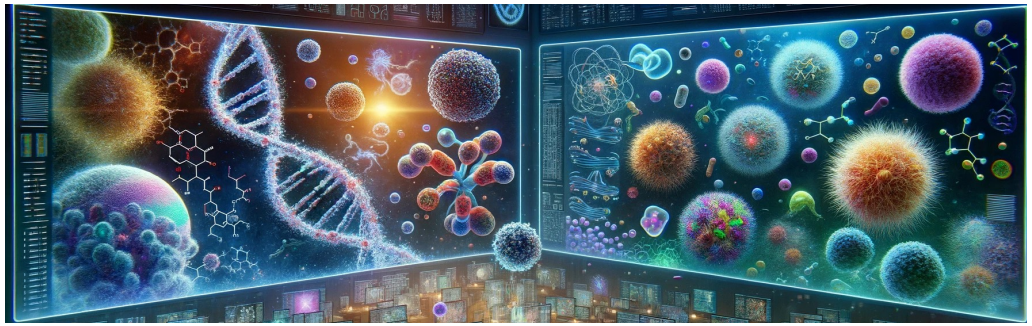
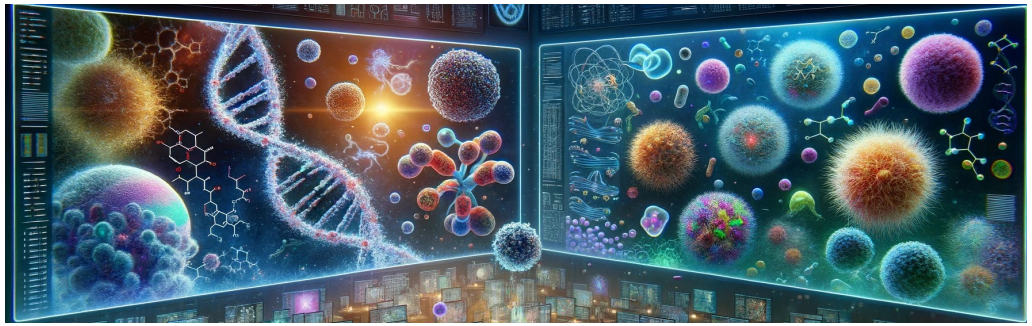


Advanced Applications of Next Generation Sequencing in Food Safety

www.ucd.ie/cfs
@CfsUcd

Dr Guerrino Macori, BSc, MSc, PhD
Assistant Professor
School of Biology and Environmental Science
University College Dublin, Ireland
UCD Centre for Food Safety
guerrino.macori@ucd.ie
@guerrinomacori



Advanced Applications of Next Generation Sequencing in Food Safety
Scientific Programme

DAY 1 **Monday 22nd January 2024**

09:00 Stellenbosch welcome UCD to laboratory facilities
12:30 Lunch

SESSION 1: INTRODUCTION TO NEXT GENERATION SEQUENCING (NGS)
technologies, applications, and an introduction to bioinformatics

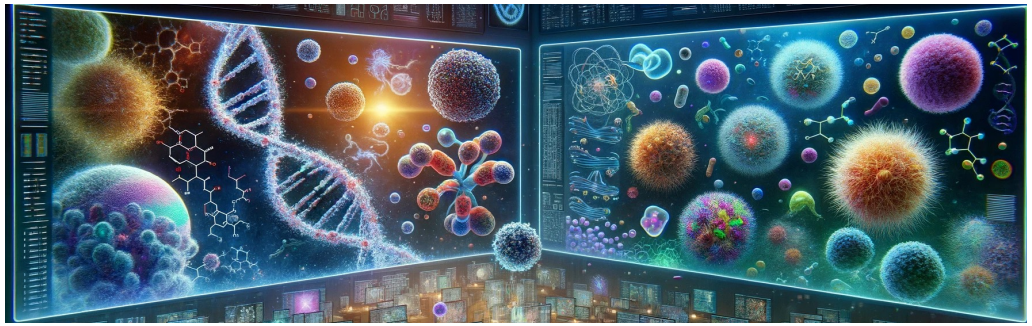
13:30 Professor Pieter A. Gouws, Stellenbosch University, South Africa
- Welcome, opening remarks and introduction to the program

14:30 Assistant Professor Guerrino Macori, University College Dublin, Ireland
- Overview of Next Generation Sequencing Technologies -what do we need to know

15:00 Professor Séamus Fanning, University College Dublin, Ireland
- Next Generation Sequencing Technologies in the context of risk assessment & food safety

15.30 Assistant Professor Guerrino Macori, University College Dublin, Ireland
- Introduction to Bioinformatics and its applications

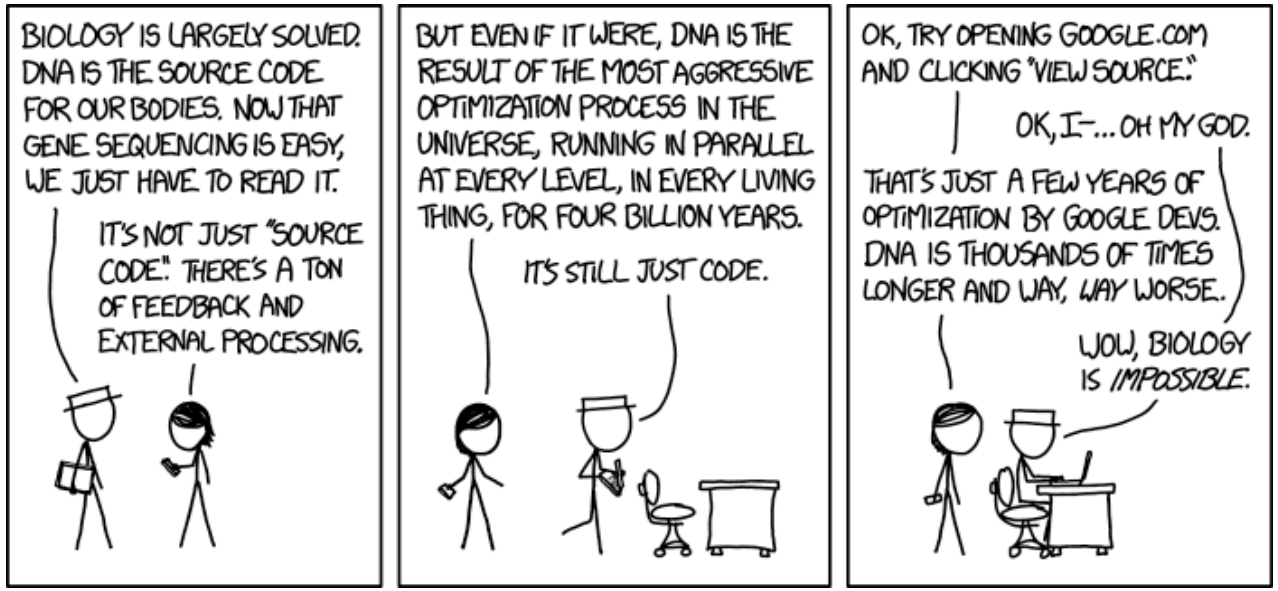
16:30 Industry talk



Data analysis to enable characterisation of *Listeria monocytogenes* & *Salmonella*

WORKSHOP A: Whole genome sequencing (WGS) library preparation & loading the Nanopore flongle

- Recap – application bioinformatics in food safety
- Introduction to Galaxy
- Quality assessment raw reads
- Assembly
- Quality assessment assemblies



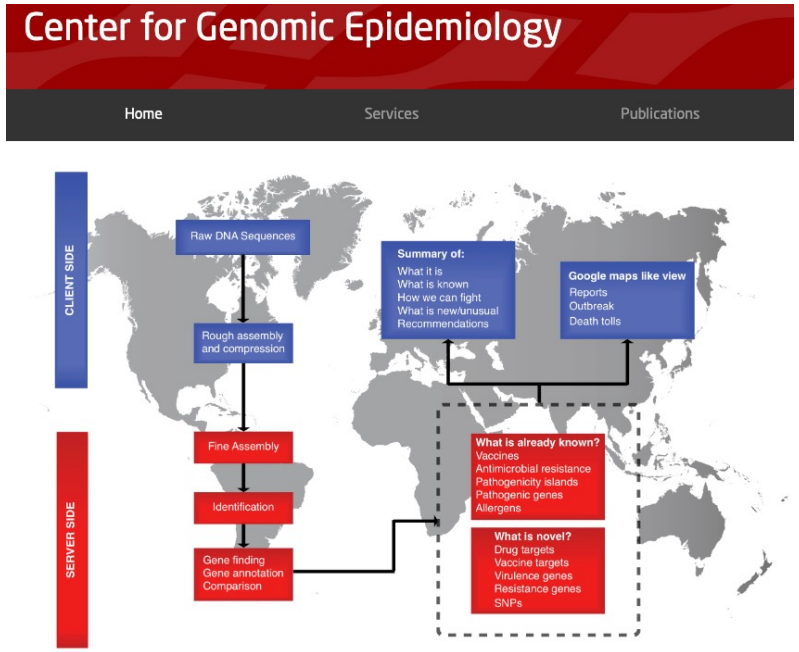
Options bioinformatics analysis for Food Safety

- Command line
- Stand alone web-resources (for example http://www.genomicepidemiology.org/services/)
- Online pipelines (for example https://galaxyproject.org/eu/)

Live exercise
Using command line/terminal

```
spades -1 read1.fastq -2 read2.fastq --careful --only-assembler -t 48 --cov-cutoff auto -o Sample1
```

```
abricate --db resfinder CFS1073_ONT.fasta
```



Welcome to the Center for Genomic Epidemiology

The use of sequencing technologies is currently transforming almost every aspect of biological science. In relation to infectious diseases, the advances are rapidly changing our scientific discoveries, as well as diagnostic and outbreak investigations. The ability to analyze sequencing data and take advantage of the rapid progress, is however, not equally distributed between institutions and countries.

The aim of the Center for Genomic Epidemiology (CGE) is to provide access to bioinformatics resources also for those with limited experience and thereby allow all countries, institutions and individuals to take advantage of the novel sequencing technologies. Doing so, we hope to facilitate more open data sharing around the world and provide more equal opportunities for all.

CGE is entirely non-commercial and operates a number of free online bioinformatics services. Funding is provided as core funding from the Technical University of Denmark (DTU) and from a range of public and private sources as listed below.

If you want to read more about us and our research activities, please visit the Global Surveillance website.

Overview of Services

Phenotyping

ResFinder
Identfcation of acquired antibiotic resistance genes.

ResFinderFG
Identfcation of functional metagenomic antibiotic resistance determinants.

LRE-finder
Identfcation of genes and mutations leading to linezolid resistance.

KmerResistance
Identfcation of acquired antibiotic resistance genes using Kmers.

PathogenFinder
Prediction of a bacteria's pathogenicity towards human hosts.

VirulenceFinder
Identfcation of acquired virulence genes.

Restriction-ModificationFinder
Determination of Restriction-Modification sites (based on REBASE.)

SPIFinder
SPIFinder identifies Salmonella Pathogenicity Islands.

ToxFinder
ToxFinder identifies genes involved in mycotoxin synthesis.

Typing

MLST
Multi Locus Sequence Typing (MLST) from an assembled genome or from a set of reads.

PlasmidFinder
PlasmidFinder identifies plasmids in total or partial sequenced isolates of bacteria.

pMLST
Multi Locus Sequence Typing (MLST) from an assembled plasmid or

Phylogeny

MINTyper
Identification of SNPs with automatic filtering, masking and site validation together with inferred phylogeny based on both long and short sequencing data.

CSIPhylogeny
CSI Phylogeny calls SNPs, filters the SNPs, does site validation and infers a phylogeny based on the concatenated alignment of the high quality* SNPs.

NDtree
NDtree constructs phylogenetic trees from Single-End or Pair-End FASTQ files.

Evergreen
Evergreen generates a forest of constantly updated phylogenetic trees with publicly available whole-genome sequencing data from foodborne, bacterial isolates that were deposited in the short sequencing read archives (NCBI SRA/ENA).

TreeViewer
Phylogeny Tree Viewer.

Metagenomics

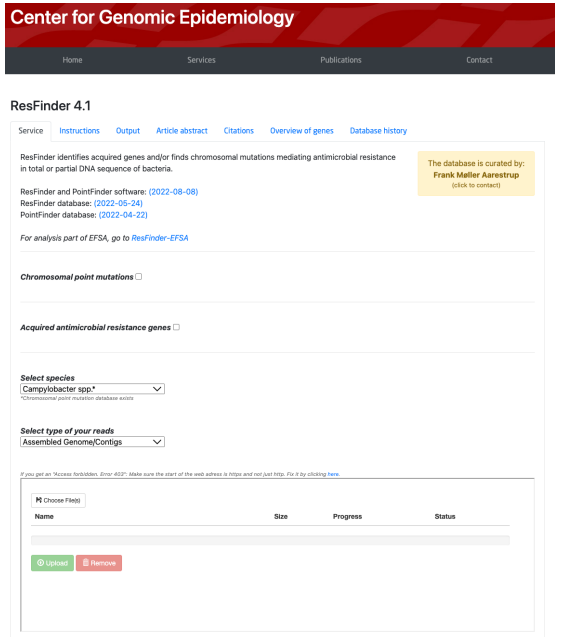
CCMetagen
CCMetagen: Comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data.

PCR-tools

RUCS
RUCS: Rapid Identification of PCR Primers Pairs for Unique Core Sequences.

Other

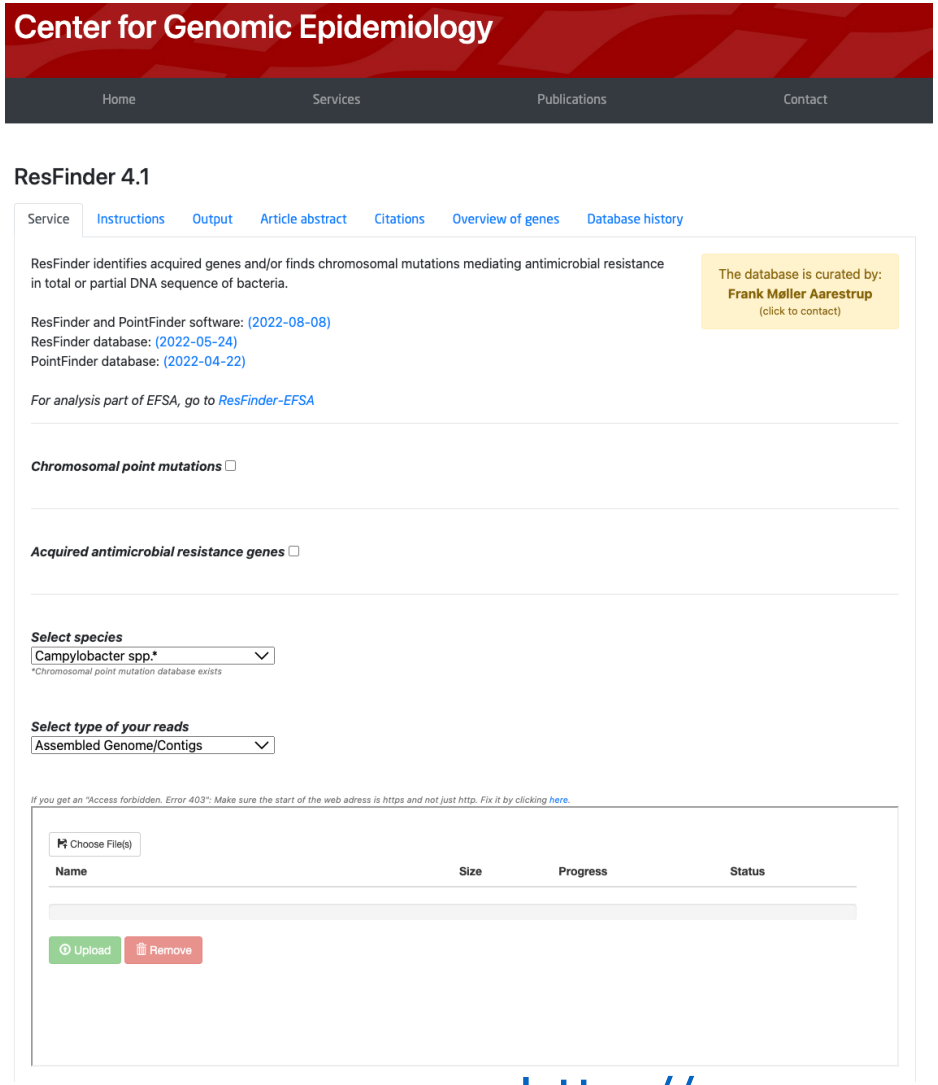
MyKMAfinder



Live exercise
web resources

http://www.genomicepidemiology.org/services/

recap



Try different services using the sequence (fasta file) included in the lecture material!

File name: assembly.fasta

Live exercise
web resources
http://www.genomicepidemiology.org/services/

Galaxy Europe

The homepage of the European Galaxy community

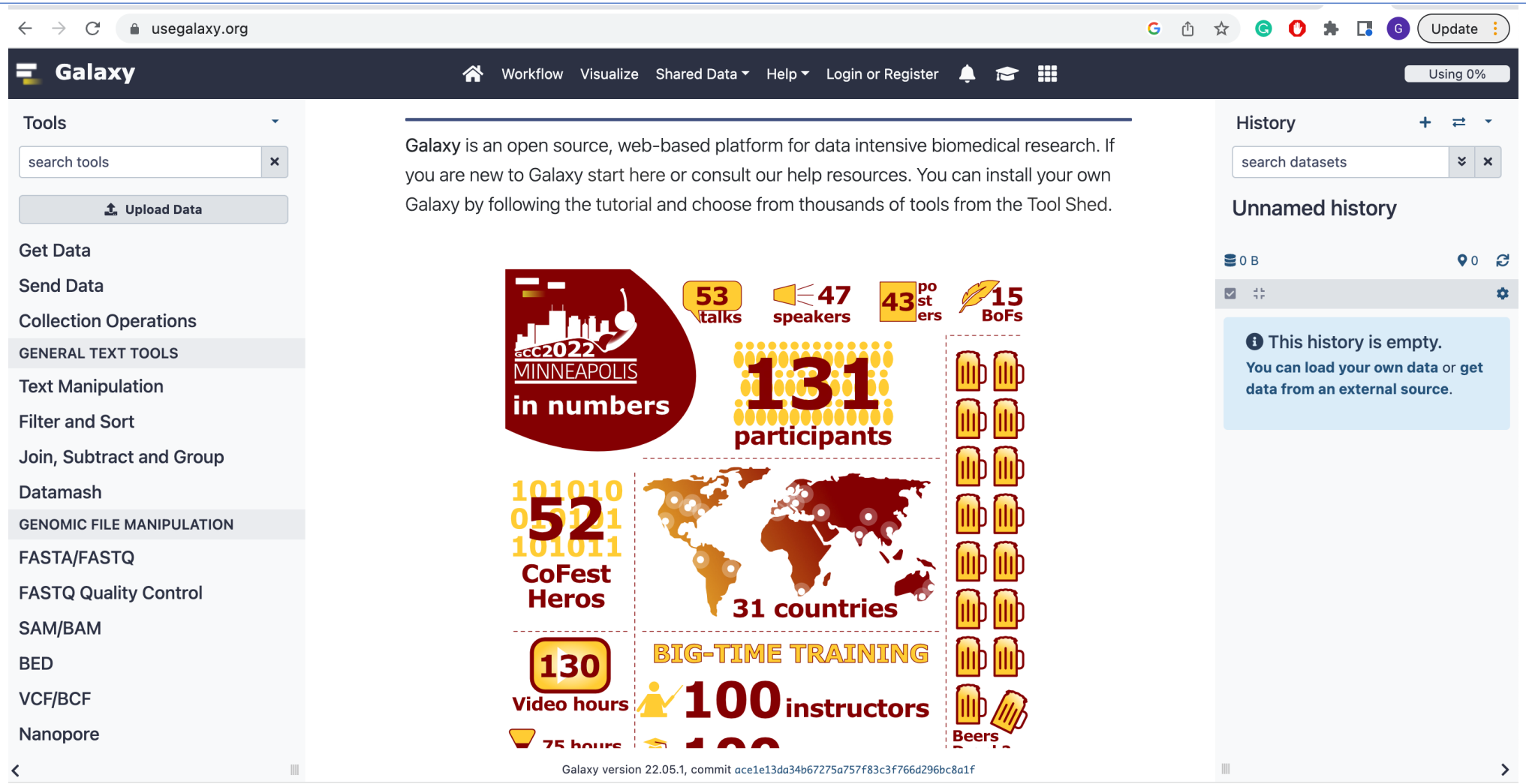
Galaxy is an **open-source** platform for **FAIR data analysis** that enables users to:

- use **tools** from various domains (that can be plugged into **workflows**) through its graphical web interface.
- run code in **interactive environments** (RStudio, Jupyter...) along with other tools or workflows.
- **manage data** by sharing and publishing results, workflows, and visualizations.
- **ensure reproducibility** by capturing the necessary information to repeat and understand data analyses.

The **Galaxy Community** is actively involved in helping the ecosystem improve and sharing scientific discoveries.

Live exercise
Using https://usegalaxy.eu/

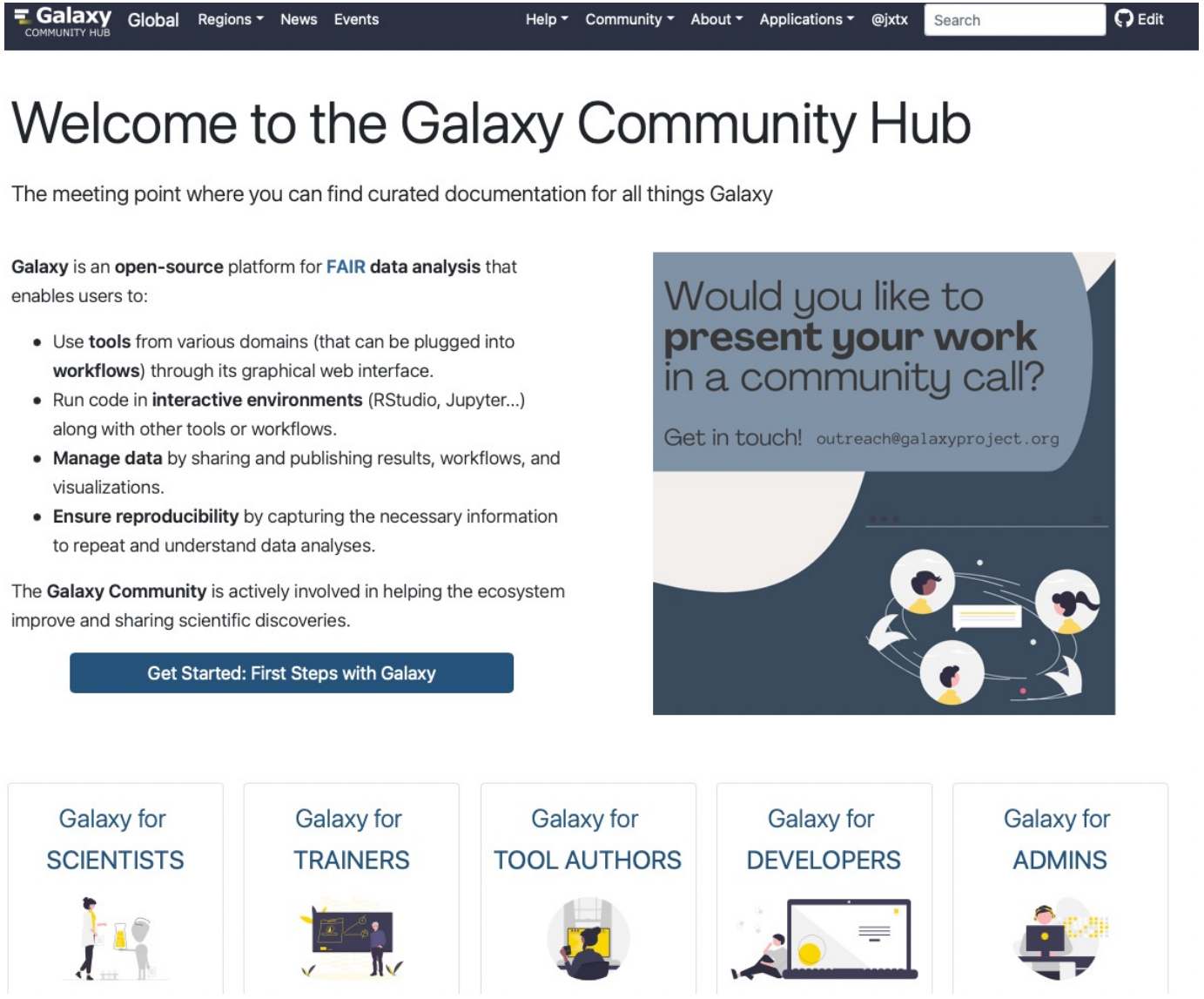
recap



Live exercise

Using https://galaxyproject.org/eu/

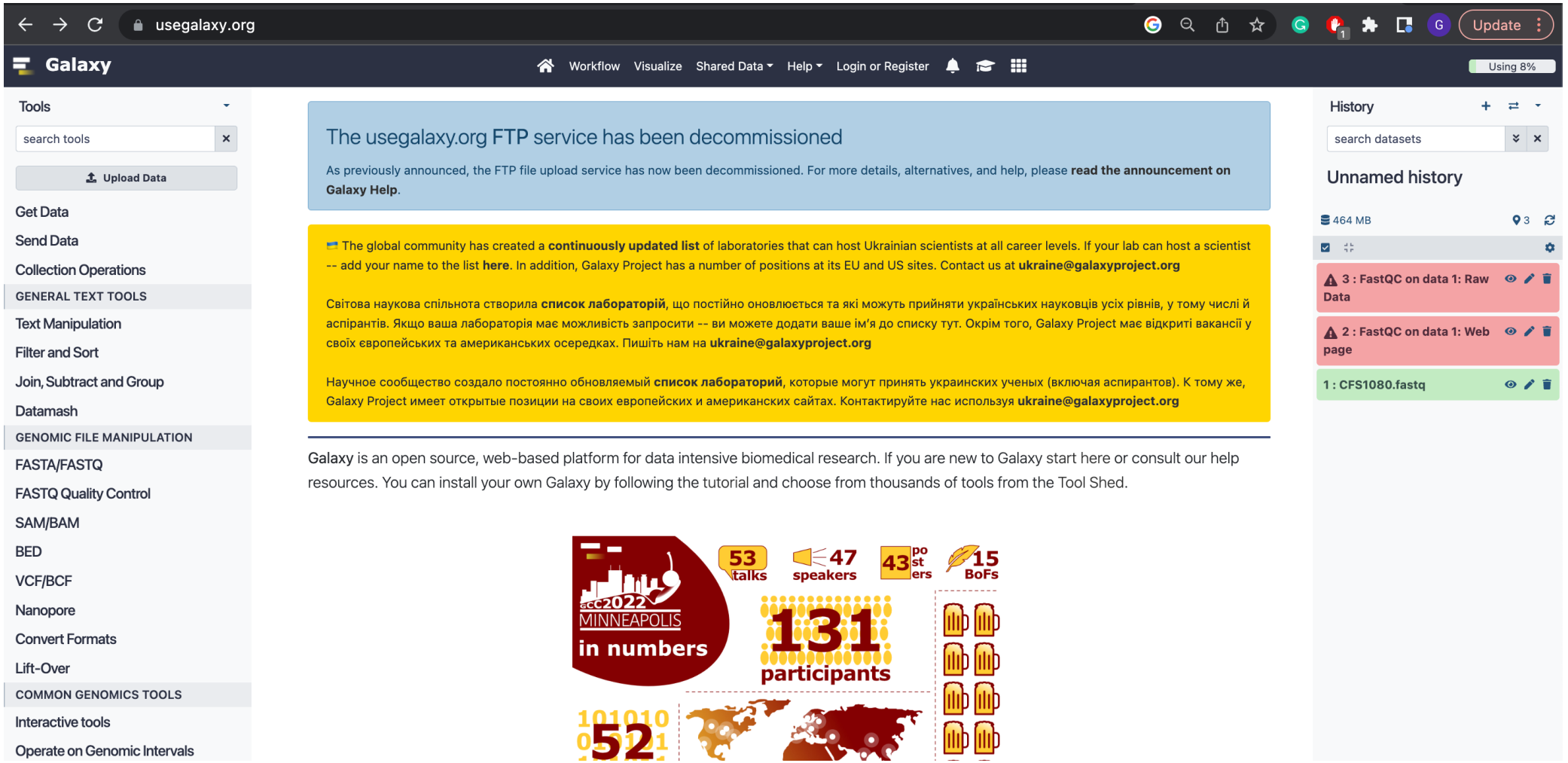
Galaxy introduction to the platform



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

Galaxy introduction to the platform

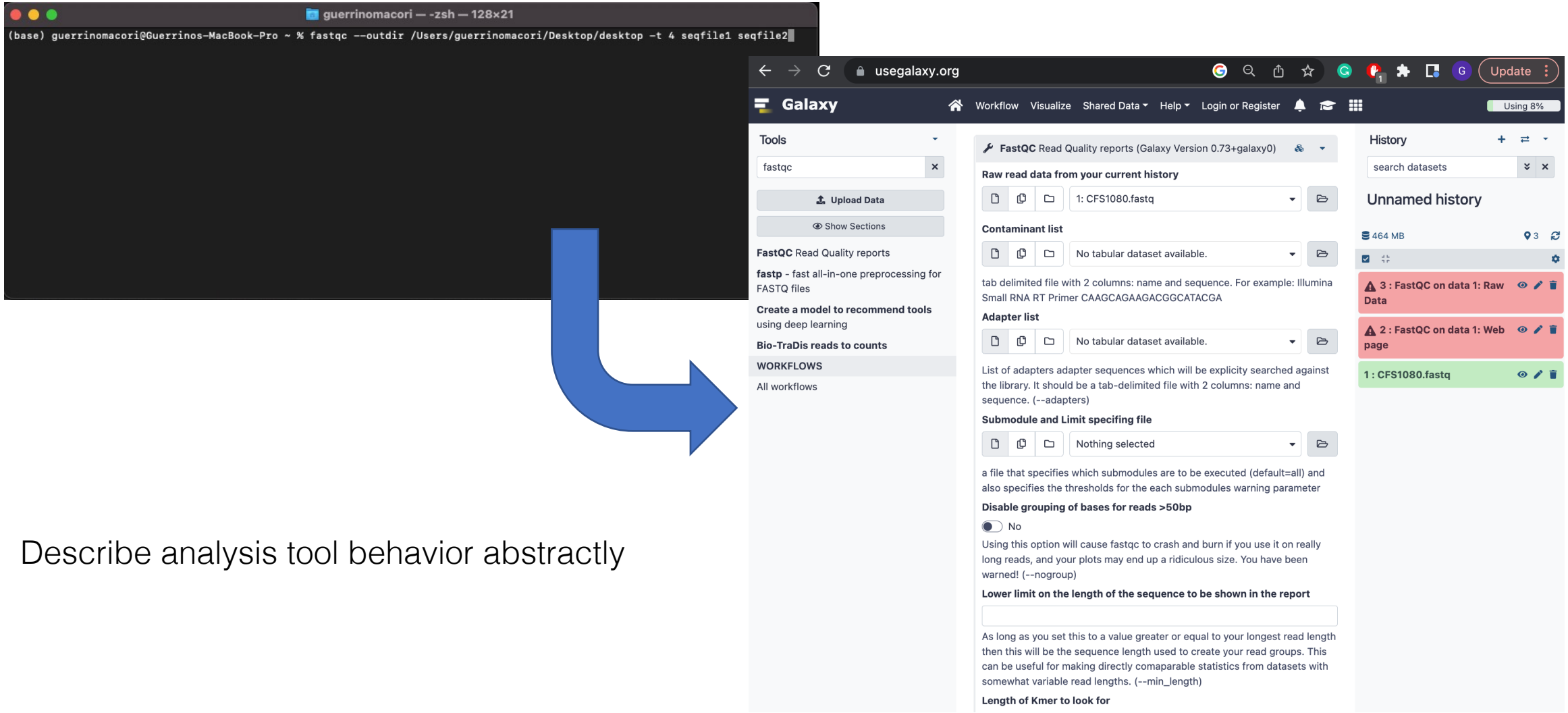
- Accessible analysis system



Galaxy introduction to the platform

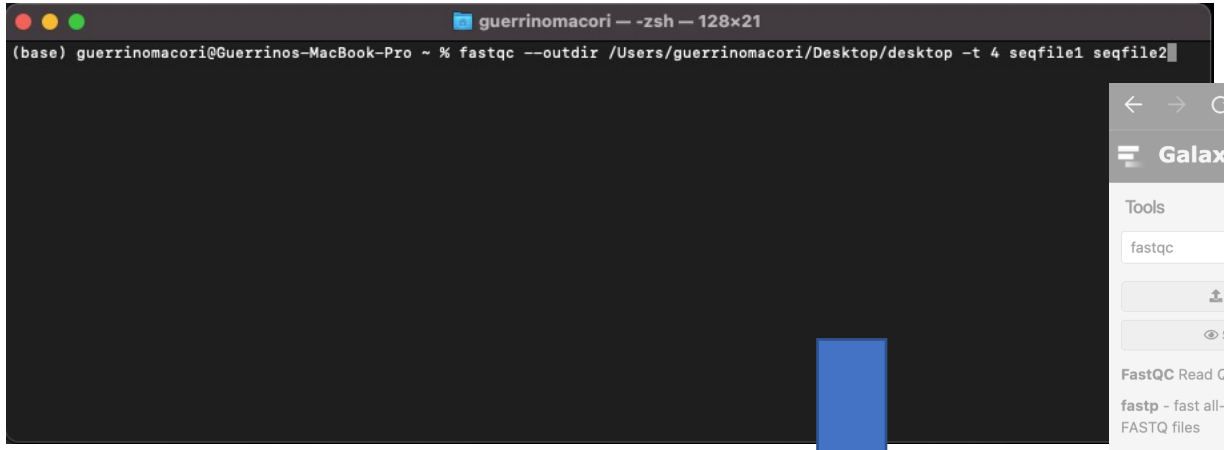
- A free (for everyone) web service integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- Open source software that makes integrating your own tools and data and customizing for your own site simple
- An open extensible platform for sharing tools, datatypes, workflows, ...

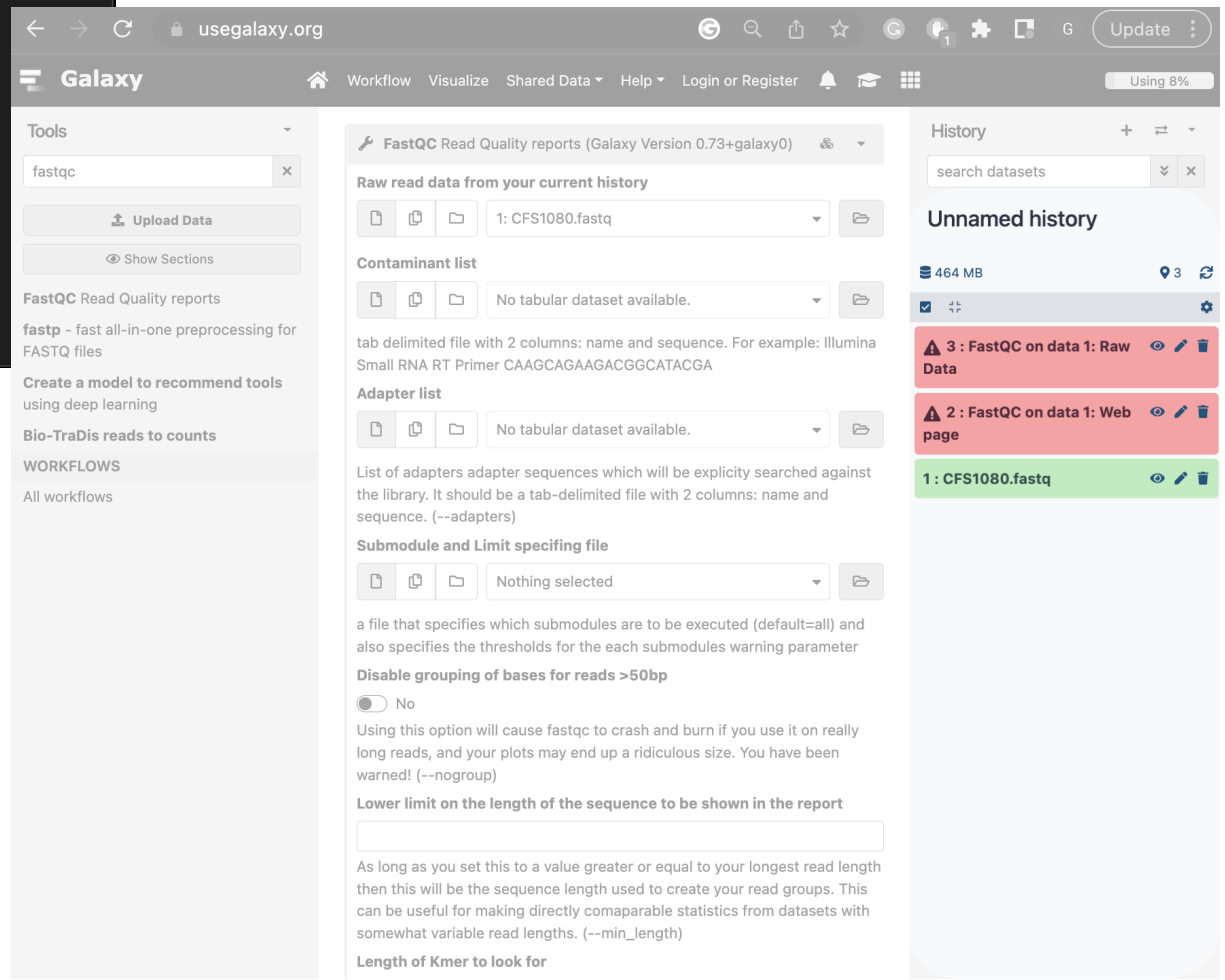
Galaxy introduction to the platform



Describe analysis tool behavior abstractly

Galaxy introduction to the platform

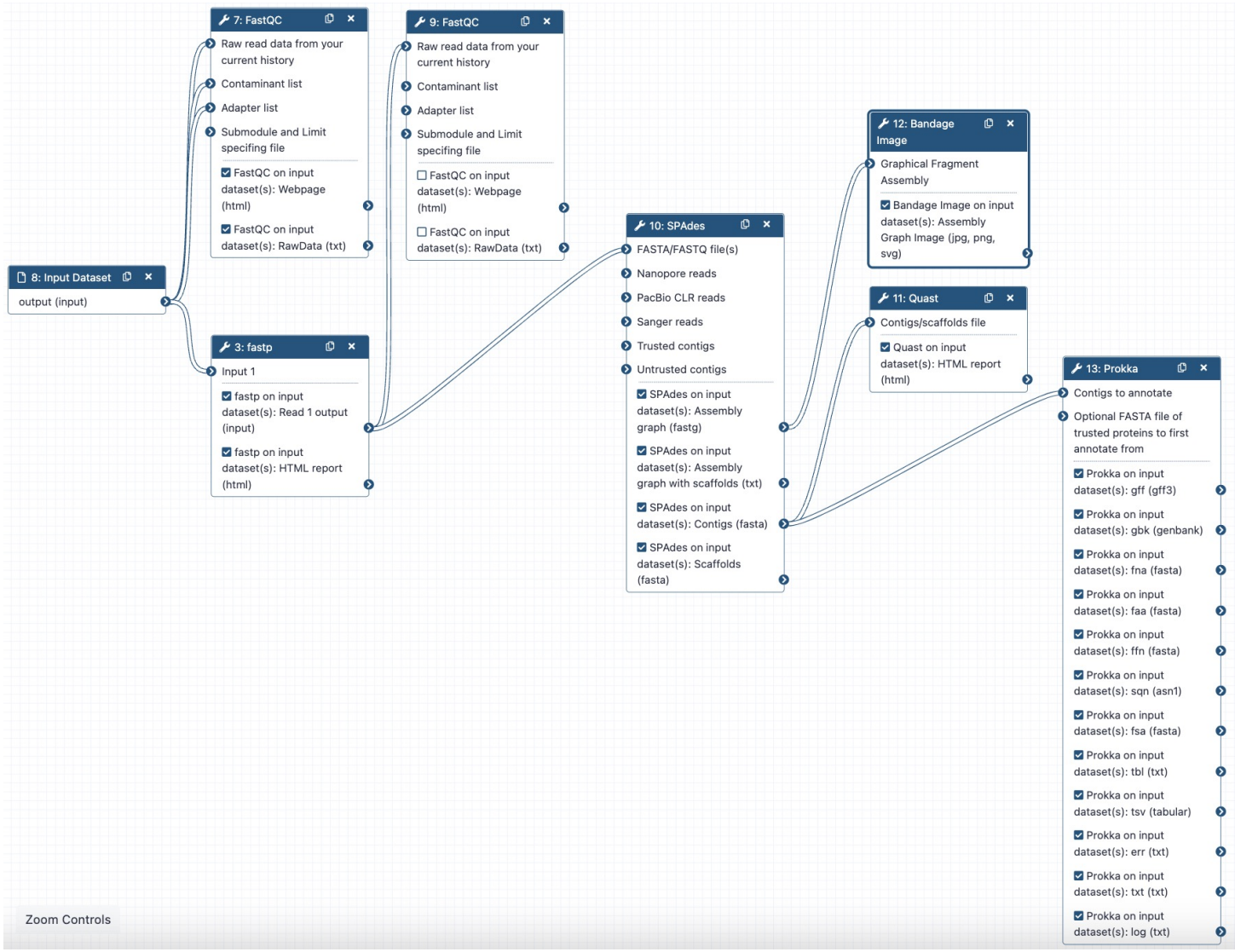




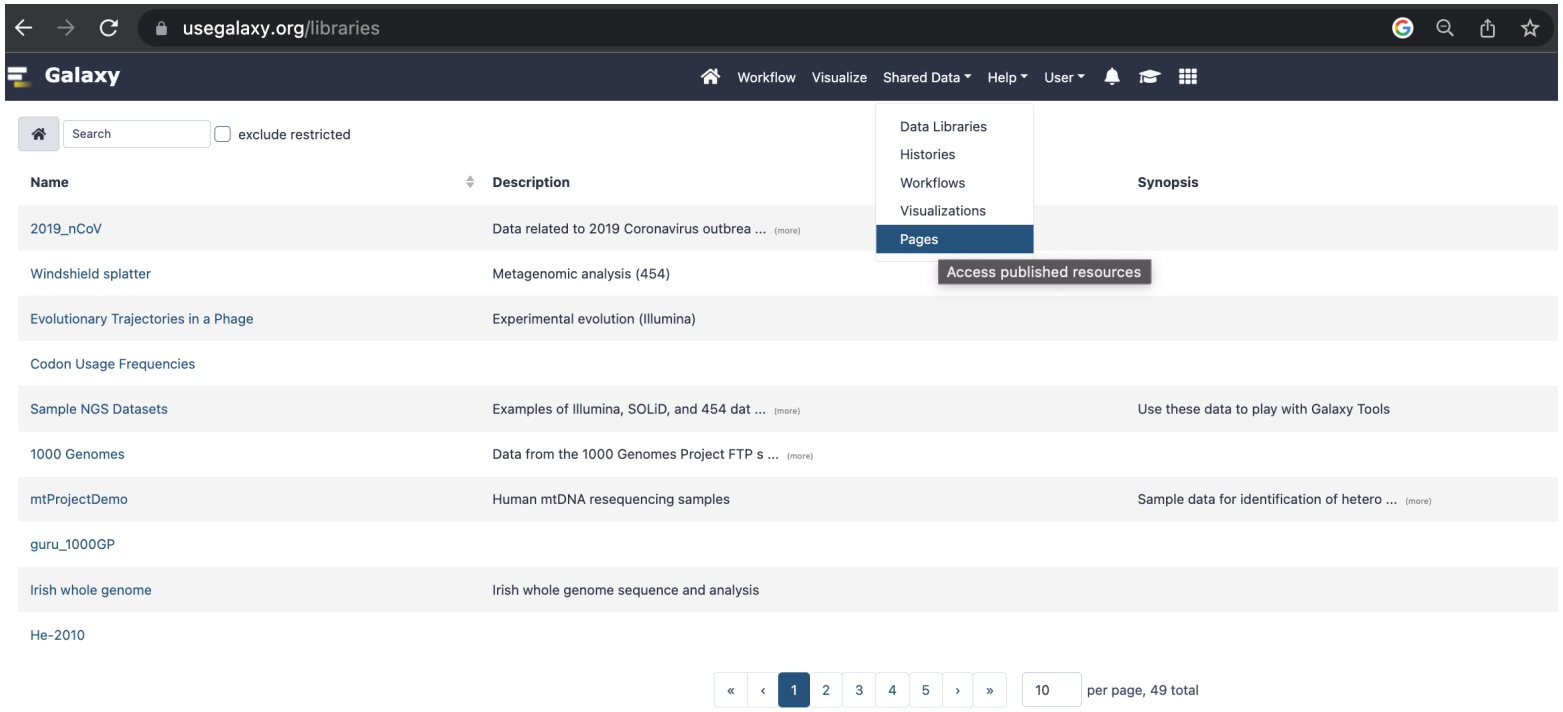
Analysis environment automatically and transparently tracks details

Galaxy introduction to the platform

Workflow system for complex analysis, constructed explicitly or automatically



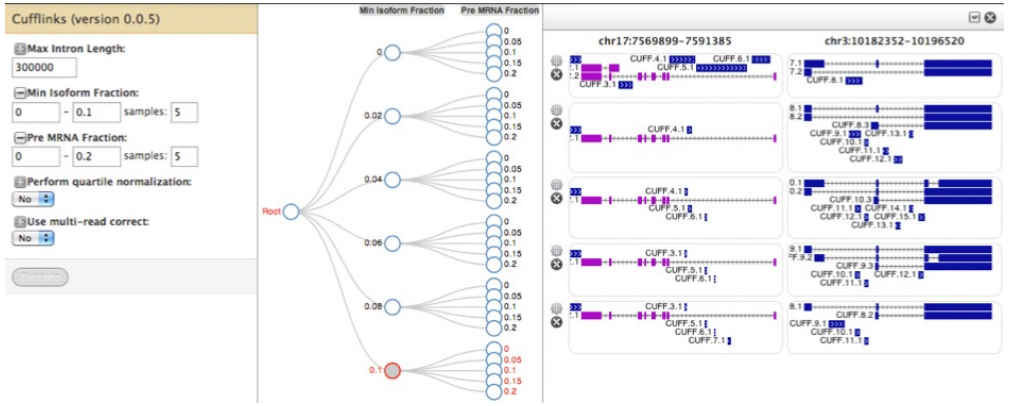
Galaxy introduction to the platform

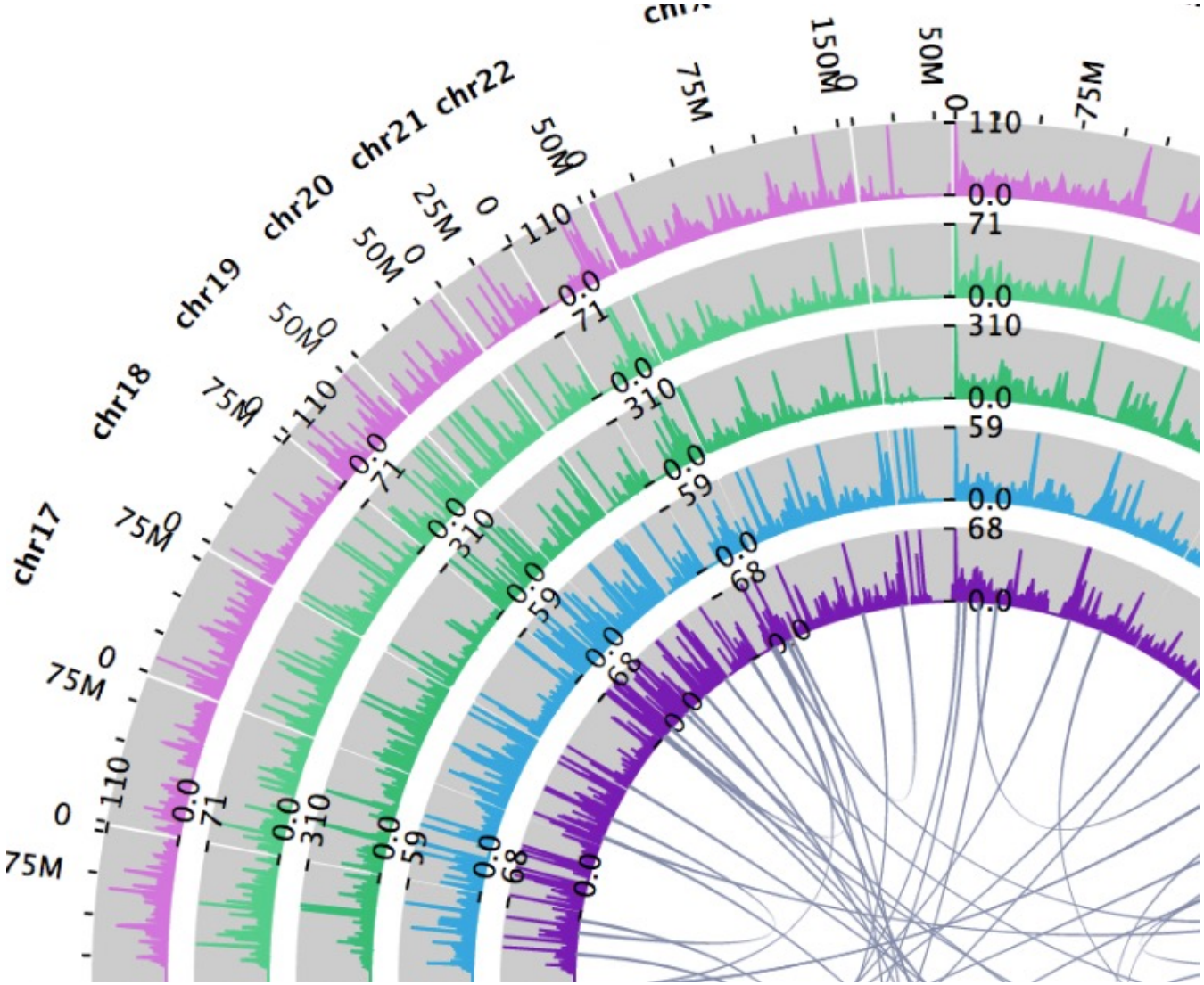


Workflow system for complex analysis, constructed explicitly or automatically

Galaxy introduction to the platform

Visualisation and visual analytics



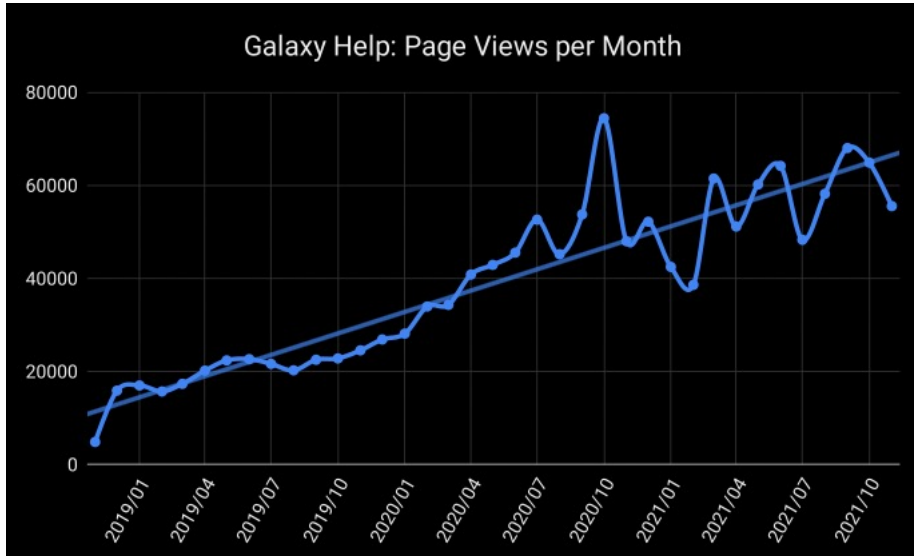


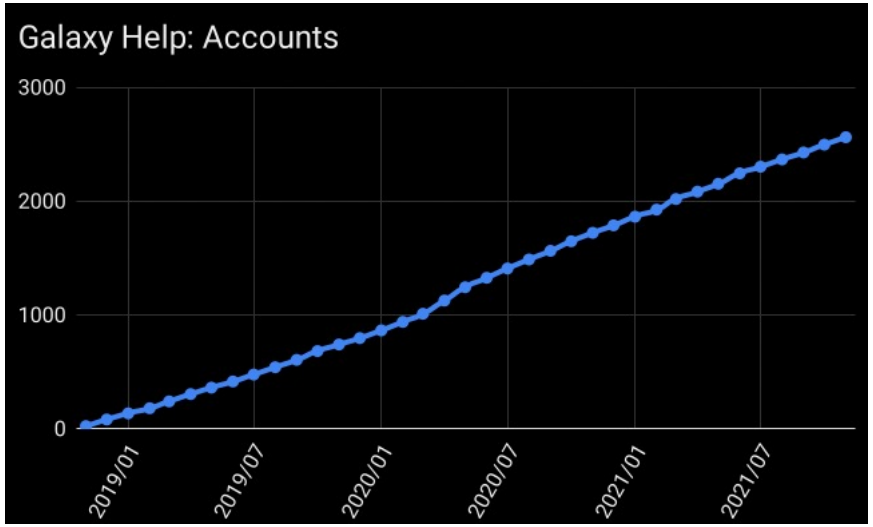
Galaxy introduction to the platform

Galaxy is available as a free (for everyone) web server integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage:
http://usegalaxy.org
However, a centralized solution cannot support the different analysis needs of the entire world!

Galaxy introduction to the platform

https://galaxyproject.org/galaxy-project/statistics/





Galaxy introduction to the platform

Alternatives to the public service
1) Local installation 2) Cloud Computing

Local Galaxy Deployment
Galaxy is designed for local installation and customization...
just download and run
Pluggable interfaces to compute resources, easily connect to one or more existing clusters
Ideally, allow users to take advantage of whatever computational resources they already have access to.

Workshop second part

- Starting from the Quast report of the assemblies
- Genome annotation: structural and functional annotation
- Identification of gene clusters
- Protein characterization
- Summary of pipeline
- Class assessment submission

References
1.Anna Syme, Torsten Seemann, Simon Gladman, **Genome annotation with Prokka (Galaxy Training Materials)**. https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/annotation-with-prokka/tutorial.html Online; accessed Mon Nov 21 2022
2.Batut et al., 2018 **Community-Driven Data Analysis Training for Biology** Cell Systems 10.1016/j.cels.2018.05.012

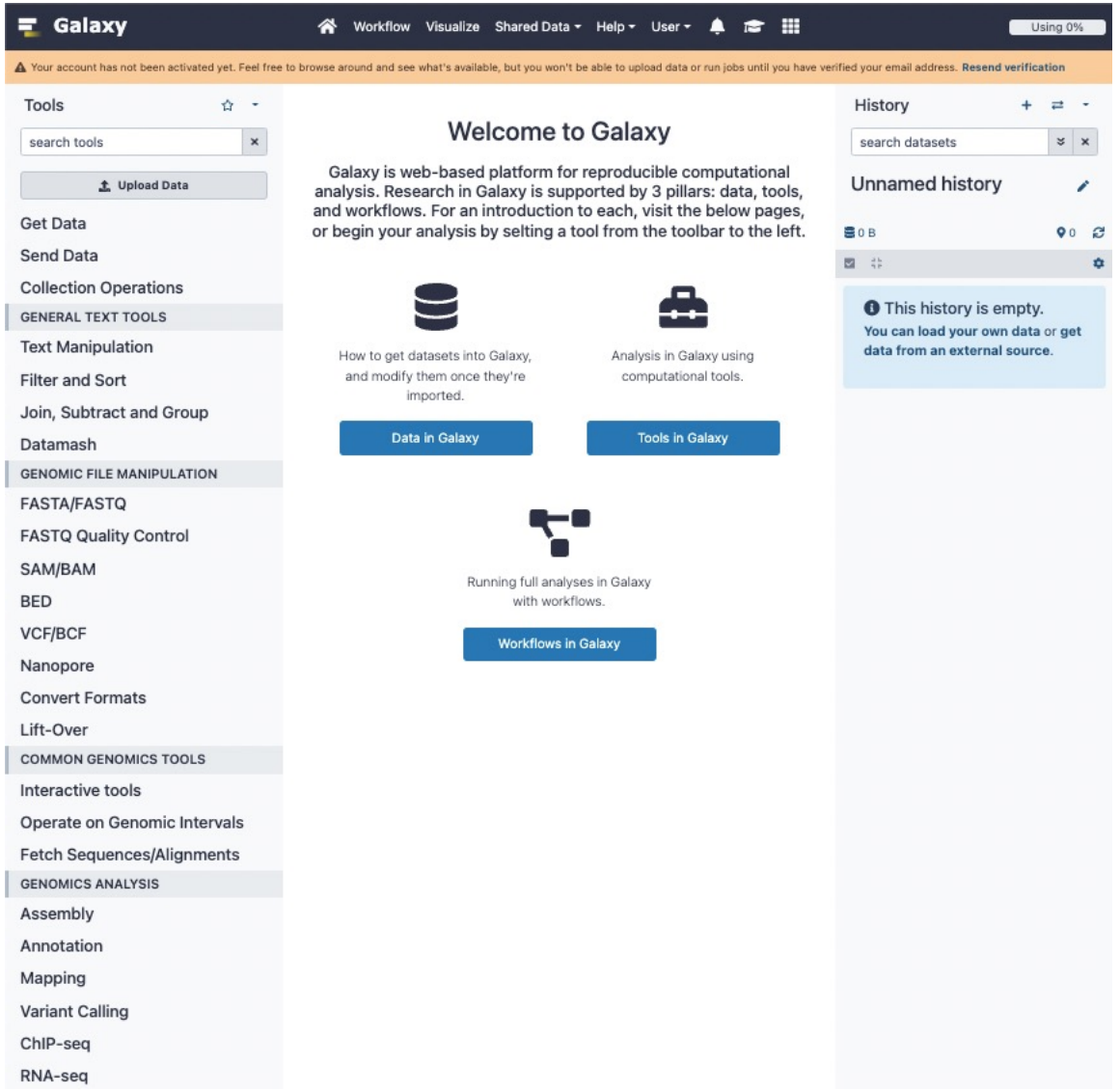
Quality assessment of raw data

Login

http://usegalaxy.eu

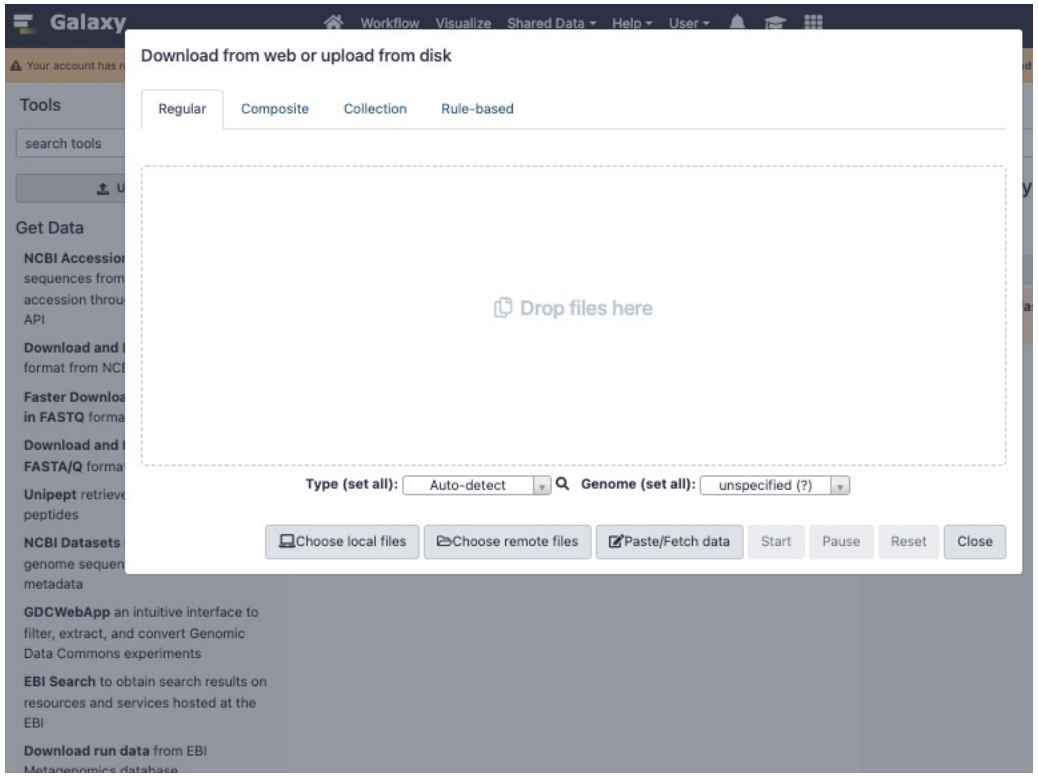
Quality assessment of raw data

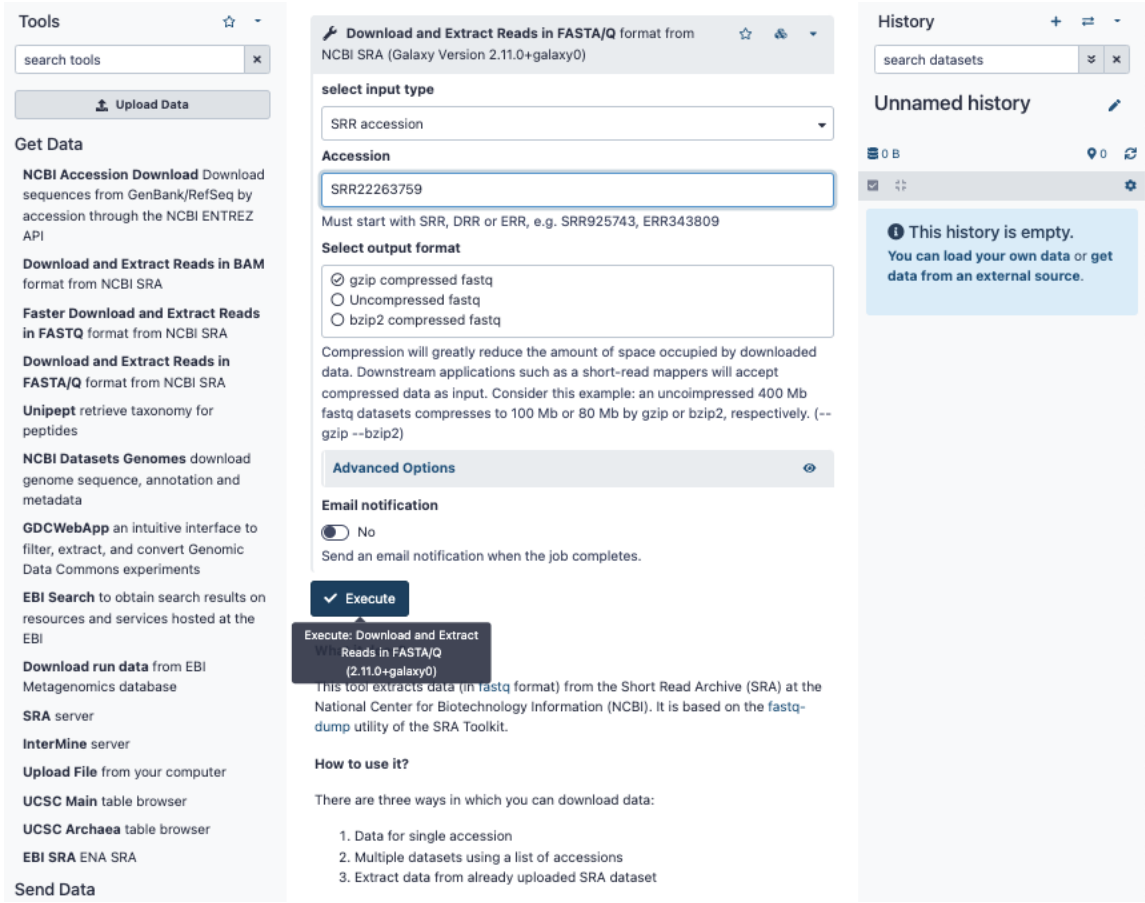
Login
http://usegalaxy.eu



Quality assessment of raw data

Upload raw data





Genome annotation: structural and functional annotation

Genome annotation is the process of attaching biological information to sequences. It consists of three main steps:

- identifying portions of the genome that do not code for proteins
- identifying elements on the genome, a process called gene prediction, and
- attaching biological information to these elements.

Genome annotation: structural and functional annotation

Data format – a recap

- FASTA

DNA and protein sequences are written in FASTA format where you have in the first line a “>” followed by the description. In the second line the sequence starts.

```
>contig_1
TGGGAATAGTAAATTATTTAGCTTAGTTAGCTTATATAAGGCTTTCTAAGCTATTTTTAA
AGAAAGAGCTAACAAAAAGCAAACATCAGCGGAATAAAAAAGGTTGTGAGAAGGTTAAGT
TATGAAAGTTAAGGAGGAATGGTTGTGGACAGACTTAAATTACTAGAGAGTTTAGCATTA
AACAATAATCTTAGTAAATCTGAAATGAACGTTGTTTATTTTTGTTACAAAAAGGGAAGA
ACTTCGAAAGAGATTGCCGTATATTTAGGTTGGGCGTCACCAAATGTTGCAAGACTTTTA
TTGTCAATGTATAATAAGGGCTTTTTAGAAAGAAAGTTGTTGGAAGAAGATAAGAAAACA
TATATATATATAACTAATGACAAAAGTGAATTGTTGCATATTAAGTAAAGTAGTAGTCAT
CCAAGAGGATGGCTTTTTTATTTAAAAGGGTTGCTTTGTTTAAAATAAGTTATTTTTTCT
ATTGACTATTAACTATATAGTGTATAGTATATATTATATATTAAGGAGGCTTGATAATAA
TGAGTTTAGAAGAGTTAGAAGAAATTGAAAGAGAGGAAGGGAAAGAAGCAGTTTGGCAGG
CAATTGAGGATTCAAAGGAATGGGAAGAAATTAATCGTGAAAATTGGGAATTTTAAATTG
GAGGAATTGGAAATGTATATTAAAGATTTGTTTGTTATAGAAAGAACAAGAAATTTATAT
TTTGATGCAGACATTAAAATTAAGCTCGCACCTTGTTATTCACGAGTAACATCAGAACAG
ATACAAGATTTTAATCGAGCTATTAGTGAGTATCGGGGAAGAGTAAATTTTTCTTTTGTA
GAGAAAGCGTATGACAGATATGTATATTCTTTATTTGCAAAGGATTTAGGAGGTCAAACC
ATGCATGATTATTTAATCGCTAATGAGACAGGTTTTACTAATAAGAAGCCAAGTAGTTTT
AGTTTATAGGAGGAACAAAAATGAAAAATATATACCGTTCATTTAGTTGGATTATGCTTG
TGGGGATTTCTCTCTTTTCCGGTTTCACCGCTAACATTGATAAGCACTATCATTGGGAGT
CTTATTCTCCTTACCTAAGTGTAAAAACATTTTTGTTATTAGCTTTGGTGCTACTGATAT
TAAGTCTGCCAGAGATTATCATTTTATTCGTTGAAAGAAAGAAATCAATAGGAGGTAAAT
AAAATGCTAAAAGATCAATGTTTGTGTTGGTTCAAAGATACTTTGTTTACAAGATTTTAT
GATAATGAGTTGTTTTATTTATCTAAGGAAATGTCAGAGAAAAAGTTGCAAGTACTTGAT
AAAATACCTGAAATGAATGAAGGCTTCAATTTCTTTTATTTTGATAATCTTTCTTTTACA
AATGAAAATTTATTATTAATTCCTAATCCACATAATACTTTGGAATTCTTTGTCTTTAAC
TATGTTAAACAGAATGACTTACTTTTTGTTAAAACCCTTGATATTACTGATTATCTTGGA
ATTTTGGAACATCTAGAAGAATTTAAAAATAAGCTTGATTTAATAATAGTAAAAGAGGGG
TATTAAAAATATGCATACTTCAAATAAATTTAAAGAAAATATTTTATTAATAATTGGTTT
GGCTTTATGTATCTTAGCAGGATTACTAAGAAATCATGGAGTAATAGCGGGGATCCTTTT
AATTCCCGGGATTCTTGCACTTTTAATTAGTTCATATTATACGATAAAAAGAAGATTTTC
```

Genome annotation: structural and functional annotation

Data format – a recap

- GFF3

The general feature format (gene-finding format, generic feature format, GFF) is a file format used for describing genes and other features of DNA, RNA and protein sequences.

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
contig_1	AUGUSTUS	gene	60515	61234	0.89	+	.	contig_1.g36
contig_1	AUGUSTUS	transcript	60515	61234	0.89	+	.	contig_1.g36.t1
contig_1	AUGUSTUS	start_codon	60515	60517	.	+	0	transcript_id "contig_1.g36.t1"; gene_id "contig_1.g36";
contig_1	AUGUSTUS	CDS	60515	61234	0.89	+	0	transcript_id "contig_1.g36.t1"; gene_id "contig_1.g36";

Genome annotation: structural and functional annotation

Data format – a recap

- Genebank

The genbank sequence format is a rich format for storing sequences and associated annotations.

Sequence 89 from Patent WO2016185057

GenBank: MS811295.1

FASTA Graphics

Go to:

```
LOCUS       MS811295                504 bp    DNA     linear   PAT 12-JUL-2017
DEFINITION  Sequence 89 from Patent WO2016185057.
ACCESSION   MS811295
VERSION     MS811295.1
KEYWORDS    .
SOURCE      Klebsiella
  ORGANISM  Klebsiella
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
            Enterobacteriaceae; Klebsiella/Raoultella group.
REFERENCE   1
  AUTHORS   LOMBO,B.F., VILLAR,G.C.J., MARLN,F.L., FERNANDEZ,F.J. and
            GUTIERREZ,D.R.M.I.
  TITLE     RECOMBINANT NUCLEIC ACID FOR USE IN THE PRODUCTION OF POLYPHENOLS
  JOURNAL   Patent: WO 2016185057-A2 89 24-NOV-2016;
            UNIV OVIEDO [ES]
FEATURES             Location/Qualifiers
     source          1..504
                     /organism="Klebsiella"
                     /mol_type="unassigned DNA"
                     /db_xref="taxon:570"
     misc_feature    1..504
                     /note="Sequence in description"
ORIGIN
        1 gaattcccgg ggatccggtg attgattgag caagctttat gcttgtaaac cgttttgtga
       61 aaaaattttt aaaataaaaa aggggacctc tagggtcccc aattaattag taatataatc
      121 tattaaaggt cattcaaaag gtcatccacc ggatcaattc ccctgctcgc gcaggctggg
      181 tgccaagctc tcgggtaaca tcaaggcccg atccttggag cccttgccct cccgcacgat
      241 gatcgtgccg tgatcgaaat ccagatcctt gacccgcagt tgcaaaccct cactgatccg
      301 gctcacggta actgatgccg tatttgcagt accagcgtac ggcccacaga atgatgtcac
      361 gctgaaaatg ccggcctttg aatgggttca tgtgcagctc catcagcaaa aggggatgat
      421 aagtttatca ccaccgacta tttgcaacag tgccgttgat cgtgctatga tcgactgatg
      481 tcatcagcgg tggagtgcaa tgtc
//
```

Genome annotation: structural and functional annotation

Structural Annotation

Sequence Features

1. Count the number of bases in your sequence (**compute sequence length**)
2. Check for sequence composition and GC content.