# Project presentations

# 5 minutes per project + 3 minutes for questions

GREEKC Training-hackathon
Marseille, Apr 23-26, 2019

# Links

- Project descriptions (editable on github)
  - https://github.com/GREEKC/hackathon-marseille/blob/master/project_descriptions/
- Web pages
  - https://greekc.github.io/hackathon-marseille/project_descriptions/
- Schedule
  - https://greekc.github.io/hackathon-marseille/schedule.html

# Projects

1. Interfacing ReMap2020 with REST API
2. Adding the RESTful interface to UniBind
3. RESTful interface for RSAT
4. Annotation of GREEKC-related tools and data types in bio.tools and EDAM ontology
5. PSICQUIC for causal interactions and network modelling
6. Transcription Factor - Target Gene regulatory interactions inference from multiomics data integration
7. Workflow for the analysis of disease-associated regulatory variants

# Suggestions for project descriptions

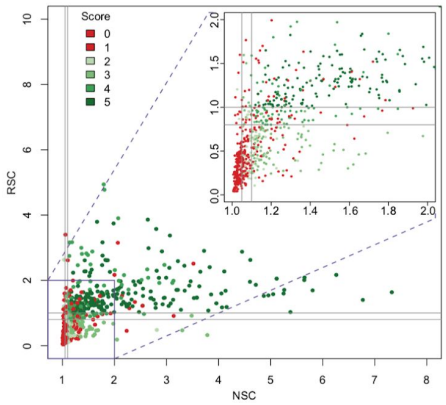
- Duration: 5’ presentation + 3’ questions per project
- For each project, please document (in the way you want)
  - Proponents
  - Motivation (context, state of the art, current limitations, …)
  - General goal
  - Detailed expectations and deliverables  (at the end of the hackathon we aim at …)
  - Resources relevant to the project (with links to their home pages or publications)
  - Expected skills and types of contributions to the project
  - Participants (hackers already recruited for some session), knowing that additional participants are free to join
  - Organisation of the sessions (what will be done at each session)

# Interfacing ReMap2020 with REST API

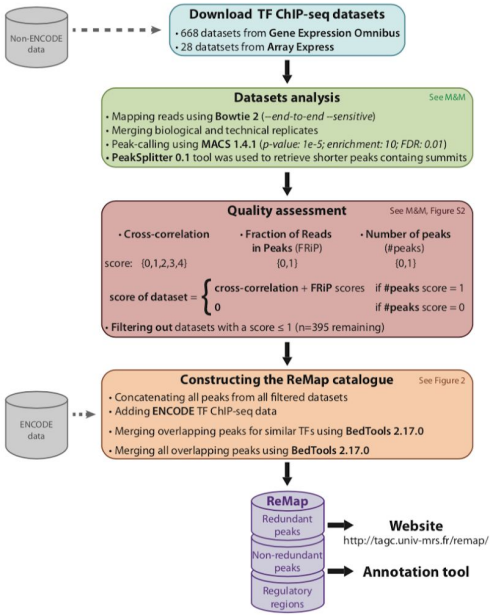
# ReMap 2015 : Data & Methods

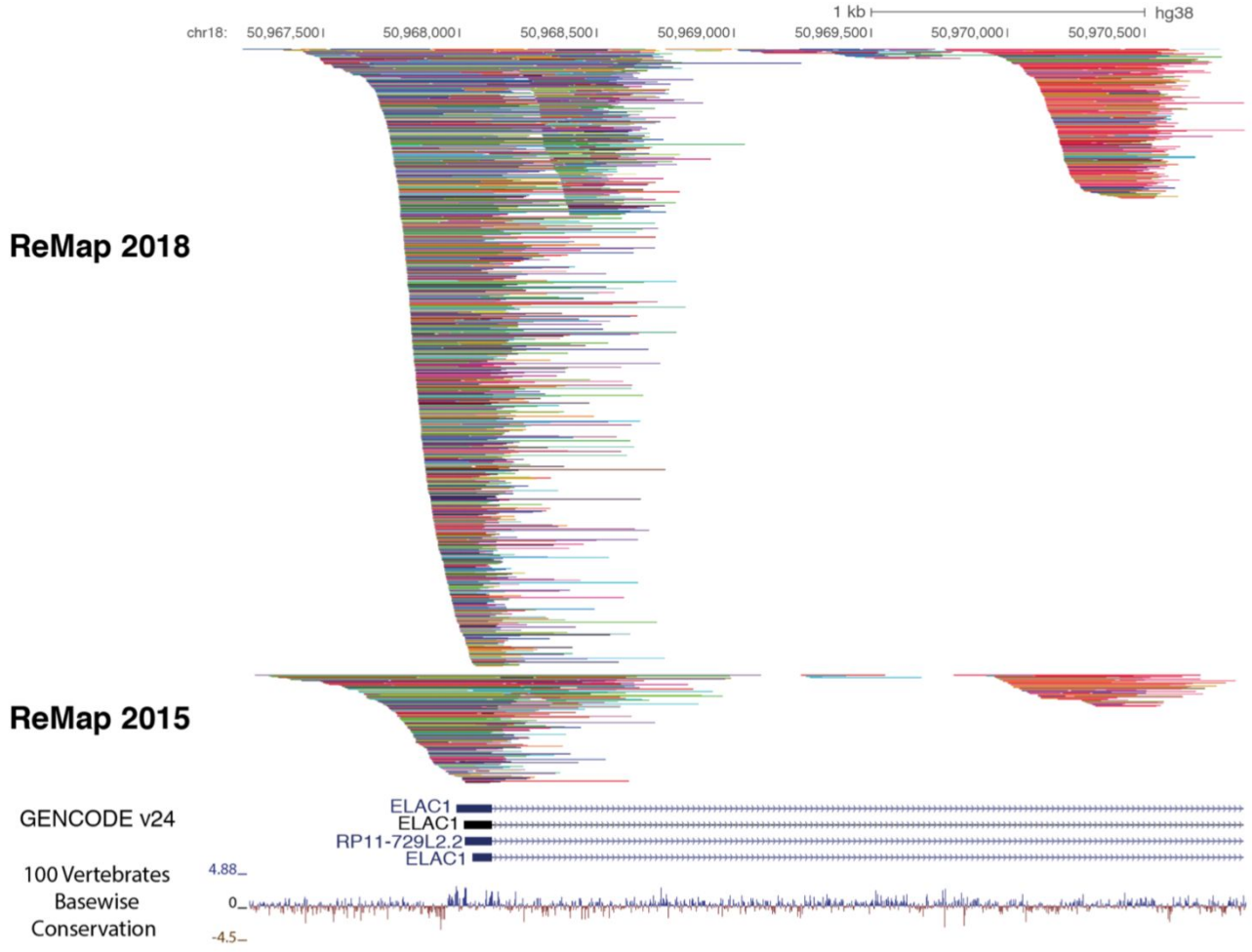
Public GEO/AE ChIP-seq:

- 696 datasets analyzed (GEO / AE)
- 395 retained for catalogue
- 132 Transcription factors




Marinov, et al. Large-scale quality analysis of published ChIP-seq data. G3, 2014



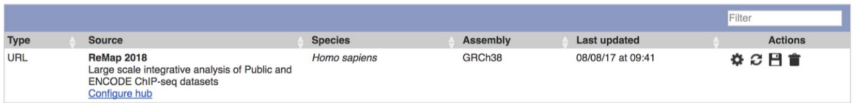

1 kb
hg38
chr18:
50,967,500
50,968,000
50,968,500
50,969,000
50,969,500
50,970,000
50,970,500
ReMap 2018
ReMap 2015
GENCODE v24
ELAC1
ELAC1
RP11-729L2.2
ELAC1
100 Vertebrates
Basewise
Conservation
4.88
0
-4.5

# ReMap 2018 : Accessibility

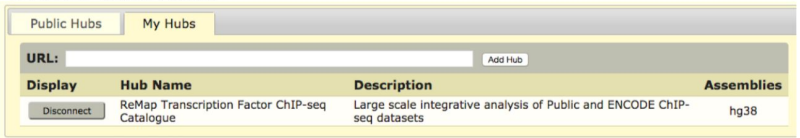
Accessibility is key to our data

## ReMap website

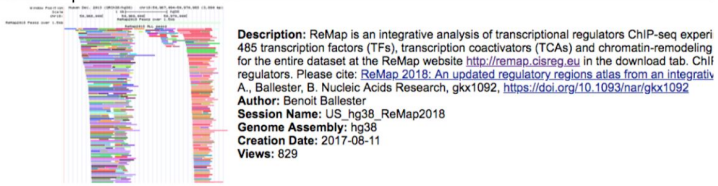

## Ensembl track hub




## UCSC track hub




## UCSC public session



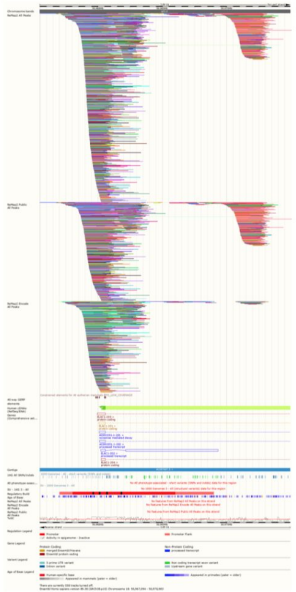

REST Accessibility ?

What does users, or power users (RSAT, JASPAR, HOCOMOCO) want from ReMap DB

# ReMap 2018 : Browsing data

Made our data available to all main browsers

UCSC



Ensembl

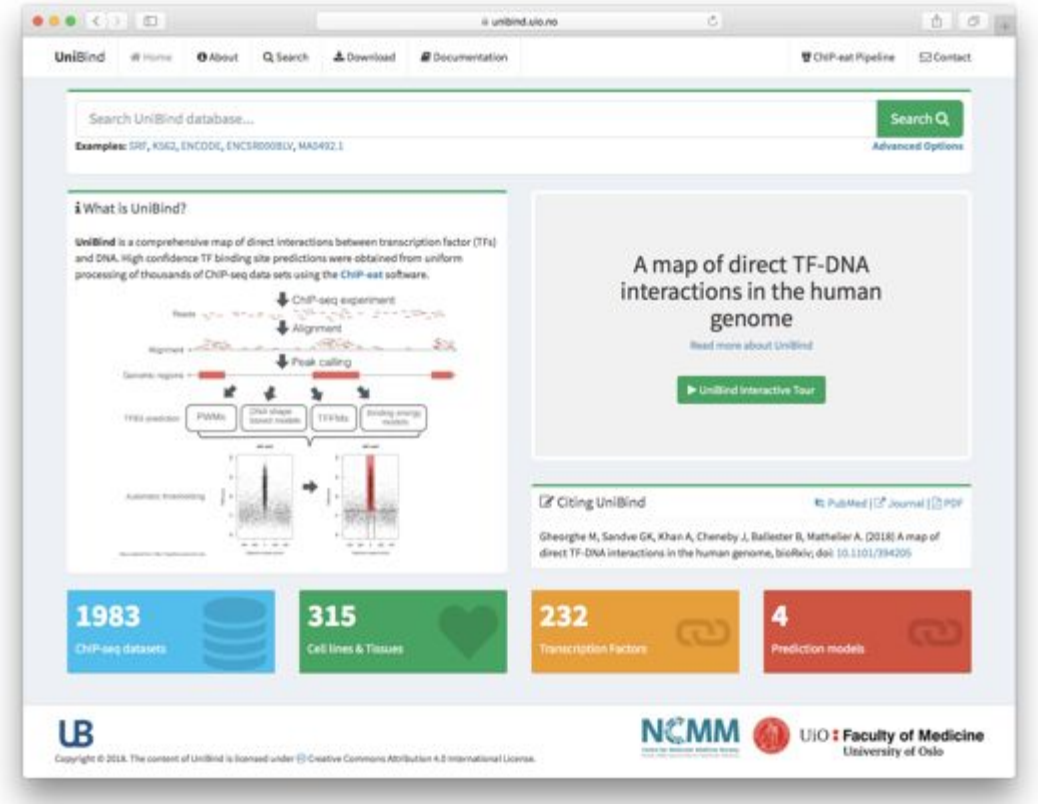
IGV data server

What do we want? REST Accessibility !!!

# Adding the RESTful interface to UniBind

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

- A comprehensive map of direct interactions between transcription factor (TFs) and DNA.
- **1983** ChIP-seq datasets
- **231** unique TFs
- Covering >4% of the human genome




**UniBind**

**unibind.uio.no**

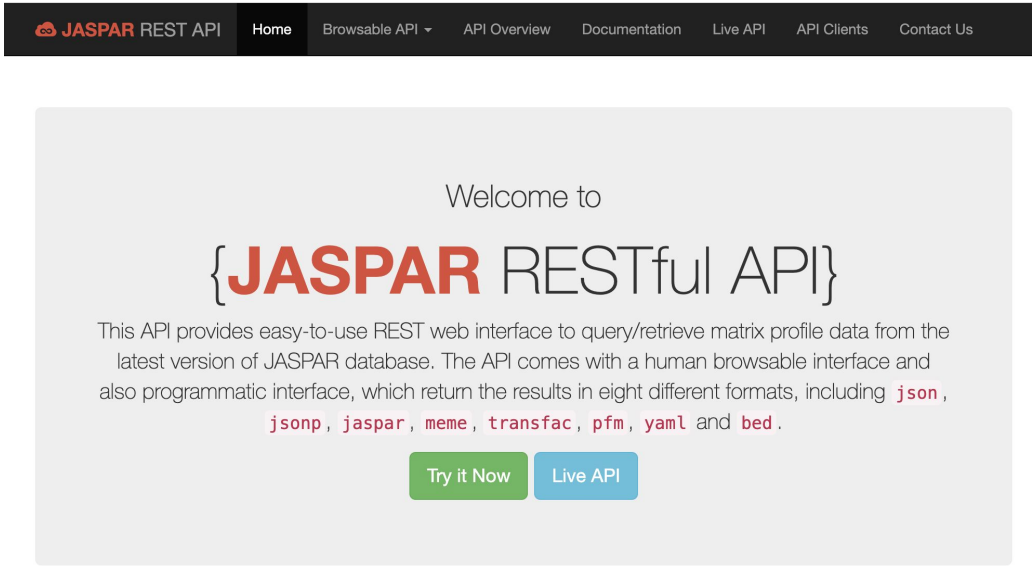
github.com/asntech/UniBind

# Introducing RESTful interface to UniBind

- A fully functional REST API
  - Django REST Framework - django-rest-framework.org
- Browsable interface
  - Django REST Framework
- Basic documentation of the API
  - Core API - coreapi.org
- Live API
  - Django REST Swagger - django-rest-swagger.readthedocs.io

# Deliverable

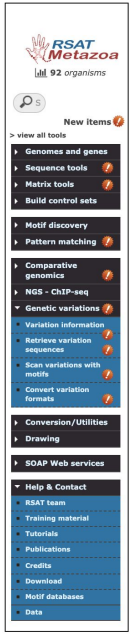
- A fully functional, documented REST API with browsable interface
- github.com/asntech/unibind
- unibind.uio.no/api

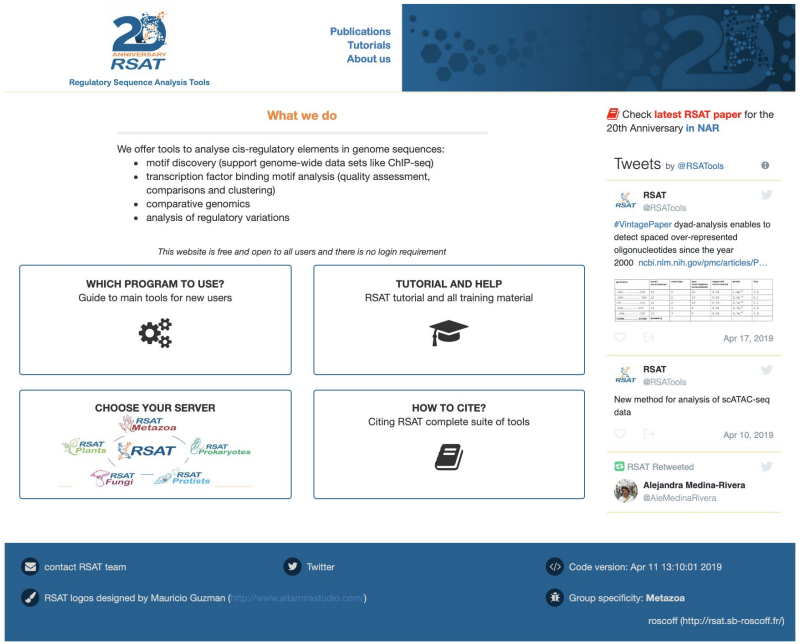


# RESTful interface for RSAT

# RSAT – Regulatory Sequence Analysis Tools

- Project proponent
  - Jacques van Helden
- URL: [http://rsat.eu/](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)*



# Motivation

- A 20-years package with some weight of its history
  - Several languages: Perl, python, R, C, C++
  - Multiple dependencies: Perl, python, C + system libraries
  - Installation procedure semi-automated for Ubuntu but not easy to transpose between systems
- Current project to facilitate local installation: porting RSAT to **conda**
  - Partly implemented during ELIXIR BioHackathon (Oct 2018)
  - To be finalized in May 2019
- Alternative (and complementary): remote access via Web services
  - No need to install RSAT locally
  - Current interface in **SOAP/WSDL**
    - not supported anymore in R
    - difficult to maintain
  - **RESTful** interfaces are adopted by more and more bioinformatics resources
    - Thi Thuy Nga NGUYEN implemented a first version of REST server
    - Currently restricted to a handful of toolss

# Goals

- Add entry points required for the other hacking projects of this hackathon (in particular the project “Workflow for the analysis of disease-associated regulatory variants”).
- Add entry points for some recently developed RSAT tools (e.g. *matrix-clustering*)
- Develop a detailed documentation for the existing and new entry points.
- Implement small test and demo scripts invoking these web services with various languages.

# Technical aspects

- Languages, libraries and tools
  - The REST server is developed in Perl with the flask library
  - The clients can be developed in any language
  - Requested skills for the hacking
- REST API development
  - Code documentation
  - Scripting of REST clients in any language (Perl, python, java, R)

| Resource name | Data types | URL |
|---|---|---|
| RSAT | Regulatory Sequence Analysis Tools | http://rsat.eu |
| RSAT REST | Prototype of the RSAT REST Web services | http://rsat-tagc.univ-mrs.fr/rest/ |

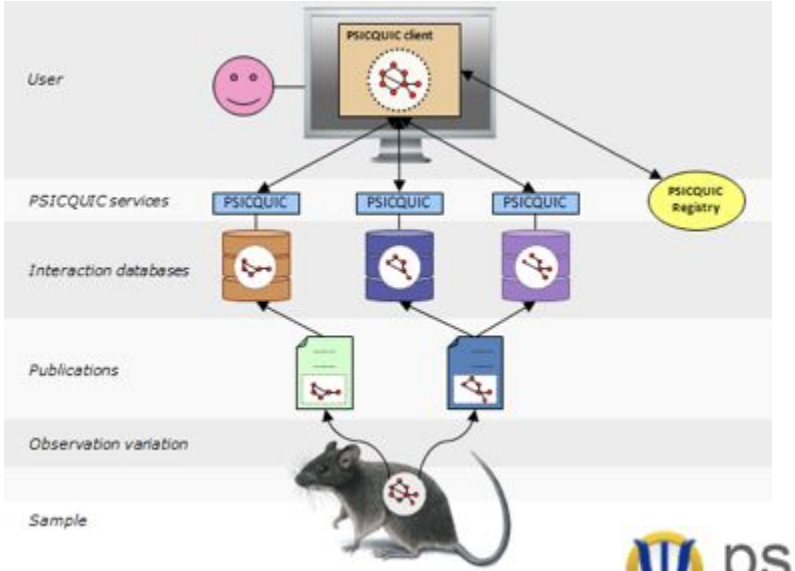
# Annotation of GREEKC-related tools and data types in bio.tools and EDAM ontology

# PSICQUIC for causal interactions and network modelling

*Noemi del Toro*

# What is PSICQUIC?

PSICQUIC is an effort from the HUPO Proteomics Standards Initiative (HUPO-PSI) to standardise programmatic access to molecular interaction databases.




psicquic

Protemics Standard Initiative Common QUery InterfaCe

# What is PSICQUIC?

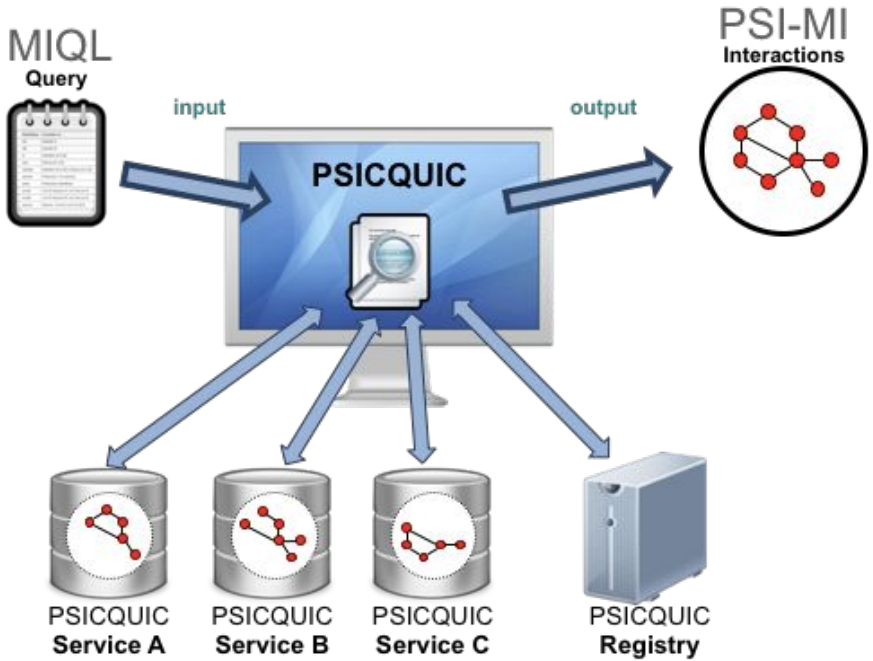
- PSICQUIC is established as a tool for the molecular interaction community. Key tools and databases such as Cytoscape, mentha, HiPPIE or GeneMania make use of it for data retrieval.
- External groups have developed R and Python packages to facilitate its use for specific communities which highlights its popularity.





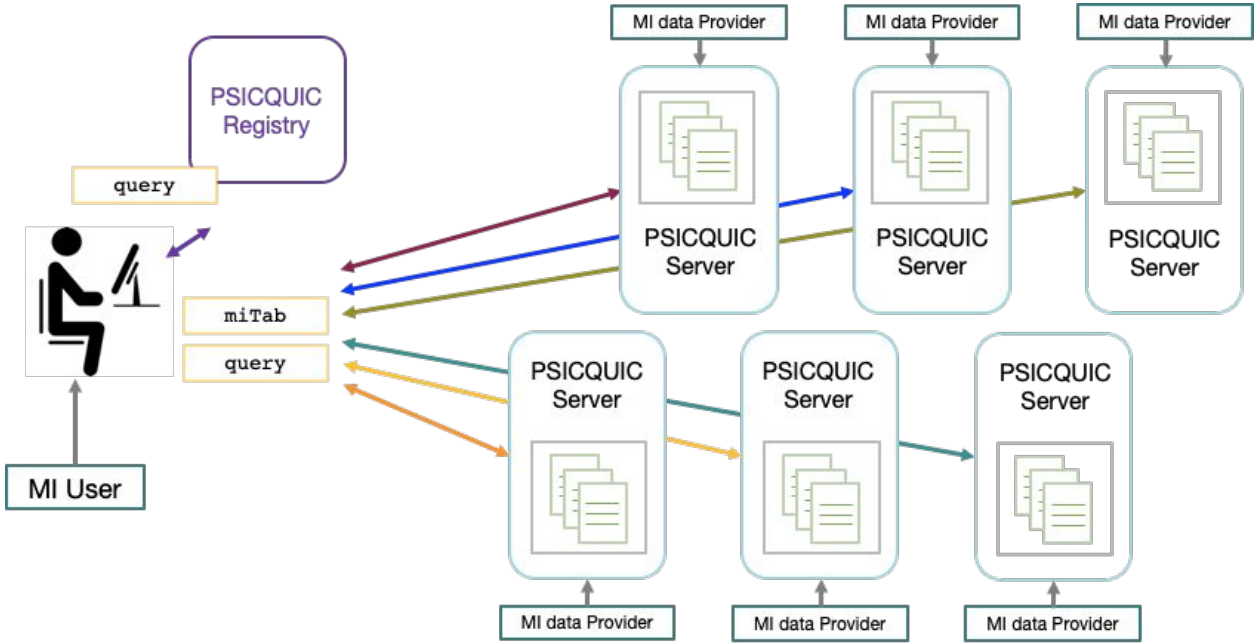

# PSICQUIC Services

- PSICQUIC defines a minimum set of standard SOAP and REST methods
- Independent data providers
  - Adopt the same specification
  - Implement the same interface
  - Same language MIQL to query services
  - Same format to provide results

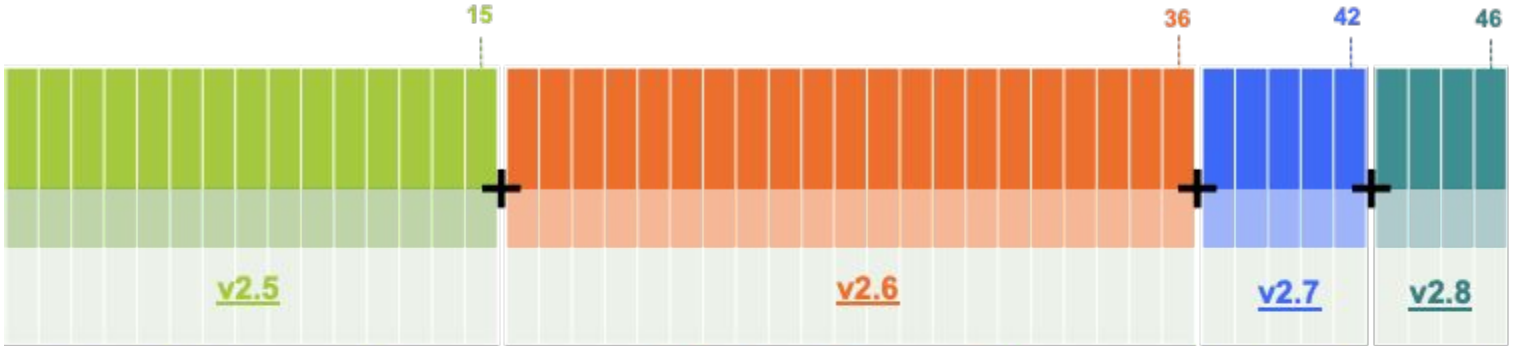


# Motivation

PSICQUIC can be reused inside of the GREEKC consortium to provide a way of distribute experimental and inferred knowledge about interactions involved in gene regulation. This knowledge could fulfill several use cases or workflows, for example, in network modelling.



# PSI-MITAB 2.8/CausalTab extension




**Standard columns (15):**

- ID(s) interactor A & B
- Alt. ID(s) interactor A & B
- Alias(es) interactor A & B
- Interaction detection method(s)
- Publication 1st author(s)
- Publication Identifier(s)
- Taxid interactor A & B
- Interaction type(s)
- Source database(s)
- Interaction identifier(s)
- Confidence value(s)

**Standard columns (21):**

- Complex expansion
- Biological role A & B
- Experimental role A & B
- Interactor type A & B
- Xrefs A, B & Int.
- Annotations A, B & Int.
- Host organism
- Parameters Int.
- Created
- Updated
- CheckSum A, B & Int.
- Negative

**Standard columns (4):**

- Binding feature A & B
- Stoichiometry A & B
- Participant identif. method A & B

**Standard columns (4):**

- Biological effect of interactor A & B
- Causal regulatory mechanism
- Causal statement

# Organisation of the hacking project

Project representation at the BioHackathon

- Noemi del Toro

Participants recruited (so far :D)

- John Zobolas
- Vasundra Touré

Expertise from Hackathon attendees

- Previous knowledge in Java, Spring and Solr is recommended but not mandatory.

# Steps and tasks

*At the end of the hackathon, we aim at providing the **first causal PSICQUIC service** based on causalTab data (e.g. SIGNOR miTab 2.8).*

**Organization**

- Day 1: extend psi-jami library to incorporate the new miTab 2.8 standard format (causalTAB) (work currently in progress)
- Day 2: have a working miTab 2.8 dataset solr-indexed
- Day 3: create a new PSICQUIC service based in miTab 2.8 in a public server and add it to the registry

**Project description**

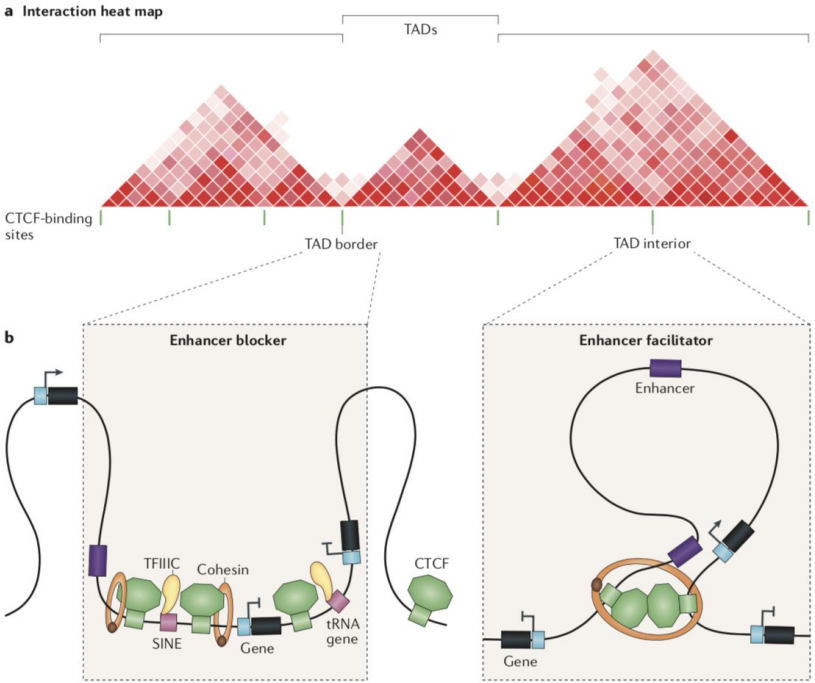
https://greekc.github.io/hackathon-marseille/project_descriptions/causal_psicquic/

# TF-TG inference from omics data integration

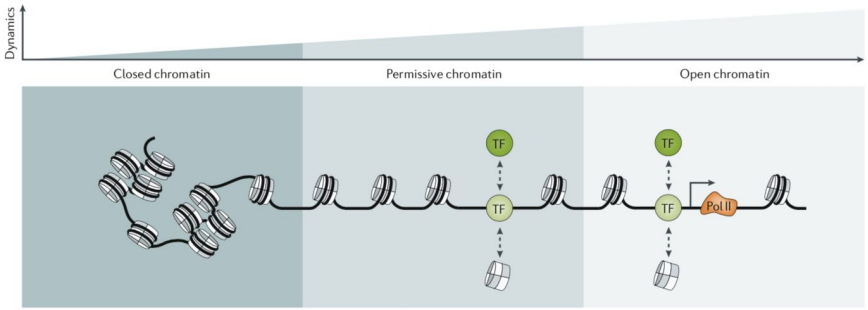
*Rafel Riudavets Puig*

# TF-TG inference from omics data integration

**Motivation:** gene transcription regulation is a complex system involving many control layers interacting between them.




*Chin-Tong Ong and Victor G. Corces, Nature reviews 2014*




*Sandy L. Klemm et.al, Nature Reviews 2019*

# TF-TG inference from omics data integration

**Objective:** to integrate information regarding the different known layers involved in the regulation of gene expression in order to build a TF-TG matrix. Validate the procedure by contrasting TF-TG matrix against transcription dynamics data.

## Data:

A549 cells treated with 100 nM dexamethasone (GR agonist) for different times. Collected data:

- Hi-C (3D architecture)
- ChIP-seq (TFs)
- ChIP-seq (histones)
- ATAC-seq
- DNAse-seq
- RNA-seq

Raw data has been checked for QC, aligned and preprocessed by ENCODE.

## Skills needed:

- R base and bioconductor packages:
  - DESeq2
  - GenomicRanges
  - rtracklayer
  - Data integration
- Experience in NGS data analysis of:
  - Hi-C
  - ChIP-seq
  - ATAC-seq/DNAse-seq
  - RNA-seq
- Molecular biology of transcription regulation.

## Expected outcome:

A workflow going from wet lab to dry lab to infer TFTG interactions.

# TF-TG inference from omics data integration

***Day 1:***

- *Create teams for the analysis of the individual data types.*
- *Distribute the data among the teams and getting familiarized with it.*
- *Individual analyses and creating consensus tracks.*
- **Discussion of the way to go for the integration on day 2.*

***Day 2:***

- *Separation in teams for the TAD vs no TAD hypotheses.*
- *Create preliminary TFTG list from TF ChIP-seq peaks (+ TAD information).*
- *Refine preliminary TFTG list with other types of data.*
- *Validation with RNA-seq data.*

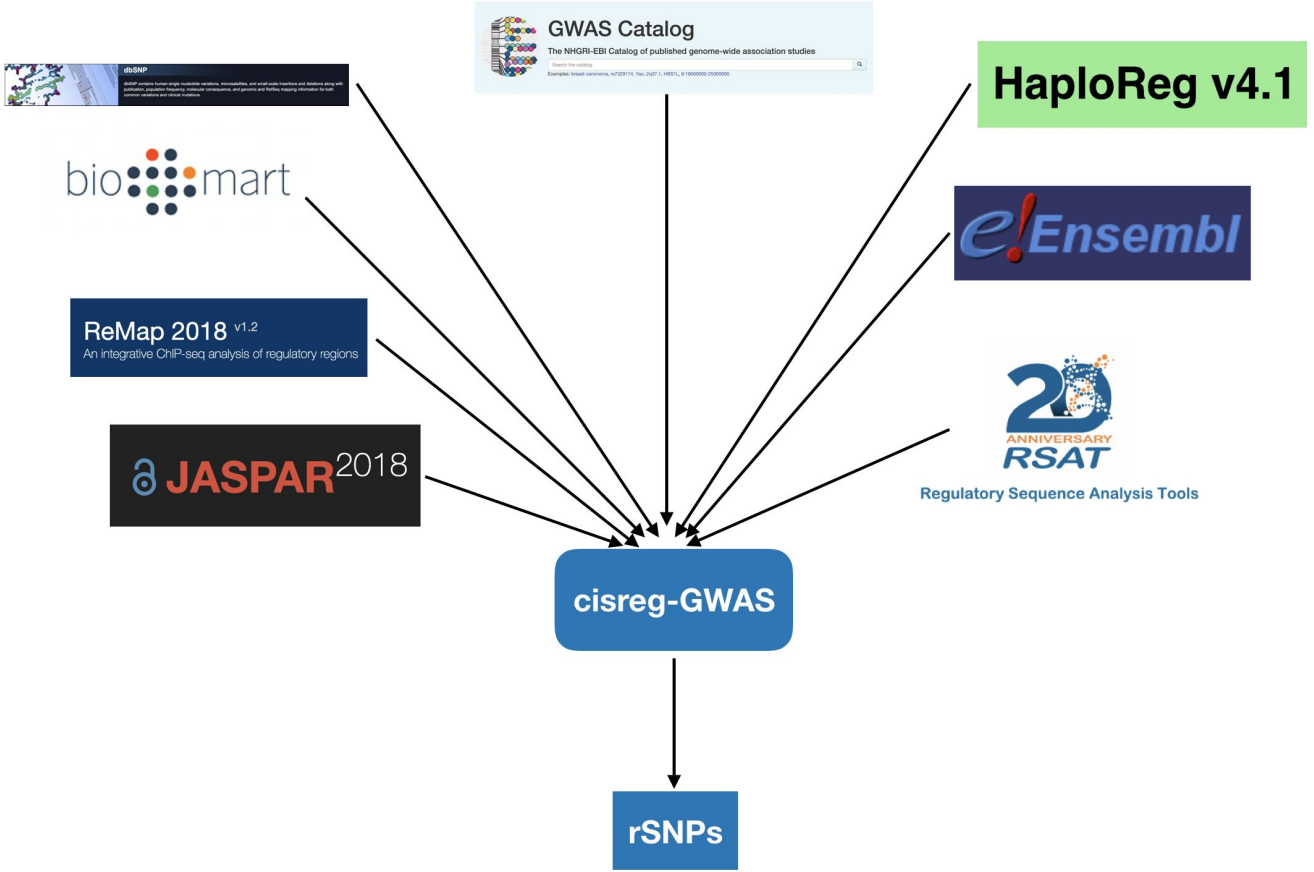
# Workflow for the analysis of disease-associated regulatory variants

*Yvon Mbouamboua & Jacques van Helden*

# Motivation

- **Project proponents**
  - Yvon Mbouamboua & Jacques van Helden
- **Motivation**
  - Develop a workflow to detect non-coding disease-associated variant that may affect transcriptional regulation by modifying transcription factor binding sites.
- **Approach**
  - Integration of information elements collected automatically from
    - genomic databases:
      - BioMart
      - dbSNP
      - Ensembl
      - HaploReg
    - bioinformatic tools:
      - Regulatory Sequence Analysis Tools (RSAT) for motif analysis
      - ReMap (ChIP-seq peaks)

# Resources


dbSNP
GWAS Catalog
The NHGRI-EBI Catalog of published genome-wide association studies
HaploReg v4.1
biomart
e!Ensembl
ReMap 2018 v1.2
An integrative ChIP-seq analysis of regulatory regions
20 ANNIVERSARY RSAT
Regulatory Sequence Analysis Tools
JASPAR 2018
cisreg-GWAS
rSNPs

# Mobilised resources

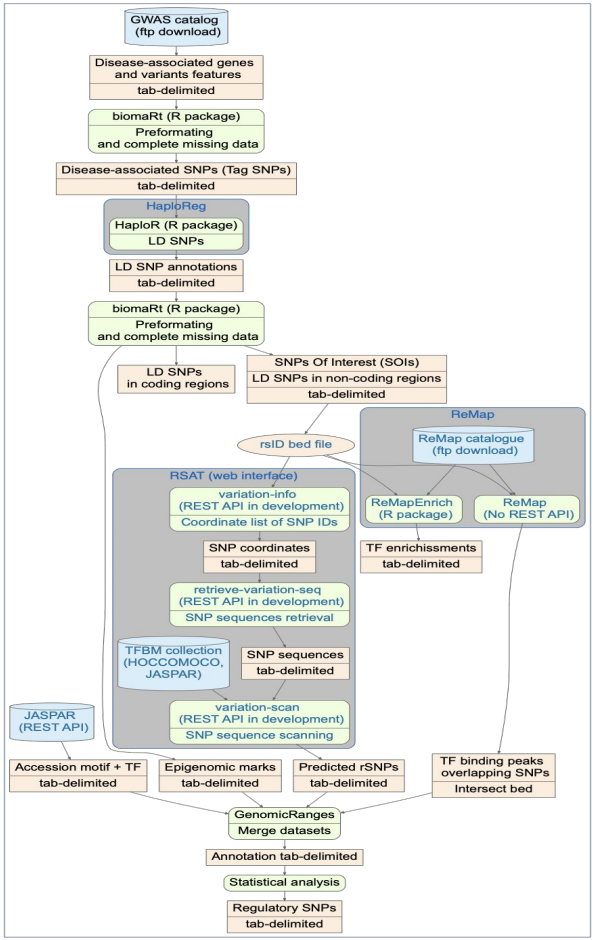
| Resource name | Data types | URL | Access mode in the workflow |
|---|---|---|---|
| GWAS catalog | SNPs associated to a query disease | https://www.ebi.ac.uk/gwas/ | ftp download |
| HaploReg | Collect the SNPs in linkage disequilibrium (LD) | https://pubs.broadinstitute.org/mammals/haploreg/ | R package |
| BioMart | Collect SNP missing data | http://www.biomart.org | R package |
| ReMap | Collect transcriptional regulators ChIP-seq experiments | http://remap.cisreg.eu/ | Web interface, to be converted to REST |
| Jaspar | Collect all matrices corresponding to transcription factor names | http://jaspar2018.genereg.net | ftp download, to be converted to REST |
| RSAT | Prediction of polymorphic variations affecting transcription factor binding | http://rsat.sb-roscoff.fr/ | Web interface, to be converted to REST |

# Interoperability issues

- **Missing interfaces**
  - Currently there is no Web services for Remap
  - RSAT Web services were originally based on SOAP/WSDL, which is not supported anymore by R
  - We are currently developing a REST interface (and the efforts are put on the tools that will be used for this hackathon)
- **Inconsistencies between IDs**

  How to establish cross-links between:
  - factor names in ReMap
  - matrix names from RSAT
  - matrices from Jaspar (names + IDs)
  - proteins in Uniprot
  - genes in Ensembl

# Flowchart


GWAS catalog
(ftp download)
Disease-associated genes
and variants features
tab-delimited
biomaRt (R package)
Preformating
and complete missing data
Disease-associated SNPs (Tag SNPs)
tab-delimited
HaploReg
HaploR (R package)
LD SNPs
LD SNP annotations
tab-delimited
biomaRt (R package)
Preformating
and complete missing data
LD SNPs
in coding regions
SNPs Of Interest (SOIs)
LD SNPs in non-coding regions
tab-delimited
ReMap
ReMap catalogue
(ftp download)
rsID bed file
RSAT (web interface)
variation-info
(REST API in development)
Coordinate list of SNP IDs
ReMapEnrich
(R package)
ReMap
(No REST API)
SNP coordinates
tab-delimited
TF enrichissments
tab-delimited
retrieve-variation-seq
(REST API in development)
SNP sequences retrieval
TFBM collection
(HOCCOMOCO,
JASPAR)
SNP sequences
tab-delimited
variation-scan
(REST API in development)
SNP sequence scanning
JASPAR
(REST API)
Accession motif + TF
tab-delimited
Epigenomic marks
tab-delimited
Predicted rSNPs
tab-delimited
TF binding peaks
overlapping SNPs
Intersect bed
GenomicRanges
Merge datasets
Annotation tab-delimited
Statistical analysis
Regulatory SNPs
tab-delimited

# Requirements for the hacking project

## Languages, libraries and tools used in the workflow

The workflow is written in R code embedded in a R markdown document, which automatically generates a report in HTML , pdf or Word .docs format.

## Needs

- Replace the downloads and manual analyses by programmatic accesses
    - Use R JASPAR package or RESTful API to download all matrices
- REST interface for RSAT
- REST interface for ReMap
- Cross-references between RSAT and Jaspar matrices
- Cross-references between ReMap factors and Jaspar

## Requested skills for the hacking

- REST API development
- Shiny interface
- Occasional help of the developers of the mobilized resources

# Expected deliveries

**Final goal**

- A fully automated workflow relying on APIs without having to download the full datasets and parse them locally.

**Deliverables**

- A workflow in R markdown document
- Yaml-based specification of the parameters
- Examples of utilisation with selected study cases
- A Shiny-based Web interface to the workflow
- Full code of the workflow available in github
- A user documentation enabling biologists to run the analyses on their own computer