

**TAKING BIOMEDICAL
COMPUTING TO THE NEXT
SCALE: THE GLOBAL
ALLIANCE AND GENOME
BRIDGE**

12 December 2013

BRIITE

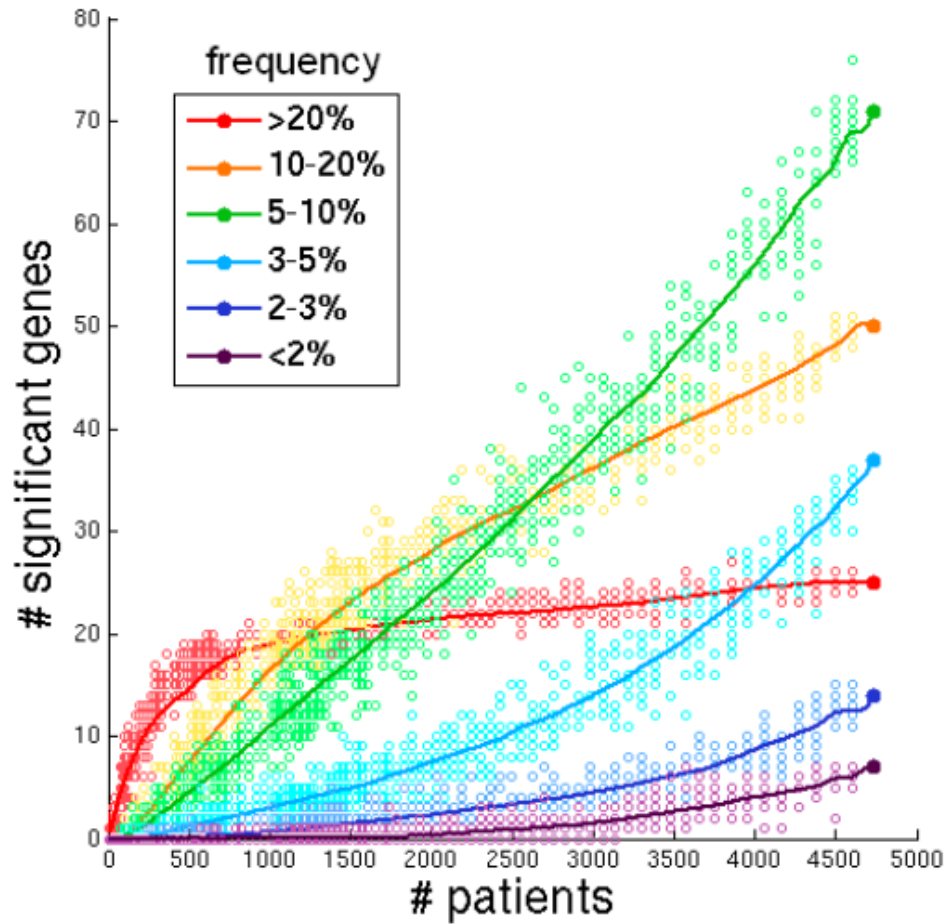


Global Alliance

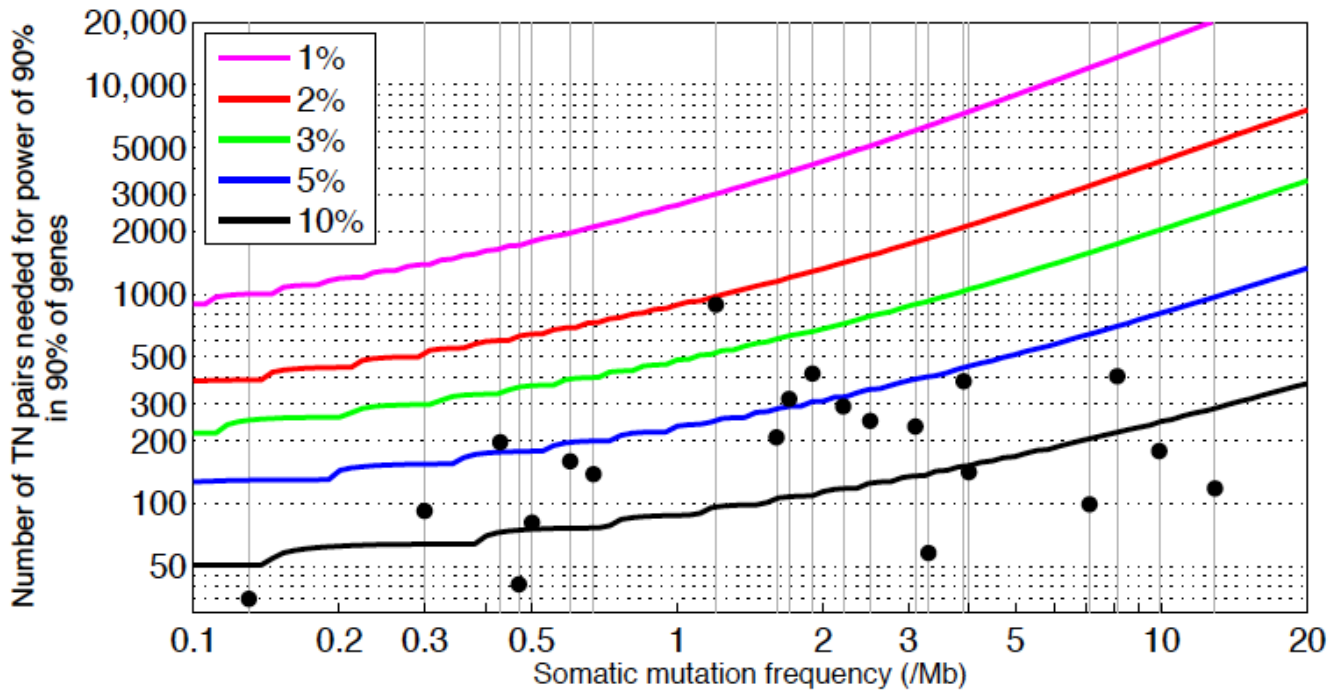


“Biomedical research is becoming more data-intensive as researchers are generating and using increasingly large, complex, and diverse data sets. This era of "Big Data" taxes the ability of biomedical researchers to locate, analyze, and interact with these data (and more generally all biomedical data) and associated software due to the lack of tools, accessibility, and training. “

Discovery of cancer genes is far from complete



$\geq 20\%$ largely discovered
 $< 20\%$ still rising rapidly




For 90% power to detect 90% of genes at frequency $\geq 2\%$:
Need mean of ~ 1700 samples

50 tumor types x 1,700 = 85,000 samples

- Rhabdoid tumor
- Medulloblastoma
- Acute myeloid leukemia
- Carcinoid
- Neuroblastoma
- Chronic lymphocytic leukemia
- Prostate
- Breast
- Multiple myeloma
- Ovarian
- Kidney clear cell
- Glioblastoma multiforme
- Endometrial
- Colorectal
- Diffuse large B-cell lymphoma
- Head and neck
- Esophageal adenocarcinoma
- Bladder
- Lung adenocarcinoma
- Lung squamous cell carcinoma
- Melanoma

Understanding *one* genome
requires *many* genomes.



“The results of the Research IT community interviews, and recent experiences working with very large data sets suggest that **Broad’s current research IT infrastructure is becoming limiting to certain research activities.**”

THE SCIENTIFIC OPPORTUNITY

We are witnessing an explosion of genomic information from individuals with known clinical characteristics and disease outcomes.

Learning from these data should accelerate progress in:

- Cancer outcomes and targeted therapy
- Understanding the basis of inherited disease
- Infectious disease
- Identifying targets for drug development

THE SCALE OF THE CHALLENGE

Need very large comparator datasets (millions)

Need new methods and organizational models:

- Data is typically in silos: by type, by disease, by institution
- Analysis methods are non-standardized, few at scale
- Approaches to regulation, consent and data sharing are needed

If we don't act: a hodge-podge of Balkanized datasets, much as developed for electronic medical records in the United States

Overcoming the challenge

Work together internationally to ensure interoperability of data and of methods, to harmonize approach to ethics and regulation, and to promote patient autonomy

Create cloud-based capabilities to store, analyze and share genomic information from patients and innovative methods



A global alliance with the mission to:

- Encourage interoperability of technology standards for managing and sharing genomic data in clinical samples;
- Facilitate harmonization of approaches to privacy and ethics in the international regulatory context;
- Engage stakeholders across sectors to encourage the responsible and voluntary sharing of data and of methods.

International partners describe global alliance to enable secure sharing of genomic and clinical data

By Broad Communications, June 4th, 2013

Over 70 leading health care, research, and disease advocacy organizations that together involve colleagues in over 40 countries have taken the first steps to form an international alliance dedicated to enabling secure sharing of genomic and clinical data. The cost of genome sequencing has fallen one-million fold, and more and more people are choosing to make their genetic and clinical data available for research, clinical, and personal use. However, interpreting these data requires an evidence base for biomedicine that is larger than any one party alone can develop, and that adheres to the highest standards of ethics and privacy. These organizations recognize that the public interest will be best served if we work together to develop and promulgate standards (both technical and regulatory) that make it possible to share and interpret this wealth of information in a manner that is both effective and responsible.

BACKGROUND

Technological advances have led to large-scale collection of data on genome sequencing and clinical outcomes, with great promise for medicine. In late January, 50 colleagues from eight countries met to discuss the current challenges and opportunities in genomic research and



[New York Times: Accord aims to create trove of genetic data](#) >

[New York Times Op Ed: Our genes, their secrets](#) >

[Boston Globe: Global alliance to create framework for sharing genomic data](#) >

[Science: Q&A: David Altshuler on how to share millions of human genomes](#) >

[Bloomberg: Scientists seek order to potential confusion of gene data](#) >

[The Guardian: DNA data to be shared worldwide in medical research project](#) >

Letters of Intent: 107 organizations in 17 countries:

American Association for Cancer Research
American Society of Clinical Oncology (US)
American Society of Human Genetics (US)
Association for Molecular Pathology (US)
A-T Children's Project (US)
Beth Israel Deaconess Medical Center (US)
BGI-Shenzhen (CN)
Boston Children's Hospital (US)
Brigham and Women's Hospital (US)
Broad Institute of MIT and Harvard (US)
California Institute of Technology (US)
Canada Health Infoway (CA)
Canadian Cancer Society (CA)
Canadian Institutes for Health Research (CA)
Cancer Commons (US)
Cancer Research UK (UK)
Center for Genomic Regulation (ES)
Murdoch University (AU)
Centre for the Advancement of Sustainable Medical Innovation (UK)
Centro Nacional de Analisis Genomico (ES)
Chinese Academy of Sciences (CN)
Coalition of Heritable Disorders of Connective Tissue (US)
Community Oncology Research Consortium
Max Planck Institute (DE)
Dana-Farber Cancer Institute (US)
Duke Health System (US)
ELIXIR (UK)
EMBL- European Bioinformatics Institute (UK)
European Molecular Biology Laboratory (DE)
European Society of Human Genetics (AT)
Garvan Institute of Medical Research (AU)
Genetic Alliance (US)
Genetic Alliance UK (UK)
Genome Canada (CA)
Genome Institute of Singapore (SG)
Global Genes | RARE Project (US)
H3ABioNet Consortium (ZA)
Harvard University (US)
Howard Hughes Medical Institute (US)
Human Genome Organization (SG)
Human Variome Project International (AU)
Huntington Society of Canada (CA)
Institut National du Cancer (FR)
Institute of Oncology Buenos Aires (AR)
Institute for Systems Biology (US)
Intermountain Healthcare (US)
International Cancer Genome Consortium
International Rare Diseases Research Consortium (FR)
Johns Hopkins University School of Medicine (US)
King's Health Partners (UK)
Knight Cancer Institute, Oregon Health & Science University (US)
Lund University (SE)
Massachusetts Eye and Ear Infirmary (US)
Massachusetts General Hospital (US)
McGill University/Université McGill (CA)
McLaughlin Centre, Faculty of Medicine, University of Toronto (CA)
Melbourne Genomic Health Alliance (AU)
Memorial Sloan-Kettering Cancer Center
National Cancer Center (JP)
National Cancer Institute (US)
National Health and Medical Research Council (AU)
National Human Genome Research Institute
National Institute of Genomic Medicine (MX)
National Institute for Health and Welfare (FI)
National Institute for Health Research (UK)
National Institutes of Health (US)
New York Genome Center (US)
O'Neill Institute for National and Global Health Law at Georgetown University (US)
Ontario Institute for Cancer Research (CA)
Ontario Personalized Medicine Network (CA)
Osaka University (JP)
Partners HealthCare (US)
Public Population Project in Genomics and Society (CA)
PHG Foundation (UK)
PXE International (US)
Queen's University Belfast (UK)
RD-Connect (UK)
RIKEN Center for Integrative Medical Sciences (JP)
Sage Bionetworks (US)
Samuel Lunenfeld Res. Institute, Mount Sinai Hospital (CA)
Simons Foundation (US)
Sleep Research Society (US)
Spanish Institute of Bioinformatics (ES)
Spanish National Cancer Research Center (ES)
St. Jude Children's Research Hospital (US)
Stanford University (US)
Sunnybrook Health Sciences Centre (CA)
SIB-Swiss Institute of Bioinformatics (CH)
The Hospital for Sick Children (CA)
The University of Cape Town (ZA)
Translational Oncology at the University Medical Center, Johannes Gutenberg University (DE)
University Health Network (CA)
University of California, Berkeley (US)
University of California Health System (US)
University of California, Los Angeles, Health Sciences (US)
University of California, San Francisco (US)
University of California, Santa Cruz (US)
University of Chicago (US)
University of Michigan (US)
University of Oxford (UK)
University of Texas M.D. Anderson Cancer Center (US)
University of Toronto (CA)
University of Washington (US)
University of Waterloo (CA)
Weill Cornell Medical College (US)
Wellcome Trust Sanger Institute (UK)
Wellcome Trust (UK)

nature biotechnology

Open to interpretation

An international alliance to enable secure sharing of human genomic and clinical data merits both broad support and financial backing from the global research and clinical communities.

Imagine a world where genomics and clinical data travel seamlessly between repositories at different institutions around the world; where harmonized standardized data formats and consent processes enable pooling of sequence data; and where standardized guidelines exist for informing patients and their families of the pathogenic significance of variations in their genome sequences. Making such a world a reality is the aim of the Global Alliance, an international initiative that aims to create universal interoperability standards and guidelines for genomics and medical data.

In recent weeks, the alliance (<http://www.broadinstitute.org/news/globalalliance>) published a white paper outlining its draft mission, goals and core principles. This was accompanied by a letter of intent with >70 signatories from medical, research and advocacy organizations in 41 countries, including such funders as the US National Institutes of Health, the UK's Wellcome Trust and Genome Canada. Since that time, another 20 organizations have signed the letter and the alliance has received numerous other expressions of interest.

in the past has been off-limits to research. But this doesn't make much sense going forward and the Global Alliance emphasizes the importance of autonomy—that patients should decide whether and how their data are shared—as potentially tens, perhaps hundreds, of thousands of exomes are sequenced in the clinic. Even today, diagnostic laboratories have more clinical information on sequence variants than is present in the literature—the question remains how to incentivize them to spend the time and effort in posting variant data in open repositories like ClinVar.

Finally, and perhaps most importantly, the forces of commercial balkanization and annexation provide even greater urgency to efforts to encourage genome data exchange in the public domain. The past year has witnessed extensive consolidation in the sequencing and molecular diagnostics sectors, with a handful of companies now monopolizing the market. In 2012, Life Technologies acquired direct-to-consumer (DTC) genomics provider Navigenics for its clinical testing services and 23andMe consolidated its stranglehold on the DTC market through lowball pricing.

- **Governance:** focused on the structural start-up needs of the alliance, including developing and making recommendations on governance structures, legal agreements and operating mechanisms.
- **Regulatory:** focused on ethics and the legal and social implications of the global alliance, including harmonizing policies and standards, and developing forward-looking consent, privacy procedures, and best-practices in data governance and transparency. (Bartha Knoppers, Katzuo Kato)
- **Genomic Data:** focused on data representation, storage, and analysis, including working with platform development partners and industry leaders to develop standards that will facilitate interoperability. (Paul Flicek)
- **Security:** leading the thinking on the technology aspects of data security, user access control, and audit functions, working to develop or adopt standards for data security, privacy protection, and user/owner access control. (David Haussler, Richard Durbin)
- **Clinical Interface:** working to establish linkages to phenotypic and clinical (health) informatics data. Rather than invent such standards, this group will focus on aligning our genomic data activities with the ongoing international standards initiatives in clinical and health data. (Charles Sawyer)



New non-profit subsidiary of Broad with the mission to create a scalable platform for storage and analysis of genomic data:

- **Cloud-based, elastic compute and cost effective data-storage**
- **Enterprise-level security with identity and granular access management to ensure privacy and manage data-use agreements and IRB protocols**
- **Provide access to best-in-class tools for analyzing and interpreting genome data developed at Broad Institute and elsewhere**
- **Rich APIs for third-parties to develop pipeline and analysis tools**

An environment for sharing data between communities of researchers, clinicians, patients and biopharmaceutical companies



Leadership Team:



- Brad Margus (CEO): experienced founder/CEO (Perlegen Sciences, Envoy Therapeutics); longstanding advocate families with genetic disease

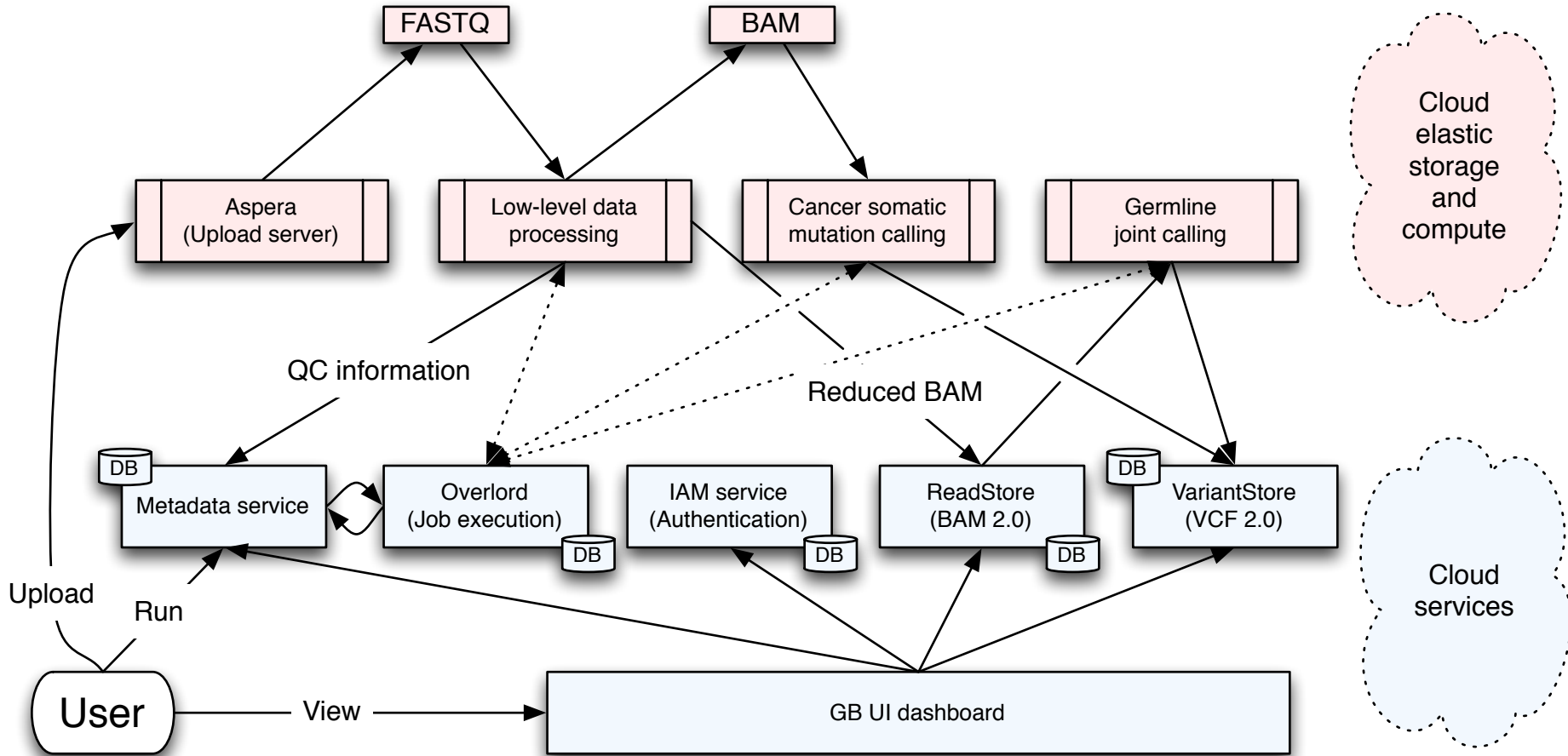


- Eric Golin (CTO): computer scientist with deep experience leading software development (former CTO of Carbonite, Eons, Broadvision)



- Anthony Philippakis (CMO): cardiologist, geneticist, expert in machine learning

A cloud-native service-oriented architecture to run at scale



LONG TERM GOALS

(1) Move all of Broad's compute to a cloud-model (including research, production and CLIA pipelines)

- To enable more efficient and scalable compute for Broad scientists and replace existing systems
- To demonstrate that a large number of scientific projects and a huge genome sequencing center can perform all of their business at Genome Bridge

(2) Move all of Broad's data and other large project data to Genome Bridge

- To enable the scientific goals of the Broad to aggregate large amounts of data and analyze them
- To demonstrate that the Genome Bridge system can handle large amounts of data (including upload, ease of use, state-of-the-art analysis and inherent security) and to attract others to contribute their data.

CANCER GENOME CHALLENGE

A Broad international project seeking to

- **Aggregate cancer genomic data (n~17,000 samples)**
 - TCGA and other available data
 - 10K-15K tumor/normal exome pairs
 - 500-2000 tumor/normal whole genome pairs
- **Pan-Cancer analysis to identify new driver mutations**
 - Identify significant somatic point mutations
 - Somatic copy-number analysis including germline variants
 - Integrate germline data to discover genetic predispositions

MOTIVATION FOR THE CANCER GENOME CHALLENGE PROJECT

- (1) Scientific – Analyze the world’s cancer genome data**
- (2) Pilot the Broad use of scalable (elastic) compute environment for reaching its scientific and production goals**
- (3) Pilot Genome Bridge offering of a generic infrastructure for analysis of genomic data**



ACKNOWLEDGEMENTS

Genome Bridge

- Anthony Philippakis
- Brad Margus
- Eric Golin

Broad

- David Altshuler
- Gad Getz
- Mike Lawrence