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.



Output Skyrocketing

Number sequenced



Baker et al. "Gene data to hit milestone." Nature. 2012 Jul 18

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

IIII Sequence Read Archive







"... 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...



isatab***

investigation

high level concept to link related studies

study

the central unit, containing information on the subject under study, its characteristics and any treatments applied.

a study has associated assays

assay

test performed either on material taken from the subject or on the whole initial subject, which produce qualitative or quantitative measurements (data)

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)

Workflow	/ Canvas bulk_zip			Options •
	Input dataset	*	Bulk Download Zipper 🟾 🛠	
	output	<u></u>	> Files to Zip 1 > file	
			→ Files to Zip 2 > file	
	Input Dataset	*	zipped_output (tar.gz) 🛛 🔅	
	output			





Data and resource **sharing is a central concept** in Refinery.

Source: *f*-*r*-*a*-*n*-*k* on *F*lickr

Sharing



Group Ownership

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

DEMO

A ► ↔ GL G → C + Pocket Refinery - Conte	Refinery Wiki Ref	inery Tracker Ster	n Cell Commons eRA Co	mmons Git Galaxy V	8102-06C5-11e1-a867 Veb Search + Delicious	-C8DCC8ed32d3/7a42U9UT-dbC5- Delicious VPN Galaxy AE	-11e1-984D-C8DCC8e C Futon Phraseban	k Lab Wiki TAI	Q* Google P Wiki TCGA Wiki	TCGA Data G	DAC Broad Dep	osit J6SE
Refinery Home	Search About	Contact Statist	ics								1 Nils Gehler	iborg Logout
Assay Files												
Search	Search	spp_bam_igvto	ools			\$ Apply					63	/1085
Tdf File			Data Type	Cell Or Tissue	Species	Data Source	Antibody 🕹	Treatment	Tissue Order	Bin Size	Target	Lab
Data Type	63	exp_file \$	Histone ChIP-seq	AH	D. melanogaster	modENCODE	H3K27ac	None	4	10	H3K27ac	NA
Cell Or Tissue		exp_fie ‡	Histone ChIP-seq	EL	D. melanogaster	modENCODE	H3K27ac	None	2	10	H3K27ac	NA
Species X H. sapiens	57	exp_file \$	Histone ChIP-seq	L3	D. melanogaster	modENCODE	H3K27ac	None	3	10	H3K27ac	NA
X D. melanogaster C. elegans	6 4	exp_file \$	Histone ChIP-seq	NH-A	H. sapiens	ENCODE	H3K27ac	None	2	50	H3K27ac	Broad
Data Source Platform		exp_file \$	Histone ChIP-seq	NHDF-Ad	H. sapiens	ENCODE	H3K27ac	None	2	50	H3K27ac	Broad
Antibody	41	exp_file \$	Histone ChIP-seq	NHEK	H. sapiens	ENCODE	H3K27ac	None	2	50	H3K27ac	Broad
X H3K27ac	22	exp_file \$	Histone ChIP-seq	NHLF	H. sapiens	ENCODE	H3K27ac	None	2	50	H3K27ac	Broad
Treatment Tissue Order		•	Histone ChIP-seq	Osteobl	H. sapiens	ENCODE	H3K27ac	None	2	50	H3K27ac	Broad
Bin Size			Histone ChIP-seq	IMR90	H. sapiens	Roadmap Epigenomics	H3K27ac	None	2	50	H3K27ac	Ren
laiget		-	1.5.	10 1500	11 mentane	ENCODE Jackus	10/07==	Mana			1101/071	

Remove Columns

X data type | X cell or tissue | X species | X data source | X antibody | X treatment | X tissue order | X bin size | X target | X lab

Add Columns

+ tdf file | + platform | + name

Copyright © 2012 Park Lab | Center for Biomedical Informatics at Harvard Medical School. All rights reserved.

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



















Park Lab @ HMS

- 🍐 Nils Gehlenborg
- 着 Richard Park
- 🎍 Psalm Haseley
- 🎍 Joe Luquette
- 🍐 Peter J Park

Hide Lab @ HSPH

Shannan Ho Sui
Ilya Sytchev
Winston Hide

Funded in part by

Agilent Technologies