Intro to HPC4Health

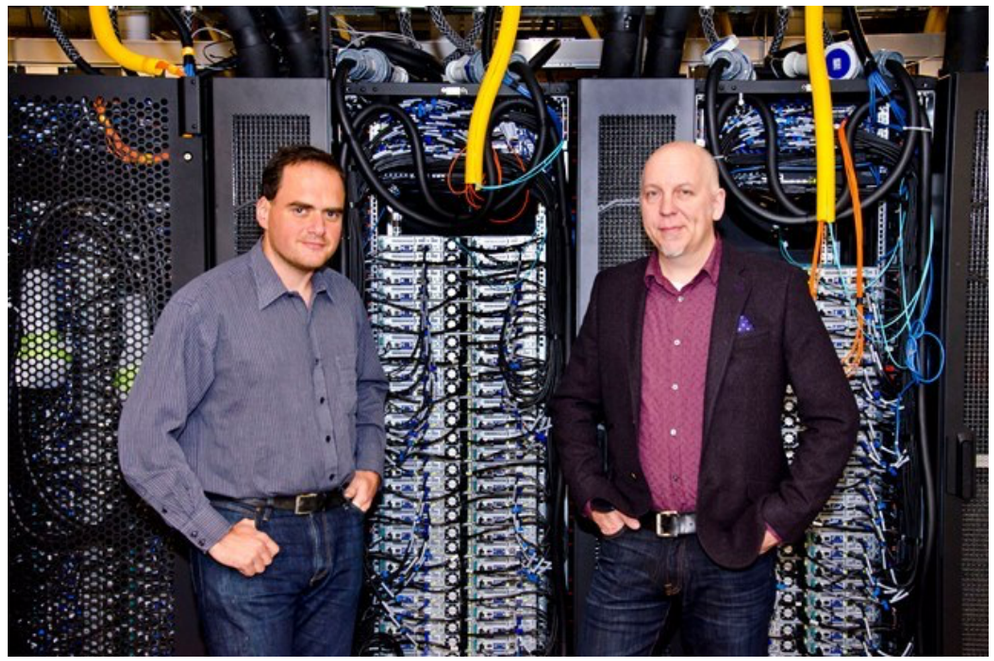
Advanced Health Research Computing Locally, Provincially, Nationally, and Internationally

CanDIG

National Analysis of Distributed Private Genomic Data

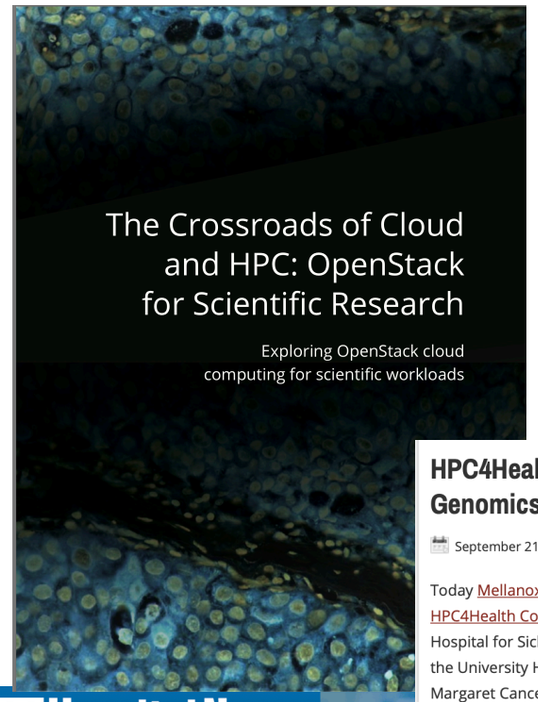
HPC4Health, 2014

- HPC4Health started in 2014 as a pilot between SickKids and UHN
  - Solving problems: space, procurement, best practices security/privacy
  - Shared computational resources for research genomics
  - Pool of compute, data, expertise



HPC4Health, 2014

- This was/is a big deal - hospitals sharing data infrastructure!
- Both technical and organizational aspects were written up widely.



HPC4Health Selects Mellanox InfiniBand for Cancer and Genomics Research

September 21, 2015 by staff Leave a Comment

Today Mellanox announced that the HPC4Health Consortium, led by The Hospital for Sick Children (SickKids) and the University Health Network's Princess Margaret Cancer Centre, has selected its InfiniBand networking solutions to improve patient care and help researchers to optimize treatment with the ultimate goal of finding a cure for cancer. The end-to-end FDR 56Gb/s InfiniBand networking solution was adopted as the foundation of the center's cancer and genomics program, to accelerate the sharing, processing and analysis of data generated from radiology imaging, medical imaging analysis, protein folding, x-ray diffraction in order to improve patient care and expedite cancer research.

Hospital News
CANADA'S HEALTH CARE NEWS AND BEST PRACTICES

NEWS & TOPICS MEDICINE COLUMNS LONG

Home › Topics › Innovation and Technology › SickKids-UHN partnership bri

Topics Innovation and Technology

SickKids-UHN partnership brings cloud technology to health care

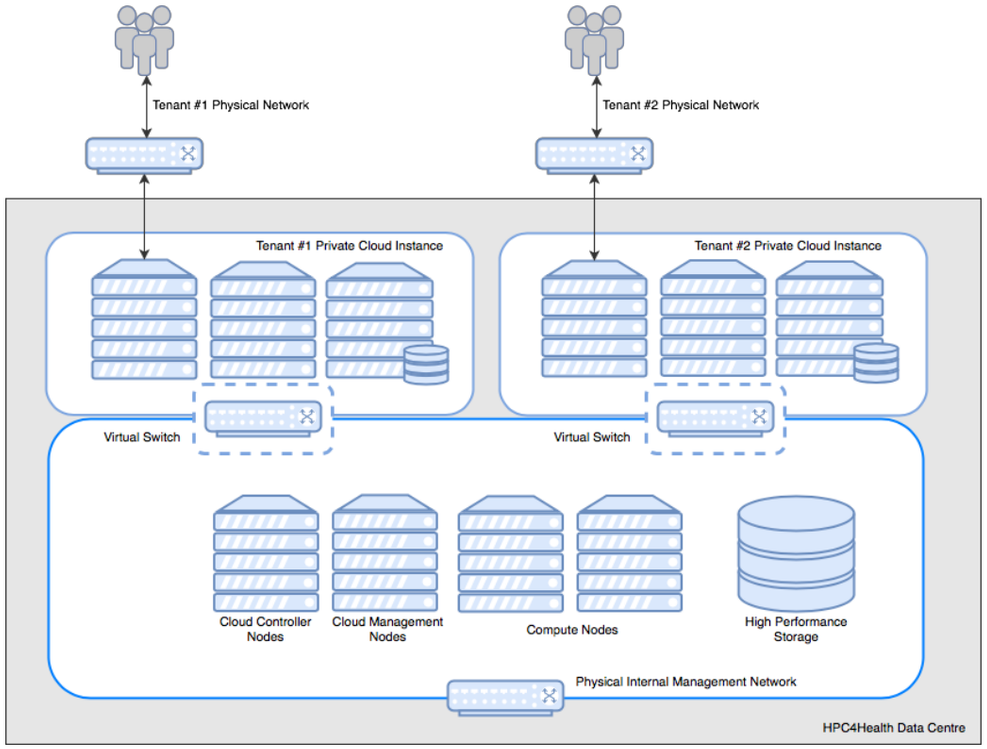
1862

Michael Brudno on the left, and Carl Virtanen on the right. Photo credit: Robert Teteruck, The Hospital for Sick Children

More and more, the fields of biology and genomics are becoming big data sciences. Today research discovery and innovation are made possible by not only experiments in the laboratory, but also through the power of high performance computing (HPC). Without access to such computational resources, it would not be possible to analyze and interpret the terabytes of data generated every day that contribute to scientific discovery.

HPC4Health Today

- Sophisticated Secure Tenant Infrastructure
- Housed on 6th floor
- Six scientific staff
- Eight technical staff
- Four+ partner institutions across Ontario
- 2 PB of secure data
- 7,000 compute cores



HPC4Health and Compute Ontario: Helping Deploy AHRC Nationally

- Reproducible Infrastructure being used:
  - Calcul Quebec secure cloud @ McGill
  - Health Cloud @ McMaster, funding secured by CO
- Best practices in security, privacy, administration of ESC

HPC4Health Provincially

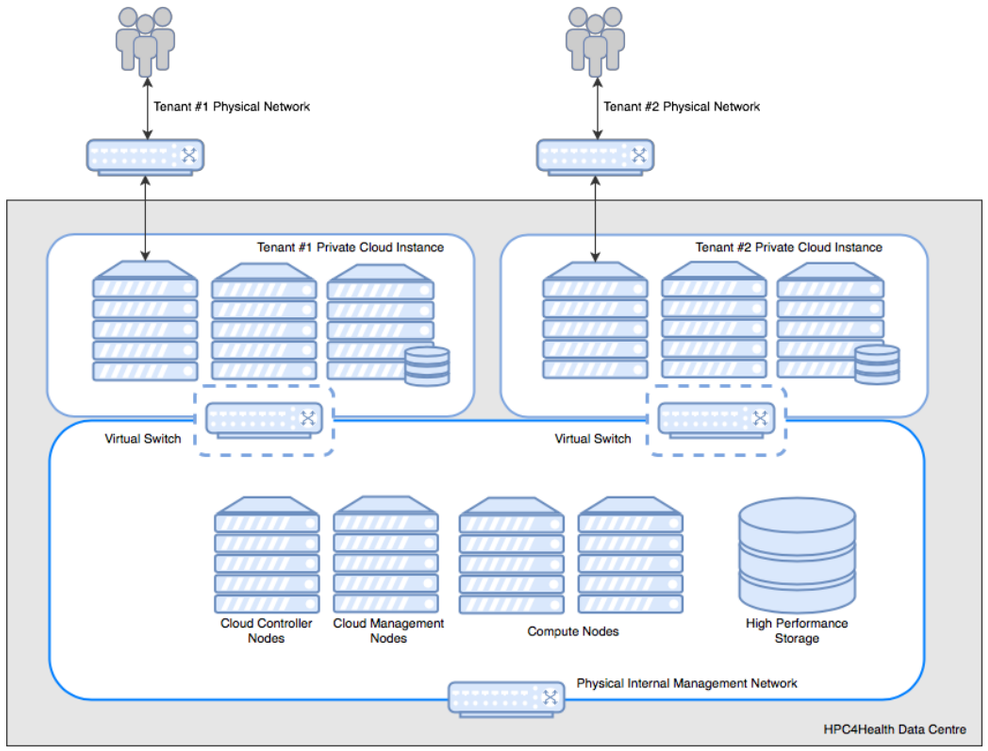
- From other speakers, will hear about HPC4Health services' outsized impact on Toronto and Ontario health research
- Bioinformatics services
- Cancer projects: OCTANE (ON)HIDAP

HPC4Health Nationally

- From other speakers, will hear about HPC4Health services' outsized impact on national health research computing
  - Bioinformatics national team
  - Extending the H4H architecture model
- I'll focus on a project I know particularly well, which is led by H4H - CanDIG

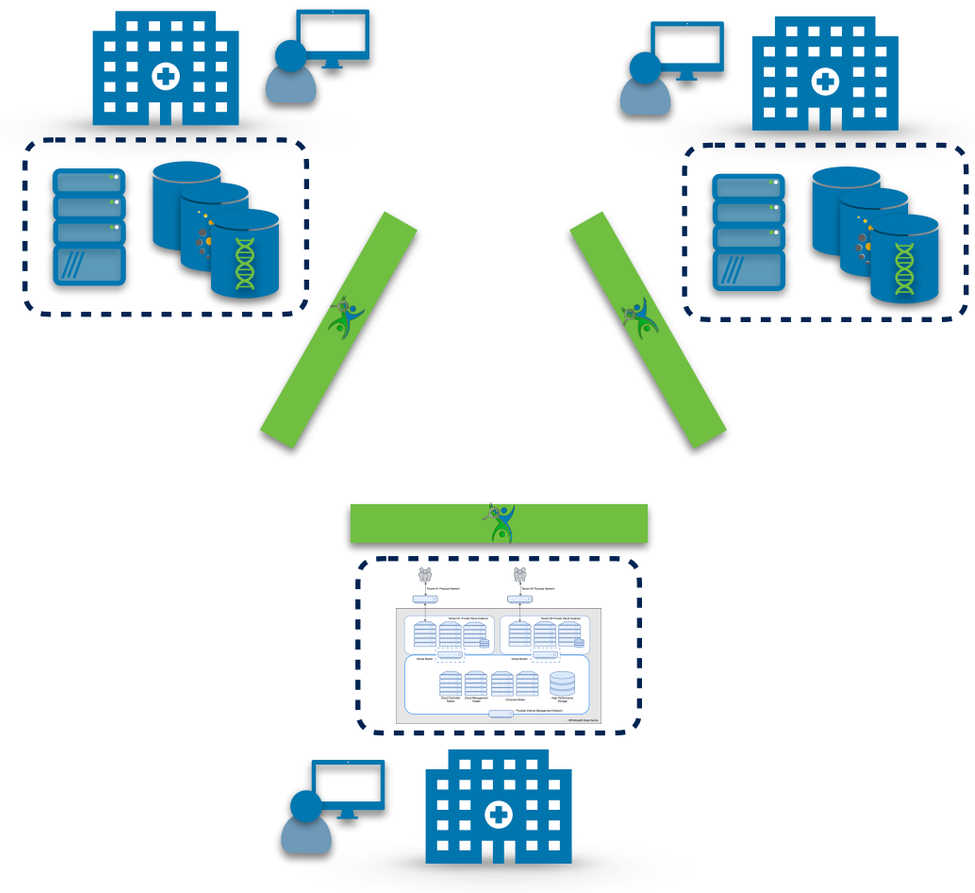
HPC4Health and CanDIG

- HPC4Health's system & organizational architecture allows combining of
  - Physical Infrastructure
  - Expertise
- It does *not* directly combine data.
- Data stewards must be able to control access to their data
- Even more complicated in Canada



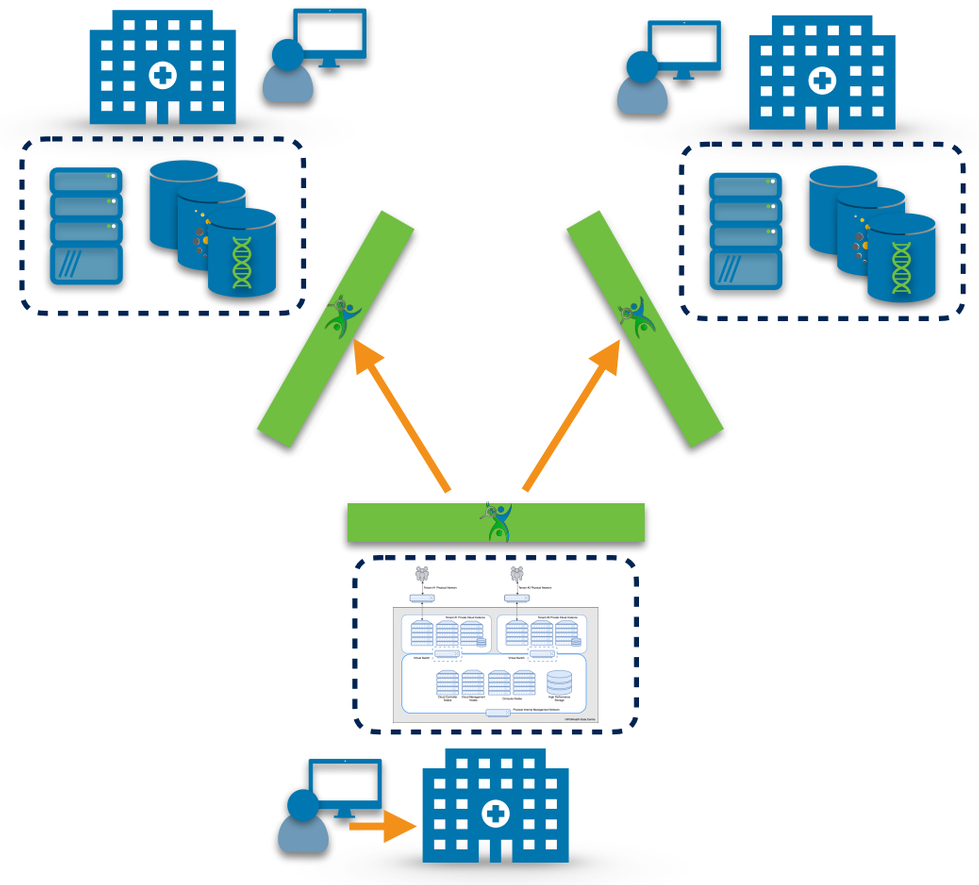
CanDIG - Led by HPC4Health

- A Canadian approach to analysis of health research data:
  - National-scale populations
  - Respecting provincial, institutional stewards local control over their data, users.
- Funded 4 year CFI Cyberinfrastructure project, ~6 FTEs and rapidly growing
- Comes from H4H cloud, bioinformatics expertise



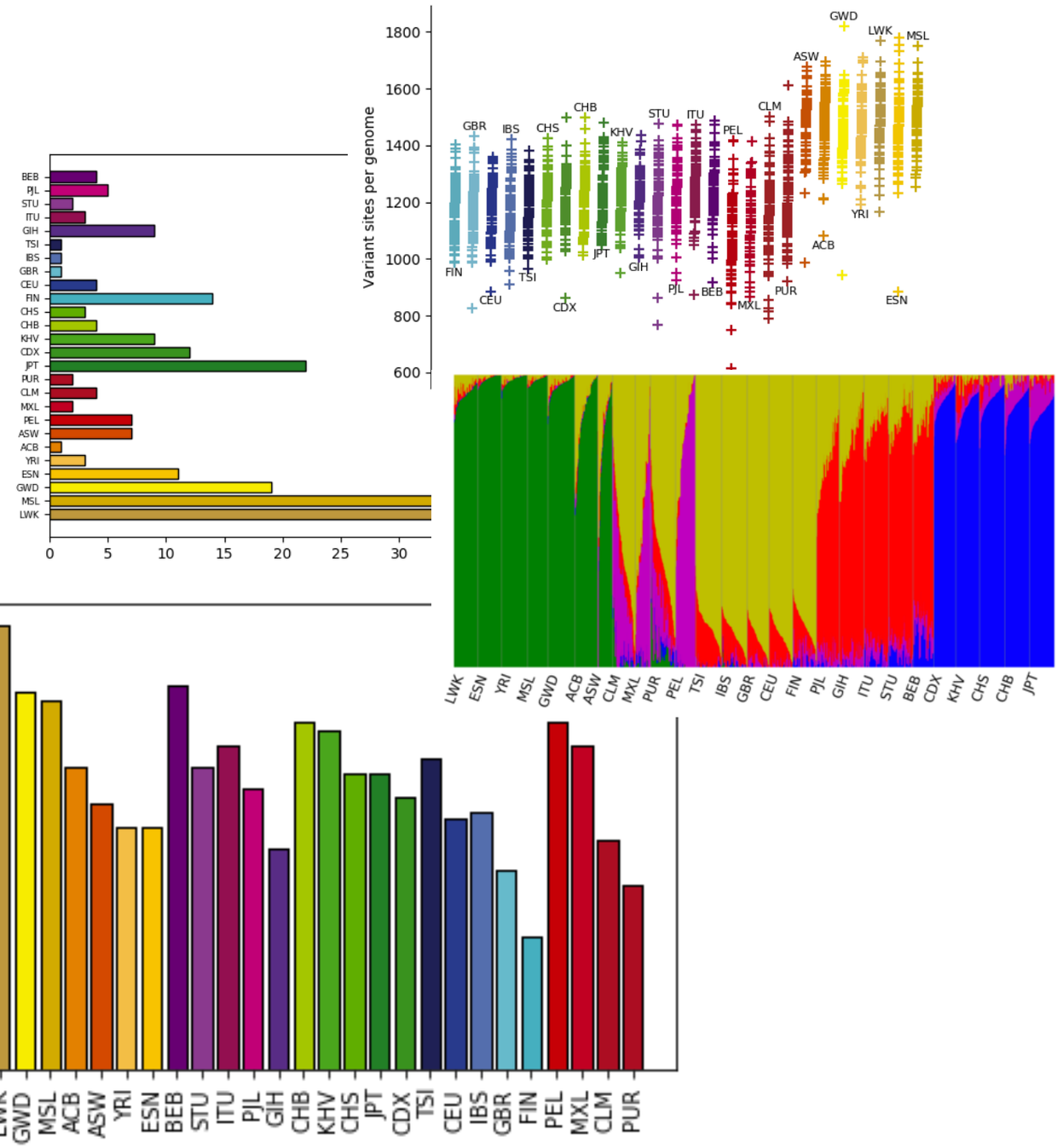
CanDIG

- Fully distributed and federated
- Send analysis to the data
- Participating sites: data providers, source of user requests
- Access to data through requests, either for data as it stands or for processing through some pipelines
- Local sites control access to their data
- Sites authenticate their users
- Modern cloud architecture



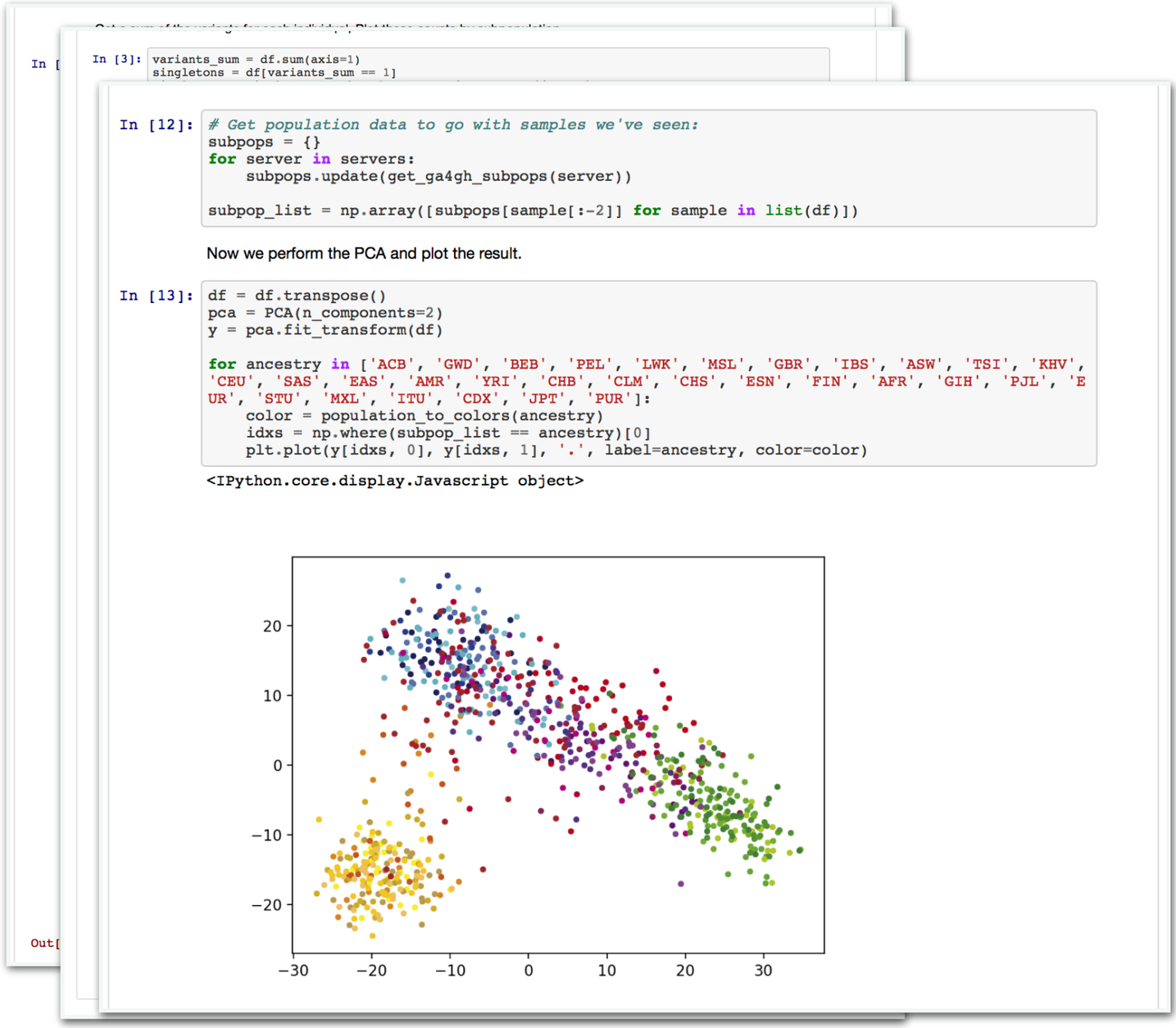
CanDIG 2017

- Demonstrated that can, e.g., re-run 1000 genomes analyses in this way
  - Randomly distribute data
  - Generate results
  - Send back only intermediate results



CanDIG 2017

- PoC Jupyter notebook access for bioinformaticians
- (Eventually BioconductoR)



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)])
```

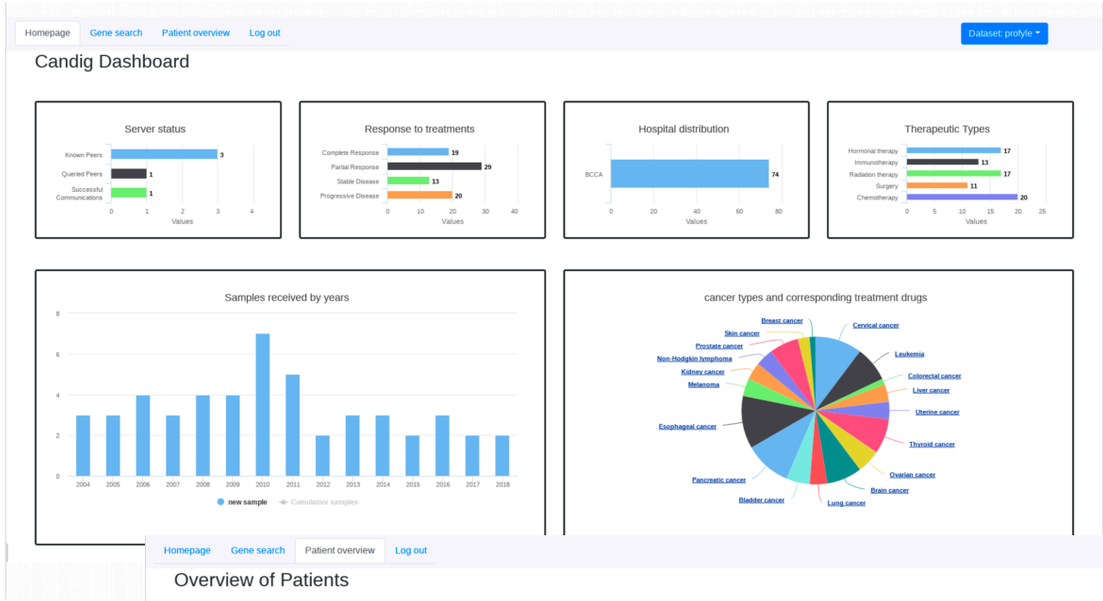
Now we perform the PCA and plot the result.

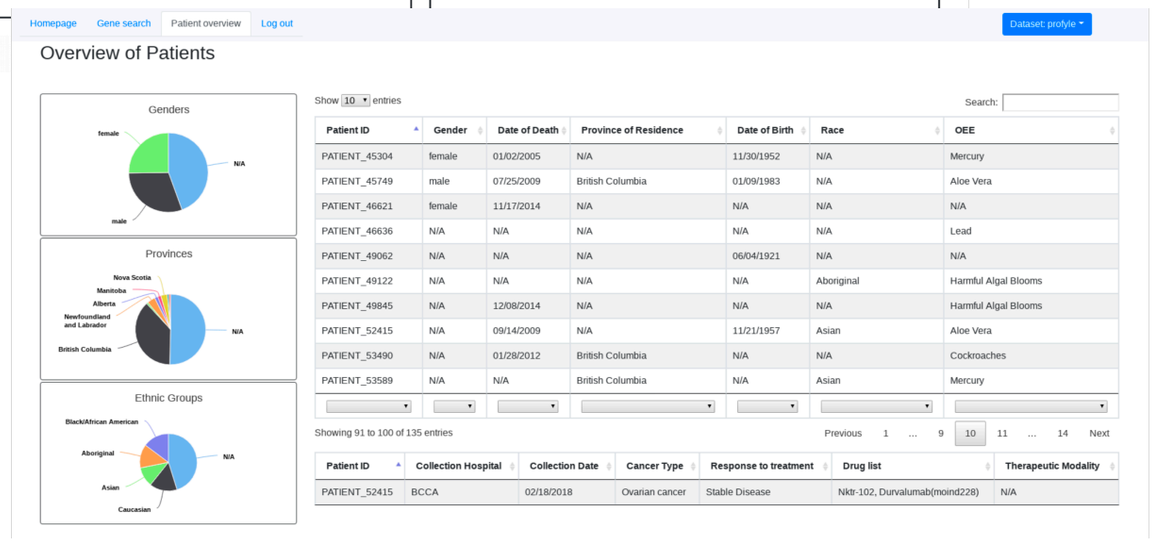
In [13]:
```
df = df.transpose()
pca = PCA(n_components=2)
y = pca.fit_transform(df)

for ancestry in ['ACB', 'GWD', 'BEB', 'PEL', 'LWK', 'MSL', 'GBR', 'IBS', 'ASW', 'TSI', 'KHV',
'CEU', 'SAS', 'EAS', 'AMR', 'YRI', 'CHB', 'CLM', 'CHS', 'ESN', 'FIN', 'AFR', 'GIH', 'PJL', 'E
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 2018

- Project dashboard
  - Support for PROFYLE and TF4CN
  - Both national-scale cancer projects
- Ability to dive deep into particular cases



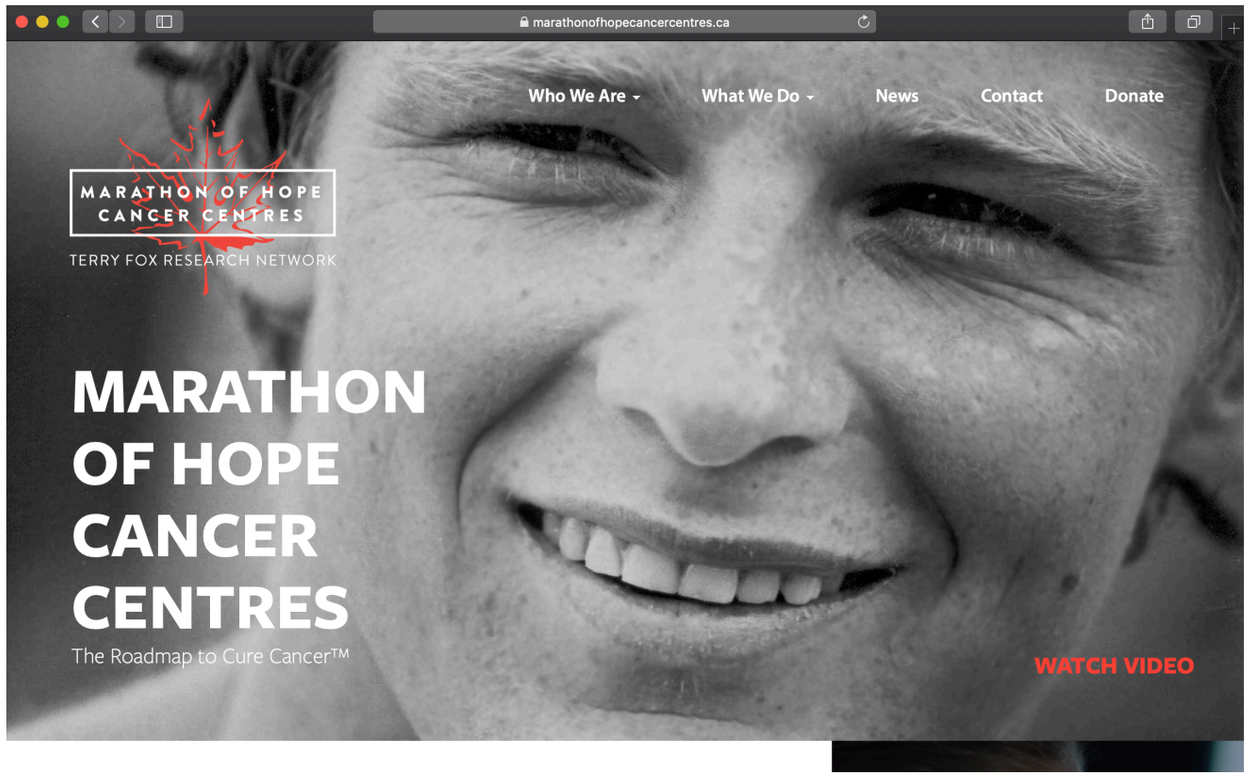


CanDIG Nationally - TF4CN

- The infrastructure project for TF4CN
- Building the data backbone for a network of comprehensive cancer centres

Marathon of Hope Cancer Centers

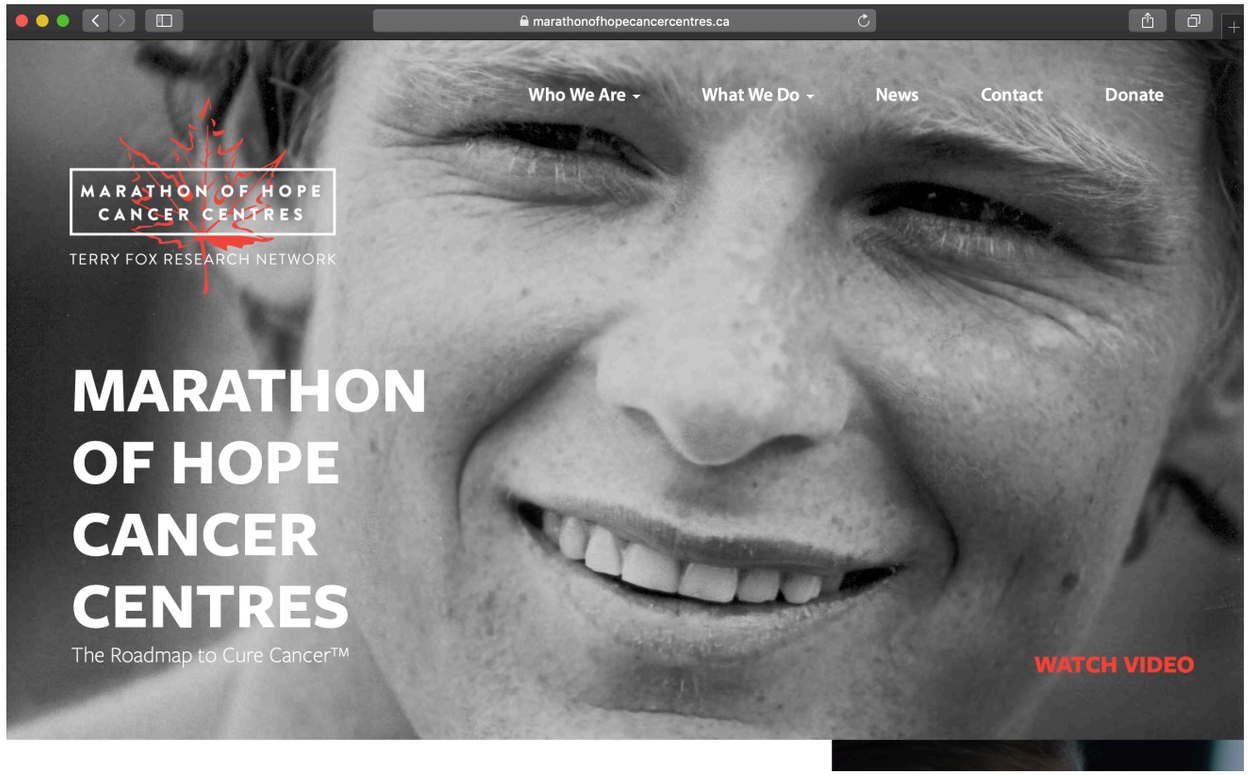
- Expanding from initial pilot (BCGSC + H4H) to include Quebec
- TFRI: aims for 10,000 patients in 5 years, 100,000 in 10
- CanDIG + H4H architecture imagined as underlying infrastructure
- Future of cancer research, research-driven care, in Canada



MarathonOfHopeCancerCentres.ca

Marathon of Hope Cancer Centers

- Expanding from initial pilot (BCGSC + H4H) to include Quebec
- TFRI: aims for 10,000 patients in 5 years, 100,000 in 10
- CanDIG + H4H architecture imagined as underlying infrastructure
- Future of cancer research, in Canada



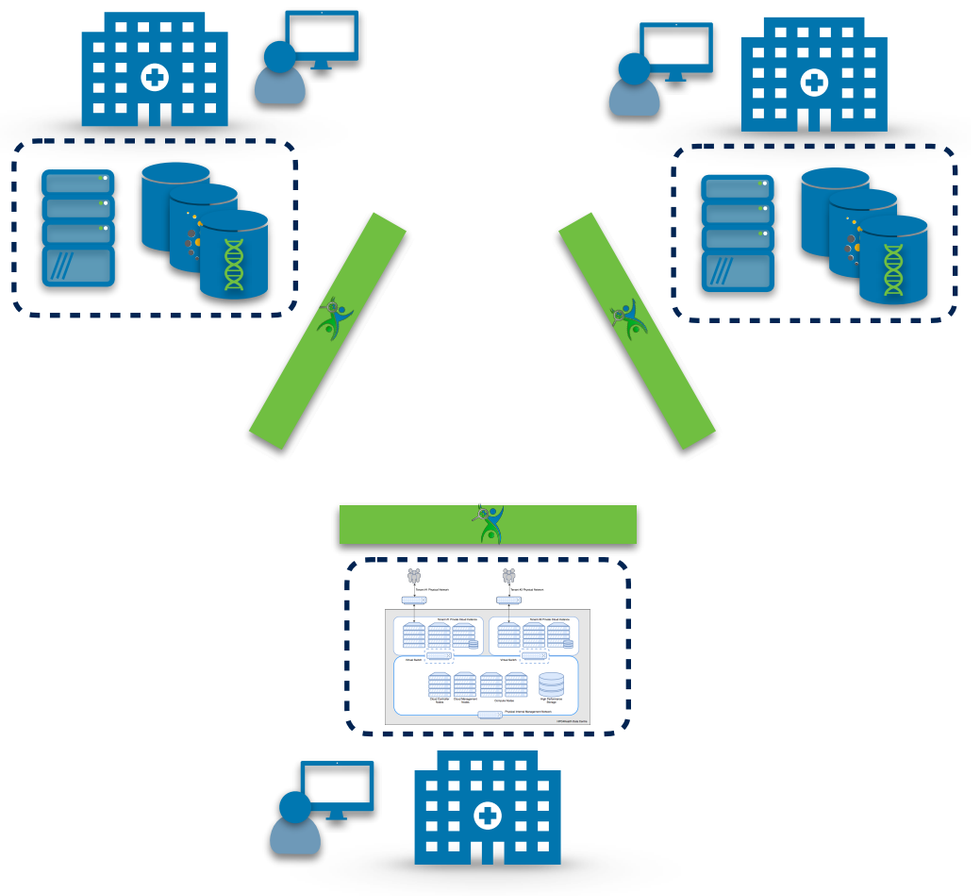
MarathonOfHopeCancerCentres.ca

Marathon of Hope Cancer Centers

- PROFYLE project
- PRecision Oncology For Young peopLE
- Goes beyond just cancer research, implications for clinical decision making
- Being supported by CanDIG + H4H

Marathon of Hope Cancer Centers: Governance

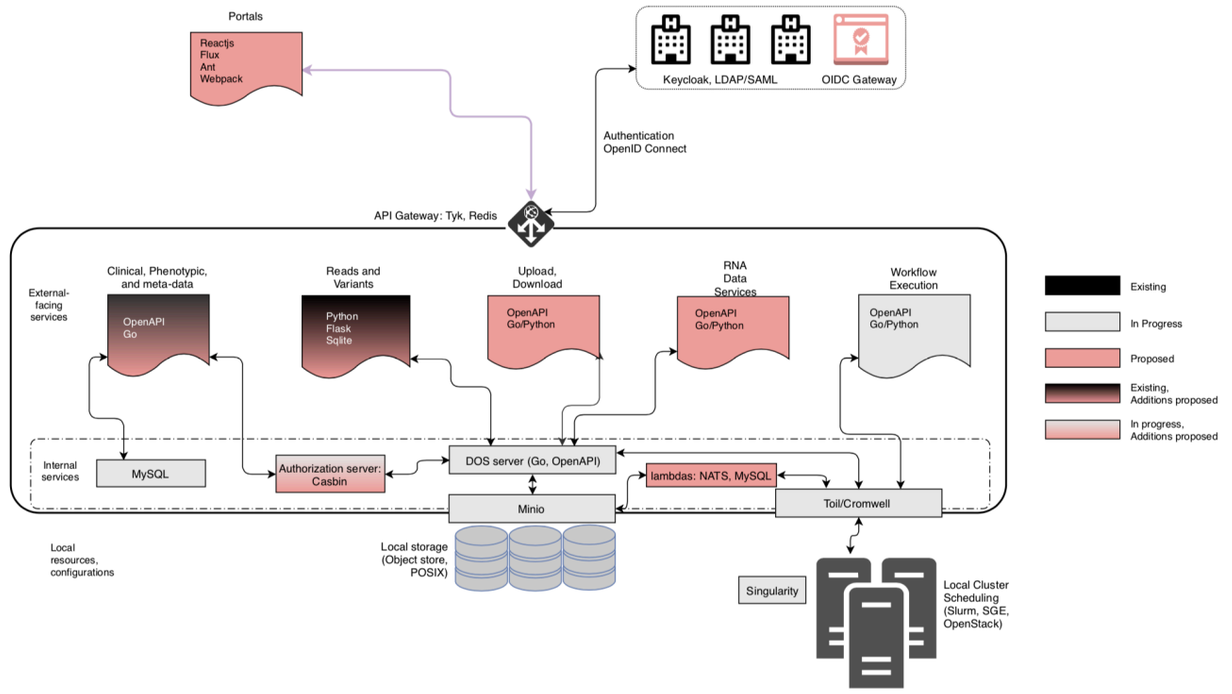
- CanDIG and HPC4Health: Key role in Data Governance discussions
- Fundamental to these national projects
- CanDIG is the substrate for data governance pilots, experiments



CanDIG Nationally - CHORD

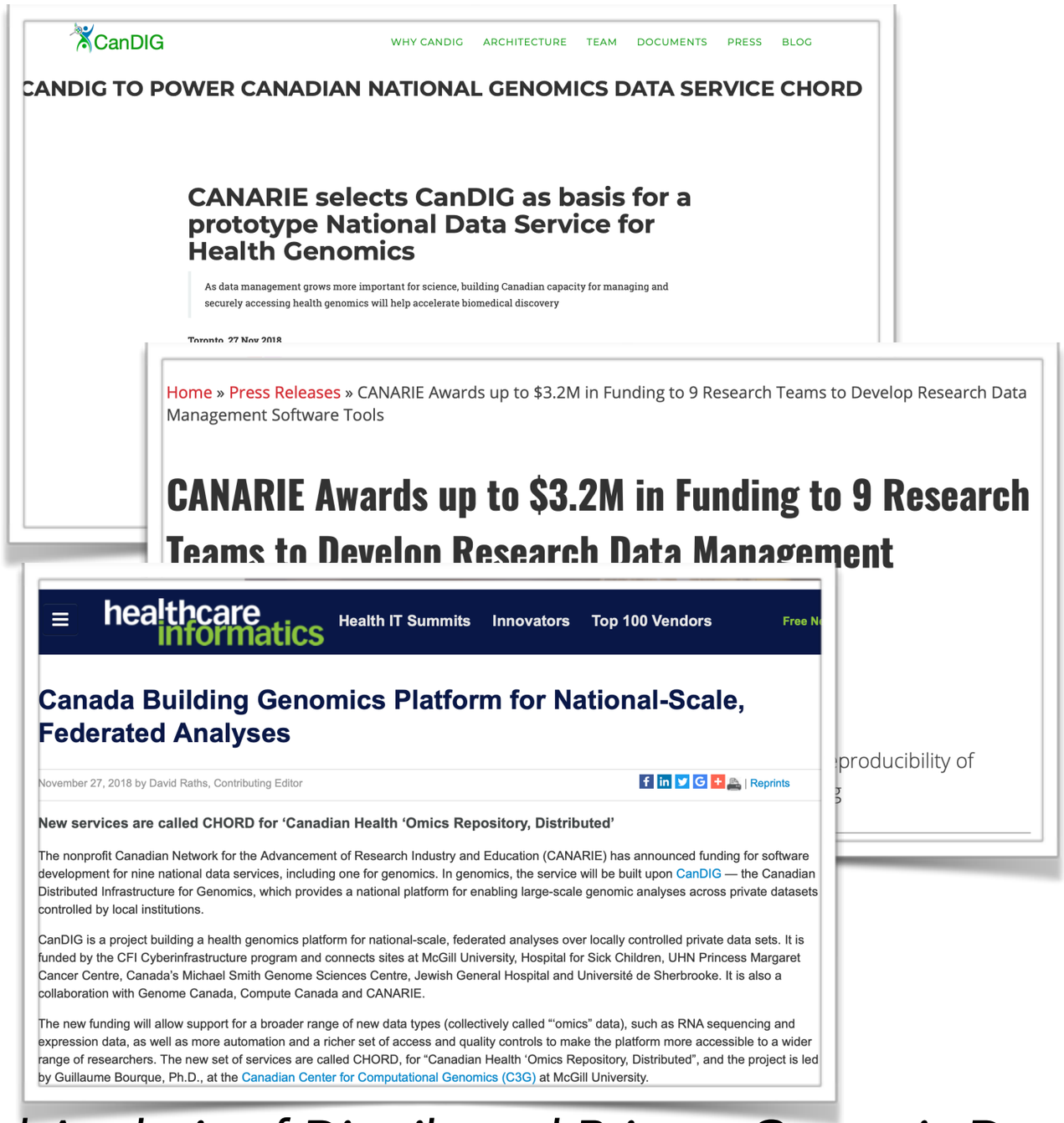
- Growing and taking a leadership role on genomics data management nationally
- CANARIE Research Data Management grant
- Build software for national data service for health genomic data

Canadian Health 'Omics Repository, Distributed



CanDIG Nationally - CHORD

- Announced earlier this week



CanDIG

National Analysis of Distributed Private Genomic Data

CanDIG Internationally - GA4GH

- Can’t do all of this alone
- Don’t want to build a single Canada-sized silo
- Work with international partners to share work, ensure interoperability, learn best practices
- GA4GH

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

GA4GH Driver Project

Australian Genomics Health Alliance

elixir

All of Us RESEARCH PROGRAM

Matchmaker Exchange

NIH NATIONAL CANCER INSTITUTE Genomic Data Commons

EUROPEAN GENOME-PHENOME ARCHIVE

BRCA CHALLENGE

ENA European Nucleotide Archive

HUMAN CELL ATLAS

monarch INITIATIVE

ClinGen Clinical Genome Resource

Genomics england

International Cancer Genome Consortium for Medicine

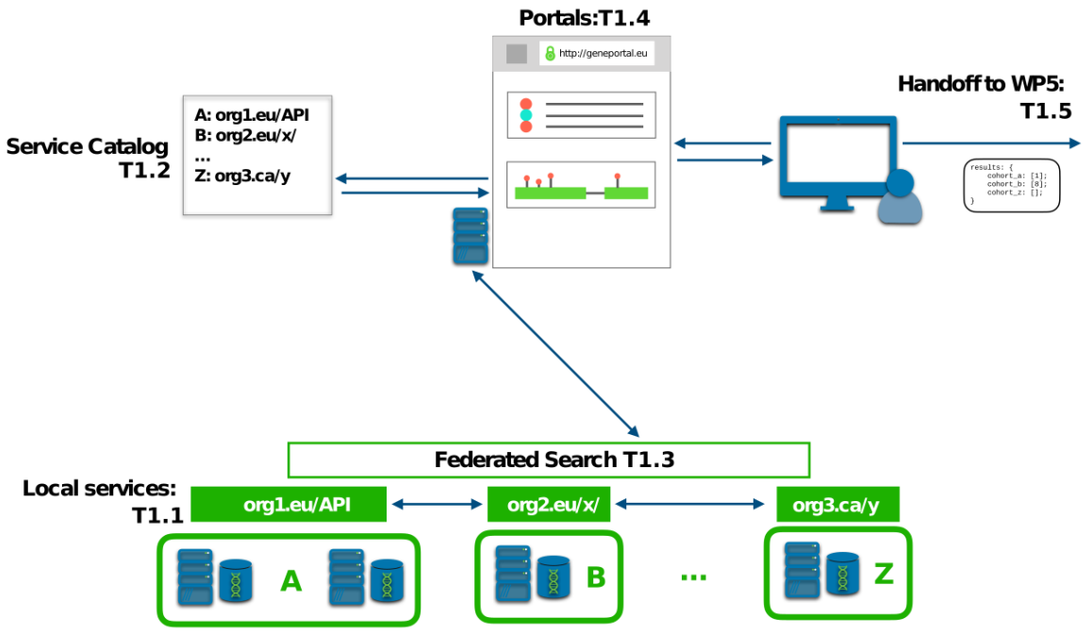
VICC Variant Interpretation for Cancer Consortium

CanDIG

National Analysis of Distributed Private Genomic Data

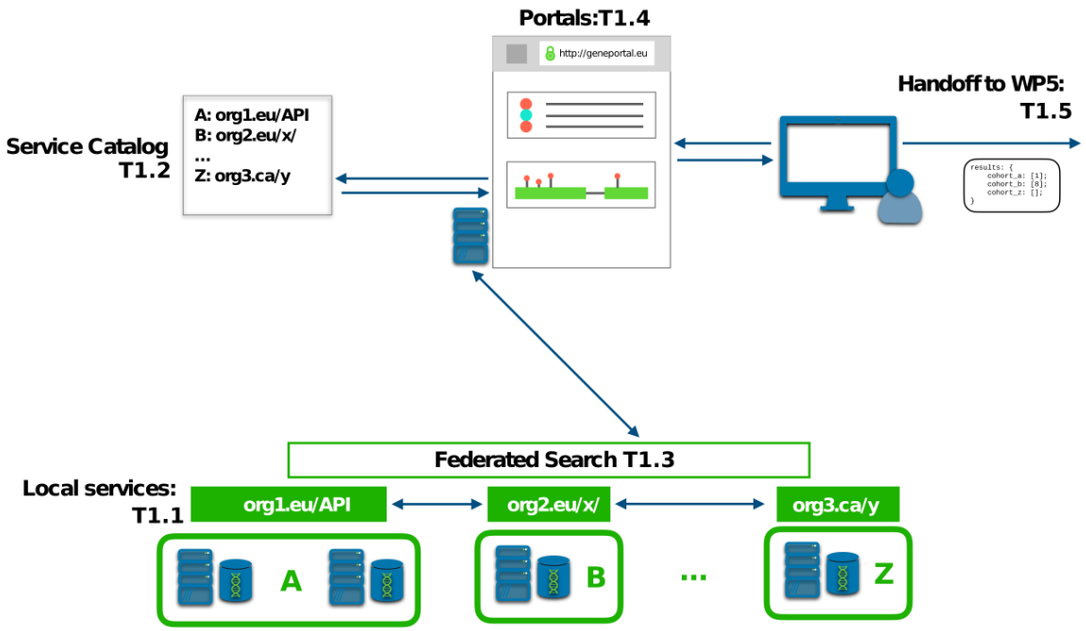
CanDIG Internationally - GA4GH

- Part of a successful EU Horizon 2020 project (not yet announced)
  - CINECA - Cohorts across Canada, EU, Canada, Africa
  - Build interoperability between peers CanDIG, ELIXIR, EGA
  - Lead work package 1 - federated queries



CanDIG Internationally - CINECA

- Canada's need for data federation - we're building tools everyone else will have to have
- And now we're teaching them how to use them
- Key role in data governance internationally



HPC4Health

- Provincially: with CO, getting groups working together w/ HPC4Health
  - Expanding the architecture model
  - Supporting key biology and bioinformatics projects - cancer, rare disease
- Nationally
  - HPC4Health used as a model
  - CanDIG as a national genomics data platform
  - Data governance for health data in Canada
  - Building capacity, skills, and HQP via training in foundational technologies like genomics, bioinformatics
- Internationally
  - CanDIG and GA4GH
  - CanDIG and ELIXIR, EGA: leading international federation, governance
- H4H, with its enormous capacity, and overlapping pools of expertise in cloud, bioinformatics, rare disease, and cancer, makes this possible