UNIVERSITY OF CALIFORNIA SAN DIEGO

Deep learning predicts the impact of non-coding genetic variants in human traits and diseases

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

in

Computer Science

by

An Zheng

Committee in charge:

Professor Melissa Gymrek, Chair
Professor Hao Su, Co-Chair
Professor Christopher Benner
Professor Christopher Glass
Professor Niema Moshiri

2022

The Dissertation of An Zheng is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

I dedicate this dissertation to my parents, Feng Zheng and Yan An, for their unconditional love and endless support.

I also dedicate this dissertation to all the medical professionals who have been battling against the coronavirus disease 2019 pandemic. I admire your bravery and fortitude, and we remain ever thankful.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AD	Alzheimer's disease
ATAC-seq	Assay for transposase-accessible chromatin using sequencing
ASD	Autism spectrum disorder
auROC	Area under the receiver operating characteristic curve
auPRC	Area under the precision-recall curve
BMDM	Bone marrow-derived macrophage
BD	Bipolar disorder
BP	Base pair
ChIPs	ChIP-seq simulator
ChIP-seq	Chromatin immunoprecipitation sequencing
CNN	Convolutional neural network
CR	Chromatin regulators
DNA	Deoxyribonucleic acid
ENCODE	Encyclopedia of DNA elements
HM	Histone modification
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	Insertions and/or deletions in DNA
KLA	Kdo2 lipid A
LD	Linkage disequilibrium
LDTF	Lineage-determining transcription factor
MAGGIE	Motif alteration genome-wide to globally investigate elements
MDD	Major depressive disorder
NOD	NOD/ShiLtJ
PWK	PWK/Ph
PCR	Polymerase chain reaction
PWM	Position weight matrix
QTL	Quantitative trait locus
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RNN	Recurrent neural network
SDTF	Signal-dependent transcription factor
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SPRET	SPRET/EiJ
TF	Transcription factors

ACKNOWLEDGEMENTS

I would first like to warmly thank my advisor, Professor Melissa Gymrek, for her invaluable mentorship and support. I am grateful for the opportunity to work with her and together push the boundary of biomedical sciences. I am greatly honored to be her student, and I will always think fondly of my time as a student in her lab. Her guidance, trust, and openness to new ideas have been indispensable to me during my research training and have pushed me to become a better researcher and independent thinker.

I would like to thank my dissertation committee for their guidance and support. My committee co-chair, Professor Hao Su, has been a fantastic collaborator and gave me tremendous help on my deep learning projects. I am also grateful that he invited me to attend his lab meetings, which allowed me to keep track of state-of-the-art machine learning methods. Professor Christopher Glass has been working closely with me on my project for brain diseases and providing invaluable insights into the result interpretations. Professor Christopher Benner has been extremely helpful in my paper editing and has provided insightful comments on how to improve my papers. Professor Niema Moshiri was my mentor in the first bioinformatics algorithm course I took at UCSD and has been giving me research advice since then.

I am blessed to have incredible research collaborators and colleagues. I would like to thank all the members of the Gymrek lab for the helpful discussions and collaborations; specifically, Michael Lamkin for his contributions to both the AgentBind and ChIPs projects, Hanqing Zhao and Cynthia Wu for their contributions to the AgentBind project. I would like to thank my collaborators, Professor Alon Goren for his mentorship and guidance in the ChIPs project, Dr. Zeyang Shen for his contribution to my project for brain disease, Yutong Qiu, and Kevin Ren for their contributions to the ChIPs project.

I would like to thank my family for their unconditional love and support, for sparking my interest in natural sciences in my childhood, for always being there for me during my ups and downs, and for always encouraging me to chase my dream. I certainly could not have achieved this feat without them.

Chapter 1, in part, is taken from "Deep neural networks identify sequence context features predictive of transcription factor binding" in Nature machine intelligence by Zheng, A., Lamkin, M., Zhao, H., Wu, C., Su, H., and Gymrek, M; and "AgentBind: Profiling Context-specific Determinants of Transcription Factor Binding Affinity" presented in the ICML 2019 Workshop on Computational Biology by Zheng, A., Lamkin, M., Wu, C., Su, H., and Gymrek, M. The dissertation author was a primary researcher and author of both papers.

Chapter 2, in full, is taken from "Deep neural networks identify sequence context features predictive of transcription factor binding" in Nature machine intelligence by Zheng, A., Lamkin, M., Zhao, H., Wu, C., Su, H., and Gymrek, M. The dissertation author was a primary investigator and author of this paper.

Chapter 3, in full, is currently being prepared for submission for publication of the material by Zheng, A., Shen, Z., Glass, C., and Gymrek, M. The dissertation author was a primary researcher and author of this material.

Chapter 4, in part, includes portions of material as it appears in "Mechanisms underlying divergent responses of genetically distinct macrophages to IL-4" in Science advances by Hoeksema, M. A., Shen, Z., Holtman, I. R., Zheng, A., Spann, N. J., Cobo, I., Gymrek, M., and Glass, C. K. The dissertation author contributed to the deep learning modeling and computational analyses in this work.

Chapter 5, in full, is a reprint of the material as it appears in "A flexible ChIP-sequencing simulation toolkit" in BMC bioinformatics by Zheng, A., Lamkin, M., Qiu, Y., Ren, K., Goren, A., & Gymrek, M. The dissertation author was a primary investigator and author of this paper.

VITA

2015 Bachelor of Engineering in Electronics Information Engineering, Huazhong University of Science and Technology, China.

2017 Master of Science in Computer Science, University of California San Diego, USA.

2022 Doctor of Philosophy in Computer Science, University of California San Diego, USA.

PUBLICATIONS

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.

Zheng, A., Lamkin, M., Qiu, Y., Ren, K., Goren, A., & Gymrek, M. (2021). A flexible ChIP-sequencing simulation toolkit. *BMC bioinformatics*, 22(1), 1-10.

Hoeksema, M. A., Shen, Z., Holtman, I. R., **Zheng, A.**, Spann, N. J., Cobo, I., Gymrek, M., & Glass, C. K. (2021). Mechanisms underlying divergent responses of genetically distinct macrophages to IL-4. *Science advances*, 7(25), eabf9808.

Zheng, A., Lamkin, M., Wu, C., Su, H., & Gymrek, M. (2021). AgentBind: Profiling Context-specific Determinants of Transcription Factor Binding Affinity. *ICML 2019 Workshop on Computational Biology*.

Zheng, A. (2017). *Use solid k-mers in minHash-based genome distance estimation*. University of California, San Diego. (Master's thesis)

Zhou, X.*, **Zheng, A.***, Yin, J.*, Chen, R., Zhao, X., Xu, W., Cheng, W., Xia, T., & Lin, S. (2015). Context-sensitive spelling correction of consumer-generated content on health care. *JMIR medical informatics*, 3(3), e4211. (*these authors contributed equally)

ABSTRACT OF THE DISSERTATION

Deep learning predicts the impact of non-coding genetic variants in human traits and diseases

by

An Zheng

Doctor of Philosophy in Computer Science

University of California San Diego, 2022

Professor Melissa Gymrek, Chair
Professor Hao Su, Co-Chair

In the human genome, the vast majority of DNA is non-coding. Although non-coding DNA does not directly encode protein sequences, they are vital to the transcriptional regulation of the protein-coding process. Recent genome-wide association studies (GWAS) have shown that ~93% of genetic variants driving common human traits and diseases lie within non-coding sequences. However, due to the complicated and indirect functions of these non-coding genetic

variants, it is difficult for traditional analysis metrics to sift through the large number of non-coding sequences and pinpoint the variants casual to human diseases and traits.

In this dissertation, I present AgentBind, a deep learning framework that identifies and interprets sequence features most predictive of regulatory activities, such as transcription factor binding, histone modification, and chromatin accessibility. I demonstrate that AgentBind is applicable to diverse types of biological tasks, including (1) pinpointing sequence features most important for transcription factor binding; (2) prioritizing genetic variants in transcriptional enhancers associated with human brain disorders; and (3) identifying the dominant combinations of lineage-determining and signal-dependent transcription factors driving enhancer activation in mice. Collectively, these studies provide a valuable deep learning framework and its use cases in decoding the rules within non-coding regulatory regions and identifying specific non-coding nucleotides with the strongest effects on human traits and diseases.

INTRODUCTION

In the human genome, only 1 percent of DNA is made up of protein-coding genes; the rest 99 percent is non-coding (ENCODE Project Consortium, 2007; Feingold et al., 2004). For a long time, non-coding DNA was controversially considered as "junk DNA" without any biological functions. However, in 2012, the Encyclopedia of DNA Elements (ENCODE) project, an international research program aiming to identify functional elements in the human genome, suggested that 76% of the non-coding DNA in the human genome is transcribed and that 42% of the genome across all cell types is accessible to genetic regulatory proteins such as transcription factors (TFs) (ENCODE Project Consortium, 2012; Pennisi, 2012). The ENCODE project also suggested that around 80% of the genome contains elements linked to some biochemical functions. These findings are aligned with recent genome-wide association studies (GWAS) which have shown that the majority (~93%) of genetic variants driving common human diseases lie in regulatory, rather than protein-coding, regions (French et al., 2020; Wells et al., 2019; Spielmann et al., 2016; Zhang et al., 2015; Maurano et al., 2012). In recent years, there has been a growing number of studies focusing on decoding the functions of non-coding DNA sequences. But while it is relatively straightforward to predict the consequences of mutations in coding regions, traditional analysis metrics are far from being able to interpret and sift through the large number of non-coding variants arising from whole-genome studies. Moreover, pinpointing individual causal variants becomes even more challenging when we factor in linkage disequilibrium (LD) which results in blocks of variants being co-inherited and difficult to be differentiated in GWAS (Eraslan et al., 2019).

Previously, several machine learning methods based on convolutional neural networks (CNNs) have been proposed to help address this challenge (Zhou & Troyanskaya et al., 2015;

Quang et al., 2016; Kelley et al., 2016; Quang et al., 2019; Avsec et al., 2021). These metrics model the effects of sequence composition on different types of markers of potential regulatory regions, including DNAseI hypersensitivity, chromatin accessibility, and transcription factor binding. Compared with traditional modeling methods, such as position weight matrix (PWM), these deep learning solutions can process DNA sequences of much wider scope (>100 kbp, Kelley et al., 2018) and capture more complex DNA sequence patterns and dependencies in large datasets (Eraslan et al., 2019). Moreover, due to the nature of deep learning models, deep learning solutions allow us to integrate heterogeneous genetic and epigenetic data in a more direct and organic way, enabling us to make use of their inner connections. Biological information, such as distance from gene regions, 3d genomic structure, recombination rate, or nucleosome occupancy, can be very informative for deep learning models to determine functions of non-coding DNA sequences.

One major challenge that deep learning applications in genomics are facing is that deep neural networks are good at recognizing sequence patterns but difficult to interpret. Pinpointing the functions of individual nucleotides from well-trained models is non-trivial and being actively studied. Several techniques, including in silico mutagenesis, DeepLIFT, and saliency maps, have previously been applied to interpret CNN results on DNA sequences (Selvaraju et al., 2017). These techniques quantitatively annotate the contribution of each nucleotide in a sequence toward the classification prediction.

However, current deep learning applications in genomics still face several limitations. First, many of them use a multi-class training procedure focusing on multiple types of biologically active regions, but do not contain inactive regions as controls. Thus, their models cannot learn general features that distinguish active vs. inactive genomic regions. Second, the model interpretation techniques they use originate from computer vision tasks and have not yet been benchmarked and

evaluated on genomic datasets. The strengths and limitations of these interpretation techniques on genomic data are largely unknown. Third, deep learning models usually require thousands of samples to train, but such a large number of DNA sequences is either very expensive or impossible to acquire in many biological experiments.

Thus, in this dissertation, I present AgentBind, a deep learning framework leveraging both neural networks and model interpretation techniques to identify, visualize, and interpret sequence features most important for determining biological activities, such as TF binding, histone modification, and chromatin accessibility. AgentBind uses a two-step transfer learning scheme, including a pre-training step and a fine-tuning step, to enable this framework to accommodate smaller datasets. AgentBind applies Grad-CAM (Selvaraju et al., 2017), a post-analytical method for neural networks, to compute importance scores for each nucleotide in the input sequences and characterize sequence features predictive of biological functions. Chapter 1 in this dissertation mainly focuses on evaluating the applicability of AgentBind on genomic data and benchmarking model interpretation techniques with a controlled simulated dataset with ground truth available.

Chapter 2 – 4 show three applications of AgentBind in modeling real-world genomic datasets and describe how my colleagues and I use AgentBind to gain biological insights from these datasets. In chapter 2, AgentBind is applied to predicting binding at motifs for 38 TFs in a lymphoblastoid cell line, scoring the importance of context sequences at base-pair resolution, and characterizing context features most predictive of binding. We also find that the choice of training data heavily influences classification accuracy and the relative importance of features such as open chromatin.

In chapter 3, my colleagues and I make use of AgentBind to develop a deep learning pipeline for prioritizing genetic variants associated with human brain disorders. These variants are

predicted to influence brain functions through transcriptional enhancer activities. Specifically, we use the AgentBind framework to model genome regions with strong H3K27ac signals, a significant marker for enhancer activities, and characterize sequence features predictive of H3K27ac activities. We then integrate the AgentBind importance scores with fine-mapping results from GWAS of brain-related traits to identify putative causal variants that may act via modulating enhancer activity.

In chapter 4, we evaluated the effects of >50 million single nucleotide polymorphisms (SNPs) and short insertions/deletions (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. By applying AgentBind to epigenetic data for macrophages from five different mouse strains, we identify the dominant combinations of lineage-determining and signal-dependent transcription factors driving IL-4 enhancer activation. The results further reveal mechanisms by which noncoding 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.

While exploring the applicability of AgentBind in different scenarios, I used a large number of ChIP-seq data for model training and for post-analyses. However, as is discussed above, DNA sequence data are very expensive and time-consuming to generate. Thus, my colleagues and I designed ChIPs, a toolkit for rapidly simulating ChIP-seq data using statistical models of key experimental steps. In chapter 5, I demonstrate how ChIPs can be used for a wide range of applications, including benchmarking analysis tools and evaluating the impact of various experimental parameters. This ChIP-seq simulation framework is highly efficient and flexible. It

can serve as an important component in various ChIP-seq analyses where ground truth data are needed.

Collectively, in this dissertation, I show that AgentBind is a versatile deep learning framework and able to recognize and pinpoint sequence features important for biological activities. Through a series of use cases, I also demonstrate how we can integrate heterogeneous genetic and epigenetic data and exploit this deep learning framework to predict the impact of non-coding genetic variants in human traits and diseases.

CHAPTER 1

AgentBind: a deep learning framework profiling determinants predictive of transcription factor binding

1.1 Introduction

Gene expression is a biological process in which cells read the genetic code written in the DNA and used its information to produce molecule products, such as proteins. In most cases, gene expression is regulated through integrated actions of various genomic cis-regulatory elements, including promoters, enhancers, silencers, insulators, and tethering elements. Among them, enhancers have a leading role in the initiation of gene expression through transcription factors (TFs). TFs are proteins that bind to specific DNA sequences and control the rate of transcription of genetic information from DNA to messenger RNA (Spitz & Furlong, 2012). Most TFs have intrinsic binding preferences to specific motifs, but these motifs cannot completely explain TF binding affinity. For example, according to analyses on ENCODE datasets, the motif for a TF named SP1 occurs more than 3.6 million times across the human genome, whereas less than 0.76% of these occurrences may be bound in vivo in a human lymphoblastoid cell line (GM12878) (The ENCODE Project Consortium, 2012; Davis et al., 2018; Zheng et al., 2021, *Nature machine intelligence*).

Many studies have shown that the sequence context of TF binding sites play an important role in the TF binding process (Westholm et al., 2008; Le et al., 2018). Identifying these features will allow us to better understand the underlying mechanisms of gene regulation as well as pathogenicity of diseases caused by gene regulation disorders. However, traditional algorithms for pattern discovery were shown to be inefficient and error-prone in solving this task, especially for

large heterogeneous datasets that contain multiple motifs exerting the same biological functions and/or depending on each other with various combinations.

Machine learning techniques have been applied to a broad range of applications in genomic tasks, such as identifying tumor samples, annotating gene functions, and predicting genetic variants. In recent years they start to grow in popularity in the analyses of TF binding sites. Several convolutional neural network (CNN) frameworks, such as DeepSEA (Zhou & Troyanskaya, 2015), DanQ (Quang & Xie, 2016) and Basset (Kelley et al., 2016), have been shown to be successful in predicting TF binding activities by taking both binding sites and their context regions as input. Furthermore, people have introduced various interpretation methods from the computer vision field to interpret the classification results from deep learning models into a human understandable format. These methods were used to quantitatively annotate each element in the input data based on its importance towards binding status. However, these frameworks integrated hundreds of TFs in a single model yet did not contain any unbound samples as control. Thus, they may ignore genomic features specific to a particular TF yet uncommon overall.

Here, we will present a novel framework, AgentBind, that takes as input ChIP-seq peaks and a motif for a single TF of interest and outputs (1) the predicted binding scores of each occurrence of the motif in the genome and (2) per-base-pair annotation scores which indicate the importance of each base in determining binding status. AgentBind extracts 1kb DNA sequences from the genome and labels each sequence as bound vs. unbound based on overlap with ChIP-seq peaks. These sequences are then fed into a CNN model for training. It then utilizes Grad-CAM (Selvaraju et al., 2017), a state-of-the-art interpretation method for CNN models, to compute scores for each base pair of each input sequence and annotate how strongly they contributed to the classification outcome.

AgentBind provides three main advantages over published deep learning methods for classifying cell-type specific regulatory elements. First, by conditioning on sequences that contain the same TF motif in both the positive and negative set, this framework is enabled to focus specifically on sequence context features that determine binding status. In contrast, other methods, such as DeepSEA and DanQ, consider bound regions only, and thus they primarily capture the sequence features of core motifs rather than contexts. Second, during model training, AgentBind uses a transfer learning technique and leverages parameters learned by existing models as a starting point. Transfer learning has been shown to effectively reduce the requirement of training data amount and improve the overall predictive performance compared to training from scratch (Avsec et al., 2019). Finally, AgentBind framework focus specifically on interpreting the resulting CNN model. By adapting computer vision interpretation techniques to the DNA sequence analysis, it can pinpoint specific context features that determine whether a given motif instance is bound.

In this chapter, we will mainly focus on the design and implementation of AgentBind framework and evaluate its performance on controlled simulated datasets with ground-truth information available. In the chapter 2 through 4, we will further introduce more applications of AgentBind on real biological datasets and discuss what biological insights people can gain through this framework.

1.2 Background

1.2.1 Deep learning models used in genomics

In recent years, deep learning methods have shown their capability in identifying DNA patterns in large genomic datasets and prioritizing nucleotide variants based on their pathogenic influences (Zhou & Troyanskaya, 2015; Quang & Xie, 2016; Kelley et al., 2016). Compared with

traditional machine learning methods used in genomics, such as support vector machine, deep neural networks models show a higher capacity and flexibility in modeling DNA sequences and identifying patterns (Zou et al., 2019) and thus have become increasingly popular in a wide range of studies in genomics. For example, in cancer genomics, deep learning was applied to extract the high-level features of combinatorial somatic mutations for various cancer types (Yuan et al., 2016) and learn prognostic information (Yousefi et al., 2017). Similarly, to investigate the pathogenicity of autism spectrum disorders, Zhou et al. used a CNN framework to predict the causal regulatory variants to autism spectrum disorders (Zhou et al., 2019, *Nature genetics*). And for nonhuman species, Sundaram et al. used a deep neural network model to analyze hundreds of genomic variants in primates and identified pathogenic variants shared between human and other primates (Sundaram et al., 2018).

There are three families of neural network architectures commonly used in identifying DNA patterns: fully connected, CNN, and recurrent neuron network (RNN) (Zou et al., 2019; LeCun et al., 2015). Fully connected neural networks are the ancestor of the other two. It consists of multiple layers and connects every neuron in one layer to every neuron in another layer. Fully connected neural networks architecture is suitable for generic prediction problems when there are no special relations among the input data features.

CNN models are slightly different from fully connected models: each neuron in a CNN model is only connected with a small and continuous subset of neurons in the previous layer. In CNN models, parameter matrices (i.e., filters) scan across the input matrix, and compute a weighted sum of local context at each position of input (Krizhevsky et al., 2012). This scanning process is similar to the traditional position weight matrix (PWM) method in which people use the PWM of a motif to scan across a DNA sequence and evaluate the resemblance of the motif with

each region in the DNA sequence. In general, CNN models are useful in cases where some spatially invariant patterns, such as sub-sequence patterns, in the input are expected.

RNN models are designed for sequential or time-series data (Goodfellow et al., 2016). Hidden layers of the RNN are memory states that retain information from the sequence previously observed and are updated at each time step. RNN models can use their internal memory to process sequences of variable length.

1.2.2 Model interpretation methods used in genomics

One major challenge that these deep learning applications are facing is that these deep neural networks are good at identifying DNA patterns but not trivially interpretable. Several techniques, including in silico mutagenesis, DeepLIFT, and saliency maps, have previously been applied to interpret CNN results on DNA sequences (Selvaraju et al., 2017). These methods quantitatively annotate the contribution of each nucleotide in a sequence toward the classification prediction.

Among these model interpretation methods, saliency maps are a group of effective efficient methods that compute gradients of neural network outputs with respect to each nucleotide. Saliency map methods (1) require only one step of forward propagation per sample, (2) can be applied to any type of neural network, and (3) can be implemented easily under deep learning frameworks such as Tensorflow (Abadi et. al, 2015) and PyTorch (Paszke et al. 2019).

However, naive implementations, such as the model interpretation method using in Basset (Kelley et al., 2016), are highly sensitive to noise and are susceptible to model saturation (Shrikumar et al., 2017). Thus, Selvaraju et al. implemented Grad-CAM (Selvaraju et al., 2017), an advanced version of saliency maps, which overcomes this challenge using an aggregated

distribution map of important sequence features and integrates this distribution map with input layer gradients through element-wise multiplication. This method is computationally efficient and has been proven to be more stable than the vanilla saliency maps in a wide variety of applications.

1.3 AgentBind framework

1.3.1 Deep learning architecture and training

To properly train the models in the AgentBind framework, we applied a two-step transfer learning scheme, including a pre-training step and a fine-tuning step, which enabled AgentBind to accommodate smaller datasets. In the pre-training step, we trained a CNN model on the dataset from DeepSEA, which consists of 4,863,024 sequences of 1kb annotated with 919 ChIP-seq and DNase-seq profiles collected from ENCODE (Encode Project Consortium et al., 2020; ENCODE Project Consortium, 2012) and the Epigenomics Roadmap Project (Roadmap Epigenomics Consortium et al., 2015) across dozens of cell types. This step allows the CNN model to capture common DNA patterns in regulatory regions and encode them into its convolutional layers.

In the fine-tuning step, we trained an individual classification model for the binary input dataset of each TF of interest. Notably, instead of training from scratch, we initialized the convolutional layers with the parameters we learned in the pre-training step, which allowed the model to inherit the encoded common DNA patterns in regulatory regions from the pre-trained model and focus only on learning the novel patterns specific to the TF of interest.

The Agentbind framework is compatible with virtually any CNN architectures. And as examples, we evaluated its performance using two popular neural network architectures DeepSEA and DanQ, separately. We implemented AgentBind with these two architectures using Tensorflow. DeepSEA consists of three convolutional layers and a fully connected layer. In these convolutional

layers, the size of filters is uniform (i.e., eight) and stride by one each time. And its fully connected layer contains 925 hidden neurons. DanQ consists of one convolutional layer, one bidirectional Long short-term memory (biLSTM) layer and a fully connected layer. There are 320 filters in its convolutional layer with the filter size as 26. And in its biLSTM layer, drop-out technique was used with a rate of 0.5. Its fully connected layer contains 925 hidden neurons same as DeepSEA.

1.3.2 Model interpretation

We implemented four separate model interpretation techniques discussed in the chapter 1.2.2, including in silico mutagenesis, vanilla saliency maps, DeepLIFT, and Grad-CAM. Each of these methods computes individual scores for each nucleotide of the input sequence indicating its importance in determining the model's prediction.

For in silico mutagenesis, we performed computational mutations to assess the importance of every base of the input sequences. More specifically, we substituted each base with its three possible nucleotide substitutions and recorded the changes made by them in terms of the output prediction scores. The greatest score change was used to represent the importance of this base.

For vanilla saliency maps, the importance of each base was quantified using the gradient of the output prediction score with respect to this base. This step was accomplished using a TensorFlow built-in function "gradients".

Grad-CAM is an advanced version of saliency maps which additionally brings in an aggregated distribution map of important k-mers and integrates it with the vanilla saliency map through element-wise multiplication. In practice, to build a k-mer distribution map for a TF, we chose the first convolutional layer as the layer of interest. This layer contains distribution maps for

various sequence features. We calculated a weighted average of them using equation (1.1), in which A^k represents the k-th feature map and its weight a_k is calculated using equation (1.2).

$$A_{Grad-CAM} = relu(\sum a_k A^k) \tag{1.1}$$

$$a_k = \frac{1}{n}\sum_i^n \frac{\partial y}{\partial A_i^k} \tag{1.2}$$

In equation (1.2), A_i^k is the i-th neuron in the k-th feature map (out of n neurons in total) and y is the output neuron of the overall neural network. Intuitively, the weight a_k represents the average of gradients that flow back to the k-th feature map.

In comparison with vanilla saliency map which evaluates the importance of each base individually, this aggregated map highlights the regions that are important to the binding activities. To combine the best aspects of these two maps, we then merged the aggregated distribution map with the vanilla saliency map through element-wise multiplication.

For DeepLIFT, we used version v0.6.10.0-alpha together with Keras v2.3.1 and applied its "revealcancel_fc, rescale_conv" mode for model interpretation. Since DeepLIFT is only compatible with Keras (Chollet, 2015), we first constructed a DeepSEA architecture in Keras matching the TensorFlow implementation and then imported the pre-trained DeepSEA model parameters into this Keras model. The training procedures of fine-tuning were the same as the TensorFlow implementation.

1.4 Performance evaluation

1.4.1 Simulation dataset for benchmarking

We adapted a previously published evaluation scheme from Shrikumar et al. (Shrikumar et al., 2017) and generated a binary dataset with TF motifs embedded at known positions. The

binary dataset consists of 50,000 samples for training, 1,000 for cross validation and 1,000 for testing, and labels are assigned evenly with same number of positives and negatives. All sequences were length 1kb and contained the GATA1 motif (http://compbio.mit.edu/encode-motifs) in the center. Context bases were generated by sampling the nucleotides A, C, G, and T at each position with probabilities 0.3, 0.2, 0.2 and 0.3 respectively.

In each positive sample, we randomly embedded 1-3 instances of the TAL1 motif (http://compbio.mit.edu/encode-motifs) in the context regions. The number of TAL1 motifs embedded in sequences followed a Poisson distribution but truncated after 3. In negative samples, there was no TAL1 motif placed in their sequences. These simulated sequences were fed into the AgentBind framework and annotated at nucleotide resolution using the model interpretation methods described in chapter 1.3.2.

1.4.2 Evaluating AgentBind on benchmarking datasets

First, we evaluated the classification module of AgentBind and examine its ability on recognizing and modeling DNA sequence patterns. The classification performance was quantified using area under the receiver operating characteristic curve (auROC) and the precision recall curve (auPRC). Both DeepSEA and DanQ performed well on this simulated dataset. DeepSEA achieved 0.97 auROC and 0.96 auPRC, while DanQ achieved 0.99 auROC and 0.99 auPRC. It is noteworthy that the embedded patterns in the simulation dataset is relatively easy to capture for deep learning models. In chapter 2, we will further evaluate the performance of these two models in more complex scenarios with real-world biological datasets and examine their ability of modeling DNA sequences.

Next, to benchmark the interpretation module with the use of different interpretation methods, we employed five metrics: (1) run time. All timing experiments were tested in a Linux environment running Centos 7.4.1708 on a server with 28 cores (Intel® Xeon® CPU E5–2660 v4 @ 2.00 GHz), NVIDIA® Tesla® K40c GPU, and 125GB RAM. Only a single core was used for timing. (2) the percentage of the top 5% scoring bases that overlap embedded motifs (accuracy). (3) the percentage of embedded motifs for which at least half of the motif bases are in the top 5% scoring bases (recall). (4) auROC when we move the threshold of top scoring bases from 100% to 0% in (3). (5) signal-to-noise ratios computed as the ratio of scores in embedded motifs to scores in background regions. We evaluated all interpretation methods with a DeepSEA architecture, and all except DeepLIFT with a DanQ architecture since DeepLIFT doesn't currently support hybridized architectures containing RNNs.

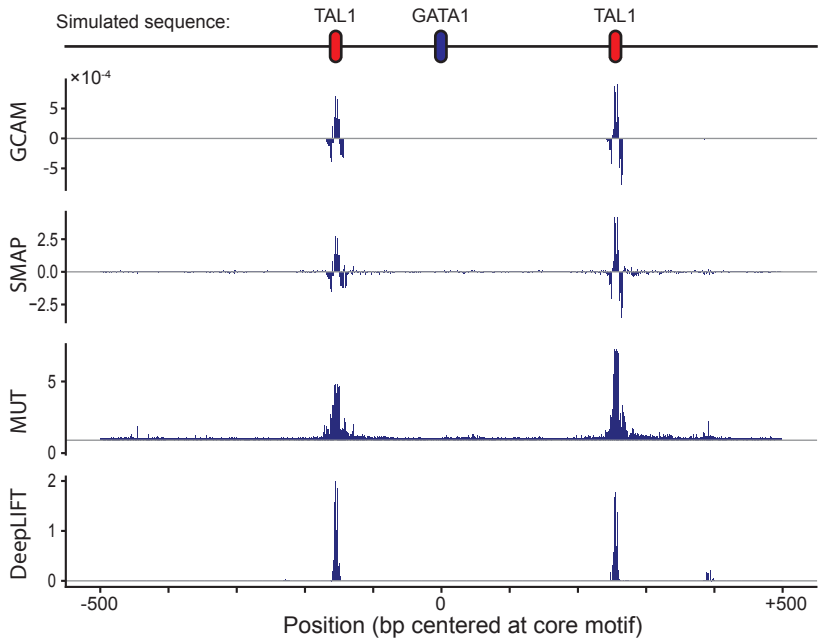


Figure 1.1: Example importance scores for a simulated region. The top shows an example simulated sequence, with a central GATA motif and two context TAL1 motifs. Importance scores are shown for each method based on a DeepSEA architecture.

The results show that all the interpretation methods we evaluated, despite of subtle performance difference, are able to pinpoint the majority of embedded TAL1 motifs (percentage of motif retrieved > 72%; Table 1.1). One example is shown in Figure 1.1, in which there is a GATA1 motif in the center and two TAL1 motifs on each side. All four interpretation methods can annotate this sequence on a nucleotide-level resolution and highlight the locations of TAL1 motifs with high importance scores.

Table 1.1: Comparison of model interpretation techniques. This table shows the performance comparison of four interpretation method (in silico saturated mutagenesis, naive saliency map, Grad-CAM, and DeepLIFT) under different evaluation metrics. The results from both model architectures, DeepSEA and DanQ, are reported.

	Model architecture						
	DeepSEA				DanQ		
Metrics	in silico saturated mutagenesis	naive saliency map	Grad-CAM	DeepLIFT	in silico saturated mutagenesis	naive saliency map	Grad-CAM
runtime (seconds per 1k sequence)	7973	18	62	113	7048	16	62
recovery performance - AUC	0.893	0.871	0.897	0.811	0.986	0.962	0.979
%motif bases recovered	58.15%	61.17%	66.60%	84.33%	91.63%	80.29%	87.43%
%motifs retrieved	71.62%	72.09%	80.01%	99.20%	97.34%	93.67%	95.40%
signal-noise ratio (median value)	1.04	14.36	74.08	2321.83	4.16	31.18	92.79

On the other hand, the results also demonstrate that each interpretation method has unique strengths and weaknesses (Table 1.1). For example, in silico mutagenesis generally shows superior classification of individual bases (recovery performance – AUC = 0.893 with DeepSEA and 0.986 with DanQ) but its run time is two orders of magnitude higher than the other methods (Runtime >

7000 seconds per 1k sequence with both model architectures). DeepLIFT identifies more embedded motif instances whereas Grad-CAM better pinpoints specific important bases (Figure 1.2). We also computed signal-to-noise ratios for each method by taking the ratio of context scores in the embedded TAL1 motif regions to context scores in background regions. Grad-CAM and DeepLIFT greatly outperformed alternative methods, indicating these two methods can more precisely identify the embedded context motifs. In summary, among these methods, Gram-CAM is the most balanced one, with fast run time, high classification accuracy at base pair resolution, and applicability to the better performing DanQ architecture. And that was why we chose this method in the AgentBind framework for the applications in the following chapters.

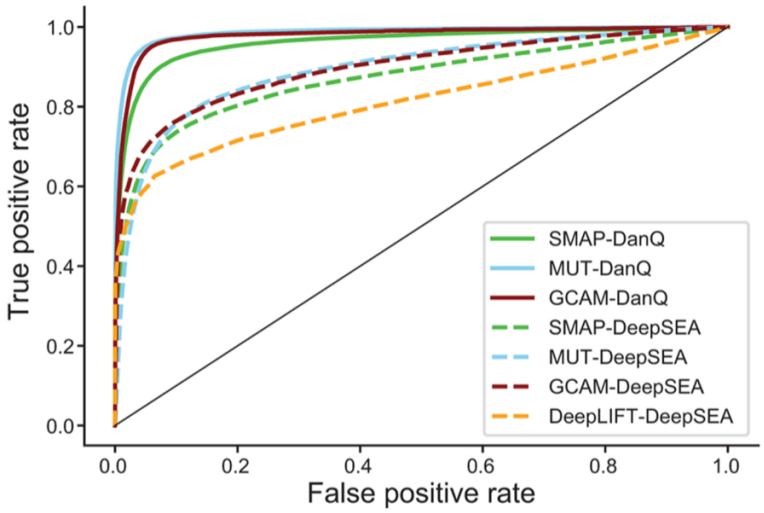


Figure 1.2: Comparison of model interpretation methods. ROC curves are shown comparing performance of each method to distinguish simulated important vs. neutral context bases. Dashed and solid lines denote DeepSEA and DanQ architectures respectively. Green=saliency map (SMAP), cyan=saturated mutagenesis (MUT), red=Grad-CAM (GCAM), orange=DeepLIFT.

1.5 Summary and acknowledgements

This chapter mainly focus on benchmarking AgentBind using a simulated dataset with ground-truth available. We evaluated four model interpretation methods with two different deep learning models. We found that Grad-CAM is the most balanced one in the aspect of runtime speed,

accuracy, and retrieval rates. In the following chapters, we will discuss more about how to apply AgentBind in analyzing real-world genomic datasets and what biological insights people can gain from this framework.

The material used in this chapter is taken from the following two papers: (1) Zheng, A., Lamkin, M., Wu, C., Su, H., & Gymrek, M. "AgentBind: Profiling Context-specific Determinants of Transcription Factor Binding Affinity," published and presented in the ICML 2019 Workshop on Computational Biology (Zheng et al. 2019); (2) Zheng, A., Lamkin, M., Zhao, H., Wu, C., Su, H., & Gymrek, M. "Deep neural networks identify sequence context features predictive of transcription factor binding," published on Nature machine intelligence (Zheng et al. 2021, *Nature machine intelligence*). The dissertation author was the primary investigator and author of this material.

These studies were supported in part by NIH/NHGRI 1R21HG010070-01, the Microsoft Genomics for Research program, and an Amazon Web Services research award. We thank NVIDIA for donating a Tesla K40 GPU to support this project. We additionally thank Christopher Benner and Alon Goren for helpful comments.

CHAPTER 2

Deep neural networks identify sequence context features predictive of transcription factor binding

2.1 Introduction

Binding of transcription factors (TFs) to DNA is one of the major transcriptional regulation mechanisms. TFs typically recognize short motifs of 6–12bp (Lambert et al., 2018). However, there is often only partial overlap between sequences matching the motif for a particular TF in the genome and experimentally determined binding sites (Lambert et al., 2018). For example, we found that <1% of approximately 3.6 million SP1 motifs across the human genome are bound in a human lymphoblastoid cell line (GM12878). Whether a particular motif instance is bound depends on multiple factors, including chromatin accessibility (Zaret et al., 2016), nucleosome positioning (Segal et al., 2006), cooperative and competitive binding with other factors (Morgunova et al., 2017), local GC content (Wang et al., 2012), local DNA tertiary structures (Zhou et al, 2015, *Proceedings of the National Academy of Sciences*; Guo et al., 2018), and inter-position dependencies within motifs (Guo et al. 2018). Many of these features are related to sequence context in the immediate vicinity of the TF motif itself (Westholm et al., 2008), implying that TF binding may be predicted directly from sequence information.

Several machine learning methods (Alipanahi et al, 2015; Kelley et al, 2018; Kelley et al, 2016, Lee et al., 2015; Quang et al., 2016; Quang et al., 2019; Zeng et al., 2016; Zhou & Troyanskaya, 2015) have proven successful in predicting TF binding from sequence. Many of these methods, such as DeepSEA (Zhou & Troyanskaya, 2015) and DanQ (Quang et al., 2016), rely on convolutional neural networks (CNNs), which infer important sequence context features and learn combinations and orientations of these features that are predictive of binding. However,

these frameworks face several limitations. First, they focus on open chromatin regions that are active in at least one cell type of interest, but do not consider regions inactive in all cell types as controls. Thus, they do not learn general features that distinguish bound vs. unbound genomic regions. Second, while these models have shown excellent prediction accuracy for a variety of marks and cell types, interpreting CNNs to derive meaningful biological insights remains challenging.

In this chapter, we will present an application of AgentBind, the deep learning framework introduced in the previous chapter, to predicting whether a particular instance of a TF motif will be bound and interpreting the specific nucleotides with the strongest influence on binding status. By conditioning on sequences that contain the core TF motif in both the positive and negative samples, this framework specifically learns context features in the vicinity of the core motif. Next, we apply Grad-CAM (Selvaraju et al., 2017) to compute importance scores for each nucleotide in the context regions and characterize sequence features predictive of TF binding. We find that TF binding is largely predicted by open chromatin, and to a lesser extent by TF-specific sequence features. The relative importance of these features depends heavily on how positive and negative training sets are chosen. Overall, this framework enables novel insights into sequence features predictive of TF binding.

2.2 Results

2.2.1 Predicting binding status of TF motif occurrences

We focused on 38 TFs active in GM12878 with ChIP-sequencing datasets available from ENCODE18 and motifs available from JASPAR19. For each TF, we scanned the human reference genome (hg19) to identify all instances of its motif, which are referred to as the core motif. We

extracted 1kb genomic sequences centered on each core motif instance and labeled each sequence as bound (positive) vs. unbound (negative) based on overlap with binding sites identified by ChIP-sequencing (Figure 2.1a). On average for each TF, we obtained 18,892 sequences as input for the baseline model.

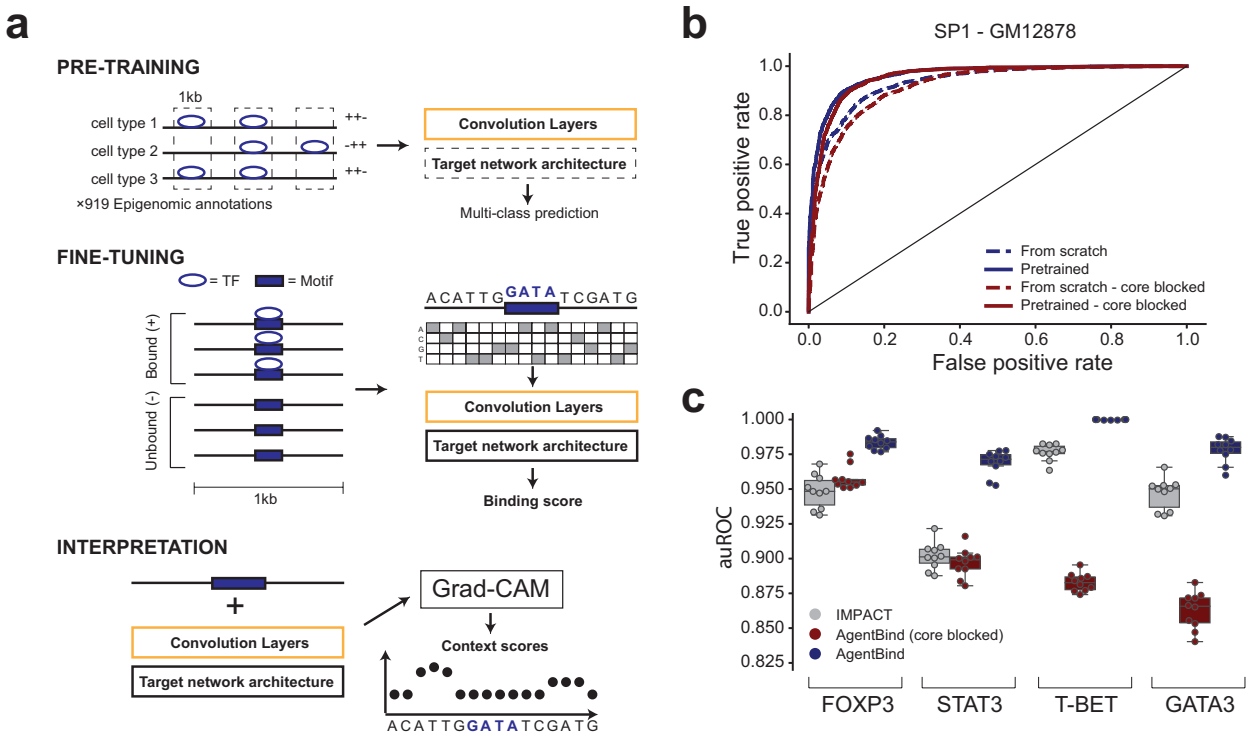


Figure 2.1: AgentBind Overview. (a) Method schematic. AgentBind pre-trains a convolutional neural network on epigenomic annotations from multiple cell types (top). It then fine-tunes on sequences containing a core motif (purple box) for a target TF that are either bound (+) or unbound (−) to learn important context features (middle). Grad-CAM is then used to score the contribution of each nucleotide to binding predictions (bottom). (b) Pre-training improves TF binding predictions. Receiver operator curves (ROC) are shown for the TF SP1 in GM12878 using baseline models with a DanQ architecture. Dashed and solid lines show performance with and without pre-training, respectively. (c) Comparison to IMPACT. We compared the ability of AgentBind and IMPACT to distinguish bound vs. unbound motifs for four TFs in CD4+ Th17 cells. Boxplots show distributions of auROC values for 10 rounds of randomly selecting training (80%) vs. testing (20%) motif instances. Middle lines give medians and boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to Q1–1.5*IQR (minima) and Q3+1.5*IQR (maxima), where IQR=Q3-Q1. Dots show the auROC values for individual rounds. Gray=IMPACT. For (b) and (c), dark blue=AgentBind without the core motif blocked, dark red=AgentBind with the core motif blocked.

In this application, the AgentBind framework consists of (1) pre-training CNNs using ChIP-sequencing and DNaseI-sequencing profiles collected from ENCODE (ENCODE Project

Consortium, 2012) and the Epigenomics Roadmap Project (Roadmap Epigenomics Consortium et al., 2015) across dozens of cell types and (2) fine-tuning an individual model for each TF to identify bound vs. unbound sequences as described above. This framework is compatible with theoretically any CNN-based architecture. As examples, we evaluated its performance using two popular architectures, DeepSEA and DanQ. We evaluated performance using area under the receiver operating characteristic curve (auROC) and the precision recall curve (auPRC), and partial auROC (pAUC) with false positive rate less than 0.121. Average performance across all TFs is high (auROC=0.941 for DanQ and 0.928 for DeepSEA), suggesting that binding is largely predictable by local sequence features within a few hundred base pairs of the core motif. In both evaluation experiments, pre-training noticeably improves the performance compared with models trained from scratch (Figure 2.1b), especially for TFs with low sample sizes such as FOS. This improvement is expected, since pre-trained DanQ and DeepSEA are models optimized for large datasets. Fine-tuned TF-specific models consistently outperform multi-class models for this classification task, with an average auROC increase of 0.033 (range from 0.002 [NFYA] to 0.123 [NRSF]) for DanQ. Because of its consistently higher performance, subsequent results are reported for DanQ unless otherwise specified.

We tested whether the classification performance could be driven by differences either in core motif sequences or nucleotide content not directly relevant to the context features we aimed to identify. We first repeated the analyses with the central core motifs masked. In most cases performance is only slightly reduced after masking (average auROC decrease 0.015). CTCF is a notable exception (auROC=0.945 and 0.798 before and after blocking, respectively), suggesting its bound vs. unbound regions have key differences within core motifs despite being similarly scored by position weight matrices (PWMs). We further observed that model performance is

correlated with the difference in GC content between bound vs. unbound regions (Pearson r=0.58; two-sided P=0.00012; n=38, Supplementary Figure 2.1a–b). To ensure that the models focus on more specific sequence features, we retrained models using negative and positive datasets with matched GC percentages. These models, which are referred to as GC-controlled, have only slightly lower performance (mean decrease in auROC=0.013).

We hypothesized that these models are learning a combination of general sequence features characteristic of active regulatory regions in open chromatin and more specific sequence features required for each TF. To determine the extent to which the predictions are driven by features of open chromatin, we re-trained GC-controlled models restricting all sequences to be within DNaseI hypersensitive sites in GM12878. As expected, overall performance for these models, which are referred to as DNaseI-controlled, decreases (mean decrease in auROC=0.133 compared to the GC-controlled model), suggesting open chromatin features make a major contribution to classification accuracy for most TFs. Notably, controlling for DNaseI results in greatly reduced sample sizes (mean decrease 50%) which may in part drive this trend (Supplementary Figure 2.1c–d). Even after controlling for DNaseI, auROC values are above 0.7 for 32/38 TFs, indicating that while open chromatin is a major predictor, it does not completely determine binding.

To additionally investigate the predictive power of chromatin accessibility alone, we used AgentBind to predict binding of each TF using GM12878 DNaseI-seq output of pre-trained DanQ models. For baseline and GC-controlled datasets, DNaseI alone is highly predictive of binding (mean auROC=0.882 and 0.841), although fine-tuned TF-specific models outperform DNaseI alone (mean auROC=0.941 and 0.928 for baseline and GC-controlled datasets). On the other hand, for DNaseI-controlled data, DNaseI alone is a relatively poor predictor as expected (mean auROC=0.634, compared to 0.795 for TF-specific models).

To further determine whether the models identify context features specific to each TF, we performed a pairwise comparison in which we used models for each TF to predict binding at motifs for all other TFs. For GC-controlled models, binding for most TFs could be predicted well using models for most other TFs (Supplementary Figure 2.1e–f). Still, most (33/38) are predicted best by their own model (median gain in auROC=0.017 compared to the next best model). For the 5 TFs predicted better by other models, the performance difference is negligible (median difference in auROC=0.0060). For DNaseI-controlled models, TF-specific models tend to show higher performance compared to the next best model (Supplementary Figure 2.1g–h). Taken together, these results suggest as expected that GC-controlled models largely learn features indicative of open chromatin but capture some TF-specific features, whereas DNaseI-controlled models better capture features specific to each TF.

We compared these results with two alternative methods, KSM (Guo et al., 2018) and IMPACT (Amariuta et al., 2019), which are not based on deep learning. KSM represents TF motifs as a set of aligned k-mers that are overrepresented at TF binding sites and more accurately predicts in vivo binding sites than PWM models. We identified KSM motifs for each TF using the same set of training and test data as used for AgentBind. For GC-controlled models, AgentBind outperforms KSM in predicting binding status for all TFs (median gain in auROC=0.261). For DNaseI-controlled models, AgentBind outperforms KSM for 33/38 TFs (median gain in auROC=0.182). IMPACT tackles a similar classification task to Agentbind but uses a broad range of experimentally determined epigenomic features including ChIP-sequencing, ATAC-sequencing, and DNaseI-sequencing profiles. While IMPACT has a variety of applications such as prioritizing causal variants for gene expression and complex traits, we only evaluate the application of binding site prediction here. We benchmarked each method on four TFs active in CD4+ T cells and applied

the same training scheme as the IMPACT study (chapter 2.3.4). AgentBind demonstrated higher auROC than IMPACT in all four cases (Figure 2.1c). This suggests that the majority of determinants of binding for these TFs can be learned directly from local sequence features. For FOXP3 and STAT3, performance was comparable to IMPACT even with core motifs blocked, meaning classification decisions were largely based on context sequence rather than differences in the core motifs themselves.

2.2.2 Identifying the context-specific determinants of TF binding

Although deep neural networks achieve high classification accuracy, compared to simpler linear models they are not trivially interpretable. In the previous chapter, we have shown that Grad-CAM is overall performing better with genomic datasets especially because of its fast run time, high classification accuracy at base pair resolution, and applicability to the better performing DanQ architecture. And thus, we chose Grad-CAM for the downstream analyses of this application.

We applied Grad-CAM to interpret the GC-controlled models for the 38 TFs and computed importance scores for each base in input sequences. As expected, aggregating scores across all input sequences for each TF shows that sequences closest to the core motif tend to have the highest impact (Figure 2.2a). However, aggregate score profiles differ noticeably for different TFs. For example, whereas the important bases for predicting CTCF binding are highly concentrated directly adjacent to the motif, important bases for YY1 are spread across the entire 1 kb region (Figure 2.2c). In concordance with the results above, differences in core motifs themselves receive high importance scores for some TFs (e.g., PU1, CTCF) but not others (e.g., MEF2A, SP1, Figure 2.2c). Context scores for bound (positive) sequences show far more distinct patterns than for

unbound (negative) sequences (Supplementary Figure 2.2). Therefore, we focus on scores for bound sequences for downstream analyses.

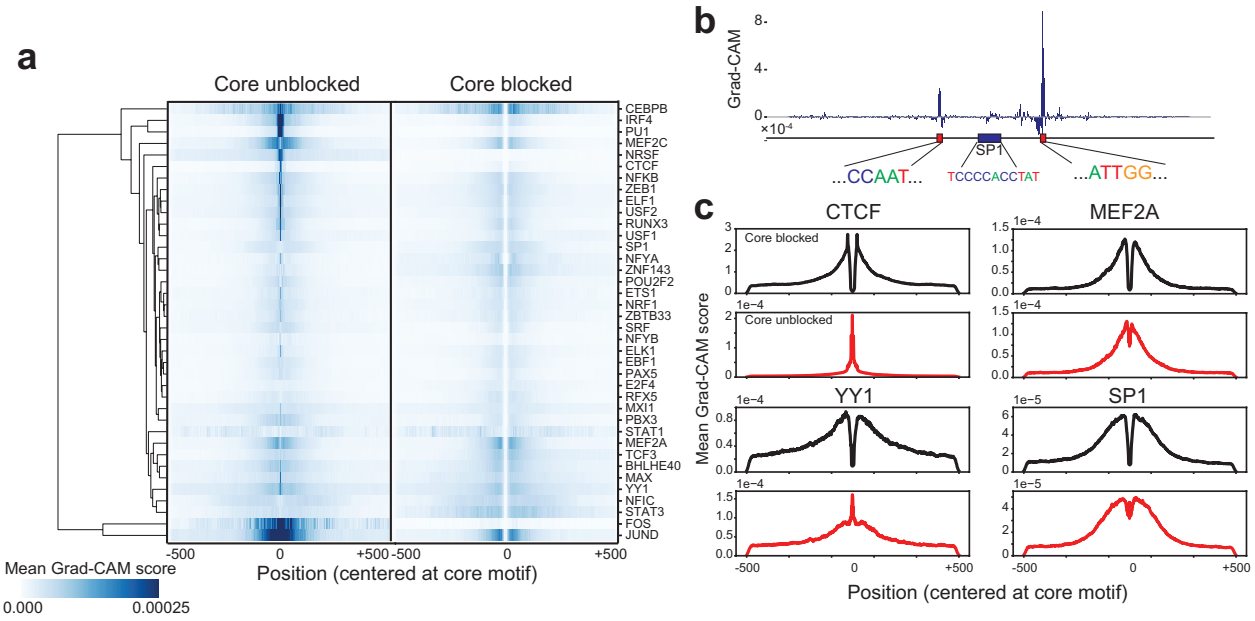


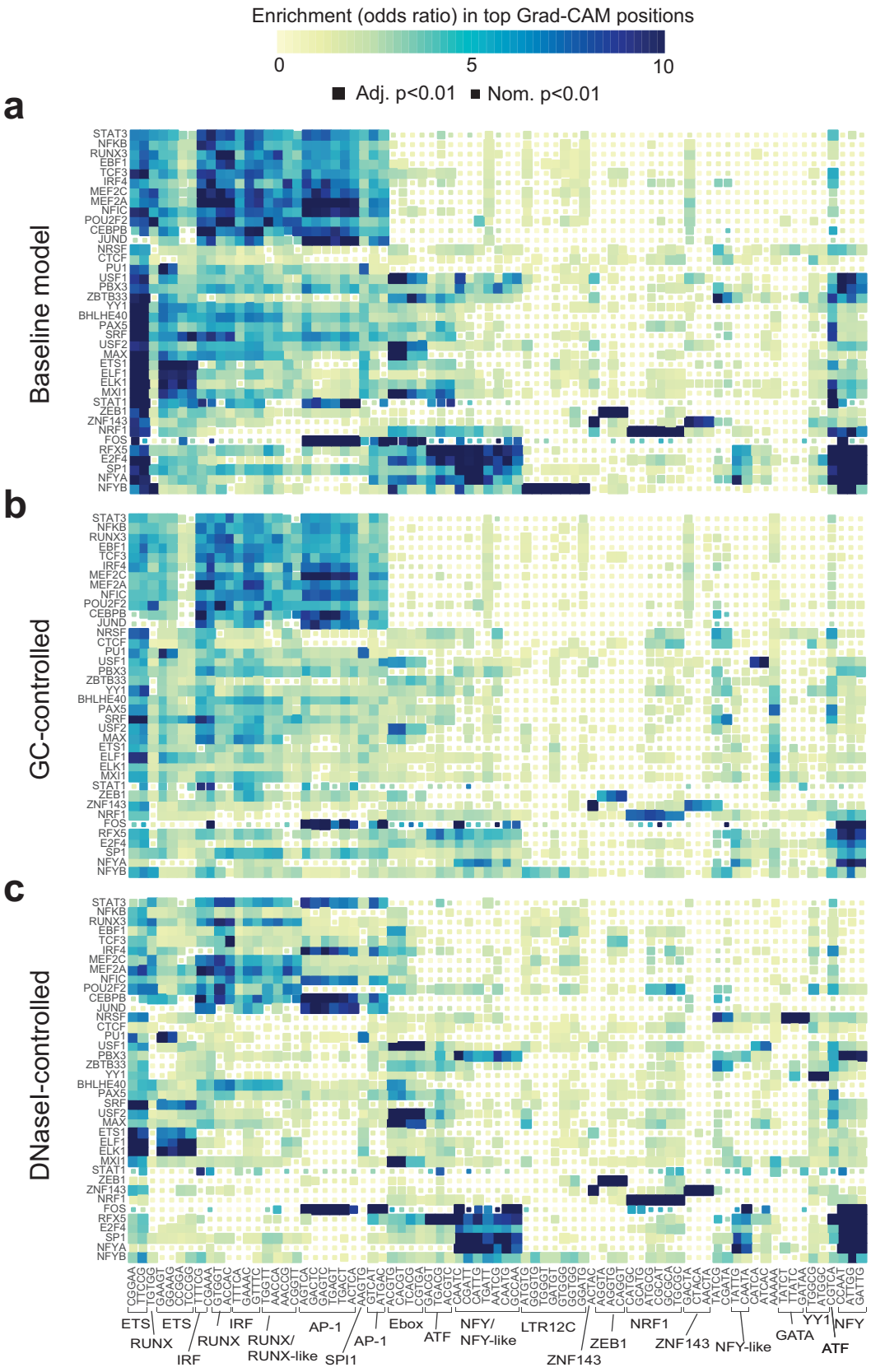
Figure 2.2: Interpreting context-specific determinants of TF binding. (a) Aggregate Grad-CAM score profiles. For each TF, we computed the average absolute value of the Grad-CAM score per position in positive sequences using GC-controlled models with the core motif unblocked (left) or blocked (right). Values were Z-normalized across rows. The dendrogram is based on hierarchical clustering of the rows. (b) Example Grad-CAM scores for a region (chr1:12289432–12290431 in hg19) containing an SP1 motif. The y-axis shows the Grad-CAM score of each nucleotide based on the GC-controlled model. Sequences are shown for the central SP1 motif and two regions with high scores corresponding to NFY motifs. (c) Example aggregate Grad-CAM profiles. For four representative TFs, average Grad-CAM scores are shown for models with the core motif blocked (dark blue) or unblocked (dark red).

2.2.3 Grad-CAM scores give insight into features of TF binding

We next sought to use Grad-CAM score profiles to identify context sequence features with strongest impact on binding status for each motif. We extracted 5-mers from positive sequences accounting for the strand of the core motif and tested whether each unique 5-mer is enriched among 5-mers with highest average Grad-CAM scores for each TF (chapter 2.3.5). The results using the baseline models recapitulate multiple known trends (Figure 2.3a). First, the top scoring sequences for a TF often closely match the core motif of the TF itself, consistent with previous literature

showing homotypic clusters of TF motifs can promote binding (Gotea et al., 2010). For example, 5-mers from the NRF1 (5'-TGCGCATGCGCA-3') and ZEB1 (5'-CAGGTG-3') motifs score highly for NRF1 (Fisher's exact test one-sided $P<10^{-200}$; OR=14.1 for ATGCG) and ZEB1 ($P<10^{-200}$; OR=18.2 for CAGGT), respectively. In some cases, these enrichments are strand specific. For instance, the ZEB1 motif is consistently enriched in important context bases for ZEB1, whereas its reverse complement is not. Similar trends are observed for other factors such as YY1 and ZNF143. Second, top 5-mers capture known co-binding relationships. For example, the NFY motif scores highly among known co-binders SP1 (Roder et al., 1999) and RFX5 (Dolfini et al., 2016). Additionally, the motif for AP-1 (5'-TGA G/C TCA-3'), bound by of a dimerization of JUN and FOS (van Dam et al., 2001), scores highly for known co-binders CEBPB (Heinz et al., 2010) and IRF4 (Li et al., 2012). These trends are also observed using GC-controlled and DNaseI-controlled models (Figure 2.3b–c).

Figure 2.3: Identifying key context sequence features for TF binding in GM12878. (a) Enrichment of 5-mers in the most influential context regions. The heatmap shows the enrichment of each sequence in regions with the highest Grad-CAM scores for each TF using baseline models. Heatmaps in (b-c) are the same as in (a) but show data for GC-controlled (b) and DNaseI-controlled (c) models. Only 5-mers corresponding to top 50 5-mers in at least one of the three models are shown. Colors denote odds ratios, and the sizes of the boxes denote statistical significance based on one-sided P-values computed using Fisher's exact tests. Adjusted P-values are based on a Bonferroni correction for the number of 5-mers tested. The color scale is capped at 10. Odds ratios higher than 10 are all colored the same. Boxed and annotated 5-mers correspond to known motifs. The order of TFs (y-axis) and 5-mers (x-axis) is the same for all plots and is based on hierarchical clustering of the odds ratio matrix for the baseline model.


Enrichment (odds ratio) in top Grad-CAM positions
0
5
10
Adj. p<0.01
Nom. p<0.01
a
Baseline model
b
GC-controlled
c
DNaseI-controlled
STAT3
NFKB
RUNX3
EBF1
TCF3
IRF4
MEF2C
MEF2A
NFIC
POU2F2
CEBPB
JUND
NRSF
CTCF
PU1
USF1
PBX3
ZBTB33
YY1
BHLHE40
PAX5
SRF
USF2
MAX
ETS1
ELF1
ELK1
MXI1
STAT1
ZEB1
ZNF143
NRF1
FOS
RFX5
E2F4
SP1
NFYA
NFYB
ETS
RUNX
ETS
RUNX
IRF
IRF
RUNX/
RUNX-like
AP-1
AP-1
SPI1
Ebox
ATF
NFY/
NFY-like
LTR12C
ZNF143
ZEB1
NRF1
ZNF143
NFY-like
GATA
YY1
ATF
NFY

While these three different models (baseline, GC-controlled, and DNaseI-controlled) capture many similar trends, each of them also highlight orthogonal context features relevant to TF binding. Baseline models identify many key elements of promoter regions (Figure 2.3a), which comprise approximately 56% of ChIP-seq peaks analyzed. For example, top-scoring 5-mers include the NFY and ETS motifs, both of which have previously been shown to act as cardinal elements of certain promoter regions (Benner et al., 2013). Both baseline and GC-controlled models identify clusters of TFs with highest scoring context 5-mers corresponding to known pioneer factors (e.g., NFY (Dolfini et al., 2016), RUNX32/AP-1 (Mevel et al., 2019), and interferon regulatory factors [IRFs] (Kröger et al., 2017)) which open chromatin and enable additional TFs to bind. These pioneer factors motifs are far more strongly enriched in the fine-tuned models compared to pre-trained DanQ models not trained on negative sequences (Supplementary Figure 2.3).

In DNaseI-controlled models, which only consider sequences already in open chromatin, motifs for pioneer factors such as AP-1, RUNX, and IRF are less prevalent in top-scoring 5-mers for many TFs (Figure 2.3c), suggesting the pioneer factors do not directly co-bind with those TFs. On the other hand, pioneer motifs remain enriched for TFs known to physically co-bind (e.g., AP-1 motif for IRF4 and CEBPB). We hypothesize these DNaseI-controlled models instead identify 5-mers that represent cooperative relationships between TFs or sequence elements near the core motif required for binding. For some TFs, top 5-mers in DNaseI-controlled models are distinct from those in the other models. For example, in the baseline model for NRSF, 5-mers corresponding to the pioneer IRF and promoter ETS motifs are most significant, whereas the GATA motif (5'-GATAA-3') is only moderately enriched (one-sided P=0.000045, OR=1.7). However, in the DNaseI-controlled model, the GATA motif is highly significant (one-sided

P=1.2×10−245, OR=11.5) for NRSF, suggesting a potential role for this sequence in promoting nearby NRSF binding after the surrounding region is made accessible by pioneer factors.

We hypothesized that the sequence context features which promote binding of a particular TF to its core motif might differ between motifs in promoter (+/− 3kb from transcription start sites [TSS], denoted as "proximal" vs. enhancer regions (>3kb from the nearest TSS, denoted as "distal"). We repeated the analysis of top 5-mers separately for proximal and distal binding sites (Supplementary Figure 2.4). In some cases, such as for SP1, the highest scoring 5-mers differ dramatically between proximal and distal sites. Overall, NFY and NFY-like motifs score highly for proximal binding sites but have less influence on distal sites. On the other hand, RUNX, AP-1, and IRF motifs show stronger scores for distal sites. These results suggest that many features influencing binding are orthogonal at promoter vs. enhancer regions and these sites are likely governed by separate sets of pioneer and other factors.

To investigate the ability of the AgentBind framework to capture cell-type specific regulatory features, we trained separate GC-controlled models to predict STAT3 binding using ChIP-sequencing data from GM12878, CD4+ Th17, and HeLa cells and used each model to predict binding in all three cell types. As expected, STAT3 binding in each cell type was best predicted by a model trained on that cell type. We computed Grad-CAM scores for each bound sequence and repeated the analysis of top scoring 5-mers as described above. Our analysis reveals that some enriched 5-mers are shared across multiple cell types whereas others are highly cell-type specific (Figure 2.4). For example, RUNX and IRF motifs are enriched only in GM12878 whereas FOX and T-box motifs are enriched only in Th17. AP-1 and BATF motifs are enriched in both HeLa and GM12878, and ETS motifs are enriched in both Th17 and GM12878. Overall, these results

are consistent with a model whereby STAT3 binds to regions made accessible by different combinations of pioneer factors in each cell type.

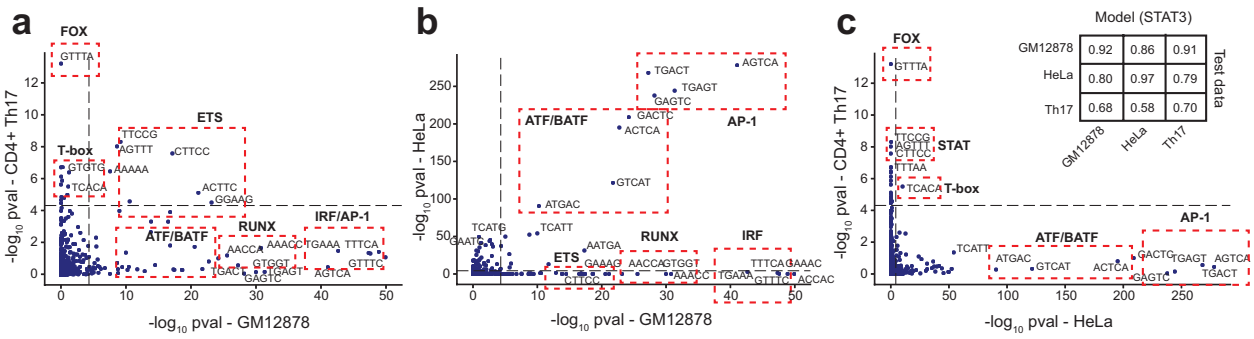


Figure 2.4: Cell-type specific enrichment of 5-mers influential for STAT3 binding. Enrichments were computed using Fisher's exact tests as in Fig. 3 using GC-controlled models trained separately in either GM12878, HeLa, or CD4+ Th17 cells. (a) Comparison of top-scoring 5-mers for GM12878 vs. CD4+ Th17 cells. (b) Comparison of top-scoring 5-mers for GM12878 vs. HeLa cells. (c) Comparison of top-scoring 5-kmers for HeLa vs. CD4+ Th17 cells. The inset table shows the auROC obtained from training each model on one cell type and using it to predict STAT3 binding status in another cell type. Dashed horizontal and vertical lines denote an adjusted P-value threshold of 0.05, based on a Bonferroni correction for the number of 5-mers tested.

Finally, we investigated whether top-scoring SNPs are enriched for properties characteristic of causal variants. We find that SNPs with top-scoring Grad-CAM scores (top 0.5%) show significantly higher signals of negative selection based on observed allele frequencies compared to other SNPs in context regions (two proportion z-test one-sided $P=3.3\times10^{-8}$, Supplementary Figure 2.5). Further, we compared Grad-CAM scores to effects of SNPs on expression measured through massively parallel reporter assays (MPRA) (Tewhey et al., 2016) in LCLs and find that Grad-CAM scores are significantly higher for SNPs that induce expression changes in MPRA (Mann-Whitney two-sided P=0.013), although still are only moderately predictive of causal variants for gene expression. Overall, these results suggest that context bases most influential for TF binding identified by the AgentBind framework may be helpful in prioritizing variants relevant to human traits.

2.3 Method details

2.3.1 ChIP-sequencing datasets and preprocessing

We used FIMO (Grant et al., 2011) v4.12.0 to identify all instances of the motif for each TF across the human reference genome (hg19). FIMO takes the reference genome and target motifs for each TF as input and returns all occurrences of the target motif (using the default p-value threshold $p<10^{-4}$). Motifs for each TF were obtained from JASPAR. We intersected motif instances with binding sites as identified by ChIP-sequencing available for each TF in GM12878 from the ENCODE Project using a custom script. ENCODE ChIP-sequencing experiments for each TF were performed in duplicate and peaks were scored against an appropriate control designated by the ENCODE Analysis Working Group. Motif instances (core motifs) were labeled as positive if they were fully within ChIP-sequencing peaks for the TF. All other instances were labeled as negative. We extended each core motif region to include 1 kb centered at the motif. For each sequence, we included it and its reverse complement sequence for the training procedure described below. In the experiments that required core motifs to be blocked, we substituted the motif region with a string of "N"s of the same length as the JASPAR motif.

The binary datasets acquired above were highly imbalanced: on average we identified 433 times more negative than positive sequences. To balance the dataset ratio while alleviating effects of differences within the core motif, we chose an identical number of negative and positive sequences for each TF while requiring the distribution of motif match p-values to be similar. To obtain p-value matched sets, we binned -log10 P-values of motif matches into bins of size 0.1. For baseline models, we randomly selected the same number of positive and negative sequences from each P-value bin. For GC-controlled models, within each P-value bin, we further binned sequences

based on their GC contents to ensure the selected sequences shared the same distribution of GC contents in positive and negative datasets. For the DNaseI-controlled models, we only considered both positive and negative sequences whose core motifs fall within DNaseI hotspots in GM12878 (ENCODE accession ENCFF491BOT). We then followed the same procedures as in the GC-controlled models to match motif P-values and GC content between positive and negative sequences.

2.3.2 CNN model architecture and training

We implemented DeepSEA and DanQ architectures using TensorFlow (Abadi et al. 2015) v1.9.0. The well-trained models and their associated code are available in a Github repository (https://github.com/Pandaman-Ryan/AgentBind). DeepSEA consists of three convolutional layers and two fully connected layers, and DanQ consists of one convolutional layer, one bi-directional recursive neural network layer and two fully connected layers. We applied sigmoid cross entropy as the loss function for both models.

The sizes of input datasets vary widely, from 182 to 107,539 total sequences per TF. To enable the AgentBind framework to accommodate smaller datasets, we applied a two-step transfer learning scheme, including a pre-training step and a fine-tuning step. Transfer learning has been shown to dramatically reduce the amount of training needed for related classification tasks and improves the overall predictive performance compared to training from scratch (Avsec et al. 2019). For pre-training, we downloaded a dataset consisting of 4,863,024 1kb sequences annotated with a total of 919 ChIP-seq and DNase-seq profiles available on the DeepSEA website. We left out sequences on chromosome 8 for cross validation and sequences on chromosome 9 for testing. We applied one-hot encoding to convert nucleotide sequences into 4-element vectors as has been done

in previous studies (Kelley et al., 2016; Zhou & Troyanskaya, 2015). "N"s were converted into vectors with entries of 0.25 for each of the four nucleotides. During training, we initialized all model parameters with random Gaussian noise with mean 0 and standard deviation 10^{-2} and trained this model on the DeepSEA compendium dataset until the loss function converged. In the fine-tuning step, we used the same architectures as in the pre-training step, and we built an independent model for each TF of interest using the labeled dataset described in the previous section. Same as the pre-training step, we left out sequences on chromosome 8 and 9 for cross validation and testing respectively. From the pre-trained model, we transferred its convolutional layers and RNN layer into the new models but initialized the fully connected layers again with random Gaussian noise. These new models were further fine-tuned on the TF binary datasets until convergence.

2.3.3 Benchmarking experiment against KSM

In the KSM experiment, to identify KSMs for each TF, we used the same set of training and test data as we used in the GC-controlled and DNaseI-controlled models and kept the central 61bp in each sequence. KMAC is a de novo motif discovery method for KSM. We applied KMAC to identify KSM motifs with k_win set as 61, k_min as 5, k_max as 13, and k_top as 10. Finally, we applied KSM to predict the TF binding status of the test data with motifs identified by KMAC as input. We quantified the performance evaluation using auROC, partial auROC, and auPRC.

2.3.4 Benchmarking experiment against IMPACT

The IMPACT study focused on TFs active in T cells and created their own binary (bound vs. unbound) datasets for TFs including FOXP3 (Treg), GATA3 (Th2), STAT3 (Th17) and T-BET (Th1). The coordinates of motif instances for these four TFs were published on the IMPACT

Github repository (https://github.com/immunogenomics/IMPACT). In the benchmarking experiment, we used an identical set of motif instances, extending them into 1 kb sequences to train the model.

We applied an identical training scheme as was used by IMPACT: we randomly selected 80% of the sequences in the input dataset for training and tested on the remaining samples. We evaluated the method in four situations using different architectures and core motif treatments (DeepSEA or DanQ architecture, with core motif blocked or unblocked), and for each situation we conducted 10 parallel trials with different selections of the test set.

2.3.5 Model interpretation methods

In the implementation of Grad-CAM, we chose the first convolutional layer as the layer of interest. This layer contains distribution maps for various sequence features. Following the weighting method proposed by the Grad-CAM authors (Selvaraju et al., 2017), we quantified the importance of these sequence features and computed a weighted summation of all the distribution maps. In comparison with vanilla saliency map which evaluates the importance of each base individually, this aggregated map highlights the regions that are important to the binding activities. To combine the best aspects of these two maps, we then merged the aggregated distribution map with the vanilla saliency map through element-wise multiplication.

2.3.6 K-mer enrichment analysis

In this analysis, we first segmented all the input sequences into 5-bp subsequences using a sliding window and removed subsequences overlapping core motifs in the center. Next, for each subsequence, we quantified its importance by averaging the Grad-CAM scores of each base. For each factor, we ranked all the subsequences based on their Grad-CAM scores and marked the top

1% as top 5-mers. We used a Fisher's Exact test to determine whether each 5-mer was enriched among top 5-mers for each TF. Fisher's Exact tests were performed using the fisher_exact function in the stats module of the Python scipy library v1.3.1. 5-mers were matched to published motifs in the Hocomoco (Kulakovskiy et al., 2018) database based on manual inspection. For Figure 2.3 and Supplementary Figure 2.3 and 2.4, we obtained the top 50-mers ranked by the maximum odds ratio across all TFs separately in each of the three models (baseline, GC-controlled, and DNaseI-controlled). We merged this set for a total of 77 unique 5-mers and clustered matrices of odds ratios for each 5-mer in each TF. For clustering, all insignificant odds ratios (nominal $P \geq 0.01$) were set to 0. To make heatmaps visually comparable, we used the ordering of 5-mers and TFs based on clustering results from the baseline models in each 5-mer heatmap. For the comparison of proximal and distal sites, proximal sites were defined as sequences for which the core motif is within +/− 3kb from the nearest transcription start site (TSS) and distal sites were defined as sequences for which the core motif is >3kb from the nearest TSS. Transcription start sites were annotated based on GENCODE v19.

2.3.7 Cross cell-type comparison of STAT3 models

GC-controlled models for STAT3 in three cell types were trained using the procedure described above. Samples were labeled as positive vs. negative based on overlap with STAT3 peaks in GM12878, HeLa (obtained from ENCODE data) and CD4+ Th17 cells (GEO accession GSM2545819). Input datasets consisted of 2,648, 792, and 7,652 sequences for GM12878, CD4+ Th17, and HeLa, respectively, with equal numbers of positive and negative sequences.

2.3.8 Allele frequency analysis

To quantify selection for a set of genomic positions, we assessed whether those positions are depleted of common genetic variation compared to nearby positions. We focused on single nucleotide polymorphisms (SNPs) present in gnomAD (Karczewski et al., 2020) overlapping sites that were scored by Grad-CAM using GC-controlled models for each TF and computed the percentage of SNPs for which the alternate allele is observed only once (termed singletons). This "percent singleton" metric has previously been used as a proxy for deleteriousness of a set of SNPs (Lek et al., 2016).

For each TF, we overlapped bound sequences scored by Grad-CAM with SNPs in the control samples reported in the gnomAD v2 dataset (Karczewski et al., 2020). For positions overlapping gnomAD SNPs, we recorded observed counts of minor alleles. We then labeled sites where the minor allele counts were 1 as singletons. We only included samples annotated in gnomAD as healthy controls (n=5,442 individuals) in the analysis and required a minimum total allele count of 1000. Sites not overlapping a gnomAD SNP (i.e., minor allele count of 0) were excluded from singleton analysis. The singleton ratio of a group of sites is then simply defined as the percentage of SNPs in that category that are singletons.

2.3.9 CNN model architecture and training

We obtained MPRA results for expression quantitative trait loci (eQTLs) tested in two lymphoblastoid cell lines from Table S1 of Tewhey, et al. (Tewhey et al., 2016). We converted SNP rsids to hg19 coordinates based on dbSNP (Sherry et al., 2001) build 147 and retained only SNP variants which overlapped positions scored by Grad-CAM in at least one TF of interest in DNaseI-controlled models. We further filtered SNPs for which the regulatory effect was not scored

in the Tewhey et al. dataset (C.Skew.fdr column set to NA) indicating one or both alleles of the SNP did not drive expression of the reporter in the MPRA experiment. We treated variants with FDR<5% in the MPRA data (based on the column C.Skew.fdr) as true positives and FDR>=5% as true negatives. A total of 116 true positive and 226 true negative SNPs were included in the analysis. We then set the Grad-CAM score for each variant as the maximum value recorded across all TFs considered at the locus.

2.4 Discussion

In this chapter, we presented AgentBind, a machine learning framework to predict whether instances of TF motifs in the genome were bound vs. unbound in the given cell type and to identify the most influential context bases. While we focused on TFs in GM12878 using the DanQ architecture, this framework can similarly be applied to a flexible range of CNN model architectures for any TF and cell type of interest for which ChIP-sequencing data is available.

The experiment results support the hypothesis that a variety of context features work together to determine whether a motif instance will be bound. The large decrease in auROC values after controlling for DNaseI (mean=0.133) suggests the most important binding determinant for most TFs is whether its motif falls in a region of active chromatin previously opened by a pioneer factor. However, in all the model settings a TF is usually predicted best by its own model. This suggests that even after a region is open, for some TFs additional context sequence features, such as additional copies of its own motif or those of co-binding TFs, are important for determining whether the core motif is bound.

We generated three different models for each TF, each of which identifies distinct sequence features most predictive of TF binding. These different settings highlight how the choice of negative vs. positive sequences for training models has a major impact on the features learned. In

the baseline and GC-controlled models, we fine-tune existing DanQ models with negative training samples from regions of the genome that are inactive in most cell types. Accordingly, the most prominent features learned correspond to known motifs for pioneer factors, which predict whether a region is open or closed. DNaseI-controlled models, which only consider both positive and negative sequences in open chromatin, give decreased importance to pioneer factors and likely highlight sequence features most directly related to TF binding. Importantly, the appropriate model may depend on the application of interest. For example, baseline models may be most appropriate for predicting the impact of a medically relevant variant, where it is simply desirable to have the highest prediction accuracy. On the other hand, for the application of learning sequence features that directly interact with the TF of interest, DNaseI-controlled models are best.

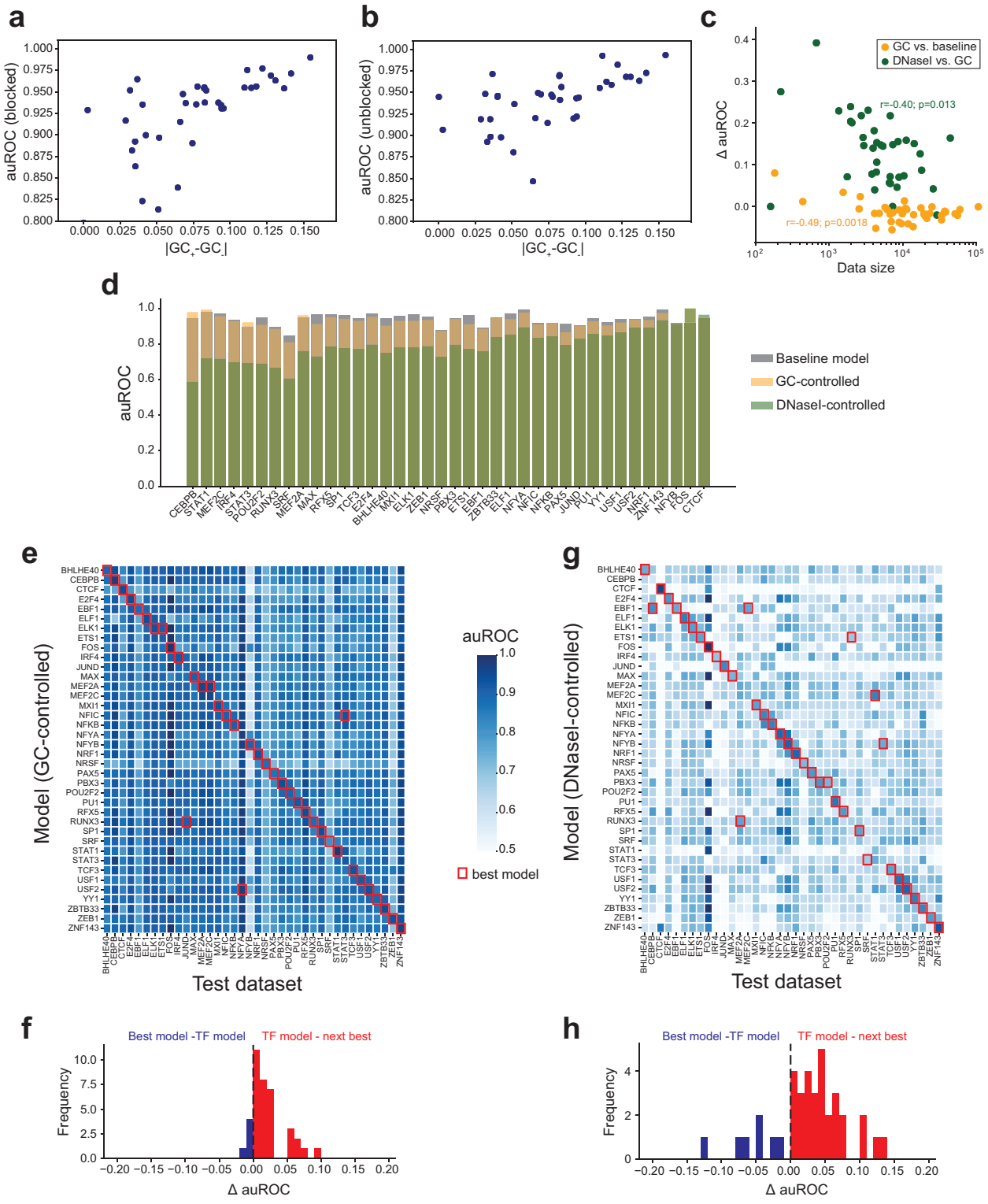
This study faced several limitations: (i) modifications to the training process, such as varying the lengths of context sequences or training separate models for distal vs. proximal regions, are likely to improve performance. (ii) Further, this application relies on PWMs to identify motif instances. PWMs suffer from known limitations, including an inability to capture dependencies between positions, which may trivially distinguish bound vs. unbound sequences in some cases. (iii) Model interpretation techniques can be further improved to extract more complex rules for TF binding such as motif spacing, orientation, and combinations. Visualization techniques such as DeepResolve (Liu et al., 2019) may reveal additional patterns such as interactions between important sequence features learned by CNNs. (iv) TF binding does not necessarily imply regulatory function and thus a high-scoring Grad-CAM site may ultimately not affect gene regulation of downstream phenotypes. Other methods based on a combination of deep-learning and k-mer based approaches have been developed to specifically predict expression from sequence content (Kelley et al., 2018; Zeng et al., 2016). In future work, the scores from AgentBind could

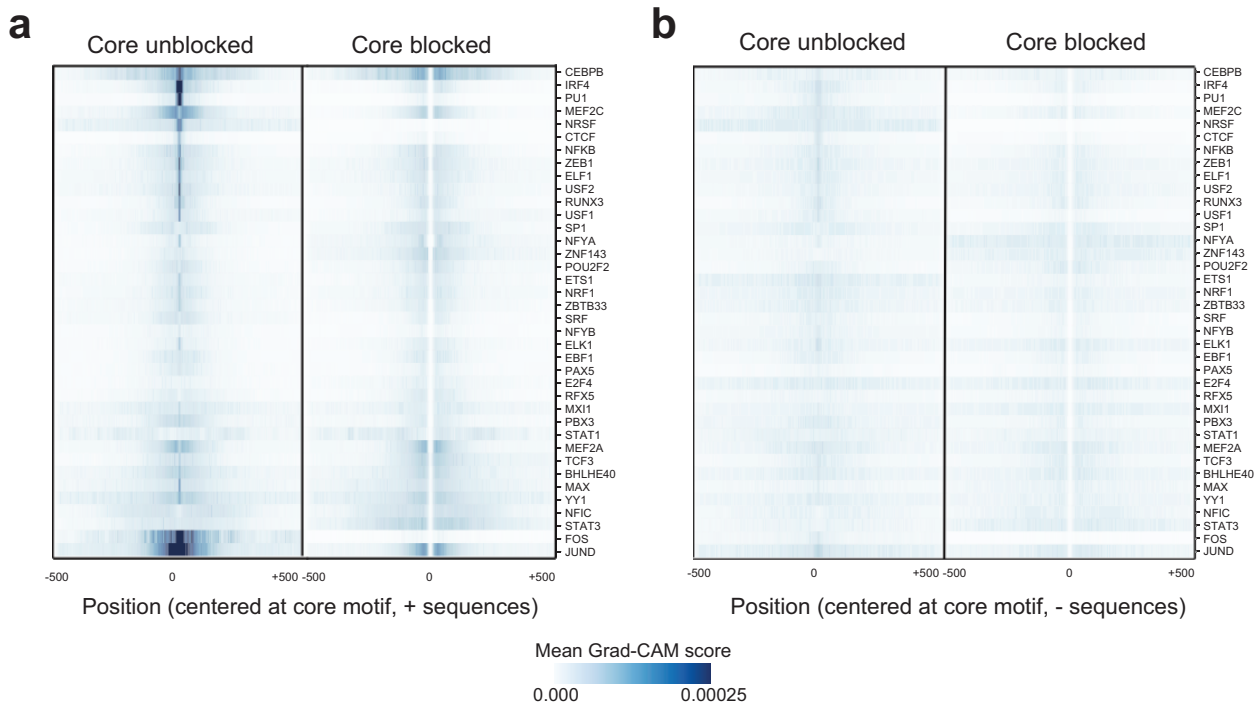
be integrated into similar frameworks to improve prioritization of disease-associated variants. (v) Finally, we mainly focused on the GM12878 cell type. While the results for STAT3 on multiple cell types indicate that important context bases are highly cell-type specific, future work is needed to further investigate other cell types.

Altogether, this study provides a valuable machine-learning framework for helping decode the rules by which TFs bind their target sites and identifying specific non-coding nucleotides with the strongest effects on binding. To facilitate future applications, Grad-CAM scores for all TF models studied here and code for running AgentBind on additional datasets are available at https://github.com/Pandaman-Ryan/AgentBind.

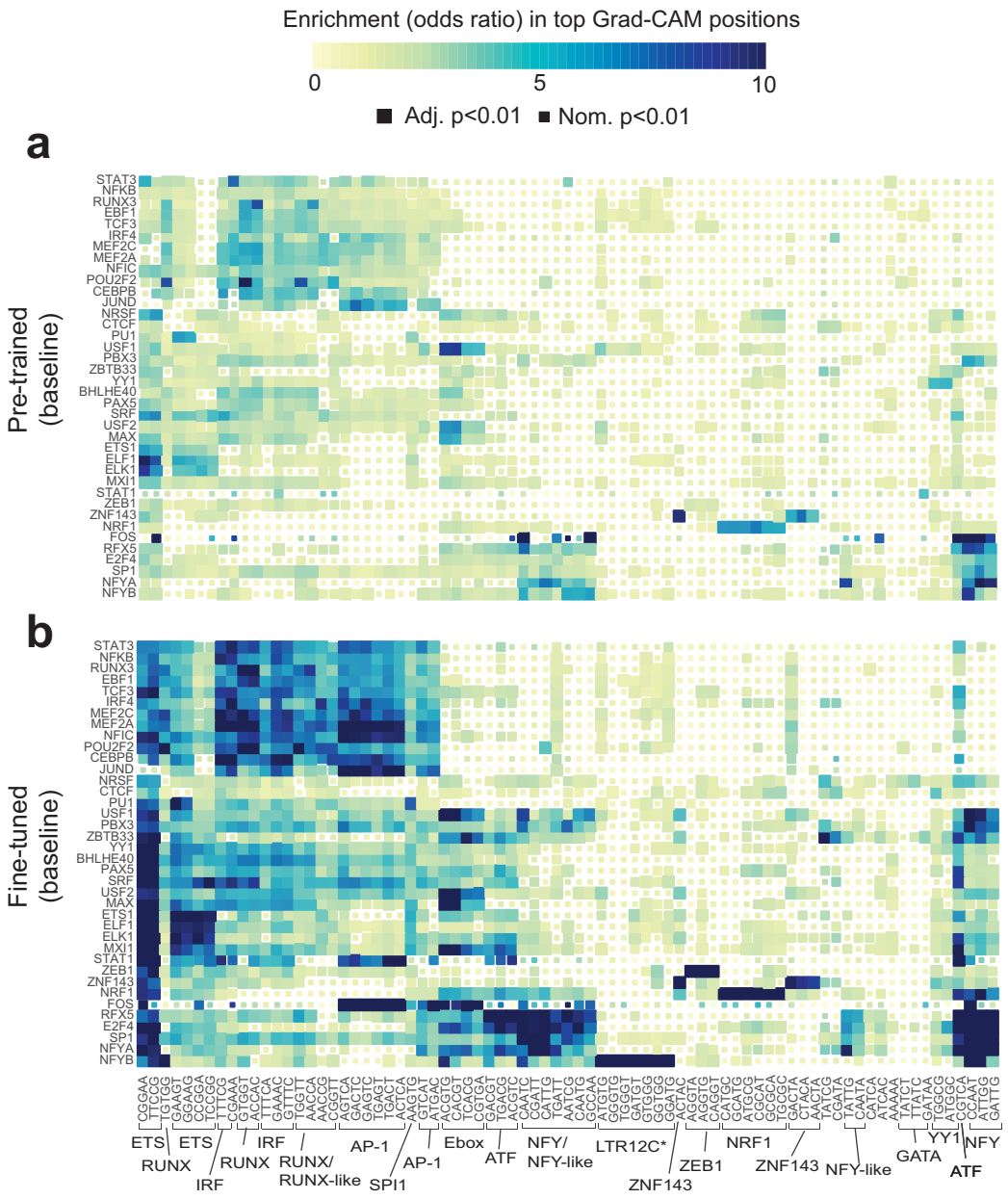
2.5 Supplementary figures

Supplementary Figure 2.1: Model performance related to GC content and open chromatin. (a-b) GC content differences correlated with model performance. The x-axis shows the absolute value of the difference in mean GC content for positive vs. negative sequences with the motif for each TF. The y-axis shows auROCs. Each dot represents one TF. Results in a-b are for baseline models with motifs blocked (a) or unblocked (b). (c) Comparison of training data size and change in model performance. The x-axis (log10 scale) shows the number of training samples. Orange points show the difference in auROC (y-axis) for baseline vs. GC-controlled models. Green points show GC-controlled vs. DNaseI-controlled models. Each dot represents one TF. (d) Model performance for each TF. The y-axis gives the auROC obtained for different models for each TF. Gray=baseline; orange=GC-controlled; green=DNaseI-controlled. TFs are ranked by the change in auROC between the DNaseI and GC-controlled models. (e) Comparison of cross-TF model performance. Heatmaps show the auROC using a GC-controlled model trained on one TF (rows) and tested on another TF (columns). Red squares denote the model with highest auROC for each TF. (f) Distribution of the difference in auROC between top models and TF-specific models. For TFs where the TF-specific model was best, we computed the difference between the TF-specific model and the next best model (red). For all other TFs, we compared performance of the best model to the TF-specific model (blue). (g-h) are the same as in e-f but based on DNaseI-controlled models.

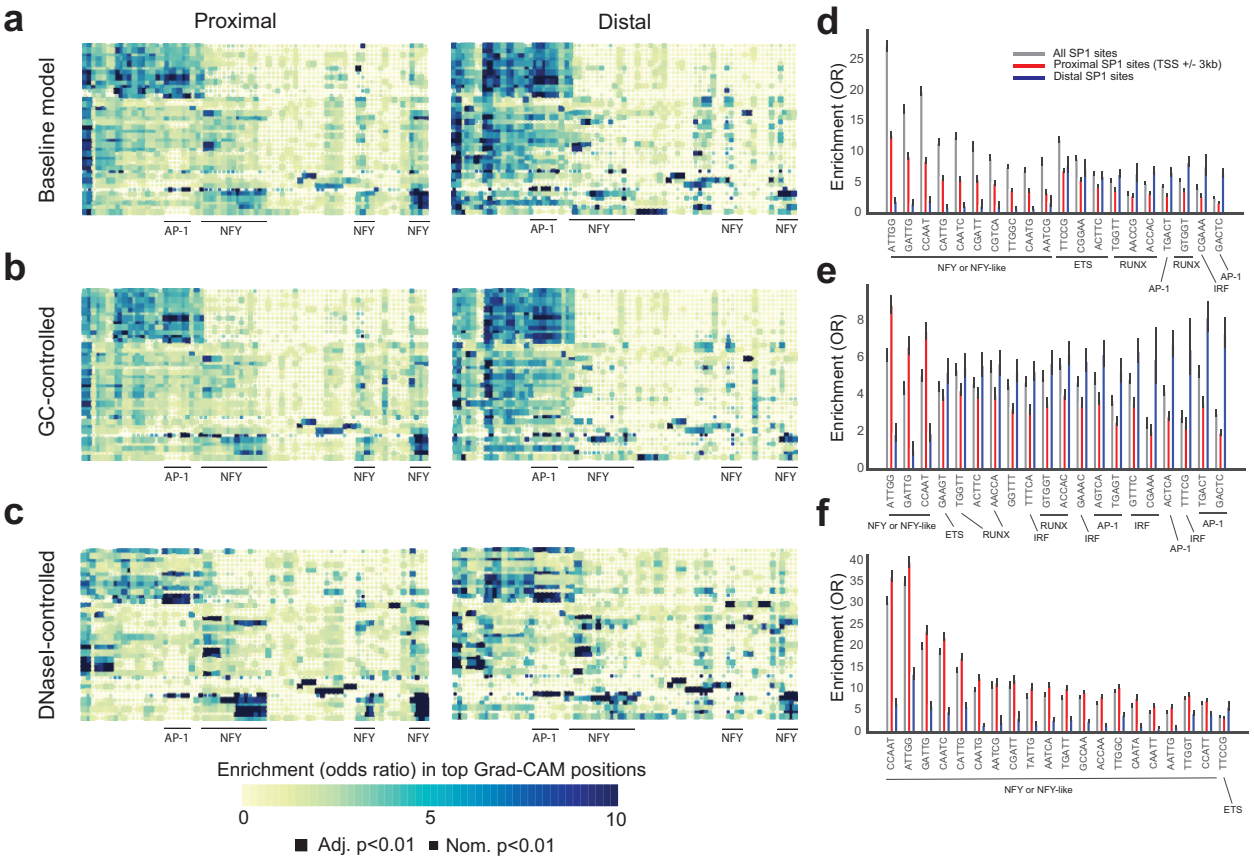

a
auROC (blocked)
|GC+-GC-|
b
auROC (unblocked)
|GC+-GC-|
c
GC vs. baseline
DNaseI vs. GC
r=-0.40; p=0.013
r=-0.49; p=0.0018
Δ auROC
Data size
d
auROC
Baseline model
GC-controlled
DNaseI-controlled
e
Model (GC-controlled)
Test dataset
auROC
best model
g
Model (DNaseI-controlled)
Test dataset
f
Best model -TF model
TF model - next best
Frequency
Δ auROC
h
Best model -TF model
TF model - next best
Frequency
Δ auROC



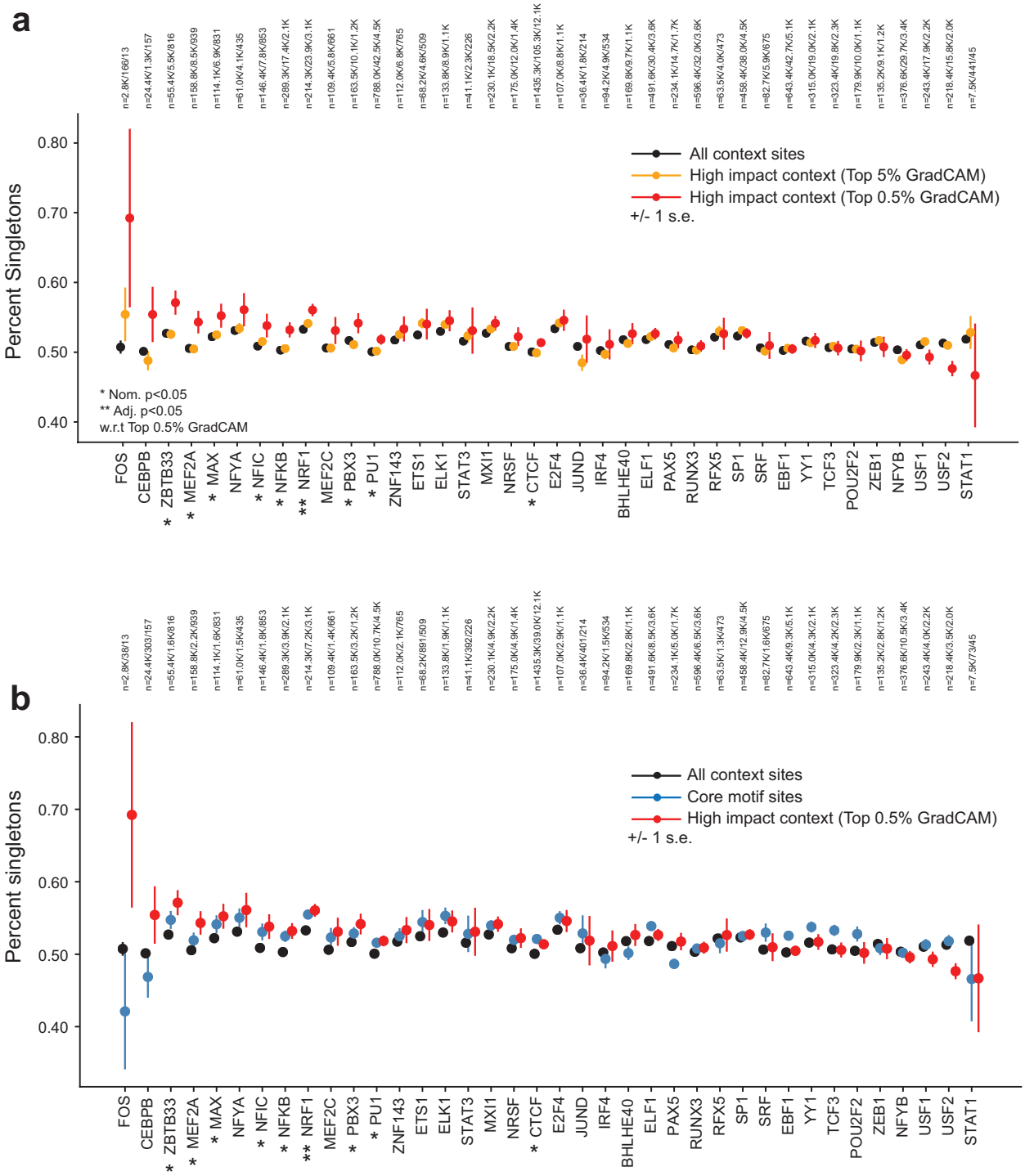
Supplementary Figure 2.2: Aggregate Grad-CAM score profiles for each TF. For each TF, we computed the average absolute value of the Grad-CAM score per position in positive sequences using either models with the core motif unblocked (left) or blocked (right). Values shown are Z-normalized across rows. (a) shows aggregate scores for sequences labeled as positive (bound) and is reproduced from Fig. 2d. (b) shows aggregate scores for sequences labeled as negative (unbound).



Supplementary Figure 2.3: Comparing key context sequence features identified in pre-trained vs. fine-tuned models. The heatmap shows the enrichment of each 5-mer in regions with the highest Grad-CAM scores for each TF using baseline models before (a) and after (b) fine-tuning. Rows and columns are ordered the same as in Fig. 3. Colors denote odds ratios, and the sizes of the boxes denote statistical significance as in Figure 2.3. Panel (b) is reproduced from Figure 2.3a for comparison



Supplementary Figure 2.4: Context sequence features specific to proximal vs. distal sites. (a) Enrichment of 5-mers in high-scoring Grad-CAM regions for proximal (left) and distal (right) binding sites. Proximal and distal TF binding sites are defined as described in Online Methods. Rows and columns are ordered the same as in Fig. 3. (b-c) are the same as in (a) but show data for GC-controlled (b) and DNaseI-controlled (c) models. For (a-c), colors denote odds ratios, and the sizes of the boxes denote statistical significance as in Figure 2.3. (d) Comparison of top scoring 5-mers in proximal vs. distal SP1 sites. Bars show the odds ratio of enrichment of each sequence in top 5-mers for all (gray), proximal (red) and distal (blue) SP1 sites. The top 20 5-mers ranked by the best odds ratio across all three SP1 models (all, proximal, and distal sites) are shown. Error bars show 95% confidence intervals on odds ratios. (e-f) are the same as in (d) but show data for GC-controlled (e) and DNaseI-controlled (f) models.



Supplementary Figure 2.5: Singleton rate of context SNPs vs. core motif regions. (a) Singleton rate of context SNPs. The plot shows the percent of SNPs in each category that are singletons. Black=all context sites, orange=context sites with top 5% Grad-CAM scores, red=context sites with top 0.5% Grad-CAM scores. Error bars show +/− 1 s.e. (b) is the same as (a), but additionally shows singleton rates for SNPs in core motif regions (blue). The number of SNPs in each category for each TF is annotated above each plot.

2.6 Acknowledgments

This study was supported in part by NIH/NHGRI 1R21HG010070-01, the Microsoft Genomics for Research program, and an Amazon Web Services research award. We thank NVIDIA for donating a Tesla K40 GPU to support this project. We additionally thank Christopher Benner and Alon Goren for helpful comments.

This chapter is a reformatted version of the material as it appears in "Deep neural networks identify sequence context features predictive of transcription factor binding," An Zheng, Michael Lamkin, Hanqing Zhao, Cynthia Wu, Hao Su, and Melissa Gymrek. The material has been published on Nature machine intelligence. The dissertation author was the primary investigator and author of this material.

CHAPTER 3

Deep learning predicts regulatory functions of variants in cell-type specific enhancers in brain

3.1 Introduction

Globally, there is nearly 1 in 6 world's population suffering from neurological and psychiatric disorders. Studies have found out that many of these brain disorders, including Alzheimer's Disease (AD), schizophrenia, and reduced intelligence, are highly inheritable and the majority of genetic variants associated with disease risk are found in transcriptional enhancer regions (Nord & West 2020; Li et al., 2018). However, unlike protein-coding regions, pathogenic mutations in enhancers are not directly involved in encoding protein sequences and are difficult to pinpoint and interpret. Moreover, the behaviors of transcriptional enhancers are highly specific to cell types, requiring epigenetic information to accurately characterize (Li, M. et al., 2018). The identification process for variant impacts is further complicated when we factor in linkage disequilibrium (LD), which results in blocks of variants being co-inherited and difficult to be differentiated in the genome-wide association studies (GWAS).

Recently, numerous fine-mapping techniques have been developed to prioritize putative causal variants from GWAS data. To analyze a genetic trait, these techniques first partition the human genome into subregions using the LD structures of the genome and then identify the SNPs most likely to be causal within each subregion. However, these techniques face several limitations: (1) it is difficult to find a causal SNP annotated with large probability when adjacent SNPs are highly correlated or the density of non-causal SNPs nearby is high (Schaid et al., 2018); (2) Due to enhancers' specificity to cell types, the pathogenic pathways of putative causal variants toward human traits are difficult to determine, even for the ones overlapping with enhancer regions.

Deep learning methods have been recently used for modeling non-coding DNA and pinpointing nucleotides predictive of regulatory functions, such as chromatin accessibility and transcription factor binding (Lai et al., 2021; Zheng et al., 2021, *Nature machine intelligence*; Corces et al., 2020; Avsec et al., 2021; Zhou, 2019, *Nature genetics*). Here, we make use of both deep learning and fine-mapping and develop a pipeline to prioritize genetic variants predicted to impact cell-type-specific enhancer activities in the brain and then identify candidate causal variants underlying association signals for a variety of brain-related traits (Figure 3.1a). Specifically, we first adapt our previously published AgentBind framework to capture features in sequences with strong H3K27ac signals and build separate models for four major brain cell types, including neurons, microglia, oligodendrocytes, and astrocytes. We additionally incorporate an improved model architecture including incorporation of spatial information (Liu et al., 2018) to boost model performance. Next, we apply Grad-CAM (Selvaraju et al., 2017) to compute importance scores at nucleotide-resolution and characterize sequence features predictive of H3K27ac activities. We find that variants predicted to have the highest impact on the H3K27ac signal are under stronger negative selection compared to low-impact variants and show a stronger allelic imbalance in the H3K27ac signal. Finally, we integrate our scores with fine-mapping results from GWAS of brain-related traits to identify putative causal variants that may act via modulating enhancer activity.

3.2 Results

3.2.1 Modeling brain cell-type specific H3K27ac signals

We obtained published H3K27ac ChIP-sequencing data for four brain cell types (microglia, neurons, oligodendrocytes, and astrocytes) (Nott et al., 2019). For each cell type, we collected genome sequences overlapping transcriptional enhancer regions (Method: chapter 3.3.1) and

acquired 12,074-21,415 non-overlapping H3K27Ac regions. Next, for each cell type, we constructed a binary dataset consisting of 1kb sequences centered at H3K27ac peaks (positive samples) and randomly chosen sequences with matched GC-content distributions (negative samples). We created multiple copies of each sample through window shifting (Method: chapter 3.3.1) to reduce model overfitting and to ensure model predictions are robust to the relative location of H3K27ac signals within each sequence.

We trained a separate model for each cell type. Our model training process consisted of two steps: pre-training and fine-tuning. Previous works (Zheng et al., 2021, *Nature machine intelligence*; Novakovsky et al., 2021) have found that pre-training could noticeably improve the performance of deep learning models in modeling genomic sequences, especially for small datasets. Similar to AgentBind, we first pre-trained our models using a large published dataset consisting of epigenomics profiles across 35 different cell types available from the DeepSEA project (Zhou et al., 2015). Next, for each brain cell type, we fine-tuned its model to predict the H3K27ac signal. Model performance was evaluated using the area under the receiver operating characteristic curve (auROC) and area under the precision-recall curve (auPRC). We left out sequences on chromosome 8 for cross-validation and sequences on chromosome 9 for testing.

We tested two different deep learning architectures: the DanQ (Quang et al., 2016) architecture used in AgentBind, and a version of ResNet (He et al., 2016) modified from that used in ChromDragoNN (Nair et al., 2019). The ResNet architecture consisted of 5 convolutional layers followed by 8 residual blocks and 2 fully connected layers (Method: chapter 3.3.2). The output of both models was a single number ranging between 0 to 1, indicating how likely the input sequence contained a H3K27ac peak. The more complex ResNet architecture allowed better performance in

modeling H3K27ac samples, resulting in an average auROC increase of 0.023 and auPRC increase of 0.025 (Figure 3.1b-e).

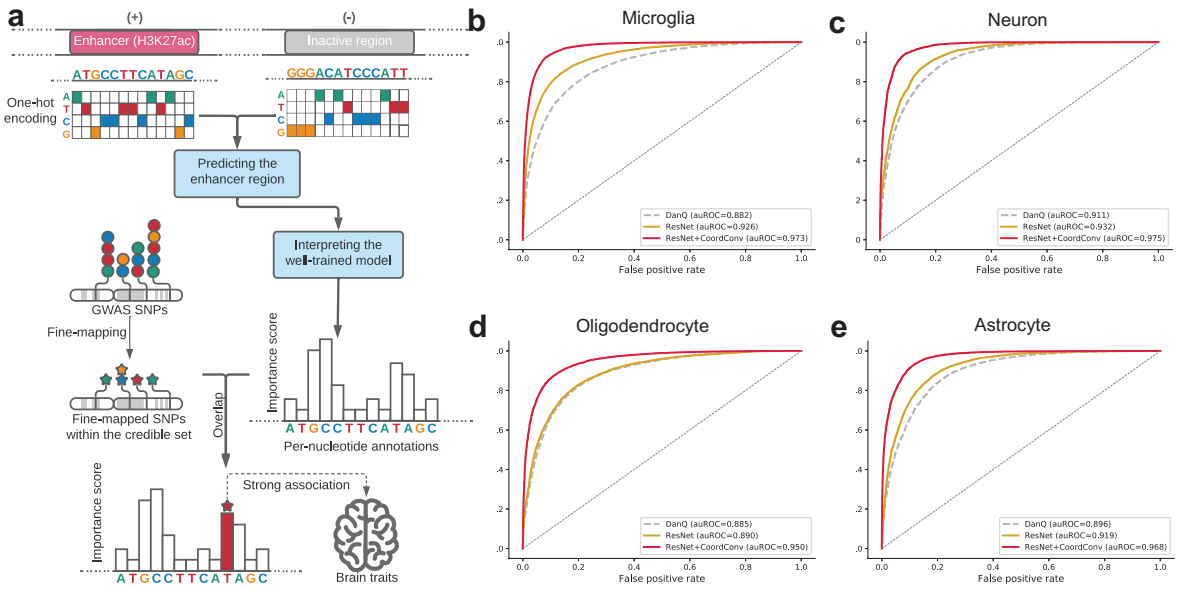


Figure 3.1: Pipeline Overview. (a) Method schematic. We construct a ResNet model for each type of brain cells and capture sequence features unique to the transcriptional enhancers in this type of cells. Next, we use Grad-CAM to score the contribution of each nucleotide to enhancer activities. On the other hand, we fine-map the SNPs from the GWAS file of each brain trait and identified the SNPs most likely to be causal. These putative causal SNPs were then overlapped with the importance scores from Grad-CAM. For each cell type and its associated brain trait, we acquire a list of SNPs likely to result in the trait through impacting enhancer activities. (b) - (e) CoordConv and ResNet improve model performance in (b) microglia, (c) neurons, (d) oligodendrocytes, (e) astrocytes. In each plot, receiver operator curves (ROC) are shown for H3K27ac predictions using DanQ (grey dashed line), ResNet (gold solid line), and ResNet + CoordConv (red solid line).

To further improve our model, we applied the CoordConv (Liu et al., 2018) technique which adds an extra coordinate channel to the input of the first convolutional layer to better encode spatial information in the input sequences (Method: chapter 3.3.2). We found this resulted in a notable performance boost, with an average auROC and auPRC increase of 0.050 and 0.051, respectively. Overall, our final models could predict H3K27ac signal with high accuracy (mean auROC=0.966, mean auPRC=0.967; Figure 3.1b-e). Full results are reported in Table 3.1.

Table 3.1: classification performance.

	w/o CoordConv				w/ CoordConv	
Model architecture	DanQ		ResNet		ResNet	
Metrics	auROC	auPRC	auROC	auPRC	auROC	auPRC
Microglia	0.882	0.877	0.926	0.921	0.973	0.969
Neuron	0.911	0.895	0.932	0.922	0.975	0.970
Oligodendrocyte	0.885	0.901	0.890	0.908	0.950	0.960
Astrocyte	0.896	0.891	0.919	0.912	0.968	0.968
Average	0.893	0.891	0.917	0.916	0.966	0.967

3.2.2 Identifying sequence features predictive of H3K27ac signal

We next applied Grad-CAM (Selvaraju et al., 2017), a model interpretation technique previously used to infer genomic sequence features learned by deep learning models (Zheng et al., 2021, *Nature machine intelligence*), to characterize key sequence features learned by our models. We used Grad-CAM to assign nucleotide-level scores to quantifying the importance of each base pair in predicting H3K27ac signal (Method: chapter 3.3.3). An example score profile for a single sequence is shown in Figure 3.2a. In this example, Grad-CAM scores highlight a short DNA sequence that matches with the known PU.1 motif from the JASPAR database (Castro-Mondragon et al., 2022) as being critical in predicting the enhancer activity of this locus in microglia.

Next, we sought to use importance score profiles to identify features most predictive of H3K27ac status for each cell type. To this end, we applied two strategies. First, we extracted 6-mers from positive sequences and tested whether each unique 6-mer is enriched within the 6-mers with the highest importance scores (Method chapter 3.3.4; Figure 3.2b). To associate 6-mers with known motifs, we used TOMTOM (Gupta et al., 2007), which is a computational tool that can align a given sequence with the known motifs. We saw substantial specificity of 6-mer enrichment in different cell types. For example, the most predictive 6-mers for microglia are associated with ETS, IRF, CEBP, and MEF2 motifs, which correspond to well documented transcription factors

important for microglia phenotype and function (Gosselin et al., 2017; Masuda et al., 2012; Holtman et al., 2017). There were also 6-mers shared by multiple cell types, as exemplified by those associated with NFI motif enriched in astrocytes, neurons, and oligodendrocytes. Different members of NFI family have been reported to play an important role in the development of these cell types in both mouse and human (Chen et al., 2017; Wilczynska et al., 2009).

We additionally applied TF-MoDISco (Shrikumar et al., 2018), which leverages per-base importance scores to infer enriched motifs. The motifs identified from TF-MoDISco were consistent with those inferred from enriched 6-mers (Figure 3.2C). Since this approach exceeds the length limitation of the k-mer approach, we were able to uncover additional motifs associated with high importance scores, as exemplified by an additional RUNX motif for microglia, consistent with the established role of Runx1 (Zusso et al., 2012).

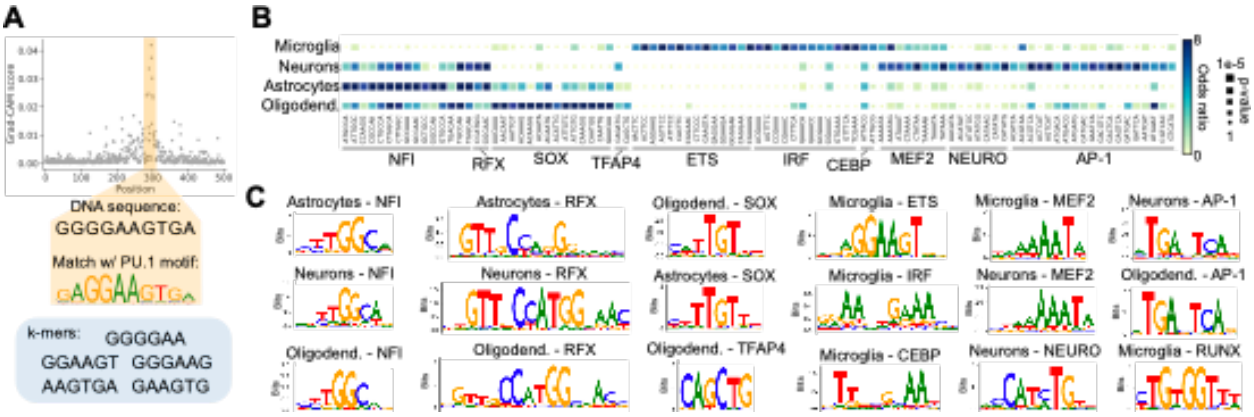


Figure 3.2: Interpreting H3K27ac regions and identifying enriched motifs. (a) Example of importance scores for a region (chr1:12289432–12290431 in hg19) containing an SP1 motif. The y-axis shows the importance score of each nucleotide based on the ResNet+Coordconv model. In this example, a short DNA sequence that matches with the known PU.1 motif is annotated with high scores. (b) Enrichment of 6-mers in the most influential context regions. The heatmap shows the enrichment of each 6-mer in regions with the highest importance scores for each cell type. (c) Enriched motifs highlighted by TF-MoDISco for four brain cells.

3.2.3 High-scoring variants show biological significance

To evaluate our variant-level scores, we next checked whether the predicted importance of a variant correlates with its biological impact. We first examined allelic imbalance based on available ATAC-seq reads for microglia (Method: chapter 3.3.6). Briefly, an imbalance of reads from each allele at a heterozygous single nucleotide polymorphism (SNP) indicates a bias in regulatory activity between the two genome copies. We found that variant-level importance scores computed based on microglia models are correlated with allelic imbalance scores (Pearson r2=0.760; two-sided $p=6.70*10^{-3}$; Figure 3.3a), whereas imbalance scores computed for other cell types show no correlation with the allelic imbalance p-values for microglia. SNPs with low allelic imbalance p-values are strongly enriched with high importance scores (Method: chapter 3.3.6; two-sided fisher exact test $p=2.48*10^{-22}$, Odds ratio=2.80). These results indicate that our scores are indeed predictive of enhancer activity.

We next tested whether variants predicted to have a high impact on H3K27ac in the brain show signals of nature selection. We hypothesized that variants with high impact on brain regulatory activity would tend to be deleterious and thus kept at low frequencies in the population. Indeed, we found that in all cell types, rare variants ($MAF<10^{-4}$ based on gnomAD; Method: chapter 3.3.5) have higher average importance scores (Figure 3.3b) and the importance scores follow a downward trend with the increase of MAF. We additional examined the percentage of variants in different Grad-CAM score bins that are singletons, meaning the variant has only been observed in a single individual. This "percent singletons" has been previously used as a proxy for the deleteriousness of different variant categories11. Variants with top-scoring importance scores (top 5% of Grad-CAM scores) show significantly higher singleton percentages (Two-sided $p = 6.04*10^{-27}$; Figure 3.3c) for all cell types. This trend is further pronounced when restricting to the top 0.5% of high-scoring variants ($p=3.48*10^{-7}$). Taken together, these results suggest that variants

with high impacts on brain enhancer activity are deleterious and are likely targeted by purifying selection.

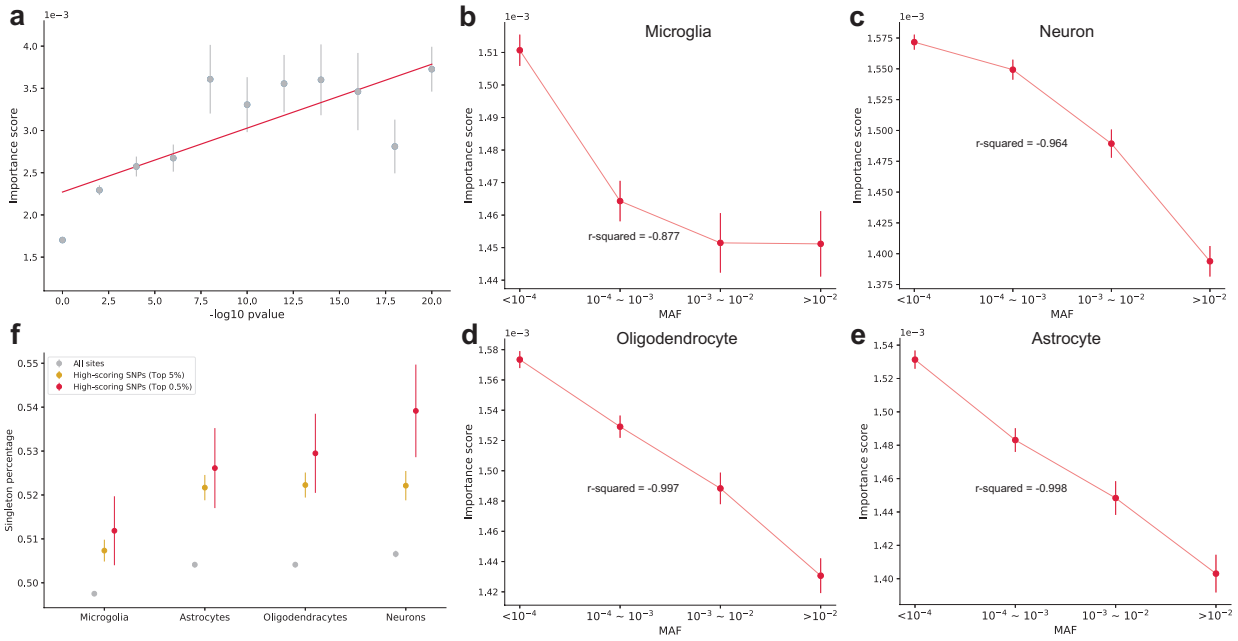


Figure 3.3: Investigating the biological significance of importance scores. In all figures, dots show the average importance scores and error bars show +/− 1 standard errors. (a) The relationship between allelic imbalance p-values and importance scores. The linear regression result is shown with the red line. (b) - (c) The relationship between MAF (based on the control samples in gnomAD v2.1.1) and importance scores in (b) microglia, (c) neurons, (d) oligodendrocytes, and (e) astrocytes. Pearson r2 values measuring the linear relationships are annotated in plots. (f) The plot shows the percent of SNPs in each category that are singletons. Grey=all sites, gold=positions with top 5% importance scores, red=positions with top 0.5% importance scores.

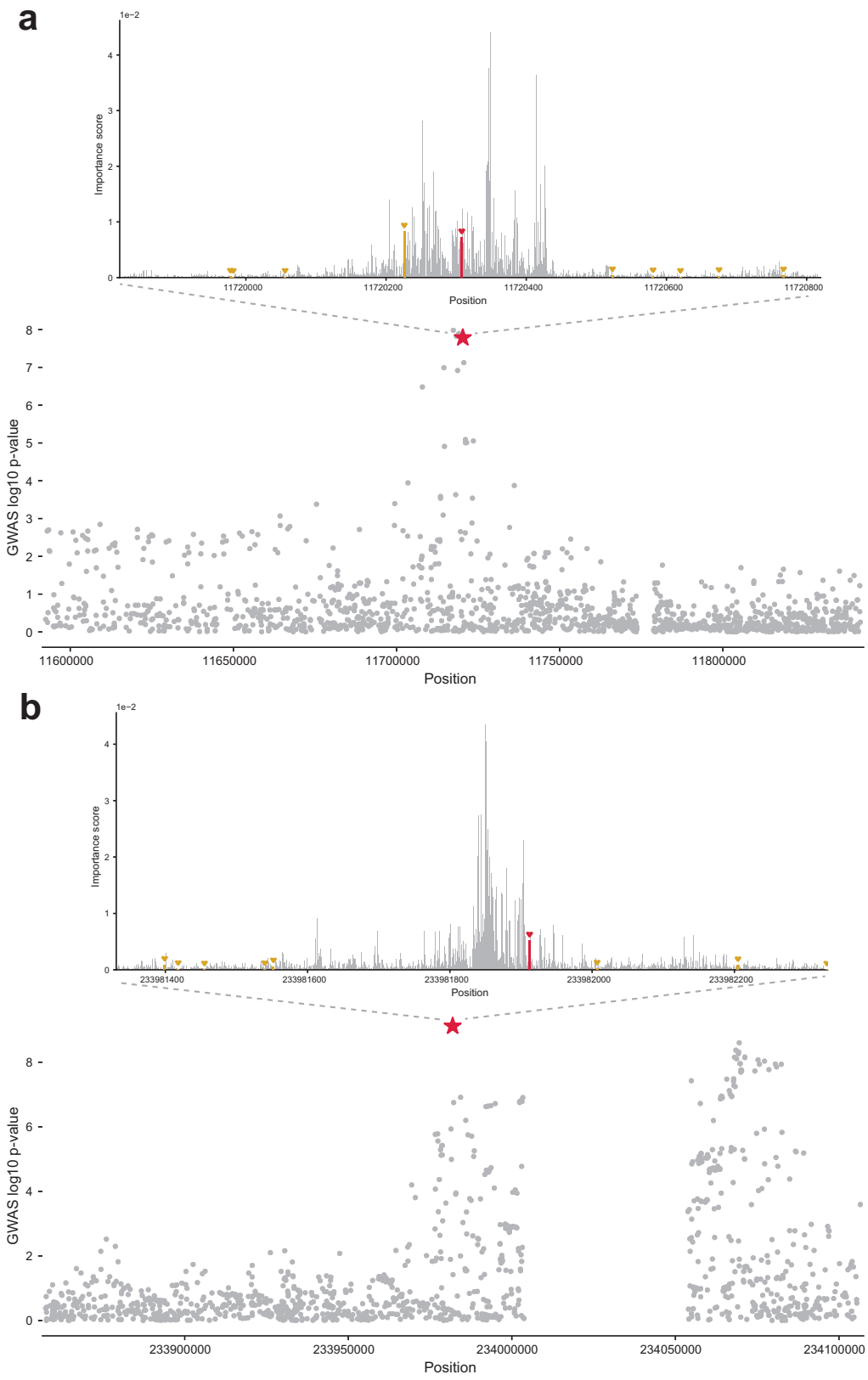
3.2.4 Linking high-scoring variants with brain traits and disorders

Previous studies have demonstrated an enrichment between cell-type specific brain enhancers and various neurological and psychiatric disorders3. To investigate whether variants predicted to disrupt enhancer activity might contribute to brain-related complex traits, we analyzed GWAS summary statistics for 8 traits and disorders (Alzheimer's disease (AD), schizophrenia, major depressive disorder (MDD), bipolar disorder (BD), autism spectrum disorder (ASD), intelligence, risky behaviors, and insomnia; Method: chapter 3.3.7). We first applied FINEMAP

12 to identify candidate causal variants for each trait (leniently defined as inclusion in 95% credible sets for each locus and posterior inclusion probability [PIP]>1%). Notably, we did not apply annotation-based fine-mapping tools since our AgentBind annotations score only a small subset of variants and are thus expected to result in unreliable results using these methods. Overall, we found only a minority (0.744% variants on average) of GWAS variants are overlapped with our H3K27ac datasets and this number continue to drop to 0.194% for variants within the 95% credible sets.

We identified candidate causal variants with high impact on H3K27ac (top 20% of scores) predicted by Grad-CAM for the relevant cell type for each trait as predicted by Nott et al (Nott et al., 2019). Several SNPs we found to be associated with AD, including rs7920721 and rs10933431, were also investigated by other studies. The importance score of SNP rs7920721 was higher than 97.2% of its neighbors and with 14.2% probability to be a causal variant to AD (Figure 3.4a). This variant is adjacent to ECHDC3, a gene whose expression was altered in AD brains compared with controls (Desikan et al., 2015). And studies showed that rs7920721 is significantly associated with increased risk for AD in genome-wide through proxy case-control analysis and meta-analysis (Desikan et al., 2015; Witoelar et al., 2018; Liu et al., 2017; Efthymiou & Goate, 2017; Yin et al., 2019). Moreover, a large transethnic GWAS study revealed that the association between this variant and AD is exclusive to humans lacking APOE ε4 alleles (Jun et al., 2017). Another SNP rs10933431 (Figure 3.4b) was also found genome-wide significant in meta-analyses (Kunkle et al., 2019; Marioni et al., 2018; Lambert et al., 2013). This variant regulates INPP5D, a gene associated with altered CSF pTau levels, which is a biomarker determining the pathologic process of AD (Tan et al., 2021).

Figure 3.4: Examples of putative casual SNPs (a) chr10:11720308; rs7920721 and (b) chr2:233981912, rs10933431 found in this study. In each Manhattan plot, the putative SNP is annotated with a red star. In each importance score distribution plot, the putative SNP is colored as red and surrounding variants not passing our filters are colored in gold.


a
1e−2
Importance score
11720000
11720200
11720400
11720600
11720800
Position
GWAS log10 p-value
11600000
11650000
11700000
11750000
11800000
Position
b
1e−2
Importance score
233981400
233981600
233981800
233982000
233982200
Position
GWAS log10 p-value
233900000
233950000
234000000
234050000
234100000
Position

For the brain trait of intelligence, there were a couple SNPs we identified that were also previously studied. We found SNP rs4500960 (importance score ranked at 91.0% and PIP=12.2%) is significantly associated with educational attainment (Kong et al., 2017). Studies also discovered that this SNP involves in not only developing intellectual disability, but also other brain disorders including schizophrenia (Bansal et al., 2018), risk for alcohol dependence (Rosoff et al., 2021), ASD (Okbay et al., 2016), and general cognitive functions (Davies et al., 2018). Okbay et al. (Okbay et al., 2016) found that this variant is an intronic variant in TBR1, an important gene for differentiation and migration of neurons during brain development. Another SNP rs61786697 (importance score ranked at 95.8% and PIP=2.1%) is also associated with intelligence. Studies (Hauberg et al., 2016; Brum et al., 2021) found this variant changes a promoter for MIR317, a conserved and high confidence miRNA linked to a wide range of brain disorders.

Two SNPs in Supplementary Table 4 are also lead SNPs in GWAS studies for schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014): rs324017 (importance score ranked at 87.8% and PIP=36.8%) and rs2071407 (importance score ranked at 90.5% and PIP=2.1%). Moreover, rs324017 is shown to be associated with the expression of LRP1, a gene that strongly implicate in blood-brain barrier function and in the etiopathology of developmental disorders, including schizophrenia (Torrico et al., 2019; Pong et al., 2020). This variant is one of the lead SNPs in the GWAS study of insomnia (Lane et al., 2019).

3.3 Method details

3.3.1 Brain cells training data

We obtained ATAC-seq and H3K27ac ChIP-seq data from previous literature (Nott et al., 2019) for four brain cell types: microglia, neurons, astrocytes, and oligodendrocytes. We mapped

these data to the hg19 genome using Bowtie2 v2.3.5 with default parameters (Langmead and Salzberg, 2012). Since every cell type has at least three biological replicates of different individuals, we first called unfiltered ATAC-seq peaks for each replicate using findPeaks script of HOMER v4.11.1 with parameters "-style factor -L 0 -C 0 -fdr 0.9 -size 200" (Heinz et al., 2010). We then used IDR v2.0.3 with a threshold at 0.05 (Li et al., 2011) to identify reproducible open chromatin regions. Since IDR works with only two replicates at a time, we applied to each pair of replicates of the same cell type and merge the reproducible peaks of each pair of replicates using mergePeaks script of HOMER with the parameter "-d 200" to reach a final set of reproducible ATAC-seq peaks for every cell type. Based on the reproducible ATAC-seq peaks, we computed the normalized number of H3K27ac ChIP-seq tags in an expanded region of 1,000 bp and added genomic annotations to these regions using annotatePeaks.pl script of HOMER with parameters "-norm 1e7 -size -500,500". The normalized tag counts were averaged across replicates of the same cell type. We eventually selected a high-confidence set of enhancers for each brain cell type by restricting ATAC-seq peaks to be within intronic or intergenic regions based on HOMER annotations and, in the meantime, associated with more than 20 averaged, normalized tags of H3K27ac. Our processing step resulted in 21,415 enhancers for microglia, 12,074 enhancers for neurons, 15,774 enhancers for astrocytes, 16,034 enhancers for oligodendrocytes. The sequences of these enhancers were positive sequences used in our model training. Our negative sequences were generated with matching repeat and GC content as the positive sequences using the "genNullSeqs" function of gkmSVM v0.81 (Ghandi et al., 2016).

To avoid the bias of model training towards open chromatin regions and the ignorance of wider contexts, we created 11 copies for each positive and negative sequence with equally distanced window shifts with a gap of 100 bp. All copies included the core H3K37ac regions but

with different amounts of context areas included upstream and downstream. Positive and negative sequences were then one-hot encoded into 1,000 by 4 matrices based on the sequence present in the hg19 reference genome. Nucleotides marked as "N" were converted to vectors with entries of 0.25 for each of the four nucleotides.

3.3.2 ResNet model architecture

Our ResNet architecture consisted of 5 standalone convolutional layers, 8 residual blocks, and 2 fully connected layers. The standalone convolutional layers have kernels with sizes ranging between 1 to 5 and the number of channels ranging between 64 to 256. The standalone convolutional layers have small kernels: the kernel size of the first two layers is 5, with 128 channels; the kernel size of the next two layers is 3, with 256 channels. The fifth layer is used for dimensionality reduction and has a kernel size of 1 with 64 channels. These convolutional layers are used for extracting basic sequence features such as motifs and motif combinations. The 8 residual blocks were constructed the same as in ChromDragoNN (Nair et al. 2019), with a standalone convolutional layer after every two blocks. Batch normalization layers were used after all convolutional layers. The final two layers were fully connected layers with 1000 neurons each. This model takes one-hot encoded 1000 bp sequences as input and outputs a number ranging between 0 to 1 for each of them to indicate whether it is predicted to contain and H3K27ac signal.

Previous works (Liu et al., 2019; El Jurdi et al., 2021; Zhu et al., 2021) have shown that adding into convolutional layers hard-coded channels for the data coordinates can improve their transitional invariance property and boost their modeling performance in pattern localization and object detection tasks. In our ResNet model, we modified the first convolutional layer into a CoordConv layer. The coordinates were defined as the distance from the center of H3K27ac

regions, with upstream nucleotides labeled as negative and downstream labeled as positive. These coordinates were then re-scaled to range from -1 to 1.

3.3.3 Model interpretation

We implemented the Grad-CAM method (Selvaraju et al. 2017) to interpret our ResNet model by computing an individual score for each nucleotide of the input sequence which indicates its importance in determining the model's prediction. In our implementation of Grad-CAM, we chose the second to the last convolutional layer prior to the ResNet blocks as the layer of interest. The receptive field of neurons is 13 in the feature maps of this layer, with enough length to cover the cores of most of the common transcription factor motifs. Following the weighting method proposed in the Grad-CAM method, we calculated the weight of each feature map in this layer and used these weights to compute a weighted combination of feature map activations. This gave us a coarse importance map for the input sequence. To acquire a finer resolution at the base-pair level, we mapped this coarse importance map onto the input sequence and multiplied it with input gradients elementwise. We define the scores in the resulting finer resolution map as importance scores which are used in downstream analyses.

3.3.4 Sequence feature analysis

To identify important sequence features, we segmented positive H3K27ac sequences from each cell type into 6-bp sequences (6-mers) using a sliding window. We computed the average importance score of each 6-mer and ranked all 6-mers based on this score. We defined top-scoring sequences as those with the top 1% of scores. Similar to in AgentBind, we then performed a Fisher's Exact Test for each 6-mer to test whether it is enriched in the top-scoring subsequences

for that cell type. Tests were performed using the fisher_exact method from the Python scipy.stats library (https://docs.scipy.org/doc/scipy/reference/stats.html). The significantly enriched 6-mers were aligned with known motifs from JASPAR database (Castro-Mondragon et al., 2022) using TOMTOM v5.1.1 (Gupta et al., 2007) to infer most likely motifs associated with every 6-mer.

We additionally used TF-MoDISco v0.5.16.0 (Shrikumar et al., 2018) to cluster and aggregate the importance scores and recover motifs occurring in the H3K27Ac regions. The core 500 bp of each H3K27ac region and its importance scores were used as input. We use its built-in LaplaceNullDist function to generate null distributions with a sampling size of 10000. To enable TF-MoDISco to find longer motifs, we set the trim_to_window_size as 500 in its seqlets-to-patterns factory.

3.3.5 Analysis of variant allele frequencies

We obtained SNP allele frequencies computed across 5,192 control samples from gnomAD (Karczewski et al. 2020) v2.1.1. For each cell type, we collected the minor allele frequencies (MAFs) for the gnomAD SNPs that were also scored in our H3K27ac dataset. We defined singletons as SNPs whose total allele counts in gnomAD were at least 1,000 and for which the alternate allele was observed only once. The singleton ratio of a set of SNPs is defined as the percentage of gnomAD SNPs in this set that are singletons.

3.3.6 Allelic imbalance analysis

We combined the original ATAC-seq data of four different individuals (Nott et al., 2019) with additional twelve ATAC-seq data of ex vivo microglia obtained from previous literature (Gosselin et al., 2017). We first masked hg19 genome with “N” at positions tested by Alzheimer’s

disease GWAS (Jansen et al., 2019) and re-mapped all the sixteen datasets to this masked genome using Bowtie2 with parameter "--np 0" meaning no penalty for "N" (Langmead and Salzberg, 2012). Then we counted the number of reads with different alleles at each position using the mpileup tool from samtools v0.1.15 (Danecek et al., 2021) followed by the mpileup2snp function of VarScan v2.4.3 (Koboldt et al., 2012). Read counts for the reference and variant allele at each masked position were compared by a binomial test to identify significant allelic imbalance.

In the fisher exact test, we defined the SNPs with top 5% scores (threshold=$5.84*10^{-3}$) as high-scoring variants and SNPs with allelic imbalance p-value lower than 10-10 as the highly imbalanced. We used these two thresholds to divide our dataset into four groups and recorded the number of SNPs in each group. We used these numbers for the two-sided fisher exact test.

3.3.7 Fine-mapping published GWAS signals

We used FINEMAP (Benner et al. 2016) v1.4 to fine-map variants in each locus. Linkage disequilibrium (LD) for each pair of input variants was computed using the LD computing script "CalcLD_1KG_VCF.py" from PAINTOR v3.0 (Kichaev et al, 2017) based on available genotypes from the 1000 Genomes Project phase 3 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). GWAS summary statistics files we used are listed in Summplementary Table 4.3. GWAS loci were acquired from their original GWAS studies. For studies only providing lead SNPs instead of a range, we defined a GWAS locus as a window of 250,000bp with a lead SNP in the center. FINEMAP was run with default parameters allowing up to 5 causal SNPs per locus.

3.4 Acknowledgements

This chapter, in full, is currently being prepared for submission for publication of the material by Zheng, A., Shen, Z., Glass, C., and Gymrek, M. The dissertation author was a primary researcher and author of this material.

CHAPTER 4

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

4.1 Introduction

Noncoding genetic variation is a major driver of phenotypic diversity and the risk of a broad spectrum of diseases. For example, of the common single-nucleotide polymorphisms and short insertions/deletions identified by genome-wide association studies to be linked to specific traits or diseases, ~90% are typically found to reside in noncoding 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 these 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 (Vierstra et al., 2020; Kilpinen et al., 2013; van der Veeken et al., 2019). Despite these major advances, it remains difficult to predict the consequences of most forms of noncoding 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 (Encode Project Consortium 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 (Deplancke et al., 2016; Grossman et al., 2017; Behera et al., 2018).

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 (Fairfax et al., 2014; Bakker et al., 2018; 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 noncoding 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 interleukin-4 (IL-4) (Figure 4.1A). The selected strains include both similar and 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 et al., 2018, *Cell*). 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 et al., 2010). The immediate transcriptional response to IL-4 is mediated by activation of signal transducer and activator of transcription 6 (STAT6) (Ostuni et al., 2013; Goenka et al., 2011), which rapidly induces the expression of direct target genes that include effector proteins such as Arginase 1 (Arg1) and transcription factors like peroxisome proliferator–activated receptor γ (PPARγ) (Huang et al., 1999; Daniel et al., 2018) and early growth response 2 (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. 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.

4.2 Results

4.2.1 The response to IL-4 is highly variable in BMDMs from genetically diverse mouse strains

To investigate how natural genetic variation affects the macrophage response to IL-4, we began by performing RNA sequencing (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. We therefore focused our analysis on the response to IL-4 in BMDMs from the five strains at this time point. Weighted gene co-expression network analysis (WGCNA; Langfelder et al., 2008) identified numerous modules of highly correlated mRNAs, most of which were driven by strain differences (Figure 4.1B). 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 (Figure 4.1B).

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 (Figure 4.1D, Supplementary Figure 4.1C and 4.1D). 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 (Figure 4.1B, red module, and Figure 4.1C, second panel). A similar pattern was observed for down-regulated genes (Figure 4.1D). Despite these differences at the level of individual genes, similar pathways/gene programs were enriched in all strains for both induced and repressed genes (Figure 4.1E). Substantial differences in IL-4 target gene expression across strains are illustrated by Arg1, Slc7a2, and Msx3 (Figure 4.1F). 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 twofold 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 pairwise comparisons indicate that 29% of the genes with decreased IL-4-induced gene expression were due to low basal expression, 36% had no differences before IL-4 stimulation (equal basal), and 35% were the result of a high basal expression level in the less responsive strain (Figure 4.1G).

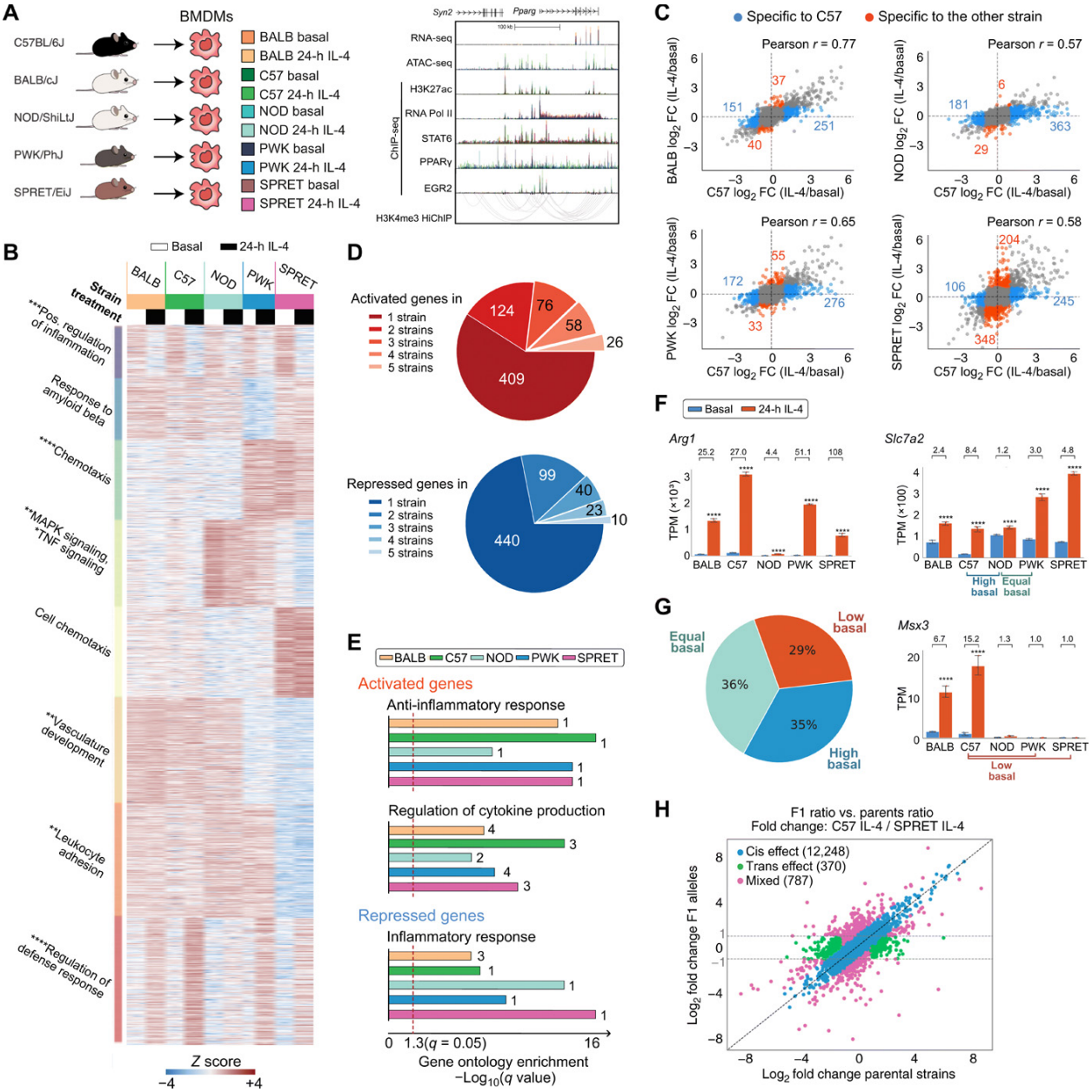


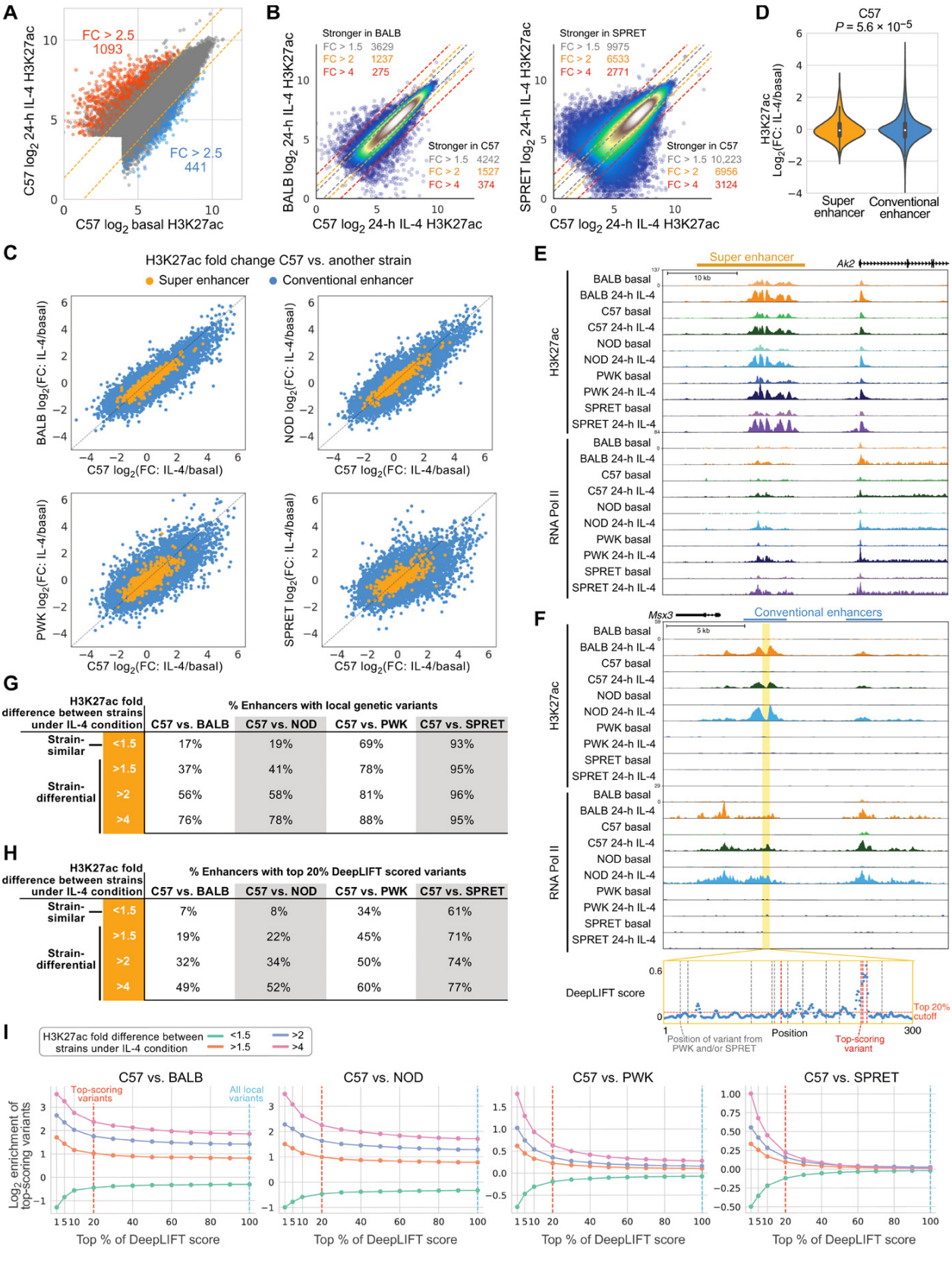
Figure 4.1: Response to IL-4 is highly divergent in bone marrow–derived macrophages from different mouse strains. (A) Overview of experimental design and main datasets. (B) WGCNA clustering focused on strain-differentially regulated genes in IL-4-treated bone marrow-derived macrophages (BMDMs). The top hit Metascape pathways are annotated for each module. *q < 0.05, **q < 0.01, ***q < 0.001, ****q < 0.0001. MAPK, mitogen-activated protein kinase; TNF, tumor necrosis factor. (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-hour 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 (C57 × SPRET F1) and parental strain under 24-hour IL-4 conditions.

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. A total of 91.4% of parental-specific RNA-seq reads in the F1 strain are within twofold of their values in C57 and SPRET (blue data points) and considered to be due to local (cis) effects of genetic variation (Figure 4.1H), while only 2.8% were 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 C57 × NOD macrophages. RNA-seq on IL-4-stimulated macrophages of F1 C57 × NOD macrophages showed strong convergence of expression of genes that were differentially regulated in the parental strain, consistently a major contribution of trans regulation. To investigate the point at which this regulation occurs, we performed chromatin immunoprecipitation sequencing (ChIP-seq) for RNA polymerase II (Pol II) under control and IL-4-stimulated conditions. In contrast to mRNA levels, examination of the IL-4-dependent changes in gene body RNA Pol II indicated similar magnitude changes in all strains, including NOD. These results suggest the presence of a transacting factor in NOD that acts downstream of transcription to attenuate mRNA levels. Collectively, these studies uncovered notable 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 affects signal-dependent gene expression.

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

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 assay for transposase-accessible chromatin using sequencing (ATAC-seq) with at least 2.5-fold increase in H3K27ac (Creyghton et al., 2010) and RNA Pol II (Bonn et al., 2012) after IL-4 treatment (Supplementary Figure 4.1, A to D). In 24-hour IL-4–stimulated C57 BMDMs, 1093 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 (Figure 4.2A). 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 (Figure 4.2B and Supplementary Figure 4.1, E and F). We further subdivided these regions into "conventional enhancers" (Figure 4.2C, blue) and "super enhancers" (Figure 4.2C, orange), on the basis of 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 coregulators 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 (Figure 4.2D and Supplementary Figure 4.1G). For example, IL-4 induction of the Ak2 super enhancer (Figure 4.2E) 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 (Figure 4.2F).

Figure 4.2: 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. FC, fold change. (B) Comparison of H3K27ac signal between C57 and BALB or SPRET under the 24-hour IL-4 condition. (C) Log2 H3K27ac fold changes after 24-hour 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. Dashed lines represent locations of PWK or SPRET variants. (G) and (H) Enhancers were categorized into strain-similar and strain-differential on the basis of 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
1093
FC > 2.5
441
C57 log2 24-h IL-4 H3K27ac
C57 log2 basal H3K27ac
B
Stronger in BALB
FC > 1.5 3629
FC > 2 1237
FC > 4 275
Stronger in C57
FC > 1.5 4242
FC > 2 1527
FC > 4 374
BALB log2 24-h IL-4 H3K27ac
C57 log2 24-h IL-4 H3K27ac
Stronger in SPRET
FC > 1.5 9975
FC > 2 6533
FC > 4 2771
Stronger in C57
FC > 1.5 10,223
FC > 2 6956
FC > 4 3124
SPRET log2 24-h IL-4 H3K27ac
C57 log2 24-h IL-4 H3K27ac
D
C57
P = 5.6 × 10−5
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
10 kb
H3K27ac
RNA Pol II
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
5 kb
DeepLIFT score
Top 20% cutoff
Position of variant from PWK and/or SPRET
Position
Top-scoring variant
G
H3K27ac fold difference between 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 between 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 between 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

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; Figure 4.2G). The fraction of enhancers containing variants at strain-similar enhancers ranged from 17 to 19% in the strains most similar to C57 (BALB and NOD) to 69 to 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 that may have reached a saturation of variation capacity (Figure 4.2G). These findings are not only consistent with local variants affecting enhancer activity but also indicate that a substantial fraction of even strongly strain-differential IL-4–induced enhancers lack these variants, consistent with previous findings for strain-specific enhancers overall (Link et al., 2018, *Cell*).

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-hour IL-4 condition on the basis of local sequence context (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 on the basis of the model's classification decision. Variants at positions with top importance scores within surrounding 300 base pair (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 Figure 4.2F illustrates 4 predicted functional variants of 14 variants in PWK and SPRET (red dashed 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 (Figure 4.2H and Supplementary Figure 4.1H). 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 (Figure 4.2I). 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.

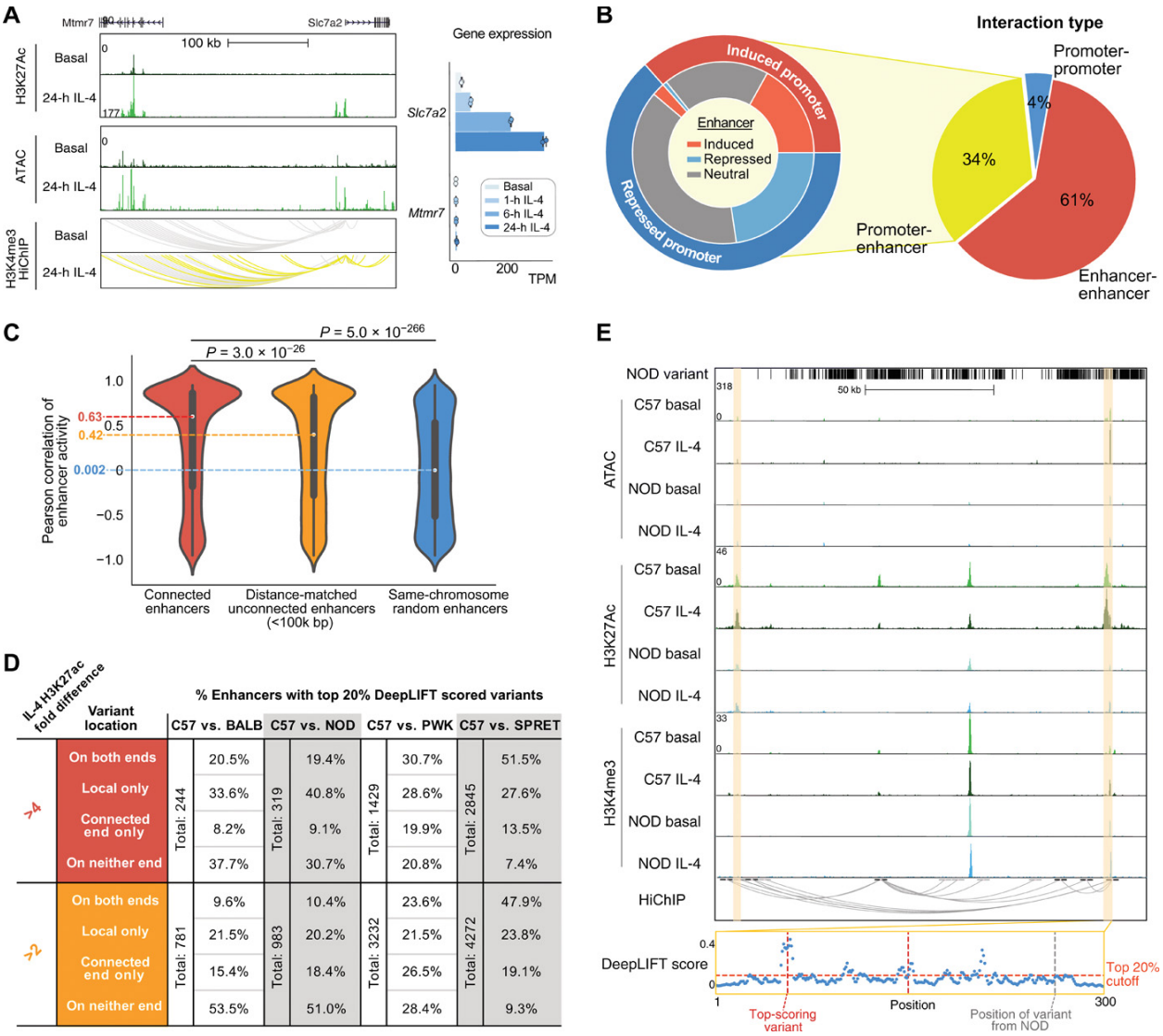
4.2.3 IL-4–activated enhancers use preexistent promoter-enhancer interactions to regulate gene activity

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 hours of IL-4 treatment. HiChIP interactions are exemplified in Fig. 3A at the Slc7a2 locus, a gene that becomes maximally activated after 24 hours of IL-4 treatment and connects primarily to an enhancer-like region within the Mtmr7 gene, which itself is expressed at negligible levels (Figure 4.3A). 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 Figure 4.2A). Moreover, enhancer-promoter interaction intensity did not correlate with IL-4–induced gene activity or the level of H3K4me3 at promoters (Supplementary Figure 4.2, B and C). However, IL-4-activated promoters mostly interact with IL-

4–activated enhancers (Fisher's exact test, P = 2.2 × 10−16), and IL-4-repressed promoters strongly interact with IL-4–repressed enhancers (P = 1.2 × 10−15; Figure 4.3B). These results suggest a preexistent and relatively stable landscape of enhancer-promoter interactions in macrophages, whose regulatory function was activated in response to IL-4.

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 (Figure 4.3B), 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 noninteractive enhancers (Figure 4.3C and Supplementary Figure 4.2D), consistent with their being functionally related. Noticeably, the closer enhancers, despite being noninteractive on the basis of 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. On the basis of 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 (Figure 4.2H). 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 (Figure 4.3D, upper left). Depending on the strain comparison, an additional 8.2 to 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 twofold 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 to 26.5%, Fisher's exact test P = 0.002 for BALB, 2.8 × 10−5 for NOD, 8.3 × 10−7 for PWK, and 3.3 × 10−10 for SPRET), suggesting that local variants have a stronger effect on inducing differential activation than variants at connected enhancers (Figure 4.3D, bottom, and Supplementary Figure 4.2E). Figure 4.3E illustrates an enhancer affected by genetic variants at the connected enhancer. The enhancer highlighted on the left is notably 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 (Waszak et al., 2015; Grubert et al., 2015).



IL-4 H3K27ac fold difference	Variant location	C57 vs. BALB		C57 vs. NOD		C57 vs. PWK		C57 vs. SPRET	
			% Enhancers with top 20% DeepLIFT scored variants						
>4	On both ends	Total: 244	20.5%	Total: 319	19.4%	Total: 1429	30.7%	Total: 2845	51.5%
>4	Local only		33.6%		40.8%		28.6%		27.6%
>4	Connected end only		8.2%		9.1%		19.9%		13.5%
>4	On neither end		37.7%		30.7%		20.8%		7.4%
>2	On both ends	Total: 781	9.6%	Total: 983	10.4%	Total: 3232	23.6%	Total: 4272	47.9%
>2	Local only		21.5%		20.2%		21.5%		23.8%
>2	Connected end only		15.4%		18.4%		26.5%		19.1%
>2	On neither end		53.5%		51.0%		28.4%		9.3%

Figure 4.3: IL-4 enhancers use preexistent promoter-enhancer interactions to regulate gene activity. (A) HiChIP indicates that the Slc7a2 promoter is highly connected with several IL-4-activated enhancers. Slc7a2 and Mtmr7 gene expression upon IL-4 stimulation was 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 noninteractive 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 dashed lines).

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

IL-4 rapidly activates a set of enhancers, most of which exhibit maximal H3K27ac at 1 or 6 hours and returns to (near) basal levels by 24 hours (Figure 4.4A, top three clusters) when most gene expression changes were found. Others are long-lasting or become activated at later time points (Figure 4.4A, bottom three clusters). De novo motif enrichment analysis of enhancers exhibiting >2.5-fold increase in H3K27ac and RNA Pol II at 1, 6, and 24 hours (Supplementary Figure 4.1A) recovered a STAT6 motif as the most enriched motif for all time points (Figure 4.4B). Motifs for the lineage-determining factors PU.1 and AP-1 (activator protein 1) family members were also recovered in all three classes of enhancers. Notably, an EGR2 motif was significantly enriched among enhancers induced at 24 hours.

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 IL-4–repressed enhancers (Figure 4.4C). The EGR motif was found as the top motif associated with enhancer activation at the 24-hour treatment time, as well as motifs of known SDTFs STAT6 and PPARγ and macrophage LDTFs PU.1, AP-1, and C/EBP (CCAAT/enhancer binding protein) (Figure 4.4C). We also found Kruppel-like factor (KLF) motifs associated with IL-4 enhancer activation, which fits with increased KLF4 expression

by IL-4, and an nuclear factor erythroid 2-related factor (NRF) motif associated with both enhancer activation (Figure 4.4C).

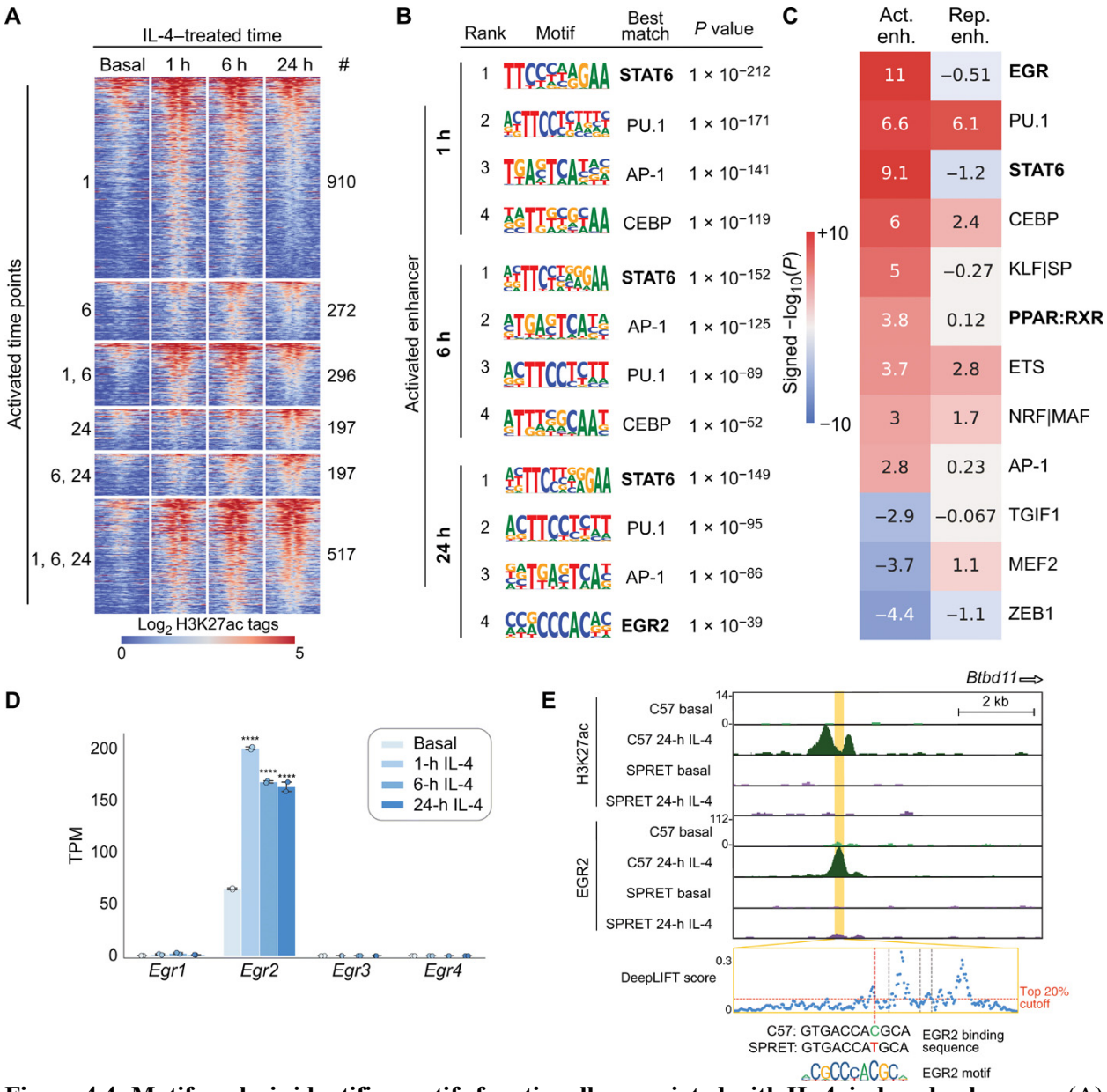


Figure 4.4: Motif analysis identifies motifs functionally associated with IL-4–induced enhancers. (A) Heatmap showing the effects of 1-, 6-, and 24-hour IL-4 stimulation on enhancer activation based on H3K27ac abundance. (B) Top motifs enriched at ATAC-seq peaks exhibiting gained H3K27ac at different time points. (C) MAGGIE motif mutation analysis on strain-differential activated and repressed enhancers after 24-hour 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 (Daniel et al., 2018; Czimmerer et al., 2018). Out of the EGR family members, only Egr2 is expressed in unstimulated BMDMs and rapidly induced after IL-4 stimulation (Figure 4.4D). 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 Pol II signal after IL-4 stimulation. These super enhancers were observed in BMDMs of all five different strains 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.

4.2.5 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 (Figure 4.1), we used H3K27ac to define three different categories of strain-differential IL-4-induced enhancers (Figure 4.5A, 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); and enhancers in which a lack of IL-4–induced activity was associated with relatively higher basal activity compared with 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 (Figure 4.5B). The closest genes for enhancers of these three categories follow similar trends as observed for enhancer activity. Low basal, equal basal, and high basal enhancers are exemplified by enhancers associated with the Treml2, Ripk2, and Cd36 genes, respectively (Figure 4.5, C to E).

Consideration of chromatin accessibility as determined by ATAC-seq further uncovered potential mechanisms that distinguished the three enhancer categories (Figure 4.5A, right column). The enhancers in the low basal category showed low to absent basal ATAC signal in noninduced strains, suggesting a lack of LDTFs under the basal condition to preoccupy chromatin required for subsequent recruitment of SDTFs after IL-4 stimulation. In contrast, high basal enhancers exhibited a higher basal level of ATAC in noninduced strains compared with the induced strains (Figure 4.5A, right column), suggesting stronger LDTF binding in noninduced 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.

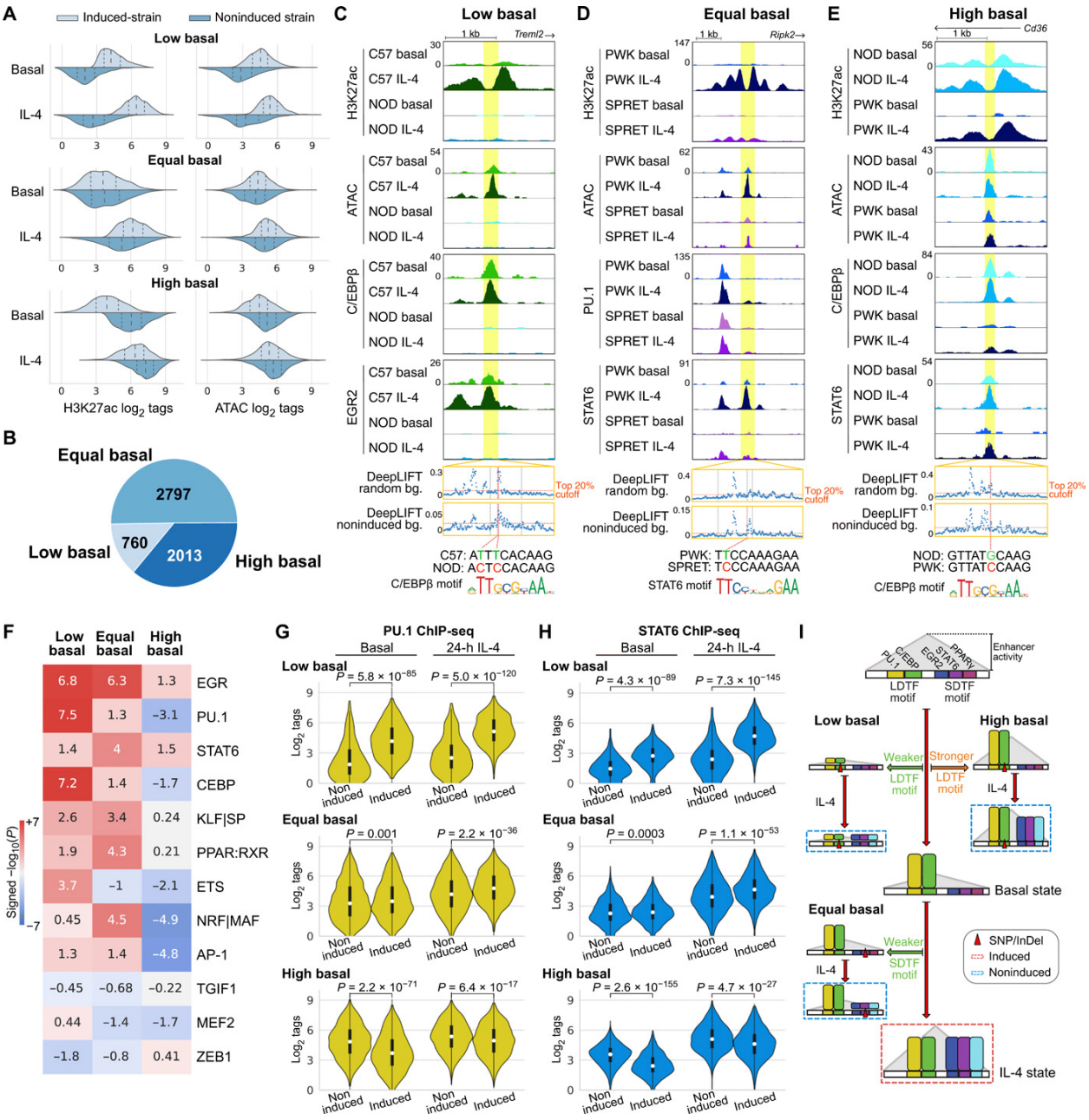


Figure 4.5: Quantitative variations in motif affinity determine dynamic responses of IL-4 enhancers. (A) Three different categories of strain-differential IL-4-activated enhancers with distributions of ATAC and H3K27ac signal. Dashed lines in each distribution indicate quartiles. (B) Numbers of enhancers in the three categories. (C to 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 and H) Binding intensities of PU.1 (G) and STAT6 (H) in noninduced and induced strains at different categories of enhancers. (I) Graphical representation of the general mechanisms for different categories of IL-4–induced enhancers. SNP, single-nucleotide polymorphism; InDel, insertion/deletion.

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. 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 (Figure 4.5F). 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 notable decrease in open chromatin under IL-4 conditions after deletion of Egr2. 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 noninduced strains (Figure 4.5F).

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 noninducible strains under both basal and IL-4 conditions (Figure 4.5G). This pattern was accompanied by significantly weaker binding of SDTFs STAT6, EGR2, and PPARγ after IL-4 stimulation (Figure 4.5H). The example in Figure 4.5C 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 noninduced strains (Figure 4.5G). Upon IL-4 stimulation, the induced strains displayed significantly stronger binding of SDTFs STAT6, EGR2, and PPARγ (Figure

4.5H). In the example in Figure 4.5D, 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 were 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 noninduced enhancers as the background, which emphasized a relatively divergent set of k-mers and focused less on those matched with LDTF motifs. 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 (Figure 4.5D). For high basal enhancers, we found stronger binding of not only the LDTFs PU.1 and C/EBPβ (Figure 4.5G) but also the SDTFs STAT6 and PPARγ (Figure 4.5H) in noninduced strains under basal conditions. For example, high basal levels of C/EBPβ and STAT6 binding were observed at the Cd36 enhancer in NOD mice (Figure 4.5E). 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 (Figure 4.5I).

4.3 Materials and methods

4.3.1 Experimental design

To investigate the influence of genetic variation on signal-dependent gene expression, enhancer activation, and transcription factor binding, we performed RNA-seq, ATAC, and ChIP-

seq to study the responses of macrophages derived from five different inbred mouse (C57, BALB, NOD, PWK, and SPRET) strains to the anti-inflammatory cytokine IL-4 (Figure 4.1A).

Female and male breeder mice for C57, BALB, NOD, PWK, and SPRET mice were purchased from the Jackson laboratory. F1 C57 × SPRET mice were crossed, and Egr2fl/fl mice were generously donated by Drs. V. Lazarevic and L. Warren (National Institutes of Health) and crossed to LyzM-Cre mice (the Jackson laboratory) to achieve myeloid-specific targeted deletion of Egr2. Mice were housed at the University of California San Diego animal facility on a 12-hour/12-hour 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- to 12-week-old healthy female mice were used for all our experiments.

Femur, tibia, and iliac bones from the different mouse strains were flushed with Dulbecco's modified Eagle's medium (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 15-cm nontissue 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 macrophage colony-stimulating factor (M-CSF), as described before (Link et al., 2018, *Cell*)], penicillin/streptomycin + l-glutamine (100 U/ml; Gibco), and amphotericin B (2.5 μg/ml; HyClone). After 4 days of differentiation, mouse M-CSF (16.7 ng/ml; Shenandoah Biotechnology) was added to the media. After an additional 2 days of culture, adherent cells were scraped and subsequently seeded onto tissue culture–treated petri dishes in DMEM containing 10% FBS, penicillin/streptomycin + l-glutamine (100 U/ml), amphotericin B (2.5 μg/ml), and M-CSF (16.7

ng/ml). Macrophages were left untreated or treated with mouse recombinant IL-4 (20 ng/ml; PeproTech) for 1, 6, or 24 hours.

Cells were fixed with Cytofix/Cytoperm Buffer (BD Biosciences, BD554714) for 10 min at room temperature. Cytofix/Cytoperm buffer was removed, and cells were washed twice with Hanks' balanced salt solution containing 2% bovine serum albumin (BSA) and 1 mM EDTA. Cells were kept in permeabilization/wash buffer (BD Biosciences, BD554714) for 1 hour at 4°C or until the experiment was performed. Fixed cells were blocked using 3% BSA, 0.1% Triton–phosphate-buffered saline (PBS) for 30 min at room temperature and then with 1/200 of the EGR2 antibody (Abcam) overnight at 4°C. The next day, cells were washed with 0.1% Triton-PBS and incubated with 1/200 donkey anti-rabbit 555 (Thermo Fisher Scientific, no. A31572) secondary antibody and phalloidin (Abcam, ab176759) for staining actin filaments, and nuclei were counterstained with 4′,6-diamidino-2-phenylindole. After washing with 0.1% Triton-PBS, slides were mounted with ProLong Gold Antifade Reagent (Life Technologies, no. 10144). Images were taken using a Leica SP8 with light deconvolution microscope.

4.3.2 Data mapping

Custom genomes were generated for BALB, NOD, PWK, and SPRET mice from the C57 or mm10 genome as before (Link et al., 2018, *Cell*) using MMARGE v1.0 (Link et al., 2018, *Nucleic acids research*) and the variant call format (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 et al., 2012) for ATAC-seq, ChIP-seq, and HiChIP data. Then, the mapped data were

shifted to the mm10 genome using the MMARGE v1.0 “shift” function (Link et al., 2018, *Nucleic acids research*) for downstream comparative analyses.

4.3.3 RNA-seq data analysis

RNA-seq data processing. Transcripts were quantified using HOMER v4.11.1 “analyzeRepeats” script (Heinz et al., 2010). Transcripts per kilobase million (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 (Zhou et al., 2019, *Nature communications*).

Categorization of strain-differential genes. Strain-differential genes were defined on the basis of 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 noninduced strain: high basal, equal basal, and low basal. In the high basal group, the noninduced 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 noninduced strain. The genes in between are categorized into the equal basal group.

F1 mice data processing. RNA-seq data from F1 mice were mapped to both parental genomes (C57 and SPRET) and analyzed in the same way as before (Link et al., 2018, *Cell*). In short, the read counts for each transcript were multiplied by the ratio of reads overlapping mutations times 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 $-1 < \log2(\text{parental fold change}) - \log2(\text{F1 fold change}) < 1$, 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.

4.3.4 ATAC-seq and ChIP-seq data analysis

On the basis of 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 Pol II were quantified within ±500 bp around the centers of reproducible ATAC-seq 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, and 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.

4.3.5 Identification of IL-4–responsive regulatory elements

IL-4–responsive enhancers were identified by the strong fold changes of H3K27ac and RNA Pol II at intergenic or intronic open chromatin. Reproducible ATAC-seq 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 Pol II within ±500 bp using HOMER v4.11.1. On the basis of the genomic annotations from HOMER annotatePeaks.pl, we classified regions at promoter-transcription start sites (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 Pol II were filtered out. For the remaining promoters and enhancers, we computed the fold changes of the normalized tags of H3K27ac and RNA Pol II 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.

4.3.6 H3K4me3 HiChIP

H3K4me3 ChIP-seq HiChIP reference preprocessing. H3K4me3 ChIP-seqs from basal and 24-hour IL-4–stimulated macrophages were performed in duplicate with input controls. Fastq files were aligned with bowtie2 (Langmead et al., 2012) to the mm10 reference genome, and peak calling was done with MACS (Zhang et al., 2008, *Genome biology*) for each replicate separately.

Significant peaks were merged using bedtools (Quinlan et al., 2010) into a general bed file that was used as corresponding peak file for MAPS.

H3K4me3 HiChIP preprocessing. HiChIP sequencing (HiChIP-seq) data were processed with MAPS (Juric et al., 2019) at 5000-bp resolution as described previously for proximity ligation-assisted ChIP-seq (PLAC-seq) (Nott et al., 2019) for all four samples separately, basal and 24-hour IL-4 duplicate samples combined, and a merge of all four samples.

Differential analysis. To identify interactions that were significantly stronger in Il-4 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 notx 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 (IL-4 versus control), with and without batch correction. No interactions were identified that were significantly different between IL-4 and control by either method (false discovery rate < 0.1, and absolute log2 fold change > 1). Hence, the combined interaction set (generated using both IL-4 and control samples) was used for downstream analysis.

4.3.7 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" 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, IL-4-repressed, and/or IL-4-neutral promoters on one end and IL-4-induced, IL-4-neutral, and/or IL-4-repressed enhancers on the other end, and quantified the number of interactions between these possible promoter-enhancer combinations in nine categories as a contingency table. Fisher's exact test was applied to the contingency table to determine whether any of the categories were significantly different for three comparisons of interest: IL-4-induced enhancer/promoter interactions versus noninduced enhancer/promoters; IL-4-repressed enhancer/promoter interactions versus nonrepressed enhancer/promoters; and IL-4-induced enhancer/promoter interactions versus IL-4-repressed enhancer/promoter interactions. For enhancer-enhancer interactions, we preselected enhancers that have at least fourfold difference in H3K27ac ChIP-seq tags between any two strains under the 24-hour 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 preselected set. To obtain noninteractive enhancers, we either randomly paired preselected 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).

4.3.8 Motif analysis

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.

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.

4.3.9 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 noninduced strains. High basal enhancers have more than twofold stronger H3K27ac in noninduced strains, while low basal enhancers have more than twofold stronger H3K27ac in induced strains (lower basal H3K27ac in noninduced strains). Equal basal enhancers are those in between.

4.3.10 Deep learning

Neural network training. We adapted a similar strategy as AgentBind (Zheng et al., 2021, *Nature machine intelligence*) for our training procedure. We started with a pretrained DeepSEA (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 versus random backgrounds (auROC = 0.894), IL-4-induced enhancers versus random backgrounds (auROC = 0.919), and IL-4–induced enhancers versus noninduced 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 noninduced enhancers were from C57 mice, while IL-4-induced enhancers were pooled from all the five strains to reach a comparable sample size. Random genomic backgrounds were generated by randomly selecting nearby guanine-cytosine (GC) content-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 1×10^{-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.

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, *Nature machine intelligence*). 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.

4.3.11 Data and code availability

All sequencing data have been made available by deposition in the Gene Expression Omnibus (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.

4.3.12 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.

4.4 Discussion

In this chapter, 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 substantial 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 and a recent study by Daniel et al. (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 noncoding 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. These 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 (Farley et al., 2015; Crocker 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 also is 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 datasets 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 these 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 and 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 eightfold enrichment in enhancers with strongest strain differences (top 1% variants for C57 versus BALB comparison; Figure 4.2I), strongly suggesting causality. Although DeepLIFT scored a substantial 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. Furthermore, 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 versus noninduced enhancers as negative training examples). This observation has important implications with respect to application of deep learning models to identify potential functional variants in

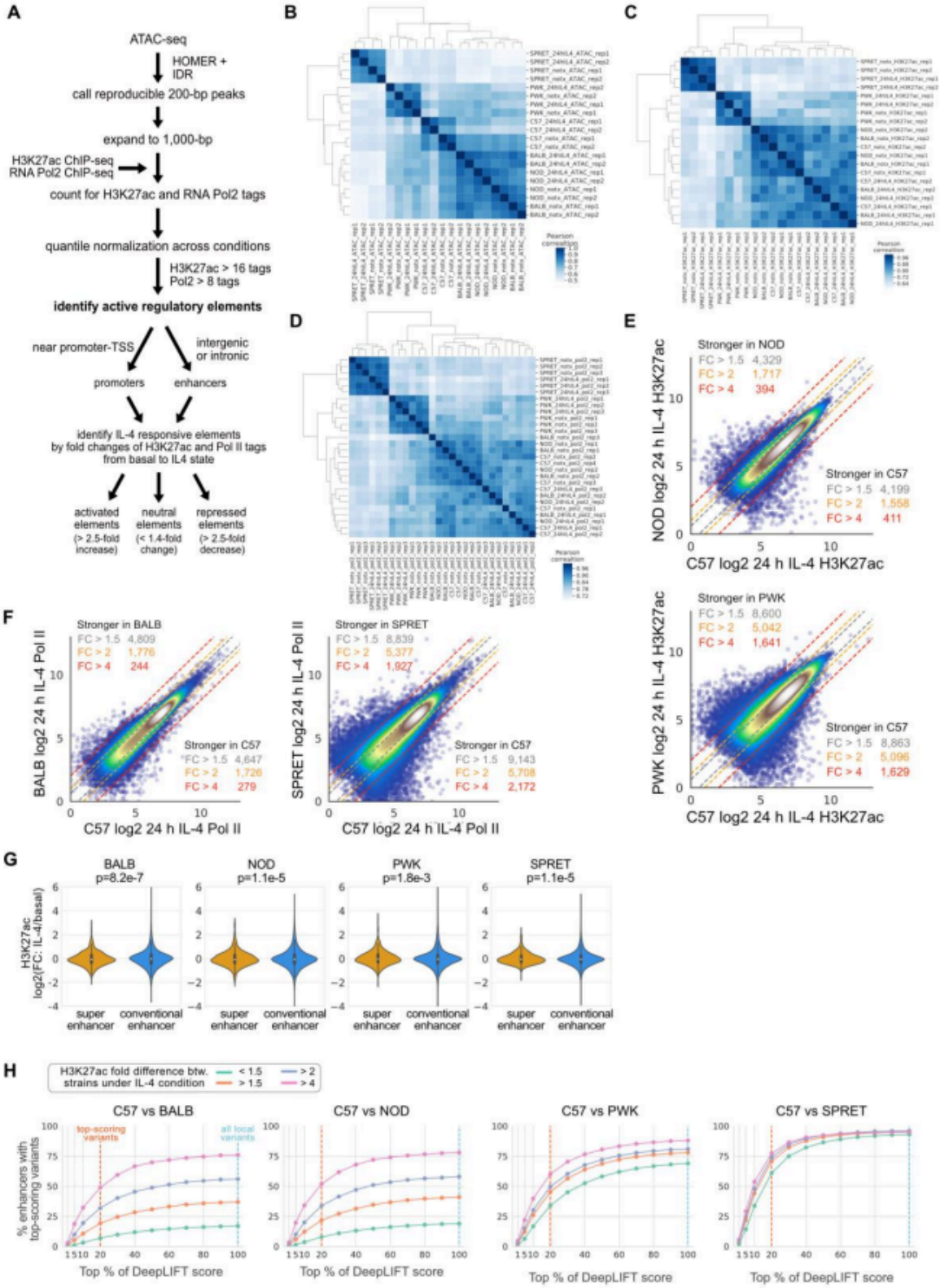
disease contexts. The datasets generated by these studies will therefore provide an important resource for further improvements in methods for interpretation of local genetic variation.

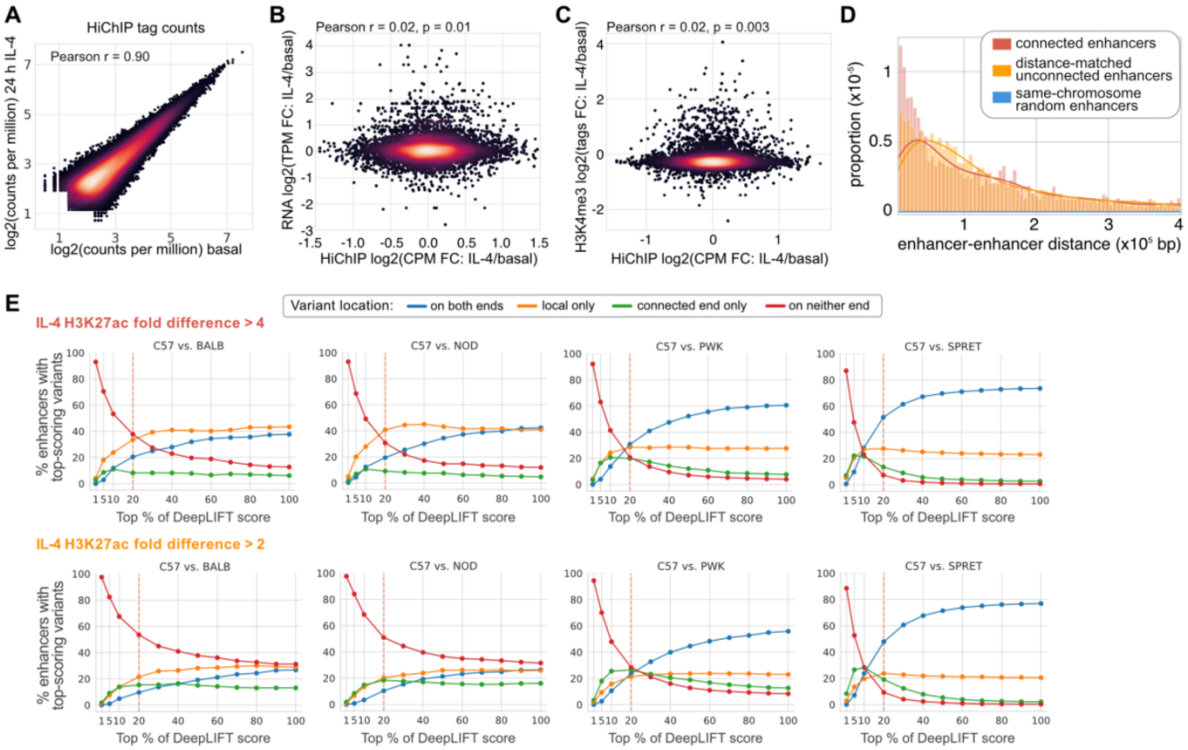
These analyses further indicated that 20 to 50% of the most divergent IL-4-responsive enhancers lacked any functional variants in the proximity of open chromatin. This fits with previous observations that variant-free enhancers can reside in cis-regulatory domains (CRD) containing functionally interacting enhancers, suggesting that a variant strongly affecting one enhancer within the CRD could have domain-wide effects (Link et al., 2019, *Cell*). This concept was supported and extended here by HiChIP experiments. In addition to demonstrating that the IL-4 response was primarily associated with preexisting enhancer-promoter connections, the HiChIP assay also captured a large number of enhancer-enhancer interactions. Examination of these connected enhancers provided evidence that a substantial 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.

4.5 Supplementary figures

Supplementary Figure 4.1: 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. 2G and 2H are based on the top 100% and 20%, respectively.


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 4.2: 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. Figure 4.3E is based on the top 20%.

4.6 Acknowledgments

This chapter, in full, is taken from "Mechanisms underlying divergent responses of genetically distinct macrophages to IL-4," Marten A. Hoeksema, Zeyang Shen, Inge R. Holtman, An Zheng, Nathan J. Spann, Isidoro Cobo, Melissa Gymrek, and Christopher K. Glass (Hoeksema et al., 2021). The material has been published on Science Advances. The dissertation author contributed to the deep learning modeling and computational analyses in this study.

We 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 Drs. L. Warren and V. Lazarevic for donating Egr2flfl mice. Funding: These studies were supported by NIH grants DK091183 and HL147835 and a Leducq Transatlantic Network grant 16CVD01 to C.K.G. Sequencing costs were partially supported by DK063491. M.A.H. was supported by a Rubicon grant from the Netherlands Organization for Scientific Research and postdoctoral grants from the Amsterdam Cardiovascular Sciences Institute and the American Heart Association. Author contributions: Conceptualization: M.A.H., Z.S., and C.K.G. Formal analysis: Z.S., M.A.H., I.R.H., and A.Z. Investigation: M.A.H., I.C., and N.J.S. Writing: M.A.H., Z.S., and C.K.G. Visualization: Z.S., M.A.H., I.R.H., and I.C. Supervision: C.K.G. and M.G. Funding acquisition: C.K.G. Competing interests: The authors declare that they have no competing interests. Data and materials availability: All data needed to evaluate the conclusions in the paper are present in the paper and the sequencing data were deposited in the GEO database: GSE159630.

CHAPTER 5

A flexible ChIP-sequencing simulation toolkit

5.1 Introduction

Chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) is a widely used technology for genome-wide mapping of the location of histone modifications (HMs) or DNA-associated proteins such as transcription factors (TFs) and chromatin regulators (CRs) (Furey et al., 2012). Dozens of methods have been developed for quantitatively analyzing ChIP-seq data, including peak callers (Zhang et al., 2008, *Genome biology*; Harmanci et al., 2014) and differential binding tools (Ross-Innes et al., 2012; Love et al., 2014). A major challenge in training and evaluating these methods as well as interpreting their results is a lack of reliable ground truth data: in most cases, the actual locations and strengths of binding sites or regions enriched for certain histone modifications are not known and cannot be reliably measured using orthogonal experimental techniques. Computational analysis of ChIP-seq is further complicated by multiple sources of noise introduced during the experimental process, including inefficiency or non-specificity of antibodies, PCR artifacts, and sequencing errors (Meyer et al., 2014; Landt et al., 2012).

Accurate simulation of ChIP-seq data can mitigate this challenge, but existing frameworks (Humburg et al., 2011; Datta et al., 2019; Zhang et al., 2008, *PLoS Computational Biology*; Subkhankulova et al., 2021) are either cumbersome to apply genome-wide or do not accurately capture important sources of variation present in real data such as pulldown non-specificity, fragment length variability, or sequencing errors. Importantly, existing simulation tools are not

capable of inferring model parameters from real ChIP-seq datasets, making it difficult to choose realistic simulation settings.

In this chapter, we will present ChIPs (ChIP-seq simulator), a flexible toolkit for rapidly simulating ChIP-seq data based on realistic statistical models. ChIPs is a computationally efficient command-line solution that allows users to easily specify a wide range of parameters modeling key experimental steps and to infer these parameters from existing datasets. We will demonstrate the applicability of ChIPs for evaluating the impact of various experimental conditions and for benchmarking computational analysis tools.

5.2 Implementation

5.2.1 Framework architecture

ChIPs models each major ChIP-seq step (shearing, immunoprecipitation, pulldown, PCR, and sequencing) as a distinct module (Figure 5.1a). It assumes binding sites for the target epitope and their binding scores (probabilities) are known. Notably, for histone modifications, we use binding to refer to genomic localization with the target modification, although the DNA itself is not typically bound by the modification. Importantly, each step is modeled in a way that key parameters can be inferred from existing datasets.

Step 1: Shearing. Cross-linked DNA is first sheared to a target fragment length, for instance by sonication or enzymatic approaches (Kidder et al., 2011). ChIPs models fragment lengths using a gamma distribution (Figure 5.1a; top) based on empirical observation of fragment distributions which have long right tails. The fragment length distribution parameters are either trivially inferred from paired end read alignments or are approximated from single end data using a heuristic method (Supplementary Figure 5.1).

Step 2: Immunoprecipitation. Sheared cross-linked DNA is subject to immunoprecipitation, during which an antibody is used to enrich the pool of fragments for those bound to the epitope of interest. To model this imperfect process, we quantify the ratio, α, of the probability of pulling down a bound versus unbound fragment. This modeled ratio is specific to each ChIP-seq experiment and depends on the antibody specificity as well as the fraction of the genome bound by the factor of interest. Let f be the fraction of the genome bound by the factor of interest and s be the fraction of pulled down reads that originate from true binding sites. ChIPs can approximate α using equation (5.1). A detailed derivation of this ratio is provided in chapter 5.4.

$$\alpha = \frac{s(1-f)}{(1-s)f} \tag{5.1}$$

The parameters f and s can be directly inferred from real data based on binding sites or enriched regions (peaks) identified by various peak-calling methods (Supplementary Figure 2).

Step 3: PCR. PCR is used to amplify pulled down fragments before sequencing. Let n_i represent the number of reads (or read pairs) with i PCR duplicates (including the original fragment). n_i is modeled using a geometric distribution, where p gives the probability that a fragment has no PCR duplicates. The parameter p is estimated as $\frac{1}{\bar{n}}$, where $\bar{n} = \frac{\sum_{i=1}^{\infty} i n_i}{\sum_{i=1}^{\infty} n_i}$.

Step 4: Sequencing. Finally, amplified fragments are subject to either paired end or single end sequencing. Sequences are based on an input reference genome using the coordinates of each fragment. We model the per-base pair substitution, insertion, and deletion rates (Supplementary Table 5.2).

5.2.2 Implementation details

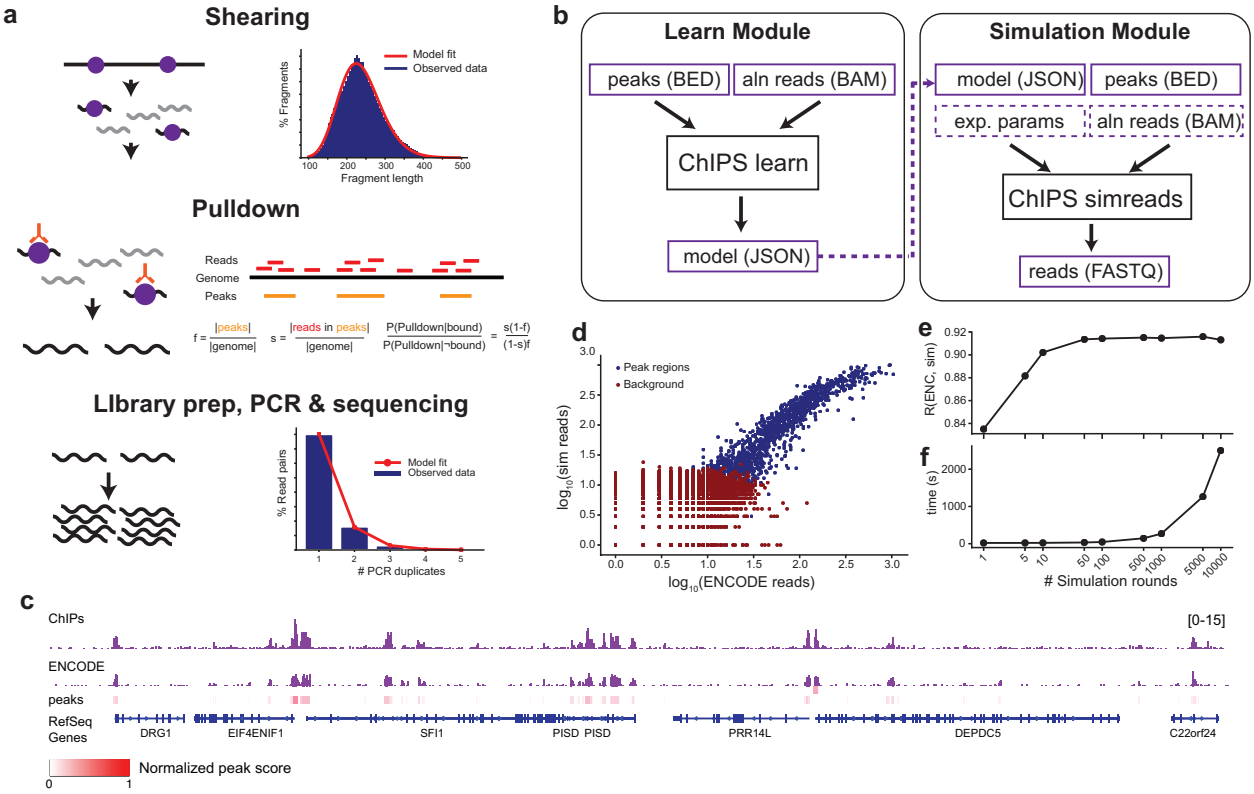
ChIPs is implemented as an open-source C++ project with source code publicly available on Github: https://github.com/gymreklab/chips. It consists of two utilities: simreads and learn (Fig. 1b). The simreads module takes in ChIP-seq model parameters and experimental settings (Supplementary Table 5.2), and outputs simulated reads. Input parameters can either be set by the user to mimic a future ChIP-seq experiment or learned from existing data using the learn module. The user must additionally specify the number of simulation rounds, which denotes the number of times the input reference genome is processed by ChIPs. Notably, this number is related, but not directly comparable, to the number of experimentally processed cells, since pulldown efficiency is not directly included in the current model. We have found that in most settings 25–100 and 1000 rounds work well for HMs and TFs, respectively. Full implementation details and methods for benchmarking experiments are provided in chapter 5.4.

5.3 Results

5.3.1 Comparison of ChIPs simulation results to real ChIP-seq data

We evaluated ChIPs using ChIP-seq data generated by the ENCODE Project (ENCODE Project Consortium, 2012) for an example histone modification H3K27ac in the GM12878 cell line. To evaluate the effect of varying the number of simulation rounds, we simulated reads on chromosome 22 using parameters inferred from real data over a range of simulation rounds (1–10,000). Run time for chromosome 22 ranged from 11 s (1 round) to 15 min (10,000 rounds). Resulting reads were aligned to the hg19 reference genome using BWA-MEM (Li, 2013), and duplicates were flagged using Picard (Broad Institute, 2018). Visual inspection of the resulting coverage profiles shows high similarity between real and simulated data (Figure 5.1c).

Figure 5.1: ChIPs overview. (a) Overview of the ChIPs model. ChIPs models four steps: shearing (top), pulldown (middle), PCR (bottom), and sequencing. Top: the dark blue histogram shows an example fragment length distribution from real paired end ChIP-seq data. The red line shows the best fit gamma distribution. Middle: pulldown is modeled using two parameters; f (the fraction of the genome bound by the factor) and s (the probability that a pulled down fragment is bound. Bottom: The dark blue histogram shows an example of a distribution of the numbers of PCR duplicates in real ChIP-seq data. The red line shows the best fit geometric distribution. (b) Schematic of ChIPs modules. The learn module takes an existing ChIP-seq experiment (aligned reads and peaks) and learns model parameters (see Supplementary Table 5.2). The simulation module takes as input a set of peaks and model parameters, simulates a ChIP-seq experiment, and returns raw reads in FASTQ format. Model parameters input to the simulation module may either be learned from an existing ChIP-seq dataset (dashed arrow) or manually specified to capture planned experimental conditions. Purple borders represent input or output files and black boxes denote ChIPs commands. Boxes with solid lines denote required inputs. Boxes with dashed borders denote optional inputs. "Exp. params" denotes experimental parameters including the number of reads, read length, and number of simulation rounds. "Aln reads" denotes aligned reads in BAM format. (c) Example coverage profiles of real versus simulated data. The bottom track shows peaks identified by ENCODE, with normalized peak scores between 0 to 1 colored based on a gradient from white to red. The middle track shows coverage profiles based on aligned reads from ENCODE, and the top track shows coverage profiles based on ChIPs simulations. Coverage profiles were generated using IGV. Coverage profiles may also be viewed interactively at https://tinyurl.com/y7x6ggdq. (d) Concordance of read counts between simulated versus real ChIP-seq data. chr22 was divided into non-overlapping 5 kb bins. The scatter plot shows the comparison of read counts per bin for bins overlapping peaks (dark blue) or background regions (dark red). The x- and y-axes are on a log10 scale. The plot shown is for 100 simulated genome copies. (e) Read count correlation between real and simulated data as a function of number of simulated genome copies. For each number of copies, the correlation was computed between read counts in 5 kb bins overlapping input peaks. The x-axis is on a log10 scale. (f) Simulation run time as a function of number of simulated genome copies. The x-axis is on a log10 scale.


a
Shearing
% Fragments
Model fit
Observed data
Fragment length
Pulldown
Reads
Genome
Peaks
Library prep, PCR & sequencing
% Read pairs
# PCR duplicates
b
Learn Module
peaks (BED)
aln reads (BAM)
ChIPS learn
model (JSON)
Simulation Module
model (JSON)
peaks (BED)
exp. params
aln reads (BAM)
ChIPS simreads
reads (FASTQ)
c
ChIPs
[0-15]
ENCODE
peaks
RefSeq Genes
DRG1
EIF4ENIF1
SFI1
PISD
PISD
PRR14L
DEPDC5
C22orf24
Normalized peak score
d
Peak regions
Background
log10(sim reads)
log10(ENCODE reads)
e
R(ENC, sim)
f
time (s)
Simulation rounds

Next, we compared read counts in bins of 5kb and found high correlation between real and simulated data in bins containing at least one peak (Fig. 1d; Pearson r=0.91; p<10−200; n=1,232 bins; 100 simulation rounds). Further, correlation with ENCODE data increased as a function of the number of simulation rounds but plateaued around 100, suggesting little gain in simulating additional rounds compared to the time tradeoff (Figure 5.1e–f). We repeated this analysis on multiple additional HMs and TFs in GM12878 with similar results (Supplementary Figure 5.3).

5.3.2 Benchmarking against existing ChIP-seq simulators

We next benchmarked ChIPs against existing ChIP-seq simulators, which are summarized in Supplementary Table 5.1. We focused on two recent methods: (1) ChIPulate (Datta et al., 2019) is a method for simulating TF ChIP-seq data using detailed modeling of locus-specific binding energies. ChIPulate only simulates reads at bound regions, and does not simulate background fragments outside of peak regions, a key feature of real ChIP-seq datasets related to the antibody specificity. (2) isChIP (Subkhankulova et al., 2021) is a command-line method for simulating ChIP-seq data based on a set of input peaks, model parameters, and sequencing parameters. While isChIP performs a similar task to ChIPs, it is not able to infer model parameters from existing datasets, which is a key feature of ChIPs. A more detailed description of model differences between these tools is provided in chapter 5.4.

We used ChIPs, ChIPulate, and isChIP to simulate ChIP-seq data based on six different ENCODE datasets including 3 HMs (H3K4me1, H3K4me3, and H3K27ac) and 3 TFs (BCLAF1, IKZF1, and NFYA) (Supplementary Table 5.3). For each dataset, we used the three methods to simulate data for chr22 based on ENCODE peaks and with settings meant to capture similar properties of the ENCODE data, including read length and read number. We additionally inferred

model parameters using the learn module of ChIPs and used these models to set appropriate simulation options for each tool when possible (details are in chapter 5.4). For each tool, we varied the number of simulation rounds (similar to the number of cells) from 1 to 10,000. ChIPulate simulations took approximately 80 min to complete regardless of the number of simulation rounds, although subsequent simulations reused intermediate files and were faster. isChIP consistently achieved the fastest run time (e.g., 0.8 min for 1000 rounds on H3K27ac compared to 4.9 min for ChIPs). For both isChIP and ChIPs, simulation time was far less than the run time of downstream steps of sequence alignment and peak calling.

For each simulated dataset, we compared to real data using two methods. First, similar to above, we aligned simulated reads to the hg19 reference genome and compared read counts in 1kb bins containing at least one peak. As expected, correlation with ENCODE increases for all tools with additional simulation rounds (Supplementary Figure 5.4a). In all evaluated conditions, we found that ChIPs showed superior correlation with ENCODE data. The performance of ChIPs was virtually unchanged when using models based on paired versus single end data (Supplementary Figure 5.4a).

Second, to evaluate how well each tool captures noise in real data, we examined the distribution of read counts in bins with and without peaks (referred to as peak and background regions, Supplementary Figure 4b) between simulated and real data. We further visualized these trends using simulated coverage profiles and ENCODE data using the Integrative Genomics Viewer (Robinson et al., 2011) (Supplementary Figure 5.5). In all cases, data simulated by ChIPs most closely matches read count distributions in peak versus background regions in the ENCODE data. As expected, almost no reads from ChIPulate align to background regions. For isChIP, we found that using the default background noise level resulted in far higher signal to noise ratios than

in the real data. We attempted to more closely match ENCODE data by performing an additional experiment with increased background noise. This in some cases alleviated the bias but still matched less closely than ChIPs data (Supplementary Figure 5.4b).

Taken together, these benchmarking results show that ChIPs most accurately captures properties of real ChIP-seq data. Further, whereas ChIPs could learn appropriate model parameters from existing datasets, the alternative tools first required detailed user involvement to determine realistic simulation settings for a particular dataset type. While it is hard to rule out that further tuning of parameters for each method could achieve higher correlation, we found that without a method to infer parameters from existing data that it was difficult to choose optimal simulation settings.

5.3.3 Demonstration of ChIPs applications

Finally, to demonstrate the ability of ChIPs to generate ground truth data for evaluating analysis tools, we compared performance of multiple peak calling methods on simulated datasets. We focused on five representative tools: MACS2 (Zhang et al., 2008, *Genome biology*), GEM (Guo et al., 2012), MUSIC (Harmanci et al., 2014), BCP (Xing et al., 2012), and HOMER (Heinz et al., 2010). We measured peak calling performance using simulated datasets representative of generic HMs or TFs as described above but with varying degrees of non-specific binding (ChIPs s parameter, commonly referred to as a SPOT or FRIP score (Landt et al., 2012); Figure 5.2e–f, Supplementary Figure 5.7). As expected, in all settings peak calling performance increased as a function of s. No method achieved superior performance across all datasets or metrics. For TFs, GEM, MACS2, and HOMER showed similarly high F1 scores for datasets with $s>0.05$. For HMs, all tools except BCP showed high F1 scores across a range of s values. Notably, this analysis

captures only a small subset of possible dataset parameters, and it is likely that results will vary depending on specific datasets. Previous work has performed an extensive evaluation of various peak calling methods (Thomas et al., 2017).

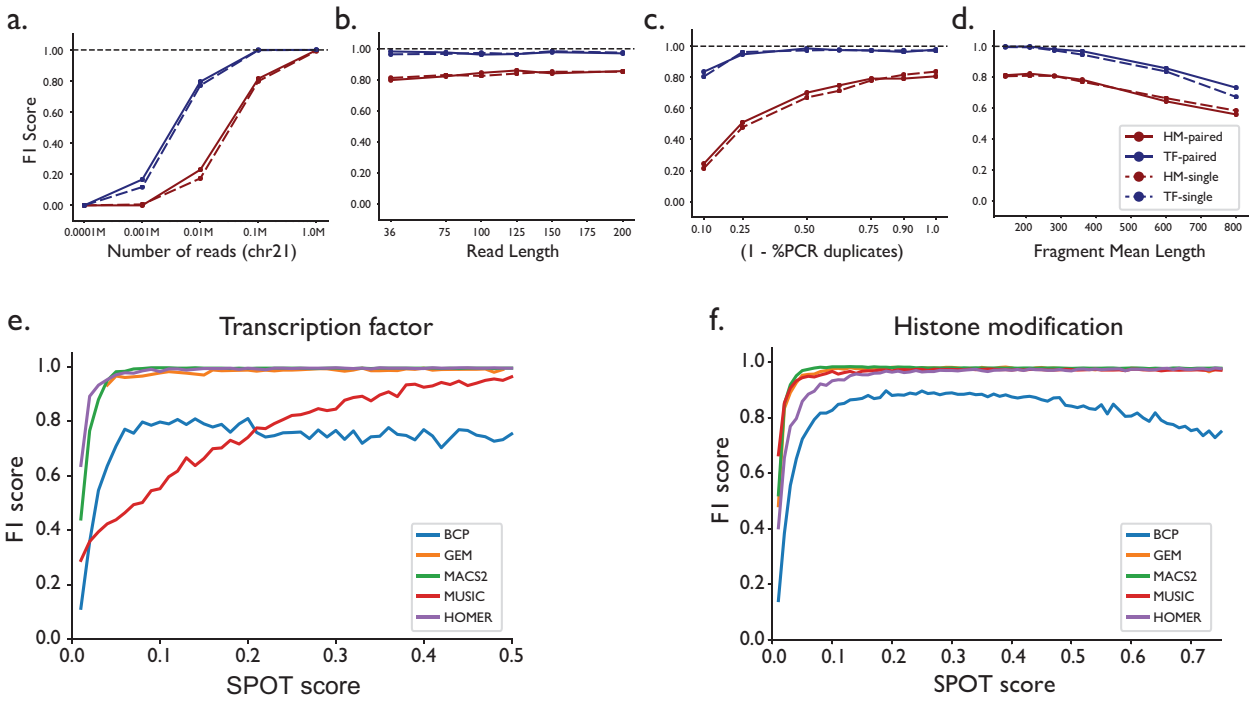


Figure 5.2: Example ChIPs applications. (a)-(d) Evaluation of the effects of varying experimental parameters on peak calling performance. Results are based on simulation of generic TF and HM datasets for chr21 as described in chapter 5.4. In each plot the y-axis shows the F1 score computed by comparing ground truth peaks to those inferred from simulated datasets using MACS2. a F1 score as a function of the total number of reads simulated from chr21. (b) F1 score as a function of read length. (c) F1 score as a function of PCR duplicates. The x-axis gives the parameter p, which can be interpreted as the percent of fragments with no PCR duplicates (Supplementary Table 5.2). (d) F1 score as a function of mean fragment length (bp). Red = HM; Blue = TF; solid lines=paired end reads; dashed lines=single end reads. (e)-(f) Evaluation of various peak calling methods on simulated TF (e) and HM (f) datasets with different noise levels. Noise levels are quantified using s, the fraction of pulled down reads that originate from true binding sites. Blue = BCP; orange = GEM; green = MACS2; red = MUSIC; purple = HOMER.

5.4 Method details

5.4.1 Model details

This section provides additional details on the ChIPs model.

Step 1: shearing. In the case of paired-end reads, fragment lengths can be determined trivially from the mapping locations of paired reads. The observed fragment length (X_i) for each read pair i can be computed based on the mapping coordinates of the two reads. The learn module randomly selects 10,000 uniquely aligned read pairs from the input BAM for fitting a gamma distribution. Read pairs are filtered to remove fragments that are unaligned, not properly paired, marked as duplicates or marked as secondary alignments. Read pairs are further filtered to remove fragments with length greater than 3 times the median length of selected fragments. The mean fragment length is easily computed as $\mu = \frac{\sum_{i=1}^{n} X_i}{n}$, where n is the number of fragments remaining after filtering. We then use the method of moments to find maximum likelihood estimates of the gamma distribution shape (k) and scale parameters (θ):

$$\theta = \frac{1}{n\mu}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{5.2}$$

$$k = \frac{\mu}{\theta} \tag{5.3}$$

For single-end reads, individual fragment lengths are not directly observed. We outline a novel method for estimating fragment length distributions from single-end data in the chapter 5.5.

Step 2: immunoprecipitation. We use B_i to denote that fragment i is bound, $\overline{B_i}$ to denote that fragment i is unbound, and D_i to denote that fragment i is pulled down. For fragment i, assuming the probability that a given bound fragment is pulled down is approximately constant over all fragments, the probability of being pulled down can then be written as:

$$P(D_i) = P(D, B_i) + P\left(D, \overline{B_i}\right) = P(D|B)P(B_i) + P\left(D\middle|\overline{B}\right)\left(1 - P(B_i)\right) \tag{5.4}$$

where $P(D|B)$ denotes the probability that a given bound fragment will be pulled down and $P(D|\overline{B})$ denotes the probability that a given unbound fragment will be pulled down.

For any fragment i, we set the probability of it being bound $P(B_i)$ based on the scores of peaks it overlaps:

$$P(B_i) = \begin{cases} 0, & \text{if no overlap} \\ 1 - \prod_{r \in R} 1 - S(r), & \text{if overlap peak(s)} \end{cases} \tag{5.5}$$

where R is the set of all peak regions overlapping fragment i and $S(r)$ is the probability that peak r is bound. A method for estimating $S(r)$ is detailed below.

We compute conditional pulldown probabilities using Bayes' Rule:

$$P(D|B) = \frac{P(B|D)P(D)}{P(B)} \tag{5.6}$$

$$P\left(D\middle|\overline{B}\right) = \frac{P\left(\overline{B}\middle|D\right)P(D)}{P\left(\overline{B}\right)} \tag{5.7}$$

where here $P(B|D)$ and $P(\overline{B}|D)$ represent averages across all fragments. There is no straightforward way to compute $P(D)$, the average probability that a fragment is pulled down, using only observed ChIP-seq reads since we do not actually observe fragments that are not pulled down. Thus, we do not attempt to compute these conditional probabilities directly. Instead, we take the ratio which cancels $P(D)$:

$$\alpha = \frac{P(D|B)}{P\left(D\middle|\overline{B}\right)} = \frac{P(B|D)P\left(\overline{B}\right)}{P\left(\overline{B}\middle|D\right)P(B)} \tag{5.8}$$

$P(B)$, or the probability that a fragment is bound on average, is equal to f, the fraction of the genome bound by the factor of interest. We can approximate f as the sum of the lengths of all peaks l_r weighted by their binding probabilities divided by the total length of the genome G:

$$f \approx \frac{1}{G}\sum_{r \in R} S(r) l_r \tag{5.9}$$

where G is the size of the reference genome, $S(r)$ is the probability peak region r is bound as described above, and l_r is the length of peak r, assuming no overlap between peaks.

$P(B|D)$, or the probability that a fragment is bound given that it is pulled down, is a measure of the specificity of the antibody. Assuming the majority of reads falling in peaks are truly bound, this can be roughly approximated as the percent of fragments falling within peaks, which is denoted as s. Using these two metrics, f, and s (Supplementary Table 5.1), we can simplify the ratio α using equation (5.1). And since the ratio α is available, for simplicity in simulations we set $P(D|B) = 1$ and $P\left(D\middle|\overline{B}\right) = \frac{1}{\alpha}$. Substituting into equation (5.8) above, we compute the probability that fragment i is pulled down as:

$$P(D_i) = P(B_i) + \frac{1}{\alpha}\left(1 - P(B_i)\right) \tag{5.10}$$

where $P(B_i)$ is based on peak scores as defined above. It is noteworthy that, in reality, $P(D|B)$ is likely to be much smaller than 1, since the pulldown process is inefficient, and many fragments are lost. Further, the number is likely to vary largely across different experiments. Estimating the absolute value of $P(D|B)$ from real datasets is a topic of future work.

5.4.2 Inferring fragment lengths from single-end reads

To estimate the fragment length distribution from single-end reads, we assume the length distribution follows a gamma distribution with mean μ and variance v, and use reads located inside ChIP-seq peaks (provided as input) to estimate μ and v which are used to compute k and θ. This is a heuristic method that provides reasonable estimates of the fragment size distribution, which is sufficient for most applications.

For each peak $peak_i$, we keep track of two lists, $\{start\}_{peak_i}$ and $\{end\}_{peak_i}$. For each read overlapping $peak_i$, if the read is on the forward strand, we add its start coordinate to $\{start\}_{peak_i}$. If the read is on the reverse strand, we add its start coordinate to $\{end\}_{peak_i}$. The center point of this peak is calculated as:

$$center_{peak_i} = \frac{mean\left(\{start\}_{peak_i}\right) + mean\left(\{end\}_{peak_i}\right)}{2} \tag{5.11}$$

For every $peak_i$, we offset the coordinates in $\{start\}_{peak_i}$ and $\{end\}_{peak_i}$ by $center_{peak_i}$, so that the coordinates of start points and end points are symmetric around zero. We then concatenate lists from each peak to form $\{start\}$ and $\{end\}$:

$$\{start\} = \bigoplus_{i=0}^{n} \left(\{start\}_{peak_i} - center_{peak_i}\right) \tag{5.12}$$

$$\{end\} = \bigoplus_{i=0}^{n} \left(\{end\}_{peak_i} - center_{peak_i}\right) \tag{5.13}$$

The mean fragment length μ can be estimated as:

$$\mu = mean(\{end\}) - mean(\{start\}) \tag{5.14}$$

We calculate the probability density functions, cumulative density functions and expected density functions for both $\{start\}$ and $\{end\}$. The expected density function $EDF(x)$ is defined as the expected deviation of a random element in the list to x:

$$EDF_{start}(x) = E(|S - x|) \tag{5.15}$$

$$EDF_{end}(x) = E(|E - x|) \tag{5.16}$$

where S is a random element in $\{start\}$ and E is a random element in $\{end\}$.

After computing μ, we reduce the density function of the fragment length distribution from $p_{\mu,v}(x)$ to $p_v(x)$, the probability density function of the fragment length variance. We construct a score function $F(v)$ as shown below.

$$F(v) = E_v\left(\left|S + \frac{L}{2}\right|\right) + E_v\left(\left|E - \frac{L}{2}\right|\right) - E\left(\left|S + \frac{\mu}{2}\right|\right) - E\left(\left|E - \frac{\mu}{2}\right|\right) \quad (5.17)$$

$$E_v\left(\left|S + \frac{L}{2}\right|\right) = \sum_{x=0}^{\infty} p_{v(x)} * EDF_{\{start\}}\left(-\frac{x}{2}\right) \quad (5.18)$$

$$E_v\left(\left|E - \frac{L}{2}\right|\right) = \sum_{x=0}^{\infty} p_{v(x)} * EDF_{\{end\}}\left(-\frac{x}{2}\right) \quad (5.19)$$

$$E\left(\left|S + \frac{\mu}{2}\right|\right) = EDF_{\{start\}}(x) \quad (5.20)$$

$$E\left(\left|E - \frac{\mu}{2}\right|\right) = EDF_{\{end\}}(x) \quad (5.21)$$

where L represents a randomly chosen fragment length. Intuitively, if v is guessed correctly, $F(v)$ should be equal to zero.

To find an optimal v that minimizes $|F(v)|$, we conduct a binary search between 1000 and 10,000. In practice, we slightly offset the last two items in the score function in equation (5.17) to get the score below, which gives slightly more accurate estimation of v on real data:

$$F(v) = E_v\left(\left|S + \frac{L}{2}\right|\right) + E_v\left(\left|E - \frac{L}{2}\right|\right) - E\left(\left|S + \frac{\mu}{2} - \frac{E - \frac{\mu}{2}}{4}\right|\right) - E\left(\left|E - \frac{\mu}{2} - \frac{S + \frac{\mu}{2}}{4}\right|\right) \quad (5.22)$$

This may be because fragment length distributions are truncated on the left end, with little or no fragments with lengths less than 100bp observed, and thus do not follow a true gamma distribution. Examples of inferred fragment length distributions from single-end data compared to actual fragment length distributions for a variety of datasets is shown in Supplementary Figure 5.1.

5.4.3 ChIPs implementation details

5.4.3.1 Peak scores

A peak score $S(r)$ is defined for each input peak, where $S(r)$ gives the probability that a fragment overlapping the region is bound. Note we cannot directly estimate this probability from bulk ChIP-seq data but assume variability in intensity across peaks is representative of different relative binding probabilities. Based on user input, ChIPs computes these binding probabilities either based on peak intensities given in an input BED file or based on read counts from an existing BAM file.

If peak intensities are provided, $S(r)$ is computed as the score of peak r divided by the maximum peak intensity. If a BAM file is provided, $S(r)$ is defined as the number of reads overlapping peak r divided by the maximum number of reads overlapping any peak. In both cases, resulting scores $S(r)$ are between 0 and 1.

By specifying the option --no-scale, users may directly input binding scores that will be treated directly as probabilities and will not be rescaled. Users may also specify --scale-outliers to remove peak intensities greater than 3 times the median score before rescaling peak intensities to reduce the effect of outlier peaks. In this case all peaks with scores greater than 3 times the median will be set to 1.

5.4.3.2 Learn implementation

The ChIPs learn module takes a set of input peaks (BED) and aligned reads (BAM) and learns parameters for (1) fragment length distribution (k, θ); (2) pull-down efficiency (f, s); and (3) PCR efficiency (p) (Supplementary Table 5.1). BAM files must have PCR duplicates marked using a tool such as Picard to enable accurate estimation of PCR parameters. The learn module outputs model parameters to a JSON file that can be used as input to the simreads module. We

found that the key pulldown parameter, α, learned, is relatively robust to the peak caller used to generate the input peak dataset (Supplementary Figure 5.2).

5.4.3.3 Simulation implementation

The simreads module takes in a set of peaks, model parameters (e.g., from the learn module), experimental parameters (including read length, number of reads), and number of simulation rounds, and simulates raw sequencing reads in FASTQ format. First, each copy of the genome (equivalent to one simulation round) is randomly fragmented based on the specified fragment length gamma distribution. Next, ChIPs decides whether to pull down each fragment based on its overlap with input peaks based on $P(B_i)$ described above. Finally, ChIPs generates reads from the resulting pool of fragments based on input parameters (using the specified mode read length, number of reads, and mode [single/paired]). ChIPs outputs "PCR duplicates" of each sequenced read based on the input PCR rate p.

In practice, in each round the genome is processed in bins (default size 100kb) to avoid storing large fragment pools in memory. In a preprocessing step, ChIPs determine how many reads to generate from each simulation round based on the target number of output reads. ChIPs is parallelized by performing different simulation rounds on separate threads simultaneously.

5.4.4 Benchmarking experiments

5.4.4.1 Evaluating ChIPs performance

To evaluate ChIPs, we compared simulated reads to reads from real ChIP-seq experiments using read alignments and peaks generated by ENCODE. All ENCODE accessions for reads and peaks are given in Supplementary Table 5.3. We first ran the chips learn module to learn model

parameters based on the ENCODE BAM and BED files for each TF or HM. We used learn parameters -c 7 -t bed --scale-outliers. The json model file output by the learn module was used as input to the simulate module. We performed various runs using an increasing number of simulation rounds (chips simulate argument --nc 1, 5, 10, 50, 100, 500, 1000, 5000, or 10000). We simulated single end reads for chr22 using the same read length as the corresponding ENCODE dataset and setting the number of reads to twice of the number of aligned reads for chr22 in the real dataset since not all simulated reads will uniquely align back to the genome. Resulting reads were aligned to the hg19 reference genome using BWA-MEM v0.7.12-r1039, and duplicates were flagged using Picard v2.18.11.

We used the bedtools (v2.27.1) makewindows command to generate a list of non-overlapping windows across chr22 and the bedtools multicov command to count the number of reads from simulated or ENCODE BAM files falling in each 1kb window. We used the bedtools intersect command to determine the intersection of each bin with the input peak files. For each bin, we determined the Pearson correlation between log10 read counts in each simulated vs. ENCODE dataset after adding a pseudocount of 1 read to each bin.

Timing experiments were performed in a Linux environment running Centos 7.4.1708 on a server with 28 cores (Intel® Xeon® CPU E5-2660 v4 @ 2.00GHz) and 125 GB RAM using the UNIX "time" command and are based on the "sys" time reported.

5.4.4.2 Comparison to ChIPulate

We compared performance of ChIPulate [5] to ChIPs based on the datasets described in Supplementary Table 5.3 and shown in Supplementary Figures 5.5-5.6. We first computed binding energies for each peak required as input. Binding probabilities $S(r)$ for each peak r were

computed as described above by scaling peak scores based on ENCODE data to be between 0 and 1, similarly setting outlier peaks with scores higher than twice the median to have score 1. Then we computed binding energies for each region r as:

$$E_r^{bound} = E_r^{unbound} + chemical\ potential + \ln\left[\frac{1 - S(r)}{S(r)}\right] \quad (5.23)$$

E_r^{bound} and $chemical\ potential$ are provided as command line inputs to ChIPulate. We set chemical potential to 0, which provided the best dynamic range across peak scores and highest correlation with real data. Background binding energy was set to the default (1.0).

ChIPulate parameters were set to achieve the same fragment length distributions, number of reads, and read lengths as in ENCODE data. Fragment length (--fragment-length) was set to the mean fragment size inferred by ChIPs learn. Read length (--read-length) was set to the read length of each dataset, ranging from 36bp to 101bp. Depth (-d) was set to the ratio of the number of reads in the ENCODE dataset mapped to chr22 divided by the number of peaks on chr22. We additionally set these parameters: p amp=0.50, p ext=0.54, --mu-A=0 based on examples shown in ChIPulate documentation. We used the parameter -n to vary the number of copies (simulation rounds) from 10 to 10,000. Runs with -n set to 1 or 5 failed and so were excluded from analysis. All other parameters were set to default values. Read counts in windows of 1kb were compared between simulated and ENCODE datasets using bedtools multicov as described above.

5.4.4.3 Comparison to isChIP

We similarly evaluated performance of isChIP v1.0 based on the datasets described in Supplementary Table 5.3 and shown in Supplementary Figures 5.4-5.5, setting parameters to mimic those of real datasets. Read length (-r) and the maximum number of reads (--rd-lim) were set based on read lengths and the number of reads aligned to chr22 in ENCODE data. The log of

the mean fragment length (-L) was set based on the mean fragment length inferred by ChIPs learn. We provided ENCODE peaks as input with option --bscore 0. Using ENCODE peak scores directly (--bscore 7) resulted in very poor correlation with real data. We used the parameter --cells to vary the number of cells from 1 to 10,000.

Since we did not have a way to estimate foreground and background parameters (--ground option), we evaluated two settings. In the first, we set foreground to the spot score and background to the default value of 1. In the second, labeled "HighBG", we increased the background noise to 4.

Unless otherwise specified, we set the number of PCR cycles to 0 (--pcr 0). Attempts to run with even small numbers of PCR cycles resulted in unreliable output. Results from setting --pcr 10 are shown in Supplementary Figure 5.4.

5.4.4.4 Simulating benchmarking peaks

For experiments shown in Figure 5.2, we generated two sets of datasets meant to represent characteristics of either histone modification (HM) or transcription factor (TF) ChIP-seq datasets. We simulated a peak file for TF and HM each, with SP1 (bam=ENCFF001TYZ) and H3K27ac (bam=ENCFF411MHX) as templates. To generate peaks, we randomly placed peaks on the hg19 genome sampling peak lengths and scores from distributions observed on real data. In total, we simulated 29,579 non-overlapping peaks for the TF dataset and 77,413 for the HM dataset. Results in Figure 2 are based on 350 TF peaks and 697 HM peaks on chr21.

5.4.4.4 Evaluating effects of experimental parameters

For evaluating experimental parameters (Figure 2a-d), we used ChIPs to simulate reads from chr21 based on either the HM or TF peak sets described above. Unless otherwise noted, all runs used ChIPs simulate parameters --scale-outliers --numcopies 100 --numreads 100000 --readlen 36 --gamma-frag 20,15 --pcr rate 0.8. For TF datasets, unless otherwise noted, all runs used model options --spot 0.40 --frac 0.001. For HM datasets, unless otherwise noted, all runs used model options --spot 0.45 --frac 0.01. For evaluating read number, the --numreads option was set to 100, 1000, 10000, 100000, or 1000000. For evaluating read length, the --readlen option was set to either 36, 75, 100, 125, 150, or 200. For evaluating PCR rate, the --pcr rate option was set to either 0.1, 0.25, 0.5, 0.625, 0.75, 0.875, or 1.0. For evaluating fragment length, the --gamma-frag option was set to either 10,14, 21,10, 40,7, 90,4, 200,3, or 400,2. In all cases separate datasets were generated for single and paired end reads using the --paired option to specify paired end output. Resulting reads were aligned to the hg19 reference genome using BWA-MEM v0.7.12-r1039, and duplicates were flagged using Picard v2.18.11.

5.4.4.5 Evaluating peak callers

We conducted multiple sets of experiments simulating paired-end ChIP-seq datasets with different SPOT scores (s ranging 0.01-0.5 for TF data and 0.01-0.75 for HM data). For each simulated dataset, we ran ChIPs simreads using the HM and TF model parameters described above and with additional options --paired --region chr21:1-48129895. We benchmarked peakcallers macs2 (v2.2.6), GEM (v3.4), MUSIC, BCP (peakranger v1.18), and HOMER. When running these peak callers, we applied their default parameter settings except flags indicating the types of input data (i.e., TF or HM).

To quantitatively evaluate the quality of the predicted peaks, we followed the metrics used in (Thomas et al., 2017) and trimmed each peak to 200bp around its summit. Then, we calculated the number of real peaks that peak callers retrieved as well as the number of predicted peaks that overlap with real peaks.

5.4.5 Function comparison

Below we compare how key steps of the ChIP-seq process are modeled by ChIPs in comparison to other recently developed ChIP-seq simulators (ChIPulate and isChIP). These differences are also summarized in Supplementary Table 5.1.

5.4.5.1 Shearing

- **ChIPulate:** uses a fixed fragment length.
- **isChIP:** models variable fragment lengths drawn from a log-normal distribution with an optional size-selection step.
- **ChIPs:** models variable fragment lengths drawn from a gamma distribution. Gamma and log-normal distributions often appear highly similar, and in practice we have not found this difference in fragment length distributions to have a significant impact on simulation results.

5.4.5.2 Pulldown and binding efficiency

- **isChIP:** models the probability of loss of selected foreground vs. background fragments compared to generated fragments.

- **ChIPulate:** models binding efficiencies of each binding site but does not consider background regions outside of specified binding sites.
- **ChIPs:** models pulldown efficiency using a Bayesian model based on the SPOT score (s) and fraction of the genome found (f), both of which can be inferred from real datasets and peaks generated by a variety of peak-calling algorithms (Supplementary Figure 5.2).

5.4.5.3 PCR

- **ChIPulate:** models the amplification efficiency of each region as well as the number of PCR cycles. Of the three tools, the ChIPulate PCR model is most advanced but also most computationally intensive.
- **isChIP:** copies each fragment 2n times, where n is the number of PCR cycles. In practice, because the total number of fragments exponentially increases, this process explodes after several cycles. In our simulation experiments, we have found that using even small numbers of PCR cycles (e.g., 10) results in unreliable output.
- **ChIPs:** models the distribution of the number of duplicate reads. This parameter can be easily inferred from existing data and used to simulate realistic PCR duplicate patterns.

5.4.5.4 Sequencing

- **ChIPulate:** does not model sequencing errors.
- **isChIP:** does not model sequencing errors.
- **ChIPs:** models base substitution and indel rates based on real Illumina data.

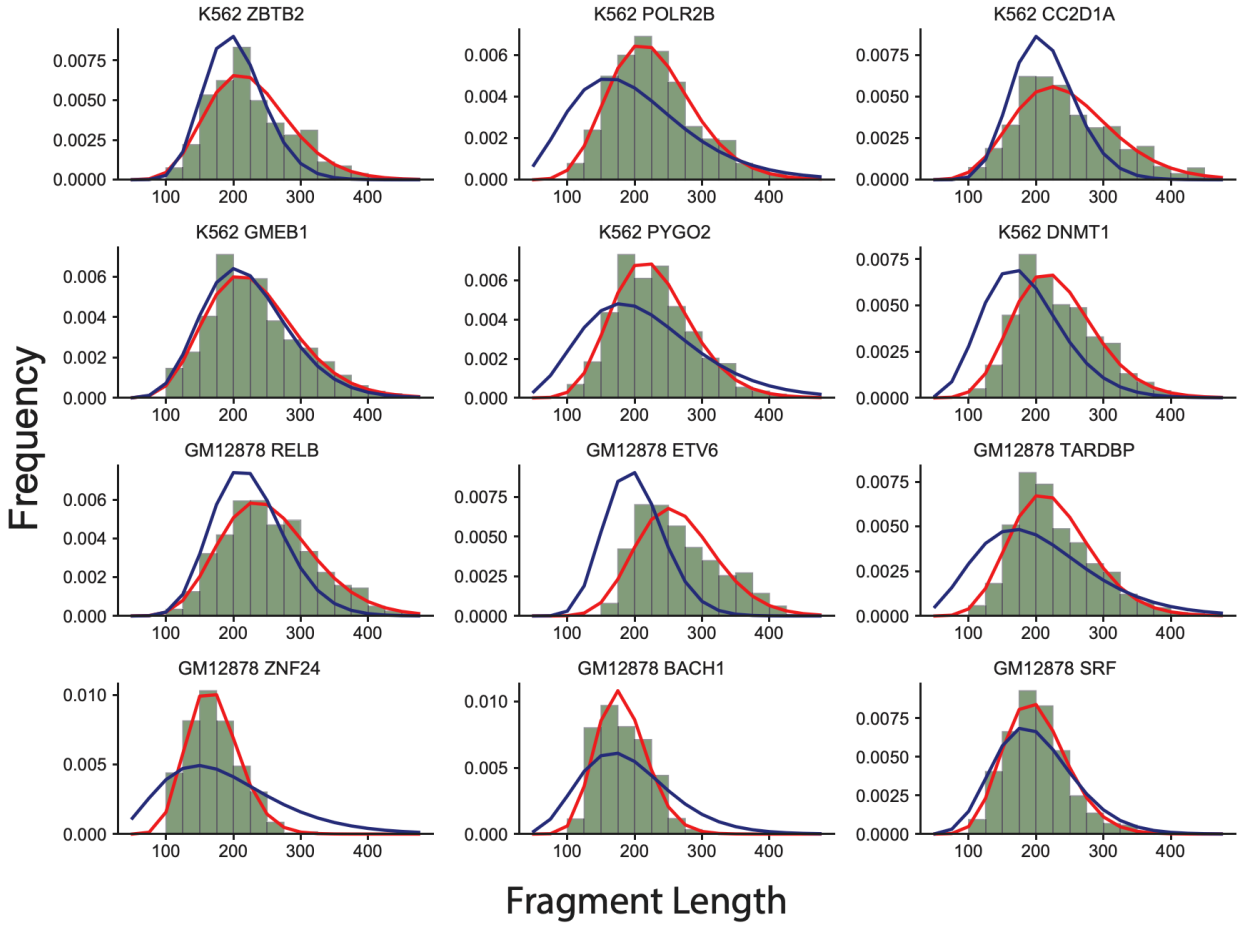
5.5 Conclusions

In summary, we present ChIPs, an efficient command-line program that can rapidly generate realistic ChIP-seq data over a wide range of experimental conditions. ChIPs can infer model parameters from real data and generate simulated data for both TF and HMs. The whole process takes just seconds to minutes for most applications. Our framework is modular, allowing future integration of alternative or improved models at various simulation steps. For example, we can further model multiple types of biases, such as the ones introduced by specific cross-linking steps. Or we can model the biases introduced during pulldown by inherent factors such as GC content or DNA accessibility.

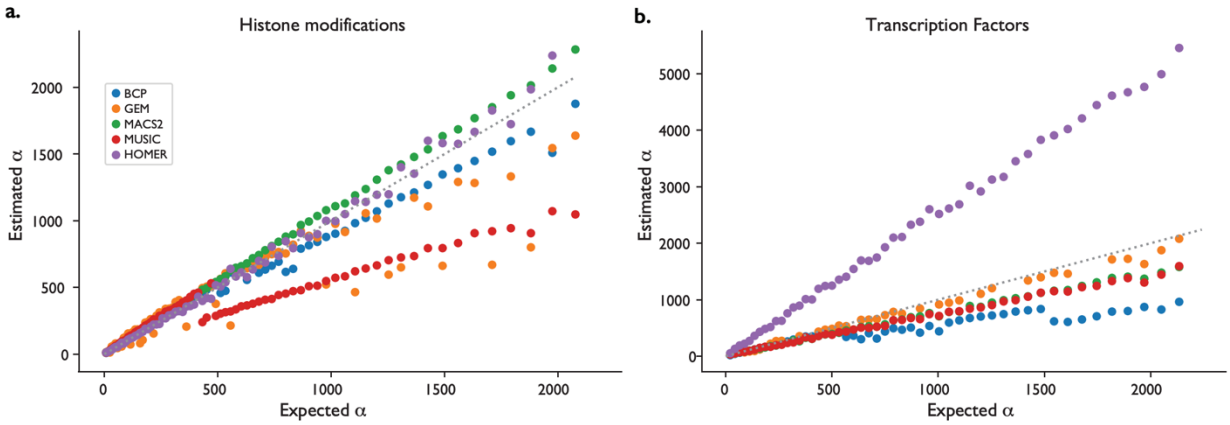
In this study, we benchmarked ChIPs against existing simulation tools and compared simulation results with a broad range of real ChIP-seq datasets as ground-truth. While all these tools could model multiple aspects of ChIP-seq data, we found that ChIPs most closely captures the properties of real ChIP-seq datasets. Another advantage of ChIPs is that, among all simulation tools benchmarked in this study, ChIPs is the only method capable of inferring model parameters from real data, allowing realistic simulation.

We demonstrated the utility of ChIPs in several usage scenarios, including benchmarking peak calling methods and measuring the effects of experimental conditions on peak detection. Some potential future applications include (1) evaluating the effects of genetic variation, such as SNPs, indels, or repeats, on observed ChIP-seq signals, (2) modeling effects of biological processes, such as DNA replication, on ChIP-seq signals, and (3) analyzing effects of spike-in normalization controls. Overall, we envision our framework will serve as a valuable resource for future efforts in ChIP-seq analysis.

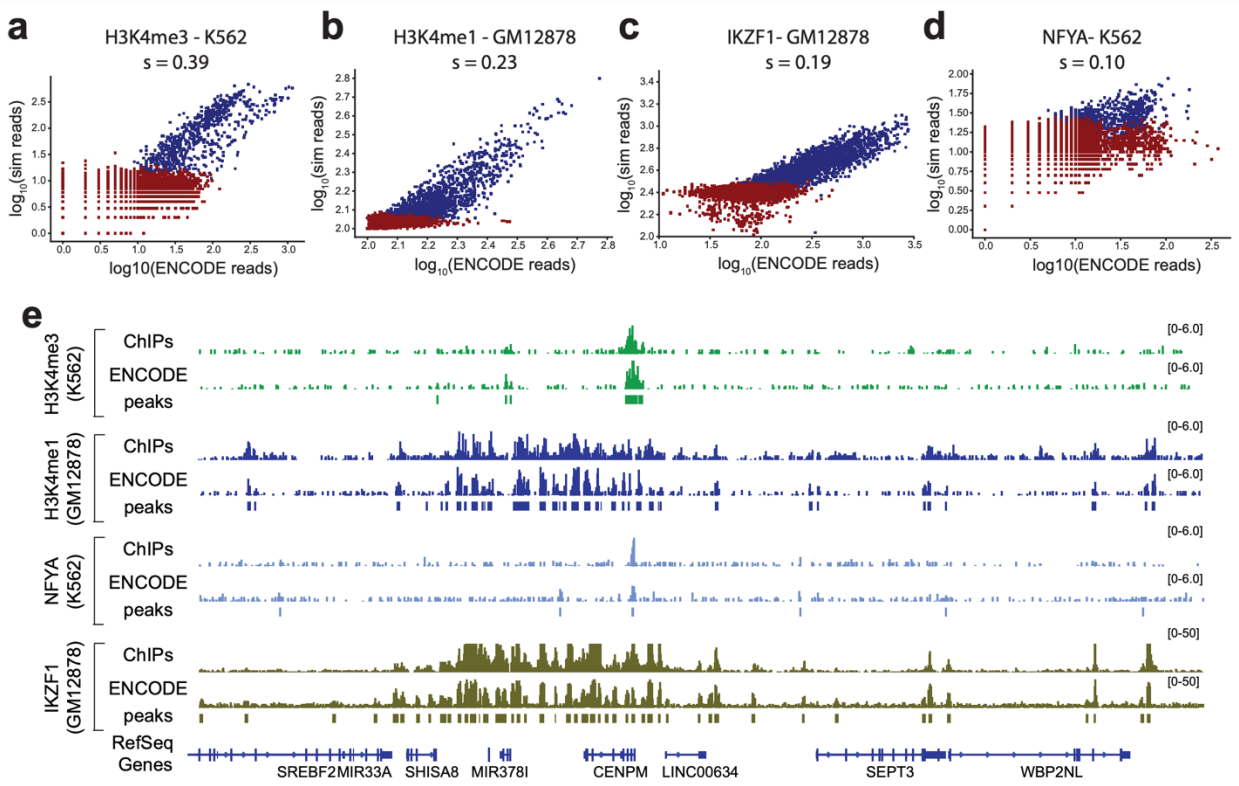
5.6 Supplementary figures



Supplementary Figure 5.1: Inferring fragment length distributions from single-end reads. Green bars show a histogram of lengths of 10,000 randomly chosen fragments from GM12878 paired end ChIP-seq experiments. Red lines give the best fit gamma distribution learned using observed fragment lengths. Blue lines give the fit inferred ignoring pair information using our novel method for learning fragment length distributions from single end data. ENCODE accessions are given in Supplementary Table 5.3.

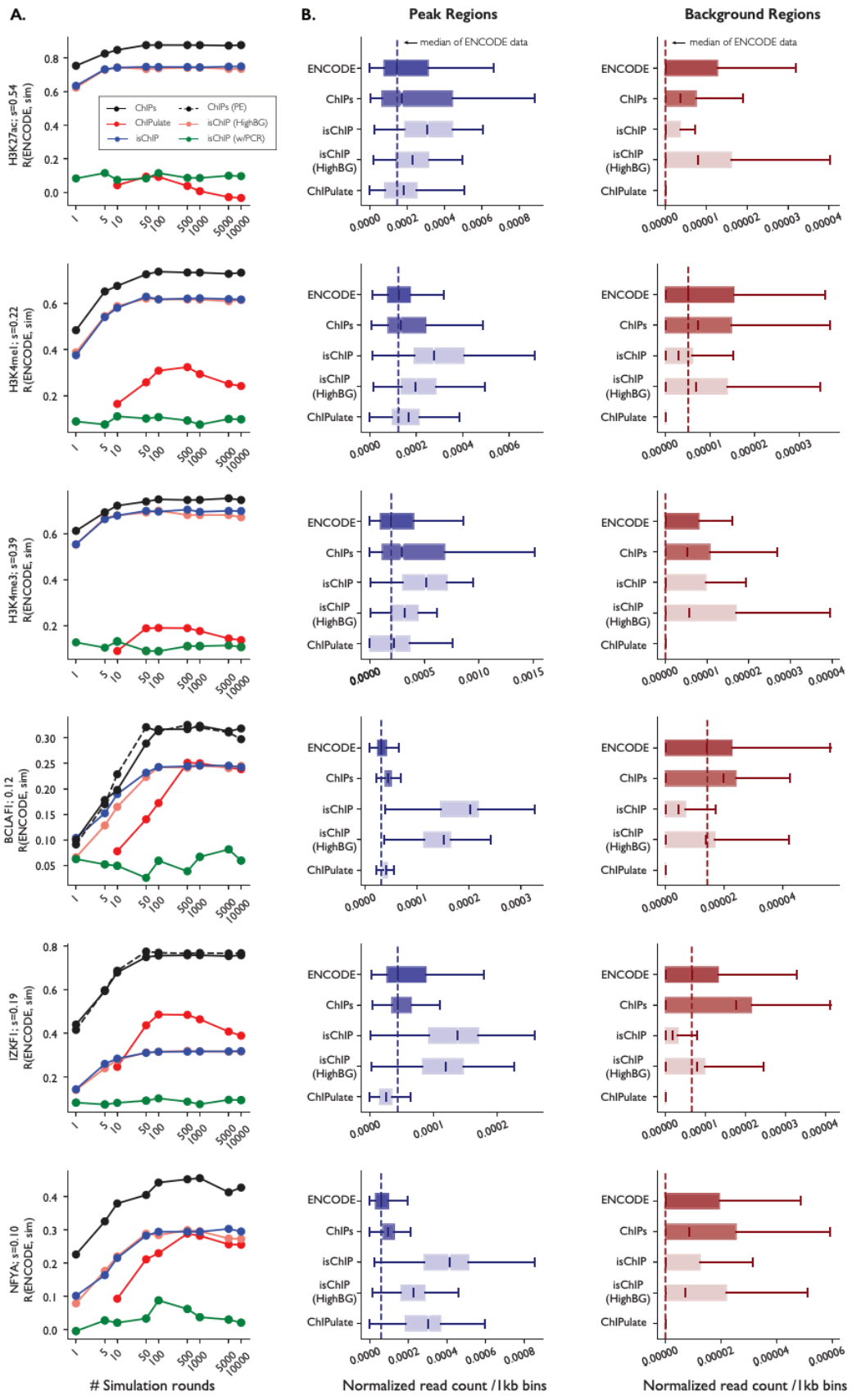


Supplementary Figure 5.2: Evaluation of pulldown parameter estimation with input TF and HM peaks from different peak callers. Each dot in the plots indicates an individual experiment and shows the comparison between estimated and expected alpha. Alpha is the ratio of pulldown probabilities of bound regions to unbound regions as described in the Supplementary Methods. The expected alpha was used in simulating reads and the estimated alpha was computed using our ChIPs learn module. Blue=BCP; orange=GEM; green=MACS2; red=MUSIC; purple=HOMER. In (a), (HM), α estimated based on MUSIC peaks becomes unreliable for simulated datasets with high s (SPOT) scores. This issue can be mitigated by applying the "scale-outliers" option in the learn module which reduces the effect of outlier peaks. In (b), (TF), the estimated alpha based on HOMER peaks is higher than expected. This is because HOMER is more stringent in deciding peak boundaries: the peak length from HOMER (mean=323.72bp) is smaller than the others (mean=746.22bp) and the ground truth (mean=448.80bp), resulting in the f value being underestimated and the α value being overestimated.

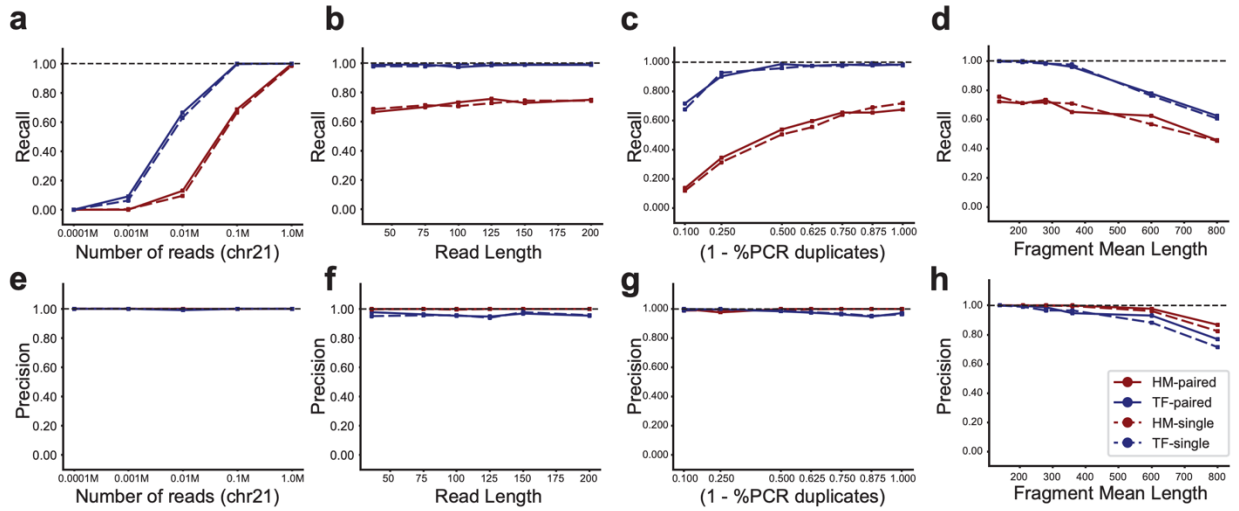


Supplementary Figure 5.3: (a)-(d) Concordance of read counts between simulated vs. real ChIP-seq data. chr22 was divided into nonoverlapping 5kb bins. The scatter plots show the comparison of read counts per bin for bins overlapping peaks (dark blue) or background regions (dark red). The x- and y-axes are on a log10 scale. The plot shown is for 100 simulated genome copies. **(e) Examples of coverage profiles of real vs. simulated data.**

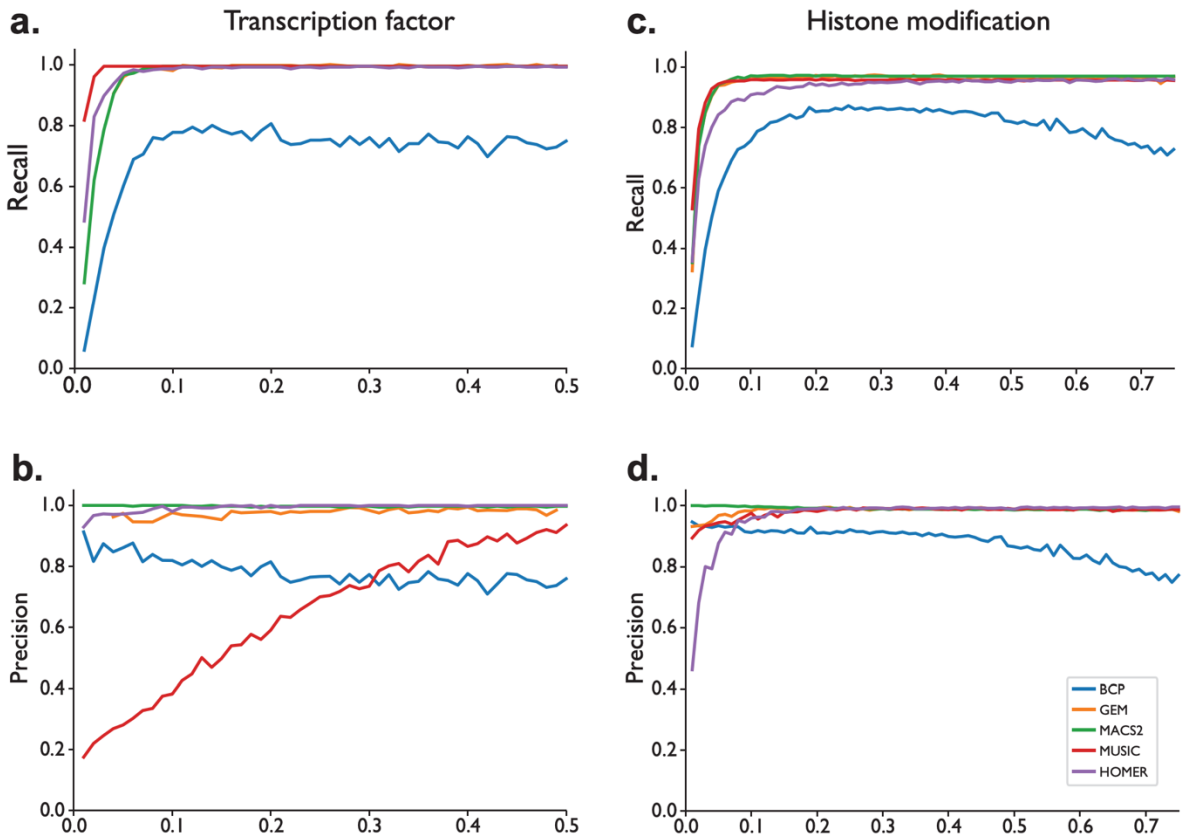
Supplementary Figure 5.4: Benchmarking of ChIPs against existing simulators. Evaluation of ChIPs against existing methods was performed by comparing the simulation results of each to real ChIP-sequencing datasets covering a range of dataset types (both histone modifications and transcription factors), sequencing settings (single vs. paired end reads), and data qualities as measured by SPOT scores (s). a. The Pearson correlation between read counts of simulated vs. real (ENCODE) datasets for histone modifications and transcription factors was measured across a range of simulation rounds. isChIP was run with multiple parameter choices (isChIP: low background and no PCR, isChIP w/PCR: low background and 10 PCR cycles, isChIP HighBG: high background and no PCR). For datasets where paired-end data was available (BLCAF1 and IZKF1), we evaluated ChIPs using models learned from paired end data (dashed lines) and ignoring paired end information (solid lines). All other datasets are single-end. b. Box plots of the relative number of reads in peaks vs. background regions in simulated vs. real data per 1kb bins. Left (blue) are the peak regions, right (red) are the background regions. The median of the real (ENCODE) data is denoted as a dashed line in each plot. Read counts in each bin were normalized by the total number of reads aligned to chr22 for each dataset.


A.
B.
Peak Regions
Background Regions
median of ENCODE data
ChIPs
ChIPulate
isChIP
ChIPs (PE)
isChIP (HighBG)
isChIP (w/PCR)
ENCODE
isChIP (HighBG)
H3K27ac; s=0.54
R(ENCODE, sim)
H3K4me1; s=0.22
R(ENCODE, sim)
H3K4me3; s=0.39
R(ENCODE, sim)
BCLAF1; 0.12
R(ENCODE, sim)
IZKF1; s=0.19
R(ENCODE, sim)
NFYA; s=0.10
R(ENCODE, sim)
# Simulation rounds
Normalized read count /1kb bins
Normalized read count /1kb bins

Supplementary Figure 5.5: Visualization of coverage profiles for different ChIP-seq simulators. Examples of coverage profiles of real vs. simulated data (brown=ENCODE, red=ChIPs, olive=isChIP, teal=ChIPulate. Brown bars show peaks called in the ENCODE data). Coverage profiles may also be viewed interactively at https://tinyurl.com/yblk3gb2.



Supplementary Figure 5.6: Evaluation of the effects of varying experimental parameters on peak calling performance. Results are based on simulation of generic TF and HM datasets for chr21 as described in the chapter 5.4. For top plots the y-axis shows the recall (percent of real peaks inferred from simulated reads). For bottom plots the y-axis shows the precision (percent of inferred peaks that match simulated real peaks). (a) and (e) Recall and precision as a function of the total number of reads simulated from chr21. (b) and (f) Recall and precision as a function of read length. (c) and (g) Recall and precision as a function of PCR rate. The x-axis gives the parameter p, which can be interpreted as the percent of fragments with no PCR duplicates (Supplementary Table 5.2). (d) and (h) Recall and precision as a function of mean fragment length. Red=HM; Blue=TF; solid lines=paired end reads; dashed lines=single end reads.



Supplementary Figure 5.7: Evaluation of peak calling methods using simulated data. Noise levels are quantified using *s*, the fraction of pulled down reads that originate from true binding sites. Blue=BCP; orange=GEM; green=MACS2; red=MUSIC; purple=HOMER. Top plots show recall and bottom plots show precision.

5.7 Supplementary tables

Supplementary Table 5.1: Comparison of features in existing ChIP-seq simulation tools.

Feature	simchip	ChIPulate	isChIP	ChIPs
Simulates background noise	Yes	No	Yes	Yes
Learns parameters from real data	No	No	No	Yes
Simulates peaks at user-specified regions	No	Yes	Yes	Yes
Simulates paired end reads	No	Yes	Yes	Yes
Simulates sequencing errors	Yes	No	No	Yes
Simulates PCR	No	Yes	Yes	Yes
Fragment length model	NA	fixed (1kb)	log-normal	gamma
User interface R package	R package	command-line		
Approximate run time relative to ChIPs	-	16	0.16	1

Supplementary Table 5.2: ChIPs parameters learned or input by users.

Parameters learned from real data (learn module)			
Parameter	ChIP step	simreads argument	Description
k	Shearing	--gamma-frag float,float	Parameters of template lengths gamma distribution.
s	Pulldown	--spot float	SPOT score (percentage of reads falling in peaks).
f	Pulldown	--frac float	The fraction of the genome bound by the target factor.
p	PCR	--pcr_rate float	The percentage of fragments with no PCR duplicates.
User-specified parameters			
Parameter	ChIP step	simreads argument	Description
N	Input	--numcopies int	The number of rounds (genome copies) to simulate.
R	Sequencing	--numreads int	Target number of reads (or read pairs) to simulate.
l	Sequencing	--readlen int	Length of output reads.
	Sequencing	--paired	Whether to output paired vs. single end data.
e	Sequencing	--sub rate float	Sequencing error rate.
d	Sequencing	--del rate float	Sequencing deletion rate.
i	Sequencing	--ins rate float	Sequencing insertion rate.

Supplementary Table 5.3: ENCODE accessions for benchmark datasets. BAM files and peak files corresponding to each accession can be found on the ENCODE Project website: https://www.encodeproject.org/files/{accession}.

Cell type	Epitope	BAM accession	PEAK accession	Result
K562	ZBTB2	ENCFF294LPO	ENCFF551YGS	Supplementary Figure 1
K562	POLR2B	ENCFF454PCU	ENCFF729LTY	Supplementary Figure 1
K562	CC2D1A	ENCFF054TBR	ENCFF051PEB	Supplementary Figure 1
K562	GMEB1	ENCFF809QWF	ENCFF154QJU	Supplementary Figure 1
K562	PYGO2	ENCFF288JVH	ENCFF078LXS	Supplementary Figure 1
K562	DNMT1	ENCFF987HMB	ENCFF958LLL	Supplementary Figure 1
GM12878	RELB	ENCFF708KIW	ENCFF355VTC	Supplementary Figure 1
GM12878	ETV6	ENCFF425VPI	ENCFF959JZX	Supplementary Figure 1
GM12878	TARDBP	ENCFF673WUM	ENCFF016QUV	Supplementary Figure 1
GM12878	ZNF24	ENCFF699QHD	ENCFF882PND	Supplementary Figure 1
GM12878	BACH1	ENCFF518TTP	ENCFF866OLZ	Supplementary Figure 1
GM12878	SRF	ENCFF387RFR	ENCFF500GHH	Supplementary Figure 1
GM12878	H3K27ac	ENCFF097SQI	ENCFF465WTH	Figure 1, Supplementary Figure 4
GM12878	IKZF1	ENCFF216YZE	ENCFF795PEX	Supplementary Figure 3,4
GM12878	H3K4me1	ENCFF252ZII	ENCFF966LMJ	Supplementary Figure 3,4
K562	H3K4me3	ENCFF681JQI	ENCFF127XXD	Supplementary Figure 3,4
K562	NFYA	ENCFF000YUR	ENCFF003WYE	Supplementary Figure 3,4
GM12878	BCLAF1	ENCFF671NSO	ENCFF222GJV	Supplementary Figure 4

5.8 Acknowledgements

We thank Dr. Christopher Benner and Bing Ren for helpful discussions of the method. We thank Rahel Wachs for assistance in preparing figures. We also thank the American Society of Human Genetics (ASHG) Annual Meeting for publishing the abstract online for a poster session. This work was supported in part by NIH/NHGRI Grant 1R21HG010070 to AG and MG. The funding agency did not participate in design or implementation of this work.

This chapter is a reformatted version of the material as it appears in "A flexible ChIP-sequencing simulation toolkit," An Zheng, Michael Lamkin, Yutong Qiu, Kevin Ren, Alon Goren, and Melissa Gymrek. (Zheng et al., 2021, *BMC bioinformatics*) The material has been published

on BMC Bioinformatics. The dissertation author was the primary investigator and author of this material.

CHAPTER 6

Conclusions

In this dissertation, I presented AgentBind, a deep learning framework leveraging neural network architectures and state of the art model interpretation techniques to identify, visualize, and interpret sequence features predictive of regulatory activities. AgentBind use a two-step transfer learning scheme, enable it to accommodate datasets as small as 100 samples. This framework applies Grad-CAM, a post-analytical method for neural networks, to compute importance scores for each nucleotide in the input sequences and characterize sequence features potentially associated with biological functions and human genetic traits. And in chapter 1, I showed the applicability of AgentBind on genomic data and benchmarked both its classification and interpretation modules using a controlled simulated dataset with ground-truth available.

Next, I demonstrated how to use AgentBind in real-world biological tasks in chapter 2 – 4. In chapter 2, I presented a research work in which my collogues and I applied AgentBind to identify and interpret sequence context features most important for predicting whether a particular motif instance will be bound by the TF of interest. We applied our framework to predict binding at motifs for 38 TFs in a lymphoblastoid cell line, score the importance of context sequences at base-pair resolution, and characterize context features most predictive of binding. Another use case I presented was to use AgentBind to prioritize genetic variants associated with human brain disorder (chapter 3). In this work, we integrated the AgentBind importance score with fine-mapping results from GWAS and identified putative causal variants that may act via modulating enhancer activity. In chapter 4, I demonstrated a use case where my collogues and I applied AgentBind to epigenetic data for macrophages from five inbred strains of mice and identified the

dominant combinations of LDTFs and SDTFs influencing IL-4 enhancer activation. Our results uncover general mechanisms by which noncoding genetic variants influences signal-dependent enhancer activity, thereby contributing to strain-differential patterns of gene expression and phenotypic diversity.

These biological findings, together with the benchmarking experiments in chapter 1, suggest that AgentBind is a reliable deep learning framework that helps in decoding the rules within non-coding DNA sequences and identifying the regulatory functions of non-coding genetic variants. This framework has been consistently shown to accurate and flexible in multiple studies across various types of biological activities, cells, and species. It has good potential to be extended to more applications and enables novel insights into regulation functions of non-coding DNA.

REFERENCES

Abadi M., Agarwal. A, Barham P., Brevdo E., Chen Z., Citro C., Corrado G. S., Davis A, Dean J., Devin M., Ghemawat S., Goodfellow I., Harp A., Irving G., Isard M, Jozefowicz R, Jia Y., Kaiser L, Kudlur M., Levenberg J., Mané D., Schuster M., Monga R., Moore S., Murray D., Olah C., Shlens J., Steiner B., Sutskever I., Talwar K., Tucker P., Vanhoucke V., Vasudevan V., Viégas F., Vinyals O., Warden P., Wattenberg M., Wicke M., Yu Y., and Zheng X.. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. *Software available from tensorflow.org*.

Amariuta, T., Luo, Y., Gazal, S., Davenport, E. E., van de Geijn, B., Ishigaki, K., Westra, H., Teslovich, N., Okada, Y., Yamamoto, K., RACI Consortium, GARNET Consortium, Price, A. L., & Raychaudhuri, S. (2019). IMPACT: genomic annotation of cell-state-specific regulatory elements inferred from the epigenome of bound transcription factors. *The American Journal of Human Genetics*, 104(5), 879-895.

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8), 831-838.

Avsec, Ž., Kreuzhuber, R., Israeli, J., Xu, N., Cheng, J., Shrikumar, A., Banerjee, A., Kim D. S., Beier T., Urban L., Kundaje A., Stegle O. & Gagneur, J. (2019). The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nature biotechnology*, 37(6), 592-600.

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R, McAnany, C., Gagneur, J., Kundaje, A., & Zeitlinger, J. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3), 354-366.

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.

Bansal, V., Mitjans, M., Burik, C. A., Linner, R. K., Okbay, A., Rietveld, C. A., Begemann, M., Bonn, S., Ripke, S., de Vlaming, R., Nivard, M. G., Ehrenreich, H.& Koellinger, P. D. (2018). Genome-wide association study results for educational attainment aid in identifying genetic heterogeneity of schizophrenia. *Nature communications*, 9(1), 1-12.

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), 1-15.

Benner, C., Konovalov, S., Mackintosh, C., Hutt, K. R., Stunnenberg, R., & Garcia-Bassets, I. (2013). Decoding a signature-based model of transcription cofactor recruitment dictated by cardinal cis-regulatory elements in proximal promoter regions. *PLoS genetics*, 9(11), e1003906.

Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S., & Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10), 1493-1501.

Bonn S., Zinzen R. P., Girardot C., Gustafson E. H., Perez-Gonzalez A., Delhomme N., Ghavi-Helm Y., Wilczyński 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.

Broad Institute. (2018). Picard Tools v2.18.11. *Software available from http://broadinstitute.github.io/picard*.

Brum, C. B., Paixão-Côrtes, V. R., Carvalho, A. M., Martins-Silva, T., Carpena, M. X., Ulguim, K. F., Luquez, K. Y. S., Salatino-Oliveira, A., & Tovo-Rodrigues, L. (2021). Genetic variants in miRNAs differentially expressed during brain development and their relevance to psychiatric disorders susceptibility. *The World Journal of Biological Psychiatry*, 22(6), 456-467.

Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Pérez, N. M., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., Vandepoele, K., Wasserman, W. W., Parcy, F., & Mathelier, A. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 50(D1), D165-D173.

Chen, K. S., Lim, J. W., Richards, L. J., & Bunt, J. (2017). The convergent roles of the nuclear factor I transcription factors in development and cancer. *Cancer letters*, 410, 124-138.

Chen, Z., Zhang, J., Liu, J., Dai, Y., Lee, D., Min, M. R., Xu, M., & Gerstein, M. (2021). DECODE: A De ep-learning Framework for Co n de nsing Enhancers and Refining Boundaries with Large-scale Functional Assays. *Bioinformatics*, 37(Supplement_1), i280-i288.

Chollet, F. (2015). keras. *Software available from https://keras.io*.

Corces, M. R., Shcherbina, A., Kundu, S., Gloudemans, M. J., Frésard, L., Granja, J. M., Louie, B. H., Eulalio, T., Shams, S., Bagdatli, S. T., Mumbach, M. R., Liu B., Montine K. S., Greenleaf W. J., Kundaje A., Montgomery S. B., Chang, H. Y. Chang & Montine, T. J. (2020). Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nature genetics*, 52(11), 1158-1168.

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*, 107(50), 21931-21936.

Czimmerer Z., Daniel B., Horvath A., Rückerl D., Nagy G., Kiss M., Peloquin M., Budai M. M., Cuaranta-Monroy I., Simandi Z., Steiner L., Jr B. N., Poliska S., Banko C., Bacso Z., Schulman I. G., Sauer S., Deleuze J.-F., Allen J. E., Benko S., 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.

Datta, V., Hannenhalli, S., Siddharthan, R. (2019). ChIPulate: A comprehensive ChIP-seq simulation pipeline. *PLoS Computational Biology*. 15(3), 1006921.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M. & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, 10(2), giab008.

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 & development*, 34(21-22), 1474-1492.

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.

Davies, G., Lam, M., Harris, S. E., Trampush, J. W., Luciano, M., Hill, W. D., ... & Stott, D. J. (2018). Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nature communications*, 9(1), 1-16.

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.

Deplancke, B., Alpern, D., & Gardeux, V. (2016). The genetics of transcription factor DNA binding variation. *Cell*, 166(3), 538-554.

Desikan, R. S., Schork, A. J., Wang, Y., Thompson, W. K., Dehghan, A., Ridker, P. M., Chasman, D. I., McEvoy, L. K., Holland, D., Chen, C., Karow, D. S., Brewer, J. B., Hess, C. P., Williams, J., Sims, R., O'Donovan, M. C., Choi, S. H., Bis, J. C., Ikram, M. A., Gudnason, V., DeStefano, A. L., van der Lee, S. J., Psaty, B. M., van Duijn, C. M., Launer, L., Seshadri, S., Pericak-Vance, M. A., Mayeux, R., Haines, J. L., Farrer, L. A., Hardy, J., Ulstein, I. D., Aarsland, D., Fladby, T., White, L. R., Sando, S. B., Rongve, A., Witoelar, A., Djurovic, S., Hyman, B. T., Snaedal, J.,

Steinberg, S., Stefansson, H., Stefansson, K., Schellenberg, G. D., Andreassen, O. A., & Dale, A. M. (2015). Polygenic overlap between C-reactive protein, plasma lipids, and Alzheimer disease. *Circulation*, 131(23), 2061-2069.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut P., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15-21.

Dolfini, D., Zambelli, F., Pedrazzoli, M., Mantovani, R., & Pavesi, G. (2016). A high definition look at the NF-Y regulome reveals genome-wide associations with selected transcription factors. *Nucleic acids research*, 44(10), 4684-4702.

Efthymiou, A. G., & Goate, A. M. (2017). Late onset Alzheimer's disease genetics implicates microglial pathways in disease risk. *Molecular neurodegeneration*, 12(1), 1-12.

El Jurdi, R., Petitjean, C., Honeine, P., & Abdallah, F. (2021). CoordConv-Unet: Investigating CoordConv for Organ Segmentation. *IRBM*, 42(6), 415-423.

ENCODE Project Consortium. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), 799.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57.

Encode Project Consortium, Snyder M. P., Gingeras T. R., Moore J. E., Weng Z., Gerstein M. B., Ren B., Hardison R. C., Stamatoyannopoulos J. A., Graveley B. R., Feingold E. A., Pazin M. J., Pagan M., Gilchrist D. A., Hitz B. C., Cherry J. M., Bernstein B. E., Mendenhall E. M., Zerbino D. R., Frankish A., Flicek P., Myers R. M. (2020). Perspectives on ENCODE. *Nature*, 583(7818), 693-698.

Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7), 389-403.

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), 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., Hafler, A. H., Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539), 337–343.

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.

Feingold, E. A., & Pachter, L. (2004). The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696), 636-640.

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.

French, J. D., & Edwards, S. L. (2020). The role of noncoding variants in heritable disease. *Trends in Genetics*, 36(11), 880-891.

Furey, T. S. (2012). ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, 13(12), 840-852.

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., Beer M. A., Ye C. J., Regev A. (2018). Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nature genetics*, 50(8), 1140-1150.

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.

Goenka, S., & Kaplan, M. H. (2011). Transcriptional regulation by STAT6. *Immunologic research*, 50(1), 87-96.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gordon, S., & Martinez, F. O. (2010). Alternative activation of macrophages: mechanism and functions. *Immunity*, 32(5), 593-604.

Gosselin, D., Skola, D., Coufal, N. G., Holtman, I. R., Schlachetzki, J. C., Sajti, E., Jaeger, B. N., O'Connor, C., Fitzpatrick, C., Pasillas, M. P., Pena, M., Adair, A., Gonda, D. D., Levy, M. L., Ransohoff, R. M., Gage, F. H. & Glass, C. K. (2017). An environment-dependent transcriptional network specifies human microglia identity. *Science*, 356(6344), eaal3222.

Gotea, V., Visel, A., Westlund, J. M., Nobrega, M. A., Pennacchio, L. A., & Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome research*, 20(5), 565-577.

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7), 1017-1018.

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.

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*, 114(7), E1291-E1300.

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.

Guo, Y., Mahony, S., & Gifford, D. K. (2012). High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS computational biology*, 8(8), 1002638.

Guo, Y., Tian, K., Zeng, H., Guo, X., & Gifford, D. K. (2018). A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction. *Genome research*, 28(6), 891-900.

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. *Genome biology*, 8(2), 1-9.

Harmanci, A., Rozowsky, J., & Gerstein, M. (2014). MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome biology*, 15(10), 1-15.

Hauberg, M. E., Holm-Nielsen, M. H., Mattheisen, M., Askou, A. L., Grove, J., Børglum, A. D., & Corydon, T. J. (2016). Schizophrenia risk variants affecting microRNA function and site-specific regulation of NT5C2 by miR-206. *European Neuropsychopharmacology*, 26(9), 1522-1526.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

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.

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(7477), 487-492.

Hoeksema, M. A., Shen, Z., Holtman, I. R., Zheng, A., Spann, N. J., Cobo, I., Gymrek, M., & Glass, C. K. (2021). Mechanisms underlying divergent responses of genetically distinct macrophages to IL-4. *Science advances*, 7(25), eabf9808.

Holtman, I. R., Skola, D., & Glass, C. K. (2017). Transcriptional control of microglia phenotypes in health and disease. *The Journal of clinical investigation*, 127(9), 3220-3229.

Huang J. T., Welch J. S., Ricote M., Binder C. J., Willson T. M., Kelly C., Witztum J. L., Funk C. D., Conrad D., Glass C. K. (1999). Interleukin-4-dependent production of PPAR-γ ligands in macrophages by 12/15-lipoxygenase. *Nature*, 400(6742), 378-382.

Humburg, P., Helliwell, C.A., Bulger, D., Stone, G. (2011). ChIPseqR: analysis of ChIP-seq experiments. *BMC Bioinformatics*, 12, 39.

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., Karlsson, I. K., Hägg, S., Athanasiu, L., Voyle, N., Proitsi, P., Witoelar, A., Stringer, S., Aarsland, D., Almdahl, I. S., Andersen, F., Bergh, S., Bettella, F., Bjornsson, S., Brækhus, A., Bråthen, G., de Leeuw, C., Desikan, R. S., Djurovic, S., Dumitrescu, L., Fladby, T., Hohman, T. J., Jonsson, P. V., Kiddle, S. J., Rongve, A., Saltvedt, I., Sando, S. B., Selbæk, G., Shoai, M., Skene, N. G., Snaedal, J., Stordal, E., Ulstein, I. D., Wang, Y., White, L. R., Hardy, J., Hjerling-Leffler, J., Sullivan, P. F., van der Flier, W. M., Dobson, R., Davis, L. K., Stefansson, H., Stefansson, K., Pedersen, N. L., Ripke, S., Andreassen, O. A. & Posthuma, D. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature genetics*, 51(3), 404-413.

Jun, G. R., Chung, J., Mez, J., Barber, R., Beecham, G. W., Bennett, D. A., ... & Pankratz, V. S. (2017). Transethnic genome‐wide scan identifies novel Alzheimer's disease loci. *Alzheimer's & Dementia*, 13(7), 727-738.

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.

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., … Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M. E., Genome Aggregation Database Consortium, Neale B. M., Daly M. J., MacArthur D. G.. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809), 434–443.

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., Oliver P. L., McIntyre R. E., Bhomra A., Nicod J., Gan X., Yuan W., van der Weyden L., Steward C. A., Bala S., Stalker J., Mott R., Durbin R., Jackson I. J., Czechanski A., Guerra-Assunção J. A., Donahue L. R., Reinholdt L. G., Payseur B. A., Ponting C. P., Birney E., Flint J., Adams D. J. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364), 289-294.

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., & Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, 28(5), 739-750.

Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7), 990-999.

Kichaev, G., Roytman, M., Johnson, R., Eskin, E., Lindstroem, S., Kraft, P., & Pasaniuc, B. (2017). Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics*, 33(2), 248-255.

Kichaev, G., Yang, W. Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., & Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS genetics*, 10(10), e1004722.

Kidder, B.L., Hu, G., Zhao, K. (2011). ChIP-Seq: technical considerations for obtaining high-quality data. *Nature Immunology*. 12(10), 918–922.

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., Hacker D., Core L. J., Lis J. T., Hernandez N., Reymond A., Deplancke B., Dermitzakis E. T. (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, 342(6159), 744-747.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L. & Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research, 22(3), 568-576.

Kong, A., Frigge, M. L., Thorleifsson, G., Stefansson, H., Young, A. I., Zink, F., Jonsdottir, G. A., Okbay, A., Sulem, P., Masson, G., Gudbjartsson, D. F., Helgason, A., Bjornsdottir, G., Thorsteinsdottir, U., & Stefansson, K. (2017). Selection against variants in the genome associated with educational attainment. *Proceedings of the National Academy of Sciences*, 114(5), E727-E732.

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, 357-379.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kröger, A. (2017). IRFs as competing pioneers in T-cell differentiation. *Cellular & Molecular Immunology*, 14(8), 649-651.

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.

Kunkle, Brian W., et al. "Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing." *Nature genetics*, 51.3 (2019): 414-430.

Lai, B., Qian, S., Zhang, H., Zhang, S., Kozlova, A., Duan, J., He, X. & Xu, J. (2021). Predicting Epigenomic Functions of Genetic Variants in the Context of Neurodevelopment via Deep Transfer Learning. *bioRxiv*.

Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., ... & Nalls, M. A. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature genetics*, 45(12), 1452-1458.

Lane, J. M., Jones, S. E., Dashti, H. S., Wood, A. R., Aragam, K. G., van Hees, V. T., Strand, L. B., Winsvold, B. S., Wang, H., Bowden, J., Song, Y., Patel, K., Anderson, S. G., Beaumont, R. N., Bechtold, D. A., Cade, B. E., Haas, M., Kathiresan, S., Little, M. A., Luik, A. I., Loudon, A. S., Purcell, S., Richmond, R. C., Scheer, F. A. J. L., Schormair, B., Tyrrell, J., Winkelman, J. W., Winkelmann, J., HUNT All In Sleep, Hveem, K., Zhao, C., Nielsen, J. B., Willer, C. J., Redline, S., Spiegelhalder, K., Kyle, S. D., Ray, D. W., Zwart, J., Brumpton, B., Frayling, T. M., Lawlor, D. A., Rutter, M. K., Weedon, M. N. & Saxena, R. (2019). Biological and clinical insights from genetics of insomnia symptoms. *Nature genetics*, 51(3), 387-393.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), 357-359.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., R. Hughes, T. R. & Weirauch, M. T. (2018). The human transcription factors. *Cell*, 172(4), 650-665.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K.I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A.J., Hoffman, M.M., Iyer, V.R., Jung, Y.L., Karmakar, S., Kellis, M., Kharchenko, P.V., Li, Q., Liu, T., Liu, X.S., Ma, L., Milosavljevic, A., Myers, R.M., Park, P.J., Pazin, M.J., Perry, M.D., Raha, D., Reddy, T.E., Rozowsky, J., Shoresh,

N., Sidow, A., Slattery, M., Stamatoyannopoulos, J.A., Tolstorukov, M.Y., White, K.P., Xi, S., Farnham, P.J., Lieb, J.D., Wold, B.J., Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*. 22(9), 1813–1831.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1), 1-13.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), 357-359.

Le, D. D., Shimko, T. C., Aditham, A. K., Keys, A. M., Longwell, S. A., Orenstein, Y., & Fordyce, P. M. (2018). Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proceedings of the National Academy of Sciences*, 115(16), E3702-E3711.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., & Beer, M. A. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nature genetics*, 47(8), 955-961.

Lek, M., Karczewski, K. J., Minikel, E. V, Samocha, K. E., Banks, E., Fennell, T., O'DonnellLuria, 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.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013). *arXiv*,1303.3997.

Li, Q., Brown, J. B., Huang, H., & Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *The annals of applied statistics*, 5(3), 1752-1779.

Li, P., Spolski, R., Liao, W., Wang, L., Murphy, T. L., Murphy, K. M., & Leonard, W. J. (2012). BATF–JUN is critical for IRF4-mediated transcription in T cells. *Nature*, 490(7421), 543-546.

Li, M., Santpere, G., Imamura Kawasawa, Y., Evgrafov, O. V., Gulden, F. O., Pochareddy, S., Sunkin, S. M. Li, Z., Shin, Y., Zhu, Y., Sousa, A. M. M., Werling, D. M., Kitchen, R. R., Kang, H., Pletikos, M., Choi, J., Muchnik, S., Xu, X., Wang D., Lorente-Galdos, B., Liu, S., Giusti-Rodríguez, P., Won, H., de Leeuw, C. A., Pardiñas, A. F., BrainSpan Consortium, PsychENCODE Consortium, PsychENCODE Developmental Subgroup, Hu, M., Jin, F., Li, Y., Owen, M. J., O'Donovan, M. C., Walters, J. T. R., Posthuma, D., Reimers, M. A., Levitt, P., Weinberger, D. R., Hyde, T. M., Kleinman, J. E., Geschwind, D. H., Hawrylycz, M. J., State, M. W., Sanders, S. J., Sullivan, P. F., Gerstein, M. B., Lein, E. S., Knowles, J. A. & Sestan, N. (2018). Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science*, 362(6420), eaat7615.

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., Ren B., 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.

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.

Liu, J. Z., Erlich, Y., & Pickrell, J. K. (2017). Case–control association mapping by proxy using family history of disease. *Nature genetics*, 49(3), 325-331.

Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., & Yosinski, J. (2018). An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31.

Liu, G., Zeng, H., & Gifford, D. K. (2019). Visualizing complex feature interactions and feature sharing in genomic deep neural networks. *BMC bioinformatics*, 20(1), 1-14.

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.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass I., Sunyaev S.R., Kaul, R. & Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099), 1190-1195.

Marioni, R. E., Harris, S. E., Zhang, Q., McRae, A. F., Hagenaars, S. P., Hill, W. D., Davies, G., Ritchie, C. W., Gale, C. R., Starr, J. M., Goate, A. M., Porteous, D. J., Yang, J., Evans, K. L., Deary, I. J., Wray, N. R. & Visscher, P. M. (2018). GWAS on family history of Alzheimer's disease. *Translational psychiatry*, 8(1), 1-7.

Masuda, T., Tsuda, M., Yoshinaga, R., Tozaki-Saitoh, H., Ozato, K., Tamura, T., & Inoue, K. (2012). IRF8 is a critical transcription factor for transforming microglia into a reactive phenotype. *Cell reports*, 1(4), 334-340.

Mevel, R., Draper, J. E., Lie-a-Ling, M., Kouskoff, V., & Lacaud, G. (2019). RUNX transcription factors: orchestrators of development. *Development*, 146(17), dev148296.

Morgunova, E., & Taipale, J. (2017). Structural perspective of cooperative transcription factor binding. *Current opinion in structural biology*, 47, 1-8.

Meyer, C.A., Liu, X.S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Review Genetics*. 15(11), 709–721.

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.

Nair, S., Kim, D. S., Perricone, J. & Kundaje, A. (2019). Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinformatics*, 35(14), i108-i116.

Nord, A. S., & West, A. E. (2020). Neurobiological functions of transcriptional enhancers. *Nature neuroscience*, 23(1), 5-14.

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., Gonda D. D., Levy M. L., Rosenfeld M. G., McVicker G., Gage F. H., Ren B., Glass C. K. (2019). Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science*, 366(6469), 1134-1139.

Novakovsky, G., Saraswat, M., Fornes, O., Mostafavi, S. & Wasserman, W. W. (2021). Biologically relevant transfer learning improves transcription factor binding prediction. *Genome biology*, 22(1), 1-25.

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., ... & Rustichini, A. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, 533(7604), 539-542.

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.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., and Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pennisi, E. (2012). ENCODE project writes eulogy for junk DNA. *Science*, 337(6099), 1159-1161.

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94(4), 559-573.

Pong, S., Karmacharya, R., Sofman, M., Bishop, J. R., & Lizano, P. (2020). The role of brain microvascular endothelial cell and blood-brain barrier dysfunction in schizophrenia. *Complex Psychiatry*, 6(1-2), 30-46.

Quang, D., & Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic acids research*, 44(11), e107-e107.

Quang, D., & Xie, X. (2019). FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods*, 166, 40-47.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841-842.

Ritchie, M. E., Phipson, B., Wu, D. I., 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.

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.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology*, 29(1), 24-26.

Roder, K., Wolf, S. S., Larkin, K. J., & Schweizer, M. (1999). Interaction between the two ubiquitously expressed transcription factors NF-Y and Sp1. *Gene*, 234(1), 61-69.

Rosoff, D. B., Clarke, T. K., Adams, M. J., McIntosh, A. M., Davey Smith, G., Jung, J., & Lohoff, F. W. (2021). Educational attainment impacts drinking behaviors and risk for alcohol dependence: results from a two-sample Mendelian randomization study with~ 780,000 participants. *Molecular psychiatry*, 26(4), 1119-1132.

Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R., Ali, S., Chin, S.F., Palmieri, C., Caldas, C., Carroll, J.S. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481(7381), 389–393.

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8), 491-504.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510), 421.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I. K., Wang, J. Z., & Widom, J. (2006). A genomic code for nucleosome positioning. *Nature*, 442(7104), 772-778.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

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, i84-i92.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1), 308-311.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017, July). Learning important features through propagating activation differences. In *International conference on machine learning* (pp. 3145-3153). PMLR.

Shrikumar, A., Tian, K., Avsec, Ž., Shcherbina, A., Banerjee, A., Sharmin, M., Nair, S., & Kundaje, A. (2018). Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5. *arXiv*.

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.

Spielmann, M., & Mundlos, S. (2016). Looking beyond the genes: the role of non-coding variants in human disease. *Human molecular genetics*, 25(R2), R157-R165.

Spitz, F., & Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics*, 13(9), 613-626.

Subkhankulova T., Naumenko F., Tolmachov O. E., Orlov Y. L. (2021). Novel ChIP-seq simulating program with superior versatility: isChIP. *Briefings in Bioinformatics*, 22(4), bbaa352.

Sundaram, L., Gao, H., Padigepati, S. R., McRae, J. F., Li, Y., Kosmicki, J. A., Fritzilas N., Hakenberg J., Dutta A., Shon J., Xu J., Batzoglou S., Li X. & Farh, K. K. H. (2018). Predicting the clinical impact of human mutation with deep neural networks. *Nature genetics*, 50(8), 1161-1170.

Tan, M. S., Yang, Y. X., Xu, W., Wang, H. F., Tan, L., Zuo, C. T., Dong, Q., Tan, L., Suckling, J., Yu, J., & Alzheimer's Disease Neuroimaging Initiative. (2021). Associations of Alzheimer's disease risk variants with gene expression, amyloidosis, tauopathy, and neurodegeneration. *Alzheimer's Research & Therapy*, 13(1), 1-11.

Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., Andersen K. G., Mikkelsen, T. S., Lander, E. S., Schaffner S. F., & Sabeti, P. C. (2016). Direct identification of

hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, 165(6), 1519-1529.

Thomas, R., Thomas, S., Holloway, A. K., & Pollard, K. S. (2017). Features that define the best ChIP-seq peak calling algorithms. *Briefings in bioinformatics*, 18(3), 441-450.

Torrico, B., Shaw, A. D., Mosca, R., Vivó-Luque, N., Hervás, A., Fernàndez-Castillo, N., Aloy, P., Bayés, M., Fullerton, J. M., Cormand, B. & Toma, C. (2019). Truncating variant burden in high-functioning autism and pleiotropic effects of LRP1 across psychiatric phenotypes. *Journal of Psychiatry and Neuroscience*, 44(5), 350-359.

van Dam, H., & Castellazzi, M. (2001). Distinct roles of Jun: Fos and Jun: ATF dimers in oncogenesis. *Oncogene*, 20(19), 2453-2464.

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.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., Rando, O. J., Birney, E., Myers, R. M., Noble, W. S., Snyder, M & Weng, Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research*, 22(9), 1798-1812.

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., Deplancke, B., Dermitzakis, E. T. (2015). Population variation and genetic control of modular chromatin architecture in humans. *Cell*, 162(5), 1039-1050.

Westholm, J. O., Xu, F., Ronne, H., & Komorowski, J. (2008). Genome-scale study of the importance of binding site context for transcription factor binding and gene regulation. *BMC bioinformatics*, 9(1), 1-14.

Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., Telenti, A., & di Iulio, J. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nature communications*, 10(1), 1-9.

Wilczynska, K. M., Singh, S. K., Adams, B., Bryan, L., Rao, R. R., Valerie, K., Wright, S., Griswold-Prenner, Irene & Kordula, T. (2009). Nuclear factor I isoforms regulate gene expression during the differentiation of human neural progenitors to astrocytes. *Stem cells*, 27(5), 1173-1181.

Witoelar, A., Rongve, A., Almdahl, I. S., Ulstein, I. D., Engvig, A., White, L. R., Selbæk, G., Stordal, E., Andersen, F., Brækhus, A., Saltvedt, I., Engedal, K., Hughes, T., Bergh, S., Bråthen, G., Bogdanovic, N., Bettella, F., Wang, Y., Athanasiu, L., Bahrami, S., Le Hellard, S., Giddaluru, S., Dale, A. M., Sando, S. B., Steinberg, S., Stefansson, H., Snaedal, J., Desikan. R. S., Stefansson, K., Aarsland, D., Djurovic, S., Fladby, T. & Andreassen, O. A. (2018). Meta-analysis of

Alzheimer's disease on 9,751 samples from Norway and IGAP study identifies four risk loci. *Scientific reports*, 8(1), 1-8.

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.

Xing, H., Mo, Y., Liao, W., & Zhang, M. Q. (2012). Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. *PLoS computational biology*, 8(7), e1002613.

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.

Yin, J., Feng, W., Yuan, H., Yuan, J., Wu, Y., Liu, X., Jin, C., & Cheng, Z. (2019). Association analysis of polymorphisms in STARD6 and near ECHDC3 in Alzheimer's disease patients carrying the APOE ε4 Allele. *Neuropsychiatric Disease and Treatment*, 15, 213.

Yousefi, S., Amrollahi, F., Amgad, M., Dong, C., Lewis, J. E., Song, C., Gutman D. A., Halani S. H., Velazquez Vega J. E., Brat D. J. & Cooper, L. A. (2017). Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific reports*, 7(1), 1-11.

Yuan, Y., Shi, Y., Li, C., Kim, J., Cai, W., Han, Z., & Feng, D. D. (2016). DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. *BMC bioinformatics*, 17(17), 243-256.

Zaret, K. S., & Mango, S. E. (2016). Pioneer transcription factors, chromatin dynamics, and cell fate control. *Current opinion in genetics & development*, 37, 76-81.

Zeng, H., Hashimoto, T., Kang, D. D., & Gifford, D. K. (2016). GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics*, 32(4), 490-496.

Zhang, F., & Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Human molecular genetics*, 24(R1), R102-R110.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R., Brown, M., Li, W., & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9), 1-9.

Zhang, Z. D., Rozowsky, J., Snyder, M., Chang, J., Gerstein, M. (2008). Modeling ChIP sequencing in silico with applications. *PLoS Computational Biology*. 4(8), 1000158.

Zheng, A., Lamkin, M., Qiu, Y., Ren, K., Goren, A., & Gymrek, M. (2021). A flexible ChIP-sequencing simulation toolkit. *BMC bioinformatics*, 22(1), 1-10.

Zheng, A., Lamkin, M., Wu, C., Su, H., & Gymrek, M. (2021). AgentBind: Profiling Context-specific Determinants of Transcription Factor Binding Affinity. *ICML 2019 Workshop on Computational Biology*.

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.

Zhou, J., Park, C. Y., Theesfeld, C. L., Wong, A. K., Yuan, Y., Scheckel, C., Fak J. J., Funk J., Yao K., Tajima Y., Packer A., Darnell R. B. & Troyanskaya, O. G. (2019). Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nature genetics*, 51(6), 973-980.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S., Bussemaker, H. J., Gordân, R. & Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. *Proceedings of the National Academy of Sciences*, 112(15), 4654-4659.

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10), 931-934.

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), 1-10.

Zhu, G., & Kim, S. C. (2021). Coord-FCN for same-class objects segmentation. In *2021 International Conference on Information and Communication Technology Convergence (ICTC)* (pp. 1672-1674). IEEE.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature genetics*, 51(1), 12-18.

Zusso, M., Methot, L., Lo, R., Greenhalgh, A. D., David, S., & Stifani, S. (2012). Regulation of postnatal forebrain amoeboid microglial cell proliferation and development by the transcription factor Runx1. *Journal of Neuroscience*, 32(33), 11285-11298.