Implementing a High Performance, Reusable Consensus Calling Pipeline for Next Generation Sequencing using Globus Genomics

RK Madduri[1,2], A Rodriguez[1], V Trubetskoy[3], LK Davis[12], PJ Dave[1], NJ Cox[3], IT Foster[1,2]

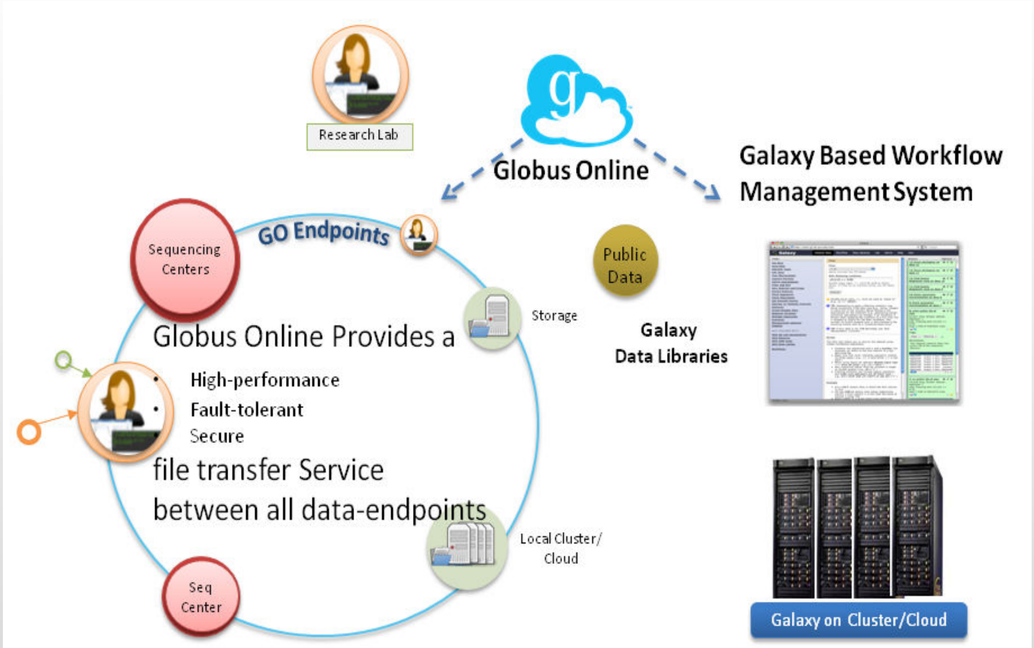
[1]Computation Institute, University of Chicago, Chicago, IL, USA. [2]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA.
[3] Section Genetic Medicine, University of Chicago, Chicago, IL.

Challenges in Next-Gen Sequencing Analysis

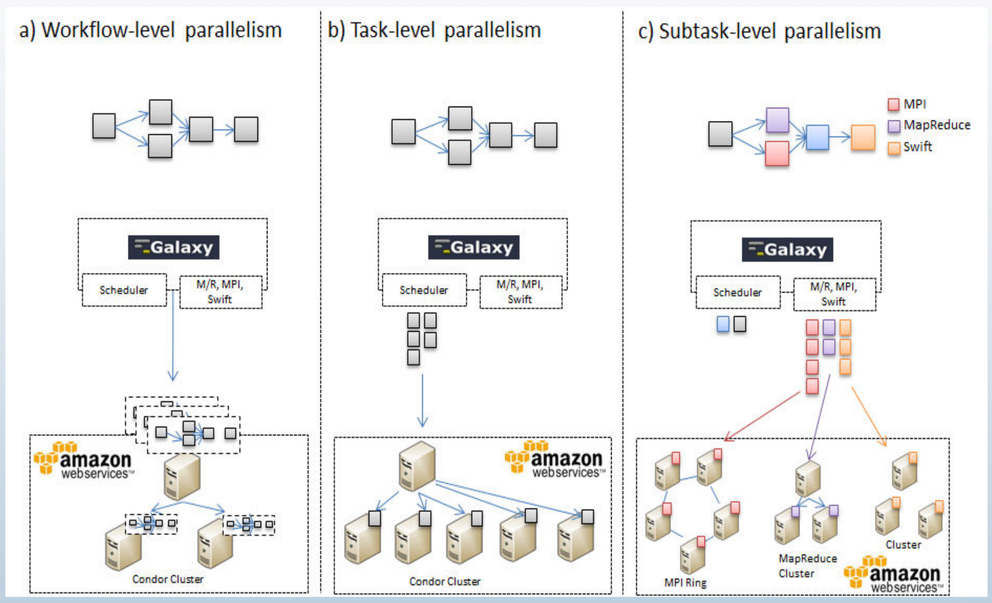
The requirements of translational research are unique and daunting: massive data, complex software, limited budgets, and demand for increased collaboration.

While "the Cloud" promises to alleviate some of these pressures, concerns about feasibility still exist for scientists and their institutional computing facilities.

- Access to distributed data: sequencing centers, collaborators, storage archival
- Share data with other researchers/collaborators
- Inefficient ways of data movement (NGS data shipped on disks)
- Data needs to be available on the local and Distributed Compute Resources (Clusters, Cloud, Grid)
- Scalable compute resources for handling growing NGS data
- Reproducible analysis in the form of workflows
- Dedicated IT support to handle ever growing data and analytical tools

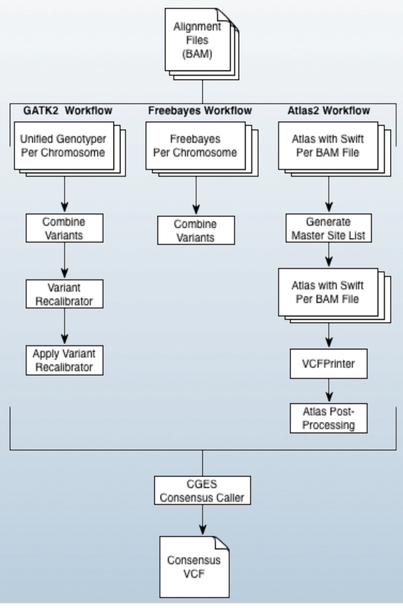


Parallel Workflows on Globus Genomics



High Performance, Reusable Consensus Calling Pipeline

- We built three highly parallelizable pipelines for variant calling analysis using best practices guidelines.
- Very little user input required once the workflow is set up and submitted.

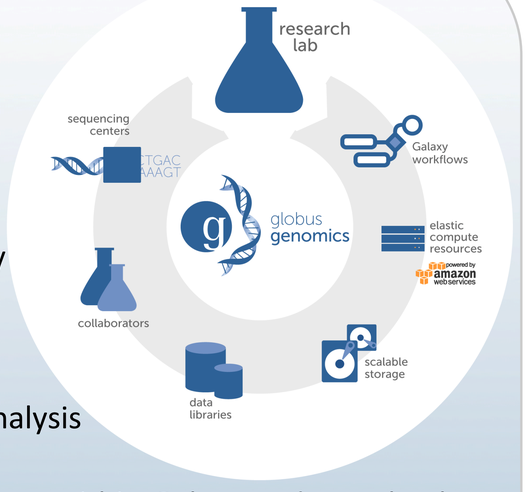


Results and Performance Details

- Over 1TB of data transferred seamlessly using Globus Online into the Globus Genomics instance which included 134 BAM files from the Autism data.
- GATK2 pipeline took about 21 hours in wall time and 240 CPU hours.
- Freebayes pipeline took about 31 hours in wall time and 720 hours in CPU time.
- Atlas2 pipeline took about 111 hours in wall time and 603 days in CPU time.
- Consensus caller generated a high quality list of variants (less than 0.01% mendel error error rate) from the three variant caller pipelines

Benefits of Globus Genomics

- Globus Genomics is an open Web-based platform for NGS research
  - Intuitive UI for Workflow creation
  - Easily integrate your own tools and scripts for analysis
  - Provides Reproducibility and Transparency
  - Shared datasets and workflows can be imported by other users for reuse
  - Uses Amazon Web Services to provide scalable compute resources for parallel analysis
- Globus Online Integration:
  - Access GO Endpoints and transfer data from within Galaxy and seamlessly move data from sequencing centers and other sources to Globus Genomics on Amazon AWS



The Computation Institute was established in 2000 to advance the study and application of computation across the University of Chicago and Argonne National Laboratory.

Its fellows — who hail from the bioinformatics, chemistry, computer science, digital media, ecology, and many other disciplines—are united by a common focus on the most challenging problems in disciplinary and interdisciplinary scholarship.

Web Resources:

- Globus Genomics: www.globus.org/genomics
- Globus Online: www.globusonline.org
- Galaxy: https://main.g2.bx.psu.edu/