CanDIG: National-scale analysis of private, locally-controlled data

CanDIG

National Analysis of Distributed Private Genomic Data

Executive Summary

- **CanDIG**: new 4-year funded Canadian project to enable analysis on genomics data for national cohorts while limiting need for data sharing (slides 4-12)
  - Will lean heavily on, extend GA4GH APIs and related work
  - Will support paediatric cancer projects like PROFYLE, basket-type clinical trial projects like CaMPACT
- **Currently** performing federated data analysis & privacy-preserving data mining of variants data with (extended) reads/variants API, testing out task executions with Funnel (slides 14-26)
  - Federated 1000 genomes re-analysis
  - Building classifiers with differential privacy
- **Proposal**: Use support for PROFYLE as a driver project in year 1, with natural extension possible to CaMPACT in year 2: (slides 28-38)
  - Y1 -"Productionize" OpenID Connect authentication, add authorization in Reads/Variants server (work with Large-Scale Genomics)
  - Y1 - Simple federated reads analysis like joint variant calling (Large-Scale Genomics)
  - Y1 - Tighter interoperability of Reads/Variants (Large-Scale Genomics) and Task/Workflow Execution Servers (Cloud)
  - Y1,2 - Leverage existing federated authentication work (Beacon Network, Access & Authentication)
  - Y2 - Authorization (Access & Authentication)
  - Y2 - Clinical data system integration (Clinical & Phenotypic)

CanDIG Overview

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
- http://CanDIG.github.io

CanDIG Founding Partners

SickKids®
RESEARCH
INSTITUTE

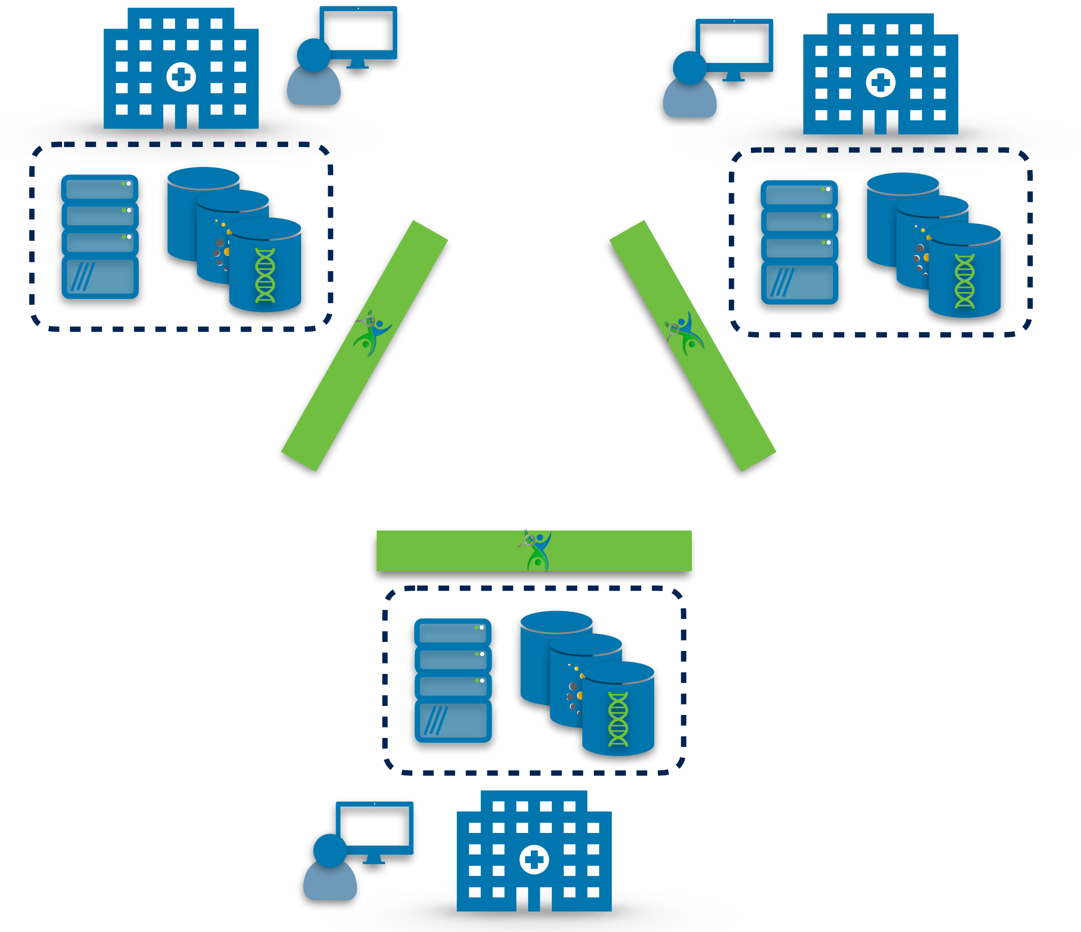
PRINCESS
MARGARET
GENOMICS
CENTRE

McGill University and Génome Québec
Innovation Centre
Innovation in action

CANADA'S MICHAEL SMITH
GENOME
SCIENCES
CENTRE

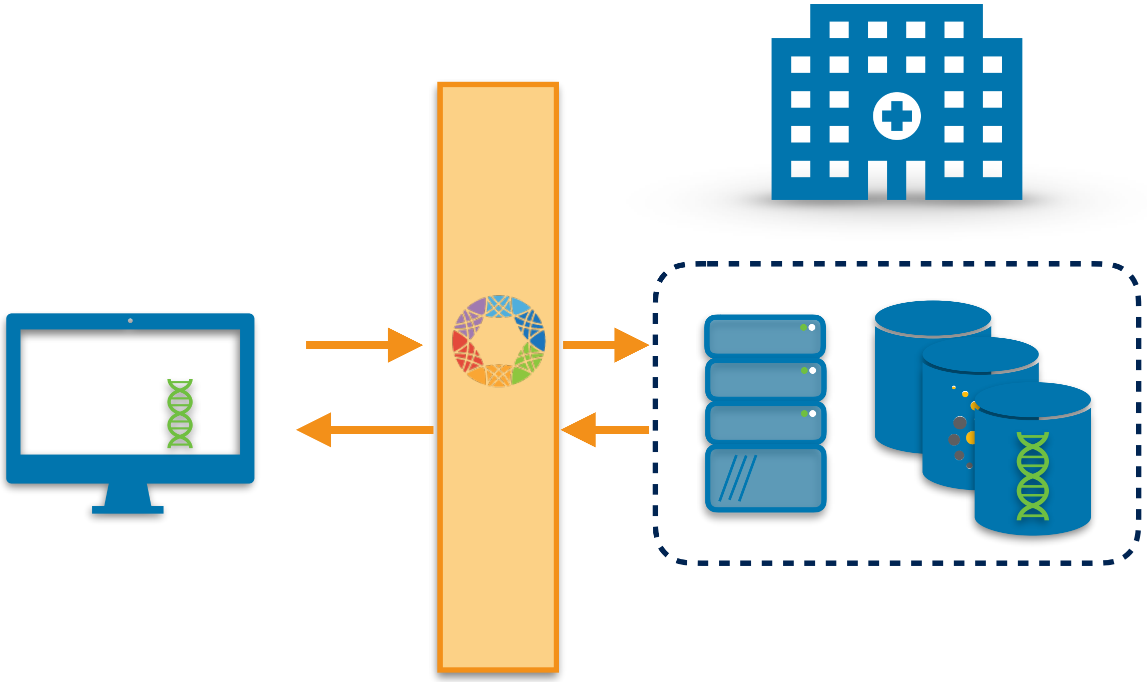
Platform Design: Overall Picture

- Fully distributed
- Participating sites: data providers, source of user requests
- Distributed synchronization of metadata, apps available, etc
- Access to data through API requests, either for data as it stands or for processing through some pipelines
- Local sites control access to their data
- Sites authenticate their users



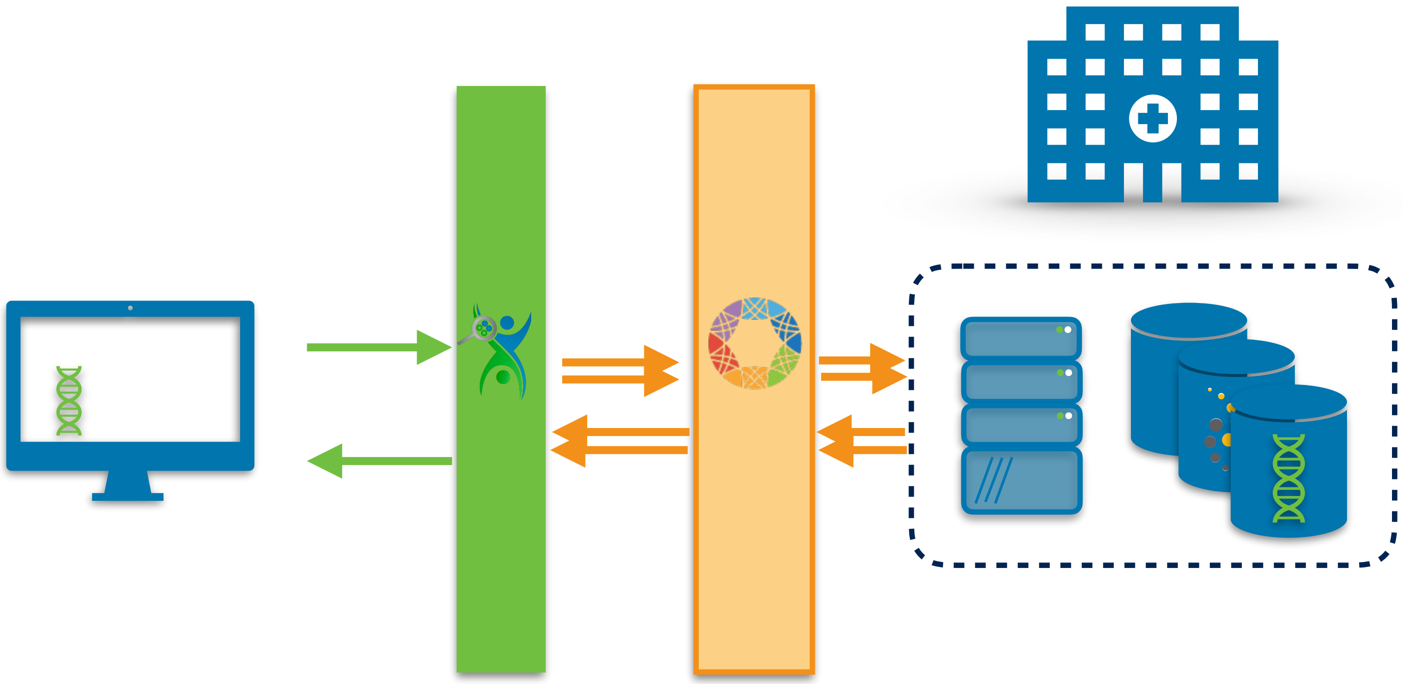
Platform Design: API Data Access

- All access to data will always be through a GA4GH API.
- Allows abstraction of underlying data store (obj store, variant DB), auditing, fine-grained permissions to particular data
- GA4GH reads/variants (etc) API
  + GA4GH task executor service
  + Beacon-network like federated auth

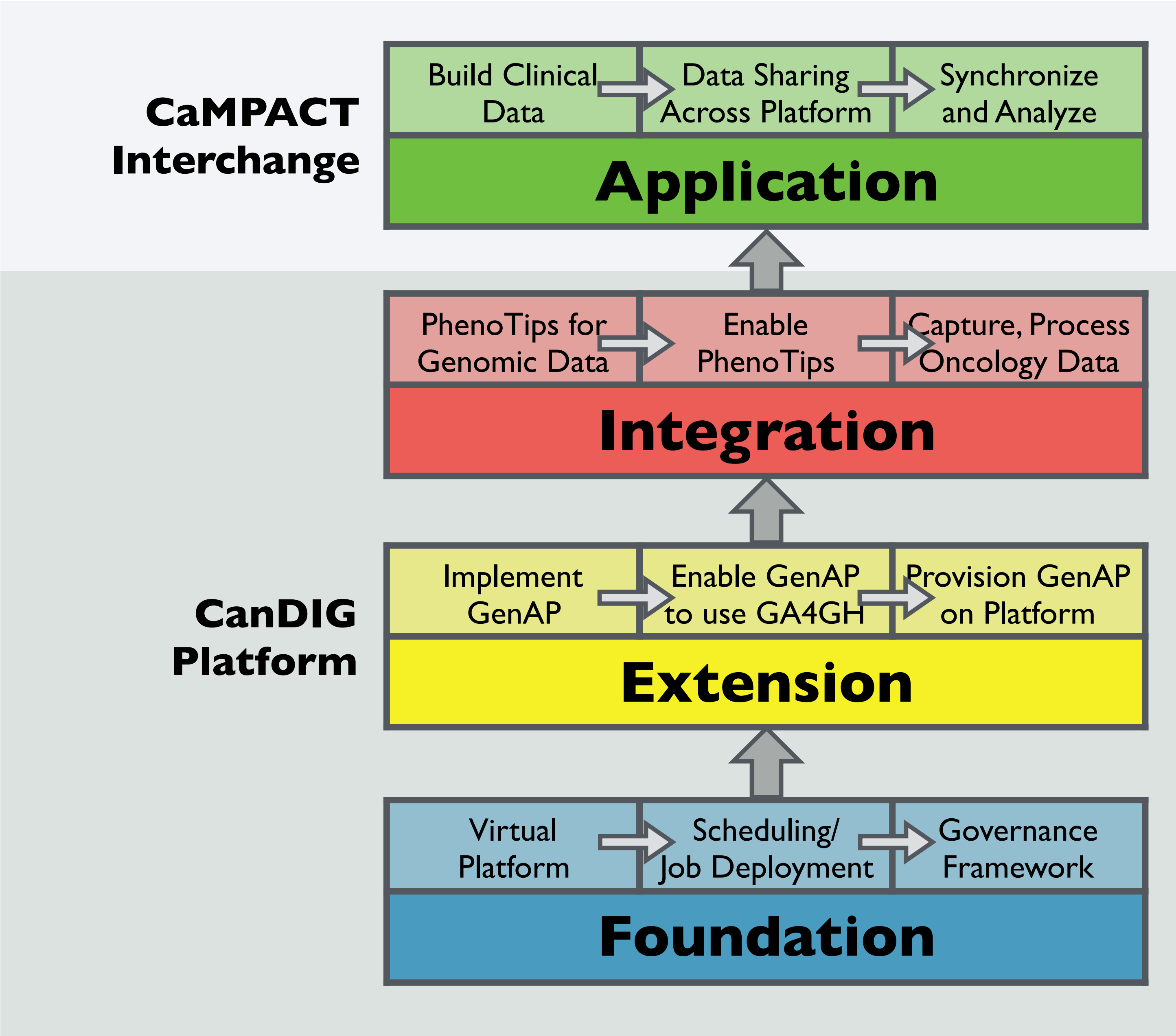


Building higher-level queries

- CanDIG layer in front of GA4GH servers to support:
  - Breaking high-level queries into subqueries
  - Returning only enough info to answer high-level queries
  - Filtering (select … where…)
  - Privacy
  - Fine-grained authorization
  - Particular federation needs
- Extensions of broad interest will be proposed to GA4GH

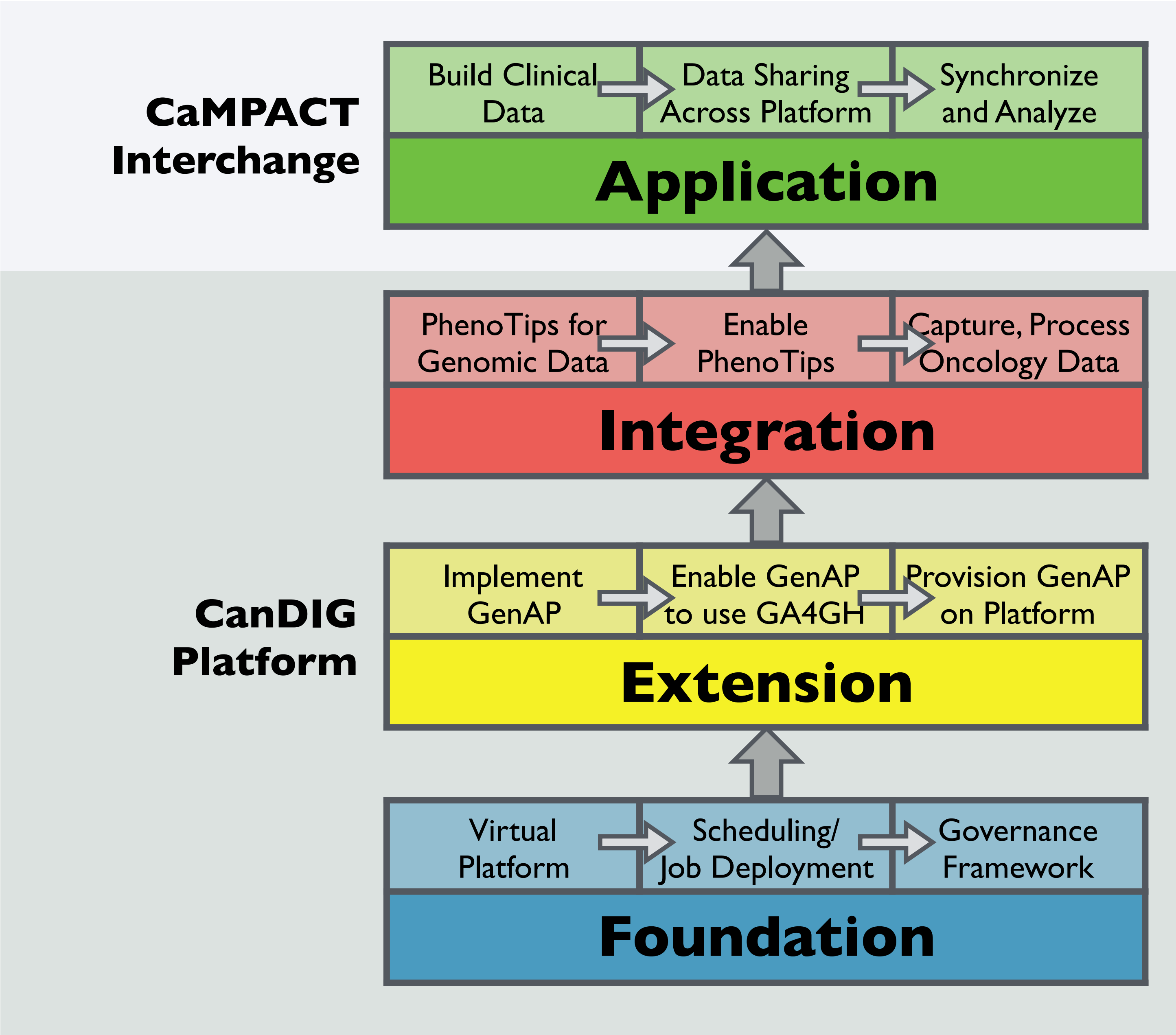


Platform Design: Data access through API



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

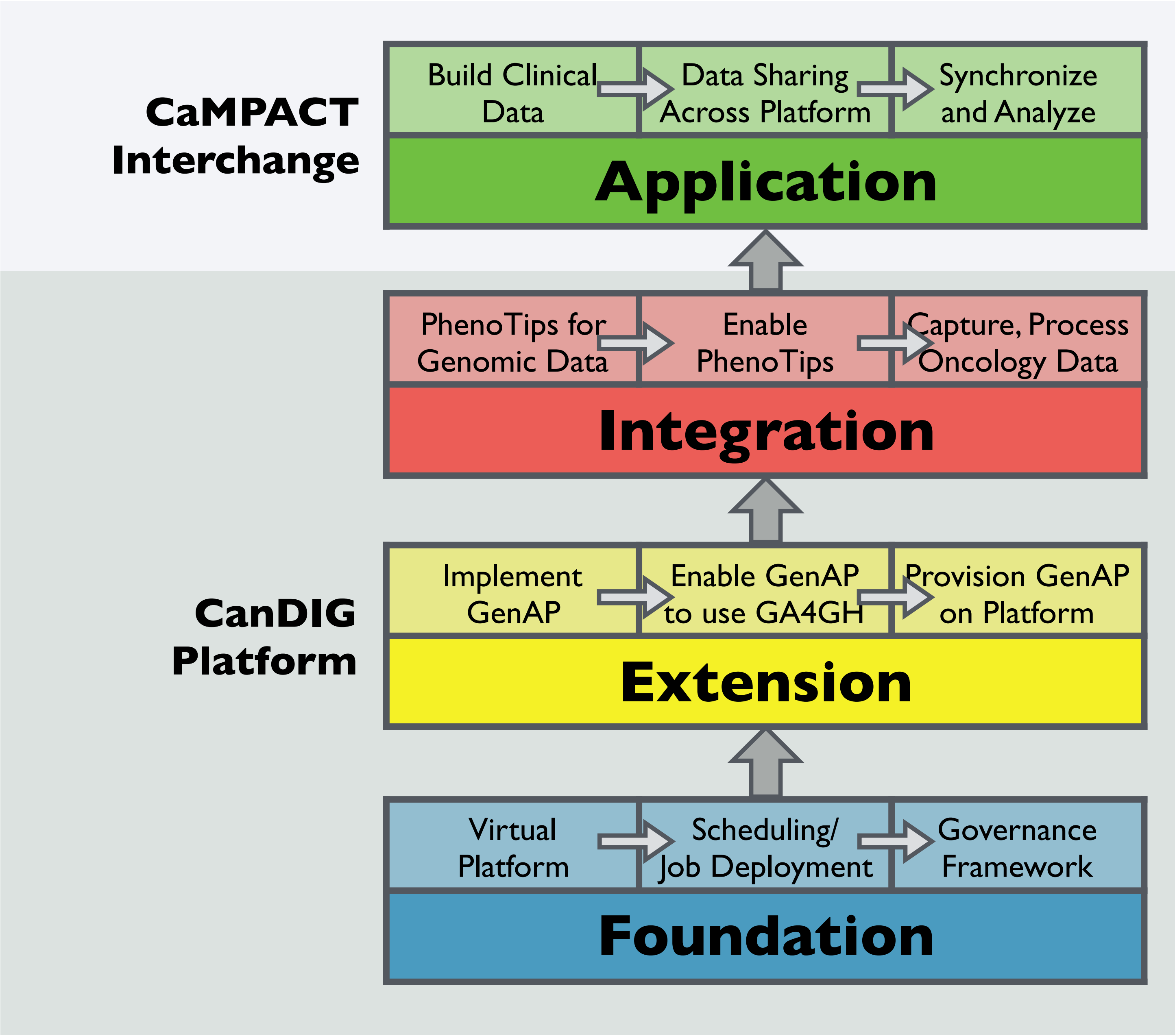
Platform Design: Data access through API



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

Platform Design: Data access through API

- Support PhenoTips, including phenotype info via EPIC/FHIR
- Then GA4GH-WES enable existing bioinformatics pipelines
- GA4GH (++) layers provide foundational data movement/access layer

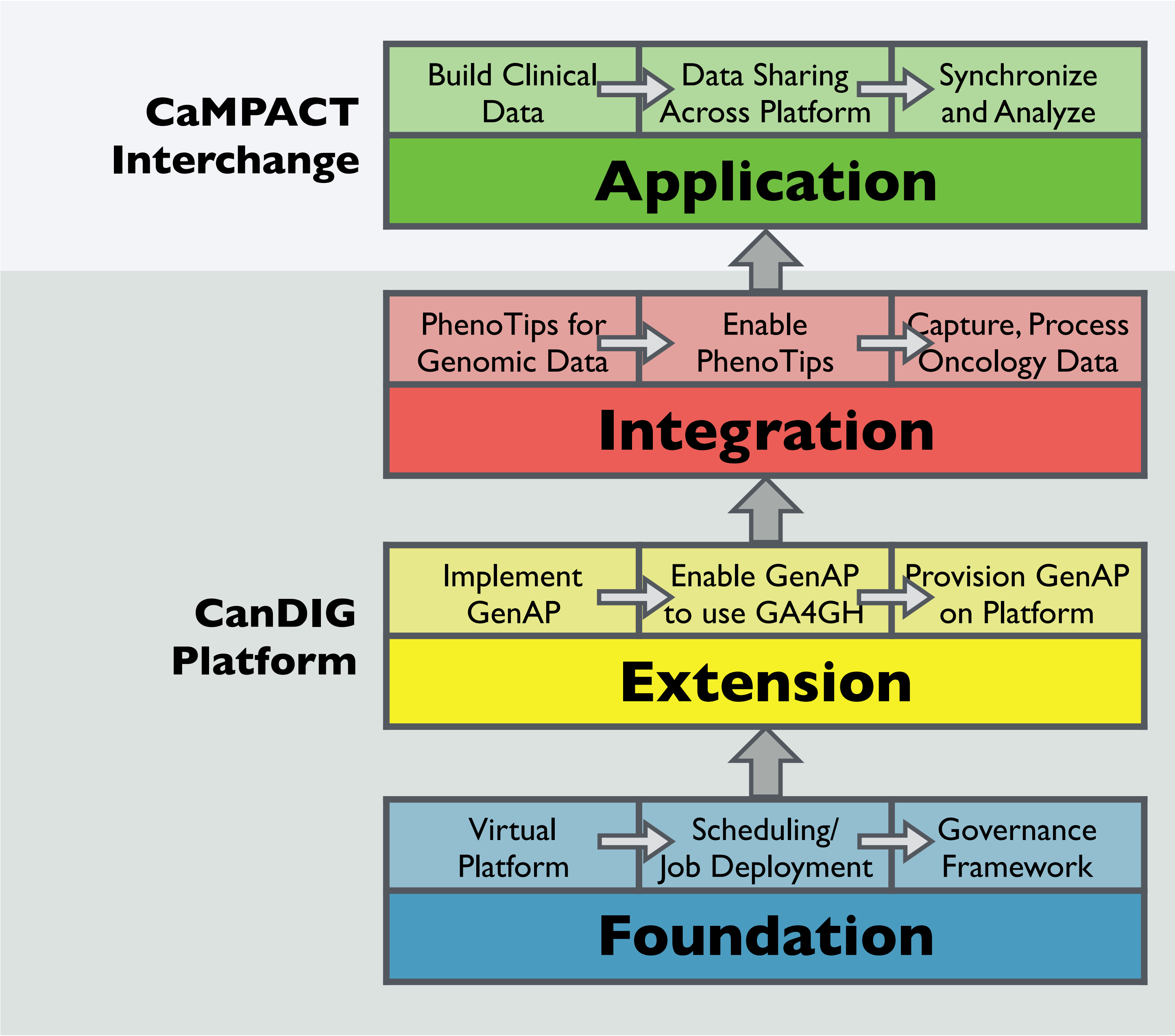


CanDIG

National Analysis of Distributed Private Genomic Data

Platform Design: Data access through API

- Enable clinical studies atop the platform
- Support PhenoTips, including phenotype info via EPIC/FHIR
- Then GA4GH-WES enable existing bioinformatics pipelines
- GA4GH (++) layers provide foundational data movement/access layer



CanDIG Status

Technical Team

UHN:

- **Kevin Chan** - Authentication
- **Duncan Hu** - Authentication
- **Zhibin Lu** - Systems

HSC:

- **Jonathan Dursi** - Coordinator
- **Justin Foong** - Data mining

MUQGIC:

- **David Bujold** - Metadata
- **Carol Gauthier** - GenAP interface
- **Quan Nguyen** - Systems

BSGSC

- **Neelam Memon** - Privacy, Data mining
- **Scott Baker** - BCGSC Project Manager
- **Brendan O'Huiggan** - Systems

Federated Analysis

- Aiming to reproduce 4 classic figures from 1000genomes papers
- Reads and Variants API - public data, no auth
- Data partitioned horizontally (by individual) over 3 sites
- Quickly ran into show-stopping performance problem, identified by Justin Foong: getting large number of calls out of the API

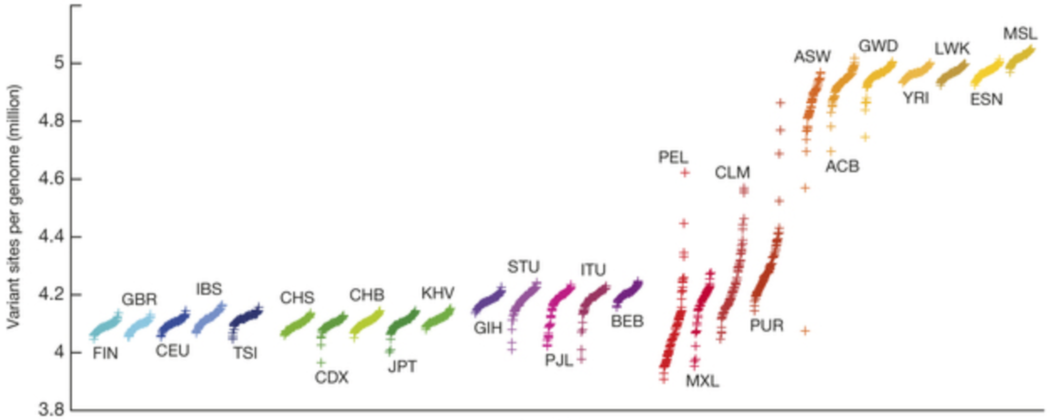


Figure 2: Population structure and demography.

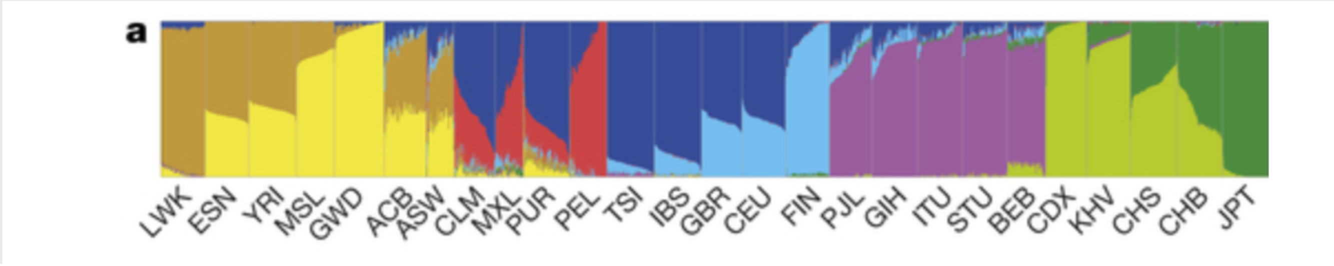
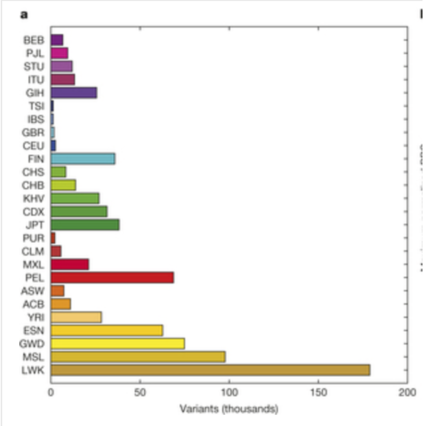
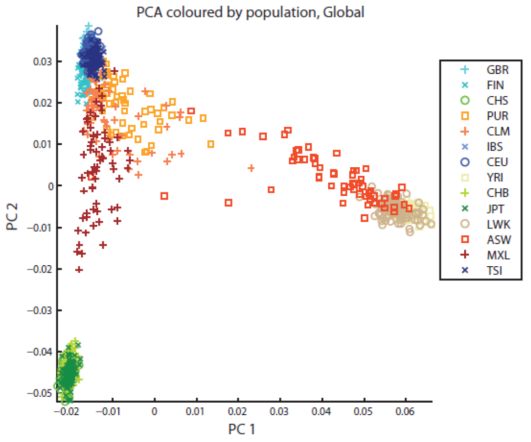


Figure 3: Population differentiation.





Federated Analysis

- Implemented, contributed two 2x performance enhancements to reference server
- Implemented new genotype matrix API for another ~12x speedup
- Can now produce these graphs (but data access still slow for significant portion of genome - imagine downloading VCFs for each calculation)

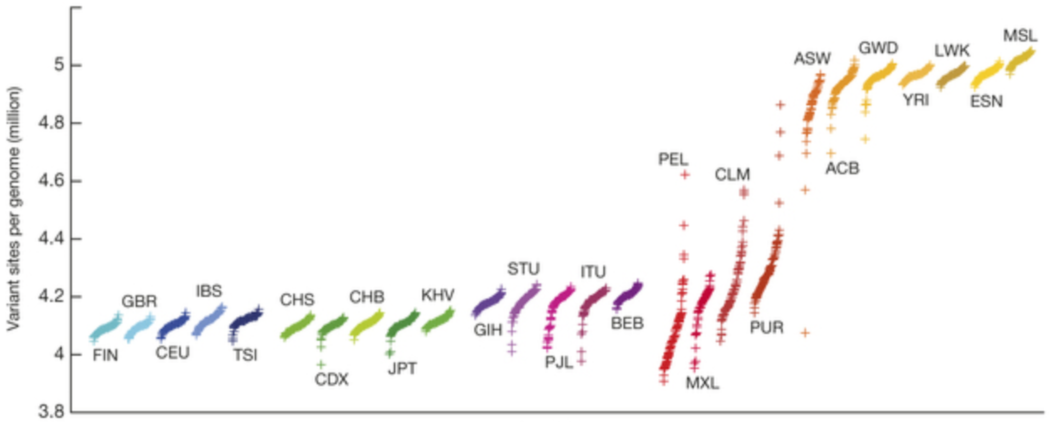


Figure 2: Population structure and demography.

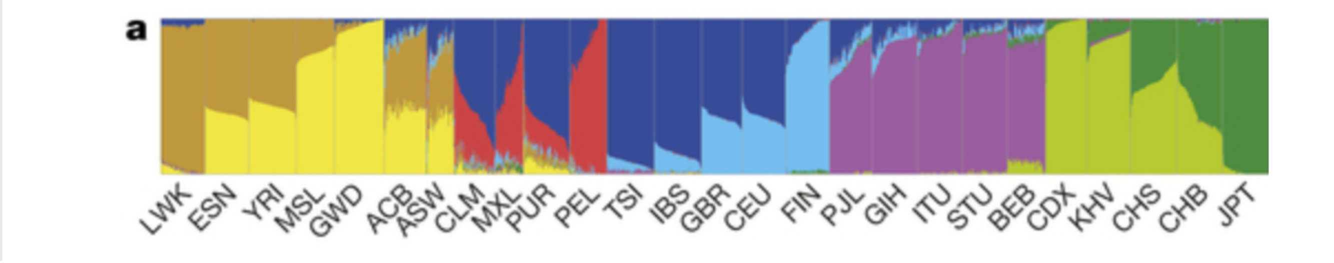
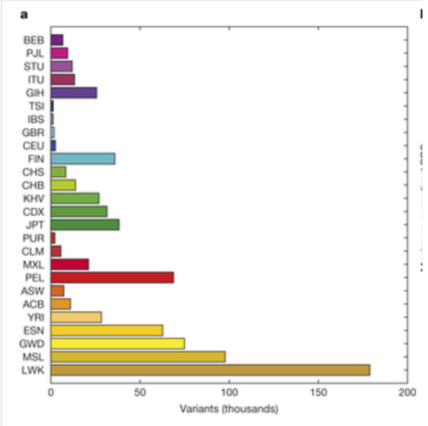
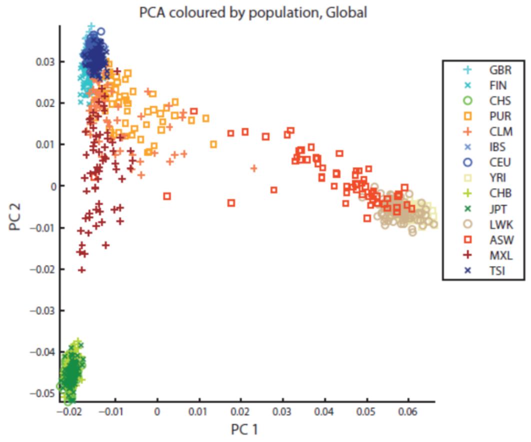


Figure 3: Population differentiation.



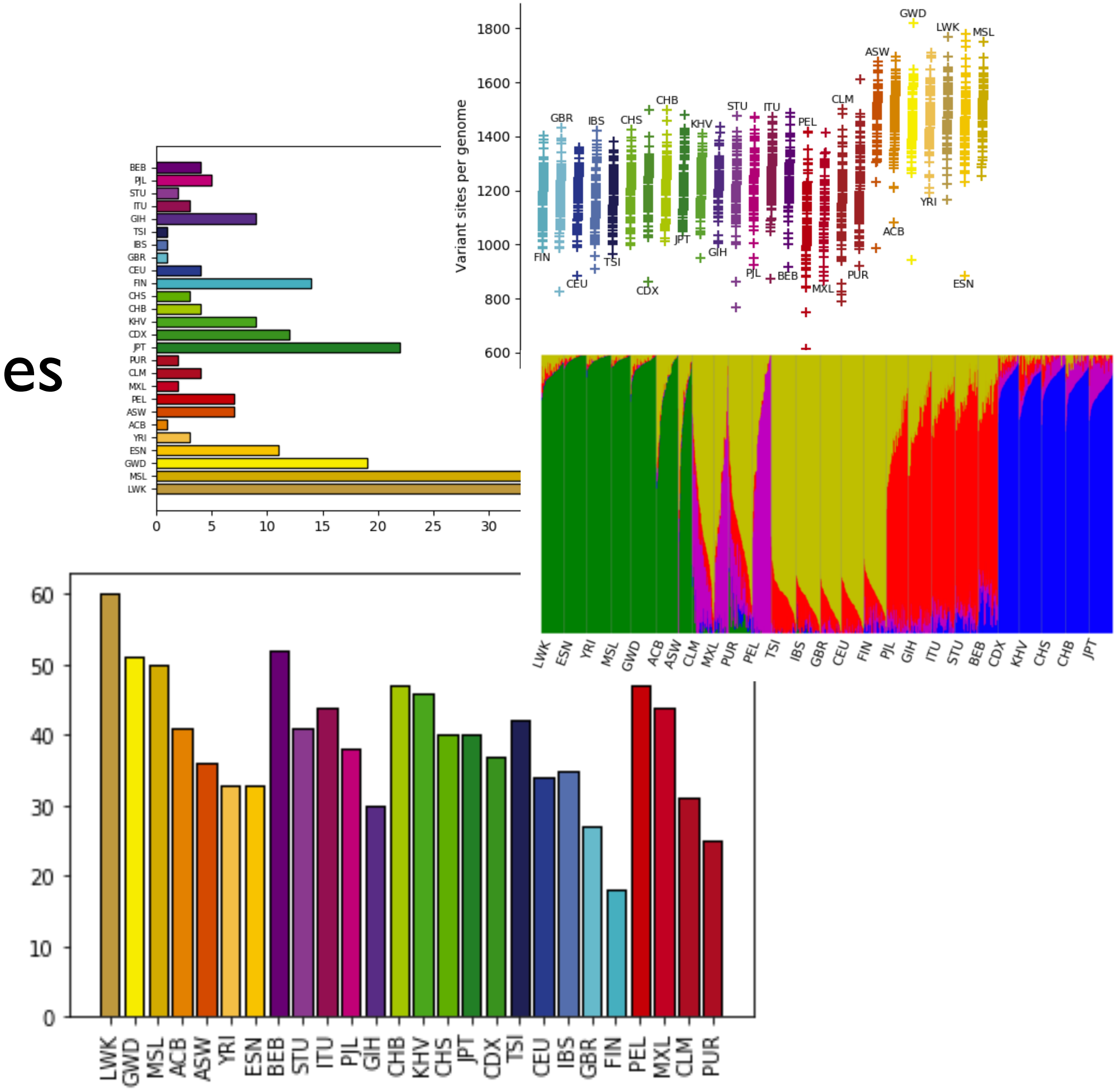


CanDIG

National Analysis of Distributed Private Genomic Data

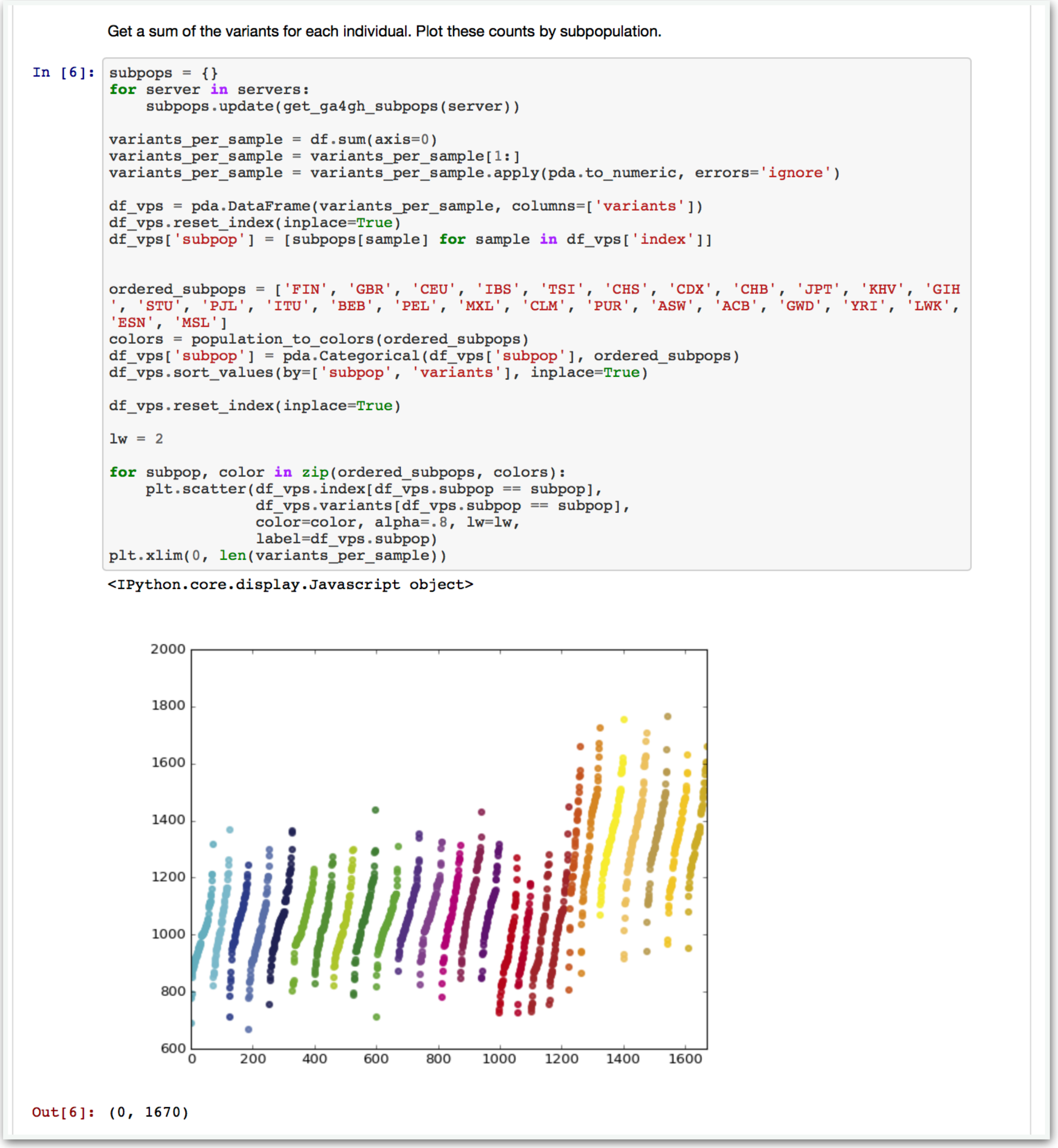
Federated Analysis

- Scripts to re-run analyses on part or whole of genome (Neelam Memon)
- http://github.com/CanDIG/federated-1kg



Federated Analysis

- Based on that work, Jupyter notebooks demonstrating making the figures interactively for ~1Mbp of chr20 (still takes a few minutes to get the data) (Justin Foong)
- http://github.com/CanDIG/federated-1kg



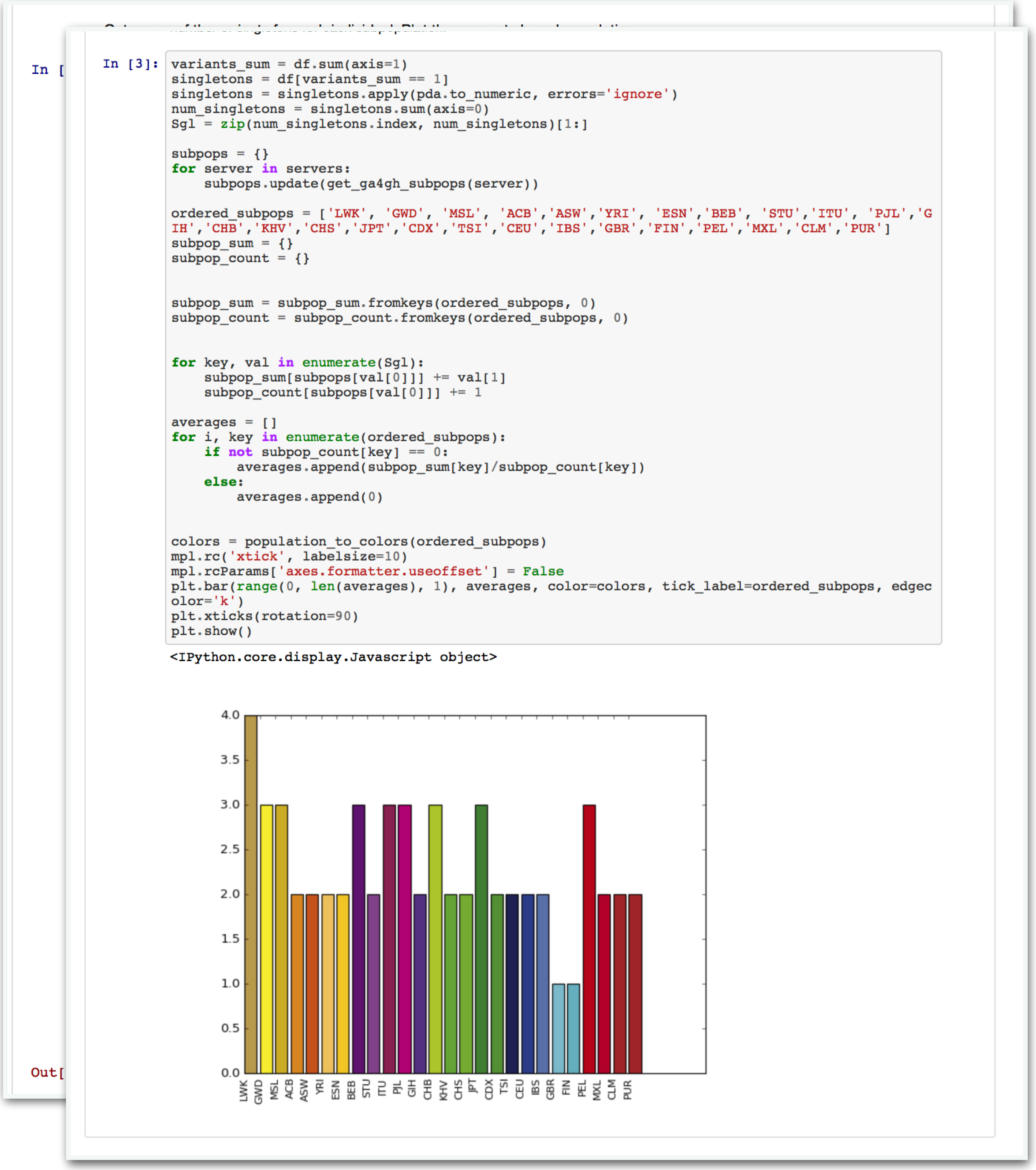
Get a sum of the variants for each individual. Plot these counts by subpopulation.

```
In [6]: subpops = {}
        for server in servers:
            subpops.update(get_ga4gh_subpops(server))

        variants_per_sample = df.sum(axis=0)
        variants_per_sample = variants_per_sample[1:]
        variants_per_sample = variants_per_sample.apply(pda.to_numeric, errors='ignore')

        df_vps = pda.DataFrame(variants_per_sample, columns=['variants'])
        df_vps.reset_index(inplace=True)
        df_vps['subpop'] = [subpops[sample] for sample in df_vps['index']]


        ordered_subpops = ['FIN', 'GBR', 'CEU', 'IBS', 'TSI', 'CHS', 'CDX', 'CHB', 'JPT', 'KHV', 'GIH
        ', 'STU', 'PJL', 'ITU', 'BEB', 'PEL', 'MXL', 'CLM', 'PUR', 'ASW', 'ACB', 'GWD', 'YRI', 'LWK',
        'ESN', 'MSL']
        colors = population_to_colors(ordered_subpops)
        df_vps['subpop'] = pda.Categorical(df_vps['subpop'], ordered_subpops)
        df_vps.sort_values(by=['subpop', 'variants'], inplace=True)

        df_vps.reset_index(inplace=True)

        lw = 2

        for subpop, color in zip(ordered_subpops, colors):
            plt.scatter(df_vps.index[df_vps.subpop == subpop],
                        df_vps.variants[df_vps.subpop == subpop],
                        color=color, alpha=.8, lw=lw,
                        label=df_vps.subpop)
        plt.xlim(0, len(variants_per_sample))

        <IPython.core.display.Javascript object>
```

```
Out[6]: (0, 1670)
```

Federated Analysis

- Based on that work, Jupyter notebooks demonstrating making the figures interactively for ~1Mbp of chr20 (still takes a few minutes to get the data) (Justin Foong)
- http://github.com/CanDIG/federated-1kg



In [

In [3]:

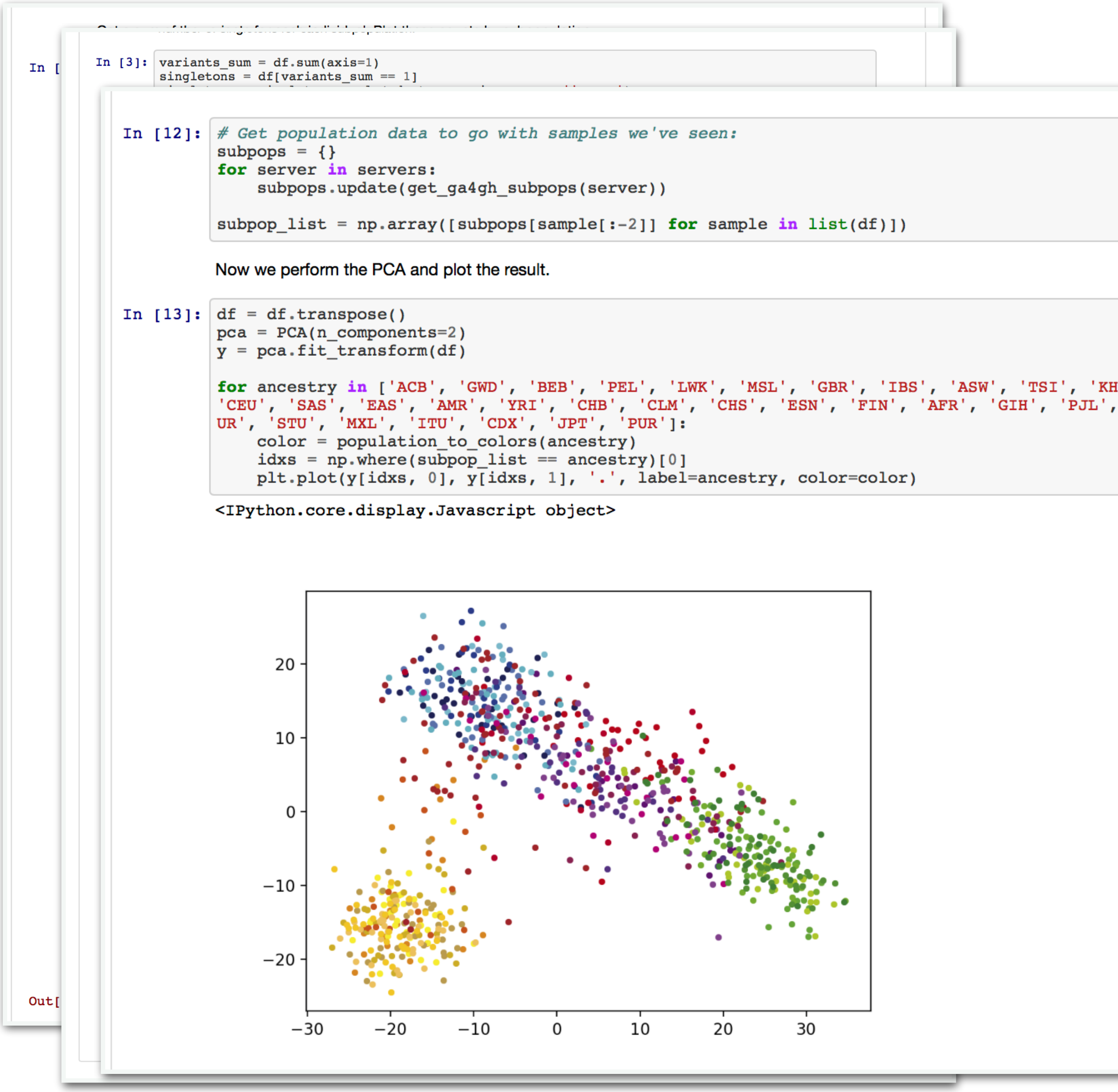
```python
variants_sum = df.sum(axis=1)
singletons = df[variants_sum == 1]
singletons = singletons.apply(pda.to_numeric, errors='ignore')
num_singletons = singletons.sum(axis=0)
Sgl = zip(num_singletons.index, num_singletons)[1:]

subpops = {}
for server in servers:
    subpops.update(get_ga4gh_subpops(server))

ordered_subpops = ['LWK', 'GWD', 'MSL', 'ACB','ASW','YRI', 'ESN','BEB', 'STU','ITU', 'PJL','G
IH','CHB','KHV','CHS','JPT','CDX','TSI','CEU','IBS','GBR','FIN','PEL','MXL','CLM','PUR']
subpop_sum = {}
subpop_count = {}


subpop_sum = subpop_sum.fromkeys(ordered_subpops, 0)
subpop_count = subpop_count.fromkeys(ordered_subpops, 0)


for key, val in enumerate(Sgl):
    subpop_sum[subpops[val[0]]] += val[1]
    subpop_count[subpops[val[0]]] += 1

averages = []
for i, key in enumerate(ordered_subpops):
    if not subpop_count[key] == 0:
        averages.append(subpop_sum[key]/subpop_count[key])
    else:
        averages.append(0)


colors = population_to_colors(ordered_subpops)
mpl.rc('xtick', labelsize=10)
mpl.rcParams['axes.formatter.useoffset'] = False
plt.bar(range(0, len(averages), 1), averages, color=colors, tick_label=ordered_subpops, edgec
olor='k')
plt.xticks(rotation=90)
plt.show()
```

```
<IPython.core.display.Javascript object>
```

Out[

Federated Analysis

- Based on that work, Jupyter notebooks demonstrating making the figures interactively for ~1Mbp of chr20 (still takes a few minutes to get the data) (Justin Foong)
- http://github.com/CanDIG/federated-1kg
- (No ancestral population graph - long calculation and requires a lot of data - not really suitable for interactive demo)



```
In [3]: variants_sum = df.sum(axis=1)
        singletons = df[variants_sum == 1]
```

```
In [12]: # Get population data to go with samples we've seen:
         subpops = {}
         for server in servers:
             subpops.update(get_ga4gh_subpops(server))

         subpop_list = np.array([subpops[sample[:-2]] for sample in list(df)])
```

Now we perform the PCA and plot the result.

```
In [13]: df = df.transpose()
         pca = PCA(n_components=2)
         y = pca.fit_transform(df)

         for ancestry in ['ACB', 'GWD', 'BEB', 'PEL', 'LWK', 'MSL', 'GBR', 'IBS', 'ASW', 'TSI', 'KH
         'CEU', 'SAS', 'EAS', 'AMR', 'YRI', 'CHB', 'CLM', 'CHS', 'ESN', 'FIN', 'AFR', 'GIH', 'PJL',
         UR', 'STU', 'MXL', 'ITU', 'CDX', 'JPT', 'PUR']:
             color = population_to_colors(ancestry)
             idxs = np.where(subpop_list == ancestry)[0]
             plt.plot(y[idxs, 0], y[idxs, 1], '.', label=ancestry, color=color)
```

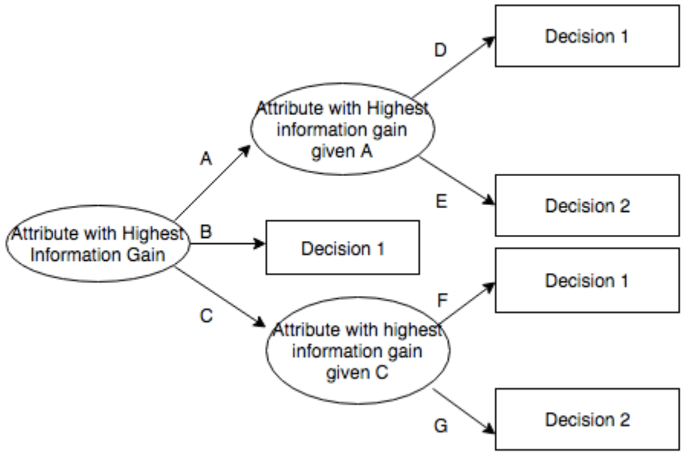
```
<IPython.core.display.Javascript object>
```

CanDIG

National Analysis of Distributed Private Genomic Data

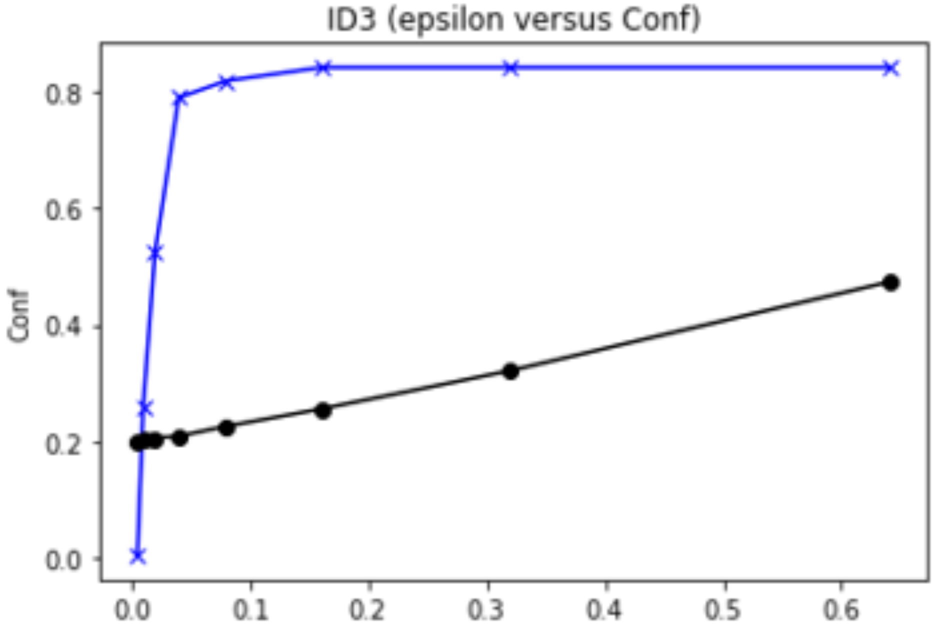
Federated Analysis

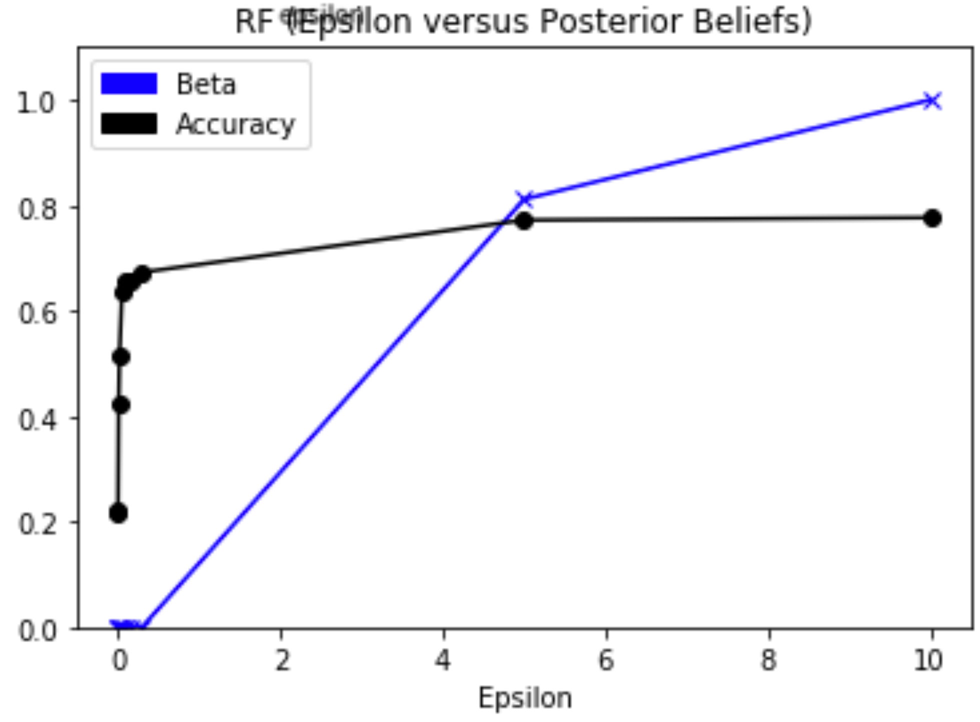
- While waiting for genotype API to be developed, investigation into building classifiers over the variant information
- Small number of variants known to be sufficient for inferring ancestry
- Using two different differentially private tree-based classifiers which can work well for partitioned data: ID3 tree, random forest
- A Practical Differentially Private Random Decision Tree Classifier, Jagannathan, Pillaipakkamnatt, Wright
- Work led by Neelam Memon


Attribute with Highest Information Gain
A
B
C
Attribute with Highest information gain given A
Decision 1
Attribute with highest information gain given C
D
E
F
G
Decision 1
Decision 2
Decision 1
Decision 2

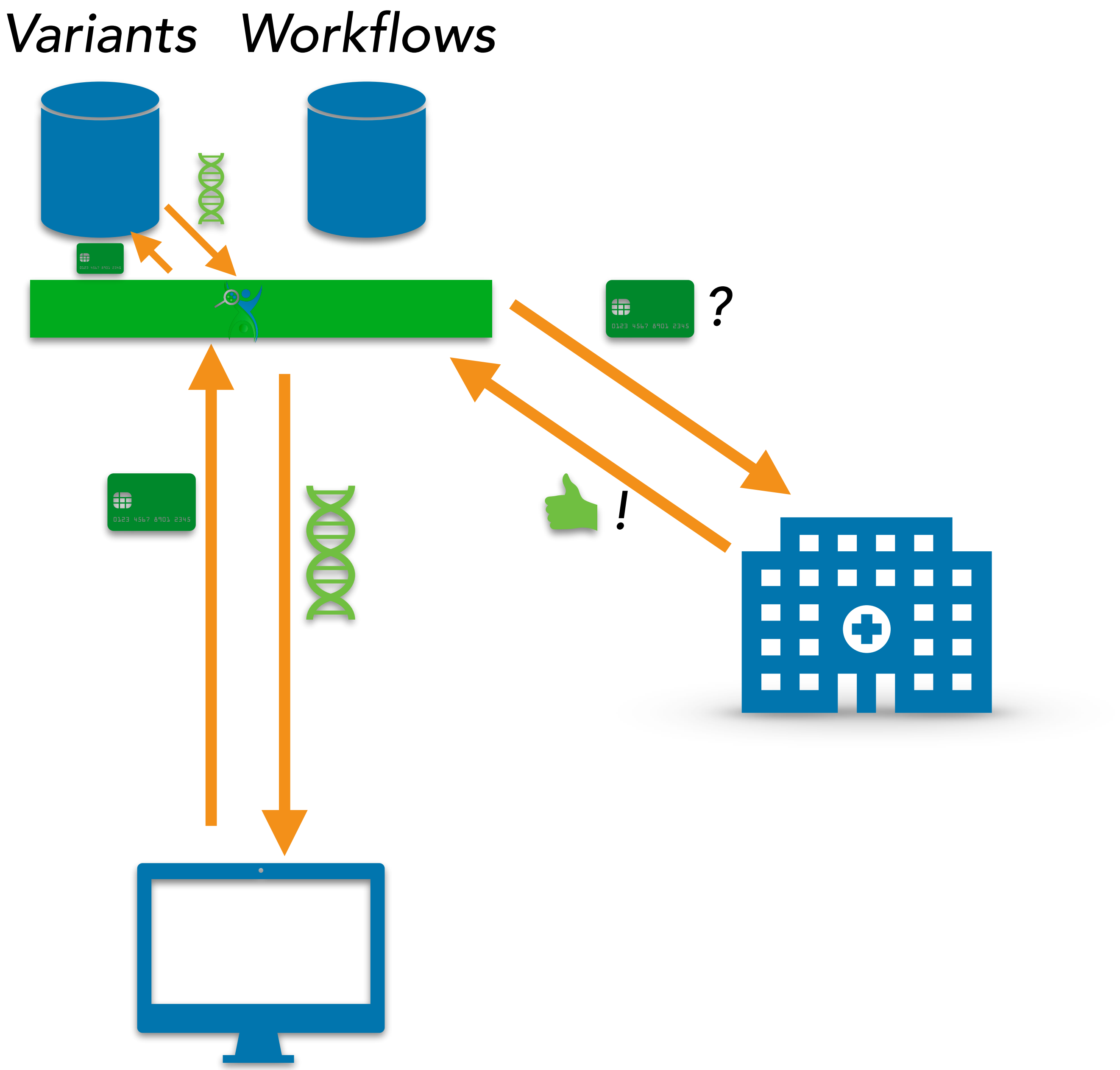
Federated Analysis

- Previous work (wasn’t distributed) claims RF much better accuracy for given “privacy” than ID3; but “privacy” metric was ϵ, the differential privacy parameter
- Differential Privacy: An Economic Method for Choosing Epsilon, Hsu *et al.*: for a given (almost-all-knowing) adversary, find out how much information is actually leaked given queries to build trees
- More complicated…



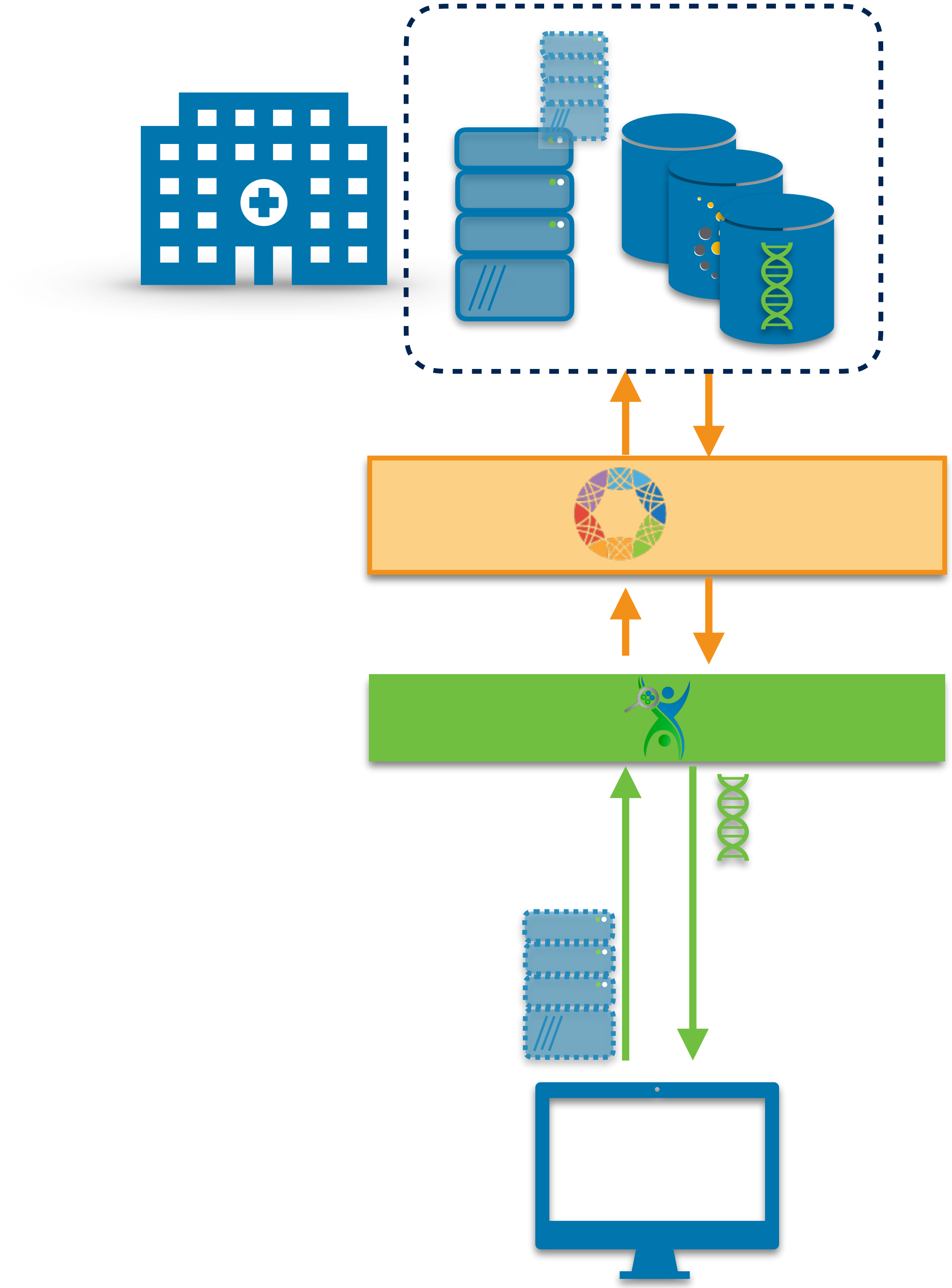


Remote images



- **Authentication**: Open ID Connect (OIDC) approach - Verifiable tokens with identity claims - particularly useful in our case, where CanDIG "server" will be several collaborating services (analogous to microservices)

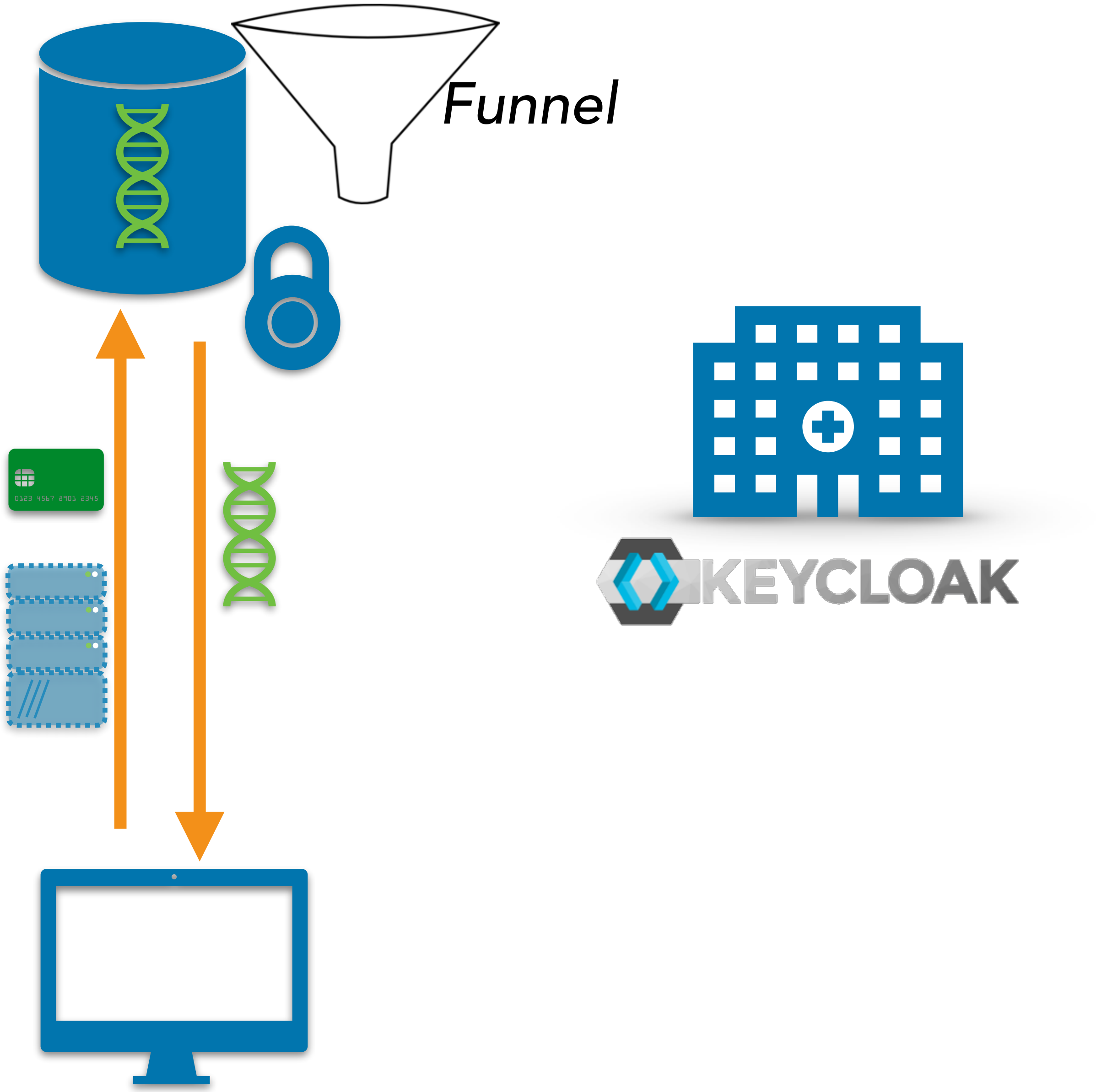
Remote images



- **Bundling of images**: using Docker for now due to tooling, will move to Singularity or rkt depending on which “wins” over next couple of years

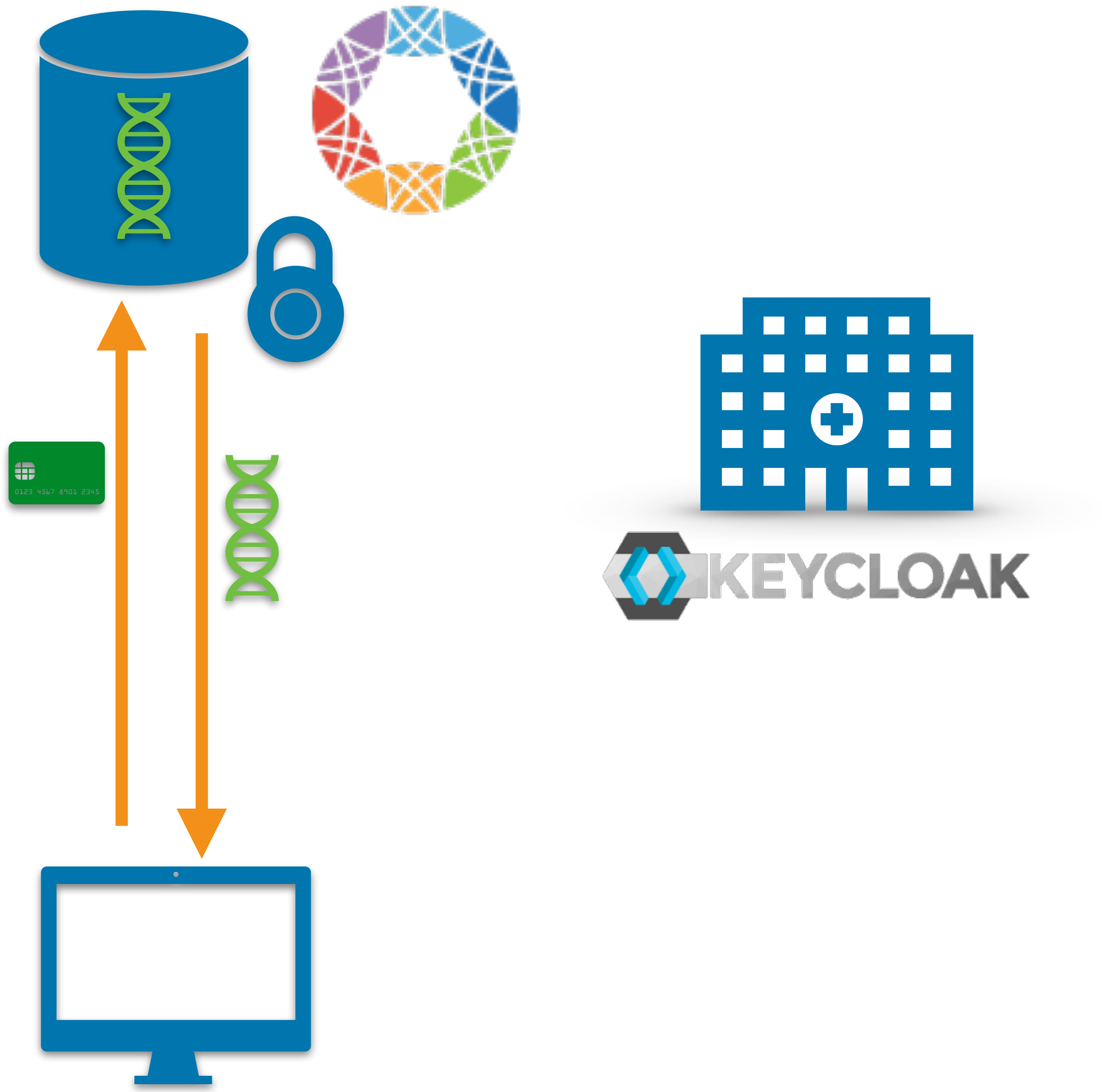
Remote images

- **Task Executor**: using Funnel out of OHSU for task executor (TES)
- Proof of concept done, with authentication (Steven Li)



OIDC for Reads/Variants

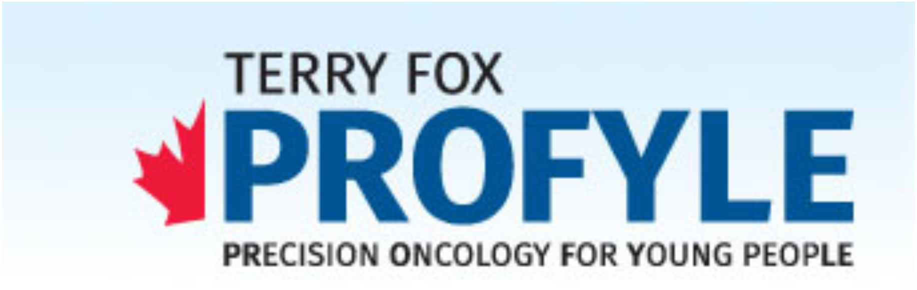
- **Reads/Variants API**: Dustin Hu and Kevin Chan have been working on using same authentication working for the Reference server for reads + variants
- Prototype working now



CanDIG Proposal

Year 1: PROFYLE

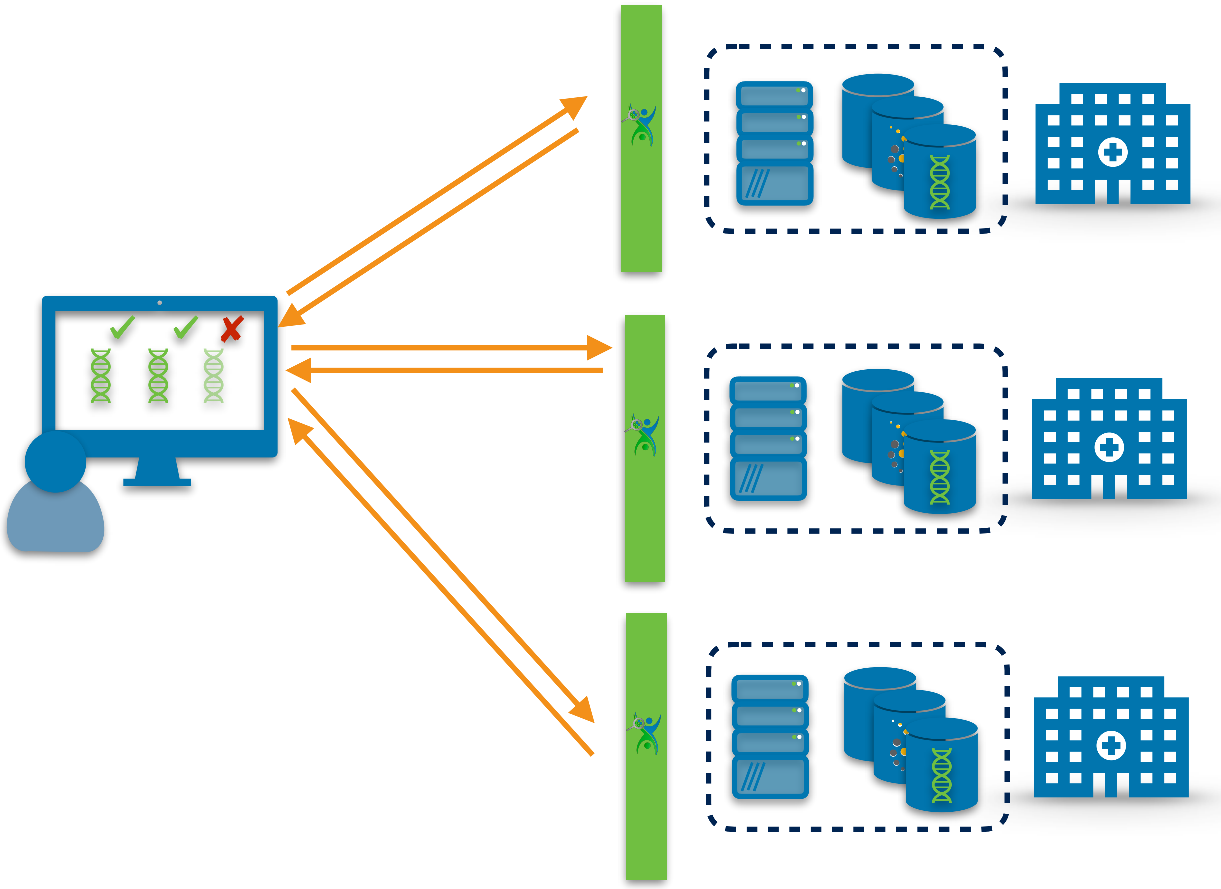
- Precision Oncology For Young PeopLE
- National paediatric oncology project
- Distributed data, many steps: need to keep track of what's available, where
- Current plans for metadata distribution: rsyncing directories of anonymized metadata files



Year 1: PROFYLE

- Proposal: start supporting PROFYLE in next several months by serving as distributed data directory/dashboard
- One of our team members (David Bujold) is developing PROFYLEs metadata schemas; is expressible in GA4GH schemas
- Over remainder of year:
  - Provide access, simple analysis of VCFs, BAMs (Large-Scale genomics)
    - Hypothesis-driven variant analysis in regions/genes in sub-cohorts - richer queries
    - Hypothesis-driven joint variant calling in specific genes
  - Pipelines - Variant Calling, RNASeq - (Cloud)

TERRY FOX PROFYLE
PRECISION ONCOLOGY FOR YOUNG PEOPLE

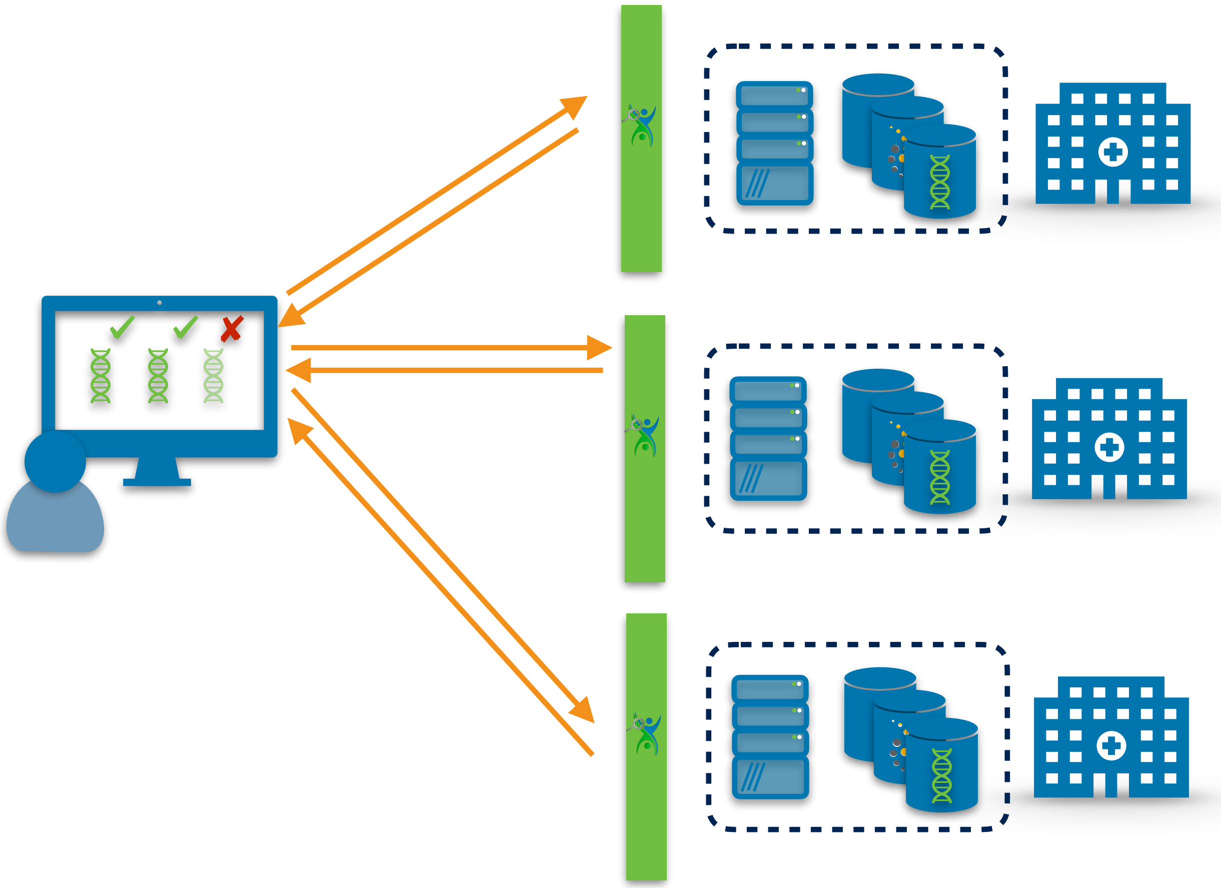


Year 1: PROFYLE

- Roadmap would look like:
  - “Productionize” our current OIDC authentication of GA4GH reads/ variants server - work with large scale genomics team
  - Develop API-powered dashboard for data directory
  - Stand up PROFYLE CanDIG servers
  - Automate ingestion of PROFYLE metadata
  - Federated authentication
  - Expand to reads and variants, enable some simple analyses/visualizations
  - Expand to pipelines

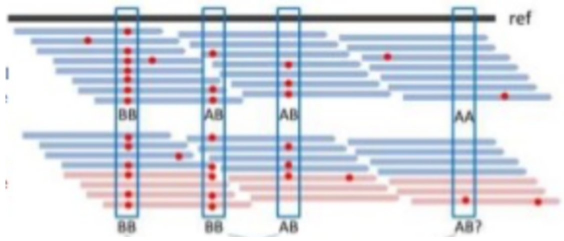
TERRY FOX PROFYLE

PRECISION ONCOLOGY FOR YOUNG PEOPLE



Joint Variant Calling

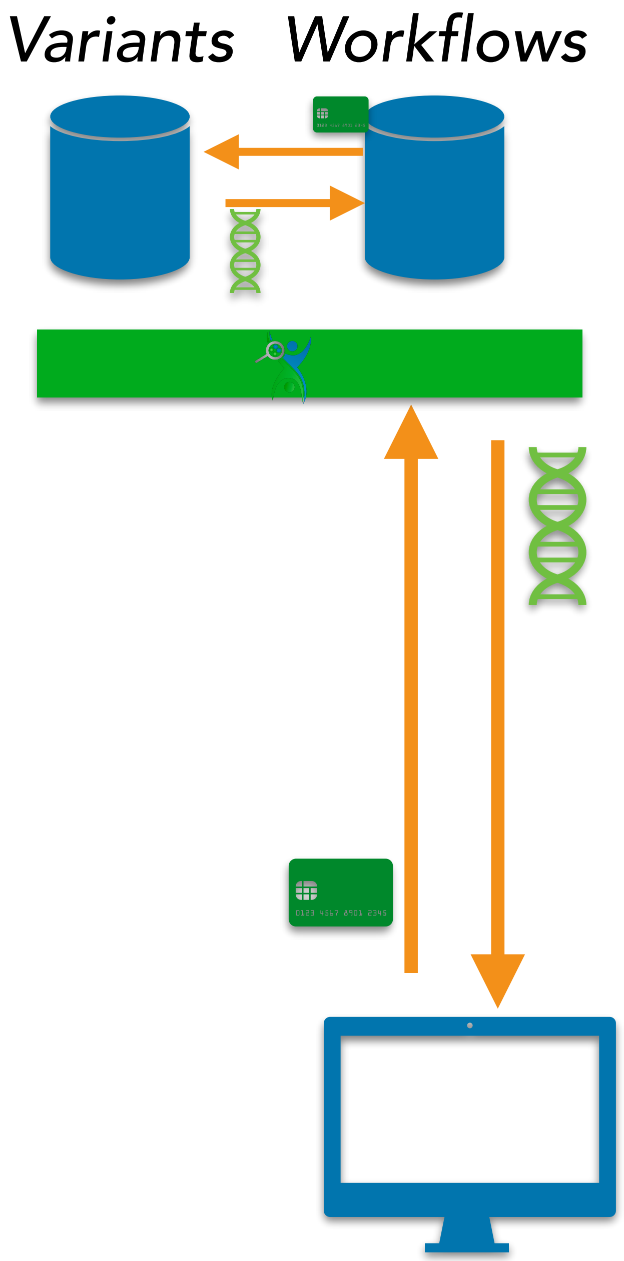
- Would be extension of federated analysis work, but on reads rather than variants
- Primarily bioinformatics methods development
  - Access reads, call locally
  - Aggregate calls, calculate MAF
  - Re-call locally given updated priors
  - Iterate as needed
- Work with Large-Scale Genomics team, and our collaborators at DNAStack



DNASTACK

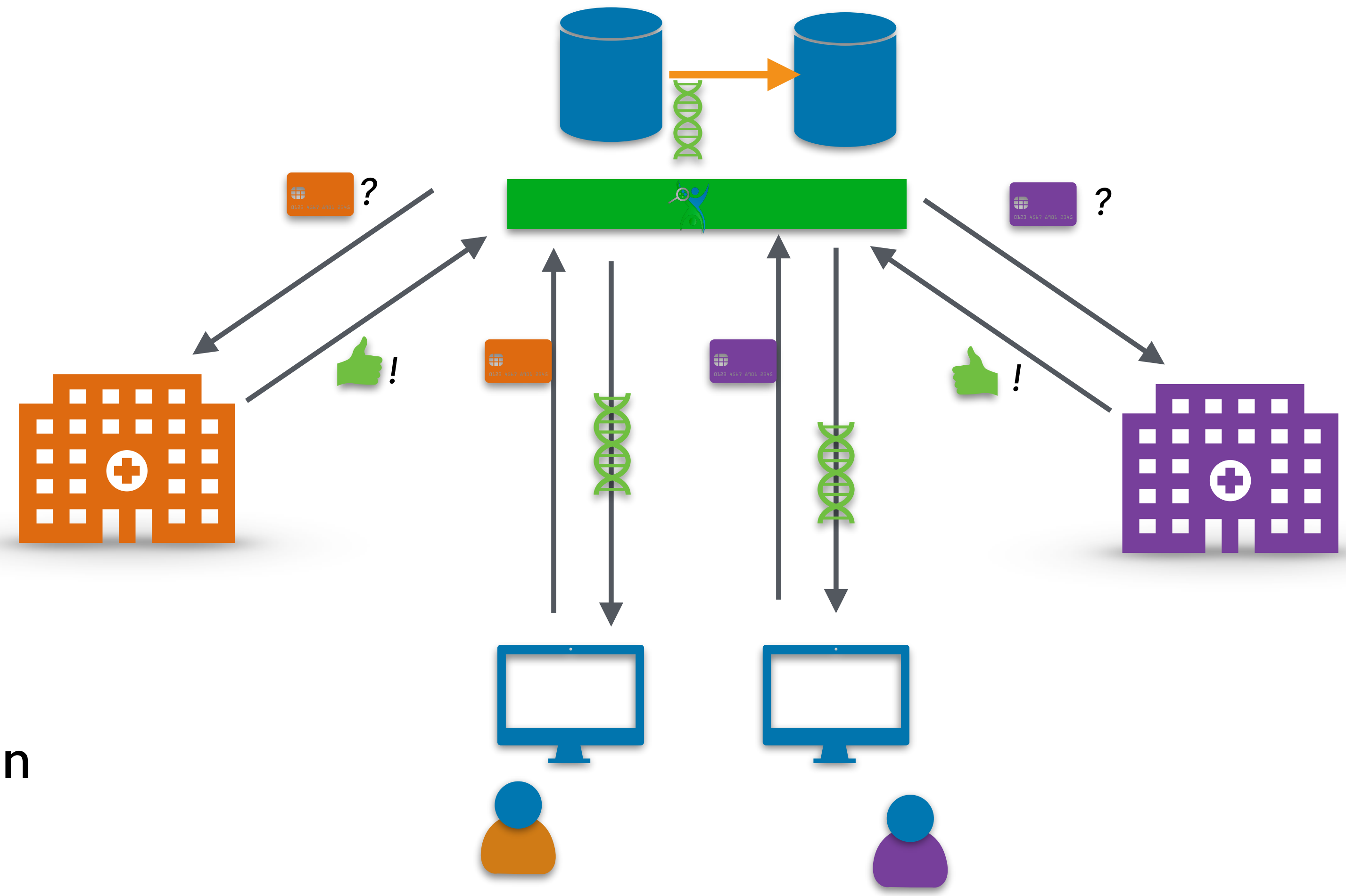
Tighter Interop Between Reads/ Variants & Tasks/Workflows

- Want *all* data access to be through APIs: logging, auditability
  - No Dockers dropped in a directory of files
  - For variants, may not *be* files: variant stores
- TES/WES (Cloud) and Reads/Variants (Large-Scale Genomics) servers currently unaware of each other
- Stage in data from Reads/Variants via htsget, make set of servers more easily deployable together



Building Federated Authentication

- Federated OIDC: work with Beacon-Network to take advantage of their efforts
- Allow local participating sites to authenticate their users as valid CanDIG users
- Accept OIDC Authentication tokens from registered sites. Local data steward makes authorization decisions based on identity, roles, dataset
- Will require beginnings of Authorization implementation in Reads/Variants - work with Large-Scale Genomics
- "Single-Sign On" behind the API - work with Access & Authentication



Year 2: CaMPACT

- Based on existing IMPACT and COMPACT trials
- CaMPACT is infrastructure for Canada-wide basket-style clinical cancer trials
- Clinicians will use cBioPortal to consider patients for assignment to trials
- Researchers will examine data for hypothesis generation
- Will require authorization as well as authentication - not all groups may access all data (Access & Authentication, SWG)
- Will require integration of clinical and phenotypic systems (Clinical and Phenotypic)

cBioPortal
FOR CANCER GENOMICS

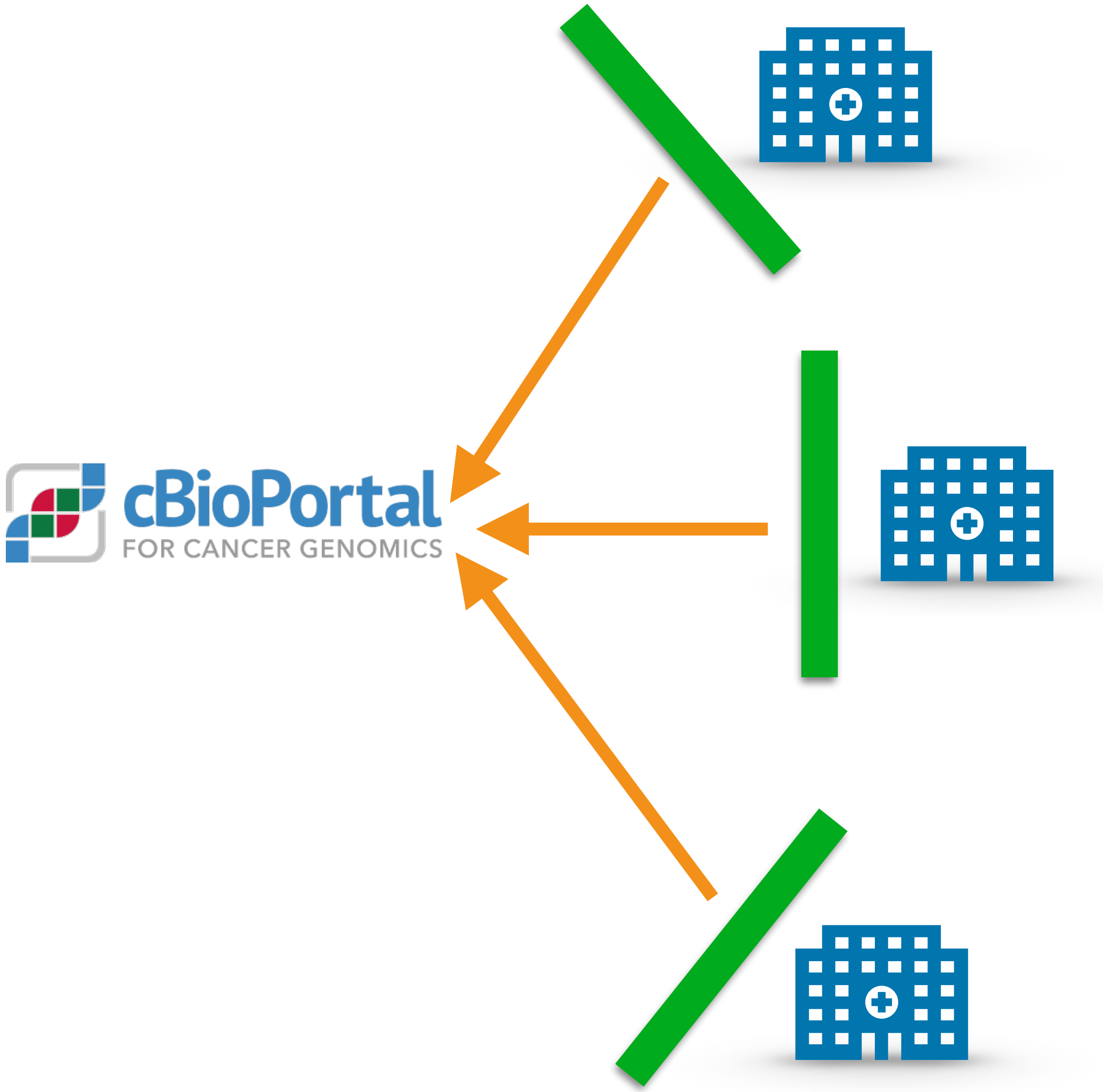
Year 2: CaMPACT

- Will benefit from all of work done for PROFYLE
- Can begin testing infrastructure for COMPACT early with synthetic GENIE data and then full GENIE data once approved
  - Builds on auth^{n}, metadata, dashboard, richer queries
- Team member has already included support for OIDC into cBioPortal upstream for future support of CanDIG authentication

PROJECTGENIE

Genomics Evidence Neoplasia Information Exchange

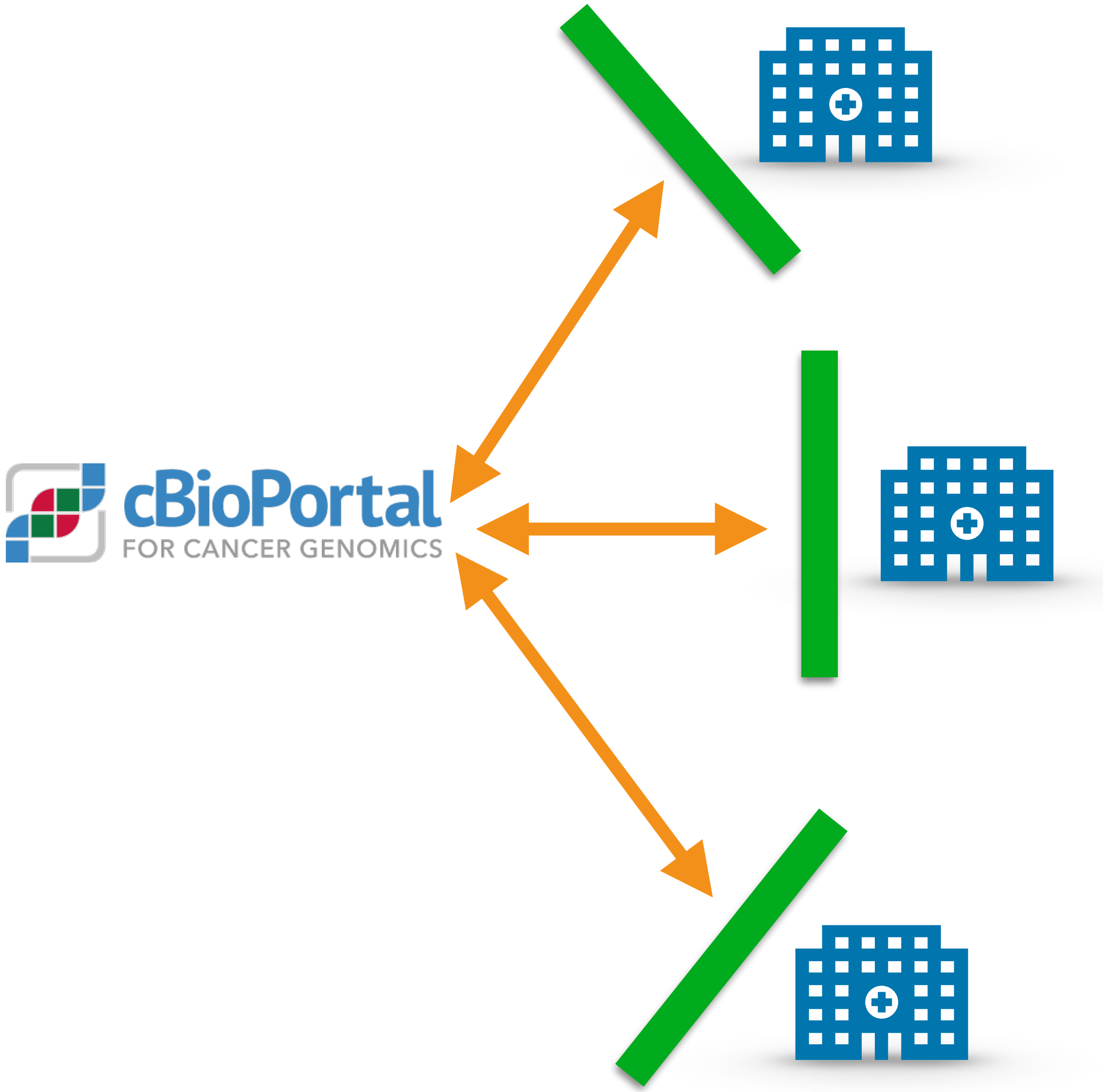
Towards CaMPACT



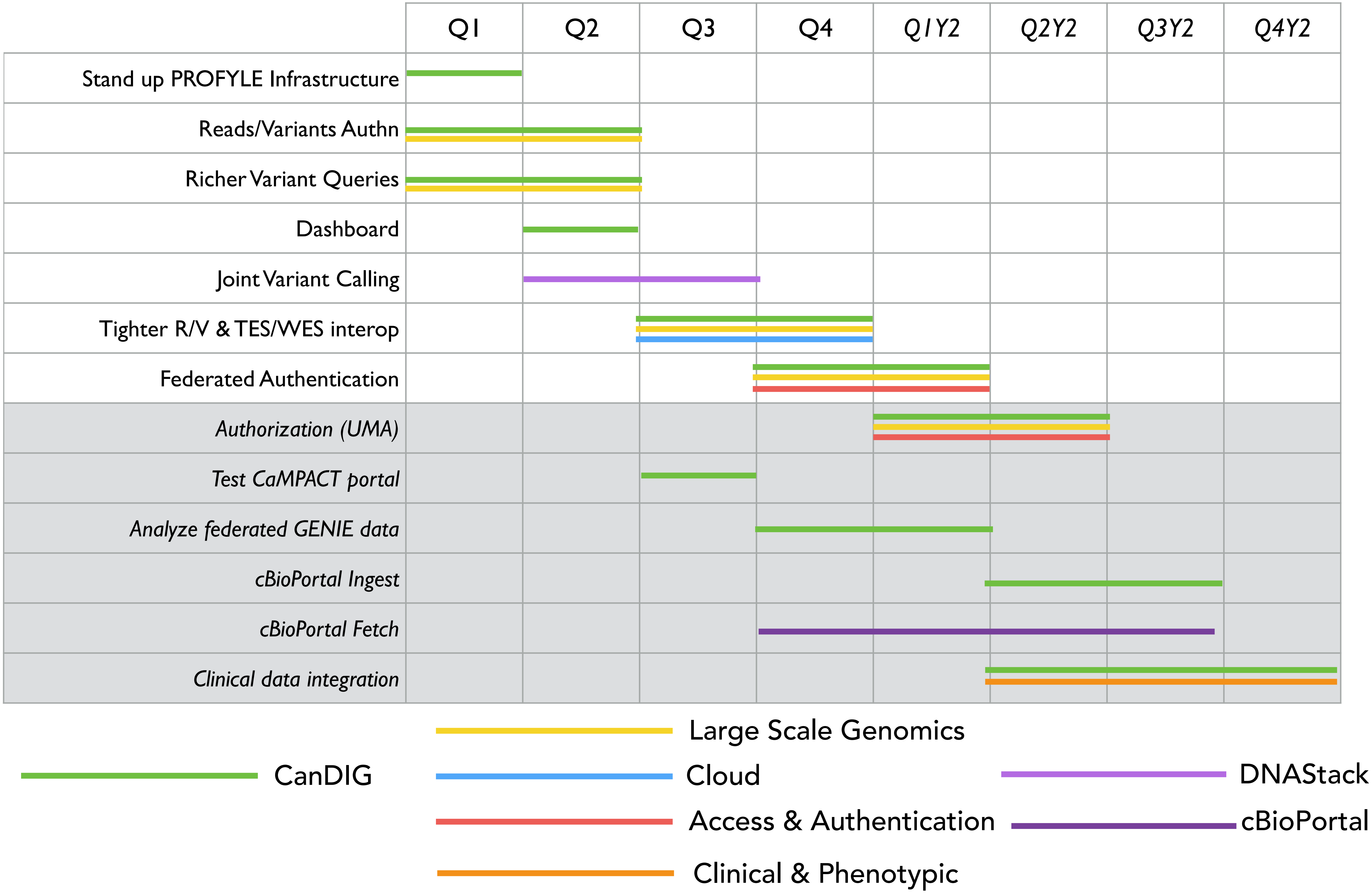
- First step: routinely ingest data from sites (htsget), process, update cBioPortal database

Towards CaMPACT

- Second step: ingest + aggregate metadata, query and display data over APIs as needed for deep dives
    - Would require extensive work in cBioPortal
- In either case, research analysis goes through Reads/Variants API + clinical data schemas



Draft Work Plan



	Q1	Q2	Q3	Q4	Q1Y2	Q2Y2	Q3Y2	Q4Y2
Stand up PROFYLE Infrastructure	🟩							
Reads/Variants Authn	🟩🟨	🟩🟨						
Richer Variant Queries	🟩🟨	🟩🟨						
Dashboard		🟩						
Joint Variant Calling		🟪	🟪					
Tighter R/V & TES/WES interop			🟩🟨🟦	🟩🟨🟦				
Federated Authentication				🟩🟨🟥	🟩🟨🟥			
Authorization (UMA)					🟩🟨🟥	🟩🟨🟥		
Test CaMPACT portal			🟩					
Analyze federated GENIE data				🟩	🟩			
cBioPortal Ingest						🟩	🟩	
cBioPortal Fetch				■	■	■	■	
Clinical data integration						🟩🟧	🟩🟧	🟩🟧

🟩 CanDIG

🟨 Large Scale Genomics

🟦 Cloud

🟥 Access & Authentication

🟧 Clinical & Phenotypic

🟪 DNAStack

■ cBioPortal