Scaling Galaxy for Big Data

Preparing for those next few orders of magnitude

NGS Data after the Gold Rush
The Genome Analysis Centre (TGAC)
Norwich, United Kingdom

7 May 2014

Dave Clements (and the Galaxy Team)
Johns Hopkins University

TGAC The Genome Analysis Centre™

http://galaxyproject.org/

Galaxy

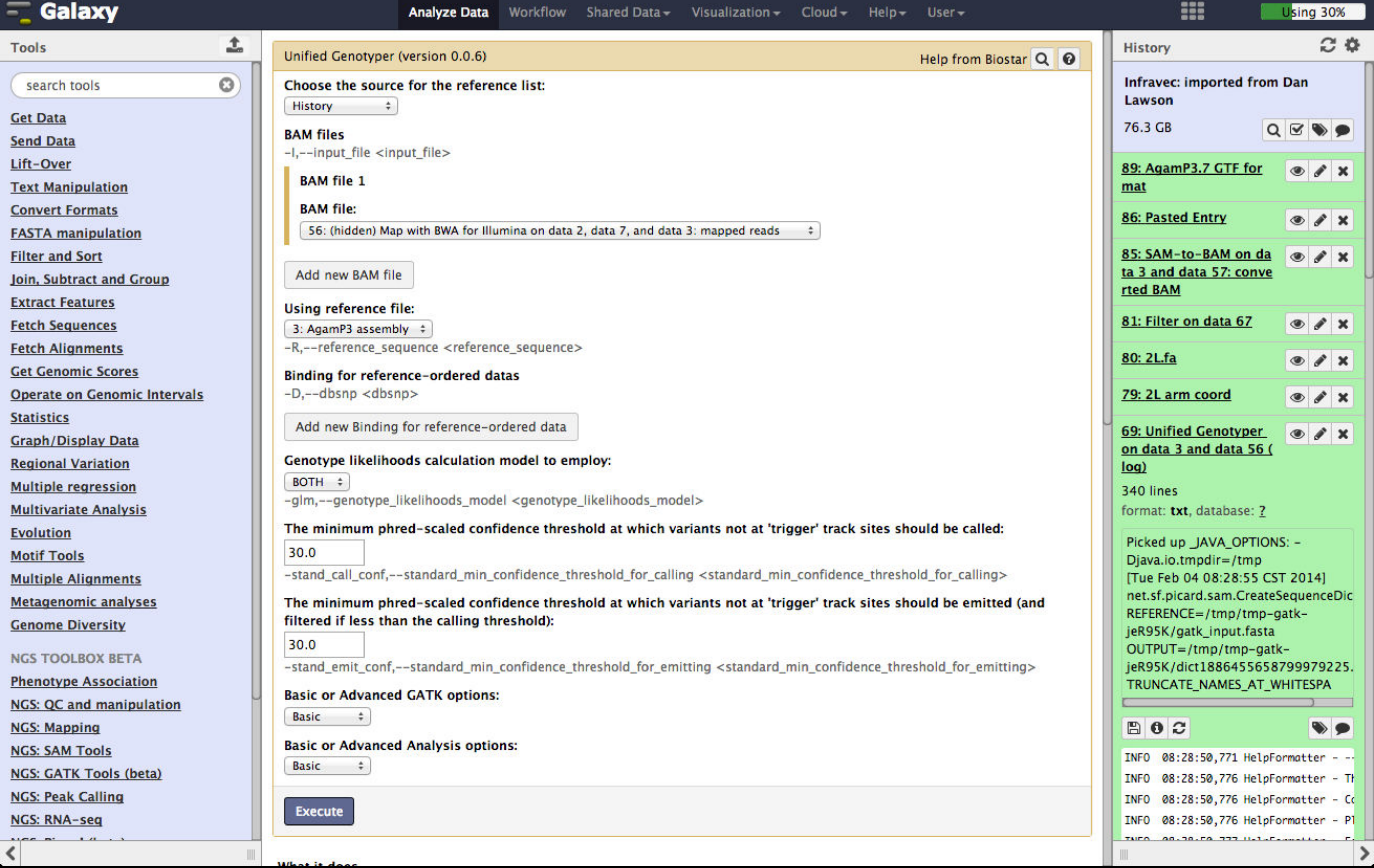
Big ...

Data generation is cheap and will stay cheap.

Scale & complexity of analysis will continue to grow.

More researchers are running bioinformatics analyses of all scales and complexities.

Data generation never sleeps



Unified Genotyper (version 0.0.6)

Choose the source for the reference list: History

BAM files
-I,--input_file <input_file>

BAM file 1

BAM file: 56: (hidden) Map with BWA for Illumina on data 2, data 7, and data 3: mapped reads

Using reference file: 3: AgamP3 assembly
-R,--reference_sequence <reference_sequence>

Binding for reference-ordered datas
-D,--dbsnp <dbsnp>

Genotype likelihoods calculation model to employ: BOTH
-glm,--genotype_likelihoods_model <genotype_likelihoods_model>

The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be called: 30.0
-stand_call_conf,--standard_min_confidence_threshold_for_calling <standard_min_confidence_threshold_for_calling>

The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be emitted (and filtered if less than the calling threshold): 30.0
-stand_emit_conf,--standard_min_confidence_threshold_for_emitting <standard_min_confidence_threshold_for_emitting>

Basic or Advanced GATK options: Basic

Basic or Advanced Analysis options: Basic

Infravec: imported from Dan Lawson

76.3 GB

89: AgamP3.7 GTF format

86: Pasted Entry

85: SAM-to-BAM on data 3 and data 57: converted BAM

81: Filter on data 67

80: 2L.fa

79: 2L arm coord

69: Unified Genotyper on data 3 and data 56 (log)

340 lines

format: txt, database: ?

```
Picked up _JAVA_OPTIONS: -
Djava.io.tmpdir=/tmp
[Tue Feb 04 08:28:55 CST 2014]
net.sf.picard.sam.CreateSequenceDic
REFERENCE=/tmp/tmp-gatk-
jeR95K/gatk_input.fasta
OUTPUT=/tmp/tmp-gatk-
jeR95K/dict1886455658799979225.
TRUNCATE_NAMES_AT_WHITESPA
```

```
INFO  08:28:50,771 HelpFormatter - --
INFO  08:28:50,776 HelpFormatter - Th
INFO  08:28:50,776 HelpFormatter - Co
INFO  08:28:50,776 HelpFormatter - Pl
```

usegalaxy.org

Showing the Infravec history published by Dan Lawson

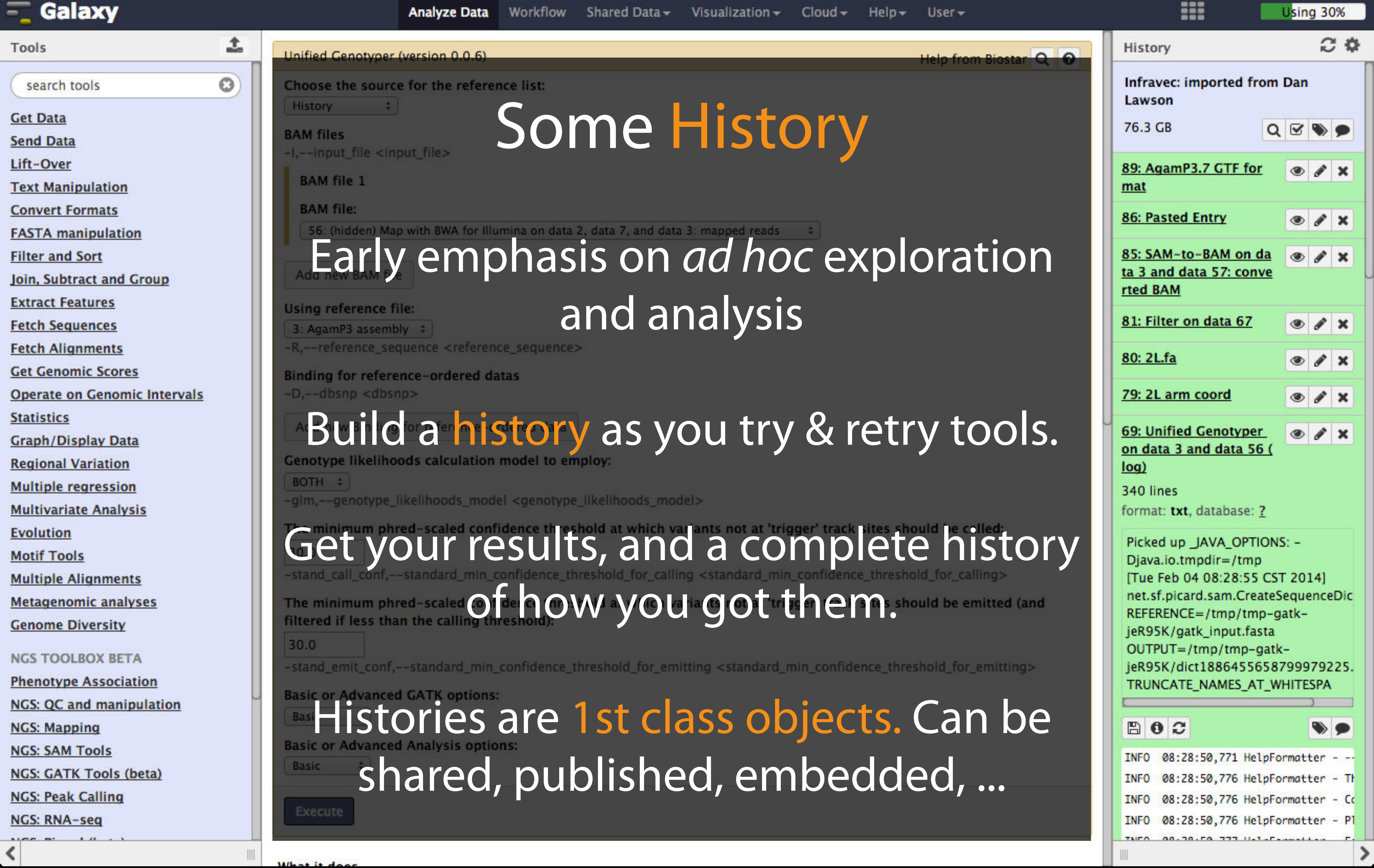
Some History

Early emphasis on *ad hoc* exploration and analysis

Build a history as you try & retry tools.

Get your results, and a complete history of how you got them.

Histories are 1st class objects. Can be shared, published, embedded, …



Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User | Using 30%

Tools

search tools

Get Data
Send Data
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
Motif Tools
Multiple Alignments
Metagenomic analyses
Genome Diversity

NGS TOOLBOX BETA

Phenotype Association
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: GATK Tools (beta)
NGS: Peak Calling
NGS: RNA-seq

Unified Genotyper (version 0.0.6) | Help from Biostar

Choose the source for the reference list: History

BAM files
-I,--input_file <input_file>

BAM file 1

BAM file: 56: (hidden) Map with BWA for Illumina on data 2, data 7, and data 3: mapped reads

Add new BAM file

Using reference file: 3: AgamP3 assembly
-R,--reference_sequence <reference_sequence>

Binding for reference-ordered datas
-D,--dbsnp <dbsnp>

Genotype likelihoods calculation model to employ: BOTH
-glm,--genotype_likelihoods_model <genotype_likelihoods_model>

The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be called:
-stand_call_conf,--standard_min_confidence_threshold_for_calling <standard_min_confidence_threshold_for_calling>

The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be emitted (and filtered if less than the calling threshold): 30.0
-stand_emit_conf,--standard_min_confidence_threshold_for_emitting <standard_min_confidence_threshold_for_emitting>

Basic or Advanced GATK options: Basic

Basic or Advanced Analysis options: Basic

Execute

What it does

History

Infravec: imported from Dan Lawson
76.3 GB

89: AgamP3.7 GTF format
86: Pasted Entry
85: SAM-to-BAM on data 3 and data 57: converted BAM
81: Filter on data 67
80: 2L.fa
79: 2L arm coord
69: Unified Genotyper on data 3 and data 56 (log)
340 lines
format: txt, database: ?

```
Picked up _JAVA_OPTIONS: -Djava.io.tmpdir=/tmp
[Tue Feb 04 08:28:55 CST 2014] net.sf.picard.sam.CreateSequenceDic
REFERENCE=/tmp/tmp-gatk-jeR95K/gatk_input.fasta
OUTPUT=/tmp/tmp-gatk-jeR95K/dict1886455658799979225.
TRUNCATE_NAMES_AT_WHITESPA
```

```
INFO  08:28:50,771 HelpFormatter - --
INFO  08:28:50,776 HelpFormatter - Th
INFO  08:28:50,776 HelpFormatter - Co
INFO  08:28:50,776 HelpFormatter - Pl
```

Traditional Strengths

ad hoc learning and exploration

Protect bench scientists from command line interfaces, programming, Unix/Linux system administration

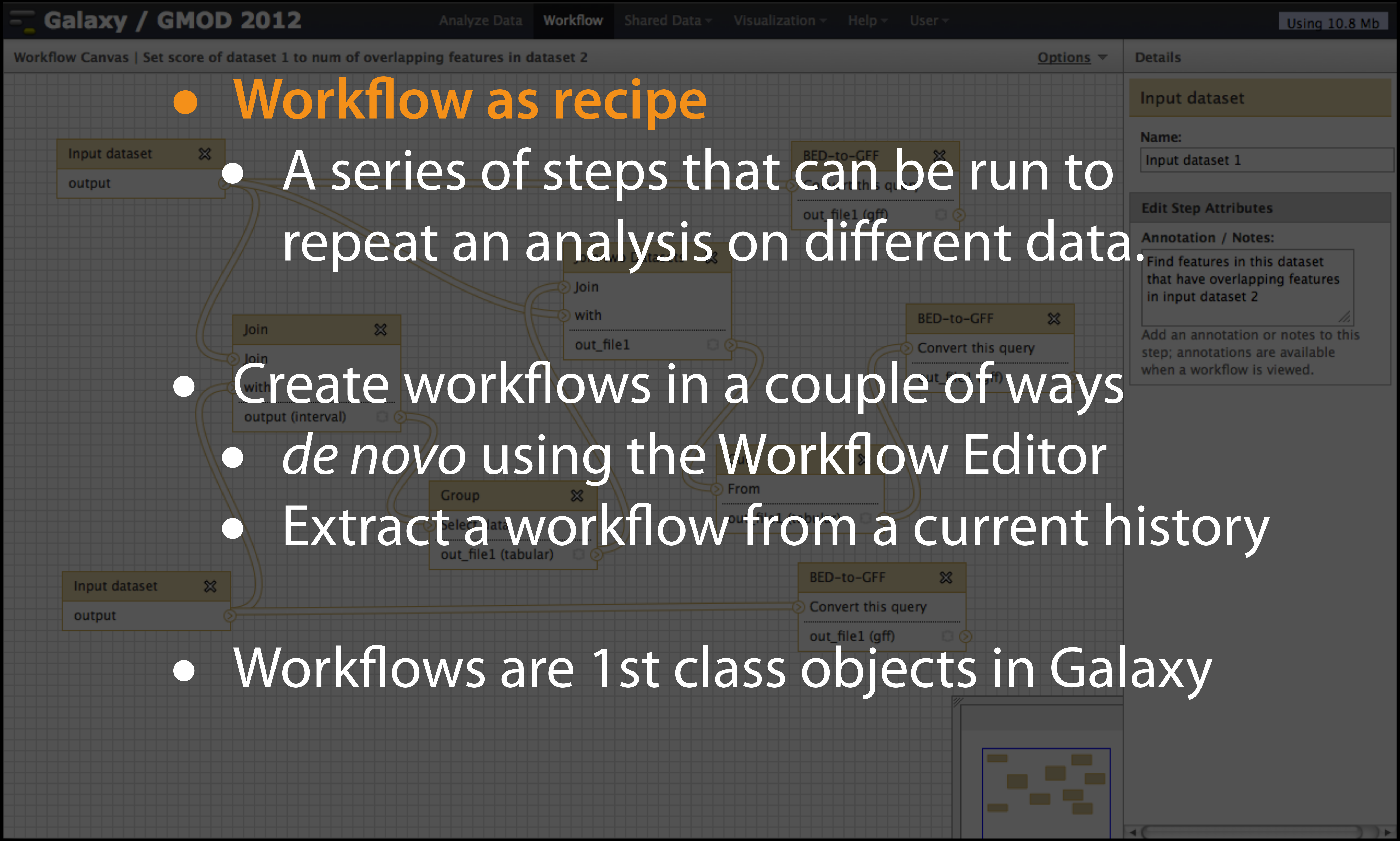
Sharing and reproducibility

Consistent and easy to use web interface

Extensible tool interface to incorporate tools

Along Came Workflows

- Workflow as recipe
  - A series of steps that can be run to repeat an analysis on different data.
- Create workflows in a couple of ways
  - *de novo* using the Workflow Editor
  - Extract a workflow from a current history
- Workflows are 1st class objects in Galaxy



Galaxy / GMOD 2012

Using 10.8 Mb

Workflow Canvas | Set score of dataset 1 to num of overlapping features in dataset 2

Details

Input dataset

Name: Input dataset 1

Edit Step Attributes

Annotation / Notes: Find features in this dataset that have overlapping features in input dataset 2

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

Some Workflow Extensions

Enable hiding of intermediate datasets

Imagine running a 25 step workflow on 20 samples.

Support for linking datasets

Avoids having to start a workflow
20 times, to process 20 samples

Still, a simple concept of workflow

The Challenge

Solutions for *ad hoc* learning and experimenting solve different issues than do solutions that make very large analyses understandable and manageable.

Can these scalability challenges be addressed without sacrificing existing strengths?

Approaches

The problem needs to be attacked from both the client side interface (the front end) and the server side implementation (the back end)

User Interface

Basics

Dynamic smooth, user interface

Provide data progressively on demand

Many fewer page loads

Better search mechanisms and scalable interfaces

Often implemented by switching from static HTML to Javascript

User Interface

Visualization

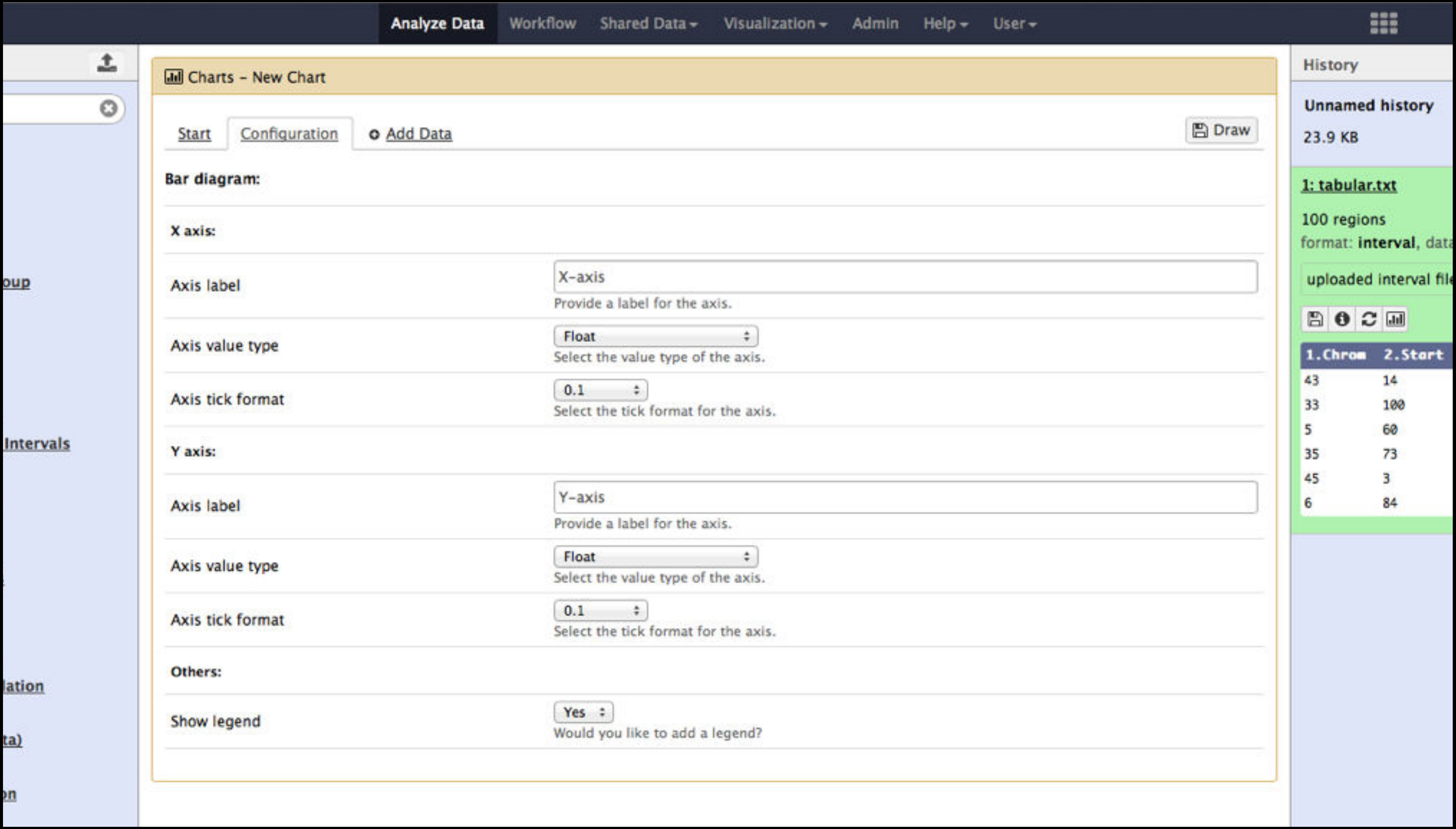
Web based visualiztion for high-throughput biology is a challenge.

Requires client side, modular, scalable components

General visualization framework implemented

Visualizations are 1st class objects

Charts



Charts - New Chart

Bar diagram:

X axis:

Axis label: X-axis
Provide a label for the axis.

Axis value type: Float
Select the value type of the axis.

Axis tick format: 0.1
Select the tick format for the axis.

Y axis:

Axis label: Y-axis
Provide a label for the axis.

Axis value type: Float
Select the value type of the axis.

Axis tick format: 0.1
Select the tick format for the axis.

Others:

Show legend: Yes
Would you like to add a legend?

Unnamed history

23.9 KB

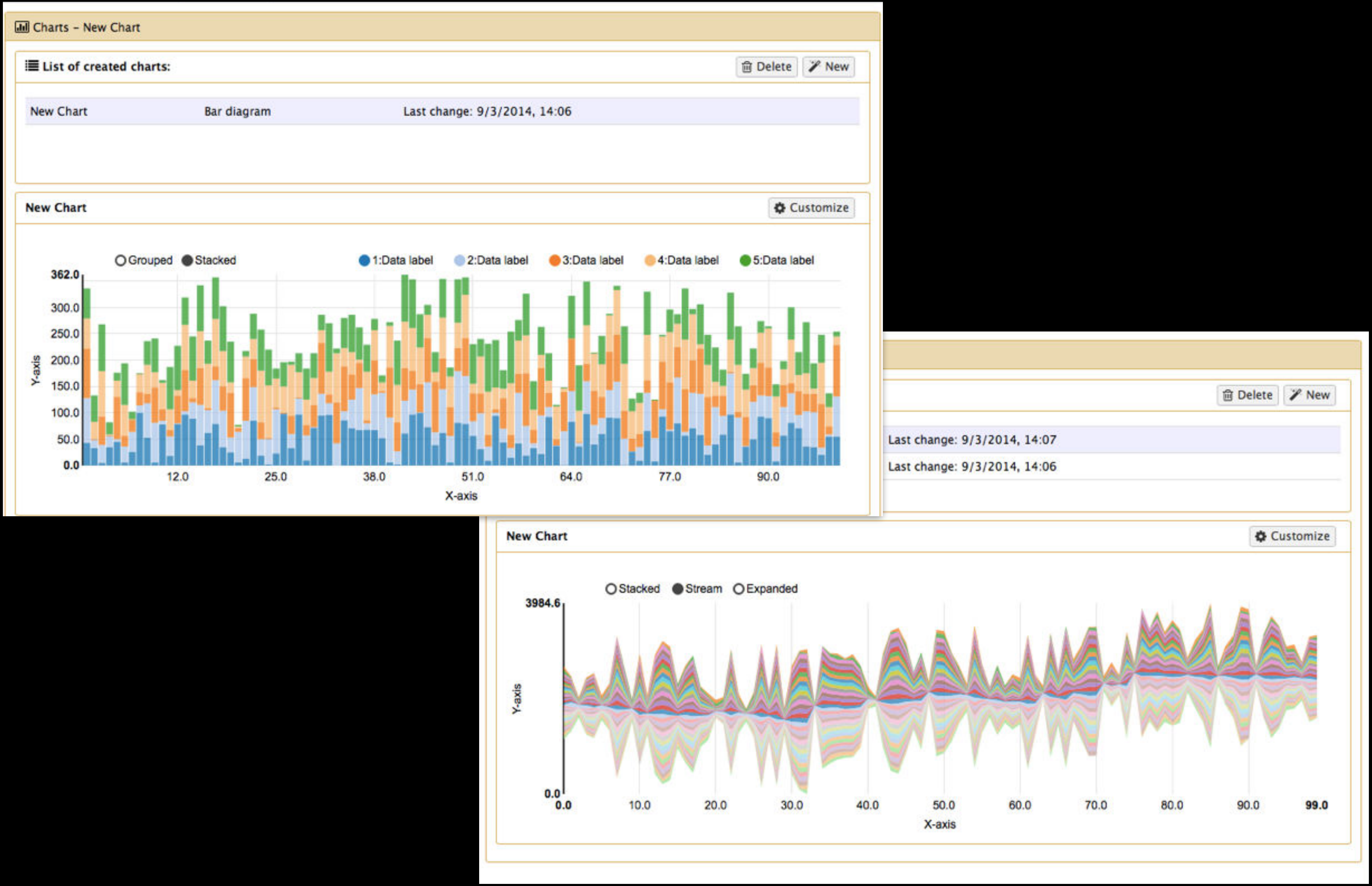
1: tabular.txt

100 regions
format: **interval**, data

uploaded interval fil

1.Chrom	2.Start
43	14
33	100
5	60
35	73
45	3
6	84

Charts



Charts – New Chart

List of created charts:

New Chart   Bar diagram   Last change: 9/3/2014, 14:06

New Chart

○ Grouped ● Stacked

Last change: 9/3/2014, 14:07

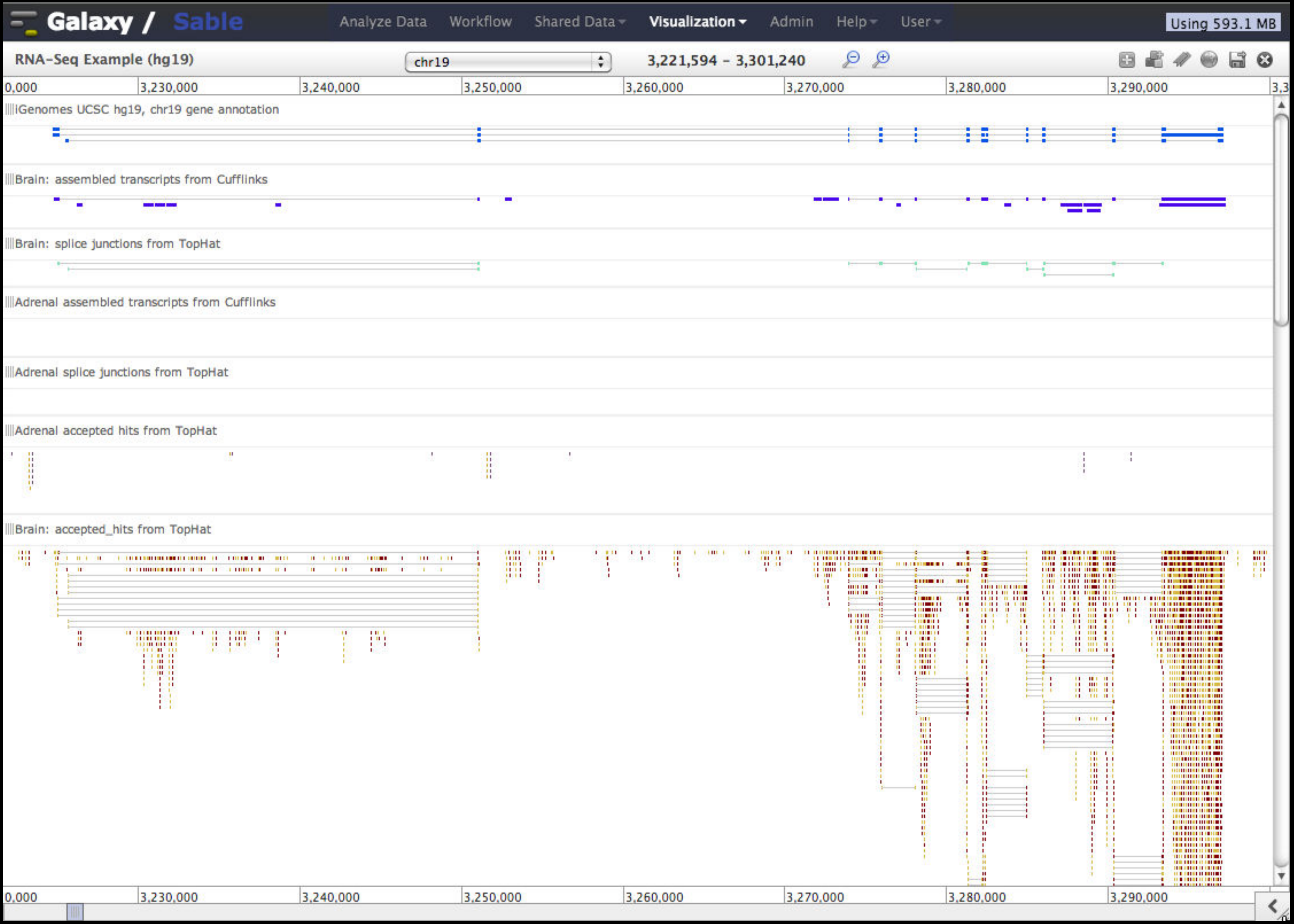
Last change: 9/3/2014, 14:06

New Chart

○ Stacked ● Stream ○ Expanded

Trackster

Genome browser embedded in Galaxy.


Galaxy / Sable
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 593.1 MB
RNA-Seq Example (hg19)
chr19
3,221,594 - 3,301,240
0,000
3,230,000
3,240,000
3,250,000
3,260,000
3,270,000
3,280,000
3,290,000
iGenomes UCSC hg19, chr19 gene annotation
Brain: assembled transcripts from Cufflinks
Brain: splice junctions from TopHat
Adrenal assembled transcripts from Cufflinks
Adrenal splice junctions from TopHat
Adrenal accepted hits from TopHat
Brain: accepted_hits from TopHat

Why do all this?

Visual Analytics

Researchers often do an analyze-summarize-visualize-repeat loop.

By bringing visualization into Galaxy we hope to tighten that loop as much as possible.

Dataset Collections

Support dataset collections as 1st class objects.
Run tools once on each dataset in the collection.
Run tools on the collection as a whole.
Tools become much more dynamic, flexible and responsive to input.
Support map/reduce paradigm.

Makes it possible to build workflows that can reason about paired datasets, technical replicates, multiple biological samples, ...

Galaxy API: Galaxy for the Bioinformatician

Scaling up also requires support for bioinformaticians and core staff.

Graphical user interfaces are great way to learn and explore tools.

But running analysis from a GUI can kinda irritate a bioinformatician who is adept at scripting and command line interfaces

Galaxy for the Bioinformatician

But if you go to the command line, you give up on Galaxy's user management, sharing, persistence, reproducibility, publishing, visualization, ... capabilities

The Galaxy API: full programmatic access to Galaxy, without going the a GUI.

Allows bioinformaticians to get the best of both worlds.

Back End Support: Workflow System

Replace current workflow system with, well, a workflow system.
Current system could be, um, more robust

Define a workflow engine API so that the workflow engine is interchangeable.

Back-End: Scaling Resources: Compute

Better support heterogeneous back ends:
Local cluster, institutional cloud, public cloud, national supercomputing centres, ...

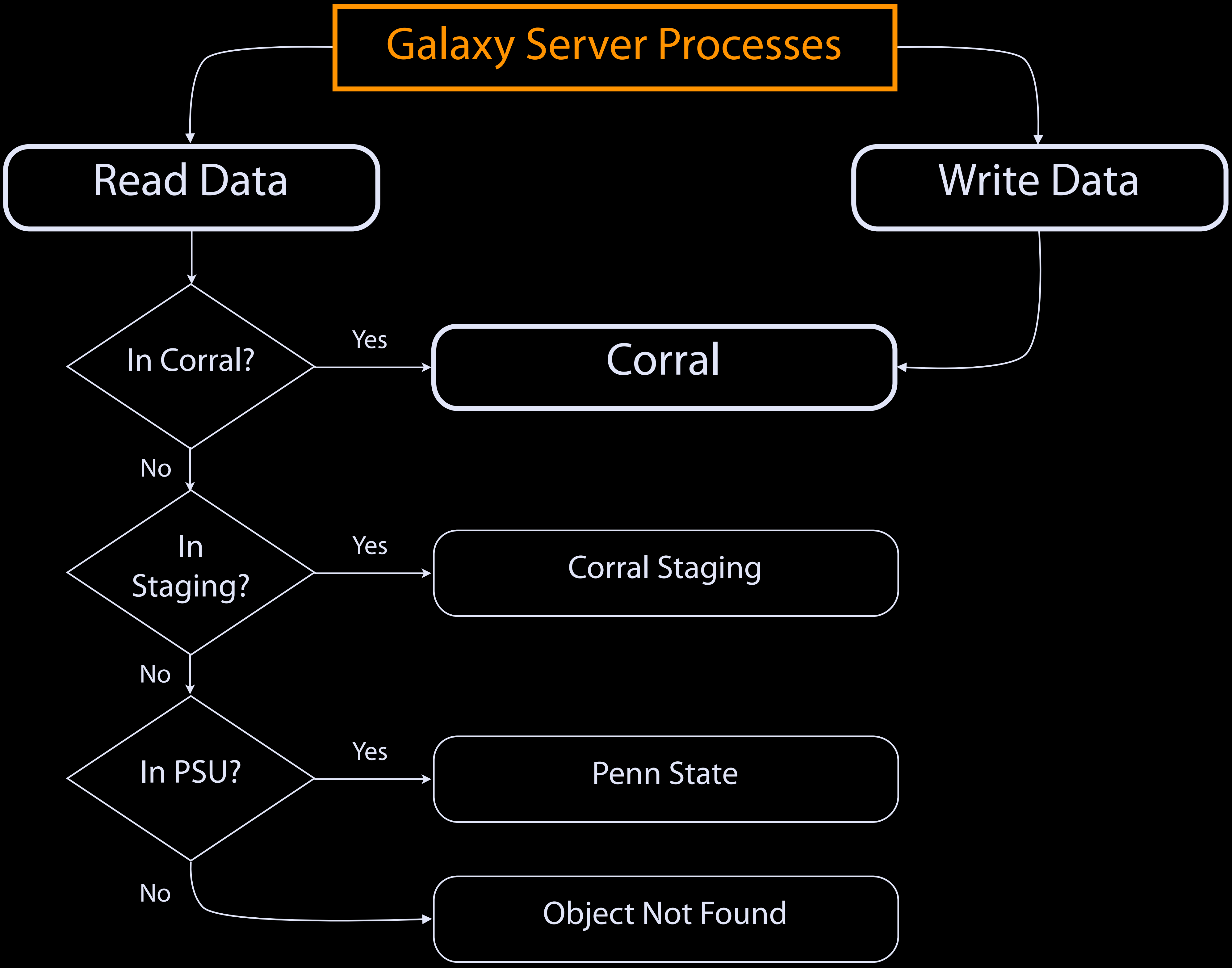
Enable Galaxy instances to be setup to exploit these resources.

Back-End: Scaling Resources: Storage

ObjectStore: Galaxy API for connecting to different storage backends

NFS, iRODS, S3, ...

ObjectStore in action: usegalaxy.org move to TACC



Scaling for Big Demand

So far all about big data
That's part of the challenge

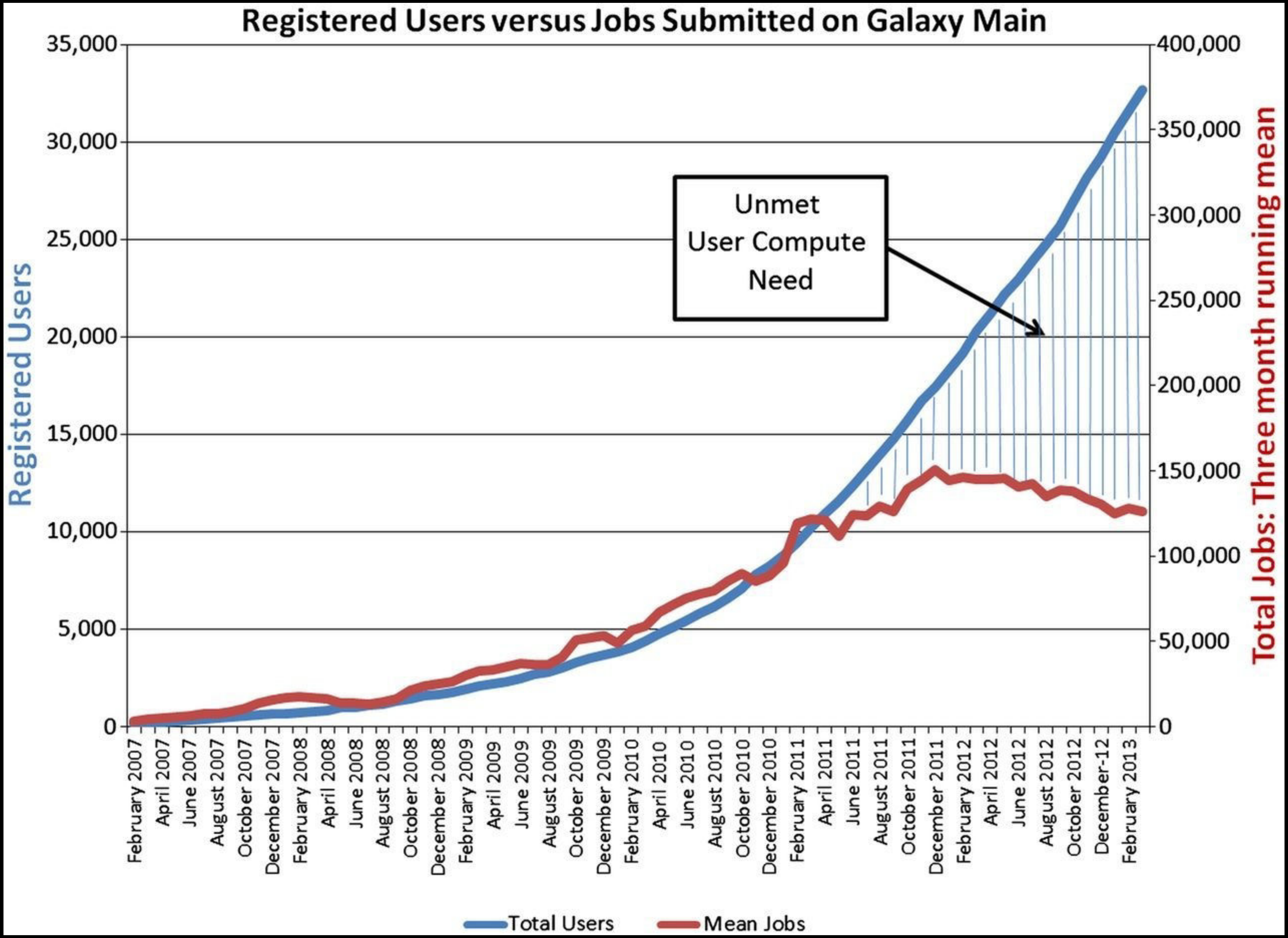
An orthogonal challenge is the sheer number of researchers now interested in doing bioinformatics analysis

Scaling for Big Demand: usegalaxy.org

When people think of Galaxy they often think of usegalaxy.org, the project's free (for everyone) web server. This integrates a wealth of tools, compute resources, terabytes of reference data and permanent storage.

However, ***a centralized solution cannot support the different analysis needs of the entire world.***

It's good to be popular, isn't it?



Leveraging the national cyberinfrastructure for biomedical research

LeDuc, *et al. J Am Med Inform Assoc doi:10.1136/amiajnl-2013-002059*

Scaling for Big Demand: Open Source Software

http://getgalaxy.org

Galaxy is installed at organizations around the world

Working hard to make installing a local Galaxy easy

Galaxy ToolShed

Data Managers

Scaling for Big Demand: Public Galaxy Instances

http://bit.ly/gxyServers

Interested in:

ChIP-seq?

- ✓ Cistrome, Nebula

Statistical Analysis?

- ✓ Genomic Hyperbrowser

Protein Synthesis?

- ✓ GWIPS-viz

de novo assembly?

- ✓ GigaGalaxy

Reasoning with ontologies?

- ✓ GO Galaxy

Repeats?

- ✓ RepeatExplorer

Over 60 public Galaxy servers

Scaling for Big Demand: Galaxy on the Cloud

amazon
web services™
openstack™
CLOUD SOFTWARE
g
globus
genomics
OpenNebula.org
The Open Source Toolkit for Cloud Computing

https://wiki.galaxyproject.org/Cloud

Scaling for Big Demand: Commercial Support

A ready-to-use appliance
(BioTeam)

Cloud-based solutions
(ABgenomica, AIS, Appistry, GenomeCloud)

Consulting & Customization
(Arctix, BioTeam, Deena Bioinformatics)

Scaling for Big Demand: Support

Tens of thousands of users leads to a lot of questions.

Absolutely have to encourage community support.

Project traditionally uses mailing list

Just moved the user support list to Galaxy Biostar, an online forum, that uses the Biostar platform

Scaling for Big Demand: Mailing Lists

wiki.galaxyproject.org/MailingLists

Galaxy-Dev

Questions about developing for and deploying Galaxy

High volume (5200 posts in 2013, 900+ members)

Galaxy-Announce

Project announcements, low volume, moderated

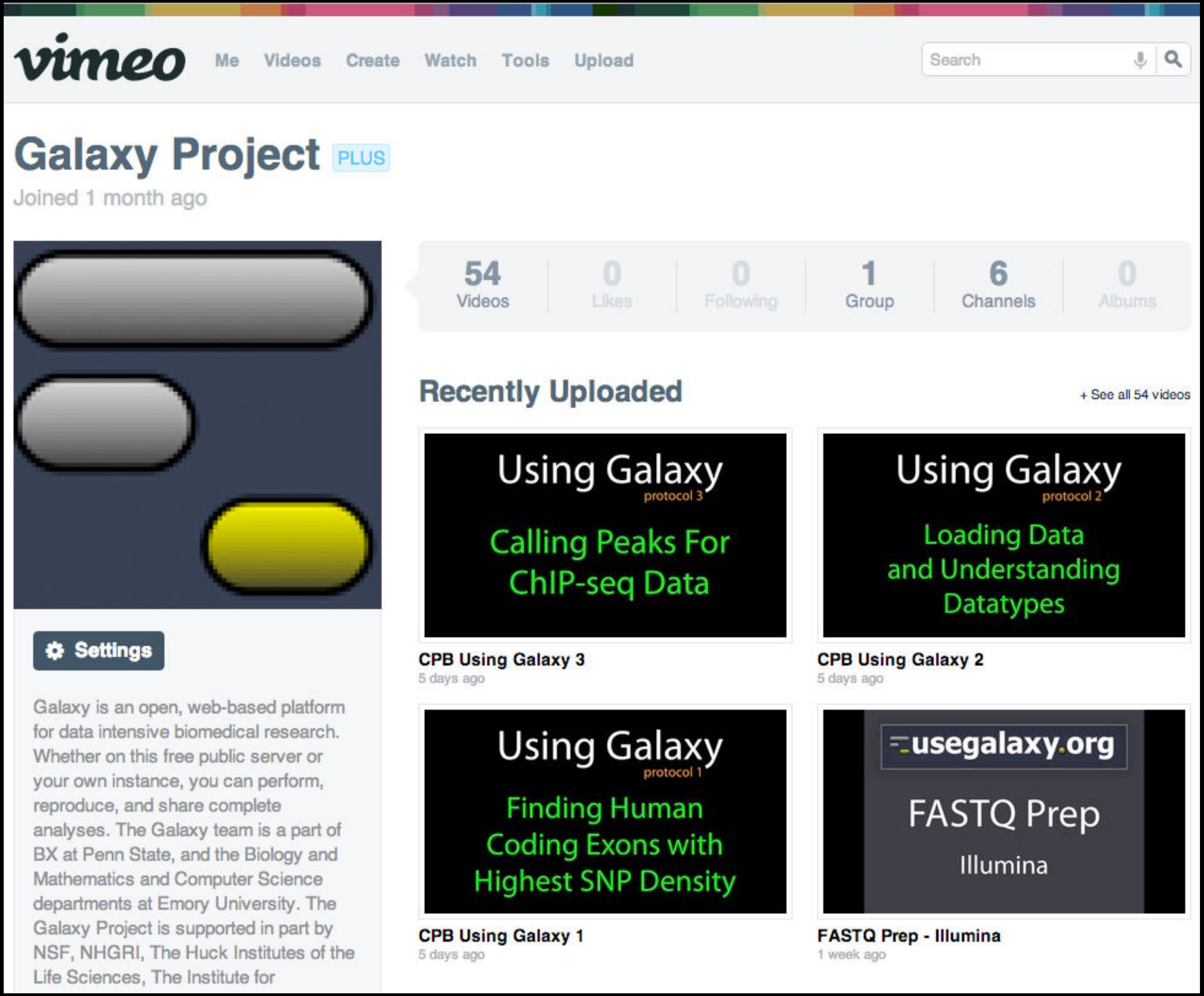
Low volume (47 posts in 2013, 3400+ members)

Galaxy-User (deprecated)

Questions about using Galaxy and usegalaxy.org

High volume (1328 posts in 2013, 2600+ members)

Scaling for Big Demand: Screencasts



Galaxy Project PLUS

Joined 1 month ago

54	0	0	1	6	0
Videos	Likes	Following	Group	Channels	Albums

Recently Uploaded

Using Galaxy protocol 3 Calling Peaks For ChIP-seq Data

CPB Using Galaxy 3
5 days ago

Using Galaxy protocol 2 Loading Data and Understanding Datatypes

CPB Using Galaxy 2
5 days ago

Using Galaxy protocol 1 Finding Human Coding Exons with Highest SNP Density

CPB Using Galaxy 1
5 days ago

usegalaxy.org FASTQ Prep Illumina

FASTQ Prep - Illumina
1 week ago

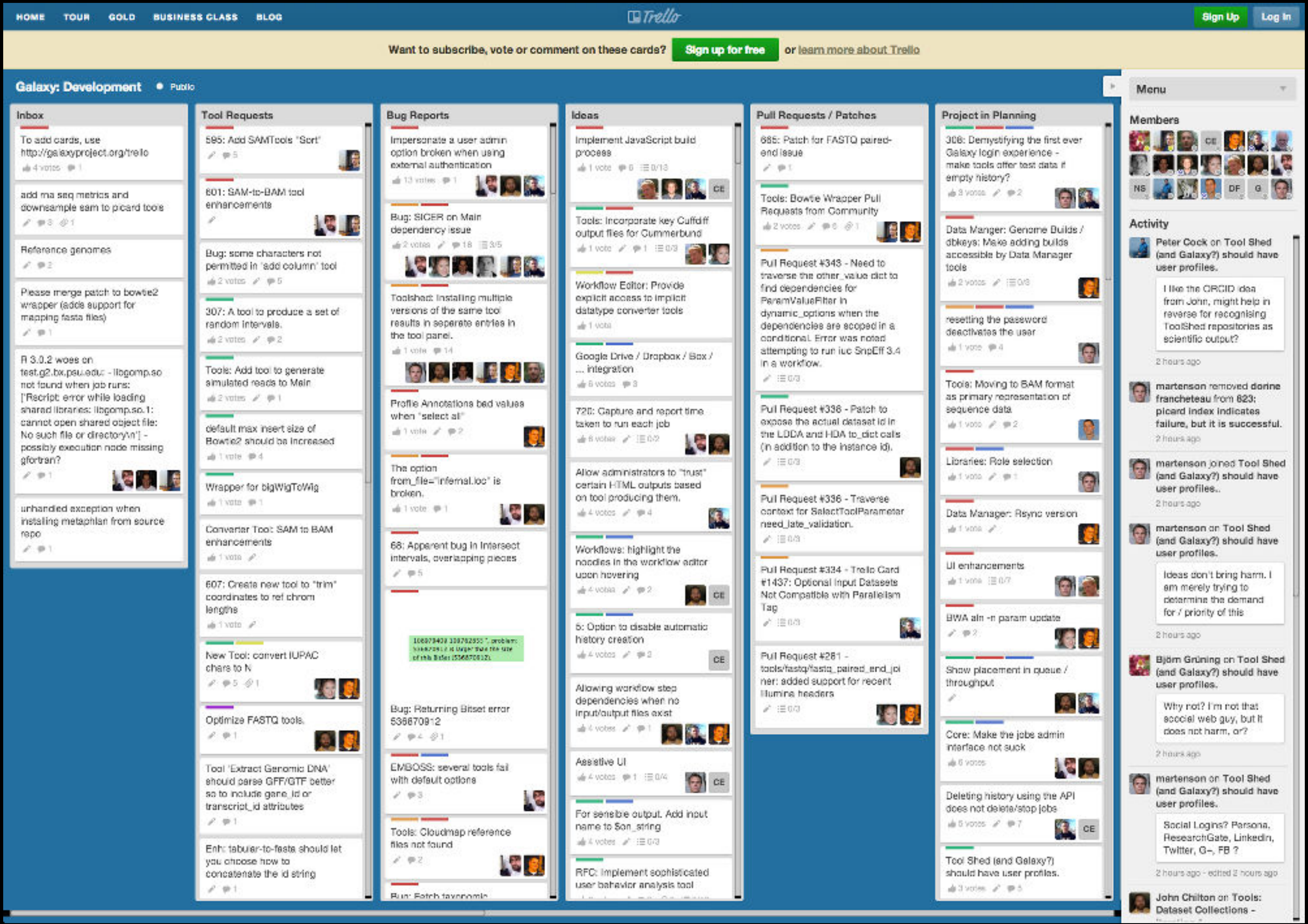
Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for

"How to" screencasts on using and deploying Galaxy

Talks from previous meetings.

http://vimeo.com/galaxyproject

Scaling for Big Demand: Feedback and guidance



Inbox

- To add cards, use http://galaxyproject.org/trello — 4 votes, 1
- add rna seq metrics and downsample sam to picard tools — 3, 1
- Reference genomes — 2
- Please merge patch to bowtie2 wrapper (adds support for mapping fasta files) — 1
- R 3.0.2 woes on test.g2.bx.psu.edu - libgomp.so not found when job runs: ['Rscript: error while loading shared libraries: libgomp.so.1: cannot open shared object file: No such file or directory\n'] - possibly execution node missing gfortran? — 1
- unhandled exception when installing metaphlan from source repo — 1

Tool Requests

- 595: Add SAMTools "Sort" — 5
- 601: SAM-to-BAM tool enhancements
- Bug: some characters not permitted in 'add column' tool — 2 votes, 5
- 307: A tool to produce a set of random intervals. — 2 votes, 2
- Tools: Add tool to generate simulated reads to Main — 2 votes, 1
- default max insert size of Bowtie2 should be increased — 1 vote, 4
- Wrapper for bigWigToWig — 1 vote, 1
- Converter Tool: SAM to BAM enhancements — 1 vote
- 607: Create new tool to "trim" coordinates to ref chrom lengths — 1 vote
- New Tool: convert IUPAC chars to N — 5, 1
- Optimize FASTQ tools. — 1
- Tool 'Extract Genomic DNA' should parse GFF/GTF better so to include gene_id or transcript_id attributes — 1
- Enh: tabular-to-fasta should let you choose how to concatenate the id string — 1

Bug Reports

- Impersonate a user admin option broken when using external authentication — 13 votes, 1
- Bug: SICER on Main dependency issue — 2 votes, 18, 3/5
- Toolshed: Installing multiple versions of the same tool results in seperate entries in the tool panel. — 1 vote, 14
- Profile Annotations bad values when "select all" — 1 vote, 2
- The option from_file="infernal.loc" is broken. — 1 vote, 1
- 68: Apparent bug in Intersect intervals, overlapping pieces — 5
- 106079409 108782955 ", problem: 536870912 is larger than the size of this Bitset (536870912). Bug: Returning Bitset error 536870912 — 4, 1
- EMBOSS: several tools fail with default options — 3
- Tools: Cloudmap reference files not found — 2
- Bug: Fetch taxonomic

Ideas

- Implement JavaScript build process — 1 vote, 6, 0/13
- Tools: Incorporate key Cuffdiff output files for Cummerbund — 1 vote, 1, 0/3
- Workflow Editor: Provide explicit access to implicit datatype converter tools — 1 vote
- Google Drive / Dropbox / Box / ... integration — 6 votes, 3
- 720: Capture and report time taken to run each job — 6 votes, 0/2
- Allow administrators to "trust" certain HTML outputs based on tool producing them. — 4 votes, 4
- Workflows: highlight the noodles in the workflow editor upon hovering — 4 votes, 2
- 5: Option to disable automatic history creation — 4 votes, 2
- Allowing workflow step dependencies when no input/output files exist — 4 votes, 1
- Assistive UI — 4 votes, 1, 0/4
- For sensible output. Add input name to $on_string — 4 votes, 0/3
- RFC: Implement sophisticated user behavior analysis tool

Pull Requests / Patches

- 685: Patch for FASTQ paired-end issue — 1
- Tools: Bowtie Wrapper Pull Requests from Community — 2 votes, 6, 1
- Pull Request #343 - Need to traverse the other_value dict to find dependencies for ParamValueFilter in dynamic_options when the dependencies are scoped in a conditional. Error was noted attempting to run iuc SnpEff 3.4 in a workflow. — 0/3
- Pull Request #336 - Patch to expose the actual dataset id in the LDDA and HDA to_dict calls (in addition to the instance id). — 0/3
- Pull Request #336 - Traverse context for SelectToolParameter need_late_validation. — 0/3
- Pull Request #334 - Trello Card #1437: Optional Input Datasets Not Compatible with Parallelism Tag — 0/3
- Pull Request #281 - tools/fastq/fastq_paired_end_joiner: added support for recent Illumina headers — 0/3

Project in Planning

- 308: Demystifying the first ever Galaxy login experience - make tools offer test data if empty history? — 3 votes, 2
- Data Manger: Genome Builds / dbkeys: Make adding builds accessible by Data Manager tools — 2 votes, 0/3
- resetting the password deactivates the user — 1 vote, 4
- Tools: Moving to BAM format as primary representation of sequence data — 1 vote, 2
- Libraries: Role selection — 1 vote, 1
- Data Manager: Rsync version — 1 vote
- UI enhancements — 1 vote, 0/7
- BWA aln -n param update — 2
- Show placement in queue / throughput
- Core: Make the jobs admin interface not suck — 6 votes
- Deleting history using the API does not delete/stop jobs — 5 votes, 7
- Tool Shed (and Galaxy?) should have user profiles. — 3 votes, 5

Activity

Peter Cock on Tool Shed (and Galaxy?) should have user profiles.

> I like the ORCID idea from John, might help in reverse for recognising ToolShed repositories as scientific output?

2 hours ago

martenson removed dorine francheteau from 623: picard index indicates failure, but it is successful.

2 hours ago

martenson joined Tool Shed (and Galaxy?) should have user profiles..

2 hours ago

martenson on Tool Shed (and Galaxy?) should have user profiles.

> Ideas don't bring harm. I am merely trying to determine the demand for / priority of this

2 hours ago

Björn Grüning on Tool Shed (and Galaxy?) should have user profiles.

> Why not? I'm not that social web guy, but it does not harm, or?

2 hours ago

martenson on Tool Shed (and Galaxy?) should have user profiles.

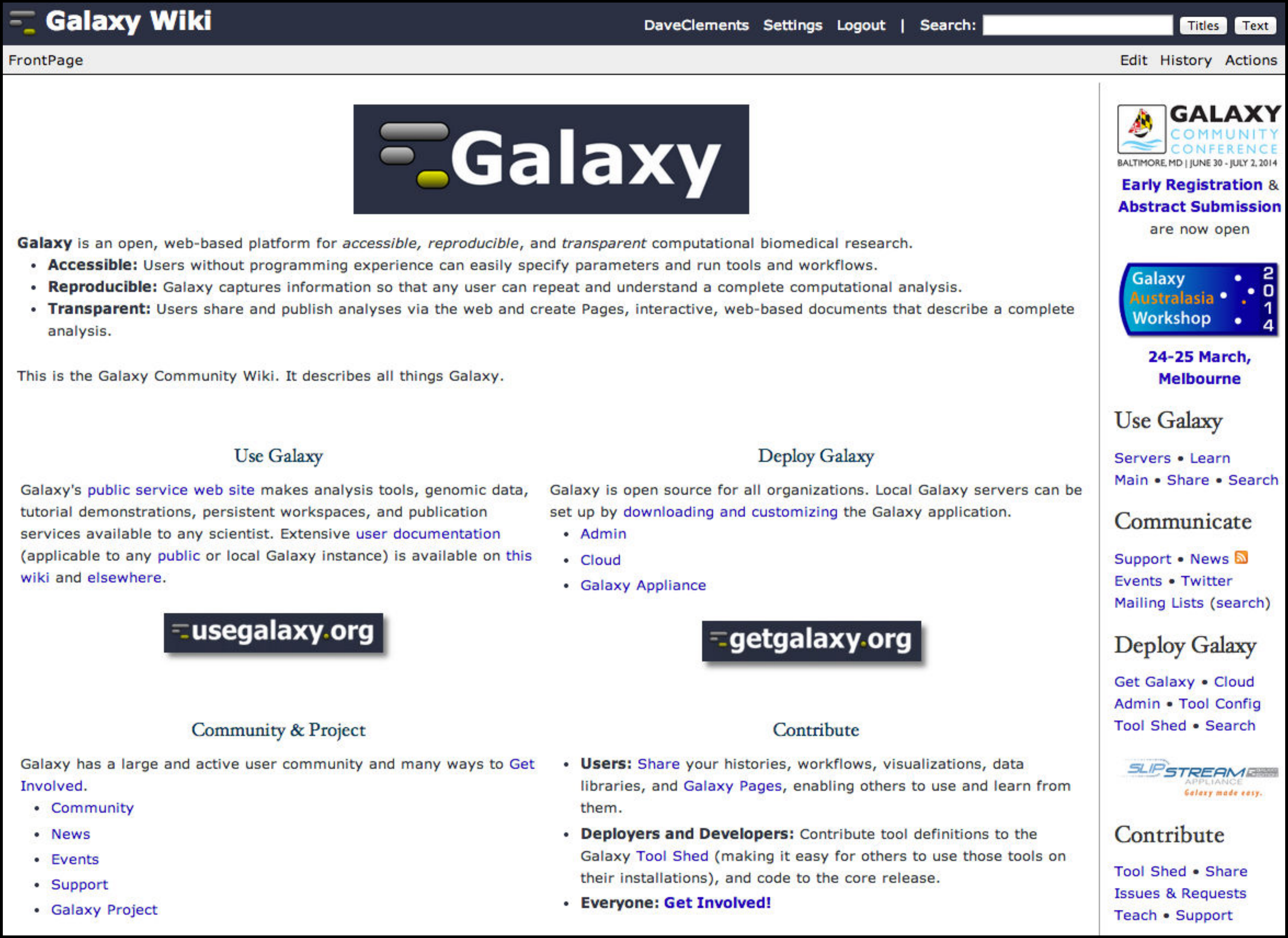
> Social Logins? Persona, ResearchGate, LinkedIn, Twitter, G+, FB ?

2 hours ago - edited 2 hours ago

John Chilton on Tools: Dataset Collections -

http://bit.ly/gxytrello

http://wiki.galaxyproject.org



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

- **Accessible:** Users without programming experience can easily specify parameters and run tools and workflows.
- **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, interactive, web-based documents that describe a complete analysis.

This is the Galaxy Community Wiki. It describes all things Galaxy.

Use Galaxy

Galaxy's public service web site makes analysis tools, genomic data, tutorial demonstrations, persistent workspaces, and publication services available to any scientist. Extensive user documentation (applicable to any public or local Galaxy instance) is available on this wiki and elsewhere.

Deploy Galaxy

Galaxy is open source for all organizations. Local Galaxy servers can be set up by downloading and customizing the Galaxy application.

- Admin
- Cloud
- Galaxy Appliance

Community & Project

Galaxy has a large and active user community and many ways to Get Involved.

- Community
- News
- Events
- Support
- Galaxy Project

Contribute

- **Users:** Share your histories, workflows, visualizations, data libraries, and Galaxy Pages, enabling others to use and learn from them.
- **Deployers and Developers:** Contribute tool definitions to the Galaxy Tool Shed (making it easy for others to use those tools on their installations), and code to the core release.
- **Everyone:** **Get Involved!**

GALAXY COMMUNITY CONFERENCE
BALTIMORE, MD | JUNE 30 - JULY 2, 2014

Early Registration & **Abstract Submission** are now open

24-25 March, Melbourne

Use Galaxy

Servers • Learn
Main • Share • Search

Communicate

Support • News
Events • Twitter
Mailing Lists (search)

Deploy Galaxy

Get Galaxy • Cloud
Admin • Tool Config
Tool Shed • Search

Contribute

Tool Shed • Share
Issues & Requests
Teach • Support

Scaling for Big Demand: Gather

GALAXY COMMUNITY CONFERENCE

BALTIMORE, MD | JUNE 30 - JULY 2, 2014

http://bit.ly/gcc2014

Scaling for Big Demand: Gather

Support community organized efforts and events.

Scaling for Big Demand: Training

Workshops in Norwich (this Friday)
and Edinburgh (next week)

https://wiki.galaxyproject.org/Events

The Galaxy Team

Enis Afgan

Dannon Baker

Dan Blankenberg

Dave Bouvier

Marten Cech

John Chilton

Dave Clements

Nate Coraor

Carl Eberhard

Dorine Francheteau

Jeremy Goecks

Sam Guerler

Jen Jackson

Greg von Kuster

Ross Lazarus

Anton Nekrutenko

Nick Stoler

James Taylor

https://wiki.galaxyproject.org/GalaxyTeam

Galaxy is hiring post-docs and software engineers

Please help.

http://wiki.galaxyproject.org/GalaxyIsHiring

Also Thanks To

TGAC

The Genome Analysis Centre™

Matt Drew
Vicky Schneider-Gricar
Helen Tunney

THE UNIVERSITY *of* EDINBURGH

igmm
INSTITUTE OF GENETICS
& MOLECULAR MEDICINE

edinburgh genomics.
design execution analysis

ROSLIN

Wellcome Trust
Centre for Cell Biology
WTCCB

Edinburgh Bioinformatics

Thank you

Dave Clements

Galaxy Project

Johns Hopkins University

outreach@galaxyproject.org