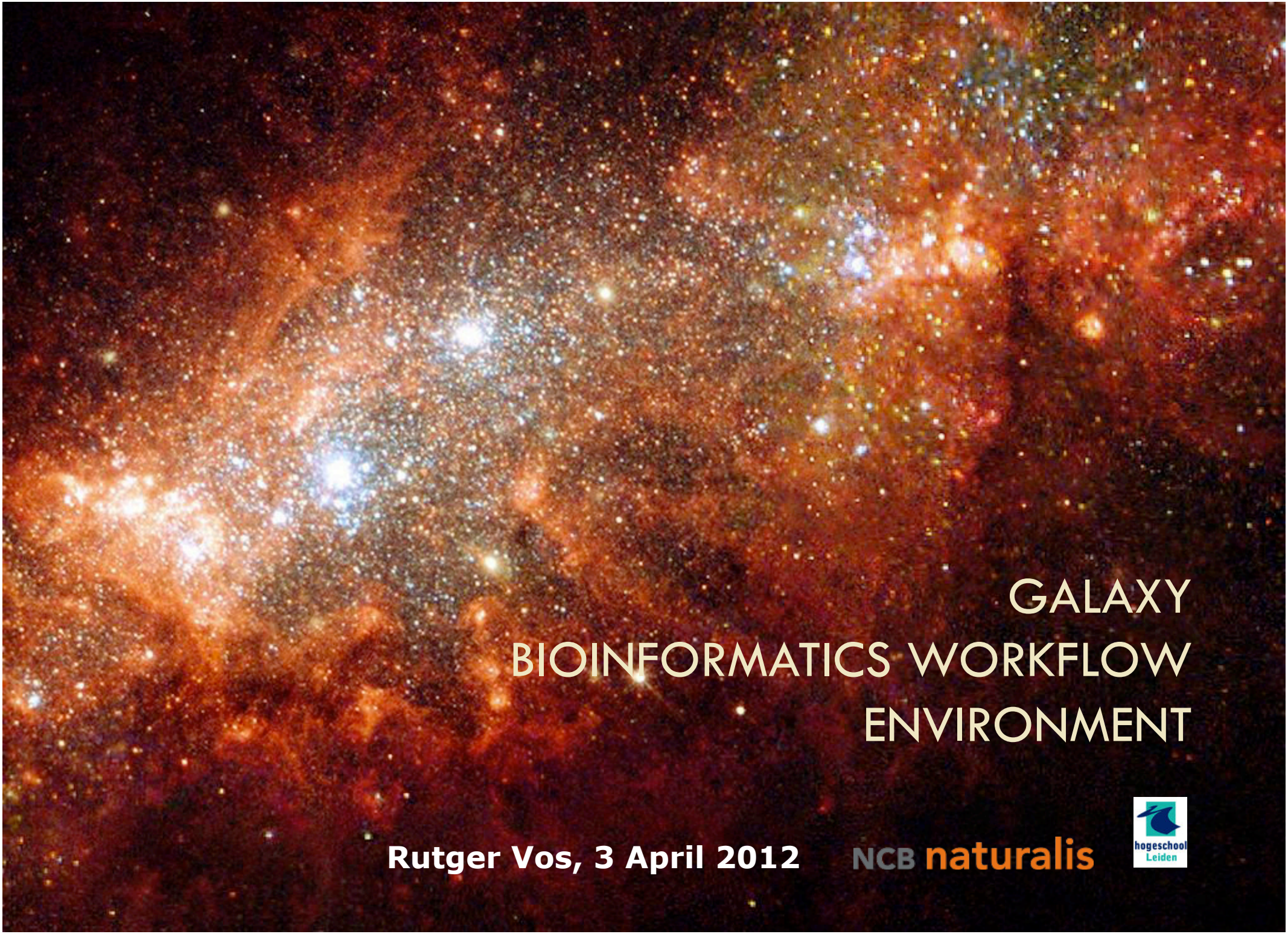


# GALAXY BIOINFORMATICS WORKFLOW ENVIRONMENT

**Rutger Vos, 3 April 2012**

NCB naturalis

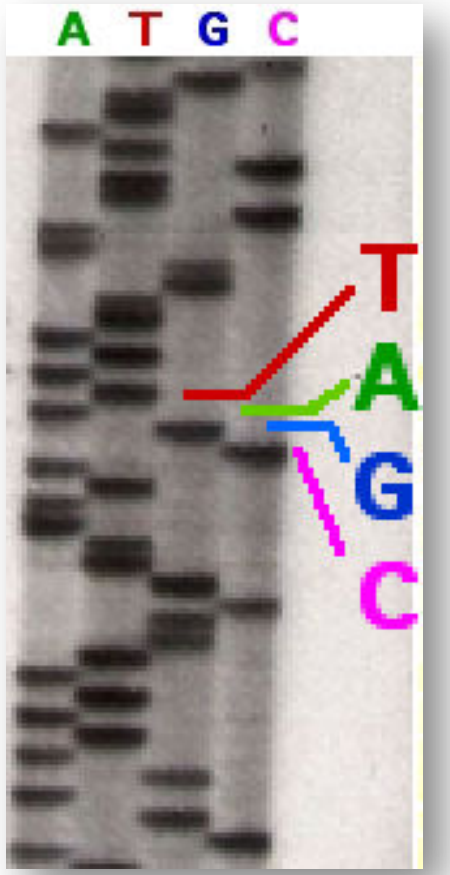
hogeschool Leiden

# Overview

Informatics in the post-genomic era

# The past (?)

- Analyses glued together using scripting languages, directly on the CLI or in GUI
- Sanger sequencing
- Smaller data volumes
- Fewer remote data resources
- Hypothesis-driven



# The present

- Graphical or text-based workflow tools
- “Next generation” sequencing
- Large data sets
- Many remote data resources
- “Data-driven”

# NGS – Roche 454 pyrosequencing

**Pyrosequencing**

- "Emulsion PCR"
- Bead with primer in each droplet
- Each bead is placed in a well with luciferase
- Plate is analyzed by fiber-optic chip

**Genome Sequencer FLX**

# NGS – Illumina/Solexa

**Reversible dye-terminator seq**

- DNA attaches to primer on slide and is amplified
- 4 RT-bases are added
- Camera detects labeled nucleotides
- Next 1-base cycle

**HiSeq2000 (BGI)**

# NGS – IonTorrent

**Ion semiconductor sequencing**

- “sequencing by synthesis”
- Not light-based, sensor detects H+ ions during synthesis
- Longer reads

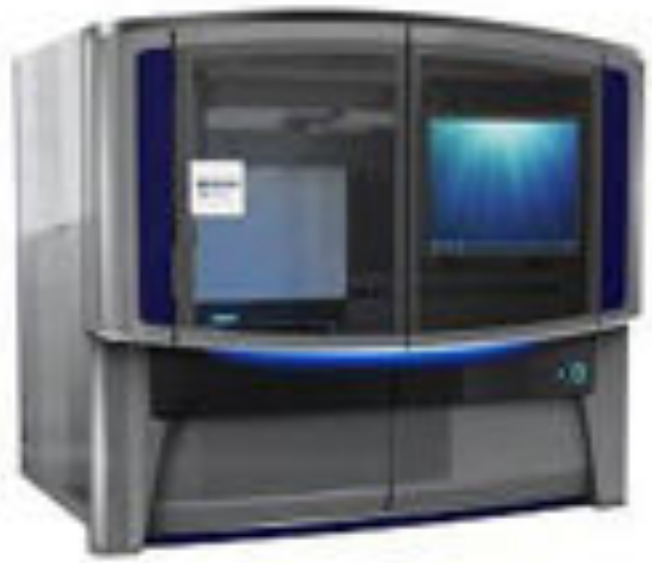
**Ion Torrent PGM**

# NGS – SOLiD Sequencing

**Oligonucleotide ligation and detection**

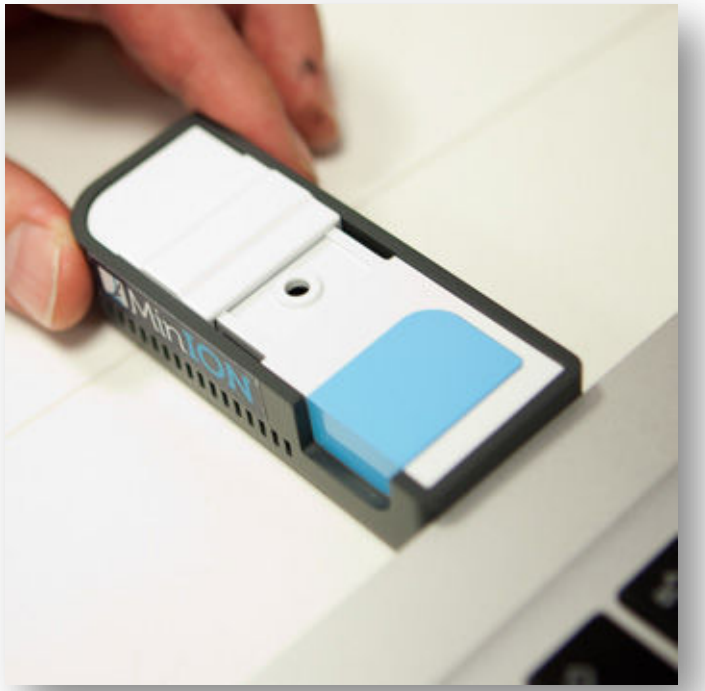
- Beads with DNA fragments
- Universal adapter attached to fragments
- PCR product attaches to slide
- Fluorescent probes ligate to the primer

**SOLiD 5500 Genetic Analyzer**



# The future

- "Next-next generation" sequencing (MinION?)
- Smaller data sets?
- Semantic web?
- Back to hypotheses?

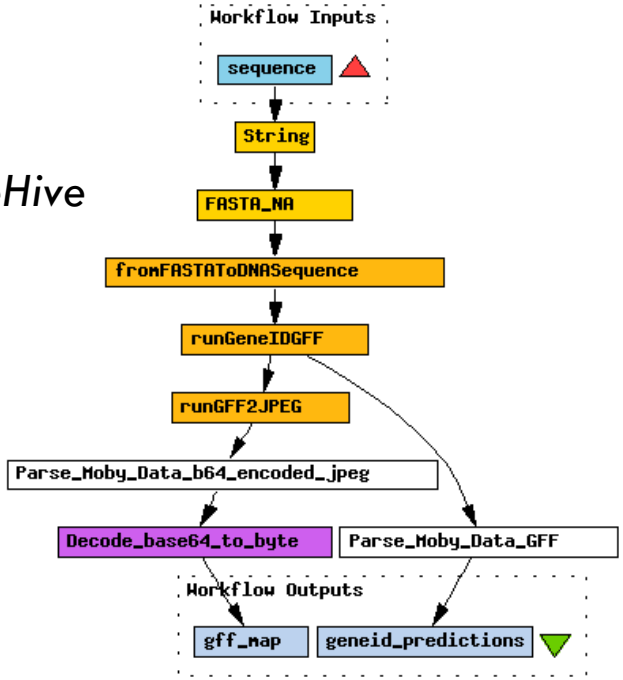


# Workflows

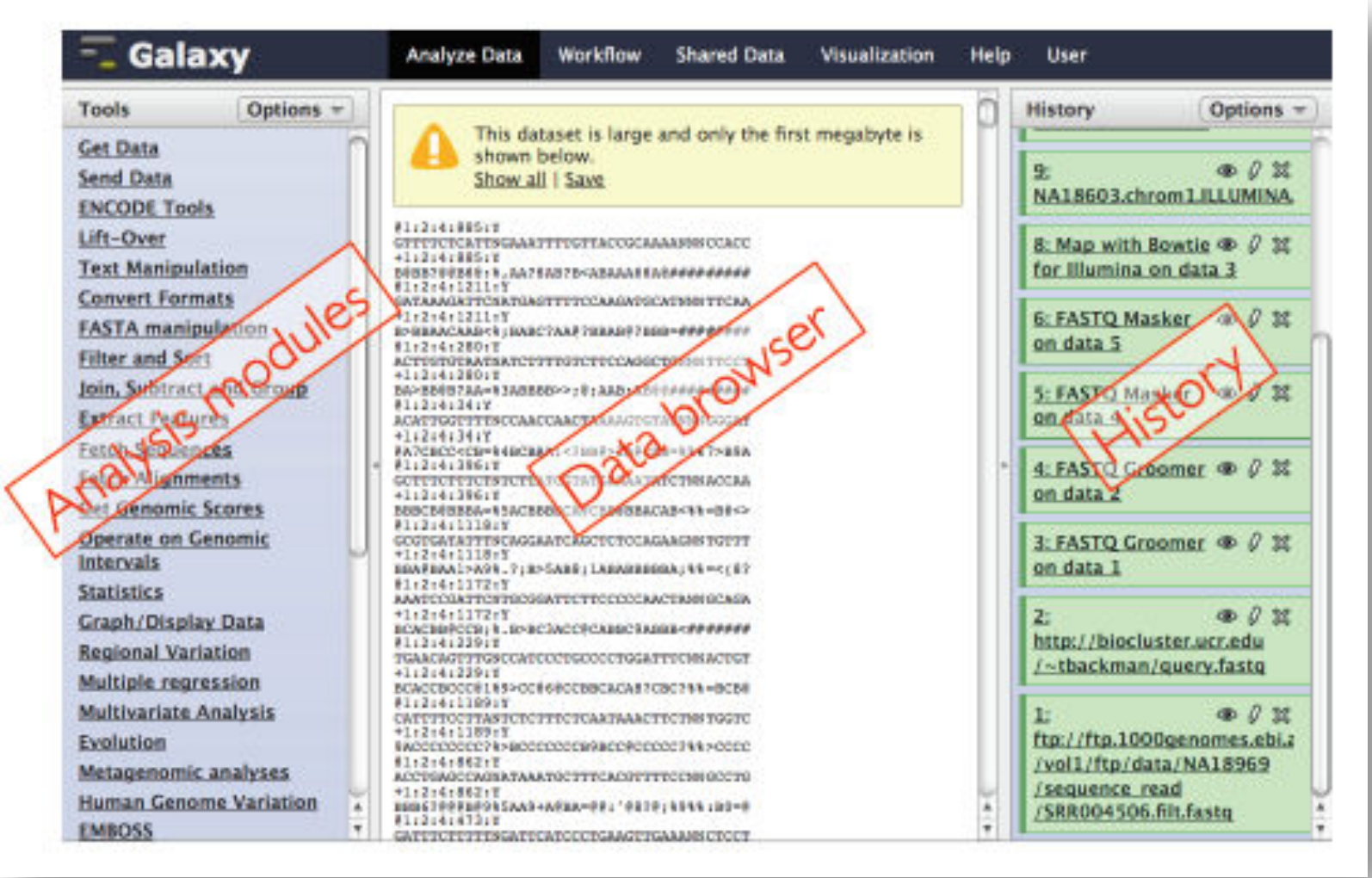
Tools for automating bioinformatics analyses

# Examples of workflow tools

| | |
|---|---|
| Taverna | *taverna.org.uk* |
| Galaxy | *usegalaxy.org* |
| eHive | *ensembl.org/info/docs/eHive* |
| Mobyle | *mobyle.pasteur.fr* |
| Cipres | *phylo.org* |
| Yahoo! Pipes | *pipes.yahoo.com* |
| “make” | *gnu.org/software/make* |

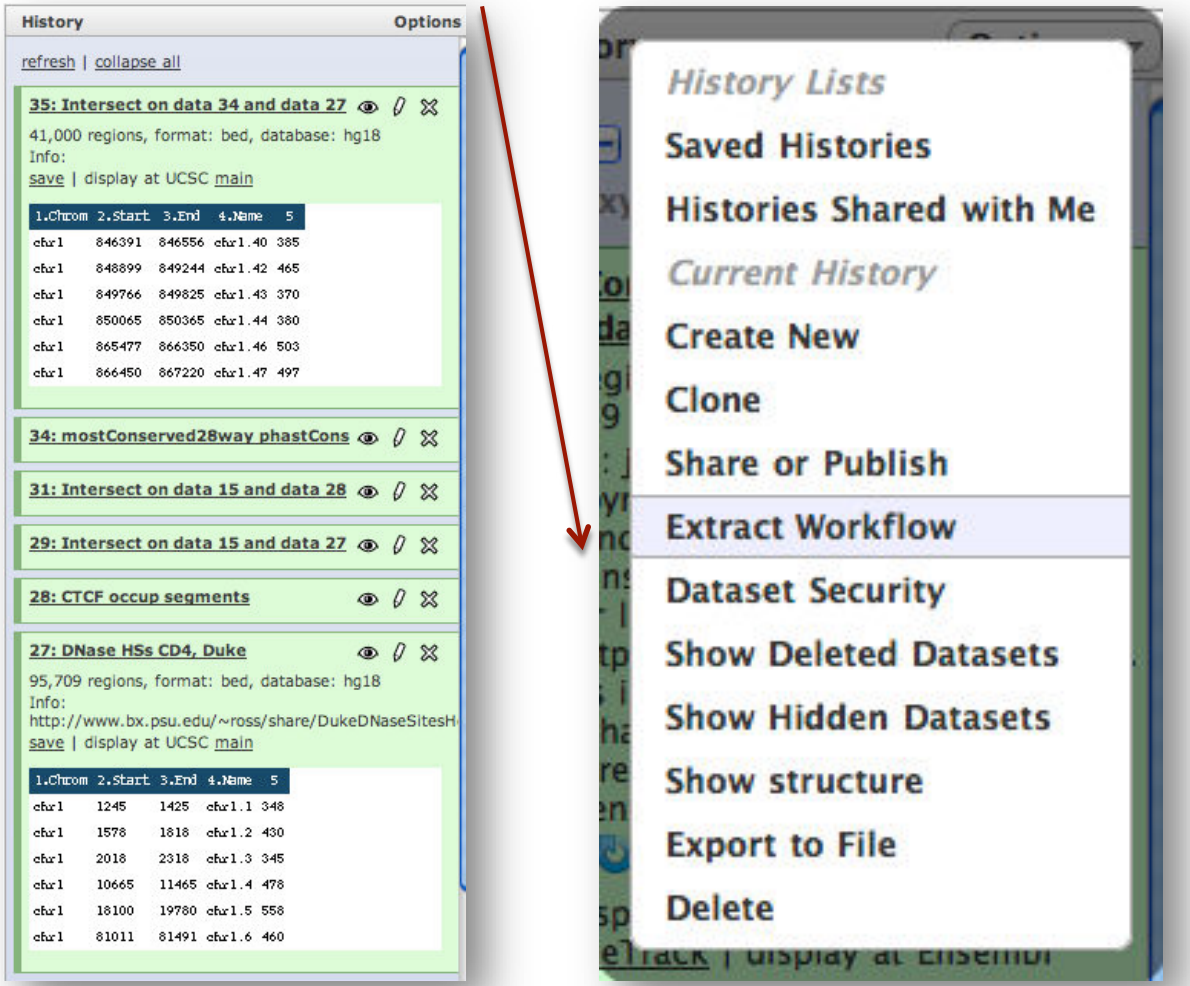


# Galaxy

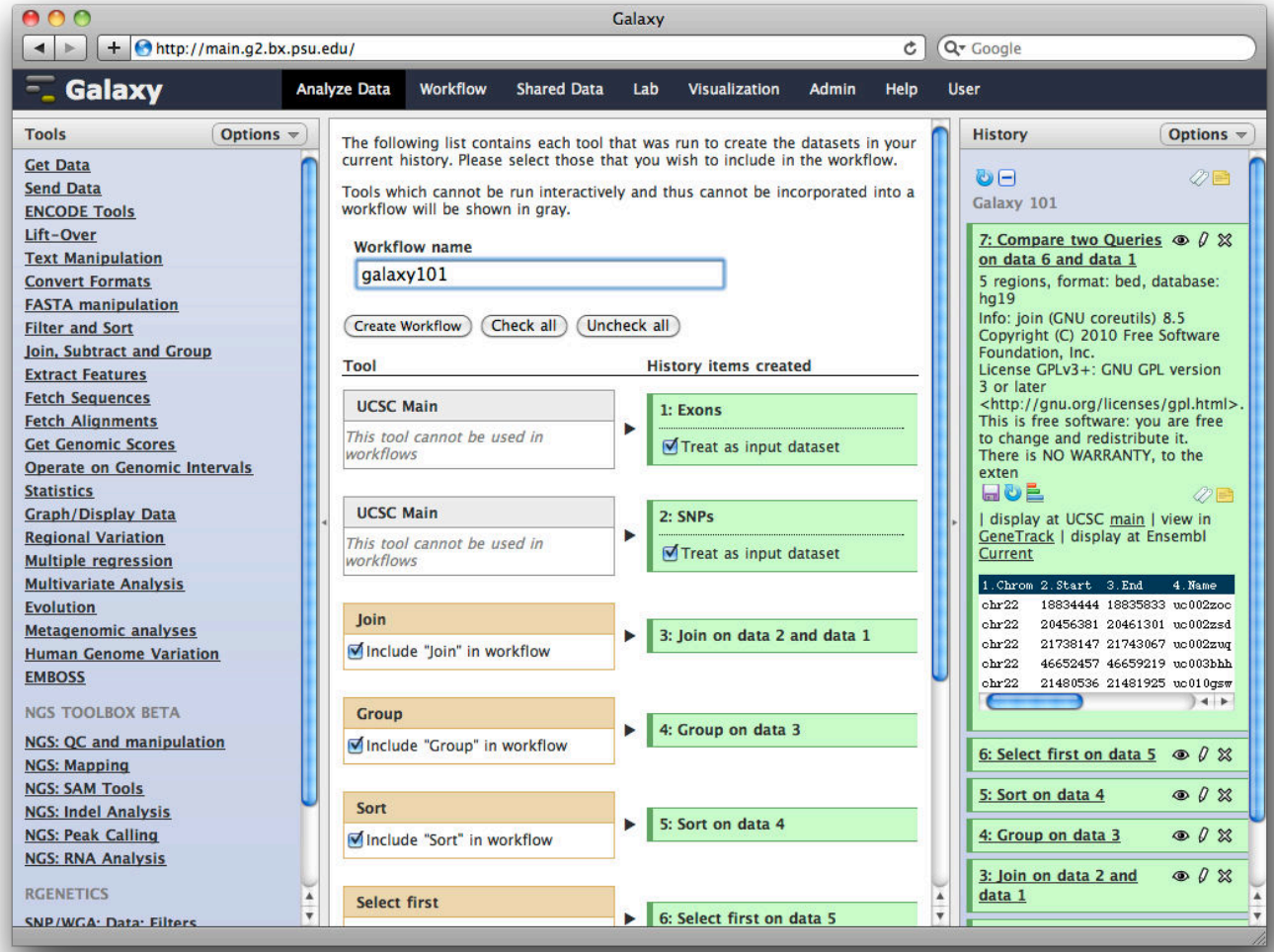

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
Options
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Human Genome Variation
EMBOSS
This dataset is large and only the first megabyte is shown below.
Show all | Save
History
Options
9: NA18603.chrom1.ILLUMINA.
8: Map with Bowtie for Illumina on data 3
6: FASTQ Masker on data 5
5: FASTQ Masker on data 4
4: FASTQ Groomer on data 2
3: FASTQ Groomer on data 1
2: http://biocluster.ucr.edu/~tbackman/query.fastq
1: ftp://ftp.1000genomes.ebi.a/vol1/ftp/data/NA18969/sequence_read/SRR004506.filt.fastq
Analysis modules
Data browser
History

# Provenance, histories

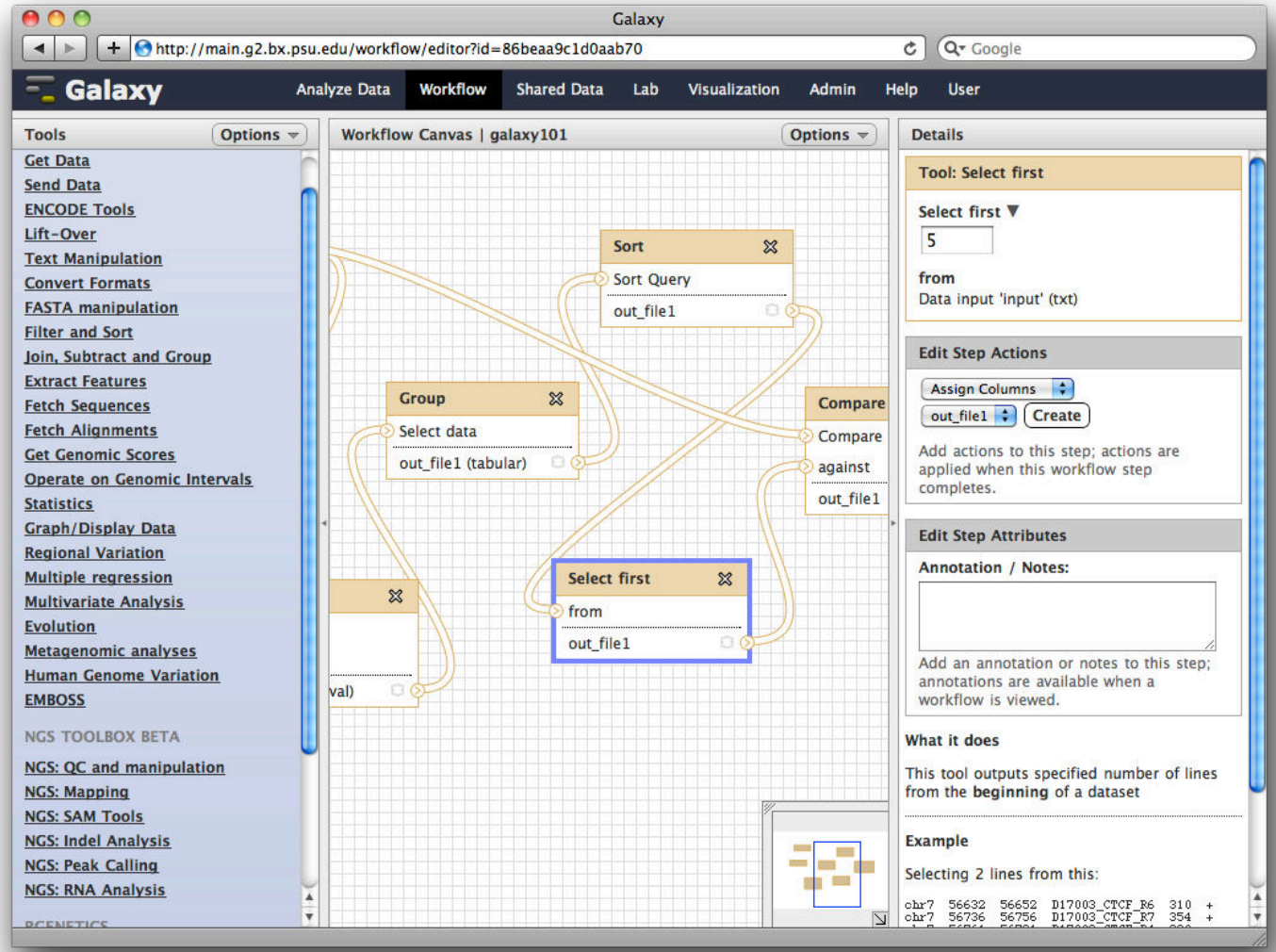
- Where do the data come from?
- How were they altered?
- Galaxy tracks the history of data.
- Histories can be converted to workflows



# Creating a workflow

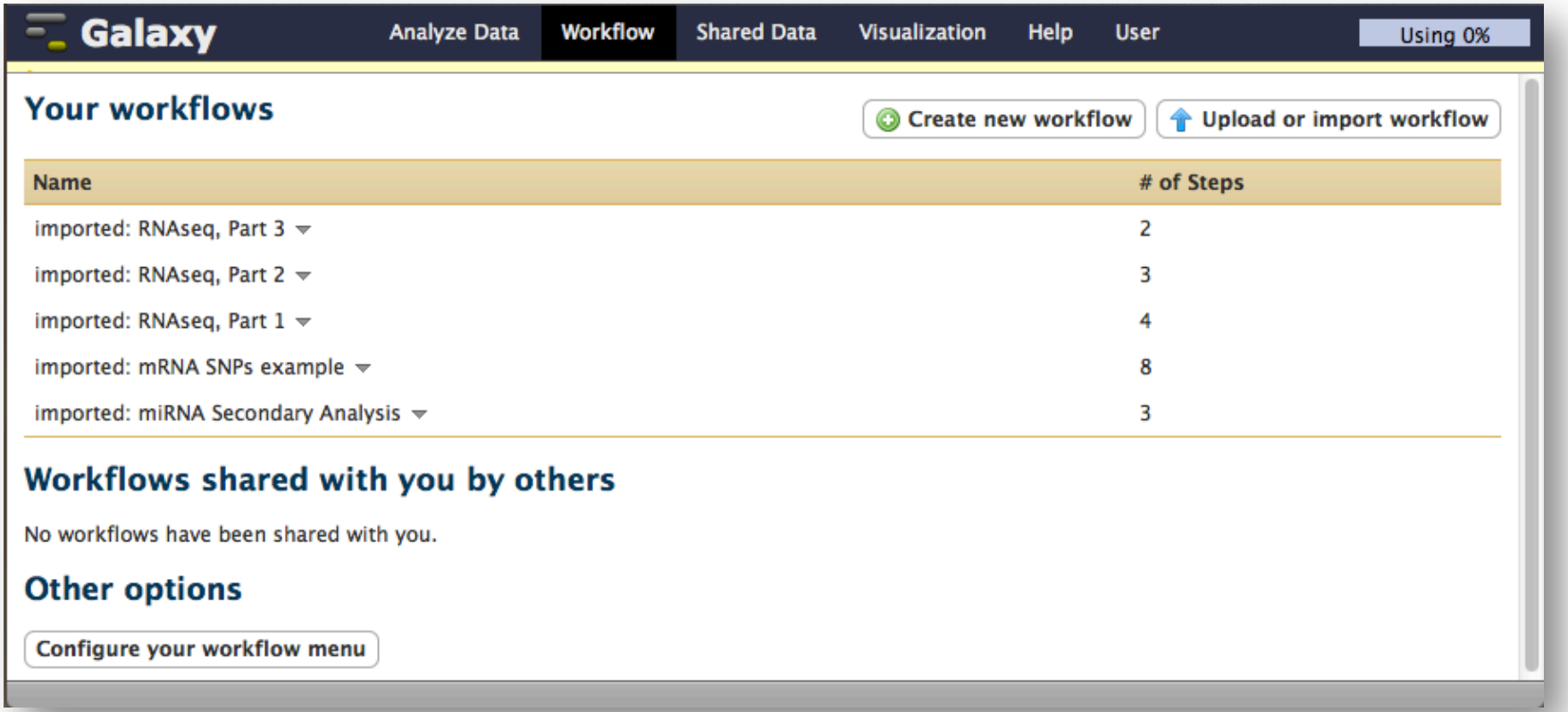


# Workflow editor



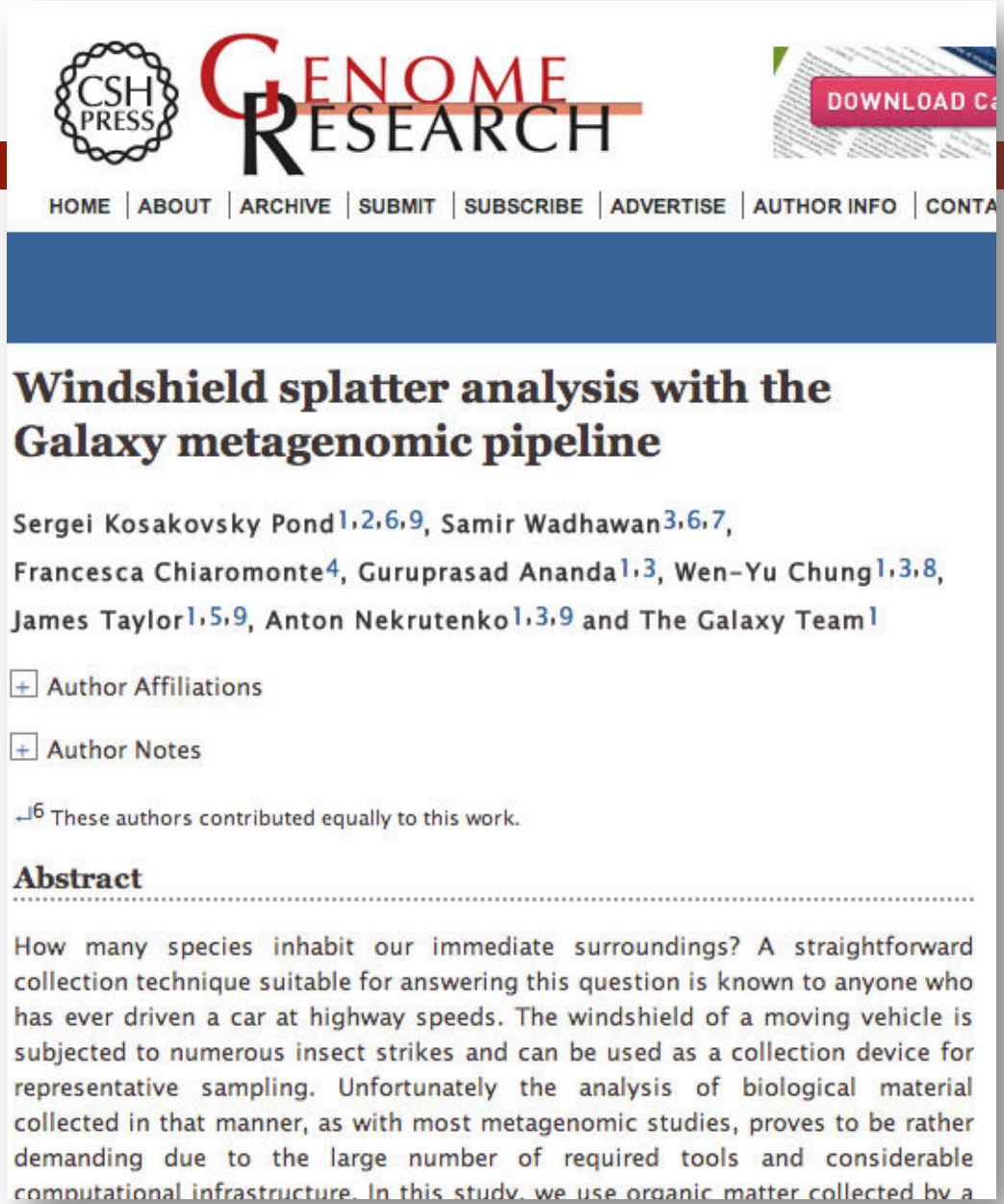
# Reproducibility

- Good science requires that results be reproducible
- Some analyses are run many times



# Sharing

- "Standing on the shoulders of giants"
- Not re-inventing the wheel
- "Executable papers" (doi: 10.1101/gr.094508.109)



CSH PRESS GENOME RESEARCH

DOWNLOAD Ca

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTA

**Windshield splatter analysis with the Galaxy metagenomic pipeline**

Sergei Kosakovsky Pond[1,2,6,9], Samir Wadhawan[3,6,7], Francesca Chiaromonte[4], Guruprasad Ananda[1,3], Wen-Yu Chung[1,3,8], James Taylor[1,5,9], Anton Nekrutenko[1,3,9] and The Galaxy Team[1]

[+] Author Affiliations

[+] Author Notes

[6] These authors contributed equally to this work.

**Abstract**

How many species inhabit our immediate surroundings? A straightforward collection technique suitable for answering this question is known to anyone who has ever driven a car at highway speeds. The windshield of a moving vehicle is subjected to numerous insect strikes and can be used as a collection device for representative sampling. Unfortunately the analysis of biological material collected in that manner, as with most metagenomic studies, proves to be rather demanding due to the large number of required tools and considerable computational infrastructure. In this study, we use organic matter collected by a

# Data

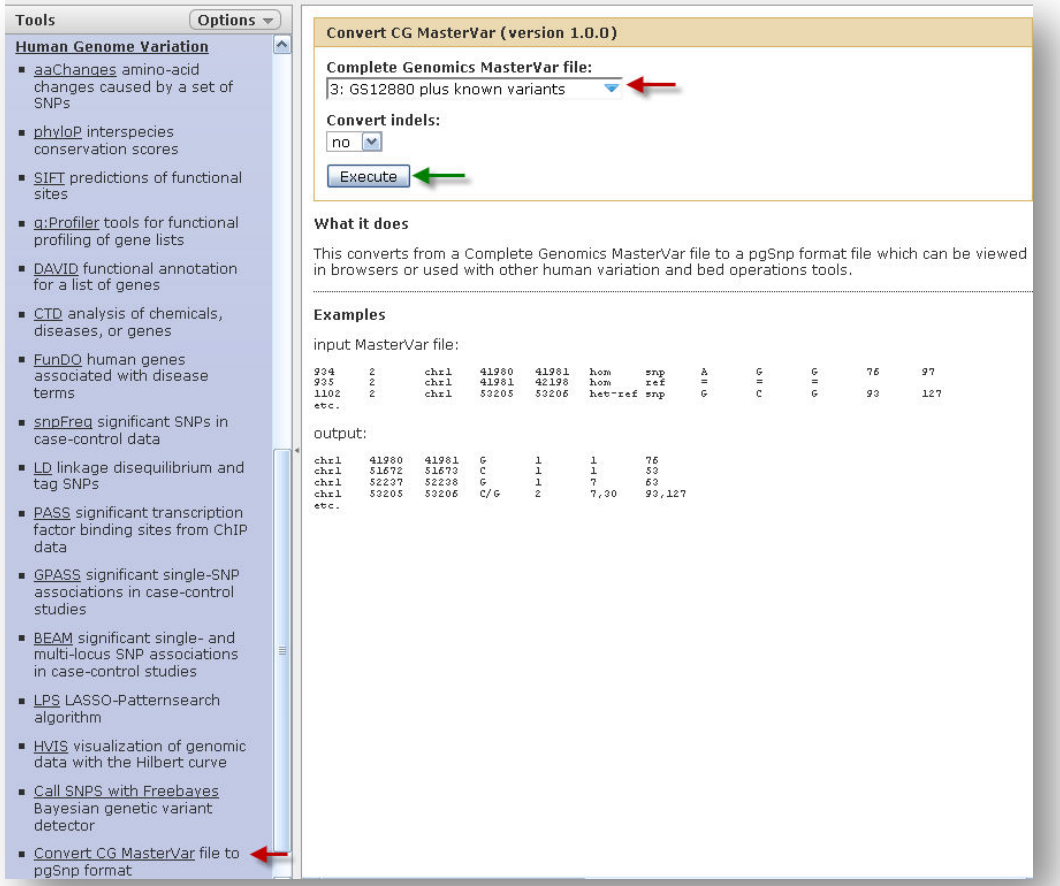
What data types, expressed in which formats, does Galaxy operate on? How do I get data into and out of Galaxy?

# Data types

- Sequences
- Alignments
- Intervals
- Tabular data

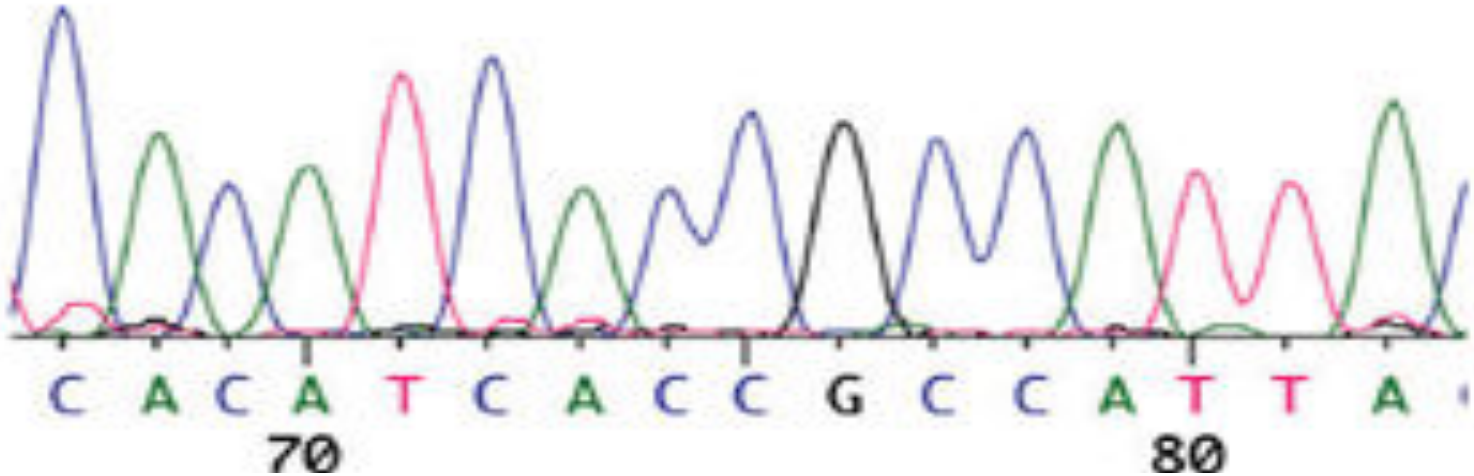
# File format conversion

- Built-in converters between related file formats are provided
- Additional converters can be added



# Galaxy sequence formats

- **FASTA –** *def line, sequence*
- **FASTQ –** *FASTA + quality - SOLEXA:*
- **ABI/SCF –** *binary sequence trace (see below)*
- **SFF (454) –** *binary flowgram format*



# Galaxy alignment formats

- **MAF –** *multiple alignment format (see below)*
- **(S|B)AM –** *text or binary format for reads and ref*
- **AXT –** *pairwise alignment, from LAV*
- **LAV –** *BLASTZ output, pairwise alignment*

```
##maf version=1
a score=606741.000000
s hg18.chr7       3532968 27 - 158821424 ATGCTGTCCCTCTTCCCCAGCCCAGGG
s panTro2.chr7    3633994 27 - 160261443 ATGCTGTCCCTCTTCCCCAGCCCAGGG
s rheMac2.chr3    3465688 27 - 196418989 ATGCTGCCCCTCTTCCCCAGCCCGGGG
s canFam2.chr16 40964059 24 -  62570175 ATGCCCCCC---CCTCCCACCTCAGTG
```

line indicator | species | chromosome | start position* | number of bases | strand | chromosome length | sequence

* Start positions on negative strand are relative to the reverse complement of the source sequence

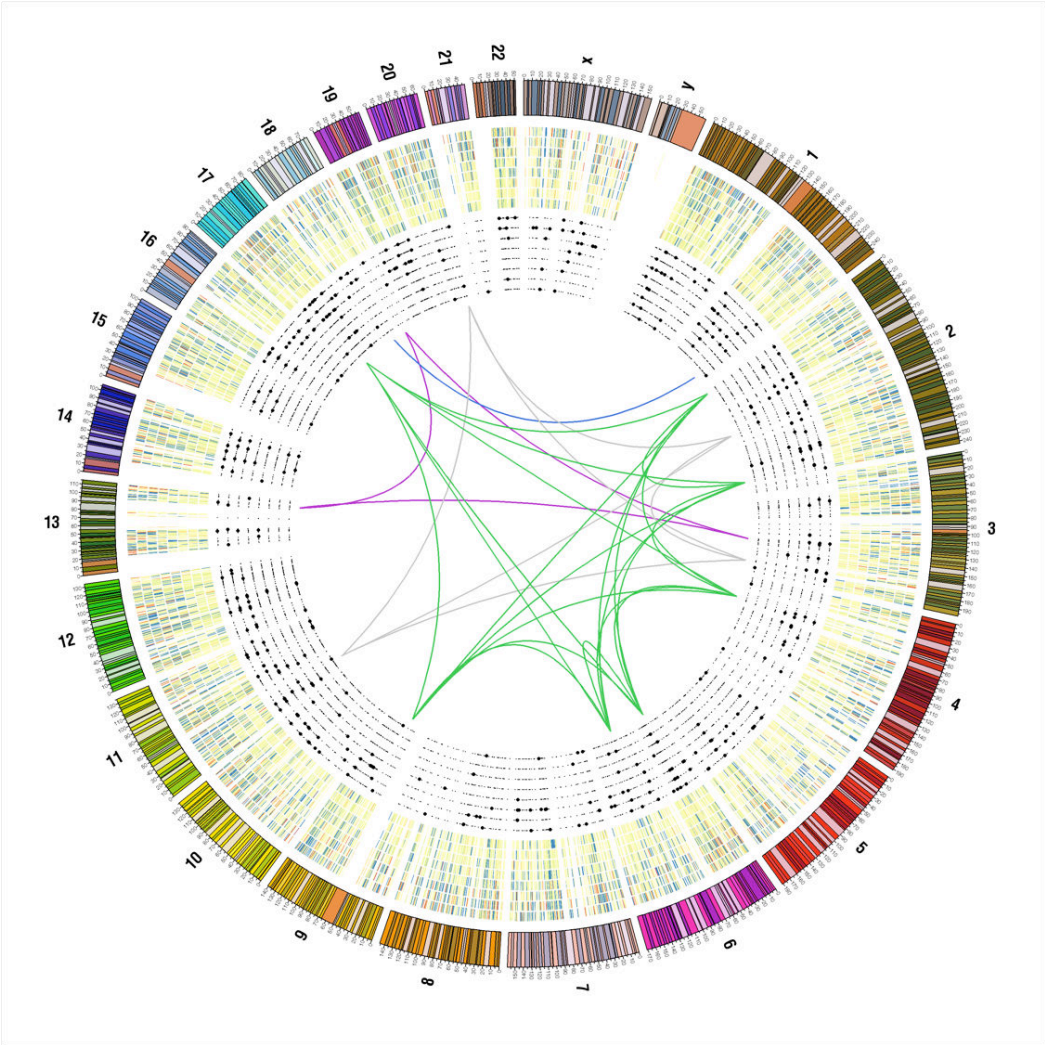
```
##maf version=1
a score=23767937.0
s hg18.chr10      29548679 3 - 135374737 G---CG
s canFam2.chr28   24796549 6 -  44191819 GGAACA
s mm9.chr19       13586550 6 -  61342430 GGAACT
s rheMac2.chr9    29548600 6 - 133323859 G---CG
s panTro2.chr10   30467642 3 - 135001995 G---CG

a score=23767937.0
s hg18.chr10      26460595 3 - 135374737 TCC
s rheMac2.chr9    26448564 3 - 133323859 TCG
s panTro2.chr10   27296115 3 - 135001995 TCG
s mm9.chr19       10589604 3 -  61342430 TCG

a score=23767937.0
s hg18.chr10     112028937 3 + 135374737 CTG
s panTro2.chr10  110891932 3 + 135001995 CTG
s rheMac2.chr9   109970341 3 + 133323859 CTG
s mm9.chr19       53444041 3 +  61342430 CTG
s canFam2.chr28   24648318 3 +  44191819 CTG
```

# Galaxy interval/feature formats

- **BED**
  - *chrom, start, end, …*
- **INTERVAL**
  - *BED with headers*
- **GFF (GFF3)**
  - *like BED and interval, but 1-based inclusive*
- **WIG(GLE)**
  - *dense, continuous-valued tracks*



# Other data formats

In addition to interval/feature tabular data as listed previously, other files with similar properties can be processed by some tools:

- *.txt tab-separated values (e.g. tabular FASTA)
- HTML (for additional prose)
- LPED/PBED (to describe SNPs, really two files, one for coordinates, other for alleles)

# Data I/O

- Upload via FTP
- Fetch by URL
- Import from data library
- Provided by interoperable web service

# Upload data using FTP

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 351.3 Gb

Tools | Options

Get Data
- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- BX main browser
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- modENCODE worm server
- Wormbase server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server

Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments

Upload File (version 1.1.3)

File Format:
Auto-detect
Which format? See help below

File:
Browse...
TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:
Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

| File | Size | Date |
|---|---|---|
| masterVarBeta-GS12880-1100-37-ASM.tsv.bz2 | 233.9 Mb | 11/29/2011 01:47:30 PM |

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at **main.g2.bx.psu.edu** using your Galaxy credentials (email address and password).

Convert spaces to tabs:
Yes
Use this option if you are entering intervals by hand.

Genome:
Human Feb. 2009 (GRCh37/hg19) (hg19)

Execute

Auto-detect

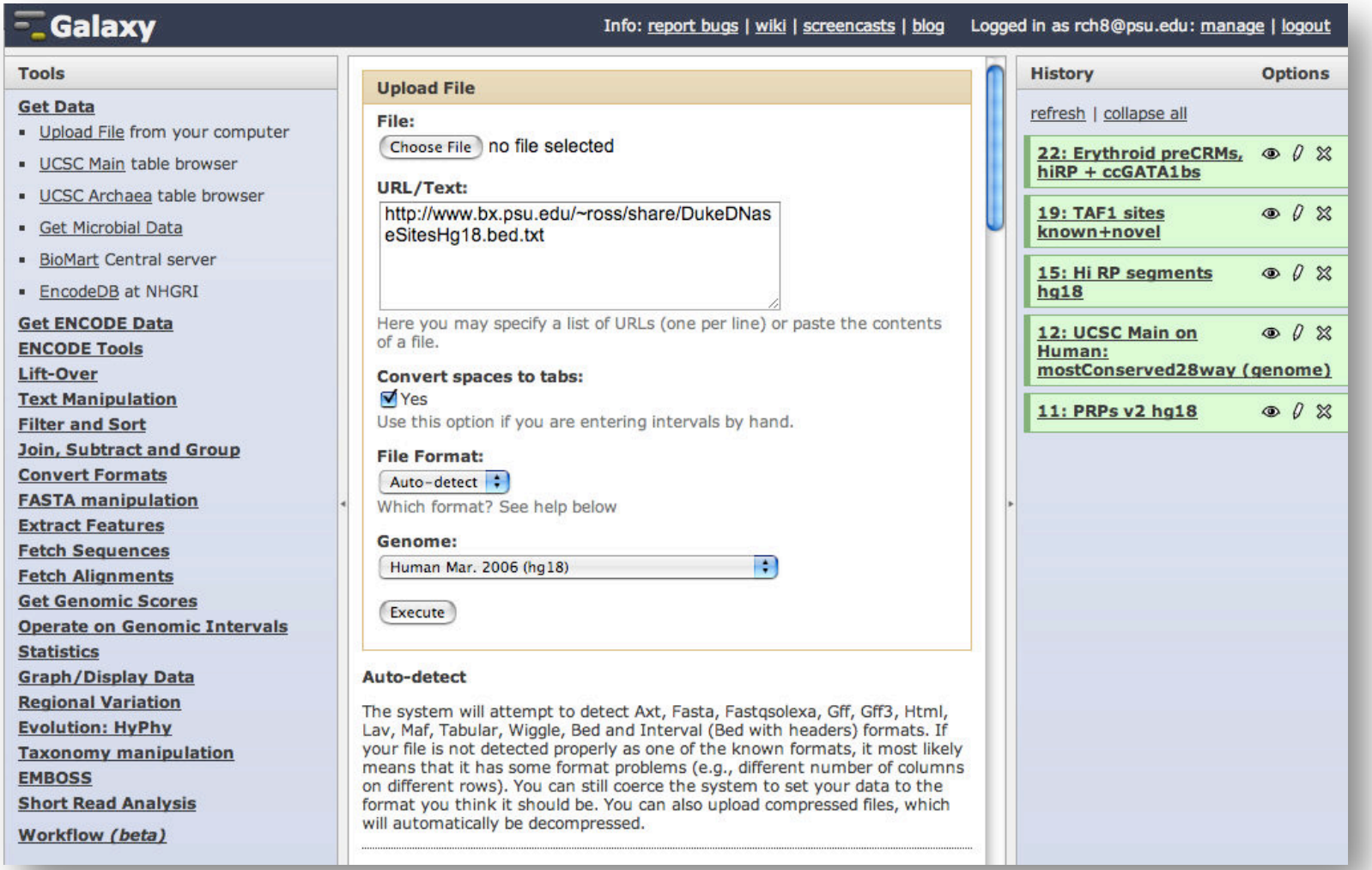
The system will attempt to detect Axt, Fasta, Fastqsolexa, Gff, Gff3, Html, Lav, Maf, Tabular, Wiggle, Bed and Interval (Bed with headers) formats. If your file is not detected properly as one of the known formats, it most likely means that it has some format problems (e.g.,

History | Options

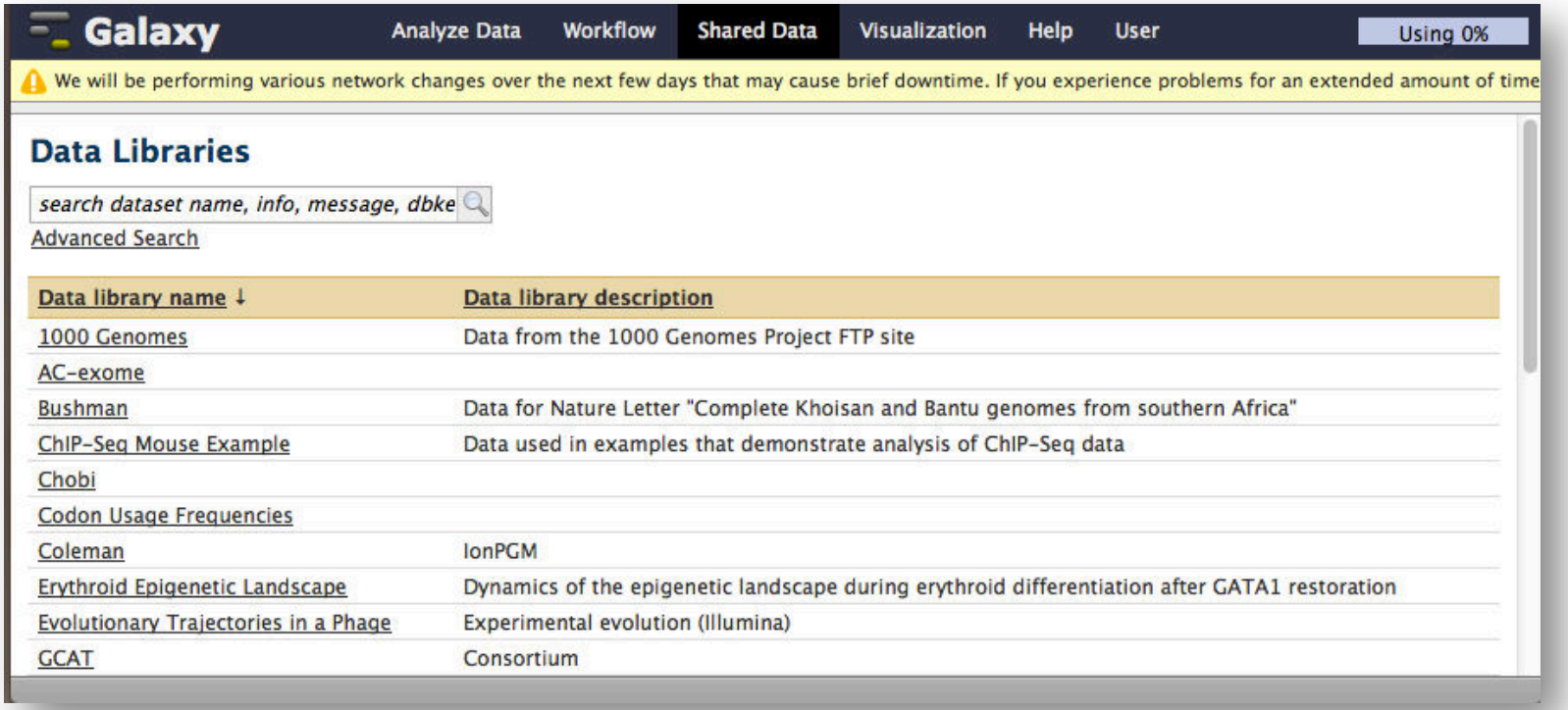
Unnamed history | 0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

# Fetch data from a URL

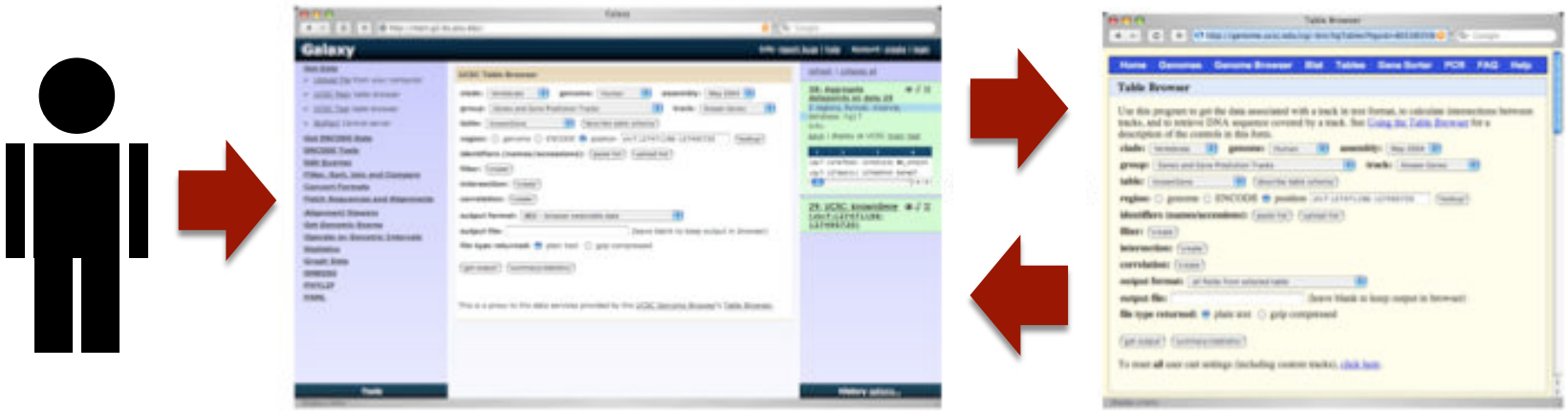


# Import from data library



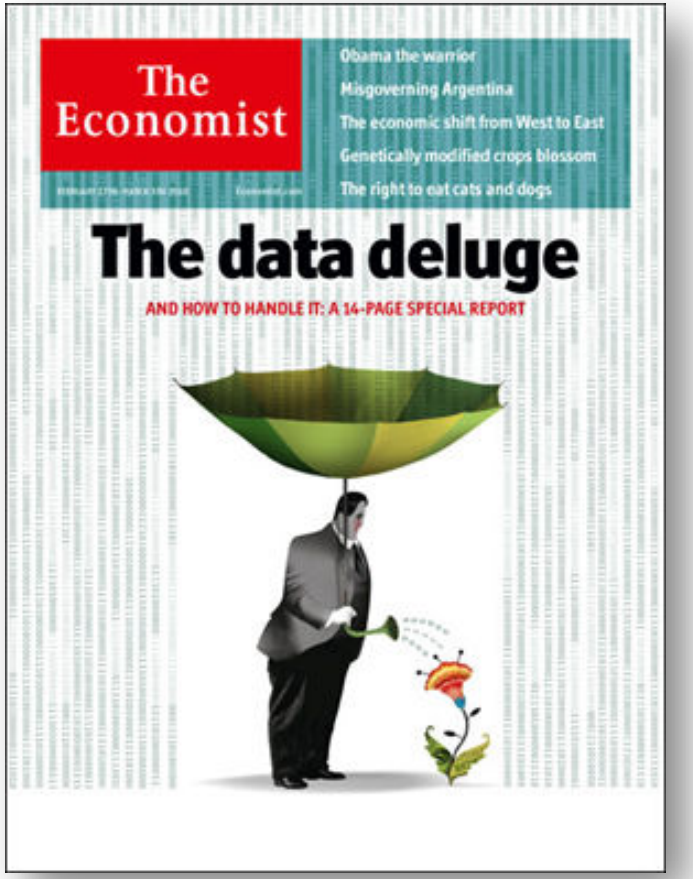
# Fetch data from proxy service

- User submits proxy request to Galaxy
- Galaxy forwards request to remote service
- Service returns data
- Galaxy infers data type and presents results



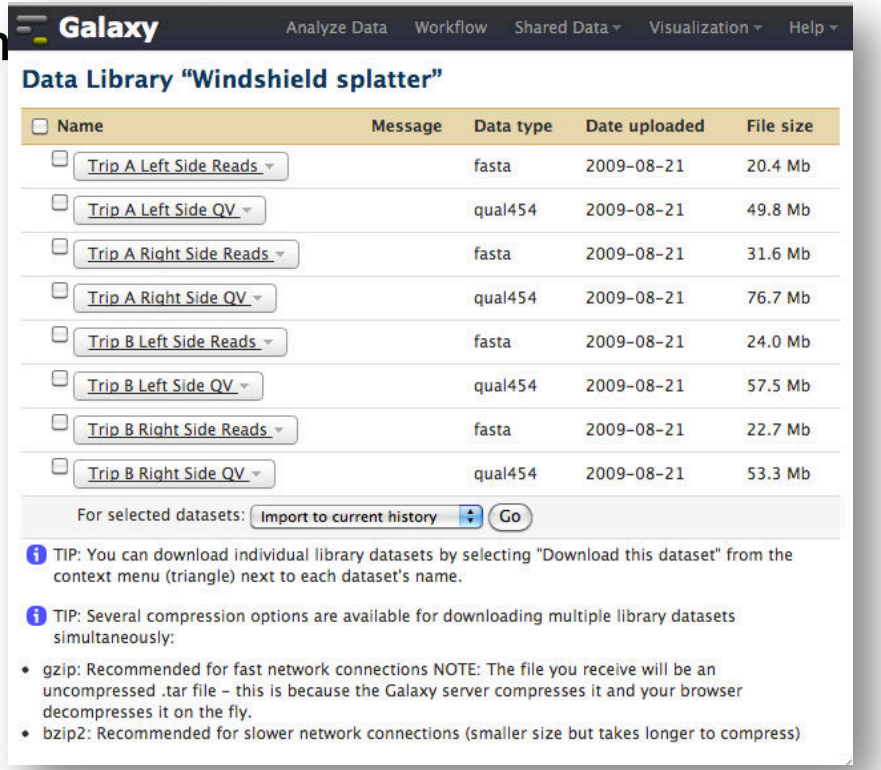
# “Big Data”

- NGS has led to massive data sets
- Data formats are simple, binary, and/or compressed
- Still, people drive around with USB hard disks



# Data sharing/publishing

- The Galaxy platform allows users to publish and share their data, for example as supplemental materials to a publication*




* example: http://genome.cshlp.org/content/19/11/2144

# Tools

Which operations can I run on Galaxy?

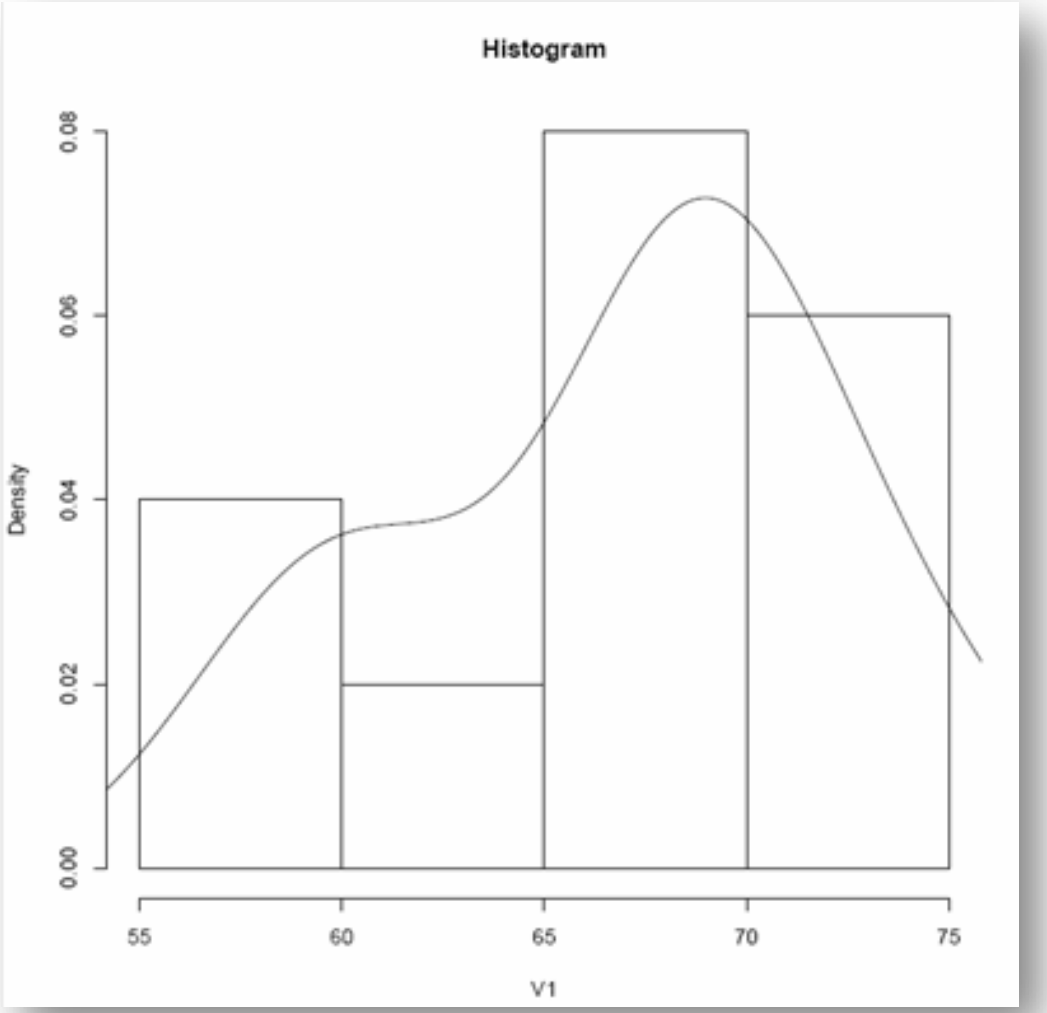
# Galaxy tools

- **Send and Get data** – *Upload, fetch, send, submit*
- **Data manipulation** – *Join, sort, filter*
- **Format conversion** – *FASTA, other format operations*
- **Statistics** – *Regressions, simulations, model tests*
- **NGS** – *BAM, FASTQ, SOLiD, 454 file operations*
- **RNA analysis** – *cufflinks, tophat*
- **Evolution** – *branch lengths, NJ, HyPhy*

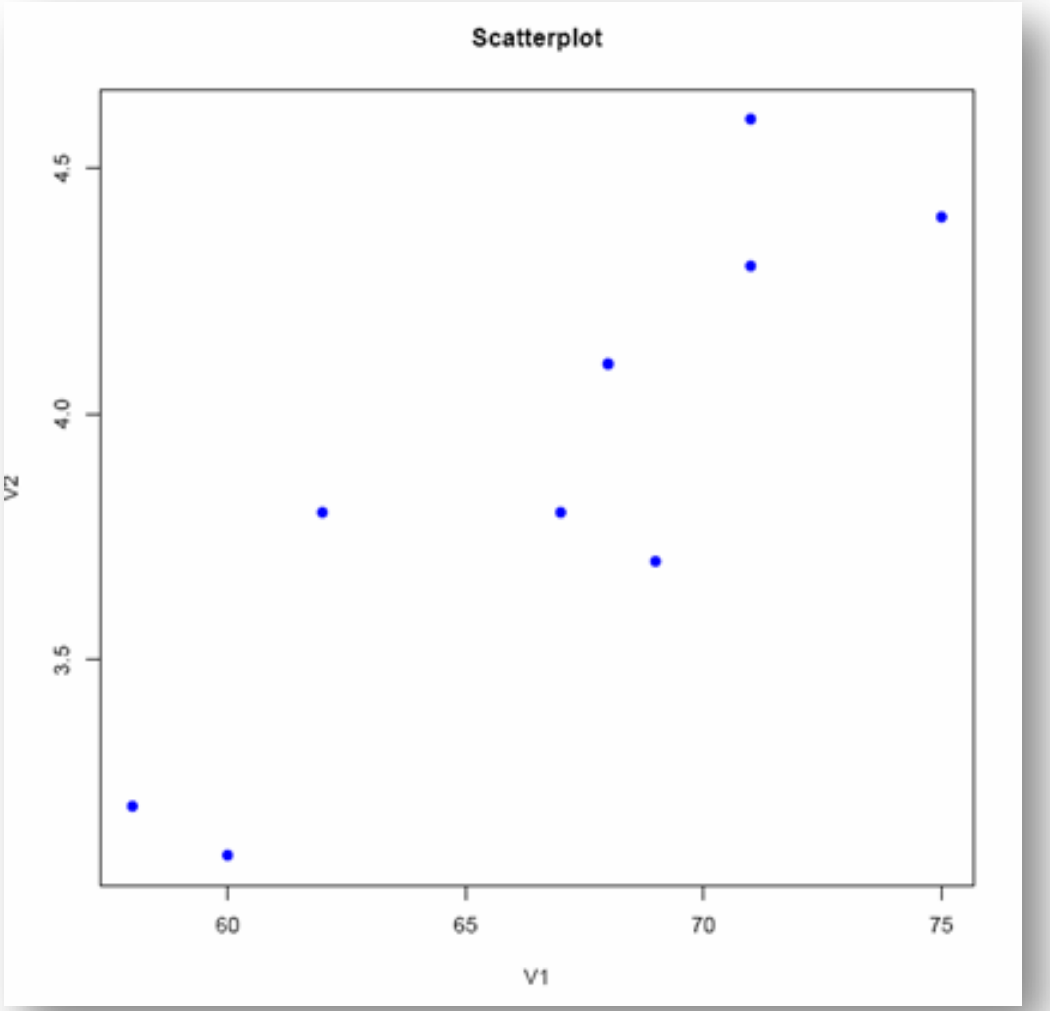
# Visualization

What data can I view in Galaxy?

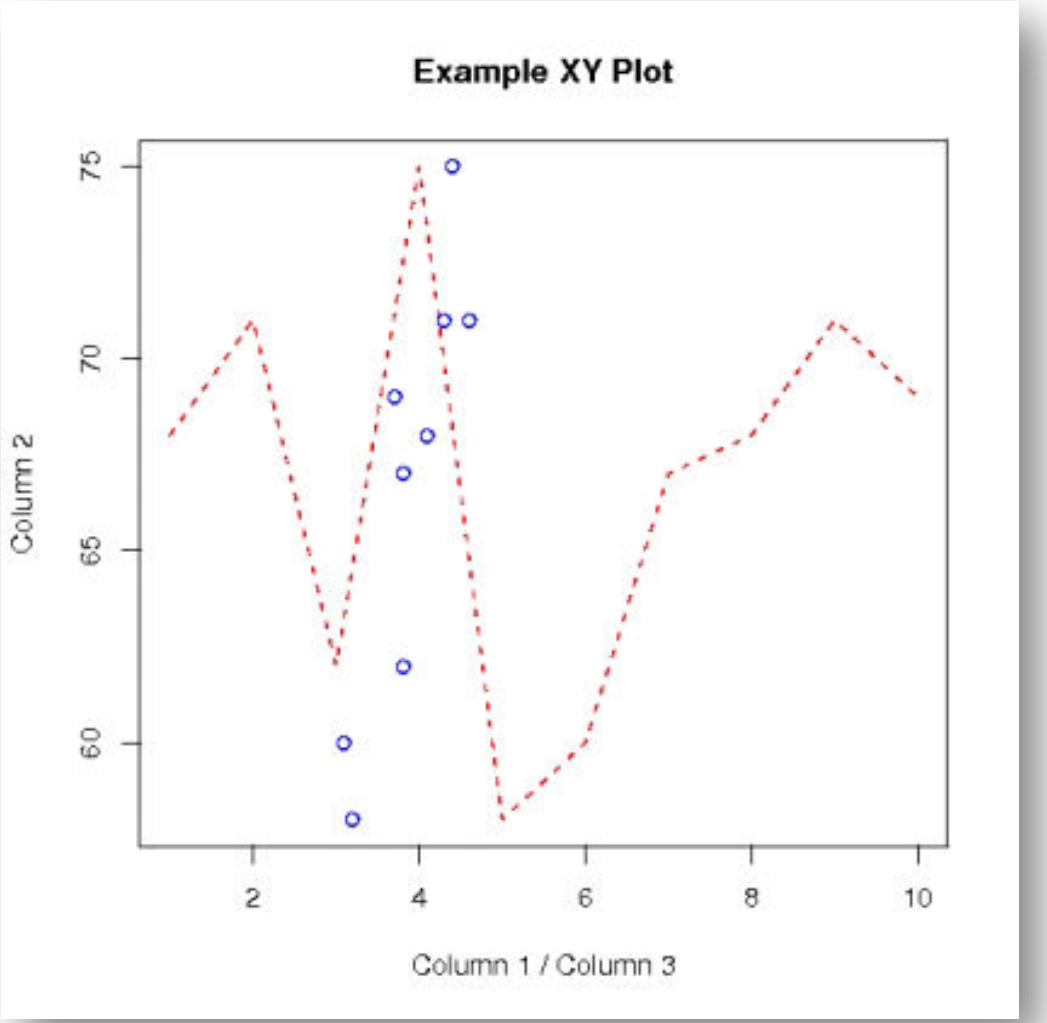
# Galaxy visualization - histograms



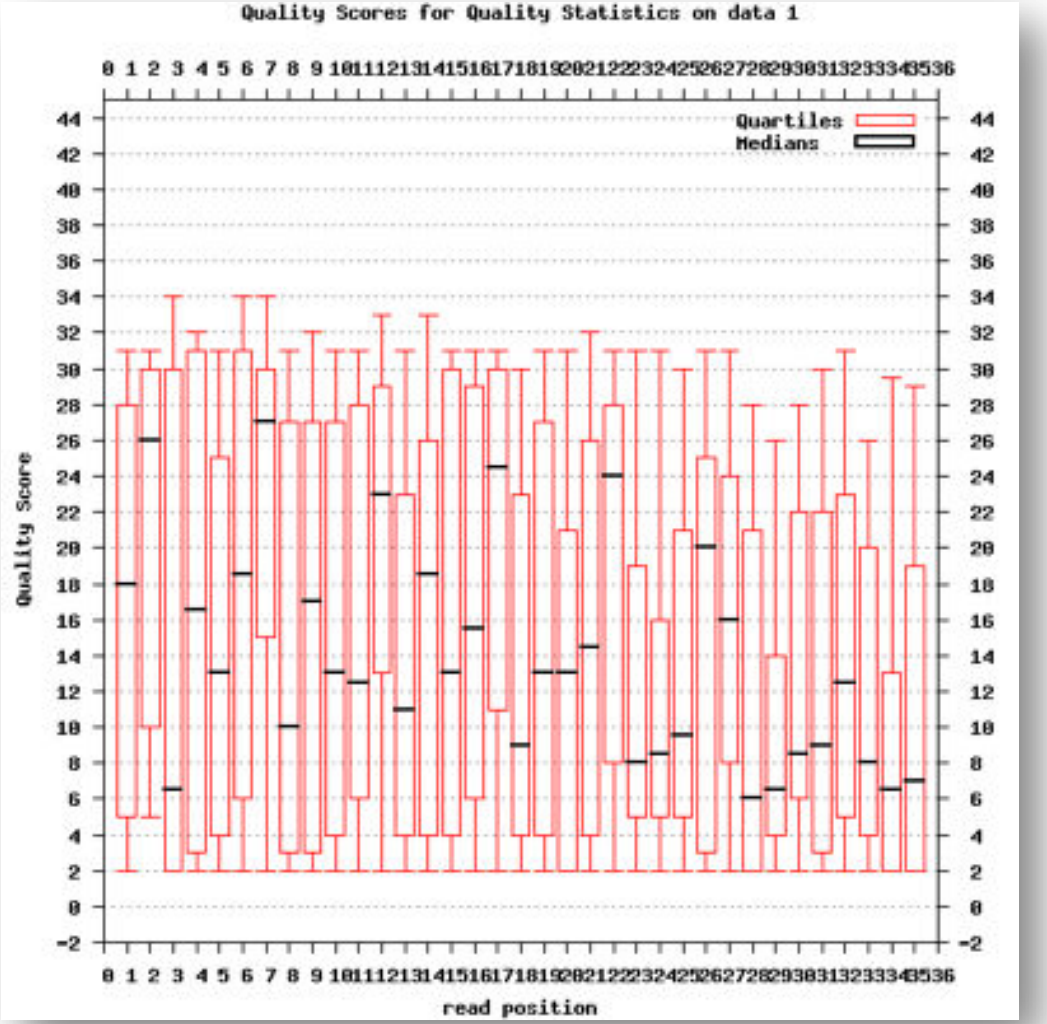
# Galaxy visualization - scatterplots



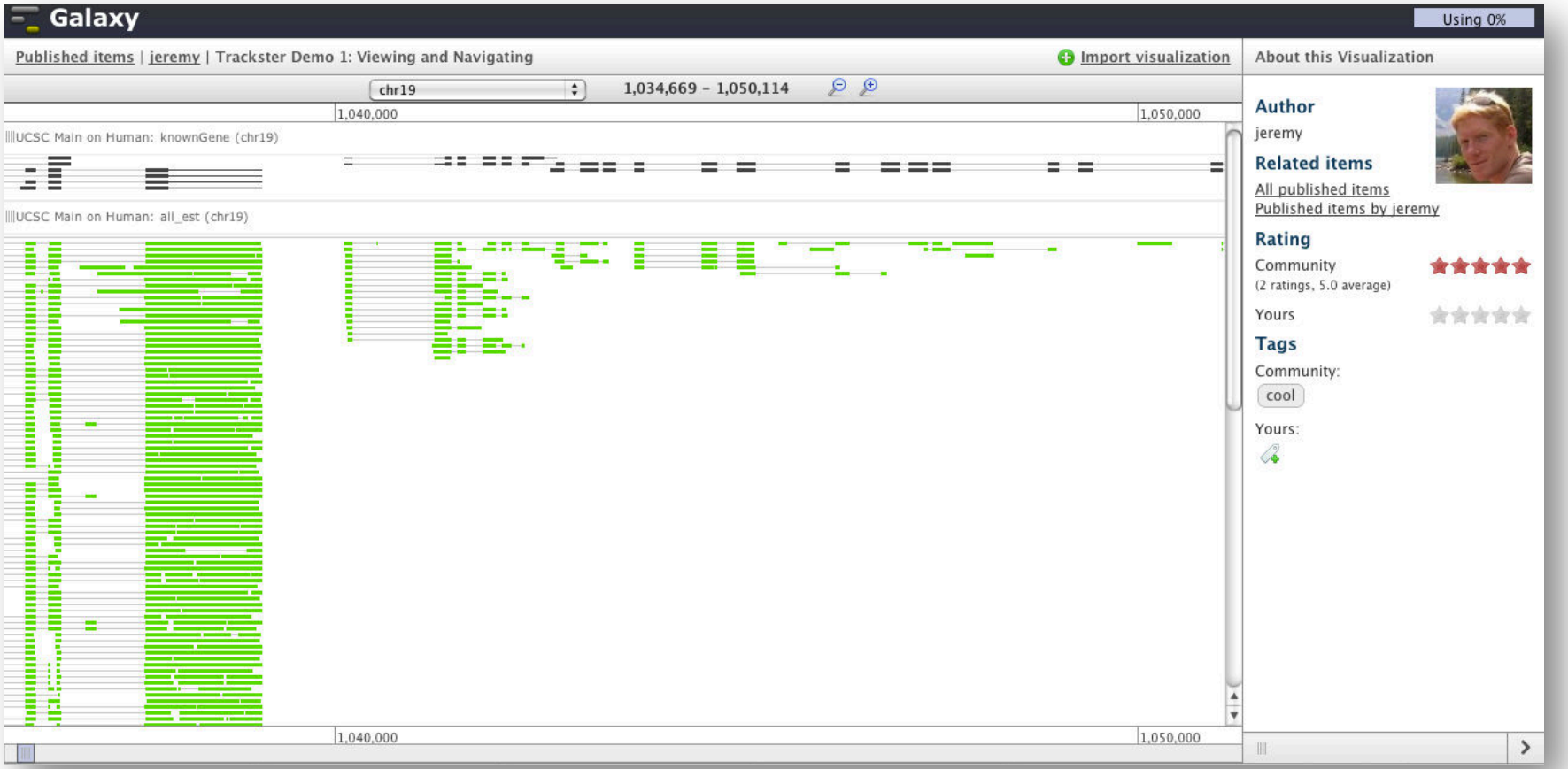
# Galaxy visualization – XY plot



# Galaxy visualization – box plot



# Galaxy visualization - trackster



# Deployment

How to deploy your analyses on Galaxy, on public servers, in the cloud or locally

# Public servers

- “Main” at UseGalaxy.org
- NBIC
- Others listed on Galaxy wiki

# Local server

**Pro:**

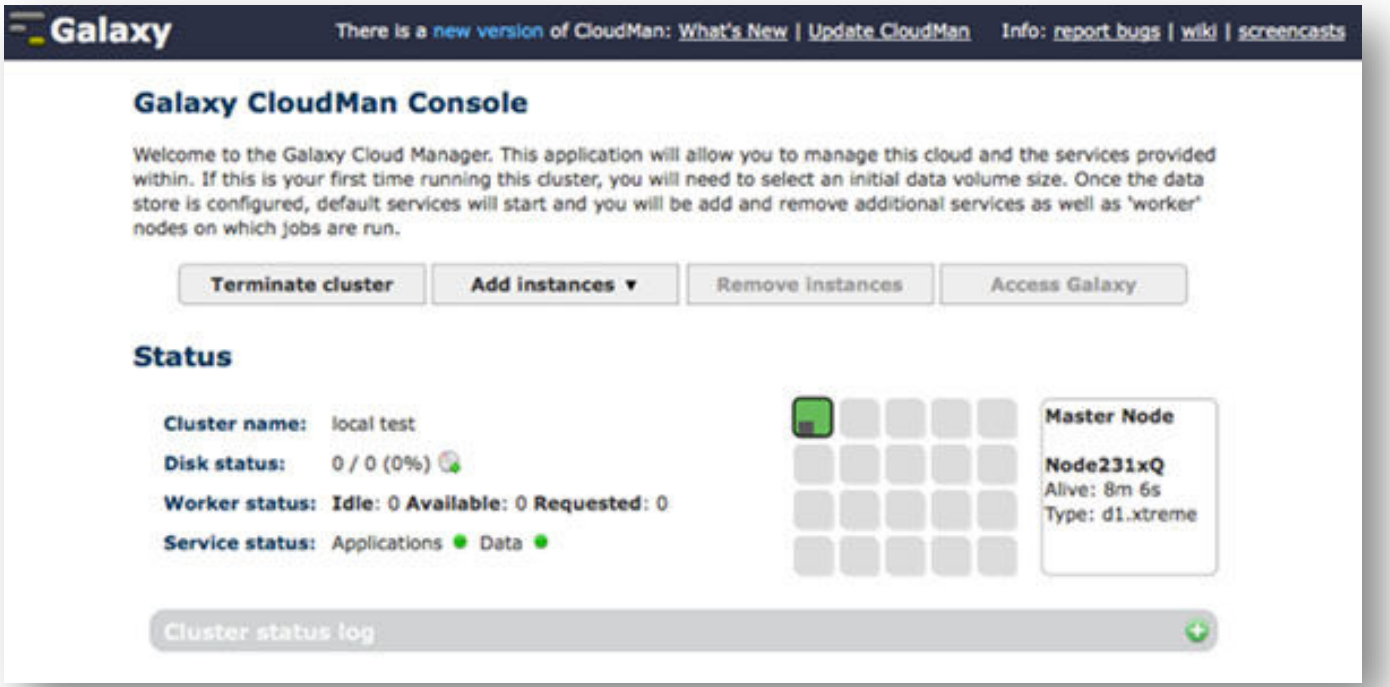
- Data close by
- Can add your own tools
- Can develop Galaxy further
- UNIX-based

**Con:**

- Complicated install
- Many dependencies
- UNIX-based

# Cloud Galaxy

- Galaxy can be installed in the (Amazon EC2 cloud)
- Private data without the hardware hassle
- Uploading and storing data can be costly, however



# Implementation

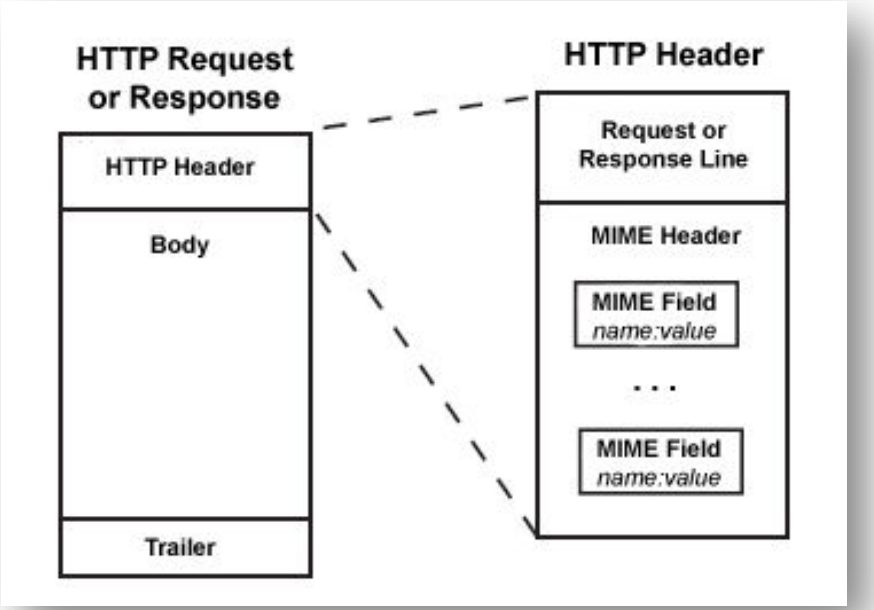
How does it work under the hood?

# Galaxy under the hood


1. Parses HTTP request
2. Identifies which tool to use
3. Reads tool description
4. Queues tool
5. Parses result
6. Returns HTML representation of result

# Web server

- Simple Galaxy installs use a built-in, python-based HTTP server
- More robust installs typically use the Apache httpd server



# Code base

- Most of the framework and the wrapper code is written in python
- Some wrappers in other languages, e.g. perl

# Interface language

- Under the hood, Galaxy executes command-line programs and scripts
- Their interfaces and tool tips are described in XML files

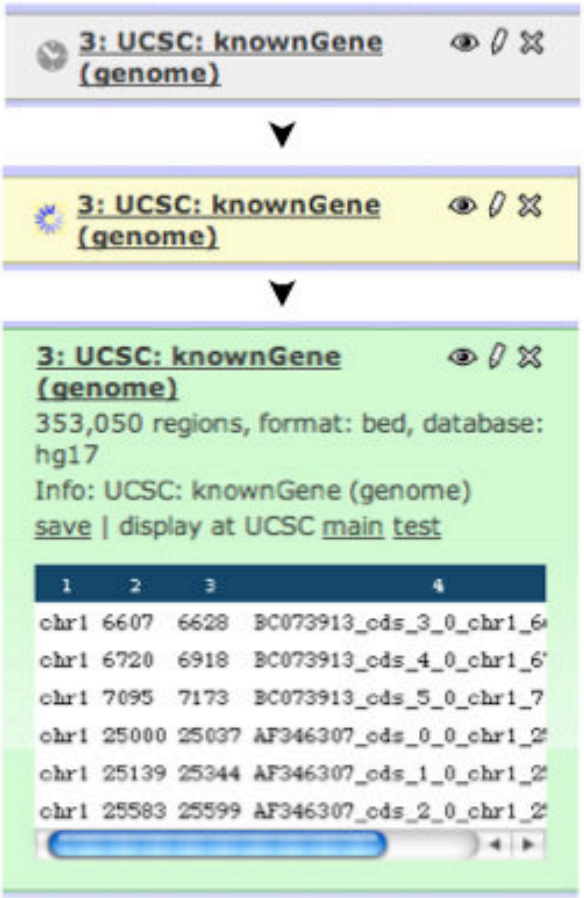
<?xml?>

# Queuing

Queued:

Running:

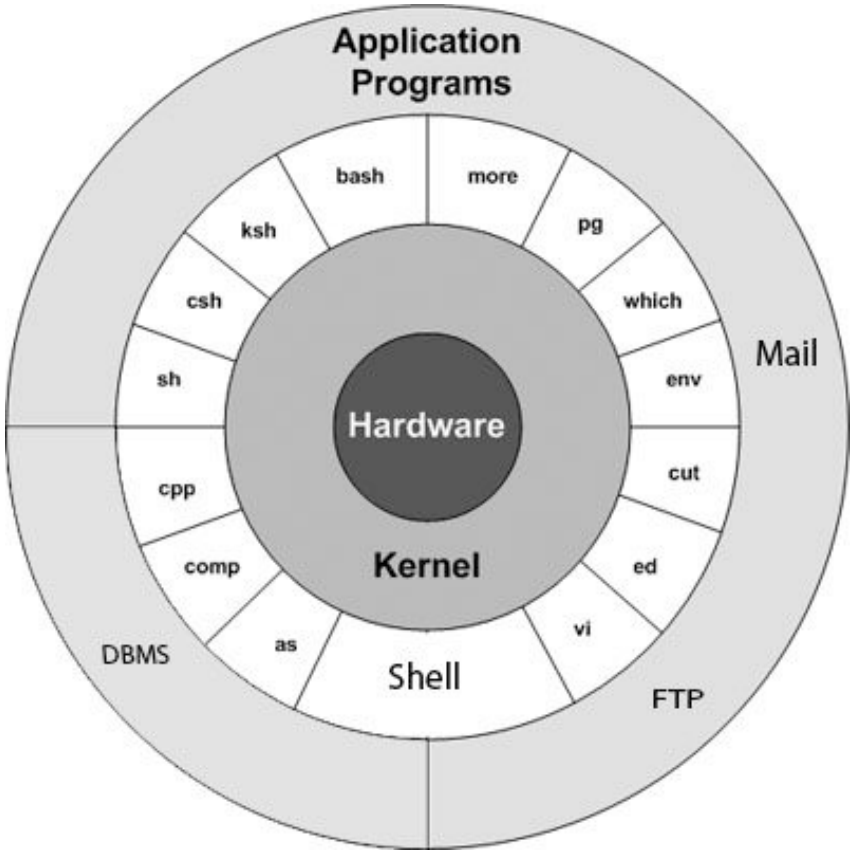
Complete:




- Jobs are executed asynchronously
- Progress is shown in the data browser
- On big servers (e.g. “Main”), queuing is managed by the dedicated “Torque” system

# UNIX

- Galaxy (simply)executes command-line programs within a UNIX-like environment
- Galaxy doesn't have to "know" how to run those programs, it finds out from their descriptions at runtime



# Database

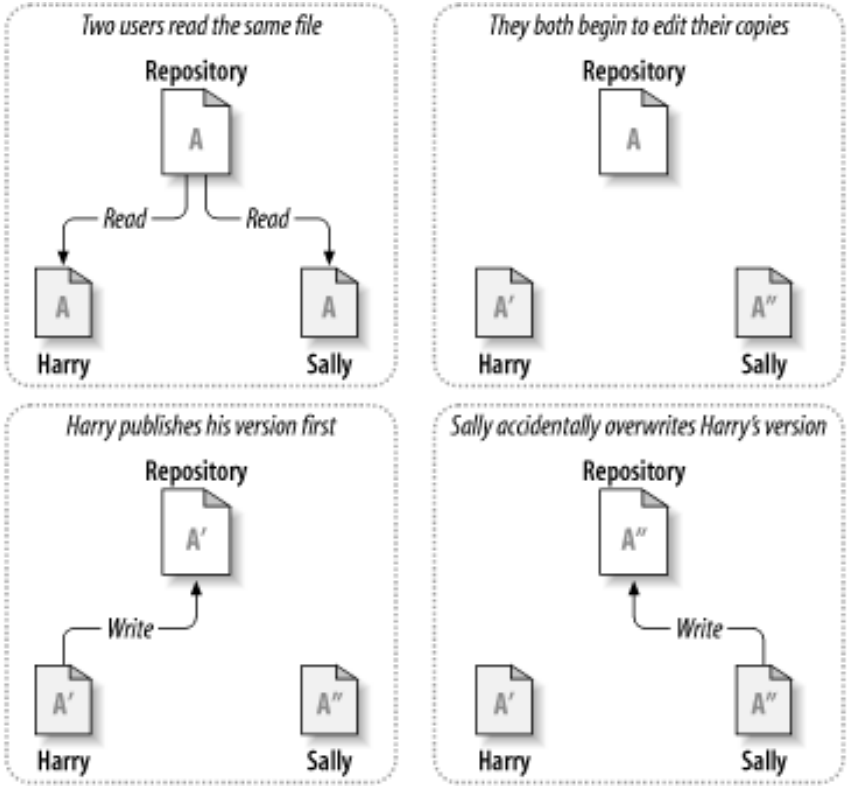
- Analysis metadata is stored in a database, by default this is SQLite
- More robust installs use PostgreSQL

# Configuration

- Galaxy has many moving parts that can be configured
- Configuration is done using simple text-based INI files

# Version control

- Revision control (or version control) provides unlimited undo and detailed tracking of changes
- Galaxy uses Mercurial
- Popular now are svn, git and hg



# Community

How to get in touch with the world-wide community to get the most out of Galaxy?

# Wiki

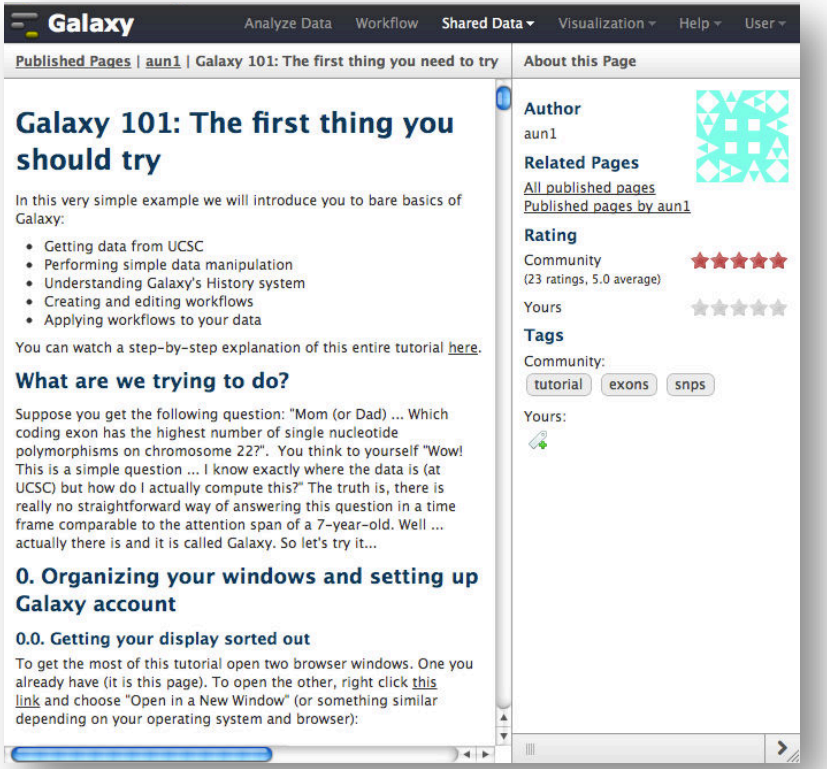

Galaxy Wiki
Login | Search:
Galaxy Project
Hub page for information within this wiki about the Galaxy project itself. See the Learn page for help on using Galaxy, and the Admin page for help on setting up your own instance.
You can take a look at the current members of the core project team. We are funded by NIH, NSF, Penn State, Emory, and the Pennsylvania Department of Public Health.
Project home page
Citing Galaxy
News Briefs
News
Events
Wiki Home Page
Big Picture
Future
Galaxy Team
Project Statistics
Galaxy
web search
Search all Galaxy resources

# Mailing lists

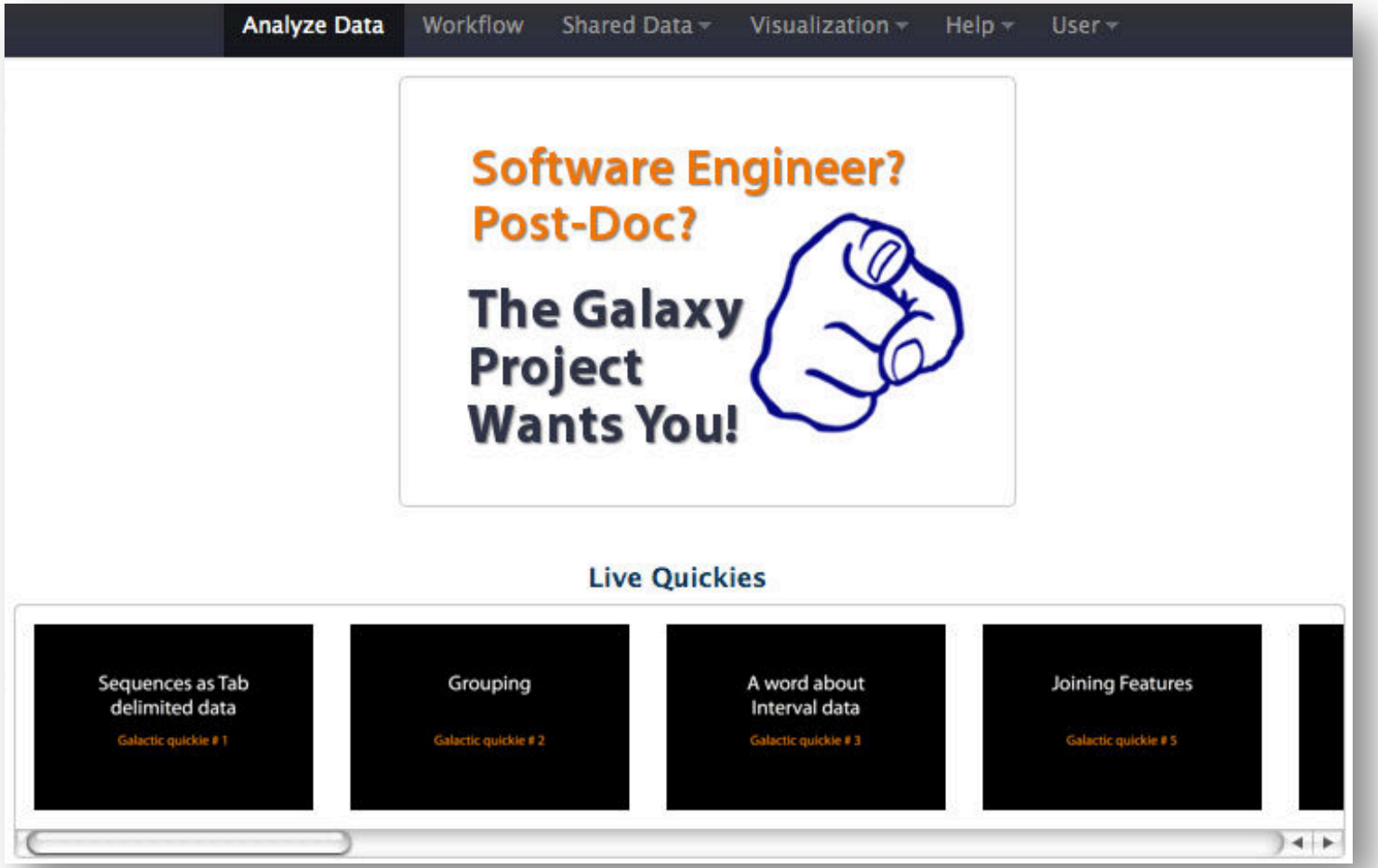
- @lists.bx.psu.edu:
  - galaxy-user
  - galaxy-dev
  - galaxy-announce
  - galaxy-commits

# Tutorials

- **Galaxy 101:**
  - Getting data from UCSC
  - Performing simple data manipulation
  - Understanding Galaxy's History system
  - Creating and editing workflows
  - Applying workflows to your data



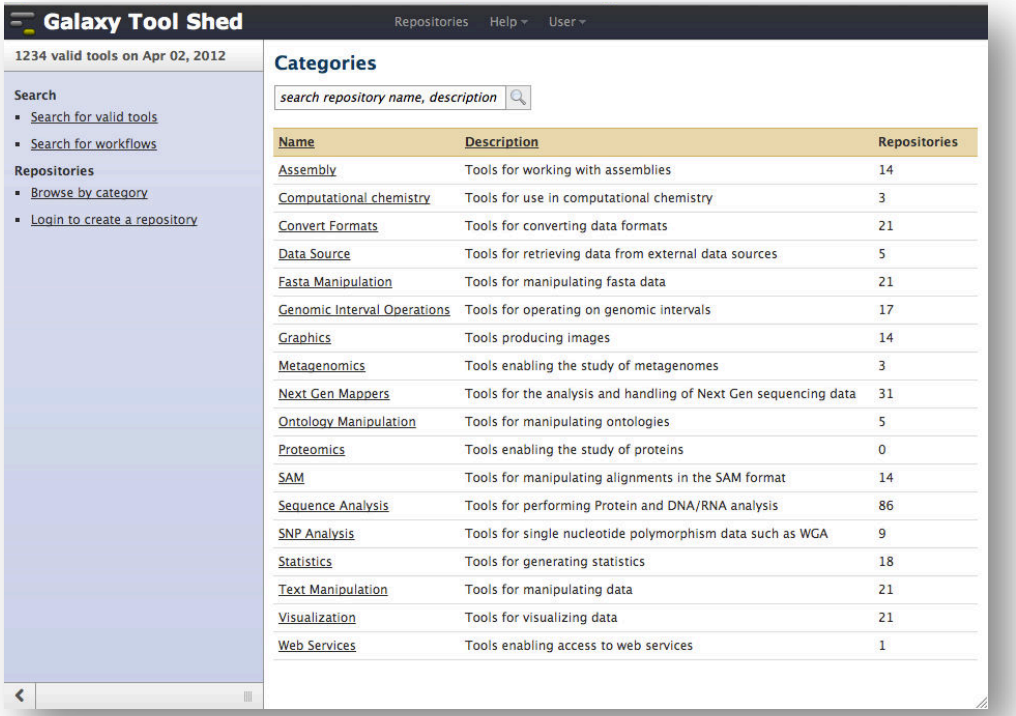
# Screencasts



# Events

- Galaxy Community Conference
- ISMB
- ECCB
- BOSC
- PAG
- Bio-IT World
- GMOD Meetings

# “Tool shed”




| Name | Description | Repositories |
|---|---|---|
| Assembly | Tools for working with assemblies | 14 |
| Computational chemistry | Tools for use in computational chemistry | 3 |
| Convert Formats | Tools for converting data formats | 21 |
| Data Source | Tools for retrieving data from external data sources | 5 |
| Fasta Manipulation | Tools for manipulating fasta data | 21 |
| Genomic Interval Operations | Tools for operating on genomic intervals | 17 |
| Graphics | Tools producing images | 14 |
| Metagenomics | Tools enabling the study of metagenomes | 3 |
| Next Gen Mappers | Tools for the analysis and handling of Next Gen sequencing data | 31 |
| Ontology Manipulation | Tools for manipulating ontologies | 5 |
| Proteomics | Tools enabling the study of proteins | 0 |
| SAM | Tools for manipulating alignments in the SAM format | 14 |
| Sequence Analysis | Tools for performing Protein and DNA/RNA analysis | 86 |
| SNP Analysis | Tools for single nucleotide polymorphism data such as WGA | 9 |
| Statistics | Tools for generating statistics | 18 |
| Text Manipulation | Tools for manipulating data | 21 |
| Visualization | Tools for visualizing data | 21 |
| Web Services | Tools enabling access to web services | 1 |

- Easy sharing of new tools
- Based on Mercurial
- Turns Galaxy into a modular ecosystem

# CiteULike group

- **Social citation manager**
- **There is a Galaxy group:**
  - citeulike.org/group/16008
- **Articles are tagged by:**
  - PROJECT, ISGALAXY, SHARED, HOWTO, METHODS, REPRODUCIBILITY, WORKFLOW

# Organizations

- Development:
  - Penn State University
  - Emory University
- Support:
  - NSF
  - NHGRI
- Power user:
  - NBIC





# Links

- **UseGalaxy.org** – *the Main server*
- **GetGalaxy.org** – *for local installs*
- **UseGalaxy.org/galaxy101** – *intro tutorial*
- **Galaxy.nbic.nl** – *Sombrero, the NBIC server*
- **CiteULike.org/group/16008** – *references*
- **genome.cshlp.org/content/19/11/2144** – *windshield paper*
- **SlideShare.net/rvosa** – *these slides*