

# Large-scale Research Data Management and Analysis Using Globus Services

Ravi Madduri

Argonne National Lab

University of Chicago

@madduri



# Outline

- Who we are
- Challenges in Big Data Management and Analysis
- Sustainability and Reproducibility
- Globus Research Data Management Service
  - Numbers, Usage Stats
- Globus Genomics
  - Description
  - Novel Pipelines
  - User segments
  - Adoption
  - 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**

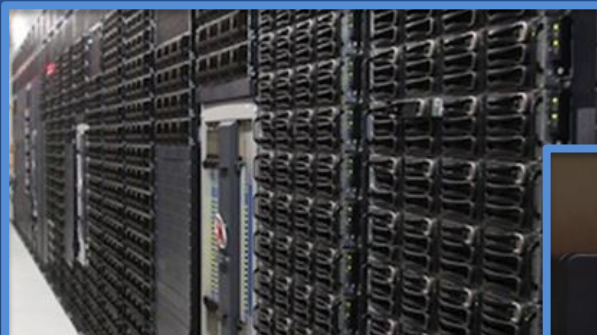


# Research data management scenarios and challenges

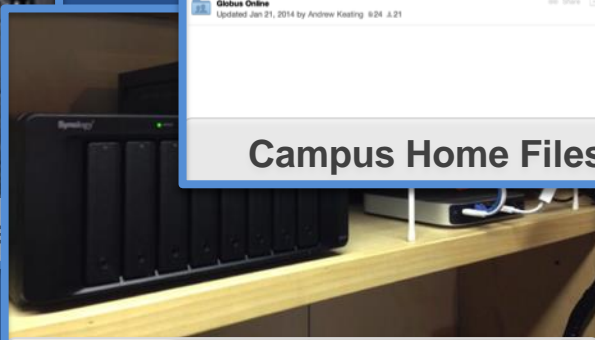
In Big, Medium *and* Small data



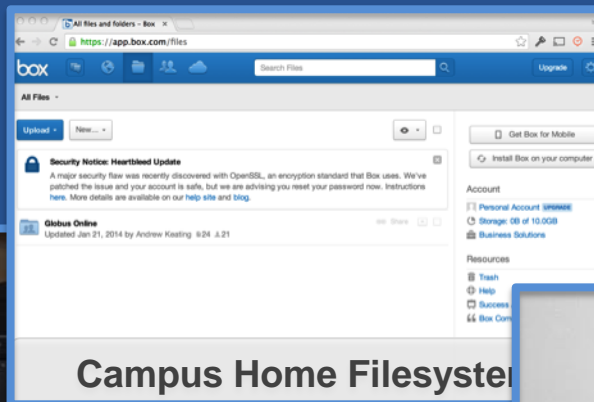
“I need to easily, quickly, & reliably move or mirror portions of my data to other places.”



Research Computing HPC Cluster



Lab Server



Campus Home Filesystem



Personal Laptop



Desktop Workstation



XSEDE Resource



Public Cloud

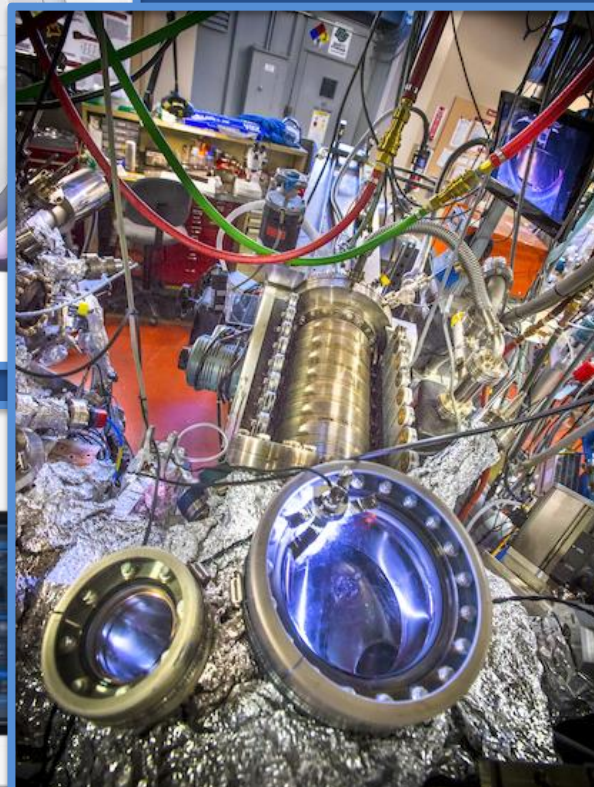


“I need to get data from a scientific instrument to my analysis server.”

MRI



Advanced  
Light Source



Next Gen  
Sequencer



Light Sheet Microscope



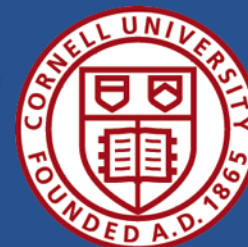
“I need to easily and securely share my data with my colleagues at other institutions.”



NCAR



Computation  
Institute







“I need a good place to store / backup / archive my (big) research data, at a reasonable price.”



Campus Store



Mass Store



Public Cloud Archive



“I need to publish my data so that others can find it and use it.”



Scholarly  
Publication

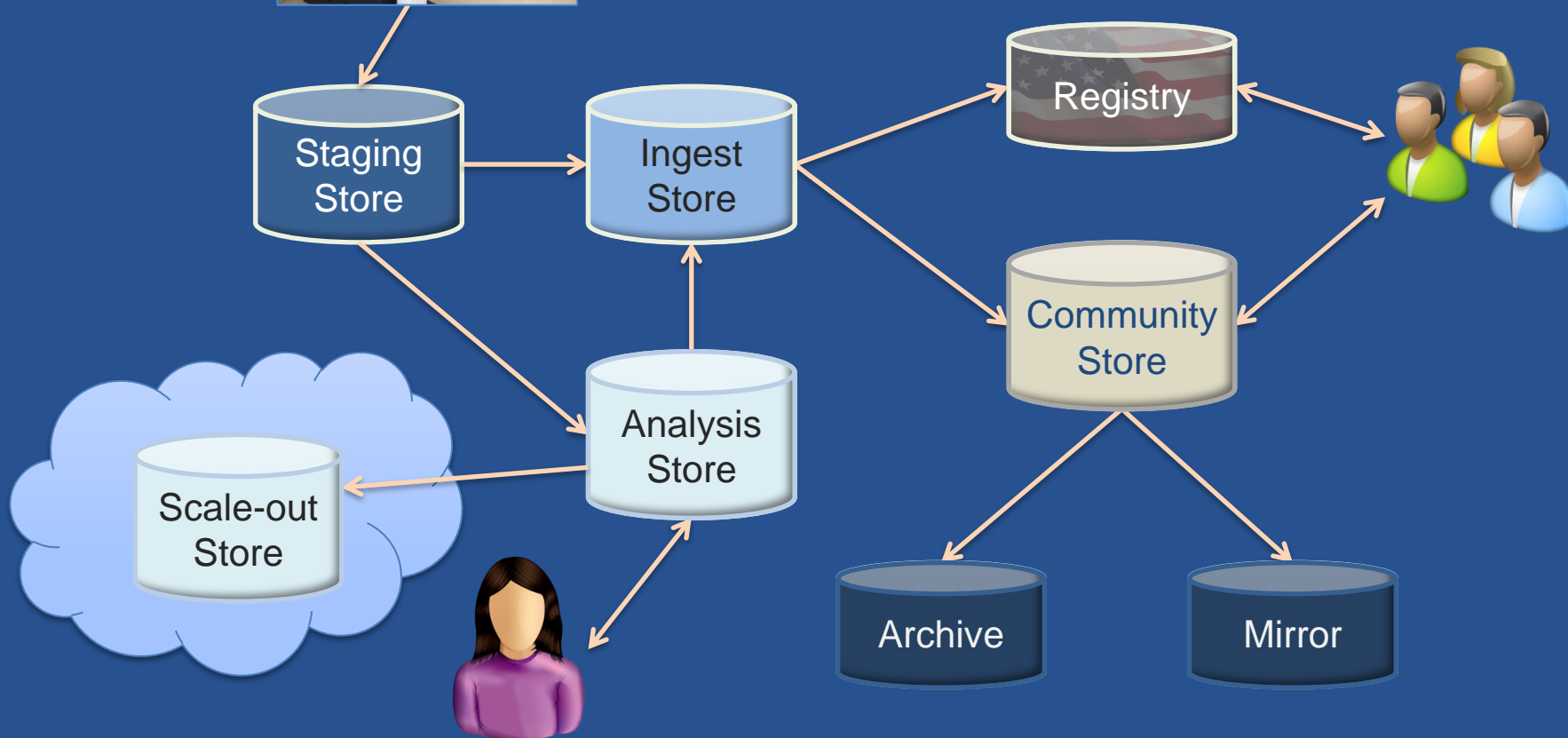
Reference  
Dataset



Active  
Research  
Collaboration

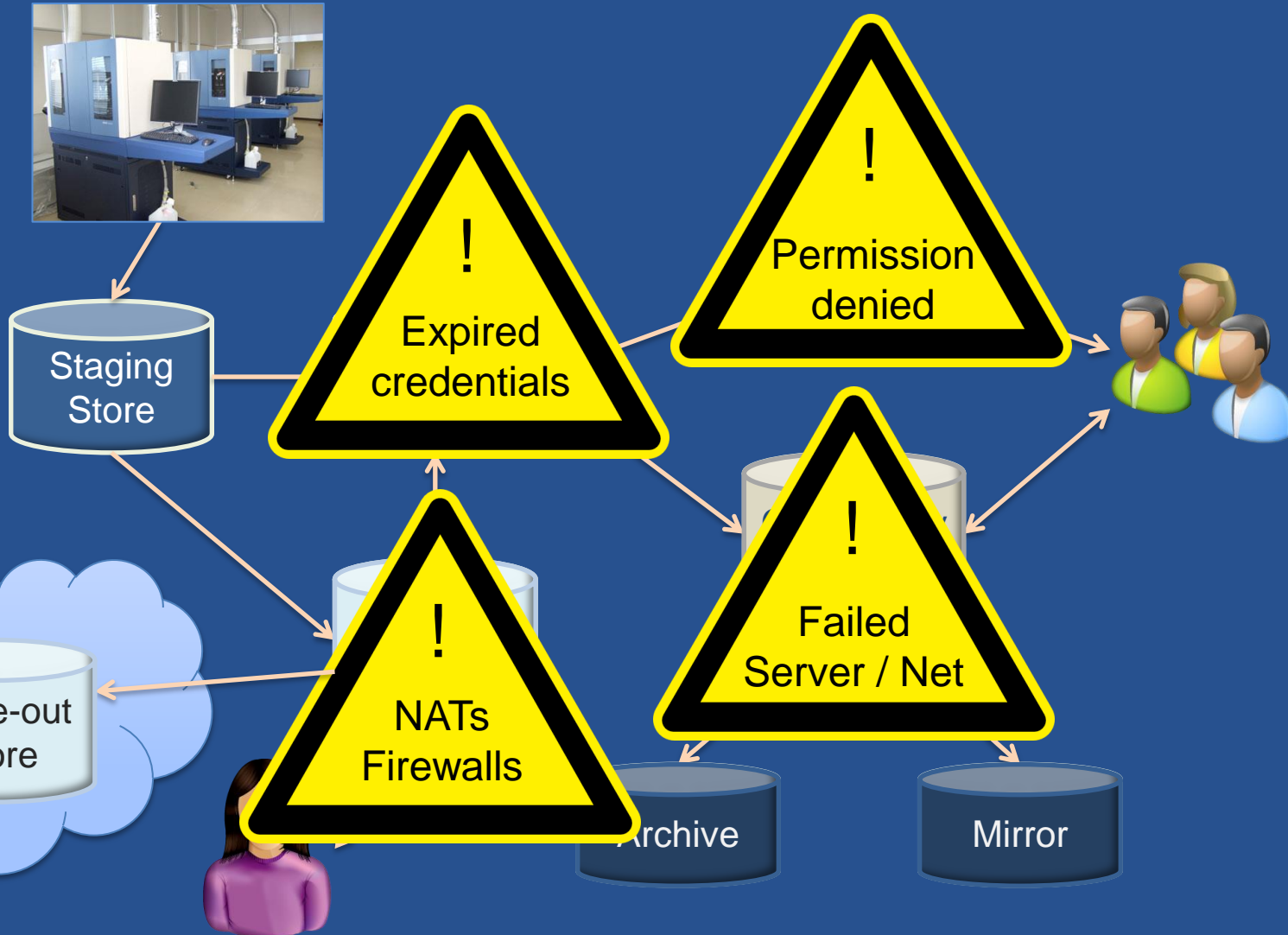


# Managing data should be easy ...





# ... but it's hard and frustrating!

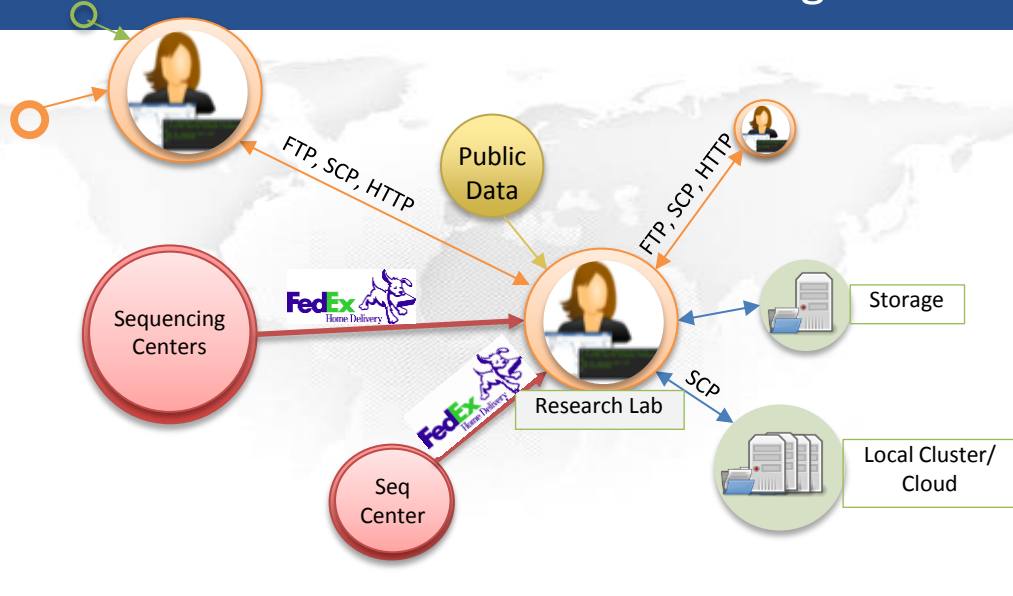






# Challenges in Sequencing Analysis

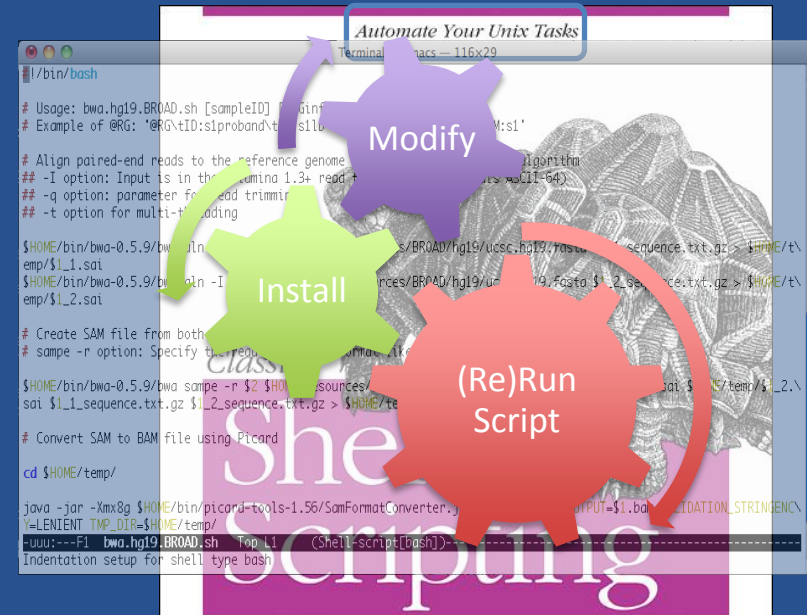
## Data Movement and Access Challenges



- 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

- 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



Globus is...

Research data  
management...

...delivered via SaaS



Globus delivers...

Big data transfer, sharing,  
publication, and discovery...

...directly from your own  
storage systems





It's about the user experience...

**flickr** ...for your photos

**Gmail** ...for your e-mail  
by Google

**NETFLIX** ...for your entertainment

**globus** ...for your research data



# 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

# Managing the research data lifecycle with Globus



## Light Source



Globus transfers files reliably, securely

2

## Compute Facility



4 Globus controls access to shared files on existing storage; no need to move files to cloud storage!



7 Curator reviews and approves; data set published on campus or other system

1 PI initiates transfer request; or requested automatically by script, science gateway



3 PI selects files to share, selects user or group, and sets access permissions

6 Researcher assembles data set; describes it using metadata (Dublin core and domain-specific)



## Publication Repository

- SaaS → Only a web browser required
- Access using your campus credentials
- Globus monitors and informs throughout

5 Researcher logs in to Globus and accesses shared files; no local account required; download via Globus



## Personal Computer

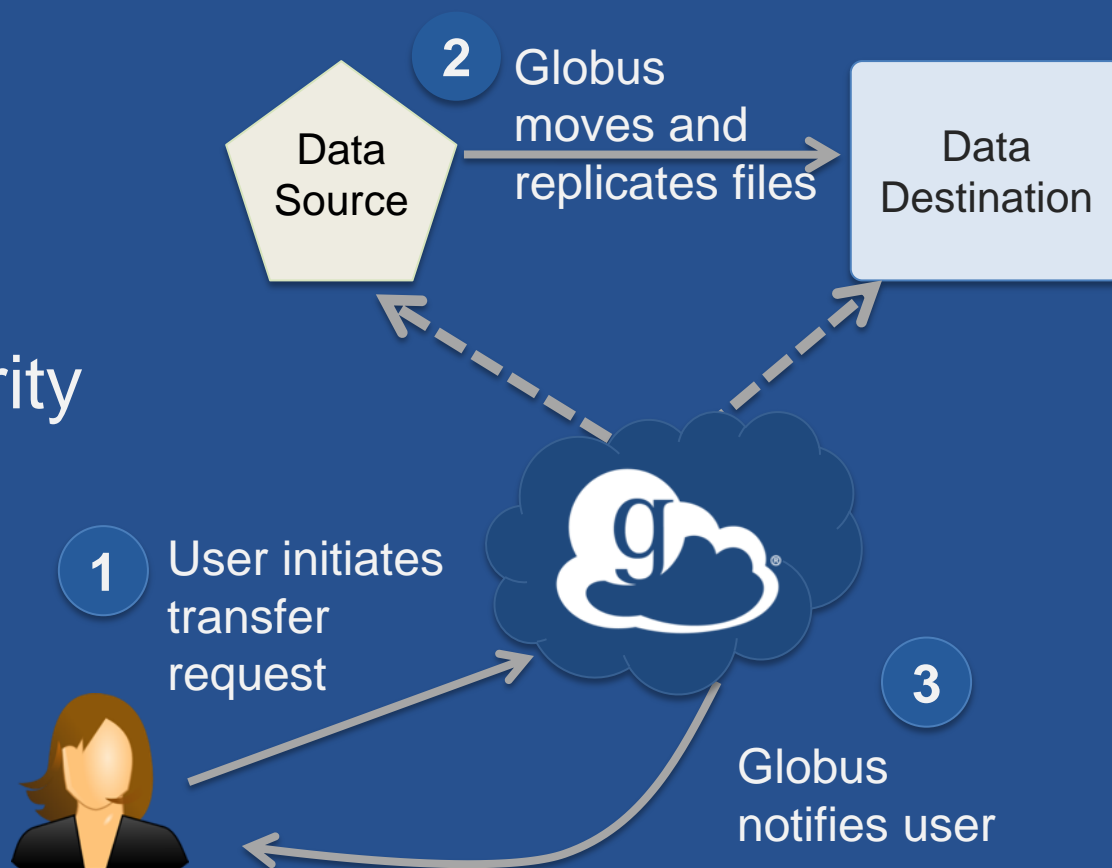
8 Peers, collaborators search and discover datasets; transfer and share using Globus





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

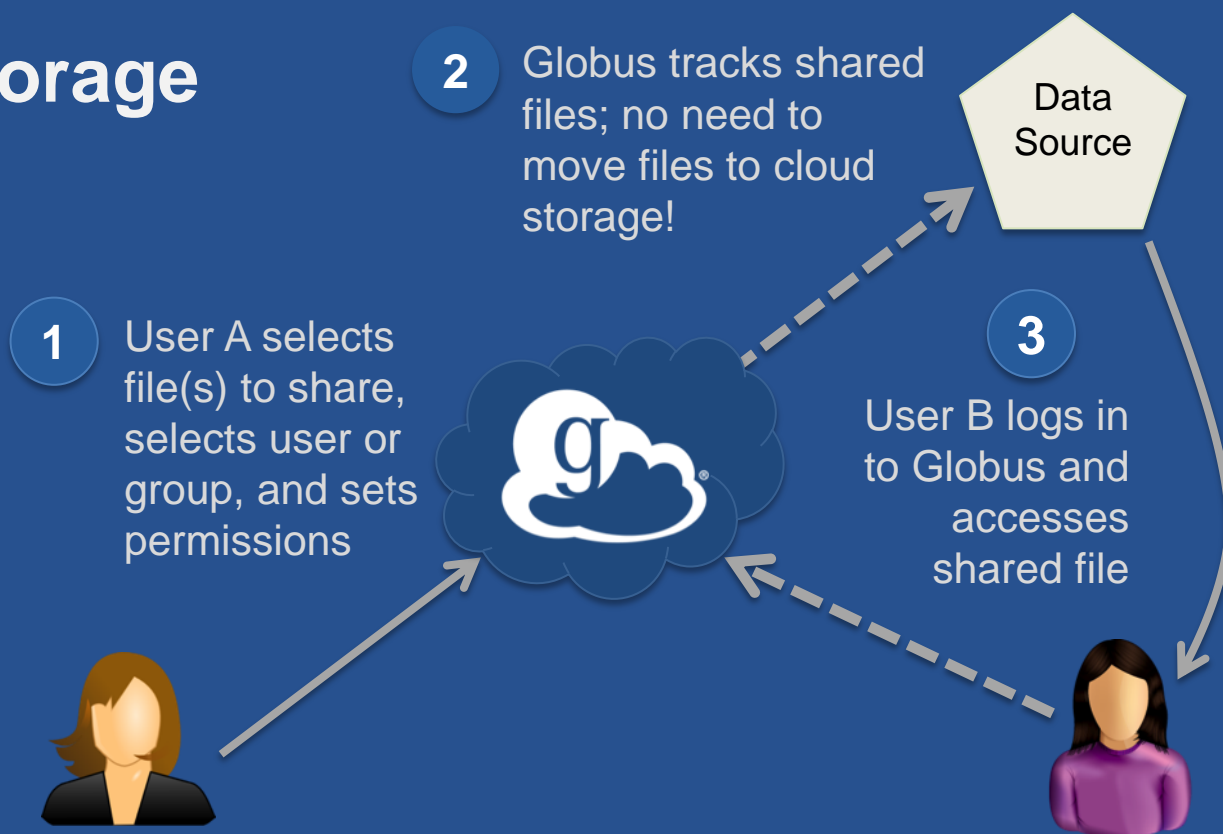
- “Fire-and-forget” transfers
- Automatic fault recovery
- Seamless security integration
- Powerful GUI and APIs





# Simple, secure *sharing* off existing storage systems

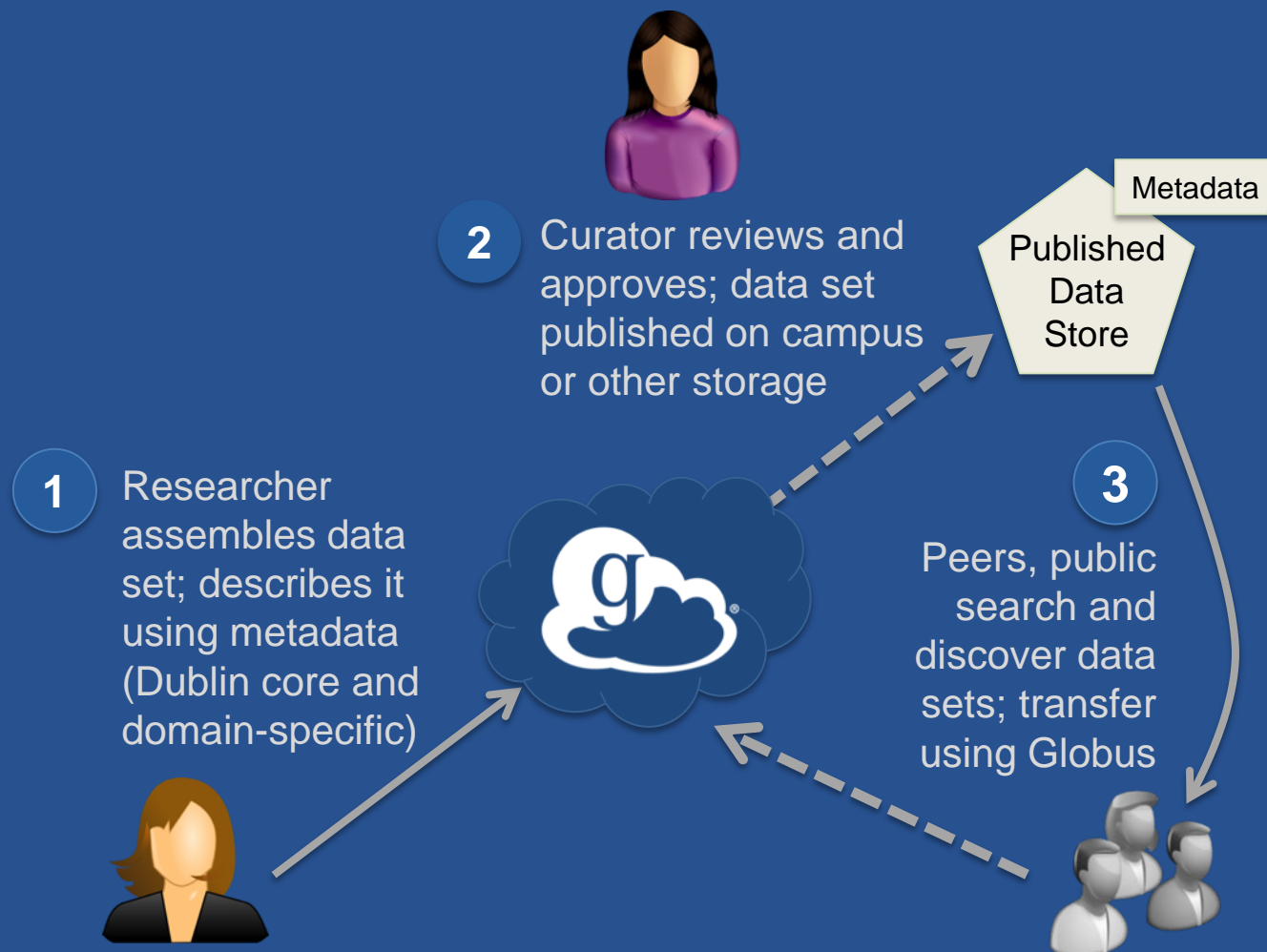
- Easily share large data with any user or group
- No cloud storage required





# Curated *publication* of data, with relevant metadata for *discovery*

- Identify
- Describe
- Curate
- Verify
- Access
- Preserve





# Globus Adoption and Usage

- 166,449 active Globus endpoints
- 27,961 users registered
- Biggest transfer: 500.42TB
- Longest running transfer: 182 days.
- Fastest transfer: 58.5Gbps (average)
- 55TB moved per day, on average, since the service was launched in November 2010
- Average throughput: 637.7Mbps (since service launch)



**Flexible, scalable,  
affordable genomics  
analysis for all biologists**





# Challenges in Scaling Up

- Rapidly evolving state-of-the-art in tools
- Things work reasonably well for small-scale
  - Local and on cloud
- Large-scale analysis requires
  - A computationally gifted postdoc or two
  - Co-location with a large compute facility hungry for justifying purchase
  - Understanding different kinds of parallelism
    - Tool level
    - Workflow level
  - And relate it to science
    - Chromosome level
    - Sample level

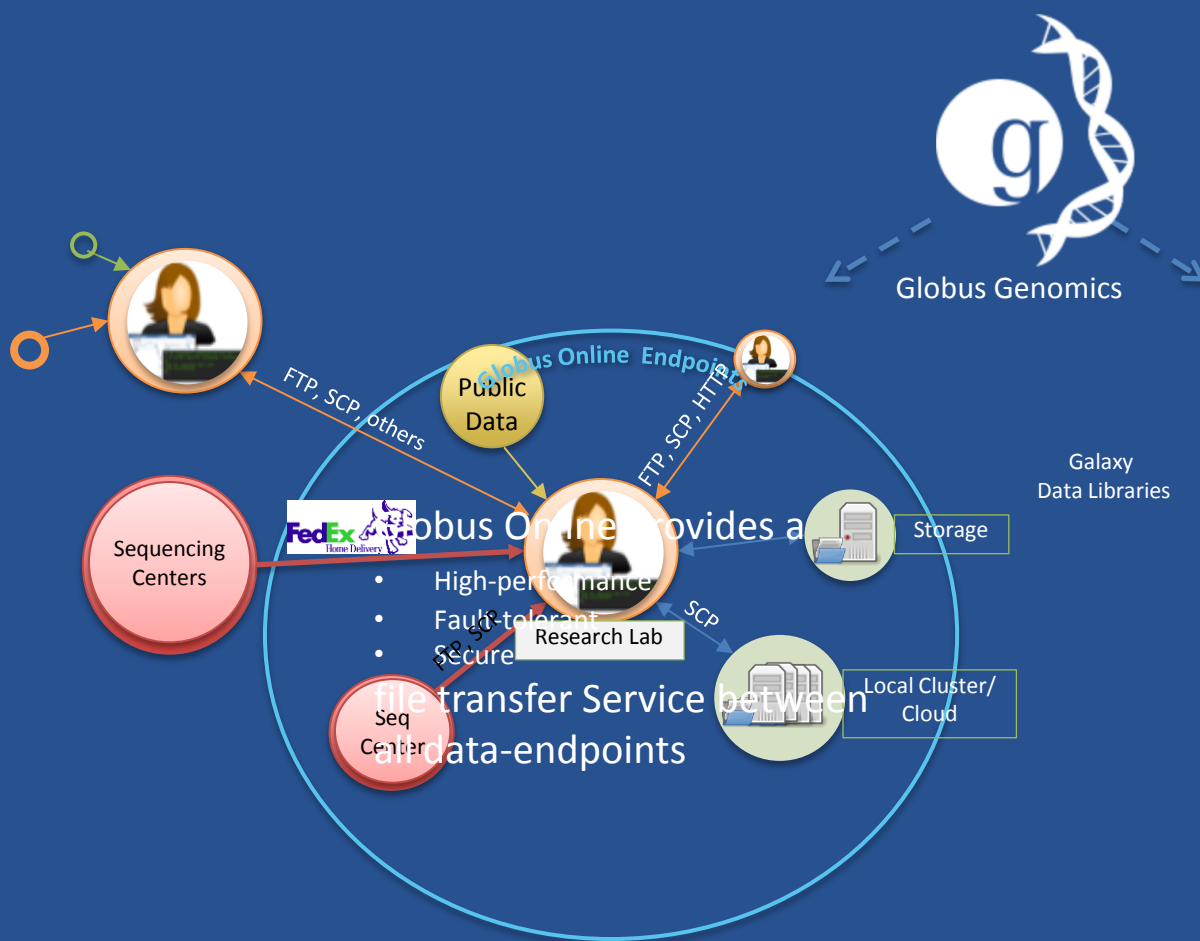


# Challenges in Scaling Up

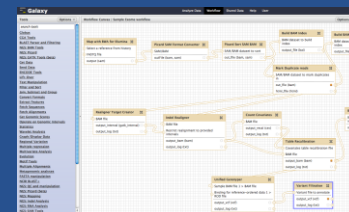
- Doing it right once
- Reproducing it
- Doing it again for the same dataset or a new dataset
- Sharing
- Publishing
- Economics
- Expertise



# Globus Genomics



## Galaxy Based Workflow Management System



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



Galaxy on Cluster/Cloud

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

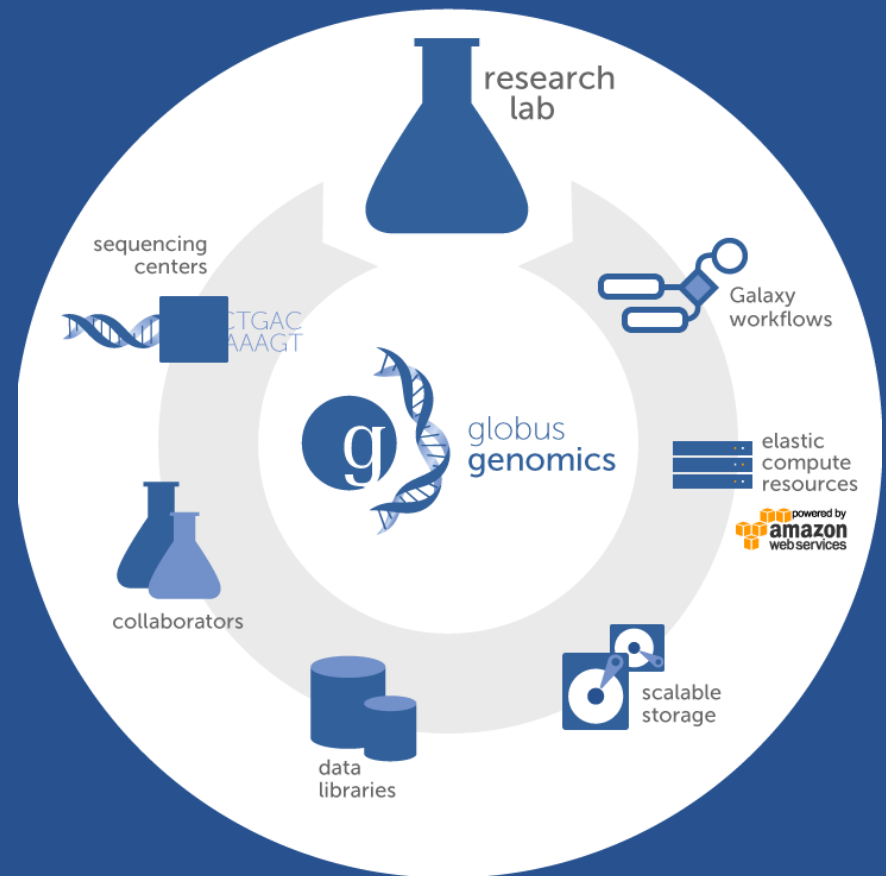
Data Management

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





# 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.)
- Cost-effective solution with subscription-based pricing



# Adoption of Globus Genomics

- Individual Research Groups
- Informatics cores at various universities
- Health Care providers
- Sequencing Service Providers



Cox lab, UChicago

## **Consensus Genotyper for Exome Sequencing: Improving the Quality of Exome Variant Genotypes**

Vassily Trubetskoy<sup>1</sup>, Ravi Madduri<sup>2</sup>, Alex Rodriguez<sup>2</sup>, Jeremiah Scharf<sup>3</sup>, Paul Dave<sup>2</sup>, Ian Foster<sup>2</sup>, Nancy Cox<sup>1</sup>, Lea Davis<sup>1</sup>

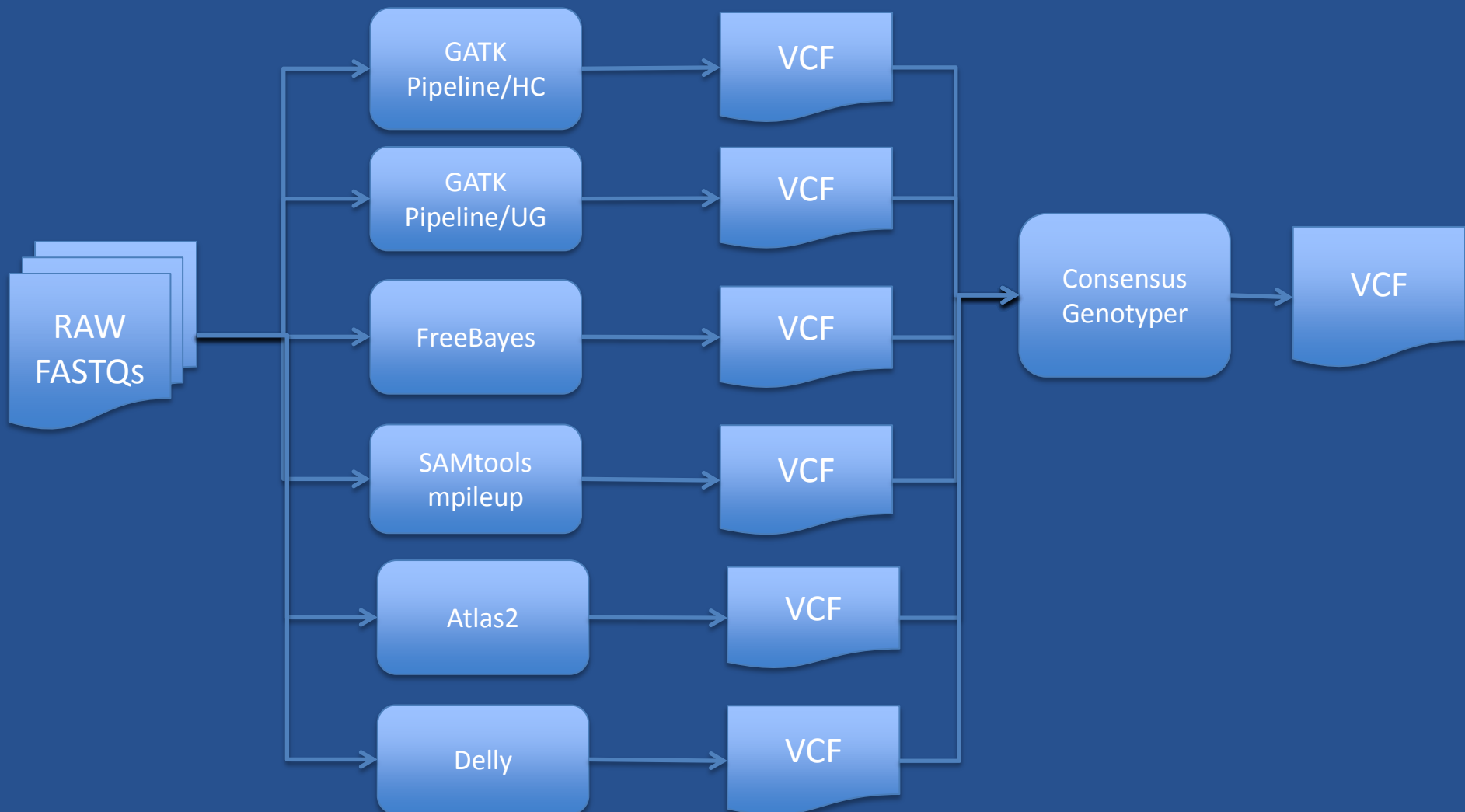
1) Section Genetic Medicine, University of Chicago, Chicago, IL; 2) Computation Institute, University of Chicago, Chicago, IL;

3) Department of Neurology, Massachusetts General Hospital, Boston, MA

- 134 samples and 4 workflows
- 4 TB data
- 2200 core hours in 6 days



# Consensus Genotyper – Version 1







# Olopade lab, UChicago

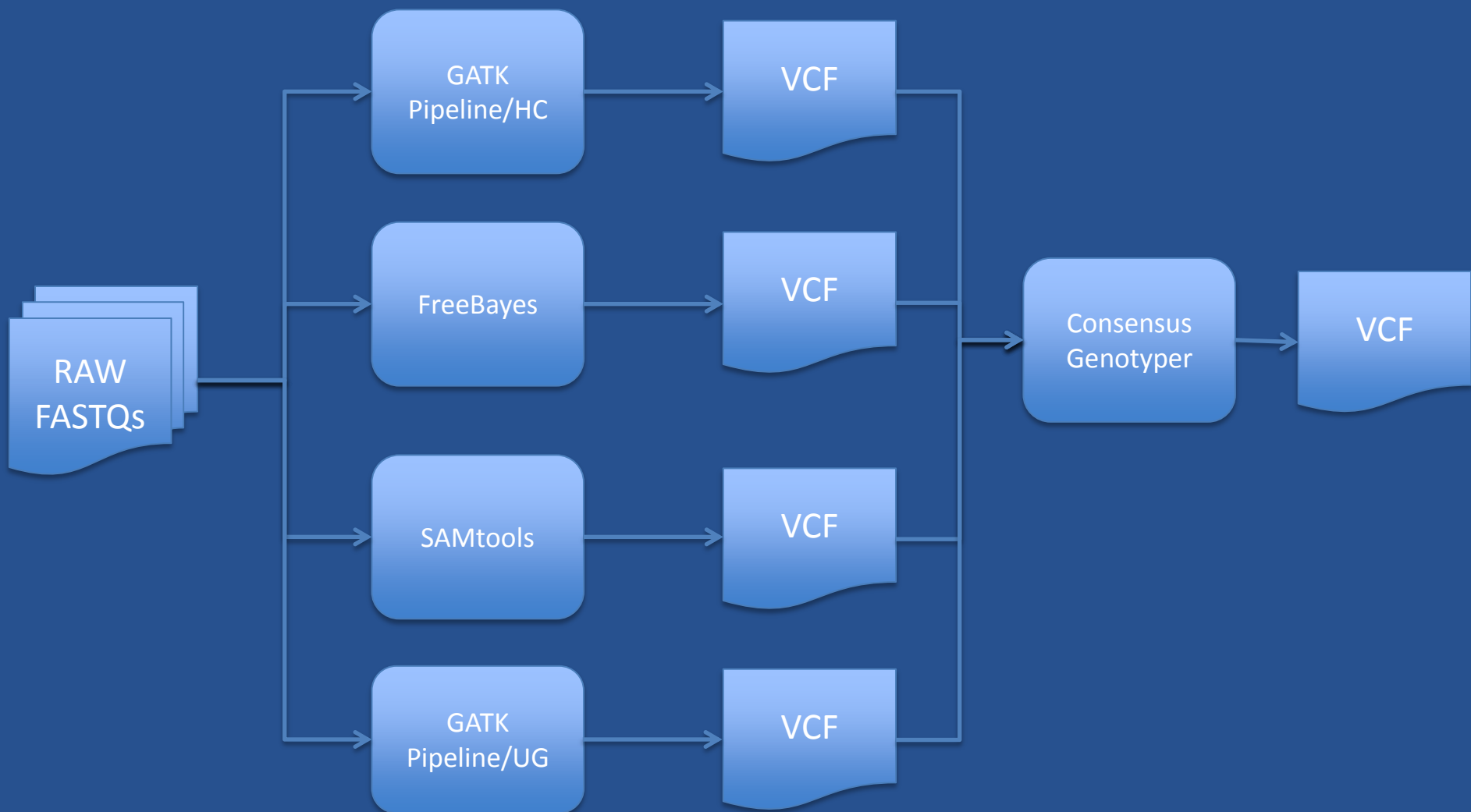
## A profile of inherited predisposition to breast cancer among Nigerian women

Y. Zheng, T. Walsh, F. Yoshimatsu, M. Lee, S. Gulsuner, S. Casadei, A. Rodriguez, T. Ogundiran, C. Babalola, O. Ojengbede, D. Sighoko, R. Madduri, M.-C. King, O. Olopade

- 200 targeted exomes
- 200 GB data
- 76,920 core hours in 1.25 days



# Consensus Genotyper – Version 2





# Innovation Center for Biomedical Informatics - Georgetown

A case study for high throughput analysis of NGS data for translational research using Globus Genomics

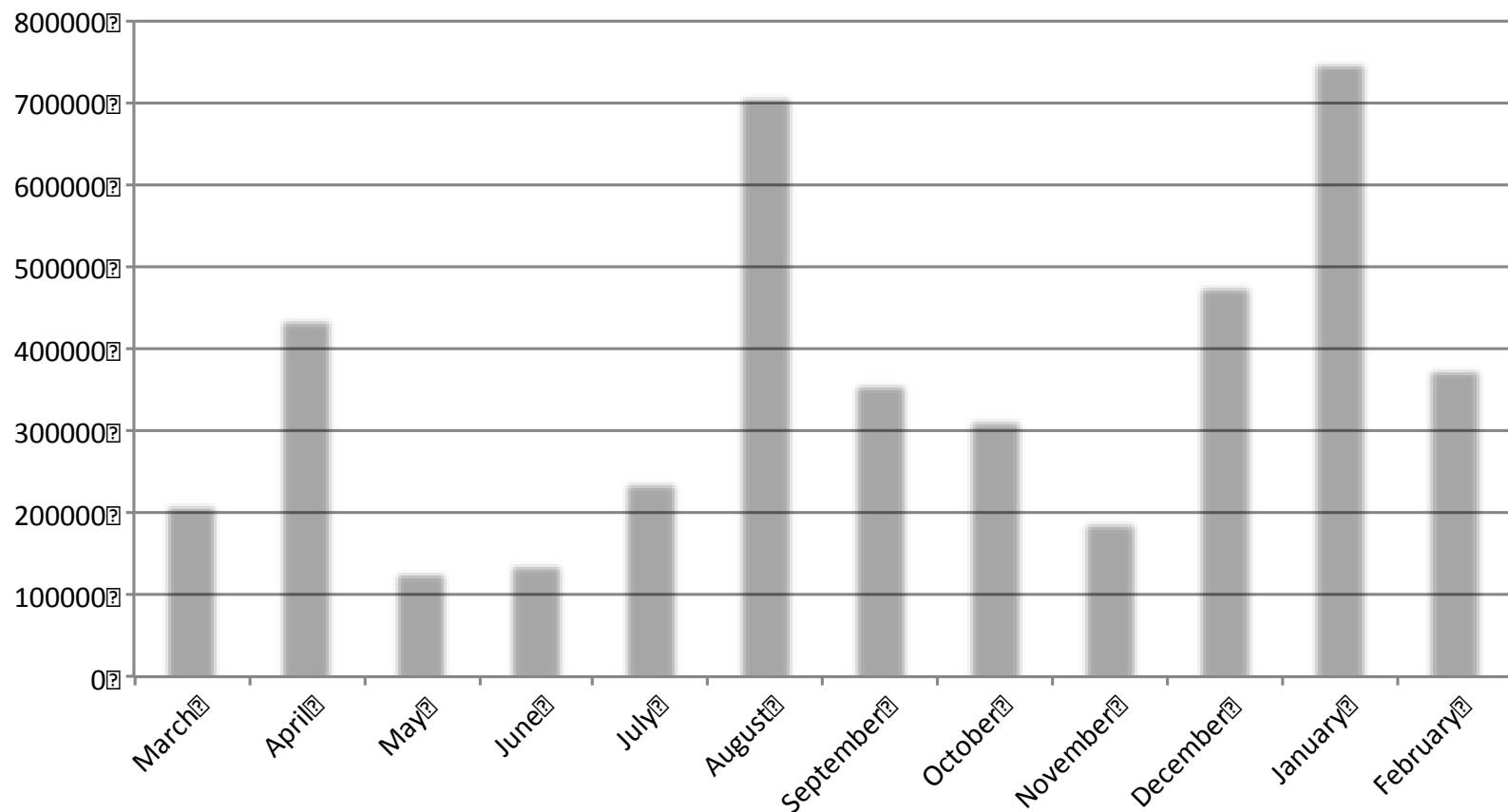
D. Sulakhe, A. Rodriguez, K. Bhuvaneshwar, Y. Gusev, R. Madduri, L. Lacinski, U. Dave, I. Foster, S. Madhavan

- 78 exomes from lung cancer study
- 2 TB data
- 125,936 core hours in 1.7 days



# Usage is high (and variable)

CPU Hours in the last 12 months





# Globus Genomics Pricing



[About Us](#) [Publications](#) [Technologies](#) [Sign Up](#)

## Pricing

As we are a non-profit entity, our offerings are priced to enable us to recover costs of providing Globus Genomics and for helping us sustain efforts to continue to support and enhance the underlying platform for the advancement of biomedical research.

We currently support numerous best-practice pipelines and allow researchers and core labs to modify, enhance and/or create their own custom pipelines for their genomics analysis needs. Actual pricing can vary based on several factors (e.g. complexity of the analysis pipeline, coverage, size of input data, duration of storage, volume of analysis).

Our pricing includes estimated compute, storage (one month), Globus Genomics platform usage, and technical support.

ACAAGATGCCATTGTCCCCGGGCTCTGCTGTCTGCTCTCCGGGGCCACGGGCCACCGCTGCCCTGCCCCCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCCGAGACAGCCGAGCATGAGGGTGTGCCCCACCGGCCGAGACAGC

### 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.



# Diversity of Collaborations



Seattle Children's  
HOSPITAL • RESEARCH • FOUNDATION



LABioMed

Los Angeles  
Biomedical  
Research Institute  
at Harbor-UCLA Medical Center

KU MEDICAL CENTER  
The University of Kansas



Washington  
University  
in St. Louis



Avera



THE UNIVERSITY OF  
CHICAGO

Cox Lab  
Volchenbom Lab  
Olopade Lab



Wexner Medical Center



INOVA®

Join the future of health.



GEORGETOWN UNIVERSITY



UNIVERSITY OF MINNESOTA



ngxbio



Boston University Medical Center

Genome  
Science  
Institute



PerkinElmer®  
For the Better



CEDARS-SINAI MEDICAL CENTER.



# Genomics and Clouds

Clouds are here to stay!








## NCI Cancer Genomics Cloud Pilots

***Bringing data and computation together to create knowledge that accelerates cancer research and enables precision medicine***

The traditional model for analyzing genomic data involves individual researchers downloading data stored at a variety of locations, adding their own data, attempting to harmonize the data, and then computing over these data on local hardware. This model has been successful for many years, but has become unsustainable given the enormous growth of biomedical data since the advent of large-scale

scientific programs that use next-generation sequencing technology. The size of the data makes access and analysis difficult for anyone but the best-resourced institutions, in terms of both storage and computing capability.

### Useful Links

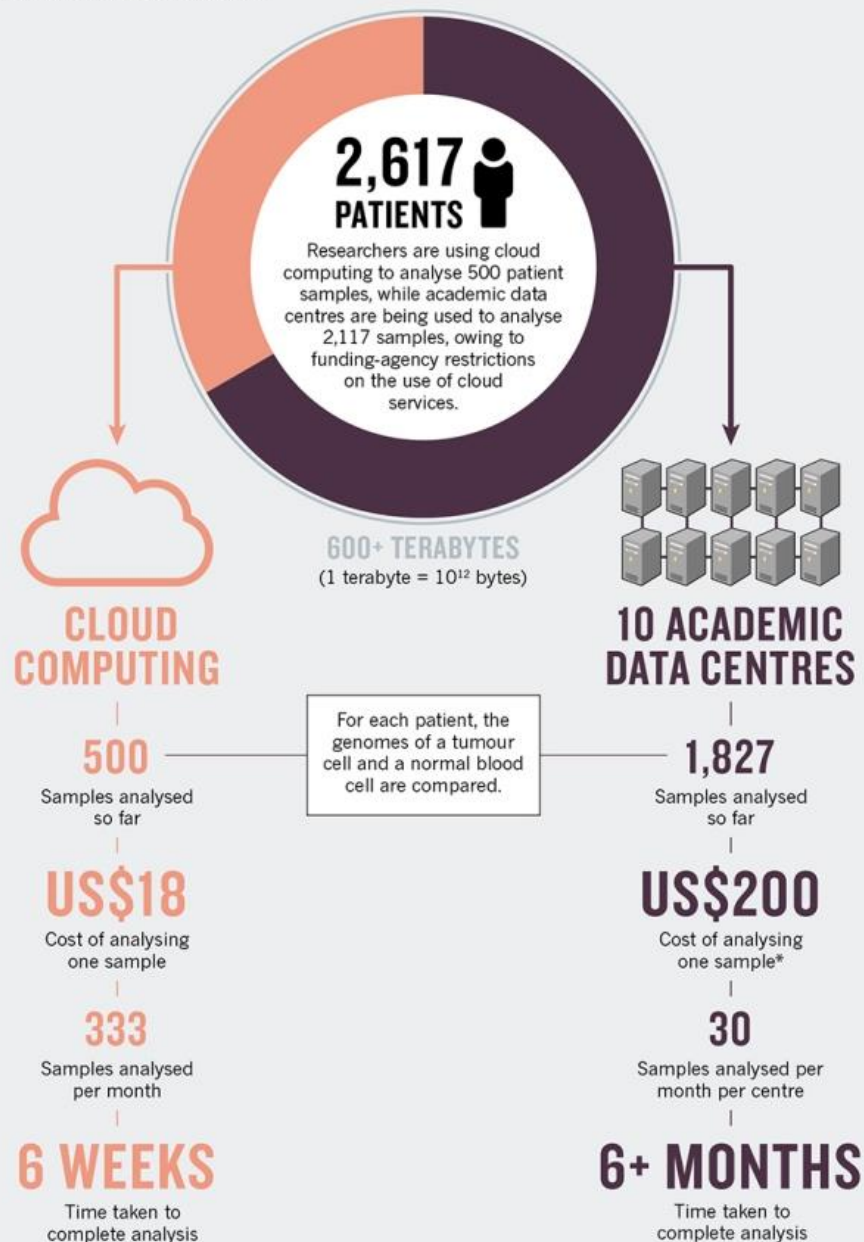
- [Broad Institute](#) 
- [Institute for Systems Biology](#) 
- [Seven Bridges Genomics](#) 
- [Tanja Davidsen, Ph.D.: NCI Cloud Pilots Report](#) 
- [NCI Center for Cancer Genomics](#)
- [ScienceLife: Transforming Cancer Research: The Genomic Data Commons](#) 
- [Cancer Genomics Cloud Pilots Fact Sheet](#)





## EXPRESS LANE

The Pan Cancer Analysis of Whole Genomes project (in which L.D.S., P.C., G.G. and J.O.K. are involved), an effort to investigate the role of non-coding parts of the genome in cancer, demonstrates how much faster and cheaper it is to use cloud computing than to use conventional academic data centres when analysing vast biological data sets.



\*If using a standard university computer system and buying the hardware.



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
<b>Cumulative Analysis Workload*</b> (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
<b>Technical Support</b>	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
<b>Access to Enhanced Workbench</b>	Yes	Yes	Yes
<b>Multi-sample submission</b>	Yes	Yes	Yes
<b>Usage Dashboard</b>	Yes	Yes	Yes
<b>Price/Performance Controls</b>	Basic	Advanced	Advanced
<b>On-Demand Tool Wrapping</b>	No	Limited	Yes
<b>HIPAA / optional BAA</b>	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.*



- More information on Globus Genomics and to sign up for a **free** trial :  
[www.globus.org/genomics](http://www.globus.org/genomics)
- More information on Globus:  
[www.globus.org](http://www.globus.org)



# Thank you to our sponsors!



U.S. DEPARTMENT OF  
**ENERGY**



THE UNIVERSITY OF  
**CHICAGO**

**Argonne**  
NATIONAL LABORATORY



powered by  
**amazon**  
web services