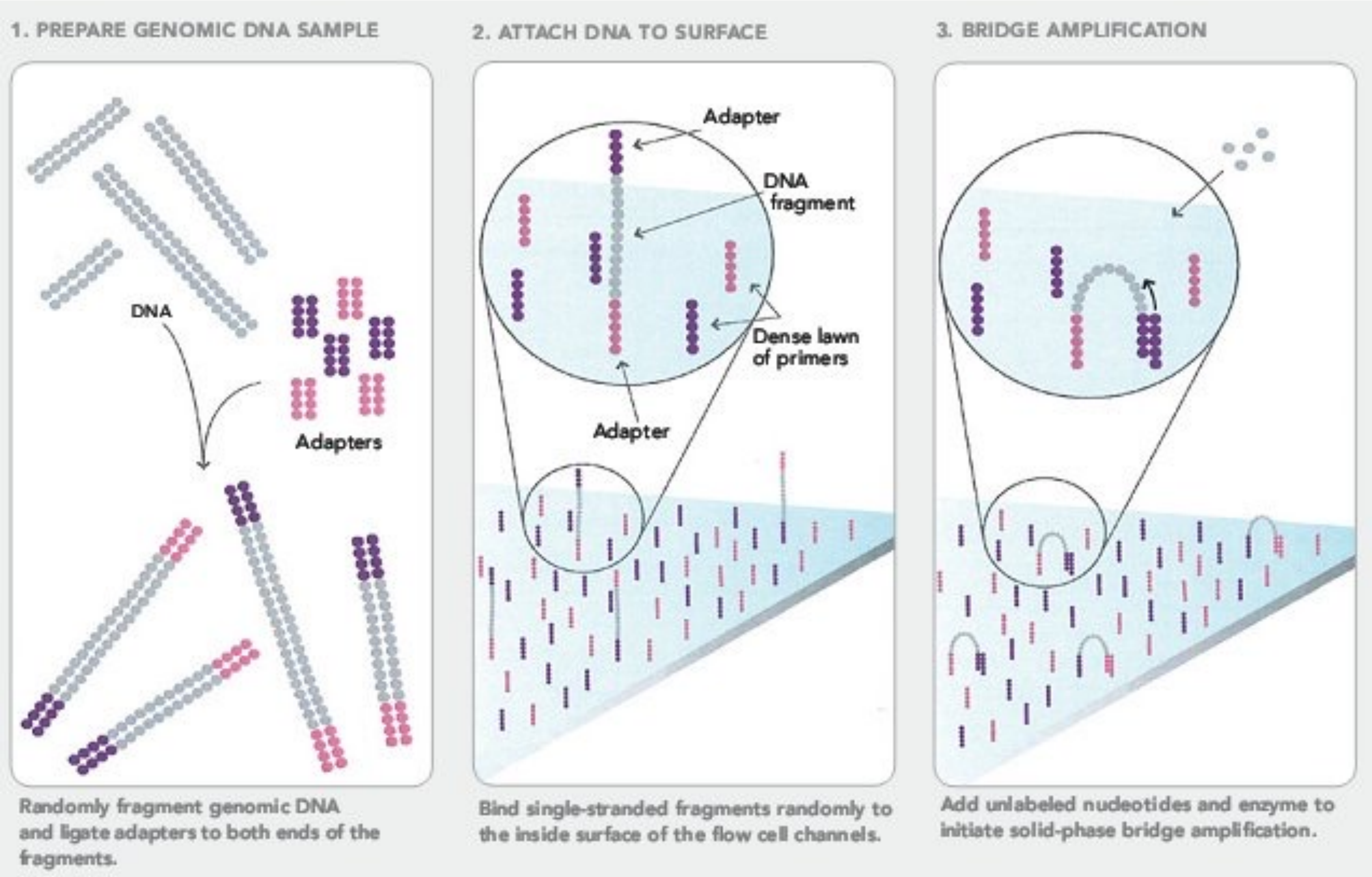
Read Alignment

Visualization

SARS-Cov-2 Alpha variant reference sequence
SARS-Cov-2 Delta variant NGS sample

NGS details

Next Generation Sequencing (NGS)



1. PREPARE GENOMIC DNA SAMPLE

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE

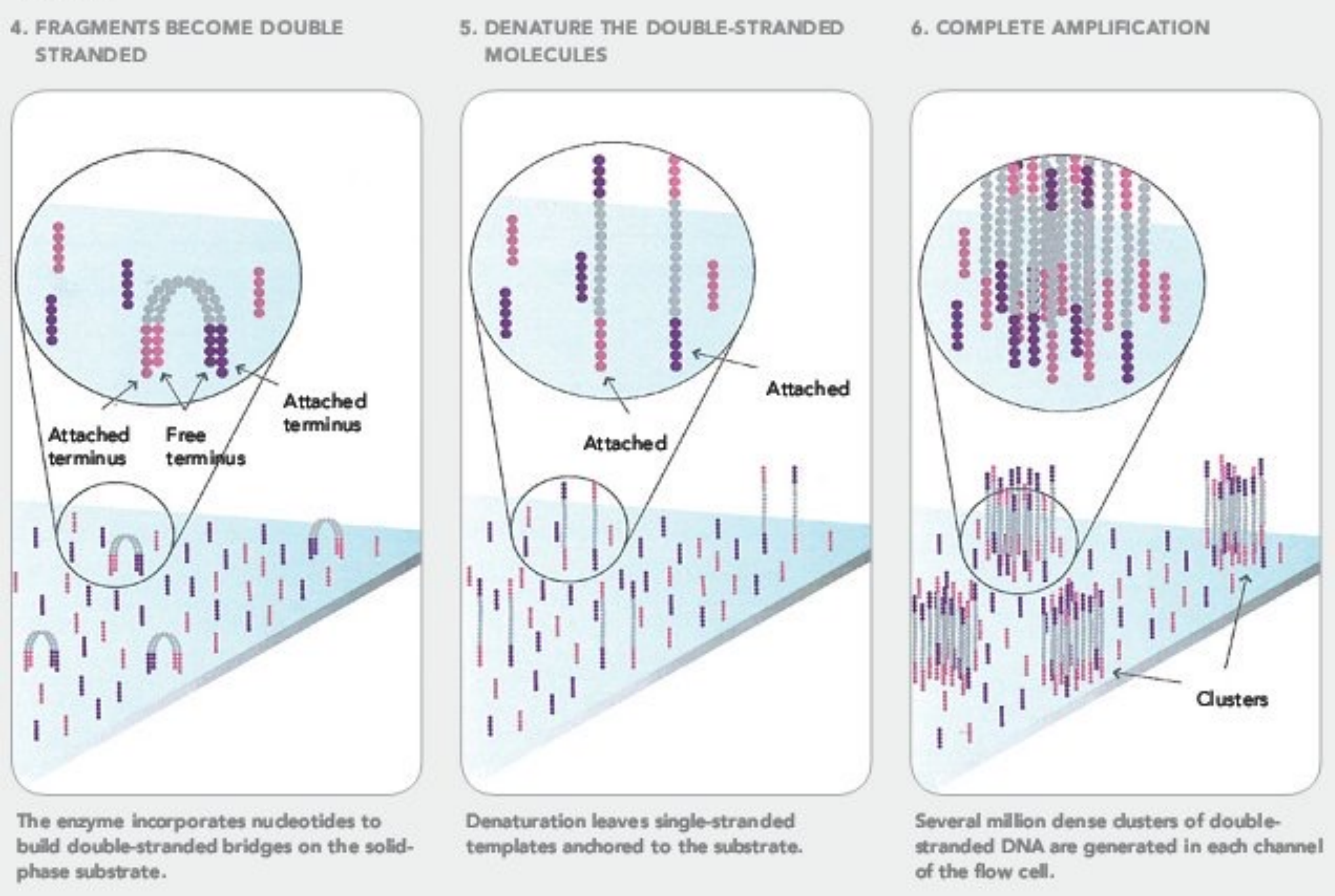
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

https://sites.google.com/site/himbcorelab/illumina_sequencing

Next Generation Sequencing (NGS)



4. FRAGMENTS BECOME DOUBLE STRANDED

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES

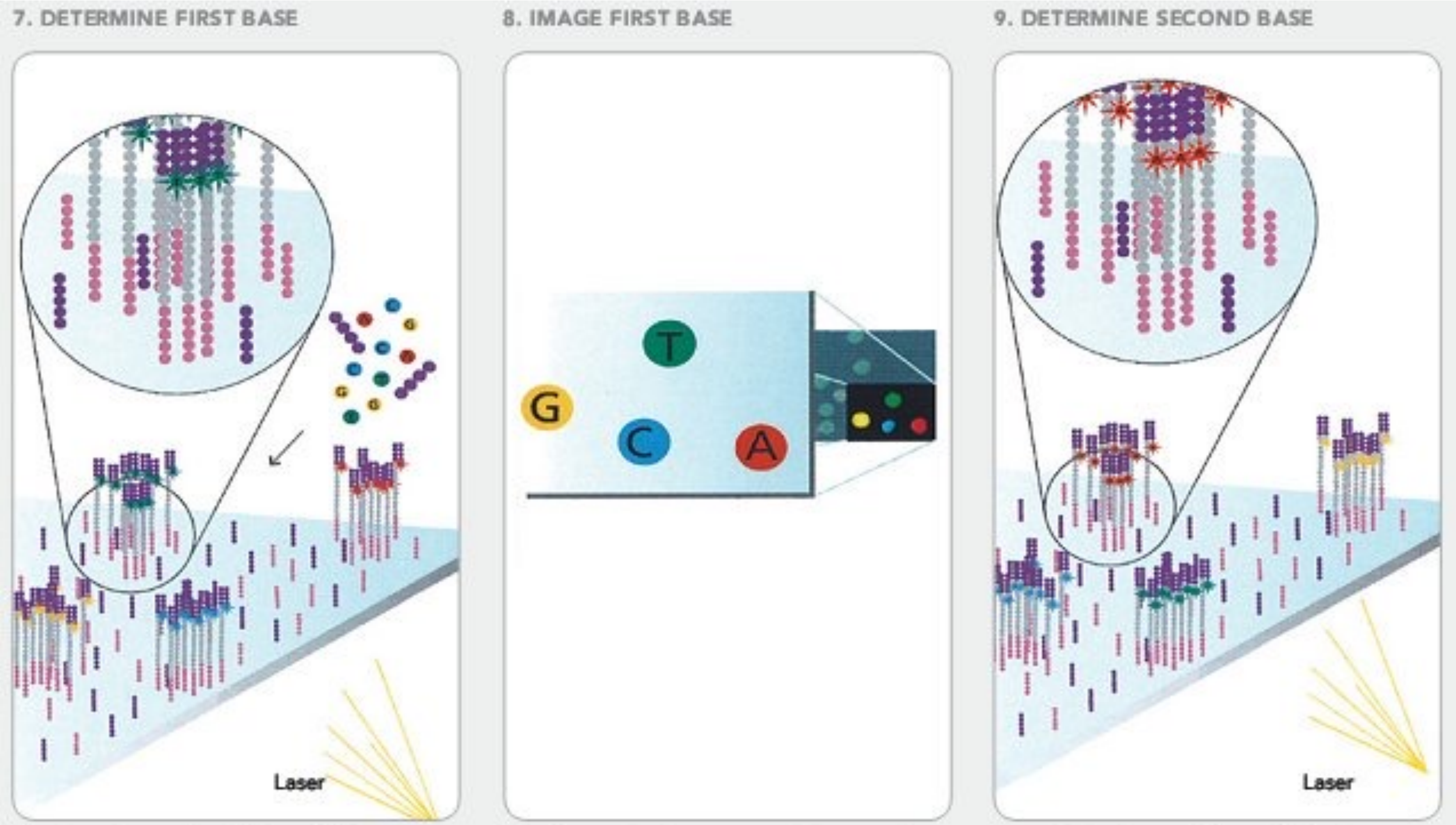
Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

https://sites.google.com/site/himbcorelab/illumina_sequencing

Next Generation Sequencing (NGS)



7. DETERMINE FIRST BASE

First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE

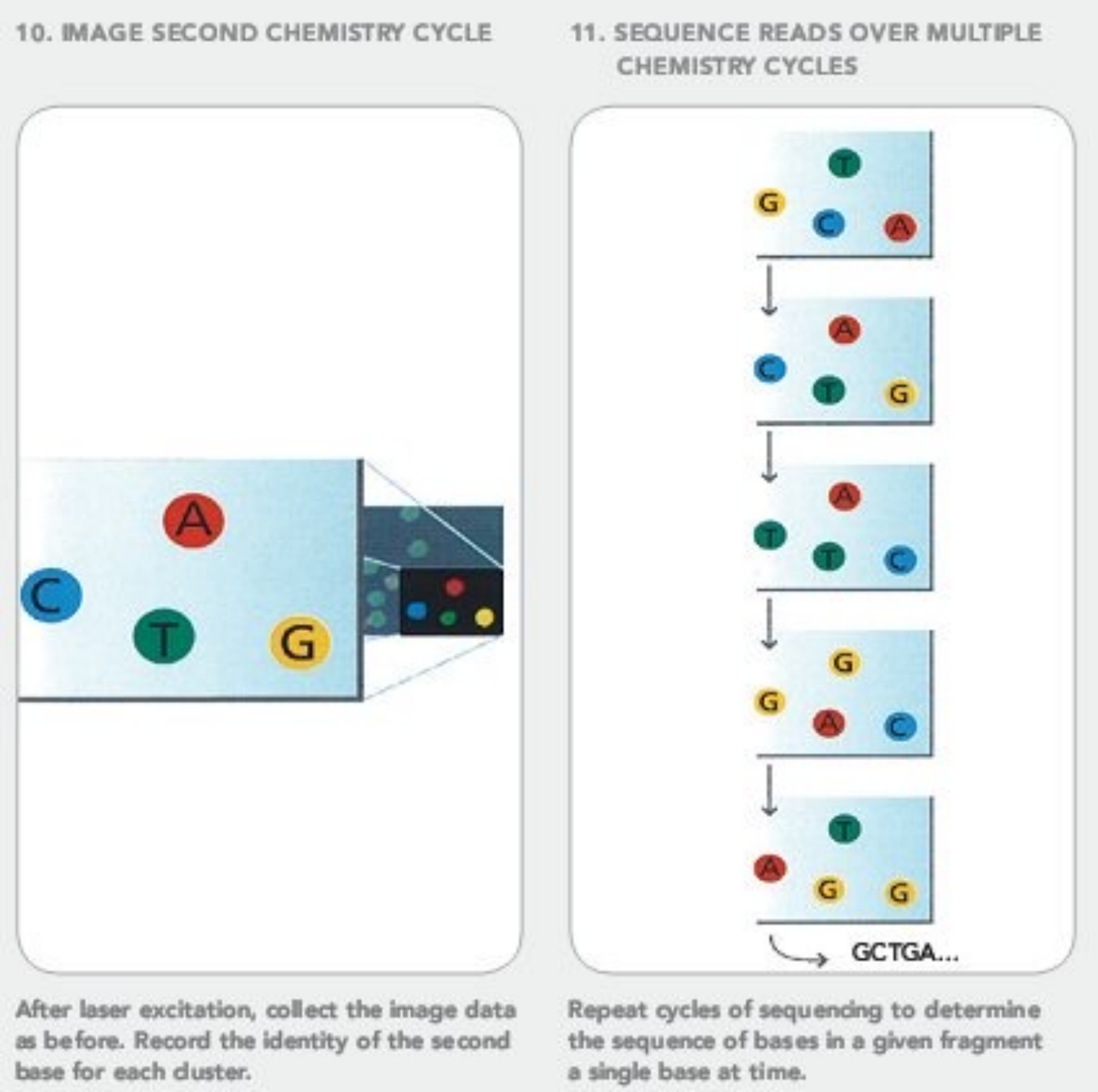
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

https://sites.google.com/site/himbcorelab/illumina_sequencing

Next Generation Sequencing (NGS)



10. IMAGE SECOND CHEMISTRY CYCLE

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES

After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

This Illumina video is great for visualization!

https://sites.google.com/site/himbcorelab/illumina_sequencing