






RESEARCH ARTICLE

CiiiDER: A tool for predicting and analysing transcription factor binding sites

Linden J. Gearing^{1,2}, Helen E. Cumming^{1,2}, Ross Chapman^{1,2}, Alexander M. Finkel^{1,2}, Isaac B. Woodhouse^{1,2}, Kevin Luu^{1,2}, Jodee A. Gould^{1,2}, Samuel C. Forster^{1,2}, Paul J. Hertzog^{1,2}

1 Centre for Innate Immunity and Infectious Diseases, Hudson Institute of Medical Research, Clayton, Victoria, Australia, **2** Department of Molecular Translational Science, Monash University, Clayton, Victoria, Australia

 These authors contributed equally to this work.

 These authors also contributed equally to this work.

* paul.hertzog@hudson.org.au.



 OPEN ACCESS

Citation: Gearing LJ, Cumming HE, Chapman R, Finkel AM, Woodhouse IB, Luu K, et al. (2019) CiiiDER: A tool for predicting and analysing transcription factor binding sites. *PLoS ONE* 14(9): e0215495. <https://doi.org/10.1371/journal.pone.0215495>

Editor: Manuela Helmer-Citterich, Universita degli Studi di Roma Tor Vergata, ITALY

Received: March 31, 2019

Accepted: August 5, 2019

Published: September 4, 2019

Copyright: © 2019 Gearing et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files or at www.ciiider.org.

Funding: The authors would like to acknowledge the support and contributions of the Bio21 Undergraduate Research Opportunity Program (A.M.F. and S.C.F.) and the Victorian Life Sciences Computation Initiative (S.C.F.) for support in this project. Research at Hudson Institute of Medical Research is supported by the Victorian Government's Operational Infrastructure Support

Abstract

The availability of large amounts of high-throughput genomic, transcriptomic and epigenomic data has provided opportunity to understand regulation of the cellular transcriptome with an unprecedented level of detail. As a result, research has advanced from identifying gene expression patterns associated with particular conditions to elucidating signalling pathways that regulate expression. There are over 1,000 transcription factors (TFs) in vertebrates that play a role in this regulation. Determining which of these are likely to be controlling a set of genes can be assisted by computational prediction, utilising experimentally verified binding site motifs. Here we present CiiiDER, an integrated computational toolkit for transcription factor binding analysis, written in the Java programming language, to make it independent of computer operating system. It is operated through an intuitive graphical user interface with interactive, high-quality visual outputs, making it accessible to all researchers. CiiiDER predicts transcription factor binding sites (TFBSs) across regulatory regions of interest, such as promoters and enhancers derived from any species. It can perform an enrichment analysis to identify TFs that are significantly over- or under-represented in comparison to a bespoke background set and thereby elucidate pathways regulating sets of genes of pathophysiological importance.

Introduction

Contemporary transcriptomic technologies such as microarrays and RNA-sequencing provide reliable methods to identify genes differentially expressed across cell types, tissues or in response to different stimuli. These methods reveal many co-expressed genes or gene networks that are together predicted to determine the observed biological responses.

Transcription factors (TFs) bind to specific DNA sequences (transcription factor binding sites; TFBSs) within promoter and enhancer regions of genomic DNA and either activate or repress gene expression. These interactions can be determined experimentally, for example

Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

using chromatin immunoprecipitation (ChIP) techniques, and are typically represented as position frequency matrices (PFMs) [1].

Other more complicated models of TFBSs have been developed, for example based on hidden Markov models [2, 3] or machine learning [4, 5]. However, PFMs are still the most widely used models of transcription factor binding: not only is it easy to construct PFMs from a set of validated sequences, but there are also several curated databases of PFMs, applicable to a wide range of species, including the commercial TRANSFAC database [6] and the open-access JASPAR database [7, 8]. While using PFMs alone can predict TFBSs within regulatory sequences, in eukaryotic organisms, this typically results in high false positive prediction rates, since predicted sites may not be accessible to the transcriptional machinery due to chromatin structure or the epigenetic landscape.

An enrichment analysis, which compares the distribution of TFBSs predicted in a set of regulatory DNA regions to the distribution in a set of background sequences, can be utilised to more accurately identify true TFBSs. With the appropriate choice of background, it is possible to identify TFBSs that are statistically over- or under-represented. The TFs with over-represented TFBSs in a set of co-expressed genes are more likely to be involved in regulating the expression of these genes [9].

While there are existing and publicly available programs that can perform enrichment analyses, these tools are typically web-based (e.g. MotifViz [10]), restricted to command-line use (e.g. Clover [11] and HOMER [12]) or offer web-based and command-line versions (e.g. Pscan [13], oPOSSUM [14] and the MEME suite [15]). Online tools can be convenient, but it is beneficial to have the security of a downloadable program, which can be run on a local computer and to have saved projects that can easily be revisited. Downloadable applications for TFBS prediction that are run through the command line often require additional tools to be installed or lack effective visualisations, providing static or text-based results, which limits their utility for a wide audience.

There is a need for a downloadable program, independent of computer operating system, that provides the required flexibility to perform accurate, integrated, customisable analysis and data exploration. To address this, we developed CiiiDER, a user-friendly analysis toolkit for predicting and analysing putative TFBSs within regulatory regions. CiiiDER is operated through an intuitive graphical user interface (GUI), which is designed with ease of use in mind. The established MATCH algorithm [16] has been implemented to map potential TFBSs within a set of regulatory sequences. TF enrichment analysis can be used to enable identification of key TFs that are statistically enriched and are therefore more likely to be biologically relevant.

Methods

Workflow and algorithms

The CiiiDER workflow (Fig 1) can accept query and background regions in a variety of input formats. It consists of two main analyses, with results presented in an interactive format.

Inputs. CiiiDER has been designed to ensure that data input is easy and that a wide variety of formats are accepted. All data input and parameter selection is facilitated through simple interfaces, with pop-up information boxes available to help users make the appropriate selections. Data can be loaded from file or by pasting the sequence information directly into a box in the GUI.

CiiiDER will read sequences directly from data entered in FASTA format. Genomic location formats (GFF, GTF or BED) require an associated genome file for the relevant species to

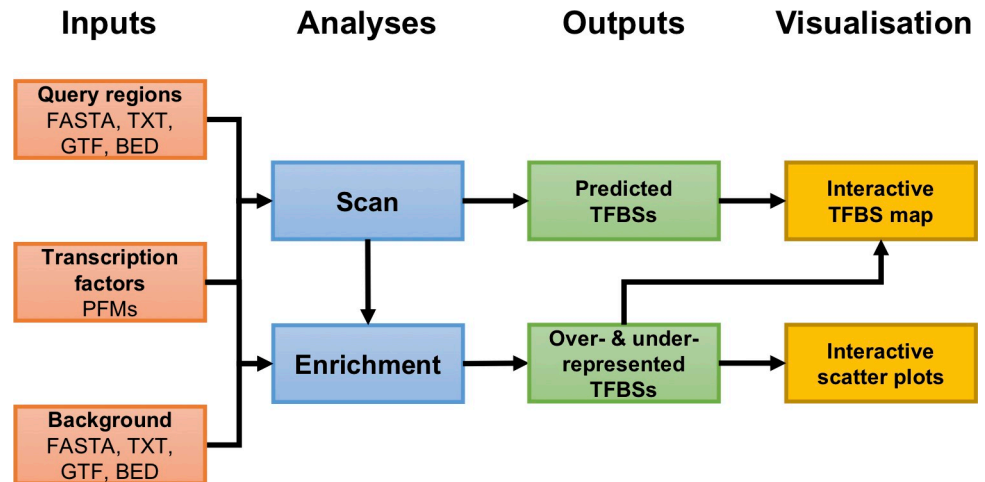


Fig 1. CiiiDER workflow. A typical analysis involves submitting gene sets for scanning against known TF models, followed by the identification of sites that are statistically enriched relative to a submitted background gene list.

<https://doi.org/10.1371/journal.pone.0215495.g001>

obtain DNA sequences. These formats can be used to analyse any regulatory region of interest including promoters, enhancers and untranslated regions (UTRs).

As promoters tend to be the regulatory regions of interest for the majority of users, CiiiDER will automatically extract promoter sequences from a genome file, given a list of gene symbols, Ensembl gene IDs or Ensembl transcript IDs. This requires an additional annotation file, denoted the gene look-up manager (GLM), which contains the location of the transcription start sites (TSSs) for each gene and transcript. The user can specify the size of the sequence relative to the TSS, with the default set at -1500 bases upstream to $+500$ bases downstream of the TSS. CiiiDER can be downloaded with human GRCh38 and mouse GRCm38 genomes and GLM files. For alternative genomes, the user will need to provide the appropriate genome and an Ensembl GTF file, from which CiiiDER can automatically generate the necessary GLM file.

Scan. During the Scan stage, CiiiDER uses an implementation of the MATCH algorithm [16] to predict potential TFBSs in regions of interest. This approach is compatible with PFMs in JASPAR [17] or TRANSFAC [6] format. The mapping of each TFBS is performed with a user-specified deficit that determines the stringency of the scan.

In brief, since PFMs often have a highly conserved core binding region, which is flanked by areas of higher variability, a core PFM is created for the five most conserved consecutive bases (which is calculated using the sum of information vector values, as described in [16]).

To search for TFBSs, sequences are split into overlapping five-base regions, which are compared with the core PFM. If the similarity score between a five-base sequence and the core PFM meets a defined threshold, then the sequence window is increased to the full length of the TFBS and the similarity score with the full PFM is calculated.

CiiiDER uses a deficit value, which is the difference between the MATCH score of a TFBS and the maximum possible score, which is 1. The default deficit is 0.15, which means the scan will accept any TFBSs that have MATCH scores of 0.85 or above (the same cut-off is applied to both the core and matrix scores from the MATCH algorithm).

Enrichment. The Enrichment stage identifies those TFBSs that are significantly over- or under-represented in query regions compared to relevant, user-specified background regions. This analysis scans the background sequences for TFBSs using the same criteria as used for the query sequences. Over- and under-represented TFs are determined by comparing the numbers of sequences with predicted TFBSs to the number of those without, using a Fisher's exact test;

alternatively, the distributions of the number of sites per sequence in the query and background sets can be compared using a Mann-Whitney U test.

For the enrichment plots, if a given transcription factor has binding sites in n_S out of N_S search regions and n_B out of N_B background regions, then:

$$\text{Average.Log2.Proportion.Bound} = \frac{1}{2} \log_2 \left(\frac{n_S + 1/2}{N_S + 1/2} \right) + \frac{1}{2} \log_2 \left(\frac{n_B + 1/2}{N_B + 1/2} \right)$$

$$\text{Log2.Enrichment} = \log_2 \left(\frac{n_S + 1/2}{N_S + 1/2} \right) - \log_2 \left(\frac{n_B + 1/2}{N_B + 1/2} \right)$$

$$\text{Significance.Score} = -\text{sign}(\text{Log2.Enrichment}) \times \log_{10}(P\text{-value})$$

Outputs and visualisation. CiiDER produces clear graphical displays to help interpret the results of TFBS prediction. Putative TFBSs are displayed in an interactive map (Fig 2). By default, ten TFs are shown, but the user can choose to add or remove TFs from the image. There are also options to filter the TFs displayed according to the scan stringency or enrichment P -value, for intuitive exploration of the data.

Promoters or other regulatory regions can be re-arranged or removed and the colour of each TF can be customised for the production of figures. The enrichment analysis produces an additional interactive plot that displays the fold enrichment, average abundance and P -value associated with all TFs (Fig 3). The images created can be saved as publication-quality files and the binding site data and enrichment statistics can be saved as text files for subsequent analysis using additional tools.

The entire project can be saved for further analysis and interpretation.

Implementation. CiiDER has been implemented in Java for platform independence with the GUI utilising the Swing libraries to deploy a simple, intuitive interface. The multi-threading capabilities of Java are used to take advantage of all available computer processors, significantly improving analysis speeds (S1 Fig); alternatively, CiiDER can also be restricted to use only a certain number of processors. Enrichment of transcription factors is also displayed using interactive HTML plots generated using the Plotly JavaScript library [18].

CiiDER is available for download as a JAR file with supporting files; other software dependencies do not need to be installed. Two PFM libraries are supplied with the software: the JASPAR 2018 CORE non-redundant vertebrate matrices [7] and matrices from Jolma *et al.*, a large experimental dataset [19]. Genomes and associated GLM files are also available for extracting promoter sequences using gene names or Ensembl IDs.

CiiDER is distributed under the GNU GPLv3 licence. The program and documentation are available from www.ciiider.org and the source code is available at <https://gitlab.erc.monash.edu.au/ciiid/ciiider>.

Experimental methods

CiiDER analyses were performed using the Ensembl 89 or Ensembl 94 annotations of the human GRCh38 and mouse GRCm38 genomes, respectively, with the 2011 version of the TRANSFAC non-redundant vertebrate database [6] or the 2018 JASPAR core non-redundant vertebrate matrices [7]. All promoter regions were defined as spanning -1500 bases to $+500$ bases relative to the transcription start site.

ChIP-seq scan. Experimentally verified transcription factor binding site regions were obtained from publicly available ChIP-seq experiments: a CTCF dataset (ENCSR000DLG)

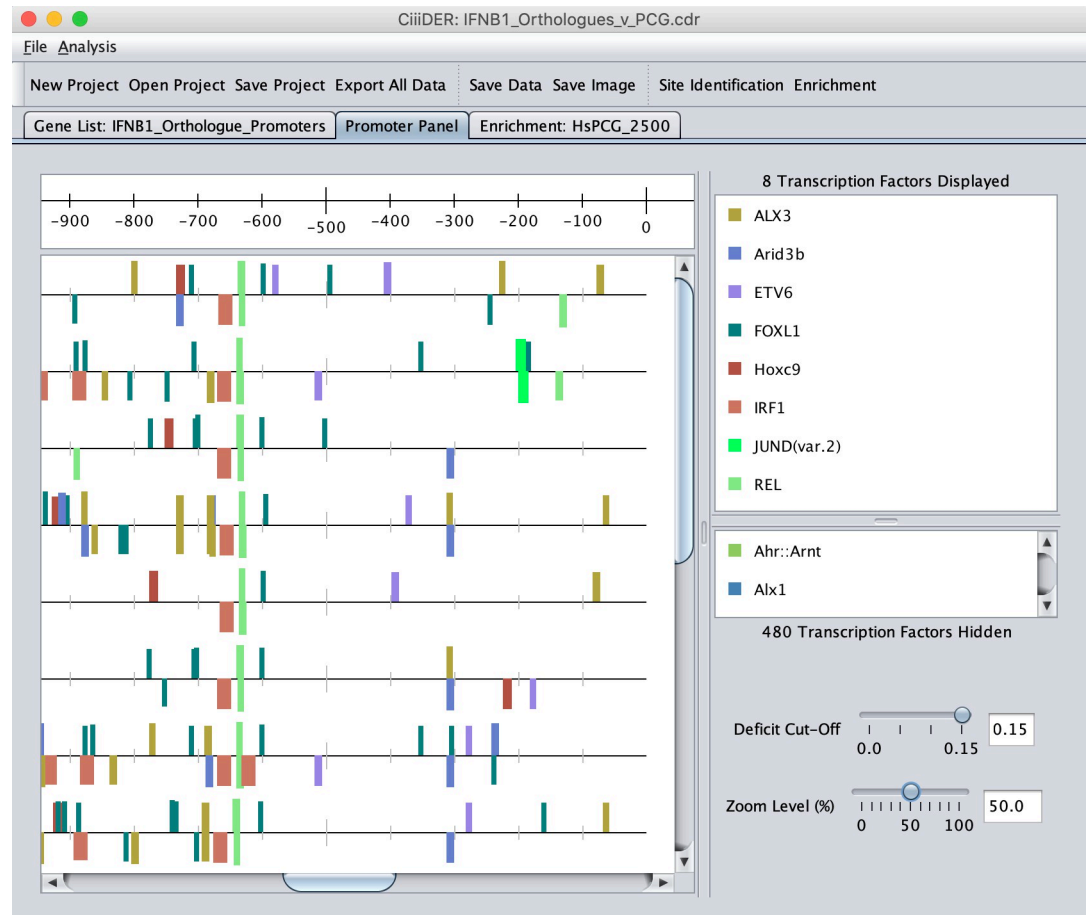


Fig 2. CiiiDER interactive site map. The scan and enrichment algorithms produce a graphical display of the TFBS locations on the sequences. There are many options to edit the images, including adjusting the deficit and *P*-value thresholds for displaying TFBSs, selecting or removing TFs to be viewed, editing the colour scheme for TFs and rearranging the order of the sequences.

<https://doi.org/10.1371/journal.pone.0215495.g002>

from ENCODE [20] and STAT3 dataset (GSM288353 [21]) from GEO [22]. These were selected because they were available in narrow peak format with peak max values, to give the highest probability of focusing on the true binding site, and because matching high-quality TRANSFAC TF models were available.

Sequences corresponding to 50 bases either side of the maximum signal of each ChIP-seq peak were obtained. This length was chosen to allow sufficient sequence to identify TFBSs, while minimising extraneous sequence. Backgrounds were produced by using 101 base genomic sequences 10,000 bases away from the peak, ensuring that none of the background sequences overlapped with surrounding peaks. CiiiDER scans were performed using deficits of 0.2. TRANSFAC scans were performed with equivalent core and matrix similarity score cut-offs of 0.8. Clover analyses were performed with default values. Prism 7 was used to generate ROC curves and associated area under curve (AUC) values for each program.

Microarray enrichment analysis. Robust multiarray averaging (RMA)-normalised microarray data from Bidwell *et al.* (GSE37975) [23] were downloaded using GEOquery [24] and processed through the limma package [25] to identify differentially expressed transcripts between the primary and metastatic tumour samples, with \log_2 fold change > 1 or < -1 and a *P*-value < 0.05 (adjusted for false discovery rate using the Benjamini-Hochberg correction).

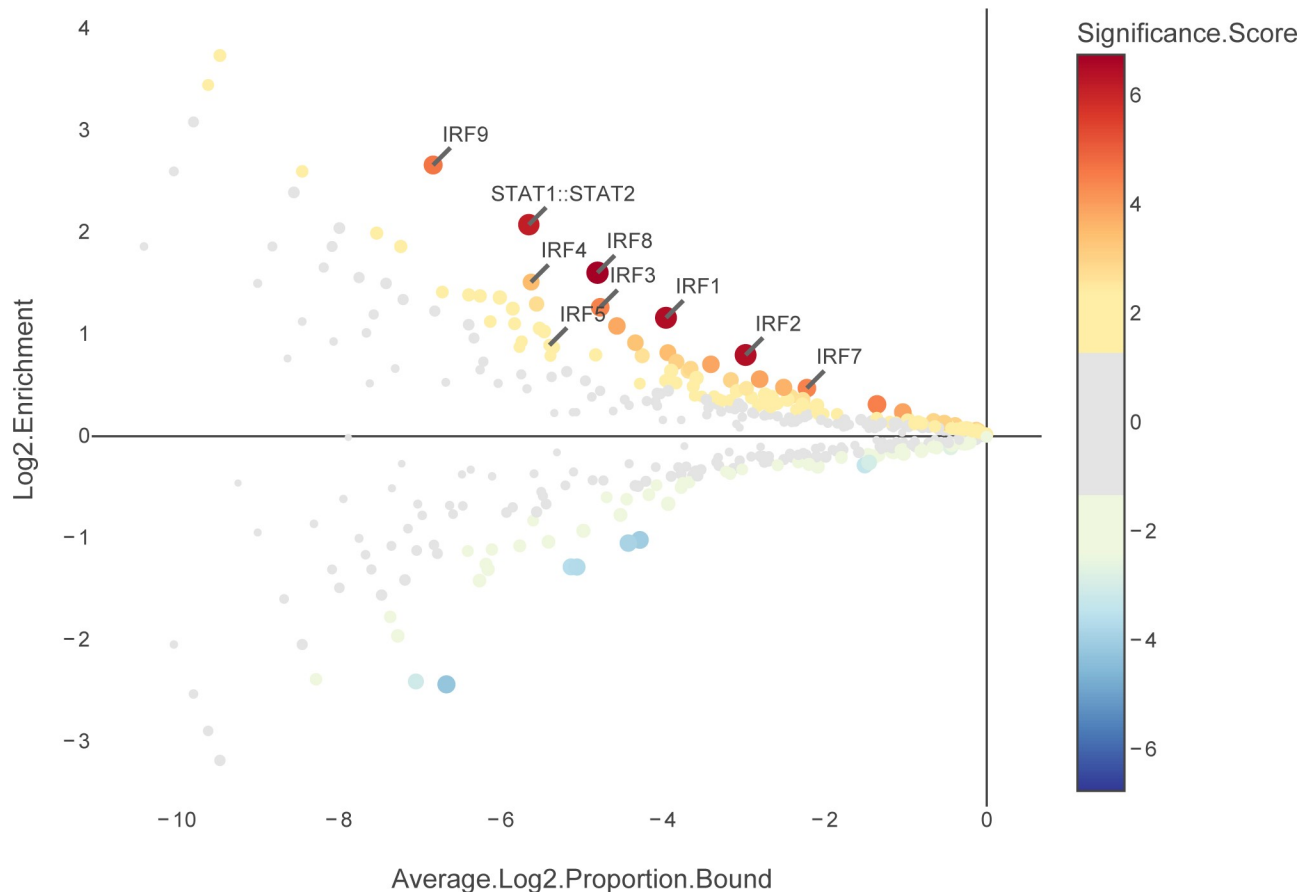


Fig 3. CiiiDER enrichment results for the breast cancer metastasis dataset. The data are derived from the proportion of regions bound for each TF, which is the number of bound regions divided by the total number of regions. The plot shows the enrichment (ratio of proportion bound) and average log proportion bound. Size and colour show $\mp \log_{10}(P\text{-value})$ (significance score); it is greater than zero if the TF is over-represented and less than zero if under-represented. Underlying data are provided in [S1 Data](#).

<https://doi.org/10.1371/journal.pone.0215495.g003>

The query gene list consisted of significantly down-regulated genes that were defined as interferon-inducible in mouse using the Interferome v2.0 [26] (up-regulated at least two-fold by interferon treatment). The background gene list was derived from transcripts with an absolute \log_2 fold change < 0.05 and average expression greater than the first quartile. CiiiDER analyses were performed on the promoter sequences using deficits of 0.15 with JASPAR TF models.

Phylogenetic scan. The IFN β promoter sequence was obtained from 16 species of placental mammals, using Ensembl 94 annotations. A background containing 2,500 human protein-coding genes was used for enrichment. CiiiDER analysis was performed using JASPAR TF motifs with deficits of 0.15.

Results and discussion

Comparing CiiiDER to other software with ChIP-seq data

In order to demonstrate the utility of CiiiDER and that it correctly implements the MATCH algorithm, we compared CiiiDER against other downloadable TF site prediction software, TRANSFAC (which also uses the MATCH algorithm), as well as Clover. As examples, we obtained ChIP-seq data for CTCF and STAT3 and compared the performances of the

programs to detect true positive TFBSs using ROC curves (S2 Fig). The Area Under Curve (AUC) was calculated for each program and CiiiDER showed the same ability to predict true positive sites as TRANSFAC, demonstrating that the CiiiDER implementation of the MATCH TFBS prediction algorithm functions as expected.

Applications of CiiiDER

Identification of enriched TFs in co-expressed gene sets. Transcriptomic experiments such as microarray and RNA-sequencing are an excellent source of co-regulated genes for CiiiDER analyses. In order to perform an enrichment analysis for a gene set of interest, it is important to choose an appropriate background [27]. Comparing an experimentally derived, co-expressed gene list to a genome-wide background may lead to the enrichment of some TFs that are not specifically related to the experiment. For example, if the query were promoters of genes showing a significant change in expression following a particular treatment, then an appropriate background might be promoters of genes that were expressed in the appropriate cell or tissue type, but showed no significant response to the stimulation [9].

The ability for CiiiDER to predict key regulatory TFs was demonstrated by reanalysing a published study of the regulation of the immune system in breast cancer [23]. Bidwell *et al.* compared the gene expression in primary and metastatic tumour cells in a mouse model of spontaneous bone metastasis. A set of approximately 3,000 genes were down-regulated in the metastasised cells relative to the primary tumour, 540 of which were determined to be interferon-regulated genes (IRGs) from the Interferome v1.0 database [28]. In that study, Clover was used to show that these genes were enriched for Irf7 binding sites. The role of Irf7 was confirmed by showing an increase in interferon signalling and a reduction of tumour metastases following restoration of Irf7 expression in the tumour cells in the bone metastasis model.

We reanalysed the normalised expression data from this experiment to create a list of IRGs down-regulated in metastases using the updated Interferome v2.0 [26]. A CiiiDER enrichment analysis was performed on the promoters of these IRGs using a background gene set of expressed, unchanged genes (Fig 3). This showed that CiiiDER was able to identify Irf7 as a key TF, potentially regulating the expression of immune system genes within the breast cancer tumour ($P = 3.46E-05$), in agreement with the published prediction and experimental validation.

In this example, many other IRF-family TFs were also significantly over-represented (e.g. IRF8, $P = 1.76E-07$). It is often difficult to accurately distinguish between TFBSs of TFs belonging to a family, since their binding site preferences can be very similar. In this case, cross-referencing with the published expression data revealed that Irf7 was the most significantly suppressed IRF-family TF in metastases, which added supporting evidence to its role.

Identifying phylogenetically conserved TFBSs. Since gene orthologues often retain similar functions throughout evolution and maintain a similar method of regulation [29], CiiiDER could potentially be used to examine phylogenetic conservation, through prediction of enriched TFs, and by creating visualisations to help distinguish patterns in TFBSs. Prediction of phylogenetically conserved TFBSs has previously been shown to be able to identify functionally important TFBSs [30, 31].

To test the capacity of CiiiDER to identify evolutionary conserved regulatory elements, we selected the interferon β (IFN β) promoter, the transcriptional regulation of which has been very well characterised [32]: in brief, IRF-family TFs, NF- κ B and AP-1 (which is comprised of ATF2 and JUN) together allow remodelling of the local chromatin structure to promote gene transcription.

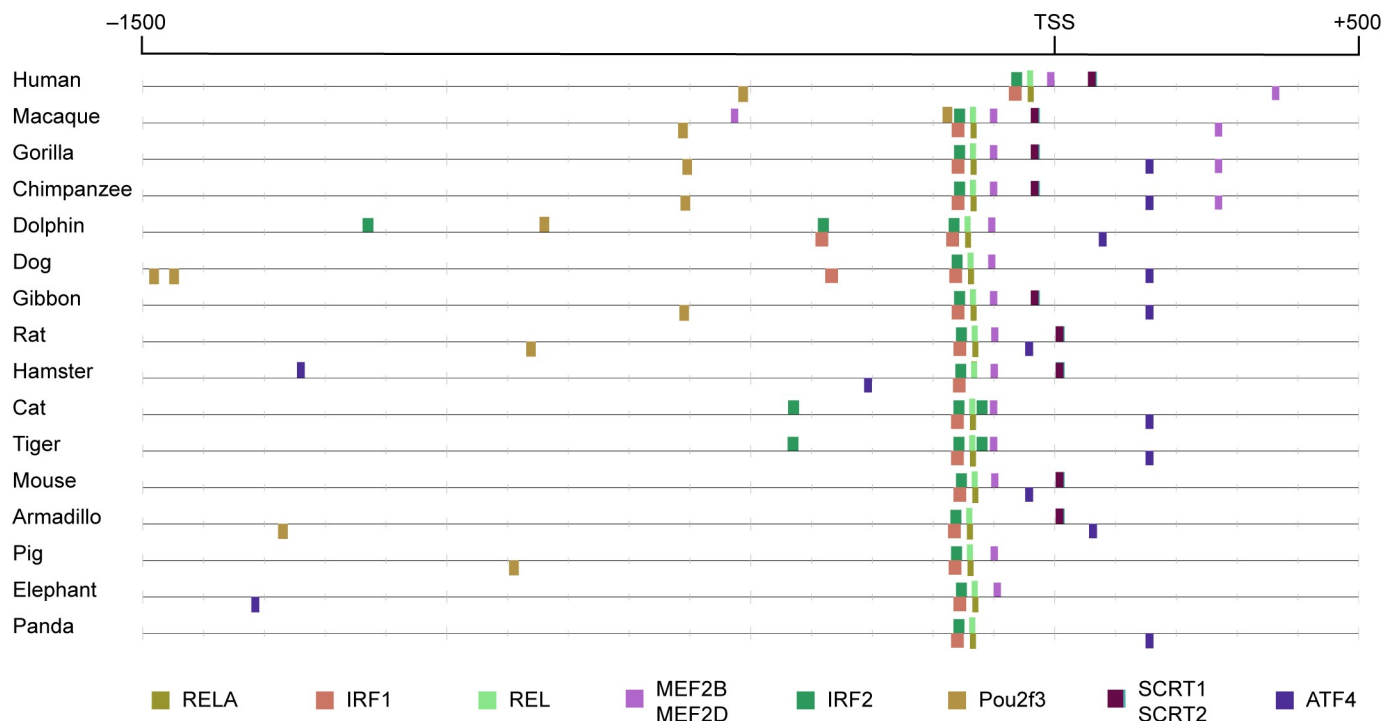


Fig 4. Phylogenetic conservation of TF binding sites in the IFN β promoter. The results of the enrichment algorithm, displaying the ten most significantly enriched TFs present in at least half of the promoters. Underlying data are provided in [S3 Data](#).

<https://doi.org/10.1371/journal.pone.0215495.g004>

Initially, the scan method was used to identify TFBSs in the IFN β promoters from placental mammal species detailed in [Fig 4](#). This identified a great number of potential TFBSs for hundreds of TFs (see [Fig 2](#)), many of which are likely to be false positives, which makes it difficult to identify likely candidate transcriptional regulators. An enrichment analysis was then performed to compare the TFBSs in the IFN β promoters against a background of human protein coding genes. The top ten over-represented TFs that occurred in at least half of the promoters were selected for display ([Fig 4](#)). Spatially conserved TFBSs are immediately apparent, particularly those in a cluster within 200 bases of the TSSs. This includes TFBSs for the best characterised IFN β regulators—NF- κ B components RELA and REL ($P = 2.38E-19$ and $1.10E-10$) and IRF-family TFs IRF1 and IRF2 ($P = 9.54E-14$ and $4.46E-10$). These are the most significant TFs that are predicted in all promoters, whereas other top significant TFs do not show the same consistent pattern.

The combination of enrichment analysis and effective visualisation can allow rapid identification of TFBSs that are phylogenetically and spatially conserved. This gives greater support when choosing candidate TFs that are most likely to be involved in regulatory elements.

Further analyses

CiiiDER can also be used to search for TFBSs associated with regions of the genome marked with epigenetic modifications obtained from CHIP-seq data or open chromatin regions derived from ATAC-seq data. For example, we have published using CiiiDER to examine transcriptional enhancers (marked by histone H3 lysine 4 mono-, di- and tri-methylation) in effector and memory T-cells, compared to those common between effector, memory and naïve T-cells, to show an enrichment of BATF, JUN and FOS motifs, among others [33].

The power of CiiiDER analyses can be increased by linking the results to other data. As with the breast cancer example, it is worth considering all members of a TF family when choosing TFs for further validation. TFBS enrichment results may be assessed in the context of gene expression data to determine which TFs are detectable or have altered expression levels in the experimental system of interest.

Conclusion

CiiiDER is an intuitive new tool for analysing TFBSs in regulatory regions of interest. It can efficiently scan sequences for potential TFBSs and identify TFBSs that are statistically under- or over- represented. It is user-friendly and produces quality visual outputs to assist researchers to uncover signalling pathways and their controlling TFs in a wide variety of biological contexts. The program, user manual and example data are available at www.ciiider.org.

Supporting information

S1 Fig. CiiiDER analysis time. Example plot of analysis time and CPU usage of CiiiDER when performing site identification and enrichment using the *Irf7* breast cancer gene set. Gene sets were loaded into the GUI and promoters were obtained (A), TFBSs were predicted across the query promoters (B) and collated (C), background sites were predicted (D), the enrichment calculation was performed (E) and the final graphical outputs were created (F). The site prediction steps take advantage of multiple computer processors. The maximum memory usage was 4.53 GB. Measurements were made on an iMac with four i7 4.0 GHz processors and 32 GB RAM. Underlying data are provided in [S1 Data](#).
(TIFF)

S2 Fig. CiiiDER scan algorithm. The accuracy of CiiiDER was compared with Clover and TRANSFAC software using ROC curves for (A) CTCF and (B) STAT3. The curves represent the ratio of true binding sites predicted against the number of false binding sites predicted. The locations of true binding sites have been validated previously using ChIP-seq experiments. Note that, due to almost complete overlap with the TRANSFAC curves, the CiiiDER curves for both CTCF and STAT3 were shifted down by -0.01. Underlying data are provided in [S2 Data](#).
(TIFF)

S1 Data. Breast cancer enrichment data. IRGs down-regulated in metastases, background genes, CiiiDER enrichment results and analysis times are provided.
(XLSX)

S2 Data. ChIP-seq scan data. Data used to generate ROC curves are provided. For each program tested, scores are given for predicted sites in ChIP-seq peaks and background regions. If more than one site was predicted in a peak or background region, the maximum score is given.
(XLSX)

S3 Data. IFN β promoter enrichment data. IFN β promoter sequences, human background genes and CiiiDER enrichment results are provided.
(XLSX)

Acknowledgments

The authors would like to thank Gordon Smyth and Jonathan Keith for statistical advice and Michael Gantier, Mirana Ramialison, Christine Wells and Jesse Balic for evaluating the

software. The authors would like to acknowledge the support and contributions of the Bio21 Undergraduate Research Opportunity Program, the Victorian Life Sciences Computation Initiative, Monash e-Research and Philip Chan for support in this project. Research at Hudson Institute of Medical Research is supported by the Victorian Government's Operational Infrastructure Support Program. CiiiDER is available from www.ciiider.org.

Author Contributions

Conceptualization: Samuel C. Forster, Paul J. Hertzog.

Data curation: Linden J. Gearing, Helen E. Cumming, Isaac B. Woodhouse, Jodee A. Gould.

Formal analysis: Linden J. Gearing, Helen E. Cumming, Jodee A. Gould, Samuel C. Forster, Paul J. Hertzog.

Funding acquisition: Paul J. Hertzog.

Investigation: Linden J. Gearing, Helen E. Cumming, Alexander M. Finkel, Kevin Luu, Samuel C. Forster.

Methodology: Linden J. Gearing, Helen E. Cumming, Ross Chapman, Alexander M. Finkel, Isaac B. Woodhouse, Kevin Luu, Jodee A. Gould, Samuel C. Forster.

Project administration: Samuel C. Forster, Paul J. Hertzog.

Resources: Helen E. Cumming, Paul J. Hertzog.

Software: Linden J. Gearing, Helen E. Cumming, Ross Chapman, Alexander M. Finkel, Isaac B. Woodhouse, Kevin Luu, Samuel C. Forster.

Supervision: Samuel C. Forster, Paul J. Hertzog.

Validation: Linden J. Gearing, Ross Chapman, Alexander M. Finkel, Isaac B. Woodhouse.

Visualization: Linden J. Gearing, Helen E. Cumming, Ross Chapman, Alexander M. Finkel, Isaac B. Woodhouse, Kevin Luu.

Writing – original draft: Linden J. Gearing, Helen E. Cumming, Ross Chapman, Samuel C. Forster, Paul J. Hertzog.

Writing – review & editing: Linden J. Gearing, Helen E. Cumming, Samuel C. Forster, Paul J. Hertzog.

References

1. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res.* 1982; 10(9):2997–3011. <https://doi.org/10.1093/nar/10.9.2997> PMID: 7048259; PubMed Central PMCID: PMC320670.
2. Marinescu VD, Kohane IS, Riva A. MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics.* 2005; 6:79. <https://doi.org/10.1186/1471-2105-6-79> PMID: 15799782; PubMed Central PMCID: PMC1131891.
3. Mathelier A, Wasserman WW. The next generation of transcription factor binding site prediction. *PLoS Comput Biol.* 2013; 9(9):e1003214. <https://doi.org/10.1371/journal.pcbi.1003214> PMID: 24039567; PubMed Central PMCID: PMC3764009.
4. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol.* 2015; 33(8):831–8. <https://doi.org/10.1038/nbt.3300> PMID: 26213851.
5. Khamis AM, Motwalli O, Oliva R, Jankovic BR, Medvedeva YA, Ashoor H, et al. A novel method for improved accuracy of transcription factor binding site prediction. *Nucleic Acids Res.* 2018; 46(12):e72. <https://doi.org/10.1093/nar/gky237> PMID: 29617876; PubMed Central PMCID: PMC6037060.

6. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 2006; 34(Database issue):D108–10. <https://doi.org/10.1093/nar/gkj143> PMID: 16381825; PubMed Central PMCID: PMC1347505.
7. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* 2018; 46(D1):D260–D6. <https://doi.org/10.1093/nar/gkx1126> PMID: 29140473; PubMed Central PMCID: PMC5753243.
8. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 2004; 32(Database issue):D91–4. <https://doi.org/10.1093/nar/gkh012> PMID: 14681366; PubMed Central PMCID: PMC308747.
9. Boeva V. Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. *Front Genet.* 2016; 7:24. <https://doi.org/10.3389/fgene.2016.00024> PMID: 26941778; PubMed Central PMCID: PMC4763482.
10. Fu Y, Frith MC, Haverty PM, Weng Z. MotifViz: an analysis and visualization tool for motif discovery. *Nucleic Acids Res.* 2004; 32(Web Server issue):W420–3. <https://doi.org/10.1093/nar/gkh426> PMID: 15215422; PubMed Central PMCID: PMC441564.
11. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z. Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.* 2004; 32(4):1372–81. <https://doi.org/10.1093/nar/gkh299> PMID: 14988425; PubMed Central PMCID: PMC390287.
12. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell.* 2010; 38(4):576–89. <https://doi.org/10.1016/j.molcel.2010.05.004> PMID: 20513432; PubMed Central PMCID: PMC2898526.
13. Zambelli F, Pesole G, Pavesi G. Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res.* 2009; 37(Web Server issue):W247–52. <https://doi.org/10.1093/nar/gkp464> PMID: 19487240; PubMed Central PMCID: PMC2703934.
14. 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).* 2012; 2(9):987–1002. <https://doi.org/10.1534/g3.112.003202> PMID: 22973536; PubMed Central PMCID: PMC3429929.
15. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. *Nucleic Acids Res.* 2015; 43(W1):W39–49. <https://doi.org/10.1093/nar/gkv416> PMID: 25953851; PubMed Central PMCID: PMC4489269.
16. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* 2003; 31(13):3576–9. <https://doi.org/10.1093/nar/gkg585> PMID: 12824369; PubMed Central PMCID: PMC169193.
17. Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 2016; 44(D1):D110–5. <https://doi.org/10.1093/nar/gkv1176> PMID: 26531826; PubMed Central PMCID: PMC4702842.
18. Collaborative data science [Internet]. Plotly Technologies Inc. 2015. Available from: <https://plot.ly/>.
19. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* 2010; 20(6):861–73. <https://doi.org/10.1101/gr.100552.109> PMID: 20378718; PubMed Central PMCID: PMC2877582.
20. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012; 489(7414):57–74. <https://doi.org/10.1038/nature11247> PMID: 22955616; PubMed Central PMCID: PMC3439153.
21. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell.* 2008; 133(6):1106–17. <https://doi.org/10.1016/j.cell.2008.04.043> PMID: 18555785.
22. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 2013; 41(Database issue):D991–5. <https://doi.org/10.1093/nar/gks1193> PMID: 23193258; PubMed Central PMCID: PMC3531084.
23. Bidwell BN, Slaney CY, Withana NP, Forster S, Cao Y, Loi S, et al. Silencing of Irf7 pathways in breast cancer cells promotes bone metastasis through immune escape. *Nat Med.* 2012; 18(8):1224–31. <https://doi.org/10.1038/nm.2830> PMID: 22820642.
24. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics.* 2007; 23(14):1846–7. <https://doi.org/10.1093/bioinformatics/btm254> PMID: 17496320.

25. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015; 43(7):e47. <https://doi.org/10.1093/nar/gkv007> PMID: 25605792; PubMed Central PMCID: PMC4402510.
26. Rusinova I, Forster S, Yu S, Kannan A, Masse M, Cumming H, et al. Interferome v2.0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Res.* 2013; 41(Database issue):D1040–6. <https://doi.org/10.1093/nar/gks1215> PMID: 23203888; PubMed Central PMCID: PMC3531205.
27. Timmons JA, Szkop KJ, Gallagher IJ. Multiple sources of bias confound functional enrichment analysis of global -omics data. *Genome Biology.* 2015; 16(1):1–3. <https://doi.org/10.1186/s13059-015-0761-7> PMID: 26346307
28. Samarajiwa SA, Forster S, Auchettl K, Hertzog PJ. INTERFEROME: the database of interferon regulated genes. *Nucleic Acids Res.* 2009; 37(Database issue):D852–7. <https://doi.org/10.1093/nar/gkn732> PMID: 18996892; PubMed Central PMCID: PMC2686605.
29. Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, et al. Surveying *Saccharomyces* Genomes to Identify Functional Elements by Comparative DNA Sequence Analysis. *Genome Res.* 2017; 11(1175–1186).
30. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, et al. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics.* 2005; 21(13):2933–42. <https://doi.org/10.1093/bioinformatics/bti473> PMID: 15860560.
31. Eisermann K, Tandon S, Bazarov A, Brett A, Fraizer G, Piontkivska H. Evolutionary conservation of zinc finger transcription factor binding sites in promoters of genes co-expressed with WT1 in prostate cancer. *BMC Genomics.* 2008; 9:337. <https://doi.org/10.1186/1471-2164-9-337> PMID: 18631392; PubMed Central PMCID: PMC2515153.
32. Honda K, Takaoka A, Taniguchi T. Type I interferon gene induction by the interferon regulatory factor family of transcription factors. *Immunity.* 2006; 25(3):349–60. Epub 2006/09/19. <https://doi.org/10.1016/j.immuni.2006.08.009> PMID: 16979567.
33. Russ BE, Olshansky M, Li J, Nguyen MLT, Gearing LJ, Nguyen THO, et al. Regulation of H3K4me3 at Transcriptional Enhancers Characterizes Acquisition of Virus-Specific CD8(+) T Cell-Lineage-Specific Function. *Cell Rep.* 2017; 21(12):3624–36. <https://doi.org/10.1016/j.celrep.2017.11.097> PMID: 29262339.