



globus online

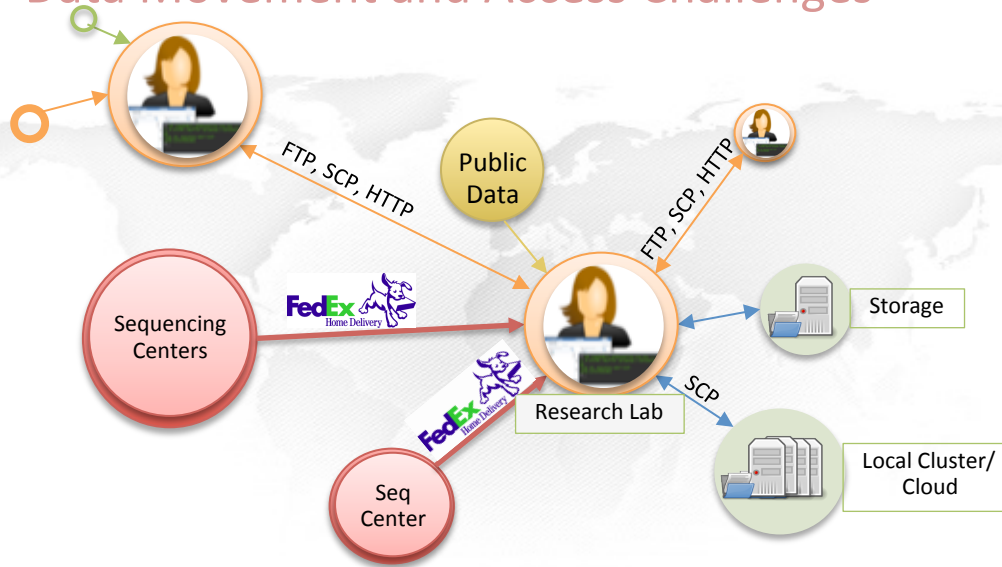


Globus Genomics

- **Challenges in Next Gen Sequencing analysis**
- **Description of Globus Genomics**
- **Building Globus Genomics**
- **Success Stories**
- **Future Steps**
- **Capability Walkthrough**
- **Q & A**

Challenges in Sequencing Analysis

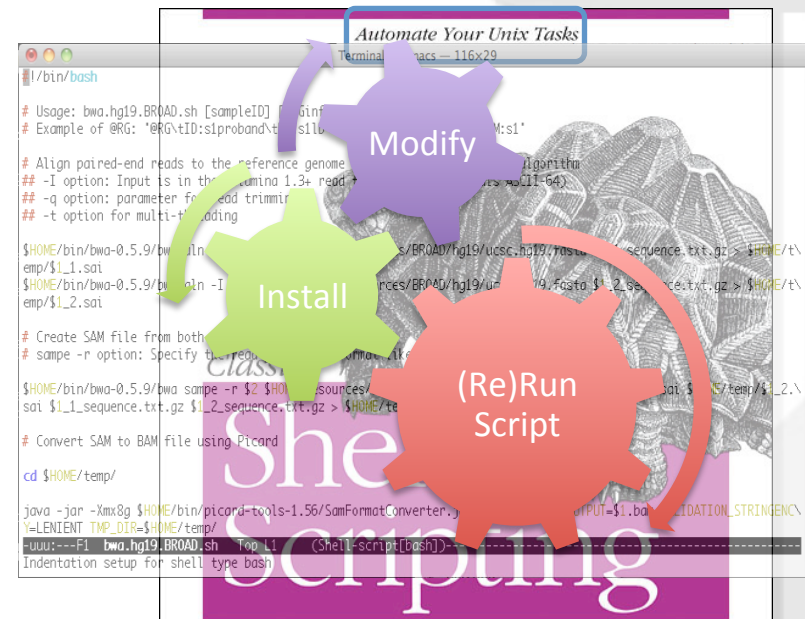
Data Movement and Access Challenges



- 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

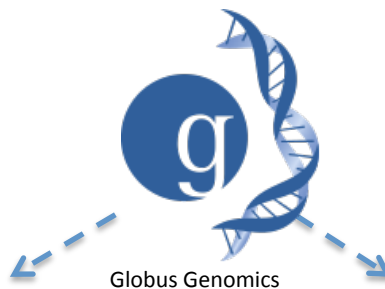
- Manually move the data to the Compute node
- Install all the tools required for the Analysis
 - BWA, Picard, GATK, Filtering Scripts, etc.
- Shell scripts to sequentially execute the tools
- Manually modify the scripts for any change
 - Error Prone, difficult to keep track, messy..
- Difficult to maintain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



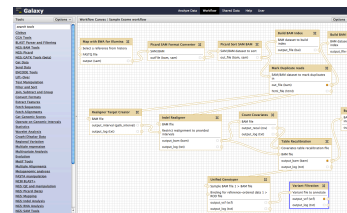
```

/bin/bash
# Usage: bwa.hg19.BROAD.sh [sampleID] [bin]
# Example of @RG: '@RG\tID:slproband\tSI:slb'
# Align paired-end reads to the reference genome
## -I option: Input is in the format of 1.3+ read
## -q option: parameter for read trimming
## -t option for multi-threading
$HOME/bin/bwa-0.5.9/bwa mem -t 8 -I 1000000 -q 15 -t 8 $HOME/resources/BROAD/hg19/ucsc/hg19.fasta $1_2_sequence.txt.gz > $HOME/temp/$1_1.sai
$HOME/bin/bwa-0.5.9/bwa mem -I 1000000 -q 15 -t 8 $HOME/resources/BROAD/hg19/ucsc/hg19.fasta $1_2_sequence.txt.gz > $HOME/temp/$1_2.sai
# Create SAM file from both
# sampe -r option: Specify the read pair name
$HOME/bin/bwa-0.5.9/bwa sampe -r $HOME/resources/SLB/BROAD/SLB11-64X $1_1.sai $1_2.sai > $HOME/temp/$1_2.sam
# Convert SAM to BAM file using Picard
cd $HOME/temp/
java -jar -Xmx8g $HOME/bin/picard-tools-1.56/SamFormatConverter.jar -LENIENT TMP_DIR=$HOME/temp/ INPUT=$1.bam OUTPUT=$1.bam VALIDATION_STRINGENCY=LENIENT --F1 bwa.hg19.BROAD.sh Top Line (Shell-script[bash])
Indentation setup for shell type bash
    
```

Manual Data Analysis



Galaxy Based Workflow Management System



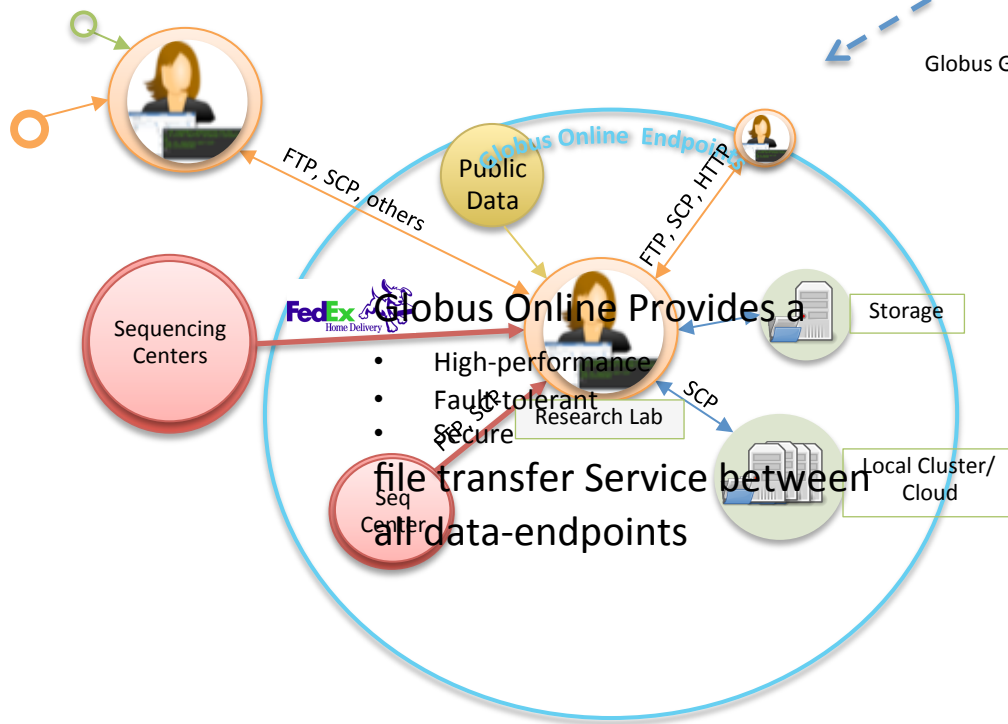
- Globus Online Integrated within Galaxy
- Web-based UI
- Drag-Drop workflow creations
- Easily modify Workflows with new tools



Globus Genomics on Amazon EC2

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

Galaxy Data Libraries



Data Management

Data Analysis

- **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**



- **Professionally managed and supported platform**
- **Access to scalable and elastic compute resources via AWS**
- **Best practice pipelines**
- **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, etc.)**
- **Cost-effective solution with subscription-based pricing**

- **Leveraging Globus Online as PaaS**
 - “Outsourcing” Identity Management, Access control, Reliable Data Transfer, Sharing and more
 - “Outsourcing” compute and storage using Amazon Web Services
- **Product built out using reliable services, APIs allows one to focus on the “opportunity”**
- **Lots of lessons learned in creating and scaling a vertical product offering**

- **Integration with Globus Catalog**
 - Better data discovery and metadata management
- **Integration with Globus Sharing**
 - Easy and Secure method to share large datasets with collaborators
- **Integration with Amazon Glacier for data archiving**
- **Support for high throughput computational modalities**
 - MapReduce
 - MPI

Onel-Skol Lab



Background: Cancer researchers sequencing normal and relapse genomes from cancer patients to investigate genetic factors in cancer relapse

Approach: Replaced outsourced analysis with Globus Genomics

Results: Achieved greater than 10X speed-up in analysis of NGS data and 10X cost savings compared to alternative solutions

Future Plans: Leverage flexibility in Globus Genomics to extend analysis pipelines and compare results utilizing recently added algorithms

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 10X speed-up in analysis of exome data

Future Plans: Leverage scale-out capability of Globus Genomics by running increasingly larger data sets

- **Cosmology Galaxy at NERSC**
 - Globus Transfer, Nexus
 - NEWT for execution
- **Proteomics**
- **CVRG**
- **Image Analysis**





globus online



Globus Genomics Demo

Translational Research..

- Cohort of patients are selected
- Patient Consent forms
- Specimen Collection

Diagnosis To Treatment

- Molecular-based medicine
- Improved Diagnosis
- Improved treatment



(Phenotype)

- **Collections of Clinical Data**
 - Demographics
 - Histories
 - Routine Tests
 - Diagnoses
 - Treatments
- **Imaging Data (EEG, ECG, etc)**

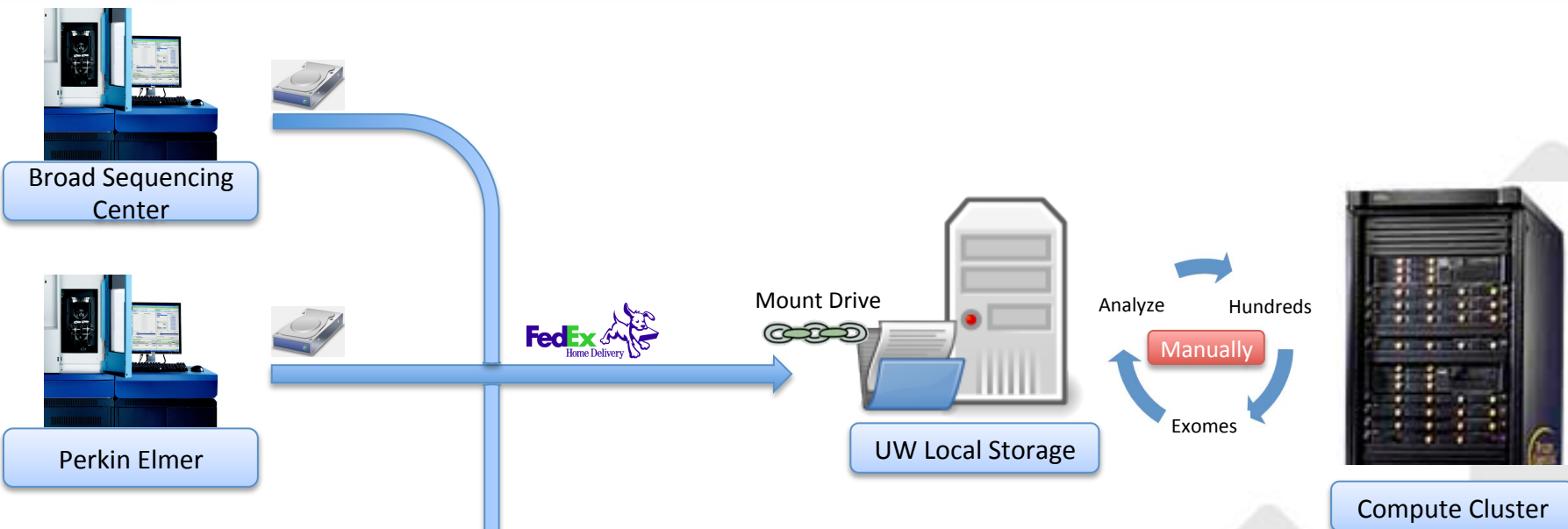
(Genotype-Phenotype)

- Identification of Molecular mechanisms
- Hypothesis Generation
- Features (ontologies) based:
 - *Enrichment and Prioritization of Disease Genes*
- Networks based Prioritization

(Genotype)

- **Genomic Data**
- **Next-gen Sequencing data**
 - Whole, Exome..
- **Preliminary Analysis**
 - Variant Calling (SNPS, CNVs)

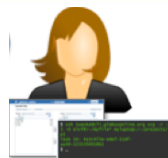
Dobyns Lab Use case..



- Hundreds of Exomes sequence data in the very near future
- Data coming from multiple sequencing centers
- Need scalable infrastructure for:
 - Data Management
 - Data Analysis
- Easily move data from sequencing center to Compute nodes
- Automatically analyze data over and over again
- Share data and analyses with collaborators

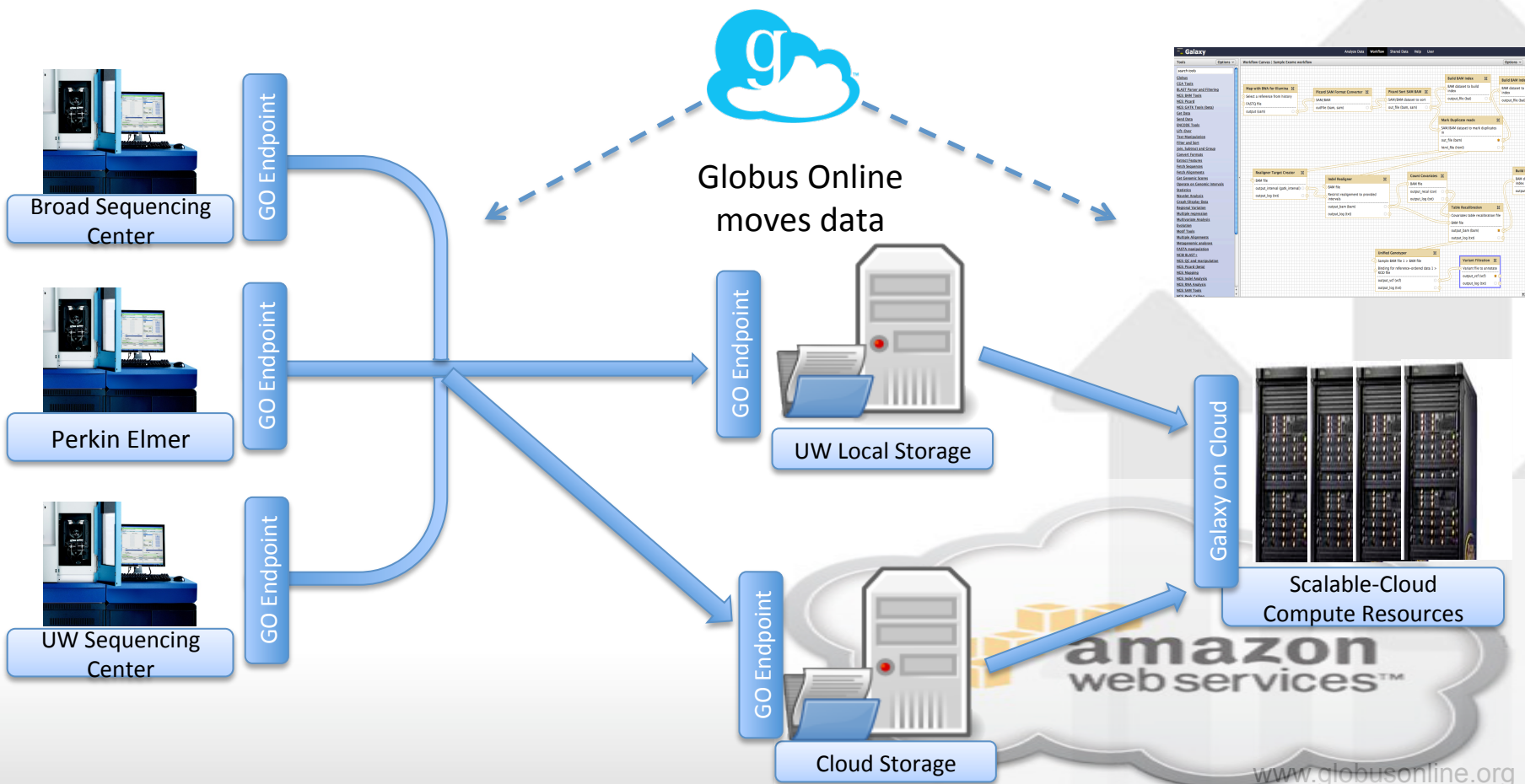
- <http://demo.galaxycloud.org>





Transfer Exome data from "Sequencing center" to "Galaxy Endpoint"

Run "Exome Analysis Pipeline"



- Sign up at www.globus.org/genomics
- If you have more questions? Ask...

