Computation Institute

Integrating Galaxy with Globus Online: Lessons Learned from the CVRG Project

Bo Liu

Computation Institute, University of Chicago
Argonne National Laboratory

CardioVascular Research Grid | Argonne NATIONAL LABORATORY | THE UNIVERSITY OF CHICAGO | www.ci.anl.gov
www.ci.uchicago.edu

Agenda

- Introduction
- New Galaxy Tools
  - Globus Online
  - CRData
  - Picard/GATK via Condor
  - CummeRbund
  - Miso
- System Implementation
  - Amazon Cloud
  - Globus Provision
- CVRG Use Case
- Integration of Galaxy and Globus
- Conclusions & Future Plan

Introduction

- Computation Institute
  - A joint institute of Argonne National Laboratory and the University of Chicago
  - Interdisciplinary approaches to challenging systems problems
  - Development and application of advanced computational methods
  - Co-hosts of the GCC 2012

CardioVascular Research Grid

- CVRG project
  - Funded by NHLBI
  - Aim: create an infrastructure for sharing cardiovascular data and data analysis tools
  - Collaborator: Institute for Computational Medicine at the Johns Hopkins University

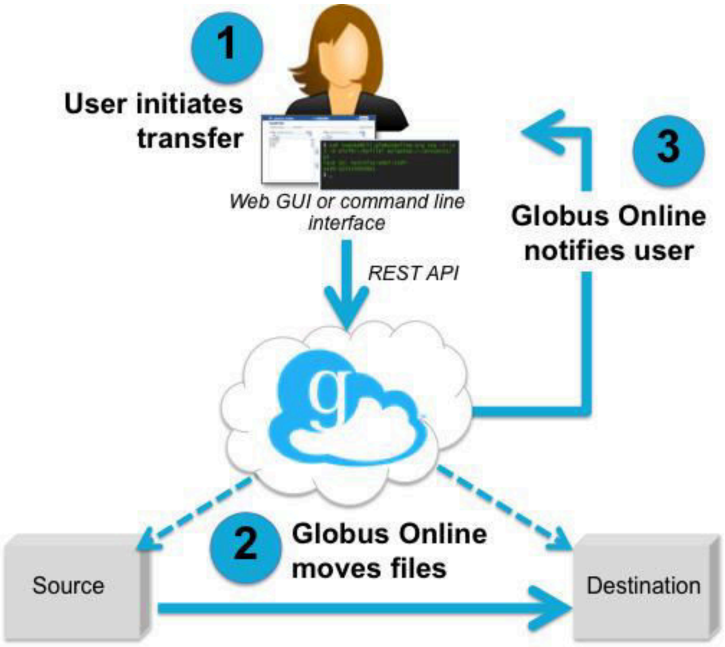
http://cvrgrid.org/

Introduction

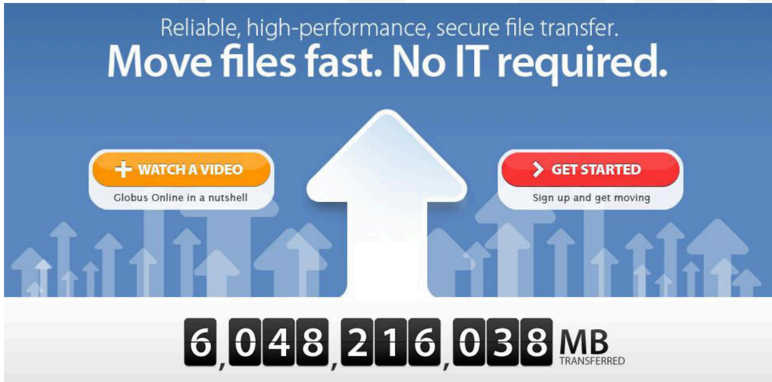
- Galaxy
  - Web-based platform, ease of use and distribution
  - Various sequencing analysis tools
  - Easy data/workflow/history sharing
- Challenges
  - Distributed data at different locations
    - Transfer and sharing data with other researchers
  - Inefficient ways of data movement
    - Browser Upload: 2GB limit
    - FTP: not stable enough
  - Distributed computing for high performance
  - Requirements for new tools
    - CRData, execute R script
    - CummeRbund , analyze Cufflinks RNA-Seq output
    - Miso, compute Psi values for genes/events
  - Automatic deployment of Galaxy on Cloud

Globus Transfer

- Globus Online - No IT required
  - Software as a Service (SaaS)
  - Consolidated support & troubleshooting
  - Works with existing GridFTP servers
  - Globus Connect solves “last mile problem”
- Reliable file transfer
  - Easy “fire-and-forget” transfers
  - Automatic fault recovery
  - High performance
  - Across multiple security domains
- > 5500 registered users, > 6PB moved
- Recommended and used by DOE Facilities, NSF Supercomputing centers, many campuses


1
User initiates transfer
Web GUI or command line interface
REST API
3
Globus Online notifies user
2
Globus Online moves files
Source
Destination

https://www.globusonline.org/

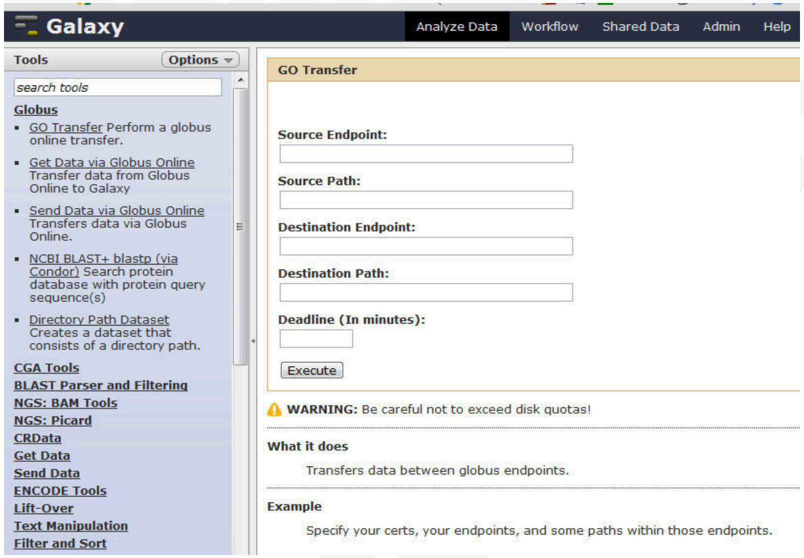

Reliable, high-performance, secure file transfer.

Move files fast. No IT required.

6,048,216,038 MB TRANSFERRED

Integration of Galaxy and Globus Online

- Transfer large quantities of data in and out of Galaxy fast and reliably
- 3 Globus Online tools
  - GO transfer
    - transferring data between different endpoints
  - Get data via Globus Online
    - getting data from GO endpoint to Galaxy server
  - Send data via Globus Online
    - sending data from Galaxy to GO endpoint
- Set "Deadline" for transfer
- Email notification
- Transfer monitoring



Globus
- GO Transfer Perform a globus online transfer.
- Get Data via Globus Online Transfer data from Globus Online to Galaxy
- Send Data via Globus Online Transfers data via Globus Online.
- NCBI BLAST+ blastp (via Condor) Search protein database with protein query sequence(s)
- Directory Path Dataset Creates a dataset that consists of a directory path.

CGA Tools
BLAST Parser and Filtering
NGS: BAM Tools
NGS: Picard
CRData
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort

GO Transfer

Source Endpoint:

Source Path:

Destination Endpoint:

Destination Path:

Deadline (In minutes):

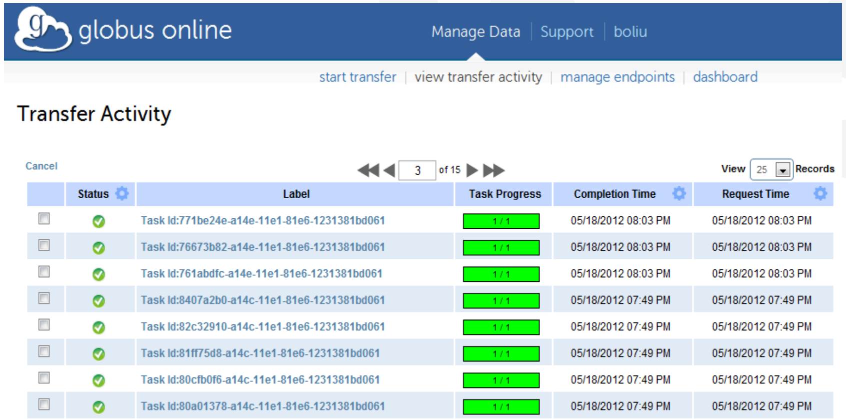
WARNING: Be careful not to exceed disk quotas!

What it does

Transfers data between globus endpoints.

Example

Specify your certs, your endpoints, and some paths within those endpoints.

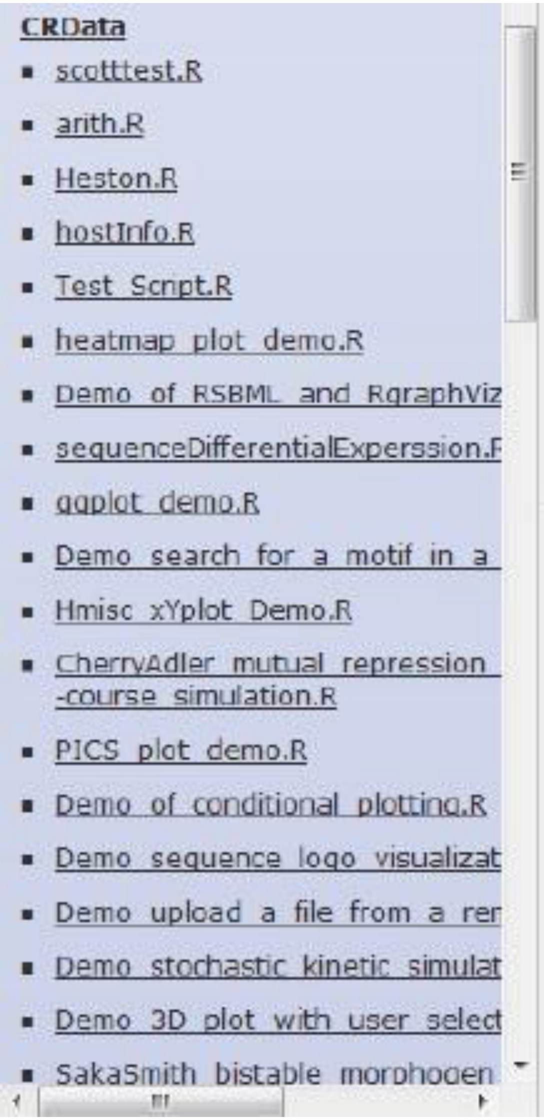

globus online
Manage Data | Support | boliu
start transfer | view transfer activity | manage endpoints | dashboard
Cancel
3 of 15
View 25 Records

Transfer Activity

	Status	Label	Task Progress	Completion Time	Request Time
□	✓	Task Id:771be24e-a14e-11e1-81e6-1231381bd061	1 / 1	05/18/2012 08:03 PM	05/18/2012 08:03 PM
□	✓	Task Id:76673b82-a14e-11e1-81e6-1231381bd061	1 / 1	05/18/2012 08:03 PM	05/18/2012 08:03 PM
□	✓	Task Id:761abdfc-a14e-11e1-81e6-1231381bd061	1 / 1	05/18/2012 08:03 PM	05/18/2012 08:03 PM
□	✓	Task Id:8407a2b0-a14c-11e1-81e6-1231381bd061	1 / 1	05/18/2012 07:49 PM	05/18/2012 07:49 PM
□	✓	Task Id:82c32910-a14c-11e1-81e6-1231381bd061	1 / 1	05/18/2012 07:49 PM	05/18/2012 07:49 PM
□	✓	Task Id:81ff75d8-a14c-11e1-81e6-1231381bd061	1 / 1	05/18/2012 07:49 PM	05/18/2012 07:49 PM
□	✓	Task Id:80cfb0f6-a14c-11e1-81e6-1231381bd061	1 / 1	05/18/2012 07:49 PM	05/18/2012 07:49 PM
□	✓	Task Id:80a01378-a14c-11e1-81e6-1231381bd061	1 / 1	05/18/2012 07:49 PM	05/18/2012 07:49 PM

CRData

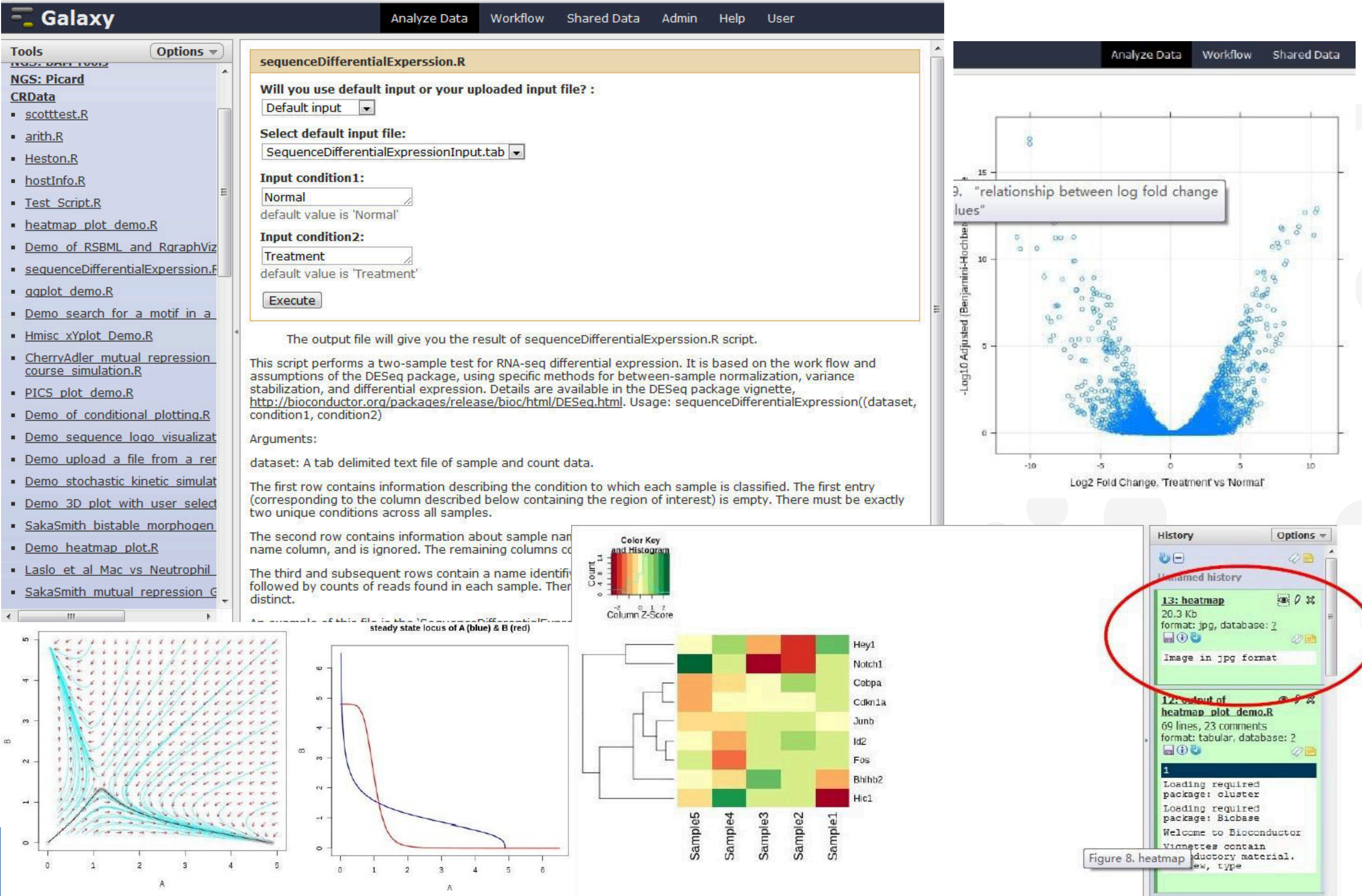
- R
  - A software for statistical computing and graphics
- CRData
  - Originally created as part of crdata.org
  - Wrap 35 tools
    - heatmap_plot_demo.R
    - sequenceDifferentialExperssion.R
    - affyClassify.R
    - ……
  - Execute a set of BioConductor R scripts
  - Used for analyzing ECG Data in CVRG
- Complements the functionality of executing CRData R scripts in Galaxy



CRData

- scotttest.R
- arith.R
- Heston.R
- hostInfo.R
- Test_Script.R
- heatmap_plot_demo.R
- Demo_of_RSBML_and_RgraphViz
- sequenceDifferentialExperssion.R
- ggplot_demo.R
- Demo_search_for_a_motif_in_a
- Hmisc_xYplot_Demo.R
- CherryAdler_mutual_repression_-course_simulation.R
- PICS_plot_demo.R
- Demo_of_conditional_plotting.R
- Demo_sequence_logo_visualizat
- Demo_upload_a_file_from_a_rer
- Demo_stochastic_kinetic_simulat
- Demo_3D_plot_with_user_select
- SakaSmith_bistable_morphogen

CRData



Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Tools
Options
NGS: Picard
CRData
scotttest.R
arith.R
Heston.R
hostInfo.R
Test_Script.R
heatmap_plot_demo.R
Demo_of_RSBML_and_RgraphViz
sequenceDifferentialExperssion.R
ggplot_demo.R
Demo_search_for_a_motif_in_a
Hmisc_xYplot_Demo.R
CherryAdler_mutual_repression_course_simulation.R
PICS_plot_demo.R
Demo_of_conditional_plotting.R
Demo_sequence_logo_visualizat
Demo_upload_a_file_from_a_rer
Demo_stochastic_kinetic_simulat
Demo_3D_plot_with_user_select
SakaSmith_bistable_morphogen
Demo_heatmap_plot.R
Laslo_et_al_Mac_vs_Neutrophil
SakaSmith_mutual_repression_G

sequenceDifferentialExperssion.R

Will you use default input or your uploaded input file? : Default input

Select default input file: SequenceDifferentialExpressionInput.tab

Input condition1: Normal
default value is 'Normal'

Input condition2: Treatment
default value is 'Treatment'

Execute

The output file will give you the result of sequenceDifferentialExperssion.R script.

This script performs a two-sample test for RNA-seq differential expression. It is based on the work flow and assumptions of the DESeq package, using specific methods for between-sample normalization, variance stabilization, and differential expression. Details are available in the DESeq package vignette, http://bioconductor.org/packages/release/bioc/html/DESeq.html. Usage: sequenceDifferentialExpression((dataset, condition1, condition2)

Arguments:

dataset: A tab delimited text file of sample and count data.

The first row contains information describing the condition to which each sample is classified. The first entry (corresponding to the column described below containing the region of interest) is empty. There must be exactly two unique conditions across all samples.

The second row contains information about sample nan
name column, and is ignored. The remaining columns cc

The third and subsequent rows contain a name identifi
followed by counts of reads found in each sample. Ther
distinct.

13: heatmap
20.3 Kb
format: jpg, database: ?

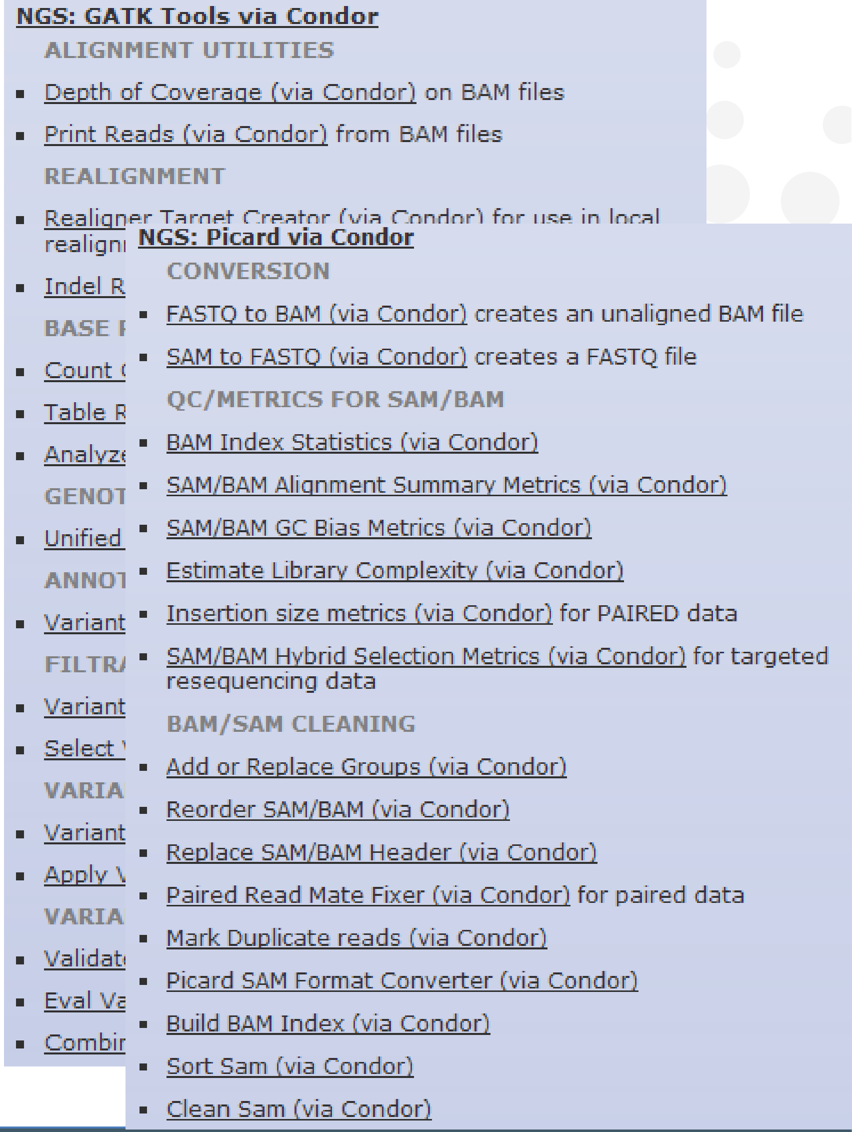
Image in jpg format

12: output of heatmap_plot_demo.R
69 lines, 23 comments
format: tabular, database: ?

```
1
Loading required
package: cluster
Loading required
package: Biobase
Welcome to Bioconductor
Vignettes contain
ductory material.
ew, type
```

Picard/GATK via Condor

- Wrap all the Picard and GATK tools to run through Condor job scheduler
- Leverage local cluster or cloud based scalable computational resources for parallelizing the tools
- Condor
  - Parallel execution
  - Automatically job scheduling
  - Condor pool: multiple worker nodes, easily add or remove
- Condor Runner for Galaxy
  - The executable along with command line options get passed to a condor runner
  - Applications are run on a worker node instead of the galaxy node
  - Enable faster and more efficient execution of Galaxy jobs



NGS: GATK Tools via Condor

ALIGNMENT UTILITIES

- Depth of Coverage (via Condor) on BAM files
- Print Reads (via Condor) from BAM files

REALIGNMENT

- Realigner Target Creator (via Condor) for use in local realigni
- Indel R

BASE I

- Count (
- Table R
- Analyze

GENOT

- Unified

ANNOT

- Variant

FILTRA

- Variant
- Select \

VARIA

- Variant
- Apply \

VARIA

- Validate
- Eval Va
- Combir

NGS: Picard via Condor

CONVERSION

- FASTQ to BAM (via Condor) creates an unaligned BAM file
- SAM to FASTQ (via Condor) creates a FASTQ file

QC/METRICS FOR SAM/BAM

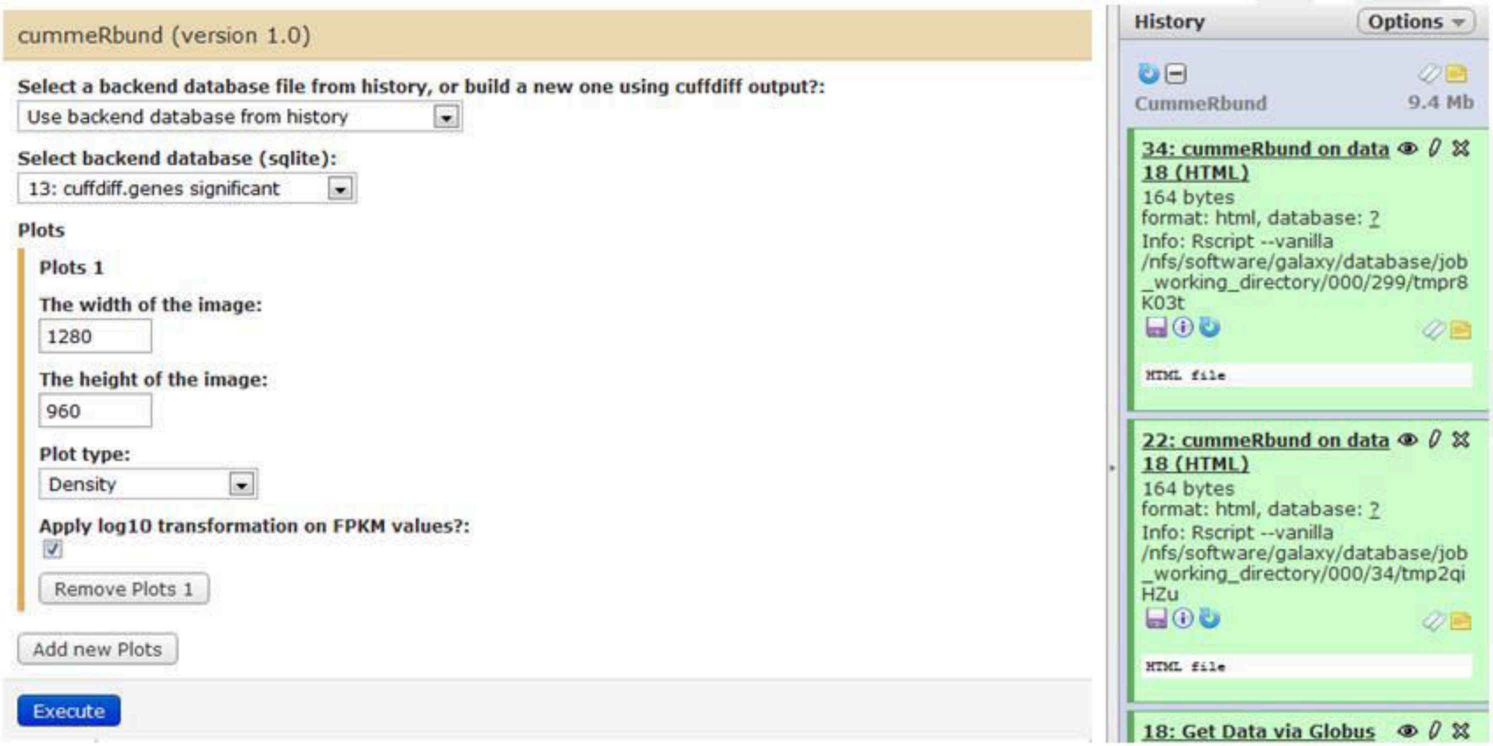
- BAM Index Statistics (via Condor)
- SAM/BAM Alignment Summary Metrics (via Condor)
- SAM/BAM GC Bias Metrics (via Condor)
- Estimate Library Complexity (via Condor)
- Insertion size metrics (via Condor) for PAIRED data
- SAM/BAM Hybrid Selection Metrics (via Condor) for targeted resequencing data

BAM/SAM CLEANING

- Add or Replace Groups (via Condor)
- Reorder SAM/BAM (via Condor)
- Replace SAM/BAM Header (via Condor)
- Paired Read Mate Fixer (via Condor) for paired data
- Mark Duplicate reads (via Condor)
- Picard SAM Format Converter (via Condor)
- Build BAM Index (via Condor)
- Sort Sam (via Condor)
- Clean Sam (via Condor)

CummeRbund

- An R package designed to aid and simplify the task of analyzing Cufflinks RNA-Seq output
- Input: a backend database file from the history, or build a new database using cuffdiff output
- Plot type: Density, Boxplot, Scatter, Volcano, Heatmap, Cluster, Expression Plot, Expression Bar Plot



cummeRbund (version 1.0)

Select a backend database file from history, or build a new one using cuffdiff output?: Use backend database from history

Select backend database (sqlite): 13: cuffdiff.genes significant

Plots

Plots 1

The width of the image: 1280

The height of the image: 960

Plot type: Density

Apply log10 transformation on FPKM values?: ☑

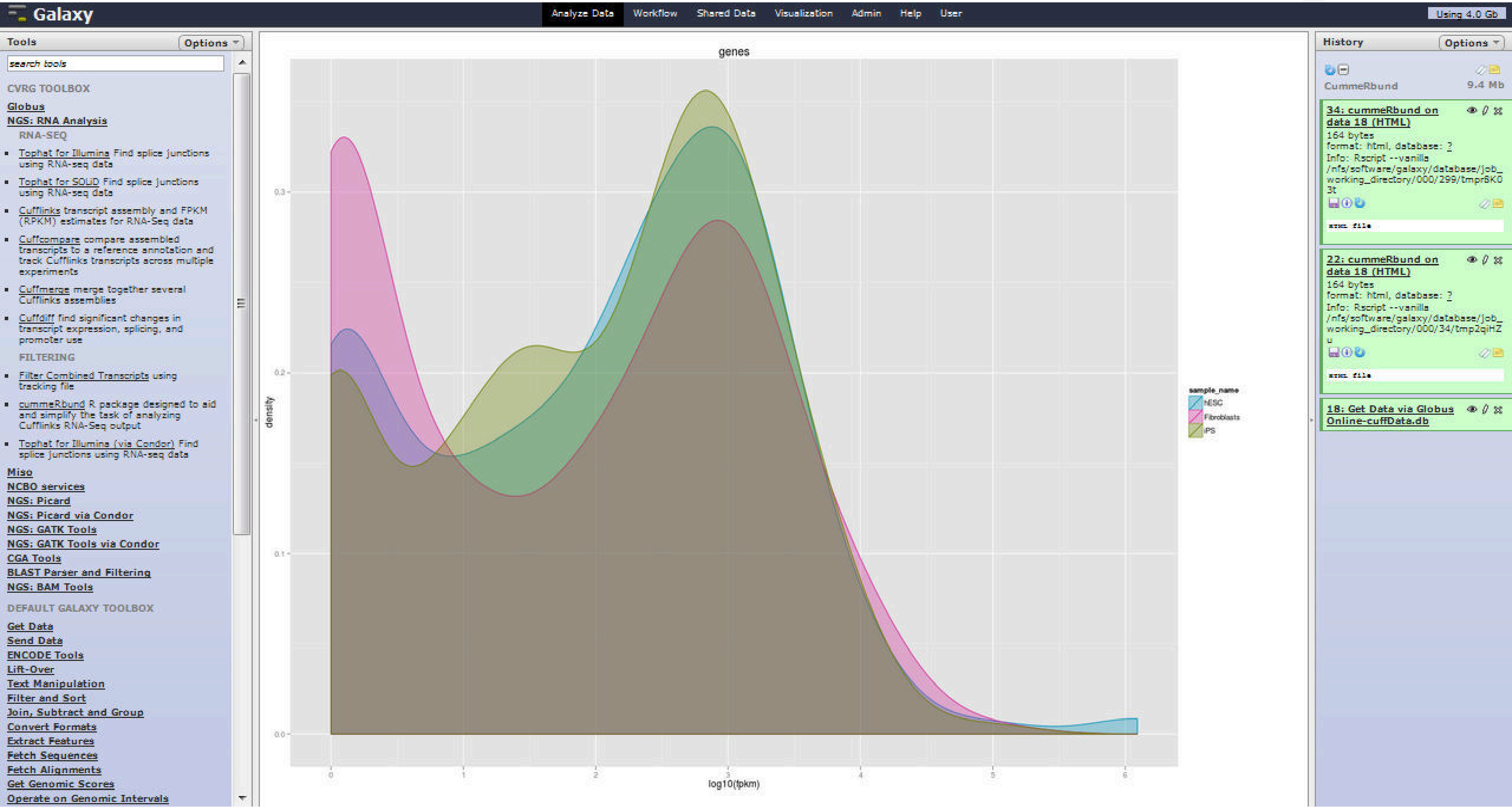
CummeRbund 9.4 Mb

34: cummeRbund on data 18 (HTML)
164 bytes
format: html, database: ?
Info: Rscript --vanilla /nfs/software/galaxy/database/job_working_directory/000/299/tmpr8K03t
HTML file

22: cummeRbund on data 18 (HTML)
164 bytes
format: html, database: ?
Info: Rscript --vanilla /nfs/software/galaxy/database/job_working_directory/000/34/tmp2qiHZu
HTML file

18: Get Data via Globus

CummeRbund



CVRG TOOLBOX

Globus

NGS: RNA Analysis

RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data
- Tophat for SOLiD Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffmerge merge together several Cufflinks assemblies
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

FILTERING

- Filter Combined Transcripts using tracking file
- cummeRbund R package designed to aid and simplify the task of analyzing Cufflinks RNA-Seq output
- Tophat for Illumina (via Condor) Find splice junctions using RNA-seq data

Miso
NCBO services
NGS: Picard
NGS: Picard via Condor
NGS: GATK Tools
NGS: GATK Tools via Condor
CGA Tools
BLAST Parser and Filtering
NGS: BAM Tools

DEFAULT GALAXY TOOLBOX

Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals

CummeRbund 9.4 Mb

34: cummeRbund on data 18 (HTML)
164 bytes
format: html, database: ?
Info: Rscript --vanilla /nfs/software/galaxy/database/job_working_directory/000/299/tmpr8K03t

error file

22: cummeRbund on data 18 (HTML)
164 bytes
format: html, database: ?
Info: Rscript --vanilla /nfs/software/galaxy/database/job_working_directory/000/34/tmp2qiHZu

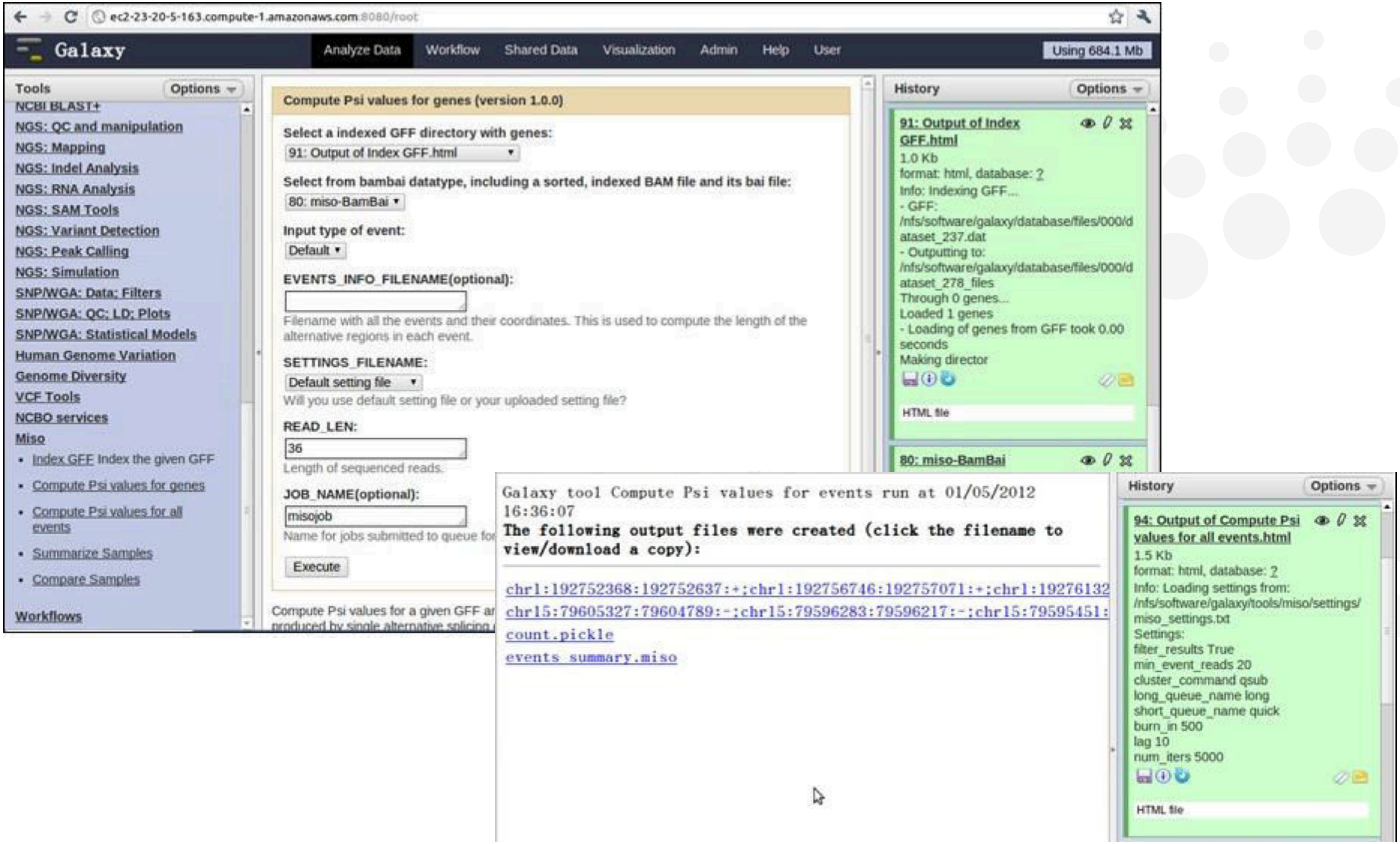
error file

18: Get Data via Globus Online-cuffData.db

Miso

- MISO (Mixture-of-Isoforms)
  - a probabilistic framework
  - Quantitate the expression level of spliced genes from RNA-Seq data
  - Identify regulated isoforms or exons across samples
- Wrap 5 Miso tools
  - Index GFF
    - create an index for the GFF annotations
  - Compute Psi values for genes
  - Compute Psi values for all events
    - compute Psi values for a given GFF annotation of either whole mRNA isoforms or isoforms produced by single alternative splicing events
  - Summarize Samples
    - summarize MISO output and obtain confidence intervals for Psi values
  - Compare Samples
    - compute comparison statistics between the two given samples

Miso



Compute Psi values for genes (version 1.0.0)

Select a indexed GFF directory with genes: 91: Output of Index GFF.html

Select from bambai datatype, including a sorted, indexed BAM file and its bai file: 80: miso-BamBai

Input type of event: Default

EVENTS_INFO_FILENAME(optional):

Filename with all the events and their coordinates. This is used to compute the length of the alternative regions in each event.

SETTINGS_FILENAME: Default setting file

Will you use default setting file or your uploaded setting file?

READ_LEN: 36

Length of sequenced reads.

JOB_NAME(optional): misojob

Name for jobs submitted to queue for

Compute Psi values for a given GFF ar produced by single alternative splicing

91: Output of Index GFF.html
1.0 Kb
format: html, database: ?
Info: Indexing GFF...
- GFF: /nfs/software/galaxy/database/files/000/dataset_237.dat
- Outputting to: /nfs/software/galaxy/database/files/000/dataset_278_files
Through 0 genes...
Loaded 1 genes
- Loading of genes from GFF took 0.00 seconds
Making director
HTML file

80: miso-BamBai

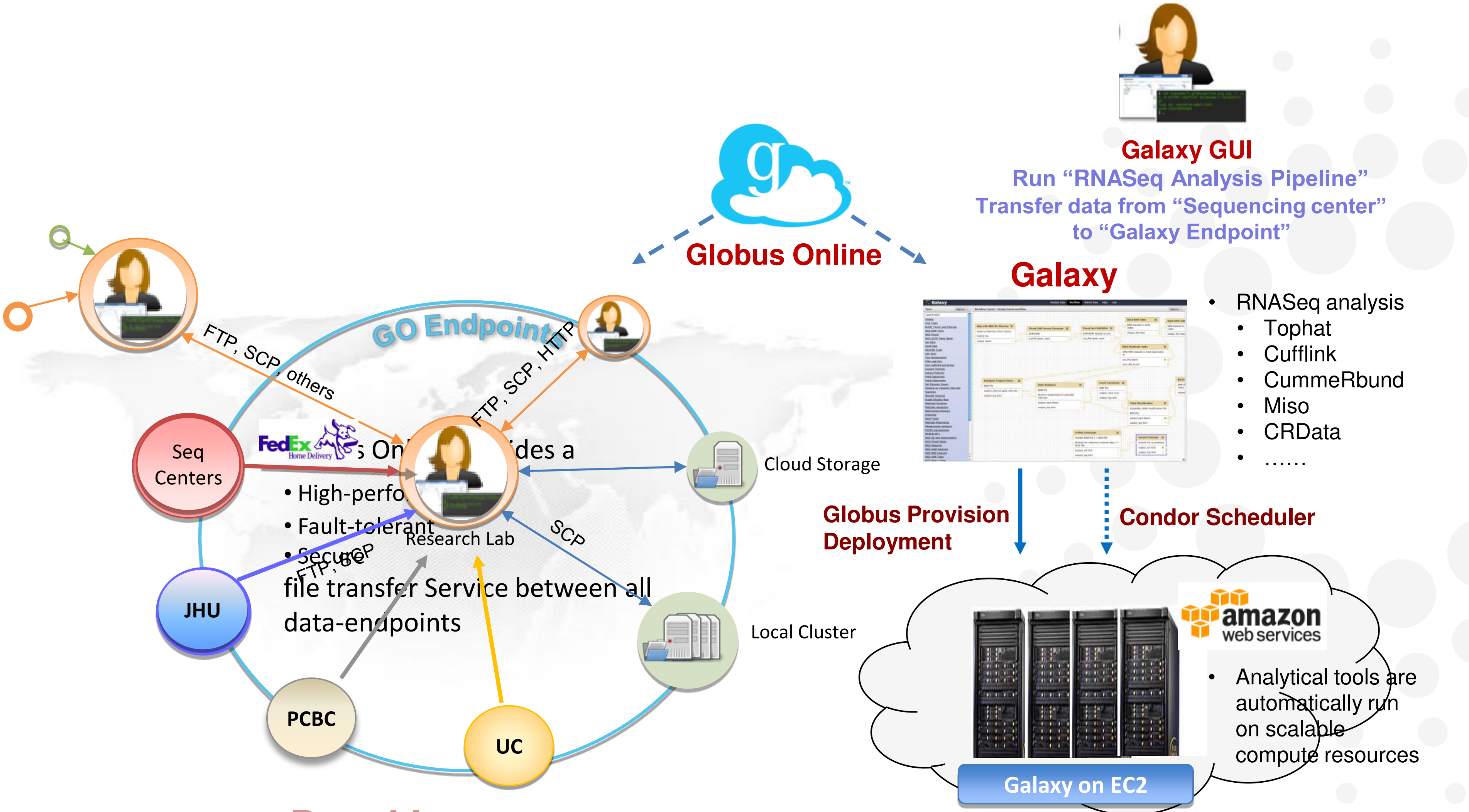
```
Galaxy tool Compute Psi values for events run at 01/05/2012
16:36:07
The following output files were created (click the filename to
view/download a copy):

chr1:192752368:192752637:+;chr1:192756746:192757071:+;chr1:19276132
chr15:79605327:79604789:-;chr15:79596283:79596217:-;chr15:79595451:
count.pickle
events_summary.miso
```

94: Output of Compute Psi values for all events.html
1.5 Kb
format: html, database: ?
Info: Loading settings from: /nfs/software/galaxy/tools/miso/settings/miso_settings.txt
Settings:
filter_results True
min_event_reads 20
cluster_command qsub
long_queue_name long
short_queue_name quick
burn_in 500
lag 10
num_iters 5000
HTML file

System Implementation

- Amazon EC2/EBS
  - scalable computing and storage capabilities
- Deploying Galaxy on Cloud
  - on-demand resource configuration
  - usage-based pricing
- Globus Provision
  - a tool for deploying fully-configured Globus systems
  - Automate the process of deploying Galaxy on EC2
  - Provision a EC2 cluster with Galaxy, Globus Online, Condor, GridFTP server and a set of users in <15mins
  - Reusable Chef recipes to provision production galaxy instances on demand

CVRG Use Case



s On ides a

- High-perfo
- Fault-tolerant
- Secure

file transfer Service between all data-endpoints

Galaxy GUI

Run "RNASeq Analysis Pipeline"
Transfer data from "Sequencing center" to "Galaxy Endpoint"

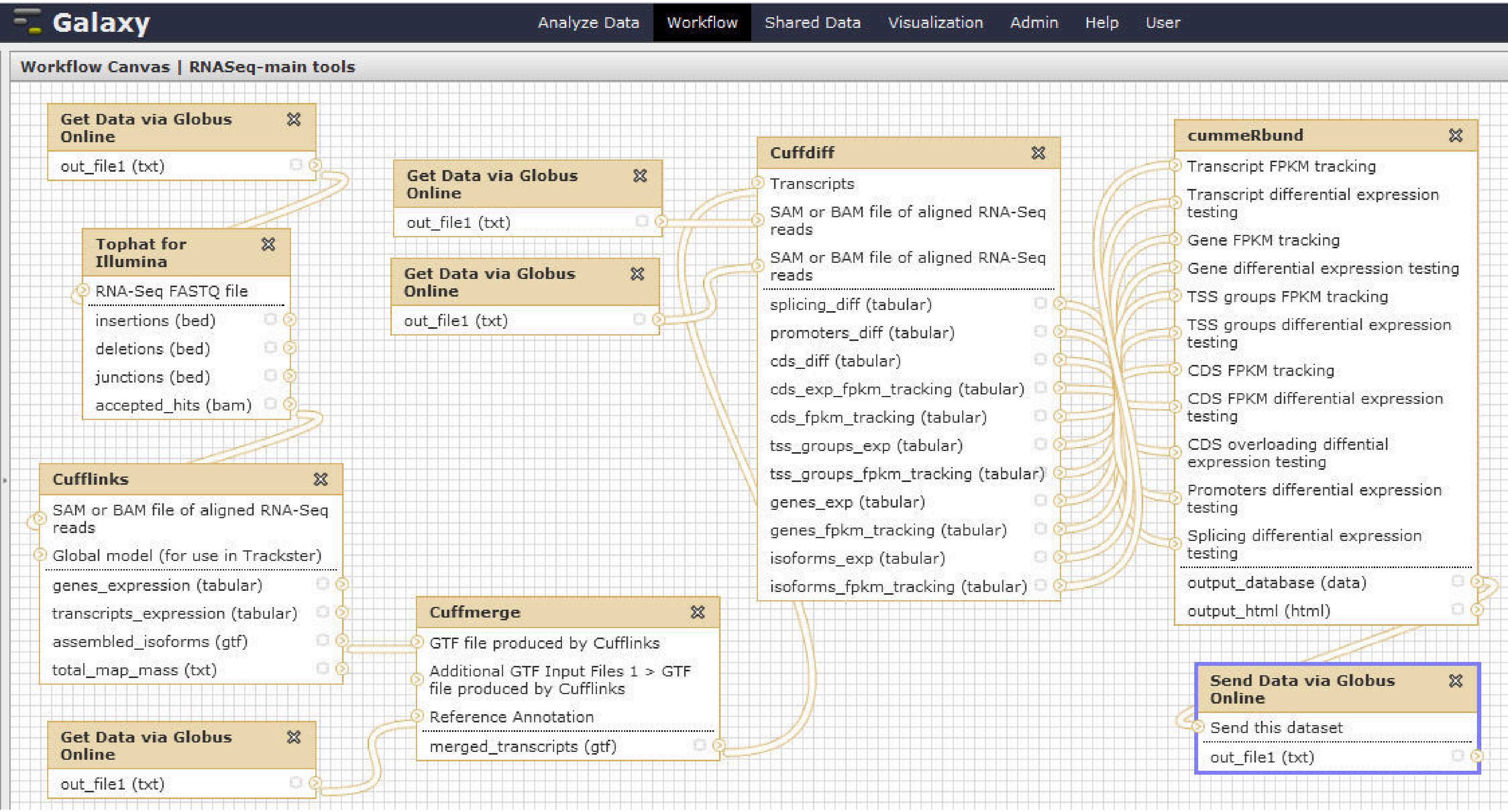
- RNASeq analysis
  - Tophat
  - Cufflink
  - CummeRbund
  - Miso
  - CRData
  - ……

- Analytical tools are automatically run on scalable compute resources

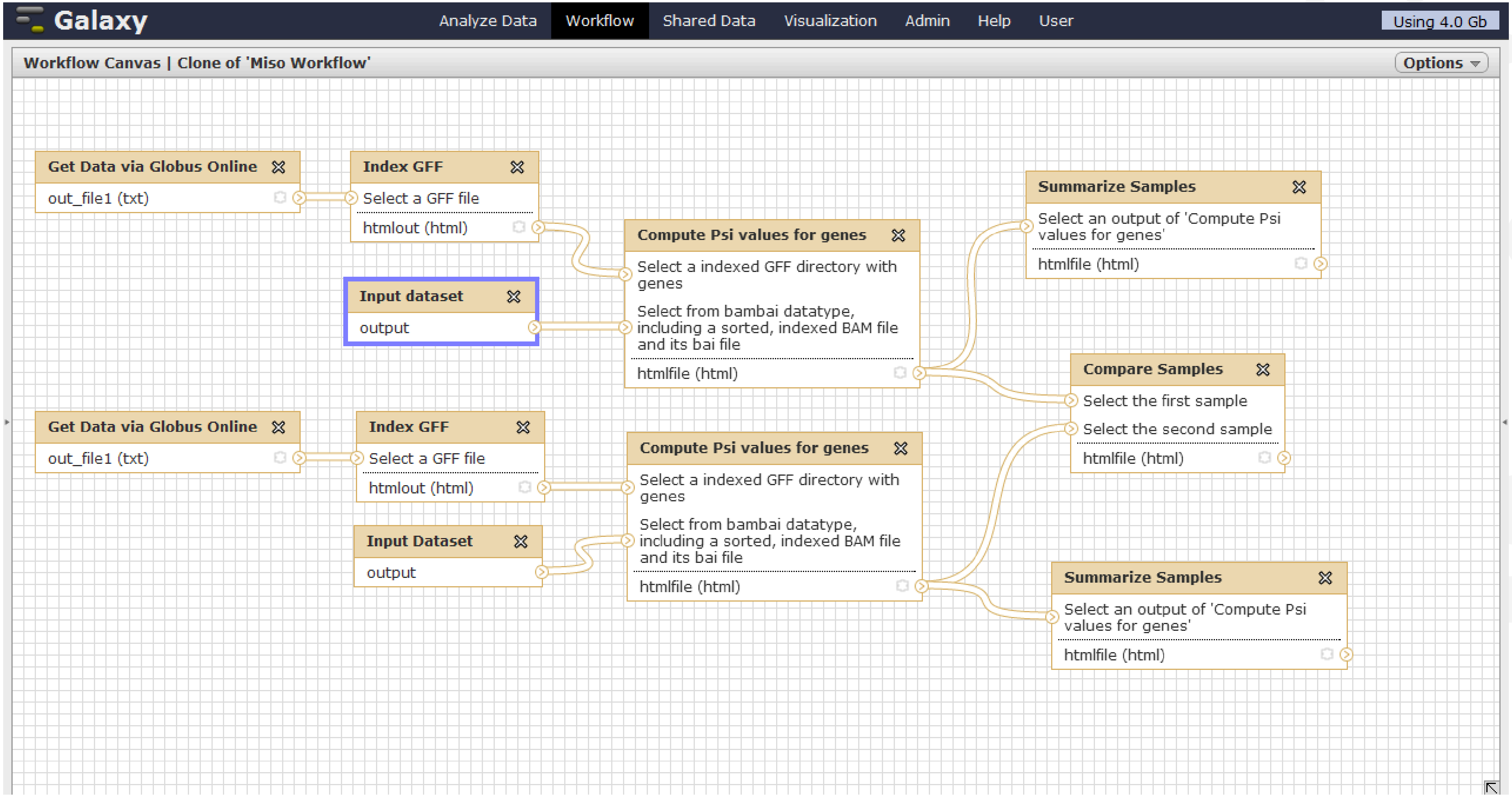
Data Management

Data Analysis

RNASeq Analysis Pipeline


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Workflow Canvas | RNASeq-main tools
Get Data via Globus Online
out_file1 (txt)
Tophat for Illumina
RNA-Seq FASTQ file
insertions (bed)
deletions (bed)
junctions (bed)
accepted_hits (bam)
Cufflinks
SAM or BAM file of aligned RNA-Seq reads
Global model (for use in Trackster)
genes_expression (tabular)
transcripts_expression (tabular)
assembled_isoforms (gtf)
total_map_mass (txt)
Get Data via Globus Online
out_file1 (txt)
Get Data via Globus Online
out_file1 (txt)
Get Data via Globus Online
out_file1 (txt)
Cuffmerge
GTF file produced by Cufflinks
Additional GTF Input Files 1 > GTF file produced by Cufflinks
Reference Annotation
merged_transcripts (gtf)
Cuffdiff
Transcripts
SAM or BAM file of aligned RNA-Seq reads
SAM or BAM file of aligned RNA-Seq reads
splicing_diff (tabular)
promoters_diff (tabular)
cds_diff (tabular)
cds_exp_fpkm_tracking (tabular)
cds_fpkm_tracking (tabular)
tss_groups_exp (tabular)
tss_groups_fpkm_tracking (tabular)
genes_exp (tabular)
genes_fpkm_tracking (tabular)
isoforms_exp (tabular)
isoforms_fpkm_tracking (tabular)
cummeRbund
Transcript FPKM tracking
Transcript differential expression testing
Gene FPKM tracking
Gene differential expression testing
TSS groups FPKM tracking
TSS groups differential expression testing
CDS FPKM tracking
CDS FPKM differential expression testing
CDS overloading diffential expression testing
Promoters differential expression testing
Splicing differential expression testing
output_database (data)
output_html (html)
Send Data via Globus Online
Send this dataset
out_file1 (txt)

Miso Pipeline



Workflow Canvas | Clone of 'Miso Workflow'

Get Data via Globus Online
out_file1 (txt)

Index GFF
Select a GFF file
htmlout (html)

Input dataset
output

Compute Psi values for genes
Select a indexed GFF directory with genes
Select from bambai datatype, including a sorted, indexed BAM file and its bai file
htmlfile (html)

Summarize Samples
Select an output of 'Compute Psi values for genes'
htmlfile (html)

Compare Samples
Select the first sample
Select the second sample
htmlfile (html)

Get Data via Globus Online
out_file1 (txt)

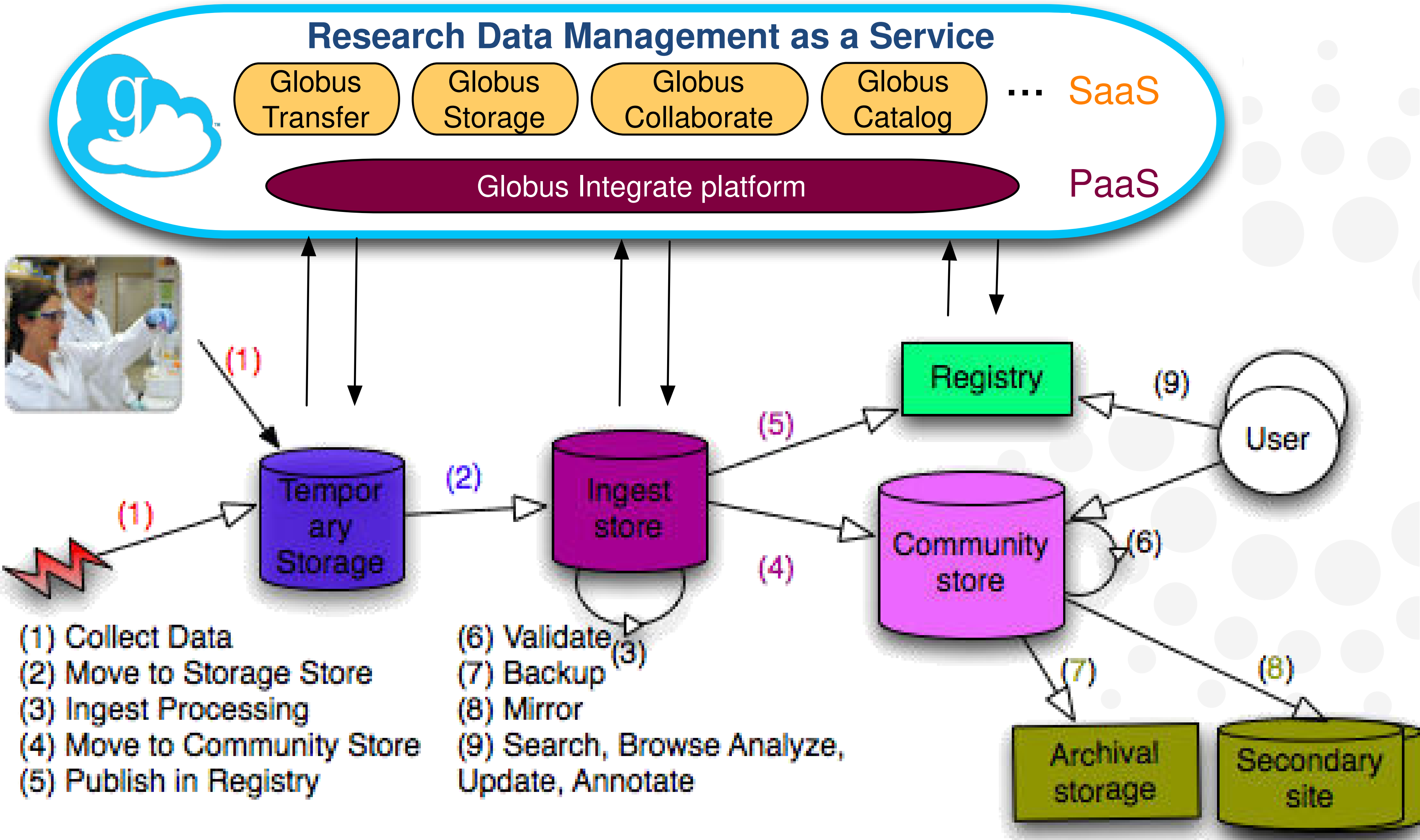
Index GFF
Select a GFF file
htmlout (html)

Input Dataset
output

Compute Psi values for genes
Select a indexed GFF directory with genes
Select from bambai datatype, including a sorted, indexed BAM file and its bai file
htmlfile (html)

Summarize Samples
Select an output of 'Compute Psi values for genes'
htmlfile (html)

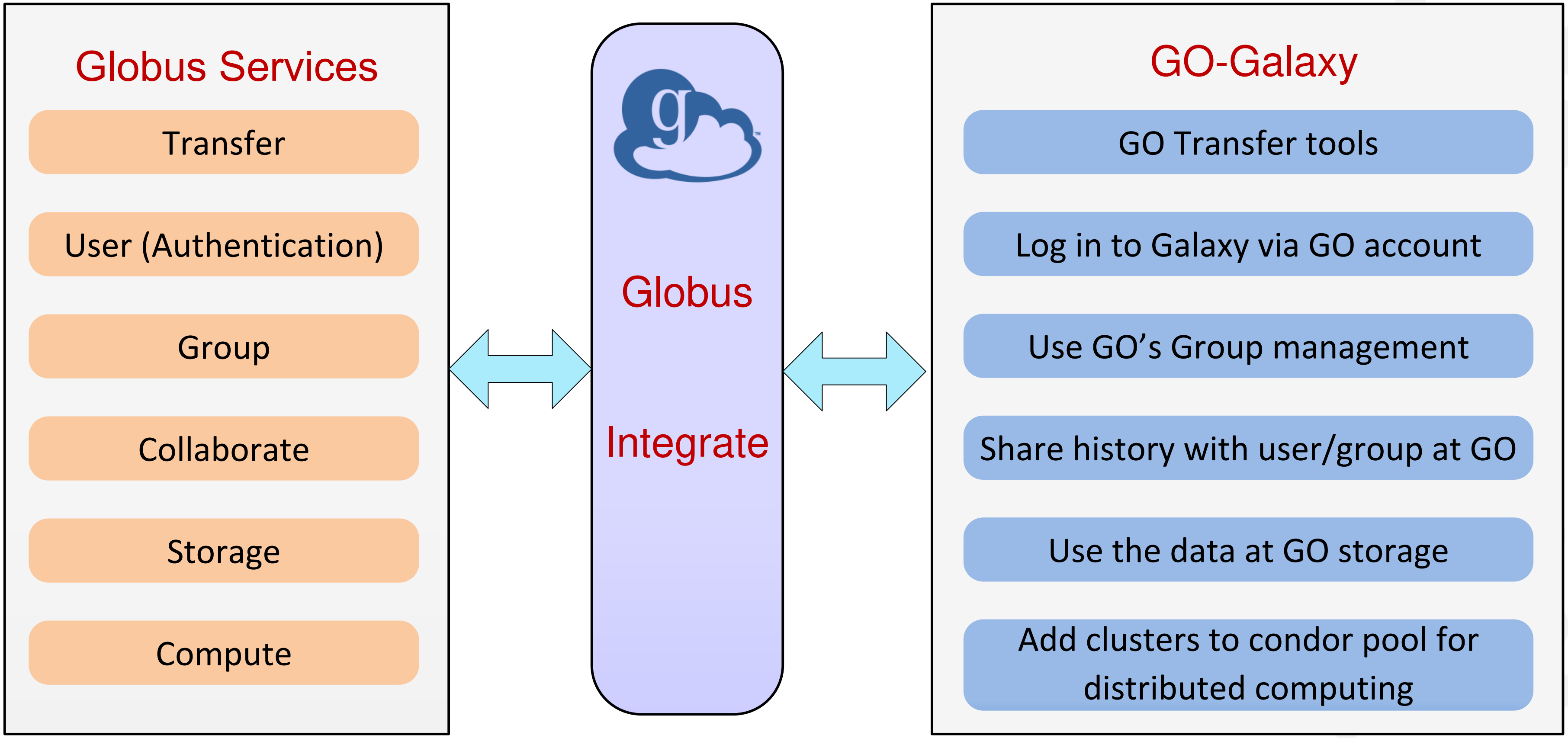
Towards "research IT as a service"



(1) Collect Data
(2) Move to Storage Store
(3) Ingest Processing
(4) Move to Community Store
(5) Publish in Registry

(6) Validate
(7) Backup
(8) Mirror
(9) Search, Browse Analyze, Update, Annotate

Integrate Galaxy and Globus Online



Globus Services

- Transfer
- User (Authentication)
- Group
- Collaborate
- Storage
- Compute

Globus

Integrate

GO-Galaxy

- GO Transfer tools
- Log in to Galaxy via GO account
- Use GO’s Group management
- Share history with user/group at GO
- Use the data at GO storage
- Add clusters to condor pool for distributed computing

Conclusions & Future Plan

- Galaxy + Globus Online
  - Transfer large-scale datasets in and out of Galaxy in a secure, efficient and fast way
- New Galaxy tools
  - CRData: processing R scripts
  - CummeRbund, Miso: RNA analysis
  - Condor: distributed computing capabilities
- Globus Provision
  - Automatically deploy Galaxy on EC2 with user-specific configuration
- Future Plan
  - Thorough integration of Galaxy and Globus
  - Performance optimization
- Galaxy community Toolshed contributions

Thank you for your attention

Bo Liu: boliu@uchicago.edu
Ravi Madduri: madduri@mcs.anl.gov

This project incorporates tools whose development was funded in part by the NIH through the NHLBI grant: The Cardiovascular Research Grid (R24HL085343).