

Presentation of bioinformatics resources for gene regulation

Max 3 minutes per resource

GREEKC Training-hackathon
Marseille (France) Apr 23-26, 2019



Andrew Parton

23 April 2019

GreekC Hackathon Marseille

Data in Ensembl



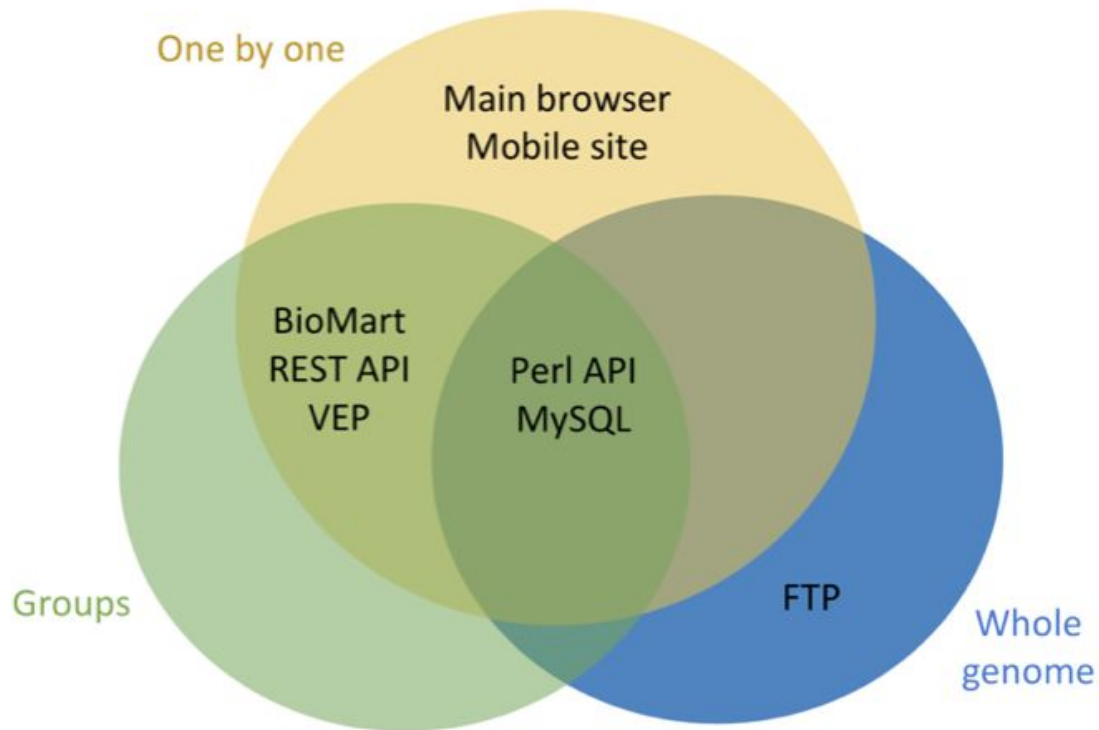
What is Ensembl?

- Genome Browser, providing a single point of access to annotated genomes
- Aim of Ensembl used to be to annotate human genome with gene models and other available data
- Genome annotation
- Added in species - human/mouse/birds/reptiles/fish
- EG species – fungi, plants, bacteria, metazoa

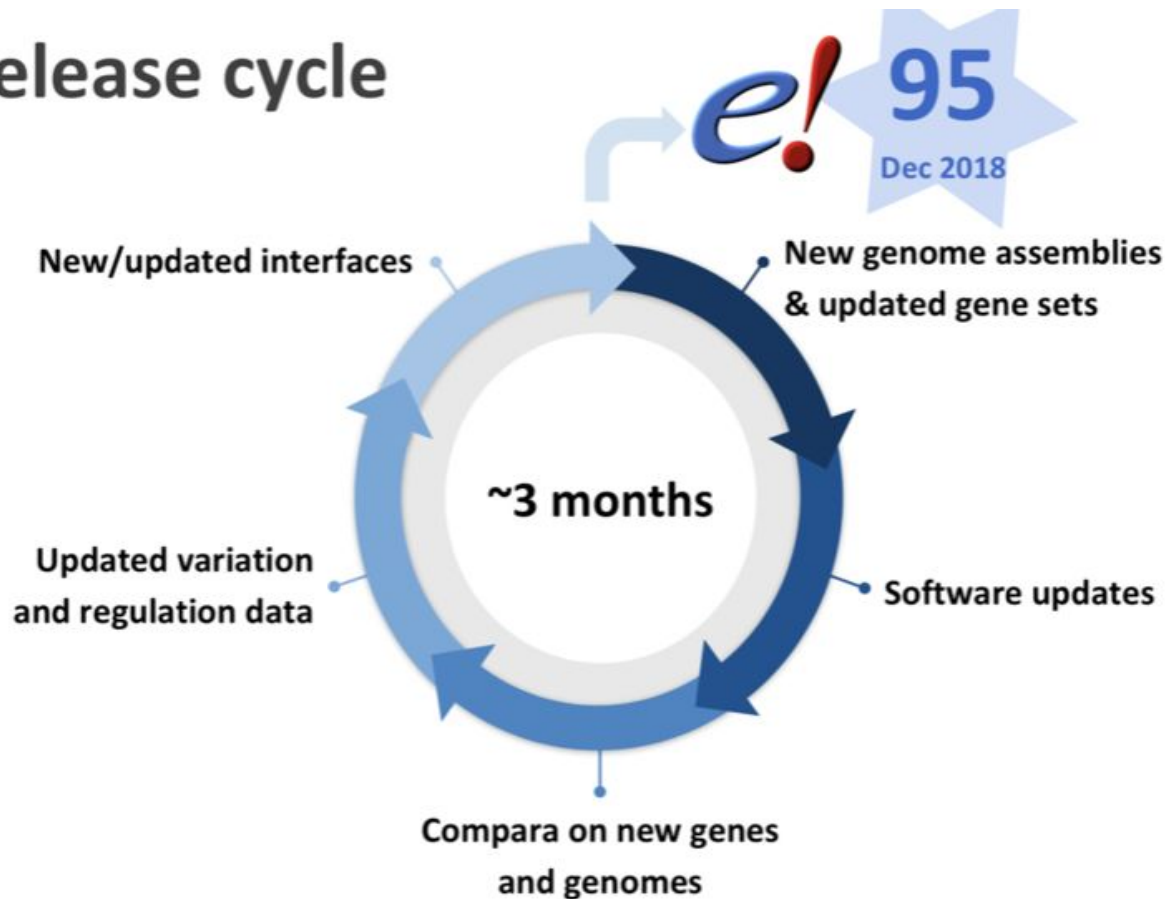
Data in Ensembl

- Genome assemblies for ~140 species
- Gene Trees
- Regulatory Build (Human and mouse)
- Variation display and VEP
- BioMart (Custom bulk data export)
- Programmatic access via the APIs
- Completely open source

Many ways to access Ensembl data



Release cycle



The goal of the Ensembl Regulation team is to annotate the genome with features that may play a role in regulation of gene expression.








- Predicted open/closed chromatin
 - DNase I sensitivity
- Transcription factor binding sites
- Epigenetic marks
 - Histone modifications
 - DNA methylation



- **Regulatory regions** \Rightarrow part of the DNA sequence
 - Promoters & flanking regions
 - Enhancers
 - CTCF binding sites
 - Transcription factor binding sites
 - Open chromatin regions
- **Epigenetic marks** \Rightarrow base DNA sequence unchanged
 - DNA methylation data
 - Histone modifications



Where does the data come from?

| Species | Data source | Assay types | Epigenomes |
|---|--|--|--|
|  |   | <ul style="list-style-type: none">• ChIP-seq (histone mods)• TF binding sites• RNApol• DNase sensitivity (open chromatin) | Cultured cell lines |
|  |  | <ul style="list-style-type: none">• ChIP-seq (histone mods)• TF binding sites• RNApol• DNase sensitivity (open chromatin) | Cultured cell lines |
|  |  | <ul style="list-style-type: none">• ChIP-seq (histone mods)• DNase sensitivity (open chromatin) | Primary cells from haematopoietic cell lineage (direct from human cells) |

What does Ensembl not do?

- We do not link regulatory features to genes
 - We allow you see the location of features.
- We do not link regulatory features to gene expression.
 - We have cell-line specific regulation data and tissue specific expression data

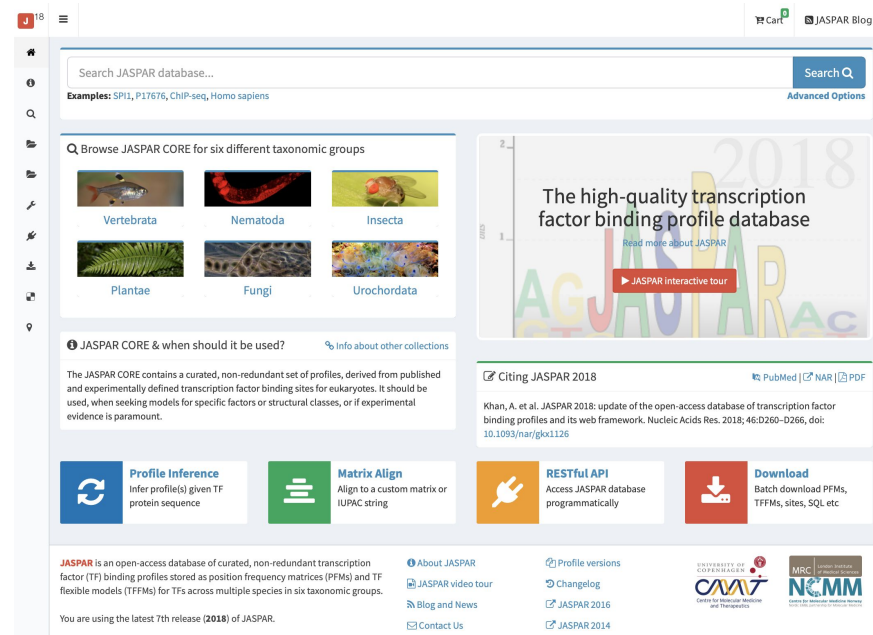
You are required to make your own inferences about this data

Regulatory data is incredibly complex and still in relative infancy. There is no comprehensive database of regulation data...yet!

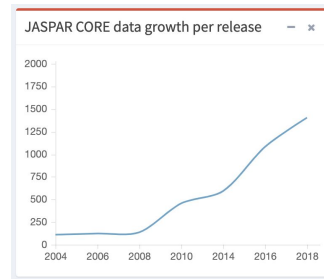
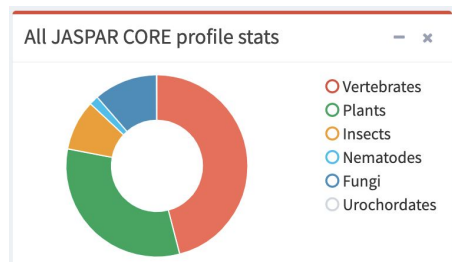
Thank you

JASPAR - database for eukaryotic TF binding profiles

- URL: jaspar.genereg.net
- TF profiles stored as position frequency matrices (PFMs) and TF flexible models (TFFMs)
- TFs across multiple species in six taxonomic groups
- 2018 release has 1404 non-redundant PFMs
 - 579 for vertebrates, 489 for plants, 176 for fungi, 133 for insects, 26 for nematodes and 1 for urochordata
 - >1.8 million pageviews by >100K users from 144 countries
- Access to data
 - REST API: jaspar.genereg.net/api
 - Download: <http://jaspar.genereg.net/downloads>
 - R/Bioconductor package: **JASPAR2018**



The screenshot shows the JASPAR database homepage. At the top, there is a search bar with the text "Search JASPAR database..." and a "Search" button. Below the search bar, there are examples of search terms: "SPI1, P17676, ChIP-seq, Homo sapiens". The main content area is divided into several sections. On the left, there is a "Browse JASPAR CORE for six different taxonomic groups" section with images and labels for Vertebrata, Nematoda, Insecta, Plantae, Fungi, and Urochordata. To the right of this is a large banner for "The high-quality transcription factor binding profile database" with a "JASPAR interactive tour" button. Below the banner, there is a section for "Citing JASPAR 2018" with a link to a PubMed entry. At the bottom, there are several buttons for "Profile Inference", "Matrix Align", "RESTful API", and "Download". The footer contains links for "About JASPAR", "JASPAR video tour", "Blog and News", "Contact Us", "Profile versions", "Changelog", "JASPAR 2016", and "JASPAR 2014".



UniBind - A map of direct TF-DNA interactions in the human genome

- URL: unibind.uio.no
- A comprehensive map of direct interactions between transcription factor (TFs) and DNA.
- High confidence TF binding site predictions were obtained from uniform processing of thousands of ChIP-seq data
- UniBind covers >4% of the human genome
- Access to data
 - Download: unibind.uio.no/downloads
- Hackathon project
 - Develop a fully functional REST API

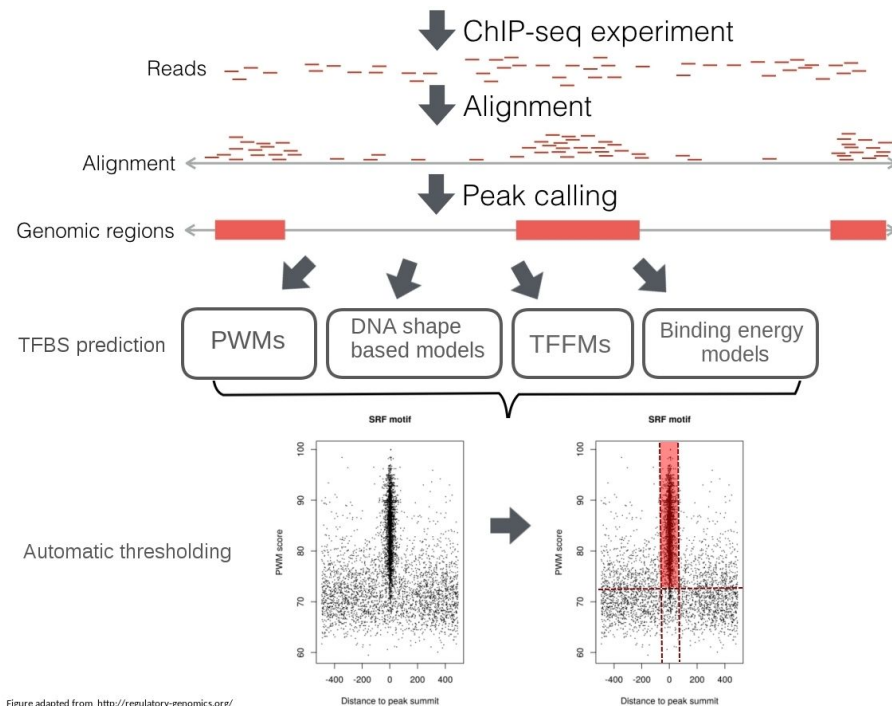


Figure adapted from <http://regulatory-genomics.org/>

1983

ChIP-seq datasets



315

Cell lines & Tissues



231

Transcription Factors



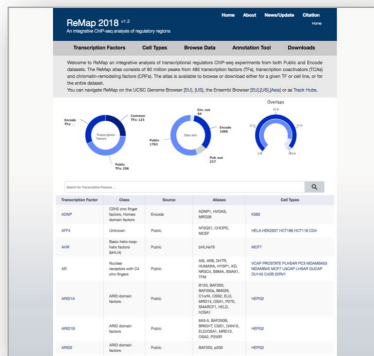
4

Prediction models



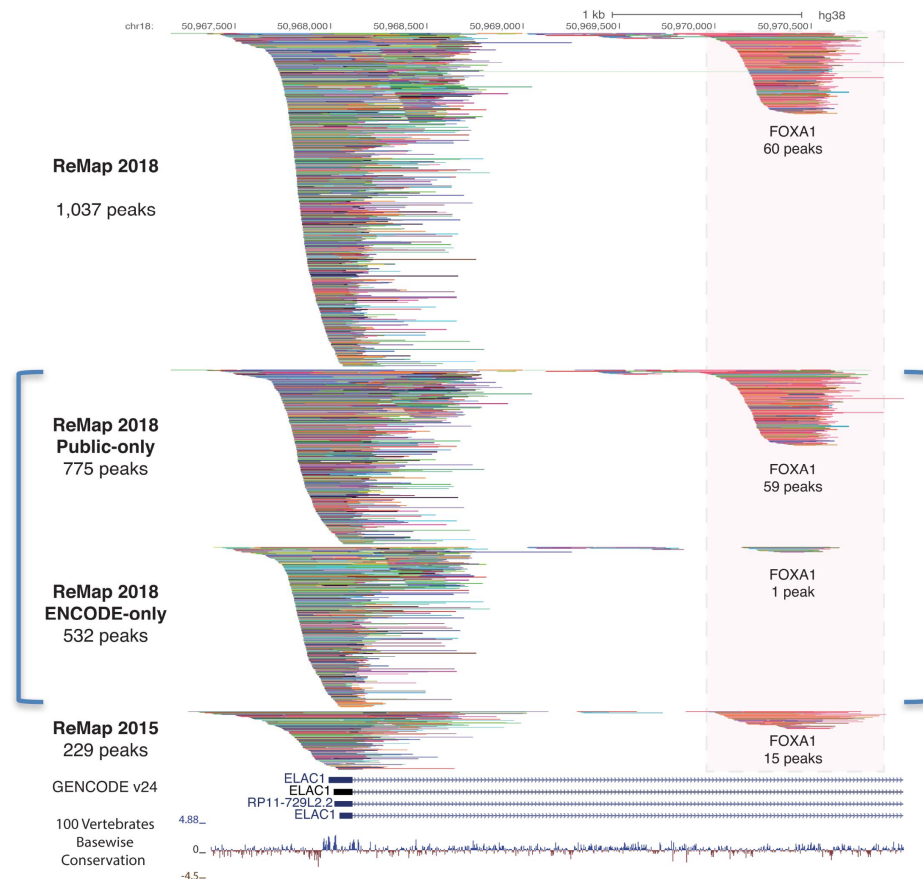
ReMap : Catalogue of ChIP-seq peaks for TR

ReMap website



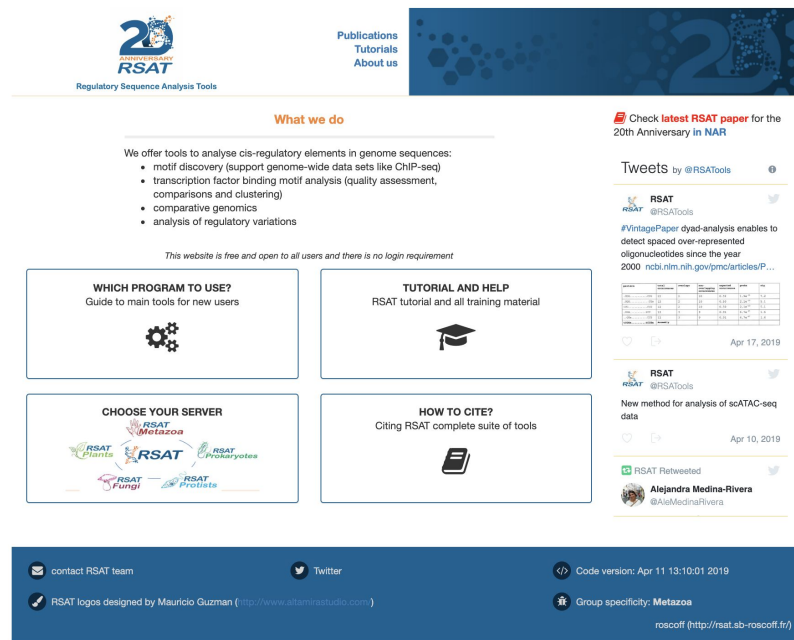
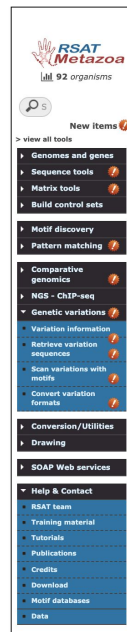
<http://remap.cisreg.eu>

- Public data (GEO + ENCODE)
 - manual curation
- 485 TR (mainly TFs)
- 346 cell lines
- 80 millions peaks



RSAT – Regulatory Sequence Analysis Tools

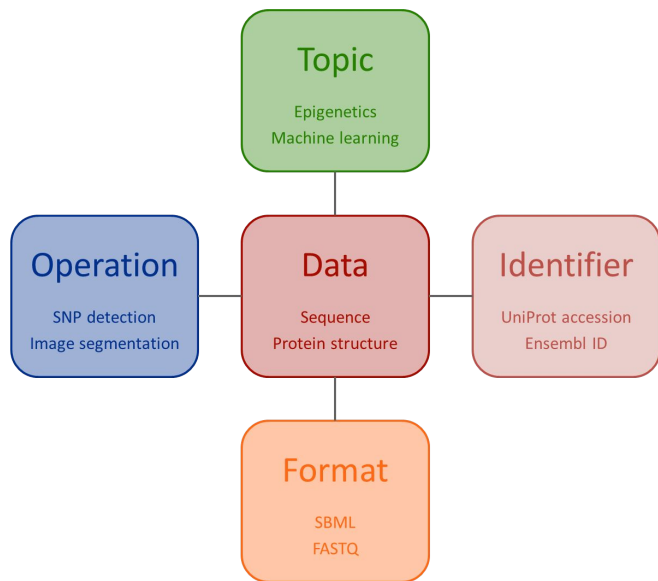
- URL: <http://rsat.eu/>
- Software suite for the detection of cis-regulatory elements in non-coding sequences.
- 6 servers worldwide
 - Bacteria, Fungi, Protists, Plants, Metazoa
 - Teaching
- Access modes
 - Web interface: 52 tools
 - Command line: many more
 - SOAP/WSDL Web services since 2008
 - REST Web services: in development
(will evolve during the hackathon)



EDAM ontology (<http://edamontology.org>)

EDAM bioinformatics operations, types of data, data formats, identifiers, and topics

Last uploaded: July 30, 2018



3,354 bioinformatics concepts

Summary

Classes

Properties

Notes

Mappings

Widgets

Jump to:

- Data
- DeprecatedClass
- Format
- Operation
 - Alignment
 - Analysis
 - Enrichment analysis
 - Expression analysis
 - Genetic variation analysis
 - Image analysis
 - Pathway or network analysis
 - Expression profile pathway mapping
 - Gene regulatory network analysis**
 - Pathway or network comparison
 - Pathway or network prediction
 - Pathway or network visualisation
 - Protein interaction network analysis
 - Weighted correlation network analysis
 - Phylogenetic tree analysis
 - Protein function analysis
 - Sequence analysis
 - Spectral analysis
 - Structure analysis
 - Text mining
 - Transmembrane protein analysis
- Annotation
- Calculation
- Classification
- Clustering

Details

Visualization

Notes (0)

Class Mappings (2)



| | |
|----------------|--|
| Preferred Name | Gene regulatory network analysis |
| Definitions | Analyse a known network of gene regulation. |
| ID | http://edamontology.org/operation_1781 |
| Created in | beta12orEarlier |
| hasDefinition | Analyse a known network of gene regulation. |
| inSubset | http://purl.obolibrary.org/obo/edam#edam http://purl.obolibrary.org/obo/edam#operations |
| label | Gene regulatory network analysis |
| prefixIRI | operation_1781 |
| prefLabel | Gene regulatory network analysis |
| subClassOf | Pathway or network analysis |

<https://bioportal.bioontology.org/ontologies/EDAM>

EDAM ontology browser (<https://ifb-elixirfr.github.io/edam-browser>)

← → ↺ https://ifb-elixirfr.github.io/edam-browser/#topic_0204

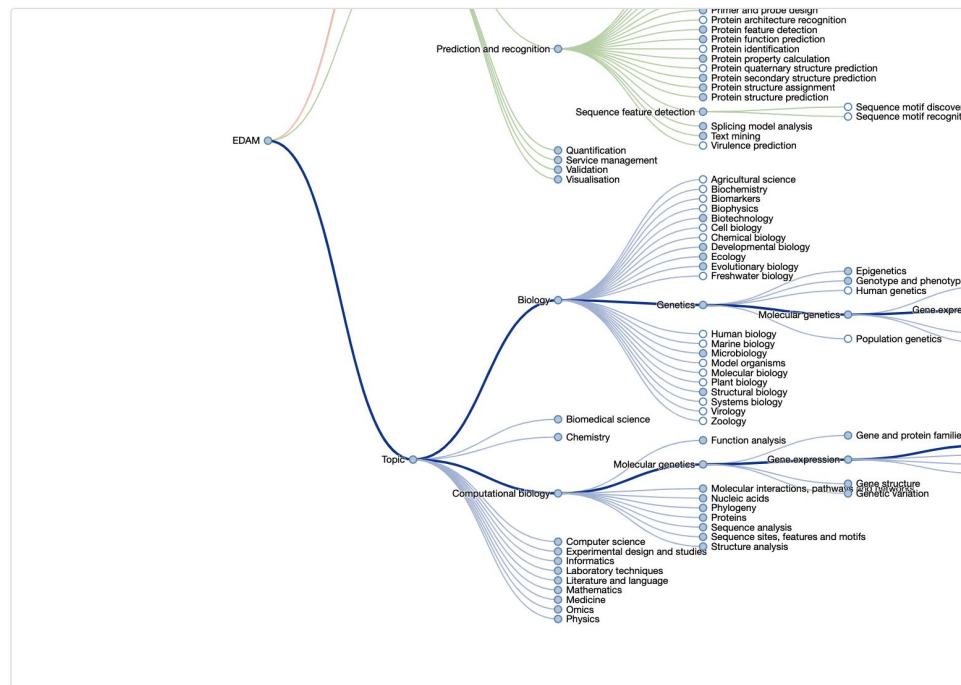


EDAM ontology

EDAM ▼ Custom

Gene regulation

EDAM is a simple ontology of well established, familiar concepts that are prevalent within bioinformatics [edamontology.org]



Collapse not selected elements

Reset zoom



1.21_dev

Details of term "Gene regulation"



| | |
|-----------------|---|
| Term | Gene regulation |
| Definition | The regulation of gene expression. |
| Comment | |
| Exact synonyms | |
| Narrow synonyms | Regulatory genomics |
| URI | http://edamontology.org/topic_0204 |
| Parents | Gene expression Gene expression |
| Community usage | |
| bio.tools | 156 times |
| Biosphere | 1 times by appliances, 1 times by tools. |
| BioWeb | not used |
| TeSS | 1 times |
| Links | Open in AberOWL , BioPortal , OLS or WebProtégé . |

Previously seen:

Details of term "Sequence motif discovery"






JASPAR RESTful API (biotools:jaspar_api)

ID Verified

<http://jaspar.genereg.net/api/>

[Data acquisition >](#)
[DNA >](#)
[Protein interactions >](#)

GPL-3.0

[Web API](#)
[Python](#)






1

9

Widely used open-access database of curated, non-redundant transcription factor binding profiles. Currently, data from JASPAR can be retrieved as flat files or by using programming language-specific interfaces. Here, we present a programming language-independent application programming interface (API) to access JASPAR data using the Representational State Transfer (REST) architecture.

[Query and retrieval >](#)

Credits & Support

[Anthony Mathelier](#)
 Primary contact | ✉ anthony.mathelier@ncmm.uio.no

[Aziz Khan](#)
 Primary contact | ✉ aziz.khan@ncmm.uio.no

Documentation

[General >](#)

12198 registered tools

- URL: <https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php>
- Tool for exploring annotations of non-coding genome at variant on haplotype blocks, such as candidate regulatory SNPs at disease-associated loci.
- Access modes
 - Web interface
 - R package (haploR)

HaploReg v4.1



Broad Institute
Homepage

HaploReg is a tool for exploring annotations of the noncoding genome at variants on haplotype blocks, such as candidate regulatory SNPs at disease-associated loci. Using LD information from the 1000 Genomes Project, linked SNPs and small indels can be visualized along with chromatin state and protein binding annotation from the Roadmap Epigenomics and ENCODE projects, sequence conservation across mammals, the effect of SNPs on regulatory motifs, and the effect of SNPs on expression from eQTL studies. HaploReg is designed for researchers developing mechanistic hypotheses of the impact of non-coding variants on clinical phenotypes and normal variation.

Update 2015.11.05: Version 4.1 GWAS and eQTL have been updated; a simpler pruning strategy is applied when combining GWAS; and links out to other NHGRI/EBI GWAS hits and GRASP QTL hits are provided.

Update 2015.09.15: Version 4.0 now includes many recent eQTL results including the GTEx pilot, four different options for defining enhancers using Roadmap Epigenomics data, and a complete set of source files for download and local analysis. Older versions available: [v3](#), [v2](#), [v1](#).

[Build Query](#) [Set Options](#) [Documentation](#)

Use one of the three methods below to enter a set of variants. If an r^2 threshold is specified (see the Set Options tab), results for each variant will be shown in a separate table along with other variants in LD. If r^2 is set to NA, only queried variants will be shown, together in one table.

Query (comma-delimited list of rsIDs OR a single region as chrN:start-end):

or, upload a text file (one refSNP ID per line): Choisir un fichier Aucun fichier choisi

or, select a GWAS:

[Valider](#)

-