

# **An Introduction to Galaxy**

Rémi Marenco & Anne Pajon & Jing Su  
CRUK Cambridge Institute

# Today

- Morning session from 9:30 to 12:30
  - The Galaxy interface
    - Tutorial sections 1 & 2
  - Loading data
    - Tutorial section 3
  - Operations on Genomics intervals
    - Tutorial section 4
- Afternoon session from 13:30 to 16:30
  - Workflows
    - Tutorial section 5
  - Data visualisation
    - Tutorial section 6

# Today

- What you'll gain
  - How to navigate around Galaxy
  - How to import data
  - How to run bioinformatics tools on data
  - How to create pipelines (workflow)
  - How to visualise data
- ... but you won't gain
  - a PhD in Bioinformatics!

# What is Galaxy?

- Web-platform for bioinformatics analysis
- Availability
  - Public servers
    - Galaxy Main <http://usegalaxy.org/>
    - Other Public Accessible Galaxy Servers
      - <https://wiki.galaxyproject.org/PublicGalaxyServers>
  - On Amazon cloud
    - <https://wiki.galaxyproject.org/CloudMan>
  - In VirtualBox
    - <https://wiki.galaxyproject.org/Events/GCC2014/TrainingDay/VMs?highlight=%28virtualbox%29>
  - As a local installation at CRUK-CI
    - <http://galaxy.cruk.cam.ac.uk/>

# Which server to use?

- Compare and choose based on:
  - size of datasets, available storage, backup
  - data security
  - computational requirements
  - tools installed

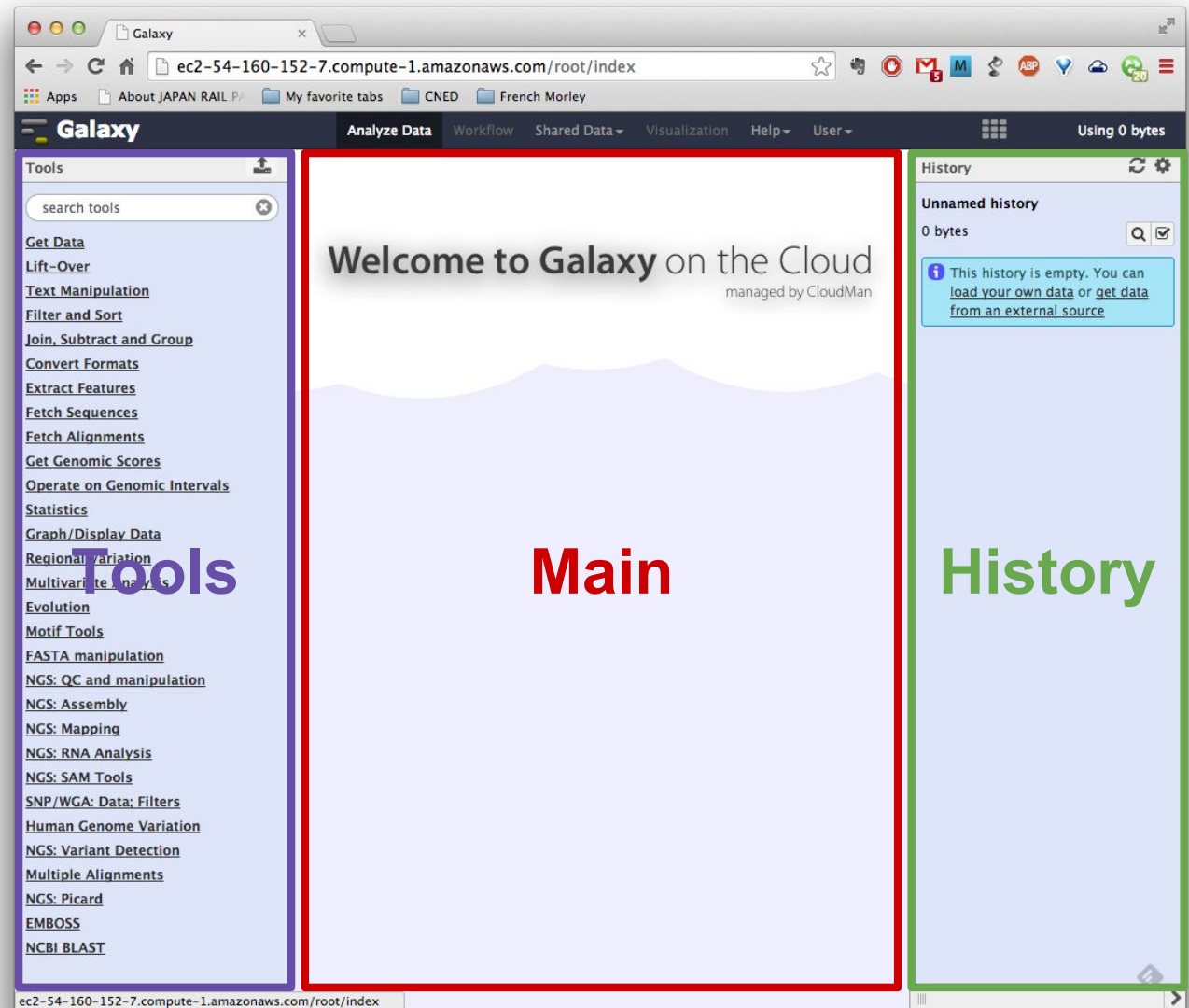
	Main	Local	Cloud	Appliance	Other
Your data sets are moderately sized	Yes	Yes	Yes	Yes	?
Your computational requirements are moderate	Yes	Yes	Yes	Yes	?
You want to share your Galaxy objects with others	Yes	Yes	Yes	Yes	?
All needed Tools are installed on Main.	Yes	?	Yes	Yes	?
Your data sets are very large	No	?	Yes	Yes	?
Your computational requirements are very large	No	?	Yes	Yes	?
You have absolute data security requirements	No	Yes	Yes	Yes	?
No network transfer of data	No	Yes	No	Yes	Yes

# Why Galaxy?

- Accessible
  - Users without programming experience can easily specify parameters, run tools, workflows and parse/filter data.
- Reproducible
  - Galaxy captures information so that any user can repeat and understand a complete computational analysis.
- Transparent
  - Users share and publish analyses via the web and create Pages and workflows - interactive, web-based documents that describe a complete analysis.

# The Galaxy interface

- Divided into 3 panels
  - Tools
  - History
  - Main



# The Menu



The Menu is where we look for other items in Galaxy.

- *Shared data* - your main source of data.
- *Workflow* - shows your workflows - editable diagrammatic pipeline (more later).
- *Visualization* - shows your visualisations.
- *User* – things specific to you (histories, datasets, pages, etc).



# Tools

- Many tools available
- Galaxy Toolshed <https://toolshed.g2.bx.psu.edu/>
- Need a tool that's not in Galaxy?
  - Ask a bioinformatician! At CRUK-CI, we can install available command-line tools into Galaxy and also develop custom software for you to use.



The screenshot shows the Galaxy web interface. At the top, there's a dark blue header with the 'Galaxy' logo. Below it, a light blue sidebar contains a 'Tools' section with a search bar labeled 'search tools'. A long list of tool categories is displayed, each with a blue underlined link. The categories include: Get Data, Lift-Over, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Graph/Display Data, Regional Variation, Multivariate Analysis, Evolution, Motif Tools, FASTA manipulation, NGS: QC and manipulation, NGS: Assembly, NGS: Mapping, NGS: RNA Analysis, NGS: SAM Tools, SNP/WGA: Data; Filters, Human Genome Variation, NGS: Variant Detection, Multiple Alignments, NGS: Picard, and EMBOS.

# Tutorial

- Start tutorial – Introduction to Galaxy
  - Go through Sections 1 and 2
    - The Galaxy interface
    - Getting started

<http://tinyurl.com/GalaxyCamPractical>

# Loading data

## Section 3

# Importing data

- Copy/paste from a file
- Upload data from a local Computer
- Upload data from internet
- Upload data from database queries
  - UCSC, BioMart, ENCODE, modENCODE, Flymine etc.
- Download shared data from public libraries or shared:
- Data libraries, Histories, Workflows, Visualizations, Pages
- Upload data from FTP (>2GB)
  
- Be aware of the data attribute: Datatype and genome assembly

# Genome build and Data types

- Genome build specifies which genome assembly this dataset is associated with. e.g. mm9, hg19.
- Genome build can be detected and assigned or user specified.
- User can define their own custom genome build.
- New genome assembly can be added by the galaxy server/instance admin.
- Data type can be detected and assigned or user specified. e.g.
  - Edit Attributes
  - Convert Formats
- Data type is also assigned by tools when output is created.
- Many tools will only accept as input datasets with the appropriate data types assigned.
- New genome assembly and data type can be added by the galaxy server/instance admin.

## Convert Formats

Tabular-to-FASTA converts tabular file to FASTA format

FASTA-to-Tabular converter

FASTQ to FASTA converter

AXT to concatenated FASTA

Converts an AXT formatted file to a concatenated FASTA alignment

AXT to FASTA Converts an AXT formatted file to FASTA format

AXT to LAV Converts an AXT formatted file to LAV format

BED-to-GFF converter

GFF-to-BED converter

LAV to BED Converts a LAV formatted file to BED format

MAF to BED Converts a MAF formatted file to the BED format

MAF to Interval Converts a MAF formatted file to the Interval format

MAF to FASTA Converts a MAF formatted file to FASTA format

Wiggle-to-Interval converter

SFF converter

GTF-to-BEDGraph converter

Wig/BedGraph-to-bigWig converter

BED-to-bigBed converter

# Data types

- Sequence files:
  - Ab1, Fasta, Scf
- Sequencing files:
  - FASTQ, FastqSolexa, sff
- Alignment files:
  - Axt, SAM/BAM, MAF, LAV
- Intervals:
  - Bed (0 based), GFF, GTF(GFF2), GFF3
- Graph:
  - BedGraph, WIG/BigWIG (1 based)
- Variant files
  - VCF/BCF
- Tabular and Text

1. Ab1
2. Axt
3. BAM
4. Bed
5. BedGraph
6. BCF
7. Fasta
8. Fastq
9. FastqSolexa
10. GFF
11. GTF
12. GFF3
13. Interval
14. Lav
15. MAF
16. SAM
17. Scf
18. Sff
19. Tabular (tab delimited)
20. VCF
21. Wig and bigWig
22. Plain text

# Tutorial

- Start tutorial – Introduction to Galaxy
  - Go through Section 3
    - Loading data

<http://tinyurl.com/GalaxyCamPractical>

# Interval Operations

## Section 4



# Operations on Genomic intervals

- Join
- Intersect
- Get flanks
- Coverage
- Complement
- Cluster
- Base Coverage
- Subtract
- Fetch closes non-overlapping feature
- Merge
- Arithmetic Operations on tables
- Subtract Whole Dataset
- Concatenate
- Converting into interval format

# Operations on Genomic intervals

## A Intersect



## B Subtract



# Operations on Genomic intervals

## C Merge



## D Concatenate

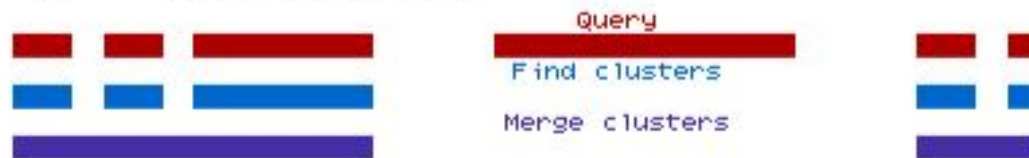


# Operations on Genomic intervals

## E Complement



## F Cluster



# Operations on Genomic intervals

- Join
- Intersect
- Get flanks
- Coverage (merge)
- Complement
- Cluster
- Base Coverage (intersect)
- Subtract
- Fetch closes non-overlapping feature
- Merge
- Arithmetic Operations on tables
- Subtract Whole Dataset
- Concatenate
- Converting into interval format

# Operations on Genomic intervals

Input						
Query 1:						
chr1	10	100	Query1.1			
chr1	500	1000	Query1.2			
chr1	1100	1250	Query1.3			
Query 2:						
				chr1	20	80 Query2.1
				chr1	2000	2204 Query2.2
				chr1	2500	3000 Query2.3
Output						
(Return only records that are joined)						
chr1	10	100	Query1.1	chr1	20	80 Query2.1
(Return all records of first query)						
chr1	10	100	Query1.1	chr1	20	80 Query2.1
chr1	500	1000	Query1.2	.	.	.
chr1	1100	1250	Query1.3	.	.	.
(Return all records of second query)						
chr1	10	100	Query1.1	chr1	20	80 Query2.1
.	.	.	.	chr1	2000	2204 Query2.2
.	.	.	.	chr1	2500	3000 Query2.3
(Return all records of both queries)						
chr1	10	100	Query1.1	chr1	20	80 Query2.1
chr1	500	1000	Query1.2	.	.	.
chr1	1100	1250	Query1.3	.	.	.
.	.	.	.	chr1	2000	2200 Query2.2
.	.	.	.	chr1	2500	3000 Query2.3

# Tutorial

- Start tutorial – Introduction to Galaxy
  - Go through Section 4
    - Interval Operations
      - Basic operations
      - Identify promoter regions containing TAF1 binding sites
      - Finding coding exons with highest SNP density

<http://tinyurl.com/GalaxyCamPractical>

# Workflows

## Section 5



# Understanding Histories

- In Galaxy your analyses live in histories
- Histories can be very large, you can have as many histories as you want, and all history behavior is controlled by the History options on the top of the History pane
- Many of the options here are self explanatory. If you create a new history, your current history does not disappear.
- If you would like to list all of your histories just choose Saved Histories and you will see a list of all your histories in the center pane.

## HISTORY LISTS

Saved Histories

Histories Shared with Me

## CURRENT HISTORY

Create New

Copy History

Copy Datasets

Share or Publish

Extract Workflow

Dataset Security

Resume Paused Jobs

Collapse Expanded Datasets

Include Deleted Datasets

Include Hidden Datasets

Unhide Hidden Datasets

Delete Hidden Datasets

Purge Deleted Datasets

Show Structure

Export Citations

Export to File

Delete

Delete Permanently

## OTHER ACTIONS

Import from File

# Workflows

- Converting histories into workflows
  - One of the history options listed is very special. It allows you to easily convert existing histories into analysis workflows.
  - Why would you want to create a workflows out of a history? To redo the analysis again with minimal clicking.

## HISTORY LISTS

Saved Histories

Histories Shared with Me

## CURRENT HISTORY

Create New

Copy History

Copy Datasets

Share or Publish

Extract Workflow

Dataset Security

Resume Paused Jobs

Collapse Expanded Datasets

Include Deleted Datasets

Include Hidden Datasets

Unhide Hidden Datasets

Delete Hidden Datasets

Purge Deleted Datasets

Show Structure

Export Citations

Export to File

Delete

Delete Permanently

## OTHER ACTIONS

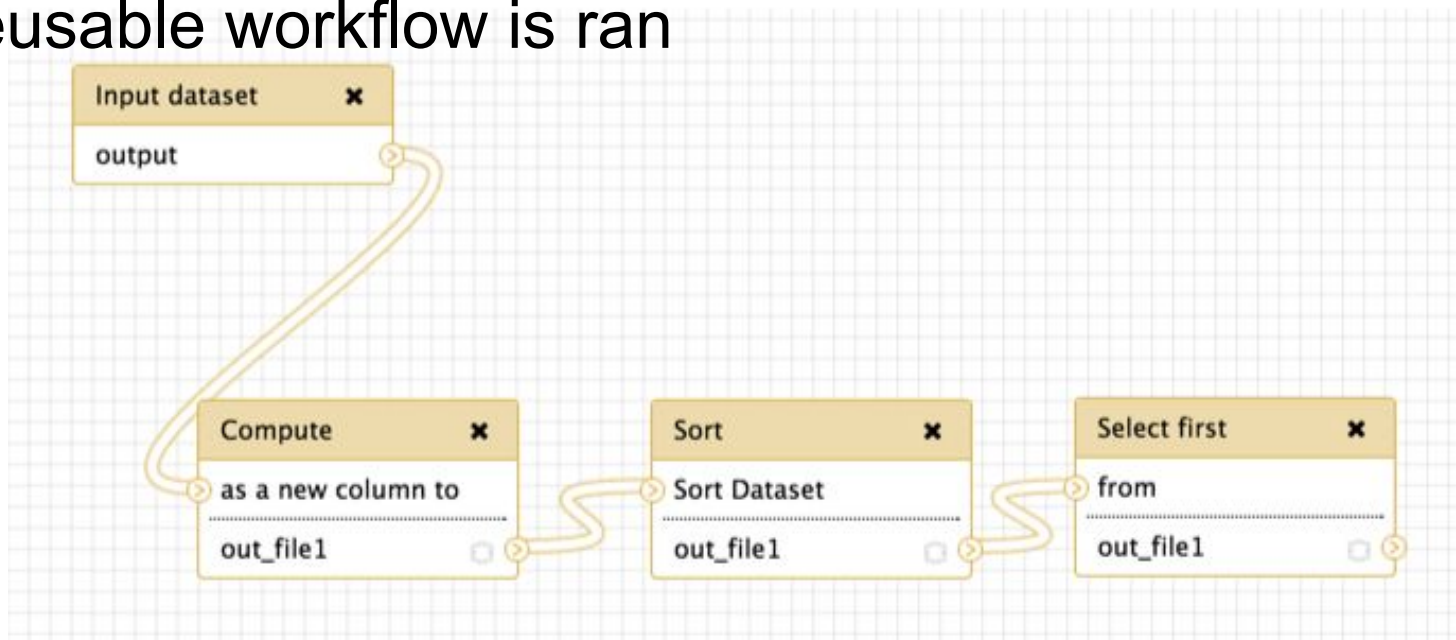
Import from File

# Building workflows

- Workflow allows analysis containing multiple tools to be built, run, extracted from histories, and rerun.
- Can be built
  - manually by adding tools on the workflow canvas,
  - from an existing history, or
  - by importing an existing workflow.
- Collaborations, publications, pipelines.

# Building workflows

- Publish and share objects: Dataset, history, workflow, Page, visualization
- Workflow can be created: from existing project, or from scratch, or by downloading from a publicly accessible workflow
- The parameters can be modified before each time the reusable workflow is ran



# Running workflows

- All the jobs are submitted to Galaxy
- Each job is run in turn
- Each job waits for output of previous tool
- Workflow can be run over and over again on any suitable datasets
- Workflow can be published, shared or downloaded from [https://usegalaxy.org/workflow/list\\_published](https://usegalaxy.org/workflow/list_published) [menu: 'Shared Data' > 'Published Workflows']

# Tutorial

- Start tutorial – Introduction to Galaxy
  - Go through Section 5
    - Workflows
      - Create workflow from history
      - Create a new workflow
      - Importing a workflow

<http://tinyurl.com/GalaxyCamPractical>

# Visualisation

## Section 6

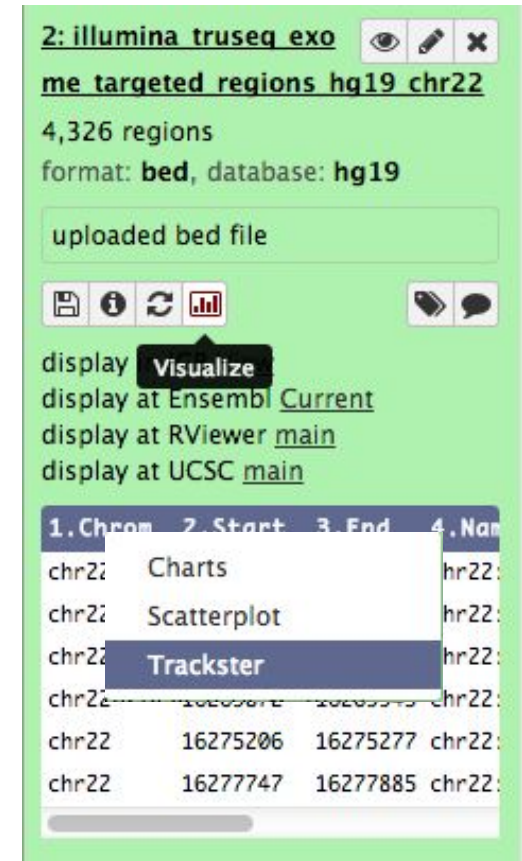
# Data sharing

- You can share your Galaxy items - histories, workflows, visualizations, and pages - with other people in three different ways:
  - Individual users: directly using a Galaxy account's email addresses on the same instance
  - One or more users: using a web link, with anyone who knows the link
  - Everyone: using a web link plus publishing into Shared Data
- Galaxy Toolshed <https://toolshed.g2.bx.psu.edu/>



# Biological visualisation

- Galaxy incorporates a track browser called Trackster.
  - This can be used to visualize genomic data within Galaxy in a tightly integrated way.
  - The browser also currently supports (and aims to support maximally) visual analytics, where visualization is used iteratively to provide feedback on analysis.



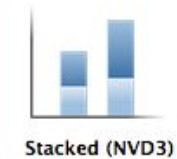
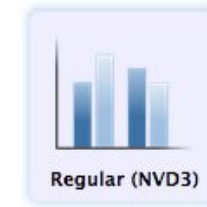
# Numerical visualisation

- Additionally, Galaxy enables you to create bar diagrams, pie charts, scatter plots and other visualisations using the Charts plugin.

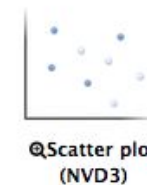
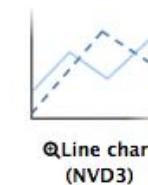
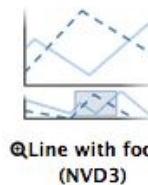
The screenshot shows the Galaxy web interface. At the top, a file named `1: http://www.compsysb.io.org/bacteriome/dataset/functiona l interactions.txt` is listed with 3,989 lines and a tabular format. Below the file list, there is a section for visualizing the data. A dropdown menu is open, showing options: **Charts**, **Scatterplot**, and **Trackster**. The **Charts** option is selected. The data table below shows columns of IDs and values.

1	2	Visualize
B1882	B1888	1.000000
B0728	B0729	
B1812	B3360	
B4200	B4202	
B0779	B4058	
B0032	B0033	0.933183

## • Bar diagrams



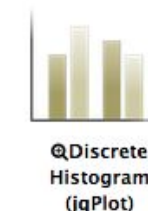
## • Others



## • Area charts



## • Data processing (requires 'charts' tool from Toolshed)



# Tutorial

- Start tutorial – Introduction to Galaxy
  - Go through Section 6
    - Data visualisation
      - Biological visualisation
      - Numerical visualisation

<http://tinyurl.com/GalaxyCamPractical>

# Thank you!

- Questions
- Please feel free to give us any feedback on this form <http://tinyurl.com/galaxy-feedback>

# Acknowledgements

- Graham Etherington, Sainsbury Laboratory  
Norwich
  - 'An Introduction to Galaxy' <http://tsltraining.tsl.ac.uk/>
- Galaxy Team
  - 'Galaxy 101' <https://usegalaxy.org/u/aun1/p/galaxy101>
- CloudMan
  - <https://wiki.galaxyproject.org/CloudMan>