Scalable Data Management and Computable Framework for Large Scale Longitudinal Studies

Chris Allan[1,2], Gianmauro Cuccuru[3], Simone Leo[3], Luca Lianas[3], Josh Moore[2],
Maristella Pitzalis[4], Serena Sanna[4], Ilenia Zara[3], Jason Swedlow[1,2], Gianluigi Zanetti[3]

[1]Wellcome Trust Centre for Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dundee, Scotland, UK. [2]Glencoe Software Inc., Seattle, Washington, USA. [3]CRS4, Pula, Italy. [4]Istituto Ricerca Genetica e Biomedica – CNR, Monserrato, Italy.

GCC 2012, July 25-27, Chicago

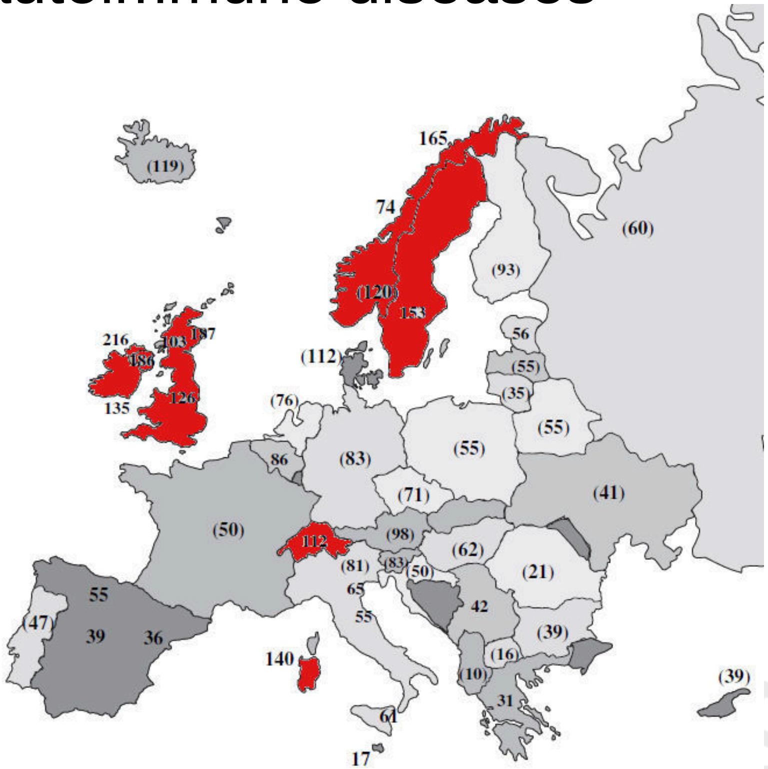
CRS4

- Center for Research, Development, and Advanced Studies in Sardinia
- Interdisciplinary research center focused on computational sciences
- Located in the POLARIS Science and Technology Park (Pula, Sardinia, Italy)
- Operational since 1992
- RTD staff of ~180 people

CRS4
POLARIS Edificio 1
C.P. 25
09010 Pula (CA), ITALY
www.crs4.it

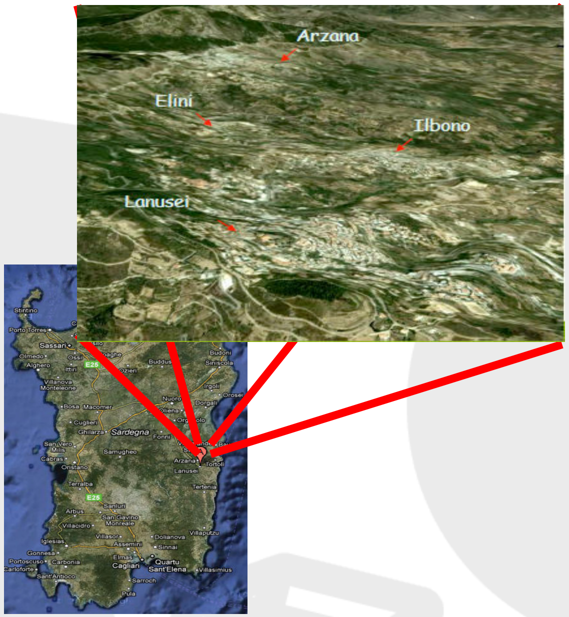
Joint Effort CRS4-CNR

CRS4 cooperates with CNR-IRGB on study on autoimmune diseases



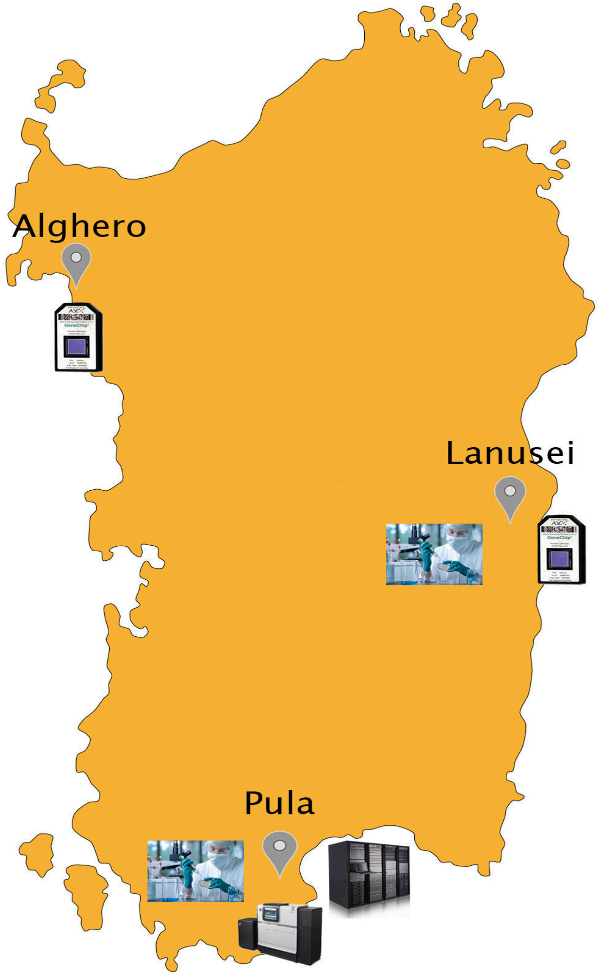
Autoimmune diseases such as type I diabetes and MS have in Sardinia one of the highest incidences worldwide

CRS4 cooperates with CNR-IRGB and NIH-NIA on large scale study on longevity



http://*sardinia*.nia.nih.gov/

Scale of the Problem



Separate labs that need access to the same data and computational resources

- Geographically distributed biosamples
- Data generated and used by multiple sites
  - Need to provide a global namespace with fast network data transfer
- Multiple genomic technologies
- Multiple clinical data sources
  - Need to document actions applied to the data
- Users with varying computing skills need different tools
  - Traditional, queue-based
  - Hadoop-based
  - Database queries

Scale of the Problem

- ~16,500 volunteers
- ~28,200 biological samples

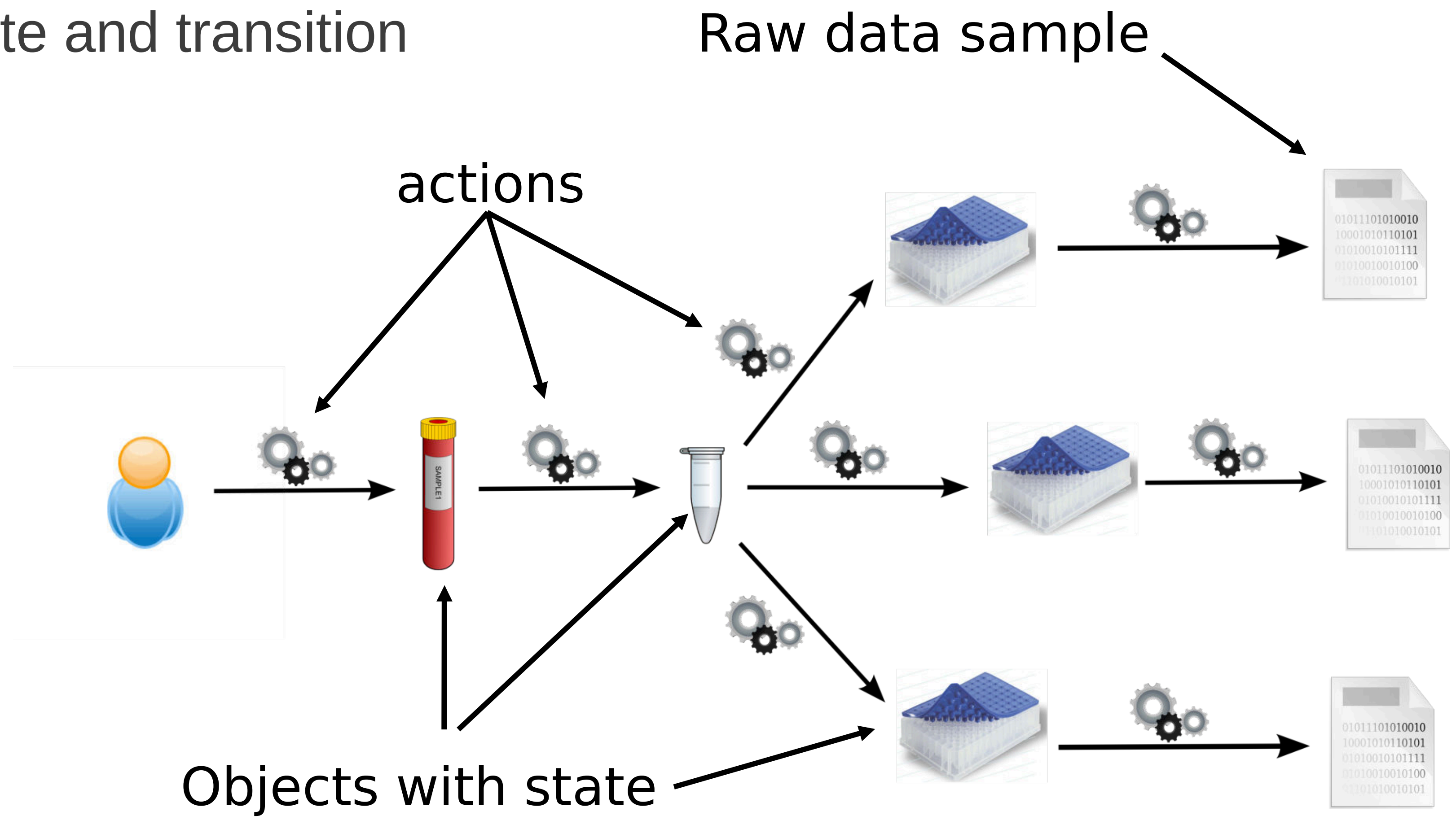
- Genotyping
  - 2 different vendors: Affymetrix and Illumina
  - 4 different chips:
  - Illumina Immunochip, 10,000 genotypes
  - Illumina OmniExpress, 3,000 genotypes
  - Illumina Exome chip, 5,000 genotypes
  - Affymetrix Genome-Wide Human SNP Array 6.0, 7,000 genotypes

- Sequencing
  - 3 Illumina HiSeq2000 (the largest sequencing center in Italy)
  - capable to produce more than 10 TB/month
  - Sequenced ~2,800 samples: 85% whole-genome resequencing, rest RNA and Exome

Details of the Problem

We have to model:

- Different data type
- Objects state and transition



Raw data sample

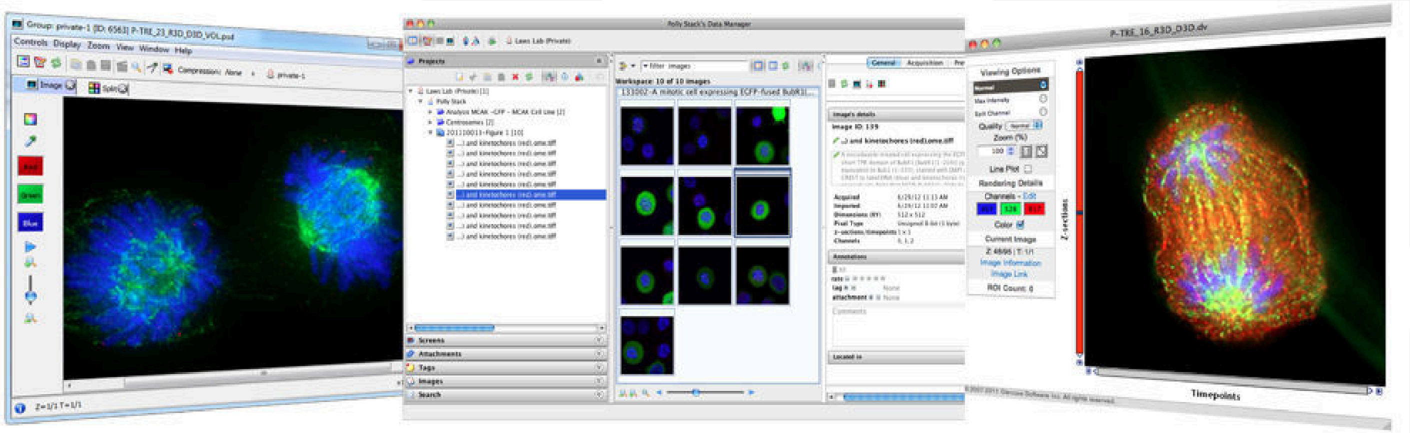
actions

Objects with state

Omero as our Knowledge Base

OMERO is a flexible, client-server, model-driven data management platform for experimental biology

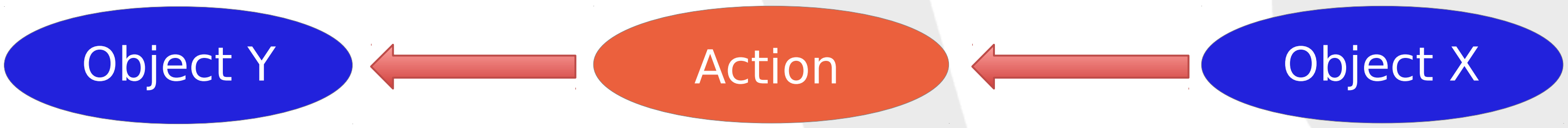
- includes several storage mechanisms
- middleware to facilitate access to stored data through API
- client applications for biological image data management
- Developed with bioimages in mind but not limited to those, by the Open Microscopy Environment Consortium (University of Dundee, Glencoe Software, Harvard Medical School, LOCI)



http://openmicroscopy.org

Omero.biobank

- OMERO.biobank is our specialization to support large-scale high-resolution genome-wide association studies
  - Extends OMERO with customized models and data structures for biomedical data handling
    - Genotyping data, clinical records, vessels, ...
  - network of objects connected by actions
  - can track transformations performed on the data
  - provides a rich API and a software suite with tools for data input and queries



If object X is the result of an event that occurred on object Y, the action has a reference to it

An action is a generic event that produces data that is recorded into the system

Every object knows the action that generated it

Specialized computational tools

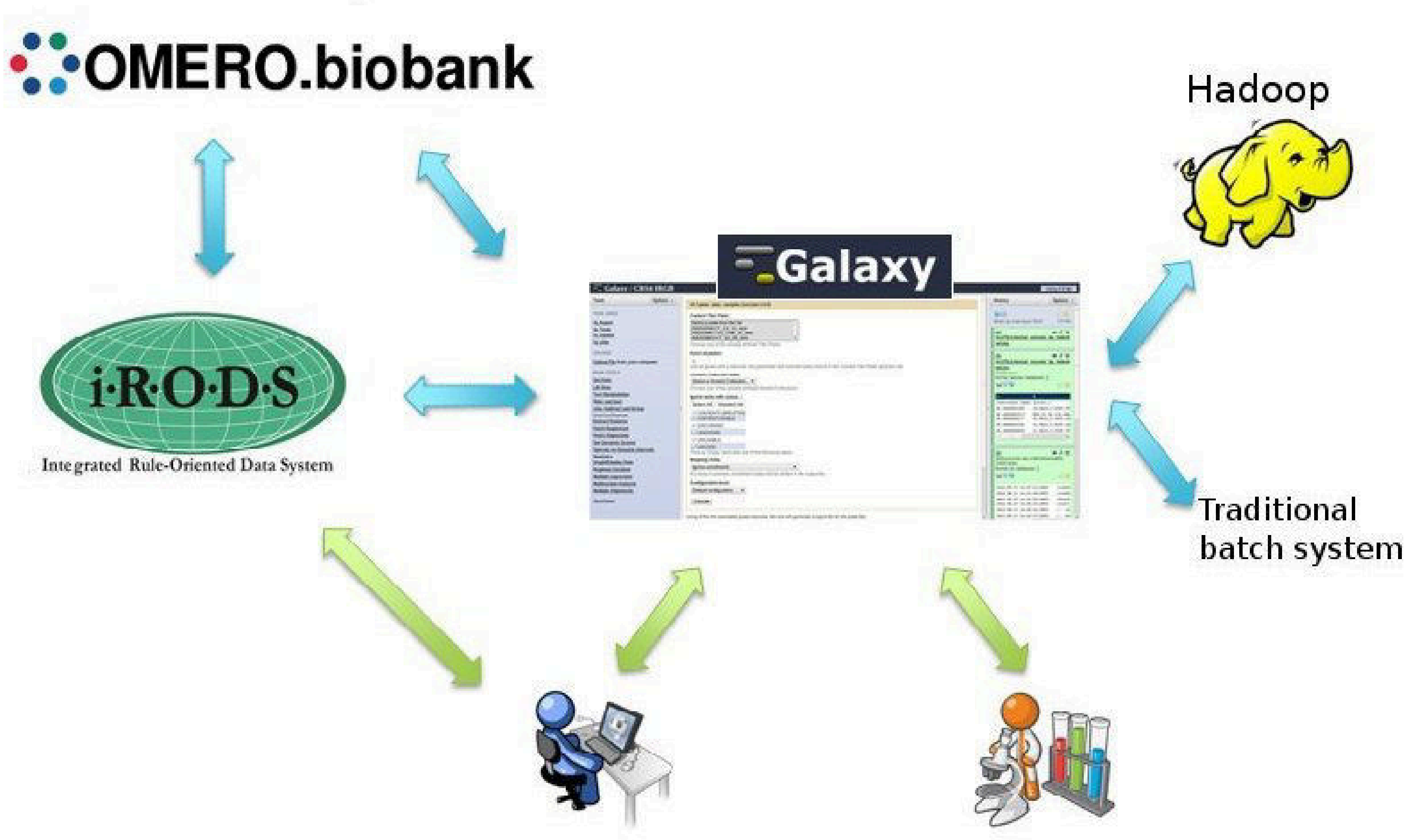
- We developed specialized Hadoop tools to compute on large dataset, for instance:

  - Seal
    - is a suite of distributed applications for aligning short DNA reads, and manipulating and analyzing short read alignments.
    - made to scale well in the amount of computing nodes available and the amount of the data to process
    - *http://biodoop-seal.sourceforge.net/*

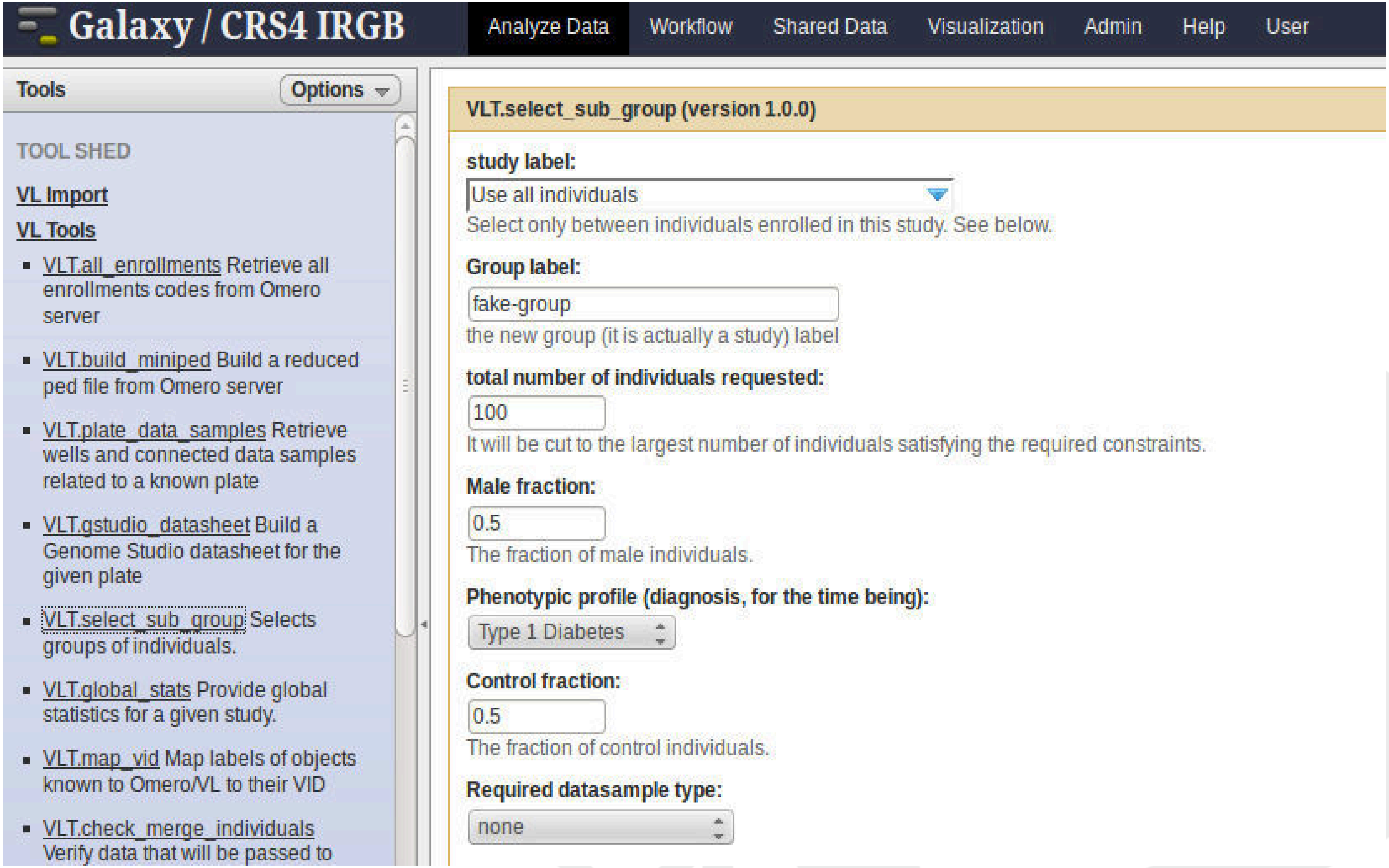
...still not enough

- Programmatic/script interface too complex for casual user
- Users are in different locations:
  - From the same island to different continents
- Depth of operation tracking
- Need to access multiple computing environments
  - Batch system
  - Hadoop
  - largest cluster 3200 cores, uses an 'elastic' hadoop-grid-engine resource allocation scheme
  - Different filesystems

Our Galaxy "ecosystem"



Simplify 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

Short-term vs Long-term memory

- Typical workflows have several steps and may fail
- Don’t want to commit intermediate data to repository

- Solution:
  - Short-term memory → Galaxy history
    - Tracks steps while the computation is running
    - Permits to iteratively build the “perfect protocol”
  - Long term memory → OMERO.biobank
    - Record history in OMERO.biobank

- Working in progress
    - We currently have only API support, no integrated Galaxy UI

iRODS as a Decoupling System

- IRODS is an integrated Rule-Oriented Data-management System (http://www.irods.org, developed by DICE UNC)
  - transfers data across the network in an integrated manner (parallel threads for large files)
  - 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

- We use IRODS as a front end to our heterogeneous storage system (about 4.5PB in various boxes)

iRODS and Galaxy

- Manual import
  - (I)put data on iRODS
  - Retrieve them through the Galaxy upload form

- Automatic import
  - Registering data in a specific iRODS collection will trigger a Galaxy library upload
  - e.g., sequencing run completion triggers an iRODS rule that upload data into a Galaxy library
  - In principle, it could also run galaxy workflows on the data

Conclusions

Example of a specialized Galaxy "ecosystem"

- OMERO.Biobank as knowledge base
- iRODS as decoupling systems
- Hadoop as computational workhorse

Future work:

- automated, data-tracking, scalable HTSeq pipeline
  - based on omero.biobank, irods, galaxy API and UI
- histories export to Omero.biobank as a integrated tool in Galaxy

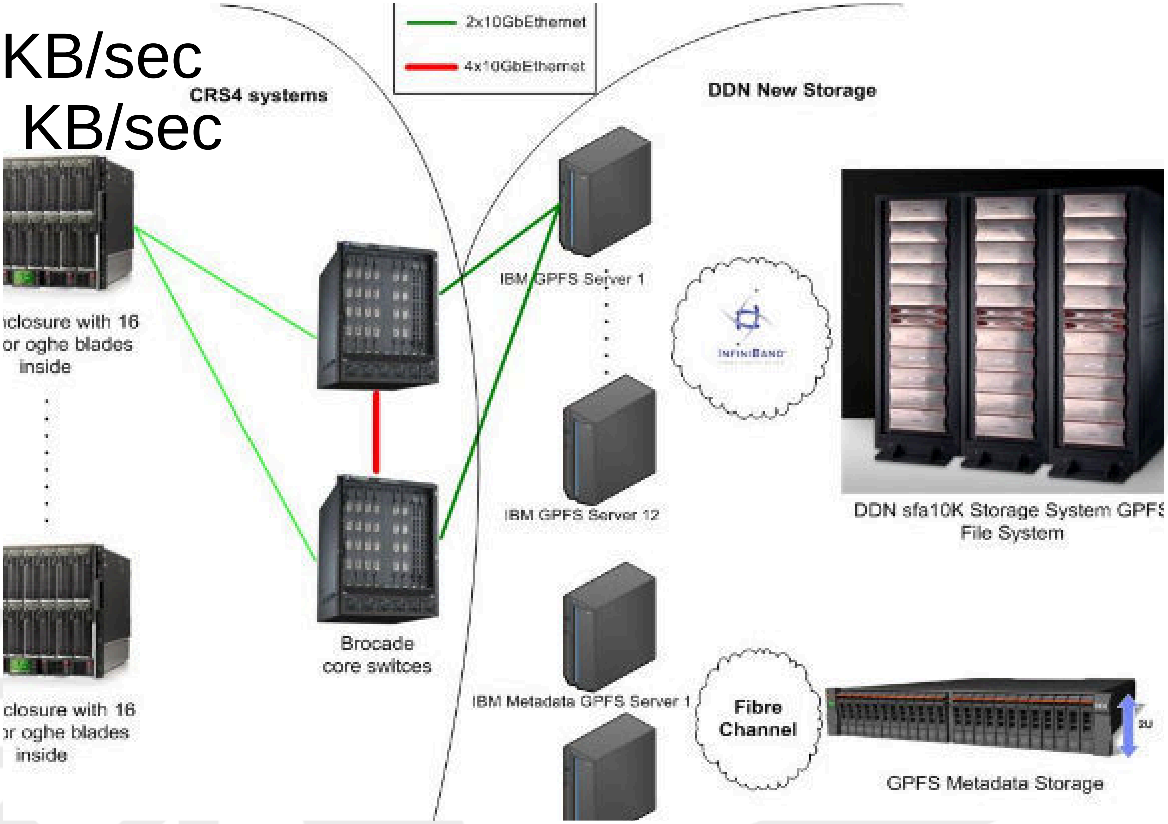
Thank you for your time !

Why Hadoop?

- Hadoop provides a framework for easy development of distributed and robust applications
- Both are necessary for scalability
  - More machines = more failures
  - More jobs = more failures
- A robust system is required to sustain a good throughput
- Hadoop provides a resilient mechanism that resists many hardware failures and transient cluster conditions

GPFS benchmark

- Direct Data networks sfa10k
- 5.4 PB raw / 4.3 PB raid6
- GPFS filesystems
- 352 hosts benchmark:
  - random read = 17438042.89 KB/sec
  - random write = 27511029.63 KB/sec
  - more than 15GB/sec


🟩 2x10GbEthernet
🟥 4x10GbEthernet
CRS4 systems
DDN New Storage
nclosure with 16 or oghe blades inside
IBM GPFS Server 1
INFINIBAND
IBM GPFS Server 12
DDN sfa10K Storage System GPFS File System
Brocade core switces
nclosure with 16 or oghe blades inside
IBM Metadata GPFS Server 1
Fibre Channel
2U
GPFS Metadata Storage