UNIVERSITY OF CALIFORNIA
SANTA CRUZ | Genomics Institute

Center for Big Data in Translational Genomics (Genomics Center)

September 30th 2016

NIH/NHGRI 5U54HG007990-03

Genomics Center

- Collaboration of:
  - UCSC
    - David Haussler, Benedict Paten
  - UCB
    - Dave Patterson, Anthony Joseph, Lior Pachter
  - OHSU
    - Brian Druker, Adam Margolin
  - UCSF
    - Laura van’t Veer
  - CalTech
    - Barbara Wold, Mitch Guttman
  - Sage Bionetworks
    - Justin Guinney

Genomics Center Research Aims

1: **APIs**

Pioneer common Application Programming Interfaces (APIs) for big genomics data in biomedicine.

2: **Benchmarking**

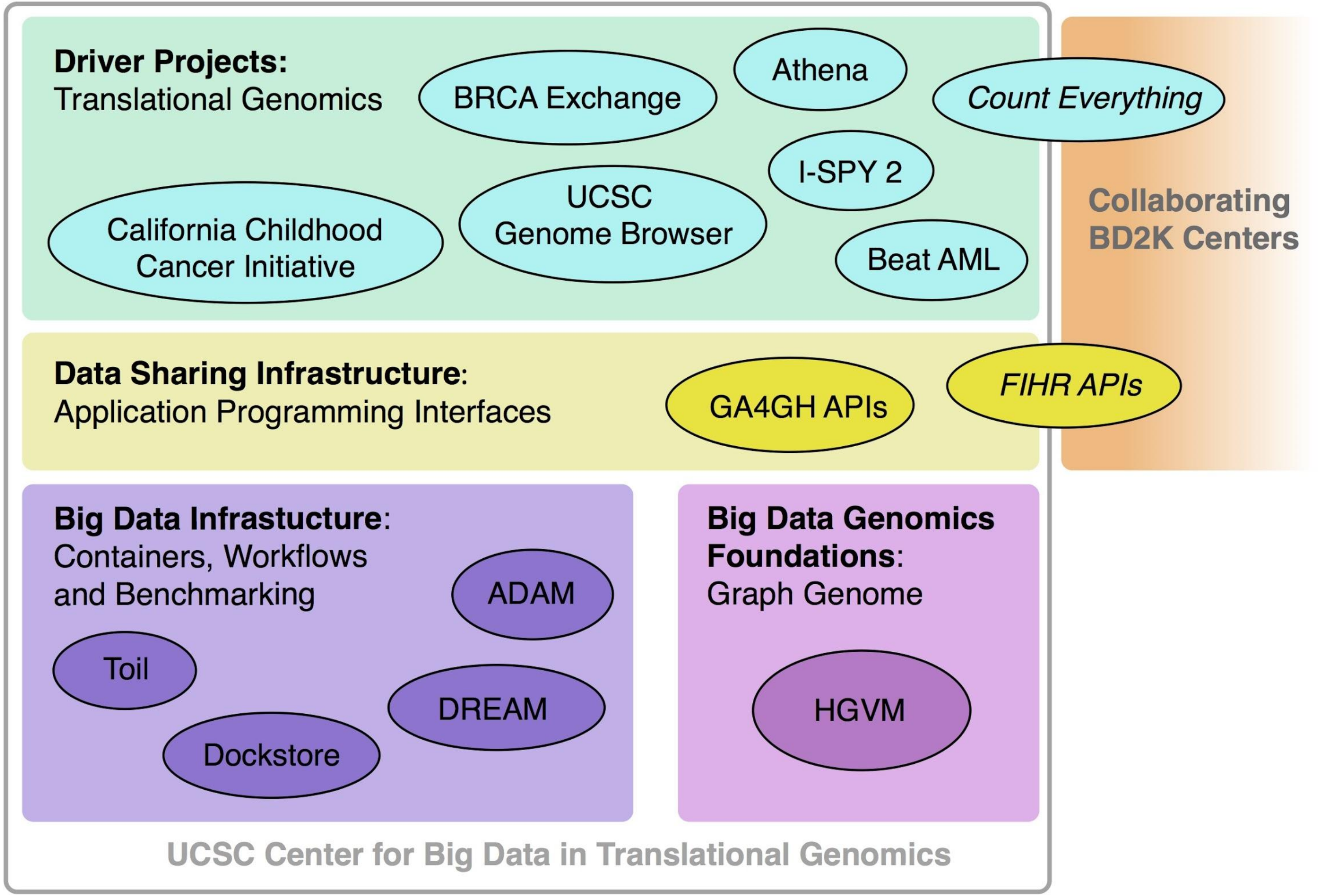
Create a continuously operating benchmarking platform for methods of large-scale genomics analysis.

3: **Big Data Genomics Software**

Develop large-scale genomics analysis tools that interact with the genomics data APIs.

4: **Driver projects**

Pilot APIs and tools in a variety of large and small projects in different areas.



Driver Projects:
Translational Genomics

BRCA Exchange

Athena

Count Everything

I-SPY 2

California Childhood Cancer Initiative

UCSC Genome Browser

Beat AML

Collaborating BD2K Centers

Data Sharing Infrastructure:
Application Programming Interfaces

GA4GH APIs

FIHR APIs

Big Data Infrastucture:
Containers, Workflows and Benchmarking

ADAM

Toil

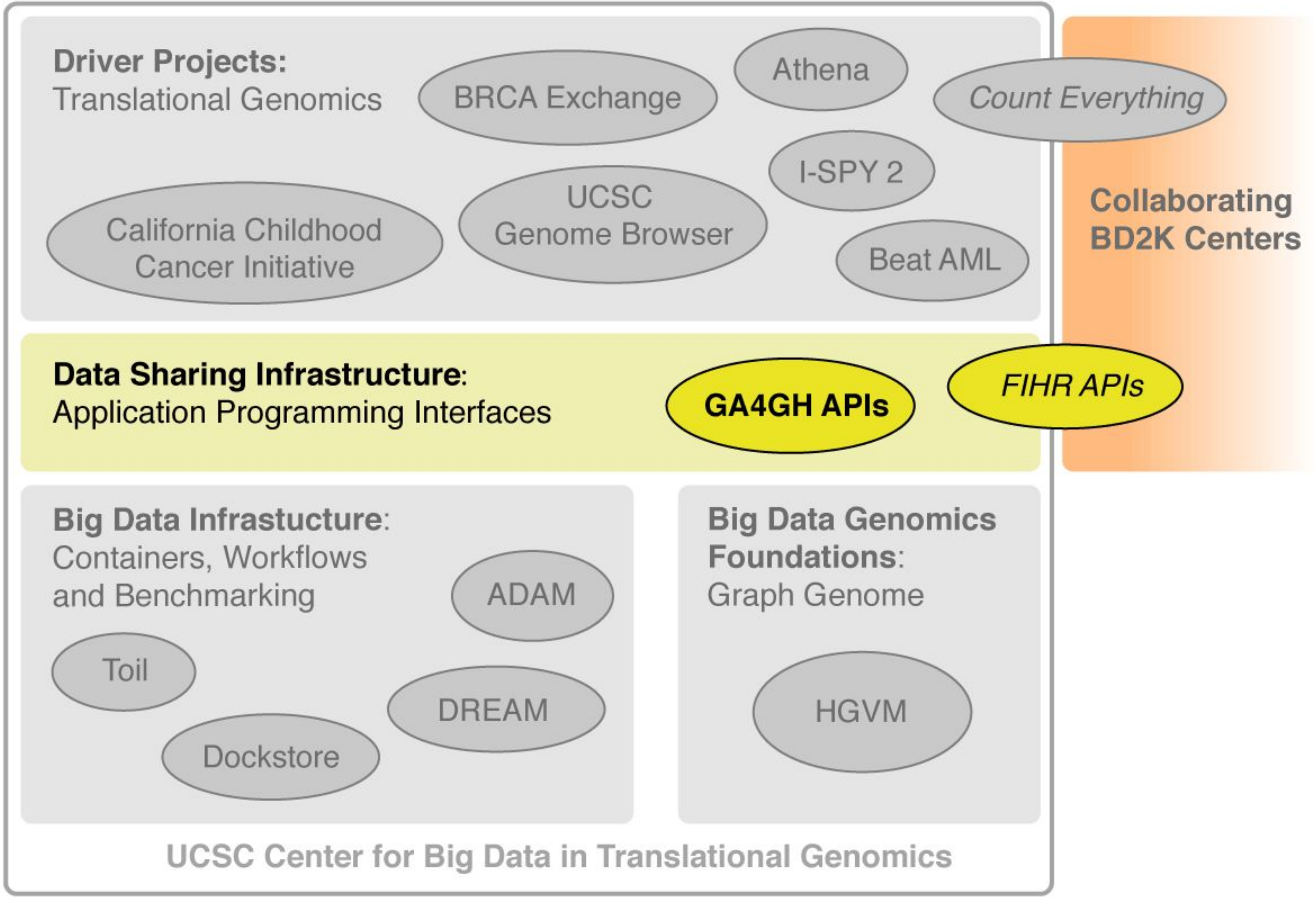
DREAM

Dockstore

Big Data Genomics Foundations:
Graph Genome

HGVM

UCSC Center for Big Data in Translational Genomics



Driver Projects:
Translational Genomics

BRCA Exchange

Athena

Count Everything

I-SPY 2

California Childhood Cancer Initiative

UCSC Genome Browser

Beat AML

Collaborating BD2K Centers

Data Sharing Infrastructure:
Application Programming Interfaces

GA4GH APIs

FIHR APIs

Big Data Infrastucture:
Containers, Workflows and Benchmarking

ADAM

Toil

DREAM

Dockstore

Big Data Genomics Foundations:
Graph Genome

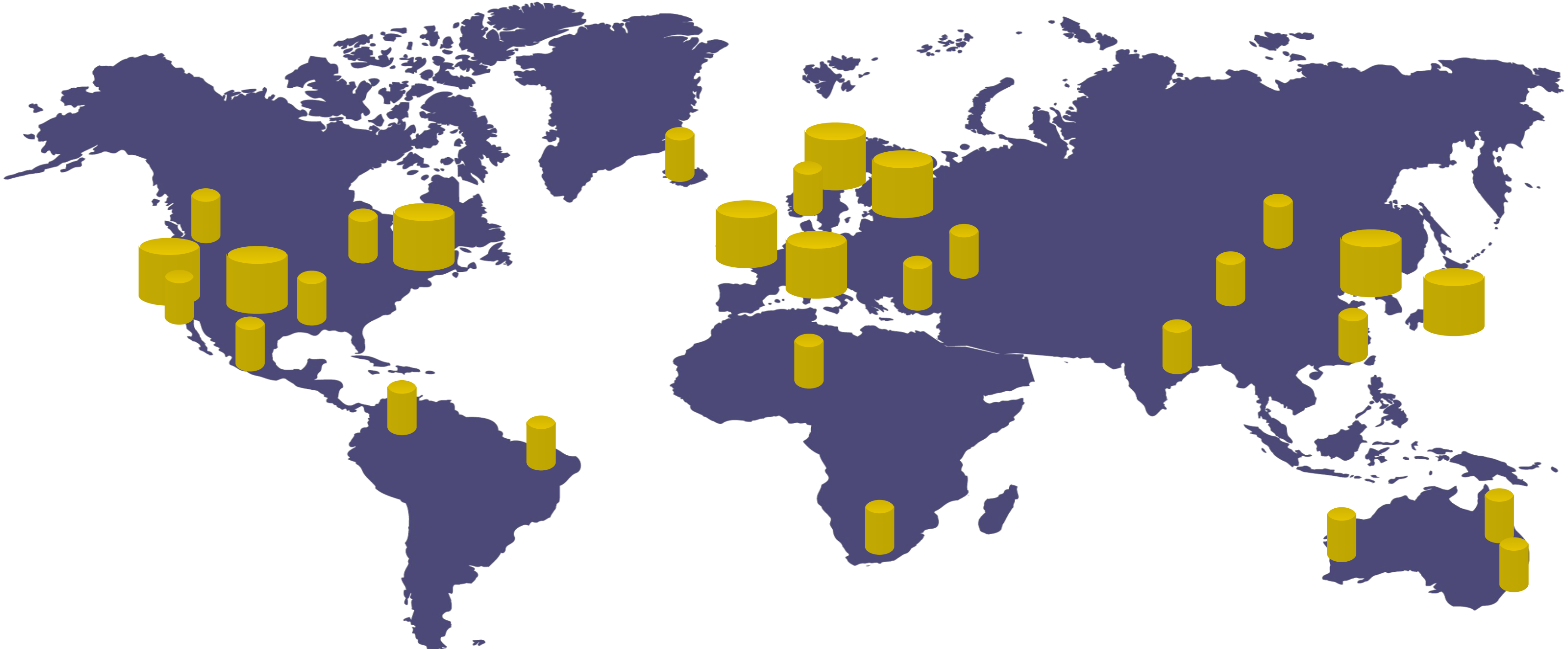
HGVM

UCSC Center for Big Data in Translational Genomics

Problem: Genome Data Held in Silos, Unshared, not Standardized for Exchange

No one institute has enough on its own to make progress.

Every clinician should be able to compare their genomes to others.



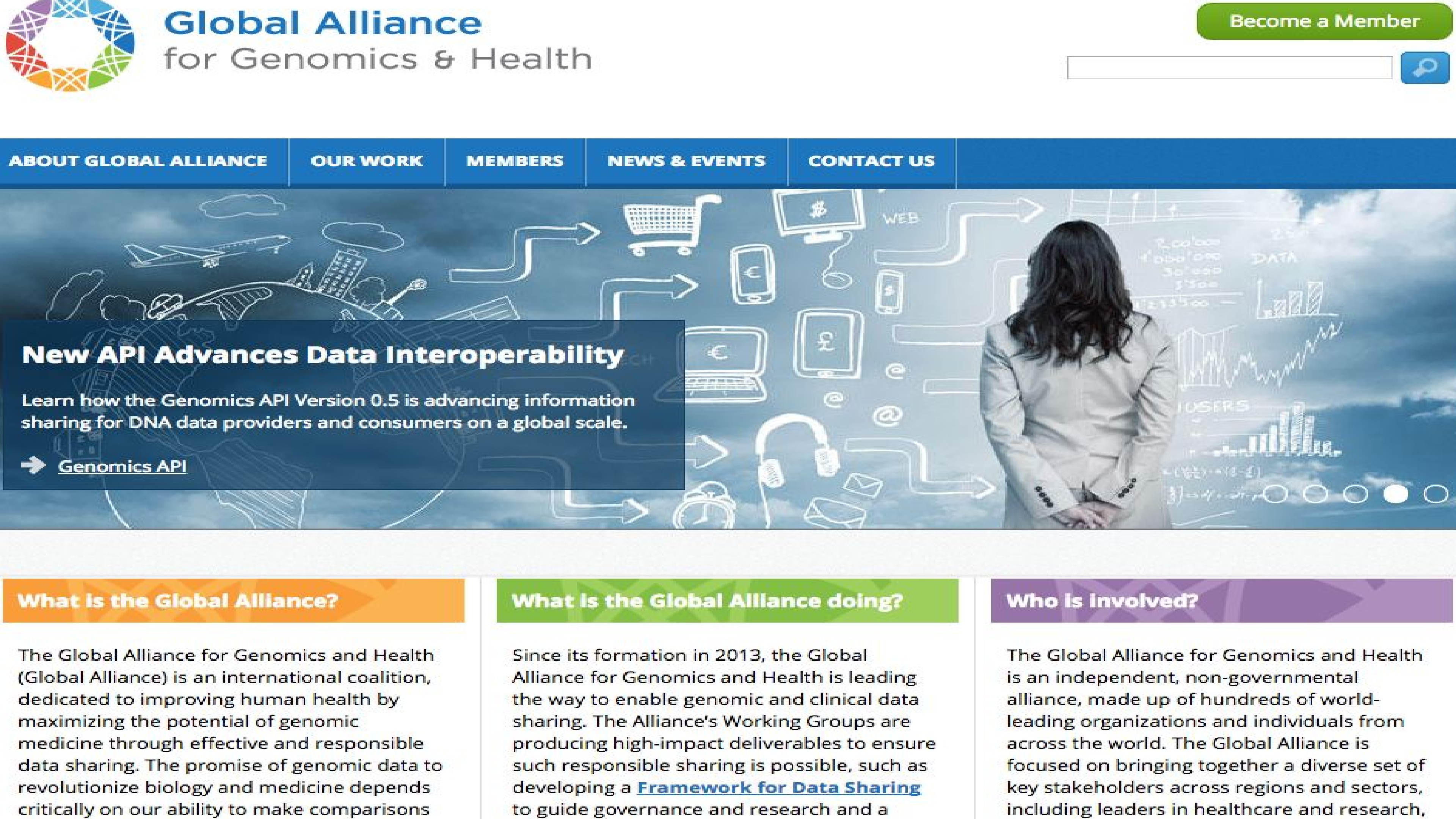
We need a network for sharing

ATTTATCTGCTCTCGTTG

GAAGTACAAAATTCATTAAT

GCTATGCACAAAATCTGTAG
C

TAGTGTCCCATCTATTT



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.

➔ Genomics API

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

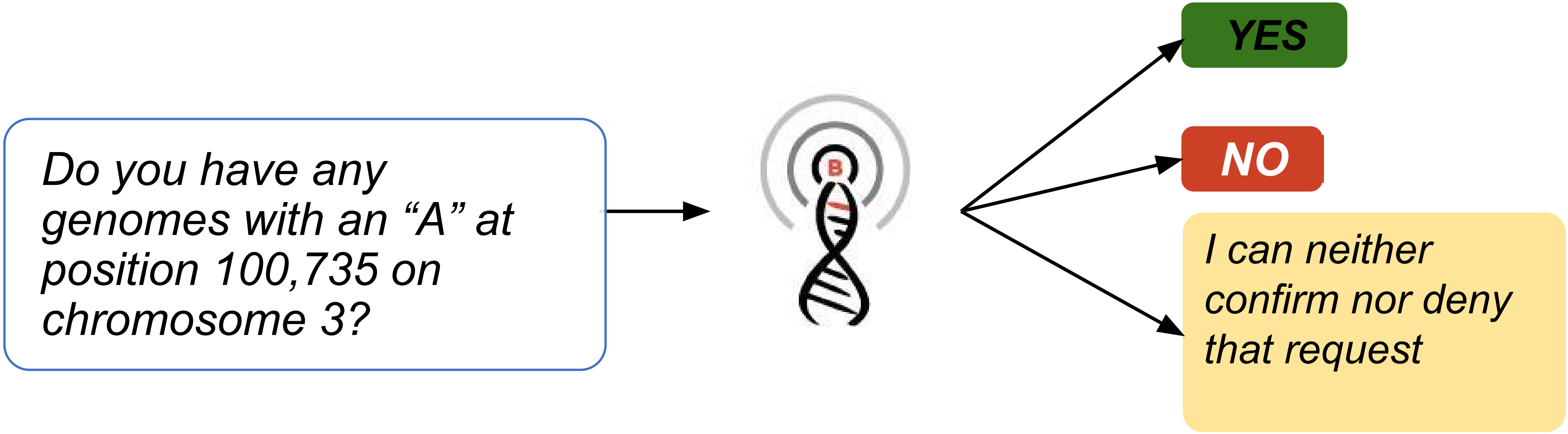
What is the Global Alliance doing?

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

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,

GA4GH Driver Project:
Beacons to Discover Data



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

YES

NO

I can neither confirm nor deny that request

Protocol Adopters (Fall 2015)

OICR
Ontario Institute
for Cancer Research

UCSC

amp
lab

University of
Leicester

BROAD
INSTITUTE

Institute for
Systems Biology

NCBI

LOVD
Leiden Open Variation Database

wellcome trust
sanger
institute

EMBL-EBI

SolveBio

Curoverse

DNASTACK

Google

BioReference
LABORATORIES

SFARI
SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE

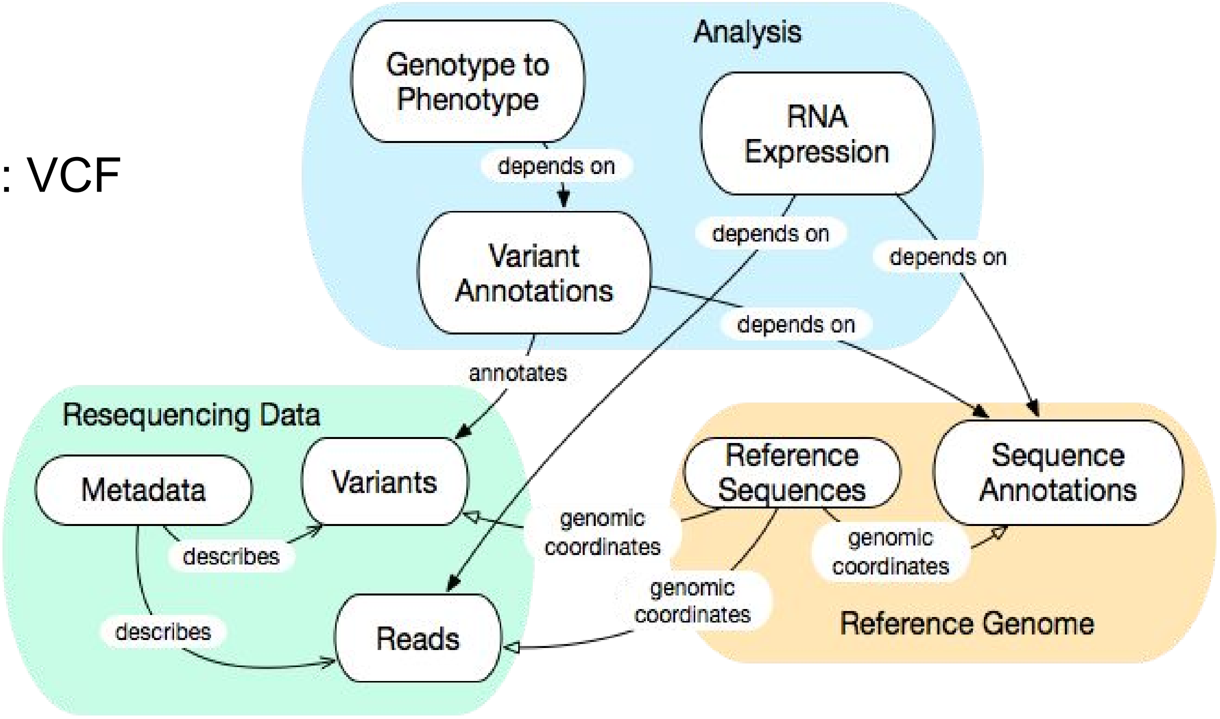
Accomplishments to date

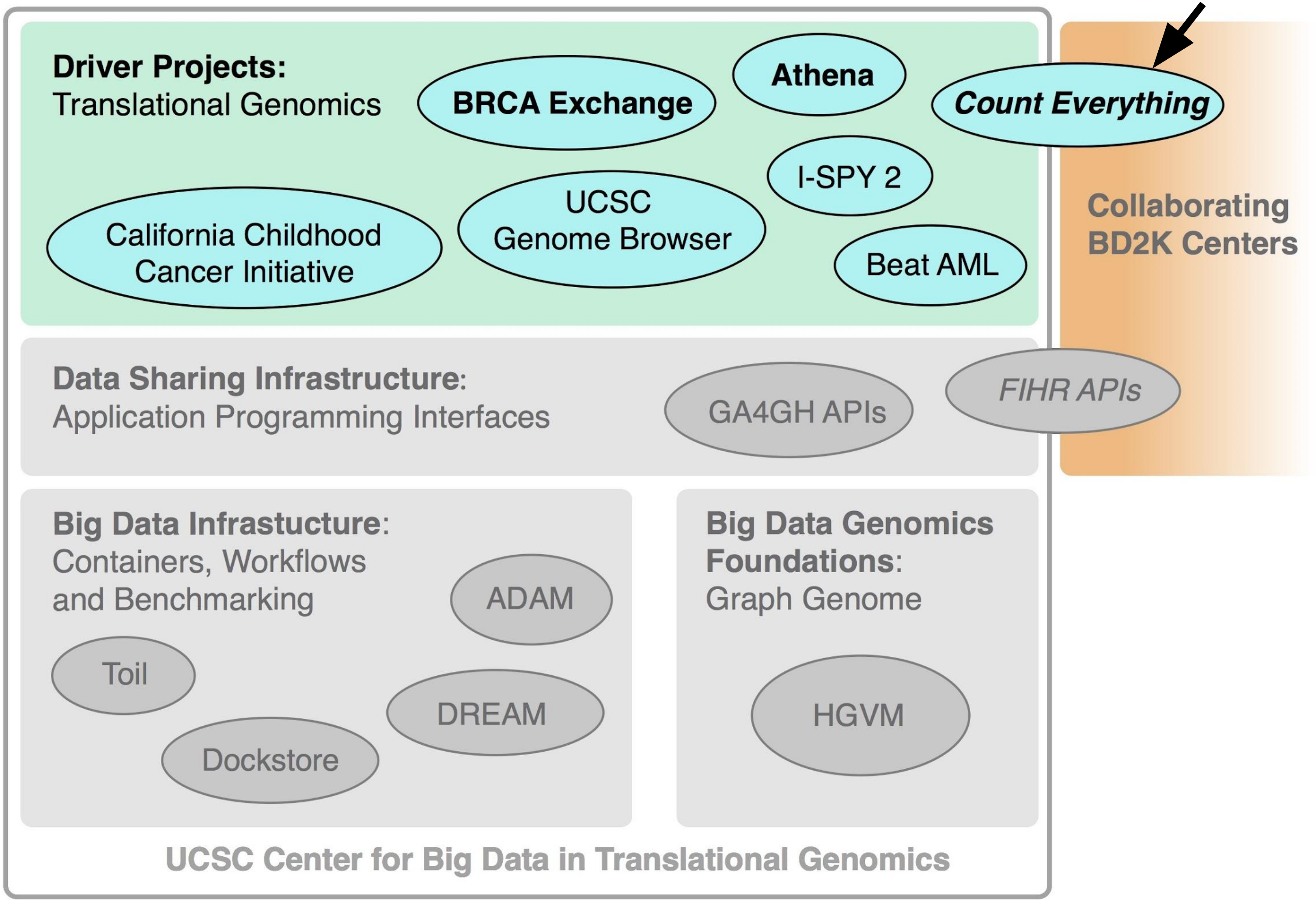
Functional support for:

- Reads : BAM
- Variants and annotations : VCF
- References : FASTA
- Seq Annotations: GFF3
- RNA
- Genotype to Phenotype
- Metadata

Coming soon:

- Other data sources: BED, wiggle




Driver Projects:
Translational Genomics
BRCA Exchange
Athena
Count Everything
I-SPY 2
California Childhood Cancer Initiative
UCSC Genome Browser
Beat AML
Collaborating BD2K Centers
Data Sharing Infrastructure:
Application Programming Interfaces
GA4GH APIs
FIHR APIs
Big Data Infrastucture:
Containers, Workflows and Benchmarking
ADAM
Toil
DREAM
Dockstore
Big Data Genomics Foundations:
Graph Genome
HGVM
UCSC Center for Big Data in Translational Genomics

Problem Statement

Secure aggregating counts of relevant patients

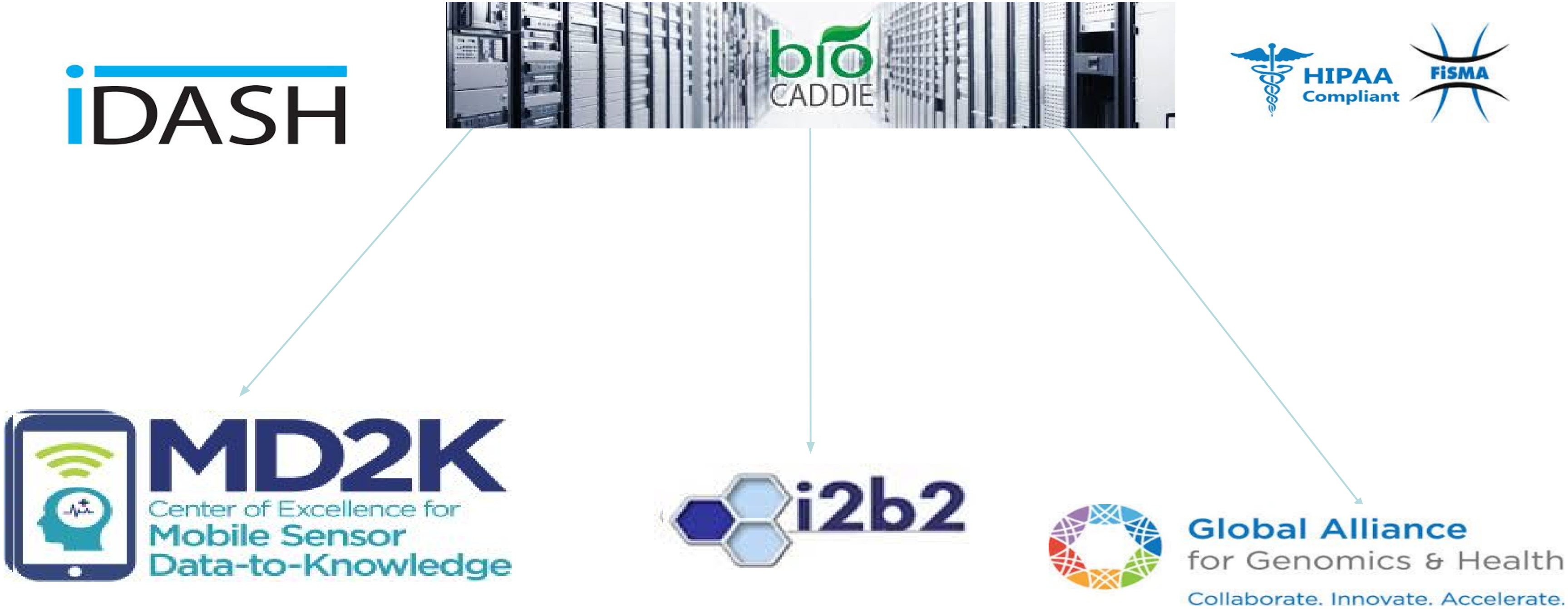
iDASH

HIPAA Compliant

FISMA

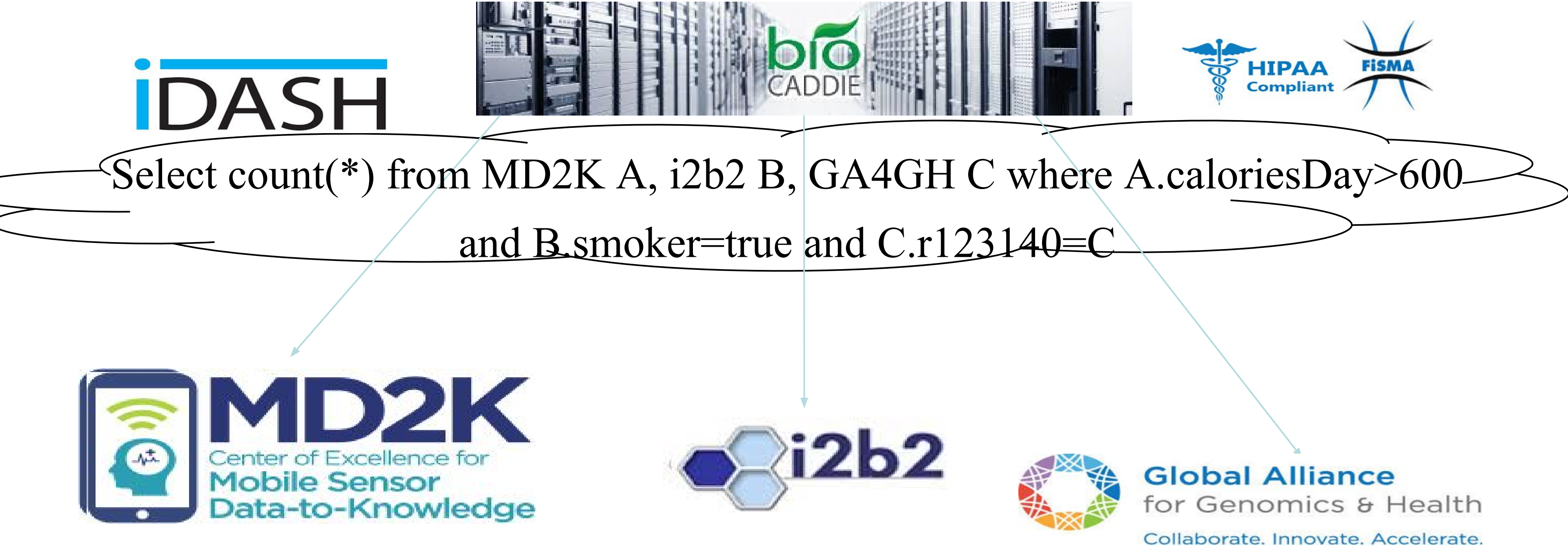
Problem Statement

Secure aggregating counts of relevant patients


iDASH
bio CADDIE
HIPAA Compliant
FISMA
MD2K
Center of Excellence for Mobile Sensor Data-to-Knowledge
i2b2
Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

Problem Statement

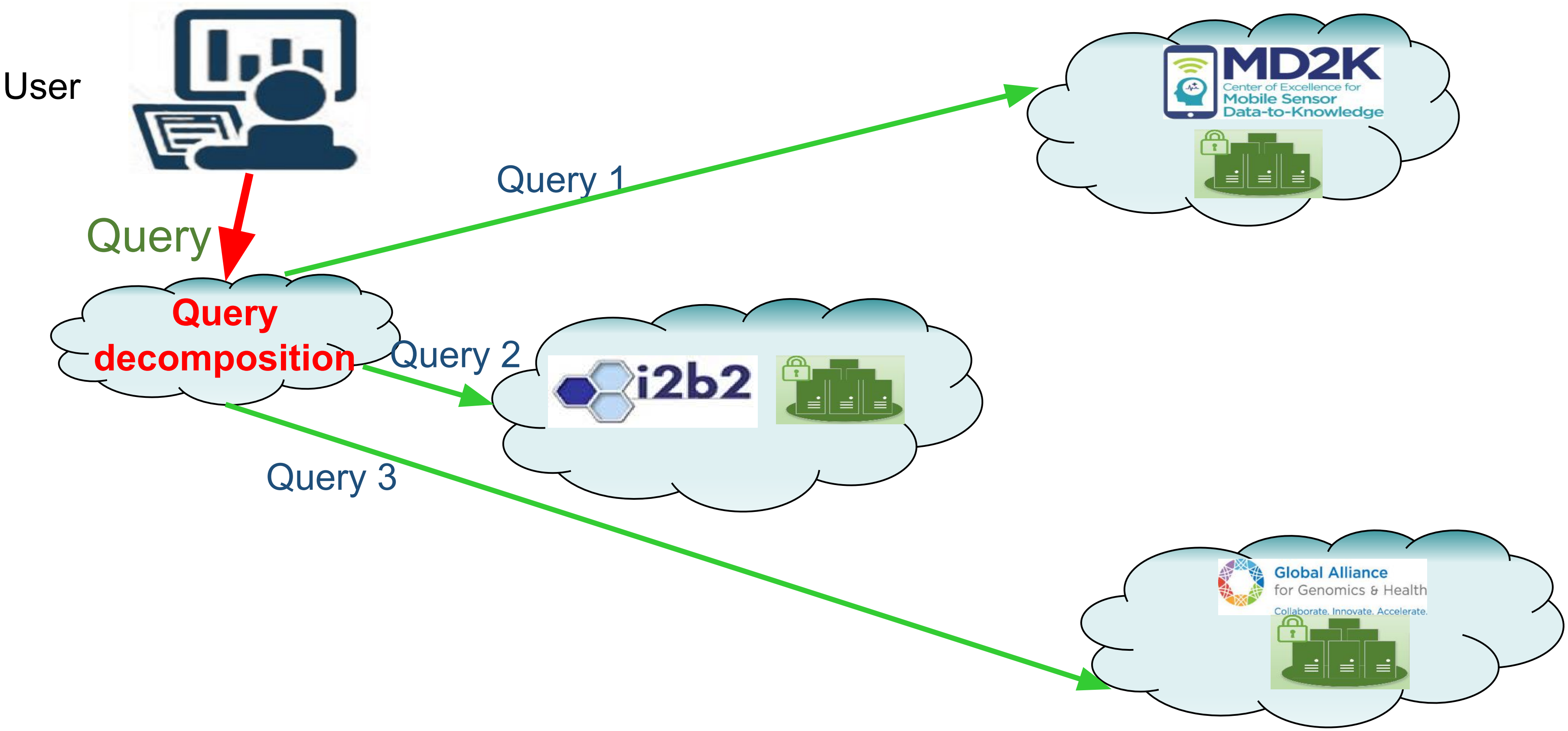
Secure aggregating counts of relevant patients



Select count(*) from MD2K A, i2b2 B, GA4GH C where A.caloriesDay>600 and B.smoker=true and C.r123140=C

iDASH
bio CADDIE
HIPAA Compliant
FISMA
MD2K
Center of Excellence for Mobile Sensor Data-to-Knowledge
i2b2
Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

Framework Overview



User

Query

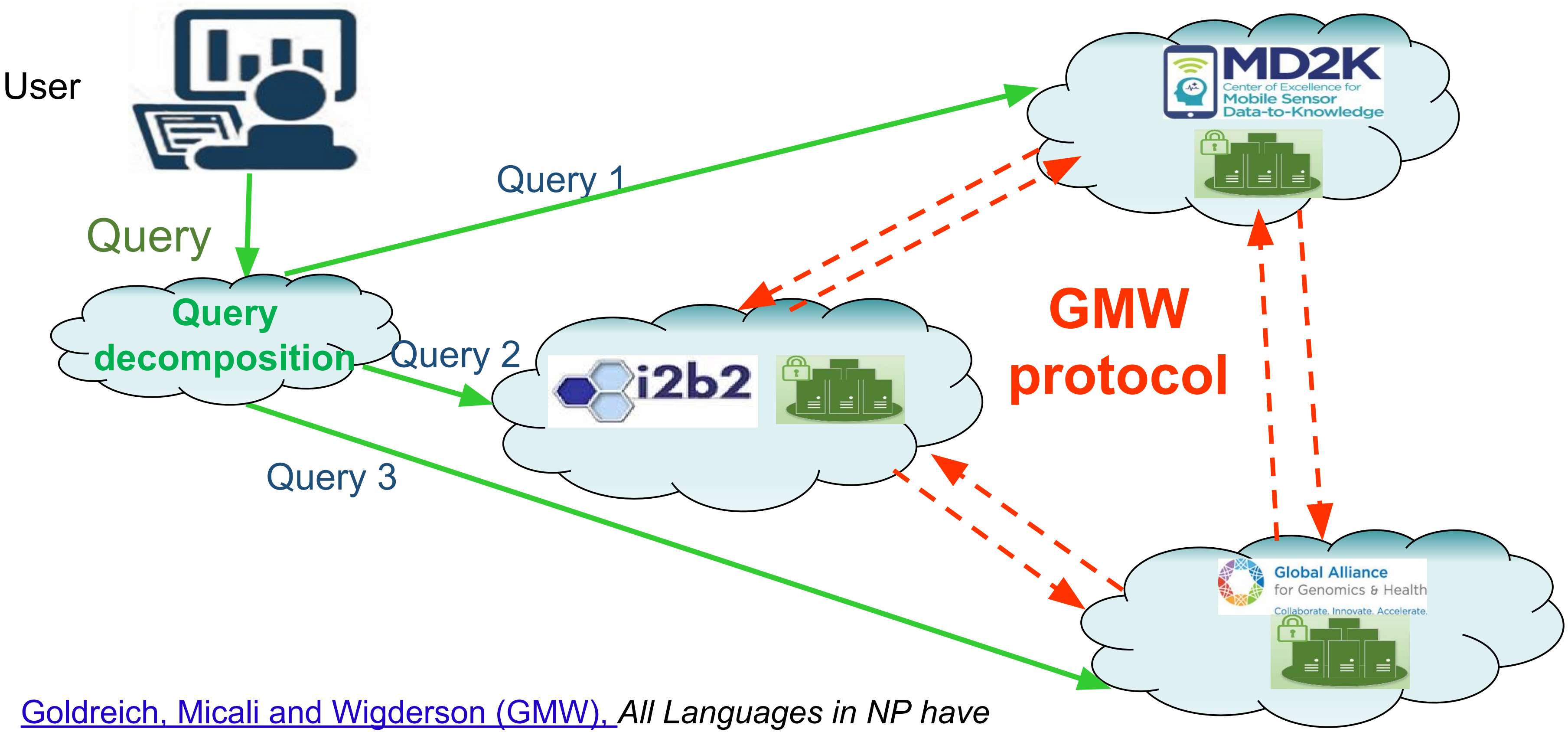
Query decomposition

Query 1

Query 2

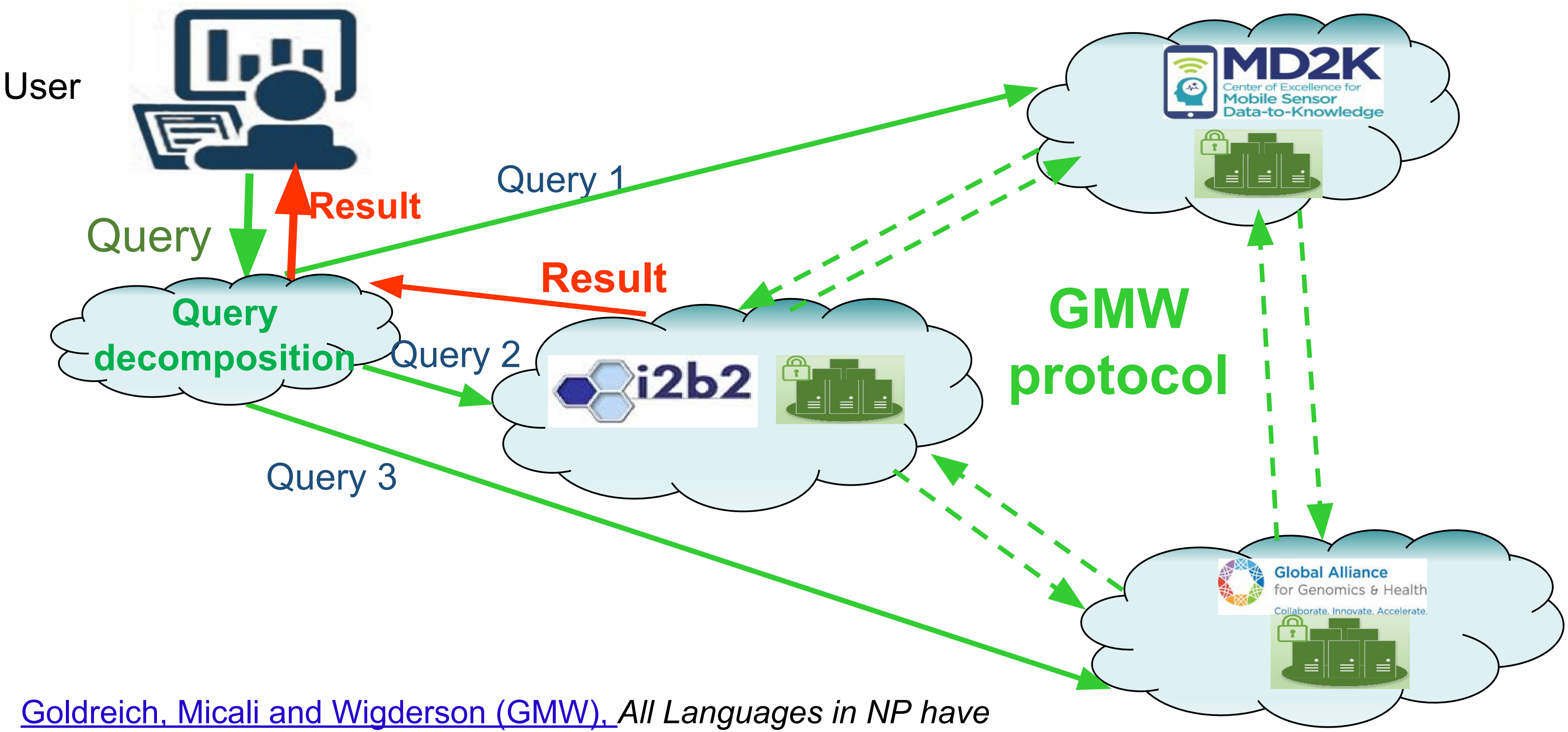
Query 3

Framework Overview


User
Query
Query decomposition
Query 1
Query 2
Query 3
MD2K
Center of Excellence for Mobile Sensor Data-to-Knowledge
i2b2
GMW protocol
Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

Goldreich, Micali and Wigderson (GMW), *All Languages in NP have Zero-Knowledge Proofs*, JACM, July 1991

Framework Overview



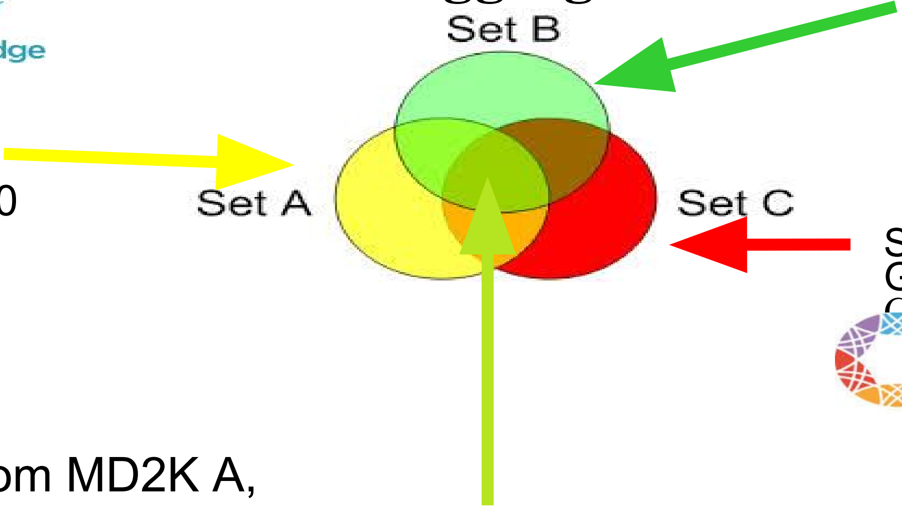
Goldreich, Micali and Wigderson (GMW), *All Languages in NP have Zero-Knowledge Proofs*, JACM, July 1991

Decomposition of the query

MD2K
Center of Excellence for
Mobile Sensor
Data-to-Knowledge

Select IDs from
MD2K where
A.caloriesDay>600

Secure aggregation



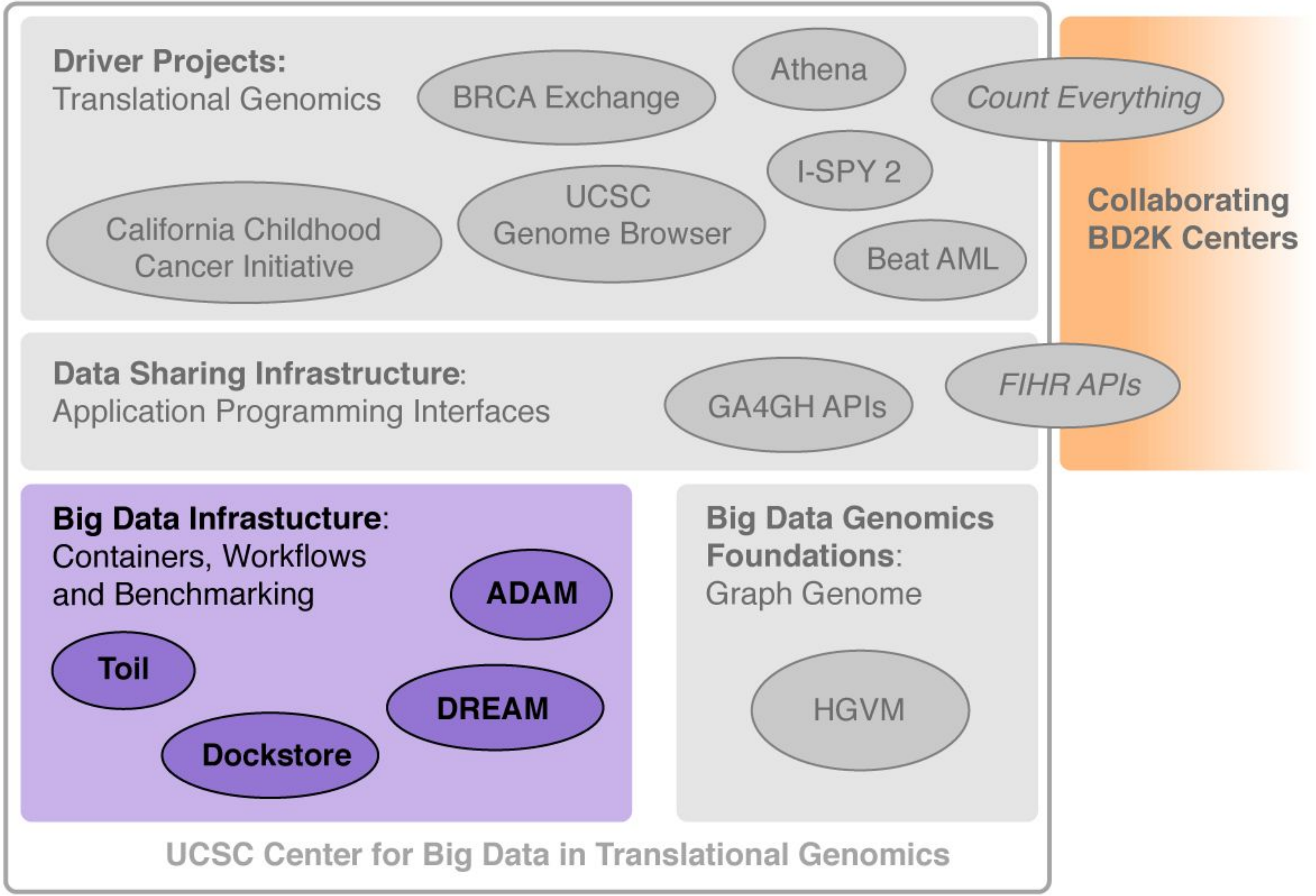
i2b2

Select IDs from
i2b2 where
B.smoker=true

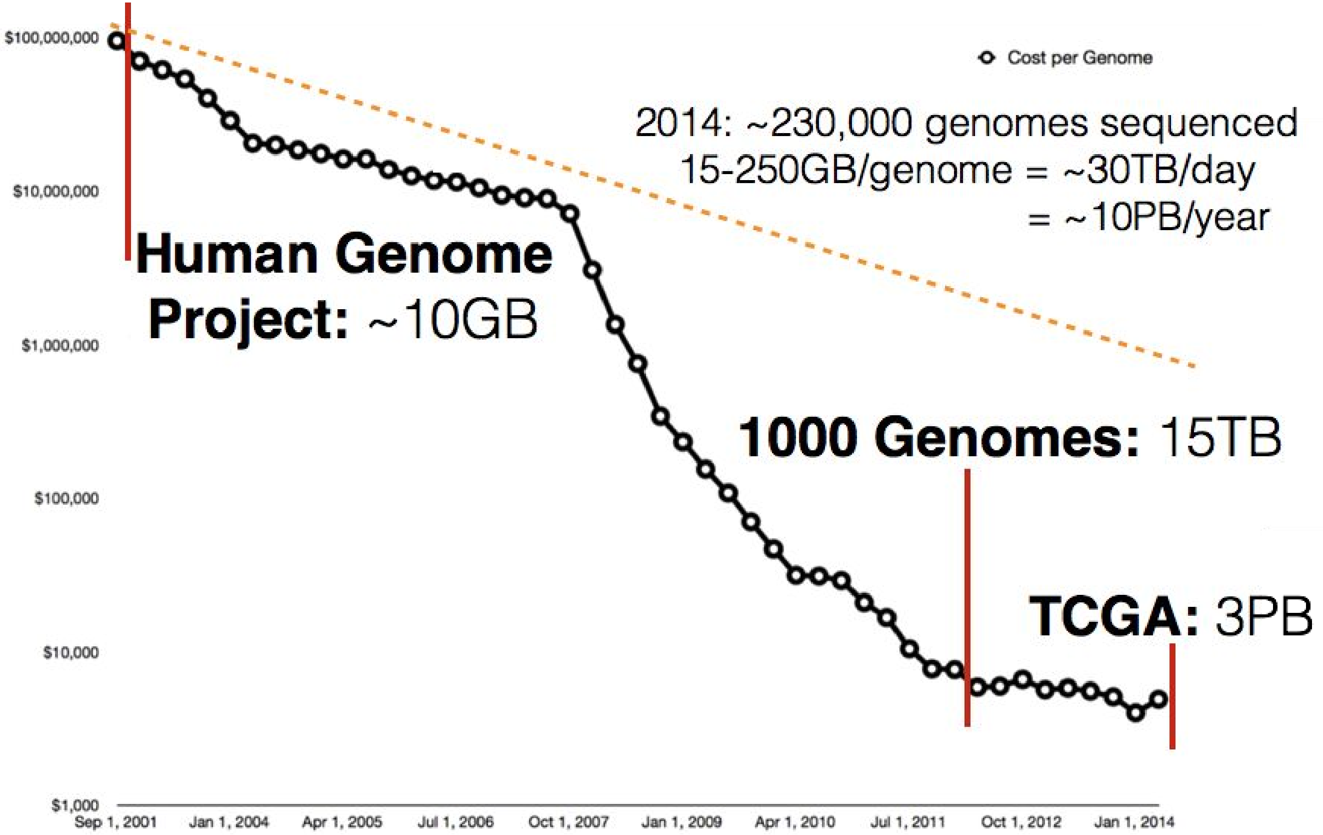
Select IDs from
GA4GH where
C.r123140=C

Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

Select count(*) from MD2K A, i2b2 B, GA4GH C where A.caloriesDay>600 and B.smoker=true and C.r123140=C


Driver Projects:
Translational Genomics
BRCA Exchange
Athena
Count Everything
I-SPY 2
California Childhood Cancer Initiative
UCSC Genome Browser
Beat AML
Collaborating BD2K Centers
Data Sharing Infrastructure:
Application Programming Interfaces
GA4GH APIs
FIHR APIs
Big Data Infrastucture:
Containers, Workflows and Benchmarking
ADAM
Toil
DREAM
Dockstore
Big Data Genomics Foundations:
Graph Genome
HGVM
UCSC Center for Big Data in Translational Genomics

Compulsory Moore's Law Slide



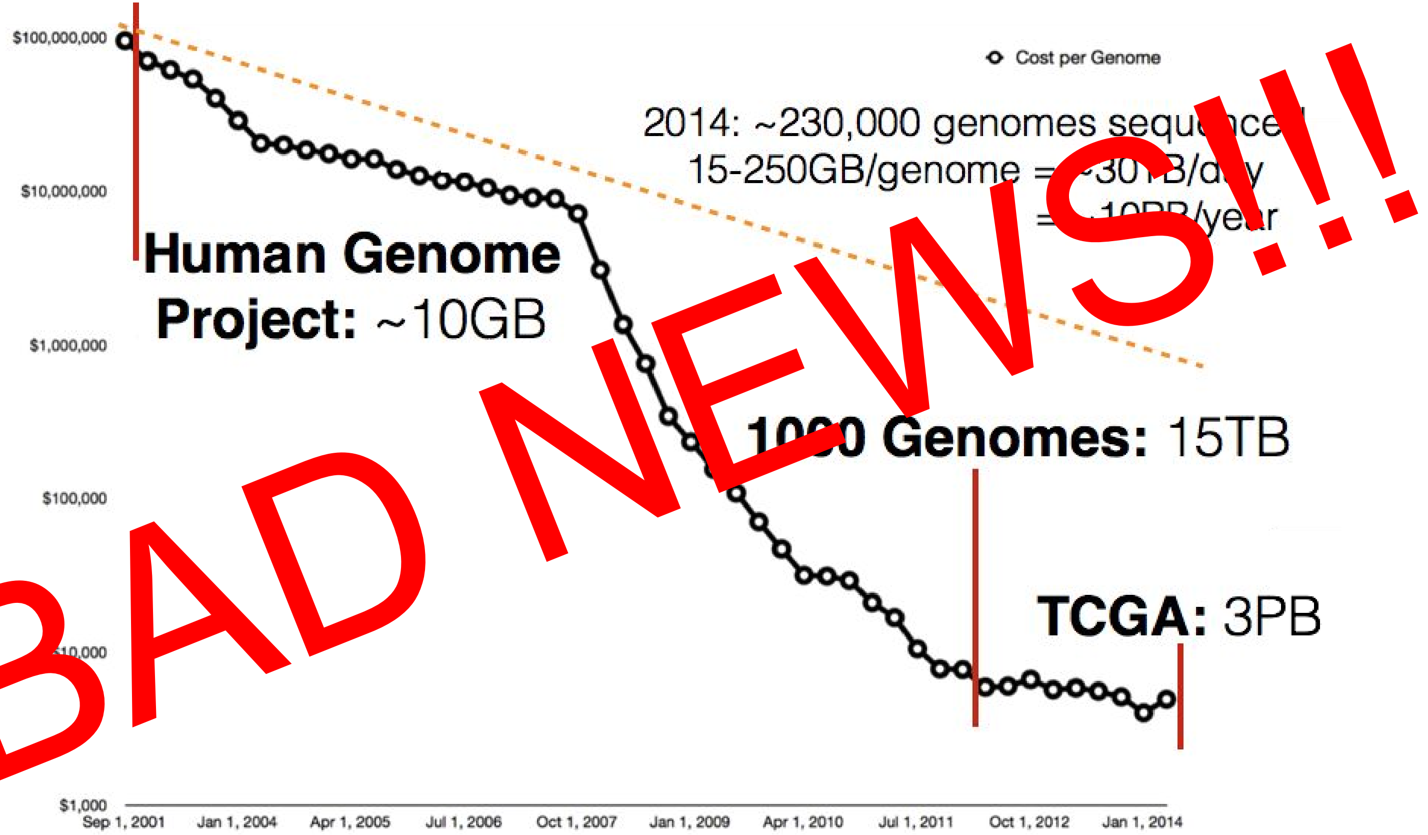
2014: ~230,000 genomes sequenced
15-250GB/genome = ~30TB/day
= ~10PB/year

Human Genome Project: ~10GB

1000 Genomes: 15TB

TCGA: 3PB

Compulsory Moore’s Law Slide



Human Genome Project: ~10GB

2014: ~230,000 genomes sequenced!
15-250GB/genome = ~30TB/day
= ~10PB/year

1000 Genomes: 15TB

TCGA: 3PB

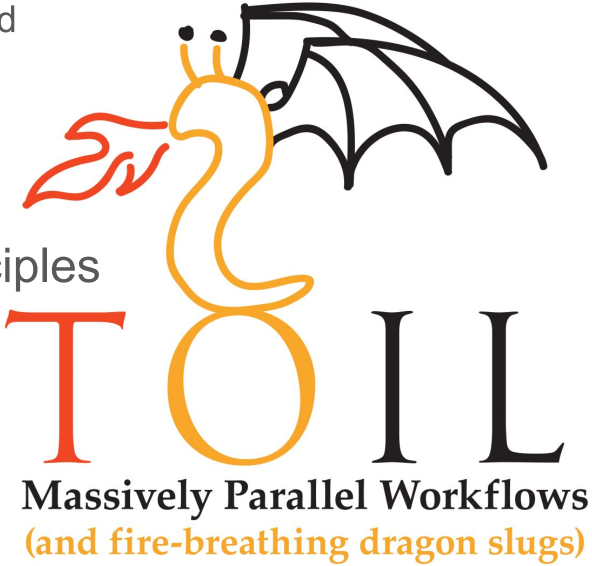
BAD NEWS!!!

Toil — Pipeline Architecture for Genomic Workflows

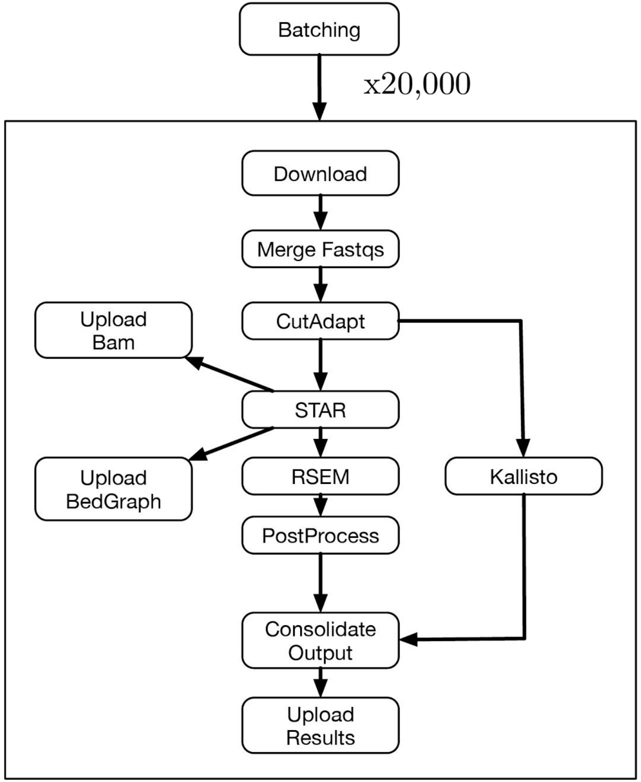
- Massively Scalable — tested on 32,000 cores
- Resumable after failure of *any* node
- Portable — installed with a single command
  - Runs on Amazon, OpenStack, Azure and (soon) Google, and existing HPC environments
- Simple — built entirely in Python
- Supports CWL and (soon) WDL
- Simple API, based on functional programming principles
- Open-source — Fork us!

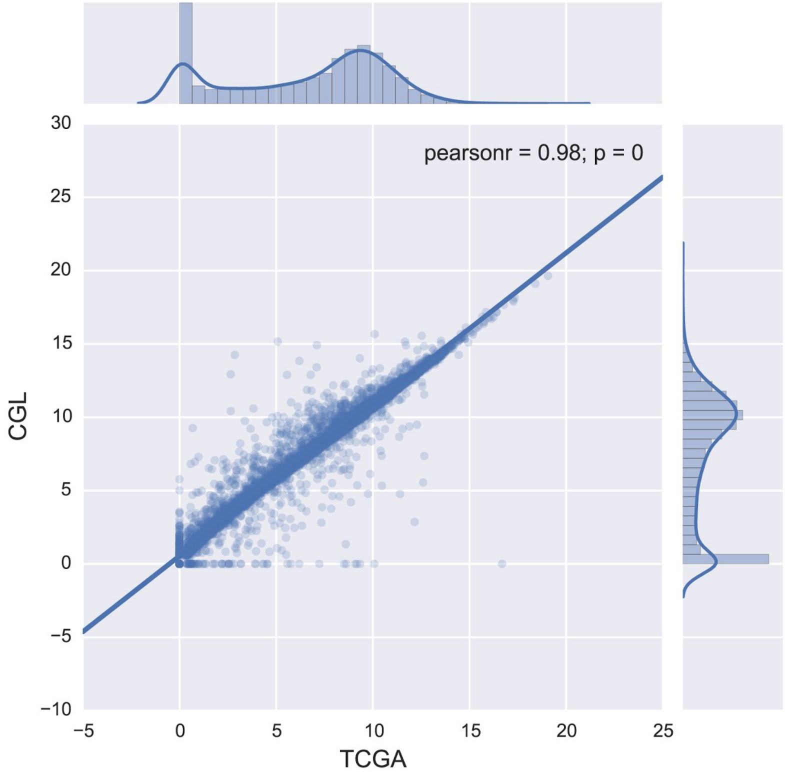
Develop workflows locally…

Deploy at scale without changing source code!

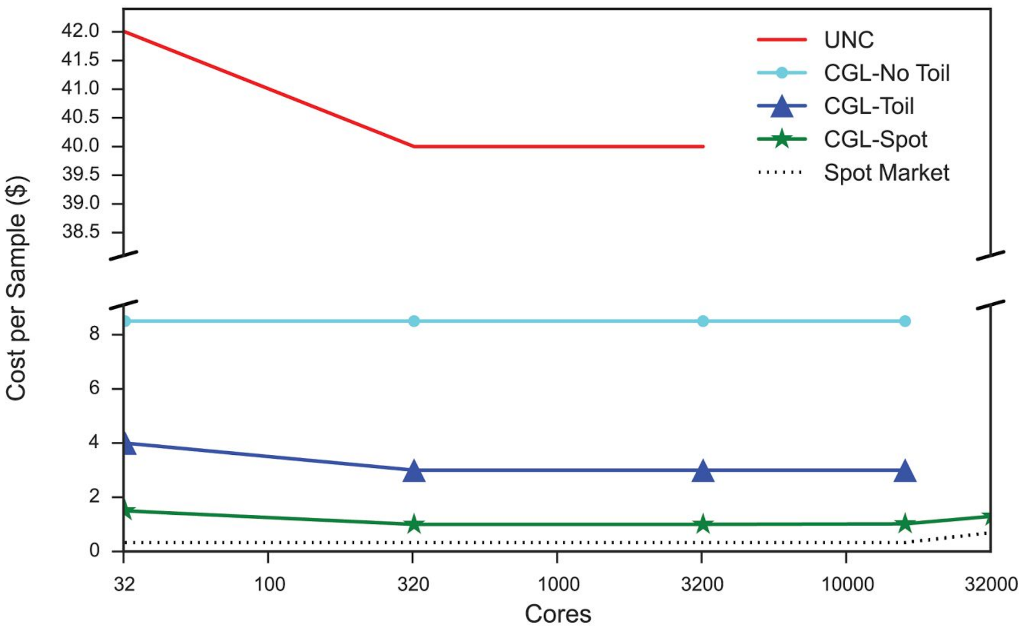

TOIL
Massively Parallel Workflows
(and fire-breathing dragon slugs)

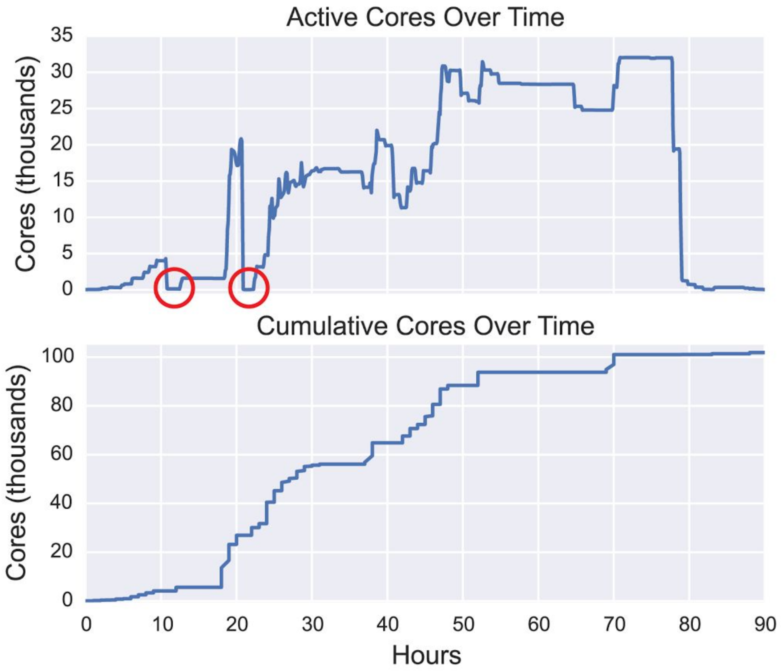
Toil RNASeq Recompute





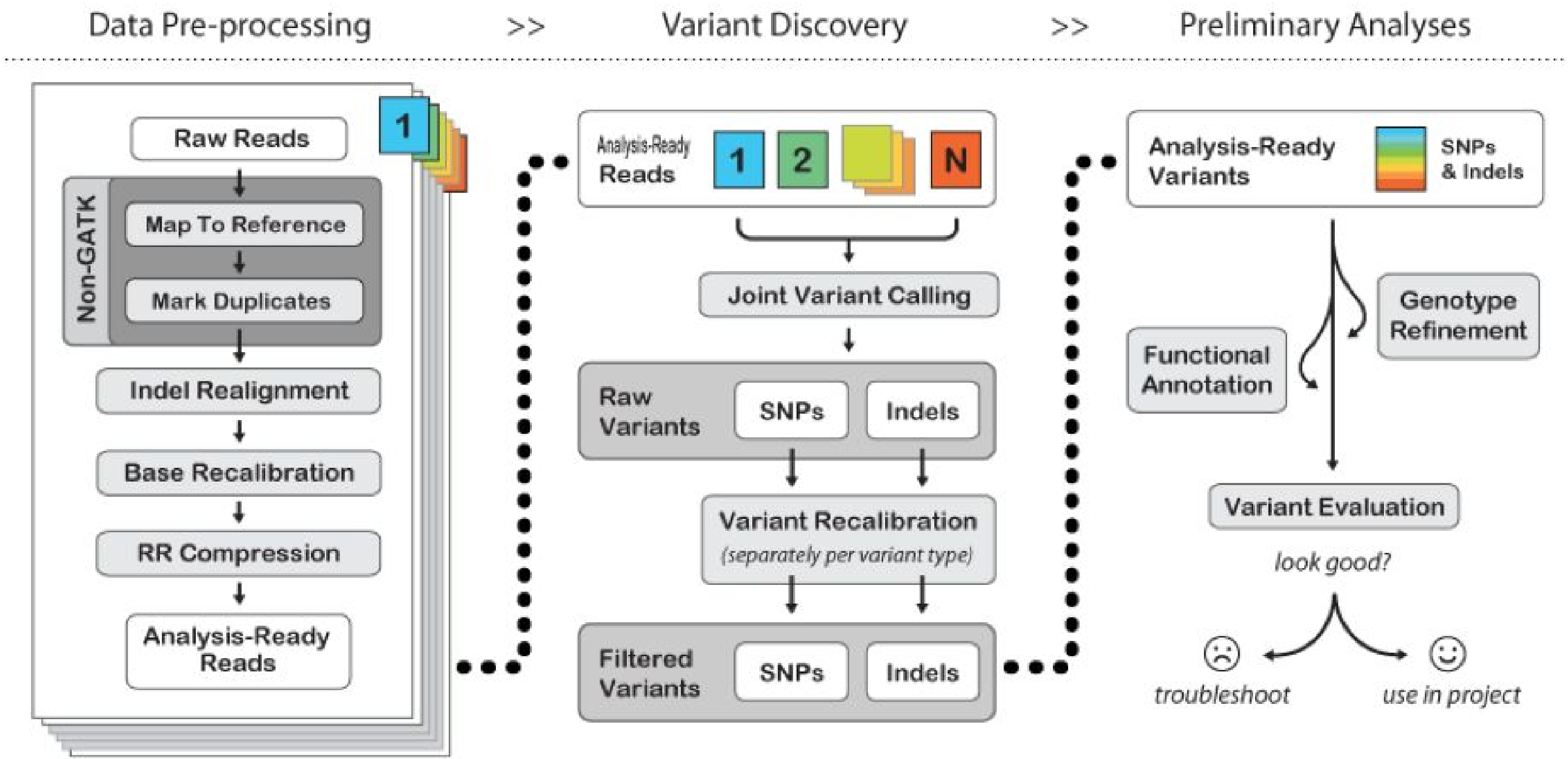
Toil RNA-Seq Recompute



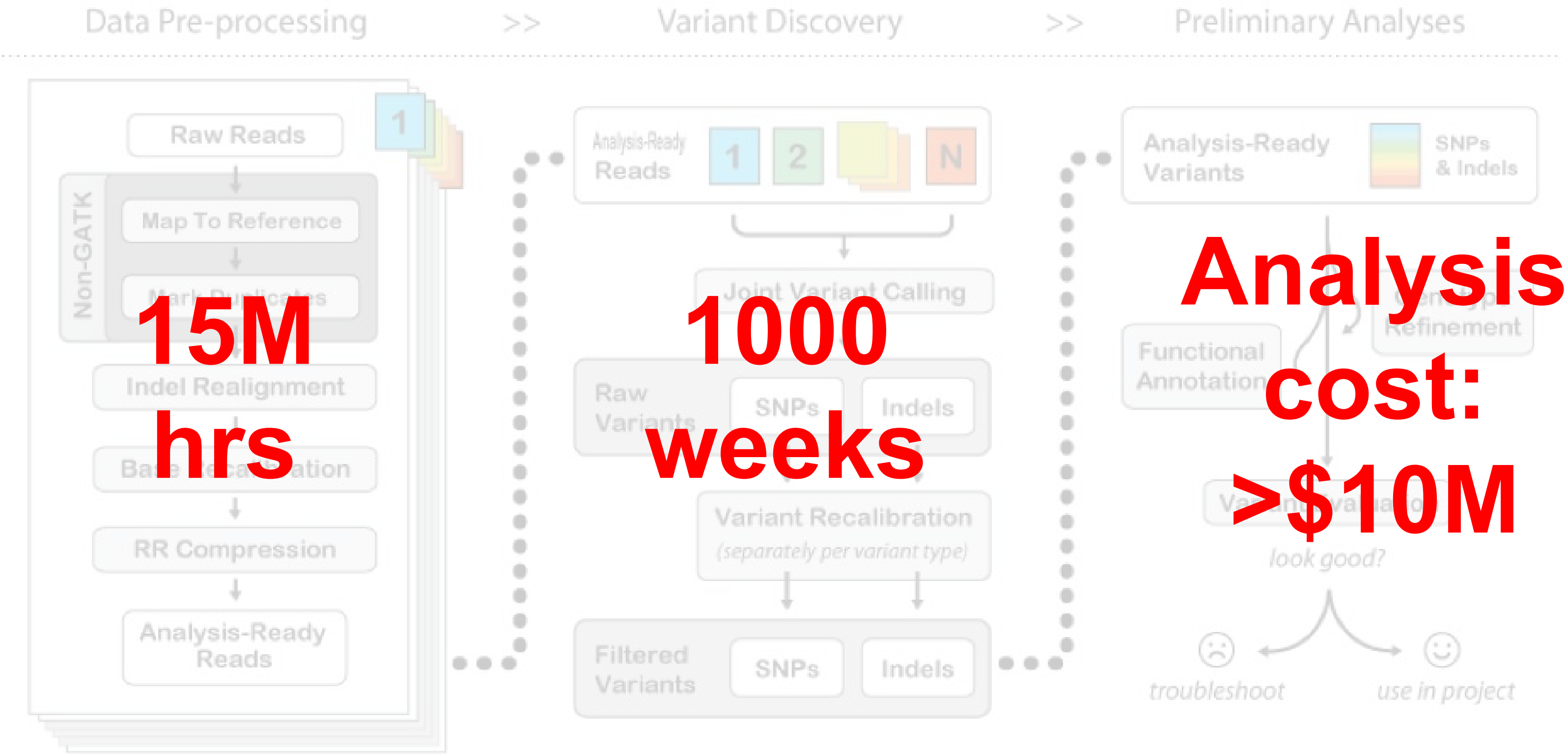


http://biorxiv.org/content/early/2016/07/07/062497

End-to-end variant analysis



End-to-end variant analysis does not scale



15M hrs

1000 weeks

Analysis cost: >$10M

ADAM provides a stack model for genomics

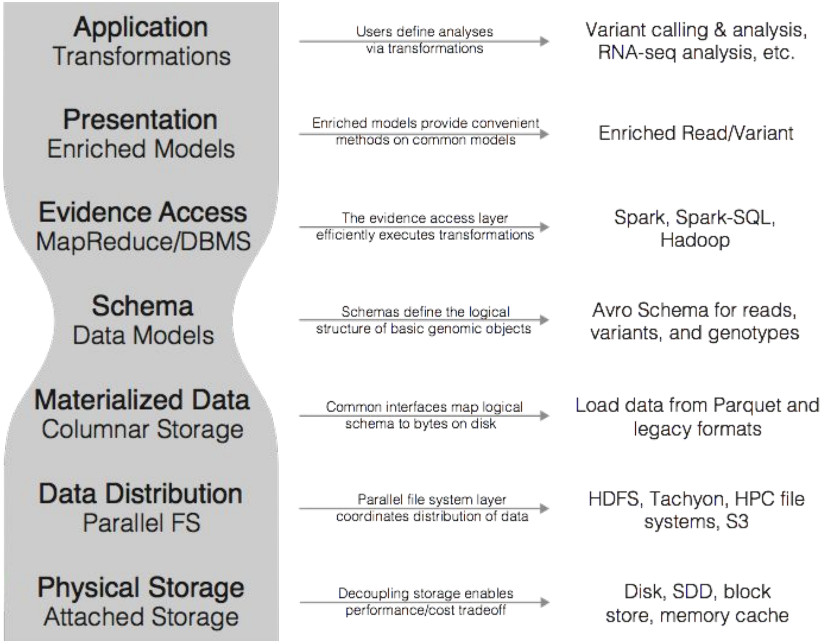
Genomics is built around legacy file formats:

- E.g., SAM/BAM → alignment, VCF → variants, BED/GTF/etc → features
- Manually curated text/binary flat files

We want a narrow waist:

- Can change storage medium, execution system
- Can make use of horizontally scalable systems like Apache Spark

ADAM makes it easy to write parallel algorithms on top of RDDs, instead of against “Genome walker”



Application Transformations — Users define analyses via transformations → Variant calling & analysis, RNA-seq analysis, etc.

Presentation Enriched Models — Enriched models provide convenient methods on common models → Enriched Read/Variant

Evidence Access MapReduce/DBMS — The evidence access layer efficiently executes transformations → Spark, Spark-SQL, Hadoop

Schema Data Models — Schemas define the logical structure of basic genomic objects → Avro Schema for reads, variants, and genotypes

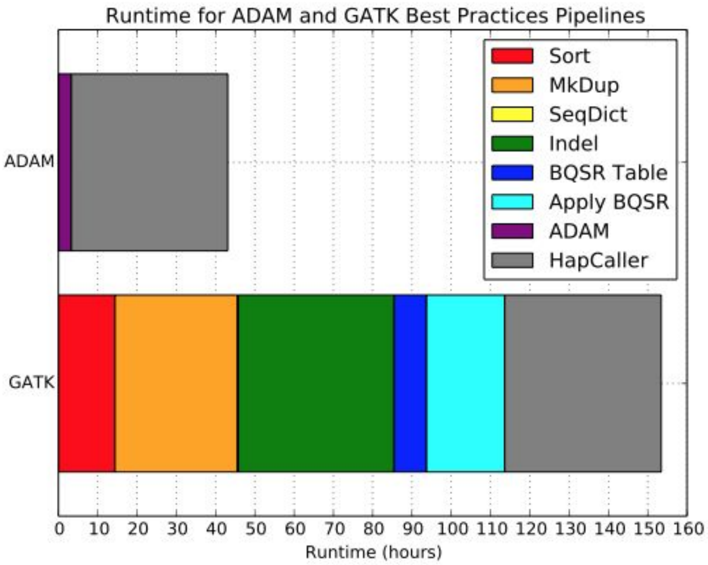
Materialized Data Columnar Storage — Common interfaces map logical schema to bytes on disk → Load data from Parquet and legacy formats

Data Distribution Parallel FS — Parallel file system layer coordinates distribution of data → HDFS, Tachyon, HPC file systems, S3

Physical Storage Attached Storage — Decoupling storage enables performance/cost tradeoff → Disk, SDD, block store, memory cache

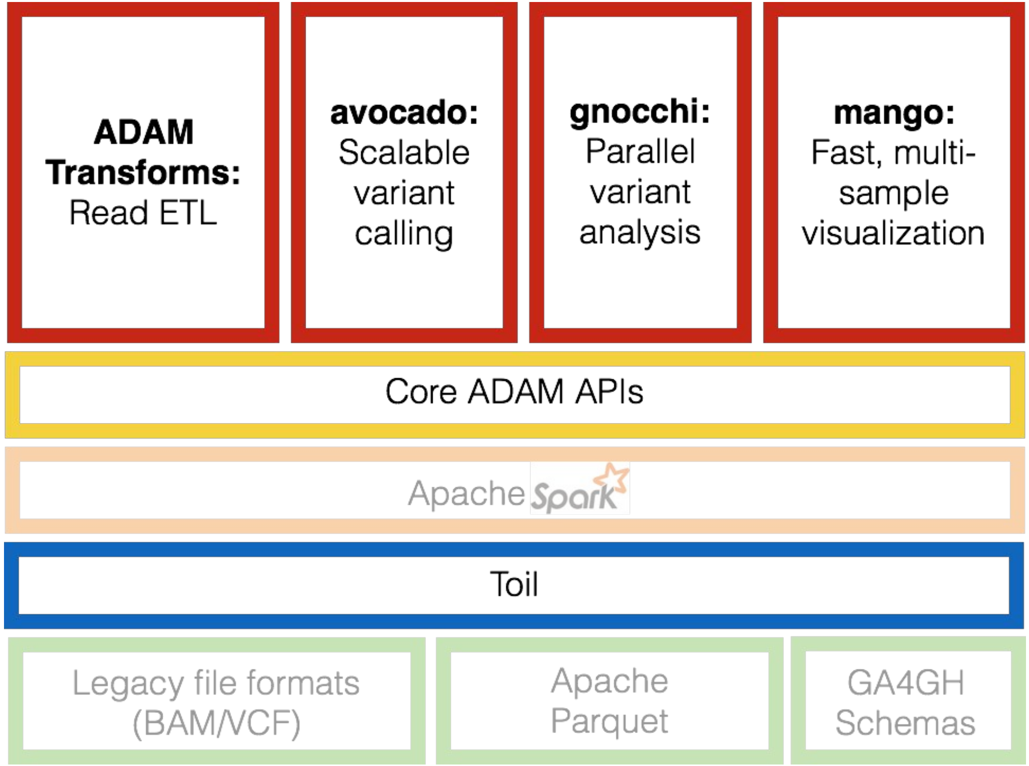
End results compared to legacy systems

- ADAM produces statistically equivalent results to the GATK best practices pipeline
- Our end-to-end pipeline is 3.5x faster while also being 4x cheaper
- In the process of recalling the Simons Genome Diversity Project using ADAM
- We have a working pipeline using both HG19 and GRCh38



Ongoing work

- Completing validation study by recalling SGDP against GRCh38
- ADAM will compete in the VariantDB challenge
- Ongoing work on downstream analysis tools:
  - Avocado
  - Gnocchi
  - Mango


ADAM Transforms: Read ETL
avocado: Scalable variant calling
gnocchi: Parallel variant analysis
mango: Fast, multi-sample visualization
Core ADAM APIs
Apache Spark
Toil
Legacy file formats (BAM/VCF)
Apache Parquet
GA4GH Schemas

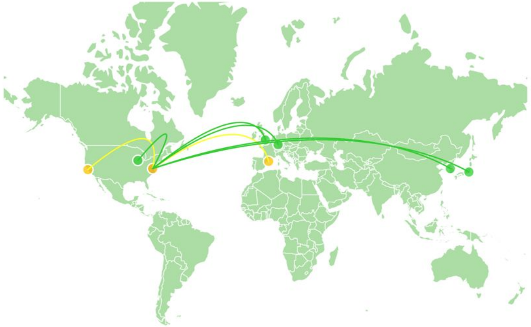
UNIVERSITY OF CALIFORNIA SANTA CRUZ | Genomics Institute

Dockstore

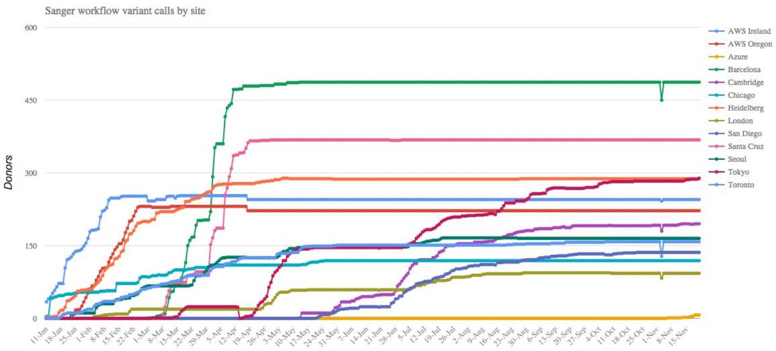
Brian O’Connor
Technical Director, Analysis Core - UCSC Genomics Institute
Consultant - OICR

PCAWG Drove Portable Tool Development

PCAWG
PanCancer Analysis
OF WHOLE GENOMES

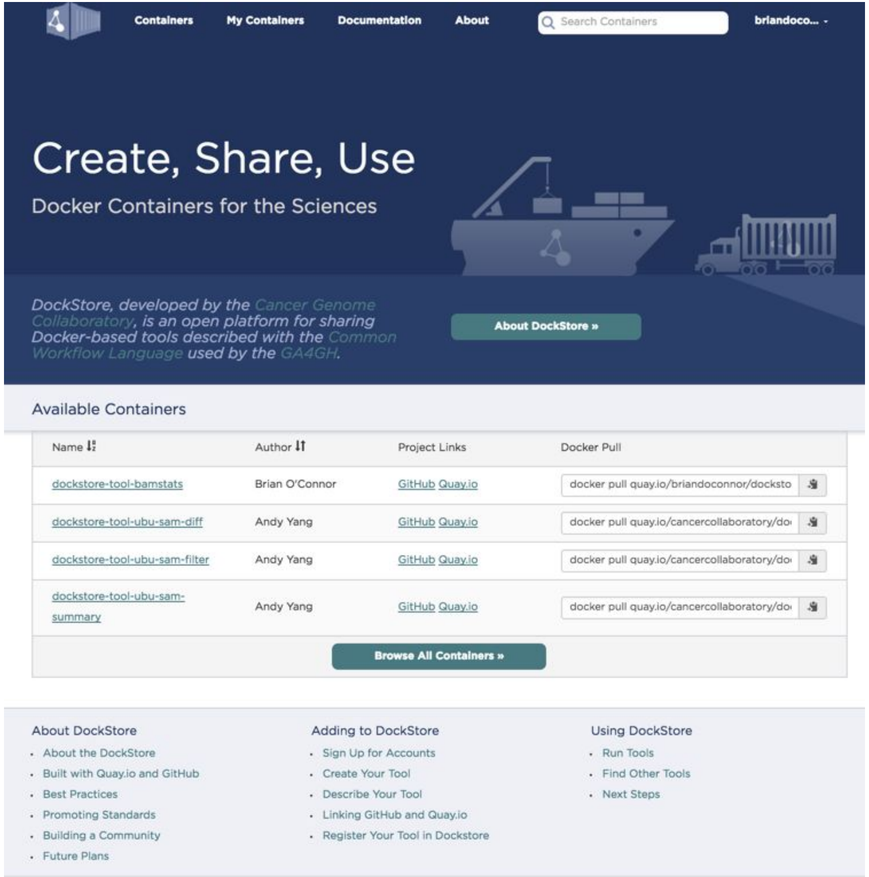


http://pancancer.info
http://dcc.icgc.org/pcawg



- **International Cancer Genome Consortium (ICGC)**
- **~2,800 Cancer Donors**
  - **~1,300 with RNASeq data**
  - **~5,800 Whole Genomes**
  - ***Goal is to consistently analyze data***
- **14** Cloud (and HPC) environments
  - 3 Commercial, 7 OpenStack, 4 HPC
  - ~630 VMs, ~15K cores, ~60TB of RAM
- 8 sites storing and sharing data via GNOS

Dockstore Tour



Create, Share, Use

Docker Containers for the Sciences

DockStore, developed by the Cancer Genome Collaboratory, is an open platform for sharing Docker-based tools described with the Common Workflow Language used by the GA4GH.

Available Containers

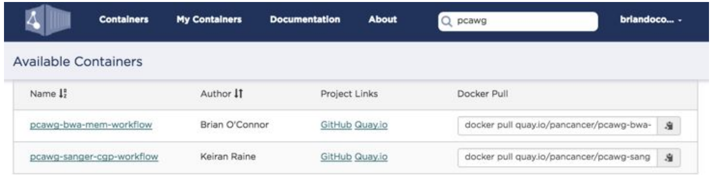
Name	Author	Project Links	Docker Pull
dockstore-tool-bamstats	Brian O'Connor	GitHub Quay.io	docker pull quay.io/briandoconnor/docksto
dockstore-tool-ubu-sam-diff	Andy Yang	GitHub Quay.io	docker pull quay.io/cancercollaboratory/do
dockstore-tool-ubu-sam-filter	Andy Yang	GitHub Quay.io	docker pull quay.io/cancercollaboratory/do
dockstore-tool-ubu-sam-summary	Andy Yang	GitHub Quay.io	docker pull quay.io/cancercollaboratory/do

About DockStore
- About the DockStore
- Built with Quay.io and GitHub
- Best Practices
- Promoting Standards
- Building a Community
- Future Plans

Adding to DockStore
- Sign Up for Accounts
- Create Your Tool
- Describe Your Tool
- Linking GitHub and Quay.io
- Register Your Tool in Dockstore

Using DockStore
- Run Tools
- Find Other Tools
- Next Steps

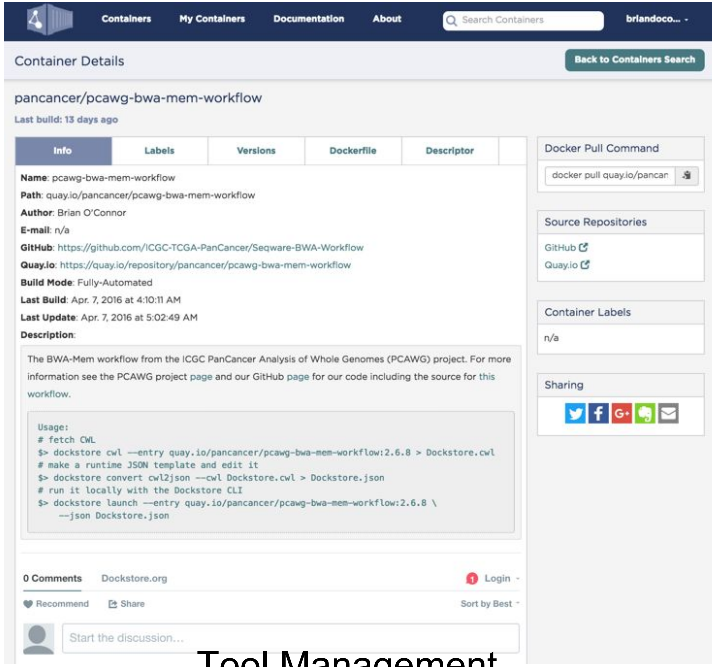
Main Page



Available Containers

Name	Author	Project Links	Docker Pull
pcawg-bwa-mem-workflow	Brian O'Connor	GitHub Quay.io	docker pull quay.io/pancancer/pcawg-bwa-
pcawg-sanger-cgp-workflow	Keiran Raine	GitHub Quay.io	docker pull quay.io/pancancer/pcawg-sang

Search



Container Details

pancancer/pcawg-bwa-mem-workflow

Last build: 13 days ago

Name: pcawg-bwa-mem-workflow
Path: quay.io/pancancer/pcawg-bwa-mem-workflow
Author: Brian O'Connor
E-mail: n/a
GitHub: https://github.com/ICGC-TCGA-PanCancer/Seqware-BWA-Workflow
Quay.io: https://quay.io/repository/pancancer/pcawg-bwa-mem-workflow
Build Mode: Fully-Automated
Last Build: Apr. 7, 2016 at 4:10:11 AM
Last Update: Apr. 7, 2016 at 5:02:49 AM
Description:

The BWA-Mem workflow from the ICGC PanCancer Analysis of Whole Genomes (PCAWG) project. For more information see the PCAWG project page and our GitHub page for our code including the source for this workflow.

```
Usage:
# fetch CWL
$> dockstore cwl --entry quay.io/pancancer/pcawg-bwa-mem-workflow:2.6.8 > Dockstore.cwl
# make a runtime JSON template and edit it
$> dockstore convert cwl2json --cwl Dockstore.cwl > Dockstore.json
# run it locally with the Dockstore CLI
$> dockstore launch --entry quay.io/pancancer/pcawg-bwa-mem-workflow:2.6.8 \
    --json Dockstore.json
```

Docker Pull Command

docker pull quay.io/pancan

Source Repositories

GitHub
Quay.io

Container Labels

n/a

Sharing

Tool Management

DREAM CHALLENGES
powered by Sage Bionetworks

Sage BIONETWORKS

Building communities to advance science.

DREAM Challenges: crowdsourcing quantitative solutions in biomedicine

Future of DREAM: continuous benchmarking

Continuous evaluation and comparison of methods as new data or new algorithms become available.

What is required to do continuous benchmarking?

- Challenge management & data store
- Containerized tools
- Rerunnable workflows
- Cloud compute

Synapse

docker

COMMON WORKFLOW LANGUAGE

Google Cloud Platform

amazon web services

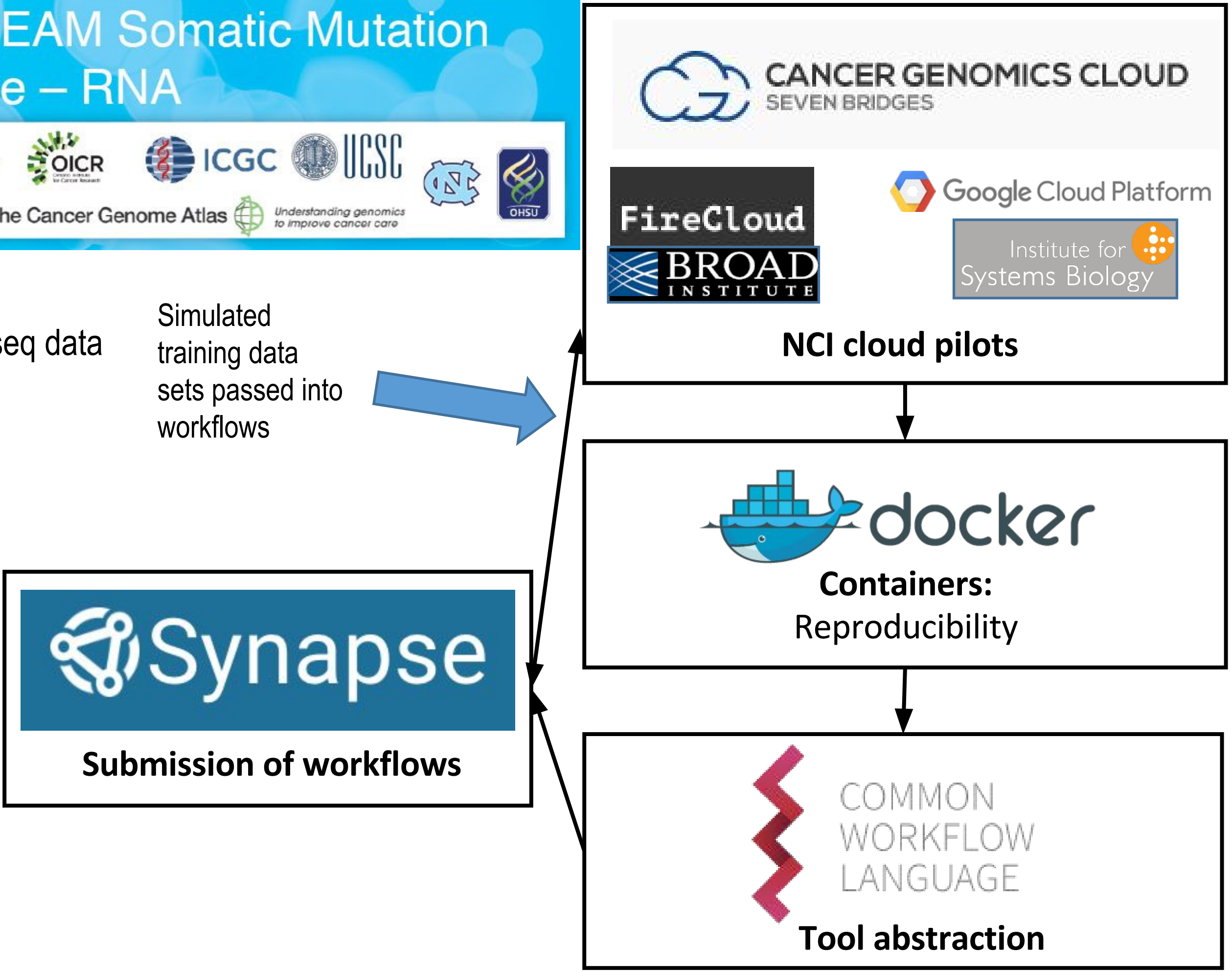
ICGC-TCGA DREAM Somatic Mutation Calling Challenge – RNA

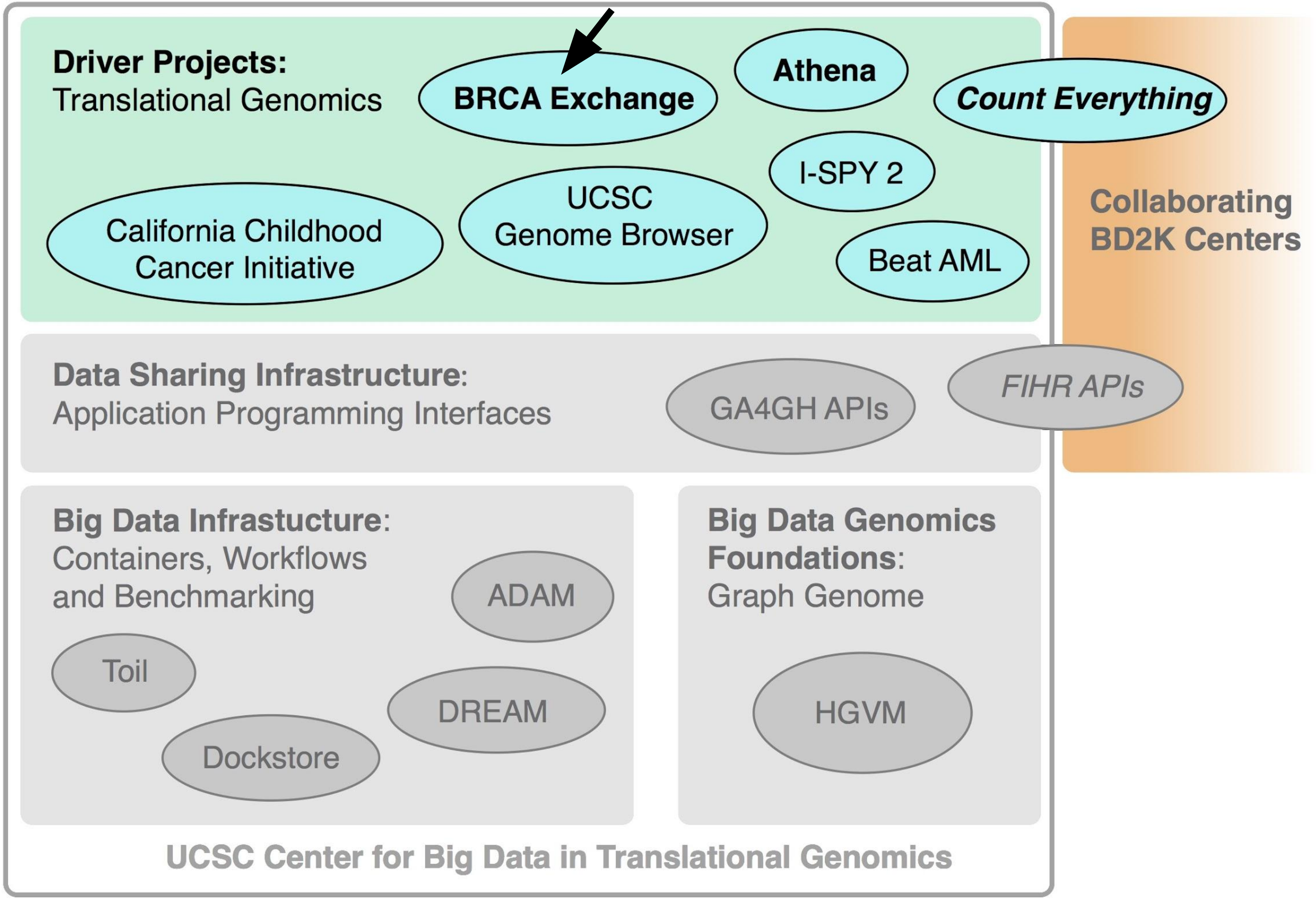
Goals:
Benchmark methods for prediction in RNAseq data on oncology datasets

Challenges:
Identify fusion genes
Quantitate transcript isoforms

Datasets:
Simulated data: in silico data generated from prostate model, with varying number of events from diploid genome at depths 30-250 million reads.

Cell-line spike-ins: breast cancer cell line samples with known isoforms and fusions spiked-in. Sequenced at 50 million reads.





Driver Projects:
Translational Genomics

BRCA Exchange

Athena

Count Everything

I-SPY 2

California Childhood Cancer Initiative

UCSC Genome Browser

Beat AML

Collaborating BD2K Centers

Data Sharing Infrastructure:
Application Programming Interfaces

GA4GH APIs

FIHR APIs

Big Data Infrastucture:
Containers, Workflows and Benchmarking

ADAM

Toil

DREAM

Dockstore

Big Data Genomics Foundations:
Graph Genome

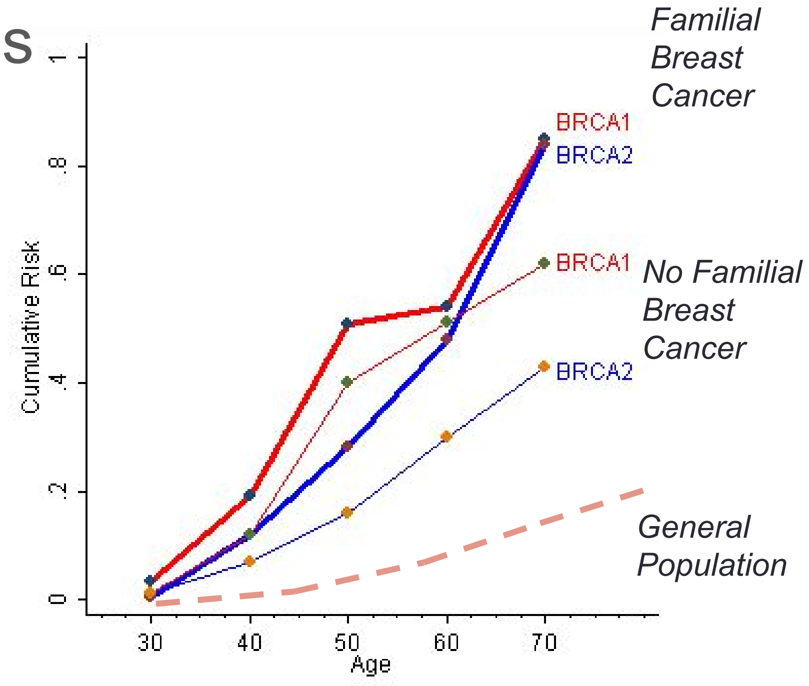
HGVM

UCSC Center for Big Data in Translational Genomics

Motivation for the BRCA Exchange

BRCA variation is relatively common with well known medical implications

- Lifetime risk of developing breast or ovarian with pathogenic BRCA mutation
- Men with pathogenic BRCA mutations are also at risk for prostate cancer
- Drug treatment: PARP inhibitors show effectiveness for BRCA1/2 patients

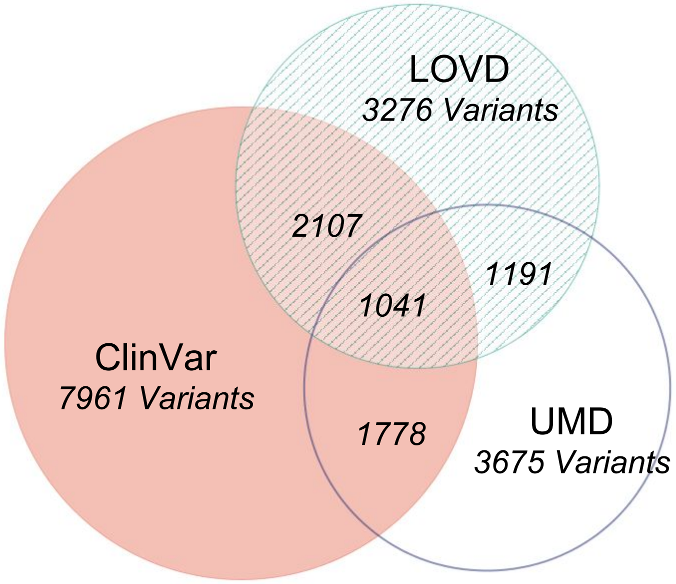


Motivation for the BRCA Exchange

BRCA variation is relatively common with well known medical implications

No single source for BRCA variant information

- ClinVar is incomplete:
  - European projects
  - Individual papers and submitters
  - Some organizations can't pay the cost of preparing a submission

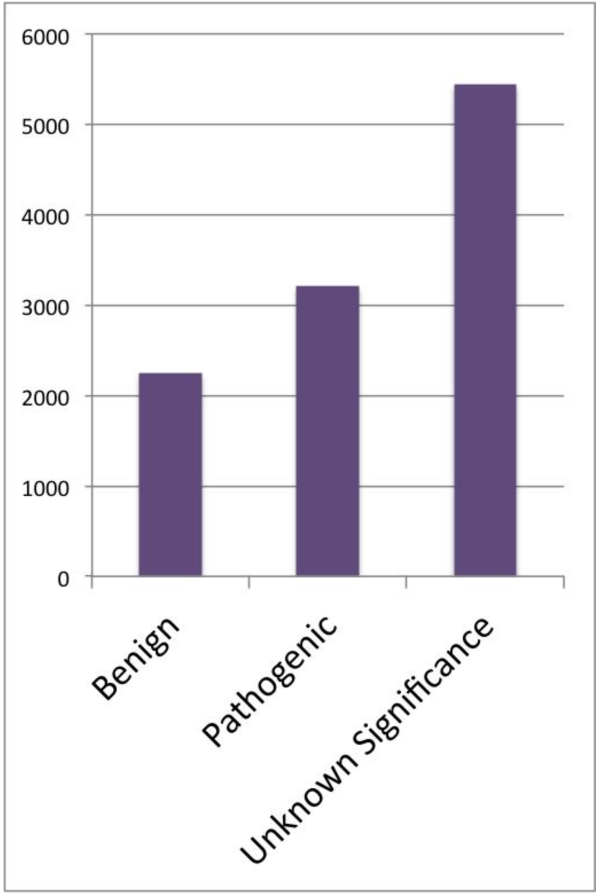


Motivation for the BRCA Exchange

BRCA variation is relatively common with well known medical implications

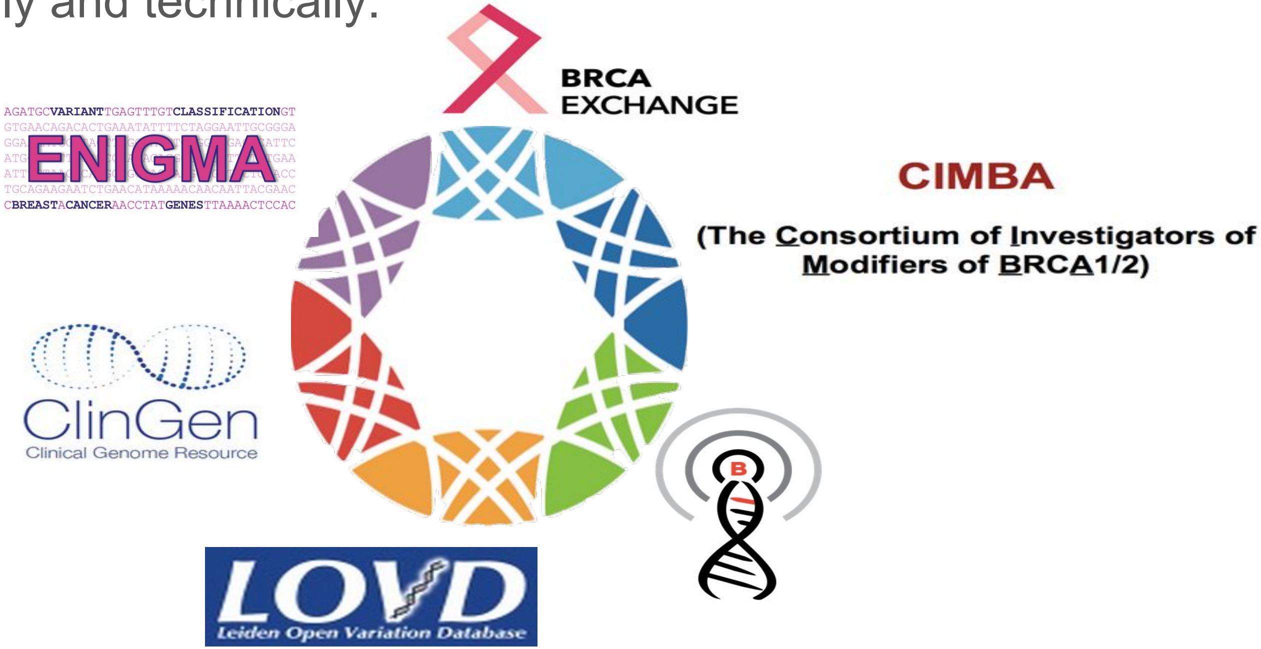
There is no single source for BRCA variant information

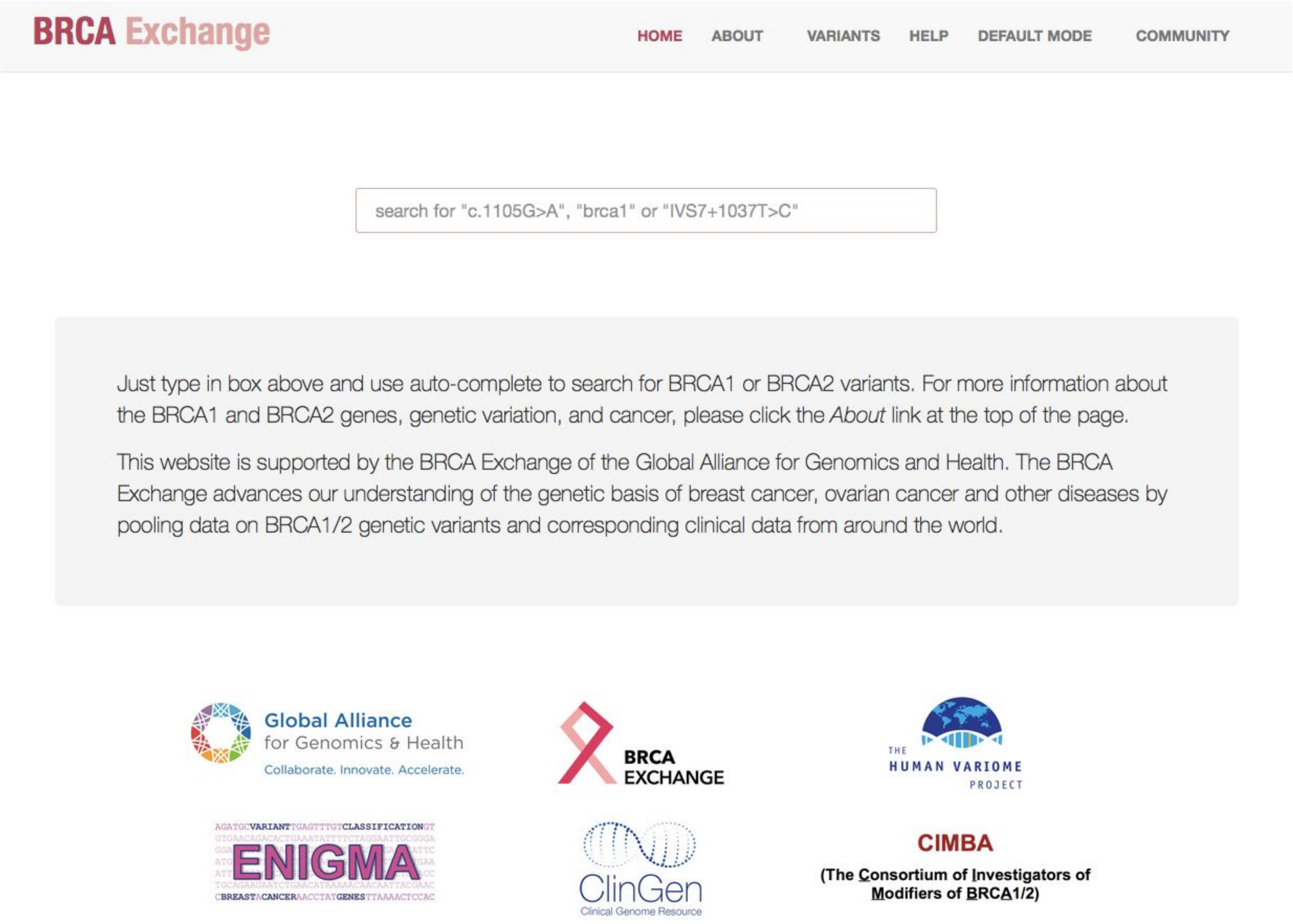
High numbers of VUSs, where classification is limited by being unable to see all the data on a variant



Motivation for the BRCA Exchange within GA4GH/BD2K

Focus on two genes to efficiently and effectively lay the foundations for GA4GH sharing, both legally and technically.


BRCA EXCHANGE
ENIGMA
CIMBA
(The Consortium of Investigators of Modifiers of BRCA1/2)
ClinGen
Clinical Genome Resource
B
LOVD
Leiden Open Variation Database



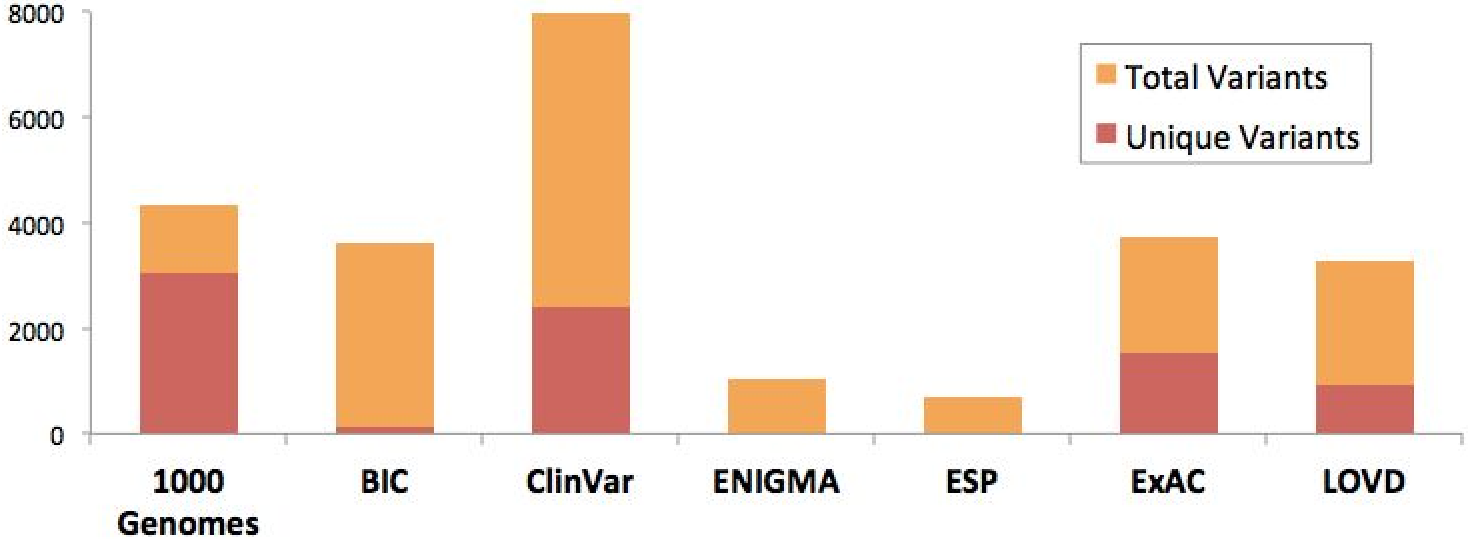
Just type in box above and use auto-complete to search for BRCA1 or BRCA2 variants. For more information about the BRCA1 and BRCA2 genes, genetic variation, and cancer, please click the *About* link at the top of the page.

This website is supported by the BRCA Exchange of the Global Alliance for Genomics and Health. The BRCA Exchange advances 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.

CIMBA

(The Consortium of Investigators of Modifiers of BRCA1/2)

Each repository contributes distinct information on BRCA variation



Combined, BRCA Exchange has
13,500 individual deduplicated variants.

Acknowledgements

Molly Zhang

Charlie Markello

Benedict Paten

Mary Goldman

Brian Craft

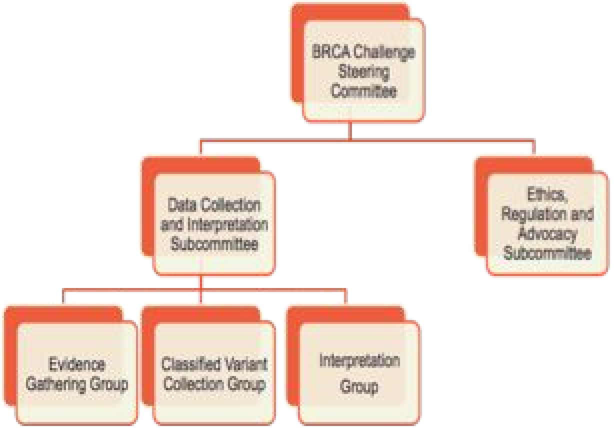
Gunnar Rasch

Rachel Liao

BRCA Challenge Steering Committee

BRCA Challenge Evidence Gathering Group

BRCA Challenge Interpretation Group



BRCA Challenge Steering Committee

Data Collection and Interpretation Subcommittee

Ethics, Regulation and Advocacy Subcommittee

Evidence Gathering Group

Classified Variant Collection Group

Interpretation Group

NIH Big Data to Knowledge (BD2K)

UCSC

Memorial Sloan Kettering Cancer Center

National Cancer Institute

Global Alliance for Genomics & Health