CanDIG

Jonathan Dursi, Hospital for Sick Children

CanDIG

National Analysis of Distributed Private Genomic Data

The CanDIG Platform

Goal:

- A **Canadian** approach to analysis of health research data:
  - **National**-scale populations
  - **Respecting** provincial, institutional stewards **local control** over their data, users.

Project:

- Funded 4 year cyberinfrastructure project, ~5 FTEs and staffing up

The CanDIG Platform

SickKids®
RESEARCH INSTITUTE

PRINCESS MARGARET GENOMICS CENTRE

McGill University and Génome Québec
Innovation Centre
Innovation in action

Canada's Michael Smith
Genome Sciences Centre

CanDIG

National Analysis of Distributed Private Genomic Data

Health Care Data is Provincial

- Each province has made it's own decisions about privacy trade-offs
- Putting data in one place challenging even if it scaled

CanDIG

Health Care Data is Provincial

- National-scale data needed for:
  - Population-scale studies (*e.g.*, cancers)
  - Supporting researchers with national projects

CanDIG

National Analysis of Distributed Private Genomic Data

Health Care Data is Everywhere

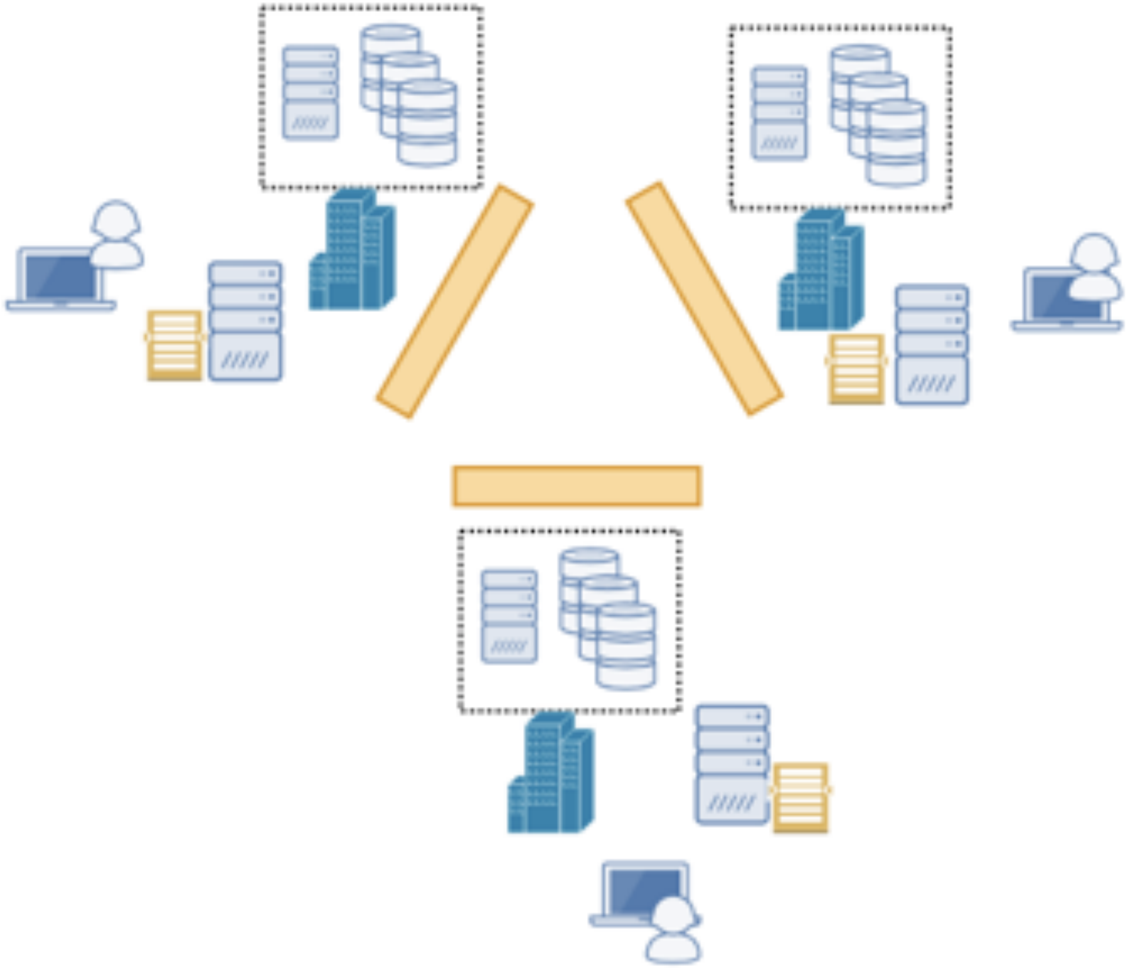
- Torrent of next-generation sequencing data:
  - How to discover it?
  - How to analyze it?
  - How to make it accessible while still maintaining security, privacy?

CanDIG

National Analysis of Distributed Private Genomic Data

CanDIG

- Allow each site to control its own data, users
- Trust authentication of users from other sites, but make own authorization decisions
  - Users may be able to see everything in one set (Co-Is on a national project), only little, with differential privacy, or nothing
- Researchers send queries, aggregate intermediate results to get final answers



CanDIG

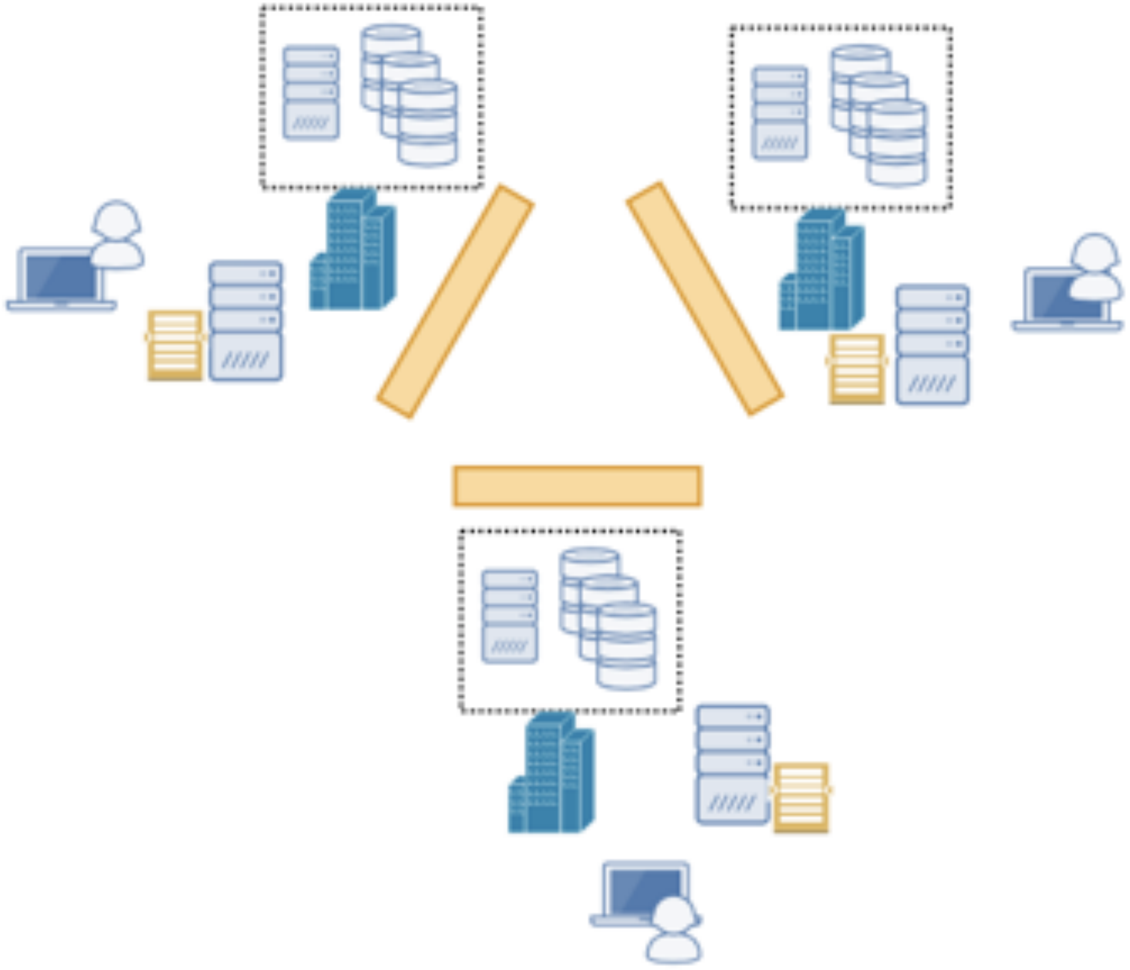
National Analysis of Distributed Private Genomic Data

Platform Principles

- As decentralized as possible
  - Local control of data
  - Minimize centralized infrastructure (maintenance, security)

Distributed Infrastructure for Genomics

- **Fully** distributed
- Participating sites: data providers, compute providers, source of user requests
- Access to data through API requests, directly or via pipelines
- Local sites control access to their data
- Sites authenticate their users
- Researcher queries need only ever see intermediate results, aggregated.



CanDIG *National Analysis of Distributed Private Genomic Data*

Platform Principles

- As decentralized as possible
  - Local control of data
  - Minimize centralized infrastructure (maintenance, security)
- Reduce, reuse, recycle
  - Lots of interesting and new work to do, including challenging algorithmic/privacy work
  - Don’t add to that by re-inventing wheels

CanDIG and the GA4GH

- CanDIG makes use of APIs and data standards from GA4GH (Global Alliance for Genomics and Health)
  - RESTful APIs for variants, reads data, metadata…
  - Schemas for data exchange
  - Security best practices
- Part of several successful projects
- Google Genomics, Microsoft, …

Global Alliance
for Genomics & Health

Matchmaker Exchange

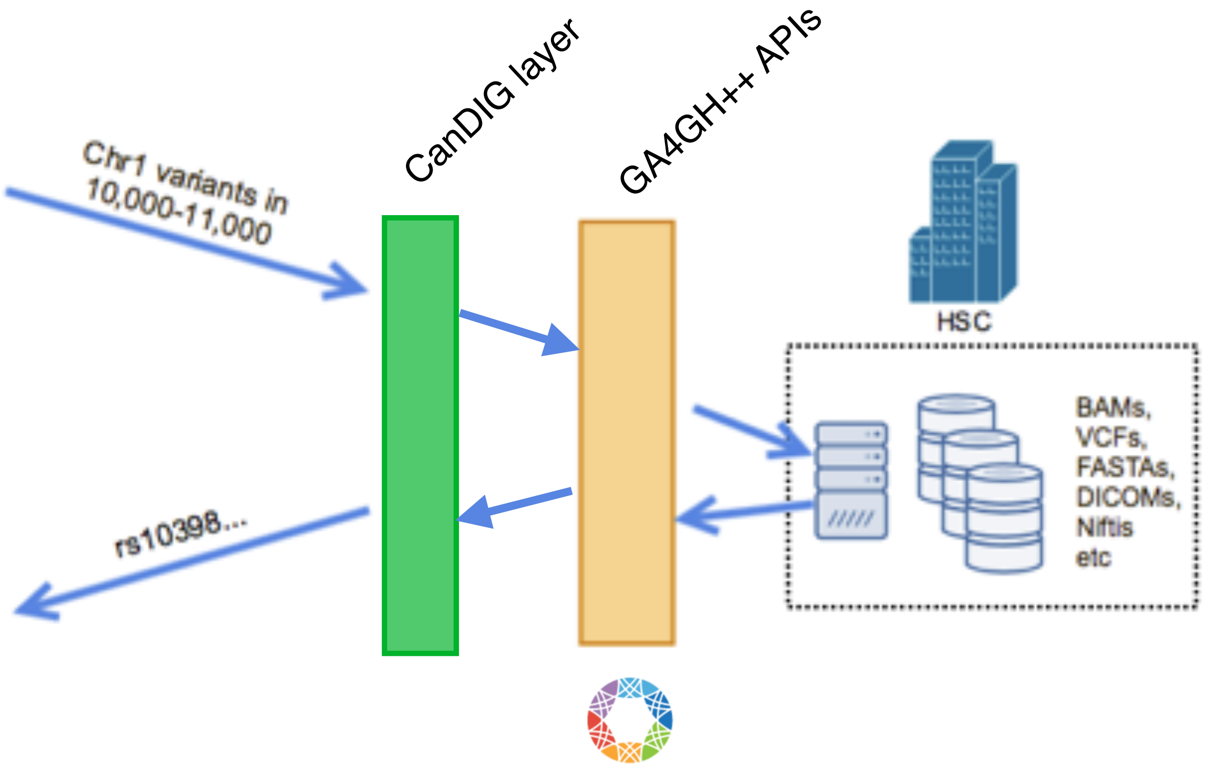
BRCA CHALLENGE

Beacon

Cancer Gene Trust

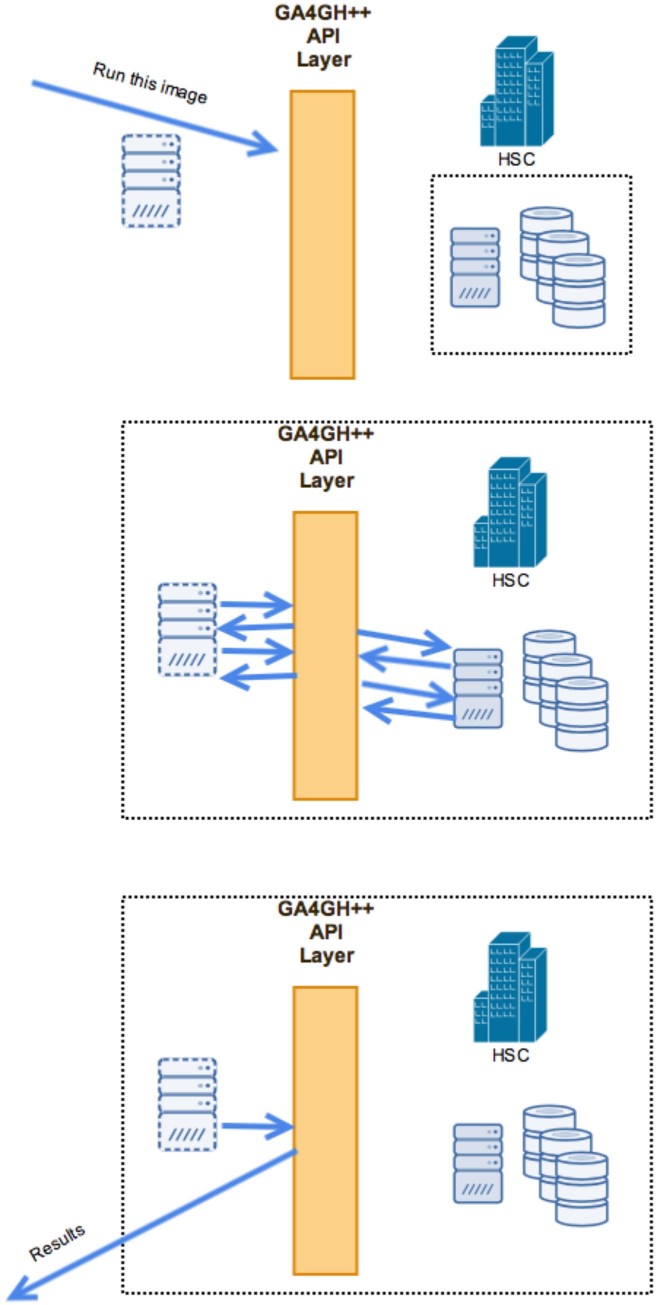
CanDIG and the GA4GH

- All access to data through API.
- Allows abstraction of underlying data store, fine-grained permissions to particular data
- Thin CanDIG layer on top:
  - Richer queries
  - Federation of queries
  - Authentication/Authorization
  - Differential Privacy



CanDIG and the GA4GH

- Queries can be simple queries, handled by the API layer immediately
- Or analyses requiring substantial computation
  - Task Executor Service: run one (or chain) of images against local data
  - Return results through API



CanDIG and OpenID Connect

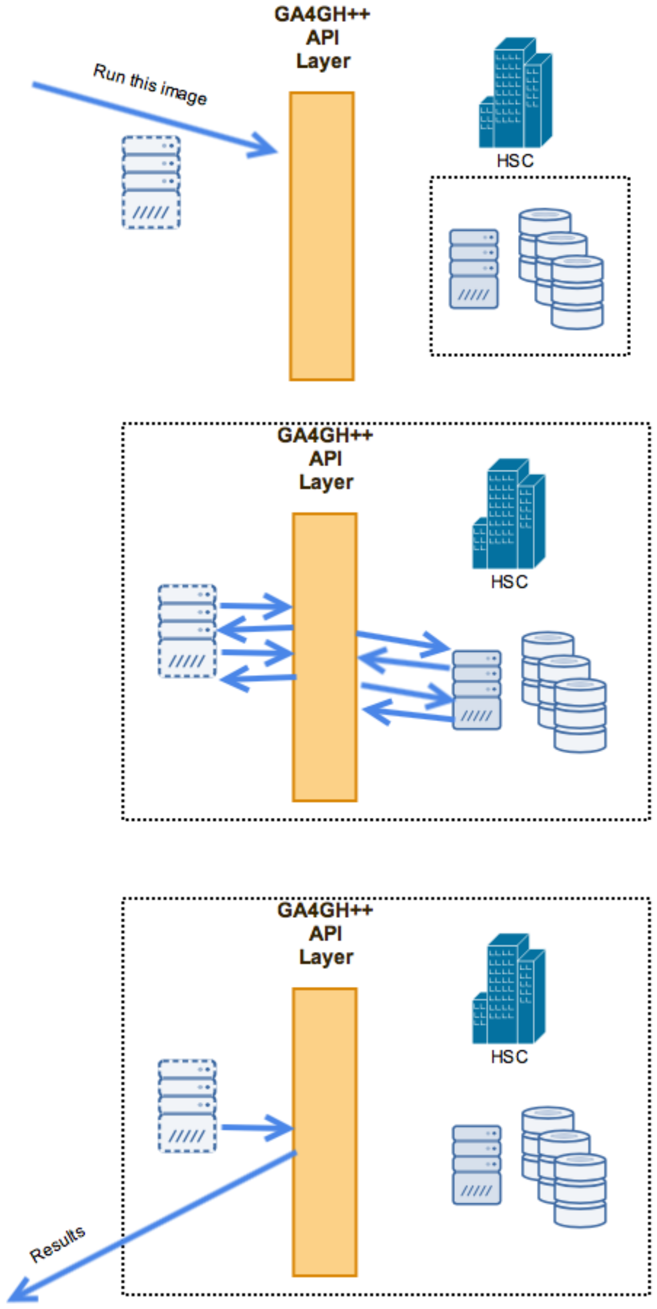
- Use existing well-tested web technologies
  - OpenID Connect for federated authentication
  - KeyCloak to serve OIDC from existing LDAP/AD/etc for each IdP

Platform Principles

- As decentralized as possible
  - Local control of data
  - Minimize centralized infrastructure (maintenance, security)
- Reduce, reuse, recycle
  - Lots of interesting and new work to do, including challenging algorithmic/privacy work
  - Don’t add to that by re-inventing wheels
- Start simple
  - Get simple, working things up and running first
  - Iterate towards desired applications

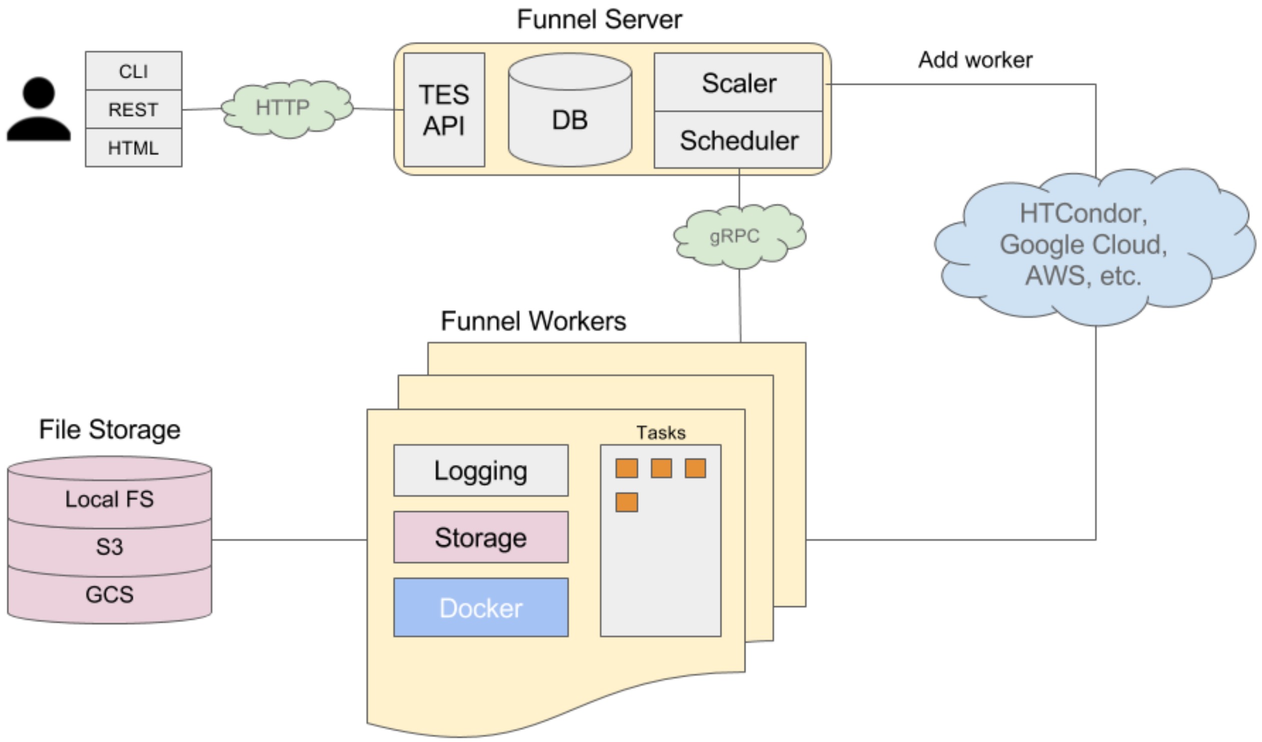
Remote Task Execution

- Want to be able to:
  - Authenticate in
  - Run a bioinformatics task against one of the remote data sets
  - Work done by Steven Li, co-op student, UHN



Remote Task Execution

- Using Funnel, an implementation of GA4GH Task executor definition
- Using Keycloak for OIDC authentication
  - Access to underlying LDAP
- Proof of concept completed



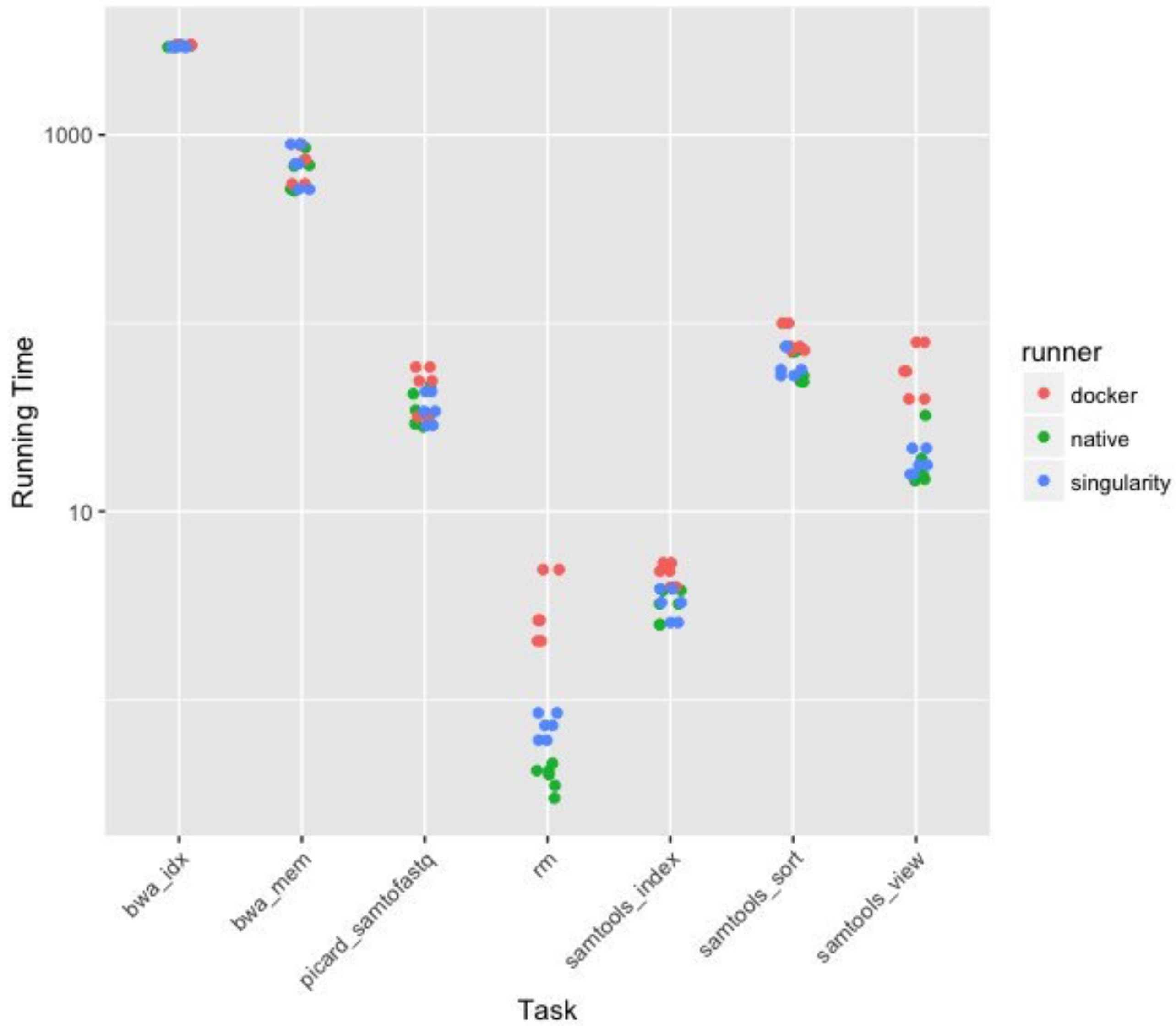
https://ohsu-comp-bio.github.io/funnel/

Containers

- For tasks to be cataloged, distributed, and run on several systems, must be bundled
- Looked at VMs, Docker, rkt, Singularity, Intel Clear Containers
- Need some sort of packaging
- Don’t necessarily need isolation; can handle that at job running time w/ unprivileged users, sandboxes

Containers

- VMs are an awkward fit for discrete, short-lived jobs
- With different container options, performed some benchmarking (https://github.com/CanDIG/images_bakeoff)
- Modest startup cost for docker, perhaps some very modest I/O penalty
- Otherwise quite good performance across all solutions



Singularity or Rkt

- Docker gives us lots of great tooling, and we will use it in the short term (*e.g.*, funnel support)
- But medium term will move to Singularity or Rkt
  - Focus on packaging rather than isolation
  - rkt can easily dial up/down isolation w/o root daemons
- If really needed VM-like isolation, Intel clear containers would be a good choice.

Privacy and Queries

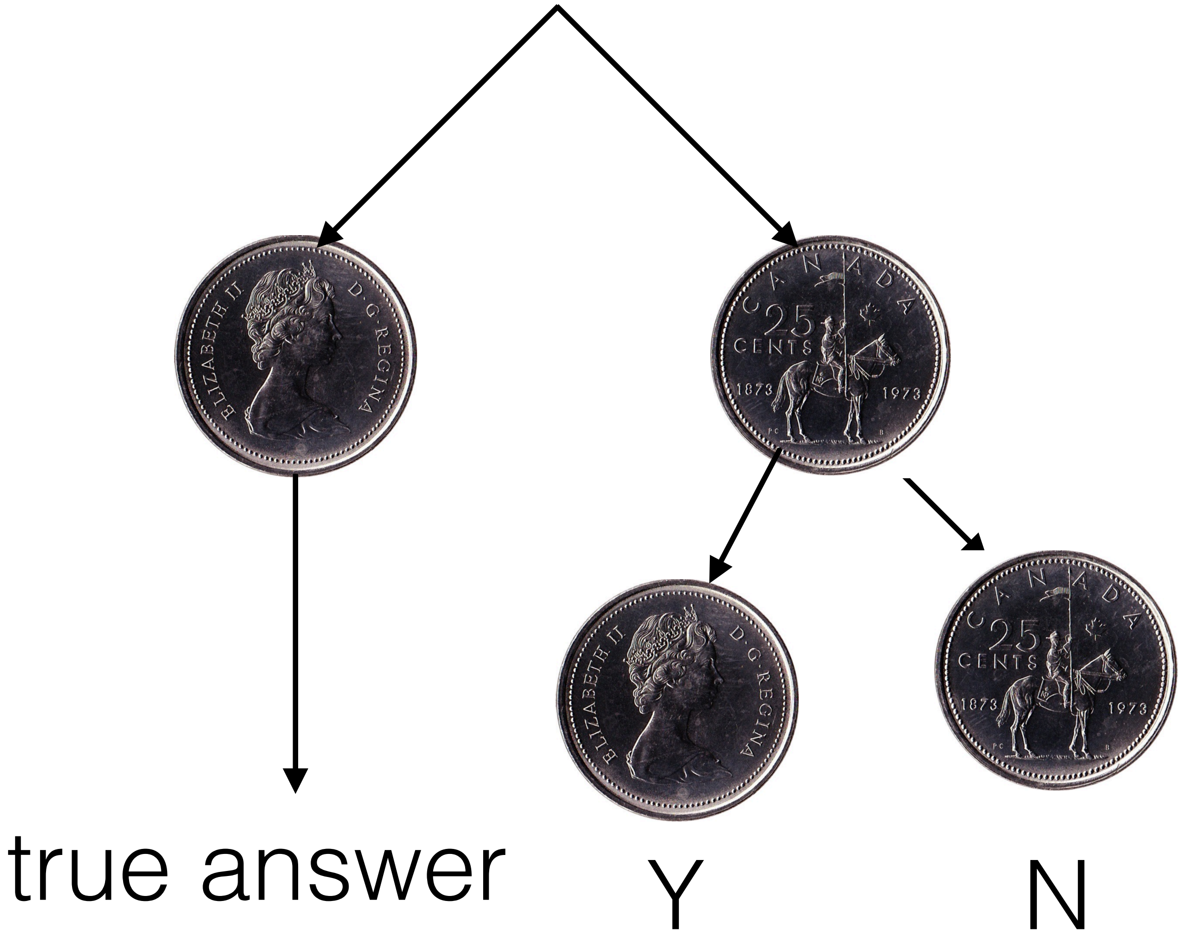
- In many cases, a researcher in a particular project will have complete access to data set
- In other cases, data set can only be accessed at all if privacy of all individuals can be guaranteed
- How can we allow analysis of data while not exposing information of any individual?

Privacy and Queries

- Two approaches:
  - Build queries and applications that only the minimal results are returned - don’t leak extraneous data
  - Add differential privacy for sensitive data sets

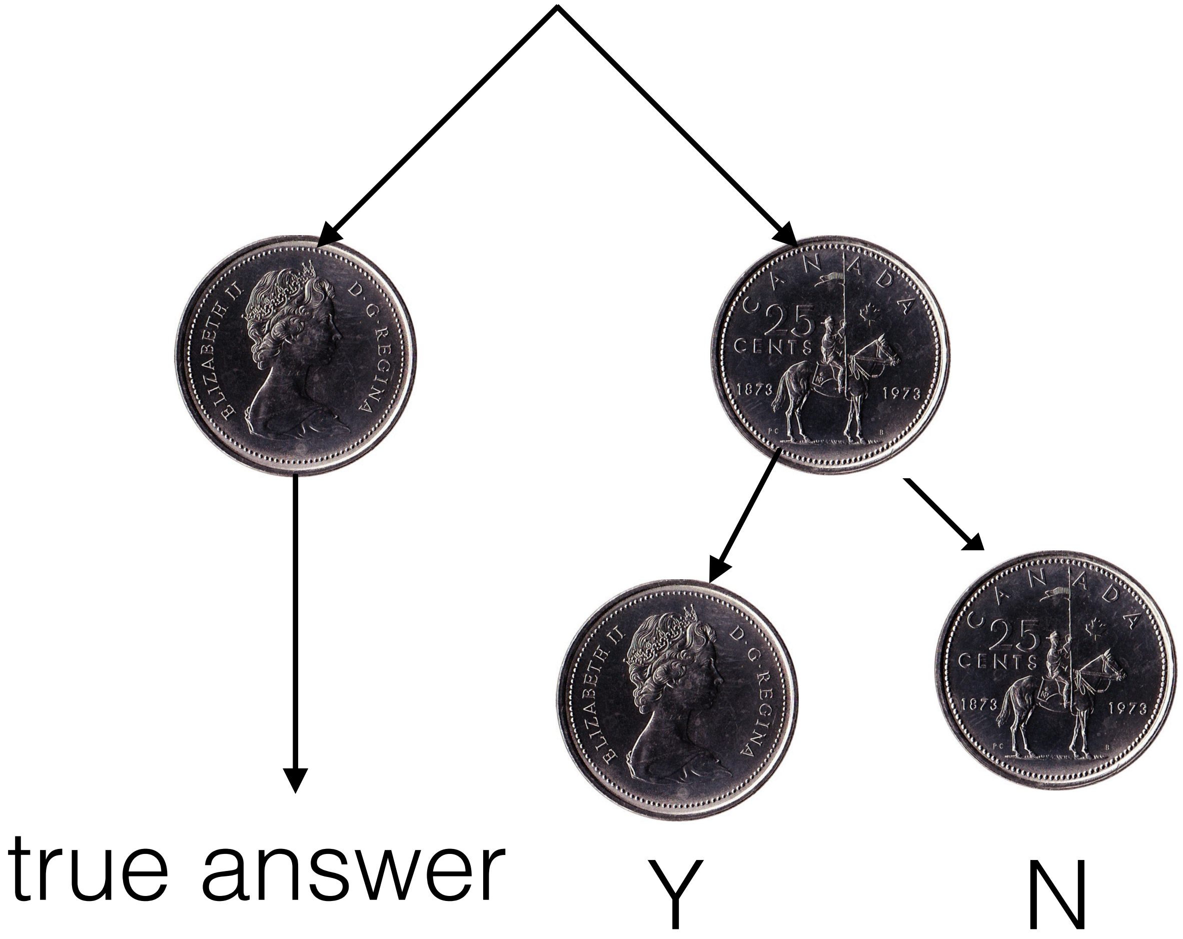
Randomized Response

- Old technique for surveying for behaviours which are illegal or have other stigma attached.
- “Have you, in the last week, listened to Nickleback.”
- p = 0.5 - true answer
- p = 0.5 - random answer
- “bad” answer occurs w/ p = 0.25; “plausible deniability” for any survey respondent.



Randomized Response

- But at the same time, can estimate true *overall* frequencies (and correlations!) knowing the noise model.
- If obtain a frequency f' from the survey instrument, can calculate true frequency f = 2(f' - 1/4)
- Need more samples for given variance, but can get accurate results while protecting each individual's privacy.

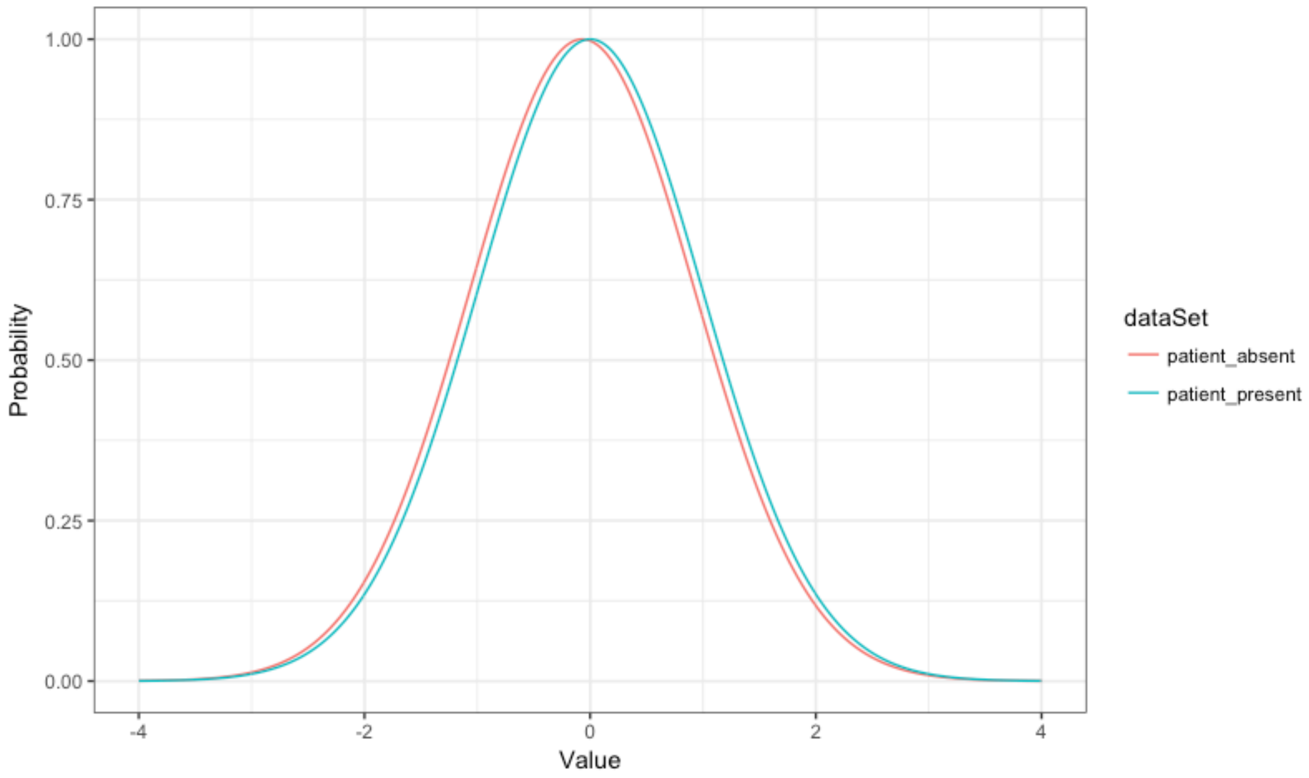


Differential Privacy

- Patient: “There is a quantifiable, minimal, cost to my privacy by participating in this database”.
- Researcher: “I would get an essentially equal distribution of answers from this query if any one row had been absent from the database”.

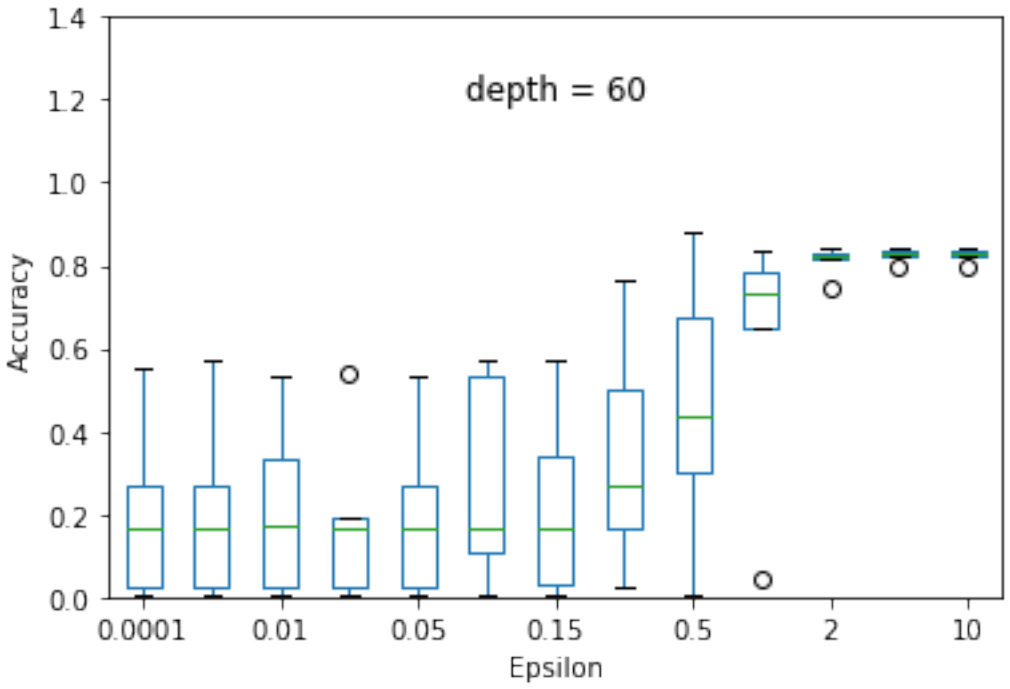
Differential Privacy

- Typical way of implementing: keep *inputs* unperturbed, add noise to *outputs*.
- For any query, add enough random noise that contribution of any one row can't be ascertained



Differential Privacy

- Work by Neelam Memon and Justin Foong: differentially private calculation of classifiers trained on federated thousand genomes data
- Can you perform complex analyses while keeping data private? (Yes!)
- Built with counting queries.
- Informing:
    - How we'll build our differential privacy layer
    - What should go in user queries and what should go in server/API



Federated Analysis of Data

- First task - demonstrate that we can successfully analyze federated data over API.
- Thousand Genome Project - now-classic (2010-2015) sequencing of 2,504 individuals across the world; public data.
- Attempt to reproduce several important population genetics results using simple queries.

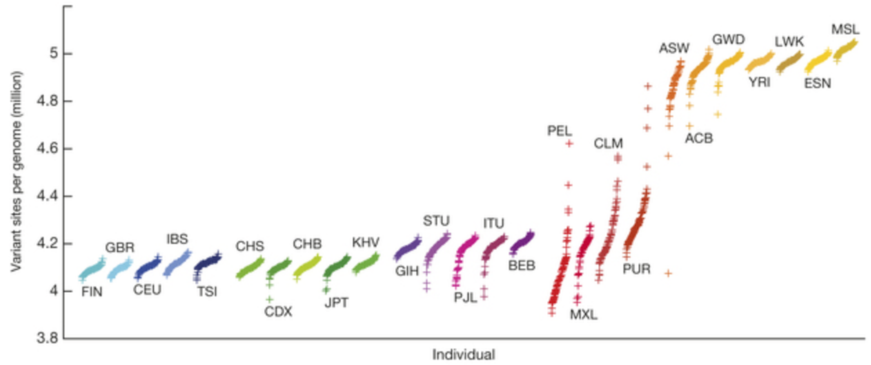


Figure 2: Population structure and demography.

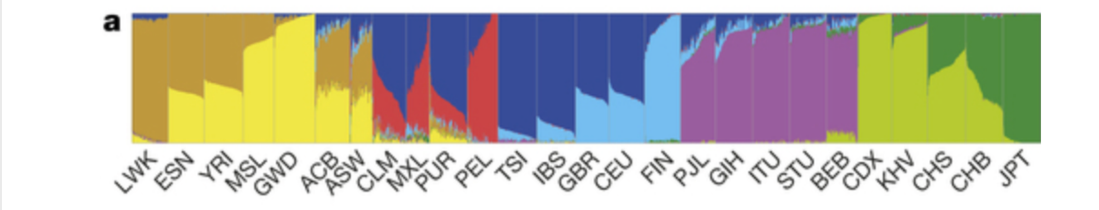
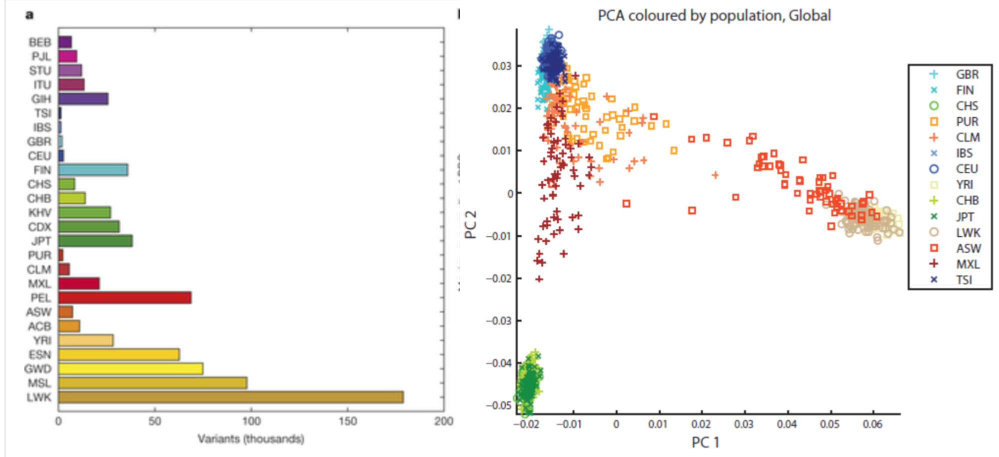


Figure 3: Population differentiation.



Federated Analysis of Data

- Work done by Neelem Memon (BCGSC), Jason Foong (HSC)
- Several of the analyses are straightforward; one is more complicated
- All come down to being able to readily access genotype matrix (does individual *i* have variant *j*)

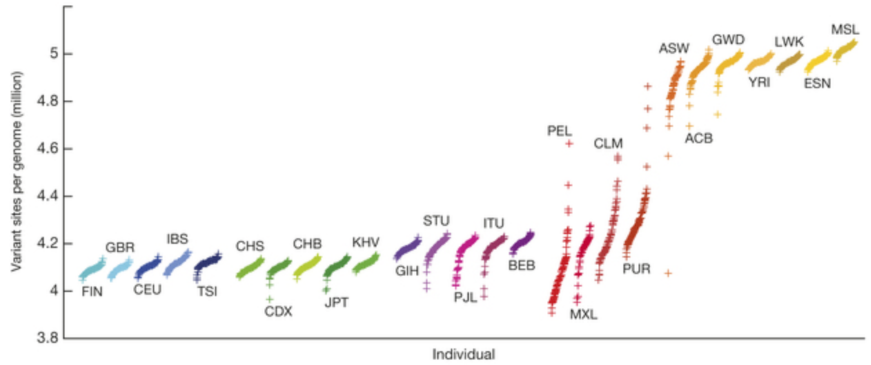


Figure 2: Population structure and demography.

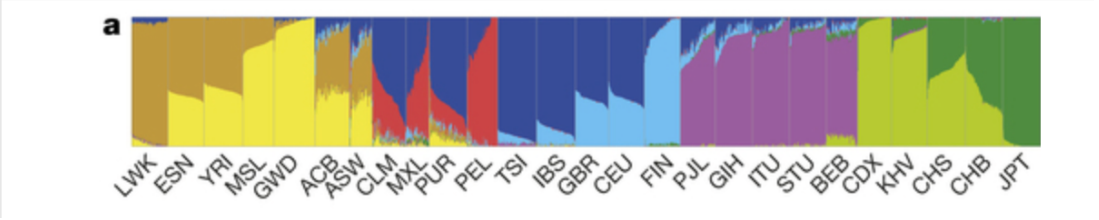
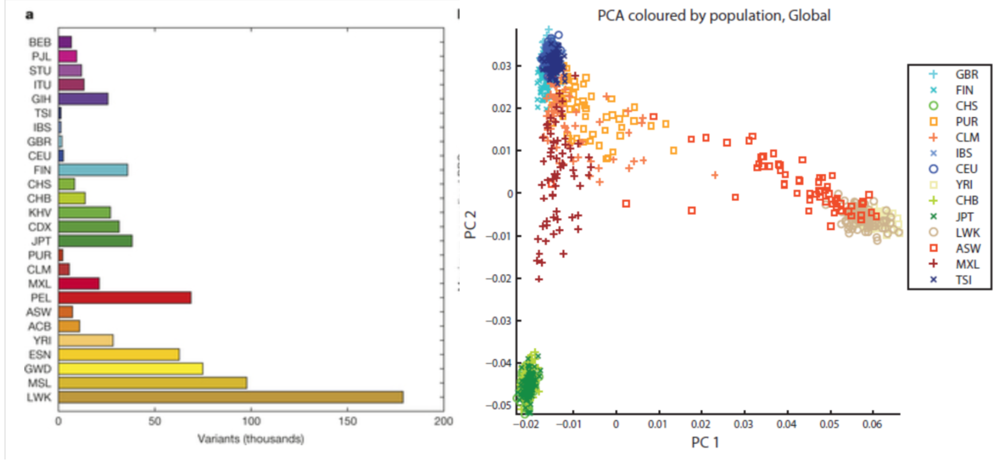


Figure 3: Population differentiation.



CanDIG

National Analysis of Distributed Private Genomic Data

Federated Analysis of Data

- Existing API too slow!
  - Made 4x speed improvements, contributed back to GA4GH
  - Added specific genotype API. (15x)
  - Developed process distributing updated servers across the network
- Analysis now being completed (code already in place)

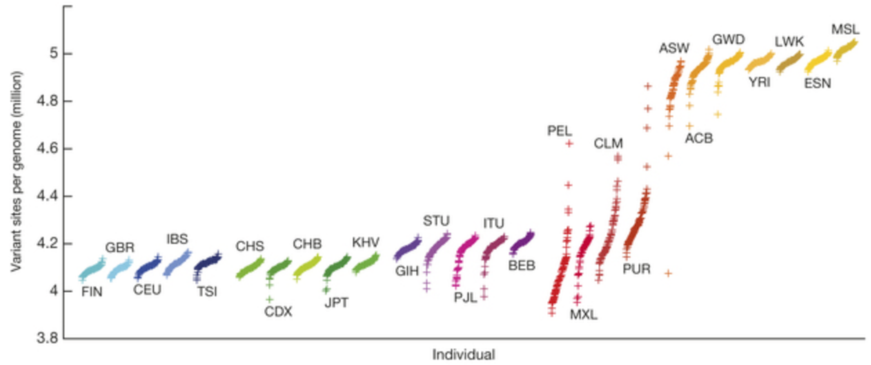


Figure 2: Population structure and demography.

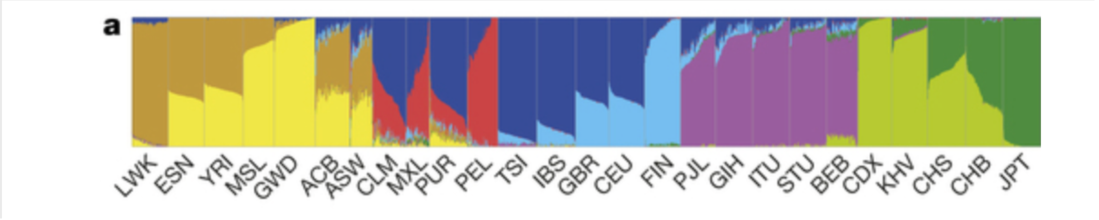
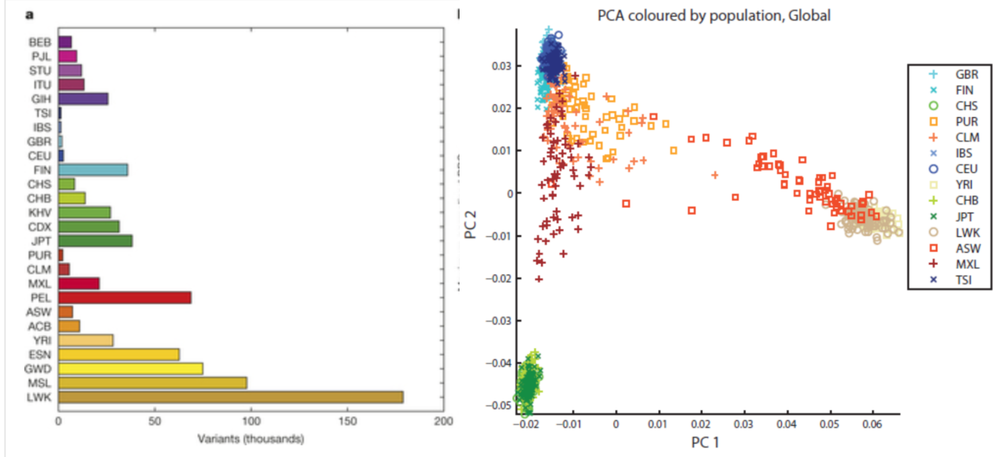
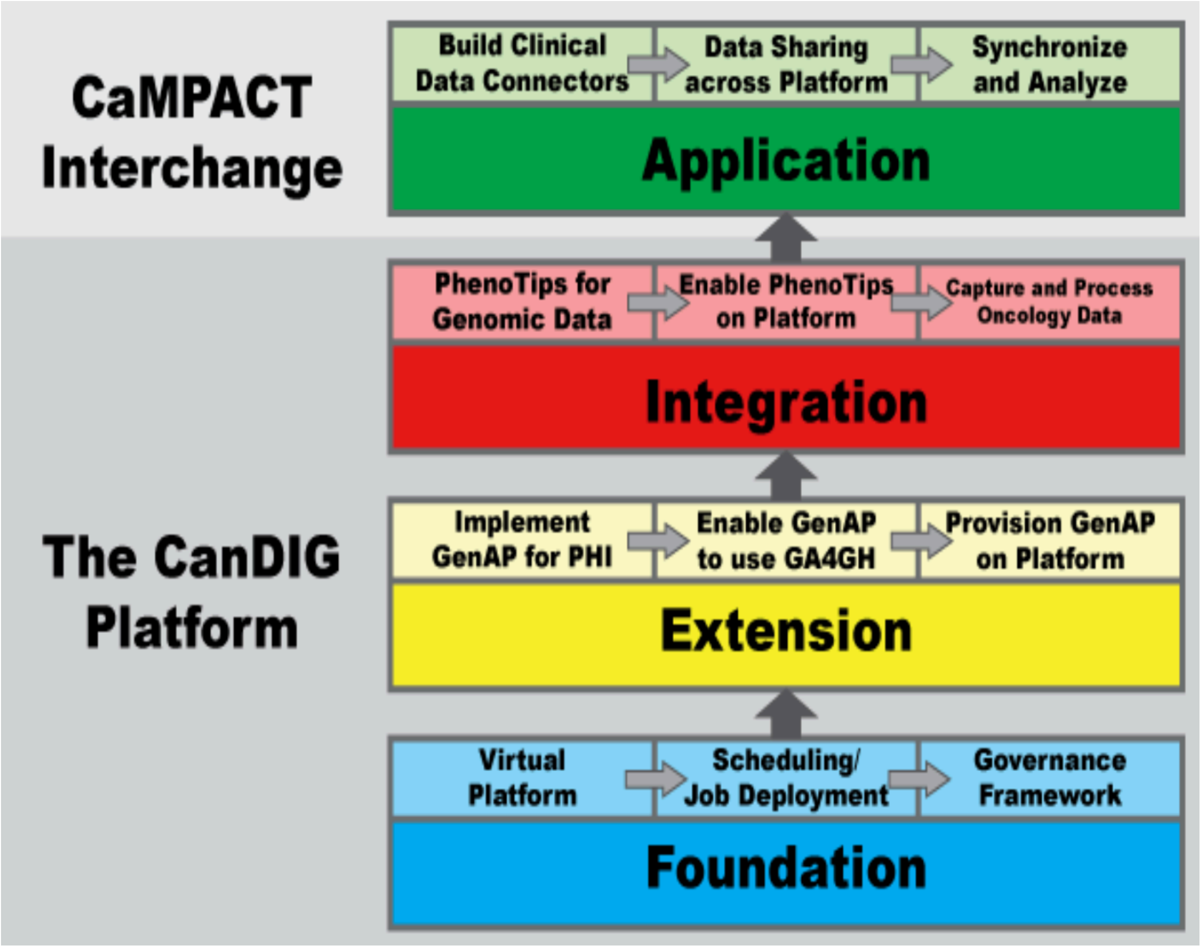


Figure 3: Population differentiation.



Layered Design

- GA4GH (++) layers provide foundational data movement/ access layer



CaMPACT Interchange

Build Clinical Data Connectors → Data Sharing across Platform → Synchronize and Analyze

Application

PhenoTips for Genomic Data → Enable PhenoTips on Platform → Capture and Process Oncology Data

Integration

The CanDIG Platform

Implement GenAP for PHI → Enable GenAP to use GA4GH → Provision GenAP on Platform

Extension

Virtual Platform → Scheduling/ Job Deployment → Governance Framework

Foundation

Foundation TODOs:

- Support the PROFYLE project (Precision Oncology For Young peopLE)
  - Dynamic data directory
  - David Bujold, McGill
- Full authentication:
  - Dustin Hu, Kevin Chan, UHN
  - GENIE data
- Portal:
  - Carol Gauthier, Sherbrooke

TERRY FOX PROFYLE PRECISION ONCOLOGY FOR YOUNG PEOPLE

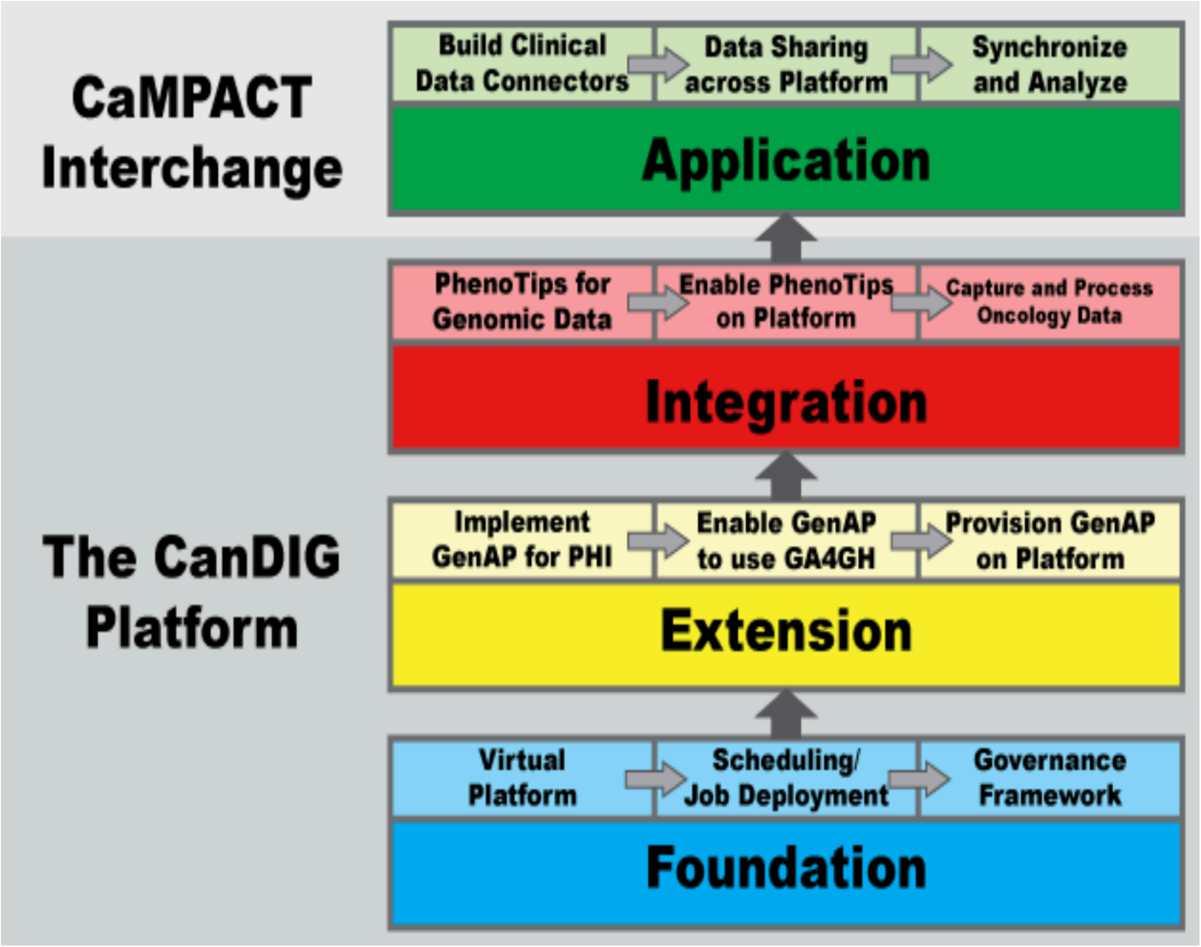
OpenID Connect

AACR American Association for Cancer Research FINDING CURES TOGETHER™

PROJECTGENIE Genomics Evidence Neoplasia Information Exchange

Layered Design

- Then CanDIG-enable existing bioinformatics pipelines
- GA4GH (++) layers provide foundational data movement/ access layer



CaMPACT Interchange

Build Clinical Data Connectors → Data Sharing across Platform → Synchronize and Analyze

Application

The CanDIG Platform

PhenoTips for Genomic Data → Enable PhenoTips on Platform → Capture and Process Oncology Data

Integration

Implement GenAP for PHI → Enable GenAP to use GA4GH → Provision GenAP on Platform

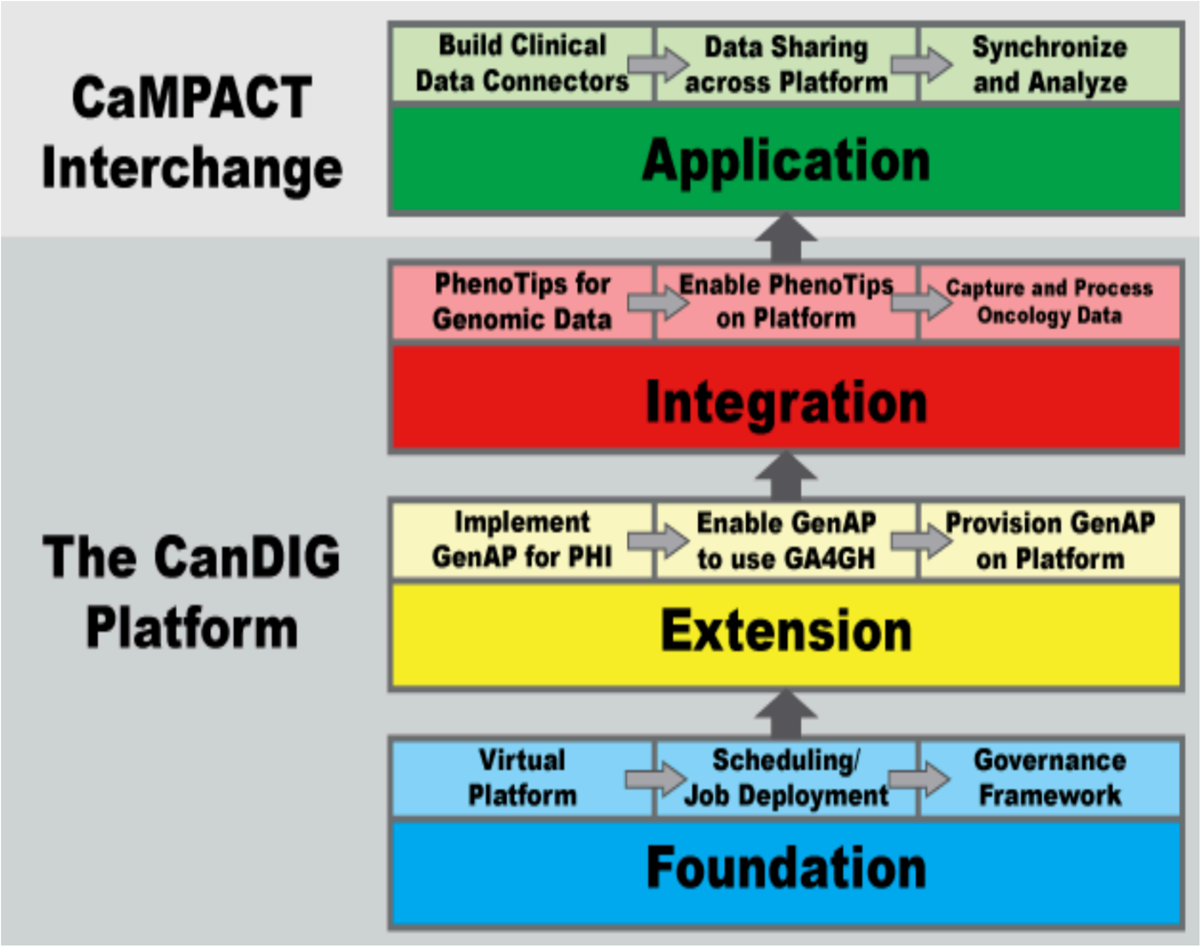
Extension

Virtual Platform → Scheduling/ Job Deployment → Governance Framework

Foundation

Layered Design

- Support PhenoTips (allow integration of phenotypic data)
- Then CanDIG-enable existing bioinformatics pipelines
- GA4GH (++) layers provide foundational data movement/ access layer


CaMPACT Interchange

Build Clinical Data Connectors → Data Sharing across Platform → Synchronize and Analyze

Application

The CanDIG Platform

PhenoTips for Genomic Data → Enable PhenoTips on Platform → Capture and Process Oncology Data

Integration

Implement GenAP for PHI → Enable GenAP to use GA4GH → Provision GenAP on Platform

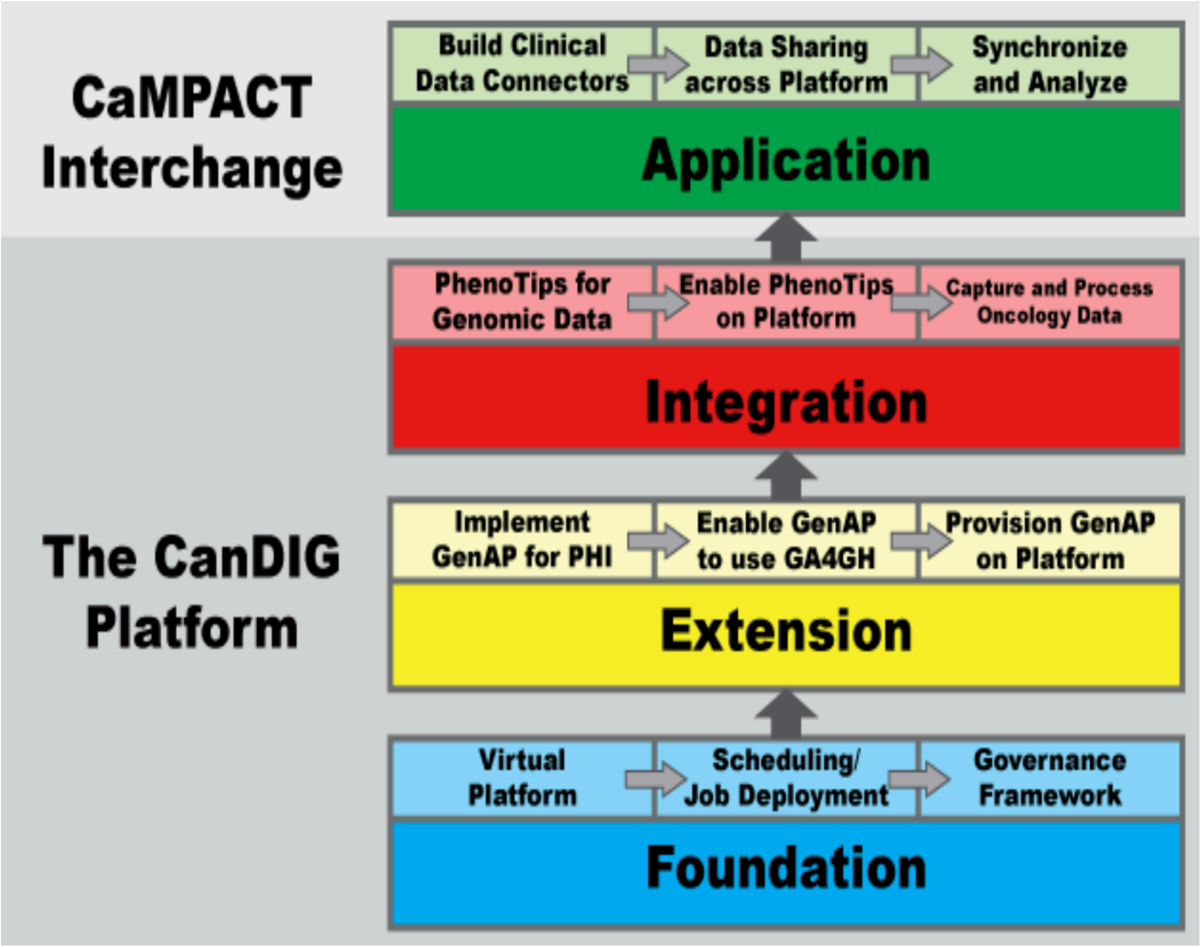
Extension

Virtual Platform → Scheduling/ Job Deployment → Governance Framework

Foundation

Layered Design

- Enable clinical studies atop the platform
- Support PhenoTips (allow integration of phenotypic data)
- Then CanDIG-enable existing bioinformatics pipelines
- GA4GH (++) layers provide foundational data movement/ access layer



CaMPACT Interchange

Build Clinical Data Connectors → Data Sharing across Platform → Synchronize and Analyze

Application

The CanDIG Platform

PhenoTips for Genomic Data → Enable PhenoTips on Platform → Capture and Process Oncology Data

Integration

Implement GenAP for PHI → Enable GenAP to use GA4GH → Provision GenAP on Platform

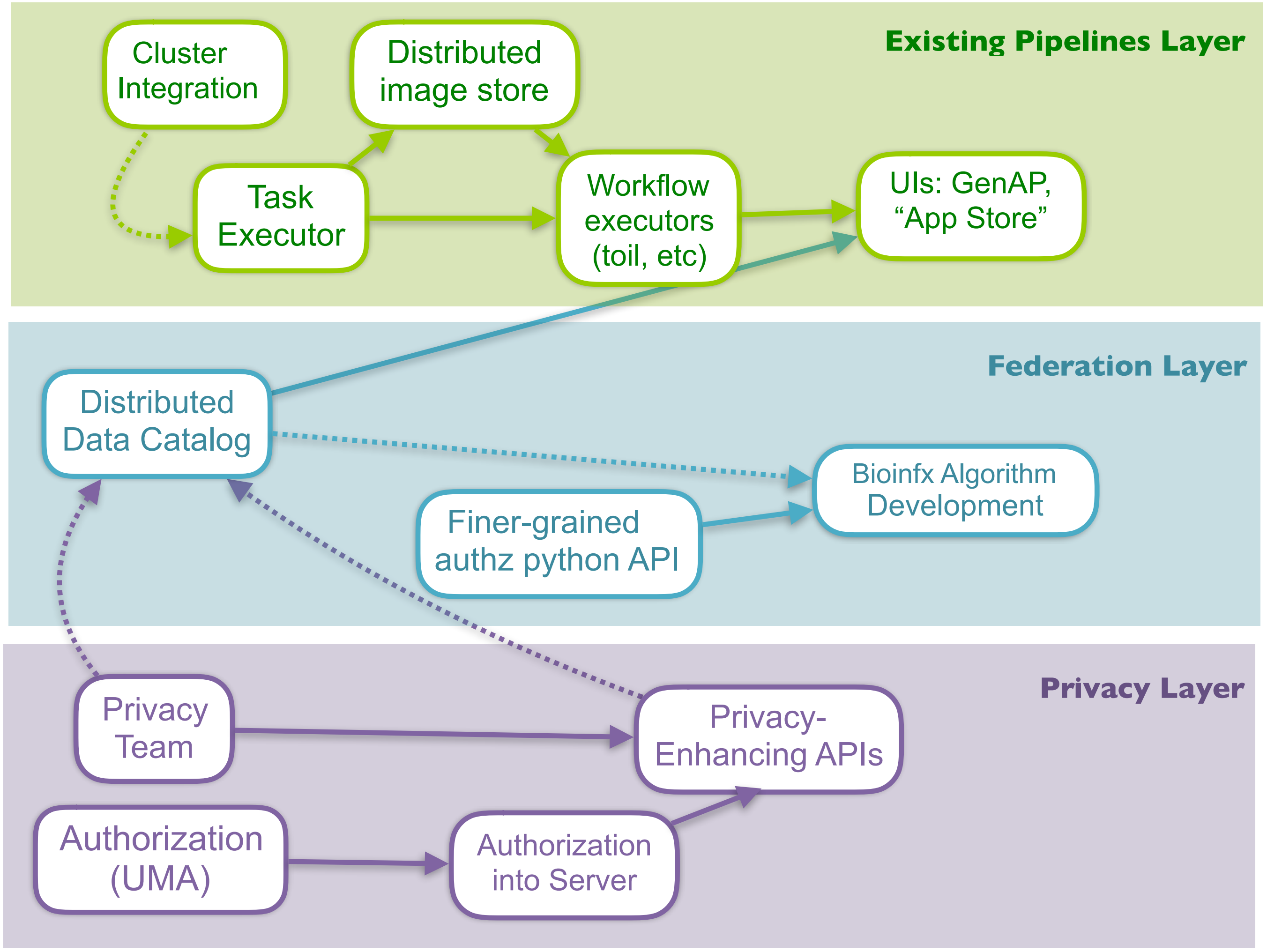
Extension

Virtual Platform → Scheduling/ Job Deployment → Governance Framework

Foundation

Future Plans

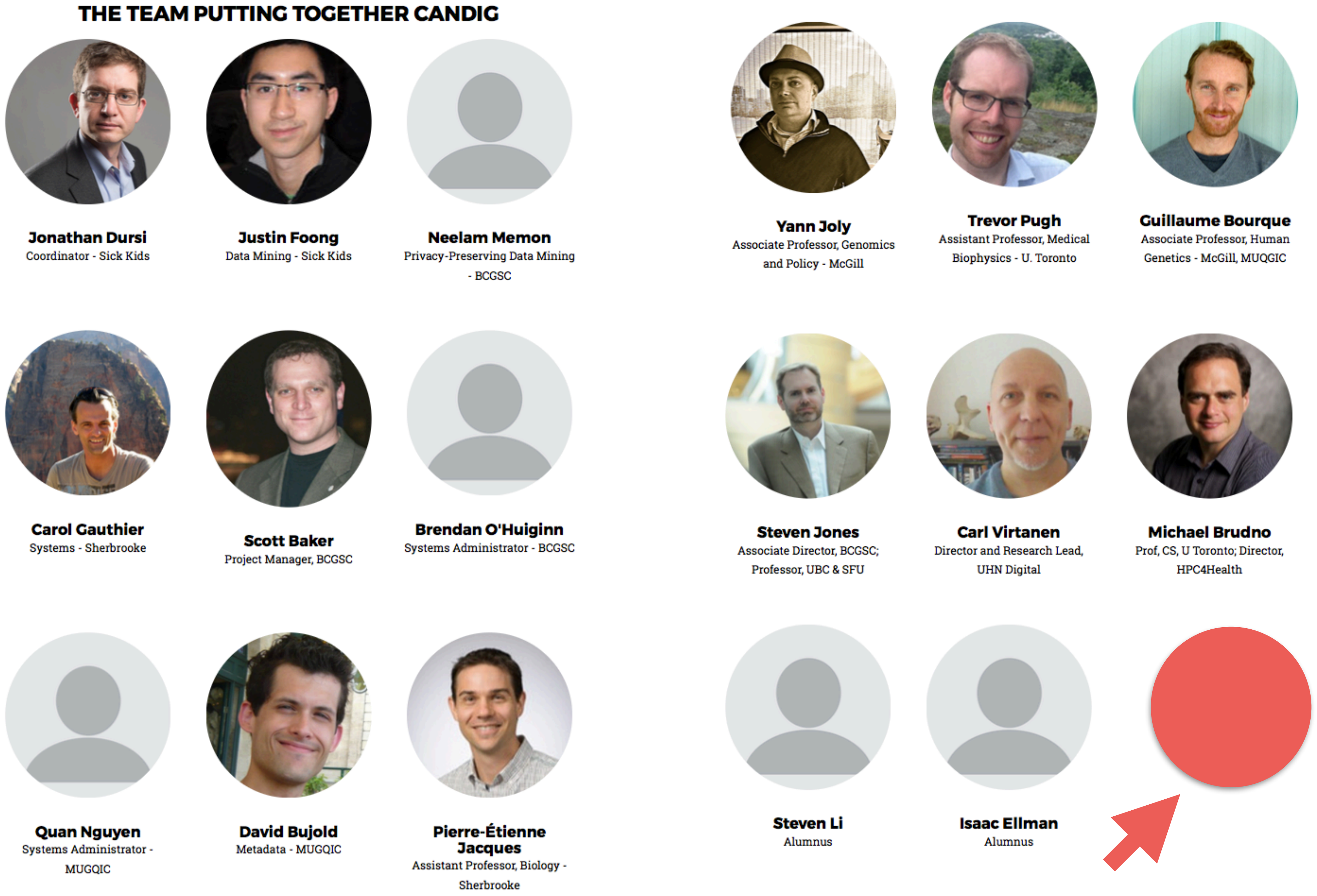
- Lots of cool, important, hard problems ahead:
  - Bioinformatics algorithms (joint calling?) on distributed data
  - Federated Authorization (UMA)
  - Orchestrating workflows across independent sites



CanDIG

National Analysis of Distributed Private Genomic Data

Come Work With Us!



THE TEAM PUTTING TOGETHER CANDIG

Jonathan Dursi
Coordinator - Sick Kids

Justin Foong
Data Mining - Sick Kids

Neelam Memon
Privacy-Preserving Data Mining - BCGSC

Carol Gauthier
Systems - Sherbrooke

Scott Baker
Project Manager, BCGSC

Brendan O'Huiginn
Systems Administrator - BCGSC

Quan Nguyen
Systems Administrator - MUGQIC

David Bujold
Metadata - MUGQIC

Pierre-Étienne Jacques
Assistant Professor, Biology - Sherbrooke

Yann Joly
Associate Professor, Genomics and Policy - McGill

Trevor Pugh
Assistant Professor, Medical Biophysics - U. Toronto

Guillaume Bourque
Associate Professor, Human Genetics - McGill, MUQGIC

Steven Jones
Associate Director, BCGSC; Professor, UBC & SFU

Carl Virtanen
Director and Research Lead, UHN Digital

Michael Brudno
Prof, CS, U Toronto; Director, HPC4Health

Steven Li
Alumnus

Isaac Ellman
Alumnus

You (could be) here

CanDIG

National Analysis of Distributed Private Genomic Data

Come Work With Us!

- **CanDIG**: CanDIG
  - Keep an eye on https://CanDIG.github.io
- **C3G:** Canadian Centre for Computational Genomics
  - Jr/Sr Bioinformatician
  - Postdoc/RA
  - https://ccm.sickkids.ca; ccm.admin@sickkids.ca