Galaxy data libraries and sample tracking at NGS facilities

Greg Von Kuster
Galaxy Team
What are data libraries?
a hierarchical container for datasets
How do I put data in?
How do I put data in?

- Upload a single file
- Import datasets from a Galaxy history
- Upload a directory of files
- Upload files from file system paths
uploading a directory

- The directory location is configurable
- Includes an option to not copy files into Galaxy’s normal files directory, leaving them in their original location
uploading from file system paths

• Allows for any number of file system paths (files or directories), saving the directory structure if desired

• Includes an option to not copy files into Galaxy’s normal files directory, leaving them in their original location
Oops, I uploaded the wrong data, what can I do?
delete it!

- Folders and datasets can be deleted at any level (state is saved for contents, if any)
- Deleted items are not displayed
Can I undelete?
yes! Show deleted items
deleted items are red
...and you can undelete
How do I use the data in a library?
using the data

- Users that can access a library dataset can import it into their Galaxy history for analysis
using the data

- Users that can access a library dataset can import it into their Galaxy history for analysis.
- Importing a library dataset into a history creates a pointer to the same single disk file, minimizing disk space.
using the data

- Users that can access a library dataset can import it into their Galaxy history for analysis
- Importing a library dataset into a history creates a pointer to the same single disk file, minimizing disk space
- Versioning is supported for library datasets
Can I protect the data?
data library security

• Restricts access to the entire library or specific datasets contained within it

• The default is no restriction, making items “public”
so Dick can see...
...but Jane can see
data library security also...

• Grants permission to specific users to perform actions on library items
• The default is no permission to do anything (except access)
so Dick can...
but Jane can...
How does this work?
flexible security policies
flexible security policies

- Use built-in role-based authorization
flexible security policies

- Use built-in role-based authorization
- Library level: access library, add, modify, manage permissions
flexible security policies

- Use built-in role-based authorization
- Library level: access library, add, modify, manage permissions
- Folder level: add, modify, manage permissions
flexible security policies

- Use built-in role-based authorization
- Library level: access library, add, modify, manage permissions
- Folder level: add, modify, manage permissions
- Datasets: access, modify, manage permissions
and...

- Security settings are automatically inherited downward in the hierarchy, but can be overridden, providing distinct security policies at any level in the hierarchy
What’s the difference between the “library access” permission on a library and the “access” permission on a dataset?
subtle, but important

- If a library is public but a contained dataset is not, anyone can see the library, but not everyone will see the dataset
subtle, but important

• If a library is public but a contained dataset is not, anyone can see the library, but not everyone will see the dataset

• If a library is restricted, but a contained dataset is public, only those that can see the library will have access to the dataset
subtle, but important

• If a library is public but a contained dataset is not, anyone can see the library, but not everyone will see the dataset

• If a library is restricted, but a contained dataset is public, only those that can see the library will have access to the dataset

• Users having any role associated with “library access” can see the library, but users must have all roles associated with “access” on the dataset in order to see the dataset
data library templates
data library templates

- Associate information with a library and its contents (e.g., associate sequencing run parameters with the resulting dataset)
data library templates

- Associate information with a library and its contents (e.g., associate sequencing run parameters with the resulting dataset)
- Template layout can easily be defined or changed at any time
data library templates

- Associate information with a library and its contents (e.g., associate sequencing run parameters with the resulting dataset)
- Template layout can easily be defined or changed at any time
- Templates can be inherited downward, but can be overridden at any level
data library templates

• Associate information with a library and its contents (e.g., associate sequencing run parameters with the resulting dataset)

• Template layout can easily be defined or changed at any time

• Templates can be inherited downward, but can be overridden at any level

• Template inheritance can be turned on or off at any level
data library templates

- Associate information with a library and its contents (e.g., associate sequencing run parameters with the resulting dataset)
- Template layout can easily be defined or changed at any time
- Templates can be inherited downward, but can be overridden at any level
- Template inheritance can be turned on or off at any level
- Templates can be deleted from any item
a dataset template

This is the latest version of this library dataset

Information about Alu insertions in KB1

Message:

Uploaded by:
greg@bx.psu.edu

Date uploaded:
2010-02-10

Build:
hg18

Miscellaneous information:
uploaded bed file
500 regions

 Peek:
 chr1 4055016 4055320 AluYb8
 chr1 26362411 26362721 AluYa5
 chr1 28120394 28120699 AluYb8
 chr1 57015512 57015523 AluYa5
 chr1 62163190 62163528 AluY
 chr1 62924421 62924746 AluYb8

Other information about Alu insertions in KB1

Column Assignments

This is a tab separated file which reports the detected Alu insertions in the human reference genome (NCBI Bld. 36.1) relative to the KB1 genome. The file has the following columns:

1. Chrom  chr  chromosome
2. Start   start 0 based start position on the chromosome 'chr'
3. End     end 0 based end position on the chromosome (0 based half-open start,end)
4. Name    alu name of the ALU element
Galaxy sample tracking streamlines the delivery of data from sequencing runs to customers
how?
example facility process

- A customer submits a sequencing request to a facility and delivers the samples, selecting a data library for the run results
example facility process

- A customer submits a sequencing request to a facility and delivers the samples, selecting a data library for the run results
- The facility bar codes the samples, scanning the tubes at each station in the lab
example facility process

- A customer submits a sequencing request to a facility and delivers the samples, selecting a data library for the run results
- The facility bar codes the samples, scanning the tubes at each station in the lab
- The customer can watch the progress
example facility process

• A customer submits a sequencing request to a facility and delivers the samples, selecting a data library for the run results

• The facility bar codes the samples, scanning the tubes at each station in the lab

• The customer can watch the progress

• When the run is complete, the resulting data is transferred to the requested data library
who does what?

customer
- create Galaxy sequencing request
- add samples
- submit request

facility manager
- transfer datasets to Galaxy data library
- update sample states
- assign barcodes to samples
example sequencing request lifecycle

<table>
<thead>
<tr>
<th>State</th>
<th>Description</th>
<th>Last Update</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Received</td>
<td>Sample Received</td>
<td>27 minutes ago</td>
<td></td>
</tr>
<tr>
<td>Run Started</td>
<td>Run Started</td>
<td>Feb 25, 2010</td>
<td>bar code scanner</td>
</tr>
<tr>
<td>Library Complete</td>
<td>Library Complete</td>
<td>Feb 25, 2010</td>
<td>bar code scanner</td>
</tr>
<tr>
<td>Library Started</td>
<td>Library Started</td>
<td>Feb 25, 2010</td>
<td>bar code scanner</td>
</tr>
<tr>
<td>Sample Received</td>
<td>Sample Received</td>
<td>Feb 25, 2010</td>
<td>Sample added to the system</td>
</tr>
<tr>
<td>New</td>
<td>New</td>
<td>Feb 25, 2010</td>
<td>Samples created.</td>
</tr>
</tbody>
</table>

Change current state

Select the new state of the sample from the list of possible state(s)

Sample Received (Current)

Comments

Optional

Save
transferring the data

- Remote file browser to select datasets on the sequencer and transfer them to the requested Galaxy data library
What if my facility does things differently?
You define the process

- The layout of the request and sample forms is defined by the lab and can be changed over time
You define the process

- The layout of the request and sample forms is defined by the lab and can be changed over time

- Lifecycle “states” of the request are defined by the “stations” in the lab
You define the process

- The layout of the request and sample forms is defined by the lab and can be changed over time
- Lifecycle “states” of the request are defined by the “stations” in the lab
- Bar code scanners can be used, but manual data entry is also supported
You define the process

- The layout of the request and sample forms is defined by the lab and can be changed over time
- Lifecycle “states” of the request are defined by the “stations” in the lab
- Bar code scanners can be used, but manual data entry is also supported
- Configure the sequencer information for communication with Galaxy
You define the process

- The layout of the request and sample forms is defined by the lab and can be changed over time
- Lifecycle “states” of the request are defined by the “stations” in the lab
- Bar code scanners can be used, but manual data entry is also supported
- Configure the sequencer information for communication with Galaxy
- Permissions to submit requests is granted to specific users
What if my facility uses a LIMS? Will it work with Galaxy?
Yes!

- Galaxy sample tracking complements existing LIMS applications, it is not intended to replace them.
Yes!

- Galaxy sample tracking complements existing LIMS applications, it is not intended to replace them.

- Galaxy uses a generic messaging engine with a very simple XML api for communication with the sequencer. This same messaging engine can communicate with a LIMS application.
For more information about data libraries...
For more information about data libraries...

- See me at the bar
bonus topic
 Wouldn’t it be cool if there was a place that allowed me to upload tools that I’ve developed, and share them with others?
How about if this place allowed me to easily find someone else’s tool that I’m interested in using in my own Galaxy instance?
Maybe allow me to browse tools by category, or search for tools using names or descriptions?
Would’t it be nice if I could upload new and improved versions of tools that are already available there...
...you know, since I'm a better programmer than the one that wrote it the first time?
the time has come...
http://usegalaxy.org/community