Computation Institute

Globus Galaxies: Experiences in building *sustainable* services for Science

Ravi K Madduri and Globus Genomics Team
University of Chicago and Argonne National Laboratory
@madduri

THE UNIVERSITY OF CHICAGO

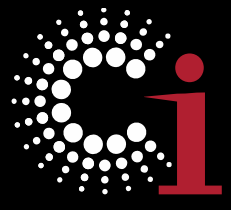
Argonne NATIONAL LABORATORY

Our work is supported by:

U.S. DEPARTMENT OF ENERGY

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY



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

Computation Institute

Trying to find a single causative *gene* for diseases with a complex genetic background is like looking for the proverbial *needle* in a haystack

– Nancy Cox
(Vanderbilt)

THE UNIVERSITY OF CHICAGO

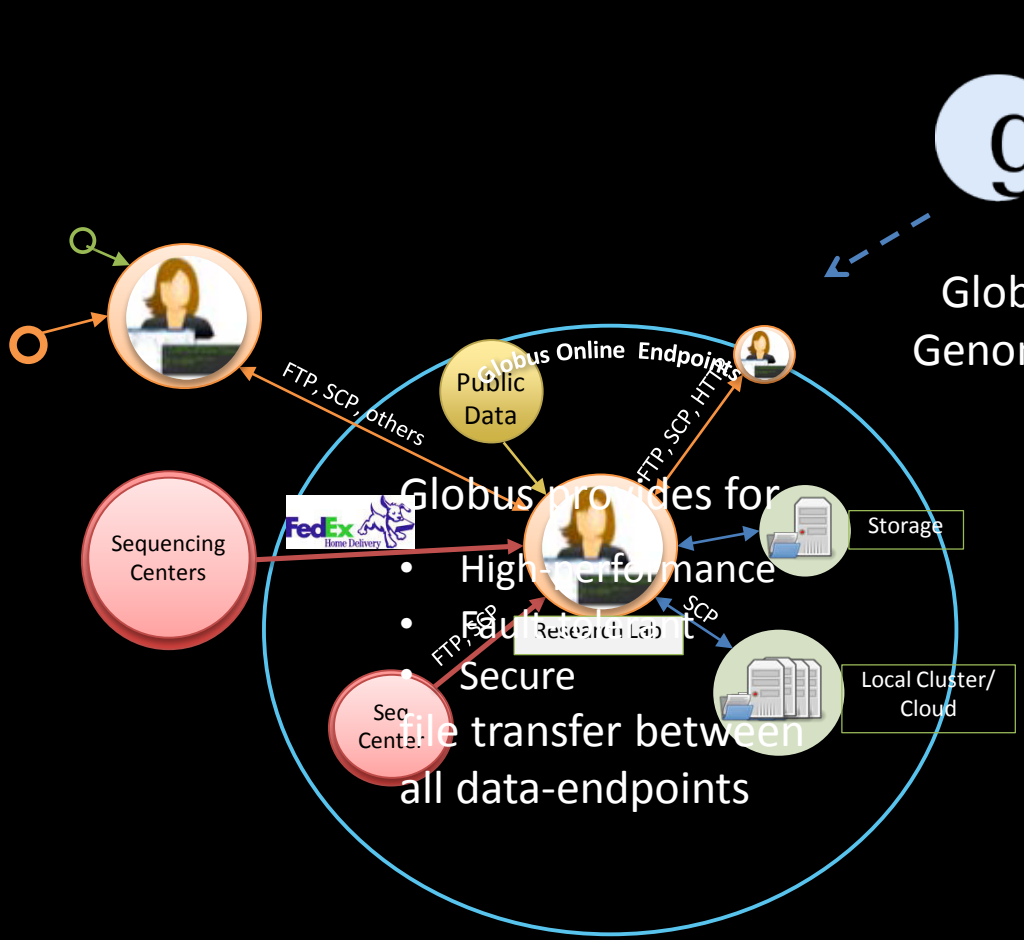
Argonne NATIONAL LABORATORY

computationinstitute.org

How do we accelerate discovery without requiring that every lab acquire a haystack-sorting machine?

Clayton & Shuttleworth thresher, 1910. Museum Victoria, Australia

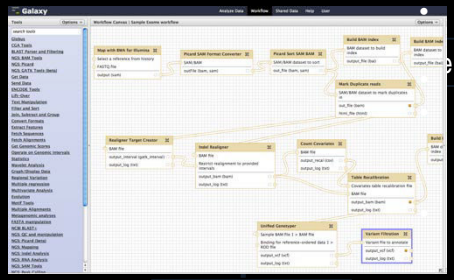
Our answer: Globus Genomics



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 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

Globus Genomics users

Cox lab, UChicago

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

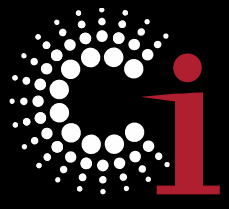
- 134 samples and 4 workflows
- 4 TB data
- 2200 core hours in 6 days

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

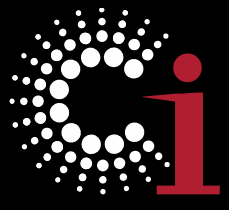


Innovation Center for Biomedical Informatics - Georgetown

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

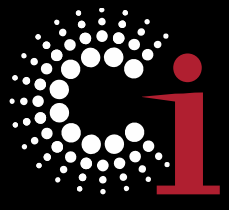
D. Sulakhe, A. Rodriguez, K. Bhuvaneshwar, Y. Gusev, R. Madduri, L. Lacinski, U. Dave, I. Foster, S. Madhavan

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



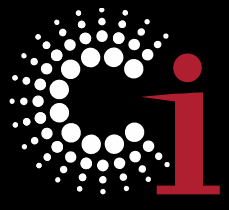
Building a Sustainable Science Service

Globus Genomics



Our Experience in One Slide

- Reverse the model – From a Grant-funded infrastructure model to a Service Provider model with subscription revenue for sustainability
  - Incentivize *value* creation for users
- Create a "Stack"
- Leverage Industry best-practices to deploy, manage, upgrade
- Professional Support



Our Science Stack

- Galaxy
  - Interactive execution
  - Creation, Execution, Sharing, Discovering Workflows
- Globus
  - Data management
  - Identity Management
- AWS
  - HTCondor, Chef, EC2, EBS, S3, SNS
  - Spot, Route 53, Cloud Formation

SaaS

PaaS

IaaS

Decisions.. Decisions..

- Things to consider for an “Informatics Core”
  - Build Vs Buy
  - Total Cost of Ownership
  - Available Expertise
  - ROI
  - Long Term Sustainability
  - Core Competencies
  - Support
  - Compliance Requirements

Subscription Pricing

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

Costs are remarkably low

Exome	Whole Genome	RNA-Seq.
$5 - $30	$20 - $100	$5 - $10
Pricing based on example of paired-end fastq files with 5 Gbases.	Pricing based on example of paired-end fastq files with 80 Gbases.	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.	Pipeline includes quality control, alignment, variant calling, and annotation.	Pipeline includes quality control, alignment, exon count using cufflinks, and HT-Seq count.

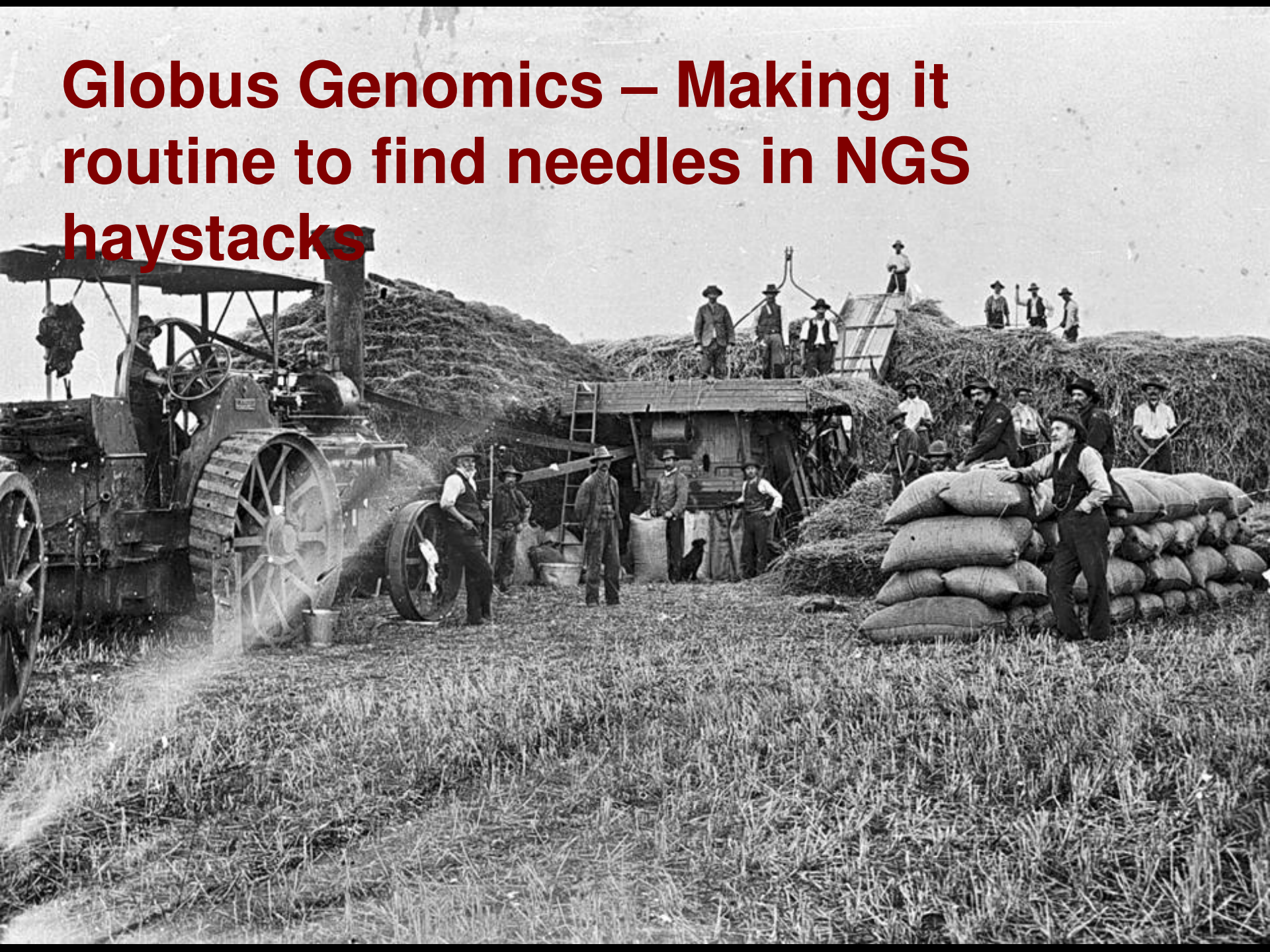
Pricing includes

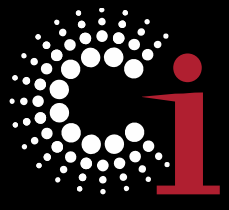
- Estimated compute
- Storage (one month)
- Globus Genomics platform usage
- Support

A Reminder About Us

- Globus Genomics 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

Globus Genomics – Making it routine to find needles in NGS haystacks





Thank you!

@madduri