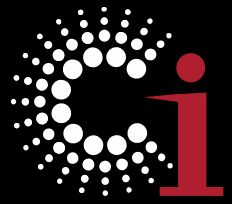


Lessons Learned Analyzing Thousands of Samples for Clinical Use Cases Using Amazon Web Services

Ravi Madduri

Argonne National Laboratory, University of Chicago

madduri@anl.gov

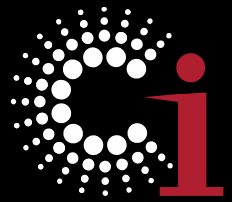
BioIT 2016

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

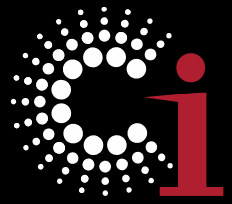
Galaxy

globus.org/genomics



Outline

- Introduction (who we are) – 2 mins
- Setting up the problem
  - Scale
  - Cost
  - Compliance
- Lessons learned

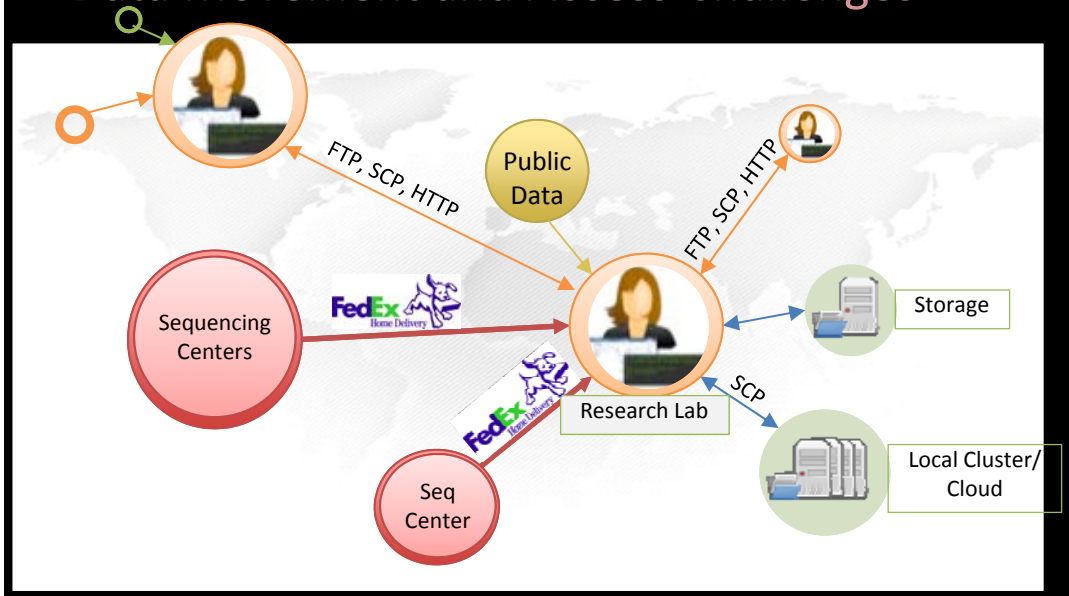


Who We Are

- Globus is developed, operated, and supported by researchers, developers, and bioinformaticians at the Computation Institute – University of Chicago/Argonne National Lab
- We are a non-profit organization building solutions for non-profit researchers
- Our goal is to support the advancement of science by bringing together our strengths and capabilities to help meet the unique needs of researchers and research institutions

Challenges In Large Scale NGS Analysis

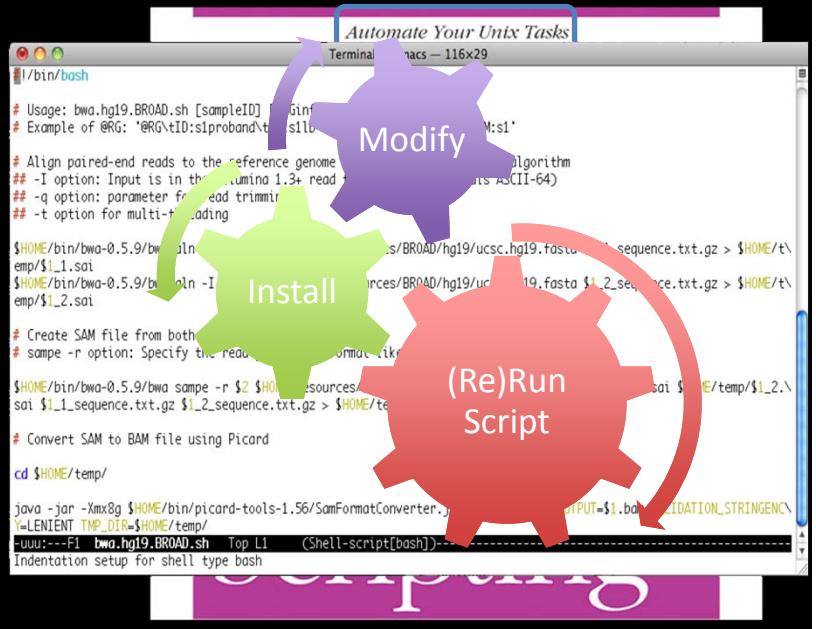
Data Movement and Access Challenges



- 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

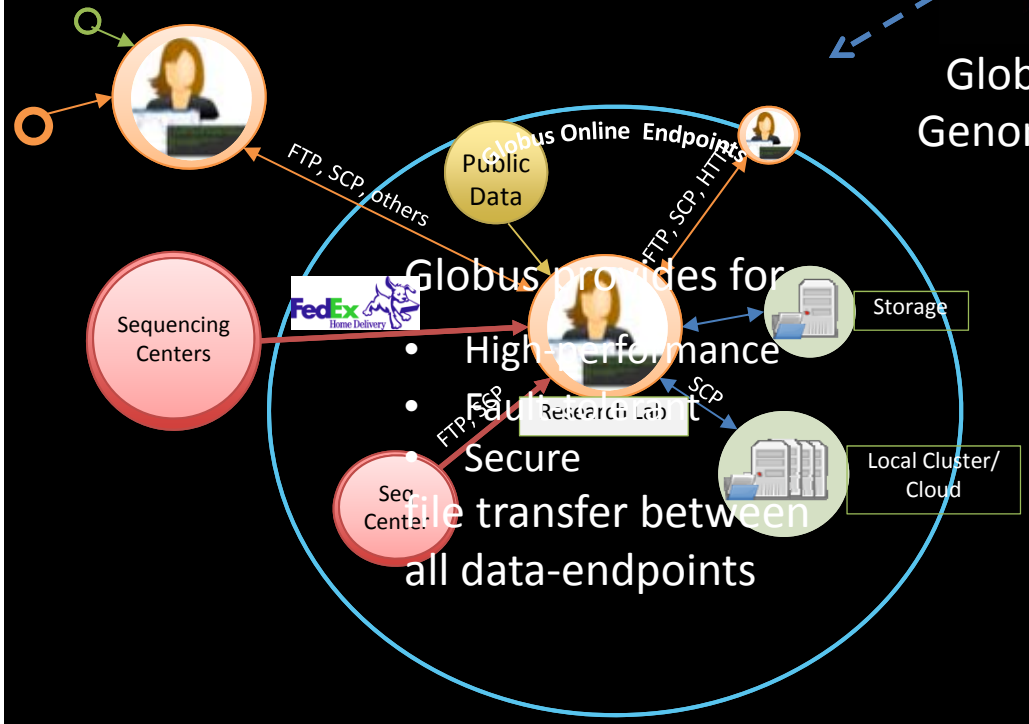
- Manually move the data to the Compute node
- Install all the tools required for the Analysis
  - BWA, Picard, GATK, Filtering Scripts, etc.
- Shell scripts to sequentially execute the tools
- Manually modify the scripts for any change
  - Error Prone, difficult to keep track, messy..
- Difficult to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



Manual Data Analysis

globus.org/genomics

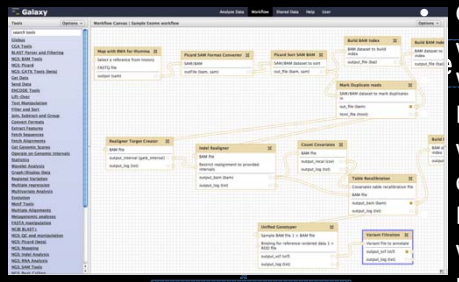
Globus Genomics



Data management

Globus Genomics

Galaxy-based workflow management



Galaxy Data Libraries

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

Globus Genomics on Amazon EC2

Data analysis

globus.org/genomics

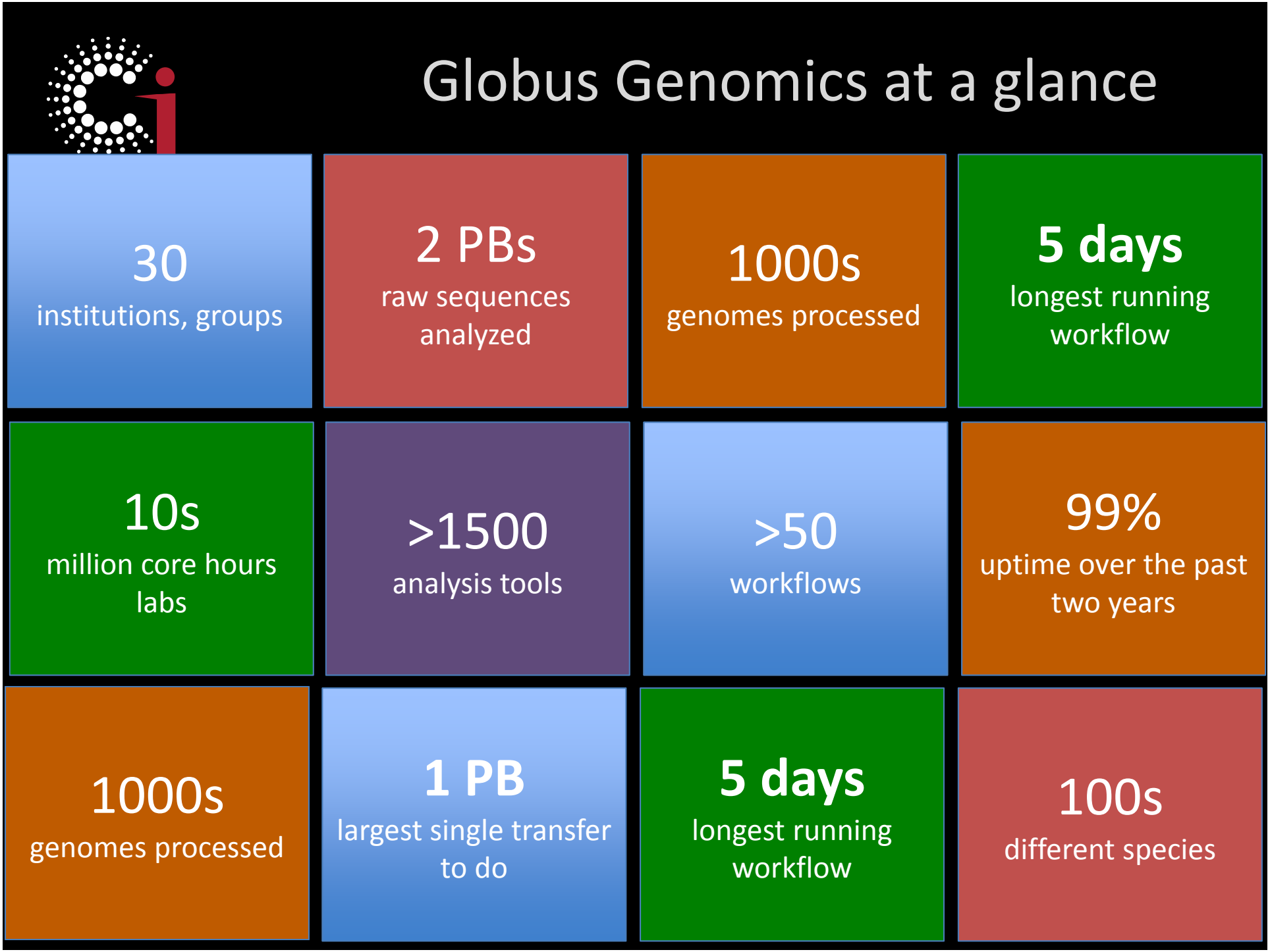
Globus Genomics

- 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

globus.org/genomics

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


Globus Genomics at a glance
30
institutions, groups
2 PBs
raw sequences analyzed
1000s
genomes processed
5 days
longest running workflow
10s
million core hours labs
>1500
analysis tools
>50
workflows
99%
uptime over the past two years
1000s
genomes processed
1 PB
largest single transfer to do
5 days
longest running workflow
100s
different species

Typical Usecases

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

Vassily Trubetskoy[1], Ravi Madduri[2], Alex Rodriguez[2], Jeremiah Scharf[3], Paul Dave[2], Ian Foster[2] , Nancy Cox[1] , Lea Davis[1]

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

A profile of inherited predisposition to breast cancer among Nigerian women

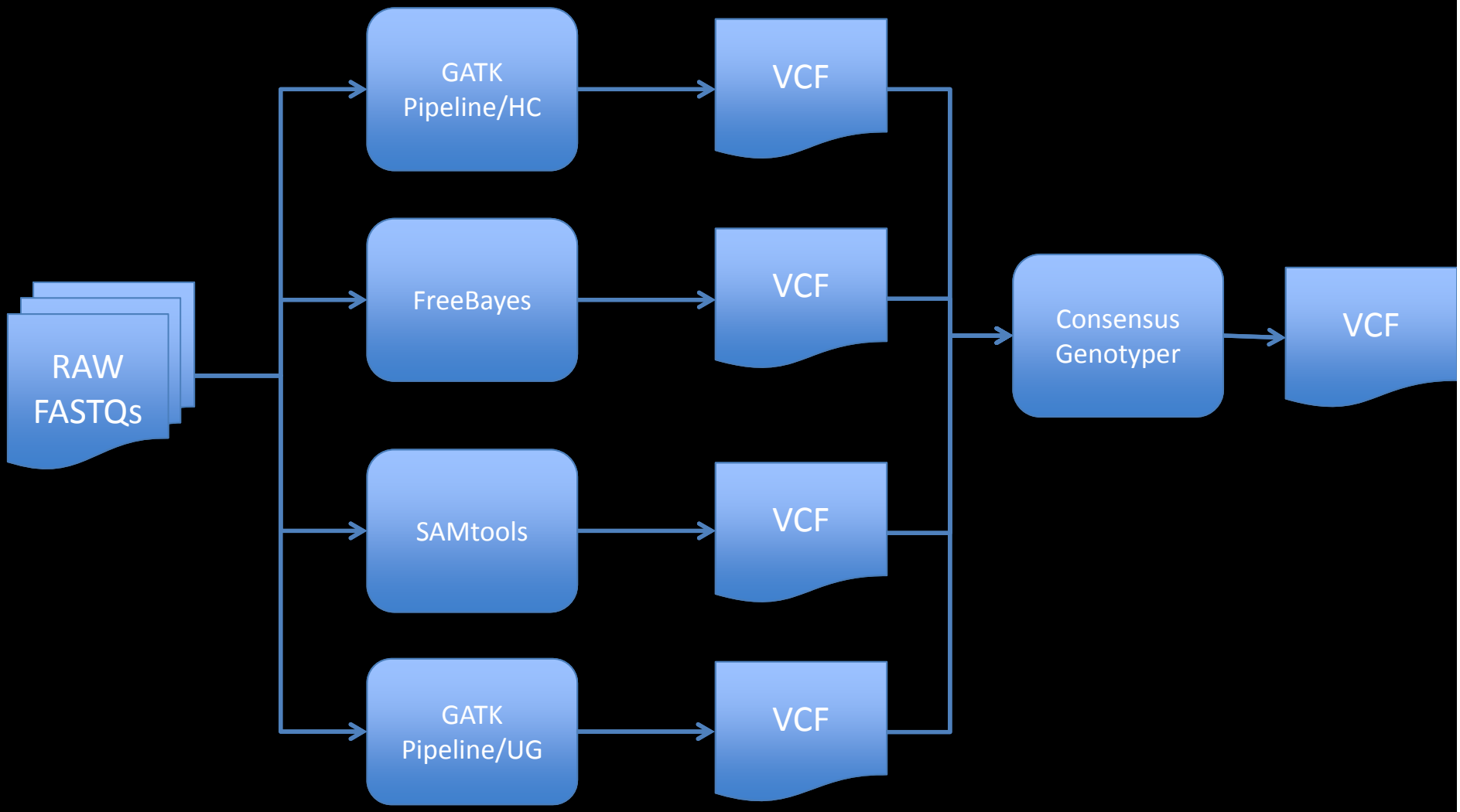
Y. Zheng, T. Walsh, F. Yoshimatsu, M. Lee, S. Gulsuner, S. Casadei, A. Rodriguez, T. Ogundiran, C. Babalola, O. Ojengbede, D. Sighoko, R. Madduri, M.-C. King, O. Olopade

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

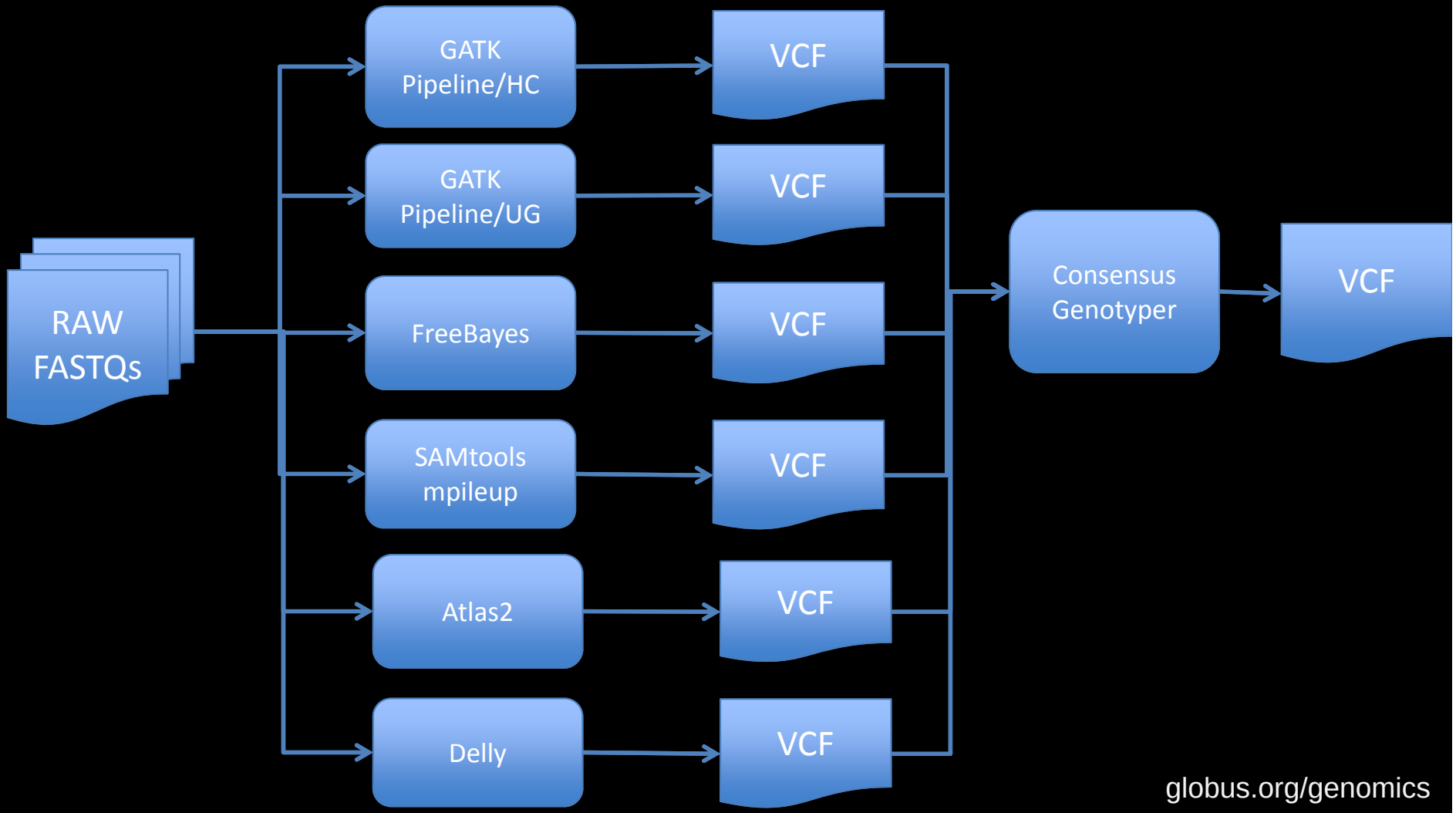
D. Sulakhe, A. Rodriguez, K. Bhuvaneshwar, Y. Gusev, R. Madduri, L. Lacinski, U. Dave, I. Foster, S. Madhavan

globus.org/genomics

How we realize them?



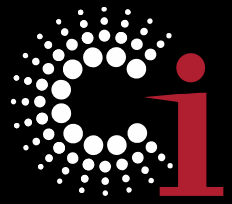
How we realize them?



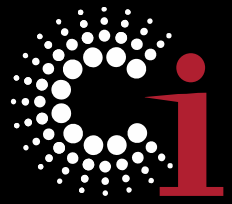
globus.org/genomics

Early Results

14 deleterious SNVs and 11 damaging Indels (BRCA1: 15, BRCA2: 4, PALB2: 2, BRIP1: 1, CHEK2: 1, NBN: 1, TP53: 1) were found in 29 subjects, and they were all confidently detected among 5 callers. Identified SNVs and Indels were all confirmed by Sanger sequencing.



Now we want to run these pipelines on 1000s of samples on a budget



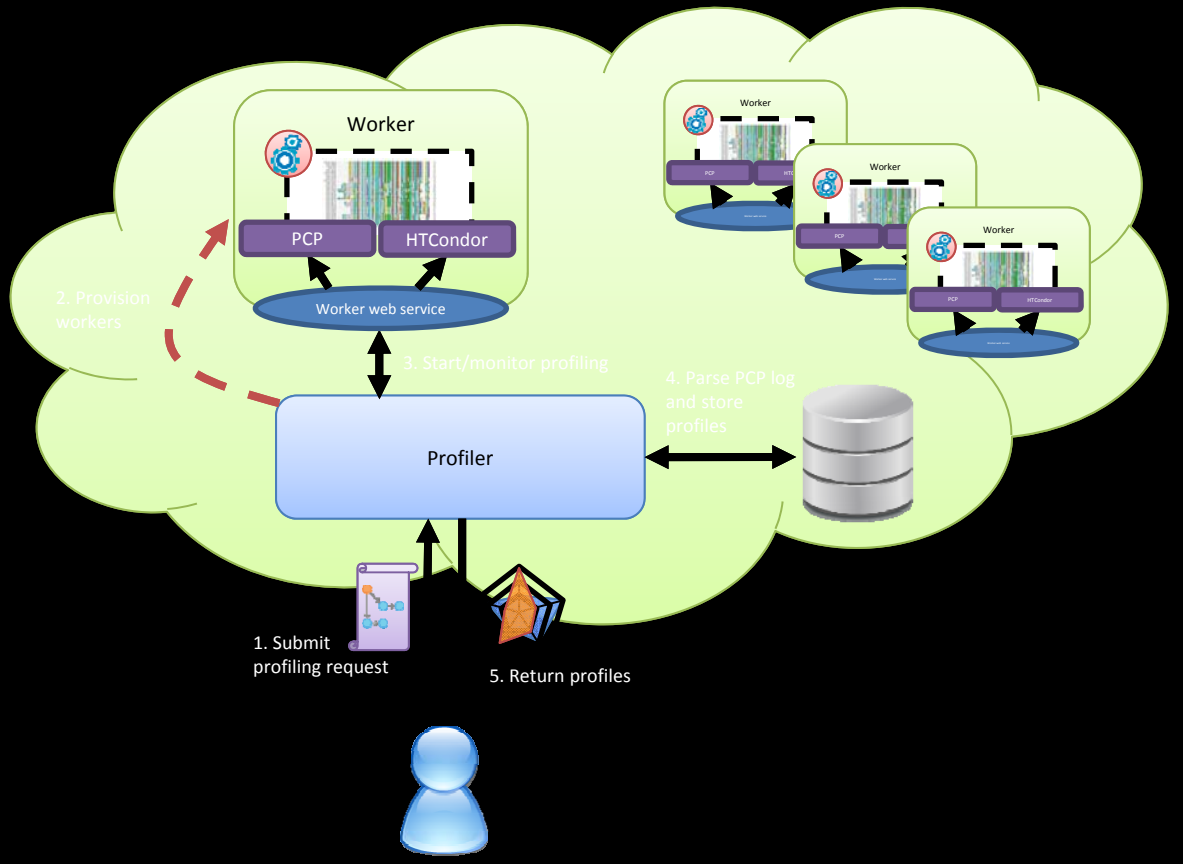
We built two important capabilities

Application Profiling Service

Cost-Aware Scheduling

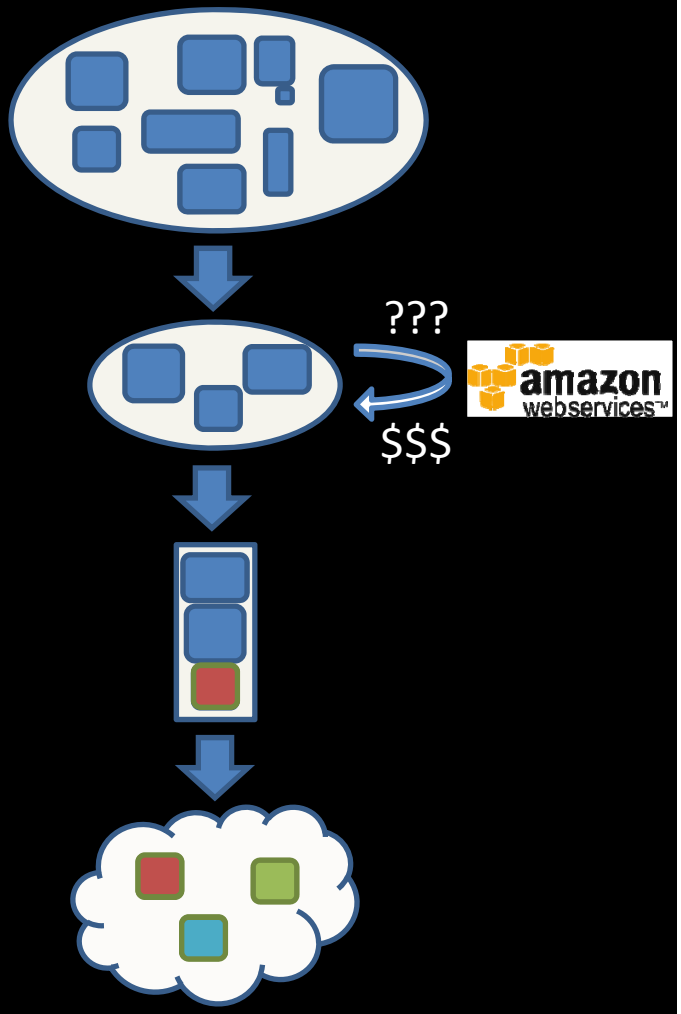
A Cloud Tool Profiling Service

- Describe profile requests in JSON
- Provision resources and apply a profiling Web Service
- Use Performance Co-pilot (PCP) to capture usage
- Capture and process PCP logs
- Return profiles as JSON (or logs via s3)

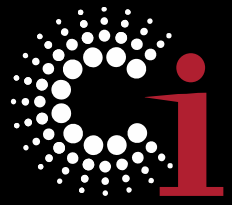


Cost-aware Provisioning

1. Filter instance types with profiles
2. Determine price for each instance type across all AZs
3. Rank potential requests
4. Make requests and monitor
5. Cancel or repurpose excess active requests once one is fulfilled



R. Chard et al. Cost-aware cloud provisioning,
Proceedings of the 11th IEEE International Conference on e-Science (e-Science), 2015.



Now we want to run these pipelines on 1000s of samples on a budget in a clinical setting

What's Different

Research

- Modular
- Flexible parameters
- Applications
  - RNA-Seq
  - Exome-seq (Whole and Targeted)
  - Whole-genome sequencing
  - Methylation sequence
- Input size
  - 1GB - 500GB per sample
- Sample quantity and parallelization level
  - Hundreds
- Time to treatment
  - No rush needed
- Not CLIA-CAP compatible
- AWS resource
  - Spot instances

Clinical

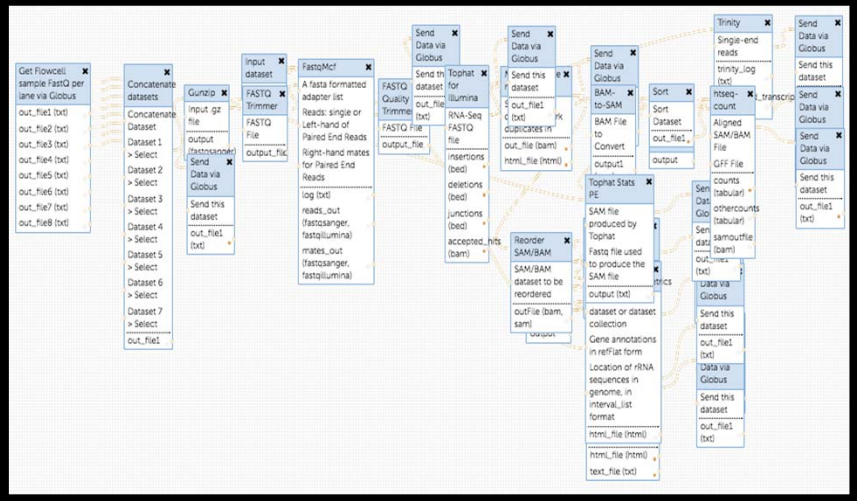
- Black box
- Fixed parameters
- Applications
  - Exome-seq (Target array panels)
  - .
  - .
  - .
- Input size
  - <1GB per sample
- Sample quantity and parallelization level
  - Thousands
- Time to treatment
  - As fast as possible
- CLIA-CAP compatible
- AWS resource
  - Dedicated instances

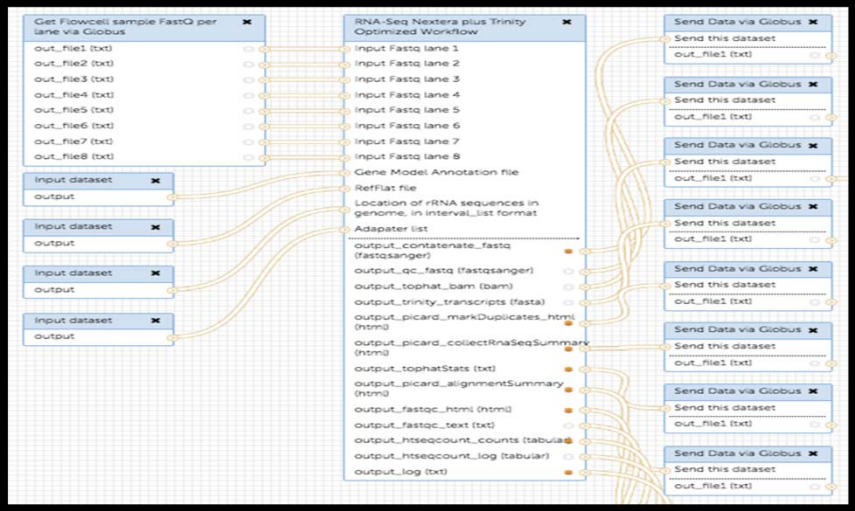
What's Different

Research

vs

Clinical





What can be modified?

- Input files
- Output files
- Parameters
- Tools and tool versions
- Reference genomes

What can be modified?

- Input files

How can this be done?

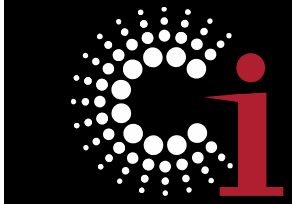
- One step process to convert a resear based workflow to an optimized (CLIA CAP) workflow

globus.org/genomics

Clinical workflows - challenges

- Creating Patient cohorts is still a big challenge
  - Data Silos
- Variant filtration and prioritization
- PDF reports
  - Lack of integration into EMRs
  - Lack of well-adopted standards for genomics data models

Other Globus Genomics users



Dobyns Lab

Seattle Children's
HOSPITAL • RESEARCH • FOUNDATION

Nagarajan Lab

Washington University in St. Louis

Wexner Medical Center

UPMC
University of Pittsburgh Medical Center

THE UNIVERSITY OF CHICAGO

Cox Lab
Volchenboum Lab
Olopade Lab

BD

INOVA
Join the future of health.

Avera

GEORGETOWN UNIVERSITY

LA BioMed
Los Angeles Biomedical Research Institute at Harbor-UCLA Medical

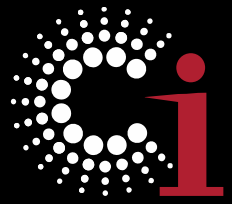
UCSF

Genome Science Institute

Boston University Medical Center

PerkinElmer
For the Better

GEORGETOWN UNIVERSITY



- More information on Globus Genomics:www.globus.org/genomics
- More information on Globus: www.globus.org

Our work is supported by:

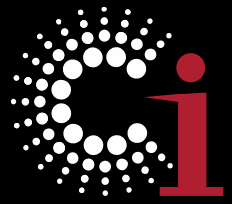
U.S. DEPARTMENT OF
ENERGY

THE UNIVERSITY OF
CHICAGO

Argonne
NATIONAL LABORATORY

amazon
web services

globus.org/genomics



Thank you!

@madduri