

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.

1. 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.

Refinery Platform



Major Components

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

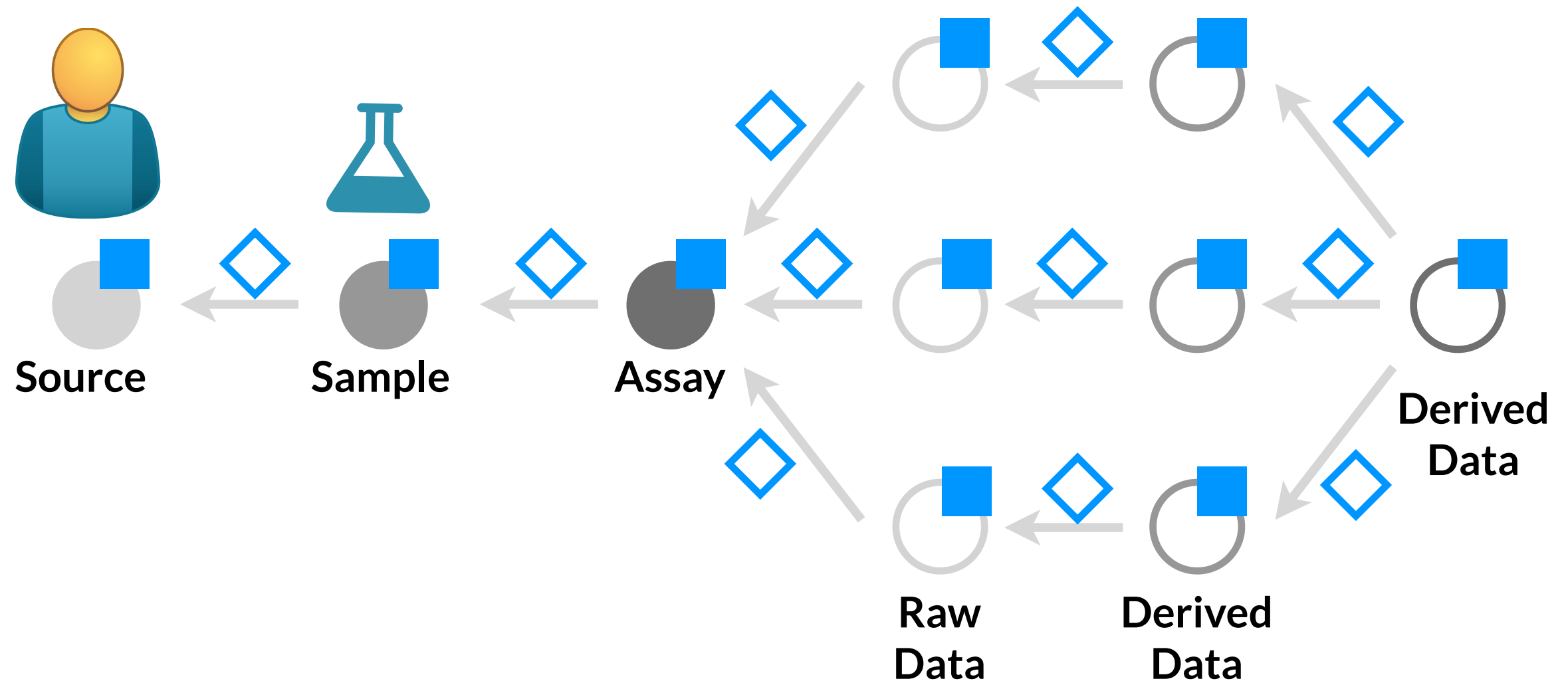
Refinery is a flexible analysis platform that is designed to accommodate diverse data and workflows for individuals and teams; our current implementation focuses on epigenomics and cancer genomics. One goal for this system is to serve as a platform for the development of novel **visual exploration tools**, 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 tab-delimited text files).

ISA-Tab: <http://www.isa-commons.org>

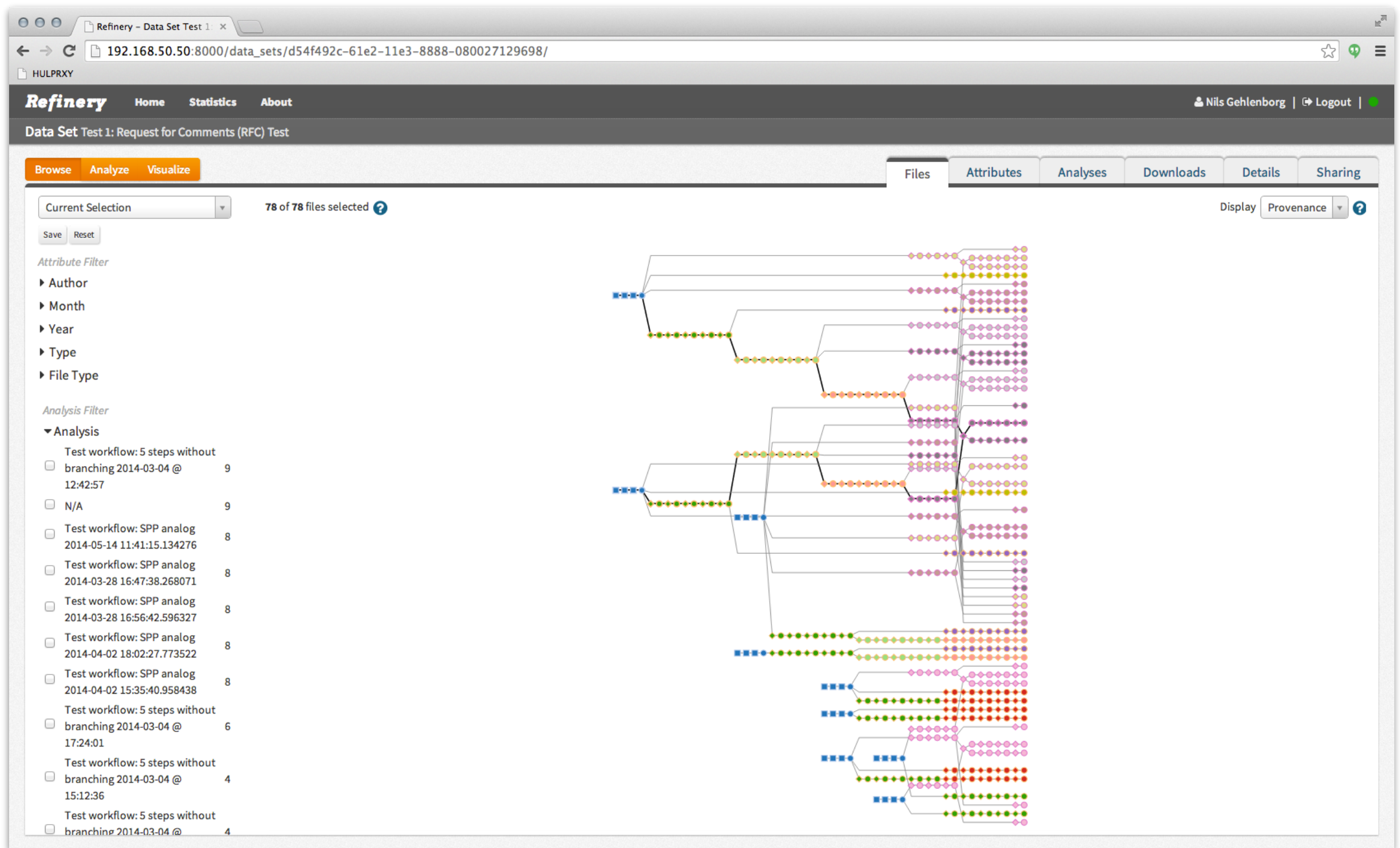
Experiment Graph



■ Meta Data
◇ Provenance Information

ISA-Tab

Provenance Visualization



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.

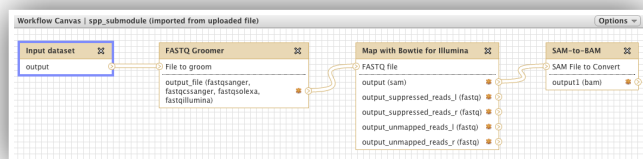
Workflow Templates

Galaxy

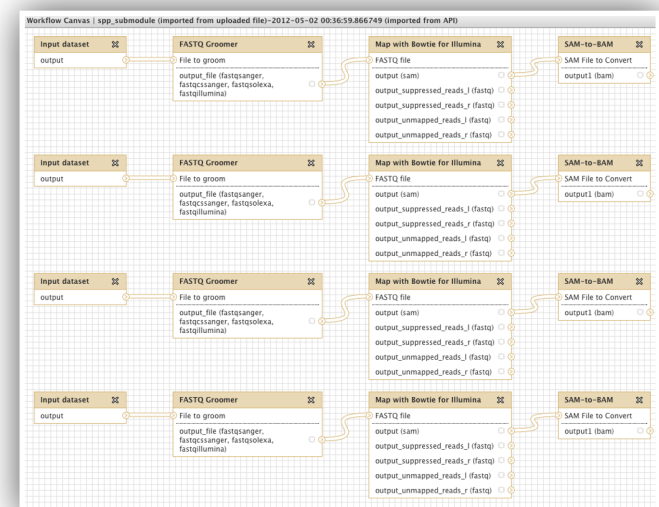
Workflow
Template

API

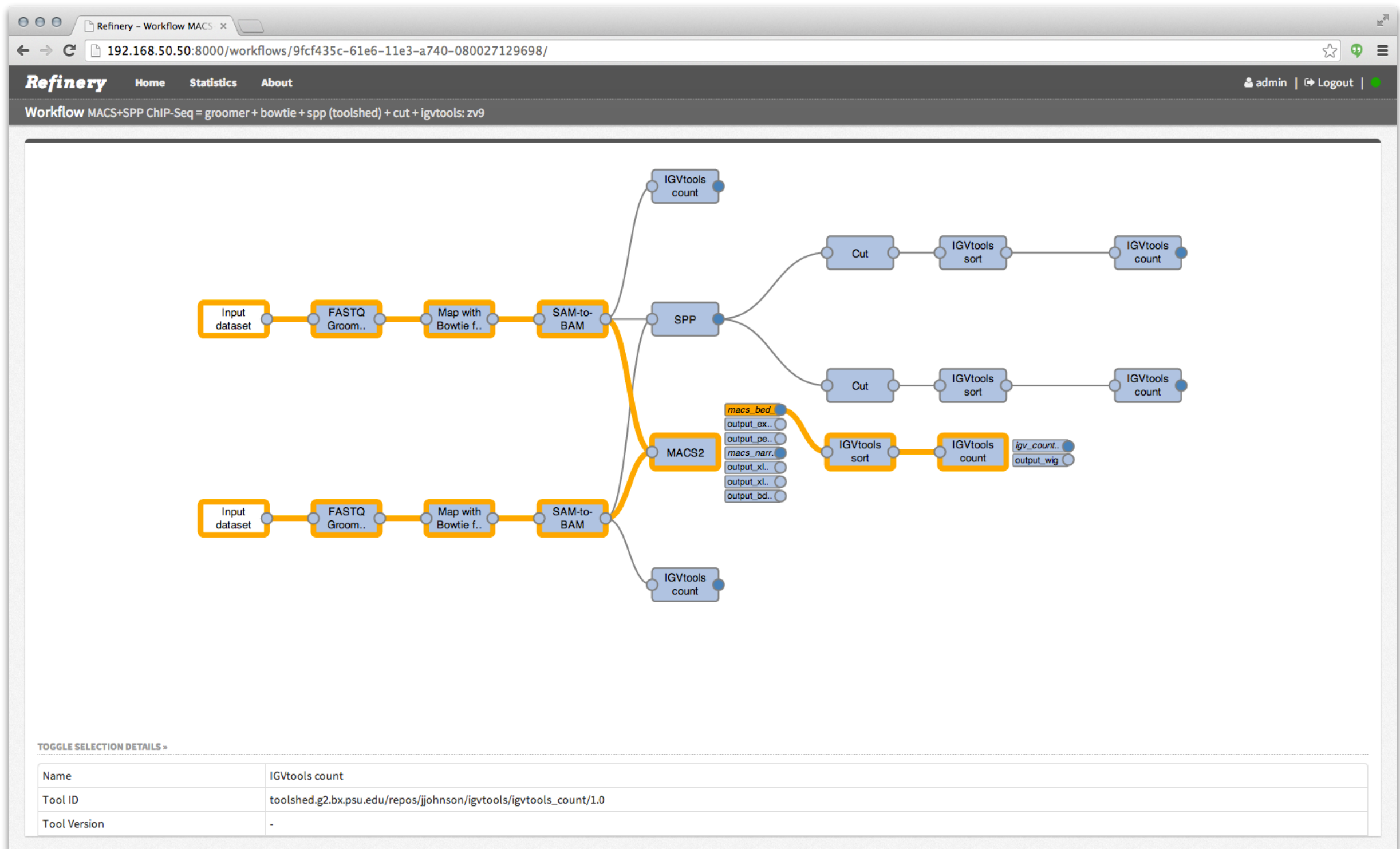
Workflow
Instance



Refinery



Workflow Visualization



Analysis Setup User Interface

Refinery - Data Set 12373

192.168.50.50:8000/data_sets/249e8bf8-7c25-11e3-b515-080027129698/#/content

Refinery

HomeStatisticsAbout

admin | Logout

Data Set 12373: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells (Refinery mod v2)

BrowseAnalyzeVisualize

FilesAttributesAnalysesDownloadsDetailsSharing

ModeEditOverviewMappingManualAutomatic

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
New ... Rename Delete
Launch Analysis

Current Mapping | 1 of 1 relationships

NewDelete PairDelete All Pairs

input_fileexp_file

Namehttp://bloodprogram.hsci.harvard.edu/sites/default/files/datafiles/3236.3.all.fastq.gzhttp://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_fileexp_file

Grab in table below to drag file on dropzone to update the assignment.

Current Selection

SaveResetDownload

Attribute Filter

Organism

Sex

Cell Type

Embryonic stem cell (ESC) 12

Fibroblast 0

Neural precursor cell 0

Strain

129SvJae x C57BL/6 28

129SvJae x M.castaneus F1 12

Analysis Filter

Analysis

12 of 75 files selected

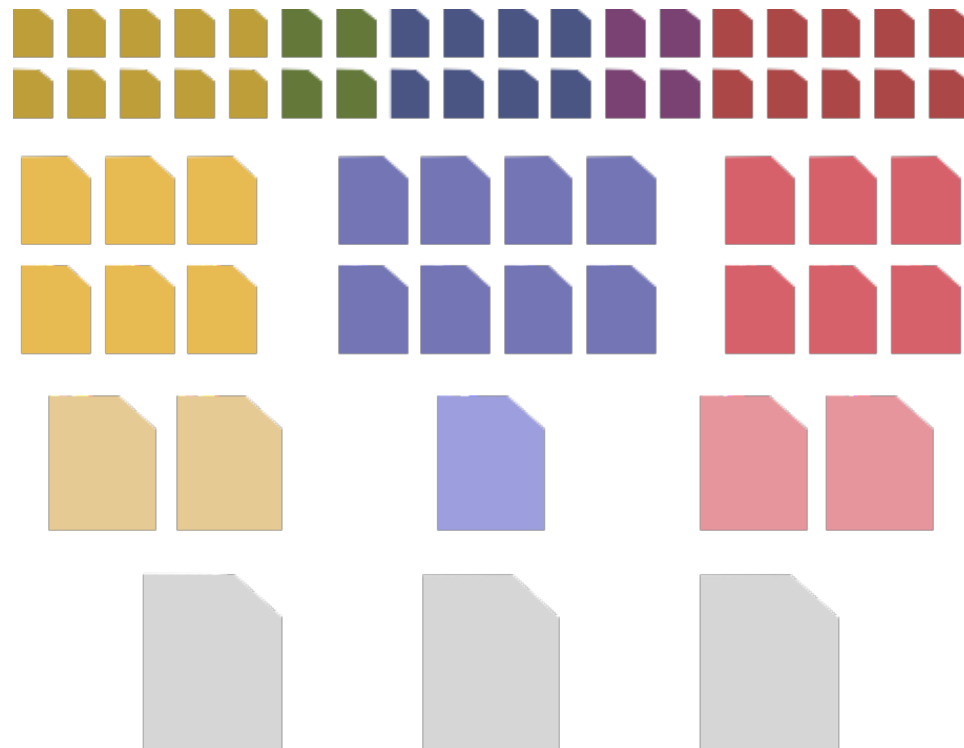
Columns102050100500

	Organism	Sex	Cell Type	Strain	Organism Part	Development Stage	Name
☒	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.castaneus F1	Embryo	Embryo	http://bloodprogram.hsci. ...
☒	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.castaneus F1	Embryo	Embryo	http://bloodprogram.hsci. ...
☒	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.castaneus F1	Embryo	Embryo	http://bloodprogram.hsci. ...
☒	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.castaneus F1	Embryo	Embryo	http://bloodprogram.hsci. ...
☒	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.castaneus F1	Embryo	Embryo	http://bloodprogram.hsci. ...
☒	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.castaneus F1	Embryo	Embryo	http://bloodprogram.hsci. ...

ViewDataAnnotationDisplayTable

Foundation for Visual Exploration

Refinery Platform



Vertical Integration *across processing stages*

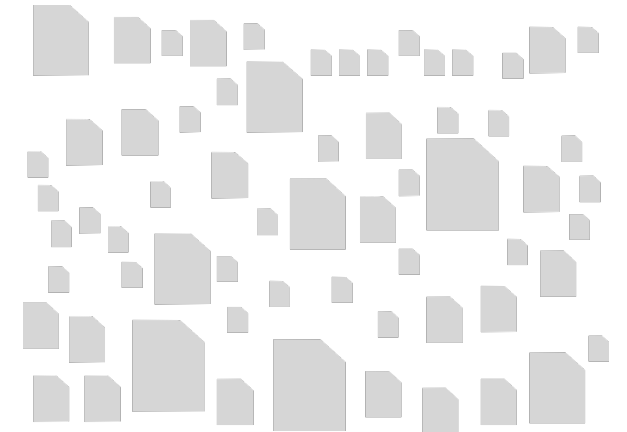
- drill down from high-level results to raw data
- richer filtering capabilities, e.g., filter high-level results based on low-level analysis QC score

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**

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

IGV Integration

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

