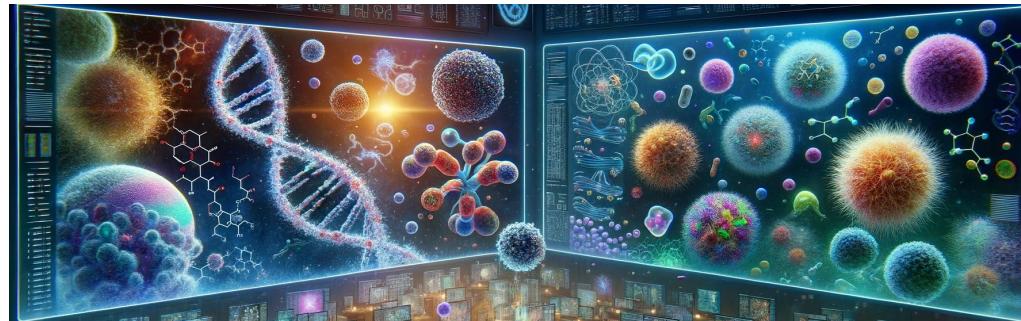


# Advanced Applications of Next Generation Sequencing in Food Safety



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](mailto:guerrino.macori@ucd.ie)  
@guerrinomacori



---

## **Advanced Applications of Next Generation Sequencing in Food Safety Scientific Programme**

**DAY 1**

**Monday 22<sup>nd</sup> 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

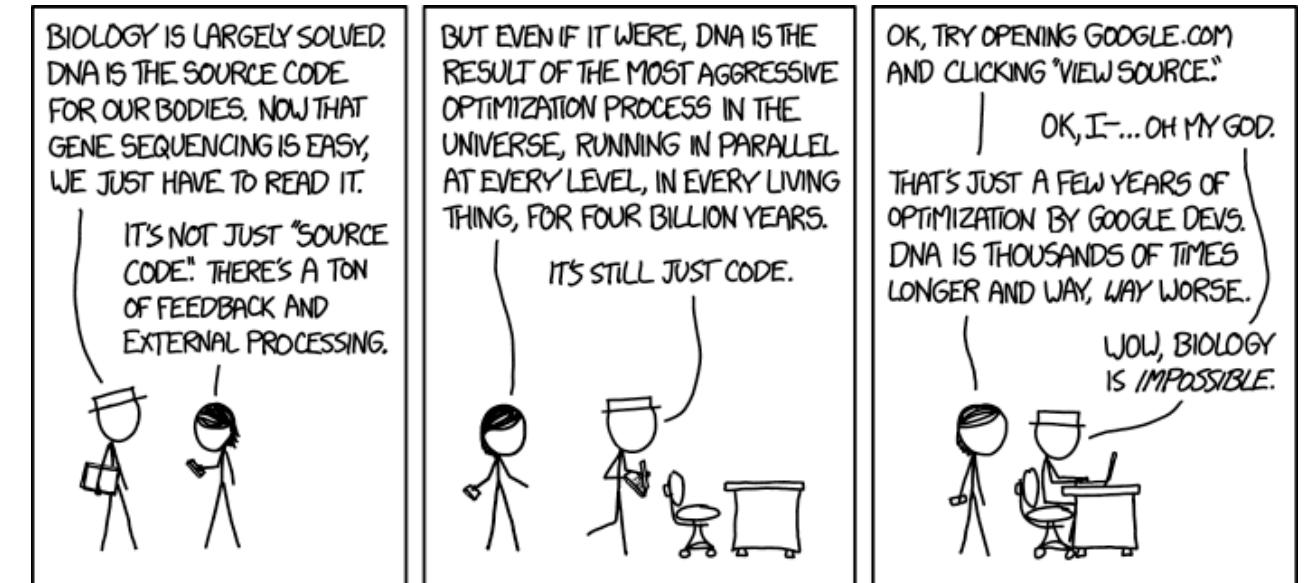
**16:30 Industry talk**



## 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/>)

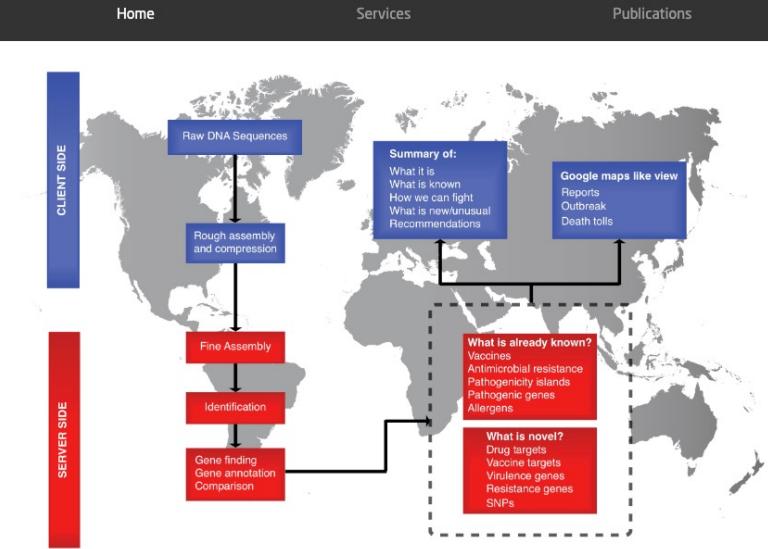
## Live exercise

## 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
```

## Center for Genomic Epidemiology



### 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](#)

Identification of acquired antibiotic resistance genes.

##### [ResFinderFG](#)

Identification of functional metagenomic antibiotic resistance determinants.

##### [LRE-finder](#)

Identification of genes and mutations leading to linezolid resistance.

##### [KmerResistance](#)

Identification of acquired antibiotic resistance genes using Kmers.

##### [PathogenFinder](#)

Prediction of a bacteria's pathogenicity towards human hosts.

##### [VirulenceFinder](#)

Identification 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

##### [MINType](#)

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

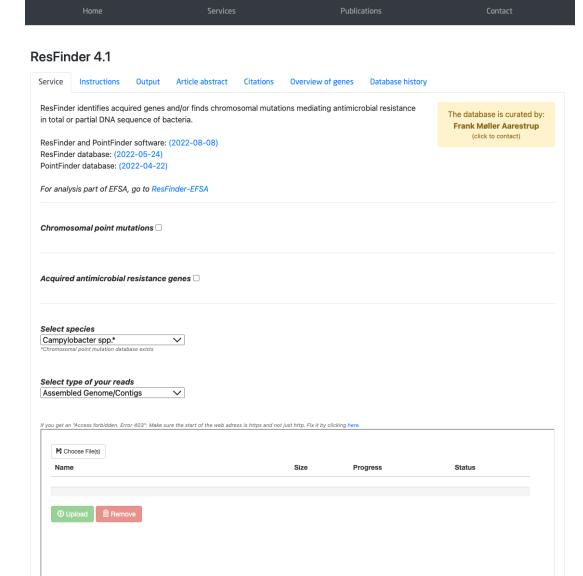
##### [RUCS](#)

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

#### Other

##### [MyKMAfinder](#)

## Center for Genomic Epidemiology



The screenshot shows the ResFinder 4.1 service page. At the top, there are links for Home, Services, Publications, and Contact. Below that, there are tabs for Service, Instructions, Output, Article abstract, Citations, Overview of genes, and Database history. A note says "The database is curated by Frank Møller Aarestrup click to contact". Under the Service tab, there is information about the software (ResFinder and PointFinder), databases (ResFinder and PointFinder), and a note about EFSA. There are sections for Chromosomal point mutations, Acquired antimicrobial resistance genes, and Select species (Campylobacter spp.). A table for Select type of your reads (Assembled Genome/Contigs) is shown with columns for Choose Field, Name, Size, Progress, and Status. Buttons for Upload and Remove are at the bottom.

## Live exercise web resources

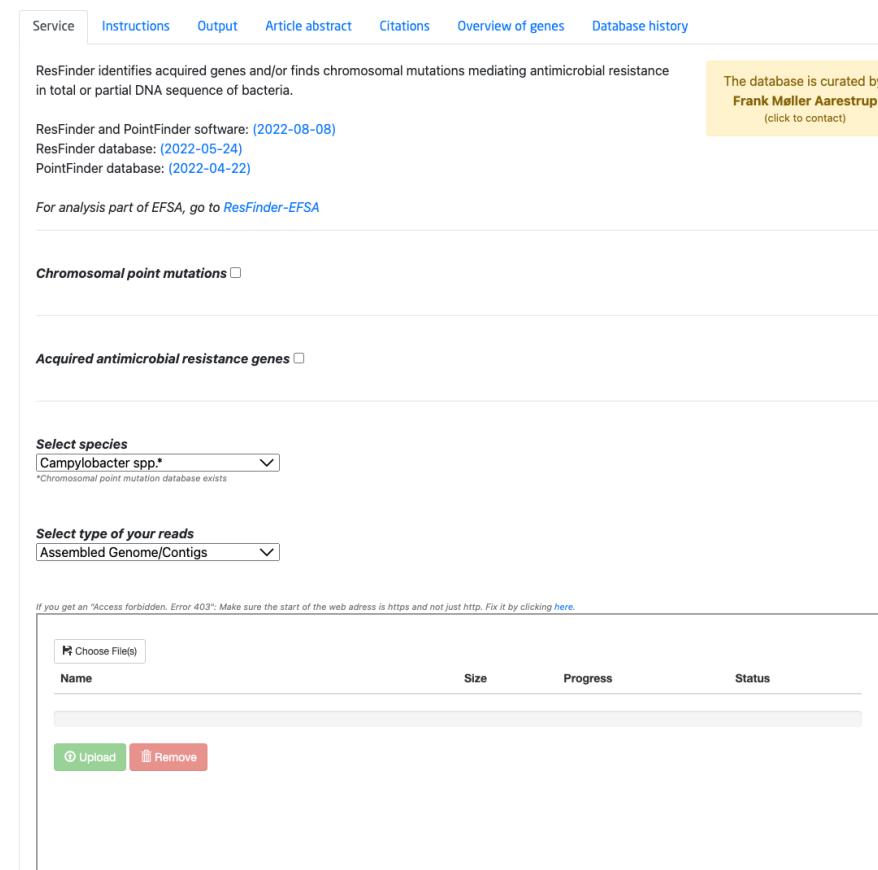
<http://www.genomicepidemiology.org/services/>



Center for Genomic Epidemiology

Home Services Publications Contact

### ResFinder 4.1



Service Instructions Output Article abstract Citations Overview of genes Database history

ResFinder identifies acquired genes and/or finds chromosomal mutations mediating antimicrobial resistance in total or partial DNA sequence of bacteria.

ResFinder and PointFinder software: (2022-08-08)  
ResFinder database: (2022-05-24)  
PointFinder database: (2022-04-22)

For analysis part of EFSA, go to [ResFinder-EFSA](#)

The database is curated by:  
**Frank Møller Aarestrup**  
(click to contact)

**Chromosomal point mutations** □

**Acquired antimicrobial resistance genes** □

Select species  
Campylobacter spp.\*

\*Chromosomal point mutation database exists

Select type of your reads  
Assembled Genome/Contigs

If you get an "Access forbidden. Error 403": Make sure the start of the web address is https and not just http. Fix it by clicking [here](#).

Choose File(s)  
Name Size Progress Status  
  
Upload Remove

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/>

[usegalaxy.org](https://usegalaxy.org)
G Update

**Galaxy**
Using 0%

Tools + ↗  
 x  
Upload Data

Get Data  
 Send Data  
 Collection Operations  
**GENERAL TEXT TOOLS**  
 Text Manipulation  
 Filter and Sort  
 Join, Subtract and Group  
 Datamash  
**GENOMIC FILE MANIPULATION**  
 FASTA/FASTQ  
 FASTQ Quality Control  
 SAM/BAM  
 BED  
 VCF/BCF  
 Nanopore

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.



**in numbers**  
**52** CoFest Heros  
**131** participants  
**31** countries  
**130** Video hours  
**75** hours

**53 talks** **47 speakers** **43 posters** **15 BoFs**

**BIG-TIME TRAINING** **100 instructors**

**Beers**

Galaxy version 22.05.1, commit ace1e13da34b67275a757f83c3f766d296bc8a1f

History + ↗  
 x  
**Unnamed history**  
0 B 0 Beers  
This history is empty.  
You can load your own data or get data from an external source.

## Live exercise

Using <https://galaxyproject.org/eu/>

# Galaxy introduction to the platform



The screenshot shows the Galaxy Community Hub homepage. At the top, there's a navigation bar with links for "Galaxy", "Global", "Regions", "News", "Events", "Help", "Community", "About", "Applications", a user handle "@jxtx", a search bar, and an "Edit" button. The main title "Welcome to the Galaxy Community Hub" is displayed prominently. Below it, a sub-headline says "The meeting point where you can find curated documentation for all things Galaxy". A text block explains that Galaxy is an open-source platform for FAIR data analysis, enabling users to use tools from various domains, run code in interactive environments, manage data by sharing and publishing results, and ensure reproducibility by capturing necessary information. A large call-to-action button at the bottom left says "Get Started: First Steps with Galaxy".

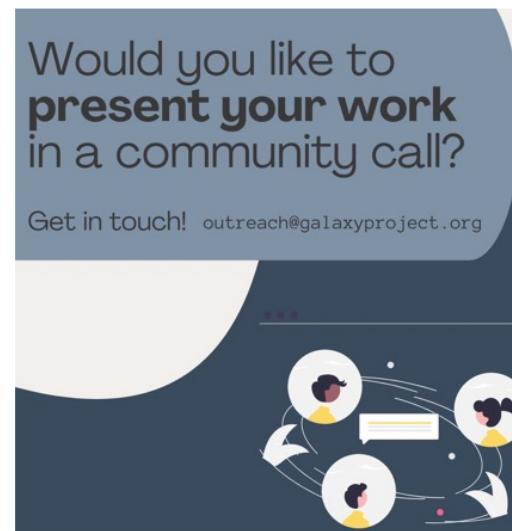
## Welcome to the Galaxy Community Hub

The meeting point where you can find curated documentation for all things Galaxy

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.



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



The grid consists of five cards, each representing a different user group for Galaxy:

- Galaxy for SCIENTISTS**: An icon shows two scientists in a lab setting.
- Galaxy for TRAINERS**: An icon shows a person giving a presentation.
- Galaxy for TOOL AUTHORS**: An icon shows a person working on a computer.
- Galaxy for DEVELOPERS**: An icon shows a person working on a laptop.
- Galaxy for ADMINS**: An icon shows a person working on a computer with gears.

# Galaxy introduction to the platform

- Accessible analysis system

The screenshot shows the usegalaxy.org web interface. On the left, a sidebar lists various genomic tools and data manipulation options. The main content area displays a message about the decommissioning of the FTP service, followed by a yellow box containing text in English and Russian. Below this, a section about Galaxy's open source nature and available resources is shown. At the bottom, a graphic titled "in numbers" provides a summary of the event: 53 talks, 47 speakers, 43 posters, 15 BoFs, 131 participants, and 52 countries represented.

The usegalaxy.org FTP service has been decommissioned

As previously announced, the FTP file upload service has now been decommissioned. For more details, alternatives, and help, please [read the announcement on Galaxy Help](#).

The global community has created a [continuously updated list](#) of laboratories that can host Ukrainian scientists at all career levels. If your lab can host a scientist -- add your name to the list [here](#). In addition, Galaxy Project has a number of positions at its EU and US sites. Contact us at [ukraine@galaxyproject.org](mailto:ukraine@galaxyproject.org)

Світова наукова спільнота створила [спісок лабораторій](#), що постійно оновлюється та які можуть прийняти українських науковців усіх рівнів, у тому числі аспірантів. Якщо ваша лабораторія має можливість запросити -- ви можете додати ваше ім'я до списку тут. Окрім того, Galaxy Project має відкріті вакансії у своїх європейських та американських осередках. Пишіть нам на [ukraine@galaxyproject.org](mailto:ukraine@galaxyproject.org)

Научное сообщество создало постоянно обновляемый [список лабораторий](#), которые могут принять украинских ученых (включая аспирантов). К тому же, Galaxy Project имеет открытые позиции на своих европейских и американских сайтах. Контактируйте нас используя [ukraine@galaxyproject.org](mailto:ukraine@galaxyproject.org)

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

**in numbers**

- 53 talks
- 47 speakers
- 43 posters
- 15 BoFs
- 131 participants
- 52 countries

# 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



The image shows a comparison between a terminal session and the Galaxy web interface.

**Terminal Session:** On the left, a terminal window titled "guerrinomacori -- zsh -- 128x21" shows the command: "(base) guerrinomacori@Guerrinos-MacBook-Pro ~ % fastqc --outdir /Users/guerrinomacori/Desktop/desktop -t 4 seqfile1 seqfile2".

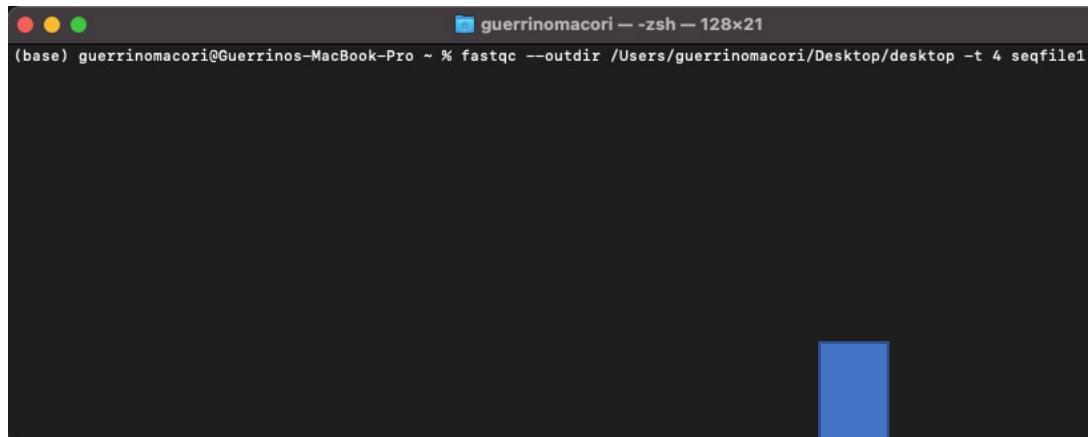
**Galaxy Web Interface:** On the right, the Galaxy homepage is displayed at "usegalaxy.org". The search bar contains "FastQC Read Quality reports (Galaxy Version 0.73+galaxy0)". The interface includes sections for "Raw read data from your current history" (containing "1: CFS1080.fastq"), "Contaminant list" (noting "No tabular dataset available."), "Adapter list" (noting "No tabular dataset available."), "Submodule and Limit specifying file" (set to "Nothing selected"), "Disable grouping of bases for reads >50bp" (radio button set to "No"), and "Lower limit on the length of the sequence to be shown in the report" (input field empty). The "History" panel shows an "Unnamed history" with a dataset named "1: CFS1080.fastq".

Describe analysis tool behavior abstractly

# Galaxy introduction to the platform



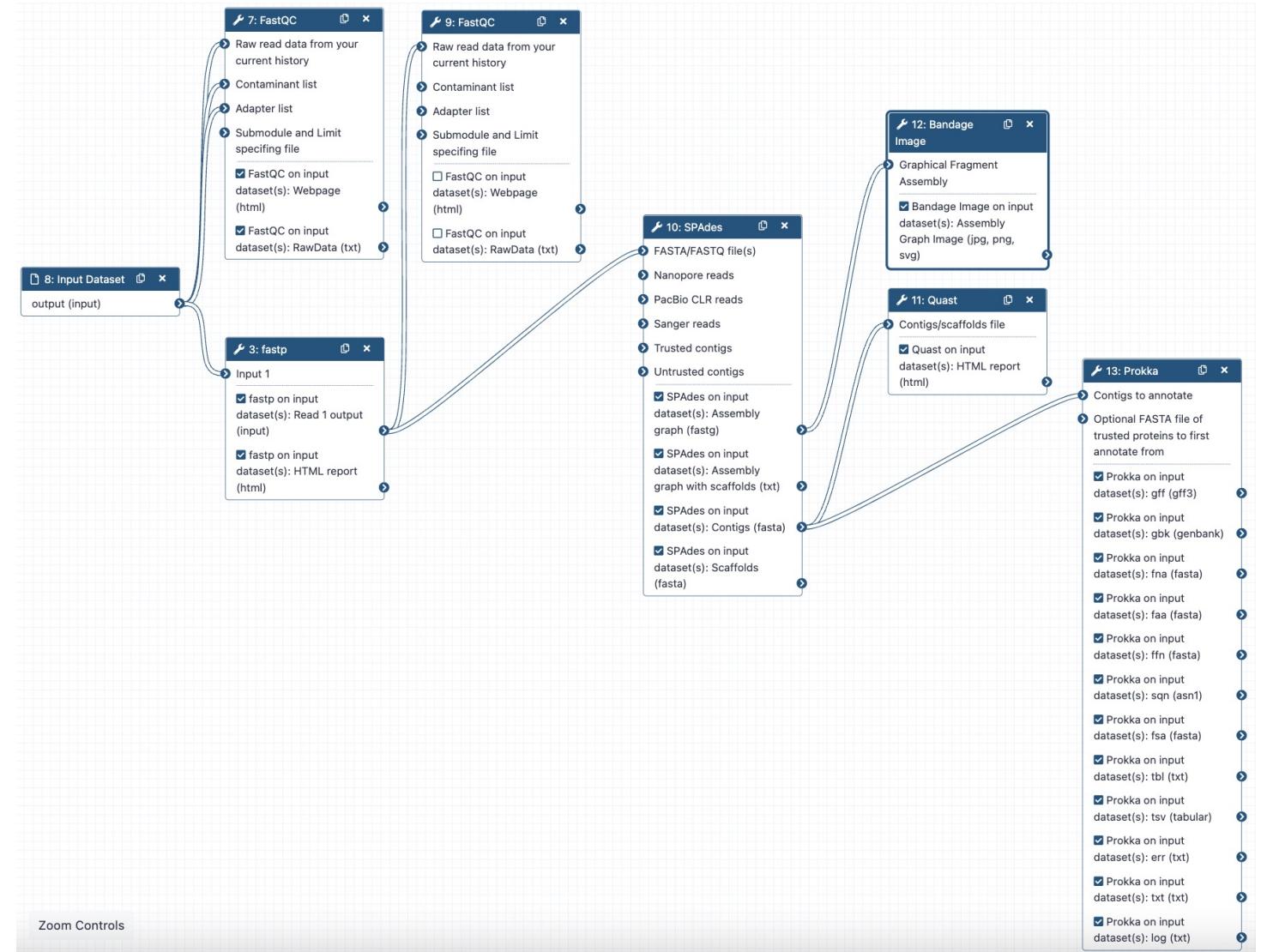
Analysis environment automatically and transparently tracks details



guerrinomacori@Guerrinos-MacBook-Pro ~ % fastqc --outdir /Users/guerrinomacori/Desktop/desktop -t 4 seqfile1 seqfile2

The Galaxy web interface shows the 'FastQC Read Quality reports' tool configuration. It includes fields for Raw read data (1: CFS1080.fastq), Contaminant list, Adapter list, Submodule and Limit specifying file, Disable grouping of bases for reads >50bp (set to No), and Lower limit on the length of the sequence to be shown in the report (set to 1). The History panel shows three recent datasets: '3 : FastQC on data 1: Raw Data' (red), '2 : FastQC on data 1: Web page' (red), and '1 : CFS1080.fastq' (green).

# Galaxy introduction to the platform



Workflow system for complex analysis,  
constructed explicitly or automatically

# Galaxy introduction to the platform

The screenshot shows the Galaxy Data Libraries page at [usegalaxy.org/libraries](https://usegalaxy.org/libraries). The page displays a list of datasets categorized by name and description. A sidebar on the right provides links to Data Libraries, Histories, Workflows, Visualizations, and Pages, with 'Pages' currently selected. A button labeled 'Access published resources' is also visible. At the bottom, there is a navigation bar with page numbers and a per-page selection.

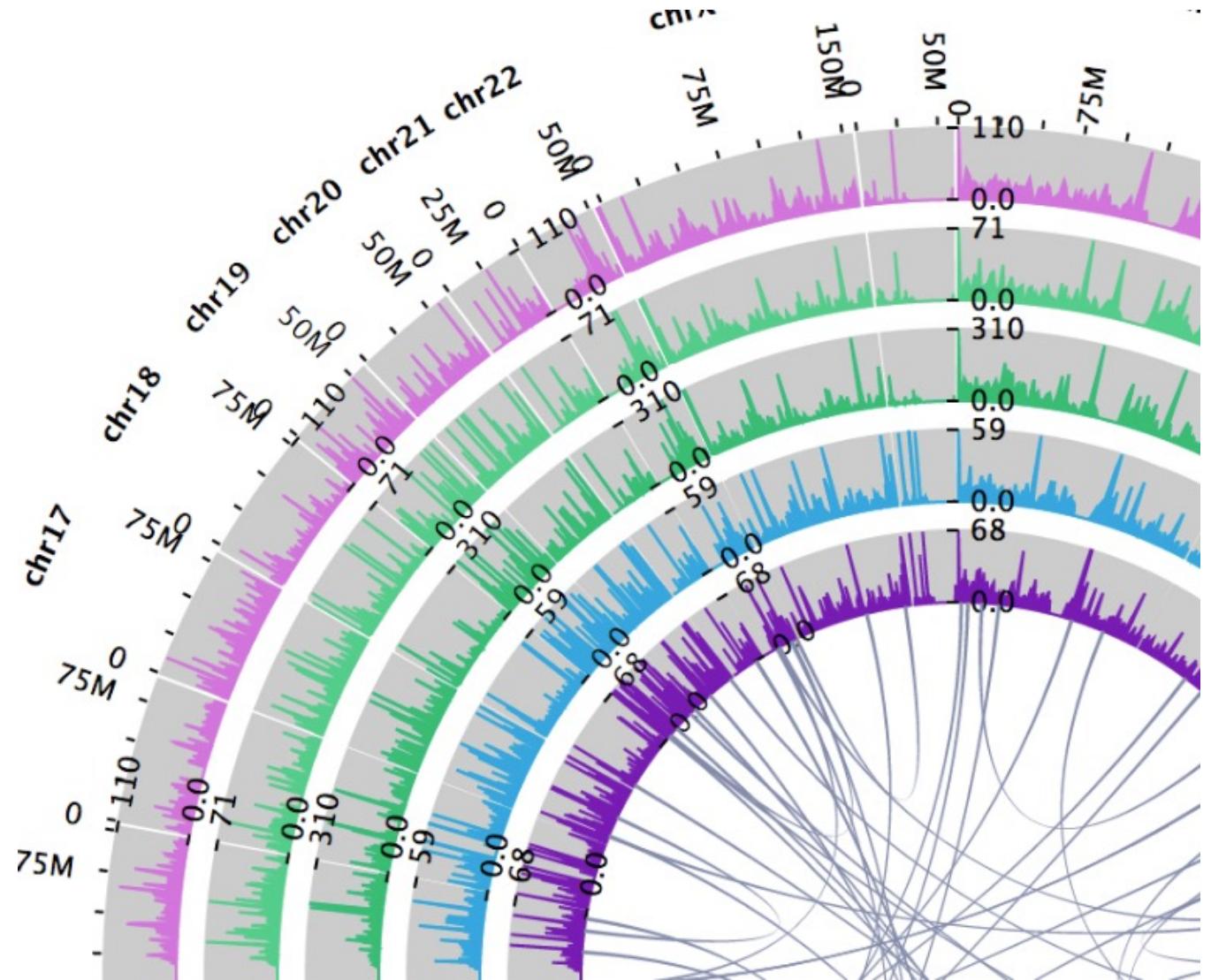
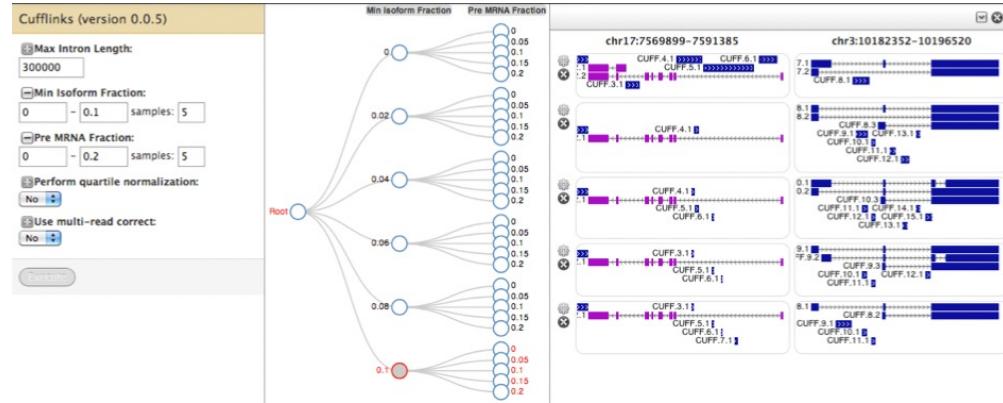
Name	Description	Synopsis
2019_nCoV	Data related to 2019 Coronavirus outbreak ... (more)	
Windshield splatter	Metagenomic analysis (454)	
Evolutionary Trajectories in a Phage	Experimental evolution (Illumina)	
Codon Usage Frequencies		
Sample NGS Datasets	Examples of Illumina, SOLiD, and 454 dat ... (more)	Use these data to play with Galaxy Tools
1000 Genomes	Data from the 1000 Genomes Project FTP s ... (more)	
mtProjectDemo	Human mtDNA resequencing samples	Sample data for identification of hetero ... (more)
guru_1000GP		
Irish whole genome	Irish whole genome sequence and analysis	
He-2010		

« < 1 2 3 4 5 > » 10 per page, 49 total

Workflow system for complex analysis,  
constructed explicitly or automatically

# Galaxy introduction to the platform

Visualisation and visual analytics



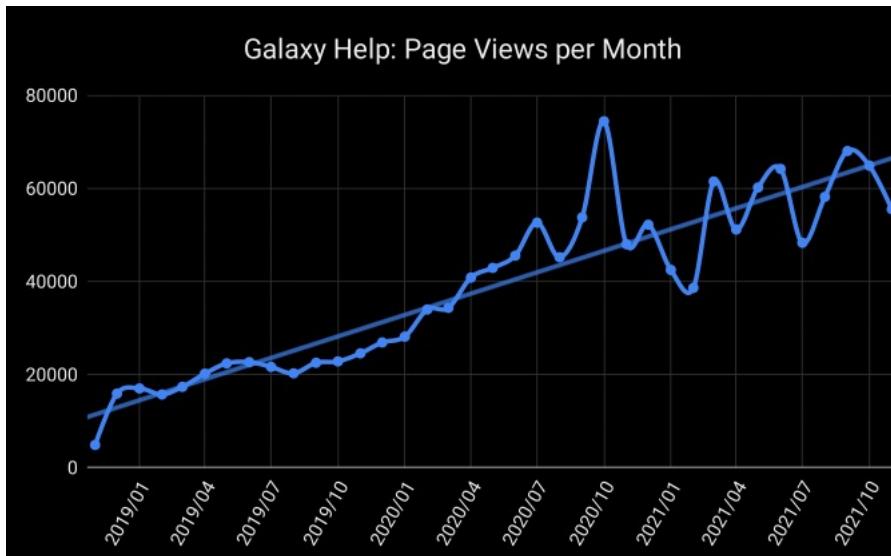
Galaxy is available as a free (for everyone) web server integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage:

<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

---

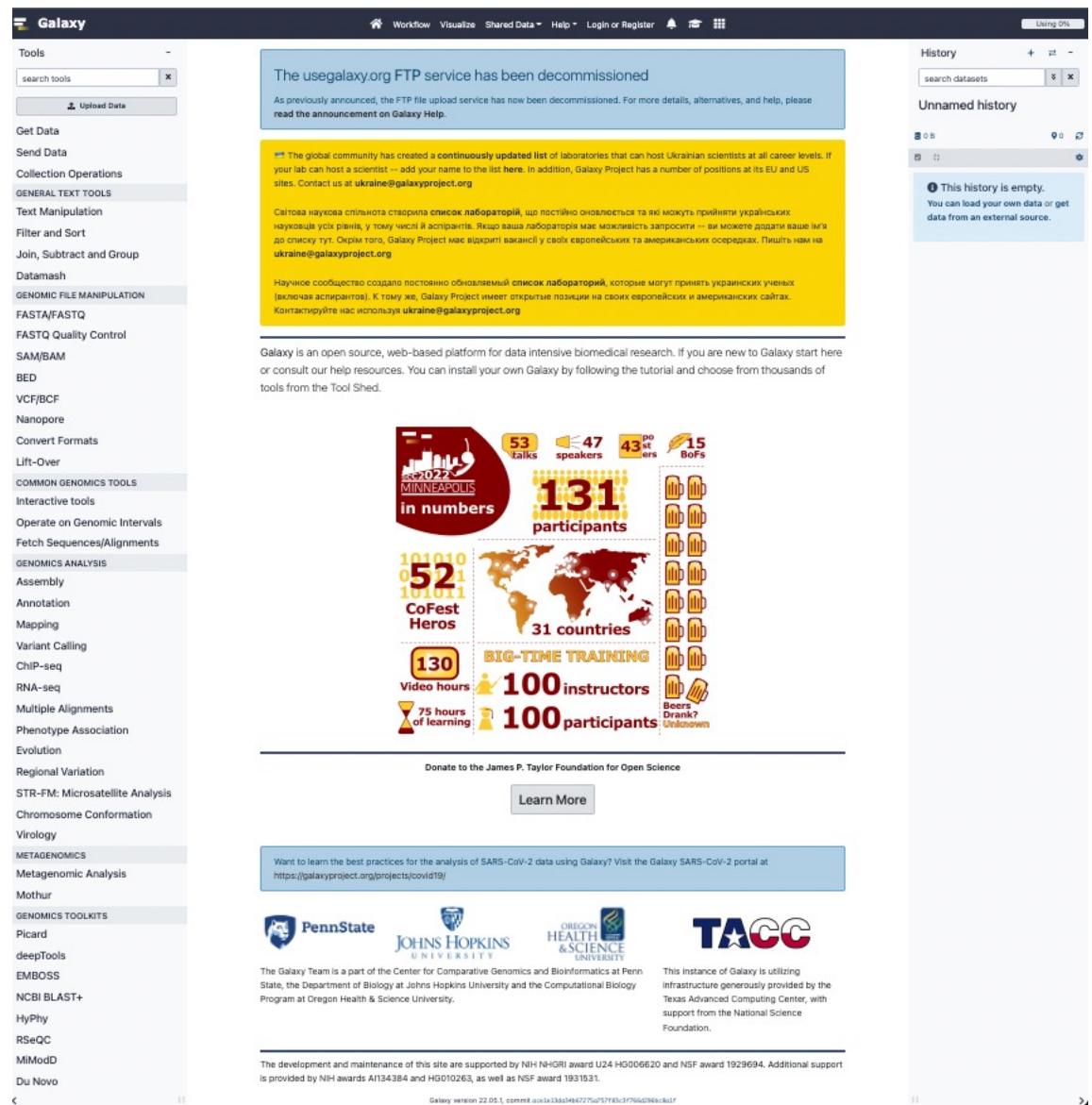
- Starting from the Quast report of the assemblies
- Genome annotation: structural and functional annotation
- Identification of gene clusters
- Protein characterization
- Summary of pipeline
- Class assessment submission

## 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](https://doi.org/10.1016/j.cels.2018.05.012)

# Quality assessment of raw data

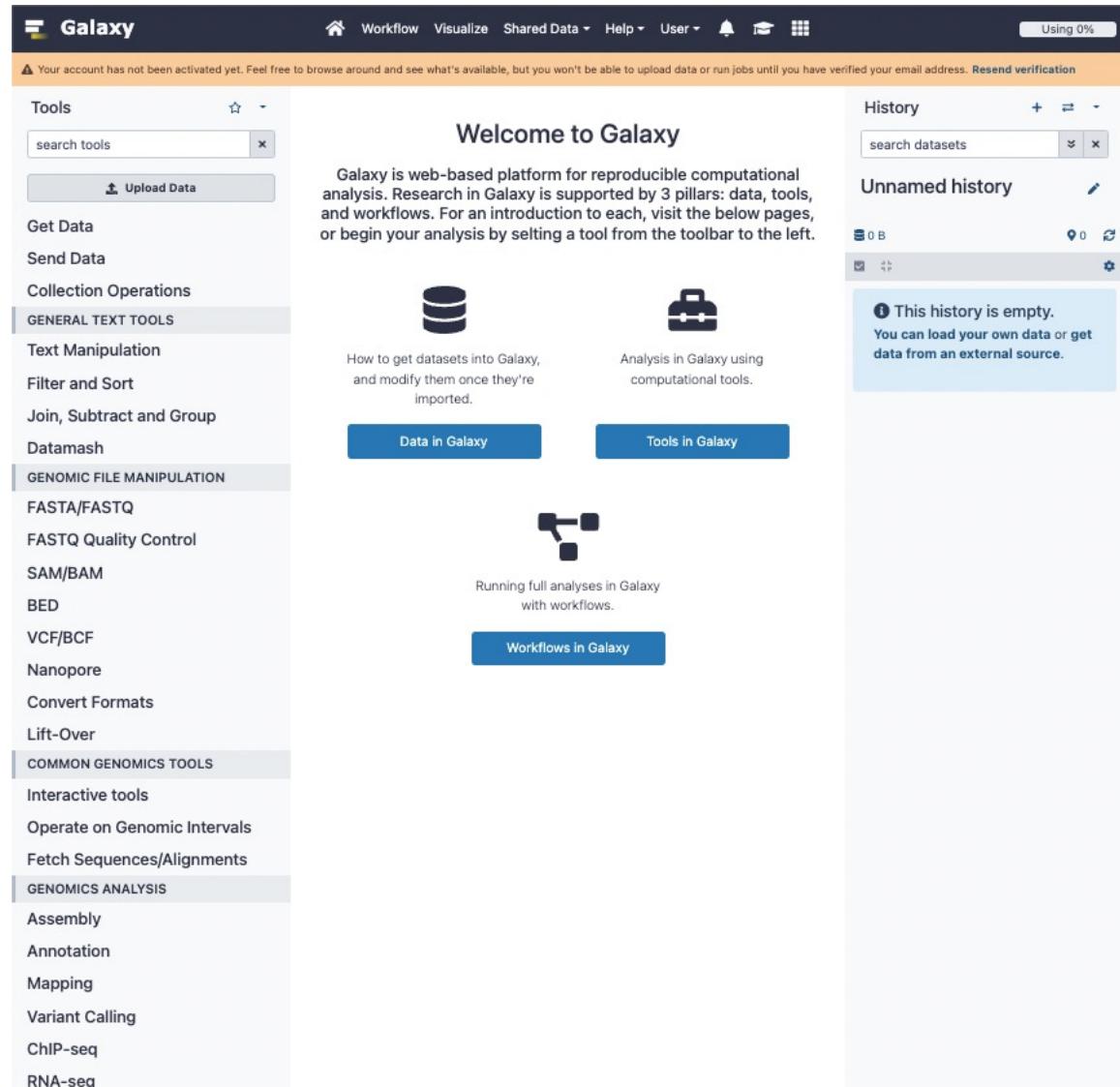
Login  
<http://usegalaxy.eu>



The screenshot shows the usegalaxy.org web interface. On the left, a sidebar lists various Galaxy tools categorized under General Text Tools, Genomic File Manipulation, Common Genomics Tools, Genomics Analysis, Metagenomics, Genomic Toolkits, and others. The main content area features a prominent yellow banner announcing the decommissioning of the FTP service. Below this, there's a section for Ukrainian scientists listing laboratories that can host them, followed by a general introduction to Galaxy. A central graphic titled "in numbers" provides a summary of recent activity: 53 talks, 47 speakers, 43 posters, 15 beers, 131 participants, 52 CoFest Heros, 31 countries, 130 video hours, 75 hours of learning, 100 instructors, and 100 participants. At the bottom, logos for Penn State, Johns Hopkins University, Oregon Health & Science University, and TACC are displayed, along with a note about infrastructure support.

# Quality assessment of raw data

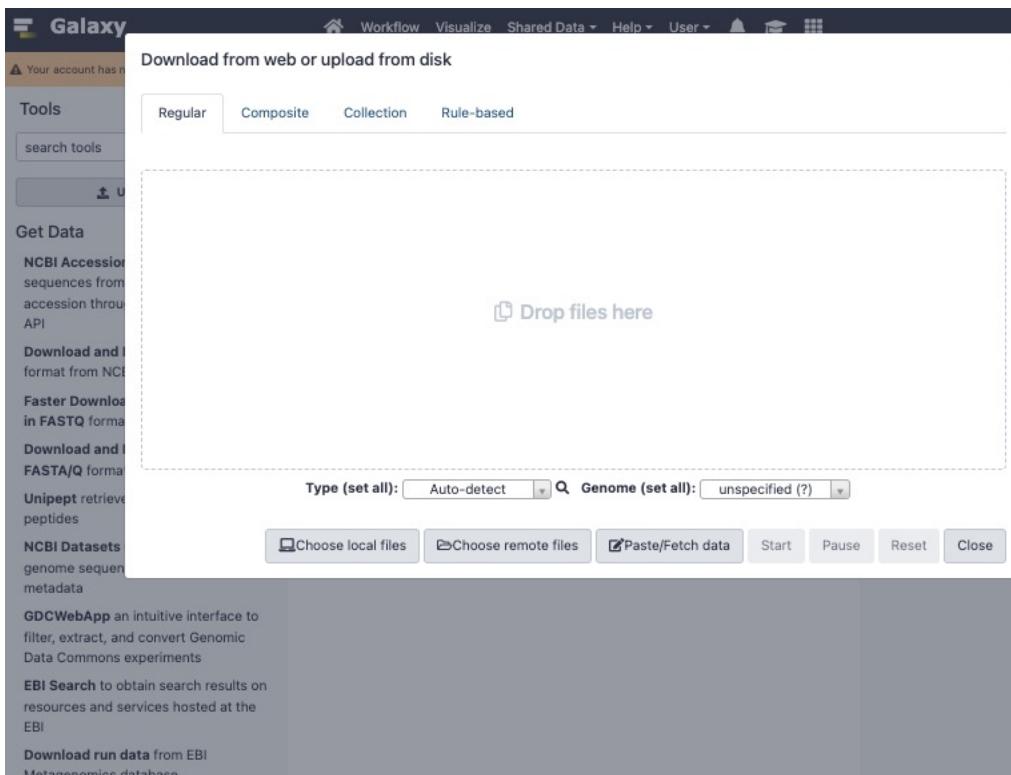
Login  
<http://usegalaxy.eu>



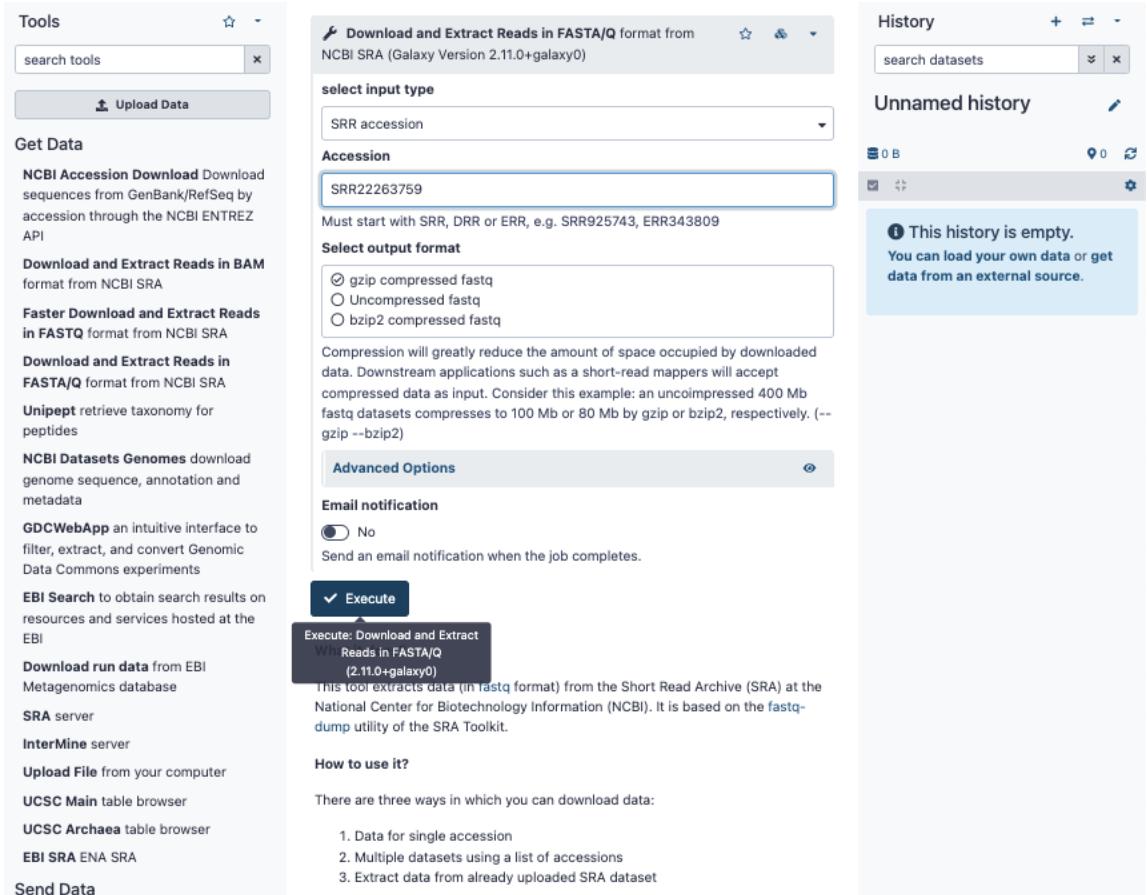
The screenshot shows the Galaxy web interface. At the top, there's a navigation bar with links for Home, Workflow, Visualize, Shared Data, Help, User, and a bell icon. A message at the top left says, "Your account has not been activated yet. Feel free to browse around and see what's available, but you won't be able to upload data or run jobs until you have verified your email address. Resend verification". On the left, a sidebar titled "Tools" lists various categories: Tools, Get Data, Send Data, Collection Operations, GENERAL TEXT TOOLS, Text Manipulation, Filter and Sort, Join, Subtract and Group, Datamash, GENOMIC FILE MANIPULATION, FASTA/FASTQ, FASTQ Quality Control, SAM/BAM, BED, VCF/BCF, Nanopore, Convert Formats, Lift-Over, COMMON GENOMICS TOOLS, Interactive tools, Operate on Genomic Intervals, Fetch Sequences/Alignments, GENOMICS ANALYSIS, Assembly, Annotation, Mapping, Variant Calling, ChIP-seq, and RNA-seq. The "GENERAL TEXT TOOLS" section is currently selected. In the center, a "Welcome to Galaxy" section explains the platform's purpose: "Galaxy is web-based platform for reproducible computational analysis. Research in Galaxy is supported by 3 pillars: data, tools, and workflows. For an introduction to each, visit the below pages, or begin your analysis by selecting a tool from the toolbar to the left." It features three cards: "Data in Galaxy" (database icon), "Tools in Galaxy" (briefcase icon), and "Workflows in Galaxy" (robot icon). On the right, a "History" panel shows an "Unnamed history" with 0 B and an empty dataset. A message in the history panel says, "This history is empty. You can load your own data or get data from an external source."

# Quality assessment of raw data

## Upload raw data



The screenshot shows the Galaxy web interface with the 'Tools' menu open. A specific tool, 'Download from web or upload from disk', is highlighted. This tool allows users to upload files directly from their computer or choose files from a local or remote source. It includes fields for specifying the type of data and the genome being used, along with various execution options like 'Start', 'Pause', and 'Reset'.



The screenshot shows the Galaxy web interface with the 'Tools' menu open. A specific tool, 'Download and Extract Reads in FASTA/Q format from NCBI SRA', is highlighted. This tool is used to download sequence data from the Short Read Archive (SRA) using an accession number. The configuration page shows the accession 'SRR22263759' entered. The 'Select output format' section offers options for gzip compressed, uncompressed, or bzip2 compressed FASTQ files, with 'gzip compressed fastq' currently selected. The 'Execute' button is prominently displayed at the bottom of the tool configuration. To the right, a 'History' panel is visible, showing an empty history list.

# 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.

## 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
TGGGAATAGTAAATTATTTAGCTTAGTTAGCTTATAAGGCTTCTAACGCTATTTTAA
AGAAAGAGCTAACAAAAGCAAACATCAGCGGAATAAAAAGGTTGTGAGAAGGTTAAGT
TATGAAAGTTAAGGAGGAATGGTTGTGGACAGACTTAATTTACTAGAGAGTTAGCATTA
AACATAATCTTAGTAATCTGAATGAACGTTTTATTTTGTACAAAAGGGAAAGA
ACTTCGAAAGAGATTGCCGTATTTAGGTTGGCGTCACCAATGTTCAAGACTTTA
TTGTCATGTATAATAAGGGTTTTAGAAAGAAAGTTGTGGAGAAAGATAAGAAACA
TATATATATATAACTAATGACAAAAGTGAAATTGTTGCATATTAAGTAAGTAGTAGTCAT
CCAAGAGGATGGTTTTTATTTAAAAGGGTTGCTTGTAAATAAGTTATTTTCT
ATTGACTATTAACATATAGTGTATAGTATATTATATTAAAGGAGGCTTGATAATAA
TGAGTTAGAAGAGTTAGAAGAAATTGAAAGAGAGGAAGGGAAAGAAGCAGTTGGCAGG
CAATTGAGGATTCAAAGGAATGGGAAGAAATTAAATCGTAAAATTGGGAATTAAATTG
GAGGAATTGAAATGTATATTAAAGATTGTTAGAAGAAACAAGAAATTTATAT
TTTGATGCAGACATTAAATTAAGCTCGCACCTGTTATTACGAGTAACATCAGAACAG
ATACAAGATTAAATCGAGCTATTAGTGAGTATCGGGGAAGAGTAAATTCTTTGTA
GAGAAAGCGTATGACAGATATGTATATTCTTATTGCAAAAGGATTAGGAGGTCAAACC
ATGCATGATTATTAATCGCTAATGAGACAGGTTTACTATAAGAACCCAAGTAGTTT
AGTTTATAGGAGGAACAAAATGAAATATACCGTTCATTTAGTGGATTATGTTG
TGGGGATTCTCTCTTCCGGTTCACCGCTAACATTGATAAGCACTATCATTGGAGT
CTTATTCTCTTACCTAAGTGTAAAAACATTGTTATTAGCTTGGCTACTGATAT
TAAGTCTGCCAGAGATTCACTTCTCGTGAAGAAAGAAATCAATAGGAGGTAAT
AAAATGCTAAAGATCAATGTTGTGTTGGTCAAAGATACTTGTAAAGATTAT
GATAATGAGTTGTTTATTCTAAGGAAATGTCAGAGAAAAGTTGCAAGTAGTTGAT
AAAATACCTGAAATGAATGAAGGCTCAATTCTTTATTGATAATCTTCTTTACA
AATGAAAATTATTATTAATTCTAACACATAACTTGGAAATTCTTGTCTTAAAC
TATGTTAACAGAATGACTTACTTTGTTAAACCCCTGATATTACTGATTATCTGG
ATTTTGGACATCTAGAAGAATTAAAAAAGCTTGATTTAATAATAGTAAAGAGGGG
TATTAAAAATATGCATACTTCAAATAATTAAAGAAAATTTTATTAAATAATTGGTT
GGCTTTATGTATCTAGCAGGATTACTAAGAAATCATGGAGTAATAGCGGGGATCCTT
AATTCCGGGATTCTGCACTTTAATTAGTTCATATTACGATAAAAAGAAGATTTC
```

## 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";

## 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 gaatccccgg ggatccggtg attgatttag caagctttat gcttgcataac cgttttgtga
           61 aaaaatttttt aaaaataaaaa agggacacctc tagggccccca attaatttag taatataatc
           121 tattaaagggt cattcaaaag gtcatccacc ggatcaattt ccctgcgtcg gcaggctggg
           181 tgccaaagctc tcgggttaaca tcaaggcccg atcccttgag cccttgccct cccgcacgat
           241 gatcgccgtc tgatcgaaat ccagatccctt gacccgcagt tgcaaaacctt cactgatccg
           301 gctcacggta actgtatcccg tatttgcaat accagcgatc ggccacaga atgtatgtcac
           361 gctgaaaaatgc cggccatcg aatggttca tggtcgatcc catcgcaaa agggatgtat
           421 aatgttatcatca ccacccacta ttgcacacag tgccgttgat cgtgctatga tcgactgtat
           481 tcatcagccgg 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.