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)
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
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!
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

Manual Data Analysis

- 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
Solutions for Biomedical analysis at scale
Research Data Management as a Service

Facility data acquisition

Globus transfer service

Globus sharing service

Globus data publication service

Reduced data

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

1. User initiates transfer request
2. Globus moves and syncs files
3. Globus notifies user

Data Source

Data Destination
Simple, secure *sharing* off existing storage systems

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

1. User A selects file(s) to share, selects user or group, and sets permissions
2. Globus tracks shared files; no need to move files to cloud storage!
3. User B logs in to Globus and accesses shared file

• 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

Globus Online provides a file transfer service between all data-endpoints.

- High-performance
- Fault-tolerant
- Secure

Data Management

Data Analysis

Galaxy Based Workflow Management System

- 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

Reference: Galaxy on Cluster/Cloud
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
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

Data Transfer

Quality Control

Alignment

Read / Gene Count

Differential Expression

Data Transfer
RNA-Seq Analysis Workflow (Stage 1)

Data Transfer:
- Get Data via Globus Online
- FastQC: Read QC

Quality Control:
- Sickle
- Paired-End Forward Strand FastQC
- Paired-End Reverse Strand FastQC

Read Alignment:
- FastQC: Read QC
- Short read data from your current history
- Contaminant list

Transcript Assembly:
- Transcript Assembly

Transcript Count:
- Transcript Count

Quality Control:
- Junction Saturation
- Read Duplication
- Read GC
- Gene Body Coverage
- Read Quality
- bam2bed
- bam2bed.py

Workflow Curves | Imported: illumina RNA-Seq Stage 1 - EU
RNA-Seq Analysis Workflow (Stage 2)

Stage 1 Inputs

Tuxedo Package
- Cuffmerge
- Cuffdiff

Stage 2 Outputs
- DEXSeq
- CummeRBund

Inputs:
- Input dataset
- Output
Exome / Whole Genome Analysis

Workflow Canvas | Illumina Complete Exome Analysis Pipeline

Stage 1 Inputs
- Input Dataset
  - output

Alignment
- Map with BWA for Illumina
  - Forward FASTQ file
  - Reverse FASTQ file
  - output (sam)

BAM Cleanup
- Add or Replace Groups
  - SAM/BAM dataset to add or replace read groups in
  - output (bam, sam)

- Mark Duplicates
  - SAM/BAM dataset to mark duplicates in
  - output (bam, sam)

GATK Realignment
- Realigner Target Creator
  - SAM/BAM dataset to be realigned
  - output (bam, sam)

- Indel Realigner
  - BAM file
  - Restrict realignment to provided intervals
  - output (bam)

- Table Recalibration
  - Converts table recalibration file
  - output (bam)

Variant Calling
- Unified Genotyper
  - BAM file 1 > BAM file
  - Binding for reference-ordered data 1 > ROD file
  - output (vcf)

- Variant filtration
  - Variant file to annotate
  - output (vcf)

Inputs
- Variant Calling
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)

- Current API key: 286872d6!
- Generate a new key now (invalidates old key)

An API key will allow you to access Galaxy via its web API (documentation forthcoming). Please note that this key acts as an alternate means to access your account, and should be treated with the same care as your login password.
Running Batch Job (Download)

Your workflows

Illumina RNA-seq Analysis
API batch test workflow - with transfer
API batch test workflow

Edit
Run
Share or Publish
Download or Export
Submit via API batch mode
Copy
Rename
View
Delete

Export Workflow Parameters for API Batch Submission: Workflow 'API'

Export parameters of workflow for API batch submission

1. API Key - You will need to generate an API key to identify yourself with the Galaxy server. Please follow the instructions.

2. Workflow parameters table - You can create a workflow through the workflow generator. To do so, you will need your input files and parameters that are specific to your workflow. Please don't modify the parameter names.

Export Workflow Parameters for batch submission
Running Batch Job (Fill out)

# Data Export for Workflow Batch Submission Through the API

### 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 workflow
Workflow id    ef3b0c28aebea40
Project Name   <Your_project_name>

### TABLE DATA

<table>
<thead>
<tr>
<th>SampleName</th>
<th>SourceType:SourceName:Ref1</th>
<th>SourceType:SourceName:Ref2</th>
<th>Param:defaultValue:exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>TestSample1</td>
<td>library:API Test Library:Tabular1.bed</td>
<td>library:API Test Library:Tabular2.bed</td>
<td>expressionToAdd1</td>
</tr>
<tr>
<td>TestSample2</td>
<td>library:API Test Library:Tabular1.bed</td>
<td>library:API Test Library:Tabular2.bed</td>
<td>Second submission</td>
</tr>
<tr>
<td>TestSample3</td>
<td>library:API Test Library:Tabular1.bed</td>
<td>library:API Test Library:Tabular2.bed</td>
<td>Just for kicks</td>
</tr>
</tbody>
</table>
Running Batch Job (Upload)

# Data Export for Workflow Batch Submission Through the API

### 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.

- **Source Type** - which can be library or history
- **Source Name** - the name of the library or history
- **Filename** - specify the name of the file as it exists in the library or history.

### METADATA

- **Workflow Name**: API batch test workflow
- **Workflow id**: b21250b21b11a56
- **Project Name**: <Your_project_name>

### TABLE DATA

<table>
<thead>
<tr>
<th>SampleName</th>
<th>SourceName</th>
<th>SourceType</th>
<th>Ref1</th>
<th>SourceName</th>
<th>SourceType</th>
<th>Ref2</th>
<th>Param1</th>
<th>Param2</th>
<th>Param3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run1</td>
<td>library:API TEST LIBRARY::Tabular1.bed</td>
<td>library:API TEST LIBRARY::Tabular2.bed</td>
<td>Value1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run2</td>
<td>library:API TEST LIBRARY::Tabular1.bed</td>
<td>library:API TEST LIBRARY::Tabular2.bed</td>
<td>Value2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run3</td>
<td>library:API TEST LIBRARY::Tabular1.bed</td>
<td>library:API TEST LIBRARY::Tabular2.bed</td>
<td>Value3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Running Batch Job (Submit)
Example Collaborations

Dobyns Lab

**Background**: 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

Background: 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
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

<table>
<thead>
<tr>
<th>Proof of Concept</th>
<th>Pilot</th>
<th>Production</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–2 weeks</td>
<td>1–3 months</td>
<td>Ongoing</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>Starter</th>
<th>Standard</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cumulative Analysis Workload</strong> (over a 12-month subscription)</td>
<td>~ 800 exomes</td>
<td>~80 whole genomes</td>
<td>~ 400 RNA-seqs</td>
</tr>
<tr>
<td><strong>Technical Support</strong></td>
<td>M-F, 9-5 CT 2-business day response</td>
<td>M-F, 9-5 CT, 1-business day response</td>
<td>M-F, 9-5 CT 1-business day response</td>
</tr>
<tr>
<td><strong>Access to Enhanced Workbench</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Multi-sample submission</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Usage Dashboard</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Price/Performance Controls</strong></td>
<td>Basic</td>
<td>Advanced</td>
<td>Advanced</td>
</tr>
<tr>
<td><strong>On-Demand Tool Wrapping</strong></td>
<td>Not Available</td>
<td>Limited</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>HIPAA / optional BAA</strong></td>
<td>Not Available</td>
<td>Available</td>
<td>Available</td>
</tr>
</tbody>
</table>

*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!