CHARACTERIZING GLOBAL REGULATORY PATTERNS OF TRANSCRIPTION FACTORS ON SYSTEMS-WIDE SCALE USING MULTI-OMICS DATASETS AND MACHINE LEARNING

by

NEEL PATEL

Submitted in partial fulfillment for the requirements for the degree of

Doctor of Philosophy

Systems Biology and Bioinformatics Program

CASE WESTERN RESERVE UNIVERSITY

August, 2021

CASE WESTERN RESERVE UNIVERSITY
SCHOOL OF GRADUATE STUDIES

We hereby approve the thesis/dissertation of

NEEL PATEL

candidate for the degree of Doctor of Philosophy

Committee Chair

DAVID LODOWSKI

Committee Members

WILLIAM S. BUSH

MARK CAMERON

RONG XU

Date of Defense

July 2, 2021

We also certify that written approval has been obtained for any proprietary material contained therein.

TABLE OF CONTENTS

LIST OF FIGURES

Overview

Chapter 2

Chapter 3

Chapter 4

Chapter 5

LIST OF TABLES

Chapter 2

Chapter 4

Chapter 5

Characterizing Global Regulatory Patterns of Transcription Factors on a Systems-Wide Scale Using Multi-Omics Datasets and Machine Learning

Abstract

By

NEEL PATEL

Transcription factors(TFs) are specialized DNA binding proteins, that regulate target gene (TG) expression by driving the process of transcription. Disruption of TF binding sites can cause significant changes in TG expression, which has been shown to be associated with several diseases. Moreover, besides binding of TFs, which is a local/*cis* regulatory mechanism, *trans*-acting mechanisms such as cooperativity among different combinations of TFs and co-regulation of multiple TGs by the same set of TFs can also influence TG expression. Integrative approaches that can incorporate information from these mechanisms, obtained from different data sources, are needed to comprehend TF based TG expression regulation on a systems-wide level. Furthermore, there is also a need to integrate this regulatory information in tests for statistical associations of genetic variants with complex disease traits to unravel mechanisms responsible for causing these diseases, while also discovering novel risk TGs.

In this dissertation, I develop an integrative gene regulatory network based approach utilizing information from different *cis* and *trans* regulatory mechanisms to model TG expression using machine learning algorithms. Furthermore, I use these models to calculate effect estimates of individual TFs as well as of combinations of TFs forming TF regulatory modules. Lastly, I build neural networks to quantify influence of non-coding

variants on TF binding and integrate them with effect estimates of the TFs in order to derive their impact on TG regulation. I utilize these aggregated scores, as weights for common variants (allele frequency > 5%), to build TG expression prediction models based on individual level genotype information to perform transcriptome wide association study(TWAS) within my novel framework TFXcan. I show that such models are more accurate compared to state-of-the-art TWAS models using broad epigenetic priors as variant weights. Furthermore, I describe a novel weighted kernel association test TFKin, which uses kinship matrix computed for individuals based on TF regulatory scores of rare variants. I show that this kind of a weighting approach, is better at TG-expression association, compared to conventional allele frequency derived weights. Both TFXcan and TFKin can be utilized to derive influence of common and rare variants on TF based TG expression regulation.

OVERVIEW

The central dogma of life, which involves replication of DNA, transcription of DNA into RNA and translation of RNA into protein product, is essential for the existence of almost all organisms on this planet. About a decade ago, researchers mainly studied these processes in context of cell survival, proliferation and death as it pertained to a small set of genes within the cell. Recently however, availability of big "omics" data sources corresponding to the central dogma processes has provided researchers with the opportunity for studying them in context of organismal level physiological traits such as height, body mass index(BMI), cholesterol levels etc. as shown in **Figure 1.1**. Such studies have been immensely helpful in identification of mechanisms behind the occurrence of such traits. The big "omics" data sources mainly include: genomics for studying DNA related processes such as the impact of genetic variation; transcriptomics for analyzing transcription regulation, mRNA quantification, and characterization; and lastly proteomics for looking at the process of translation, protein quantification and analysis of protein function. Additionally, epigenomics involving study of how chemical modifications of, and protein binding to, the DNA leads to systems-wide gene regulation, and metabolomics which is the study of small molecules(metabolites) and their associations with cellular metabolism are also part of the big "omics" data sources essential for analysis and characterization of systems-level physiological traits. Furthermore, disruptions in the central dogma processes such as genetic variants changing the DNA constitution, transcriptional dysregulation leading to altered mRNA levels and protein dysfunction

causing changes in several metabolic processes have been known to cause serious conditions such as different types of cancer, several immunological, neurological and developmental disorders. These diseases have been extensively studied in context of individual disruptions by using data corresponding to the respective central dogma process (see **Figure 1.1**). For instance, researchers have leveraged genomics data sources such as whole genome genotyping data for individuals to find associations with several complex disease traits(traits with multiple risk genes/variants associated with them). The diseases have also been studied with regards to transcription dysregulation and protein dysfunction utilizing transcriptomics and proteomics data sources respectively. However, a more comprehensive understanding of the mechanistic underpinnings of these diseases requires an integrative approach capable of leveraging data from multiple big omics data sources. Unfortunately, in the field of computational biology, there is a serious lack of such integrative approaches.

In this dissertation, I will describe novel methodologies that mainly utilize genomics, epigenomics, transcriptomics and proteomics datasets to derive the relationship between genetic variants and transcription dysregulation as well as to characterize their associations with complex disease traits. The central theme/hypothesis of the dissertation will be that utilizing multi-omics datasets to characterize disruptions in the central dogma

processes will lead to a better understanding of disease mechanisms compared to utilizing

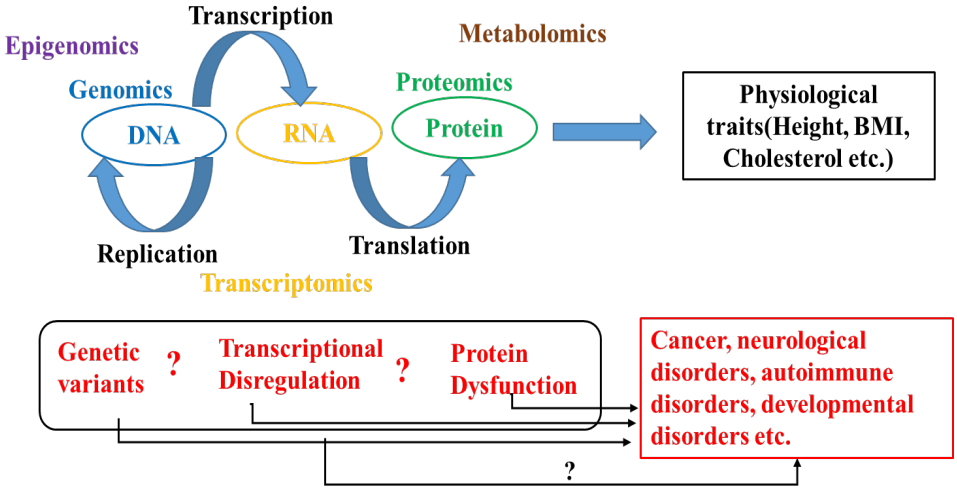

Figure 1.1: *Era of big omics datasets and the need for integrative approaches.Schematic diagram shows the big omics datasets such as genomics, transcriptomics,epigenomics, proteomics and metabolomics and their association with processes of the central dogma of life. These big omics datasets help in comprehending several systems-level physiological traits such as height, BMI and cholesterol levels. Additionally, the disruptions in these processes(shown in red) in the central processes have been known to cause severe diseases shown in red and their associations with these diseases have been studied individually. However, as shown in the figure, relationships among these disruptions and their combined associations with disease occurrence require integrative approaches.*

information from single omics data sources.

At the foundation of this work are transcription factors(TFs), specialized DNA binding proteins that regulate target gene (TG) expression by driving the process of transcription. These proteins recognize and bind to specific regulatory elements located proximally(promoters) or distally(enhancers) to a TG's transcription start site(TSS) in its *cis*-regulatory region. Proximal regulatory elements contain binding sites for pivotal TFs responsible for transcription initiation, while distal regulatory elements are mainly occupied by auxiliary TFs that interact with TG promoters via chromatin looping to enhance or repress the rate of transcription. Disruption of TF binding sites(TFBS) can cause significant changes in TG expression, which has been shown to be associated with several diseases. Moreover, besides binding of TFs, which is a local/*cis* regulatory mechanism, there are other ways in which TFs can regulate TG expression from a greater

genomic distance. These *trans*-acting mechanisms include co-operativity among different combinations of TFs and co-regulation of multiple TGs by the same set of TFs. Integrative approaches that can incorporate information for these *cis* and *trans* acting mechanisms, obtained from different data sources, are needed to comprehend TF based TG expression regulation on a systems-wide level. Furthermore, there is also a need to integrate this regulatory information in tests for statistical associations of genetic variants with complex disease traits. By doing so, one may be able to unravel mechanisms responsible for causing these diseases, while also discovering novel risk TGs.

I first develop an integrative gene regulatory network(GRN) based approach utilizing information from different *cis* and *trans* regulatory mechanisms to model TG expression using machine learning algorithms. I also incorporate information corresponding to distal regulatory elements interacting with TG promoters via long distance chromatin looping in these models. I illustrate that such an integrative modelling approach leads to more accurate TG expression prediction compared to models using single-source mechanism information. Furthermore, I use these models to calculate effect estimates of individual TFs as well as of combinations of TFs forming TF regulatory modules(TRMs). I use linear regularized regression models to compute the average linear influence of individual TFs, on TG expression, which coincides very well with their activating and repressing functional roles. On the other hand, the non-linear effects of TRMs on TG expression, calculated using neural network models, led me to the discovery of many novel TF interactions that mostly occur over long distances via chromatin looping. I also characterize the influence of different regulatory elements on TG expression using my modelling approach, which helped me in recapitulating the established role of

promoters, and in discovering novel role of introns, in transcriptional regulation. Lastly, I built complex neural network models to quantify influence of non-coding *cis*-genetic variants on TF binding. Subsequently, I integrate the scores obtained from these models with previously computed average effect estimates of the TFs in order to derive the impact of the *cis*-variants on TF based TG expression regulation. I further utilize these aggregated scores, as weights for common variants (allele frequency > 5%), to build TG expression prediction models based on individual level genotype information to perform transcriptome wide association study(TWAS) within my novel framework TFXcan. I show that such models are more accurate compared to the state-of-the-art TWAS models using broad epigenetic priors as variant weights. Furthermore, I describe a novel weighted kernel association test TFKin, which uses kinship matrix computed for individuals based on TF regulatory scores of *cis*-rare variants. I show that this kind of a weighting approach, is better at TG-expression association, compared to conventional allele frequency derived weights, for both discovery and replication analyses. Both TFXcan and TFKin can be utilized to derive influence of common and rare variants on TF based TG expression regulation. Additionally, one can use these methodologies to characterize regulatory mechanisms and identify risk TGs associated with heritable complex disease traits where significant amount of variability could be explained by genetic variants.

CHAPTER 1: INTRODUCTION

1A: Transcription factors and their regulatory mechanisms

Transcription is a process of converting a segment of DNA into an mRNA transcript, decoding the message residing within the genome of organisms of producing functional copies of genes[1]. This process is mainly driven by specialized proteins called transcription factors(TFs) that recognize and bind to specific DNA sequences called regulatory elements within the *cis*-regulatory region of a target gene(TG)[1]. A *cis*-regulatory region, in the context of TF activity, is roughly defined as a small(~50Kbp) region around the gene body where most of the TF based regulation takes place[1]. TFs possess DNA-binding domains(DBD), which help recognition of the regulatory elements; approximately 1600 proteins with DBD have been identified, out of which about 2/3rd have been characterized with regard to their functional roles[1]. The regulatory elements containing TF binding sites(TFBS) are located proximally as well as distally to the transcription initiation machinery. Proximal regulatory elements mainly consist of TG promoters, which are regulatory regions with a maximum length of about 1000bp present upstream of the TG coding region[2]. Promoters contain TFBS for the most pivotal TFs, also known as pioneer factors, such as TATA-binding protein(TBP) and TBP-associated factors(TAFs)[2]. These TFs are responsible for recruiting RNA polymerase II and assembling the pre-initiation complex(PIC) for beginning the process of transcription[2]. On the other hand, distal regulatory elements contain TFBS for auxiliary TFs, also known as co-factors, that interact with the promoter binding pioneer TFs. Such co-factors are versatile and can enhance or repress TG expression based on the TFs with which they interact[3].

The binding preference of a TF, also known as a motif, is generally described graphically in the form of a sequence logo which contains the four nucleotide bases("A", "T", "C", "G") displayed in different sizes, often stacked on top of each other, for the length of the motif [1]. The sizes of the nucleotide bases are correlated with the probability of these bases occurring at each given position of the motif [1]. These probabilities are in turn derived from a position weight matrix(PWM) containing tabulated weights for each of the four bases for each position of the motif [1]. TF binding affinity for a given regulatory region is most commonly calculated by scanning the DNA sequence underneath the TFBS using the PWM specific to that TF. JASPAR is a large publicly accessible database that contains PWMs for many TFs across different organisms[4]. The PWMs themselves are derived by aggregating information imparted by DNA sequences bound by the TFs utilizing different in vitro and in vivo experimental techniques. Chromatin immunoprecipitation sequencing(ChIP-Seq) is the gold standard method for deriving TF motif information, which is based on assaying TFBS across the whole genome by first pulling down the TF of interest using antibodies followed by sequencing of the DNA fragments bound to them[5]. ChIP-seq data has been collected for several TFs across different tissues and cell types and is publicly accessible through large scale databases such as encyclopedia of DNA regulatory elements(ENCODE)[6].

ChIP-Seq data has been extremely useful for defining motif preferences for a given TF as well as for finding its binding sites on a genome-wide level. However, one of the major drawbacks of ChIP-Seq is that it is extremely difficult and very expensive to perform. Hence, researchers have made an effort to utilize data from assays such as DNAse-Seq and ATAC-Seq, which provide information regarding open chromatin regions

across the genome which are accessible to TFs, in order to predict TFBS[7]. In eukaryotes, the regulatory elements of most TGs, except for the ones involved in housekeeping, are present in condensed chromatin form which prevents TFs from accessing them[2]. This condensation is caused by DNA being wrapped around an octameric complex of histone proteins forming a protein-DNA complex called nucleosomes[7]. Chemical modifications of these histone proteins, in the form of acetylation, methylation, phosphorylation and ubiquitination, leads to slight disintegration of the nucleosome complex which ultimately results in relaxed open chromatin regions[7]. Since histones are also DNA binding proteins, various histone modifications or "marks" across the genome have been studied extensively utilizing ChIP-Seq assays. Based on these studies, it has been determined that specific histone modifications can significantly relax or condense the chromatin leading to increased or decreased binding of the TFs to the regulatory elements. Thus, both motif and nucleosome composition within the *cis*-regulatory elements of a TG affect binding of TFs and ultimately its expression and are considered *cis*-acting regulatory mechanisms.

Apart from the *cis*-regulatory mechanisms mentioned above, there are also other *trans* acting mechanisms that significantly impact TG expression. For instance, different combinations of TFs cooperate among themselves forming regulatory complexes which are essential for TG expression regulation[1]. As a matter of fact, almost all TFs require such complexes to exert their influence over TG expression regulation in eukaryotes[2]. Information regarding cooperative interactions among TFs, which are mainly studied from the perspective of protein-protein interactions(PPI), can be obtained by several in vitro and in vivo techniques[8]. Large-scale databases such as BioGRID[9], STRING[10] and IntAct[11] contain curated PPI information obtained from such techniques. Additionally, a given set

of TFs can co-regulate expression of multiple TGs, which are functionally related to each other[12,13]. Co-regulation is an essential TF based regulatory mechanism that leads to a concerted control of a set of TGs whose protein products may participate in a pathway or in forming a complex. Usually, TGs that are co-regulated are also co-expressed which means that their expression patterns are highly correlated[12]. Moreover, one can also determine the set of TFs that control the expression of these TGs by analyzing their expression patterns and correlating them with that of the TGs, since TFs co-regulating a set of TGs will be expressed in a similar direction to those TGs. Thus, expression data is the main source for studying TF driven TG co-regulation. ENCODE and Gene Expression Omnibus(GEO)[14] contain data collected from a large number of expression studies based on micro-array and bulk RNA-sequencing experiments. Lastly, chromatin confirmation changes mentioned above, in the context of changing accessibility to the TG regulatory elements, can also regulate expression based on another mechanism. TFs binding distal regulatory elements (enhancers and repressors) are brought in contact with the promoter binding PIC via chromatin looping [15,16]. These interactions are essential for maintaining constant contact of the PIC and the pioneer factors with the TG promoters enhancing the rate of transcription[15]. On the other hand, chromatin looping mediated interactions between distally binding TFs and TG promoter could also lead to recruitment of repressive factors ultimately silencing its expression[7]. Thus, chromatin looping also plays an essential role in regulating TG expression. High-throughput chromatin capture (Hi-C) is an experimental technique that aggregates contacts occurring between distal regions of the genome, which are assumed to be caused by 3D chromatin confirmation changes[17]. One can utilize Hi-C

data, stored in GEO and ENCODE, to infer chromatin looping contacts occurring between distal regions of the genome.

1B: Current computational modelling methods for modelling the impact of TFs on gene expression and their limitations

All the cis and trans acting mechanisms described in **1A**, drive TG expression regulation via a complex regulatory program involving some or all of them working in concert. Thus, studying eukaryotic transcriptional regulation is a very difficult task experimentally on a systems-wide scale. Computational approaches have helped alleviate this issue to a certain extent by utilizing information from the regulatory mechanisms obtained from the publicly available databases described in **1A** . Such computational approaches build TG expression prediction models using different modelling algorithms based on the information obtained from the regulatory mechanisms. In this prediction framework, input features consist of the quantified information corresponding to the regulatory mechanism (TF binding, histone modifications etc.) and the output is TG expression obtained from a microarray or an RNA-Seq experiment[18–21]. The output is then divided into training and test set, where the former set of expression values is used to train the models and tune their hyper-parameters and the latter set is used to assess the accuracy of the trained models. After training the prediction models, one can explicitly quantify the effect that the regulatory mechanism has on TG expression regulation by computing the correlation between the observed and the predicted expression. If this correlation is really high, then the regulatory mechanism is assumed to have a high impact on TG expression[18]. One can also derive biologically relevant information from these prediction models that could pertain to specific TFs having a higher regulatory potential.

Computational approaches using the prediction framework described above have mostly utilized cis-regulatory mechanism information, such as TF binding affinity or histone modifications, in their models. 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"[21]. Cheng et al.[20] and Zhang et al.[22] 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; using affinity scores as input features, they used regularized linear regression models to predict gene expression[19]. More recently, deep learning models have become popular for this task, although inferring biologically relevant information from these complex models has remained a challenge[23,24].

The abovementioned approaches have mainly studied impact of individual TFs on TG expression. However, as mentioned in **1A** a significant amount of TG expression regulation is carried out by interactions among different combinations of TFs forming TF regulatory modules(TRMs)[1]. These TRMs influence TG expression both additively and non-additively as seen in model organisms[25,26]. Additionally, the interactions among TFs, which are the basis for the formation of these TRMs have mostly been studied by detecting proximally binding co-localizing sets of TFs using ChIP-Seq data. Gerstein et al. analyzed the co-localization maps of different TFs in K562 and GM12878 cell lines to detect significantly co-associating TFs using a discriminative machine learning approach[27]. They

detected several well-characterized TF interactions such as the GATA1-complex(GATA1-GATA2-TAL1), MYC complex(MYC-MAX-E2F6) and the AP1-factors (FOS-JUN-JUND-FOSL) as well as some novel TF interactions such as GATA1-CCNT2–HMGN3 and GATA1-NRSF-REST using their approach[27]. Others have used non-parametric modeling approaches to identify pairwise or higher-order interactions of TFs. For example, Guo and Gifford developed a topic modeling approach called Regulatory Motif Discovery(RMD) that identifies different TF interactions utilizing TF co-localization information[28]. They detected multiple well known TF interactions such as the cohesin complex (CTCF-RAD21-SMC3) complex, the transcription pre-initiation complex (POL2-TBP-TAF1) and the AP1 factor complex[28]. Bailey et. al. identified several literature-annotated interactions by identifying closely binding TFs based on significant spacings between their sequence motifs[29]. Lastly, soft and hard clustering methods such as k-means clustering, non-negative matrix factorization and self-organizing maps have also been used to identify co-localizing TFs across the genome[30–32].

1C: Analyses of transcription factor binding site altering non-coding genetic variants

As described in **1A**, TG expression regulation in eukaryotes is driven by a complex regulatory program based on several TF-based regulatory mechanisms. As a result, disruption in these mechanisms can result in drastic changes in TG expression which could ultimately lead of occurrence of diseases in humans[33]. Since TF binding is the most influential TG expression regulatory mechanism, many commonly occurring human diseases have been known to be caused by changes in TFBS caused by presence of genetic variants[34,35]. Furthermore, mutations in the TFs themselves, which are caused by variants in the coding regions, have also been associated with several diseases[33,36]. However, in this

project, I will focus on TFBS altering variants only, which fall in the broad category of non-coding/regulatory variants as they occur outside of the protein coding regions of the genome. Despite their association with several serious diseases, characterizing the functional role of such non-coding variants has remained a challenge[37].

TFBS across the genome could help in delineating the functional role of non-coding variants, as TFs are effectors in the process of transcriptional regulation. Thus, by characterizing the impact of the non-coding variants on the TFBS, one can implicitly derive its influence on TG expression regulation and can gain a better mechanistic understanding of the diseases associated with these variants. To that end, several studies have tried to quantify impact of genetic variants on TF binding using both experimental and computational approaches. Experimental techniques used to derive influence of non-coding variants on TFBS are based on electrophoretic mobility shift assays(EMSA), differential ChIP-seq of the regions containing the reference allele and the alternate allele and enhanced yeast-one hybrid assays[38]. These methods, although accurate, are low-throughput, time consuming and expensive. Thus, computational approaches taking advantage of large-scale databases containing TF binding and motif information described in **1B** to predict variant impact over TFBS have been used widely in recent times. Such predictive algorithms mostly use TF binding preference, in the form of PWMs, to predict the probability of the TF binding a given DNA sequence containing the alternate allele corresponding to the non-coding variant. The prediction algorithms either use pre-existing sets of PWMs or train models using ChIP-seq data to learn de novo PWMs. Methods such as FIMO(Find Individual Motif Occurrence)[39], RSAT(Regulatory Sequence Analysis Tools)[40], Clover[41] and QBiC-PRED[42] score DNA sequences containing reference and

alternate alleles based on the modifications to the TF motif. However, since these methods must be pre-trained on a set of variants, they cannot be used to annotate novel TFBS altering variants. In order to overcome this limitation, modern TFBS variant annotation algorithms have started using deep learning neural networks within the prediction framework. These complex neural networks consist of multiple layers of computational neurons producing non-linear outputs based on inputs received from the previous layers. A type of neural network called convolutional neural network(CNN) has been used widely in algorithms such as DeepSEA[43], DeepBIND[44], DANQ[45], and FactorNet[46] to predict impact of regulatory variants on TFBS. CNNs process input DNA sequences using kernel filters, to scan for TF motifs, to predict the probability of a TF binding them. These algorithms are trained in a supervised fashion using hundreds of thousands of DNA sequences obtained from DNAse-Seq and ChIP-seq datasets obtained from ENCODE. Furthermore, these CNN based algorithms have been shown to outperform conventional approaches for predicting TFBS because of their inherent capability to automatically extract abstract features from the DNA sequences, in addition to TF motif, that can affect TF binding. Furthermore, these algorithms can efficiently quantify the effect of changes within DNA sequences, based on the presence of a genetic variant, without being explicitly trained on a pre-defined set of TFBS altering variants. Due to this property, CNN based methods can also be used to annotate *de novo* and novel TFBS altering variants.

Non-coding genetic variants can be generally classified into two types, based on their minor allele frequency(MAF) in a population: common variants (MAF > 5%) and rare variants(MAF > 5%). Single variants association tests, based on linear (continuous outcome) and logistic (binary outcome) regression models, are generally used to analyze

common variants. Furthermore, non-coding common variants significantly associated with TG expression from single variants association tests are called expression quantitative trait loci(eQTLs)[47]. While there have been extensive studies, with regard to eQTL discovery, functional fine mapping of these variants is still lagging[47]. Transcriptome wide association studies(TWAS) attempt to overcome this limitation by predicting TG expression using information from the genotypes of eQTLs[48] and using it for gene based association testing. Reference datasets containing genotype and TG expression information, such as GTEx[49], are used to build such prediction models. The information learned from these models is stored in the form of weights/effect sizes, which correspond to the magnitude of the influence that the eQTLs have on TG expression. The weights are then used to predict TG expression in an independent dataset containing genotype and phenotype information. By performing gene based association testing using predicted TG expression and phenotype data, one can identify possibly TG targets for a given disease while overcoming the disadvantages of single variant association studies[48]. Several TWAS approaches have been developed to date that differ in the type of models as well as the kind of information used to predict TG expression. For instance, PrediXcan[50] and EpiXcan[51] use ENET regularized regression to build the prediction models, while the latter also uses epigenetic priors corresponding to different chromatin states/regulatory elements containing eQTLs in the models. On the other hand, FUSION[52] uses eQTL information along with their effect sizes to predict TG expression using BLUP and BSLMM models. FUSION, along with another method S-PrediXcan[53], also have the capability of the predicting TG expression using summary-level data. Among these methodologies, only EpiXcan has the capability of

including functional annotations for eQTLs in the TWAS models, which has been shown to result in more accurate TG expression models.

Because of their low population frequencies, single variant association tests, normally used for the common variants, cannot be applied to rare variants as they are underpowered to detect any association[54]. Thus, collapsing tests using a set of rare variants within a pre-defined unit/region of the genome(e.g. gene) combined either in a form of a kernel/kinship matrix or as a single combined score aggregated for all the rare variants are used.[55] The former type of test, uses variance component statistic to perform rare variants association analysis e.g. sequence kernel association test(SKAT)[56]. On the other hand, burden test is used to find associations between the combined scores of a group of rare variants and a given trait[57]. While burden tests make the assumption that all the variants in a set have positive effect on the train, the variance component tests make no such assumption[55]. Additionally, combinative methods have been developed to efficiently combine the two types of tests[58–60]. These tests weight rare variants based on their MAF within the association tests, such that the rarer variants are assumed to have higher effect on the trait. Moreover, it has been shown that weighting rare variants based on their functional annotations helps improve the power of the set based association tests[57,61–64]. There have been several approaches developed for integrating functional scores into rare variants association test based on different types of models. The FunSPU method[64] uses adaptive sum of powered test, while STAAR[61], SMART[63] and FST[62] utilize linear mixed-models(LMM), to incorporate functional annotations for coding and non-coding variants, from multiple sources, into rare variants association tests. These methods, along with EpiXcan use broad functional annotations for rare and common variants in the association

tests and don't include any specific information regarding how they influence TF binding or the associated regulatory mechanisms.

1D: Overcoming the limitation of current methods by developing integrative approaches that can utilize big omics data sources to study transcription factor based gene regulation.

Despite availability of big "omics" datasets corresponding to different transcriptional regulatory mechanisms, there is a dearth of algorithms that can integrate them into a coherent framework. Such integrative approaches are essential to model influence of multiple regulatory mechanisms on TG expression, which could provide a more complete picture of TF driven TG regulation. Additionally, non-coding variants occurring within different regulatory elements could exert their influence over TG expression and ultimately over several diseases via modulating the TFBS. An integrative framework that involves weighting of variants based on their influence on TFBS aided by information derived from the other regulatory mechanism would lead to a better understanding of the disruption of these mechanisms causing different diseases. In this project, my primary goal is to build an integrative method to model impact of multiple TF based regulatory mechanisms, as well as of chromatin looping, on TG expression regulation. I developed such a framework utilizing gene regulatory networks as well as machine learning algorithms. Additionally, I used this framework to enhance annotations of non-coding variants influencing TF binding scored using complex deep learning based neural network models.

Previous computational approaches aimed at quantify influence of transcriptional regulatory mechanisms on TG expression have mainly focused on TF binding and histone

modifications. In **Aim 1** (see Chapter 2), I hypothesize that utilizing information derived from trans acting mechanisms such as TF cooperativity, TG co-regulation and chromatin looping based interactions between distal TFBS binding sites and TG promoters in addition to TF binding would lead to a better estimation of variance in TG expression compared to just using TF binding information. In this aim, I will develop a gene regulatory network(GRN) based prediction framework to predict TG expression utilizing input features derived from information corresponding to multiple regulatory mechanisms.

Interactions among TFs significantly influencing TG expression lead to formation of TF regulatory modules (TRMs), which are the cornerstone of transcriptional regulation in eukaryotes. However, previous computational approaches have only utilized TF peak co-localization information to define these interactions and also have not identified their influence on TG expression. In **Aim 2** (see Chapter 3), I hypothesize that utilizing information beyond TF co-localization within a TG expression prediction framework would lead to identification of TRMs based on TF interactions also occurring over long distances between distally binding TFs. I used the GRN based framework capturing influence of multiple TF based regulatory mechanisms on TG expression and complex non-linear neural network based multilayer perceptron(MLP) to predict TG expression. Additionally, I used the trained MLP models to compute interaction effects of different TF interactions and to detect TRM. Such a multi-omics TRM identification framework could provide a blueprint for researchers who want to study disruption in TRM formation leading to downstream changes in TG expression and occurrence of diseases.

Despite availability of several machine learning algorithms that can quantify the impact of non-coding genetic variants on TF binding, statistical tests for association of

common and rare variants that can integrate this information are virtually non-existent. In **Aim 3**, I hypothesize that including variant annotations, reflecting their influence on TF based TG expression, will lead to more accurate and more replicable TG associations with respect to their expression. In **Aim 3a**(see Chapter 4), I develop a TWAS framework to incorporate functional scores for common variants reflecting their influence on TF based TG expression regulation. To that end, I will develop a scoring algorithm that uses modern deep learning architecture to compute the effect of non-coding variants on TF binding affinity. I will integrate these scores with average influence of TF based regulatory mechanisms on TG expression and subsequently use them in a TWAS framework. Utilizing such as framework will aid in functional fine-mapping of eQTLs along with finding risk TGs associated with complex diseases. In **Aim 3b** (see Chapter 5), I will develop a weighted kernel based association test, containing regulatory scores for rare variants based on their influence on TF based TG expression, to analyze their regulatory potential. Even though set based rare variants annotation tests have the ability to integrate functional annotations from multiple sources, fine mapped TF based influence of these variants is still missing from them. In this aim, I will develop a kernel based rare variants association test, which uses kinship matrix computed based on similarity among individuals derived from the *cis*-regulatory potential of rare variants.

CHAPTER 2: MODELLING THE INFLUENCE OF CIS AND TRANS TRANSCRIPTION FACTOR BASED REGULATORY MECHANISMS ON GENE EXPRESSION USING GENE REGULATORY NETWORKS AND MACHINE LEARNING

In press as: N. Patel and W. S. Bush, "Modeling transcriptional regulation using gene regulatory networks based on multi-omics data sources," *BMC Bioinformatics*, vol. 22, no. 1, p. 200, 2021, doi: 10.1186/s12859-021-04126-3.

2.2A: Introduction

Dysregulation of transcription and gene expression has been linked to conditions such as diabetes[65], different subtypes of cancer[66] and neurological[67], autoimmune[68] and developmental disorders[33]. However, due to the complexity of the process of transcriptional regulation in eukaryotes, the mechanistic underpinnings of many of these diseases are yet unknown. Databases such as the Encyclopedia of DNA elements(ENCODE)[6], FANTOM5[69] and gene expression omnibus(GEO)[14] 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. Furthermore, with significant advancement in sequencing technology in the past decade, more and more genetic variants associated with the aforementioned disorders have been identified[70–74]. A majority of these variants are present within the transcriptional regulatory elements and TF binding sites(TFBS) [70–74]. However, despite the availability of the transcriptomic and genomic data, there is a dearth of integrative algorithms that consolidate these data types to better explain regulatory mechanisms of the aforementioned diseases. Such algorithms would first require generation of feature weights corresponding to transcriptional regulatory elements reflecting their influence over gene expression. Next, these weights would be used to annotate the regulatory variants, which could then be used downstream to perform weighted genetic association tests.

Current computational approaches, described in **1B** estimate regulatory feature weights by modelling gene expression utilizing information corresponding to cis/local regulatory mechanisms such as histone modification and TF binding strengths [18–22] . 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[75]. 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[76]. Published approaches, except for a recent extension of the TEPIC framework[77], 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[17]. 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[16].

In this chapter, I utilized multi-omics PANDA GRN based TF-TG features derived from multiple cis and trans acting transcriptional regulatory mechanisms to predict gene expression in GM12878 immortalized lymphoblastoid cell line as well as in K562 chronic myelogenous leukemia cell line. I further derived TF feature weights in form of

linear effect estimates from my learned models in order to characterize individual influence of various TFs on gene expression. In addition, I 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 interactions between TF binding distal regulatory elements and promoters on gene regulation by integrating Hi-C data into my GRNs and prediction models. Finally, in order to show the utility of my framework, I utilized the TF feature weights to perform weighted collapsing rare cis-regulatory variants based test using Depression Genes and Network(DGN) dataset for discovery and the Genotype Tissue and Expression(GTEx) dataset for replication. My 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. It also provides a blueprint for researchers of incorporating functional transcriptomic and genomic data in order to gain mechanistic understanding of diseases.

2.2B: Methods and Materials

2.2B1: Datasets used in this chapter

ChIP-Seq Files

In order to obtain information about the TFBS across genomes for GM12878 and K562, I used the Encyclopedia of DNA elements(ENCODE) database[6]. I downloaded processed ChIP-seq narrow peak bed files corresponding to 149, 382 and 234 TFs corresponding to GM12878, K562 and HepG2 cell lines respectively that were aligned with hg19/GRCh37 reference assembly of the human genome and that had passed the

optimal IDR(Irreproducible Discovery Rate) threshold as defined by the ENCODE consortium.

Gene Annotations

I used the GRCh37/hg19 reference genome build from the biomaRt library(*version .2.44.1*)[78] in R to derive gene annotations such as transcription start sites(TSS), length of the gene body, transcript ids, exon ids, transcript lengths, gene ids etc. for all the protein coding genes.

PWM Files

I downloaded the position weight matrix(PWM) files corresponding to 469 human TFs from the JASPAR database(*version..2020*)[4] in MEME format. I later used these files as inputs for running the FIMO algorithm [39]in order to find statistically TFBS across the genome.

TFBS sequences

I used the GRCh37/hg19 reference build to obtain sequences corresponding the transcription factor binding sites(TFBS) in the regulatory region of each gene. I later used these sequences as inputs along with the PWMs for running the FIMO algorithm[39] in order to find statistically TFBS across the genome.

Protein-Protein Interaction Data

I used the BioGrid database(*version.3.5.188*)[9] to download PPI data in order to build the PANDA GRNs. I only used the high confidence experimentally validated using experimental techniques such as co-fractionation, co-immunoprecipitation, yeast two-hybrid and affinity capture in BioGrid. I further filtered out the PPIs that did not contain

TFs, which ultimately provided us with 1937 PPIs among the GM12878 TFs, 3025 interactions among the K562 TFs and 2807 interactions among HepG2 TFs.

Co-expression data

The other source of information that I needed to build the PANDA GRNs was the co-expression matrix. I used different expression data sets in order to build these matrices for the two cell types.

For the GM12878 lymphoblastoid cell line, I used data from the GEUVADIS project[79], which contains lymphoblastoid RNA-seq and genotype data derived from individuals belonging to European and African ancestry groups who participated in the 1000 genomes project. I used the Log normalized expression values(log FPKM) for the 15,785 protein coding genes from the lymphoblastoid cells of 462 individuals in the GEUVADIS dataset with variant effects regressed out using mixed-linear models with a genome-wide genetic relationship matrix(GRM). My
models could be described using the equation below:

$$y = X\beta + Zu + \epsilon \qquad (2.1)$$

Here, y is the vector containing log FPKM expression values for the 462 individuals, X is the matrix of size 462 by N , where N represents the number of common variants (minor allele frequency > 0.05) present in the dataset (6,326,925), containing the additive genotypes for each variant for each individual, β is the vector of size N by 1 containing the effect estimates/coefficients of each variant obtained from the fitted regression models; Z is the GRM of size 462 by 462 built using the number of alleles shared by each pair of individuals at the loci representing all the 6,326,925 variants across the genome; u is the random effects vector of size 462 capturing the random variance for each individual

from the GRM and finally ϵ is the residual vector of size 462 containing the effects not explained by the model. After fitting the models across all the genes, I extracted the ϵ term for each gene which contained the residual expression values. I used these values for building the co-expression matrix.

For the K562 leukemia and the HepG2 hepatocellular carcinoma cell-lines, I downloaded expression data corresponding to four different experiments and five different experiments respectively.

K562 expression dataset consisted of 8 different samples while that for HepG2 consisted of 9 different samples. I used the normalized FPKM values corresponding to 12,209 and 13,390 protein coding genes for K562 and HepG2 cell-lines respectively to build the co-expression matrix.

2.2B2: Processing expression data for ENET prediction models

I downloaded RNA-seq data for GM12878(ENCSR889TRN), K562(ENCSR545DKY) and HepG2(ENCSR181ZGR) from the ENCODE database. Each one of these experiments contained processed TG quantification data for two technical replicates. I used the Log10 normalized mean FPKM values as outcome for the ENET prediction models.

2.2B3: Defining Transcription Factor Binding Sites

I used three methods to define the TFBS between the TFs and the TGs for both the cell types using ChIP-seq data and Ensembl gene annotations from GrCh37 human genome assembly:

1)Positional TFBS: I 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. I 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 **Figure 2.1**.[80].

2) FIMO TFBS: I applied the FIMO algorithm[39] from the latest release of the MEME-suite tools(*version.5.1.1*) on the "Positional TFBS" data to find statistically significant set of TFBS. I extracted genomic sequence underneath the TF peak corresponding to each TFBS and the JASPAR(*version.2020*) based TF position weight matrices(PWM) to find statistically significant TFBS at the p-value threshold of 0.01.

3)TEPIC TFBS: I downloaded the TEPIC software (https://github.com/SchulzLab/TEPIC) along with the position specific energy matrices(PSEMS) for all TFs[19]. I used these

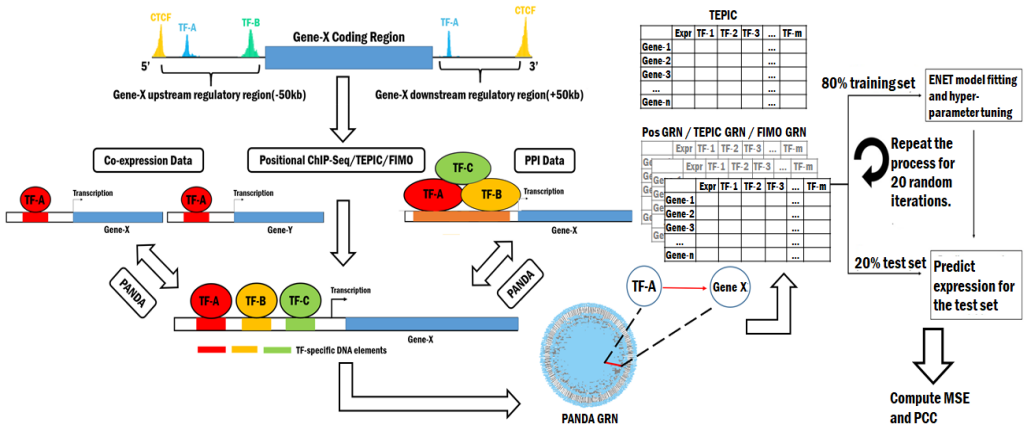

Figure 2. 1 *Workflow for building prediction models using multi-omics GRNs. ChIP-seq data for 153 TFs(GM12878) and 382 TFs(K562) having peaks passing the optimal irreproducible discovery rate(IDR) threshold defined by ENCODE were mapped to the regulatory region of each gene to define TFBS. The most 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.*

PSEMS, the Ensembl Homo_sapiens.GRCh37.87.gtf annotation, and I predefined

Positional TFBS to find affinity scores for TFs binding in the 50Kb window around each TG's TSS.

2.2B4: Generating Gene Regulatory Network Weightings

I converted the unique TF-TG interactions obtained from each TFBS identification method into weighted(TEPIC) and unweighted (Pos ChIP-Seq and FIMO) adjacency matrices. I used these matrices, along with BioGrid(*version.3.5.188*)[9], a method for defining protein-protein interactions (PPI), and cell-type specific co-expression networks to generate three different PANDA outputs. After 25 iterations, I obtained convergence by setting the threshold for Hamming's distance at 0.001 and by using the value of 0.1 for the update parameter for each GRN.

2.2B5: Generating training and test data sets for the prediction models

I used four different input datasets, for each cell type, for my prediction models based on PANDA GRN edgeweights ("Pos GRN", "FIMO GRN", "TEPIC GRN") and TEPIC affinity scores ("TEPIC") as shown in **Figure 2.1**. Using these matrices as inputs, I predicted the expression for independent datasets of GM12878(ENCSR889TRN) and K562(ENCSR545DKY) using the linear regularized elastic net(ENET) regression models. I used the python-based implementation of the ENET model from the scikit-learn library to build the prediction models, setting the value of α (the ratio between the lasso and ridge norms) at 0.5.

I used the log10-normalized FPKM(fragments per kilobase of transcripts per million) for TGs, that were common among different input matrices described in **Table 2.1** 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. I also applied my approach to 12,013 TGs for HepG2 for additional validation and generalization.

I split the input feature matrix and the output expression vector into 80% training data and 20% test data. I used the training data to train the ENET models, using 20-fold inner cross validation. I 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. I repeated this process for 20 iterations as shown in **Figure 2.1**.

2.2B6: Calculating TF average effect estimates

I calculated the average effect estimate for TF T $\bar{\beta}_T$ using the following equation:

$$\bar{\beta}_T = \frac{1}{|N|}\sum_{n \epsilon N} \beta_{T,n} \qquad (2.2)$$

Here, N is the set of random instances that I used to build my prediction models and $\beta_{T,n}$ is the effect estimate of T for instance n. I only used the GM12878 and K562 Pos GRN prediction models in order to calculate these estimates. I further divided the TFs based on these mean effect estimates using the xtile function of R(*version.3.4.2*) into 5 roughly equal bins.

2.2B7: Additional gene regulatory elements analyses

I 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 (**Figure 2.2A**). The number of corresponding TFBS and TF-TG interactions for each cell-type representing these regions is provided in **Table 2.1**. In order to get the intronic regions for each TG, I 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(**Figure 2.2B**). I 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. I used TF-TG interactions based on these

Table 2.1:*Number of TFs, TGs and TFBS obtained from different TFBS identification algorithms for GM12878, K562 and Hep2 cell lines. The "Pos ChIP-Seq" row contains TFBS identified by simply extracting the TF peaks in the cis regulatory regions around each gene, "FIMO" row contains statistically significant positional TFBS identified using the FIMO algorithm and the TEPIC row contains positional TFBS extracted based on the TEPIC affinity scores. The remaining rows contain the positional TFBS present within different regulatory elements utilized for the subsequent analyses in the paper. All the ChIP-seq data for the analysis was downloaded from the ENCODE database*

	GM12878				K562				HepG2			
	TFs	TGs	TFBS	Unique TF-TG Pairs	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	234	12,879	7,333,759	1,689,110
FIMO	85	16,850	2,444,195	714,167	110	18,173	7,349,429	1,138,823	104	12,867	3,665,921	838,999
TEPIC	80	11,784	-	517,226	86	10,239	-	880,554	73	12,841	-	937,393
Promoter	149	11,509	458,959	276,138	308	15,668	1,293,933	681,847	234	12,013	873,856	
Distal	149	16,964	3,750,174	1,128,079	309	18,152	10,320,315	2,312,490	234	12,013	6,256,152	1,128,079
Intronic	149	17,106	5,896,338	1,378,129	309	18,224	14,764,766	2,820,604	234	12,885	9,180,036	1,378,129

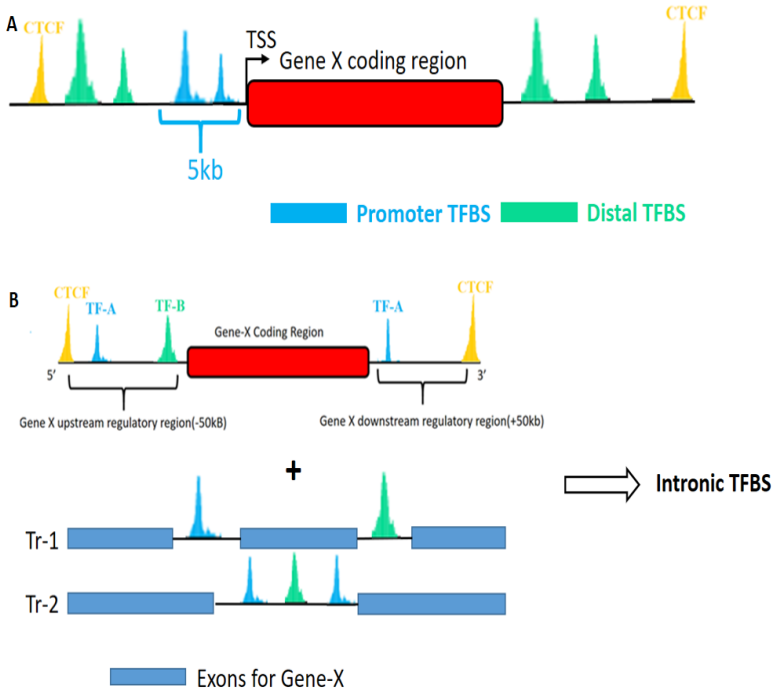

Figure 2.2: *Identifying TFBS in different regulatory elements. A) I used the 5Kb region upstream of the TSS for each gene to identify promoter TFBS. B) I extracted all the TFBS outside of the promoter region to isolate the distal TFBS. The number of promoter and distal TFBS for both cell-types have been provided in Table-2.1*

additional TFBS datasets to create motif-based adjacency matrices and used them to build additional PANDA GRNs, which I ultimately used to predict gene expression for TGs common between the models I was comparing.

2.2B8: Generating Hi-C Weightings

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

$$C_{i,g} = 1 + \text{scaled}\left(\frac{1}{N_{i,g}} \sum_{p \epsilon P_{i,g}} c_p\right) \tag{2.3}$$

Here, $C_{i,g}$ is the Hi-C adjusted edge weight between TF i and TG g , $N_{i,g}$ is the number of ChIP-seq peaks corresponding to i in the regulatory region of g, $P_{i,g}$ is the set of peaks corresponding to i in the regulatory region of g and c_p is the number of KR normalized contacts made by peak p with the promoter of g. I 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 $C_{i,g}$ would be equal to 1 and the maximum value for $C_{i,g}$ would be 1.99. I generated the cell type specific "Hi-C DP" motif adjacency matrix using these scaled interactions. I 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. I 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 2.1.**

2.2B9: QBiC-Pred-GRN rare variant association analysis

I followed the workflow shown in **Figure 2.3** for the rare variant analysis. I generated GM12878 GRN utilizing the intronic TFBS for motif network and HiC up weighting scheme described previously. I 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 (2.2). For the initial discovery analysis, I used the depression genes and networks(DGN) data set, which contains genotypes and RNA-seq data for 922 individuals of European descent[40].I further imputed variant genotypes using 1000 genomes reference panel and the University of Michigan imputation server[81,82]. I 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, I was able to find trained QBiC-Pred models for 59 TFs I scored these variants using the offline version of the QBiC-Pred software[42] which I downloaded from the github repository (https://github.com/vincentiusmartin/QBiC-Pred). I used the p-value threshold of 0.0001 to identify the variants significantly impacting the TFBS. I identified 118,789 rare variants that were present within their binding sites.

I 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.

$$Z_{v,t,g} = \bar{\beta}_t \times \frac{\sum_{p_{t,g} \in P_{T,g}} z_{v,p_{t,g}}}{|P_{t,g}|} \tag{2.4}$$

$$S_{v,g} = \frac{\sum_{t \in T_g} Z_{v,t,g}}{|T_g|} \tag{2.5}$$

Here, $z_{v,p_{t,g}}$ is the QBiC-Pred z-score for variant v significantly impacting the peak region(TFBS) $p_{t,g}$, which is a subset of all the peak regions $P_{t,g}$ belonging toTF t within the regulatory/intronic regions of TG g. $\bar{\beta}_t$ is the average ENET effect estimate obtained from the learned ENET models for TF t and $Z_{v,t,g}$ is the scaled QBiC-Pred z-score for variant v corresponding to TF t binding cis-regulatory/intronic regions for TG g. $S_{v,g}$ is the merge score for variant v for each TG g computed by averaging the scaled z-scores for all the TFs present within the cis-regulatory/intronic regions of TF g(T_g). I also computed aggregate QBiC-Pred z-scores for each variant present within all the TFBS for each TG g without utilizing the average effect estimates. In other words, I simply removed the effect estimate ($\bar{\beta}_t$) from the set of equations described above. I scaled both aggregated z-scores and merge scores within the range [-1,1] and used them for weighting the variants.

I used the R implementation of the SKAT algorithm [56](*version. 2.0.0*) in order to find association between these sets of variants and the TG expression levels normalized by HCP(hidden covariates prior). I 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, I utilized the Genotype-Tissue Expression(GTEx) dataset containing whole genome sequencing and RNA-seq data for 369 individuals[49] (**Figure 2.3**). I 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, I utilized the 65 covariates provided within the dataset to fit the SKAT model.

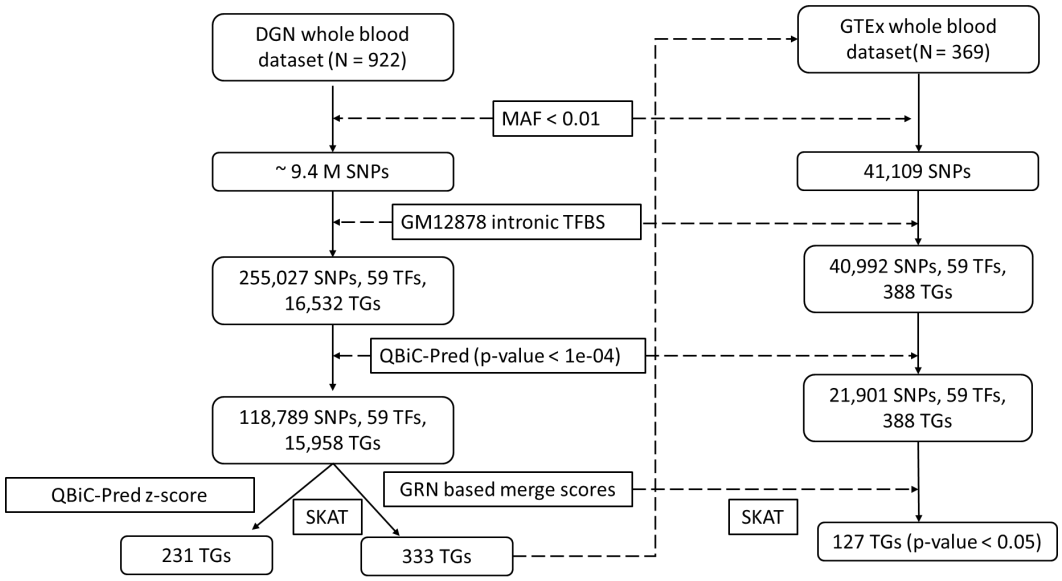

Figure 2.3: *Workflow for the QBiC-Pred based rare variant analysis. I used the DGN dataset for initial discovery analysis and the GTEx dataset for the replication analysis*.

2.2B10: Statistical Evaluations

I used R v.3.4.2 to perform all the statistical analyses in my study. Assuming a non-normal distribution of the PCC and MSE produced by the prediction models, I used the Wilcoxon signed-rank test to compare medians of these performance measures for different models. I used the gseapy package in python for Gene Ontology(GO) enrichment analyses. I 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. I specifically looked for significant enrichment terms (adjusted p-value < 0.05) for each bin for both the GO categories.

2.2C: Results

2.2C1: Accounting for trans acting mechanisms in addition to cis regulatory mechanisms improved gene expression prediction significantly

I built gene expression prediction models using TF-TG features derived from PANDA GRN edge-weights as well as from TEPIC affinity scores(**Figure 2.1)**. The PANDA GRNs were generated utilizing the information from cis and trans TF based regulatory mechanisms, while the TEPIC affinity scores only captured the influence of TF binding on TG regulation, a cis/local regulatory mechanism. I hypothesized that accounting for trans-acting mechanisms in addition to the cis acting ones would improve model accuracy.

PANDA algorithm needs three separate networks as inputs for GRN generation: motif network, PPI network and co-expression network. I have provided the details on how I generated these GRNs for each cell line in **2.2B4: Generating Gene Regulatory Network**

Weightings . I first identified TFs interacting with the cis-regulatory region of each protein coding TG, by isolating the TFBS occurring within the regulatory window demarcated by the most upstream and downstream occurring CTCF ChIP-seq peaks in a 50Kb region surrounding the gene body (**Figure 2.1**). I further filtered these positional TFBS based on statistical significance using the FIMO algorithm and TF binding affinity using the TEPIC algorithm. More details regarding TFBS identification and filtering are provided in the subsection **2.2B3**. The number of TFs, TGs and TFBS corresponding to different TFBS identification algorithms for both cell lines are also provided in **Table 2.1.**

After identifying different sets of TFBS, I created corresponding adjacency matrices to generate the motif networks for building the PANDA GRNs. I created binary (binding/no-binding) TF-TG adjacency matrices using the positional and FIMO TFBS. For the TEPIC based adjacency matrix, I used affinity scores of the TEPIC TFBS as weights. I combined these matrices with PPI data and cell type specific co-expression to fit a GRN using the PANDA algorithm.

I then used the edge-weights from each of the three GRNs (Pos GRN, FIMO GRN and TEPIC GRN) as well as the TEPIC affinity scores as features for predicting expression of TGs using elastic-net based regularized linear regression (ENET) for each cell line. Predictive performance for the models was measured using mean-squared error (MSE), and Pearson's correlation coefficient (PCC) between predicted and observed expression values of the test set TGs within a 5-fold cross-validation framework repeating for 20 iterations. I have provided more details regarding the generation of the training and test set of TGs as well as the building of the prediction models in **2.2B5.**

As already mentioned, I built the GRNs using adjacency matrices derived from CTCF

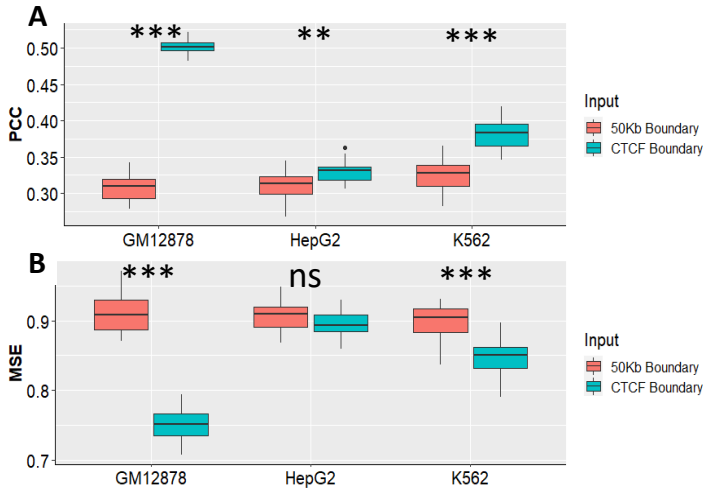

Figure 2.4: *Boxplots showing influence of using CTCF defined regulatory windows on gene expression prediction. Boxplots showing results from predicting gene expression using GRNs built using TFBS defined based on gene regulatory windows based upon CTCF peaks vs. 50Kb regions around TG body with respect to A)PCC and B)MSE*

boundary defined cis-regulatory TFBS(**Figure 2.1**). Specifically, I used the most upstream and downstream CTCF peaks to define the regulatory window of the TGs in order to look for overlapping TFBS. The median regulatory window distance across all the TGs for GM12878, K562 and HepG2 was 46,031 bp, 46,003 bp and 46,099 bp respectively. I also generated a set of TFBS for all the cell-lines based on a 50Kbp window around the gene body to compare the effect of using biologically relevant regulatory windows to those defined using traditional genomic distances on the prediction performance. For all the three cell-lines, the 50Kbp based TFBS set were significantly larger compared to the ones based on CTCF peaks (GM12878: 1,170,644 additional TFBS; K562: 2,096,025 additional TFBS and HepG2: 1,607,755 additional TFBS). In spite of this difference, the prediction performance of the CTCF defined TFBS based GRNs was significantly better to that obtained from the GRNs constructed using a traditional 50Kbp window to find TFBS as

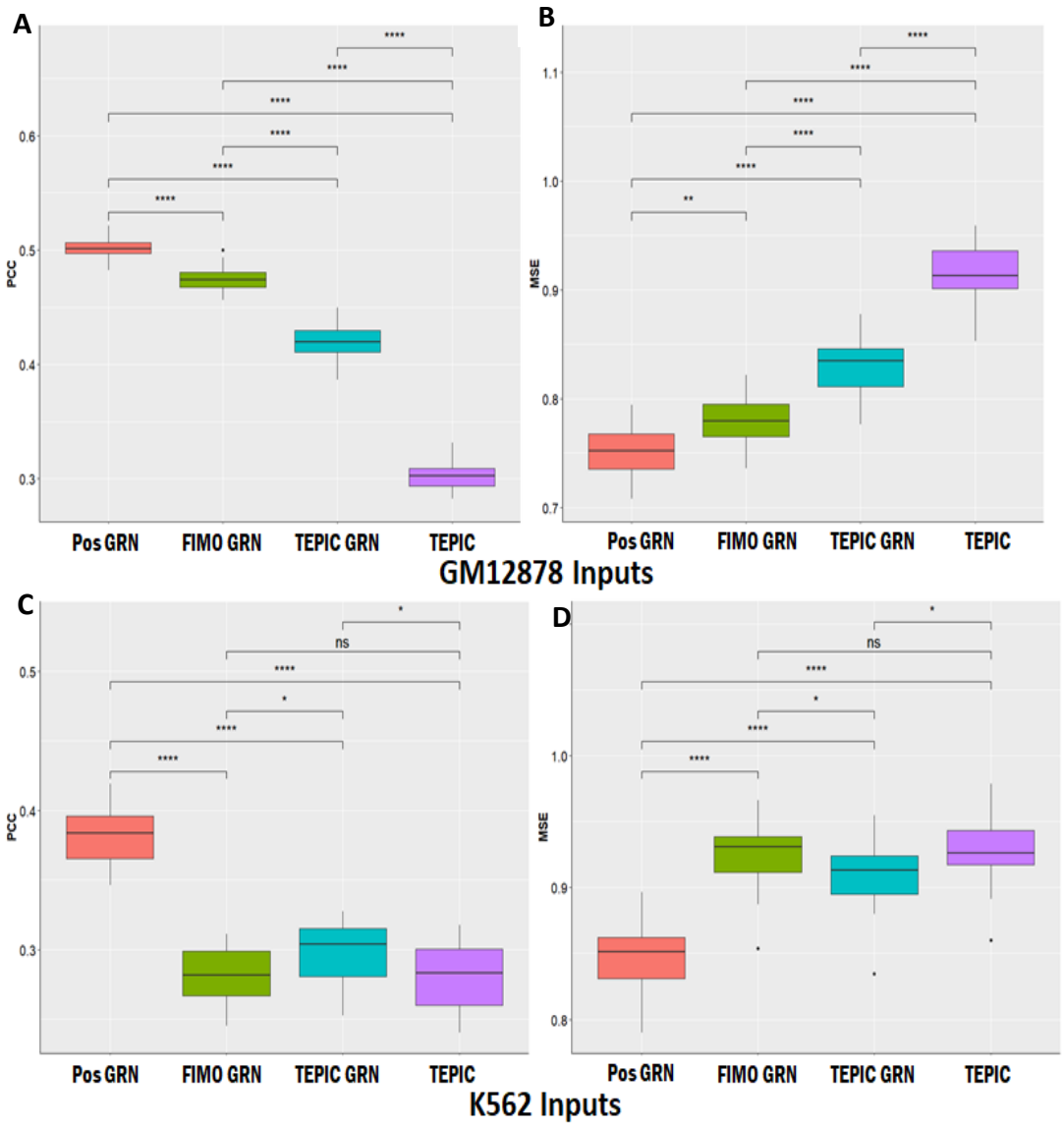

Figure 2.5: *Boxplots showing GRN based prediction models outperform those built using TEPIC affinity scores. A and B correspond to prediction performance for 20 random sets of 1729 GM12878 TGs while C and D were obtained from 1892 K562 TGs. Prediction performances for models corresponding to different inputs were compared using Wilcoxon signed-rank test (***-p < 0.0001, **-pvalue < 0.001, *-pvalue<0.05, ns-not significant)*

shown in **Figure 2.4**. Thus, using biologically defined regulatory boundaries for finding TFBS is essential to build GRNs and predict gene expression.

As shown in **Figure 2.5**, GRN based prediction models containing cis and trans regulatory mechanisms were more accurate e than models built using only cis-regulatory TF-TG TEPIC affinity scores. Specifically, the median PCC for TEPIC GRN based models

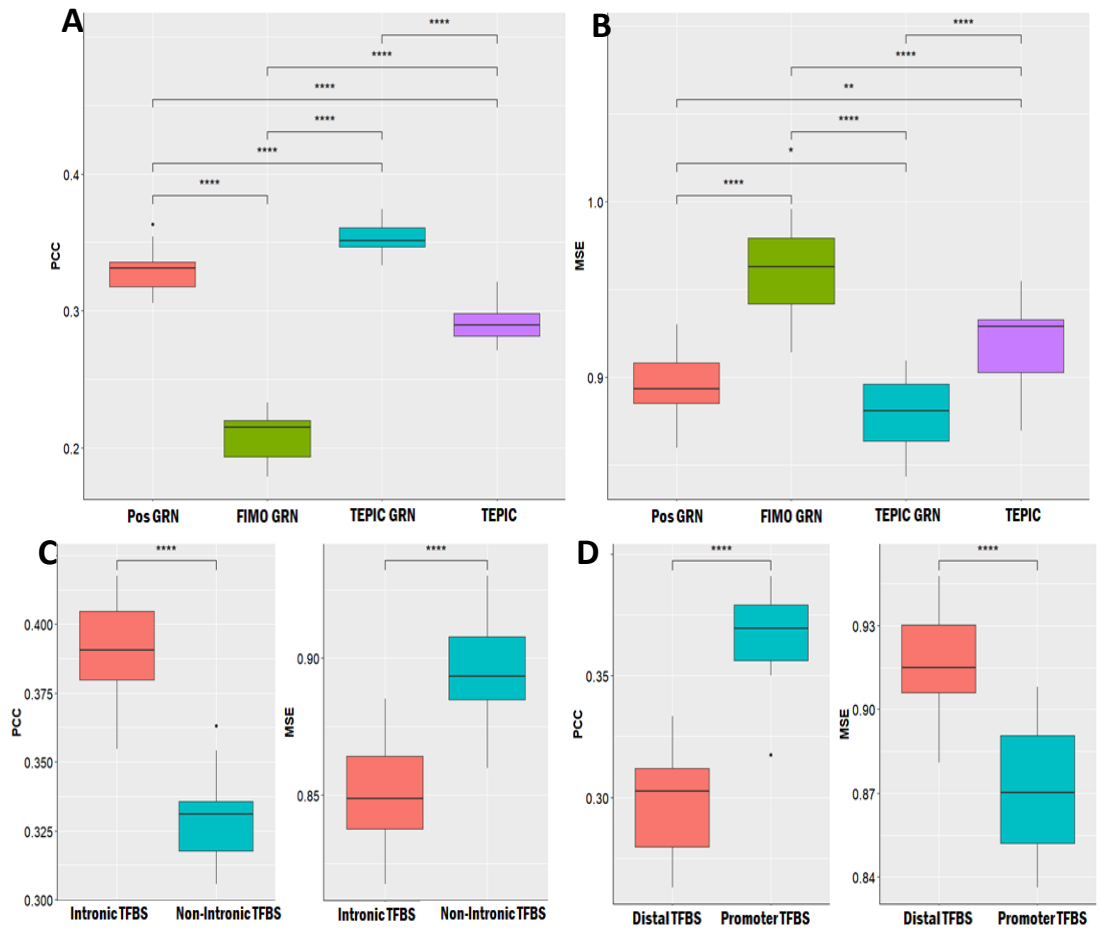

Figure 2.6: *Plots showing results from GRN based TG expression prediction for HepG2. I downloaded ENCODE data corresponding to 234 HepG2 TFs and found TFBS for 12,887 TGs. A) After generating GRNs based on positional ChIP-seq data, FIMO based statistically significant TFBS and calculating TEPIC scores, I compared prediction performance of the corresponding ENET models. B) MSE for the above models. Boxplots showing the comparison of the intronic vs. non-intronic TFBS based prediction model performance in C and of distal vs. promoter TFBS based prediction models in D*

was higher compared to that of TEPIC for GM12878 (0.42 vs. 0.30, Wilcoxon signed-rank

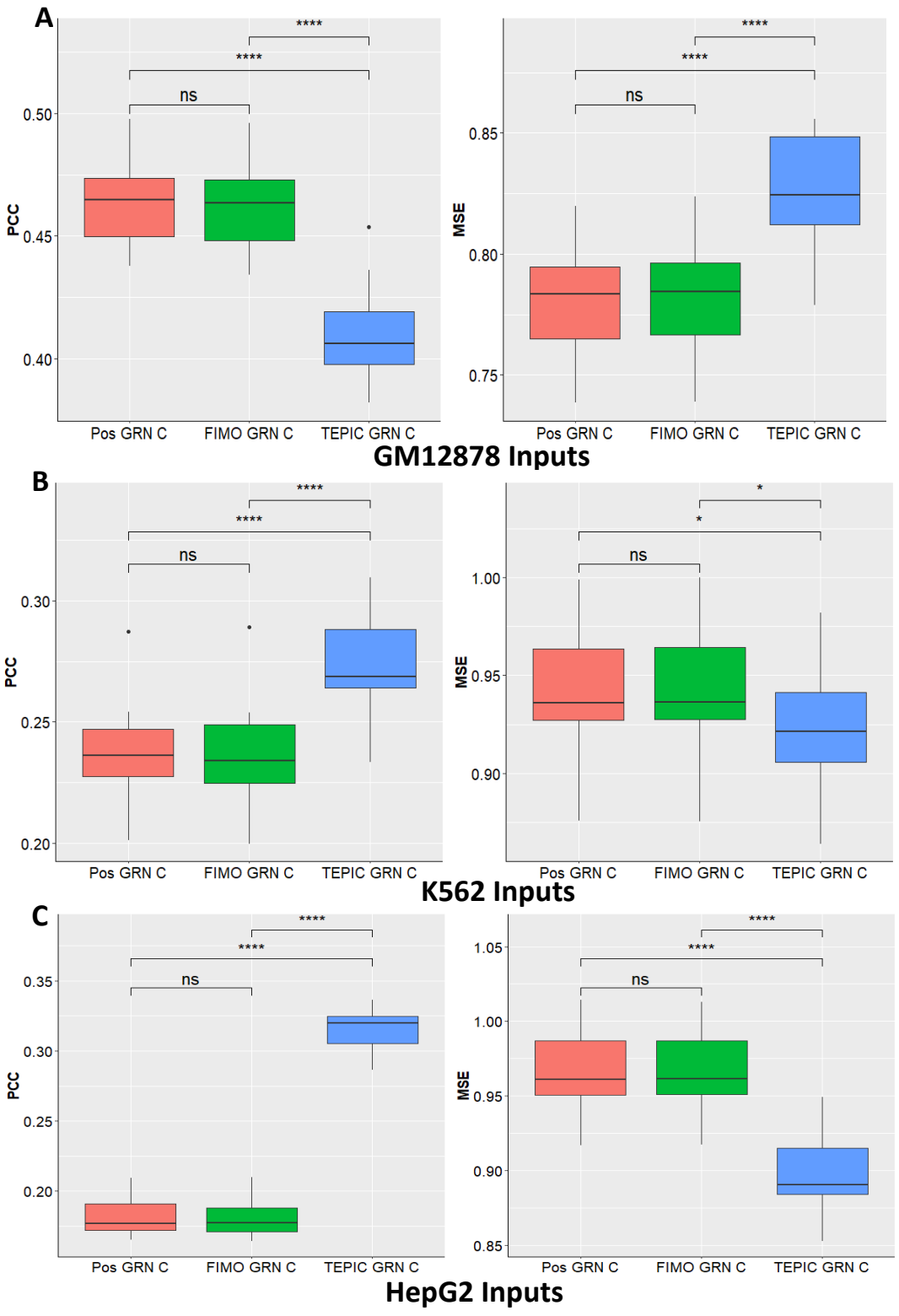

Figure 2.7: *The results from fitting prediction models with the same number of TF features for the inputs shown in* ***Figure 2.5****. The box plots are showing results from models containing 77 GM12878 TF features(A), 86 K562 TFs(B) and 73 HepG2 TFs(C). Here, I have added "C" in order to represent the comparable models and differentiate them from those shown in* ***Figure 2.1***

test p-value =1.45e-11 **2.5A**), K562(0.30 vs. 0.28, Wilcoxon signed-rank test p-value

=3.50e-2 **2.5C**) while the median MSE for the former was lower than that for the latter for GM12878(0.83 vs. 0.91, Wilcoxon signed-rank test p-value = 4.35e-10 **2.5B**), K562 (0.91 vs. 0.93, Wilcoxon signed-rank test p-value =3.26e-02 **2.5D**). Apart from the blood based cell lines GM12878 and K562, I also applied my modelling approach to the liver carcinoma cell line HepG2 to assess the generalizability and the robustness of the results. As shown in the **Figures 2.6A-B**, the results from the HepG2 analyses are qualitatively similar to the ones described above.

In addition to the main conclusion pointing towards the superiority of the GRN based features for predicting gene expression, I made the following additional observations from the aforementioned analyses:

1) Pos GRN models for GM12878 and K562 had the best performance of all models tested, while the TEPIC GRN models performed the best for HepG2 (**Figures 2.5** and **2.6A-B**). This could be due to the overfitting caused by the highest number of TF features in these models. To test this, I 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 (**Figure 2.7**). Furthermore, in this sensitivity analysis I 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.

2) Prediction performance of all GM12878 based models was better overall compared to that obtained from the other two cell lines. In order to consolidate for the difference in the number of TFs among the cell-lines, I utilized a common set of 61 TFs among the 3 cell-

lines as well as a set of 110 TFs common between GM12878 and K562 to generate the PANDA GRNs and subsequently predict TG expression. As shown in **Figures 2.8A** and **2.8B**, even in this restricted analysis, the performance of the GM12878 based models was significantly superior both for the 3 cell comparison($PCC_{GM12878}$ = 0.32, PCC_{K562}=0.22, PCC_{HepG2}=0.20) as well as for the 2 cell comparison($PCC_{GM12878}$ = 0.37, PCC_{K562} = 0.24). I also assessed the performance of the GM12878 GRN models for predicting cross-cell type TG expression. For this analysis, I tested the prediction performance of the models trained using regulatory information from one cell-line on the test genes of the other cell-lines. As shown in **Figures 2.8C** and **2.8D**, using TFs and TGs common between GM12878

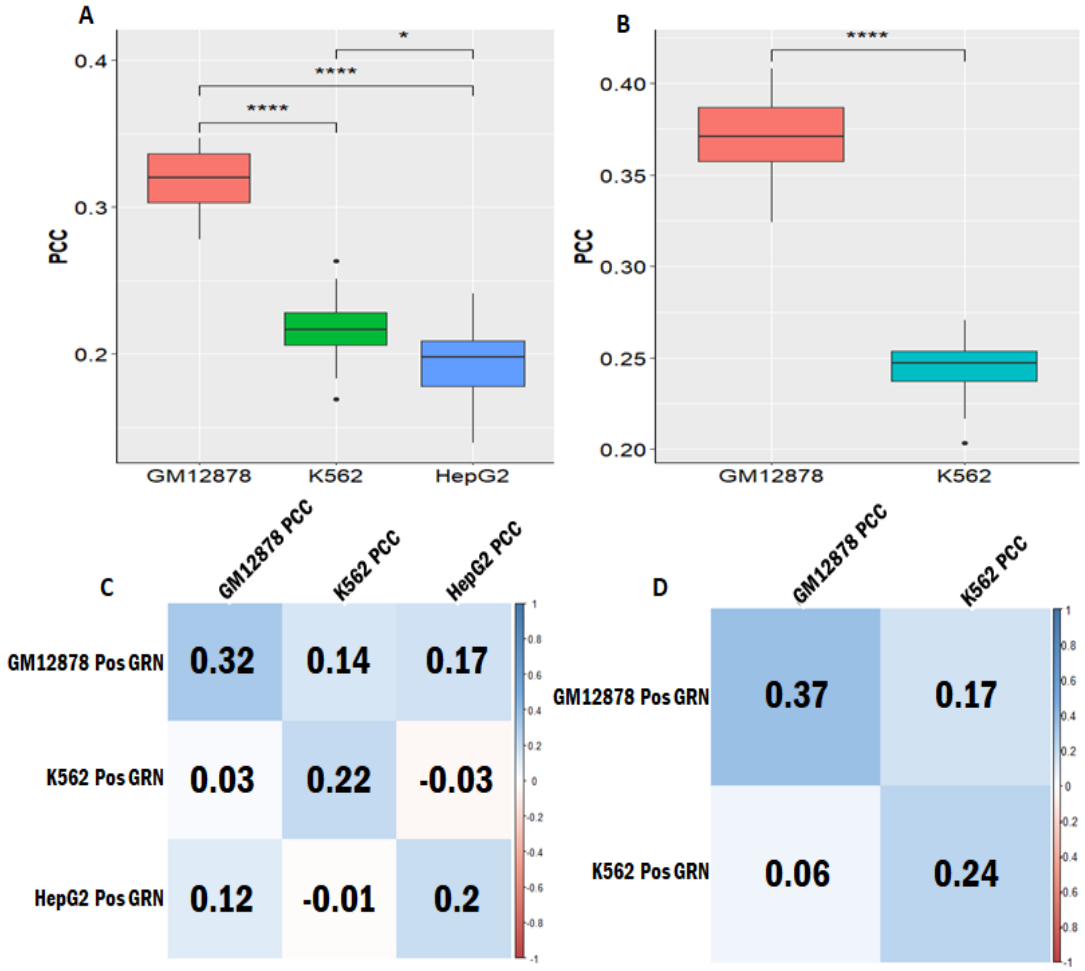

Figure 2.8: *Results from cross-cell-type TG expression prediction for GM12878, K562 and HepG2. Boxplots showing the prediction performance of GRNs generated using information corresponding to A) 61 TFs commons among GM12878, K562 and HepG2 and B) those generated using 110 TFs common between GM12878 and K562. C and D show the correlation plot for the cross-cell-type prediction performance for these models.*

and K562, I was not able to obtain decent prediction performance. However, the GM12878 Pos GRN based model performed better for K562 TGs than vice-versa. Similarly, GM12878 produced median PCC of 0.17 for K562 cell-line for the pairwise comparison and that of 0.14 for the three cell type comparison. On the other hand, despite being derived from a different lineage, the HepG2 cell-type produced decent median PCC of 0.12 for the GM12878 TGs but the prediction performance was very poor for the K562 TGs (Median PCC = -0.014). K562 based GRNs did not produce good cross-cell type prediction

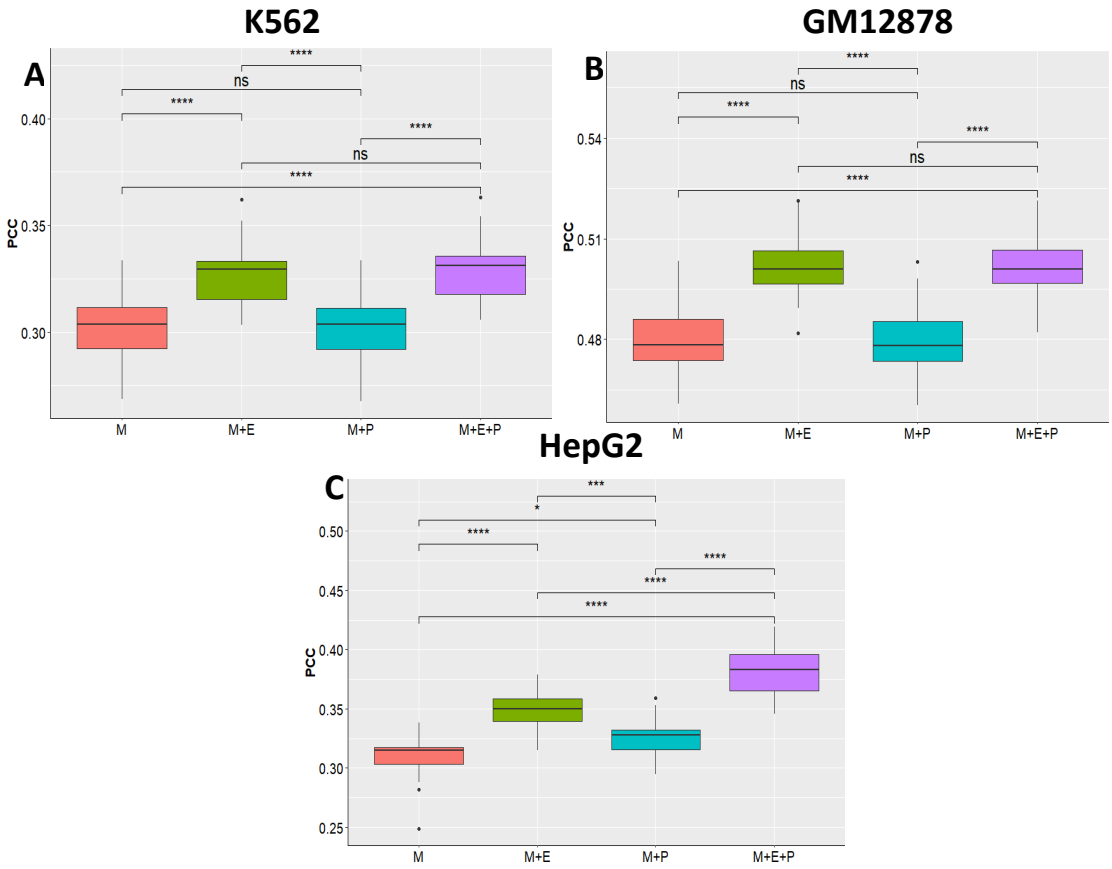

Figure 2.9: *Impact of removing different datasets from the PANDA GRN on prediction performance. "M" represents GRN containing Motif network, "M+E" represents one containing Motif and Co-expression datasets; "M+P" represents one containing Motif and PPI datasets and finally "M+E+P" represents the GRN containing all three datasets. The boxplots were created from the PCC obtained from predicting expression for 20 instances of A)1895 K562 B)1751 GM12878 and C) 2403 HepG2 test genes.*

performance in both the pairwise and the three way comparisons. Thus, GM12878 GRN based models had the best within cell-type as well as cross cell-type prediction

performance. I hypothesized that this stark difference in the prediction performance was due to the markedly smaller sample size of the PANDA expression data for K562 and HepG2 relative to GM12878 (9 and 8 vs 462). In order to detect the impact of the co-expression dataset on the prediction performance of the models, I eliminated PPI and co-expression datasets individually and together from the GRN and replaced them with an identity matrix. As shown in **Figure 2.9**, the prediction performance for the GRN containing just the TFBS based motif information was the poorest for all the cell-lines. This was expected as this GRN was devoid of the information from other regulatory

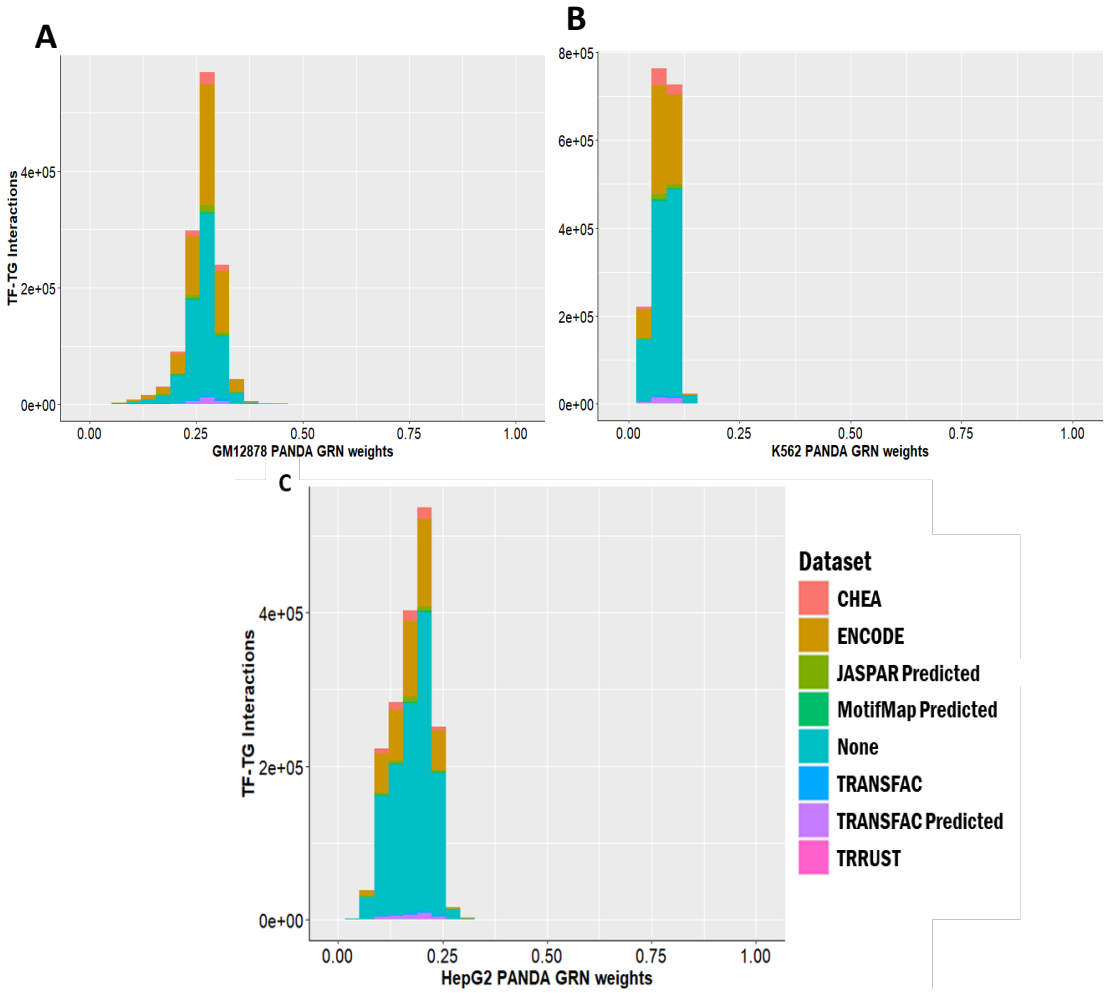

Figure 2.10: *Significant proportion of the edges present in the PANDA GRN networks represented known and predicted TF-TG interactions. I looked for annotated and predicted TF-TG interactions present in the Harmonizome and TRRUST databases corresponding to the edges in the PANDA GRNs for the three cell-lines. Histograms in the figure show the scaled Pos GRN edge-weights for the unique TF-TG interactions binned according to the annotation dataset shown in the legend for A) GM12878, B) K562 and C) HepG2 cell-lines.*

mechanisms. Moreover, the GRN containing motif and PPI information produced worse prediction models compared to the ones containing motif and co-expression datasets for the three cell-lines. The performance of "M+E" GRN was comparable to the one containing all the three types of networks, and that of "M+P" was comparable to the one containing just motif information. Thus, I was able to conclude that the co-expression datasets provided important information to generate PANDA GRN as they captured the correlation patterns for genes that were co-regulated by the same set of TFs.

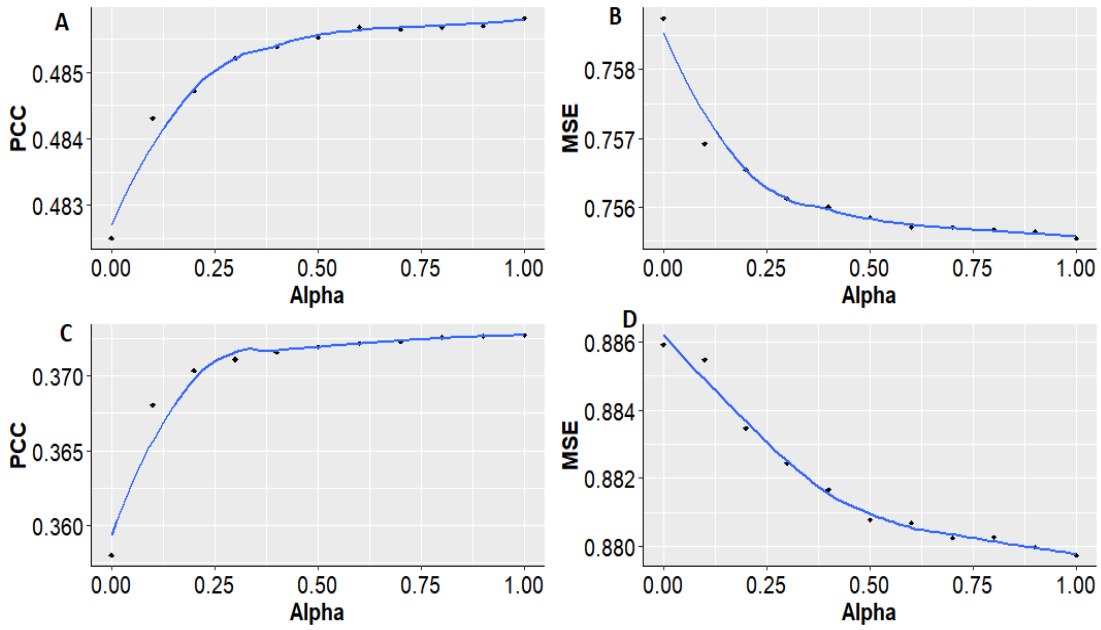

Figure 2.11:*Plots showing the impact of alpha("l1 ratio") on prediction performance of ENET models.A) and B) show the prediction performance of the ENET regression models with regard to PCC and MSE respectively for different values of alpha("l1_ratio") for 1 iteration built after predicting expression for GM12878 TGs . C) and D) show the performance for PCC and MSE respectively after predicting expression for K562 TGs.*

Lastly, I also explored the accuracy of the TF-TG regulatory information captured by the Pos GRNs corresponding to the three cell-types. I downloaded the TF-TG interactions defined within the Harmonizome[83] and the TRRUST(*version.2.0*)[84] datasets based on literature annotations and motif based computational predictions. I overlapped the PANDA GRN based TF-TG edges on top of these interactions and plotted them, along with their edge-weights, in the histograms shown in **Figure 2.10**. For the GM12878 cell-

line, I observed that 527,636 (43%) PANDA edges were present within the annotated TF-TG interactions, while for K562 and HepG2 the number of edges present within the annotated TF-TG interactions were 567,325(24%) and 443,080(26%) respectively. The higher number of literature annotated TF-TG interactions for the GM12878 GRN could be due to the more extensive co-expression dataset used to build the network leading to a more accurate estimation of TF-TG regulatory relationships compared to the other two cell-lines.

2.2C2: 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). I 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 **Figure 2.11**, moving towards a higher value of α improved the prediction of 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. Thus, Using an α of 0.5 (balancing L1 and L2 penalties) and equation (2.2), I averaged the effect estimates of 149 TFs(GM12878) and 309 TFs(K562) learned the Pos GRN models fit for 20 iterations.

ENET models use LASSO and ridge penalty terms to shrink effect estimates corresponding to features that don't contribute towards explaining the variance in the outcome[85]. This shrinkage of weights helps in reducing a complex model into a simple one and in overcoming the multi-collinearity within features. Using the Pos GRN ENET models learned from predicting the expression of TGs for K562 and GM12878 cell-lines, I calculated the average effect estimates for each one of the TFs. Of the 149 GM12878 TFs, 45(30%) had really small effect estimates in the range [-0.01,0.01], while for the 309 K562 TFs, 173 (56%) had their weights shrunk to really small size. In order to determine the effect of the correlation structure of the PANDA GRN feature weights on this shrinkage, I performed k-means clustering to define 5 clusters based on the correlation matrix derived from these weights. These clusters shown in **Figures 2.12A** and **2.12C** contained TFs with highly correlated GRN based feature weights. For each cluster, I plotted the corresponding mean effect estimates shown in the histograms in **Figures 2.12B** and **2.12D**. These histograms represent the mean ENET effect estimates in relation to the correlation structure within the TF features for the two cell-lines. Each cluster can be seen containing some TFs with high effect estimates in both directions, while some had effect estimates close to zero. For instance, cluster 5 in GM12878 contained 63 TFs, of which 26 had mean effect estimates in the range [-0.01,0.01], while estimates for 16 TFs were either greater than 0.1

or smaller than -0.05. On the other hand, cluster 2 for K562 cell-line contained 190 TFs, out of which 108 had effect estimates in the range [-0.01,0.01] and 34 contained effect

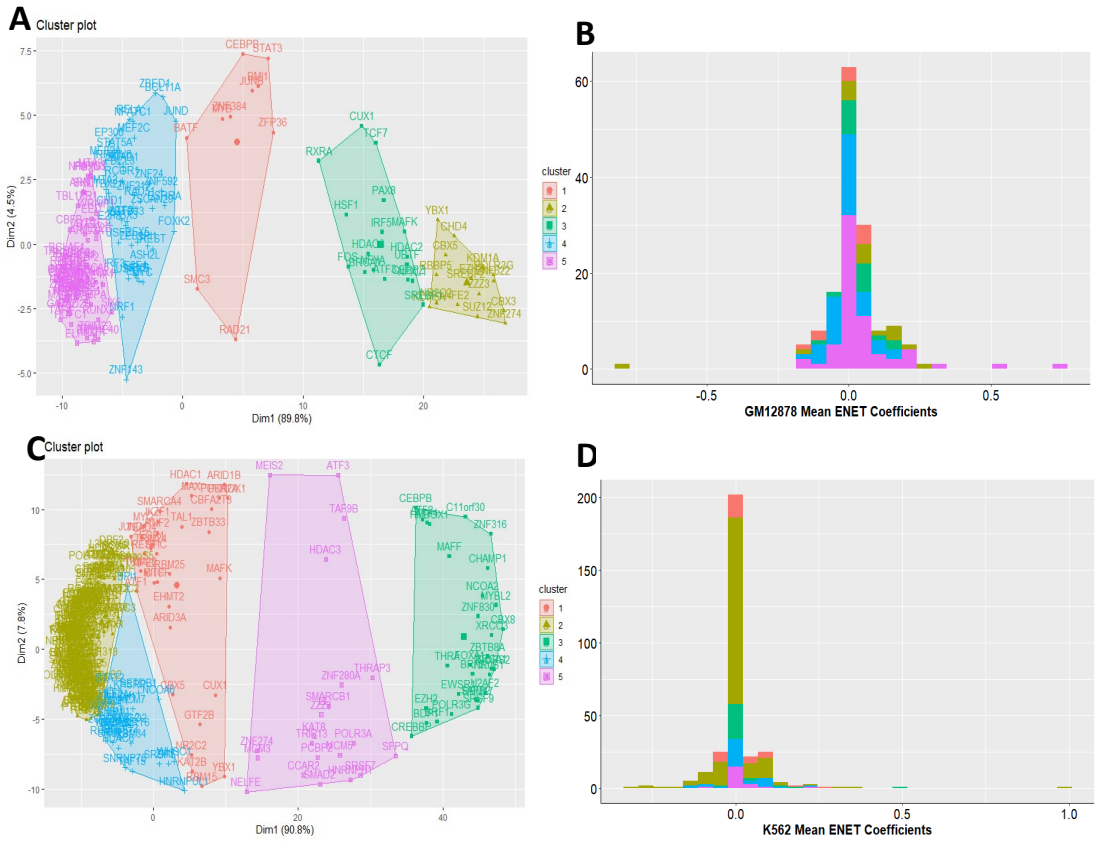

Figure 2.12: *The correlation structure for the TFs within PANDA GRN features is captured by ENET models. Plots showing the clusters defined by k-means clustering method using the correlation matrix among the 149 GM12878 TFs in A and the 309 K562 TFs in C. B and D show the histograms containing the mean ENET coefficients for the GM12878 and K562 TFs respectively binned by the clusters in which they were present.*

estimates either greater than 0.01 or smaller than -0.05.

Histograms in **Figure 2.13** show 5 roughly equal bins created using the mean effect estimates for TFs in GM12878 and K562. I performed a GO enrichment analysis for TFs in each bin and reported the top 5 enrichment terms for biological processes and molecular functions in **Figure 2.14** for both cell types. I observed that as I 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, I

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), I predict its

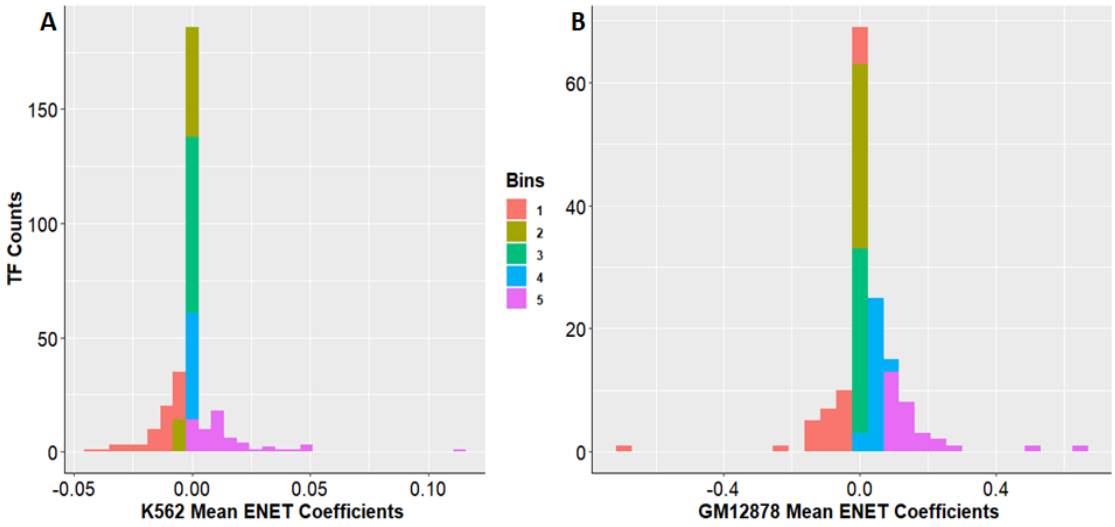

Figure 2.13: *Mean ENET effect estimates reflect the important functional roles of various TFs. Histograms of the average effect estimates for calculated for A) 309 K562 TFs and B) 149 GM12878 TFs 3 using the "Pos GRN" ENET models. I also created 5 bins(quintiles) based on the effect estimates, which are color coded in the histogram.*

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), POLR2.2A(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. I 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.

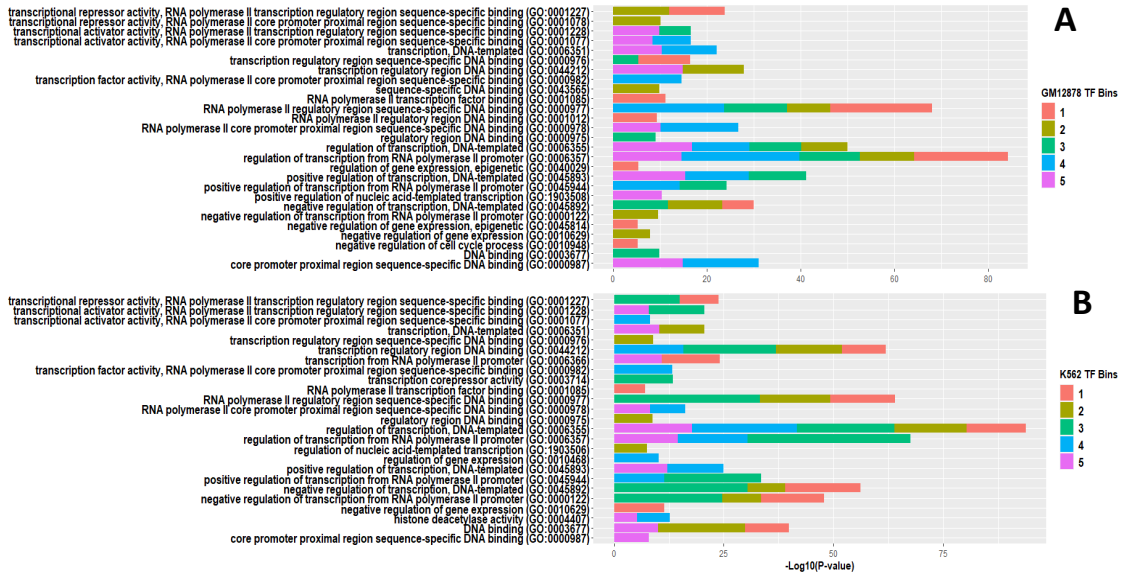

Figure 2.14: *GO Enrichment results for the TFs placed in different bins for GM12878 and K562: I divided the TFs into 5 bins based on their average effect estimates. A) shows the top 5 significant GO BP and GO MF enrichment terms for 149 GM12878 TFs and B) shows the same for 309 K562 TFs.*

I also did a similar aggregation analysis for the TF effect estimates learned from the TEPIC GRN and the TEPIC models for the two cell lines in order to explain the improvement in the prediction performance of the former compared to the latter observed in the earlier results. Additionally, I rank-ordered the TFs based on these effect estimates,

Table 2.2: *Table showing the comparison between the mean effect estimates obtained from ENET models of TEPIC and TEPIC GRN.The mean ENET effect estimates, ranks and the change in ranks for the 33 TFs(GM12878) and 38 TFs(K562) obtained from comparing the TEPIC and TEPIC GRN models. Here, TFs with the change in rank of at least 10 positions are shown.*

K562					
TF_Name	Mean ENET Coefficient(TEPIC)	Rank(TEPIC)	Mean ENET Coefficient(TEPIC GRN)	Rank (TEPIC GRN)	change_in_rank
THAP1	0.94	2	0.08	49	-47
NFIC	0.09	24	0.05	56	-32
SRF	0.01	37	0.84	8	29
CUX1	-0.03	65	0.2	37	28
SPI1	-0.01	58	0.29	30	28
MITF	0.09	25	0.07	53	-28
EWSR1	0	44	0.44	19	25
MYBL2	0.01	36	0.71	11	25
FOXA1	-0.09	71	0.09	48	23
NFE2	-0.01	59	0.2	36	23
CEBPB	0	45	0.38	22	23
PKNOX1	0.01	38	0.52	16	22
E2F8	0	46	0.35	26	20
FOXK2	0.03	31	0.08	51	-20
NFYA	0.13	20	0.13	40	-20
MNT	-0.03	64	-0.2	81	-17
EGR1	0.01	39	0.38	23	16
IRF1	0	47	0.01	62	-15
SMAD2	0.01	35	0.38	20	15
NEUROD1	0.09	26	0.12	41	-15
GABPA	0.63	3	0.49	18	-15
ESRRA	-0.12	73	0.04	59	14
E2F7	0	49	0.01	63	-14
NR2F1	0.2	17	0.29	31	-14
CREB3L1	0.25	16	1.05	2	14
USF1	-0.03	62	-0.06	75	-13
TAL1	0.03	30	0.11	43	-13
MEF2A	0.07	27	0.59	14	13
NRF1	0.39	12	0.35	25	-13
RUNX1	0	41	0.33	29	12
MEIS2	0.03	32	0.1	44	-12
STAT1	0.1	23	0.22	35	-12
GATA1	-0.19	79	-0.02	68	11
ZBED1	0	50	0.02	61	-11
TEAD4	0.1	22	0.24	33	-11
ELK1	-0.03	63	-0.04	73	-10
TCF7L2	0.04	29	0.15	39	-10
E2F1	0.27	14	0.99	4	10
GM12878					
RELB	-0.13	68	0.19	24	44
SREBF1	0.04	37	-0.72	79	-42
NR2C2	0.00E+00	57	0.35	16	41
POU2F2	-0.28	75	0.05	36	39
NFATC1	0.67	4	0.02	42	-38
ETS1	-0.22	74	0.04	38	36
RXRA	0.07	29	-0.04	65	-36
BHLHE40	0	43	-0.16	74	-31
USF1	0.02	40	-0.08	71	-31
RUNX3	0.41	11	0.03	40	-29
BATF	0	46	-0.1	72	-26
RELA	-0.16	70	0	46	24
MXI1	-0.06	65	0.02	43	22
STAT3	0.04	34	-0.01	56	-22
FOS	-0.38	77	-0.01	58	19
SMAD5	0.05	33	0.39	14	19
RFX5	-0.01	61	0.01	44	17
E2F8	0.06	30	0	47	-17
ARNT	0.3	14	0.14	31	-17
JUNB	0	47	0.09	33	14
MAFK	0	49	-0.03	62	-13
USF2	0.15	21	0.46	9	12
NR2C1	0	53	-0.03	64	-11
MEF2A	0	50	0.03	39	11
MAX	0.15	22	0.44	11	11
MEF2C	0.27	17	0.56	6	11
ZNF384	0.54	6	0.35	17	-11
TCF3	-0.37	76	-0.04	66	10
FOXK2	-0.07	66	-0.21	76	-10
PAX8	0	42	0	52	-10
CEBPB	0.02	38	0.18	28	10
SPI1	0.2	20	0.17	30	-10
CUX1	0.28	16	0.18	26	-10

such that the top ranking TFs had the most positive effect estimates and the bottom ranking

ones had the most negative effect estimates, and tabulated them in **Table 2.2**. In comparison to the TEPIC models, I observed an improvement in ranks for TFs associated with transcriptional activation as well as a decrease in ranks for the repressive TFs in the TEPIC GRN models. In other words, the effect estimates learned from the TEPIC GRN models more accurately represented the functional roles of the TFs compared to the TEPIC models leading to the better prediction of TG expression.

Thus, the aggregation of the TF effect estimates from my learned models not only illuminated the functional roles of the lesser known TFs but also helped us explain the difference in predictive performance of the features derived from the TEPIC models and those derived from the TEPIC GRN models.

2.2C3: 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, I built prediction models using GRNs containing TFBS found in those regions (**Table 2.1**) and assessed their predictive performance. Details regarding the definition of these regulatory regions and subsequent identification of the TFBS have been provided in **2.2B7**.

I first analyzed TFBS within the promoter regions (5Kb upstream of the TSS of the genes), intronic TFBS, and distal ones present outside these areas. 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 shown in **Figures 2.15A** and **2.15B** 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 2.15C, 2.15D**) with respect to median MSE (GM12878 p = 4.72e-04) and median PCC(GM12878 p = 1.33e-08; K562 p = 2.45e-02).

I next used Hi-C data corresponding to GM12878 and K562 in order to capture long distance interactions between distal TF binding and gene promoters. I 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 (2.3) as shown in **Figure 2.16A** and described in **2.2B8**. Prediction models including Hi-C adjusted distal TFBS were significantly more accurate compared to the ones built using normal distal TFBS as shown in **Figure 2.16B** with regard 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.

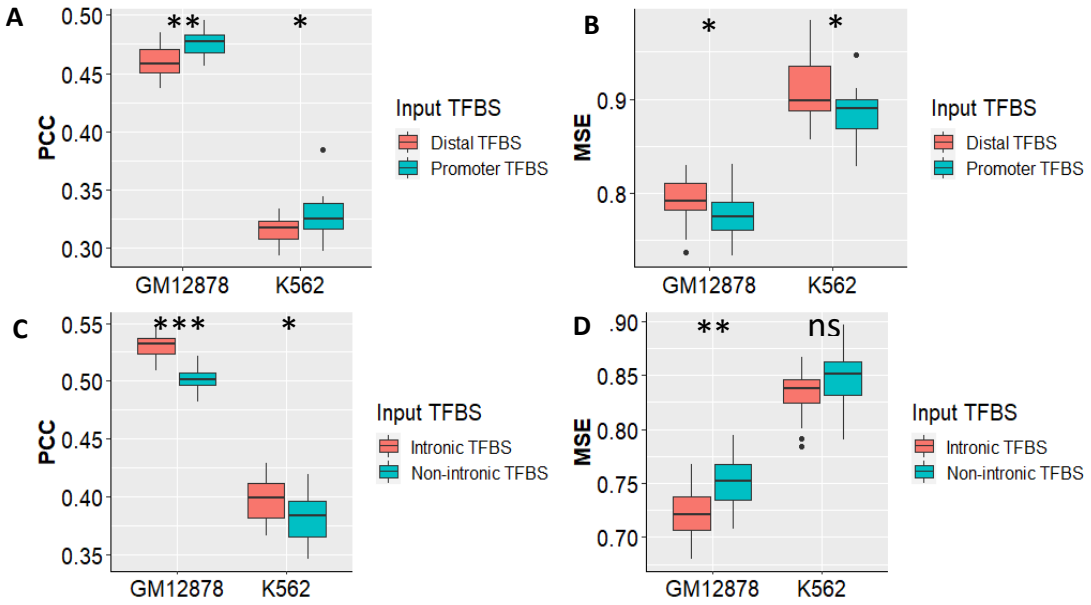

Figure 2.15: *Intronic and Promoter TFBS are important for predicting gene expression. A)PCC and B) MSE obtained from the expression prediction of GM12878 and K56 TGs using models built from GRNs containing promoter and distal TFBS. C) PCC and D) MSE produced by models predicting expression for GM12878 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.*

Next, I expanded this weighting scheme to include promoter TFBS. As promoters are regions of high TFBS activity, I expected a high degree of Hi-C contact points within promoter regions. Unexpectedly, these models performed significantly worse; I observed a large number of promoter TFBS (59% for GM12878 and 90% for K562) that showed no

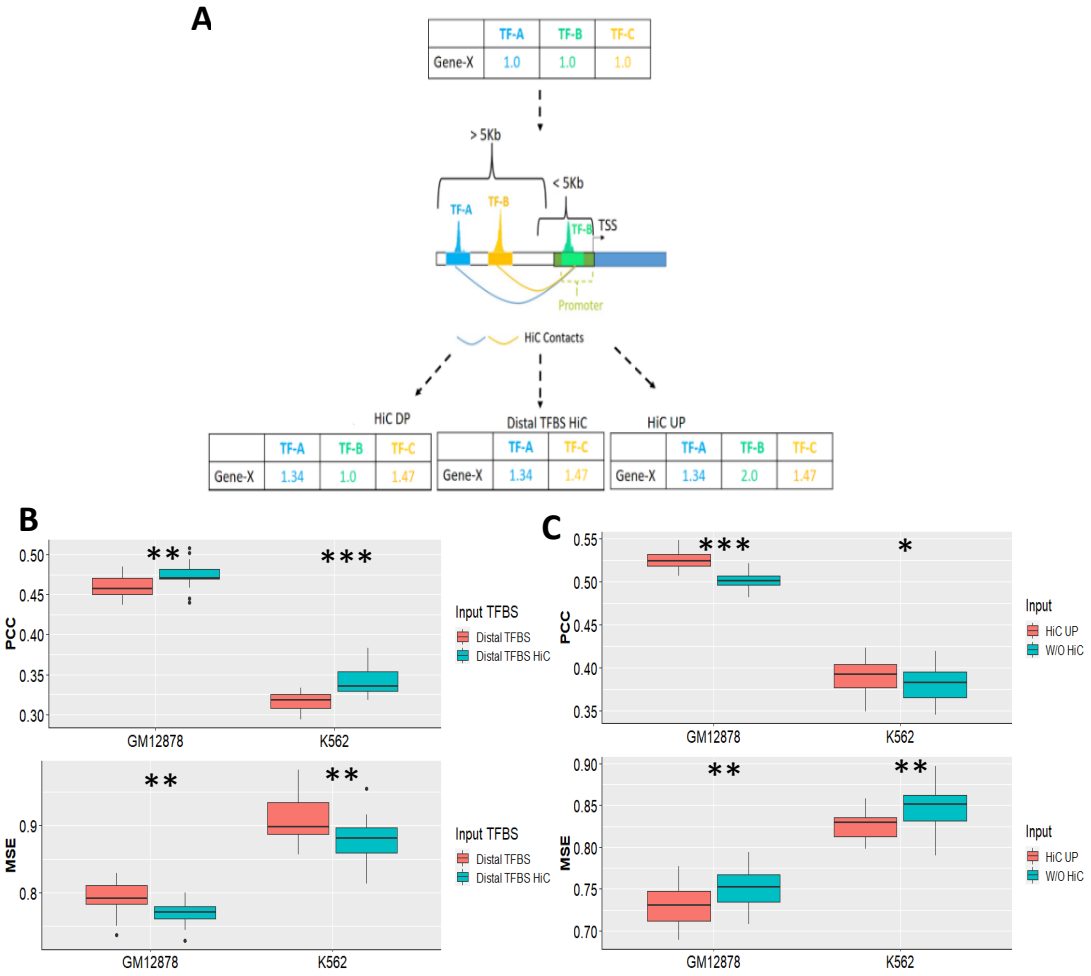

Figure 2.16: *HiC data is capable of capturing the effect of long distance interactions between TF binding within distal TFBS and gene's promoter on gene expression. A) I used the cell line specific Hi-C data to weight the distal TF-TG interactions in the motif adjacency matrix. I also down-weighted 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) I 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.*

evidence of within-promoter contacts, and using this weighing approach effectively down-weighted promoter TF-TG interactions (Hi-C DP). I therefore also considered an approach

that applies the maximum Hi-C weight to all promoter TFBS (Hi-C UP), shown in **Figure 2.16A**. These Hi-C UP based prediction models significantly outperformed all the other models for both cell types as shown in **Figure 2.16C**. These Hi-C UP based prediction models significantly outperformed all the other models for both cell types as shown in **Figure 2.16C** and **Figures 2.17A-B.** Thus, Hi-C data added important regulatory information to my models capturing the effect of long distance interactions between TFs binding to distal regulatory elements and the TG promoter.

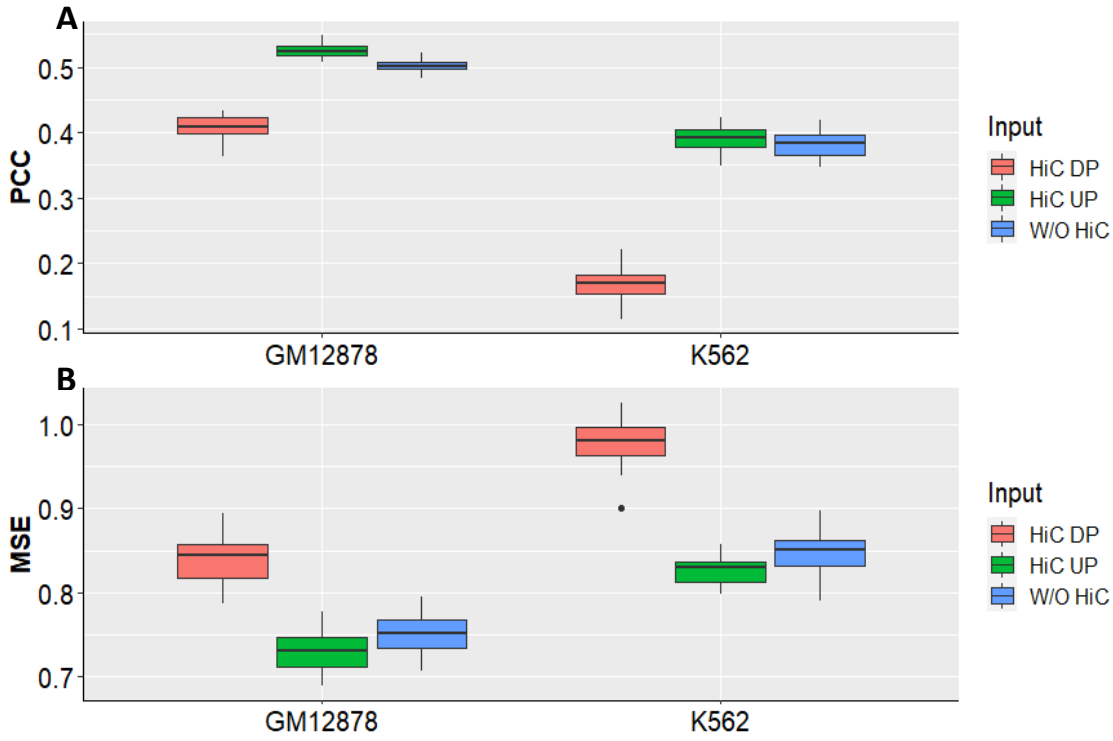

Figure 2.17: *Boxplots capturing the effect of long distance interactions between TF peaks and TG promoters on expression prediction. This figure is the extension of* ***Figure 2.16C*** *with the results from HiC DP GRN based prediction models added to the analysis. The prediction performance of HiC UP GRN prediction models was significantly better than the ones constructed without HiC information and the HiC DP normalized TF-TG motif network. A shows the prediction performance w.r.t median PCC while B shows the performance w.r.t median MSE. differently weighted HiC motif matrices (HiC DP and HiC UP) in comparison to those using GRNs based on unweighted motif matrices(Pos GRNs) for both cell types.*

2.2C4: 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[54]. Here, I present the utility of the TF features derived from my integrative GRN based prediction framework for weighting rare variants

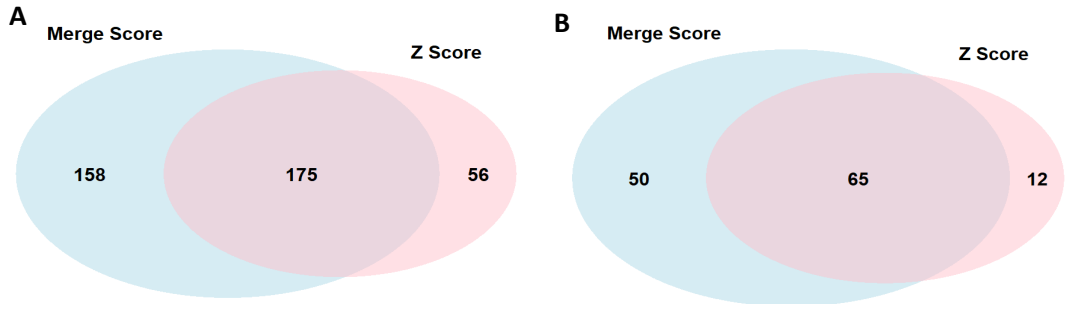

Figure 2.18: *Results from rare variants analysis based on merging QBiC-Pred scores with GRN derived TF effect estimates. QBiC-pred z-scores with GRN derived TF ENET effect estimates enriches identifications of TGs significantly associated with expression trait.A) shows the venn diagram containing significant TGs(N=389, p-value < 4.18e-06) obtained from the initial discovery analysis based on fitting the merge score and z-score SKAT models using the DGN dataset. B) shows the significant TGs(N=127, p-value < 0.05)identified within the replication analysis done using the GTEx dataset.*

within kernel-based association tests to improve their power. I used the DGN dataset[86] containing HRC-imputed variant genotypes and RNA-seq from the whole blood of 922 individuals in order to perform SKAT[56] based rare variant analysis. I 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 feature weights in form of effect estimates. I 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[42]. By merging this score with the average effect estimates of the corresponding TFs, based on equations (2.3) and (2.4), I created a variant scoring metric representing the estimated average effect of a base-pair change on TF-TG regulation. More details about how I derived these merged scores for the rare variants are provided in **2.2B9.**

I used the merged scores to perform SKAT for the normalized expression of TGs in the DGN dataset. I 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 **Figure 2.18A**, 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. I also performed a replication analysis using the whole blood sequencing and expression data from 369 individuals within the GTEx dataset[49]. I was 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 (**Figure 2.18B**). Thus, utilizing TF-TG regulatory information learned from my GRN framework for weighting rare variants enriched the identification of TGs, which would have been missed if I had only utilized variant influence over TF binding.

2.2D: Discussion

In this chapter, I 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. My approach explained more variance in gene expression compared to models built using TF-TG affinity scores for cis-regulatory features alone. My approach significantly outperformed models built using TF-TG affinity scores for cis-regulatory features alone.

I further estimated the influence of individual TFs on gene expression outcomes based on their effect coefficients learned from my 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. I 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 my 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, I observed that the highest ranking TFs are crucial for transcriptional initiation and activation, binding 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[87] . Promoter TFBS based models were also significantly more accurate at predicting gene expression than models using distal TFBS alone. These results validate my modelling strategy, as these findings are consistent with observations from previous studies[20,88], 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[77].I 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.

My 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[25,89]. Second, introns could house alternate promoters for a gene, as noted by analyses of GTEx and FANTOM datasets[90]. For I analyses, I used the upstream TSS of the longest transcript to define gene promoter regions.

Finally, I utilized the TF-TG regulatory information learned from my 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 I used linear regression based QBiC-Pred to score TF binding affinity, more complex scoring approaches could also be used within the framework. These analyses demonstrate the utility of my models for annotating otherwise difficult to characterize regulatory variants.

The most direct comparison of predictive performance for my models against published methods is the TEPIC method, which I outperformed. Other approaches have included either more complex modeling techniques or additional histone modification data to improve model performance[20,22]. Non-linear prediction models such as support vector regression or multi-layer perceptrons applied within my 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[22]. Thus, inclusion of both histone modification data and TF binding as predictors could diminish the effect of individual TFs, clouding the interpretation of my predictions.

At present, my 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[1] . This dearth in data availability is due to the difficult and expensive nature of the ChIP-seq experiments themselves[91]. 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[43,46]. In future work, these TF binding predictions could supplement the set of inputs to my GRN-based framework to produce better models.

CHAPTER 3: DETECTING TF REGULATORY MODULES UTILIZING MULTI-OMICS GENE REGULATORY NETWORK BASED DEEP LEARNING MODELS

3.3A: Introduction

Concerted and combinatorial binding of transcription factors (TF) within the cis-regulatory elements of target genes (TG) in humans gives rise to transcriptional regulatory modules(TRMs), which are essential for regulating TG expression[1]. These TRMs influence TG expression both additively and non-additively as seen in model systems[25,26]. Physical interaction among TFs, which is the basis for the formation of these TRMs, has been theorized to occur based on different models[92]. "Enhanceosome" and "Billboard" models represent linear co-operative interactions among TFs brought about by their DNA sequence based motif proximity, with "Billboard" models allowing a more flexible motif orientation and spacing than the "Enhanceosome"[9293]. Alternatively, the "TF collective" model comprises of non-linear TF interactions, independent of the DNA sequence motif composition. While this model was originally based in part on protein-protein interactions among TFs, it may also be due to distally binding TFs brought together via chromatin looping[92]. Previous computational approaches have mainly focused on characterizing TF interactions using the "Enhanceosome" and "Billboard" models[27–29,94]. However, the influence of TFs interacting via the non-linear "TF collective" model on TG expression is not well understood. Disruption in TRM based TG expression regulation, caused by genetic mutations in the TFs forming these TRMs, has been associated with several diseases. For instance, genes encoding TFs forming the BAF chromatin remodeling complex and the cohesin complex are found mutated in some congenital disorders[95]. Similarly, mutations in the TFs mediating the interaction of distally binding TFs with TG promoters and in those present within the heterochromatin forming polycomb-repressive complex(PRC) have been shown to cause different types of tumors[33]. Genetic disruption in the AP-1 factor

complex based TG regulation has been found to cause neurodevelopmental disorders as well as autoimmune diseases [9697]. While most of these examples have been studied in isolation, a systems-wide understanding of TG expression regulation driven by TRMs will likely unravel regulatory mechanisms underlying a range of other diseases.

Availability of high-throughput ChIP-Seq datasets, which provide the sequence specific binding information for each TF, has enabled researchers to detect TF interactions across the whole genome. Although these approaches, which have been described in **1B** , have helped in systems-wide detection and characterization of TF interactions, they have the following limitations: 1) Interactions that non-additively influence TG expression via distally binding TFs caused by chromatin looping (the "TF collective" model) cannot be detected using the abovementioned methods, as they rely upon TF co-localization information to identify proximally co-binding TFs (more consistent with the "Enhanceosome" and "Billboard" models). 2) The unsupervised clustering and topic modelling methods require the user to pre-determine the number of TF interactions to be identified preventing the agnostic discovery of TF interactions. 3) Lastly, and most importantly, for most of these studies the quantitative impact of TF interactions on TG expression remains unknown.

In this chapter, I use a multi-omics machine learning framework to model the impact of multiple TF based regulatory mechanisms on TG expression and detect TRMs based on the interaction effects learned from these models. I generated a gene regulatory network(GRN) containing information from datasets representing TF-TG, TF cooperativity and TG co-regulation. The TF-TG interactions in my multi-omic GRN were also weighted based on chromatin looping interactions made by distally binding TFs with

the TG promoters to appropriately capture their effect on TG regulation. I used the features from this GRN to predict TG expression values in the GM12878 lymphoblastoid cell line(LCL) using non-linear deep learning multilayer perceptron(MLP) prediction models. By aggregating interaction effects among different combinations of TFs from my learned models, I was able to identify specific TRMs that had high impact on TG expression. I validated the TF interactions, that I discovered within these TRMs, based on long distance chromatin looping contacts between their distal binding sites and significant spacing between their motifs for proximal binding sites. I also characterized the transcriptional regulatory programs for these modules based on the orientation and interaction of the corresponding ChIP-seq peaks relative to the promoters of TGs. Using I flexible multi-omics machine learning framework, I was able to detect TRMs significantly influencing TG expression, while characterizing their regulatory architectures using biologically relevant information.

3.3B: Methods and Materials

3.3B1: Datasets and algorithms used in this chapter

Most of the datasets(ENCODE, BioGRID, JASPAR, GEUVADIS and GEO) and the algorithms (ENET and PANDA) used in the analyses described in this chapter have already been described in **2.2B.** Additionally, I used the SpaMo(spaced motif analysis) to identify significant spacing between motifs to two proximally binding co-localizing TFs[29]. It works on the hypothesis that if two TFs bind at a fixed distance in the given set of input DNA sequences then there is a high probability that they form a complex with each other. Working on this hypothesis, SpaMo scans a given set of DNA sequences for the presence of a primary motif corresponding to a TF and identifies hits based on position weight matrix

scores. It then scans the sequences for the presence of a set of secondary motifs corresponding to a set of TFs and finds significant hits using position weight matrices again. It then calculated displacement between the primary motif and the set of secondary motifs and derives p-value using the null hypothesis that the displacements between a pair of primary and secondary will follow a uniform distribution if there is no significant interaction between the TFs. On the other hand, a binomial distribution of the displacement would reflect significant interaction between a pair of TFs.

3.3B2: Building multi-omics GRN

I utilized the Passing Attributes between Networks for Data Assimilation (PANDA) algorithm to build the GRN. This algorithm uses a TF binding site(TFBS) based motif network, a PPI network and a co-expression network for building the GRN(**Figure 3.1**). I generated these three networks using the following approach:

Motif network: I 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. I 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 **Figure 3.1**. Furthermore, I added the TFBS found in the intronic regions of each TG to this set in order to capture the effect of introns on transcriptional regulation. I have shown previously in **2.2C3** that inclusion of intronic TFBS in the GRN framework ultimately improves the model prediction accuracy, as introns are hypothesized to have regulatory influence over TG expression[25,89]. I then weighted each TFBS based on the number of Hi-C contacts(1Kb) it makes with the TG's promoter(**Figure 3.1**) using the weighting scheme described in **2.2B8**. Using such a

weighting scheme helps to capture regulatory information provided by long distance interactions of distal TFBS with TG promoters created via chromatin looping while preserving the influence of proximal promoter based TFBS as I have shown in **2.2C3**. I created a weighted motif network using the unique TF-TG interactions and the average Hi-C weight for them.

PPI network: I downloaded PPI data from the BioGRID database(*version.3.5.188*) to generate the PPI network.

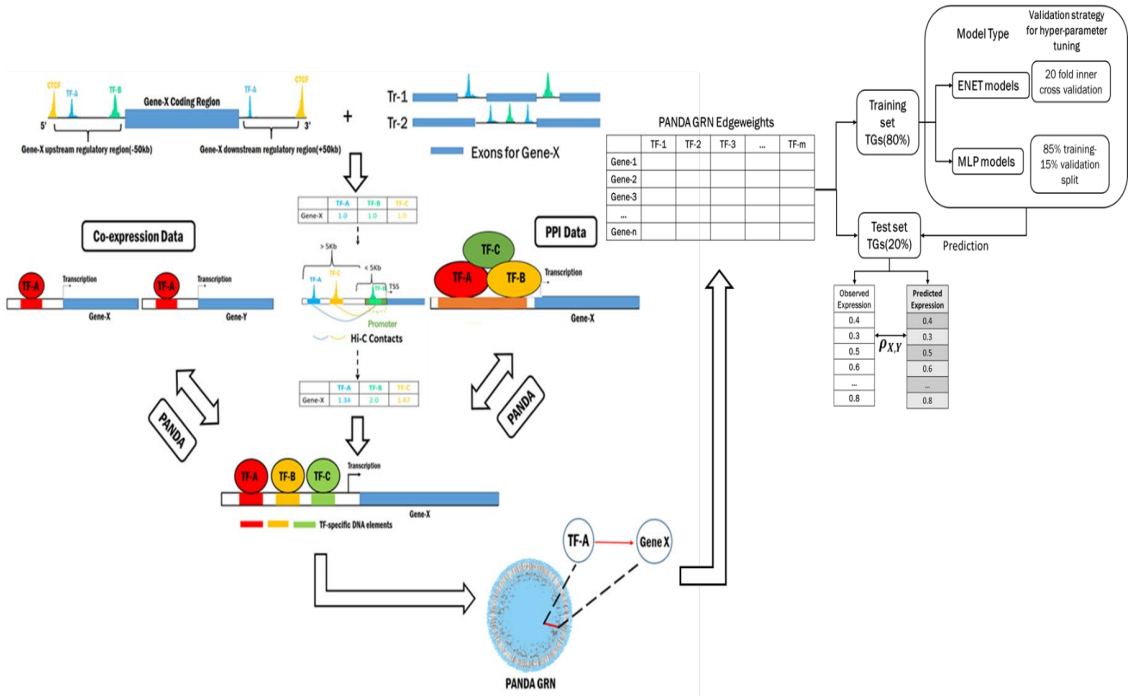

Figure 3.1: *Using a multi-omics GRN framework to predict gene expression based on MLP models for TRM detection. I downloaded ChIP-seq data for 149 GM12878 TFs from the ENCODE consortium. I used the peaks that passed the optimal IDR(Irreproducible Discovery Rate) threshold defined by the consortium and mapped them onto the regulatory region of each gene to define TFBs. I used CTCF peaks within a 50Kb window upstream and downstream of the gene body in order to demarcate the regulatory boundaries. Furthermore, I weighted the TF-TG interactions based on the number of contacts made by the corresponding peaks with the promoter of TGs. I used a weighting scheme where promoter TFBS were automatically up-weighted because of the inability of HiC data to capture them due to limited resolution. I created PANDA GRNs using the weighted adjacency matrices, the PPI data corresponding to the TFs obtained from BioGRID and the lymphoblastoid co-expression data obtained from GEUVADIS. After generating the PANDA GRN, I built elastic net(ENET) and multilayer perceptron(MLP) models that used them as input features to predict log FPKM values(gene expression) of an independent dataset. I used two different internal cross-validation strategies to train the ENET and the MLP models and assessed their accuracy by computing Pearsons correlation coefficient(PCC) between observed and predicted expression.*

Co-expression network: I extracted expression residuals for the 462 LCL samples within the GEUVADIS datasets using a genome-wide genetic relationship matrix(GRM) based mixed-linear regression model and used them to generate the co-expression network as described in **2.2B1**.

I used the above networks to generate GRN utilizing the R(*version.3.4.2*) implementation of the PANDA algorithm. After 25 iterations, I obtained convergence by setting the threshold for Hamming's distance at 0.001 and by using the value of 0.1 for the update parameter.

3.3B3: MLP network architecture and building the prediction models

I utilized two different MLP architectures in my paper: 1)MLP-U(MLP-Univariate) and 2)Traditional MLP as shown in **Figure 3.2**. The MLP-U architecture contained individual univariate MLPs receiving inputs corresponding to each TF in addition to the traditional MLP. All the univariate MLPs had 3 layers containing 10 nodes each and the traditional MLP also contained 3 layers with 800, 500 and 1000 nodes for each model. The non-linear activation function for all the layers was Rectified Linear Unit(ReLU).

I built the ENET and the MLP prediction models using log10 FPKM expression values of 11,780 protein coding TGs, where I used 80% of the data(9,424 TGs) for training the models and the remaining 20%(2356 TGs) to test the models and assess their prediction accuracy. I used two different internal cross-validation strategies to train the two types of models: 1) For the MLP-U and MLP models, I further divided the training data into 85% training and 15% validation sets. I then trained these models using the backpropagation algorithm. Additionally, I summed the output from all the individual univariate MLPs and the traditional MLP at the last node for training the MLP-U models. I note here that the

traditional MLP architecture was only used as a comparison in the paper and most of the analyses were done using the trained MLP-U models. 2) For the ElasticNet(ENET) prediction model, I used an alpha of 0.5 and trained the models based on 20 fold inner cross-validation. I trained and tested the models for 20 iterations(**Figure 3.1**), and computed Pearson's Correlation Coefficient(PCC) each time to assess model performance.

Thus, I had an input matrix **X** of size N × T, containing N TGs and T TFs. The values in this matrix were scaled edge-weights corresponding to the vertex $TF_t \longrightarrow TG_n$, where n ϵ {N} and t ϵ {T} derived from the learned PANDA GRN network. The output was a column vector **y** of size N containing scaled and centered log FPKM (Fragments per kilobase per

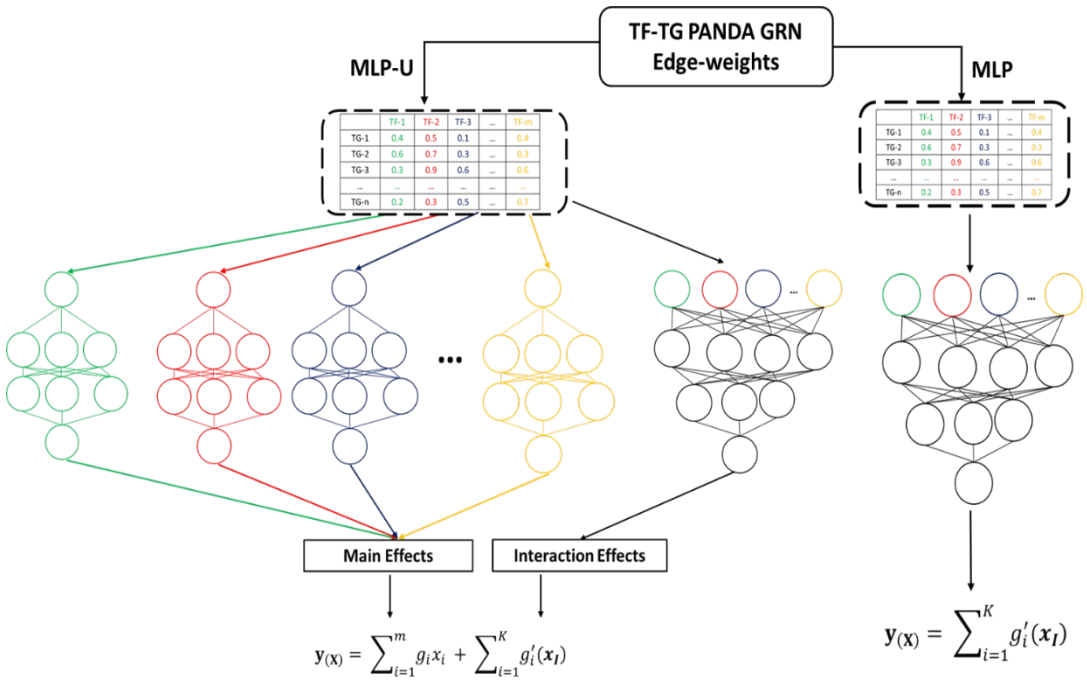

Figure 3.2: *MLP-U and MLP architecture used in Chapter 3. I trained two different types of MLP architectures in the paper: MLP-Uor MLP-Univariate(Left) and traditional MLP(Right). For the MLP-U models, I utilized individual TF edge-weights as inputs for the corresponding MLPs and trained them together with a traditional MLP receiving inputs corresponding to all the TFs. This ensemble model represented a generalized additive model where the main effects were derived from each individual MLP while the interaction effects were derived from traditional MLP. The traditional MLP model was trained without the individual MLPs and was assumed to just model the interaction effects. The modelling process involved partitioning the TG expression dataset into test and training sets using a 20-80 split and then training the models for 20 iterations. For the MLP-U model, I trained the individual univariate MLPs and the traditional MLP together via backpropagation.*

million) expression values of the N genes. For the MLP-U models, it was derived based on a generalized additive model:

$$\mathbf{y}_{(\mathbf{X})} = \sum_{i=1}^{T} g_i x_i + \sum_{i=1}^{K} g_i'(\mathbf{x_I}) \quad (3.1)$$

3.3B4: Obtaining main and interaction effects from the MLP-U models

For each trained MLP-U model, I performed an additional 5-fold prediction task in order to capture the prediction performance over all the TGs within each iteration. Thus, I essentially conducted 100 prediction rounds for which I stored the model weights learned during the training process.

In order to calculate the main effect corresponding to each TF, I utilized the learned MLP-U models. Specifically, I extracted layer weights from each one of the univariate MLP corresponding to each TF feature and aggregated them across all the prediction iterations. These iterations corresponded to a set S of 20 random numbers s each representing an instance/state for bootstrapping test set genes for each prediction task.

$$S = \{ s \mid s \in \mathbb{R}, k > 0 \} \text{ and } |S| = 20 \quad (3.2)$$

For each random state s, I picked 5 non-overlapping sets of test genes

$$G_{si} = \{g_{si} | g_{si} \in N\};\ s \in S;\ 1 \leq i \leq 5\ ; \quad (3.3)$$

$$|G_{si}| = \frac{|N|}{5}$$

For each G_{si} , I then used the remaining genes as the training set G_{si_train} such that

$$G_{si\_train} = \{g_{si\_train} | g_{si_{train}} \in N\} ; G_{si} \not\subset G_{si\_train} \quad (3.4)$$

I then predicted the expression values of G_{si} genes according to the following equation:

$$\mathbf{y}_{(\mathbf{X_{si}})} = \sum_{t=1}^{T} \Phi_{M_{sit}} x_{sit} + \Phi_{M_{siK}}(\mathbf{x_K}) ; \quad (3.5)$$

$$s \in S;\ 1 \leq i \leq 5$$

Here $\mathbf{y}_{(\mathbf{X_{si}})}$ is the vector containing predicted expression for gene set G_{si} using the input matrix $\mathbf{X_{si}}$ by the model trained using the input from genes in set G_{si_train}. The first part of equation (5) captures the main effect of each one of the TF t with M_{sit} representing the corresponding univariate MLP while the second part captures the interaction effect of K interactions, via the traditional MLP M_{siK},on the gene expression trait. Thus, for each iteration s, the expression vector of gene set G_{si} $\mathbf{y}_{(\mathbf{X_{si}})}$ is derived from a generalized additive model M_{si} containing main effects and interaction effects derived from a collection of complex non-linear functions $\Phi_{M_{sit}}$ and $\Phi_{M_{siK}}$ respectively. The parameters for this model were learned during the training process using the training set G_{si_train}. Furthermore, I had 5 models for each random iteration each containing a different set of test genes.

$$M_s = \{M_{si} | 1 \leq i \leq 5\};\ |M_s| = 5 \quad (3.6)$$

The architecture for each model, w.r.t the number of hidden units in each layer and the number of hidden layers was similar. Each model M_{sit} and M_{sik} contained **L** hidden layers, and there were p_l units/neurons in the l-th layer. The input layer for the univariate MLP M_{sit} was the vector $\mathbf{x}_{sit}$ containing edge-weights for TGs corresponding to TF t ($p^0_{M_{sit}} = \mathbf{x}_{sit}$). On the other hand, the input layer for the traditional MLP M_{siK} was the matrix $\mathbf{X_{si}}$ containing the edge-weights corresponding to all the TFs($p^0_{M_{siK}} = \mathbf{X_{si}}$). In each model, there were **L** weight matrices containing the weights learned during the training process such that $\mathbf{W}^{(l)}_{M_{sit}}, \mathbf{W}^{(l)}_{M_{siK}} \in \mathbb{R}^{\mathbf{p_l} \times \mathbf{p_{l-1}}}, l = 1,2,..L$ and $L+1$ bias vectors $b^{(l)}_{M_{sit}}, b^{(l)}_{M_{siK}} \epsilon \mathbb{R}^{pl}, l = 0,1,2..L$. Furthermore, there is a non-linear activation function $\phi(.)$ associated with each unit and weights $\mathbf{w}^{y}_{M_{sit}}$, $\mathbf{w}^{y}_{M_{siK}}$ and biases $b^{y}_{M_{sit}}$, $b^{y}_{M_{siK}}$ associated with the output layer for each model. The hidden units$\mathbf{h}^{(l)}_{M_{sit}}$, $\mathbf{h}^{(l)}_{M_{siK}}$ and the outputs $y_{M_{sit}}$, $y_{M_{siK}}$for the models can be mathematically described as :

$$\mathbf{h}^{(0)}_{\mathbf{M_{sit}}} = \mathbf{x_{sit}}; \mathbf{h}^{(0)}_{\mathbf{M_{siK}}} = \mathbf{X_{si}};$$

(3.7)

$$\mathbf{y_{M_{sit}}} = \left(\mathbf{w}^{\mathbf{y}}_{\mathbf{M_{sit}}}\right)^{\mathbf{T}} \mathbf{h}^{(\mathbf{L})}_{\mathbf{M_{sit}}} + \mathbf{b}^{\mathbf{y}}_{\mathbf{M_{sit}}};\ \mathbf{y_{M_{siK}}} = \left(\mathbf{w}^{\mathbf{y}}_{\mathbf{M_{siK}}}\right)^{\mathbf{T}} \mathbf{h}^{(\mathbf{L})}_{\mathbf{M_{siK}}} + \mathbf{b}^{\mathbf{y}}_{\mathbf{M_{siK}}}$$

(3.8)

$$\mathbf{h}^{(\mathbf{l})}_{\mathbf{M_{sit}}} = \boldsymbol{\phi}\left(\mathbf{W}^{(\mathbf{l})}_{\mathbf{M_{sit}}} \mathbf{h}^{(\mathbf{l-1})}_{\mathbf{M_{sit}}} + \mathbf{b}^{(\mathbf{l})}_{\mathbf{M_{sit}}}\right) \quad \mathbf{h}^{(\mathbf{l})}_{\mathbf{M_{siK}}} = \boldsymbol{\phi}\left(\mathbf{W}^{(\mathbf{l})}_{\mathbf{M_{siK}}} \mathbf{h}^{(\mathbf{l-1})}_{\mathbf{M_{siK}}} + \mathbf{b}^{(\mathbf{l})}_{\mathbf{M_{ki}}}\right),$$

(3.9)

$$\forall \mathbf{l} = \mathbf{1}, \mathbf{2} \ldots \mathbf{L}.$$

I note here that the **L** = 3 for all the models in my case.

I utilized the learned models M_{sit} and M_{siK} to calculate the main effect for each TF t and the interaction effect of K interactions respectively. I used an extension of the neural interaction detection(NID) developed by Tsang et al. in order to compute these effects[98].

For each random state s, I first aggregated the layer weights across all the models

$$\overline{\mathbf{W}_{\mathbf{st}}^{(\mathbf{l})}} = \frac{\mathbf{1}}{|\mathbf{M_s}|}\sum_{\mathbf{i=1}}^{\mathbf{5}} \mathbf{W}_{\mathbf{M_{sit}}}^{(\mathbf{l})} \; ; \; \overline{\mathbf{W}_{\mathbf{sK}}^{(\mathbf{l})}} = \frac{\mathbf{1}}{|\mathbf{M_s}|}\sum_{\mathbf{i=1}}^{\mathbf{5}} \mathbf{W}_{\mathbf{M_{siK}}}^{(\mathbf{l})}$$

(3.10)

$$\overline{\mathbf{w}_{\mathbf{st}}^{\mathbf{y}}} = \frac{\mathbf{1}}{|\mathbf{M_s}|}\sum_{\mathbf{i=1}}^{\mathbf{5}} \mathbf{w}_{\mathbf{M_{sit}}}^{\mathbf{y}} \; ; \quad \overline{\mathbf{w}_{\mathbf{sK}}^{\mathbf{y}}} = \frac{\mathbf{1}}{|\mathbf{M_K}|}\sum_{\mathbf{i=1}}^{\mathbf{5}} \mathbf{w}_{\mathbf{M_{siK}}}^{\mathbf{y}}$$

(3.11)

$$1 \leq i \leq 5, \forall l = 1,2 \ldots L$$

Here, $\overline{\mathbf{W}_{st}^{(l)}}$, $\overline{\mathbf{W}_{sK}^{(l)}}$ and $\overline{\mathbf{w}_{st}^{y}}$, $\overline{\mathbf{w}_{sK}^{y}}$ represent the weights of each hidden layer and the output layers respectively averaged across all the models in M_s.

The main effect for each TF t and the interaction effect of the TF_m-TF_n interaction at unit j of the first layer across all the models for a random state s was calculated using the following equations:

$$w_{st} = z_{st}^{1}\overline{w_{(st)}^{1}} \tag{3.12}$$

$$\mathbf{w_{j(sK:m,n)}} = \mathbf{z_{j(sK)}^{1}}\ \mathbf{min}\left(\left|\overline{\mathbf{w}_{\mathbf{j(sK:m)}}^{\mathbf{1}}}, \overline{\mathbf{w}_{\mathbf{j(sK:n)}}^{\mathbf{1}}}\right|\right) \tag{3.13}$$

Here, $w_{(st)}$ is the main effect of the transcription factor t obtained from the first layer of univariate model corresponding to random state s, $\overline{w^1_{j(st)}}$ is the mean weight of all the connections made by the input node in the first layer. Similarly, $w_{j(sK:m,n)}$ is the interaction effect for the interaction between TF_m and TF_n at the hidden unit j of the first layer aggregated across all the models in M_s and $\overline{w^1_{j(sK:m)}}$ and $\overline{w^1_{j(sK:n)}}$ are the aggregated weights corresponding to the connections(indices) of TF_m and TF_n respectively at node j. z^1_{st} and $z^1_{j(sK)}$ represent the influence of the input node and the hidden unit j respectively, which are calculated using the following formulae:

$$\mathbf{z^1_{j(sK)} = \left|\overline{w^y_{sK}}\right|^T \left|\overline{W^{(L)}_{sK}}\right| \left|\overline{W^{(L-1)}_{sK}}\right| \dots \left|\overline{W^{(1)}_{j(sK)}}\right|, j \in p^1} \tag{3.14}$$

$$z^1_{(st)} = \left|\overline{w^y_{st}}\right|^T \left|\overline{W^{(L)}_{st}}\right| \left|\overline{W^{(L-1)}_{st}}\right| \dots \left|\overline{W^{(1)}_{(st)}}\right| \tag{3.15}$$

The aggregated weight of interaction between TF_m and TF_n across all the nodes in the first layer was calculated using the following equation:

$$w_{(sK:m,n)} = \left| \sum_{j=1}^{|p^1|} w_{j(sK:m,n)} \right| \tag{3.16}$$

This step was not necessary for the main effects calculation since I only had one input node in each univariate MLP corresponding to each TF.

Since I averaged the calculations over all the models that contained different sets of test genes for each random state, I assumed that $w_{(sK:m,n)}$ and w_{st} represented average interaction effect between TF_m and TF_n and average main effect of TF t respectively over all the genes. I then averaged this effect over all the random states to produce the final NID interaction effects and main effects:

$$w_{(K:m,n)} = \frac{1}{|S|}\sum_{s \in S} w_{(sK:m,n)} \tag{3.17}$$

$$w_{(t)} = \frac{1}{|S|}\sum_{s \in S} w_{(st)} \tag{3.18}$$

3.3B5: Calculating TF average ENET main effects.

I calculated the average effect estimate for TF T $\bar{\beta}_T$ using equation (2.2) as described in **2.2B6**.

3.3B6: Detecting co-binding TF ChIP-Seq peaks

In order to identify statistically significantly co-binding pairs of TF ChIP-Seq peaks, I utilized the SpaMo algorithm (meme suite version 5.1.1)[29], which looks for significantly enriched spacings between a primary motif and a secondary motif by within a set of sequences. I isolated all the overlapping peak pairs corresponding to the 32 pairwise TF modules present within the TG's cis-regulatory regions. I centered and modified these regions so that they are no longer than 500bp, which is the required size for sequences for SpaMo. I utilized the position weight matrices(PWMs) downloaded from HOCOMOCO(*version.11*)[99] and JASPAR(*version.2020*)[4] in order to scan the sequences

for motifs corresponding to TFs in each pairwise TRM. I ran the SpaMo command line version and extracted peak pairs representing co-localizing TFs at a p-value threshold of 0.05.

3.3B7: Detecting TF ChIP-Seq peaks interacting via chromatin looping

I used the Hi-C data downloaded for GM12878(GEO accession: GSM1551688) in order to look for TF peaks interacting via chromatin looping. I used data corresponding to 1Kb and 5Kb resolution, and overlapped the peak pairs of pairwise TRMs with the Hi-C contact points. I also generated a random set of peak pairs corresponding to pairwise TFs not forming a pairwise TRM representing the background set for performing $\chi 2$ test of enrichment. I tested for enrichment of Hi-C contacts within the peak pairs corresponding to the TRMs detected using this test at a p-value threshold of 0.05.

3.3C: Results

3.3C1: Target gene expression could be better predicted by modelling complex non-linear interactions among transcription factors.

I hypothesized that information beyond sequence co-localization of TFs would be useful for detecting TRMs, formed by the "TF collective" model described in **3.3A** essential for TG expression regulation. As a basis to examine this hypothesis, I developed a multi-omics machine learning framework shown in **Figure 3.1**, which modelled the influence of multiple TF based regulatory mechanisms (TF co-operativity, TF-TG binding and TF-TG co-regulation) on TG expression, in the GM12878 LCL, by using features derived from a gene regulatory network(GRN) built using the PANDA algorithm[76]. I have

shown previously in **2.2C1** that features obtained from such a GRN explain more variance in TG expression compared to using TF-TG binding information alone.

In order to generate these GRNs, I first extracted all the cis-regulatory and intronic TF binding sites(TFBS) corresponding for 149 TFs for each TG. Next, I created an adjacency matrix weighted by the number of chromatin looping interactions between each TF and the TG promoter based on GM12878 high throughput chromatin capture(Hi-C) data. I used this weighted adjacency matrix to create a motif network reflecting GM12878 specific co-expression and PPI networks.

I used the TF-TG edge-weights from this GRN to build I prediction models where I tested for purely additive as well non-additive influence of TF features on TG expression. As a baseline comparison, I first built ElasticNet(ENET) regularized regression models which assume an additive(linear) influence of TF features over TG expression without considering any non-additive effects of TF interactions. I next used a traditional neural network based multilayer perceptron(MLPs) capable of modelling non-additive(non-linear) interaction effects of TFs on TG expression, which I hypothesized will help identify "TF collective" based TRMs. I further evaluated a hybrid model (MLP-U) that can decompose the effects into additive and non-additive components. This model was composed of set of univariate MLP models capturing individual TF influence over TG expression along with a traditional MLP to capture all possible interaction effects (see **Figure 3.2**). Thus, I used 3 different prediction models :1) an ENET to model TF main effects only 2) an MLP to model complex interaction effects, and 3) an MLP-U model that can be decomposed into additive and non-additive components. Further details about these models, especially the MLP and MLP-U architectures have been provided in **3.3B3** .

I used an independent GM12878 LCL expression dataset(accession: ENCSR889TRN) to train and test the prediction models. I used 80% of the total TG set for training the models as well as for internal cross validation and used the remaining 20% for testing the prediction accuracy (**Figure 3.1**). I performed the prediction task for 20 iterations, each time using a different set of test TGs. I used the median Pearson's correlation coefficient(PCC) aggregated across all of these iterations to compare the performance of the 3 different models. As shown in **Figure 3.3A**, the models capable of capturing interaction effects of TFs (MLP and MLP-U) perform significantly better (median $PCC_{MLP} = 0.68$; median $PCC_{MLP\text{-}U} = 0.63$) compared to the ENET models (median $PCC_{ENET} = 0.57$), which model linear influence of individual TFs on gene expression. This improvement in performance was also statistically significant (median PCC_{MLP} vs. median PCC_{ENET} p-value = 1.91e-06; median $PCC_{MLP\text{-}U}$ vs. median PCC_{ENET} p-value = 1.91e-06) as calculated by performing paired Wilcoxon sign rank tests.

The MLP-U model architecture shown in **Figure 3.2** contained individual univariate MLPs corresponding to each one of the 149 TFs receiving input features from these TFs as well as a traditional fully connected MLP receiving input features corresponding to all the TFs at the same time. Each MLP-U model was trained using the generalized additive equation shown in **Figure 3.2** containing the univariate main effects as well as the interaction effects

used to predict TG expression. In order to partition the variance in TG expression explained by these two components of the MLP-U models, I followed the following steps: 1) I

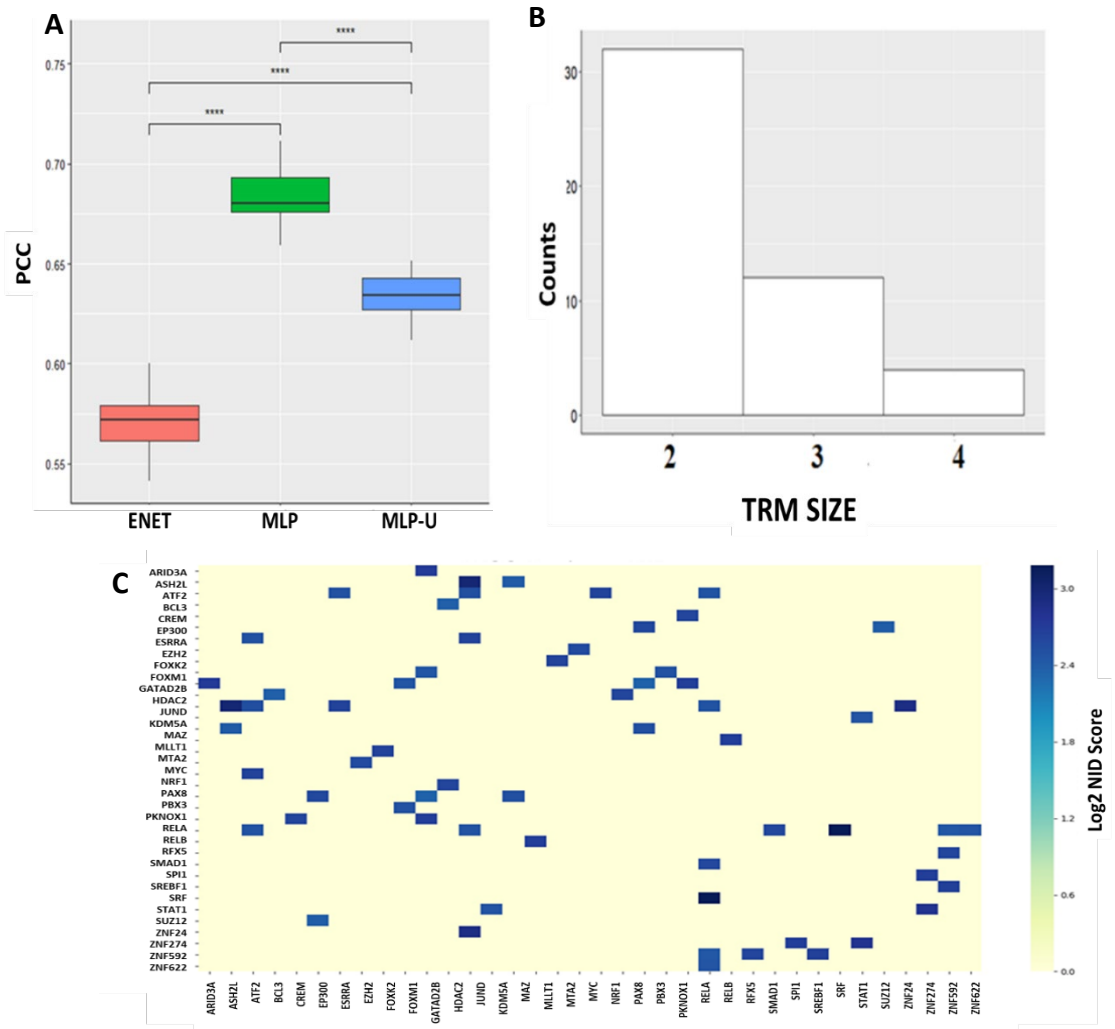

Figure 3.3: *Learning global transcriptional regulatory patterns from multi-omics GRN based machine learning approach. A) Boxplot showing the performance of the MLP, MLP-U and ENET models obtained from the prediction of 2356 TGs over 20 iterations (***p-value < 0.0001). B) Barplot showing the sizes of the 48 TRMs detected from the learned MLP-U models based on the NID algorithm. B) Heat-map showing the strength of the interactions for 32 pairwise TFs calculated based on the log2 NID scores.*

obtained the learned MLP-U models trained from predicting TG expression for the 20 random states and extracted weights corresponding to the layers of the univariate MLPs as well as of the traditional MLP for each model. 2) I used the layer weights to generate two separate MLP models based on univariate and traditional MLP neural networks. 3) Lastly, I used these two models to independently predict expression for the test set TGs defined

based on the random states and for each prediction task I calculated the R^2 that reflected the variance explained in the TG expression for the two models. I plotted these R^2 for each prediction round for the two models in **Figure 3.4.** I observed that over all the prediction iterations for the MLP-U models, the main effects, obtained from their univariate component, were more predictive of TG expression, explaining about 34% of the variance on average, than the interaction effects, captured by their MLP component, which explained 23% of the variance in TG expression.

In conclusion, accurate prediction of TG expression requires efficient modelling of main effects of individual TFs as well as interaction effects of TRMs.

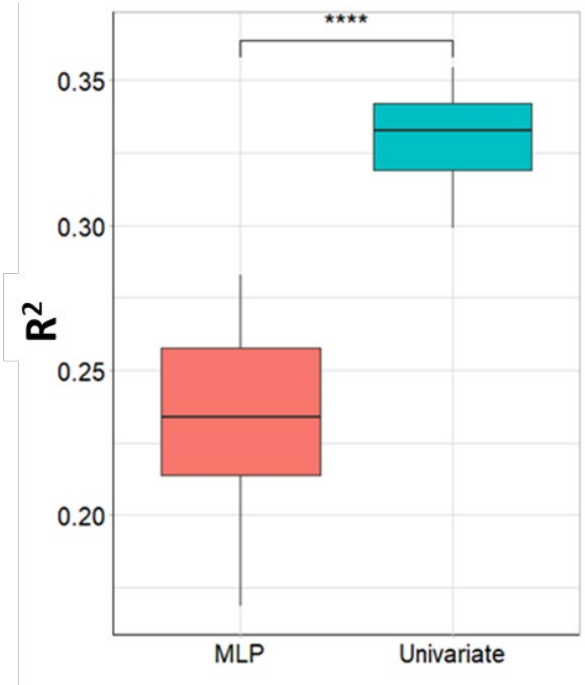

Figure 3.4: *Boxplots showing the amount of variance in TG expression explained by the two components of the MLP-U models.The boxplot shows the prediction performance calculated in the form of R^2(variance explained) by predicting TG expression(N = 2,356) over 20 iterations using the univariate and the MLP components of the learned MLP-U models. (****p-value < 0.0001 calculated using paired t-test)*

3.3C2: Context dependent influence of individual transcription factors on target gene expression could be discerned from my models

The MLP-U architecture, described in **Figure 3.2**, allowed us to model main effects of individual TFs separately from the interaction effects of TF combinations. I used equations(3.1)-(3.18) to calculate these main and interaction effects from the trained MLP-U models(see **3.3B4**).

My learned MLP-U models contained individual univariate MLPs corresponding to each one of the 149 TFs. I aggregated all the learned connection weights at the first layer of these MLPs and multiplied them with the nodal influence score for each node in that layer. After averaging these nodal scores, I calculated an average main effect for each TF across all the prediction iterations followed by scaling it in the range (-1,1).

In order to examine the validity of my main effects aggregation approach, I divided the TFs into 5 bins based on their scaled main effects (**Figure 3.5**). The bin placement of the TFs derived from their main effects reflected their functional roles. For instance, activating TFs such as TAF1, MYC, TBLXR1, RELA and BCL11A were present in the right most bin(5) because of their highly positive main effects. On the other hand, transcription repressors such as MXI1, HDAC2, SMC3, MAZ and ZNF592 had strongly negative main effects placing them in the left most bin (1). I compared the main effects obtained from the MLP-U models to those obtained from the ENET model by computing the difference in ranks(DIR) of the TFs based on their effects for the two modelling approaches. Positive DIR for a TF reflected decrease in the MLP-U main effect, while a negative DIR represented increase in the MLP-U main effect, compared to that obtained from the ENET models.

I found that TFs with extremely negative DIR based on their MLP-U main effects, such as ZNF143(-128), TBLXR1(-121), DPF2(-115), E4F1(-115) and YY1(-110), were

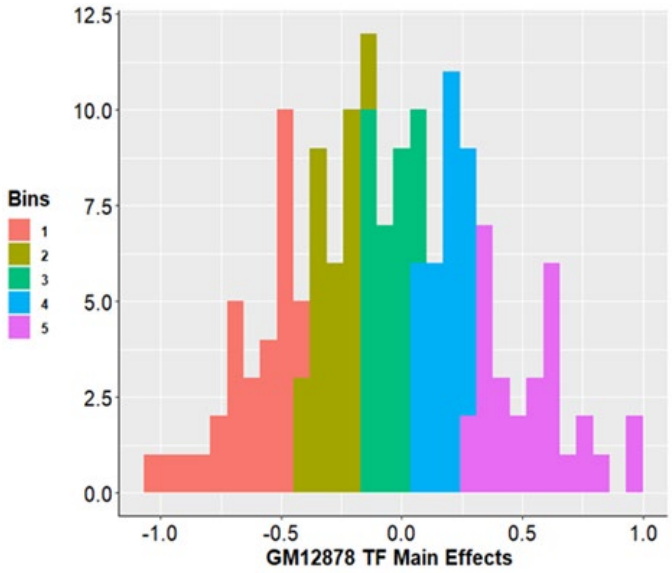

Figure 3.5: *Histogram of the scaled main effects for each TF obtained from the MLP-U models. These effects were calculated by aggregating the layer weights of the MLP-U models corresponding to each TF across the 20 iterations. The histogram was further divided into 5 equal bins based on the scaled main effects.*

transcriptional activators in specific contexts representing their interactions with other TFs[100–102]. Alternatively, TFs with extremely positive DIR , such as ZBTB40(112), HDAC2(112), SIN3A(124), SMAD1(98) and KDM1A(125) could act as repressors when interacting with other TFs[103–107]. I also found an extremely positive DIR for the well-known transcriptional activator TBP(125), which requires other promoter binding TFs such as the TBP-associating factors(TAFs) to recruit RNA polymerase II and to exert its effect[108]. Thus, while ENET models captured influence of TFs assuming independent effects on TG expression, main effects obtained from the MLP-U models are adjusted for context in which the TF binding event occurs (see equation (3.1)).

3.3C3: Interaction effects aided the detection of well-known and novel transcription factor regulatory modules.

The MLP component of the MLP-U models quantify the non-additive interaction effects of different combinations of TFs on TG expression. These effects could reflect the influence of non-linear "TF collective" interactions on TG expression. I applied the NID algorithm[98] to compute interaction effects in the form of NID scores for such TRMs. This calculation is done at each node of the first layer, for all the possible combinations/orders of the interactions and only the top ranked interactions for each order are retained. The interactions are aggregated such that lower order redundant interactions are removed and higher order top ranking interactions are retained giving a final set of highly impactful interactions of different orders. I defined these interactions along with their average NID scores as TRMs. I applied Log2 normalization to the average NID scores calculated for each TRM across all the 20 prediction iterations.

I detected 48 unique TRMs out of which 32 were pairwise interactions, 12 were 3-way and 4 were 4-way interactions as shown in **Figure 3.3B**. The pairwise TRMs were formed by 36 unique TFs, 3-way TRMs were formed by 22 TFs and the 4-way TRMs were formed by 12 TFs. Furthermore, I observed that among the higher order (3-way or higher) TRMs, the "nestedness" or the proportion of all the possible pairwise TRMs being also detected was never 100%(**Figure 3.6**). I found that JUND formed the largest number of TRMs(11) followed by GATAD2B(10), RELB(10) and ATF2(9). All of these TFs are versatile DNA binding proteins capable of affecting cell proliferation, division and apoptosis, which explains their presence in a large number of TRMs.

Multiple literature annotated TF interactions were present in the TRMs I detected.

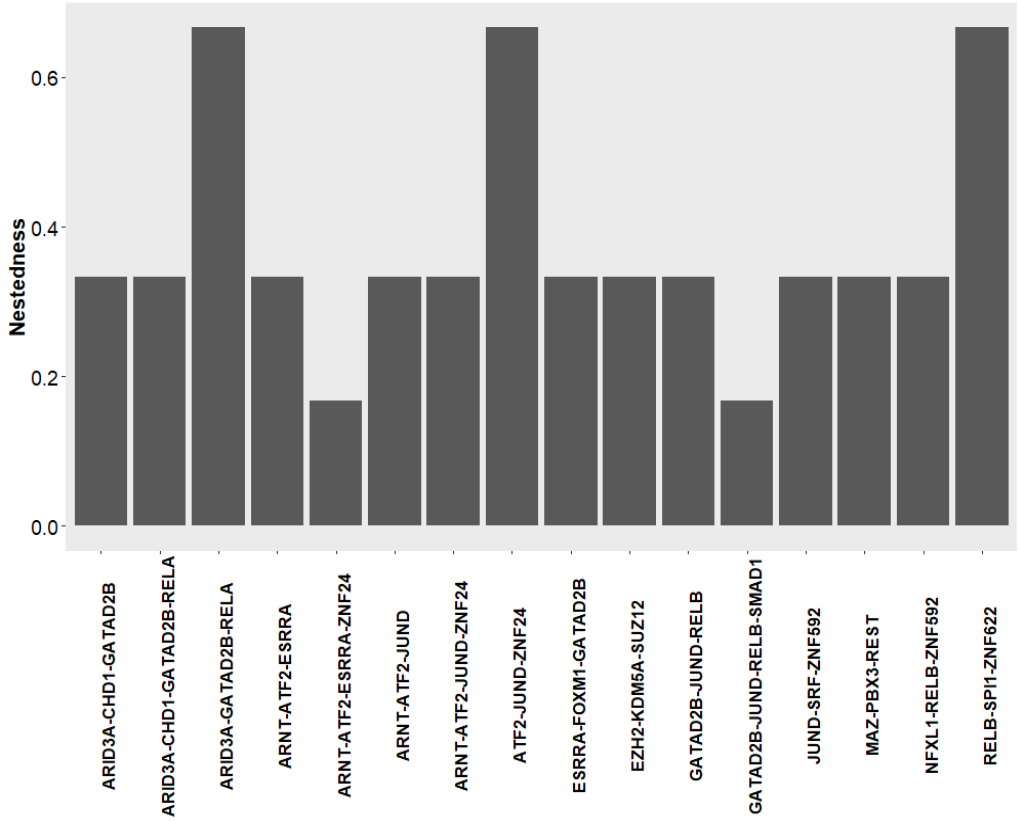

Figure 3.6: *Barplot showing the nestedness for each higher-order(three-way or higher) TRM. It was calculated as the proportion of all the possible pairwise interactions being also present in the detected set of TRMs. The nestedness of none of the higher order TRMs was 100%. In other words, none of the higher order TRMs could be completely explained by their subset pairwise interactions.*

For instance, the pairwise TRM of ATF2-JUND(Log_2NID score = 2.57) where both the TFs are part of the well-known AP-1 factor complex, which is involved in expression regulation of multiple TGs[109–111]. TF GATAD2B is known to form a repressive complex involving nucleosome remodeling and deacetylase activity with the CHD family of TFs[112]. I discovered that GATD2B and CHD1 were present in two different TRMs: ARID3A-CHD1-GATAD2B(Log_2NID score = 2.64) and ARID3A-CHD1-GATAD2B-RELA(Log_2NID score = 2.63). The presence of ARID3A and RELA in these TRMs has not been validated by the existing literature, although both of them have been associated with immune cell proliferation[113,114]. I also discovered the three way TRM EZH2-KDM5A-

SUZ12(Log_2NID score =2.40, where the methyltransferase EZH2 and scaffolding protein SUZ12 are known to form the polycomb-repressive complex PRC2 , which interacts and competes with H3K4me3 demethylase KDM5A during the process of angiogenesis and hematopoiesis[115]. I also discovered the pairwise TRM KDM5A-SUZ12(Log_2NID score = 2.47) indicating that KDM5A and SUZ12 may be the primary interactors within the three-way TRM.

I also detected several TRMs containing previously uncharacterized TF interactions. For example, the TRM with the highest influence over TG expression was RELB-STAT1 with the largest Log_2NID score of 3.18. Both of these TFs play an important role in immune response and lymphocyte development [116,117]. Thus, their closely related functions could point to the possibility of their interaction in vivo. Another intriguing, albeit unvalidated interaction, that I discovered was EP300-TAF1(Log_2NID score = 2.39). Both of these TFs are well known lysine acetyltransferases and are responsible for activating and regulating transcription of several TGs and were also found to have the highest frequency of oncogenic mutations among all the other lysine acetyltransferases[118]. The Log_2NID scores for all pairwise TRMs are shown in the form of a heat-map in the **Figure 3.3C**. Thus, I detected TRMs containing many previously uncharacterized as well as some well-known TF interactions using the NID algorithm.

3.3C4: Chromatin looping plays an essential role in forming transcription factor regulatory modules and in mediating their regulation of target genes.

Apart from some well-known interactions, I discovered TRMs contained a significant number of previously uncharacterized TF interactions. As described in **3.3A**, TRM formation can be brought about by either co-localization of proximally binding TFs

based on motif proximity or by distally binding TFs brought in close proximity by long distance chromatin looping. Thus, to characterize the TRMs, I used chromatin looping and TF motif co-occurrence information to identify TFs interacting with each other via chromatin looping (long-distance interactions) or by binding in close proximity(see **3.3B6** and **3.3B7**).

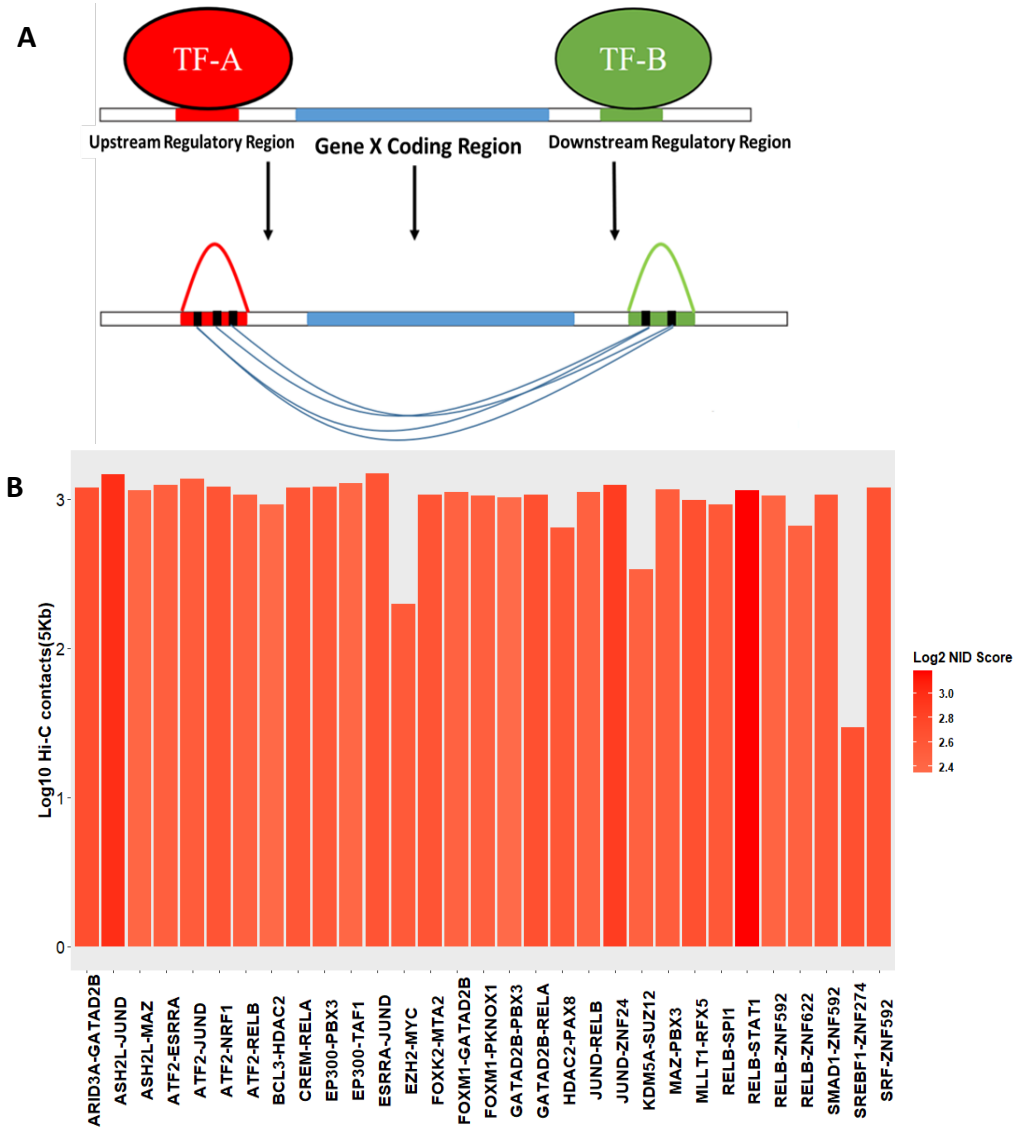

Figure 3.7: *Pairwise TRMs interact via long distance chromatin looping. A) We overlapped the GM12878 Hi-C data at 5Kb resolution with the ChIP-seq peak pair regions corresponding to the 32 pairwise TRMs within the cis-regulatory regions of the TGs. B)Barplot showing the mean log10 Hi-C contacts(5Kb resolution) between peak regions of the pairwise TRMs shaded according to the respective Log2 NID scores across all the TG. I wasn't able to detect any HiC contacts between the peak pairs of the TRM SUZ12-ZNF284*

Using GM12878 specific high throughput chromatin capture(Hi-C) data, I looked for long distance interactions between ChIP-seq peaks, present within TG's cis-regulatory regions, corresponding to all pairwise TF modules I detected (**Figure 3.7A**). I compared the enrichment of Hi-C contacts for these peaks with that obtained from a background set of peak pairs, within the TG's cis-regulatory regions, corresponding to random pairwise combinations of TFs not present in the detected set of pairwise TRMs using a chi-square test. I observed significant enrichment of Hi-C contacts at 5Kb resolution among 36,734 ChIP-seq peak pairs corresponding to 31 pairwise TF modules (χ^2 p-value = 9e-04) within TG's cis regulatory region as shown in **Figure 3.7B**. The only pairwise TRM that did not

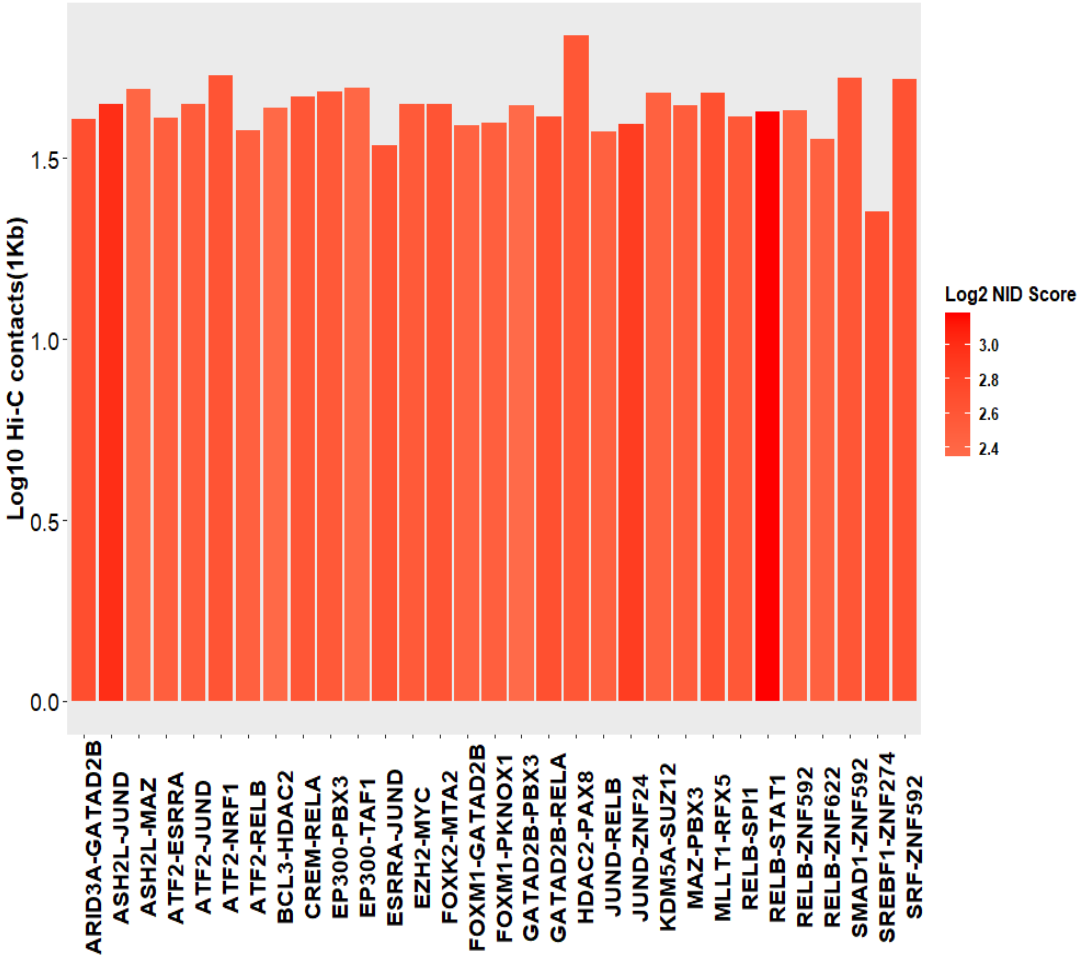

Figure 3.8: *Barplot showing the mean log10 Hi-C contacts(1Kb resolution) between peak regions of the pairwise TRMs. The plot is shaded according to the respective log2 NID scores across all the TG. I wasn't able to detect any HiC contacts between the peak pairs of the TRM SUZ12-ZNF284.*

contain any Hi-C contact points between the peak pairs was SUZ12-ZNF284. The enrichment of Hi-C contacts at 1Kb resolution was not statistically significant, however with the χ^2 p-value of 0.3423(see **Figure 3.8**).

In order to identify co-localizing TF interactions based on their sequence/motifs, I used the SpaMo tool from the MEME suite(*version.5.1.1*)[29] to examine pairwise TRMs,. I looked for significant spacing between TF motifs occurring within their overlapping peak pair regions. I found significant motif co-occurrence for 60 peak pairs corresponding to 6 pairwise modules (adjusted p-value < 0.05). Additionally, I did not find these co-binding TRMs in the set of modules previously described by other approaches[27,28,94,119].

To further characterize the regulatory architecture of the TRMs, I defined four transcription regulatory programs shown in **Figure 3.9A** based on their interactions with TG promoters. I first identified 2,038 TGs where TFs peaks were interacting with each other either via Hi-C or via motif co-occurrence. I then determined the regulatory programs

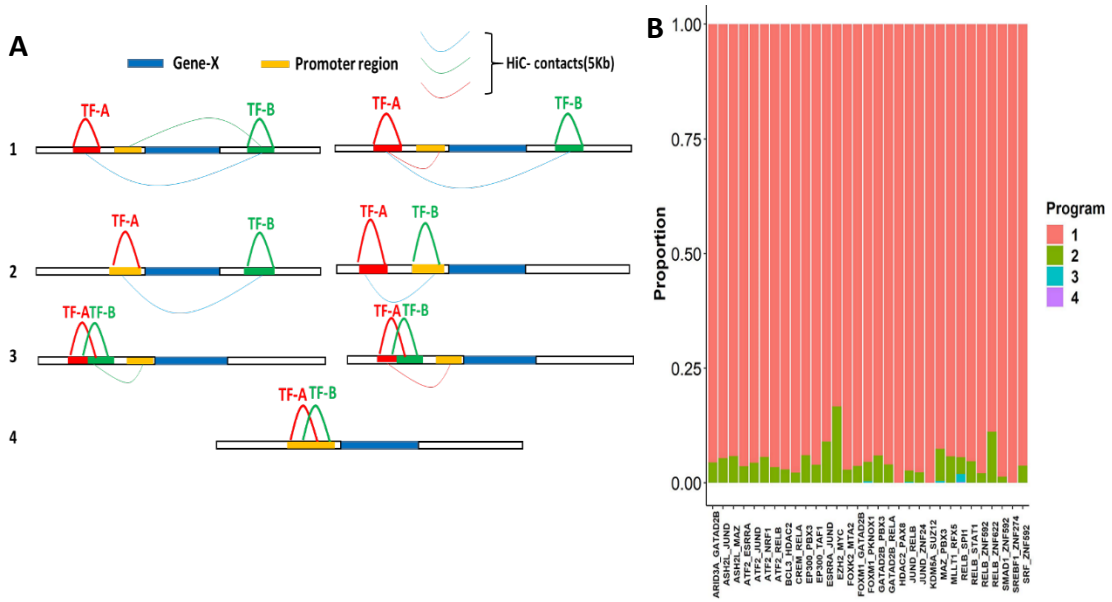

Figure 3.9: *Pairwise TF TRMs follow different regulatory programs for different TGs. A) We utilized HiC and co-binding data to define 4 TF regulatory patterns/programs for the pairwise modules for different TGs. B) Barplot shows the proportion of the total peak pairs for each pairwise TRM following each of the 4 transcription regulatory programs shown in A*

followed by the TRM peak pairs for each TG. As shown in **Figure 3.9B**, on an average

95% of the peak pairs corresponding to each pairwise TRM followed a configuration where at least one is interacting with the TG promoter and the two peaks interact with each other via long distance chromatin looping. Furthermore, TRMs HDAC2-PAX8, KDM5A-SUZ12 and SREBF1-ZNF274, for which the TFs are not known to directly bind to the TG promoters, regulated all their TGs using this program exclusively. I observed that for the remaining TRMs, about 4.5% of the peak pairs followed the second regulatory program which constituted one of them being present directly within the TG promoter while interacting with the other one via chromatin looping. About 17% of the peak pairs corresponding to the TRM EZH2-MYC, which contained TFs with known TG promoter binding activity, followed this regulatory program. Lastly, I found only 25 co-localizing peak pairs corresponding to 4 pairwise modules (RELB-SPI1, JUND-RELB, FOXM1-PKNOX1 and MAZ-PBX3) interacting with the promoters of 15 TGs via chromatin looping and 1 instance of co-localizing peak pair for the TRM RELB-SPI1 directly binding the promoter of 1 TG. Hence, only RELB-SPI1, which contained TFs important for lymphocyte development, contained peak pairs following all four types of transcription regulatory programs.

Thus, based on the above analyses, I conclude that the pairwise TRMs identified from the MLP-U learned models almost exclusively contained TF peak interactions occurring over long distance via chromatin looping. In addition, these TRMs mostly regulated their TGs also via long distance chromatin interactions with the TG promoters.

3.3D: Discussion

In this chapter, I designed a machine learning prediction framework for identifying TRM for the GM12878 immortalized LCL utilizing multiple big “omics” data sources. I

used a modified form of the neural network MLP architecture called MLP-U in order to account for the influence of individual TFs as well as of TF interactions on TG expression within the same model. I found that accounting for both these effects resulted in more accurate TG expression prediction compared to accounting for just the linear effects of TFs using the ENET regularized regression models. The traditional MLP models produced better prediction than the MLP-U models because of the recapitulation of the main effects of TFs. In other words, both main effects and interaction effects were being modelled using complex non-linear functions in the traditional MLP architecture leading to perhaps an overestimation of the main effects resulting in the better TG expression prediction.

One of the biggest drawbacks of a neural network model is that it is usually considered a "black-box" as features learned during the training as well as testing of the models are difficult to interpret. I overcome this limitation and extracted biologically relevant information using the NID algorithm[98]. I calculated main effects of individual TFs as well as interaction effects of TF combinations. I observed that the direction of the TF main effects correlated well with their known functional roles. However, these effects were largely different compared those obtained from ENET models as the MLP-U captured context/interaction dependent TF main effects, while the ENET models estimate TF main effects only.

Furthermore, I also detected highly influential TF interactions forming TRMs via statistical interactions in models of TG expression. I derived literature-based annotations for some of these TRMs, while many were novel TF interactions not identified by other approaches. This could be due to two reasons. First, the non-additive non-linear nature of the TF interactions, reflecting the "TF collective model" I detected is fundamentally

different from that of the linear, co-localizing TFs, reflecting the "Enhanceosome" and "Billboard" models identified by the previous approaches. Second, I strategy for identifying TF interactions was to model their influence on TG expression, which was largely ignored by the previous approaches. Thus, a co-localizing set of TFs not significantly impacting TG expression would be missed using my approach, though these TFs presumably have little influence on the expression of nearby genes. Additionally, I found that a significant proportion of the TF peaks for the pairwise TRMs interacted with each other and with the promoters of the TGs they regulated via chromatin looping. Therefore, long distance chromatin interactions likely play a large role in formation of TRMs as well as in their regulation of the TGs. This further validates the idea that TF interactions are not limited to proximally binding co-localizing sets of TFs. I used Hi-C chromatin looping data in two mutually independent contexts; I first included Hi-C contacts made by distal TFs with the TG promoters while building I GRNs, and further validated these chromatin interactions by examining Hi-C contact enrichments between the TF peak regions themselves. While the former Hi-C data aggregation was done to quantify the influence of distally binding TFs on TG regulation via promoter interaction, the latter instance reflected characterization of pairwise TFs interacting over long distances.

I focused I analyses on the GM12878 LCL in this study due to the density of TF binding data available, however my approach is flexible enough to analyze TRM based TG regulation in other commonly studied human cell-lines when these data are available A key limitation of my approach is the need for high-density omics assay data that often require large input DNA quantities that likely limit their application to cell-lines only. In different cellular contexts and environmental conditions, additional higher order TRMs may exist,

and the precise models underlying these interactions will be difficult to elucidate. However, I did identify pairwise TF interactions that form a basis for higher order interactions that could act as a starting point for further experimental validation or examination under different environmental conditions.

CHAPTER 4: TFXCAN: A NOVEL TRANSCRIPTOME WIDE ASSOCIATION STUDY(TWAS) APPROACH BASED ON REGULATORY INFORMATION DERIVED FROM TRANSCRIPTION FACTORS AND THEIR INFLUENCE ON GENE EXPRESSION

4.4A: Introduction

As described in **1A** , TFs are the primary effectors of the process of gene expression regulation. Disruption in their binding, brought about by genetic variants, could lead to severe consequences for the cell and for the organism as described in **1C**. Thus, it is essential to model the effects of these variants on TFBS as well as on TG expression to gain mechanistic understanding of several diseases. The former task can be accomplished by using various algorithms described in **1C**, which aid in quantifying the effect of non-coding variants on TFBS. However, deriving the influence of TFBS altering non-coding variants on TG expression is not as straightforward.

As described in **1C**, apart from EpiXcan, none of the other TWAS methods can integrate functional information regarding the non-coding cis-variants in the prediction models. As a result of this integration, EpiXcan has been shown to produce more accurate TG expression models compared to PrediXcan. However, EpiXcan based functional annotations have the following disadvantages: 1) cis-eQTL summary statistics are used to obtain epigenetic priors for different chromatin states, which may not always be readily available. Additionally, the requirement to generate cis-eQTL effects beforehand may potentially introduce confounding in the downstream TG expression prediction models. 2) The epigenetic priors assigned to the variants are broad, in that all variants present within a certain regulatory element are assigned similar priors. However, finer functional annotation of the variants is necessary to model their distinctive effects on TG expression.

In this chapter, I describe a novel TWAS approach called TFXcan, which takes advantage of my integrative GRN based TF effect estimates derived in **2.2C2** as well as non-coding variant influence over TFBS calculated using complex neural network models

to build TG expression models. I extensively validated I neural network based variant scoring approach by comparing its performance for predicting TF binding intensity with a similar method called DeFine[120]. Additionally, I also compared the performance of my models for predicting TFBS alterations induced by non-coding variants with seven other popular methods used for TFBS variant annotation. Lastly, I used reference datasets such as DGN[86] and GTEx[49] to build TG expression prediction models and compared their accuracy with that obtained from EpiXcan based models. TFXcan utilizes biologically relevant information corresponding to TF driven transcriptional regulation, without needing the summary level cis-eQTL data used in EpiXcan. Additionally, since the variants are scored using deep learning models, functional fine-mapping of event the novel TFBS altering variants can be derived and utilized within the TFXcan framework.

4.4B: Materials and Methods

4.4B1: Training the AGNet neural network models

In order to predict the variant effects on TF binding, I trained neural network models, utilizing the attentive gated neural network (AGNet) architecture described by Guo et. al[121], that predict the TF binding intensity using DNA sequence information. I used processed ChIP-Seq peak data downloaded from ENCODE, described in **2.2B1**, corresponding to the GM12878 LCL in order to train the models. I only used data corresponding to autosomes (chromosomes 1-22) and aligned to the build GRCh37/hg19 human genome reference assembly for training the models. Following steps were involved in pre-processing the data before training the models:

1) For each TF, I first removed the peak regions corresponding to top 1% TF binding intensity, as they represent binding regions with low complexity[120]. I further applied Log10 normalization to the intensity values and scaled them in the range (0,1).
2) Furthermore, I set the maximum length of the peak regions at 2000bp and trimmed the longer regions, on both ends, to bring them to this length.
3) I then downloaded the 1x normalized DNAase-seq tracks for GM12878 trimmed down to the first nucleotide from the 5' used in the ENCODE dream challenge. Using this data, I filtered out the low accessibility peak regions corresponding to the bottom 10% with respect to DNAase-seq intensity.
4) The above step gave us a positive set of peak regions for training the models pruned by their DNA accessibility. In addition to these I created a negative set of genomic regions not bound by any TF, which followed the same length and chromosome distribution as that of the positive set for each TF. The ratio of the positive and negative regions was kept at 1.0.
5) Lastly, I downloaded build 37 sequences for each peak region and corrected them for the presence NA12878 genomic variants using the "FastaAlternateReferenceMaker" tool of gatk(*version: 4.1.9*) and corresponding VCF file (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/NA12878_HG001/latest/GRCh37/HG001_GRCh37_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-10X-SOLID_CHROM1-X_v.3.3.2_highconf_PGandRTGphasetransfer.vcf.gz). This corrected for the NA12878 specific variants present within each peak region.
6) The set of sequences generated above were then used to train the GloVe model[122] in order to learn vector embeddings/representation for each sub-sequence/k-mer. In

order to accomplish this, I divided each TFBS sequence into k-mers of length 6 and stride 2, which were then aggregated into a big corpus of k-mers containing sequence information regarding all the TFBS for 149 TFs. This corpus contained 4,258 unique k-mers/words. I used the github code for training the GloVe model(https://github.com/stanfordnlp/GloVe). I used the vector size of 100 and trained the GloVe model for 100 iterations.

After generating the GloVe vector embeddings for the k-mers, I utilized the AGNet architecture with the keras package(v 2.3.1) and the tensorflow (v 1.14.0) backend to train the neural network models. I trained individual models for each TF using the k-mer vector embeddings corresponding to its peak regions as input and the normalized scaled ChIP-Seq intensities as the output. Below I have described the architecture of AGNet briefly, and I refer the readers to the AGNet paper[121] for further details.

The first layer of the AGNet models was an embedding layer receiving an input in form of indexed k-mer vectors for each TF peak region. The embedding layer was followed by a layer of multi-scale CNN layer of three 1D CNNs all connected to the embedding layer and learning local informative features from the input k-mers in parallel. Each one of the convolutional layers contained 64 filters and the kernel size for them was 3,5 and 7. Each multi-scale CNN was followed by a max-pooling layer with a pooling size of 3.

Apart from the multi-scale CNNs, the embedding layer was also connected to a position embedding layer that contained vector embeddings for the k-mers derived based on their positions within each input sequence. The position embeddings were of the similar size(100) to that of the GloVe based k-mer embeddings and the two were added to produce

the output of the position embedding layer. The output of the position embedding is then passed on to a dual attention layer meant to extract important sequence features.

The outputs from the multi-scale CNNs were concatenated and were passed on to gated convolutional network (GCN) layer. The GCN layer consisted of a conventional 1D CNN with a "sigmoid" activation function and a novel 1D CNN with a scaled exponential linear unit(SELU) activation function for improved gating and control of information flow in form of features to the next layer. Both the CNNs in the GCN layer contained 192 filters and the kernel size of 3. The outputs from these two CNNs were multiplied and a maxpooling of size 3 was applied to product.

The GCNs were followed by two gated recurrent network(GRN) layers containing stacked bi-directional gated recurrent units(BIGRU). Each GRN layer contained 256 nodes and the output dimension was 128. The output from the GCN layer is directly passed on to both the GRNs, and it is also concatenated to their outputs.

The GRN-GCN concatenate output is passed onto another dual attention layer to extract important abstract features. These features are then concatenated to the ones obtained from the dual attention layer containing position embeddings. The concatenated sequence and abstract features are then passed on to a fully connected dense layer containing 128 nodes and the SELU activation function, which is then connected to an output node.

The parameters for the AGNet models are initialized using the Lecun normal initializer. Each AGNet model was trained using the mean squared error(MSE) loss and optimized using the Adam optimizer for a maximum of 100 epochs. Furthermore,

overfitting in each model was controlled using dropout layers, L2 regularization and early stopping after seeing no improvement in validation loss for 10 straight epochs.

Each TF peak region set containing negative and positive sets was divided into 70%-15%-15% training, validation and test set. After model training was finished, the TF binding intensities of the test set were predicted and were correlated with the actual binding intensities to evaluate the accuracy of the models.

4.4B2: Cross- cell type AGNet model training

In order to generate generalizable AGNet models, which could be applied to TF binding data corresponding to multiple different cell-types, I trained them using a cross-cell type training strategy. Specifically, I downloaded ChIP-Seq data from the Tier-1 cell lines (K562 immortalized chronic myelogenous leukemia cell line and H1 human embryonic stem cell line) and Tier-2 cell lines(HeLa-S3 immortalized cervical cancer cell line and HepG2 liver carcinoma cell line). I identified 20 TFs, whose data was available for these 4 cell lines and processed it using the steps described in **4.4B1**. I then used the processed data for cross-cell type training using the following procedure: I divided the total processed data for each TF for each one of the 4 cell lines into 70-15-15 training-test-validation set. I then pooled the validation set for all the 4 cell-lines together. I began training the pre-trained models for each cell-line for each TF using the pre-trained GM12878 AGNet model. I used data training data specific to each cell-line and the validation data from the other three cell-lines to train each TF model. Thus, I trained 4 different models, corresponding to the 4 cell-lines, for each TF where the training data was cell-line specific and the validation data was obtained from the other 3 cell lines. The training parameters for the cross-cell type training were the same as that used for the

GM12878 model training in **4.4B1**. I assessed the accuracy for each TF cell line model using the test data specific to that cell line.

4.4B3: Comparison of AGNet models with the DeFine models.

DeFine models are a set of CNNs trained using the ENCODE ChIP-Seq datasets to predict TF binding intensity[120]. I compared the accuracy of my AGNet models for predicting TF binding intensity to that obtained from the DeFine models. I first downloaded DeFine models corresponding to 67 TFs, for which trained AGNet models were available in my dataset, for the GM12878 cell line. I also downloaded the peak sequences and normalized TF binding intensity used to train the DeFine models. I then selected 15% of the peak regions used to train the DeFine models and predicted their intensity values using both DeFine and AGNet models. I did the same thing with the peak regions used to train the AGNet models. I compared the prediction accuracy of the two types of models for each TF by calculating the PCC between the actual and the predicted intensities from the two different sets of peak regions.

4.4B4: Determining the accuracy of the AGNet models for predicting allele specific binding(ASB) TF binding events

In order to assess the accuracy of the AGNet models for classifying allele specific(ASB) TF binding events, I downloaded the ASB data aggregated by Wagih et al.[123] based on differential TF binding identified by ChiP-Seq experiments for 81 TFs. In this dataset, there were 32,252 ASB events ($P_{binomial} < 0.01$) and 79,827 non-ASB events($P_{binomial} > 0.5$). I compared the performance of AGNet models to 7 other methods: QBiC-Pred[42], DeepSEA[43], DeepBind[44], PWM(Meme)[39], PWM(Jaspar)[4], GERV[124] and deltaSVM[125]. I identified 1,915 ASB events(968 Gain-of-Binding and 947

Loss-of-Binding) corresponding to 613 single nucleotide polymorphisms(SNPs) variants and 10 TFs for which prediction models were present for all the algorithms. For these 10 TFs, I had 2,170 non-ASB events(1132 Gain-of-Binding and 1038 Loss-of-Binding) corresponding to 1,001 SNPs. I scored both the ASB and non-ASB SNPs using the AGNet models by first centering the variants within the TFBS identified for the 10 TFs in my original GM12878 based peak set. I then generate a pair of sequences for each variant containing an alternate allele(S_{ALT}) and a reference allele(S_{REF}) and scored both the sequences. The variant influence on the TFBS(S_v) was then derived from the difference in the two scores as shown in equation (4.1).

$$\mathbf{S_v = S_{ALT} - S_{REF}} \qquad (4.1)$$

I also scored the variants using the QBiC-Pred algorithm described in **2.2B9** and used the z-scores as the ASB and non-ASB variant influence on TF binding. For the remaining 6 methods, I simply used the scores compiled by Wagih et. al.[123] for each ASB and non-ASB variant for the 10 TFs. Furthermore, for each algorithm, I used thresholds of 25^{th} , 20^{th} and 15^{th} percentiles in order to classify the ASB events based on the scores such that events with scores in the top portion of that percentile were considered Gain-of-Binding ASB events, while the events in the bottom portion of the percentile were considered Loss-of-Binding ASB events. The remaining events were considered as non-ASB events. I then used AUROC to calculate the accuracy of each algorithm for correctly identifying an ASB event using the ground truths from the data compiled by Wagih et. al.

4.4B5: Genotype and expression datasets used for building the TFXcan framework

Apart from the DGN dataset described in **2.2B9**, I utilized two additional datasets in order to build and validate the TFXcan framework. I used normalized whole blood expression and whole genome sequencing data from GTEx[49] V.7 for the European descent individuals with sample size of 313. Additionally, I used HRC imputed Non-hispanic White(NHW) genotype and whole blood expression data for 241 ascertained at the University of Miami as part of the Alzheimer's Disease Genetics Consortium(ADGC). All the genotype and expression data were aligned to GRCh build 37 to be consistent with the AGNet models. Furthermore, I adjusted the GTEx expression data using the 3 pre-computed genetic PCs, sex, sequencing platform and 35 PEER factors. I quantile normalized NHW transcripts-per-million(TPM) expression values using the GTEx RNA-seq processing pipeline. Additionally, I used 6 genetic PCs and age of the individual when the sample was drawn as covariates to adjust the expression.

4.4B6: Scoring variants for the TFXcan framework utilizing AGNet models and TF effect estimates

In order to score genomic variants, based on their influence on TG expression, I used the aggregation schematic described in **Figure 4.1A**. Specifically, I first calculated the influence of each variant on each TFBS located within a TG's cis-regulatory region by using the corresponding AGNet model trained in **4.4B1** and equation 4.1 . I note here that

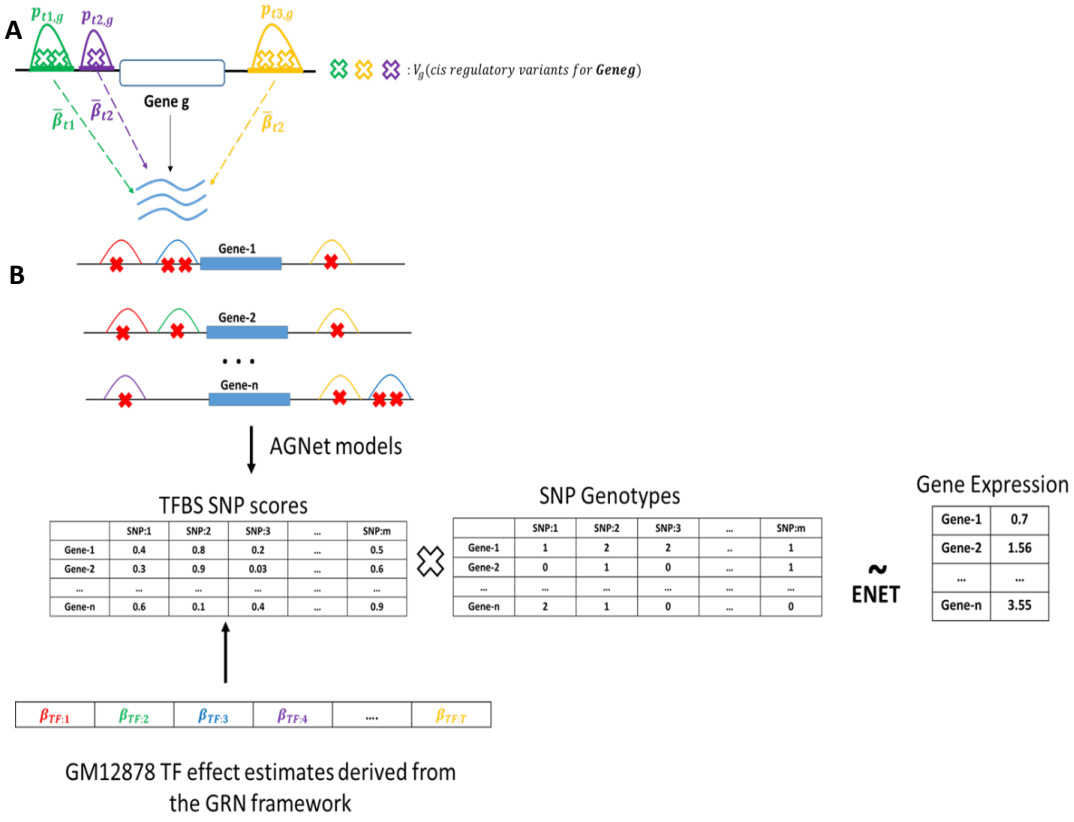

Figure 4.1: *Overview of the TFXcan framework. A) Schematic diagram of the scoring method used to calculate influence of cis-regulatory variants on TG expression by aggregating AGNet variant scores with the TF effect estimates. B) The overflow of the TFXcan framework where the aggregate scores derived from A are used in conjunction with the SNP genotype information in order to build TG expression prediction models using the ENET algorithm.*

I have used CTCF boundaries within a 50kb window around the TG body to define it's cis-regulatory region as described in **2.2B3**. The variant scores are then aggregated with the TF effect estimates calculated in **2.2B6**, using the following two strategies:

$$S_{v,g} = \frac{\bar{\beta}_t \times \sum_{t \in T_g} S_{v,p_{t,g}}}{|T_g|} \quad (4.2)$$

$$S_{v,g} = \bar{\beta}_t \times \sum_{t \in T_g} S'_{v,p_{t,g}} \quad (4.3)$$

Here, $S_{v,g}$ is the influence of cis-regulatory variant v on the expression of TG g. It is calculated in two different manners using equations (4.2) and (4.3). The "Diff-Mean" aggregated scores described in equation (4.2) is computed by adding the the TFBS score for the variant v for each peak region belonging to a TF t in TG g's cis-regulatory region, $p_{t,g}$ and then multiplying it with the average effect estimate for t $\bar{\beta}_t$ calculated in **2.2B6**. This product is then divided by the total number of TFs in g's cis-regulatory region. Furthermore, the variant-TFBS score for each peak region, $S_{v,p_{t,g}}$ using the difference in sequences containing the alternate and reference allele for v as described in equation (4.1). On the other hand, the "Log-Add" aggregation method uses the inverse transformed scores for the reference and alternate alleles scaled back to the original Log10 transformation described in **4.4B1** using a scaler specific to each TF to calculate the variant-TFBS score $S'_{v,p_{t,g}}$. The sum of all the variant-TFBS scores belonging to t is then multiplied with $\bar{\beta}_t$. The aggregated scores are then either cube root transformed or are scaled in the range (-1,1) to generate the final variant scores for each TG in the dataset.

4.4B7: Training and validating TG expression prediction models using the TFXcan framework

I used the genotype and expression datasets described in **4.4B5** and the framework described in **Figure 4.2** to train and validate TFXcan based TG expression prediction models. Specifically, I first extracted and scored all the common variants(MAF > 5%)

present within the TFBS in each TG's cis-regulatory region using scoring mechanism described in **4.4B6**. I then multiplied these scores with the variant genotypes obtained from the corresponding dataset to generate the weighted allele dosage table, which was directly used to build the train the ENET TG expression prediction models. I followed 10-fold cross-validation strategy used by PrediXcan[126] and EpiXcan[51] to train each TG model for each dataset. Specifically, I divided the samples into 10 roughly equal sized bins and used 9 bins to train the model and tune the parameters and validate it on the left out bin. For each TG prediction model, the predictions are made across all the folds and average CV R^2 is stored along with the variant weights. These weights are then used for predicting TG expression for samples in an independent dataset. I generated such weights using DGN and GTEx datasets and predicted TG expression for the NHW samples.

4.4B8: Building TG expression prediction models using the EpiXcan framework and comparing them to the TFXcan models

I utilized the EpiXcan framework described by Zhang et. al.[51] to build TG expression prediction models and then compared them to those obtained from the TFXcan framework. I first used the common cis-regulatory variants present within the TFBS identified for each TG in **4.4B6** to build linear regression models using the MatrixEqtl R package(*version 2.3*)[127] to compute their effects. I also downloaded GM12878 specific 25-epigentic state annotation bed file from the Roadmaps epigenome project website[128]. Using the annotation bed file and the eQTL summary statistics, I computed priors for each of the 22 states corresponding to active transcription regions, using qtlBHM[129]. These priors were then used as penalty factors for building the EpiXcan TG models after scaling them using the quadratic Bezier function as described by Zhang et. al.[51] The accuracy of the EpiXcan

models was compared to that obtained from the TFXcan models using prediction improvement ratio(PIR) for cross-validation correlation R^2_{cv} and for prediction correlation R^2_p for each TG's observed and predicted expression using the following equation:

$$\text{Prediction Improvement Ratio(PIR)} = \frac{R^2_{\text{TFXcan}} - R^2_{\text{EpiXcan}} > 0}{R^2_{\text{TFXcan}} - R^2_{\text{EpiXcan}} < 0} \qquad (4.4)$$

4.4C: Results

4.4C1: AGNet models were more accurate at predicting TF binding intensity compared to conventional deep learning models.

AGNet model architecture, proposed by Guo et. al., has been shown to perform better than conventional CNN-RNN model for predicting chromatin accessibility using DNA sequence information in form of k-mer embeddings[121]. It has several novel properties such as :1) Position embedding to capture the relative importance of each k-mer within a TFBS sequence 2) Dual attention layers capable of extracting important information from the sequence and feature based inputs 3) GCN and GRNs for deriving informative local

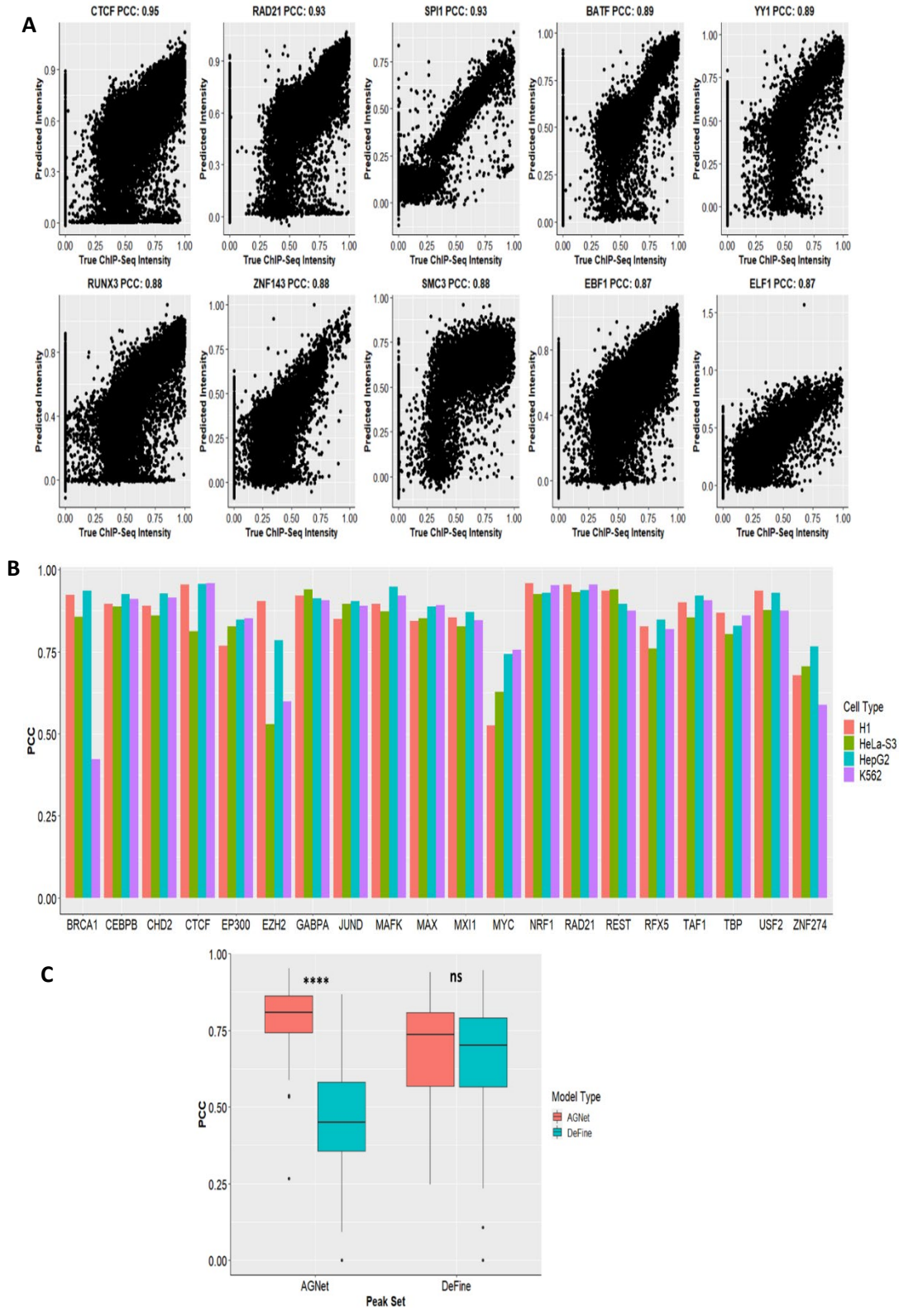

Figure 4.2: *The AGNet TF models were very accurate at predicting TF binding intensity. A) Scatterplots showing the prediction accuracy for the most accurate TF models B) PCC for the 20 TF models trained using the cross-cell type data. C) Boxplots showing the comparison between AGNet and DeFine models from predicting intensity values for AGNet and DeFine peak sets. (****-p-value < 10^{-5})*

features within and global relatedness among the input k-mers respectively. Here, I have

adapted the architecture in order to predict in vivo ChIP-Seq TF binding intensity values. Specifically, I trained AGNet models for each one of the 149 TFs using data from the GM12878 immortalized lymphoblastoid cell-line(see **4.4B1**). I tested the accuracy of each TF model by calculating PCC between observed and predicted scaled log normalized intensity values for a test set of peak regions. The median PCC for all the 149 TFs was 0.768, with just 5 models producing PCC less than 0.5 corresponding to TFs CBX3(0.395), CHD4(0.396), KDM1A(0.383), NFXL1(0.372) and SREBF2(0.267). These TFs were eliminated from all the subsequent analyses as a result of their poor performance. On the other hand, models corresponding to 28 TFs were highly accurate(PCC > 0.85), some of which have been shown in **Figure 4.2A**. I also trained models for 4 other cell-lines(K562, H1, HepG2 and HeLa-S3) for 20 TFs, that had data for available for all of these cell-lines in ENCODE using a cross-cell type training strategy described in **4.4B2**. As shown in **Figure 4.2B,** these cross-cell type models were largely accurate except for the one corresponding to BRCA1-K562 (PCC = 0.422).

I compared the performance of the AGNet models to those built using the DeFine architecture[120] consisting of a pair of identical conventional CNN layers reading the input DNA sequence for a TFBS in the forward and reverse directions to produce features passed on to a set of fully connected layers for predicting in vivo ChIP-Seq intensity. I hypothesized that due to the aforementioned properties of the AGNet models, they would perform better than the DeFine models. In order to test this hypothesis, I predicted intensity values using both models for a set of peak regions used to train DeFine models and a set of regions used to train the AGNet models for 67 GM12878 TFs. As shown in **Figure 4.2B**, both the set of models perform similarly on the DeFine peak regions(Median-PCC_{AGNet} =

0.736; Median-PCC$_{DeFine}$ = 0.701, Wilcoxon p-value = 0.07). Howevern, the AGNet models outperformed the DeFine models significantly(Median-PCC$_{AGNet}$ = 0.806; Median-PCC$_{DeFine}$ = 0.449, Wilcoxon p-value = 6.9e-12), when predictions were made on the peak regions used for training AGNet models.

Thus, the AGNet TF models were very accurate at predicting in vivo TF binding intensity values, while outperforming the conventional CNN bases DeFine models.

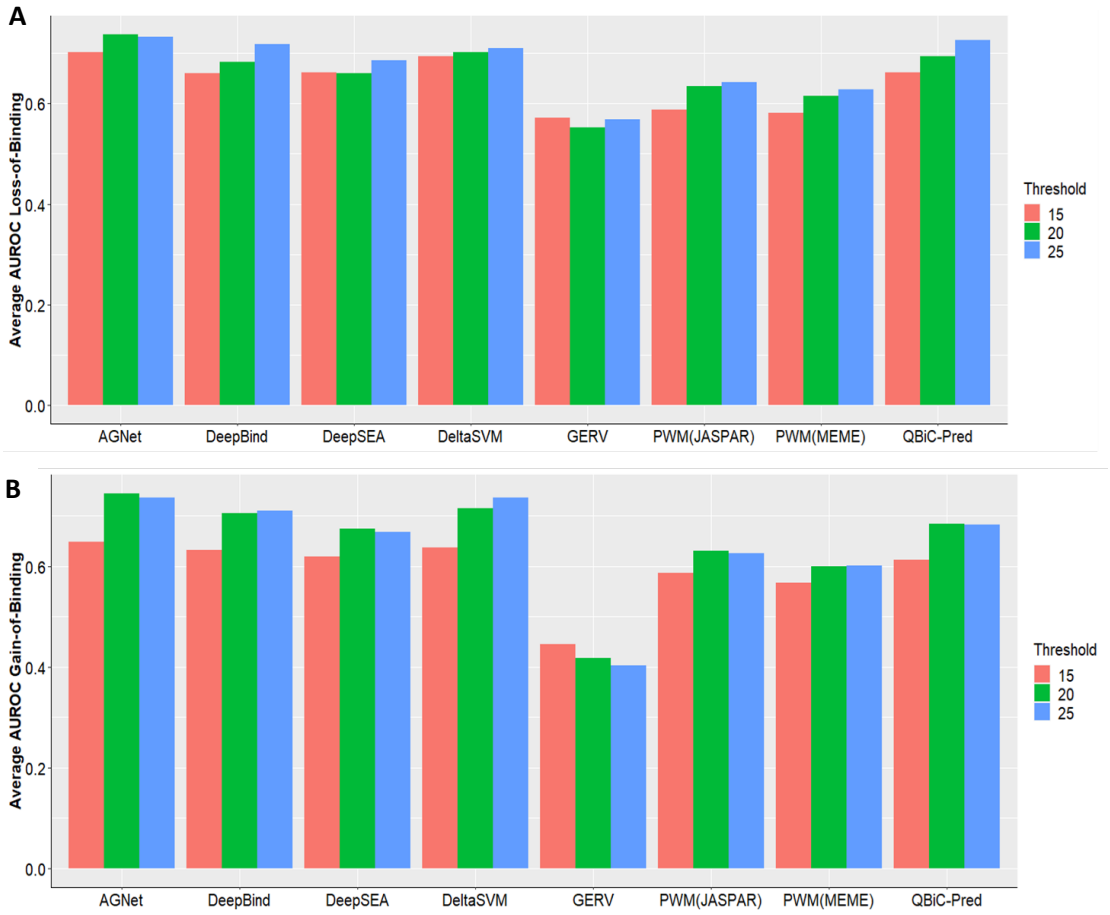

Figure 4.3: *AGNet models were more accurate than other methods for classifying ASB events. Barplots showing the accuracy obtained from classifying ASB events, by the means of average AUROC calculated over 10 different TF models, using different methods for A) 947 Loss-of-Binding events and B) 968 Gain-of-Binding events.*

4.4C2: Variants altering TF binding sites *in vivo* were more accurately classified by the AGNet models compared to other variant annotation algorithms

The AGNet models were more accurate in predicting in vivo TF binding intensity than the conventional CNN deep learning models. Next, I assessed the ability of these models to classify variants that can influence binding of TFs leading to an allele specific binding (ASB) events. Such ASB events can correspond to either increase (Gain-of-Binding variants) or a decrease(Loss-of-Binding variants) in the binding affinity. I used in vivo differential TF binding changes measured using significant changes between reads of alternate and reference alleles to classify ASB events[123]. Furthermore, I compared the performance of the AGNet models to 7 other TF binding variant annotation tools using the data compiled by Wagih et. al.[123]

I identified 10 TFs for which models were available for all the algorithms and used pre-compiled scores, except for the QBiC-Pred method, to classify Gain-of-Binding and Loss-of-Binding ASB events(see **4.4B4**). I utilized three different thresholds(25%, 20%, and 15%) on both ends of the distribution of these scores to call Gain-of-Binding and Loss-of-Binding events for each algorithm. As shown in **Figure 4.3**, the average AUROC for the AGNet models was the highest among all the 8 algorithms, for the three thresholds for both Gain-of-Binding($AUROC_{15}$ = 0.647, $AUROC_{20}$ = 0.745, $AUROC_{25}$ = 0.737) and Loss-of-Binding($AUROC_{15}$ = 0.702, $AUROC_{20}$ = 0.738, $AUROC_{25}$ = 0.732) ASB events. Additionally, methods such as DeepSEA[43],DeepBind[44], deltaSVM[125] and QBiC-Pred[42] performed comparably to the AGNet models, while JASPAR and MEME based PWM scores and GERV[124] produced poor classification results.

Thus, the AGNet models were superior at predicting ASB specific binding events brought about by the presence of variants within the TFBS in comparison to other popular non-coding variant annotation tools.

4.4C3: Utilizing TF based regulatory information in conjunction with AGNet derived variant influence over TFBS produced more accurate TG expression prediction models compared to using broadly defined epigenetic priors.

Table 4.1: *Table showing prediction performance of TFXcan in comparison with the EiXcan models. Table showing the PIRs for different datasets calculated using the prediction results obtained TFXcan models based on different aggregation and scaling methods and comparing them with those obtained from the EpiXcan models. The numbers in the parenthesis show the number of models used for each comparison. The cross-validation results show* PIR_{CV}*, while the prediction results show* $PIR_{Prediction}$

Cross-Validation Results			
Dataset	Aggregation Method	Cube root score(CRS)	Scaled score(SS)
DGN	Diff-Mean	1.393(9,936)	1.387(9,936)
	Log-Add	1.398(9,936)	1.400(9,936)
GTEx	Diff-Mean	1.304(6,714)	1.306(6,714)
	Log-Add	1.303(6,714)	1.298(6,714)
Prediction Results(Training Dataset-Prediction Dataset)			
DGN-GTEx	Diff-Mean	1.475(7,931)	1.494(7,931)
	Log-Add	1.481(7,931)	1.485(7,931)
DGN-NHW	Diff-Mean	1.374(8,450)	1.373(8,450)
	Log-Add	1.380(8,450)	1.378(8,450)
GTEx-DGN	Diff-Mean	1.481(4,920)	1.458(4,920)
	Log-Add	1.471(4,920)	1.461(4,920)
GTEx-NHW	Diff-Mean	1.406(5,051)	1.411(5,051)
	Log-Add	1.367(5,051)	1.399(5,051)

As described in previous sections, the AGNet TF models were highly accurate at predicting TF binding intensities and for classifying ASB events. Using these models, and

information derived from influence of multiple TF based regulatory mechanisms on TG expression I developed a novel TWAS approach called TFXcan. My approach models the

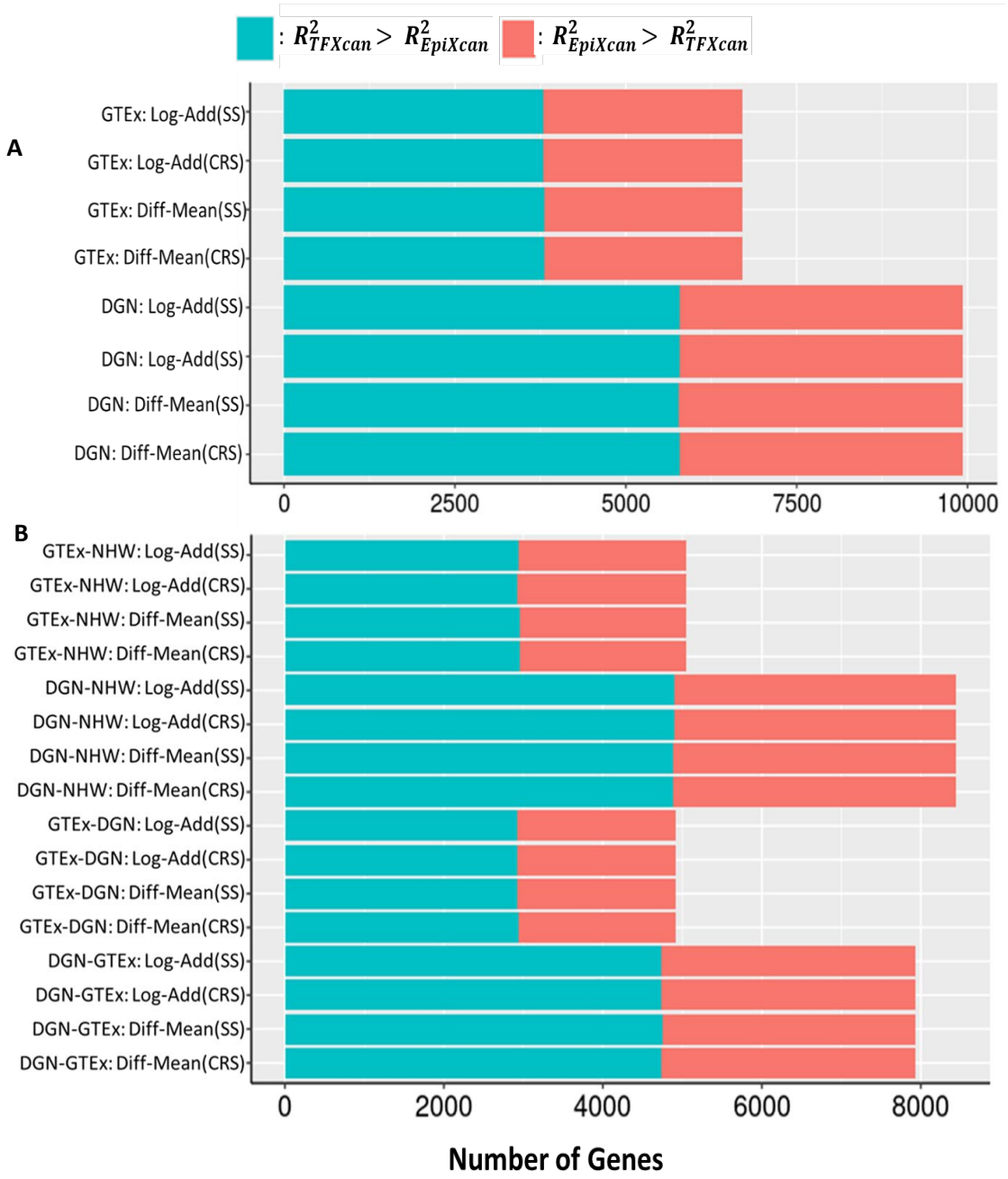

Figure 4.4: *TFXcan produced more accurate TG expression models compared to EpiXcan. A) The prediction performance of obtained fro m 10-fold cross validation within DGN and GTEx models using different aggregation and scaling methods. The color indicates the TGs for which TFXcan models performed better(in blue) or the ones for which the EpiXcan models better(in blue) B) Performance of the TFXcan and EpiXcan models from predicting TG expression in independent datasets.*

impact of common non-coding TFBS variants on TG expression regulation by using an aggregate score derived from AGNet based variant influence on TFBS and the average

impact of TFs on TG expression (see **4.4B6** and **Figure 4.1A**). The effect estimates derived from the ENET TG expression prediction models, trained using integrative PANDA GRN TF regulatory features in GM12878 in **2.2C2**, were used to capture the average influence of TFs on TG expression. I aggregated these AGNet-TF scores using two different methods: "Diff-Mean" and "Log-Add" and also subsequently transformed them into cube root scores and scaled scores in the range(-1,1). I used these scores, and the genotype and expression data obtained from DGN and GTEx whole blood datasets, to build TWAS style ENET based TG expression prediction models as described in **Figure 4.1B** and **4.4B7** . Furthermore, I predicted the TG expression for one of these datasets using variant weights derived from the training TFXcan models using the other, in addition to predicting expression for an independent NHW dataset.

I compared the results obtained from the TFXcan models to those obtained from the EpiXcan models[51](see **4.4B8**), which contain broad annotations, in the form of

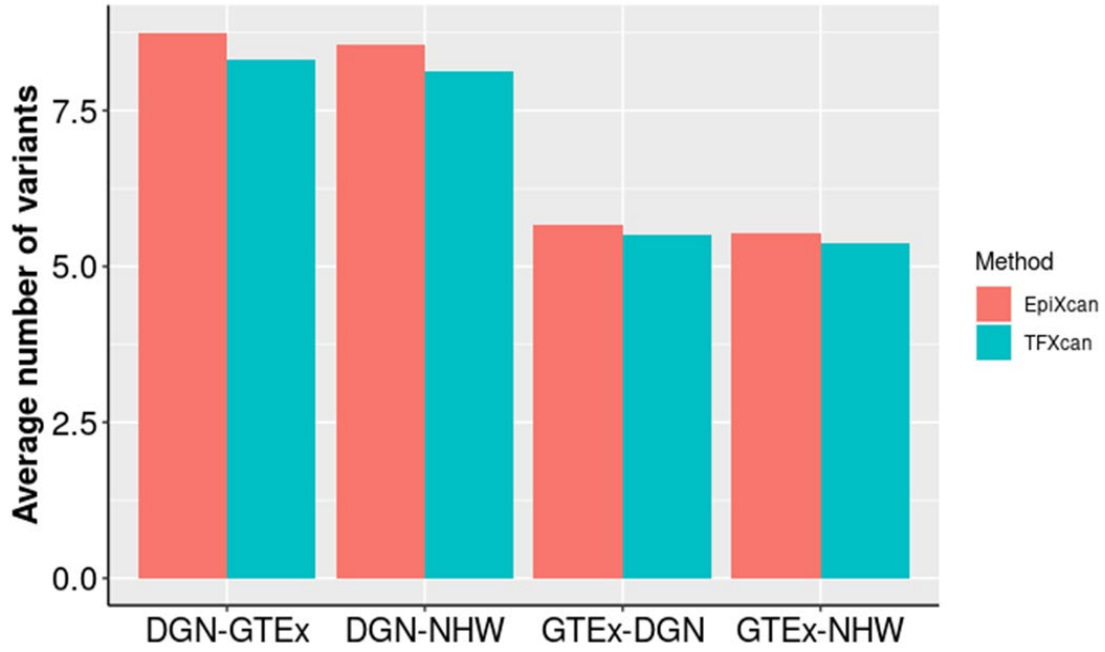

Figure 4.5: *Barplots showing the average number of variants used while predicting TG expression in independent datasets.The average number of variants across all the TGs used for TFXcan and EpiXcan prediction models were almost equal.*

Bayesian priors, for non-coding *cis*-regulatory variants derived using eQTL summary statistics and epigenetic annotations for different chromatin states. I used 143,204 variants and 180,857 variants to train the DGN and GTEx models respectively. I utilized PIRs calculated based on correlation R^2 from 10-fold inner cross-validation (PIR_{CV}) as well as from R^2 obtained from predicting TG expression for an independent dataset($PIR_{Prediction}$) to make the comparison between EpiXcan and different TFXcan models. These PIRs for each model for these datasets have been described in **Table 4.1**, along with the number of TGs used to make the comparisons. Furthermore, I have also shown the barplots in **Figure 4.4** to graphically represent the TGs for which TFXcan performed better ($R^2_{TFXcan} > R^2_{EpiXcan}$) and the ones showing the reverse results for CV as well as prediction models.

As shown in **Table 4.1**, the PIR_{cv} for both DGN and GTEx across different scaling and aggregation methods was greater than 1, implying a better performance of the TFXcan models compared to the EpiXcan models over all the TGs. Furthermore, the $PIR_{Prediction}$ produced from predicting TG expression using variant weights derived from DGN and GTEx TFXcan models was also greater than 1 for all the four scenarios ($DGN_{Training}$-$GTEx_{Prediction}$, $DGN_{Training}$-$NHW_{Prediction}$, $GTEx_{Training}$-$NHW_{Prediction}$ and $GTEx_{Training}$-$DGN_{Prediction}$). This prediction improvement in TFXcan models over EpiXcan models was observed despite the average number of variants, used for predicting TG expression in an independent set, being approximately equal for the two as shown in **Figure 4.5**. Thus, weighting variants using the TF based TG regulatory information produced more accurate TG expression prediction models compared to using broad epigenetic priors derived from eQTL summary statistics.

4.4D: Discussion

In this chapter, I described a novel TWAS approach built using TF based regulatory information in conjunction with variant influence over TF binding. I utilized neural network models trained using the novel AGNet architecture to quantify this influence. I showed that these models were more accurate at predicting TF binding intensity, compared to conventional CNN models and were also better at classifying variants causing significant changes in TF binding via ASB than many commonly used algorithms. This high accuracy and improved performance of the AGNet models over other approaches is mainly due to its capability to capture local and global features from the input k-mers using GCNs and GRNs. Furthermore, the positional information of each k-mer within a TFBS sequence is also captured in AGNet and is concatenated with the said features to obtain a better prediction of TF binding. Lastly, the dual attention layers in the AGNet models derive the importance scores for the GNN based features and the k-mer position embeddings to better inform the TF binding prediction module. I would like to note here that, the accuracy of the AGNet models comes at a cost of interpretability. In other words, due to the complicated architecture of these models, deriving motif information from them is extremely difficult, if not impossible. On the other hand, CNN models such as DeepSEA, DeepBind and DeFine are much more suited for motif discovery analysis due to their simpler architectures. However, I main area of focus for this project was to develop accurate TF-variant annotation models so I was not limited by the lack of interpretability of the AGNet models.

I built TWAS style TG expression prediction models using the AGNet based variant influence on TF binding and the average effect of TFs on TG expression. My

method, TFXcan, outperformed the current state-of-the-art approach EpiXcan which uses broad epigenetic prior based annotation to weight variants in the TWAS models. This improvement in performance was not drastic, as most of the PIR values were only slightly greater than 1. However, one must take into account that EpiXcan framework utilizes eqtl summary statistics beforehand to derive influence of different epigenetic states on TG expression. On the other hand, TFXcan is agnostic to any prior information corresponding to eqtls and only uses TF regulatory information based biological principles to derive variant weights.

Since I trained the AGNet and the GRN models using data from GM12878, the TFXcan is mostly suited for analysis of whole blood datasets. However, I did train the AGNet models using data from other cell-lines and models for other TFs can be conveniently trained using I code provided high quality ChIP-Seq data is available for them. Additionally, the GRN models can also be built using the framework described earlier in and the cell type specific multi-omics datasets. Thus, even though I have built and validated the TFXcan approach in whole blood, it can be easily generalized to other cell types and tissues.

The TFXcan approach takes advantage of the biologically relevant TF based regulatory information to derive genetic variant influence on TG expression. My models, unlike other TWAS approaches, don't depend upon availability of eQTL summary statistics and can also be used to analyze novel variants as I scoring algorithm was not trained on any reference panel of variants.

CHAPTER 5: TFKIN: A NOVEL KERNEL BASED APPROACH TO STUDY ASSOCIATION BETWEEN GENE EXPRESSION AND RARE VARIANTS ALTERING TRANSCRIPTION FACTOR BASED REGULATION OF GENE EXPRESSION

5.5A: Introduction

Rare variants(MAF < 5%) have been studied in context of many diseases and in most cases have helped elucidate the missing mechanistic link or the heritability for complex disease traits[130–134]. More and more whole genome and whole exome studies have been performed in the large decade enabling the analysis of the rare variants on a genome-wide level. The types of collapsing test generally used to study rare variants associations with complex phenotypes have already been described in **1C** . Briefly, collapsing tests use a set of rare variants within a pre-defined unit/region of the genome(e.g. gene) combined either in the form of a kernel/kinship matrix or as a single combined score aggregated from all the rare variants.[55] While these tests, assume that rare variants influence a given trait based on their MAF, there have been several other approaches that have combined functional annotations with the MAF to produce improved statistical power. Example of such methods include FunSPU[64], STAAR[61], SMART[63] and FST[62] However, there are certain disadvantages of these methods: 1) While the coding variants are appropriately weighted in these approaches, the non-coding variants weights are mostly derived from algorithms that are imprecise and don't capture their regulatory potential. For instance, all of these methods utilize annotation tools such as CADD[135] and FunSeq2[136], which have been shown to perform inferiorly to the modern deep learning non-coding variant annotation tools[43,120], to calculate weights for the non-coding variants. The STAAR method does use epigenetic and TF based functional annotations for these variants, but they need to be more fine-mapped to obtain a better functional relevance. 2) Due to the inclusion of multiple annotation scores, the

rare variants association tests become too complicated to resolve the mechanism behind the trait occurrence.

In order to overcome the aforementioned limitations of the annotation based rare variants association tests, I developed a novel approach called TFKin. My approach weights rare variants based on their influence on TF driven TG expression regulation and uses a variance component test to perform association between the kinship kernel generated based on these scores and the TG expression trait. The variant scores are derived by aggregating the AGNet model(described in **Chapter 4**) based variant influence on TF binding with effect estimates reflecting generic influence of different TFs on TG expression. My approach uses sophisticated deep learning models for deriving highly accurate rare variant annotation scores (see **4.4C2**) and merges them with TF effect estimates based on TG expression models trained using multiple regulatory mechanism information (see **2.2C2**). I performed extensive simulation analyses to calculate Type-I error and power of TFKin. Additionally, I used whole blood expression and genotype data from DGN[86] an GTEx[49] to build and validate the TFKin approach. My approach can be easily adapted and applied for studying regulatory mechanistic underpinnings of complex disease traits and for characterizing influences of novel variants on TG expression regulation.

5.5B: Methods and Materials

5.5B1: Weighting rare variants using TF regulatory information

I utilized the AGNet models for scoring *cis*-regulatory rare variants based on their influence on TF binding and then aggregated them with average effects of each TF on TG expression as described in **4.4B6**. I used the same aggregation schemes of "Diff-Mean" and

"Log-Add" as before, and then transformed the aggregated variant scores using the cube root scaling based on *equation (5.1)*.

$$S_{v,g} = \text{Scale}(\sqrt[3]{S_{v,g}}, 1,10) \tag{5.1}$$

The cube root scaled TFAGNet scores were then used for downstream analysis in the TFKin framework.

5.5B2: TFKin variance components test

I briefly review LMM used in association testing and then introduce the construction of a weighted linear kernel for TFKin. Assume I have n individuals for whom I have p non-genetic covariates, genotypes for m SNPs, and phenotype information. I refer to the phenotype, y, an $n \times 1$ vector, genotype, **G**, an $n \times m$ matrix, and covariates **X**, an $n \times p$ matrix.

An LMM includes a fixed effect from covariates, annotated by $\mathbf{X}\beta$, and a random effect annotated by $\mathbf{Z}u$ and an error term ϵ. The response y is fit with a high-dimension normal distribution (Eq. 2). The random effect can be further divided into two parts: an environmental and a genetic effect denoted as $\sigma_e^2 I$ and $\sigma_1^2 K_g$ respectively. $\mathbf{K_g}$ is the kernel reflecting the genetic similarity between individuals. σ_1^2 is the amount of variance of y explained by $\mathbf{K_g}$.

$$y = \mathbf{X}\beta + \mathbf{Z}u + \epsilon \tag{5.2}$$

$$y \sim N\left(\mathbf{X}\beta, \sigma_1^2\mathbf{K_g} + \sigma_e^2\mathbf{I}\right) \quad (5.3)$$

The null hypothesis $\sigma_1 = 0$ indicates that the $\mathbf{K_g}$ does not explain any variance of y. The score statistic Q is defined as the partial differential for log-likelihood on σ_1^2. Under the null hypothesis, the Q follows a mixed chi-squared distribution:

$$\frac{Q}{\sigma_e^2} = y^T\mathbf{S}\mathbf{K_g}\mathbf{S}y \sim \sum_{i=1}^{n}\lambda_i\chi_1^2 \quad (5.4)$$

where **S** projects y into a space orthogonal to covariates and λ_i's are the eigenvalues of $\mathbf{S}\mathbf{K_g}\mathbf{S}$. Similar to the SKAT approach[56], I utilized a weighted linear kernel function to compute the relatedness among individuals in a given dataset w.r.t. their rare variant-TF scores:

$$\mathbf{K} = \mathbf{G}\mathbf{W}\mathbf{G}' \quad (5.5)$$

$$\mathbf{K}(\mathbf{G_i}\mathbf{G_{i'}}) = \sum_{j=1}^{p} w_j\, G_{ij}G_{i'j} \quad (5.6)$$

While SKAT uses MAF to weight rare variants in the similarity kernel: $\sqrt{w_j} = \text{Beta}(MAF_j, 1,25)$, I utilized the aggregate TFAGNet scores computed in **5.5B1** instead.

5.5B3: Utilizing the TFKin framework for TG expression-rare variants association

I used the reference datasets, described in **4.4B5** , for building TFKin models to find association between rare non-coding variants and TG expression based on the framework shown in **Figure 5.1**. All the rare variants (MAF < 5%) present within the TFBS in *cis*-regulatory regions of the TGs were scored for their influence on TF binding affinity using the AGNet models(see **4.4B4**). These scores were then aggregated with the TF effect estimates followed by different scaling and transformation schemes as described in **5.5B1**. I used *equation (5.6)* to build kinship matrix based on a linear kernel function containing the rare variant genotypes and their aggregate scores. I found significant associations between TG expression and this matrix using the variance components test described in

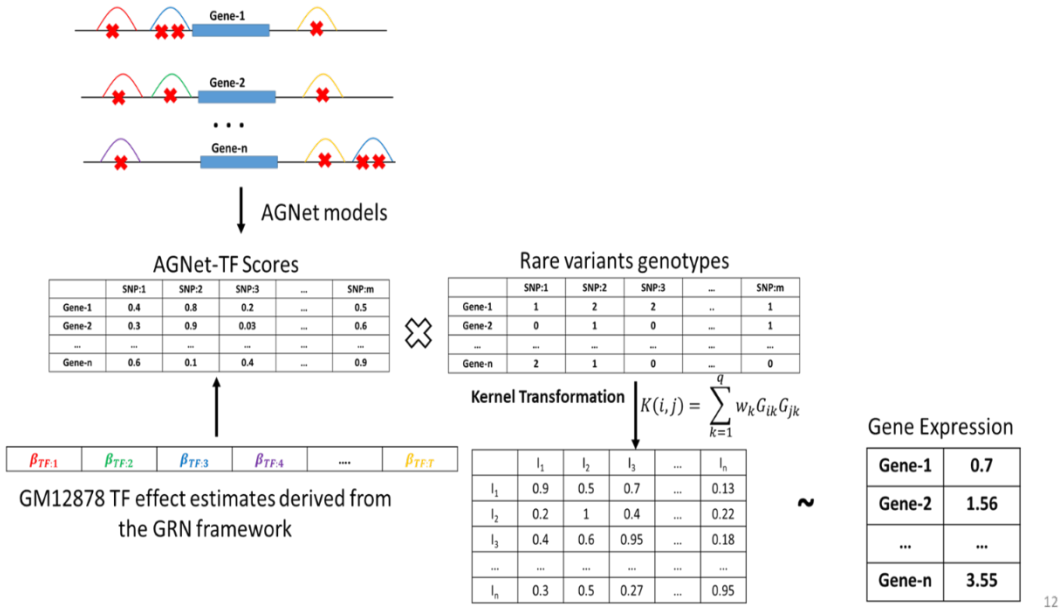

Figure 5.1: *Overview of the TFKin framework. Iutilized AGNet models and TF effect estimates to score rare variants present in the cis-regulatory region of each TG. These scores, along with the variant genotypes in a reference dataset(DGN, GTEx) to build a linearly weighted kinship kernel matrix. Ithen use variance components test to find association between this matrix and TG expression.*

5.5B2. I performed discovery analysis using DGN and GTEx datasets and replicated the results for each one of them in the other. In addition, I used the NHW dataset for second replication. The expression data was each of them was pre-adjusted for all the covariates, as described in **4.4B5** before fitting the TFKin models.

5.5B4: Simulation analysis

I used the DGN dataset to extract genotype information corresponding to *cis*-regulatory rare variants(N = 428,782) in order to perform simulation analysis to estimate power and type-I error. The workflow for this simulation analysis is presented in **Figure 5.2** encompassing the following steps:

1) I selected a random set of 50 rare variants from a random 1Mb region of one of the autosomes (Chromosome 1-22). The number of variants to be used in each simulation was based upon the median number of *cis*-regulatory rare variants used for building the TFKin for all the TGs, which was 45.

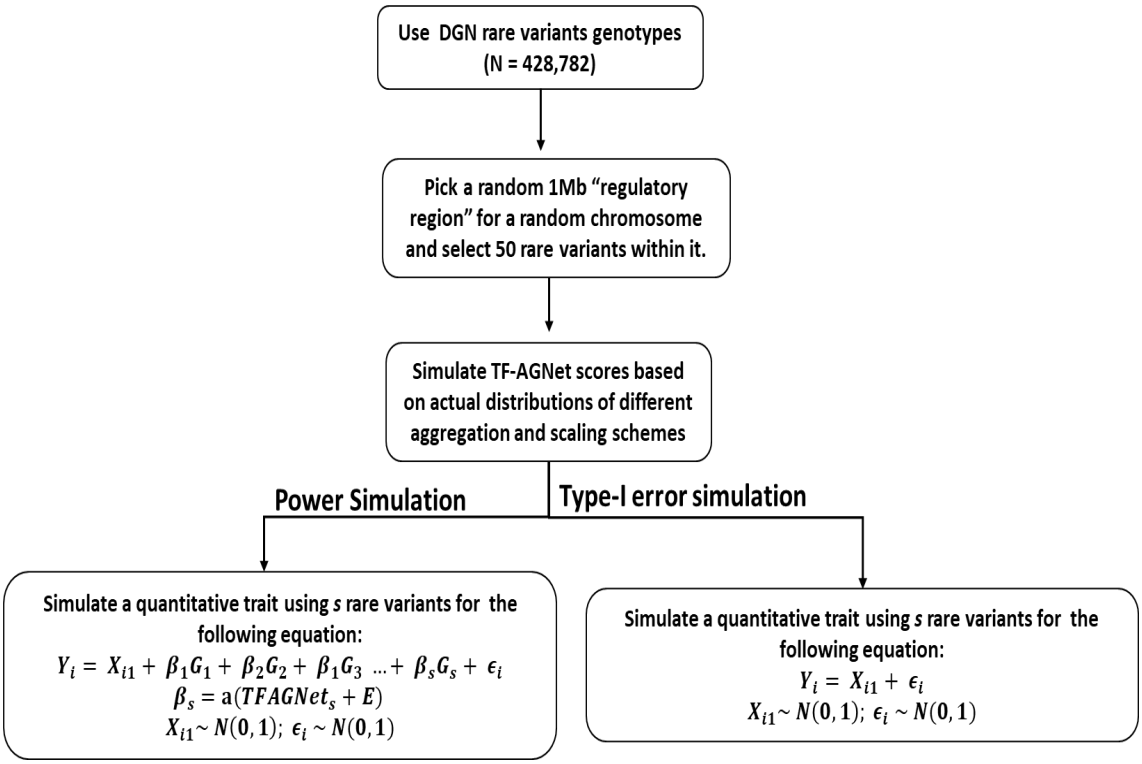

Figure 5.2: *Flowchart describing the steps used in simulation analysis for estimating empirical power and Type-I error. I used HRC imputed DGN genotype data corresponding to 428,782 cis-regulatory rare variants in order to perform the simulation. The equations in the last two box plots were used to estimate the phenotype for calculating power and Type-I error.*

2) Next, I simulated TFAGNet scores for these variants by selecting them from the distribution of the real aggregate scores utilized while fitting the TFKin models in **5.5B3**.
3) Using the genotypes for the 50 rare variants and the simulated scores, I then used the following equation to simulate the phenotype:

$$\mathrm{Y_i} = \mathrm{X_{i1}} + \beta_1 \mathrm{G_1} + \beta_2 \mathrm{G_2} + \beta_1 \mathrm{G_3} \ldots + \beta_s \mathrm{G_s} + \epsilon_i \qquad (5.7)$$

Here, Y_i is the simulated phenotype for iteration *i,* X_{i1} and ϵ_i are random values chosen from a standard normal distribution ($X_{i1} \sim N(0,1)$; $\epsilon_i \sim N(0,1)$). G_s corresponds to the number of minor alleles for rare variant *s* in the DGN dataset, while β_s is its effect estimate simulated using the following equation:

$$\beta_s = \mathrm{a}(\mathrm{TFAGNet}_s + \mathrm{E_s}) \qquad (5.8)$$

Here, *a* is a scaling factor, which was set at 0.15 and $TFAGNet_s$ is the random simulated score for the variant. E_s is an error term which was scaled and aggregated with the $TFAGNet_s$ score to reflect the error in quantifying the influence of variants on TF binding affinity using AGNet models. This error term was derived using the following equation:

$$\mathrm{E_s} = \mathrm{Scale}(\mathrm{Aggregate}(\mathrm{AGNet_T}(|\mathrm{y} - \hat{\mathrm{y}}|))) \qquad (5.9)$$

For each variant score, the error term was calculated using the AGNet model corresponding to the TF *T*, whose TFBS contains the variant. The absolute error values calculated between observed(y) and predicted($\hat{y}$) intensity for the

test set, while training the models (see **4.4B1**) were used to generate a distribution of errors. From this distribution, a random error value was selected for each variant based on the magnitude of error used in the simulation, such that an error of "high" magnitude represented value chosen from the higher end of the distribution, while "moderate" error was selected from the middle and "low" error was picked from the lower end of the distribution based on the following sets of equations:

$$E_{s-high} \geq 75th(\mu_{E_s}, \sigma_{E_s}) \tag{5.10}$$

$$75th(\mu_{E_s}, \sigma_{E_s}) < E_{s-moderate} \geq 50th(\mu_{E_s}, \sigma_{E_s}) \tag{5.11}$$

$$50th(\mu_{E_s}, \sigma_{E_s}) < E_{s-low} \geq 25th(\mu_{E_s}, \sigma_{E_s}) \tag{5.12}$$

A set of error values for each variant, based on the number of different TFBS, were then aggregated and scaled into one value which was added to its TFAGNet score in order to compute the effect estimates.

4) The phenotype for a set of 50 variants was then computed using equation (5.8) based on these error containing TFAGNet scores. TFKin models were fit using the simulated phenotype, where the kinship matrix was estimated with just the TFAGNet scores for the variants without the error term.
5) For type-I error, the phenotype was simulated using two random variables picked from a standard normal distribution based on the following equation:

$$Y_i = X_{i1} + \epsilon_i \quad (5.13)$$

The above steps were repeated for 10,000 iterations for both power and type-I error rate calculation. Additionally, for power calculation, four different scenarios were simulated, where the error term was kept high, moderate and low along with a scenario where no error term was used to calculate the effect estimates. The empirical power was calculated as the proportion of significant associations found at the p-value threshold of 5e-06. The type-I error rate was calculated similarly, at different values of alpha(0.05, 0.01, 0.001, 0.0001, 1E-05, 1E-06).

5.5C: Results

5.5C1: Weighting rare variants using TF based regulatory information improves their association with TG expression compared to traditional MAF based weighting method

As described in **Figure 5.1**, I used TF regulatory information and AGNet models to quantify the influence of rare non-coding variants on TG expression regulation using different aggregation methods. These scores, along with variant genotypes were used to compute a *cis*-regulatory kinship matrix which was then associated with TG expression using the variance components test described in **5.5B2**. I compared the results of these models to those containing kinship matrices built using MAF beta weights, similar to the SKAT approach. I used the DGN ($N_{variants} = 428{,}782;\ N_{TGs} = 12{,}027$) and GTEx ($N_{variants} = 614{,}693, N_{TGs} = 13{,}951$) datasets for discovery analysis and replicated the results for each one of them in the other, while I used NHW($N_{variants} = 442{,}262, N_{TGs} = 13{,}433$) for secondary replication. I note here that a single variant can have different aggregate scores for different TGs. The results from fitting the TFKin models are shown in **Table 5.1**, in form of significant TGs found in discovery analyses, and the two follow up replication analyses.

Table 5.1: *Table containing the results from the TFKin analysis of discovery, 1st and 2nd replication datasets. DGN and GTEx were used for discovery analyses, while the NHW dataset was purely used for 2nd replication. The number of significant TGs in the discovery analysis were calculated using the genome-wide significance threshold of 0.05/total number of TGs, while the ones in subsequent replications were calculated using the nominal p-value threshold of 0.05.*

Discovery Dataset(Replication-1, Replication-2)	Aggregation Method	Number of Significant Genes – TFAGNet scores	No. of Significant Genes – Beta MAF scores
DGN(GTEx, NHW)	Diff-Mean	2363(1120, 165)	2648(1093,148)
	Log-Add	2440(1123,178)	
GTEx(DGN, NHW)	Diff-Mean	169(140, 48)	146(117, 37)
	Log-Add	194(160, 51)	

As shown in **Table 5.1**, although the enrichment of significant TGs(p-value < 4.16e-06) was higher for kernels containing MAF based beta scores, the number of these

TGs that were found to be significant in 1st and 2nd replication analyses(p-value < 0.05) were higher for the TFAGNet score based kernel. For instance, 2440 TGs were found to be significant in the DGN discovery analysis based on kernels constructed using the *cis*-regulatory rare variants weighted by the "Log-Add" aggregation scheme. Out of these, 1123 replicated in the GTEx dataset, out of which 178 replicated in the NHW dataset. On the other hand, although MAF beta score based kernels produced 2648 significant DGN TGs, 1093 of them replicated in GTEx and of them only 148 replicated in the NHW dataset. On the other hand, when the GTEx dataset was used for discovery analysis, the number of TGs found to be significant(p-value < 3.58e-06) were higher for the TFAGNet weighted kernels compared to the ones built using the MAF beta score. Furthermore, the 1st and 2nd replication analyses for these TGs also yielded higher enrichment of significant TGs for the TFAGNet kernel based models than the ones based upon the MAF beta scores. Overall, the "Log-Add" aggregation scheme based models resulted in more significant TGs compared to the ones built using scores derived from the "Diff-Mean" aggregation scheme.

Thus, the *cis*-regulatory kinship matrices generated using rare variant weights derived from TF based regulatory information resulted in higher number of significant associations compared to conventional weighting scheme involving MAF.

5.5C2: The TFKin approach produced reasonable power and well controlled type-I error rate

I performed simulation analyses, described in **5.5B4**, to estimate empirical power and type-I error rate for the TFKin approach. Power was estimated using simulated phenotype containing effect estimates based on different magnitudes of error in the TFAGNet scores used to generate the kinship matrices. Moreover, I also calculated power for the scenario where the phenotype was simulated with the effect estimates directly computed from the TFAGNet scores without any error term. The results from this power analysis are shown in **Figure 5.3**. As expected, power was the highest for

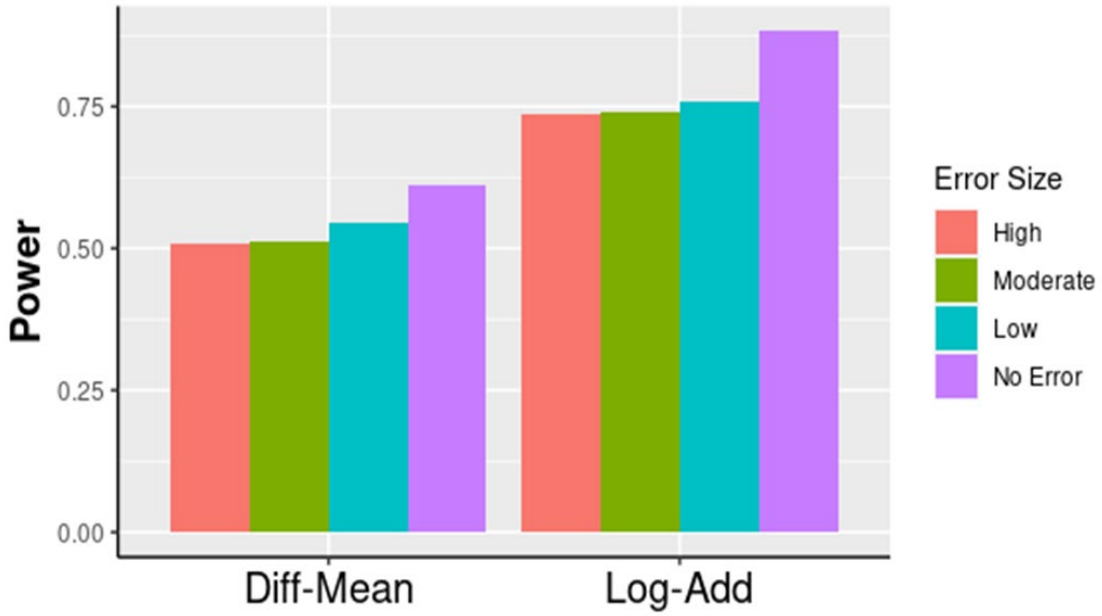

Figure 5.3: *Barplots showing the power obtained from simulation analysis. The power for each scenario was calculated as the proportion of significant associations found at a p-value threshold of 5e-06 based on 10,000 iterations.*

TFKin models when the phenotype was simulated without any error term using scores derived from the two aggregation schemes($Power_{Log\text{-}Add:No\ Error}$ = 88%; $Power_{Diff\text{-}Mean:No\ Error}$ = 61%). I observed sufficient power even when using the highest magnitude of the error term for the "Log-Add" aggregation scheme($Power_{Log\text{-}Add:High}$ = 73%), but not for

the simulation models based on "Diff-Mean"($Power_{Log\text{-}Add:No\ Error}$ = 51%). The same trend of obtaining higher power from the "Log-Add" models compared to the "Diff-Mean" models was seen for moderate($Power_{Log\text{-}Add:Moderate}$ = 74%; $Power_{Diff\text{-}Mean:Moderate}$ = 51%) and for low("$Power_{Log\text{-}Add:Low}$ = 76%; $Power_{Diff\text{-}Mean:Low}$ = 54%) error, with the expected pattern of power increasing with decreasing the magnitude of the error term.

Table 5.2: *Table showing the type-I error rate estimated at different values of alpha using the TFKin models. Type-I error rate was calculated based on 10,000 iterations using the TFKin models containing scores aggregated using the "Log-Add" and "Diff-Mean" schemes.*

Aggregation Method	Alpha	Type-I Error Rate
Log-Add	0.05	0.0472
	0.01	0.0088
	0.001	0.0014
	0.0001	0
	1.00E-05	0
	1.00E-06	0
Diff-Mean	0.05	0.0476
	0.01	0.0098
	0.001	0.001
	0.0001	0
	1.00E-05	0
	1.00E-06	0

As shown in **Table 5.2**, the type-I error rate was well controlled across different values of alpha for the simulation models corresponding to the two aggregation schemes. I did not see any inflation in the type-I error rate, in that it was always below the alpha threshold for different values of alpha.

5.5D: Discussion

In this chapter, I describe a novel weighted kernel association test TFKin for TG expression, where *cis*-regulatory variants are scored based on their influence on TF based TG expression regulatory mechanisms. Since I scoring method is based on complex neural network models, novel rare variants can be annotated and analyzed using my approach.

Furthermore, I have shown in my results that utilizing such regulatory scores in the kernel association tests leads to a better enrichment of significant associations in discovery analyses, and to a higher replication of these significant TGs, in blood based datasets, compared to using conventional weighting strategy based on MAF of rare variants. This was not surprising as previous studies[58,61] have shown that including functional annotation scores in tests for rare variants association lead to an increase in power. However, these studies use broadly defined functional scores for *cis*-regulatory non-coding variants, while my approach is based on a more fine-mapped scoring scheme for these variants. In addition, TFKin is completely independent of MAF scores, unlike other rare variants association approaches[60–62,64]. Currently, TFKin doesn't have the capability of merging the regulatory scores with the MAF beta scores. However, one can use the extended variance components tests developed by He *et. al.*[62] for inclusion of multiple functional annotations along with MAF scores in order to accomplish this goal.

TFKin models were also able to preserve sufficient power, at least for the "Log-Add" aggregation scheme based kernels, when the amount of error produced while scoring the variants was assumed to be high. The "Log-Add" aggregation strategy, as opposed to "Diff-Mean", inverse transforms the AGNet based scores to their original log scales before merging them with the TF effect estimates as described in **4.4B6**. By doing so, the scaling for the AGNet scores for different TF models becomes more uniform compared to the simpler approach taken in the "Diff-Mean" aggregation scheme. Thus, "Log-Add" based TFKin models produced higher power than the ones based on "Diff-Mean". Additionally, this pattern was also observed for the enrichment and replication of significant TGs in the whole blood based discovery and replication analyses.

In this chapter, I have built rare variants association tests for continuous TG expression trait in whole blood. Extension of my method would include modelling the influence of *cis*-regulatory variants on a dichotomous trait outcome. By training AGNet models using ENCODE data corresponding to different cell lines, one can extend the TFKin approach to other tissues as well. In addition, TFKin can also be used for TG based association testing with the phenotypic trait. One can impute TG expression using kinship matrices computed based on *cis*-regulatory rare variants scores and can associate the imputed expression with the trait outcome to identify risk TGs associated with a given complex disease based on the regulatory mechanisms perturbed by rare variants.

CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS

6.6A: Conclusions

In this dissertation, I have developed different methodologies that take advantage of multiple big "omics" data sources to derive influence of different TF based regulatory mechanisms on TG expression. I use different tools and analytical approaches, such as gene regulatory networks, machine learning, deep learning, PWM analyses along the way to accomplish this goal. My approaches also make use of variety of different data sources, such as ChIP-Seq, Hi-C, RNA-Seq, genotype data, whole genome sequencing, DNAase-Seq and PPI to construct a comprehensive picture of gene expression regulation.

In **Chapter 2**, I described a GRN based machine learning approach to detect global average influence of different TFs on TG expression in GM12878, K562 and HepG2 cell-lines. My integrative approach took advantage of several big "omics" data sources to derive influence of different TF based regulatory mechanisms on TG expression. I observed that utilizing TF based features from a GRN generated based on multiple data-sources produces more accurate TG expression prediction models, compared to utilizing single-source mechanism information such as TF binding. Furthermore, using I learned expression prediction models, I was able to derive average influence of different TFs on TG expression, which correlated very well with their biological functions. I was also able to compute effects of different regulatory elements (promoters, enhancers, introns etc.) on TG expression and found that promoters are extremely predictive of and essential for TG expression among all of them. Additionally, I also discovered novel regulatory role for intronic TFBS, previously seen only in lower eukaryotes. Lastly, I validated I integrative predictive models, by weighting rare variants using the average influence of TFs on TG

expression and building SKAT like collapsing tests for association with expression in whole blood datasets such as DGN and GTEx. I discovered that utilizing such a weighting scheme produces better enrichment and replication of significant TGs. Thus, the modelling approach, I have presented in this chapter, has multiple applications for studying general factors influencing gene expression. My 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. My framework also allows for the inclusion of additional functional genomics information, such as higher resolution chromatin interaction data, to evaluate their effect on gene expression. As I 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, my approach provides a roadmap for integrating multiple "omics" data sources and assembling fundamental aspects of transcriptional regulation into a coherent portrait of gene expression, which could ultimately help in elucidating mechanisms causing several diseases.

In **Chapter 3**, I extended I GRN based framework by computing influence of different combinations of TFs, forming TF regulatory modules(TRMs), on TG expression. I utilized trained multi-layer perceptrons(MLPs), built using GRN based TF features and TG expression, to calculate the non-linear effects of different TRMs. My discovered set of TRMs contained many novel, as well as some well-known TF interactions with varying

effects on TG expression. I further characterized the nature of these interactions, and discovered that most, if not all, of them occur via long distance chromatin looping. On the other hand, I found little evidence of co-binding among my discovered set of TF interactions. Lastly, I defined different architectures of TG regulation for my TRMs, which mainly consisted of TRMs interacting with the TG promoters via chromatin looping. In conclusion, I detected TRMs significantly impacting TG expression using neural network based prediction models containing multi-omics GRN derived TF regulatory features. I demonstrated multiple ways in which long distance chromatin looping plays a role in TRM based TG regulation. My approach for detection, characterization and validation of TRMs provides a roadmap for a multi-omics analysis to study the complex phenomenon of transcription regulation genome-wide, and may provide insights into the impact of transcriptional dysregulation in the genetic basis of human phenotypes.

In **Chapter 4**, I integrated information derived from impact of common variants on TF based TG expression regulation in my framework. Specifically, I developed a set of complex neural network models to predict the impact of common *cis*-regulatory variants on TF binding affinity. These models were more accurate than commonly used non-coding variant annotation tools at classifying variants significantly affecting TF binding. I integrated average influence of TFs on TG expression with the variant influence on TF binding to generate aggregate scores quantifying the regulatory potential of non-coding variants. Subsequently, I used these aggregate scores to build TG expression prediction models utilizing the novel TWAS framework TFXcan and individual level common variants genotype data in whole blood datasets of DGN and GTEx. In comparison to the state-of-the-art TWAS method EpiXcan, TFXcan models were more accurate at predicting

TG expression within and across datasets. This was especially surprising considering the fact that EpiXcan explicitly uses eQTL summary statistics to derive epigenetic priors in order to weight variants in the prediction models. Thus, utilizing biologically relevant TF based regulatory information to weight variants results in accurate TG expression models. My approach can be easily applied to complex traits with a significant genetic influence in order to identify their mechanisms. Furthermore, since my scoring algorithm doesn't depend upon any reference panels, one can use them to annotate novel variants without any prior eQTL information.

In **Chapter 5** , I developed a weighted rare variants kernel association test based on kinship matrices computed using the *cis*-regulatory rare variants scores which reflect their influence on TF based TG expression regulation. I showed that these matrices are better able to capture associations of rare variants with TG expression compared to those containing conventional rare variants MAF based weights. My discovery and replication analyses were based on whole blood expression and genotype datasets described in The TFKin approach also produced decent power, while preserving sufficient type-I error rate for different values of alpha. Furthermore, I hypothesize that merging the two types of scores based on the unified variance components tests developed for inclusion of multiple different functional annotation scores, would lead to a further enhancement of the significant associations. The TFKin doesn't make any assumptions regarding the importance of MAF of the rare variants in their association tests and instead uses biologically relevant information to derive their regulatory potential. Such an approach could be used to study regulatory mechanisms perturbed by rare variants in the context of complex disease traits.

The methodologies described in this dissertation have the capability to bridge the gap between the presence of genetic variants and transcriptional dysregulation by leveraging information from genomics, transcriptomics, epgenomics and proteomics datasets. Such methodologies will hopefully provide blueprint for researchers to design integrative approaches capable of further enhancing the field of multi-omics data integration to better understand several disease mechanisms.

6.6B: Future directions

6.6B1: Including regulators, other than TFs, in the GRN framework

In this dissertation, I have presented different methodologies that make use of multiple big-omics data sources, modelling and analytical approaches to determine global regulatory patterns of TFs and their influence on TG expression. Although my methodologies and results filled an important gap in the existing literature regarding TG expression regulation, one still needs to consider regulatory other mechanisms. For instance, different types of histone modifications mentioned in **1A** also play an important role in regulating TG expression. In addition, small non-coding microRNA(miRNA) bind to 3' UTR of mRNA transcripts to significantly alter their expression. Both of these regulators, besides TFs, also have a significant influence over TG expression regulation. One extension of my integrative approach will be to incorporate histone modifications and miRNA as regulators in the PANDA GRN while modelling TG expression thereby improving the overall prediction performance.

6.6B2: Identifying TRMs in cell lines other than GM12878

Using TF features and neural network based multi-layer perceptrons I computed the effects of different TRMs on TG expression. My preliminary analysis was only focused

on blood based GM12878 cell line. However, one can easily apply the TRM detection algorithm to other cell lines, provided that all the requisite data types are available for them. Data sources like ENCODE and GEO are extremely rich with ChIP-seq, RNA-Seq and Hi-C datasets for multitude of cell lines. Using these datasets, and I code, one can detect TRMs in cell lines other than GM12878. In addition, the modular regulation of TFs differs from one cell-line to other. This difference can be further analyzed by first detecting TRMs in different cell-lines and comparing them. Such comparison may shed light on cell-type specific TRM based TG regulation. Lastly, due to the non-linear nature of the MLPs, I was only able to detect the non-additive TF interactions for the TRMs as evidenced by an extremely small number of TF interactions interacting via additive co-binding being present in my detected set of TRMs. A more complete TRM detection should include both distally interacting and proximally co-binding sets of TFs. Thus, one can run several TF co-binding detection algorithms described in **1B** in parallel with my MLP based TRM detection approach to identify and characterize both types of TF interactions.

6.6B3: Training AGNet models for different cell lines and extending the TFXcan framework for other tissue types

The AGNet models described in **4.4B1** were trained primarily in the GM12878 cell line, with additional training done for 20 TFs using cross-cell type training strategy. Since GM12878 is a blood derived cell lines, currently the TFXcan approach is limited to only blood based datasets. However, in order to be more generalizable, one can train the AGNet models using TF ChIP-Seq data for other cell types following I training method. By doing so, one can easily adapt the TFXcan approach for application in other cell and tissue types by leveraging multi-tissue expression and genotype data sources such as GTEx. ENCODE

currently contains ChIP-Seq data for hundreds of TFs for commonly used cell lines such as K562, HepG2, H1-ESC, Hela-S3. However, these cell lines still don't represent complex tissues such as brain for which the epigenetic data is extremely limited. However, brain open chromatin atlas(BOCA) contains chromatin accessibility data for different brain cells derived from ATAC-Seq experiments. One can use this data to map out the open chromatin regions in these cells, and can use one of several TFBS prediction tools described in **1C** to impute TF binding data. Once that's done, building TFXcan models is straightforward. Lastly, the standalone AGNet models can be used for functional fine mapping of several eQTLs by leveraging data from large databases such as the GTEx portal and the eQTL catalogue. Thus, one can determine if any of the previously unannotated eQTLs, significantly associated with TG expression, present in these large datasets are present within and significantly impact TF binding.

6.6B4: Applying the methodologies described in the dissertation to study complex disease mechanisms.

In the era of big omics, there has been a push in the biomedical community to develop and utilize integrative multi-omics approaches to study a disease in order to gain better understanding of the flow of its occurrence, from its cause(genetic, environmental or developmental) to the functional ramifications. Moreover, unlike Mendelian disorders where a few genetic variants located within the coding region of the genome can cause the disease, most complex diseases occur due to complicated mechanisms involved in TG regulation. Thus, the approaches I have developed and described in this dissertation can be utilized to unravel such mechanisms by taking advantage of big omics data sources such as genomics, transcriptomics, epigenomics and proteomics.

Besides the methodologies themselves, the data that I have aggregated in the process of developing can also prove to be very useful for studying complex diseases. For instance, I have compiles TFBS corresponding to several TFs in atleast 6 different cell-lines. Along with it, I also have Hi-C data for two of these cell-lines. If one can identify risk loci for their disease of trait, then they could easily use these data sources to derive functional relevance of these variants. More specifically, one determine if any of the TFBS are being perturbed by the risk variants by using the TFBS data and our AGNet models. Furthermore, the TGs for these variants can also be determined by integrating the chromatin looping data and identifying the contact points of the TFBS with the TG promoters. Using the epigenetic profiles for *cis*-regulatory regions from REMC and transcriptomic data available for several tissues from GTEx, one can further solidify the functional annotation of the risk variants. Additionally, interacting pairs of epistatic eQTLs(ieQTLs) found associated with complex diseases, can be characterized based on the pairwise TRMs, we had detected in **Chapter 3**. Furthermore, one can also derive the mode of regulation for these ieQTLs based on analyses presented in **3.3C4**. Such a characterization effort of ieQTLs could prove to be very useful to study the concept of epistasis, which has been rather controversial in the field of human genetics.

Furthermore, both TFXcan and TFKin can also be combined with other methods described in this dissertation. TFXcan and EpiXcan can be merged to build TWAS models containing TF specific regulatory scores and REMC derived broad epigenetic priors for common non-coding variants. Additionally, one can also utilize methods such as STAAR and FST, in conjunction with TFKin, to incorporate other functional annotations for rare variants, besides TF based regulatory information, to perform rare variants association

tests. Lastly, TFXcan and TFKin can be utilized to analyze the respective influence of common and rare variants on TF based TG regulation for complex diseases. Alzheimer's Disease(AD) is a debilitating neurological condition affecting millions of people around the world every year. The late onset AD(LOAD) is highly heritable (60% heritability), with many known genetic variants associated with the trait. However, the mechanism of LOAD is still unknown with several hypothesis being put forward. Using TFXcan and TFKin, one can fill in that gap by identifying common and rare variants significantly influencing TF based regulatory mechanisms leading to TG expression dysregulation subsequently. Additionally, both TFXcan and TFKin are gene based tests, in that they produce association results for common and rare variants with respect to genes and not phenotypic traits. However, one can easily adapt these approaches for finding significant trait associations. TFXcan can be used to impute TG expression for individuals within a dataset containing phenotype information. This imputed expression reflects the influence of common variants on TG expression via different TF based regulatory mechanisms. A gene based association test can be performed using this imputed TG expression with the trait outcome to identify significantly dysregulated TGs based on common variant information. The kernel based approach TFKin can be similarly used to estimate expression values from kinship matrices reflecting the influence of rare variants on TG expression regulation. A gene based association test, similar to the one described for TFXcan, can then be used to identify TGs associated with a given phenotypic trait based on *cis*-regulatory rare variants. Additionally, TFKin does not have to be limited to rare variants only. One can also include common variants while estimating the kernel to generate a more comprehensive kinship matrix capturing the effect of all the *cis*-regulatory variants. I plan on applying all of these

extensions of TFXcan and TFKin to AD datasets such as ADGC and ADSP(Alzheimer's Disease Sequencing Project), which contain whole genome sequencing data for thousands of individuals. Estimating TG expression using models containing information corresponding to common variants (TFXcan) or both common and rare variants(TFKin) using these datasets can ultimately lead to identification of novel TGs associated with AD, which may serve as potential drug targets.

BIBLIOGRAPHY

1. Lambert SA, Jolma A, Campitelli LF, et al. The Human Transcription Factors. *Cell*. 2018;172(4):650-665. doi:https://doi.org/10.1016/j.cell.2018.01.029

2. Wray GA, Hahn MW, Abouheif E, et al. The Evolution of Transcriptional Regulation in Eukaryotes. *Mol Biol Evol*. 2003;20(9):1377-1419. doi:10.1093/molbev/msg140

3. Ye B, Yang G, Li Y, Zhang C, Wang Q, Yu G. ZNF143 in Chromatin Looping and Gene Regulation. *Front Genet*. 2020;11:338. doi:10.3389/fgene.2020.00338

4. Fornes O, Castro-Mondragon JA, Khan A, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2019;48(D1):D87-D92. doi:10.1093/nar/gkz1001

5. Park PJ. ChIP–seq: advantages and challenges of a maturing technology. *Nat Rev Genet*. 2009;10(10):669-680. doi:10.1038/nrg2641

6. Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res*. 2017;46(D1):D794-D801. doi:10.1093/nar/gkx1081

7. Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin*. 2014;7(1):33. doi:10.1186/1756-8935-7-33

8. Lu H, Zhou Q, He J, et al. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduct Target Ther*. 2020;5(1):213. doi:10.1038/s41392-020-00315-3

9. Oughtred R, Stark C, Breitkreutz B-J, et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res*. 2018;47(D1):D529-D541. doi:10.1093/nar/gky1079

10. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47(D1):D607-D613. doi:10.1093/nar/gky1131

11. Kerrien S, Aranda B, Breuza L, et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Res*. 2012;40(Database issue):D841-D846. doi:10.1093/nar/gkr1088

12. Yeung KY, Medvedovic M, Bumgarner RE. From co-expression to co-regulation: how many microarray experiments do we need? *Genome Biol*. 2004;5(7):R48. doi:10.1186/gb-2004-5-7-r48

13. Snel B, van Noort V, Huynen MA. Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. *Nucleic Acids Res*. 2004;32(16):4725-4731. doi:10.1093/nar/gkh815

14. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*. 2012;41(D1):D991-D995. doi:10.1093/nar/gks1193

15. Kadauke S, Blobel GA. Chromatin loops in gene regulation. *Biochim Biophys Acta*. 2009;1789(1):17-25. doi:10.1016/j.bbagrm.2008.07.002

16. Li B, Carey M, Workman JL. The Role of Chromatin during Transcription. *Cell*. 2007;128(4):707-719. doi:10.1016/j.cell.2007.01.015

17. Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods*.

2012;58(3):268-276. doi:https://doi.org/10.1016/j.ymeth.2012.05.001

18. Budden DM, Hurley DG, Crampin EJ. Predictive modelling of gene expression from transcriptional regulatory elements. *Brief Bioinform*. 2014;16(4):616-628. doi:10.1093/bib/bbu034

19. Schmidt F, Gasparoni N, Gasparoni G, et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res*. 2016;45(1):54-66. doi:10.1093/nar/gkw1061

20. Cheng C, Gerstein M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res*. 2011;40(2):553-568. doi:10.1093/nar/gkr752

21. Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci*. 2009;106(51):21521 LP - 21526. doi:10.1073/pnas.0904863106

22. Zhang L-Q, Li Q-Z. Estimating the effects of transcription factors binding and histone modifications on gene expression levels in human cells. *Oncotarget*. 2017;8(25):40090-40103. doi:10.18632/oncotarget.16988

23. Robins G, Lanchantin J, Singh R, Qi Y. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*. 2016;32(17):i639-i648. doi:10.1093/bioinformatics/btw427

24. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet*. 2018;50(8):1171-1179. doi:10.1038/s41588-018-0160-6

25. Fuxman Bass JI, Tamburino AM, Mori A, et al. Transcription factor binding to Caenorhabditis elegans first introns reveals lack of redundancy with gene promoters. *Nucleic Acids Res*. 2014;42(1):153-162. doi:10.1093/nar/gkt858

26. Prazak L, Fujioka M, Gergen JP. Non-additive interactions involving two distinct elements mediate sloppy-paired regulation by pair-rule transcription factors. *Dev Biol*. 2010;344(2):1048-1059. doi:10.1016/j.ydbio.2010.04.026

27. Gerstein MB, Kundaje A, Hariharan M, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*. 2012;489(7414):91-100. doi:10.1038/nature11245

28. Guo Y, Gifford DK. Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. *BMC Genomics*. 2017;18(1):45. doi:10.1186/s12864-016-3434-3

29. Whitington T, Frith MC, Johnson J, Bailey TL. Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res*. 2011;39(15):e98-e98. doi:10.1093/nar/gkr341

30. Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*. 2015;32(1):1-8. doi:10.1093/bioinformatics/btv544

31. Mortazavi A, Pepke S, Jansen C, et al. Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps. *Genome Res*. 2013;23(12):2136-2148. doi:10.1101/gr.158261.113

32. Giannopoulou EG, Elemento O. Inferring chromatin-bound protein complexes from genome-wide binding assays. *Genome Res*. 2013;23(8):1295-1306. doi:10.1101/gr.149419.112

33. Lee TI, Young RA. Transcriptional Regulation and Its Misregulation in Disease. *Cell*. 2013;152(6):1237-1251. doi:10.1016/j.cell.2013.02.014

34. Laurila K, Lähdesmäki H. Systematic analysis of disease-related regulatory mutation classes reveals distinct effects on transcription factor binding. *In Silico Biol*. 2009;9(4):209-224.

35. Huo Y, Li S, Liu J, Li X, Luo X-J. Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. *Nat Commun*. 2019;10(1):670. doi:10.1038/s41467-019-08666-4

36. Latchman DS. Transcription-Factor Mutations and Disease. *N Engl J Med*. 1996;334(1):28-33. doi:10.1056/NEJM199601043340108

37. Zhang F, Lupski JR. Non-coding genetic variants in human disease. *Hum Mol Genet*. 2015;24(R1):R102-10. doi:10.1093/hmg/ddv259

38. Carrasco Pro S, Bulekova K, Gregor B, Labadorf A, Fuxman Bass JI. Prediction of genome-wide effects of single nucleotide variants on transcription factor binding. *Sci Rep*. 2020;10(1):17632. doi:10.1038/s41598-020-74793-4

39. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. 2011;27(7):1017-1018. doi:10.1093/bioinformatics/btr064

40. Thomas-Chollier M, Defrance M, Medina-Rivera A, et al. RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res*. 2011;39(suppl_2):W86-W91. doi:10.1093/nar/gkr377

41. Frith MC, Fu Y, Yu L, Chen J-F, Hansen U, Weng Z. Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res*. 2004;32(4):1372-1381. doi:10.1093/nar/gkh299

42. Martin V, Zhao J, Afek A, Mielko Z, Gordân R. QBiC-Pred: quantitative predictions of transcription factor binding changes due to sequence variants. *Nucleic Acids Res*. 2019;47(W1):W127-W135. doi:10.1093/nar/gkz363

43. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods*. 2015;12(10):931-934. doi:10.1038/nmeth.3547

44. 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):831-838. doi:10.1038/nbt.3300

45. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res*. 2016;44(11):e107. doi:10.1093/nar/gkw226

46. Quang D, Xie X. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods*. 2019;166:40-47. doi:https://doi.org/10.1016/j.ymeth.2019.03.020

47. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci*. 2013;368(1620):20120362. doi:10.1098/rstb.2012.0362

48. Wainberg M, Sinnott-Armstrong N, Mancuso N, et al. Opportunities and challenges for transcriptome-wide association studies. *Nat Genet*. 2019;51(4):592-599. doi:10.1038/s41588-019-0385-z

49. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-585. doi:10.1038/ng.2653

50. van Bömmel A, Love MI, Chung H-R, Vingron M. coTRaCTE predicts co-occurring transcription factors within cell-type specific enhancers. *PLOS Comput Biol*. 2018;14(8):e1006372. https://doi.org/10.1371/journal.pcbi.1006372.

51. Zhang W, Voloudakis G, Rajagopal VM, et al. Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. *Nat Commun*. 2019;10(1):3834. doi:10.1038/s41467-019-11874-7

52. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. 2016;48(3):245-252. doi:10.1038/ng.3506

53. Barbeira AN, Dickinson SP, Bonazzola R, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun*. 2018;9(1):1825. doi:10.1038/s41467-018-03621-1

54. Bocher O, Génin E. Rare variant association testing in the non-coding genome. *Hum Genet*. 2020;139(11):1345-1362. doi:10.1007/s00439-020-02190-y

55. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am J Hum Genet*. 2014;95(1):5-23. doi:https://doi.org/10.1016/j.ajhg.2014.06.009

56. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82-93. doi:10.1016/j.ajhg.2011.05.029

57. Byrnes AE, Wu MC, Wright FA, Li M, Li Y. The value of statistical or bioinformatics annotation for rare variant association with quantitative trait. *Genet Epidemiol*. 2013;37(7):666-674. doi:10.1002/gepi.21747

58. Sun J, Zheng Y, Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. *Genet Epidemiol*. 2013;37(4):334-344. doi:10.1002/gepi.21717

59. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet*. 2019;104(3):410-421. doi:https://doi.org/10.1016/j.ajhg.2019.01.002

60. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 2012;13(4):762-775. doi:10.1093/biostatistics/kxs014

61. Li X, Li Z, Zhou H, et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet*. 2020;52(9):969-983. doi:10.1038/s41588-020-0676-4

62. He Z, Xu B, Lee S, Ionita-Laza I. Unified Sequence-Based Association Tests Allowing for Multiple Functional Annotations and Meta-analysis of Noncoding Variation in Metabochip Data. *Am J Hum Genet*. 2017;101(3):340-352. doi:https://doi.org/10.1016/j.ajhg.2017.07.011

63. Hao X, Zeng P, Zhang S, Zhou X. Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. *PLOS Genet*. 2018;14(1):e1007186. https://doi.org/10.1371/journal.pgen.1007186.

64. Ma Y, Wei P. FunSPU: A versatile and adaptive multiple functional annotation-based association test of whole-genome sequencing data. *PLOS Genet*. 2019;15(4):e1008081. https://doi.org/10.1371/journal.pgen.1008081.

65. Pedersen HK, Gudmundsdottir V, Brunak S. Pancreatic Islet Protein Complexes and Their Dysregulation in Type 2 Diabetes . *Front Genet* . 2017;8:43. https://www.frontiersin.org/article/10.3389/fgene.2017.00043.

66. Gonda TJ, Ramsay RG. Directly targeting transcriptional dysregulation in cancer. *Nat Rev Cancer*. 2015;15(11):686-694. doi:10.1038/nrc4018

67. Chen ZS, Chan HYE. Transcriptional dysregulation in neurodegenerative diseases: Who tipped the balance of Yin Yang 1 in the brain? *Neural Regen Res*. 2019;14(7):1148-1151. doi:10.4103/1673-5374.251193

68. Ramsingh AI, Manley K, Rong Y, Reilly A, Messer A. Transcriptional dysregulation of inflammatory/immune pathways after active vaccination against Huntington′s disease. *Hum Mol Genet*. 2015;24(21):6186-6197. doi:10.1093/hmg/ddv335

69. Forrest ARR, Kawaji H, Rehli M, et al. A promoter-level mammalian expression atlas. *Nature*. 2014;507(7493):462-470. doi:10.1038/nature13182

70. Lettre G, Rioux JD. Autoimmune diseases: insights from genome-wide association studies. *Hum Mol Genet*. 2008;17(R2):R116-R121. doi:10.1093/hmg/ddn246

71. Liang B, Ding H, Huang L, Luo H, Zhu X. GWAS in cancer: progress and challenges. *Mol Genet Genomics*. 2020;295(3):537-561. doi:10.1007/s00438-020-01647-z

72. Tan M-S, Jiang T, Tan L, Yu J-T. Genome-wide association studies in neurology. *Ann Transl Med*. 2014;2(12):124. doi:10.3978/j.issn.2305-5839.2014.11.12

73. Xue A, Wu Y, Zhu Z, et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat Commun*. 2018;9(1):2941. doi:10.1038/s41467-018-04951-w

74. Niemi MEK, Martin HC, Rice DL, et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature*. 2018;562(7726):268-271. doi:10.1038/s41586-018-0566-4

75. Emmert-Streib F, Dehmer M, Haibe-Kains B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks . *Front Cell Dev Biol* . 2014;2:38. https://www.frontiersin.org/article/10.3389/fcell.2014.00038.

76. Glass K, Huttenhower C, Quackenbush J, Yuan G-C. Passing Messages between Biological Networks to Refine Predicted Interactions. *PLoS One*. 2013;8(5):e64832. https://doi.org/10.1371/journal.pone.0064832.

77. Schmidt F, Kern F, Schulz MH. Integrative prediction of gene expression with chromatin accessibility and conformation data. *Epigenetics Chromatin*. 2020;13(1):4. doi:10.1186/s13072-020-0327-0

78. Durinck S, Moreau Y, Kasprzyk A, et al. BioMart and Bioconductor: a powerful link

between biological databases and microarray data analysis. *Bioinformatics*. 2005;21(16):3439-3440. doi:10.1093/bioinformatics/bti525

79. Lappalainen T, Sammeth M, Friedländer MR, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013;501(7468):506-511. doi:10.1038/nature12531

80. Ong C-T, Corces VG. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet*. 2014;15(4):234-246. doi:10.1038/nrg3663

81. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*. 2012;44(8):955-959. doi:10.1038/ng.2354

82. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. *Bioinformatics*. 2015;31(5):782-784. doi:10.1093/bioinformatics/btu704

83. Rouillard AD, Gundersen GW, Fernandez NF, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*. 2016;2016. doi:10.1093/database/baw100

84. Han H, Cho J-W, Lee S, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res*. 2018;46(D1):D380-D386. doi:10.1093/nar/gkx1013

85. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. *J R Stat Soc Ser B (Statistical Methodol*. 2005;67(2):301-320. http://www.jstor.org/stable/3647580.

86. Battle A, Mostafavi S, Zhu X, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res*. 2014;24(1):14-24. doi:10.1101/gr.155192.113

87. Robinson PJ, Trnka MJ, Bushnell DA, et al. Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex. *Cell*. 2016;166(6):1411-1422.e16. doi:https://doi.org/10.1016/j.cell.2016.08.050

88. Schacht T, Oswald M, Eils R, Eichmüller SB, König R. Estimating the activity of transcription factors by the effect on their target genes. *Bioinformatics*. 2014;30(17):i401-i407. doi:10.1093/bioinformatics/btu446

89. Rose AB. Introns as Gene Regulators: A Brick on the Accelerator . *Front Genet* . 2019;9:672. https://www.frontiersin.org/article/10.3389/fgene.2018.00672.

90. Reyes A, Huber W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res*. 2017;46(2):582-592. doi:10.1093/nar/gkx1165

91. Keilwagen J, Posch S, Grau J. Accurate prediction of cell type-specific transcription factor binding. *Genome Biol*. 2019;20(1):9. doi:10.1186/s13059-018-1614-y

92. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*. 2012;13(9):613-626. doi:10.1038/nrg3207

93. Vockley CM, McDowell IC, D'Ippolito AM, Reddy TE. A long-range flexible billboard model of gene activation. *Transcription*. 2017;8(4):261-267. doi:10.1080/21541264.2017.1317694

94. Yang G, Ma A, Qin ZS, Chen L. Application of topic models to a compendium of ChIP-Seq datasets uncovers recurrent transcriptional regulatory modules. *Bioinformatics*. 2020;36(8):2352-2358. doi:10.1093/bioinformatics/btz975

95. Izumi K. Disorders of Transcriptional Regulation: An Emerging Category of Multiple Malformation Syndromes. *Mol Syndromol*. 2016;7(5):262-273. doi:10.1159/000448747

96. Pennypacker KR. AP-1 transcription factor complexes in CNS disorders and development. *J Fla Med Assoc*. 1995;82(8):551-554.

97. Trop-Steinberg S, Azar Y. AP-1 Expression and its Clinical Relevance in Immune Disorders and Cancer. *Am J Med Sci*. 2017;353(5):474-483. doi:10.1016/j.amjms.2017.01.019

98. Tsang M, Cheng D, Liu Y. Detecting Statistical Interactions from Neural Network Weights. 2017.

99. Kulakovskiy I V, Vorontsov IE, Yevshin IS, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res*. 2018;46(D1):D252-D259. doi:10.1093/nar/gkx1106

100. Vaqueiro AC, de Oliveira CP, Cordoba MS, et al. Expanding the spectrum of TBL1XR1 deletion: Report of a patient with brain and cardiac malformations. *Eur J Med Genet*. 2018;61(1):29-33. doi:https://doi.org/10.1016/j.ejmg.2017.10.008

101. Gordon S, Akopyan G, Garban H, Bonavida B. Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. *Oncogene*. 2006;25(8):1125-1142. doi:10.1038/sj.onc.1209080

102. Rodier G, Kirsh O, Baraibar M, et al. The Transcription Factor E4F1 Coordinates CHK1-Dependent Checkpoint and Mitochondrial Functions. *Cell Rep*. 2015;11(2):220-233. doi:10.1016/j.celrep.2015.03.024

103. Hill CS. Transcriptional Control by the SMADs. *Cold Spring Harb Perspect Biol*. 2016;8(10):a022079. doi:10.1101/cshperspect.a022079

104. de Dieuleveult M, Miotto B. DNA Methylation and Chromatin: Role(s) of Methyl-CpG-Binding Protein ZBTB38. *Epigenetics insights*. 2018;11:2516865718811117-2516865718811117. doi:10.1177/2516865718811117

105. Ropero S, Ballestar E, Alaminos M, Arango D, Schwartz S, Esteller M. Transforming pathways unleashed by a HDAC2 mutation in human cancer. *Oncogene*. 2008;27(28):4008-4012. doi:10.1038/onc.2008.31

106. Ismail T, Lee H-K, Kim C, Kwon T, Park TJ, Lee H-S. KDM1A microenvironment, its oncogenic potential, and therapeutic significance. *Epigenetics Chromatin*. 2018;11(1):33. doi:10.1186/s13072-018-0203-3

107. Icardi L, Mori R, Gesellchen V, et al. The Sin3a repressor complex is a master regulator of STAT transcriptional activity. *Proc Natl Acad Sci*. 2012;109(30):12058 LP - 12063. doi:10.1073/pnas.1206458109

108. Akhtar W, Veenstra GJC. TBP-related factors: a paradigm of diversity in transcription initiation. *Cell Biosci*. 2011;1(1):23. doi:10.1186/2045-3701-1-23

109. Giannoudis A, Malki MI, Rudraraju B, et al. Activating transcription factor-2 (ATF2) is a key determinant of resistance to endocrine treatment in an in vitro model of breast cancer.

Breast Cancer Res. 2020;22(1):126. doi:10.1186/s13058-020-01359-7

110. Shaulian E, Karin M. AP-1 in cell proliferation and survival. *Oncogene*. 2001;20(19):2390-2400. doi:10.1038/sj.onc.1204383

111. Hernandez JM, Floyd DH, Weilbaecher KN, Green PL, Boris-Lawrie K. Multiple facets of junD gene expression are atypical among AP-1 family members. *Oncogene*. 2008;27(35):4757-4767. doi:10.1038/onc.2008.120

112. Shieh C, Jones N, Vanle B, et al. GATAD2B-associatedneurodevelopmental disorder (GAND): clinical and molecular insights into a NuRD-relateddisorder. *Genet Med*. 2020;22(5):878-888. doi:10.1038/s41436-019-0747-z

113. Pahl HL. Activators and target genes of Rel/NF-κB transcription factors. *Oncogene*. 1999;18(49):6853-6866. doi:10.1038/sj.onc.1203239

114. Ward JM, Ratliff ML, Dozmorov MG, et al. Human effector B lymphocytes express ARID3a and secrete interferon alpha. *J Autoimmun*. 2016;75:130-140. doi:10.1016/j.jaut.2016.08.003

115. Chen J, Liang X, Zhang S, et al. Two faces of bivalent domain regulate VEGFA responsiveness and angiogenesis. *Cell Death Dis*. 2020;11(1):75. doi:10.1038/s41419-020-2228-3

116. Yang M, Sun L, Han J, et al. Biological characteristics of transcription factor RelB in different immune cell types: implications for the treatment of multiple sclerosis. *Mol Brain*. 2019;12(1):115. doi:10.1186/s13041-019-0532-6

117. Lee CK, Smith E, Gimeno R, Gertner R, Levy DE. STAT1 affects lymphocyte survival and proliferation partially independent of its role downstream of IFN-gamma. *J Immunol*. 2000;164(3):1286-1292. doi:10.4049/jimmunol.164.3.1286

118. Jiang Y, Guo X, Liu L, et al. Metagenomic characterization of lysine acetyltransferases in human cancer and their association with clinicopathologic features. *Cancer Sci*. 2020;111(5):1829-1839. doi:10.1111/cas.14385

119. Oh YM, Kim JK, Choi S, Yoo J-Y. Identification of co-occurring transcription factor binding sites from DNA sequence using clustered position weight matrices. *Nucleic Acids Res*. 2011;40(5):e38-e38. doi:10.1093/nar/gkr1252

120. Wang M, Tai C, E W, Wei L. DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Res*. 2018;46(11):e69-e69. doi:10.1093/nar/gky215

121. Guo Y, Zhou D, Li W, Nie R, Hou R, Zhou C. Attentive gated neural networks for identifying chromatin accessibility. *Neural Comput Appl*. 2020;32(19):15557-15571. doi:10.1007/s00521-020-04879-7

122. Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. In: *Empirical Methods in Natural Language Processing (EMNLP)*. ; 2014:1532-1543. http://www.aclweb.org/anthology/D14-1162.

123. Wagih O, Merico D, Delong A, Frey BJ. Allele-specific transcription factor binding as a benchmark for assessing variant impact predictors. *bioRxiv*. January 2018:253427. doi:10.1101/253427

124. Zeng H, Hashimoto T, Kang DD, Gifford DK. GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics*. 2016;32(4):490-496. doi:10.1093/bioinformatics/btv565

125. Lee D, Gorkin DU, Baker M, et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet*. 2015;47(8):955-961. doi:10.1038/ng.3331

126. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47(9):1091-1098. doi:10.1038/ng.3367

127. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. 2012;28(10):1353-1358. doi:10.1093/bioinformatics/bts163

128. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317-330. doi:10.1038/nature14248

129. Li YI, van de Geijn B, Raj A, et al. RNA splicing is a primary link between genetic variation and disease. *Science (80-)*. 2016;352(6285):600 LP - 604. doi:10.1126/science.aad9417

130. Hernandez RD, Uricchio LH, Hartman K, Ye C, Dahl A, Zaitlen N. Ultrarare variants drive substantial cis heritability of human gene expression. *Nat Genet*. 2019;51(9):1349-1355. doi:10.1038/s41588-019-0487-7

131. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*. 2004;305(5685):869-872. doi:10.1126/science.1099870

132. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet*. 2008;40(6):695-701. doi:10.1038/ng.f.136

133. Ji W, Foo JN, O'Roak BJ, et al. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet*. 2008;40(5):592-599. doi:10.1038/ng.118

134. Lord J, Lu AJ, Cruchaga C. Identification of rare variants in Alzheimer's disease. *Front Genet*. 2014;5:369. doi:10.3389/fgene.2014.00369

135. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315. doi:10.1038/ng.2892

136. Fu Y, Liu Z, Lou S, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol*. 2014;15(10):480. doi:10.1186/s13059-014-0480-5