TrackFind

– FAIR Search of Genomic Tracks

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

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

***FAIR track search** – In the context of the ELIXIR Implementation Study: "FAIRification of Genomic Tracks", we have developed TrackFind, a track search engine and metadata FAIRification service. We believe TrackFind to be an important contribution, both to maintainers of genomic annotation track data, as well as to researchers and tool developers interested in making use of the wealth of genomic annotation track data publicly available.*

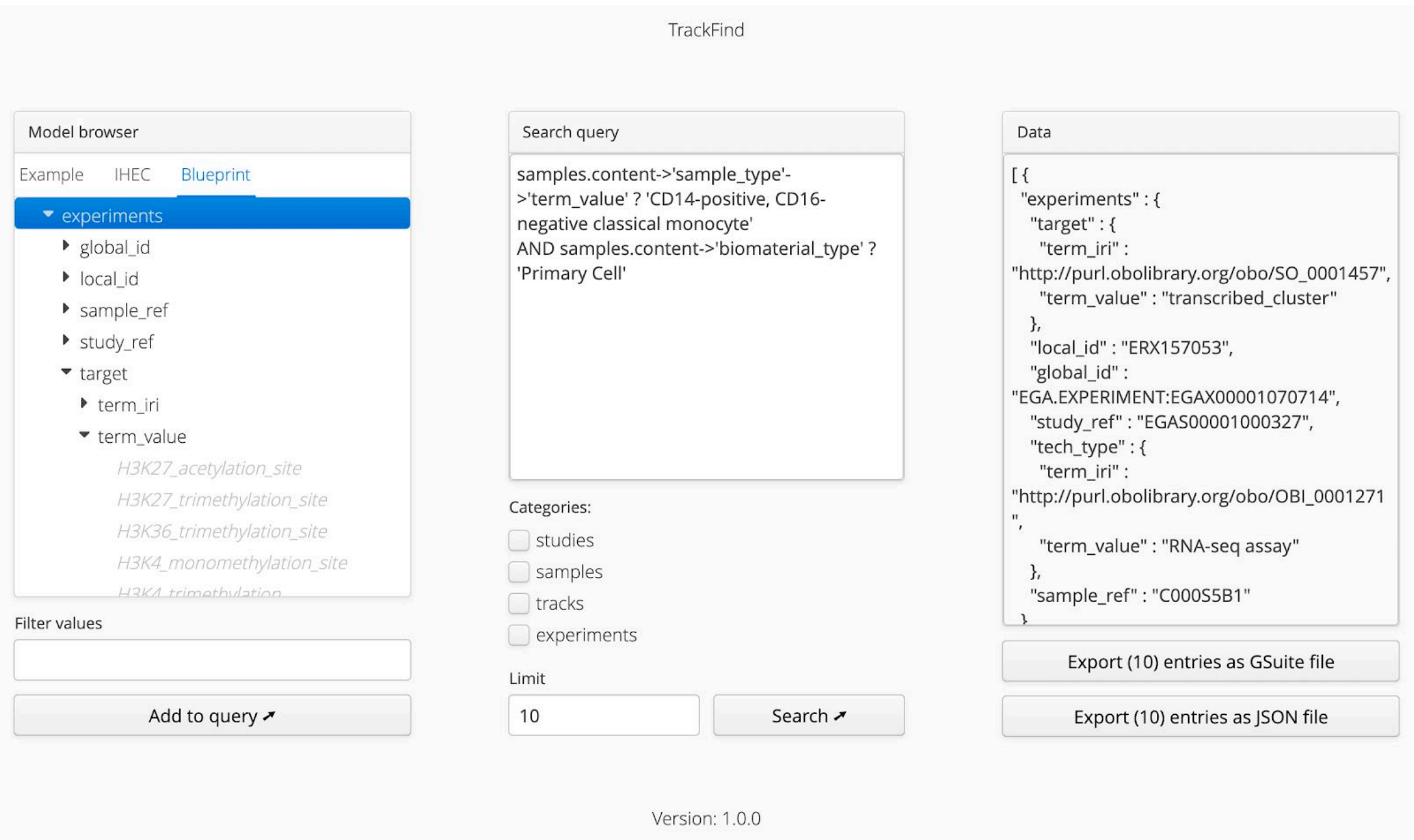
Motivation

Thousands of genomic annotation tracks have been generated the recent years, many in the context of larger undertakings such as BLUEPRINT and ENCODE. Several data portals for tracks are providing search services to researchers, but the underlying metadata is diverse and often poorly curated. The Track Hub Registry (11) provides a unified access point, but currently only supports limited search capabilities.

Overview of TrackFind

TrackFind (1) supports crawling of the Track Hub Registry (11) and other data portals to fetch track metadata. Crawled metadata can be accessed through hierarchical browsing or by search queries, both through a web-based user interface, and as a REST API (2). TrackFind supports advanced SQL-based search queries that can be easily built in the user interface, and the search results can be browsed and exported in JSON or GSuite (8) format (see Figure 1). The RESTful API allows downstream tools and scripts to easily integrate TrackFind search, currently demonstrated by the GSuite HyperBrowser (3), and soon by EPICO (12).

In addition to supporting most metadata models directly, TrackFind also supports the transformation of metadata into a JSON Schema currently named "Fairtracks" (4), as defined in the "FAIRification of Genomic Tracks" Implementation Study. Such transformation can be achieved on a per-trackhub basis through online scripting, thus providing a simple path for data managers to FAIRify their track metadata. TrackFind also maintains a version history of all metadata changes, including all recrawlings and transformations. We are also planning to add functionality for curating existing track metadata content.


TrackFind

Model browser

Search query

```
samples.content->'sample_type'-
>'term_value' ? 'CD14-positive, CD16-
negative classical monocyte'
AND samples.content->'biomaterial_type' ?
'Primary Cell'
```

Categories:
□ studies
□ samples
□ tracks
□ experiments

Limit: 10

Data

```
[ {
  "experiments" : {
    "target" : {
      "term_iri" :
"http://purl.obolibrary.org/obo/SO_0001457",
      "term_value" : "transcribed_cluster"
    },
    "local_id" : "ERX157053",
    "global_id" :
"EGA.EXPERIMENT:EGAX00001070714",
    "study_ref" : "EGAS00001000327",
    "tech_type" : {
      "term_iri" :
"http://purl.obolibrary.org/obo/OBI_0001271
",
      "term_value" : "RNA-seq assay"
    },
    "sample_ref" : "C000S5B1"
  }
```

Export (10) entries as GSuite file
Export (10) entries as JSON file

Version: 1.0.0

Figure 1: TrackFind (1) user interface for browsing, performing search queries, and exporting track metadata. The model browser on the left shows a hierarchical overview of all metadata attributes and values present in the selected track hub. Attributes or values from this browser can be dragged to the middle part of the screen, automatically becoming terms in the search query. Such on-the-fly query building allows for easy construction of complicated queries. After possibly applying filter and clicking the "Search" button, the search results (if any) will appear on the right, currently in JSON format. The results can be briefly reviewed and exported to a file, in either JSON or GSuite (8) formats.

Underlying technology stack

TrackFind (1, 5, 6) is a Java-based web-application powered by Spring Boot as a back-end technology and Vaadin framework as a front-end technology. The database engine used by TrackFind is PostgreSQL, which have been chosen due to support for storing arbitrary JSON documents while retaining the possibility to perform regular SQL queries to access data.

The application is distributed as a Docker image (9), which, together with the PostgreSQL image and the image for the JSON-to-GSuite converter (10), allows for easy deployment of the whole stack with Docker Compose.

Web user interface

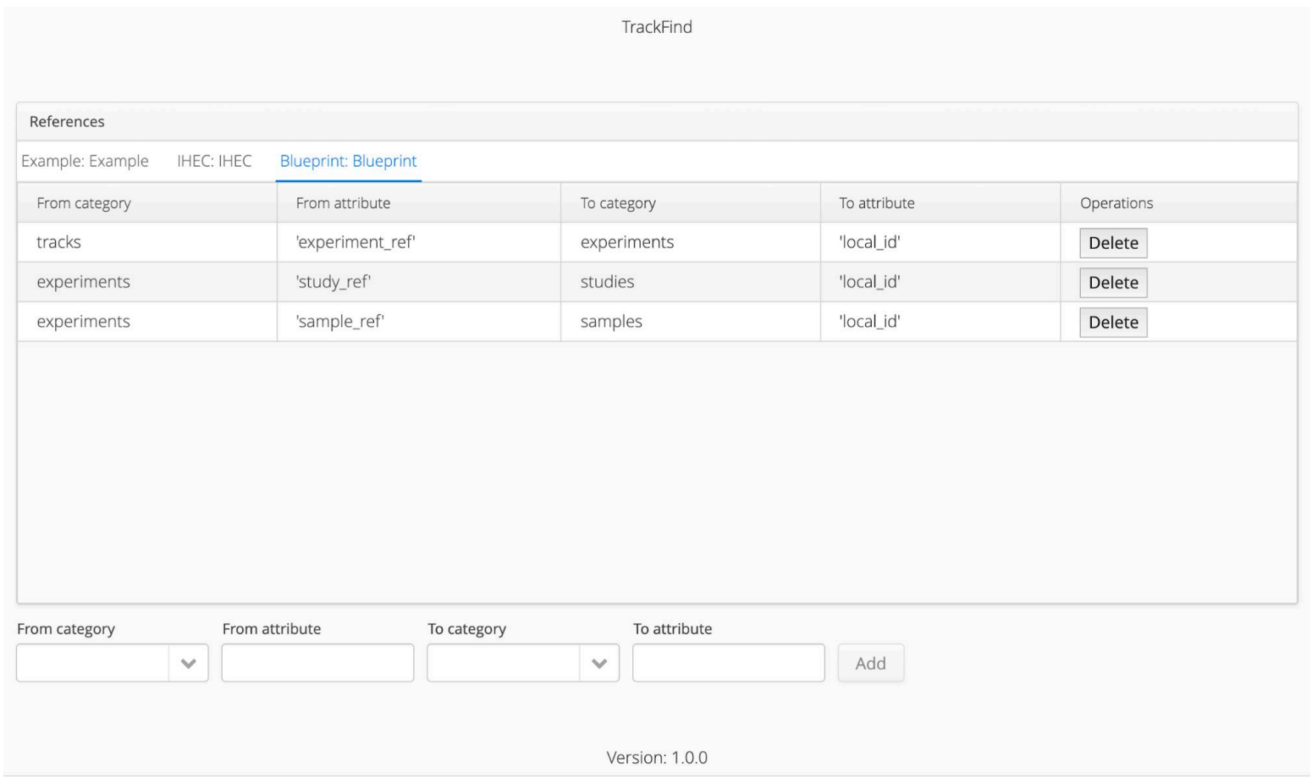
The Web user interface (UI) of TrackFind currently contains four pages:

1. The main TrackFind page (Figure 1), where users can browse, search, and export metadata.
2. The Hub admin page (not shown) allows to initiate the process of crawling, i.e., fetching metadata from an individual track hub or remote resource. This admin page also supports activation/deactivation of individual hubs.
3. The References admin page (Figure 2) provides a handy UI for managing cross-references between entities of different types.
4. The Mappings admin page (not shown) provides solutions for applying custom batch mappings for the imported metadata. This will include web-based scripting functionality. Such functionality could, for example, be used to map existing metadata onto the Fairtracks standard (4).

Future plans include separate page(s) to simplify curation of metadata content.

REST API

In addition to the Web UI, TrackFind is also accessible via a set of REST endpoints (Figure 3).

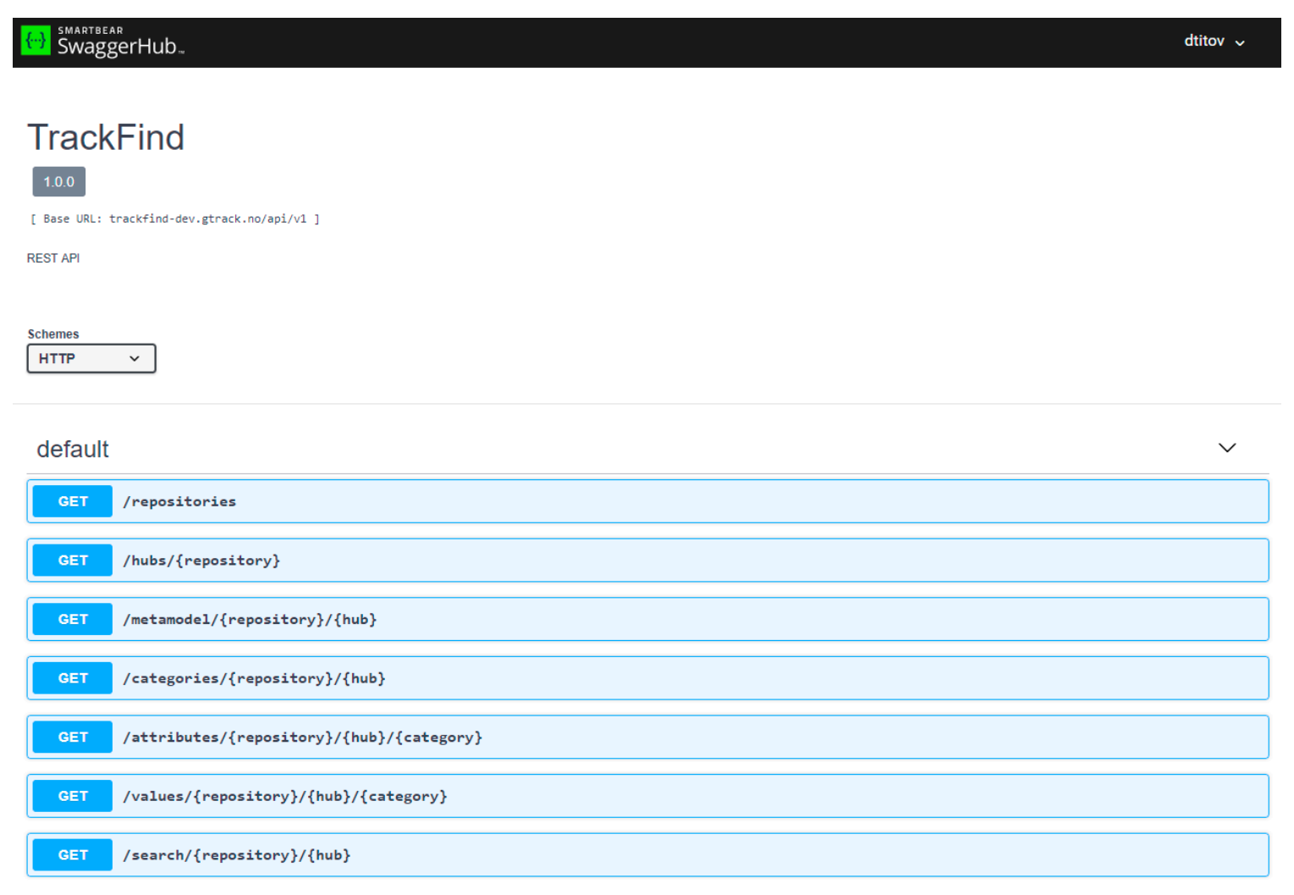

TrackFind

References

From category	From attribute	To category	To attribute	Operations
tracks	'experiment_ref'	experiments	'local_id'	Delete
experiments	'study_ref'	studies	'local_id'	Delete
experiments	'sample_ref'	samples	'local_id'	Delete

Version: 1.0.0

Figure 2: The References admin page. In this example, BLUEPRINT metadata has been mapped to the Fairtracks standard (4), which provides four object types for track metadata: studies, samples, experiments, and tracks. Metadata attributes in the different types are here linked together to allow for search and retrieval across object types.



TrackFind
1.0.0
[Base URL: trackfind-dev.gtrack.no/api/v1]
REST API

Schemes: HTTP

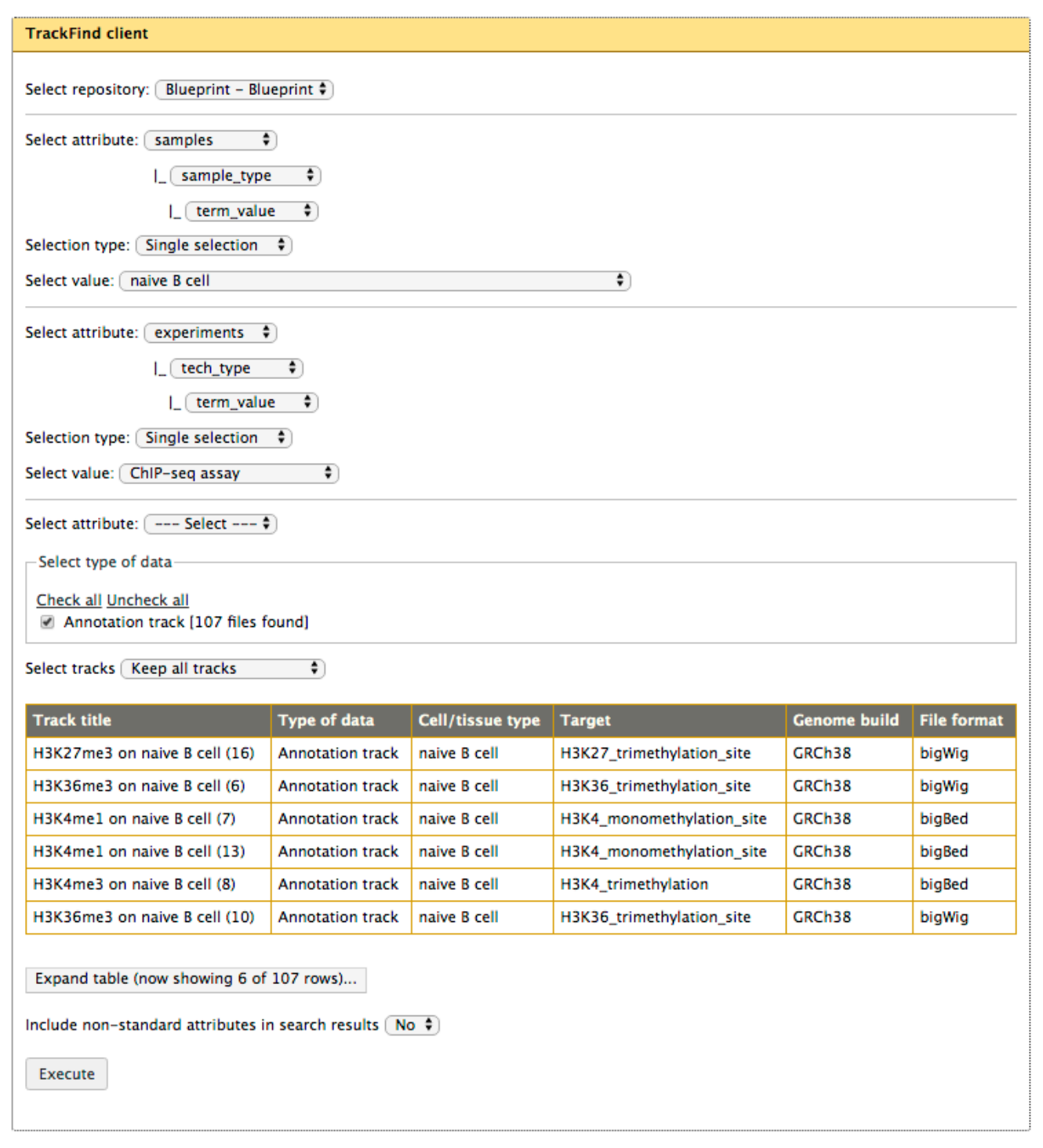
default

```
GET /repositories
GET /hubs/{repository}
GET /metamodel/{repository}/{hub}
GET /categories/{repository}/{hub}
GET /attributes/{repository}/{hub}/{category}
GET /values/{repository}/{hub}/{category}
GET /search/{repository}/{hub}
```

Figure 3: Specification of the TrackFind REST API, including functionality to manually try out the API (2)

Tool integration

One of the crucial features of TrackFind is its integration with downstream tools and scripts, mainly by making use of the REST API (2). A first example of such downstream use is a TrackFind client implemented in the GSuite HyperBrowser (3, 7), which builds upon the Galaxy platform. Figure 4 shows an example of search and retrieval from the BLUEPRINT track hub using the HyperBrowser client. TrackFind supports exporting track metadata into the GSuite tabular file format, which is a part of the BioXSD/GTrack family of file formats (8). GSuite metadata files can easily be manipulated and filtered in order to define a suite of related tracks to be downloaded and analyzed in bulk. GSuite files also provides a simple way for metadata to persist throughout the analysis process.


TrackFind client

Select repository: Blueprint – Blueprint

Select attribute: samples
|_ sample_type
|_ term_value
Selection type: Single selection
Select value: naive B cell

Select attribute: experiments
|_ tech_type
|_ term_value
Selection type: Single selection
Select value: ChIP-seq assay

Select attribute: --- Select ---

Select type of data
Check all Uncheck all
☑ Annotation track [107 files found]

Select tracks: Keep all tracks

Track title	Type of data	Cell/tissue type	Target	Genome build	File format
H3K27me3 on naive B cell (16)	Annotation track	naive B cell	H3K27_trimethylation_site	GRCh38	bigWig
H3K36me3 on naive B cell (6)	Annotation track	naive B cell	H3K36_trimethylation_site	GRCh38	bigWig
H3K4me1 on naive B cell (7)	Annotation track	naive B cell	H3K4_monomethylation_site	GRCh38	bigBed
H3K4me1 on naive B cell (13)	Annotation track	naive B cell	H3K4_monomethylation_site	GRCh38	bigBed
H3K4me3 on naive B cell (8)	Annotation track	naive B cell	H3K4_trimethylation	GRCh38	bigBed
H3K36me3 on naive B cell (10)	Annotation track	naive B cell	H3K36_trimethylation_site	GRCh38	bigWig

Expand table (now showing 6 of 107 rows)...

Include non-standard attributes in search results: No

Figure 4: A TrackFind client implemented within the GSuite HyperBrowser (3). A user has searched for ChIP-seq tracks targeting various histone modifications in naive B cells, and will receive a GSuite tabular file (8) containing the relevant track metadata. The actual track data can then, with a few clicks, be transferred from the BLUEPRINT FTP server to the HyperBrowser server, converted into the efficient binary BTrack format (8), and analyzed in bulk via the powerful statistical analysis tools provided by the HyperBrowser.

Resource availability (URLs)

TrackFind and related services are still under development, but are available as test instances, as follows:

1. TrackFind: *http://trackfind-dev.gtrack.no*
2. REST API docs: *https://apidocs.trackfind-dev.gtrack.no*
3. GSuite HyperBrowser TrackFind client: *https://hyperbrowser.uio.no/trackfind_test* (search for "trackfind" under Tools)

All source code is open source and available in GitHub as follows:

4. Fairtracks standard and validation tools: *https://github.com/fairtracks/fairtracks_standard*
5. TrackFind: *https://github.com/elixir-no-nels/trackfind*
6. JSON-to-GSuite converter: *https://github.com/elixir-no-nels/rest_gsuite*
7. GSuite HyperBrowser: *https://github.com/hyperbrowser/genomic-hyperbrowser* (TrackFind client currently in a separate branch)
8. GSuite specification (within GTrack ecosystem repo): *http://gtrack.no*

Docker images:

9. TrackFind: *https://hub.docker.com/r/nels/trackfind*
10. JSON-to-GSuite converter: *https://hub.docker.com/r/nels/rest_gsuite*

Other integrated resources:

11. Track Hub Registry: *https://trackhubregistry.org*
12. EPICO: *https://github.com/inab/epico-data-analysis-portal*

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
elixir NORWAY