1. Introduction

EpiShare is an open science project involving the International Human Epigenome Consortium (IHEC) and ENCODE that aims to develop tools and APIs to increase accessibility of epigenomic data. It does so by using and contributing to standards established by the Global Alliance for Genomics and Health (GA4GH), such as the maget API specification to store and query transcriptomic data in formats that are compatible with other large scale projects (e.g. Human Cell Atlas). EpiShare objectives:

- Offer methods to easily discover available epigenomes, and to launch multi-omics analyses on restricted access datasets
- Develop an infrastructure of nodes to securely store and make epigenomic (meta)data available
- Deploy data analysis containers to be executed at data storage location when possible (bring computation to the data), or to otherwise bring them to selected compute services

2. EpiShare and GA4GH

As an epigenomics-focused project, EpiShare brings a unique perspective to GA4GH. Many of the existing standards are focussed on genomic and health information, meaning that EpiShare has an opportunity to contribute new insight to epigenomics data standardization.

<table>
<thead>
<tr>
<th>GA4GH Standard</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workflow Execution Service (WES)</td>
<td>API specification to standardize how workflow runs are submitted and monitored.</td>
</tr>
<tr>
<td>Data Repository Service (DRS)</td>
<td>API specification for serving data objects (i.e. files) and bundles (i.e. folders.)</td>
</tr>
<tr>
<td>Phenopackets</td>
<td>A transfer format for clinical and phenotypic metadata.</td>
</tr>
<tr>
<td>Tool Registry Service (TRS)</td>
<td>API specification for communicating a platform’s workflows and tools.</td>
</tr>
<tr>
<td>Discovery workflow standards</td>
<td>Service Info endpoint, Service Registries, and eventually Search.</td>
</tr>
</tbody>
</table>

Some GA4GH standards used by EpiShare. For example, the Workflow Execution Service (WES), running Workflow Description Language (WDL)-powered workflows, is used for data analysis and ingestion pipeline execution. In the future, this will be integrated with a Data Repository Service schema (DRS) compatible service to provide persistent, easily controlled access URIs for data objects.

2. RNAget

EpiShare was involved in the development of the original maget API specification, and has implemented it in the IHEC Data Portal. Using the /continuous endpoint, portal users can extract slices of epigenomic experiments signal tracks for RNA-Seq and ChIP-Seq, and download them in the LOOM format.

3. REWS (Regulatory & Ethics Work Stream)

We are involved in the GA4GH REWS, contributing to the development of Data Access Committee Review Standards (DACReS).

DACReS will identify areas of best practices for procedural standards to drive consistency and robust reviews for data access requests to genomic, epigenomics and health-related data.

To join this effort, please contact rews-coordinator@ga4gh.org

4. D-PATH Data Privacy Assessment Tool for Health

We performed an in-depth study of the relevant legislation, ethical standards, and best practice documents to develop the Data Privacy Assessment Tool for Health (D-PATH). D-PATH is a unique tool to ensure that EpiShare’s data sharing activities meet the applicable ethical, legal, professional requirements. While the current version is taking into account the particular needs of EpiShare (physically located in Quebec while processing data from Canadian and international cohorts), D-PATH could be transposed to other data sharing projects, as we plan on expanding it to account for laws and policies from other jurisdictions.

5. Clinical/Phenotypic/Experiments Metadata standards

We implemented Katsu, an open-source clinical and phenotypical metadata service around Phenopackets, offering interoperability between other data exchange formats (FHIR, mCODE). This service promotes usage of existing biomedical ontologies and controlled vocabularies for annotations, and captures experiments metadata and their relationship to phenotypic descriptions. Consequently, we are participating in the GA4GH Clin/Pheno Work Stream to help develop the Phenopackets standard.

EpiShare puts emphasis on using common standards, such as those proposed by GA4GH and the W3C:

- The Data Use Ontology (DUO) captures metadata related to a dataset usage restrictions and conditions.
- DATS Schema defines how and where the dataset was produced, its subjects, creators, location in a repository, etc.
- Patient’s profile is extended with oncology related metadata (mCODE).
- Phenotypic (GA4GH Phenopackets) and Experiment metadata (mCODE), both are interoperable (to some extent) with FHIR.
- Phenopackets schema is based on the IHEC specifications and linked to Phenopackets.

Katsu source code available here: https://github.com/bento-platform/katsu

7. Future work

- Adapt the D-PATH tool to support data sharing projects from other jurisdictions.
- Adapt the Katsu clin/phenometadata service to support the newest 2.0 revision of Phenopackets.
- Release EpiShare’s node architecture as an open-source product to support external epigenomic data sharing initiatives. Future developments will be made available in the https://github.com/bento-platform repository.

8. Partners and sponsors

The EpiShare prototype and resources are hosted by GenAP (genap.ca) and funded by Genome Canada, and leverage resources from the IHEC Data Portal, an online resource funded under the CEEHRC, by the CIHR and Genome Quebec, with additional support from Genome Canada. The computing and networking infrastructure, and part of the software development, are provided by Compute Canada and CANARIE.