




# Practical Solutions for Big Data Analytics

Ravi Madduri

(madduri@anl.gov)

Paul Dave

(pdave@uchicago.edu)

Dinanath Sulakhe

(sulakhe@uchicago.edu)

Alex Rodriguez

(arodri7@uchicago.edu)

*Urban Science*

**Genomics**

High energy physics

Molecular biology

**Climate change**

Cosmology

*Linguistics*

**Metagenomics**

Visual arts

**Economics**

We are a non-profit organization of researchers, developers, and bioinformaticians, building solutions for the advancement of research in various fields

# Our vision for a 21st century discovery infrastructure

To provide **more** capability for **more** people at **substantially lower cost**

# Agenda

12:30pm Challenges for biomedical analysis at scale

12:45pm Best practices and solution components

1:00pm Introduction to Globus Genomics, Globus Transfer

1:30pm Exercise: Transferring raw NGS datasets from sequencing centers and sharing

2:00pm Refreshment Break

2:15pm Demonstration of QC pipeline, Exome, RNA, Whole Genome analysis

2:30pm Exercise: Running example pipelines with sample data sets

2:45pm Running analysis at scale using Globus Genomics

3:15pm Exercise/Demonstration: Executing Exome, RNA, QC pipelines at scale

3:45pm Interactive Q&A and Session Wrap-up

All materials are available at:
**http://tinyurl.com/lnsoy49**
Please complete the sign-up form.

Thank you!

# Challenges in Biomedical analysis at scale

# “I need…

- …to get my sequence data from the NGS core to the lab for analysis.”
- …to easily, quickly, and reliably move or mirror (some or all of) my data to other places
  - Lab server, HPC cluster, desktop, public cloud server
- …to easily and securely share my data with my colleagues at other institutions.”
- …to make my data available for others to replicate my experiments.”
- …a good place to backup and archive my research data, at a reasonable price.”

# Managing data should be easy …

# … but it's hard and frustrating!

Staging Store

! Expired credentials

! Permission denied

Scale-out Store

! NATs Firewalls

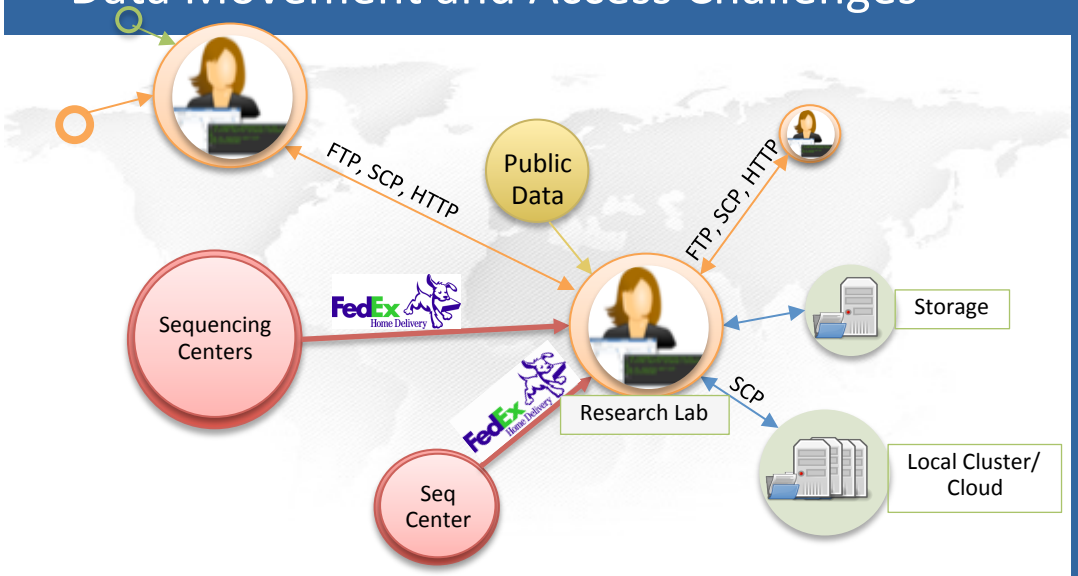
! Failed Server / Net

Archive

Mirror

# Challenges in Sequencing Analysis

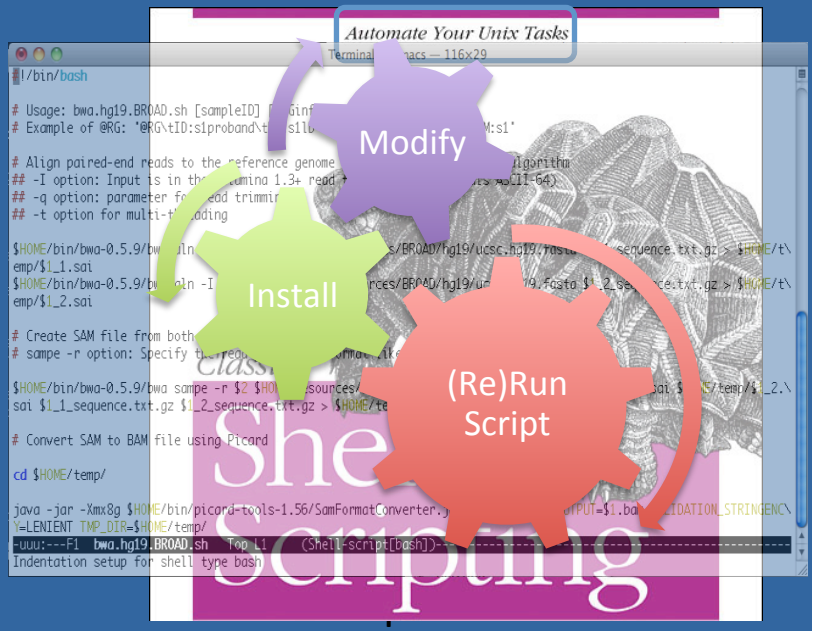
## Data Movement and Access Challenges




- Data is distributed in different locations
- Research labs need access to the data for analysis
- Be able to Share data with other researchers/collaborators
  - Inefficient ways of data movement
- Data needs to be available on the local and Distributed Compute Resources
  - Local Clusters, Cloud, Grid

Once we have the Sequence Data

- Manually move the data to the Compute node
- Install all the tools required for the Analysis
  - BWA, Picard, GATK, Filtering Scripts, etc.
- Shell scripts to sequentially execute the tools
- Manually modify the scripts for any change
  - Error Prone, difficult to keep track, messy..
- Difficult to maintain and transfer the knowledge




Manual Data Analysis

# Solutions for Biomedical analysis at scale

# Research Data Management as a Service

globus

Facility data acquisition

Reduced data

Globus transfer service

Globus data publication service

Globus sharing service

Analysis/ Sharing

# What is Globus?

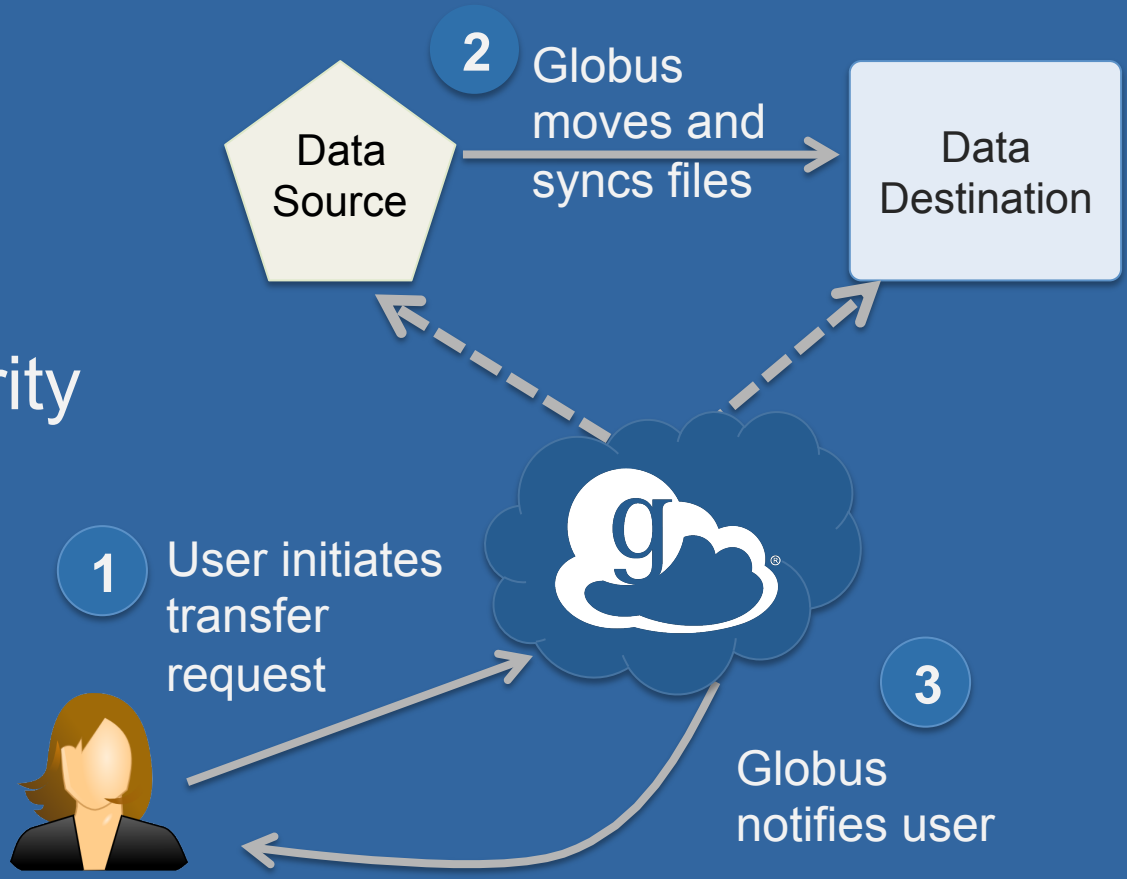
Big data transfer, sharing, publication and discovery…

…simply, securely, and fast…

…directly from your own storage systems

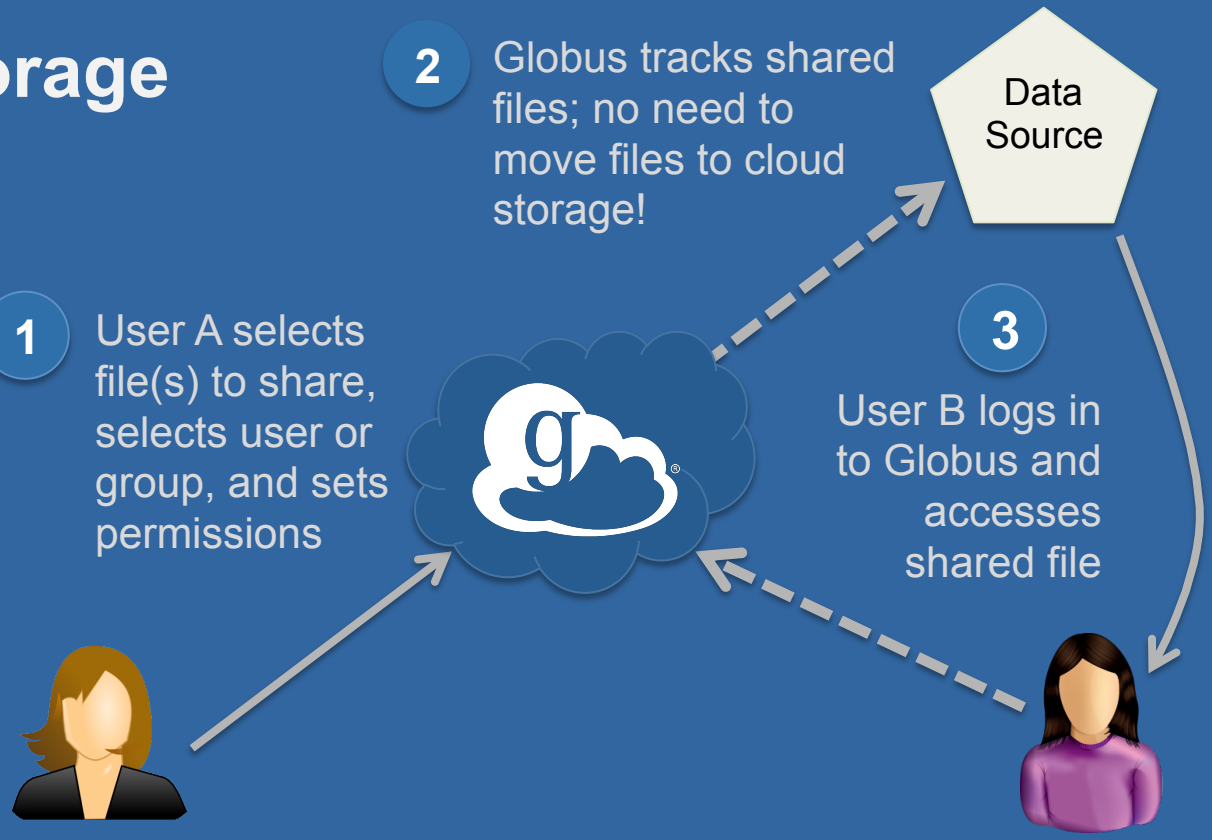
# Reliable, secure, high-performance ***file transfer*** and ***synchronization***

- "Fire-and-forget" transfers
- Automatic fault recovery
- Seamless security integration



# Simple, secure ***sharing*** off existing storage systems

- **Easily share large data with any user or group**
- **No cloud storage required**

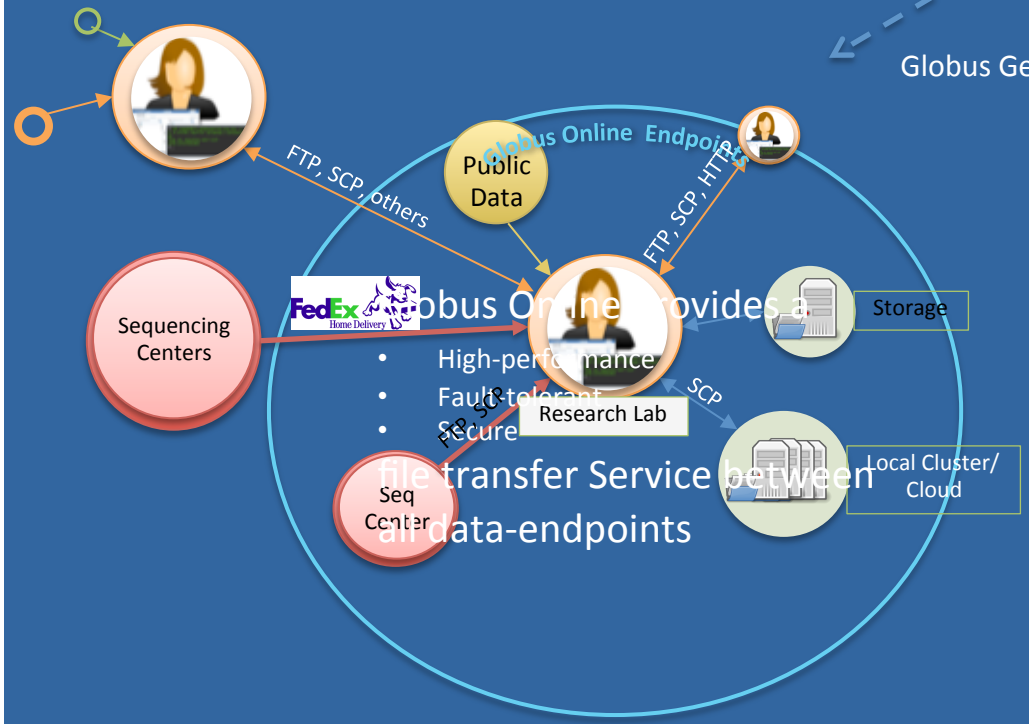


# Globus is SaaS

- Web, command line, and REST interfaces
- Reduced IT operational costs
- New features automatically available
- Consolidated support & troubleshooting
- Easy to add your laptop, server, cluster, supercomputer, etc. with Globus Connect


# Flexible, scalable, affordable genomics analysis for all biologists

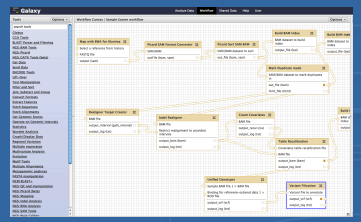
# Globus Genomics




Data Management

Globus Genomics

Galaxy Based Workflow Management System



Galaxy Data Libraries

- Globus Online Integrated within Galaxy
- Web-based UI
- Drag-Drop workflow creations
- Easily modify Workflows with new tools

Analytical tools are automatically run on the scalable compute resources when possible

Galaxy on Cluster/Cloud

Data Analysis

# Globus Genomics

- Workflows can be easily defined and automated with integrated Galaxy Platform capabilities
- Data movement is streamlined with integrated Globus file-transfer functionality
- Resources can be provisioned on-demand with Amazon Web Services cloud based infrastructure

# Agenda

12:30pm Challenges for biomedical analysis at scale

12:45pm Best practices and solution components

1:00pm Introduction to Globus Genomics, Globus Transfer

1:30pm Exercise: Transferring raw NGS datasets from sequencing centers and sharing

2:00pm Refreshment Break

2:15pm Demonstration of QC pipeline, Exome, RNA, Whole Genome analysis

2:30pm Exercise: Running example pipelines with sample data sets

2:45pm Running analysis at scale using Globus Genomics

3:15pm Exercise/Demonstration: Executing Exome, RNA, QC pipelines at scale

3:45pm Interactive Q&A and Session Wrap-up

# Exercise: Transferring raw NGS datasets from sequencing centers and sharing

# Exercise 1: Account Signup

1. Go to: globus.org/signup
2. Create your Globus account
3. Validate e-mail address
4. Optional: Login with your campus/InCommon identity

# Exercise 2: Transfer, Sharing, Group Management

1. Install Globus Connect Personal
2. Move file(s) from esnet#anl-diskpt1 to your laptop
3. Check your email for a notification on successful transfer
4. Go to globus.org/Groups and search for group named BioIT2015. Click Join the Group

# Exercise 3: Transferring data from a Sequencing center

1. Login to https://bioit.globusgenomics.org
2. Click on Browse and Get Data using Globus Online tool on the left hand panel
3. Start typing the name of the endpoint sulakhe#SequencingCenter
4. Log in to the endpoint with username genomics, password globus
5. Select the the forward Exome files under Exome-seq-sample data and click Execute
6. Repeat the above step for reverse file

# Break

# Resume at 2:15pm

# Agenda

12:30pm Challenges for biomedical analysis at scale

12:45pm Best practices and solution components

1:00pm Introduction to Globus Genomics, Globus Transfer

1:30pm Exercise: Transferring raw NGS datasets from sequencing centers and sharing

2:00pm Refreshment Break

2:15pm Demonstration of QC pipeline, Exome, RNA, Whole Genome analysis

2:30pm Exercise: Running example pipelines with sample data sets

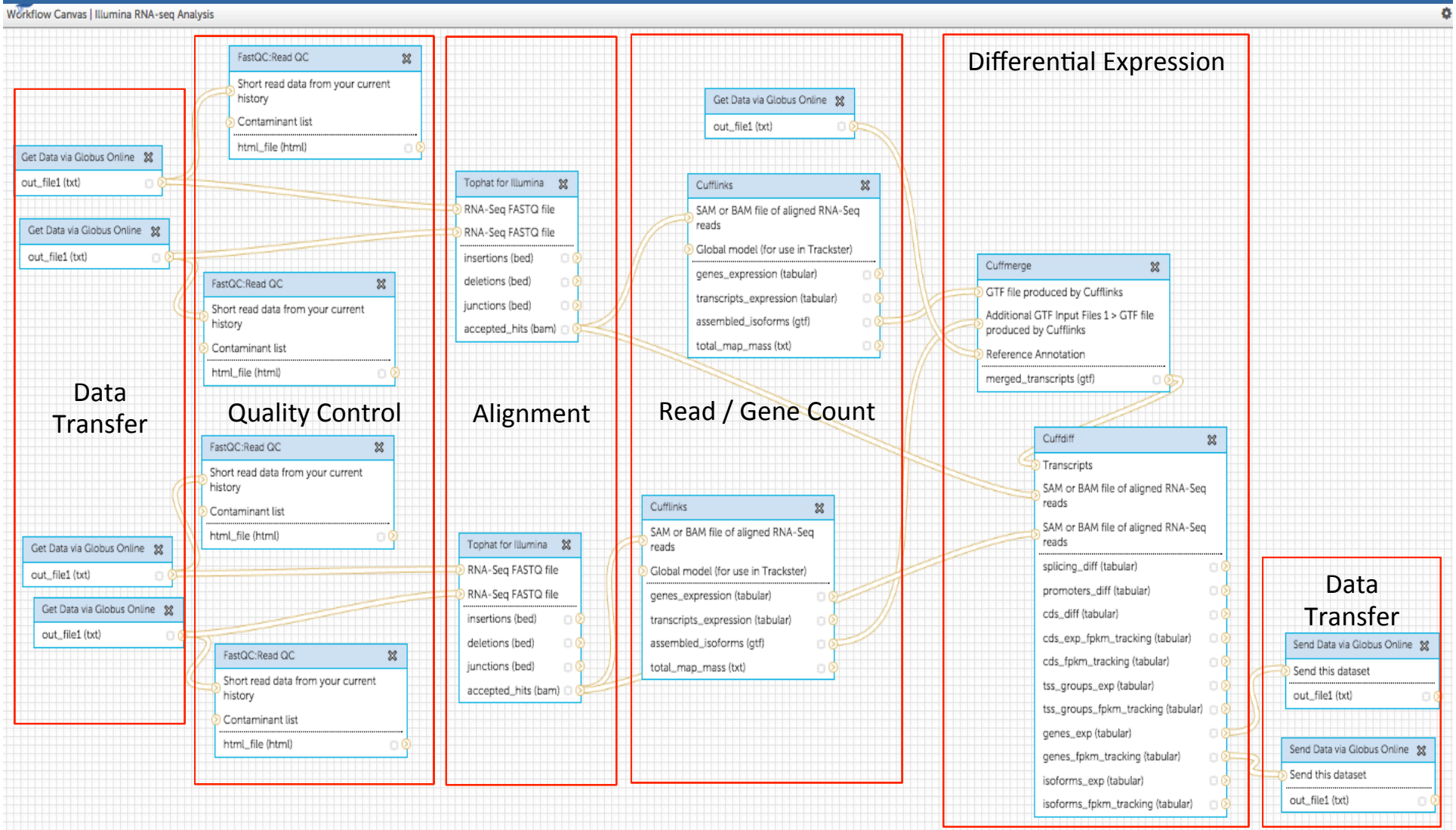
2:45pm Running analysis at scale using Globus Genomics

3:15pm Exercise/Demonstration: Executing Exome, RNA, QC pipelines at scale

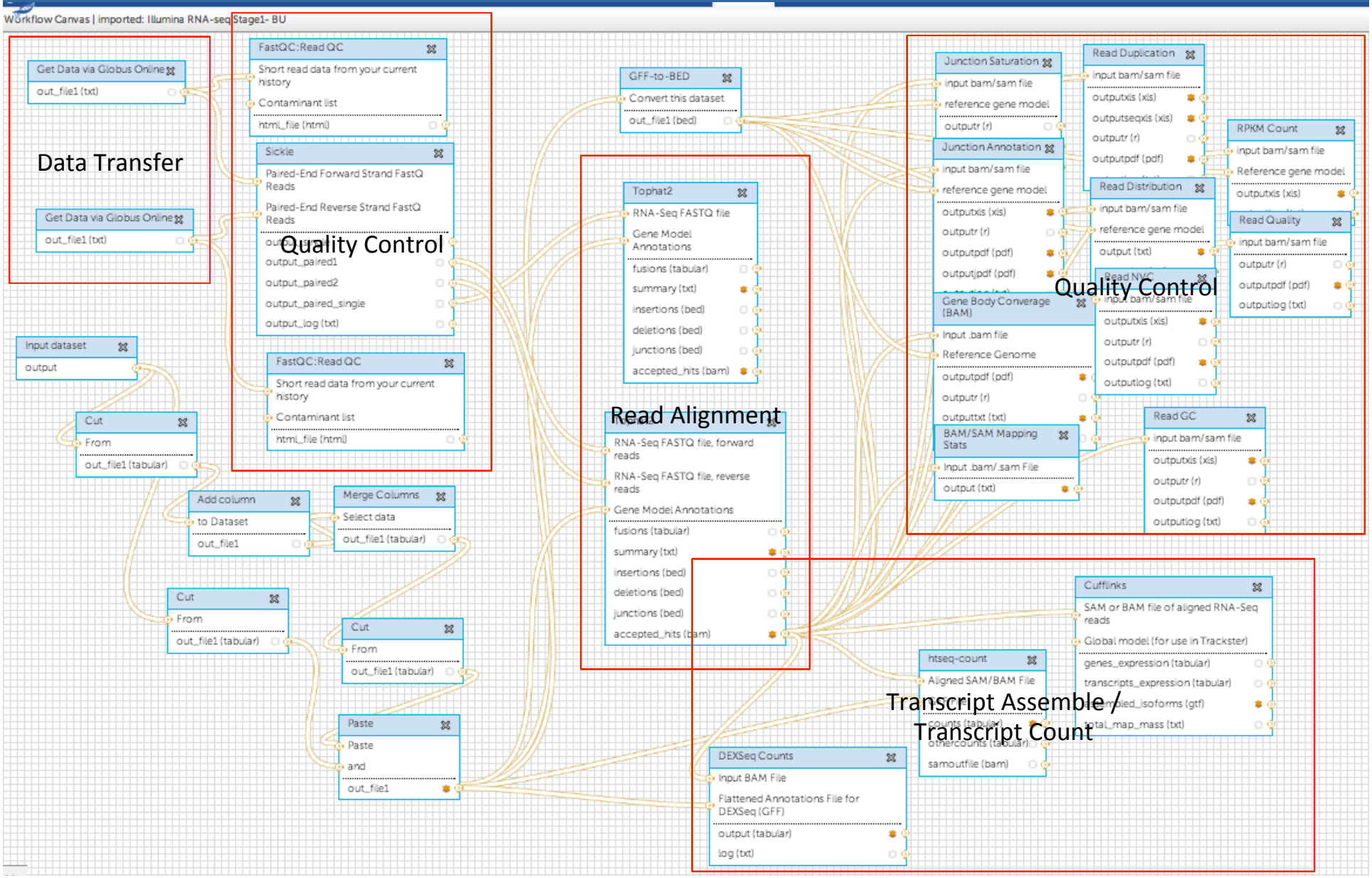
3:45pm Interactive Q&A and Session Wrap-up

# Demonstration of QC pipeline, Exome, RNA, Whole Genome analysis

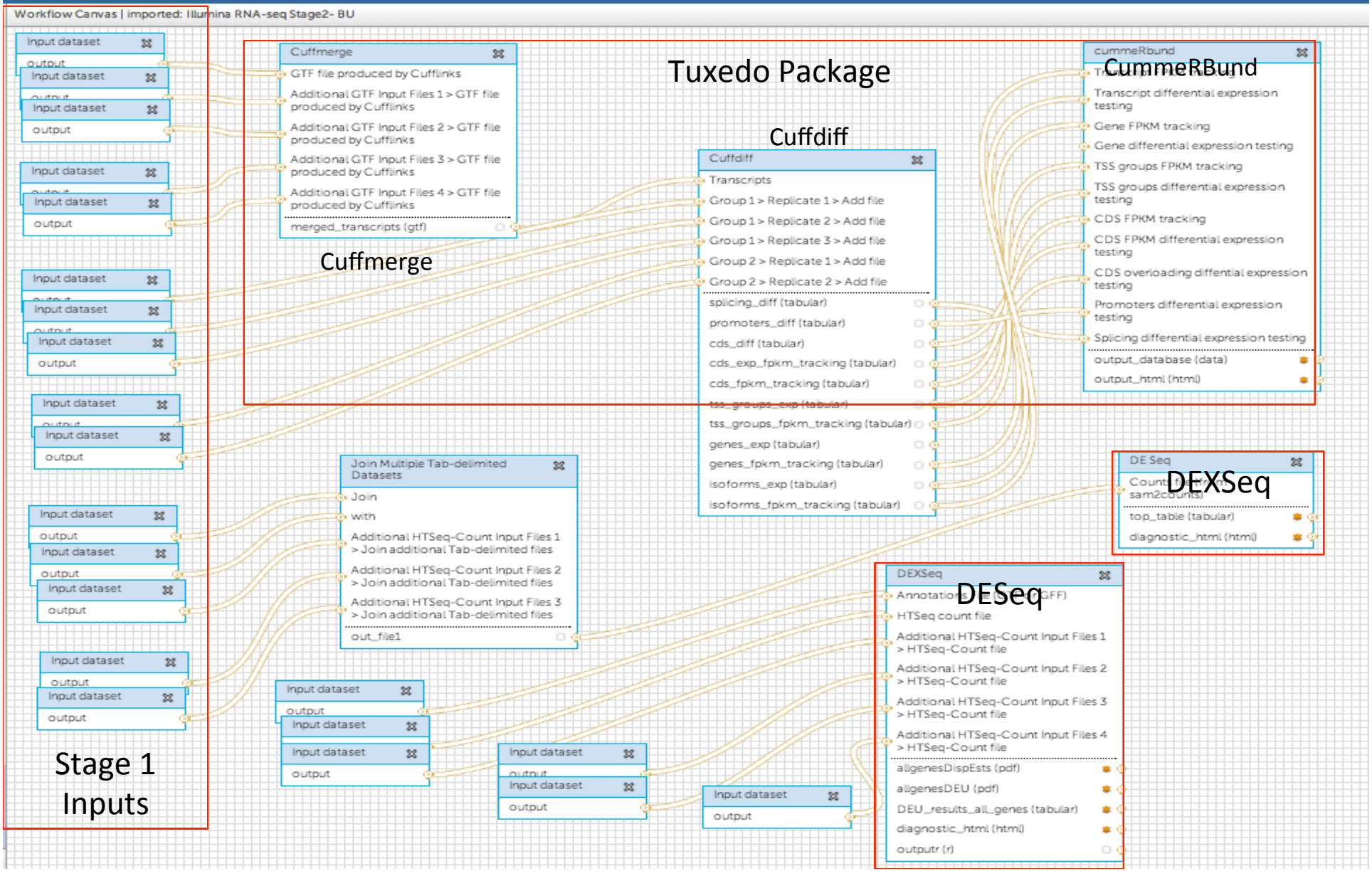
# RNA-Seq Analysis Workflow



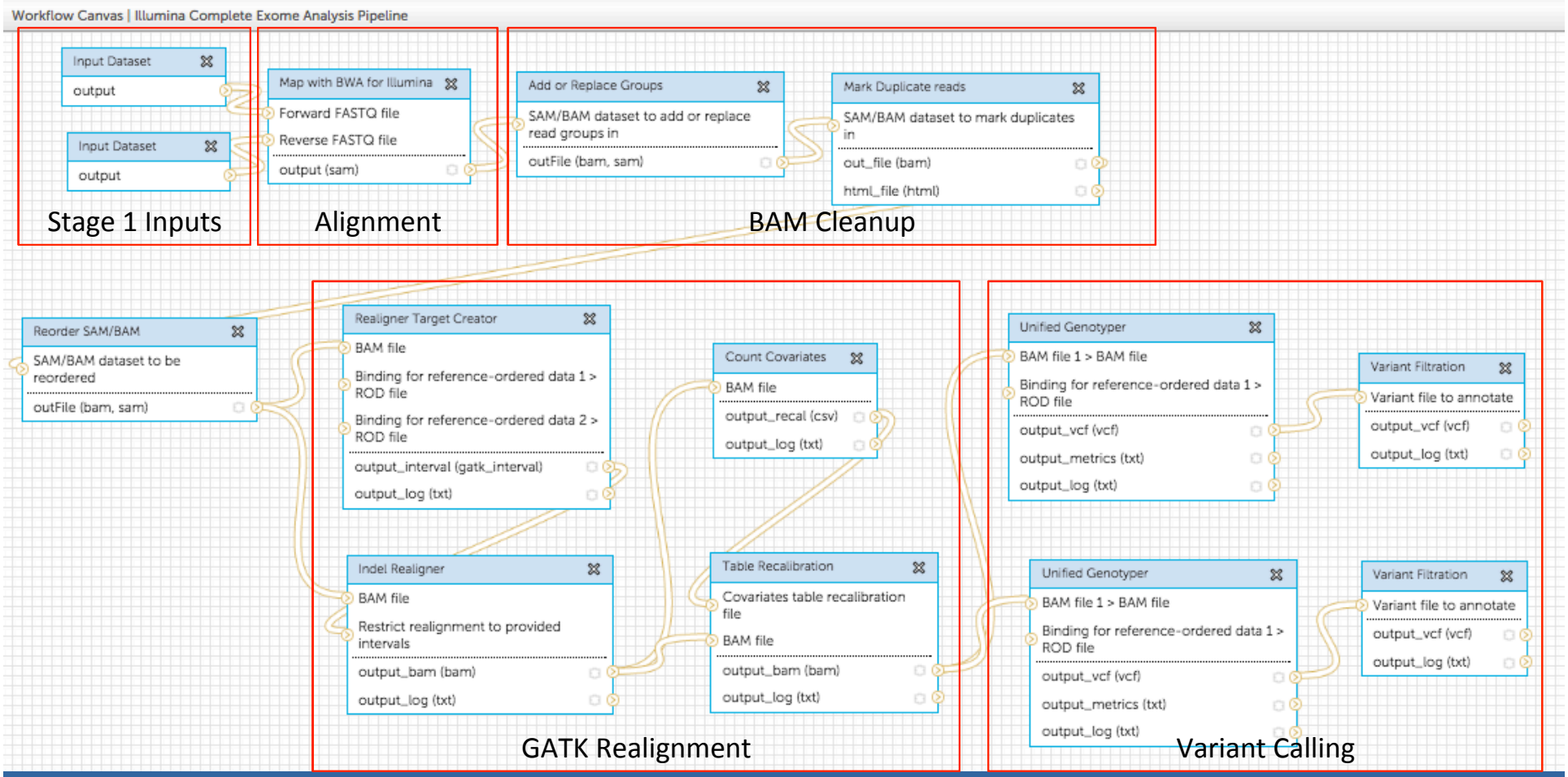
# RNA-Seq Analysis Workflow (Stage 1)



# RNA-Seq Analysis Workflow (Stage 2)



# Exome / Whole Genome Analysis



# Agenda

12:30pm Challenges for biomedical analysis at scale

12:45pm Best practices and solution components

1:00pm Introduction to Globus Genomics, Globus Transfer

1:30pm Exercise: Transferring raw NGS datasets from sequencing centers and sharing

2:00pm Refreshment Break

2:15pm Demonstration of QC pipeline, Exome, RNA, Whole Genome analysis

2:30pm Exercise: Running example pipelines with sample data sets

2:45pm Running analysis at scale using Globus Genomics

3:15pm Exercise/Demonstration: Executing Exome, RNA, QC pipelines at scale

3:45pm Interactive Q&A and Session Wrap-up

# Running example pipelines with sample data sets

# Exercise 4 : Exome Analysis Workflow

1. Login to https://bioit.globusgenomics.org
2. Copy “dbsnp” and “1000G” files from: “Shared Data -> Data Libraries -> Reference Data Library” (into same history)
3. On the Main page, click on the “Workflow for Illumina Exome-Seq” and “import workflow”
4. Run Workflow: Click Workflow tab and select the imported Exome workflow and click “Run”
5. Input Parameters: Select appropriate input files (Forward & Reverse as well as reference files for each step.
6. Click “Run Workflow” button.

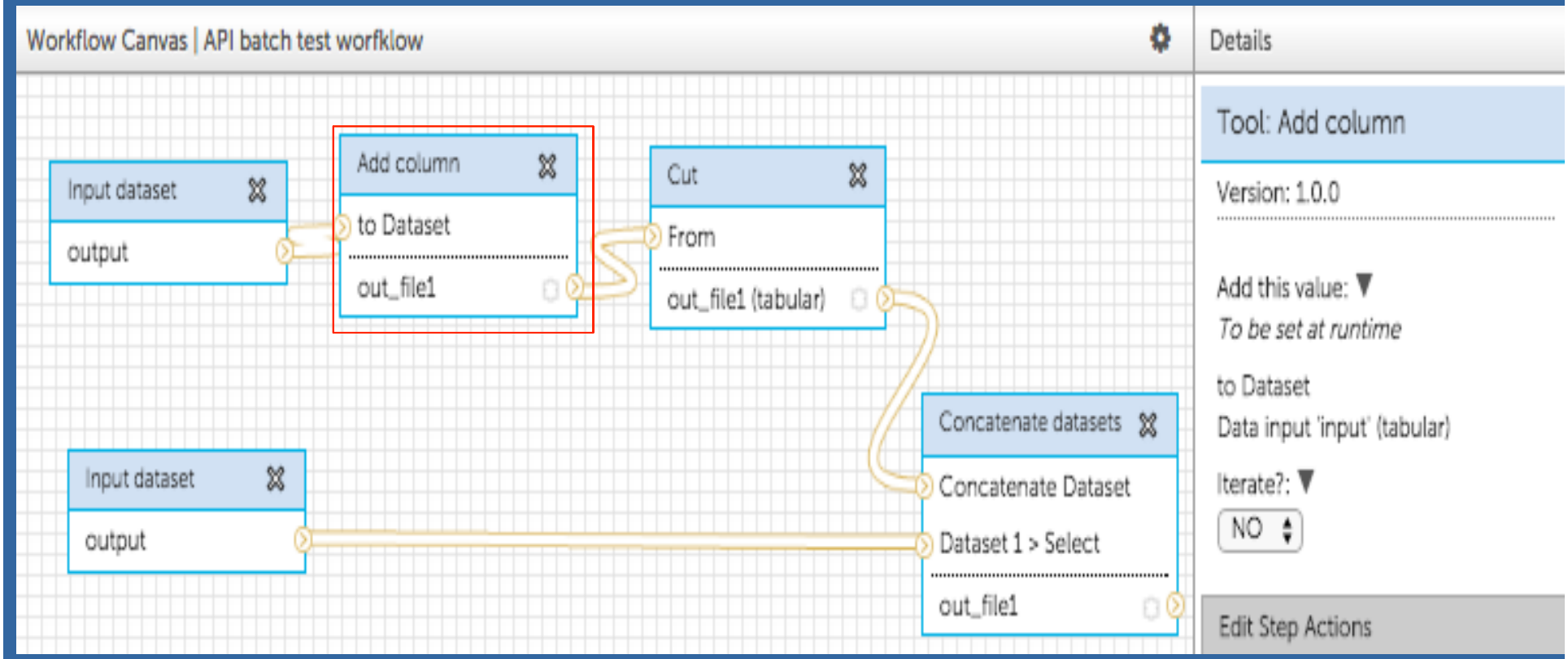
# Exercise 5 : Exome Analysis Workflow

- Try the Exome Workflow with Transfers jobs as inputs (Under Published Workflows)

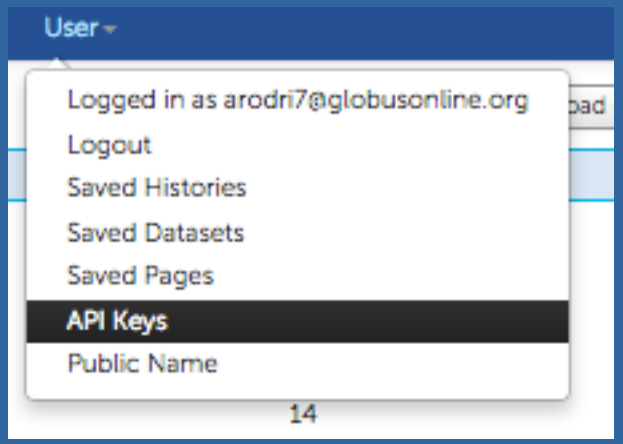
# Running Batch Job

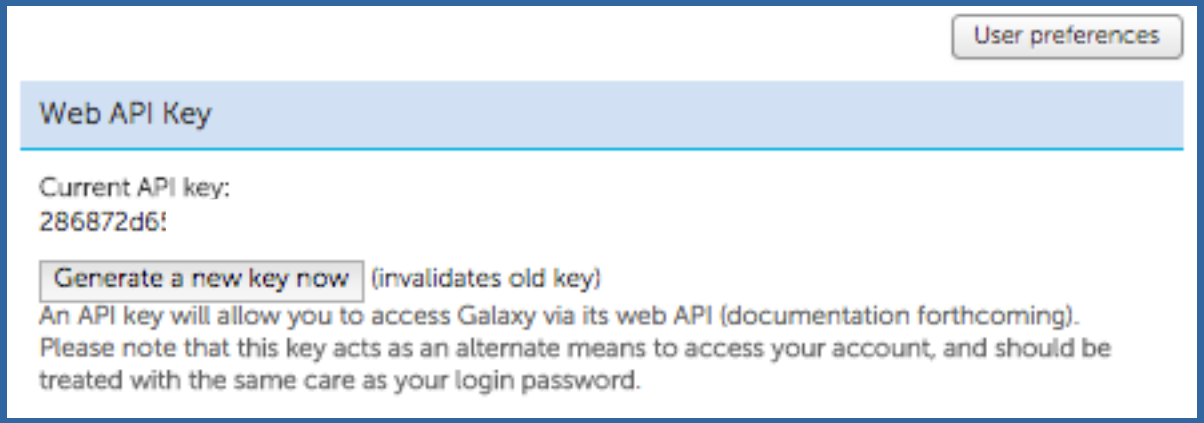
- Create workflow
- Generate user API Key (if necessary)
- Download and fill out workflow table file
- Upload table file
- Submit

# Running Batch Job (Workflow)

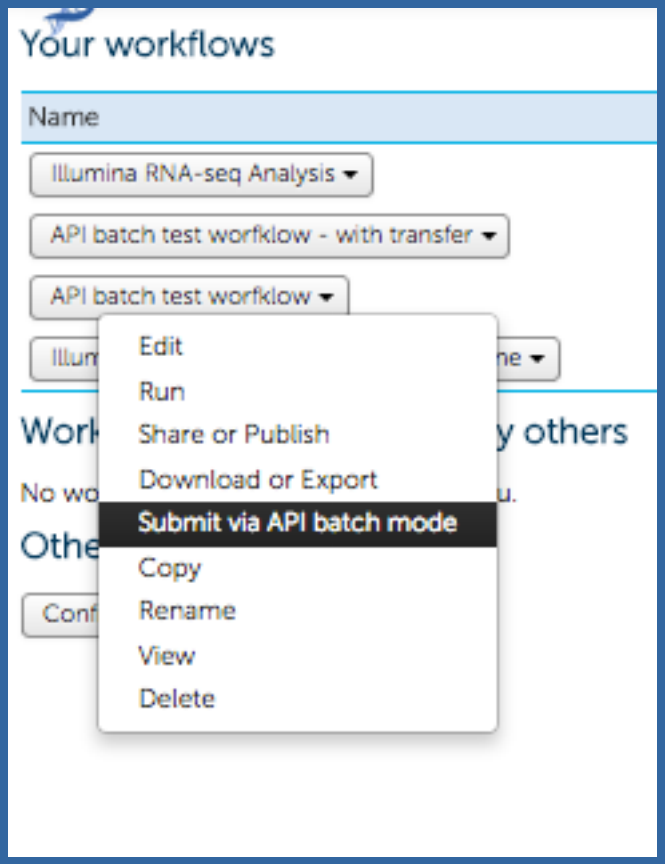


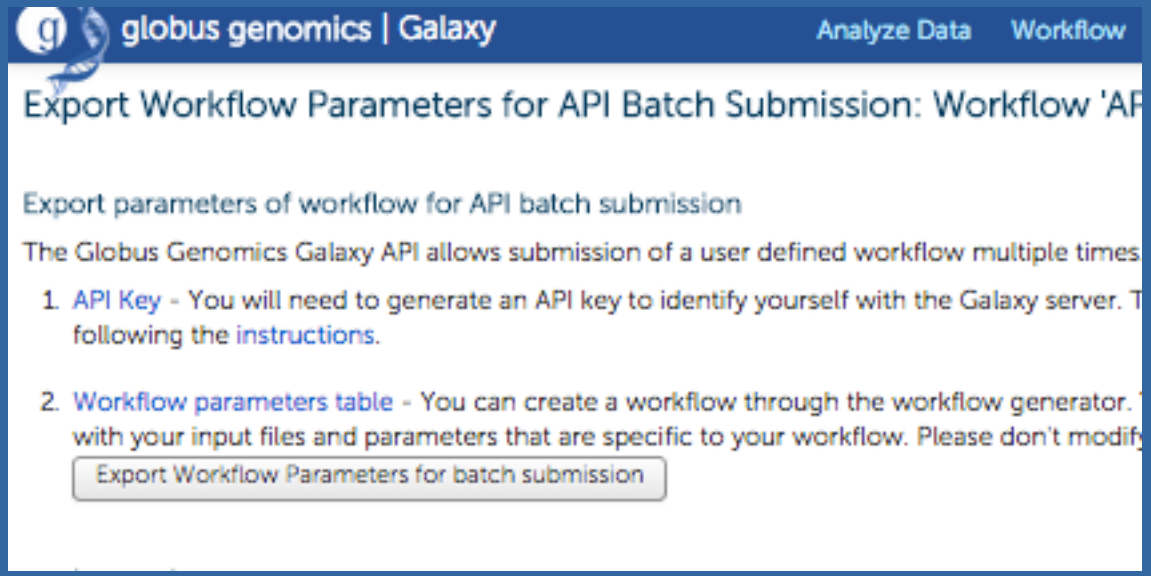
# Running Batch Job (API Key)



# Running Batch Job (Download)



# Running Batch Job (Fill out)

Galaxy-API-Workflow-API_batch_test_worfklow (1).txt

```
#Data Export for Workflow Batch Submission Through the APII


### INSTRUCTIONS
########################################
#The following data can be used to input the parameters you have previously determined to be
#set at runtime. Please specify the library or history where the input data can be found.
#Once you have filled out the table you can run the API script to submit the jobs through Galaxy
#via the API.

#NOTE: If you make any changes to the workflow or edit the name of the workflow, you will need
#to recreate the table before submitting the job via the API since some metadata parameters will
#be modified.

#NOTE: It is up to the user to make sure the input files are in the correct format for each
#parameter being filled out.

#NOTE: You will need to specify three items for input files to an application.
#The format for an input file should be [SourceType:SourceName:file_name]:
#1. Source Type - which can be library or history
#2. Source Name - the name of the library or history.
#3. Filename - specify the name of the file as it exists in the library or history.


########################################


### METADATA
########################################
Workflow Name  API batch test worfklow
Workflow id    ef5b0cd28aaeba40
Project Name   <Your_project_name>
########################################


###TABLE DATA
########################################
SampleName      SourceType:SourceName:Ref1     SourceType:SourceName:Ref2     Param:addValue:exp
TestSample1      library:API Test Library:Tabular1.bed    library:API Test Library:Tabular2.bed    expressionToAdd1
TestSample2      library:API Test Library:Tabular1.bed    library:API Test Library:Tabular2.bed    Second submission
TestSample3      library:API Test Library:Tabular1.bed    library:API Test Library:Tabular2.bed    Just for kicks
```

# Running Batch Job (Upload)

```
#Data Export for Workflow Batch Submission Through the APII


### INSTRUCTIONS
#######################################
#The following data can be used to input the parameters you have previously determined to be
#set at runtime. Please specify the library or history where the input data can be found.
#Once you have filled out the table you can run the API script to submit the jobs through Galaxy
#via the API.

#NOTE: If you make any changes to the workflow or edit the name of the workflow, you will need
#to recreate the table before submitting the job via the API since some metadata parameters will
#be modified.

#NOTE: It is up to the user to make sure the input files are in the correct format for each
#parameter being filled out.

#NOTE: You will need to specify three items for input files to an application.
#The format for an input file should be [SourceType::SourceName::file_name]:
#1. Source Type - which can be library or history
#2. Source Name - the name of the library or history.
#3. Filename - specify the name of the file as it exists in the library or history.


########################################


### METADATA
#######################################
Workflow Name    API batch test worfklow
Workflow id      b21250b21b11a056
Project Name     <Your_project_name>
#######################################


###TABLE DATA
#######################################
SampleName       SourceType::SourceName::Ref1    SourceType::SourceName::Ref2    Param::262::addValue::exp
Run1    library::API TEST LIBRARY::Tabular1.bed library::API TEST LIBRARY::Tabular2.bed Value1
Run2    library::API TEST LIBRARY::Tabular1.bed library::API TEST LIBRARY::Tabular2.bed Value2
Run3    library::API TEST LIBRARY::Tabular1.bed library::API TEST LIBRARY::Tabular2.bed Value3
```

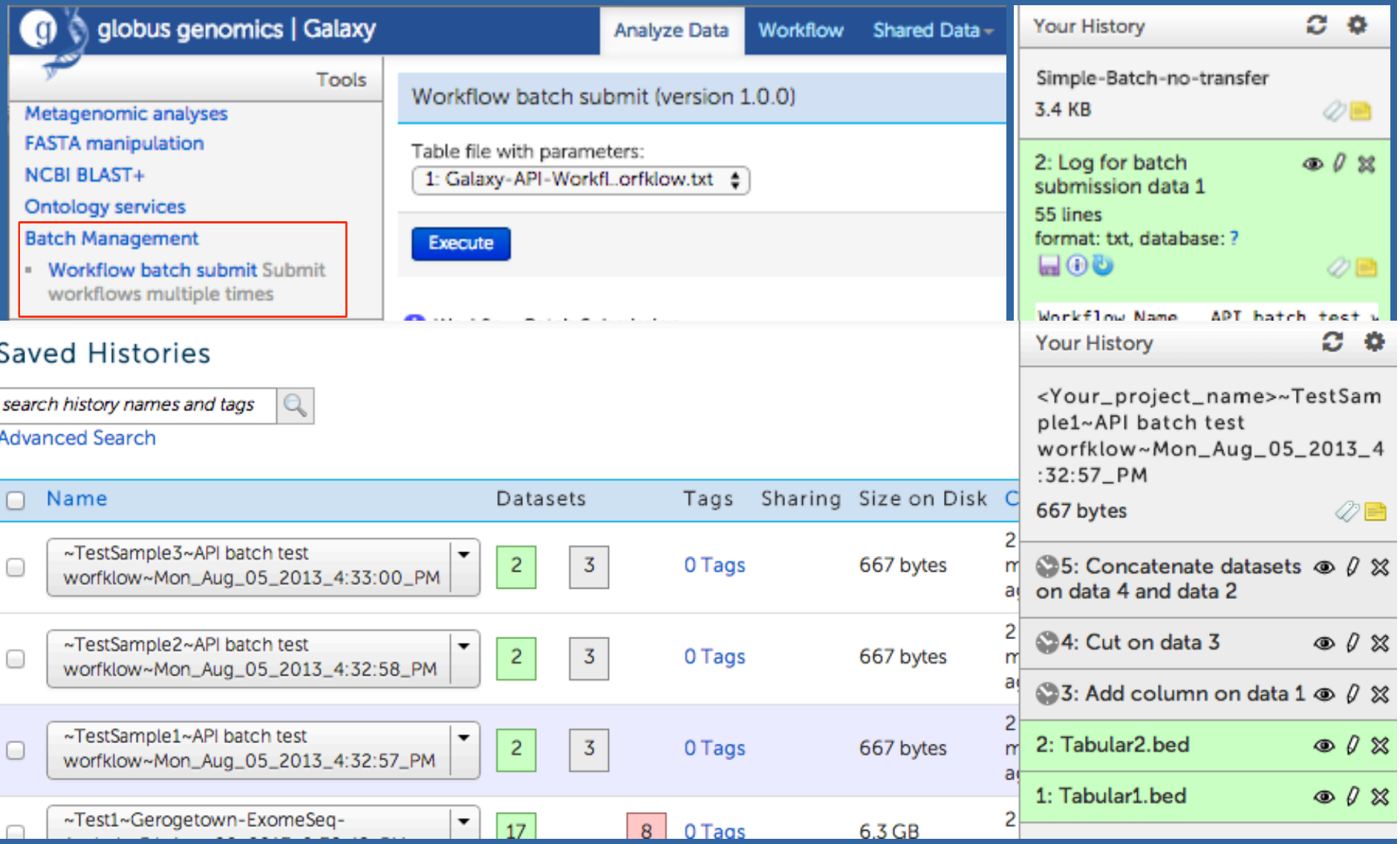
Your History

Simple-Batch-no-transfer
3.4 KB

1: Galaxy-API-Workflow-API_batch_test_worfklow.txt
7 lines
format: txt, database: ?
Globus transfer summary: From: arodri7#ci-arodri-laptop To: Local galaxy instance.

```
#Data Export for Workflow Batch
### INSTRUCTIONS
#################################
#The following data can be used
```

# Running Batch Job (Submit)



# Example Collaborations

## Dobyns Lab

UNIVERSITY OF WASHINGTON

**Backround**: Investigate the nature and causes of a wide range of human developmental brain disorders

**Approach**: Replaced manual analysis with Globus Genomics

**Results**: Achieved greater than 20X speed-up in analysis of exome data

**Future Plans**: Leverage scale-out capability of Globus Genomics on 150 exome data set and seek to achieve 50X speed-up in analysis

# Example Collaborations

## Georgetown Medical Center

**Backround**: Innovation Center for Biomedical Informatics is an academic hub for innovative research in the field of biomedical informatics.

**Approach**: Augment current team and tools with a NGS analysis platform to support standard and best-practice pipelines while leveraging elastic cloud-based resources.

**Results**: Pilot effort is complete – improved quality and performance results on whole genome, exome and RNA-Seq pipelines utilizing Globus Genomics

**Future Plans**: Provide Globus Genomics as a well-managed platform-as-a-service for ICBI collaborators and users

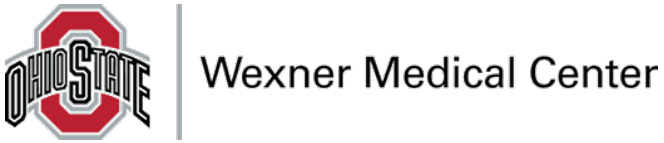
# Diversity of Collaborations







Cox Lab
Volchenboum Lab
Olopade Lab




GEORGETOWN UNIVERSITY






# Typical Engagement

## Proof of Concept

- Limited scale
- Existing or slightly modified pipeline
- Setup
- Training
- Testing

## Pilot

- Multi-endpoint
- Additional tools
- Pipeline validation
- Scale-out analysis
- Optimization
- Training
- Staged transition to production

## Production

- Monthly, annual subscription
- Startup help
- Training
- Support

# Wrapping up…

We are a non-profit service provider to various research communities

We are a non-profit service provider to various research communities

We offer multiple subscription tiers to provide a cost-effective solution and ensure sustainability of our service

# Subscription Pricing

| | Starter | Standard | Large |
|---|---|---|---|
| **Cumulative Analysis Workload* (over a 12-month subscription)** | ~ 800 exomes \| ~80 whole genomes \| ~ 400 RNA-seqs | ~ 4000 exomes \| ~ 400 whole genomes \| ~ 2000 RNA-seqs | ~ 20000 exomes \| ~ 2000 whole genomes \| ~ 10000 RNA-seqs |
| **Technical Support** | M-F, 9-5 CT 2-business day response | M-F, 9-5 CT, 1-business day response | M-F, 9-5 CT 1-business day response |
| **Access to Enhanced Workbench** | Yes | Yes | Yes |
| **Multi-sample submission** | Yes | Yes | Yes |
| **Usage Dashboard** | Yes | Yes | Yes |
| **Price/Performance Controls** | Basic | Advanced | Advanced |
| **On-Demand Tool Wrapping** | No | Limited | Yes |
| **HIPAA / optional BAA** | Not Available | Available | Available |

*Annual subscriptions start at $5,000 for individual PIs and $10,000 for core labs*

* *Representative workloads based on human genome, GATK variant calling pipeline (whole genome, exome), Tuxedo suite of tools (RNA-Seq), etc.*

# Thank you to our sponsors!


U.S. DEPARTMENT OF
**ENERGY**



THE UNIVERSITY OF
CHICAGO

Argonne
NATIONAL LABORATORY