Scalable methods for data analysis and sharing in genomics

Guillaume Bourque

Dept. Human Genetics, McGill University

McGill Genome Center

Canadian Center for Computational Genomics (C3G)

@guilbourque

RDC Webinar Series

26 Novembre, 2020

Canadian Center for Computational Genomics (C3G)

*C3G is a Genome Canada Genomics Technology Platform (GTP) launched in 2015 that provides **bioinformatics analysis, HPC services and software solutions** to the life sciences research community.*

Guillaume Bourque Ph.D.
Director, C3G Montreal

McGill

Michael Brudno Ph.D.
Director, C3G Toronto

SickKids

GenomeCanada

Large genomic projects at C3G

Cancer:

PROFYLE/Signature

MCC/QCC

DHDP (w/ Imagia)

Marathon of Hope

...

Infection/Immunity:

CoVSeQ/CanCOGeN

BQC19

MI4

Recodid

...

Epigenomics/genomics:

IHEC

4DN (NIH)

LSARP (Jabado)

CGEn

BRIDGET

Secure Cloud (w/ CQ)...

Software/Others:

CanDIG

GenAP

EpiShare

ClinDIG

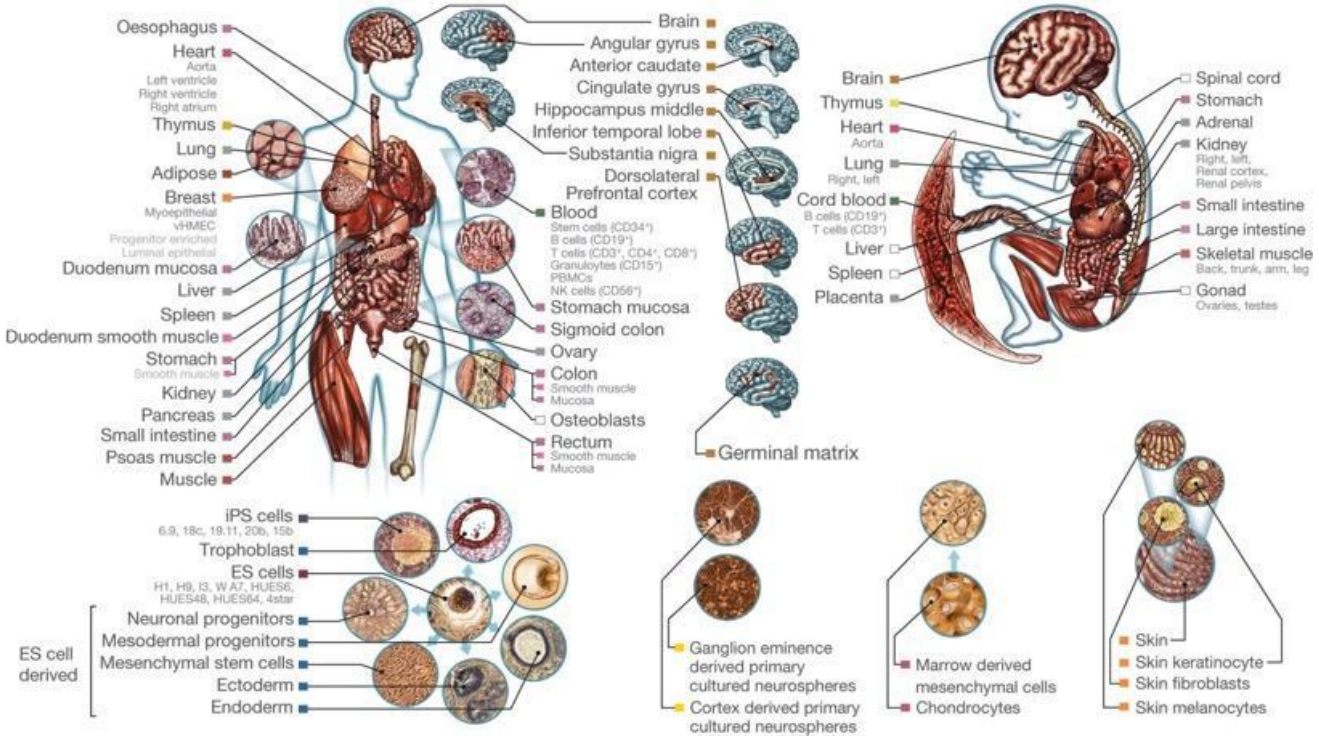
CHORD

CINECA

Covid19 Resource Canada

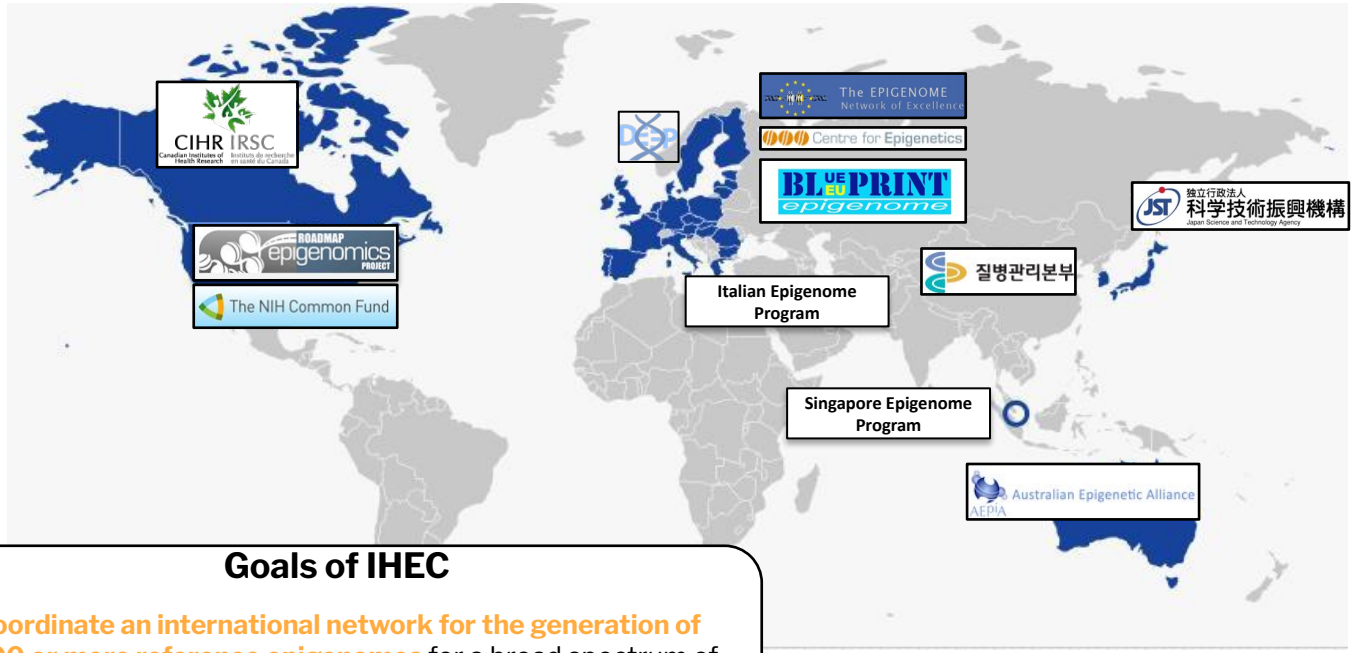
...

One genome… Many epigenomes



Roadmap Epigenomics Consortium *et al. Nature* **518**, 317-330 (2015) doi:10.1038/nature14248

IHEC
International Human Epigenome Consortium



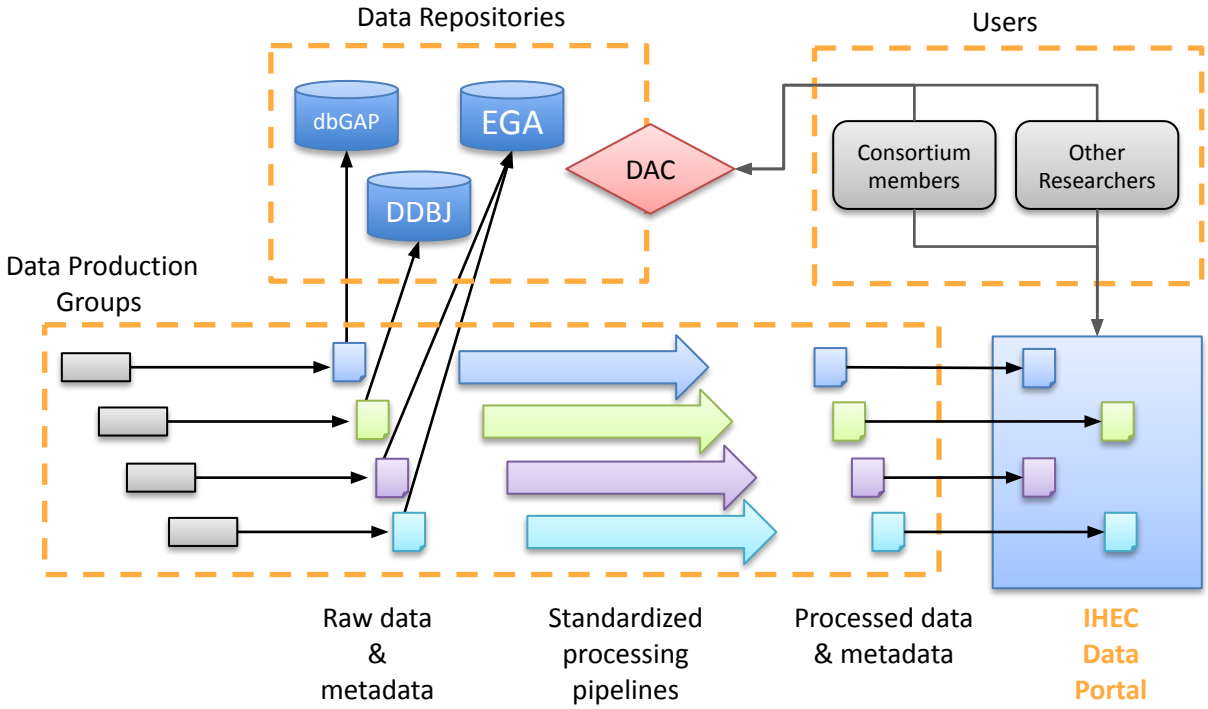
Goals of IHEC

Coordinate an international network for the generation of 1000 or more reference epigenomes for a broad spectrum of human cell types and a wide range of developmental stages, laying the foundation to study the epigenetic mechanisms of human diseases.

IHEC-generated data

- Data generated by IHEC is aimed to be available for everyone's own research, following the consortium's Open Science mission.
- Human data produced by most IHEC members has different levels of access:
  - Controlled access data
    - Potentially personally identifiable
    - Access is limited to protect study participants
    - Archived at repositories such as EGA and DDBJ
    - Obtaining the data involves applying for access
  - Fully public data
    - Meta-information considered non-personally identifiable
    - Processed data (analysis pipelines output)

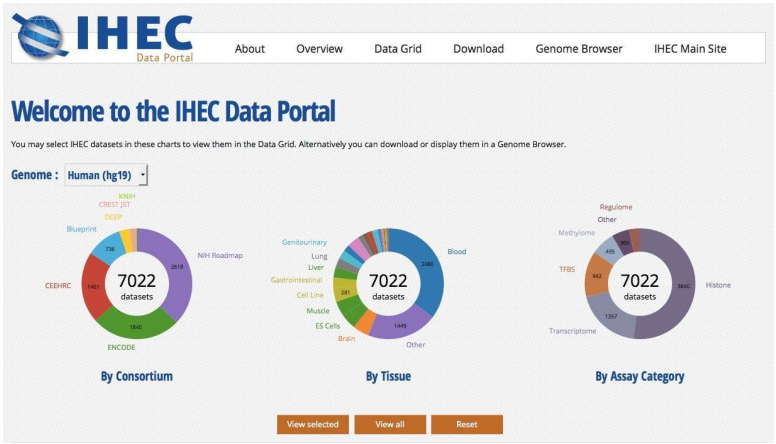
IHEC data integration and sharing strategy



IHEC Data Portal

- Launched in June 2014, http://epigenomesportal.ca/ihec
- Includes:
  - Over **10,800** human epigenomic datasets (hg19 and hg38)
  - Over **280** mouse and primate datasets
  - Over **>290** full reference epigenomes
  - Data from: Blueprint, CEEHRC, CREST, DEEP, ENCODE, KNIH, NIH Roadmap

David
Bujold



Bujold et al.
Cell Systems
2016

The epiATLAS project

- IHEC Integrative Analysis workgroup initiative
- Initial goals:
  a. establish an analyst-friendly and widely sharable compendium of **quality-controlled, consistently processed, reference epigenomic maps** from all areas of IHEC.
     - Prepare a gold standard dataset
     - Develop ways to ease access to the raw data
     - Improve the overall experience of accessing and analysing IHEC data
  b. initiate and support numerous hypothesis-driven as well as exploratory analysis projects based on the IHEC epigenome compendium.
  c. coordinate the publication of the resulting analyses in the form of a flagship and companion papers.

Challenges of current model

There are multiple challenges bound to using controlled access data, even before getting to the bulk of the analysis!

- Obtaining access
  - Application to a Data Access Committee (DAC)
- Downloading
  - Getting the data from a controlled access repository
- Comparing datasets across projects
  - Metadata is often hard to collate across projects
- Analysing the data
  - Heavy use of resources

Accessing raw data

	Korea Epigenome Consortium	Blueprint Consortium	McGill EMC	BC Cancer Agency	DEEP (Germany)	Singapore Epigenome Project	CREST (Japan)
a) Acknowledgements	■	■	■	■	■	■	■
b) Application renewal			■		■		
c) Confidentiality	■	■	■	■	■	■	■
d) Data destruction	■	■	■	■	■	■	
e) Data usage	■	■	■	■	■	■	■
f) Ethics review			■	■	■		■
g) Evidence of PI's experience			■	■	■		
h) Intellectual property	■	■	■	■	■	■	
i) IT security requirements			■				
j) Naming of jurisdiction for disputes			■	■		■	
k) Liability exclusions & warranty limitations	■	■	■	■	■	■	
l) Process in the face of access requests	■	■	■	■			
m) Publication embargos	■	■	■		■		
n) Rapid publication				■			
o) Progress report	■	■		■		■	
p) References to laws			■	■		■	
q) References to policies	■	■	■		■		■
r) Rules on student access to data			■	■	■	■	

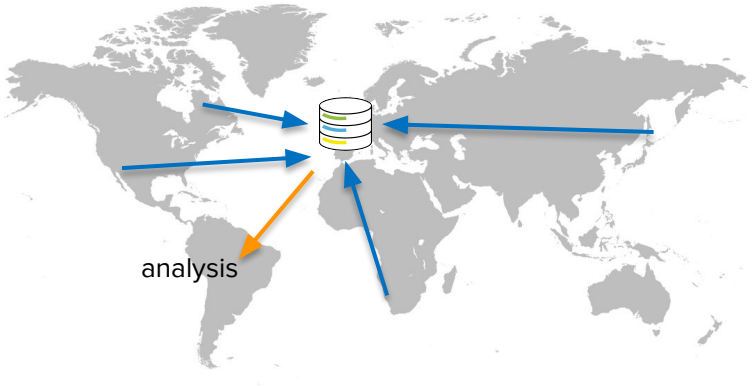
Fig. 1 Terms and Conditions identified in IHEC DAAs.

Saulnier et al. *Scientific Data*, 2019

Federation

Global Alliance for Genomics & Health

Open Research Data

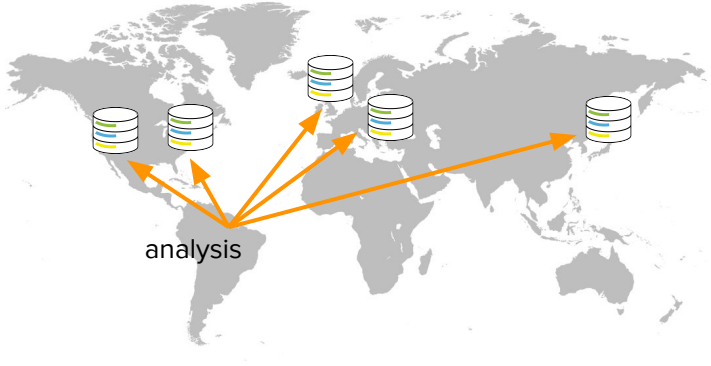


Aggregate data globally

Download, analyze locally

Continues for basic research

Healthcare Data with Research Use



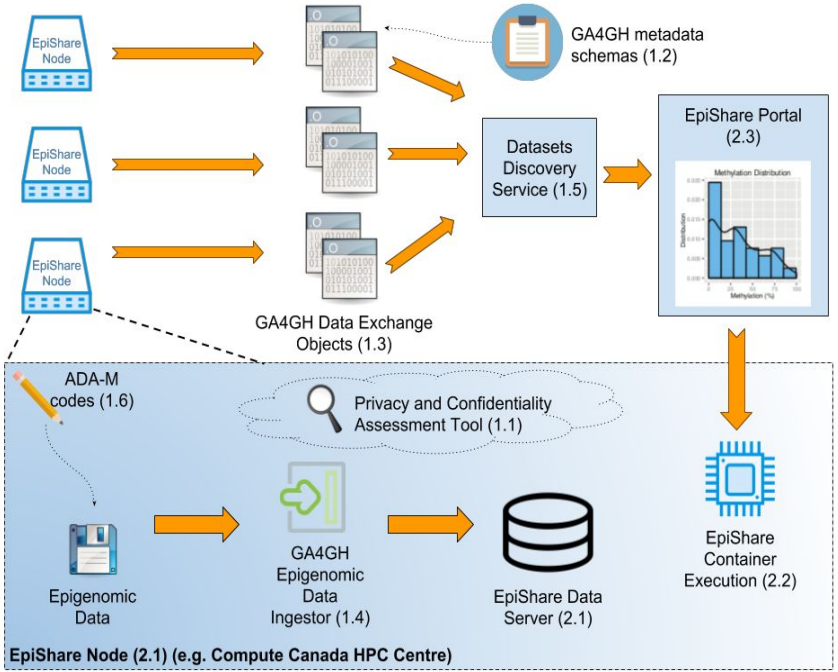
Aggregate data locally (via VMs)

Collate analyses

New approach for research and healthcare

EpiShare

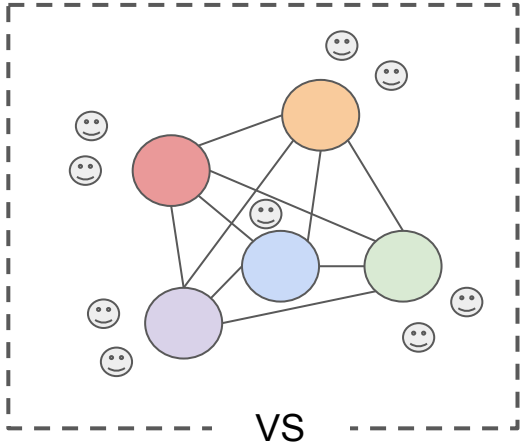
- Genome Canada funded project (2018-2021)
- Aims at extending the GA4GH APIs, etc. for epigenomic data
- Will create a resource to make data more easily discoverable
- Will enable the launch of multi-omics analyses on controlled-access datasets at their storage location
- Now a GA4GH Driver Project



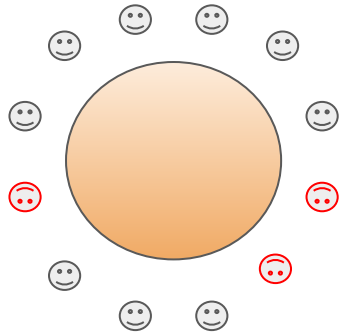
Global Alliance for Genomics & Health

Federated platform

- Enables **fine-grained control over data** while participating in data sharing
  - Shared normalization/QC pipelines
- **Resource allocations** from e.g. Compute Canada can be used **at a node level**
- Federation **maintain locality** while optionally **sharing in a network**
  - Some data may not be allowed to be stored in another jurisdiction



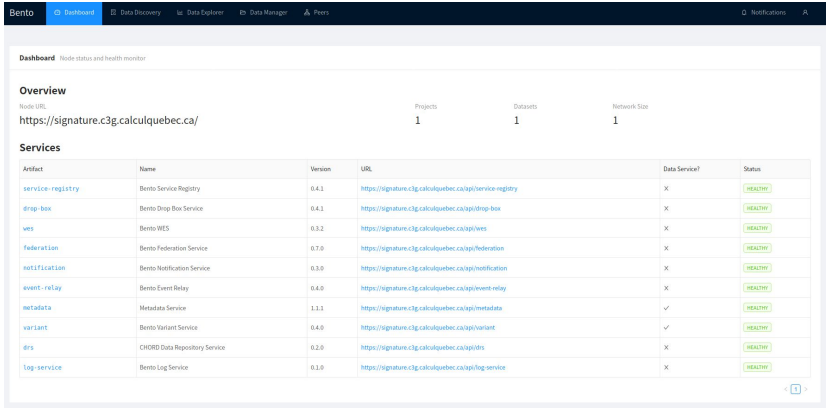
VS



CanDIG/Bento platform

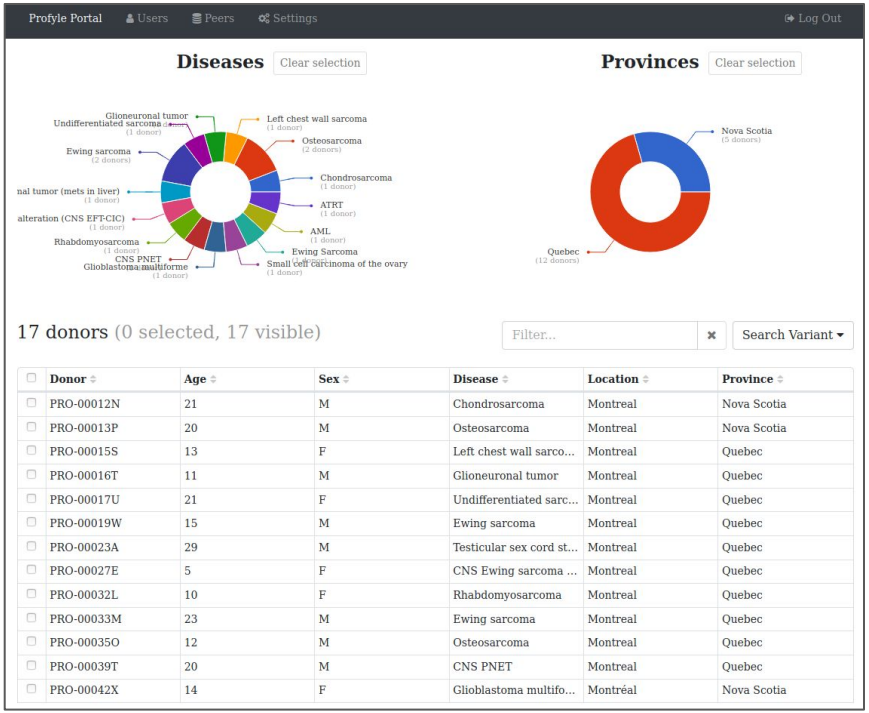
- Released as a self-hosted software infrastructure
    - SIGNATURE instance deployed at the Calcul Quebec Secure Cloud

- Data types-specific microservices include:
    - Clinical/phenotypic metadata
    - Genomic variants (VCF)
    - RNA-Seq expression data (in progress)



CanDIG demonstration project

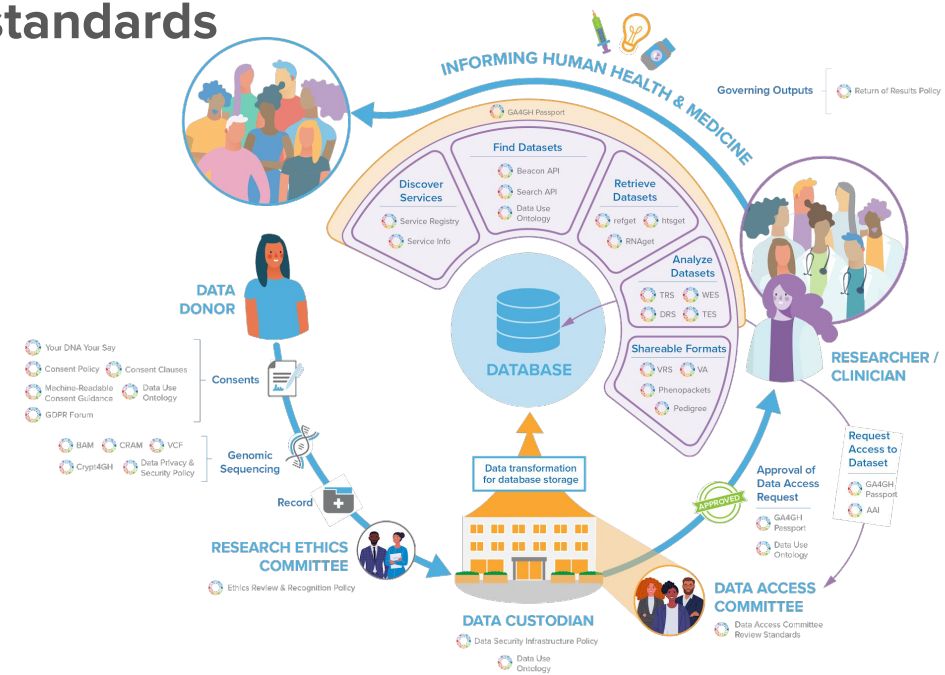
Visualization tools and a overview of the project in the dashboard



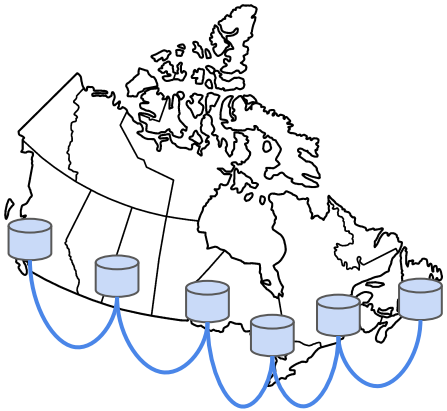
Donor	Age	Sex	Disease	Location	Province
PRO-00012N	21	M	Chondrosarcoma	Montreal	Nova Scotia
PRO-00013P	20	M	Osteosarcoma	Montreal	Nova Scotia
PRO-00015S	13	F	Left chest wall sarco...	Montreal	Quebec
PRO-00016T	11	M	Glioneuronal tumor	Montreal	Quebec
PRO-00017U	21	F	Undifferentiated sarc...	Montreal	Quebec
PRO-00019W	15	M	Ewing sarcoma	Montreal	Quebec
PRO-00023A	29	M	Testicular sex cord st...	Montreal	Quebec
PRO-00027E	5	F	CNS Ewing sarcoma ...	Montreal	Quebec
PRO-00032L	10	F	Rhabdomyosarcoma	Montreal	Quebec
PRO-00033M	23	M	Ewing sarcoma	Montreal	Quebec
PRO-00035O	12	M	Osteosarcoma	Montreal	Quebec
PRO-00039T	20	M	CNS PNET	Montreal	Quebec
PRO-00042X	14	F	Glioblastoma multifo...	Montréal	Nova Scotia

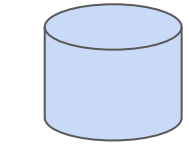
Global Alliance
for Genomics & Health

GA4GH sets of standards



Data Federation using GA4GH in Canada





International Projects and Consortia

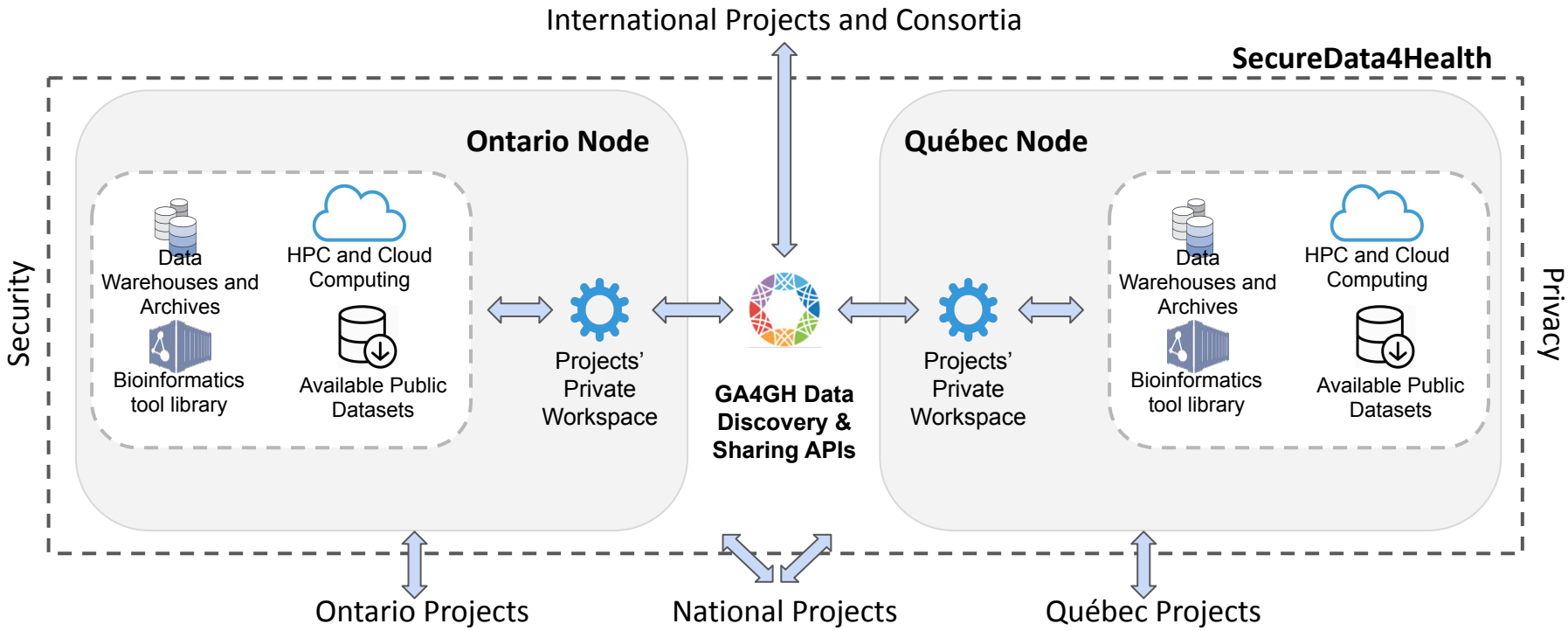
GA4GH Standards

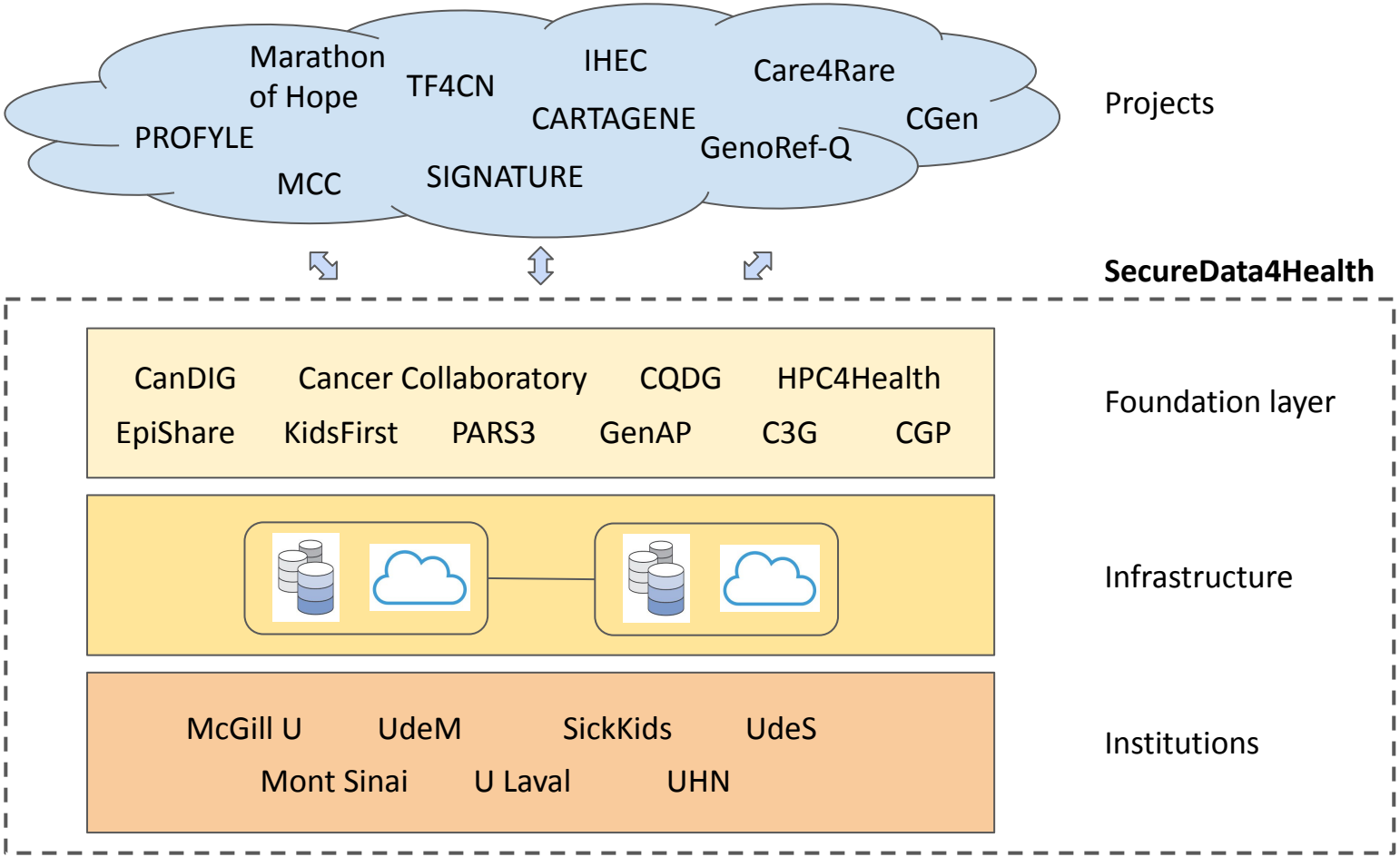
SecureData4Health project

- A secure cloud infrastructure for analysis and sharing of genomic and health data
- 20M proposal submitted to CFI's Innovation fund 2020
- Team:

Guillaume Bourque	McGill University
Vincent Ferretti	Université de Montréal
Michael Brudno	University Health Network
Anne-Claude Gingras	Sinai Health System
Anna Goldenberg	The Hospital for Sick Children
Benjamin Haibe-Kains	University Health Network
Julie Hussin	Université de Montréal
Pierre-Étienne Jacques	Université de Sherbrooke
Bartha Knoppers	McGill University
Jacques Simard	Université Laval

SecureData4Health project




Marathon of Hope
TF4CN
IHEC
Care4Rare
PROFYLE
CARTAGENE
CGen
GenoRef-Q
MCC
SIGNATURE
Projects
SecureData4Health
CanDIG
Cancer Collaboratory
CQDG
HPC4Health
EpiShare
KidsFirst
PARS3
GenAP
C3G
CGP
Foundation layer
Infrastructure
McGill U
UdeM
SickKids
UdeS
Mont Sinai
U Laval
UHN
Institutions

Acknowledgements

Analysis team
Francois Lefebvre
Pascale Marquis
Gary Lévesque
Emmanuel Gonzalez
Senthil Duraikannu
Alain Pacis
Robert Syme

Development team
Mathieu Bourgey
Edouard Henrion
Robert Eveleigh
Rola Dali
Hector Galvez Lopez
Paul Strenenowich
Pierre-Olivier Quirion
Pubudu Nawarathna
Ulysse Fortier-Gauthier

Data team
David Bujold
Romain Gregoire
David Lougheed
David Anderson
Ksenia Zaytseva
David Brownlee
Simon Chenard
Adrielle Houweling

Lab
Patricia Goerner-Potvin
Maxime Caron
Jeffrey Hyacinthe
Cristian Groza
Qinwei Zhuang
Mio Shibata
Xun Chen
Lindsay Dayton
Ksenia Egorova
Arber Kacollja
Rachade Hmamouchi

The group is recruiting!
guil.bourque@mcgill.ca

@guilbourque

McGill · Calcul Québec · canarie · compute • calcul CANADA · CIHR IRSC Canadian Institutes of Health Research / Instituts de recherche en santé du Canada · GenomeCanada · TOP GLOBAL UNIVERSITY JAPAN · Kyoto University Founded 1897 · ASHBi WPI Kyoto University