



GALAXY BIOINFORMATICS WORKFLOW ENVIRONMENT

Rutger Vos, 3 April 2012

NCB **naturalis**

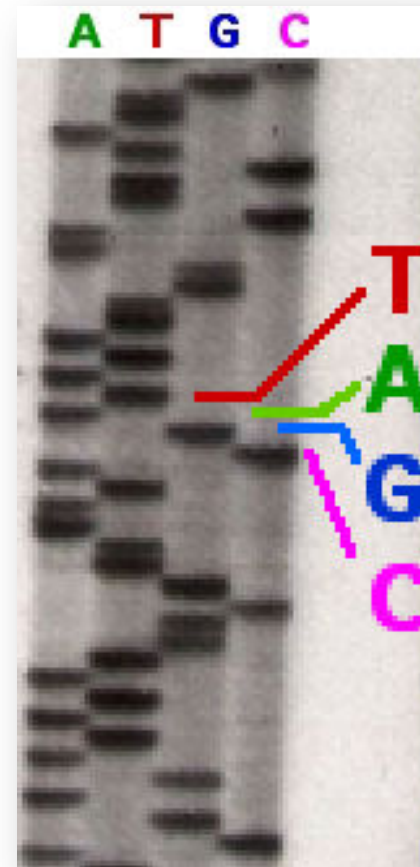


Overview

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

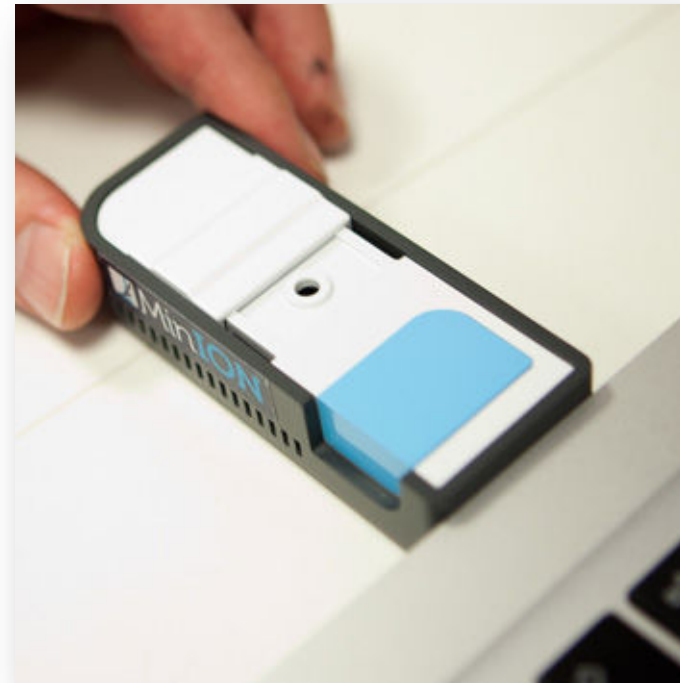
- ❑ Beads with DNA fragments
- ❑ Universal adapter attached to fragments
- ❑ PCR product attaches to slide
- ❑ Fluorescent probes ligate to the primer

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

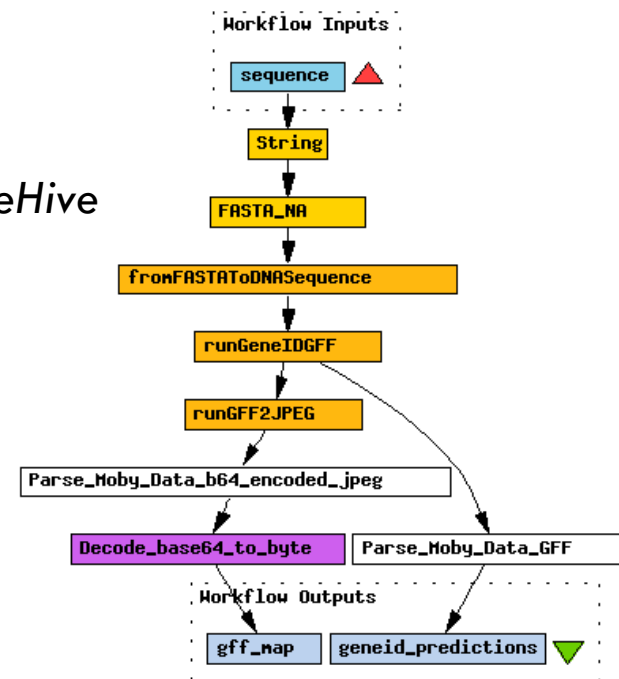
phylo.org

Yahoo! Pipes

pipes.yahoo.com

“make”

gnu.org/software/make

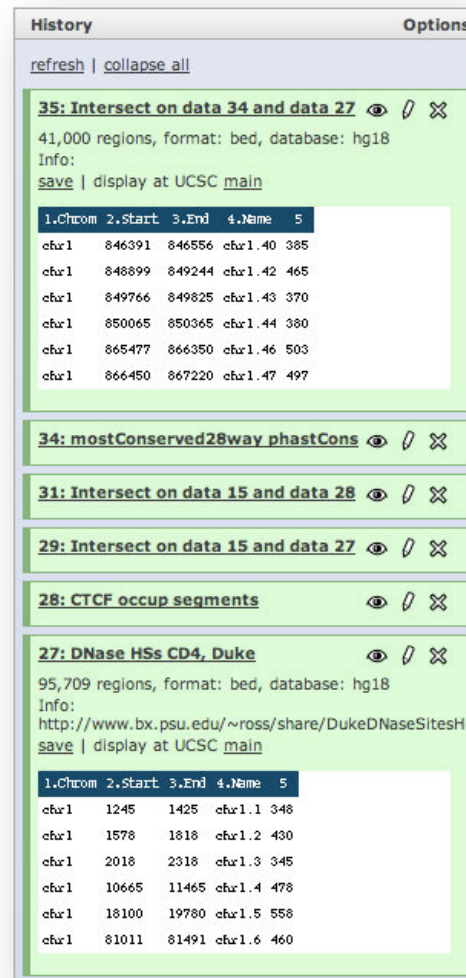


100%

[Link-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](#)

Provenance, histories

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



History Options

[refresh](#) | [collapse all](#)

35: Intersect on data 34 and data 27

41,000 regions, format: bed, database: hg18
Info:
[save](#) | [display at UCSC main](#)

1.Chrom	2.Start	3.End	4.Name	5
chr1	846391	846556	chr1.40	385
chr1	848899	849244	chr1.42	465
chr1	849766	849825	chr1.43	370
chr1	850065	850365	chr1.44	380
chr1	865477	866350	chr1.46	503
chr1	866450	867220	chr1.47	497

34: mostConserved28way phastCons

31: Intersect on data 15 and data 28

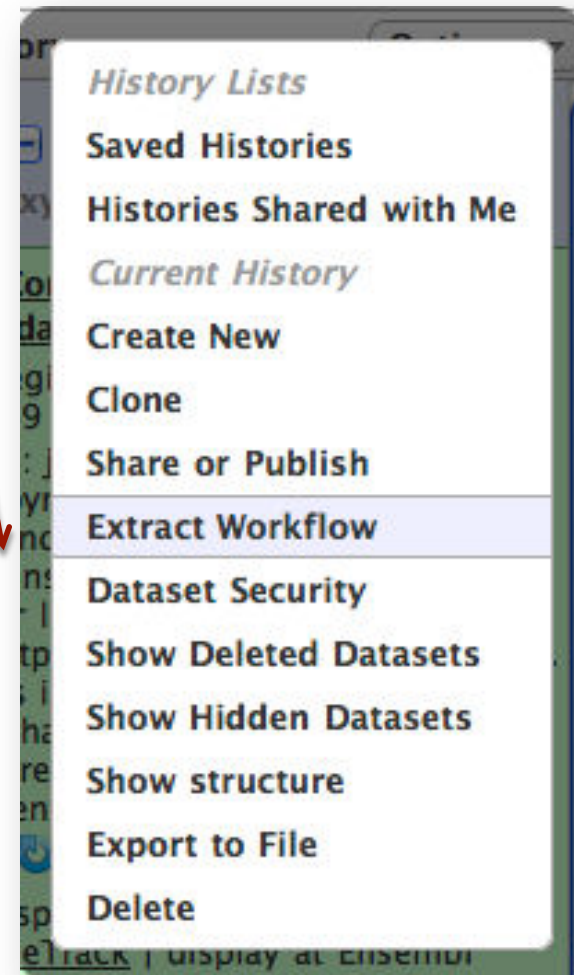
29: Intersect on data 15 and data 27

28: CTCF occup segments

27: DNase HSs CD4, Duke

95,709 regions, format: bed, database: hg18
Info:
<http://www.bx.psu.edu/~ross/share/DukeDNaseSitesH>
[save](#) | [display at UCSC main](#)

1.Chrom	2.Start	3.End	4.Name	5
chr1	1245	1425	chr1.1	348
chr1	1578	1818	chr1.2	430
chr1	2018	2318	chr1.3	345
chr1	10665	11465	chr1.4	478
chr1	18100	19780	chr1.5	558
chr1	81011	81491	chr1.6	460



Creating a workflow

The screenshot displays the Galaxy web interface at <http://main.g2.bx.psu.edu/>. The interface is divided into several sections:

- Tools:** A sidebar on the left lists various tools categorized by function, such as "Get Data", "Text Manipulation", "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", "NGS TOOLBOX BETA", "NGS: QC and manipulation", "NGS: Mapping", "NGS: SAM Tools", "NGS: Indel Analysis", "NGS: Peak Calling", "NGS: RNA Analysis", "RGENETICS", and "SNP/WGA: Data: Filters".
- Workflow Creation:** The main panel shows the process of creating a new workflow. It includes a "Workflow name" field with the value "galaxy101". Below this are buttons for "Create Workflow", "Check all", and "Uncheck all". A table lists the tools and their status in the workflow:

Tool	History items created
UCSC Main <i>This tool cannot be used in workflows</i>	1: Exons <input checked="" type="checkbox"/> Treat as input dataset
UCSC Main <i>This tool cannot be used in workflows</i>	2: SNPs <input checked="" type="checkbox"/> Treat as input dataset
Join <input checked="" type="checkbox"/> Include "Join" in workflow	3: Join on data 2 and data 1
Group <input checked="" type="checkbox"/> Include "Group" in workflow	4: Group on data 3
Sort <input checked="" type="checkbox"/> Include "Sort" in workflow	5: Sort on data 4
Select first	6: Select first on data 5

The right sidebar shows the "History" section, which lists the workflow steps and their details. The first step is "7: Compare two Queries on data 6 and data 1", which includes information about the tool (5 regions, format: bed, database: hg19), the version (Info: join (GNU coreutils) 8.5), the license (Copyright (C) 2010 Free Software Foundation, Inc. License GPLv3+: GNU GPL version 3 or later), and a link to the license page. Below this, a table shows the results of the query:

1. Chrom	2. Start	3. End	4. Name
chr22	18834444	18835833	uc002zoc
chr22	20456381	20461301	uc002zsd
chr22	21738147	21743067	uc002zuq
chr22	46652457	46659219	uc003bhh
chr22	21480536	21481925	uc010gsw

Workflow editor

The screenshot displays the Galaxy Workflow Editor interface. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Lab, Visualization, Admin, Help, and User. The left sidebar lists various tools categorized under 'Tools', 'NGS TOOLBOX BETA', and 'SCENETICS'. The central 'Workflow Canvas' shows a workflow for 'galaxy101' with steps: 'Group' (Select data: out_file1 (tabular)), 'Sort' (Sort Query: out_file1), 'Select first' (from: out_file1), and 'Compare' (Compare against: out_file1). The 'Details' panel on the right shows the configuration for the 'Select first' tool, including a 'Select first' dropdown set to 5, a 'from' field set to 'Data input 'input' (txt)', and an 'Edit Step Actions' section with 'Assign Columns' and 'out_file1' buttons. The 'Edit Step Attributes' section includes an 'Annotation / Notes' field. The 'What it does' section states: 'This tool outputs specified number of lines from the beginning of a dataset'. The 'Example' section shows a table of genomic data with columns for chromosome, start, end, and score, and a '+' sign in the last column.

Galaxy

http://main.g2.bx.psu.edu/workflow/editor?id=86beaa9c1d0aab70

Galaxy

Analyze Data Workflow Shared Data Lab Visualization Admin Help User

Tools Options Workflow Canvas | galaxy101 Options Details

Tools

- 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
- NGS TOOLBOX BETA
- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS: Indel Analysis
- NGS: Peak Calling
- NGS: RNA Analysis
- SCENETICS

Workflow Canvas | galaxy101

Group

Select data

out_file1 (tabular)

Sort

Sort Query

out_file1

Select first

from

out_file1

Compare

Compare

against

out_file1

Details

Tool: Select first

Select first ▼

5

from

Data input 'input' (txt)

Edit Step Actions

Assign Columns

out_file1

Create

Add actions to this step; actions are applied when this workflow step completes.

Edit Step Attributes

Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

What it does

This tool outputs specified number of lines from the beginning of a dataset

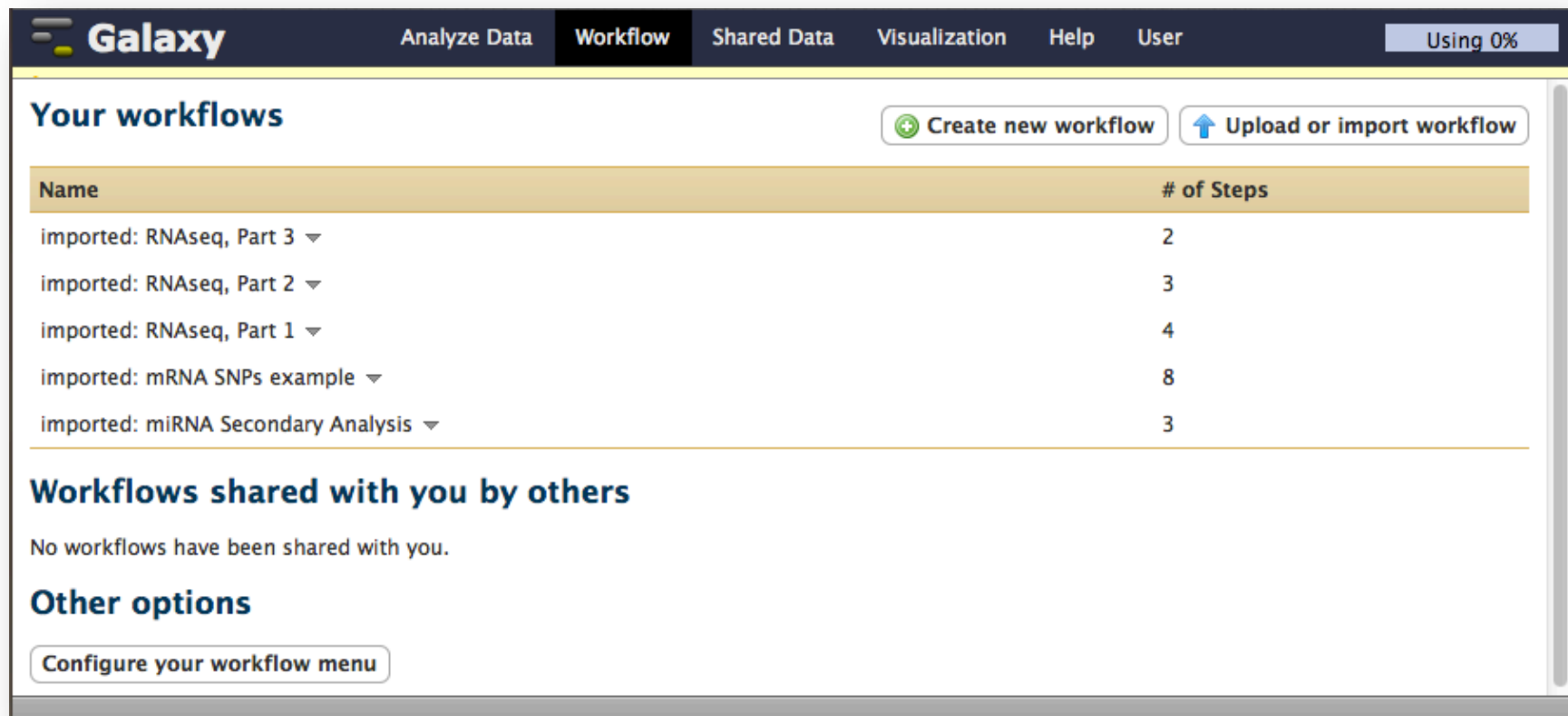
Example

Selecting 2 lines from this:

chr7	56632	56652	D17003_CTCF_R6	310	+
chr7	56736	56756	D17003_CTCF_R7	354	+

Reproducibility

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

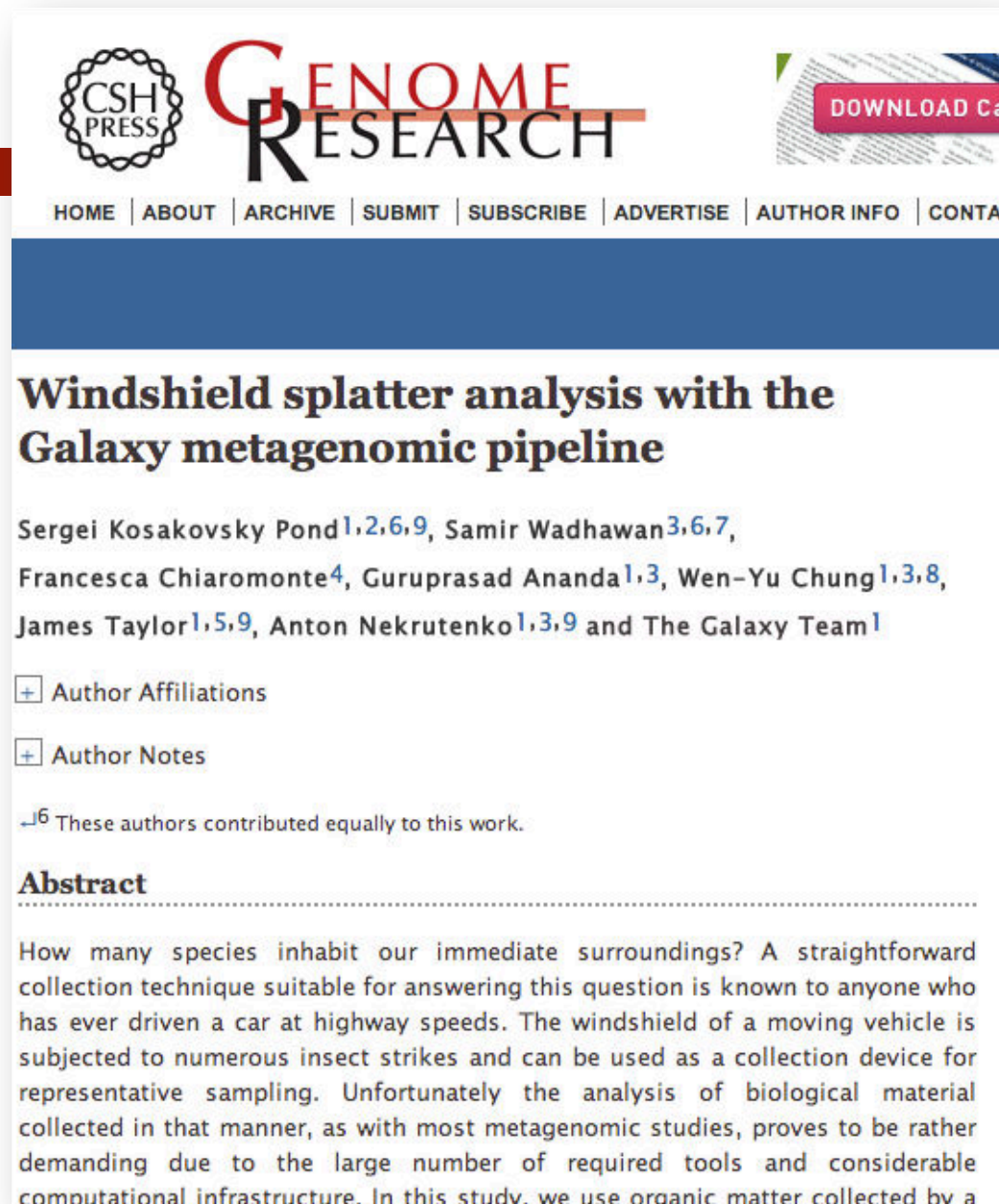


The screenshot shows the Galaxy web interface. The top navigation bar includes links for 'Analyze Data', 'Workflow' (which is active), 'Shared Data', 'Visualization', 'Help', and 'User'. A 'Using 0%' indicator is on the right. Below the navigation bar, the 'Your workflows' section is displayed. It features two buttons: 'Create new workflow' and 'Upload or import workflow'. A table lists the user's workflows with columns for 'Name' and '# of Steps'. The table contains five entries, all marked as 'imported' with a dropdown arrow. Below the table, the 'Workflows shared with you by others' section states 'No workflows have been shared with you.' and the 'Other options' section includes a button for 'Configure your workflow menu'.

Name	# of Steps
imported: RNAseq, Part 3 ▼	2
imported: RNAseq, Part 2 ▼	3
imported: RNAseq, Part 1 ▼	4
imported: mRNA SNPs example ▼	8
imported: miRNA Secondary Analysis ▼	3

Sharing

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



The screenshot shows the top of a web page for Genome Research, published by CSH Press. The header includes the CSH Press logo, the 'GENOME RESEARCH' title, and a 'DOWNLOAD' button. A navigation bar contains links: HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT. The article title is 'Windshield splatter analysis with the Galaxy metagenomic pipeline'. The authors listed are Sergei Kosakovsky Pond^{1,2,6,9}, Samir Wadhawan^{3,6,7}, Francesca Chiaromonte⁴, Guruprasad Ananda^{1,3}, Wen-Yu Chung^{1,3,8}, James Taylor^{1,5,9}, Anton Nekrutenko^{1,3,9}, and The Galaxy Team¹. There are expandable sections for 'Author Affiliations' and 'Author Notes'. A note indicates that six authors contributed equally to the work. The abstract begins with the question 'How many species inhabit our immediate surroundings?' and describes a collection technique using car windshields for insect sampling, noting the complexity of metagenomic analysis.

CSH PRESS GENOME RESEARCH

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

DOWNLOAD

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond^{1,2,6,9}, Samir Wadhawan^{3,6,7},
Francesca Chiaromonte⁴, Guruprasad Ananda^{1,3}, Wen-Yu Chung^{1,3,8},
James Taylor^{1,5,9}, Anton Nekrutenko^{1,3,9} and The Galaxy Team¹

+ 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

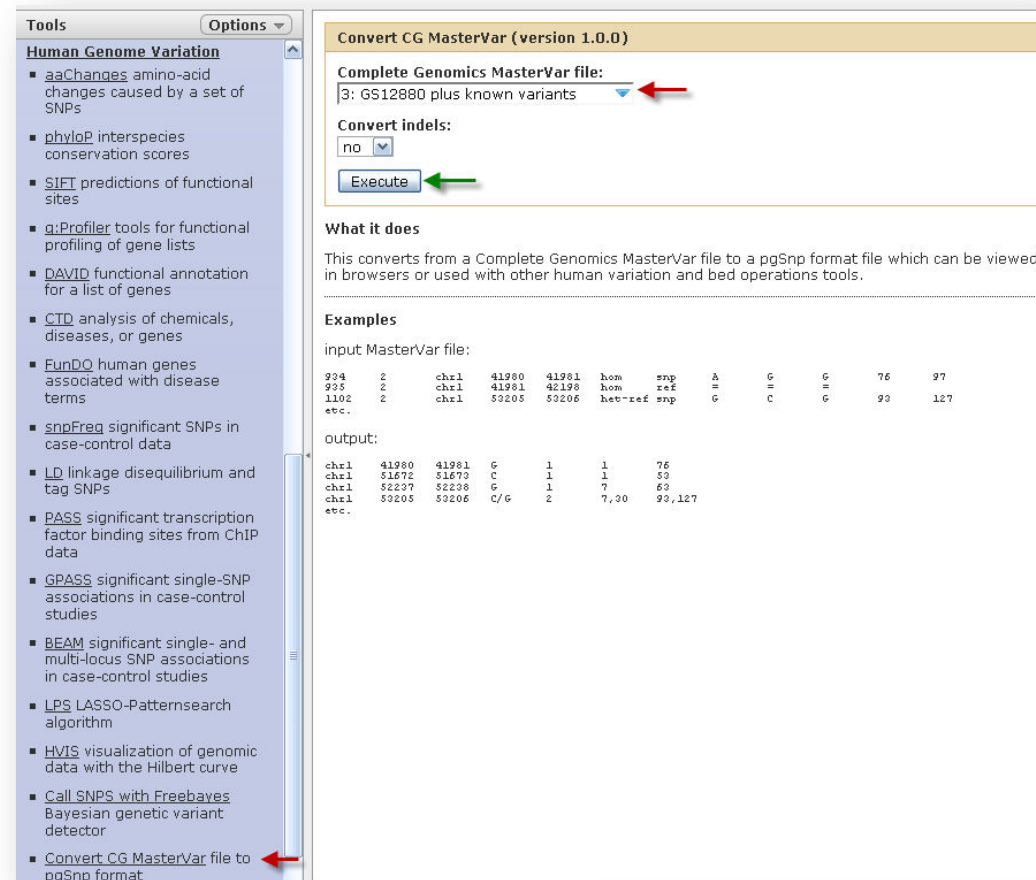
What data types, expressed in which formats, does Galaxy operate on? How do I get data into and out of Galaxy?

Data types

- Sequences
- Alignments
- Intervals
- Tabular data

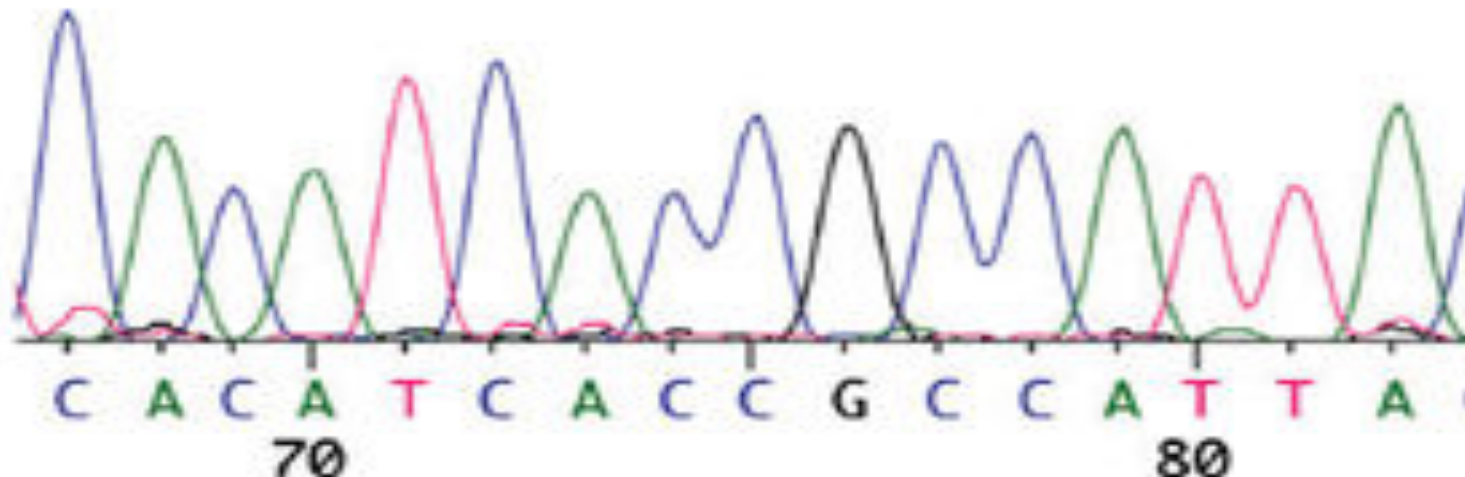
File format conversion

- 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 hgl8.chr7      3532968 27 - 158821424 ATGCTGTCCCTCTTCCCCAGCCCAGGG
s panTro2.chr7   3633994 27 - 160261443 ATGCTGTCCCTCTTCCCCAGCCCAGGG
s rheMac2.chr3   3465688 27 - 196418989 ATGCTGCCCCCTTCCCCAGCCCGGGG
s canFam2.chr16 40964059 24 - 62570175  ATGCCCCC---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

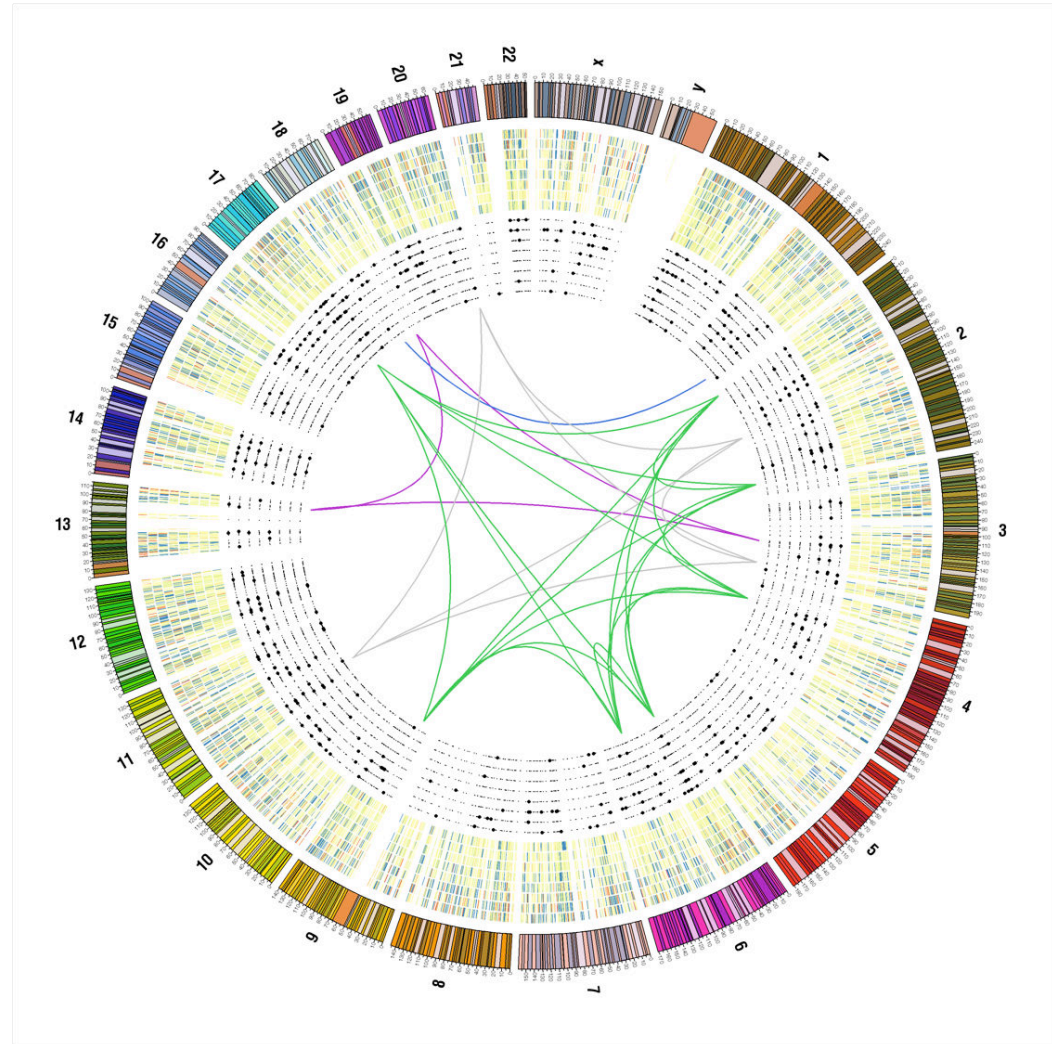
```
##maf version=1
a score=23767937.0
s hgl8.chr10     29548679 3 - 135374737 G---CG
s canFam2.chr28  24796549 6 - 44191819  GGAACA
s mm9.chr19      13586550 6 - 61342430  GGAAC
s rheMac2.chr9   29548600 6 - 133323859 G---CG
s panTro2.chr10  30467642 3 - 135001995 G---CG

a score=23767937.0
s hgl8.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 hgl8.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
<input checked="" type="checkbox"/> 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

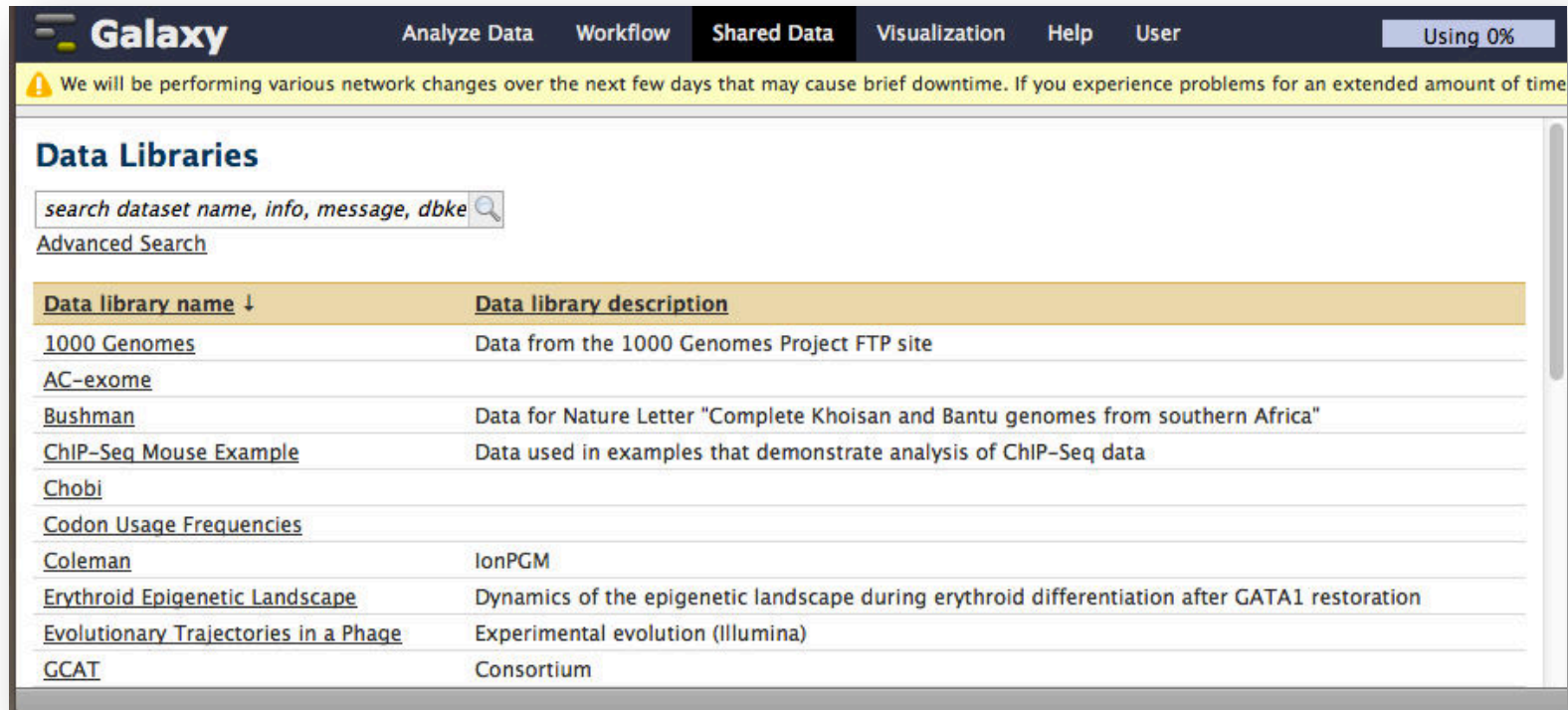
The screenshot displays the Galaxy web interface. At the top, the 'Galaxy' logo is on the left, and navigation links for 'report bugs', 'wiki', 'screencasts', and 'blog' are in the center. On the right, it shows the user is logged in as 'rch8@psu.edu' with links for 'manage' and 'logout'.

The left sidebar contains a 'Tools' section with various categories: 'Get Data' (including Upload File, UCSC Main table browser, UCSC Archaea table browser, Get Microbial Data, BioMart Central server, and EncodeDB at NHGRI), 'Get ENCODE Data', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'FASTA manipulation', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Graph/Display Data', 'Regional Variation', 'Evolution: HyPhy', 'Taxonomy manipulation', 'EMBOSS', 'Short Read Analysis', and 'Workflow (beta)'.

The main content area is titled 'Upload File'. It has two input sections: 'File:' with a 'Choose File' button and 'no file selected' text, and 'URL/Text:' with a text box containing the URL 'http://www.bx.psu.edu/~ross/share/DukeDNaseSitesHg18.bed.txt'. Below the text box is a note: 'Here you may specify a list of URLs (one per line) or paste the contents of a file.' There are three checkboxes: 'Convert spaces to tabs:' (checked 'Yes'), 'File Format:' (set to 'Auto-detect'), and 'Genome:' (set to 'Human Mar. 2006 (hg18)'). An 'Execute' button is at the bottom.

The right sidebar shows a 'History' section with a list of previous jobs. Each job entry includes a name, a refresh icon, a collapse icon, and a delete icon. The jobs listed are: '22: Erythroid preCRMs, hiRP + ccGATA1bs', '19: TAF1 sites known+novel', '15: Hi RP segments hg18', '12: UCSC Main on Human: mostConserved28way (genome)', and '11: PRPs v2 hg18'.

Import from data library

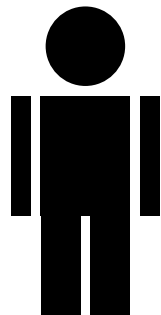


The screenshot shows the Galaxy web interface. At the top is a dark blue navigation bar with the Galaxy logo and links for Analyze Data, Workflow, Shared Data, Visualization, Help, and User. A status bar on the right indicates 'Using 0%'. Below the navigation bar is a yellow warning banner. The main content area is titled 'Data Libraries' and features a search bar with the placeholder text 'search dataset name, info, message, dbke' and a magnifying glass icon. Below the search bar is a link for 'Advanced Search'. A table lists various data libraries with two columns: 'Data library name' and 'Data library description'.

Data library name ↓	Data library description
1000 Genomes	Data from the 1000 Genomes Project FTP site
AC-exome	
Bushman	Data for Nature Letter "Complete Khoisan and Bantu genomes from southern Africa"
ChIP-Seq Mouse Example	Data used in examples that demonstrate analysis of ChIP-Seq data
Chobi	
Codon Usage Frequencies	
Coleman	IonPGM
Erythroid Epigenetic Landscape	Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration
Evolutionary Trajectories in a Phage	Experimental evolution (Illumina)
GCAT	Consortium

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



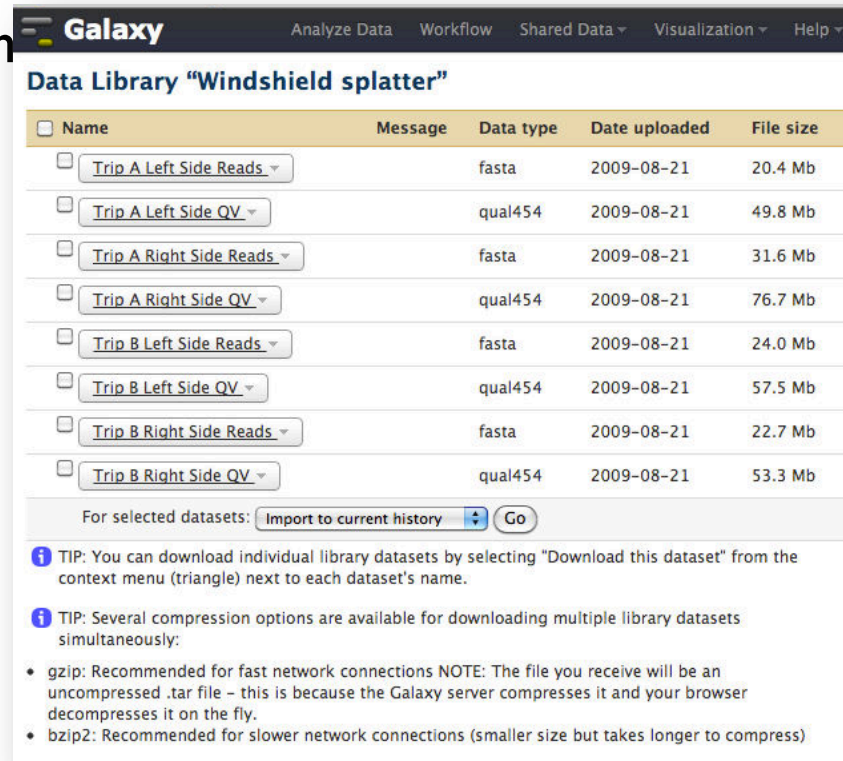
“Big Data”

- ❑ 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*



The screenshot shows the Galaxy Data Library interface for a dataset named "Windshield splatter". The interface includes a table with columns for Name, Message, Data type, Date uploaded, and File size. The table lists eight datasets, each with a checkbox for selection and a dropdown menu for actions. Below the table, there is a section for selected datasets with an "Import to current history" button and a "Go" button. Two tips are provided: one about downloading individual datasets and another about compression options for multiple datasets.

<input type="checkbox"/>	Name	Message	Data type	Date uploaded	File size
<input type="checkbox"/>	Trip A Left Side Reads ▾		fasta	2009-08-21	20.4 Mb
<input type="checkbox"/>	Trip A Left Side QV ▾		qual454	2009-08-21	49.8 Mb
<input type="checkbox"/>	Trip A Right Side Reads ▾		fasta	2009-08-21	31.6 Mb
<input type="checkbox"/>	Trip A Right Side QV ▾		qual454	2009-08-21	76.7 Mb
<input type="checkbox"/>	Trip B Left Side Reads ▾		fasta	2009-08-21	24.0 Mb
<input type="checkbox"/>	Trip B Left Side QV ▾		qual454	2009-08-21	57.5 Mb
<input type="checkbox"/>	Trip B Right Side Reads ▾		fasta	2009-08-21	22.7 Mb
<input type="checkbox"/>	Trip B Right Side QV ▾		qual454	2009-08-21	53.3 Mb

For selected datasets:

TIP: You can download individual library datasets by selecting "Download this dataset" from the context menu (triangle) next to each dataset's name.

TIP: Several compression options are available for downloading multiple library datasets simultaneously:

- gzip: Recommended for fast network connections NOTE: The file you receive will be an uncompressed .tar file – this is because the Galaxy server compresses it and your browser decompresses it on the fly.
- bzip2: Recommended for slower network connections (smaller size but takes longer to compress)

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

Tools

Which operations can I run on Galaxy?

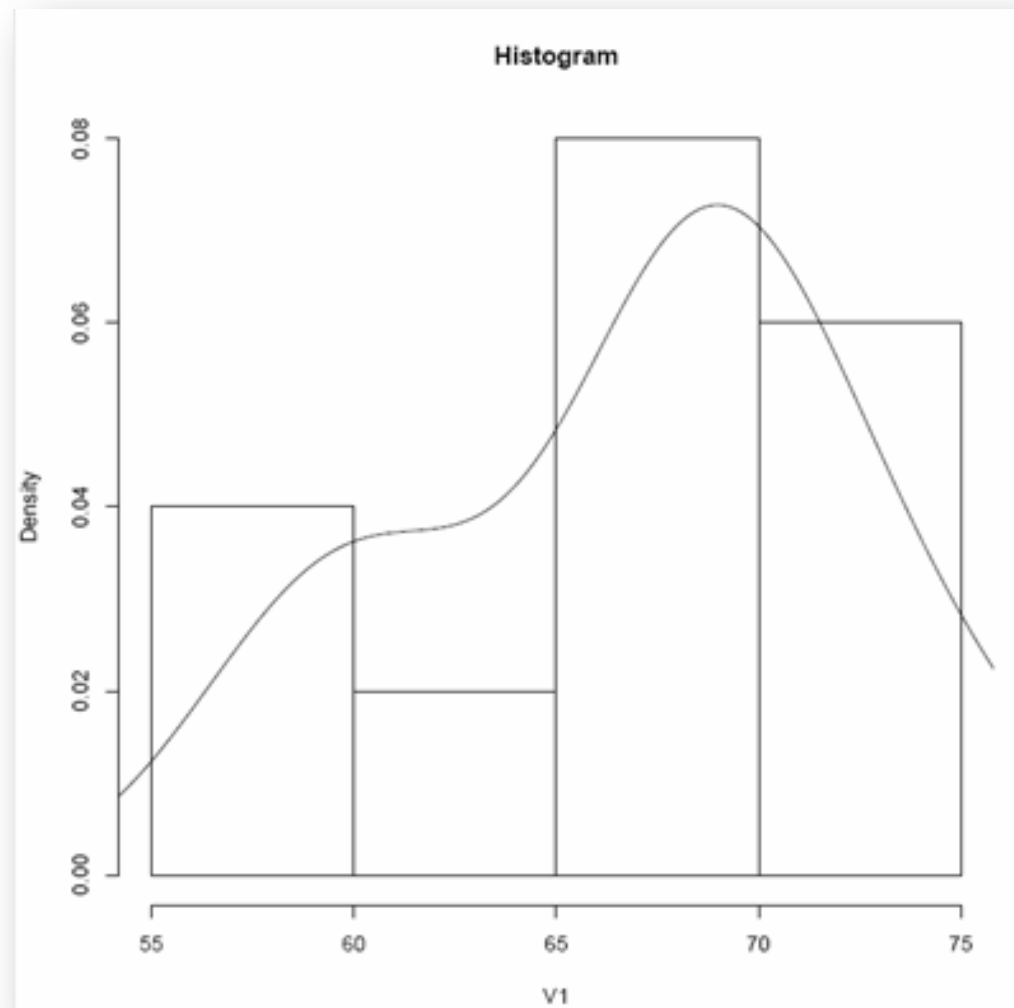
Galaxy tools

- ❑ **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*

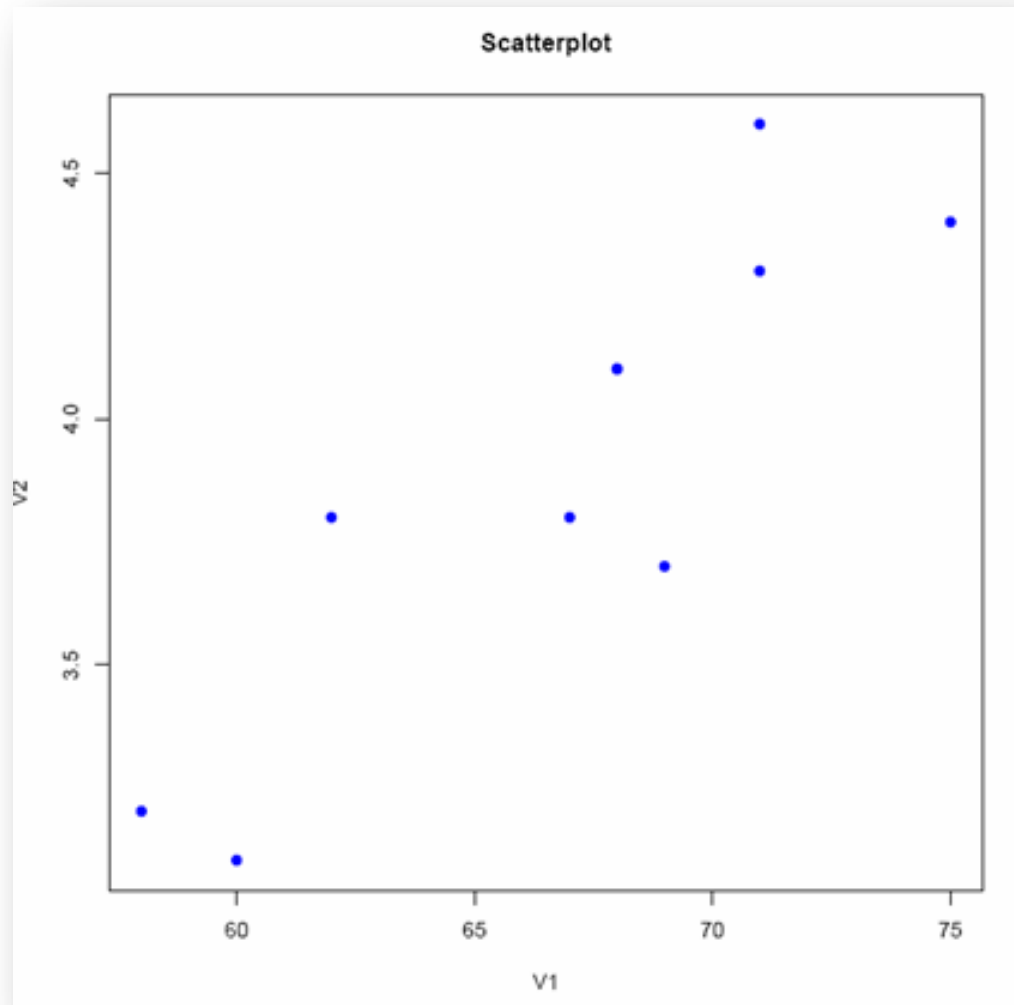
Visualization

What data can I view in Galaxy?

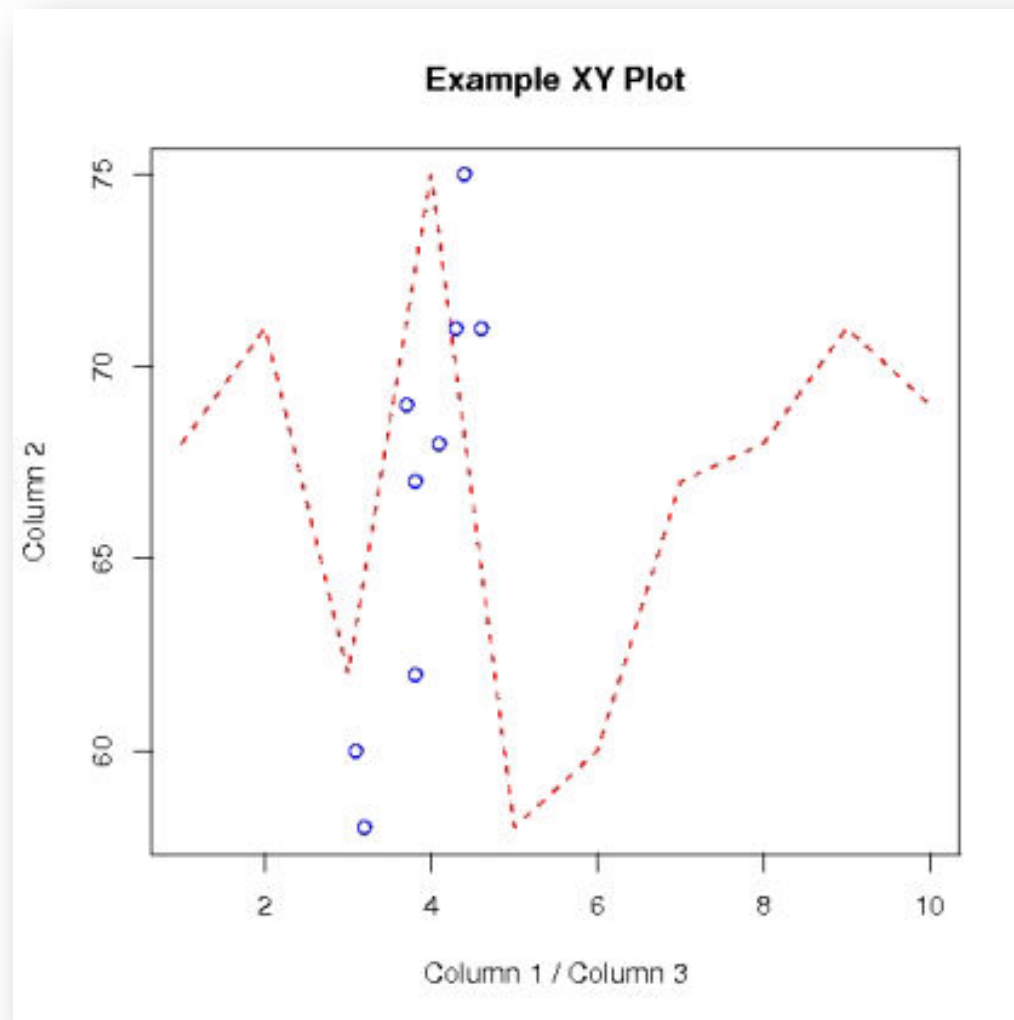
Galaxy visualization - histograms



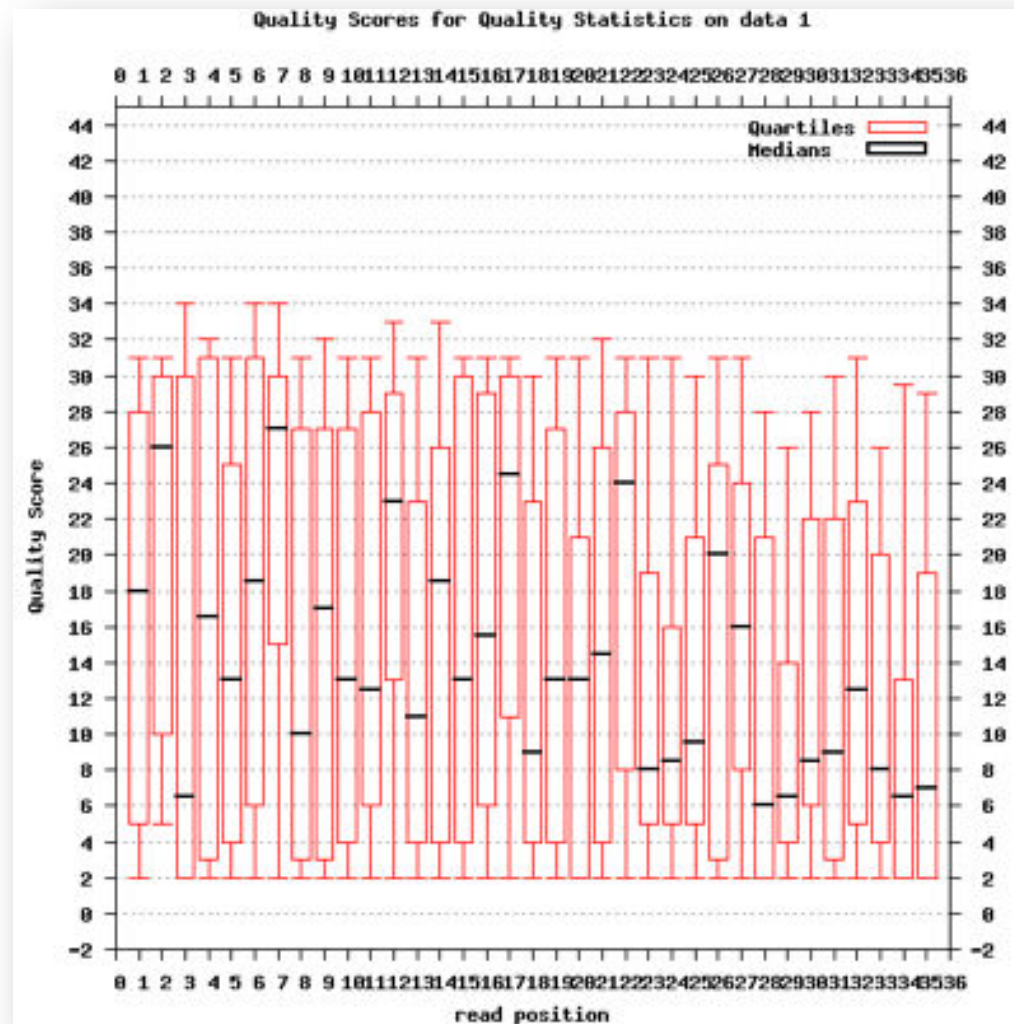
Galaxy visualization - scatterplots



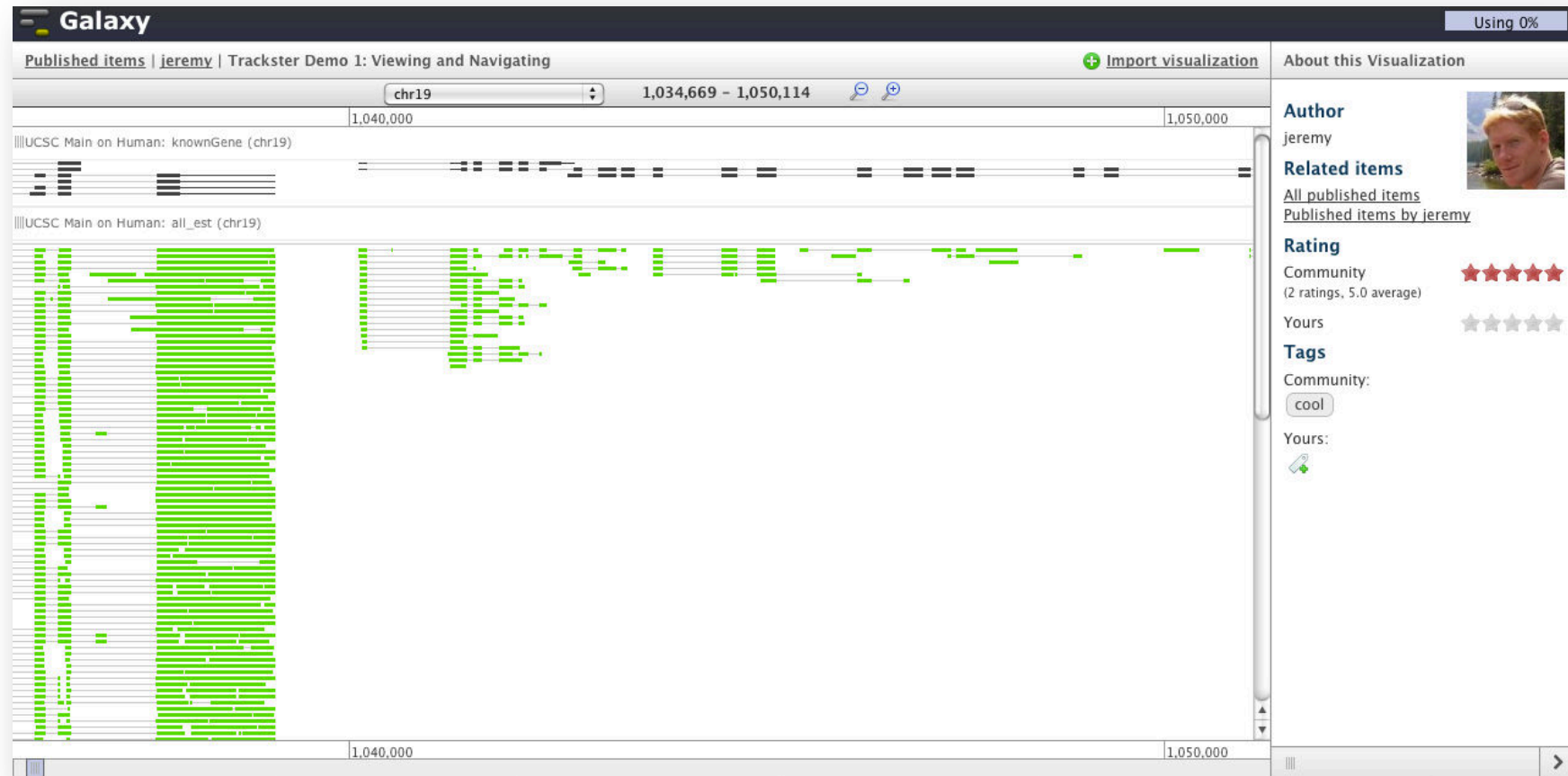
Galaxy visualization – XY plot



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:

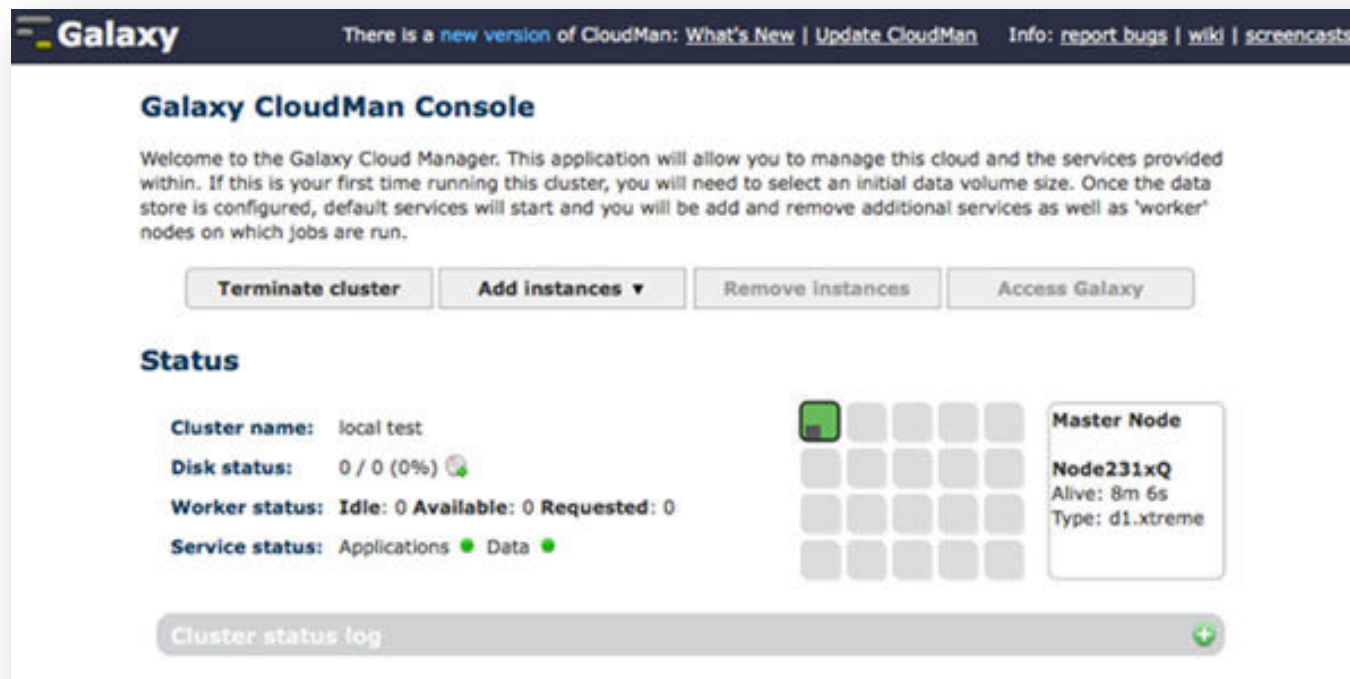
- ❑ Data close by
- ❑ Can add your own tools
- ❑ Can develop Galaxy further
- ❑ UNIX-based

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



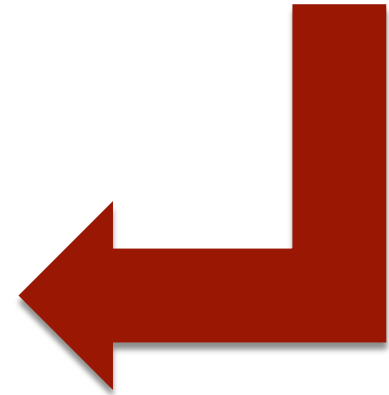
Implementation

How does it work under the hood?

Galaxy under the hood

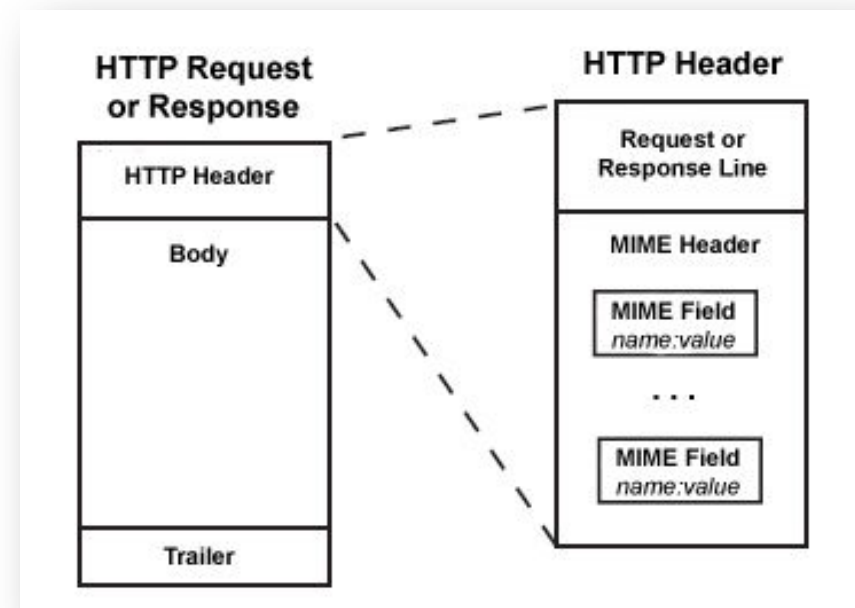


1. Parses HTTP request
2. Identifies which tool to use
3. Reads tool description
4. Queues tool
5. Parses result
6. Returns HTML representation of result



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

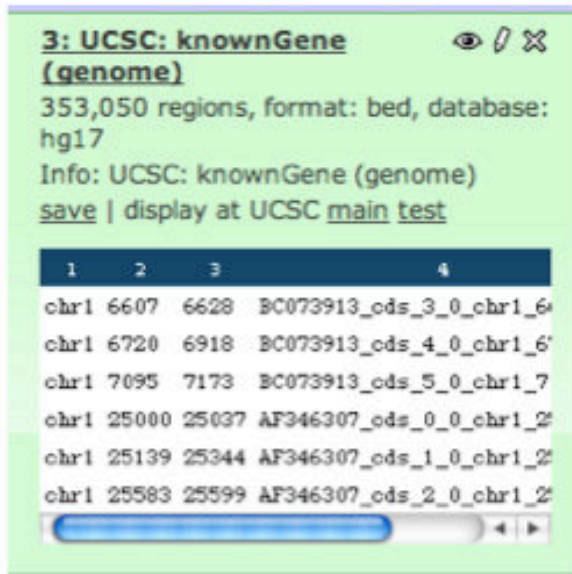
Queued:



Running:



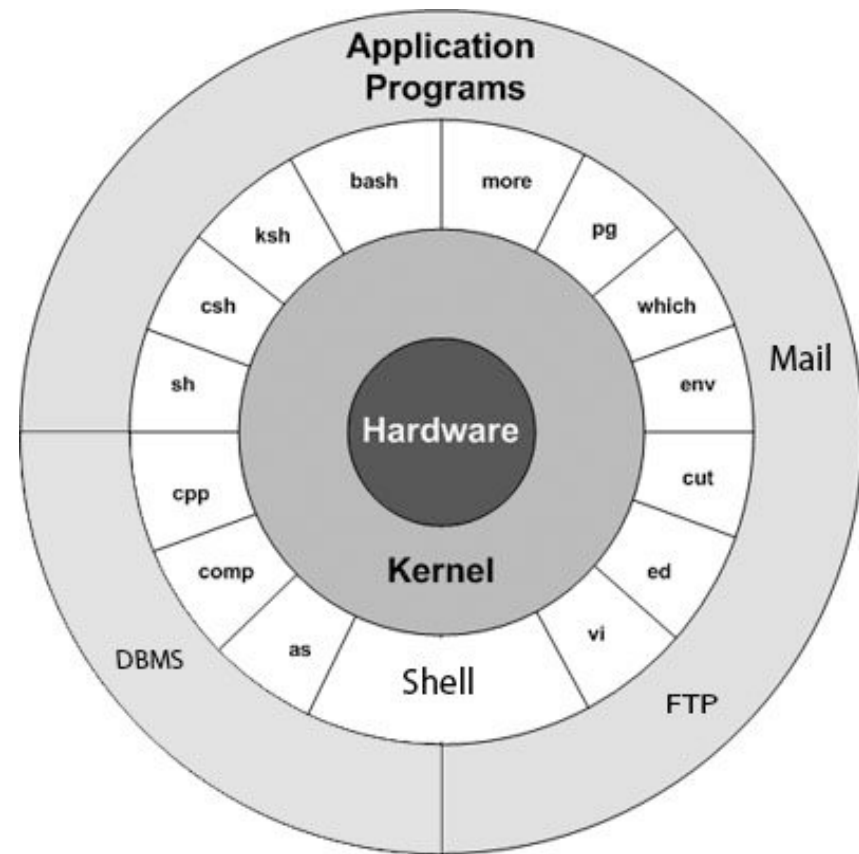
Complete:



- ❑ 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



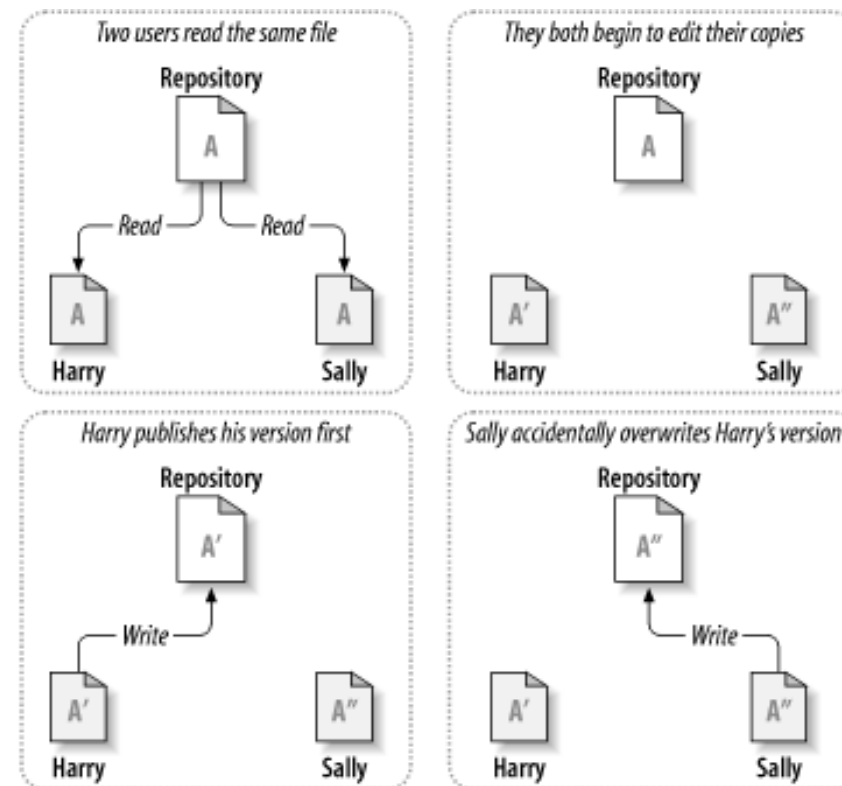
Configuration

- ❑ 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



The screenshot shows the Galaxy Wiki homepage. At the top is a dark blue header with the 'Galaxy Wiki' logo on the left and 'Login | Search:' with a text input field on the right. Below the header is a light gray bar with the text 'Galaxy Project'. The main content area is white and contains a paragraph about the hub page, a paragraph about funding, and a bulleted list of links. At the bottom, there is a 'Galaxy web search' button and a link to 'Search all Galaxy resources'.

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](#)

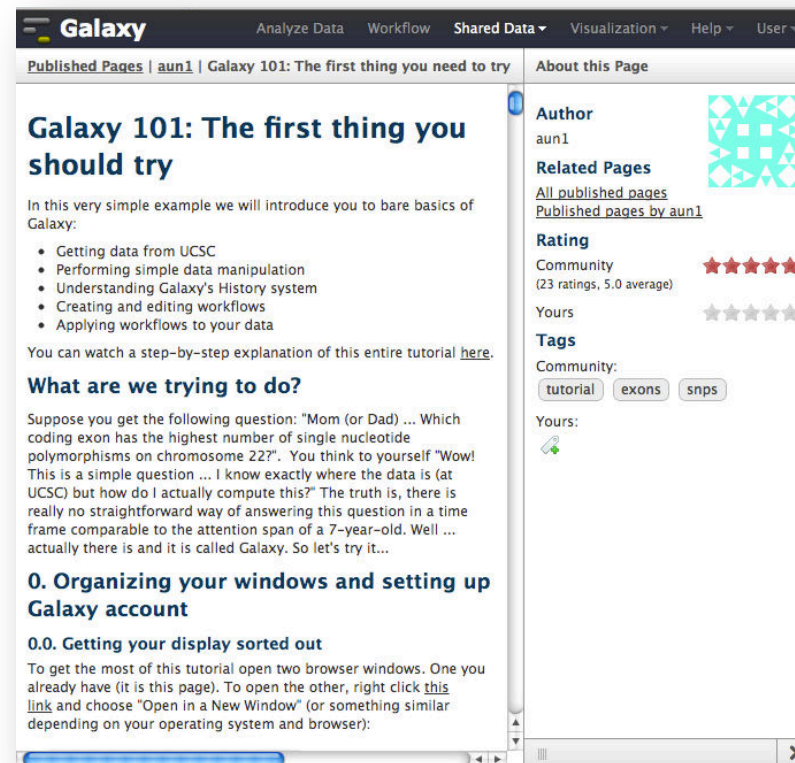
Mailing lists

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

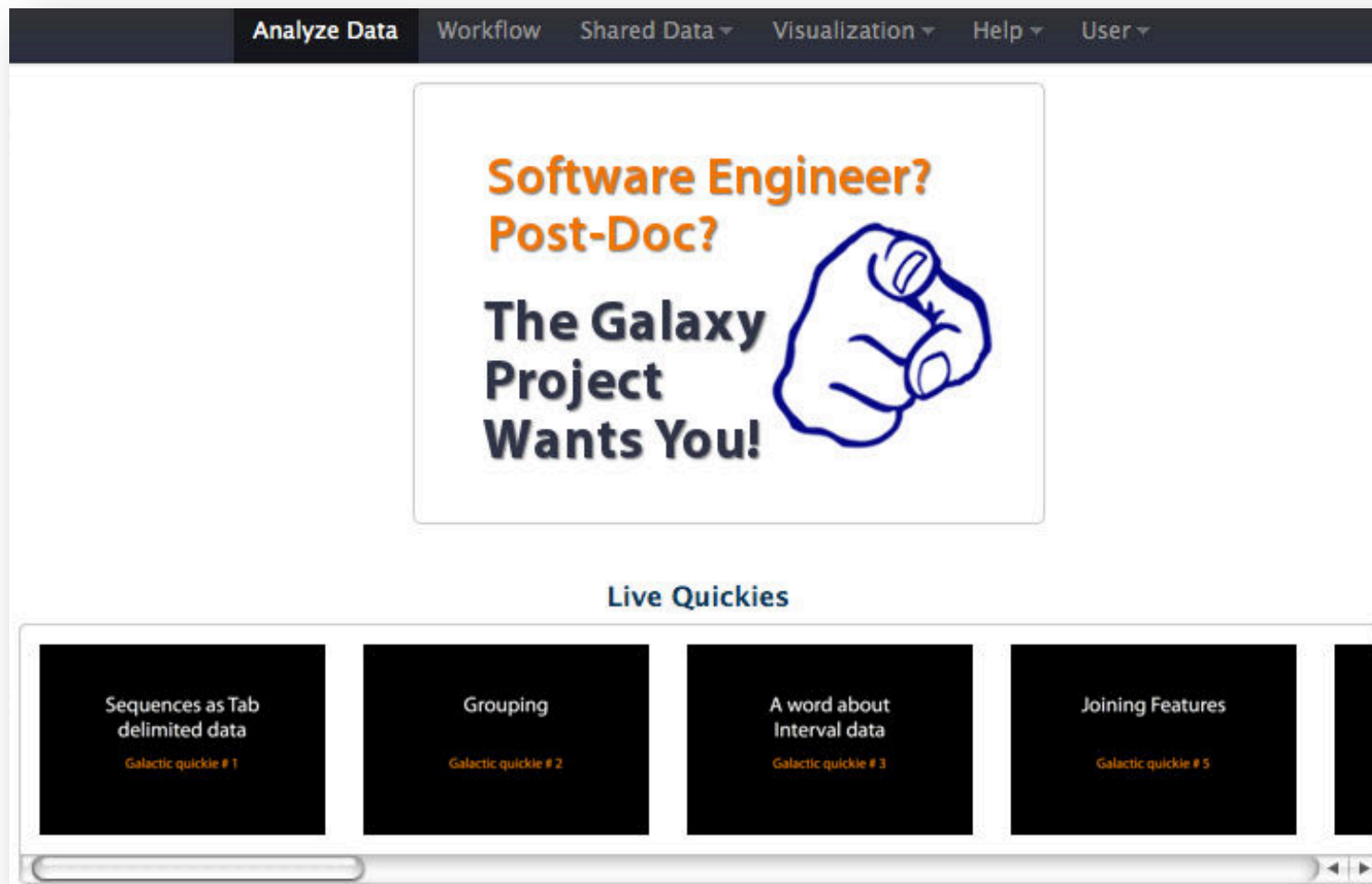


Tutorials

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

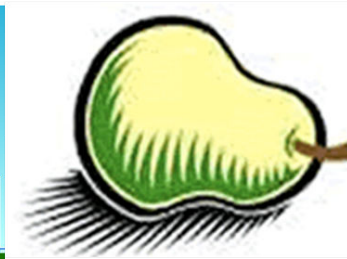


Screencasts



Events

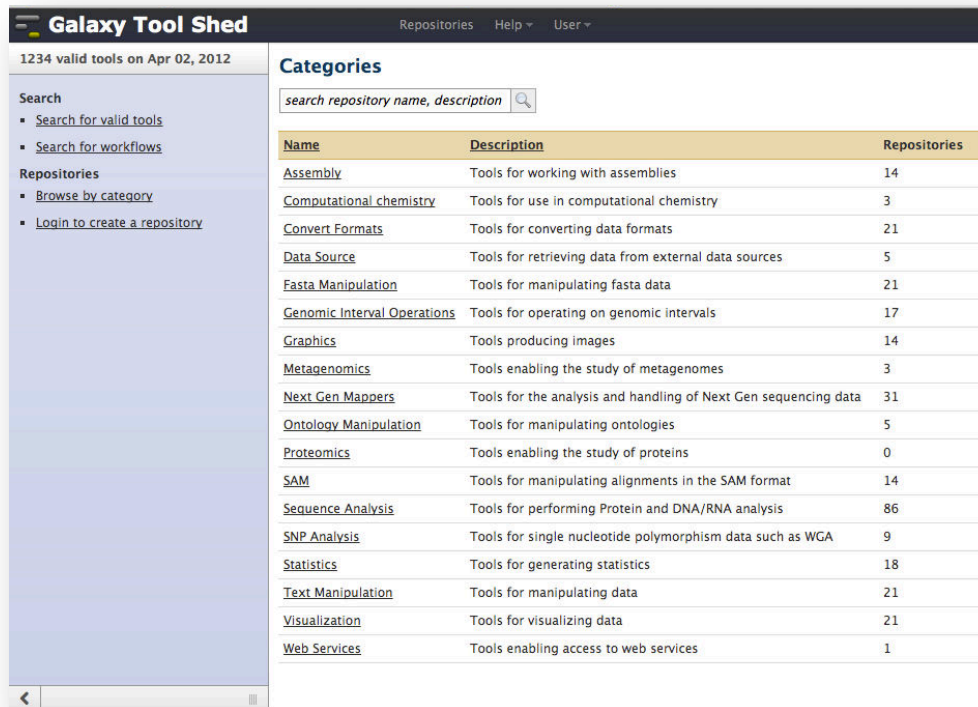
- Galaxy Community Conference
- ISMB
- ECCB
- BOSC
- PAG
- Bio-IT World
- GMOD Meetings



Bio-IT World



“Tool shed”



The screenshot displays the Galaxy Tool Shed web interface. The header includes the title "Galaxy Tool Shed" and navigation links for "Repositories", "Help", and "User". A sidebar on the left contains a search bar and links for "Search for valid tools", "Search for workflows", "Repositories", "Browse by category", and "Login to create a repository". The main content area is titled "Categories" and features a search bar for repository names and descriptions. Below this is a table listing various tool categories and the number of repositories for each.

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:

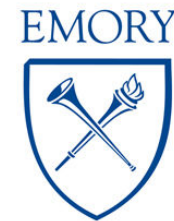
- ▣ Penn State University
- ▣ Emory University

- Support:

- ▣ NSF
- ▣ NHGRI

- Power user:

- ▣ NBIC



National Human
Genome Research
Institute



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*