Galaxy 101

EMC Galaxy Course

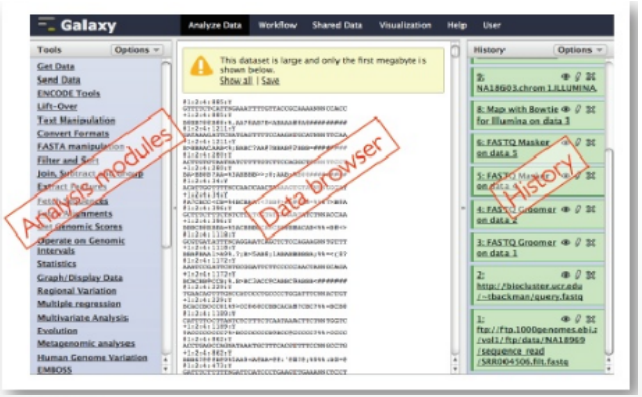
November 24-25, 2014

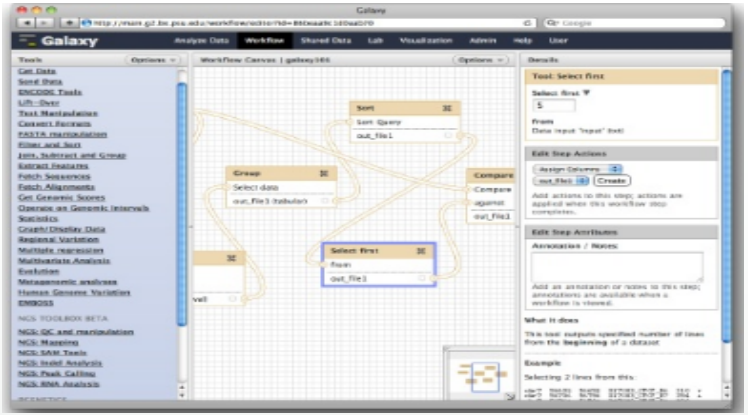
Youri Hoogstrate, David van Zessen, Saskia Hiltemann
Guido Jenster, Andrew Stubbs

What is Galaxy?

Galaxy is an open, web-based platform for data intensive biomedical research.

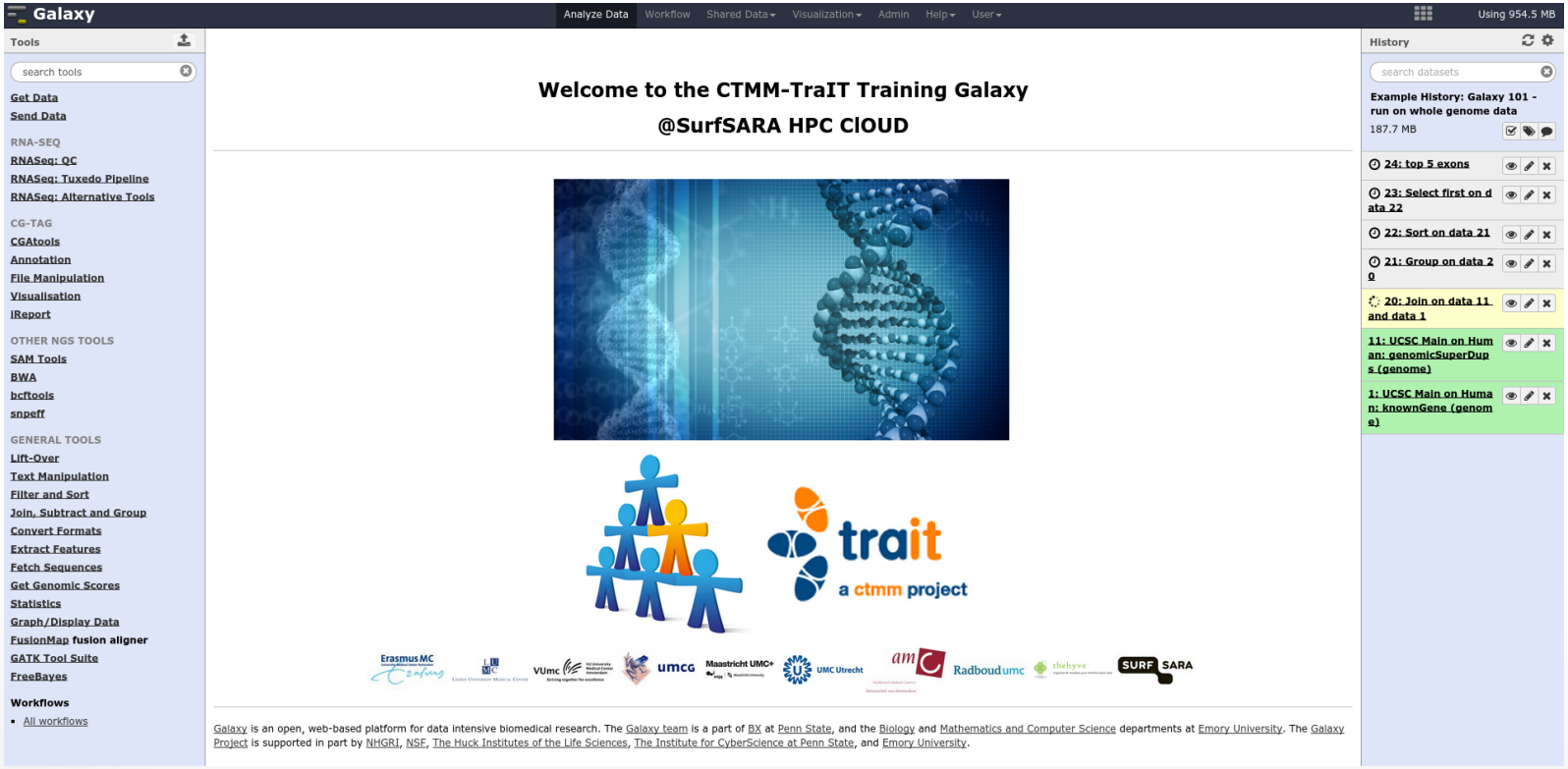
- Provides graphical user interface for command-line programs
- User can build **workflows** graphically
- Analysis steps performed on datasets stored in the ***History***
- **Sharing** of datasets and workflows amongst users
- Free and open-source, large user/developer community





What is Galaxy

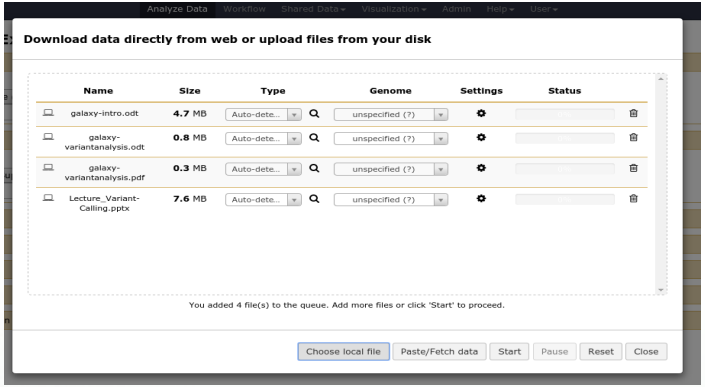
Tool list | view tools/data | History



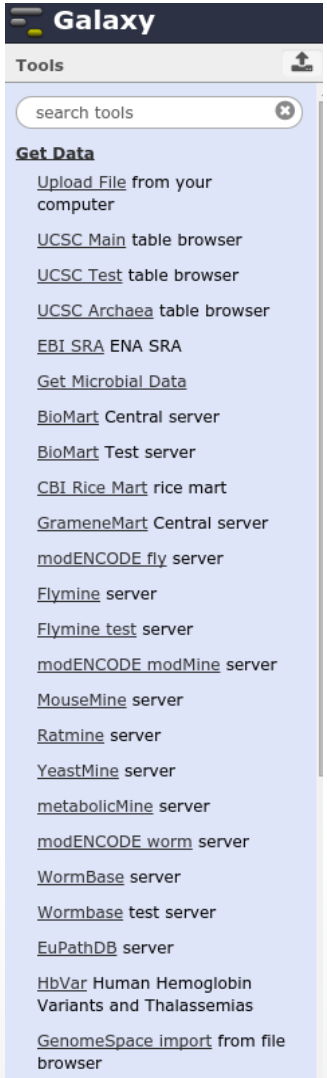
Getting Data

- **Upload** from your machine or supply URL

- **Batch upload** 
  - supports drag and drop
  - supply multiple files/URLs at once



- Get Data from **external sources** (UCSC, EBI, BioMart, ...)



Histories

Galaxy keeps track of all analysis steps taken in a so-called *history*.
Users can have as many histories as they like and switch between them easily

Items coloured by status:

- **Green**: completed successfully
- **Yellow**: running
- **Grey**: queued
- **Red**: completed with error

Buttons:

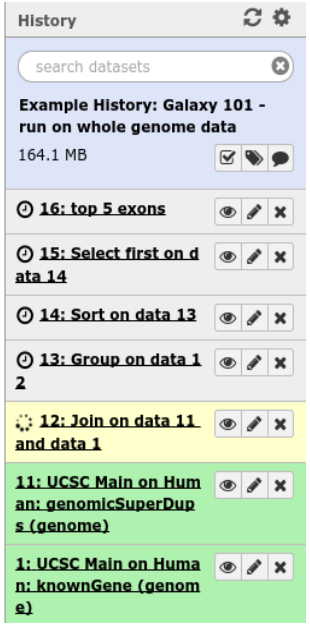
- Eye: view contents of file
- Pencil: edit attributes (name, reference genome, file type, convert format, etc..)

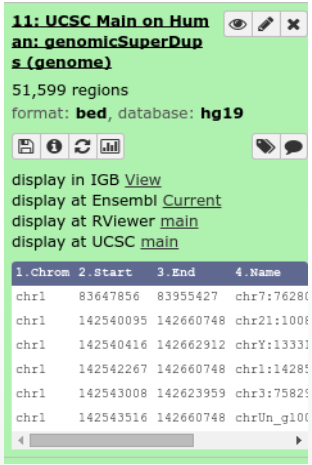
Expand history item for more options

- Download
- View metadata
- View output/error logs
- Visualisation options
- Rerun tool

Metadata is recorded:

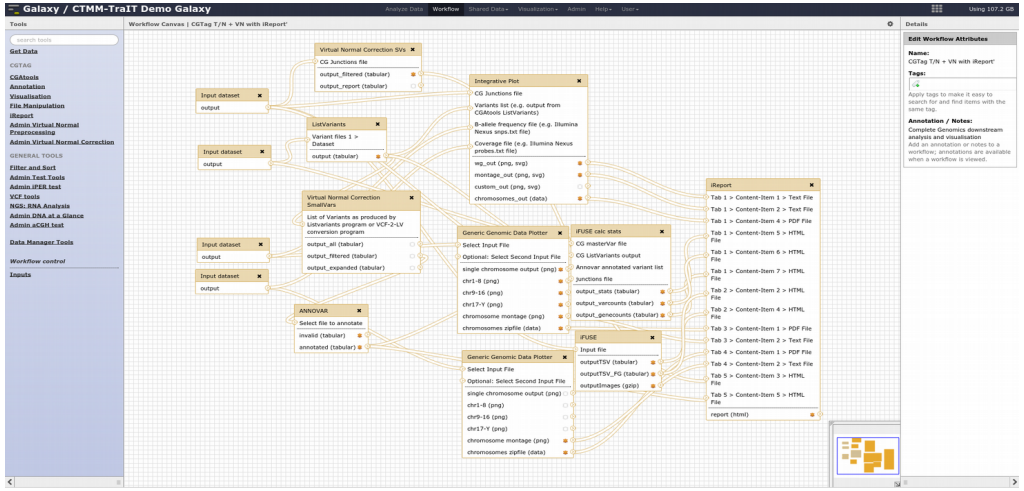
- Tool versions
- Reference genome
- Full parameter settings
- Input files used





Workflows

Galaxy lets you easily create pipelines in the **workflow editor**

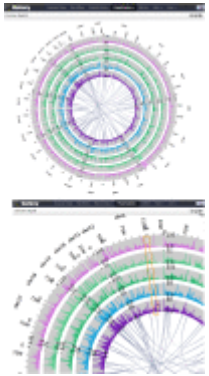


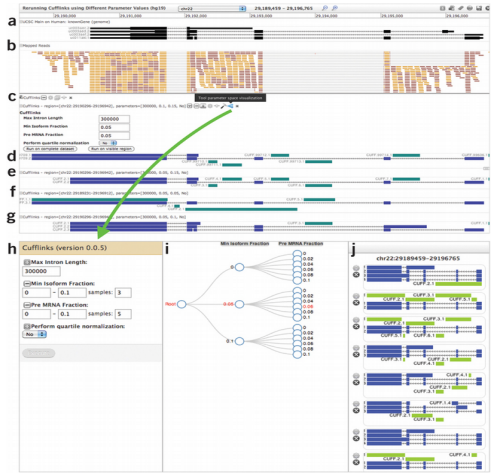
You can also **extract workflow from your history**

→ Perform analysis manually once, then extract workflow to make it easier next time (just supply input files and hit *Execute*)

Visualisations

Galaxy has many visualisation options:





Built-in Features:

- Genome browser: **Trackster**
- Circos-like views with **Circster**
- Custom plots with **Galaxy Charts**

Links to external display applications

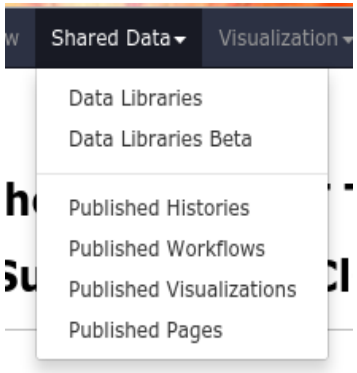
- UCSC
- IGV
- IGB
- Ensemble
- Rviewer
- ..

iReport

Create your own interactive HTML reports in Galaxy (more about this later)

Sharing

Data Libraries
Admins use data libraries to share often-used data with all users.



Sharing histories, workflows, visualisations
Users can share their work by:

- Sharing with specific Galaxy users
- Make accessible via link. Anybody with this link can see your shared data (even without an account)
- Publishing on Galaxy (will appear in shared data menu)

Published Pages
To describe an analysis or tool to others, users can make a page. This is made within galaxy in a simple text editor. Histories, workflows and visualisations can be included directly in the page

CGtag: Complete Genomics Analysis Toolkit and Annotation in Galaxy

Here we describe several of the tools contained in the CGtag toolkit by way of example. Example data is provided in a shared data library, or can be imported to your history directly from this page.

Example Data

Complete Genomics provides free public access to a variety of whole human genome data sets generated from Complete Genomics' sequencing service (link). Here we will be using the breast cancer cell line HCC1187.

Data is also available from the CGtag data library which can be accessed by going to "Shared Data" at the top of your screen.

Alternatively, here is an example history containing all the shared files, which can be copied to your own history:

Galaxy History | CGTag: Input Data

CGTag: Input Data

0 bytes

search datasets

Dataset	Annotation
1: HCC1187 Normal highConfidenceJunctions hg19.tsv	
2: HCC1187 Tumour highConfidenceJunctions hg19.tsv	
3: HCC1187 Normal varfile hg19	
4: HCC1187 Tumour varfile hg19	
5: VCaP CNV Details Diploid	

Small Variants (SNPs and small insertion, deletions and substitutions) are reported in the CG varfile format. SV breakpoint (junctions) are reported in the CG junctions file.

CGAtools

Complete Genomics provides a suite of command-line tools, cgatools, for downstream analysis of their data.

Let us perform a Tumour-Normal Comparison of the small variants and the SV.

For SVs we use the cgatool "jumctionDiff". This tool output those junctions present in the first input file, but not in the second.

Galaxy History | CGtag: SV analysis

For small variants we will use the ListVariants and TestVariants tools to perform the comparison. Listvariants takes as input a varfile, and will output a list of all the variants found in the sample. This can then be compared to an arbitrary number of varfiles using the TestVariants tool. The output will consist of all variants in the variantlist, with an additional column per varfile indicating whether that variant was present in the other sample or not. In our case we are interested in the variants present in the tumour, and wish to compare these to the variants found in the normal sample. We will therefore use the tumour varfile as input to ListVariants, and then compare the resulting list of variant to the varfile of the normal sample with the TestVariants tool:

Galaxy History | CGtag: small variants

Notice that if we had selected the wrong build (as we did in dataset 3), the tool would have let us know and the job would have failed.

We could also have performed some preprocessing on the varfile before we ran the listvariants tool. This can be done on the varfile using the cgatools VarFilter.

Annotation

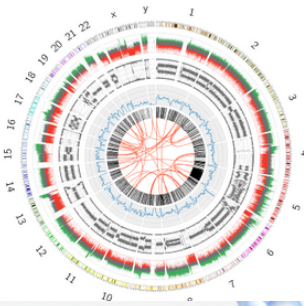
There are several annotation tools available. We start by running ANNOVAR on the result of our T/N comparison we did in the previous section. Because our input is neither a VCF file nor a varfile, we select the option "other" and specify which columns contain the necessary information for ANNOVAR. We next select the annotation we would like to perform and hit excecute. Note that for a file of this size (some 3,5 million variants), the execution may take quite some time, so any prefiltering that can be done will save a lot of time at this step.

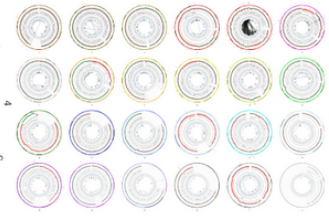
Galaxy History | CGtag: Annotation

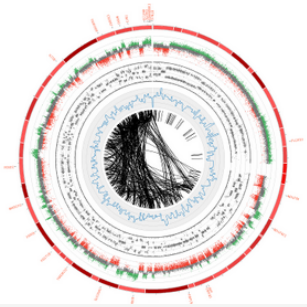
We can now run condel to obtain the consensus deliteriousness score from the SIFT and polyphen2 annotations we got from ANNOVAR. MutationAssessor may also be run to get more information about SNPs in exonic regions. Because this connects to an online service, the queries are limited to 3 per second, so filtering your list of variants here is highly recommended.

Visualisation

Integrated Circos plot can visualize SVs and small variants from Complete Genomics files, and B-allelefrequency and coverage data from Illumina SNP array data processed by Nexus. Output can be either a whole-genome plot, a zip file containing all individual chromosome plots, a single image containing all single-chromosome plots, and/or a cutom defined region. Below is example of the output for the VCaP sample, all necessary data for these plots can be found in the CGtag shared data library. On the left is the whole-genome plot, in the middle the montage of the chromosome plots, and on the right the indivudual plot for chromosome 5. These images clearly shows the chromothripsis present on chromosome 5.







Author
saskia-hiltemann

Related Pages
All published pages
Published pages by saskia-hiltemann

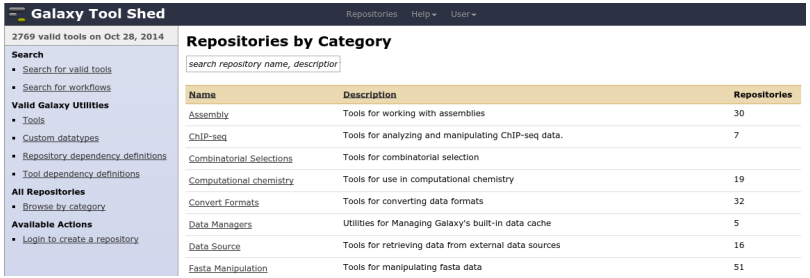
Rating
Community
(0 ratings, 0.0 average)
Yours

Tags
Community: none
Yours:

Advanced Topics

Tool Shed: The Galaxy App Store

- Anybody can add tools to Galaxy and make them available to the world via the Tool Shed



- ~ 3000 tools and counting in main tool shed
- Very easy to install to other Galaxy instances

API: Programmatic (automated) access to Galaxy

- Example: Pathology department connected their IonTorrent to Galaxy. Whenever a sequence run is finished, files are sent to Galaxy, workflow is run, results are pulled back, all automatically.

Servers

Training: galaxy-training1.trait-ctmm.cloudlet.sara.nl
galaxy-training2.trait-ctmm.cloudlet.sara.nl
galaxy-training3.trait-ctmm.cloudlet.sara.nl
...

(these servers will stay up until the end of the week, after that everything is erased)

Penn State Galaxy: https://usegalaxy.org/

EMC in-house: http://bioinf-galaxy4 (only accessible from within EMC)

CTMM TraIT demo Galaxy: http://galaxy.ctmm-trait.nl

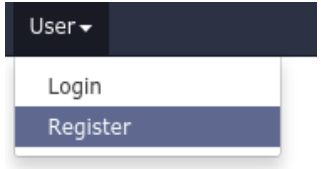
You can use any of these, but there may be limitation in resources or available tools. If you want to do large analyses in Galaxy, please contact us.

Practical Session

Learn by doing it yourself!

Servers: galaxy-training1.trait-ctmm.cloudlet.sara.nl
galaxy-training2.trait-ctmm.cloudlet.sara.nl
galaxy-training3.trait-ctmm.cloudlet.sara.nl
..

Register for an account



All handouts and slides can be found under Shared Data → Data Libraries

Manual: [Course Manual] EMC Galaxy Training 1: Introduction to Galaxy.pdf