Évaluation de méthodes de prédictions de relations *enhancer-gène* & application à l'étude des mécanismes de régulation de l'hémochromatose

Tristan Hoellinger

IRSD, Inserm

14 janvier 2022

Contents

Table of Contents

Table of Contents

Current acceptation

Biological definition of enhancers

In our view, an actual enhancer is: a DNA sequence that serves to enhance the transcription of at least one distal gene, *in vivo* and its native genomic context [Gasperini, Tome, and Shendure 2020].

Introduction

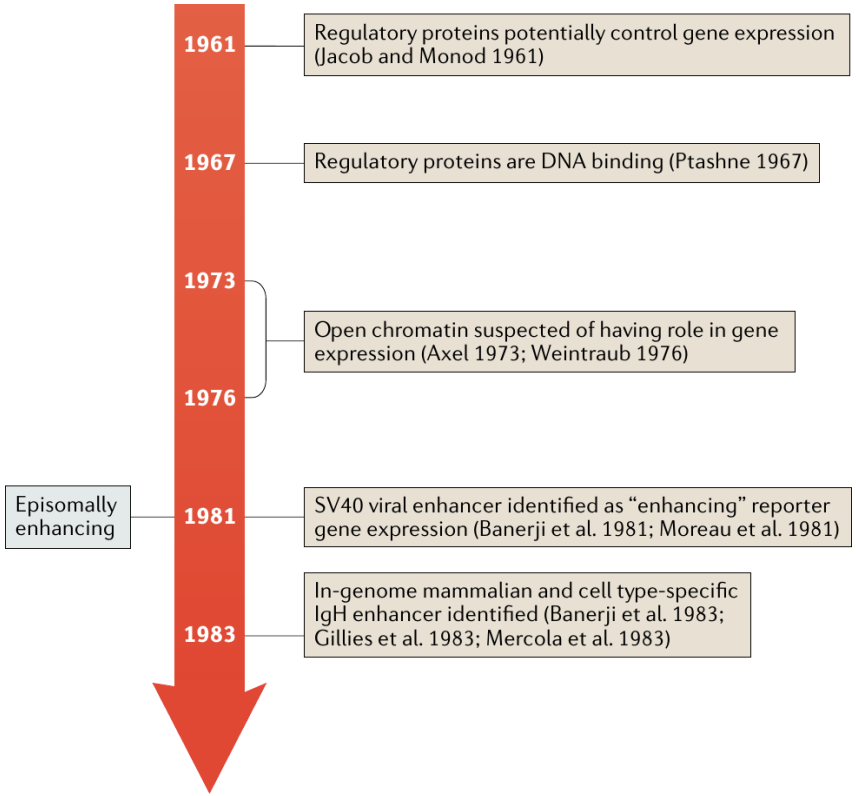

Figure 1: A history of operational definitions of enhancers. Figure adapted from Gasperini, Tome, and Shendure 2020.

- 1899, 1900, 1901: Duclaux, Dienert and Wert found that certain enzymes of micro-organisms are formed only in the presence of a specific substrate
- [Jacob and Monod 1961] discovered the first example of prokaryotic gene regulation (regulation of the lactose operon)
- The term ***enhancer*** first appeared in 1981: [Moreau et al. 1981] identified a non-coding sequence *enhancing* a *cis*-encoded reporter gene expression in SV40
- [Mercola et al. 1983] **first discovery of a eukaryotic enhancer**

Introduction

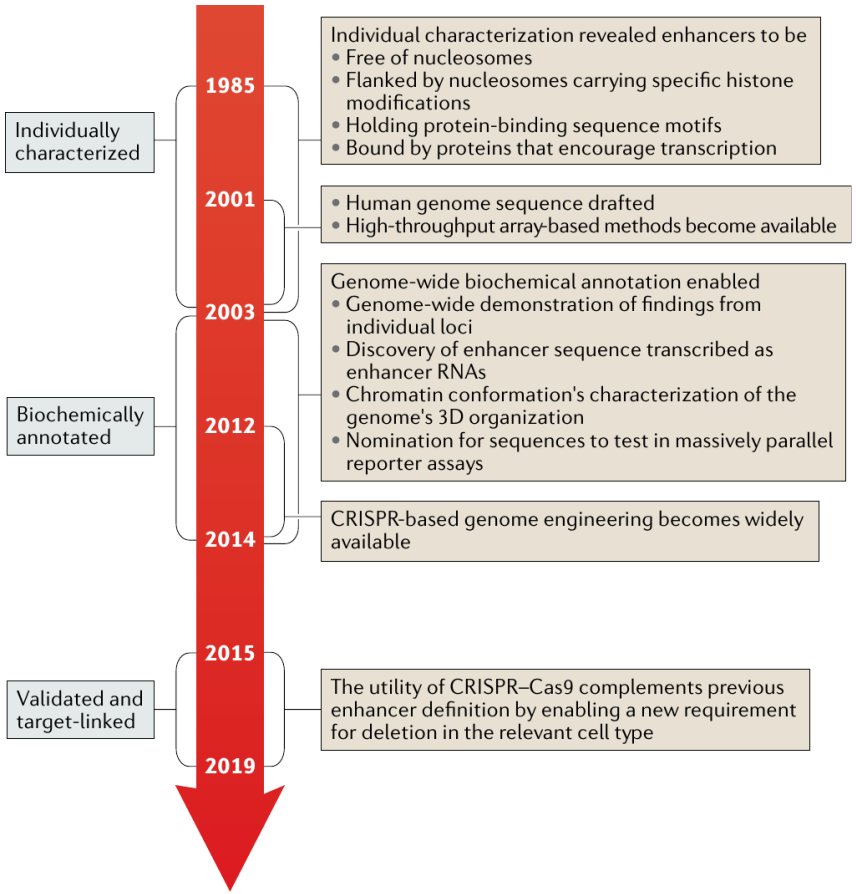

Figure 2: A history of operational definitions of enhancers. Figure adapted from Gasperini, Tome, and Shendure 2020.

- 1985-2003: New enhancers discovered and experimentally characterized on a one-by-one basis. Shared features: **free of nucleosomes**, flanked by nucleosomes with transcriptional-activity -associated **histone modifications**, **sequence motifs for transcription factors (TFs)**, bound by these TFs
- 2003-2014: Biochemical features associated with enhancers were measured genome-wide in selected cell types and tissues => annotation of cell type-specific candidate enhancers on a genome-wide basis, generally without demonstration of enhancing activity
- 2015-today: the emergence of CRISPR–Cas9 genome engineering provide a way to assert, for a candidate enhancer, whether the deletion from its native genomic context results in altered expression of a potential target gene

2020: The ENCODE Registry of candidate cis-Regulatory Elements (ccRE)

[The ENCODE Project Consortium 2020] published a registry (V2) of 926,535 human and 339,815 mouse candidate cis-regulatory elements, covering 7.9 and 3.4% of their respective genomes, by integrating selected datatypes associated with gene regulation. Today, the Registry (V3) contains **1,063,878 ccRE in 1,518 cell types and tissues.**

ccRE are a subset of representative DHS (rDHS, see below) supported by high DNase and high H3K4me3 (promoters signature) and/or high H3K27ac (enhancers signature), and/or high CTCF signal in at least one biosample.

Namely, they started with 93 million individual DHS computed across 706 DNase-seq profiles (475 distinct biosamples) in human. Then, iteratively, they clustered the DHS accross all profiles, kept the DHS with the highest signal for each cluster as rDHS, removed all the overlapping DNase peaks for subsequent iterations, until it finally results in a list of non-overlapping rDHS representing all DNase peaks in all samples, totalling 2,3 million human rDHS. At the end of the day, a rDHS is a region which has a DNase peak in at least one of the 706 DNase-seq profiles.

2020: The ENCODE Registry of candidate cis-Regulatory Elements (ccRE)

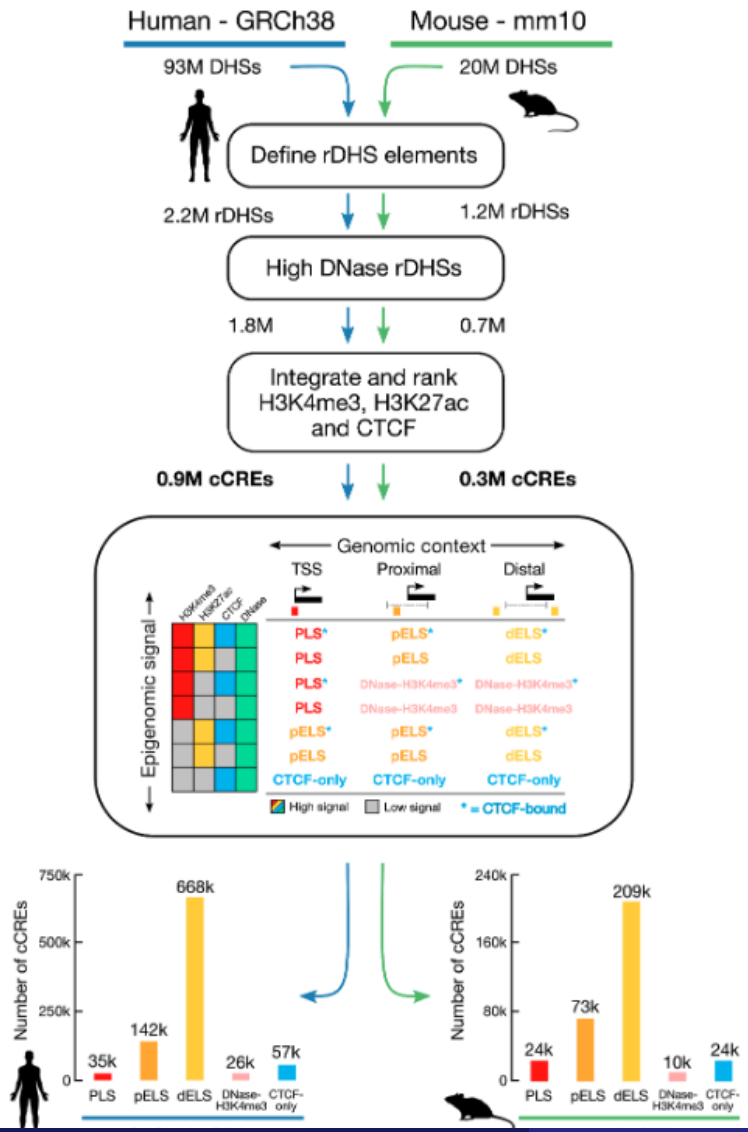

Figure 3: Selection and classification of ccRE to build the Registry of candidate cis-regulatory elements. Figure taken from screen.encodeproject.org.

ccRE are a subset of rDHS supported by high DNase and high H3K4me3 (promoters signature) and/or high H3K27ac (enhancers signature), and/or high CTCF signal in at least one biosample.

When some data are missing, ENCODE3 defined an ordered list of most-probable ccRE type.

2020: The ENCODE Registry of candidate cis-Regulatory Elements (ccRE)

- The ~ 1 million ccRE have a good resolution: $150 <$ length < 350 bp.

- **ccRE-ELS (candidate enhancers) are ccRE that, in any of the biosamples, have high DNase and high H3K27ac signal**, *and must additionnaly have a low H3K4me3 signal* ***if*** *they are within 200 bp of any annotated TSS. The same definition holds for cell-type specific classification of ccRE.*

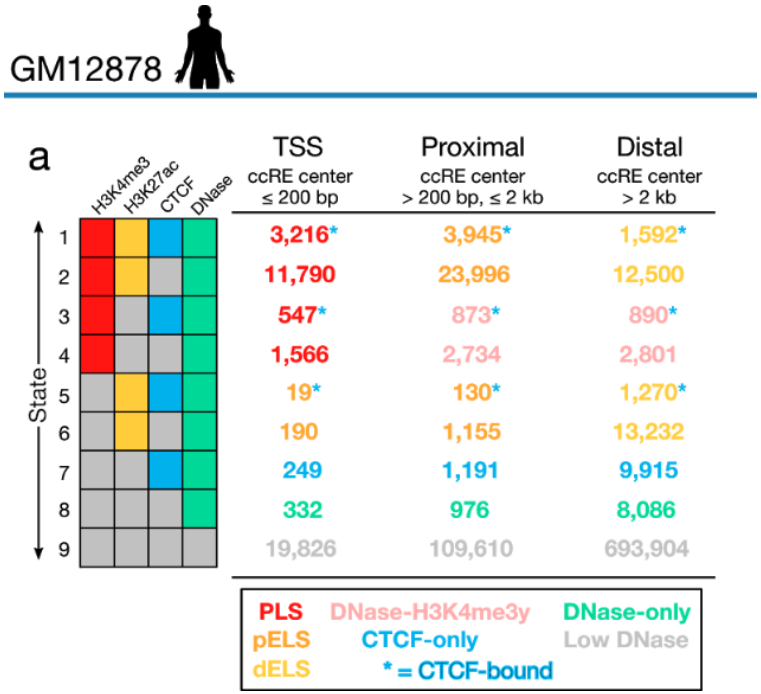

Figure 4: Classification of ccRE in GM12878. Figure taken from screen.encodeproject.org.

Table of Contents

Definition of enhancers

Biological definition of enhancers

In our view, an actual enhancer is: a DNA sequence that serves to enhance the transcription of at least one distal cis-located gene, *in vivo* and its native genomic context [Gasperini, Tome, and Shendure 2020].

We lack ideal operational definitions of enhancers, though we have solid reasons to think that almost all enhancers are associated with regions of open chromatin (otherwise transcription factors cannot cling to it) and are flanked by histones carrying H3K27ac and/or H3K4me1 histone marks. They interact with the cognate promoters in 3D space and can be active or not. Hence, as DHS - regions sensitive to cleavage by DNase I - are believed to charaterize open chromatin [Boyle et al. 2008], it is reasonable to think that most enhancers are to be found among ENCODE ccRE.

Regulation mechanisms

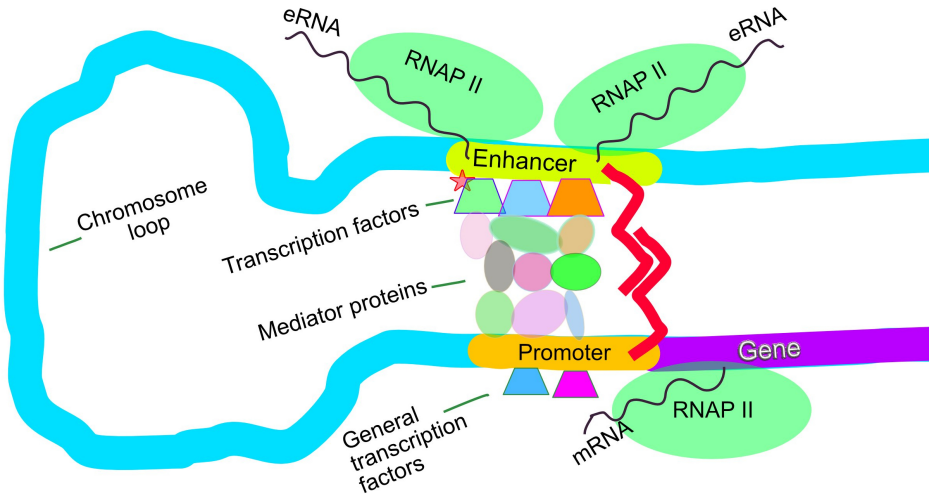

Figure 5: Enhancers in mammals, en.wikipedia.org/wiki/Enhancer_(genetics)

Although the exact mechanisms by which enhancers operate may be numerous and are still poorly known, one is believed to be through 3D contact [Schoenfelder and Fraser 2019]. Specific regulatory transcription factors bind to DNA sequence motifs in the enhancer, and the mediator proteins communicate regulatory signals from the enhancer to the promoter. The loop is stabilized by connector proteins (for instance CTCF) anchored to binding motifs in the enhancer and the promoter. Note that the mechanistic importance of such enhancer–promoter proximity is far from settled [Gasperini, Tome, and Shendure 2020].

Table of Contents

Position of the problem

We recall the definition of candidate enhancers:

Definition of candidate enhancers (ccRE-ELS)

ccRE-ELS are ccRE that, in any of the biosamples curated by ENCODE, have **high DNase and high H3K27ac signal**, *and must additionnaly have a low H3K4me3 signal* ***if*** *they are within 200 bp of any annotated TSS. The same definition holds for cell-type specific classification of ccRE.*

Although the Registry of ccRE provides a comprehensive framework for identifying enhancers, it remains a major challenge to find what genes these candidate enhancers mediate their effect through, thereby to identify E-G interactions.

Main approaches

There are four broad types of approaches in this field:

(A) cell type specific gene expression QTL analysis combined with a set of cell type specific enhancers

(B) computational prediction from several types of high-throughput functional genomic 1D data (RNA-seq, ATAC-seq, histone marks, etc)

(C) computational prediction from a single type of chromosome conformation 3D data (promoter capture HiC, Pol II ChIA-PET, etc) combined with a set of cell type specific enhancers

(D) cell type specific genetic screening data analysis (notably CRISPR -based experimental approach (see section 2))

Main approaches

Given (A), (C) and (D) are costly, (A) is not sensitive, (D) is not genome-wide, (C) is not only not precise but might also not be sensitive (see 2), and many international projects have now generated plenty of functional genomic 1D data, (B) seems more promising.

Yet, the main drawback of (B) (and (C)) is that one needs good testing datasets to evaluate the accuracy of the predictions.

Table of Contents

Importance of E-G relations in complex genetic disorders

If GWAS have identified hundreds of variants associated to a broad range of traits and diseases, little insight has been gained about the biological mechanisms underlying these associations. Part of this gap in knowledge is likely explained by the fact that the large majority of these variants do not impact protein-coding regions but are rather located in regulatory regions active in disease-related cell types, and in particular in enhancers.
Therefore it is crucial to be able to better identify E-G relationships. In particular, identifying E-G relationhsips seems crucial for a better understanding of complex genetic disorders such as Parkinson's [Farh et al. 2015], and might be of great importance towards a better understanding of the severity of haemochromatosis.

Table of Contents

Table of Contents

Introduction

We investigated the performance of two of the most recent and promising methods of the field: the heuristic Activity-by-contact (ABC) model from Fulco, Nasser, et al. [2019] and the Average-rank method from Moore et al. [2020].

We evaluated the predictions of both methods over the two most recent reference sets of the field: the CRISPRi-FlowFISH set based on approach (D) in the K562 cell line for 30 genes spanning 1.1–4.0 Mb in five genomic regions aggregated with other CRISPR-based element-gene pairs involving 29 other genes Fulco, Nasser, et al. 2019, and the Benchmark of E-G interactions (BENGI) set made of data from (A), (C) and (D) Moore et al. 2020

Table of Contents

Recall on Precision-Recall!

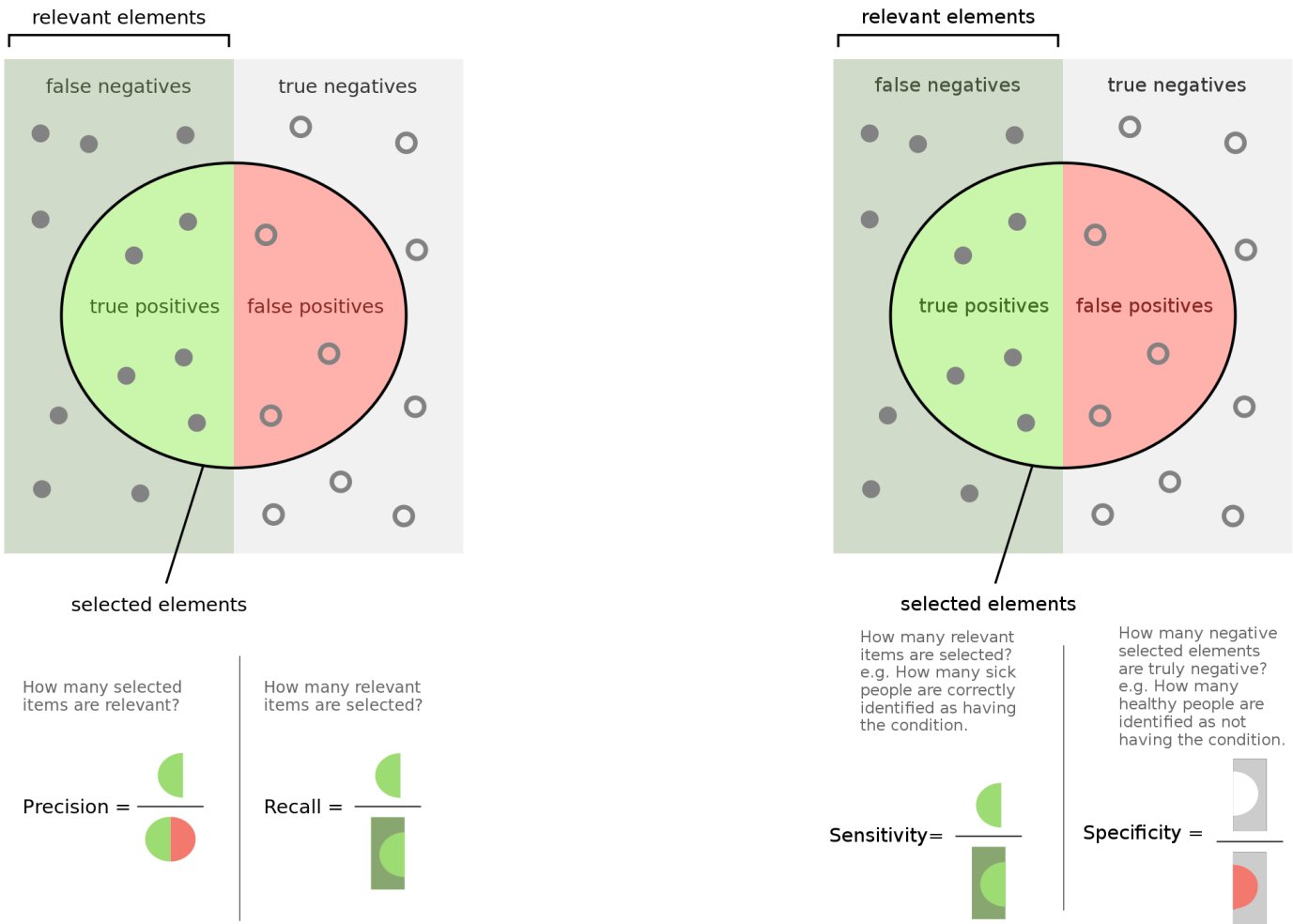

Figure 6: `en.wikipedia.org/wiki/Precision_and_recall`, `en.wikipedia.org/wiki/Sensitivity_and_specificity`

Table of Contents

BENGI in a nutshell

Essentially: according to the cell type, up to 3 types of sets. Namely, 3D-based sets, QTL-based sets, and CRISPR-based sets.

The CRISPRi-FlowFISH set in a nutshell

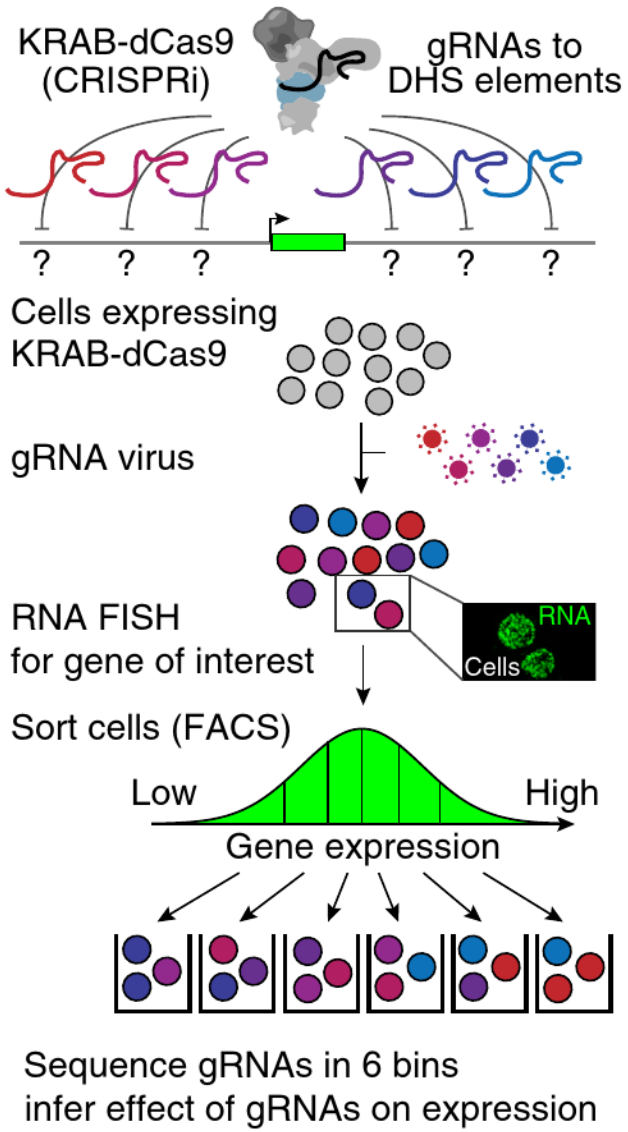

Figure 7: CRISPRi-FlowFISH method for identification of gene regulatory elements. Cells expressing KRAB-dCas9 are infected with a pool of gRNAs targeting DHS elements near a gene of interest, labeled using RNA FISH against that gene and sorted into bins of fluorescence signal by FACS. The quantitative effect of each gRNA on the expression of the gene is determined by sequencing the gRNAs within each bin. Inset: example of K562 cells labeled for RPL13A. [Fulco, Nasser, et al. 2019]

The CRISPRi-FlowFISH set in a nutshell

The CRISPRi-FlowFISH set, based on (D), seems more reliable, however it is not genome-wide. It is based on CRISPR interference with stringent criteria to ensure statistical significance. Broadly speaking, it defines positive E-G pairs as pairs for which enhancer inhibition leads to a decrease in gene expression, at an $FDR < 0.05$ (Fulco, Nasser, et al. 2019, Fulco, Munschauer, et al. 2016).

The ABC model

The Activity-By-Contact (ABC) model is a heuristical model based "on the simple biochemical notion that an element's quantitative effect on a gene should depend on its strength as an enhancer (Activity) weighted by how often it comes into 3D contact with the promoter of the gene (Contact), and that the relative contribution of an element on a gene's expression (as assayed by the proportional decrease in expression following CRISPR-inhibition) should depend on that element's effect divided by the total effect of all elements" (note that the latter assumption is a superpositinon/linearity/absence of interactions assumption, which is very questionable but make things easy).

The ABC model

Formally, the ABC score between element E and gene G is defined as follows:

$$\mathrm{ABC}_{EG} = \frac{a_E \cdot c_{EG}}{\sum_{|eG| \leq 5Mb} a_e \cdot c_{eG}}$$

where thed Activity (a) is estimated as the geometric mean of the read counts of DHS and H3K27ac chromatin immunoprecipitation sequencing (ChIP–seq) at element E, and Contact (c) as the (normalized) Hi-C contact frequency between E and the promoter of gene G at 5-kb resolution.

The ABC model

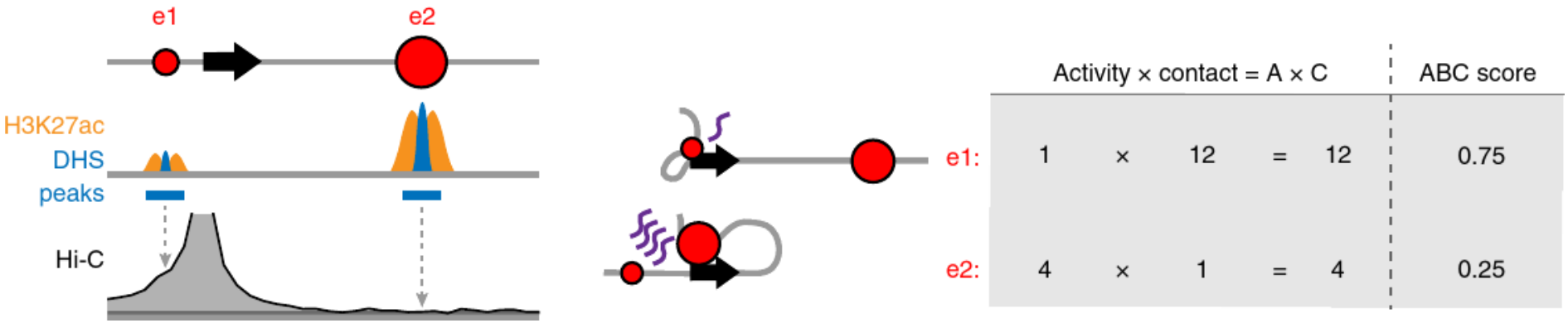

Figure 8: Calculation of the ABC score. Extracted from [Fulco, Nasser, et al. 2019]

Note that the authors have shown that cell-type specific Hi-C data is not strictly speaking necessary, as it can be well estimated with a power law with scaling exponent (somewhere around -1), for a very similar performance of the ABC model.

Table of Contents

Introduction

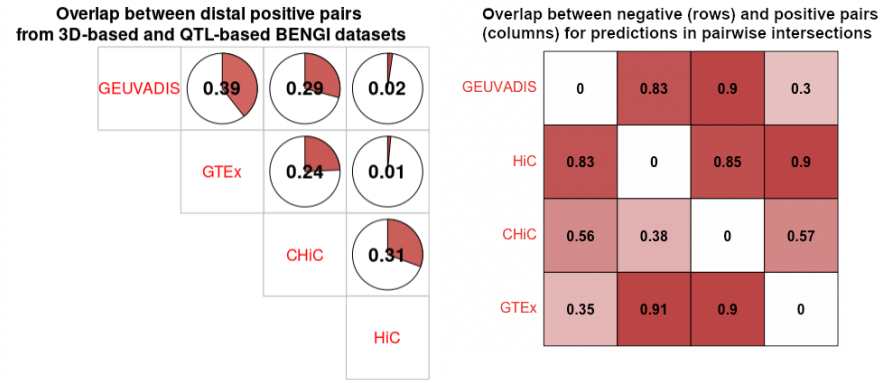

Figure 9: **(left)** Overlap between distal (distance $\geq$ 20-percentile of the distance distribution of HiC pairs) pairs from 3D-based sets (HiC, CHiC) and eQTL-based sets (GEUVADIS, GTEx). **(right)** Overlap between negative pairs (rows) and positive pairs (columns). The overlap is computed only taking into account predictions (positives + negatives) that are in the pairwise intersections.

- 1-2%, (resp. 24-29%) overlap between distal positives pairs from the HiC (resp. CHiC) set and eQTL sets ($\sim$ same with all pairs)
- $\sim$ 90% HiC negative pairs overlapping eQTL pairs, overlap a **positive** eQTL pair, supporting the suspected poor sensitivity of (C)-based methods

Table of Contents

Over the BENGI sets

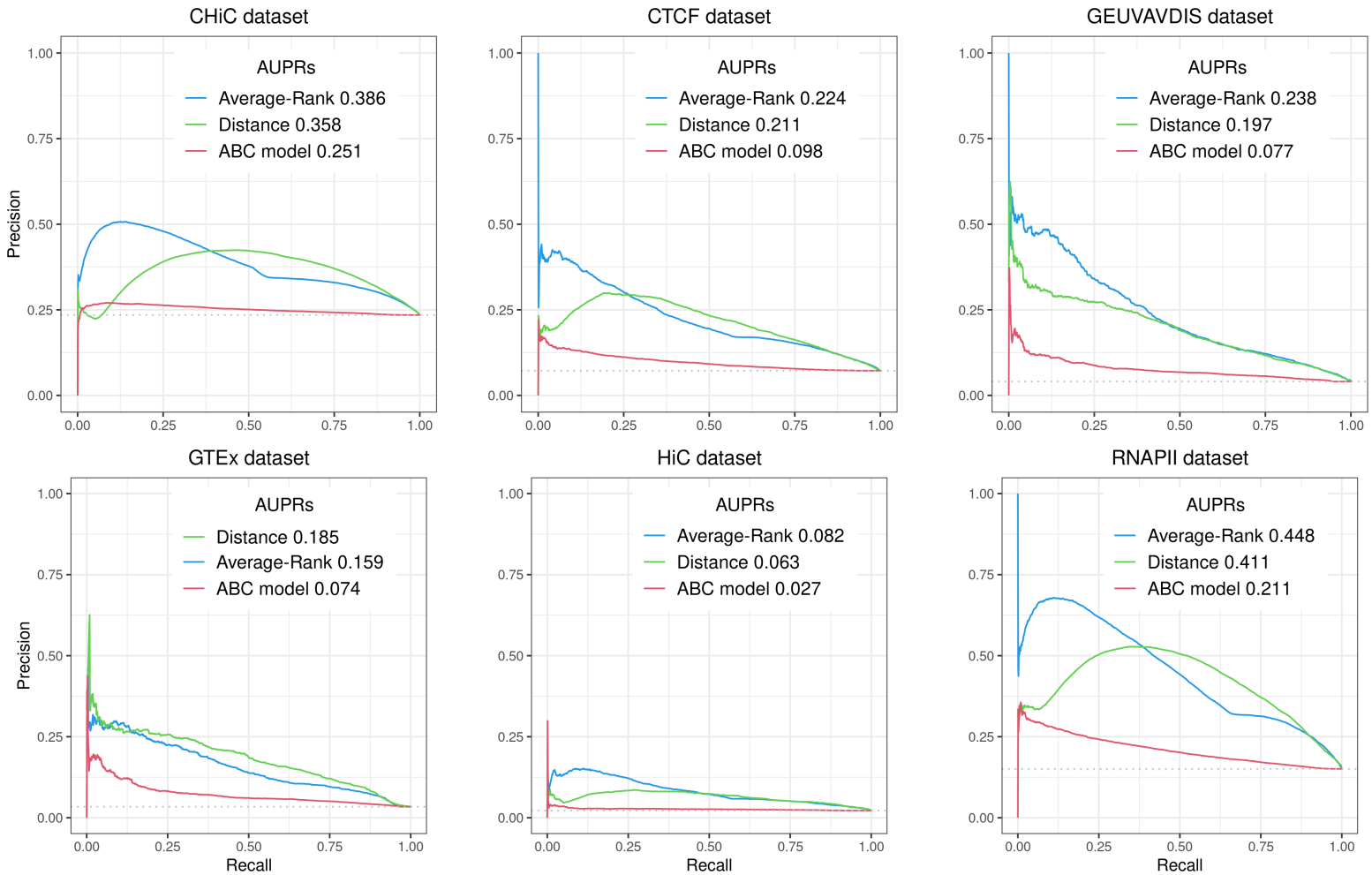

Figure 10: Precision-recall curves for baseline distance method, ABC model applied to ccRE-dELSs, and Average-rank method from Moore et al. 2020, over the 6 BENGI datasets for GM12878. ABC performs even much worst than the baseline distance method.

Over the CRiFF set

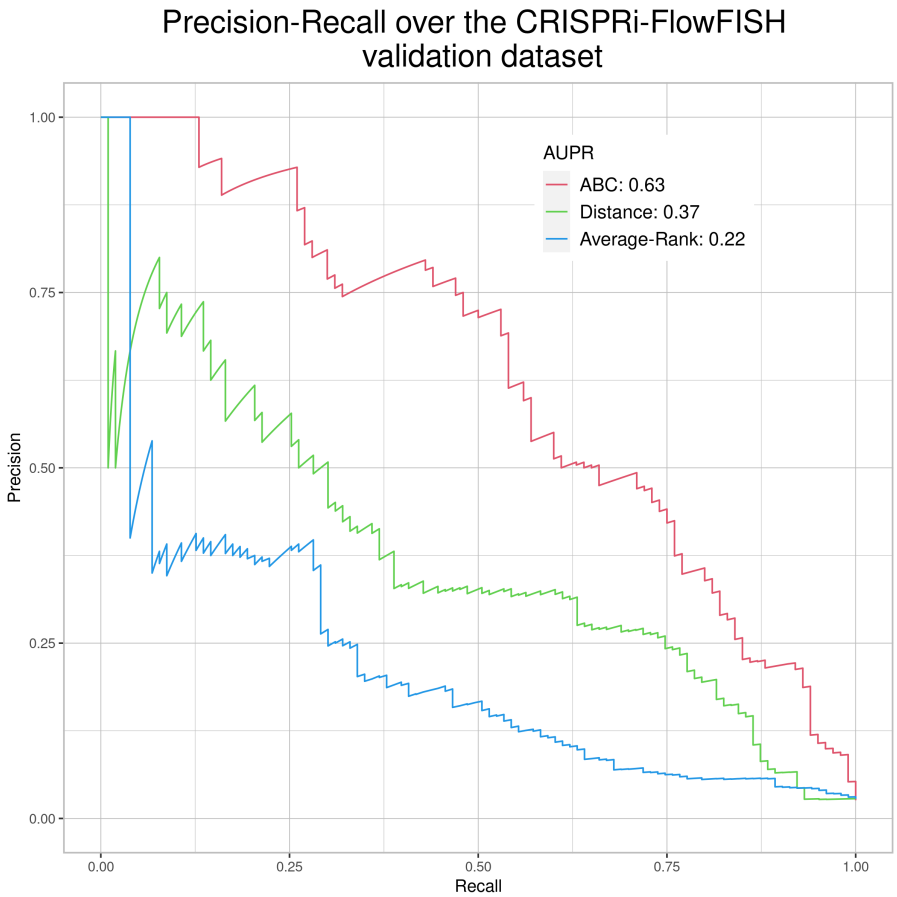

Figure 11: Precision-recall curves of distance method, Average-Rank method and ABC model over the CRISPRi-FlowFISH validation dataset.

Yet the ABC model has a good predictive power for interactions based on CRISPRi-FlowFISH.

In addition, if 3D contact was mandatory for E-G interactions, it would either imply a high overlap coefficient between distal positive E-G pairs from HiC and eQTL datasets - which is not the case as shown on Fig. 9, or that many positive eQTL pairs are in fact false positives.

Table of Contents

Conclusion

We have seen that, despite the outstanding effort of Moore et al. [2020] to produce comprehensive reference sets, our results suggest that one cannot only rely on BENGI sets to reliably evaluate the ability of a given method to find regulatory interactions of the E-G type.

Despite not being genome-wide, the CRiFF dataset from Fulco, Nasser, et al. [2019] validates the performance of the ABC model at least when it comes to predicting one precise type of enhancer - the same as that contained in the CRiFF dataset. Moreover, Nasser et al. [2021] evaluated the ABC model over more pairs obtained with the CRISPRi-FlowFISH method (~ 2 times more), which resulted in the same performance.

Hence we chose to use E-G predictions obtained with the ABC model (from [ibid.]) for our investigations on haemochromatosis.

Table of Contents

Table of Contents

Objective

We aim at finded new genes potentially related to haemochromatosis / involved in the severity of haemochromatosis, hopefully different from the ones already known to be strongly involved in iron metabolism.

The idea is to collect the list of enhancers regulating genes already known to be causally related to haemochromatosis, or strongly related with iron metabolism. Then, we can compute the list of all genes regulated by those putative enhancers.

List of initial genes

name	*Role in*	*chr*
HFE	**hemochromatosis**	6
TFR2	**hemochromatosis**	7
HFE2	**hemochromatosis**	1
HAMP	**hemochromatosis**	19
SLC40A1	**hemochromatosis**	2
BMP6	**hemochromatosis**	6
TMPRSS2	iron metabolism	21
TFR1 (= TFRC)	iron metabolism	3
DMT1	iron metabolism	12
DCYTB	iron metabolism	2
NEO1	iron metabolism	15
CIAPIN1	iron metabolism	16
ZIP14	iron metabolism	8

Table of Contents

Collecting data

We first downloaded all ABC model -predictions computed in [Nasser et al. 2021] accross 131 distinct biosamples, among which there are 74 distinct primary cell types, tissues and cell lines. They consist of the 7,717,393 E-G pairs for which $\mathrm{ABC.Score} \geq 0.015$ (empirical threshold chosen by the authors. In a previous study they use a 0.02 threshold which led to $\sim 70\%$ precision and recall on their CRiFF validation dataset).

Extracting data of interest

Note that all the genes involved in all E-G pairs are expressed in their respective biosamples. Over the 7,717,393 E-G pairs, we first filtered out the $\sim$ 1 million E-G pairs for which E is the promoter of G. The remaining pairs involve 2,463,310 distinct "enhancers" (see next slides), the majority of which are genic or intergenic, and some of which are promoters.

Of those 131 biosamples, only 5 interest us: 3 in liver (`hepatocyte-ENCODE`, `HepG2-Roadmap`, `liver-ENCODE`) and 2 in intestine (`large_intestine_fetal-Roadmap`, `small_intestine_fetal-Roadmap`).

How to define enhancers?

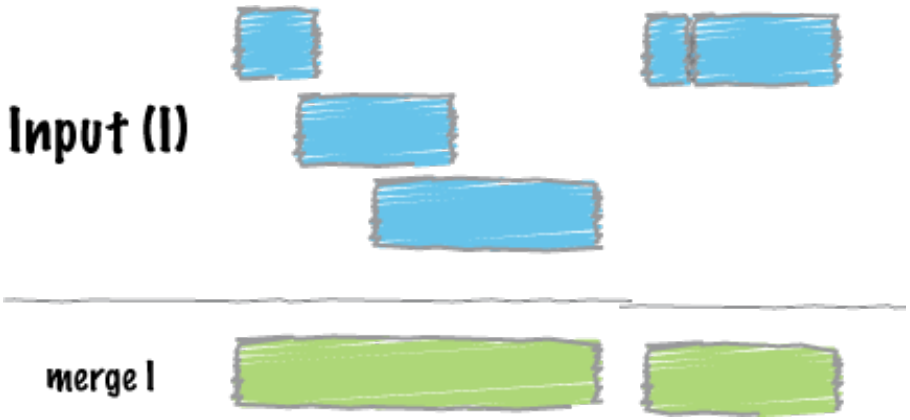

Figure 12: Adapted from `bedtools` documentation

- computes the 150,000 strongest peaks (in term of number of reads) using a tool called `bedtools`,
- resizes all of them to a fixed length (eg 500 bp) region centered on the peak summit,
- removes any regions included in a blocklist
- merge any regions overlapping

By default, the ABC model defines its own "elements" (understand: "candidate enhancers") starting from DNase peaks in DNase-seq data.Namely, it:

This is the procedure that has been used to define enhancers in the data we downloaded, except the final regions have been resized troncated by 150 bp on each side.

How to define enhancers?

There are multiple issues with using enhancers defined this way:

- Their boundaries are not invariant wrt sample, that is: two elements defined this way in 2 biosamples, can overlap w/o being the same
  - how to gather E-G pairs from multiple biosamples?
  - from a biological perspective, it makes little sense: enhancers shan't be cell-type specific, but the corresponding signals / activity can be
- Open chromatin is necessary to have an enhancer, but high H3K27ac is required too, and this is not taken into account
  - why not using cell-type specific ccRE-ELS, which are fewer and more likely to be actual enhancers / better curated?
  - what if we need to refer on ccRE or ccRE-ELS in futher studies?

How to define enhancers?

We tried several ways to address these issues.

- Merge all the 2,463,310 enhancers from the 131 biosamples together, to obtain a list of 269,254 non-overlapping enhancers
- Merge all the 102,996 enhancers among the 5 biosamples of interest, to obtain a list of 57,398 enhancers
- Merge the enhancers from the 3 liver biosamples together, and those from the 2 intestine biosamples together
- Repeat these 3 approaches + intersect with ccRE-ELS

In any case, the idea was then to replace each $E_{\text{origin}} - G$ pair with the corresponding $E_{\text{merged}} - G$ pair.

How to define enhancers?

The sizes of enhancers obtained are distributed as follows:

Enhancers	1st Qu.	Median	3rd Qu.	Max
All 131 biosamples	200	308	637	6991
All 131 biosamples (merged)	288	631	1078	11616
Liver + intestine	200	353	644	4626
Liver + intestine (merged)	200	423	756	6298
Liver + intestine (merged 131)	520	982	1668	11616
Liver	200	296	585	4626
Liver (merged)	200	348	650	4737
Liver (merged 131)	528	1027	1780	11616
Intestine	200	428	724	3679
Intestine (merged)	221	488	814	4461
Intestine (merged 131)	710	1209	1978	11073
ccRE-ELS	**247**	**352**	**519**	**16633**

How to define enhancers?

Merging enhancers accross biosamples of interest made no big difference in size distributions, but merging accross all 131 biosamples did.

Still, we chose to use enhancers merged accross the 131 biosamples, so that the list is not modified when we add or remove biosamples in further analysis.

Note that 32% of the merged putative enhancers across all 131 biosamples, do not match any ccRE-ELS, but only 9% do not match any ccRE. This suggests that a lot of those "putative enhancers" are actually other types of regulatory element as defined by ENCODE. This is to be expected: ccRE-ELS are a subset of ccRE, which are themselves a subset of rDHS, and putative enhancers defined in [Nasser et al. 2021] are based (only) on DHS.

Which gene annotation should we use?

ABC predictions from [Nasser et al. 2021] use gene symbols, which are not well suited for further analysis. We somehow had to replace them whether with `ensembl` or `entrez` id to perform GO enrichment analysis.

There are 23,220 distinct gene names in the 131 biosamples.

At first, we needed `entrez` id of genes in order to compute GO enrichment. To obtain `entrez` id, we first had to obtain `ensembl` id of the genes. In the `ensembl` annotation we use, we found an entry for 20,692 genes among the 23,220 from Nasser et al. [ibid.].

Which gene annotation should we use?

In order to investigate why there are 2,528 genes for which we found no `ensembl` id, given Nasser et al. [2021] work with `RefSeq` annotation, we downloaded a RefSeq annotation over the GRCh37 assembly, which is the same as that used by the authors, except they did not specify the precise version.

In this RefSeq annotation, we found an entry for 21,845 genes among the 23,220 from Nasser et al. [ibid.], 17,665 of which are protein-coding.

Hopefully, among the 2,528 genes for which we did not find any ensembl id, 1,197 have an entry in our RefSeq annotation. It appeared that only a small part of them (52) are protein-coding, whereas most of them are long non-conding RNA.

At the end of the day, starting from 23,220 distinct gene names in the 131 biosamples, we obtained 20,692 `ensembl` id and 19,538 `entrez` id.

Networks: implementation

We aimed not only at finding genes regulated by the same enhancers as those regulating our 13 (12 at the time we performed our experiments, we still have to add ZIP14) initial genes, but also at having a nice representation of them. To that purpose, we used the R package `visNetwork`.

We found 422 genes (including the initial 12) regulated by the enhancers regulating the 10 initial genes.

Networks: results

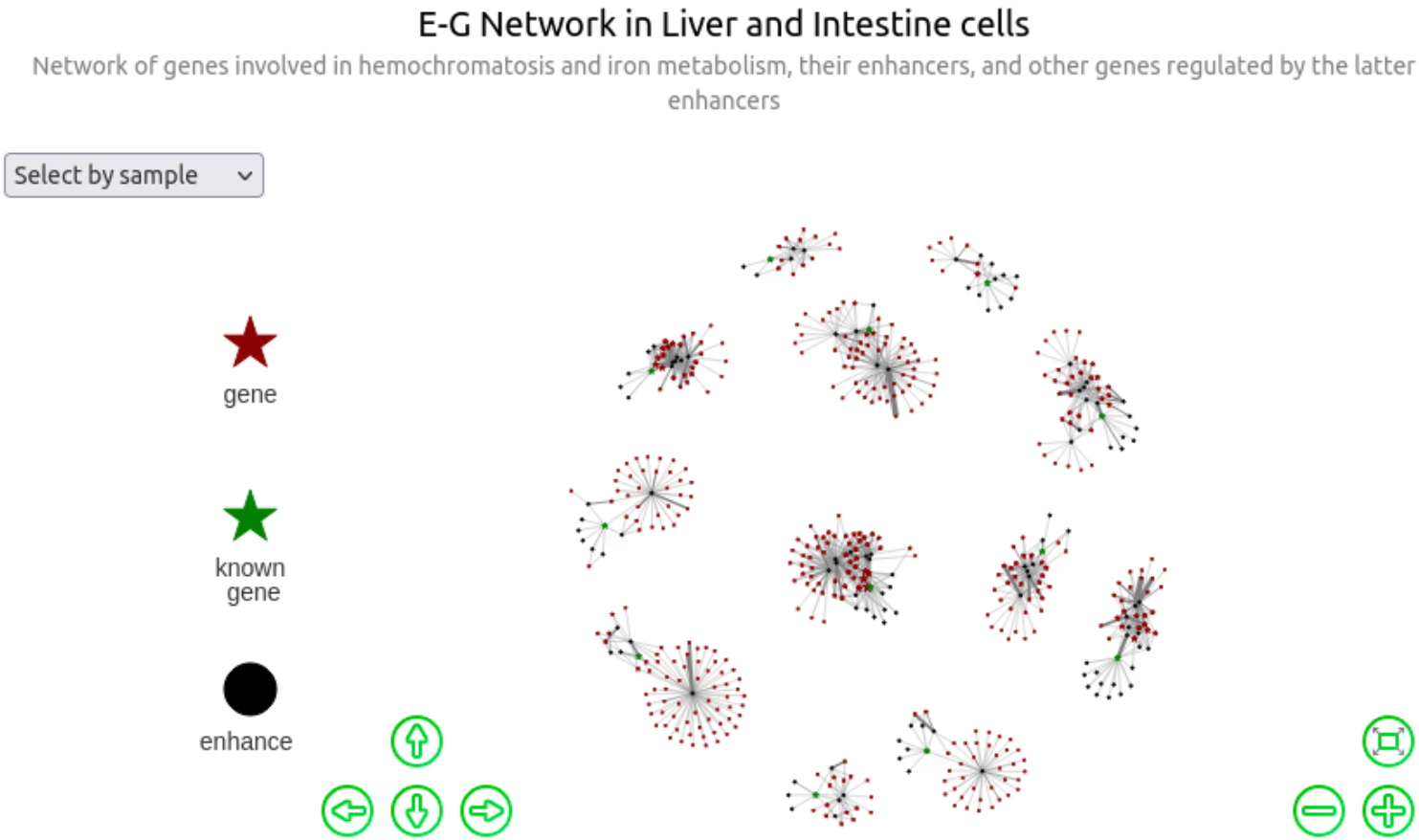

Figure 13: Graph obtained starting from the 12 initial genes. The weights of the edges are proportional to the corresponding max ABC scores (over all the instances of the E-G pairs found in the different biosamples. Note that, for a given biosample, the "ABC.score" column contains ABC scores that have already been averaged once when merging the enhancers).
`http://genoweb.toulouse.inra.fr/~thoellinger/2022/12_genes_liver_and_intestine_preliminary_analysis_v7.html`

Networks: results

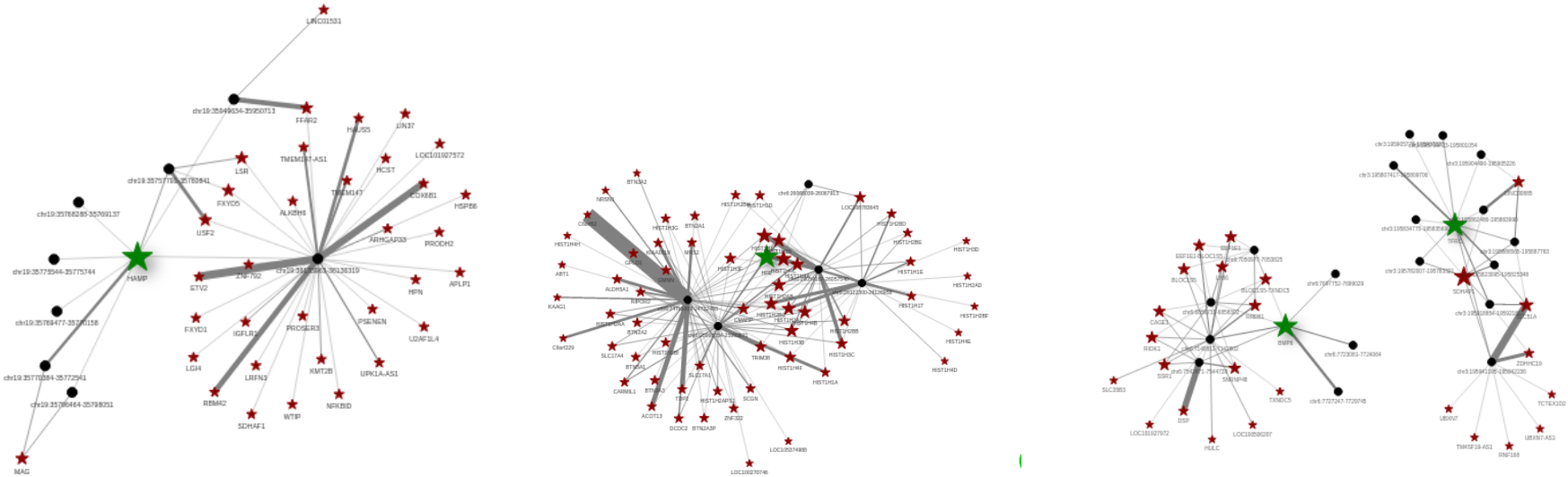

Figure 14: 3 connected compounds. `http://genoweb.toulouse.inra.fr/~thoellinger/fall_2021/10genes_liver_and_intestine_preliminary_analysis_v6.html`

Table of Contents

Data availability

All data were taken from [Jung et al. 2019].

Extracting data of interest

In the multiple biosamples provide by [Jung et al. 2019], only one corresponds to a tissue of interest for us, namely the liver.

Difficulties

- No difficulties for genes (thanks to the fact that the authors used ENSEMBL ids instead of gene symbols!)
- Same difficulties for putative enhancers as with Nasser et al. [2021]. Yet, Jung et al. [2019] use significantly larger enhancers (median 5,596 bp), so we decided to directly cast out those not overlapping any ccRE-ELS, and the corresponding data. At the end of the day, over 48,038 initial putative E-G pairs in liver, we kept 39,252 remaining pairs.

Networks: results

We found 42 genes using CHiC data.

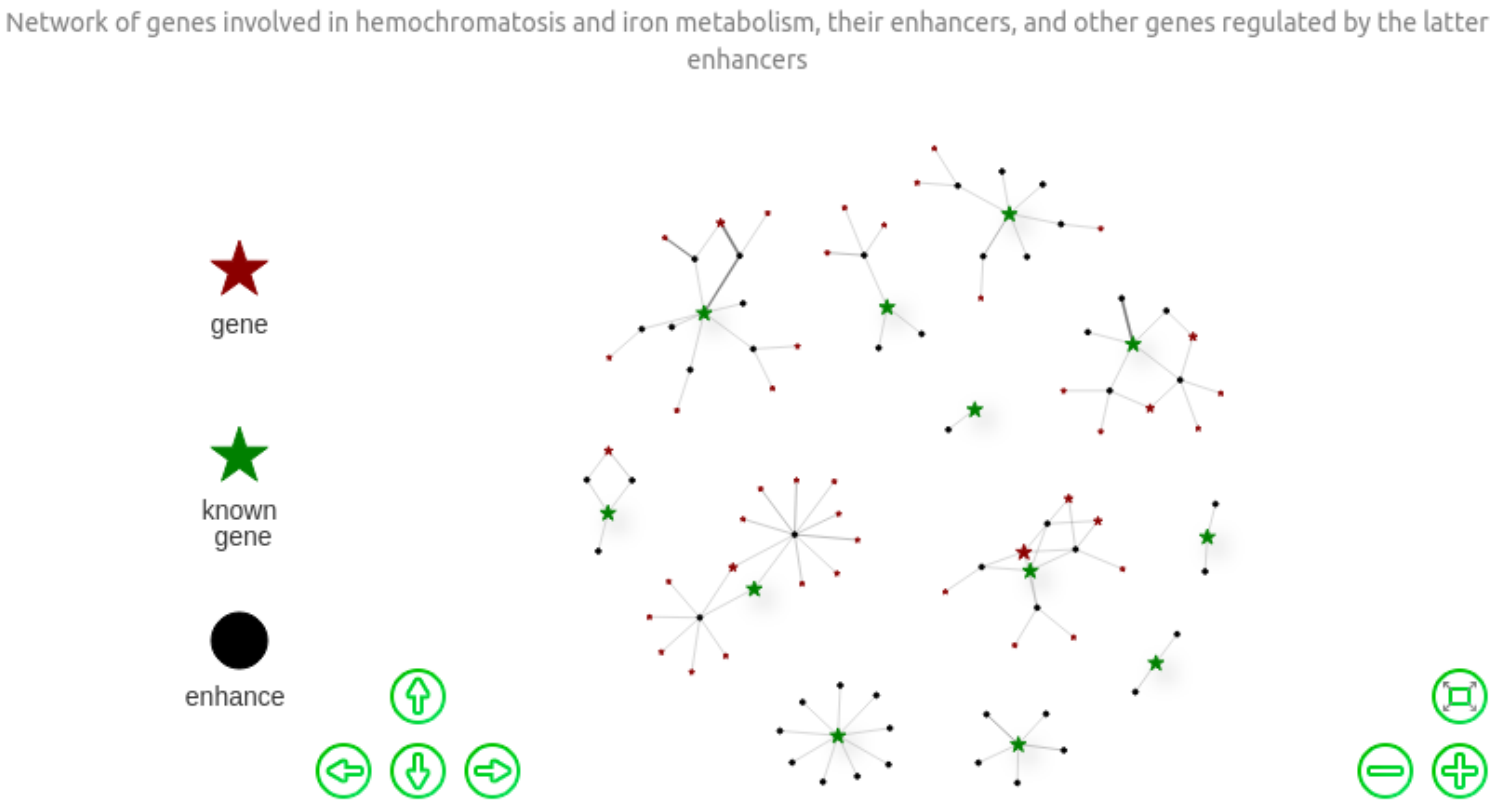

Figure 15: Graph obtained starting from the 12 initial genes. The weights of the edges are proportional to the corresponding distances). `http://genoweb.toulouse.inra.fr/~thoellinger/2022/12_genes_preliminary_analysis_chic_v1.html`

Table of Contents

Perspective

- We could use E-G pairs based on eQTL [null et al. 2020]
- Add a confidence score to predictions based on some of the following criteria:
  - Score of Enhancers from 0 to 2 if they intersect >= 1 ccRE-dELS, 0 ccRE-dELS but >=1 ccRE, 0 ccRE
  - Score of Genes from 0 to 2 if they are found in 1, 2 or 3 approaches (ABC, CHiC, eQTL)
  - Score of E-G pairs according to ABC scores
  - Score of E-G pairs according to contact scores
  - etc

  => Work in progress!

At the end of the day, we will obtain a list of genes pontentially linked to the severity of hemochromatosis, labeled by confidence.

Table of Contents

The Gene Ontology (GO)

The Gene Ontology (GO) is an ontology (in the form of an acyclic graph structure) of terms representing gene product properties, covering 3 domains:

- **biological process**, molecular events related to the functioning of integrated living units: cells, tissue, organs and organisms
- **molecular function**, the elemental activities of a gene product at the molecular level, such as binding or catalysis
- **cellular component**, the parts of a cell or a cellular environment

(Wikipedia)

We focus on biological processes.

GO terms enrichment

GO term enrichment is a technique for interpreting sets of genes making use of the Gene Ontology system of classification, in which the input gene set is compared with each of the bins (terms) in the GO – a statistical test can be performed for each bin to see if it is enriched for the input genes.

(Wikipedia)
The statistical test performed is generally a Fisher exact test or a hypergeometric test.

Correction for multiple testing

As the sets of genes are tested against multiple GO terms, we need to perform a multiple testing correction to control either the Family-Wise Error Rate (FWER) or the (positive) False Discovery Rate (FDR).

Let us denote V the number of false positives (test declared significant whereas the null hypothesis is true), and R the total number of test declared significant. We have:

$$\mathrm{pFDR} = \mathbb{E}[\frac{V}{R}|R > 0]$$

the expected proportion of false discoveries (for ≥ 1 discoveries)

$$\mathrm{FWER} = \mathbb{P}(V \geq 1)$$

the probability of making one or more false discoveries

Correction for multiple testing

Two common means to estimate and control the positive FDR are the q-value or the Benjamini-Hochberg (BH) procedure.

Issue about gene symbols

We wrote a script to compute GO enrichment using the R `clusterProfiler` package. The issue is that, over 422 genes (10 initial + 410 predicted) symbols, regargless if we try to first convert them to `entrez` id or to `ensembl` id, the script only finds between 299 and 310 of those genes in its GO database.

For that reason, we also used the online tool GOrilla: `http://cbl-gorilla.cs.technion.ac.il/`, which performs the same statistical test as we did but managed to recognize 396 out of 422 genes! (361 of which are associated with GO terms).

GOrilla results

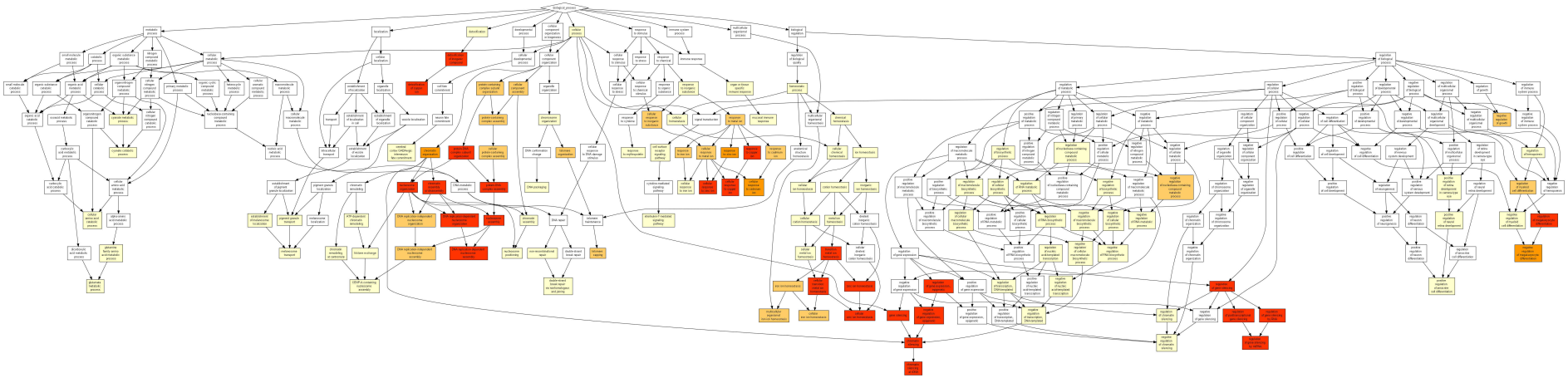

Figure 16: GO tree obtained with GOrilla, with respectively 324 target genes symbols and 23,221 background genes symbols as input (GOrilla recognized resp. 396 and 21,304 - 20,990 unique - of them). Only 17,621 background genes and 361 target genes were associated to $\geq$ GO term.

GOrilla results

Figure 17: GO treemap obtained with GOrilla, with respectively 324 target genes symbols and 23,221 background genes symbols as input (GOrilla recognized resp. 396 and 21,304 - 20,990 unique - of them). Only 17,621 background genes and 361 target genes were associated to $\geq$ GO term.

GOrilla results

Figure 18: GO treemap obtained with GOrilla, with respectively 410 target genes symbols and 23,221 background genes symbols as input (GOrilla recognized resp. 384 and 21,304 - 20,990 unique - of them). Only 17,621 background genes and 349 target genes were associated to $\geq$ GO term.

GOrilla results

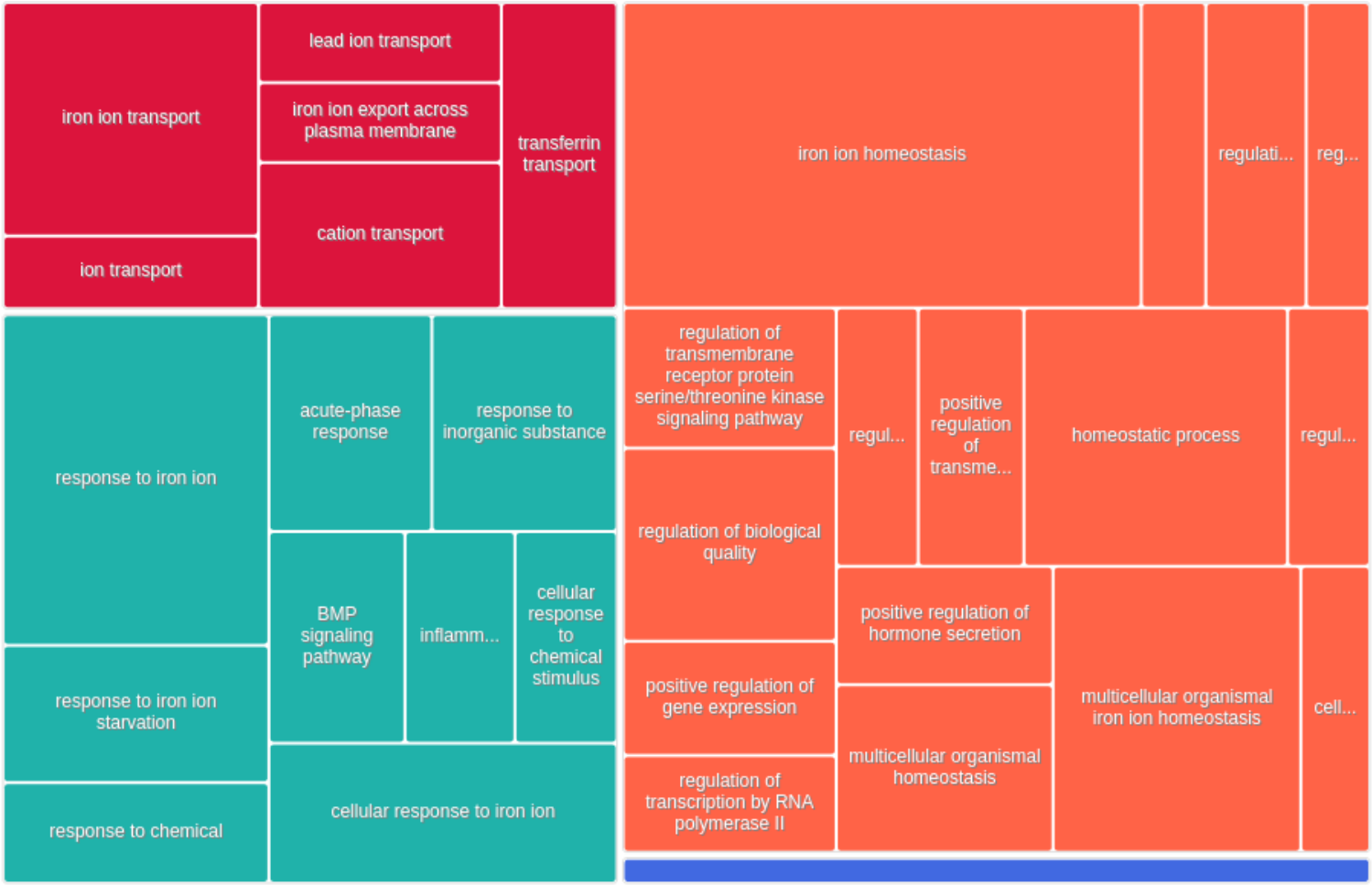

Figure 19: GO treemap obtained with GOrilla, with respectively 12 target genes symbols and 23,221 background genes symbols as input (GOrilla recognized resp. 12 and 21,304 - 20,990 unique - of them). Only 17,621 background genes and 12 target genes were associated to $\geq$ GO term.

Custom script: results

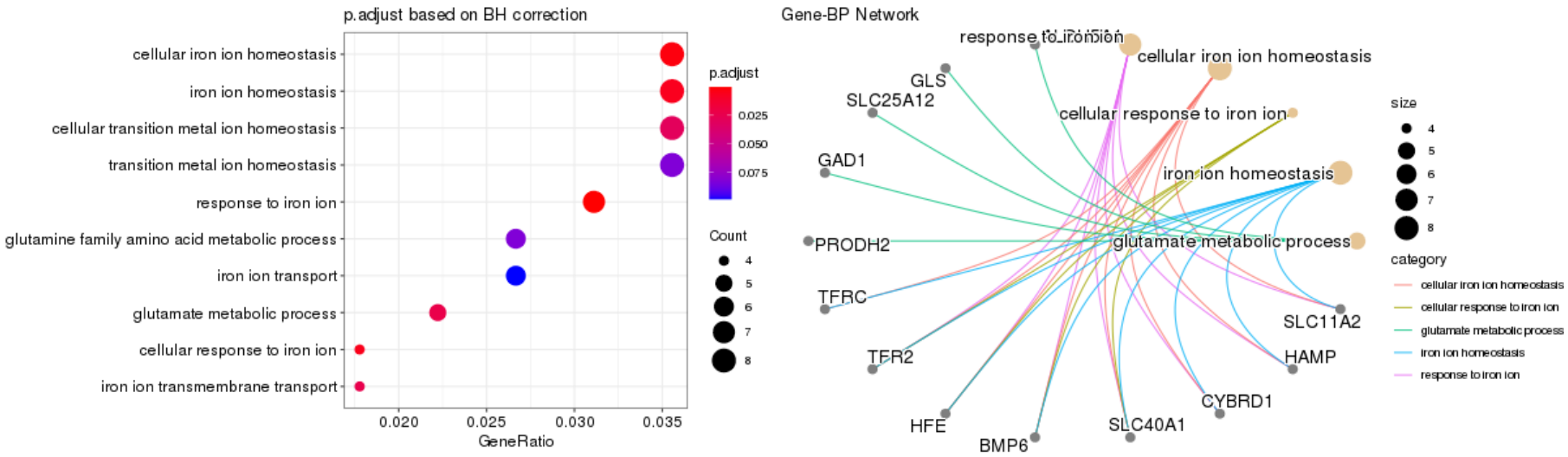

Figure 20: Enrichment in GO terms with custom script. 324 genes symbols provided. ≥ 1 GO terms found in database for 225 of them. Default universe of 18,666 background genes. Enrichment was tested for 2712 distinct GO terms. The 10 significant results have a FDR $\leq 10\%$.

Custom script: results

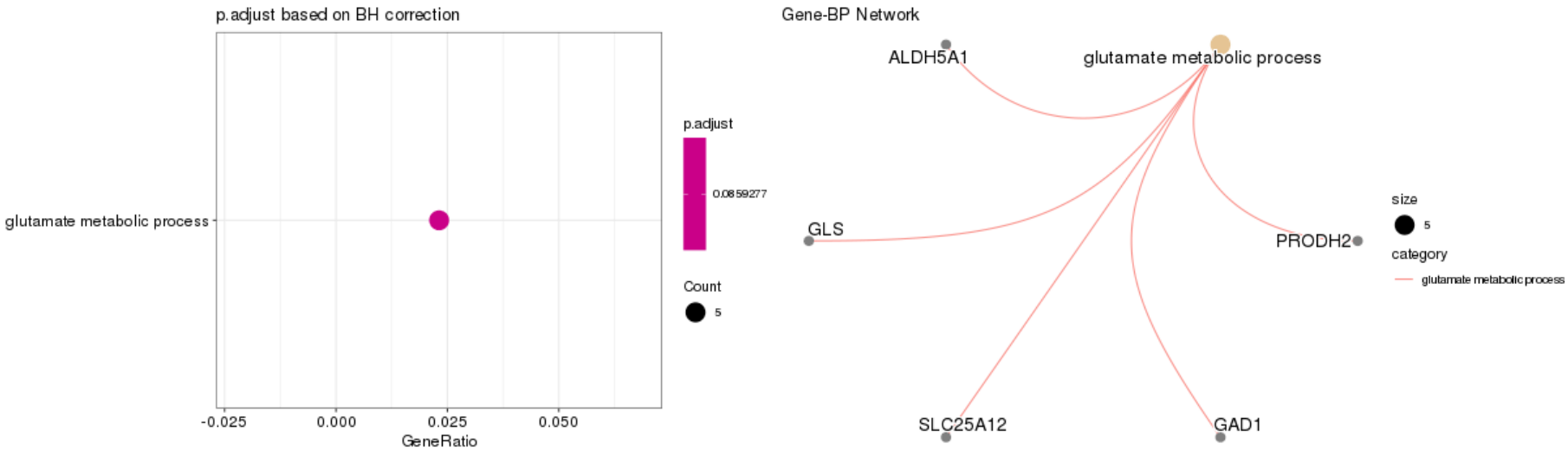

Figure 21: Enrichment in GO terms with custom script. 314 genes symbols provided. ≥ 1 GO terms found in database for 225 of them. Default universe of 18,666 background genes. Enrichment was tested for 2503 distinct GO terms. There is only 1 significant results have a FDR $\leq 10\%$ (actually, this is the only signifant results even at a FDR $\leq 20\%$).

Custom script: issues

It is a major issue that the GO database used by the package invoked in our script recognize only 225 out of our 314+10 infered genes. This can make a huge difference in computed enrichment, and it does, as a lot of GO terms are found with GOrilla against 10 resp. 1 only when using our script on the list of 324 resp. 314 genes.

So we shall better rely on GOrilla results.

Table of Contents

Perspective

- Compute the most comprehensive possible list of infered genes potentially linked through enhancers to that involved in haemochromatosis / in haemochromatosis severity, with multiple labels and scores for each of them according to the set of enhancers used, the number of enhancers regulating each genes, the ABC scores, etc
- Investigate specific properties of the initial genes (for instance: are they connected to more genes than other genes are?), and compare them to those of new infered genes
- Other perspectives?

Discussion

Thanks for listening!

Table of Contents

Bibliography I

[1] Alan P. Boyle et al. "High-resolution mapping and characterization of open chromatin across the genome". eng. In: *Cell* 132.2 (Jan. 2008). S0092-8674(07)01613-3[PII], pp. 311–322. ISSN: 1097-4172. DOI: 10.1016/j.cell.2007.12.014. URL: https://doi.org/10.1016/j.cell.2007.12.014.

[2] Kyle Kai-How Farh et al. "Genetic and epigenetic fine mapping of causal autoimmune disease variants". In: *Nature* 518.7539 (Feb. 2015), pp. 337–343. ISSN: 0028-0836. DOI: 10.1038/nature13835. URL: https://europepmc.org/articles/PMC4336207.

[3] Charles P. Fulco, Mathias Munschauer, et al. "Systematic mapping of functional enhancer–promoter connections with CRISPR interference". In: 354.6313 (2016), pp. 769–773. ISSN: 0036-8075. DOI: 10.1126/science.aag2445. URL: science.sciencemag.org/content/354/6313/769.

Bibliography II

[4] Charles P. Fulco, Joseph Nasser, et al. "Activity-by-Contact model of enhancer specificity from thousands of CRISPR perturbations". In: *Nature genetics* 51 (2019), pp. 1664–1669. DOI: 10.1038/s41588-019-0538-0. URL: www.nature.com/articles/s41588-019-0538-0.

[5] Molly Gasperini, Jacob M. Tome, and Jay Shendure. "Towards a comprehensive catalogue of validated and target-linked human enhancers". In: *Nature Reviews Genetics* 21.5 (May 2020), pp. 292–310. ISSN: 1471-0064. DOI: 10.1038/s41576-019-0209-0. URL: https://doi.org/10.1038/s41576-019-0209-0.

Bibliography III

[6] François Jacob and Jacques Monod. "Genetic regulatory mechanisms in the synthesis of proteins". In: *Journal of Molecular Biology* 3.3 (1961), pp. 318–356. ISSN: 0022-2836. DOI: https://doi.org/10.1016/S0022-2836(61)80072-7. URL: https://www.sciencedirect.com/science/article/pii/S0022283661800727.

[7] Inkyung Jung et al. "A compendium of promoter-centered long-range chromatin interactions in the human genome". In: *Nature Genetics* 51.10 (Oct. 2019), pp. 1442–1449. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0494-8. URL: https://doi.org/10.1038/s41588-019-0494-8.

[8] M Mercola et al. "Transcriptional enhancer elements in the mouse immunoglobulin heavy chain locus". In: *Science* 221.4611 (1983), pp. 663–665. ISSN: 0036-8075. DOI: 10.1126/science.6306772. URL: science.sciencemag.org/content/221/4611/663.

Bibliography IV

[9] Jill E. Moore et al. "A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods". In: *Genome Biol* 21.17 (2020). DOI: 10.1186/s13059-019-1924-8. URL: genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1924-8.

[10] P. Moreau et al. "The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants". In: *Nucleic Acids Research* 9.22 (Nov. 1981), pp. 6047–6068. ISSN: 0305-1048. DOI: 10.1093/nar/9.22.6047. eprint: https://academic.oup.com/nar/article-pdf/9/22/6047/7051707/9-22-6047.pdf. URL: https://doi.org/10.1093/nar/9.22.6047.

Bibliography V

[11] Joseph Nasser et al. "Genome-wide enhancer maps link risk variants to disease genes". In: *Nature* 593.7858 (May 2021), pp. 238–243. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03446-x. URL: https://doi.org/10.1038/s41586-021-03446-x.

[12] null null et al. "The GTEx Consortium atlas of genetic regulatory effects across human tissues". In: *Science* 369.6509 (2020), pp. 1318–1330. DOI: 10.1126/science.aaz1776. eprint: https://www.science.org/doi/pdf/10.1126/science.aaz1776. URL: https://www.science.org/doi/abs/10.1126/science.aaz1776.

[13] Stefan Schoenfelder and Peter Fraser. "Long-range enhancer–promoter contacts in gene expression control". In: *Nature Reviews Genetics* 20.8 (Aug. 2019), pp. 437–455. ISSN: 1471-0064. DOI: 10.1038/s41576-019-0128-0. URL: https://doi.org/10.1038/s41576-019-0128-0.

Bibliography VI

[14] The ENCODE Project Consortium. “Expanded encyclopaedias of DNA elements in the human and mouse genomes”. In: 583 (2020), pp. 699–710. URL: https://www.nature.com/articles/s41586-020-2493-4.