Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Modeling Transcriptional Regulation Using Gene Regulatory Networks Based on Multi-Omics Data Sources

Neel Patel
Department of Nutrition, Case Western Reserve University, OH
William Bush (wsb36@case.edu)
Department of Population and Quanitative Health Sciences, Cleveland, OH

Research Article

Keywords: GM12878, K562, demonstration, SKAT

Posted Date: December 1st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-112300/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at BMC Bioinformatics on April 19th, 2021.
See the published version at https://doi.org/10.1186/s12859-021-04126-3.

Abstract

Background

Transcriptional regulation is complex, requiring multiple *cis*(local) and *trans* acting mechanisms working in concert to drive gene expression, with disruption of these processes linked to multiple diseases. Previous computational attempts to understand the influence of regulatory mechanisms on gene expression have used prediction models containing input features derived from *cis* regulatory factors. However, local chromatin looping and *trans*-acting mechanisms are known to also influence transcriptional regulation, and their inclusion may improve model accuracy and interpretation.

Results

We describe a computational framework to model gene expression for GM12878 and K562 cell lines. This framework weights the impact of transcription factor-based regulatory data using multi-omics gene regulatory networks to account for both *cis* and *trans* acting mechanisms, and the local chromatin context. These prediction models perform significantly better compared to models containing *cis*-regulatory features alone. Models that additionally integrate long distance chromatin interactions (or chromatin looping) between distal transcription factor binding regions and gene promoters also show improved accuracy. As a demonstration of their utility, effect estimates from these models were used to weight *cis*-regulatory rare variants for SKAT(sequence kernel association test) analyses of gene expression.

Conclusions

Our models generate refined effect estimates for individual transcription factors, allow characterization of their roles across the genome, and provide a framework for integrating multiple data types into a single model of transcriptional regulation.

Introduction

Dysregulation of transcription and gene expression has been linked to conditions such as diabetes[1], different subtypes of cancer[2] and neurological[3], autoimmune[4] and developmental disorders[5]. However, due to the complexity of the process of transcriptional regulation in eukaryotes, the mechanistic underpinnings of many of these diseases are yet unknown. Multiple regulatory mechanisms acting in concert to control transcript abundance and gene expression have been identified[6]. Databases such as the Encyclopedia of DNA elements(ENCODE)[7], FANTOM5[8] and GEO[9] have provided researchers with the opportunity to explore gene expression regulation using computational methods. These databases contain information about the binding sites of transcription factors(TFs), coordinates of regulatory elements such as promoters and enhancers as well as epigenetic markers, and changes in expression patterns in response to external stimuli on a genome-wide level.

Despite the availability of data and many potential computational modelling techniques[10], approaches to date have been limited to predictions based on “local features”. These involve predicting gene expression using observed transcription factor and histone modification signal strengths as predictors[10]–[14]; such models estimate the average influence of regulatory elements on expression of autosomal genes. Early work conducted by Ouyang et al. built linear regression models to predict gene expression in Embryonic Stem Cells (ESCs) using TF association strengths (ChIP-Seq intensity relative to transcription start site) of 12 essential TFs and principal components to capture their “multi-collinearity”[14]. Cheng *et al.*[13] and Zhang *et al.*[11] extended this work by including ChIP-seq data for histone modifications overlapping transcription start and termination sites and applying support vector regression. Schmidt *et al.* developed the TEPIC method to calculate TF-target gene(TG) affinity scores using a biophysical model of binding based on open chromatin assay data[12]; using affinity scores as input features, they used regularized linear regression models to predict gene expression. More recently, deep learning models have become popular for this task[15]–[17], although inferring biologically relevant information from these complex models has remained a challenge. All of these approaches have produced prediction models with varying accuracy. Additionally, none of these models have accounted for the influence of trans-acting factors, such as the expression levels of and the co-operative interactions among TFs themselves.

Weighted gene regulatory networks(GRNs) attempt to fill this gap by capturing information corresponding to multiple *cis* and *trans*-acting transcriptional regulatory mechanisms in the form of edge-weights between a regulator and its TG[18]. The Passing Attributes between Networks for Data Assimilation (PANDA) algorithm generates such a GRN by extracting information from heterogeneous networks built using multiple big “omics” data sources corresponding to different TF-based regulatory mechanisms[19]. Published approaches, except for a recent extension of the TEPIC framework[20], have also not yet considered the impact of chromatin conformation on transcriptional regulation despite its increasing availability from high throughput assays such as Hi-C[21]. Condensed chromatin within the cell is heavily restructured during the process of transcription, leading to increased accessibility of gene promoters and closer physical proximity of distal transcription machinery and enhancer elements[22].

In this study, we developed a computational framework to assess the impact of multiple *cis* and *trans* regulatory mechanisms on gene expression using weighted PANDA GRNs, and to identify general mechanisms influencing the regulation of gene expression. We predicted expression for genes in the immortalized lymphoblastoid cell line GM12878 and the chronic myelogenous leukemia cell line K562. Our prediction models estimate the linear influence of individual TFs on gene expression and characterize their regulatory roles in the two cell lines. We compared the prediction performance of models built using TF binding sites(TFBS) found within various regulatory elements such as introns, promoters and distal regulatory regions, and further assessed the impact of long distance (> 1Kb) interactions between TF binding distal regulatory elements and promoters on gene regulation by integrating Hi-C data into our GRNs and prediction models. Finally, we applied our model estimates to a weighted analysis of rare cis-regulatory variants in whole-blood. Our *in-silico* prediction framework has the flexibility of including

datatypes from multiple heterogeneous sources for estimating the relative influence of multiple regulatory mechanisms on gene expression.

Methods

All the published algorithms and datasets used in this study have been described in *supplementary data*.

Defining Transcription Factor Binding Sites

We used three methods to define the TFBS between the TFs and the TGs for both the cell types using ChIP-seq data described in **Supplementary Table S1** and Ensembl gene annotations from GrCh37 human genome assembly:

1)Positional TFBS: We isolated all the ChIP-Seq peaks within a 50Kb window upstream of the TSS of the longest transcript and downstream of the body of each protein coding TG. We then used the most distant CTCF peaks to demarcate the *cis*-regulatory boundaries for these TFBS, as it is a well-known insulator protecting the enhancers of TG gene from acting upon the promoters of another as shown in **Figure1**.[36].

2) FIMO TFBS: We applied the FIMO algorithm[37] from the latest release of the MEME-suite tools(*v.5.1.1*) on the "Positional TFBS" data to find statistically significant set of TFBS. We extracted genomic sequence underneath the TF peak corresponding to each TFBS and the JASPAR(*v.2020*) based TF position weight matrices(PWM) to find statistically significant TFBS at the p-value threshold of 0.01.

3)TEPIC TFBS: We downloaded the TEPIC software (https://github.com/SchulzLab/TEPIC) along with the position specific energy matrices(PSEMS) for all TFs[12]. We used these PSEMS, the Ensembl Homo_sapiens.GRCh37.87.gtf annotation, and our predefined Positional TFBS to find affinity scores for TFs binding in the 50Kb window around each TG's TSS.

Generating Gene Regulatory Network Weightings

We converted the unique TF-TG interactions obtained from each TFBS identification method into weighted(TEPIC) and unweighted (Pos ChIP-Seq and FIMO) adjacency matrices. We used these matrices, along with BioGrid(*v.3.5.188*)[38], a method for defining protein-protein interactions (PPI), and cell-type specific co-expression networks to generate three different PANDA outputs. After 25 iterations, we obtained convergence by setting the threshold for Hamming's distance at 0.001 and by using the value of 0.1 for α for each GRN.

Generating training and test data sets for the prediction models

We used four different input datasets, for each cell type, for our prediction models based on PANDA GRN edgeweights ("Pos GRN", "FIMO GRN", "TEPIC GRN") and TEPIC affinity scores ("TEPIC") as shown in **Figure 1**. Using these matrices as inputs, we predicted the expression for independent datasets of GM12878(ENCSR889TRN) and K562(ENCSR545DKY) using the linear regularized elastic net(ENET)

regression models. We used the python-based implementation of the ENET model from the *scikit-learn* library to build the prediction models, setting the value of (the ratio between the lasso and ridge norms) at 0.5.

We used the log10-normalized FPKM(fragments per kilobase of transcripts per million) for TGs, that were common among different input matrices described in **Table 1** and also contained promoter Hi-C contacts with distal TFBS, as the response vector for the ENET prediction models. Thus, the models contained 8,644 TGs for GM12878, and 9460 TGs for K562. We also applied our approach to 12,013 TGs for HepG2 for additional validation and generalization.

We split the input feature matrix and the output expression vector into 80% training data and 20% test data. We used the training data to train the ENET models, using 20-fold inner cross validation. We then predicted the expression of the test set genes, using the learned ENET models and calculated mean squared error(MSE) and Pearson's correlation coefficient(PCC) to measure the predictive performance for the models. We repeated this process for 20 iterations as shown in **Figure1**.

Calculating TF average effect estimates

We calculated the average effect estimate for TF *T* using the following equation:

$$\bar{\beta}_T = \frac{1}{|N|}\sum_{n \epsilon N} \beta_{T,n} \qquad (1)$$

Here, is the set of random instances that we used to build our prediction models and is the effect estimate of *T* for instance . We only used the GM12878 and K562 Pos GRN prediction models in order to calculate these estimates. We further divided the TFs based on these mean effect estimates using the *xtile* function of R(*v.3.4.2*) into 5 roughly equal bins.

Additional gene regulatory elements analyses

We generated additional TFBS datasets by extracting TF peaks overlapping TG intronic regions, promoter regions (5Kb upstream of the TSS) as well the ones present in distal region beyond the promoter (**Supplementary Figure S11A**). The number of corresponding TFBS and TF-TG interactions for each cell-type representing these regions is provided in **Table 1** and **Supplementary TableS2**. In order to get the intronic regions for each TG, we first obtained the exonic regions corresponding to all the transcripts for a given TG and then subtracted them from the regions spanning the respective transcript lengths using *bedtools*(**Supplementary Figure S11B**). We added the TFBS present in the intronic regions to the positional ChIP-Seq TFBS dataset to create the intronic TFBS dataset for each cell-line. We used TF-TG interactions based on these additional TFBS datasets to create motif-based adjacency matrices and used them to build additional PANDA GRNs, which we ultimately used to predict gene expression for TGs common between the models we were comparing.

Generating Hi-C Weightings

We accessed Hi-C data for K562(GSM1551620) with 5Kb resolution and for GM12878(GSM1551688) with 1Kb resolution. We defined the promoter as the 5Kb region upstream of the TSS of the longest transcript for each gene. We normalized the Hi-C interactions using the Knight Ruiz(KR) normalization and created sparse contact matrices for both cell types. We calculated the number of contact points between each TF peak within a gene's distal regulatory region and its promoter using *bedtoolsv.2.27.1*. We then calculated the HiC adjusted edge-weights between each TF and TG using the following formula:

$$C_{i,g} = 1 + scaled\left(\frac{1}{N_{i,g}} \sum_{p \epsilon P_{i,g}} c_p\right) \qquad (2)$$

Here, is the Hi-C adjusted edge weight between TF and TG , is the number of ChIP-seq peaks corresponding to in the regulatory region of , is the set of peaks corresponding to in the regulatory region of and is the number of KR normalized contacts made by peak with the promoter of . We used the MinMax scaling function of the *scikit-learn* library to scale the mean contacts within the (0,0.99) range. Thus, if the TF did not contain any peaks interacting with a gene's promoter, the would be equal to 1 and the maximum value for would be 1.99. We generated the cell type specific "Hi-C DP" motif adjacency matrix using these scaled interactions. We then extracted all the promoter-based TF-TG interactions that were down-weighted to 1.0, or were found to have no Hi-C interactions, in the "Hi-C DP" matrix and gave them maximum weight of 2.0 to create the cell-type specific "Hi-C UP" adjacency matrix. We created two new GRNs using these adjacency matrices as motif networks along with the cell-type specific PPI and co-expression data to build prediction models following the workflow described in **Figure 1.**

QBiC-Pred-GRN rare variant association analysis.

We followed the workflow shown in **Supplementary Figure S13** for the rare variant analysis. We generated GM12878 GRN utilizing the intronic TFBS for motif network and HiC up weighting scheme described previously. We then fit the ENET models using TF-TG edgeweight features from this GRN, and used the learned models to compute average TF effect estimates based on *equation (1)*. For the initial discovery analysis, we used the depression genes and networks(DGN) data set, which contains genotypes and RNA-seq data for 922 individuals of European descent[40].We further imputed variant genotypes using 1000 genomes reference panel and the University of Michigan imputation server[39], [40]. We extracted rare variants at a minor allele frequency(MAF) threshold of 1%(N ≈ 9.4M variants) and overlapped them with the GM12878 intronic TFBS.

Out of the 149 TFs, we were able to find trained QBiC-Pred models for 59 TFs We scored these variants using the offline version of the QBiC-Pred software[25] which we downloaded from the github repository (https://github.com/vincentiusmartin/QBiC-Pred). We used the p-value threshold of 0.0001 to identify the variants significantly impacting the TFBS we identified 118,789 rare variants that were present within their binding sites.

We merged the z-score obtained from the QBiC-Pred algorithm and the TF effect estimates for each rare variant present within the TFBS for each TG using the following sets of equations.

$$Z_{v,t,g} = \bar{\beta}_t \times \frac{\sum_{p_{t,g} \in P_{T,g}} Z_{v,p_{t,g}}}{|P_{t,g}|} \quad (3)$$

$$S_{v,g} = \frac{\sum_{t \in T_g} Z_{v,t,g}}{|T_g|} \quad (4)$$

Here, is the QBiC-Pred z-score for variant significantly impacting the peak region(TFBS) , which is a subset of all the peak regions TF within the regulatory/intronic regions of TG . is the average ENET effect estimate obtained from the learned ENET models for TF and is the scaled QBiC-Pred z-score for variant corresponding to TF binding *cis*-regulatory/intronic regions for TG . is the merge score for variant for each TG computed by averaging the scaled z-scores for all the TFs present within the *cis*-regulatory/intronic regions of TF ( ). We also computed aggregate QBiC-Pred z-scores for each variant present within all the TFBS for each TG without utilizing the average effect estimates. In other words, we simply removed the effect estimate ( ) from the set of equations described above. We scaled both aggregated z-scores and merge scores within the range [-1,1] and used them for weighting the variants.

We used the R implementation of the SKAT algorithm [24](*v 2.0.0*) in order to find association between these sets of variants and the TG expression levels normalized by HCP(hidden covariates prior). We used the merge scores and QBiC-Pred aggregated z-scores as variant weights for the SKAT kernel matrices and fit the models for 11,650 TGs using 74 additional biological and technical covariates provided within the DGN dataset.

For replication analysis, we utilized the Genotype-Tissue Expression(GTEx) dataset containing whole genome sequencing and RNA-seq data for 369 individuals[26] (**Supplementary Figure S13**). We repeated the analysis done for the DGN dataset to extract and score variants and then performed SKAT using the normalized expression of TGs that were found significant in the DGN analysis and whose expression values were present in the GTEx dataset(N = 388). For GTEx analysis, we utilized the 65 covariates provided within the dataset to fit the SKAT model.

Statistical Evaluations

We used R *v.3.4.2* to perform all the statistical analyses in our study. Assuming a non-normal distribution of the PCC and MSE produced by the prediction models, we used the Wilcoxon rank sum test to compare medians of these performance measures for different models. We used the *gseapy* package in python for Gene Ontology(GO) enrichment analyses. We divided the TFs into 5 bins (quintiles) based on their average effect estimates and ran the enrichment analysis for GO Biological Processes (GO BP) and GO Molecular Functions (GO MF) terms using all cell-type specific TFs as background. We specifically looked for significant enrichment terms (adjusted p-value < 0.05) for each bin for both the GO categories. We

then extracted the top 5 significant enrichment terms for each bin (provided in **Supplementary TablesS6C** and **S6D**) based on their p-values and plotted them in **Supplementary Figure S10**.

Results

Accounting for *trans* acting mechanisms in addition to *cis* regulatory mechanisms improved gene expression prediction significantly.

We built gene expression prediction models using information from GRNs, applying TF-TG edge-weights from these networks as predictors, under the hypothesis that accounting for *trans*-acting mechanisms would improve model accuracy. We then compared our results to those obtained from using TF-TG affinity scores calculated based on *cis*-regulatory features using the TEPIC algorithm. The workflow for our approach is provided in **Figure 1.** The number of TFs, TGs and TFBS corresponding to different ChIP-seq processing algorithms for both cell lines is provided in **Table 1.** All validation results within the HepG2 cell-line are shown in the supplementary section and in **Supplementary Figure S1**.

Table 1: Number of TFs, TGs and TFBS obtained from different ChIP-seq processing algorithms for GM12878 and K562 cell lines. All the ChIP-seq data for the analysis was downloaded from the ENCODE database. We have also included the TFBS and interactions identified for latter analysis.

	GM12878				K562			
	TFs	**TGs**	**TFBS**	**Unique TF-TG Pairs**	**TFs**	**TGs**	**TFBS**	**Unique TF-TG Pairs**
Pos ChIP-Seq	149	17,106	4,209,133	1,216,272	309	18,190	11,614,248	2,372,274
FIMO	85	16,850	2,444,195	714,167	110	18,173	7,349,429	1,138,823
TEPIC	80	11,784	-	517,226	86	10,239	-	880,554
Promoter	149	11,509	458,959	276,138	308	15,668	1,293,933	681,847
Distal	149	16,964	3,750,174	1,128,079	309	18,152	10,320,315	2,312,490
Intronic	149	17,106	5,896,338	1,378,129	309	18,224	14,764,766	2,820,604

We created binary (binding/no-binding) TF-TG adjacency matrices using the positional and FIMO TFBS. For the TEPIC based adjacency matrix, we used affinity scores as weights. We combined these matrices with PPI data and cell type specific co-expression to fit a GRN using the PANDA algorithm. We fit four models using elastic-net based regularized linear regression (ENET) for each cell line. The first three models (FIMO GRN, Pos GRN, TEPIC GRN) used the edge weights from the respective GRN as input features to predict expression for test set TGs. The fourth model (TEPIC) used only TF-TG affinity scores without GRN information. Predictive performance was measured using mean-squared error (MSE), and

Pearson's correlation coefficient (PCC) between predicted and observed expression values within a 5-fold cross-validation framework repeating for 20 iterations.

We built the GRNs using motif networks derived from CTCF boundary defined *cis*-regulatory TFBS(**Figure-1**). We observed that models using such GRNs performed statistically significantly better than or similar to those built using a 50Kb window to define the *cis*-regulatory regions of the TGs (**Supplementary Figure S2A**). Furthermore, the model performance for GRNs built using binary adjacency matrices for motif networks was comparable to those utilizing a weighted motif network based on the number of TFBS for each TF-TG interaction (**Supplementary Figure S2B**).

As shown in **Figure 2** and in **Supplementary Figure S1**, GRN based prediction models containing *cis* and *trans* regulatory mechanisms were more accurate than models built using only *cis*-regulatory TF-TG TEPIC affinity scores. Specifically, the median PCC for TEPIC GRN based models was higher compared to that of TEPIC for GM12878 (0.42 vs. 0.30, p=1.45e-11 **2A**), K562(0.30 vs. 0.28, p=3.50e-2 **2C**) while the median MSE for the former was lower than that for the latter for GM12878(0.83 vs. 0.91, p = 4.35e-10 **2B**), K562 (0.91 vs. 0.93, p=3.26e-02 **2D**). Additionally, we achieved a statistically significantly better prediction performance while using the GRNs constructed from TFs, for which we could calculate TEPIC affinity scores, used in the original TEPIC paper for each cell-line compared to TEPIC affinity scores as shown in **Supplementary Figure S3**. These results generalized to the HepG2 cell-line as well (**Supplementary Figure S1**).

While the top three and bottom three ranked TFs, based on their average ENET effect estimates, did not change in our TEPIC GRN models compared to the TEPIC models, 33 of the 80 for GM12878 and 38 of 86 K562 had a change of rank greater than 10 positions (**SupplementaryTablesS3***)*. TFs that improved in rank were associated with transcriptional activation, and those that decreased in rank were associated with transcriptional repression. Thus, capturing *trans*-acting regulatory mechanisms in their cellular context produces better predictions of gene expression and a better understanding of the relative importance of TFs in the transcription process. Results from all analyses (median measures, p-values and TF ranks) are provided in **Supplementary Tables S4A** – **S4I** and **S5A-B**.

Performance of all GM12878 based models was better overall compared to those obtained from the other two cell lines. This observation was true even when the Pos GRN models were constructed using a common set of TFs among the three cell-lines (**Supplementary Figure S4A-B**). Moreover, these models also performed the best when used to predict cross-cell type TG expression (**Supplementary Figure S5C-D)**. Furthermore, the GM12878 Pos GRN also contained the highest proportion of literature annotated TF-TG interactions of the three cell-lines (**Supplementary Figure S6**). Thus, GM12878 GRNs were capturing the most accurate regulatory information between TFs and TGs and the models based on them were explaining the most variance in the gene expression trait. We hypothesized that this may have occurred due to the better estimates of the expression correlation network in the GM12878 line. To explicitly test this, we examined the influence of co-expression networks over the TF-TG regulatory relationships learned by the PANDA GRN by replacing them with networks built using identity matrices to build prediction

models (**Supplementary FigureS7)**. The models built using motif and co-expression information performed statistically significantly better compared to those built using just motif information or motif and PPI information, illustrating the importance of having a robust co-expression dataset for generating the GRN weights.

Pos GRN models for GM12878 and K562 had the best performance of all models tested. This could be due to overfitting caused by the higher number of TF features in these models. To test this, we restricted the analysis to GRNs built using the same number of TFs. The performance of Pos GRN in this restricted analysis was similar to that of FIMO GRN for all cell-lines (**Supplementary Figure S4**). Furthermore, in this sensitivity analysis we found that TEPIC GRN-based models were the most accurate for K562 and HepG2. In other words, when restricted a common set of TFs, TEPIC affinity scores were able to capture more regulatory information between TFs and TGs in comparison to simple positional ChIP-seq data and statistically defined FIMO-based TFBS for K562 and HepG2. Thus, the difference in performance of TF binding assessed by the three methods is either due to the specific collection of TFs assayed for a particular cell-type or just biological differences between the cell-types.

Expression prediction highlights the regulatory roles of transcription factors

Transcription factors may influence gene expression as core activating factors, as responsive factors to environmental stimuli, or as repressors. ENET regression models allow for this heterogeneity by linearly combining two penalizing terms, LASSO(L1) and Ridge(L2), that identify the most influential features(TFs) and shrink the weights of lesser features by either reducing them to 0 (L1) or to a very small number (L2). We examined the influence of the hyper-parameter α, which controls the ratio between the two terms in ENET models, on the prediction performance of genes for each cell type. As shown in the **Supplementary Figure S8**, moving towards a higher value of α improved the gene expression with a plateau at 0.5 after which improvement was not significant. This indicates a balance between a sparse regulatory model where certain TFs have large effects on gene regulation, and a distributed regulatory model where multiple TFs contribute small effects.

Using an α of 0.5 (balancing L1 and L2 penalties) and equation (1), we averaged the feature weights of 149 TFs(GM12878) and 309 TFs(K562) for the Pos GRN models fit for 20 iterations, which are shown in **Figure 3** and in **Supplementary Tables S6A** and **S6B**.

For some collections of correlated TFs, collinearity was reduced in the learned models by shrinking effect estimates of some TFs close to zero (in the range [-0.01, 0.01]) (**Supplementary Figure S9**), as such these TFs may be representative of larger TF complexes. Histograms in **Figure 3** show 5 roughly equal bins created using mean effect estimates. The overall range of the mean effect estimates for both cell types, as well as the ranges within each bin, are provided in **Supplementary TableS7.** We performed a GO enrichment analysis for TFs in each bin and reported the top 5 enrichment terms for biological processes and molecular functions in the **Supplementary Tables S6C** and **S6D** and in **Supplementary Figure S10** for both cell types. We observed that as we moved from positive to negative TF effect coefficients (bin 5 to bin 1), the corresponding GO terms changed from reflecting transcriptional activation to those indicating

transcriptional repression. Thus, we could derive functions of unannotated TFs based on the bins in which they are placed. For instance, K562 bin 1 contained MYNN(β_{K562} = -0.0059) whose function is largely unknown. However, based on its placement in the bin containing strong repressors such as CBX1(β_{K562} = -0.0188), HDAC6(β_{K562} = -0.0045) and BMI1(β_{K562} = -0.0341), we predict its function is related to transcriptional repression. Similarly, bin 5 for both K562 and GM12878 contained TFs related to core promoter activity and positive gene expression regulation such as TAF1($\beta_{GM12878}$ = 0.6334), TBP($\beta_{GM12878}$ = 0.2142), ELF1($\beta_{GM12878}$ = 0.2249), POLR2A(β_{K562} = 0.1123), POLR2G(β_{K562} = 0.0233), CHD1(β_{K562} = 0.0492)and MYC($\beta_{GM12878}$ = 0.1481). Relatively lesser known TF ZZZ3($\beta_{GM12878}$ = 0.1359; β_{K562} = 0.0375), which was also present in that bin may most likely play a similar transcriptional activation role. We also note that TFs with mean effect estimates very close to or equal to zero were present in bin 2 for GM12878 and in bins 2 and 3 for K562. These TFs were enriched for cofactor activity, and their functional annotations reflected their roles as secondary TFs that required binding of the primary TFs to the DNA in order to exert their influence. Thus, analyzing the prediction models and characterizing TFs based on their effect estimates enables us to hypothesize about the transcriptional regulatory roles of lesser known TFs.

Accounting for chromatin interactions between TFBS and gene promoters improves expression prediction

In order to determine the effect of TF binding within different regulatory regions on gene expression, we built prediction models using GRNs containing TFBS found in those regions and calculated their predictive performance. We first analyzed TFBS within the promoter regions (5Kb upstream of the TSS of the genes), intronic TFBS, and distal ones present outside these areas as shown in **Supplementary Figure S11**. The promoter region near the TSS of the gene is important for transcription initiation and regulation and it contains binding sites for pivotal pioneer TFs such as TAFs, POL2 subunits, and TBP. As expected (**Figures 4A,4B** and the **Supplementary Tables S4A- S4I**), the median PCC and MSE for the promoter TFBS based ENET models were significantly better than that of the ones containing the distal TFBS alone for GM12878(MSE p = 3.26e-02; PCC p =2.92e-04), K562(MSE p =3.75e-02, PCC p = 3.26e-02). Also, models containing intronic TFBS performed significantly better than those without (**Figures 4C, 4D**) with respect to median MSE (GM12878 p = 4.72e-04) and median PCC(GM12878 p = 1.33e-08; K562 p = 2.45e-02). The results corresponding to the HepG2 cell-line reflected similar predictive patterns (**Supplementary Figure S1**).

We next used Hi-C data corresponding to GM12878 and K562 in order to capture long distance interactions between distal TF binding and gene promoters. We used the motif adjacency matrices and weighted them based on the number of normalized Hi-C contacts between TF peaks and TG promoters for both cell lines using equation (2) as shown in **Figure 5A**. Prediction models including Hi-C adjusted distal TFBS were significantly more accurate compared to the ones built using normal distal TFBS as shown in **Figure5B** and **Supplementary TablesS4A – S4I** with regards to both PCC(GM12878 p =7.33e-03; K562 p = 2.00e-06) and MSE(GM12878 p = 1.43e-03; K562 p =5.61e-03) for both cell types.

Next, we expanded this weighting scheme to include promoter TFBS. As promoters are regions of high TFBS activity (as seen in our models, **Figure 4A** and **4B**), we expected a high degree of Hi-C contact points within promoter regions. Unexpectedly, these models performed significantly worse; we observed a large number of promoter TFBS (59% for GM12878 and 90% for K562) that showed no evidence of within-promoter contacts, and using this weighing approach effectively down-weighted promoter TF-TG interactions (Hi-C DP). We therefore also considered an approach that applies the maximum Hi-C weight to all promoter TFBS (Hi-C UP), shown in **Figure 5A**. These Hi-C UP based prediction models significantly outperformed all the other models for both cell types as shown in **Figure 5C** and **Supplementary Figure S12**. Thus, Hi-C data added important regulatory information to our models capturing the effect of long distance interactions between TFs binding to distal regulatory elements and the TG promoter.

Weighting rare variants using GRN derived effect estimates enriches the SKAT based identification of significant TGs

Determining the impact of rare non-coding variants on TG regulation is a major challenge in the field of human genetics[23]. Here, we present the utility of our GRN based ENET models for weighting rare variants within kernel-based association tests to improve their power. We used the DGN dataset[43] containing HRC-imputed variant genotypes and RNA-seq from the whole blood of 922 individuals in order to perform SKAT[24] based rare variant analysis(**Supplementary Figure S13**). We generated a PANDA GRN for GM12878 based on intronic TFBS motif network weighted using HiC-UP weighting scheme described earlier and then used it to build ENET prediction models and subsequently derived average TF effect estimates. We extracted approximately 9.4 million rare SNPs(MAF < 0.01) from the DGN dataset and scored them based on their impact on TF binding intensity using the QBiC-Pred algorithm[25]. By merging this score with the average effect estimates of the corresponding TFs, based on *equations (3)* and *(4)*, we created a variant scoring metric representing the estimated average effect of a base-pair change on TF-TG regulation. We used these merged scores to perform SKAT for the normalized expression of TGs in the DGN dataset. We compared the performance of this model to that obtained from aggregated QBiC-Pred z-scores, representing the effect of rare variants on TF-binding alone. As shown in **Supplementary Figure S14A**, both SKAT models were able to detect 175 common TGs at the multiple hypothesis correction significance threshold of p-value < 4.18e-06. Merge score based SKAT model was able to detect 158 unique TGs while z-score based model detected 56 unique TGs at this threshold. We also performed a replication analysis using the whole blood sequencing and expression data from 369 individuals within the GTEx dataset[26]. We were able to replicate 32% of the TGs uniquely identified by merge score based SKAT model (p-value < 0.05), while only 21% of the TGs uniquely identified by the QBiC-Pred z-score SKAT model replicated (**Supplementary Figure S14B**). We have provided the results from all the SKAT models in **Supplementary Tables S8A-S8C**. Thus, utilizing TF-TG regulatory information learned from our GRN framework for weighting rare variants enriched the identification of TGs, which would have been missed if we had only utilized variant influence over TF binding.

Discussion

In this study, we developed a modelling framework to predict gene expression within two cellular contexts using gene regulatory networks to capture the *trans* effect of cooperativity and co-regulation on *cis* regulatory factors relative to their TGs. Our approach explained more variance in gene expression compared to models built using TF-TG affinity scores for *cis*-regulatory features alone. The prediction models for the HepG2 and K562 cell lines were significantly less accurate than for GM12878. One of the reasons for this observation could be the markedly smaller sample size of the PANDA expression data for K562 and HepG2 relative to GM12878 (9 and 8 vs 462). In fact, the prediction performance of GRNs containing motif and co-expression information was comparable to those containing information corresponding to all three datasets for all the cell-lines. Thus, having a more extensive co-expression dataset is essential to estimate accurate TF-TG regulatory relationships, which ultimately leads to better prediction models. The dearth of expression datasets for the K562 and HepG2 cell-lines could be potentially alleviated by using more robust scRNA-seq instead of total RNA-seq datasets. However, we also realize that another explanation for the superior performance of the GM12878 based models could be bias (technical or happenstance). This bias may have led to GM12878 models containing TFs essential for transcriptional regulation resulting in a better capturing of variance in the gene expression trait.

We further estimated the influence of individual TFs on gene expression outcomes based on their effect coefficients learned from our models. This led to a ranked list of activating and repressive factors influencing transcriptional regulation in both cell lines, including classifications of TFs with previously unknown effects. We observed substantial changes to the ranking of TFs relative to analyses using *cis*-factors alone, illustrating the importance of accounting for the cellular context in interpreting TF effects. While TFs with the strongest and the weakest effects were roughly the same between our baseline TEPIC model and the model overlaid with GRN weights, many TFs with activating and repressive properties show stronger effect estimates after accounting for information captured by the GRN.

As expected, we observed that the highest ranking TFs are crucial for transcriptional initiation and activation, falling within promoter regions of a majority of protein coding genes. The process by which transcriptional machinery forms at the promoter regions of genes has been extensively studied[27] . Promoter TFBS based models were also significantly more accurate at predicting gene expression than models using distal TFBS alone. These results validate our modeling strategy, as these findings are consistent with observations from previous studies[13], [28], and further highlight the important role that promoter regions play in regulating gene expression.

Hi-C data was useful for characterizing long distance interactions between distal TFBS and the gene's promoter. Integrating this data into the PANDA GRNs improved the prediction performance of the models when scaled relative to promoter TFBS. This improvement was also observed in the recently published extension of the TEPIC framework[20].We observed significant improvement in both cell lines despite differences in Hi-C resolution (1Kb for GM12878 and 5Kb for K562), however the resolution difference may account for the greater improvement in prediction for GM12878 relative to K562.

Our results also indicate that intronic TFBS provide significant prediction power to the models. There are two likely explanations for this observation. First, introns may bind regulatory TFs or splicing factors that alter the rate of transcription. Previous studies looking at the role of first introns in regulating transcription in *C.elegans* found genome wide occurrence of TFBS in these regions are important in driving gene expression[29], [30]. Second, introns could house alternate promoters for a gene, as noted by analyses of GTEx and FANTOM datasets[31]. For our analyses, we used the upstream TSS of the longest transcript to define gene promoter regions.

Finally, we utilized the TF-TG regulatory information learned from our GRN based framework in order to weight rare variants. This weighting approach led to a significant improvement in power of kernel based SKAT models to detect significant associations with TG expression relative to using weights capturing TF binding affinity alone. While we used QBiC-Pred to score TF binding affinity, multiple other scoring approaches could also be used within the framework. These analyses demonstrate the utility of our models for annotating otherwise difficult to characterize regulatory variants.

The most direct comparison of predictive performance for our models against published methods is the TEPIC method, which we outperformed. Other approaches have included either more complex modeling techniques or additional histone modification data to improve model performance[11], [13]. Non-linear prediction models such as support vector regression or multi-layer perceptrons applied within our framework may capture more complex interactions among TFs and improve performance. It also remains unclear to what extent the epigenetic context influences the effect a transcription factor has on gene expression. Zhang et al. have demonstrated some redundancy between histone modification and TF binding intensities with respect to gene expression prediction[11]. Thus, inclusion of both histone modification data and TF binding as predictors could diminish the effect of individual TFs, clouding the interpretation of our predictions.

At present, our approach is limited by the availability of ChIP-seq data. Although large scale efforts such as the ENCODE consortium have produced binding data for a large number of TFs in different cell types, this number is still small compared to the actual TFs being expressed in a cell at any given time[32] . This dearth in data availability is due to the difficult and expensive nature of the ChIP-seq experiments themselves[33]. One way to potentially incorporate histone modification and chromatin accessibility data is through the imputation of TF binding not directly measured by ChIP-seq experiments for a given cellular context through techniques like DeepSEA or FactorNet[34], [35]. In future work, these TF binding predictions could supplement the set of inputs to our GRN-based framework to produce better models.

Conclusion

The modelling approach, we have presented here, has multiple applications for studying general factors influencing gene expression. Our models provide an approach for annotating the regulatory structure of a given gene in a tissue or cell-type specific manner, for ranking TFs in order of their likely impact on gene expression, and for clustering genes based on their weighted regulatory features. Our framework also

allows for the inclusion of additional functional genomics mechanisms, such as higher resolution chromatin interaction data, to evaluate their effect. As our understanding of chromatin accessibility and conformation grows, the framework can also be used to better define the *cis*-regulatory window surrounding a gene, which can be useful for eQTL mapping and other downstream analyses. Finally, prioritizing TFs relative to gene expression allows for better prioritization of genetic variants and their influence on nearby gene expression traits. More generally, our approach provides a roadmap for integrating multiple “omics” data sources and assembling fundamental aspects of transcriptional regulation into a coherent portrait of gene expression.

Abbreviations

TF: Transcription Factors

TG: Target Gene

ChIP-Seq: Chromatin Immunoprecipitation Sequencing

PANDA: Passing Attributes between Networks for Data Assimilation

GRN: Gene Regulatory Network

TFBS: Transcription Factor Binding Site/s

FIMO: Find Individual Motif Occurences

PWM: Position Weight Matrix

ENET: Elasticnet

FPKM: Fragments per kilobase of transcripts per million

GO: Gene Ontology

PCC: Pearsons Correlation Coefficient

MSE: Mean Squared Error

QBiC-Pred: Quantitative TF Binding Change Predictions Due to Sequence Variants

eQTL: Expression quantitative trait loci

SKAT: Sequence kernel association test

DGN: Depression genes and networks

SNP: Single Nucleotide Polymorphism

GTEx: Genotype-Tissue Expression

GO MF: Gene Ontology Molecular Function

GO BP: Gene Ontology Biological Process

PSEMS: Position Specific Eneergy Matrices

TSS: Transcription Start Site

PPI: Protein-protein interactions

ENCODE: Encyclopedia of DNA elements

Declarations

- **Ethics approval and consent to participate**

The study conducted a secondary analysis of de-identified data from the DGN and GTEx datasets.

- **Consent for publication**

Not applicable

- **Availability of data and materials**

We have created a docker container as well as a github repository with all the data files and scripts used for analysis. (https://hub.docker.com/orgs/bushlab/repositories/centos_tf_grn , https://github.com/bushlab-genomics/TF_GRN) Bio-samples and/or data for this publication were obtained from NIMH Repository & Genomics Resource, a centralized national biorepository for genetic studies of psychiatric disorders. The GTEx dataset used for the analyses described in this manuscript were obtained from dbGaP at

http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v8.p2.

- **Competing interests**

Not applicable

- **Funding**

This research was supported in part by grants R01 AG061351 (Below, Naj, Bush) , R01 AG059716 (Hohman), and U01 AG058654 (Haines, Bush, Martin, Farrer, Pericak-Vance) from the National Institute on Aging, and by T32 HL007567 (Zhu) from the National Heart Lung and Blood Institute.

- **Authors' contributions**

NP and WSB conceptualized and designed the research project, as well as wrote the manuscript. NP collected the data and performed all the subsequent analyses.

- **Acknowledgements**

We would like to thank Dr. Penelope Benchek and Dr. Yeunjoo Song for helping us with DGN genotype imputation. The computational analysis was done using the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. We have included detailed acknowledgements for the DGN and GTEx dataset within the supplemental information.

References

1. H. K. Pedersen, V. Gudmundsdottir, and S. Brunak, “Pancreatic Islet Protein Complexes and Their Dysregulation in Type 2 Diabetes ,” *Frontiers in Genetics* , vol. 8. p. 43, 2017.
2. T. J. Gonda and R. G. Ramsay, “Directly targeting transcriptional dysregulation in cancer,” *Nat. Rev. Cancer*, vol. 15, no. 11, pp. 686–694, 2015, doi: 10.1038/nrc4018.
3. Z. S. Chen and H. Y. E. Chan, “Transcriptional dysregulation in neurodegenerative diseases: Who tipped the balance of Yin Yang 1 in the brain?,” *Neural Regen. Res.*, vol. 14, no. 7, pp. 1148–1151, Jul. 2019, doi: 10.4103/1673-5374.251193.
4. A. I. Ramsingh, K. Manley, Y. Rong, A. Reilly, and A. Messer, “Transcriptional dysregulation of inflammatory/immune pathways after active vaccination against Huntington's disease,” *Hum. Mol. Genet.*, vol. 24, no. 21, pp. 6186–6197, Aug. 2015, doi: 10.1093/hmg/ddv335.
5. T. I. Lee and R. A. Young, “Transcriptional Regulation and Its Misregulation in Disease,” *Cell*, vol. 152, no. 6, pp. 1237–1251, Mar. 2013, doi: 10.1016/j.cell.2013.02.014.
6. K. M. Lelli, M. Slattery, and R. S. Mann, “Disentangling the Many Layers of Eukaryotic Transcriptional Regulation,” *Annu. Rev. Genet.*, vol. 46, no. 1, pp. 43–68, Nov. 2012, doi: 10.1146/annurev-genet-110711-155437.
7. C. A. Davis *et al.*, “The Encyclopedia of DNA elements (ENCODE): data portal update,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D794–D801, Nov. 2017, doi: 10.1093/nar/gkx1081.
8. A. R. R. Forrest *et al.*, “A promoter-level mammalian expression atlas,” *Nature*, vol. 507, no. 7493, pp. 462–470, 2014, doi: 10.1038/nature13182.
9. T. Barrett *et al.*, “NCBI GEO: archive for functional genomics data sets—update,” *Nucleic Acids Res.*, vol. 41, no. D1, pp. D991–D995, Nov. 2012, doi: 10.1093/nar/gks1193.
10. D. M. Budden, D. G. Hurley, and E. J. Crampin, “Predictive modelling of gene expression from transcriptional regulatory elements,” *Brief. Bioinform.*, vol. 16, no. 4, pp. 616–628, Sep. 2014, doi: 10.1093/bib/bbu034.
11. L.-Q. Zhang and Q.-Z. Li, “Estimating the effects of transcription factors binding and histone modifications on gene expression levels in human cells,” *Oncotarget*, vol. 8, no. 25, pp. 40090–40103, Jun. 2017, doi: 10.18632/oncotarget.16988.

12. F. Schmidt *et al.*, “Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction,” *Nucleic Acids Res.*, vol. 45, no. 1, pp. 54–66, Nov. 2016, doi: 10.1093/nar/gkw1061.
13. C. Cheng and M. Gerstein, “Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells,” *Nucleic Acids Res.*, vol. 40, no. 2, pp. 553–568, Sep. 2011, doi: 10.1093/nar/gkr752.
14. Z. Ouyang, Q. Zhou, and W. H. Wong, “ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells,” *Proc. Natl. Acad. Sci.*, vol. 106, no. 51, p. 21521 LP-21526, Dec. 2009, doi: 10.1073/pnas.0904863106.
15. G. Robins, J. Lanchantin, R. Singh, and Y. Qi, “DeepChrome: deep-learning for predicting gene expression from histone modifications,” *Bioinformatics*, vol. 32, no. 17, pp. i639–i648, Aug. 2016, doi: 10.1093/bioinformatics/btw427.
16. J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyanskaya, “Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk,” *Nat. Genet.*, vol. 50, no. 8, pp. 1171–1179, Aug. 2018, doi: 10.1038/s41588-018-0160-6.
17. R. Xie, J. Wen, A. Quitadamo, J. Cheng, and X. Shi, “A deep auto-encoder model for gene expression prediction,” *BMC Genomics*, vol. 18, no. 9, p. 845, 2017, doi: 10.1186/s12864-017-4226-0.
18. F. Emmert-Streib, M. Dehmer, and B. Haibe-Kains, “Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks ,” *Frontiers in Cell and Developmental Biology* , vol. 2. p. 38, 2014.
19. K. Glass, C. Huttenhower, J. Quackenbush, and G.-C. Yuan, “Passing Messages between Biological Networks to Refine Predicted Interactions,” *PLoS One*, vol. 8, no. 5, p. e64832, May 2013.
20. F. Schmidt, F. Kern, and M. H. Schulz, “Integrative prediction of gene expression with chromatin accessibility and conformation data,” *Epigenetics Chromatin*, vol. 13, no. 1, p. 4, 2020, doi: 10.1186/s13072-020-0327-0.
21. J.-M. Belton, R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker, “Hi–C: A comprehensive technique to capture the conformation of genomes,” *Methods*, vol. 58, no. 3, pp. 268–276, 2012, doi: https://doi.org/10.1016/j.ymeth.2012.05.001.
22. B. Li, M. Carey, and J. L. Workman, “The Role of Chromatin during Transcription,” *Cell*, vol. 128, no. 4, pp. 707–719, Feb. 2007, doi: 10.1016/j.cell.2007.01.015.
23. O. Bocher and E. Génin, “Rare variant association testing in the non-coding genome,” *Hum. Genet.*, vol. 139, no. 11, pp. 1345–1362, 2020, doi: 10.1007/s00439-020-02190-y.
24. M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, “Rare-variant association testing for sequencing data with the sequence kernel association test.,” *Am. J. Hum. Genet.*, vol. 89, no. 1, pp. 82–93, Jul. 2011, doi: 10.1016/j.ajhg.2011.05.029.
25. V. Martin, J. Zhao, A. Afek, Z. Mielko, and R. Gordân, “QBiC-Pred: quantitative predictions of transcription factor binding changes due to sequence variants,” *Nucleic Acids Res.*, vol. 47, no. W1, pp. W127–W135, Jul. 2019, doi: 10.1093/nar/gkz363.

26. J. Lonsdale *et al.*, “The Genotype-Tissue Expression (GTEx) project,” *Nat. Genet.*, vol. 45, no. 6, pp. 580–585, 2013, doi: 10.1038/ng.2653.
27. P. J. Robinson *et al.*, “Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex,” *Cell*, vol. 166, no. 6, p. 1411–1422.e16, 2016, doi: https://doi.org/10.1016/j.cell.2016.08.050.
28. T. Schacht, M. Oswald, R. Eils, S. B. Eichmüller, and R. König, “Estimating the activity of transcription factors by the effect on their target genes,” *Bioinformatics*, vol. 30, no. 17, pp. i401–i407, Sep. 2014, doi: 10.1093/bioinformatics/btu446.
29. J. I. Fuxman Bass *et al.*, “Transcription factor binding to Caenorhabditis elegans first introns reveals lack of redundancy with gene promoters,” *Nucleic Acids Res.*, vol. 42, no. 1, pp. 153–162, Jan. 2014, doi: 10.1093/nar/gkt858.
30. A. B. Rose, “Introns as Gene Regulators: A Brick on the Accelerator ,” *Frontiers in Genetics* , vol. 9. p. 672, 2019.
31. A. Reyes and W. Huber, “Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues,” *Nucleic Acids Res.*, vol. 46, no. 2, pp. 582–592, Nov. 2017, doi: 10.1093/nar/gkx1165.
32. S. A. Lambert *et al.*, “The Human Transcription Factors,” *Cell*, vol. 172, no. 4, pp. 650–665, 2018, doi: https://doi.org/10.1016/j.cell.2018.01.029.
33. J. Keilwagen, S. Posch, and J. Grau, “Accurate prediction of cell type-specific transcription factor binding,” *Genome Biol.*, vol. 20, no. 1, p. 9, Jan. 2019, doi: 10.1186/s13059-018-1614-y.
34. J. Zhou and O. G. Troyanskaya, “Predicting effects of noncoding variants with deep learning–based sequence model,” *Nat. Methods*, vol. 12, no. 10, pp. 931–934, 2015, doi: 10.1038/nmeth.3547.
35. D. Quang and X. Xie, “FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data,” *Methods*, vol. 166, pp. 40–47, 2019, doi: https://doi.org/10.1016/j.ymeth.2019.03.020.
36. C.-T. Ong and V. G. Corces, “CTCF: an architectural protein bridging genome topology and function,” *Nat. Rev. Genet.*, vol. 15, no. 4, pp. 234–246, 2014, doi: 10.1038/nrg3663.
37. C. E. Grant, T. L. Bailey, and W. S. Noble, “FIMO: scanning for occurrences of a given motif,” *Bioinformatics*, vol. 27, no. 7, pp. 1017–1018, Feb. 2011, doi: 10.1093/bioinformatics/btr064.
38. R. Oughtred *et al.*, “The BioGRID interaction database: 2019 update,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D529–D541, Nov. 2018, doi: 10.1093/nar/gky1079.
39. B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis, “Fast and accurate genotype imputation in genome-wide association studies through pre-phasing,” *Nat. Genet.*, vol. 44, no. 8, pp. 955–959, 2012, doi: 10.1038/ng.2354.
40. C. Fuchsberger, G. R. Abecasis, and D. A. Hinds, “minimac2: faster genotype imputation,” *Bioinformatics*, vol. 31, no. 5, pp. 782–784, Mar. 2015, doi: 10.1093/bioinformatics/btu704.
41. Battle, A. et al(2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 24, 14–24.

Figures

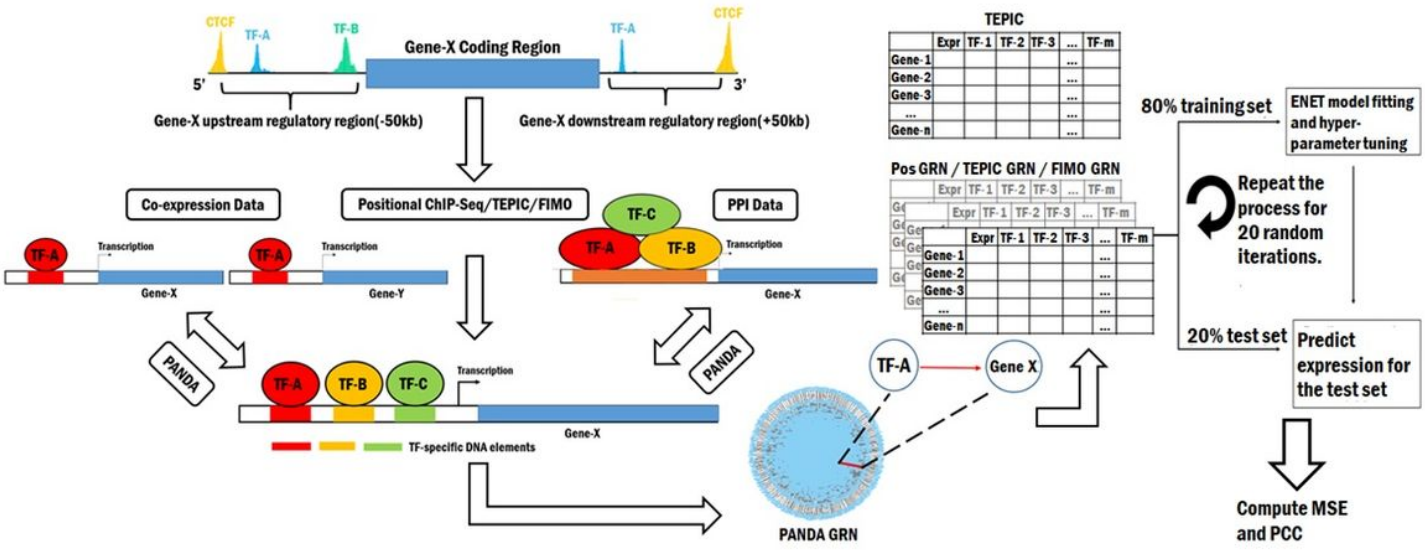

Figure 1

Workflow for building prediction models using multi-omics GRNs. ChIP-seq data for 153 TFs(GM12878) and 382 TFs(K562) having peaks passing the optimal IDR threshold defined by ENCODE were mapped to the regulatory region of each gene to define TFBS. The most distant CTCF peaks within a 50Kb window upstream and downstream of the gene body were used to demarcate regulatory boundaries. Statistically significant TFBS from these regions were identified by FIMO and TEPIC based TF-TG affinity scores were calculated. PANDA GRNs were then generated using weighted and unweighted adjacency matrices. PPI data from BioGRID corresponding to TFs for each cell lines and cell line specific co-expression were obtained from GEUVADIS(GM12878) and ENCODE(K562). Elastic Net(ENET)-based regularized regression models were built from the resulting input features to predict log FPKM values(gene expression) of independent datasets for the two cell lines.

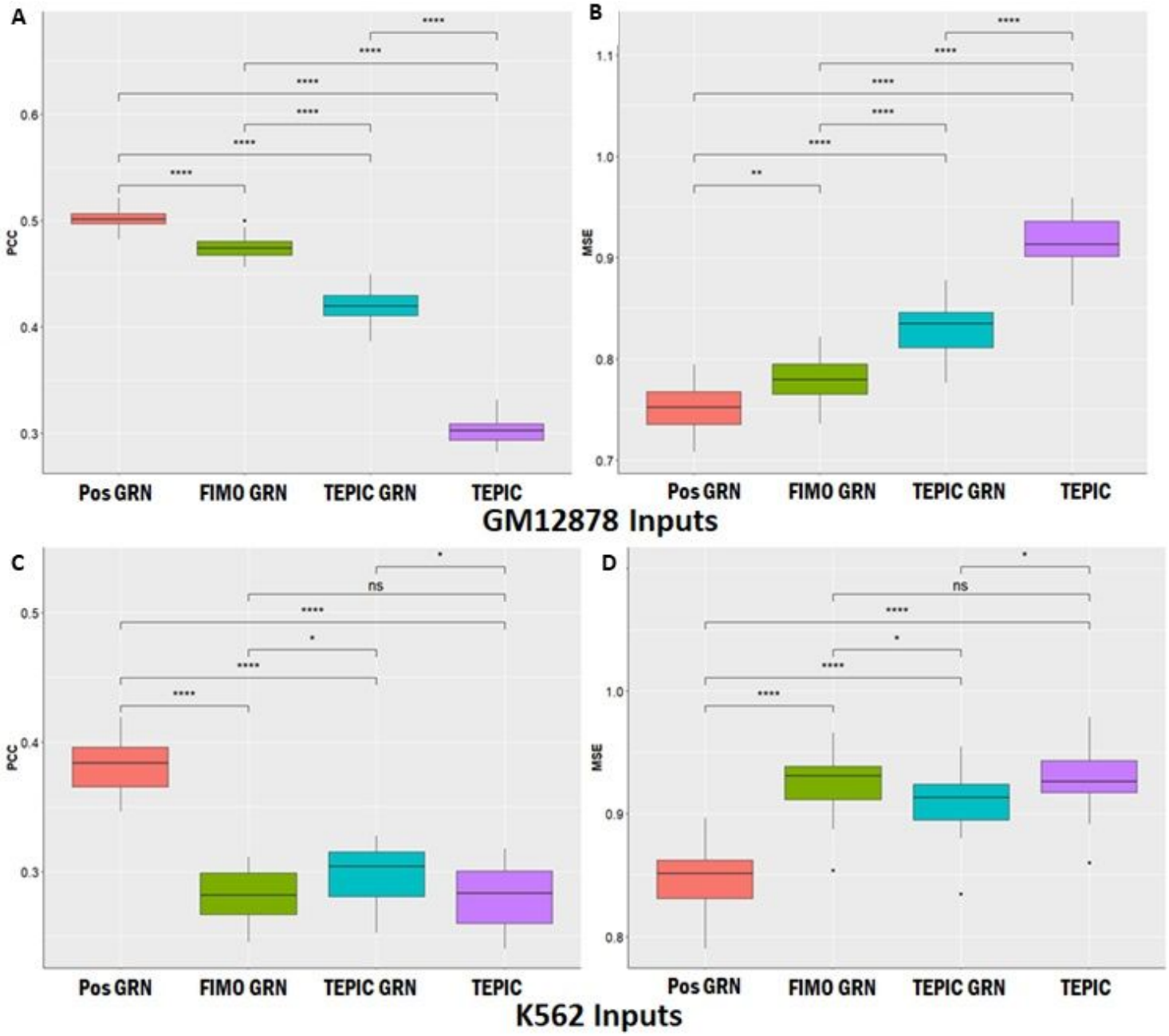

Figure 2

GRN based prediction models outperform those built using TEPIC affinity scores. A and B correspond to prediction performance for 20 random sets of 1729 GM12878 TGs while C and D were obtained from 1892 K562 TGs. Prediction performances for models corresponding to different inputs were compared using Wilcoxon rank sum test (***-p < 0.0001,**-pvalue < 0.001, *-pvalue<0.05, ns-not significant)

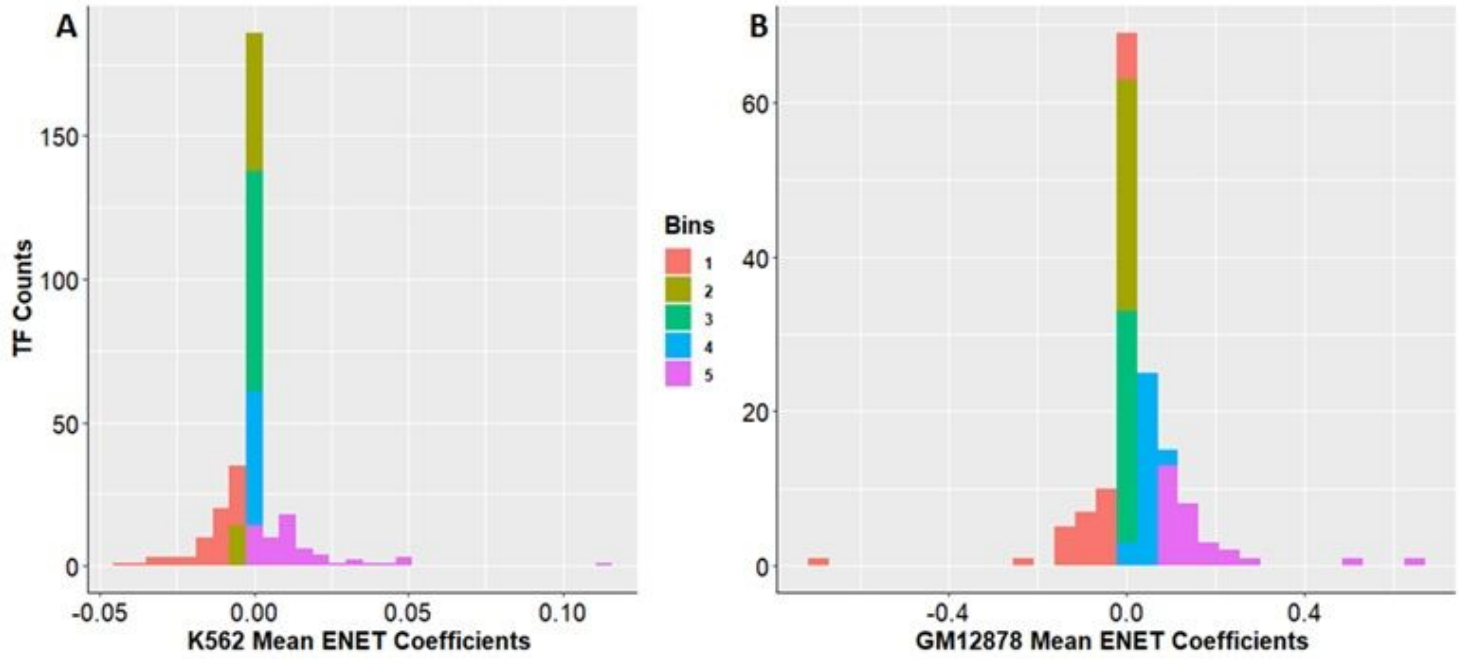

Figure 3

Mean ENET effect estimates reflect the important functional roles of various TFs. Histograms of the average effect estimates for calculated for A) 309 K562 TFs and B) 149 GM12878 TFs 3 using the “Pos GRN” ENET models. We also created 5 bins(quintiles) based on the effect estimates, which are color coded in the histogram.

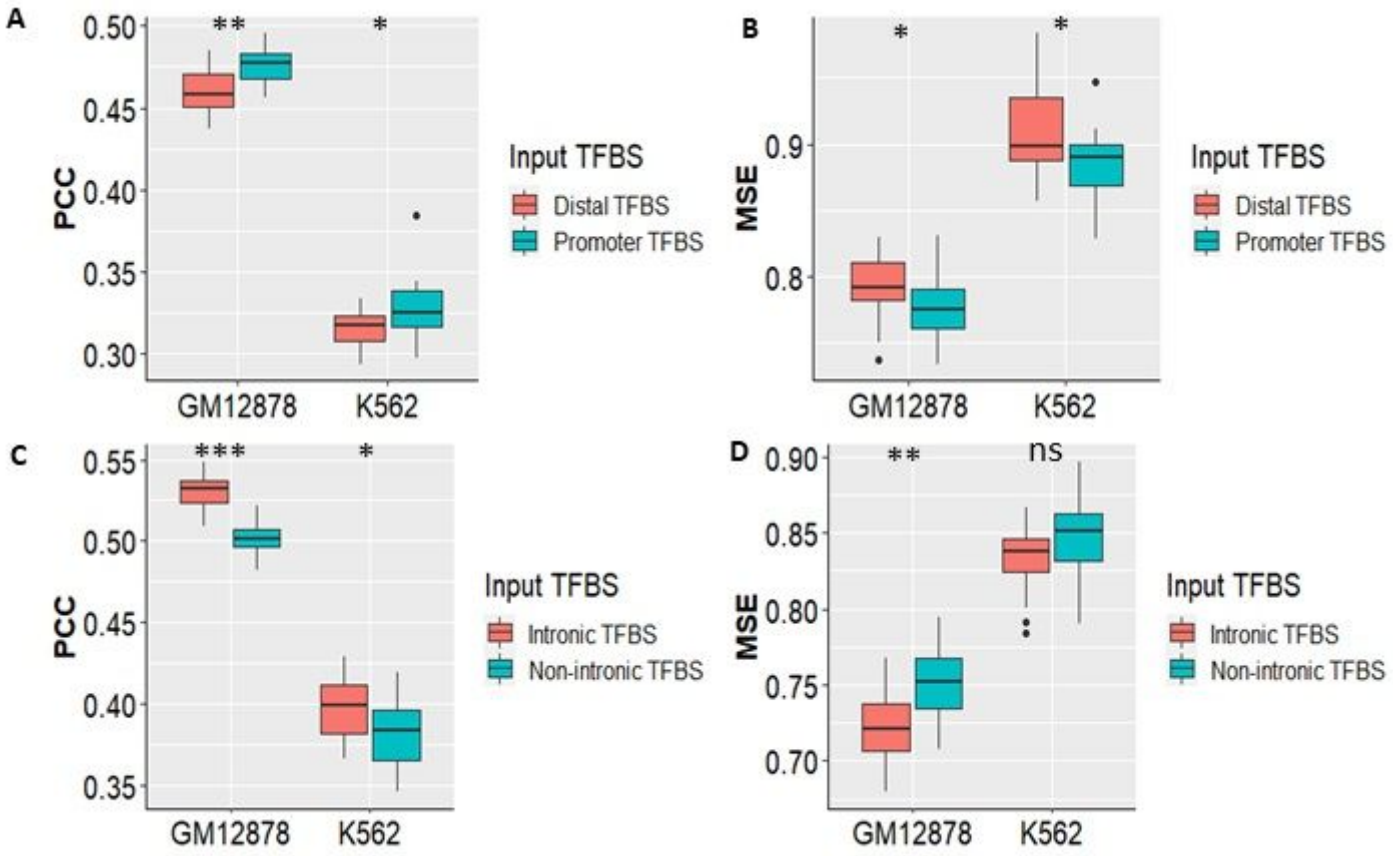

Figure 4

Intronic and Promoter TFBS are important for predicting gene expression. A)PCC and B) MSE obtained from the expression prediction of GM12878 and K56 TGs using models built from GRNs containing promoter and distal TFBS. C) PCC and D) MSE produced by models predicting expression for GM12878 and K562 TGs built using GRNs containing intronic TFBS vs. those built without them. The non-intronic TFBS input weights were derived from Pos GRN for both cell types.

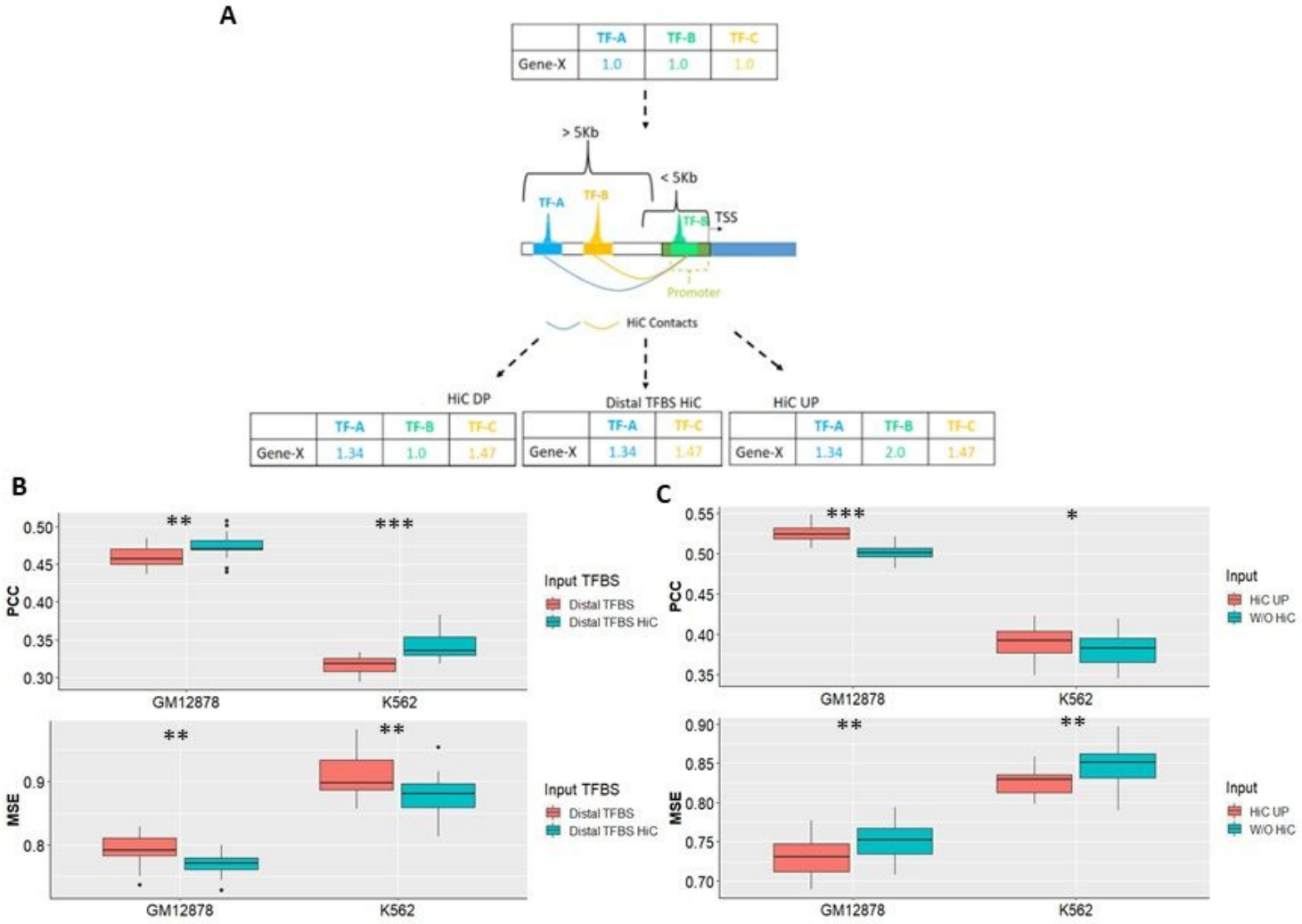

Figure 5

HiC data is capable of capturing the effect of long distance interactions between TF binding within distal TFBS and gene's promoter on gene expression. A) We used the cell-line specific Hi-C data to weight the distal TF-TG interactions in the motif adjacency matrix. We also down-weighted or up-weighted the interactions with the promoter TFs which would have been missed otherwise due to the low resolution nature of Hi-C data. B) We predicted expression of GM12878 and K562 TGs using distal TFBS based GRNs with and without HiC data integration in order to evaluate its predictive value for the models. C shows the predictive performance of the models using GRNs containing HiC normalized motif edges

based on the Hi-C UP weighting scheme compared to those built using unweighted binary motif network without HiC information.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.xlsx
- TFGRNRevisedSupplementaryBMC.docx
- TFGRNRevisedSupplementaryBMC.docx
- TFGRNRevisedSupplementaryBMC.docx
- S1.xlsx
- S1.xlsx
- S4.xlsx
- S4.xlsx
- S5.xlsx
- S5.xlsx
- S6.xlsx
- S6.xlsx
- S8.xlsx
- S8.xlsx