Track hubs and their Registry Remote Data Integration made easy?

Alessandro Vullo
avullo@ebi.ac.uk

Ensembl Core Team

EMBL – European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD, UK

Outline

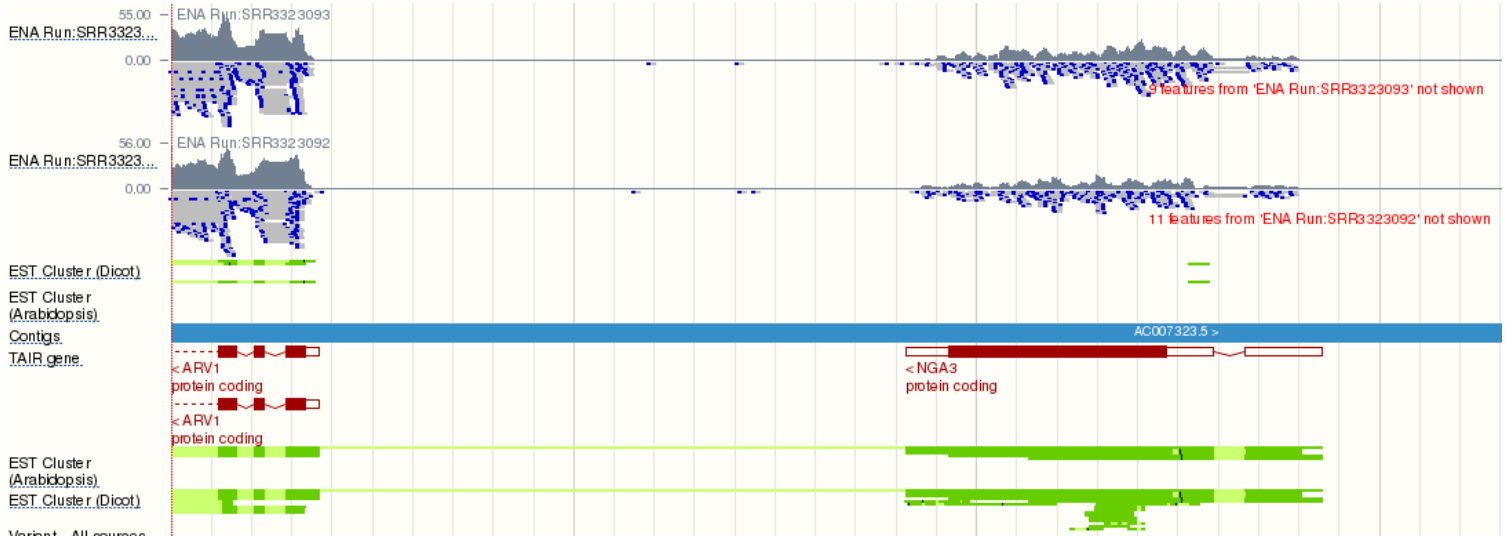
- Prelude: track data hubs
- How do we use them in Ensembl?
- The **Track Hub Registry**

Background

- High-throughput seq challenges genomic data visualisation tools
- UCSC/Ensembl browsers improved
  - visualise remotely hosted large data sets
  - support binary indexed data: BigBed, BigWig, BAM, VCF/tabix etc.
- UCSC introduced **track data hubs** (2011)
    - integrate remote data sets into the browser

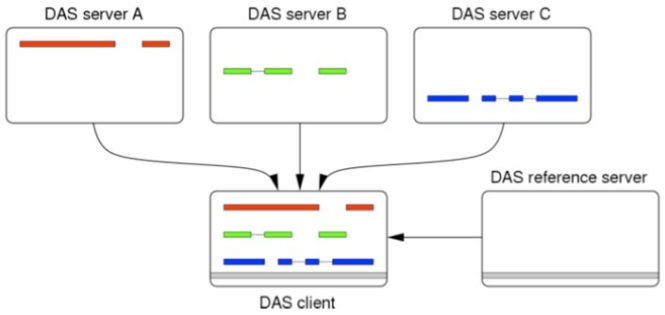
What are Track Hubs?

- Internet-accessible collections of genome annotations
- Alternative to DAS
  - no need for special software
  - WWW/FTP server and some text



How about DAS?

- Workhorse of external data integration (so far)
- Can no longer support new features or scale to modern data sets sizes
  - very high feature densities
  - query large feature-rich regions
  - fast zooming
- No longer supported (>=e!84)



Track Hubs

- Collate related data sets (tracks) through a single attachable URL
- Annotations in binary indexed file formats
  - partial downloads
  - caching
- Hosted on HTTP/FTP servers + text files

Hub Configuration

hub.txt

```
hub Blueprint_Hub_20150820
shortLabel Blueprint Hub
longLabel Blueprint Epigenomics Data Hub
genomesFile genomes.txt
email blueprint-info@ebi.ac.uk
descriptionUrl
http://www.blueprint-epigenome.eu/index.cfm?p=31AD6D30-9B3C-BB9
```

UCSC assembly name

genomes.txt

```
genome hg38
trackDb grch38/tracksDb.txt
```

```
track miRNA_new_pre
bigDataUrl
  https://www.broadinstitute.org/ftp/pub/vgb/dog/trackHub/canFam3/miRNA/cf3_miRNA_novel_pr
  ecursors.bb
shortLabel miR New Hairpins
longLabel Novel miRNAs - Hairpin Structures
html html/miRNA
priority 5.53
type bigBed 6
visibility pack
useScore 1


track miRNA_new_mat
bigDataUrl
  https://www.broadinstitute.org/
  ture.bb
shortLabel miR New Mature
longLabel Novel miRNAs - Mature
html html/miRNA
priority 5.54
type bigBed 6
visibility pack
useScore 1
```

hierarchical

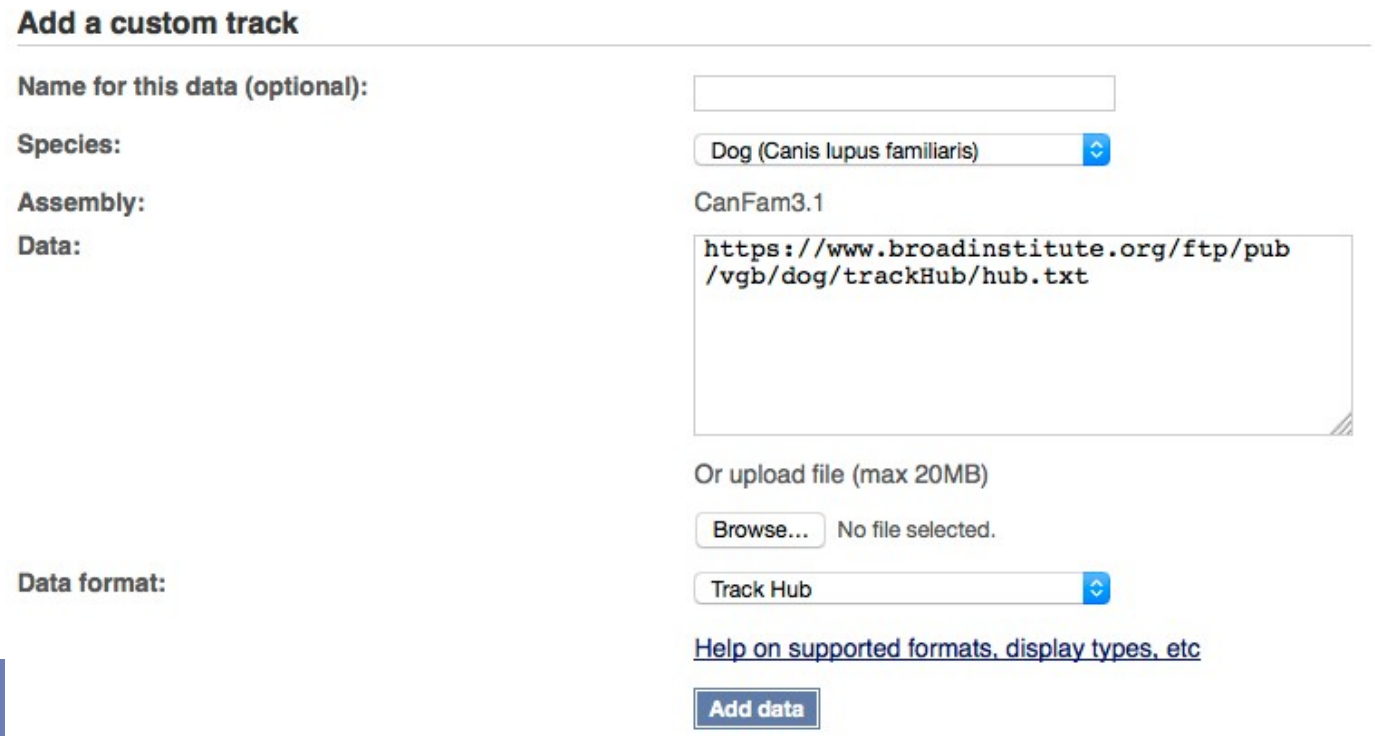
```
track CTVT_variation
superTrack on
group CTVT
shortLabel CTVT Variation
longLabel CTVT Variations from Germline and Somatic
html html/CTVT
priority 5.8

      track CTVT_ind_Som
      bigDataUrl
 https://www.broadinstitute.org/ftp/pub/vgb/dog/trackHub/canFam3/ostrander/CTVT_ind
 els_somatic.vcf.gz
      parent CTVT_variation
      shortLabel  CTVT Indels Somat
      longLabel  CTVT Indels Somatic
      html html/CTVT
      priority 5.81
      type vcfTabix
      visibility dense


   track CTVT_ind_Ger
      bigDataUrl
 https://www.broadinstitute.org/ftp/pub/vgb/dog/trackHub/canFam3/ostrander/CTVT_ind
 els_germline.vcf.gz
      parent CTVT_variation
      shortLabel  CTVT Indels Germl
      longLabel  CTVT Indels Germline
      html html/CTVT
      priority 5.82
      type vcfTabix
      visibility dense
```

Attaching a hub (Ensembl)

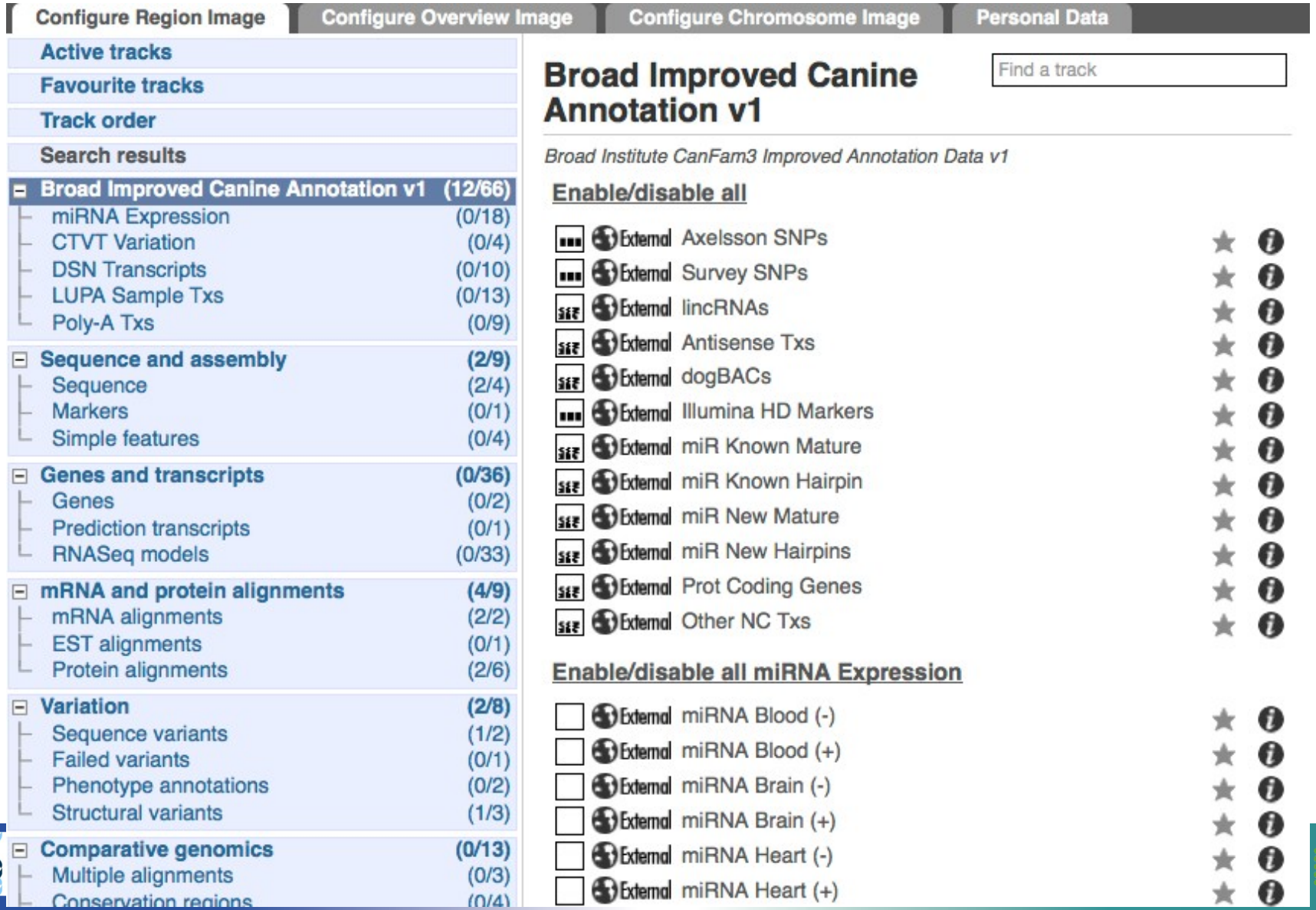
- Via a URL:

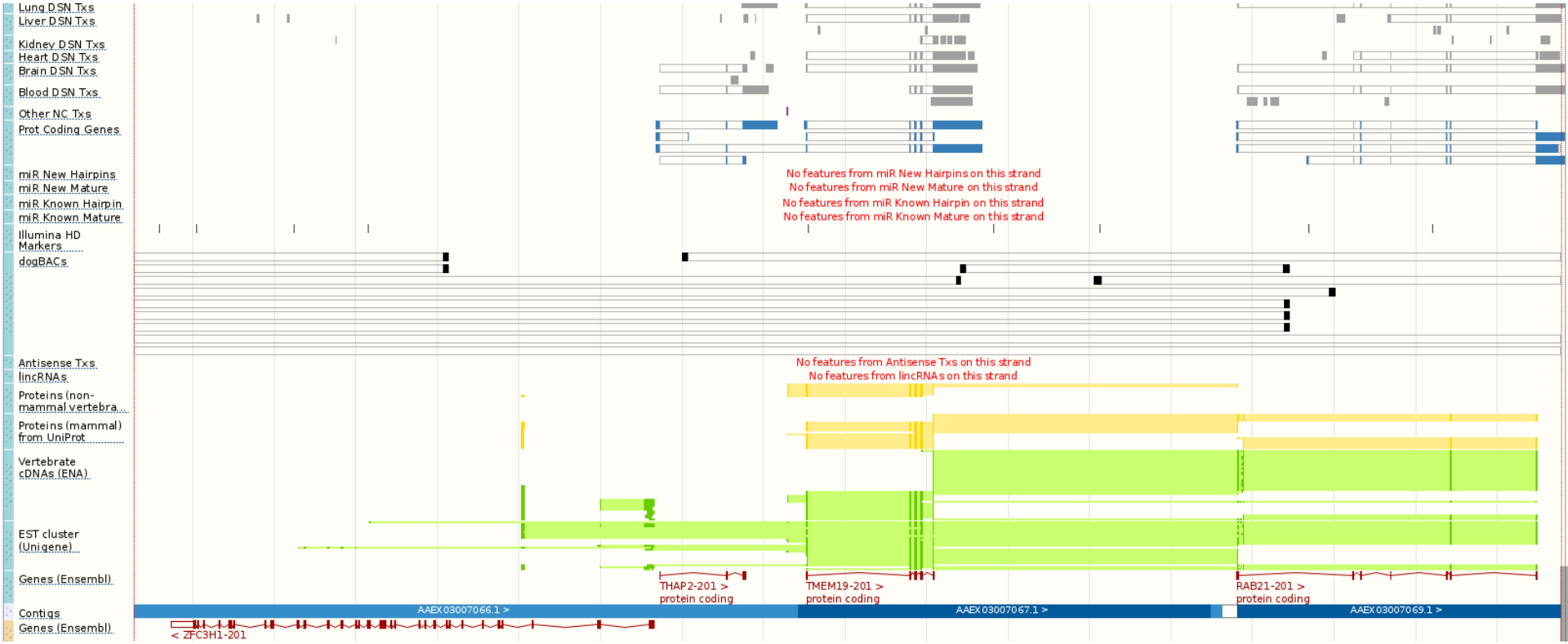
  /Trackhub?url=https://www.broadinstitute.org/ftp/pub/vgb/dog/trackHub/hub.txt

- Via Configuration Panel



Configuring a Hub




Lung DSN Txs
Liver DSN Txs
Kidney DSN Txs
Heart DSN Txs
Brain DSN Txs
Blood DSN Txs
Other NC Txs
Prot Coding Genes
miR New Hairpins
miR New Mature
miR Known Hairpin
miR Known Mature
Illumina HD Markers
dogBACs
Antisense Txs
lincRNAs
Proteins (non-mammal vertebra...
Proteins (mammal) from UniProt
Vertebrate cDNAs (ENA)
EST cluster (Unigene)
Genes (Ensembl)
Contigs
Genes (Ensembl)
No features from miR New Hairpins on this strand
No features from miR New Mature on this strand
No features from miR Known Hairpin on this strand
No features from miR Known Mature on this strand
No features from Antisense Txs on this strand
No features from lincRNAs on this strand
THAP2-201 > protein coding
TMEM19-201 > protein coding
RAB21-201 > protein coding
AAEX03007066.1 >
AAEX03007067.1 >
AAEX03007069.1 >
< ZFC3H1-201

Track Hubs: cons

- Browser integration
  - copy-paste of known URLs
- Discovery:
  - word of mouth
  - manually curated pages

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Track Data Hubs

Track data hubs are collections of external tracks that can be imported into the UCSC Genome Browser. Hub tracks show up under the hub's own blue label bar on the main browser page, as well as on the configure page. "Connect" button below.

NOTE: Because Track Hubs are created and maintained by external sources, UCSC is not responsible for their content.

Public Hubs | My Hubs

Enter search terms to find in public track hub description pages:

Search Public Hubs

Clicking Connect redirects to the gateway page of the selected hub's default assembly.

Display	Hub Name	Description	Assemblies
Connect	Roadmap Epigenomics Data Complete Collection at Wash U VizHub	Roadmap Epigenomics Human Epigenome Atlas Data Complete Collection, VizHub at Washington University in St. Louis	hg19
Connect	UMassMed ZHub	UMassMed H3K4me3 ChIP-seq data for Autistic brains	hg19
Connect	Cancer genome polyA site & usage	An in-depth map of polyadenylation sites in cancer (matched-pair tissues and cell lines)	hg19
Connect	ENCODE Analysis Hub	ENCODE Integrative Analysis Data Hub	hg19
Connect	miRcode microRNA sites	Predicted microRNA target sites in GENCODE transcripts	hg19
Connect	Translation Initiation Sites (TIS)	Translation Initiation Sites (TIS) track	hg19
Connect	SDSU NAT	Sense/antisense gene/exon expression using Affymetrix exon array from South Dakota State University, USA	hg19, mm9, rn4
Connect	DNA Methylation	Hundreds of analyzed methylomes from bisulfite sequencing data	[+] hg19, hg18, mm9, mm10, panTro2, danRer7...
Connect	Plants	CSHL Biology of Genomes meeting 2013 demonstration assembly hub	araTha1, braRap1, ricCom1
Connect	Blueprint Hub	Blueprint Epigenomics Data Hub	hg19
Connect	CEMT (CEEHRC)	Epigenomic Data tracks from BCGSC, Vancouver	hg19

Track Hub Registry

- Global collection of publicly available TrackHub servers
  - external parties advertise/publish THs
  - users discover interesting data
- Services:
  - hub registration (RESTful API)
  - search track hubs (WWW/API)
  - track hub still available

REST API

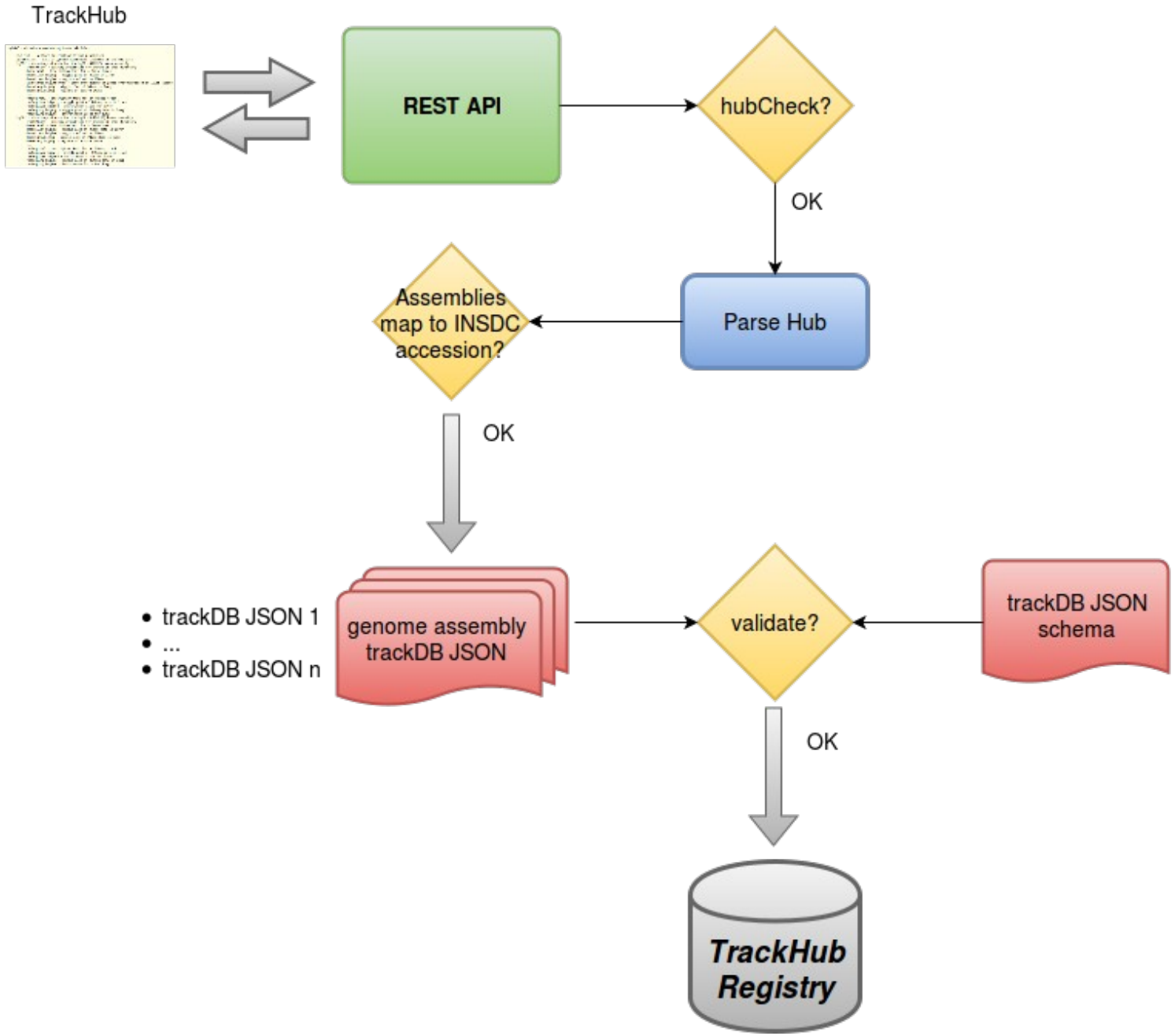
- Hub Registration, 8 endpoints (CRUD)
  - HTTP SSL-secured/Basic Authentication
  - token based authorisation
- Basic service information, 5 endpoints
  - e.g. server alive? species/assembly/hub list
- Search, 2 endpoints
  - genome browser support

POST /api/trackhub

- Register/update remote public hub
- **Requirement**: INSDC assemblies (GCA)
- Body:
  - Hub URL (required)
  - *(assembly name → GCA accession)* map (required if not UCSC native)
  - data type (optional)

```
POST https://www.trackhubregistry.org/api/trackhub
User: exampleuser
Auth-Token: 6l5/GuIiOSCywuSI9HF1VU97clwb/CXPDFS0MyAB/HCZuxtjQBj4uORZL8NY3Yhi
{
  "url": "http://genome-test.cse.ucsc.edu/~hiram/hubs/Plants/hub.txt",
  "assemblies": {
    "araTha1": 'GCA_000001735.1',
    "ricCom1": 'GCA_000151685.2',
    "braRap1": 'GCA_000309985.1'
  }
}
```

```
201 Created
Content-type: application/json; charset=utf-8
Location: [ 'https://www.trackhubregistry.org/api/trackdb/KRBr5PS7RmapaFr7ofpTBA, 'https://www.tra
...
[
  {
    // ricCom1 trackDb configuration
    'owner': 'exampleuser',
    'source': {
                'checksum': 'f9561ae6f7883add3698fad7abab7e13',
                'url': 'http://genome-test.cse.ucsc.edu/~hiram/hubs/Plants/ricCom1/trackDb.txt'
              },
    'hub': {
              'shortLabel': 'Plants',
              'name': 'cshl2013',
              'url': 'http://genome-test.cse.ucsc.edu/~hiram/hubs/Plants/hub.txt',
              'longLabel': 'CSHL Biology of Genomes meeting 2013 demonstration assembly hub'
            },
    'species': {
                 'scientific_name': 'Ricinus communis',
                 'common_name': 'castor bean',
                 'tax_id': '3988'
               },
    'assembly': {
                  'synonyms': 'ricCom1',
                  'name': 'JCVI_RCG_1.1',
                  'accession': 'GCA_000151685.2'
                },
    'configuration': {
```


TrackHub
REST API
hubCheck?
OK
Parse Hub
Assemblies map to INSDC accession?
OK
trackDB JSON 1
...
trackDB JSON n
genome assembly trackDB JSON
validate?
trackDB JSON schema
OK
TrackHub Registry

Web Front-end 

- Track hub aware, intuitive interface for searching interesting tracks
- Simple dashboard for track hub providers
- Track hub submission instructions and REST API docs

http://beta.trackhubregistry.org

The Track Hub Registry

A global centralised collection of publicly accessible track hubs

The goal of the Track Hub Registry is to allow third parties to advertise track hubs, and to make it easier for researchers around the world to discover and use track hubs containing different types of genomic research data.

Search by keywords: hg19, epigenomics, mouse ...

Submit Data

I want maximum visibility for my track hubs.

External track hub providers can register and submit their track databases to the registry. Registration is web-based and done on this site; submission happens programatically via our RESTful API. Once submitted and successfully validated, the track dbs become available for search by other users worldwide, allowing for automatic and rapid integration into a genome browser.

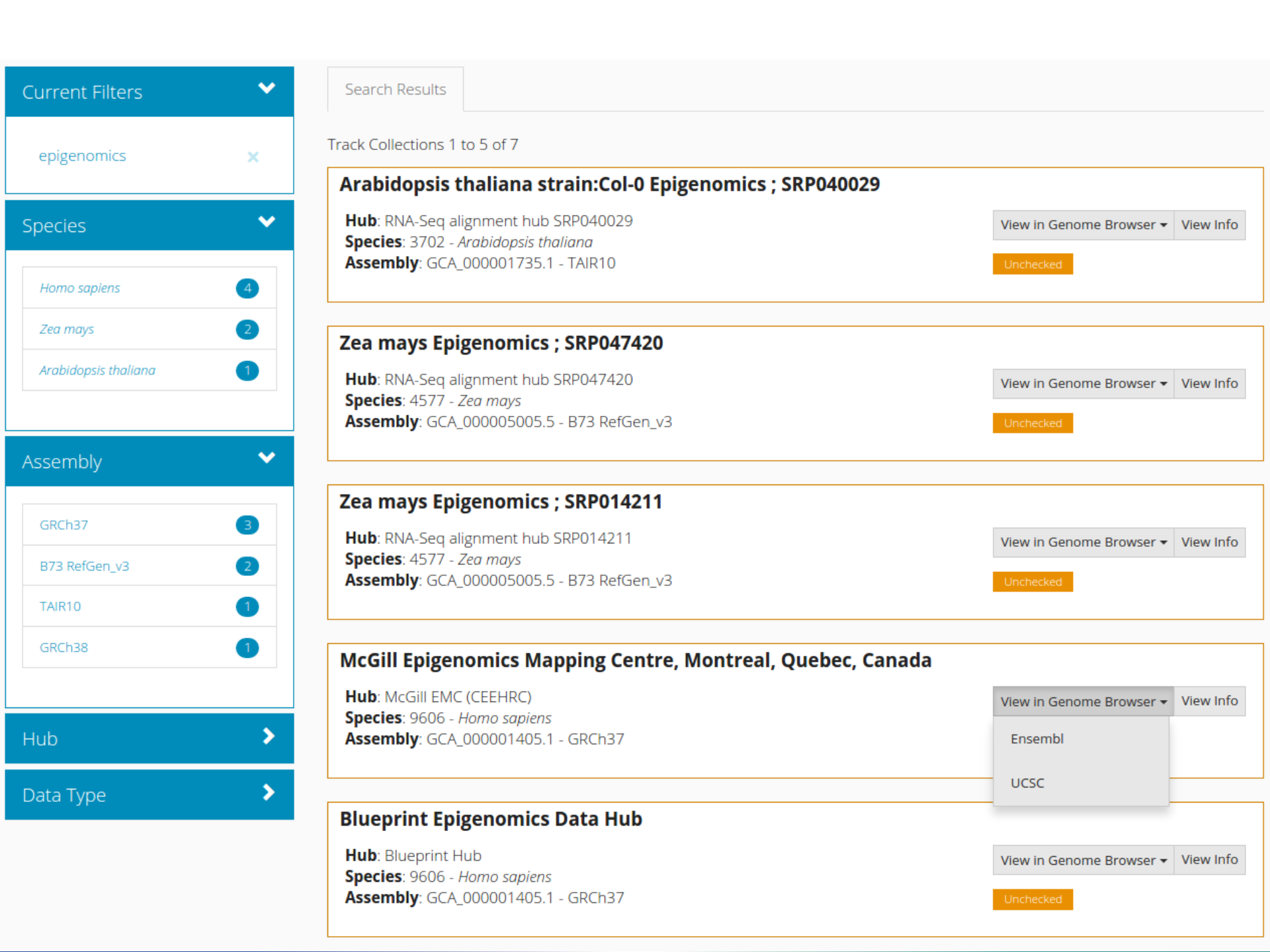
How to Submit

Access Data

How do I find omics tracks for an assembly of my favourite organism?

Track hubs can be searched based on metadata information. Free text search is provided from the search box in the header of all track hub Registry web pages and in the middle of this page. Advanced search options are available for more specific and customised searches.

Help on Advanced Search


Current Filters
epigenomics
Species
Homo sapiens 4
Zea mays 2
Arabidopsis thaliana 1
Assembly
GRCh37 3
B73 RefGen_v3 2
TAIR10 1
GRCh38 1
Hub
Data Type
Search Results
Track Collections 1 to 5 of 7
Arabidopsis thaliana strain:Col-0 Epigenomics ; SRP040029
Hub: RNA-Seq alignment hub SRP040029
Species: 3702 - Arabidopsis thaliana
Assembly: GCA_000001735.1 - TAIR10
View in Genome Browser
View Info
Unchecked
Zea mays Epigenomics ; SRP047420
Hub: RNA-Seq alignment hub SRP047420
Species: 4577 - Zea mays
Assembly: GCA_000005005.5 - B73 RefGen_v3
View in Genome Browser
View Info
Unchecked
Zea mays Epigenomics ; SRP014211
Hub: RNA-Seq alignment hub SRP014211
Species: 4577 - Zea mays
Assembly: GCA_000005005.5 - B73 RefGen_v3
View in Genome Browser
View Info
Unchecked
McGill Epigenomics Mapping Centre, Montreal, Quebec, Canada
Hub: McGill EMC (CEEHRC)
Species: 9606 - Homo sapiens
Assembly: GCA_000001405.1 - GRCh37
View in Genome Browser
View Info
Ensembl
UCSC
Blueprint Epigenomics Data Hub
Hub: Blueprint Hub
Species: 9606 - Homo sapiens
Assembly: GCA_000001405.1 - GRCh37
View in Genome Browser
View Info
Unchecked

McGill EMC (CEEHRC) | Homo sapiens | GRCh37

General Info

Total number of tracks	1189
Remote data tracks	1185
Data type: genomics	
File type(s): BigBed, BigWig	
Source URL: View	

Unchecked

View in Genome Browser

Hub

Name: McGill_EMC_Hub

Short Label: McGill EMC (CEEHRC)

Long Label: McGill Epigenomics Mapping Centre, Montreal, Quebec, Canada

Assembly Hub: ✖

Public URL: View

Species

Taxonomy 9606
Scientific name: *Homo sapiens*
Common name: human

Assembly Information

Accession	Name	Long Name	Synonyms
GCA_000001405.1	GRCh37	Genome Reference Consortium Human Build 37 (GRCh37)	hg19

Registration API

Authentication

The Registry requires requests to the Registration API to be authenticated. The Registration API implements Basic HTTP Authentication, as defined by RFC 2617, which enables a client to authenticate individual HTTP requests by including an authentication header in the request. In order to make authorized calls to the API, your application must first obtain an access token, and use that token in the authentication header of all subsequent requests.

A Simplified RESTful workflow

API clients and the Registry communicates over HTTPS, exchanging JSON representations of API objects.

This simplified example of a RESTful workflow includes requests to register remote track hubs, list registered track data hubs with the Registry and update a registered track hub.

These examples assume the track hub data provider using the API has signed up and created an account with the Registry. To create an account, go to registration.

Logging in

All requests to the service must be authenticated. The first step in any Registration API RESTful workflow is to obtain an authentication token.

Registering track hubs

You've just signed up and you've got some remote public track hubs that you want to register with and make available for search on the Track Hub Registry.

Retrieve the list of registered track hubs

You've submitted some of your public track hubs to the Registry. You want to know which ones by retrieving the list of registered track hubs from the Registry.

Update registered track hubs

You've updated the structure or content of one of your remote public hubs registered with us. You obviously want the changes to appear on the Registry as well.

Delete registered track hubs

One of your remote public hubs does not exist any more. Or you simply don't want to make it available for search in the Track Hub Registry.

Log out

This terminates the client session and ends any possible workflow of interaction between the client and the Registration API.

Example Clients

Perl | Python | Ruby

```python
    r = requests.get(server+'/api/logout', headers={ 'user': user, 'auth_token': auth_token })
    if not r.ok:
        print "Couldn't logout, reason: %s [%d]" % (r.text, r.status_code)
        sys.exit
    print 'Logged out'

auth_token = login(server, user, password)
headers = { 'user': user, 'auth_token': auth_token }
payload = { 'url': hub_url, 'assemblies': { 'araTha1': 'GCA_000001735.1', 'ricCom1': 'GCA_000151685
r = requests.post(server+'/api/trackhub', headers=headers, json=payload, verify=False)
if not r.ok:
    print "Couldn't register track hub at %s, reason: %s [%d]" % (hub_url, r.text, r.status_code)
    sys.exit
print "I have registered hub at %s" % hub_url

logout(server, user, auth_token)

```

GET /api/login

Authenticate the client and obtain an access token in order to make subsequent requests to the Registration API.

If the request is successful, the response is formatted as a JSON object with a single key (`auth_token`), whose value is the access token. This token must be included as an `Auth-Token` header in all subsequent requests.

Resource Information

Response formats	JSON
Authentication	Basic, MIME Base64
Rate Limited	No

Parameters

None.

Example Request

```
GET https://www.trackhubregistry.org/api/login
Authorization: Basic ZXhhbXBsZXVzZXI6ZXhhbXBsZXBhc3N3b3Jk
```

Example Response

```
HTTP/1.0 200 OK
Content-type: application/json; charset=utf-8
...
{
  "auth_token":"615/GuIiOSCywuSI9HF1VU97clwb/CXPDFS0MyAB/HCZuxtjQBj4uORZL8NY3Yhi"
}
```

HTTP Status Codes

Code	Description	Reason
200	OK	Request successful
401	Unauthorized	The request requires user authentication
500	Internal Server Error	Request cannot be fulfilled due to unexpected condition
503	Service Unavailable	Request cannot be fulfilled due to temporary overloading or maintenance of the server

Ensembl as a Registry Client

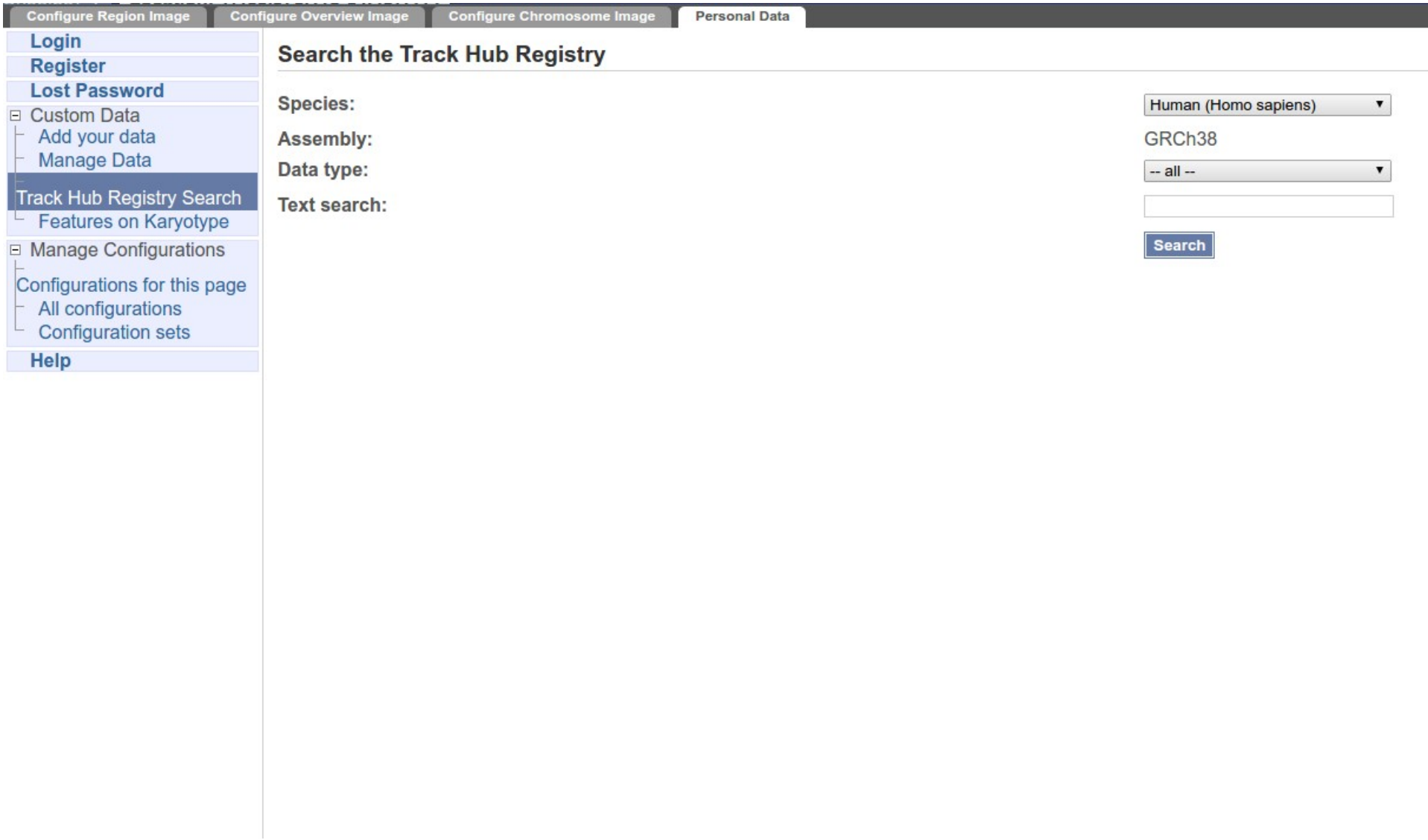
Phase 1 – Spring '16

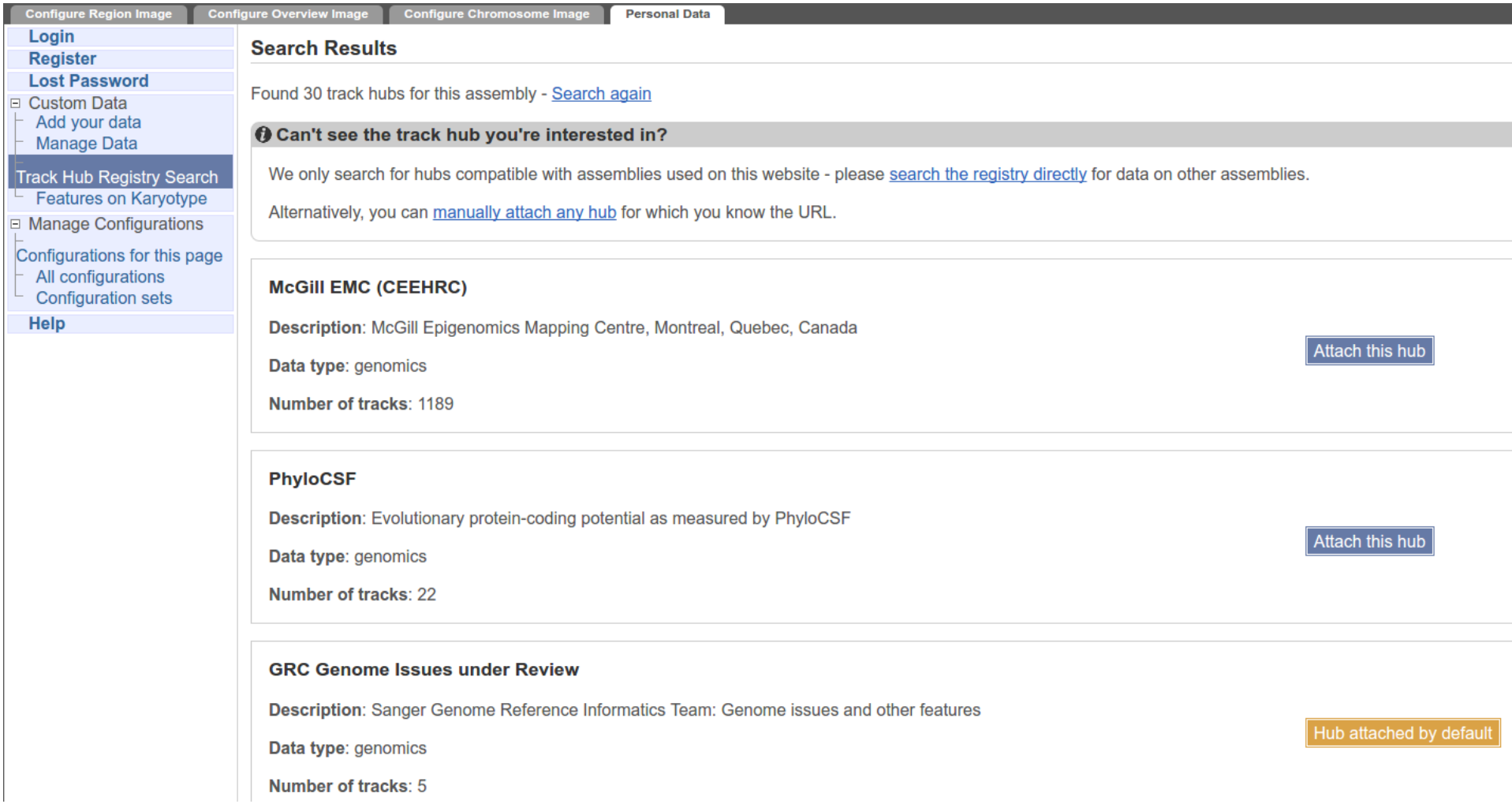
Search interface using Registry API

- one-click attachment of chosen hub
- text files still retrieved/parsed by web server

Phase 2 – Spring '16

Ensembl consumes registry's hub JSON


Configure Region Image
Configure Overview Image
Configure Chromosome Image
Personal Data
Login
Register
Lost Password
Custom Data
Add your data
Manage Data
Track Hub Registry Search
Features on Karyotype
Manage Configurations
Configurations for this page
All configurations
Configuration sets
Help
Search the Track Hub Registry
Species:
Human (Homo sapiens)
Assembly:
GRCh38
Data type:
-- all --
Text search:
Search


Configure Region Image
Configure Overview Image
Configure Chromosome Image
Personal Data
Login
Register
Lost Password
Custom Data
Add your data
Manage Data
Track Hub Registry Search
Features on Karyotype
Manage Configurations
Configurations for this page
All configurations
Configuration sets
Help
Search Results
Found 30 track hubs for this assembly - Search again
Can't see the track hub you're interested in?
We only search for hubs compatible with assemblies used on this website - please search the registry directly for data on other assemblies.
Alternatively, you can manually attach any hub for which you know the URL.
McGill EMC (CEEHRC)
Description: McGill Epigenomics Mapping Centre, Montreal, Quebec, Canada
Data type: genomics
Number of tracks: 1189
Attach this hub
PhyloCSF
Description: Evolutionary protein-coding potential as measured by PhyloCSF
Data type: genomics
Number of tracks: 22
Attach this hub
GRC Genome Issues under Review
Description: Sanger Genome Reference Informatics Team: Genome issues and other features
Data type: genomics
Number of tracks: 5
Hub attached by default

Present/Future Directions

- Stats:
  - track hubs: 1173
  - species: 78
  - assemblies: 83
- Out of Beta soon!
- Embed Biodalliance
- Track hub RDF
  - link hub to other data via ontology mapping
  - query hubs via SPARQL

Acknowledgements

Ensembl

Anne Lyle, Steve Trevanion

Andy Yates

Magali Ruffier

Paul Flicek and the entire e! Team

UCSC

Kate Rosenbloom

Funding

wellcome trust

EMBL

National Human Genome Research Institute

BBSRC bioscience for the future

European Commision FP7

SEVENTH FRAMEWORK PROGRAMME