Computation Institute

Globus Genomics: An End-to-End NGS Analysis Service on the Cloud for Researchers and Core Labs

Ravi K Madduri, University of Chicago and Argonne National Laboratory

madduri@uchicago.edu

@madduri

Argonne NATIONAL LABORATORY

Galaxy

globus.org/genomics

globus.org/genomics

Our vision for a 21st century discovery infrastructure

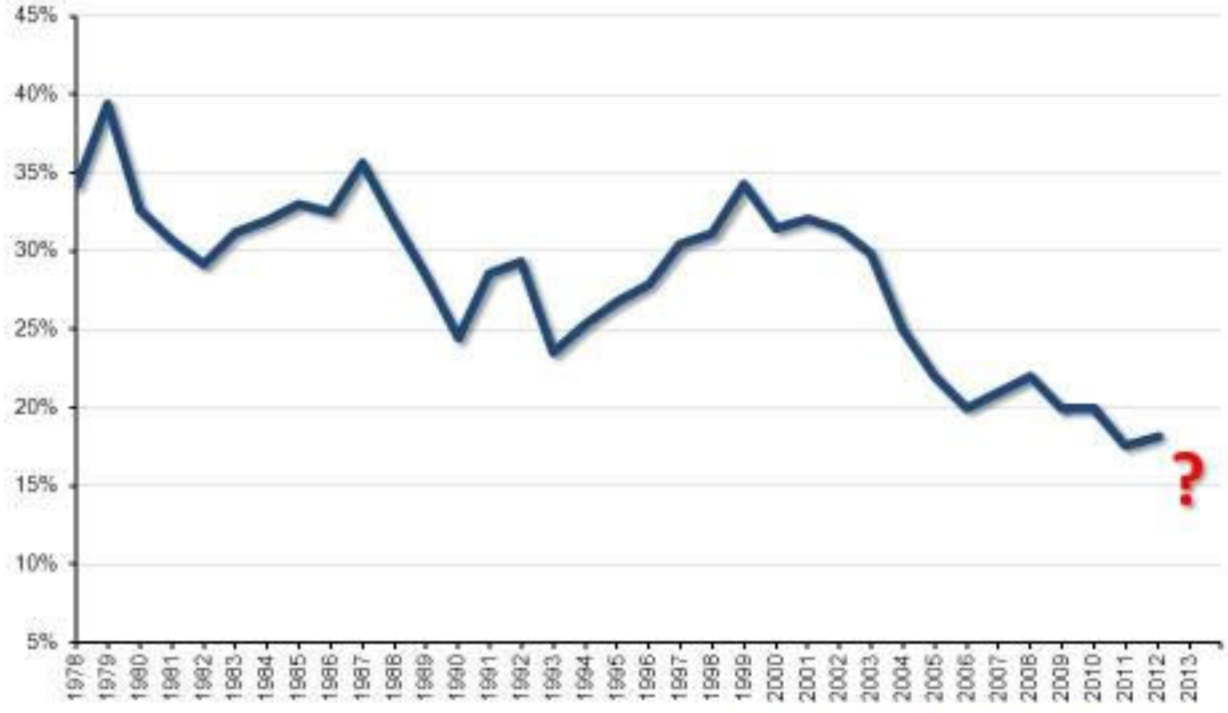
Provide **more** capability for **more** people at **lower cost** by **delivering "Science as a service"**

www.globus.org

Why is this Important?

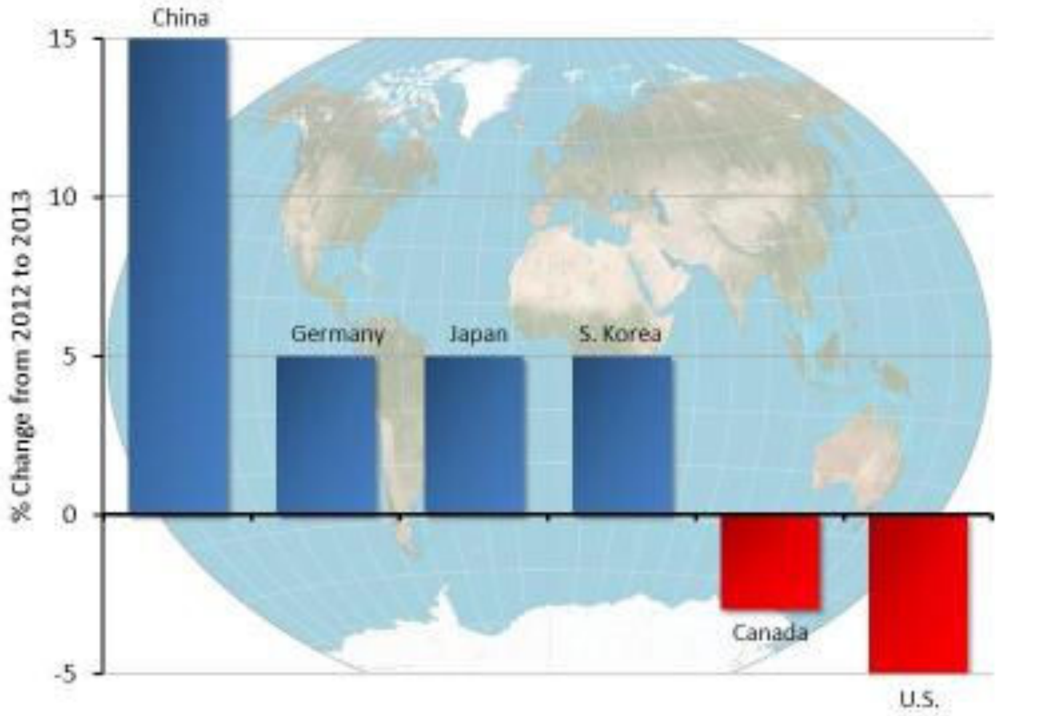
NIH Grant Application Success Rates

FY 1978-2013



Source: NIH http://report.nih.gov/success_rates/

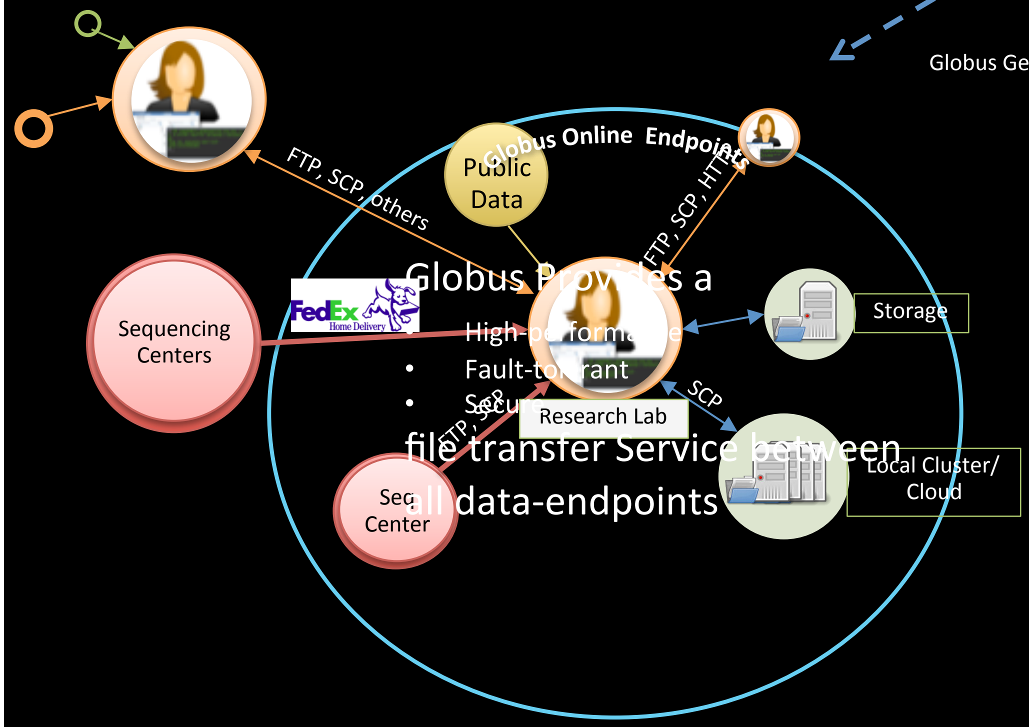
Scientific R&D Spending



Source: *Cell.* 2013 Jul 3;154(1):16-9.

Globus Genomics

Globus Genomics



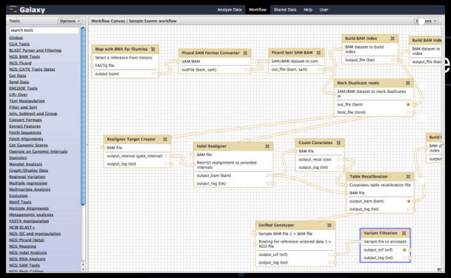
Globus Provides a

- High-performance
- Fault-tolerant
- Secure

file transfer Service between all data-endpoints

Galaxy Data Libraries

Galaxy Based Workflow Management System



- Globus Integrated wit Galaxy
- Web-based UI
- Drag-Drop workflow creations
- Easily modify Workflo with new tools

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

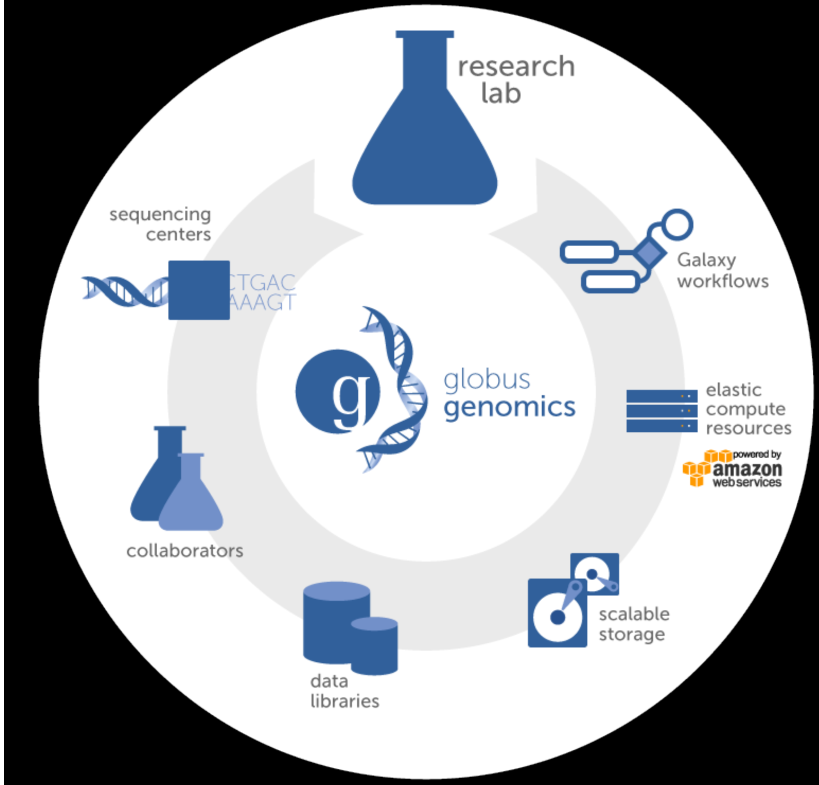
Globus Genomics on Amazon EC2

Data Management

Data Analysis

globus.org/genomics

Core Capabilities



- Computational profiles for various analysis tools to provide optimal performance
- Resources can be provisioned on-demand with Amazon Web Services cloud based infrastructure
- High performance, Reliable Data movement is streamlined with integrated Globus file-transfer functionality
- Integrated Globus endpoints and Campus login

globus.org/genomics

Diversity of Collaborations

Seattle Children's
HOSPITAL • RESEARCH • FOUNDATION

Dobyns Lab

KU MEDICAL CENTER
The University of Kansas

Washington University in St. Louis

Nagarajan Lab

UPMC
University of Pittsburgh Medical Center

BROAD INSTITUTE

THE UNIVERSITY OF CHICAGO

Cox Lab
Volchenboum Lab
Olopade Lab

GEORGETOWN UNIVERSITY

OHIO STATE | Wexner Medical Center

INOVA®
Join the future of health.

BD

UCSF

BU
Genome Science Institute
Boston University Medical Center

PerkinElmer®
For the Better

CEDARS-SINAI MEDICAL CENTER®

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.

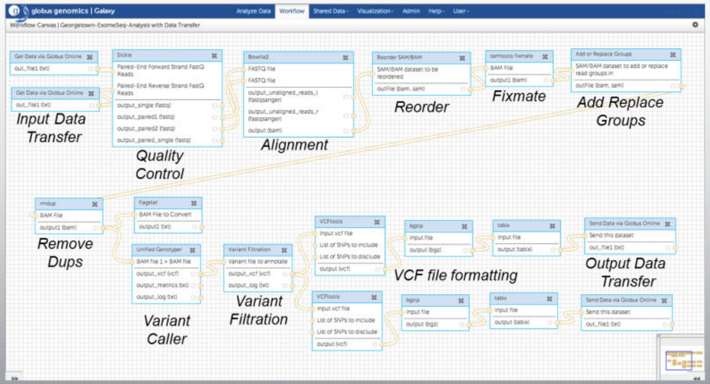
Future Plans: Provide Globus Genomics as a well-managed platform-as-a-service for ICBI collaborators and users

Main NGS workflows in use at ICBI

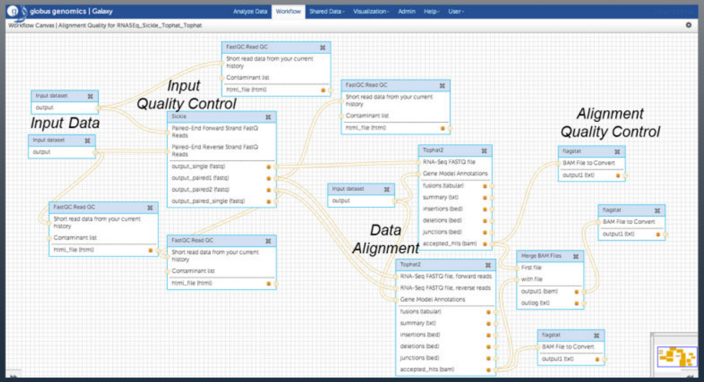
ICBI Workflows

- **Whole-genome and Exome Seq Workflow**
  - In collaboration, the Globus Genomics and ICBI teams, tested and benchmarked the analytical workflows
  - Workflows include data transfer from data source to analysis platform using Globus Online.
- **RNA-Seq Workflow**
  - Performed quality control for alignment of Transcriptome data.
  - Includes multiple read filtering tools (Fastx toolkit, native Galaxy filtering tools) to achieve optimal alignment statistics.
  - Includes comparing performance of Tophat2 and RSEM alignment.

Whole-Genome and Exome Analysis Workflows



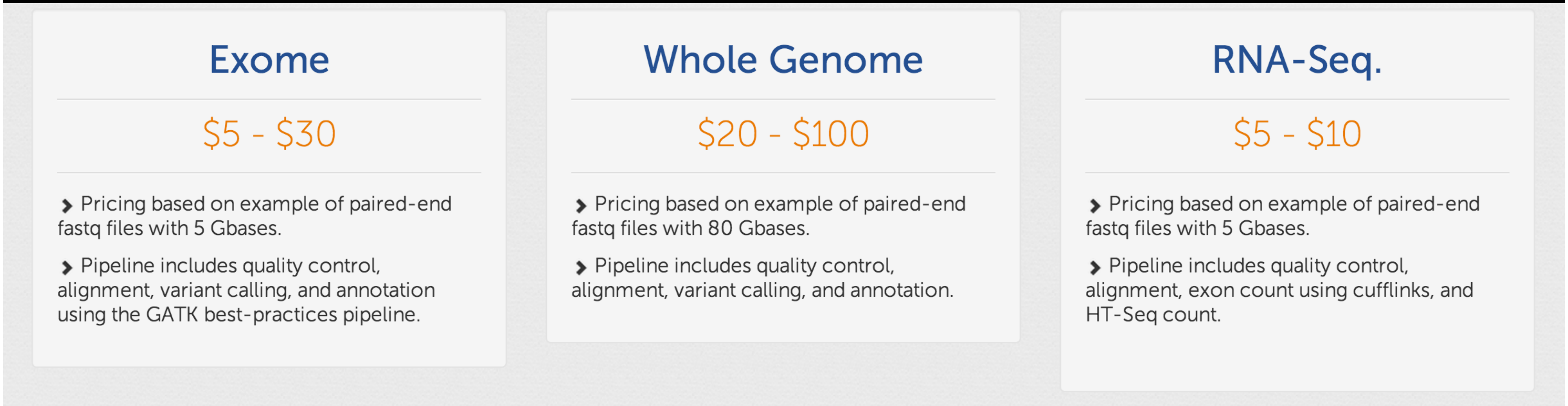
Alignment Quality Control for RNA-Seq Analysis



Summary of Results

- Completed setup of Globus endpoints and validated data transfer capabilities
- Wrapped additional tools and validated execution of Whole Genome, Exome, and RNA-Seq pipelines utilizing Globus Genomics
- Ran all three targeted pipelines at scale against large data sets demonstrating significant speed-up of execution compared to serial approaches
- Optimized the Globus Genomics environment in AWS to efficiently handle burst requirements through elastic provisioning / de-provisioning of compute capacity
- Gathered performance and quality data associated with running all three pipelines at scale on the optimized Globus Genomics instance
- **Jointly prepared and presented several posters: ICBI Symposium 2013; NIH translational Genomics Symposium etc.**
- **Developed Platform to share & learn bioinformatics best practices and technical expertise**

Affordability



Exome

$5 - $30

› Pricing based on example of paired-end fastq files with 5 Gbases.

› Pipeline includes quality control, alignment, variant calling, and annotation using the GATK best-practices pipeline.

Whole Genome

$20 - $100

› Pricing based on example of paired-end fastq files with 80 Gbases.

› Pipeline includes quality control, alignment, variant calling, and annotation.

RNA-Seq.

$5 - $10

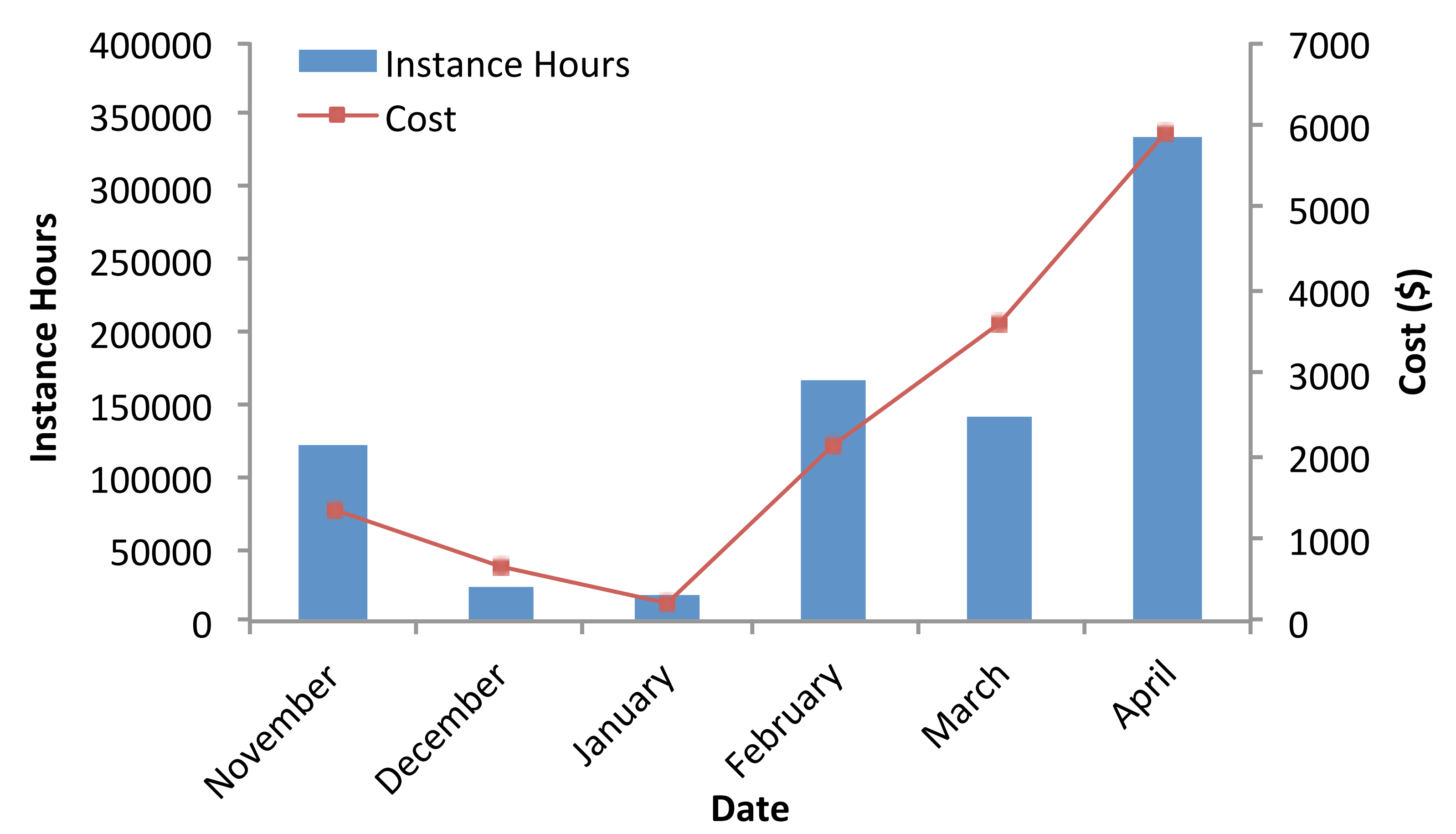
› Pricing based on example of paired-end fastq files with 5 Gbases.

› Pipeline includes quality control, alignment, exon count using cufflinks, and HT-Seq count.

- Pricing includes
  - Estimated compute
  - Storage (one month)
  - Globus Genomics platform usage
  - Support

globus.org/genomics

Scalability



800K Core hours in last 6 months

globus.org/genomics

Security and Privacy

Globus Genomics compliance with the NCBI Database of Genotypes and Phenotypes (dbGaP) security best practices

Protecting the Security of Controlled Data on Servers

- All Globus Genomics servers are protected by Amazon Security Groups and by stateful packet inspection firewalls. Only necessary services are allowed
- All relevant security patches are applied as soon as they are available
- Globus Genomics and Globus provide sharing solutions that are secure and user controlled
- Globus Genomics uses HTTPS and GridFTP protocol with authentication and encryption when transferring the files
- Data access is strictly restricted to individual users and only users can share the data with other users. We provide detailed instructions to our users on data security and access control.
- The data sharing and access policies on Globus Genomics are retained across all the systems involved

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 (CLI based tools
- Collection of tools (Tools Panel) that reflect good practices and community insights
- Access every step of analysis and intermediate results: View, Download, Visualize, Reuse

Reproducibility and Reuse

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

Collaboration

- Users from different institutions can come together and meet in the middle
- Jointly create and share analytical pipelines
- Securely share data
- Verify results

Collaboration

Globus Genomics facilitates meta-analysis across sequence datasets. Large-scale meta-analysis has been hugely important in driving the success of GWAS. With GWAS, investigators could simply share summary results without losing much, but for sequencing, we do much better when we jointly call the samples and reanalyze. There are only a few places that can do this at scale now, and creating resources that allow groups to come together spontaneously to do this is hugely important. This is a really important opportunity for the scientific community to have a very distributed approach to large-scale analysis from the control perspective, while still being centralized and cost-effective from the hardware and software end. It is straightforward for groups to come together to do meta-analysis over many large sequence datasets in which data can be secured so that raw data are not directly shared (but rather each group maintains control over access to their raw data) but variant calls can be made over all data with a shared pipeline that can then be used to conduct analysis over all of the new variant calls. Power to the people!!

-- Nancy Cox, PhD
University of Chicago

Sustainability

- Our goal is to build service that lives beyond a funded proposal
- Two pricing options and multiple usage tiers.
  - Targeted users include individual research groups and bioinformatics cores
  - Platform pricing (includes only subscription to the Globus Genomics platform)
  - Bundled pricing (includes Globus Genomics platform subscription and AWS usage costs)

- More information on Globus Genomics and to sign up for a free trial : www.globus.org/genomics

- More information on Globus: www.globus.org

Our work is supported by:

U.S. DEPARTMENT OF ENERGY

NSF

NATIONAL INSTITUTES OF HEALTH

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

amazon web services

globus.org/genomics

Thank you!

@madduri

globus.org/genomics