TFEA.ChIP

A tool kit for transcription factor binding site enrichment analysis capitalizing on ChIP-seq datasets

Laura Puente-Santamaria, Luis del Peso

February 15, 2018

Abstract

The identification of transcription factor (TF) responsible for the co-regulation of an specific set of genes is a common problem in transcriptomics. With the development of TFEA.ChIP we aim to provide a tool to estimate and visualize TF enrichment in a set of differentially expressed genes that takes into account the wide variation in TF's behavior across different cell types and stimuli. To that end, ChIP-Seq experiments from the ENCODE Consortium and GEO Datasets were gathered, and a database linking TFs with the genes they interact with in each ChIP-Seq experiment was generated. In its current state, TFEA.ChIP covers 333 different transcription factors in 1122 ChIP-Seq experiments, with over 150 cell types being represented.

Include here comparison with Oposssum and results.

1 Introduction

In the most simple scenario, the comparison of the transcriptome of cells or organisms in two conditions leads to the identification of a set of differentially expressed (DE) genes, and the underlying assumption is that one or a few TFs regulate the expression of those genes. Traditionally, the identification of relevant TFs has relied on the use of position weight matrices (PWMs) to predict transcription factor binding sites (TFBSs) proximal to the DE genes[1]. The comparison of predicted TFBS in DE versus a set of control genes, reveals factors that are significantly enriched in the DE gene set. The prediction of TFBS using these approaches have been useful to narrow down potential binding sites, but can suffer from high rates of false positives. In addition, this approach is limited by design to sequence-specific transcription factors (TF) and thus unable to identify cofactors that bind indirectly to target genes. To overcome these limitations we developed the R package TFEA.ChIP, which exploits the vast amount of publicly available ChIP-Seq datasets to determine TFBS proximal to a given set of genes and computes enrichment analysis based on this experimentally-derived rich information. Specifically TFEA.ChIP, uses information derived from the hundreds of ChIP-Seq experiments from the ENCODE Consortium[2] expanded to include additional datasets contributed to GEO database[3][4] by individual laboratories representing the binding sites of factors not assayed by ENCODE. The package includes a set of tools to customize the ChIP data, perform enrichment analysis and visualize the results. Herein we describe the main characteristics of the package and compare the results produced by TFEA.ChIP vs those generated by Oppossum, an state of the art TFBS identification software based on PWMs

[5]. Our data indicate that the results of TFEA.ChIP and Opossum are coincident for those datasets where Oppossum identifies clear TFBS candidate(s). In addition, TFEA.ChIP identified enriched factors for some data sets where Opossum was unable to find a significant match.

The package implements two enrichment analysis methods:

- Analysis of the association of TFBS and differential expression from 2x2 tables recording the presence of binding sites for a given TF in DE and control genes. The statistical significance of the association for each factor determined by a Fishers exact test.
- GSEA analysis, based on the core function of the GSEA algorithm for R[6][7], GSEA.EnrichmentScore.

TFEA.ChIP includes a TF-gene interaction database containing 1122 datasets from ChIP-Seq experiments testing 333 different human transcription factors. Although the package is mainly focused towards analyzing expression data generated from human cells, TFEA.ChIP includes the option to use datasets coming from experiments in mice, translating mouse gene names to their equivalent ID on the human genome.

Indicate that the factors found by TFEA.ChIP are the ones expected according to the experimental manipulations in each dataset.

I would move the following paragraphs to results.

Adding mention to web implementation if done.

2 Building our TFBS database

The first source of ChIP-Seq datasets is Encode Uniform TFBS database, which guarantees a standard procedure to gather, filter, and share information from ChIP-Seq experiments. However, taking into consideration that the current estimations of the amount of transcription factors in the human genome range from 1,391[8] -manually curated candidates- to 2886 -predicted through computational methods by DBD[9][10] -, the 157 transcription factors covered by Encodes database were not considered enough to build a comprehensive TF enrichment analysis tool. In order to expand the scope of our TFBS database, we also included datasets from ChIP-Seq experiments stored in GEO. In total, 1122 ChIP-Seq datasets, 689 from Encode and 433 from GEO DataSets, make up the source of information to generate this database, covering 333 different transcription factors in a variety of cell types and experimental conditions. The process to establish a link

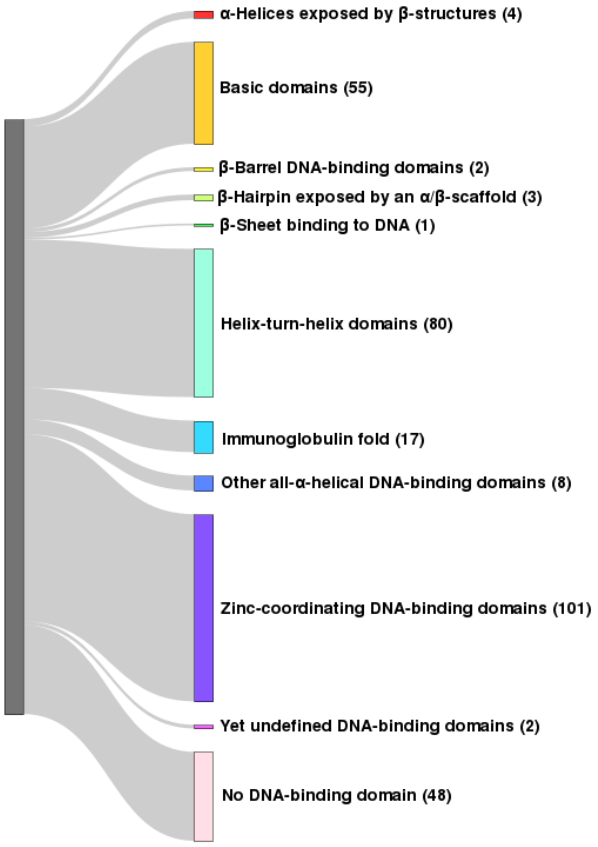

Figure 1: Structural diversity according to DNA binding domains of the 333 transcription factors included in the TFBS database.

between a peak in a ChIP-Seq experiment and a specific gene goes as follows:

1. Generating a Dnase Hypersensitive Sites database, linking each Dnase HS to the nearest gene of those included in UCSC Known Gene database[11]. During this process, DHSs that were farther than 1Kb from any gene were discarded, so as to avoid highly uncertain connections that would undermine the robustness of any analysis. In the case of a DHSs close enough to more than one gene, both were assigned to the site. For this purpose we used Encodes Master DNaseI HS database[12][13].

2. Selecting from each ChIP-Seq dataset those peaks that overlap a DHSs. Each of these peaks will be assigned the same gene as the DHS they overlap.

3. Storing the list of genes assigned to a peak in each of the ChIP-Seq experiments. With this lists we generated a binary matrix which rows correspond to all the human genes in the Known Gene database, and its columns, to every ChIP-Seq experiment; the values assigned are 1 for a peak assigned to that gene or 0.

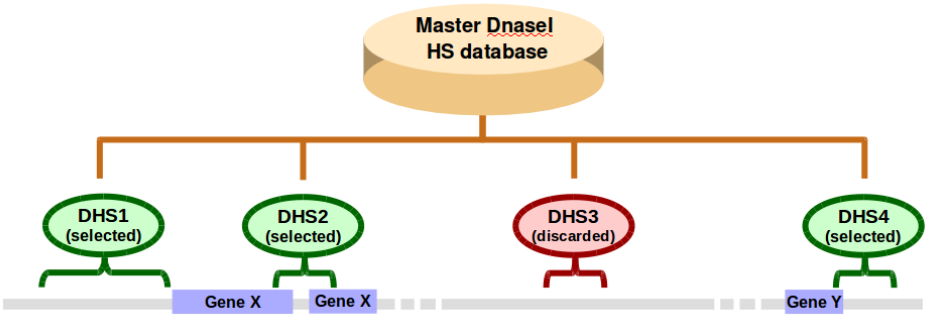

Figure 2: Building a DHS database. From Encodes Master DNaseI HS database, that gathers all DHSs found in several cell lines, we selected those DHSs that are 1Kb or closer to a gene according to the genes location on UCSCs Known Gene database. DHS further than 1Kb from any gene are discarded to avoid highly uncertain links. In this illustration, both DHS1 and DHS2 would be assign to gene X, and DHS4 to gene Y, while DHS3 would be discarded.

3 Analyzing TF enrichment with TFEA.ChIP

TFEA.ChIP is designed to take the output of a differential expression analysis and identify transcription factors enriched in the list of differentially expressed genes. The core premise of our method is that key effectors of a regulatory response will have more target genes among the differentially expressed than among the unresponsive genes. In the case of analysis of association, the only required input is a set of DE genes and, optionally, a set of control genes whose expression is not altered by the experimental conditions under the study. For the GSEA analysis a ranked list of genes is required. This is supplied as a matrix or data frame containing a column with gene names and a numerical column with the ranking metric, which typically are log2(Fold change) for the gene expression changes in the two conditions under evaluation.

Figure 3: Building the TFBS database. For every peak in a ChIP-Seq dataset is tested whether it overlaps any of the DHSs located close to a gene. If the result is positive, the gene corresponding to said DHS gene X in this illustration will be assigned to the ChIP-Seq experiment, if its negative, that peak will be discarded. The ChIP-Seq dataset Y has three peaks that overlap DHS1 and DHS2, so the Entrez ID of gene X would be associated to the ChIP-Seq Y.

3.1 Association analysis

The simplest approach to transcription factor enrichment consist on comparing how many targets of a given transcription factor are in two lists of genes. This is the course of action taken in this method, focused on finding differences in transcription factor enrichment between differentially expressed genes (be it up-regulated, down-regulated or both) and a control group. To that end, the program generates contingency matrices for every ChIP-Seq experiment in the database, estimating statistical significance using Fisher's exact test is performed.

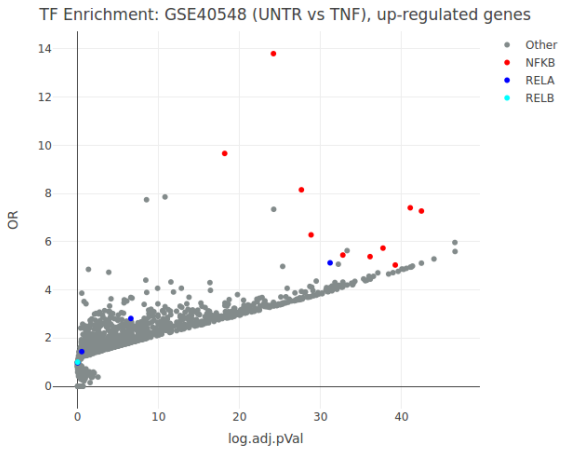

Figure 4: Association analysis performed on up-regulated vs un-responsive genes in neutrophils after TNF addition. As expected, the main enriched TF -NFKβ and RELA- are related with inflammatory processes.

3.2 GSEA-like analysis

This method is based on the same principle than GSEA[6][7]: having the list of genes of an RNA-seq experiment ordered by log2(Fold Change), for every ChIP-Seq in the database, the algorithm starts by calculating how many matches and mismatches there are between the gene list provided and the genes associated to each ChIP-Seq experiment in the database. With the amounts of matches and mismatches, it generates a match score and a mismatch score, so that Enrichment Score is always delimited between 1 and -1. The result is an array in which every time a gene in the list provided is associated

to the ChIP-Seq experiment, the match score is added, while on the rest of cases, the mismatch score is subtracted -Running Enrichment Score or *RES*. The final Enrichment Score given to each experiment is the absolute maximum of the RES.

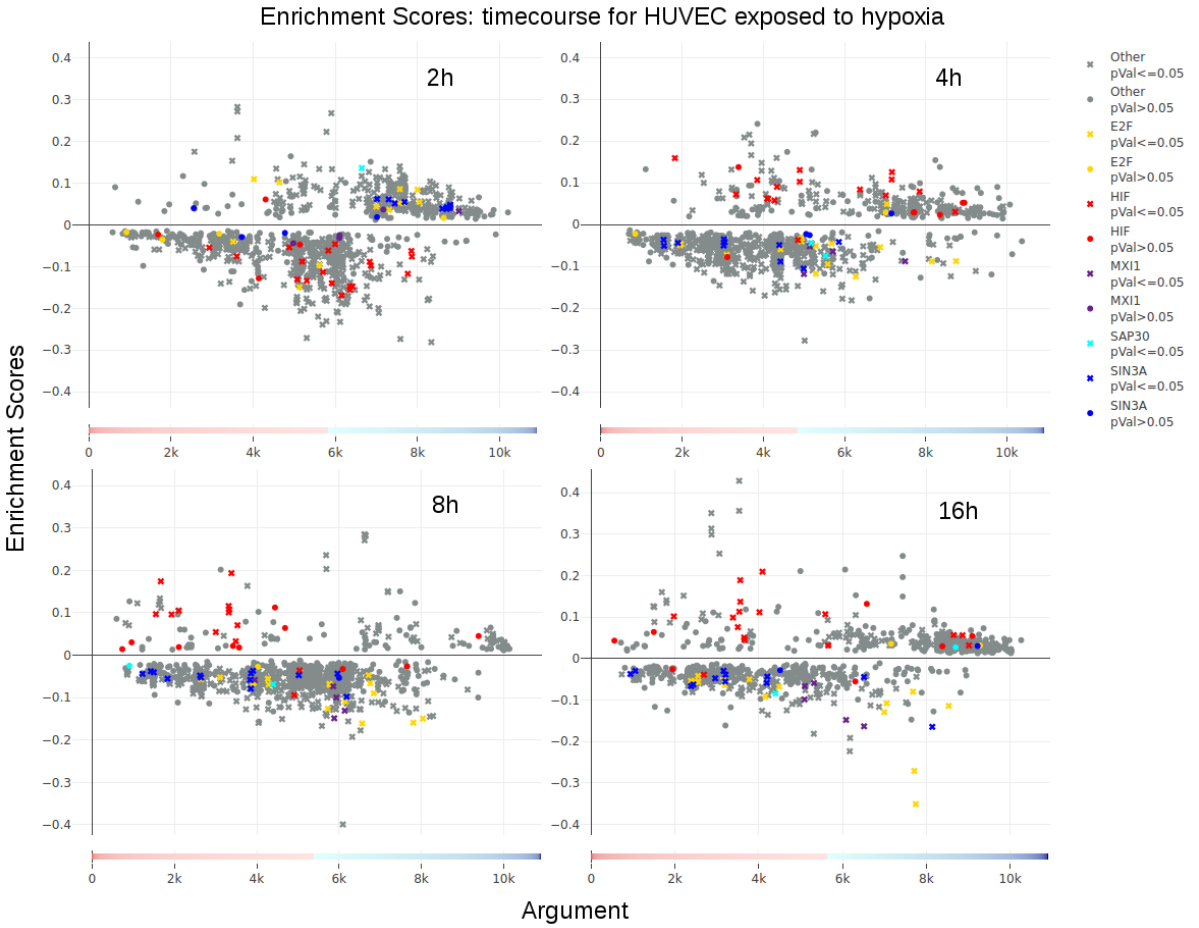

Figure 5: GSEA-like analysis of gene expression on HUVEC exposed to hypoxia vs normoxia at several times. As is shown in the figure, HIF-mediated response to hypoxia starts at 4h, and reaches it's peak at 8h, when most of it's targets are among up-regulated genes. The same pattern of behavior can be observed in repressors, that reach their maximum enrichment among down-regulated genes between 8 and 16h after exposure.

4 Comparison with similar software: oPOSSUM's single site analysis

To test TFEA.ChIP's performance, we decided to compare its TF enrichment estimations with a program that uses a related approach to measuring TF enrichment: oPOSSUM's Single site analysis (SSA)[14]. In contrast with our package, oPOSSUM's SSA searches for potential TF binding sites in the sequences of a given set of genes compared to a background set.

Several validation sets were used, mainly focusing on upregulated vs. unresponsive genes (with one case of downregulated vs. unresponsive genes during hypoxia). These validation sets include:

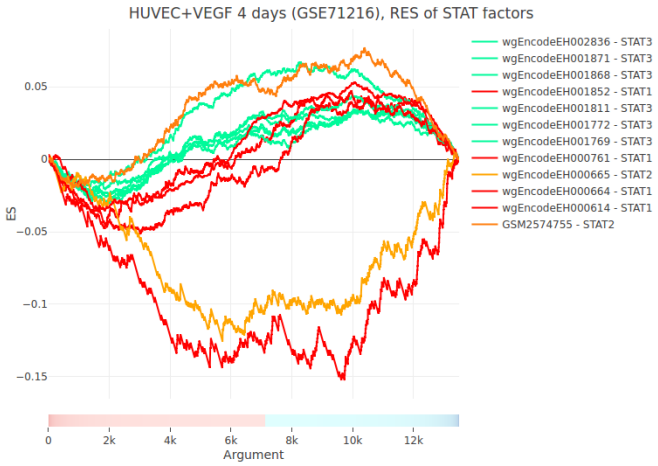

Figure 6: plot of the Running Enrichment Score in an RNA-seq experiment on HUVEC cells supplemented with VEGF for 4 days for STAT1, STAT2 and STAT3 ChIP-Seqs. The ChIP-Seqs wgEncodeEH000664 and wgEncodeEH000665 were done with a IFN treatment while the rest of STAT1 and STAT2 ChIPs were done using IFN. Since IFN is a known angiogenesis inhibitor, the targets of this two ChIPs are significantly different than those of the ChIPs done with IFN, and in this case, amongst the down-regulated genes.

- Hypoxia vs normoxia in HUVEC cells.
- TNF addition in neutrophils and adipocytes.
- Left ventricular non-compaction in cardiomyocytes.
- INF addition in hESC cells for 15 and 21 days.

To compare both methods we generate contingency matrices with the number of target hits, target non-hits, background hits, and background non-hits for every profile and ChIP- Seq experiment, and then performed Fishers exact test. The resulting p-values are adjusted for multiple testing using FDR method.

Hypoxia vs normoxia in HUVEC cells

oPOSSUM 3.0 its limited by design to transcription factors that bind directly to a specific DNA sequence, so important cofactors during hypoxia, like SIN3A, that dont have a DNA binding domain, are excluded from the analysis. Since the source of information for TFEA.ChIP are ChIP-Seq experiments, were able to collect binding information for TFs that arent sequence specific or depend on other cofactors to function.

Using significantly up-regulated genes, both TFEA.ChIP and oPOSSUM are able to detect and enrichment in ARTN targets, but TFEA.ChIP also detects significant enrichment in HIF1A, EPAS1, and EZH2. The last transcription factor is part of a histone modification complex, and doesnt bind directly to DNA, so is not included in oPOSSUMs database.

TNF addition in neutrophils and adipocytes

In case of genes upregulated by TNF addition, both oPOSSUM and TFEA.ChIP are able to detect a significant enrichment in NfB, REL and between 5 and 20 more TFs depending on the experiment. Most of the transcription factors detected only by one of the methods are not included in the other ones database. This shows that, when a transcription factors response doesnt change much depending on stimuli or tissue, both methods perform similarly.

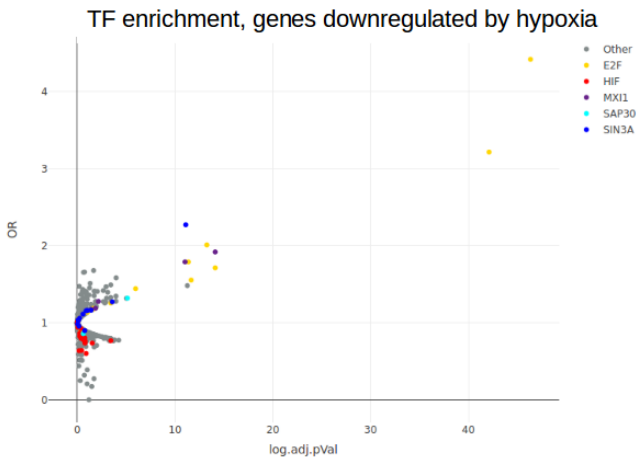

Figure 7: For down-regulated genes, after adjusting p-values, oPOSSUM doesnt find any TF significantly enriched, while TFEA.ChIP detects significant enrichment in 111 different TFs, being the most important E2F4, E2F7, SIN3A, MXI1, and E2F1.

Left ventricular non-compaction in cardiomyocytes

In a complex case, such as left ventricular non-compaction, both methods are able to detect a significant enrichment in HIF factors as expected, since LVNC is associated to hipoxemic conditions along with another 17 significantly enriched or depleted factors. Of the 71 TF enriched TFs detected by oPOSSUM, 29 are not in TFEA.ChIPs database, while the same happens with most of the enriched transcription factors detected by TFEA.ChIP, such as KDM3A or EZH2.

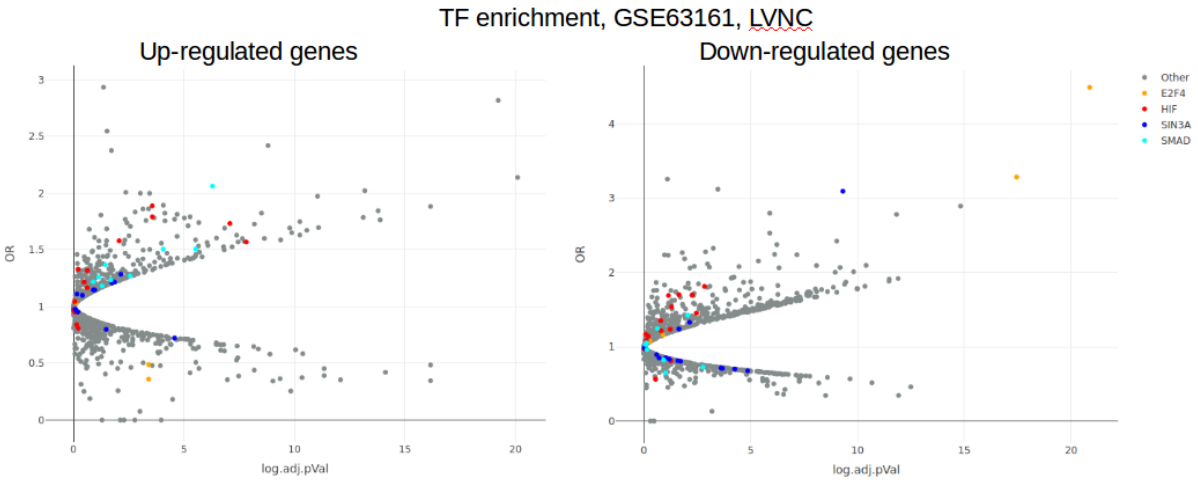

Figure 8: Enrichment analysis in LVNC. TFEA.ChIP is able to find several enriched TF. Some of the most important are HIF factors (in red) along with SMAD2, SMAD4 (cyan) and E2F4 (yellow), while SIN3A only has one ChIP-Seq experiment among those highly enriched.

INF addition in hESC cells for 15 and 21 days

As previous studies have concluded[15], regulation networks can differ greatly depending on the tissue and the stimuli applied. An example of this behavior would be Signal Transducer And Activator Of Transcription STAT factors, such as STAT1 Or STAT2, whose targets change depending on the interferon type used as stimulus. This behavior can be seen in figure 9.

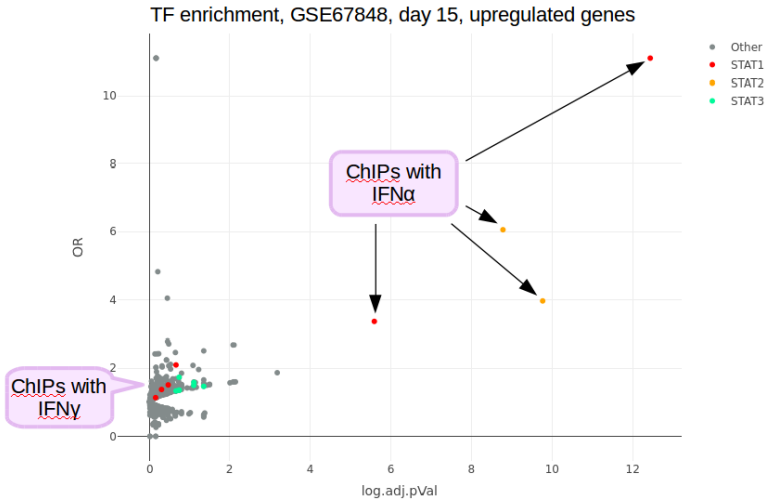

Figure 9: Validation dataset GSE67848, hESC + IFN, day 15, upregulated genes. Two of the STAT1 ChIP-Seq experiments and the STAT2 ones were done after IFN addition while the rest of STAT1 ChIPs were treated with IFN. In both association analysis and GSEA-like analysis, TFEA.ChIP is able to detect conditional TF enrichment.

Conclusion

In general, TFEA.ChIP allows to take into account cellular context (tissue, stimuli) on which an expression experiment is done, making possible to detect enrichment in transcription factors in cases where oPOSSUM cant, due to:

- Indirect binding: in case of complexes such as the Polycomb Repressive Complex or SIN3A, that is a co-factor of MXI1.
- Stimuli-dependent binding: as seen previously with STAT and IRF factors.

On another note, oPOSSUM allows the user to have more control over some of the parameters used for the analysis (conservation cutoff or matrix score threshold, for instance) while TFEA.ChIP analysis only takes as input two lists of genes. More over,since TFEA.ChIPs sources are treatment and tissue specific, in some cases the result might not be applicable to the dataset used as input (i.e., tissues or cell types that have a very particular regulatory pattern, when there are no ChIP-Seq experiments done in said cell type).

5 Dependencies and requirements

The following R packages are used to run TFEA.ChIP:

1. Packages part of Bioconductor[16]:
    - GenomicRanges, IRanges, and GenomicFeatures[17]
    - biomaRt[18][19]
    - TxDb.Hsapiens.UCSC.hg19.knownGene[20]
    - org.Hs.eg.db[19]
5. plotly[21]
6. dplyr[22]

7. knitr[23]
8. rmarkdown[24]
9. S4Vectors[25]
10. scales[26]

I'll add here what R version the package runs on (currently in Bioconductor's development version is R $\geq$ 3.5, but in the release branch is usually a lower version). If a web implementation it's done it will be mentioned here to.

References

[1] WW Wasserman and A Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5:276, 2004.

[2] ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57–74, 2004.

[3] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30:207–210, 2002.

[4] Tanya Barrett et al. NCBI GEO: archive for functional genomics data sets - update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.

[5] Andrew T Kwon, David J Arenillas, Rebecca Worsley Hunt, and Wyeth W Wasserman. oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3 (Bethesda, Md.)*, 2(9):987–1002, 2012.

[6] Aravind Subramanian, Pablo Tamayo, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005.

[7] Vamsi K Mootha et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34:267–273, 2003.

[8] Juan M. Vaquerizas, Sarah K. Kummerfeld, Sarah A. Teichmann, and Nicholas M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10:252–263, 2009.

[9] Sarah K. Kummerfeld and Sarah A. Teichmann. DBD: a transcription factor prediction database. *Nucleic Acids Research*, 34(suppl 1):74–81, 2006.

[10] Derek Wilson, Varodom Charoensawan, Sarah K. Kummerfeld, and Sarah A. Teichmann. DBD - taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Research*, 36(suppl 1):88–92, 2008.

[11] Fan Hsu, W. James Kent, Hiram Clawson, Robert M. Kuhn, Mark Diekhans, and David Haussler. The ucsc known genes. *Bioinformatics*, 22(9):1036–1046, 2006.

[12] Sam John et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics*, 43:264–268, 2011.

[13] Robert E. Thurman et al. The accessible chromatin landscape of the human genome. *Nature*, 489:75–82, 2012.

[14] Andrew T. Kwon, David J. Arenillas, Rebecca Worsley Hunt, and Wyeth W. Wasserman. oPOSSUM-3: Advanced Analysis of Regulatory Motif Over-Representation Across Genes or ChIP-Seq Datasets. *G3: Genes, Genomes, Genetics*, 2(9):987–1002, 2012.

[15] Abhijeet Rajendra Sonawane. Understanding tissue-specific gene regulation. *Cell Reports*, 21(4):10771088, 2017.

[16] W. Huber et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121, 2015.

[17] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin Morgan, and Vincent Carey. Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9, 2013.

[18] Steffen Durinck, Paul T. Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols*, 4:1184–1191, 2009.

[19] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21:3439–3440, 2005.

[20] Marc Carlson and Bioconductor Package Maintainer. *TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s)*, 2015.

[21] Carson Sievert, Chris Parmer, Toby Hocking, Scott Chamberlain, Karthik Ram, Marianne Corvellec, and Pedro Despouy. *plotly: Create Interactive Web Graphics via 'plotly.js'*, 2017.

[22] Hadley Wickham, Romain Francois, Lionel Henry, and Kirill Mller. *dplyr: A Grammar of Data Manipulation*, 2017.

[23] Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2017.

[24] JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, and Winston Chang. *rmarkdown: Dynamic Documents for R*, 2017.

[25] H. Pags, M. Lawrence, and P. Aboyoun. *S4Vectors: S4 implementation of vector-like and list-like objects*, 2017.

[26] Hadley Wickham. *scales: Scale Functions for Visualization*, 2017.

Todo list