FAIRification of Genomic Tracks
Implementation Study

ELIXIR All Hands meeting 2019, June 17–20, Lisbon, Portugal

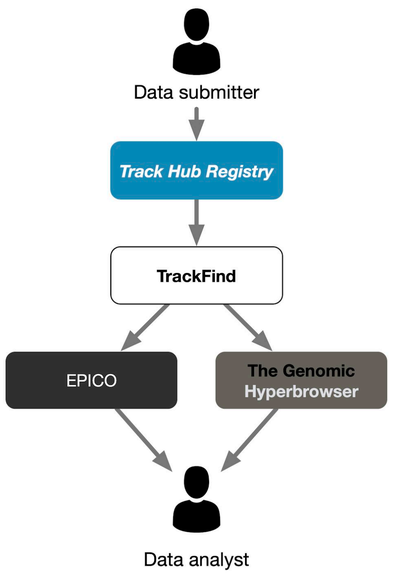
Salvador Capella-Gutierrez [1], Finn Drabløs [2], José M. Fernández [1], Sveinung Gundersen [3], Eivind Hovig [3, 4], Radmila Kompova [3], Kieron Taylor [5], Dmytro Titov [3], Daniel Zerbino [5]

[1] *INB Coordination Node / ELIXIR ES, Barcelona Supercomputing Center, Spain*
[2] *Norwegian University of Science and Technology (NTNU), Norway*
[3] *Center for Bioinformatics, University of Oslo (UiO), Norway*
[4] *Department of Tumor biology, Institute for Cancer Research, Oslo University Hospital (OUH), Norway*
[5] *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), United Kingdom*

FAIR track metadata *- The FAIRification of Genomic Tracks Implementation Study is focused on "FAIRifying" the metadata related to genomic annotation track files contained in track hubs. To achieve this, we developed a common data model and technical solutions compatible with the existing TrackHub exchange format for genome browser tracks and implemented demonstrators to show the feasibility of this proposal across systems and programming languages.*

Motivation

While there exist substantial efforts to ensure that large bioinformatic resources and consortia adhere to FAIR principles, an increasing amount of genomic data is now being produced in independent laboratories and distributed informally as genomic track files. These files, typically in BAM, VCF, BigBed or BigWig formats, are primarily designed to be displayed on a genomic browser such as Ensembl (1) or UCSC (2), but can also be reused for analysis. They are often bundled and distributed within track hubs, which are configuration files with metadata and links to the actual data files. Also, even in cases where FAIR metadata already exist, they may be poorly curated, and their structure and vocabulary typically differ substantially, making it difficult to search and combine track data from different sources.

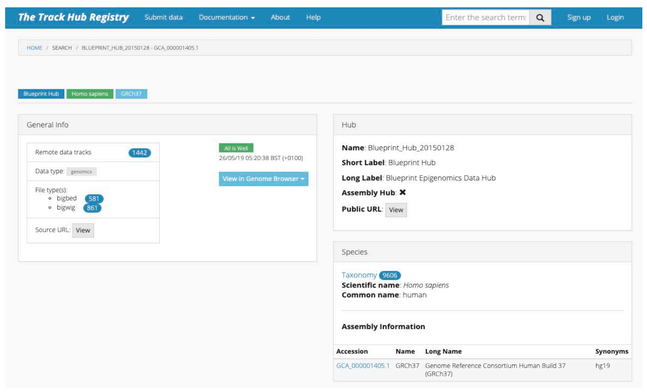


Common data model

We started by modelling a comprehensive JSON Schema, currently named "Fairtracks", that defines the most important metadata of a genomic track and ensures that the metadata is FAIR. We have considered important existing standards, authorities, and consumers of track metadata, including INSDC (3), IHEC (4), FAANG (5), TrackHub file format (6), ISA Model (7), Zenbu (8), and GSuite HyperBrowser (9), in order to identify the fields to be included in our FAIR standard. Also, we enforce the use of specific ontologies for several of the fields, as well as requiring the use of CURIE identifiers registered in identifiers.org. We have developed validators in several programming languages that are able to check additional constraints specific to the genomic tracks metadata model, declared using reference extensions on the JSON Schema vocabulary. Our JSON schema and extended validators are available from GitHub (10).

Metadata aggregation: Track Hub Registry

The Track Hub Registry services (11), maintained by EMBL-EBI, allows independent researchers to distribute their track hubs. Each track hub is a set of text files with links to data files, display configuration for each file, but also some metadata, which is used by the browsers to dynamically create selection menus. We extended the TrackHub Registry so that its content could be queried by outside services.

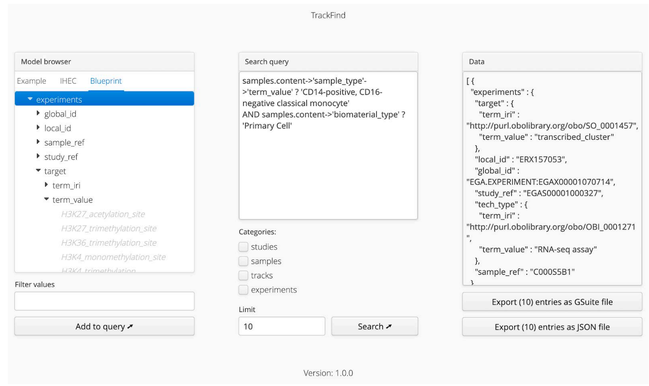


Demonstration

To complete the demonstration, a TrackHub file was updated and fed into this pipeline. To account for scale and complexity, we produced a track hub of public BLUEPRINT data files, which was loaded into TrackFind. The TrackFind client in GSuite HyperBrowser was used to query the metadata and extract a related collection of tracks, which was then used to carry out an analysis in the HyperBrowser.

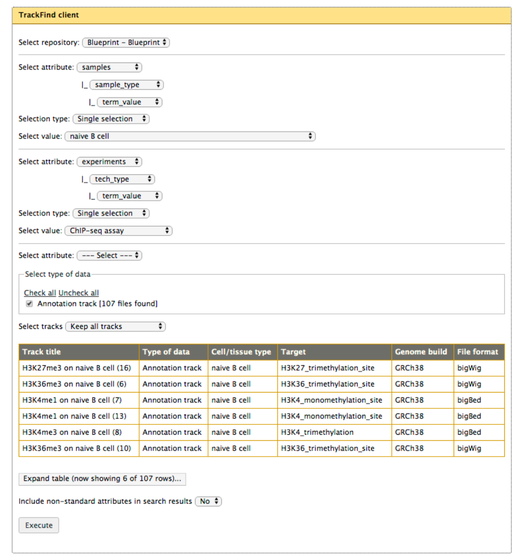
Metadata curation and distribution: TrackFind

ELIXIR Norway at UiO implemented a track search service called TrackFind (12) that integrates metadata from various sources, including the Track Hub Registry (11), according to the proposed recommendations. TrackFind supports hierarchical browsing of metadata and advanced search queries, both through a web-based user interface, and as a RESTful API (13), and the search results can be browsed and exported in JSON or GSuite (14) format. Future improvements include the support for custom mappings and editing of metadata towards the recommendations – for data sources that are either historical or have not yet applied the recommendations.



Metadata re-use I: GSuite HyperBrowser

The GSuite HyperBrowser (9) is a general-purpose web-based platform for rigorous statistical analysis of track data, built upon the Galaxy framework. The HyperBrowser already has support for a track search mechanism (limited prototype), making use of the GSuite format (14) to move collections of track data (typically resulting from a track search operation) through both basic and advanced data manipulation and analysis steps. A TrackFind client (15) has been implemented to replace the existing prototype, and proved to work with BLUEPRINT data.



Metadata re-use II: EPICO

EPICO (16) is an open-access reference set of tools, libraries and APIs to develop comparative epigenomic data portals, as well as a data and metadata validator and database loader. EPICO components work with a customizable, rich data model where ontology term checks can be introduced for specific fields, as a generalization to enumerated values. EPICO has been used to implement the BLUEPRINT Data Analysis Portal (BDAP) (17). BDAP provides a virtual desktop for the comparative analysis of epigenetic features, recorded features (genes, transcripts, etc.) and pathways in the context of differentiation of hematopoietic lineages. The EPICO system will be modified to upload the FAIR metadata from TrackFind.

Online resources

1. https://www.ensembl.org
2. https://genome.ucsc.edu
3. http://www.insdc.org
4. http://ihec-epigenomes.org
5. https://www.animalgenome.org/community/FAANG
6. https://genome.ucsc.edu/goldenpath/help/hgTrackHubHelp.html
7. https://isa-specs.readthedocs.io/en/latest/isamodel.html
8. http://fantom.gsc.riken.jp/zenbu
9. https://hyperbrowser.uio.no
10. https://github.com/fairtracks/fairtracks_standard
11. https://trackhubregistry.org
12. http://trackfind-dev.gtrack.no
13. https://apidocs.trackfind-dev.gtrack.no
14. http://gtrack.no
15. https://hyperbrowser.uio.no/trackfind_test
16. https://github.com/inab/epico-data-analysis-portal
17. http://blueprint-data.bsc.es

Contact

Eivind Hovig
ehovig@ifi.uio.no
Oslo node of ELIXIR-NO (head)
Dep of Informatics, Univ. of Oslo,
Pb 1080 Blindern, 0316 Oslo, Norway

Spanish National Bioinformatics Institute
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
With funding from The Research Council of Norway
EMBL-EBI
NTNU Norwegian University of Science and Technology
UiO : University of Oslo