REFINERY PLATFORM

A Foundation for Integrative Data Visualization Tools

Nils Gehlenborg¹, Richard Park¹, Ilya Sytchev², Psalm Haseley¹, Stefan Luger³, Anton Xue¹, Marc Streit³, Shannan Ho Sui², Winston Hide², Peter J Park¹

¹ Harvard Medical School ² Harvard School of Public Health ³ Johannes Kepler University Linz

Funding: NIH/NHGRI Pathway to Independence K99 HG007583, Agilent Technologies Emerging Insights Grant, Harvard Stem Cell Institute

http://www.refinery-platform.org

Challenges

Large datasets with dozens or hundreds of samples are now common in biology.

- Setting up and launching analyses with dozens or hundreds of samples across multiple data types is a complex task.
- 2. Manually keeping track of software, parameters, and data used in analyses is tedious and error prone.
- 3. Using visual exploration tools to study the results of such analyses is currently not well supported.

Major Components

(1) Repository(2) Workflows(3) Visualization

Refinery Platform

Refinery is a flexible analysis platform that is designed to accommodate diverse data and workflows for individuals and teams; our current implementation focuses on epigenomics and cancer genomics. One goal for this system is to serve as a platform for the development of novel **visual exploration tools**, that can directly access large and complex datasets and analysis results, and trigger new analyses on these data.

Provenance & Meta Data

ISA-Tab ("Investigation-Study-Assay") is an extremely flexible general purpose file format to describe biological experiments. The data model of Refinery is modeled after the ISA-Tab file format and provides extensive provenance information in an "experiment graph", which links all files to the inputs that they were derived from. Refinery provides a faceted-browsing interface and a flexible matrix view to enable users to quickly filter through datasets with thousands of samples annotated in ISA-Tab (or tabdelimited text files).

Experiment Graph



Provenance Visualization

→ C □ 192.168.50.50:8000/data_sets/0	d54f492c-61e2-11e3-8888-080027129698/
Refinery Home Statistics About	t Logout
ata Set Test 1: Request for Comments (RFC) Test	
Browse Analyze Visualize	Files Attributes Analyses Downloads Details Sharing
Current Selection 78 of	Display Provenance V
Save Reset	
Attribute Filter	
▶ Author	
▶ Month	
▶ Year	
• Туре	
▶ File Type	**************************************
Analysis Filter	
▼Analysis	202021 P++++++
Test workflow: 5 steps without	
branching 2014-03-04 @ 9	
12:42:57	
□ N/A 9	
Test workflow: SPP analog 2014-05-14 11:41:15.134276	
 Test workflow: SPP analog 	
2014-03-28 16:47:38.268071	
Test workflow: SPP analog	
2014-03-28 16:56:42.596327	
2014-04-02 18:02:27.773522 8	
Test workflow: SPP analog	+0+0+0 +0 +0+0+0+0
2014-04-02 15:35:40.958438	
Test workflow: 5 steps without	
Dranching 2014-03-04 @ 6 17:24:01	
Test workflow: 5 steps without	
branching 2014-03-04 @ 4	
15:12:36	
est workflow: 5 steps without	

Workflow Execution

Refinery Platform workflows are executed by **Galaxy**. The Galaxy workflow editor is used to create a "workflow" template" that is imported by the Refinery Platform, automatically instantiated based on the inputs selected by the user, and exported back into Galaxy through its API. Workflow results are downloaded into Refinery from Galaxy, added to the experiment graph, and made available for visualization and as input for further analyses.



Workflow Visualization



Analysis Setup User Interface

Refinery - Data Set 12373: X													
← → C [] 192.168.50.50:8000/data	a_sets/249e8bf8-7c25-11e3-b515	-080027	129698/#/conte	ent							☆ ♀ ≡		
Refinery Home Statistics	About										🛎 admin 🕩 Logout 🍵		
Data Set 12373: Genome-wide maps of chi	romatin state in pluripotent and linea	ge-commit	ed cells (Refinery	mod v2)								
Browse Analyze Visualize							Files	Attributes	Analyses	Downloads	Details Sharing		
MACS+SPP ChIP-Seq = groo NAME MACS+SPP ChIP-Seq = groomer + bowtie + spp (toolshed) + cut + igvtools: zv9 SUMMARY Peak-calling with both SPP and MACS. INPUTS (1-1 FILE MAPPING) input_file exp_file	Current Mapping 1 of 1 relationships			exp_file		Mode	Edit Overview	Mapping	Manual Automatic ?				
	Name http://bloodprogram.hsci.harvard.edu/sites/default/files/datafiles/3236.3.all.fastq.gz http://bloodprogram.hsci.harvard.edu/sites/default/files/datafiles/3992.5.all.fastq.gz												
	Shared Organism: Mus musculus, Sex: male, Cell Type: Embryonic stem cell (ESC), Strain: 129SvJae x M.castaneus F1, Analysis: N/A, Subanalysis: -1, Output Type: N/A, Type: Raw Data File, Organism Part: Embryo, Development Stage: Embryo, File Type: gz,												
	input_file Grab≡in table below to drag file on dropzone to update the assignment. Grab≡in table below to drag file on dropzone to update the assignment.												
Current Mapping New Rename Delete Launch Analysis	Current Selection	12 of 75 files selected ✓ Columns → 10 20 50 100 500 ✓ Columns → Columnian Columns → Columnian Columns → Columnian Columns → Columnia Columns → Columnia				Strain	Pisplay Table						
	 ▶ Organism ▶ Sex ▼ Cell Type ☑ Embryonic stem cell (ESC) □ Fibroblast □ Neural precursor cell □ Neural precursor cell □ 129SvJae x C57BL/6 28 ☑ 129SvJae x M.castaneus F1 12 Analysis Filter ▶ Analysis 	≡ 🗖	Mus musculus	male	Embryonic stem cell (ESC)	129SvJae x M.casta	ineus F1	Embryo	Embryo	http:/	/bloodprogram.hsci		
		≡ ø	Mus musculus	male	Embryonic stem cell (ESC)	129SvJae x M.castaneus F1		Embryo	Embryo	http:/	/bloodprogram.hsci		
		≡ ⊿	🗹 Mus musculus male Embryonic stem cell (ESC) 129SvJae x M		129SvJae x M.casta	M.castaneus F1 Embryo		Embryo	ıbryo http://bloodprogram.hs				
		≡ ₫	Mus musculus	male	Embryonic stem cell (ESC)	129SvJae x M.castaneus F1 Em		Embryo	Embryo	http:/	/bloodprogram.hsci		
		≡ 🗹 Mus musculus male Embryonic stem ce		Embryonic stem cell (ESC)	129SvJae x M.castaneus F1		Embryo	Embryo	http:/	/bloodprogram.hsci			
		≡ ⊿	Mus musculus male Embryonic stem cell (ESC		Embryonic stem cell (ESC)	129SvJae x M.castaneus F1		Embryo	Embryo		/bloodprogram.hsci		
		≡ ⊘	Mus musculus	male	Embryonic stem cell (ESC)	129SvJae x M.castaneus F1		Embryo	Embryo	http:/	/bloodprogram.hsci		
		Embryonic stem cell (ESC)		129SvJae x M.castaneus F1		Embryo Embryo		http:/	http://bloodprogram.hsci				
		≡ ∅	Mus musculus	male	Embryonic stem cell (ESC)	129SvJae x M.casta	ineus F1	Embryo	Embryo Embryo		/bloodprogram.hsci		
					Embryonic stem cell (ESC)) 129SvJae x M.castaneus F1		Embryo Embryo		http:/	http://bloodprogram.hsci		
		≡ ⊗	Mus musculus	male	Embryonic stem cell (ESC)	129SvJae x M.casta	ineus F1	Embryo	Embryo	http:/	/bloodprogram.hsci		
		≡ ∅	Mus musculus	male	Embryonic stem cell (ESC)	129SvJae x M.casta	ineus F1	Embryo	Embryo	http:/	/bloodprogram.hsci		

Foundation for Visual Exploration

Refinery Platform

Vertical Integration across processing stages

- drill down from highlevel results to raw data
- richer filtering capabilities, e.g., filter high-level results based on low-level analysis QC score

Status Quo



- no frameworks to develop visual exploration tools for large & heterogeneous data sets
- most tools are "data type-centric"
- often multiple tools are needed even for basic exploration
- user needs to track & enter provenance information by hand

Horizontal Integration across data types & analyses

- combine observations from multiple data types
- compare analysis approaches
- integrative visualizations

Integrative tools for visual exploration of large & heterogeneous biological data sets

IGV: http://www.broadinstitute.org/igv

IGV Integration

• • • • • Refinery - Data Set 13309: ×	000			IGV – Session: http	://refine	ry-dev.m	ed.harvard	.edu/file_sto	pre/23/2d/tm	pAll7PM.xml		
← → C 192.168.50.50:8000/data	D. melanogaster (d 🗘 🛛 All	;		G	• 👚	۹ ►	 Image: Constraint of the second second	X 🏳				
Refinery Home Statistics												
Data Set 13309: Genome-wide analyses of		5	i									
Browse Analyze Visualize		E source or tissue of function	type ies ody	2L	2R	2R	Het	3L	3LHet	3R 3R	Het X 4	U
	·	NAM geno data platfi facto facto	diata antib treat			1				1	11 1	
Current Selection 🔻	L3_H3K9me3_FE.wig_3.tdf						V		M		1	
Unable to detect species and genome build. Please select the	L3_H3K36me3_FE.wig_3.tdf			المريبة ويعرف لمادين	land)	a.l.	Ale ale	ر ماريك المراجع	Harbert	and the state	والمعا ومطلقه والمع	
correct genome and launch IGV.	L3_H3K27me3_FE.wig_3.tdf			at last as to so	ıl		h e sh i		dua	de ta an	, half an	
D. melanogaster (dm3)	L3_H3K4me3_FE.wig_3.tdf			a lather and			L.L. m		d du	Shark at	add and so held	
I Launch IGV	L3_H3K4me1_FE.wig_3.tdf			Label and the state	Haller	L da N	the second	a to alstance	ni - Libra d	alle at and to a life	A CALLER AND A CAL	
	HC_EL_2.00											
	iHMM.M1K16.fly_L3.bb											
	iHMM.M1K16.fty_EL.bb											
	LAD.Kc_2.bb											
	Gene			- han and a	(m) (m, a)	1				And the local sector		
	11 tracks loaded :1											
			-			_						