promiscuous omero

aka

omero as a general purpose framework for biomedical data management

PG-13

Our first goals (about 3 years ago)

- **to have scalable, uniform, computational access to large amounts of *-omic heterogeneous data**
  - From bio-samples to next gen sequencing data
- **to be able to track data dependencies**
  - model both objects and actions that connect them
- **to support computation on meta information and data dependency tree**
  - E.g., plan optimal titer-plate loading for next experiment
- **to support data access from multiple, geographically distributed, labs**
  - {Pula,Alghero,Lanusei,Monserrato}@sardinia, …
- **but first and foremost: no more excel sheet (!)**

omero.biobank

- **specialization of the "omero framework" to the handling of *omic data**
  - customized models and data structures for biomedical data handling:Genotyping data, clinical records, vessels, … (49 customized models)
  - network of objects connected by actions
  - can track transformations performed on the data
  - provides a rich API and tools for data input and queries
- **heavy use of omero tables**
  - snp markers, markers set, alignments, phenotype records
- **all client side code (~30k lines)**
  - mostly syntactic sugar
  - mostly boring stuff (importers/exporters/...)

omero.biobank: use

- **Data mainly from two large scale studies**
  - autoimmune disease (CNR-IRGB)
  - longevity (CNR-IRGB, NIH-NIA)
- **Currently handling:**
  - > 38000 individuals (~16.500 with parental relationships)
  - 26.800 clinical records
  - ~28.200 vessels, ~330 Titer Plates
  - 4 Genotyping technologies
    - Affymetrix GWH 6.0 (~935.000 markers, ~7.000 gtypes)
    - Illumina Immunochip (~196.000 markers, ~10.000 gtypes)
    - Illumina Hu OmniExpress (~730.000 markers, ~3.000 gtypes)
    - Illumina Hu Exome (~ 240.000 markers, ~5.000 gtypes)

omero.biobank: problems

- **Not particularly biologist-friendly**
  - Programmatic/script interface too complex for casual user
  - Tracking complex operations (action(s)) is rather cumbersome
- **Need to access multiple computing environments**
  - Batch system
  - Hadoop
    - largest cluster 3200 cores, uses an 'elastic' hadoop-grid-engine resource allocation scheme
  - Different filesystems
- **Users are in different locations:**
  - From the same island to different continents

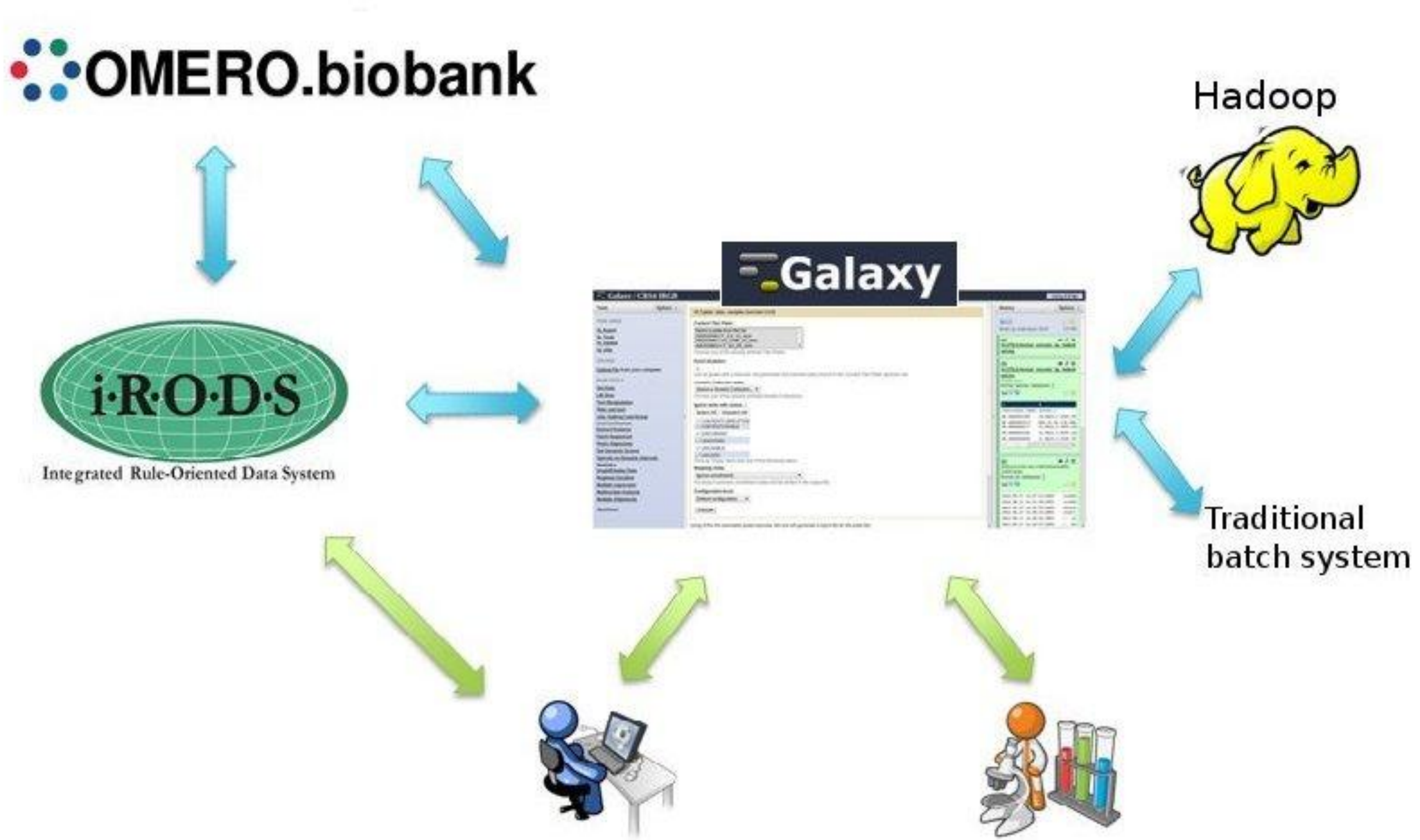
omero.biobank: omero specific problems

- **no omero integrated solution for dependency graph navigation**
  - We are currently using client side solution (pygraph) [slow]
  - Next: external graph handling service [fast, but dangerous]
- **slow on large data (tables) operations**
  - improved with ColumnArray<X>
  - more on this later
- **external file handling headaches**
  - DataObjects point to physical files not directly managed by omero

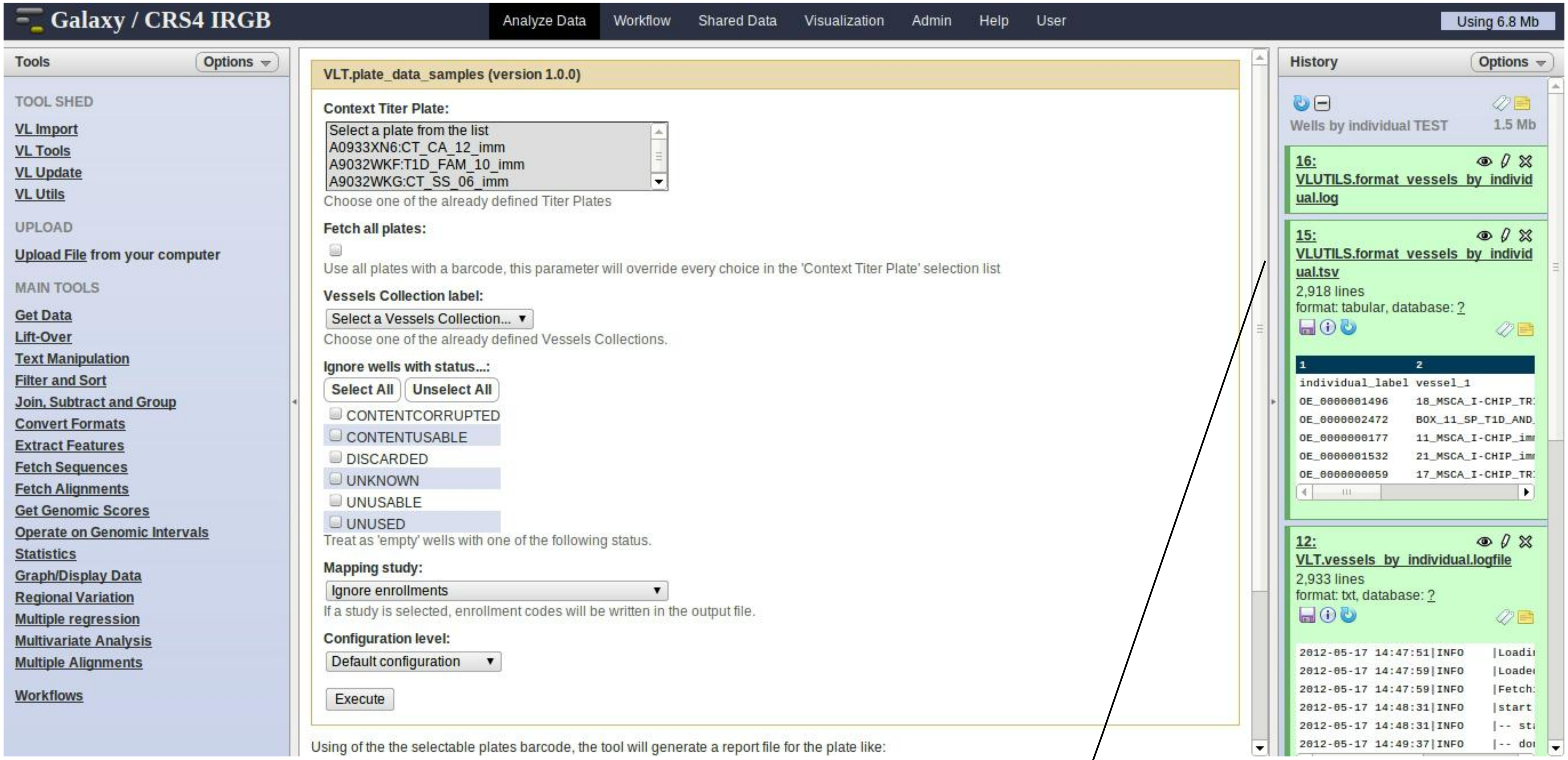
refined goals (18 months ago)

- to have a simple, biologist friendly, user interface
- to simplify standard data processing
  - facade to hadoop, batch job submission
- tools to build and share workflows
- maintain history of operations performed
  - share histories, save histories in omero,...
- decouple logical file view from file system details
  - meta-information based file system

omero.biobank + galaxy + iRODS



Galaxy (usegalaxy.org)
web interface for CLI tools



VLT.plate_data_samples (version 1.0.0)

Context Titer Plate:
Select a plate from the list
A0933XN6:CT_CA_12_imm
A9032WKF:T1D_FAM_10_imm
A9032WKG:CT_SS_06_imm
Choose one of the already defined Titer Plates

Fetch all plates:
□
Use all plates with a barcode, this parameter will override every choice in the 'Context Titer Plate' selection list

Vessels Collection label:
Select a Vessels Collection... ▾
Choose one of the already defined Vessels Collections.

Ignore wells with status...:
□ CONTENTCORRUPTED
□ CONTENTUSABLE
□ DISCARDED
□ UNKNOWN
□ UNUSABLE
□ UNUSED
Treat as 'empty' wells with one of the following status.

Mapping study:
Ignore enrollments ▾
If a study is selected, enrollment codes will be written in the output file.

Configuration level:
Default configuration ▾

Using of the the selectable plates barcode, the tool will generate a report file for the plate like:

Wells by individual TEST 1.5 Mb

16: VLUTILS.format_vessels_by_individual.log

15: VLUTILS.format_vessels_by_individual.tsv
2,918 lines
format: tabular, database: ?

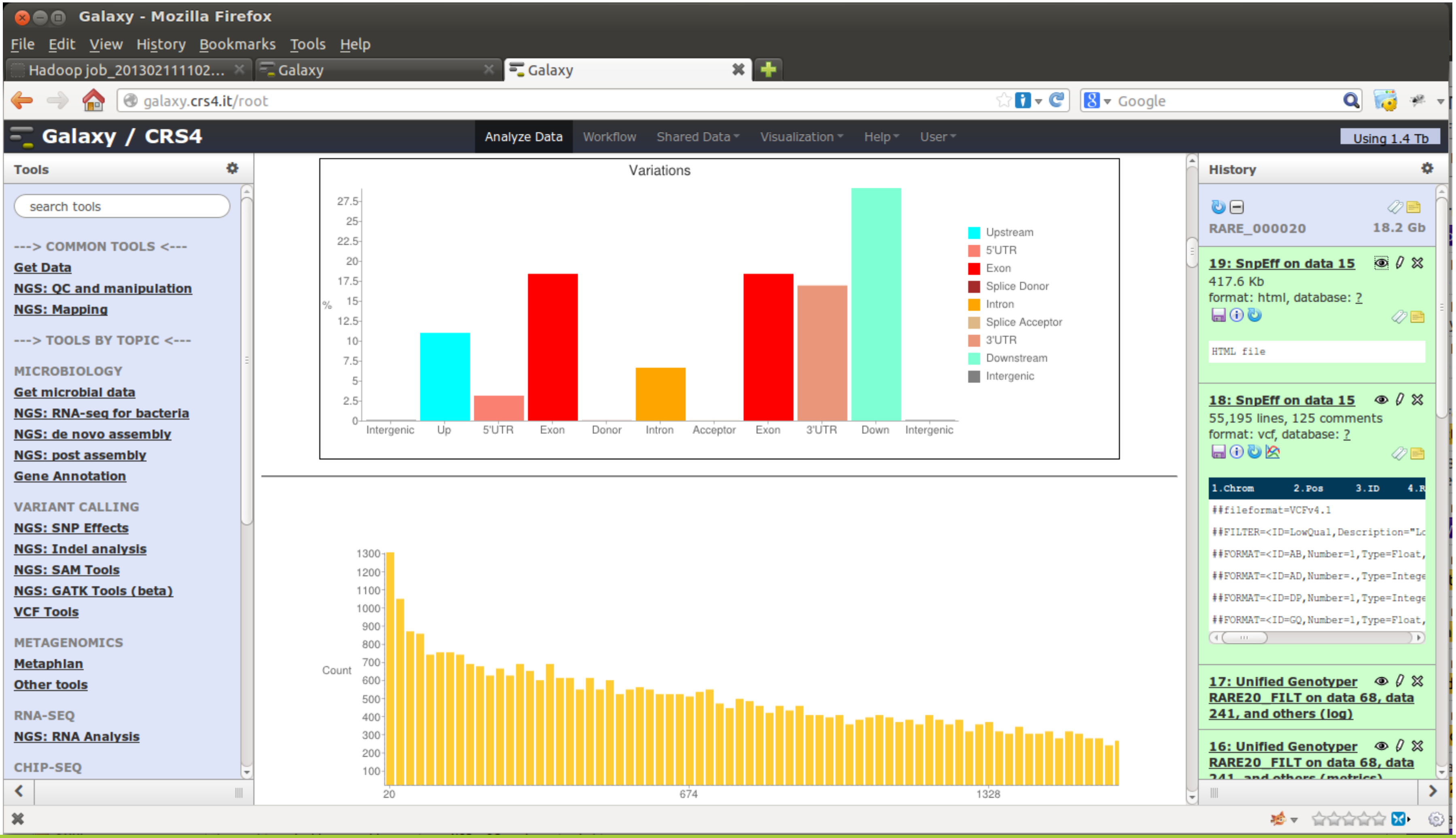
```
1                2
individual_label vessel_1
OE_0000001496    18_MSCA_I-CHIP_TR
OE_0000002472    BOX_11_SP_T1D_AND_
OE_0000000177    11_MSCA_I-CHIP_im
OE_0000001532    21_MSCA_I-CHIP_im
OE_0000000059    17_MSCA_I-CHIP_TR
```

12: VLT.vessels_by_individual.logfile
2,933 lines
format: txt, database: ?

```
2012-05-17 14:47:51|INFO     |Loadi
2012-05-17 14:47:59|INFO     |Loade
2012-05-17 14:47:59|INFO     |Fetch
2012-05-17 14:48:31|INFO     |start
2012-05-17 14:48:31|INFO     |-- st
2012-05-17 14:49:37|INFO     |-- do
```

History of operations performed

Galaxy: quasi-lab-book



RARE_000020 18.2 Gb

19: SnpEff on data 15
417.6 Kb
format: html, database: ?

HTML file

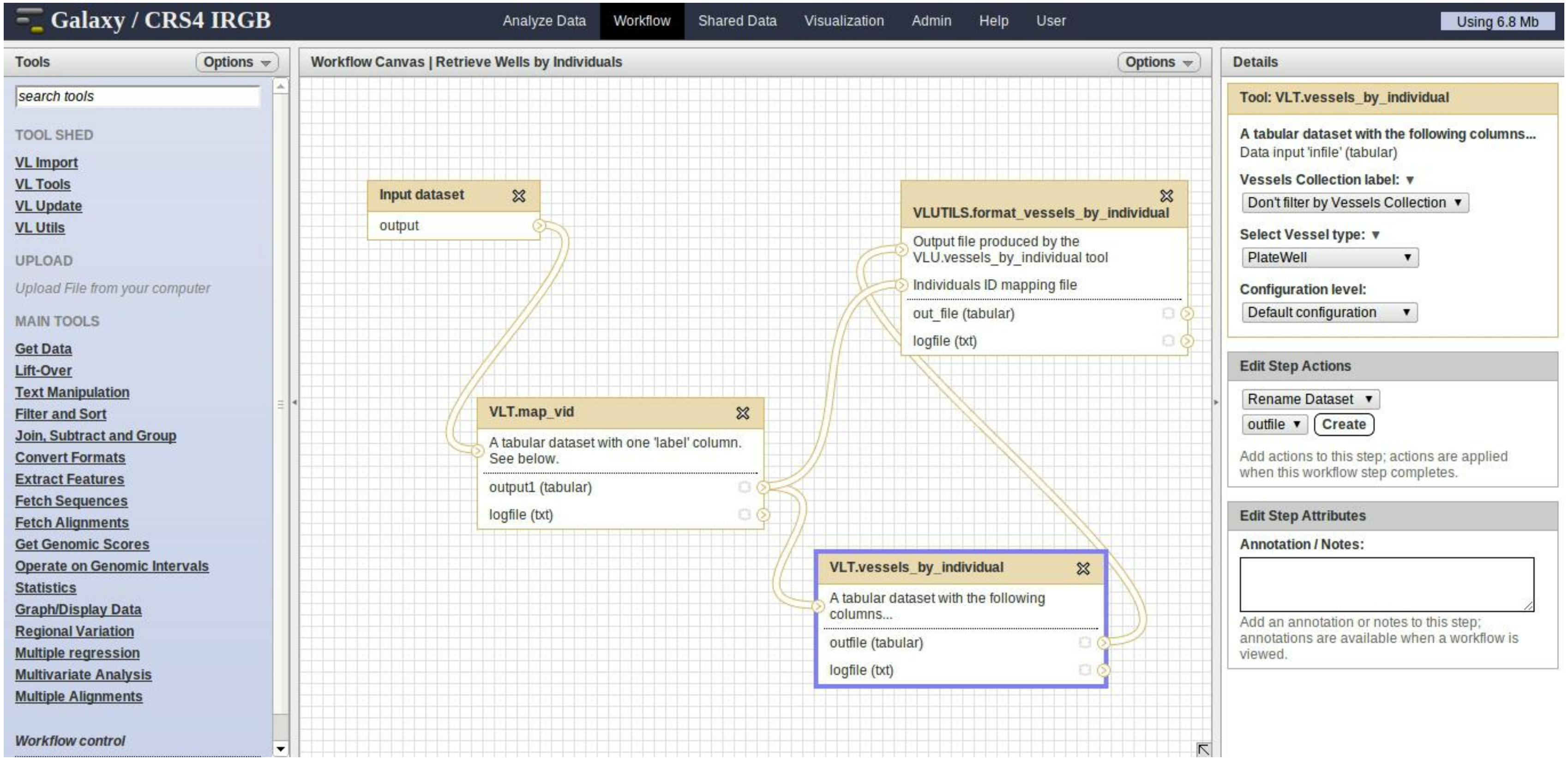
18: SnpEff on data 15
55,195 lines, 125 comments
format: vcf, database: ?

```
1.Chrom    2.Pos    3.ID    4.R
##fileformat=VCFv4.1
##FILTER=<ID=LowQual,Description="Lc
##FORMAT=<ID=AB,Number=1,Type=Float,
##FORMAT=<ID=AD,Number=.,Type=Intege
##FORMAT=<ID=DP,Number=1,Type=Intege
##FORMAT=<ID=GQ,Number=1,Type=Float,
```

17: Unified Genotyper RARE20_FILT on data 68, data 241, and others (log)

16: Unified Genotyper RARE20_FILT on data 68, data 241, and others (metrics)

Galaxy: workflow editor



Workflow Canvas | Retrieve Wells by Individuals

VLT.map_vid: A tabular dataset with one 'label' column. See below.

VLUTILS.format_vessels_by_individual: Output file produced by the VLU.vessels_by_individual tool; Individuals ID mapping file

VLT.vessels_by_individual: A tabular dataset with the following columns...

Tool: VLT.vessels_by_individual

A tabular dataset with the following columns...
Data input 'infile' (tabular)

Vessels Collection label: Don't filter by Vessels Collection

Select Vessel type: PlateWell

Configuration level: Default configuration

Edit Step Actions

Rename Dataset ▾ outfile ▾ Create

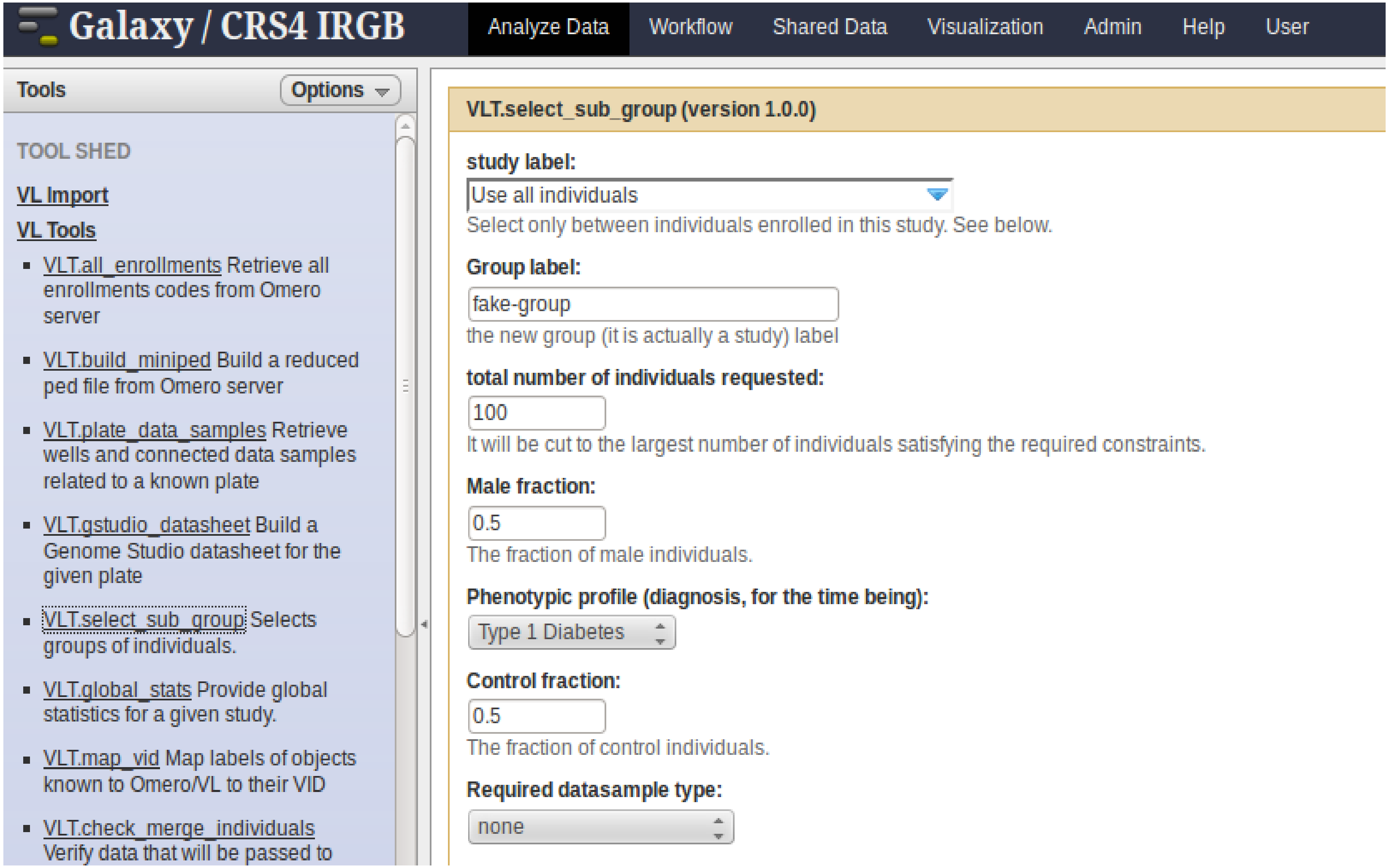
Add actions to this step; actions are applied when this workflow step completes.

Edit Step Attributes

Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

Interaction with omero.biobank



TOOL SHED

VL Import

VL Tools

- VLT.all_enrollments Retrieve all enrollments codes from Omero server
- VLT.build_miniped Build a reduced ped file from Omero server
- VLT.plate_data_samples Retrieve wells and connected data samples related to a known plate
- VLT.gstudio_datasheet Build a Genome Studio datasheet for the given plate
- VLT.select_sub_group Selects groups of individuals.
- VLT.global_stats Provide global statistics for a given study.
- VLT.map_vid Map labels of objects known to Omero/VL to their VID
- VLT.check_merge_individuals Verify data that will be passed to

VLT.select_sub_group (version 1.0.0)

study label:

Use all individuals

Select only between individuals enrolled in this study. See below.

Group label:

fake-group

the new group (it is actually a study) label

total number of individuals requested:

100

It will be cut to the largest number of individuals satisfying the required constraints.

Male fraction:

0.5

The fraction of male individuals.

Phenotypic profile (diagnosis, for the time being):

Type 1 Diabetes

Control fraction:

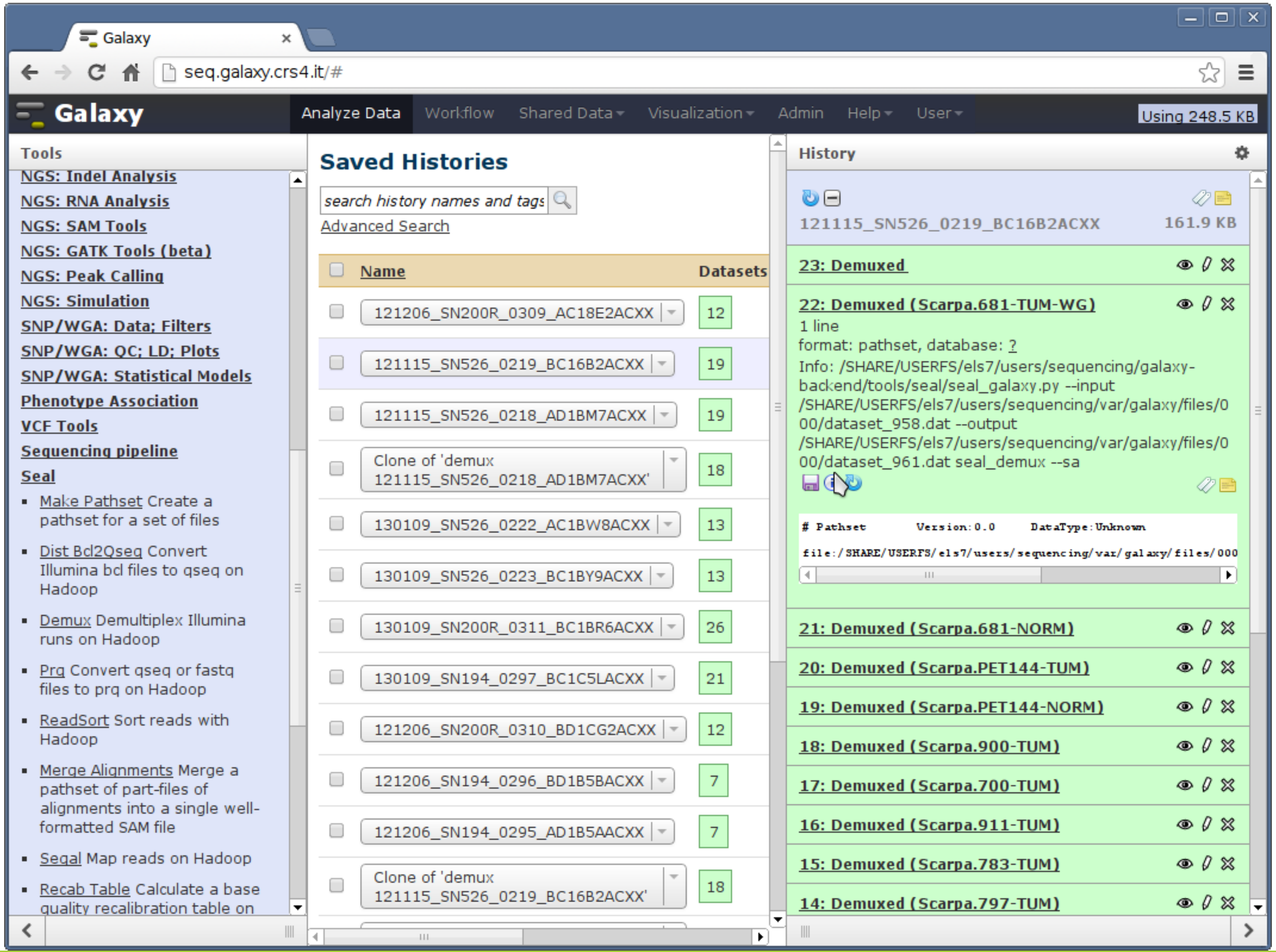
0.5

The fraction of control individuals.

Required datasample type:

none

Façade to hadoop tools



Tools

- NGS: Indel Analysis
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: GATK Tools (beta)
- NGS: Peak Calling
- NGS: Simulation
- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots
- SNP/WGA: Statistical Models
- Phenotype Association
- VCF Tools
- Sequencing pipeline
- Seal
  - Make Pathset Create a pathset for a set of files
  - Dist Bcl2Qseq Convert Illumina bcl files to qseq on Hadoop
  - Demux Demultiplex Illumina runs on Hadoop
  - Prq Convert qseq or fastq files to prq on Hadoop
  - ReadSort Sort reads with Hadoop
  - Merge Alignments Merge a pathset of part-files of alignments into a single well-formatted SAM file
  - Seqal Map reads on Hadoop
  - Recab Table Calculate a base quality recalibration table on

Saved Histories

search history names and tags

Advanced Search

□	Name	Datasets
□	121206_SN200R_0309_AC18E2ACXX	12
□	121115_SN526_0219_BC16B2ACXX	19
□	121115_SN526_0218_AD1BM7ACXX	19
□	Clone of 'demux 121115_SN526_0218_AD1BM7ACXX'	18
□	130109_SN526_0222_AC1BW8ACXX	13
□	130109_SN526_0223_BC1BY9ACXX	13
□	130109_SN200R_0311_BC1BR6ACXX	26
□	130109_SN194_0297_BC1C5LACXX	21
□	121206_SN200R_0310_BD1CG2ACXX	12
□	121206_SN194_0296_BD1B5BACXX	7
□	121206_SN194_0295_AD1B5AACXX	7
□	Clone of 'demux 121115_SN526_0219_BC16B2ACXX'	18

History

121115_SN526_0219_BC16B2ACXX 161.9 KB

23: Demuxed

22: Demuxed (Scarpa.681-TUM-WG)

1 line
format: pathset, database: ?
Info: /SHARE/USERFS/els7/users/sequencing/galaxy-backend/tools/seal/seal_galaxy.py --input /SHARE/USERFS/els7/users/sequencing/var/galaxy/files/000/dataset_958.dat --output /SHARE/USERFS/els7/users/sequencing/var/galaxy/files/000/dataset_961.dat seal_demux --sa

```
# Pathset       Version:0.0     DataType:Unknown
file:/SHARE/USERFS/els7/users/sequencing/var/galaxy/files/000
```

21: Demuxed (Scarpa.681-NORM)

20: Demuxed (Scarpa.PET144-TUM)

19: Demuxed (Scarpa.PET144-NORM)

18: Demuxed (Scarpa.900-TUM)

17: Demuxed (Scarpa.700-TUM)

16: Demuxed (Scarpa.911-TUM)

15: Demuxed (Scarpa.783-TUM)

14: Demuxed (Scarpa.797-TUM)

iRODS as a Decoupling System

- **IRODS is an integrated Rule-Oriented Data-management System**
  - uses unique logical names that are separate from the names as stored physically, providing a global ‘logical name-space’
  - Rules to automatically treat data on insertion and retrieval
  - Ability to tag data sets (e.g., sample id, data format)
  - Web based and command line interfaces
  - transfers data across the network in an integrated manner (parallel threads for large files)
- **We use IRODS as a front end to our heterogeneous storage system**
  - about 4.5PB in various boxes

iRODS is developed by DICE UNC (http://www.irods.org)

Short-term vs long term memory

- **Typical workflows**
  - have several steps and may fail
  - unwise to commit intermediate data to repository
- **Solution:**
  - Short-term memory → Galaxy history
    - Tracks steps while the computation is running
    - Permits to iteratively build a “good protocol”
  - Long term memory → OMERO.biobank
    - Record history in OMERO.biobank

galaxy + omero + iRods: glue

- **extensions to galaxy**
  - support communication with omero.biobank
  - improved galaxy histories API to support omero consumption
  - Almost all relevant tools galaxy wrapped
    - omero.biobank import/export/query tools
    - hadoop based tools for NGS and genotyping
    - ….
  - we are extending galaxy objectstore to directly support iRODS objects (files and collections)
- **iRODS**
  - external reference data is moving to iRODS
  - omero.biobank is moving to irods:// file paths
  - iRODS rules to simplify registration of huge dataset and galaxy integration

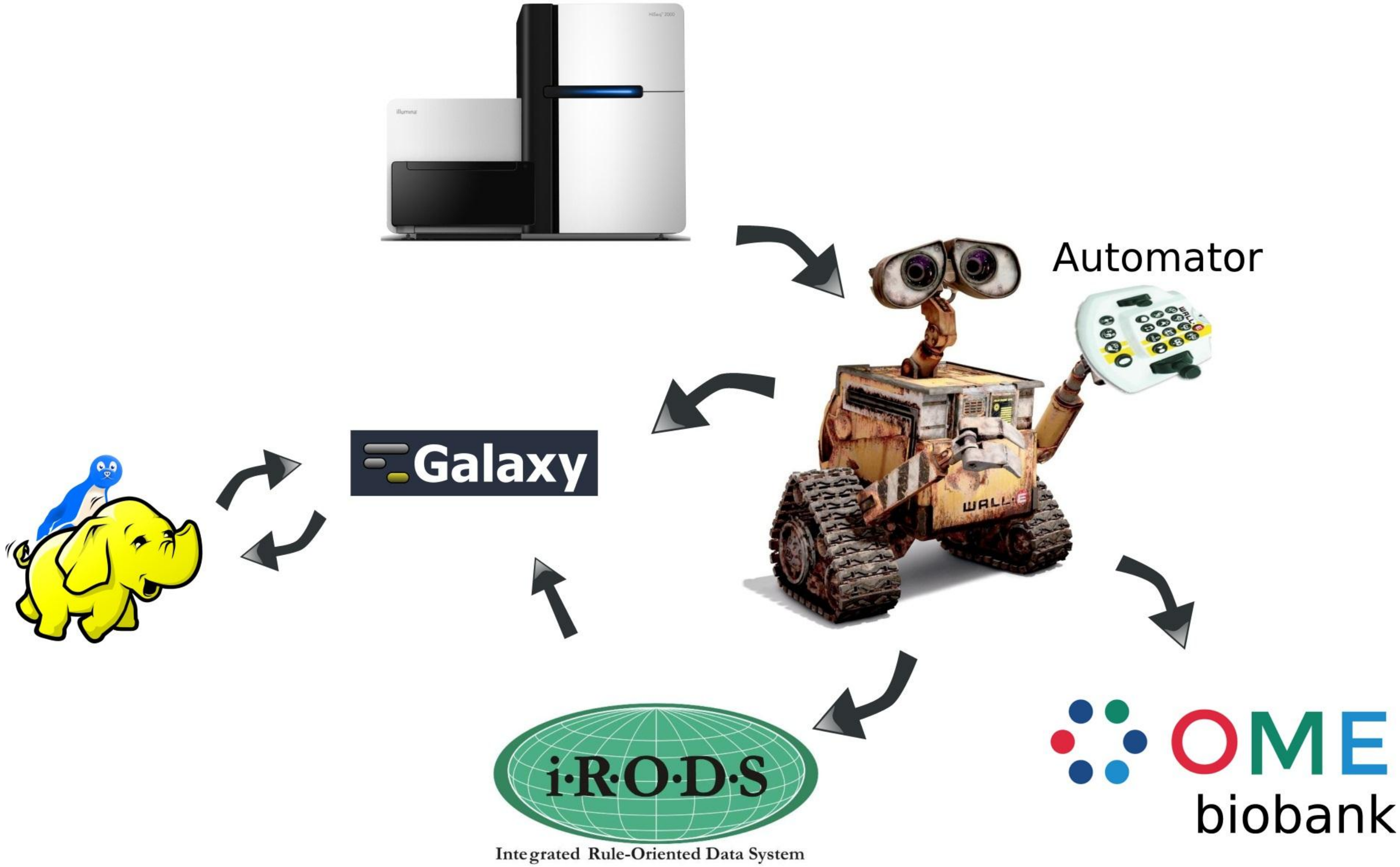
galaxy + omero + iRods

- **User community: biologist/bioinformaticians**
  - About 50 external, 10 internal users
  - All omero.biobank import, most export and queries
- **Problems:**
  - «designed» to have a human in command
    - Manage complex workflows chains, handle failures
  - Boring, dangerous and expensive for large scale production runs

new goals (5 months ago)

- **support the running of the CRS4 next generation sequencing service (3 Hiseq-2000)**
  - From biological sample in the mail to digital data in the cloud
  - automatize anything that would be cost-effective to automatize

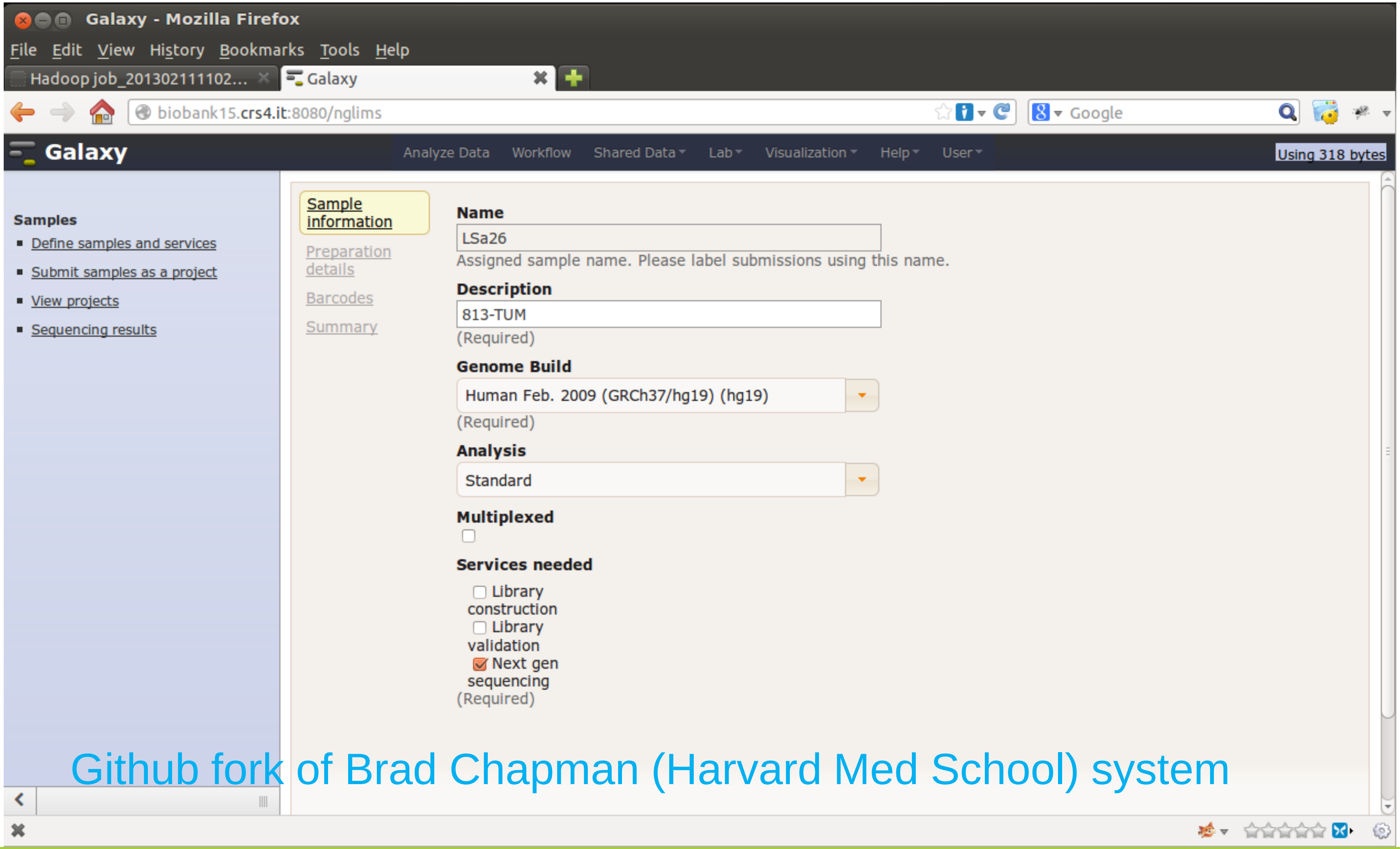
Yet anOther full Data cycle Automator



Automation

- **Galaxy front-end for biosample submission and analysis request**
- **All data operations described as galaxy workflows**
- **Automation layer that chains together workflows and integrates the various system components:**
  - Illumina sequencers
  - Galaxy (-> Hadoop cluster)
  - omero.biobank
  - iRODS
- **Basic pipelines up and running**
  - Flowcell to per-sample fastq datafiles in production

Sample submission front-end



Samples

- Define samples and services
- Submit samples as a project
- View projects
- Sequencing results

Sample information
Preparation details
Barcodes
Summary

Name
LSa26
Assigned sample name. Please label submissions using this name.

Description
813-TUM
(Required)

Genome Build
Human Feb. 2009 (GRCh37/hg19) (hg19)
(Required)

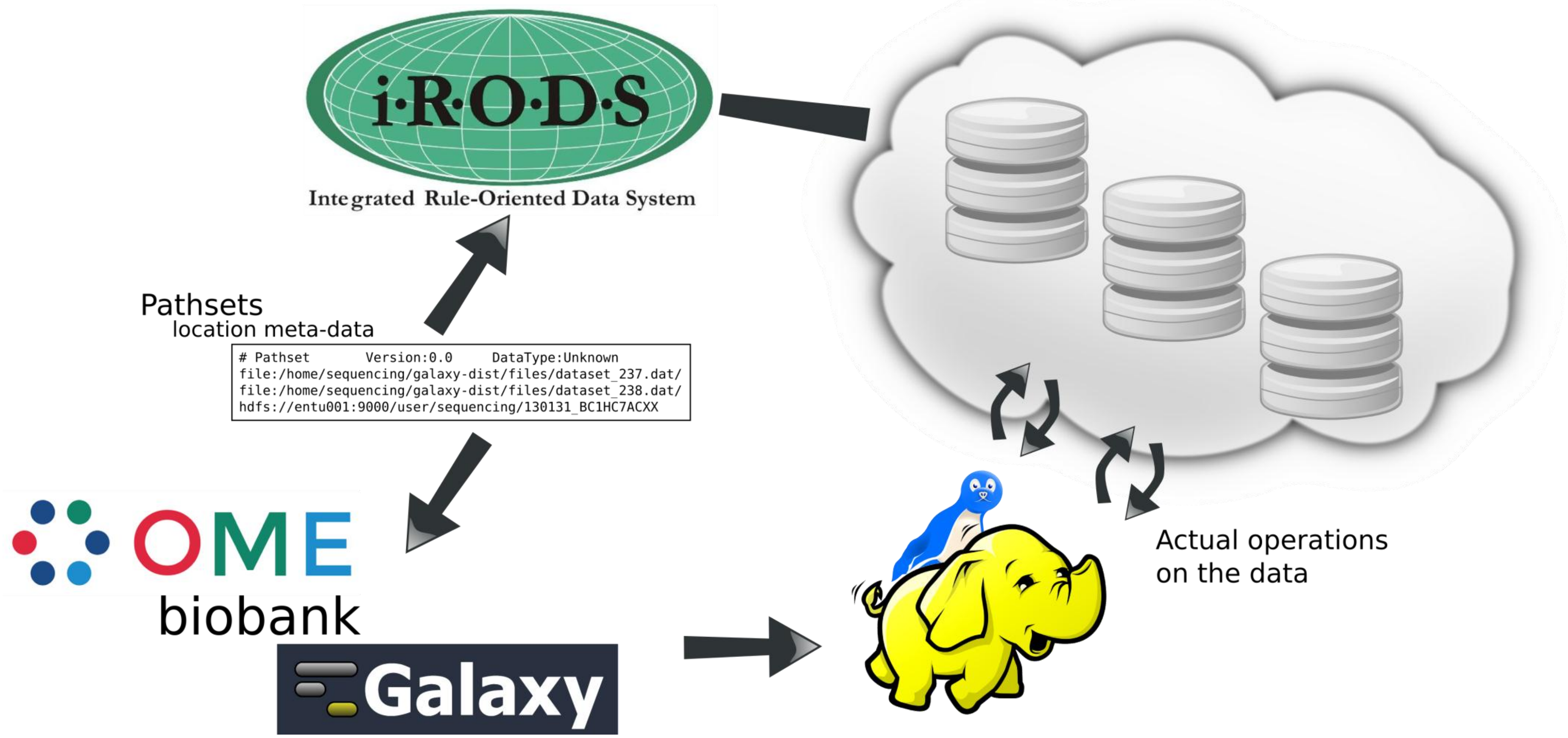
Analysis
Standard

Multiplexed
□

Services needed
□ Library construction
□ Library validation
☑ Next gen sequencing
(Required)

Github fork of Brad Chapman (Harvard Med School) system

Big data workflow

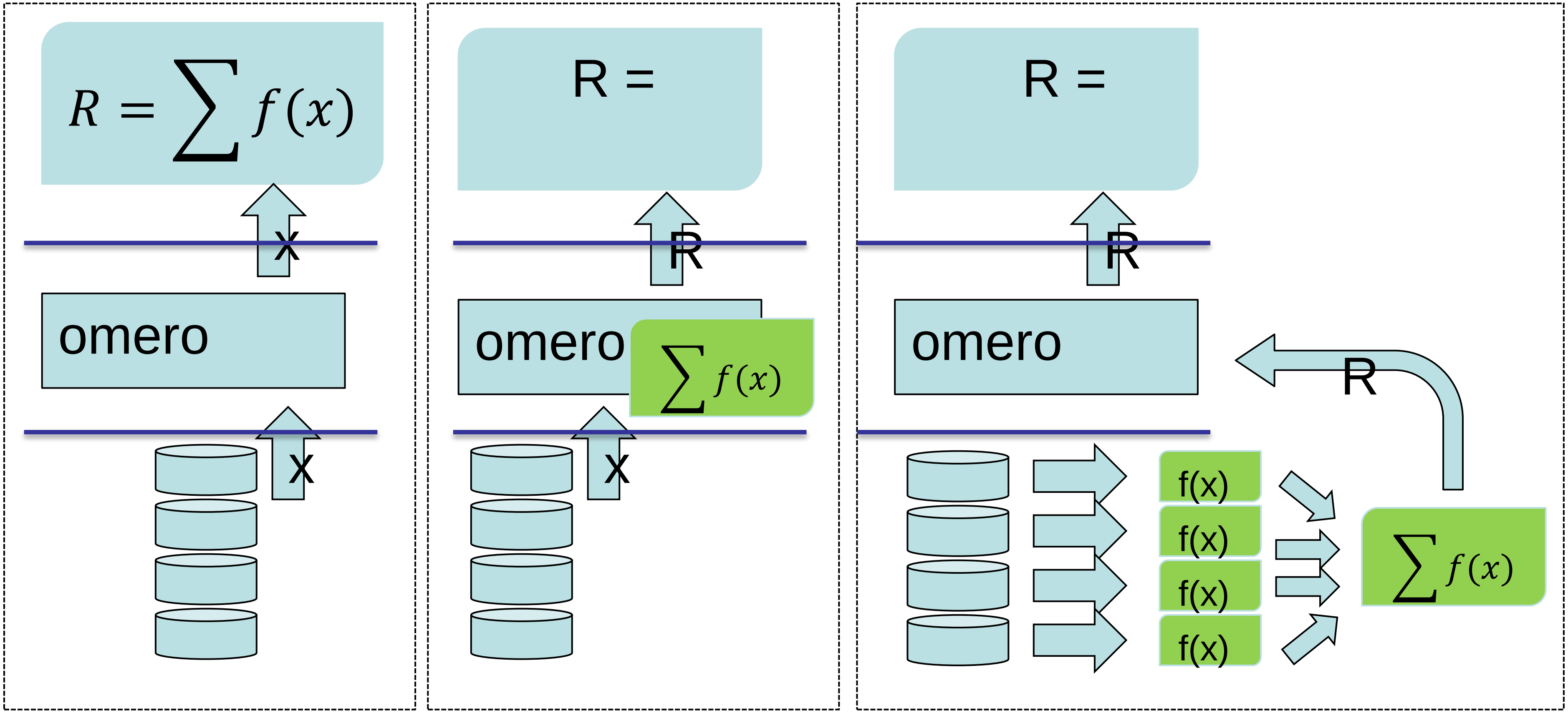


```
# Pathset        Version:0.0      DataType:Unknown
file:/home/sequencing/galaxy-dist/files/dataset_237.dat/
file:/home/sequencing/galaxy-dist/files/dataset_238.dat/
hdfs://entu001:9000/user/sequencing/130131_BC1HC7ACXX
```

to summarize: our mantra

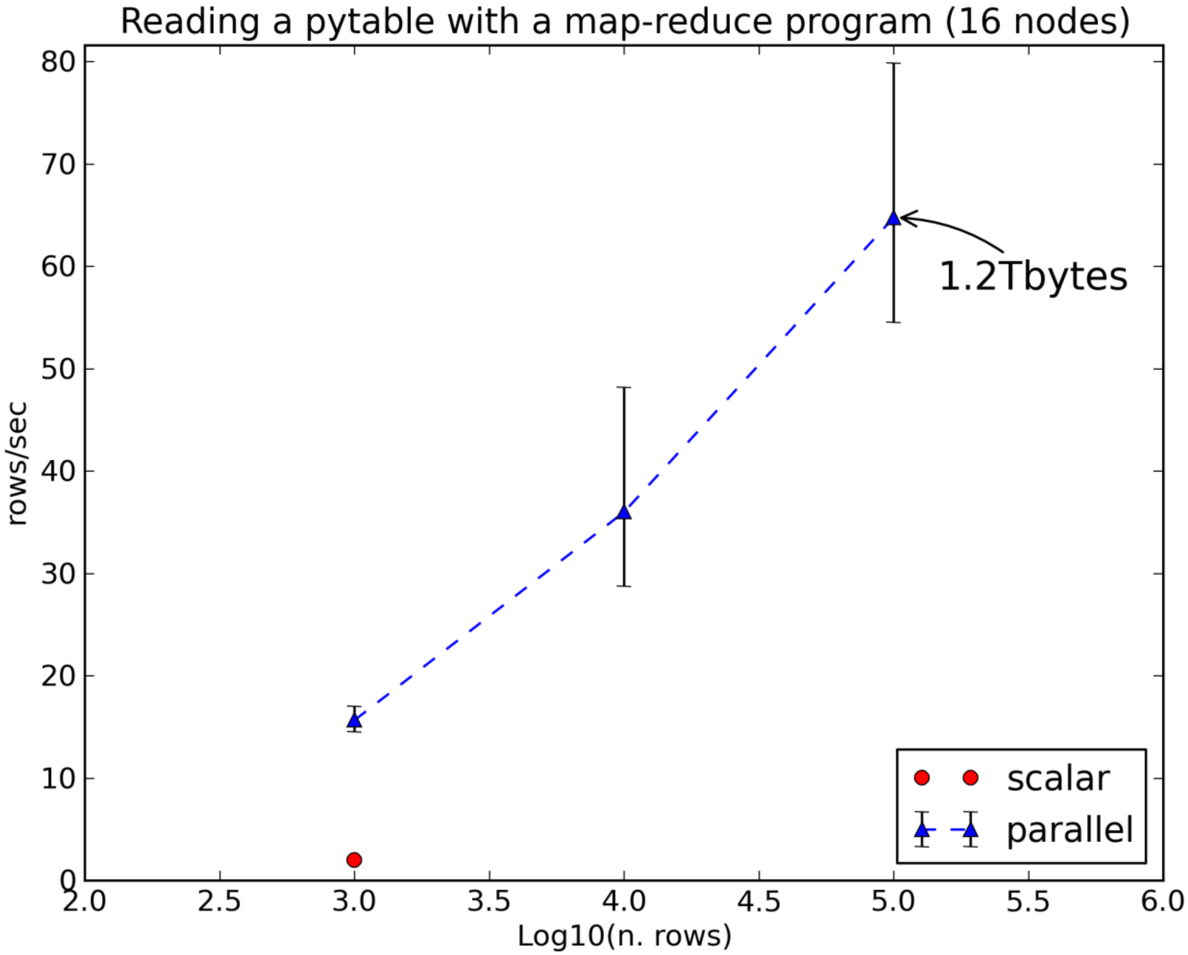
- omero.biobank knows what things are
- iRods knows where things are
- galaxy knows how to operate on them

Back to one of our slowness problems



Client side (current) Server side (classic) Server side (map-reduce)

Processing rates



Structured objects file system

- **Possible to instruct/delegate computing framework on how computational load should be distributed**
- **HDF5 natural candidate to impose «scientific data» structure on file system**
  - Implementation details
    - using H5FD_SPLIT it is possible to separate data from metadata in two different files
    - In principle possible to have HDF5 on top of HDFS, QFS better?
    - We wrote a minor pytables extension to support H5FD_SPLIT, so we can easily try on HDFS (and later on QFS)
- **BTW- For this class of objects, e.g., big SNP arrays, HBASE is not a good solution.**

new goals: back to images!

- **We are moving toward "pathology" applications support**
  - Integration of sequencing + proteomics + digital pathology

THANK YOU FOR YOUR TIME!