Experiences In Building Globus Genomics Using Galaxy, Globus Online and AWS

Ravi K Madduri

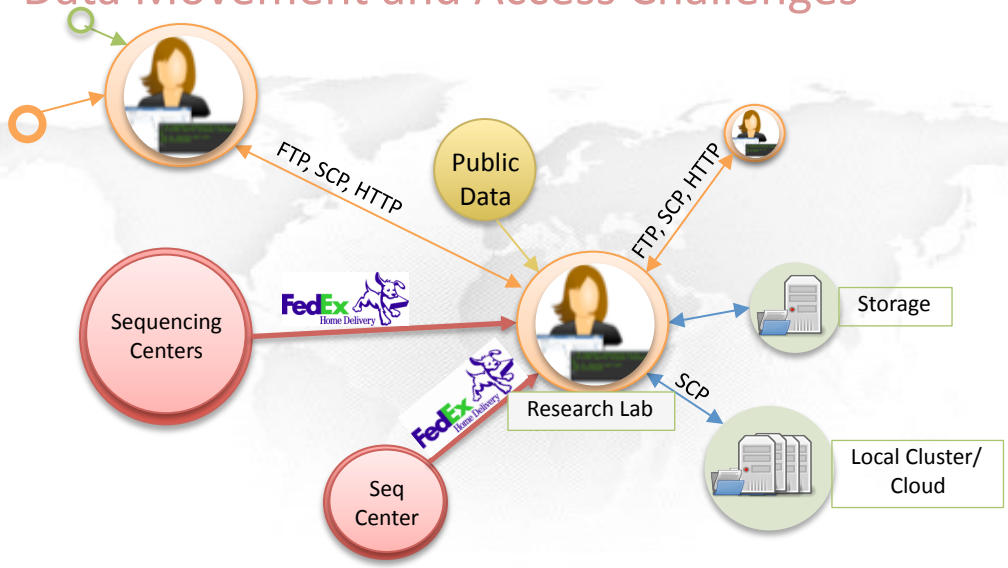
University of Chicago and ANL

Outline

- Challenges in Sequencing Analysis
- Proposed Approach Using Globus Genomics
- Example Collaborations
- Relevance to XSEDE
- Q&A

Challenges in Sequencing Analysis

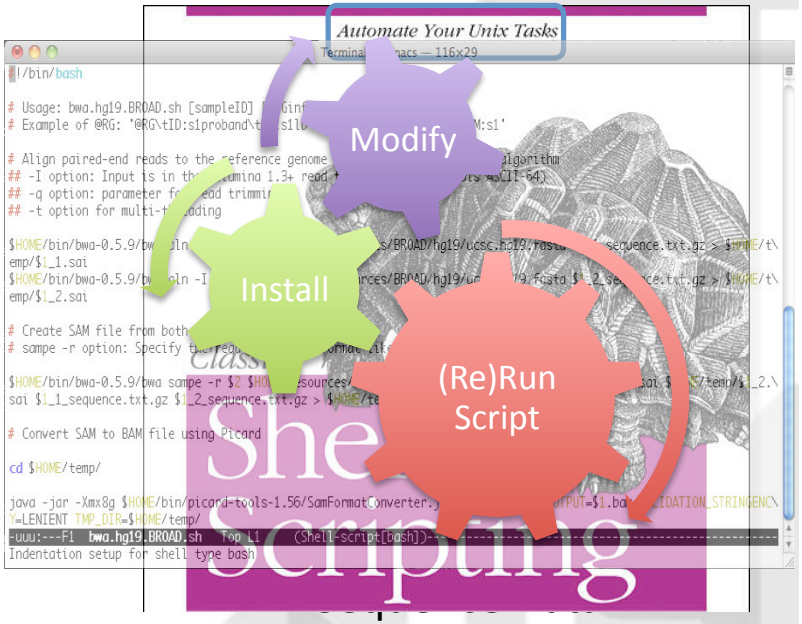
Data Movement and Access Challenges



- 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

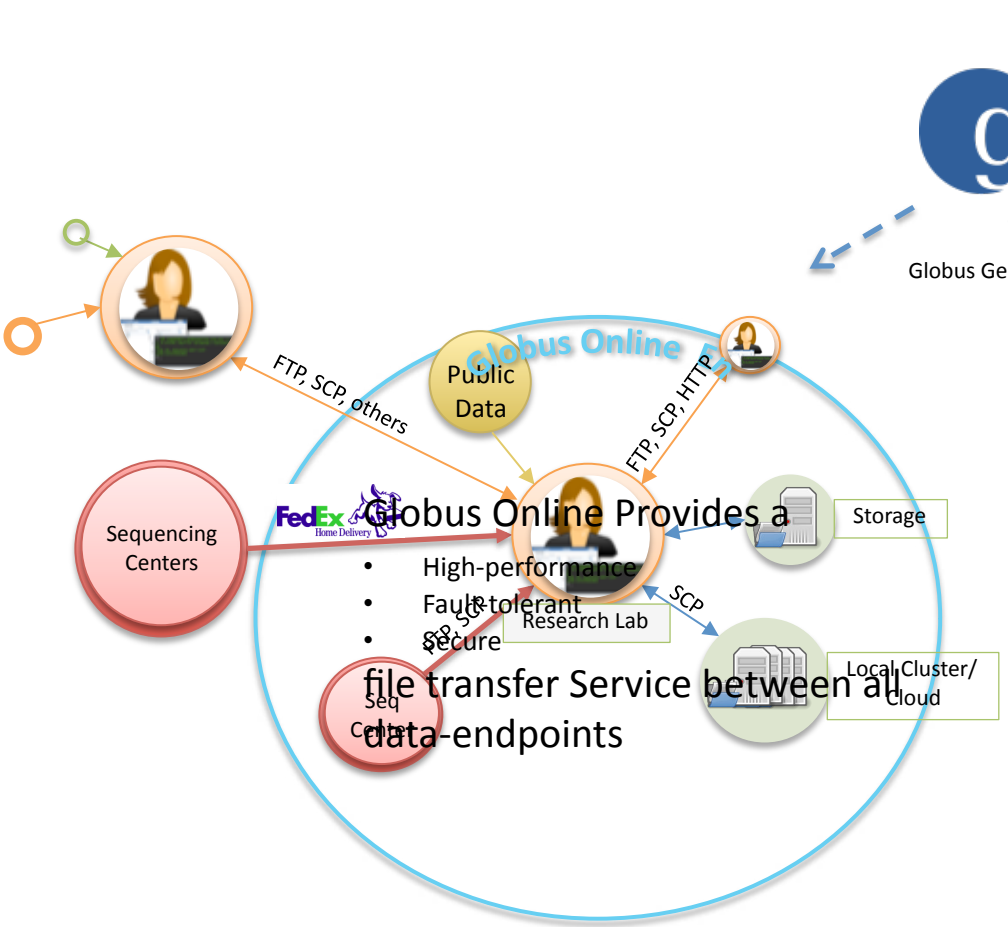
Once we have the Sequence Data

- 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

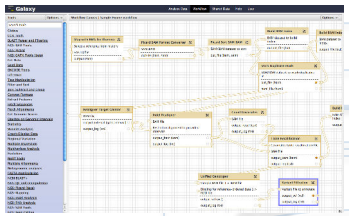


Manual Data Analysis

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 Data Libraries

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

Globus Genomics on Amazon EC2

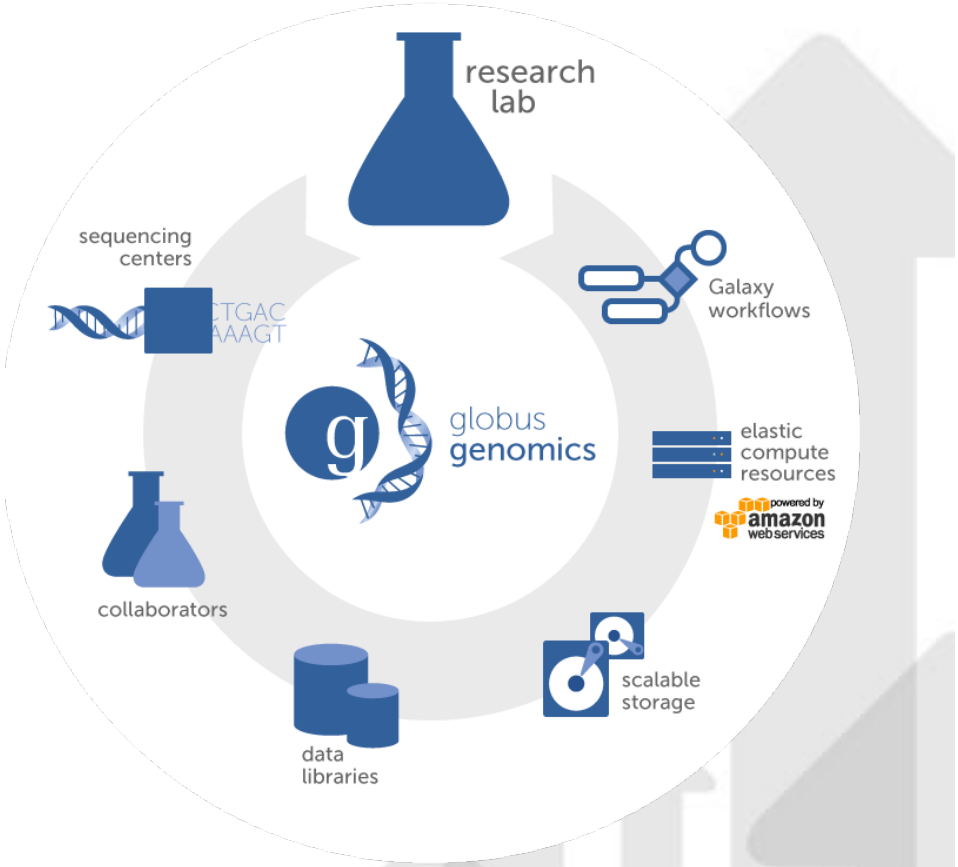
Data Management

Data Analysis

www.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



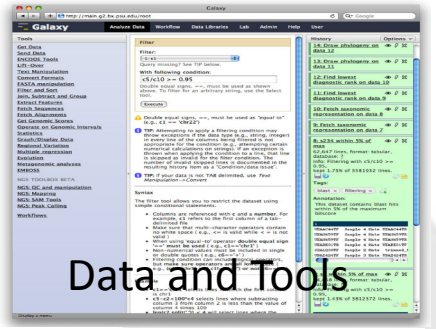
Additional Capabilities

- Professionally managed and supported platform
- 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

Globus Genomics – A flexible, scalable, simplified analysis platform

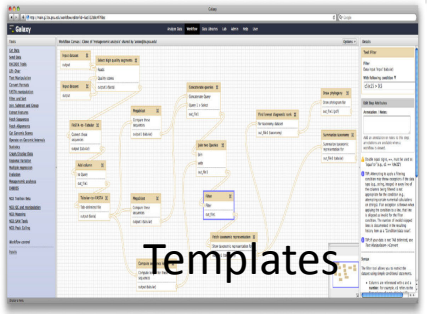
Accessibility

- 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 (History Panel)



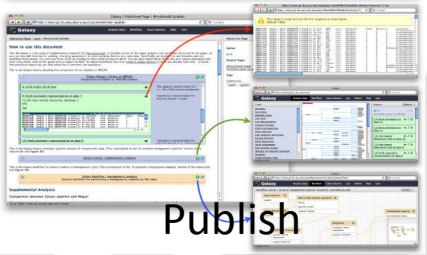
Reproducibility

- 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



Transparency

- 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



Globus Online Integration

- Access GO Endpoints and transfer data from within Galaxy UI and into Galaxy workspace
- Leverage local cluster or cloud based scalable computational resources for parallelizing the tools

www.globus.org/genomics

Example Collaborations

Dobyns Lab

UNIVERSITY OF WASHINGTON

Backround: 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

Relevance to XSEDE

XSEDE's Mission Statement

accelerat[ing] open scientific discovery by enhancing the productivity of researchers, engineers, and scholars and making advanced digital resources easier to use."

Key XSEDE Goals That Globus Genomics Addresses

- "Deepen and extend the impact of eScience infrastructure on research and education; in particular, to reach communities that have not previously made use of it; and
- Expand the environment through the integration of new capabilities and resources such as instruments and data repositories based on the identified needs of the community."

Relevance to XSEDE (Cont..)

- **Globus Genomics leverages an XSEDE service**
  - Globus Transfer for data movement
  - Globus Nexus for identity management
  - Globus Groups for group-based access management
- **Integrates advanced digital resources**
  - sequencing centers, a commercial cloud provider, and NGS analysis pipelines
- **Reduces the cost and complexity of scientific discovery for a new community (NGS researchers) who have not historically made much use of advanced eScience infrastructures.**

XSEDE Vs AWS

- **Globus Genomics achieves these goals without making use of XSEDE supercomputers**
- **Choice to use Amazon cloud services rather than XSEDE systems for Globus Genomics computations is deliberate**
  - scales at which our target users operate today, the costs associated with the use of Amazon cloud computers are modest, and Amazon's on-demand, pay-as-you-go storage and computing capabilities match user needs better than the proposal- and queue-based access policies provided by XSEDE computers.
- **We plan to explore using XSEDE resources to execute Globus Genomics pipelines**

Acknowledgments

- This work was supported in part by the NIH through the NHLBI grant: The Cardiovascular Research Grid (R24HL085343) and by the U.S. Department of Energy under contract DE-AC02-06CH11357. We are grateful to Amazon, Inc., for an award of Amazon Web Services time that facilitated early experiments.
- The Globus Genomics and Globus Online teams at University of Chicago and Argonne National Laboratory

For more information

- More information on Globus Genomics and to sign up: www.globus.org/genomics
- More information on Globus Online: www.globusonline.org
- Questions?
- Thank you!