UNIVERSITY OF CALIFORNIA SAN DIEGO

Machine learning approaches for relating genomic sequence to enhancer activity and function

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Jenhan Tao

Committee in charge:

Professor Christopher K. Glass, Chair
Professor Christopher Benner, Co-Chair
Professor Olivier Harismendy
Professor Bing Ren
Professor Wei Wang
Professor Sheng Zhong

2018

The Dissertation of Jenhan Tao is approved and is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California San Diego

2018

DEDICATION

I dedicate this work to everyone that gave my ideas a chance. You all could have chosen a more reasonable project and collaborator. I am thankful that you did not.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge my advisor, Chris Glass, for the energy and guidance that has allowed for the research described in this dissertation. For each of the ideas that I explored, the reasonable ones and the risky ones, Chris Glass had both encouraging words as well as meaningful advice that helped me to translate a biological problem into a rigorous computational problem. I would also like to thank my co-advisor, Chris Benner, for his role as a mentor. His honest advice and emphasis on interpretable analysis were critical for improving the coherence and rigor of my work.

Each of my committee members, Olivier Harismendy, Bing Ren, Wei Wang, and Sheng Zhong, helped me to find excitement in different aspects of my research. I am thankful that they have always found time to provide me with advice and challenge me to be a better scientist.

I would like to thank members of the Glass Lab and my collaborators who gave my ideas a chance. In particular, Greg Fonseca, who ought to be awarded another Ph.D. considering the efforts he invested in our projects. Joint efforts with Zeyang Shen were critical to jump starting my explorations into deep learning and complex neural network models. Leslie Van El's administrative efforts were also much appreciated.

Lastly, I would like to thank my friends and family for their encouragement, patience, and support. My mother and father never questioned whether this arduous Ph.D. process was right for me and only encouraged to push onward. Discussions with Evan Appleton, Justin Huang, Ernst Oberortner, and Yuan Zhao often yielded useful suggestions as well as good humor. I thank Michelle Dow for her enduring patience and for tolerating the many weekends I spent working.

Chapter 2, in part, has been submitted for publication. Fonseca, G.J.*, Tao J.*, Westin, E.M., Duttke, S.H., Spann, N.J., Strid, T., Shen, Z., Stender, J.D., Sakai, M., Link, V.M., Benner, C., Glass, C.K. Diverse motif ensembles specify non-redundant DNA binding activities of AP-1 family members in macrophages. (* These authors contributed equally to this work). The dissertation author was one of the primary investigators and authors of this study.

Chapter 3, in part, has been submitted for publication. Tao, J., Bennett, H., Fonseca. G.J.,

Shen, Z., Benner, C., Glass, C.K. A method for describing transcription factor binding specificity as a set of DNA motifs. The dissertation author was the primary investigator and author of this study.

Chapters 4, in part, will be submitted for publication. Tao, J.*, Fonseca, G.J.*, Duttke, S.H., Hoeksema, M.A., Shen, Z.,Bennett, H., Benner, C., Glass, C.K. Learning composition rules for macrophage enhancers with neural attention. (* These authors contributed equally to this work). The dissertation author was one of the primary investigators and authors of this study.

VITA

2009	Folsom High School, Folsom, CA
2012	B.S. Bioengineering, University of California Berkeley, Berkeley, CA
2018	Ph.D. Bioinformatics, University of California San Diego, San Diego, CA

PUBLICATIONS

J. Tao*, G. J. Fonseca*, C. Benner, and C. K. Glass.Identifying composition rules for transcription factor circuits that control macrophage signal response with deep learning.*International Workshop on Bio-Design Automation Proceedings*, August 2018.

J. Tao*, G. J. Fonseca*, C. Benner, and C. K. Glass. Learning Composition Rules for Mammalian Circuits with Neural Attention.*Association for the Advancement of Artificial Intelligence Fall Symposium 2018 Proceedings*, October 2018.

J. Tao*, G. J. Fonseca*, C. Benner, and C. K. Glass.Identifying composition rules for transcription factor circuits that control macrophage signal response with deep learning.*International Workshop on Bio-Design Automation Proceedings*, August 2018.

A. Shemer, J. Grozovski, T. Leng Tay, **J. Tao***, A. Volaski, P. Suess, A. Ardura-Fabregat, M. Gross, J.-S. Kim, E. David, L. Chappell-Maor, L. Thielecke, C. K. Glass, K. Cornils, M. Prinz, and S. Jung. Engrafted parenchymal brain macrophages differ from host microglia in transcriptome, epigenome and responsiveness to challenge. *bioRxiv*, 2018.

J. C. M. Schlachetzki*, I. Prots*, **J. Tao***, H. B. Chun, K. Saijo, D. Gosselin, B. Winner, C. K. Glass, and J. Winkler. A monocyte gene expression signature in the early clinical course of Parkinson's disease. *Scientific Reports*, 8(1):10757, 2018.

E. D. Muse*, S. Yu*, C. R. Edillor+, **J. Tao+**, N. J. Spann, T. D. Troutman, J. S. Seidman, A. Henke, J. T. Roland, K. A. Ozeki, B. M. Thompson, J. G. McDonald, J. Bahadorani, S. Tsimikas, T. R. Grossman, M. S. Tremblay, and C. K. Glass.Cell-specific discrimination of desmosterol and desmosterol mimetics confers selective regulation of LXR and SREBP in macrophages.*Proceedings of the National Academy of Sciences*, page 201714518, 2018

V. M. Link, S. H. Duttke, H. B. Chun, I. R. Holtman, E. Westin, M. A. Hoeksema, Y. Abe, D. Skola, C. E. Romanoski, **J. Tao**, G. J. Fonseca, T. D. Troutman, N. J. Spann, T. Strid, M. Sakai, M. Yu, R. Hu, R. Fang, D. Metzler, B. Ren, and C. K. Glass. Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide Mechanisms that Control Transcription Factor Binding and Function. *Cell*, 173(7):1796–1809.e17, 2018.

G. J. Fonseca*, **J. Tao***, E. M. Westin, S. H. Duttke, N. J. Spann, T. Strid, Z. Shen, J. D. Stender, V. M. Link, C. Benner, and C. K. Glass. Diverse motif ensembles specify non-redundant DNA binding activities of AP-1 family members in macrophages. *BioRxiv*, pages 1–28, 2018.

Y. Oishi, N. J. Spann, V. M. Link, E. D. Muse, T. Strid, C. Edillor, M. J. Kolar, T. Matsuzaka, S. Hayakawa, **J. Tao**, M. U. Kaikkonen, A. F. Carlin, M. T. Lam, I. Manabe, H. Shimano, A. Saghatelian, and C. K. Glass. SREBP1 Contributes to Resolution of Pro-inflammatory TLR4 Signaling by Reprogramming Fatty Acid Metabolism. *Cell Metabolism*, pages 1–16, 2016.

D. Z. Eichenfield, T. D. Troutman, V. M. Link, M. T. Lam, H. Cho, D. Gosselin, N. J. Spann, H. P. Lesch, **J. Tao**, J. Muto, R. L. Gallo, R. M. Evans, and C. K. Glass. Tissue damage drives co-localization of NF-κB, Smad3, and Nrf2 to direct Rev-erb sensitive wound repair in mouse macrophages. *eLife*, 5(JULY):1–30, 2016.

E. Appleton, **J. Tao** H, T. Haddock, and D. Densmore. Interactive assembly algorithms for molecular cloning. *Nat Methods*, 11(6):657–+, 2014.

E. Appleton, **J. Tao**, F. C. Wheatley, D. H. Desai, T. M. Lozanoski, P. D. Shah, J. A. Awtry, S. S. Jin, T. L. Haddock, and D. M. Densmore. Owl: Electronic datasheet generator. *ACS Synthetic Biology*, 3(12):966–968, 2014.

ABSTRACT OF THE DISSERTATION

Machine learning approaches for relating genomic sequence to enhancer activity and function

by

Jenhan Tao

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2018

Professor Christopher K. Glass, Chair
Professor Christopher Benner, Co-Chair

Despite the advent of high throughput genomics technology and the wealth of data characterizing transcription that followed, it remains difficult to relate genomic sequence to transcriptional activity. Next generation sequencing techniques, including ChIP-seq, RNA-seq, and ATAC-seq, have enabled high resolution mapping of transcriptional activity, including RNA expression and histone modifications, as well as the localization of transcription factors and DNA binding proteins that regulate transcription. By integrating of these activity maps using statistical methods and high-performance computing, a model has emerged in which transcription factors recognize and bind to short DNA sequence motifs (words) to recruit cellular machinery such as RNA polymerase, which is necessary for transcription. Previous studies have also demonstrated that transcription

factors often bind together in a cell type and context specific manner, setting the foundation for a genomic grammar in which combinations of transcription factors recognize ”sentences” that specify cell type and context specific transcriptional activity. Using this foundational model as our starting point, we devised a machine learning framework named TBA (a Transcription factor Binding Analysis), for investigating the sequence specificity of transcription factors by jointly weighing the contributions of hundreds of DNA motifs. We applied TBA to a systematic map of the binding profiles for the AP-1 transcription factor family, which share a conserved DNA binding domain. We observed that each family member demonstrated interactions with distinct sets of motifs, which varied from cell type to cell type, and in different cellular states. Next we applied the TBA framework to hundreds of transcription factor ChIP-seq data sets, demonstrating that like AP-1, transcription factors generally interact with dozens of other transcription factors genome-wide and with 3-4 transcription factors at a given locus in a cell-type specific manner. We used these findings describing transcription factor behavior to devise a neural network with an attention mechanism that calculates locus specific maps of how motifs interact to predict transcriptional activity. These studies demonstrate machine learning approaches that reveal additional insight into a transcriptional grammar that coordinates eukaryotic gene expression.

Chapter 1

Introduction

The Central Dogma of biology describes a flow of information that begins with the the genome, an instruction set that describes the attributes of a living organism, and that ultimately arrives at proteins, one of the physical manifestations of these instructions. Using the four DNA nucleotides - A (adenosine), C (cytosine), G (guanine), and T (thymine) - that serve as the alphabet of the genome, nature has composed a rich library of genomes that specifies the unique appearance and behavior of the diverse organisms that roam the earth. Additionally in complex organisms, such as mammals, there may be hundreds of different cell types, many of which are conserved, that each express the genome in a cell type specific manner. Within mammalian genomes, there are on the order of tens of thousands of genes (sequences of nucleotides) that are transcribed to RNA, which may then be translated into proteins that give rise to cellular function. Each cell type expresses a distinct repertoire of these genes, allowing for cell type specific behavior and function[20] . It remains an ongoing challenge to explain how cell type specific gene expression patterns are encoded by the genome.

A critical piece of this puzzle is RNA polymerase, an enzyme that synthesizes RNA by using DNA as a template. And so, RNA polymerase must be recruited to each gene in a cell type specific manner in order to effect cell type specific gene expression. RNA polymerase itself cannot recognize the genes, and more specifically promoters, the DNA sequences at the start of genes where RNA transcription starts[40]. RNA polymerase's localization to the promoter of genes depends in part on interactions with general transcription factors (TFIIB, TFIID, TFIIE, TFIIF, and TFIIH),

which recognize and bind to short DNA sequences located at promoters[40]. As general transcription factors are broadly expressed across all cell types, they cannot be the primary determinant of cell type specific gene expression. In addition to the general transcription factors, there are hundreds of other transcription factors that can have cell type specific patterns of expression[96]. Transcription factors are categorized into families according to conserved protein domains including their DNA binding domains[104]. Each family of transcription factors may contain dozens of members which bind to highly similar DNA sequences[33;82]. Despite recognizing similar DNA sequences, members of a transcription factor family can have distinct biological functions and bind to distinct locations in the genome, suggesting that a transcription factor's DNA binding domain is not not the sole determinant of its binding targets[27;35;37;73;101].

Previous work has identified combinations of transcription factors that are critical to the identity of a cell type. These combinations of transcription factors, which are also referred to as lineage determining transcription factors. Examples of lineage determining transcription factors include the Yamanaka factors, Oct2, SOX4, and Nanog, which are critical for the identity of pluripotent stem cells as well as PU.1 and CEBPa, which are important in macrophage cells[28;29;89]. Studies have shown that the binding sites of each lineage determining transcription factors often are in close proximity, suggesting collaborative binding activity[29]. Indeed, studies on the effect of natural genetic variation in diverse mouse strains have shown that a disruption in the binding site of one lineage determining transcription factor can affect the binding of another[28;65]. Lineage determining transcription factors are also referred to as pioneer factors as they are thought to be able to make an closed chromatin, inaccessible region of the genome, accessible to other transcription factors, allowing transcription factors to bind to cell type specific regions[31]. Thus, collaborative binding of transcription factors is likely an important mechanism necessary for activating chromatin regions and imparting cell type specific gene expression[29]. The phenomena of collaborative binding suggests that cell type specific gene expression can be encoded as arrangements of transcription factor motifs at promoters as well as enhancers. Enhancers are regulatory DNA sequences located throughout the genome that are thought to modulate gene expression levels by looping to interact

with promoters[31].

High throughput sequencing technology has allowed for us to examine the workings of the genome at unprecedented resolution[83]. In addition to maps of the genome that have been completed for humans and model organisms, sequencing technology has allowed for the interrogation of the state of the genome, including at promoters and enhancers[79]. Sequencing of RNA allows for the construction of cell type specific gene expression profiles. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) allows for the isolation of DNA bound by a specific transcription factor, with sequencing to map the binding sites of a transcription factor. Analysis of sequences enriched at a transcription factor's binding sites allows for the identification of DNA motifs recognized by that transcription factor[31]. The DNA binding motif of hundreds of transcription factors are available in several databases[38;68;102]. ChIP-seq can also be used to epigenetic modifications of the genome such as histone methylation and acetylation[110]. Assays such as Assay for Transposase Accessible Chromatin (ATAC-Seq), MNase-Seq and DNase-Seq allow for the assessment of which regions of the genome are accessible for transcription factor binding[12;36;45;106].

The scale of high throughput sequencing data, has made the study of the genome an increasingly computational field of study where big data has been matched with high performance computing. Machine learning has emerged as a promising technique for relating genomic attributes such as DNA sequence, epigenetic modifications, and transcriptional activity using high throughput sequencing data[1;22;46;55;76]. Machine learning is a natural choice for analyzing high throughput genomics data for two primary reasons. First, machine learning requires the use of large data sets, which is easily met by the scale of sequencing data. Second, machine learning models can learn complex, nonlinear relationships in large data sets, which allows for the modelling of sophisticated biological phenomena influenced by multiple factors such as the combinatorial binding of transcription factors or enhancer activation[1;46;76]. Despite the scale of sequencing data, it remains a challenge to systematically profile all the components that may play a role in regulating transcription. Just within the nucleus, there are hundreds of transcription factors that would need to be profiled using ChIP-seq. And so, sequence based models that leverage existing databases of

DNA binding motifs to infer transcription factor binding are a powerful tool for investigating how transcriptional regulation is encoded by the genome [1;22;46;55;76].

This work began with the broad question - how is the genome interpretted or read? Over time, this work became focused on the identification of sequence elements that can be used to predict genomic features such as transcription factor binding, accessible regions of the genome, and enhancer activation. We began with the puzzle of how members of the AP-1 transcription factor family can recognize the same DNA sequence, yet bind to distinct regions in the genome. As part of our solution to this puzzle, we developed a machine learning model, which used a library of DNA motifs to learn combinations of DNA motifs that are enriched at the binding sites of each AP-1 family member. In addition to interactions with lineage determining transcription factors, we found that each family member potentially interacted with dozens of other transcription factors. This was true not just for the AP-1 family members, but broadly for hundreds of transcription factors whose binding sites were mapped by the ENCODE consortium. Using the same machine learning approach, we were able to accurately predict which regions of the genome are accessible. These studies led us to assess whether combinations of DNA motifs, and implicitly the collaborative binding of transcription factors, can be used to predict enhancer activity. We then devised a neural network with an attention mechanism that computes high resolution maps of how DNA motifs interact with one another to predict transcriptional activity. Collectively, these studies demonstrate both the promise of machine learning techniques for interpreting high throughput sequencing data as well as the importance of combinations of DNA motifs in a transcriptional grammar that coordinates eukaryotic gene expression.

Chapter 2

Diverse motif ensembles specify DNA binding activities of AP-1 family members

2.1 Abstract

Mechanisms by which members of the AP-1 family of transcription factors play both redundant and non-redundant biological roles despite recognizing the same DNA sequence remain poorly understood. To address this question, we investigated the molecular functions and genome-wide DNA binding patterns of AP-1 family members in mouse macrophages. ChIP-sequencing showed overlapping and distinct binding profiles for each factor that were remodeled following TLR4 ligation. Development of a machine learning approach that jointly weighs hundreds of DNA recognition elements yielded dozens of motifs predicted to drive factor-specific binding profiles. Machine learning-based predictions were confirmed by analysis of the effects of mutations in genetically diverse mice and by loss of function experiments. These findings provide evidence that non-redundant genomic locations of different AP-1 family members in macrophages largely result from collaborative interactions with diverse, locus-specific ensembles of transcription factors and suggest a general mechanism for encoding functional specificities of their common recognition motif.

2.2 Introduction

Gene expression is controlled by sequence-specific transcription factors (TFs) which bind to promoters and distal enhancer elements[31;59;84]. Genome wide studies of regulatory regions in diverse cell types suggest the existence of hundreds of thousands of enhancer sites within mammalian genomes. Each cell type selects a unique combination of ∼20,000 such sites that play essential roles in determining that cells identity and functional potential[2;53;79;99]. Selection and activation of cell-specific enhancers and promoters is achieved through combinatorial actions of the available sequence-specific TFs[5;23;28;42;54;95;103].

TFs are organized into families according to conserved protein domains including their DNA binding domains (DBD)[104]. Each family may contain dozens of members which bind to similar or identical DNA sequences[33;82]. An example is provided by the AP-1 family, which is composed of 15 monomers subdivided into five subfamilies based on amino acid sequence similarity: Jun (Jun, JunB, JunD), Fos (Fos, FosL1, FosL2, FosB), BATF (BATF, BATF2, BATF3), ATF (ATF2, ATF3, ATF4, ATF7) and Jdp2[8;32;35;78;92]. AP-1 binds DNA as an obligate dimer through a conserved bZIP domain. All possible dimer combinations can form with the exception of dimers within the Fos subfamily[75]. The DBD of each monomer of the AP-1 dimer recognizes half of a palindromic DNA motif separated by one or two bases (TCASTGA and TCASSTGA)[26;33;57;72;82]. Previous work has shown that dimers formed from Jun and Fos subfamily members bind the same motif[33]. Given a conserved DBD, and the ability to form heterodimers, it naturally follows that different AP-1 dimers share regulatory activities. However, co-expressed family members can play distinct roles[27;35;37;73;101]. For example, Jun and Fos are co-expressed during hematopoiesis, but knockout of Jun results in an increase in hematopoiesis whereas knockout of Fos has the opposite effect[35;37;73;101]. The basis for non-redundant activities of different AP-1 dimers and heterodimers remains poorly understood.

Specific AP-1 factors have been shown to form ternary complexes with other TFs such as IRF, NFAT and Ets proteins, resulting in binding to composite recognition elements with fixed

spacing[9;71;98]. However, recent studies examining the effects of natural genetic variation suggested that perturbations in the DNA binding of Jun in bone marrow derived macrophages are associated with mutations in the motifs of dozens of TFs that occurred with variable spacing[65]. These observations raise the general question of whether local ensembles of TFs could be determinants of differential binding and function of specific AP-1 family members. To explore this possibility, we examined the genome-wide functions and DNA binding patterns of co-expressed AP-1 family members in resting and activated mouse macrophages. In parallel, we developed a machine learning model, called a Transcription Factor Binding Analysis (TBA), that integrates the affinities of hundreds of TF motifs and learns to recognize motifs associated with the binding of each AP-1 monomer genome-wide. By interrogating our model, we identified DNA binding motifs of candidate collaborating TFs that influence specific binding patterns for each AP-1 monomer that could not be identified with conventional motif analysis. We confirmed these predictions functionally by leveraging the natural genetic variation between C57BL/6J and BALB/cJ mice, and observing the effects of single nucleotide polymorphisms (SNPs) and short insertions or deletions (InDels) on AP-1 binding. Finally, we confirm the models prediction of PPARγ binding being specifically associated with the selection of a single family member, Jun, using PPARγ-deficient macrophages.

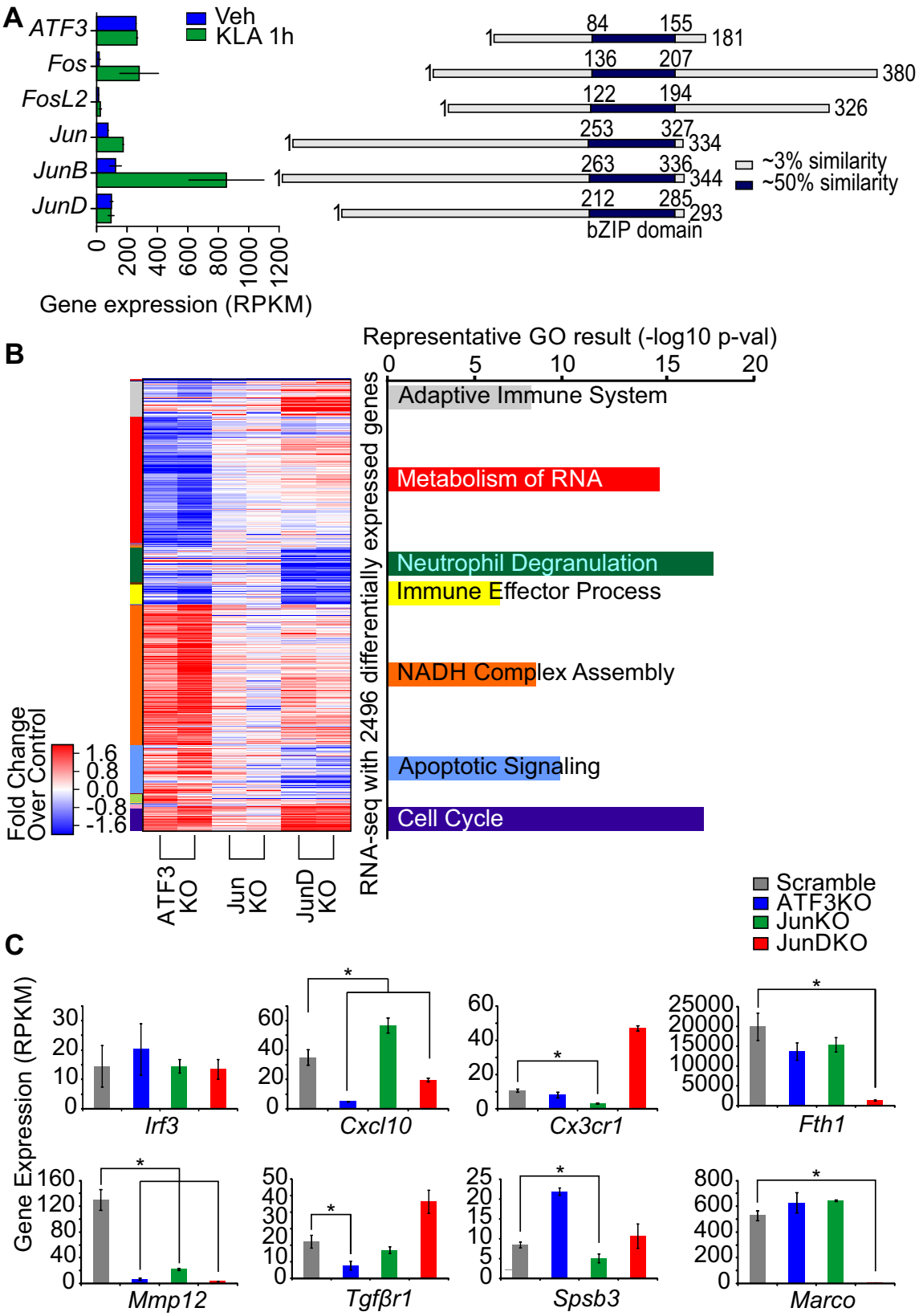
2.3 Results

2.3.1 AP-1 family members have distinct regulatory functions in macrophages

AP-1 family members are ubiquitously expressed with each cell type selecting a subset of family members (monomers), which make up the AP-1 dimer. Each family member shares a conserved DNA binding and dimerization domain but are dissimilar outside of the basic leucine zipper (bZIP domain, Fig. 2.1A). For this study, we will focus on Thioglycollate elicited macrophages (TGEMs). TGEMs, which are a classical primary macrophage population, are produced by injection of thioglycolate into the peritoneal space. Macrophages are then recruited to the peritoneum and can be easily isolated by flushing the peritoneal cavity three days after treatment. RNA-seq performed

on TGEMs revealed ATF3, Jun, and JunD as the most expressed AP-1 family members under basal conditions (Veh, Fig. 2.1A, Fig. 2.2A). Following activation of TGEMs with Kdo2 lipid A (KLA), a specific agonist of TLR4[77], there is a marked increase in Fos, Jun and JunB expression, consistent with AP-1 family members having context-specific roles (Fig. 2.1A).

Figure 2.1 ***(next page)*.** AP-1 proteins have overlapping and distinct transcriptional functions in macrophages. **A**. Protein alignment of monomers (right) and mRNA expression of monomers in TGEMs before and after 1-hour KLA treatment (left). **B.** Hierarchical clustering of genes that are differentially expressed in iBMDMs subjected to CRISPR mediated knockdown of the indicated AP-1 monomer with respect to scramble control. Expression values are given as the fold change with respect to scramble; values are Z-score normalized across each row. Representative functional annotations for each gene cluster are calculated using Metascape and the enrichment of each term is quantified as the negative log transform of the p-value. **C.** Expression of a subset of genes in AP-1 protein knockouts. * indicates FDR < 0.05.


A
ATF3
Fos
FosL2
Jun
JunB
JunD
Veh
KLA 1h
0
200
400
600
800
1000
1200
Gene expression (RPKM)
84
155
181
136
207
380
122
194
326
253
327
334
263
336
344
212
285
293
bZIP domain
~3% similarity
~50% similarity
B
Representative GO result (-log10 p-val)
0
5
10
15
20
Adaptive Immune System
Metabolism of RNA
Neutrophil Degranulation
Immune Effector Process
NADH Complex Assembly
Apoptotic Signaling
Cell Cycle
RNA-seq with 2496 differentially expressed genes
Fold Change Over Control
1.6
0.8
0.0
-0.8
-1.6
ATF3 KO
Jun KO
JunD KO
C
Scramble
ATF3KO
JunKO
JunDKO
Gene Expression (RPKM)
Irf3
Cxcl10
Cx3cr1
Fth1
Mmp12
Tgfβr1
Spsb3
Marco

To examine the regulatory function of individual family members, knockout cell lines for ATF3, Jun and JunD were produced using CRISPR/Cas9-mediated mutagenesis in immortalized bone marrow-derived macrophages (iBMDMs). Knockout efficiency was confirmed by western blotting (Fig. 2.2B). RNA-seq analysis identified 2496 genes differentially expressed when comparing the knockout to control cells (FDR$<$0.05, fold change $>$2, RPKM$\geq$16 Fig. 2.1B, Fig. 2.2C). Clustering of differentially expressed genes revealed distinct clusters that were affected in individual knockout cell lines, demonstrating that each family member can have distinct as well as redundant activity within a single cell type and corroborating previous studies[27;35;37;73;101]. The Jun knockout had a more modest effect on gene expression than the ATF3 and JunD knockout (125, 651, and 1564 differentially expressed genes respectively), suggesting that Jun may have more redundant activity (Fig. 2.1B and Fig. 2.2C). Each of the gene clusters was enriched for Gene Ontology terms for differing biological functions, including cell cycle, immune effector process and NADPH complex assembly (Fig. 2.1B). Examples of affected genes are shown in Figure 1C. *Mmp12* is affected by knockdown of all three factors, whereas *Marco* and *Fth1* exhibit minimal changes in expression in ATF3 and Jun KO, but decreased expression in the JunD KO iBMDMs.

2.3.2 AP-1 family members can target distinct loci in addition to overlapping loci

Given the distinct roles of individual family members in regulating macrophage transcription, we used chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) to map the binding of each family member in resting TGEMs treated with vehicle (Veh) or KLA for one hour (activated TGEMs). Not surprisingly, these experiments detected a substantial number of binding sites (n $>$ 10000, IDR $<$ 0.05) for family members with the highest mRNA expression (Fig. 2.2A, Fig. 2.4A). ATF3, Jun, and JunD binding sites were detected in both Veh and KLA treatment whereas Fos, Fosl2 and JunB bind predominantly after KLA treatment (Fig. 2.4A). Despite high RNA expression in Veh treatment, JunB protein expression was not detected in the nucleus by western blot, explaining a lack of ChIP-seq signal (Fig. 2.4B). Though ATF4 is highly expressed by RNA, we were anable to detect ATF4 by ChIP-seq using several conditions and several different antibodies

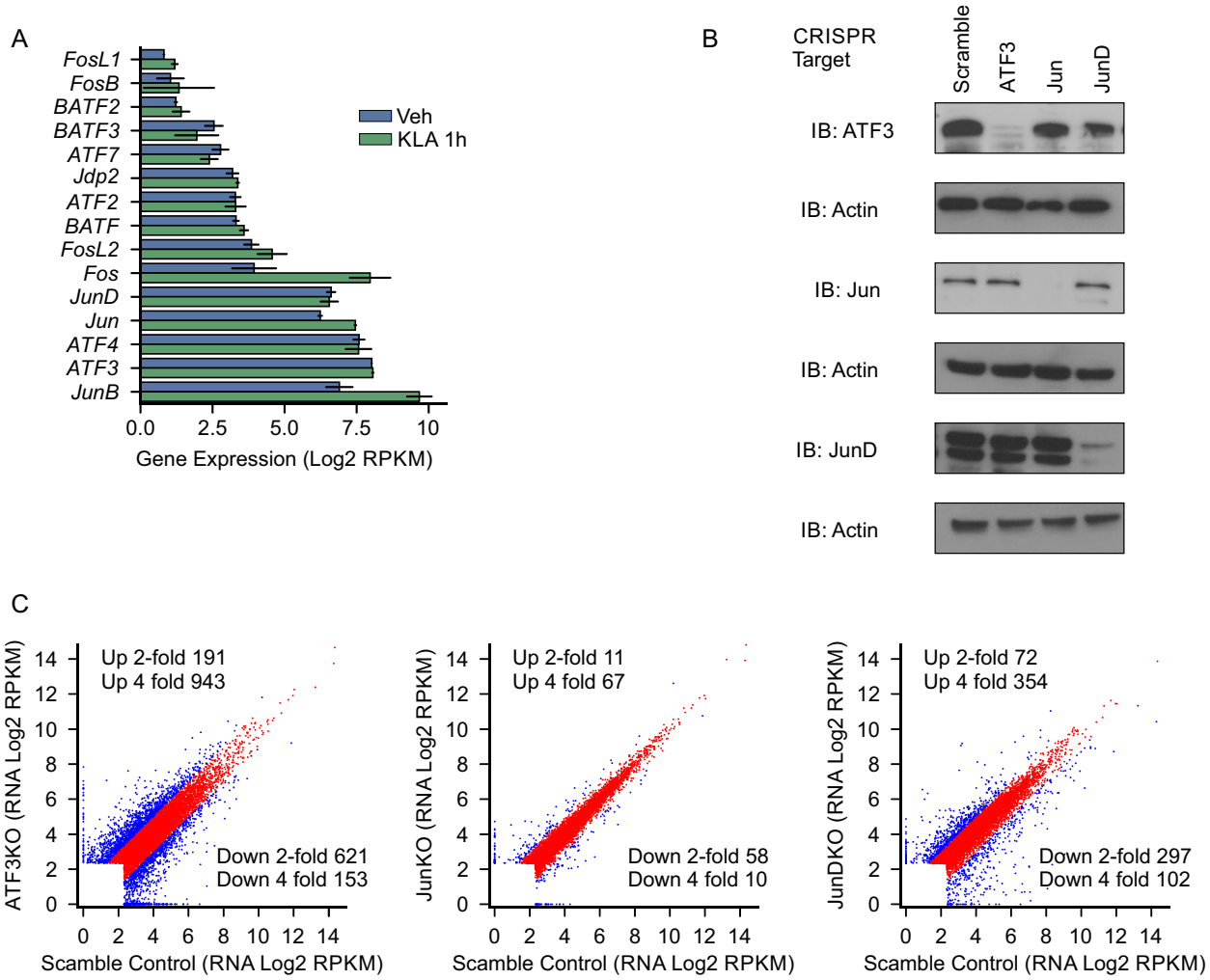
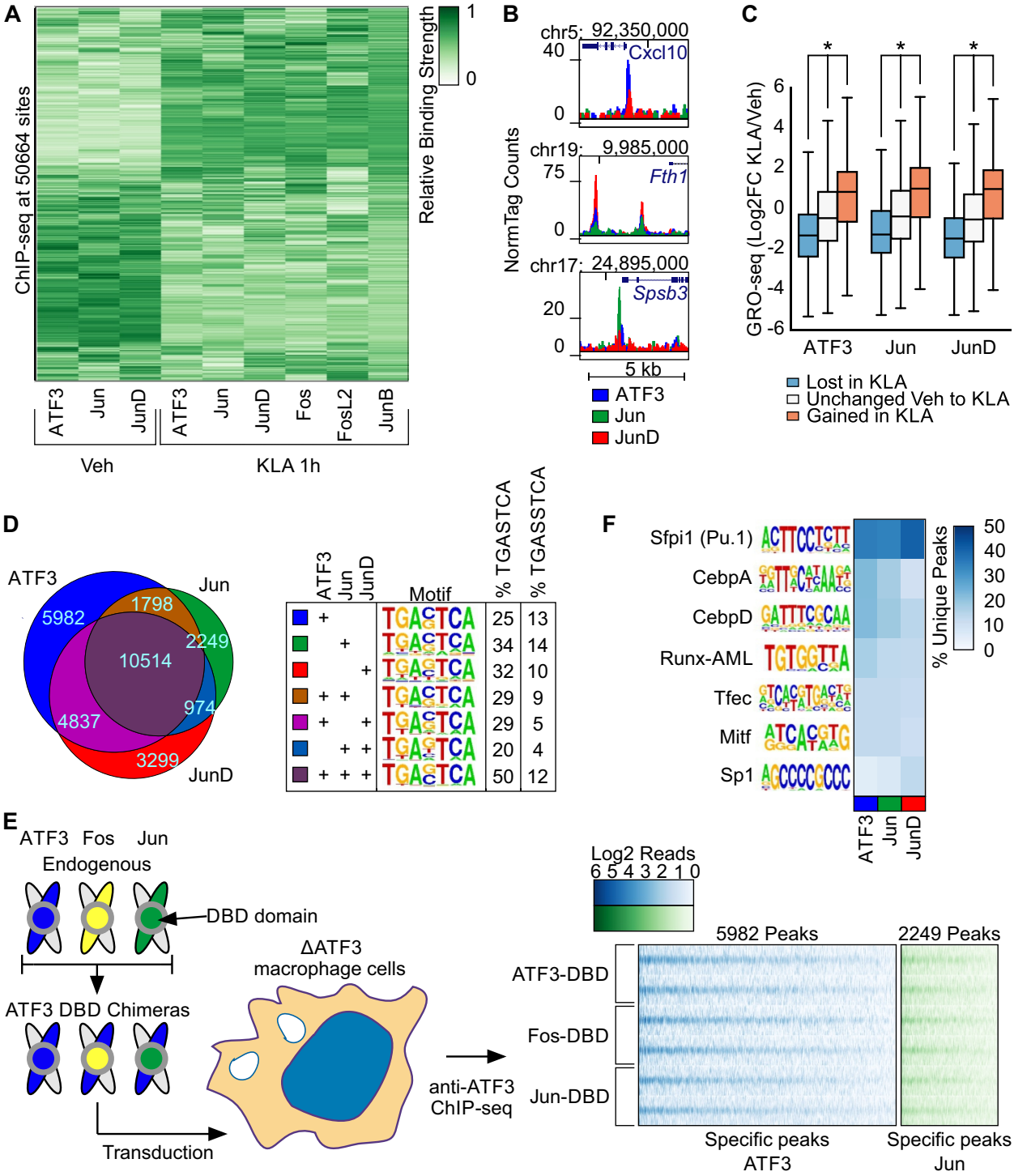


Figure 2.2. A genome-wide map of AP-1 activity in macrophages. **A.** mRNA expression of all AP-1 monomers in Vehicle and 1-hour KLA treated TGEMs by RNA-seq. **B.** Western blot analysis of ATF3, Jun and JunD expression in CRISPR mediated knockout of ATF3, Jun and JunD in iBMDM cells. **C.** Scatterplots showing RNA-seq expression between scramble control and CRISPR mediated knockout of ATF3, Jun and JunD.

(data not shown). Hierarchical clustering of all 50664 AP-1 binding sites (Fig. 2.3A) found in either Veh or KLA treated TGEMs according to the relative binding strength of the family members (normalized to a maximum of 1 at each locus) yielded distinct subclusters that highlight the specific binding patterns of AP-1 family members as well as the reorganization of AP-1 cistromes in KLA treated macrophages (Fig. 2.3A). Representative regions that show distinct binding patterns of AP-1 family members are shown (Fig. 2.3B, Fig. 2.4C).

The gain and loss of binding sites of ATF3, Jun and JunD after KLA treatment provided an opportunity to correlate changes in their DNA occupancy with local changes in enhancer activity. Changes in the expression of enhancer-associated RNAs (eRNAs) are highly correlated with changes in enhancer function and nearby gene expression[42]. To detect eRNAs, we performed Genome Run-On Sequencing (GRO-seq) in TGEMs, which provides a quantitative measure of nascent RNA[13]. We examined GRO-seq signal at ATF3, Jun and JunD binding sites exhibiting gain, loss or no change in binding after KLA treatment. In each case, AP-1 occupancy was associated with greater GRO-seq signal (Fig. 2.3C). These findings suggest that ATF3, Jun and JunD primarily function as transcriptional activators.

Figure 2.3 ***(next page).*** AP-1 monomers bind at unique loci that cannot be explained by differences in the DNA binding domain. **A.** Hierarchical clustering of the relative strength of binding of each monomer at all AP-1 binding sites in Vehicle and 1-hour KLA treatment conditions. **B.** Representative browser shots of ChIP-seq peaks for Veh specific monomers ATF3, Jun and JunD. **C.** GRO-seq at sites where ATF3, Jun and JunD was lost, gained or unchanged after one hour KLA treatment. **D.** Venn diagram of ATF3, Jun and JunD peaks in Vehicle (left) and table indicating the de novo AP-1 motifs found in each subset of peaks and the percent of peaks in each subset that contain one of the two AP-1 motif variants (right). **E.** Binding strength comparison of ATF3 chimeras. The ATF3 DNA binding domain (blue) is replaced the DNA binding domains of Fos (yellow) or Jun (Green) and then transduced into ATF3-deficient iBMDM cells with a lentivirus vector (left). The binding of each chimera is shown as a heatmap of ChIP-seq tags centered on ATF3 chimera binding sites (replicates indicated in separate rows) that were found to be specific for ATF3 (blue) or Jun binding in TGEMs (Green). **F.** Heatmap showing the percent of unique binding sites for each monomer that contain a de novo motif calculated from each set of unique peaks.


A
ChIP-seq at 50664 sites
Relative Binding Strength
1
0
ATF3
Jun
JunD
ATF3
Jun
JunD
Fos
FosL2
JunB
Veh
KLA 1h
B
chr5: 92,350,000
Cxcl10
chr19: 9,985,000
Fth1
chr17: 24,895,000
Spsb3
NormTag Counts
5 kb
ATF3
Jun
JunD
C
GRO-seq (Log2FC KLA/Veh)
ATF3
Jun
JunD
Lost in KLA
Unchanged Veh to KLA
Gained in KLA
D
ATF3
Jun
JunD
5982
1798
2249
10514
974
4837
3299
Motif
% TGASTCA
% TGASSTCA
F
Sfpi1 (Pu.1)
CebpA
CebpD
Runx-AML
Tfec
Mitf
Sp1
% Unique Peaks
ATF3
Jun
JunD
E
ATF3 Fos Jun
Endogenous
DBD domain
ATF3 DBD Chimeras
Transduction
ΔATF3
macrophage cells
anti-ATF3
ChIP-seq
Log2 Reads
5982 Peaks
2249 Peaks
ATF3-DBD
Fos-DBD
Jun-DBD
Specific peaks
ATF3
Specific peaks
Jun

2.3.3 Family member specific binding sites are associated with the same AP-1 motif

While 10514 of the binding sites of ATF3, Jun and JunD in the vehicle condition are shared by all three factors, a greater number of binding sites (11530) are not (Fig. 2.3D). To ensure that the unique sites were not technical artifacts, we ranked the peaks of each family member according to the number of ChIP-seq tags detected and then calculated the percent of peaks that were unique after filtering away binding sites that fell below a given percentile threshold. We found that unique peaks were present even at higher thresholds, supporting our observation that AP-1 family members can bind to distinct loci (Fig. 2.4D).

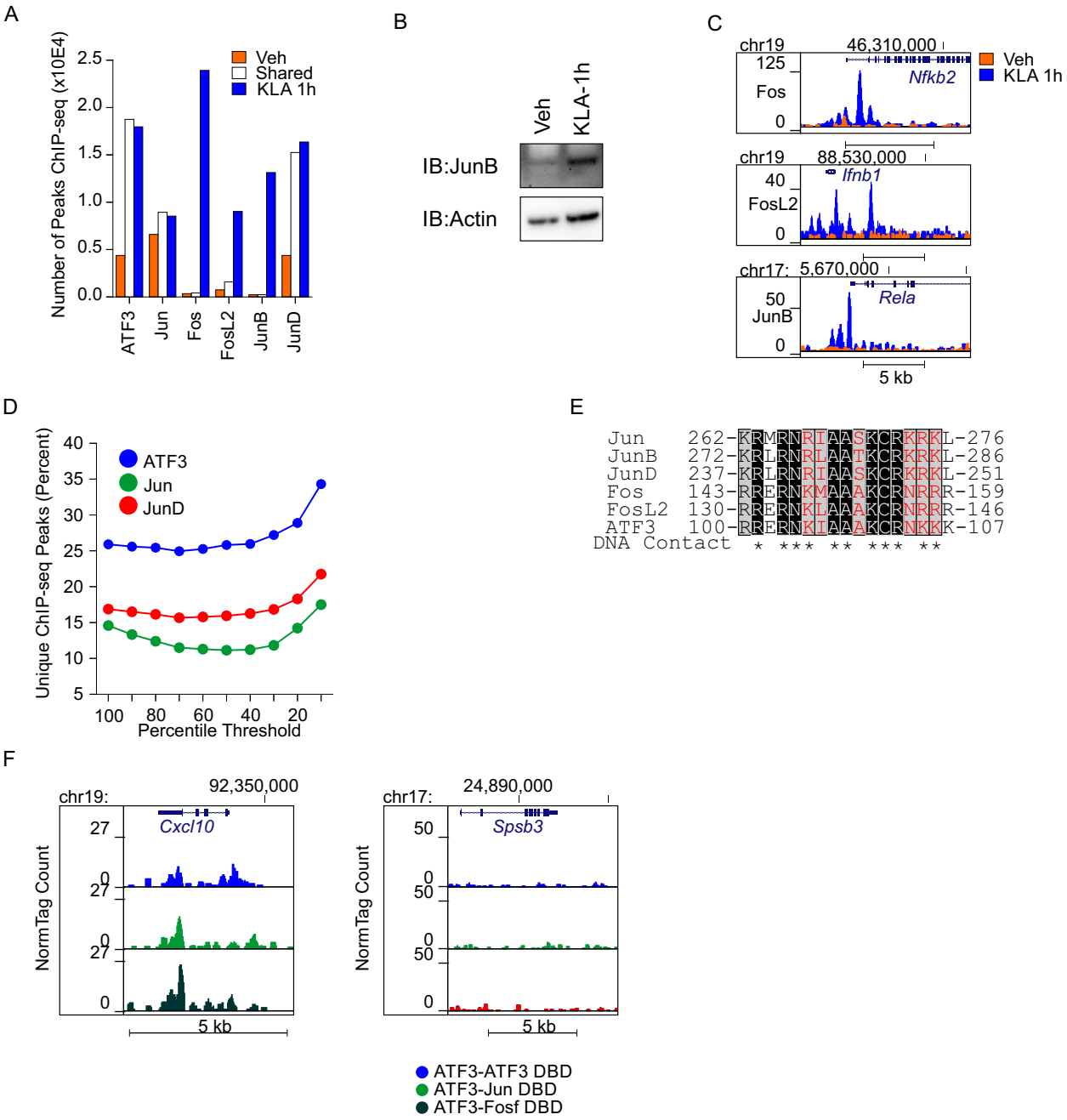
Using de novo motif enrichment analysis, we observed that the binding motif for each combination of monomers was nearly identical (Fig. 2.3D). To investigate whether family members preferred either variant of the AP-1 motif, we calculated the percent of peaks bound by each combination of monomers that had the TRE variant of the AP-1 motif (TGASTCA) and the CRE variant of the motif (TGASSTCA)[25;33]. Consistent with previous studies, we found both variants of the AP-1 motif at regions bound by each combination of monomers, but there was a preference for the TRE motif (Fig. 2.3D)[33]. These results suggest that differences in the AP-1 DBD cannot explain the majority of family member specific binding.

To test the prediction that differences in the AP-1 DBD do not explain binding patterns, we created ATF3 chimeras by replacing the DBD of ATF3 with that of Fos and Jun (Fig. 2.3E, 2.4E). The DBDs of these three factors are highly conserved, with identity at 8 and charge conservation at 3 of 11 amino acids directly involved in DNA interaction (Fig. 2.4E)[26]. We transduced expression vectors for ATF3 chimeras with either an ATF3, Fos or Jun DBD into ATF3 KO iBMDMs and then measured the genome-wide binding patterns of each chimera by performing ChIP-seq using an antibody specific for ATF3 (Fig. 2.3E). Globally, we observed that the chimeras had stronger binding at ATF3 specific sites in comparison to Jun specific sites and that each chimera exhibited similar binding across all loci visualized as normalized tag counts in a heatmap (Fig. 2.3E). Representative browser shots showing similar binding between chimeras are shown at *Cxcl10* and

Spsb1 which are loci specifically bound by ATF3 and Jun respectively (Fig. 2.4F).

Given that the family members all recognized a common DNA binding motif, we hypothesized that differential interactions with locally bound factors mediated by non-conserved protein contact surfaces may explain unique monomer binding sites. We calculated de novo motifs enriched at the unique peaks for ATF3, Jun, and JunD individually, and then calculated the percent of each family members specific binding sites that contained a match to each de novo motif. We identified motifs for key TFs in macrophages[28;65] such as PU.1, CEBP, and Runx (Fig. 2.3F). Composite motifs for AP-1 and IRF or NFAT occurred at similar frequencies at the unique peaks for each family member ($\sim$5% and $\sim$3% of peaks respectively). However, we found no significant differences in the relative enrichment of motifs associated with ATF3, Jun, and JunD specific peaks that would explain their specific binding profiles (Fig. 2.3F).

Figure 2.4 ***(next page)*.** Extended characterization of AP-1 binding cistrome. **A.** Quantification of the number of binding sites for each monomer that are present in vehicle, 1-hour KLA, or shared in both treatment conditions. **B.** Western blot testing the nuclear protein expression of JunB in Vehicle and KLA-1h treated TGEMs. **C.** Representative browser shots of ChIP-seq peaks for KLA specific monomers Fos (top), FosL2 (middle), and JunB (bottom). **D.** Percent of peaks that are unique to each monomer at different thresholds for the number of reads at each peak. 100 indicates 100% of the peaks and 10 indicates the top 10% of peaks when sorting by the number of reads. **E.** Protein alignment of AP-1 monomer DNA binding domains by Clustal Omega. Black background denotes identity while grey background denotes similar charge. Amino acids involved in DNA binding are indicated by stars. **F.** Representative browser shots of ChIP-seq peaks from ATF3 DNA binding domain chimeras at an ATF3 specific site (left) and Jun specific site (right).

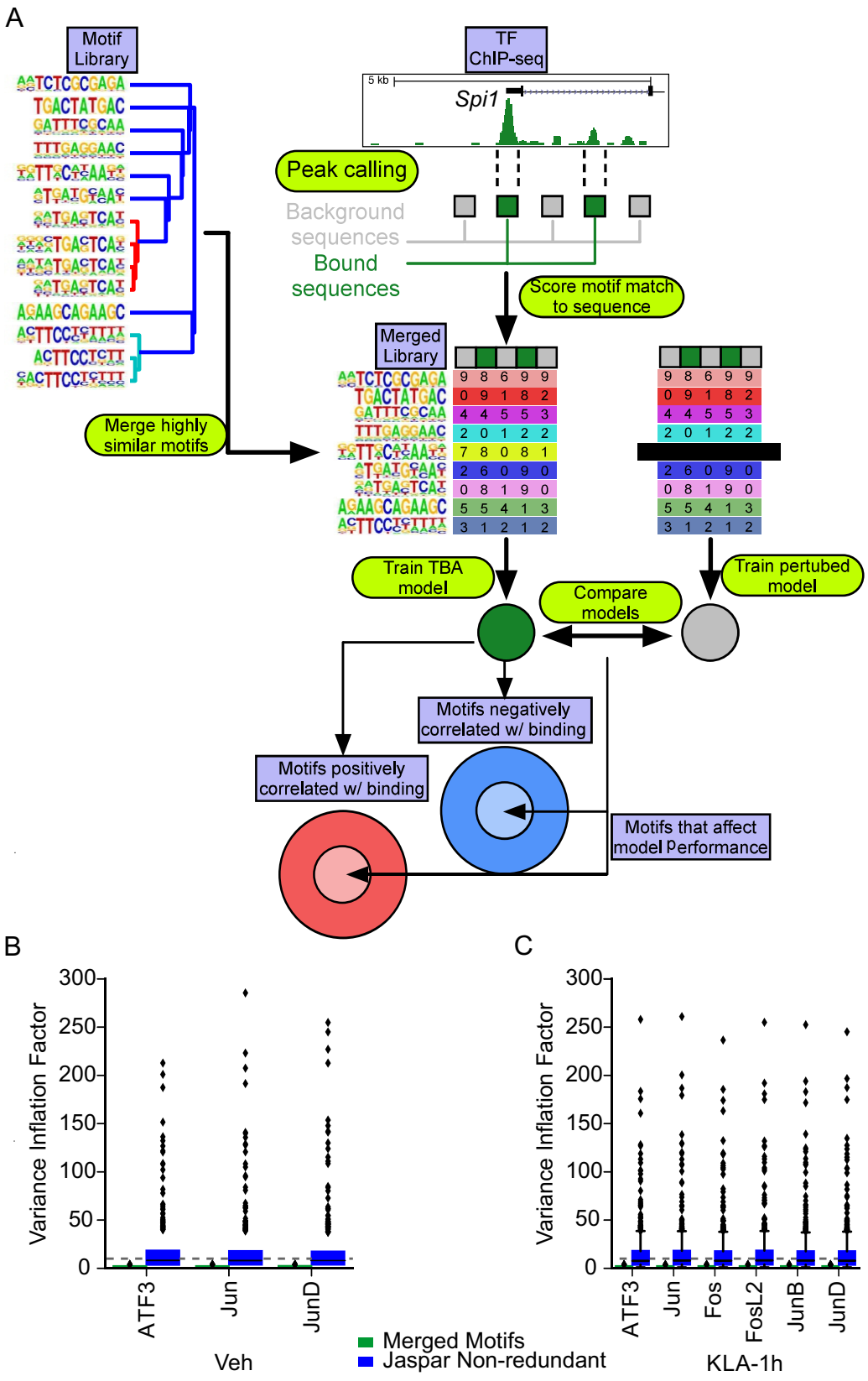

A
Number of Peaks ChIP-seq (x10E4)
2.5
2.0
1.5
1.0
0.5
0.0
Veh
Shared
KLA 1h
ATF3
Jun
Fos
FosL2
JunB
JunD
B
Veh
KLA-1h
IB:JunB
IB:Actin
C
chr19
46,310,000
125
Fos
0
Nfkb2
Veh
KLA 1h
chr19
88,530,000
40
FosL2
0
Ifnb1
chr17:
5,670,000
50
JunB
0
Rela
5 kb
D
Unique ChIP-seq Peaks (Percent)
40
35
30
25
20
15
10
5
ATF3
Jun
JunD
100
80
60
40
20
Percentile Threshold
E
Jun 262-KRMRNRIAASKCRKRKL-276
JunB 272-KRLRNRLAATKCRKRKL-286
JunD 237-KRLRNRIAASKCRKRKL-251
Fos 143-RRERNKMAAAKCRNRRR-159
FosL2 130-RRERNKLAAAKCRNRRR-146
ATF3 100-RRERNKIAAAKCRNKKK-107
DNA Contact * *** ** *** **
F
chr19:
92,350,000
Cxcl10
27
0
27
0
27
0
NormTag Count
5 kb
chr17:
24,890,000
Spsb3
50
0
50
0
50
0
NormTag Count
5 kb
ATF3-ATF3 DBD
ATF3-Jun DBD
ATF3-Fosf DBD

2.3.4 A machine learning model that relates combinations of motifs to transcription factor binding

Given the robustness of the family member specific peaks (Fig. 2.4D), we considered additional biological mechanisms that might be leveraged for detection of motifs differentially associated with each family member. Current methods for calculating enriched motifs analyze each motif individually despite data demonstrating that TFs bind cooperatively in groups[9;31]. Additionally, collaborative binding by TFs allows for partners to bind to more degenerate motifs, which are ignored in de novo motif analysis[28]. We incorporated these concepts into a machine learning model that relates the presence of multiple TF motifs, which may be degenerate, to the binding of a TF. Machine learning models are often considered difficult to interpret due to their complexity. In building our model, we emphasized simplicity and as a consequence, interpretability.

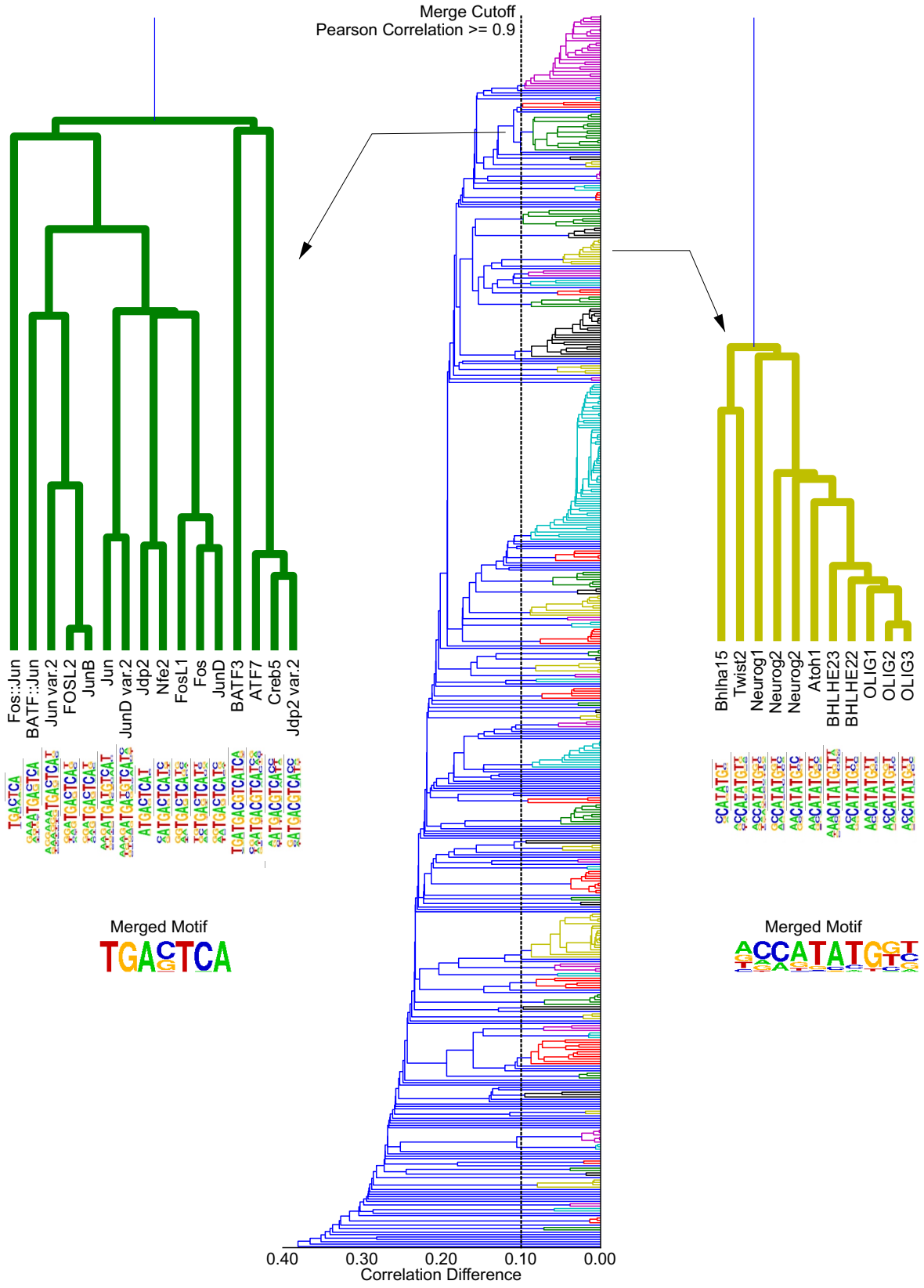
Figure 3A summarizes our model, TBA (Transcription factor Binding Analysis). TBA uses logistic regression to learn to distinguish the binding sites of a TF from a set of GC-matched background loci. For each binding site and background locus, TBA calculates the best match to hundreds of DNA binding motifs, drawn from the JASPAR library, and quantifies the quality of the match as the motif score (aka log likelihood ratio score). To allow for degenerate motifs, all motif matches scoring over zero are considered. The motif scores are then used to train the TBA model to distinguish TF binding sites from background loci. TBA scores the probability of observing binding at a sequence by computing a weighted sum over all the motif scores for that sequence. By considering all motifs simultaneously, TBA can learn to recognize combinations of motifs that are co-enriched at TF binding sites but that are not individually enriched over genomic background. The weight for each motif is learned by iteratively modifying the weights until the models ability to differentiate binding sites from background loci no longer improves. The final motif weight measures whether the presence of a motif is correlated with TF binding. The significance of a given motif can be assigned by comparing the predictive performance of a trained TBA model and a perturbed model that cannot recognize that one motif with the likelihood ratio test.

Figure 2.5 ***(next page)*.** TBA, a Transcription factor Binding Analysis. **A.** Schematic workflow of TBA. Binding sites for a transcription factor (green boxes) are mixed with random GC-matched background sequences (grey boxes). Motifs from the JASPAR library are merged to create a non-redundant motif library. Motif scores are calculated for all sequences at all binding sites and GC-matched background and then used to train a TBA model. Model weights from the trained model indicate whether a motif is positively or negatively correlated with the occupancy of a transcription factor. The performance of the full model and a perturbed model with one motif removed are compared to identify motifs that are important to the model. The intersection of important motifs that affect model performance and the model weights learned by the classifier can be used to infer the binding partners of a transcription factor. **B-C**. Distribution of Variance Inflation Factor for each motif in the TBA merged motif library and JASPAR motif library for experiments performed in **(B)** Vehicle and **(C)** KLA treated TGEMs.


A
Motif Library
TF ChIP-seq
5 kb
Spi1
Peak calling
Background sequences
Bound sequences
Score motif match to sequence
Merged Library
Merge highly similar motifs
Train TBA model
Train pertubed model
Compare models
Motifs negatively correlated w/ binding
Motifs positively correlated w/ binding
Motifs that affect model Performance
B
Variance Inflation Factor
300
250
200
150
100
50
0
ATF3
Jun
JunD
Veh
Merged Motifs
Jaspar Non-redundant
C
Variance Inflation Factor
ATF3
Jun
Fos
FosL2
JunB
JunD
KLA-1h

Machine learning models, including TBA, can be confounded by collinearity, which in our case corresponds to the presence of motifs that are highly similar or redundant[6]. Collinearity can cause inaccurate weight and significance to be assigned to motifs. To assess the extent of collinearity, we calculated the Variance Inflation Factor (VIF)[6] for the scores of each motif in the JASPAR library at AP-1 binding sites. A VIF above 10 would indicate problematic collinearity and that the scores for a motif are highly correlated with the scores of another motif. We found that a substantial number of motifs were collinear with at least one other motif (VIF $>$ 10) (Fig. 3B, 3C). To address the presence of redundant motifs we clustered the JASPAR library, identifying groups of motifs that are highly similar (Fig. 2.6, colored clades), and merged these motifs together (Pearson Correlation $>$ 0.9, Fig. 2.6, Fig. 3), resulting in a condensed library of 196 motifs formed from 519 JASPAR motifs. Multiple collinearity was substantially reduced in our condensed library (VIF $<$ 10, Fig. 3B, 3C).

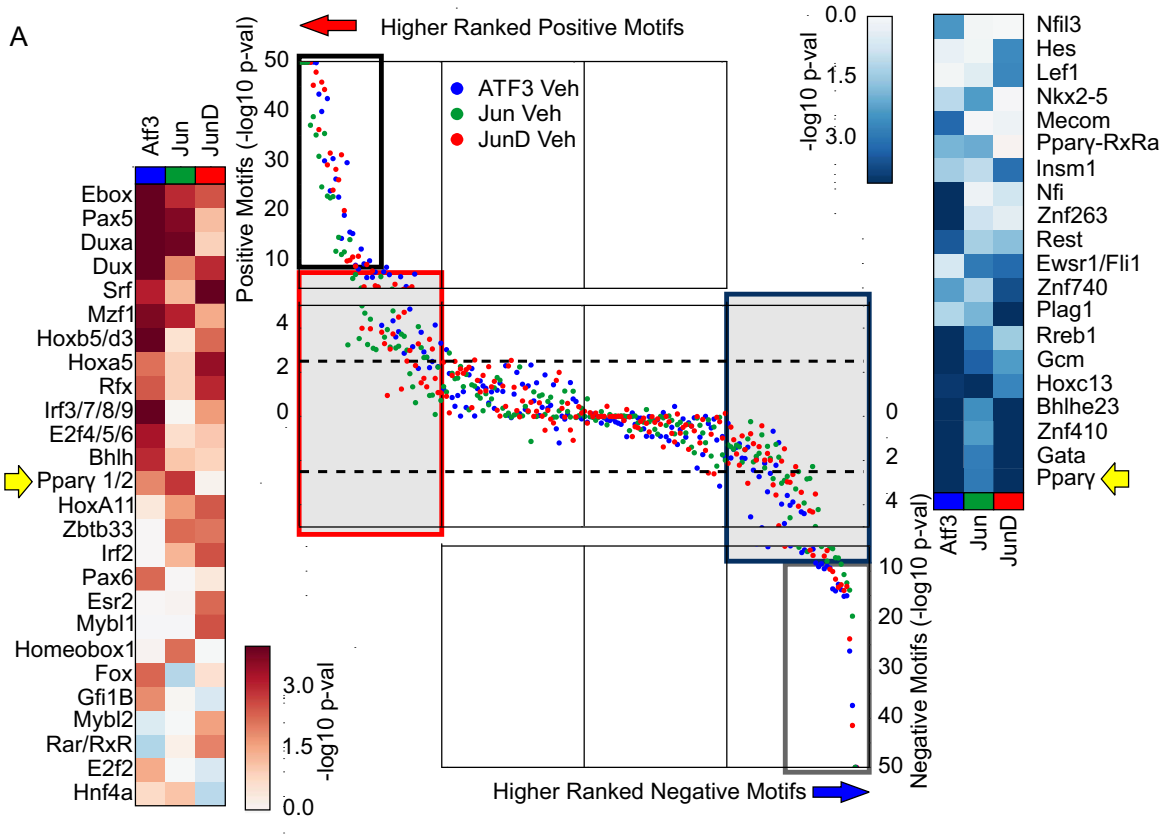
Figure 2.6 ***(next page)*.** Curation of the JASPAR motif library to eliminate highly similar motifs. Hierarchical clustering of all motifs in the JASPAR non-redundant library. Colored clades give highly similar motifs (Pearson Correlation ≥ 0.9, indicated by the dotted line) and blue leaf nodes give motifs that are distinct from all other motifs (Pearson Correlation < 0.9). Representative clades are indicated on the left and right (with the motif logos indicated).


Merge Cutoff
Pearson Correlation >= 0.9
Fos::Jun
BATF::Jun
Jun var.2
FOSL2
JunB
Jun
JunD var.2
Jdp2
Nfe2
FosL1
Fos
JunD
BATF3
ATF7
Creb5
Jdp2 var.2
Merged Motif
Bhlha15
Twist2
Neurog1
Neurog2
Neurog2
Atoh1
BHLHE23
BHLHE22
OLIG1
OLIG2
OLIG3
Merged Motif
0.40
0.30
0.20
0.10
0.00
Correlation Difference

2.3.5 TBA identifies combinations of binding motifs that coordinate AP-1 recruitment

To identify motifs associated with specific AP-1 family members, we trained TBA models for each monomer in resting TGEMs, and probed for differences in the identified motifs. Ranking each motif according to the mean p-value, we found that all family members shared a core set of highly significant motifs both positively and negatively correlated with binding (Fig. 2.7A, i and ii, respectively). The motifs exhibiting strong positive correlation included the AP-1 motif as well as motifs of macrophage collaborative binding partners for AP-1, such as PU.1 and CEBP[28;42;65]. To determine a significance threshold for more moderately ranked motifs, we compared significance values calculated by TBA models trained on replicate ChIP-seq experiments. We determined that motifs with a mean p-value$<$1e-2.5 tended to have similar significance values (absolute likelihood ratio $\sim$1, Fig. 2.8A). The motif weights that exceeded this threshold were highly correlated between replicate experiments (Fig. 2.8B). Outside of the core group of motifs shared by all monomers, we observed $\sim$50 motifs with differential affinities (likelihood ratio $>$ 100 between at least 2 monomers) for each monomer as defined by TBA (Fig. 2.7A, center panel, shaded regions). Differential motifs positively correlated with binding (Fig. 2.7A left heatmap in red) included motifs unique to a monomer such as the PPAR half site with Jun. The full PPARγ motif was negatively correlated with both ATF3 and JunD, suggesting that PPARγ positively influences the binding of Jun to a greater extent than the other AP-1 monomers (Fig. 2.7A right heatmap in blue). These results suggest that AP-1 monomers have distinct sets of collaborating TFs that affect their binding patterns.

Figure 2.7 ***(next page)*****.** TBA identifies motifs predicted to specify differential AP-1 monomer-binding in resting TGEMs. **A.** DNA motifs rank order based on the significance of the motif according to the likelihood ratio test. The black box represents the most significant motifs positively correlated with binding for all AP-1 monomers and are listed in (i) and the most significant motifs negatively correlated with binding for all AP-1 monomers are shown in the grey box and are listed in (ii). The significance of motifs positively correlated with binding that show a 100-fold likelihood difference between two monomers are shown on the left heatmap (red); the right heatmap (blue) gives the significance of corresponding motifs negatively correlated with binding. **B.** Comparison of the performance of TBA against the AP-1 motif score alone, Bayesian Markov Model (BaMM) motif score, and gapped k-mer SVM as measured by the area under the Receiver Operating Characteristic curve (aucROC). Error bars indicate the standard deviation of aucROC across 5 cross validation sets. **C.** Number of motifs that pass an in-silico mutagenesis test for significance (the likelihood ratio test comparing the performance of a full model that uses all the motifs and a mutated model with one motif removed) at various p-value thresholds. **D.** Predictive performance of TBA when predicting ATF3, Jun and JunD binding as motifs are iteratively removed starting from the least important motif based on the weights calculated by TBA. Inset shows performance values beginning at 150 motifs removed where predictive performance begins to drop.

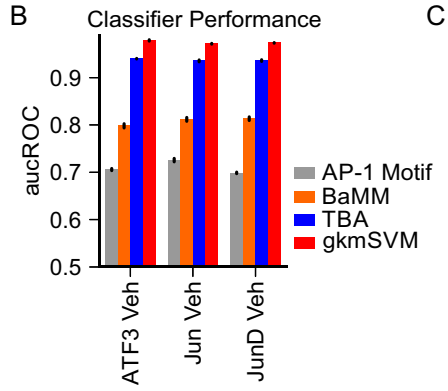


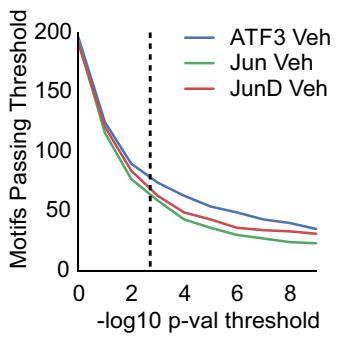
i

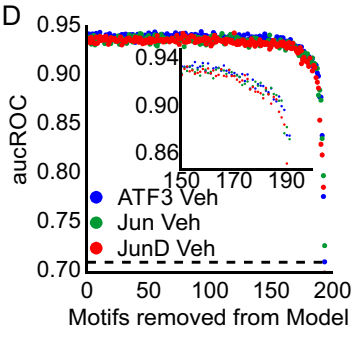
Motif	Mean p-val		
AP-1 CRE	1.00E-50		
AP-1 TRE	1.00E-50	Mef2a/b/c	9.38E-24
CEBP	1.00E-50	Irf1	1.70E-23
PU.1	1.00E-50	Elf	1.73E-15
Pax2	4.62E-40	Sp1	1.19E-12
Elk-Etv	2.55E-38	Tcfl5	1.61E-12
SpiB	2.43E-36	Ctcf	5.44E-12
Runx	5.49E-32	Nr2e1	7.79E-11
Sp4	4.64E-31	Usf	6.73E-08
Egr	9.34E-26	Znf143	4.22E-07
Maf-Nrl	6.88E-24	Tfcp2	6.24E-06

ii

Motif	Mean p-val
Zeb	1.00E-50
Yy1	2.15E-20
Figla-id4-snai1-tcf3/4	4.51E-15
Msc-Myf6-Tfap4	1.55E-13
Zbtb7	6.58E-13
Onecut	8.63E-10
Hic	1.57E-08
Arid3a	1.79E-08
Nr5a2	5.22E-07
Tbp	1.22E-06
Tead	2.00E-06







2.3.6 Evaluation of collaborating TF motifs that coordinate AP-1 binding

To assess whether the additional motifs identified by TBA are useful for identifying AP-1 sites, we compared TBAs ability to predict the binding of each monomer to several other sequence based approaches. Predicting TF binding using just the AP-1 TRE motif score had the worst performance as measured by the area under the receiver operating characteristic curve (aucROC (Fig. 2.7B). Bayesian Markov Model motifs (BaMM)[86], which assesses dependencies between the positions within the binding motif, improved upon the simple AP-1 motif score by $\sim$15% (Fig. 2.7B). The TBA model and the gkm-SVM model achieved even higher performance, demonstrating that additional sequences outside of a TFs motif may contribute to binding site selection (Fig. 2.7B). The performance of gkm-SVM exceeded that of TBA (by $\sim$3%). However, a greater number of motifs related to the binding of a TF can be extracted from TBA in comparison to gkmSVM. The authors of gkm-SVM described a procedure to retrieve up to three PWMs from k-mers ranked by gkmSVM[22], while TBA identified over 50 motifs that passed a significance threshold of p$<$1e-2.5 (Fig. 2.7C). To examine the impact of statistically significant (p$<$1e-2.5) but moderately ranked motifs, we calculated TBAs performance while iteratively removing motifs from the model (starting with the least significant motif) (Fig. 2.7D). The performance of the model started declining when the motifs from the top 50 were removed, demonstrating that the local sequence environment outside of the AP-1 motif affects AP-1 binding (Fig. 2.7D, inset).

2.3.7 Cell type specific binding preferences of JunD

To further test the hypothesis that distinct sets of collaborating TFs can affect AP-1 binding, we examined JunD binding in a panel of cell lines. Each cell type expresses a distinct repertoire of TFs that are available as binding partners for JunD. We trained TBA models for ChIP-seq of JunD in each cell line and then extracted the 20 most significant motifs from each model. Motifs which are bound by TFs known to be important for particular cell lines were found to be correlated with JunD binding. For example, the Gata motif was positively correlated with JunD binding in K562 cells, an erythroid lineage erythroleukemia, while Pou motifs (e.g. OCT4) were important in h1-hESCs

(Fig. 2.8C)[34]. Differences in the motifs identified by TBA for each cell line corresponded to large differences in the loci bound by JunD (Fig. 2.8D), suggesting that JunD interacts with different TFs depending on the expressed binding partners available in each cell type[100].

2.3.8 KLA treatment changes the collaborating TFs available to AP-1 and remodels the AP-1 cistrome

Given that AP-1 binds collaboratively with other TFs, the selection of binding sites for each monomer will depend on the available of collaborating partners. To study effects of changes in collaborating TF availability, we examined AP-1 binding before and after KLA treatment. Treatment of TGEMs with KLA resulted in 178 mRNAs increasing 2-fold (FDR$<$0.05) or greater (Fig. 2.10A). A total of 29 genes encoding TFs with known binding motifs (20 upregulated and 9 downregulated) had a significant change in expression (FDR $<$ 0.05) including AP-1 monomers Fos, Fra2 and JunB (Fig. 2.10A, blue points). In addition, TLR4 activation by KLA results in the activation of several latent transcription factors, including NFκB and interferon regulatory factors (IRFs). Correspondingly, AP-1 monomers showed changes in their global binding patterns with Fos and JunB displaying drastic upregulation in binding sites (Fig. 2.4A, Fig. 2.9A).

To examine motifs associated with AP-1 binding after KLA treatment, we trained TBA models for each monomer in KLA treated TGEMs. Again, we observed that all AP-1 monomers shared a common group of highly significant motifs positively correlated with binding, including AP-1, CEBP, PU.1, REL, and Egr, and negatively correlated with binding, such as the Zeb1 motif (Fig. 2.10B, Table 2.1, Table 2.2). Many of the moderately ranked motifs showed large differences in significance between the monomers (Fig. 2.10B, C: likelihood ratio $>$ 100).

We found that AP-1 monomers with substantive binding before KLA treatment (ATF3, Jun, and JunD) showed changes in their preference (as measured by the likelihood ratio for each motif when comparing the KLA and Vehicle TBA models) for motifs bound by upregulated TFs such as Rel, Irf3/7/8/9, Irf2 and Nfat (Fig. 2.9B, likelihood ratio $>$ 10e4). Conversely, down regulated TFs were found to have reduced significance for all AP-1 monomers after 1-hour KLA treatment including Usf (Fig. 2.9B, likelihood ratio $<$ 1e-4). AP-1 monomers activated after 1-hour KLA

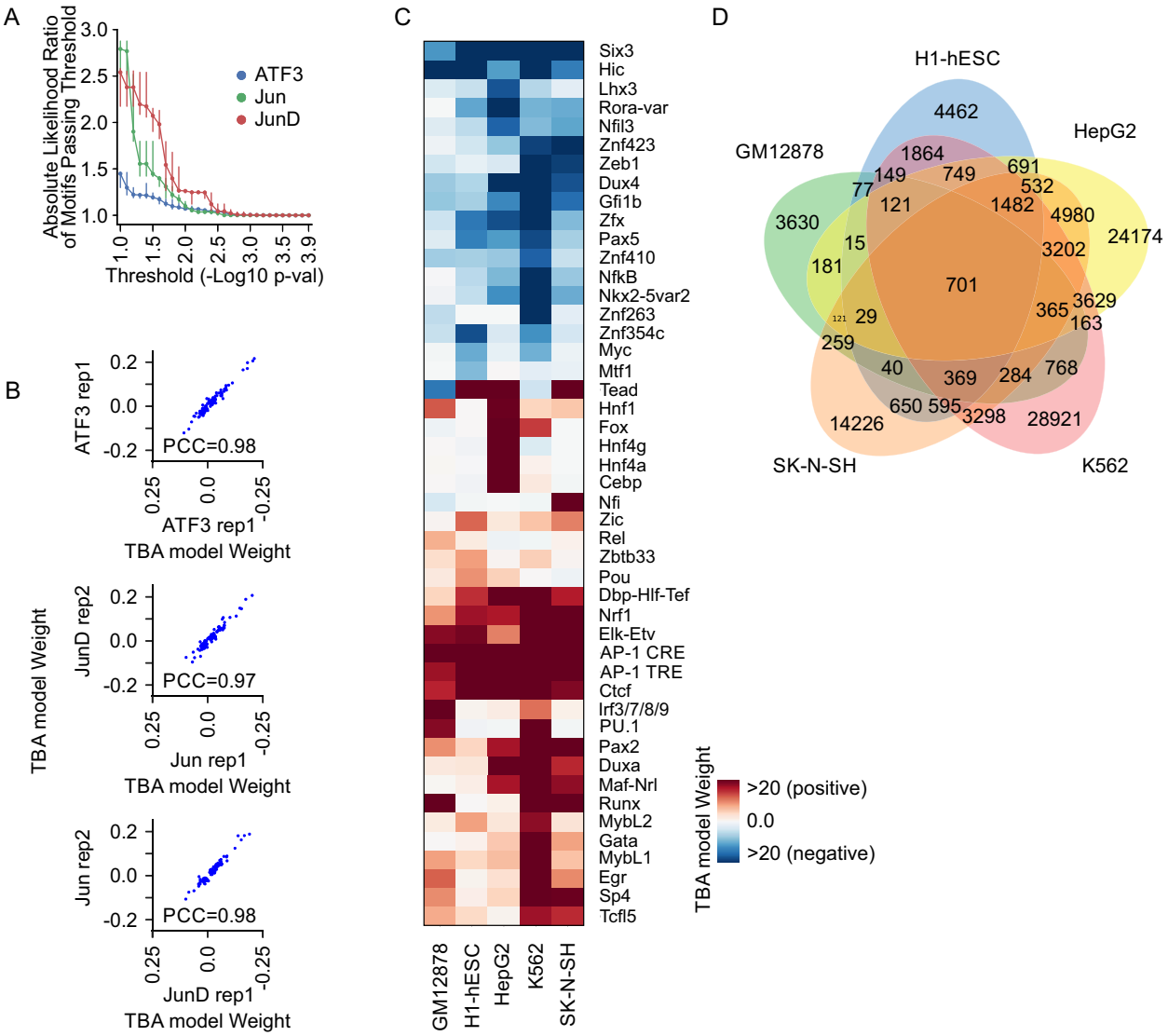


Figure 2.8. Characterization of TBA on individual replicate experiments and JunD ChIP-data from different cell lines. **A.** Distribution of the absolute ratio of the likelihood value for each motif that has mean likelihood below the threshold indicated on the horizontal axis, when comparing models trained on individual replicate experiments. Likelihood values for each individual replicate are calculated using the likelihood ratio test, and then averaged across 5 cross validation sets. Error bars indicate the standard deviation. **B.** Comparison of the weights of significant motifs ($p<1e\text{-}2.5$) from TBA models calculated from individual experiments. The similarity of each pair of models, measured by the Pearson correlation coefficient, is annotated in each panel. **C.** Significance values for the 20 most significant motifs for TBA models trained for JunD binding in a several of cell lines. Red hues indicate motifs positively correlated with binding and blue hues indicate motifs negatively correlated with binding. **D.** Overlap of JunD binding sites from various cell lines. Labels indicate the number of binding sites that overlap between a combination of the cell lines.

treatment (Fos, FosL2 and JunB) (Fig. 2.3A, 2.9A) also showed an affinity for the Rel, Nfat, Irf3/7/8/9 and NFκB motifs (Fig. 2.9B).

To assess the extent to which individual TF motifs could explain the change in binding after KLA treatment, we calculated the correlation of each motifs score to the change in binding after KLA treatment at all loci (Fig. 2.9C). We found that motifs with large changes in significance when comparing the Vehicle and KLA TBA models for each monomer showed higher correlations to the change in binding after KLA treatment and that these motifs corresponded to well established TLR4 activated TFs such as Rel, NFAT, and NFκB (Fig. 2.9B, C)[9;42]. To demonstrate that combinations of TFs can better explain the change in AP-1 binding after KLA treatment, we used TBA to predict the change in binding after KLA treatment. We calculated a predicted change in binding by taking the difference of the predicted binding strength given by the Vehicle and KLA model for each monomer (Fig. 2.9D-F). We found that TBA could predict the change in binding after KLA treatment better than any individual motif (Fig. 2.9C).

Figure 2.9 ***(next page).*** AP-1 binding is context-dependent and affected by the availability of binding partners. **A.** Heatmap of the change in binding of AP-1 monomers after 1-hour KLA treatment quantified as the Log2 ratio of KLA binding to Vehicle binding. **B.** Heatmap showing the TBA assigned significance of DNA motifs that had a 10e4 absolute likelihood ratio between the KLA and Vehicle value for each monomer. **C.** Pearson correlation of individual motif scores and TBA predictions with the change in binding after one hour KLA treatment. **D. E. F.** TBA predicted change in ATF3, Jun, and JunD binding after KLA-1h treatment versus actual change in binding. PCC indicates the Pearson Correlation coefficient of TBA predictions to the log2 fold change in binding of each monomer after one hour KLA treatment.

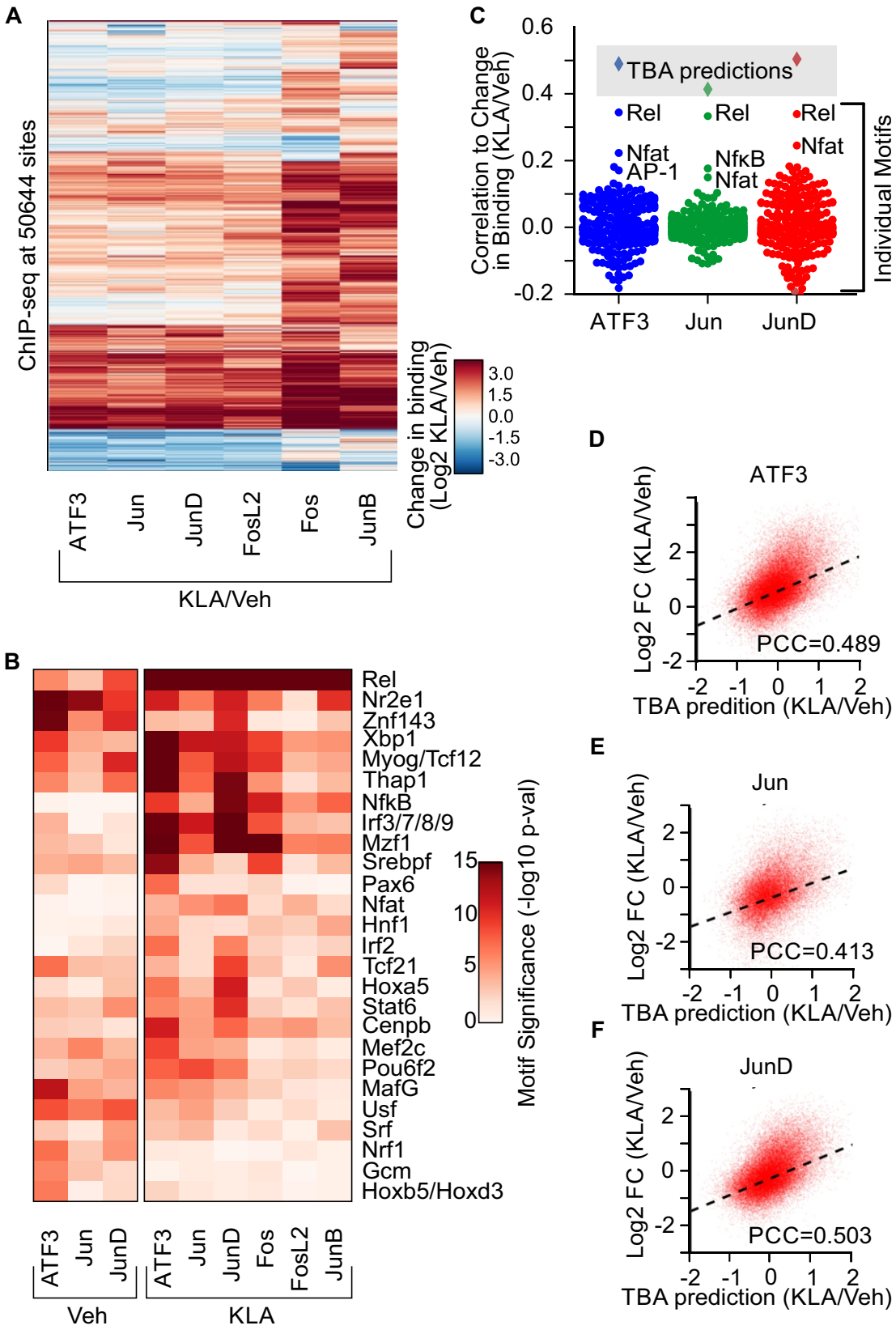

A
ChIP-seq at 50644 sites
ATF3
Jun
JunD
FosL2
Fos
JunB
KLA/Veh
Change in binding (Log2 KLA/Veh)
3.0
1.5
0.0
-1.5
-3.0
B
Rel
Nr2e1
Znf143
Xbp1
Myog/Tcf12
Thap1
NfkB
Irf3/7/8/9
Mzf1
Srebpf
Pax6
Nfat
Hnf1
Irf2
Tcf21
Hoxa5
Stat6
Cenpb
Mef2c
Pou6f2
MafG
Usf
Srf
Nrf1
Gcm
Hoxb5/Hoxd3
15
10
5
0
Motif Significance (-log10 p-val)
ATF3
Jun
JunD
Veh
ATF3
Jun
JunD
Fos
FosL2
JunB
KLA
C
Correlation to Change in Binding (KLA/Veh)
0.6
0.4
0.2
0.0
-0.2
TBA predictions
Rel
Rel
Rel
Nfat
AP-1
NfκB
Nfat
Nfat
Individual Motifs
ATF3
Jun
JunD
D
ATF3
Log2 FC (KLA/Veh)
2
0
-2
PCC=0.489
-2 -1 0 1 2
TBA predition (KLA/Veh)
E
Jun
Log2 FC (KLA/Veh)
2
0
-2
PCC=0.413
-2 -1 0 1 2
TBA prediction (KLA/Veh)
F
JunD
Log2 FC (KLA/Veh)
2
0
-2
PCC=0.503
-2 -1 0 1 2
TBA prediction (KLA/Veh)

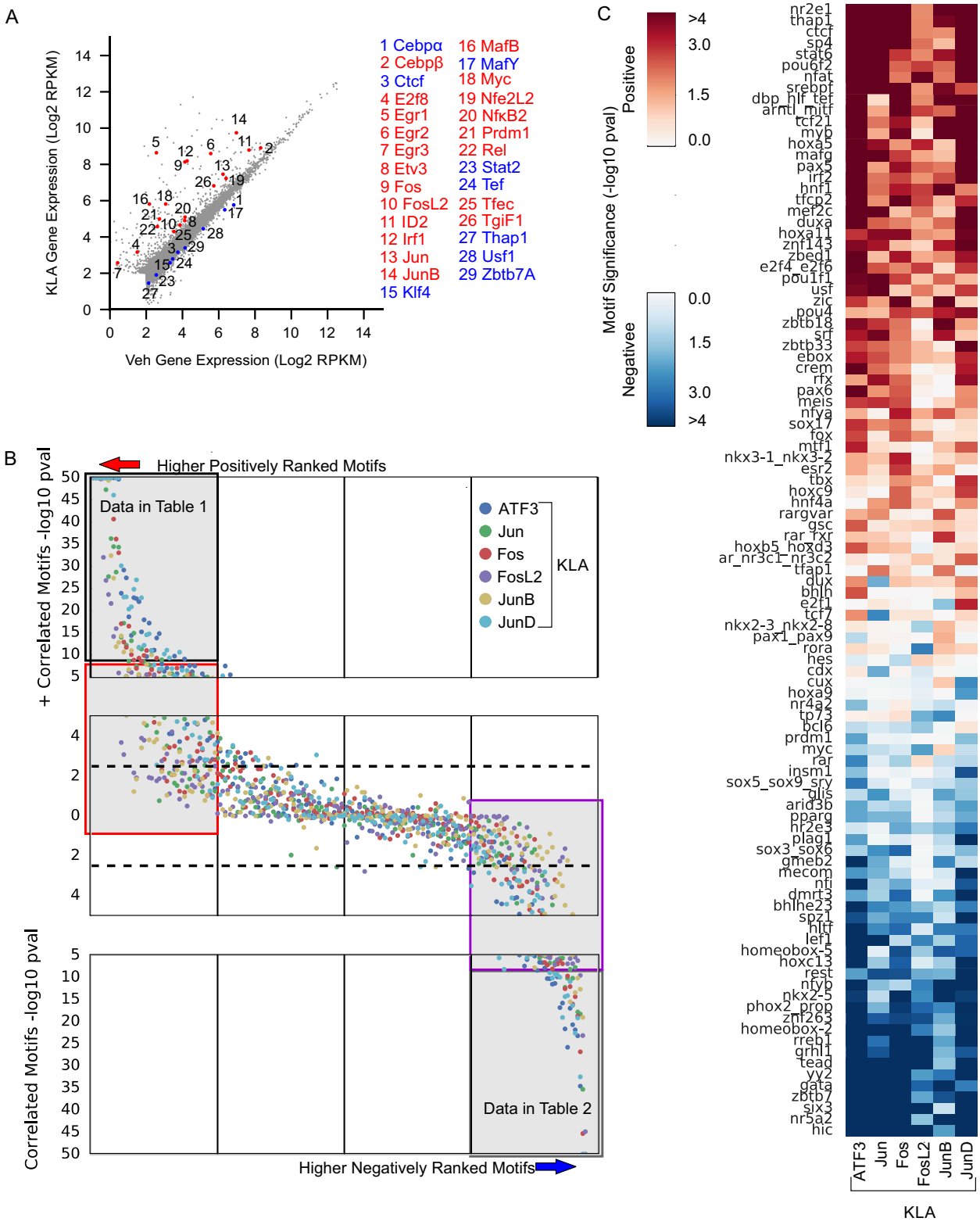
Table 2.1. Table of highly significant motifs, positively correlated with binding for all AP-1 monomers in KLA treated TGEMs.

Motif	**Mean p-val**
ap-1	1.00E-50
cebp	1.00E-50
spi1-c	1.00E-50
rel	1.00E-50
egr	1.15E-50
atf7_batf3_creb5	1.64E-33
pax2	2.76E-28
runx	4.26E-25
elk_etv	4.48E-23
maf_nrl	1.34E-15
irf1	2.70E-13
mzf1	1.14E-07
tcfl5	1.82E-07
spib	2.32E-07
xbp1	1.39E-06
elf	2.77E-06

Table 2.2. Table of highly significant motifs, negatively correlated with binding for all AP-1 monomers in KLA treated TGEMs.

Motif	**Mean p-val**
zeb1	2.20E-46
yy1	1.53E-20
figla_id4_snai2_tcf3_tcf4	4.78E-11
onecut	1.04E-07
tbp	3.04E-07
msc_myf6_tfap4	6.38E-07

Figure 2.10 ***(next page)*.** TBA identifies motifs that coordinate the binding of each AP-1 monomer in KLA-1h treatment. **A.** mRNA expression of transcripts before and after KLA-1h treatment. Differentially expressed (FDR$<$0.05) factors with known DNA motifs are highlighted in red (up-regulated) and blue (down-regulated) , labeled and listed to the right. **B.** DNA motifs rank order based on the significance of the motif according to the likelihood ratio test. **C.** Heatmap representing the negative log10 p-value of each motif that shows a 100 fold likelihood ratio between two monomers when using the likelihood ratio test.


A
KLA Gene Expression (Log2 RPKM)
Veh Gene Expression (Log2 RPKM)
1 Cebpα
2 Cebpβ
3 Ctcf
4 E2f8
5 Egr1
6 Egr2
7 Egr3
8 Etv3
9 Fos
10 FosL2
11 ID2
12 Irf1
13 Jun
14 JunB
15 Klf4
16 MafB
17 MafY
18 Myc
19 Nfe2L2
20 NfκB2
21 Prdm1
22 Rel
23 Stat2
24 Tef
25 Tfec
26 TgiF1
27 Thap1
28 Usf1
29 Zbtb7A
B
Higher Positively Ranked Motifs
+ Correlated Motifs -log10 pval
Data in Table 1
ATF3
Jun
Fos
FosL2
JunB
JunD
KLA
Correlated Motifs -log10 pval
Data in Table 2
Higher Negatively Ranked Motifs
C
Positivee
Negativee
Motif Significance (-log10 pval)
>4
3.0
1.5
0.0
nr2e1
thap1
ctcf
sp4
stat6
pou6f2
nfat
srebpf
dbp_hlf_tef
arntl_mitf
tcf21
myb
hoxa5
mafg
pax5
irf2
hnf1
tfcp2
mef2c
duxa
hoxa11
znf143
zbed1
e2f4_e2f6
pou1f1
usf
zic
pou4
zbtb18
srf
zbtb33
ebox
crem
rfx
pax6
meis
nfya
sox17
fox
mtf1
nkx3-1_nkx3-2
esr2
tbx
hoxc9
hnf4a
rargvar
gsc
rar_rxr
hoxb5_hoxd3
ar_nr3c1_nr3c2
tfap1
dux
bhlh
e2f1
tcf7
nkx2-3_nkx2-8
pax1_pax9
rora
hes
cdx
cux
hoxa9
nr4a2
tp73
bcl6
prdm1
myc
rar
insm1
sox5_sox9_sry
glis
arid3b
pparg
hr2e3
plag1
sox3_sox6
gmeb2
mecom
nfi
dmrt3
bhlhe23
spz1
hltf
lef1
homeobox-5
hoxc13
rest
nfyb
nkx2-5
phox2_prop
znf263
homeobox-2
rreb1
grhl1
tead
yy2
gata
zbtb7
six3
nr5a2
hic
ATF3
Jun
Fos
FosL2
JunB
JunD
KLA

2.3.9 Leveraging natural genetic variation between mouse strains to validate TBA results

To validate the results of our machine learning model genome wide, we used natural genetic variation found between C57BL6/J and BALBc/J mice, which differ genetically by $\sim$5 million single nucleotide polymorphisms (SNPs) and insertions/deletions (InDels)[44]. We have previously shown that mutations which occur within DNA binding motifs can be used to predict genetic interactions between TFs[28;65]. We performed ChIP-seq targeting expressed AP-1 monomers, ATF3, Fos, FosL2, Jun, JunB and JunD in TGEMs isolated from BALB/cJ mice. Mutations can be found in $\sim$17% of each monomers binding sites, and one third of those loci show strain specific binding (fold change $>$2), as shown for ATF3 (Fig. 2.11A). These binding differences cannot be attributed to differences in mRNA or protein expression levels, which are highly similar (Fig. 2.12A, B). We observed that TBA models trained on either strain could be used to predict binding in the other with no loss of predictive ability (Fig. 2.12C), suggesting that each monomer, which has identical protein sequence in both strains, interacts with the same repertoire of collaborating TFs in both strains.

To assess the extent to which SNPs/InDels in individual motifs explain strain-specific binding, we calculated the difference between the best matching motif score at every loci between the strains and then calculated the Pearson Correlation to the change in binding (Fig. 2.11B, 2.12D). Mutations in individual motifs showed a weak correlation to strain specific binding (Fig. 2.11B, 2.12D). We found that motifs identified with TBA ($p<$1e-2.5) are enriched at strain specific peaks in comparison to non-strain specific peaks, but that mutations in any individual motif do not occur frequently enough to explain the majority of strain specific binding (Fig. 2.11C, 2.12E). We integrated the contributions of multiple motifs to strain specific binding, by weighting the motif score difference with the TBA calculated weight, and were able to predict strain specific binding with a 2-fold improvement in performance in comparison to using the AP-1 motif score (Fig. 2.11B, 2.12D)

Next, we created a variant of our model, which we call TBA-2Strain, that directly learns

from genetic variation (Fig. 2.11D). TBA-2Strain takes genetic variation as input (quantified as the change in motif scores between the two strains) and the extent of strain specific binding for each AP-1 monomer. Using TBA-2strain, we predicted strain specific binding at all binding sites with a mutation (Fig. 2.11B). In comparison to TBA, TBA-2Strain has better predictive performance (Fig. 2.11B). This may be attributed to TBA-2Strain being able to observe sites that contain mutations but do not exhibit strain specific binding. The ability of TBA-2Strain to predict strain specific binding improves upon deltaSVM, a state of the art tool for predicting the effect of genetic variation[22] (2.11B, 2.12D).

We then extracted significant motifs from TBA-2Strain using the F-test ($p < 0.05$) and intersected these motifs with motifs identified by TBA model (Fig. 2.11E, 2.7A). We found that the motifs from both models overlapped substantially (Fig. 2.11E, $p < 0.05$, Fisher's exact test), reinforcing the notion that dozens of motifs contribute to coordinating the targeting of AP-1 monomers. Significance values for motifs identified by both models are shown from resting and activated TGEMs (Fig. 2.11E, 2.12F). Notably, the PPARγ half-site was detected by both the TBA and TBA-2Strain models.

Figure 2.11 ***(next page)*.** Leveraging the effects of genetic variation to validate TBA predictions in resting macrophages. **A.** Comparison of the mean strength of binding (number of quantile normalized ChIP-seq tags) for ATF3 in resting TGEMs isolated from C57Bl/6J and Balb/cJ versus the extent of strain specific binding. Loci with a mutation are indicated in blue (fold change $\geq$2) when there is strain specific binding and grey otherwise. **B.** Comparison of different models for predicting strain specific binding of each monomer as measured by the Pearson correlation of a models predictions versus the extent of strain specific binding in resting TGEMs. Models that integrate multiple motifs deltaSVM, TBA, TBA-2Strain, are represented as diamonds. Individual motifs are indicated using round points. **C.** Frequency of mutations in significant motifs (from TBA model, p$<$1e-2.5) at strain specific (fold change $\geq$2) versus non-strain specific peaks resting TGEMs. **D.** Schematic of TBA-2Strain model. Binding sites for a transcription factor with at least one SNP or indel (red boxes) and binding sites with no mutation (grey) are identified. Next, genetic variation is quantified as the difference in the motif scores between the sequences from the two strains and then used as input to train the TBA-2Strain model to predict the extent of strain specific binding. Model weights from the trained model indicate whether a mutation in a motif is correlated with strain specific binding. **E.** Heatmap of significance values for motifs that intersected between the TBA and TBA-2Strain model for each monomer in resting TGEMs. Blue indicates motifs negatively correlated with binding and red indicates positively correlated motifs.

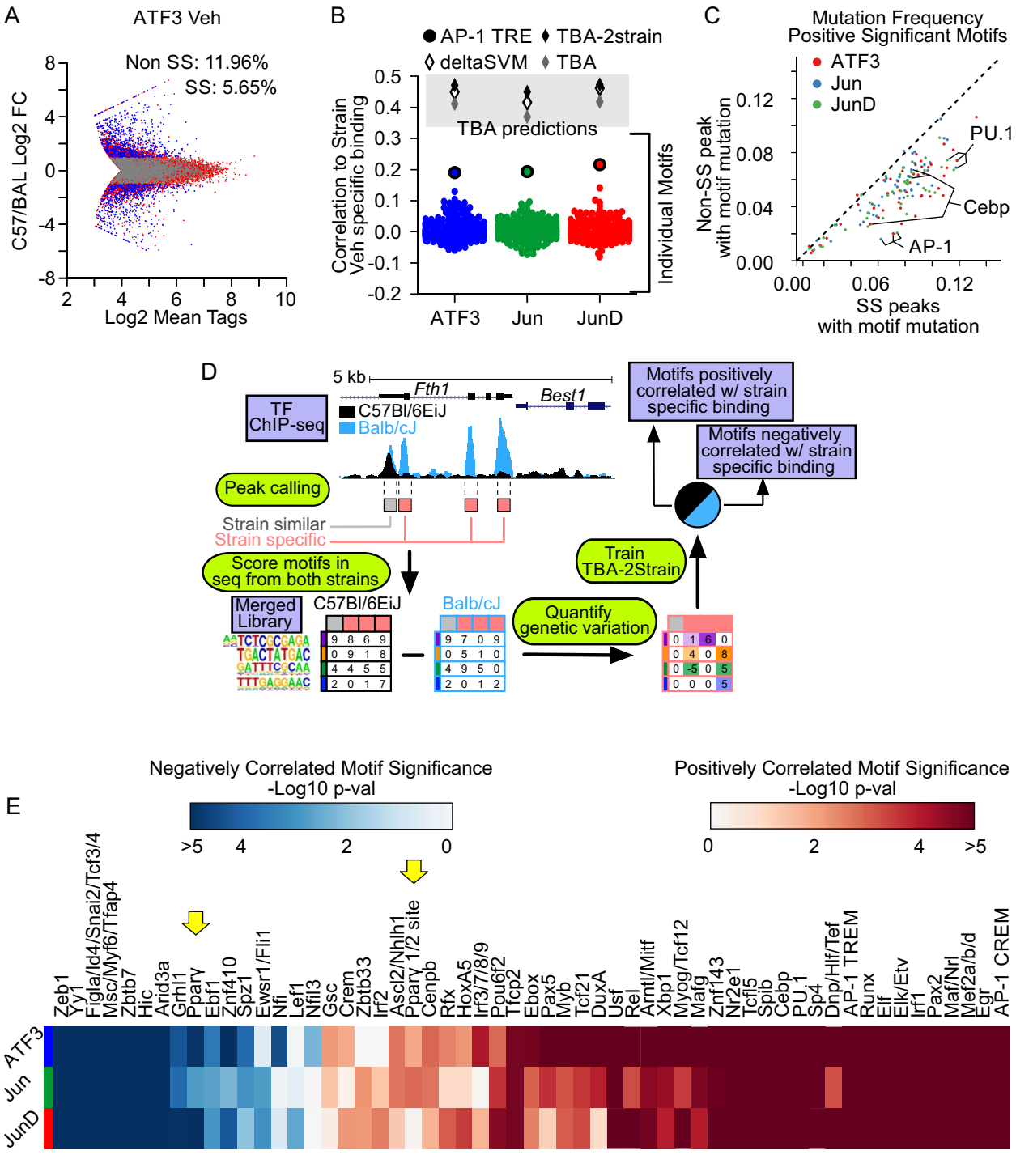
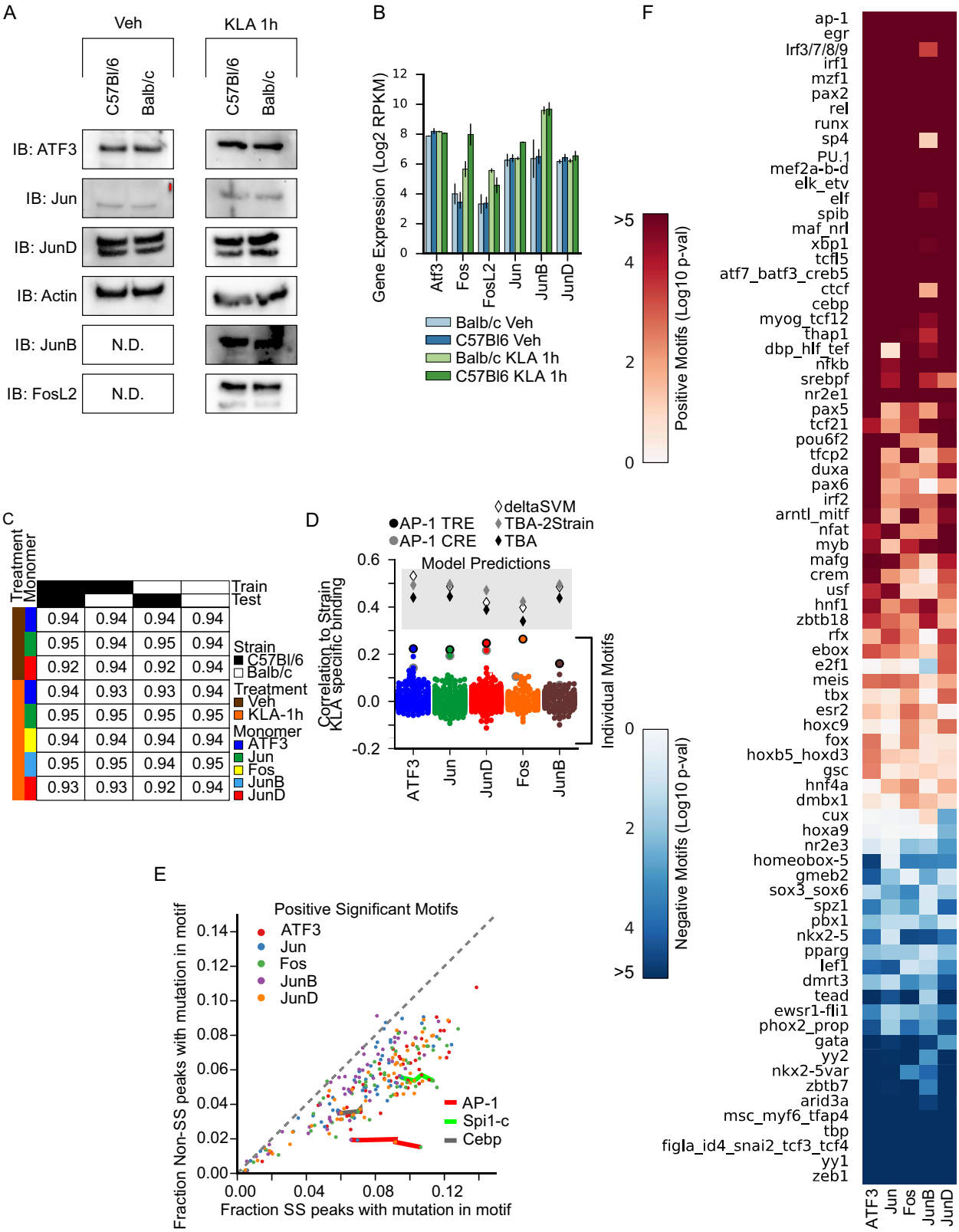

A
ATF3 Veh
Non SS: 11.96%
SS: 5.65%
C57/BAL Log2 FC
Log2 Mean Tags
B
AP-1 TRE
TBA-2strain
deltaSVM
TBA
TBA predictions
Correlation to Strain Veh specific binding
Individual Motifs
ATF3
Jun
JunD
C
Mutation Frequency
Positive Significant Motifs
ATF3
Jun
JunD
PU.1
Cebp
AP-1
Non-SS peak with motif mutation
SS peaks with motif mutation
D
5 kb
Fth1
Best1
C57Bl/6EiJ
Balb/cJ
TF ChIP-seq
Peak calling
Strain similar
Strain specific
Score motifs in seq from both strains
Merged Library
C57Bl/6EiJ
Balb/cJ
Quantify genetic variation
Train TBA-2Strain
Motifs positively correlated w/ strain specific binding
Motifs negatively correlated w/ strain specific binding
E
Negatively Correlated Motif Significance
-Log10 p-val
Positively Correlated Motif Significance
-Log10 p-val
Zeb1
Yy1
Figla/Id4/Snai2/Tcf3/4
Msc/Myf6/Tfap4
Zbtb7
Hic
Arid3a
Grhl1
Ppary
Ebf1
Znf410
Spz1
Ewsr1/Fli1
Nfi
Lef1
Nfil3
Gsc
Crem
Zbtb33
Irf2
Ascl2/Nhlh1
Ppary_1/2 site
Cenpb
Rfx
HoxA5
Irf3/7/8/9
Pou6f2
Tfcp2
Ebox
Pax5
Myb
Tcf21
DuxA
Usf
Rel
Arntl/Mitf
Xbp1
Myog/Tcf12
Mafg
Znf143
Nr2e1
Tcfl5
Spib
Cebp
PU.1
Sp4
Dnp/Hlf/Tef
AP-1 TREM
Runx
Elf
Elk/Etv
Irf1
Pax2
Maf/Nrl
Mef2a/b/d
Egr
AP-1 CREM
ATF3
Jun
JunD

Figure 2.12 ***(next page)*.** Leveraging the effects of genetic variation to validate TBA predictions in activated macrophages. **A.** Western blot showing protein expression of AP-1 family members in TGEMs after Veh and one hour KLA treatment using nuclear extracts. **B.** mRNA expression of monomers before and after 1-hour KLA treatment in C57Bl/6J and BALB/cJ. **C.** Model performance when varying the strain of the data used for training and testing the TBA model. TBA was trained on AP-1 monomers in either Veh or KLA and on one strain and tested for predictive ability on either strain. **D.** Predictive performance of various models for predicting strain specific binding of each monomer as measured by the Pearson correlation of a models predictions versus the log2 fold change in binding between the Balb/cJ and C57Bl/6J at all AP-1 binding in activated, one hour KLA treated, TGEMs. The performance of models that integrate multiple motifs deltaSVM, TBA, TBA-2Strain, are represented as diamonds. The correlation of the change in an individual motifs score (due to a mutation) to strain specific binding is indicated using round points. **E.** Frequency of mutations in significant motifs (from TBA model, $p<1e\text{-}2.5$) at strain specific versus non-strain specific peaks for each monomer in KLA treated macrophages. **F.** Heatmap of significance values for motifs that intersected between the One Strain and Two Strain model.

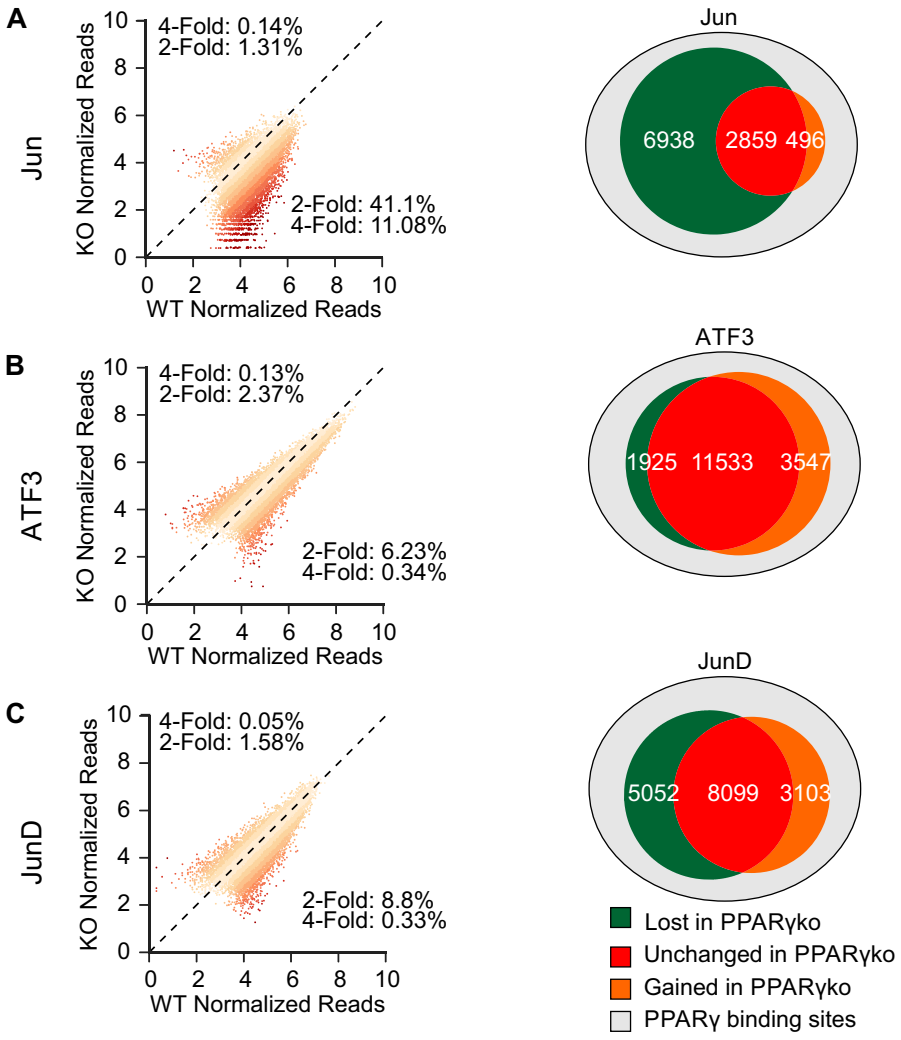

A
Veh
KLA 1h
C57Bl/6
Balb/c
IB: ATF3
IB: Jun
IB: JunD
IB: Actin
IB: JunB
IB: FosL2
N.D.
N.D.
B
Gene Expression (Log2 RPKM)
Atf3
Fos
FosL2
Jun
JunB
JunD
Balb/c Veh
C57Bl6 Veh
Balb/c KLA 1h
C57Bl6 KLA 1h
F
Positive Motifs (Log10 p-val)
Negative Motifs (Log10 p-val)
ap-1
egr
Irf3/7/8/9
irf1
mzf1
pax2
rel
runx
sp4
PU.1
mef2a-b-d
elk_etv
elf
spib
maf_nrl
xbp1
tcf15
atf7_batf3_creb5
ctcf
cebp
myog_tcf12
thap1
dbp_hlf_tef
nfkb
srebpf
nr2e1
pax5
tcf21
pou6f2
tfcp2
duxa
pax6
irf2
arntl_mitf
nfat
myb
mafg
crem
usf
hnf1
zbtb18
rfx
ebox
e2f1
meis
tbx
esr2
hoxc9
fox
hoxb5_hoxd3
gsc
hnf4a
dmbx1
cux
hoxa9
nr2e3
homeobox-5
gmeb2
sox3_sox6
spz1
pbx1
nkx2-5
pparg
lef1
dmrt3
tead
ewsr1-fli1
phox2_prop
gata
yy2
nkx2-5var
zbtb7
arid3a
msc_myf6_tfap4
tbp
figla_id4_snai2_tcf3_tcf4
yy1
zeb1
ATF3
Jun
Fos
JunB
JunD
C
Treatment
Monomer
Train
Test
Strain
C57Bl/6
Balb/c
Treatment
Veh
KLA-1h
Monomer
ATF3
Jun
Fos
JunB
JunD
D
AP-1 TRE
AP-1 CRE
deltaSVM
TBA-2Strain
TBA
Model Predictions
Correlation to Strain KLA specific binding
Individual Motifs
E
Positive Significant Motifs
ATF3
Jun
Fos
JunB
JunD
AP-1
Spi1-c
Cebp
Fraction Non-SS peaks with mutation in motif
Fraction SS peaks with mutation in motif

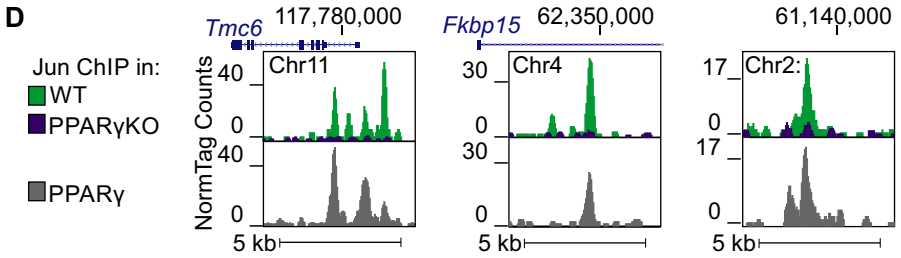
2.3.10 Validation of PPARγ as a preferential modifier of Jun binding

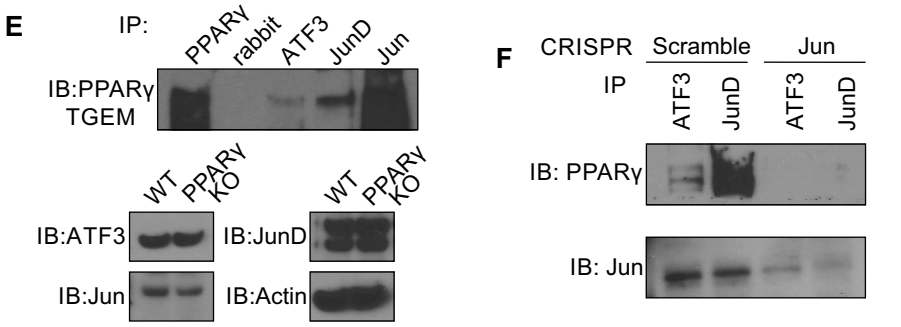
TBA and TBA-2Strain predicted that PPARγ is a preferential collaborating TF specific to Jun in resting macrophages (Fig. 2.7A, Fig. 2.11E). To confirm this prediction, we performed ChIP-seq for ATF3, Jun, JunD and PPARγ in wild type and PPARγ knockout mouse TGEMs (Fig. 2.13 A-C)[67]. Representative browser tracks are shown for Jun binding in wild-type and PPARγ knockout macrophages (Fig. 2.13D). The protein expression of ATF3, Jun and JunD are unchanged in PPARγ knockout TGEMs in comparison to wild type (Fig. 2.13E). ChIP-seq experiments in PPARγ knockout TGEMs show a marked reduction in Jun binding (Fig. 2.13A). In contrast, ATF3 and JunD show little change in binding (Fig. 2.13B, C). We found that PPARγ bound loci where Jun binding is lost in the PPARγ knockout tended to score higher for the PPARγ half site motif in comparison to Jun bound loci that did not overlap with PPARγ binding (independent T-test $p<5e\text{-}05$). To verify the specificity of the Jun antibody we also performed ChIP-seq on Jun in CRISPR mediated Jun knockout iBMDM cells and iBMDM transduced with scramble control. We observed substantial loss of Jun binding in the Jun KO cells in comparison to iBMDM cells transduced with scramble control (12 versus 25041 peaks detected with IDR < 0.05) (Fig. 2.14). Collectively, these results confirm that PPARγ specifically affects Jun recruitment.

We then probed the interactions between PPARγ and AP-1 family members by co- immunoprecipitation. ATF3, Jun, and JunD co-precipitated with PPARγ (Fig. 2.13E). As AP-1 binds as a dimer, ATF3 and JunD may be interacting with PPARγ indirectly by dimerizing with Jun. To confirm that Jun is required for interaction of ATF3 and JunD with PPARγ, we performed Co-IP from iBMDM cells in which Jun was knocked out using CRISPR/Cas9 (Fig. 2.2B). We found a loss of interaction between PPARγ and ATF3 or JunD in JunKO cells as compared to scramble control (Fig. 2.13F). This suggests that ATF3 and JunD do not interact with PPARγ in the absence of Jun.

Figure 2.13 ***(next page)*.** The Jun-specific DNA binding program is preferentially altered in PPARγ knockout macrophages. **A-C**. Changes in binding strength all binding sites in wild type macrophages in PPARγ-KO macrophages (left) and venn diagrams summarizing the change in binding at binding sites that overlap with PPARγ (right) for Jun **(A)**, ATF3 **(B)** and JunD **(C)**. D. Representative browser shots of Jun in WT and PPARγ-KO TGEMs and PPARγ in WT TGEMs. **E.** Western blot analysis of co-immunoprecipiation experiments between AP-1 monomers ATF3, Jun and JunD and PPARγ in TGEMs. **F.** Western blot analysis of co-immunoprecipiation experiments between AP-1 monomers ATF3 and JunD and PPARγ in scramble iBMDM or CRISPR mediated Jun knockout iBMDM


A
Jun
KO Normalized Reads
WT Normalized Reads
4-Fold: 0.14%
2-Fold: 1.31%
2-Fold: 41.1%
4-Fold: 11.08%
Jun
6938
2859
496
B
ATF3
4-Fold: 0.13%
2-Fold: 2.37%
2-Fold: 6.23%
4-Fold: 0.34%
ATF3
1925
11533
3547
C
JunD
4-Fold: 0.05%
2-Fold: 1.58%
2-Fold: 8.8%
4-Fold: 0.33%
JunD
5052
8099
3103
Lost in PPARγko
Unchanged in PPARγko
Gained in PPARγko
PPARγ binding sites


D
Jun ChIP in:
WT
PPARγKO
PPARγ
Tmc6
117,780,000
Fkbp15
62,350,000
61,140,000
NormTag Counts
Chr11
Chr4
Chr2:
5 kb


E
IP:
PPARγ
rabbit
ATF3
JunD
Jun
IB:PPARγ
TGEM
WT
PPARγ KO
IB:ATF3
IB:JunD
IB:Jun
IB:Actin
F
CRISPR
Scramble
Jun
IP
ATF3
JunD
IB: PPARγ
IB: Jun

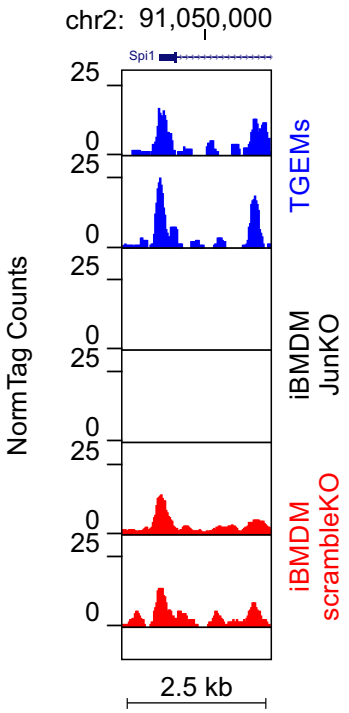


Figure 2.14. CRISPR mediated knockout of Jun leads to a drastic reduction in Jun binding by ChIP-seq. ChIP-seq was performed on iBMDM where Jun knockout was performed using CRISPR, leaving 12 peaks in comparison to the 250041 detected in scramble CRISPR treated iBMDMs and 15548 in TGEMs. Representative browser shot for Jun ChIP-seq at the Spi1 locus is shown.

2.4 Discussion

We demonstrate that AP-1 monomers have both distinct and overlapping transcriptional functions and genome-wide binding patterns in macrophages. Monomer-specific differences in DNA binding are not due to differences in the DBD contact residues as demonstrated by ATF3 chimeras with Jun or Fos DBDs. These observations led us to hypothesize that monomer-specific DNA binding patterns result from locus-specific interactions with different ensembles of collaborating TFs. To address this question, we developed a machine learning model that identified combinations of motifs that are correlated with the binding of a TF. Through this approach, we inferred TF cooperation via the presence of DNA motifs correlated with the binding of each AP-1 monomer. Leveraging the natural genetic variation found between C57BL/6J and BALB/cJ, we confirmed that mutations in motifs predicted by TBA affect AP-1 binding. Finally, we confirmed that PPARγ plays a preferential role in coordinating Jun binding in TGEMs.

In designing our machine learning model, we optimized for interpretability. We leveraged logistic regression, a relatively simple method, to accurately predict TF binding, and we were able to extract TF motifs underlying these predictions, allowing for the generation of biological hypotheses that can be experimentally validated. A secondary benefit of this approach is that the software can be readily used without specialized computing equipment or a high level of computational understanding. To improve the ability of TBA to robustly identify motifs of interest, we programmatically curated a library that "captures" the core of each motif, thereby mitigating collinearity, which can cause machine learning models to produce inaccurate results. By jointly weighing this library of motifs, TBA enables the detection of combinations of TF binding sites that can predict the distinct and overlapping DNA binding of families of TFs that recognize similar sequences. More broadly, TBA can be applied to predict of the effects of mutations on TF binding, and identify determinants of enhancer activation and open chromatin.

There are additional complexities in TF binding and enhancer activation we have not explored. Transcriptional regulation may be encoded by the spacing between motifs as well as the specific arrangement of motifs. Recent neural network architectures, such as CapsuleNets, could allow modeling of these complex properties[17;81;97]. Although more complex machine learning techniques can be applied to predict TF binding and chromatin state[1;46;76], it is challenging to extract insights from these models. Efforts to build more advanced methods to extract information from machine learning models will allow not only for interpretation of future models of greater complexity, but also better understanding of existing models[85]. For example, the procedure used by Ghandi and Lee et al to retrieve motifs from gkm-SVM can likely be improved to retrieve additional PWMs[22].

Collectively, our findings suggest two classes of collaborative TFs: 1) highly ranked TFs that are strongly correlated with the binding of all AP-1 monomers, including TFs important to macrophage identity such as such as PU.1 and C/EBPs[28;30;42;69;91;95] (Fig. 2.7A, black and grey boxes), and 2) moderately ranked TFs that specify the binding of individual AP-1 monomers (Fig. 2.7A, red and blue boxes). The former likely consists of TFs that play a role in opening chromatin

while the latter class of TFs may allow for tuning the optimal level of transcriptional activation or response. These two classes of motifs were also seen in TLR4 activated macrophages where highly ranked motifs, such as NFκB, were correlated with the binding of all AP-1 family members (Supp Table 1), while a large set of moderately ranked motifs distinguished each AP-1 monomer (Fig. 2.10C). Overall, these studies provide evidence that collaborative interactions of TFs allow a single DNA motif to be used in a wide variety of contexts, which may be a general principle for how transcriptional specificity is encoded by the genome.

2.5 Methods

2.5.1 Statistical Analyses

In Fig. 2.1C, differences in gene expression was tested using the independent T-test (degree of freedom=1, two-tailed) on two replicate experiments (n=2). Differentially expressed genes in Fig. 2.1B were identified using EdgeR[80] with default parameters, and using the cut offs FDR$<$0.05 and log2 fold change $\geq$2. In Fig. 2.3C, differences between each group (Veh, Shared, and KLA 1h) were examined using independent T-test (degree of freedom=1, two-tailed); the number of loci in each group for each monomer are as follows ATF3 (Veh=1447, Shared=7460, KLA=6997), Jun (2390, 3751, 3401), JunD (1351, 5976, 6422). Significance for motifs in Fig. 2.7A was calculated using the likelihood ratio test (degree of freedom=1) comparing the predictions made by the full TBA model and the perturbed TBA model at all loci bound in Veh treated macrophages for Atf3 (n=23160), Jun (n=15548), and JunD (n=19653). Significance for motifs in Fig. 2.8C was calculated using the likelihood ratio test (degree of freedom=1) comparing the predictions made by the full TBA model and the perturbed TBA model at all loci bound by JunD in GM12878 (n=7451), H1-hESC (n=12931), HepG2 (n=41318), K562 (n=47477), and SK-N-SH (38960). Significance for motifs in Fig. 2.9B, Fig. 2.10B, and Fig. 2.10C were calculated using the likelihood ratio test (degree of freedom=1) comparing the predictions made by the full TBA model and the perturbed TBA model at all loci bound in KLA treated macrophages for Atf3 (n=36745), Jun (n=17481), JunD (n=31641), Fos (n=24365), Fosl2 (n=10619), and JunB (n=13376). Significance values for

Fig. 2.11F and 2.12F were calculated using the F-test; the number of loci analyzed for monomers in Vehicle treated macrophages are: ATF3 (n=4163), Jun (n=3004), and JunD (n=4148); the number of loci analyzed for monomers in KLA treated macrophages are: Atf3 (n=4577), Jun (n=3232), JunD (n=4366), Fos (n=4477), and JunB (n=3616).

2.5.2 Generating Custom Genome for BALB/cJ

A custom genome for BALB/cJ by replacing invariant positions of the mm10 genome with alleles reported by the Mouse Genomes Project (version 3 VCF file)[44]. For C57BL/6J the mm10 reference genome from the UCSC genome browser was used. To allow for comparisons between BALB/cJ and C57BL/6J during analysis, the coordinates for the custom genome for BALB/cJ was shifted to match the positions of the mm10 reference genome using MARGE[65]. We did not analyze any reads that fell within deletions in BALB/cJ. Reads that overlapped with an insertion were assigned to the last overlapping position in the reference strain.

2.5.3 Analysis of ChIP-seq Peaks

Sequencing reads from ChIP-seq experiments were mapped to the mm10 assembly of the mouse reference genome (or the BALBc/J custom genome) using the latest version of Bowtie2 with default parameters[52]. Mapped ChIP-seq reads to identify putative transcription factor binding sites with HOMER[29] findPeaks command (with parameters -size 200 -L 0 -C 0 -fdr 0.9), using the input ChIP experiment corresponding to the treatment condition. In order to reduce the number of false positive peaks, we calculated the Irreproducible Discovery Rate (IDR) at each peak (using version 2.0.3 of the idr program) with the HOMER peak score calculated for each replicate experiment as the input to IDR and then filtered all peaks that had IDR ≥ 0.05[61]. De novo motifs were calculated with the HOMER findMotifsGenome.pl command with default parameters. Enrichment of de novo motifs was calculated using the findKnownMotifs.pl program in HOMER with default parameters.

Quantification of RNA Expression Reads generated from RNA-seq experiments were aligned to the mm10 mouse reference genome (or the BALBc/J custom genome) using STAR aligner with default parameters[16]. To quantify the expression level of each gene, we calculated the Reads Per

Kilobase of transcript per Million mapped reads (RPKM) with the reads that were within an exon. Un-normalized sequencing reads were used identify differentially expressed genes with EdgeR[80]; we considered genes with FDR < 0.05 and a change in expression between two experimental conditions two fold or greater differentially expressed. To quantify the expression of nascent RNAs we annotated our ChIP-seq peaks with the number of GRO-seq reads (normalized to 10 million) that were within 500 bps of the peak center using the HOMER annotatePeaks.pl command.

2.5.4 TBA Model Training

For each AP-1 monomer under each treatment condition, we trained a model to distinguish binding sites for each monomer from a set of randomly selected genomic loci. The set of random background loci used to train each model was selected according to the following criteria: 1) the GC content distribution of the background loci matches the GC content of the binding sites for a given monomer, 2) contain no ambiguous or unmappable positions, and 3) the number of background sequences matches the number of binding sites k. For each of the sequences in the combined set of the binding sites and background loci, we calculated the highest log-odds score (also referred to as motif score) for each of the n motifs that will be included in the model[88] Motif matches in both orientations were considered. Log-odds scores less than 0 were set to 0. Per standard preprocessing procedures prior to training a linear model, we standardized the log-odds scores for each motif, scaling the set of scores for each motif so that the mean value is 0, and the variance is 1. Standardization scales the scores for all motifs to the same range (longer motifs have a larger maximum score) and also helps to reduce the effect of multi-collinearity on the model training. And so, the features used for training our model is an n by $2k$ matrix of log-odds scores standardized across each row. To generate the corresponding array of labels, we assigned each binding site a label of 1 and each background loci a label of 0. Using this feature matrix, and label array, we trained weights for each motif using an L1 penalized logistic regression model as implemented by the scikit-learn Python package[74]. Motif weights shown in our analysis are the mean values across five rounds of cross validation, using 80% of the data for training and 20% for testing in each round.

Models were trained for ChIP-seqs generated in this study as well as data downloaded from the NCBI Gene Expression Omnibus (accession number GSE46494) and the ENCODE data portal (www.encodeproject.org).

2.5.5 Quantification of Multiple Collinearity

To assess the extent of multi-collinearity in the motif score features we used to train our models, we took each feature matrix corresponding to each experiment and calculated the Variance Inflation Factor (VIF) for each motif[6]. To calculate the VIF, we first determine the coefficient of determination, R2, for each motif by regressing the log-odds scores for one motif against the log-odds scores of the remaining motifs. Next using the coefficient of determination, the tolerance for each motif can be calculate as the difference between 1 and the coefficient of determination $(1-R^2)$. The VIF is the reciprocal of the tolerance $\frac{1}{1-R^2}$. We used the linear_model module of sklearn Python package to calculate the coefficient of determination.

2.5.6 Motif Clustering and Merging

We scored the similarity of all pairs of DNA sequence motifs by calculating the Pearson correlation of the aligned position probability matrices (PPMs) corresponding to a given pair of motifs[66]. The Pearson correlation for a pair of motifs A and B of length i is calculated using the formula:

$$\frac{\Sigma_i \Sigma_j^4 (A_{ij} - \bar{A}_i)(B_{ij} - \bar{B}_i)}{\sqrt{(\Sigma_i \Sigma_j^4 (A_{ij} - \bar{A}_i))^2} \sqrt{(\Sigma_i \Sigma_j^4 (B_{ij} - \bar{B}_i))^2}}$$

PPMs were first aligned using the Smith-Waterman alignment algorithm[87]. Shorter motifs are padded with background frequency values prior to alignment. Gaps in the alignment were not allowed and each position in the alignment was scored with the Pearson correlation. The Pearson Correlation was then calculated using the optimal alignment. Next, sets of motifs that have PPMs with a Pearson correlation of 0.9 or greater were merged by iteratively aligning each PPM within the set, and then averaging the nucleotide frequencies at each position.

2.5.7 Assessing Significance of Motifs for TBA

p-values for TBA were calculated using the log likelihood ratio test. Each motif was removed from the set of features used to train a perturbed TBA model (using five-fold cross validation). We then used the full model (containing all motifs) and the perturbed model to calculate the likelihood of observing binding on all binding sites and background sequences for a given monomer and all the background regions. The difference in the likelihoods calculated by the full model and the perturbed model was then used to perform the chi-squared test for each motif. The chi-squared test was performed using the scipy python package[39]

2.5.8 Comparison to other Methods

BaMM motif and gkm-SVM were both run with default parameters. We used the latest version of the larger scale gkm-SVM, LS-GKM (compiled from source code downloaded from github.com/Dongwon-Lee/lsgkm on 8/25/16), and BaMM motif (v1.0 downloaded from github.com/soedinglab/BaMMmotif[56;86]. Both models were trained using five-fold cross validation. Model performance was scored using roc_auc_score and precision_score functions from the metrics module of sklearn.

2.5.9 Predicting changes in AP-1 binding after one-hour KLA treatment

To predict the change in binding after KLA treatment, we leveraged the motif weights learned for each of the n motifs (w_n) by a TBA model trained on the Vehicle treated data ($W_veh = [w_{veh,1}, ...w_{veh,n}]$) and a TBA model trained on the one-hour KLA treated data ($W_kla = [w_{kla,1}, ...w_{kla,n}]$) for each AP-1 monomer. The predicted change in binding for each sequence is then then the difference between the dot product of the standardized motif scores calculated for the sequence each of the k binding sites ($S_k = [s_{1,k}, ..., s_{n,k}]$) with the KLA motif weights and the dot product of the motif scores and the Veh motif weights ($\Delta_{kla-veh,k} = W_{kla} \cdot S_k - W_{veh} \cdot S_k$). Predictions were made for all genomic loci that intersected with a peak for one of the AP-1 monomers in either the vehicle or KLA treatment condition.

2.5.10 Predicting strain specific binding with TBA

To predict strain specific binding, we leveraged the motif weights learned for each of the n motifs (w_n) by a TBA model ($W = [w_1, ..., w_n]$) for each AP-1 monomer using the C57BL/6J data, and the motif scores calculated for each of the k binding sites using the genomic sequence for C57BL/6J and BALBc/J ($S_{C57,k} = [s_{C57,1,k}, ..., s_{C57,n,k}], S_{BAL,k} = [s_{BAL,1,k}, ..., s_{BAL,n,k}]$). Next, we computed the difference of the motif scores for C57BL6/J and BALBc/J ($[D_n = [s_{C57,n,1} - s_{BAL,n,1}, ..., s_{C57,n,k} - s_{BAL,n,k}]$) and then standardized the score differences for each motif across all the k binding sites that had a mutation when comparing BALBc/J to C57BL/6J, yielding standardized motif score differences for each binding site ($Z_n = standardize(D_n) = [z_{n,1}, ..., z_{n,k}]$). Finally, we then made a prediction for strain specific binding by computing the dot product of the motif weights and the standardized difference of the motif scores between C57BL6/EiJ and BALBc/J for the k^{th} mutated binding site ($\Delta_{C57-BAL} = W \cdot [z_{1,k}, ..., z_{n,k}]$).

2.5.11 TBA-2Strain Model Training

For each genomic loci that intersected with a peak for one of the AP-1 monomers, in either C57BL/6J or BALBc/J, we calculated the highest log-odds score for each of the n motifs that will be included in the model, using the genomic sequence from both strains, yielding a two sets of motif scores for each of the k binding sites ($S_{C57,k} = [s_{C57,1,k}, ..., s_{C57,n,k}], S_{BAL,k} = [s_{BAL,1,k}, ..., s_{BAL,n,k}]$). Motif matches in both orientations were considered. Log-odds scores less than 0 were set to 0. Using the motif scores, we compute the standardized difference of the motif scores across the two strains as described in the above section ($Z_n = [z_{n,1}, ..., z_{n,k}]$).And so, the features used for training our model is an n by k matrix of log-odds scores standardized across each row. Next, we calculated the log2 fold ratio of the number of ChIP-seq reads in C57BL/6J compared to BALBc/J to represent the extent of strain specific binding. Using this feature matrix, and setting the log2 fold ratio of binding between the two strains as the dependent variable, we trained weights for each motif using linear regression as implemented by the scikit-learn Python package. Motif weights shown in our analysis are the mean values across five rounds of cross validation, using 80% of the data for training

and 20% for testing in each round. Predictions for strain specific binding can be made using the calculated weights following the procedure in the previous section.

2.5.12 Code Availability

All algorithms relating to training and testing our model, TBA, has been implemented using Python. Source code and executable files are available at: https://github.com/jenhantao/tba.

2.5.13 ChIP protocol

Protein A and G Dynabeads 50/50 mix from Invitrogen are sued for ChIP (10001D, 10003D). IP mix consists of 20ul beads/2ug antibody per 2 million cell ChIP. Antibodies against AP-1 family members were chosen for targeting of non-conserved regions to minimize the potential for non-specific binding. Antibodies are listed in 2.3. For preparation, beads were washed with 2x with 0.5%BSA-PBS, then beads-antibody were incubated with 0.5%BSA-PBS for at least 1h on rotator (4°C). Wash 2X with 0.5%BSA-PBS, then resuspended in dilution buffer (1% Triton, 2mM EDTA, 150 mM NaCl, 20 mM Tris-HCl (pH 7.4), 1X Protease Inhibitors). Double Crosslinking for ChIP. Media was decanted from cells in 10 cm plates, wash once briefly with PBS (RT). Disuccinimidyl glutarate (Pierce Cat # 20593) (diluted in DMSO at 200mM)/PBS (RT) was used for 10 min. Then Formaldehyde was added to a final concentration of 1% for an additional 10 min. Reaction was quenched with 1:10 1M Tris pH 7.4 on ice. Cells were collected and washed twice with cold PBS, spinning at 1000xg for 5 min. Nuclei Isolation and Sonication. Resuspend cell pellets in 1 ml of Nuclei Isolation Buffer (50 mM Tris-Ph 8.0, 60 mM KCl, 0.5% NP40) + PI and incubate on ice for 10 minutes. Centrifuge 2,000xg for 3 minutes at 4°C. Resuspend nuclei in 200 ul of fresh Lysis Buffer (0.5% SDS, 10mM EDTA, 0.5mM EGTA, 50mM Tris-HCl (pH 8))+ PI. Sonication. Nuclei were then sonicated (10 million cells) for 25 minutes in a Biorupter (settings= 30 seconds=On, 30 seconds=Off, Medium) using thin wall tubes (Diagenode Cat# C30010010). After sonication spin max speed for 10 minute at 4°C. ChIP set up. Sonicated DNA was diluted 5X with 800 Dilution Buffer (1% Triton, 2mM EDTA, 150 mM NaCl, 20 mM Tris-HCl (pH 7.4), 1X Protease Inhibitors). An aliquot is removed for input samples (5%). Samples ON at 4° C while rotating.

Table 2.3. A List of Antibodies used in this study.

Reactivity	Description	Company	Cat #
ATF2	Rabbit polyclonal	Santa Cruz	sc-187
Atf3	Rabbit polyclonal	Thermo	PA5-36244
Atf4	Rabbit polyclonal	Cell Signaling	11815
Atf4	Rabbit polyclonal	Sigma	ABE387
Fos	Rabbit polyclonal	Santa Cruz	sc-7202
FosL1	Rabbit polyclonal	Santa Cruz	sc-605
FosL2	Mouse monoclonal	Santa Cruz	sc-166102
Fosb	Rabbit polyclonal	Cell Signaling	2251
Jun	Rabbit polyclonal	Santa Cruz	sc-1694
JunB	Rabbit polyclonal	Santa Cruz	sc-73
JunD	Rabbit polyclonal	Santa Cruz	sc-74
Jdp2	Rabbit polyclonal	Thermo	PA5-19692
Batf	Rabbit polyclonal	Brookwood Biomedical	PAB4003
Batf2	Rabbit polyclonal	Santa Cruz	sc-241891
Batf3	Rabbit polyclonal	Abnova	H00055509-M04
CEBPa	Rabbit polyclonal	Santa Cruz	sc-61
Pu.1	Rabbit polyclonal	Santa Cruz	sc-352
PPARg	Rabbit polyclonal	Santa Cruz	sc-7196
PPARg	Rabbit monoclonal	Cell Signaling	C26H12
PPARg	Rabbit polyclonal	Diagenode	C15410133

Washing. ChIP are washed 1X with TSE I (20mM Tris-HCl pH7.4, 150mM NaCl, 0.1%SDS, 1% Triton X-100, 2mM EDTA), 2X with TSE III (10mM Tris-HCl pH7.4, 250mM LiCl, 1%IGEPAL, 1%Deoxycholate, 1mM EDTA), 1X with TE+0.1%TritonX-100, transfer to new tube and then wash another time with TE+0.1%TritonX-100. Elution. Elute with 200 μL Elution Buffer (1% SDS, 10mM Tris pH7.5) for 20 minutes at RT, shaking on the vortex or a nutator or rotator. De-crosslinking. Add 10 μL of 5 M NaCl and incubate ON at 65 C (or at least 8 hours). Clean up samples using Zymo ChIP DNA Clean and Concentrator. Elute in 100 μL. Take 40 μL and proceed to library prep protocol.

2.5.14 PolyA RNA Isolation and Fragmentation

RNA isolation. RNA was isolated using TRIZOL-reagent (ambion cat# 15596018) and DIRECT-ZOL RNA mini-prep kit (cat# 11-330MB). Poly-A RNA isolation. Use 0.2 Total RNA as

starting material for ideal mapping efficiency and minimal clonality. Collect 10 μL oligo (dT) (NEB cat# S1419S) beads per RNA sample. Beads were washed twice with 1x DTBB (20mM Tris-HCl pH7.5, 1M LiCl, 2mM EDTA, 1% LDS, 0.1% Triton X-100). Beads were resuspended in 50 μL of 2x DTBB. 50 μL of beads were mixed with 50 μL RNA and Heated to 65 °C for 2 min. RNA-beads were then incubated for 10 min at RT while rotating. RNA-beads were then collected on a magnet and washed 1x each with RNA WB1 (10mM Tris-HCl pH7.5, 0.12 M LiCl, 1mM EDTA, 0.1% LDS, 0.1% Triton X-100)and WB3 (10mM Tris-HCl pH7.5, 0.5M LiCl, 1mM EDTA). Add 50μL Tris-HCl pH7.5 and heat to 80°C for 2 min to elute. Collect RNA and perform a second Oligo-dT bead collection. After washing the second collection, instead of eluting was 1X with 1X First strand buffer (250 mM Tris-HCl (pH 8.3), 375 mM KCl, 15 mM $MgCL_2$ (ThermoFisher SSIII kit Cat# 18080093). Fragmentation. Then Add 10μL of 2X First strand buffer plus 10mM DTT and fragment DNA at 94°C for 9 min. Collect beads on magnet and transfer eluate containing fragmented mRNA to a new PCR strip. Should recover 10 μL fragmented RNA. First strand synthesis. We mixed fragmented RNA with 0.5μL Random Primer (3 μg/μL) Life Tech #48190-011, 0.5μL oligo-dT (50μM from SSIII kit), 1μL dNTPs (10mM Life Tech, cat 18427088) and 0.5μL SUPERase-In (ThermoFisher Cat#AM2696) and heat 50°C for 1 min. Immediately place on ice. We then added 5.8μL ddH_2O, 0.1 μL Actinomycin (2ug/μL Sigma cat#A1410), 1μL DTT (100mM Life Tech cat# P2325), 0.2μL of 1% Tween and 0.5 μL of Superscript III and incubate 25°C for 10 min, then 50°C for 50 min. Bead clean up. We added 36 μL of RNAClean XP (ampure XP) and mixed, incubating for 15 min on ice. The beads were then collected on a magnet and washed 2X with 75% ethanol. Beads were then air-dried for 10 min and elute with 10 μL nuclease free H_2O. Second strand synthesis. 10μL of cDNA/RNA was mixed with 1.5μL 10X Blue Buffer (Enzymatics cat# B0110L), 1μL dUTP/dNTP mix (10mM Affymatrix cat# 77330), 0.1μL dUTP (100mM Affymatrix cat# 77206), 0.2μL RNase H (5U/μL Enzymatics cat# Y9220L), 1μL DNA polymerase I (10U/μL Enzymatics cat#P7050L), 0.15 μL 1% Tween-20 and 1.05μL nuclease free water. Reaction was incubated at 16°C for 2.5 hours. Bead clean up. DNA was purified by adding 1μL Seradyn "3 EDAC" SpeedBeads (Thermo 6515-2105-050250) per reaction in 28μL 20% PEG8000/2.5 M NaCl

(13% final concentration) and incubating at RT for 10min. Beads were then collected on a magnet and washed 2X with 80% Ethanol. Beads were air-dried for 10min and eluted in 40μL of nuclease free water. DNA is ready for library prep.

2.5.15 Library Prep Protocol

dsDNA End Repair. We mixed 40μL of DNA from ChIP or RNA protocols with 2.9μL of H_2O, 0.5μL 1% Tween-20, 5μL 10X T4 ligase buffer (Enzymatics cat# L6030-HC-L), 1μL dNTP mix (10 mM Affymetrix 77119), 0.3 μL T4 DNA pol (Enzymatics P7080L), 0.3μL T4 PNK (Enzymatics Y9040L), 0.06μL Klenow (Enzymatics P7060L) and incubated for 30min at 20°C. 1μL of Seradyn 3 EDAC SpeedBeads (Thermo 6515-2105-050250) in 93 μL 20% PEG8000/2.5M NaCl (13% final) was added and incubated for 10 min. Bead clean-up. Beads were collected on a magnet and washed 2X with 80% ethanol. Beads were air-dried for 10 min and then eluted in 15μL ddH_2O. dA-Tailing. DNA was mixed with 10.8μL ddH_2O, 0.3μL 1% Tween-20, 3μL Blue Buffer (Enzymatics cat# B0110L), 0.6μL dATP (10mM Tech 10216-018), 0.3μL Klenow 3- 5 Exo (Enzymatics P7010-LC-L) and incubated for 30min at 37°C. 55.8μL 20% PEG8000/2.5 M NaCl (13% final) was added an incubated for 10 min. Then bead clean up was done. Beads were eluted in 14μL. Y-Shape Adapter Ligation. Sample was mixed with 0.5μL of a BIOO barcode adapter (BIOO Scientific cat# 514104), 15μL Rapid Ligation Buffer (Enzymatics cat@ L603-LC-L), 0.33μL 1% Tween-20 and 0.5μL T4 DNA ligase HC (Enzymatics L6030-HC-L) and incubated for 15 min at RT. 7 μL of 20% PEG8000/2.5 M NaCl was added and incubated for 10min at RT. Bead clean up was performed and beads were eluted in 21μL. 10μL was then used for PCR amplification (14 cycles) with IGA and IGB primers (AATGATACGGCGACCACCGA, CAAGCAGAAGACGGCATACGA).

2.5.16 GRO-seq

Nascent transcription was captured by global nuclear run-on sequencing (GRO-seq). Nuclei were isolated from TGEMs using hypotonic lysis (10mM Tris-HCl (pH 7.5), 2mM $MgCl_2$, 3mM $CaCl_2$; 0.1% IGEPAL CA-630) and flash frozen in GRO-freezing buffer (50mM Tris-HCl (pH 7.8),

5mM $MgCL_2$, 40% Glycerol). Run-on. 3-5 x 10^6 BMDM nuclei were run-on with BrUTP-labelled NTPs with 3x NRO buffer (15mM Tris-Cl (pH 8.0), 7.5mM $MgCl_2$, 1.5mM DTT, 450mM KCl, 0.3U/μL of SUPERase In, 1.5% Sarkosyl, 366μM ATP, GTP (Roche), Br-UTP (Sigma 40 Aldrich) and 1.2μM CTP (Roche, to limit run-on length to $\sim$40 nucleotides)). Reactions were stopped after five minutes by addition of 500μL Trizol LS reagent (Invitrogen), vortexed for 5 minutes and RNA extracted and precipitated as described by the manufacturer. RNA pellets were resuspended in 18μL ddH_2O + 0.05% Tween (dH_2O+T) and 2μL fragmentation mix (100 mM$ZnCL_2$, 10mM Tris-HCl (pH 7.5)), then incubated at 70°C for 15 minutes. Fragmentation was stopped by addition of 2.5μL of 100mM EDTA. BrdU enrichment. BrdU enrichment was performed using BrdU Antibody (IIB5) and AC beads (Santa Cruz, sc-32323 AC, lot #A0215 and #C1716). Beads were washed once with GRO binding buffer (0.25xsaline-sodium-phosphate-EDTA buffer (SSPE), 0.05% (vol/vol) Tween, 37.5mM NaCl, 1mM EDTA) + 300mM NaCl followed by three washes in GRO binding buffer and resuspended as 25% (vol/vol) slurry with 0.1 U/μL SUPERase-in. To fragment RNA, 50μL cold GRO binding buffer and 40μL equilibrated BrdU antibody beads were added and samples slowly rotated at 4°C for 80 minutes. Beads were subsequently spun down at 1000xg for 15 seconds, supernatant removed and the beads transferred to a Millipore Ultrafree MC column (UFC30HVNB; Millipore) in 2x 200μL GRO binding buffer. The IP reaction was washed twice with 400μL GRO binding buffer before RNA was eluted by incubation in 200μL Trizol LS (Thermo Fisher) under gentle agitation for 3 minutes. The elution was repeated a second time, 120μL of dH_2O+T added to increase the supernatant and extracted as described by the manufacturer. End repair and decapping. For end-repair and decapping, RNA pellets were dissolved in 8μL TET (10mM Tris-HCl (pH 7.5), 1mM EDTA, 0.05% Tween20) by vigorous vortexing, heated to 70°C for 2 minutes and placed on ice. After a quick spin, 22μL Repair master mix (3μL 10x PNK buffer, 15.5μL dH_2O+T, 0.5μL SUPERase-In RNase Inhibitor (10 U), 2μL PNK (20U), 1μL RppH (5U)) was added, mixed and incubated at 37°C for 1 hour. To phosphorylate the 5'end, 0.5μL 100mM ATP was subsequently added and the reactions were incubated for another 45 minutes at 37°C (the high ATP concentration quenches RppH activity). Following end repair, 2.5μL 50mM EDTA was added, reactions mixed

and then heated to 70°C for 2 minutes before being placed on ice. A second BrdU enrichment was performed as detailed above. RNA pellets were dissolved in 2.75μL TET + 0.25μL Illumina TruSeq 3'Adapter (10μM), heated to 70°C for 2 minutes and placed on ice. 7 of 3'master mix (4.75μL 50% PEG8000, 1μL 10x T4 RNA ligase buffer, 0.25μL SUPERase-In, 1μL T4 RNA Ligase 2 truncated (200U; NEB)) was added, mixed well and reactions incubated at 20°C for 1 hour. Reactions were diluted by addition of 10μL TET + 2μL 50mM EDTA, heated to 70°C for 2 minutes, placed on ice and a third round of BrUTP enrichment was performed. RNA pellets were transferred to PCR strips during the 75% ethanol wash and dried. Samples were dissolved in 4μL TET (10mM Tris-HCl (pH 7.5), 0.1mM EDTA, 0.05% Tween 20) + 1μL 10μM reverse transcription (RT) primer. To anneal the RTprimer, the mixture was incubated at 75°C for 5 minutes, 37°C for 15 minutes and 25°C for 10 minutes. To ligate the 5' Illumina TruSeq adapter, 10μL 5'master mix (1.5μL dH_2O + 0.2% Tween20, 0.25μL denaturated 5'TruSeq adapter (10μM), 1.5μL 10x T4 RNA ligase buffer, 0.25μL SUPERase-In, 0.2μL 10mM ATP, 5.8μL 50% PEG8000, 0.5μL T4 RNA ligase 1 (5U; NEB)) was added and reactions were incubated at 25°C for 1 hour. Reverse transcription was performed using Protoscript II (NEB) (4μL 5x NEB FirstStrand buffer (NEB; E7421AA), 0.25μL SUPERase-In, 0.75μL Protoscript II (150U; NEB)) at 50°C for 1 hour. After addition of 30μL PCR master mix (25μL 2X LongAmp Taq 2X Master Mix (NEB), 0.2μL 100μM forward primer, 2.8μL 5M Betaine and 2μL 10μM individual barcoding primer), mixtures were amplified (95°C for 3 minutes, (95°C for 60 seconds, 62°C for 30 seconds, 72°C for 15 seconds) x13, 72°C for 3 minutes). PCR reactions were cleaned up using 1.5 volumes of SpeedBeads (GE Healthcare) in 2.5M NaCl/20% PEG8000. Libraries were size selected on PAGE/TBE gels to 160-225 base pairs. Gel slices were shredded by spinning through a 0.5 ml perforated PCR tube placed on top of a 1.5ml tube. 150μL Gel EB (0.1% LDS, 1M LiCl, 10mM Tris-HCl (pH 7.8)) was added and the slurry incubate under agitation overnight. To purify the eluted DNA, 700μL Zymogen ChIP DNA binding buffer was added into the 1.5ml tube containing the shredded gel slice and the Gel EB, mixed by pipetting and the slurry transferred to a ZymoMiniElute column. Samples were first spun at 1000xG for 3 minutes, then 10,000xG for 30 seconds. Flow through was removed, and samples

washed with 200 ul Zymo WashBuffer (with EtOH). Gel remainders were removed by flicking and columns washed by addition of another 200μL Zymo WashBuffer (with EtOH). Flow through was removed, columns spun dry by centrifugation at 14,000xG for 1 minute and DNA eluted by addition of 20μL pre-warmed Sequencing TET (10mM Tris-HCl (pH 8.0), 0.1 mM EDTA, 0.05% Tween 20). Libraries were sequenced.

2.5.17 Western Blotting

Cells were lysed with Igepal lysis buffer (50mM Tris pH8.0, 150mM NaCl, 0.5% Igepal) and protein concentrations were determined with BioRad protein assay reagent using BSA as a standard. Proteins were separated on NuPage 4-12% Bis-Tris gradient gels (Invitrogen) and transferred onto a nitrocellulose membrane (Amersham). Membranes were blocked in TBS with 0.1% Tween-20 and 5% BSA. Membranes were blotted with the indicated primary overnight at 4oC. Horseradish peroxidase conjugated secondary antibodies were detected using ECL plus western blotting detection system (Amersham).

2.5.18 Animals and Cell Culture

TGEMs were collected 3 days after injection from male 8 week C57Bl/6J, or BALB/cJ mice, and plated at 20 x 106 cells per 15 cm Petri dish in DMEM plus 10% FBS and 1x penicillin-streptomycin. One day after plating, cells were supplemented with fresh media and treated with PBS (Veh) or 100 ng/mL KLA for 1 hour, and then directly used for downstream analyses. iBMDM are produced by infection of BMDM with a retrovirus containing myc and Braf V600E[21]. The immortalized cells are then grown out over several weeks. All animal experiments were performed in compliance with the ethical standards set forth by University of California, San Diegos Institutional Annual Care and Use Committee (IUCAC).

2.5.19 Lentivirus Production

pLentiguide was modified to contain a U6-bsmbi-spgRNA scaffold and a CMV promoter driving tagBFP2. 2 CRISPR guides were inserted for each target via PCR amplification with the

Table 2.4. Guide RNAs used for CRISPR experiments

Target	Guide Sequence	Source
ATF3-mouse	GTCAAATACCAGTGACCCAGG	This study
ATF3-mouse	GCTTGGTGACTGACATCTCCA	This study
Jun-mouse	gcttcccagtgtcacctccg	This study
Jun-mouse	GCTCTCGGACTGGAGGAACGG	This study
JunD-mouse	gctcaggttggcgtagaccg	This study
JunD-mouse	gccgagtctcgaaagagtccg	This study
Scramble-mouse	GCACTACCAGAGCTAACTCA	This study

H1 promoter (bsmbi site/guide1/scaffold/H1 promoter/guide 2/bsmb1 site) for a total of 2 guides per virus (U6 and H1 driven) (2.4). Virus was made with pVSVg/ppAX2 system. 2 days post transfection, media was collected and centrifuged at 4°C for 2 hours at 20,000xg. Cell pellet was reconstituted overnight at 4°C in OPTI-MEM and stored at -80°C.

2.5.20 Production of CRISPR KO iBMDMs

KO iBMDMs were produced using lentiviral infection. iBMDM-CAS9-IRES-EGFP were infected with MOI 100, as measured on 293T cells, with Lentiblast (OZ biosciences) (5μL each reagent) in OPTI-MEM. This was then centrifuged at 1300g for 1h at room temperature. Media was then removed and cells were supplemented in bone marrow media (30% L-cell, 20% FBS, 1% penicillin/streptomycin in DMEM) for two days. Cells were then sorted for infection by expression of a transgene on the viral sequence (tagBFP2).

2.5.21 Data Availability

Data generated for this study has been deposited to the NCBI Gene Expression Omnibus (GEO) under the accession number GSE111856. Previously published data was downloaded from GEO (accession number GSE46494) and the ENCODE data portal:
(https://www.encodeproject.org).

2.6 Acknowledgements

We thank L. Van Ael for assistance with manuscript preparation and J. Collier, M. Pasillas and Z. Ouyang for technical assistance. These studies were supported by NIH grants DK091183, CA17390 and GM085764 and Leducq Transatlantic Network grant 16CVD01 to CKG. DNA sequencing was supported by NIH grant DK063491. SHD is a CRI-Irvington Postdoctoral Fellow. TS was supported by the Swedish Society for Medical Research. GJF was supported by a Canadian Institute of Health Research Postdoctoral Fellowship, FME-135475. MS was supported by the Manpei Suzuki Diabetes Foundation of Tokyo, Japan, and the Osamu Hayaishi Memorial Scholarship for Study Abroad, Japan.

Chapter 2, in part, has been submitted for publication. Fonseca, G.J.*, Tao J.*, Westin, E.M., Duttke, S.H., Spann, N.J., Strid, T., Shen, Z., Stender, J.D., Sakai, M., Link, V.M., Benner, C., Glass, C.K. Diverse motif ensembles specify non-redundant DNA binding activities of AP-1 family members in macrophages. (* These authors contributed equally to this work). The dissertation author was one of the primary investigators and authors of this study.

Chapter 3

A method for describing transcription factor binding specificity as a set of DNA motifs

3.1 Abstract

The sequence specificity of DNA binding transcription factors (TFs) has typically been described using position weight matrices (PWMs). While PWMs are intuitive, PWMs poorly discriminate TF binding sites from random genomic sequences. In contrast, machine learning models are more complex and difficult to interpret. Here, we describe a machine learning approach inspired by the biological phenomena that TFs bind collaboratively with one another. Our machine learning model learns to characterize the binding specificity of a TF as an ensemble of enriched and depleted PWMs. Information extracted from our model can yield mechanistic insight into the activity of TFs as well as the non-coding regulatory regions bound by multiple TFs. We present a machine learning tool, ABTBA (A Bigger TF Binding Analysis), for the analysis of TF binding motif interactions in the context of non-coding regulatory elements. ABTBA uses a programmatically curated library of motifs formed from the JASPAR and CISBP databases that reduces multiple collinearity and enhances the interpretability of the model. ABTBA then analyzes genomic sequence to learn combinations of motifs that are associated with regulatory elements. We apply ABTBA to 363 ENCODE ChIP-seq data sets, and extract information suggesting that each TF interacts with dozens of other TFs genome-wide and 3-4 TFs at a single locus in a cell type specific

manner. Additionally, we demonstrate that ABTBA can detect cell type specific combinations of motifs that can be integrated with RNA-seq to identify TFs that establish accessible chromatin landscape within the hematopoietic stem cell lineage.

3.2 Introduction

Gene expression in mammalian organisms is a complex process that allows for the same genome to be interpreted uniquely in a spectrum of environmental contexts encountered by many different cell types. This diversity in transcriptional activity is, in part, enabled by the activity of sequence specific transcription factors (TFs) that bind to gene proximal regulatory sequences (promoters) and distal regulatory sequences (enhancers)[31;59;84]. TFs bind collaboratively with one another[28;31;65;103] to tens of thousands of enhancers in a given cell type out of hundreds of thousands of possible enhancers present in mammalian genomes[2;79;99]. The target DNA sequence motif of a TF has typically been described using a Position Weight Matrix (PWM), which summarizes the frequency of each nucleotide at a given position of a TF's binding site. The quality with which a sequence matches to the PWM can be described using the log-odds score (also referred to as motif score)[88]. Experimentally, a PWM for a TF can be inferred using chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq), protein binding microarrays, as well as HT-SELEX[3;11;38]. Thousands of PWMs are currently available in TF motif databases such as JASPAR and CISBP[47;51;68;102].

While use of the PWM is widespread due to its intuitive nature, more recent machine learning based models have been introduced that can more accurately discriminate TF binding sites from unbound genomic sequences. Machine learning models can learn to relate sequence based features, such as the frequency of k-mers per sequence, to TF binding[1;22;55]. In comparison to PWMs, which can be described using tens of parameters, machine learning models use thousands of parameters and, as a consequence, are more difficult to interpret. Additionally, although machine learning models considerably improve upon the performance of the PWM, the biological insight extracted from these models has been limited; often restricted to a couple DNA motifs represented

as PWMs. And so, no information beyond a PWM is learned or represented. However, given the superior performance of machine learning models, there is likely further insights into TF mechanism and behavior that can be gleaned using machine learning approaches.

Here we present a machine learning tool that builds upon our previous work[19] that leverages public motif PWM databases to model TF binding with high performance and interpretability. Our improved tool, ABTBA (A Bigger TF Binding Analysis) features additional capabilities for analyzing motif databases and characterizing multiple collinearity that allows ABTBA to leverage hundreds of PWMs drawn from the JASPAR and CISBP databases as well as additional user specified databases without diminishing the interpretability of the model. Modeling the collaborative binding behavior of TFs, ABTBA uses logistic regression to learn to represent the binding specificity of a TF as an ensemble of motifs drawn from a programmatically curated library of motifs. In our application of logistic regression, the probability that a genomic sequence is bound by a TF is calculated as the weighted sum of how well that sequence matches to each motif in our curated motif library. The weight of each motif can be used to determine the importance of each motif to the model. The interpretation of the model weights from logistic regression can be confounded by multiple collinearity, or the presence of highly correlated features. While multiple collinearity does not affect the performance of the model, it severely limits the amount of useful information that can be extracted from the model coefficients[6]. To reduce multiple collinearity, we merge highly similar motifs, using the Variance Inflation Factor (VIF) to assess the presence of collinear features, thereby stabilizing the weights learned by ABTBA[6]. The ensemble of motifs learned by ABTBA for a TF describes the binding preference of that TF as not only the TF's PWM but also the PWMs of potential binding partners and depleted motifs, giving additional insight into the binding behavior of that TF beyond a single PWM. We apply our approach to hundreds of TF ChIP-seq datasets from ENCODE and extract information from our model that suggests that on average each TF interacts with dozens of other TFs genome-wide and with 3-4 other DNA binding TFs on a per locus basis. Additionally, we demonstrate that our model can be used to extract lineage determining TFs for the hematopoietic cell lineage from open chromatin regions identified by ATAC-seq.

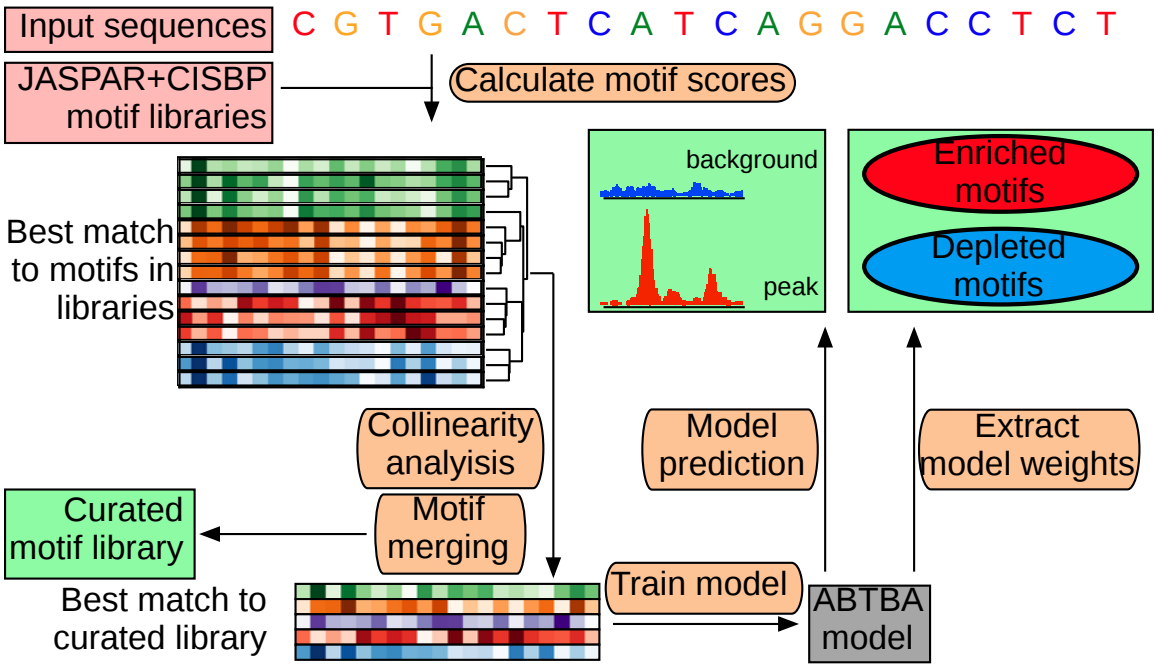


Figure 3.1. Overview of ABTBA model. Input sequences from ChIP-seq or ATAC-seq peaks (mixed with background sequences) are first scored for the best match to hundreds of motifs drawn from the JASPAR and CISBP databases. After examining the extent of collinearity in the set of motif scores, highly similar motifs are merged together to form a curated motif library. Motif scores for matches to the curated motif library are used to train a logistic regression model to predict whether or not a sequence corresponds to a peak or a background sequence. Model weights can be extracted to identify enriched and depleted motifs in peak sequences and motifs that are depleted.

3.3 Methods

3.3.1 ABTBA Model

ABTBA learns to distinguish target regions of interest, such as a set of TF binding regions T_n identified with ChIP-seq or a set of open chromatin regions determined using ATAC-seq, from genomic background regions (Fig. 3.1). For each target region, a background region (that contains no ambiguous or unmappable positions) is randomly selected from the genome such that the entire set of background regions B_n is matched for GC content with respect to the regions of interest. The genomic sequence of target and background of regions is retrieved and then the best match to each motif in a library M is calculated for each sequence. How well a sequence s (encoded as a one-hot vector) within the combined set, $S = T \cup B$, matches to a motif m of length K can be calculated

using the log-odds score, which is also known as the motif score (equation 3.1)[88].

$$l = \Sigma_{k=0}^{K} log_2(\frac{s_k \cdot m_k}{0.25}) \tag{3.1}$$

We calculate the motif score for each subsequence at position i within each sequence and take the best score considering the sequence in the forward and reverse complement orientation (equation 3.2).

$$\forall s \in S, \forall m \in Ml_{s,m} = max(max(s_{fwd,0:k}, \ldots, s_{fwd,i:i+k}), max(s_{rev,0:k}, \ldots, s_{rev,i:i+k})) \tag{3.2}$$

We take the $||M||$ by $||S||$ matrix of motif scores, $L_{m,s}$ and a corresponding label vector $Y \in 0,1$ of length $||S||$,indicating whether a sequence is a region of interest or a background region, to train a logistic regression model using the scikit-learn library[74]. Prior to training, we standardize the scores for each motif, which helps to reduce the effect of multiple-collinearity and ensures that the scores for each motif fall within a similar range (larger motifs have a larger possible motif score). Our logistic regression model learns to calculate a weighted sum (using learned weights W) over all the motif scores calculated for a sequence to score the probability that a sequence is a region of interest as opposed to random genomic background (equation 3.3). To discourage the model from selecting too many motifs, we use an L1 penalty term, giving the objective function in equation 3.4 (where λ is a scaling factor).

$$p(y_s = 1) = \frac{1}{1 + e^{-(\Sigma_{m \in M} w_m \cdot L_{m,s})}} \tag{3.3}$$

$$\min_W \lambda ||W||_1 + \Sigma_{s \in S} log(e^{-y_s \cdot W^T \cdot L_{m,s}} + 1) \tag{3.4}$$

By default, ABTBA models are trained using five-fold cross validation (80% of the data is used for training the model 20% for evaluating the model in each iteration of cross-validation) and the mean values across all iterations are reported.

3.3.2 Assessment of Multiple Collinearity

We can quantify the extent of multiple collinearity using the Variance Inflation Factor (VIF)[6]. For each motif m represented in the motif score matrix $L_{m,s}$, we regress L_m against the motif scores of all other motifs $L_{M-m,s}$. The VIF for motif m can then be calculated using the coefficient of determination R^2 (equation 3.5).

$$VIF = \frac{1}{1-R^2} \tag{3.5}$$

A VIF greater than 4 is considered problematic[6]. The VIF for each motif is calculated separately for each data set.

3.3.3 Motif Library and Curation

The JASPAR CORE non-redundant motif collection and the CISBP database forms the basis of the motif library used for training ABTBA models[38;102]. The similarity of each pair of motifs is calculated as the Pearson Correlation coefficient of the aligned motifs[66]. Motifs are aligned with the Smith Waterman algorithm and using the Pearson correlation to score each position within the alignment[87]. Smaller motifs are padded with background frequency values when they are aligned with larger motifs. We empirically determined that merging all motifs that have Pearson Correlation of 0.80 or greater would reduce the mean VIF of each motif across all datasets we analyzed to less than 4. Values for a merged motif are computed by averaging the value for a given position across all motifs that are merged together. We performed two successive rounds of motif merging to obtain a library of 299 motifs.

3.3.4 Extracting Motif Sets from ABTBA

Weights for each motif within in the ABTBA model can be directly interpreted. As we apply an L1 penalty when training our model, most motifs will have a weight close to 0. Motifs that have a positive weight are enriched in target regions T and motifs with a negative weight are depleted. Motifs with a weight of greater magnitude indicate greater enrichment in target regions. To assess the importance of a motif in the context of all other motifs within the model, we implement the

likelihood ratio test. In the likelihood ratio test, the importance of a motif is quantified as the difference in the performance of a perturbed model missing that motif versus the full model with all motifs, and then a significance value is assigned using the chi-squared test. By default, ABTBA performs the likelihood ratio test five times using separate iterations of cross validation.

3.3.5 ChIP-seq Data Processing

We downloaded all TF ChIP-seq data sets for GM12878, HepG2, and K562 human cell lines from the ENCODE data portal (www.encodeproject.org). We excluded all data sets that were flagged for a potential data quality issue. As we were interested in examining cell type specific properties of TF binding, we excluded all data sets that involved a genetically modified variant of a TF. In total we used TF binding sites (BED files using hg38 genomic coordinates) filtered according to the optimal IDR threshold calculated by ENCODE, from 363 ENCODE data sets for analysis with ABTBA.

3.3.6 ATAC-seq and RNA-seq Data Processing

We downloaded ATAC-seq and RNA-seq data from primary human hematopoietic lineage cell types from the Gene Expression Omnibus (accession GSE75384)[12]. To improve read mapping the ATAC-seq and RNA-seq we trimmed all reads to 40 bp using the HOMER command "homerTools trim"[29]. Trimmed ATAC-seq and RNA-seq reads were then mapped the hg38 genome build using Bowtie2 and STAR, respectively[16;52]. For RNA data, the HOMER analyzeRepeats.pl (using the parameters -count exons -condenseGenes -rpkm) was used to quantify gene expression in terms of Reads Per Kilobase Mapped (RPKM). For ATAC-seq data, putative open chromatin regions were called using the HOMER findPeaks command (with parameters -L 0 -C 0 -fdr 0.9 -style factor -size 200). High confidence open chromatin regions were identified using the Irreducible Discovery Rate (IDR) algorithm[60]. We then selected regions with IDR < 0.01 for analysis with ABTBA.

3.4 Results

3.4.1 Characterizing TF behavior with ABTBA

To test the performance of our model, we examined TF binding in three different cell lines GM12878 (lymphoblastoid), HepG2 (liver carcinoma) and K562 (chronic myeloid leukemia). For 363 ENCODE TF ChIP-seq experiments across the three cell lines, we trained ABTBA to discriminate TF binding sites from genomic background and compared the performance of ABTBA against the PWM with the best performance from the JASPAR or CISBP motif database (Fig. 3.2a). ABTBA generally had better performance (as measured by the area under the Receiver Operating characteristic curve (auROC). ABTBA's performance was highly correlated with the performance of a state-of-the-art machine learning model, gkmSVM (Fig. 3.3a), suggesting that ABTBA is broadly applicable for analyzing TF ChIP-seq data despite using considerably less parameters than gkmSVM ($\sim$300 parameters for ABTBA versus tens of thousands of parameters). For several TFs, the performance of the single PWM was similar to that of ABTBA ($p < 0.05$, Z-test on the ratio of ABTBA performance to PWM performance, blue points, Fig. 3.2a). We observed that the motifs that could be well-predicted by a single PWM were those that had a strong difference between the mean score for that PWM at the binding sites and the background regions; this suggests that TFs that can be accurately predicted with a single PWM recognize motifs that occurs less frequently throughout the genome. We found that these TFs with highly predictive PWMs did not necessarily recognize a more specific motif (with fewer degenerate positions). The degeneracy of a TF's best-matching PWM (quantified using information content) showed only a weak correlation to predictive performance (Fig. 3.3b).

For each TF, we counted the number of enriched and depleted motifs that pass a series of significance thresholds for the likelihood ratio test (Fig. 3.2c). At a threshold of 1e-10, we observed that on average genome-wide each TF interacts with genome-wide with dozens of motifs of significantly enriched motifs. Typically, only a few motifs are significantly depleted at a given TF's binding sites. To identify the number of motifs per sequence that are used by ABTBA to predict

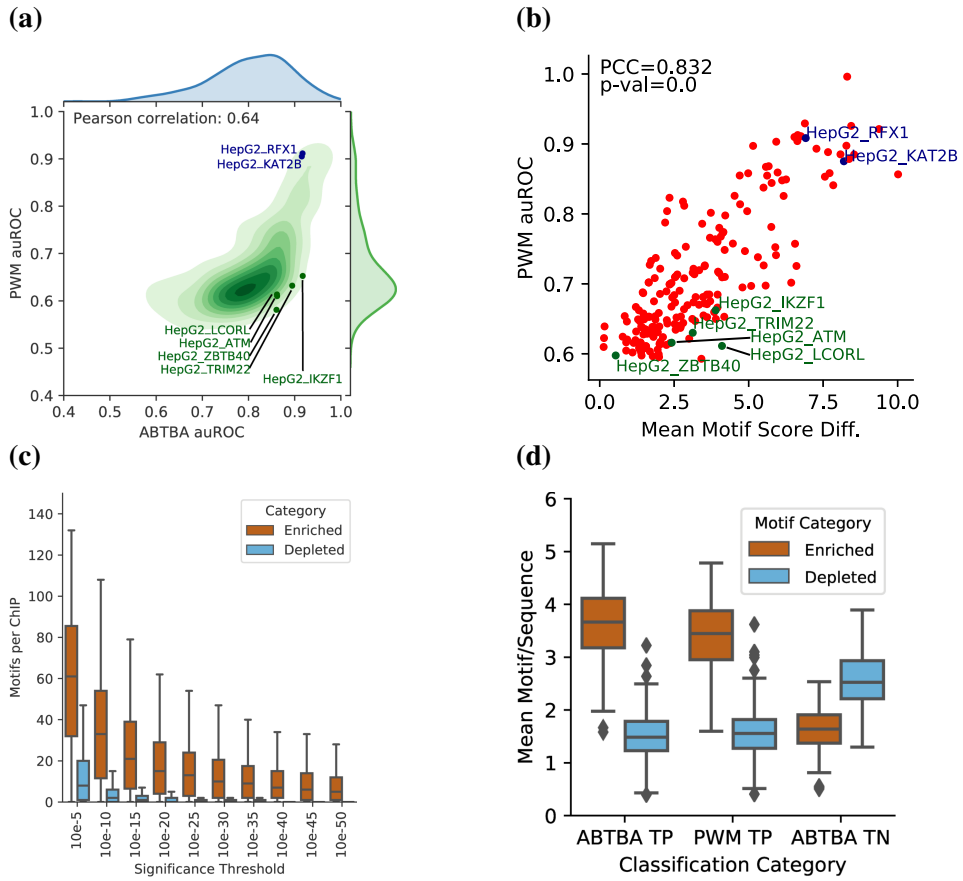


Figure 3.2. ABTBA learns to predict TF binding sites by learning ensembles of enriched and depleted motifs. **(a)** Performance of ABTBA versus the best matching PWM as measured by the area under the ROC curve for 363 TF ChIP-seq experiments. Higher color intensity indicates higher concentration of data points. Marginal distributions are indicated on the adjacent axes. Experiments where ABTBA and the PWM had highly similar performance (blue points) and where they had highly dissimilar performance (green) are indicated. **(b)** Distribution of the number of enriched and depleted motifs for each ChIP-seq experiment that pass a given significance threshold. **(c)** The mean difference between motif score of the best matching PWM found at peak sequences and that of background sequences versus the predictive performance of the PWM. ChIP-seq experiments where ABTBA and the PWM had highly similar (blue points) and highly dissimilar performance (green) are indicated. **(d)** Distribution of the mean number of significant motif terms observed at each sequence for peak sequences and background sequences correctly classified by ABTBA (ABTBA TP and ABTBA TN, respectively) and peak sequences correctly classified by the best matching PWM (PWM TP)

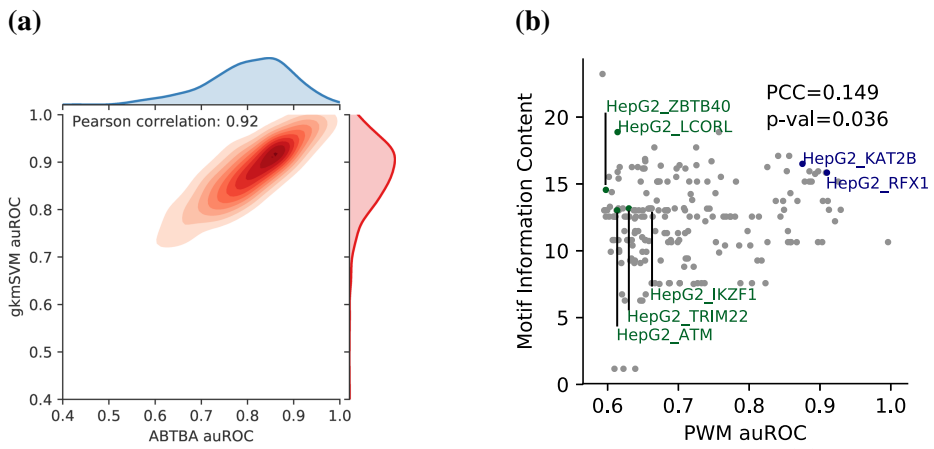


Figure 3.3. Supplementary characterization of ABTBA performance. **(a)** Performance of ABTBA versus the best matching gkmSVM as measured by the area under the ROC curve for 363 TF ChIP-seqs. Higher color intensity indicates higher concentration of data points. Marginal distributions are indicated on the adjacent axes. **(b)** The information content of best matching PWM versus the predictive performance of that PWM. ChIPs where ABTBA and the PWM had highly similar (blue points) and highly dissimilar performance (green) are indicated.

TF binding, we examined the individual terms in the model (each term is the product of a motif's weight and the log-odds score of the best-matching instance) at each sequence. We considered each term that was 3 standard deviations from the mean value of all terms (which are normally distributed and pass D'Agostino and Pearson's test for normality) to be significant, indicating that ABTBA is using the motif corresponding to the term to make its prediction. At TF binding sites correctly classified by ABTBA, ABTBA typically observes 3-4 enriched motifs per sequence, and less than 2 depleted sequences (ABTBA TP, Fig. 3.2d). In contrast, at background sequences correctly classified by ABTBA, less than 2 enriched motifs and 2 to 3 depleted sequences are observed by ABTBA (ABTBA TN, Fig. 3.2d). Notably, at TF binding sites that can be identified using the best matching PWM, ABTBA still observed 3-4 enriched motifs per binding site (PWM TP, Fig. 3.2d); this suggests that while some sites have a good match to a PWM, groups of TFs recognizing combinations of 3-4 motifs determine the collaborative binding of most TFs.

Next, we examined whether the motifs enriched at the binding sites of a TF varied in a cell type specific manner. Data was available for 9 TFs in GM12878, HepG2, and K562 cells. We identified enriched motifs that were highly significant in at least one cell type (likelihood ratio test,

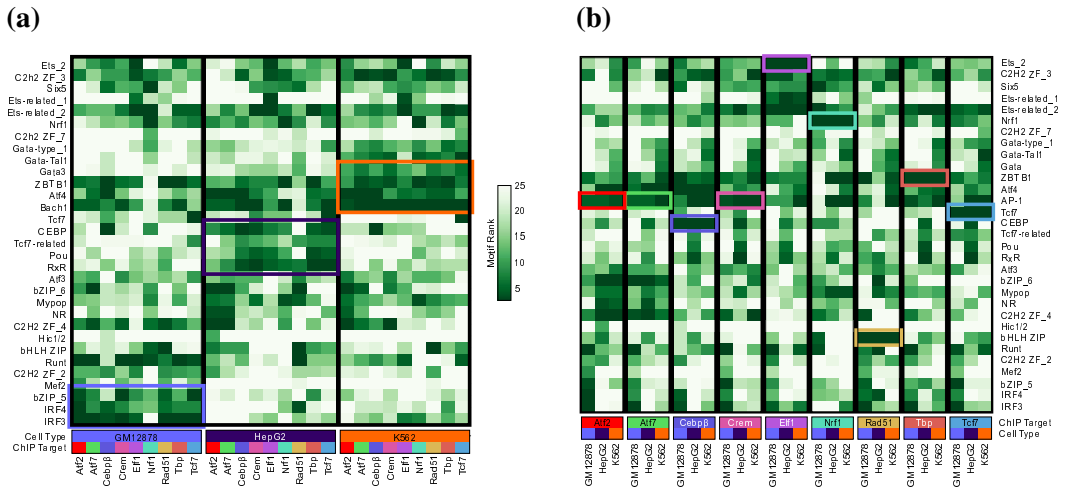


Figure 3.4. Comparison of motifs identified in ABTBA for TF ChIP-seqs performed in HepG2, GM12878, and K562 cell lines. **(a)** Heatmap of motif rankings calculated by ABTBA (darker hues indicate higher ranked motifs) grouped by cell type. Boxed regions indicate motifs common to TFs for a given cell type. **(b)** Heatmap of motif rankings calculated by ABTBA (darker hues indicate higher ranked motifs) grouped by TF. Boxed regions indicate the best matching motif for each TF.

$p < 1e-50$), and then ranked the identified motifs for each cell type according to the p-value. When we clustered the experiments by cell type, we observed that within each cell type, 3-4 motifs were highly ranked for all 9 TFs (boxed regions, Fig. 3.4a), which indicates the presence of TFs that are of general importance to a cell type. For example, POU and TCF7 motifs were generally enriched at TF binding cells in HepG2 cells. Both Oct4, which is required for HepG2 cells to maintain stemness, and TCF7, which is constitutively active due to beta-catenin activation, are known to be important activators in HepG2 cells[50;60;70]. When we clustered the experiments according to ChIP target, we find that the motif of the ChIP target is enriched for all experiments (boxed regions, Fig. 3.4b. However, additional motifs, presumably bound co-factors associated with each factor, varied considerably amongst the different cell lines. This suggests that while a TF consistently binds its target sequence, the local DNA environment, modulated by the TFs expressed in a given cell type, can affect the binding specificity of a TF.

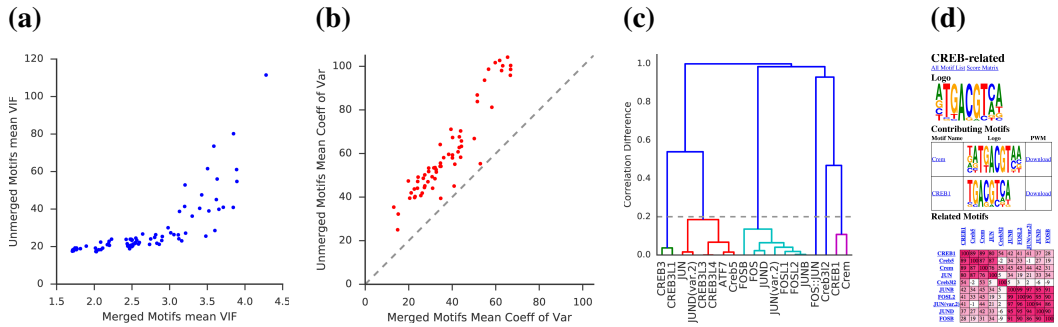


Figure 3.5. ABTBA generates a curated library of motifs that improves model stability. **(a)** The mean extent of multiple collinearity (measured using the VIF - Variance Inflation Factor) for motif scores calculated using the merged motif library curated by ABTBA for the peak sequences (and corresponding background sequences) of 91 HepG2 TF ChIP-seqs versus the mean extent of multiple collinearity for the unmerged motif library. **(b)** Mean stability of non-zero coefficients (Z-score $\geq$1), measured using the coefficient of variation, for 68 high performing ABTBA models (that achieved auROC $\geq$0.80) trained using the merged motif library curated by ABTBA versus models trained using the unmerged motif library **(c)** Example visualization of motif clustering produced by ABTBA. Colored clades indicate motifs that would be merged together. **(d)** Example HTML report produced by ABTBA that summarizes the motifs merged together (Crem and CREB1) as well as other similar motifs (scored by Pearson correlation)

3.4.2 Motif library curation using ABTBA

The presence of highly correlated features, multiple collinearity, can confound the interpretation of linear models such as ABTBA[6]. The extent to which a single feature contributes to multiple collinearity, which in our case corresponds to the presence of similar PWMs within our model, can be quantified using the VIF. ABTBA reduces multiple collinearity by aligning and merging highly similar motifs. The motif library used by ABTBA is formed using PWMs from the JASPAR and CISBP motif databases. To quantify ABTBA's effectiveness at reducing multiple collinearity, we calculated the mean VIF for every motif for each of the HepG2 datasets (91 TFs) using ABTBA's merged motif library and the unmerged set of PWMs from JASPAR and CISBP, observing a reduction in the mean VIF by more than an order of magnitude (Fig. 3.5a. This reduction in multiple collinearity corresponds to increased stability of the coefficients learned by the ABTBA model. For the 71 HepG2 datasets where we were able to fit ABTBAs with auROC $\geq$ 0.80, we quantified the stability of sizable coefficients (with absolute value one standard deviation greater than mean coefficient value) using the coefficient of variation and plotted the mean coefficient of

variation for each ABTBA model (lower values indicate greater stability, Fig. 3.5b).

While ABTBA uses a merged library formed from JASPAR and CISBP by default, ABTBA features several tools to help users incorporate motifs from other databases. ABTBA can produce dendrograms to visualize the similarity (measured using Pearson Correlation) of a set of motifs (Fig. 3.5c). Using the dendrogram, users can specify the level of similarity at which they would like to merge motifs together (colored clades that fall below the dotted line would be merged together, Fig. 3.5d). The results of motif merging in ABTBA are presented as a series of HTML reports that details the logo and PWM of the merged motif, the contributing motifs that were merged to form the merged motif, as well as related motifs (Fig. 3.5d).

3.4.3 Integration of ABTBA results and RNA-seq

Given ABTBA's ability to discern cell type specific ensembles of motifs using ChIP-seq, we applied ABTBA to ATAC-seq data (from GEO accession GSE75384) to assess whether ABTBA can identify important TFs for a cell type from open chromatin regions. Using HOMER and IDR, we identified between 53012 and 144164 open chromatin regions from each hematopoietic cell type. We trained ABTBA models using the open chromatin regions for each cell type, achieving an average performance of auROC 0.873 (Table 3.1). In addition to being able to distinguish open chromatin regions from each cell type from random genomic background, ABTBA was also able to distinguish the open chromatin regions unique to a cell type from regions that are already accessible in hematopoietic stem cells (Table 3.1). We selected the top 5 motifs identified using the likelihood ratio test as representative motifs for each cell type. Principal component analysis (PCA) of these top coefficients for the top motifs across the 12 cell types resolved the cells into clusters consistent with hematopoiesis lineages Fig. 3.6a). Additionally, the resulting clusters are qualitatively similar to those formed when performing PCA on quantile normalized gene expression (quantified as RPKM, genes expressed less than < 4 RPKM are not considered) for each cell type. In both analyses, erythroblasts clustered distinctly from leukocyte lineage cells (Fig. 3.6b). Notably, PCA of ABTBA motif weight effectively clustered terminally differentiated lymphoid lineage cells,

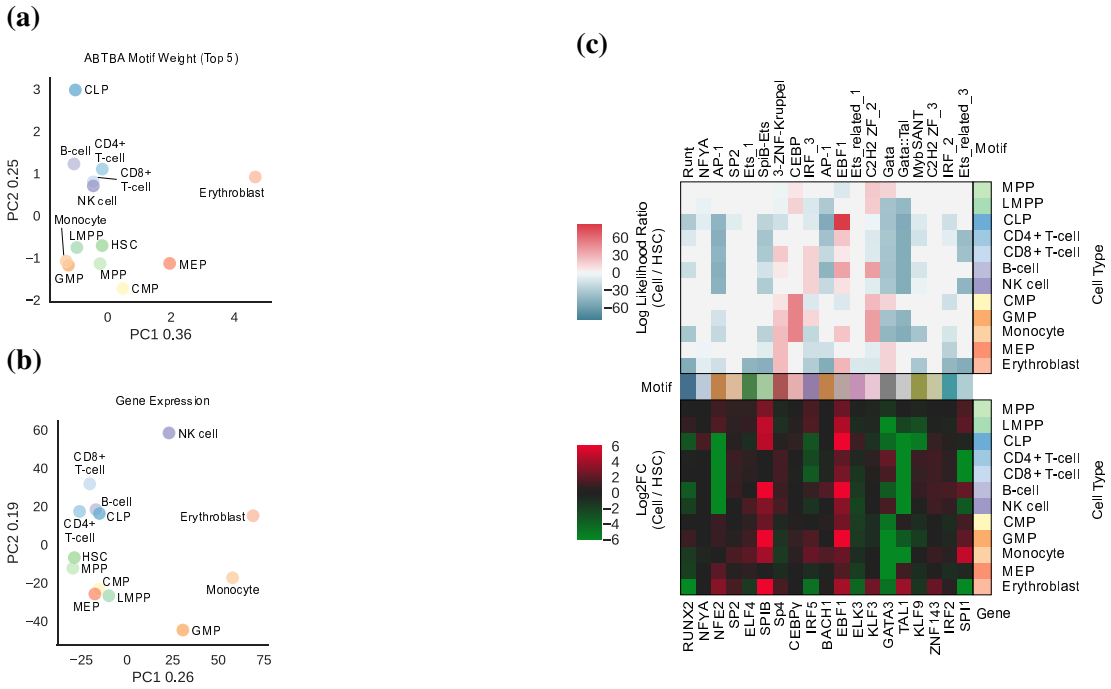


Figure 3.6. ABTBA Analysis identifies TFs in hematopoietic cell differentiation. **a, b** Principal component analysis of gene expression and ABTBA motif weights reveal similar clustering patterns. ATAC and RNA-seq generally separate lymphoid cells from myeloid and erythroid lineages. Gene expression is quantified using RPKM, transcripts with expression < 4RPKM are filtered and then results are quantile normalized. Motif weights were restricted to the five most significant motifs identified using the likelihood ratio test for each cell type. The amount of variance explained by each principal component is indicated in parentheses on the axes. **c** Each of the top 20 TF motifs identified by ABTBA were linked to a TF with the highest average expression in the hematopoietic cell dataset, revealing specific TFs known to be essential for differentiation of particular hematopoietic cell lineages, such as SPI1. Motif significance (upper heatmap) is expressed as a log-likelihood ratio relative to the motif's significance in hematopoietic stem cells. TF expression (lower plot) is expressed as log2 fold change in expression relative to TF expression in hematopoietic stem cells. For motifs with multiple expressed TFs, the TF with the highest average expression is displayed.

Table 3.1. ABTBA performance for predicting open chromatin regions in hematopoietic lineage cell types.

Cell Type	Background	auROC
hematopoietic Stem Cell	Genomic background	0.866
B-Cell	Genomic background	0.867
	HSC peaks	0.896
CD4 T-Cell	Genomic background	0.875
	HSC peaks	0.870
CD8 T-Cell	Genomic background	0.869
	HSC peaks	0.857
Common Lyphoid Progenitor	Genomic background	0.864
	HSC peaks	0.900
Common Myeloid Progenitor	Genomic background	0.868
	HSC peaks	0.890
Erythroblast	Genomic background	0.890
	HSC peaks	0.888
Granulocyte Myeloid Progenitor	Genomic background	0.875
	HSC peaks	0.875
Lymphoid-primed Multipotent Progenitor	Genomic background	0.869
	HSC peaks	0.850
Megakaryocyte-Erythoid Progenitor	Genomic background	0.869
	HSC peaks	0.904
Monocyte	Genomic background	0.897
	HSC peaks	0.864
Multipotent Progenitor	Genomic background	0.861
	HSC peaks	0.842
Natural Killer Cell	Genomic background	0.868
	HSC peaks	0.877

including CD4+ T cells, CD8+ T cells, B cells, and natural killer cells, separately from myeloid and erythroid lineage cells.

In order to match each enriched TF motif with an expressed TF responsible for cell type-specific open chromatin profiles, we integrated the enriched motifs identified by ABTBA with gene expression profiles measured using RNA-seq for each cell type. The 20 motifs we identified as being critical for hematopoiesis corresponded to 81 distinct TFs, but only 49 of these TFs were expressed at an RPKM > 4 in any hematopoietic cell type. We reasoned that the TFs driving hematopoesis that recognized these 20 critical motifs is likely to be highly expressed in at least one cell type. And so for each motif, we selected the TF that had the highest expression level when considering all

the cell types. Comparing ABTBA motif significance with RPKM expression identified TFs with expression profiles that support ABTBA motif significance values. For example, the EBF1 motif is highly significant in common lymphoid progenitor (CLP) cells, which express greater than 4-fold more EBF1 mRNA than any other hematopoietic cell line we examined, while also being necessary for CLP lineage function[64;70;109].

Integrating mRNA expression with ABTBA motif weights identified previously reported TFs that play important roles across several cell types. The Ets-related_factors_3_merged motif is highly significant in hematopoietic stem cells, B-cells, and monocyte lineage cells. Of the two TFs that bind the Ets-related_factors_3_merged motif, only Spi1 is expressed at an RPKM > 4 in any of the cell types examined. Spi1, also known as PU.1, has been well characterized as a lineage determining TF important for establishing both B-cell and monocyte cistromes[29]. ABTBA analysis also identified the EBF1 motif as an enriched motif in B-cells in addition to CLP cells. Both CLP and B-cells show a high level of expression of EBF1. Previous work has shown that EBF1 is required for B-cell differentiation from CLPs, suggesting that EBF1 is a necessary lineage determining transcription factor in B-cell development[64;70;109]. In erythroblasts, the EBF1 motif is enriched and EBF1 expression is high, possibly indicating a similar role in development of erythroblasts from CMPs.

Our analysis also identified motifs that are thought to be signal responsive TFs as opposed to constitutively expressed lineage factors. The interferon regulatory factor 5 motif was identified in our set of 20 top motifs and was a significant predictor of binding in B-cells and monocytes. IRF has a defined role in the differentiation of B-cells into plasma cells[49;62]. In the myeloid lineage, IRF5 expression increases as cells differentiate into monocytes. IRF4, 5, and 8 are also essential in cellular response to Toll-like receptor (TLR) signaling. Monocyte TLR signaling is a key pathway that mediates cellular responses to bacterial and viral compounds in their environment[90;108]. Taken together, these results demonstrate that integrating mRNA expression data with ABTBA results identifies biologically relevant TFs and motifs with cell specific functions.

3.5 Discussion

While ABTBA models allow for greater insight into TF binding, there remains other aspects of the genomic grammar read by TFs that we have not incorporated into our model. For example, previous studies have indicated that multiple instances of a TF's motif may be required for binding[24;63]. Multiplicity of a motif at a single locus can be captured by k-mer counts, potentially contributing to gkmSVM's improved performance over ABTBA. Our model also does not consider dependencies between positions within a motif, such as those calculated in Bayesian Markov models[86]. Supposing a similarity metric between two TF binding sites captured by Bayesian Markov models can be developed, the overall framework of ABTBA can be used for Bayesian Markov models as well. Recent reports have indicated that the spacing and arrangement of TF motifs can play an important role in the functional output of an enhancer[18]. Modeling spacing and arrangement of motifs may necessitate the use of more sophisticated models such as neural networks, which have been demonstrated to be highly flexible function approximates in a wide set of problem domains.

The ABTBA model describes the binding specificity of a TF as a set of motifs that include the binding motif of the TF as well as the motifs of potential collaborative binding partners. Exploring the motifs of potential collaborating TFs provide an opportunity for the integration of orthogonal data types in the analysis of ATAC-seq and ChIP-seq datasets. We demonstrate the utility of combining RNA-seq data with an ABTBA analysis of ATAC-seq data by identifying previously validated TFs acting in specific hematopoietic lineages such as the TF PU.1, which is a known lineage determining factor in B-cells and monocytes[29]. Beyond RNA-seq, other data types could be readily combined with ABTBA to yield biological insight. For example, whole cell or nuclear mass spectrometry (MS) data could be used to integrate protein quantity information with the motif weights calculated by ABTBA.

Biological network data could also be used to select motifs identified by ABTBA with a high likelihood of physical or biological interactions. For example, affinity purification mass

spectrometry (AP-MS) data could be readily combined with ChIP-seq ABTBA analysis to predict site specific larger scale transcriptional complexes.

TFs with a high ABTBA score that also exhibit physical interactions identified by AP-MS could be acting collaboratively as part of a larger activation or inactivation complex. Such a hypothesis could be readily tested using co-immunoprecipitation studies or by knocking out the collaborative partner.

The genome contains potentially millions of copies of each DNA motif, yet ChIP-seq experiments typically find that only tens of thousands of these motifs are actually bound by a TF. It has been suggested that TFs bind in combination as a means to control and direct binding. Here, we find that open target chromatin by ATAC and ChIP -seq predicts 3-4 quality DNA motifs in each distinct region, suggesting 3-3 TFs are necessary to activate a region of chromatin(Fig. 3.2d). Further, we predict dozens of DNA motifs as significant in each cell line (Fig. 3.2c). Though some DNA motifs are predicted with much higher significance, indicating ABTBA found these at a higher frequency, the identification of dozens of significant motifs by ABTBA suggests a large number of possible transcriptional complexes. Allowing for a large number of possible transcriptional complexes based on 3-4 DNA binding TFs from a panel of 20-30 would provide a cell with a large repertoire of available transcriptional units to generate responses to diverse signals. Future work will involve designing machine learning networks to determine the groups of collaborating TFs which are able to form transcriptional units within a cell and the frequency at which they occur.

3.6 Conclusion

With the aim of balancing the simplicity of a PWM and the performance of machine learning models, we developed a machine learning tool that draws from a programmatically curated motif library to describe TF binding specificity as a set of enriched and depleted motifs. We applied our tool to an extensive collection of TF datasets and by inspecting the behavior of our model, estimated that each TF interacts with dozens of other TFs genome wide and with 3-4 other TFs at a single locus. These results corroborates previous findings that the combinatorial binding of TFs forms

the basis of a genomic grammar in a cell type specific manner[18;28;29;31]. As our tool can model the combinatorial interaction of TFs, we demonstrated that ABTBA can be applied to ATAC-seq data sets to retrieve combinations of motifs that can be cross referenced with RNA-seq data to identify TFs that establish the open chromatin landscape in a variety of cell types. Our tool, ABTBA, is built to be compatible with future advances in motif databases and can be readily applied to study motif interactions at other genomic regions of interest.

3.7 Acknowledgements

We thank L. Van Ael for assistance with manuscript preparation and Z. Ouyang for technical assistance. These studies were supported by NIH grants DK091183, CA17390 and GM085764 and Leducq Transatlantic Network grant 16CVD01 to CKG. GJF was supported by a Canadian Institute of Health Research Postdoctoral Fellowship, FME-135475. HB was supported by NIH grant 2T32DK007202-42A1.

Chapter 3, in part, has been submitted for publication. Tao, J., Bennett, H., Fonseca. G.J., Shen, Z., Benner, C., Glass, C.K. A method for describing transcription factor binding specificity as a set of DNA motifs. The dissertation author was the primary investigator and author of this study.

Chapter 4

Learning Composition Rules for Mammalian Circuits with Neural Attention

4.1 Abstract

The expression of each gene in mammalian cells is controlled by regulatory sequences called enhancers. Regulatory logic encoded at enhancers is interpreted by transcription factors (TFs), which recognize individual 'words' in the genome. Here we describe a neural network with an attention mechanism that learns to discriminate enhancers from random genomic sequences as well as predict the activity of an enhancers. We distill the parameters learned by our model to identify combinations of TF target sequences that signal activation of an enhancer in response to specific cellular stimuli.

4.2 Introduction

Regulation of gene expression in mammalian cells is mediated in part by hundreds of sequence specific TFs that bind to their individual binding motifs at enhancers, which are distal regulatory elements located thousands of basepairs away from a gene[31]. The binding of TFs at enhancers mediates the recruitment of machinery necessary for transcription such as RNA polymerase. Prior studies have suggested two classes of TFs: 1) lineage determining TFs (LDTFs) and and 2) signal dependent TFs (SDTFs)[28]. LDTFs bind to cell type specific enhancers while SDTFs bind at enhancers bound by LDTFs in response to a cellular stimuli, resulting in cell type

specific activation of an enhancer in response to stimuli[28]. These studies suggest that context specific gene expression in a cell type is genetically encoded by combinations of TF binding motifs at millions of enhancers scattered throughout the genome[11].

Given the evidence that TFs act collaboratively to activate enhancers, it follows that individual TF motifs are poor predictors of whether or not a sequence is an enhancer. The biological activity of an enhancer may depend on the composition of TF motifs - arrangement and spacing between TF motifs, as well as the degeneracy of each motif[18]. And so, we endeavored to teach an attentive neural network (ANN) to distinguish accessible enhancer elements from background genomic sequences as well as to predict enhancer activation (measured using ChIP-seq targeting H3K27Ac) in macrophage cells, a cell of the innate immune system. Neural networks have been previously applied to predict enhancers[1;46;76]. However, most of these previous models report only individual motifs that are enriched and do not describe how TF motifs interact. Our ANN uses a neural mechanism, which is at the heart of current models for modeling sequences of words in natural language processing applications such as language translation and sentiment analysis[10;97]. In our ANN, neural attention allows the model to focus on the TF motifs that are functional at an enhancer and ignore dozens of other nonfunctional motifs (Figure4.1). By extracting information learned by our network, we can identify combinations of TF motifs that signal the activation of an enhancer in response to cellular stimuli as well as learn how these TF motifs interact (Figure4.1).

4.3 Methods

4.3.1 Data Processing

Raw sequencing data (fastq files) were mapped to the mm10 build of the mouse genome using version 2.2.9 of Bowtie2[52]. Open chromatin regions were identified with the HOMER (version 4.8.3) findPeaks command using replicate ATAC-seq experiments[29]. Open chromatin regions were scored for consistency between replicate experiments using the Irreproducible Discovery Rate (IDR)[61] and only peaks with IDR < 0.01 were used for analysis. IDR scores were calculated using version 2.02 of the IDR program. The activity at an open chromatin region was quantifed as the

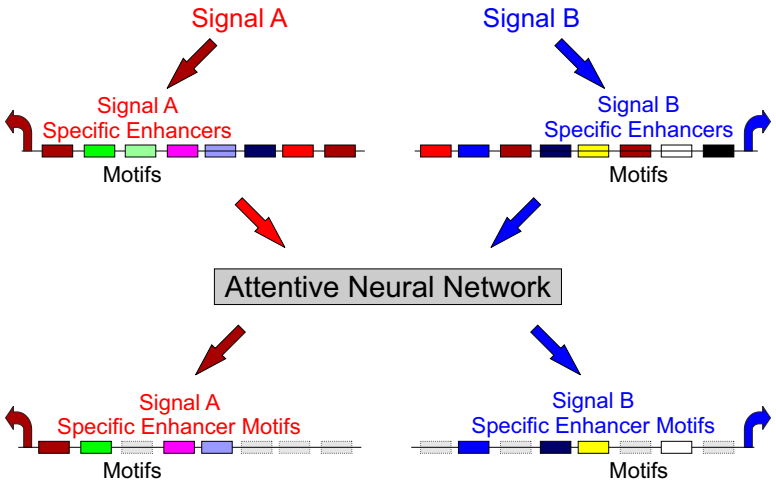


Figure 4.1. Attentive neural network learns to ignore non-functional motifs (faded gray boxes) thereby revealing TF motifs that control activation in response to signal A and B respectively

number of H3K27Ac ChIP-seq reads that fell within a 1000 basepair window centered at each open chromatin region. Differentially acetylated regions were identified using EdgeR (version 3.7) using biological replicates and defined as sites with False Disovery Rate < 0.05 and a fold change between two treatment conditions greater than two[80].

4.3.2 Model Architecture

The architecture of our ANN model is shown in (Figure4.2). To learn motifs recognized by TFs, our ANN model applies a 1-dimensional convolution, $conv_m$ over 4 channels to the input sequence, s, encoded as a one hot vector[1;46]. We uses 150 convolution kernels in total. To learn relationships between motifs, we eschew recurrent layers, which require many parameters that are hard to interpret, and use neural dot product self attention only[76;97]. The rectified convolution output, $R = rect(conv_m(s)$, which quantifies how well each position in a sequence matches to a motif, is then fed to the attention layer. Using the notation of Vaswani et al, we project R using 3 separate sets of weights, forming RW^Q, RW^K, RW^V. The product $A = (RW^Q)(RW^K)^T$ forms an attention matrix, which can be used to identify interactions between positions within a sequence. The attention matrix also highlights which positions of the sequence are being used by the model to make a prediction. The output of the attention layer, $(RW^Q)(RW^K)^T(RW^V)$ is a weighted sequence

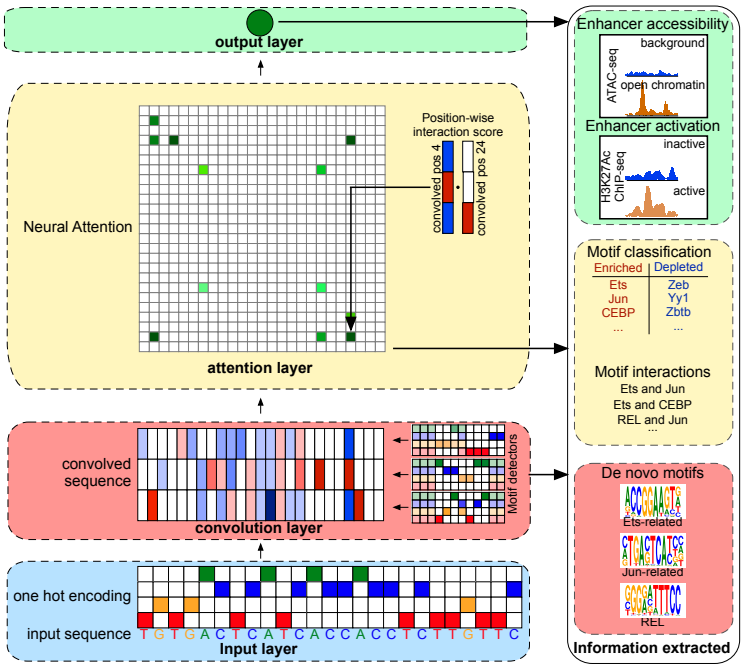


Figure 4.2. Overview of attentive neural network model. A sequence is encoded as a binary one-hot vector and then convolved to quantify how well each subsequence matches to a DNA motif. The motifs present at each pair of positions are used to calculate an attention matrix, which weights the importance of each position. The output of the convolutional layer is weighted using the attention matrix and the fed to a single dense neuron which then makes a prediction

of motifs where higher magnitude weight indicate that a position is more important; a positive weight would indicate that a motif is positively correlated with the prediction target and conversely, a negative weight would indicate that a motif is negatively correlated with the prediction target. The output of the attention layer is then fed to the output layer, which quantifies the number of motifs positively and negatively correlated with the prediction target. The structure of the output layer will vary according to the prediction target. For binary classification (eg. open chromatin versus genomic background, we use a single dense neuron with a sigmoid activation function. For regression (eg. predicting the level of H3K27Ac) we would use a dense neuron with a Relu activation function instead. For multi-target tasks, the prediction layer would use a dense neuron for each prediction target. Models were implemented version 2.2.0 of Keras and version 1.7.0 of TensorFlow using the Python (version 3.6.1) API.

4.3.3 Motif Library Model Variant

Although, our ANN model uses relatively few parameters in comparison to a comparable neural network that uses recurrent neurons, our ANN model would still require roughly ten thousand data points (Parms, 4.3). In response to a stimuli, only thousands of loci may become activated in a macrophage. And so, we created a variant of our ANN model that can use a mix known motifs drawn from public databases[38;102] and de novo motifs learned using convolution kernels. To use existing motifs we initialize a convolution kernel for each known motif and set the weights of that kernel to the log-transformed values of the position weight matrix for that known motif in the same fashion as[4]. We used a motif library formed from motifs from the JASPAR and CISBP motif libraries where highly similar motifs have been merged together[19]. As each de novo motif costs four times the size of the motif to learn in terms of parameters, the amount of parameters saved can be considerable for a model that considers many motifs (Table 4.3).

4.3.4 Model Training

Each model for predicting open chromatin was trained using 20,000 open chromatin regions that are accessible in each treatment condition and a 20,000 GC matched genomic background regions. We employ five-fold cross validation when training all models, and so 80% of the data is employed for training the model and 20% of the data is used for evaluating the performance of the model. Models were trained by optimizing performance of the model using cross entropy as the loss function and using the ADAM optimizer[48].

Models trained to predict the level of H3K27Ac were trained using all open chromatin sites detected in any treatment condition. The log_2 transform of the number of H3727Ac reads at each open chromatin region. These models were trained using five-fold cross-validation. Models for predicting H3K27Ac levels were trained by optimizing performance of the model using mean absolute error as the loss function and using the RMSProp optimizer.

4.3.5 Identification of subgroups of open chromatin regions

Subgroups of open chromatin regions were identified using the output of the attention layer of our ANN model for each open chromatin region. The output of the attention layer is a weighted version of the output of the convolutional layer, a matrix that gives motif scores for each motif (or convolution kernel) considered by the model at each position within the open chromatin region. In the attention layer, important positions are assigned greater weight and motif scores at that position would increase in magnitude. Conversely, positions that are unimportant will have motif scores closer to zero after weighting. The sign of the attention weight indicates whether a motif is enriched (positive weights) or depleted at open chromatin sites. For each motif, we took value with the greatest absolute value across all positions within a sequence, giving us an attended score matrix with length equal to the number of open chromatin regions and width equal to the number of motifs/kernels. We then calculate the principal components of this score matrix and use the principal components to cluster the open chromatin regions. The principal components, which are linear combinations the attended scores for each motif, can also be used to identify motifs that co-occur. We performed clustering using the k nearest neighbor algorithm implemented in the Seurat R package[7;58;107]. Results of the clustering is then visualized using t-Distributed Stochastic Neighbor Embedding (t-SNE). We used five principal components and perplexity of 100 for t-SNE visualization.

4.3.6 Software and Code Availability

Source code for our ANN model is available at: github.com/jenhantao/genomic_grammar.

4.4 Results

4.4.1 Profiling macrophage chromatin landscape

We profiled the open chromatin landscape of bone marrow derived macrophages (BMDMs) from C57Bl6/J mice using ATAC-seq[12]. To assay for active regions of the genome, we performed ChIP-seq targeting H3K27 acetylation (H3K27Ac). We performed these experiments in resting

(vehicle treated) and macrophages stimulated with either Kdo2 lipid A (KLA, one hour treatment, a model for M1 polarized macrophages) or IL4 (24 hour treatment, a model for M2 polarized macropahges). These experiments identified tens of thousands of open chromatin regions (Table 4.1) in each treatment context as well as thousands of deferentially acetylated regions (Table 4.2).

Table 4.1. Number of open chromatin regions detected using ATAC-seq

Treatment	Open Chromatin Regions
Vehicle	31553
KLA-1h	21783
IL4-24h	33581

Table 4.2. Number of differentially acetylated regions detected using H3K27Ac ChIP-seq

Treatment	Up-regulated	Down-regulated
KLA vs. Veh	7311	8264
IL4 vs. Veh	1121	297
KLA vs IL4	4869	6386

4.4.2 Open chromatin prediction

We began by training a model to discriminate open chromatin regions present in each condition from GC-content matched genomic background regions. To assess the performance of our ANN model architecture, we compared the performance of our ANN model against the current state of the art, a convolutional network. We trained our ANN model and an implementation of DeepBind, a previously described convolutional network[1], to distinguish active enhancers from random genomic sequences. The performance of our ANN model exceeded that of the convolutional model, in terms of model accuracy and precision (Figure4.3). To ensure that the improvement in the performance of our ANN model is not due to the greater number of free parameters (Figure4.3), we also trained a large convolutional network (with 54 convolution kernels and 108 dense neurons versus 16 convolution kernels and 32 neurons in the original model). We also quantified a variant of our ANN model where the convolution kernels (Figure 4.2) are initialized using a library of known motifs (see Model Design section). We found that the performance of the model variant that

only used the motif library achieved similar performance to the large convolutional neural network while using substantially less parameters (LargeConv, Lib, Table 4.3. We found that the difference in performance between the library model and the large convolutional network can be narrowed by introducing just 16 convolutional kernels that learn de novo motifs to the motif library based variant of our ANN model (LargeConv, Att+Lib, Table 4.3). Overall, models using the attention mechanism were able to achieve better performance, suggesting that the attention mechanism is capable of extracting additional useful information beyond what is learned by the convolutional neural networks[93;94]. We believe that the the neural network with an attention mechanism is able to learn a sparse representation of each sequence where uninformative motifs are ignored (Fig. 4.1). An example of an attention matrix (used by our ANN model to weight raw motif scores) is shown on the left in Fig. 4.3 adjacent to the unweighted motif scores on the right.

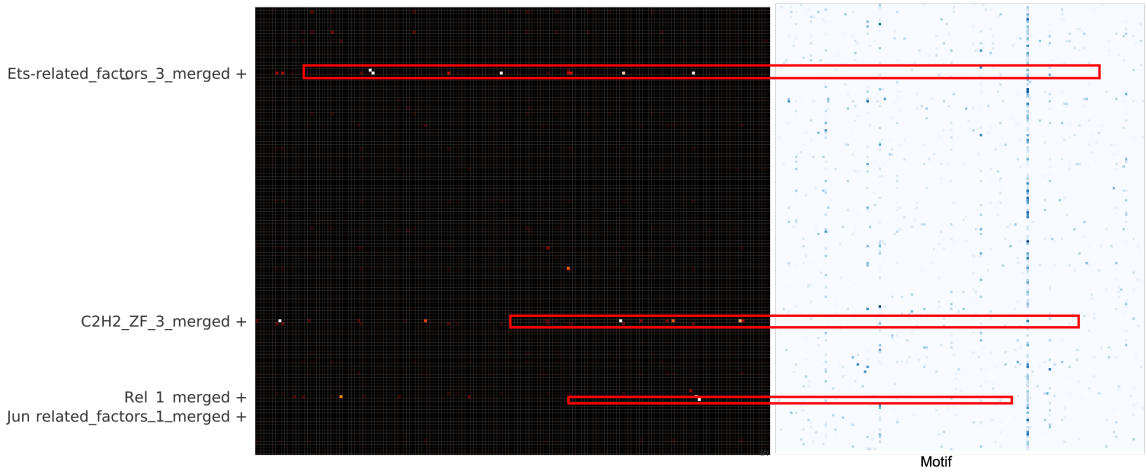


Figure 4.3. Example of attention matrix calculated for an accessible open chromatin region in KLA treated macrophages. The attention matrix is on the left and lighter values indicate that the model is focusing on a particular position. Unweighted motif scores are shown on the right, and darker hues indicate better matches to each motif (columns). The most attended/important motif at each position is indicated on the left side of the panel. Red boxes are drawn to indicate that the attention values at each position corresponds to a set of motif scores for each motif.

Table 4.3. Comparative performance of various models for predicting open chromatin. Performance metrics (n=3), accuracy and precision, of our attentive neural network (Att.), a convolutional network (Conv), a large convolutional network (LargeConv), a variant of our attentive neural network using a library of 299 motifs and 16 convolution kernels for learning 16 de novo motifs (Att+Lib), and a variant of our attentive neural network that uses only a motif library of 299 motifs (Lib) are shown for 3 treatment conditions (Veh, KLA, IL4)

			Model				
			Att	Conv	LargeConv	Att+Lib	Lib
		Params	12750	2129	14657	4590	2990
Tx	Veh	Acc.	0.896	0.870	0.880	0.895	0.888
		Prec.	0.900	0.877	0.877	0.907	0.889
		auROC	0.965	0.944	0.954	0.960	0.956
	KLA	Acc.	0.863	0.833	0.846	0.852	0.843
		Prec.	0.862	0.810	0.850	0.852	0.852
		auROC	0.936	0.914	0.927	0.928	0.919
	IL4	Acc.	0.888	0.878	0.885	0.888	0.881
		Prec.	0.899	0.865	0.880	0.881	0.886
		auROC	0.957	0.949	0.958	0.959	0.954

4.4.3 Subtypes of open chromatin regions

To begin examining the motifs that drive the performance of the classifier, we adapted several clustering techniques typically used for single cell RNA-seq analysis, principal component analysis and t-SNE, to analyze the attended motif scores calculated by our ANN model. The attended motif scores calculated by our ANN model is structurally similar to single cell RNA-seq data in that both types of data are sparse and are high dimensional. Instead of finding clusters of cells that express different sets of genes, we hope to identify clusters of open chromatin sites that enrich for different sets of motifs. To facilitate ease of analysis, we analyzed the variant of our ANN model that used the motif library only (Lib in Table 4.3 so that we did not have to match each convolution kernel to a motif as was done in previous studies[1;46;76]. Following previously described single cell RNA-seq analysis procedures, we began by performing principal components analysis on the attended motif scores calculated by our ANN model (see methods) and then used these principal components for visualization using t-SNE[93;94]. To determine the number of principal components to use, we examined the amount of variance explained by each principal component in each treatment condition and observed that the amount of variance explained by each principal

component dropped considerably after the fifth principal component 4.4. And so, we used t-SNE to visualize the first five principle components. Visually, we observe that t-SNE separates the open chromatin sites detected in each condition into distinct clusters (Fig. 4.5c, 4.6c, 4.7c). Cluster assignments were determined using the k-nearest algorithm[7].

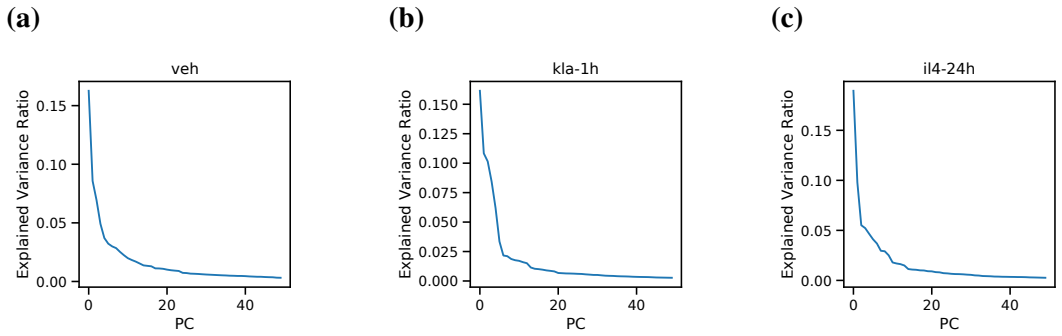


Figure 4.4. Amount of variance explained by each principal component for attended motif scores. Individual panels are shown for each treatment: vehicle (4.4a), KLA-1h (4.4b), and IL4-24h (4.4c)

Visually, the clusters of open chromatin regions formed using the attended motif scores learned by our ANN model were more distinct than clusters formed using un-weighted motif scores (Fig. 4.5a, 4.6a, 4.7a). Additionally, we calculated clusters formed using the output of the dense layer of the large convolutional network 4.5b, 4.6b, 4.7b). Each neuron in the dense layer of the convolutional network calculates linear combination of the convolutional layer and is essentially a combination of motif scores. The effectiveness of clustering where the ground truth clusters are unknown (we do not actually know how the open chromatin regions cluster together biologically) can be quantified using the Silhouette Coefficient. The Silhouette Coefficient is defined using b, the mean distance between each data point and the points in the nearest cluster and a, the mean distance between each data point and the points in the cluster, as: $\frac{b-a}{max(a,b)}$. Data that is clustered with distinct clusters that have little overlap between clusters would have Silhouette Coefficient approaching 1. Clustered data that has clusters that overlap would have a Silhouette Coefficient approaching 0. Negative Silhouette Coefficients indicate poor clustering where many data points have been potentially assigned to the incorrect cluster. Open chromatin regions clustered using attended motif scores calculated by our ANN model had higher Silhouette Coefficients than when

the same regions were calculated using motif scores or the output of the dense layer of the large convolutional network (Table 4.4. These results suggests that our ANN model is better able to identify subgroups of open chromatin regions.

Table 4.4. Comparison of cluster structure using Silhouette Coefficient. Higher coefficients indicate better clustering.

		Representation		
		Motif Score	Attended motif Score	LargeConv
Tx	Veh	0.12	0.21	0.19
	KLA	0.13	0.26	0.18
	IL4	0.13	0.19	0.17

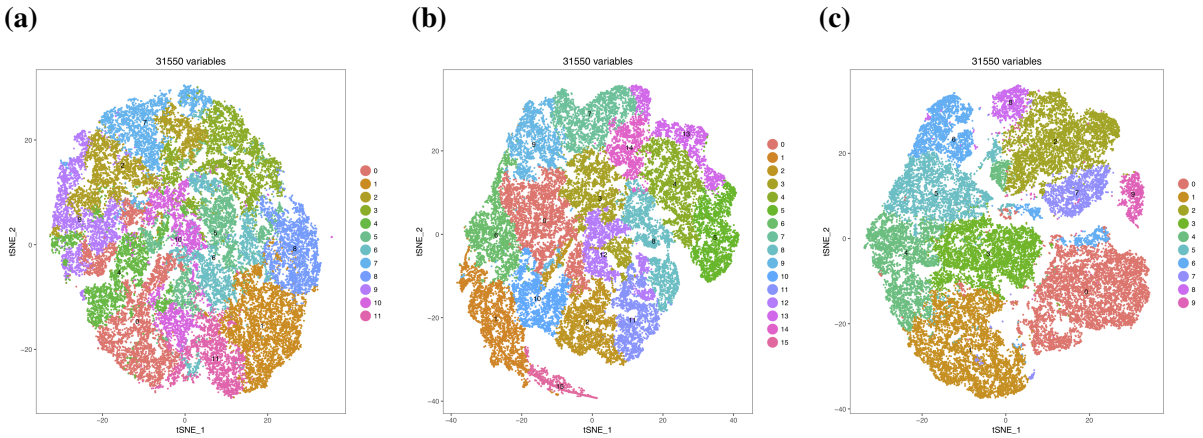


Figure 4.5. t-SNE visualization of sequence representations learned by various models for open chromatin regions accessible in resting macrophages (vehicle treatment). Panel 4.5a gives t-SNE clustering results using unweighted motif scores (299 motifs). Panel 4.5b gives t-SNE clustering results for the sequence representation calculated by the large convolutional neural network (output of dense layer). Panel 4.5c gives t-SNE clustering results using the sequence representation learned by our neural network with an attention mechanism (output of attention layer).

To examine the motifs that are enriched within each cluster, we used the Wilcoxon Rank Sum test to compare the attended motif scores in one cluster versus the rest of the clusters. As most of the open chromatin sites in vehicle treated, KLA treated and IL4 treated macrophages overlap, we show just the top 25 motifs that are significantly enriched in at least one cluster for vehicle treated macrophages (Fig. 4.8). We observed 10 clusters that each enriched for a distinct set of motifs, suggesting that different combinations of TFs are responsible for establishing the open chromatin landscape in macrophages.

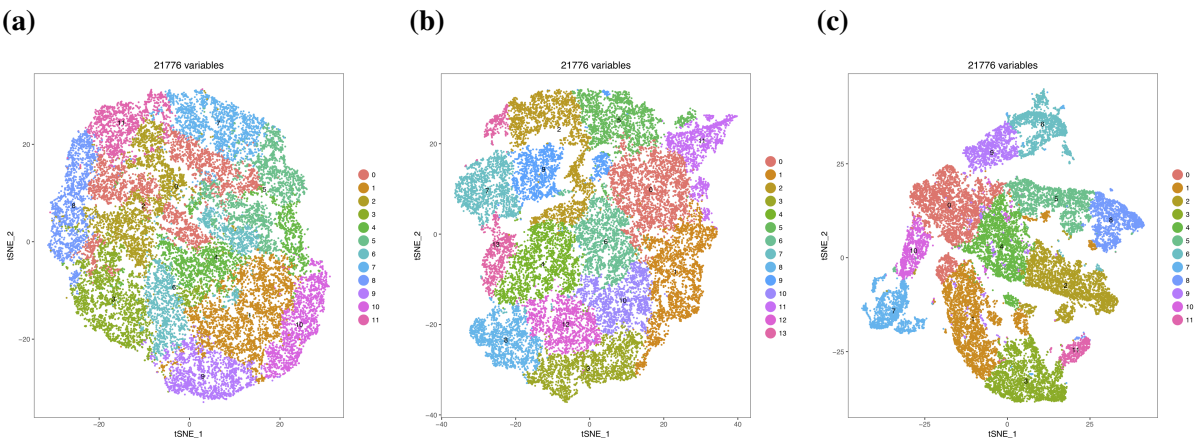


Figure 4.6. t-SNE visualization of sequence representations learned by various models for open chromatin regions accessible in KLA treated macrophages. Panel 4.6a gives t-SNE clustering results using unweighted motif scores (299 motifs). Panel 4.6b gives t-SNE clustering results for the sequence representation calculated by the large convolutional neural network (output of dense layer). Panel 4.6c gives t-SNE clustering results using the sequence representation learned by our neural network with an attention mechanism (output of attention layer).

4.4.4 Enhancer Activity Prediction

Encouraged by the performance of our ANN model at predicting open chromatin, we applied our ANN model to predicting the level of H3K27Ac under a given treatment at all open chromatin regions detected in any treatment condition. Again, we observe that our ANN model (Att, Table 4.5) performs better than both the convolutional network (Conv) and the large convolutional network (LargeConv). Variants of our ANN model that were initialized using a motif library performed similarly to the large convolutional network despite using considerably less parameters (Table 4.5). Notably, each of the models performed far better at discriminating open chromatin from genomic background (Table 4.3. This could be due to the fact that while chromatin accessibility is largely determined by TF binding, which can be indirectly observed via motifs, there are additional biological processes that govern enhancer/promoter activation. The global conformation of the genome and trans interactions between different regulatory elements may play a role[15;105]. Additionally, protein modifications such as phosphorylation may affect the activity of both TFs and cellular machinery recruited by TFs[43].

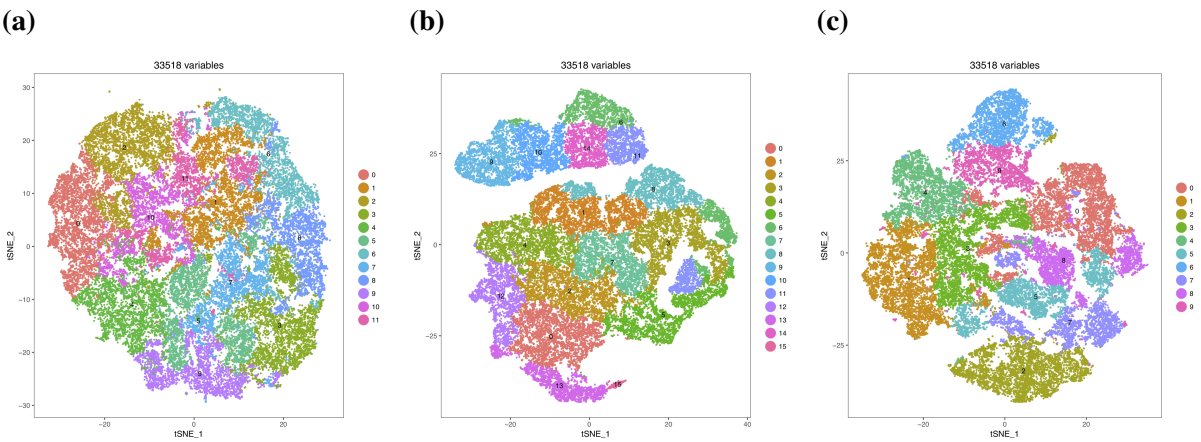


Figure 4.7. t-SNE visualization of sequence representations learned by various models for open chromatin regions accessible in IL4 treated macrophages. Panel 4.7a gives t-SNE clustering results using unweighted motif scores (299 motifs). Panel 4.7b gives t-SNE clustering results for the sequence representation calculated by the large convolutional neural network (output of dense layer). Panel 4.7c gives t-SNE clustering results using the sequence representation learned by our neural network with an attention mechanism (output of attention layer).

4.4.5 Subtypes of differentially acetylated regions

Considering the limited performance of our ANN model (and that of other models), we decided to specifically examine open chromatin regions with differential acetylation after treatment with KLA-1h or IL4-24h instead of all open chromatin sites. We reasoned that examining activated and de-activated regions would give us the best chance to observe signal specific motifs. t-SNE representations created using attended motif scores calculated by our ANN model of differentially acetylated regions after KLA-1h treatment (Fig. 4.9a and IL4-24h treatment (Fig. 4.9b.

And in Figure 4.10, we show the enriched motifs within each cluster of differentially acetylated regions. In differentially aceltylated regions in KLA treated macrophages, we of course find that several clusters are defined by the enrichment of the Rel motif, which is the recognition target of p65 (Fig. 4.10a). Additionally we find that the clusters are also stratified according to the enrichment of Ets and IRF motifs which have been previously shown to play a role in the macrophage TLR4 response[41;42].

In IL4 stimulated macrophages, we find that several clusters enrich for STAT motifs; STAT factors play an important role in the macrophage's IL4 response[14]. Interestingly, two distinct STAT

Figure 4.8. Motifs enriched in each open chromatin cluster in Vehicle treated macrophages. Color intensities give the negative log10 transformed p-value, indicating greater enrichment of a particular motif in a given cluster. Scores are calculated separately for both orientations of a motif (indicated using +/-).

motifs were enriched in separate clusters (STAT 1 merged and Stat6). Additional motifs that that define different clusters of differentially acetylated regions in IL4 treatment included IRF and bZip factor motifs (Fig. 4.10b).

In KLA treated macrophages, we found that the clusters of differentially acetylated regions formed using the attended motif scores also showed differences in activity before and after treatment with KLA (Fig. 4.11a. Clusters 6, 8, 9, 10, and 11 appear to contain most of the regions that are deactivated in response to KLA. Clusters, 6, 8, 10, and 11 show a depletion of the Rel/p65 motif (Fig 4.10a). Interestingly, cluster 9 does show some enrichment for the Rel motif, which may

Table 4.5. Comparative performance of various models for predicting enhancer activity. Performance metrics (n=3) for our attentive neural network (Att.), a convolutional network (Conv), a large convolutional network (LargeConv), a variant of our attentive neural network using a library of 299 motifs and 16 convolution kernels for learning 16 de novo motifs (Att+Lib), and a variant of our attentive neural network that uses only a motif library of 299 motifs (Lib) are shown for 3 treatment conditions (Veh, KLA, IL4). Performance was quantified as the Pearson Correlation of the predicted level of H3K27Ac and the actual level of H3K27Ac at all chromatin regions detected in any of the three treatment conditions.

		Model				
		Att	Conv	LargeConv	Att+Lib	Lib
	Params	12750	2129	14657	4590	2990
Tx	Veh	0.403	0.339	0.368	0.398	0.384
	KLA	0.517	0.493	0.505	0.497	0.484
	IL4	0.397	0.366	0.372	0.387	0.373

suggest that the Rel motif is not the sole determinant of activation in response to KLA. Changes in acetylation activity in IL4 treated macrophages did not stratify clearly across the different clusters (Fig. 4.10b).

4.5 Discussion

In this study, we propose the a novel neural network architecture for analyzing regulatory elements located throughout the genome. Our neural network with an attention mechanism is capable of discriminating context specific (different treatment conditions) open chromatin regions from genomic background at state of the art levels. Additionally, our neural network shows some improvement over existing convolutional networks at predicting the level of H3K27Ac, a proxy for promoter/enhancer activity. Additionally, we demonstrate that techniques for identifying groups of functionally distinct cell groups using single cell RNA-seq can be adapted to identify groups of open chromatin regions as well as motifs that are differentially enriched in each of these groups/clusters.

While this study began with the intention of showing that multiple combinations of motifs are important for mediating signal response in macrophages, we currently have not achieved the level of results to make that claim. There are however, several useful ideas that can be taken from this work in the current state. First, we believe that it is an important problem to distinguish functional

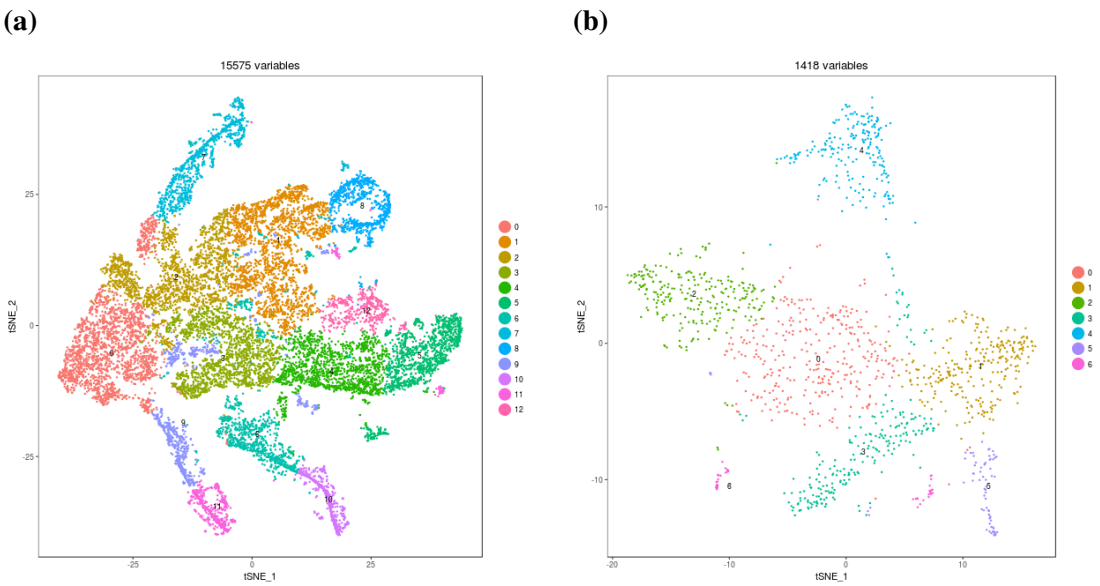


Figure 4.9. t-SNE visualization of sequence representations learned by our ANN model for differentially acetylated regions. Panel 4.9a gives t-SNE visualization of differentially acetylated regions in KLA treated macrophages using the sequence representation learned by our neural network with an attention mechanism (output of attention layer). Panel 4.9b gives t-SNE visualization of differentially acetylated regions in KLA treated macrophage

motifs from motifs that are present (Fig. 4.1). In this study, we have tried to tie function to motifs by examining motifs in several signaling contexts. We believe the neural attention mechanism we have described is a potential solution. Second, given growing evidence for the collaborative binding of TFs [19;28;29;65], we believe it is important to look for groups of motifs, and not individual motifs.

4.6 Future Work

4.6.1 Data Analysis

While we are encouraged by the performance of our ANN model, we believe that additional efforts need to invested in extracting insights from the parameters learned by the neural network and refining our analysis approach.

First, there is remaining work to be done with respect to the design of our model. The model described by Vaswani et al used a position encoding, which we did not implement. Previous studies have noted that motif positioning can have a periodicity that is potentially due to the turn

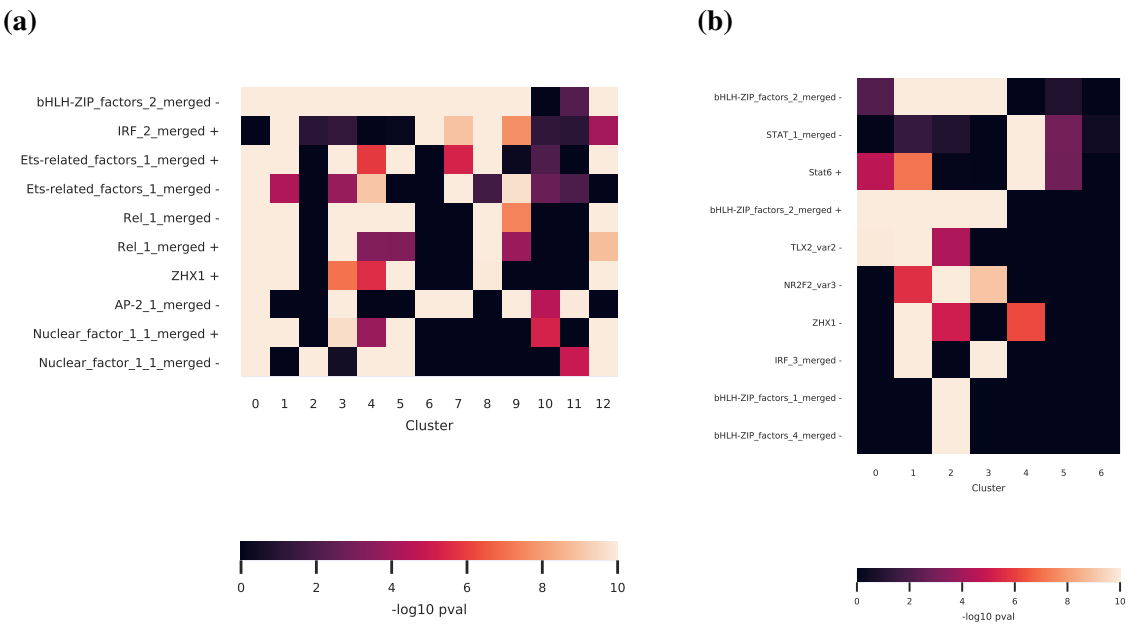


Figure 4.10. Motifs enriched in cluster of differentially acetylated regions with respect to Vehicle treatment in KLA-1h and IL4-24h treated macrophages. Color intensities give the negative log10 transformed p-value, indicating greater enrichment of a particular motif in a given cluster. Scores are calculated separately for both orientations of a motif (indicated using +/-).

of the DNA helix; the sin based positional encoding used by Vaswani could be used to model this periodicity. The model used by Vaswani et al featured multiple attention matrices, which can be used to model the behavior of a transcription factor in multiple signaling contexts. We have previously demonstrated that a transcription factor's binding partner may change depending on the cell type as well as the signaling context (see Chapter 2 and 3). The original intuition behind using data from multiple signal contexts stems from the observation that Vaswani et al constructed a model that excelled at coupling a dense representation of a sequence (a sentence in one language) with a corresponding dense representation of that sequence (a sequence in a different language). In our application of an attention only model, we have a dense representation of a sequence, a matrix of motif scores, and our original intuition was that we would require a corresponding dense vector of activity states to identify functional motifs using this approach. And so, a multiclass learning task (predicting the response of multiple signals simultaneously) may be more effective for tackling the challenge of trying to predict enhancer/promoter activity.

Second, the tools we applied downstream of our ANN model have a number of parameters

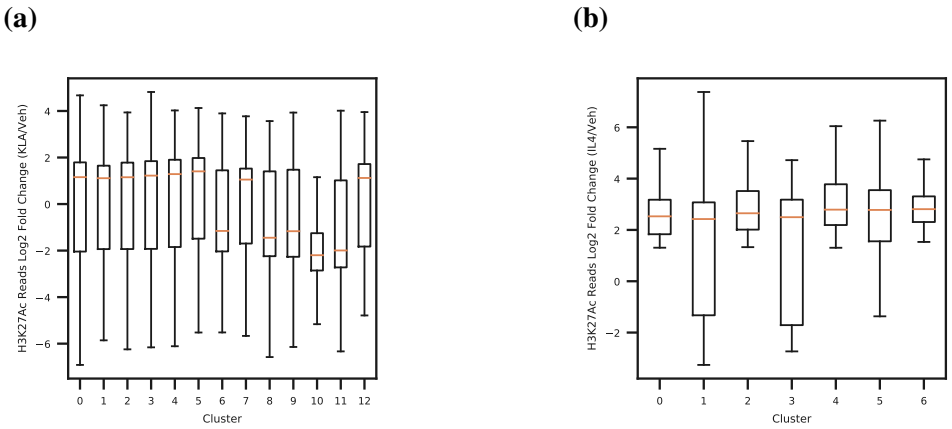


Figure 4.11. Change in H3K27Ac signal at differentially acetylated regions clustered using attended motif scores. The vertical axis gives the log2 transformed fold change in H3K27Ac signal between KLA and vehicle treated macrophages (Fig. 4.11a) and between IL4 and vehicle treated macrophages (Fig. 4.11b)

that require tuning. The k nearest neighbor algorithm has a resolution parameter that controls the number of clusters identified by the algorithm. In this study, we did not rigorously calculate the expected number of clusters; this can be done by clustering the open chromatin regions according to their activity state across many treatment contexts. Additionally, the t-SNE algorithm has a perplexity parameter that approximates the number of neighbors a data point is expected to have. Typical values for this perplexity parameters ranges from 5-50. However, we anticipate that hundreds if not thousands of open chromatin regions may respond to a signal similarly and have similar sequence compositions. The choice of the wilcoxon rank sum test, a nonparametric, for identifying enriched motifs is a suitable and conservative choice, but there are likely better options that can be selected upon more careful examination of the data.

4.6.2 Experimental Validation

Given the preliminary nature of this work, caution needs to be taken when considering follow up experiments. To demonstrate that a TF is associated with the activity of a specific cluster, one could perform knockdown of a specific TF and perform ChIP-seq to determine if the activity of that cluster is specifically affected. knockdown can be effectively performed using CRISPR-Cas9 technology. Alternatively, one could try to over express a TF using an expression vector and attempt

to specifically activate a cluster of enhancers. To identify the correct TF to target, further data describing the availability of TFs need to be integrated into this study because multiple TFs can recognize the same motif.

Assuming that sequene content is the primary determinant of enhancer activity (it may not be), we can also use motifs enriched in each cluster to design artificial transcriptional units that are cell type and signal specific. To do so would require additional analysis examining the spacing and arrangement of the motifs. Such artificial transcriptional units may have clinically relevant applications.For example, this would allow us to use well characterized, general delivery systems such as adenovirus, which target cells indiscriminately. We can specify treatment activity by using a synthetic target specific transcriptional unit designed using information from our attentive neural network that drives the expression of a therapeutic transgene. As an example, a transcriptional unit specific to cancer driving the expression of a immune reactive transgene would allow for immune specific targeting of the cancer.

4.7 Acknowledgements

We thank L. Van Ael for assistance with manuscript preparation and Z. Ouyang for technical assistance. These studies were supported by NIH grants DK091183, CA17390 and GM085764 and Leducq Transatlantic Network grant 16CVD01 to CKG. GJF was supported by a Canadian Institute of Health Research Postdoctoral Fellowship, FME-135475. HB was supported by NIH grant 2T32DK007202-42A1.

Chapters 4, in part, will be submitted for publication. Tao, J.*, Fonseca, G.J.*, Duttke, S.H., Hoeksema, M.A., Shen, Z.,Bennett, H., Benner, C., Glass, C.K. Learning composition rules for macrophage enhancers with neural attention. (* These authors contributed equally to this work) The dissertation author was one of the primary investigators and authors of this paper.

Chapter 5

Conclusions

Here we describe machine learning models for interpreting genomic sequences that are grounded by biological realities that were uncovered prior to this study. Large scale epigenomic studies suggest the presence hundreds of thousands of regulatory elements - enhancers and promoters - in mammalian genomes[79]. Each cell type selects tens of thousands of these elements, which play important roles in determining the identity and function of each cell type[2;53]. Previous studies established that the activity of promoters and enhancers depends in part on the binding of sequence-specific transcription factors to their target sequence[31;84]. In mammals, there are on the order of hundreds of transcription factors, which are organized into families that can share highly conserved DNA binding domains[104]. The target binding motif of the majority of transcription factors has been characterized using a combination of in vitro and in vivo assays and these DNA motifs have been deposited in a variety of databases[29;38;102]. Given these prior results, the aim of this study was to leverage machine learning to learn how these components of transcriptional regulation fit together to enable cell type and context specific activity at enhancers and promoters.

Chapter 2 of this work detailed a generalization of the collaborative hierarchical model, a previously described model for transcription factor binding and enhancer promoter activation[28]. The collaborative hierarchical model suggests that transcription factors bind to cell type specific regions by targeting sites already bound by simple combinations of lineage determining transcription factors[5;23;28;29;42;54;95;103]. As we could not explain differences in the binding profiles of the AP-1 transcription factor family using simple combinations of motifs, we constructed a machine learning

model that jointly considers hundreds of motifs simultaneously to predict transcription factor binding. This model resulted in the realization that a single transcription factor may interact with dozens of other transcription factors genome wide. Additionally, we identified two classes of collaborative motifs: 1) highly ranked TFs that are likely to be important for cellular identity[28;29;42;69;91;95] and 2) moderately ranked TFs that specify the binding of individual AP-1 family members. In chapter 3, we extended this result to many transcription factors profiled by the ENCODE Consortium. Overall, these two studies provide evidence that collaborative binding of TFs allow a single DNA motif to be used in a wide variety of contexts; this may be a general principle for how transcriptional regulation is encoded by the genome.

In chapter 4, we applied lessons from chapter 2 and 3 to construct a neural network with an attention mechanism that allows us to better model combinations of transcription factor motifs. The capability of this neural network is qualitatively distinct from previous models as it can account for multiple occurrences of a motif at a single locus as well as leverage a mixture of known motifs (drawn from public databases) and de novo motifs that are learned. We applied this neural network to study macrophage signal response. Leveraging the attention mechanism, which computes high resolution maps of how DNA motifs interact with one another, the model learns to ignore instances of a motif that may not be functional while focusing on instances of motifs that co-occur with other important motifs. Genome wide, there are millions of copies of each DNA motif, most of which are not bound by a transcription factor. Generalizing this observation to a single locus, we would expect that most occurrences of motifs at a single locus are nonfunctional and so it is important for a model to learn to ignore motifs. Using this approach, we were able to identify subtypes of regulatory elements that activate in response to a particular signal that are each identified by distinct combinations of transcription factors. The results from these studies demonstrate that combinations of transcription factor motifs are not only important for determining transcription factor binding (as described in Chapters 2 and 3) but also for determining the activation of regulatory elements themselves.

In these studies, we have largely focused on combinations of transcription factor motifs,

which play a role in directing transcription factor binding and enhancer activation. However, there are additional aspects of a how regulatory information is encoded at promoters and enhancers that we have not explored. Transcriptional regulation may also be influenced by the spacing between motifs and the specific arrangement of motifs[18]. As efforts to model these additional aspects of the genomic grammar progress, which may require increasingly complex model architectures, equal efforts need to be invested in developing methods to extract information from these complex machine learning models. Models described in this study do not explicitly phenomena such as protein to protein interactions, post translational modifications on transcription factors, or interactions between multiple loci (eg. promoter enhancer looping); integrating this information could also potentially result in more complex models. The proper application of machine learning for interpreting the genome ought to be a two step process. After constructing a model with good performance, one also needs to explain - in a biological fashion - how the the model has achieved good performance. In order to test a hypothesis generated by a computational model while fully accounting for potentially confounding biological processes, a biological experiment must be conducted as the complexity of a living cell vastly exceeds that of any current computational model.And so, further studies of the genome and transcriptional regulation will continue depend on the joint efforts of both experimental scientists and computational scientists as much as these studies have.

Bibliography

[1] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*, 33(8):831–838, 2015.

[2] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, Evgenia Ntini, Erik Arner, Eivind Valen, Kang Li, Lucia Schwarzfischer, Dagmar Glatz, Johanna Raithel, Berit Lilje, Nicolas Rapin, Frederik Otzen Bagger, Mette Jørgensen, Peter Refsing Andersen, Nicolas Bertin, Owen Rackham, A Maxwell Burroughs, J Kenneth Baillie, Yuri Ishizu, Yuri Shimizu, Erina Furuhata, Shiori Maeda, Yutaka Negishi, Christopher J Mungall, Terrence F Meehan, Timo Lassmann, Masayoshi Itoh, Hideya Kawaji, Naoto Kondo, Jun Kawai, Andreas Lennartsson, Carsten O Daub, Peter Heutink, David A Hume, Torben Heick Jensen, Harukazu Suzuki, Yoshihide Hayashizaki, Ferenc Müller, Alistair R R Forrest, Piero Carninci, Michael Rehli, and Albin Sandelin. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, mar 2014.

[3] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk. Diversity and Complexity in DNA Recognition by Transcription Factors. *Science*, 324(5935):1720–1723, 2009.

[4] Nicholas E Banovich, Yang I Li, Anil Raj, Michelle C Ward, Peyton Greenside, Diego Calderon, Po Yuan Tung, Jonathan E Burnett, Marsha Myrthil, Samantha M. Thomas, Courtney K Burrows, Irene Gallego Romero, Bryan J Pavlovic, Anshul Kundaje, Jonathan K Pritchard, and Yoav Gilad. Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Research*, 28(1):122–131, jan 2018.

[5] Iros Barozzi, Marta Simonatto, Silvia Bonifacio, Lin Yang, Remo Rohs, Serena Ghisletti, and Gioacchino Natoli. Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA Features of Mammalian Enhancers. *Molecular Cell*, may 2014.

[6] David A. Belsley, Edwin Kuh, and Roy E. Welsch. *Regression Diagnostics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, jun 1980.

[7] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Inte-

grating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, 2018.

[8] Huihui Chen and Zhengfan Jiang. The essential adaptors of innate immune signaling. *Protein & cell*, 4(1):27–39, jan 2013.

[9] Lin Chen, J. N. Mark Glover, Patrick G. Hogan, Anjana Rao, and Stephen C. Harrison. Structure of the DNA-binding domains from NFAT, Fos and Jun bound specifically to DNA. *Nature*, 392(6671):42–48, mar 1998.

[10] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long Short-Term Memory-Networks for Machine Reading. *arXiv*, jan 2016.

[11] The ENCODE Project Consortium, Ian Dunham, Anshul Kundaje, Shelley F Aldred, Patrick J Collins, Carrie a Davis, Francis Doyle, Charles B Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R Lajoie, Stephen G Landt, Burn-Kyu Bum-Kyu Lee, Florencia Pauli, Kate R Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M Simon, Lingyun Song, Nathan D Trinklein, Robert C Altshuler, Ewan Birney, James B Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Jason Ernst, Terrence S Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C Hardison, Robert S Harris, Javier Herrero, Michael M Hoffman, Sowmya Iyer, Manolis Kelllis, Pouya Kheradpour, Timo Lassman, Qunhua Li, Xinying Lin, Georgi K Marinov, Angelika Merkel, Ali Mortazavi, Stephen C J Stephanie L Parker, Timothy E Reddy, Joel Rozowsky, Felix Schlesinger, Robert E Thurman, Jie Wang, Lucas D Ward, Troy W Whitfield, Steven P Wilder, Weisheng Wu, Hualin S Xi, Kevin Y Yip, Jiali Zhuang, Bradley E Bernstein, Eric D Green, Chris Gunter, Michael Snyder, Michael J Pazin, Rebecca F Lowdon, Laura a L Dillon, Leslie B Adams, Caroline J Kelly, Julia Zhang, Judith R Wexler, Peter J Good, Elise a Feingold, Gregory E Crawford, Job Dekker, Laura Elinitski, Peggy J Farnham, Morgan C Giddings, Thomas R Gingeras, Roderic Guigó, Timothy J Tomothy J Hubbard, Manolis Kellis, W James Kent, Jason D Lieb, Elliott H Margulies, Richard M Myers, John a Starnatoyannopoulos, Scott a Tennebaum, Zhiping Weng, Kevin P White, Barbara Wold, Yanbao Yu, John Wrobel, Brian a Risk, Harsha P Gunawardena, Heather C Kuiper, Christopher W Maier, Ling Xie, Xian Chen, Tarjei S Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J Coyne, Timothy Durham, Manching Ku, Thanh Truong, Matthew L Eaton, Alex Dobin, Timo Lassmann, Andrea Tanzer, Julien Lagarde, Wei Lin, Chenghai Xue, Brian a Williams, Chris Zaleski, Maik Röder, Felix Kokocinski, Rehab F Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakrabortty, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J Luo, Eddie Park, Jonathan B Preall, Kimberly Presaud, Paolo Ribeca, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandu, Lorain Schaeffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Hao Wang, Yoshihide

Hayashizaki, Alexandre Reymond, Stylianos E Antonarakis, Gregory J Hannon, Yijun Ruan, Piero Carninci, Cricket a Sloan, Katrina Learned, Venkat S Malladi, Matthew C Wong, Galt P Barber, Melissa S Cline, Timothy R Dreszer, Steven G Heitner, Donna Karolchik, Vaness M Kirkup, Laurence R Meyer, Jeffrey C Long, Morgan Maddren, Brian J Raney, Linda L Grasfeder, Paul G Giresi, Anna Battenhouse, Nathan C Sheffield, Kimberly a Showers, Darin London, Akshay a Bhinge, Christopher Shestak, Matthew R Schaner, Seul Ki Kim, Zhuzhu Zhancheng Zhengdong Zhang, Piotr a Mieczkowski, Joanna O Mieczkowska, Zheng Liu, Ryan M McDaniell, Yunyun Ni, Naim U Rashid, Min Jae Kim, Sheera Adar, Tianyuan Wang, Deborah Winter, Damian Keefe, Vishwanath R Iyer, Kljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Jason Gertz, Jost Vielmetter, E Christopher Partridge, Katherine E Varley, Clarke Gasper, Anita Bansal, Shirley Pepke, Preti Jain, Henry Amrhein, Kevin M Bowling, Michael Anaya, Marie K Cross, Michael a Muratet, Kimberly M Newberry, Kenneth McCue, Amy S Nesmith, Katherine I Fisher-Aylor, Barbara Pusey, Gilberto DeSalvo, Suganthi Sreeram Balasubramanian, Nicholas S Davis, Sarah K Meadows, Tracy Eggleston, J Scott Newberry, Shawn E Levy, Devin M Absher, Wing H Wong, Matthew J Blow, Axel Visel, Len a Pennachio, Laura Elnitski, Hanna M Petrykowska, Alexej Abyzov, Bronwen Aken, Daniel Barrell, Gemma Barson, Andrew Berry, Alexandra Bignell, Veronika Boychenko, Govanni Bussotti, Claire Davidson, Gloria Despacio-Reyes, Mark Diekhans, Iakes Ezkurdia, Adam Frankish, James Gilbert, Jose Manuel Gonzalez, Ed Griffiths, Rachel Harte, David a Hendrix, Toby Hunt, Irwin Jungreis, Mike Kay, Ekta Khurana, Jing Leng, Michael F Lin, Jane Loveland, Zhi Lu, Deepa Manthravadi, Marco Mariotti, Jonathan Mudge, Gaurab Mukherjee, Cedric Notredame, Baikang Pei, Jose Manuel Rodriguez, Gary Saunders, Andrea Sboner, Stephen Searle, Cristina Sisu, Catherine Snow, Charlie Steward, Electra Tapanari, Michael L Tress, Marijke J van Baren, Stefan Washieti, Laurens Wilming, Amonida Zadissa, Zhang Zhengdong, Michael Brent, David Haussler, Alfonso Valencia, Alexandre Raymond, Nick Addleman, Roger P Alexander, Raymond K Auerbach, Keith Bettinger, Nitin Bhardwaj, Alan P Boyle, Alina R Cao, Philip Cayting, Alexandra Charos, Yong Cheng, Catharine Eastman, Ghia Euskirchen, Joseph D Fleming, Fabian Grubert, Lukas Habegger, Manoj Hariharan, Arif Harmanci, Susma Iyenger, Victor X Jin, Konrad J Karczewski, Maya Kasowski, Phil Lacroute, Hugo Lam, Nathan Larnarre-Vincent, Jin Lian, Marianne Lindahl-Allen, Renqiang Min, Benoit Miotto, Hannah Monahan, Zarmik Moqtaderi, Xinmeng J Mu, Henriette O'Geen, Zhengqing Ouyang, Dorrelyn Patacsil, Debasish Raha, Lucia Ramirez, Brian Reed, Minyi Shi, Teri Slifer, Heather Witt, Linfeng Wu, Xiaoqin Xu, Koon-Kiu Yan, Xinqiong Yang, Kevin Struhl, Sherman M Weissman, Scott a Tenebaum, Luiz O Penalva, Subhradip Karmakar, Raj R Bhanvadia, Alina Choudhury, Marc Domanus, Lijia Ma, Jennifer Moran, Alec Victorsen, Thomas Auer, Lazaro Centarin, Michael Eichenlaub, Franziska Gruhl, Stephan Heerman, Burkard Hoeckendorf, Daigo Inoue, Tanja Kellner, Stephan Kirchmaier, Claudia Mueller, Robert Reinhardt, Lea Schertel, Stephanie Schneider, Rebecca Sinn, Beate Wittbrodt, Jochen Wittbrodt, Gaurav Jain, Gayathri Balasundaram, Daniel L Bates, Rachel Byron, Theresa K Canfield, Morgan J Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Kavita Garg, Erica Gist, R Scott Hansen, Lisa Boatman, Eric Haugen, Richard Humbert, Audra K Johnson, Ericka M Johnson, Tattayana M Kutyavin, Kristin Lee, Dimitra Lotakis, Matthew T Maurano, Shane J Neph, Fiedencio V Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Eric Rynes, Minerva E Sanchez, Richard S Sandstrom, Anthony O Shafer, Andrew B Stergachis, Sean Thomas, Benjamin Vernot, Jeff Vierstra,

Shinny Vong, Molly a Weaver, Yongqi Yan, Miaohua Zhang, Joshua a Akey, Michael Bender, Michael O Dorschner, Mark Groudine, Michael J MacCoss, Patrick Navas, George Stamatoyannopoulos, John a Stamatoyannopoulos, Kathryn Beal, Alvis Brazma, Paul Flicek, Nathan Johnson, Margus Lukk, Nicholas M Luscombe, Daniel Sobral, Juan M Vaquerizas, Serafim Batzoglou, Arend Sidow, Nadine Hussami, Sofia Kyriazopoulou-Panagiotopoulou, Max W Libbrecht, Marc a Schaub, Webb Miller, Peter J Bickel, Balazs Banfai, Nathan P Boley, Haiyan Huang, Jingyi Jessica Li, William Stafford Noble, Jeffrey a Bilmes, Orion J Buske, Avinash O Sahu, Peter V Kharchenko, Peter J Park, Dannon Baker, James Taylor, and Lucas Lochovsky. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, sep 2012.

[12] M. Ryan Corces, Jason D. Buenrostro, Beijing Wu, Peyton G. Greenside, Steven M. Chan, Julie L. Koenig, Michael P. Snyder, Jonathan K. Pritchard, Anshul Kundaje, William J. Greenleaf, Ravindra Majeti, and Howard Y. Chang. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics*, 48(10):1193–1203, 2016.

[13] L. J. Core, J. J. Waterfall, and J. T. Lis. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science*, 322(5909):1845–1848, dec 2008.

[14] Zsolt Czimmerer, Bence Daniel, Attila Horvath, Dominik Rückerl, Gergely Nagy, Mate Kiss, Matthew Peloquin, Marietta M. Budai, Ixchelt Cuaranta-Monroy, Zoltan Simandi, Laszlo Steiner, Bela Nagy, Szilard Poliska, Csaba Banko, Zsolt Bacso, Ira G. Schulman, Sascha Sauer, Jean-Francois Deleuze, Judith E. Allen, Szilvia Benko, and Laszlo Nagy. The Transcription Factor STAT6 Mediates Direct Repression of Inflammatory Enhancers and Limits Activation of Alternatively Polarized Macrophages. *Immunity*, 48(1):75–90.e6, 2018.

[15] Yarui Diao, Rongxin Fang, Bin Li, Zhipeng Meng, Juntao Yu, Yunjiang Qiu, Kimberly C Lin, Hui Huang, Tristin Liu, Ryan J Marina, Inkyung Jung, Yin Shen, Kun-Liang Guan, and Bing Ren. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nature Methods*, 14:629, apr 2017.

[16] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[17] Dzmitry Bahdana, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation By Jointly Learning To Align and Translate. *Iclr 2015*, pages 1–15, 2014.

[18] Emma K. Farley, Katrina M. Olson, Wei Zhang, Daniel S. Rokhsar, and Michael S. Levine. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proceedings of the National Academy of Sciences*, 113(23):6508–6513, jun 2016.

[19] Gregory J Fonseca, Jenhan Tao, Emma M Westin, Sascha H Duttke, Nathaneal J Spann,

Tobias Strid, Zeyang Shen, Joshua D Stender, Verena M Link, Christopher Benner, and Christopher K Glass. Diverse motif ensembles specify non-redundant dna binding activities of ap-1 family members in macrophages. *bioRxiv*, 2018.

[20] V Galvao, J G Miranda, R F Andrade, J S Andrade Jr., L K Gallos, and H A Makse. Modularity map of the network of human cell differentiation. *Proc Natl Acad Sci U S A*, 107(13):5750–5755, 2010.

[21] Lucia Gandino and Luigi Varesio. Immortalization of macrophages from mouse bone marrow and fetal liver. *Experimental Cell Research*, 188(2):192–198, 1990.

[22] Mahmoud Ghandi, Dongwon Lee, Morteza Mohammad-Noori, and Michael A. Beer. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *PLoS Computational Biology*, 10(7), 2014.

[23] David Gosselin, Verena M. Link, Casey E. Romanoski, Gregory J. Fonseca, Dawn Z. Eichenfield, Nathanael J. Spann, Joshua D. Stender, Hyun B. Chun, Hannah Garner, Frederic Geissmann, and Christopher K. Glass. Environment Drives Selection and Function of Enhancers Controlling Tissue-Specific Macrophage Identities. *Cell*, 159(6):1327–1340, dec 2014.

[24] Valer Gotea, Axel Visel, John M. Westlund, Marcelo A. Nobrega, Len A. Pennacchio, and Ivan Ovcharenko. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Research*, 20(5):565–577, 2010.

[25] T Hai and T Curran. Cross-family dimerization of transcription factors Fos/Jun and ATF/CREB alters DNA binding specificity. *Proceedings of the National Academy of Sciences of the United States of America*, 88(9):3720–4, may 1991.

[26] T D Halazonetis, K Georgopoulos, M E Greenberg, and P Leder. c-Jun dimerizes with itself and with c-Fos, forming complexes of different DNA binding affinities. *Cell*, 55(5):917–24, dec 1988.

[27] Sebastian C. Hasenfuss, Latifa Bakiri, Martin K. Thomsen, Evan G. Williams, Johan Auwerx, and Erwin F. Wagner. Regulation of steatohepatitis and PPARγ signaling by distinct AP-1 dimers. *Cell Metabolism*, 19(1):84–95, 2014.

[28] S Heinz, C E Romanoski, C Benner, K A Allison, M U Kaikkonen, L D Orozco, and C K Glass. Effect of natural genetic variation on enhancer selection and function. *Nature*, oct 2013.

[29] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*, 38(4):576–89, may 2010.

[30] Sven Heinz and Christopher K Glass. Roles of lineage-determining transcription factors in establishing open chromatin: lessons from high-throughput studies. *Current topics in microbiology and immunology*, 356:1–15, jan 2012.

[31] Sven Heinz, Casey E. Romanoski, Christopher Benner, and Christopher K. Glass. The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology*, 16(3):144–154, mar 2015.

[32] Jochen Hess, Peter Angel, and Marina Schorpp-Kistner. AP-1 subunits: quarrel and harmony among siblings. *Journal of cell science*, 117(Pt 25):5965–73, dec 2004.

[33] Alina Isakova, Romain Groux, Michael Imbeault, Pernille Rainer, Daniel Alpern, Riccardo Dainese, Giovanna Ambrosini, Didier Trono, Philipp Bucher, and Bart Deplancke. SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nature Methods*, 14(3):316–322, jan 2017.

[34] Makiko Iwafuchi-Doi and Kenneth S. Zaret. Pioneer transcription factors in cell reprogramming. *Genes & Development*, 28(24):2679–2692, dec 2014.

[35] Wolfram Jochum, Emmanuelle Passegué, and Erwin F. Wagner. AP-1 in mouse development and tumorigenesis. *Oncogene*, 20(19 REV. ISS. 2):2401–2412, 2001.

[36] David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.

[37] R S Johnson, B M Spiegelman, and V Papaioannou. Pleiotropic effects of a null mutation in the c-fos proto-oncogene. *Cell*, 71(4):577–86, nov 1992.

[38] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M. Vaquerizas, Renaud Vincentelli, Nicholas M. Luscombe, Timothy R. Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, 2013.

[39] Eric Jones, Travis Oliphant, Pearu Peterson, and Others. SciPy: Open source scientific tools for Python.

[40] James T. Kadonaga. Perspectives on the RNA polymerase II core promoter. *Wiley Interdisciplinary Reviews: Developmental Biology*, 1(1):40–51, 2012.

[41] Minna U Kaikkonen, Michael T Y Lam, and Christopher K Glass. Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovascular research*, 90(3):430–40, jun 2011.

[42] Minna U Kaikkonen, Nathanael J Spann, Sven Heinz, Casey E Romanoski, Karmel A Allison,

Joshua D Stender, Hyun B Chun, David F Tough, Rab K Prinjha, Christopher Benner, and Christopher K Glass. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Molecular cell*, 51(3):310–25, aug 2013.

[43] Michael Karin and Tony Hunter. Transcriptional control by protein phosphorylation: signal transmission from the cell surface to the nucleus. *Current Biology*, 5(7):747–757, 1995.

[44] Thomas M Keane, Leo Goodstadt, Petr Danecek, Michael A White, Kim Wong, Binnaz Yalcin, Andreas Heger, Avigail Agam, Guy Slater, Martin Goodson, Nicholas A Furlotte, Eleazar Eskin, Christoffer Nellåker, Helen Whitley, James Cleak, Deborah Janowitz, Polinka Hernandez-Pliego, Andrew Edwards, T Grant Belgard, Peter L Oliver, Rebecca E McIntyre, Amarjit Bhomra, Jérôme Nicod, Xiangchao Gan, Wei Yuan, Louise van der Weyden, Charles A Steward, Sendu Bala, Jim Stalker, Richard Mott, Richard Durbin, Ian J Jackson, Anne Czechanski, José Afonso Guerra-Assunção, Leah Rae Donahue, Laura G Reinholdt, Bret A Payseur, Chris P Ponting, Ewan Birney, Jonathan Flint, and David J Adams. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–94, sep 2011.

[45] Michael A Keene and Sarah C R Elgin. Micrococcal nuclease as a probe of DNA sequence organization and chromatin structure. *Cell*, 27(1, Part 2):57–64, 1981.

[46] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7):990–9, jul 2016.

[47] Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A. Castro-Mondragon, Robin Van Der Lee, Adrien Bessy, Jeanne Chèneby, Shubhada R. Kulkarni, Ge Tan, Damir Baranasic, David J. Arenillas, Albin Sandelin, Klaas Vandepoele, Boris Lenhard, Benoît Ballester, Wyeth W. Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, 46(D1):D260–D266, 2018.

[48] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv*, pages 1–15, 2014.

[49] Ulf Klein, Stefano Casola, Giorgio Cattoretti, Qiong Shen, Marie Lia, Tongwei Mo, Thomas Ludwig, Klaus Rajewsky, and Riccardo Dalla-Favera. Transcription factor IRF4 controls plasma cell differentiation and class-switch recombination. *Nature Immunology*, 7(7):773–782, 2006.

[50] Frank T. Kolligs, Guido Bommer, and Burkhard Göke. Wnt/beta-catenin/Tcf signaling: A critical pathway in gastrointestinal tumorigenesis. *Digestion*, 66(3):131–144, 2002.

[51] Ivan V. Kulakovskiy, Ilya E. Vorontsov, Ivan S. Yevshin, Ruslan N. Sharipov, Alla D. Fedorova, Eugene I. Rumynskiy, Yulia A. Medvedeva, Arturo Magana-Mora, Vladimir B.

Bajic, Dmitry A. Papatsenko, Fedor A. Kolpakov, and Vsevolod J. Makeev. HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research*, 46(D1):D252–D259, 2018.

[52] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–359, 2012.

[53] D. Lara-Astiaso, A. Weiner, E. Lorenzo-Vivas, I. Zaretsky, D. A. Jaitin, E. David, H. Keren-Shaul, A. Mildner, D. Winter, S. Jung, N. Friedman, and I. Amit. Chromatin state dynamics during blood formation. *Science*, 345(6199):943–9, aug 2014.

[54] Yonit Lavin, Deborah Winter, Ronnie Blecher-Gonen, Eyal David, Hadas Keren-Shaul, Miriam Merad, Steffen Jung, and Ido Amit. Tissue-Resident Macrophage Enhancer Landscapes Are Shaped by the Local Microenvironment. *Cell*, 159(6):1312–1326, dec 2014.

[55] D. Lee, R. Karchin, and M. A. Beer. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Research*, 21(12):2167–2180, dec 2011.

[56] Dongwon Lee. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics (Oxford, England)*, 32(14):2196–8, 2016.

[57] Sung-Young Lee, Jaeho Yoon, Mee-Hyun Lee, Sung Keun Jung, Dong Joon Kim, Ann M Bode, Jaebong Kim, and Zigang Dong. The role of heterodimeric AP-1 protein comprised of JunD and c-Fos proteins in hematopoiesis. *The Journal of biological chemistry*, 287(37):31342–8, sep 2012.

[58] Jacob H. Levine, Erin F. Simonds, Sean C. Bendall, Kara L. Davis, El Ad D. Amir, Michelle D. Tadmor, Oren Litvin, Harris G. Fienberg, Astraea Jager, Eli R. Zunder, Rachel Finck, Amanda L. Gedman, Ina Radtke, James R. Downing, Dana Pe'er, and Garry P. Nolan. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–197, 2015.

[59] Mike Levine. Transcriptional Enhancers in Animal Development and Evolution. *Current Biology*, 20(17):R754–R763, sep 2010.

[60] Jun Li, Jingyi Li, and Bingbo Chen. Oct4 was a novel target of Wnt signaling pathway. *Molecular and Cellular Biochemistry*, 361(1-2):233–240, 2012.

[61] Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, 5(3):1752–1779, 2011.

[62] Chunyang Lien, Chee-Mun Fang, David Huso, Ferenc Livak, Runqing Lu, and Paula M. Pitha. Critical role of irf-5 in regulation of b-cell differentiation. *Proceedings of the National Academy of Sciences*, 107(10):4664–4668, 2010.

[63] A. P. Lifanov. Homotypic Regulatory Clusters in Drosophila. *Genome Research*, 13(4):579–588, 2003.

[64] Yin C. Lin, Suchit Jhunjhunwala, Christopher Benner, Sven Heinz, Eva Welinder, Robert Mansson, Mikael Sigvardsson, James Hagman, Celso A. Espinoza, Janusz Dutkowski, Trey Ideker, Christopher K. Glass, and Cornelis Murre. A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nature Immunology*, 11(7):635–643, 2010.

[65] Verena M. Link, Sascha H. Duttke, Hyun B. Chun, Inge R. Holtman, Emma Westin, Marten A. Hoeksema, Yohei Abe1, Dylan Skola, Casey E. Romanoski, Jenhan Tao, Greg Fonseca, Ty D. Troutman, Nathanael Spann, Tobias Strid, Mashito Sakai, Miao Yu, Hu Rong, Rongxin Fang, Dirk Metzler, Bing Ren, and Christopher K. Glass. Transcription Factor Landscapes in Macrophages from Genetically Diverse Mice Reveal Extensive Connected Regulatory Domains. *Cell*, CELL-D-18-, 2018.

[66] Shaun Mahony and Panayiotis V. Benos. STAMP: A web tool for exploring DNA-binding motif similarities. *Nucleic Acids Research*, 35(SUPPL.2):1–6, 2007.

[67] Kimihiko Matsusue, Martin Haluzik, Gilles Lambert, Sun-Hee Yim, Oksana Gavrilova, Jerrold M Ward, Bryan Brewer, Marc L Reitman, and Frank J Gonzalez. Liver-specific disruption of PPARgamma in leptin-deficient mice improves fatty liver but aggravates diabetic phenotypes. *The Journal of clinical investigation*, 111(5):737–47, mar 2003.

[68] V. Matys. TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34(90001):D108–D110, 2006.

[69] S R McKercher, B E Torbett, K L Anderson, G W Henkel, D J Vestal, H Baribault, M Klemsz, A J Feeney, G E Wu, C J Paige, and R A Maki. Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities. *The EMBO journal*, 15(20):5647–58, oct 1996.

[70] N. Miyoshi, H. Ishii, K.-i. Nagai, H. Hoshino, K. Mimori, F. Tanaka, H. Nagano, M. Sekimoto, Y. Doki, and M. Mori. Defined factors induce reprogramming of gastrointestinal cancer cells. *Proceedings of the National Academy of Sciences*, 107(1):40–45, 2010.

[71] Theresa L Murphy, Roxane Tussiwand, and Kenneth M Murphy. Specificity through cooperation: BATF-IRF interactions control immune-regulatory networks. *Nature reviews. Immunology*, 13(7):499–509, jul 2013.

[72] Y Nakabeppu and D Nathans. The basic region of Fos mediates specific DNA binding. *The EMBO journal*, 8(12):3833–41, dec 1989.

[73] K Okazaki and N Sagata. The Mos/MAP kinase pathway stabilizes c-Fos by phosphorylation and augments its transforming activity in NIH 3T3 cells. *The EMBO journal*, 14(20):5048–59, oct 1995.

[74] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2012.

[75] D Porte, P Oertel-Buchheit, M John, M Granger-Schnarr, and M Schnarr. DNA binding and transactivation properties of Fos variants with homodimerization capacity. *Nucleic acids research*, 25(15):3026–33, aug 1997.

[76] Daniel Quang and Xiaohui Xie. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic acids research*, 44(11):e107, jun 2016.

[77] Christian R. H. Raetz, Teresa A. Garrett, C. Michael Reynolds, Walter A. Shaw, Jeff D. Moore, Dale C. Smith, Anthony A. Ribeiro, Robert C. Murphy, Richard J. Ulevitch, Colleen Fearns, Donna Reichart, Christopher K. Glass, Chris Benner, Shankar Subramaniam, Richard Harkewicz, Rebecca C. Bowers-Gentry, Matthew W. Buczynski, Jennifer A. Cooper, Raymond A. Deems, and Edward A. Dennis. Kdo2-Lipid A of Escherichia coli, a defined endotoxin that activates macrophages via TLR-4. *Journal of Lipid Research*, 47(5):1097–1111, may 2006.

[78] Sekhar P M Reddy and Brooke T Mossman. Role and regulation of activator protein-1 in toxicant-induced responses of the lung. *American journal of physiology. Lung cellular and molecular physiology*, 283(6):L1161–78, dec 2002.

[79] Anshul Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, Viren Amin, John W Whitaker, Matthew D Schultz, Lucas D Ward, Abhishek Sarkar, Gerald Quon, Richard S Sandstrom, Matthew L Eaton, Yi-Chieh Wu, Andreas R Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R Alan Harris, Noam Shoresh, Charles B Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K Canfield, R Scott Hansen, Rajinder Kaul, Peter J Sabo, Mukul S Bansal, Annaick Carles, Jesse R Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R Mercer, Shane J Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C Sallari, Kyle T Siebenthall, Nicholas A Sinnott-Armstrong, Michael Stevens, Robert E Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E Beaudet, Laurie A Boyer, Philip L De Jager, Peggy J Farnham, Susan J Fisher, David Haussler, Steven J M Jones, Wei Li, Marco A Marra, Michael T McManus, Shamil Sunyaev, James A Thomson, Thea D Tlsty, Li-Huei Tsai, Wei Wang, Robert A Waterland, Michael Q Zhang, Lisa H Chadwick, Bradley E Bernstein, Joseph F Costello, Joseph R Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30,

feb 2015.

[80] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009.

[81] Sara Sabour, Nicholas Frosst, and Geoffrey Hinton. Dynamic Routing between Capsules. *Nips*, 2017.

[82] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(Database issue):D91–4, jan 2004.

[83] Jay Shendure, Gregory J Porreca, Nikos B Reppas, Xiaoxia Lin, John P McCutcheon, Abraham M Rosenbaum, Michael D Wang, Kun Zhang, Robi D Mitra, and George M Church. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science*, 309(5741):1728–1732, 2005.

[84] Daria Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews. Genetics*, 15(4):272–86, apr 2014.

[85] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. *arXiv*, 2017.

[86] Matthias Siebert, Johannes Soeding, and S Johannes. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research*, 44(13):6055–6069, 2016.

[87] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *Molecular Biology*, 147:195–197, 1981.

[88] Gary D. Stormo. Consensus patterns in DNA. *Methods in Enzymology*, 183(C):211–221, 1990.

[89] Kazutoshi Takahashi and Shinya Yamanaka. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, 126(4):663–676, 2006.

[90] Akinori Takaoka, Hideyuki Yanai, Seiji Kondo, Gordon Duncan, Hideo Negishi, Tatsuaki Mizutani, Shin Ichi Kano, Kenya Honda, Yusuke Ohba, Tak W. Mak, and Tadatsugu Taniguchi. Integral role of IRF-5 in the gene induction programme activated by Toll-like receptors. *Nature*, 434(7030):243–249, 2005.

[91] Sigal Tavor, Peter T Vuong, Dorothy J Park, Adrian F Gombart, Arthur H Cohen, and H Phillip Koeffler. Macrophage functional maturation and cytokine production are impaired in C/EBP epsilon-deficient mice. *Blood*, 99(5):1794–801, mar 2002.

[92] D Tempé, E Vives, F Brockly, H Brooks, S De Rossi, M Piechaczyk, and G Bossis. SUMOylation of the inducible (c-Fos:c-Jun)/AP-1 transcription complex occurs on target promoters to limit transcriptional activation. *Oncogene*, 33(7):921–7, feb 2014.

[93] L J P Van Der Maaten and G E Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[94] Laurens van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*, 15:3221–3245, 2014.

[95] Chris van Oevelen, Samuel Collombet, Guillermo Vicent, Maarten Hoogenkamp, Cyrille Lepoivre, Aimee Badeaux, Lars Bussmann, Jose Luis Sardina, Denis Thieffry, Miguel Beato, Yang Shi, Constanze Bonifer, and Thomas Graf. C/EBPα Activates Pre-existing and De Novo Macrophage Enhancers during Induced Pre-B Cell Transdifferentiation and Myelopoiesis. *Stem Cell Reports*, 5(2):232–47, jul 2015.

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

[97] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Nips*, jun 2017.

[98] A Verger, E Buisine, S Carrère, R Wintjens, A Flourens, J Coll, D Stéhelin, and M Duterque-Coquillaud. Identification of amino acid residues in the ETS transcription factor Erg that mediate Erg-Jun/Fos-DNA ternary complex formation. *The Journal of biological chemistry*, 276(20):17181–9, may 2001.

[99] Diego Villar, Camille Berthelot, Sarah Aldridge, Tim F. Rayner, Margus Lukk, Miguel Pignatelli, Thomas J. Park, Robert Deaville, Jonathan T. Erichsen, Anna J. Jasinska, James M.A. Turner, Mads F. Bertelsen, Elizabeth P. Murchison, Paul Flicek, and Duncan T. Odom. Enhancer Evolution across 20 Mammalian Species. *Cell*, 160(3):554–566, jan 2015.

[100] Jie Wang, Jiali Zhuang, Sowmya Iyer, Authors Jie Wang, Xinying Lin, Troy W Whitfield, Melissa C Greven, Brian G Pierce, Xianjun Dong, Anshul Kundaje, Yong Cheng, Oliver J Rando, Ewan Birney, Richard M Myers, William S Noble, Michael Snyder, and Zhiping Weng Comments. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors Repository Citation Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, 9:1798 – 1812, 2012.

[101] Z Q Wang, C Ovitt, A E Grigoriadis, U Möhle-Steinlein, U Rüther, and E F Wagner. Bone and haematopoietic defects in mice lacking c-fos. *Nature*, 360(6406):741–5, dec 1992.

[102] Matthew T. Weirauch, Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-

Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate Cook, Hong Zheng, Alejandra Goity, Harm van Bakel, Jean-Claude Lozano, Mary Galli, Mathew G. Lewsey, Eryong Huang, Tuhin Mukherjee, Xiaoting Chen, John S. Reece-Hoyes, Sridhar Govindarajan, Gad Shaulsky, Albertha J.M. Walhout, François-Yves Bouget, Gunnar Ratsch, Luis F. Larrondo, Joseph R. Ecker, and Timothy R. Hughes. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, 158(6):1431–1443, 2014.

[103] Warren A Whyte, David A Orlando, Denes Hnisz, Brian J Abraham, Charles Y Lin, Michael H Kagey, Peter B Rahl, Tong Ihn Lee, and Richard A Young. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–19, apr 2013.

[104] Edgar Wingender, Torsten Schoeps, Martin Haubrock, Mathias Krull, and Jürgen Dönitz. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic acids research*, 46(D1):D343–D347, jan 2018.

[105] Elzo De Wit and Wouter De Laat. A decade of 3C technologies-insights into nuclear organization. *Genes & development*, pages 11–24, 2012.

[106] Carl Wu. The 5 ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature*, 286:854, aug 1980.

[107] Tianlei Xu, Ben Li, Meng Zhao, Keith E. Szulwach, R. Craig Street, Li Lin, Bing Yao, Feiran Zhang, Peng Jin, Hao Wu, and Zhaohui S. Qin. Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. *Nucleic Acids Research*, 43(5):2757–2766, 2015.

[108] Michio Yamamoto, Takayuki Kato, Chie Hotta, Akira Nishiyama, Daisuke Kurotaki, Masahiro Yoshinari, Masamichi Takami, Motohide Ichino, Masatoshi Nakazawa, Toshifumi Matsuyama, Ryutaro Kamijo, Seiichi Kitagawa, Keiko Ozato, and Tomohiko Tamura. Shared and distinct functions of the transcription factors IRF4 and IRF8 in myeloid cell development. *PLoS ONE*, 6(10):2–11, 2011.

[109] S. Zandi, R. Mansson, P. Tsapogas, J. Zetterblad, D. Bryder, and M. Sigvardsson. EBF1 Is Essential for B-Lineage Priming and Establishment of a Transcription Factor Network in Common Lymphoid Progenitors. *The Journal of Immunology*, 181(5):3364–3372, 2008.

[110] Zhenhai Zhang and B Franklin Pugh. High-Resolution Genome-wide Mapping of the Primary Structure of Chromatin. *Cell*, 144(2):175–186, 2011.