Galaxy: The Basics

How to get started with Galaxy

Ann Black-Ziegelbein
annblack@eng.uiowa.edu
Senior Application Development
Molecular Otolaryngology and Renal Research Laboratories
How do I use Galaxy? The basics.
Accessing Galaxy

- Galaxy is hosted publically at Penn State:
  - http://usegalaxy.org/
  - Can use anonymously
  - Registration for an account is simple

- Galaxy is also hosted at Iowa
  - https://galaxy.hpc.uiowa.edu/
  - Hawkid & password Secure
  - No Registration Necessary
    - Account is auto created on first access
  - Only accessible via campus or hospital intranets
    - Includes VPN (Virtual Private Network) access
      - http://its.uiowa.edu/apps2/support/vpn
Inspecting Galaxy

Top menu is for switching between different types.
Getting Data Into Galaxy @ Iowa

- Upload from local Computer (< 2 GB)
- SFTP to galaxy.hpc.uiowa.edu
- UCSC Browser @ Iowa
- Shared or Imported Histories
- Shared Data Libraries
- URL (Iowa DNA Core)
Galaxy: Where is my data?

- **DataSets: The Basics**
  - All data files in Galaxy are represented as a galaxy DataSet. (Imports, inputs, outputs from analysis steps, intermediate files, etc).
  - A dataset’s content, once created, will NEVER change. Each new operation on the data will create a NEW Galaxy DataSet.
    - DataSet properties such as name, reference database, file type **CAN** be changed – but does not affect the content within the DataSet.
  - DataSets are accessed in User “Histories”.
    - ID Number based on order of creation within a History.
    - Named – Can be renamed to anything you like.

**Tip: Mouse click dataset name to expand/collapse summary**
Galaxy Histories

- History is a collection of inputs and actions taken in Galaxy
  - Serves as an audit trail of all steps performed
- You can have multiple histories
  - The active history shown in the right-hand panel is “Current History”
- Histories can be annotated, tagged, and shared
- DataSets in a history are numbered based on creation order within that history.

**Best Practices:**
- Name your histories
- Create a new history for each high-level task/analysis you want to perform
Where did my history go?

- Whenever Galaxy can not detect a pre-existing History, you get an “Unnamed history”
- Browser Cookies tell Galaxy about your last active History
- Accessing Galaxy from a new computer, browser, clear your browser cookies/cache, etc. You get a new “Unnamed history”
- Oh, and History names do not need to be unique. You can have 50 + “Unnamed history” histories
Navigating Galaxy Histories

- All Histories once created are automatically saved. Any action taken to datasets in the history are saved within that history
  - IE there is no need to concretely hit a “Save” button/menu action, etc.
- To see all histories and toggle your “current History” to a previously created one, access the “Saved Histories” menu item
- Select the name of the history you want to make active
Best Practices:

- Name your histories something meaningful
- Leverage tags (keywords) for searching
- Create a new history for each high level task/analysis you want to perform

Iowa Initiative in Human Genetics :: http://www.medicine.uiowa.edu/humangenetics
University of Iowa Bioinformatics :: http://bioinform-div.healthcare.uiowa.edu/
Using data in Galaxy @ Iowa

- All data in your “current history” can be inputs to Galaxy Tools.
  - Datasets can be “copied” between Histories
- A dataset’s content can never be changed once created (imported or output of a tool)
  - Only deleted or have attributes of the file change (like name)
- A dataset can be shared between users

Data in Galaxy is NOT Backed up!
Archiving Data From Galaxy

- All DataSets from Galaxy can be downloaded for local archival.
  - From History: Manual procedure – no batch method.
  - From Data Library: Can download everything in library to a compressed archive
- Histories can also be exported from Galaxy and re-imported
  - This will preserve an audit trail of all steps taken … but not file content
Managing DataSets (animation)

- **(X)** Icon marks the dataset for deletion and hides it in history view.
- **Disk icon** will download the file to your local computer.

Additional text:

Eye icon shows a portion of the file content in the browser.
Pencil icon changes attributes of dataset (not content).
(i) icon shows the file's metadata details and information about how it was generated.

Re-Run icon will pre-populate work area with the tool and all configuration options used to generate the dataset.
Disk icon will download the file to your local computer.

**Eye icon** marks the dataset as hidden in history view.
Managing DataSets (no animation)

- **Eye icon**: shows a portion of the file content in the browser.
- **Pencil icon**: changes attributes of the dataset (not content).
- **Re-Run icon**: will pre-populate work area with the tool and all configuration options used to generate the dataset.
- **(X) Icon**: marks the dataset for deletion and hides it in history view.
- **Disk icon**: will download the file to your local computer.
Removing Datasets

- Can view storage consumption
  - Total (over all datasets)
  - Per History
  - For each dataset

- Datasets & Histories can be deleted
  - Best Practice: Delete anything you don’t need
    - Shared pool of storage.
    - We may ask you to clean up your data on a deadline
  - Deleted Datasets can be Restored
    - Datasets are not really removed from storage and Galaxy’s database until backend clean/purge scripts are ran by galaxy administrator.
    - Datasets can be “undeleted” and “restored” unless Purged from disk.
    - Only restorable during window of time
Accessing Tools in Galaxy

- Tools are accessible in the “Analyze Data” work area, left hand panel
  - Tip: Can search for Tools by keyword
  - Tools are grouped together by type
- Current Galaxy at Iowa is a subset of Penn State Galaxy tools.
- Click on tool category to expand/collapse
- Click tool name to run
Executing a Tool in Galaxy

- Select the tool from the Tool menu
- Center Work Area Panel will populate with a wizard for executing the tool
- Configure all the options (help typically exists at the bottom or in line)
- Click “Execute”
Galaxy: Other cool stuff

- **Workflows**
  - Way to automate a series of steps, can be extracted from a history
  - Can be shared, imported, exported, published
  - Can receive emails upon completion

- **Sharing/Collaboration**
  - Histories and datasets can be shared with other Galaxy users or via a URL link to any other person on campus
  - All datasets/histories can be annotated with notes
  - Failures in Galaxy provide you a mechanism to email the Galaxy administrator
    - This will automatically provide history details and access to the administrator as well as open up a problem ticket
Other technology approaches

- **Helium: Iowa’s High Performance Compute Cluster**
  - Power User
  - Good for custom or cutting edge tools
  - Command Line
    - Several bioinformatics tools available
    - Can install additional tools locally
    - [https://www.icts.uiowa.edu/confluence/display/ICTSit/Helium+Cluster+Overview+and+Quick+Start+Guide](https://www.icts.uiowa.edu/confluence/display/ICTSit/Helium+Cluster+Overview+and+Quick+Start+Guide)

- **Galaxy in the cloud**
  - Access to tools not yet available in Galaxy @ Iowa
  - Does cost – based on how much you use
  - Skirt Helium queuing
  - [http://wiki.g2.bx.psu.edu/CloudMan](http://wiki.g2.bx.psu.edu/CloudMan)
Galaxy @ Iowa

- Galaxy @ Iowa
  - Built into the Iowa Helium HPC Cluster
  - No data storage caps
  - Subset of Penn State Tools & Reference Genomes
    - More available upon request
    - Goal: is to be superset of Penn State Tools

- Is there a charge for Galaxy access? No 😊
Lab 1

Do stuff in Galaxy to become familiar with:

1. Finding tools
2. Manage Datasets
3. Manipulating Data in Galaxy
Introduction to Lab 1

1. Import/Load Data into Galaxy
2. Explore Existing Target and Bait Interval Files
3. Reverse Engineer Target Regions to Gene Names
4. Extract a target bed file from a list of gene names to use in analysis
5. Compare targeted capture design (bed) file against a whole exome bait interval file
Lab Material

• Slides & Instructions found off of the IIHG Bioinformatics Course II link from:
  • http://tinyurl.com/iowa-galaxy-ed
  • Or the longer version:
  • http://wiki.uiowa.edu/display/galaxyusers/IIHG+Galaxy+Eduction

• Can view on Class Room Monitors and Use Laptops for Running Labs (or vice versa)

• Please ASK questions. We are hear to help! 😊
Goals of Lab 1

- By the end of Lab 1 you should:
  - Be familiar with the overall Galaxy web interface
  - Understand how to access Shared Data Libraries
  - Familiarity on invoking Galaxy tools and manipulating datasets within Galaxy
  - Have an introduction to the UCSC Browser and how to download data from it into Galaxy
Course Shared Library

- All lab data is stored and accessible out of a shared Galaxy Data Library
- **Shared Data Libraries**
  - Shared common repository of data between users
  - Must be initially created by an administrator
  - Can have user & group permissions for post-creation management
    - Adding Files
    - Administering privileges
    - Viewing Files
    - Etc
  - Accessible via top menu
Lab 1: Let’s begin!

1. Import/Load Data into Galaxy
2. Explore Existing Target and Bait Interval Files
3. Reverse Engineer Target Regions to Gene Names
4. Extract a target bed file from a list of gene names to use in analysis
5. Compare targeted capture design (bed) file against a whole exome bait interval file
Step 1A: Log into Galaxy

- Login to Galaxy:
  - \(\text{http://galaxy.hpc.uiowa.edu}\)
  - Enter HawkID & HawkPswd when prompted
  - Do not check “Remember credentials” unless on private computer
Step 1A: Log into Galaxy

New users will have an empty history. Existing users may see their last current history.
Step 1B: Create and Name a New History

- Best Practice is to create a new history for each new high level task.
- Let's create a new history and name it.
  1. Select Options -> Create New
  2. Click on History Name to edit. Enter “Lab 1 Analysis” & hit Enter key.
- Histories are always saved automatically. Let's view all saved histories
  3. Select Options -> Saved Histories
Step 1C: Import Data To History

- Navigate to Shared Library View:
  1. Select “Shared Data -> Data Libraries”
  2. Find and select “IIHG Bioinformatics Course II”
  3. Select both .bed files
     1. SureSelect_50MB_exome.bed
     2. Otoscope_v4.bed
  4. Drop Down Box “Import to Current history”
  5. Click “Go”

Iowa Initiative in Human Genetics :: http://www.medicine.uiowa.edu/humangenetics
University of Iowa Bioinformatics :: http://bioinformatics.healthcare.uiowa.edu/
Step 1C Cont.: Verifying Data

• After “Go” button clicked, you should see a confirmation message.
• Go back to history to see datasets
  1. Select “Analyze Data” in top menu
  2. Verify you have two new datasets in history

Iowa Initiative in Human Genetics :: http://www.medicine.uiowa.edu/humangenetics
University of Iowa Bioinformatics :: http://bioinform-div.healthcare.uiowa.edu/
Step 1D: Inspecting Bed Files

- You can view File Content (First couple MB for large files) within the Galaxy Browser
  1. Select the “eye” icon next to the file name
     • “Poke it in the eye!”
  2. Investigate both BED files
     • OtoSCOPE_v4.bed : Targetted Capture Design File
     • SureSelect_50MB_exome.bed: BAIT Interval File for whole exome capture
Step 1E: Inspecting Bed Files Cont.

- You can view File MetaData for each file within the Galaxy Browser.
  - Meta, what? Metadata are associated attributes of a file like type, associated genome build, structure information, etc.
  1. Expand the Collapsed Dataset by clicking on the name.
  2. Select the “Pencil” Icon next to the Dataset name.
Understanding the Bed Format

- **BED in NGS:** not something you sleep in.
- **Secret? Basic Tab Delimited File**
  - With expected data in certain orders
  - Basic 3 Column Bed Format:
    - `{CHROM}` {START} {END}
  - Optional 9 more columns:
    - Name: Name of Bed Line
    - Score: 0-1000
    - Strand: + or –
    - ...
    - [http://genome.ucsc.edu/FAQ/FAQformat.html#format1](http://genome.ucsc.edu/FAQ/FAQformat.html#format1)
Baits vs. Target Interval Files

- Bait files depict the sequence capture regions. (Can download this from the capture kit manufacturer)
- Target interval files depict any area you think is biologically relevant.
- A Target interval file can be a bait file, or a subset of Bait intervals
  - Example: For whole exome sequencing you may only care about a specific target sub-region of what was sequenced
UCSC Browser @ Iowa

- UCSC Browser contains reference genomes for several species with annotations
  - Can search, download, visualize areas of interest compared to your aligned sequences
  - Powerful queries and exporting capabilities
- UCSC Browser Main
  - http://genome.ucsc.edu/
  - https://cgwb.nci.nih.gov/goldenPath/help/hgTablesHelp.html#Introduction
- Iowa maintains a copy of the UCSC browser
  - https://research.icts.uiowa.edu/genomebrowser/gbpublic
  - Updated December 2011 (semi-recently)
Lab 1: Practicum

Follow Lab Handouts. Please ask questions 😊
Introduction to Lab 1

1. Import/Load Data into Galaxy
2. Explore Existing Target and Bait Interval Files
3. Reverse Engineer Target Regions to Gene Names
4. Extract a target bed file from a list of gene names to use in analysis
5. Compare targeted capture design (bed) file against a whole exome bait interval file
Galaxy Terminology

- Do you speak Galaxy?
  - **DataSet**: Any input, output or intermediate set of data. A representation of a physical file.
  - **History**: a collection of datasets: inputs, analysis steps, intermediate steps, and outputs. Histories are named (by the user) and a user can have multiple Histories.
  - **Workflow**: a repeatable, ordered series of computational steps where the output of one step is automatically fed into the input of another.
  - **Share**: make something available to one or more explicitly identified galaxy users.
  - **Publish**: make something accessible to all galaxy users.
  - **Data Library**: a collection of datasets in a common shared location for easy group access.