UNIVERSITY OF CALIFORNIA SAN DIEGO

Investigating the Effects of Genetic Variation on Transcriptional Regulation

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

in

Bioengineering

by

Zeyang Shen

Committee in charge:

Professor Christopher Glass, Chair
Professor Kun Zhang, Co-Chair
Professor Ludmil B. Alexandrov
Professor Sanjoy Dasgupta
Professor Melissa Ann Gymrek

2021

The dissertation of Zeyang Shen is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

This dissertation is dedicated to Lily Yuxuan Zhu

for her support and company through this incredible journey.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ATAC-seq	Assay for transposase-accessible chromatin using sequencing
BALB	BALB/cJ
BMDM	Bone marrow-derived macrophage
C57	C57BL/6J
ChIP-seq	Chromatin immunoprecipitation sequencing
DNA	Deoxyribonucleic acid
DNase-seq	DNase I hypersensitive sites sequencing
GWAS	Genome-wide association study
H3K27ac	Acetylation of histone H3 lysine 27
H3K4me2	Di-methylation of histone H3 lysine 4
H3K4me3	Tri-methylation of histone H3 lysine 4
IL-4	Interleukin 4
InDel	short insertion and deletion
KLA	Kdo2 lipid A
LDTF	Lineage-determining transcription factor
MAGGIE	Motif Alteration Genome-wide to Globally Investigate Elements
NF-kB	Nuclear factor-kappa B
NOD	NOD/ShiLtJ
PWK	PWK/PhJ
PWM	Position weight matrix
QTL	Quantitative trait locus
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
SDTF	Signal-dependent transcription factor
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SPRET	SPRET/EiJ
TF	Transcription Factor

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Christopher Glass for the tremendous support, incredible energy, and valuable guidance as my advisor. I am grateful for the opportunity to work with him on exciting research and am honored to be his student. Prof. Glass is a role model in science to me and taught me that a dedicated scientist should always seek to answer the significant and difficult questions. His persistence, open-mindedness, and trust have been a gift to me during my research training.

Each of my committee members, Professor Kun Zhang, Professor Ludmil Alexandrov, Professor Sanjoy Dasgupta, and Professor Melissa Gymrek, provided me with valuable advice that helped improve different aspects of my research. I am thankful that they have always found time to discuss my research and am grateful for the unique insights and inspirations they brought. I am also grateful for Professor Chris Benner being a strong support outside of my committee.

I would like to thank my mentor, Dr. Jenhan Tao, who did not give up on me when I had zero knowledge about bioinformatics and provided continuous support while I was growing. His mentorship and encouragement helped me learn new computational techniques in a short period of time and offered me a strong sense of safety to overcome the times of uncertainty.

I would like to acknowledge members of the Glass Lab and my collaborators who trusted me with their data and time. Special thanks to Dr. Marten Hoeksema, who is a great scientist and friend, and a collaborator one could ever wish for. His positivity and frankness maintained my excitement about doing research. I would also like to thank An Zheng and Rick Zhenzhi Li for always offering their best and giving my ideas a chance.

Lastly, I would like to thank Jan Lenington, Leslie Van Ael, and Vanessa Hollingsworth for their administrative efforts. It would not be possible for me to reach each of the milestones

without their support.

Chapter 2, in full, is a reprint of the material as it appears in Shen Z, Hoeksema MA, Ouyang Z, Benner C & Glass CK. (2020). MAGGIE: leveraging genetic variation to identify DNA sequence motifs mediating transcription factor binding and function. Bioinformatics. 36(Supplement_1). The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, has been accepted for publication of the material as it will appear in Hoeksema MA*, Shen Z*, Holtman IR, Zheng A, Spann N, Cobo I, Gymrek M & Glass CK. (2020). Mechanisms underlying divergent responses of genetically distinct macrophages to IL-4. Science Advances. (*These authors contributed equally to this work). The dissertation author was a primary investigator and author of this paper.

Chapter 4, in full, is currently being prepared for submission for publication of the material. Shen Z, Li RZ, Prohaska T, Hoeksema MA, Spann N & Glass CK. The dissertation author was the primary investigator and author of this material.

VITA

2016	Bachelor of Engineering, Southeast University, China
2018-2019	Teaching Assistant, University of California San Diego
2019	Master of Science, University of California San Diego
2021	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Hoeksema MA*, **Shen Z***, Holtman IR, Zheng A, Spann N, Cobo I, Gymrek M & Glass CK. Mechanisms underlying divergent responses of genetically distinct macrophages to IL-4. *equal contribution. *Science Advances*. In press.

Shen Z, Hoeksema MA, Ouyang Z, Benner C & Glass CK. (2020). MAGGIE: leveraging genetic variation to identify DNA sequence motifs mediating transcription factor binding and function. *Bioinformatics*. 36(Supplement_1)

Shen Z, Tao J, Fonseca GJ & Glass CK. (2020). Natural genetic variation affecting transcription factor spacing at regulatory regions is generally well tolerated. *BioRxiv*.

Nott A, Holtman IR, Coufal NG, Schlachetzki JCM, Yu M, Hu R, Han CZ, Pena M, Xiao J, Wu Y, Keuelen Z, Pasillas MP, O'Connor C, Nickl CK, Schafer ST, **Shen Z**, Rissman RA, Brewer JB, Gosselin D, Gonda DD, Levy ML, Rosenfeld MG, McVicker G, Gage FH, Ren B & Glass CK. (2019). Cell type-specific enhancer-promoter connectivity maps in the human brain and disease risk association. *Science*. 366(6469)

Fonseca GJ, Tao J, Westin EM, Duttke SH, Spann NJ, Strid T, **Shen Z**, Stender JD, Link VM, Benner C & Glass CK. (2019). Diverse motif ensembles specify non-redundant DNA binding activities of AP-1 family members in macrophages. *Nature Communications*. 10(414).

FIELDS OF STUDY

Major Field: Bioinformatics

Studies in Genomics
Professor Christopher Glass

ABSTRACT OF THE DISSERTATION

Investigating the Effects of Genetic Variation on Transcriptional Regulation

by

Zeyang Shen

Doctor of Philosophy in Bioengineering

University of California San Diego, 2021

Professor Christopher Glass, Chair
Professor Kun Zhang, Co-Chair

Thousands of genetic variants have been found to increase disease risk based on genome-wide association studies. Many of these variants are located outside of protein-coding regions, suggesting their regulatory effects on gene transcription. However, it is not fully understood the effects of non-coding genetic variation on transcriptional regulation. One way of interpreting these variants is to link with the specific DNA sequences recognized by transcription factors

(TFs), which are also called motifs. I developed MAGGIE, a bioinformatic approach to identify functional motifs that mediate TF binding and function. Unlike many other motif analysis tools, MAGGIE associates motif mutations caused by non-coding variants with the changes in TF binding or regulatory function to provide more direct insights into the regulatory effects of genetic variation. I showed the outstanding performance of MAGGIE in various applications, including its ability to distinguish the divergent functions of distinct NF-kB factors in pro-inflammatory macrophages. As a detailed case study of the effects of non-coding variants, I applied MAGGIE to identify functional motifs for anti-inflammatory macrophages and discovered dominant TFs driving the anti-inflammatory response, which are also the frequent targets of genetic variation to influence such response. In combination with an integrative analysis of transcriptomic and epigenomic data, I revealed quantitative variations in motif affinity underlying the divergent anti-inflammatory responses observed in genetically different mouse strains. By leveraging deep learning approaches, I pinpointed functional variants altering functional motifs and provided strong evidence supporting the promise of using deep learning to identify functional variants. Finally, I went beyond motifs to systematically analyze the spacing between motifs and investigated its significance in the context of variant interpretation. I found most collaborative TFs do not require a constrained spacing but allow a relaxed range of spacing in between. Based on synthetic genetic variations from mutagenesis experiments and millions of naturally occurring variations, I showed that spacing alterations are generally tolerated by TF binding and regulatory function at TF binding sites. Collectively, these findings advance our understanding of how non-coding genetic variation influences gene transcription and phenotypic diversity.

Chapter 1. Introduction

DNA is the most basic material that makes every species and every person unique, but "basic" doesn't mean "simple". On the contrary, human DNA is a string of 3 billion base pairs made up of four different nucleotides – adenosine (A), cytosine (C), guanine (G), and thymine (T) – embedded with complex puzzles, so far, remaining to be fully solved. One of the successful examples is the codons within exons of genes that are used to encode proteins. A codon is composed of three nucleotides. Besides the start and stop codons that signal the initiation and the termination of transcription, respectively, each of the rest codons encodes a specific amino acid, which enables the direct prediction of amino acid sequence from DNA sequence and the potential interpretation of a gene's function. However, protein-coding regions compose only 2% of the genome. They answer the question of "what is transcribed?", but not "how much is transcribed?" or "when is a gene transcribed?". The latter questions are largely related to the regulation of gene transcription involving the rest of the DNA, non-coding regions. Among non-coding regions, regulatory elements including promoters and enhancers are defined as regions directly involved in transcriptional regulation, which leads to the large variety of gene expression in different cell types and cell states. Promoters are those at the transcription start sites of genes, while enhancers are further away, either at introns of genes or at intergenic. Studies found that these regulatory elements contain the majority of genetic variation associated with many common diseases (Farh et al., 2015; MacArthur et al., 2017), including cardiovascular diseases, neurodegenerative diseases, psychiatric diseases, etc. Therefore, it has long been one of the central problems in the genomic field how to decode DNA sequences of regulatory elements so that we can better understand those disease-associated variants. Apparently, the solution is not as

straightforward as three-nucleotide codons for protein-coding regions and remains to be fully understood.

Thanks to the rapid development of sequencing technologies and assays in the past twenty years, it is now possible to study the relationship between non-coding sequences and transcriptional regulation from various aspects. RNA-seq (RNA sequencing) measures the number of transcripts, in other words, the outcomes of the transcriptional regulation. ChIP-seq (chromatin immunoprecipitation sequencing) is able to measure the genome-wide binding sites of DNA-binding proteins, including transcription factors (TFs), RNA polymerases, and specific histone modifications (Reuter et al., 2015). TFs recognize specific DNA sequences when binding to DNA and play an important role in activating regulatory function. RNA polymerases and histone modifications can link TF binding to the actual regulatory function. Commonly measured histone modifications include acetylation of histone H3 lysine 27 (H3K27ac) representing regulatory activity (Creyghton et al., 2010), di-methylation of histone H3 lysine 4 (H3K4me2) enriched at regulatory regions, and tri-methylation of histone H3 lysine 4 (H3K4me3) enriched at promoters (Koch et al., 2007). Unlike ChIP-seq which has specific target proteins and histone modifications, ATAC-seq (assay for transposase-accessible chromatin using sequencing) and DNase-seq (DNase I hypersensitive sites sequencing) both measure chromatin regions accessible for TF binding (Reuter et al., 2015). Going beyond linear sequences to 3D structures, techniques like Hi-C (chromosome conformation capture coupled with sequencing), ChIA-PET (chromatin interaction analysis by paired-end tag sequencing), and HiChIP can measure chromatin interactions and potentially pinpoint the target genes of regulatory elements (Mumbach et al., 2016). An integrative analysis of these data measuring different aspects of transcriptional regulation are important to understand how non-coding sequences are related to regulatory

function. In Chapter 3, I integrate different types of sequencing data to study the anti-inflammatory response of macrophages and dissect the rules for how non-coding genetic variation can affect such response.

One of the key breakthroughs for decoding regulatory sequences is the study of DNA sequences recognized and bound by TFs, which are also called TF binding motifs or motifs for short. Even though motifs have a direct link to TF binding and show specificity for different TFs, the identification and interpretation of motifs are not easy. First, unlike codons, which have clear one-to-one relationships, one motif can be recognized by several different TFs and is usually not just a single sequence, but an aggregation of sequences allowing variation at some of the positions. Another complication is that motifs are important to be considered together with each other. Studies have shown intensive collaborations between TFs, each of which binds to their respective motifs within hundreds of bases (Lambert et al., 2018). Certain TFs are found specific to cell types and cell states and, in many cases, need to collaborate with each other to activate regulatory function (Spitz & Furlong, 2012).

In Chapter 2, I introduce a bioinformatic approach, MAGGIE (https://github.com/zeyang-shen/maggie) to identify functional DNA sequence motifs that mediate TF binding and regulatory function. Instead of looking at the presence or absence of a motif like many other motif analysis tools, this method focuses on how genetic variation at motifs link with the changes in TF binding or regulatory function. As a result, this work builds a more direct connection between motif and its impact on regulatory function and provides more direct insights into the interpretation of non-coding variants. I demonstrate how MAGGIE outperforms existing motif analysis tools in identifying functional motifs and is applicable to some complex problems (e.g., the response of enhancers to stimulus). As part of Chapter 3, I applied MAGGIE to identify

functional motifs for the anti-inflammatory response of macrophages. It helps discover a TF called EGR2 in parallel with a recently published work (Daniel et al., 2020) as an important regulator for the anti-inflammatory response. MAGGIE also exclusively reveals the mechanisms of how motif affinity can determine not only the basal level of enhancer activity but also the fold change of enhancer activity in response to stimulus, providing an explanation for the divergent anti-inflammatory responses observed in genetically different mouse strains.

If most motif analysis tools including MAGGIE still focus on individual motifs one at a time, the recent application of deep learning or deep neural networks in the genomic field has significantly changed the way people study non-coding sequences (Eraslan et al., 2019). Deep neural networks can consider a much larger range of sequences maintaining multiple motifs and their relative locations. After iterations of training, neural networks are able to learn patterns among a given set of sequences corresponding to certain function. One major part of Chapter 3 describes the application of deep learning in dissecting enhancer sequences relevant to the anti-inflammatory response of macrophages. I demonstrate the superior power of neural networks in predicting active enhancers based on DNA sequences. Using model interpretation techniques, I show that the important sequence patterns captured by neural networks match with the known functional motifs and, more importantly, highlight the vulnerable positions for high-impact genetic variation. Based on the strong enrichment of the variants associated with changes in regulatory function at these vulnerable positions, this work provides further evidence to support the potential of using deep learning to predict the impact of non-coding variants and, therefore, pinpoint functional variants.

In Chapter 4, I put the focus on the spacing relationships between motifs with a special focus on the impacts of spacing alterations on TF binding and regulatory function. The impacts

of motif mutations have been widely studied in different scenarios for various TFs, while the impacts of spacing altercations are much less studied, especially not on a systematic level. Based on an analysis on dozens of TFs available from the ENCODE database (Davis et al., 2018), I summarize three major categories of spacing relationships between collaboratively binding TFs, among which the majority follows a "relaxed" spacing relationship, meaning that they frequently bind close to each other while allowing for variable spacing as well. Two case studies of TFs with relaxed spacing relationships, one for human endothelial cells and the other for mouse macrophages, provide back-to-back evidence supporting general tolerance of spacing alterations for TF binding and regulatory function. Mutagenesis experiments that test the effects of variable sizes of synthetic spacing alterations further show the robustness of such tolerance.

Collectively, this work investigates the regulatory effects of genetic variation using different bioinformatic approaches, integrating various types of sequencing data, and analyzing DNA sequences from multiple aspects, including motif and spacing.

Chapter 2. Leveraging genetic variation to identify DNA sequence motifs mediating transcription factor binding and function

2.1 Abstract

Genetic variation in regulatory elements can alter transcription factor (TF) binding by mutating a TF binding motif, which in turn may affect the activity of the regulatory elements. However, it is unclear which motifs are prone to impact transcriptional regulation if mutated. Current motif analysis tools either prioritize TFs based on motif enrichment without linking to a function or are limited in their applications due to the assumption of linearity between motifs and their functional effects. We present MAGGIE (Motif Alteration Genome-wide to Globally Investigate Elements), a novel method for identifying motifs mediating TF binding and function. By leveraging measurements from diverse genotypes, MAGGIE uses a statistical approach to link mutations of a motif to changes of an epigenomic feature without assuming a linear relationship. We benchmark MAGGIE across various applications using both simulated and biological datasets and demonstrate its improvement in sensitivity and specificity compared with the state-of-the-art motif analysis approaches. We use MAGGIE to gain novel insights into the divergent functions of distinct NF-kB factors in pro-inflammatory macrophages, revealing the association of p65–p50 co-binding with transcriptional activation and the association of p50 binding lacking p65 with transcriptional repression. The Python package for MAGGIE is freely available at https://github.com/zeyang-shen/maggie. The accession number for the NF-kB ChIP-seq data generated for this study is Gene Expression Omnibus: GSE144070.

2.2 Introduction

Genome-wide association studies (GWASs) have identified thousands of genetic variants associated with an increase in disease risk (MacArthur et al., 2017). Many of these variants fall

within regulatory elements such as promoters and enhancers, implicating an effect on transcriptional regulation (Farh et al., 2015; Khurana et al., 2016). Transcription factors (TFs) play an essential role in mediating the activity of regulatory elements. Many TFs possess DNA-binding domains that recognize specific DNA sequences, called TF binding motifs. Alterations of TF binding motifs have been established as an important mechanism for genetic variants to affect transcriptional regulation (Deplancke et al., 2016; Grossman et al., 2017; Heinz et al., 2013). However, it is not always straightforward which TF binding motifs are prone to have an impact on transcriptional regulation if mutated. First of all, a genetic variant is able to alter multiple motifs. Binding motifs for hundreds of TFs are currently available in the public databases (Fornes et al., 2020; Kulakovskiy et al., 2018; Matys et al., 2006). Many motifs correspond to similar or overlapping DNA sequences, which can be altered by the same variant simultaneously. The second complication is due to the strong dependency of TF binding on conditions. Multiple TF binding motifs are usually packed at regulatory elements across 100–200 base pairs (Lambert et al., 2018) but can become functional under different conditions depending on cell type, developmental time point, stimulus etc. (Spitz & Furlong, 2012) Knowing the function of motifs for a given condition can help prioritize TFs prone to be affected by genetic variation and ultimately have an impact on transcriptional regulation.

Numerous motif analysis tools have been published in the past decade to prioritize important TFs for experimental validation (Boeva, 2016; Jayaram et al., 2016). One major category of tools identifies enriched motifs that appear more frequently at given regions of interest than random genomic regions (Heinz et al., 2010; Machanick & Bailey, 2011; Siebert & Söding, 2016). Due to the development of high-throughput sequencing assays, these approaches can now be applied to various types of epigenomic features, such as chromatin accessibility

measured by the assay for transposase-accessible chromatin using sequencing (ATAC-seq) or DNase I hypersensitive sites sequencing (DNase-seq), and TF binding and histone modification measured by chromatin immunoprecipitation sequencing (ChIP-seq) etc. (Reuter et al., 2015). However, this category of methods does not connect motif enrichment to a function, so the identified motifs may not have any functional impact on the epigenomic feature of interest.

Another category of motif analysis tools prioritize TFs by leveraging measurements and genetic variation of multiple human individuals or animal strains. Many of these methods depend on an assumption of linearity between the motif and the signal of epigenomic features (Fonseca et al., 2019; Grubert et al., 2015; Link, Romanoski, et al., 2018; Mcvicker et al., 2013). This assumption worked for TF binding but likely does not hold for many other epigenomic features like histone modification or stimulus response of regulatory elements, which result from the interactions between multiple TFs and may not possess a simple linear relationship with TF binding motifs.

Here, we developed a novel approach, MAGGIE (Motif Alteration Genome-wide to Globally Investigate Elements), to identify DNA motifs mediating TF binding and function. Considering the increasing amount of genotype and epigenomic data for different individuals and animal strains, we are able to identify genomic regions associated with a biased epigenomic feature of interest between different genotypes, labeling them as positive or negative for sequences with or without the feature, respectively (Fig. 2.1A). We propose to associate these biased regions with changes of TF binding motifs caused by genetic variation to gain insights into the functions of motifs. Unlike conventional motif enrichment methods, MAGGIE is independent of the background frequency of motifs and gains power in capturing the functional impacts of motifs by leveraging motif mutations at the same regions between individuals or

strains. MAGGIE differs from other methods that associate motif mutations with epigenomic features by eliminating the assumption of linearity between motifs and testing features. The design of this framework is flexible in accommodating any type of epigenomic feature, including but not limited to the ones to be discussed in this article, such as TF binding, open chromatin, histone modification and stimulus response of regulatory elements.

We evaluated the performance of MAGGIE in both simulated datasets and biological datasets and compared our results to HOMER (Heinz et al., 2010), MMARGE (Link, Romanoski, et al., 2018) and TBA (Fonseca et al., 2019), which are representative for the existing motif analysis tools. The results demonstrated the superior sensitivity and specificity of MAGGIE for detecting the effects of motif mutations in all of the experiments. By applying MAGGIE to the regulatory elements of macrophages in response to proinflammatory stimulus, we captured divergent functions of distinct NF-kB (nuclear factor-kappa B) factors despite the similarity of their motifs. These results were further validated by the NF-kB binding sites measured by ChIP-seq experiments, showing the promise of MAGGIE in identifying highly specific motifs and discovering novel functions of TFs.

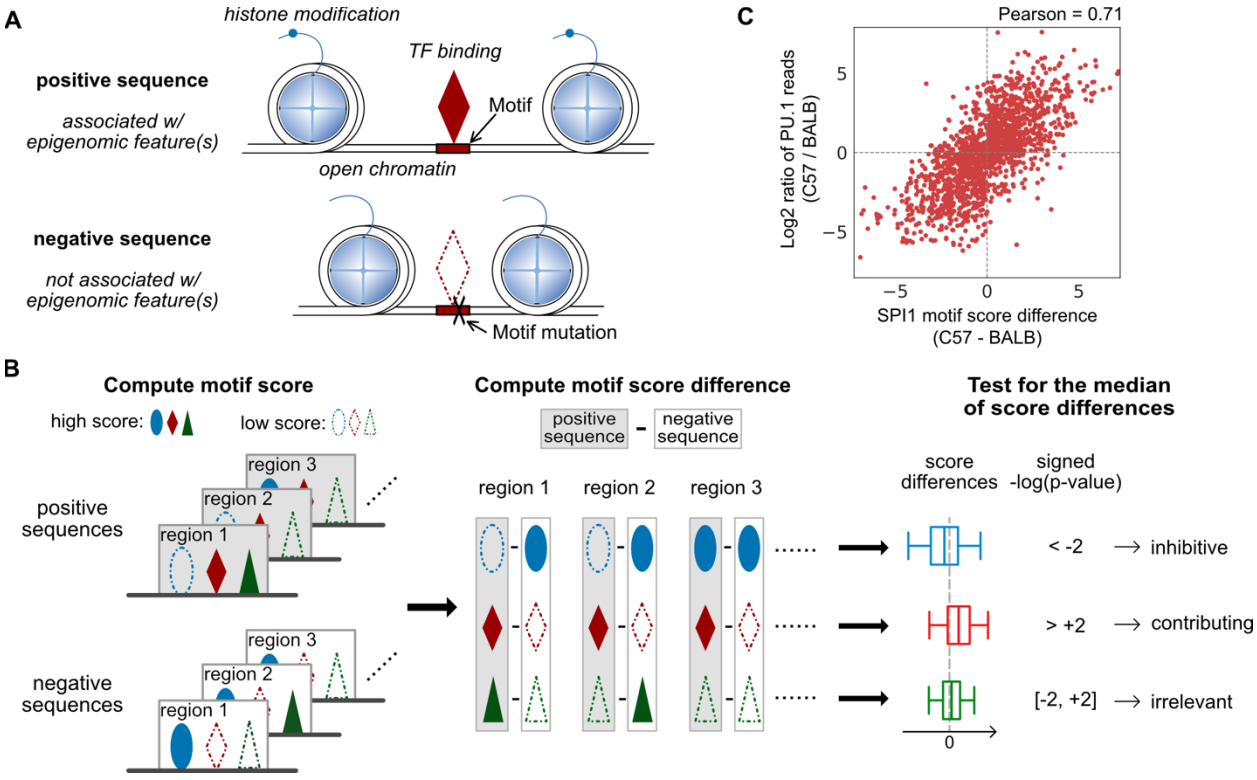


Figure 2.1: Overview of MAGGIE. (A) Schematic depicting how the epigenomic features of regulatory elements are related to the inputs of MAGGIE. Positive sequences are defined to be associated with epigenomic feature(s) of interest, such as TF binding, open chromatin, histone modification etc. Each positive sequence has a negative counterpart, which has a loss of the chosen epigenomic feature(s) due to mutations on TF binding motifs. **(B)** Flowchart of MAGGIE. Positive and negative sequences are used to compute motif scores as an estimated likelihood of being bound by certain TF. A representative motif score is obtained for each sequence by taking the maximum, displayed by different shapes (ellipse, diamond and triangle) for different TFs. High motif scores are shown as solid shapes and low scores as dashed shapes. Next, differences of representative motif scores are computed for every TF by subtracting scores of negative from positive sequences. Finally, the score differences for each TF are aggregated, and the median value is tested by Wilcoxon signed-rank test to evaluate whether there is a bias in the changing direction from positive to negative sequences. The examples demonstrate a significant bias of increase (ellipse) or decrease (diamond) or an insignificant bias (triangle), which implicates the inhibitive, contributing, or irrelevant role of TF, respectively. **(C)** Correlation between motif score differences of SPI1 motif and log2-fold changes of PU.1 binding activity between BALB and C57 mice. Each dot represents one of the 1641 PU.1 binding sites that have SPI1 motif mutations between the two strains.

2.3 Materials and methods

2.3.1 Overview of MAGGIE

The overall framework of MAGGIE is illustrated in Fig. 2.1B. MAGGIE takes pairs of sequences as inputs. Positive sequences are identified to be associated with an epigenomic feature of interest, while negative sequences are from different alleles or the same regions of a

different genome where the epigenomic feature is not found. Depending on the genetic difference of genomes, every pair of input sequences can have a variable number of genetic variants like single-nucleotide polymorphisms (SNPs) and short insertions and deletions.

The basic assumption for MAGGIE is that the allele specificity of an epigenomic feature is derived from the genetic variation between positive and negative sequences that mutate TF binding motifs. This assumption is supported by the findings that motif mutations due to local genetic variation is the major explanation for the gain or loss of TF binding sites (Link, Duttke, et al., 2018; Roadmap Epigenomics Consortium et al., 2015). Considering the importance of TFs for other epigenomic features like promoter and enhancer function (Reiter et al., 2017; Spitz & Furlong, 2012), we hypothesized that our framework could help identify motifs mediating both TF binding and other epigenomic features affected by TF binding.

The computation of MAGGIE is centered on the motif score based on position weight matrix (PWM), which is the widely used metric to approximate the likelihoods of being bound by certain TF (Stormo, 2000). Given pairs of positive and negative sequences associated with a chosen allele-specific epigenomic feature, MAGGIE computes motif scores for hundreds of TFs whose PWMs are currently available in the JASPAR database (Fornes et al., 2020). For each TF, a representative motif score is calculated for every sequence by taking the maximal score across the sequence. MAGGIE then computes differences of representative motif scores by subtracting scores of negative from positive sequences to obtain the changes of binding likelihood. Score differences should have a bias toward positive values (i.e., higher motif scores in positive sequences) if the corresponding TF is contributing to the chosen epigenomic feature. On the contrary, if the TF is potentially inhibitive for the chosen feature, the aggregated differences will tend to have negative values (i.e., lower motif scores in positive sequences). Irrelevant TFs will

have their motifs randomly mutated by genetic variation, so the score differences should be overall balanced around zero. A non-parametric Wilcoxon signed-rank two-sided test is used to statistically test the significance of the association between motif mutations and the chosen epigenomic feature by asking whether the median of all the non-zero motif score differences is close to zero. A signed P-value combining the sign of median score difference with the P-value from statistical tests implicates the function of TF to be either contributing (positive) or inhibitive (negative) if called significant.

2.3.2 Computation of motif score and motif score

Motif score is a reliable metric to measure the likelihood of TF binding and can well reflect the binding activity of the corresponding TF (Boeva, 2016; Ji et al., 2018). A PWM stores the log likelihoods for the four possible nucleotides (A, C, G and T) to be bound by a TF at each position (Stormo, 2000):

$$M_{k,n} = \log2(\frac{P_{k,n}}{b_n})$$

where $P_{k,n}$ is the probability of seeing nucleotide n at the kth position of the motif, and b_n is the background probability for different nucleotides. Given a DNA sequence, we can compute motif scores for any TF by adding up the log likelihoods of seeing certain nucleotides at every position:

$$S_i = \sum_{k=0}^{L-1} M_{k,n_{i+k}} = \sum_{k=0}^{L-1} \log2(\frac{P_{k,n_{i+k}}}{b_{n_{i+k}}})$$

where S_i is the motif score for a segment of the given sequence from position i to position i+L-1, supposing L is the length of the motif and i starts at 1, and n_{i+k} is the nucleotide at position i+k. For a sequence longer than the motif (i.e., the biggest possible i > L), instead of dealing with a

list of motif scores, we obtain the maximal motif score to represent the binding likelihood of the entire sequence:

$$S_R = \max\{S_i \; i = 1,2, \dots\} = \sum_{k=0}^{L-1} \log2(\frac{P_{k,n_{r+k}}}{b_{n_{r+k}}})$$

where r is the starting position of the maximal motif score. Every sequence pair will yield two representative motif scores whose starting positions are notated by r_P and r_N for positive and negative sequence, respectively:

$$S_R^{Pos} = \sum_{k=0}^{L-1} \log2(\frac{P_{k,n_{r_P+k}^{Pos}}}{b_{n_{r_P+k}^{Pos}}})$$

$$S_R^{Neg} = \sum_{k=0}^{L-1} \log2(\frac{P_{k,n_{r_N+k}^{Neg}}}{b_{n_{r_N+k}^{Neg}}})$$

Then, the log-fold change of binding likelihood within the sequence pair can be computed by subtracting the representative motif score of the negative sequence from that of the positive sequence:

$$S_R^{Pos} - S_R^{Neg} = \sum_{k=0}^{L-1} \log2\left(\frac{P_{k,n_{r_P+k}^{Pos}}}{b_{n_{r_P+k}^{Pos}}}\right) - \sum_{k=0}^{L-1} \log2(\frac{P_{k,n_{r_N+k}^{Neg}}}{b_{n_{r_N+k}^{Neg}}})$$

If we set the background probability as the same for the four types of nucleotides (i.e., 0.25), the difference of representative motif score turns out to be the log-fold change of the binding likelihood between positive and negative sequences:

$$S_R^{Pos} - S_R^{Neg} = \sum_{k=0}^{L-1} \log2\left(\frac{P_{k,n_{r_P+k}^{Pos}}}{P_{k,n_{r_N+k}^{Neg}}}\right) = \log2\left(\frac{\prod_{k=0}^{L-1} P_{k,n_{r_P+k}^{Pos}}}{\prod_{k=0}^{L-1} P_{k,n_{r_N+k}^{Neg}}}\right)$$

Here, we compute the motif score difference based on the maximal score of each sequence, which may or may not at the same location (r_P not necessarily equal to r_N). This

strategy is able to compensate for the effects from nearby variants and the interactions between multiple motifs. Any representative motif score less than zero is replaced by zero before computing a score difference in order to reduce impacts from poorly matched motifs. Motif score difference has been used as an indicator of the change in TF binding (Martin et al., 2019; Spivakov et al., 2012). For example, by comparing PU.1 binding in macrophages of C57BL/6J (C57) and BALB/cJ (BALB) mice (Link, Duttke, et al., 2018), we observed a strong positive correlation between the score difference of SPI1 motif and the change in PU.1 (encoded by SPI1) binding quantified by ChIP-seq reads (Fig. 2.1C). This relationship is independent of the actual motif score (Supplementary Fig. 2.1). We saw a diminished correlation using non-uniform background probabilities (Supplementary Fig. 2.2) or restricting motifs at the same locations ($r_P = r_N$) instead of their respective best matches (Supplementary Fig. 2.3). These intrinsic characteristics of motif score difference support the hypotheses that (i) motif score difference can indicate change in binding of the corresponding TF, and (ii) aggregated motif score differences can reflect whether the presence of specific epigenomic feature is associated with the gain or loss of TF binding.

2.3.3 Applications and data preparation

2.3.3.1 Simulated data

To characterize the performance of MAGGIE and systematically compare with other methods, we conducted simulated experiments. Positive sequences were generated by first randomly selecting A, C, G or T to form sequences of 200-base pair (bp). Then we created TF binding motifs by sampling nucleotides based on their probabilities derived from PWMs and inserted these motifs at non-overlapping random positions. To obtain counterpart negative

sequences, SNPs were simulated inside hypothetic 'contributing' motifs by changing the existing nucleotides.

During the generation of simulated data, we inserted 'irrelevant' motifs, which experienced either no mutation or random mutation, to evaluate the specificity of MAGGIE. The sensitivity of MAGGIE was tested by changing the number of simulated sequences (i.e., sample size) or the fraction of sequences having motif mutations [i.e., signal-to-noise ratio (SNR)].

2.3.3.2 TF binding sites

We tested MAGGIE to identify TF binding motifs for corresponding TF binding. Allele-specific binding sites of 12 TFs were obtained from two cell types, GM12878 and HeLa-S3 (Shi et al., 2016). We extracted 100-bp sequences around the SNPs associated with allele-specific binding sites and labeled the sequences with the binding alleles as positive sequences and those with the non-binding alleles as negative.

MAGGIE was then used to identify collaborative TFs. We downloaded the ChIP-seq data of PU.1 and C/EBPb for C57 and BALB mice from the Gene Expression Omnibus (GEO) database with accession number GSE109965 (Link, Duttke, et al., 2018), and the ChIP-seq data of ATF3 for the same mouse strains with the accession number GSE46494 (Fonseca et al., 2019). The data for C57 were mapped to the mm10 genome using Bowtie2 v2.3.5.1 (Langmead & Salzberg, 2012), whereas the data for BALB were first mapped to the BALB genome and then shifted to the mm10 genome using the MMARGE v1.0 'shift' function (Link, Romanoski, et al., 2018). The reproducible TF binding sites were identified by using HOMER v4.9.1 to call unfiltered 200-bp peaks (Heinz et al., 2010) and running IDR v2.0.3 on replicates with the default parameters (Li et al., 2011). The TF binding sites found only in one of the strains were defined to be strain-specific, yielding 13,099 PU.1, 8,127 C/EBPb and 13,347 ATF3 strain-

specific binding sites between BALB and C57. The sequences of strain-specific binding sites were extracted from both strains using the MMARGE v1.0 'extract_sequences' function (Link, Romanoski, et al., 2018). Sequences associated with TF binding are labeled as positive regardless of which strain they are originated from, and their counterpart sequences from the other strain are labeled as negative.

2.3.3.3 Chromatin quantitative trait loci

We applied MAGGIE to discover motifs mediating chromatin accessibility and histone modification. DNaseI sensitivity quantitative trait loci (dsQTLs) were downloaded from the GEO database with accession number GSE31388 (Degner et al., 2012). Histone QTLs (hQTLs) were acquired for three types of histone modifications (Grubert et al., 2015), local acetylation of histone H3 lysine 27 (H3K27ac), monomethylation of histone H3 lysine 4 (H3K4me1) and tri-methylation of histone H3 lysine 4 (H3K4me3). All QTLs were originally analyzed for lymphoblastoid cell lines (LCLs). We obtained more stringent hQTLs based on a P-value < 1e-6 and a distance from the associated peak <1000 bp. QTLs were further separated based on HOMER annotations into promoter, intronic and distal subsets to investigate functional motifs of different genomic regions. Distal QTLs are those within intergenic regions and >2000 bp from the nearest transcription start sites. Similar to the pre-processing for the allele-specific binding sites, we extracted 100-bp sequences centering around the variants and labeled the alleles associated with a higher trait level as positive and the other alleles as negative.

2.3.3.4 Stimulus responses of regulatory elements

The application of MAGGIE was further extended to the stimulus response of regulatory elements. We downloaded ATAC-seq and H3K27ac ChIP-seq data from macrophages at both basal state and pro-inflammatory state induced by 1-h treatment of the TLR4-specific ligand

Kdo2 lipid A (KLA) from four diverse strains of mice (Link, Duttke, et al., 2018): C57BL/6J (C57), NOD/ShiLtJ (NOD), PWK/PhJ (PWK) and SPRET/EiJ (SPRET). Similar to the preprocessing of ChIP-seq data for TFs, the raw reads were mapped and shifted to the mm10 genome. Based on ATAC-seq data, we obtained 200-bp reproducible open chromatin and filtered for intergenic and intronic regions to obtain potential enhancers. Open chromatin regions of the two conditions from the same strain were merged and extended from 200 to 1000 bp to quantify their activity by the count of H3K27ac ChIP-seq reads. We filtered for active regulatory elements (>16 reads in at least one condition; Supplementary Fig. 2.6) and computed the change of activity from basal to KLA-treated condition by the fold change of reads. Regions showing a higher or lower level of H3K27ac >2.5-fold after KLA treatment were labeled as 'activated' or 'repressed', respectively (Fig. 2.4A), and those with <40% change were labeled as 'neutral'. Based on pairwise comparisons across the four mouse strains, regulatory elements labeled as 'activated' or 'repressed' only in one of the compared strains were called strain-specific and were pooled for analysis.

2.3.4 Comparative methods

We compared the performance of MAGGIE against several existing methods in identifying functional TF binding motifs. The most obvious competitors are those that also leverage measurements from diverse genotypes, including a recently developed method called MMARGE (Link, Romanoski, et al., 2018), which fits a linear mixed model between the motif score and the signal of epigenomic features (e.g., TF binding activity). Unlike other approaches based on linear assumptions, MMARGE additionally corrects for individual variance while leveraging genetic variation between individuals. MMARGE v1.0 was downloaded from https://github.com/vlink/marge. Since the existing linear methods do not directly work with

binary-labeled datasets (e.g., simulated data, QTLs), we implemented a replacement model that fit motif scores against binary labels in the simulated experiments using statsmodels package (Seabold & Perktold, 2010).

Another big category of motif analysis tools is based on motif enrichment algorithms, such as HOMER (Heinz et al., 2010), MEME Suite (Machanick & Bailey, 2011), BaMM (Siebert & Söding, 2016), etc. We performed comparisons between enriched and functional motifs identified by MAGGIE. We expect any one of those methods to be representative for the others, so we picked HOMER in our experiments, which was downloaded from http://homer.ucsd.edu/homer/data/software/homer.v4.9.1.zip. Besides using HOMER to find enriched motifs, we extended its application to calculate differential enrichment between positive and negative sequences and evaluated the performance of enrichment algorithms in detecting motif mutations.

We also adapted a machine learning-based approach, TBA, to detect motif mutations between positive and negative sequences (Fonseca et al., 2019). We trained a logistic regression model with representative motif scores, from which P-values were generated from the likelihood-ratio test to represent the importance of each motif in classifying binary labels. The model training modules were downloaded from https://github.com/jenhantao/tba. All of the comparative methods above were run with the default parameters. Since none of these methods output signed P-values as MAGGIE does, we reported only P-values from MAGGIE in any comparative studies.

2.3.5 Validation experiment

Bone marrow was isolated from C57 mice and differentiated for 7 days using media containing M-CSF to generate bone marrow-derived macrophages (BMDMs) as described

previously (Link, Duttke, et al., 2018). BMDMs were maintained at basal conditions or treated with KLA for 1 h. For p65 (Santa Cruz, sc-372X) and p50 (Abcam, ab32360) ChIP-seq experiments, 8 million untreated or KLA-treated BMDMs per assay were double-crosslinked using disuccinimidyl glutarate and formaldehyde (FA). ChIP-seq was performed using 2 lg of antibody as described previously (Heinz et al., 2018). ChIP DNA was prepared for sequencing using the NEBNext Ultra II DNA library prep kit (NEB, E7645) and sequencing was performed on the HiSeq4000 (75 bp SR, Illumina). The binding sites of p65 and p50 were identified using HOMER 'findPeaks -size 200' (Heinz et al., 2010) and then merged to obtain co-binding sites and p65- or p50-only binding sites. The binding activity of p65 and p50 was quantified by the count of ChIP-seq reads. The raw and processed data have been deposited to the NCBI GEO under the accession number GSE144070.

2.4 Results

2.4.1 MAGGIE shows superior specificity and sensitivity on simulated datasets

Table 2.1: Top motifs output from different motif analysis tools evaluated on the simulated datasets. Log10 P-values are shown in parentheses. (*) indicates motifs that passed FDR < 0.05 after the Benjamini–Hochberg controlling procedure. As the true positive, SPI1 motif is highlighted in bold.

Rank	MAGGIE	Linear model	Logistic regression (TBA)	HOMER— pos versus bg	HOMER— neg versus bg
1	**SPI1*** (13)	**SPI1*** (8.6)	**SPI1** (1.0)	CEBPG* (198)	CEBPG* (195)
2	SPIB* (10)	SPIB* (6.5)	ZSCAN10 (0.8)	CEBPD* (194)	CEBPB* (183)
3	SPIC* (4.7)	ETV6 (3.6)	EWSR1-FLI1(0.8)	CEBPB* (192)	CEBPD* (181)
4	ETV6* (4.5)	SPIC (3.6)	STAT5A (0.6)	CEBPE* (191)	CEBPE* (178)
5	ELF1* (4.2)	EHF (3.2)	SIX6 (0.4)	**SPI1*** (177)	**SPI1*** (127)

To evaluate the performance of MAGGIE, we stochastically simulated one thousand DNA sequences of 200 bp embedded with an arbitrary pair of motifs, SPI1 and CEBPB, labeled as positive sequences. Negative sequences were then derived from this set by switching single

nucleotides of the SPI1 motif for half of the positive sequences. Table 1.1 shows the top motifs output from MAGGIE and three comparative approaches: linear model, logistic regression adapted from TBA and HOMER. Both MAGGIE and linear model identified SPI1 and its similar motifs as the most significant hits. Logistic regression that was trained to classify positive and negative sequences lacked the sensitivity to detect SPI1 motif. On the contrary, HOMER identified both SPI1 and CEBPB as significantly enriched over the default random backgrounds for both positive ('pos versus bg' column) and negative sequences ('neg versus bg' column). As expected, enriched motifs failed to distinguish the mutated motif from the unmutated motif, which was only captured by methods that leveraged motif mutations resulted from synthetic genetic variation.

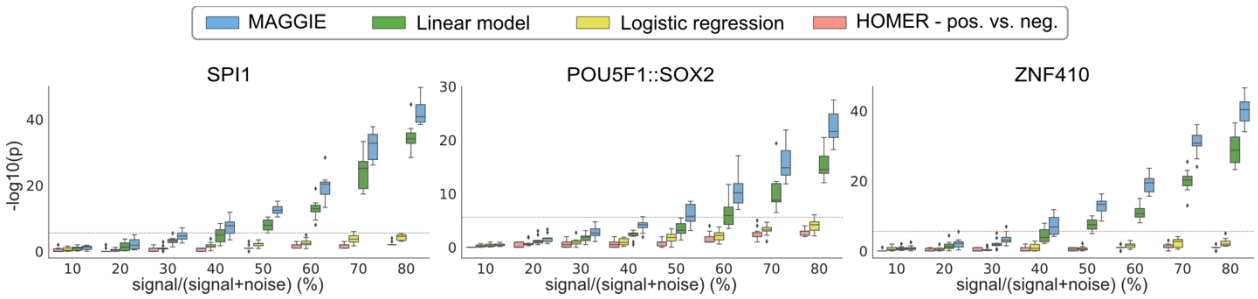


Figure 2.2: Comparison of sensitivity between MAGGIE and other approaches on simulated datasets. Each boxplot aggregates the significance values from 10 simulations. Boundary lines show the median and quartiles of each distribution. Every simulation generated a thousand sequences inserted with a pair of motifs, which serve as the positive set. 10–80% of these sequences had a single nucleotide changed within the SPI1 motif for SPI1-CEBPB pair, or the POU5F1::SOX2 motif for POU5F1::SOX2-KLF4 pair or the ZNF410 motif for ZNF410-IRF1 pair, whereas the rest were kept untouched, resulting in the negative set. The dashed lines indicate the significance threshold after multiple testing correction.

To assess the sensitivity of MAGGIE, we tested its performance when different fractions of sequences experienced motif mutations (i.e., SNR). For every SNR ranging from 10% to 80%, we repeated simulation of sequences 10 times and aggregated P-values for embedded motifs from the comparative methods. Here, we also assessed the performance of the motif enrichment algorithm implemented by HOMER in detecting motif mutations by setting positive sequences as inputs and negative sequences as backgrounds. We evaluated the comparative methods on three

arbitrary pairs of motifs: SPI1-CEBPB, POU5F1::SOX2-KLF4 and ZNF410-IRF1. For each experiment, one motif pair was inserted into sequences, but only the first motif (SPI1, POU5F1::SOX2 and ZNF410) was mutated by synthetic genetic variation. MAGGIE consistently outperforms the other methods in identifying the mutated motif (Fig. 2.2) and not the unmutated motif (Supplementary Fig. 2.4). Even though the other methods could potentially pass the significance threshold with a higher SNR or an increasing sample size (Supplementary Fig. 2.5), MAGGIE showed superior sensitivity when motifs are mutated in <50% of the finite samples.

2.4.2 MAGGIE identifies known mediators for TF binding sites and QTLs

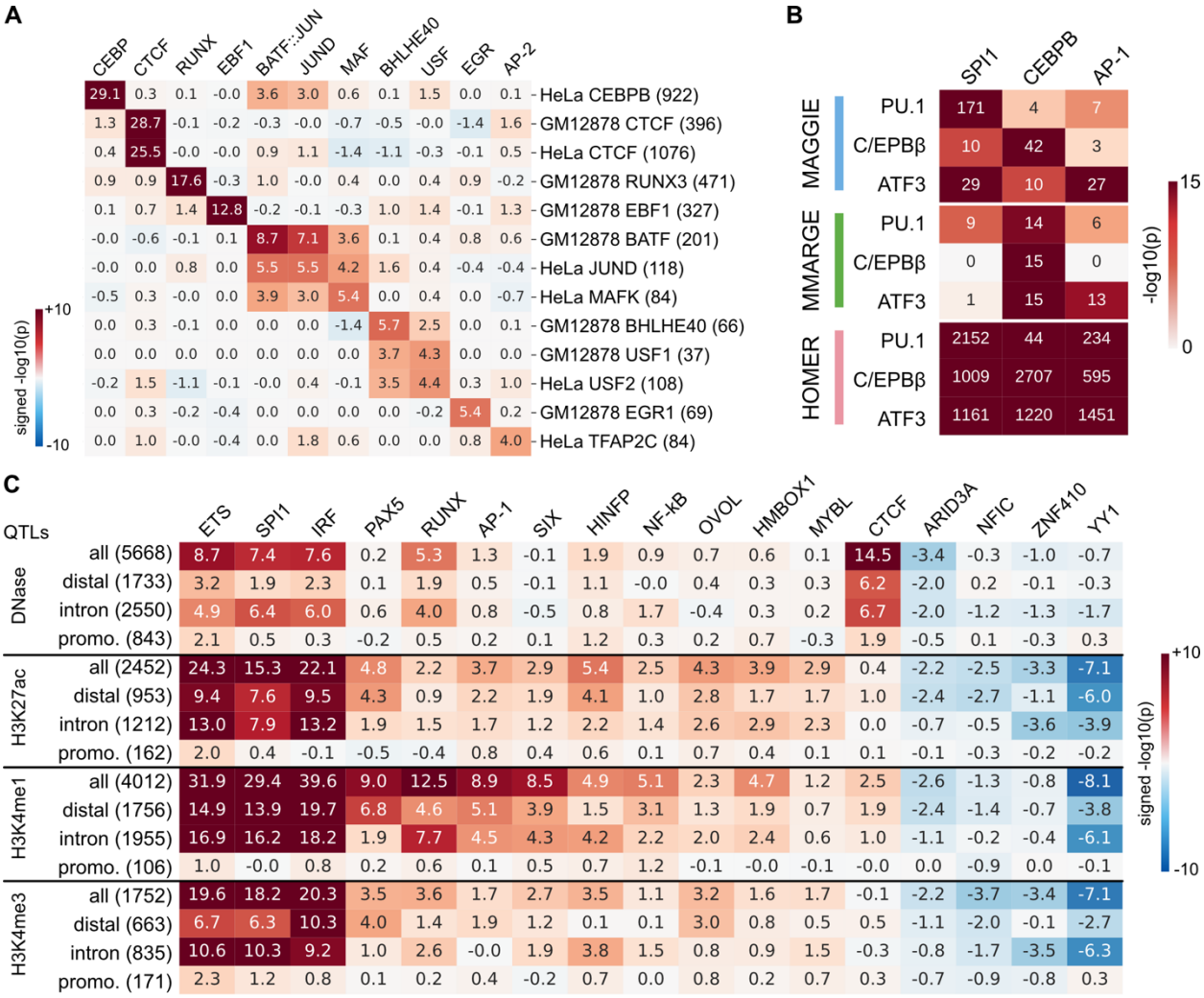


Figure 2.3: Functional motifs identified by MAGGIE for various epigenomic features using biological datasets. (A) Signed P-values for allele-specific TF binding sites. In total, 13 datasets were analyzed covering 12 different TFs from two cell types: GM12878 and HeLa-S3 (HeLa). Datasets are arranged vertically with their sample sizes displayed in brackets, and motifs are shown horizontally on top by their gene names. **(B)** Comparative results for strain-specific TF binding sites. P-values from different motif analysis methods are shown for the corresponding motifs of the three LDTFs (PU.1, C/EBPb and ATF3). **(C)** Significant motifs identified for chromatin QTLs of LCLs. Signed P-values from MAGGIE are shown for the entire sets as well as the subsets based on the locations of QTLs. The number of QTLs in each set is shown in brackets. Motifs shown here were tested significant for at least one type of the QTLs. All the results in this figure have been averaged for similar motifs and are displayed by their family names (e.g., ETS, AP-1).

After observing the superior performance of MAGGIE on simulated data, we tested our method with several biological datasets. First, we analyzed the allele-specific TF binding sites associated with SNPs (Shi et al., 2016). Among the 13 experiments tested, MAGGIE identified

the corresponding motifs of the bound TFs for all of them (Fig. 2.3A). Even though P-values vary due to the quality and the sample size of each dataset, the corresponding motifs were recognized as the most significant even for TFs with as few as 37 allele-specific binding sites like USF1.

Next, we evaluated whether MAGGIE is able to recover the collaborative binding between TFs. Regulatory elements are usually bound by multiple TFs together, which form a complex with other co-activators to regulate functions (Reiter et al., 2017). For example, lineage-determining TFs (LDTFs) of macrophages such as PU.1, C/EBP, and AP-1 factors were frequently found to co-bind at macrophage-specific enhancers (Glass & Natoli, 2016; Heinz et al., 2015). Previous studies showed that the binding of specific LDTFs was not only dependent on each factor's own motif, but also on nearby motifs recognized by collaborative factors (Heinz et al., 2013; Link, Duttke, et al., 2018). To verify this conclusion with MAGGIE, we downloaded ChIP-seq data for PU.1 (encoded by SPI1), C/EBPb (encoded by CEBPB) and ATF3, which binds to AP-1 motif, from two genetically diverse strains of mice, C57 and BALB (Fonseca et al., 2019; Link, Duttke, et al., 2018). Strain-specific TF binding sites were identified for each factor and analyzed with MAGGIE. As comparison, we used MMARGE to find motifs correlated with TF binding activity quantified by ChIP-seq read counts and used HOMER to find enriched motifs among positive sequences in comparison to random backgrounds. The outputs from the three approaches for the relevant motifs are summarized in Fig. 2.3B. MAGGIE recognized PU.1 binding to be mostly dependent on its own motif instead of any other motif, while C/EBPb binding was highly dependent on CEBPB motif but also significantly dependent on SPI1 motif. The results were consistent with the pioneer role of PU.1 in opening chromatin and guiding the binding of other TFs (Barozzi et al., 2014). On the contrary, the comparative

methods failed to distinguish the different functions between the bound TF and its collaborative factors. HOMER assigned strong significance to all the motifs because it was designed to identify enriched motifs without considering functions. MMARGE showed a lack of power in detecting collaborative factors using the data of two mouse strains as it requires more data or larger genetic difference to confidently identify a correlation between motif and TF binding.

The general framework of MAGGIE can also be applied to QTL datasets for epigenomic features that are influenced by TF binding. We downloaded QTLs of several epigenomic features for LCLs, including dsQTLs for chromatin accessibility (Degner et al., 2012) and hQTLs for three types of histone modifications, H3K27ac, H3K4me1 and H3K4me3 (Grubert et al., 2015). MAGGIE identified motifs with different specificity for the testing features (Fig. 2.3C). CTCF was output at top for dsQTLs but was insignificant for each type of hQTLs, supporting the major role of CTCF in maintaining chromatin structures instead of inducing active chromatin states (Arzate-Mejía et al., 2018). PU.1 together with other ETS factors were significant for both chromatin accessibility and histone modifications, indicating a pioneer role in opening chromatin as well as an important role in activating regulatory elements in LCLs (Scott et al., 1994). MAGGIE also identified many other motifs important for histone modifications, which have been found to maintain the cell identity and function of LCLs from previous studies, such as PAX5 (Glimcher & Singh, 1999), RUNX (Mevel et al., 2019) and NF-kB (Nagel et al., 2014). It is intriguing that several motifs showed up with potentially inhibitive functions, although these will need to be confirmed in later studies.

2.4.3 MAGGIE captures divergent functions of NF-kB factors for the stimulus responses of regulatory elements

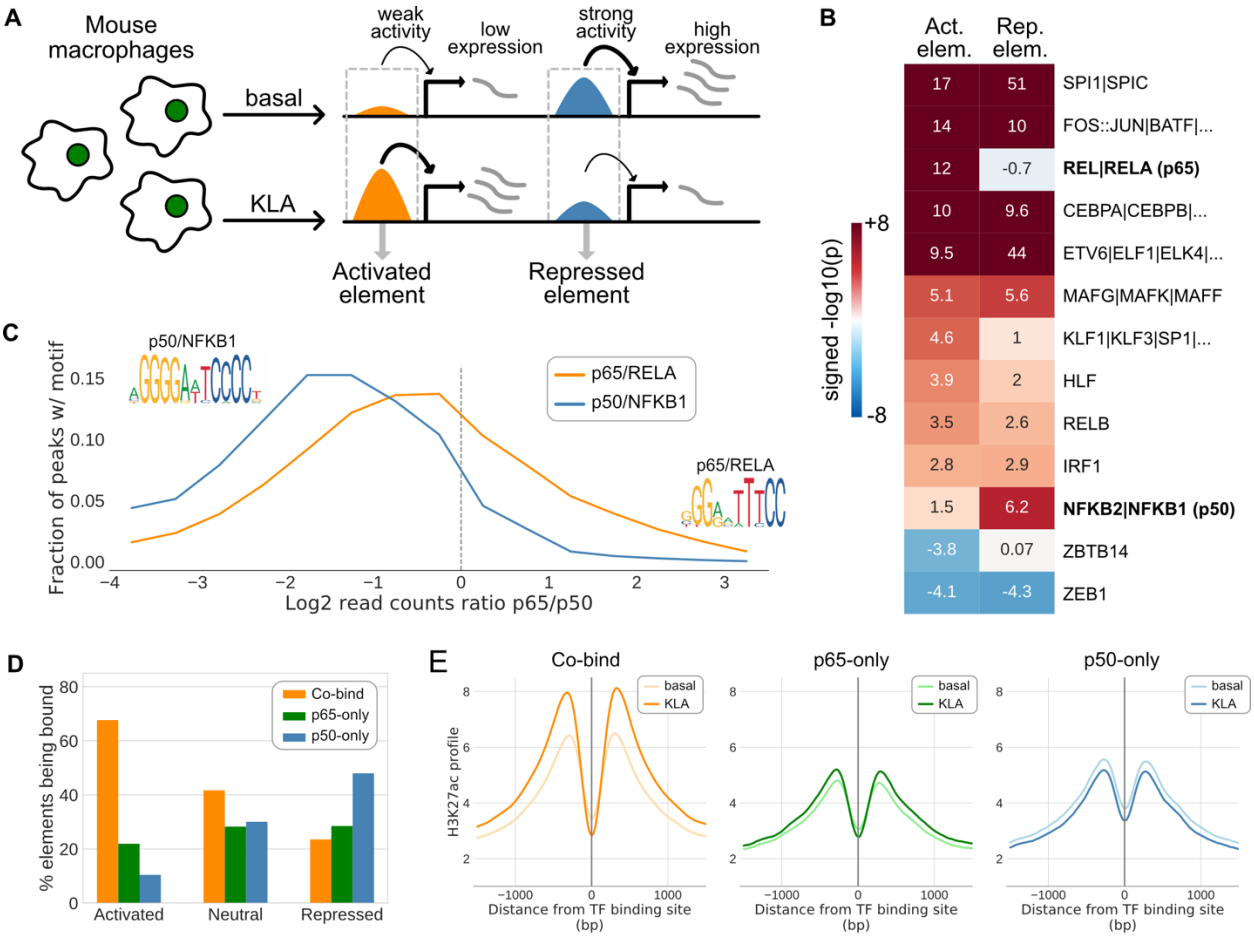


Figure 2.4: Divergent functions of NF-kB factors in pro-inflammatory macrophages captured by MAGGIE and validated by experiments. (A) Sketch of KLA-activated and KLA-repressed regulatory elements defined by >2.5-fold changes of H3K27ac from basal to KLA-treated conditions. **(B)** Significant functional motifs identified by MAGGIE for activated and repressed regulatory elements. Similar motifs are separated by '|' and shown with their average results. Protein names of RELA and NFKB1 are shown in the brackets, corresponding to p65 and p50, respectively. **(C)** Binding activities of NF-kB factors associated with RELA or NFKB1 motifs. Motifs were searched within the 200-bp binding sites of p65 and p50 at the KLA-treated condition measured by ChIP-seq experiments. ChIP-seq reads for both p65 and p50 were counted to quantify binding activities. Regions with at least 32 reads of either factor were used to compute the log2 ratio of reads between p65 and p50. The distributions of log ratios are displayed in orange for sites having RELA motif (10,549 sites) and in blue for sites having NFKB1 motif (2,144 sites). The logos of motif PWMs are demonstrated. **(D)** Co-existence of NF-kB binding and the KLA responses of regulatory elements. NF-kB binding sites were separated into sites bound by p65 alone (p65-only), p50 alone (p50-only) or both (Co-bind). Among the regulatory elements that overlap with NF-kB binding sites, the bar plots summarized the fractions of elements bound by different NF-kB factors for activated, neutral, and repressed elements. **(E)** Change of H3K27ac at NF-kB binding sites after KLA treatment. H3K27ac ChIP-seq reads were counted within +/-1500 bp around the three categories of NF-kB binding sites using a bin size of 25 bp and were averaged to show the overall change of H3K27ac profiles.

Next, we tested MAGGIE with a more complex epigenomic feature: stimulus responses of regulatory elements. ATAC-seq and H3K27ac ChIP-seq data from four genetically diverse strains of mice were downloaded for macrophages at basal conditions and at proinflammatory conditions induced by 1-h treatment of KLA (Link, Duttke, et al., 2018). We used ATAC-seq data to locate open chromatin regions accessible for TF binding and H3K27ac ChIP-seq data to quantify the activity of these regions and identify active regulatory elements (Supplementary Fig. 2.6). By filtering for 2.5-fold change of activity from basal to KLA-treated conditions, we identified KLA-activated and KLA-repressed regulatory elements for each mouse strain (Fig. 2.4A). Among those, about 12,000 activated elements and 18,000 repressed elements were specific to one of the strains based on pairwise comparisons. Strain-specific activated and repressed regulatory elements were separately tested by MAGGIE to identify their mediators. Interestingly, besides SPI1, CEBP and AP-1 (e.g., FOS::JUN) motifs that were known to be important for the KLA responses of macrophages (Glass & Natoli, 2016), MAGGIE assigned divergent functions for NF-kB factors (Fig. 2.4B). RELA corresponding to p65 subunit was output as functional for the activation response, while NFKB1 corresponding to p50 subunit was found significant for the repressed elements. On the contrary, due to the similarity of these motifs (Supplementary Fig. 2.7), HOMER found both motifs enriched in the activated elements compared with random backgrounds and neither enriched in the repressed elements (Supplementary Fig. 2.8). Previous studies have shown that p65 frequently forms heterodimers with p50 to act as a transcriptional activator, while p50 homodimers result in transcriptional repression (Brignall et al., 2019; Cheng et al., 2011; Natoli et al., 2005). However, the genome-wide functions and binding patterns of these factors remain unknown.

To validate the functions of p65 and p50 for the KLA responses of macrophages, we conducted ChIP-seq experiments in C57 mice for p65 and p50 to measure their genome-wide binding sites in KLA-treated macrophages. Based on the measured TF binding sites, we first investigated the binding patterns of these factors. We searched for sites with RELA or NFKB1 motifs and computed the binding activities of NF-kB factors at those sites by counting ChIP-seq reads. Regions with relatively strong binding of either factor (>32 ChIP-seq reads of p65 or p50; Supplementary Fig. 2.9) were used to calculate the log ratios of read counts between p65 and p50 (Fig. 2.4C). RELA motif was enriched at the co-binding sites of p65 and p50, while NFKB1 motif was more strongly bound by p50. To connect the binding patterns to the regulatory elements used in MAGGIE, we overlapped the TF binding sites with the activated and repressed elements previously defined for C57 mice and found that the majority of activated elements were co-bound by both p65 and p50, while repressed elements were more often bound by p50 alone (Fig. 2.4D). By quantifying the regulatory activity around the binding sites of p65 and p50 by the level of H3K27ac, we found an overall decrease in H3K27ac around sites only bound by p50 and an overall increase around the co-binding sites of p65 and p50 after KLA treatment (Fig. 2.4E). These findings suggest a genome-wide role of p65–p50 heterodimers as a transcriptional activator and p50 homodimers as a repressor for KLA-treated macrophages. More importantly, our experimental results validated the predictions from MAGGIE regarding the divergent functions of p65 and p50 subunits for pro-inflammatory macrophages, showing promise of using MAGGIE to discover novel functions of TFs for complex epigenomic features.

2.5 Discussion

To our knowledge, MAGGIE is the first work to associate the mutation of TF binding motif with various types of epigenomic features. In contrast to motif enrichment methods, such as HOMER, in which identified motifs may or may not be functionally related to epigenomic features, MAGGIE determines the significance of motifs based on putative functional consequences of local motif mutations. Due to this qualitative difference, MAGGIE and motif enrichment methods recover overlapping but non-identical sets of significant motifs. As the major difference in methodology, MAGGIE focuses on the change of motif score and intentionally ignores the actual motif score due to the strong correlation between motif mutation and change of TF binding (Fig. 2.1C) and the independency of this relationship from the actual motif score (Supplementary Fig. 2.1). Another reason not to incorporate the actual motif score is that many epigenomic features do not possess a simple relationship with motif score. Instead of assuming a linear relationship between motif scores and epigenomic features like many current methods do, MAGGIE tests for a bias in the changing direction of motif mutations. We demonstrated that MAGGIE is able to identify known functional motifs for TF binding (Fig. 2.3A and B), chromatin accessibility (Fig. 2.3C), and histone modification (Fig. 2.3C). MAGGIE also helped to discover divergent functions of distinct NF-kB factors for the KLA response of regulatory elements in macrophages (Fig. 2.4), which was not found by any other motif analysis tools. It is worth noting that the motifs of NF-kB factors are usually too similar to be distinguished by motif enrichment methods, but the strategy of focusing on the change of motif score instead of the actual motif score is sensitive enough to capture the difference.

MAGGIE takes binary-labeled datasets as inputs (i.e., positive and negative sequences), which facilitates application to most publicly available data, including aggregated datasets like

QTLs and processed data from sequencing experiments like ChIP-seq and ATAC-seq. However, for the framework to work, MAGGIE requires additional measurements and genetic variation information for at least two different genotypes, which may not be currently available for some biological problems. The primary limitation to the discovery power of MAGGIE is the degree of genetic variation provided by the samples being analyzed. Another limitation is the inevitable cutoff accompanied with binary labels, which might affect the results especially when there are concerns about insufficient sample size or low data quality.

The flexibility of our statistical framework makes it applicable to any type of epigenomic feature that is potentially affected by TF binding. Given the stand-alone tool we provided for the motif analysis methods described here, it will be interesting to investigate the performance of MAGGIE in other features, such as chromatin interaction and DNA methylation. Another future extension is to switch the PWM score used in this study to other types of motif scores, such as more advanced representations of TF binding motifs based on hidden Markov models. It will also be promising to incorporate state-of-the-art machine learning approaches (e.g., deep learning) into our framework to consider complex interactions between motifs. For instance, we can integrate the prediction scores of variant impacts from deep learning models and associate those predictions with biased changes of motifs.

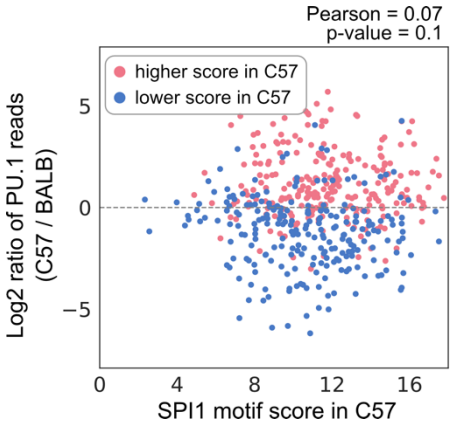
In summary, we presented a novel method for identifying DNA sequence motifs mediating TF binding and function, which goes beyond enriched or correlated motifs that are frequently analyzed nowadays. Given the growing interest in the function of TFs and the unprecedented generation of epigenomic data for different individuals and animal strains, we expect MAGGIE to be an effective bioinformatic tool that can be included in the regular routine of motif analysis.

2.6 Acknowledgements

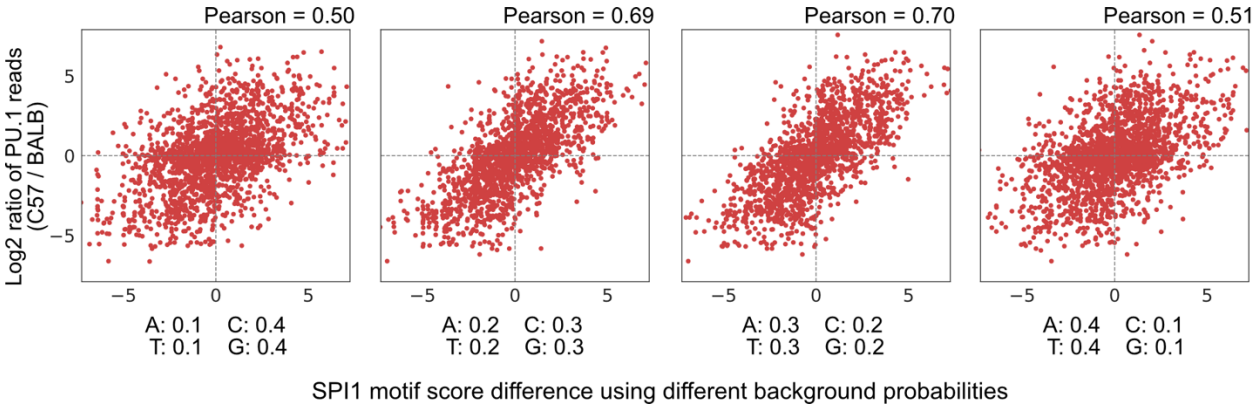
I would like to express my great appreciation to Melissa Gymrek, Jenhan Tao and Ludmil B. Alexandrov for friendly reviews. Our special thanks are extended to Inge R. Holtman for beta testing and feedback on the software package, and Jana Collier for technical assistance. This work was supported by the following grants: National Institutes of Health/National Institute of Diabetes and Digestive and Kidney Diseases R01 DK091183 and Foundation Leducq Grant 16CVD01.

Chapter 2, in full, is a reprint of the material as it appears in Shen Z, Hoeksema MA, Ouyang Z, Benner C & Glass CK. (2020). MAGGIE: leveraging genetic variation to identify DNA sequence motifs mediating transcription factor binding and function. Bioinformatics. 36(Supplement_1). The dissertation author was the primary investigator and author of this paper.

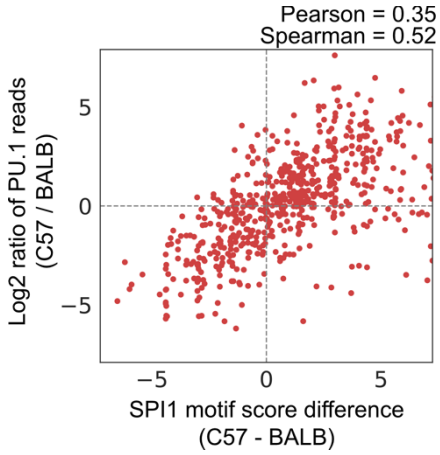
2.7 Supplementary figures



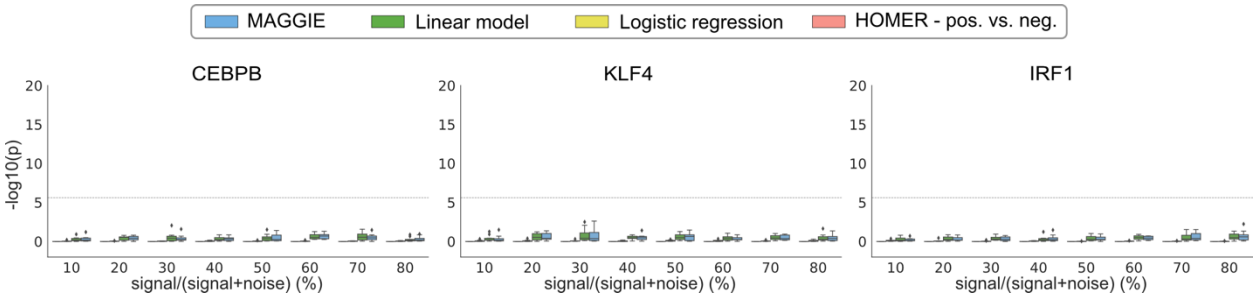
Supplementary Figure 2.1 Relationship between change of PU.1 binding and SPI1 motif score at a similar level of SPI1 motif mutation. Each dot represents a PU.1 binding site that has mutation on SPI1 motif between BALB and C57 by a difference of motif score between 1 and 1.5. Red dots are binding sites with a higher motif score in C57, while blue dots are for sites with lower scores in C57. Change of PU.1 binding activity was calculated by the fold change of PU.1 ChIP-seq reads between BALB mice and C57 mice. Sites with a stronger PU.1 motif in BALB has an increase in PU.1 binding in general, but the level of binding increase is not affected by the actual motif score (Pearson coefficient = 0.07 with an insignificant p-value).



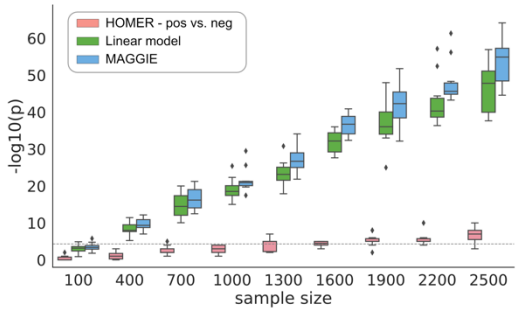
Supplementary Figure 2.2 Diminished correlations between motif score differences of SPI1 motif and fold changes of PU.1 binding activity using non-uniform background probabilities. Background probabilities are displayed under each plot with GC contents ranging from 20% (rightmost) to 80% (leftmost).



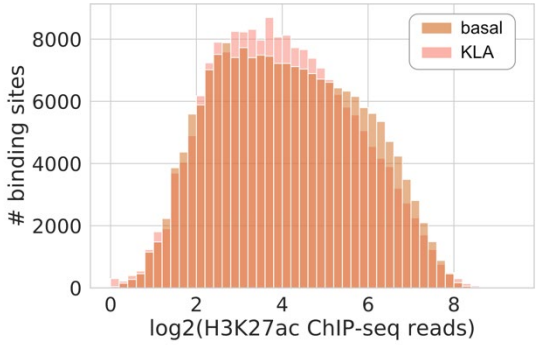
Supplementary Figure 2.3 Diminished correlation between motif score differences of SPI1 motif and fold changes of PU.1 binding activity using motif at the same locations. By restricting motifs at the same locations, we ended up with fewer PU.1 binding sites having mutations on SPI1 motif.



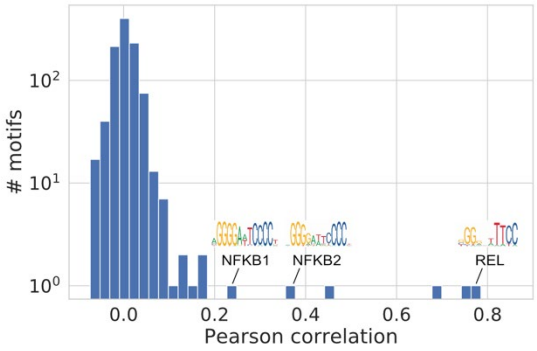
Supplementary Figure 2.4 Significance values of simulated experiments from the comparative approaches for the unmutated motifs. All of the methods recognize unmutated motifs as insignificant for all levels of signal-to-noise ratio.



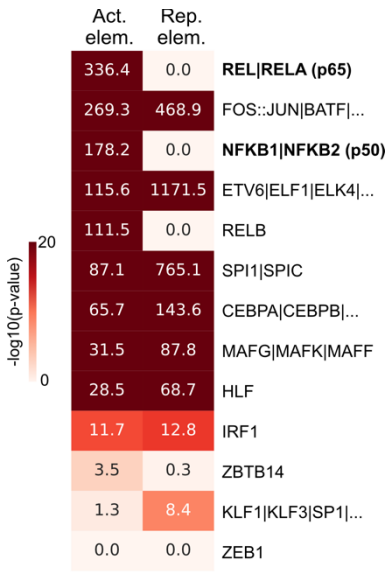
Supplementary Figure 2.5 Effect of sample size on the outputs from the comparative approaches. Different number of input sequences were simulated and embedded with SPI1 and CEBPB motifs, among which 50% of total sequences experienced mutation on the SPI1 motifs. Ten simulations were repeated for each sample size. Overall, significance values increased with a larger sample size for all the methods.



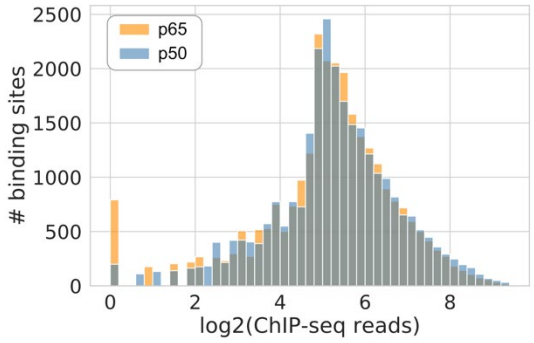
Supplementary Figure 2.6 Distribution of H3K27ac ChIP-seq reads for extended open chromatin of macrophages at basal and KLA-treated conditions. ChIP-seq reads were counted within 1000-bp extended regions around open chromatin regions identified from ATAC-seq. Read counts were pooled for the four testing strains of mice. Regions with larger than 16 reads are within the more active half of the total regions.



Supplementary Figure 2.7 Correlations of motif score differences between RELA motif and other testing motifs. NFKB1 and NFKB2 motif have relatively better correlations with RELA motif due to their motif similarity. However, it is not comparable to the extreme similarity between REL and RELA motif, which has a Pearson correlation coefficient larger than 0.7. Motif score differences are based on strain-specific KLA-activated regulatory elements.



Supplementary Figure 2.8 Motif enrichment results from HOMER for KLA-activated and KLA-repressed regulatory elements of C57 mice. Motif enrichment was calculated by comparing to random backgrounds with default parameters. The motifs are ranked by the enrichment p-values for activated elements.



Supplementary Figure 2.9 Distribution of ChIP-seq reads for p65 and p50 at their respective binding sites. ChIP-seq reads were counted within 200-bp binding sites called for p65 and p50 using HOMER "findpeaks -size 200". The median counts for both factors are roughly at 5 = log2(32 reads).

Chapter 3. Mechanisms underlying divergent responses of genetically distinct macrophages to IL-4

3.1 Abstract

Mechanisms by which non-coding genetic variation influences gene expression remain only partially understood but are considered to be major determinants of phenotypic diversity and disease risk. Here, we evaluated effects of >50 million SNPs and InDels provided by five inbred strains of mice on the responses of macrophages to interleukin 4 (IL-4), a cytokine that plays pleiotropic roles in immunity and tissue homeostasis. Remarkably, of >600 genes induced >2-fold by IL-4 across the five strains, only 26 genes reached this threshold in all strains. By applying deep learning and motif mutation analyses to epigenetic data for macrophages from each strain, we identified the dominant combinations of lineage determining and signal-dependent transcription factors driving IL-4 enhancer activation. These studies further revealed mechanisms by which non-coding genetic variation influences absolute levels of enhancer activity and their dynamic responses to IL-4, thereby contributing to strain-differential patterns of gene expression and phenotypic diversity.

3.2 Introduction

Non-coding genetic variation is a major driver of phenotypic diversity as well as the risk of a broad spectrum of diseases. For example, of the common single nucleotide polymorphisms (SNPs) and short insertions/deletions (InDels) identified by genome-wide association studies (GWAS) to be linked to specific traits or diseases, ~90% are typically found to reside in non-coding regions of the genome (Farh et al., 2015). The recent application of genome-wide approaches to define the regulatory landscapes of many different cell types and tissues allows intersection of such variants with cell-specific regulatory elements and strongly supports the

concept that alteration of transcription factor binding sites at these locations is an important mechanism by which they influence gene expression (Kilpinen et al., 2013; van der Veeken et al., 2019; Vierstra et al., 2020). Despite these major advances, it remains difficult to predict the consequences of most forms of non-coding genetic variation. Major challenges that remain include defining the causal variant within a block of variants that are in high linkage disequilibrium, identifying the gene that is regulated by the causal variant, and understanding the cell type and cell state specific regulatory landscape in which a variant might have a functional consequence (Abascal et al., 2020). For example, a variant that affects the binding of a signal-dependent transcription factor (SDTF) may only be of functional importance in a cell that is responding to a signal that activates that factor (Soccio et al., 2015). Also, sequence variants can have a range of effects on transcription factor binding motifs, from abolishing or inducing binding by affecting critical nucleotides to quantitatively changing binding by affecting an intermediate affinity motif (Behera et al., 2018; Deplancke et al., 2016; Grossman et al., 2017).

Studies of the impact of natural genetic variation on signal-dependent gene expression have demonstrated large differences in absolute levels of gene expression under basal and stimulated conditions, which result in corresponding differences in the dynamic range of the response (Bakker et al., 2018; Fairfax et al., 2014; Gate et al., 2018). The molecular mechanisms by which genetic variation results in these qualitatively and quantitatively different signal-dependent responses remain poorly understood but are likely to be of broad relevance to understanding how non-coding variation influences responses to signals that regulate development, homeostasis and disease-associated patterns of gene expression.

To investigate the influence of genetic variation on signal-dependent gene expression, we performed transcriptomic and epigenetic studies of the responses of macrophages derived from

five different inbred mouse strains to the anti-inflammatory cytokine IL-4 (Fig. 3.1A). The selected strains include both similar as well as highly divergent strain pairs, allowing modeling of the degree of variation between two unrelated individuals (~4 million variants) and that observed across large human populations (>50 million variants). Using this approach, we previously showed that strain-specific variants that disrupt the recognition motif for one macrophage lineage determining transcription factor (LDTF, e.g., PU.1), besides reducing binding of the LDTF itself, also result in decreased binding of other collaborative factors and SDTFs (Heinz et al., 2013; Link, Duttke, et al., 2018). Collectively, these findings supported a model in which relatively simple combinations of LDTFs collaborate with an ensemble of additional transcription factors to select cell-specific enhancers that provide sites of action of broadly expressed SDTFs (Heinz et al., 2010).

IL-4 has many biological roles, including regulation of innate and adaptive immunity (Gieseck et al., 2018). In macrophages, IL-4 drives an 'alternatively activated' program of gene expression associated with inhibition of inflammatory responses and promotion of wound repair (Gordon & Martinez, 2010). The immediate transcriptional response to IL-4 is mediated by activation of STAT6 (Goenka & Kaplan, 2011; Ostuni et al., 2013), which rapidly induces the expression of direct target genes that include effector proteins such as Arginase 1 (*Arg1*) and transcription factors like PPARγ (Daniel et al., 2018) and EGR2 (Daniel et al., 2020). However, the extent to which natural genetic variation influences the program of alternative macrophage activation has not been systematically evaluated. Here, we demonstrate highly differential IL-4 induced gene expression and enhancer activation in bone marrow-derived macrophages (BMDMs) across the five mouse strains, thereby establishing a robust model system for quantitative analysis of the effects of natural genetic variation on signal-dependent gene

expression. Through the application of deep learning methods and motif mutation analysis of strain-differential IL-4 activated enhancers, we provide functional evidence for a dominant set of LDTFs and SDTFs required for late IL-4 enhancer activation, which include STAT6, PPARγ and EGR2, and validate these findings in *Egr2*-knockout BMDMs. Importantly, assessment of the quantitative effects of natural genetic variants on recognition motifs for LDTFs and SDTFs suggests general principles by which such variation affects enhancer activity patterns and dynamic signal responses.

3.3 Results

3.3.1 The response to IL-4 is highly variable in BMDMs from genetically diverse mice

To investigate how natural genetic variation affects the macrophage response to IL-4, we began by performing RNA-seq in BMDMs derived from female BALB/cJ (BALB), C57BL/6J (C57), NOD/ShiLtJ (NOD), PWK/PhJ (PWK) and SPRET/EiJ (SPRET) mice under basal conditions and following stimulation with IL-4. Time course experiments in C57 BMDMs indicated a progressive increase in the number of differentially expressed genes from 1 to 24 hours (Supplementary Fig. 3.1A-B). We therefore focused our analysis on the response to IL-4 in BMDMs from the five strains at this timepoint. Weighted Co-expression Network Analysis (WGCNA) identified numerous modules of highly correlated mRNAs, the majority of which were driven by strain differences (Fig. 3.1B). Genes that were positively regulated by IL-4 across strains (red module, bottom) were enriched for functional annotations related to negative regulation of defense responses. Conversely, the purple (top) module captured genes that were negatively regulated by IL-4 and were enriched for pathways associated with positive regulation of inflammation (Fig. 3.1B).

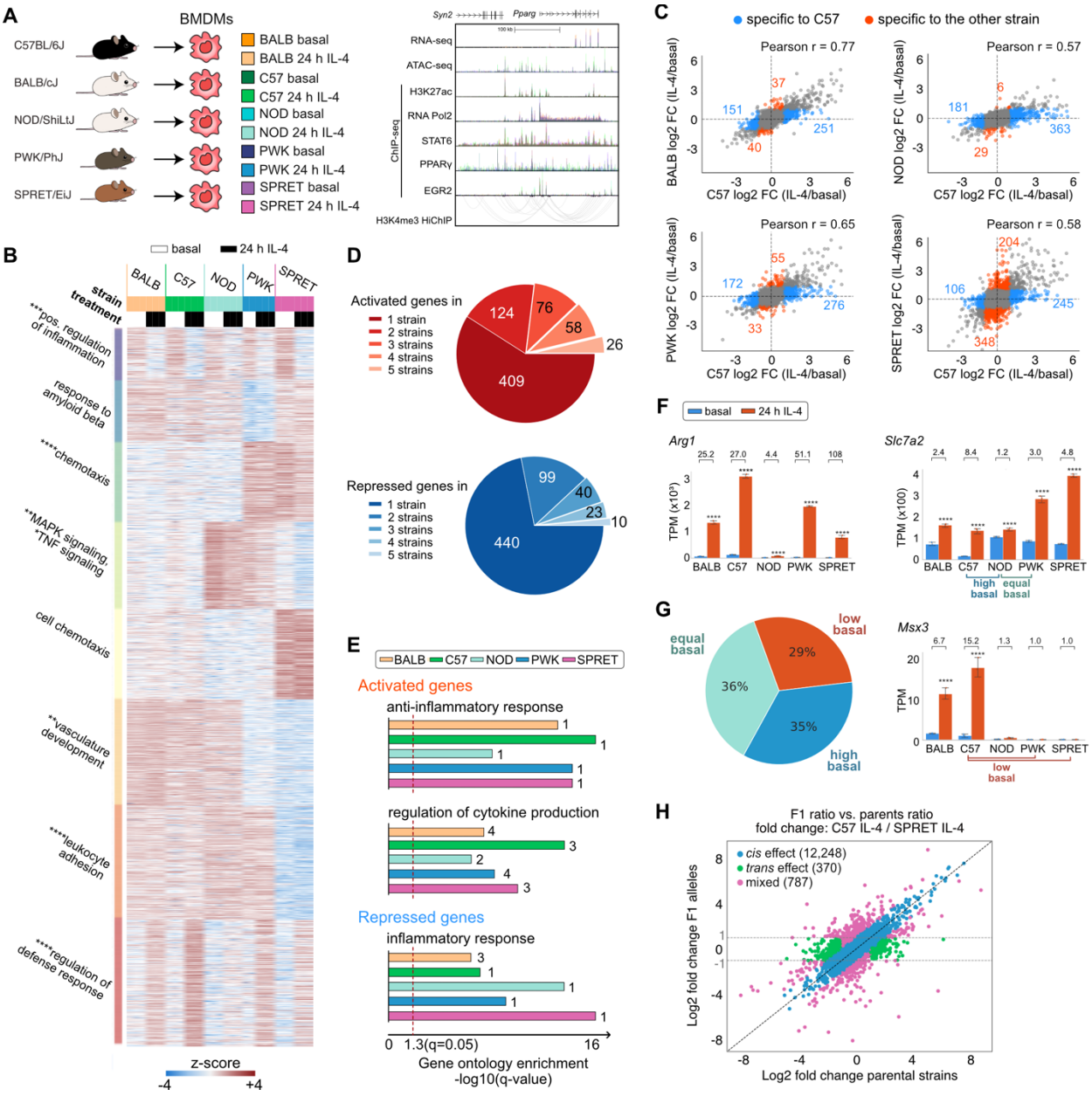


Figure 3.1 Response to IL-4 is highly divergent in BMDMs from different mouse strains. (A) Overview of experimental design and main data sets. **(B)** WGCNA clustering focused on strain-differentially regulated genes in IL-4 treated BMDMs. The top hit Metascape pathways are annotated for each module. *q<0.05, **q<0.01, ***q<0.001, ****q<0.0001. **(C)** Ratio-ratio plots demonstrating the mRNA response to IL-4 in pairwise comparisons. **(D)** Overlap of genes significantly induced or repressed (q<0.05, >2-fold) after IL-4 treatment in BMDMs from all strains. **(E)** Gene ontology terms enriched in up- and down-regulated genes after 24 h IL-4 stimulation in BMDMs from all strains. Numbers indicate the rank order in pathway analysis. **(F)** *Arg1*, *Slc7a2* and *Msx3* as example genes differentially up-regulated by IL-4 in strains. TPM, transcripts per kilobase million. ****q<0.0001, compared to basal. Numbers indicate fold change by IL-4. **(G)** Categories of strain-differential IL-4 up-regulated genes based on the differences in basal gene expression. **(H)** Average log2 gene expression fold change between alleles in hybrid (C57xSPRET F1) and parental strain under 24 h IL-4 conditions.

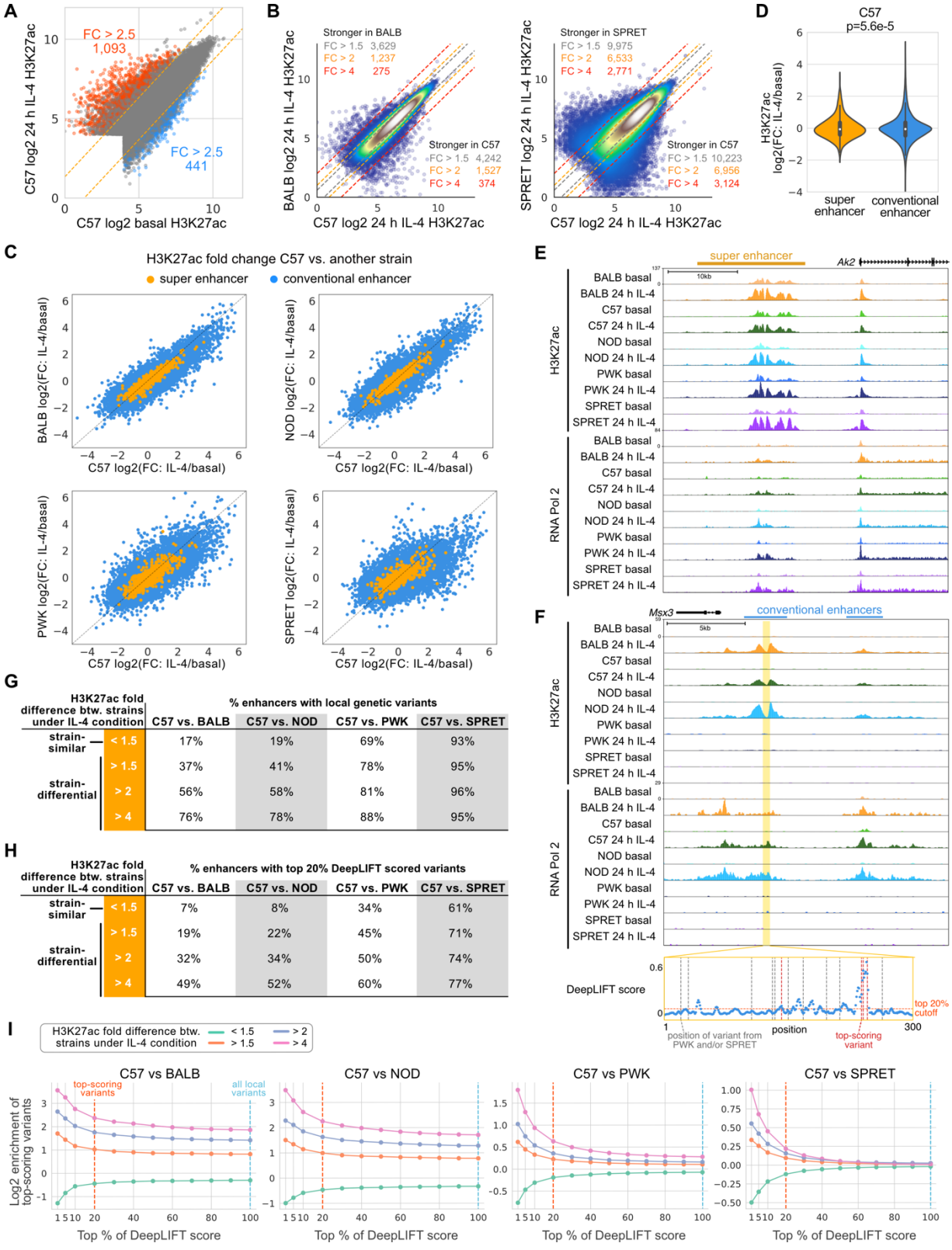
Remarkably, of the 693 genes induced >2-fold in at least one strain, only 26 (3.75%) were induced at this threshold in all five strains (Fig. 3.1D, Supplementary Fig. 3.1C-D). Conversely, more than half of the IL-4-responsive genes identified were induced >2-fold in only a single strain. NOD BMDMs were notable for a generally attenuated response to IL-4 (Fig. 3.1B, red module, 1C, second panel). A similar pattern was observed for down-regulated genes (Fig. 3.1D). Despite these differences at the level of individual genes, similar pathways/gene programs were enriched in all strains for both induced and repressed genes (Fig. 3.1E). Substantial differences in IL-4 target gene expression across strains are illustrated by *Arg1*, *Slc7a2* and *Msx3* (Fig. 3.1F, fig S1E). BMDMs from all strains exhibit a significant induction of *Arg1* expression, but the absolute basal levels and induction folds vary by more than an order of magnitude. *Slc7a2* exhibits similar levels of expression in C57 and NOD BMDMs after IL-4 treatment, but its differences at the basal level result in an 8.4-fold and 1.2-fold change, respectively. We refer to the pattern of reduced responsiveness to IL-4 in this comparison of C57 and NOD as being associated with 'high basal' activity in the less responsive strain. Conversely, NOD and PWK BMDMs exhibit similar levels of basal *Slc7a2* expression, but IL-4 only increased *Slc7a2* expression more than 2-fold in PWK. We refer to this pattern of reduced responsiveness to IL-4 in NOD compared to PWK as being associated with 'equal basal' activity. A third category is exemplified by *Msx3*, which is induced in C57 but not in PWK and SPRET BMDMs. In this case, lack of responsiveness is associated with low expression of *Msx3* under basal conditions. We refer to this pattern as 'low basal' in the less responsive strain. Quantitative analyses of pair-wise comparisons indicate that 29% of the genes with decreased IL-4 induced gene expression were due to low basal expression, 36% had no differences prior to

IL-4 stimulation (equal basal), and 35% were the result of a high basal expression level in the less responsive strain (Fig. 3.1G).

To investigate local versus distant effects of genetic variation on the differential responses to IL-4, we crossed C57 mice with the most genetically distinct SPRET mice to generate F1 offspring containing each parental chromosome. 91.4% of parental-specific RNA-seq reads in the F1 strain are within 2-fold of their values in C57 and SPRET (blue data points) and considered to be due to local (*cis*) effects of genetic variation (Fig. 3.1H, Supplementary Fig. 3.1F), while only 2.8% was divergent between the parental strains but not in F1 BMDMs (green data points), indicating *trans* regulation. As NOD macrophages exhibited a broadly attenuated response to IL-4 on the level of gene expression, we followed the same strategy using F1 C57xNOD macrophages. Interestingly, RNA-seq on IL-4 stimulated macrophages of F1 C57xNOD macrophages showed strong convergence of expression of genes that were differentially regulated in the parental strain (green data points in Supplementary Fig. 3.1G), consistent a major contribution of *trans* regulation. To investigate the point at which this regulation occurs, we performed ChIP-Seq for RNA-Pol2 under control and IL-4 stimulated conditions. In contrast to mRNA levels, examination of the IL-4-depenent changes in gene body RNA Pol2 indicated similar magnitude changes in all strains, including NOD (Supplementary Fig. 3.1H). These results suggest the presence of a transacting factor in NOD that acts downstream of transcription to attenuate mRNA levels. Collectively, these studies uncovered striking variation in the cell autonomous responses of BMDMs to IL-4 across these five strains, providing a powerful experimental system for investigating mechanisms by which natural genetic variation impacts signal-dependent gene expression.

3.3.2 Strain-differential IL-4 induced gene expression is associated with differential IL-4 enhancer activation

Figure 3.2 Divergent IL-4 response is associated with strain-differential IL-4 enhancer activation. (A) Log2 H3K27ac signal at ATAC peaks in C57 BMDMs under basal and IL-4 conditions. **(B)** Comparison of H3K27ac signal between C57 and BALB or SPRET under the 24 h IL-4 condition. **(C)** Log2 H3K27ac fold changes after 24 h IL-4 in C57 versus other strain in enhancers. **(D)** Distributions of IL-4 H3K27ac log2 fold changes, Levene's test was performed to test response differences in conventional versus super enhancers. **(E)** *Ak2* super enhancer responsive to IL-4 and conserved across all strains. **(F)** *Msx3* IL-4 induced enhancer in C57, BALB and NOD, but not PWK and SPRET BMDMs. Absolute DeepLIFT scores indicate predicted importance of single nucleotides for enhancer activity. Dotted lines represent locations of PWK or SPRET variants. **(G, H)** Enhancers were categorized into strain-similar and strain-differential based on fold differences in H3K27ac between C57 and one of the other strains. Table with percentages of enhancers containing local genetic variants **(G)** and the percentage of enhancers that contain predicted functional variants **(H)**. **(I)** Log2-scaled enrichment of enhancers with variants at top-scoring positions based on DeepLIFT scores. The enrichment was calculated by (% enhancers in one category with top variants) / (% all enhancers with top variants). **G** and **H** are based on the top 100% and 20%, respectively.


A
FC > 2.5
1,093
FC > 2.5
441
C57 log2 24 h IL-4 H3K27ac
C57 log2 basal H3K27ac
B
Stronger in BALB
FC > 1.5 3,629
FC > 2 1,237
FC > 4 275
Stronger in C57
FC > 1.5 4,242
FC > 2 1,527
FC > 4 374
BALB log2 24 h IL-4 H3K27ac
C57 log2 24 h IL-4 H3K27ac
Stronger in SPRET
FC > 1.5 9,975
FC > 2 6,533
FC > 4 2,771
Stronger in C57
FC > 1.5 10,223
FC > 2 6,956
FC > 4 3,124
SPRET log2 24 h IL-4 H3K27ac
C57 log2 24 h IL-4 H3K27ac
D
C57
p=5.6e-5
H3K27ac log2(FC: IL-4/basal)
super enhancer
conventional enhancer
C
H3K27ac fold change C57 vs. another strain
super enhancer
conventional enhancer
BALB log2(FC: IL-4/basal)
NOD log2(FC: IL-4/basal)
PWK log2(FC: IL-4/basal)
SPRET log2(FC: IL-4/basal)
C57 log2(FC: IL-4/basal)
E
super enhancer
Ak2
H3K27ac
RNA Pol 2
BALB basal
BALB 24 h IL-4
C57 basal
C57 24 h IL-4
NOD basal
NOD 24 h IL-4
PWK basal
PWK 24 h IL-4
SPRET basal
SPRET 24 h IL-4
F
Msx3
conventional enhancers
DeepLIFT score
top 20% cutoff
position of variant from PWK and/or SPRET
position
top-scoring variant
G
H3K27ac fold difference btw. strains under IL-4 condition
% enhancers with local genetic variants
C57 vs. BALB
C57 vs. NOD
C57 vs. PWK
C57 vs. SPRET
strain-similar
< 1.5 17% 19% 69% 93%
strain-differential
> 1.5 37% 41% 78% 95%
> 2 56% 58% 81% 96%
> 4 76% 78% 88% 95%
H
H3K27ac fold difference btw. strains under IL-4 condition
% enhancers with top 20% DeepLIFT scored variants
C57 vs. BALB
C57 vs. NOD
C57 vs. PWK
C57 vs. SPRET
strain-similar
< 1.5 7% 8% 34% 61%
strain-differential
> 1.5 19% 22% 45% 71%
> 2 32% 34% 50% 74%
> 4 49% 52% 60% 77%
I
H3K27ac fold difference btw. strains under IL-4 condition
< 1.5
> 1.5
> 2
> 4
C57 vs BALB
C57 vs NOD
C57 vs PWK
C57 vs SPRET
top-scoring variants
all local variants
Log2 enrichment of top-scoring variants
Top % of DeepLIFT score

To investigate the impact of *cis* variation on putative transcriptional regulatory elements, we defined high confidence IL-4 activated enhancers as intronic or intergenic open chromatin regions (based on ATAC-seq) with at least 2.5-fold increase in H3K27ac (Creyghton et al., 2010) and RNA Pol2 (Bonn et al., 2012) after IL-4 treatment (Supplementary Fig. 3.2A to D). In 24-hour IL-4 stimulated C57 BMDMs, 1,093 regions exhibited a >2.5-fold increase in H3K27ac, whereas 441 regions exhibited a >2.5-fold decrease, corresponding to putative IL-4-activated and IL-4-repressed enhancers, respectively (Fig. 3.2A). Comparison of C57 enhancers to those of other strains under IL-4 treatment conditions revealed marked differences that scaled with the degree of genetic variation (Fig. 3.2B, Supplementary Fig. 3.2E and F). We further subdivided these regions into 'conventional enhancers' (blue, Fig. 3.2C) and 'super enhancers' (orange, Fig. 3.2C), based on the density distribution of normalized H3K27ac tag counts (Whyte et al., 2013). Super enhancers represent regions of the genome that are highly enriched for cell-specific combinations of transcription factors and co-regulators and control the expression of genes required for cellular identity and critical functions. In comparison to conventional enhancers, super enhancers exhibited significantly less variation in H3K27ac in response to IL-4 (Fig. 3.2D, Supplementary Fig. 3.2G). For example, IL-4 induction of the *Ak2* super enhancer (Fig. 3.2E) is highly conserved between the five strains. In contrast, a typical example of strain specificity is provided by the conventional enhancers associated with the *Msx3* gene. These enhancers are IL-4 inducible only in BALB, C57 and NOD and absent in PWK and SPRET macrophages (Fig. 3.2F).

We next compared the fractions of enhancers containing variants in strain-similar enhancers (<1.5-fold differences in H3K27ac between strains) to strain-differential enhancers at increasing levels of difference (fold differences >1.5 to >4; Fig. 3.2G). The fraction of enhancers

containing variants at strain-similar enhancers ranged from 17-19% in the strains most similar to C57 (BALB and NOD) to 69-93% in the most genetically divergent strains (PWK and SPRET). As expected, the fraction of enhancers containing variants increased with increasing levels of difference, except for SPRET which may have reached a saturation of variation capacity (Fig. 3.2G). These findings are consistent with local variants affecting enhancer activity, but also indicate that a substantial fraction of even strongly strain-differential IL-4 induced enhancers lack such variants, consistent with previous findings for strain-specific enhancers overall (Link, Duttke, et al., 2018).

In an effort to distinguish silent variants from those affecting enhancer activities, we trained a DeepSEA convolutional neural network to classify enhancers as active or inactive under the 24 h IL-4 condition based on local sequence context (J. Zhou & Troyanskaya, 2015). The training data consisted of enhancers active under IL-4 conditions (positive data) and random background (negative data). The area under the receiver operating characteristic curve (auROC) was 0.894 on test data. We then used DeepLIFT (Shrikumar et al., 2017) to compute the importance score of each nucleotide based on the model's classification decision. Variants at positions with top importance scores within surrounding 300-bp enhancer regions are hypothesized to affect enhancer activity. We considered variants residing in the top 20% of importance scores for each region as predicted functional variants. The *Msx3* enhancer in Fig. 3.2F illustrates four predicted functional variants out of fourteen variants in PWK and SPRET (red dotted lines). By focusing on top-scoring variants rather than all local variants, we saw an expected overall decreased percentage of enhancers with top-scoring variants (Fig. 3.2H, Supplementary Fig. 3.2H). On the other hand, enrichment of predicted functional variants increases as a function of importance score threshold and is strongest for enhancers that show the

highest differences across strains (Fig. 3.2I). This is true when considering all strains, including SPRET. These results reveal a quantitative impact of variants affecting enhancer under IL-4 treatment conditions and suggest the extent to which a deep learning approach can distinguish potentially functional variants from the silent variants.

3.3.3 IL-4 activated enhancers use pre-existent promoter-enhancer interactions to regulate gene activity

Interpretation of effects of genetic variation on distal regulatory elements is facilitated by knowledge of cell-specific enhancer-promoter interactions (Nott et al., 2019). To identify connections of IL-4-responsive enhancers to target promoters, we performed HiChIP using an antibody to H3K4me3 (Mumbach et al., 2016) in C57 BMDMs under basal conditions and after 24 h of IL-4 treatment. HiChIP interactions are exemplified in Fig. 3.3A at the *Slc7a2* locus, a gene that becomes maximally activated after 24 h of IL-4 treatment and connects primarily to an enhancer-like region within the *Mtmr7* gene which itself is expressed at negligible levels (Fig. 3.3A). Although we observed instances of IL-4-specific interactions (e.g., yellow loops), a differential interaction analysis was unable to identify significantly different interactions between basal and IL-4 conditions, supported by the high correlation of interaction intensity between the two conditions (Supplementary Fig. 3.3A). Moreover, enhancer-promoter interaction intensity did not correlate with IL-4 induced gene activity or the level of H3K4me3 at promoters (Supplementary Fig. 3.3B-C). However, IL-4 activated promoters mostly interact with IL-4 activated enhancers (Fisher's exact test, p=2.2e-16), and repressed promoters strongly interact with repressed enhancers (p=1.2e-15, Fig. 3.3B). These results suggest a pre-existent and relatively stable landscape of enhancer-promoter interactions in macrophages, whose regulatory function was activated in response to IL-4.

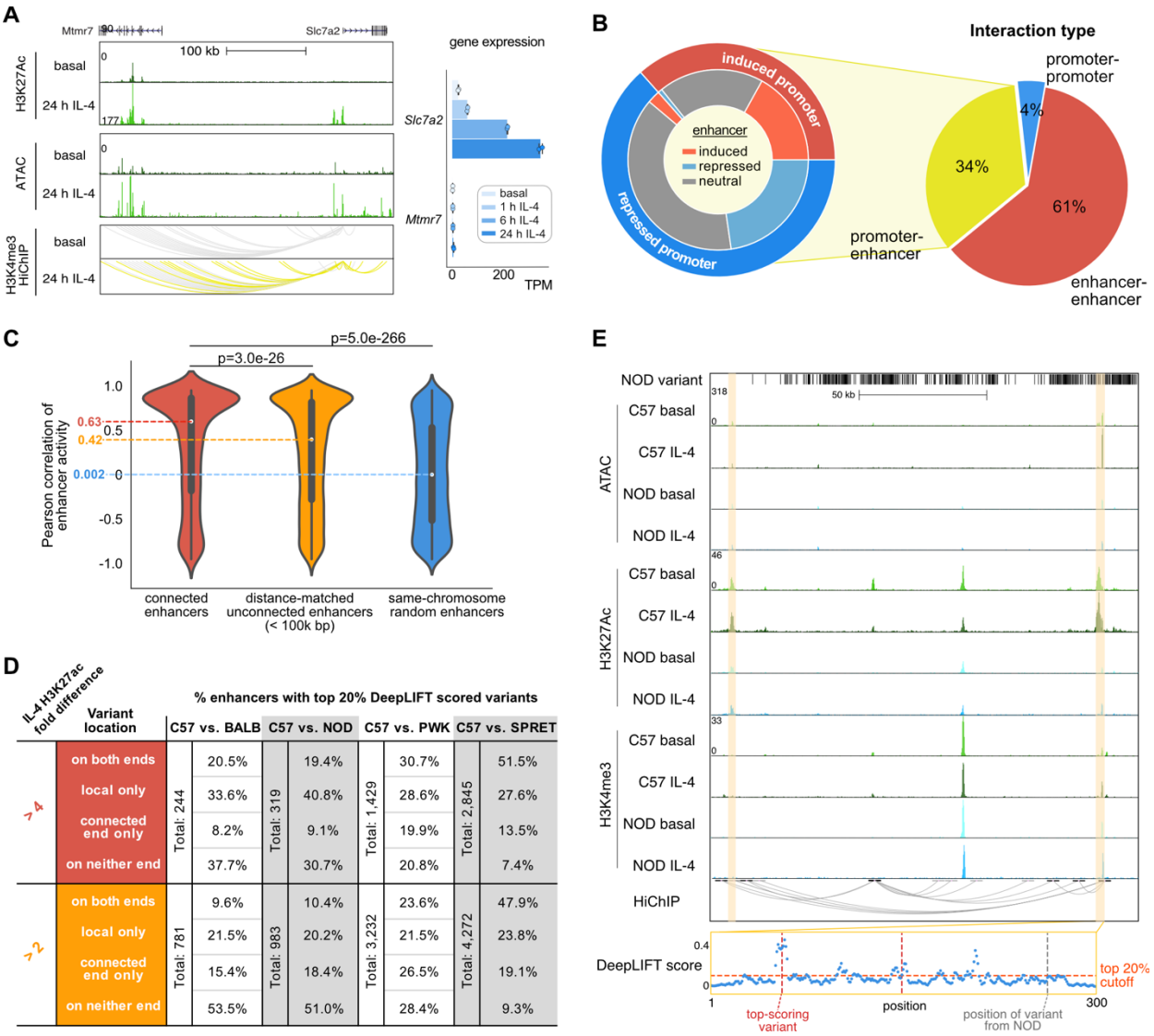


Figure 3.3 IL-4 enhancers use pre-existent promoter-enhancer interactions to regulate gene activity. (A) HiChIP indicates the *Slc7a2* promoter is highly connected with several IL-4 activated enhancers. *Slc7a2* and *Mtmr7* gene expression upon IL-4 stimulation were shown. **(B)** Different categories of HiChIP interactions (right), and enhancer-promoter connections overlapping with IL-4 responsive regulatory elements in C57 BMDMs (left). Outer ring indicates induced or repressed promoters, while inner ring indicates their connected enhancers associated with IL-4-induced, IL-4 repressed or IL-4 neutral H3K27ac. **(C)** Correlations of H3K27ac signal between connected enhancers compared to non-interactive enhancers using Mann–Whitney U test. **(D)** Table representing enhancers containing DeepLIFT high-scored genetic variants locally or at connected elements in pairwise comparisons between C57 and other strains. **(E)** Strain-differential enhancer between C57 and NOD where genetic variants were absent locally but present at a connected enhancer with two DeepLIFT high-scored variants (red dotted lines).

Although the HiChIP assay is designed to capture promoter-enhancer interactions based on preferential occurrence of H3K4me3 at promoters, we also recovered 145,907 pairs of interactive enhancers (Fig. 3.3B), consistent with more than one enhancer being in local

proximity of a target promoter. The H3K27ac correlations between interactive enhancers were significantly stronger than those between non-interactive enhancers (Fig. 3.3C, Supplementary Fig. 3.3D), consistent with their being functionally related. Noticeably, the closer enhancers despite being non-interactive based on our data still have much stronger correlation than completely random enhancers, which might be due to more frequent contacts of nearby regions within the same interactive domain that were not captured by H3K4me3 HiChIP. Based on the high correlation of enhancer activity between connected enhancers, we hypothesized that enhancer-enhancer interactions could explain strain-differential enhancer when local genetic variants were absent (Fig. 3.2H). Among 224 interactive enhancers exhibiting a >4-fold difference in H3K27ac signal between BALB and C57 under the IL-4 condition, the original ~50% of strain-differential enhancers with predicted functional variants was further split into 20.5% that had top-scoring variants on both ends and 33.6% that had only local top-scoring variants (Fig. 3.3D, upper left). Depending on the strain comparison, an additional 8.2%-19.9% of differential enhancers could be explained by genetic variants in interacting enhancers, indicating that enhancers may be affected by functional variants in other connected enhancers. Reducing the fold change requirement to 2-fold yielded a smaller proportion of strain-differential enhancers containing local variants overall but significantly increased the proportion having top-scoring variants on the connected ends only (15.4%-26.5%, Fisher's exact test p=0.002 for BALB, 2.8e-5 for NOD, 8.3e-7 for PWK, 3.3e-10 for SPRET), suggesting that local variants have a stronger effect on inducing differential activation than variants at connected enhancers (Fig. 3.3D lower panels, Supplementary Fig. 3.3E). Fig. 3.3E illustrates an enhancer affected by genetic variants at the connected enhancer. The enhancer highlighted on the left is significantly more active in C57 than NOD. This region lacks local variants in NOD but is connected to

another enhancer ~100 kb away (highlighted on the right) containing multiple variants that are predicted to affect activity by deep learning. These findings are consistent with genetic variants at an enhancer influencing the activity states of other enhancers that lack local functional variants within the same connected network (Grubert et al., 2015; Waszak et al., 2015).

3.3.4 Motif mutation analysis identifies motifs that are functionally associated with IL-4 induced enhancer activity

IL-4 rapidly activates a set of enhancers, the majority of which exhibit maximal H3K27ac at 1 h or 6 h and returns to (near) basal levels by 24 h (Fig. 3.4A, top three clusters) when most gene expression changes were found (Supplementary Fig. 3.1B). Others are long-lasting or become activated at later timepoints (Fig. 3.4A, bottom three clusters). *De novo* motif enrichment analysis of enhancers exhibiting >2.5-fold increase in H3K27ac and RNA Pol2 at 1 h, 6 h and 24 h (Supplementary Fig. 3.2A) recovered a STAT6 motif as the most enriched motif for all timepoints (Fig. 3.4B). Motifs for the lineage determining factors PU.1 and AP-1 family members were also recovered in all three classes of enhancers. Notably, an EGR2 motif was significantly enriched among enhancers induced at 24 h.

As a genetic approach to identify functional transcriptional factor binding motifs, we assessed the quantitative impact of the genetic variation provided by the five different strains of mice on the IL-4 response of enhancers using the motif mutation analysis tool MAGGIE. MAGGIE associates changes of epigenomic features at homologous sequences (e.g., enhancer activation or enhancer repression) with motif mutations caused by genetic variation so that it can prioritize motifs that likely contribute to the regulatory function (Shen et al., 2020). This analysis identified more than a dozen motif clusters in which motif mutations were significantly associated with strain-differential IL-4 activated or repressed enhancers (Fig. 3.4C,

Supplementary Fig. 3.4A). The EGR motif was found as the top motif associated with enhancer activation at the 24 h treatment time, as well as motifs of known SDTFs STAT6 and PPARγ and macrophage LDTFs PU.1, AP-1 and CEBP (Fig. 3.4C). We also found KLF motifs associated with IL-4 enhancer activation, which fits with increased KLF4 expression by IL-4 (Supplementary Fig. 3.4B), and an NRF motif associated with both enhancer activation (Fig. 3.4C).

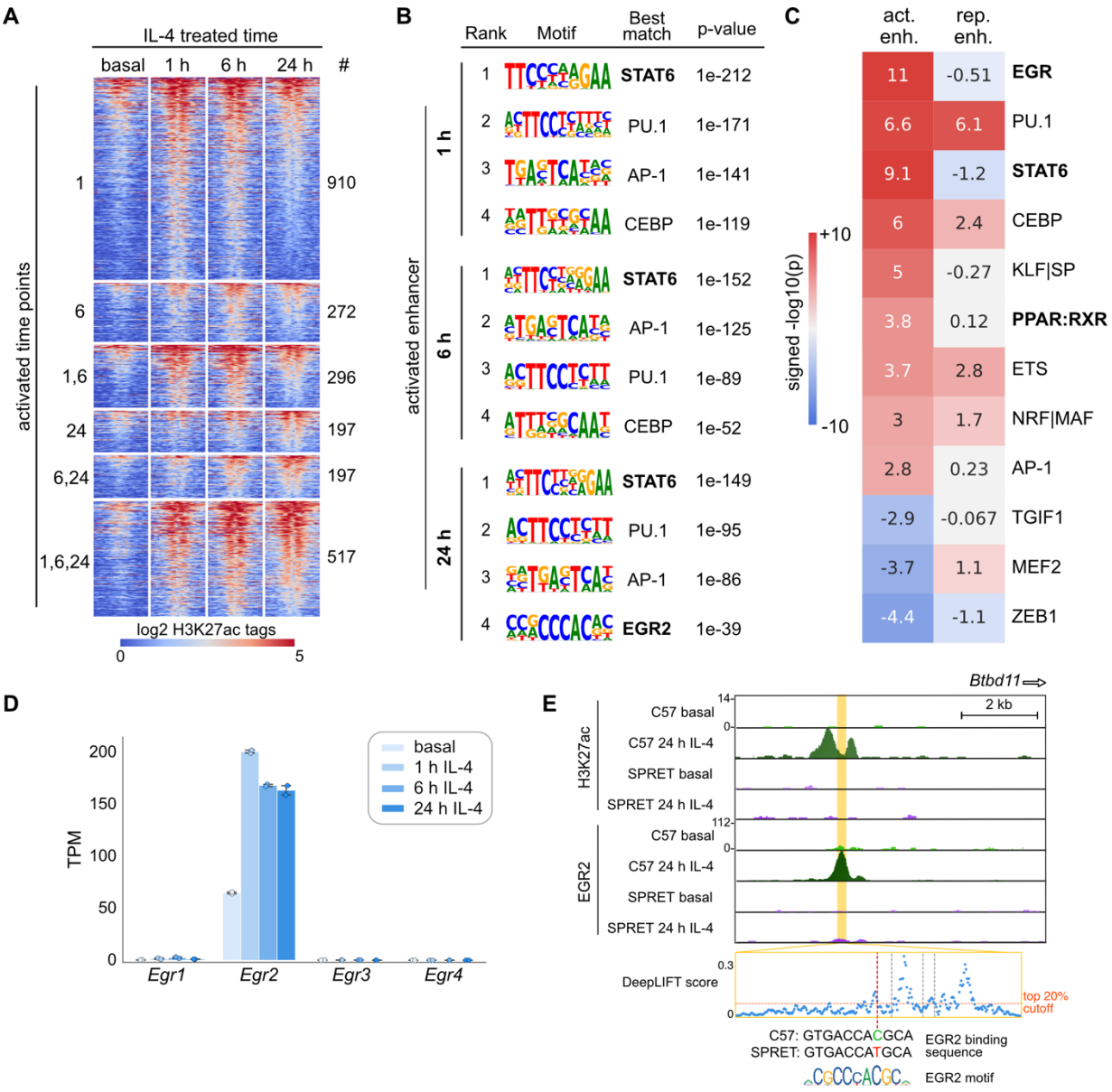


Figure 3.4 Motif analysis identifies motifs functionally associated with IL-4 induced enhancers. (A) Heatmap showing the effects of 1 h, 6 h and 24 h IL-4 stimulation on enhancer activation based on H3K27ac abundance. **(B)** Top motifs enriched at ATAC-seq peaks exhibiting gained H3K27ac at different time points. **(C)** MAGGIE motif mutation analysis on strain-differential activated and repressed enhancers after 24 h IL-4. **(D)** *Egr* gene expression in C57 BMDMs under basal conditions and after stimulation with IL-4, ****q<0.0001, compared to basal. **(E)** Example of a strain-differential activated enhancer upstream of the *Btbd11* gene based on IL-4-induced H3K27ac signal in C57 but not in SPRET BMDMs, supported by binding of EGR2 and a functional variant predicted by DeepLIFT that mutates the EGR2 motif.

The identification of STAT6 and PPARγ motif mutations as being functionally associated with strain-differential IL-4 activation is consistent with substantial prior work demonstrating the importance of these factors in regulating IL-4-dependent gene expression (Czimmerer et al.,

2018; Daniel et al., 2018). Out of the Early Growth Response (EGR) family members only *Egr2* is expressed in unstimulated BMDMs and rapidly induced after IL-4 stimulation (Fig. 3.4D, Supplementary Fig. 3.4C). *Egr2* has also been associated with late IL-4 enhancer activation in a recent study (Daniel et al., 2020). Examination of the *Egr2* locus indicates IL-4 induced binding of STAT6 and PPARγ to a set of upstream super enhancers that gain H3K27ac and RNA Pol2 signal after IL-4 stimulation (Supplementary Fig. 3.4D). These super enhancers were observed in BMDMs of all five different strains (Supplementary Fig. 3.4E) that are strongly connected to the *Egr2* promoter in C57 BMDMs as indicated by H3K4me3 HiChIP interactions. Overall, these findings suggest a functionally important role of EGR2 in contributing to IL-4 induced enhancer activation in BMDMs.

3.3.5 IL-4 induced EGR2 contributes to late IL-4 enhancer activation

In order to incorporate EGR2 into a comprehensive analysis of the impact of genetic variation on IL-4 responses, we next performed ChIP-seq for EGR2 under basal and 24 h IL-4 treatment conditions. This confirmed the prediction that mutations in EGR binding sites contribute to strain-differential enhancer activation by altering the binding of EGR2. An example is provided by the *Btbd11* enhancer, which is IL-4 inducible in C57, but not in SPRET BMDMs (Fig. 3.4E). Consistent with the loss of EGR2 binding in SPRET, a C-to-T variant in SPRET mutated an EGR2 motif.

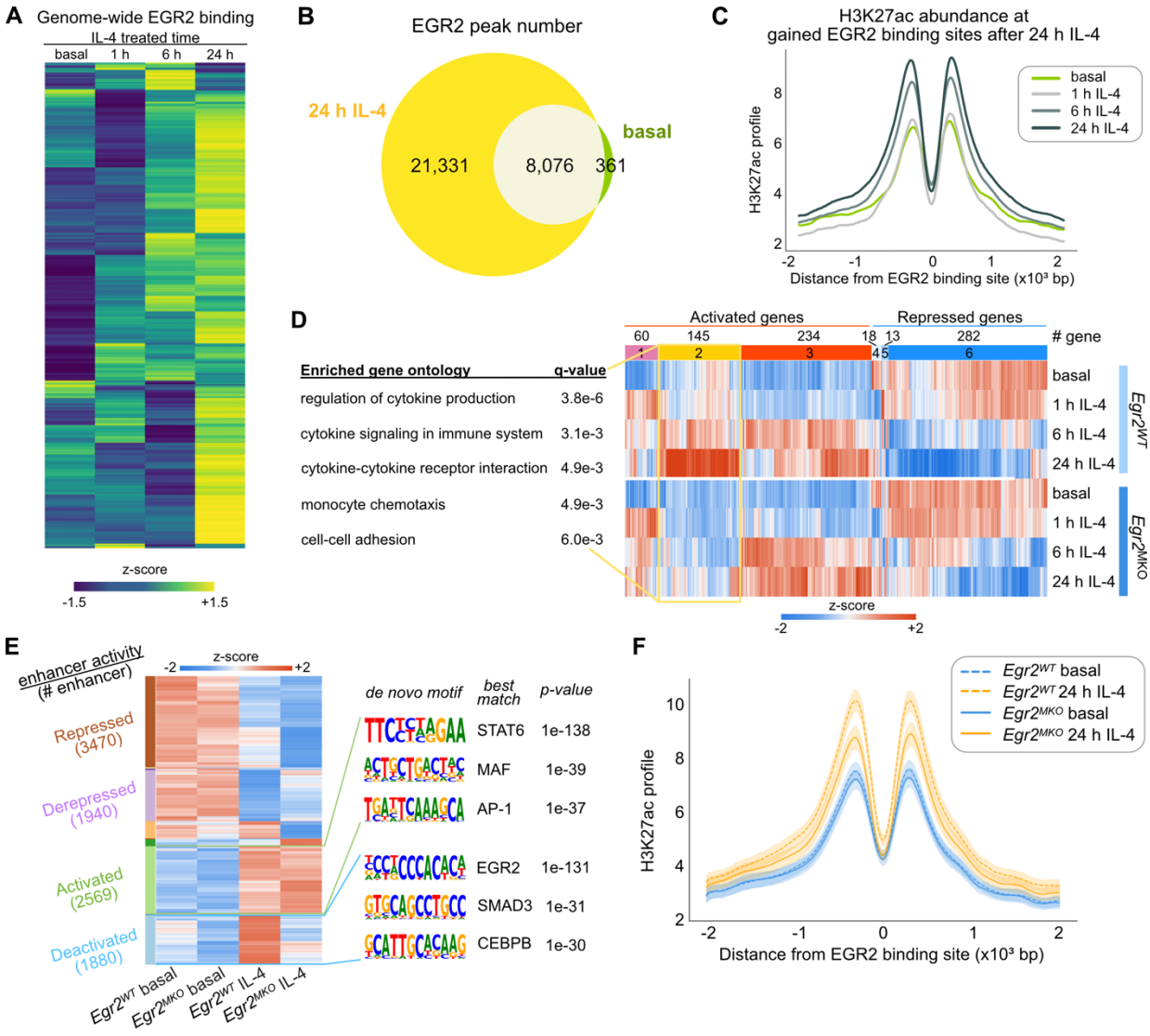


Figure 3.5 IL-4 induced EGR2 contributes to late IL-4 enhancer activation. (A) Heatmap displaying EGR2 ChIP-seq binding intensity after IL-4 stimulation over time in C57 BMDMs. **(B)** Number of EGR2 binding sites after 24 h IL-4 compared to the basal condition in C57 BMDMs. **(C)** H3K27ac profiles at 24 h IL-4 induced intergenic and intronic EGR2 peaks in C57 BMDMs. **(D)** Expression of IL-4 regulated genes in *Egr2*^{WT} and *Egr2*^{MKO} BMDMs with top gene ontology terms displayed for cluster 2. **(E)** Enhancer activity of IL-4 regulated enhancers in *Egr2*^{WT} and *Egr2*^{MKO} BMDMs. Enriched motifs at EGR2-dependent and EGR2-independent IL-4 induced enhancers using each other as backgrounds. **(F)** H3K27ac profiles at IL-4 induced EGR2 binding sites in *Egr2*^{WT} and *Egr2*^{MKO} BMDMs. 90% confidence intervals are shown together with the average profiles.

To study EGR2 binding and its effects on gene regulation over time, we additionally measured EGR2 binding at 1 h and 6 h after IL-4. We saw a marked expansion of the EGR2 cistrome after stimulation with IL-4 (Fig. 3.5A-B). Most EGR2 binding sites had an increasing binding intensity over time and reached maximum values at 24 h (Fig. 3.5A, Supplementary Fig.

3.5A). In constrast, STAT6 binding was strong immediately after 1 h and slowly decreased in its binding intensity (Supplementary Fig. 3.5A). At 24 h, there were over 20,000 newly gained EGR2 binding sites (Fig. 3.5B), which were associated with an increase in H3K27ac and RNA Pol2 signals over time, supporting a major role of EGR2 in late enhancer activation (Fig. 3.5C, Supplementary Fig. 3.5B).

To extend these analyses, we crossed *Egr2*^{flfl} (Du et al., 2014) (*Egr2*^{WT}) with *LyzM*-Cre^{+} mice to obtain *LyzM*-Cre^{+} *Egr2*^{flfl} (*Egr2*^{MKO}) mice. This resulted in efficient deletion of *Egr2* in BMDMs (Supplementary Fig. 3.5C-D). RNA-seq data from *Egr2*^{WT} and *Egr2*^{MKO} BMDMs indicated that at the mRNA level, EGR2 is regulating ~40% of the 24 h IL-4 induced genes (Fig. 3.5D, Supplementary Fig. 3.5E), many of which correspond to gene ontology terms related to cytokine production and signaling (Fig. 3.5D). At the regulatory level, we found that ~40% of the IL-4 induced enhancers at 24 h had significantly decreased activity in *Egr2*^{MKO} BMDMs (blue cluster, Fig. 3.5E, Supplementary Fig. 3.5F-H). In concordance, the IL-4 induction in H3K27ac was found to be decreased at IL-4 induced EGR2 binding sites in *Egr2*^{MKO} BMDMs (Fig. 3.5F). Based on a motif enrichment analysis of the EGR2-dependent IL-4 activated enhancers (blue cluster "deactivated") with the EGR2-independent enhancers as background (green cluster "activated"), we found EGR2 as the most significantly enriched motif and SMAD3 and CEBP motifs as the second and the third hits (Fig. 3.5E). When comparing the EGR2-independent activated enhancers ("activated") to the EGR2-dependent ones ("deactivated"), the STAT6 motif was most significantly enriched (Fig. 3.5E), suggesting STAT6 could work independently of EGR2 to maintain the late activation for a subset of enhancers. Together, these findings establish an essential role of EGR2 in IL-4 dependent enhancer activation and gene expression and are in full agreement with the recent studies of Daniel, et al., (Daniel et al., 2020).

3.3.6 Collaborative and hierarchical transcription factors interact at IL-4 dependent enhancers

Analysis of the genome-wide binding patterns of EGR2, STAT6 and PPARγ indicated intensive co-binding at IL-4 activated enhancers (Supplementary Fig. 3.6A). To achieve highly strain-differential responses to IL-4 that are observed at the level of gene expression (Fig. 3.1) and enhancer activation (Fig. 3.2), these factors are hypothesized to exert their transcriptional effects via correspondingly divergent genomic binding patterns. About 4,000 EGR2 binding sites exhibited more than 4-fold differences in normalized tag counts between C57 and BALB, and about 10,000 between C57 and SPRET (

Fig. 3.6A). Similar relationships are observed for STAT6 (Supplementary Fig. 3.6B). Within the strain comparisons of C57 to BALB and SPRET, C57-specific EGR2 binding sites are associated with stronger H3K27ac signal in C57 (Fig. 3.6B, green boxes). At these C57-specific EGR2 binding sites, H3K27ac signal is more strongly downregulated in *Egr2*^{MKO} macrophages (Fig. 3.6B, orange boxes). An example of this is demonstrated in Supplementary Fig. 3.6C in which C57-specific binding of EGR2 is associated with an IL-4 induced increase in H3K27ac, which is absent in SPRET (where the EGR2 motif is disrupted, Fig. 3.4E) and is decreased in *Egr2*^{MKO} BMDMs.

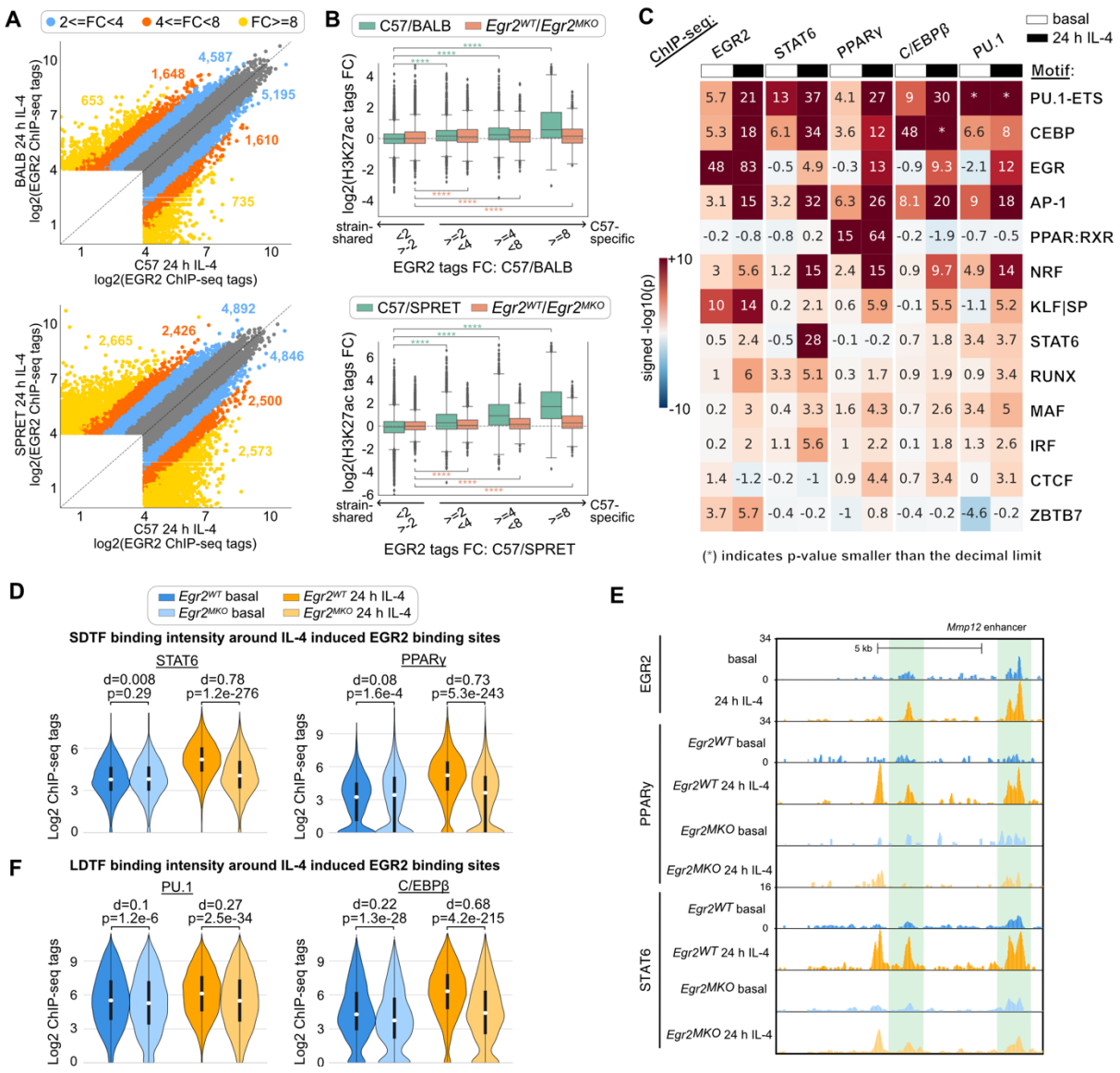


Figure 3.6 Collaborative and hierarchical transcription factors interact at IL-4 enhancers. (A) Scatter plots comparing binding of EGR2 in C57 versus BALB and C57 versus SPRET IL-4 stimulated BMDMs. **(B)** Log2 fold changes of H3K27ac signal between different strains (green boxes) or between *Egr2*^{WT} and *Egr2*^{MKO} BMDMs (orange boxes) at C57-specific and strain-shared EGR2 binding sites. Distributions of C57-specific sites were compared to those of strain-shared sites in the same category using the two-sample t-test. ****p<0.0001. **(C)** Functional motifs from MAGGIE analysis at EGR2, STAT6, PPARγ, C/EBPβ and PU.1 binding sites. **(D)** STAT6 and PPARγ binding at IL-4 induced EGR2 peaks in *Egr2*^{WT} and *Egr2*^{MKO} BMDMs. Cohen's d effect size and p-values from Mann–Whitney U tests are shown. **(E)** Co-binding of STAT6, EGR2, and PPARγ at the *Mmp12* enhancer. **(F)** C/EBPβ and PU.1 binding at IL-4 induced EGR2 peaks in *Egr2*^{WT} and *Egr2*^{MKO} BMDMs.

The strain-differential binding patterns of SDTFs and LDTFs enabled motif mutation analysis to study the importance of motifs for SDTF and LDTF binding (Fig. 3.6C). As a validation, LDTF and SDTF binding depended on their own motifs (e.g., PU.1 motif mutation

was significantly associated with PU.1 binding, indicating that when PU.1 binding is lost in one strain, it is often found that the PU.1 motif score is reduced in that strain compared to the other). In addition, mutations in the motifs of LDTFs PU.1, C/EBP and AP-1 influence the binding of all LDTFs and SDTFs, which fits with earlier observations (Heinz et al., 2013). We found that the PPAR motif is only significant for PPARγ binding, and likewise, the STAT6 motif is not associated with binding of other SDTFs or LDTFs but STAT6. Interestingly, we found that mutations in EGR2 motifs are significantly associated with binding of SDTFs STAT6 and PPARγ and LDTFs PU.1, C/EBPβ under IL-4 conditions, but not under basal conditions. These analyses also provided evidence for functional roles of several additional transcription factors. Mutations in NRF motifs were strongly associated with the IL-4-dependent binding of all SDTFs and LDTFs. Both NRF1 and NRF2 are expressed in BMDMs (Supplementary Fig. 3.4B) and are involved in lipid metabolism and stress responses (Kobayashi et al., 2016; Widenmaier et al., 2017). Mutations in KLF motifs were strongly associated with EGR2 binding under both basal and IL-4 conditions. KLF2, KLF4 and KLF6 are expressed in BMDMs (Supplementary Fig. 3.4B) and KLF4 has previously been associated with anti-inflammatory roles in macrophages (Liao et al., 2011). Mutations in IRF motifs were moderately associated with IL-4-dependent STAT6 binding. Multiple IRFs, including IRF4, are expressed in BMDMs (Supplementary Fig. 3.6D) and IRF4 has previously been linked to macrophage polarization by IL-4 (El Chartouni et al., 2010; Satoh et al., 2010).

A prediction emerging from the analysis results above is that EGR2 should have a small effect on the co-binding of SDTFs and LDTFs under basal conditions and a significant effect following 24 h of IL-4 treatment. To examine this prediction, we performed ChIP-seq for STAT6, PPARγ, PU.1 and C/EBPβ in *Egr2*^{WT} and *Egr2*^{MKO} BMDMs and evaluated their binding

in the vicinity of IL-4 induced EGR2 binding sites. Deletion of *Egr2* had little effect on PPARγ and STAT6 binding under basal conditions and a much greater effect following 24 h of IL-4 treatment (Fig. 3.6D). As an example, in *Egr2*^{MKO} BMDMs, PPARγ and STAT6 binding was found decreased at the *Mmp12* enhancer at sites where EGR2 normally binds (Fig. 3.6E). Similarly, PU.1 and C/EBPβ binding was more significantly affected by *Egr2* deletion under the IL-4 condition than the basal condition (Fig. 3.6F). In concert, these findings provide evidence for collaborative interactions between PU.1, C/EBPs, AP-1, STAT6, PPARγ and EGR2 as major drivers of late enhancer activation in response to IL-4. EGR2 is a strong collaborative factor as it promotes binding of LDTFs PU.1 and C/EBPβ and SDTFs STAT6, PPARγ after IL-4 stimulation.

3.3.7 Quantitative variations in motif affinity determine dynamic responses of IL-4 enhancers

We next investigated the possibility that the mutational status of the dominant motifs recovered by MAGGIE analysis was sufficient to predict qualitative patterns of strain-differential responses of IL-4 induced enhancers. Following the classification of strain-differential mRNA responses (Fig. 3.1), we used H3K27ac to define three different categories of strain-differential IL-4-induced enhancers (Fig. 3.7A, left column): enhancers exhibiting lower levels of basal activity in the lowly induced strain (**low basal**); enhancers with a similar level of basal activity (**equal basal**); enhancers in which a lack of IL-4 induced activity was associated with relatively higher basal activity compared to the more responsive strain (**high basal**). Using these criteria, we identified 760 low basal, 2797 equal basal and 2013 high basal enhancers from all pairwise comparisons of the five strains that exhibited >2-fold differences in H3K27ac induction (Fig. 3.7B). The closest genes for enhancers of these three categories follow similar

trends as observed for enhancer activity (Supplementary Fig. 3.7A). Low basal, equal basal and high basal enhancers are exemplified by enhancers associated with the *Treml2, Ripk2* and *Cd36* genes, respectively (Fig. 3.7C-E, Supplementary Fig. 3.7B-D).

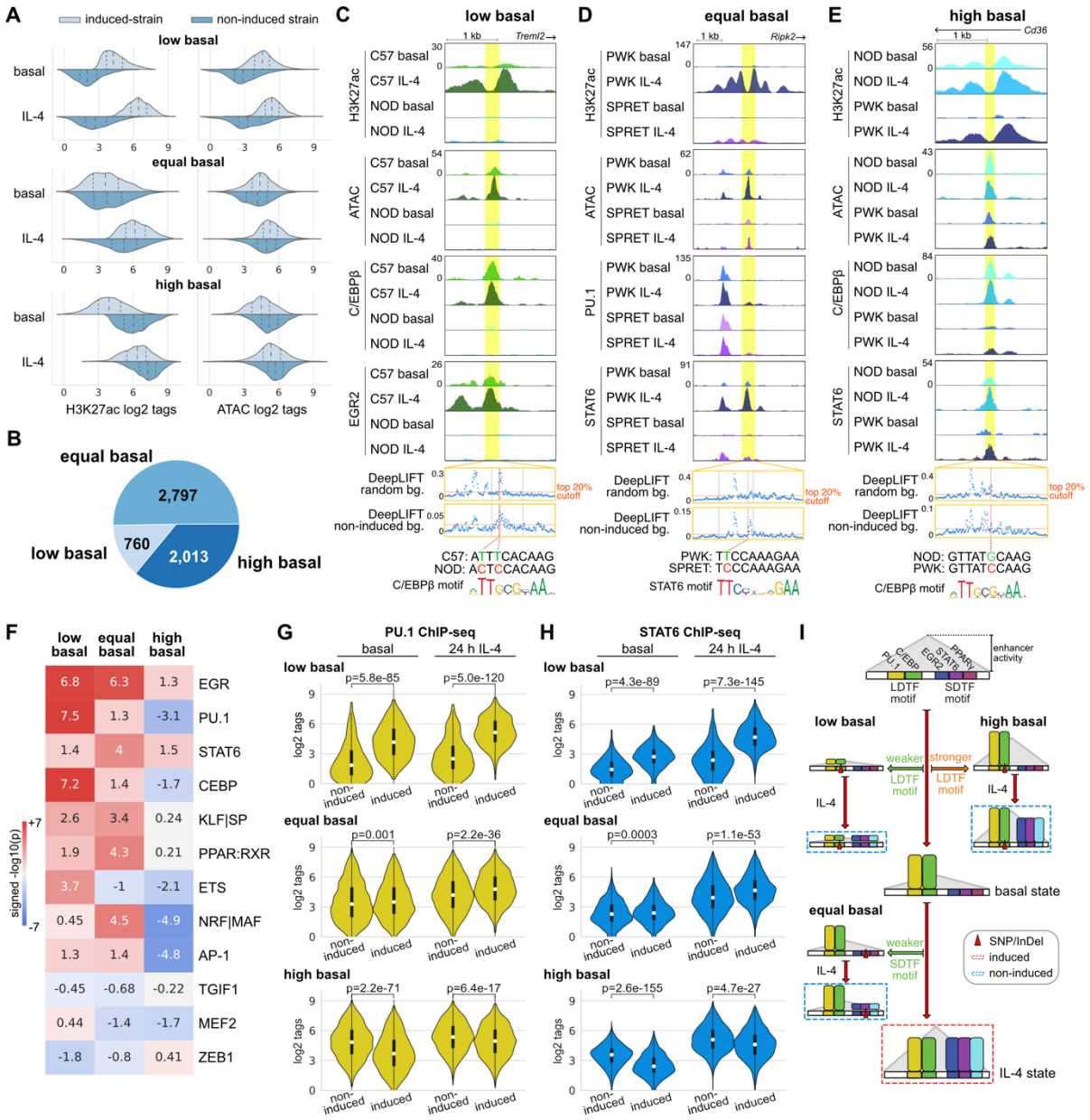


Figure 3.7 Quantitative variations in motif affinity determine dynamic responses of IL-4 enhancers. (A) Three different categories of strain-differential IL-4 activated enhancers with distributions of ATAC and H3K27ac signal. Dashed lines in each distribution indicate quartiles. **(B)** Numbers of enhancers in the three categories. **(C-E)** Example of low (**C**), equal (**D**) and high (**E**) basal enhancers with high impact variants predicted by DeepLIFT. **(F)** MAGGIE motif mutation analysis on different categories of enhancers. **(G, H)** Binding intensities of PU.1 (**G**) and STAT6 (**H**) in non-induced and induced strains at different categories of enhancers. **(I)** Graphical representation of the general mechanisms for different categories of IL-4 induced enhancers.

Consideration of chromatin accessibility as determined by ATAC-seq further uncovered potential mechanisms that distinguished the three enhancer categories (Fig. 3.7A, right column). The enhancers in the low basal category showed low to absent basal ATAC signal in non-induced strains, suggesting a lack of LDTFs under the basal condition to pre-occupy chromatin required for subsequent recruitment of SDTFs after IL-4 stimulation. In contrast, high basal enhancers exhibited a higher basal level of ATAC in non-induced strains compared to induced strains (Fig. 3.7A, right column), suggesting stronger LDTF binding in non-induced strains under the basal condition. Different from the other categories, equal basal enhancers exhibited similar levels of chromatin accessibility under both basal and IL-4 conditions between comparative strains, suggesting that the recruitment of SDTFs might be the key determinant for the strain difference instead of basal LDTF binding.

To test the hypotheses above regarding the different determinants for the three categories of enhancers, we performed MAGGIE motif mutation analysis on each category of enhancers that contain motif mutations (Supplementary Fig. 3.7E). We found that mutations in motifs of LDTFs PU.1/ETS and C/EBP were associated with low basal enhancers and resulted in better motifs in induced strains, while mutations in motifs of SDTFs EGR, STAT6, PPAR and NRF/MAF were associated with the equal basal category leading to better motifs in induced strains (Fig. 3.7F, Supplementary Fig. 3.7F). Mutations in EGR motifs were also associated with the low basal category, suggesting another role of EGR2 as a strong collaborative factor under the IL-4 condition, which is supported by the significant decrease in open chromatin under IL-4 conditions after deletion of *Egr2* (Supplementary Fig. 3.5F). Of particular interest, the high basal category of enhancers was most strongly associated with negative significance scores for LDTF

PU.1, C/EBP and AP-1 as well as NRF/MAF, meaning higher motif affinity in non-induced strains (Fig. 3.7F).

We validated these findings with our ChIP-seq data by examining the binding profiles of PU.1, C/EBPβ, STAT6, PPARγ and EGR2 in three categories of enhancers. In low basal enhancers, we saw significantly reduced binding of PU.1 and C/EBPβ in non-inducible strains under both basal and IL-4 conditions (Fig. 3.7G, Supplementary Fig. 3.7G). This pattern was accompanied by significantly weaker binding of SDTFs STAT6, EGR2 and PPARγ after IL-4 stimulation (Fig. 3.7H, Supplementary Fig. 3.7H). The example in Fig. 3.7C showed the absence of C/EBPβ binding in NOD under the basal condition likely due to two local variants at high-scored positions according to DeepLIFT that together mutated a C/EBP motif. Upon IL-4 stimulation, neither C/EBPβ nor EGR2 was further recruited. For equal basal enhancers, we found that PU.1 and C/EBPβ binding was similar under basal conditions in induced and non-induced strains (Fig. 3.7G, Supplementary Fig. 3.7G). Upon IL-4 stimulation, the induced strains displayed significantly stronger binding of SDTFs STAT6, EGR2 and PPARγ (Fig. 3.7H, Supplementary Fig. 3.7H). In the example in Fig. 3.7D, STAT6 binding was strongly induced by IL-4 at the *Ripk2* enhancer in PWK but was absent in SPRET. Despite the clear difference in STAT6 binding, none of the local variants between the two strains was predicted functional when using a neural network model trained with random genomic backgrounds. To better capture the sequence patterns relevant for enhancer activation, we retrained neural networks using non-induced enhancers as the background, which emphasized a relatively divergent set of k-mers and focused less on those matched with LDTF motifs (Supplementary Fig. 3.7I). As a result, our retrained model assigned a high DeepLIFT score to one of the nucleotides in a STAT6 motif that was mutated by a variant in SPRET (Fig. 3.7D). For high basal enhancers, we found

stronger binding of not only the LDTFs PU.1 and C/EBPβ (Fig. 3.7G, Supplementary Fig. 3.7G) but also the SDTFs STAT6 and PPARγ (Fig. 3.7H, Supplementary Fig. 3.7H) in non-induced strains under basal conditions. For example, high basal levels of C/EBPβ and STAT6 binding were observed at the *Cd36* enhancer in NOD mice (Fig. 3.7E). The only local variant in PWK was at a predicted functional position and mutated a C/EBP motif likely causing the low basal C/EBPβ binding in PWK. In concert, these analyses validated the importance of LDTF motif mutations as primary determinants of differential enhancer activation in low basal and high basal enhancers, while also demonstrating the expected consequences of SDTF motif mutations in determining strain-differential activation of equal basal enhancers (Fig. 3.7I).

3.4 Discussion

Here, we report a systematic investigation of the effects of natural genetic variation on signal-dependent gene expression by exploiting the highly divergent responses of BMDMs from diverse strains of mice to IL-4. Unexpectedly, despite broad conservation of IL-4 signaling pathways and downstream transcription factors in all five strains, only 26 of more than 600 genes observed to be induced >2-fold by IL-4 at 24 hours reached that level of activation in all five strains and more than half were induced in only a single strain. To the extent that this remarkable degree of variation observed in BMDMs occurs in tissue macrophages and other cell types *in vivo*, it is likely to have significant phenotypic consequences with respect to innate and adaptive immunity, tissue homeostasis and wound repair. Notably, only ~25% of the variation in response to IL-4 was due to altered dynamic ranges in the context of an equivalent level of basal expression. Nearly half of the genes showing strain-specific impairment in IL-4 responsiveness exhibited low basal activity, whereas lack of induction was associated with constitutively high

basal levels of expression in the remaining ~25%. These qualitatively different patterns of strain responses to IL-4 imply distinct molecular mechanisms by which genetic variation exerts these effects.

Motif mutation analysis of strain-differential enhancer activation recovered a dominant set of motifs recognized by known LDTFs PU.1, C/EBPβ and AP-1 family members, as well as motifs recognized by SDTFs STAT6 and PPARγ that have been previously established to play essential roles in the IL-4 response. In addition, effects of mutations in motifs for EGR, NRF and KLF also strongly implicate these factors as playing important roles in establishing basal and induced activities of IL-4 responsive enhancers, which was genetically confirmed for EGR2 in this study as well as a recent study (Daniel et al., 2020). It will be of interest in the future to perform analogous studies of NRF and KLF factors.

Analysis of strain-differentially activated enhancers revealed qualitative differences in basal and IL-4-dependent activity that were analogous to the qualitative differences observed for strain-differentially activated genes. As expected, sequence variants reducing the affinity of SDTFs STAT6, PPARγ and EGR2 were the major forms of variation resulting in strain-differential IL-4 induction of equal basal enhancers. From the standpoint of interpreting the effects of non-coding variation, these types of sequence variants are silent in the absence of IL-4 stimulation. As also expected, sequence variants strongly reducing the binding affinity of LDTFs prevented the generation of open chromatin required for subsequent binding of SDTFs. Such variants are thus expected to result in loss of enhancer function in a signal-independent manner. Of particular significance, these analyses also provide strong evidence that quantitative variation in suboptimal motif scores for LDTFs is a major determinant of differences in the absolute levels and dynamic range of high basal enhancers across strains. The importance of low affinity motifs

in establishing appropriate quantitative levels of gene expression within a given cell type and cell specificity across tissues has been extensively evaluated (Crocker et al., 2015; Farley et al., 2015; Kribelbauer et al., 2019). Here we present evidence that improvement of low affinity motifs for LDTFs not only increases basal binding of the corresponding transcription factor but is also associated with increased basal binding of STAT6 and PPARγ, thereby rendering their actions partially or fully IL-4 independent. These findings thus provide evidence that quantitative effects of genetic variation on LDTF motif scores play major roles in establishing different absolute enhancer activity levels and dynamic ranges of their responses to IL-4 that are observed between strains.

To go beyond the discovery of mechanisms mediating the IL-4 response using natural genetic variation, a major objective of these studies was to use the resulting data sets as the basis for interpreting and predicting the effects of specific variants. As expected, enhancers exhibiting strain specific differences in IL-4 responses were significantly enriched for sequence variants. However, the background frequencies of variants in the much larger sets of strain-similar enhancers ranged from 17% to 93%, consistent with the vast majority of such variants being silent and underscoring the challenges of discriminating them from functional variants. The application of recently developed deep learning approaches illustrates both the potential of these methods to improve predictive power as well as their current limitations. Nucleotides predicted by DeepLIFT to be of functional importance frequently intersected with variants at strain-differential enhancers that significantly altered LDTF or SDTF motifs, with over 8-fold enrichment in enhancers with strongest strain differences (top 1% variants for C57 vs. BALB comparison, Fig. 3.2I), strongly suggesting causality. Even though DeepLIFT scored a significant fraction of variants present in strain-similar enhancers with low importance, a large

fraction of remaining strain-similar enhancers contained variants associated with high DeepLIFT scores, most likely representing false positives. Further, we found that the highest scoring variants in some cases depended on the choice of data used to train the convolutional neural network (e.g. using random vs. non-induced enhancers as negative training examples). This observation has significant implications with respect to application of deep learning models to identify potential functional variants in disease contexts. The data sets generated by these studies will therefore provide an important resource for further improvements in methods for interpretation of local genetic variation.

These analyses further indicated that 20%-50% of the most divergent IL-4-responsive enhancers lacked any functional variants in the proximity of open chromatin. This fits with previous observations that variant-free enhancers can reside in cis regulatory domains (CRD) containing functionally interacting enhancers, suggesting that a variant strongly affecting one enhancer within the CRD could have domain-wide effects (Link, Duttke, et al., 2018). This concept was supported and extended here by HiChIP experiments. In addition to demonstrating that the IL-4 response was primarily associated with pre-existing enhancer-promoter connections, the HiChIP assay also captured a large number of enhancer-enhancer interactions. Examination of these connected enhancers provided evidence that a significant fraction of strain-differential enhancers lacking local variants were connected to strain-differential enhancers containing functional variants. An important future direction will be to further investigate the significance and mechanisms underlying these associations.

Collectively, these studies reveal general mechanisms by which noncoding genetic variation influences signal-dependent enhancer activity, thereby contributing to strain-differential patterns of gene expression and phenotypic diversity. A major future goal will be to

incorporate these findings into improved algorithms for prediction of absolute levels and dynamic responses of genes to IL-4 at the level of individual genes.

3.5 Materials and methods

3.5.1 Experimental Design

To investigate the influence of genetic variation on signal-dependent gene expression, enhancer activation and transcription factor binding, we performed RNA-sequencing, ATAC and ChIP-sequencing to study the responses of macrophages derived from five different inbred mouse (C57BL/6J, BALB/cJ, NOD/ShiLtJ, PWK/PhJ, and SPRET/EiJ) strains to the anti-inflammatory cytokine IL-4.

3.5.2 Mice

Female and male breeder mice for C57BL/6J, BALB/cJ, NOD/ShiLtJ, PWK/PhJ, and SPRET/EiJ mice were purchased from Jackson Laboratory. F1 C57 x SPRET mice were crossed and *Egr2*$^{fl/fl}$ mice were generously donated by dr. Lazarevic and dr. Warren (NIH) and crossed to *LyzM*-Cre mice (Jackson) to achieve myeloid specific targeted deletion of *Egr2*. Mice were housed at the UCSD animal facility on a 12h/12h light/dark cycle with free access to normal chow food and water. All animal procedures were in accordance with University of California San Diego research guidelines for the care and use of laboratory animals. 8-12-week-old healthy female mice were used for all our experiments.

3.5.3 Bone marrow-derived macrophage (BMDM) culture

Femur, tibia and iliac bones from the different mouse strains were flushed with DMEM high glucose (Corning) and red blood cells were lysed using red blood cell lysis buffer (eBioscience). After counting, 20 million bone marrow cells were seeded per 15cm non-tissue

culture plates in DMEM high glucose (50%) with 20% fetal bovine serum (FBS, Omega Biosciences), 30% L929-cell conditioned laboratory-made media (as source of M-CSF, as described before (Link, Duttke, et al., 2018)), 100 U/ml penicillin/streptomycin+L-glutamine (Gibco) and 2.5µg/ml Amphotericin B (HyClone). After 4 days of differentiation, 16.7 ng/ml mouse M-CSF (Shenandoah Biotechnology) was added to the media. After an additional 2 days of culture, adherent cells which were scraped and subsequently seeded onto tissue culture-treated petri dishes in DMEM containing 10% FBS, 100 U/ml penicillin/streptomycin+L-glutamine, 2.5µg/ml Amphotericin B and 16.7 ng/ml M-CSF. Macrophages were left untreated or treated with 20 ng/mL mouse recombinant IL-4 (Peprotech) for 1, 6 or 24 hours.

3.5.4 Immunofluorescence

Cells were fixed with Cytofix/Cytoperm Buffer (BD, BD554714) for 10 min at room temperature. Cytofix/Cytoperm buffer was removed, and cells were washed twice with HBSS containing 2% BSA and 1mm EDTA. Cells were kept in permeabilization/wash buffer (BD, BD554714) for one hour at 4C or until the experiment was performed. Fixed cells were blocked using 3% BSA, 0.1% Triton-PBS for 30 min at room temperature and then with 1/200 of the EGR2 antibody (abcam) overnight at 4C. Next day, cells were washed with 0.1% Triton-PBS, incubated with 1/200 donkey anti-rabbit 555 (ThermoFisher, #A31572) secondary antibody, phalloidin (abcam, ab176759) for staining actin filaments and nuclei were counter-stained with DAPI. After washing with 0.1% Triton-PBS, slides were mounted with Prolong Gold Antifade Reagent (Life Technology, #10144). Images were taken using a Leica SP8 with light deconvolution microscope.

3.5.5 RNA-seq library preparation

Total RNA was isolated from cells and purified using RNA Directzol micro prep columns and RNase-free DNase digestion according to the manufacturer's instructions (Zymo Research). Sequencing libraries were prepared in biological replicates from polyA enriched mRNA as previously described (Link, Duttke, et al., 2018). Libraries were PCR-amplified for 9-14 cycles, size selected using TBE gels or one-sided 0.8X Ampure clean-up, quantified by Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) and 75bp single-end sequenced on a HiSeq 4000 or NextSeq 500 (Illumina).

3.5.6 Crosslinking for ChIP-seq

For histone marks, PU.1, C/EBPβ and RNA Pol2 ChIP-seqs, culture media was removed and plates were washed once with PBS and then fixed for 10 minutes with 1% formaldehyde (Thermo Fisher Scientific) in PBS at room temperature and reaction was then quenched by adding glycine (Thermo Fisher Scientific) to 0.125M. For STAT6, PPARγ, and EGR2 ChIP-seq, cells were cross-linked for 30 minutes with 2mM DSG (Pierce) in PBS at room temperature. Subsequently cells were fixed for 10 minutes with 1% formaldehyde at room temperature and the reaction was quenched with 0.125M glycine. After fixation, cells were washed once with cold PBS and then scraped into supernatant using a rubber policeman, pelleted for 5 minutes at 400xG at 4°C. Cells were transferred to Eppendorf DNA LoBind tubes and pelleted at 700xG for 5 minutes at 4°C, snap-frozen in liquid nitrogen and stored at -80°C until ready for ChIP-seq protocol preparation.

3.5.7 Chromatin immunoprecipitation

Chromatin immunoprecipitation (ChIP) was performed in biological replicates as described previously (Seidman et al., 2020). Samples were sonicated using a probe sonicator in

500 µl lysis buffer (10 mM Tris/HCl pH 7.5, 100 mM NaCl, 1 mM EDTA, 0.5mM EGTA, 0.1% deoxycholate, 0.5% sarkozyl, 1 × protease inhibitor cocktail). After sonication, 10% Triton X-100 was added to 1% final concentration and lysates were spun at full speed for 10 minutes. 1% was taken as input DNA, and immunoprecipitation was carried out overnight with 20 µl Protein A Dynabeads (Invitrogen) and 2 µg specific antibodies for PU.1 (Santa Cruz, sc-352X), H3K4me2 (Millipore, 07-030), H3K4me3 (Millipore, 04-745), H3K27ac (Active Motif, 39135), RNA Pol2 (Genetex, GTX102535), STAT6 (Santa Cruz, sc-374021), EGR2 (abcam, ab43020) and C/EBP-β (Santa Cruz, sc-150). Beads were washed three times each with wash buffer I (20mM Tris/HCl, 150mM NaCl, 0.1% SDS, 1% Triton X-100, 2mM EDTA), wash buffer II (10mM Tris/HCl, 250mM LiCl, 1% IGEPAL CA-630, 0.7% Na-deoxycholate, 1mM EDTA), TE 0.2% Triton X-100 and TE 50mM NaCl and subsequently resuspended 25 µl 10 mM Tris/HCl pH 8.0 and 0.05% Tween-20 and sequencing libraries were prepared on the Dynabeads as described below.

For PPAR-γ ChIP-seq, fixed cells were lysed in 500 µl RIPA lysis buffer (20 mM Tris/HCl pH7.5, 1 mM EDTA, 0.5 mM EGTA, 0.1% SDS, 0.4% Na-Deoxycholate, 1% NP-40 alternative, 0.5 mM DTT, 1x protease inhibitor cocktail (Sigma)) and chromatin was sheared using a probe sonicator. 1% was taken as input DNA, and immunoprecipitation was carried out overnight with 20 µl Protein A Dynabeads (Invitrogen) and 2 µg of both PPAR-γ antibodies (Santa Cruz, sc-271392 and sc-7273). Beads were then collected using a magnet and washed with 175 µl ice cold buffer as indicated by incubating samples on ice for 3 minutes: three times RIPA wash buffer (20 mM Tris/HCl pH7.5, 1 mM EDTA, 0.5 mM EGTA, 0.1% SDS, 0.4% Na-Deoxycholate, 1% NP-40 alternative, 0.5 mM DTT, 1x protease inhibitor cocktail (Sigma)), six times LiCl wash buffer (10 mM Tris/HCl pH7.5, 250mM LiCl, 1 mM EDTA, 0.7% Na-

Deoxycholate, 1% NP-40 alternative, 1x protease inhibitor cocktail (Sigma)), twice with TET (10 mM Tris/HCl pH 8.0, 1 mM EDTA, 0.2% Tween-20, 1x protease inhibitor cocktail (Sigma)), and once with TE-NaCl (10 mM Tris/HCl pH 8.0, 0.1 mM EDTA, 50 mM NaCl, 1x protease inhibitor cocktail (Sigma)). Bead complexes were resuspended in 25 µl TT (10 mM Tris/HCl pH 8.0, 0.05% Tween-20) and sequencing libraries were prepared on the Dynabeads as described below.

3.5.8 ChIP-seq library preparation

ChIP libraries were prepared while bound to Dynabeads using NEBNext Ultra II Library preparation kit (NEB) using half reactions. DNA was polished, polyA-tailed and ligated after which dual UDI (IDT) or single (Bioo Scientific) barcodes were ligated to it. Libraries were eluted and crosslinks reversed by adding to the 46.5 µl NEB reaction 16 µl water, 4 µl 10% SDS, 4.5 µl 5M NaCl, 3 µl 0.5 M EDTA, 4 µl 0.2M EGTA, 1 µl RNAse (10 mg/ml) and 1 µl 20 mg/ml proteinase K, followed by incubation at 55C for 1 hour and 75C for 30 minutes in a thermal cycler. Dynabeads were removed from the library using a magnet and libraries were cleaned up by adding 2 µl SpeedBeads 3 EDAC (Thermo) in 124 µl 20% PEG 8000/1.5 M NaCl, mixing well, then incubating at room temperature for 10 minutes. SpeedBeads were collected on a magnet and washed two times with 150 µl 80% ethanol for 30 seconds. Beads were collected and ethanol removed following each wash. After the second ethanol wash, beads were air dried and DNA eluted in 12.25 µl 10 mM Tris/HCl pH 8.0 and 0.05% Tween-20. DNA was amplified by PCR for 14 cycles in a 25 µl reaction volume using NEBNext Ultra II PCR master mix and 0.5 µM each Solexa 1GA and Solexa 1GB primers. Libraries were size selected using TBE gels for 200 – 500 bp and DNA eluted using gel diffusion buffer (500 mM ammonium acetate, pH 8.0, 0.1% SDS, 1 mM EDTA, 10 mM magnesium acetate) and purified using ChIP DNA Clean

& Concentrator (Zymo Research). Sample concentrations were quantified by Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) and 75bp single-end sequenced on HiSeq 4000 or NextSeq 500 (Illumina).

3.5.9 ATAC-seq library preparation

Approximately 80k cells were lysed in 50 µl room temperature ATAC lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630), 2.5 µL DNA Tagmentation Enzyme mix (Nextera DNA Library Preparation Kit, Illumina) was added. The mixture was incubated at 37°C for 30 minutes and subsequently purified using the ChIP DNA purification kit (Zymo Research) as described by the manufacturer. DNA was amplified using the Nextera Primer Ad1 and a unique Ad2.n barcoding primers using NEBNext High-Fidelity 2X PCR MM for 8-14 cycles. PCR reactions were size selected using TBE gels for 175 – 350 bp and DNA eluted using gel diffusion buffer (500 mM ammonium acetate, pH 8.0, 0.1% SDS, 1 mM EDTA, 10 mM magnesium acetate) and purified using ChIP DNA Clean & Concentrator (Zymo Research). Samples were quantified by Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) and 75bp single-end sequenced on HiSeq 4000 or NextSeq 500 (Illumina).

3.5.10 H3K4me3 HiChIP

For H3K4me3 HiChIP, 10 million formaldehyde crosslinked cells per condition in biological replicates were used. HiChIP was performed as described before (Mumbach et al., 2016). In our experiments, 375 U of MboI (NEB, R0147M) restriction enzyme was used for chromatin digestion. Shearing was performed in three Covaris microtubes per sample and using the following parameters on a Covaris E220 (Fill Level = 6, Duty Cycle = 5, PIP = 140, Cycles/Burst = 200, Time = 200s). H3K4me3 IP was performed using 7.5 µg of antibody (Millipore, 04-745). Final PCR was performed using NEBNext High-Fidelity PCR MM and

Nextera general Primer Ad1 and specific Nextera Primer Ad2.n. PCR product was run on a TBE gel (Invitrogen) and libraries were size selected from 250bp to 700bp and cleaned up using 150 ul gel diffusion buffer (500 mM ammonium acetate, pH 8.0, 0.1% SDS, 1 mM EDTA, 10 mM magnesium acetate) and purified using ChIP DNA Clean & Concentrator (Zymo Research). Samples were quantified by Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) and 75bp paired-end sequenced on a NextSeq 500 (Illumina).

3.5.11 Data mapping

Custom genomes were generated for BALB/cJ, NOD/ShiLtJ, PWK/PhJ, and SPRET/EiJ mice from the C57BL/6J or mm10 genome as before (Link, Duttke, et al., 2018) using MMARGE v1.0 (Link, Romanoski, et al., 2018) and the VCF files from the Mouse Genomes Project (Keane et al., 2011). Data generated from different mouse strains were first mapped to their respective genomes using STAR v2.5.3 (Dobin et al., 2013) for RNA-seq data, or bowtie2 v2.2.9 (Langmead & Salzberg, 2012) for ATAC-seq, ChIP-seq, and HiChIP data. Then the mapped data was shifted to the mm10 genome using the MMARGE v1.0 'shift' function (Link, Romanoski, et al., 2018) for downstream comparative analyses.

3.5.12 RNA-seq data analysis

3.5.12.1 RNA-seq data processing

Transcripts were quantified using HOMER v4.11.1 "analyzeRepeats" script (Heinz et al., 2010). TPM values were reported by using the parameters -count exons -condenseGenes -tpm. Log-scaled TPM values were computed by log2(TPM+1). Raw read counts within transcripts were reported by using the parameters -count exons -condenseGenes -noadj. Differentially expressed genes were identified by feeding raw read counts into DESeq2 (Love et al., 2014) through the "getDiffExpression" script of HOMER. IL-4-induced and IL-4-repressed genes were

called by fold changes greater than 2 or less than half, respectively, together with q-values smaller than 0.05. Gene ontology analysis was performed using Metascape (Y. Zhou et al., 2019).

3.5.12.2 Categorization of strain-differential genes

Strain-differential genes were defined based on pairwise comparisons between C57 and one of the other strains as being called IL-4-induced or IL-4-repressed in one strain but not in the other. Strain-differential IL-4-induced genes were further classified into three categories based on the relative level of basal expression between the induced strain versus the non-induced strain: high basal, equal basal, and low basal. In the "high basal" group, the non-induced strain has at least 1.5-fold greater basal expression level than the induced strain. The direction of difference flipped for the "low basal" group where the induced strain has over 1.5-fold greater basal expression than the non-induced strain. The genes in between are categorized into the "equal basal" group.

3.5.12.3 F1 mice data processing

RNA-seq data from F1 mice was mapped to both parental genomes (C57 and SPRET) and analyzed in the same way as before (Link, Duttke, et al., 2018). In short, the read counts for each transcript were multiplied by the ratio of reads overlapping mutations time 10 and assigned to the parental genomes. Transcripts without any assigned reads in one of the F1 alleles were filtered out. To determine *cis* versus *trans* effects of genetic variation on gene expression, the difference of fold change between parental alleles and F1 alleles were calculated. The genes with majorly *cis* effects were defined by $-1 < \log_2(\text{parental fold change}) - \log_2(\text{F1 fold change}) < 1$, while those with majorly trans effects were defined by F1 fold change < parental fold change for genes with over +/- 2 fold-change in parental alleles.

3.5.13 WGCNA analysis

For each strain a differential gene expression analysis was performed to compare IL-4 to basal with Limma Voom (Law et al., 2014). A linear model was fit for all 5 differential comparisons at once, and 1912 genes that were significant with q-value below 0.05 and an absolute fold change of 1.5 in any comparison where included in a Weighted gene co-expression network analysis (WGCNA) (Langfelder & Horvath, 2008). WGCNA was performed with a softpower value of 20, and a signed network was generated. Modules were cut with min module size of 50 and cut-height of 0.999 including PAM-stage. 9 modules were detected of which 2 genes were part of the grey (non-connected) module which was subsequently excluded. Module Eigengenes were calculated and visualized using the verbose-boxplots function that also performed a Kruskall Wallis significance test to test whether all ME values belong to the same distribution and all modules were significantly different between conditions (all P-values below < 0.0012). Two modules exhibited consistent differential expression between IL-4 and notx across strains, while the other 6 modules were most prominently influenced in a strain specific manner. Modules were annotated with Metascape (Y. Zhou et al., 2019).

3.5.14 ATAC-seq and ChIP-seq data analysis

Based on the HOMER tag directories created from mapped sequencing data, the reproducible ATAC-seq and transcription factor ChIP-seq peaks were identified by using HOMER to call unfiltered 200-bp peaks (parameters -L 0 -C 0 -fdr 0.9 -size 200) and running IDR v2.0.3 on replicates of the same sample with the default parameters (Li et al., 2011). The levels of histone modifications and RNA polymerase II were quantified within +/- 500 bp around the centers of ATAC-seq reproducible peaks using HOMER annotatePeak.pl with parameters "-size -500,500 -norm 1e7". The transcription factor binding intensities were quantified within +/-

300 bp around the identified ChIP-seq peaks using parameters "-size -150,150 -norm 1e7". For comparisons across multiple samples (e.g., different time points, mouse strains, transcription factors), we merged the set of peaks first using HOMER mergePeaks "-d given" before quantifying the features above. To visualize the average profile of a dataset around a certain set of peaks, we used HOMER annotatePeaks.pl with parameters "-norm 1e7 -size 4000 -hist 20" to help compute the histograms of 20-bp bins within +/- 2000 bp regions.

3.5.15 Identification of IL-4 responsive regulatory elements

IL-4 responsive enhancers were identified by the strong fold changes of H3K27ac and RNA Pol2 at intergenic or intronic open chromatin. Reproducible ATAC peaks called from each mouse strain for the basal and IL-4 conditions were first merged and then annotated for genomic positions and the enrichment of H3K27ac and RNA Pol2 within +/- 500 bp using HOMER v4.11.1. Based on the genomic annotations from HOMER annotatePeaks.pl, we classified regions at promoter-TSS as promoters and regions at intergenic or intronic positions as enhancers. Regions with less than 16 normalized tags of H3K27ac or less than 8 normalized tags of RNA Pol2 were filtered out. For the remaining promoters and enhancers, we computed the fold changes of the normalized tags of H3K27ac and RNA Pol2 between basal and IL-4 conditions for each mouse strain. Regions were called IL-4 induced or IL-4 repressed if there were at least 2.5-fold increases or decreases, respectively, from basal to IL-4 state for both histone markers. Regions with less than 1.4-fold changes were called neutral elements.

3.5.16 Super enhancer

We used ROSE to call super enhancers for the five mouse strains (Whyte et al., 2013). The active enhancers were first merged within each strain for both basal and IL-4 conditions to obtain a set of starting conventional enhancers. Then the ROSE algorithm was run for each strain

on the mapped H3K27ac ChIP-seq data with parameter "-t 2500" to exclude TSS. The overall activity of a super enhancer was quantified by the H3K27ac ChIP-seq read counts within the entire identified super enhancer region.

3.5.17 H3K4me3 HiChIP

3.5.17.1 H3K4me3 ChIP-seq HiChIP reference preprocessing

H3K4me3 ChIP-seqs from basal and 24 h IL-4 stimulated macrophages were performed in duplicate with input controls. Fastq files were aligned with bowtie2 (Langmead & Salzberg, 2012) to the mm10 reference genome and peak calling was done with MACS (Zhang et al., 2008) for each replicate separately. Significant peaks were merged using bedtools (Quinlan & Hall, 2010) into a general bed file that was used as corresponding peak-file for MAPS.

3.5.17.2 H3K4me3 HiChIP preprocessing

HiChIP-seq data was processed with MAPS (Juric et al., 2019) at 5000-bp resolution as described previously for PLAC-seq (Nott et al., 2019) for all four samples separately, basal and 24 h IL-4 duplicate samples combined, and a merge of all four samples.

3.5.17.3 Differential analysis

In order to identify interactions that were significantly stronger in IL4 or control, a differential analysis was performed as described in (Nott et al., 2019). Briefly, significant interactions that were identified in the combined duplicate analysis of IL-4 and control were merged in a general interaction set. Paired end read counts that fell within these interactions were quantified for each sample separately. The quantified matrix of all significant interactions for all cell types was used as input for Limma (Ritchie et al., 2015) differential interaction analysis. A linear model was fit, with one pairwise contrast (IL4 vs control), with and without batch correction. No interactions were identified that were significantly different between IL4 and

control by either method (FDR < 0.1, and absolute log2 FC > 1). Hence, the combined interaction set (generated using both IL4 and control samples) was used for downstream analysis.

3.5.18 Interactions among promoters and enhancers

Significant interactions captured by HiChIP-seq were overlapped with previously identified active promoters and enhancers for the five mouse strains using HOMER mergePeaks "-d 2500" in order to identify three categories of interactive pairs: enhancer-enhancer, enhancer-promoter, and promoter-promoter. Enhancer-promoter interactions have enhancers on one end and promoters on the other end, while enhancer-enhancer or promoter-promoter interactions are the linked pairs of enhancers or promoters, respectively. We ended up with 145,907 enhancer-enhancer interactions, 81,411 enhancer-promoter interactions, and 10,710 promoter-promoter interactions. To better understand the regulatory landscape associated with IL-4 stimulation, we subsequently focused on enhancer-promoter interactions that contained IL-4 induced, repressed and/or neutral promoters on one end, and IL-4 induced, neutral, and or repressed enhancers on the other end, and quantified the number of interactions between these possible promoter-enhancer combinations in 9 categories as a contingency table. Fisher's exact test was applied to the contingency table to determine if the any categories were significantly different for three comparisons of interest: IL-4 induced enhancer/promoter interactions vs non-induced enhancer/promoters; IL-4 repressed enhancer/promoter interactions vs non-repressed enhancer/promoters; and IL-4 induced enhancer/promoter interactions vs IL4 repressed enhancer/promoter interactions. For enhancer-enhancer interactions, we pre-selected enhancers that have at least 4-fold difference in H3K27ac ChIP-seq tags between any two strains under the 24 h IL-4 condition to obtain a set of strongly strain-differential enhancers. We then computed the Pearson correlation of H3K27ac tags across the five strains for every pair of interactive

enhancers among the pre-selected set. To obtain non-interactive enhancers, we either randomly paired pre-selected enhancers on the same chromosome (same-chromosome random enhancers) or looked for enhancers within certain distances but not connected based on our data (distance-matched random enhancers).

3.5.19 Genetic variants at local and connected enhancers

Genetic variation between C57 and the other four strains at strain-differential enhancers was extracted using MMARGE annotate_mutations (Link, Romanoski, et al., 2018), which was based on the VCF files from the Mouse Genomes Project (Keane et al., 2011). Variants were searched within +/- 150 bp around the centers of enhancers. At least one genetic variant from the comparative strain needs to be present within the search area for such enhancer to be counted as having variants.

3.5.20 Motif analysis

3.5.20.1 Motif enrichment analysis

Given a certain set of peaks, we used HOMER findMotifsGenome.pl with parameters "-size 200 -mask" to identify de novo motifs and their matched known motifs (Heinz et al., 2010). The background sequences were either the default random sequences or a different set of peaks from a comparative condition in the main text and in the figure legends.

3.5.20.2 Motif mutation analysis

To integrate the genetic variation across mouse strains into motif analysis, we used MAGGIE, which is able to identify functional motifs out of the currently known motifs by testing for the association between motif mutations and the changes in specific epigenomic features (Shen et al., 2020). The known motifs are obtained from the JASPAR database (Fornes et al., 2020). We applied this tool to strain-differential IL-4-responsive enhancers and

transcription factor binding sites. Strain-differential IL-4 responsive enhancers were defined as previously described for KLA-responsive enhancers (Shen et al., 2020). In brief, from every pairwise comparison across the five strains, enhancers identified as “IL-4 activated” or “IL-4 repressed” only in one of the compared strains were called strain-differential and were pooled together. For enhancer sites to be included in the analysis, enhancer activity had to be differentially regulated between two strains. As required by MAGGIE, sequences from the genomes of the responsive strains were input as “positive sequences”, and those from the other strains as “negative sequences”. Strain-differential transcription factor binding sites were defined by reproducible ChIP-seq peaks called in one strain but not in the other. “Positive sequences” and “negative sequences” were specified as sequences from the bound and unbound strains, respectively. The output p-values with signs indicating directional associations were averaged for clusters of motifs grouped by a maximum correlation of motif score differences larger than 0.6. Only motif clusters with at least one member showing a corresponding gene expression larger than 2 TPM in BMDMs were shown in figures.

3.5.21 Categorization of IL-4-induced enhancers

Among the strain-differential IL-4-induced enhancers as described above, we further split them into three categories based on the level of H3K27ac under the basal condition in non-induced strains. “High basal” enhancers have more than 2-fold stronger H3K27ac in non-induced strains, while “low basal” enhancers have more than 2-fold stronger H3K27ac in induced strains (lower basal H3K27ac in non-induced strains). “Equal basal” enhancers are those in between.

3.5.22 Deep learning

3.5.22.1 Neural network training

We adapted a similar strategy as AgentBind (Zheng et al., 2021) for our training procedure. We started with a pre-trained DeepSEA (J. Zhou & Troyanskaya, 2015) model consisting of three convolutional layers and two fully connected layers and then fine-tuned it to generate three models based on our data: IL-4 active enhancers vs. random backgrounds (auROC = 0.894), IL-4 induced enhancers vs. random backgrounds (auROC = 0.919), and IL-4 induced enhancers vs. non-induced enhancers (auROC = 0.796). The enhancer sequences were extended to 300-bp long. In all experiments, we left out sequences on chromosome 8 for cross validation and sequences on chromosome 9 for testing. IL-4 active enhancers and non-induced enhancers were from C57 mice, while IL-4 induced enhancers were pooled from all the five strains in order to reach a comparable sample size. Random genomic backgrounds were generated by randomly selecting nearby GC-matched equal-length sequences on the mm10 genome. We applied binary cross-entropy as the loss function. During each training, the initial learning rate was set as 1e-4 and reduced by a factor of 0.9 when learning stagnated. The training process stopped when the loss value had not decreased for more than 20 epochs.

3.5.22.2 DeepLIFT and importance score

We used DeepLIFT (Shrikumar et al., 2017) to generate importance scores with single-nucleotide resolution using uniform nucleotide backgrounds. For each input sequence, we generated two sets of scores, one for the original sequence and the other for its reverse complement. The final scores were the absolute maximum at each aligned position. We defined predicted functional nucleotides by the top 20% (i.e., top 60) positions within each input 300-bp sequence. To interpret the most important sequence patterns learned by neural networks, we

computed the odds ratio of each 5-mer within top 10% of all 5-mers (Zheng et al., 2021). Fisher's Exact test was performed to determine whether 5-mers were enriched. We used TOMTOM (Gupta et al., 2007) to match 5-mers with known transcription factor binding motifs.

3.5.23 Data and code availability

All sequencing data have been made available by deposition in the GEO database: GSE159630. The UCSC genome browser was used to visualize sequencing data. The codes for neural network model training and interpretation are available on our Github repository: https://github.com/zeyang-shen/macrophage_IL4Response.

3.5.24 Statistical Analysis

Two independent groups were tested using Mann–Whitney U test for medians and using Levene's test for variance. Gene expression comparisons were reported by adjusted p-values (i.e., q-values) from DESeq2 (Love et al., 2014). Enrichment was computed by odds ratio and tested by Fisher's exact test. Effect sizes were reported by Cohen's d. All gene expression data are displayed as means with 95% confidence interval. All data distributions are shown with means, 25th percentiles, and 75th percentiles.

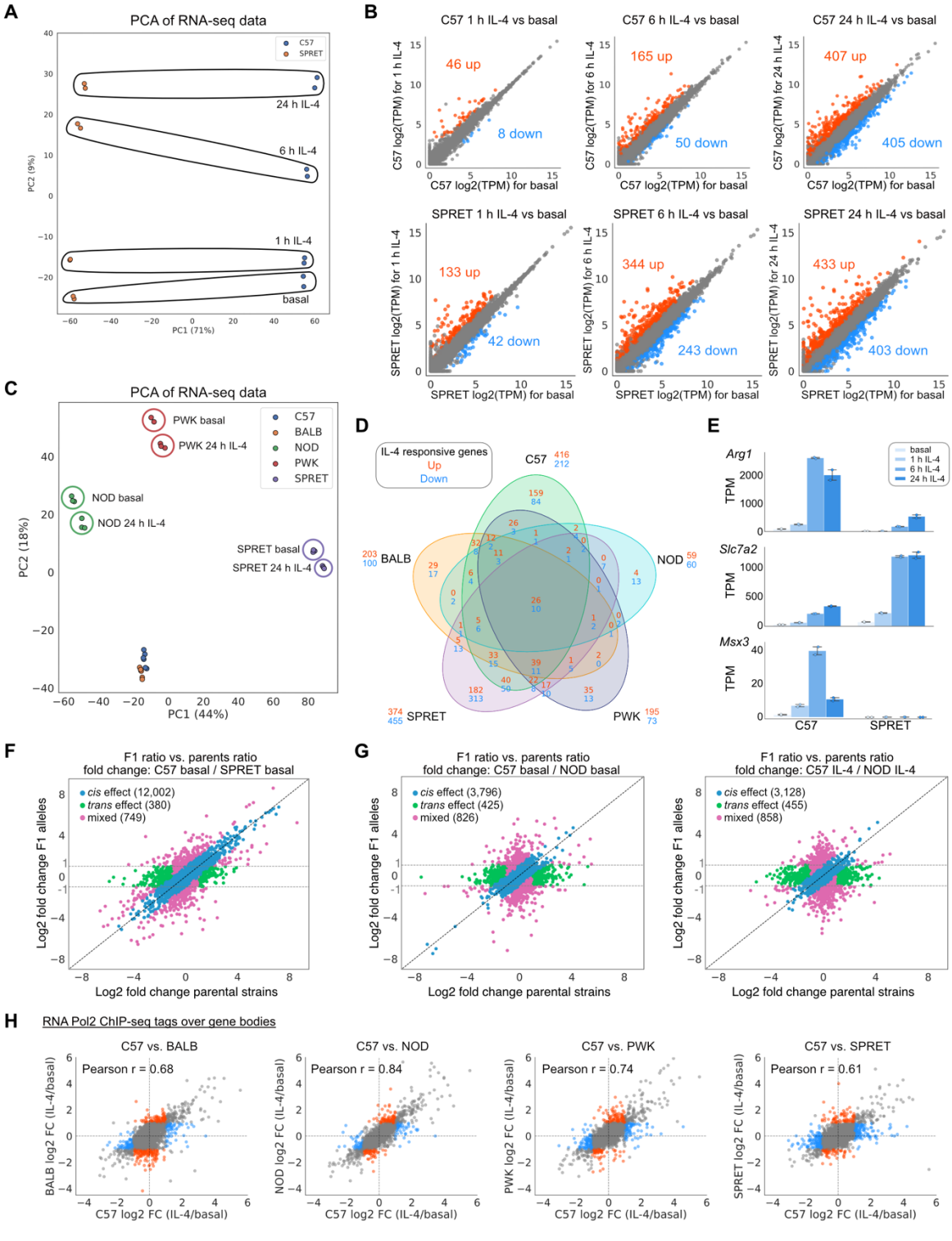
3.6 Acknowledgements

I would like to thank J. Collier and J. Chang for technical assistance, the IGM core for library sequencing, L. Van Ael for assistance with manuscript preparation and Dr. Warren and Dr. Lazarevic for donating *Egr2*$^{fl/fl}$ mice. These studies were supported by NIH grants DK091183 and HL147835 and a Leducq Transatlantic Network grant 16CVD01. Sequencing costs were partially supported by DK063491.

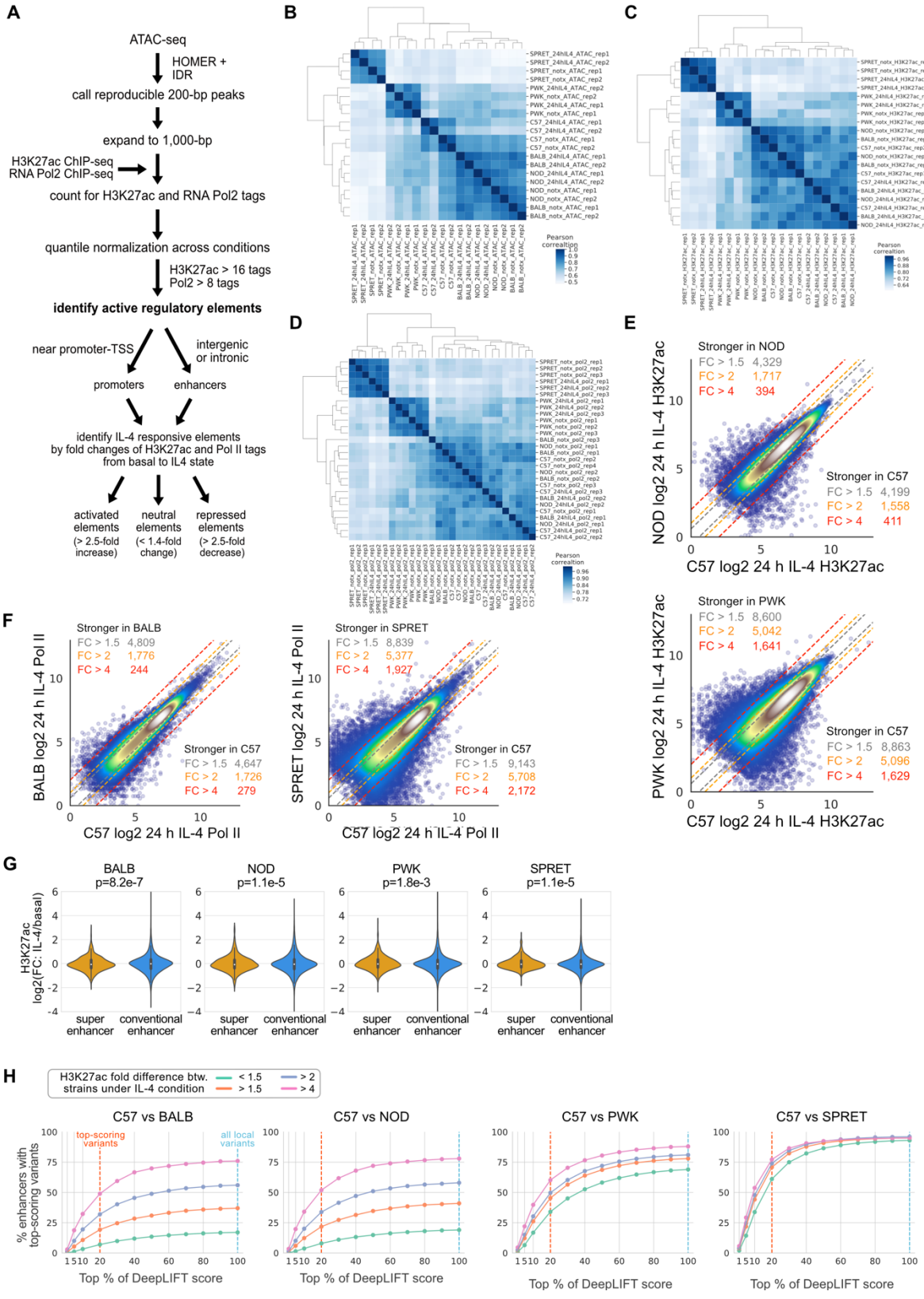
Chapter 3, in full, has been accepted for publication of the material as it will appear in Hoeksema MA*, Shen Z*, Holtman IR, Zheng A, Spann N, Cobo I, Gymrek M & Glass CK. (2020). Mechanisms underlying divergent responses of genetically distinct macrophages to IL-4. Science Advances. (*These authors contributed equally to this work). The dissertation author was a primary investigator and author of this paper.

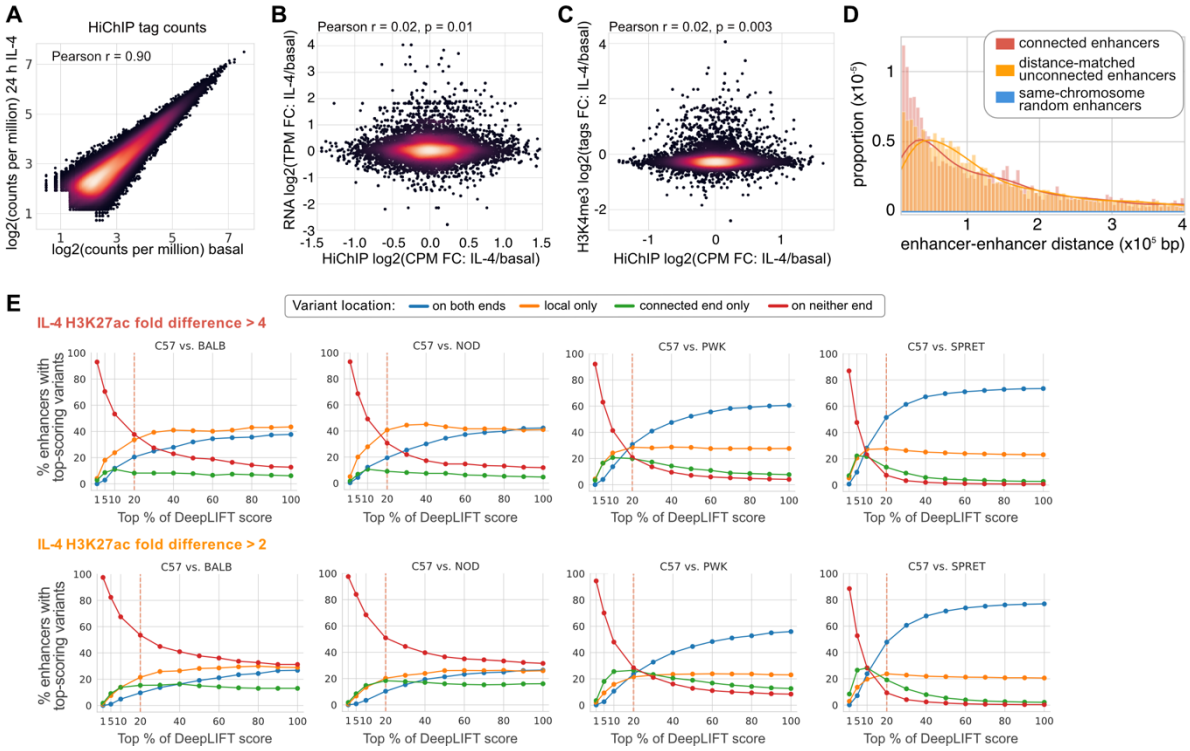
3.7 Supplementary figures

Supplementary Figure 3.1 Response to IL-4 is slow and highly divergent in macrophages from different mouse strains. (A) PCA plot showing the variance in RNA-seq IL-4 time course data in C57 and SPRET macrophages. **(B)** Scatter plot showing the effects of 1 h, 6 h and 24 h IL-4 stimulation on gene expression in C57 and SPRET BMDMs (n=2 per condition). **(C)** PCA plot showing the variation in macrophages from different strains in response to 24 h IL-4. **(D)** Venn diagram of the IL-4 response in macrophages from the five different strains. Repressed and activated genes are plotted that have a twofold change and an q-value<0.05 between untreated and IL-4 stimulated conditions. **(E)** Kinetics of IL-4 induced gene expression for *Arg1*, *Slc7a2* and *Msx3* in C57 and SPRET macrophages. **(F)** Ratio-ratio fold change plots of allele-specific RNA-seq reads in F1 (C57xSPRET) vs parents C57 or SPRET under basal conditions. **(G)** Ratio-ratio fold change plots of allele-specific RNA-seq reads in F1 (C57xNOD) vs parents C57 or NOD under basal and 24h IL-4 conditions. **(H)** Ratio-ratio plots demonstrating the RNA Pol2 over gene bodies in response to IL-4 in pairwise comparisons.


A
PCA of RNA-seq data
C57
SPRET
24 h IL-4
6 h IL-4
1 h IL-4
basal
PC1 (71%)
PC2 (9%)
B
C57 1 h IL-4 vs basal
46 up
8 down
C57 6 h IL-4 vs basal
165 up
50 down
C57 24 h IL-4 vs basal
407 up
405 down
SPRET 1 h IL-4 vs basal
133 up
42 down
SPRET 6 h IL-4 vs basal
344 up
243 down
SPRET 24 h IL-4 vs basal
433 up
403 down
C57 log2(TPM) for basal
SPRET log2(TPM) for basal
C
PCA of RNA-seq data
PWK basal
PWK 24 h IL-4
NOD basal
NOD 24 h IL-4
SPRET basal
SPRET 24 h IL-4
C57
BALB
NOD
PWK
SPRET
PC1 (44%)
PC2 (18%)
D
IL-4 responsive genes
Up
Down
C57
BALB
NOD
SPRET
PWK
E
Arg1
Slc7a2
Msx3
TPM
basal
1 h IL-4
6 h IL-4
24 h IL-4
C57
SPRET
F
F1 ratio vs. parents ratio
fold change: C57 basal / SPRET basal
cis effect (12,002)
trans effect (380)
mixed (749)
Log2 fold change F1 alleles
Log2 fold change parental strains
G
F1 ratio vs. parents ratio
fold change: C57 basal / NOD basal
cis effect (3,796)
trans effect (425)
mixed (826)
F1 ratio vs. parents ratio
fold change: C57 IL-4 / NOD IL-4
cis effect (3,128)
trans effect (455)
mixed (858)
H
RNA Pol2 ChIP-seq tags over gene bodies
C57 vs. BALB
Pearson r = 0.68
C57 vs. NOD
Pearson r = 0.84
C57 vs. PWK
Pearson r = 0.74
C57 vs. SPRET
Pearson r = 0.61
BALB log2 FC (IL-4/basal)
NOD log2 FC (IL-4/basal)
PWK log2 FC (IL-4/basal)
SPRET log2 FC (IL-4/basal)
C57 log2 FC (IL-4/basal)

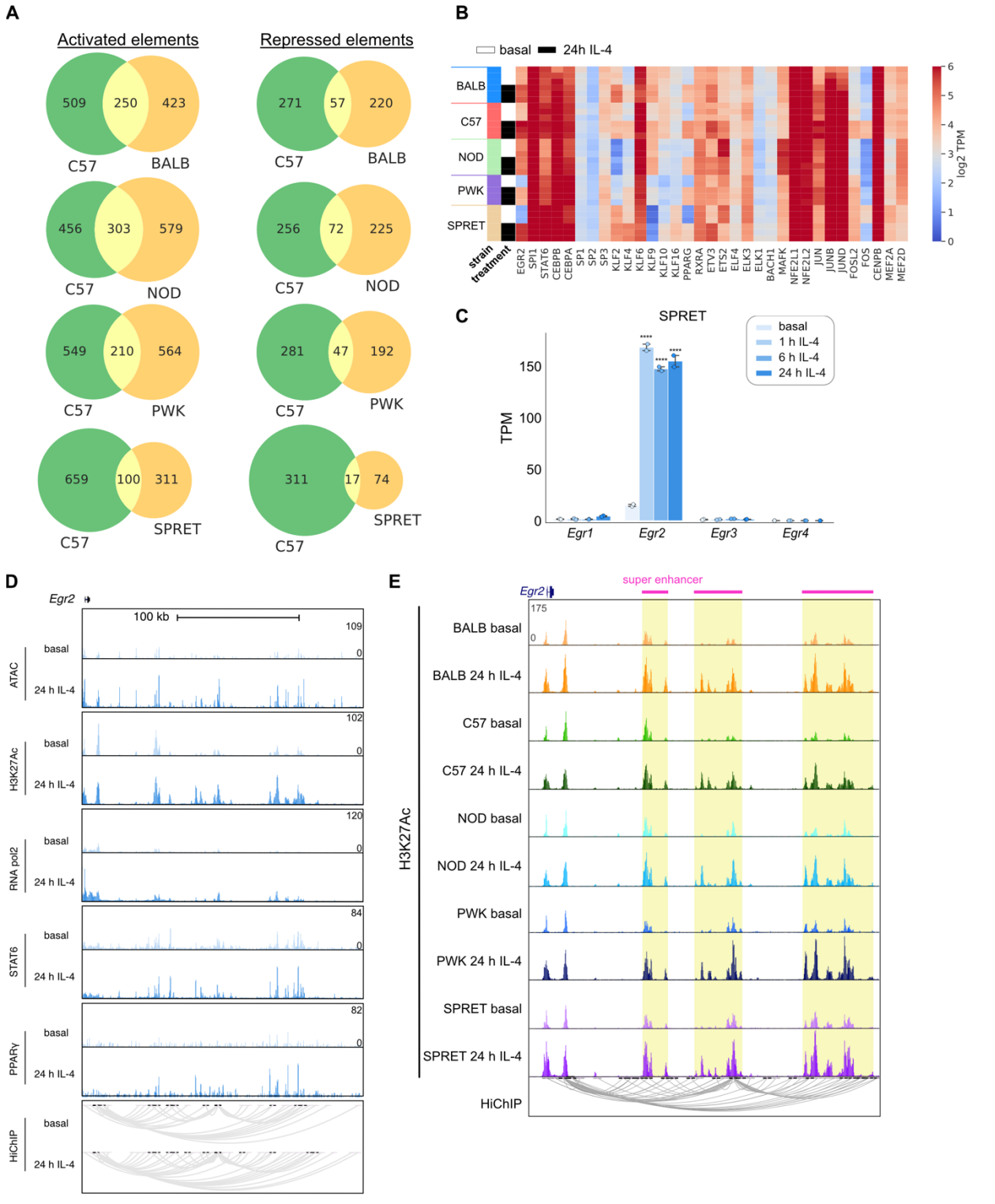
Supplementary Figure 3.2 Strain-differential IL-4 induced gene expression is the result of differential IL-4 enhancer activation in macrophages derived from genetically diverse mice. (A) Enhancer and promoter selection criteria, including criteria for activated, neutral or repressed elements. **(B)** Clustering of ATAC-seq data in strains macrophages stimulated with IL-4 for 24 h. **(C)** Clustering of H3K27ac ChIP-seq data in strains macrophages stimulated with IL-4 for 24 h. **(D)** Clustering of RNApolII ChIP-seq data in strains macrophages stimulated with IL-4 for 24 h. **(E)** Comparison of C57 ATAC peaks with H3K27ac signal to those of NOD or PWK under IL4 treatment conditions. **(F)** Comparison of C57 ATAC peaks with RNA Pol2 signal to those of BALB or SPRET under IL4 treatment conditions. **(G)** Violin plots showing the difference in H3K27ac in response to 24 h IL-4 between super enhancers and conventional enhancers in BALB, NOD, PWK and SPRET macrophages. Mann–Whitney U test was performed to test the difference between super enhancers and conventional enhancers. **(H)** Percentages of enhancers that contain variants at high-ranked positions based on DeepLIFT scores using different cut-offs.

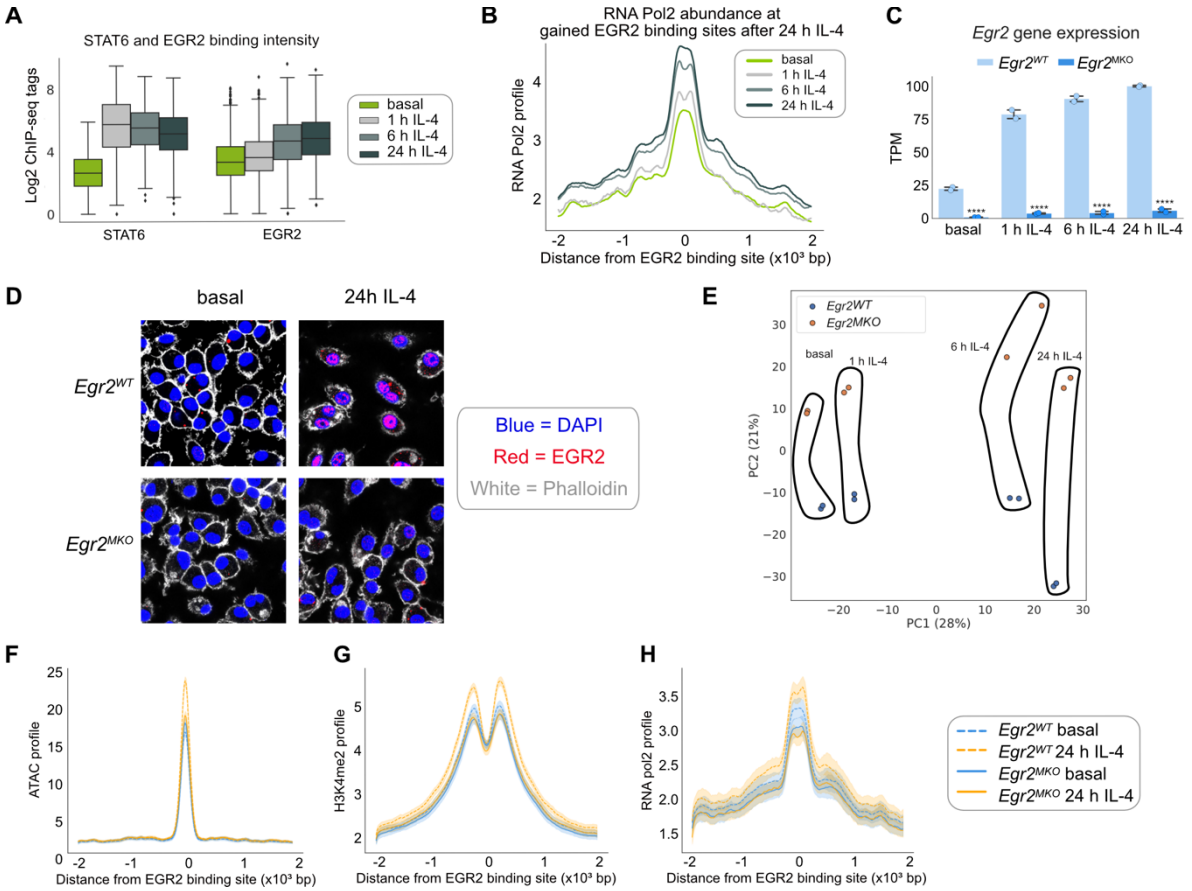

A
ATAC-seq
HOMER + IDR
call reproducible 200-bp peaks
expand to 1,000-bp
H3K27ac ChIP-seq
RNA Pol2 ChIP-seq
count for H3K27ac and RNA Pol2 tags
quantile normalization across conditions
H3K27ac > 16 tags
Pol2 > 8 tags
identify active regulatory elements
near promoter-TSS
intergenic or intronic
promoters
enhancers
identify IL-4 responsive elements by fold changes of H3K27ac and Pol II tags from basal to IL4 state
activated elements (> 2.5-fold increase)
neutral elements (< 1.4-fold change)
repressed elements (> 2.5-fold decrease)
B
C
D
E
Stronger in NOD
FC > 1.5 4,329
FC > 2 1,717
FC > 4 394
Stronger in C57
FC > 1.5 4,199
FC > 2 1,558
FC > 4 411
NOD log2 24 h IL-4 H3K27ac
C57 log2 24 h IL-4 H3K27ac
Stronger in PWK
FC > 1.5 8,600
FC > 2 5,042
FC > 4 1,641
Stronger in C57
FC > 1.5 8,863
FC > 2 5,096
FC > 4 1,629
PWK log2 24 h IL-4 H3K27ac
C57 log2 24 h IL-4 H3K27ac
F
Stronger in BALB
FC > 1.5 4,809
FC > 2 1,776
FC > 4 244
Stronger in C57
FC > 1.5 4,647
FC > 2 1,726
FC > 4 279
BALB log2 24 h IL-4 Pol II
C57 log2 24 h IL-4 Pol II
Stronger in SPRET
FC > 1.5 8,839
FC > 2 5,377
FC > 4 1,927
Stronger in C57
FC > 1.5 9,143
FC > 2 5,708
FC > 4 2,172
SPRET log2 24 h IL-4 Pol II
C57 log2 24 h IL-4 Pol II
G
BALB p=8.2e-7
NOD p=1.1e-5
PWK p=1.8e-3
SPRET p=1.1e-5
H3K27ac log2(FC: IL-4/basal)
super enhancer
conventional enhancer
H
H3K27ac fold difference btw. strains under IL-4 condition
< 1.5
> 1.5
> 2
> 4
C57 vs BALB
C57 vs NOD
C57 vs PWK
C57 vs SPRET
top-scoring variants
all local variants
% enhancers with top-scoring variants
Top % of DeepLIFT score



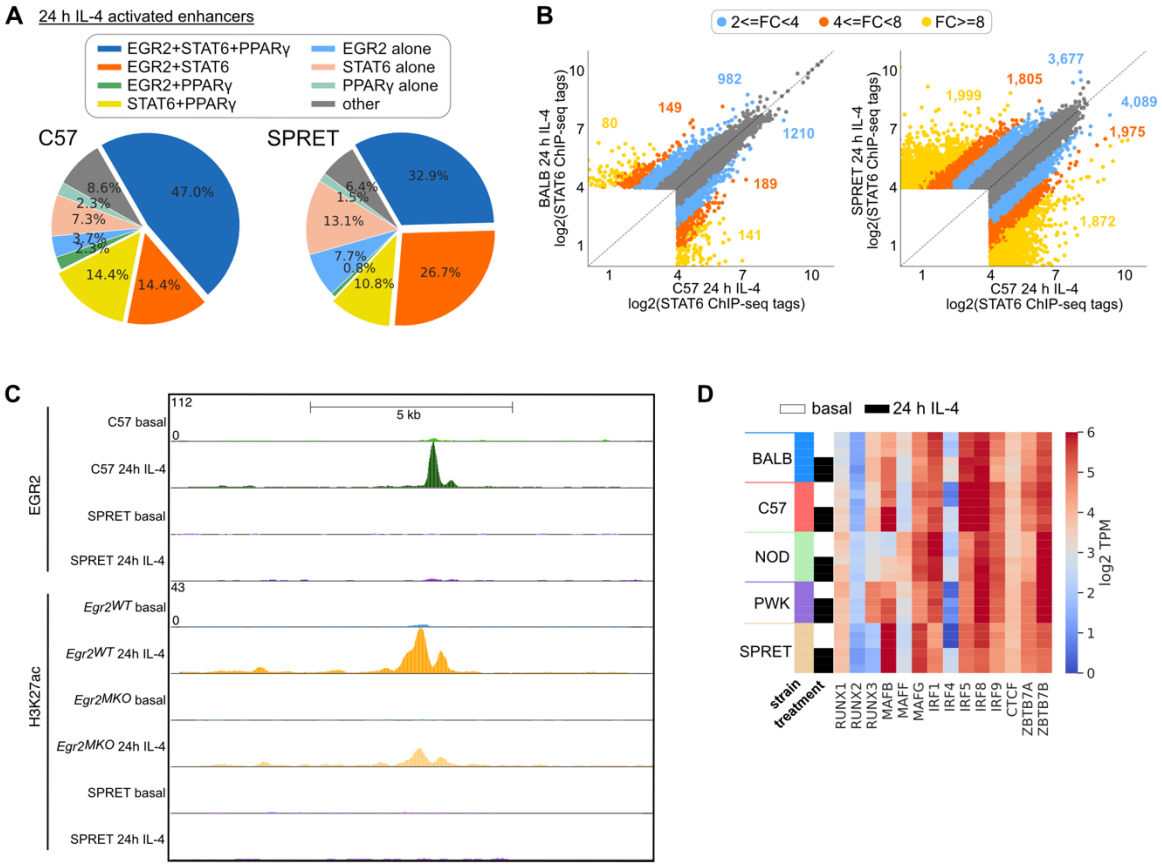
Supplementary Figure 3.3 IL-4 enhancers use pre-existent promoter-enhancer interactions to regulate gene activity. (A) The correlation of HiChIP reads between basal and 24 h IL-4 stimulated C57 macrophages. The reads were counted within the bins of both sides of the connection. Each dot represents a connection. **(B)** Comparison between HiChIP read changes and gene expression changes. **(C)** Comparison between HiChIP read changes and H3K4me3 signal changes. **(D)** Distance distributions of enhancer pairs. Distance-matched random enhancers have similar distances compared to connected enhancers, while the distances between same-chromosome random enhancers are spread out. **(E)** Percentages of interactive enhancers that contain predicted functional variants using different cut-offs for C57 versus the other strains.

Supplementary Figure 3.4 Enhancer mutation motif analysis identifies EGR2 to be strongly associated with late IL-4 enhancer activation. (A) Overlap in IL-4 enhancer activation and repression in pairwise comparisons between C57 and one of the other strains. **(B)** Gene expression of transcription factors whose motifs were identified significant by MAGGIE for enhancer activation in response to 24 h IL-4. **(C)** Gene expression of all Egr family members in SPRET BMDMs under basal conditions and after stimulation with IL-4 for 1 h, 6 h or 24 h. ****q<0.0001, compared to basal. **(D)** *Egr2* promoter connected to several upstream enhancers in C57 BMDMs as determined by H3K4me3 HiChIP. Connected enhancers bound by SDTFs STAT6 and PPARγ display increased H3K27ac and RNA Pol2 by IL-4. **(E)** Super enhancers of the *Egr2* gene that are strongly conserved in macrophage of the five different strains.


A
Activated elements
Repressed elements
509
250
423
C57
BALB
271
57
220
C57
BALB
456
303
579
C57
NOD
256
72
225
C57
NOD
549
210
564
C57
PWK
281
47
192
C57
PWK
659
100
311
C57
SPRET
311
17
74
C57
SPRET
B
basal
24h IL-4
BALB
C57
NOD
PWK
SPRET
strain
treatment
EGR2
SPI1
STAT6
CEBPB
CEBPA
SP1
SP2
SP3
KLF2
KLF4
KLF6
KLF9
KLF10
KLF16
PPARG
RXRA
ETV3
ETS2
ELF4
ELK3
ELK1
BACH1
MAFK
NFE2L1
NFE2L2
JUN
JUNB
JUND
FOSL2
FOS
CENPB
MEF2A
MEF2D
log2 TPM
C
SPRET
basal
1 h IL-4
6 h IL-4
24 h IL-4
TPM
Egr1
Egr2
Egr3
Egr4
D
Egr2
100 kb
ATAC
H3K27Ac
RNA pol2
STAT6
PPARγ
HiChIP
basal
24 h IL-4
E
super enhancer
Egr2
H3K27Ac
BALB basal
BALB 24 h IL-4
C57 basal
C57 24 h IL-4
NOD basal
NOD 24 h IL-4
PWK basal
PWK 24 h IL-4
SPRET basal
SPRET 24 h IL-4
HiChIP

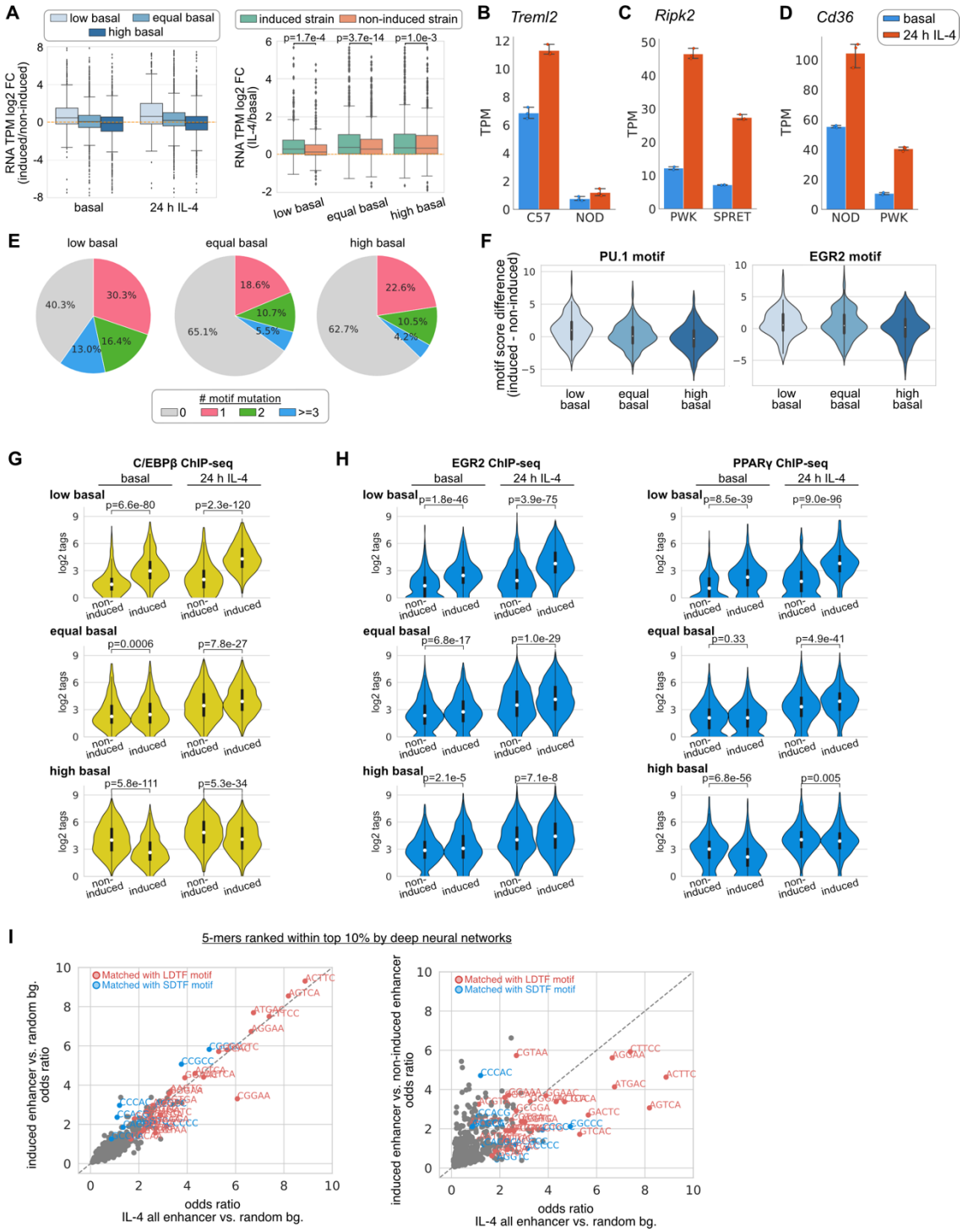


Supplementary Figure 3.5 *Egr2* deletion results in decreased IL-4 induced enhancer activation and gene expression. (A) STAT6 and EGR2 binding intensity as measured with ChIP-seq after IL-4 stimulation over time in C57 BMDMs. **(B)** Enhancer activity as measured with RNA Pol2 binding at 24 h IL-4 induced at intergenic and intronic EGR2 peaks in C57 BMDMs. **(C)** Efficient deletion of *Egr2* in *LyzM*-Cre^{+} *Egr2*^{flfl} (*Egr2* macrophage knock-out, *Egr2*^{MKO}) macrophages, one out of two representative experiments is shown, n=2 per condition. One out of two replicates is shown, ****q<0.0001, compared to *Egr2*^{WT} macrophages. **(D)** Immunofluorescence of EGR2 in combination with DAPI and Phalloidin in untreated and IL-4 stimulated *Egr2*^{WT} and *Egr2*^{MKO} macrophages, one out of two representative experiments is shown. **(E)** PCA plot showing the variance in RNA-seq samples IL-4 time course data in *Egr2*^{WT} and *Egr2*^{MKO} macrophages. **(F)** ATAC profile over IL-4 induced EGR2 peaks in *Egr2*^{WT} and *Egr2*^{MKO} macrophages under basal conditions and after 24 h IL-4 stimulation. **(G)** H3K4me2 profile over IL-4 induced EGR2 peaks in *Egr2*^{WT} and *Egr2*^{MKO} macrophages under basal conditions and after 24 h IL-4 stimulation. **(H)** RNA Pol2 binding at IL-4 induced EGR2 peaks in *Egr2*^{WT} and *Egr2*^{MKO} macrophages under basal conditions and after 24 h IL-4 stimulation. 90% confidence intervals are shown together with the average profiles.



Supplementary Figure 3.6 Collaborative and hierarchical transcription factor interactions at IL-4 dependent enhancers. (A) Overlap of binding as determined with ChIP-seq of the SDTFs STAT6 and PPARγ with EGR2 at IL-4 activated enhancers in C57 and SPRET BMDMs. **(B)** Scatter plots comparing binding of STAT6 in C57 versus BALB and C57 versus SPRET IL-4 stimulated BMDMs. **(C)** Example of a strain-differential activated enhancer upstream of the *Btbd11* gene which is decreased in *Egr2*^{MKO} BMDMs as well. **(D)** Gene expression of transcription factors that were identified significant from the MAGGIE analysis.

Supplementary Figure 3.7 Determinants of absolute levels and dynamic responses of IL-4 responsive enhancers. (A) Fold changes of gene expression between induced and non-induced strains for the three categories of enhancers (left). Gene expression changes between basal and 24 h IL-4 conditions (right). P-values from two-sample t-test were reported for the comparison between induced and non-induced strains. **(B)** *Treml2* gene expression in C57 and NOD macrophages. **(C)** *Ripk2* gene expression in PWK and SPRET macrophages. **(D)** *Cd36* gene expression in NOD and PWK macrophages. **(E)** Percentages of enhancers with motif mutations of the positively significant motifs according to MAGGIE results. (**F**) Score differences of PU.1 and EGR2 motifs between induced and non-induced strains for the three categories of enhancers. **(G)** C/EBPβ binding in non-induced and induced strains in the three different categories of enhancers. **(H)** EGR2 and PPARγ binding in non-induced and induced strains in the three different categories of enhancers. **(I)** Enrichment of 5-mers at top-ranked positions based on different neural network models. 5-mers matched with LDTF (red) or SDTF (blue) motifs based on significant results from TOMTOM are highlighted.


A
low basal
equal basal
high basal
RNA TPM log2 FC (induced/non-induced)
basal
24 h IL-4
induced strain
non-induced strain
p=1.7e-4
p=3.7e-14
p=1.0e-3
RNA TPM log2 FC (IL-4/basal)
low basal
equal basal
high basal
B
Treml2
TPM
C57
NOD
C
Ripk2
TPM
PWK
SPRET
D
Cd36
basal
24 h IL-4
TPM
NOD
PWK
E
low basal
equal basal
high basal
40.3%
30.3%
16.4%
13.0%
65.1%
18.6%
10.7%
5.5%
62.7%
22.6%
10.5%
4.2%
motif mutation
0
1
2
>=3
F
PU.1 motif
EGR2 motif
motif score difference (induced - non-induced)
low basal
equal basal
high basal
G
C/EBPβ ChIP-seq
basal
24 h IL-4
low basal
p=6.6e-80
p=2.3e-120
equal basal
p=0.0006
p=7.8e-27
high basal
p=5.8e-111
p=5.3e-34
log2 tags
non-induced
induced
H
EGR2 ChIP-seq
low basal
p=1.8e-46
p=3.9e-75
equal basal
p=6.8e-17
p=1.0e-29
high basal
p=2.1e-5
p=7.1e-8
PPARγ ChIP-seq
low basal
p=8.5e-39
p=9.0e-96
equal basal
p=0.33
p=4.9e-41
high basal
p=6.8e-56
p=0.005
I
5-mers ranked within top 10% by deep neural networks
Matched with LDTF motif
Matched with SDTF motif
induced enhancer vs. random bg. odds ratio
odds ratio
IL-4 all enhancer vs. random bg.
induced enhancer vs. non-induced enhancer odds ratio
odds ratio
IL-4 all enhancer vs. random bg.

Chapter 4. Natural genetic variation affecting transcription factor spacing at regulatory regions is generally tolerated

4.1 Abstract

Regulation of gene expression requires the combinatorial binding of sequence-specific transcription factors (TFs) at promoters and enhancers. Prior studies showed that alterations in the spacing between TF binding sites can influence promoter and enhancer activity. However, the relative importance of TF spacing alterations resulting from naturally occurring insertions and deletions (InDels) has not been systematically analyzed. To address this question, we first characterized the genome-wide spacing relationships of 75 TFs in K562 cells as determined by ChIP-sequencing. We found a dominant pattern of a relaxed range of spacing between collaborative factors, including forty-five factors exclusively exhibiting relaxed spacing with their binding partners. Next, we exploited millions of InDels provided by genetically diverse mouse strains and human individuals to investigate the effects of altered spacing on TF binding and local histone acetylation. Spacing alterations resulting from naturally occurring InDels are generally tolerated in comparison to genetic variants directly affecting TF binding sites. A remarkable range of tolerance was further established for PU.1 and C/EBPβ, which exhibit relaxed spacing, by introducing synthetic spacing alterations ranging from 5-bp increase to >30-bp decrease using CRISPR/Cas9 mutagenesis. These findings provide implications for understanding mechanisms underlying enhancer selection and for interpretation of non-coding genetic variation.

4.2 Introduction

Genome-wide association studies (GWASs) have identified thousands of genetic variants associated with diseases and other traits (MacArthur et al., 2017; Visscher et al., 2017). Single

nucleotide polymorphisms (SNPs) and short insertions and deletions (InDels) represent common forms of these variants. The majority of GWAS variants fall at non-protein-coding regions of the genome, implicating their effects on gene regulation (Farh et al., 2015; Ward & Kellis, 2012). Gene expression is regulated by transcription factors (TFs) in cell-type-specific manner. TFs bind to short, degenerate sequences at promoters and enhancers, often referred to as TF binding motifs. Active promoters and enhancers are selected by combinations of sequence-specific TFs that bind in an inter-dependent manner to closely spaced motifs. SNPs and InDels can create or disrupt TF binding motifs and are a well-established mechanism for altering gene expression and biological function (Behera et al., 2018; Deplancke et al., 2016; Grossman et al., 2017; Heinz et al., 2013). InDels can additionally change spacing between motifs, but it remains unknown the extent to which altered spacing are relevant for interpreting natural genetic variation in human population or between animal species.

Previous studies reported two major categories of motif spacing between inter-dependent TFs (Slattery et al., 2014). One category requires specific spacing, or "constrained" spacing. These are mainly TFs that form ternary complexes recognizing composite binding sites, exemplified by GATA, Ets and E-box in mouse hematopoietic cells (Ng et al., 2014), MyoD and other cell-type-specific factors in muscle cells (Nandi et al., 2013), Sox2 and Oct4 in embryonic stem cells (Rodda et al., 2005), and 315 interactive TF pairs displaying cooperative binding based on CAP-SELEX studies (Jolma et al., 2015). Similar constrained spacing was also found between independent motifs at the interferon-β enhanceosome for the optimal binding and function of interacting TFs (Panne, 2008). In comparison to constrained spacing, another major category of motif spacing allows TFs to interact over a relatively broad range (e.g., 100-200 bp), which we call "relaxed" spacing. This type of spacing relationship is frequently observed in

collaborative TFs that do not target promoters or enhancers as a cooperative unit (Heinz et al., 2010; Jiang & Singh, 2014; Slattery et al., 2014).

Substantial evidence showed that the two different categories of spacing requirement can experience a divergent level of impact from genetic variation. Reporter assays examining synthetic alterations of motif spacing revealed examples of TFs that require constrained spacing and have high sensitivity of transcription factor binding (Ng et al., 2014; Panne, 2008) and gene expression (Farley et al., 2015) on spacing. On the contrary, flexibility in motif spacing has been demonstrated using parallel reporter assays in Drosophila (Menoret et al., 2013) and HepG2 cells (Smith et al., 2013). However, these studies did not distinguish the impact of altered spacing on transcription factor binding or subsequent recruitment of co-activators required for gene activation. Moreover, it remains unknown the extent to which these findings are relevant to interpret spacing alterations resulting from naturally occurring genetic variation.

To investigate the effects of altered spacing on TF binding and function, we first characterized the genome-wide binding patterns of seventy-five TFs based on their binding sites determined by chromatin immuno-precipitation sequencing (ChIP-seq). We developed a computational framework that assigned each spacing relationship to "constrained" and "relaxed" category and associated spacings to the naturally occurring InDels observed in human population to study the selective constraints of different spacing relationships. As specific case studies, we leveraged natural genetic variation from five strains of mice and numerous human samples to study the effect size of spacing alterations on TF binding activity and local histone acetylation. We find that InDels altering TF spacing have selective constraints similar to motif mutations when they occur between TF pairs with a constrained spacing relationship but are generally less constrained and well tolerated when they occur between TF pairs exhibiting relaxed spacing

relationships. Finally, we established remarkable tolerance in spacing for PU.1 and C/EBPβ by introducing a wide range of InDels between their respective binding sites at representative endogenous genomic loci using CRISPR/Cas9 mutagenesis.

4.3 Results

4.3.1 Most transcription factor pairs bind with a relaxed spacing relationship

We downloaded and processed over 70 TF ChIP-seq data from ENCODE data portal for K562 cells (Davis et al., 2018). After obtaining reproducible TF binding sites based on replicates, we first used the position weight matrix (PWM) of corresponding TF from JASPAR database (Fornes et al., 2020) to scan through the sequence of every binding site and identified the locations of high-affinity motifs (Fig. 4.1A, Supplementary Fig. 4.1). We then merged the binding sites of every pair of TFs and computed the edge-to-edge motif spacing for all the co-binding sites. Spacings of the co-binding sites were eventually aggregated in a density plot showing the distribution of motif spacing within +/- 100 bp. To categorize a spacing relationship, we used permutation tests on the averaged gradients to test for specific spacing constraints and used Kolmogorov–Smirnov test (KS test) to test for a relaxed spacing relationship against random distribution.

We applied this computational framework to all possible TF pairs in K562 cells. Overall, more TF pairs follow relaxed spacing relationships in comparison to constrained spacing relationships (Fig. 4.1B). Among the TF pairs with constrained spacing relationships, we saw examples binding very close to each other like GATA1 and TAL1 (Fig. 4.1C), which is consistent with the frequently observed composite motif of GATA and E-box (Ng et al., 2014). here are also TF pairs, exemplified by EGR1 and JUND, that bind relatively further away from

each other but still require some specific spacing (Fig. 4.1C). Previous studies demonstrated interactions between EGR1 and AP-1 factors (Levkovitz & Baraban, 2002; Nakashima et al., 2003), but the underlying mechanism for such constrained spacing at 29 bp needs to be further investigated. In addition, both constrained and relaxed spacing relationships are usually invariable for different motif orientations (Fig. 4.1C), consistent with previous findings (Lis & Walther, 2016). The same TF pairs, however, can have similar or different spacing relationships in different cell types (Supplementary Fig. 4.2). By dissecting each TF's binding sites based on their spacing relationships with co-binding TFs, we found that many TFs interact only in a relaxed spacing relationship, and some can interact in two distinct relationships depending on co-binding TFs (Fig. 4.1D). Very few TFs interact with only constrained spacing, some of which might show relaxed spacing relationships by expanding the current set of TFs.

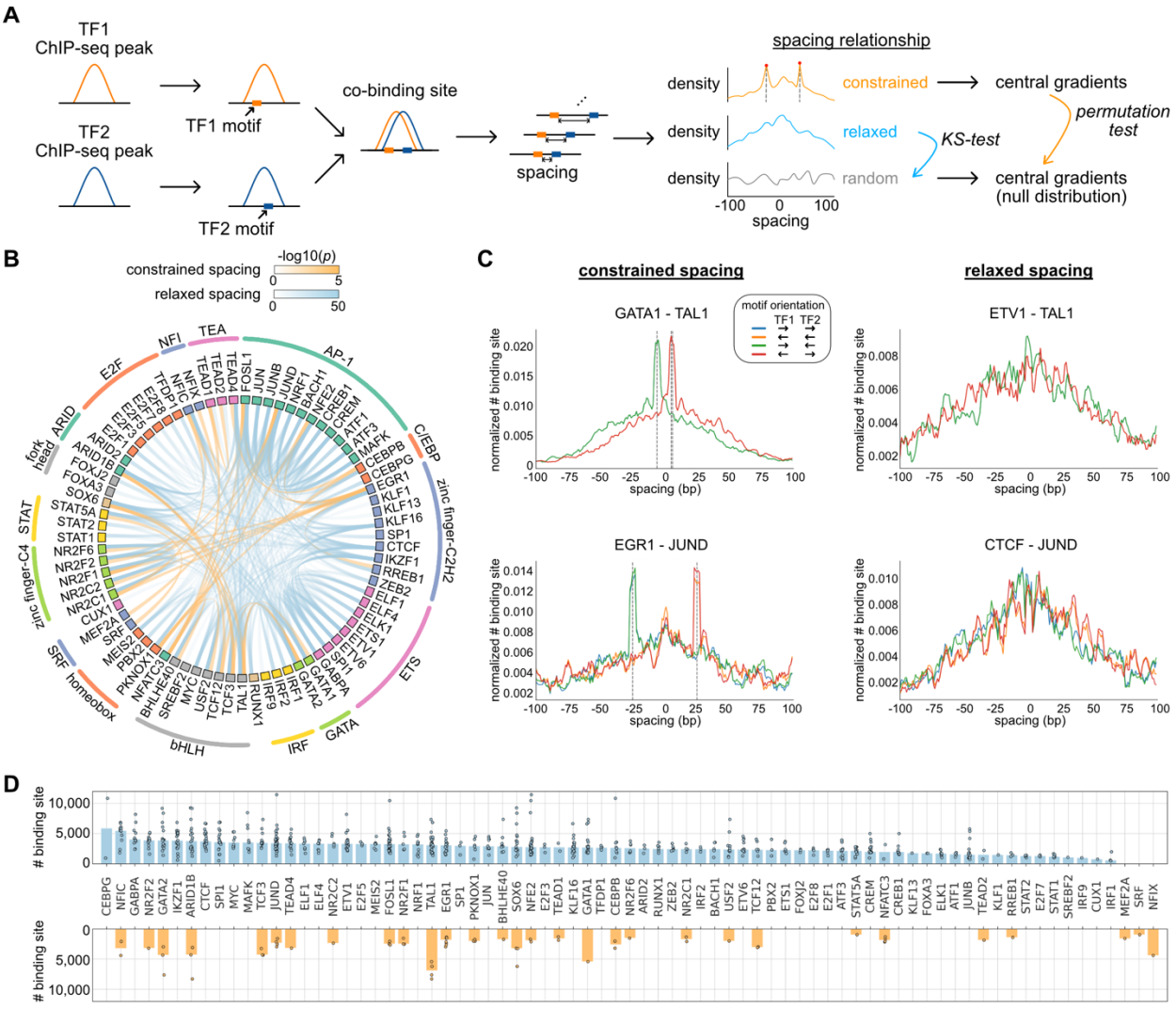


Figure 4.1 Characterization of spacing relationships for transcription factor pairs. (**A**) Schematic of data analysis pipeline for characterizing the spacing relationships based on TF ChIP-seq data. (**B**) Circos plot summarizing spacing relationships for TFs in K562 cells. Orange and blue bands represent significant constrained and relaxed spacing relationships, respectively. TFs are grouped and colored by the same family. (**C**) Examples of TF pairs with constrained spacing relationships or relaxed spacing relationships. Since TAL1 motif is completely palindromic, the motif orientation is only differentiated by its co-binding partners. (**D**) Dissection of TF binding sites based on spacing relationships. Each dot represents the co-binding peak number of the corresponding TF and one other TF with certain spacing relationship. Bar heights indicate means among all the TFs with the same spacing relationship.

4.3.2 Spacings between transcription factors with a relaxed spacing relationship are under less selective constraint

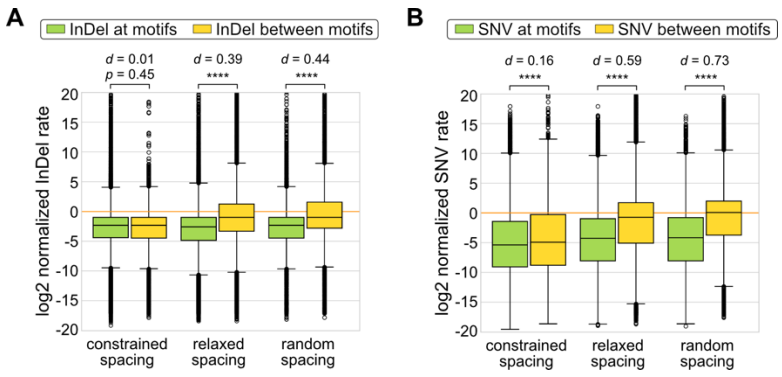


Figure 4.2 Comparison of selective constraints for different spacing relationships. (A) Aggregated normalized InDel rates for TF pairs of the same spacing relationship. InDel rates at motifs and those between motifs are compared within each spacing relationship using two-sample t-tests. ****p<0.0001. Cohen's d was displayed to indicate effect size. (B) Aggregated normalized SNV rates for the same set of TF pairs as used for calculation of InDel rates.

After a global view of the TF spacing relationships, we studied whether these relationships associate with different levels of sensitivity to spacing alterations. Here, we leveraged more than 60 million InDels from gnomAD data (Karczewski et al., 2020), which were based on over 75,000 genomes from unrelated individuals. We first compared the InDel rates at the binding sites of TF pairs representative of the different spacing relationships. The InDel rate is calculated as the total allele count of InDels per base pair occurring at motifs or between motifs divided by that occurring at background regions. For TF pairs with a constrained spacing relationship, we saw that the InDel rates at motifs are similar to those between motifs, while TF pairs with a relaxed or random spacing relationship have significantly lower InDel rates at motifs than those between motifs (Fig. 4.2A). Since common variants are associated with less deleteriousness and rare variants with more deleteriousness (Lek et al., 2016), our data suggest a weak effect of InDels that alter spacing of TFs with relaxed spacing relationships. In addition, the InDel rates at motifs are generally lower than those at background regions (log2 rate < 0),

consistent with the likely damaging effects of motif mutations. It also implicates that InDels between motifs of TFs with constrained spacing could be just as damaging as those at motifs.

As a comparative study, we applied the same approach to single nucleotide variants (SNVs) from gnomAD data (Karczewski et al., 2020) and calculated the SNV rates at the same TF co-binding sites as previously used for calculation of InDel rates (Fig. 4.2B). Relaxed and random spacing relationships showed significantly more SNVs between motifs compared to those at motifs. On the contrary, TFs with constrained spacings had much weaker difference in the SNV rate. Even though the aggregated results showed generally more SNVs occurring between motifs than those at motifs (Fig. 4.2B), we found TF pairs with either direction of change, exemplified by GATA1-TAL1 for more SNVs at motifs and FOSL1-NFATC3 for more SNVs between motifs. These results indicate that sequences between motifs of constrained TFs are generally under strong selective constraints, suggesting a comparable deleterious effect of changes at motifs and between motifs. All these findings regarding InDel rate and SNV rate suggest a weak selective constraint at sequences between motifs of TFs with a relaxed spacing relationship and potentially a small effect on TF binding by spacing alterations.

4.3.3 Spacing alterations by natural genetic variation of mouse strains are generally tolerated for transcription factor binding and promoter and enhancer function

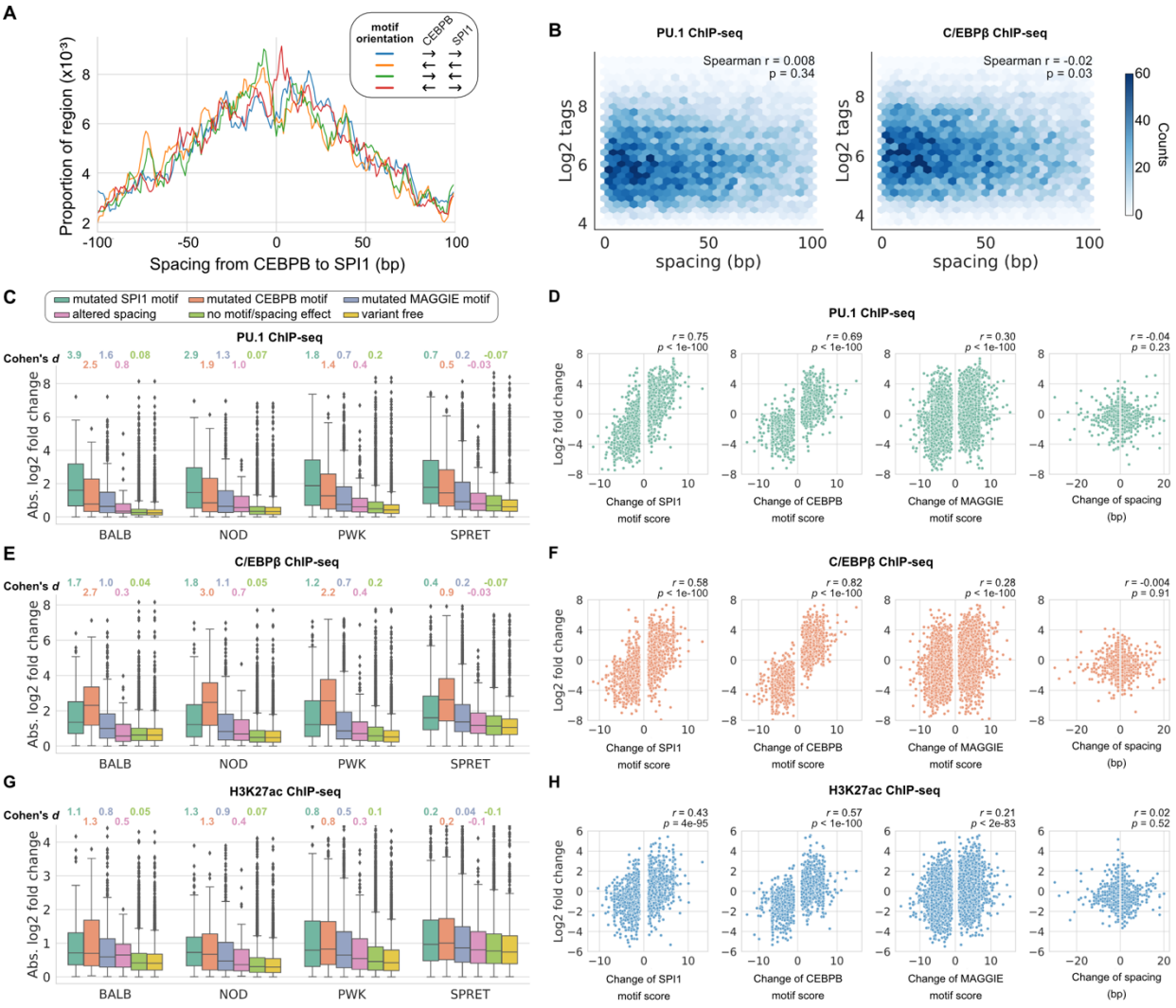


Figure 4.3 Effects of spacing alterations resulting from natural genetic variation across mouse strains. (**A**) Spacing distributions of PU.1 and C/EBPβ motif at co-binding sites. (**B**) Density plots showing the relationship between TF binding activity and motif spacing for the co-binding sites. Log2 ChIP-seq tags were calculated within 300 bp to quantify the binding activity of PU.1 and C/EBPβ. The color gradients represent the number of sites. Spearman correlation coefficients together with p-values are displayed to show the level of correlation. (**C**, **E**, **G**) Absolute log2 fold changes of ChIP-seq tags between C57 and another strain for (**C**) PU.1 binding, (**E**) C/EBPβ binding, or (**G**) H3K27ac level. Boxplot shows the median and quartiles of every distribution with its Cohen's d effect size displayed on top comparing against variant-free regions. (**D**, **F**, **H**) Correlations between change of motif spacing or motif score and change of (**D**) PU.1 binding, (**F**) C/EBPβ binding, or (**H**) H3K27ac level. Pearson correlation coefficients together with p-values are displayed to show the level of correlation.

To investigate the potential of using motif spacing alterations to interpret natural genetic variation, we leveraged more than 50 million SNPs and 5 million InDels from five genetically

diverse mouse strains and the ChIP-seq data of key TFs and histone modifications for the macrophages in each one of the five strains (Link, Duttke, et al., 2018). The five mouse strains include C57BL/6J (C57), BALB/cJ (BALB), NOD/ShiLtJ (NOD), PWK/PhJ (PWK), and SPRET/EiJ (SPRET). We first characterized the spacing relationship between the macrophage lineage-determining TFs (LDTFs), PU.1 and C/EBPβ, which have been found to bind in a collaborative manner at regulatory regions of macrophage-specific genes (Heinz et al., 2010). Based on our computational framework (Fig. 4.3A), these two TFs follow a relaxed spacing relationship regardless of their motif orientations (Fig. 4.3A; KS p-value < 1e-6). Moreover, both PU.1 and C/EBPβ binding activities quantified by the ChIP-seq tags were not correlated with the motif spacing, suggesting no direct association between spacing and TF binding (Fig. 4.3B).

We then conducted independent comparisons between C57 and one of the other four strains to investigate the effects of spacing alterations caused by natural genetic variation. We first identified the co-binding sites of PU.1 and C/EBPβ for each strain and then, for each pairwise analysis, pooled the co-binding sites of C57 and the compared strain to obtain the testing set of regions. Based on the impacts of genetic variants with regard to motif affinity and motif spacing, we categorized the testing regions into the following non-overlapping groups: 1) mutated PU.1 (i.e., SPI1) motif, 2) mutated C/EBPβ (i.e., CEBPB) motif, 3) mutated other functional motifs (i.e., MAGGIE motif), 4) altered spacing, 5) no motif affinity/spacing effect, and 6) variant free. Functional motifs were identified from PU.1 and C/EBPβ binding sites respectively using MAGGIE (Shen et al., 2020), which is a computational tool that can prioritize motifs whose affinity changes are associated with TF binding changes based on the ChIP-seq tags across different mouse strains (Supplementary Fig. 4.3). The effect of genetic variation was quantified by the log2 fold difference of ChIP-seq tags between strains (Fig. 4.3C). All the four

independent comparisons showed that PU.1 binding is most strongly affected by PU.1 motif mutation (average Cohen's d=2.3), followed by C/EBPβ motif mutation (average Cohen's d=1.6) and other functional motif mutation (average Cohen's d=0.95). Spacing alterations have much smaller effect size than any of these motif mutations (average Cohen's d=0.54), but still a larger effect than variants affecting neither motif affinity nor spacing (average Cohen's d=0.07). Despite the moderate effect size of spacing alterations, we found such effect was independent from the size and direction of InDels (Fig. 4.3D), suggesting the effects of InDels in this category might not be directly resulted from the spacing alterations but from other reasons. On the contrary, changes of motif affinity are strongly correlated with changes of PU.1 ChIP-seq tags (Fig. 4.3D). The effects of motif mutation and spacing alteration did not vary by the initial spacing between PU.1 and C/EBPβ motifs (Supplementary Fig. 4.4). The similar relationships were found in C/EBPβ binding, except that C/EBPβ motif mutation had the largest effect size and strongest correlation with changes in C/EBPβ binding activity (Fig. 4.3E, F, Supplementary Fig. 4.4).

To investigate whether the effects of altered spacing on PU.1 and C/EBPβ binding can be generalized to hierarchical interactions with signal-dependent transcription factors, we leveraged the ChIP-seq data of PU.1, the NF-kB subunit p65, and an AP-1 factor cJun for macrophages treated with the TLR4-specific ligand Kdo2 lipid A (KLA) in the same five strains of mice (Link, Duttke, et al., 2018). Upon macrophage activation with KLA, p65 enters the nucleus and primarily binds to poised enhancer elements that are selected by LDTFs including PU.1 and AP-1 factors (Heinz et al., 2015). We observed a relaxed spacing relationship between PU.1 and p65 and also between cJun and p65 (Supplementary Fig. 4.5). In addition, InDels altering motif

spacing had a much smaller effect size on TF binding than motif mutations (Supplementary Fig. 4.6), consistent with our finding from PU.1 and C/EBPβ.

Although alterations in motif spacing had generally weak effects at the level of DNA binding, it remained possible that changes in motif spacing could influence subsequent steps in enhancer and promoter activation. To examine this, we extended our analysis to local acetylation of histone H3 lysine 27 (H3K27ac), which is a histone modification that is highly correlated with enhancer and promoter function (Creyghton et al., 2010). We leveraged the H3K27ac ChIP-seq data of untreated macrophages in the five strains of mice (Link, Duttke, et al., 2018) and calculated the log fold changes of H3K27ac level within the extended 1000-bp regions of the PU.1 and C/EBPβ co-binding sites. Similar to what was observed for TF binding, altered spacing demonstrated weaker effects on histone acetylation than motif mutations (Fig. 4.3G, Supplementary Fig. 4.2), which is supported by the high consistency between change of TF binding and change of histone acetylation (Supplementary Fig. 4.7). The relative tolerance of spacing alteration was further reflected by a weak correlation between the change of acetylation level and the size of InDels, in comparison to a much stronger correlation with changes in motif affinity (Fig. 4.3H).

4.3.4 Human quantitative trait loci are depleted of variants changing motif spacing

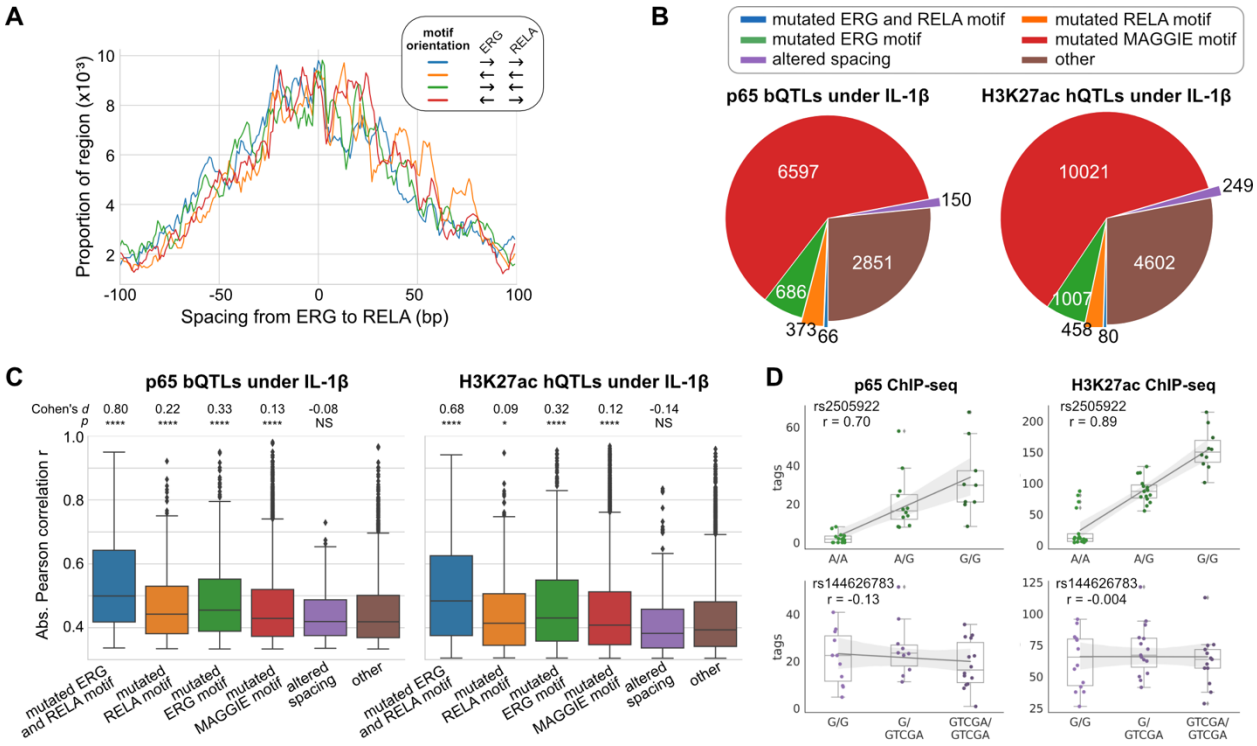


Figure 4.4 Effects of chromatin QTLs in human endothelial cells. (A) Spacing distributions of ERG and RELA motif at co-binding sites. (B) Classification of chromatin QTLs based on the effects on motif and spacing. (C) Absolute correlation coefficients of different QTLs. Cohen's d effect sizes together with p-values comparing against the "other" group are displayed on top. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. (D) Example QTLs for large effect size due to ERG motif mutation (top) and trivial effect due to spacing alteration (bottom).

To study the effects of spacing alteration in human, we leveraged the ChIP-seq data of ERG, p65, and H3K27ac in endothelial cells from many individuals (Stolze et al., 2020). ERG is a predominant ETS factor as well as an LDTF in endothelial cells that selects poised enhancers where p65 binds in a hierarchical manner upon interleukin one beta (IL-1β) stimulation (Hogan et al., 2017). ERG and p65 follow a relaxed spacing relationship according to our computational framework (Fig. 4.4A). Based on 22 ERG samples in untreated endothelial cells and 35 p65 samples in IL-1β-treated endothelial cells, we identified 2,669 TF binding quantitative trait loci (bQTLs) for ERG and 10,723 bQTLs for p65 (Supplementary Fig. 4.8). By overlaying 42 H3K27ac samples of IL-1β-treated endothelial cells at a pooled set of p65 binding sites and

another 42 H3K27ac samples of untreated endothelial cells at a pooled set of ERG binding sites, we identified 7,693 and 16,419 histone modification QTLs (hQTLs) for untreated and IL-1β-treated cells, respectively. We further classified bQTLs and hQTLs based on their impacts on motif affinity and spacing: 1) mutated both ERG and p65 (i.e., RELA) motif, 2) mutated ERG motif only, 3) mutated p65 motif only, 4) mutated other functional motifs identified by MAGGIE (Shen et al., 2020), 5) altered spacing, 6) none of the above. To find functional motifs, we fed MAGGIE with 100-bp sequences around QTLs before and after swapping alleles at the center (Supplementary Fig. 4.9). As a result, only a small portion of bQTLs and hQTLs directly mutates an ERG or RELA motif (Fig. 4.4B, Supplementary Fig. 4.10) even though such motif mutations are still enriched in QTLs compared to non-QTLs (Fisher's exact $p < 1e\text{-}4$). On the contrary, InDels that alter motif spacing are significantly depleted in QTLs (Fisher's exact $p < 1e\text{-}9$). The majority of QTLs have an impact on other functional motifs, implicating the complexity of TF interactions. Roughly a quarter of the QTLs affect neither motif affinity nor motif spacing, which can be explained by the high correlation of non-functional variants with functional variants due to the linkage disequilibrium.

We further compared the effect sizes of different categories of QTLs. Despite being the minority group among all QTLs, variants that mutate both ERG and RELA motif have the strongest effects on both p65 binding and histone acetylation in IL-1β-treated endothelial cells (Fig. 4.4C). In comparison, ERG binding and the basal level of histone acetylation are mainly affected by ERG motif mutations in untreated endothelial cells, while p65 motif mutations have trivial effects, consistent with the hierarchical interaction of p65 only upon IL-1β stimulation (Supplementary Fig. 4.11). In both states of endothelial cells, spacing alterations have the least effect size among all the categories and are not significantly different from likely non-functional

variants in the “other” group. The examples showed a variant being both a p65 bQTL and a hQTL due to its impact on an ERG motif, and also another variant associated with no change in p65 binding and H3K27ac despite its impact on increasing the motif spacing between ERG and p65 by 4 bp (Fig. 4.4D).

4.3.5 Transcription factor binding tolerates synthetic spacing alterations

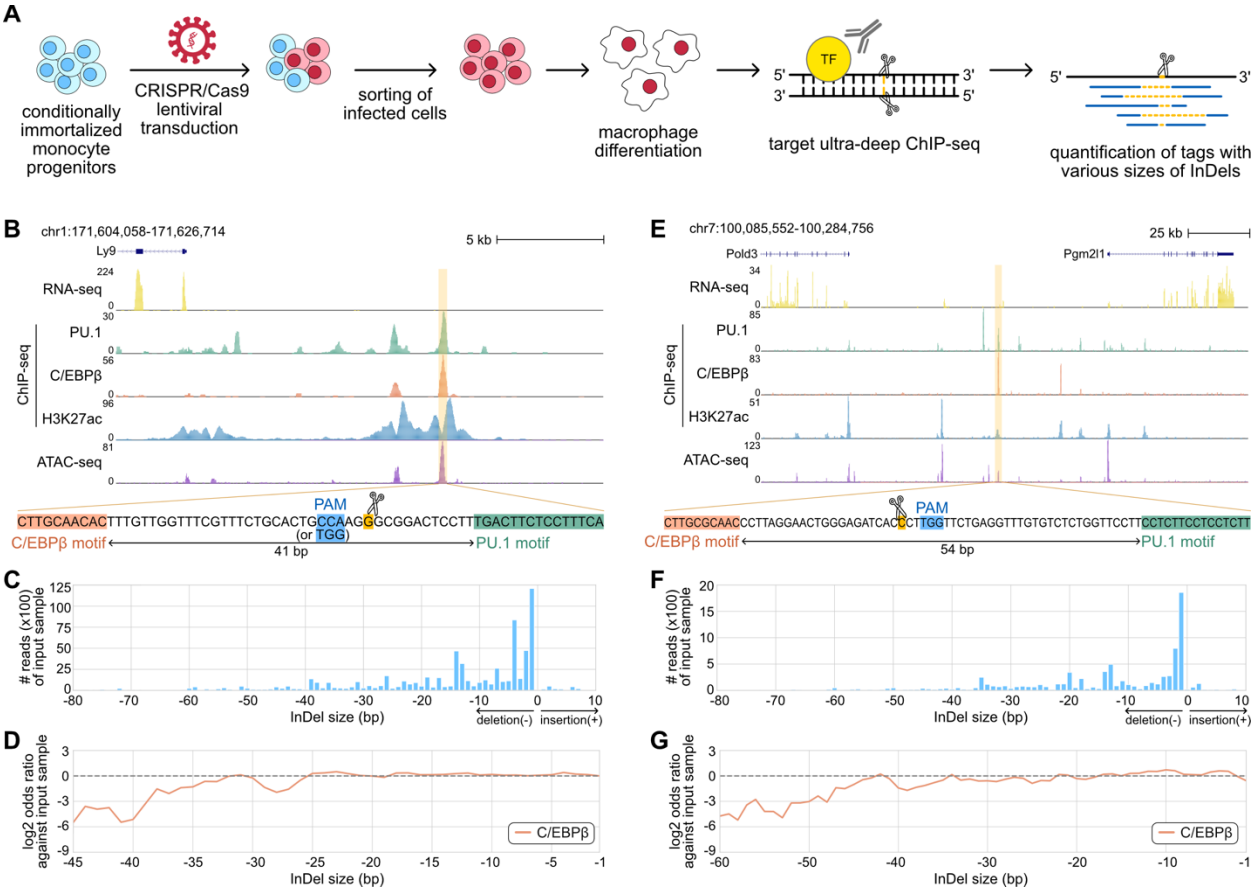


Figure 4.5 Effects of variable sizes of synthetic spacing alterations. (A) Schematic for generating and analyzing synthetic spacing alterations. (B, E) Experimental co-binding site of PU.1 and C/EBPβ. The sequencing data were based on ER-Hoxb8 cells. The target enhancer region is highlighted, and its DNA sequence from PU.1 motif to C/EBPβ motif is shown. (C, F) The distribution of valid read counts from the input sample based on the InDel sizes of the reads. Negative InDel size means deletion, while positive size means insertion. (D, G) Log2 odds ratios by comparing TF ChIP-seq reads and input sample reads at certain InDel sizes. Y=0 line indicates where TF binding has an expected amount of activity.

After seeing generally small regulatory effects of spacing alterations based on naturally occurring InDels, we tested the robustness and the extent of such tolerance using experimentally synthetic InDels. We tested on the LDTFs PU.1 and C/EBPβ of mouse macrophages using a

combination of CRISPR, deep target sequencing, and bioinformatic techniques (Fig. 4.5A). We started with ER-Hoxb8 cells, which is an immortalized macrophage progenitor cell line, and transduced the cells with the CRISPR/Cas9 system that targeted certain co-binding sites of PU.1 and C/EBPβ at the sequences between motifs. The Cas9 nuclease was supposed to cut at specific positions and generate various sizes of InDels in different cells. After sorting the successfully infected ER-Hoxb8 cells and differentiating them into macrophages, we then conducted ChIP-seq experiments plus very deep sequencing to amplify the target regions with the aim of capturing signals coming from different cells, which have different InDels. Lastly, the reads were mapped to the target regions by allowing various sizes of gaps at the cut sites and were quantified by comparing to the input DNA samples.

Here, we showed two testing regions, one with supportive evidence from naturally occurring InDels of mouse strains (Fig. 4.5B-D) and one without (Fig. 4.5E-G). For the region with supportive evidence, the highlighted *Ly9* enhancer has a 5-bp insertion between PU.1 and C/EBPβ motifs in BALB, NOD, and PWK, and shows unaffected binding of PU.1 and C/EBPβ in the BMDMs of these strains. This region also has strong binding of both PU.1 and C/EBPβ in ER-Hoxb8 cells and strong signals of H3K27ac and chromatin accessibility by ATAC-seq, indicating a potential regulatory function of this region (Fig. 4.5B). The PU.1 and C/EBPβ motif at this region are 41 bp apart. Based on the bioinformatic analysis of the ultra-deep sequencing reads from the input DNA sample, we saw the CRISPR/Cas9 system mostly generated deletions, more likely leading to shorter deletions (Fig. 4.5C). By calculating the odds ratio between C/EBPβ ChIP-seq reads with certain deletion and input sample reads with the same size of deletion, we produced the effect size of deletion on C/EBPβ binding in the function of deletion size (Fig. 4.5D). Overall, deletion ranging from 1 to 35 bp did not have much effect on TF

binding, indicated by a log2 odds ratio close to 0. On the contrary, deletions greater than this range resulted in a decrease in TF binding activity, likely due to the elimination of at least one of the two motifs. The similar results were found at the other region without supportive evidence from naturally occurring InDels (Fig. 4.5E-G). This region is located near genes *Pold3* and *Pgm2l1* and has strong signals of TF binding and regulatory activity (Fig. 4.5E). The distribution of input sample reads shows a higher chance of seeing shorter deletions from the Cas9 (Fig. 4.5F). The TF binding activity was not generally affected by deletions less than 45 bp (Fig. 4.5G). Beyond this range, the deletions likely crossed over motifs and diminished TF binding activity.

4.4 Discussion

By classifying the genome-wide spacing relationships of 75 co-binding TFs as "constrained" or "relaxed", we revealed that relaxed spacing relationships were the dominant pattern of interaction for majority of these factors. Among these factors, approximately half could also participate in constrained spacing relationships with specific TF partners. We confirmed TF pairs known to exhibit constrained relationships (e.g., GATA1-TAL1) and identified previously unreported constrained relationships for additional pairs, including EGR1 and JUND. Overall, this finding of a subset of constrained TF interactions on a genome wide level is consistent with the locus-specific examples provided by functional and structural studies of the interferon-β enhanceosome (Panne, 2008) and in vivo studies of synthetically modified enhancer elements in Ciona (Farley et al., 2015). Each of these examples represents genomic regulatory elements in which key TF motifs are tightly spaced in their native contexts (i.e., 0-9 bp between motifs). Direct protein-protein interactions are observed between bound TFs at the

interferon-β enhanceosome, analogous to interactions defined for cooperative TFs that form ternary complexes (Morgunova & Taipale, 2017). In the present studies, InDels between TF pairs exhibiting constrained spacings were under large selective constraints that were comparable to InDels at motifs, suggesting a deleterious effect of these spacing alterations on TF binding. However, the spacing analyses in this study did not directly consider the possible overlap or lack of spacing between TF binding sequences. Thus, we are not able to clearly distinguish effects of spacing alterations from effects of InDels on motifs at sites of tightly spaced composite motifs.

The observation that most TF pairs exhibited relaxed spacing relationships has intriguing implications for the mechanisms by which functional enhancers and promoters are selected from chromatinized DNA. In contrast to ternary complexes of TFs that cooperatively bind to composite elements as a unit, relaxed spacing relationships appear to not require specific protein-protein interactions between TFs for collaborative binding at most genomic locations. Although pioneering TFs necessary for selection of cell-specific enhancers have been reported to recognize their motifs within the context of nucleosomal DNA, the basis for collaborative binding interactions between TFs with relaxed spacings remains poorly understood.

While the current studies relying on natural genetic variation and mutagenesis experiments concluded clear tolerance of spacing alterations between motifs of TFs with relaxed spacings, the extent to which this set of binding sites is representative of all regulatory elements is unclear. For example, we observed outliers in which significant differences in TF binding between mouse strains were associated with InDels occurring between motifs. However, the proportion of outliers was generally similar to that observed at genomic regions lacking such InDels, and such strain differences may be driven by distal effects of genetic variation on interacting enhancer or promoter regions (Link, Duttke, et al., 2018). The remarkable tolerance

of synthetic InDels at two independent endogenous genomic locations between PU.1 and C/EBPβ binding sites strongly support the generality of relaxed binding interactions for these two proteins. Intriguingly, while the densities of C/EBP motifs increase with decreasing distance to PU.1 motifs over a 100 bp range (Fig. 4.3A), deletions from 1 to >30 bp in the context of PU.1-C/EBPβ pairs 41 or 54 bp apart did not result in improved binding. Instead, relatively constant binding was observed with progressive deletions bringing two motifs close together until the deletions started to cause mutations in one or both motifs. This is consistent with the lack of correlation between DNA binding strengths and distances between these factors (Figure 4.3B). A limitation of these studies is that very few insertions were obtained, preventing conclusions as to the extent to which increases in spacing are tolerated.

In concert, the present studies provide a basis for estimation of the potential phenotypic consequences of naturally occurring InDels in non-coding regions of the genome. In most cases, InDels between motifs for TFs that have relaxed binding relationships are unlikely to alter TF binding and function. In contrast, InDels between motifs for TFs that have constrained binding relationships have the potential to result in biological consequences. Application of these findings to the interpretation of non-coding InDels that are associated with disease risk will require knowledge of the relevant cell type in which the InDel exerts its phenotypic effect and the types of TF interactions driving the selection and function of the affected regulatory elements.

4.5 Methods

4.5.1 Sequencing data processing

We downloaded two replicates for each TF ChIP-seq data from ENCODE data portal (Davis et al., 2018). The mouse macrophage data and the human endothelial cell data were downloaded from the GEO database with accession number GSE109965 (Link, Duttke, et al., 2018) and GSE139377 (Stolze et al., 2020), respectively. We mapped the ChIP-seq reads using Bowtie2 v2.3.5.1 with default parameters (Langmead & Salzberg, 2012). All the human data downloaded from ENCODE were mapped to the hg38 genome. Data from C57BL/6J mice were mapped to the mm10 genome. Data from other mouse strains and endothelial cell data from different individuals were mapped to their respective genomes built by MMARGE v1.0 (Link, Romanoski, et al., 2018). More details are described below.

Based on the mapped ChIP-seq data, we then called TF binding sites or peaks using HOMER v4.9.1 (Heinz et al., 2010). For data with replicates including ENCODE data and mouse data, we first called unfiltered 200-bp peaks using HOMER "findPeaks" function using parameters "-style factor -L 0 -C 0 -fdr 0.9 -size 200" and then ran IDR v2.0.3 with default parameters (Li et al., 2011) to obtain reproducible peaks. For data without replicates including human endothelial cell data, we called peaks using HOMER "findPeaks" with the default setting and parameters "-style factor -size 200".

Activity of TF binding was quantified by the ChIP-seq tag counts within 300-bp around peak centers and normalized by library size using HOMER "annotatePeaks.pl" script with parameters "-norm 1e7 -size -150,150". Activity of promoter and enhancer was quantified by normalized H3K27ac ChIP-seq tags within 1000-bp regions around TF peak centers using parameters "-norm 1e7 -size -500,500".

4.5.2 Motif identification

Based on DNA sequences of the TF binding sites, we calculated motif scores by the dot products between position weight matrices (PWMs) from the JASPAR database (Fornes et al., 2020) and sequence vectors using Biopython package (Cock et al., 2009). The PWMs were trimmed to obtain only the core motifs from the first position where information content greater than 0.3 to the last position of information content greater than 0.3 (Ng et al., 2014). The valid motifs were identified by a motif score passing a false positive rate 0.1% and a location within 50 bp close to the peak center. The motif spacing is computed as the edge-to-edge distance between two motifs at TF co-binding sites. If there are multiple valid motifs for one or both TFs, we computed the spacing between all possible combinations of valid motifs.

4.5.3 Characterization of different motif spacing relationships

To test for the constrained spacing relationship between any two TFs, we developed a method to identify "spikes" in the spacing distribution. We first counted the TF pair distances at single-base-pair resolution ranging from -100 bp to +100 bp. Next, we computed the slope at each position using the following formula:

$$S_i = \frac{\Delta_{i,i-1} + \Delta_{i,i+1}}{2}, i \in [-99,99]$$

$$\Delta_{i,i-1} = N_i - N_{i-1}$$

S_i is the average of single-step forward and backward slope at position i. N_i represents the number of TF pair at position i, and Δ is the difference in the number of TF pairs between two locations. We conducted permutation tests to compare each S_i to a simulated null distribution to determine a p-value based on the percentile rank. P-value smaller than 6.25e-05 is called significant (familywise error rate=0.05/200/4), indicating a spike is found among motif spacing

between the testing TF pair. The null distribution was generated by 1,000 iterations of 1,000 random spacing between 0 and 100 bp.

To test for the relaxed spacing relationship, we used Kolmogorov–Smirnov (KS) test to compare a spacing distribution to the random distribution. We randomly sampled integers between -100 and 100 to match the same size of the testing spacing distribution and then tested the spacing distribution against the distribution of the random integers to obtain a p-value. We repeated the above process 100 times and computed an average p-value for the final report. Significance threshold was adjusted by familywise error rate: $0.05/4/\mathrm{C}_{\#\mathrm{TF}}^{2}$.

4.5.4 Calculation of variant rate

We obtained SNPs and InDels from gnomAD v3.1 (Karczewski et al., 2020) and overlapped the gnomAD variants with TF co-binding sites, specifically with two TF motifs (denoted as motif1 and motif2) and their intermediate sequences (denoted as mid). To account for local background, we also overlapped variants with 100 bp upstream and downstream region outside of the motifs (denoted as surrounding). For each co-binding site, the normalized variant rate for motif1, motif2 and mid region is calculated as:

$$F_i = \frac{C_i/S_i + \text{pseudo}}{C_{surrounding}/S_{surrounding} + \text{pseudo}}, i \in \{motif1, motif2, mid\}$$

where C_i denotes variant count in certain region (motif1, motif2, or mid), S_i represent size of the region. $C_{surrounding}$ represents variant count within the surrounding regions outside of motifs, and $S_{surrounding}$ represents the size of surrounding regions, set as 200 bp in this study. To avoid division by zero, a small pseudo-rate 0.005 is added.

4.5.5 Genetic variation processing and genome building

Genetic variation of the five mouse strains was obtained from (Keane et al., 2011), and that of the human individuals from which endothelial cell data were generated was derived from

(Stolze et al., 2020). We used MMARGE v1.0 with default variant filters (Link, Romanoski, et al., 2018) to build separate genomes for each mouse strain and human individual. The sequencing data from different samples were respectively mapped to the corresponding genomes and were then shifted to a common reference genome using MMARGE "shift" function to facilitate comparison at homologous regions. The reference genome is mm10 for mouse strains and hg19 for human individuals.

4.5.6 Identification of QTLs

ChIP-seq tags from human endothelial cells were counted surrounding every genetic variant within +/- 150 bp for ERG and p65 or +/- 1000 bp for H3K27ac. Tag counts of the same genotype from different individuals were aggregated and regressed across genotypes 0/0, 0/1 (or 1/0), and 1/1. Variants associated with no more than 16 tags in any individual were removed. QTLs were identified as the variants associated with a linear regression p-value smaller than the familywise error rate corrected by the total number of variants within the peaks of each signal.

4.5.7 Motif mutation analysis

We used MAGGIE (Shen et al., 2020) to identify functional motifs for different TF binding. To prepare the inputs into MAGGIE based on the mouse strains data, we adapted a similar strategy as described in Shen et al., 2020. In brief, we conducted pairwise orthogonal comparisons of TF peaks between each possible pair of the five strains to find strain-differential peaks. We then extracted pairs of 200-bp sequences around the centers of the differential peaks from the genomes of two comparative strains, the ones with TF binding as positive sequences paired with those without TF binding as negative sequences. For the QTLs of human endothelial cells, MAGGIE can directly work with a VCF file of QTLs with effect size and effect direction indicated in a column of the file. We ran MAGGIE separately for each type of QTLs. The

significant hits passed FDR < 0.05 after the Benjamini–Hochberg controlling procedure, and the p-values output from MAGGIE were reported.

4.5.8 Categorization of genetic variation

We categorized genetic variation based on its impact on motif affinity and motif spacing. Motif mutations were defined by at least 1-bit difference in the motif score, which is equivalent to 2-fold difference in the binding likelihood. Mutations of other functional motifs identified by MAGGIE required that at least one of the functional motifs had motif mutations. InDels were first classified into motif mutation categories if eligible before being considered in motif spacing. Spacing alterations included InDels between target motifs, which did not have any mutations. Variants fitting neither motif mutation nor spacing alteration were gathered in a separate group as a control. Another control category during analysis of mouse strains data was defined by TF binding sites that have no genetic variation between strains.

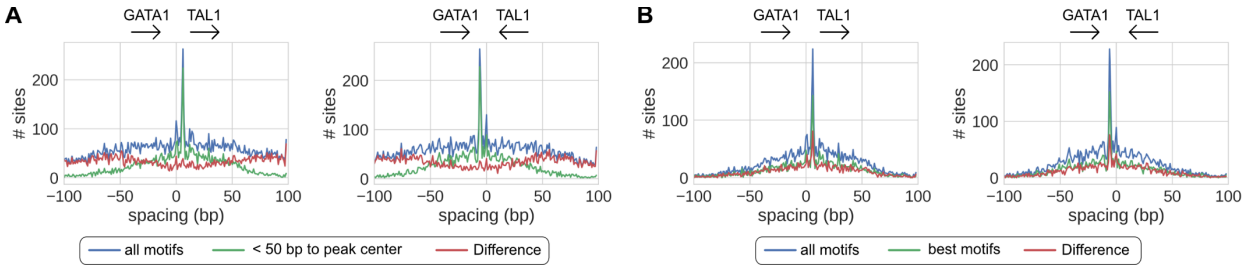
4.5.9 Statistical testing of effect size

Effect size of genetic variation was computed by the ratio of ChIP-seq tag counts at orthogonal sites between two comparative samples followed by log2 transformation. We conducted Mann-Whitney tests between “Variant free” and other categories to test significant differences in their distributions. We also obtained the Cohen’s d between the sampled variant-free set and the testing category as the effect size (Sullivan & Feinn, 2012).

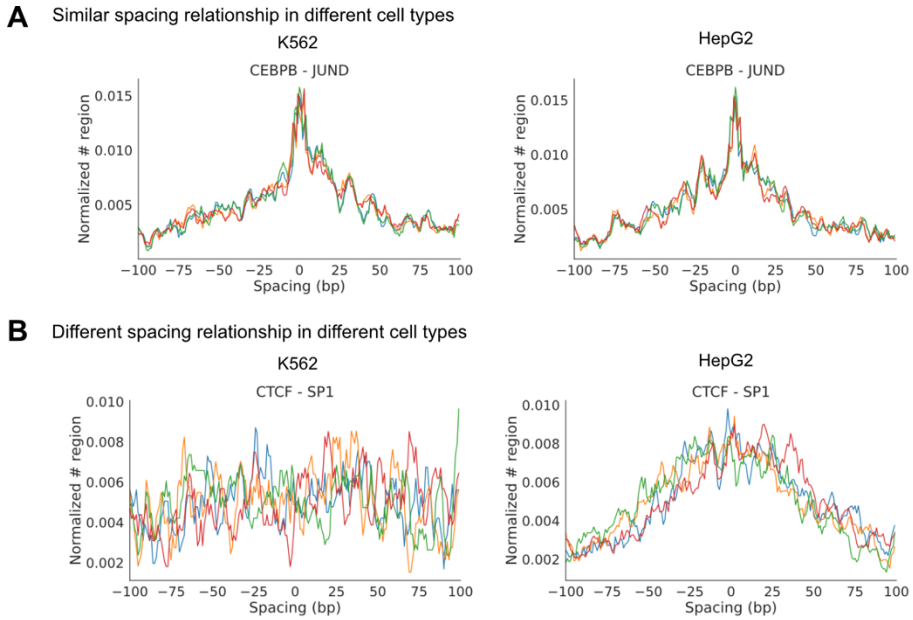
4.6 Acknowledgements

Chapter 4, in full, is currently being prepared for submission for publication of the material. Shen Z, Li RZ, Prohaska T, Hoeksema MA, Spann N & Glass CK. The dissertation author was the primary investigator and author of this material.

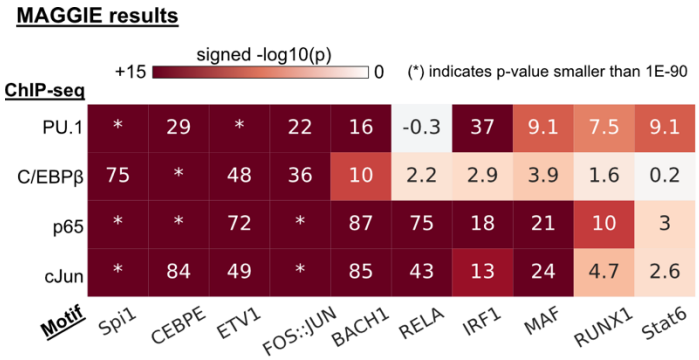
4.7 Supplementary figures



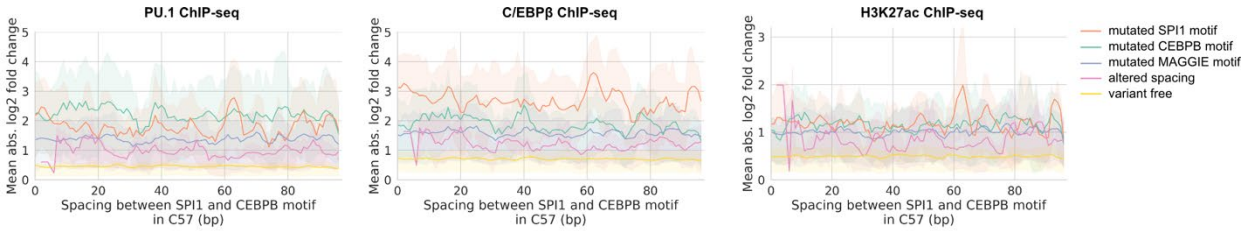
Supplementary Figure 4.1 Effects of different motif scanning criteria. (A) Motifs proximal to peak centers are potentially more confident than motifs distal from peak centers. (B) All motifs passing FPR < 0.001 are potentially as confident as the best motif of every peak.



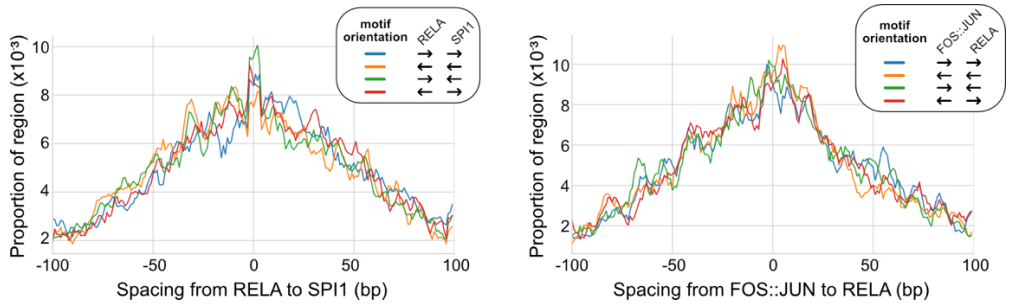
Supplementary Figure 4.2 Comparison of the spacing relationships of same TF pairs in different cell types. Example of TF pairs with (A) similar or (B) different spacing relationships in different cell lines.



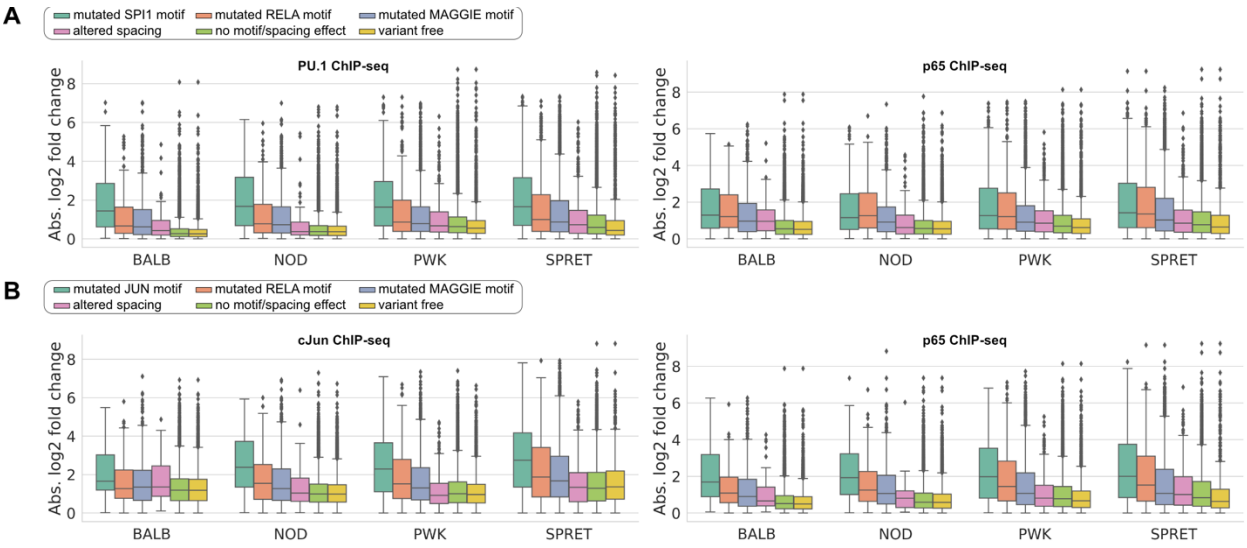
Supplementary Figure 4.3 Functional motifs identified by MAGGIE for different TF binding.



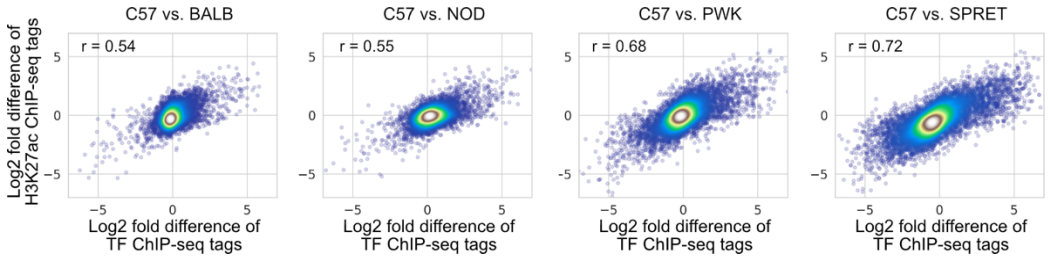
Supplementary Figure 4.4 Absolute log2 fold changes of ChIP-seq tags in relationship with the initial spacing between PU.1 and C/EBPβ motif.



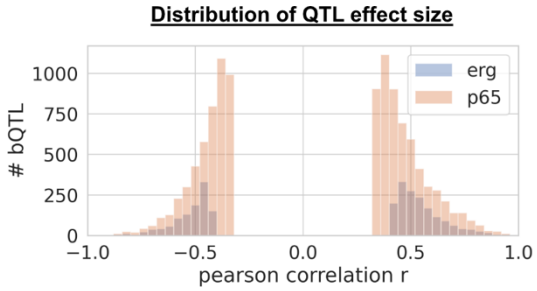
Supplementary Figure 4.5 Spacing distributions between LDTFs and SDTFs. Left: p65 and PU.1. Right: p65 and cJun.



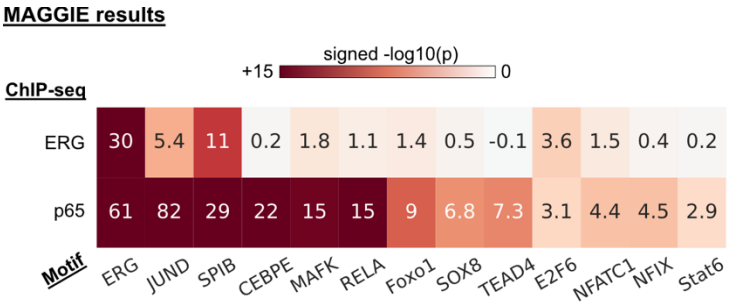
Supplementary Figure 4.6 Absolute log2 fold changes of ChIP-seq tags between C57 and another strain for LDTFs and SDTFs. (A) PU.1 and p65 binding at their co-binding sites and (B) cJun and p65 binding at their co-binding sites.



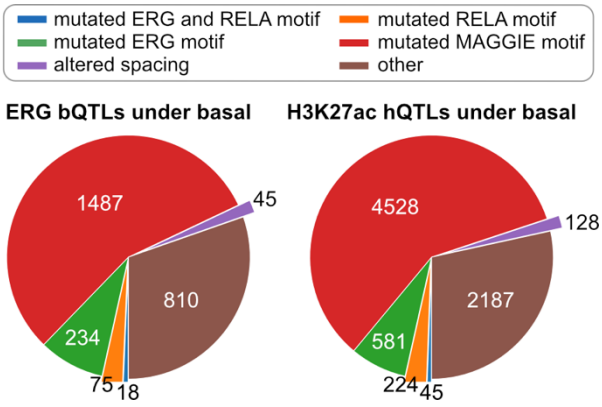
Supplementary Figure 4.7 Correlations between changes in TF binding activity and changes in the H3K27ac level.



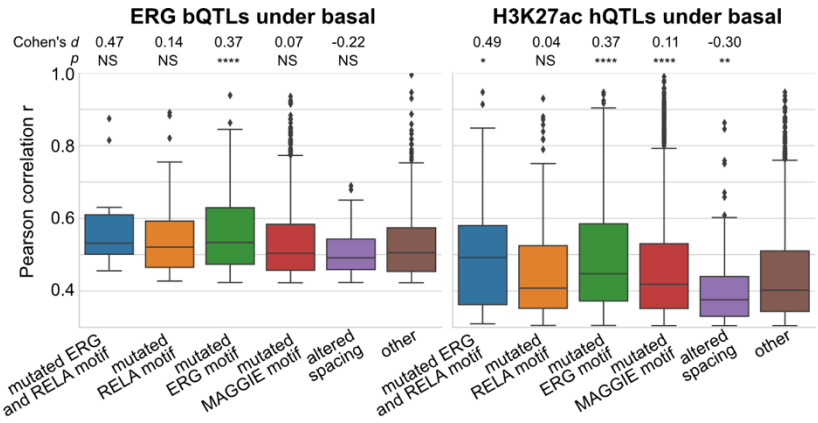
Supplementary Figure 4.8 Distribution of effect sizes of TF binding QTLs.



Supplementary Figure 4.9 Functional motifs identified by MAGGIE based on bQTLs.



Supplementary Figure 4.10 Classification of chromatin QTLs based on the effects on motif and spacing for basal condition.



Supplementary Figure 4.11 Absolute correlation coefficients of different QTLs for basal condition. Cohen's d effect sizes together with p-values comparing against the "other" group are displayed on top. $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$, $^{****}p<0.0001$.

Chapter 5. Conclusion

Here I investigated the effects of genetic variation on different aspects of transcriptional regulation, including TF binding and regulatory activity, and linked the regulatory effects to gene expression. TF plays as an important anchor in this work that connects DNA sequences with transcriptional regulation via TF binding motifs. Relevant to motifs, I studied the effects of genetic variation that alters motif or motif spacing. In Chapter 4, I summarized my findings that some TFs require constrained spacing since they form ternary complex and recognize composite motifs, but most collaborative TFs do not have such constraint and can tolerate spacing alterations in the scale of naturally occurring InDels and beyond. On the contrary, motif mutations have much stronger impacts regardless of whether the alterations occur at the motifs of bound TFs or collaborative TFs. These findings suggest the universal relevance of interpreting non-coding genetic variation in the context of motif mutations, while for a few TFs with constrained spacing relationships, spacing alterations may also be relevant.

To facilitate the interpretation of variants in the context of motif mutations, I described a new bioinformatic approach called MAGGIE in Chapter 2 that can identify functional motifs for TF binding and regulatory function based on motif mutations. Using this approach, I reproduced known important TFs for the regulatory function of lymphoblastoid cell lines and two different states of macrophages. The significant motifs from MAGGIE often overlap with those found by motif enrichment methods, providing additional support for them to be the targets of genetic variation that influences transcriptional regulation. In some cases, exemplified by the pro-inflammatory state of macrophages, MAGGIE showed exceptional ability to distinguish the functions of relatively similar motifs based on their associations with different functions. In Chapter 3, such capability of MAGGIE helped discover quantitative variations in motif affinity

underlying the divergent anti-inflammatory response of macrophages. It would be important to validate this finding using mutagenesis experiments and study this phenomenon in other systems in the future. Overall, this finding suggests the importance of considering motif as a continuous and quantitative variable instead of a binary variable with only the status of presence or absence.

Besides the consideration of quantitative motif mutations, this work also support that the effects of motif mutations should be extended to collaborative TFs. The functional motifs of collaborative TFs identified by MAGGIE based on TF binding sites (Fig. 2.3 and 3.6) and the strong impacts on TF binding from motif mutations at collaborative TFs' motifs (Fig. 4.2 and 4.3) all indicate the inter-dependence of different motifs. In Chapter 3, I leveraged deep learning techniques to model an entire non-coding sequence so that multiple motifs together with their relative locations could be preserved and incorporated into the model. This approach successfully prioritized non-coding variants that are associated with changes in regulatory function, showing a strong promise of using deep learning to identify functional non-coding variants. Many of these prioritized variants are located at known functional motifs, which adds more confidence to these variants being functional. As a future direction, it would be interesting to further classify the prioritized variants to find those that do not influence motifs. The mechanisms underlying these non-motif mutations might provide insights for novel aspects of variant interpretation.

Either motif-focused approaches or deep learning application in this study discussed the effects of genetic variation within at most hundreds of base pairs. The analysis of chromatin conformation data in Chapter 3 indicated the potential of extending variant interpretation from local effects to a longer range of effects on interactive regions. This work showed a significant correlation of regulatory activity between interactive enhancers. Many of these interactive

enhancers had genetic variation only on one side but were under a synchronized effect from the genetic variation. Variants prioritized by deep learning are significantly enriched at the interactive enhancers of those that do not have local variants. Further investigation is needed to validate the significance and mechanisms underlying these associations, but still, it implicates the possibilities of studying the effects of non-coding variants by taking the chromatin conformation into consideration. The insights provided by our findings could inspire an extension of current deep learning application to model long-range associations of regulatory elements and potentially improve the performance in predicting the effects of non-coding variants.

Building upon many great works studying the effects of genetic variation on transcriptional regulation, this work pushed the boundaries a bit further by applying state-of-the-art computational tools and conducting novel analyses that integrate different types of data. As much as I hope my work has demonstrated deep dissection of non-coding genetic variation from a couple of aspects, I hope it will provoke thinking on how we can study this problem by going wide to integrate a variety of data types and going deep into the current data with novel analysis approaches.

Bibliography

Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Aken, B., Akiyama, J. A., Jammal, O. Al, Amrhein, H., Anderson, S. M., Andrews, G. R., Antoshechkin, I., Ardlie, K. G., Armstrong, J., Astley, M., Banerjee, B., Barkal, A. A., Barnes, I. H. A., Barozzi, I., … Myers, R. M. (2020). Perspectives on ENCODE. *Nature*, *583*(7818), 693–698. https://doi.org/10.1038/s41586-020-2449-8

Arzate-Mejía, R. G., Recillas-Targa, F., & Corces, V. G. (2018). Developing in 3D: the role of CTCF in cell differentiation. *Development (Cambridge, England)*, *145*(6). https://doi.org/10.1242/dev.137729

Bakker, O. B., Aguirre-Gamboa, R., Sanna, S., Oosting, M., Smeekens, S. P., Jaeger, M., Zorro, M., Võsa, U., Withoff, S., Netea-Maier, R. T., Koenen, H. J. P. M., Joosten, I., Xavier, R. J., Franke, L., Joosten, L. A. B., Kumar, V., Wijmenga, C., Netea, M. G., & Li, Y. (2018). Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses. *Nature Immunology*, *19*(7), 776–786. https://doi.org/10.1038/s41590-018-0121-3

Barozzi, I., Simonatto, M., Bonifacio, S., Yang, L., Rohs, R., Ghisletti, S., & Natoli, G. (2014). Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA Features of Mammalian Enhancers. *Molecular Cell*, *54*(5), 844–857. https://doi.org/10.1016/j.molcel.2014.04.006

Behera, V., Evans, P., Face, C. J., Hamagami, N., Sankaranarayanan, L., Keller, C. A., Giardine, B., Tan, K., Hardison, R. C., Shi, J., & Blobel, G. A. (2018). Exploiting genetic variation to uncover rules of transcription factor binding and chromatin accessibility. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-03082-6

Boeva, V. (2016). Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in Eukaryotic cells. *Frontiers in Genetics*, *7*(FEB). https://doi.org/10.3389/fgene.2016.00024

Bonn, S., Zinzen, R. P., Girardot, C., Gustafson, E. H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczyåski, B., Riddell, A., & Furlong, E. E. M. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics*, *44*(2), 148–156. https://doi.org/10.1038/ng.1064

Brignall, R., Moody, A. T., Mathew, S., & Gaudet, S. (2019). Considering abundance, affinity, and binding site availability in the NF-κB target selection puzzle. *Frontiers in Immunology*, *10*(MAR), 1–14. https://doi.org/10.3389/fimmu.2019.00609

Cheng, C. S., Feldman, K. E., Lee, J., Verma, S., Huang, D. Bin, Huynh, K., Chang, M., Ponomarenko, J. V., Sun, S. C., Benedict, C. A., Ghosh, G., & Hoffmann, A. (2011). The specificity of innate immune responses is enforced by repression of interferon response elements by NF-κB p50. *Science Signaling*, *4*(161), 1–12. https://doi.org/10.1126/scisignal.2001501

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & De Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., & Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(50), 21931–21936. https://doi.org/10.1073/pnas.1016071107

Crocker, J., Abe, N., Rinaldi, L., McGregor, A. P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., Mann, R. S., & Stern, D. L. (2015). Low affinity binding site clusters confer HOX specificity and regulatory robustness. *Cell*, *160*(1–2), 191–203. https://doi.org/10.1016/j.cell.2014.11.041

Czimmerer, Z., Daniel, B., Horvath, A., Rückerl, D., Nagy, G., Kiss, M., Peloquin, M., Budai, M. M., Cuaranta-Monroy, I., Simandi, Z., Steiner, L., Nagy, B., Poliska, S., Banko, C., Bacso, Z., Schulman, I. G., Sauer, S., Deleuze, J. F., Allen, J. E., … Nagy, L. (2018). The Transcription Factor STAT6 Mediates Direct Repression of Inflammatory Enhancers and Limits Activation of Alternatively Polarized Macrophages. *Immunity*, *48*(1), 75-90.e6. https://doi.org/10.1016/j.immuni.2017.12.010

Daniel, B., Czimmerer, Z., Halasz, L., Boto, P., Kolostyak, Z., Poliska, S., Berger, W. K., Tzerpos, P., Nagy, G., Horvath, A., Hajas, G., Cseh, T., Nagy, A., Sauer, S., Francois-Deleuze, J., Szatmari, I., Bacsi, A., & Nagy, L. (2020). The transcription factor EGR2 is the molecular linchpin connecting STAT6 activation to the late, stable epigenomic program of alternative macrophage polarization. *Genes and Development*, *34*(21–22), 1474–1492. https://doi.org/10.1101/gad.343038.120

Daniel, B., Nagy, G., Czimmerer, Z., Horvath, A., Hammers, D. W., Cuaranta-Monroy, I., Poliska, S., Tzerpos, P., Kolostyak, Z., Hays, T. T., Patsalos, A., Houtman, R., Sauer, S., Francois-Deleuze, J., Rastinejad, F., Balint, B. L., Sweeney, H. L., & Nagy, L. (2018). The Nuclear Receptor PPARγ Controls Progressive Macrophage Polarization as a Ligand-Insensitive Epigenomic Ratchet of Transcriptional Memory. *Immunity*, *49*(4), 615-626.e6. https://doi.org/10.1016/j.immuni.2018.09.005

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y., & Cherry, J. M. (2018). The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Research*, *46*(D1), D794–D801. https://doi.org/10.1093/nar/gkx1081

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J. B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E., Stephens, M., Gilad, Y., & Pritchard, J. K. (2012). DNase-I sensitivity QTLs are a major determinant of human expression variation. *Nature*, *482*(7385), 390–394. https://doi.org/10.1038/nature10808

Deplancke, B., Alpern, D., & Gardeux, V. (2016). The Genetics of Transcription Factor DNA Binding Variation. *Cell*, *166*(3), 538–554. https://doi.org/10.1016/j.cell.2016.07.012

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Du, N., Kwon, H., Li, P., West, E. E., Oh, J., Liao, W., Yu, Z., Ren, M., & Leonard, W. J. (2014). EGR2 is critical for peripheral naïve T-cell differentiation and the T-cell response to influenza. *Proceedings of the National Academy of Sciences*, *111*(46), 16484 LP – 16489. https://doi.org/10.1073/pnas.1417215111

El Chartouni, C., Schwarzfischer, L., & Rehli, M. (2010). Interleukin-4 induced interferon regulatory factor (Irf) 4 participates in the regulation of alternative macrophage priming. *Immunobiology*, *215*(9–10), 821–825. https://doi.org/10.1016/j.imbio.2010.05.031

Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-019-0122-6

Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., & Knight, J. C. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, *343*(6175). https://doi.org/10.1126/science.1246949

Farh, K. K. H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., … Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, *518*(7539), 337–343. https://doi.org/10.1038/nature13835

Farley, E. K., Olson, K. M., Zhang, W., Brandt, A. J., Rokhsar, D. S., & Levine, M. S. (2015). Suboptimization of developmental enhancers. *Science*, *350*(6258), 325–328. https://doi.org/10.1126/science.aac6948

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. (2019). Diverse motif ensembles specify non-redundant DNA binding activities of AP-1 family members in macrophages. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-018-08236-0

Fornes, O., Castro-Mondragon, J. A., Khan, A., Van Der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W., & Mathelier, A. (2020). JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. *Nucleic Acids Research*, *48*(D1), D87–D92. https://doi.org/10.1093/nar/gkz1001

Gate, R. E., Cheng, C. S., Aiden, A. P., Siba, A., Tabaka, M., Lituiev, D., Machol, I., Gordon, M. G., Subramaniam, M., Shamim, M., Hougen, K. L., Wortman, I., Huang, S. C., Durand, N. C., Feng, T., De Jager, P. L., Chang, H. Y., Aiden, E. L., Benoist, C., … Regev, A. (2018). Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nature Genetics*, *50*(8), 1140–1150. https://doi.org/10.1038/s41588-018-0156-2

Gieseck, R. L., Wilson, M. S., & Wynn, T. A. (2018). Type 2 immunity in tissue repair and fibrosis. *Nature Reviews Immunology*, *18*(1), 62–76. https://doi.org/10.1038/nri.2017.90

Glass, C. K., & Natoli, G. (2016). Molecular control of activation and priming in macrophages. *Nature Immunology*, *17*(1), 26–33. https://doi.org/10.1038/ni.3306

Glimcher, L. H., & Singh, H. (1999). Transcription factors in lymphocyte development T and B cells get together. *Cell*, *96*(1), 13–23. https://doi.org/10.1016/S0092-8674(00)80955-1

Goenka, S., & Kaplan, M. H. (2011). Transcriptional regulation by STAT6. *Immunologic Research*, *50*(1), 87–96. https://doi.org/10.1007/s12026-011-8205-2

Gordon, S., & Martinez, F. O. (2010). Alternative activation of macrophages: Mechanism and functions. *Immunity*, *32*(5), 593–604. https://doi.org/10.1016/j.immuni.2010.05.007

Grossman, S. R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B. E., Mikkelsen, T. S., & Lander, E. S. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(7), E1291–E1300. https://doi.org/10.1073/pnas.1621150114

Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V., Martin, A. R., Greenside, P., Srivas, R., Phanstiel, D. H., Pekowska, A., Heidari, N., Euskirchen, G., Huber, W., Pritchard, J. K., Bustamante, C. D., Steinmetz, L. M., Kundaje, A., & Snyder, M. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, *162*(5), 1051–1065. https://doi.org/10.1016/j.cell.2015.07.048

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biology*, *8*(2), R24. https://doi.org/10.1186/gb-2007-8-2-r24

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, *38*(4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Heinz, S., Romanoski, C. E., Benner, C., Allison, K. A., Kaikkonen, M. U., Orozco, L. D., & Glass, C. K. (2013). Effect of natural genetic variation on enhancer selection and function. *Nature*, *503*, 487. https://doi.org/10.1038/nature12615

Heinz, S., Romanoski, C. E., Benner, C., & Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology*, *16*(3), 144–154.

https://doi.org/10.1038/nrm3949

Heinz, S., Texari, L., Hayes, M. G. B., Urbanowski, M., Chang, M. W., Givarkes, N., Rialdi, A., White, K. M., Albrecht, R. A., Pache, L., Marazzi, I., García-Sastre, A., Shaw, M. L., & Benner, C. (2018). Transcription Elongation Can Affect Genome 3D Structure. *Cell*, *174*(6), 1522-1536.e22. https://doi.org/10.1016/j.cell.2018.07.047

Hogan, N. T., Whalen, M. B., Stolze, L. K., Hadeli, N. K., Lam, M. T., Springstead, J. R., Glass, C. K., & Romanoski, C. E. (2017). Transcriptional networks specifying homeostatic and inflammatory programs of gene expression in human aortic endothelial cells. *ELife*, *6*(Cvd), 1–28. https://doi.org/10.7554/eLife.22536

Jayaram, N., Usvyat, D., & Martin, A. C. (2016). Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*, *17*(1), 1–12. https://doi.org/10.1186/s12859-016-1298-9

Ji, Z., He, L., Rotem, A., Janzer, A., Cheng, C. S., Regev, A., & Struhl, K. (2018). Genome-scale identification of transcription factors that mediate an inflammatory network during breast cellular transformation. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-04406-2

Jiang, P., & Singh, M. (2014). CCAT: Combinatorial Code Analysis Tool for transcriptional regulation. *Nucleic Acids Research*, *42*(5), 2833–2847. https://doi.org/10.1093/nar/gkt1302

Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., & Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, *527*(7578), 384–388. https://doi.org/10.1038/nature15518

Juric, I., Yu, M., Abnousi, A., Raviram, R., Fang, R., Zhao, Y., Zhang, Y., Qiu, Y., Yang, Y., Li, Y., Ren, B., & Hu, M. (2019). MAPS: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. *PLOS Computational Biology*, *15*(4), e1006982. https://doi.org/10.1371/journal.pcbi.1006982

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … Consortium, G. A. D. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N. A., Eskin, E., Nellåker, C., Whitley, H., Cleak, J., Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T. G., … Adams, D. J. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, *477*(7364), 289–294. https://doi.org/10.1038/nature10413

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., & Gerstein, M. (2016). Role

of non-coding sequence variants in cancer. *Nature Reviews Genetics*, *17*(2), 93–108. https://doi.org/10.1038/nrg.2015.17

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-arcelus, M., Panousis, N. I., Yurovsky, A., Lappalainen, T., Romano-palumbo, L., Planchon, A., Bielser, D., Bryois, J., Padioleau, I., Udin, G., Thurnheer, S., … Lis, J. T. (2013). Coordinated Effects of Sequence. *Science*, *342*(November), 744–747.

Kobayashi, E. H., Suzuki, T., Funayama, R., Nagashima, T., Hayashi, M., Sekine, H., Tanaka, N., Moriguchi, T., Motohashi, H., Nakayama, K., & Yamamoto, M. (2016). Nrf2 suppresses macrophage inflammatory response by blocking proinflammatory cytokine transcription. *Nature Communications*, *7*(May), 1–14. https://doi.org/10.1038/ncomms11624

Koch, C. M., Andrews, R. M., Flicek, P., Dillon, S. C., Karaöz, U., Clelland, G. K., Wilcox, S., Beare, D. M., Fowler, J. C., Couttet, P., James, K. D., Lefebvre, G. C., Bruce, A. W., Dovey, O. M., Ellis, P. D., Dhami, P., Langford, C. F., Weng, Z., Birney, E., … Dunham, I. (2007). The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Research*, *17*(6), 691–707. https://doi.org/10.1101/gr.5704207

Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J., & Mann, R. S. (2019). Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. *Annual Review of Cell and Developmental Biology*, *35*(1), 357–379. https://doi.org/10.1146/annurev-cellbio-100617-062719

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., Medvedeva, Y. A., Magana-Mora, A., Bajic, V. B., Papatsenko, D. A., Kolpakov, F. A., & Makeev, V. J. (2018). 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. https://doi.org/10.1093/nar/gkx1106

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., & Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, *172*(4), 650–665. https://doi.org/10.1016/j.cell.2018.01.029

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559. https://doi.org/10.1186/1471-2105-9-559

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357. https://doi.org/10.1038/nmeth.1923

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

Lek, M., Karczewski, K. J., Minikel, E. V, Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E.,

Berghout, J., … Consortium, E. A. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Levkovitz, Y., & Baraban, J. M. (2002). A Dominant Negative Egr Inhibitor Blocks Nerve Growth Factor-Induced Neurite Outgrowth by Suppressing c-Jun Activation: Role of an Egr/c-Jun Complex. *Journal of Neuroscience*, *22*(10), 3845–3854. https://doi.org/10.1523/jneurosci.22-10-03845.2002

Li, Q., Brown, J. B., Huang, H., & Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, *5*(3), 1752–1779. https://doi.org/10.1214/11-AOAS466

Liao, X., Sharma, N., Kapadia, F., Zhou, G., Lu, Y., Hong, H., Paruchuri, K., Mahabeleshwar, G. H., Dalmas, E., Venteclef, N., Flask, C. A., Kim, J., Doreian, B. W., Lu, K. Q., Kaestner, K. H., Hamik, A., Clément, K., & Jain, M. K. (2011). Krüppel-like factor 4 regulates macrophage polarization. *Journal of Clinical Investigation*, *121*(7), 2736–2749. https://doi.org/10.1172/JCI45444

Link, V. M., Duttke, S. H., Chun, H. B., Holtman, I. R., Westin, E., Hoeksema, M. A., Abe, Y., Skola, D., Romanoski, C. E., Tao, J., Fonseca, G. J., Troutman, T. D., Spann, N. J., Strid, T., Sakai, M., Yu, M., Hu, R., Fang, R., Metzler, D., … Glass, C. K. (2018). Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide Mechanisms that Control Transcription Factor Binding and Function. *Cell*, *173*(7), 1796-1809.e17. https://doi.org/10.1016/j.cell.2018.04.018

Link, V. M., Romanoski, C. E., Metzler, D., & Glass, C. K. (2018). MMARGE: Motif mutation analysis for regulatory genomic elements. *Nucleic Acids Research*, *46*(14), 7006–7021. https://doi.org/10.1093/nar/gky491

Lis, M., & Walther, D. (2016). The orientation of transcription factor binding site motifs in gene promoter regions: Does it matter? *BMC Genomics*, *17*(1), 1–21. https://doi.org/10.1186/s12864-016-2549-x

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., MayPendlington, Z., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., & Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, *45*(D1), D896–D901. https://doi.org/10.1093/nar/gkw1133

Machanick, P., & Bailey, T. L. (2011). MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics*, *27*(12), 1696–1697. https://doi.org/10.1093/bioinformatics/btr189

Martin, V., Zhao, J., Afek, A., Mielko, Z., & Gordân, R. (2019). QBiC-Pred: Quantitative predictions of transcription factor binding changes due to sequence variants. *Nucleic Acids*

Research, *47*(W1), W127–W135. https://doi.org/10.1093/nar/gkz363

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., & Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, *34*(Database issue), 108–110. https://doi.org/10.1093/nar/gkj143

Mcvicker, G., Geijn, B. Van De, Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., & Myrthil, M. (2013). Identification of Genetic Variants. *Science*, *342*(November), 747–749.

Menoret, D., Santolini, M., Fernandes, I., Spokony, R., Zanet, J., Gonzalez, I., Latapie, Y., Ferrer, P., Rouault, H., White, K. P., Besse, P., Hakim, V., Aerts, S., Payre, F., & Plaza, S. (2013). Genome-wide analyses of Shavenbaby target genes reveals distinct features of enhancer organization. *Genome Biology*, *14*(8), R86. https://doi.org/10.1186/gb-2013-14-8-r86

Mevel, R., Draper, J. E., Lie-A-Ling, M., Kouskoff, V., & Lacaud, G. (2019). RUNX transcription factors: Orchestrators of development. *Development (Cambridge)*, *146*, 1–19. https://doi.org/10.1242/dev.148296

Morgunova, E., & Taipale, J. (2017). Structural perspective of cooperative transcription factor binding. *Current Opinion in Structural Biology*, *47*, 1–8. https://doi.org/10.1016/j.sbi.2017.03.006

Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., & Chang, H. Y. (2016). HiChIP: Efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods*, *13*(11), 919–922. https://doi.org/10.1038/nmeth.3999

Nagel, D., Vincendeau, M., Eitelhuber, A. C., & Krappmann, D. (2014). Mechanisms and consequences of constitutive NF-κB activation in B-cell lymphoid malignancies. *Oncogene*, *33*(50), 5655–5665. https://doi.org/10.1038/onc.2013.565

Nakashima, A., Ota, A., & Sabban, E. L. (2003). Interactions between Egr1 and AP1 factors in regulation of tyrosine hydroxylase transcription. *Molecular Brain Research*, *112*(1–2), 61–69. https://doi.org/10.1016/S0169-328X(03)00047-0

Nandi, S., Blais, A., & Ioshikhes, I. (2013). Identification of cis-regulatory modules in promoters of human genes exploiting mutual positioning of transcription factors. *Nucleic Acids Research*, *41*(19), 8822–8841. https://doi.org/10.1093/nar/gkt578

Natoli, G., Saccani, S., Bosisio, D., & Marazzi, I. (2005). Interactions of NF-κB with chromatin: The art of being at the right place at the right time. *Nature Immunology*, *6*(5), 439–445. https://doi.org/10.1038/ni1196

Ng, F. S. L., Schütte, J., Ruau, D., Diamanti, E., Hannah, R., Kinston, S. J., & Göttgens, B. (2014). Constrained transcription factor spacing is prevalent and important for

transcriptional control of mouse blood cells. *Nucleic Acids Research*, *42*(22), 13513–13524. https://doi.org/10.1093/nar/gku1254

Nott, A., Holtman, I. R., Coufal, N. G., Schlachetzki, J. C. M., Yu, M., Hu, R., Han, C. Z., Pena, M., Xiao, J., Wu, Y., Keulen, Z., Pasillas, M. P., O'Connor, C., Nickl, C. K., Schafer, S. T., Shen, Z., Rissman, R. A., Brewer, J. B., Gosselin, D., … Glass, C. K. (2019). Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science*, *366*(6469), 1134 LP – 1139. https://doi.org/10.1126/science.aay0793

Ostuni, R., Piccolo, V., Barozzi, I., Polletti, S., Termanini, A., Bonifacio, S., Curina, A., Prosperini, E., Ghisletti, S., & Natoli, G. (2013). Latent enhancers activated by stimulation in differentiated cells. *Cell*, *152*(1–2), 157–171. https://doi.org/10.1016/j.cell.2012.12.018

Panne, D. (2008). The enhanceosome. *Current Opinion in Structural Biology*, *18*(2), 236–242. https://doi.org/10.1016/j.sbi.2007.12.002

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Reiter, F., Wienerroither, S., & Stark, A. (2017). Combinatorial function of transcription factors and cofactors. *Current Opinion in Genetics and Development*, *43*, 73–81. https://doi.org/10.1016/j.gde.2016.12.007

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-Throughput Sequencing Technologies. *Molecular Cell*, *58*(4), 586–597. https://doi.org/10.1016/j.molcel.2015.05.004

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47–e47. https://doi.org/10.1093/nar/gkv007

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–329. https://doi.org/10.1038/nature14248

Rodda, D. J., Chew, J. L., Lim, L. H., Loh, Y. H., Wang, B., Ng, H. H., & Robson, P. (2005). Transcriptional regulation of Nanog by OCT4 and SOX2. *Journal of Biological Chemistry*, *280*(26), 24731–24737. https://doi.org/10.1074/jbc.M502573200

Satoh, T., Takeuchi, O., Vandenbon, A., Yasuda, K., Tanaka, Y., Kumagai, Y., Miyake, T., Matsushita, K., Okazaki, T., Saitoh, T., Honma, K., Matsuyama, T., Yui, K., Tsujimura, T., Standley, D. M., Nakanishi, K., Nakai, K., & Akira, S. (2010). The Jmjd3-Irf4 axis regulates M2 macrophage polarization and host responses against helminth infection. *Nature Immunology*, *11*(10), 936–944. https://doi.org/10.1038/ni.1920

Scott, E. W., Simon, M. C., Anastasi, J., & Singh, H. (1994). Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science*, *265*(5178), 1573 LP – 1577. https://doi.org/10.1126/science.8079170

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference*, *Scipy*, 92–96. https://doi.org/10.25080/majora-92bf1922-011

Seidman, J. S., Troutman, T. D., Sakai, M., Gola, A., Spann, N. J., Bennett, H., Bruni, C. M., Ouyang, Z., Li, R. Z., Sun, X., Vu, B. C. T., Pasillas, M. P., Ego, K. M., Gosselin, D., Link, V. M., Chong, L. W., Evans, R. M., Thompson, B. M., McDonald, J. G., … Glass, C. K. (2020). Niche-Specific Reprogramming of Epigenetic Landscapes Drives Myeloid Cell Diversity in Nonalcoholic Steatohepatitis. *Immunity*, *52*(6), 1057-1074.e7. https://doi.org/10.1016/j.immuni.2020.04.001

Shen, Z., Hoeksema, M. A., Ouyang, Z., Benner, C., & Glass, C. K. (2020). MAGGIE: leveraging genetic variation to identify DNA sequence motifs mediating transcription factor binding and function. *Bioinformatics*, *36*(Supplement_1), i84–i92. https://doi.org/10.1093/bioinformatics/btaa476

Shi, W., Fornes, O., Mathelier, A., & Wasserman, W. W. (2016). Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Research*, *44*(21), 10106–10116. https://doi.org/10.1093/nar/gkw691

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). *Learning Important Features Through Propagating Activation Differences*. http://arxiv.org/abs/1704.02685

Siebert, M., & Söding, J. (2016). Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research*, *44*(13), 6055–6069. https://doi.org/10.1093/nar/gkw521

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R., & Rohs, R. (2014). Absence of a simple code: How transcription factors read the genome. *Trends in Biochemical Sciences*, *39*(9), 381–399. https://doi.org/10.1016/j.tibs.2014.07.002

Smith, R. P., Taher, L., Patwardhan, R. P., Kim, M. J., Inoue, F., Shendure, J., Ovcharenko, I., & Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics*, *45*(9), 1021–1028. https://doi.org/10.1038/ng.2713

Soccio, R. E., Chen, E. R., Rajapurkar, S. R., Safabakhsh, P., Marinis, J. M., Dispirito, J. R., Emmett, M. J., Briggs, E. R., Fang, B., Everett, L. J., Lim, H. W., Won, K. J., Steger, D. J., Wu, Y., Civelek, M., Voight, B. F., & Lazar, M. A. (2015). Genetic Variation Determines PPARγ Function and Anti-diabetic Drug Response in Vivo. *Cell*, *162*(1), 33–44. https://doi.org/10.1016/j.cell.2015.06.025

Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: From enhancer binding to developmental control. *Nature Reviews Genetics*, *13*(9), 613–626.

https://doi.org/10.1038/nrg3207

Spivakov, M., Akhtar, J., Kheradpour, P., Beal, K., Girardot, C., Koscielny, G., Herrero, J., Kellis, M., Furlong, E. E. M., & Birney, E. (2012). Analysis of variation at transcription factor binding sites in Drosophila and humans. *Genome Biology*, *13*(9), R49. https://doi.org/10.1186/gb-2012-13-9-r49

Stolze, L. K., Conklin, A. C., Whalen, M. B., López Rodríguez, M., Õunap, K., Selvarajan, I., Toropainen, A., Örd, T., Li, J., Eshghi, A., Solomon, A. E., Fang, Y., Kaikkonen, M. U., & Romanoski, C. E. (2020). Systems Genetics in Human Endothelial Cells Identifies Non-coding Variants Modifying Enhancers, Expression, and Complex Disease Traits. *American Journal of Human Genetics*, *106*(6), 748–763. https://doi.org/10.1016/j.ajhg.2020.04.008

Stormo, G. D. (2000). DNA binding sites: Representation and discovery. *Bioinformatics*, *16*(1), 16–23. https://doi.org/10.1093/bioinformatics/16.1.16

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough . *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/jgme-d-12-00156.1

van der Veeken, J., Zhong, Y., Sharma, R., Mazutis, L., Dao, P., Pe'er, D., Leslie, C. S., & Rudensky, A. Y. (2019). Natural Genetic Variation Reveals Key Features of Epigenetic and Transcriptional Memory in Virus-Specific CD8 T Cells. *Immunity*, *50*(5), 1202-1217.e7. https://doi.org/10.1016/j.immuni.2019.03.031

Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., Rynes, E., Reynolds, A., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Kaul, R., Meuleman, W., & Stamatoyannopoulos, J. A. (2020). Global reference mapping of human transcription factor footprints. *Nature*, *583*(7818), 729–736. https://doi.org/10.1038/s41586-020-2528-x

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Ward, L. D., & Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology*, *30*(11), 1095–1106. https://doi.org/10.1038/nbt.2422

Waszak, S. M., Delaneau, O., Gschwind, A. R., Kilpinen, H., Raghav, S. K., Witwicki, R. M., Orioli, A., Wiederkehr, M., Panousis, N. I., Yurovsky, A., Romano-Palumbo, L., Planchon, A., Bielser, D., Padioleau, I., Udin, G., Thurnheer, S., Hacker, D., Hernandez, N., Reymond, A., … Dermitzakis, E. T. (2015). Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell*, *162*(5), 1039–1050. https://doi.org/10.1016/j.cell.2015.08.001

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., & Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, *153*(2), 307–319.

https://doi.org/10.1016/j.cell.2013.03.035

Widenmaier, S. B., Snyder, N. A., Nguyen, T. B., Arduini, A., Lee, G. Y., Arruda, A. P., Saksi, J., Bartelt, A., & Hotamisligil, G. S. (2017). NRF1 Is an ER Membrane Sensor that Is Central to Cholesterol Homeostasis. *Cell*, *171*(5), 1094.e15-1109.e15. https://doi.org/10.1016/j.cell.2017.10.003

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9), R137. https://doi.org/10.1186/gb-2008-9-9-r137

Zheng, A., Lamkin, M., Zhao, H., Wu, C., Su, H., & Gymrek, M. (2021). Deep neural networks identify sequence context features predictive of transcription factor binding. *Nature Machine Intelligence*, *3*(2), 172–180. https://doi.org/10.1038/s42256-020-00282-y

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, *12*(10), 931–934. https://doi.org/10.1038/nmeth.3547

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., Benner, C., & Chanda, S. K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*, *10*(1), 1523. https://doi.org/10.1038/s41467-019-09234-6