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

Laura Puente-Santamaria[1], Luis del Peso[1,2,3*]

1 Departamento de Bioquímica, Universidad Autónoma de Madrid (UAM) and Instituto de Investigaciones Biomédicas 'Alberto Sols' (CSIC-UAM), 28029 Madrid , Spain
2 IdiPaz, Instituto de Investigación Sanitaria del Hospital Universitario La Paz, 28029 Madrid, Spain.
3 CIBER de Enfermedades Respiratorias (CIBERES), Instituto de Salud Carlos III, 28029 Madrid, Spain.

* luis.peso@uam.es

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 from 1122 ChIP-Seq experiments, with over 150 cell types being represented. The analysis of publicly available RNAseq datasets, including hypoxic transcriptomes and response to cytokines, show that TFEA.ChIP accurately identifies the relevant transcription factors in each case. The use of ChIP-Seq data instead of PWM-based methods allows to expand the analysis of TF enrichment to include cofactors that lack a DNA binding domain as well as chromatin modifiers, in addition to provide a biological context to infer tissue and stimuli-dependent TF behavior. TFEA.ChIP is available as a Bioconductor package at:
`https://www.bioconductor.org/packages/devel/bioc/html/TFEA.ChIP.html`

Include
parison
sum an

Author summary

[I believe this is required for research articles, but not software papers.]

Method
and Sof
cles req
submiss
quiries.

Introduction

In the most simple scenario, the comparison of the transcriptome of samples 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.

Design and implementation

Database

TFEA.ChIP package includes analysis and visualization tools intended for the identification of TFBS enriched in a set of DE genes. To this end, the package uses experimental information derived from 1122 ChIP-seq datasets, generated by the ENCODE consortium and individual researchers, testing a total of 333 individual human transcription factors in a variety of cell types and experimental conditions. The collection of TF includes both sequence-specific transcription factors as well as molecules that bind DNA indirectly (e.g. transcription cofactors and chromatin modifiers) or in a sequence-independent fashion (e.g. RNA Polymerases). This is an important difference with methods based on PWMs that, by design, are restricted to sequence-specific factors. The database includes a total of XX of sequence-specific TF (XX% of total factors). Thus, this compiled database covers XX-XX% of the 1,391 [6] to 1600 [7] transcription factors estimated to be encoded by the human genome. In addition, the set includes proteins from all the major classes of DNA binding domains (Fig 1).

The supplementary table S1 Table contains the complete list of the datasets included in the package along with their GEO accession numbers. . ChIPseq datasets contain the coordinates of TF binding sites throughout the genome. Thus, in order to use this information in gene enrichment analyses, we first need to associate binding regions (ChIP-peaks) to specific genes. In the absence of three-dimensional contact information, such as that produced by chromosome conformation capture carbon copy (Hi-C) experiments, the peaks are usually assigned to the nearest gene. However, Hi-C experiments indicate that only a small fraction of the looping interactions of distant regulatory regions are with the nearest gene [8]. Accordingly, uncertainty of peak assignation decreases as the distance to the nearest gene increases [9]. A potential solution is to assign ChIP-peaks only to overlapping genes and remove all binding sites located far from genes. This strategy is expected to successfully work for TFs that tend to bind promoter and intronic regulatory regions, but it is bound to have poor performance for factors that preferentially associate to distant enhancers. To overcome these difficulties, we exploited the extensive map of enhancer-target gene pairs

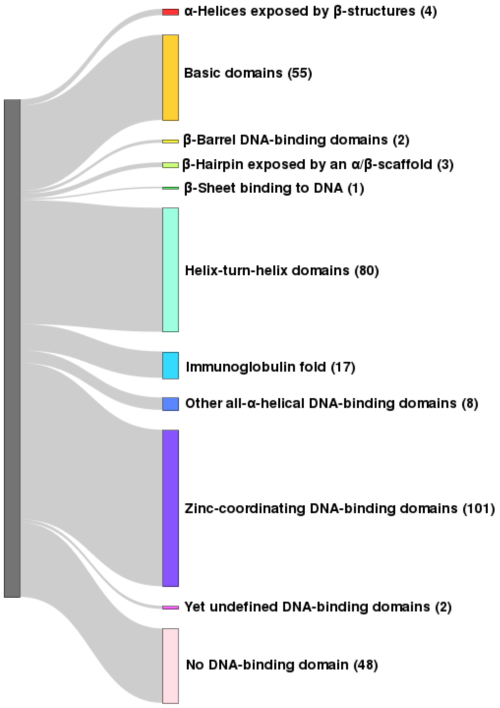
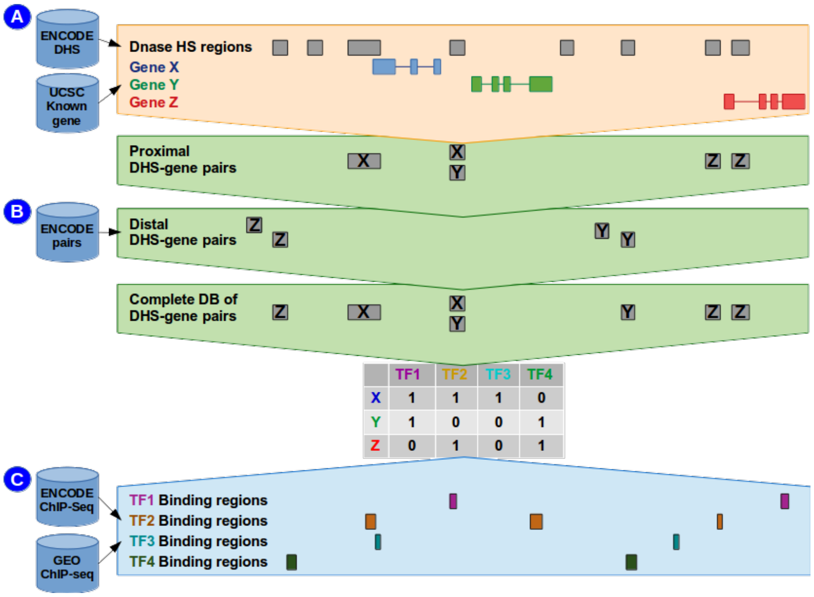


Fig 1. Structural diversity according to DNA-binding domains of the transcription factors included in the TFBS database. The 333 TFs included in TFEA.ChIP database were classified into families according to their DNA-binding domain composition. InterPro parent–child relationships between DNA-binding domains were used as the basis for TF family definition (Supplementary information S1 (PDF)). TFs with multiple DNA-binding domains were classified in each of their respective families. Families with less than five members were classified as 'other'.

generated by the ENCODE project through the analysis of correlation between the DHS signal at distant sites and gene promoters regions across 79 cell lines [**?**]. Thus, we first generated a database pairing open chromatin regions, as defined by clusters of Dnase Hypersensitive Sites (DHSs) e [10] [11], and genes in the UCSC Known Gene database (version 3.2.2) [12]. DHSs were assigned to genes overlapping with the open chromatin region tolerating a 1Kb margin from the gene boundaries and allowing for multiple gene assignation for those DHSs overlapping two genes. This process resulted in a database of DHSs-gene pairs that only retained DHSs that were assigned to at least one gene (Fig 2, step A). Next, we added to this database the list of statistically significant (Pearson's correlation coefficient ¿0.8) enhancer DHSs-gene pairs generated by ENCODE [**?**] (Fig 2, step B). Then, for each ChIP-seq dataset we selected those peaks that were statistically significant (FDR¡0.001 for ENCODE datasets and FDR¡0.05 for the rest of datasets) and overlapped an open chromatin region in the DHSs-gene database. Each of these peaks was assigned to the same gene as the DHS they overlapped (Fig 2, step C). Finally, we integrated the peak-gene information from all ChIP-dataset into a binary matrix with rows corresponding to all the human genes in the Known Gene database, and a columns for every ChIP-Seq experiment analyzed; the entry values were assigned to 1 when the row gene had at least one peak assigned in the ChIP-Seq column and 0 otherwise (Fig 2). It is worth noting that, as a result of the matching strategy, the interaction matrix contains a large fraction of TFBS-gene pairs involving distant regulatory interactions including intronic [9] and enhancer regions.



	TF1	TF2	TF3	TF4
X	1	1	1	0
Y	1	0	0	1
Z	0	1	0	1

Fig 2. Building Database of TF-gene associations. A: We selected open chromatin regions (defined as clusters of DHSs identified by the ENCODE project) that are 1Kb or closer to any of the genes in the UCSC Known Gene database and assign them to the nearest gene(s). B: for every peak in each of the ChIP datasets we selected those overlapping any of the DHSs indicated in A and assigned the peak to the gene represented by the DHSs.

Enrichment analysis

TFEA.ChIP is designed to take the output of gene expression profiling 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. TFEA.ChIP implements to types of tests to identify enriched TF. The first one analyzes 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 transcription factor determined by a Fisher's exact test. This analysis only requires a list of DE genes as input. In the second method, the association of TF to DE genes is determined using a Gene Set Enrichment Analysis (GSEA) [13]. This analysis requires a sorted list of genes where gene order reflects the relative expression in the two conditions being compared.

Software features

Include brief de tion of functio software See RO as exam

Results

The benefits of the use of ChIP-seq data over PWM for the identification of TFBS have been recently reported [14]. The implementation of this approach in the package TFEA.Chip described herein greatly simplifies its application to any general case. Here, we used four case studies to demonstrate the performance of the strategy implemented in this package when applied to different study settings. These cases include the original dataset where a primitive version of this strategy was first tested (transcriptional response of HUVEC cells to hypoxia), two additional datasets where the transcriptional response to the cytokines TNFalpha and IFNalpha was analyzed in different cell lines and a final dataset that, unlike the previous ones, recorded the transcriptional response to a complex pathological situation in vivo rather than to a defined stimulus. In all the cases we compared the output of TFEA.ChIP with the results of oPOSSUM, a PWM-based state-of-the-art tool [5]. To compare both methods we took the raw output from oPOSSUM and generated contingency matrices with the number of target hits, target non-hits, background hits, and background non-hits for every PWM, and then performed Fisher's exact test. The resulting p-values were adjusted for multiple testing using FDR method.

0.1 Hypoxia vs normoxia in HUVEC cells

The strategy behid TFEA.ChIP was first tested in an study aiming to determine transcription factor involved in gene repression induced by hypoxia [14]. In this work, the authors determined the effect of hypoxia on the transcription rate of all mRNAs expressed in human umbilical vein endothelial cells by means of pulse-labeling with 4-thiouridine followed by RNA-sequencing of labeled transcripts. Here, we reanalyzed this dataset (GSE89831) using TFEA.ChIP and compared the results to those produced by Opossum. For these analyses we identified DE genes using DEseq2 [15] and selected genes whose transcription was significantly induced by hypoxia (log-fold change hypoxia over normoxia ¿0 and FDR¡0.05). Then, we searched for overrepresented TFBS in this list of genes using TFEA.ChIP (Fig 3 A) and Opossum (Fig 3 B).

0.2 TNF addition in neutrophils and adipocytes

0.3 Left ventricular non-compaction in cardiomyocytes

0.4 INF addition in hESC cells for 15 and 21 days

Discussion

LIMITATION OF THE CURRENT VERSION: Although the package is mainly focused towards analyzing expression data generated from human cells, TFEA.ChIP includes the option to use datasets derived from experiments in mice, translating mouse gene names to their equivalent ID on the human genome. INDICATE THAT WE ARE ASSUMING THAT TF regulate same genes in different species DISCUSS LIMITATION OF 1Kb for association and the possibility of custom-generated DB DISCUSS THAT IT CAN BE APPLIED NOT ONLY TO GENE PROFILING BUT TO OTHER SUCH AS CHANGES IN EPIGENETIC MARKS (me.g.methyloma)

Conclusion

CO_2 Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit.

Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat. Ut neque ipsum, luctus id lacus ut, laoreet scelerisque urna. Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl.

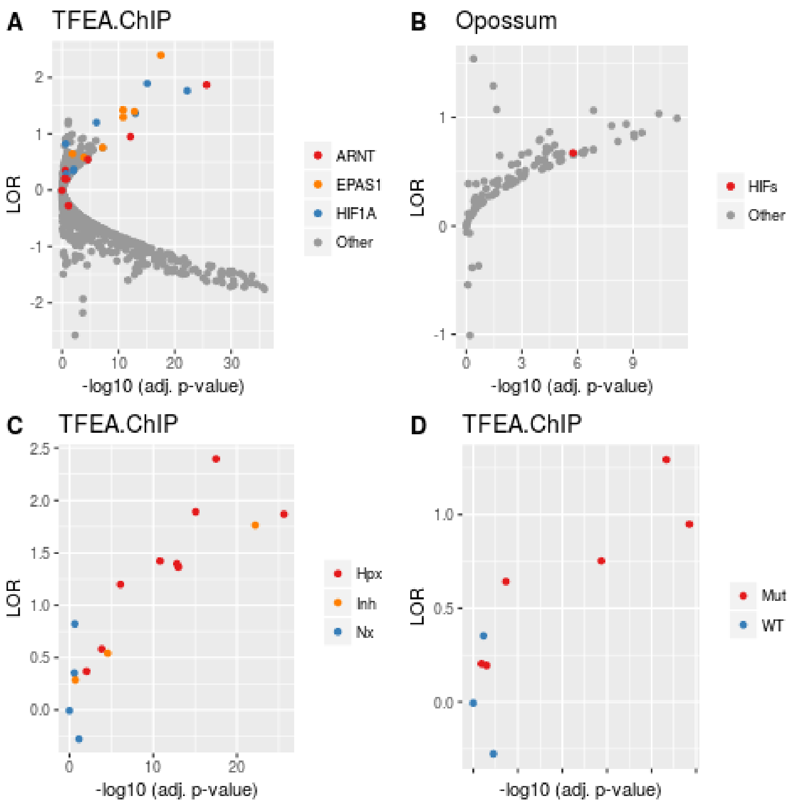


Fig 3. Identification of the TF responsible for transcriptional upregulation induced by hypoxia. A set of genes whose transcription was significantly induced in response to hypoxia was analyzed with TFEA.ChIP (A,C and D) or Opossum (B). The graphs represent the adjusted p-value (-log10 FDR) and the log-odds ratio (LOR) for the association of ChIP datasets (A,C and D) or PWM-motifs (B). A: datasets corresponding to TF of the Hypoxia Inducible Factor (”HIF1A”, ”EPAS1” and ”ARNT”) family and other TF (”other”); B: PWM-motif corresponding to Hypoxia Inducible Factors (”HIFs”) vs rest of motifs (”Other”). C: subset of datasets shown in panel A comparing normoxic (”Nx”), hypoxic (”Hpx”) or inhibitor-treated (”Inh”) samples. D: subset of datasets shown in panel A comparing samples from VHL-competent (”WT”) or deficient (”Mut”) cells.

Ut ornare mauris tellus, vel dapibus arcu suscipit sed. Nam condimentum sem eget mollis euismod. Nullam dui urna, gravida venenatis dui et, tincidunt sodales ex. Nunc est dui, sodales sed mauris nec, auctor sagittis leo. Aliquam tincidunt, ex in facilisis elementum, libero lectus luctus est, non vulputate nisl augue at dolor. For more information, see S1 Appendix.

Supporting information

S1 Fig. Bold the title sentence. Add descriptive text after the title of the item (optional).

S2 Fig. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 File. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Video. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Appendix. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Table. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

Acknowledgments

Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae.

References

1. Wasserman W, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nature Reviews Genetics. 2004;5:276. doi:10.1038/nrg1315.

2. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2004;489:57–74. doi:10.1038/nature11247.

3. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research. 2002;30:207–210.

4. Barrett T, et al. NCBI GEO: archive for functional genomics data sets - update. Nucleic Acids Research. 2013;41(D1):D991–D995. doi:10.1093/nar/gks1193.

5. Kwon AT, Arenillas DJ, Worsley Hunt R, Wasserman WW. oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. G3 (Bethesda, Md). 2012;2(9):987–1002. doi:10.1534/g3.112.003202.

6. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nature Reviews Genetics. 2009;10:252–263. doi:10.1038/nrg2538.

7. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018;172(4):650–665. doi:10.1016/j.cell.2018.01.029.

8. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489(7414):109–13. doi:10.1038/nature11279.

9. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nature Genetics. 2015;47(6):598–606. doi:10.1038/ng.3286.

10. John S, et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nature Genetics. 2011;43:264–268. doi:10.1038/ng.759.

11. Thurman RE, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82. doi:10.1038/nature11232.

12. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC Known Genes. Bioinformatics. 2006;22(9):1036–1046. doi:10.1093/bioinformatics/btl048.

13. Subramanian A, Tamayo P, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS. 2005;102(43):15545–15550. doi:10.1073/pnas.0506580102.

14. Tiana M, Acosta-Iborra B, Puente-Santamaría L, Hernansanz-Agustin P, Worsley-Hunt R, Masson N, et al. The SIN3A histone deacetylase complex is required for a complete transcriptional response to hypoxia. Nucleic acids research. 2018;46(1):120–133. doi:10.1093/nar/gkx951.

15. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550. doi:10.1186/s13059-014-0550-8.