Center for Big Data in Translational Genomics

David Haussler
Benedict Paten
Brian O’Connor

UC Santa Cruz Genomics Institute
Problem: most genome data held in silos, unshared, not standardized for exchange

No one institute has enough on its own to make progress. Every researcher and clinician should be able to compare their genomes to others.
We need a network for sharing
New API Advances Data Interoperability

Learn how the Genomics API Version 0.5 is advancing information sharing for DNA data providers and consumers on a global scale.

What is the Global Alliance?

The Global Alliance for Genomics and Health (Global Alliance) is an international coalition, dedicated to improving human health by maximizing the potential of genomic medicine through effective and responsible data sharing. The promise of genomic data to revolutionize biology and medicine depends critically on our ability to make comparisons and draw conclusions based on diverse data sets. The Global Alliance is working to ensure that genomic data is shared in a way that is ethically responsible, scientifically valuable, and technologically feasible.

What is the Global Alliance doing?

Since its formation in 2013, the Global Alliance for Genomics and Health is leading the way to enable genomic and clinical data sharing. The Alliance’s Working Groups are producing high-impact deliverables to ensure such responsible sharing is possible, such as developing a Framework for Data Sharing to guide governance and research, and a set of recommendations for the development of a global data sharing framework.

Who is involved?

The Global Alliance for Genomics and Health is an independent, non-governmental alliance, made up of hundreds of world-leading organizations and individuals from across the world. The Global Alliance is focused on bringing together a diverse set of key stakeholders across regions and sectors, including leaders in healthcare and research, governmental and non-governmental organizations, academic institutions, and industry.
Enabling Responsible Sharing of Genomic and Clinical Data

- GA4GH Founded on June 5, 2013
  - Founding partners: 70+ leading healthcare research, and disease advocacy orgs
  - Now: 446 institutional members from 42 countries, more than 150 companies

- Mission: to enable rapid progress in biomedicine

- Focus:
  - Interoperability standards
  - Privacy and ethics
  - Networks of voluntary sharing of data and of methods
One data model, API, and public data set. Many data stewards and apps.

Data Model, Application Programming Interface, Public Dataset

- Medical research institution
- Advocate’s patient registry
- Hospital
- Health/IT company
- National health service

- Query to discover data
- Perform statistics
- Authenticate for deep dive
- Integrated health app for consumer
- Integrated app for professional
GA4GH’s Data Working Group process

• We work together in an open source software development environment on the web: [https://github.com/ga4gh](https://github.com/ga4gh). There are now 179 registered developers, and thousands of code contributions.

• All groups are welcome to participate

• Decision making is done by protocols developed by Apache Open Source Software Foundation

• Leadership is determined by amount of contribution

• Simple Mantra: collaborate on interface, compete on implementation
Standardizing the computational representation of genetic variation

Variant Modeling Collaborative led by Reece Hart, with synergy efforts led by Gil Altero, Mark Diekhans (UCSC), Larry Babb (Clingen), Jennifer Lee (ClinVar), Zak Kohane group
Data Schemas and Applications Programming Interface

Kevin Osborne, David Steinberg, Jerome Kelleher, and
GA4GH API Usage Today

- UCSC 1000 Genomes server
- Ensembl/EBI
- Google
- Australian Phenomics Facility Beacon service layer
- Microsoft, Cornell, Wash U.
GA4GH Demonstration Project: Beacons to Discover Data

Do you have any genomes with an “A” at position 100,735 on chromosome 3?

Marc Fiume, U. Toronto
Some Beacon Protocol Adopters (25 Institutes)
Count Everything: Linking GA4GH APIs to BD2K APIs

Demonstrating cross cutting, secure, privacy preserving count statistics

Select count(*) from MD2K A, i2b2 B, GA4GH C where
A.blood Glucose>110 and B.smoker=true and
C.r123140=C
Count Everything: Linking GA4GH APIs to BD2K APIs

Demonstrating cross cutting, secure, privacy preserving count statistics
Sharing Analyses via GA4GH API

Tool Registry API

- Not just data, but analyses – described in containers and workflows - that must be shared to achieve FAIRness
- Formalizing sharing with the GA4GH **Containers and Workflows Task Team**
- Started with discovery API with extended support for write and search

http://dockstore.org

APIS

- GET list
- GET search
- POST register
Emerging GA4GH API Standards

Next step: Containers and Workflows: Execution APIs – a standard way to invoke workflows, regardless of environment

API Standard to Execute
- POST new task
- GET task status
- GET task stderr/stdout/file(s)
GA4GH Containers & Workflow TT Vision

1. Search & browse for tools & workflows

2. Portable workflows run on platform of choice

3. Platforms can execute multi-site/multi-cloud federated analysis

GA4GH Tools Registry API

Common Workflow Language (CWL) or Workflow Definition Language (WDL)

BioShadock
Dockstore.org
Synapse

GA4GH Workflow Execution API

CCC
Consonance
Curoverse
Firecloud
Seven Bridges

GA4GH Task Execution API

AWS
Azure
Google
HPC

GA4GH File Formats & APIs

AWS S3
Azure Blobs
Google Storage
Arvados Keep
iRODS
NFS

Co-led by Peter Amstutz (Curoverse), Jeff Genrty (Broad), Brian O’Connor (UCSC)
C&W Scale Demonstration: Toil RNA-seq Compute

- **20,000 RNA Seq samples – all TCGA, gTEX, PNOC, TARGET and SPY2.**
- **Computed in < 4 days**
- **32,000 Cores, 64 TBs of ram, >1 Million “spot” CPU hours**
- **1/40th cost of previous equivalent efforts**

See Vivian et al. 2016, Rapid and efficient analysis of 20,000 RNA-seq samples with Toil, Nature Biotech in press
Human Genome Variation Map: Merge diverse reference genomes into one graph

The major histocompatibility complex—Gil McVean
Demo: Variant calling on Illumina Platinum Genomes

Take Home: Graphs (red) + simple algorithm outperforms existing, more sophisticated methods (black) and controls (green)
Variant Interpretation for Cancer Consortium (VICC)

Building on the work of Genotype-to-Phenotype (G2P) Task Team led by Adam Margolin from our BD2K Center

http://ga4gh.org/#/vic
ga4gh-dwg-vic@genomicsandhealth.org

VICC Co-chairs:
• Obi Griffith
• Nuria Lopez-Bigas
• David Tamborero
• Malachi Griffith

Goals/Principles:
• Clinical cancer variant interpretation
• Standards and guidelines
• Open content
• Interoperability
BRCA Variation: The Challenge

- And yet, there is no single source for BRCA variant information!
  - Data is globally distributed
  - Major databases in Europe, US, other regions growing fast
  - Mixture of private and public-sector
  - Many difference health care providers/payers in system
  - Patient privacy is concern, so data must be consented for sharing

Source: BRCA Exchange, As of May 2016 (UMD last public release (May 2015))
BRCA Variation : The Challenge

• The explosion in BRCA testing:
  • Result of supreme court decision to overturn Myriad Genetics gene-patent
  • Cost reduced from $3000 to $250 in three years
The BRCA Exchange aims to advance our understanding of the genetic basis of breast cancer, ovarian cancer and other diseases by pooling data on BRCA1/2 genetic variants and corresponding clinical data from around the world. Search for BRCA1 or BRCA2 variants above.

This website is supported by the BRCA Exchange of the Global Alliance for Genomics and Health.
To date:

- > 17,800 variants
- Integrates 8 different international databases
- Gold standard curation from international team, now of > 3,500 variants

Ultimate goal:
Goal to become FDA approved site for BRCA interpretations - a single, trusted source
Cross-cancer comparisons in a prospective clinical setting

Patient 1: 8 year-old boy, dural-based sarcoma
Treated with aggressive chemotherapy and local radiation

Two years later—metastases to lungs
• No standard treatment options, so Patient 1 enrolled in a personalized genomics clinical trial
• No clinically actionable leads

S. Rod Rassekh, Rebecca Deyell, Stephen Yip and Marco Marra
JAK/STAT pathway is a potential therapeutic target for Patient 1

Activating receptor tyrosine kinases; Cytokines

FGFR1
ALK
IL6R

PI3K
AKT1

JAK1
STAT3/S TAT1

STAT6

Transcriptional targets

mTOR

C-MYC
CCND1
BCL-XL

VEGFA
MCL1
CDKN1A

Expression outlier
Top 5% expressed genes in Patient 1

IL6 Proteomics analysis
Success Last Year

“Our mutual patient ….. is still alive…. I restarted him on ruxolitinib again to inhibit JAK and he feels amazing again! eating, increased energy and feels like a million bucks again.”

—Treating Oncologist
2016: cancer recurred and the boy passed away.

His physician estimates precision treatment led to 1.5 years extension of life, much of it high quality.
Much more work to do

Let’s do it together.
## Acknowledgements

**GA4GH Data API:**
Kevin Osborn  
David Steinberg  
Edwin Jacox  
Mark Diekhans  
Danny Colligan

**Big Data Genomics:**
Brian O'Connor  
Walt Shands  
Hannes Schmidt  
Christopher Ketchum  
John Vivian  
Arjun Rao  
Carlos Mendez  
Audrey Musselman-Brown

**BRCA Exchange:**
Melissa Cline  
Mary Goldman  
Joe Thomas  
Brian Craft  
Max Haeussler  
Molly Zhang

**HGVM:**
Adam Novak  
Jordan Eizenga  
Yohei Rosen  
Charles Markello  
Glenn Hickey  
Anthony Joseph

**External collaborators:**
**Sanger:**  
Richard Durbin  
Erik Garrison

**University of California Berkeley:**
Frank Nothaft  
Justin Paschall  
Michael Heuer

**Oregon Health & Science University:**
Cristina Tognon  
Brian Druker  
Adam Margolin  
Brian Walsh

**Caltech:**
Lior Pachter  
Barbara Wold  
Sean Upchurch

**UCSF:**
Laura Van't Veer  
Denise Wolf

**Sage Bionetworks:**
Justin Guinney

**Global Alliance for Health and Genomics:**
Reece Hart  
Stephen Keenan  
Richard Durbin  
Peter Goodhand,  
Kathryn North  
Ewan Birney