Applying ConVarCal, A High Performance Consensus Variant Calling, For Discovery of Prevalence of Inherited Mutations in Breast Cancer Predisposition Genes among Uganda and Cameroon Women

2019 ITCR Annual Meeting

Ravi Madduri
Computational Scientist
Argonne National Laboratory and University of Chicago

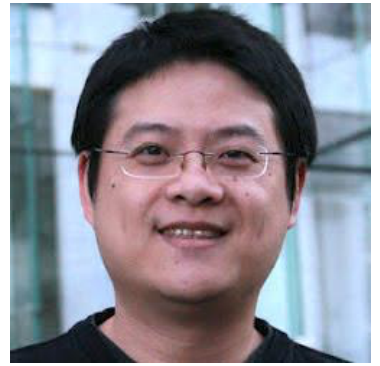
Outline

- **Acknowledgments**
- **Background**
- **Results from ConVarCal**
- **How did we get the results ?**
  - Step 1: data transfer using Globus
  - Step 2: Building and validating the pipeline
  - Step 3: Scaling up the analysis using parallel computing and cost-aware scheduling Globus Genomics
- **New Capabilities: Globus for managing PHI data**
- **Q&A**

Acknowledgements

Center for Clinical Cancer Genetics and Global Health, Department of Medicine, University of Chicago



Globus at University of Chicago, Data Science and Learning at Argonne National Laboratory





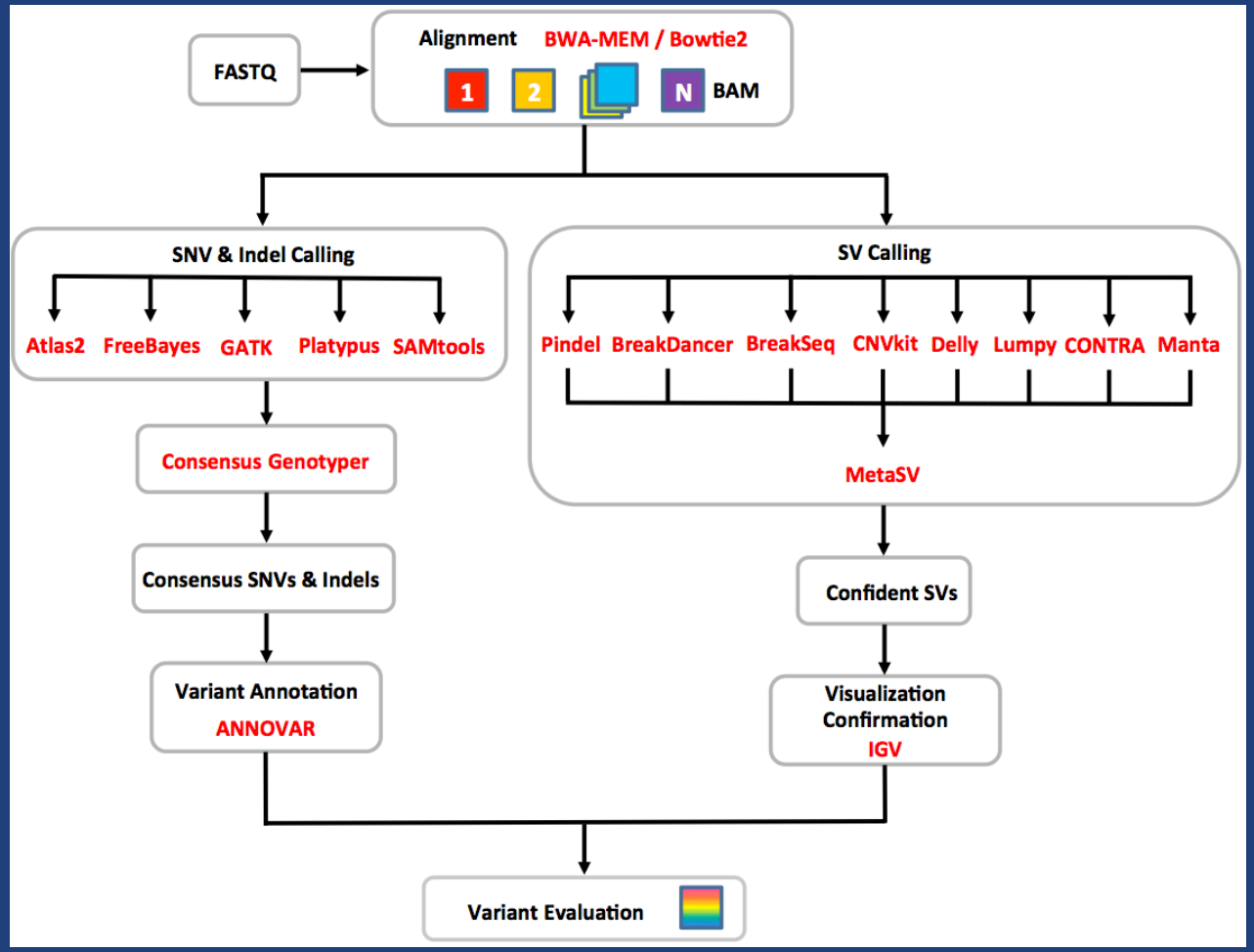
And colleagues from Cameroon and Uganda…

Motivation

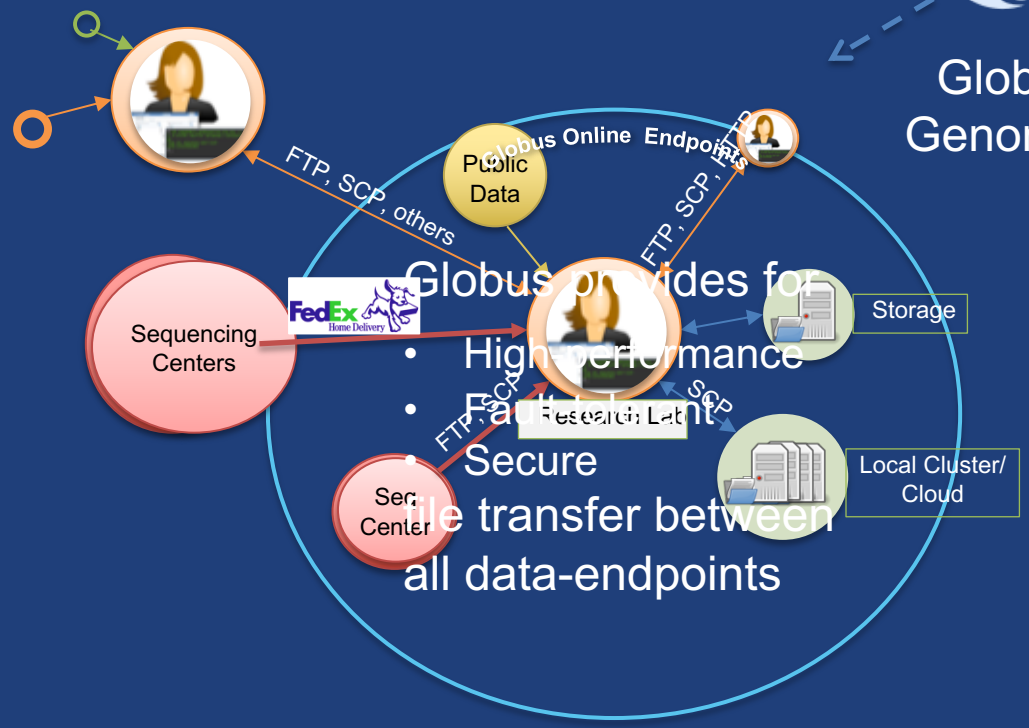
- Sub-Saharan Africa has a high proportion of premenopausal hormone receptor negative breast cancer - Previous studies* reported a strikingly high prevalence of germline mutations in BRCA1 and BRCA2 among Nigerian breast cancer patients
- It is unknown if this exists in other Sub-Saharan Africa countries
- We examined the burden of inherited breast cancer and the spectrum of germline variants in breast cancer susceptibility genes using a case-control study in Cameroon and Uganda
- 196 breast cancer cases, unselected for age at diagnosis and family history, were recruited from tertiary hospitals in Kampala, Uganda and Yaoundé, Cameroon. 185 controls were women without breast cancer recruited from the same hospitals and age-matched to cases
- A 30-gene hereditary cancer risk panel developed by Color Genomics (Burlingame, CA) was used for germline variant detection. Single nucleotide variants, small insertions and deletions, and large structure variants were called at the Color laboratory

Approach

- **In 2015, we developed Consensus Genotyper for Exome Sequencing for improving the quality of exome variant genotypes**
  - We used ensemble of variant-calling algorithms and a two-stage voting scheme among four algorithm implementations
- **We continued developing the approach by adding additional variant calling algorithms along with structural variant detection algorithms**
- **The output VCF files are normalized, and a set of highly confident variants are obtained through refinement by Consensus Genotyper for ANNOVAR or VEP annotation.**
- **SVs are precisely detected with MetaSV workflow that integrates Pindel, BreakDancer, BreakSeq, CNVkit, and Manta. DELLY, LUMPY and CONTRA are also available.**



Globus Genomics



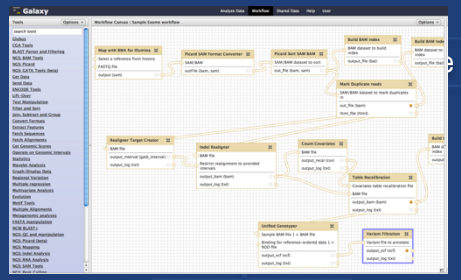
Data management

Globus Genomics

Galaxy-based workflow management

- Globus 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 using Cost-Aware Provisioning

Globus Genomics on Amazon EC2

Data analysis

Validation

- **We tested the performance of ConVarCal by analyzing germline targeted sequencing data (1.3Mbp, ave. 260x) of 200 Nigerian breast cancer patients.**
- **ConVarCal identified 25 deleterious SNVs/Indels in 29 subjects, and all have been confirmed experimentally**

Results

- The analysis of data from Cameroon and Uganda resulted in 45770 computational jobs that were launched on Amazon cloud. The analysis used a total of 25279 CPU hours from 900 nodes that were dynamically launched generating 8.76TBs of output datasets in the last year
- We found that, among cases, 15.8% carried a pathogenic or likely pathogenic mutation in a breast cancer susceptibility gene, while among controls, 1.6% carried a mutation in one of these genes. Cases were 11-fold more likely to carry a mutation compared to controls.
- Our findings confirm the earlier report of a high proportion of mutations in BRCA1 and BRCA2 among breast cancer patients in Sub-Saharan African countries
- We submitted this work for publication in May 2019, the manuscript is now being peer reviewed in Cancer Epidemiology, Biomarkers & Prevention.

nature
International journal of science

SPOTLIGHT · 10 APRIL 2018

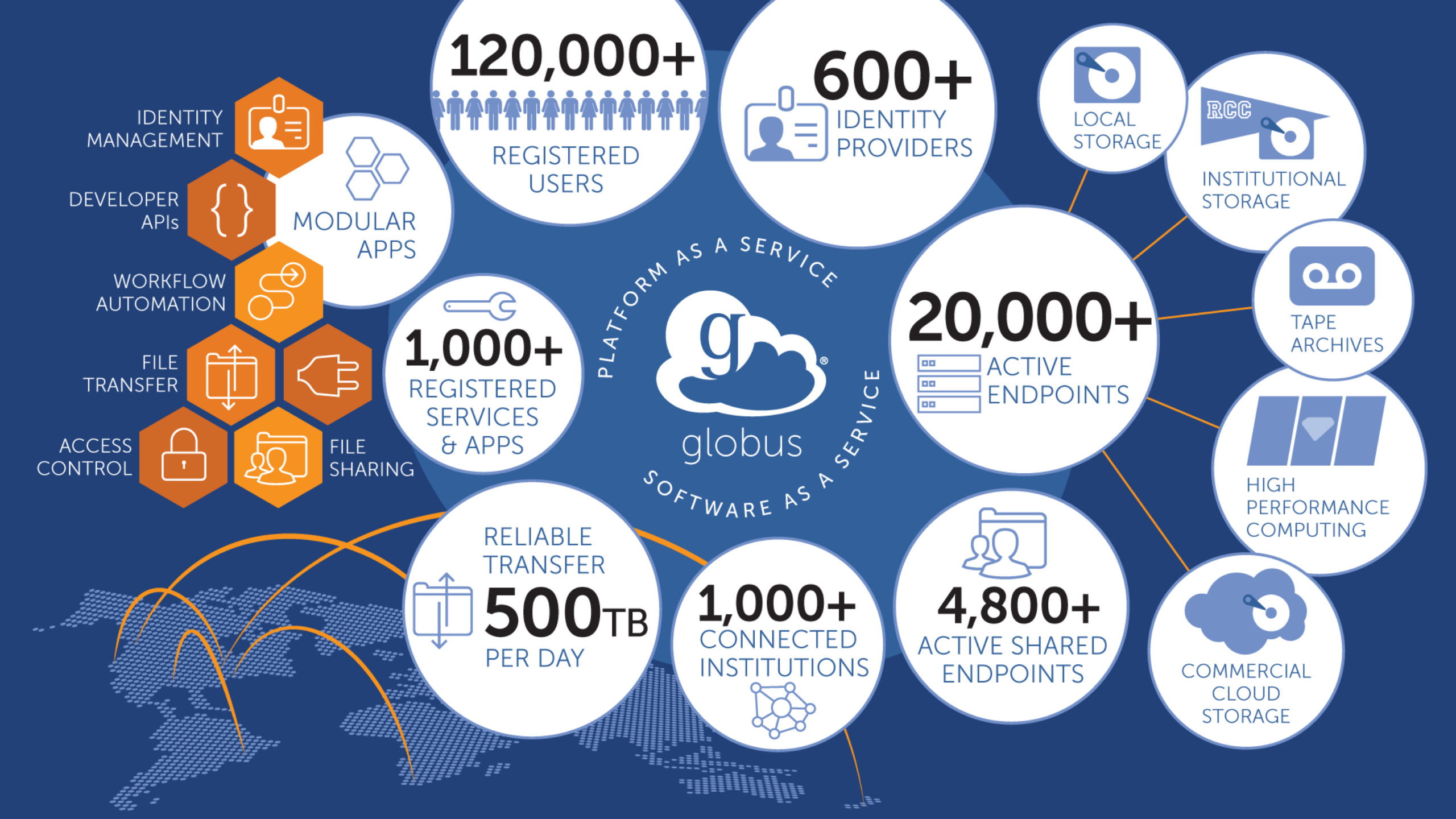
Collaboration is the key to cancer research

Lowering barriers to collaboration accelerates science.

Savage, N. Nature 556, (2018).

Mission

Globus lowers barriers to collaborative science and discovery by delivering easy, reliable, and secure research data movement, sharing, analysis and synchronization.


120,000+
REGISTERED USERS
600+
IDENTITY PROVIDERS
IDENTITY MANAGEMENT
DEVELOPER APIs
MODULAR APPS
WORKFLOW AUTOMATION
FILE TRANSFER
ACCESS CONTROL
FILE SHARING
1,000+
REGISTERED SERVICES & APPS
PLATFORM AS A SERVICE
globus
SOFTWARE AS A SERVICE
LOCAL STORAGE
RCC
INSTITUTIONAL STORAGE
20,000+
ACTIVE ENDPOINTS
TAPE ARCHIVES
HIGH PERFORMANCE COMPUTING
RELIABLE TRANSFER
500TB
PER DAY
1,000+
CONNECTED INSTITUTIONS
4,800+
ACTIVE SHARED ENDPOINTS
COMMERCIAL CLOUD STORAGE

Protected Hematologic Cancer Research Network

New Capabilities: Globus for managing PHI data

- **Access Control**
  - Identities provided and managed by institution
  - Acts as identity broker only, does not access or store any institutional user credentials
  - Institution controls all access policies (at multiple levels)
    - who can access what data and with what permissions
    - who can share what data and with what permissions
    - all access policies can be changed or revoked at any time
  - Researchers can overlay sharing permissions
- **Data remain at institution, not hosted by Globus**
- **Automated integrity checks of transferred data**
- **High service availability**
- **Monitoring**
- **Encryption (all communications, optional data in transit)**

Thank you for your time and attention

Questions ?

madduri@uchicago.edu