Connecting Galaxy to a Data Repository

Park Lab @ Harvard Medical School
Richard Park

In collaboration with
Hide Lab @ Harvard School of Public Health

Funded in part by an Agilent Technologies Emerging Insights Grant
Goals

Support data analysis on large collections of samples
Both on private and public data sets

Support efficient deployment of new algorithms and tools
For collaborators to run analyses on their data without our help
For the general public as supplement to publications

Support exploration and visualization of analysis results
With interactive, browser-based visualization tools
Difficulties with Goals

- Searching/Finding relevant data
- Downloading and processing data
- Dealing with the magnitude of data
- Sharing analyses and data
Explosion of Big Data

**DATA DUMP**
The number of gene-expression data sets in publicly available databases has climbed to nearly one million over the past decade.

- **ArrayExpress**
- **GEO**

![Graph showing data deposition growth](chart1.png)

**Output Skyrocketing**

**Number sequenced**

- 1,000,000 genomes sequenced
- 100,000
- 10,000
- 1,000
- 100
- 1

![Graph showing output growth](chart2.png)

This figure of 30,000 comes from a Nature estimate and excludes genomes sequenced by biotechnology companies.

Source: U.S. National Human Genome Research Institute, Nature

Explosion of Big Data

**Sequence Read Archive (SRA)** (7/25/2012)
- **Open**: 255 terabases, 295 TB
- **Total**: 608 terabases, 506 TB

**Gene Expression Omnibus (GEO)** (7/25/2012)
- 10,314 platforms, 777,398 samples
- 2,720 datasets, 31,544 series

**Cancer Genomics Hub (CGHub)** (7/16/2012)
- **Current**: 31 TB, 22,760 files
- **Planned**: 20k cases, 200 GB / case, 100TB

**EBI ArrayExpress** (7/25/2012)
- 31,465 experiments
- 914,184 assays
“... a production facility composed of a group of [...] processes [...] refining certain materials or converting raw material into products of value ...”

Source: Wikipedia & *geekly on Flickr*
Implementation

Refinery web application is built on top of Django

Django is a popular Python web application framework
Using Postgresql database as backend
Long running tasks are managed through Celery & RabbitMQ

Search enhanced using Apache Solr

Typeahead functionality, Facets to find and filter for similar experiments

Web application controls Galaxy through its API

Refinery and Galaxy need to be able to access the same disk space
Overview
Repository

Meta information stored in ISA-TAB format

- Automated import of public data from ArrayExpress/GEO and SRAs
- Automated import from Harvard Stem Cell Commons (Hide Lab)

Private data sets and analysis results stored locally

- e.g. BAM files and intermediate results

Public data stored remotely

- Previews available, automated download for analyses or loaded from local cache
Repository

The *Investigation, Study, Assay* (ISA) tab-delimited (TAB) format is a framework to collect and communicate complex metadata.

Builds on the existing *MAGE-TAB* format.

**Adopted by:** Pride (proteomics), ENA (genomics), ArrayExpress (transcriptomics), Novartis Institutes for BioMedical Research, NCI caBIG, etc...

Sansone et al.. The first RSBI (ISA-TAB) workshop: "Can a simple format work for complex studies?". OMICS 2008;12:143-149.
Why not (just) use Galaxy?

Lacks support to store structured meta information about studies

Becomes an issue when data sets contain large numbers of samples with complex relationships

Difficulty with large number of samples

Significant manual intervention required to load 100s of files. Difficult to organize and keep track of outputs.

Hard to estimate progress of long running analyses
Extensions to Galaxy API

Additional features

- Workflows can be downloaded and imported
- Dynamic workflows are generated based on the number of inputs
- Tool parameters are configurable (*to some degree*)
- Selected output files from workflows are stored

Improved support through Python class

- Creation and deletion of libraries and histories
- Monitoring of workflow progress through API (using history information)
Galaxy Workflow (Expansion)

Original (1x)

Expanded (4x)
Galaxy Workflow (Reduction)

Original

Reduction (6x)
Data and resource **sharing is a central concept** in Refinery.

*Source: f-r-a-n-k on Flickr*
Sharing

Resources owned by a group are managed by one or more group managers.

Group Ownership

Resources owned by a group are managed by one or more group managers.
## Assay Files

<table>
<thead>
<tr>
<th>Tdf File</th>
<th>Data Type</th>
<th>Cell Or Tissue</th>
<th>Species</th>
<th>Data Source</th>
<th>Antibody</th>
<th>Treatment</th>
<th>Tissue Order</th>
<th>Bin Size</th>
<th>Target</th>
<th>Lab</th>
</tr>
</thead>
<tbody>
<tr>
<td>exp_file</td>
<td>Histone ChIP-seq</td>
<td>AH</td>
<td>D. melanogaster</td>
<td>modENCODE</td>
<td>H3K27ac</td>
<td>None</td>
<td>4</td>
<td>10</td>
<td>H3K27ac</td>
<td>NA</td>
</tr>
<tr>
<td>exp_file</td>
<td>Histone ChIP-seq</td>
<td>EL</td>
<td>D. melanogaster</td>
<td>modENCODE</td>
<td>H3K27ac</td>
<td>None</td>
<td>2</td>
<td>10</td>
<td>H3K27ac</td>
<td>NA</td>
</tr>
<tr>
<td>exp_file</td>
<td>Histone ChIP-seq</td>
<td>L3</td>
<td>D. melanogaster</td>
<td>modENCODE</td>
<td>H3K27ac</td>
<td>None</td>
<td>3</td>
<td>10</td>
<td>H3K27ac</td>
<td>NA</td>
</tr>
<tr>
<td>exp_file</td>
<td>Histone ChIP-seq</td>
<td>NH-A</td>
<td>H. sapiens</td>
<td>ENCODE</td>
<td>H3K27ac</td>
<td>None</td>
<td>2</td>
<td>50</td>
<td>H3K27ac</td>
<td>Broad</td>
</tr>
<tr>
<td>exp_file</td>
<td>Histone ChIP-seq</td>
<td>NHEK</td>
<td>H. sapiens</td>
<td>ENCODE</td>
<td>H3K27ac</td>
<td>None</td>
<td>2</td>
<td>50</td>
<td>H3K27ac</td>
<td>Broad</td>
</tr>
<tr>
<td>exp_file</td>
<td>Histone ChIP-seq</td>
<td>NHLF</td>
<td>H. sapiens</td>
<td>ENCODE</td>
<td>H3K27ac</td>
<td>None</td>
<td>2</td>
<td>50</td>
<td>H3K27ac</td>
<td>Broad</td>
</tr>
<tr>
<td>exp_file</td>
<td>Histone ChIP-seq</td>
<td>Orsodiol</td>
<td>H. sapiens</td>
<td>ENCODE</td>
<td>H3K27ac</td>
<td>None</td>
<td>2</td>
<td>50</td>
<td>H3K27ac</td>
<td>Broad</td>
</tr>
<tr>
<td>exp_file</td>
<td>Histone ChIP-seq</td>
<td>IMF900</td>
<td>H. sapiens</td>
<td>Roadmap Epigenomics</td>
<td>H3K27ac</td>
<td>None</td>
<td>2</td>
<td>50</td>
<td>H3K27ac</td>
<td>Ren</td>
</tr>
<tr>
<td>exp_file</td>
<td>Histone ChIP-seq</td>
<td>H-1NSC</td>
<td>ENCODE_Joshua</td>
<td></td>
<td>H3K27ac</td>
<td>None</td>
<td>1</td>
<td>60</td>
<td>H3K27ac</td>
<td>Broad</td>
</tr>
</tbody>
</table>

**Remove Columns**
- data type
- cell or tissue
- species
- data source
- antibody
- treatment
- tissue order
- bin size
- target
- lab

**Add Columns**
- tdf file
- platform
- name
Next Steps

Release an open source alpha version

Development still very much IN PROGRESS

Integrate Galaxy API extensions into Galaxy code base

Any help from the Galaxy team would be appreciated

Implement a version on our cluster

Allowing collaborators an initial version to use