Epigenomic data discovery with the IHEC Data Portal

David Bujold
Canadian Centre for Computational Genomics
2019-01-24

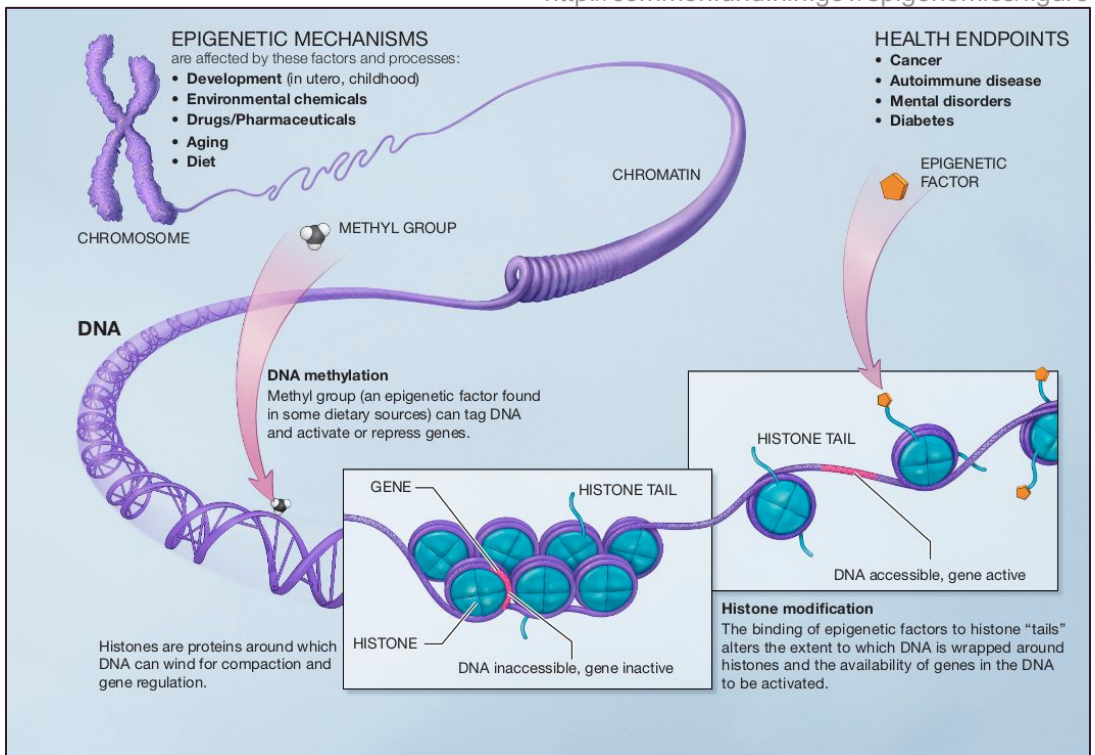
Outline

- Today's focus:
  - Our work so far to make epigenomic data Findable
  - Challenges ahead and strategies
- To be Findable: (https://www.force11.org/group/fairgroup/fairprinciples)
  - F1. (meta)data are assigned a globally unique and eternally persistent identifier.
  - F2. data are described with rich metadata.
  - F3. (meta)data are registered or indexed in a searchable resource.
  - F4. metadata specify the data identifier.

What is epigenomics?

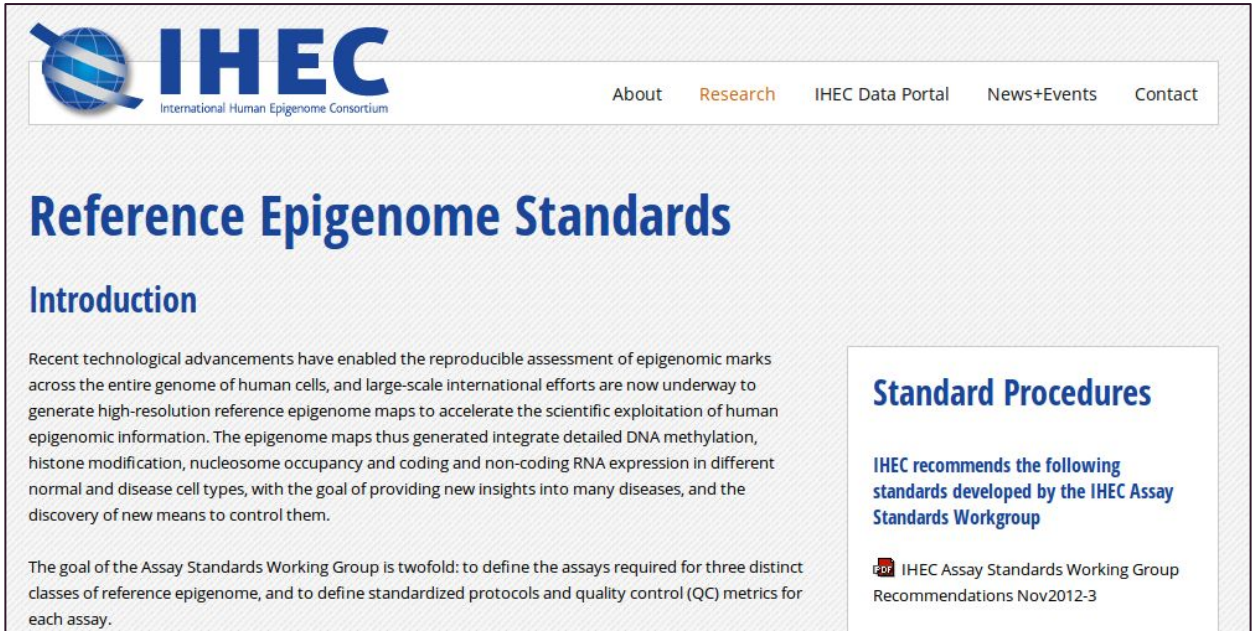
- Study of epigenetic modifications on genetic material of cells
  - Reversible modifications on cell DNA or histones
  - Affect gene expression without altering DNA sequence
  - Partly inherited, partly imputable to environment

http://commonfund.nih.gov/epigenomics/figure



What is IHEC?

- International effort with several funding agencies
  - Workgroups develop standards and toolboxes (assays, data ecosystem, integrative analysis, ethics…)
- Goal: Providing standardized reference epigenomes for a variety of normal and disease tissues
- Canadian effort, funded by CEEHRC (CIHR)



IHEC
International Human Epigenome Consortium

About | Research | IHEC Data Portal | News+Events | Contact

Reference Epigenome Standards

Introduction

Recent technological advancements have enabled the reproducible assessment of epigenomic marks across the entire genome of human cells, and large-scale international efforts are now underway to generate high-resolution reference epigenome maps to accelerate the scientific exploitation of human epigenomic information. The epigenome maps thus generated integrate detailed DNA methylation, histone modification, nucleosome occupancy and coding and non-coding RNA expression in different normal and disease cell types, with the goal of providing new insights into many diseases, and the discovery of new means to control them.

The goal of the Assay Standards Working Group is twofold: to define the assays required for three distinct classes of reference epigenome, and to define standardized protocols and quality control (QC) metrics for each assay.

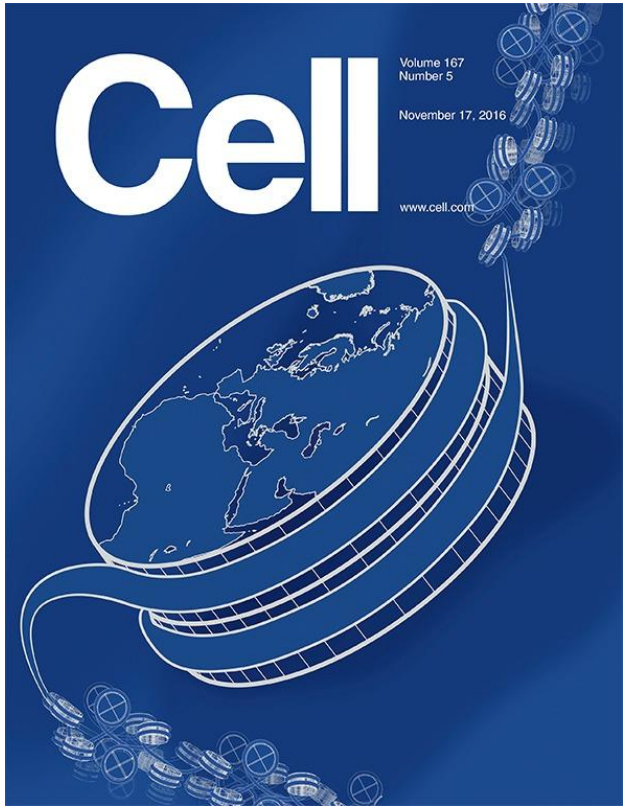
Standard Procedures

IHEC recommends the following standards developed by the IHEC Assay Standards Workgroup

IHEC Assay Standards Working Group Recommendations Nov2012-3

IHEC Full Reference Epigenome

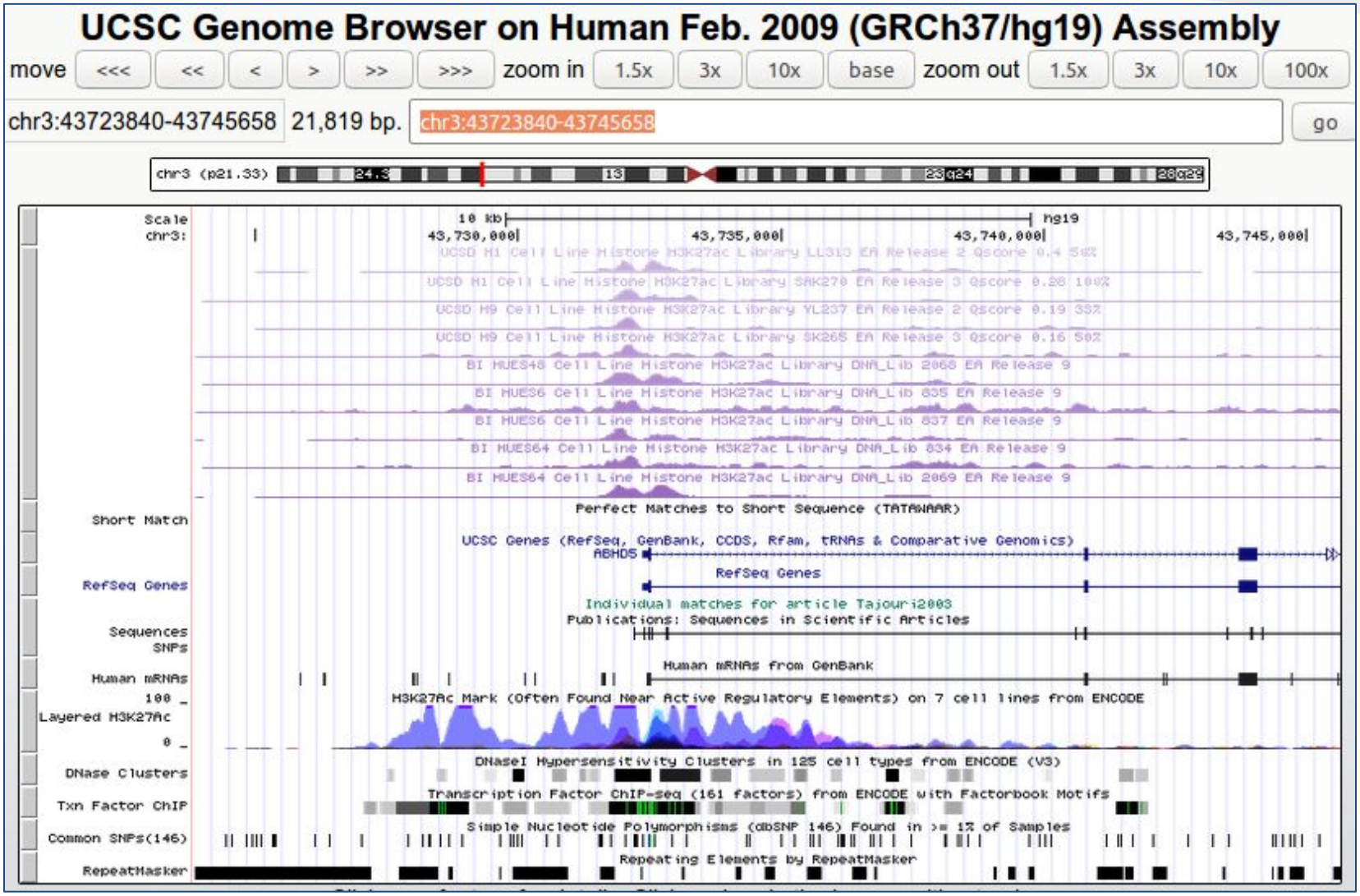
- Methylome: Whole-Genome Bisulfite Sequencing
- Transcriptome: RNA-Seq
- Histone tails modifications:
  - H3K4me1
  - H3K4me3
  - H3K36me3
  - H3K27ac
  - H3K9me3
  - H3K27me3



Generated datasets

- Datasets are made available for everyone's own research
- Human data offered by such consortia falls in one of two categories:
  - Controlled access data
    - Raw data from sequencers
    - Clinical/sensitive information such as phenotypes
    - Archived at repositories such as EGA and dbGaP
  - Public data (Freely downloadable)
    - Annotation tracks, to use in tools such as UCSC Genome Browser, Ensembl and IGV.
    - Some donor, sample and library metadata

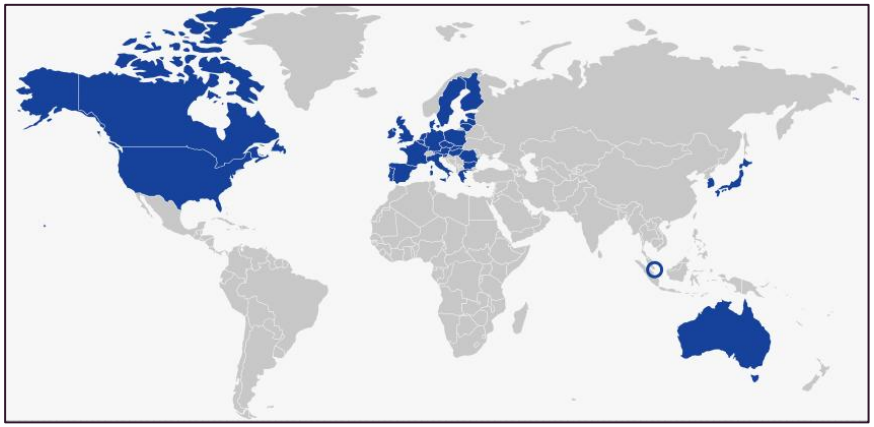
Annotation tracks in the UCSC Genome Browser



IHEC Data Portal

- Goal of the Portal: Integrate epigenomic datasets produced and released by IHEC, and making them Findable

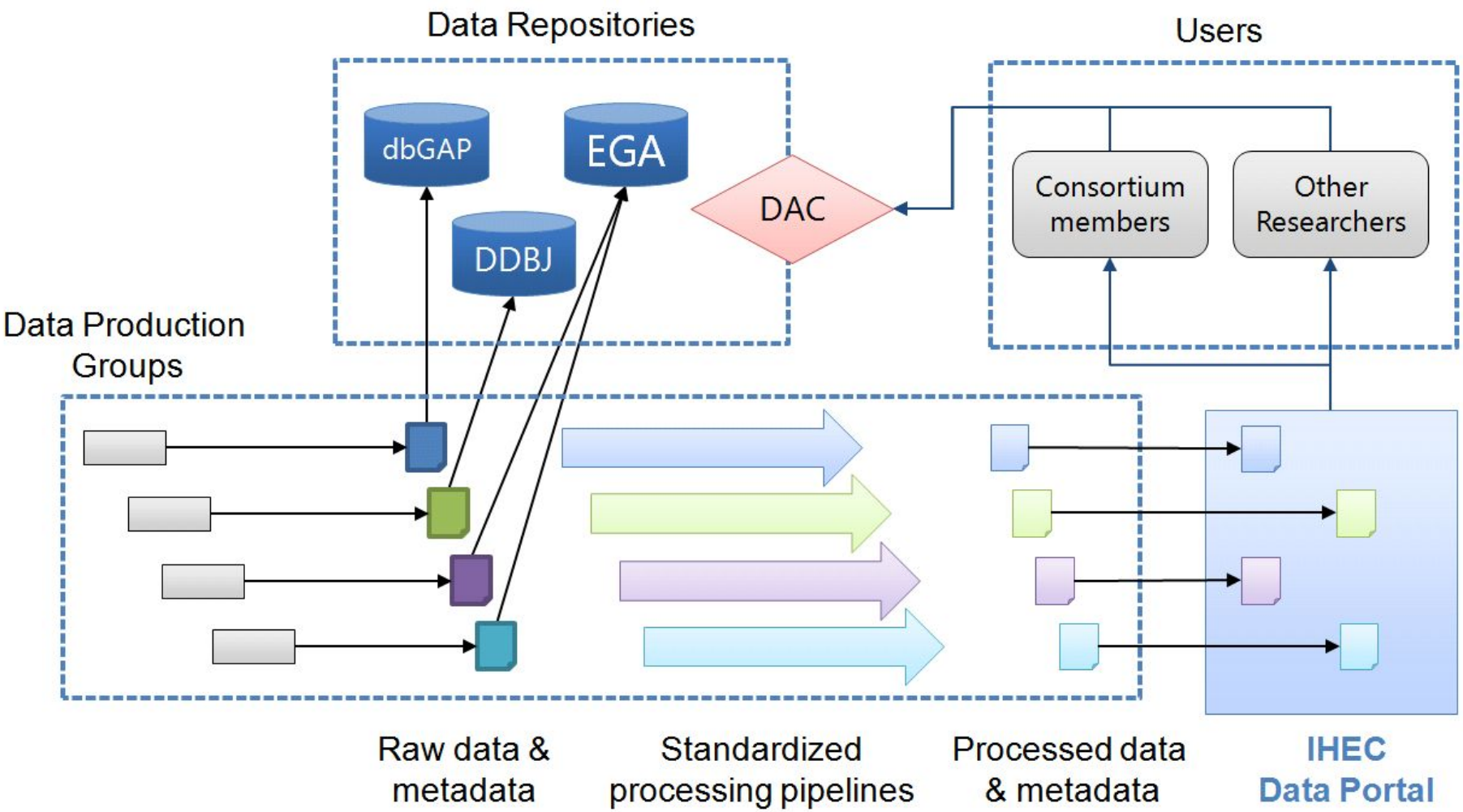
- Data currently available from these members:
  - AMED-CREST (Japan)
  - Blueprint (Europe)
  - CEEHRC (Canada)
  - DEEP (Germany)
  - ENCODE (USA)
  - GIS (Singapore)
  - KNIH (South Korea)
  - Roadmap (USA)



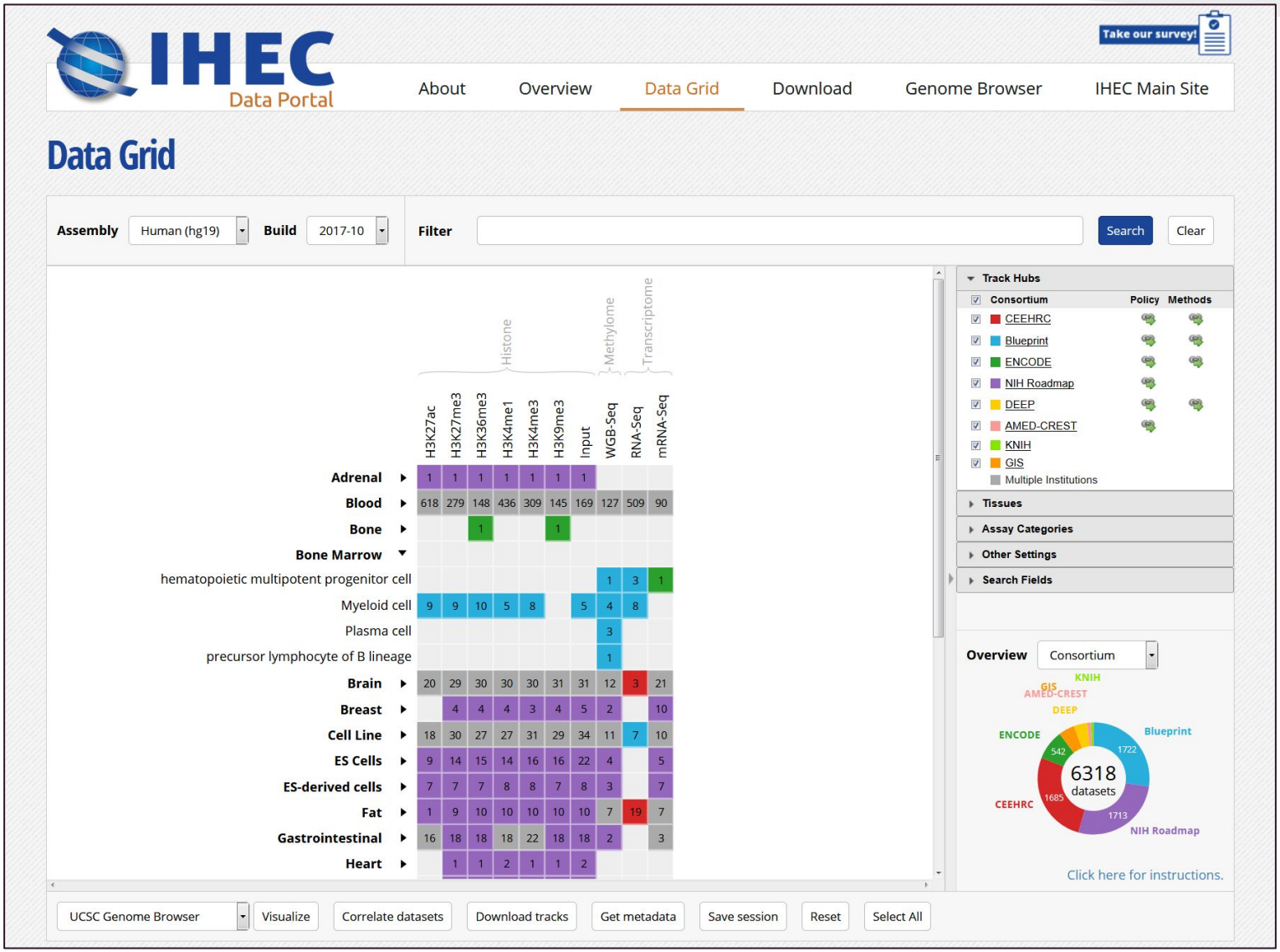
IHEC Data Portal

- Launched in June 2014 (epigenomesportal.ca/ihec)
- Includes:
  - Over 10,800 human epigenomic datasets (hg19 and hg38)
  - Over 280 mouse and primate datasets
  - Over >290 full reference epigenomes
- Centralizes the storage of public access tracks produced within IHEC, and points users to controlled access repositories to obtain the raw data
- Proposes tools for IHEC data/metadata navigation, visualization, sharing and analysis
- Connects to multiple external data visualization and analysis resources
- Allows permanent sessions creation to be shared with collaborators or for publications

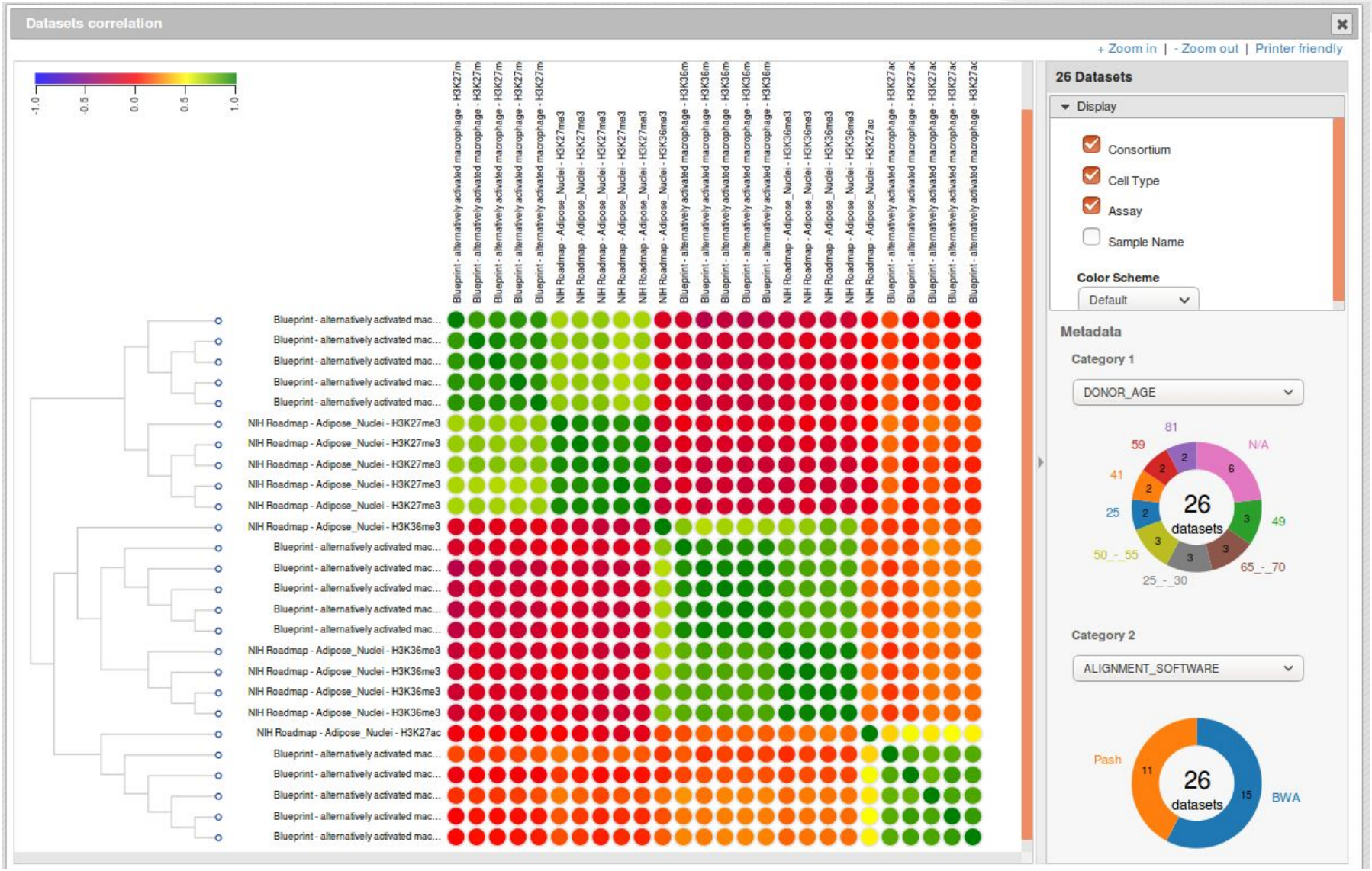
IHEC data integration and sharing strategy



IHEC Data Portal

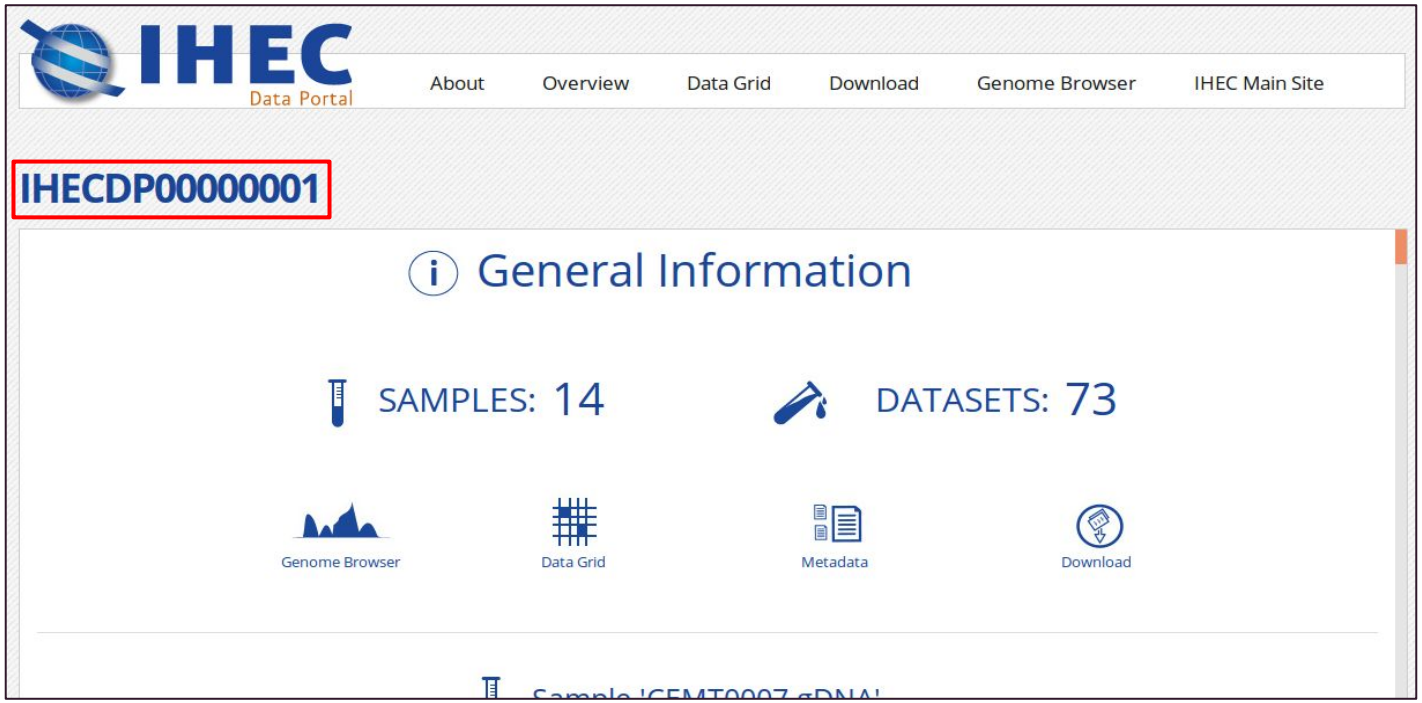


IHEC Data Portal



Permanent sessions

- An accession number can be used to restore grid selection and filtering options
- Improves shareability and enables citations



Session report

- Generation of reports on session data content

IHECDP00000001

Sample 'CEMT0007.gDNA'

Sample Metadata

biomaterial_type	Cell Line
differentiation_stage	NA
disease	Mammary gland-Breast Fibrocystic Disease
disease_ontology_uri	
line	MCF10A
lineage	mammary gland epithelia
medium	Growth media: DMEM/F12 +5% horse serum (it also includes FGF (20 ug/ml), Hydrocortizone (0.5 mg/ml), Cholera Toxin (10 ug/ml), insulin (1 mg/ml), Pen/Strep 1%)
molecule	genomic DNA
sample_id	CEMT0007
sample_ontology_uri	http://www.ebi.ac.uk/efo/EFO_0001200

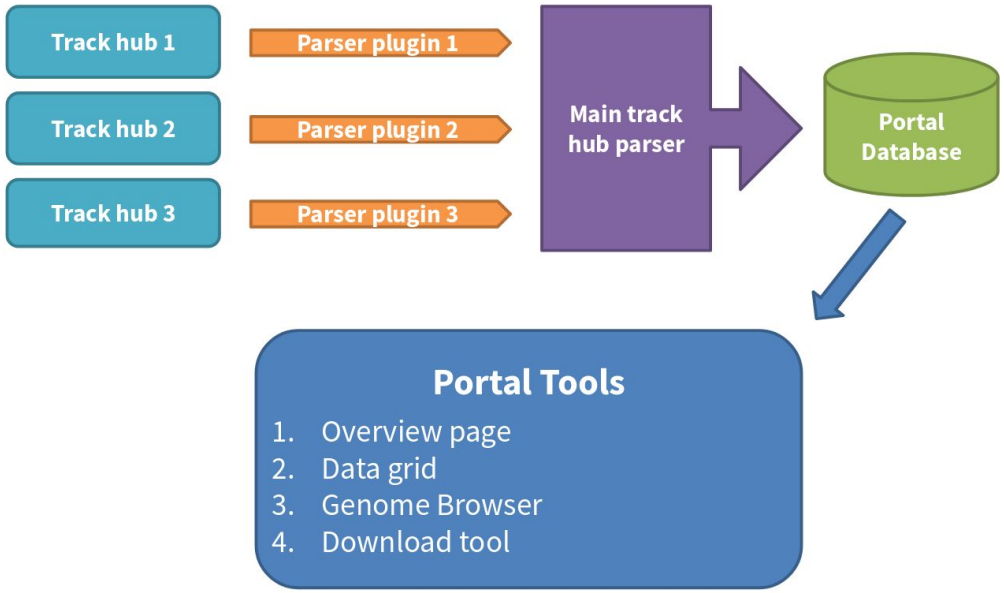
Web API

- A Web API enables users to download all available metadata in JSON format
  - Offering connectivity to other front-end applications


```
{
 ▾ datasets: {
    ▾ ERX197180: {
       ▾ analysis_attributes: {
            alignment_software: "BWA",
            analysis_software: "WIGGLER"
         },
       ▾ browser: {
          ▾ peak_calls: [
             ▾ {
                 big_data_url: http://ftp.ebi.ac.uk/pub/databases/blueprint/data/homo_sapiens/GRCh37/Cord_blood/C002YM/CD8-positive_alpha-
                 beta_T_cell/ChIP-Seq/NCMLS/C002YMH1.H3K4me3.ppqt_macs2_v2.20130415.bb,
                 md5sum: null,
                 primary: true
               }
            ],
          ▾ signal: [
             ▾ {
                 big_data_url: http://ftp.ebi.ac.uk/pub/databases/blueprint/data/homo_sapiens/GRCh37/Cord_blood/C002YM/CD8-positive_alpha-
                 beta_T_cell/ChIP-Seq/NCMLS/C002YMH1.H3K4me3.wiggler.20130415.bw,
                 md5sum: null,
                 primary: true
               }
            ]
         },
       ▾ experiment_attributes: {
            epirr_id: "IHECRE00000035.1",
```

Gathering the data

- Initially a challenge
  - completely decentralized consortium
  - no common data sharing standards across sites
- Most sites were using UCSC Track Hubs to display data in the UCSC Browser
- Manually implemented filters for each data producer



IHEC data sharing now

- Common JSON schemas to share metadata
- Use of ontologies whenever possible to replace controlled vocabularies that tend to grow
- Examples:
  - Donor Health Status: NCI Metathesaurus
  - Experiment Type: Ontology for Biomedical Investigations
  - Molecule (e.g. Genomic DNA, RNA): Sequence Ontology
  - Tissue type: UBERON


```
{
 ▾ datasets: {
    ▾ ERX197180: {
       ▾ analysis_attributes: {
            alignment_software: "BWA",
            analysis_software: "WIGGLER"
         },
       ▾ browser: {
          ▾ peak_calls: [
             ▾ {
                  big_data_url: http://ftp.ebi.ac.uk/pub/databases/blueprint/data/homo_sapiens/GRCh37/Cord_blood/C002YM/CD8-positive_alpha-
                  beta_T_cell/ChIP-Seq/NCMLS/C002YMH1.H3K4me3.ppqt_macs2_v2.20130415.bb,
                  md5sum: null,
                  primary: true
               }
            ],
          ▾ signal: [
             ▾ {
                  big_data_url: http://ftp.ebi.ac.uk/pub/databases/blueprint/data/homo_sapiens/GRCh37/Cord_blood/C002YM/CD8-positive_alpha-
                  beta_T_cell/ChIP-Seq/NCMLS/C002YMH1.H3K4me3.wiggler.20130415.bw,
                  md5sum: null,
                  primary: true
               }
            ]
         },
       ▾ experiment_attributes: {
            epirr_id: "IHECRE00000035.1",
```

Registered assays

- Raw data need to be deposited at a controlled access repository
- Afterward, dataset can be registered in EpiRR
  - Database of IHEC epigenomes with available raw data

IHECRE00000002.3

Type	Single donor
Status	Partial
Project	BLUEPRINT
Local name	
Description	Acute Promyelocytic Leukemia - MC2884 (4h) from bone marrow of donor: pz 289
Is live version?	yes
Other versions	previous

Metadata

biomaterial_type	Primary Cell
donor_id	pz 289
disease_ontology_uri	http://ncimeta.nci.nih.gov/ncimbrowser/ConceptReport.jsp?dictionary=NCI%20MetaThesaurus&code=C0023487
passage_if_expanded	NA
donor_ethnicity	Caucasian
genetic_characteristics	Translocation t(15;17)
donor_sex	Female
gender	female
tissue_type	bone marrow
taxon_id	9606
species	Homo sapiens
disease	Acute Promyelocytic Leukemia
sample_ontology_uri	http://purl.obolibrary.org/obo/CL_0000763
donor_age	30 - 35
treatment	MC2884 (4h)
phenotype	CL_0000763;C0023487;UBERON_0002371
markers	NA
cell_type	myeloid cell
donor_health_status	Acute Promyelocytic Leukemia
biomaterial_provider	Prof: Lucia Altucci (SECONDA UNIVERSITA' di NAPOLI- IT)

Raw data

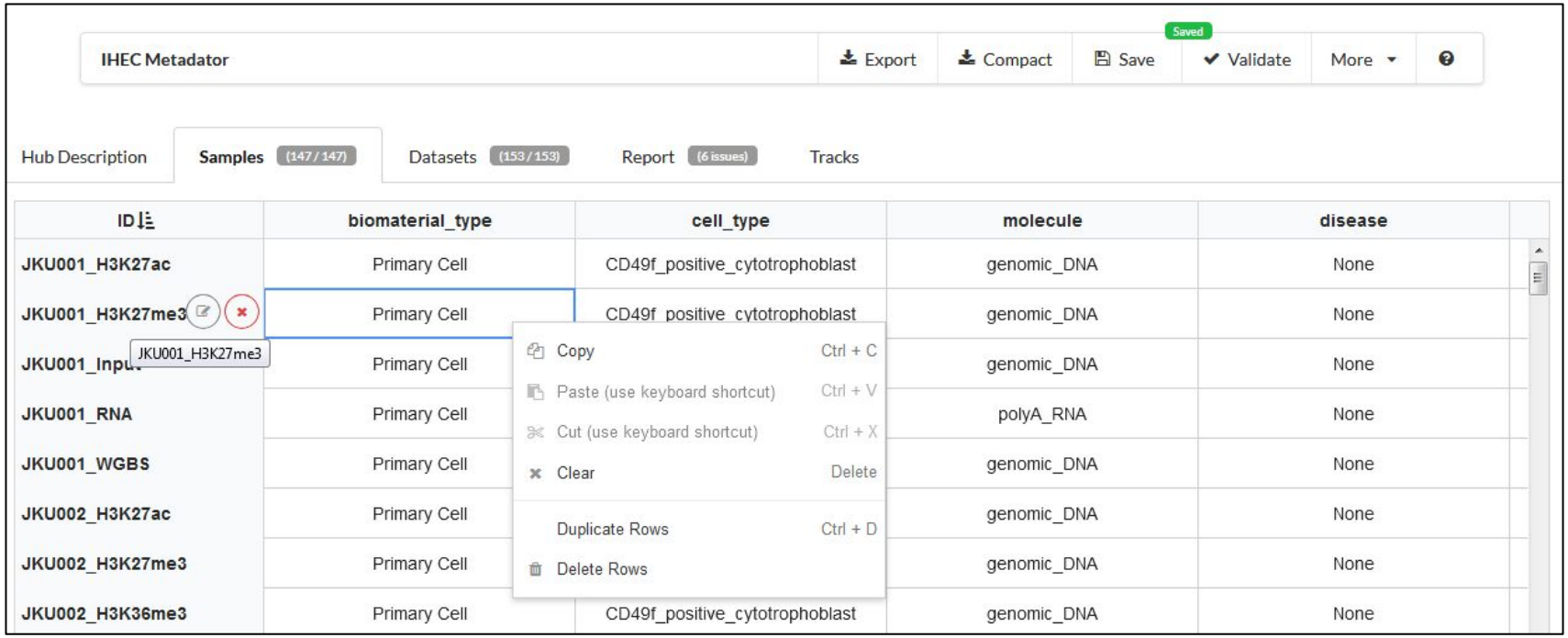
Assay type	Experiment type	Archive	Primary ID	Secondary ID	Link
ChIP-Seq	Histone H3K4me1	EGA	EGAX00001215591	EGAD00001002372	View in archive
ChIP-Seq	Histone H3K27me3	EGA	EGAX00001138954	EGAD00001002372	View in archive

EpiRR Epigenome Summary

Project	Complete	Partial	Total Epigenome count
AMED-CREST	21	2	23
BLUEPRINT	112	1137	1249
CEEHRC	80	316	396
DEEP	19	24	43
ENCODE	0	103	103
GIS	0	291	291
Korea Epigenome Project (KNIH)	0	11	11
NIH Roadmap Epigenomics	62	49	111
Total	294	1933	2227

IHEC Metadator

- Visual interface to create/edit IHEC Data Hubs (JSON annotation files)
- Includes validation features
- Includes an ontology lookup service to resolve terms in the interface



IHEC Metadator

IHEC Metadator | Export | Compact | Save (Saved) | Validate | More

Hub Description | Samples (147 / 147) | Datasets (153 / 153) | Report (6 issues) | Tracks

Issues

There are issues with you data

Please review and fix the issues described below. Proposed actions to resolve the issues are displayed on the right.

Dataset "JKU001_H3K4me3": Sample "JKU001_H3K4me3" does not exists — Add Sample "JKU001_H3K4me3" | Edit

Dataset "JKU001_H3K4me1": Sample "JKU001_H3K4me1" does not exists — Add Sample "JKU001_H3K4me1" | Edit

Dataset "JKU001_H3K9me3": Sample "JKU0 — Edit

Dataset "JKU001_H3K36me3": Sample "JKU — Edit

Dataset "JKU016_WGBS": Sample "JKU016_ — Edit

Dataset "JKU017_WGBS": Sample "JKU017_ — Edit

Edit Item

Primary Cell

IHEC Data Hub Sample schema

id *

Uniquely identifying ID

JKU001_H3K27ac

biomaterial_type *

Primary Cell

sample_ontology_uri *

Ontology term that links to sample ontology information. Depending on the biomaterial_type, will be either an UBERON or CL ontology term.

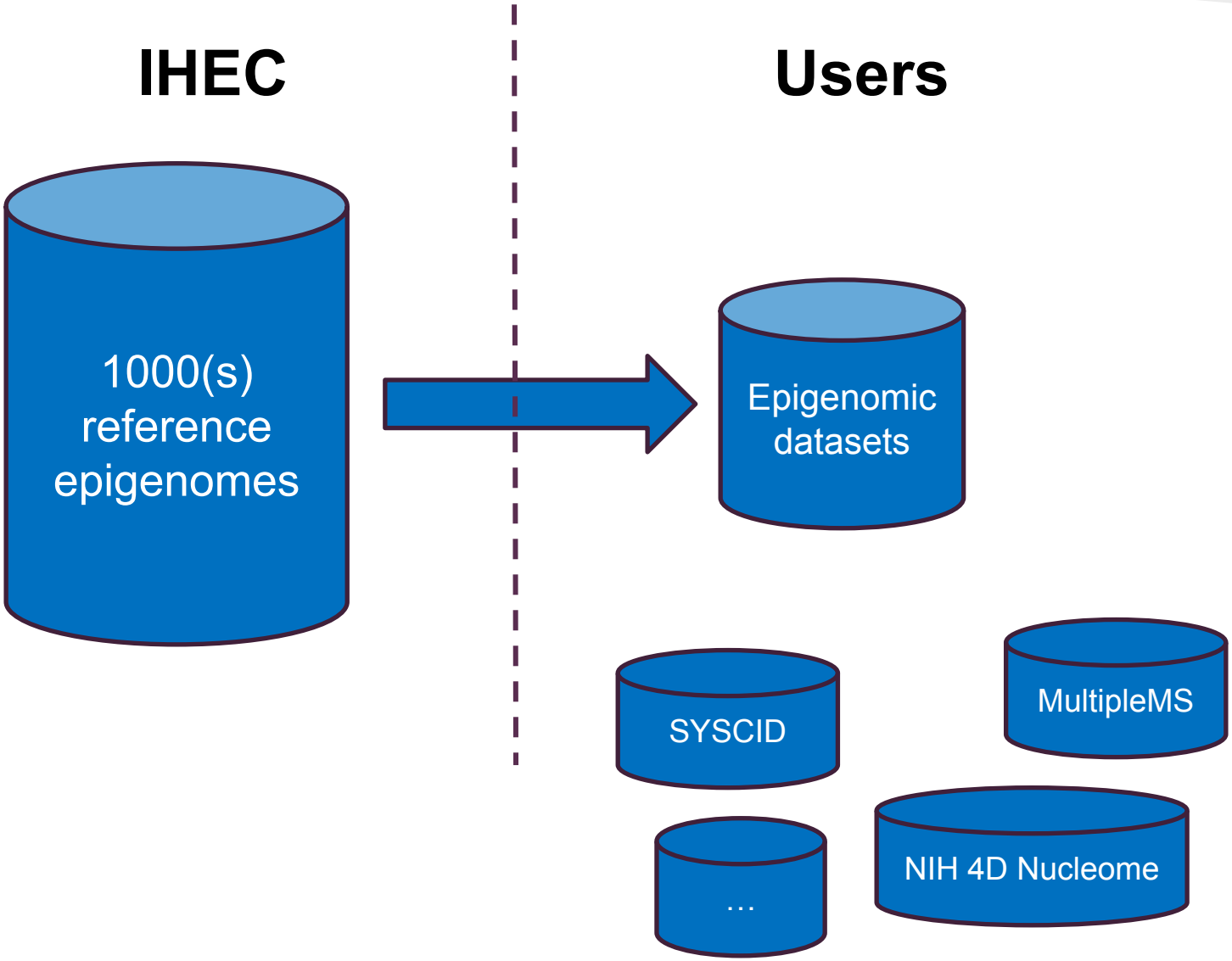
http://purl.obolibrary.org/obo/CL_0000523

molecule *

The type of molecule that was extracted from the biological material. Include one of the following: total RNA, polyA RNA, cytoplasmic RNA, nuclear RNA, genomic DNA, protein, or other.

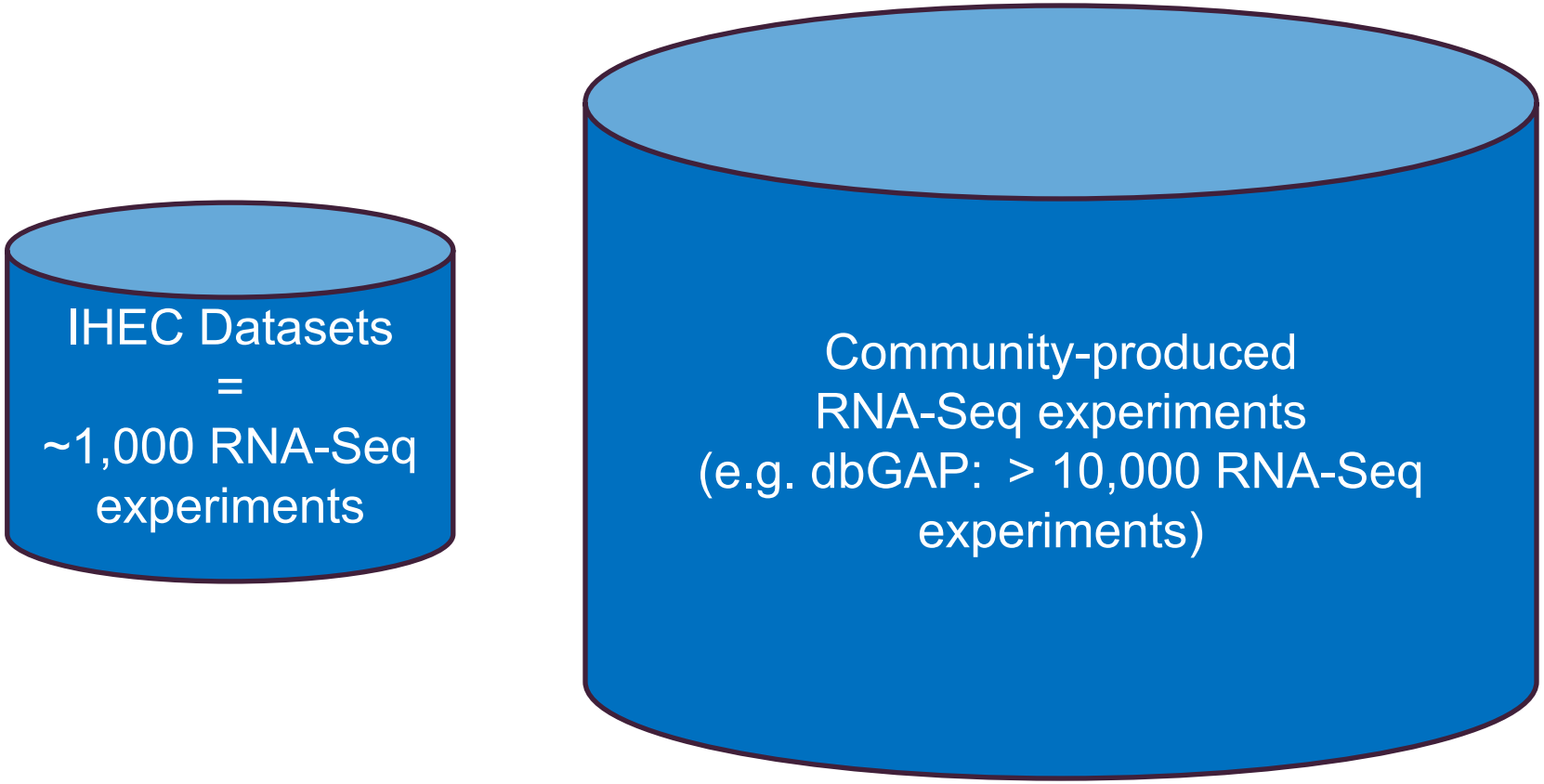
genomic_DNA

Next challenge: Integrating community datasets



Many more epigenomic datasets are out there

- How do we put IHEC data in context with all other available epigenomic datasets?



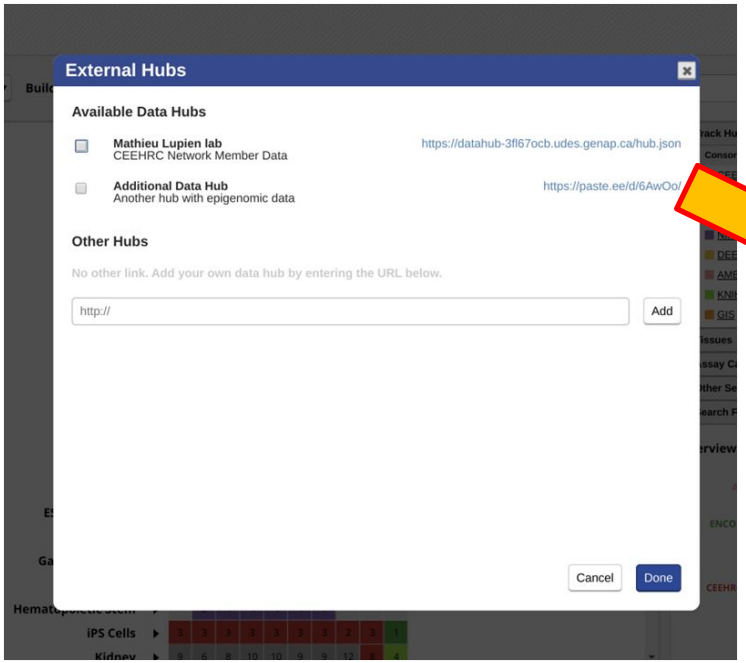
Community hubs

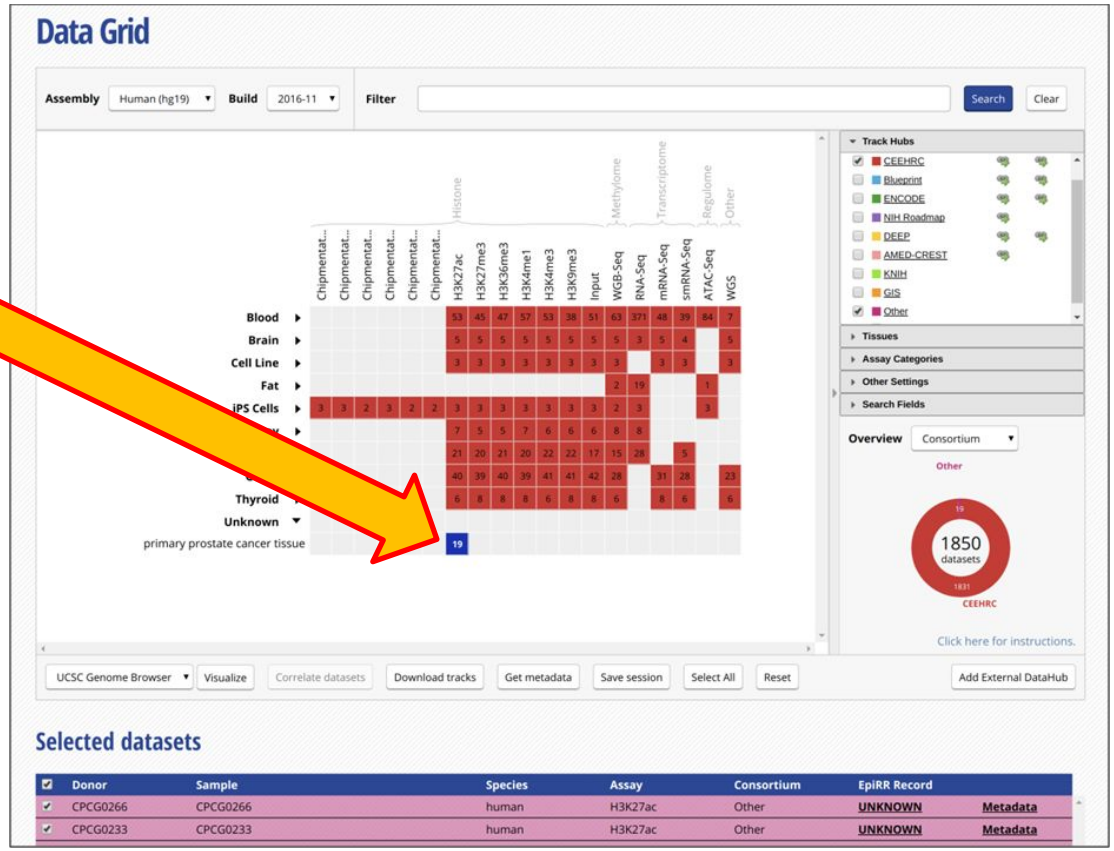
Using IHEC Data Hubs, the Portal enables external data integration

- Add pre-built data hubs from non-IHEC sources to Portal sessions
- Upload your own data hub to your session
- Possible to create IHEC Data Hubs easily using the Metadator
- Allow users to load custom datasets in the portal, and compare them to IHEC ones

Community hubs

- Data is integrated in the Portal interface, and usable in external tools





Another challenge

- Processed (public) data is generated by different groups
  - Differences in pipelines and tools
  - Differences in tools parameters
  - Differences in output data types
  - Differences in quality thresholds

The epiMAP project

- IHEC Integrative Analysis workgroup initiative
- Goals of epiMAP:
  a. establish an analyst-friendly and widely sharable compendium of **quality-controlled, consistently processed, reference epigenomic maps** from all areas of IHEC.
  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 large scale analysis

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

Table 1: Clauses Identified across IHEC Agreements

	#1	#2	#3	#4	#5	#6	#7
Constraints on Use	■	■	■	■	■	■	
Application Renewal			■		■		
Evidence of Competence			■	■	■		
Student Access			■	■	■	■	
Specific External Laws			■	■		■	
Specific Policies	■	■	■		■		■
Jurisdiction			■	■		■	
External Access	■	■	■	■			
Acknowledgements	■	■	■	■	■	■	■
Liability	■	■	■	■	■	■	
Report to Project	■	■		■		■	
Publication Delays	■	■	■	■	■	■	
Destruction of Data	■	■	■	■	■	■	
Ethics Review			■	■	■		■
IT Practices			■				
Intellectual Property	■	■	■	■	■	■	
Unique Provisions			■	■		■	

Joly, Y. The challenge of harmonizing data access agreements for IHEC. Presented at the IHEC Annual Meeting 2017.

Downloading/analysing raw data

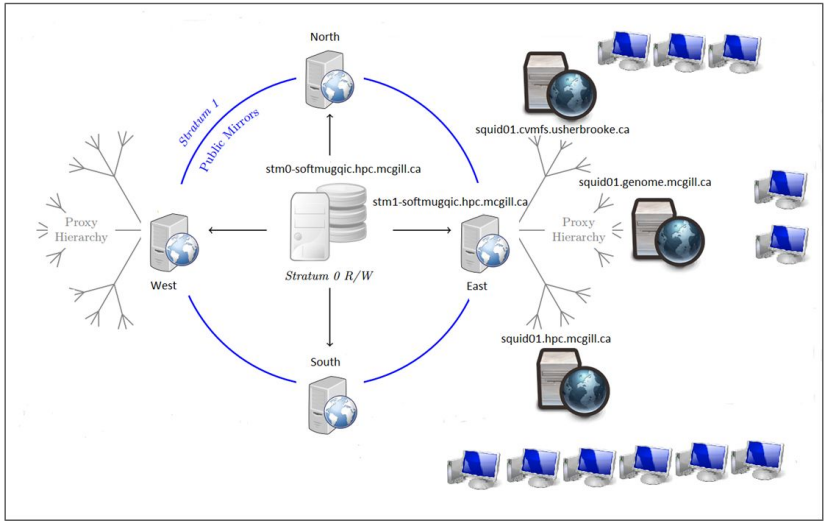
- Downloading the data can be a very long endeavour
  - For large datasets, downloading from a controlled access repository can take several months (years)
- For big datasets, large amount of space is required
  - To download the whole IHEC raw data, hundreds of terabytes are required
- Analyses are often processor and memory intensive
  - Not something that can be done on one's laptop…
- Several resources exist to address this issue, such as:
  - Commercial solutions (e.g. AWS)
  - Compute Canada

GenPipes

- Free, open-source software with Python
- Many pipelines available, including for epigenomics experiments:
  - RNA-Seq
  - RNA-Seq Denovo
  - ChIP-Seq
  - Methyl-Seq (Bisulfite-Seq)
- All software requirements are pre-installed at many Compute Canada HPCs

Software through GenAP CVMFS

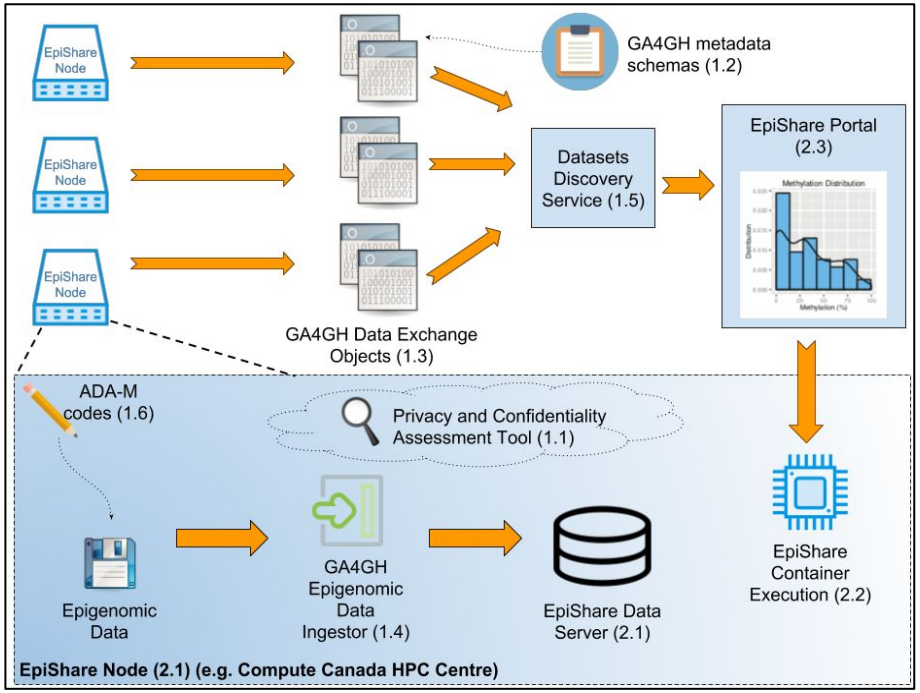
- Bioinformatics software and libraries are distributed to several Compute Canada HPCs using GenAP CVMFS
  - CVMFS is a distributed file system originally developed for CERN experiments computation
  - Software and libraries are configured in the exact same way at all locations supporting CVMFS



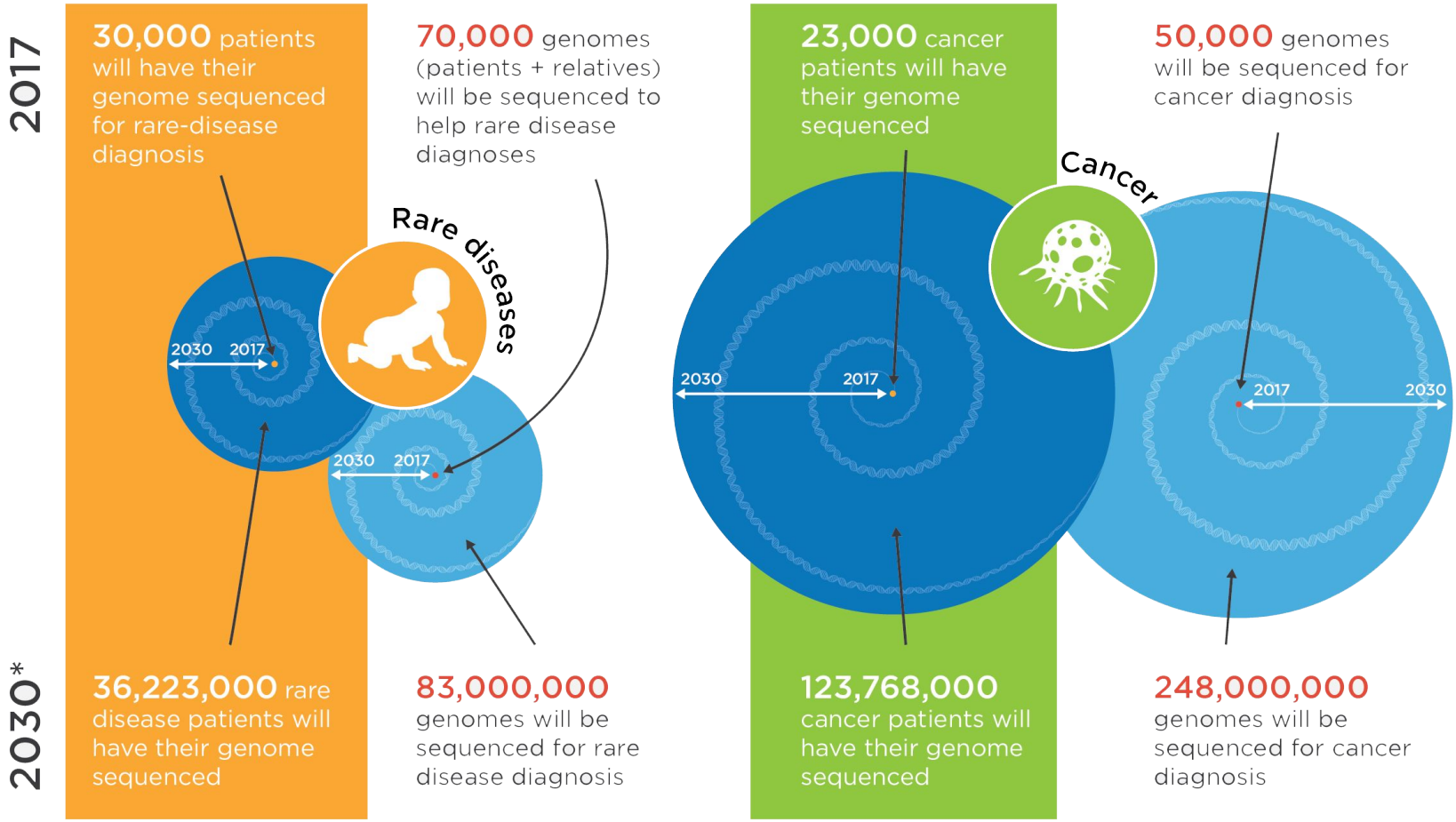
CVMFS code distribution, from central repository to local user caches

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



Global Alliance for Genomics and Health (GA4GH)



* Projected figures, based on current data and known status of genomics initiatives worldwide.

Global Alliance for Genomics and Health (GA4GH)

Members

Experts in healthcare, research, patient advocacy, life science, and information technology

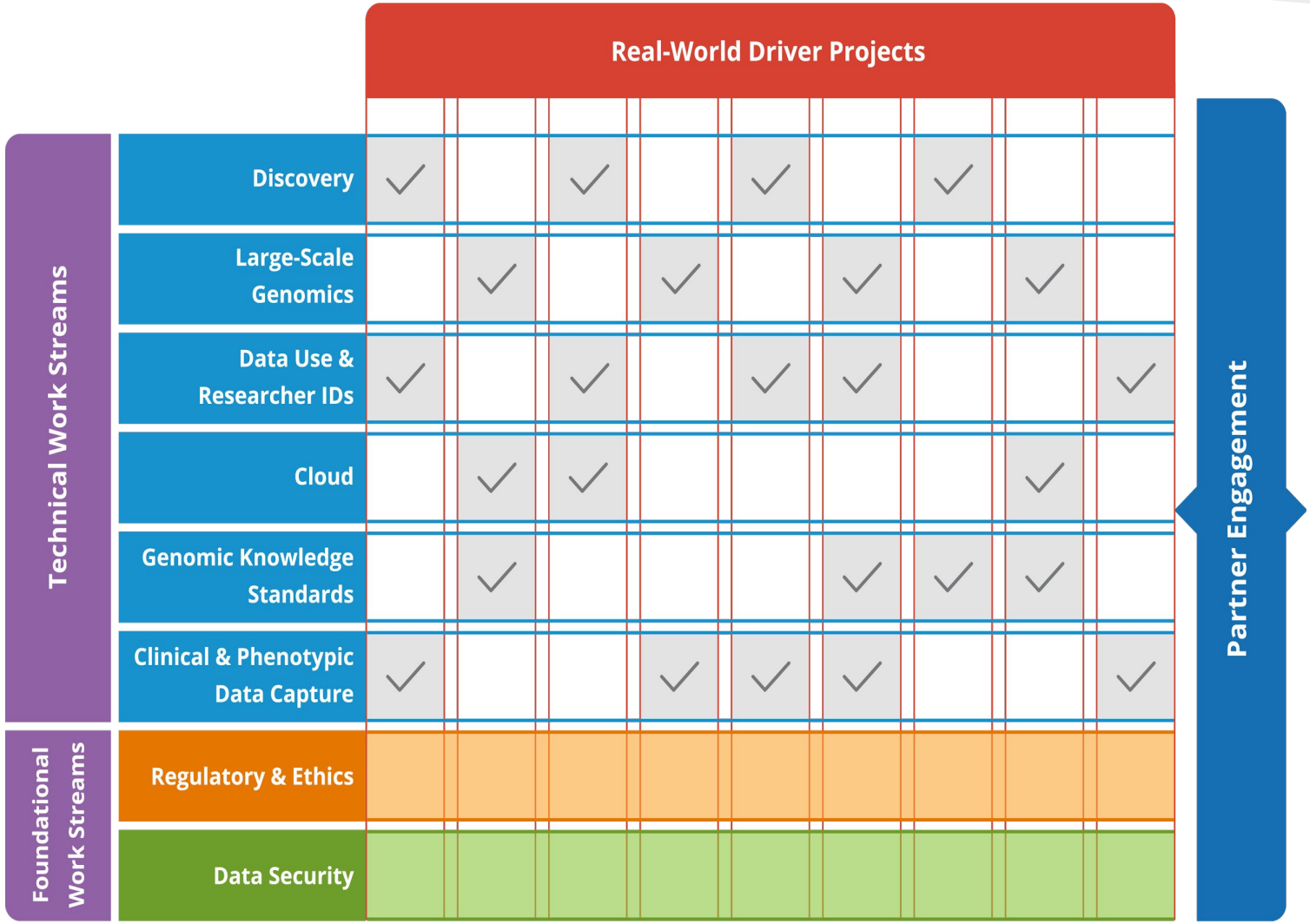
500+ **Organizational members**

2000+ **Subscribers**

71 **Countries**

BECOME A MEMBER

GA4GH Structure



Conclusion

- IHEC Data Portal has adopted different strategies to make data Findable
- Still, making restricted access data truly available to the research community has been challenging
- New emerging standards within GA4GH will be adopted to further improve the ways we share genomic data
- Through EpiShare, we hope to make meaningful contributions to GA4GH for epigenomic-specific issues

Team, Partners and sponsors

Team:

- David Bujold (1)
- Romain Grégoire (1)
- David Anderson (1)
- Michel Barrette (2)
- Carol Gauthier (2)
- Tony Kwan (1)
- Alain Veilleux (2)
- Pierre-Etienne Jacques (2)
- Guillaume Bourque (1)

1. McGill University, Montreal, Quebec, Canada
2. Université de Sherbrooke, Sherbrooke, Quebec, Canada

McGill EPIGENOMICS MAPPING CENTRE

GenAP

GenomeQuébec

canarie

GenomeCanada

compute canada | calcul canada

NSERC CRSNG

C3G

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada