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

What it does

This tool converts a Variant Call Format (VCF) file into a Multiple Alignment Format (MAF) custom track file suitable for display at genome browsers. This file should be used for display purposes only (e.g., as a UCSC Custom Track). Performing an analysis using the output created by this tool as input is not recommended: the source VCF file should be used when performing an analysis.

Unknown nucleotides are represented as ‘?’ as required to allow the display to draw properly; these include e.g. reference bases which appear before a deletion and are not available without querying the original reference sequence.

Example

Starting with a VCF:

```
#fileformat=v3.3
#source=Human_Population
#source=population

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878 NA12877 NA12876
1 13320 T . T 3020-18:10.0.10000000 10000000 10000000 10000000
```

Under the following conditions: VCF Source Type: Per Population (file), Name for this population: CHS-JPT Results in the following MAF custom track:

```
test name=Galaxy Custom Track' visibility=pack
```
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
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)
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”

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.

<table>
<thead>
<tr>
<th>Name</th>
<th>Information</th>
<th>Uploaded By</th>
<th>Date</th>
<th>File Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>All SNPs in personal genomes</td>
<td>Summary table of SNPs in all individuals</td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-01-28</td>
<td>676.8 Mb</td>
</tr>
<tr>
<td>Alu insertions in KB1</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-10</td>
<td>14.9 Kb</td>
</tr>
<tr>
<td>Alu insertions in NB1</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-10</td>
<td>6.5 Kb</td>
</tr>
<tr>
<td>KB1_microsatellites.txt</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-15</td>
<td>3.5 Mb</td>
</tr>
<tr>
<td>NB1_microsatellites.txt</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-15</td>
<td>828.5 Kb</td>
</tr>
<tr>
<td>Amino acid differences with functional predictions</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-05</td>
<td>1.1 Mb</td>
</tr>
<tr>
<td>Gene copy numbers in KB1 and other personal genome</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-15</td>
<td>2.1 Mb</td>
</tr>
<tr>
<td>Indels in ABT</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-03</td>
<td>105.3 Kb</td>
</tr>
<tr>
<td>Indels in KB1</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-03</td>
<td>14.2 Mb</td>
</tr>
<tr>
<td>Indels in MD8</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-03</td>
<td>109.8 Kb</td>
</tr>
<tr>
<td>Indels in NB1</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-03</td>
<td>519.5 Kb</td>
</tr>
<tr>
<td>Indels in TK1</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-03</td>
<td>123.2 Kb</td>
</tr>
<tr>
<td>Novel SNPs in ABT</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-09</td>
<td>9.4 Mb</td>
</tr>
<tr>
<td>Novel SNPs in KB1</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-09</td>
<td>16.9 Mb</td>
</tr>
<tr>
<td>Novel SNPs in MD8</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-09</td>
<td>594.1 Kb</td>
</tr>
<tr>
<td>Novel SNPs in NB1</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-09</td>
<td>4.1 Mb</td>
</tr>
<tr>
<td>Novel SNPs in TK1</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-09</td>
<td>722.6 Kb</td>
</tr>
<tr>
<td>Sequenced exon-containing intervals</td>
<td></td>
<td><a href="mailto:greg@bx.psu.edu">greg@bx.psu.edu</a></td>
<td>2010-02-03</td>
<td>3.1 Mb</td>
</tr>
</tbody>
</table>

For selected items: [Import into your current history](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
Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement

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 names 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
11: Join two Queries on data 9 and data 10
12: Filter on data 11
13: Fetch taxonomic representation on data 12
14: Find lowest diagnostic rank on data 13
15: Summarize taxonomy on data 14
16: Draw phylogeny on data 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 Item 1, 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 (Ibuson 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 tools) can be imported as text.

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

>chr1
taaccctaac
taaccctaac
taaccctaac
taaccctaac
taaccctaac
taaccctaac
taaccctaac
taaccctaac
taaccctaac
taaccctaac

chr1 0 10 first_ten_bases 0 +

chromosome start end name score strand

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

**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.

---

**Screeencasts!**

See Galaxy Interval Operation [Screeencasts](#) (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 output.
- **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**

```
| Query | Find clusters | Merge clusters |
```

---

**TIP:** If your query does not appear in the pulldown menu -> it is not in interval format.
Input Parameter types

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

- 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

Galaxy Cloud Console

The Galaxy cloud console allows you to manage this instance of Galaxy. From here you can start the main Galaxy interface (including an initial set of "worker" nodes on which jobs will be run), as well as add and remove workers while the main interface is running.

- Terminate Galaxy
- Access Galaxy

Scale

- Add more instances
- Remove idle instances

Status

Cluster name: galaxy-cluster
Cluster status: Ready
Disk status: 59G / 100G (59%)
Instance status: Idle: 9 Available: 12 Requested: 12

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convert Formats</td>
<td>Tools for converting data formats</td>
<td>4</td>
</tr>
<tr>
<td>Data Source</td>
<td>Tools for retrieving data from external data sources</td>
<td>1</td>
</tr>
<tr>
<td>Fasta Manipulation</td>
<td>Tools for manipulating fasta data</td>
<td>5</td>
</tr>
<tr>
<td>Next Gen Mappers</td>
<td>Tools for the analysis and handling of Next Gen sequencing data</td>
<td>5</td>
</tr>
<tr>
<td>Ontology Manipulation</td>
<td>Tools for manipulating ontologies</td>
<td>1</td>
</tr>
<tr>
<td>SAM</td>
<td>Tools for manipulating alignments in the SAM format</td>
<td>0</td>
</tr>
<tr>
<td>Sequence Analysis</td>
<td>Tools for performing Protein and DNA/RNA analysis</td>
<td>7</td>
</tr>
<tr>
<td>SNP Analysis</td>
<td>Tools for single nucleotide polymorphism data such as WGA</td>
<td>1</td>
</tr>
<tr>
<td>Statistics</td>
<td>Tools for generating statistics</td>
<td>1</td>
</tr>
<tr>
<td>Text Manipulation</td>
<td>Tools for manipulating data</td>
<td>3</td>
</tr>
<tr>
<td>Visualization</td>
<td>Tools for visualizing data</td>
<td>1</td>
</tr>
</tbody>
</table>

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

http://galaxy.psu.edu/gcc2011