Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

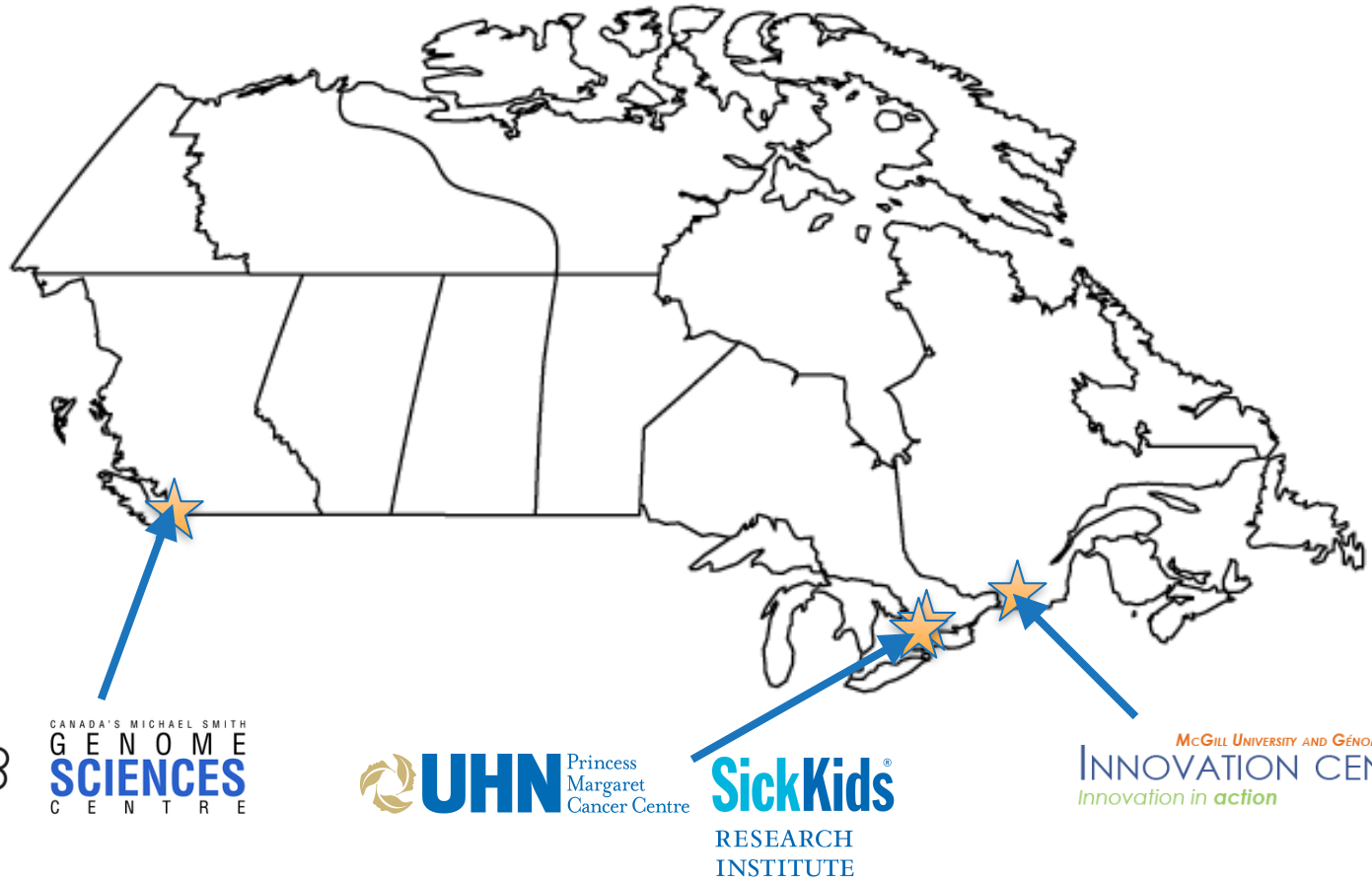
CanDIG

Distributed national analyses of locally-controlled genomic data

http://distributedgenomics.ca

Canadian Distributed Infrastructure for Genomics (CanDIG)

Global Alliance for Genomics & Health



New (start date: this spring) 4-year funded Canadian project to enable batch and interactive analysis over national cohorts with provincially controlled private genomic data - send analyses to data.

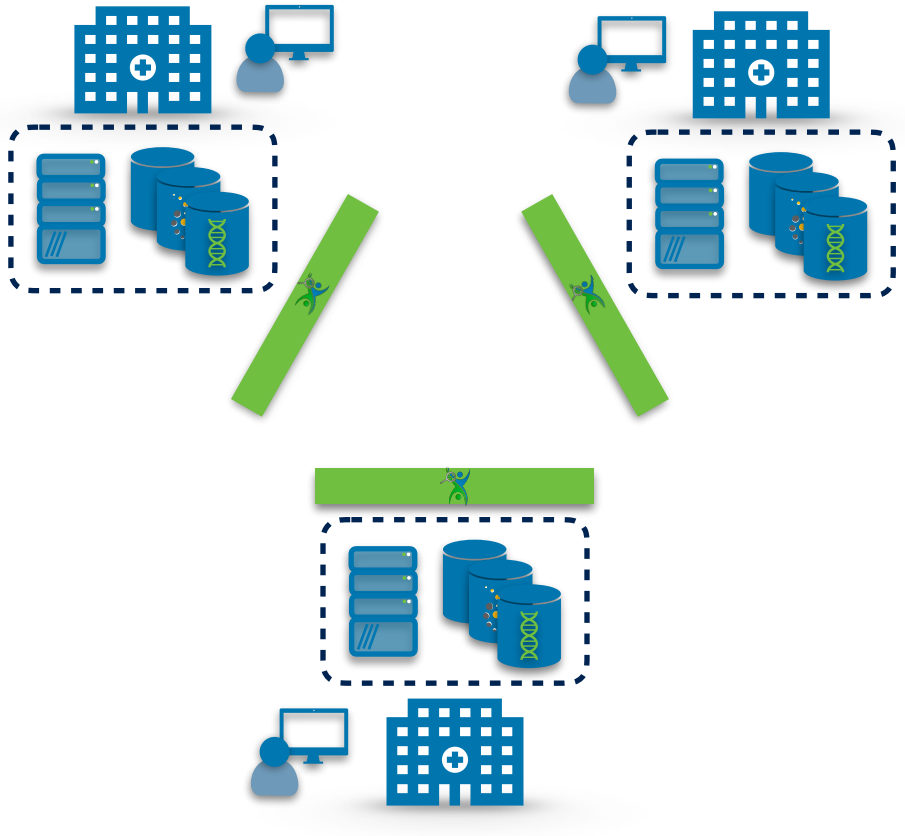
CanDIG:

- Over coming months:
  - Support paediatric cancer project (PROFYLE)
    - Provide data directory, dashboard, coordinate processing
    - Expand to directly supporting analyses
  - Support for basket-type cancer clinical trial project (CaMPACT)
    - Distributed data platform
    - Support clinician decision-making by interfacing with cBioPortal

- By year 4:
  - Large scale data directory
  - Analysis interface to large amount of research & clinical genomics data
  - “App store” of available analyses - interactive and batch
  - Privacy layer
  - Programatic access for development of new distributed analyses methods

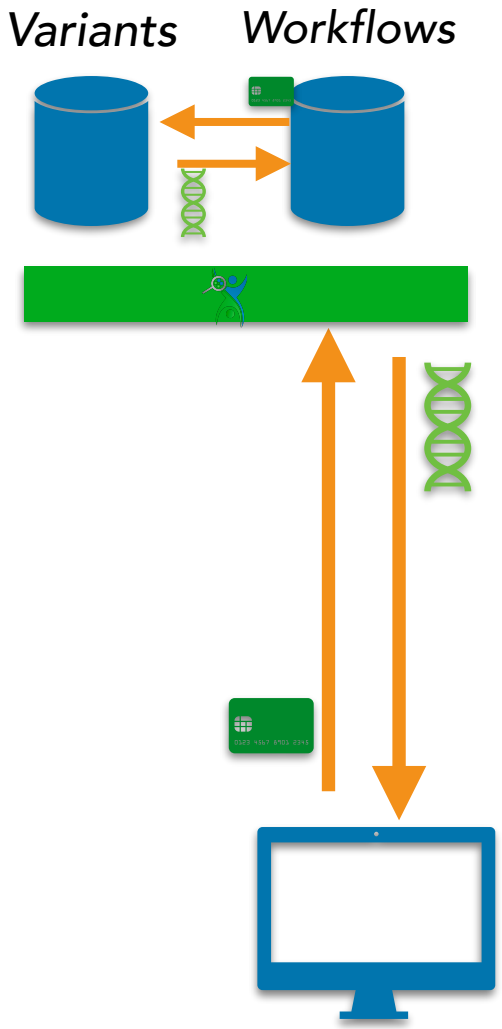
Platform Goals - Fully Distributed:

- Participating sites: provide access to data, source of user requests
- Distributed synchronization of apps available, project membership, etc.
- Sites authenticate their users
- Local sites control access to their data



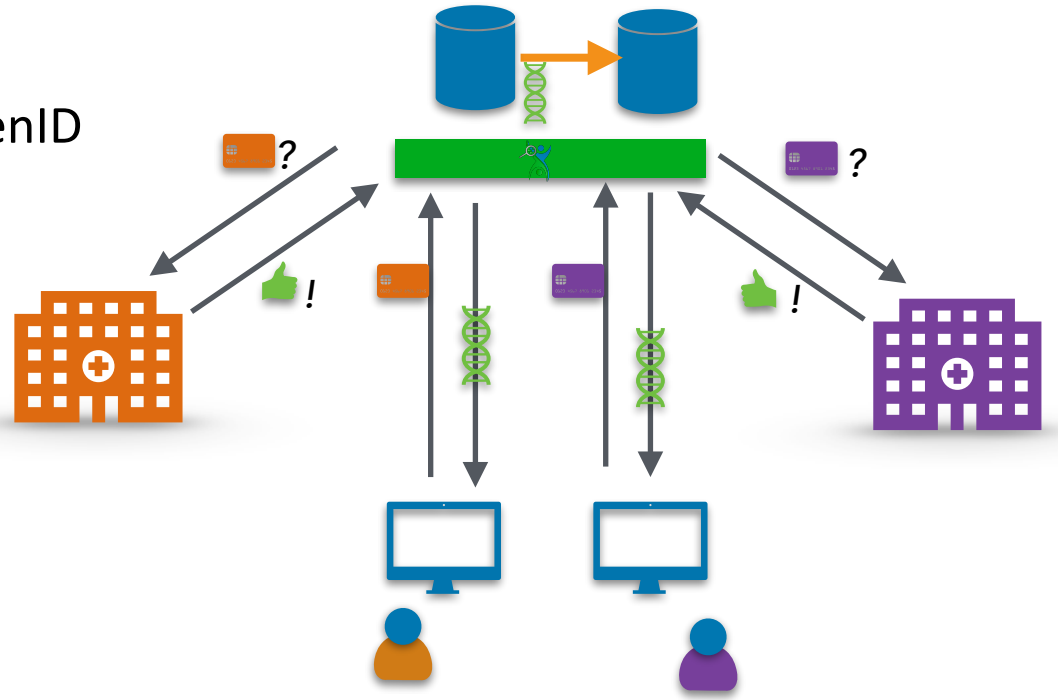
Platform Goals - API access:

- Want all data access to be through APIs: logging, audibility; no processes dropped in directory of files.
- Maybe no files: opaque back-end to different data stores (files, variant data bases, etc)
- WES (**Cloud**) and Reads/Variants servers communicating internally via htsget (**Large-Scale Genomics**)
- Metadata/clinical data standards (**Clinical & Pheno Data Capture**)



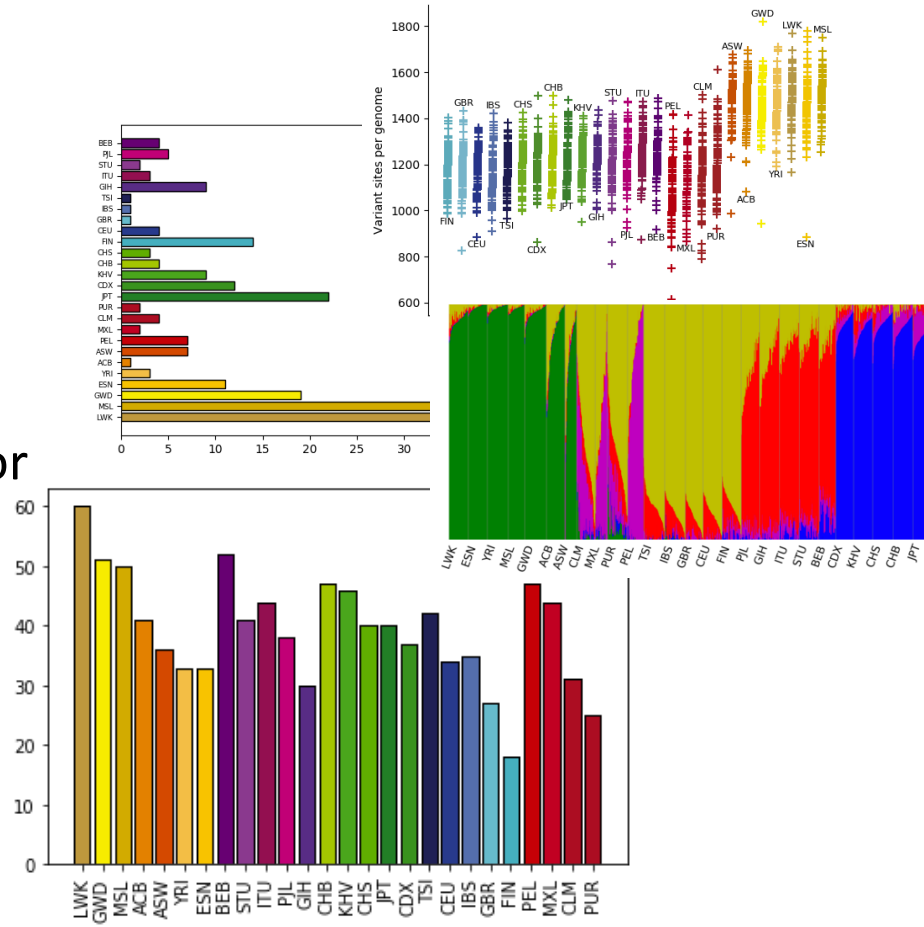
Platform Goals - AAI:

- Authentication: Federated OpenID Connect
- Local site authorizes based on remote ID and distributed role information
- Verified tokens used internally amongst services
- Build with eye towards future interoperability with **DURI**



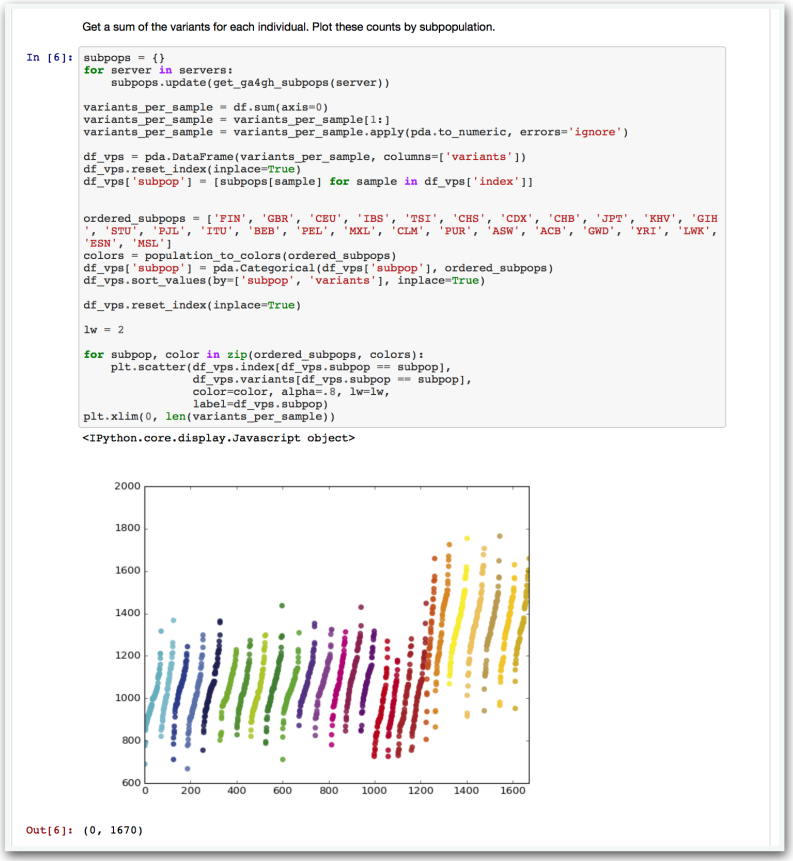
Work so far - interactive analysis

- Less obvious it would work nicely in our federated context
- E.g., re-creating some classic thousand genomes figures across federated datasets - small regions for interactivity



Work so far - interactive analysis

- Less obvious it would work nicely in our federated context
- E.g., re-creating some classic thousand genomes figures across federated datasets - small regions for interactivity



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))
```

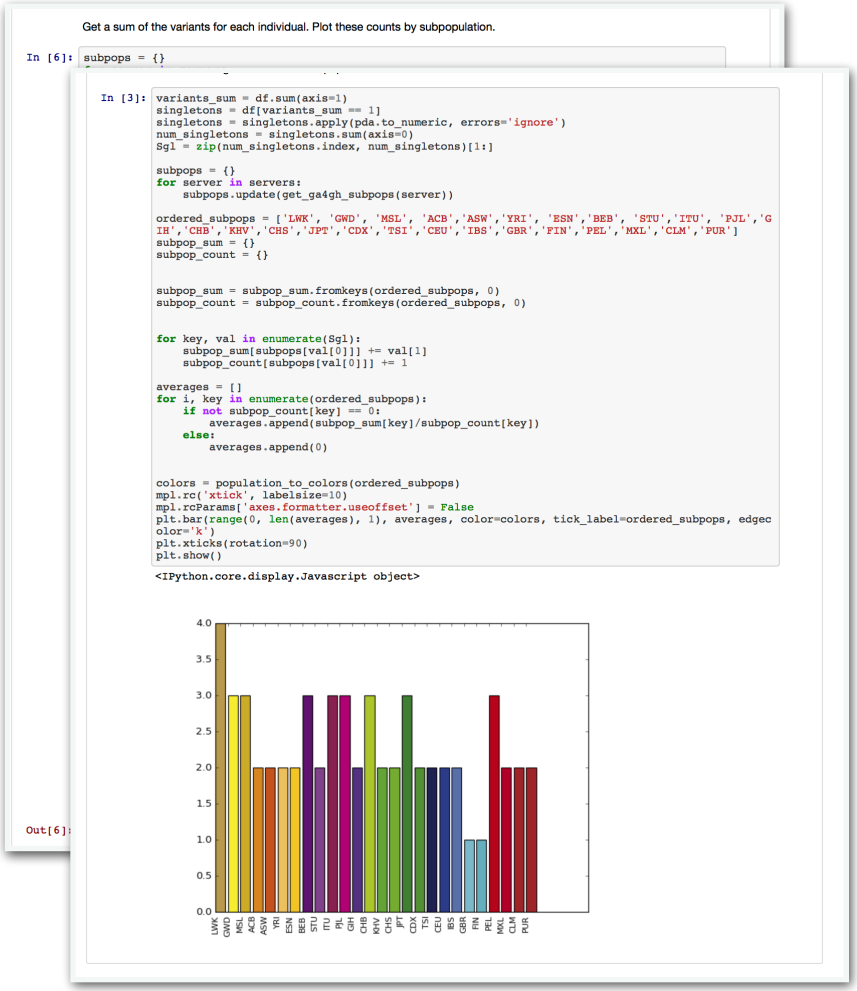
```
<IPython.core.display.Javascript object>
```



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

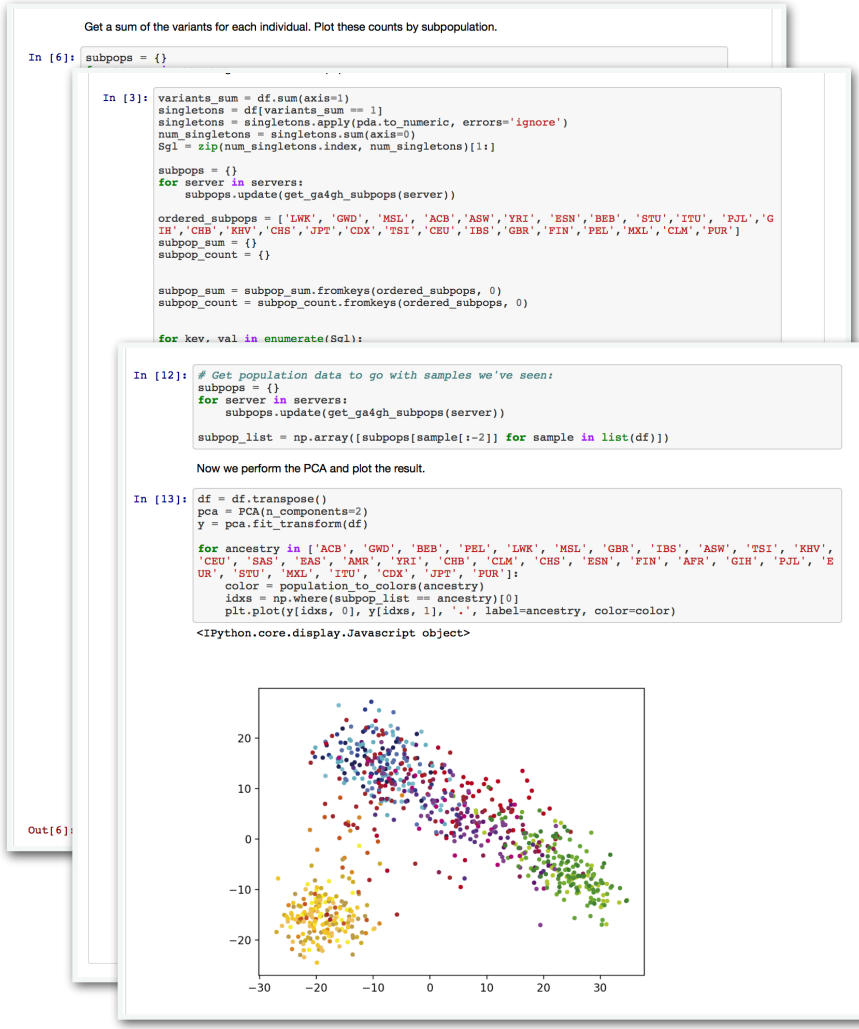
Work so far - interactive analysis

- Less obvious it would work nicely in our federated context
- E.g., re-creating some classic thousand genomes figures across federated datasets - small regions for interactivity



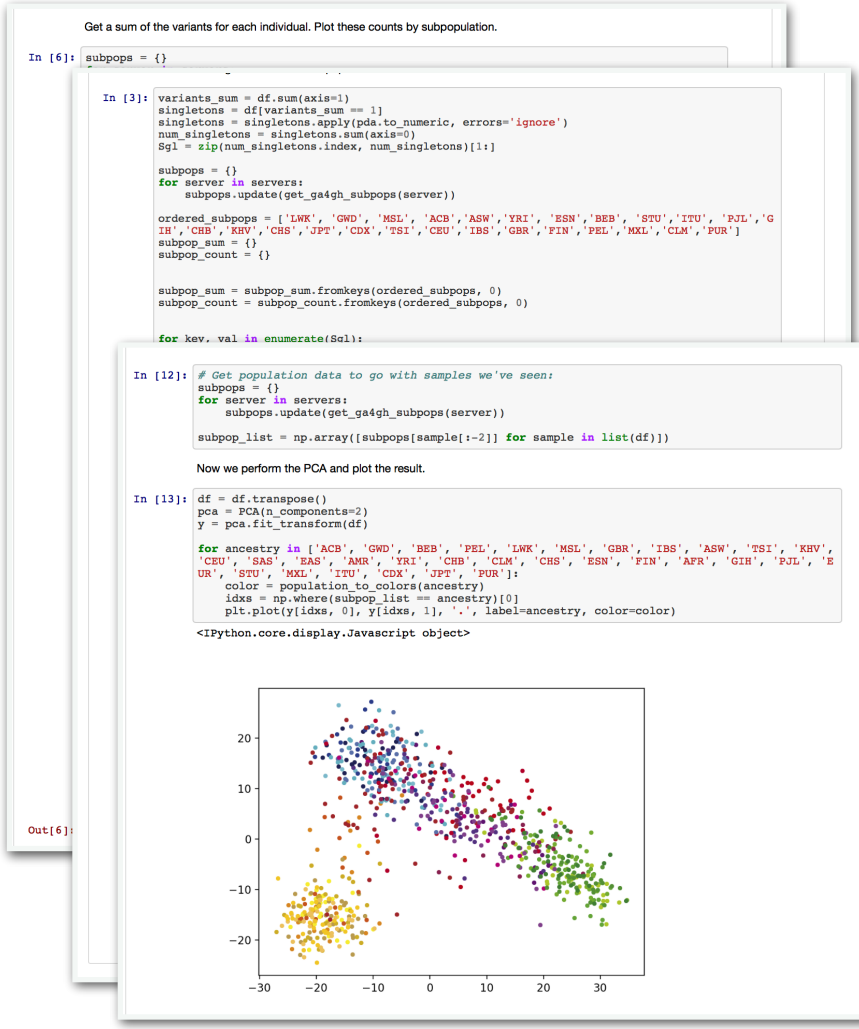
Work so far - interactive analysis

- Less obvious it would work nicely in our federated context
- E.g., re-creating some classic thousand genomes figures across federated datasets - small regions for interactivity



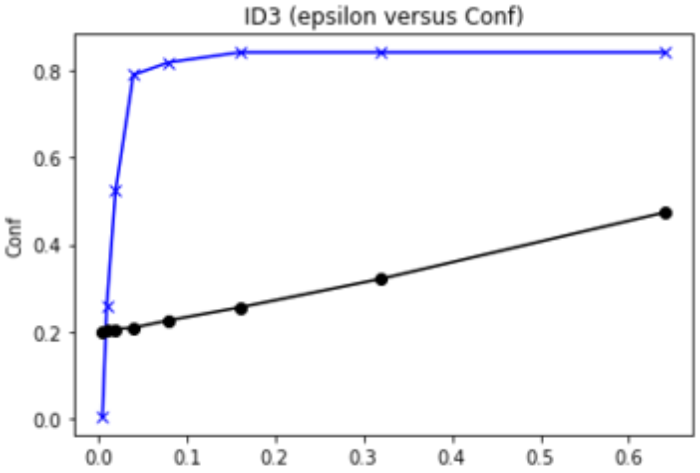
Work so far - interactive analysis

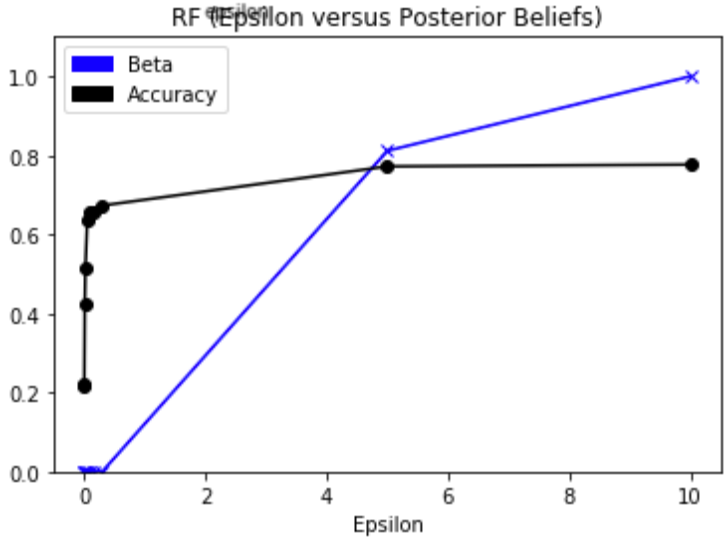
- Needed to greatly enhance R & V server performance
  - Serialization
  - “Column-oriented” approach to (e.g.) FORMAT fields
  - Contributed back
  - J. Foong, HSC
- Gives good indication on where **aggregation**, **filtering** queries will be needed
- Federated queries in a CanDIG layer



Work so far - differential privacy

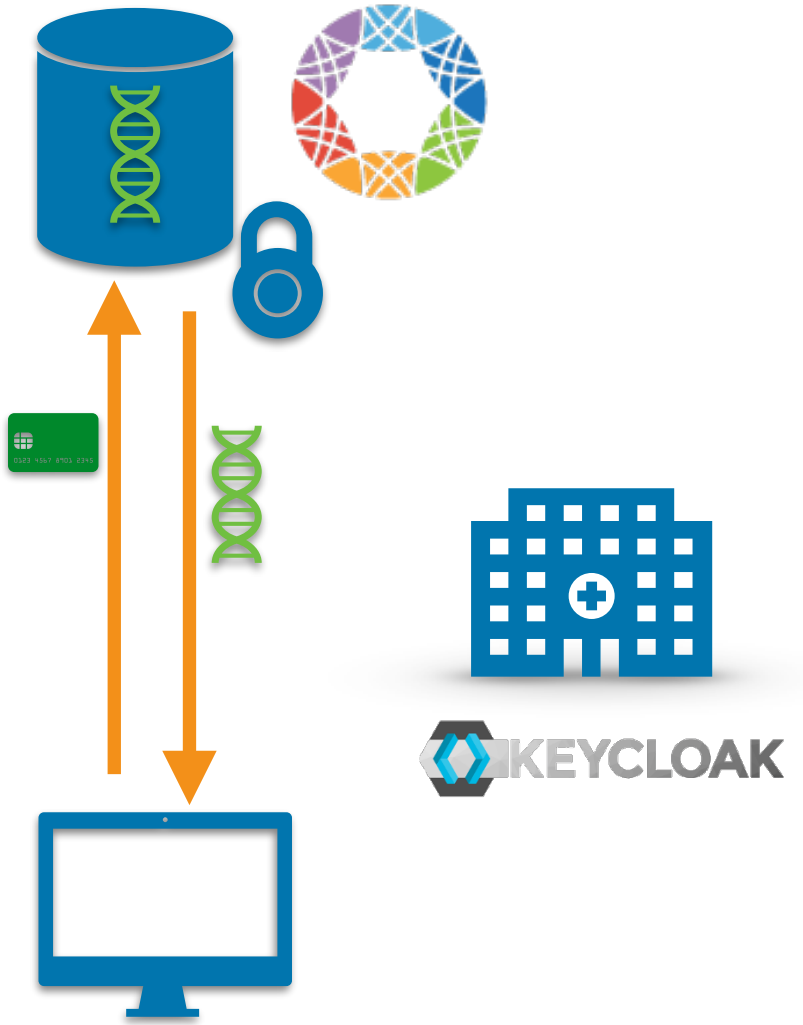
- With counting queries, raises possibility for introducing (*e.g.*) differential privacy
- Make it easier for sites to make available data they might not otherwise
- Federated classifier training with differential privacy over R&V API:
  - What approach works best, with real privacy model?
  - What happens when different sites have different privacy requirements?
  - N. Memon, BCGSC





Work so far - authentication

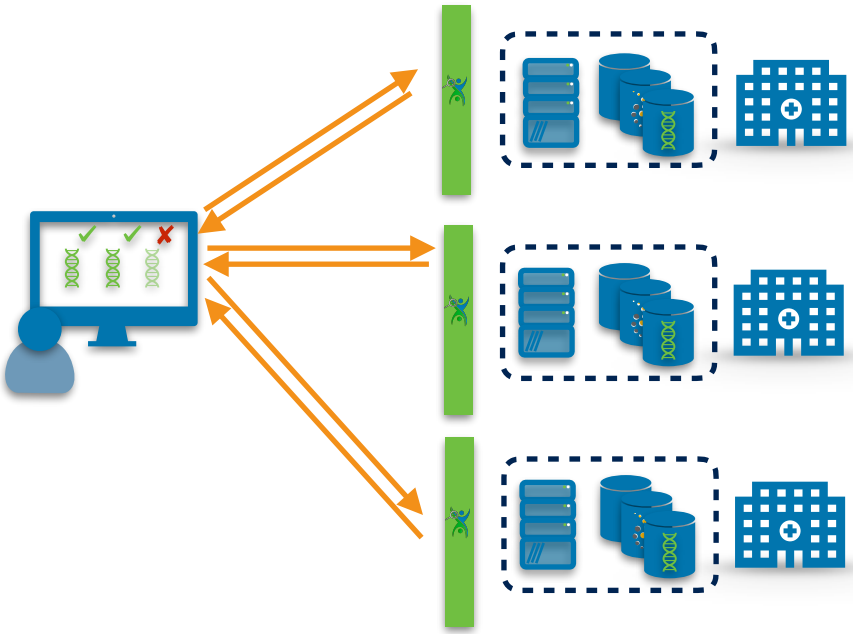
- Robust, standards-based OIDC authentication for R&V server
- R. deBorja and others, UHN



Current work - PROFYLE

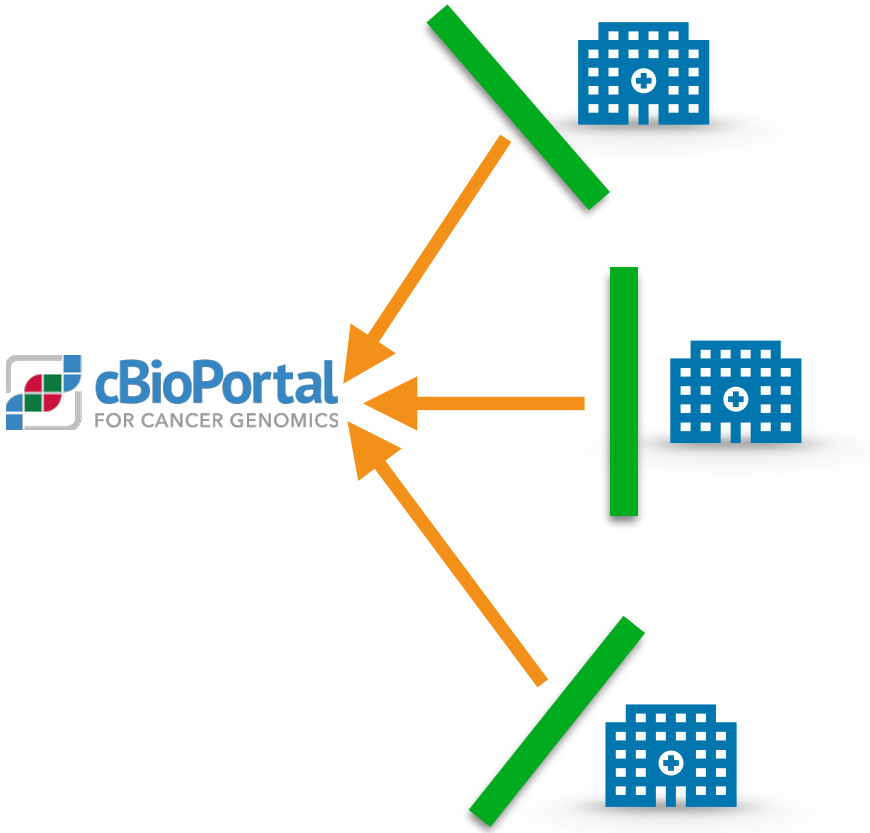
- National paediatric precision oncology project
- Data catalog/dashboard for project
- Extend to analyses, data access
  - Existing work w/ IGV.html, simple analyses (joint variant calling at locus)
- Extended support for metadata access
- Schemas for experiments / analyses will need continued work

TERRY FOX PROFYLE
PRECISION ONCOLOGY FOR YOUNG PEOPLE



Current work - CaMPACT

- Oncology basket trial
- cBioPortal for clinician data exploration
- Remote data access, ingest into cBioPortal
- Extend to remote data API?



Coming months

- Begin building on work of **Cloud** team for batch processing/analysis:
  - TES (Funnel), WES; DOS?
- Continue building on work of **LSG** team:
  - Incorporate htsget for internal transfers
- Building AAI API gateway
- Building on, contributing to metadata standards, EHR ingest (**Clinical & Pheno capture**)

Longer-term work

- Reads API: search by content of reads (string), quality, and not just mapped location
- Work towards interoperability with **DURI** for Researcher ID and data use/authorization
- Interoperability between **LSG** & **Cloud** team genomic data access models
- **Discovery** APIs atop our platform