An Introduction to Galaxy

Daniel Blankenberg
Postdoctoral Research Associate
The Galaxy Team
http://UseGalaxy.org

Overview

- What is Galaxy?
- Galaxy for Experimental Biologists
- Galaxy for Bioinformaticians

Galaxy, a web-based genome analysis platform

- An open-source **framework** for integrating various computational tools and databases into a cohesive workspace
- A web-based **service** we provide, integrating many popular tools and resources for comparative genomics
- A completely **self-contained application** for building your own **Galaxy** style sites

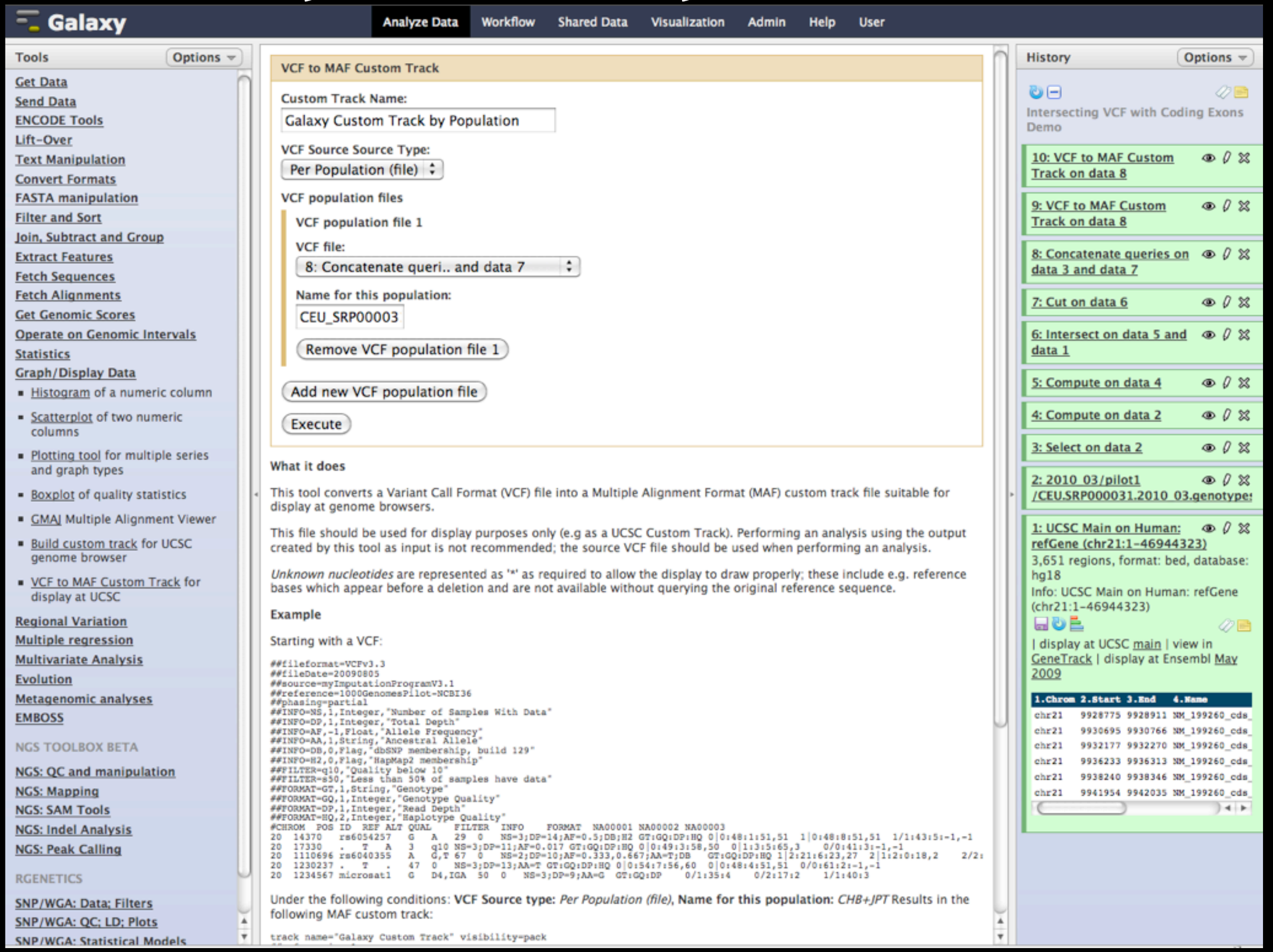
Overview

- What is Galaxy?
- Galaxy for Experimental Biologists
- Galaxy for Bioinformaticians

Galaxy: the one-stop shop for Genome Analysis

- Analyze
  - Retrieve data directly from popular data resources or upload your own
  - Interactively manipulate genomic data with a comprehensive and expanding "best-practices" toolset
- Visualize
  - Send data results to external Genome Browsers
  - Build reusable AJAX-based custom Genome Browsers (Trackster)
- Publish and Share
  - Results and step-by-step analysis record (Data Libraries and Histories)
  - Customizable pipelines (Workflows)
  - Complete protocols (Pages)

Galaxy's Analysis Interface



Tools and Datasources

- Datasources
  - Upload File from your computer
  - UCSC table browser
  - BioMart
  - interMine / modMine
  - EuPathDB server
  - EncodeDB at NHGRI
  - EpiGRAPH server
- Tool Suites
  - Text Manipulation
  - Format Converters
  - Filtering and Sorting
  - Join, Subtract, Group
  - Sequence Tools
  - Multiple-species Alignment Tools
  - Genomic Interval Operations
  - Summary Statistics
  - Graphing / Plotting
  - Regional Variation
  - EMBOSS
  - Evolution / Phylogeny
  - NGS
  - RGenetics
  - ...and more

Visualize

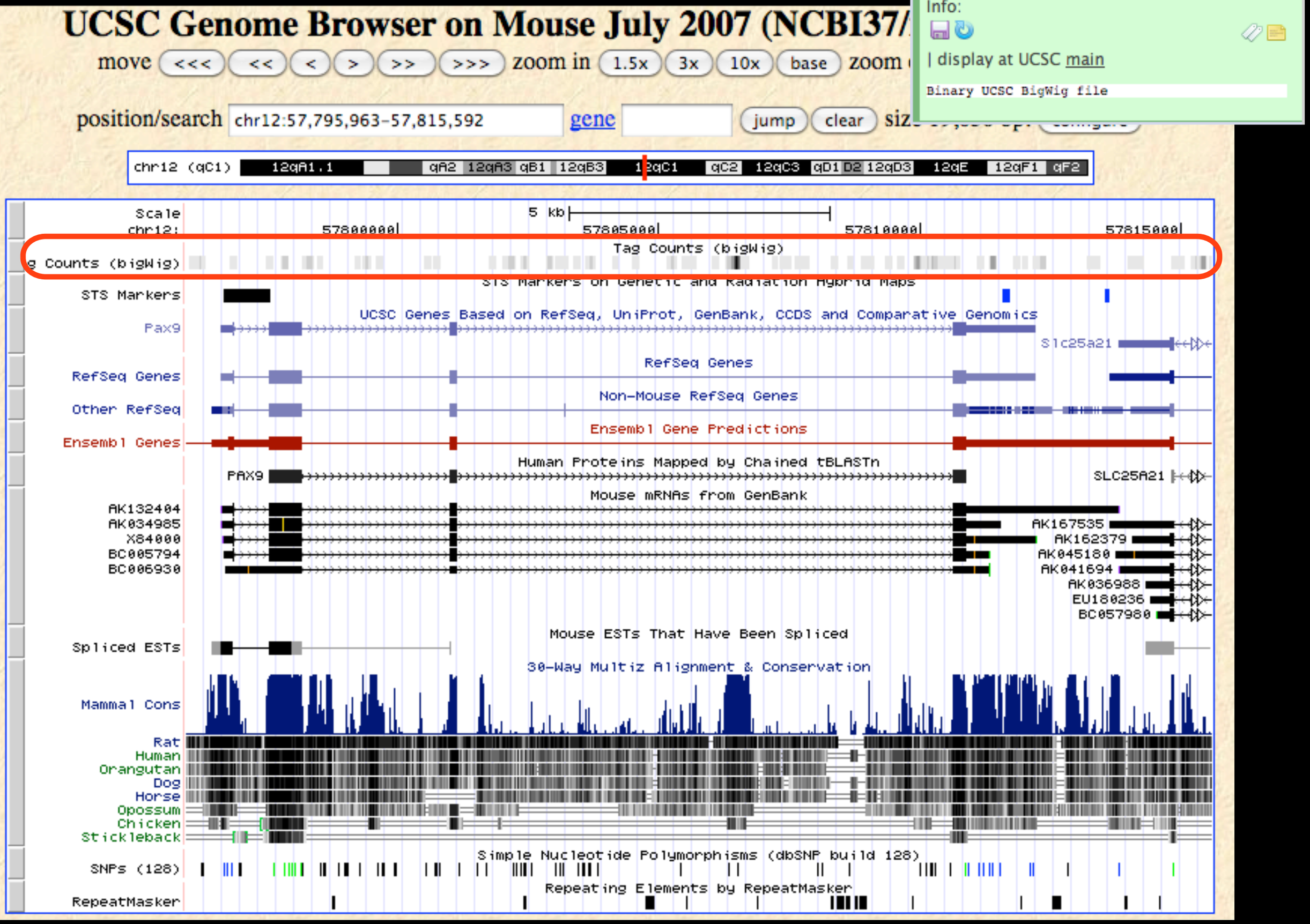
- Send data results to external Genome Browsers
- Build reusable and sharable custom Genome Browsers (Trackster)

External Genome Browsers

- UCSC
- Ensembl
- GBrowse

- Adding more is easy!
  - https://bitbucket.org/galaxy/galaxy-central/wiki/ExternalDisplayApplications/Tutorial

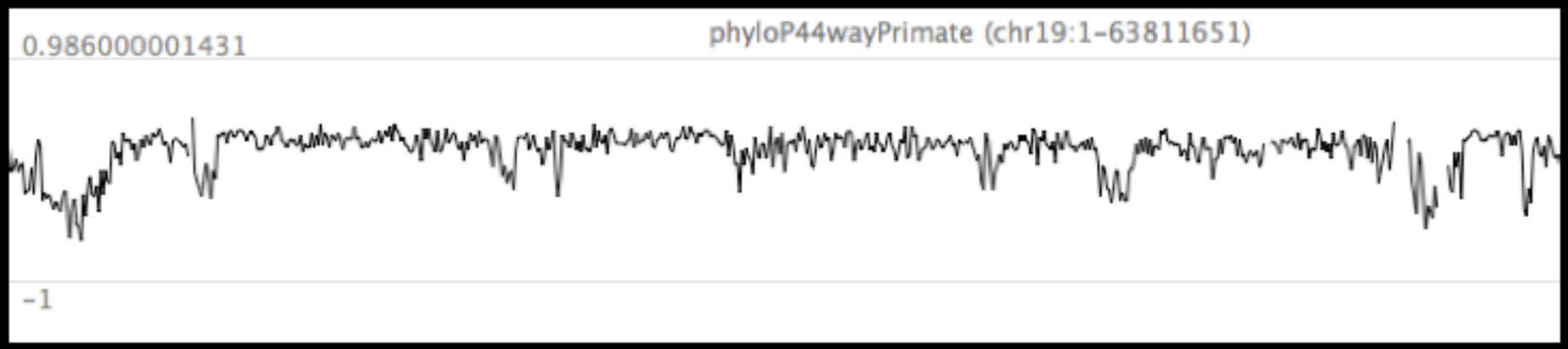
Visualize at UCSC



Trackster

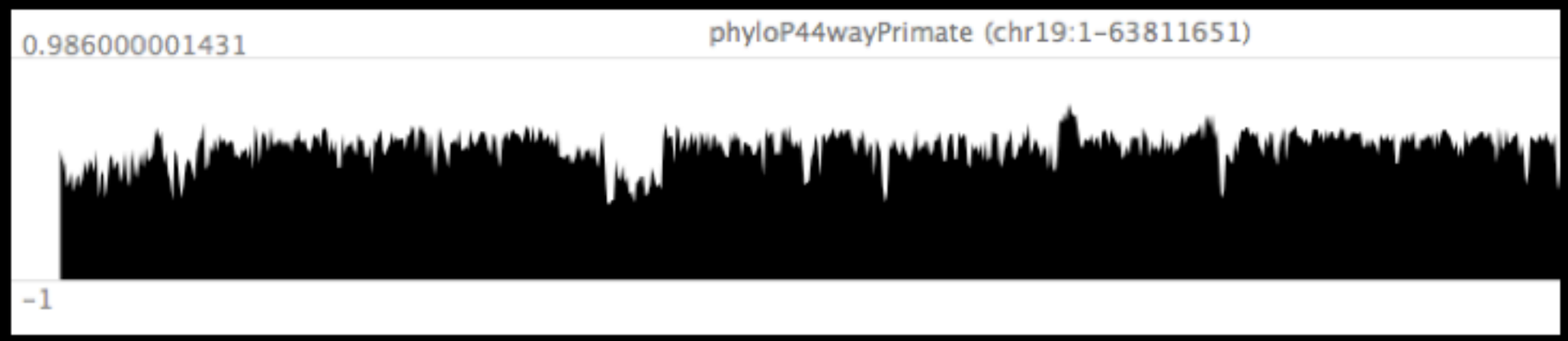
- Track/data viewer in web browser
  - HTML5 Canvas, jQuery
  - Renders in browser, not on server
- View your data from within Galaxy
  - No file transfers to third party
  - Use it locally, even without internet access
- Fast, responsive, interactive UI

Wig, Bedgraph (Line Tracks)



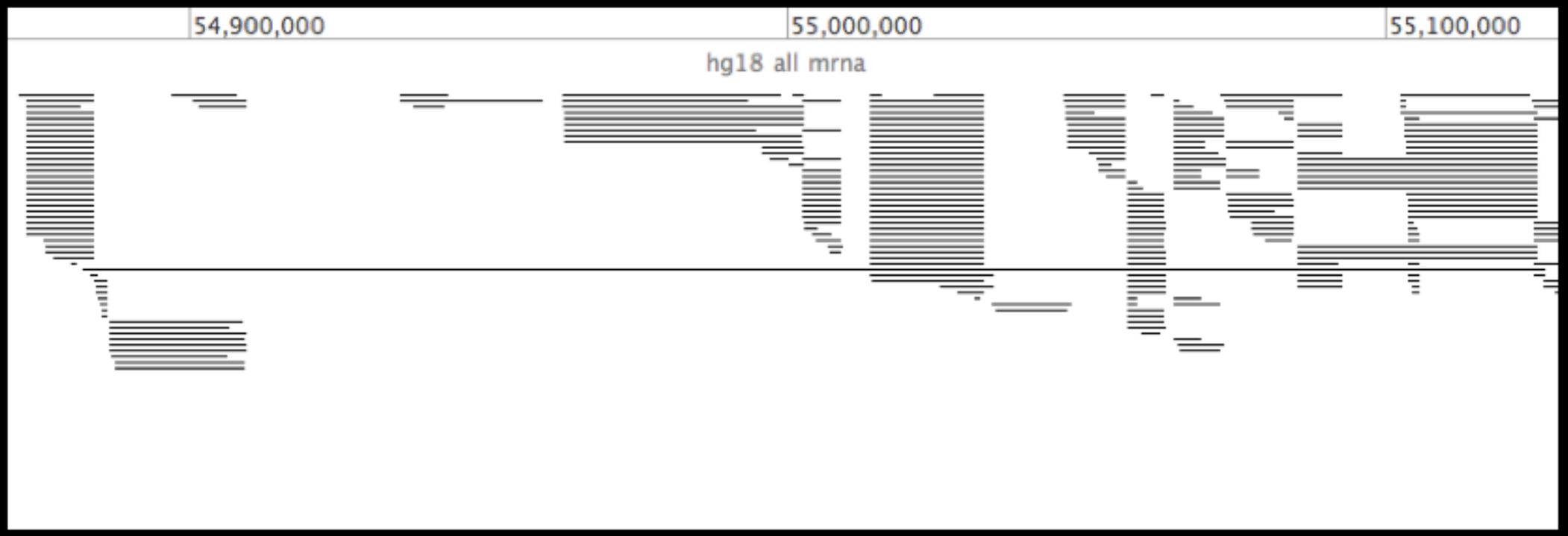
Regular line graph display

Intensity display (shades of gray)



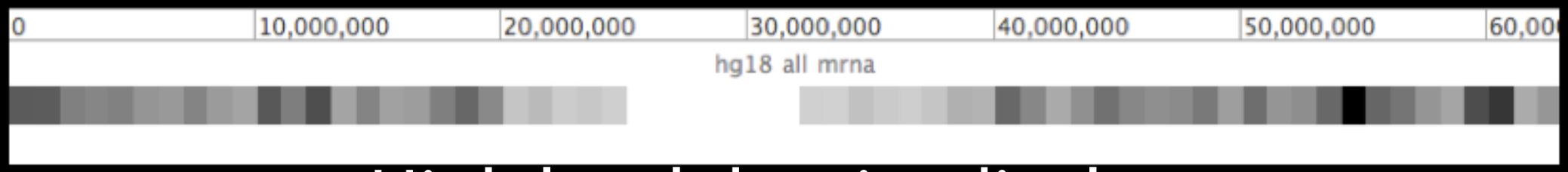
Filled line graph display

Bed (Feature Track)

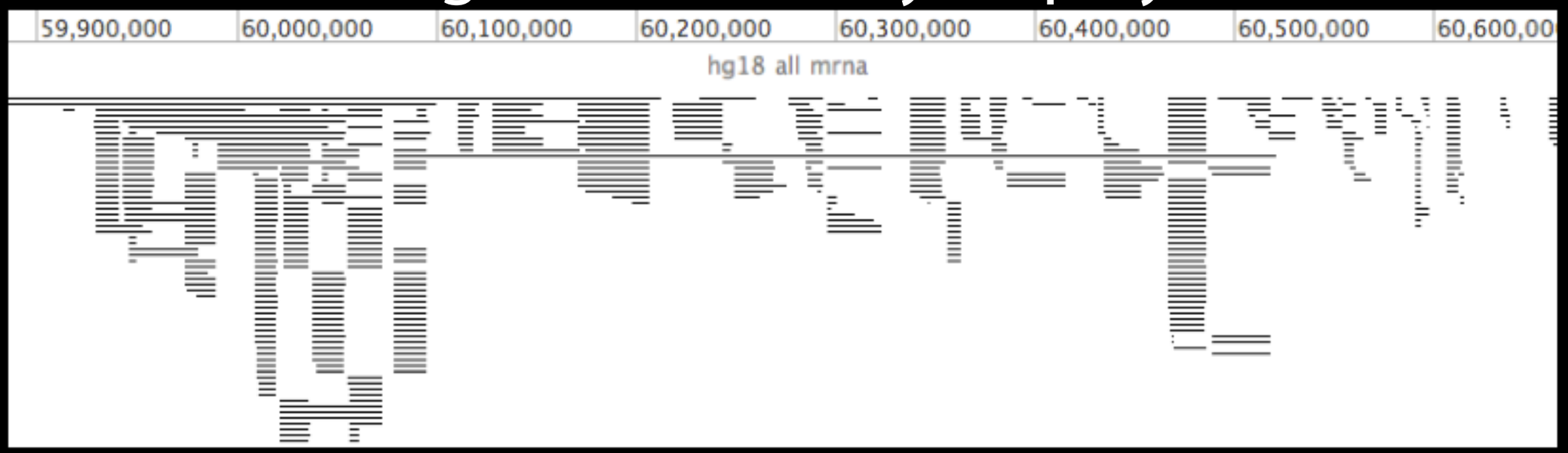


Snippet of hg18 all_mrna feature track

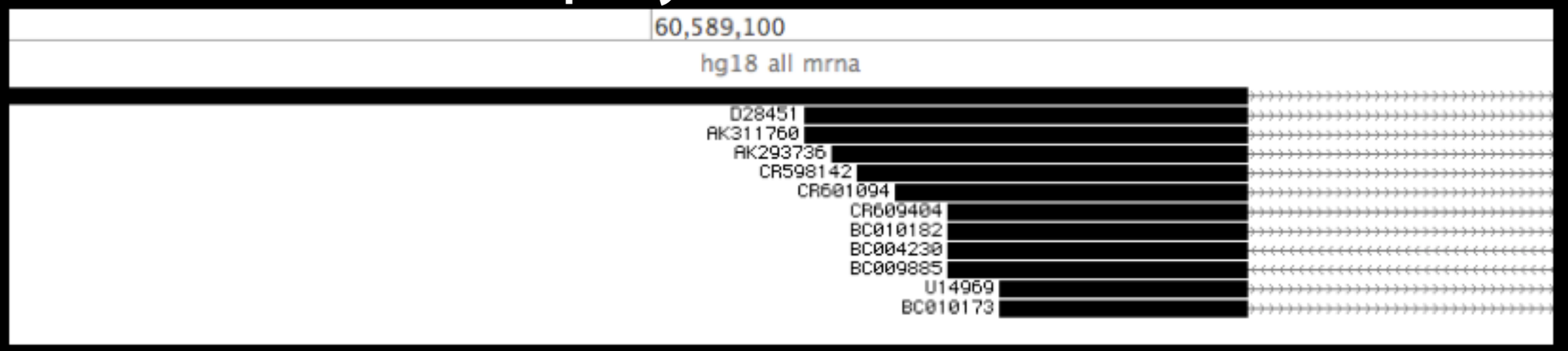
3 levels of detail: automatically adjusts based on what can fit on the screen



High level density display

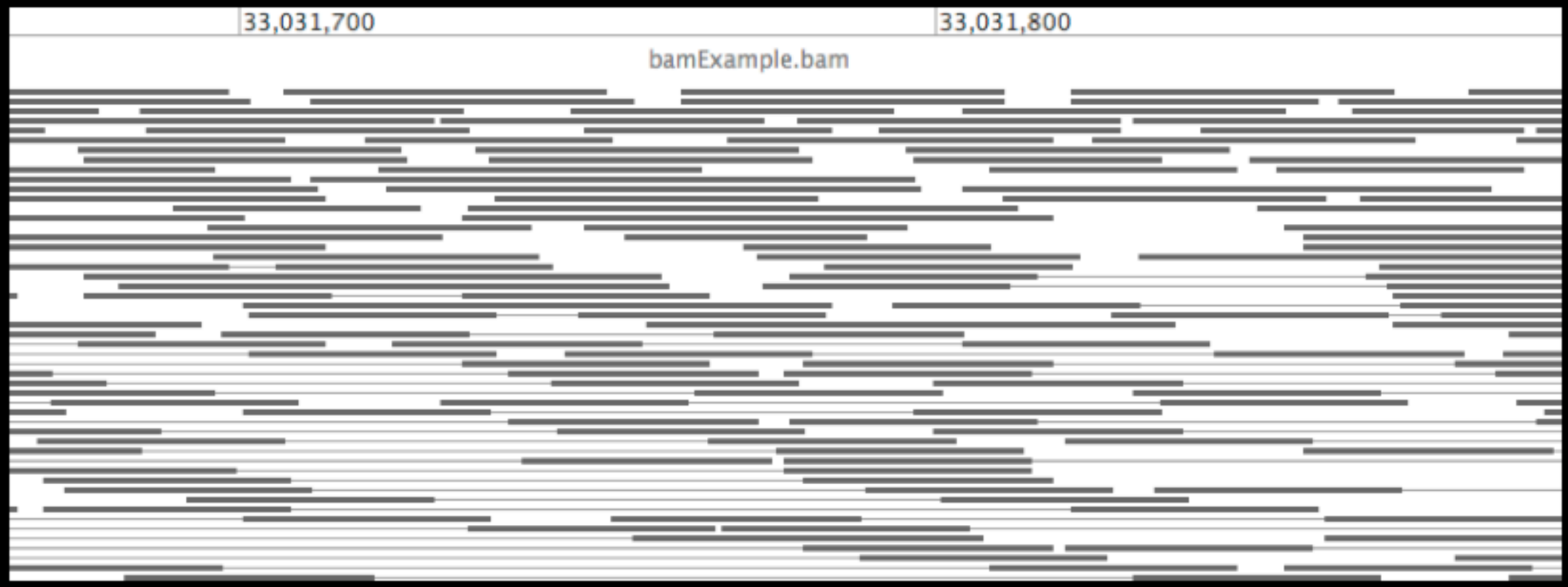


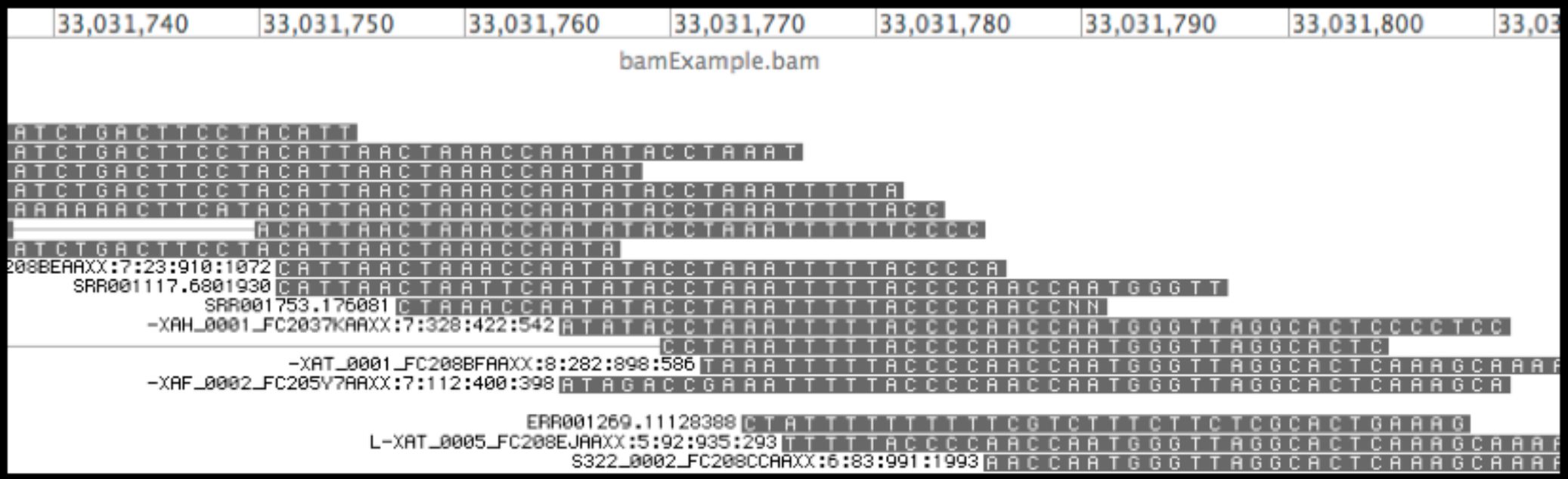
Feature display with no labels/detail



Feature display with labels, intron indicators, exon indicators

BAM (Aligned Reads)





Data Libraries

Data Library "Bushman"

Library Actions

These are the data underlying the analyses reported in the paper "Complete Khoisan and Bantu genomes from southern Africa" by S. C. Schuster et al., published in the journal Nature, February 18, 2010. Each data set can be downloaded and/or imported into a Galaxy history. Data will be updated as the project progresses.

Name	Information	Uploaded By	Date	File Size
All SNPs in personal genomes	Summary table of SNPs in all individuals	greg@bx.psu.edu	2010-01-28	676.8 Mb
Alu insertions in KB1		greg@bx.psu.edu	2010-02-10	14.9 Kb
Alu insertions in NB1		greg@bx.psu.edu	2010-02-10	6.5 Kb
KB1_microsatellites.txt		greg@bx.psu.edu	2010-02-15	3.5 Mb
NB1_microsatellites.txt		greg@bx.psu.edu	2010-02-15	828.5 Kb
amino acid differences with functional predictions		greg@bx.psu.edu	2010-02-05	1.1 Mb
gene copy numbers in KB1 and other personal genome		greg@bx.psu.edu	2010-02-15	2.1 Mb
indels in ABT		greg@bx.psu.edu	2010-02-03	105.3 Kb
indels in KB1		greg@bx.psu.edu	2010-02-03	14.2 Mb
indels in MD8		greg@bx.psu.edu	2010-02-03	109.8 Kb
indels in NB1		greg@bx.psu.edu	2010-02-03	519.5 Kb
indels in TK1		greg@bx.psu.edu	2010-02-03	123.2 Kb
novel SNPs in ABT		greg@bx.psu.edu	2010-02-09	9.4 Mb
novel SNPs in KB1		greg@bx.psu.edu	2010-02-09	16.9 Mb
novel SNPs in MD8		greg@bx.psu.edu	2010-02-09	594.1 Kb
novel SNPs in NB1		greg@bx.psu.edu	2010-02-09	4.1 Mb
novel SNPs in TK1		greg@bx.psu.edu	2010-02-09	722.6 Kb
sequenced exon-containing intervals		greg@bx.psu.edu	2010-02-03	3.1 Mb

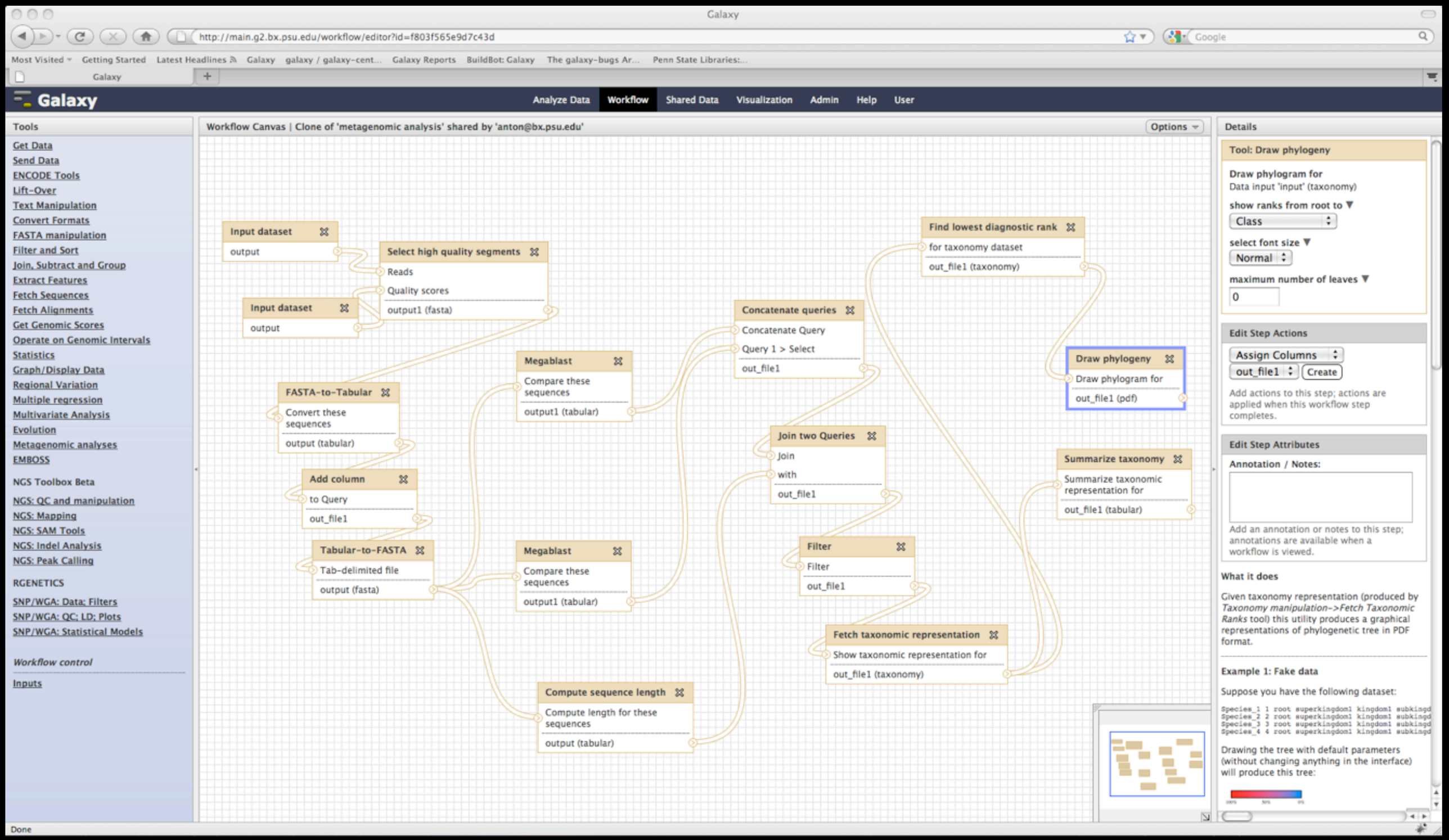
For selected items: Import into your current history Go

http://usegalaxy.org/bushman

Managing Libraries

- Loading Data
  - Upload a single file
  - Import datasets from a Galaxy history
  - Upload a directory of files
  - Directly from Sequencer using Sample Tracking System
- Accessing Data
  - Data contents on disk are not copied
  - Dataset security
    - Public
    - Role-based access control (RBAC)
- Annotating Library Data: Library Templates
  - Build user fillable forms
  - Associate at Library, Folder or Dataset level

Workflows



http://main.g2.bx.psu.edu/u/aun1/w/metagenomic-analysis

Pages

Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement

SERGEI KOSAKOVSKY POND[1,2,*], SAMIR WADHAWAN[3,6*], FRANCESCA CHIAROMONTE[4], GURUPRASAD ANANDA[1,3], WEN-YU CHUNG[1,3,7], JAMES TAYLOR[1,5], ANTON NEKRUTENKO[1,3] and THE GALAXY TEAM[1,*]

Correspondence should addressed to SKP, JT, or AN.

How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to the **exact** analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own data. Specifically, we provide the two histories and one workflow found below. You can view these items by clicking on their name to expand them. You can also import these items into your Galaxy workspace and start using them; click on the green plus to import an item. To import workflows you must create a Galaxy account (unless you already have one) – a hassle-free procedure where you are only asked for a username and password.

This is the Galaxy history detailing the comparison of our pipeline to MEGAN:

Galaxy History | Galaxy vs MEGAN
Comparison of Galaxy vs. MEGAN pipeline.

This is the Galaxy history showing a generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3A**):

Galaxy History | metagenomic analysis

10: Concatenate queries on data 8 and data 7	Merge Megablast runs to produce a single dataset for reads compared to both WGS and NT
11: Join two Queries on data 9 and data 10	Combine sequence length data with results from Megablast runs
12: Filter on data 11	Filter suboptimal sequence alignments using the expression (sequence_alignment_length/sequence_read_length > 0.5); sequences that do not meet this criterion are filtered out
13: Fetch taxonomic representation on data 12	Get taxonomic representation for filtered, aligned sequences
14: Find lowest diagnostic rank on data 13	Get reads specific to ranks below Kingdom level
15: Summarize taxonomy on data 13	Tabulate list of taxonomic groups contained in reads from dataset 14
16: Draw phylogeny on data 14	Build and draw phylogenetic tree from ranks in dataset 14

This is the Galaxy workflow for generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3B**):

Galaxy Workflow | metagenomic analysis
Generic workflow for performing a metagenomic analysis on NGS data.

Supplemental Analysis

Comparison between Galaxy pipeline and Megan

(*Use this link to see Galaxy history representing this analysis. Individual elements of this history are referred to as* **History item1, 2** *and so on using* **bold** *typeface*)

The first step of a homology-based metagenomic analysis is to contrast a collection of sequencing reads against a database whose entries are assigned to taxonomic ranks. Following the procedure of (Huson et al. 2007) we used the non-redundant protein database (NR) from the National Center for Biotechnology Information. There are several avenues for importing large sets of alignments into Galaxy. First, alignments can be generated directly within Galaxy (see the following section). Alternatively, alignments generated elsewhere (e.g., using local

About this Page

Author: aun1

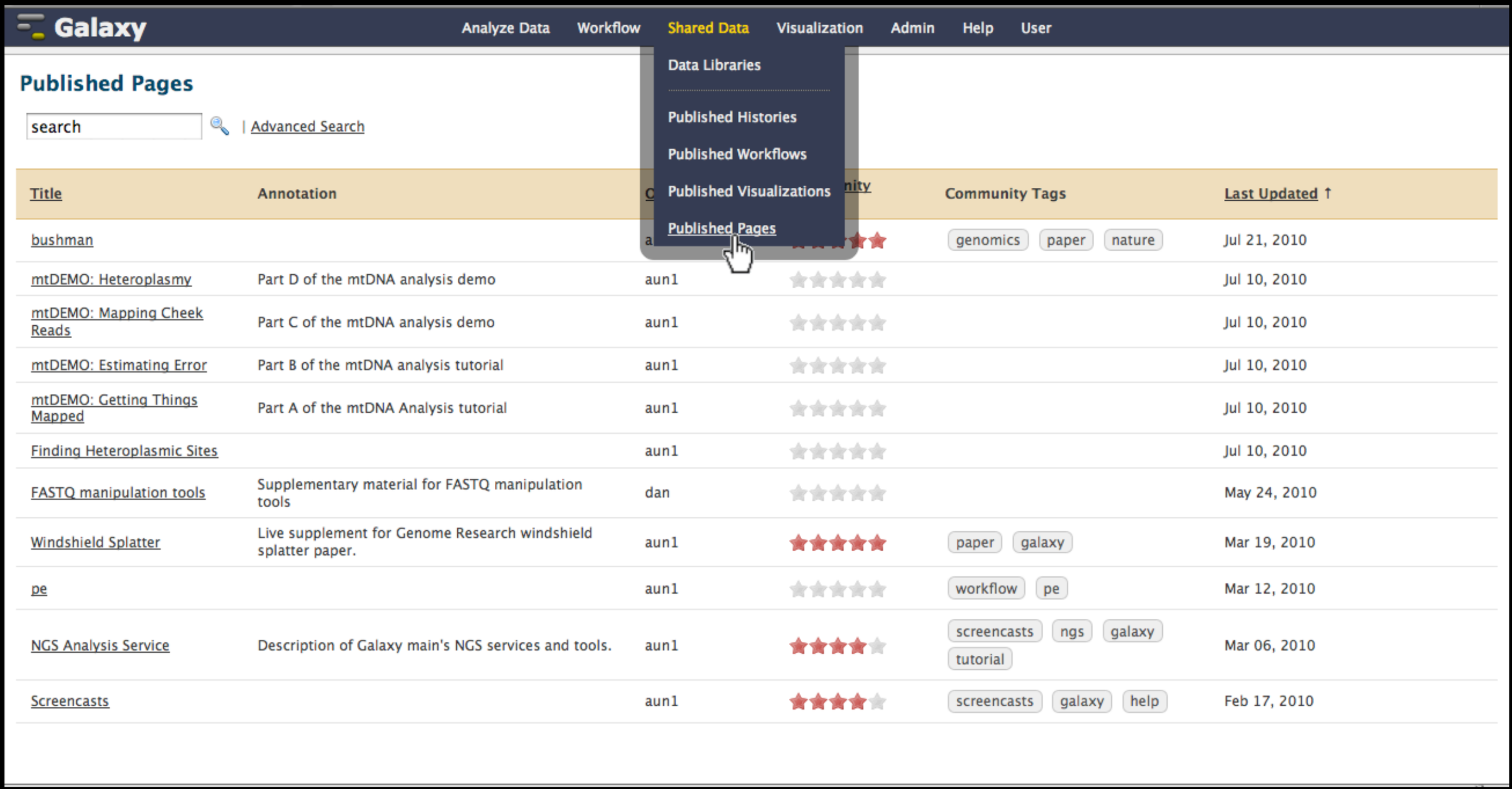
Related Pages
All published pages
Published pages by aun1

Rating
Community (2 ratings, 5.0 average)
Yours

Tags
Community: paper, galaxy
Yours:

http://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter

Sharing



Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010 Aug 25;11(8):R86.

Galaxy 101

http://usegalaxy.org/galaxy101

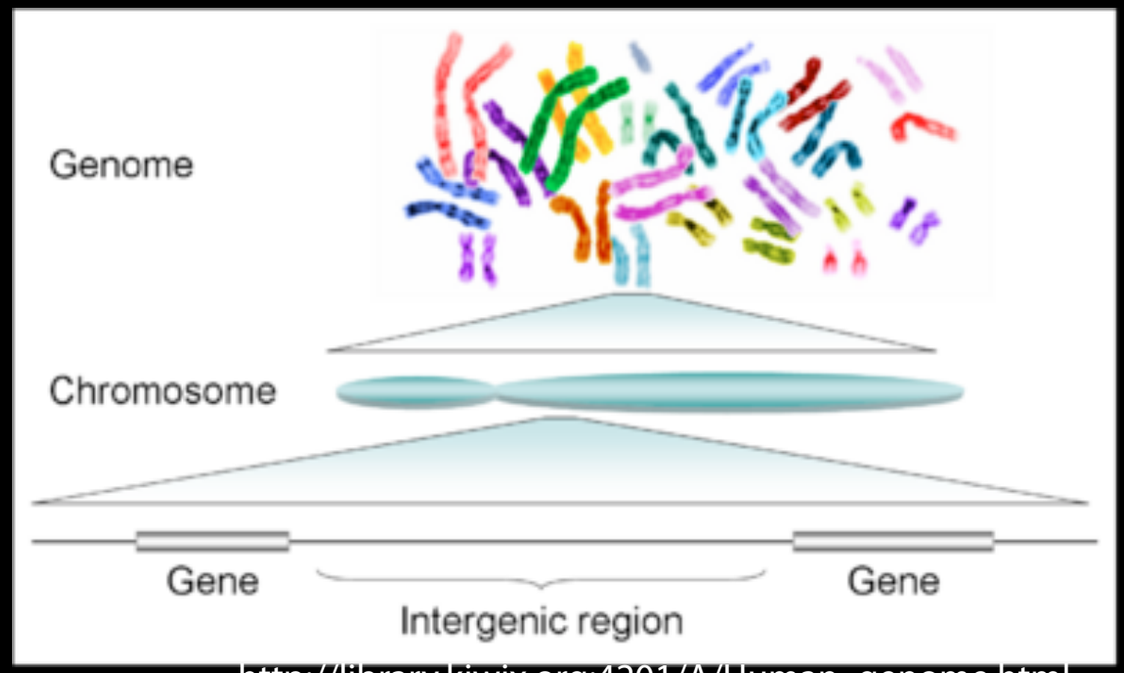
- A simple question...
- Which coding exons have highest number of single nucleotide polymorphisms?

Galaxy 101

http://usegalaxy.org/galaxy101

- Overview
  - Interactively Analyze Data
  - Create reusable generic Workflow
    - Share analysis Results, History, Workflow
- Required Data
  - Genomic Coordinates of
    - Coding Exons
    - SNPs

Genomic Coordinates



http://library.kiwix.org:4201/A/Human_genome.html

>chr1
taaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac

chr1	0	10	first_ten_bases	0	+
chromosome	start	end	name	score	strand

see also: https://bitbucket.org/galaxy/galaxy-central/wiki/GopsDesc https://bitbucket.org/galaxy/galaxy-central/wiki/zero_based_coordinates.pdf

Galaxy 101

http://usegalaxy.org/galaxy101

Interactive Analysis Steps

- Get Genomic data from UCSC Table Browser
- Determine each SNP that overlaps with a specific coding exon
- Calculate count of overlapping SNPs for each exon
- Sort and select exons by greatest SNP counts

Overview

- What is Galaxy?
- Galaxy for Experimental Biologists
- Galaxy for Bioinformaticians

Galaxy: the instant web-based tool and data resource integration platform

- Open Source downloadable package that can be deployed in individual labs
- Zero Configuration, but highly configurable
  - Modularized
    - Easy to plug in your own components
- Straightforward to run your own private Galaxy Server

Why a private Galaxy instance?

- Add new Tools
- Integrate new Data Sources
- Add new display applications
- Secure your private instance for working with sensitive data

The Problem

- You have written a Python script to analyze genomic data and you want to share it with command-line averse colleagues

The Galaxy Solution

- Solution: Integrate the script as a new Tool into your own Galaxy server
- Steps:
  - Obtain and install Galaxy source code (GetGalaxy.org)
  - Write an XML file describing the inputs and outputs and how to execute the script
  - Instruct Galaxy to load the tool

Quick Install

1. Get the latest copy from the repository:

The latest source code can be downloaded from the anonymous Mercurial repository with this command:

```
% hg clone http://www.bx.psu.edu/hg/galaxy galaxy_dist
```

If you don't have Mercurial, tarballs can be downloaded instead: zipped, bzipped or gzipped. However, this makes it more difficult to stay updated in the future since there's no simple way to update your copy.

2. Enable configuration files and download eggs:

Once the source code is downloaded, cd to the `galaxy_dist` directory and run the `setup.sh` script. This will copy sample configuration files and download the proper eggs for your platform:

```
% cd galaxy_dist
% sh setup.sh
```

This step requires Internet access to download the eggs. If the system on which you are installing Galaxy does not have Internet access, please follow the instructions for offline systems on Config/Eggs before attempting this step.

3. Start it up:

At this point Galaxy is ready to run. Simply run the following command:

```
% sh run.sh
```

This will start up the server on localhost and port 8080, so Galaxy can be accessed from your web browser at http://localhost:8080 . To stop the Galaxy server, just hit `Ctrl-c` in the terminal from which Galaxy is running.

GetGalaxy.org http://localhost:8080

Cluster

Cluster intervals of: 6: UCSC Main on Human: knownGene

max distance between intervals: 1 (bp)

min number of intervals per cluster: 2

Return type: Merge clusters into single intervals

Execute

TIP: If your query does not appear in the pulldown menu -> it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns

---

Screencasts!

See Galaxy Interval Operation Screencasts (right click to open this link in another window).

Syntax

- **Maximum distance** is greatest distance in base pairs allowed between intervals that will be considered "clustered". **Negative** values for distance are allowed, and are useful for clustering intervals that overlap.
- **Minimum intervals per cluster** allow a threshold to be set on the minimum number of intervals to be considered a cluster. Any area with less than this minimum will not be included in the ouput.
- **Merge clusters into single intervals** outputs intervals that span the entire cluster.
- **Find cluster intervals; preserve comments and order** filters out non-cluster intervals while maintaining the original ordering and comments in the file.
- **Find cluster intervals; output grouped by clusters** filters out non-cluster intervals, but outputs the cluster intervals so that they are grouped together. Comments and original ordering in the file are lost.

Example

cluster.xml

```
<tool id="gops_cluster_1" name="Cluster">
  <description>[[Cluster]] the intervals of a query</description>
  <command interpreter="python2.4">
    gops_cluster.py $input1 $output -1 $input1_chromCol,$input1_startC
                   -d $distance -m $minregions -o $returntype
  </command>
  <inputs>
    <param format="interval" name="input1" type="data">
      <label>Cluster intervals of</label>
    </param>
    <param name="distance" size="5" type="integer" value="1" help="(bp
      <label>max distance between intervals</label>
    </param>
    <param name="minregions" size="5" type="integer" value="2">
      <label>min number of intervals per cluster</label>
    </param>
    <param name="returntype" type="select" label="Return type">
      <option value="1">Merge clusters into single intervals</option>
      <option value="2">Find cluster intervals; preserve comments and
      <option value="3">Find cluster intervals; output grouped by clus
      <option value="4">Find the smallest interval in each cluster</op
      <option value="5">Find the largest interval in each cluster</opt
    </param>
  </inputs>
  <help>

.. class:: infomark

**TIP:** If your query does not appear in the pulldown menu -> it is n

-----

**Screencasts!**

See Galaxy Interval Operation Screencasts_ (right click to open this l

.. _Screencasts: http://www.bx.psu.edu/cgi-bin/trac.cgi/wiki/GopsDesc

-----

**Syntax**

- **Maximum distance** is greatest distance in base pairs allowed betw
- **Minimum intervals per cluster** allow a threshold to be set on the
- **Merge clusters into single intervals** outputs intervals that span
- **Find cluster intervals; preserve comments and order** filters out
- **Find cluster intervals; output grouped by clusters** filters out n
```

Line: 87 Column: 8 XML Soft Tabs: 2

Input Parameter types

Basic

- Text
- Integer
- Float
- Select
  - Static
  - Dynamic
- Boolean

Genome

- Genome build
- Data column
- Data
- Hidden
- Base URL
- File
- Drill down

Grouping

- Conditional
- Repeat
- Config Files

Running a Production Server

- Use a *real* database server: PostgreSQL, MySQL
- Run on compute cluster resources
- External Authentication: LDAP, Kerberos, OpenID
- Load balancing; proxy support

https://bitbucket.org/galaxy/galaxy-central/wiki/Config/ProductionServer

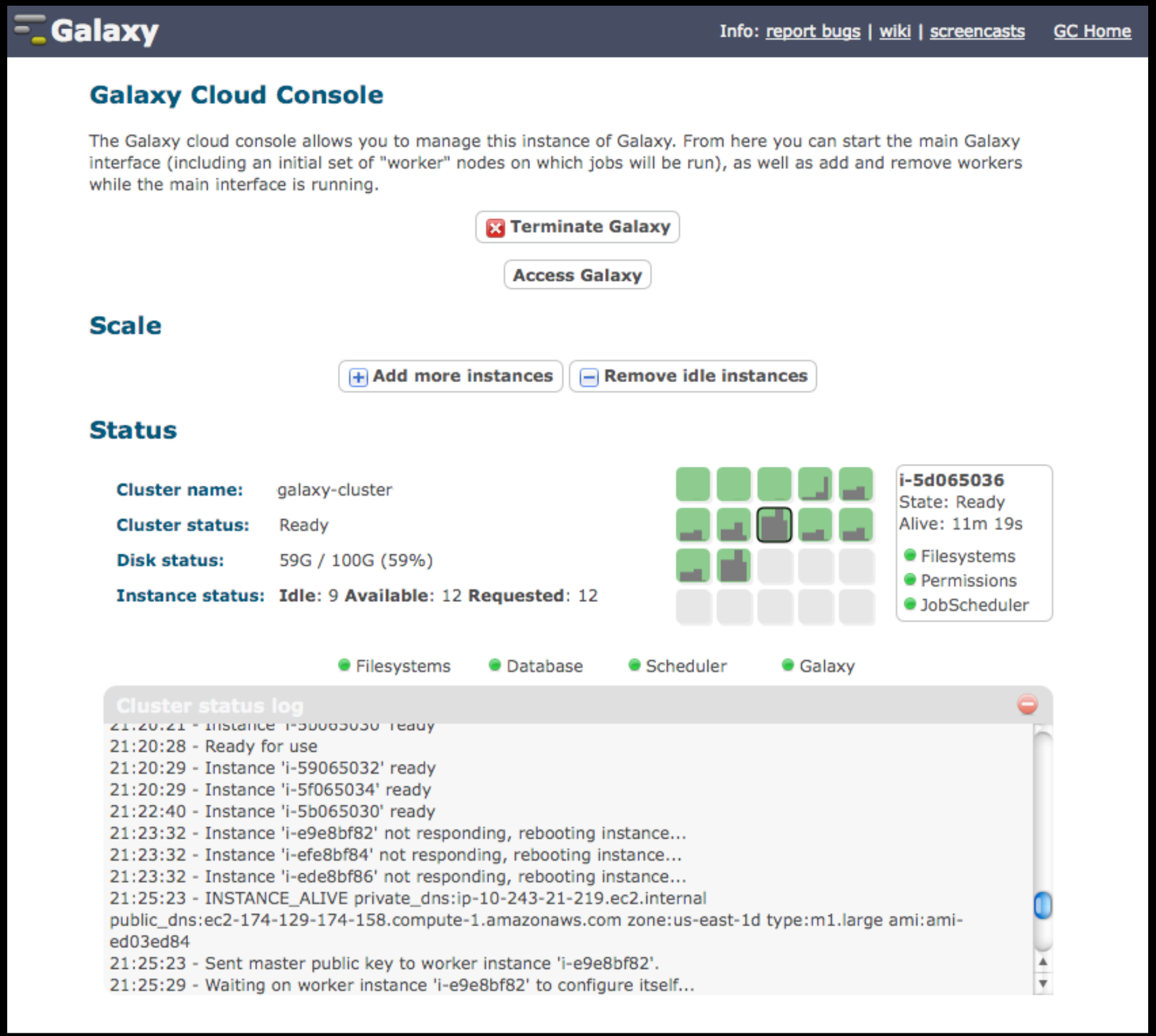
Lack IT knowledge or resources?

- No problem, just use the Cloud

Galaxy on the Cloud

- Availability of Resources are not a Problem
  - Virtually unlimited resources: storage, computing, services
    - No need to maintain machines or personnel
    - Only pay for what you use
- Amazon Elastic Compute Cloud (EC2) and Eucalyptus
- Web-based Galaxy instantiation

Point, Click, Cloud

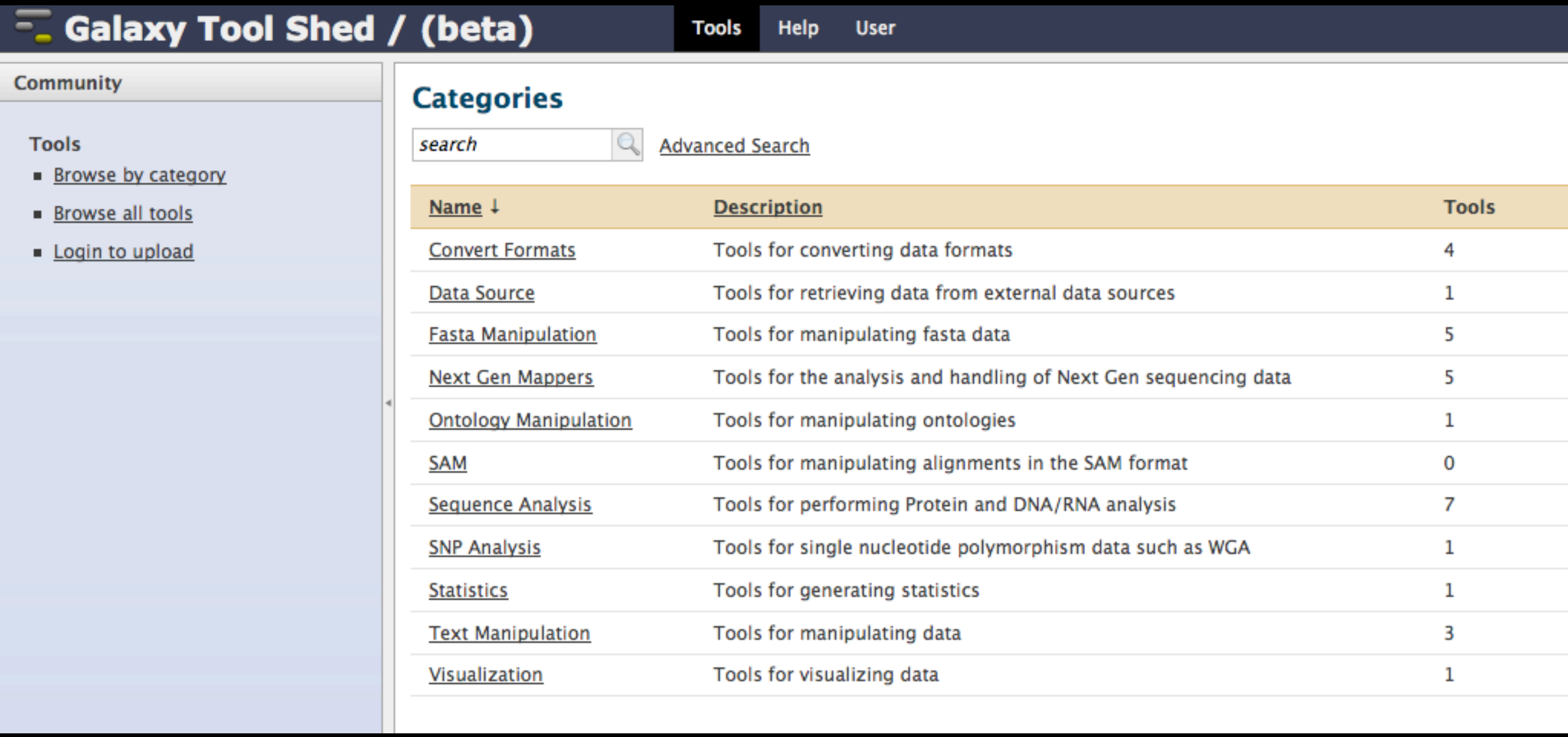


http://usegalaxy.org/cloud

You added a tool, now what?

- Share it with the community!
- Galaxy Tool Shed
  - Upload and Download contributed tools
  - Rate and provide comments and feedback

Get and Contribute Tools



http://usegalaxy.org/community

Using Galaxy

- Use public Galaxy server: UseGalaxy.org
- Download Galaxy source: GetGalaxy.org
- Galaxy Wiki: GalaxyProject.org
- Screencasts: GalaxyCast.org
- Public Mailing Lists
  - galaxy-bugs@bx.psu.edu
  - galaxy-user@bx.psu.edu
  - galaxy-dev@bx.psu.edu

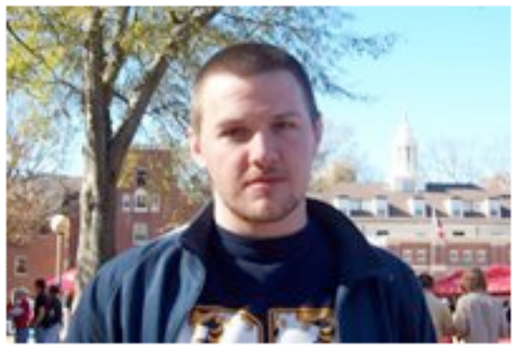
Acknowledgments

- All Members of the Galaxy Team (see them at https://bitbucket.org/galaxy/galaxy-central/wiki/GalaxyTeam)
- Thousands of our users
- GMOD Team
- UCSC Genome Informatics Team
- BioMart Team
- FlyMine/InterMine Teams
- Funding sources
  - NSF-ABI
  - NIH-NHGRI
  - Beckman Foundation
  - Huck Institutes at Penn State
  - Pennsylvania Department of Public Health
  - Emory University

Galaxy Team

Enis Afgan | Emory

Guru Ananda | Penn State



Dannon Baker | Emory

James Taylor | Emory

Jeremy Goecks | Emory

Sergei Kosakovsky Pond | UCSD

Greg von Kuster | Penn State

Dave Clements | Emory

Nate Coraor | Penn State

+ Jennifer Jackson

Ross Lazarus | Harvard | BakerIDI

Kanwei Li | Emory

Anton Nekrutenko | Penn State

Kelly Vincent | Penn State

Two full days of presentations,
workshops, and conversations by
and for Galaxy community members

Galaxy 2011

Community Conference

25-26 May Lunteren, The Netherlands

http://galaxy.psu.edu/gcc2011