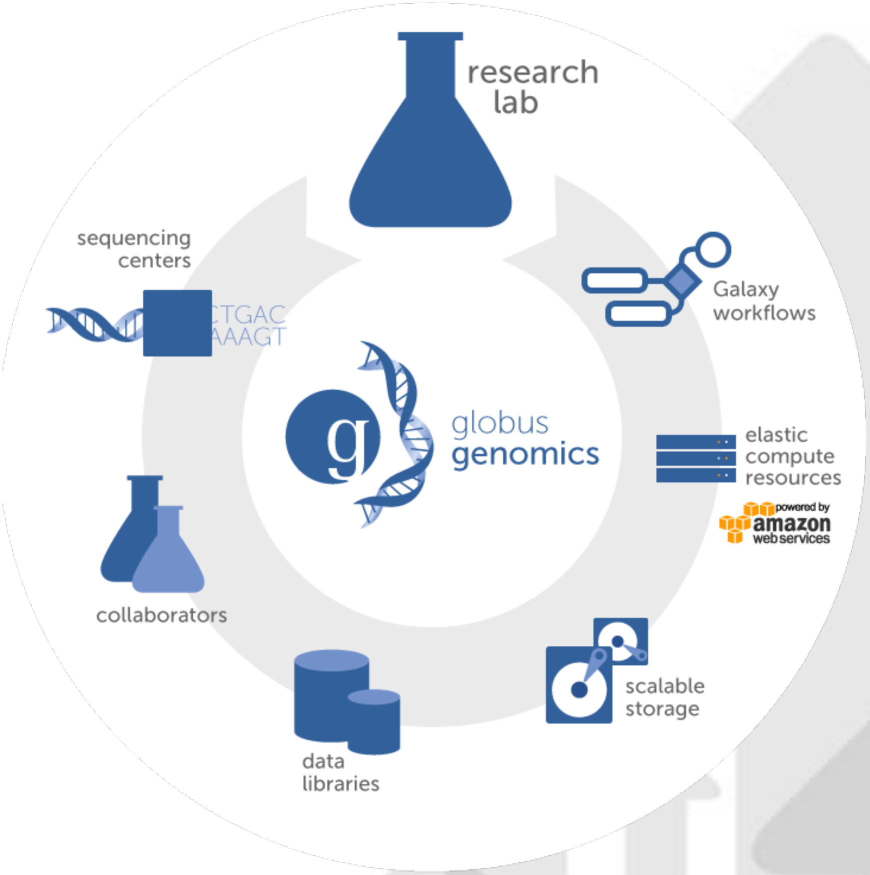
Globus Genomics Tutorial – GlobusWorld 2014

Agenda

- Overview of Globus Genomics
- Example Collaborations
- Demonstration
  - Globus Genomics interface
  - Globus Online integration
  - Scenario 1: Using Globus Genomics for Bioinformatics Core
  - Scenario 2: Using Globus Genomics for Individual Research labs
- Hands-On Experience

What Is Globus Genomics?

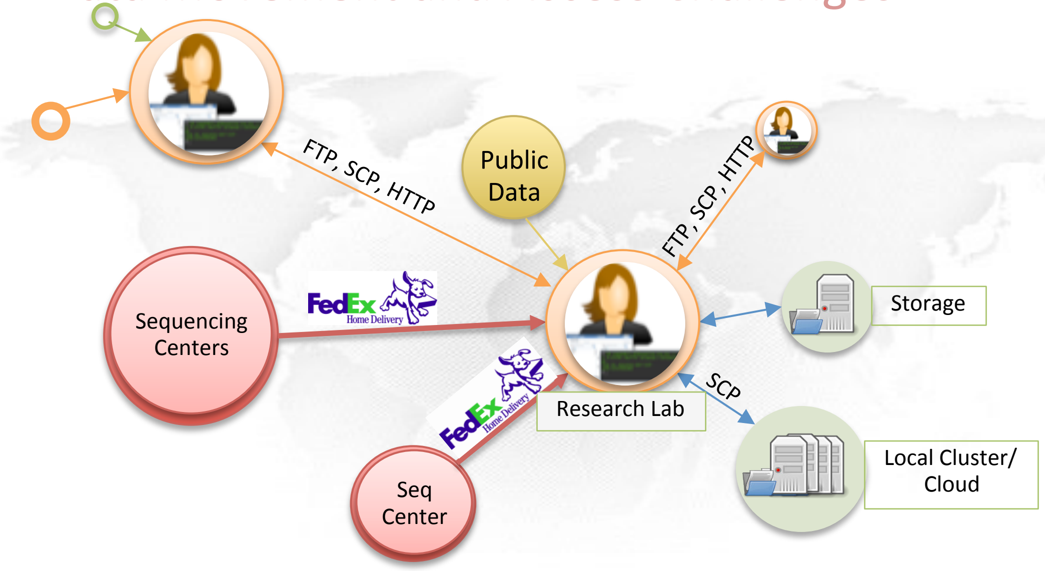
- Flexible, powerful SaaS-based genomics analysis platform
- Workflows can be easily defined and automated with integrated Galaxy capabilities
- Data movement is streamlined with integrated Globus file-transfer functionality
- Resources can be provisioned on-demand with Amazon Web Services cloud based infrastructure



www.globus.org/genomics

Challenges in Sequencing Analysis

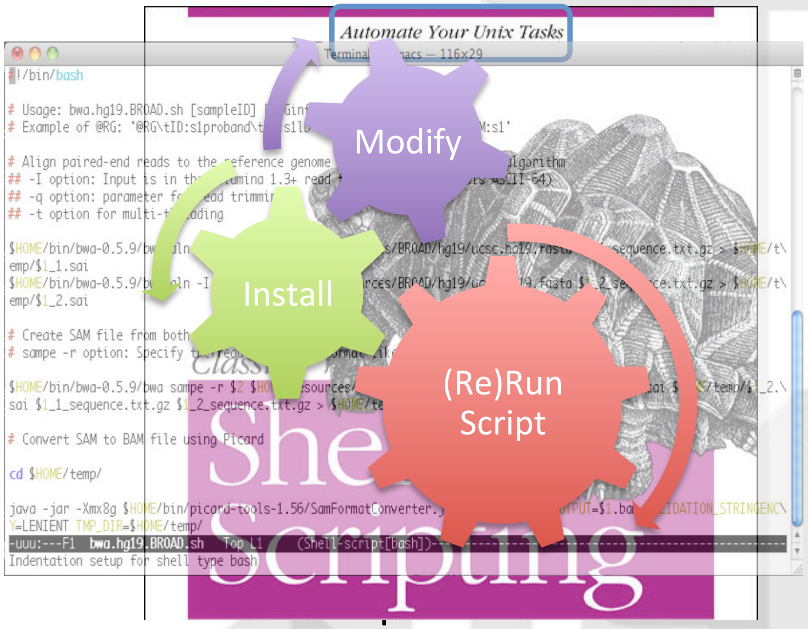
Data Movement and Access Challenges



- 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

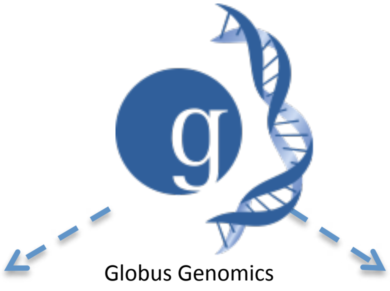


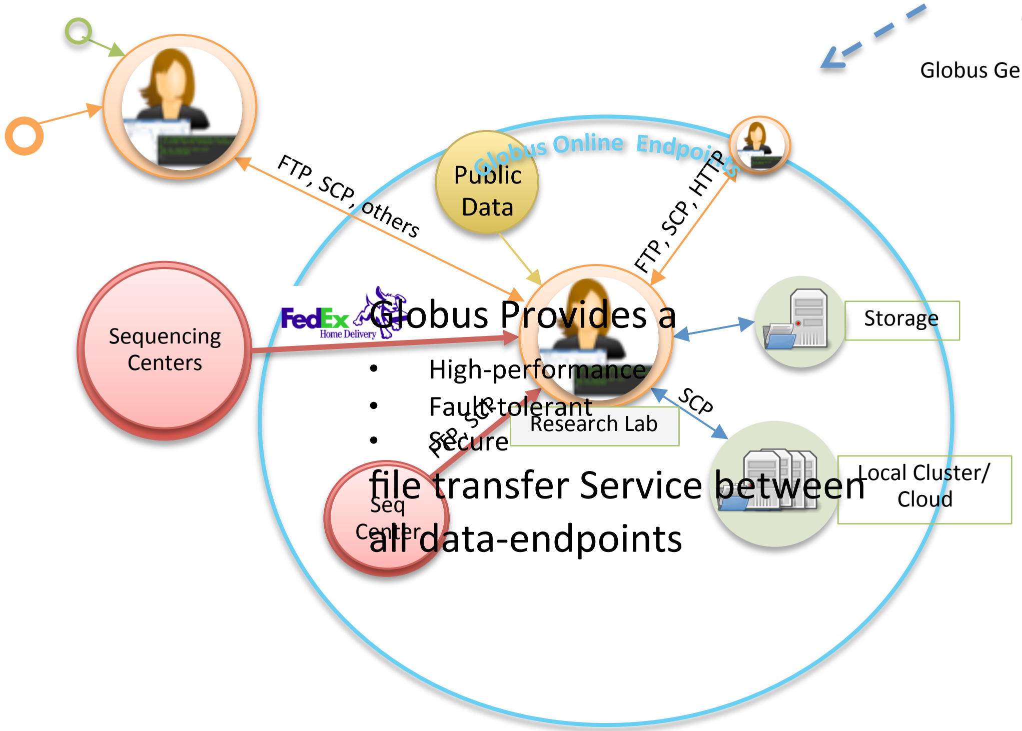
```
#!/bin/bash

# Usage: bwa.hg19.BROAD.sh [sampleID] [ Ginf
# Example of @RG: '@RG\tID:s1proband\t s1lb M:s1'

# Align paired-end reads to the reference genome lgorithm
## -I option: Input is in th umina 1.3+ read ASCII-64)
## -q option: parameter fo ead trimmi
## -t option for multi-t ading

$HOME/bin/bwa-0.5.9/bw aln s/BROAD/hg19/ucsc.hg19.fasta sequence.txt.gz > $HOME/t\
emp/$1_1.sai
$HOME/bin/bwa-0.5.9/bw aln -I rces/BROAD/hg19/uc 19.fasta $1_2_se ce.txt.gz > $HOME/t\
emp/$1_2.sai

# Create SAM file from both
# sampe -r option: Specify the requ ormat like

$HOME/bin/bwa-0.5.9/bwa sampe -r $2 $HO sources sai $ E/temp/$1_2.\
sai $1_1_sequence.txt.gz $1_2_sequence.txt.gz > $HOME/te

# Convert SAM to BAM file using Picard

cd $HOME/temp/

java -jar -Xmx8g $HOME/bin/picard-tools-1.56/SamFormatConverter.j UTPUT=$1.ba IDATION_STRINGENC\
Y=LENIENT TMP_DIR=$HOME/temp/
-uuu:---F1  bwa.hg19.BROAD.sh   Top L1    (Shell-script[bash])-------------------------------
Indentation setup for shell type bash
```

Manual Data Analysis

www.globus.org/genomics

Globus Genomics





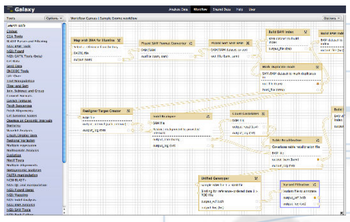
Globus Provides a

- High-performance
- Fault-tolerant
- Secure

file transfer Service between all data-endpoints

Galaxy Data Libraries

Galaxy Based Workflow Management System



- Globus 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

Data Management

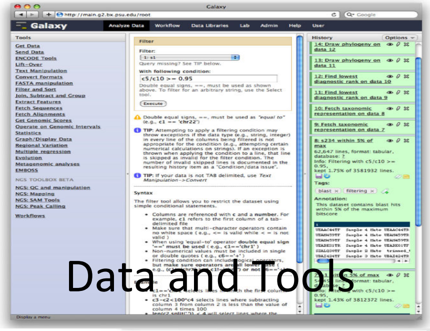
Data Analysis

www.globus.org/genomics

Globus integrated with Galaxy – A flexible, scalable, simplified analysis platform

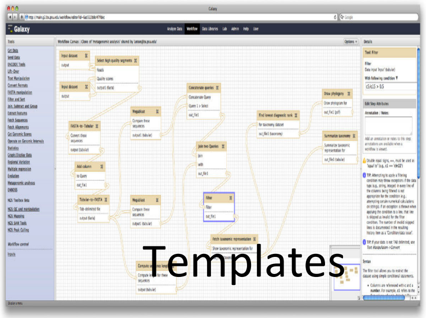
Accessibility

- Unified Web-interface for obtaining genomic data and applying computational tools to analyze the data
- Easily integrate your own tools and scripts for analysis
- Collection of tools (Tools Panel) that reflect good practices and community insights
- Access every step of analysis and intermediate results:
  - View, Download, Visualize, Reuse (History Panel)



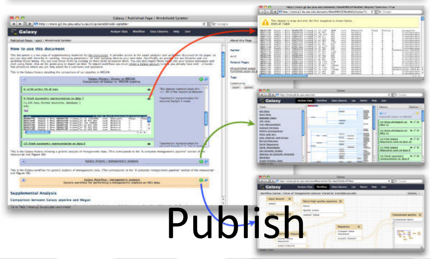
Reproducibility

- Track provenance and ensure repeatability of each analysis step:
  - input datasets, tools used, parameter values, and output datasets
- Intuitive Workflow Editor to create or modify complex workflows and use them as templates – Reusable and Reproducible



Transparency

- Publish and share metadata, histories, and workflows at multiple levels
- Store public and generated datasets as Data Libraries – e.g: hg19 Ref Genome
- Shared datasets and workflows can be imported by other users for reuse



Globus Integration

- Access Globus Endpoints and transfer data from within Galaxy UI and into Galaxy workspace
- Leverage local cluster or cloud based scalable computational resources for parallelizing the tools

Additional Capabilities

- Professionally managed and supported platform
- Best practice pipelines
  - Whole Genome, Exome, RNA-Seq, ChIP-Seq, …
- Enhanced workbench with breadth of analytic tools
- Technical support and bioinformatics consulting
- Access to pre-integrated end-points for reliable and high-performance data transfer (e.g. Broad Institute, Perkin Elmer, university sequencing centers, etc.)

globus genomics

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 – significantly improved 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

- Dobyns Lab – Seattle Children’s Hospital
- Cox Lab – University of Chicago
- ICBI / Georgetown University
- Kansas University Medical Center
- Volchenboum Lab – University of Chicago
- Olopade Lab – University of Chicago
- Inova Translational Medicine Institute
- Becton Dickinson
- Perkin Elmer
- Nagarajan Lab – Washington University St. Louis
- Genome Sciences Institute – Boston University
- Cedars-Sinai Medical Center – Los Angeles
- University of California – Irvine
- University of California – San Francisco
- University of Pittsburgh Medical Center
- Poroyko Lab – University of Chicago
- The Ohio State University Wexner Medical Center
- Broad Institute
- Many others…

Globus Genomics Platform Overview

DEMO

- Overview of the Globus Genomics interface
  - Interface (Tools, Histories)
  - Sharing Histories and Workflows
- Globus Integration in Galaxy
  - Globus interface
  - Globus transfers within Galaxy
  - View/Track Transfers

Scenario 1 – Bioinformatics Core Workflows

Use Case: Running workflows with all the tools and parameters predefined.

- Introduction to Exome seq pipeline
  - Import the best practices workflow
  - Scientific pipeline details
- Running a pre-defined exome seq pipeline with Globus transfers with one Sample
  - Submit a workflow
- Batch Submission with multiple-samples

Globus Genomics Demonstration
Scenario 1

Scenario 2 – Individual Researchers

Use Case: Running individual tools and creating/modifying workflows and the parameters

- Running individual tools
  - E.g: FastQC and Flagstat
- Importing a workflow
- Modifying the tools in the workflow
  - E.g: Change the aligner, Add/Remove Data transfer
- Modify the parameters of the tools

globus genomics

Globus Genomics Demonstration
Scenario 2

Questions?

Hands-On Exercise

1. Register with www.globus.org
2. Join the “Globus Genomics Workshop” group at https://www.globus.org/Groups
3. Login to http://demo.globusgenomics.org

4. Browse and Get Data from “SequencingCenter” endpoint
   Endpoint Name: **sulakhe#SequencingCenter**
   Username/Passwd: **genomics/globus**
   Input files: **Exome-Sample_Forward_1.fastq.gz**
   Exome-Sample_Reverse_2.fastq.gz

5. Change datatype of the input files to “fastqsanger” (click on the pencil sign)

6. Import a workflow from Shared Data
   Name: **ExomeSeq-Analysis-no-transfer_short_version**

7. Run the workflow