2020

Computational approaches to discover and characterize transcription regulatory complex binding from protein-binding microarray-based experiments

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

COMPUTATIONAL APPROACHES TO DISCOVER AND CHARACTERIZE TRANSCRIPTION REGULATORY COMPLEX BINDING FROM PROTEIN-BINDING MICROARRAY-BASED EXPERIMENTS

by

DAVID BRAY

B.S., McGill University, 2014

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2020

Approved by

First Reader ____
Trevor Siggers, Ph.D.
Associate Professor of Biology

Second Reader ____
Juan Fuxman Bass, Ph.D.
Assistant Professor of Biology

DEDICATION

For Cerberus “Cerbie” Glattly,

whose curiosity and creativity continue to inspire me.

ACKNOWLEDGMENTS

A big thank you to my family for their unconditional love and support. In particular, I would like to thank Mathieu for the graduate career advice and for the laughs. Thank you to my parents Lynn and Roy for supporting my decision to leave the country to pursue a degree in a field they had never heard of at the time. Thank you to Chantal and Mark for their support here in Boston. And thank you to Amanda for the music recommendations and for always being there for me to vent my frustrations to.

Thank you to all of my friends. In particular, thank you to Will Hackett and Dakota Hawkins for welcoming me into their home when I was technically homeless for a day. Thank you to Diana and Cerbie Glattly for their support during the first few years of the program and for helping me navigate some of the more difficult stretches of graduate school. Thank you to Alan Pacheco, Nahomie Rodriguez-Sastre, and Hannah Hulteen for their unwavering support and for helping me through the pandemic. And a big thank you to Eric Reed for quite literally showing me what bioinformatics research is all about.

Thank you to all of my thesis advisory committee members. Thank you to Drs. Andrew Emili and David J. Waxman for challenging me to be a better biologist and thank you to Dr. Juan Fuxman Bass for being an early believer in the technologies I was developing.

Thank you to the staff and administrators of the Bioinformatics Program. Thank you to Drs. Tom Tullius and Gary Benson for their commitment to fostering a fun and dynamic student culture.

Thank you to Dr. Adam Labadorf who inspired me to diversify my technical skillset and encouraged me to pursue and embrace the more entrepreneurial aspects of science. The hTF array and accompanying interactive analysis software would not exist without this inspiration.

Thank you to all of my labmates (past and present) for creating such a fun and inspiring work environment. In particular, it was a great privilege to get to work with such a phenomenal wet lab team including Ashley Penvose, Jessica Keenan, Rose Zhao, and Nima Mohaghegh. And a very big thank you to Heather Hook whose talent, dedication, and creativity really brought CASCADE to life.

And lastly, a big thank you to Dr. Trevor Siggers, my advisor and mentor, for fostering an environment of innovation these past few years. Thank you for supporting my wacky technology development ideas and for encouraging me to approach science with creativity. Your advice to "put on some Iron Maiden and power through it" will undoubtedly help me for years to come as I venture out into the post-graduate world.

COMPUTATIONAL APPROACHES TO DISCOVER AND CHARACTERIZE TRANSCRIPTION REGULATORY COMPLEX BINDING FROM PROTEIN-BINDING MICROARRAY-BASED EXPERIMENTS

DAVID BRAY

Boston University Graduate School of Arts and Sciences and College of Engineering,

2020

Major Professor: Trevor Siggers, Associate Professor of Biology

ABSTRACT

Gene regulation is controlled by DNA-bound complexes of transcription factors (TFs) and indirectly recruited transcriptional cofactors (COFs). Understanding how and where these TF-COF complexes bind in the genome is fundamental to our understanding of the role of cis-regulatory elements (CREs) in gene regulation and our mechanistic interpretation of non-coding variants (NCVs) known to impact gene expression levels. In this thesis, I present three related array-based techniques for the high-throughput profiling of DNA-bound TFs and TF-COF complexes directly from cell nuclear extracts.

First, I describe the nuclear extract protein-binding microarray (nextPBM) approach to profile TF-DNA binding using nuclear extracts to account for cell-specific post-translational modifications and cofactors. By analyzing cooperative binding of PU.1/SPI1 and IRF8 in monocytes, I demonstrate how nextPBM can be used to delineate DNA-sequence determinants of cell-specific cooperative TF complexes.

Second, I present the CASCADE (Comprehensive ASsessment of Complex Assembly at DNA Elements) approach to simultaneously discover DNA-bound TF-COF

complexes and quantify the impact of NCVs on their binding. To demonstrate applicability of CASCADE to screen NCVs, I profile differential TF-COF binding to ~1,700 single-nucleotide polymorphisms in human macrophages and discover a prevalence of perturbed ETS-related TF-COF complexes at these quantitative trait loci.

Third, I present the human TF array (hTF array) as a general platform for surveying COF recruitment to a panel of 346 non-redundant consensus TF binding sites (TFBSs). Using the hTF array, one can examine the activity of a diverse panel of TFs by profiling TF-COF complexes in a cell state-specific manner. In addition to the hTF microarray design, I have developed analysis and visualization software that allows users to explore COF recruitment profiling results interactively.

Collectively, nextPBM, CASCADE, and the hTF array represent a suite of new approaches to investigate TF-COF complex binding and their application will refine our understanding of CREs by linking NCVs with the biophysical complexes that mediate gene regulatory functions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A	Adenine
ATCC	American Type Culture Collection
ATI	Active Transcription Factor Identification
BG	Background
bp	Base Pair(s)
C	Cytosine
caQTL	Chromatin Accessibility Quantitative Trait Locus
CASCADE	Comprehensive Assessment of Complex Assembly at DNA Elements
cat	Catalog
CBP	CREB-Binding Protein
ChIP-seq	Chromatin Immunoprecipitation Followed by Sequencing
COF	Transcriptional Cofactor
CoRec	Cofactor Recruitment
CRE	cis-Regulatory Element
DMEM	Dulbecco’s Modified Eagle Medium
DNA	Deoxyribonucleic Acid
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic Acid
EICE	ETS-IRF Composite Element
EMSA	Electrophoretic Mobility Shift Assay

eQTL	Expression Quantitative Trait Locus
FBS	Fetal Bovine Serum
G	Guanine
GST	Glutathione S-Transferase
GWAS	Genome-Wide Association Study
H3K27ac	Histone 3 Lysine 27 Acetylation
H3K4me1	Histone 3 Lysine 4 Mono-Methylation
HAT	Histone Acetyltransferase
HBS	HEPES Buffered Saline
HDAC	Histone Deacetylase
HEPES	4-(2-Hydroxyethyl)-1-Piperazineethanesulfonic Acid
HT	High-Throughput
hTF	Human Transcription Factor
HT-SELEX	High-Throughput Systematic Evolution of Ligands by Exponential Enrichment
IDR	Irreproducible Discovery Rate
IFNγ	Interferon Gamma
IGEPAL	Octylphenoxypolyethoxyethanol
IP mass spec	Immunoprecipitation Mass Spectrometry
ISRE	Interferon-Stimulated Response Element
IVT	*In vitro* Transcribed/Translated
kb	kilobase

KO	Knockout
LDTF	Lineage-Determining Transcription Factor
LPS	Lipopolysaccharide
M2H	Mammalian Two-Hybrid
MPRA	Massively Parallel Reporter Assay
NCoR	Nuclear Receptor Co-Repressor
ncSNP	Non-Coding Single-Nucleotide Polymorphism
NCV	Non-Coding Variant
NE	Nuclear Extract
nextPBM	Nuclear Extract Protein-Binding Microarray
PBM	Protein-Binding Microarray
PBS	Phosphate Buffered Saline
PMA	Phorbol 12-Myristate 13-Acetate
PTM	Post-Translational Modification
PWM	Position-Weight Matrix
REF	Reference
RIPA	Radioimmunoprecipitation Assay
RNA	Ribonucleic Acid
RNA-seq	Ribonucleic Acid Sequencing
RPKM	Reads Per Kilobase Million
RPMI	Roswell Park Memorial Institute

SELEX-seq	Systematic Evolution of Ligands by Exponential Enrichment followed by Sequencing
SINGLE REP	Single Replicate
SNP	Single-Nucleotide Polymorphism
SNP-QTL	Single-Nucleotide Polymorphism Quantitative Trait Locus
SNV	Single-Nucleotide Variant
SV	Single Variant
T	Thymine
TCR	T-Cell Receptor
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TF-COF	Transcription Factor-Cofactor Complex
TF-TF	Cooperative Transcription Factor Complex
THP-1	Human Monocytic Cell Line
UT	Untreated
Y2H	Yeast Two-Hybrid

CHAPTER ONE: Introduction

1.1 The role of transcription factors in enhancer selection and gene regulation

Transcription factors (TFs) are a class of DNA-binding proteins that are expressed and activated in response to developmental or environmental cues to regulate the expression of their target genes. TFs in turn influence gene regulatory activities through interactions with non-DNA-binding proteins and protein complexes referred to collectively as transcriptional cofactors (COFs). These COFs that are indirectly recruited to DNA via TFs are considered effectors of gene regulation since many have enzymatic activity required to facilitate gene expression, such as modification of histones and remodeling of chromatin (Fig. 1.1) (Reiter et al., 2017; Zabidi and Stark, 2016; Shlyueva et al., 2014). TFs bind throughout the genome in a DNA sequence-specific manner. The DNA sequence preferences of TF binding are typically modeled using position-weight matrices (PWMs) that represent binding using the probability of observing a given nucleotide at a given position within the binding sites of the TF (Fig. 1.1) (Stormo and Fields, 1998; Siggers and Gordân, 2014). These DNA-binding preferences can also be visualized using sequence logos (Fig. 1.1) (Schneider and Stephens, 1990). As the genome across all cell types within an individual is largely identical and the nucleotide composition of a DNA sequence determines which TFs can bind a given segment, deciphering the logic of gene regulation in diverse cell types and stimulus responses fundamentally depends on our ability to determine which TFs are expressed and active in a given context, what their binding preferences are, and which COFs are subsequently recruited to these sites.

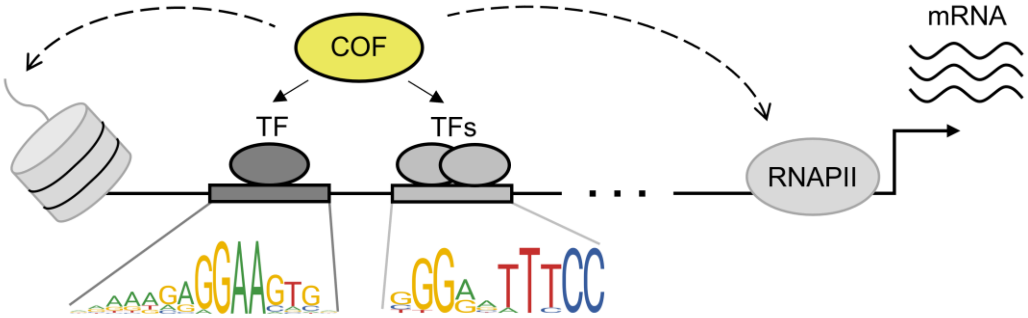


Figure 1.1: Overview of transcriptional regulation by TFs and COFs
COFs are recruited to TF sites on DNA (solid arrows) and have enzymatic activity such as histone modification and interaction with RNA polymerase II (RNAPII) and the general transcriptional machinery to influence how genes are expressed (mRNA). Different TFs display distinct DNA-binding preferences that can be modeled using PWMs and visualized using sequence logos.

TFs can be broadly classified by their functional properties and the role the play in gene regulation (Vaquerizas et al., 2009; Lambert et al., 2018; Smale, 2012). Many TFs for example are expressed in response to developmental cues and influence the regulatory potential of a given cell type or lineage (Heinz et al., 2013; Lin et al., 2010; Heinz et al., 2010; Johnson et al., 2018). For example, in monocytes and macrophages, important sentinel white blood cells of the immune system, the enhancer landscape is thought to be established by a small panel of lineage-determining transcription factors (LDTFs) including SPI1/PU.1 (Heinz et al., 2010). Once expressed, PU.1 uses its "pioneer" activity to bind to closed chromatin in a process thought to remodel local chromatin by displacing histones and exposing proximal regulatory binding sites to be targeted by TFs that do not possess this same pioneer capability (Heinz et al., 2010; Heinz et al., 2013; Heinz et al., 2015). This developmental process effectively establishes the regulatory potential of different cell types through the coordinated "selection" of lineage- and cell type-specific enhancers that confer and maintain the identity of a given cell type. An additional class of TFs includes those that are activated or expressed in response to environmental stimuliA well-characterized example in macrophages is the NF-κB complex that is activated in response to diverse stimuli including the detection of bacterial lipopolysaccharide (LPS) found at the surface of harmful pathogens (Xie et al., 1994; Hwang et al., 1997, Heinz et al., 2010; Heinz et al., 2013; Heinz et al., 2015). NF-κB binding to the regulatory elements established by pioneer factors such as PU.1 results in the robust stimulus-dependent activation of programs of proinflammatory genes that function in concert to respond to the threat of a potential pathogen (Heinz et al., 2010; Natoli et al., 2011; Heinz et al., 2013; Ostuni et al., 2013; Heinz et al., 2015).

As TFs like PU.1 and NF-κB recruit transcriptional cofactors with enzymatic activities, regulatory elements established by these and other TFs can be located genome-wide through profiling the chemical modifications that are introduced by these recruited COFs. For example, many of the myeloid regulatory elements established by PU.1 and subsequently bound by signal-dependent TFs such as NF-κB in macrophages are also marked with histone modifications characteristic of primed and active enhancers such as histone 3 lysine 4 mono-methylation (H3K4me1) and histone 3 lysine 27 acetylation (H3K27ac) as well as by the presence of the general histone acetyltransferase (HAT) p300 (Ghisletti et al., 2010; Natoli et al., 2011; Ostuni et al., 2013; Heinz et al., 2013). Beyond these two enhancer marks, there exists a number of additional histone modifications established by recruited enzymes whose combinations are thought to roughly delineate different types of regulatory elements (such as active promoters, primed enhancers, poised enhancers, and active enhancers) (Heintzman et al., 2007; Ernst et al., 2011; Ernst and Kellis, 2012). Understanding the grammar of regulatory element selection and maintenance genome-wide is therefore a complex problem of understanding where and why a TF can bind to a given DNA site and which enzymatic COFs can be consequently recruited to these TF sites.

Due to the important role of TFs in establishing and maintaining regulatory elements, there exists massive cooperative efforts to map the genome-wide binding sites of different TFs and locations of non-coding regulatory elements in diverse cell types and contexts (for example ENCODE and modENCODE) (Feingold et al., 2004; Birney et al., 2007; Gerstein et al., 2010). In addition, as TFs are sequence-specific DNA binding proteins, there also exists community resources and databases such as JASPAR (Sandelin et al., 2004; Fornes et al., 2020) , CIS-BP (Weirauch et al., 2014), Transfac (Wingender et al., 2000), UNIPROBE (Newburger and Bulyk, 2009), HOCOMOCO (Kulakovskiy et al., 2013; Kulakovskiy et al., 2018), and MotifDb (Shannon and Richards, 2018) dedicated to the collection of PWM models for different TFs that can be used to predict the binding sites of these factors. These community efforts to dissect the complex grammar of gene regulation at the level of non-coding DNA elements underscore the importance of TFs and exist due to the availability of diverse and effective methods to map the genome-wide binding sites of TFs and characterize the nucleotide determinants of their sequence-specific DNA binding.

1.2 Current methods to profile TF-DNA binding preferences

Several methods currently exist to profile the binding locations of TFs such as ChIP-seq (reviewed in Furey, 2012) and elucidate the nucleotide determinants of their binding (reviewed in Inukai et al., 2017). Each of these methods have their own advantages and drawbacks. The protein-binding microarray (PBM) for example is an *in vitro* method for measuring TF-DNA binding preferences (Mukherjee et al., 2004; Berger and Bulyk, 2006). In brief, a protein of interest is either expressed/tagged and incubated on a double-stranded DNA microarray consisting of up to hundreds of thousands of unique sequences. The TF of interest is subsequently probed with a primary antibody (specific to that TF) and binding of the primary antibody is detected in turn using a fluorophore-conjugated secondary antibody. Intensity of the fluorescence is measured and has been shown to be proportional to the TF's affinity for a given DNA probe. A distinct advantage of the PBM over other technologies, such as the widely used genomic chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) assay (discussed below), is the customizability of the DNA probes on the array. As any set of customized DNA probes can be included on the array (within a probe number limit), groups have developed creative designs to elucidate the DNA-binding preferences of TFs using the PBM platform (Berger and Bulyk, 2009; Newburger and Bulyk, 2009). The Universal PBM design for example uses a k-mer-based approach to directly compare TF binding intensities at a “seed” reference DNA probe and all of its single nucleotide variants providing a method to determine the single nucleotide determinants of TF binding. This profiling resolution comes at the cost however of having to profile in a noncellular context (an *in vitro* setting) where the potential influence of all cellular post-translational modifications (PTMs) and cooperative/collaborate protein cofactors on the binding of a TF of interest cannot be taken into account. Nonetheless, the PBM has facilitated efforts to characterize the DNA-binding models of TFs. The aforementioned Universal PBM design for example has been used to characterize the binding models of hundreds of TFs which are compiled in the widely used UNIPROBE resource (Newburger and Bulyk, 2009) which is in turn included in meta-databases such as MotifDb (Shannon and Richards, 2018).

Similar *in vitro* techniques to the PBM exist, such as high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX) and the related SELEX-seq, which in place of a microarray use a library of DNA probes (Jolma et al., 2010; Riley et al., 2014). These probes are incubated with TFs of interest and undergo multiple rounds of selection in order to identify the DNA probes preferentially bound by the protein of interest. These probes are then sequenced and analyzed to determine the binding preferences of the given TF. As with the PBM, the probe selection does not need to be limited to genomic sequences, allowing researchers to screen the impact of different DNA variants on the binding of a TF and develop customized library designs to fit their research questions. However, HT-SELEX and SELEX-seq suffer from many of the same caveats as the PBM since they are both *in vitro* techniques. The impact of cell-specific phenomena such as PTMs and cofactors are therefore not taken into account in these SELEX-based techniques either.

In addition to *in vitro* techniques used to study TF-DNA binding, such as the PBM and HT-SELEX/SELEX-seq, there exists *in vivo* techniques such as the widely used ChIP-seq (Birney et al., 2007; Johnson et al., 2007). Briefly, TFs in cells are cross-linked to DNA using a fixing agent, the DNA is fragmented, and DNA segments bound by a TF of interest are precipitated using an antibody (similar to how the TF of interest is detected in the PBM assay). The precipitated DNA is then purified, amplified, and sequenced. Computational methods to map these sequencing reads back to a reference genome and determine areas of locally enriched mapped reads (often termed ‘peaks’) allow researchers to determine likely binding sites for their TF of interest (Zhang et al., 2008; Guo et al., 2012; Xing et al., 2012). Models for the DNA-binding preferences of TFs can be computed using *de novo* motif inference techniques based predominantly on the occurrence and frequency of k-mers within the peaks uncovered (Heinz et al., 2010; Bailey et al., 2009; Machanick and Bailey, 2011; Ma et al., 2014). Relative to *in vitro* techniques, ChIP-seq offer the distinct advantage of profiling in the cellular context where the influence of cell-specific phenomena such as PTMs and cooperative/collaborative cofactors. In addition, determining the genomic binding sites of a TF can inform additional analyses and integration with other modalities such as RNA-seq to infer the target genes of a TF and map gene regulatory networks in response to stimuli (reviewed in Adigun et al., 2015). However, in regards to determining the TF binding specificities, the technique possesses certain limitations. For example, many of the ‘peaks’ found in a ChIP-seq experiment will not contain an identifiable binding site as determined by the DNA-binding model and many will contain multiple possible binding sites where the individual contributions of these sites toward TF binding and cis-regulatory element (CRE) function cannot be distinguished without additional experimentation (Inukai et al., 2017; MacQuarrie et al., 2011). Furthermore, the binding data is limited to the sequences naturally occurring within the genetic sequence within the organism being profiled. Therefore, elucidating the impact of a specific hypothetical DNA nucleotide variant of interest is not possible if the variant does not already occur in the cell being interrogated whereas with *in vitro* assays, binding to DNA variants can be directly interrogated. Due to its many advantages and applications, ChIP-seq has been widely employed by the community and coordinated efforts to map the genome-wide binding of TFs in diverse cell types and contexts, such as the ENCODE (Feingold et al., 2004) and modENCODE consortia (Gerstein et al., 2010), have contributed immeasurably to our understanding of TF binding and gene regulatory function.

Thus far, the methods of profiling TF binding discussed have been limited to one TF per experiment. Recently, a high-throughput (HT) electrophoretic mobility shift assay (EMSA) and mass spectrometry-based technique called Active TF Identification (ATI) was developed to profile the binding preferences of all TFs in the cell (Wei et al., 2018). This is achieved by incubating DNA oligomers with nuclear extracts, isolating all protein-DNA complexes by EMSA, and then analyzing the captured DNA sequences using HT sequencing and the captured proteins by mass spectrometry. Computational approaches are then used to infer the various DNA binding models (i.e. PWMs) and to match them with the TFs identified by mass spectrometry. The ATI method offers an interesting alternative to other techniques like ChIP-seq where only a single TF can be profiled at a time. This increased profiling throughput at the TF-level however comes with the cost of not being able to directly identify the genomic binding sites of the TFs with certainty. Together, these methods are effective but demonstrate a historical focus at characterizing TF binding preferences and binding sites over the functional potential of different sites. Comparatively less effort has been placed in developing techniques for determining mapping the nucleotide determinants of functional higher-order regulatory complexes and the recruitment sites of transcriptional COFs, the enzymatic effectors of gene regulatory processes like histone modification and chromatin remodeling.

1.3 Importance of studying DNA-binding preferences of TFs and higher-order complexes

An important reason to study the DNA-binding preferences of TFs and higher order gene regulatory complexes is their potential role in aberrant cell states and disease. A key finding from over a decade's worth of genome-wide association studies (GWAS) is that the vast majority of DNA sequence polymorphisms linked to disease occur in non-coding sections of the genome (Gallagher and Chen-Plotkin, 2018). These non-coding polymorphisms do not affect the sequence of an expressed protein but can instead perturb how the expression of a protein (or multiple proteins) is/are regulated resulting in consequences at the phenotype level. Of the disease-associated polymorphisms that occur within the non-coding genome, a substantial portion have been demonstrated to occur within cell type-specific gene regulatory elements (Maurano et al., 2012; Farh et al., 2015). Understanding the role of genetics in disease therefore depends on our ability study the mechanistic effects of DNA variants on the function gene regulatory elements such as promoters and enhancers and with particular focus on cell type-specific regulatory elements.

A primary mechanism by which DNA variants alter gene expression is by altering the DNA binding of TFs. Previous studies conducted to estimate the prevalence of disease-associated variants in suspected TF binding sites have found that only 10-20% likely perturb the binding of a TF with a previously characterized binding model (Farh et al., 2015). Our inability to link variants with altered TF binding has contributed to a widening gap between the ease with which we can statistically associate a non-coding DNA variant with a phenotypic outcome and our ability to functionally determine the molecular mechanisms that explain these associations (Gallagher and Chen-Plotkin, 2018). For example, there exists HT techniques to link DNA variants uncovered in GWAS with gene expression outcomes. RNA-seq can be combined with genotyping for example in a cell-specific manner to uncover expression quantitative trait loci (eQTLs) to statistically associate DNA variants with some allelic change in gene expression (Sun and Hu, 2013; Majewski and Pastinen, 2011). Recently, massively parallel reporter assays (MPRAs) have become a popular HT tool to move beyond statistical association and directly measure the impact that single DNA variant has on reporter activity (Tewhey et al., 2016). These types of studies that seek to link GWAS variants (and other DNA variants such as somatic mutations suspected to drive cancer development) with gene expression changes however do not explain the mechanism through which these variants act to affect gene expression.

Assaying the effects of a DNA variant on the binding of a TF or the subsequent recruitment of a higher-order gene regulatory complex is comparatively a more difficult task than associating variants with gene expression changes. Methods such as EMSA exist to study protein-DNA interaction and can be used to study the mechanistic effects of a variant on protein binding (Hellman and Fried, 2007) but is intractable to perform larger scale analyses on variants of interest. Larger scale analysis techniques that use ChIP-seq data exist to study the effects of different alleles on the binding of a TF exist (Harley et al., 2018; Van De Geijn et al., 2015) but only one TF or chromatin feature can be investigated at a time. These allelic imbalance approaches are therefore not suitable for more discovery-based work which is the nature of trying to uncover potential molecular mechanisms for GWAS and/or disease-associated variants. To address the widening gap between studies performed to uncover these variants and those performed as functional follow-ups, there is a need for more high-throughput and discovery-based methods to study the link between DNA variants and protein-DNA binding.

1.4 Transcriptional cofactors – linking DNA variants and TFs to gene regulatory activities

As altered TF binding alone is thought to be an underlying molecular mechanism of only a fraction of disease-associated DNA variants (Farh et al., 2015), there must be an alternate (or several) mechanism(s) to functionally explain why a given variant can affect the expression of a gene. A potential explanation could be that the allosteric effect of a DNA variant could affect the optimal recruitment of a COF complex without itself affecting the binding efficacy of an underlying TF (or TF complex). This is supported by the finding that though disease-associated variants often do not occur within a known DNA-binding motif, they often appear proximal (Farh et al., 2015).

COFs are thought of as the effectors of gene expression change as they have enzymatic activities that mediate diverse gene regulatory functions such as histone modification, chromatin remodeling, as well as formation of the transcription preinitiation complex (Reiter et al., 2017; Shlyueva et al., 2014;, Zabidi and Stark, 2018, Haberle and Stark, 2018). An example of a COF with a critical function in gene regulation is the coactivator p300 (Goodman and Smolik, 2000; Gerritsen et al., 1997). Though it is often thought of as a histone acetyltransferase, it has been shown to confer acetyl groups to many additional proteins in a gene regulatory context (Weinert et al., 2018). P300 is a key activator widely found at enhancers (Raisner et al., 2018) and known to be recruited to DNA through protein-protein interactions with DNA-bound TFs through various interaction domains (Vo and Goodman, 2001). The p300 example demonstrates several general principles of the transcriptional COFs; they are promiscuously recruited by diverse TFs and since they are recruited to DNA indirectly, their sequence specificity is conferred by the TFs with which they interact (Zabidi and Stark, 2016; Haberle and Stark, 2018; Reiter et al., 2017). Though COFs are thought of as the components with effector function, focus has traditionally been placed on TFs with massive international efforts and community resources dedicated to understanding their binding preferences (Feingold et al., 2004) though TF binding alone does not necessarily result in known function (MacQuarrie et al., 2011) and COFs are known in turn to influence the binding of TFs (Siggers et al., 2011; Siggers and Gordân, 2014).

As the binding preferences of COFs have been historically understudied compared to TFs, the potential role of perturbed COF recruitment at disease-associated DNA variants is not completely understood (Lee and Young, 2013). In addition, consistent with the current methods to profile the role of DNA variants in TF binding, the methods that exist to profile COF recruitment are not suited for the analyses that need to be performed to associate DNA variants with altered COF recruitment or TF-COF complex formation. Similar to TFs, ChIP-seq can also be used to determine the genome-wide COF recruitment locations of a COF of interest. Methods to examine allelic imbalance however are impractical (as discussed above) and would not suggest a complete mechanism. For example, measuring perturbed COF recruitment at a locus would not indicate which TF or TF(s) is/are involved. The approach also requires the variant to naturally occur in the population of cells being studied. In comparison, techniques such as yeast two-hybrid (Y2H) (Kohalmi et al., 1998) and mammalian two-hybrid (M2H) (Riegel et al., 2017) can precisely map the interactors of a COF of interest but these techniques are limited to assaying binary protein-protein interactions and are not suited for analysis involving single DNA variants. The related immunoprecipitation mass spectrometry (IP mass spec) identifies the TFs that interact with a COF but does not explicitly assay DNA-bound complexes (Mohammed et al., 2016). Despite the availability of these methods, there does not currently exist any methods to interrogate the individual nucleotide determinants of COF recruitment events in a cell-based and HT manner.

1.5 The need for new cell-based allelic-resolution COF recruitment profiling and analysis techniques

Our inability to mechanistically explain why a DNA variant correlates with a disease outcome and the unstudied potential role of higher order protein-DNA complexes (Lee and Young, 2013) suggest that the current methods available are not suited to study the impact of single nucleotide variants on COF recruitment in a cell-specific manner.

As many of the potential COF recruitment requirements, such as nuclear localization, PTMs, and auxiliary TFs, depend on the cellular context (Zabidi and Stark, 2016; Haberle and Stark, 2018; Reiter et al., 2017), *in vitro* methods would potentially miss important COF recruitment events or provide misleading or incorrect results. New methods and investigations into the role of DNA variants in the recruitment of COFs (or the assembly of higher order TF-COF complexes) should therefore strive to be more cell-based. However, cell-based assays, such as ChIP-seq for example, require that the variant or variants being interrogated exist naturally within the cells being assayed which is not always possible or realistic. An ideal new method to screen the effects of DNA variants on the recruitment of gene regulatory complex would thereby combine the advantages of both *in vitro* and cell-based modalities as is the case with MPRAs and gene expression (Tewhey et al., 2016). In order to bridge the gap between DNA variant association with disease and causal mechanism, new allelic resolution COF recruitment profiling techniques must be developed and combine a more cellular context with the customizability and ease-of-use of *in vitro* platforms.

1.6 Introduction to the novel methods developed for this dissertation

Motivated by the lack of methods to profile the nucleotide determinants of regulatory complex binding, this dissertation proposes three related methods to perform these analyses in a cell-based but highly customizable manner in terms of the DNA sequences being profiled.

In Chapter 2, we discuss the development of the nuclear extract protein-binding microarray (nextPBM) as an extension of the traditional PBM. In place of tagged or purified proteins, nextPBM profiles protein-DNA binding from whole nuclear extract. This allows researchers to characterize the binding of a protein of interest in the presence of possible cooperative-acting factors at their relative endogenous levels and with cellular their PTMs present. We use the nextPBM platform to characterize binding of myeloid LDTF PU.1 from nuclear extracts compared to *in vitro* transcribed/translated (IVT) at both its canonical binding sites as well as the composite PU.1-IRF8 binding site. We leverage the customizability of the DNA probes included to elucidate site- and context-specific single nucleotide binding preferences of PU.1 and the PU.1-IRF8 cooperative complex to these different sites. We propose nextPBM as a general purpose protein-DNA binding assay that combines the flexibility of a customizable *in vitro* system with a more biologically relevant profiling context.

In Chapter 3, we present CASCADE (Comprehensive Assessment of Complex Assembly at DNA Elements). With CASCADE, we extend the nextPBM platform from profiling TF-TF cooperative complexes to profiling indirect recruitment of COFs from cell nuclear extract. Using a single variant (SV) DNA probe approach, we show how

CASCADE can be used to profile the nucleotide determinants of COF recruitment to both large CREs as well as known functional single-nucleotide polymorphisms (SNPs) in myeloid cells. By comparing the COF recruitment models to previously characterized binding models, we infer the identity of TFs underlying COF recruitment events observed demonstrating how CASCADE can be used to characterize TF-COF binding. Importantly, we show how a 2-step approach that includes CASCADE can be used to rapidly screen the mechanistic effects of functional SNPs on TF-COF binding/recruitment in a cell- and stimulus-dependent manner thus providing the field with an HT assay to characterize SNPs.

In Chapter 4, we present the human transcription factor (hTF) array as a standardized array design to profile the recruitment of COFs to a diverse panel of TFs. To analyze results of these COF recruitment experiments, we designed a full software suite complete with an interactive user interface to enable researchers to explore their TF-COF binding data. In addition, we discuss the construction of cell state-level recruitment "signatures" and how these could eventually be used to study the TF-COF complexes active in disease cell states to inform COF- and TF-level biomarkers.

In Chapter 5, we conclude this work by summarizing the advances presented by the development of these three novel methods (nextPBM, CASCADE, and the hTF array) and outline future work. With plans to expand our approaches to characterize more COFs and further integrate our COF profiling measurements with other modalities, we hope to further establish CASCADE and the hTF array as transformative approaches that will enable researchers to investigate TF-COF complex binding in ways that were not previously achievable.

CHAPTER TWO: nextPBM - a platform to study cell-specific transcription factor binding and cooperativity

Note: A substantial portion of this chapter was previously published in a peer-reviewed journal (Mohaghegh et al., 2019) with Nima Mohaghegh (NM) and David Bray (DB) featured as co-first authors and equal contributors. The optimized nuclear extraction protocol that made the nextPBM technique possible was developed by NM with input from Trevor Siggers (TS) and DB. DB performed all of the computational work including the design and analysis of all ChIP-seq experiments, selection of DNA probes to include in the pilot nextPBM design, development of the nextPBM analysis and visualization pipelines, and the generalizable framework to discover and characterize cooperative TF-TF complex binding. All experimental work (including ChIP-seq, nextPBM, and validation assays) were performed by NM and TS. Individual contributions to the results outlined in each figure are included in the corresponding figure legend. The manuscript was written by DB, NM, and TS. Supplementary data published alongside the paper can be found in the online version of the article.

2.1 Abstract

HT *in vitro* methods for measuring protein-DNA binding have become invaluable for characterizing TF complexes and modeling gene regulation. However, current methods do not utilize endogenous proteins and, therefore, do not quantify the impact of cell-specific PTMs and cooperative cofactors. We introduce the HT nextPBM (nuclear extract protein-binding microarray) approach to study DNA binding of native cellular TFs that accounts for PTMs and cell-specific cofactors. We integrate immune-depletion

and phosphatase treatment steps into our nextPBM pipeline to characterize the impact of cofactors and phosphorylation on TF binding. We analyze binding of PU.1/SPI1 and IRF8 from human monocytes, delineate DNA-sequence determinants for their cooperativity, and show how PU.1 affinity correlates with enhancer status and the presence of cooperative and collaborative cofactors. We describe how nextPBMs, and our accompanying computational framework, can be used to discover cell-specific cofactors, screen for synthetic cooperative DNA elements, and characterize TF cooperativity.

2.2 Introduction

Defining the principles that govern TF binding and the assembly of multi-protein TF complexes remains a challenge (Siggers and Gordân, 2014; Slattery et al., 2014). HT *in vitro* techniques (both microarray- and sequencing-based) exist to characterize the DNA binding of TFs (Slattery et al., 2014; Andrilenas et al., 2015) and cooperative TF complexes (Siggers et al., 2011; Slattery et al., 2011; Jolma et al., 2015). Current approaches assay the binding of purified or in vitro produced protein samples (Slattery et al., 2011; Berger et al., 2006; Badis et al., 2009), or tagged protein overexpressed in cells (e.g. HEK293) (Jolma et al., 2013; Fang et al., 2012). Consequently, these approaches do not assay the impact of cell-specific PTMs, which are known to have diverse effects on TF binding and function (Tootle and Rebay, 2005; Filtz et al., 2014), and do not account for the impact of cell-specific cofactors that can bind cooperatively with TFs.

To characterize cell-specific TF binding features and account for the impact of cofactors and PTMs, we have developed nextPBMs. PBMs are double-stranded DNA microarrays that allow *in vitro* measurement of protein binding to tens of thousands of unique DNA sequences (Berger et al., 2006). NextPBM extends the PBM methodology by using total nuclear extracts in place of purified, IVT, or over-expressed proteins (Materials and Methods). To test the impact of specific cofactors and PTMs on binding, we have developed immune-depletion and phosphatase treatment steps into our nextPBM pipeline. We describe a computational framework based on binding to single-nucleotide variant (SNV) sites that provides a powerful approach to study DNA-binding specificity and protein cooperativity when assaying heterogenous NEs. We use nextPBMs to analyze the DNA binding of the myeloid cell-lineage factors PU.1 and IRF8, and discuss our results. We outline how nextPBMs can be used to discover cooperative TF binding and to infer the identity of cooperative-acting factors. Finally, we demonstrate how nextPBMs can be used to screen for cooperatively bound synthetic DNA elements. NextPBMs are an extendible and robust HT method to assay the binding of proteins to genomic or synthetic sites that can capture the impact of cell-specific cofactors and PTMs on TF-DNA binding.

2.3 Results

2.3.1 Genome-wide binding of PU.1, C/EBPa, and IRF8 in a human monocyte line

To demonstrate the nextPBM approach, we examined binding of PU.1/SPI1 from human monocytes as a test case. PU.1 is a master regulator of the myeloid lineage (Nerlov and Graf, 1998; Rosenbauer and Tenen, 2007; Scott et al., 1994) and functions to establish localized histone modifications that define the cell-specific enhancer repertoire (Heinz et al., 2010; Ghisletti et al., 2010; Barozzi et al., 2014). In myeloid cells, PU.1 can bind DNA autonomously to 5’-GGAA-3’ ETS motifs, or cooperatively with IRF8 to 5’-GGAANNGAAA(C/G)-3’ ETS-IRF composite elements (EICEs) (Rehli et al., 2000; Eklund et al., 1998; Merano et al., 1999). In order to select PU.1 binding sites to examine using our nextPBM assay, we first sought to determine the *in vivo* genomic instances of these sites. To define the PU.1 binding landscape in human monocytes, and co-occupancy with cofactors, we performed ChIP-seq on PU.1, C/EBPα and IRF8 in resting THP-1 cells (Materials and Methods). We observed widespread binding for each factor and significant overlap in their binding profiles (Fig. 2.1A), consistent with previous studies (Ghisletti et al., 2010; Heinz et al., 2010; Langlais et al., 2016; Mancino et al., 2015). We identified 47,799 PU.1, 26,648 C/EBPα and 2,588 IRF8 binding loci (i.e., ChIP-seq peaks) in resting THP-1 cells. The number of PU.1 binding sites is consistent with numbers reported for human peripheral blood monocytes (Pham et al., 2012) and mouse macrophages (Ghisletti et al., 2010; Heinz et al., 2010; Mancino et al., 2015). Critically, we observed near complete overlap of IRF8 binding sites (95%) with PU.1 binding sites, supporting the model that *in vivo* IRF8 must bind as a complex with PU.1 in these resting cells.

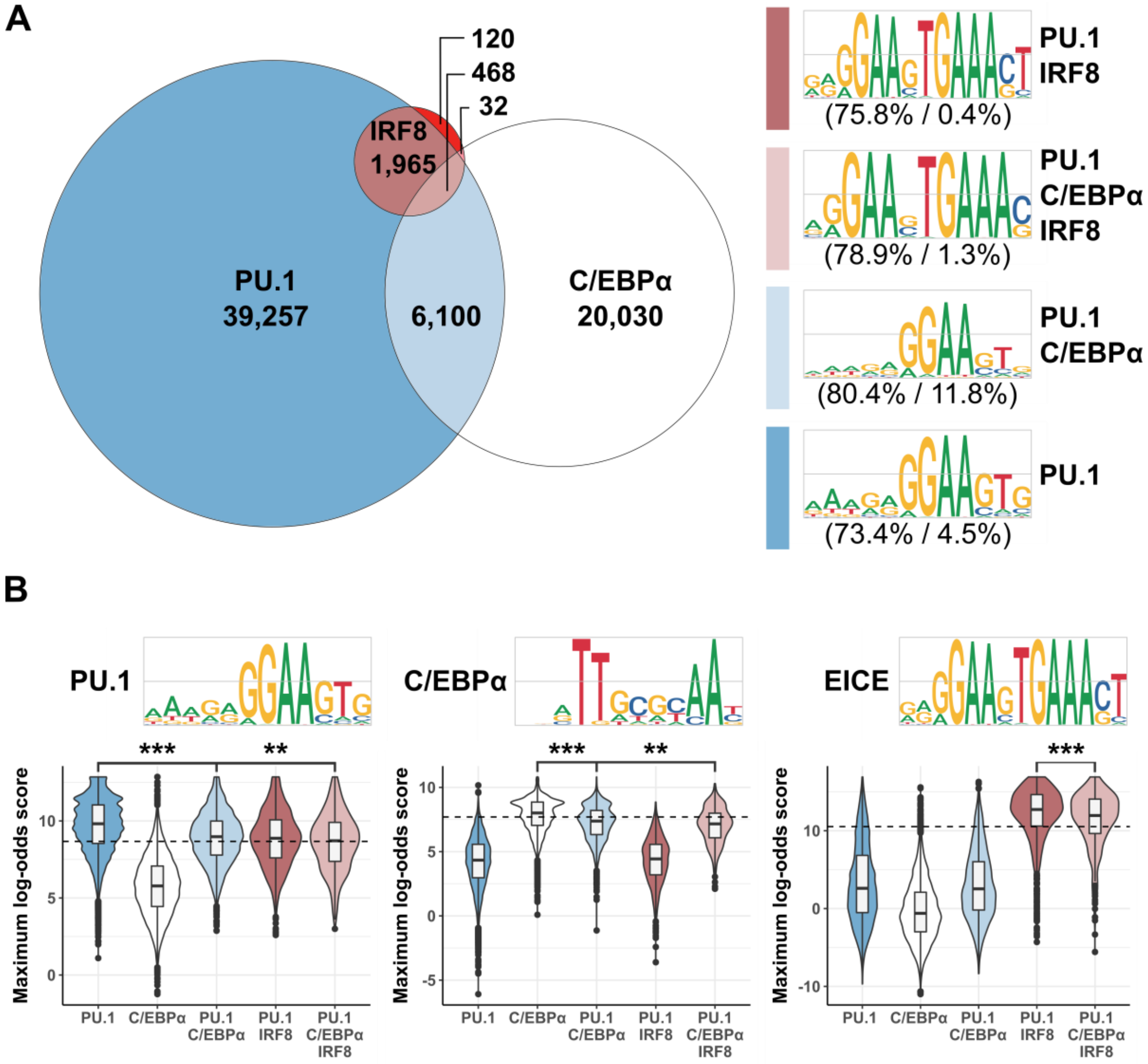


Figure 2.1: Genome-wide binding for PU.1, C/EBPα and IRF8 in human monocytes
(A) Overlap of genome-wide ChIP-seq peaks discovered for PU.1, C/EBPα, and IRF8. *De novo* motifs discovered within each PU.1-containing intersection are shown on the right. Numbers in brackets indicate the percentage of peaks containing the *de novo* motif (left) compared to background (right). Grey bars in the motif logos represent bit values of 0 (bottom), 1 (middle), and 2 (top). When a ChIP-seq peak overlapped multiple peaks from another experiment we aggregated them into a single overlapping region. (B) Distributions of motif scores obtained for ChIP-seq peaks categories described in (A). ChIP-seq peaks were scanned with each motif and assigned the maximum log-odds score (see Materials and Methods). Contributions: ChIP-seq experiments were designed by NM, DB, and TS and performed by NM. ChIP-seq analysis and subsequent motif finding and scoring was performed by DB with input from TS.

To characterize the DNA sequence motifs that define PU.1, C/EBPα and IRF8 binding we performed *de novo* motif analysis on defined subsets of the bound genomic loci (Materials and Methods). Binding motifs determined for genomic loci bound by PU.1 alone or with other factors agree well with known motifs (Fig. 2.1A). At loci bound by PU.1 alone we identify a canonical PU.1 binding motif. At loci shared with IRF8 (or IRF8 and C/EBPα) the dominant motif is the EICE bound cooperatively by IRF8 and PU.1. At loci bound with C/EBPα we identify a PU.1 motif, supporting the idea that collaboration between these factors is not via direct cooperative DNA binding but rather through synergistic effects on chromatin (Feng et al., 2008; Ghisletti et al., 2010). We find that the PU.1 motif identified on loci co-occupied with C/EBPα is slightly more degenerate than for PU.1 alone, suggesting that PU.1 binding sequences may be lower affinity at loci shared with C/EBPα.

To determine the specificity of TF motifs for their respective *in vivo* binding profiles we scored bound regions using the individual TF motifs (Fig. 2.1B). For both PU.1 and C/EBPα we observe that their binding motif is highly predictive of their ChIP-seq peaks. However, for both TFs, motif scores are lower for loci bound with the other factors. For IRF8, we find that the EICE motif scores are much higher for IRF8 ChIP-seq peaks than for peaks from other factors. These analyses demonstrate that motifs identified for each TF are specific for their genomic binding loci, and that TF binding at loci co-occupied with either a collaborating or a cooperative cofactor can be lower scoring.

2.3.2 Nuclear extract protein-binding microarrays (nextPBM)

To define PU.1 binding sites for our nextPBM assay (Fig. 2.2A), we selected 2,499 DNA sites in ChIP-positive regions that matched a PU.1 PWM (Materials and Methods, Supplementary File 1 from Mohaghegh et al., 2019). To identify composite PU.1-IRF8 EICE elements, we selected 116 EICE sites from regions bound by both PU.1 and IRF8. Nuclear extracts from human THP-1 monocyte cells were made using a detergent-based cell lysis and extraction procedure and incubated with the double-strand DNA microarrays (Materials and Methods). As proteins in the assay are not epitope-tagged, primary antibodies were used to label PU.1, followed by fluorescently labeled secondary antibodies (Materials and Methods).

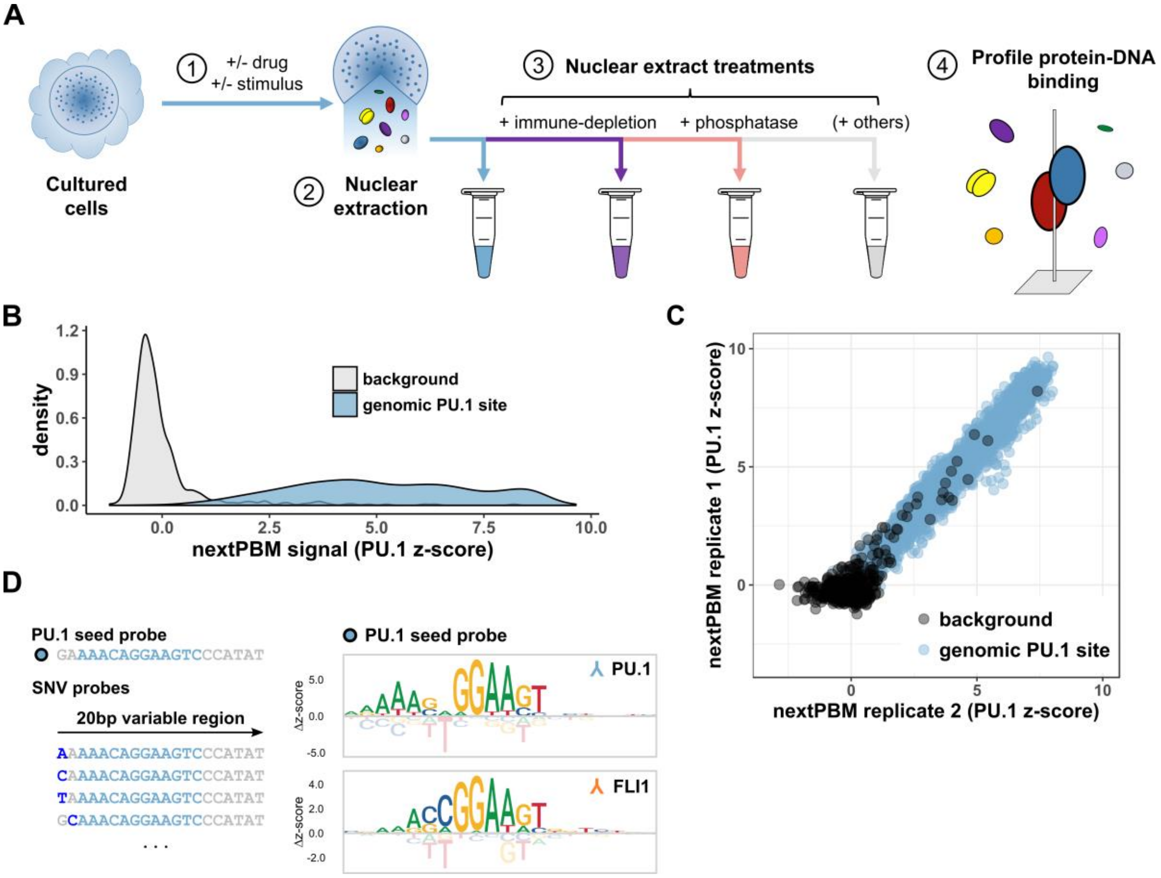


Figure 2.2: Nuclear extract protein-binding microarrays (nextPBMs)
(A) Workflow schematic for the nextPBM protocol. (1) Cultured cells can be stimulated or treated with a drug prior to nuclear extraction. (2) Total soluble protein content is harvested from cell nuclei using an optimized protocol (see Materials and Methods). (3) Nuclear extract can be treated in parallel enzymatically (i.e. by phosphatase treatment) and components of interest can be depleted (i.e. by immune-depletion using a targeted antibody) depending on goals of the experiments. (4) DNA binding affinity of one or more transcription factors of interest are profiled in parallel directly from nuclear extract. (B) Density of PU.1 nextPBM z-scores obtained at random background probes (n = 500) and at genomic PU.1 binding sites (n = 2,615). (C) Scatterplot of PU.1 binding z-scores obtained by DNA probes corresponding to random background (black) and genomic PU.1 sites (blue) in different biological replicates. (D) Left: Schematic representation of the SNV probes corresponding to an example PU.1 seed probe. Genomic sequence corresponding to the PU.1 motif is highlighted in sky blue within a larger 20bp sequence. SNVs within a given SNV probe are shown in dark blue. Right: Sequence logos obtained for the same genomic PU.1 seed probe using a PU.1 antibody (top) and an FLI1 antibody (bottom). Δz-scores are computed relative to the median score obtained within a given column. Contributions: nextPBM workflow was developed jointly by NM, DB, and TS. NM and TS performed the nextPBM experimental work. nextPBM microarray design, analysis, and visualization pipelines were developed by DB with input from TS.

PU.1 binding was detected to genome-derived sites significantly above background sites (Fig. 2.2B), demonstrating that there is sufficient endogenous protein in nuclear extracts to quantify TF binding using our assay. PU.1 binding profiles for individual replicate experiments were highly correlated, demonstrating high reproducibility between nextPBM experiments (Fig. 2.2C). To assess the sensitivity of our nextPBM assay, we generated a PU.1 DNA-binding logo using a SNV probe-based approach (Materials and Methods) (Andrilenas et al., 2018). Briefly, we measured PU.1 binding to a 20 bp-long seed sequence and all 60 SNV sequences (Fig. 2.2D); logos were generated from binding scores to each SNV sequence (Fig. 2.2D). The PU.1 binding logo agreed well with the established ETS-type motif (Wei et al., 2010), demonstrating that we can accurately measure the TF binding specificity using nextPBMs. As the nuclear extract is highly heterogenous and contains other ETS family proteins, we asked whether the binding of another ETS factor could be assayed in parallel using the same DNA sites. Probing the nextPBM with antibodies to FLI1, another ETS factor expressed in THP-1 monocytes, we were able to define the FLI1 binding motif (Wei et al., 2010) using the same seed and SNV probes as used for PU.1. We note that PU.1 and FLI1 are related ETS factors and exhibit only minor differences in their DNA binding specificity, namely the 2–3 bases upstream of the 5'-GGAA-3' core element (Fig. 2.2D); however, we were able to resolve their distinct motifs in parallel using the SNV approach. These results show that robust and sensitive quantification of TF binding can be performed for TFs at endogenous levels in heterogeneous nuclear extracts.

2.3.3 Characterizing the DNA binding of PU.1 and IRF8 in monocytes using nextPBM

To identify monocyte-specific features of PU.1 binding, we compared monocyte nextPBM data for PU.1 with binding data using IVT PU.1 (Fig. 2.3A). Binding to genomic PU.1 sites was highly correlated between extract PU.1 and IVT PU.1 (Fig. 2.3A, highlighted in blue); however, binding of extract PU.1 was enhanced to the EICEs present in genomic regions co-occupied by PU.1 and IRF8 (Fig. 2.3A, highlighted in red). We confirmed that IRF8 was also bound to the EICE sites using an IRF8 nextPBM (Fig. 2.3B). IRF8 binds almost exclusively to the EICEs, consistent with the known requirement for cooperative binding with PU.1 in monocytes. The enhanced PU.1 binding to EICEs (Fig. 2.3A) suggests cooperative binding with a monocyte-specific cofactor. These results demonstrate that using nextPBMs to compare the TF binding profiles from nuclear extracts and purified/IVT protein provides a HT approach to identify cell-specific cooperative binding.

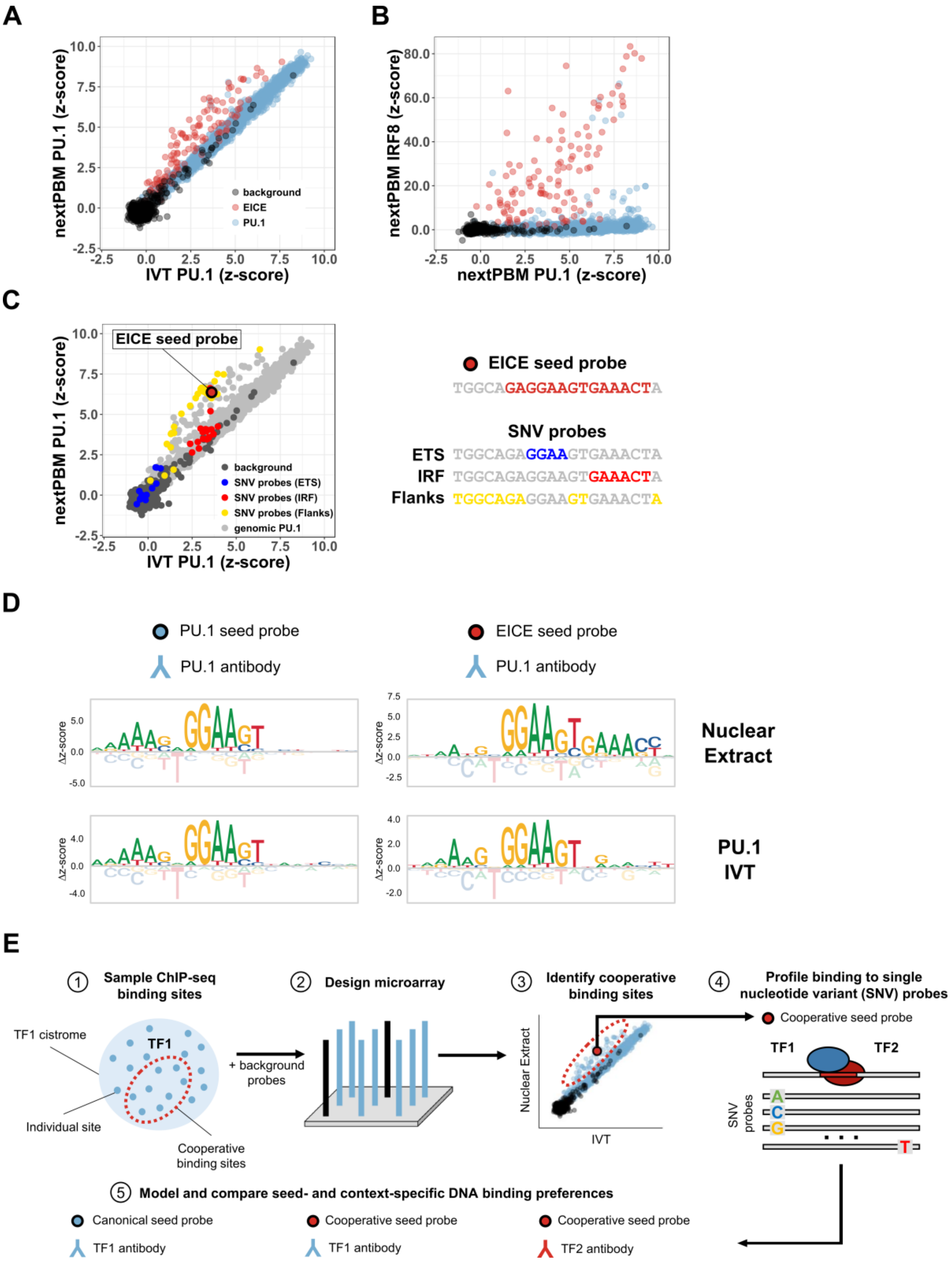


Figure 2.3: DNA sequence determinants of PU.1-IRF8 cooperative binding

(A) Scatterplot of PU.1 binding z-scores obtained from nuclear extract (nextPBM) versus IVT PU.1 for random background probes (n = 500), EICE probes (n = 116), and canonical PU.1 probes (n = 2,499). (B) Scatterplot of IRF8 binding z-scores in nuclear extract versus PU.1 binding z-scores in nuclear extract for the same sets of probes as in (A). (C) Left – scatterplot of PU.1 binding z-scores in nuclear extract versus IVT PU.1 for probes included in (A) and SNV probes corresponding to the EICE seed probe shown right. Highlighted probes correspond to SNV probes containing variations in either the ETS core half-site (blue), IRF core site (red), or flanking and linker bases (yellow). Right - schematic of EICE seed probe and bases comprising individual sub-elements. (D) Sequence logos obtained using a canonical PU.1 seed probe (left column) and a cooperative ETS-IRF composite element (EICE) probe (right column) from nuclear extract (top row) and from IVT PU.1 (bottom row). (E) Workflow schematic for identifying cooperative binding sites using nextPBM. 1 – ChIP-seq sites for a given transcription factor of interest (TF1) can be sampled and used to construct probes for a microarray design. The sample will contain sites where TF1 is cooperatively bound with other factors. 2 – TF1 sample probes are combined with a set of random background probes against which binding z-scores are computed to form the basis of a microarray design. 3 – Profiling binding of TF1 in nuclear extract versus IVT allows for the discovery of cooperative binding sites bound higher in nuclear extract (shown above the diagonal). 4 – Cooperative sites identified can be used as seed probes in a subsequent experiment where SNV probes are included in the microarray and profiled. 5 – Binding to SNV probes is used to model and compare seed- and context-specific DNA binding preferences of TF1 to identify composite elements and likely binding partners. Contributions: nextPBM experimental work was performed by NM and TS. nextPBM design, analysis, and visualization pipelines were developed by DB with input from TS.

2.3.4 Defining DNA-sequence determinants of PU.1-IRF8 cooperativity

To examine determinants of PU.1-IRF8 cooperativity we visualized the impact of SNVs on PU.1 binding (Fig. 2.3C). We highlighted SNVs that occur in different regions of an EICE site: ETS/PU.1 half-site (blue); IRF half-site (red); flanking and linker sequence (yellow) (Fig. 2.3C). SNVs in the ETS half-site abrogate PU.1 binding for both IVT and nuclear extract samples as expected (Fig. 2.3C, blue). SNVs in the IRF half-site affect the cooperative binding but do not affect the binding of IVT PU.1, capturing the impact of IRF8 present in the extract samples (Fig. 2.3C, red). SNVs in the flanking and linker sequence affect PU.1-IRF8 complex affinity but largely do not abrogate the cooperative interactions (i.e., most yellow data points are above the diagonal), demonstrating that cooperative binding does not require specific sequence features outside of the core half-sites (Fig. 2.3C, yellow). This analysis highlights that nextPBMs can be used to dissect the determinants of cooperativity for a single DNA binding site.

Binding specificity can also be visualized as DNA-binding logos, providing a way to easily reveal binding differences to distinct classes of DNA sites under different sample conditions (Fig. 2.3D). PU.1 binding logos generated for a seed sequence that was not bound cooperatively match canonical PU.1 logos for both the nuclear extract and IVT experiments (Fig. 2.3D, left). In contrast, the PU.1 binding logos for a cooperatively bound seed sequence differ between the conditions: the logo from the nuclear extract experiment resembles the composite EICE element, showing the influence of the IRF8 binding, while the logo from the IVT experiment shows just the PU.1 logo (Fig. 2.3D, right). We note that we obtain consistent motifs when using other high-scoring seed sequences (Supplementary Fig. 2.3 and Supplementary Fig. 2.4). The impact of cofactors on binding to the distinct classes of DNA sites can be easily visualized using SNV-based logo analysis. Using this approach we can analyze multiple TF binding modes in parallel in a single experiment (i.e., the PU.1 logos for cooperative and non-cooperative binding were determined using a single experiment).

2.3.5 Approach to identify and characterize cooperative binding

Our results provide an approach for the identification and characterization of cell-specific cooperative binding (Fig. 2.3E). Briefly, putative DNA binding sites of a TF can be identified from genomic data (e.g. ChIP-seq combined with motif analysis, etc.) or be designed synthetically based on prior knowledge, and can be incorporated into a nextPBM microarray (Fig. 2.3E, steps 1 and 2). For example, scanning PU.1 ChIP-seq data with a PU.1 PWM with relaxed cutoff scores can be used to identify both autonomous and cooperatively bound sites. Next, comparison of binding profiles between nuclear extract and purified TF experiments can be used to identify cooperatively bound sites (Fig. 2.3E, step 3). Based on this data, one can design SNV probes for target DNA sites and perform a follow-up nextPBM experiment to define DNA-binding logos that reveal the cooperative binding specificity and provide information about the identity of cooperatively acting factors. For example, monitoring PU.1 binding revealed the 5’-GAAACT-3’ IRF logo (Fig. 2.3D), which could be matched to PWMs from databases to make predictions about the PU.1 cooperative binding partner. The outlined approach provides a HT assay to identify and characterize cooperative TF complexes in a cell-specific manner.

2.3.6 Sensitivity of cooperative binding to nuclear extract concentration

To test the sensitivity of our results on nuclear extract concentration, we performed nextPBM experiments at successive dilutions of monocyte nuclear extract. We quantified PU.1 cooperativity as the off-diagonal displacement of the 116 EICE sites from the autonomously bound PU.1 sites (as in Fig. 2.3A, Materials and Methods). We found that PU.1-IRF8 cooperativity decreased with decreasing extracts concentrations (Fig. 2.4A). We also assessed cooperativity by monitoring the PU.1 DNA binding logo for an EICE site as extract concentration varied. We observed a consistent PU.1 element (i.e. 5’-GGAA-3’ core) with a successively weaker IRF8 element (i.e., 5’-GAACT-3’) (Fig. 2.4B, left). As PU.1 can bind to DNA in an autonomous or cooperative fashion, both the bound PU.1 and PU.1-IRF8 complexes contribute to the microarray spot intensity in a PU.1 nextPBM. Therefore, observing PU.1 cooperativity requires that the increase in spot intensity due to the presence of PU.1-IRF8 complexes must be discerned beyond the signal intensity from PU.1 binding alone, leading to the observed concentration dependence in our assay. In contrast, IRF8 is an obligate dimer in this context; therefore, all signal in an IRF8 nextPBM is due to PU.1-IRF8 complexes. As such, the binding logos for an IRF8 nextPBM are much more robust to extract concentrations and we can discern cooperative EICE logos for all extract concentrations (Fig. 2.4B, right). The results demonstrate that the concentration dependence of cooperative binding in our assay will depend on the characteristics of the individual binding partner.

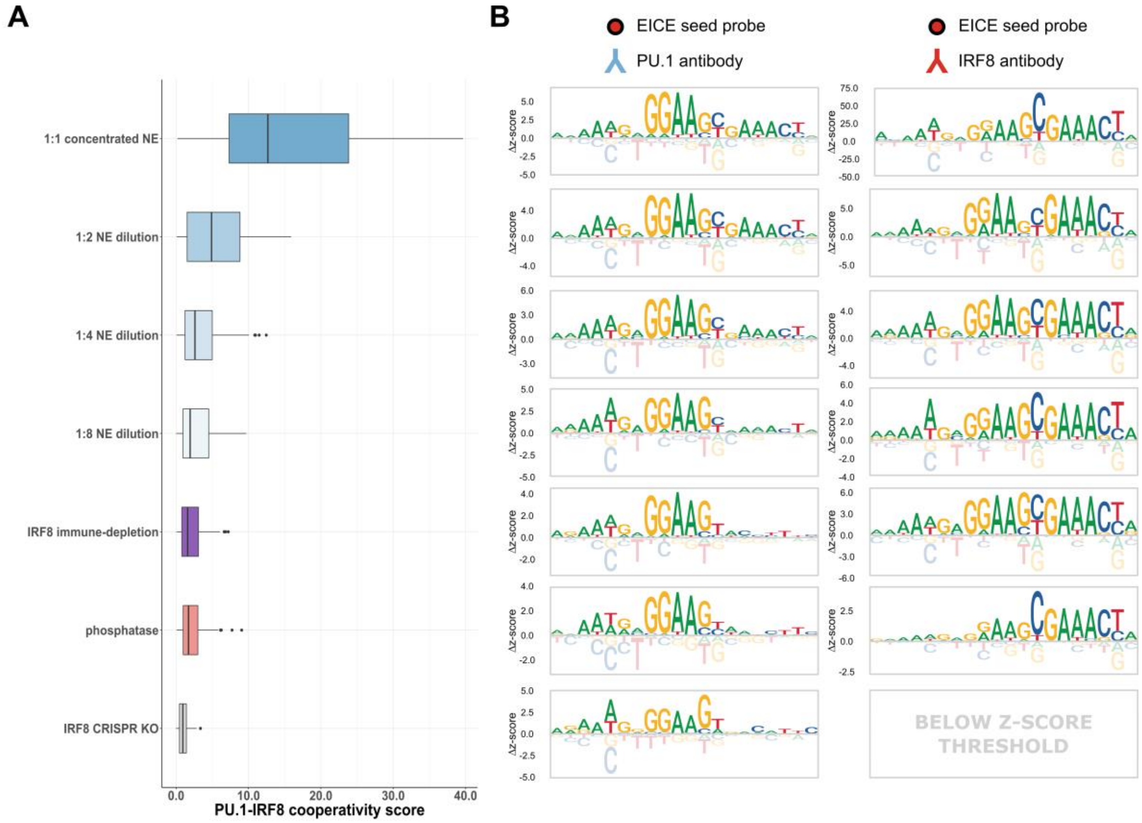


Figure 2.4: Effects of different nuclear extract treatments on PU.1-IRF8 cooperative binding
(A) Boxplot of PU.1-IRF8 cooperativity scores (see Materials and Methods) for a set of EICE probes (n = 116) in various listed nuclear extract (NE) conditions and treatments including a gradient of 2-fold dilutions (1:1, 1:2, 1:4, and 1:8), an extract where IRF8 has been immune-depleted (IRF8 immune-depletion), an extract treated with a broad-spectrum phosphatase (phosphatase), and an extract generated from a line of cells where IRF8 has been knocked out (IRF8 CRISPR KO). Boxplot elements – center line: median, box limits: first and third quartiles, whiskers: 1.5x interquartile range, individual points: data points beyond end of whiskers. (B) Sequence logos obtained by profiling PU.1 binding (left column) and IRF8 binding (right column) to the same sample EICE seed probe in the corresponding nuclear extract treatments/conditions from (A). Contributions: nextPBM experimental work including the nuclear extract treatments were performed by NM and TS. nextPBM design, analysis, and visualization pipelines were developed by DB with input from TS.

2.3.7 Assessing the impact of cofactors and post-translational modifications

To determine whether the cooperative PU.1 binding to EICEs was solely due to IRF8 we used CRISPR/Cas9 to mutate the IRF8 gene in THP-1 monocyte cells and performed nextPBM using nuclear extracts from IRF8-deficient cells. Cooperative binding of PU.1 was lost in the absence of IRF8 protein (Fig. 2.4A and Fig. 2.4B, bottom), consistent with reduced PU.1 ChIP-seq to EICEs reported for Irf8-null mouse macrophages (Mancino et al., 2015). CRISPR/Cas9-based deletion of target TFs remains a labor-intensive process; therefore, we sought to develop a more rapid approach for testing the impact of cofactors on cooperative TF binding. We developed an immune-depletion protocol to deplete a TF from the nuclear extracts in the nextPBM pipeline (Fig. 2.2A). NextPBM with IRF8-depleted extracts showed similar abrogation of the enhanced PU.1 binding (Fig. 2.4A and Fig. 2.4B), corroborating the CRISPR/Cas9-based results that IRF8 is solely responsible for PU.1 cooperativity. Our depletion step removed >90% of the IRF8 from the extract sample (Supplementary Fig. 2.5); however, an IRF8 nextPBM was still successful and we were able to generate an EICE logo, demonstrating that for obligate heterodimers such as IRF8, cooperative binding can be detected even with low levels of protein in the extract. NextPBM with an immune-depletion treatment provides rapid assay for the impact of cofactor proteins on cooperative TF complexes.

PTMs play a central role in the regulation of TF function and cooperative TFs complexes *in vivo*. Cooperative binding of PU.1 and IRF8 has been reported to involve phosphorylation of IRF8 (Sharf et al., 1997). To test the impact of phosphorylation on PU.1 cooperativity we incubated our extract sample with a broad-spectrum phosphatase prior to the nextPBM (Fig. 2.2A, Supplementary Fig. 2.6, Materials and Methods). Phosphatase treatment of our extract samples abrogated PU.1 cooperative binding to the EICEs (Fig. 2.4A), showing the dependence of PU.1-IRF8 cooperativity on phosphorylation. The disruption of cooperative binding can also be seen in the PU.1 binding logo as an absence of the IRF8 half-site (Fig. 2.4B). We note that this treatment had no effect on autonomous PU.1 binding. Therefore, nextPBM with an enzymatic treatment of the extract provides a rapid assay for the PTM-dependence of TF binding to diverse DNA sequences.

2.3.8 Screening synthetic DNA elements for cooperative binding

NextPBMs present an opportunity to screen synthetic DNA elements (i.e., mutant or novel sequences) for cooperative TF binding in a more cell-native context that may be used to probe the rules of cooperativity or to design synthetic genetic regulatory elements. We first tested our ability to screen for the impact of half-site ablations on cooperative binding. We compared the binding of PU.1 and IRF8 to 60 EICE elements and matched mutants with an ablated ETS or IRF site (Fig. 2.5A and Fig. 2.5B). Mutating the ETS half-site abrogates PU.1 binding, whereas mutating the IRF half-site only affects the observed cooperativity (Fig. 2.5B). In contrast, IRF8 binding is abrogated with mutations to either the ETS or the IRF half-site (Fig. 2.5C). These results demonstrate that IRF8 binding is dependent on cooperativity with PU.1, but not vice versa, consistent with observations *in vivo* (Rehli et al., 2000; Eklund et al., 1998; Merano et al., 1999). We next tested our ability to screen for new cooperative sites and generated 199 synthetic EICEs by combining low-affinity PU.1 sites with a consensus IRF8 site (Fig. 2.5A). An adjacent IRF8 site greatly enhanced PU.1 binding to all sites in the presence of the nuclear extract but not for IVT PU.1 (Fig. 2.5B). Similarly, IRF8 bound strongly to these synthetic EICEs, and at levels higher than seen for the genomic EICEs (Fig. 2.5C).

NextPBMs provide a platform for HT screening of DNA sequences for cooperative binding that can account for the impact of the native cell-specific protein environment.

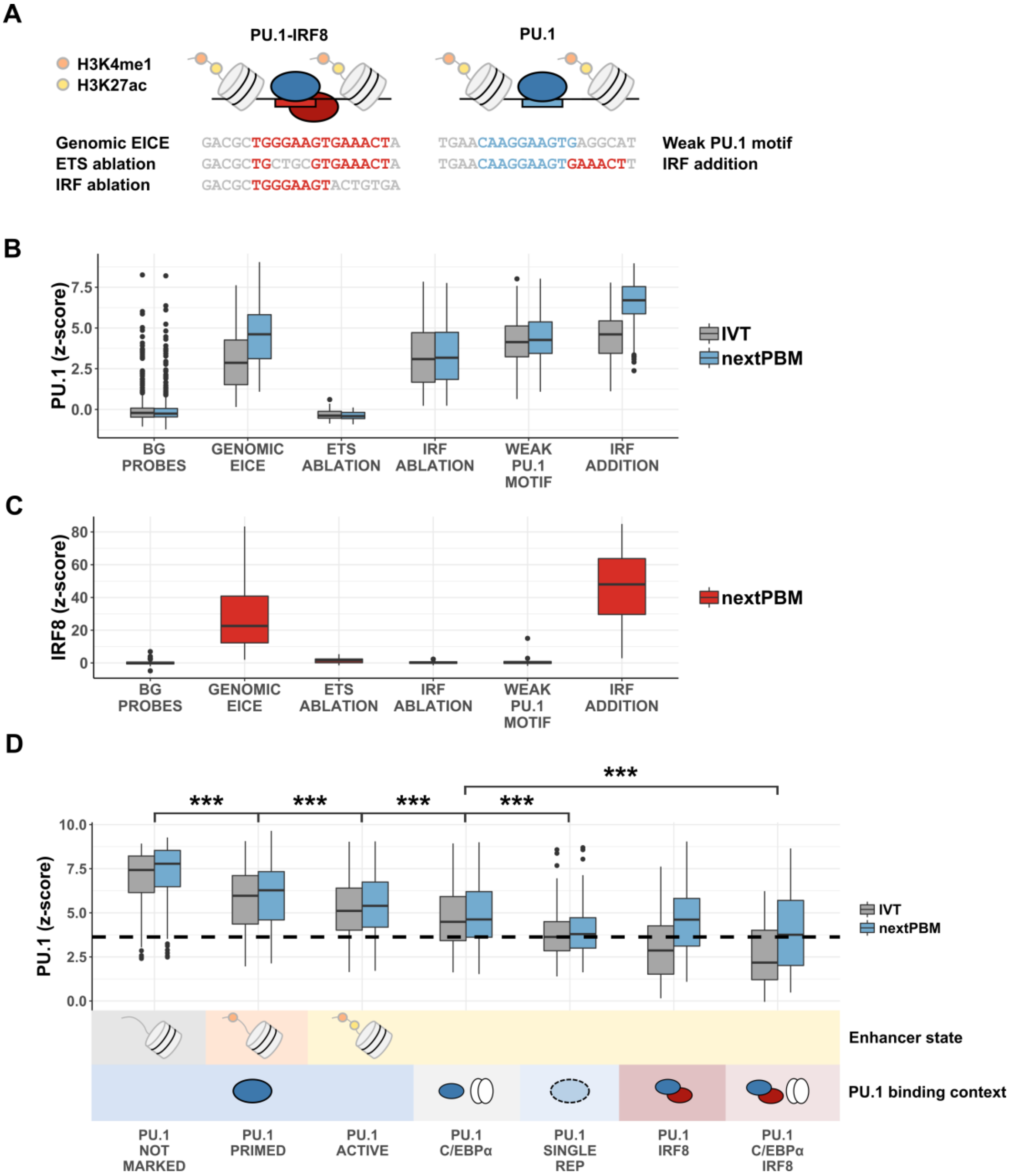


Figure 2.5: Screening synthetic cooperative elements and binding from different genomic contexts

(A) Schematic showing representative probe sequences and corresponding mutated elements. For each genomic EICE from active enhancers in our array design (n = 60), there is a corresponding probe with the ETS and IRF core sites independently mutated to contain a different k-mer. For each canonical PU.1 probe with a weak motif (n = 199), there is a corresponding probe with an IRF half-site added. (B) Distributions of PU.1 binding z-scores for DNA probe groups in (A) in nuclear extract (blue) compared to IVT. BG – random background probe set (n = 500). Boxplot elements – center line: median, box limits: first and third quartiles, whiskers: 1.5× interquartile range, individual points: data points beyond end of whiskers. (C) Distributions of IRF8 binding z-scores to the same DNA probe groups as in A and B. Boxplot elements: same as in (B). (D) Distributions of PU.1 binding z-scores for DNA probe categories defined by ChIP-seq co-occupancy with cofactors (PU.1 binding context) and/or histone modifications (enhancer state). 'NOT MARKED' indicates the absence of H3K4me1 and H3K27ac histone modifications. 'SINGLE REP' designates a category of probes designed using PU.1 ChIP-seq peaks that were discovered in a single biological replicate but were not observed in a duplicate experiment. The dashed black line denotes an approximate ChIP-seq reproducibility threshold corresponding to the median z-score obtained for the PU.1 SINGLE REP group in IVT. Boxplot elements: same as in (A) and (B). Contributions: Integrative genomics analyses and categorization of PU.1 sites was performed by DB with input from TS. nextPBM experimental work was performed by NM and TS. nextPBM design, analysis, and visualization pipelines were developed by DB with input from TS.

2.3.9 Binding affinity of PU.1 correlates with enhancer state and cofactor occupancy

To examine how nextPBM data can inform genomic analysis of TF binding, we examined PU.1 binding to sites from genomic regions defined by distinct chromatin states and cofactor occupancy. In addition to IRF8, PU.1 functions with C/EBPα to bind chromatin and establish macrophage-specific genes expression (Heinz et al., 2010; Feng et al., 2008; Laiosa et al., 2006; Xie et al., 2004). PU.1 does not bind DNA cooperatively with C/EBPα; rather they function collaboratively through mutual effects on repressive chromatin environments. We performed ChIP-seq on C/EBPα and identified PU.1 binding sites in regions co-occupied by both PU.1 and C/EBPα, or by all three factors (PU.1, C/EBPα and IRF8) (Fig. 2.1A). To examine the relation between chromatin state on PU.1 binding, we also performed ChIP-seq for H3K4me1 and H3K27ac that define poised (H3K4me1 only) and active (H3K4me1 and H3K27ac) enhancer states (Heintzman, et al., 2007; Creyghton et al., 2010).

We first examined PU.1 binding to distinct enhancer states: primed, active, or unmarked (no H3K4me1 or H3K27ac marks) (Fig. 2.5D). To control for the effect of cofactors we limited our analysis to sites from PU.1-only occupied regions. PU.1 binding affinity shows a clear trend with enhancer state. High-affinity PU.1 binding to unmarked loci is in agreement with previous studies (Pham et al., 2013) and suggests that PU.1 occupancy to less biophysically accessible chromatin regions requires high-affinity sites. Low-affinity PU.1 binding in active enhancers reveals that functional PU.1 sites are not the highest affinity, and that genome-wide analyses of the highest affinity TF sites may be enriched for non-functional binding. Binding to all sites agrees between the nuclear extract and IVT samples, suggesting that there is no influence of cooperative binding to these genomic elements and that the binding trends are defined by autonomous PU.1 binding.

We next examined PU.1 binding at active enhancers co-occupied by collaborative (C/EBPα) or cooperative (IRF8) cofactors (Fig. 2.5D). We observe a clear trend in affinity for the PU.1 IVT data that suggests an impact of cofactors on PU.1 binding. First, PU.1 binding sites are lower affinity in regions co-occupied by either cofactor than in regions occupied by PU.1 alone (Fig. 2.5D). For example, in regions co-occupied with C/EBPα, PU.1 binding sites have Δz-scores ~ 0.5 lower than for PU.1-only regions (P-value < 0.001), and in regions co-occupied with IRF8 the affinity is even lower (Δz-score ~ 2.0, P-value < 0.001). Unexpectedly, in regions co-occupied by both cofactors

(C/EBPα and IRF8) PU.1 binding is the lowest affinity (Δz-score ~ 2.5, P-value < 0.001), suggesting that the effects of collaborative and cooperative cofactors on PU.1 binding are independent and additive. However, when analyzing the nextPBM data, we observe that cooperativity with IRF8 significantly increases the PU.1 binding to EICE sites. For perspective, we examined PU.1 binding to sites from genomic regions identified as PU.1-bound in only a single ChIP-seq replicate experiment (SINGLE REP), which we found to be lower affinity than for reproducible PU.1 ChIP sites. We observe that, in the absence of IRF8, PU.1 affinity falls below this ‘reproducible level’, which may explain the drop in PU.1 ChIP-seq signal observed in IRF8 knock-out mouse macrophages (Mancino et al., 2015). Our results demonstrate that cooperative binding with IRF8 or collaborative function with C/EBPα allow PU.1 binding sites to be much lower affinity than an optimal site, and highlight the perspective gained by analyzing TF binding using both purified/IVT and nuclear extract samples.

2.4 Discussion

HT methods for characterizing TF-DNA binding provide critical biophysical data for genomic analyses of gene regulation (Siggers and Gordân, 2014; Slattery et al., 2014; Andrilenas et al., 2015). Cell-specific PTMs (Tootle and Rebay, 2005; Filtz et al., 2014) and cofactors (Siggers and Gordân, 2014; Garvie and Wolberger, 2001) can affect TF binding, but are not implicitly accounted for in current HT methods. Here we describe the nextPBM methodology for the characterization of protein-DNA binding that uses nuclear extracts to account for the impact of cell-specific PTMs and cofactors. We show that a direct comparison of binding profiles between nuclear extract of purified/IVT samples can reveal cooperative binding activity and cooperatively bound sites. Using an SNV-based approach to query sequence specificity and generate binding logos we can examine binding and cooperativity for individual genomic sites. The flexibility to analyze binding specificity for individual sites allows multiple binding modes to be directly studied in parallel in a single experiment. This approach is analogous to the seed-and-wobble approach previously described for universal PBMs that quantify TF binding to k-mers (Berger et al., 2006). We note that DNA shape is known to play an important role in TF binding specificity (Andrabi et al., 2017; Zhou et al., 2015; Yang et al., 2014), and future studies that examines the role of DNA shape in the context of multi-protein complexes and cell-specific extracts will be informative. We anticipate that this approach will be particularly useful when studying TFs that function as obligate heterodimers and may have multiple binding partners in a complex nuclear extract and, therefore, interact with DNA using distinct binding modes. To address the impact of cofactors and phosphorylation on TF binding we have incorporated immune-depletion and phosphatase-treatment steps into our nextPBM pipeline. Incorporation of additional enzymatic treatment steps will allow us to expand our assay to study other PTMs (e.g., demethylases to study impact of methylation, etc.). NextPBMs provide an extendible platform to study the DNA binding of endogenous TF complexes in a cell-specific manner. We anticipate that nextPBM-based comparison of cell-specific TF binding and cooperative assembly will be particularly informative when applied to comparisons of different cell types, cell-stimulation conditions, and to cells from disease contexts.

Our study outlines a new approach to identify cooperative binding TFs and cooperatively bound sites. First, by sampling from bound genomic loci identified by ChIP-seq experiments, one can design a nextPBM microarray to survey a diverse set of binding sites for a TF. Currently, using available microarray platforms (Materials and Methods), and accounting for replicate probes, we can assay up to ∼18 000 unique genomic sites in an experiment, which is sufficient to thoroughly sample (or even completely cover) most TF cistromes. Direct comparison of nextPBM binding profiles from nuclear extract and purified protein can then reveal differentially bound DNA sites. Enhanced binding in nextPBM experiments indicates potential cooperative binding, and by reanalyzing these binding sites with a subsequent SNV-based array design we can generate binding logos on a per-site basis that can be used to make predictions about the possible cooperative binding partners. The identity of binding partners can then be tested using nextPBMs with an immune-depletion step. Using nextPBMs to compare the binding profiles of TFs from different cells will be particularly useful for studying the TFs that function as obligate heterodimers and may utilize different partner proteins in different cell types. While the cooperative complex examined in this manuscript involves two proteins (PU.1 and IRF8), this approach can, in principle, be used to examine cooperative assembly of more than two proteins as all constituents are available in the nuclear extract. We have previously demonstrated that cooperative complexes of more than two purified proteins can be assayed using the PBM technology (Siggers et al., 2011). This approach to identify cooperative binding can also be used to screen novel DNA elements for cooperative binding activity (Fig. 2.5), providing an HT method for the design and testing of cell-specific cooperative elements that can be used to construct synthetic gene regulatory elements for mammalian cells.

We examined the binding of PU.1 and IRF8 from human monocytes, and identified the known composite EICE binding logos using nextPBMs probing either PU.1 or IRF8. Using CRISPR/Cas9-based IRF8 knockout and immune-depletion we demonstrated that IRF8 is the only cooperative binding partner for PU.1 in human monocytes. Investigating the relationship between binding, cooperativity and genomic occupancy we found that PU.1 binding affinity exhibits a clear trend with both enhancer type and cofactor co-occupancy. We found that the highest affinity PU.1 sites are in genomic regions not containing the H3K4me1 and H3K27ac histone modifications for active enhancers, and lowest affinity sites are in active enhancers. Furthermore, co-occupancy with either collaborating (C/EBPα) or cooperative (IRF8) cofactors correlated with lower affinity binding sites, suggesting that cofactor occupancy allows for the evolutionary selection of lower affinity binding sites. Surprisingly, coincident binding of PU.1 with both C/EBPα and IRF8 allowed for still lower affinity sites to be utilized. These results highlight that functional binding sites are not the highest affinity, and that genomic analyses biased to high affinity may miss functionally relevant sites. Furthermore, the nextPBM-based functional characterization of low affinity PU.1 binding sites proximal to cooperative and collaborative TF sites in active enhancers supports the findings from recent investigations into the genome-wide presence of low affinity PU.1 sites proximal to other TF binding motifs (Pham et al., 2013) and the discordance between TF binding affinity and transcriptional output (Grossman et al., 2017;

Andrilenas et al., 2018; Penvose et al., 2019). Finally, comparing binding profiles for nextPBM and IVT samples across stratified genomic sites demonstrated that PU.1 binding was autonomous on all sites except for the EICEs where it was cooperative with IRF8. NextPBM-based binding analysis of genome-derived sites provides insights into the biophysical determinants of TF binding. We anticipate that similar studies that compare TF binding profiles from different cellular conditions will provide new insights into the mechanisms of cell-specific binding and gene regulation.

2.5 Materials and Methods

2.5.1 Cell culture

THP-1 cells were purchased from ATCC (cat # TIB-202) and cultured in RPMI 1640 media with 10% FBS supplemented by 50 unit/ml Penicillin and 50 μg/ml Streptomycin. HEK293T cells for Lentivirus packaging (gift from Thomas Gilmore, Boston University) were cultured in DMEM media with 10% FBS supplemented by 50 unit/ml Penicillin and 50 μg /ml Streptomycin.

2.5.2 Protein samples

IVT samples of PU.1 (full-length, untagged) were generated using 1-Step Human Coupled IVT Kit – DNA (Thermo Fisher Scientific cat # 88881) following the provider's instructions. Protein expression was confirmed by Western analysis.

2.5.3 Antibodies

PU.1 (Santa Cruz sc-352x, used for ChIP and nextPBM); C/EBPα (Santa Cruz sc-61x, used for ChIP); IRF8 (Santa Cruz sc-6058x, used for ChIP and nextPBM); human

H3K4me1 (Abcam ab8895, used for ChIP); H3K27ac (Abcam ab177178, used for ChIP); alexa488-conjugated anti-goat (Life Technologies A11055, used for nextPBM); alexa647-conjugated anti-rabbit (Life Technologies A32733, used for nextPBM); and FLI1 (ABclonal A5644, used for nextPBM) was a gift from ABclonal.

2.5.4 Plasmids

Lentiviral plasmid constructs were prepared following Feng Zhang Lab (Massachusetts Institute of Technology) protocol. Briefly, to target IRF8 gene a pair of gRNAs were synthesized for exon 5 of the IRF8 gene (Primers: 5'-CACCGCTTCTGTGGACGATTACATG-3' and 5'-AAACCATGTAATCGTCCACAGAAGC-3') with overhangs and ligated into BsmBI digested pLentiCRISPRv2.0.

2.5.5 Nuclear extracts

5 × 10^6 THP-1 cells were pelleted at 500 × g for 5 min at 4°C in a 15 ml conical tube. The pellet was resuspended and washed twice with PBS. Cell pellet was resuspended in 1 ml of 'low-salt buffer' (10 mM HEPES (pH 7.9), 1.5 mM $MgCl_2$, 10 mM KCl plus 1 µl protease inhibitor cocktail (Sigma-Aldrich, cat # P8340) and incubated for 10 min on ice. 50 µl of 5% IGEPAL (Sigma-Aldrich, cat # I8896) was added to the cell suspension and vortexed for 10 seconds. Released nuclei were pelleted at 750 × g for 5 min at 4°C. The supernatant was saved as the 'cytosolic fraction'. To wash the remaining cytosolic proteins from the surface of the nuclear pellet, 100 µl of the low-salt buffer was gently pipetted onto the side of the tube and allowed to wash the pellet,

making sure to not disrupt the pellet. This wash was then gently transferred to the cytosolic fraction without dislodging the nuclear pellet. 200 µl of ‘high-salt buffer’ (20 mM HEPES (pH 7.9), 25% glycerol, 1.5 mM MgCl2, 0.2 mM EDTA, 420 mM NaCl plus 1 µl protease inhibitor cocktail) was pipetted on the pellet and the tube went through a vigorous vortex for 30 s followed by nutation at 4°C for 1 h. The nuclei were pelleted at 4°C for 20 min at 21,000 × g. The supernatant was transferred into another tube as the nuclear soluble protein fraction. Final nuclear extract samples used in nextPBM assays were 9.6 mg/ml.

2.5.6 CRISPR-mediated IRF8-knockout in THP-1 cells

To generate Lenti-CRISPR viruses, HEK293T cells were seeded in a 10 cm dish at 75% confluence a day before transfection. The next day, the confluent cells were co-transfected with 4µg of pCMV-VSV-G, 2 µg pCMV-ΔR8.91 and 1 µg plentiCRISPR v2-gRNA using a Lipofectamin-3000 kit and following the provider’s instructions. The transfection mixture was replaced by fresh media after 6 h and the virus-containing supernatant was collected after 48 h. Virus was concentrated by ultracentrifugation at 50,000 × g for 3 h at 4°C. The viral pellet was re-suspended in 500 µl complete medium (RPMI, 10% FBS) with 8 µg/ml Polybren and added to one million THP-1 cells in a microcentrifuge tube with 1.5 ml of complete media and shaken at 150 rpm for 30 min at room-temperature, followed by centrifugation at 850 × g for 30 min at 32°C. The THP-1 cell pellet was re-suspended in 2 ml of complete medium and was seeded in a 3 cm dish and incubated at 37°C with 5% CO_2 for 6 days. At day 6, infected cells were selected in 0.5 µg/ml puromycin (final concentration). The media was exchanged with fresh complete media containing 0.5 μg/ml puromycin every four to six days and for a total of 30 days. Cell confluence was maintained between 3 × 10^5 cells/ml to 9 × 10^5 cells/ml through the selection procedure and the culture volume was scaled up as necessary. Knockout efficiency in the pool of the infected cells was defined by Western analysis.

2.5.7 Nuclear extract treatments

Immune depletion of IRF8 – 7.5 μg of IRF8 antibody (abcam, ab207418) was added to 300 μL of diluted THP-1 nuclear extract (2 mg/ml total protein in nextPBM binding buffer (described below), 115 mM NaCl). The mixture was nutated at 4°C for 1 h. 75 μl of Dynabeads® Protein A slurry (Thermo Fisher Scientific, 10001D) was washed once using 1 ml of nextPBM binding buffer with 115 mM salt and collected by DynaMag magnet (ThermoFisher Scientific, cat # 12321D). Collected beads were re-suspended in the nuclear extract plus antibody mixture and transferred onto HulaMixer (ThermoFisher Scientific cat # 15920) to be rotated at 4°C for 2 h at 25 rpm. DynaMag magnet was used to collect the beads and the remaining nuclear extract was checked for the depletion of IRF8 by Western analysis. *Phosphatase treatment* – A general phosphatase (Lambda protein phosphatase kit, New England Biolabs, p0753) was added to 300 μl of diluted THP-1 nuclear extract (2 mg/ml total protein in nextPBM binding buffer (described below), 115 mM NaCl), and the reaction was carried out according to the provider's instructions. Phosphatase efficiency was checked by Western analysis for phospho-RNA polymerase II (abcam 5131).

2.5.8 Chromatin immunoprecipitation (ChIP-seq)

Soluble chromatin was prepared from 4×10^7 THP-1 cells according to previously described protocols (Lee et al., 2006) with some modifications (outlined below). Briefly, cells were crosslinked with 1% formaldehyde (final concentration) (Fisher Scientific, cat # F79-500) for 10 min at room temperature with gentle shaking. Crosslinking was stopped by adding 125 mM final concentration of glycine solution in PBS. Fixed cells were pelleted at 800 × g for 5 min at 4°C and washed twice with 10 ml of cold PBS in a 15 ml conical tube and pelleted at 800 × g for 5 min at 4°C. Washed cell pellet was re-suspended in 10 ml of Lysis Buffer 1 (Lee et al., 2006), nutated for 10 min at 4°C, and pelleted at 2,000 × g for 5 min at 4°C. The same procedure was repeated with lysis buffer 2 at room temperature followed by pelleting at 2,000 × g for 5 min at 4°C. To release nuclei from hard-to-disrupt THP-1 membranes, cells were re-suspended in 10 ml of Lysis Buffer 3 (Lee et al., 2006) and were shaken vigorously (225 rpm) at room temperature for 30 min. Cells were then passed through an 18-gauge needle (VWR, cat # BD305195) 25 times using a 10ml syringe. Nuclei were pelleted at 3,000 × g for 20 min at 4°C and re-suspended in 500 µl of Lysis Buffer 3 and then transferred into a 1.5 ml microfuge tube placed in Benchtop 1.5 ml Tube Cooler (Active Motif, cat # 53076). The nuclei were sonicated using Active Motif Q120AM sonicator with a 3.2 mm Probe (Active motif cat # 53053) at 25% amplitude for 15 min with 20 s ON and 30 s OFF cycles (45 cycles total). Cell debris was pelleted at 21,000 × g for 30 min at 4°C. 50 µl of the combined soluble chromatin was saved to be used as the input DNA upon reverse-crosslinking. For IP, 500 µl of the soluble chromatin was mixed with 30 µg of either PU.1, C/EBPα, H3K4me1 or H3K27ac antibodies (60 µg of IRF8 antibody was mixed with 1 ml of the soluble chromatin), and tubes were rotated at 25 rpm for one hour at 4°C using HulaMixer (ThermoFisher Scientific cat # 15920). 125 µl of the protein A Dynabead slurry (ThermoFisher Scientific cat # 10001D) per each rabbit antibody (PU.1. C/EBPα, H3K4me1 or H3K27ac), and 250 µl of the protein G Dynabead slurry (ThermoFisher Scientific cat # 10003D) for the goat-IRF8 antibody, were transferred into 1.5 ml microfuges and placed on DynaMag magnet (ThermoFisher Scientific, cat # 12321D) until all beads collected on the side of tubes. The solution was gently aspirated off from each tube and the beads were re-suspended in 1 ml of the Lysis Buffer 3 with several gentle inversions; beads were re-pelleted using the magnet and the lysis buffer was aspirated. Beads were then re-suspended in 50 µl of Lysis Buffer 3 and returned to HulaMixer to rotate at 35 rpm overnight at 4°C. Beads were collected and washed 6 times with 1 ml of the Lysis Buffer 3 and two times with 1 ml of the Wash Buffer (RIPA). All ChIP samples along with the 50 µl of the soluble chromatin were reverse-crosslinked by adding 200 µl of the Elution buffer and 3 µl of 20 mg/ml Proteinase K (ThermoFisher Scientific, cat # AM2546) and incubated at 65°C for overnight. Beads were collected and the solutions were transferred into a new 1.5 microfuge tube containing 1 µl of 10 mg/ml RNase A (ThermoFisher Scientific, cat # EN0531) and left at room temperature for an hour. The ChIP and input DNA were purified using QIAquick PCR Purification Kit (QIAGEN, cat # 28104) and eluted in 50 µl of 50°C Nuclease-Free Water (Thermo Fisher Scientific, AM9932). The concentration and size distribution of the ChIP-DNA samples were defined using Agilent 2100 Bioanalyser. DNA libraries were prepared using NEBNext Ultr II DNA Library Prep kit (NEB, E7645S) following the provider's instruction manual. Amplified libraries were Bioanalyzed again to check the size selection efficiency and to define the concentrations of libraries before preparing the library pool involving the same molarity of each library and sequenced by Illumina HiSeq 4000. An additional biological replicate for IRF8 (and corresponding input DNA) was sequenced using the Illumina NextSeq 500.

2.5.9 ChIP-seq analysis

ChIP-seq reads were aligned to the human reference genome (hg19) using Bowtie2 (Langmead and Salzberg, 2012). Aligned reads were filtered for high quality and uniquely mappable reads (MAPQ > 30) using samtools (Li et al., 2009). Peak calling for TFs was performed using MACS2 (Zhang et al., 2008) with relaxed parameters on single experiments (P-value < 0.01) and peaks were filtered using the irreproducible discovery rate (IDR < 0.05) across biological duplicates (Landt et al., 2012). Peak calling for histone marks was performed using MACS2 (Zhang et al., 2008) with relaxed parameters on single experiments (P-value < 0.01) and experiments were filtered requiring identification in both biological duplicates (i.e. IDR was not used for histone marks analysis). Peaks were further filtered if they occurred in the ENCODE consortium blacklisted regions. Peak intersections were computed using bedtools (Quinlan and Hall, 2010) by first merging the peaks from all TF ChIP-seq experiments into continuous genomic loci and identifying which TF(s) contained a peak within this union set. Raw and processed ChIP-seq data is available in the NCBI GEO database (Accession: GSE123872).

2.5.10 Motif discovery and scoring

De novo motifs within peak sets were discovered using HOMER (Heinz et al., 2010) (parameters: -size given -noweight -nlen 0 -len 6,8,10,12,14,16 -S 5) and subsequently used for motif scoring across all peaks. We also performed de novo motif analysis using MEME (Bailey et al., 2009) (meme-chip parameters: -dna -meme-mod zoops) and found consistent motifs (Supplementary Fig. 2.1). Log-odds scoring thresholds determined by HOMER against a set of random background sequences were used as significance thresholds for motif scanning. Motif scans on individual peaks were performed using a custom R script that implements the same scoring scheme as HOMER and reports the maximum log-odds score in each peak (available on Github: https://github.com/david-bray/nextPBM-paper). Uniform background probability for each nucleotide (0.25) at each position was used for log-odds scoring. We chose a uniform base-frequency background model to be consistent with that used by the HOMER algorithm, and to better support our biophysical interpretation of the nextPBM data, that is based solely on the contribution of each base to binding affinity. Motif logos were generated using the ggseqlogo R package (Wagih, 2017). Motifs and thresholds used for ChIP-seq analysis and PBM microarray design are provided (Supplementary File 2 from Mohaghegh et al., 2019).

2.5.11 PBM design

PBM experiments were performed using custom-designed microarrays (Agilent Technologies Inc. AMADID 085624 and 085106, 8 × 60K format). 2,615 PU.1 binding sites identified in ChIP-seq peaks were extracted from the genome as 20-bp genomic fragments and placed into a fixed position in the PBM probe sequence. For each unique probe sequence, 5 replicate probes were included in each orientation (10 probes per unique site). For select genomic seed sequences, 60 matching SNV probes were included to assay all SNVs at the 20 positions of the binding site (Fig. 2.2). All SNV sites were also included with 5 replicates and in each orientation (10 probes per unique SNV site). Probes for assaying binding site ablations and synthetic EICE sites were similarly included with 10 probes per unique DNA site. *Selection of binding sites from ChIP-seq data* – Binding sites were only included from PU.1 ChIP-seq peaks demonstrating high reproducibility across biological duplicates (IDR < 0.01), with the exception of probes included specifically to assay binding to single-replicate regions. PU.1 ChIP-seq sites were categorized based on their log-odds motif score, proximity to cofactors, and enhancer state. PU.1 binding sites were selected from the PU.1 ChIP-seq peaks containing exactly one significant PU.1 site (see *2.5.10 Motif discovery and analysis* above). For the genomic loci in the 'weak PU.1 motif' category we identified no significant PU.1 site and, therefore, used the PU.1 site with maximum log-odds score (Fig. 2.5). EICE sites were selected from PU.1-IRF8 co-occupied regions containing exactly one EICE site (see *2.5.10 Motif discovery and analysis* above). Co-occupancy PU.1 with cofactors (C/EBPα and/or IRF8) was determined if a highly reproducible ChIP-seq peak (IDR < 0.01) for each factor overlapped by at least one base. A PU.1 ChIP-seq peak was annotated as 'PU.1-alone' if it was located greater than 200 bases away from the nearest cofactor ChIP-seq peak (in all experiments, including duplicates, with peaks called using relaxed parameters as detailed above). Enhancer states were annotated using histone modification ChIP-seq data from biological duplicates and publicly available mRNA-seq data for THP-1 monocytes (GEO accession GSM927668). PU.1 sites were annotated as active if they occurred within 200 bases of the nearest H3K4me1 and H3K27ac peaks, and if the nearest gene was located between 2–500kb away and expressed above the median RPKM value. PU.1 sites were annotated as primed if they occurred within 200 bases of the nearest H3K4me1 peak only, and if the nearest gene was located between 2–500kb away and expressed below the median RPKM value. A full list of DNA probes used, their corresponding probe category and additional annotation can be found in the supplemental data (Supplementary File 1 from Mohaghegh et al., 2019).

2.5.12 NextPBM and PBM experiments and analysis

Microarray DNA double stranding and basic PBM protocols are as previously described (Berger and Bulyk, 2009; Andrilenas et al., 2015). All wash steps were carried out in coplin jars on an orbital shaker at 125 rpm. Double-stranded DNA microarrays were first pre-washed in PBS containing 0.01% Triton X-100 (5 min), rinsed in a PBS bath, and then blocked with 2% milk in PBS for 1 hour. Following the blocking step, arrays were washed in PBS containing 0.1% Tween-20 (5 min), then in PBS containing 0.01% Triton X-100 (2 min), and finally briefly rinsed in a PBS bath. *Protein binding* – Arrays were then incubated with the protein sample (IVT protein or THP-1 nuclear lysate, details in Supplementary File 3 from Mohaghegh et al., 2019) for one hour in a binding reaction buffer containing: 2% milk (final concentration); 20 mM HEPES buffer, pH 7.9; 100 mM NaCl; 1 mM DTT; 0.2 mg/mL BSA; 0.02% Triton X-100; and 0.4

mg/mL salmon testes DNA (Sigma D7656). *Primary antibody* – After protein incubation, microarrays were washed with PBS containing 0.5% Tween-20 (3 min), then in PBS containing 0.01% Triton X-100 (2 min), followed by a brief PBS rinse. Microarrays were then incubated with 10 µg/mL of primary antibody (see Supplementary File 3 from Mohaghegh et al., 2019) in 2% milk in PBS (20 min). *Secondary antibody* - After primary antibody incubation, microarrays were washed with PBS containing 0.5% Tween-20 (3 min), then in PBS containing 0.01% Triton X-100 (2 min), followed by a brief PBS rinse. Microarrays were then incubated with 7.5 µg/mL of alexa488-conjugated secondary antibody or alexa647-conjugated secondary antibody (see Supplementary File 3 from Mohaghegh et al., 2019) in 2% milk in PBS (20 min). Excess antibody was removed by washing with PBS containing 0.05% Tween-20 (3 min), then PBS (2 min). *PBM data analysis* - Microarrays were scanned with a GenePix 4400A scanner and fluorescence was quantified using GenePix Pro 7.2. Exported data were normalized using MicroArray LINEar Regression (Berger et al., 2006). Microarray probe sequences are provided (Supplementary File 1 from Mohaghegh et al., 2019). PBM data analysis and SNV approach for logo generation is as previously described (Andrilenas et al., 2018). Similarity between the DNA binding models generated using nextPBM and those from previously published studies was computed using the PWMSimilarity function from the TFBSTools R bioconductor package (Tan et al., 2016) (Supplementary Fig. 2.2). A threshold binding z-score of 2.0 (at the seed probe) was imposed to ensure accurate binding models. Processed PBM z-score data is available in the supplementary data (Supplementary File 1 from Mohaghegh et al., 2019), and all raw PBM data has been deposited in the NCBI GEO database (Accession: GSE123946). Scatterplots and boxplots were generated using the ggplot2 R package (Wickham, 2016). Motif logos were generated using the ggseqlogo R package (Wagih, 2017). The significance of PU.1 binding affinity and motif scores between groups was calculated using the two-sided Wilcoxon–Mann-Whitney test implemented in R.

2.5.13 PU.1-IRF8 cooperativity score

PU.1-IRF8 cooperativity was scored by quantifying the deviation of the observed EICE z-scores from an extract experiment from the expected z-scores based on the IVT sample experiment. To define the expected EICE z-scores a second degree polynomial model was fit to the z-scores for the canonical PU.1 probes as follows:

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon_1$$

where y_1 is the vector of PU.1 z-scores observed in the extract sample, x_1 is the vector of PU.1 z-scores observed for the IVT sample, β_0, β_1 and β_2 are coefficients of the best-fit polynomial model and ε_1 is the vector of error terms needed to equate y_1 to the function of x_1. A polynomial model was used to fit the canonical PU.1 site z-scores in place of a linear model to allow for non-linearity due to PU.1 concentration differences between experiments.

The coefficients fit above are then used to compute the expected EICE z-scores for the extract experiment based on the IVT experiment z-scores:

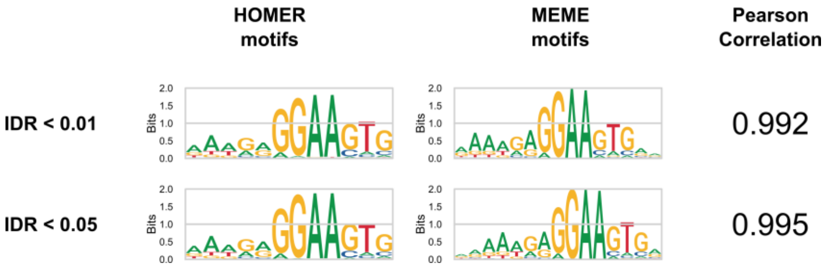
$$y_2 = \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \varepsilon_2$$

where y_2 is the vector of PU.1 z-scores observed at EICE probes in the extract sample, x_2 is the vector of PU.1 z-scores observed for the IVT sample, ε_2 is the vector of error terms needed to equate y_2 to the function of x_2 comprised of the coefficients fit using the canonical PU.1 probes.

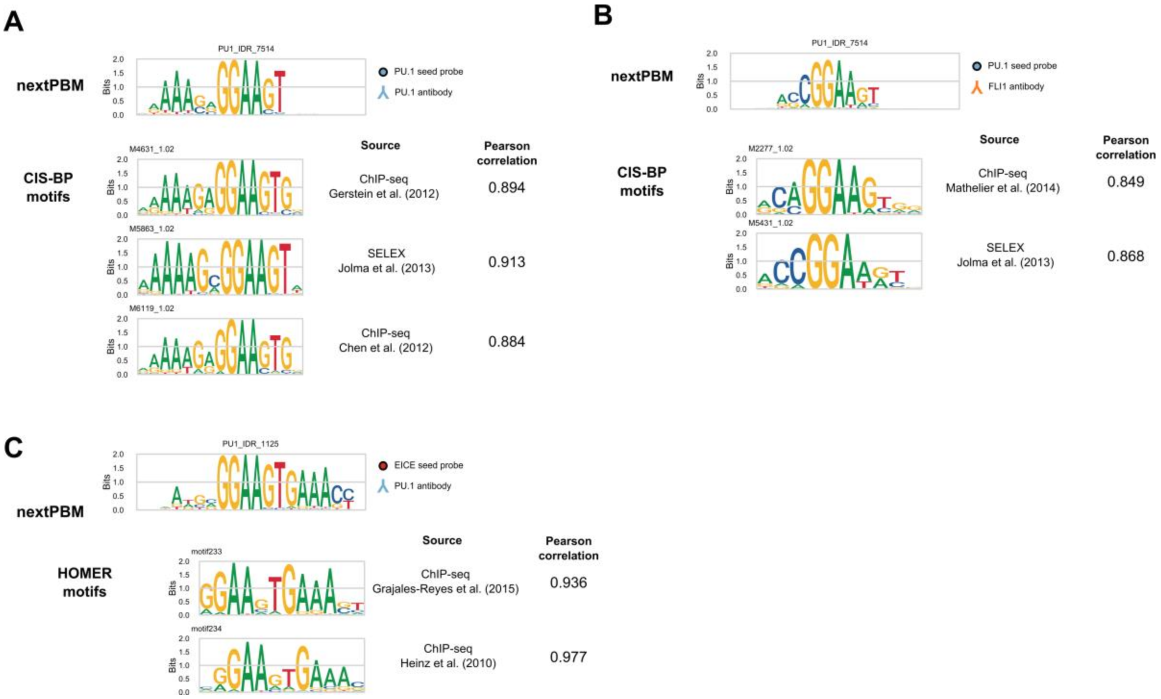
The error vectors ε_1 and ε_2 are then used to compute the PU.1-IRF8 cooperativity scores:

$$scores = \frac{|\varepsilon_2|}{variance(\varepsilon_1)}$$

2.6 Supplementary Information

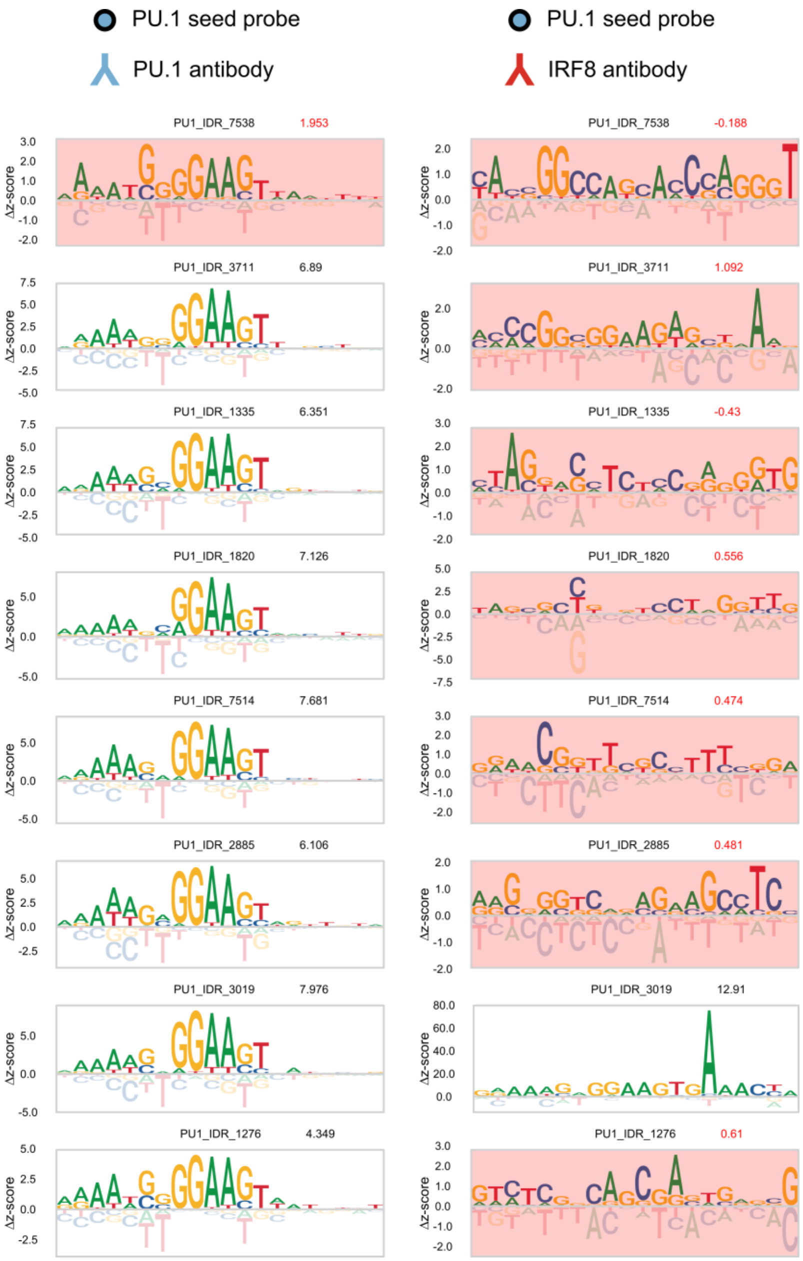


Supplementary Figure 2.1: PU.1 de novo motifs obtained using HOMER and MEME HOMER and MEME (meme-chip) de novo motifs for PU.1 ChIP-seq peaks with cross-replicate IDR < 0.01 (top row) and IDR < 0.05 (bottom row). Pearson correlation summarizing the similarity between the HOMER and MEME motifs for a given IDR threshold is shown.

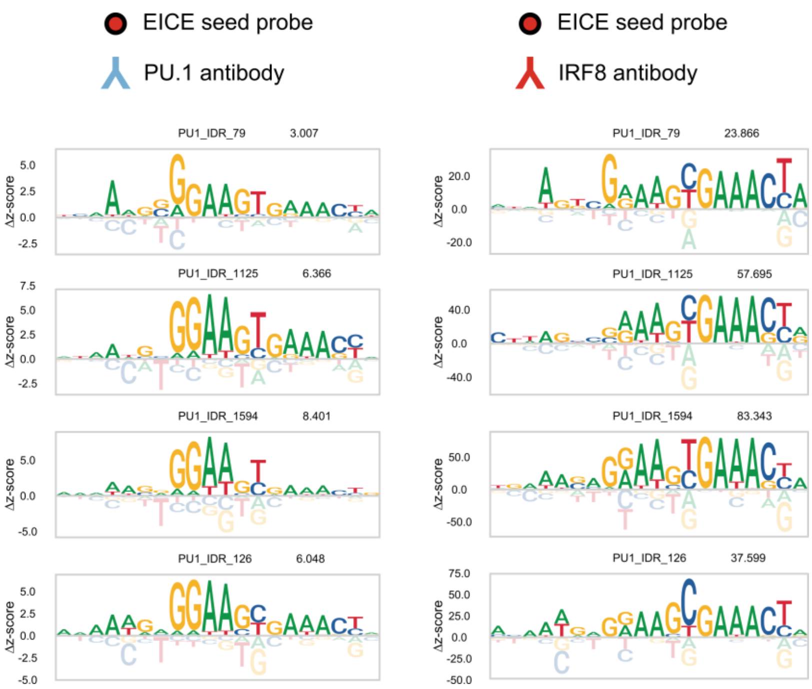


Supplementary Figure 2.2: Similarity between nextPBM binding models and existing database entries

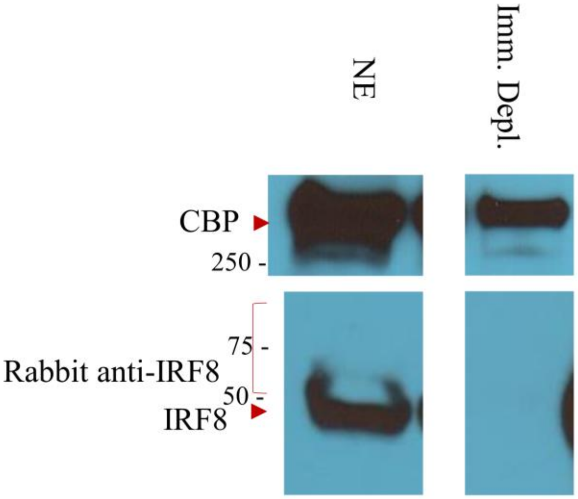
(A) Similarity between a representative PU.1 binding model obtained with nextPBM using the SNV probe approach (see 2.5 Materials and Methods) and PU.1 binding models obtained in previously published in vivo (ChIP-seq) and in vitro (SELEX) investigations (CIS-BP motifs). Similarity is measured using the maximum Pearson correlation computed between the nextPBM PWM and the database PWM. Label above the nextPBM binding model corresponds to the identifier of the ChIP-seq peak from which the PU.1 seed probe was selected (see Supplementary File 1 from Mohaghegh et al., 2019). (B) Similarity between an FLI1 binding model generated as in (A) and FLI1 binding models obtained in previous investigations. Similarity is computed as in (A). (C) Similarity between a representative PU.1 binding model using an EICE seed probe and characterized EICEs from previous investigations. Similarity is computed as in (A) and (B).



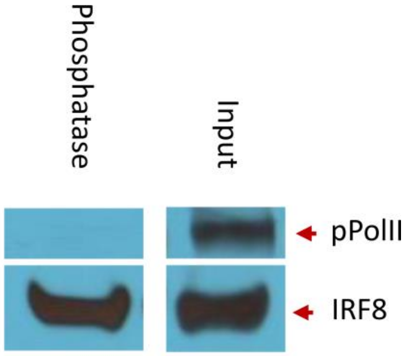
Supplementary Figure 2.3: PU.1 and IRF8 binding models at canonical PU.1 seed probes Left: nextPBM PU.1 binding models obtained using binding of PU.1 to 8 canonical seed probes and each of the single nucleotide variants of the seed sequence. Label above the model corresponds to the identifier of the ChIP-seq peak from which the PU.1 seed probe was selected. The score shown to the top right of each model is the PU.1 binding z-score obtained at the seed probe. Low-scoring models (binding with z-score < 2.0), where PU.1 does not bind well to the seed probe are highlighted in red and the corresponding model is tinted red. The Δz-score values are computed relative to the positional median score (see Methods). Right: IRF8 binding models obtained using the same 8 canonical PU.1 seeds. Low-scoring models (also with z-score < 2.0), where IRF8 does not bind well to the seed probe are also tinted red.



Supplementary Figure 2.4: PU.1 and IRF8 binding models at EICE seed probes
Left: PU.1 binding models obtained using binding of PU.1 to 4 EICE seed probes and each of the SNV probes of the seed sequence. Labels and z-scores are shown as in Supplementary Figure 2.3. Identical z-score thresholds are used as in Supplementary Figure 2.3. Right: IRF8 binding models obtained using the same 4 EICE seeds.



Supplementary Figure 2.5: Immune-depletion of IRF8 from nuclear extract
Western blot comparing IRF8 protein levels in untreated nuclear extract (NE) to NE where IRF8 has been immune-depleted (Imm. Depl.). CBP protein levels were used as a loading control. Each sample includes 60µg of total nuclear extract protein. The Western blot has been cropped for clarity. Contributions: Experimental work was performed by NM.



Supplementary Figure 2.6: Phosphatase assay validation
Western blot comparing protein levels of phosphorylated RNA polymerase II (pPolII) in untreated nuclear extract to nuclear extract treated with broad-spectrum phosphatase. IRF8 protein levels were used as a loading control. The Western blot has been cropped for clarity. Contributions: Experimental work was performed by NM.

CHAPTER THREE: CASCADE – Customizable high-throughput platform for profiling cofactor recruitment to DNA to characterize cis-regulatory elements and screen non-coding single-nucleotide polymorphisms

Note: A substantial portion of this chapter is based on a pre-print manuscript uploaded to the biorXiv with David Bray (DB) and Heather Hook (HH) featured as co-first authors and equal contributors. The CASCADE hybrid experimental/computational technique was jointly conceived by DB, HH, and Trevor Siggers (TS). All the experimental work, including the CASCADE/nextPBM array experiments and validation, was performed by HH. nextPBM and CASCADE microarrays, analysis algorithms, and visualizations were designed and implemented by DB with input from TS. Individual author contributions to figures are noted in each respective figure legend. Supplementary data published alongside the pre-print version of the manuscript can be found with the online biorXiv pre-print.

3.1 Abstract

Determining how DNA variants affect the binding of regulatory complexes CREs and non-coding single-nucleotide polymorphisms (ncSNPs) is a challenge in genomics. To address this challenge, we have developed CASCADE (Comprehensive ASsessment of Complex Assembly at DNA Elements), which is a PBM-based approach that allows for the high-throughput profiling of COF recruitment to DNA sequence variants. The method also enables one to infer the identity of the TF-COF complexes involved in COF recruitment. We use CASCADE to characterize regulatory complexes binding to CREs and SNP quantitative trait loci (SNP-QTLs) in resting and stimulated human

macrophages. By profiling the recruitment of the acetyltransferase p300 and MLL methyltransferase component RBBP5, we identify key regulators of the chemokine CXCL10, and by profiling a set of five functionally diverse COFs we identify a prevalence of ETS sites mediating COF recruitment at SNP-QTLs in macrophages. Our results demonstrate that CASCADE is a customizable, high-throughput platform to link DNA variants with the biophysical complexes that mediate functions such as chromatin modification or remodeling in a cell state-specific manner.

3.2 Introduction

Determining the impact of genetic variation on CREs, such as enhancers and promoters that control gene expression, remains a challenge in modern genomics. GWASs have identified thousands of SNPs associated with human diseases, but the causal variants and their biological effects remain largely unknown (Chen et al., 2016; Gallagher and Chen-Plotkin, 2018; Alasoo et al., 2018). Variants underlying disease risk often function by altering CRE function and gene expression. For example, >50% of causal SNPs for autoimmune diseases are ncSNPs mapping to immune gene enhancers (Farh et al., 2015). Therefore, a major challenge in understanding disease susceptibility is to determine how non-coding DNA variants disrupt CREs. A further challenge is that DNA variants, such as eQTLs, often have effects in a single cell type (Alasoo et al., 2018) or stimulation condition (Schmiedel, et al., 2018; Fairfax et al., 2014). Such studies highlight the need for experimental approaches to characterize the impact and mechanisms of non-coding DNA variants on CRE function in a cell state-specific manner.

Current high-throughput approaches to study the molecular mechanisms by which ncSNPs alter gene expression are based primarily on computational predictions of TF binding (Farh et al., 2015; Harley et al., 2018; Rojano et al., 2019) or on allelic imbalance in genomic assays of TF binding and chromatin state (Harley et al., 2018; Bailey et al., 2015; Buchkovich et al., 2015; Kumasaka et al., 2016; Shi et al., 2016; Valouev et al., 2008). However, these approaches have various limitations. Genomic assays based on allelic imbalance are impractical as a general approach to study candidate ncSNPs because each DNA variant must be present in the assayed cells and each experiment can examine only a single TF or chromatin feature. Computational approaches that use PWM models to assess the impact of ncSNPs on TF binding offer a parallelizable approach, but can predict altered TF binding for only a fraction of ncSNPs (Farh et al., 2015; Soccio et al., 2015). Additionally, PWM-based approaches do not account for changes in TF activity, such as TF nuclear localization or interactions with COFs, that occur in response to cell-state changes and are known to affect ncSNP function (Schmiedel et al., 2018; Fairfax et al., 2014).

3.3 Results

3.3.1 The CASCADE approach

To address these challenges in ncSNP annotation, we have developed CASCADE– a PBM-based high-throughput approach to profile the DNA binding of TF-COF complexes from cell nuclear extracts. PBMs are double-stranded DNA microarrays that allow protein-DNA binding to be assayed to thousands of DNA sequences (Berger et al., 2006; Berger and Bulyk, 2009). Recently, we developed nextPBM approach

(Mohaghegh et al., 2019) to study DNA binding of TFs present in nuclear extracts. However, TFs function by recruiting COFs, which subsequently alter gene expression through diverse mechanisms such as histone modification or chromatin remodeling (Fig. 3.1a) (Kouzarides, 2007). To directly interrogate TF-COF complexes, in CASCADE we extend nextPBM to profile recruitment of COFs to DNA variants using nuclear extracts (Fig. 3.1b). As many COFs, such as the acetyltransferase EP300/CBP, interact broadly with multiple TFs (Vo and Goodman, 2001; Goodman and Smolik, 2000; Janknecht and Hunter, 1996), we can assay many TF-COF complexes in a parallel manner by profiling recruitment of a single COF, without requiring previous knowledge of the TFs involved. Critically, by assaying COF recruitment to SNVs of a DNA sequence, we can determine a COF recruitment motif whose specificity allows us to infer the identity of the TF (or TF family) by comparison against TF motif databases (Fig. 3.1b). Therefore, conceptually, by profiling the recruitment of a limited set of COFs we can characterize the DNA binding of a much larger set of TF-COF complexes. Here, we demonstrate that CASCADE can be used to profile the DNA-sequence dependence of TF-COF complex binding to CREs or ncSNPs in a cell-state specific manner (Fig. 3.1c), providing a high-throughput approach to address the biophysical impact of non-coding DNA variants on gene regulatory complexes.

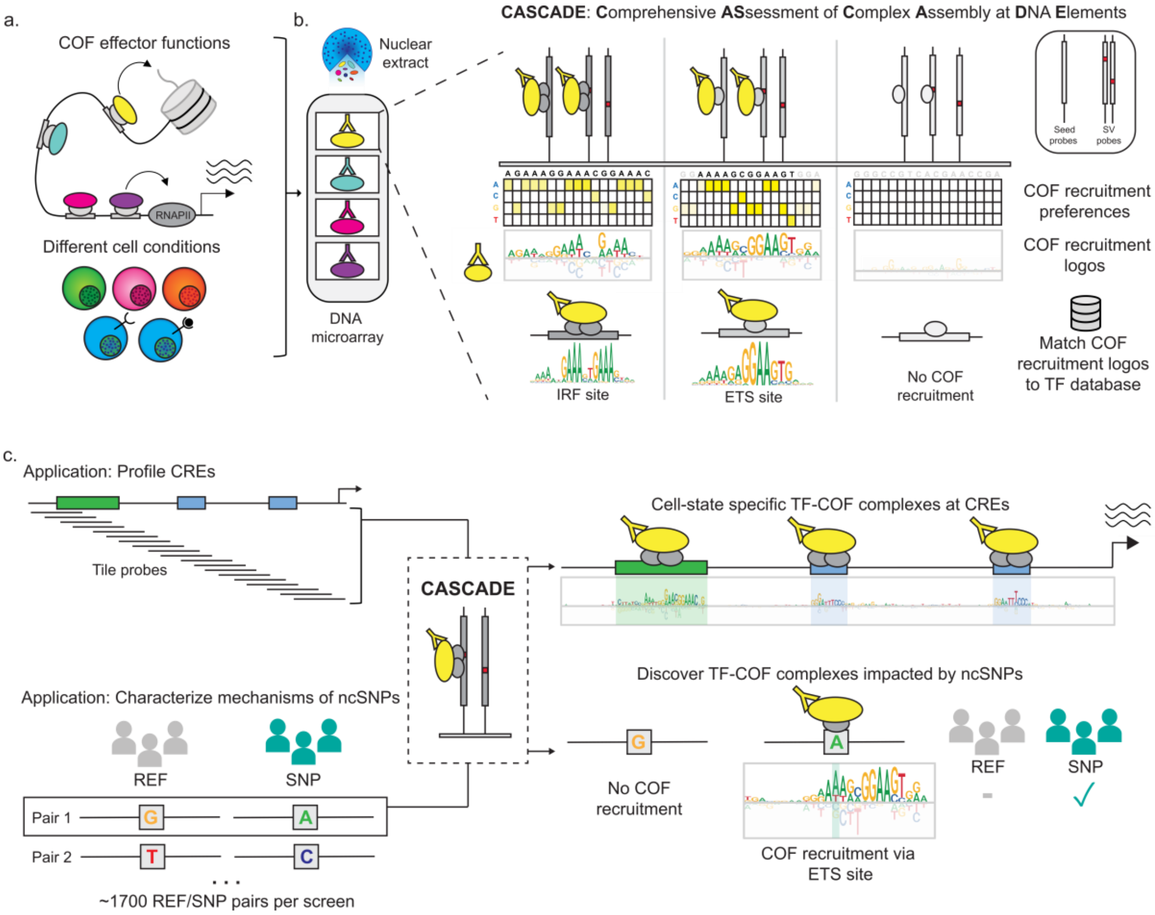

a.
COF effector functions
RNAPII
Different cell conditions
b.
Nuclear extract
DNA microarray
CASCADE: Comprehensive ASsessment of Complex Assembly at DNA Elements
Seed probes
SV pobes
AGAAAGGAAACGGAAAC
AAAAGCGGAAGT
COF recruitment preferences
COF recruitment logos
IRF site
ETS site
No COF recruitment
Match COF recruitment logos to TF database
c.
Application: Profile CREs
Tile probes
CASCADE
Cell-state specific TF-COF complexes at CREs
Application: Characterize mechanisms of ncSNPs
REF
SNP
Pair 1
Pair 2
G
A
T
C
~1700 REF/SNP pairs per screen
Discover TF-COF complexes impacted by ncSNPs
No COF recruitment
COF recruitment via ETS site
REF
SNP
-
✓

Figure 3.1: CASCADE approach and applications
(a) COFs affect transcription and chromatin state. (b) COF recruitment to DNA is assayed by nextPBM. COF recruitment is assayed to a ‘seed’ probe (e.g., genomic-derived TF binding site sequence) and all SV probes. COF recruitment to SV probes yields nucleotide preferences along DNA sequence. Preferences are transformed to COF recruitment motif (i.e., a logo). Motifs are matched to TF motif databases to infer TF identity. (c) Overview of CASCADE applications. CASCADE can be applied to CREs or reference (REF) / ncSNP pairs. For CREs, tiling probes are used to span the genomic region, and COF motifs for each tiling probe are integrated into a CRE-wide COF motif. For ncSNP/REF pairs, COF motifs are determined for both and compared. Contributions: CASCADE concept and workflow was conceived jointly by DB, HH, and TS.

3.3.2 Application of CASCADE to characterize cis-regulatory elements

To demonstrate the use of CASCADE to characterize CREs, we profiled the recruitment of the COF EP300, hereafter p300, to a promoter segment of the chemokine gene CXCL10 in resting and LPS-stimulated human THP-1 macrophages. CXCL10 is important for mediating the inflammatory response by promoting activation and recruitment of several types of immune cells, such as monocytes. The expression of CXCL10 is often dysregulated in autoimmune diseases and has been implicated in cancer pathogenesis (Lee et al., 2009; Liu et al., 2011). In LPS-induced activation of CXCL10 in macrophages, three separate TF binding sites in the promoter are required for full activation, two NF-κB binding sites and an interferon-sensitive response element (ISRE) (Majumder et al., 1998; Ohmori and Hamilton, 1993) (Fig. 3.2a), providing a test case for our CASCADE approach. p300 is a broadly acting acetyltransferase that is recruited by diverse TFs, including both NF-κB and IRF3 that function at the CXCL10 promoter Vo and Goodman, 2001; Majumder et al., 1998; Ohmori and Hamilton, 1993).

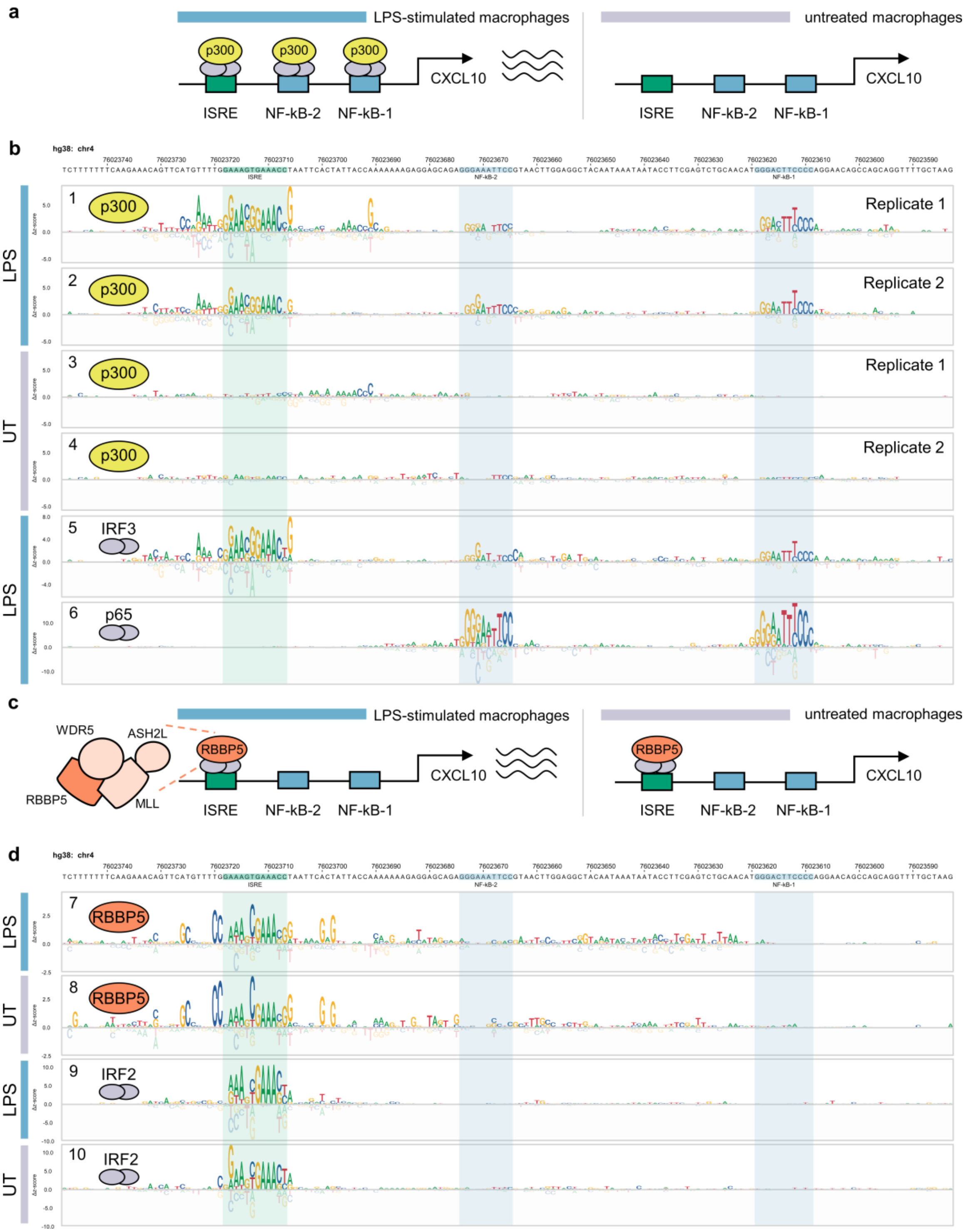


Figure 3.2: CASCADE-based characterization of COF recruitment to the CXCL10 promoter

(a) Schematic of LPS-inducible recruitment of p300 to CXCL10 promoter in macrophages. (b) CRE-wide p300 recruitment motif and TFs IRF3 and p65/RELA across CXCL10 promoter. Experiments using extracts from LPS-stimulated or untreated (UT) macrophages are indicated with colored bars. p300 motifs are shown for biological replicate experiments (Replicate 1 and 2). (c) Schematic of condition-independent recruitment of RBBP5 to CXCL10 promoter. (d) CRE-wide motifs for COF RBBP5 and TF IRF2 across the CXCL10 promoter segment. Experimental conditions as in (b). Contributions: CASCADE experimental work was performed by HH. CASCADE array design, analysis, and visualization was developed and implemented by DB with input from TS.

3.3.3 Characterization of the recruitment of COFs to the CXCL10 promoter

To query p300 recruitment across the CXCL10 promoter segment (166 bp), we assayed recruitment to 29 tiling probes (each 26 bp long) generated at 5 bp intervals across the target promoter region (Fig. 3.1c, see 3.5 Materials and Methods, Supplementary Data 1 from Bray et al., 2020). For each tiling probe on our microarray, we also included all SV probes to allow a COF recruitment motif to be determined every 5 bp (Fig. 3.1c). A CRE-wide p300 recruitment motif was then generated for each experimental condition by integrating these individual motifs across their overlapping positions (Fig. 3.2b, tracks 1-4, see 3.5 Materials and Methods).

Our CRE-wide recruitment motif revealed p300 recruitment to the three previously characterized TF binding sites occurred in an LPS-inducible manner (Fig. 3.2b, tracks 1-4). These results are consistent with previous studies that demonstrated the LPS-inducible binding of IRF3 and NF-κB to the CXCL10 promoter (Ohmori and Hamilton, 1993; Tamura et al., 2008; Medzhitov and Horng, 2009; Sakaguchi et al., 2003; Hagemann et al., 2009). To infer the identity of the TFs involved, we compared the p300 recruitment motifs to a database of previously characterized TF binding motifs (see 3.5 Materials and Methods) and identified IRF3 and NF-κB as high-scoring matches

(Supplementary Fig. 3.1a, track 1, Supplementary Fig. 3.1b, track 2) with known LPS-dependent activity (Smale, 2012; Medzhitov and Horng, 2009). To confirm the binding of NF-κB and IRF3 at these sites, we also performed CASCADE experiments directly for the TFs RELA (the p65 subunit of NF-κB) and IRF3, using antibodies against the TFs instead of p300. p65 bound specifically to the previously characterized NF-κB sites and exhibited the expected DNA binding site specificity (Fig. 3.2b, track 6, Supplementary Fig. 3.2, track 14). IRF3 bound specifically to the ISRE (Sakaguchi et al., 2003; Honda and Taniguchi, 2006) and weakly to the two NF-κB sites, which is consistent with the indirect tethering of IRF3 by NF-κB previously reported in LPS-stimulated macrophages (Fig. 3.2b, track 5) (Ogawa et al., 2005; Leung et al., 2004). Critically, the binding motifs determined for IRF3 (Fig. 3.2b, track 5) and p65 (Fig. 3.2b, track 6) agree strongly with those for p300 (Fig. 3.2b, tracks 1-2) demonstrating that COF recruitment motifs can accurately capture the binding motifs for the underlying TFs.

To determine whether additional COFs with different effector functions are also recruited to the CXCL10 promoter segment, we profiled the recruitment of RBBP5, a core subunit of the MLL histone lysine methyltransferase complex (Fig. 3.2c, Supplementary Data 1 from Bray et al., 2020). Unlike the LPS-inducible recruitment of p300, RBBP5 is constitutively recruited to the CXCL10 promoter sequences at comparable levels in the presence or absence of LPS (Fig. 3.2d, tracks 7-8). RBBP5 is recruited only to the ISRE element, and not the NF-κB sites, demonstrating a different recruitment preference than p300. However, as IRF3 binding to the ISRE is LPS-induced (Fig. 3.2b, track 5, Supplementary Fig. 3.2, track 13), our data suggests recruitment of

RBBP5 to this site is dependent on a different TF. Furthermore, the COF recruitment motifs for p300 and RBBP5 at the ISRE site exhibit clear differences in nucleotide preference (e.g., RBBP5 prefers a 5’-AAANCGAAA-3’ consensus whereas p300 prefers a 5’-GAACGGAAA-3’ consensus; Fig. 3.2b, tracks 1-2, Fig. 3.2d, tracks 7-8). Comparing the RBBP5 recruitment motifs against a TF motif database (see 3.5 Materials and Methods), we identified IRF2 as a high-scoring match (Supplementary Fig. 3.1c, track 7, Supplementary Fig. 3.1d, track 8). IRF2, and the related IRF8, are both constitutively expressed in THP-1 macrophages, which would support the LPS-independent RBBP5 recruitment. CASCADE analysis of both IRF2 and IRF8 yielded CRE-wide motifs that closely matched those obtained for RBBP5 (Fig. 3.2d, tracks 9-10, Supplementary Fig. 3.2, tracks 11-12). These results show that applying CASCADE to different COFs can reveal TF-COF complexes with distinct compositions and DNA-binding specificities.

3.3.4 A two-step CASCADE-based approach to characterize ncSNPs

To investigate the extent to which ncSNPs function by perturbing TF-COF complex binding, we used nextPBM/CASCADE approaches to screen ncSNPs for altered COF recruitment. To increase the number of ncSNPs that we could screen, we developed a hierarchical two-step approach to identify and characterize SNPs that affect binding of TF-COF complexes (Fig. 3.3a). In step one, COF recruitment to pairs of reference and SNP alleles is screened in order to identify variants that lead to significant differential COF recruitment (Fig. 3.3, step 1). In step two, to infer the identity of the TFs involved at each SNP locus, a second microarray is used to perform a CASCADE-based analysis for these significant loci (Fig. 3.3, step 2). The COF recruitment motifs generated for each SNP locus can then be compared to TF motif databases to infer the identity of the TF family and to provide additional context for assessing the impact of each SNP.

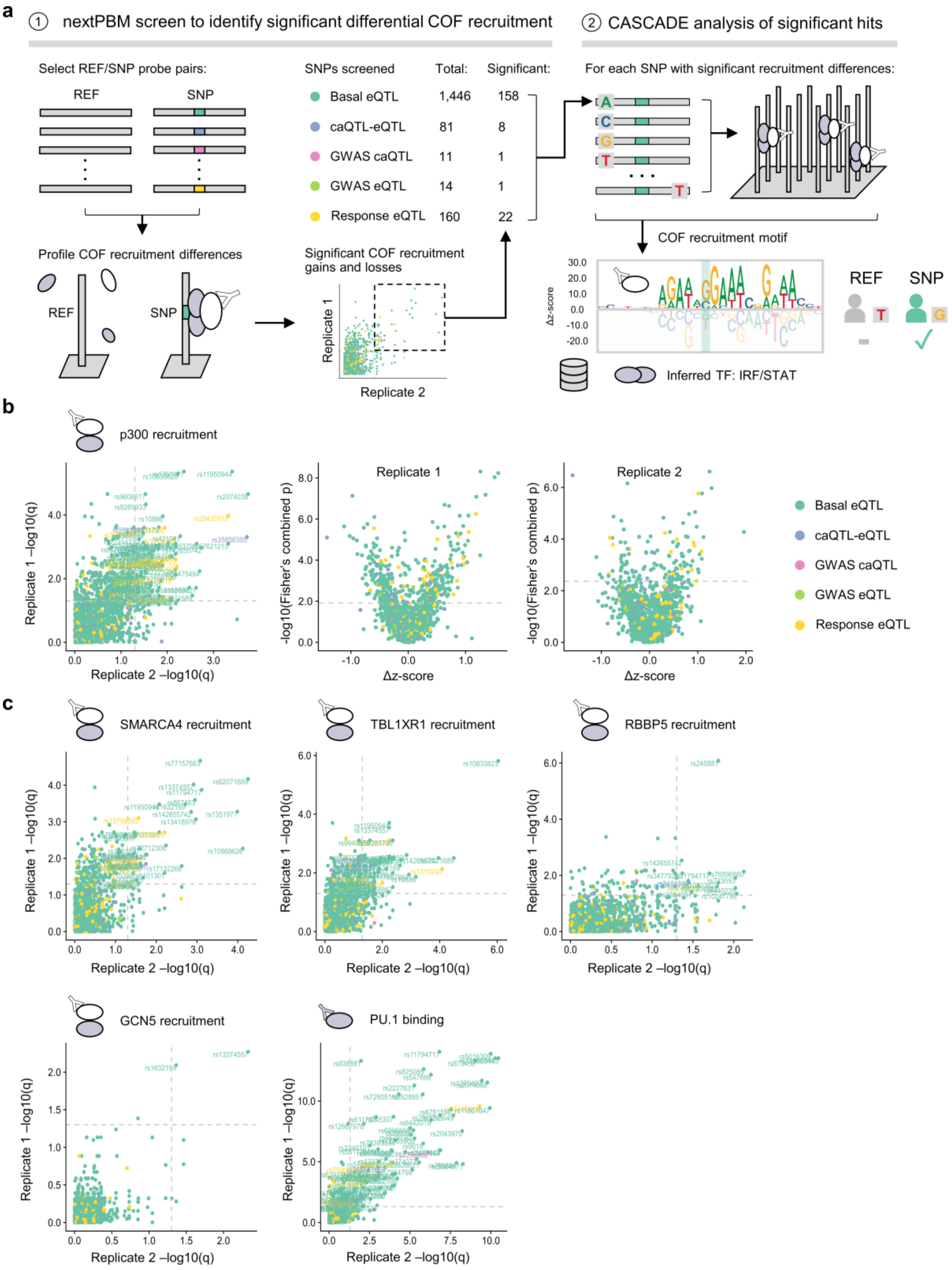


Figure 3.3: CASCADE-based analysis of SNP-QTLs in human macrophages

(a) Overview of 2-step, CASCADE-based approach to characterize 1,712 SNP-QTLs. (1) Step 1: screen for differential COF recruitment to SNP-QTL/REF probe pairs. Number of probe pairs in each QTL class for which significant COF recruitment was identified in at least one experiment. (2) Step 2: CASCADE-based motifs are generated for SNPs identified as significantly bound. COF motifs are compared against TF-motif databases to infer TF identity. (b) Comparison of p300 differential recruitment across biological replicates. Comparison of q-values for replicates is shown (left). Comparison of differential nextPBM z-scores for SNP/REF pairs against p-values (combined across probe orientations – see 3.5 Methods) is shown for replicate experiments (right). Dashed lines represent a -log10(q-value) of 1.3 (equivalent to $q < 0.05$). QTL class for each SNP is indicated. (c) Comparison of differential COF recruitment across biological replicates is shown for candidate COFs and the TF PU.1. Contributions: 2-step differential COF screen and CASCADE follow-up concept was conceived jointly by DB, HH, and TS. Experimental work was performed by HH. Design and analysis of the screening and CASCADE experiments was performed by DB with input from TS and HH.

We used this two-step approach to profile COF recruitment to 1,712 SNP-QTLs associated with gene expression (eQTLs) and chromatin accessibility (caQTLs) changes in myeloid cells (Alasoo et al., 2018; Schmiedel et al., 2018; Fairfax et al., 2014) (Fig. 3.3a, Supplementary Data 2 from Bray et al., 2020). We performed our analysis with nuclear extracts from THP-1 macrophages stimulated with IFN-γ and LPS (see 3.5 Materials and Methods). To assess the impact of SNPs on different cellular functions, we profiled recruitment of five COFs from different functional categories: p300, a histone acetyltransferase; SMARCA4/BRG1, a subunit of the SWI/SNF chromatin remodeling complex; TBL1XR1, a subunit of the nuclear receptor corepressor (NCoR) complex; RBBP5, a subunit of the MLL histone lysine methyltransferase complex; and GCN5, a histone acetyltransferase. In addition to these COFs, we screened for differential binding of the TF PU.1 due to its known role in establishing the myeloid enhancer landscape and the previously demonstrated prevalence of the PU.1 binding motif at macrophage SNP-QTLs (Alasoo et al., 2018; Ghisletti et al., 2010; Heinz et al., 2013).

3.3.5 Screening known myeloid SNP-QTLs for differential COF recruitment

Our step-one screen identified 164 total SNP alleles that reproducibly altered the recruitment of at least one of the tested COFs (Fig. 3.3b, Fig. 3.3c), representing 9.6% of the sites examined.

With the exception of the GWAS caQTL category, comparable proportions of the SNP-QTL categories tested reproducibly altered COF recruitment: 136 basal eQTLs (9.4%), 7 caQTL-eQTLs (8.6%), 1 GWAS eQTL (7.1%) and 20 response eQTLs (12.5%). Profiling the TF PU.1, we also observed widespread differential PU.1 binding at 95 SNP-QTLs (Fig. 3.3c) including 23 that coincided with the differential recruitment of at least one of the COFs screened.

By examining the direction of the differential recruitment, we identified SNPs that caused gain or loss of TF-COF binding (Fig. 3.3b, Supplementary Fig. 3.3). For example, our screen identified 63 SNP alleles that led to statistically significant gain of p300 recruitment (Fig. 3.3b, rightmost two panels, positive Δz-score) and 35 SNP alleles that led to a significant loss relative to the reference allele (Fig. 3.3b, rightmost two panels, negative Δz-score). In total, across all COFs and TFs screened, we observed differential recruitment/binding at 243 of the 1,712 SNP-QTLs (14.2%) with 134 gains, 108 losses, and one SNP demonstrating both. Of note, for each SNP exhibiting significant reproducible differential recruitment of more than one COF (40 total), the direction of the effect, either gain or loss, was consistent across each COF. These results demonstrate that our nextPBM COF-based approach can be used to reproducibly screen broad classes of ncSNPs for both gains or losses of TF-COF complex binding.

3.3.6 Inference of TF families underlying differential cofactor recruitment at ncSNPs

For step two of our SNP analysis we used CASCADE to determine COF recruitment motifs at select loci. These motifs allow us to infer the identity of the TFs mediating differential COF recruitment at each locus (Fig. 3.3a, step 2). We selected 158 basal eQTLs, 8 caQTL-eQTLs, 1 GWAS caQTL, 1 GWAS eQTL, and 22 response eQTLs, as these loci showed significant differential recruitment of one or more of the regulators screened (see 3.5 Materials and Methods, Supplementary Data 3 from Bray et al., 2020). To determine our COF recruitment motifs, we profiled the base preferences of the local genomic region (26 bp) centered at each of these SNP-QTLs. Consistent with our observed differential PU.1 binding, the COF recruitment motifs for many loci matched ETS-type binding motifs (Fig. 3.4). COF motifs were also identified that matched TBX/KLF/EGR zinc finger motifs, IRF/STAT motifs, and two motifs that did not match a known TF motif even at a relaxed stringency threshold (Fig. 4.4, see 3.5 Materials and Methods). Comparing the recruitment motifs generated at a given SNP locus, we found the motif base preferences and alignment were consistent across COF and PU.1 experiments, confirming a common underlying TF-COF complex. Examining SNPs specifically affecting ETS motifs, we found that SNPs can impact different positions along the ETS motif, including both the variable 5' flanking region (rs11940944, rs72755909, rs2526718) and the core ETS 5'-GGAA-3' element (rs873458, rs1250568). These results highlight that COF recruitment motifs can provide a means to understand the biophysical mechanism for a SNP-QTL.

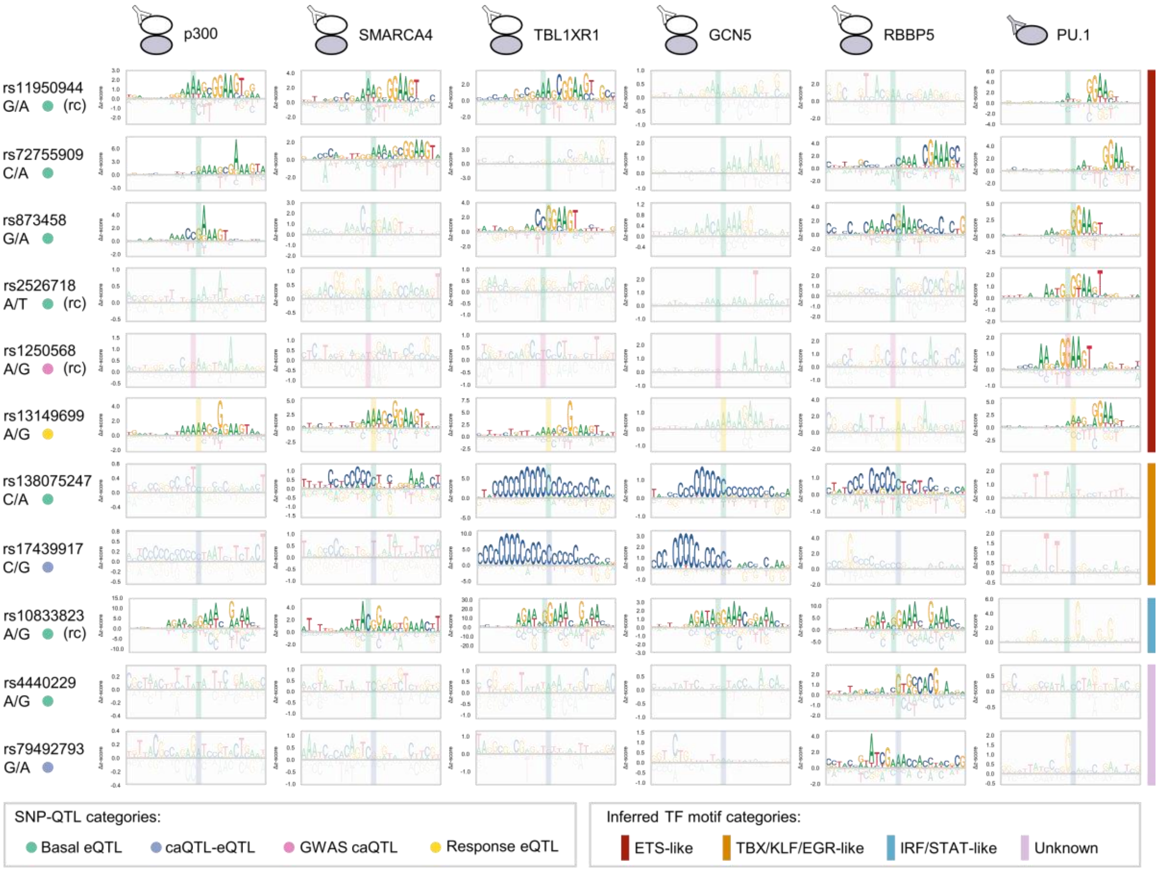


Figure 3.4: CASCADE-determined motifs at SNP loci
COF recruitment motifs for p300, SMARCA4, TBL1XR1, GCN5, and RBBP5 are shown for 10 SNP-QTL loci. PU.1 binding motifs at each locus are also shown. Position of the SNP location within each motif is shown with a shaded rectangle. QTL type of each SNP is indicated (left-hand side, colored dots). Only sites that met an imposed seed z-score threshold were plotted (see 3.5 Materials and Methods). Corresponding reference and SNP are shown beneath each rsID. (-) denotes a site plotted as its reverse complement relative to the reference strand. For these sites, the reference and SNP alleles are also indicated as their complementary nucleotides. Contributions: CASCADE experimental work was performed by HH. CASCADE analysis and motif similarity analysis was performed by DB.

3.3.7 Comparison of TF binding models associated with site-specific COF recruitment preferences

We highlight two gain-of-recruitment SNP-eQTLs identified in our screen to demonstrate how CASCADE can be used to generate mechanistic models of ncSNPs.

Our analysis for rs11950944 (G/A), a basal SNP-eQTL in myeloid cells (Schmiedel et al., 2018), found that p300 (z-score: 2.36), SMARCA4 (z-score: 2.99), and TBL1XR1 (z-score: 2.61) are recruited to the SNP allele but are either not recruited or are below our detection threshold for the reference allele (p300: z-score: - 0.13, SMARCA4: z-score: -0.38, TBL1XR1: z-score: 0.37) (Fig. 3.5a left, Supplementary Data 3 from Bray et al., 2020). The COF recruitment motifs for all three COFs matched significantly with ETS-factor motifs (Fig. 3.5a, right). Consistent with our motif-based inferences, the ETS factor PU.1 preferentially bound the SNP allele (z-score: 5.99) though it could also be detected at the reference allele (z-score: 4.04). These results suggest a model where the SNP allele enhances the DNA binding of an ETS-family TF, possibly PU.1, which leads to enhanced recruitment of these COFs (Fig. 3.5c). We note that enhanced binding of PU.1 at DNA variants in murine myeloid cells has been previously shown to correlate with increased local histone modifications characteristic of primed and active regulatory elements as well as with increased transcriptional output (Heinz et al., 2013).

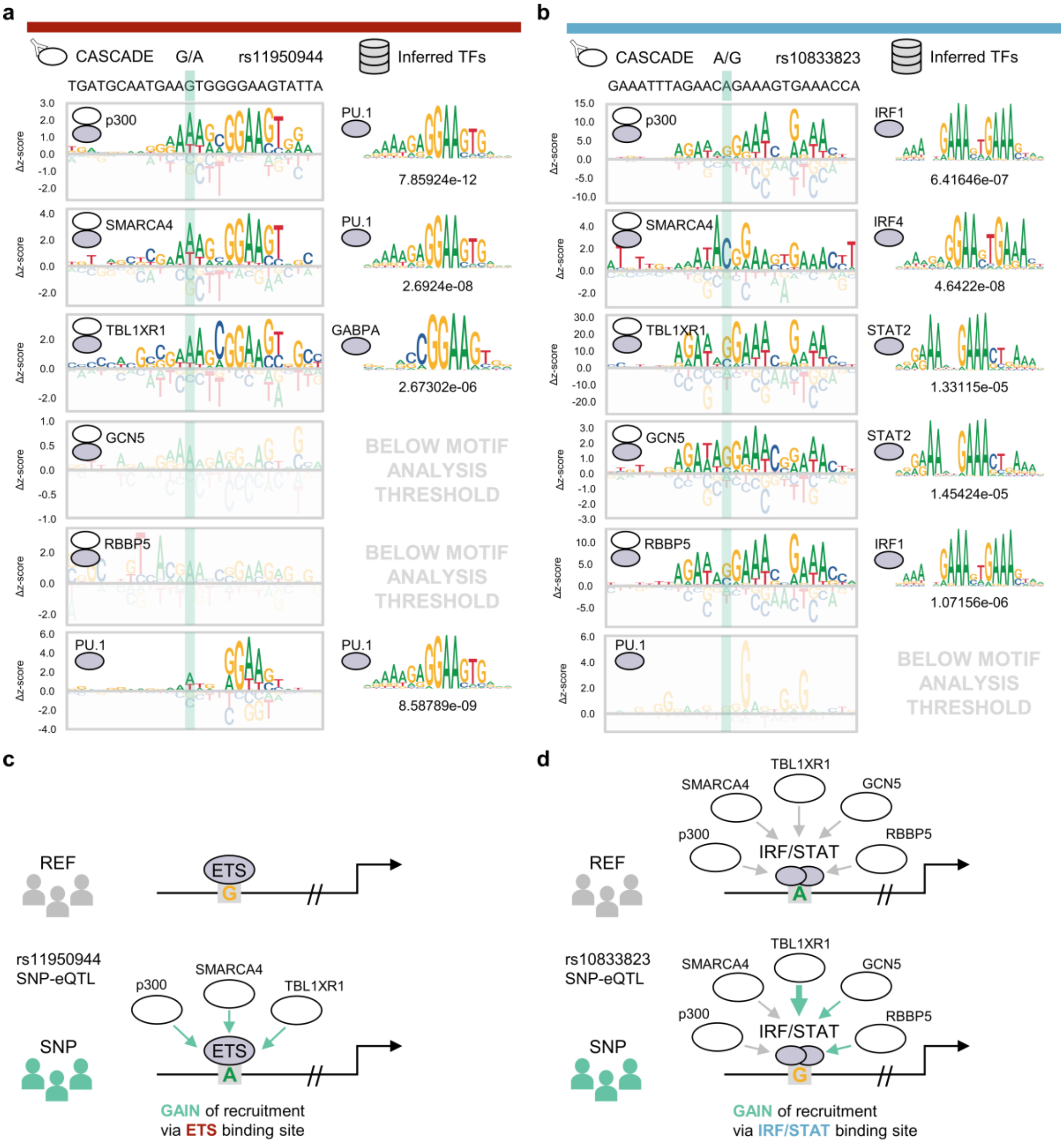


Figure 3.5: Constructing models with CASCADE for SNP-eQTLs
(a) Left column: CASCADE-determined COF recruitment motifs for p300, SMARCA4, TBL1XR1, GCN5, and RBBP5 at the local genomic region surrounding rs11950944. PU.1 binding motif is also shown. Right column: TF binding motif with the strongest association to each corresponding CASCADE COF recruitment motif. Statistical significance (p-value) for TF matching is shown below each TF motif (see 3.5 Materials and Methods). Position of the SNP location within each motif is shown in the shaded area. QTL type and inferred TF category are indicated by the same color scheme as in Fig. 3.4. (b) Same as in (a) but for the local genomic region surrounding rs10833823. Only sites that met an imposed z-score threshold were plotted and used for motif analysis (see 3.5 Materials and Methods). (c) Integrative model for COF

recruitment changes at SNP-eQTL rs11950944. (b) Same as in (c) but for SNP-eQTL rs10833823. Contributions: CASCADE experimental work was performed by HH. CASCADE analysis and motif similarity analysis was performed by DB. DB, HH, and TS jointly interpreted the results to posit mechanistic models.

Our analysis for a second basal SNP-eQTL rs10833823 (A/G) in myeloid cells (Schmiedel et al., 2018) identified a different scenario in which the entire panel of COFs tested were recruited to the reference allele, but the SNP allele caused significantly higher recruitment for three of the COFs: TBL1XR1 (z-scores: WT = 9.71, SNP = 28.49), GCN5 (z-scores: WT = 1.54 to SNP = 3.36), and RBBP5 (z-scores: WT = 10.56 to SNP = 15.82) (Fig. 3.5b left, Supplementary Data 3 from Bray et al., 2020). The COF recruitment motifs for all COFs matched GA-rich IRF/STAT-family motifs (Fig. 3.5b, right), and consistent with our inference of recruitment by IRF/STAT-type TFs, we did not observe PU.1 binding at this site (Fig. 3.5b, left). Notably, while the variant G allele enhanced COF recruitment in our assay, it occurred at a low-information position in the IRF/STAT binding motifs that did not appreciably affect the PWM scores for these TFs (Fig. 3.5b, left, highlighted position; Fig. 3.5b, right). The PWM binding models for several inferred TFs (Fig. 3.5b, right, IRF1, IRF4, STAT2) thereby predict that the variant position (Fig. 3.5b, left, highlighted position) does not affect TF binding but can alter the recruitment of several COFs (Fig. 3.5d) possibly by a mechanism involving DNA-based allostery (Meijsing et al., 2009; Gronemeyer and Bourguet, 2009). The functional consequences of this variant on COF recruitment would thereby not be captured by traditional computational annotation techniques based on PWM motif scanning and predicted differential TF binding (Grant et al., 2011; Coetzee et al., 2015;

Touzet and Varre, 2007; Claeys et al., 2012; reviewed in Gan et al., 2018). It is also possible that binding of these TFs (e.g. IRF1, IRF4, and STAT2) is optimized by the variant G at that specific binding site (flanking rs10833823) which highlights an advantage of generating site-specific recruitment and binding models compared to PWMs that are constructed by aggregating information across many binding sites and genomic contexts. These results demonstrate how the CASCADE approach, based on site-specific COF-recruitment profiling, can generate biophysical, mechanistic models for how ncSNPs can alter the binding of TF-based regulatory complexes.

3.4 Discussion

Characterizing the effects of DNA variants, such as ncSNPs, on gene regulatory complexes is a challenge in our efforts to explain the genetic contributions to human disease. A bottleneck in the field is that studies identifying the mechanisms by which ncSNPs function greatly lag studies identifying ncSNPs associated with traits or diseases (Gallagher and Chen-Plotkin, 2018). To address this need for high-throughput approaches to characterize ncSNPs, we developed CASCADE as a high-throughput, customizable platform for profiling the impact of DNA variants on TF-COF complexes. By measuring the DNA recruitment of broadly interacting COFs (i.e., that form complexes with many TFs), this approach can assay multiple TF-COF complexes in a multiplexed manner. Furthermore, as CASCADE queries the binding of TF-COF complexes, as opposed to just TFs, it can suggest a link between DNA variants and the biological functions mediated by each COF. In this work, we have applied CASCADE to the study of ncSNPs, but the approach can be customized to study any non-coding DNA variants, such as rare variants associated with disease or somatic mutations associated with cancer. We envision that using CASCADE in conjunction with other high-throughput, cell-based methods, such as MPRAs that assess gene expression (Melnikov et al., 2012; Tewhey et al., 2016; Ernst et al., 2016) will provide exciting new approaches to characterize function and mechanism of DNA variants at a genomic scale.

As disease-associated ncSNPs often reside within CREs (Alasoo et al., 2018; Farh et al., 2015; Fairfax et al., 2014; Maurano et al., 2012) the characterization of ncSNPs is directly related to the problem of delineating the mechanisms of CREs. Here, we demonstrate that CASCADE can be applied to this fundamental problem and can be used to identify TF binding sites within CREs and the TF-COF complexes that bind to these sites under different cellular conditions. Using CASCADE to characterize an LPS-inducible segment of the CXCL10 promoter, we identified the three previously validated NF-κB and IRF sites involved and TF-COF complexes bound to each individual site. In this work, we profiled a limited set of COFs, but the approach can be applied to other COFs where native antibodies are available, or COFs have been affinity tagged. We also demonstrated that we can identify site-specific recruitment of COFs that are annotated as subunits of larger, multi-protein COF complexes (e.g., RBBP5, Fig. 3.2c). Currently it is unclear the extent to which these multi-protein COF complexes are assembled on our microarrays, or whether we are assaying the recruitment of smaller sub-complexes or even single COFs (i.e., binary TF-COF interactions). Future studies will address the extent to which recruitment of larger COF complexes is being assayed and how CASCADE-identified TF-COF interactions reflect interactions critical for CRE function *in vivo*. Finally, we note that CASCADE provides the first, high-throughput approach to establish the link between individual CRE binding sites, TFs, and COFs. We anticipate that CASCADE will allow for renewed examination of the role COFs play in the cis-regulatory logic that governs CRE function, which has primarily focused on TFs and binding sites alone.

We recognize that there are technical limitations to the CASCADE approach in the detection of differential COF recruitment events and the motif-based analyses used to infer TFs underlying these differential COF recruitment sites. The paired reference/alternate allele screening procedure successfully identified differential COF recruitment sites that, when profiled using the full CASCADE procedure, demonstrated COF recruitment preferences consistent with previous TF binding models. Although, we acknowledge that agreement across technical replicates of the screening procedure was generally poor in particular when profiling COF recruitment (Fig. 3.3b, Fig 3.3c). We therefore selected only the sites that demonstrated reproducible COF recruitment differences for downstream analyses (see 3.5 Materials and Methods). Furthermore, motif-based inferences identifying TFs underlying differential COF recruitment events could largely only be resolved at the TF family-level (e.g. IRF or ETS). COF recruitment and TF binding motifs, such as those involving IRF3 at the CXCL10 promoter, demonstrated consistent nucleotide preference differences (Fig. 3.2b, tracks 1 and 5) that differed relative to the reference IRF3 binding model (Supplementary Fig. 3.1). Though these may indicate some influence of the p300 cofactor on altering the binding preference of IRF3, we cannot eliminate the possibility in this case that these preference differences are required by IRF3 to bind to the CXCL10 promoter in a site-specific manner that is not captured by a PWM (or similar binding model) that aggregates nucleotide preference information over many binding sites. Furthermore, PWMs for related TFs, such as the IRFs, can closely resemble one another and ignore additional biological contexts such as the cell state-specific activation or upregulation of signal-dependent TFs as well as their potential dimerization partners. For these reasons, associating a site-specific COF recruitment or TF binding model with a specific member of a TF family based on motifs alone remains difficult. In this study, we integrated the motif-level inferences with additional biological insight, such as the known LPS-inducibility of IRF3, to support our predictions. Exploring methods to address these limitations and improve the detection of differential COF recruitment and inference of underlying TFs will be the focus of future investigations. For example, our group has begun exploring alternate screening procedures that include additional replicate probes and allow for differential COF recruitment to be computed across several adjacent sites that we expect will result in more robust detection of differential COF recruitment based on our preliminary experiments (discussed later in Chapter 5, section 5.4). In addition, more objective and quantitative methods to integrate motif-level information with additional biological context, such as expression or activity levels of a TF within the condition being profiled, will be explored in order to improve the inference of TFs underlying observed COF recruitment sites.

The CASCADE approach introduced here is a scalable, customizable platform to study TF-COF complexes and the impact of DNA variants on these gene regulatory complexes. In this study, we demonstrate its application to the functional characterization of CREs and ncSNPs. However, it can be customized and applied to many other types of DNA variants and elements in a cell-specific manner, such as mutations in different cancers or synthetic regulatory elements designed to drive a cell-specific response. We show the application of CASCADE to nuclear extracts from a human macrophage cell line, but conceptually the approach can be used with nuclear extracts from any cell or tissue type. Finally, the ability to profile COF recruitment to DNA sites provides an opportunity to link DNA variants with therapeutic intervention. COFs are often enzymatic (e.g., methyltransferases, histone deacetylases, etc.) and therapeutic inhibitors for many COFs are available (Altucci and Rots, 2016; Cortez and Jones, 2008). Identifying the TF-COF complexes whose binding site is created by a DNA variant may allow for the identification of therapeutic antagonists to counteract their effects. Future studies applying CASCADE in these diverse scenarios should help to develop the approach and provide insights into the roles of TF-COF complexes in cell signaling and disease.

3.5 Materials and Methods

3.5.1 Cell culture

THP-1 cells, a human monocyte cell line, were obtained from ATCC (TIB-202). The cells were grown in suspension in RPMI 1640 Glutamax media (Thermofisher Scientific, Catalogue #72400120) with 10% heat-inactivated FBS (Thermofisher Scientific, Catalogue #11360070) and 1mM sodium pyruvate (Thermofisher Scientific, Catalogue #16140071). T175 (Thermofisher Scientific, Catalogue #132903) non-treated flasks were used when culturing THP-1 cells for experiments. Cells were grown in 50mL of media when being cultured in T175 flasks.

To differentiate THP-1 cells into adherent macrophages, cells were grown to a density of 8.0×10^5 cells/mL and treated with 25ng/mL Phorbol 12-Myristate 13-Acetate (PMA) (Sigma-Aldrich, Catalogue #P8139) for 4 days. Following the 4 days of PMA treatment, the media was replaced with fresh RPMI media with 10% heat-inactivated FBS and 1mM sodium pyruvate. The cells rested for two days in the fresh media before being stimulated with various reagents.

THP-1 cells differentiated with PMA were treated with either LPS (Sigma-Aldrich, L3024) or interferon gamma (IFN-γ) (Thermofisher Scientific, Catalogue #PHC4031) in combination with LPS. PMA treated THP-1 cells were treated with 1ug/mL of LPS for 45 min or with 100ng/mL IFN-γ for 2 h followed by 1ug/mL LPS for 1 h. For each condition, nuclear lysates were harvested. For all nuclear lysates assayed using PBM experiments, the expression levels of COFs and TFs profiled with CASCADE were confirmed by western blotting (Supplementary Fig. 3.4).

3.5.2 nextPBM experimental methods

The nuclear extract protocols are as previously described (Mohaghegh et al., 2019). Changes to the previously published protocols are detailed. To harvest nuclear extracts from THP-1 cells, the media was aspirated off and the cells were washed once with 1X PBS (Thermofisher Scientific, cat #100010049). Once the 1X PBS used to wash the cells was aspirated off, enough 1X PBS was mixed with 0.1mM Protease Inhibitor (Sigma-Aldrich, cat #P8340) to cover the cells was added to each flask. A cell scraper was then used to dislodge the cells from the flask. The cells were collected in a Falcon tube and placed on ice. To pellet the cells, the cell volume was centrifuged at 500xg for 5 min at 4°C. Once the cells were pelleted, the supernatant was aspirated off. The pellet was resuspended in Buffer A and incubated for 10 min on ice (10mM HEPES, pH 7.9, 1.5mM MgCl, 10mM KCl, 0.1mM Protease Inhibitor, Phosphatase Inhibitor (Santa-Cruz Biotechnology, Catalogue #sc-45044), 0.5mM DTT (Sigma-Aldrich, Catalogue #4315)) to lyse the plasma membrane. After the 10 min incubation, a final concentration of 0.1% Igepal detergent was added to the cell and Buffer A mixture and vortexed for 10 sec. To separate the cytosolic fraction from the isolated nuclei, the sample was centrifuged at 500xg for 5 min at 4°C. The cytosolic fraction was collected into a separate microcentrifuge tube. The pelleted nuclei were then resuspended in Buffer C (20mM HEPES, pH 7.9, 25% glycerol, 1.5mM MgCl, 0.2mM EDTA, 0.1mM Protease Inhibitor, Phosphatase Inhibitor, 0.5mM DTT, and 420mM NaCl) and then vortexed for 30 sec. The nuclei were incubated in Buffer C while mixing at 4°C. To separate the nuclear extract from the nuclear debris, the mixture was centrifuged at 21,000xg for 20 min at 4°C. The nuclear extract was collected in a separate microcentrifuge tube and flash frozen using liquid nitrogen. Nuclear extracts were stored at -80°C until used for experiments.

Microarray DNA double stranding and PBM protocols are as previously described (Berger et al., 2006; Berger and Bulyk, 2009; Mohaghegh et al., 2019). Any changes to the previously published protocols are detailed. Double-stranded microarrays were pre-wetted in HBS (20mM HEPES, 150mM NaCl) containing 0.01% Triton X-100 for 5 min and then de-wetted in an HBS bath. Next the array was incubated with nuclear extract for

1 h in the dark in a binding reaction buffer (20mM HEPES, pH 7.9, 100mM NaCl, 1mM DTT, 0.2mg/mL BSA, 0.02% Triton X-100, 0.4mg/mL salmon testes DNA (Sigma-Aldrich, cat #D7656)). The array was then rinsed in an HBS bath containing 0.1% Tween-20 and subsequently de-wetted in an HBS bath. After the protein incubation, the array was incubated for 20 min in the dark with 20ug/mL primary antibody for the TF or COF of interest (Supplementary Table 3.1). The primary antibody was diluted in 2% milk in HBS. After the primary antibody incubation, the array was first rinsed in an HBS bath containing 0.1% Tween-20 and then de-wetted in an HBS bath. Microarrays were then incubated with 10ug/mL of either alexa488 or alexa647 conjugated secondary antibody (Supplementary Table 3.1) for 20 min in the dark. The secondary antibody was diluted in 2% milk in HBS. Excess antibody was removed by washing the array twice for 3 min in 0.05% Tween-20 in HBS and once for 2 min in HBS in coplin jars as described above. After the washes, the array was de-wetted in an HBS bath. Microarrays were scanned with a GenePix 4400A scanner and fluorescence was quantified using GenePix Pro 7.2. Exported fluorescence data were normalized with MicroArray LINEar Regression (Berger et al., 2006).

3.5.3 CASCADE microarray designs and analyses

A known LPS-responsive segment of the CXCL10 promoter (hg38: chr4) from 76023583 to 76023748 was used for the basis of this array design (Majumder et al., 1998). The genomic region was tiled through using 26-base “target” probe sequences with a 5-base step forward between sequential tiles. In total, 29 of these tile probes were needed to span the LPS-responsive CXCL10 promoter segment. “Target” sequences corresponding to the genomic locus were obtained from the hg38 genome fasta file included with Bowtie2 (Langmead and Salzberg, 2012) using the “fastaFromBed” function from bedtools v2.26.0 (Quinlan and Hall, 2010). For each tile probe and each position along the corresponding 26-base target region, a probe was included in the array design consisting of each possible nucleotide variant (at that position) in order to employ the variant probe analysis approach (see below). A total of 2,291 targets were therefore used to model the CXCL10 promoter segment (29 tiles + 29 × 3 variant probes x 26 positions). 500 additional 26-base target regions were randomly selected from the hg38 using the bedtools “shuffleBed” function and included in the array design to build a background distribution of fluorescence intensity. Each 26-base target region in the array design was embedded in a larger 60-base PBM probe as follows:

“GCCTAG” 5’ flank – 26-base target region – “CTAG” 3’ flank –
“GTCTTGATTCGCTTGACGCTGCTG” double-stranding primer

Each target region was included in its reference (+) orientation as well as the reverse complement (-) orientation. 5 replicate spots of each probe (in each orientation) were included in the final array design. PBM microarray probes, relevant annotation for each, and the experimental results are provided (Supplementary Data 1 from Bray et al., 2020). The microarrays were purchased from Agilent Technologies Inc. (AMAID: 085605, format: 8×60K).

To design the nextPBM-based screen for differential COF recruitment at ncSNPs, the lead SNPs uncovered in previous studies were included in our high-throughput screen as follows: 1,446 basal eQTLs (Schmiedel et al., 2018) (randomly selected from the “classical monocytes” category), 81 caQTL-eQTLs, 11 GWAS caQTLs, 14 GWAS eQTLs, and 160 response eQTLs (Alasoo et al., 2018). Chromosomal coordinates (hg38) for each SNP were obtained using the biomaRt R package from Ensembl (Durinck et al., 2009). 26-base DNA probe target regions centered at the SNP position (relative to + strand: 13 bases + SNP location + 12 bases) were obtained for each reference (REF) allele using bedtools as above. For each REF allele probe, a probe with the corresponding SNP allele was also included in the design such that each rsID is represented by a pair of REF and SNP probes. 500 background target regions were also included using the same procedure as above. The 26-base target regions were embedded in larger 60-base PBM DNA probes as above. 5 replicates of each probe (in both orientations) were included in the final design. The microarrays were purchased from Agilent Technologies Inc. (AMAID: 085920, format: 8×60K).

Each REF/SNP pair was screened for differential recruitment of p300, SMARCA4, TBL1XR1, RBBP5, and GCN5 as well as differential binding of representative ETS factor PU.1 nextPBM experimental results were preprocessed as above. Z-scores were obtained for each probe as previously described (Andrilenas et al., 2018) against the distribution of fluorescence intensities obtained at the set of background probes for a given experiment. For each REF and SNP allele pair in the design, a t-test was used to compare the fluorescence intensity distributions between the 5 REF probes and 5 SNP probes for a given COF/TF assayed. To mitigate the influence of probe orientation-specific effects, t-tests were performed independently for each probe orientation with the p-values combined using Fisher’s method. The Benjamini-Hochberg method was used to adjust the individual p-values for a REF/SNP pair for multiple hypothesis testing. The fluorescence intensity z-score difference for a given REF and SNP allele probe pair (termed Δz-score) was computed by subtracting the mean REF z-score from the mean SNP z-score such that a positive Δz-score represents a gain-of-recruitment introduced by the SNP allele and a negative Δz-score represents a loss. Scatterplots based on the screening results (Fig. 3.3b-c) were plotted using the ggplot2 (Wickham, 2016), RColorBrewer (Neuwirth, 2014), and cowplot (Wilke, 2019) R packages. A full data file including the statistics from the high-throughput differential recruitment screen is included in the supplementary materials (Supplementary Data 2 from Bray et al., 2020).

Reference and SNP allele pairs exhibiting reproducible significant differential COF recruitment and/or TF binding were selected for this CASCADE array design in order to infer regulators responsible for the differential activity observed. Inclusion criteria was as follows: the difference in recruitment (or binding) of a given COF (or TF) between corresponding REF and SNP allele probes must have obtained an adjusted p-value (q-value) < 0.05 independently in both technical replicates with a concordant direction of effect. Single variant probes for the 26-base target regions (centered at the SNP position – as described above) were generated using the same procedure as above but without the tiling needed to span larger genomic loci such as the CXCL10 promoter segment used previously. In addition, only 291 background probes were included due to probe number limitations. PBM microarray probes, relevant annotation for each probe, and the experimental results are provided (Supplementary Data 3 from Bray et al., 2020). The microarrays were purchased from Agilent Technologies Inc. (AMAID: 086248, format: 4×180K).

Motif modeling using SV probes was performed as previously described (Mohaghegh et al., 2019; Andrilenas et al., 2018; Penvose et al., 2019) for the SNP-QTL sites profiled in detail using CASCADE. For the multi-tile design used to model extended loci such as the LPS-responsive CXCL10 promoter segment, a weighted mean approach was applied as follows to overlapping positions in order to integrate results across sequential tiles: all variant probes corresponding to a given nucleotide at a given position within the promoter segment were averaged using each probe's corresponding seed (reference genomic) z-score as a weight. Further, if a given SV probe's z-score was above 1.645 (above approximately 95% of the fluorescence intensities obtained using background probe distribution - assuming a normal distribution) and the SV probe's corresponding reference probe z-score was less than or equal to 1.645, the SV probe's z-score was reset to the reference seed value. This procedure ensured that the SV probe modeling approach was used to characterize true genomic recruitment sites and reduce the influence of COF recruitment sites gained specifically via a non-reference (non-genomic) variant. Sequence logo plots for the COF recruitment and TF binding motifs were generated using the ggseqlogo R package (Wagih, 2017) and arranged using cowplot (Wilke, 2019). The Δz-scores of each nucleotide represent the difference relative to the median z-score obtained across all possible nucleotides at that position and was computed after the weighted averaging procedure described previously. The Δz-score axis limits for the logo tracks (Fig. 3.2, Supplementary Fig. 3.2) were determined using the minimum and maximum Δz-scores obtained for a given COF/TF (across experiments within an array design) to enable comparisons across stimulus conditions assuming matched total protein concentrations across experiments.

3.5.4 Motif similarity analysis

For CASCADE recruitment motifs obtained at the CXCL10 locus, to simplify the analysis and reduce the number of comparisons, the promoter segment was first separated into 3 motifs broadly corresponding to each previously characterized TF site (ISRE, NF-κB-2, and NF-κB-1). For CASCADE profiling of the SNP-QTLs, a minimal seed z-score of 1.5 was enforced for motif analysis. Recruitment energy matrices obtained from CASCADE cofactor profiling (fluorescence intensity z-scores) were converted to a probability-based matrix using the Boltzmann distribution as previously described (Andrilenas et al., 2018) to be more directly comparable to previous TF binding models:

$$P_{ik} = \frac{e^{\beta z_{ik}}}{\sum_{k=1}^{4} e^{\beta z_{ik}}}$$

z_{ik} is the z-score for nucleotide variant *k* at position *i* within the motif window. *β* transformation parameters for the Boltzmann equation were scaled using the maximum z-score obtained in a given experiment using the following equation in order to account for differences in antibody efficiencies across cofactors:

$$\beta = \frac{30}{\max(z)}$$

Resulting position-weight matrices were compared against the complete HOCOMOCOv11 database (Kulakovskiy et al., 2018) of TF binding models (771 total) using TOMTOM from the MEME suite (Gupta et al, 2007) version 5.0.3. Euclidean distance was used as the similarity metric with a relaxed minimal reporting q-value of 0.25 (-dist ed -thresh .25).

3.5.5 Data availability

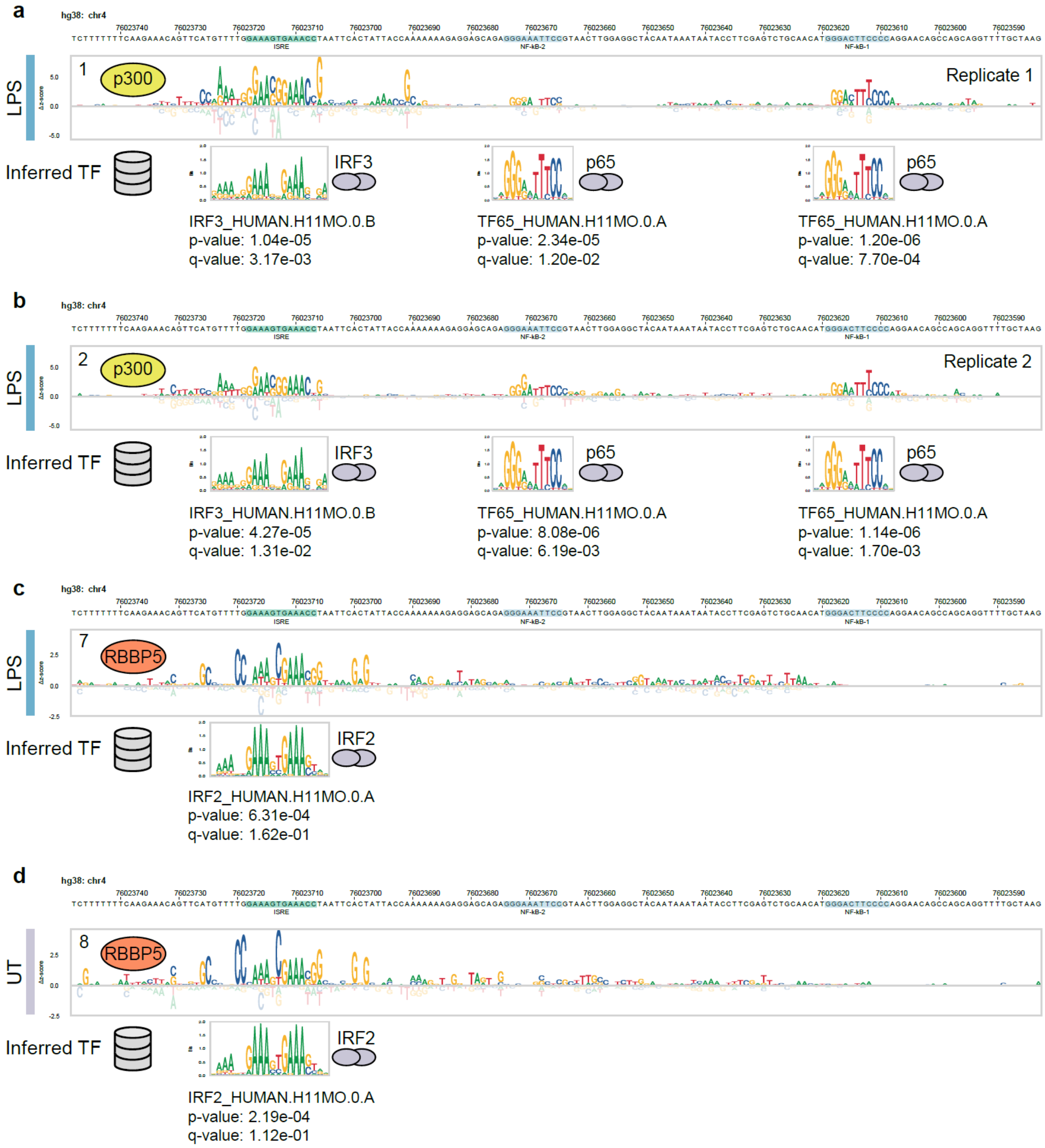
The results of all nextPBM/CASCADE array experiments performed here have been deposited in the Gene Expression Omnibus (GEO accession: GSE148945). An R script that implements CASCADE to generate the plots shown in this study using the Supplementary Data files has been made available on Github (https://github.com/Siggers-Lab/CASCADE_paper). All other data is available upon request.

3.6 Supplementary Information

Antibody	Catalog Number	Application
Primary Antibodies		
P300	ab14984	PBM Experiment/Western Blot
SMARCA4	sc17796	PBM Experiment/Western Blot
GCN5	sc-365321x	PBM Experiment/Western Blot
RBBP5	a300-109A	PBM Experiment/Western Blot
TBLX1R1	sc-100908	PBM Experiment
HDAC1	ab7028	PBM Experiment
P65	sc-372X	PBM Experiment
P65	sc-8008	Western Blot
IRF8	sc-6058X	PBM Experiment
IRF3	D83B9	PBM Experiment
IRF2	sc-374327	PBM Experiment
PU.1	sc-352X	PBM Experiment
Secondary Antibodies		

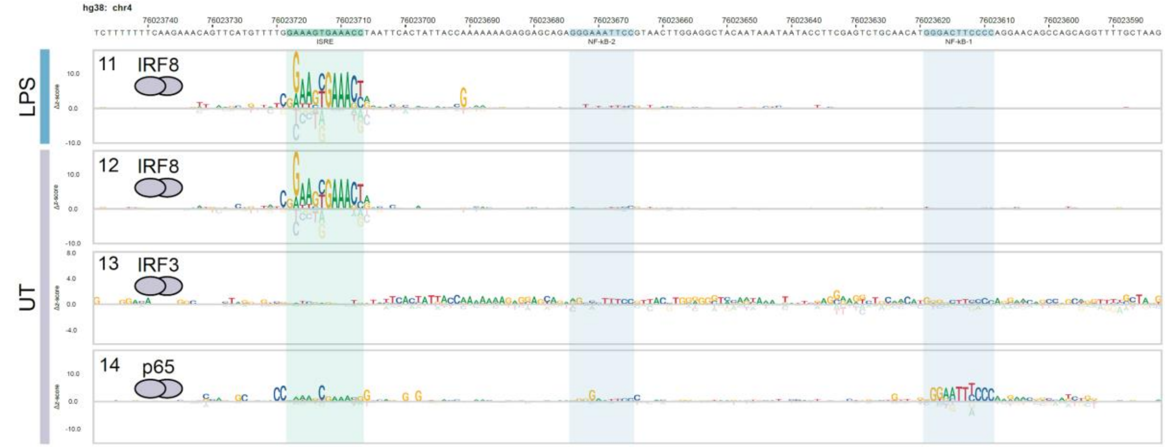
Donkey anti-goat IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 488	A11055	PBM Experiment
Goat anti-mouse IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor 488	A11029	PBM Experiment
Goat anti-rabbit IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor 488	A11034	PBM Experiment
Goat anti-mouse IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor 647	A32728	PBM Experiment
Goat anti-rabbit IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor 647	A32733	PBM Experiment
HRP conjugated Goat anti-mouse	G-21234	Western Blot
HRP conjugated Goat anti-rabbit	G-21040	Western Blot

Supplementary Table 3.1: Antibodies used for experiments
The antibodies listed were used for the PBM experiments or Western blots as listed.



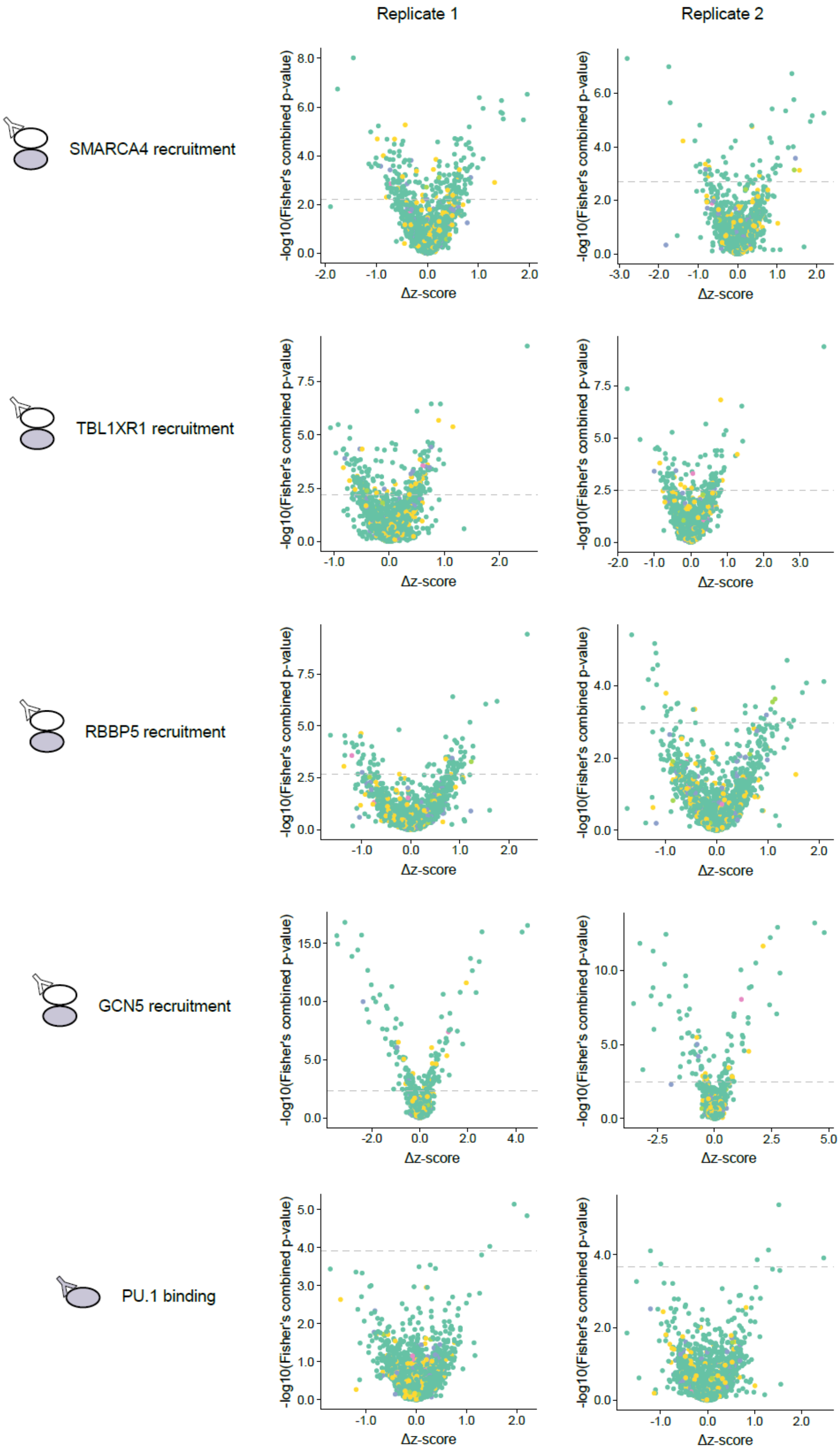
Supplementary Figure 3.1: Model-based inference of transcription factors associated with CXCL10 promoter COF recruitment motifs

(a) TF motifs matched to p300 recruitment preferences in LPS-stimulated macrophages (Replicate 1). (b) TF motifs matched to p300 recruitment preferences in LPS-stimulated macrophages (Replicate 2). (c) TF motif matched to RBBP5 recruitment preferences in LPS-stimulated macrophages (d) TF motif matched to RBBP5 recruitment preferences in untreated macrophages. All COF recruitment preference tracks were converted to probability-based models (see 3.5 Materials and Methods) prior to comparison. Similarity comparisons to known TF binding models was performed using TOMTOM and the full HOCOMOCOv11 motif database (771 total motifs – see Materials and Methods).

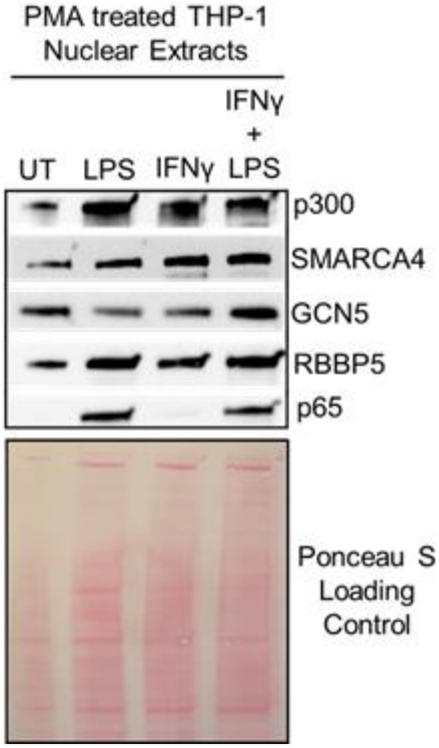


Supplementary Figure 3.2: Additional CASCADE-based analyses of TF binding to the CXCL10 promoter segment

Nucleotide binding preferences of IRF8 to the CXCL10 promoter segment in paired LPS-stimulated (track 11 - continued from Fig. 3.2) and untreated (track 12) macrophages. Binding preferences of IRF3 (track 13) and p65 (track 14) in UT macrophages.


Replicate 1
Replicate 2
SMARCA4 recruitment
TBL1XR1 recruitment
RBBP5 recruitment
GCN5 recruitment
PU.1 binding
-log10(Fisher's combined p-value)
Δz-score
0.0
2.0
4.0
6.0
8.0
-2.0
-1.0
0.0
1.0
2.0
-3.0
2.5
5.0
7.5
3.0
10.0
15.0
-2.5

Supplementary Figure 3.3: Statistical significance and direction-of-effect for changes in COF recruitment and TF binding across reference and SNP probe pairs screened
Rows represent volcano plots obtained for different COFs (SMARCA, TBL1XR1, RBBP5, GCN5) and TF PU.1. Left column shows the volcano plots obtained in a first replicate and right column shows the volcano plots obtained in a technical replicate experiment. Statistical significance threshold for each experiment (q < 0.05, see 3.5 Materials and Methods) is shown as a grey dashed line.



Supplementary Figure 3.4: Western blot of PMA-treated THP-1 NEs
The protein expression levels of p300, SMARCA4, GCN5, RBBP5, and p65 of PMA treated THP-1 cells were evaluated by western blotting. 30ug of nuclear extract were loaded for all samples. PMA treated THP-1 cells were treated with LPS for 45 min to induce p65 expression. PMA treated THP-1 cells were treated with IFNγ for 3 h to prime the immune response. PMA treated THP-1 cells were treated with IFNγ for 1 h and LPS were treated with IFNγ for 2 h followed by LPS stimulation for 45 min. Ponceau S staining was used as a loading control. Contributions: Experiments were performed by HH and interpreted jointly by HH, DB, and TS.

CHAPTER FOUR: The human TF array – surveying cofactor recruitment to the binding sites of human transcription factors

Note: The human transcription factor (hTF) microarray design idea was jointly conceived by David Bray (DB), Heather Hook (HH), Rose Zhao (RZ) and Trevor Siggers (TS) based on a pilot series of microarray designs and analyses by RZ and Jessica Keenan (JK). All pilot experimental work on the immune TF-centric coregulator recruitment (CoRec) array series of experiments prior to the development of the hTF array was performed by RZ and JK with input from TS. The pilot experimental work on the hTF microarray design was performed by RZ with input from HH and TS. The specific hTF array design, software used to design the array, and interactive array analysis and visualization software were designed and implemented by DB with input from HH, RZ, and TS. Individual author contributions to figures are noted in each respective figure legend.

4.1 Abstract

COFs recruited to TF binding sites (TFBSs) are the effectors of gene regulatory activities such as histone modification and chromatin remodeling. Despite the importance of TF-COF complex assembly on the activity of a TF, existing assays to measure TF activity have focused traditionally only on TF binding or on the presence/translocation of TFs into the nucleus. We propose that a more functionally relevant assessment of TF activity is to identify TF-COF complexes present in the cell nucleus and capable of binding to DNA. As a general framework to map such TF-COF complexes in cells, we extend our COF recruitment profiling approaches to survey COF recruitment to an

expansive panel of 346 non-redundant consensus TFBSs and their SNVs. This array design (the human TF – or hTF array) allows users to examine the activity of a wide panel of human TFs by profiling TF-COF complexes in any cell type or condition. To facilitate the application of this experimental approach, we have developed an analysis suite to allow users to explore COF recruitment results obtained using the hTF microarray. In pilot experiments exploring LPS-dependent coactivator and corepressor recruitment to the hTF panel, we use our analysis software to generate condition-specific COF recruitment comparisons at the DNA probe-level, visualize the sequence determinants of COF recruitment at the TF-level, and delineate condition-specific TF-COF complex "signatures" across all TFs included in our panel. We anticipate that the hTF platform and the concept of TF-COF recruitment signatures will provide a valuable annotation layer to refine our understanding of cell- and state-specific gene regulatory logic.

4.2 Introduction

COFs present at CREs such as promoters and enhancers are key effectors in gene regulation. Individual COFs and multi-protein COF complexes recruited to DNA by sequence-specific TFs have diverse roles in histone modification, chromatin remodeling, and assembling the transcription preinitiation complex through their enzymatic activities (Kouzarides, 2007; Vo and Goodman, 2001). As COF recruitment to cell type- or context-dependent CREs can depend on the expression, nuclear localization, and PTMs of signal-dependent TFs (Zabidi and Stark, 2016; Haberle and Stark, 2018; Reiter et al., 2017), a complete understanding of gene regulation depends on our ability to determine which TF-COF complexes are capable of assembling at CREs under diverse cellular conditions.

To address important limitations to existing methods to profile the assembly of TF-COF complexes at DNA, our group recently developed the CoRec (Cofactor Recruitment) method (Keenan, 2019). CoRec is an HT extension to our PBM-based COF recruitment approaches such as CASCADE (Bray et al., 2020). Prior to CoRec, assays to investigate the TFs active in a given cellular context have been limited to TF activation profiling arrays that assay the presence of up to hundreds of TFs in cell nuclei (Luminex 200 from Active Motif; Zhou et al., 2017; Ding et al., 2013). Unlike CoRec, these methods do not explicitly interrogate TF-COF complexes nor do they quantify the effects of DNA variants on the binding of these complexes. Methods such as M2H allow for the characterization of binary interactions between proteins but do not capture specific differences attributable to different cell-specific conditions (Riegel et al., 2017). Moreover, cell-based HT methods such as COF ChIP-seq provide stimulus-specific genome-wide maps of COF locations (Raisner et al., 2018; Blow et al., 2010; Ramos et al., 2010) but the peaks can span hundreds of bases and multiple TF binding motifs making it difficult to infer causality between TF binding and TF-COF complex assembly. Furthermore, it is difficult to manipulate the cellular and genomic context, as in assays such as ChIP-seq, to further probe the sequence- or context-dependent determinants of COF recruitment and TF-COF complex assembly beyond the nucleotides occurring naturally in the genome.

To address the aforementioned limitations of these assays, the CoRec approach was developed as an HT cell-based method to survey recruitment of a COF of interest to a panel of known TFBSs in a given cellular context. To demonstrate the utility of the method we profiled the recruitment of a diverse set of COFs (including coactivators, corepressor subunits, chromatin remodeling enzyme subunits, and more) to a panel of 91 known immune-centric TFBSs in different immune cell contexts (Keenan, 2019). By comparing the TF-COF complexes active in resting macrophages, LPS-stimulated macrophages, resting T cells, and TCR-stimulated T cells, we characterized and recapitulated known key differences in cell- and stimulus-dependent TF-COF complexes (Keenan, 2019).

Here we present the hTF array which extends our pilot CoRec array design, that was focused on 91 immune-related TFBSs, to a panel of 346 non-redundant consensus TFBSs selected algorithmically to represent the known binding repertoire of the human TFs. In addition to the standardized and open-source microarray design, we present an hTF array analysis framework designed to provide researchers with rapid insight into the TF-COF complexes active in different cell states through interactive data analysis and visualization modules. We demonstrate the utility of the hTF array and the dedicated interactive analysis software by characterizing the diversity of TF-COF complexes active in a pilot series of COF recruitment experiments in human macrophages. Furthermore, we demonstrate how the interactive analysis modules can be used in conjunction to investigate differences in COF recruitment logic at the DNA probe-level, the TF-level, and the array-level. The hTF array and accompanying analysis framework thereby represent an attempt to expand our group's COF recruitment profiling techniques in order to enable researchers to investigate the TF-COF complexes active in their system of interest (beyond the immune context) as well as enable the development of TF-COF biomarkers/signatures. We propose that these TF-COF signatures will provide a valuable annotation layer in our efforts to understand context-dependent gene regulation and understand the TF-COF complexes mediating aberrant cell states in disease.

4.3 Results

4.3.1 The human TF array – scaling CoRec up to a general panel of TFBSs

Given the initial success in using the pilot CoRec array design to profile cell type- and stimulus-dependent COF recruitment to a panel of 91 immune-related TF sites (Keenan, 2019) we sought to expand the approach to a larger panel of TFs relevant beyond the immune context. To this end, we have designed the hTF array algorithmically in order to cover as many of the TFs as we could while ensuring accurate COF recruitment models (see 4.5 Materials and Methods). Briefly, starting with the 452 non-redundant human TF binding models from the open-source JASPAR 2018 CORE (Khan et al., 2018), we further collapsed this set down to 346 non-redundant consensus sequences since multiple related TF binding models can be represented by a single consensus DNA microarray probe (Fig. 4.1, step 1). For each of these 346 non-redundant consensus sequences, on the final microarray design, we included probes corresponding to the consensus sequence itself as well as every SNV along the consensus sequence in order to allow us to determine COF recruitment motifs (as with CASCADE in Chapter 3). This probe design allows us to profile COF recruitment by any TFs that can bind one of our 346 consensus probes representative of the human TF binding repertoire. This hTF probe set can be accommodated in the Agilent 4x180K microarray format, which contains 4 replicate copies of each probe set on a single array. Therefore, using this microarray design, one can profile COF recruitment in a multiplexed fashion (Fig. 4.1, step 2). The example in the schematic shows profiling the recruitment of 2 different COFs of interest across 2 cell types in a single set of experiments (Fig. 4.1, step 2). TF-COF recruitment is profiled in a cell state-specific manner using a nextPBM-based approach as with CASCADE in Chapter 3. The hTF array design thereby allows for multiplexed COF recruitment investigations across an expansive and diverse panel of 346 TFs that can be used across any cell type or state of interest.

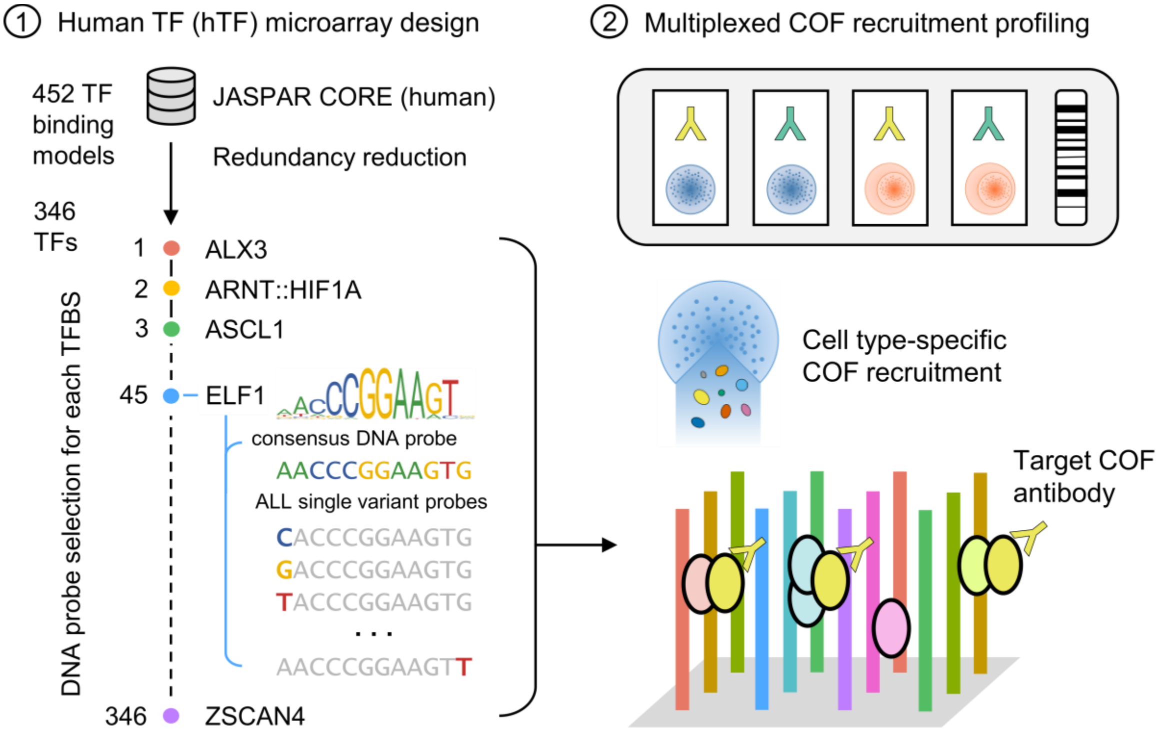


Figure 4.1: hTF array design overview
(1) Consensus binding sites from JASPAR CORE set of human TF binding models were redundancy reduced. For each consensus site, DNA probes corresponding to the site itself as well as all possible single nucleotide variants form the basis of the hTF array design. (2) The 4-

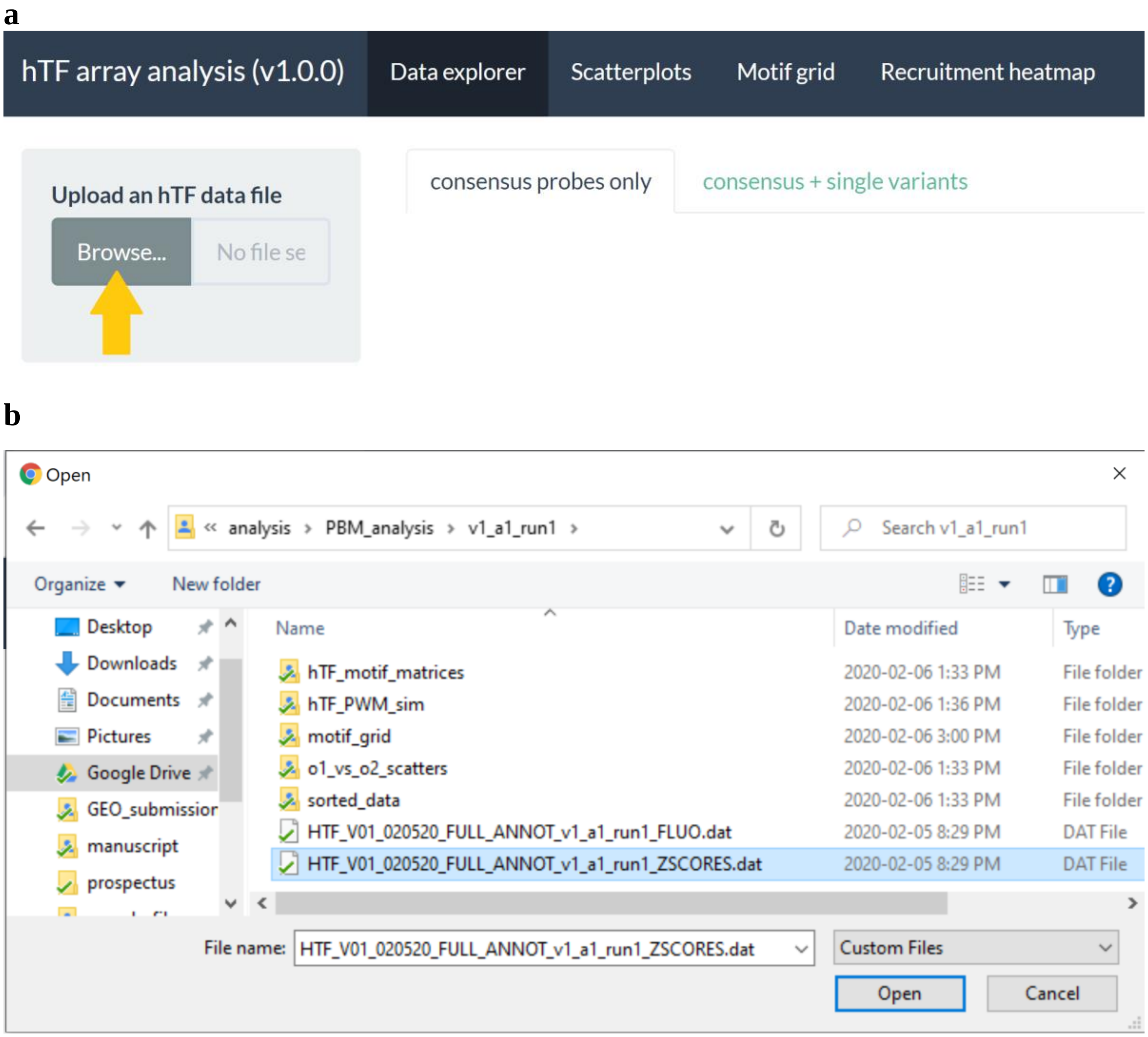
chamber microarray design allows for multiplexed COF recruitment profiling against the same core set of TFBS and SNV probes.

4.3.2 A dedicated interactive software suite to analyze hTF array data

As the hTF array was designed to be applied across cell types, conditions, and COFs of interest for any research application of interest, we sought to create an interactive software suite in order to analyze new results and integrate with previous experiments. In order to make analysis of hTF array data more interactive and accessible to the researchers who perform the experiments, we developed the hTF array analyzer in R Shiny. Through a series of interactive modules such as a dedicated data explorer, pairwise scatterplot visualizations, a grid of COF recruitment motifs, and an experiment-wide recruitment heatmap, the hTF array analyzer software allows researchers to quickly explore their results and gain valuable insight into TF-COF recruitment phenomena within their cell types or states of interest.

4.3.3 Exploration of hTF array experimental results using the array analyzer

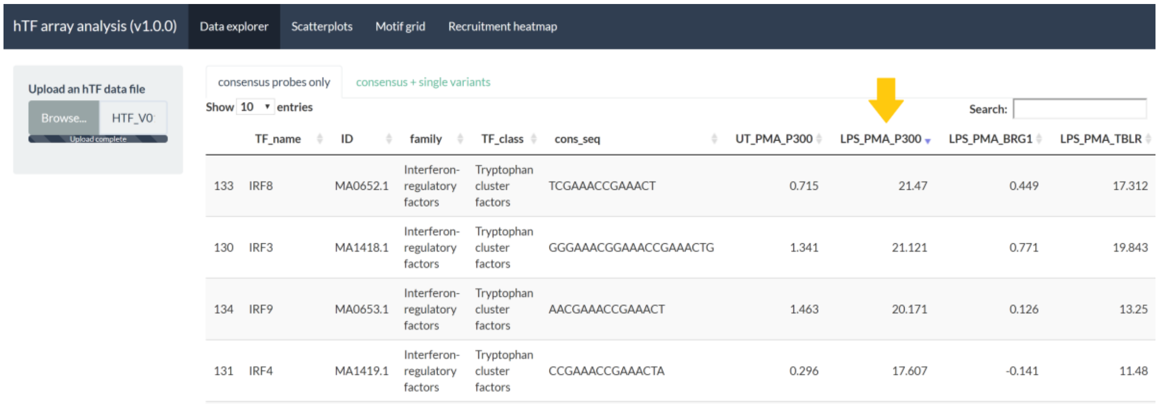
To begin using the hTF array analyzer, experimental data can be uploaded by selecting the “Browse…” button within the “Data explorer” tab (Fig. 4.2a, orange arrow). Clicking the “Browse…” button opens a file browser where the user is prompted to select their formatted hTF array experimental results (Fig. 4.2b). In this example, the user is uploading real pilot experimental data obtained using the hTF array design previously detailed (Fig. 4.2b). Specific metadata regarding the experimental details and microarray design, such as which TF sites are included and which COFs were profiled, are read directly from the uploaded data file. Opening a dataset within the analyzer generates an interactive table where the user can explore the full details of their COF recruitment data (Fig. 4.2c). Each of the 346 TF probe sets present on the pilot hTF array design contains an entry in the table and can be searched for using the “Search” bar at the top right of the analyzer window. Optionally, the user may select the “consensus + single variants” tab which displays the experimental results for each individual probe in the array design, not just the consensus TF sites.



Name	Date modified	Type
hTF_motif_matrices	2020-02-06 1:33 PM	File folder
hTF_PWM_sim	2020-02-06 1:36 PM	File folder
motif_grid	2020-02-06 3:00 PM	File folder
o1_vs_o2_scatters	2020-02-06 1:33 PM	File folder
sorted_data	2020-02-06 1:33 PM	File folder
HTF_V01_020520_FULL_ANNOT_v1_a1_run1_FLUO.dat	2020-02-05 8:29 PM	DAT File
HTF_V01_020520_FULL_ANNOT_v1_a1_run1_ZSCORES.dat	2020-02-05 8:29 PM	DAT File

File name: HTF_V01_020520_FULL_ANNOT_v1_a1_run1_ZSCORES.dat

c



	TF_name	ID	family	TF_class	cons_seq	UT_PMA_P300	LPS_PMA_P300	LPS_PMA_BRG1	LPS_PMA_TBLR
133	IRF8	MA0652.1	Interferon-regulatory factors	Tryptophan cluster factors	TCGAAACCGAAACT	0.715	21.47	0.449	17.312
130	IRF3	MA1418.1	Interferon-regulatory factors	Tryptophan cluster factors	GGGAAACGGAAACCGAAACTG	1.341	21.121	0.771	19.843
134	IRF9	MA0653.1	Interferon-regulatory factors	Tryptophan cluster factors	AACGAAACCGAAACT	1.463	20.171	0.126	13.25
131	IRF4	MA1419.1	Interferon-regulatory factors	Tryptophan cluster factors	CCGAAACCGAAACTA	0.296	17.607	-0.141	11.48

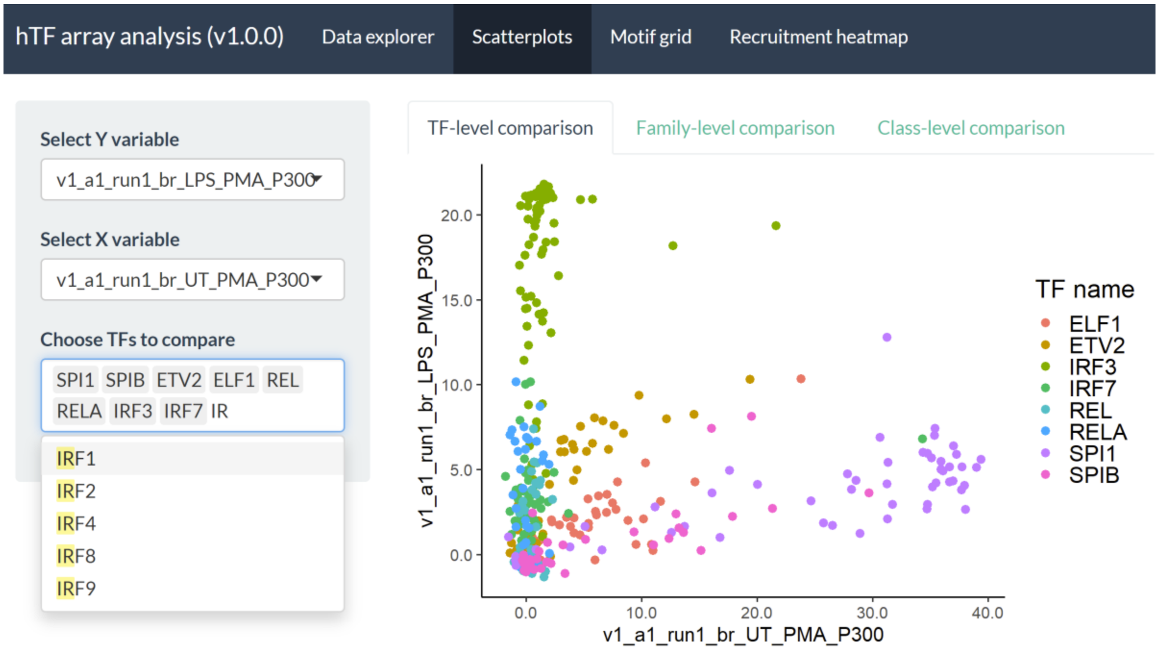
Figure 4.2: Exploring experimental data with the hTF array analyzer software
(a) Data can be uploaded to the hTF array analysis analyzer using the "Browse…" button (orange arrow). (b) Selecting "Browse…" opens a dialog box that prompts the user to select their z-score normalized hTF array results. (c) The interactive "Data explorer" is automatically generated once data has been loaded. Users can interact with their experimental results by sorting the columns by recruitment strength to the consensus site (orange arrow) or using the search feature to subset the full table. Contributions: pilot hTF experiments were performed by RZ with input from DB and TS. DB designed the hTF array, the analysis pipeline, and the interactive analyzer software.

As many of the possible applications of the hTF array platform are discovery-based in nature, an interactive sorting method is implemented in the "Data explorer" so that users may explore which TF sites produced the highest COF recruitment z-scores in the experiments performed. For example, sorting the TF entries by the "LPS_PMA_P300" column, a label assigned to an experiment used to profile the recruitment of coactivator p300 in PMA-differentiated macrophages stimulated with LPS, the highest consensus probe z-scores are obtained overwhelmingly by the IRFs (Fig 4.2c, orange arrow) which are known LPS-responsive TFs (O'Neill, 2006; Sakaguchi et al., 2003; Jefferies 2019, Lawrence and Natoli, 2011; Mogensen, 2019). The "Data explorer" module allows users to gain immediate insight into interesting TF-COF recruitment phenomena. In the same example, though the IRFs appear to be strong recruiters of p300 in the LPS-stimulated cells, in the column labeled “UT_PMA_P300”, an unstimulated control cell population, these same consensus sites do not appear to recruit p300 (Fig. 4.2c). As there are 346 total TF models to consider, the “Data explorer” tab provides an initial portal through which users can interact with their data and decide which TFs to include in the downstream analysis steps.

4.3.4 Generation of probe-level pairwise COF recruitment comparisons

Common analyses performed in TF-COF recruitment experiments are pairwise comparisons of COF recruitment data at the DNA probe level in order to visualize the relative recruitment preferences of different COFs to the full array of TFBSs, or the cell state-dependent recruitment of a given COF. To facilitate these types of pairwise analyses, the hTF array analyzer includes a “Scatterplots” module. Selecting the “Scatterplots” panel opens a page where the user is prompted to select two experiments to compare as Y and X variables in a dynamically generated scatterplot. For example, to explore whether TFs of interest recruit the COF p300 in a condition-specific or constitutive manner, a user can select the LPS-stimulated macrophage experiment as the Y variable and the experiment from untreated macrophages as the X variable (Fig. 4.3a) from the set of experiments included in their uploaded data. The TF selection box is implemented with an auto-complete feature to facilitate searching for a TF of interest from the full hTF array design (Fig. 4.3a). With each new TF site selected (or deleted), the scatterplot is dynamically regenerated to include changes in the z-score axes limits as well as the TF site color palette (Fig. 4.3a).

a



Select Y variable

v1_a1_run1_br_LPS_PMA_P300

Select X variable

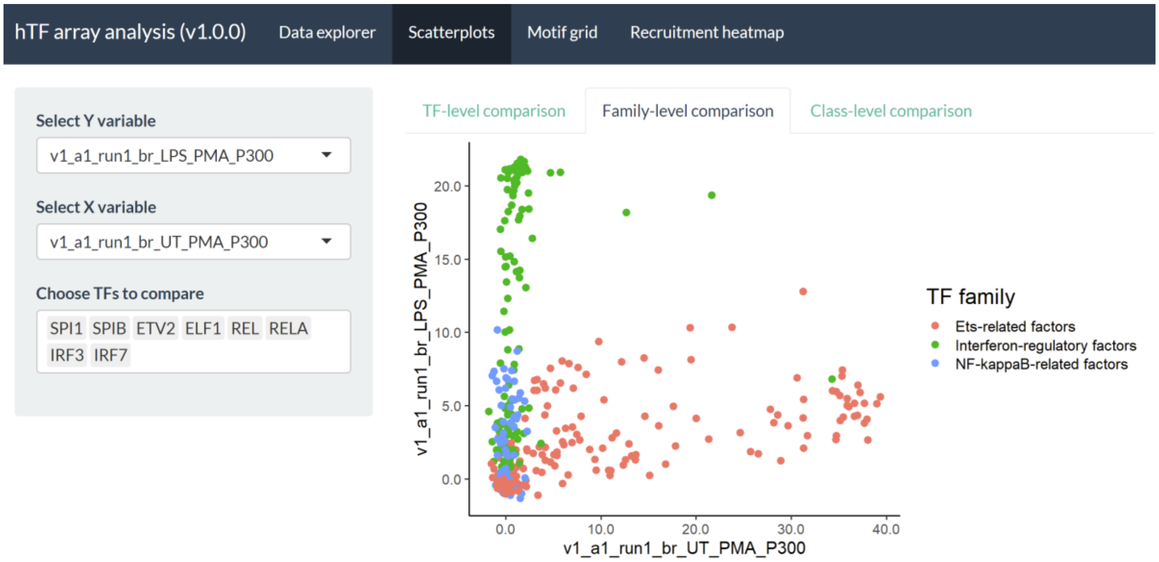
v1_a1_run1_br_UT_PMA_P300

Choose TFs to compare

SPI1 SPIB ETV2 ELF1 REL RELA IRF3 IRF7 IR

IRF1
IRF2
IRF4
IRF8
IRF9

b



Select Y variable

v1_a1_run1_br_LPS_PMA_P300

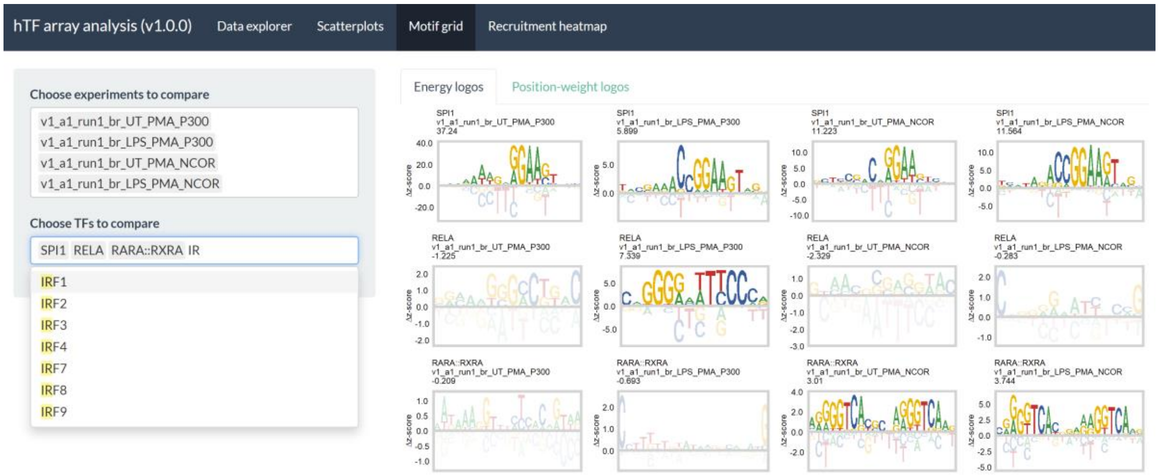
Select X variable

v1_a1_run1_br_UT_PMA_P300

Choose TFs to compare

SPI1 SPIB ETV2 ELF1 REL RELA IRF3 IRF7

c


hTF array analysis (v1.0.0)
Data explorer
Scatterplots
Motif grid
Recruitment heatmap
Energy logos
Position-weight logos
SPI1
v1_a1_run1_br_UT_PMA_P300
37.24
SPI1
v1_a1_run1_br_LPS_PMA_P300
5.899
SPI1
v1_a1_run1_br_UT_PMA_NCOR
11.223
SPI1
v1_a1_run1_br_LPS_PMA_NCOR
11.564
RELA
v1_a1_run1_br_UT_PMA_P300
-1.225
RELA
v1_a1_run1_br_LPS_PMA_P300
7.339
RELA
v1_a1_run1_br_UT_PMA_NCOR
-2.329
RELA
v1_a1_run1_br_LPS_PMA_NCOR
-0.283
RARA::RXRA
v1_a1_run1_br_UT_PMA_P300
-0.209
RARA::RXRA
v1_a1_run1_br_LPS_PMA_P300
-0.693
RARA::RXRA
v1_a1_run1_br_UT_PMA_NCOR
3.01
RARA::RXRA
v1_a1_run1_br_LPS_PMA_NCOR
3.744
Δz-score

Choose experiments to compare

v1_a1_run1_br_UT_PMA_P300
v1_a1_run1_br_LPS_PMA_P300
v1_a1_run1_br_UT_PMA_NCOR
v1_a1_run1_br_LPS_PMA_NCOR

Choose TFs to compare

SPI1 RELA RARA::RXRA IR

IRF1
IRF2
IRF3
IRF4
IRF7
IRF8
IRF9

Figure 4.3: Interactive TF-COF recruitment analyses using the hTF array analyzer
(a) Scatterplots demonstrating a pairwise probe-level comparison between p300 in LPS-stimulated macrophages (y-axis) versus unstimulated macrophages (x-axis) for a subset of TFs listed. The probe colors correspond to the consensus site and all single variant probes for the TFs listed. (b) Same as in (a) but at the TF family level. (c) Interactive motif grid displaying the full COF recruitment models for p300 recruitment in unstimulated (column 1) macrophages, LPS-stimulated (column 2), and for NCoR recruitment in unstimulated (column 3) macrophages and LPS-stimulated (column 4) for SPI1/PU.1 (top row), RELA/p65 (middle row), and RARA::RXRA (bottom row). The JASPAR CORE 2018 reference model for each TF is shown in column 5. The search field has an implemented "autocomplete" feature that suggests TF sites based on what is currently being typed (highlighted yellow). Contributions: hTF pilot data was generated by RZ with input from DB and TS. DB designed the hTF array, analysis pipeline, and interactive software.

By selecting various TF sites interactively within the "Choose TFs to compare" box, a researcher can readily identify which TFs (or TF families) likely recruits a given COF under the analyzed conditions (Fig 4.3a, p300 recruitment example). To indicate general data trends, we display the consensus site and all single nucleotide variant probes for a given TF as a single color. In the example shown, the probes representing the IRF model (Fig. 4.3a, IRF3 and IRF7) have low p300 recruitment z-scores in the untreated condition and high z-scores in the LPS condition, illustrating the strongly LPS-dependent p300 recruitment sites which is consistent with the consensus site-level results previously highlighted in the “Data explorer” tab (Fig. 4.3c). In contrast, ETS factor probes representing the models for ELF1 and ETV2 appear to recruit p300 in a more constitutive manner (i.e., these ETS probes have similar z-scores in both resting and stimulated conditions) whereas p300 recruitment to SPI1 and SPIB probes occurs more exclusively in unstimulated macrophages (Fig. 4.3a). We note that ETS factors have been previously demonstrated to interact with the CBP/p300 coactivators (Vo and Goodman, 2001; Yang et al., 1998).

In more involved analyses with increased numbers of TFs that may obfuscate the automatically generated color palette, a user can instead use the “Family-level comparison” tab to automatically assign colors to all consensus and single variant probes associated with the different TF families included in the pairwise analysis (Fig. 4.3b). As an additional layer of abstraction, the analyses can also be performed at the class level by selecting the “Class-level comparison” tab (Fig. 4.3b). Overall, the “Scatterplots” module within the hTF array analyzer provides users with a quick and user-friendly method to generate publication-quality graphs that demonstrate pairwise COF recruitment comparisons across TF probe sets of interest.

4.3.5 Integration of probe-level data to generate TF site COF recruitment models

An advantage of using the hTF array platform over other technologies, such as genomic COF recruitment profiling techniques like ChIP-seq, is the ability to use SNV probes to define COF recruitment motifs that identify the recruiting TFs and quantify the impact of nucleotide variants on COF recruitment. The “Motif grid” portal integrates the COF recruitment results and allows the user to generate the COF recruitment models and visualize them as sequence logos (Fig. 4.3c). The sequence logos are arranged in a grid where rows represent the selected TF probe sets and columns correspond to the different COF recruitment experiments of interest in order to allow researchers to visually compare recruitment models (Fig. 4.3c). The user can either select to view the models as "Energy logos", which visualize change in recruitment intensity for each nucleotide variant or as a more traditional PWM logo that represent TF binding probabilities and enables a more direct comparison to the PWM models compiled in large public database such as JASPAR (Khan et al., 2018), CIS-BP (Weirauch et al., 2014), UNIPROBE (Newburger and Bulyk, 2009; Hume et al., 2015), HOCOMOCO (Kulakovskiy et al., 2013; Kulakovskiy et al., 2018), and MotifDb (Shannon and Richards, 2018). In addition to the COF recruitment logos obtained experimentally, the "Motif grid" module displays the reference JASPAR 2018 CORE binding model for each user-selected TF to enable comparisons between empirical COF recruitment models to their corresponding expected binding model.

The "Motif grid" module allows for sophisticated comparison of COF recruitment preferences at different TFBSs of interest. For example, a user might be interested in the coactivator p300 or the corepressor NCoR is recruited to diverse TFBSs such as the SPI1, RELA, and RARA::RXRA consensus sites in both unstimulated and LPS-stimulated macrophages. As shown in the dynamically generated motif grid for this analysis example, both the coactivator p300 and the NCoR corepressor complex can be recruited to the SPI1 site albeit with distinct nucleotide preferences (Fig. 4.3c, top row). The cytosine-rich preferences 5' relative to the GGAA core element observed in three of the experiments (Fig. 4.3c, columns 3-5) match a different subclass of ETS factor (e.g., ELF1, discussed in 4.3.7 below) than the motif in column 1 which more closely resembles the SPI1 logo (Fig. 4.3c). These results suggest that different ETS factors may be recruiting p300 under different conditions and may be recruiting p300 and NCoR under unstimuluated conditions (Fig. 4.3c). In contrast to the SPI1 consensus site, the RELA site supports only the recruitment of p300 and in an LPS-dependent manner (Fig. 4.3c, middle row). This is consistent with NF-κB activation and translocation into the nucleus during the pro-inflammatory response. The recruitment model obtained for the RELA probes by integrating results over all single nucleotide variant probes associated with the RELA consensus is concordant with the expected RELA TF binding model (Fig. 4.3c, middle row, rightmost column) (Siggers et al., 2012). Finally, though the RARA::RXRA complex does not appear to recruit p300 in either condition, it demonstrates moderate constitutive recruitment of the NCoR corepressor complex (Fig. 4.3c, bottom row). The empirical NCoR recruitment models for these sites are concordant across cell states (LPS-stimulated and untreated macrophages) and strongly resemble canonical nuclear receptor binding models (Fig. 4.3c, bottom row, rightmost column) (Penvose et al., 2019). This represents only a few of the numerous comparisons possible within a single set of hTF array experiments but provides a valuable example of the insight a researcher can gain about the TF-COF complexes and individual preferences present in a cell type or state of interest.

4.3.6 Generating cell state-level recruitment signatures using the hTF array analyzer

An advantage of having a standardized platform that profiles the same set of TF consensus sites in each experiment is the possibility of developing cell state-level TF-COF signatures or biomarkers using the collection of TF-COFs active in a given cell state. This would enable researchers to then investigate TF-COF complexes that are affected in aberrant cell states (cancer, autoinflammatory, etc.) that may play a role in mediating these cell states or whose interactions may be perturbed relative to a matched "healthy" cell state. To enable investigations of cell state-specific TF-COF signatures and visually summarize the results of a set of hTF array experiments at the complete array-level, we developed the interactive COF recruitment heatmap as a module of the array analyzer software (Fig. 4.4a). The plots display the intensity of a hybrid scoring mechanism that scales the COF recruitment strength at a given TF consensus site by the Pearson similarity obtained when comparing the empirical COF recruitment model to its corresponding expected TF binding model (see 4.5 Materials and Methods). The score thereby distinguishes strong consensus models from recruitment sites of similar strength that do not produce the expected model at a given TF site. Hovering the cursor over a given element within the full heatmap displays this score obtained by a given TF-COF complex in a given experiment (Fig. 4.4a, SREBF2 shown). Representing TF-COF complexes using this hybrid scoring metric also allows for similarity-based clustering of the TF sites (rows) in addition to the collection of experiments performed in a given array run (columns) (Fig. 4.4a). Users have the option to deselect experiments (to omit them from the recruitment heatmap) which automatically recomputes the TF-level and experiment-level similarities, performs the clustering steps again, and plots an updated COF recruitment heatmap.

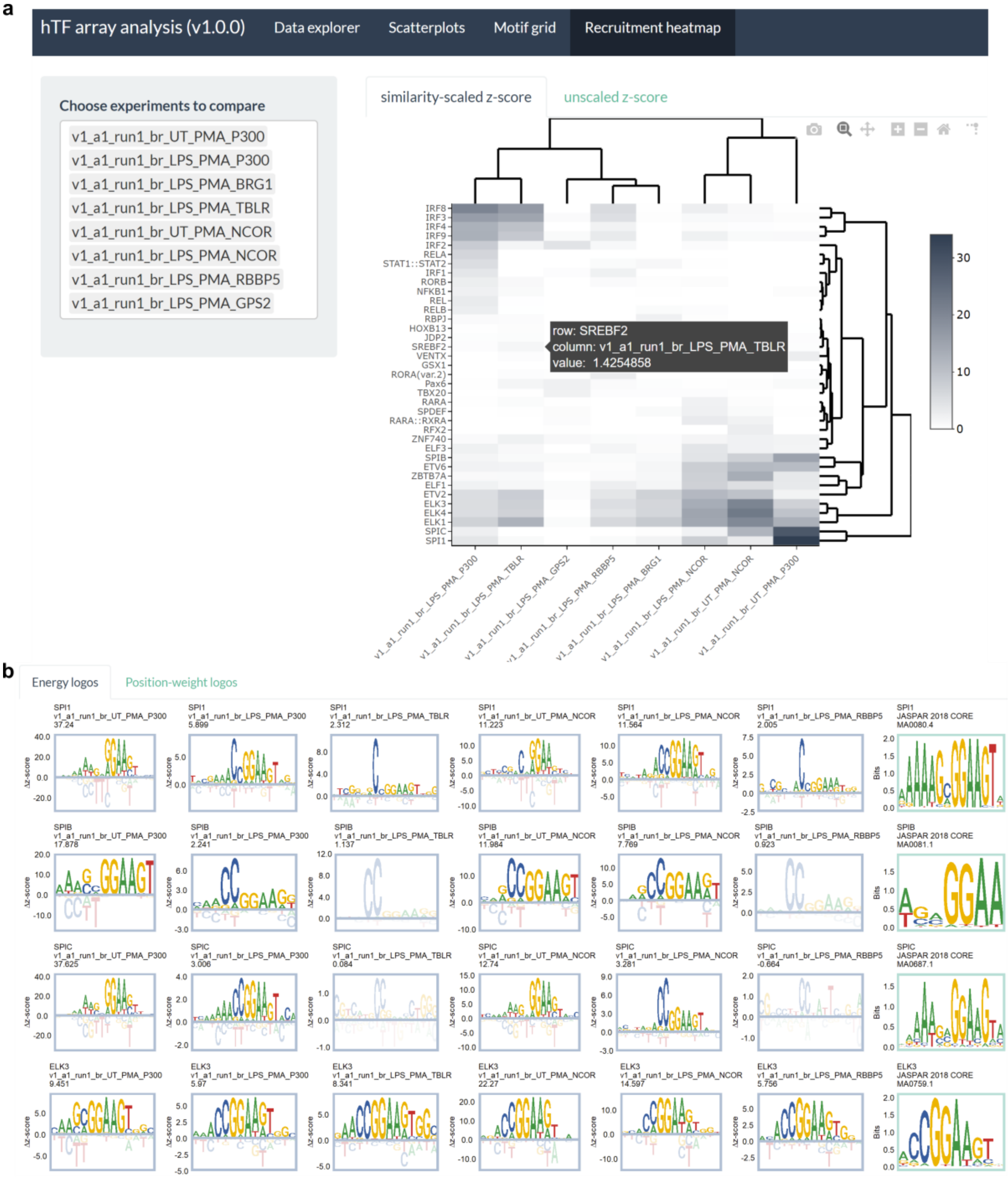


Figure 4.4: Using cell state-level TF-COF recruitment signatures to guide hTF analyses
(a) Interactive TF-COF recruitment heatmap demonstrating similarity-scaled (see 4.5 Materials and Methods) z-scores for consensus TF sites (rows) exhibiting a minimal hybrid score of 2

across the pilot experiments performed. COF recruitment experiments included in this figure are shown in the box to the left (all performed using macrophages – see 4.5 Materials and Methods). (b) Motif grid comparing the recruitment models obtained at 4 related ETS factor consensus sites (rows: SPI1/PU.1, SPIB, SPIC, ELK3). COF recruitment experiments (columns) are as follows: P300 (untreated), P300 (LPS-stimulated), BRG1 (LPS-stimulated), TBLR (LPS-stimulated), NCoR, NCoR (LPS-stimulated), RBBP5 (LPS-stimulated). The rightmost column displays the JASPAR 2018 CORE reference model for each TF. Contributions: DB developed the interactive heatmap module with input from RZ, RM, HH, JLK, and TS. Experiments were performed by RZ.

Overall, expressing the results of a given set of hTF array experiments using a summary heatmap and clustering TFs and experiments by similarity allows for immediate insight that researchers can inspect visually to determine TF-COF recruitment differences in their set of experiments. For example, it is apparent in the pilot experiments performed that there is a cluster of interferon regulatory factors (IRF8, IRF3, IRF4, IRF9) that recruit p300 in a predominantly LPS-inducible manner (Fig. 4.4a) that is stronger in intensity relative to other LPS-inducible factors such as RELA, NFKB1, REL, and RELB (Fig. 4.4a). The recruitment heatmap thereby represents a powerful method to investigate the TF-COF complexes active in a given cell state as well as the state's similarity or dissimilarity to other experiments of interest. In future experiments, this will enable researchers to rapidly define the TF-COF complexes active in a cell state and identify important differences between "healthy" and "disease" cell states to define TF-COF-level biomarkers and signatures.

4.3.7 Distinct COF recruitment logic mediated by closely related ETS factors

Given the scale of the TFBSs profiled on the hTF array (346 total), an important function of the "Recruitment heatmap" module is the ability to identify potentially interesting TFs and TF families that recruit a given COF and suggest hypotheses to further investigate. For example, the TFBSs demonstrating strong COF recruitment in experiments of interest can be further investigated using the “Motif grid” module to compare and contrast the individual COF recruitment models generated and provide insight into the TF-COF complexes active and the nucleotide preferences resulting in their recruitment. An interesting test case identified in the pilot hTF array experiments is the distinct COF recruitment logic at closely related ETS factor sites that is suggested by the recruitment heatmap (Fig. 4.4a).

The TF-level (rows) and experiment-level (columns) similarity-based clustering in the recruitment heatmap for the pilot series of hTF array experiments shows a clear difference between the COF recruitment differences between ETS factors (Fig 4.4a). For example, ETV2, ELK3, ELK4, and ELK1 all cluster together based on similarity and appear to support the recruitment of several COFs such as p300, TBLR, RBBP5, BRG1, and NCoR at moderate levels and consistent with their expected binding models (Wei et al., 2010). In comparison, the closely related ETS factors SPIB, SPIC, and SPI1/PU.1 do not appear to support the recruitment of TBLR, RBBP5 or BRG1 despite the known similarity in their binding preference to the other aforementioned ETS factors (Fig. 4.4a) (Wei et al., 2010). Furthermore, SPI1 and SPIC appear to demonstrate distinct NCoR recruitment preferences where SPI1 recruits NCoR preferentially in LPS-stimulated macrophages and SPIC recruits NCoR preferentially in the unstimulated control macrophages. To investigate these phenomena further at the level of these individual models, we compared representative ETS factors using the hTF array analyzer “Motif grid” module. Consistent with what is conveyed in the recruitment heatmap (Fig 4.4a), ELK3 recruits each of the COFs tested (with the exception of GPS2) at moderate-to-high levels with recruitment preferences consistent with its canonical binding model (Fig. 4.4b, bottom row). Similarly, the COF recruitment experiments for TBLR, BRG1, and RBBP5 do not pass the minimal motif plotting threshold (z-score of 1.5) for SPIB and SPIC as is shown in the experiment-level recruitment heatmap (Fig. 4.4b, middle rows). Though BRG1, TBLR, and RBBP5 are all recruited to the SPI1 consensus site, the preferred models for the recruitment of these COFs have a prominent C-rich preference 5' relative to the GGAA core ETS element (Fig. 4.4b, top row) which is a feature that is inconsistent with the canonical SPI1 binding model (Wei et al., 2010; Mohaghegh et al., 2019) and explains why the similarity-scaled z-score for these experiments is low (Fig. 4.4a). In this case, the site is likely being used by another ETS factor with a C-rich 5' preference flanking the core GGAA site, such as the ELK factors, as this preference is more consistent with the canonical TF binding models for this ETS sub-family (Fig. 4.4b, bottom row) (Wei et al., 2010). Similarly, comparing the NCoR recruitment preferences for the closely related SPI1 and SPIC produce a logo more consistent with the SPI1 binding model in untreated macrophages (Fig. 4.4b, rows 1 and 3, column 5) whereas the recruitment logo produced in LPS-stimulated macrophages is more consistent with the ELK factors (Fig. 4.4b, rows 1 and 3, column 6) which explains the discrepancies observed for SPIC and SPI1 NCoR recruitment observed in the experiment-level recruitment heatmap (Fig. 4.4a). Together, these observations demonstrate that the hTF array can be used to discern subtle (albeit mechanistically important) differences in COF recruitment preferences within a given TF family. Combining the experiment-level recruitment heatmap with the motif grid feature thereby enables sophistical investigations into TF-COF complex recruitment logic even for closely related TFs with similar binding preferences.

4.4 Discussion

In this work, we present the hTF array as an extension of our group's existing CoRec approach to survey TF-COF recruitment beyond a small panel of immune-centric TFBSs to an expansive panel of non-redundant TFs relevant to any cell type or context in humans. This represents a first attempt to expand our TF-COF recruitment profiling techniques through a standardized microarray design and an accompanying interactive analysis software suite that are both freely available and open-source. The hTF array design can be applied in its current iteration to investigate diverse research questions related to COF recruitment but can also be improved upon in the future. Though effort was made to algorithmically include a diverse and non-redundant panel of TFs, future experiments beyond the pilot data presented here will allow for further refinement and possibly point to redundant sites than can eliminated from the design and updated in future iterations. As the software determines the TFs included in the design at runtime, any future iterations, custom alterations, and improvements to the hTF array design will be compatible with the existing analysis software.

The hTF array analyzer was developed to allow analysis of COF recruitment datasets in a more interactive, user-friendly, and accessible to the experimentalists and researchers who perform the COF recruitment experiments. The interactive analysis software combines multiple modules that allow for the inspection of experimental results at several different scales. To inspect data at the probe-level, scatterplots that visualize the COF recruitment intensities at the consensus and SV probes for TFs of interest across pairs of experiments can be used. The data explorer and motif grid modules allow users to gain additional insight at the TF-level. And finally, the recruitment heatmaps provide an array-level summary of TF-COF recruitment across all TFs included in the hTF design and all experiments performed. Through examples using pilot hTF array data generated from unstimulated and LPS-stimulated macrophages, we demonstrated that these scales provide complementary analyses and insight. The data explorer can be used to suggest TFs to investigate further using probe-level scatterplots and the motif grid module as demonstrated with the example of investigating distinct p300 and NCoR recruitment models at diverse TF sites (SPI1, RELA, and RARA::RXR) with different binding and recruitment preferences. We further demonstrated utility of the analysis modules by inspecting the array-level recruitment heatmap that suggested a distinct recruitment logic between closely-related ETS factors. We confirmed the similar but distinct recruitment models at the motif-level using the interactive motif grid. We anticipate that these types of interactive complementary analyses will empower researchers to gain insight on the TF-COF complexes active in their samples and provide actionable hypotheses for further analysis and experimentation.

In addition to the open-source hTF array design and analysis software, we have also introduced the concept of cell state-level TF-COF recruitment signatures which we visualized using COF recruitment heatmaps across the panel of TFs. Profiling the recruitment of COFs to the same diverse panel of TFs will eventually enable investigations into differential COF recruitment between cell types/states as well as the development of COF recruitment “biomarkers” that can be used to investigate the TF-COF complexes involved in mediating or maintaining aberrant cell states compared to healthy controls. In this pilot study, we have demonstrated that our profiling approach captures TF-COF binding events that implicitly account for cell state-specific phenomena such as PTMs and differing protein levels in the nucleus. For example, the NF-κB complex translocates into the nucleus in macrophages following LPS stimulation (Smale, 2012; Medzhitov and Horng, 2009) and this presence (or absence in the case of control unstimulated macrophages) is reflected in the LPS-specific recruitment of p300 to the RELA consensus site (Fig. 4.3c, middle row). Furthermore, LPS-dependent recruitment of p300 is detected at IRF3, a TF that requires phosphorylation in order to dimerize *in vivo* (Andrilenas et al., 2018; Tamura et al., 2008; Smale, 2012; Medzhitov and Horng, 2009), thereby demonstrating that the impact of PTMs are implicitly measured (Fig. 4.3a). Additional experiments will be required to determine the extent to which changes in PTMs or protein levels across cell states are captured in the individual COF recruitment motifs (beyond simple presence/absence) and whether the profiling method is sensitive enough to detect small changes. Our previous investigations into TF binding from nuclear extracts indicate that PTM and protein level changes can be reflected in the binding motif obtained (Mohaghegh et al., 2019). Differences were especially evident at sites bound by complexes formed by more than one TF where the binding partner is only moderately expressed - such as with IRF8 at the cooperative PU.1-IRF8 composite element (Mohaghegh et al., 2019). Whether this sensitivity of the motifs to PTMs and protein levels generalizes to COF recruitment motifs across the surveyed TFBSs in the hTF array design remains an open question.

Furthermore, though the TFBSs included in the hTF array design were selected from human experiments, the JASPAR2018 CORE database, from which the models were selected, contains non-redundant consensus TF models across many vertebrate species including mouse. Of the 119 non-redundant mouse-specific models in the database, 38 (32%) have a consensus site that is equivalent to one represented by the TFBSs already included on the pilot hTF array design (Supplementary Table 4.1). As the motif database used to compute PWM similarities of the COF recruitment motifs to reference TF binding models is read at runtime, in order to investigate the application of hTF to study COF recruitment in other vertebrate models, an expanded database can be used without having to alter the hTF array design or analysis software. The applicability of the hTF array to study TF-COF binding in other vertebrate cell sources (such as mice) will be the topic of future investigations.

In addition to allowing experimentalists to define TF-COF recruitment signatures in cell states of interest, the hTF array platform will provide a means to screen compounds to “reverse” these aberrant signatures as many of these enzymatic COFs can be targeted using existing compounds (Lasko et al., 2017; Fedorov et al., 2015; Yoon and Eom, 2016). These TF-COF signatures/biomarkers will provide an important and understudied annotation layer that can be further integrated with other technologies/modalities such as RNA-seq and ChIP-seq to enable sophisticated integrative analyses into the molecular mechanisms underlying disease cell states. Future studies using the hTF array will investigate this translational potential of the platform as well as the concordance with paired COF ChIP-seq data for COFs/conditions of interest in order to compare and contrast TF-COF recruitment potential (using hTF array) with genome-wide incidence (using ChIP-seq). Overall, we anticipate the standardized hTF array design and the accompanying interactive analysis software will be applied to study diverse research questions in basic research and translational applications alike.

4.5 Materials and Methods

4.5.1 Cell culture

THP-1 human monocyte cells (ATCC TIB-202) were cultured in RPMI-1640 (Thermo #72400120) with 1 mM sodium pyruvate (Thermo #16140071) and 10% heat-inactivated FBS (Thermo #11360070) in a 37°C incubator with 5% CO_2. To prepare nuclear lysates, three 50 ml suspension cultures maintained in T-175 flasks were used for each stimulation condition. Cells were differentiated at a cell density of 8 x 10^5 cells/ml into adherent macrophages using 25 ng/ml PMA and incubated for 96 hours. After 96 hours, cells were washed with 1X PBS, fresh growth media was applied, and cells rested for 48 hours before stimulation. Differentiated THP-1 cells were stimulated with 1 ug/ml LPS for 45 min before harvesting. This section was included and modified from its original source with permission from the author (Zhao, 2020).

4.5.2 Nuclear extract preparation

To collect cells after stimulation treatments, THP-1 adherent macrophages were washed in PBS and placed on ice. Cells were dislodged from the flask using a cell scraper in cold PBS supplemented with 0.1 mM protease inhibitor cocktail (Sigma-Aldrich #P8340). Cells from three T-175 cultures were collected in 50 ml tubes and pelleted at 500xg for 5 min at 4°C. Once cell pellets were obtained, to rupture the cell membrane, the cell pellet was resuspended in 2 ml of a hypotonic Buffer A (10 mM HEPES (pH 7.9), 1.5 mM MgCl2, 10 mM KCl, 0.1 mM protease inhibitor cocktail, 0.1 mM phosphatase inhibitor cocktail (Sigma-Aldrich #4315), 0.5 mM DTT) and incubated for 10 min on ice. 20 ul of 10% IGEPAL (Sigma-Aldrich I8896) was added, and the cell suspension was vortexed for 10 s. Released nuclei were observed under a hemocytometer. Nuclei were pelleted at 500xg for 5 min at 4°C. The nuclear pellet was then resuspended in 100 ul hypertonic Buffer C (20 mM HEPES (pH 7.9), 25% glycerol, 1.5 mM MgCl2, 0.2 mM EDTA, 420 mM NaCl, 0.1 mM protease inhibitor cocktail, 0.1 mM phosphatase inhibitor cocktail, 0.5 mM DTT). The nuclei suspension was vortexed for 30 s, followed by nutation for 1 hour at 4°C on a Hula mixer. The insoluble nuclear components were pelleted at 21,000xg for 20 min at 4°C. The supernatant containing soluble nuclear proteins was collected, flash-frozen using liquid nitrogen, and stored at -80°C. Protein concentration of nuclear lysates was quantified by A280 measurement. This section was included and modified from its original source with permission from the author (Zhao, 2020).

4.5.3 nextPBM experimental methods

PBM experiments were performed on a custom designed single-stranded DNA microarray (Agilent Technologies, Design ID 082690, 4 x 180k format). DNA microarray double stranding and nextPBM protocols were performed as previously described (Berger et al., 2006; Berger and Bulyk, 2009; Mohaghegh et al. 2019). PBM wash steps were performed in coplin jars on an orbital shaker at 125 rpm, and all PBM steps were performed at room temperature. Briefly, double-stranded DNA microarrays were first washed in 0.01% Triton X-100 in HBS (HEPES-buffered saline, pH 7.4) for 5 min, followed by blocking with 2% NFDM in HBS for 1 hour. Next, arrays were rinsed in HBS and incubated with nuclear protein lysate in a binding buffer (0.3% NFDM, 20 mM HEPES, 100 mM NaCl, 1 mM DTT, 0.2 mg/ml BSA, 0.02% Triton X-100, and 0.4 mg/ml salmon testes DNA (Sigma D7656)) for 1 hour in the dark. After protein binding, arrays were incubated with 20 ug/ml of primary antibody in 2% milk in HBS for 20 min, followed by an HBS rinse and 20 ug/ml secondary antibody incubation for 20 min. Antibodies used included anti-p300 (Abcam #ab149848), anti-BRG1 (Santa Cruz #sc11796), anti-NCoR (Bethyl Laboratories #A301-145A), anti-TBL1XR1 (Santa Cruz #sc100908), anti-RBBP5 (Bethyl Laboratories #A300-109A), anti-GPS2 (Abclonal #A3901), Alexa488 anti-mouse (Invitrogen #A11131), and Alexa647 anti-rabbit (Invitrogen #A21245). Finally, arrays were washed twice in 0.05% Tween-20 for 3 min and once in HBS for 3 min before scanning. Arrays were scanned using a GenePix 4400A scanner, and fluorescence was quantified using GenePix Pro 7.2. Fluorescence data was exported and normalized using MicroArray LINEar Regression (Berger et al., 2006). This section was included and modified from its original source with permission from the author (Zhao, 2020).

4.5.4 hTF array design

Non-redundant TF binding models from the JASPAR 2018 core vertebrate set were obtained using the JASPAR2018 R bioconductor package. The total 1,564 models across model organisms were filtered to those obtained using a human source (human cell lines/tissues used for the characterization) resulting in 452 models from different TFs. The resulting motifs were then collapsed into consensus sequences using the top-scoring base preference at each position and filtered for equivalence based on nucleotide identity as well as size of the consensus sequence using a relative size filter of 0.9. Similar size was considered along with equivalent consensus sequences in order to avoid the scenario where a half-site within a composite site would be eliminated from the final design. Filtering by similar size and nucleotide identity resulted in 346 TF models to be included in the final design. To account for possible additional nucleotide determinants beyond the positions covered by the TF consensus binding sites, a random non-repeating 2 base pad was added to both ends of each consensus sequence. For each of these 346 modified consensus sequences, DNA probes corresponding to each possible SNV across these sequences were also generated. To account for size differences between probes, a 34-base backbone sequence was generated algorithmically such that the nucleotide at each position was generated randomly with the constraint that sequential positions contain non-repeating nucleotides. Each consensus and SNV sequence generated was then inserted into the backbone sequence beginning at the 5’ end such that the site being profiled is located at the end furthest away from the glass slide that the probe is fixed to.

These full 34-base targets were then embedded within a larger probe (total: 60 bases) as follows:

GC cap + 34 base target region (TF site or SNV within backbone) + 24 base primer

261 background target DNA probes were also included in the design in order to be able to estimate background fluorescence intensities in the experiments. These regions were selected as 34-base genomic segments from the human genome (hg38). The final design as well as the script used to design the array are open source and have been made freely available on Github (https://github.com/Siggers-Lab/hTF_array).

4.5.5 hTF array analyzer interactive software

The hTF array analyzer software is written using the shiny interactive web programming framework in R (RStudio Inc., 2013) with the visual theme “flatly” from the shinythemes R package (Chang, 2018). As with the final microarray design and the script used to generate the design, the hTF array analyzer is open source and has been made available on Github (https://github.com/Siggers-Lab/hTF_array). Normalized fluorescence datasets obtained from hTF array experiments are first log-transformed and z-scores are computed against the distribution of 261 background probes as previously described (Mohaghegh et al., 2019; Penvose et al., 2019; Keenan et al., 2020; Bray et al., 2020). The “Data explorer” module uses the javaScript DT package (Xie et al., 2020) to display interactive data tables of the COF fluorescence z-scores obtained at each of the 346 TF consensus sites. The “Scatterplot” module compares user-selected COF recruitment experiments at user-selected TF sites to compare z-scores obtained across experiments in a pairwise manner. The scatterplots are plotted using the ggplot2 package

(Wickham, 2016) with automatically generated colors for probes belonging to each TF group. The “family-level” and “class-level” tab regroups the user-selected probe groups based on their family or class annotation respectively and recolors the probe sets accordingly. The “Motif grid” module displays COF recruitment logos obtained at user-selected TFs and user-selected experiments plotted using the ggseqlogo R package (Wagih, 2017) and arranged as grids using the gridExtra package (Auguie, 2017). The COF recruitment “energy” logos summarize the z-scores obtained at a given TF consensus binding site (within a given experiment) as well as all of the probes corresponding to single nucleotide variants along the profiling region as previously described (Andrilenas et al., 2018; Mohaghegh et al., 2019; Penvose et al., 2019). In the “Recruitment heatmap” module, the “energy” motifs obtained at each TF (and in each experiment) within an array are first transformed into probability-based position-weight matrices using the Boltzmann energy distribution as previously described (Andrilenas et al., 2018; Mohaghegh et al., 2019; Penvose et al., 2019). To obtain the hybrid scores displayed in the “Recruitment heatmap” module, the motif database used in the initial generation of the array design (a filtered version of the JASPAR 2018 CORE) is first loaded using the universalmotif package (Tremblay, 2019). The z-score obtained experimentally for COF recruitment at each TF consensus site (and each experiment) is then scaled using the corresponding similarity of the COF recruitment PWM to its expected TF binding model (computed as a length-normalized Pearson correlation coefficient across binding site positions) to generate each hybrid score using the TFBSTools R package (Tan et al., 2016). The heatmaply package is used to generate the interactive heatmap plots (Galili et al., 2017). Default distance metric (Euclidean) and clustering function (complete) are used in the plots generated.

4.6 Supplementary Information

ID	TF name	Consensus sequence	Species	Equivalent hTF model
MA0004.1	Arnt	CACGTG	Mus musculus	BHLHE40,BHLHE41,CLOCK,HES5,HES7,HEY1,MAX,MAX::MYC,MITF,MLX,MNT,MXI1,MYC,MYCN,TFE3,USF1,USF2
MA0006.1	Ahr::Arnt	TGCGTG	Mus musculus	EGR1,EGR2,EGR3,EGR4,PAX1,Pax6
MA0029.1	Mecom	AAGATAAGATAACA	Mus musculus	
MA0063.1	Nkx2-5	TTAATTG	Mus musculus	BSX,ESX1,RAX
MA0067.1	Pax2	AGTCACGC	Mus musculus	
MA0078.1	Sox17	CTCATTGTC	Mus musculus	
MA0087.1	Sox5	ATTGTTA	Mus musculus	
MA0092.1	Hand1::Tcf3	GGTCTGGCAT	Mus musculus	
MA0111.1	Spz1	AGGGTAACAGC	Mus musculus	
MA0125.1	Nobox	TAATTGGT	Mus musculus	ESX1,GBX2
MA0135.1	Lhx3	AAATTAATTAATC	Mus musculus	
MA0142.1	Pou5f1::Sox2	CTTTGTTATGCAAAT	Mus musculus	
MA0062.2	Gabpa	CCGGAAGTGGC	Mus musculus	ZBTB7A

MA0002.2	RUNX1	GTCTGTGGTTT	Mus musculus	
MA0047.2	Foxa2	TGTTTACTTAGG	Mus musculus	
MA0065.2	Pparg::Rxra	GTAGGGCAAAGGTCA	Mus musculus	
MA0151.1	Arid3a	ATTAAA	Mus musculus	ALX3,DUX4,DUXA,GSX1,GSX2,MNX1,PHOX2A
MA0152.1	NFATC2	TTTTCCA	Mus musculus,Rattus norvegicus,Homo sapiens	NFATC2,NFATC3
MA0158.1	HOXA5	CACTAATT	Mus musculus,Homo sapiens	HOXA5
MA0160.1	NR4A2	AAGGTCAC	Mus musculus,Rattus norvegicus,Homo sapiens	ESR1,MITF,NR4A1,NR4A2
MA0164.1	Nr2e3	CAAGCTT	Mus musculus	
MA0259.1	ARNT::HIF1A	GGACGTGC	Mus musculus,Rattus rattus,Homo sapiens,Oryctolagus cuniculus	ARNT::HIF1A
MA0146.2	Zfx	GGGGCCGAGGCCTG	Mus musculus	
MA0463.1	Bcl6	TTTCCTAGAAAGCA	Mus musculus	

MA0467.1	Crx	AAGAGGATTAG	Mus musculus	
MA0480.1	Foxo1	TCCTGTTTACA	Mus musculus	
MA0482.1	Gata4	TCTTATCTCCC	Mus musculus	
MA0483.1	Gfi1b	AAATCACAGCA	Mus musculus	
MA0485.1	Hoxc9	GGCCATAAATCAC	Mus musculus	
MA0493.1	Klf1	GGCCACACCCA	Mus musculus	KLF9
MA0494.1	Nr1h3::Rxra	TGACCTAAAGTAACCTCTG	Mus musculus	
MA0499.1	Myod1	TGCAGCTGTCCCT	Mus musculus	
MA0500.1	Myog	GACAGCTGCAG	Mus musculus	
MA0503.1	Nkx2-5(var.2)	AGCCACTCAAG	Mus musculus	
MA0505.1	Nr5a2	AAGTTCAAGGTCAGC	Mus musculus	
MA0509.1	Rfx1	GTTGCCATGGCAAC	Mus musculus	RFX2
MA0514.1	Sox3	CCTTTGTTTT	Mus musculus	
MA0515.1	Sox6	CCATTGTTTT	Mus musculus	
MA0518.1	Stat4	TTTCCAGGAAATGG	Mus musculus	
MA0519.1	Stat5a::Stat5b	ATTTCCAAGAA	Mus musculus	
MA0520.1	Stat6	CATTTCCTGAGAAAT	Mus musculus	
MA0521.1	Tcf12	AACAGCTGCAG	Mus musculus	
MA0035.3	Gata1	TTCTTATCTGT	Mus musculus	

MA0150.2	Nfe2l2	CAGCATGACTCAGCA	Mus musculus	
MA0143.3	Sox2	CCTTTGTT	Mus musculus	
MA0591.1	Bach1::Mafk	AGGATGACTCAGCAC	Mus musculus	
MA0594.1	Hoxa9	GCCATAAATCA	Mus musculus	
MA0601.1	Arid3b	ATATTAATTAA	Mus musculus	
MA0602.1	Arid5a	CTAATATTGCTAAA	Mus musculus	
MA0603.1	Arntl	GGTCACGTGC	Mus musculus	
MA0604.1	Atf1	ATGACGTA	Mus musculus	FOSL1::JUND(var.2)
MA0605.1	Atf3	GATGACGT	Mus musculus	ATF7,BATF3,CREB3,CREB3L1,FOS::JUN(var.2),FOSL2::JUNB(var.2),FOSL2::JUND(var.2),XBP1
MA0607.1	Bhlha15	CCATATGT	Mus musculus	BHLHE22
MA0608.1	Creb3l2	GCCACGTGT	Mus musculus	
MA0609.1	Crem	TATGACGTAA	Mus musculus	
MA0611.1	Dux	CCAATCAA	Mus musculus	
MA0614.1	Foxj2	GTAAACAA	Mus musculus	FOXC2,FOXF2,FOXG1,FOXK1,FOXK2,SRY

MA0615.1	Gmeb1	GAGTGTACGTAAGATGG	Mus musculus	
MA1099.1	Hes1	GGCACGCGTC	Mus musculus	
MA0616.1	Hes2	TAACGACACGTGC	Mus musculus	
MA0617.1	Id2	GCACGTGA	Mus musculus	
MA0622.1	Mlxip	GCACGTGT	Mus musculus	HEY1
MA0623.1	Neurog1	ACCATATGGT	Mus musculus	OLIG2
MA0626.1	Npas2	GGCACGTGTC	Mus musculus	HEY1
MA0627.1	Pou2f3	TTGTATGCAAATTAGA	Mus musculus	
MA0629.1	Rhox11	AAGACGCTGTAAAGCGA	Mus musculus	
MA0631.1	Six3	GATAGGGTATCACTAAT	Mus musculus	
MA0632.1	Tcfl5	GGCACGTGCC	Mus musculus	HES5,HES7
MA0633.1	Twist2	ACCATATGTT	Mus musculus	BHLHE22
MA0007.3	Ar	GGGAACACGGTGTACCC	Mus musculus	
MA0643.1	Esrrg	TCAAGGTCAT	Mus musculus	ESRRB
MA0676.1	Nr2e1	AAAAGTCAA	Mus musculus	
MA0677.1	Nr2f6	GAGGTCAAAGGTCA	Mus musculus	
MA0681.1	Phox2b	TAATTTAATTA	Mus musculus	PHOX2A
MA0682.1	Pitx1	TTAATCCC	Mus musculus	PITX3

MA0075.2	Prrx2	CCAATTAA	Mus musculus	BSX,ESX1,RAX
MA0512.2	Rxra	GGGGTCA AAGGTCA	Mus musculus	RXRB
MA0704.1	Lhx4	TTAATTAA	Mus musculus	LMX1A,POU6F1
MA0705.1	Lhx8	CTAATTAG	Mus musculus	EMX2,EN1,GBX1
MA0709.1	Msx3	CCAATTAA	Mus musculus	BSX,ESX1,RAX
MA0124.2	Nkx3-1	ACCACTTAA	Mus musculus	NKX3-2
MA0720.1	Shox2	CTAATTAA	Mus musculus	ALX3,GSX1,GSX2,ISX,LHX2,MEOX1,MIXL1
MA0592.2	Esrra	TTCAAGGTCAT	Mus musculus	
MA0114.3	Hnf4a	GGGGTCA AAGTCCAAT	Mus musculus	
MA0728.1	Nr2f6(var.2)	GAGGTCA AAAGGTCA	Mus musculus	RARA
MA0739.1	Hic1	ATGCCAACC	Mus musculus	
MA0742.1	Klf12	GACCACGCCCTTATT	Mus musculus	
MA0769.1	Tcf7	AAAGATC AAAGG	Mus musculus	LEF1,TCF7L1,TCF7L2
MA0816.1	Ascl2	AGCAGCTGCT	Mus musculus	
MA0461.2	Atoh1	AACATATGTT	Mus musculus	BHLHE23,OLIG1
MA0829.1	Srebf1(var.2)	ATCACGTGAC	Mus musculus	BHLHE40
MA0832.1	Tcf21	GCAACAG CTGTTGT	Mus musculus	
MA0840.1	Creb5	AATGACGTCACC	Mus musculus	

MA0117.2	Mafb	AAAATGCTGACT	Mus musculus	
MA0851.1	Foxj3	AAAAAGTAAACAAACAC	Mus musculus	
MA0853.1	Alx4	CGCATTAATTAATTACC	Mus musculus	
MA0854.1	Alx1	CGAATTAATTAATCACC	Mus musculus	
MA0857.1	Rarb	AAAGGTCAAAAGGTCA	Mus musculus	
MA0858.1	Rarb(var.2)	AGGTCAACTAAAGGTCA	Mus musculus	
MA0859.1	Rarg	AAGGTCAAAAGGTCAA	Mus musculus	
MA0860.1	Rarg(var.2)	AAGGTCACGAAAGGTCA	Mus musculus	
MA0869.1	Sox11	AACAATTTCAGTGTT	Mus musculus	
MA0870.1	Sox1	AACAATAACATTGTT	Mus musculus	
MA0874.1	Arx	GTCCATTAATTAATGGA	Mus musculus	
MA0877.1	Barhl1	GCTAATTGCT	Mus musculus	
MA0879.1	Dlx1	CCTAATTATC	Mus musculus	
MA0880.1	Dlx3	CCAATTAC	Mus musculus	DLX6
MA0881.1	Dlx4	CCAATTAC	Mus musculus	DLX6

MA0883.1	Dmbx1	TGAACCGGATTAATGAA	Mus musculus	
MA0885.1	Dlx2	GCAATTAA	Mus musculus	
MA0896.1	Hmx1	ACAAGCAATTAATGAAT	Mus musculus	
MA0897.1	Hmx2	ACAAGCAATTAAAGAAT	Mus musculus	
MA0898.1	Hmx3	ACAAGCAATTAAAGAAT	Mus musculus	
MA0904.1	Hoxb5	ACGGTAATTAGCTCAT	Mus musculus	
MA0910.1	Hoxd8	TAAATAATTAATGGCTA	Mus musculus	
MA0911.1	Hoxa11	GGTCGTAAAATT	Mus musculus	
MA0912.1	Hoxd3	TTGAGTTAATTAACCT	Mus musculus	
MA0913.1	Hoxd9	GCAATAAAAA	Mus musculus	
MA1153.1	Smad4	TGTCTAGA	Mus musculus	SMAD3

Supplementary Table 4.1: Mouse-specific TF models from JASPAR2018 included in hTF array design

The table lists mouse-specific TF model accessions from the JASPAR2018 database (119 total) and indicates equivalent hTF array models (if applicable). TF-COF binding for mouse TF models that have an equivalent hTF model can therefore be profiled from mouse cells using hTF array probe sets

CHAPTER FIVE: Discussion and future work

5.1 New methodologies for analyzing TF and TF-COF binding

Motivated by the widening gap between the identification of disease-associated non-coding variants and our ability to mechanistically characterize them (Gallagher and Chen-Plotkin, 2018), the collection of work presented here represents a novel toolkit for researchers to begin to study the link between DNA variants and the aberrant recruitment of COFs. In Chapter 2, I presented the nextPBM as an improvement over the traditional PBM protocol using integrative genomics approaches to select the sequences profiled in the pilot microarray design. In addition to the improvements observed using nuclear extract in place of IVT proteins for profiling experiments, we investigated NE modifications such as enzymatic treatment using a phosphatase as well as immune-depletion of a cooperative factor providing researchers with a versatile toolkit to study protein-DNA complexes of interest through manipulation of the wild-type NE. In addition, the integrative genomics techniques used to select the sites included in the pilot nextPBM array to characterize the cooperative binding between PU.1 and IRF8 in monocytes provides researchers with a general framework through which they can screen suspected cooperative elements of interest and characterize the nucleotide determinants of TF-TF cooperativity to those sites. We expect that users will continue to expand the nextPBM toolkit in new and interesting ways in order to tackle their research questions as we have done in our own group with the CASCADE and hTF array approaches.

As an extension of the nextPBM platform, in Chapter 3, I presented CASCADE as a technique to profile the indirect recruitment of COFs to DNA via TFs in a stimulus-dependent manner. In our investigations using the technique to characterize COF recruitment to CREs and ncSNPs, we demonstrated that COF recruitment to these elements can be characterized at nucleotide resolution directly from cell NEs. Using an SV probe technique, we simultaneously uncovered the nucleotide determinants of COF recruitment to these locations as well as the regulators underlying their recruitment, thus providing researchers with a blueprint for future mechanistic characterizations of CREs and non-coding polymorphisms using this site-specific COF recruitment motif approach. In addition, using a combined nextPBM-based differential COF recruitment screen and subsequent CASCADE-based follow-up we also developed an HT approach to mechanistically annotate ncSNPs and link these polymorphisms to the complexes that mediate the effector functions of gene regulation. This combined 2-step screening and follow-up approach has the potential to be widely adopted for research involving DNA variants in the non-coding genome. Overall, based on our own demonstrations in Chapter 3, we envision CASCADE, and other nucleotide resolution techniques like it, will be instrumental in efforts to mechanistically annotate the effects of NCVs to provide more rapid and facilitated functional characterizations of the backlog of disease-associated variants uncovered by GWAS.

With the design of the human TF array in Chapter 4, we hoped to further standardize our COF recruitment approaches so that a breadth of research questions related to TF-COF complexes and DNA variants could be addressed using a single expansive array design. Toward this goal of expanding our approach, we designed the array using open-source TF binding models and released the full design as well as the software used to algorithmically generate it. In addition, we designed an analysis software to allow users to interact with the results of their hTF array experiments. We also present the concept of a COF recruitment "signature" and provide users with the means to build these signatures within our software. Ultimately, we hope that other researchers will begin to explore their systems using our standardized COF recruitment platform and we expect the concept of a COF recruitment signature to be important in future investigations to define TF-COF interactions within normal cell states and aberrant cell states alike.

With our future plans to further integrate our COF recruitment profiling approaches with existing orthogonal techniques and our planned improvements to the platforms presented in this work (discussed at length later in this chapter), we hope to further demonstrate the utility and versatility of these COF-centered approaches and provide the research community with a toolkit to facilitate investigation into the role of TF-COF complexes and NCVs in gene regulation and disease.

5.2 TF-COF complexes – moving beyond TF binding

High-throughput methods for studying TF-DNA binding (e.g., MITOMI, SMiLE-seq, CSI, PBM, SELEX-seq, etc.) (Berger et al., 2006; Berger and Bulyk, 2009; Maerkl and Quake, 2007; Isakova et al., 2017; Puckett et al., 2007; Warren, 2005; Jolma et al., 2010; Slattery et al., 2011) have had a tremendous impact on our understanding of TF function and genome-scale analysis of gene regulation, leading to large databases of widely-used TF binding models (Fornes et al., 2020; Weirauch et al., 2014; Wingender, 2008; Kulakovskiy et al., 2018). In addition, more recent approaches such as ATI (Wei et al., 2018) have been used to study TF-DNA binding in a more native cellular context by using TFs directly from cell nuclear extracts instead of using purified TF samples. However, COF recruitment and the assembly of TF-COF complexes in a cell type- and state-specific manner have not been examined using these approaches. Together, this demonstrates a historical focus on TF binding where the nucleotide determinants of COF recruitment and the assembly of TF-COF complexes in the cellular context has remained largely unexplored due in part to technological limitations.

In this work, we have demonstrated with CASCADE (in Chapter 3) and the hTF array (in Chapter 4) that COF recruitment, and by extension the assembly of TF-COF complexes, can be directly profiled in an HT manner using cell NEs representing an important technological and conceptual advance. Analogous to TF binding motifs, we have also introduced the concept of a COF recruitment motif that represents the DNA sequence-specificity of COF recruitment. The concept of a recruitment motif can be used to describe COF recruitment to TF sites in a local region (as with the TFBSs included on the hTF array or the genomic NCVs investigated with CASCADE) as well as larger CREs (for example, the CXCL10 promoter segment characterized with CASCADE). We have demonstrated as well that COF motifs can be used to infer the identity of the TF (or TF family) recruiting a COF to a particular DNA sequence. We note that this approach is different from COF ChIP-seq, which identifies the genomic loci to which a COF is recruited but does not identify the TFs involved at individual loci nor the DNA-sequence dependence of COF recruitment at single-nucleotide resolution. Given that COFs can be recruited by multiple TFs, by assaying recruitment of a single COF we also demonstrated the ability to profile numerous TF-COF complexes in parallel – an approach we expanded to 346 TFBSs with the hTF array.

The COF recruitment profiling techniques developed for this work, such as CASCADE and the hTF array, thereby offer conceptually new high-throughput approaches to study gene regulatory complexes that move beyond traditional TF binding investigations. We anticipate that the ability to assay COF recruitment afforded by CASCADE and the hTF array will provide a deeper general understanding of how DNA sequence and regulatory complexes control gene expression and will enable researchers to investigate new types of questions about the relationship between DNA variants, regulatory complexes, chromatin/histone modification, and gene expression.

5.3 Mapping regulatory inputs to CREs

One of the major applications of CASCADE, detailed extensively in Chapter 3 of this work, is profiling COF recruitment to large CREs to comprehensively examine the nucleotide determinants of these recruitment events in a cell type- and stimulus-dependent manner. To create more comprehensive models of the role of TF-COF complexes in gene expression, CASCADE will be integrated with modalities such as mass spec and MPRA (discussed below). In addition, to increase the throughput of CASCADE to profile more CREs, we discuss the CASCADE XL concept as a 2-step alternative to the CASCADE CRE characterization technique presented in Chapter 3.

The requirement for COF antibodies may preclude the use of CASCADE in more discovery-based applications where prior knowledge of the COFs mediating gene-regulatory function at a CRE of interest is not known. To compare our COF profiling techniques with an antibody-free proteomics approach, mass spec will be used to identify the key components of the complexes assembling at CREs of interest. Integration of CASCADE with mass spec will allow us to determine the extent to which full gene regulatory complexes are assembled on our arrays and provide the additional benefit of suggesting COFs that may be of interest to profile. Comparison of CASCADE with mass spec will thereby allow us to build more complete models of gene regulatory complex binding to CREs.

An open question in gene regulation is how transcriptional output is mediated by both TF affinity-dependent and independent mechanisms (Grossman et al., 2017; Andrilenas et al., 2018; Penvose et al., 2019; Kribelbauer et al., 2019; Louphrasitthiphol et al., 2020). To investigate the concordance (and discordance) between CASCADE-based COF recruitment results and changes in gene expression, CASCADE will be compared with MPRA reporter output. Since the input sequence library for MPRAs can be customized, probe sets including all single variants within a CRE can be screened for COF recruitment (using CASCADE) and gene expression changes (using MPRA) in matched cell types and stimulation conditions. Furthermore, cell state-specific integration of COF recruitment with reporter activity will further clarify an emerging role for PTMs in regulating TF activity given that a recent study has demonstrated that state-specific p300-mediated acetylation of a TF can both reduce its DNA-binding affinity while increasing transcriptional output (Louphrasitthiphol et al., 2020). These investigations relating COF recruitment with reporter activity will provide detailed gene regulatory models for CREs that link DNA variants to perturbations in TF-COF complex binding and gene expression changes. These gene regulatory models will also help clarify whether COF recruitment is mediated by single sites within a CRE or cooperatively distributed across several closely interspersed TF sites (Giorgetti et al., 2010; Kribelbauer et al., 2019; Louphrasitthiphol et al., 2020). We anticipate that integrating our CASCADE results with MPRA results will further position CASCADE as a useful assay to map the functional inputs of CREs.

A limitation of the CASCADE approach to characterize CREs is the number of DNA probes needed to employ the combined CRE tiling and SV probe approach to generate full CRE-wide recruitment motifs. To address this limitation, we have designed an alternate approach to screen a CRE (or group of CREs) of interest for COF recruitment prior to employing the full CASCADE approach. This concept, called CASCADE XL (Fig. 5.1), effectively leverages the success of the 2-step screening approach previously used in Chapter 3 to screen ncSNPs for differential COF recruitment. As with the differential COF ncSNP screen, a panel of COFs can be used to determine which overlapping tile probes within a CRE support the recruitment of these COFs. Only the tiles exhibiting significant COF recruitment are then included in a follow-up array where the full CASCADE SV probe approach is used to uncover the nucleotide determinants of COF recruitment and enable an inference as to the specific TF or TF family underlying the COF recruitment event (Fig. 5.1, part 2).

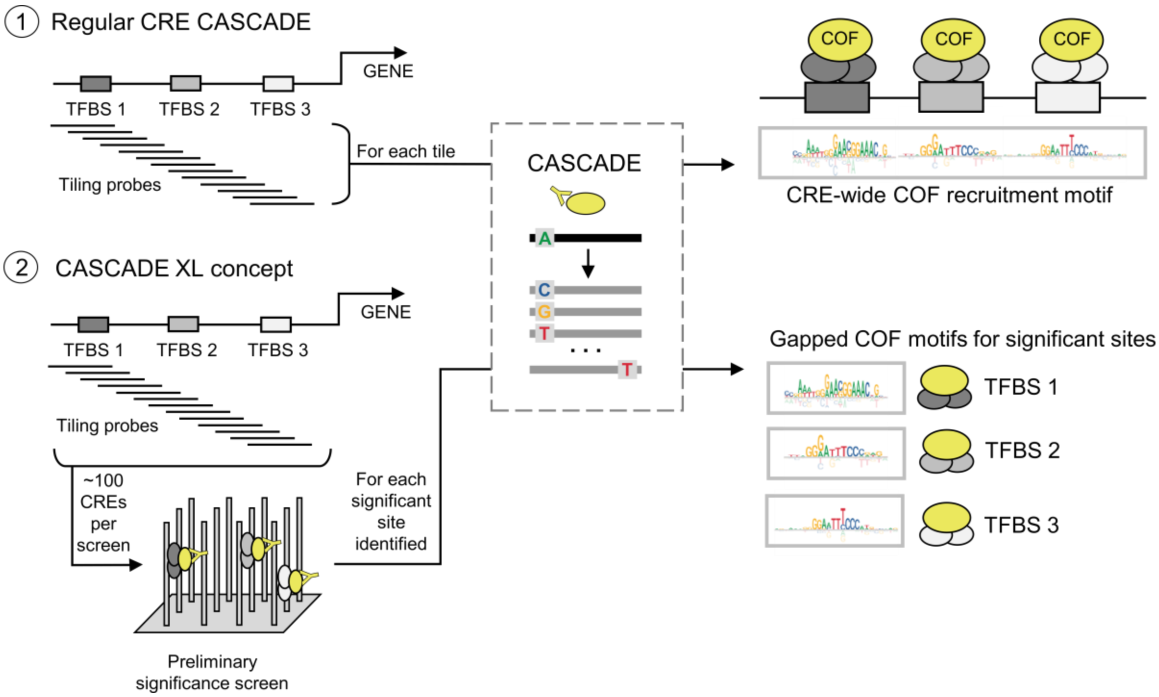


Figure 5.1: CASCADE XL concept for screening and characterizing CREs with increased throughput
(1) CRE CASCADE tiling procedure as used previously in Chapter 3 of this work. SV probes for each position on each tile were included in a CASCADE array design to generate continuous CRE-wide COF recruitment motifs. (2) CASCADE XL concept pre-screens the same tile probes for COF recruitment prior to the full CASCADE procedure enabling an increased screening throughput with the caveat that gapped CRE COF recruitment motifs are generated for significant sites only. Contributions: CASCADE XL concept was developed by DB with input from TS and HH.

Use of the CASCADE XL approach in place of the full CRE CASCADE technique outlined in Chapter 3 would result in an increased CRE screening throughput at the expense of profiling resolution. In place of a continuous CRE-wide COF recruitment motif (Fig. 5.1, part 1), the 2-step CASCADE XL approach would produce “gapped” CRE recruitment motifs that specifically examine local regions within the CRE that support the recruitment of one or more COFs screened (Fig. 5.1, part 2). Though we have yet to deploy the CASCADE XL technique, we believe that CASCADE XL will provide a useful means to investigate groups of related CREs, such as the collection of all cytokine/chemokine promoters, in a single array design to compare and contrast the use of these CREs in different cell types and conditions.

5.4 Screening NCVs for COF recruitment

As one of the major innovations presented in this work was the development of an HT differential COF recruitment screen, we have already performed a series of experiments design to test possible improvements in the detection of significant differential COF recruitment events. Within the probes used in the differential COF recruitment screen in Chapter 3, the variant position being profiled was located at the center of the profiling target region (Fig. 5.2a). Placement of the variant position within the center of the target region ultimately may result in suboptimal detection of COF recruitment at TFBS positions that are not centered. To address this possible bias, we designed an improved screen and tested it for the detection of differential COF recruitment at promoter variants associated with the development of cancer.

To circumvent the possible biases associated with having the variant position occur in the middle of the profiling region (Fig. 5.2a), we now include 3 pairs of probes representing the same variant but in different registers - with the variant position centered, shifted 5 bases to the left, and 5 bases to the right (Fig. 5.2b-c). In a pilot test where we compared the results of a triple register differential COF recruitment screen against the reference single register design for a set of NCVs associated with cancer development and control NCVs. In the triple register screening design, we observe widespread improvements in detection over our reference screen design from Chapter 3

(Fig. 5.2d). The improvements to our NCV screening method outlined here should result in a more robust differential COF recruitment detection screen that is better suited for more discovery-based research projects.

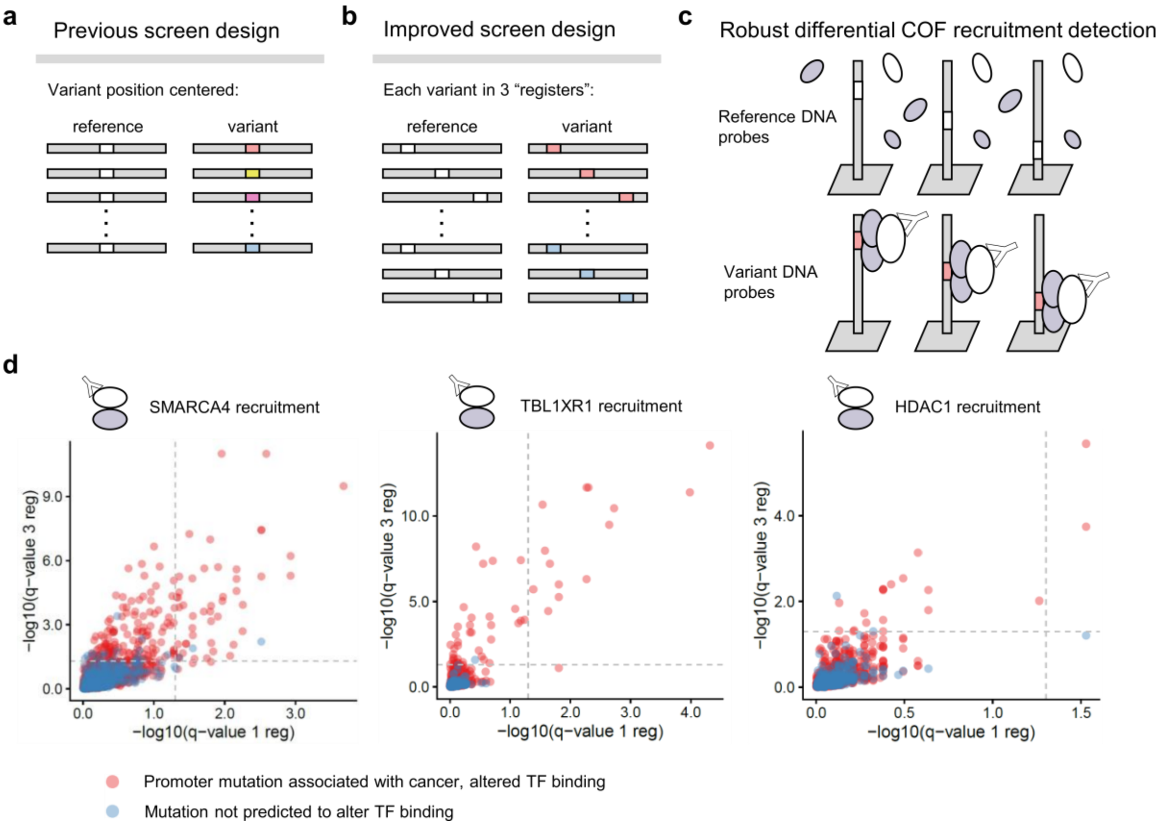


Figure 5.2: Improved COF recruitment screen overview and application to study cancer promoter variants
(a) Previous differential COF recruitment screen design used thus far. The reference and variant alleles being assayed appear exclusively at the center of the DNA probe target region. (b) An improved screen design uses 3 different "registers" to assay the variant position toward the 5' end, in the middle, and at the 3' end of the probe target region. (c) Assaying TF-COF complex formation at 3 different registers allows for more robust detection of statistically significant differences. (d) Statistically significant differential COF recruitment detected using the improved triple register design (y-axis) versus the previous single register design (x-axis) for SMARCA4 (left), TBL1XR1 (middle), and HDAC1 (right). Sites assayed represent promoter mutations associated with cancer development and altered TF binding (red) and control mutations not not associated with cancer development or predicted to alter TF binding (blue) profiled using HT29 cell nuclear extracts. Contributions: The improved triple register screen was designed by David Bray with input from Trevor Siggers and Heather Hook (HH). Sites profiled were selected by Sebastian Carrasco Pro and Juan Fuxman Bass as part of an ongoing collaboration. HH performed the experimental work.

Consistent with our efforts to improve our differential COF recruitment screens, an exciting future avenue will be to use our NCV screening approach to investigate compounds for potential therapeutic intervention. Many COFs mediate gene-regulatory functions enzymatically and commercially available compounds exist to inhibit these enzymatic functions (Yoon and Eom, 2016; Lasko et al., 2017; Fedorov et al., 2015). We explored the idea of using our platform to screen therapeutics briefly in Chapter 2 with the initial introduction of the nextPBM platform where we used a general phosphatase to disrupt the post-translational modifications required for cooperative interaction between PU.1 and IRF8. We can extend this concept in the future to our NCV differential COF recruitment screens to determine whether we can use compounds to target aberrant COF recruitment events gained with NCVs.

5.5 Expanding our repertoire of COF antibodies

A focus of our research moving forward will be the screening and validation of an increased number of COF antibodies for use in our array-based assay. The hTF array approach detailed in Chapter 4 represents the ideal platform to use in such investigations. By its nature, the hTF array can be used to rapidly screen the use of COF antibodies and since the set of TFs profiled is the same across experiments, the results across different antibodies for the same COF can in principle be compared in order to determine which would be best to use for research purposes. The hTF platform will thereby enable both rapid expansion in our repertoire of COF antibodies as well as optimization of the choice of antibody to profile for a given COF.

Though we envision that our repertoire of COF antibodies will undoubtedly increase as we scale up our investigations using the hTF array platform, we are currently inherently limited to COFs with available antibodies. As an alternative to using antibodies against the native COF complexes or subunits, we have begun investigations into using GST-tagged cloned COF subdomains in place of the native COF complexes. Not only will this allow us to expand our antibody repertoire, but this approach will also enable investigations into domain-specific recruitment of COFs to TF sites. Mapping the recruitment of subdomains within a given COF and comparing these results to those obtained with the native COF should provide a more nuanced view of the logic of TF-COF interactions. Early preliminary experiments into mapping the TFs that recruit individual subdomains of the p300 histone acetyltransferase in LPS-stimulated macrophages have been promising. Using GST-tagged subdomain clones of the p300 coactivator (Fig. 5.3a), we recapitulated known domain-TF interactions (Fig. 5.3b). Larger scale experiments are needed in order to determine whether these exogenously introduced components will accurately recapitulate the recruitment logic of the native complex from which they were designed.

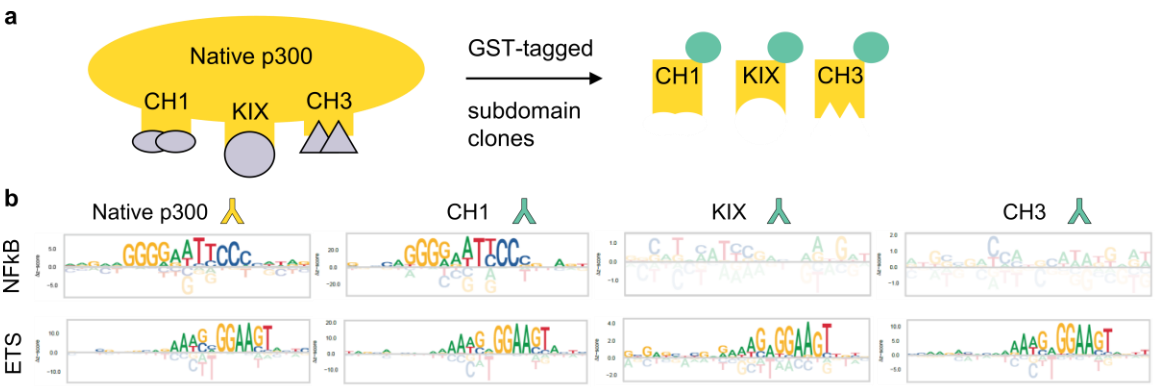


Figure 5.3: COF subdomain-specific recruitment profiling using GST-tagged clones
(a) Subdomains of a larger coactivator can be cloned and GST-tagged to enable subdomain-specific recruitment experiments. (b) Pilot recruitment data comparing the native p300 recruitment to CH1, KIX, and CH3 subdomains for NF-κB (top row) and ETS (bottom row) from the original CoRec array design. Contributions: Subdomain-specific CoRec pilot experiments were performed by Rose Zhao with input from Trevor Siggers and Jessica Keenan. David Bray designed and performed the automated CoRec analysis.

These additional investigations into expanding our COF repertoire will also enable interesting new profiling strategies. For example, pooling several of the GST-tagged subdomain components into a single experiment would allow for the examination of several COF complexes implicitly using a single anti-GST antibody. For example, pooling tagged subunits designed from various HATs such as p300, CBP, and their associated cofactors such as PCAF, and profiling their collective recruitment using the anti-GST antibody could enable to the construction of an “activation” signature that reveals the TFs responsible for mediating activator COF recruitment in a given cell type or context. Similarly, activator, repressor, and chromatin remodeling subunits could all be pooled into a single microarray chamber to build a signature of TFs broadly responsible for coordinating the major gene-regulatory effector functions in a cell type/context. As we expand our COF repertoire, we envision that these “meta”-signature approaches will become invaluable in determining which TFs most contribute to gene-regulatory function in a given condition. These signatures will provide a valuable annotation layer to be integrated with other modalities such as gene expression profiling and chromatin accessibility profiling to create detailed gene regulatory models.

BIBLIOGRAPHY

Adigun, T., Makolo, A., & Fatumo, S. (2015). Input Dataset Survey of In-Silico Tools for Inference and Visualization of Gene Regulatory Networks (GRN). Computational Biology and Bioinformatics, 3(6), 44.

Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A. J., Mann, A. L., Kundu, K., ... & Gaffney, D. J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nature genetics, 50(3), 424-431.

Altucci, L., & Rots, M. G. (2016). Epigenetic drugs: from chemistry via biology to medicine and back.

Andrabi, M., Hutchins, A., Miranda-Saavedra, D., Kono, H., Nussinov, R., Mizuguchi, K., & Ahmad, S. (2017). Predicting conformational ensembles and genome-wide transcription factor binding sites from DNA sequences. Scientific reports, 7(1), 1-16.

Andrilenas, K., Penvose, A., & Siggers, T. (2015). Using protein-binding microarrays to study transcription factor specificity: homologs, isoforms and complexes. Briefings in functional genomics, 14(1), 17-29.

Andrilenas, K., Ramlall, V., Kurland, J., Leung, B., Harbaugh, A., & Siggers, T. (2018). DNA-binding landscape of IRF3, IRF5 and IRF7 dimers: Implications for dimer-specific gene regulation. Nucleic Acids Research, 46(5), 2509-2520.

Auguie, B. (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics.

Badis, G., Berger, M., Philippakis, A., Talukder, S., Gehrke, A., Jaeger, S., . . . et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science, 324(5935), 1720-1723.

Bailey, S., Virtanen, C., Haibe-Kains, B., & Lupien, M. (2015). ABC: A tool to identify SNVs causing allele-specific transcription factor binding from ChIP-Seq experiments. Bioinformatics, 31(18), 3057-3059.

Bailey, T., Boden, M., Buske, F., Frith, M., Grant, C., Clementi, L., . . . Noble, W. (2009). MEME Suite: Tools for motif discovery and searching. Nucleic Acids Research, 37(SUPPL. 2).

Barozzi, I., Simonatto, M., Bonifacio, S., Yang, L., Rohs, R., Ghisletti, S., & Natoli, G. (2014). Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. Molecular cell, 54(5), 844-857.

Berger, M., & Bulyk, M. (2006). Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. Methods in molecular biology (Clifton, N.J.), 338, 245-260.

Berger, M., & Bulyk, M. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nature protocols, 4(3), 393.

Berger, M., Philippakis, A., Qureshi, A., He, F., Estep, P., & Bulyk, M. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nature biotechnology, 24(11), 1429-1435.

Birney, E., Stamatoyannopoulos, J., Dutta, A., Guigó, R., Gingeras, T., Margulies, E., . . . De Jong, P. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 447(7146), 799-816.

Blow, M., McCulley, D., Li, Z., Zhang, T., Akiyama, J., Holt, A., . . . Pennacchio, L. (2010). ChIP-seq identification of weakly conserved heart enhancers. Nature Genetics, 42(9), 806-812. NIH Public Access.

Bray, D., Hook, H., Zhao, R., Keenan, J. L., Penvose, A., Osayame, Y., ... & Siggers, T. (2020). Customizable high-throughput platform for profiling cofactor recruitment to DNA to characterize cis-regulatory elements and screen non-coding single-nucleotide polymorphisms. bioRxiv.

Buchkovich, M., Eklund, K., Duan, Q., Li, Y., Mohlke, K., & Furey, T. (2015). Removing reference mapping biases using limited or no genotype data identifies allelic differences in protein binding at disease-associated loci. BMC Medical Genomics, 8(1).

Chang, W. (2018). shinythemes: Themes for Shiny.

Chen, L., Ge, B., Casale, F. P., Vasquez, L., Kwan, T., Garrido-Martín, D., ... & Datta, A. (2016). Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell, 167(5), 1398-1414.

Claeys, M., Storms, V., Sun, H., Michoel, T., & Marchal, K. (2012). MotifSuite: workflow for probabilistic motif detection and assessment. Bioinformatics, 28(14), 1931-1932.

Coetzee, S. G., Coetzee, G. A., & Hazelett, D. J. (2015). motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics, 31(23), 3847-3849.

Cortez, C. C., & Jones, P. A. (2008). Chromatin, cancer and drug therapies. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 647(1-2), 44-51.

Creyghton, M., Cheng, A., Welstead, G., Kooistra, T., Carey, B., Steine, E., . . . others. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences, 107(50), 21931-21936.

Ding, C., Chan, D., Liu, W., Liu, M., Li, D., Song, L., . . . Qin, J. (2013). Proteome-wide profiling of activated transcription factors with a concatenated tandem array of transcription factor response elements. Proceedings of the National Academy of Sciences of the United States of America, 110(17), 6771-6776.

Durinck, S., Spellman, P., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. Nature Protocols, 4(8), 1184-1191.

Eklund, E., Jalava, A., & Kakar, R. (1998). PU. 1, interferon regulatory factor 1, and interferon consensus sequence-binding protein cooperate to increase gp91 phox expression. Journal of Biological Chemistry, 273(22), 13957-13965.

Ernst, J., & Kellis, M. (2012). ChromHMM: Automating chromatin-state discovery and characterization. Nature Methods, 9(3), 215-216.

Ernst, J., Kheradpour, P., Mikkelsen, T., Shoresh, N., Ward, L., Epstein, C., . . . Bernstein, B. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 473(7345), 43-49.

Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T. S., & Kellis, M. (2016). Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nature biotechnology, 34(11), 1180-1190.

Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., ... & Knight, J. C. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science, 343(6175).

Fang, B., Mane-Padros, D., Bolotin, E., Jiang, T., & Sladek, F. (2012). Identification of a binding motif specific to HNF4 by comparative analysis of multiple nuclear receptors. Nucleic acids research, 40(12), 5343-5356.

Farh, K. K. H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., ... & Hatan, M. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature, 518(7539), 337-343.

Fedorov, O., Castex, J., Tallant, C., Owen, D., Martin, S., Aldeghi, M., . . . Müller, S. (2015). Cell Biology: Selective targeting of the BRG/PB1 bromodomains impairs embryonic and trophoblast stem cell maintenance. Science Advances, 1(10).

Feingold, E., Good, P., Guyer, M., Kamholz, S., Liefer, L., Wetterstrand, K., . . . Harvey, S. (2004). The ENCODE (ENCyclopedia of DNA Elements) Project. Science, 306(5696), 636-640.

Feng, R., Desbordes, S., Xie, H., Tillo, E., Pixley, F., Stanley, E., & Graf, T. (2008). PU. 1 and C/EBPα/β convert fibroblasts into macrophage-like cells. Proceedings of the National Academy of Sciences, 105(16), 6057-6062.

Filtz, T., Vogel, W., & Leid, M. (2014). Regulation of transcription factor activity by interconnected post-translational modifications. Trends in pharmacological sciences, 35(2), 76-85.

Fornes, O., Castro-Mondragon, J., Khan, A., Van Der Lee, R., Zhang, X., Richmond, P., . . . Mathelier, A. (2020). JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. Nucleic Acids Research, 48(D1), D87-D92.

Furey, T. (2012). ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. Nature Reviews Genetics, 13(12), 840-852. Nat Rev Genet.

Galili, Tal, O'Callaghan, Alan, Sidi, Jonathan, . . . Carson. (2017). heatmaply: an R package for creating interactive cluster heatmaps for online publishing. Bioinformatics.

Gallagher, M. D., & Chen-Plotkin, A. S. (2018). The post-GWAS era: from association to function. The American Journal of Human Genetics, 102(5), 717-730.

Gan, K. A., Carrasco Pro, S., Sewell, J. A., & Fuxman Bass, J. I. (2018). Identification of single nucleotide non-coding driver mutations in cancer. Frontiers in genetics, 9, 16.

Garvie, C., & Wolberger, C. (2001). Recognition of specific DNA sequences. Molecular cell, 8(5), 937-946.

Gerritsen, M., Williams, A., Neish, A., Moore, S., Shi, Y., & Collins, T. (1997). CREB-binding protein/p300 are transcriptional coactivators of p65. Proceedings of the National Academy of Sciences of the United States of America, 94(7), 2927-2932.

Gerstein, M., Lu, Z., Van Nostrand, E., Cheng, C., Arshinoff, B., Liu, T., . . . Waterston, R. (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science, 330(6012), 1775-1787.

Ghisletti, S., Barozzi, I., Mietton, F., Polletti, S., De Santa, F., Venturini, E., . . . Natoli, G. (2010). Identification and Characterization of Enhancers Controlling the Inflammatory Gene Expression Program in Macrophages. Immunity, 32(3), 317-328.

Giorgetti, L., Siggers, T., Tiana, G., Caprara, G., Notarbartolo, S., Corona, T., . . . Natoli, G. (2010). Noncooperative Interactions between Transcription Factors and Clustered DNA Binding Sites Enable Graded Transcriptional Responses to Environmental Inputs. Molecular Cell, 37(3), 418-428.

Goodman, R. H., & Smolik, S. (2000). CBP/p300 in cell growth, transformation, and development. Genes & development, 14(13), 1553-1577.

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

Gronemeyer, H., & Bourguet, W. (2009). Allosteric effects govern nuclear receptor action: DNA appears as a player. Science Signaling, 2(73), 1-4.

Grossman, S. R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., ... & Mikkelsen, T. 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.

Guo, Y., Mahony, S., & Gifford, D. (2012). High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. PLoS Computational Biology, 8(8).

Gupta, S., Stamatoyannopoulos, J., Bailey, T., & Noble, W. (2007). Quantifying similarity between motifs. Genome Biology, 8(2), R24.

Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. Nature Reviews Molecular Cell Biology, 19(10), 621-637. Nature Publishing Group.

Hagemann, T., Biswas, S. K., Lawrence, T., Sica, A., & Lewis, C. E. (2009). Regulation of macrophage function in tumors: the multifaceted role of NF-κB. Blood, 113(14), 3139-3146.

Harley, J., Chen, X., Pujato, M., Miller, D., Maddox, A., Forney, C., . . . Weirauch, M. (2018). Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity. Nature Genetics, 50(5), 699-707.

Heintzman, N., Stuart, R., Hon, G., Fu, Y., Ching, C., Hawkins, R., . . . Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics, 39(3), 311-318.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y., Laslo, P., . . . Glass, C. (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., Benner, C., & Glass, C. (2015). The selection and function of cell type-specific enhancers. Nature Reviews Molecular Cell Biology, 16(3), 144-154. Nature Publishing Group.

Heinz, S., Romanoski, C., Benner, C., Allison, K., Kaikkonen, M., Orozco, L., & Glass, C. (2013). Effect of natural genetic variation on enhancer selection and function. Nature, 503(7477), 487-492.

Hellman, L., & Fried, M. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nature Protocols, 2(8), 1849-1861.

Honda, K., & Taniguchi, T. (2006). IRFs: master regulators of signalling by Toll-like receptors and cytosolic pattern-recognition receptors. Nature Reviews Immunology, 6(9), 644-658.

Hume, M., Barrera, L., Gisselbrecht, S., & Bulyk, M. (2015). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic acids research, 43(Database issue), D117-22.

Hwang, D., Jang, B., Yu, G., & Boudreau, M. (1997). Expression of mitogen-inducible cyclooxygenase induced by lipopolysaccharide: Mediation through both mitogen-activated protein kinase and NF-κB signaling pathways in macrophages. Biochemical Pharmacology, 54(1), 87-96.

Inukai, S., Kock, K., & Bulyk, M. (2017). Transcription factor–DNA binding: beyond binding site motifs. Current Opinion in Genetics and Development, 43, 110-119. Elsevier Ltd.

Isakova, A., Groux, R., Imbeault, M., Rainer, P., Alpern, D., Dainese, R., . . . Deplancke, B. (2017). SMiLE-seq identifies binding motifs of single and dimeric transcription factors. Nature Methods, 14(3), 316-322.

Janknecht, R., & Hunter, T. (1996). Versatile molecular glue. Transcriptional control. Current biology: CB, 6(8), 951.

Jefferies, C. (2019). Regulating IRFs in IFN driven disease. Frontiers in Immunology, 10(MAR). Frontiers Media S.A.

Johnson, D., Mortazavi, A., Myers, R., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science, 316(5830), 1497-1502.

Johnson, J., Georgakilas, G., Petrovic, J., Kurachi, M., Cai, S., Harly, C., . . . Vahedi, G. (2018). Lineage-Determining Transcription Factor TCF-1 Initiates the Epigenetic Identity of T Cells. Immunity, 48(2), 243-257.e10.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., . . . Taipale, J. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Research, 20(6), 861-873.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K., Rastas, P., . . . Taipale, J. (2013). DNA-binding specificities of human transcription factors. Cell, 152(1-2), 327-339.

Jolma, A., Yin, Y., Nitta, K., Dave, K., Popov, A., Taipale, M., . . . Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature, 527(7578), 384-388.

Keenan, J. L. (2019). Characterizing mechanisms of regulatory specificity in the nuclear receptors and general transcriptional cofactors. Boston University Theses & Dissertations.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J., Van Der Lee, R., . . . Mathelier, A. (2018). JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Research, 46(D1), D260-D266.

Kohalmi, S., Reader, L., Samach, A., Nowak, J., Haughn, G., & Crosby, W. (1998). Identification and characterization of protein interactions using the yeast 2-hybrid system. In S. Kohalmi, L. Reader, A. Samach, J. Nowak, G. Haughn, & W. Crosby, Plant Molecular Biology Manual (pp. 95-124). Springer Netherlands.

Kouzarides, T. (2007). Chromatin modifications and their function. Cell, 128(4), 693-705.

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.

Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B., & Makeev, V. J. (2013). HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic acids research, 41(Database issue), D195–D202. https://doi.org/10.1093/nar/gks1089.

Kulakovskiy, I., Vorontsov, I., Yevshin, I., Sharipov, R., Fedorova, A., Rumynskiy, E., . . . Makeev, V. (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.

Kumasaka, N., Knights, A., & Gaffney, D. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nature Genetics, 48(2), 206-213.

Laiosa, C., Stadtfeld, M., Xie, H., de Andres-Aguayo, L., & Graf, T. (2006). Reprogramming of committed T cell progenitors to macrophages and dendritic cells by C/EBPα and PU. 1 transcription factors. Immunity, 25(5), 731-744.

Lambert, S., Jolma, A., Campitelli, L., Das, P., Yin, Y., Albu, M., . . . Weirauch, M. (2018). The Human Transcription Factors. Cell, 172(4), 650-665. Cell Press.

Landt, S., Marinov, G., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., . . . Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research, 22(9), 1813-1831.

Langlais, D., Barreiro, L., & Gros, P. (2016). The macrophage IRF8/IRF1 regulome is required for protection against infections and is associated with chronic inflammation. Journal of Experimental Medicine, 213(4), 585-603.

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

Lasko, L., Jakob, C., Edalji, R., Qiu, W., Montgomery, D., Digiammarino, E., . . . Bromberg, K. (2017). Discovery of a selective catalytic p300/CBP inhibitor that targets lineage-specific tumours. Nature, 550(7674), 128-132.

Lawrence, T., & Natoli, G. (2011). Transcriptional regulation of macrophage polarization: Enabling diversity with identity. Nature Reviews Immunology, 11(11), 750-761. Nat Rev Immunol.

Lee, E., Lee, Z., & Song, Y. (2009). CXCL10 and autoimmune diseases. Autoimmunity Reviews, 8(5), 379-383.

Lee, T., & Young, R. (2013). Transcriptional regulation and its misregulation in disease. Cell, 152(6), 1237-1251. Cell.

Lee, T., Johnstone, S., & Young, R. (2006). Chromatin immunoprecipitation and microarray-based analysis of protein location. Nature protocols, 1(2), 729.

Leung, T., Hoffmann, A., & Baltimore, D. (2004). One nucleotide in a κB site can determine cofactor specificity for NF-κB dimers. Cell, 118(4), 453-464.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

Lin, Y., Jhunjhunwala, S., Benner, C., Heinz, S., Welinder, E., Mansson, R., . . . Murre, C. (2010). A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. Nature Immunology, 11(7), 635-643.

Liu, M., Guo, S., & Stiles, J. (2011). The emerging role of CXCL10 in cancer. Oncology Letters, 2(4), 583-589.

Louphrasitthiphol, P., Siddaway, R., Loffreda, A., Pogenberg, V., Friedrichsen, H., Schepsky, A., ... & Lisle, R. (2020). Tuning Transcription Factor Availability through Acetylation-Mediated Genomic Redistribution. Molecular Cell.

Ma, W., Noble, W., & Bailey, T. (2014). Motif-based analysis of large nucleotide data sets using MEME-ChIP. Nature Protocols, 9(6), 1428-1450.

Machanick, P., & Bailey, T. (2011). MEME-ChIP: Motif analysis of large DNA datasets. Bioinformatics, 27(12), 1696-1697.

MacQuarrie, K., Fong, A., Morse, R., & Tapscott, S. (2011). Genome-wide transcription factor binding: Beyond direct target regulation. Trends in Genetics, 27(4), 141-148. Elsevier Current Trends.

Maerkl, S., & Quake, S. (2007). A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. Science, 315(January), 233-238.

Majewski, J., & Pastinen, T. (2011). The study of eQTL variations by RNA-seq: From SNPs to phenotypes. Trends in Genetics, 27(2), 72-79.

Majumder, S., Zhou, L. Z. H., Chaturvedi, P., Babcock, G., Aras, S., & Ransohoff, R. M. (1998). p48/STAT-1α-containing complexes play a predominant role in induction of IFN-γ-inducible protein, 10 kDa (IP-10) by IFN-γ alone or in synergy with TNF-α. The Journal of Immunology, 161(9), 4736-4744.

Mancino, A., Termanini, A., Barozzi, I., Ghisletti, S., Ostuni, R., Prosperini, E., . . . Natoli, G. (2015). A dual cis-regulatory code links IRF8 to constitutive and inducible gene expression in macrophages. Genes & development, 29(4), 394-408.

Maurano, M., Humbert, R., Rynes, E., Thurman, R., Haugen, E., Wang, H., . . . Stamatoyannopoulos, J. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science, 337(6099), 1190-1195.

Medzhitov, R., & Horng, T. (2009). Transcriptional control of the inflammatory response. Nature Reviews Immunology, 9(10), 692-703.

Meijsing, S., Pufall, M., So, A., Bates, D., Chen, L., & Yamamoto, K. (2009). DNA binding site sequence directs glucocorticoid receptor structure and activity. Science, 324(5925), 407-410.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., ... & Kellis, M. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nature biotechnology, 30(3), 271-277.

Mogensen, T. (2019). IRF and STAT transcription factors - From basic biology to roles in infection, protective immunity, and primary immunodeficiencies. Frontiers in Immunology, 10(JAN). Frontiers Media S.A.

Mohaghegh, N., Bray, D., Keenan, J., Penvose, A., Andrilenas, K. K., Ramlall, V., & Siggers, T. (2019). NextPBM: a platform to study cell-specific transcription factor binding and cooperativity. Nucleic acids research, 47(6), e31-e31.

Mohammed, H., Taylor, C., Brown, G., Papachristou, E., Carroll, J., & D'Santos, C. (2016). Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) for analysis of chromatin complexes. Nature Protocols, 11(2), 316-326. Nature Publishing Group.

Mukherjee, S., Berger, M., Jona, G., Wang, X., Muzzey, D., Snyder, M., . . . Bulyk, M. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nature Genetics, 36(12), 1331-1339.

Natoli, G., Ghisletti, S., & Barozzi, I. (2011). The genomic landscapes of inflammation. Genes and Development, 25(2), 101-106.

Nerlov, C., & Graf, T. (1998). PU. 1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. Genes & development, 12(15), 2403-2412.

Neuwirth, E. (2014). RColorBrewer: ColorBrewer Palettes.

Newburger, D. E., & Bulyk, M. L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic acids research, 37(Database issue), D77–D82. https://doi.org/10.1093/nar/gkn660.

Ogawa, S., Lozach, J., Benner, C., Pascual, G., Tangirala, R., Westin, S., . . . Glass, C. (2005). Molecular determinants of crosstalk between nuclear receptors and toll-like receptors. Cell, 122(5), 707-721.

Ohmori, Y., & Hamilton, T. A. (1993). Cooperative interaction between interferon (IFN) stimulus response element and kappa B sequence motifs controls IFN gamma-and lipopolysaccharide-stimulated transcription from the murine IP-10 promoter. Journal of Biological Chemistry, 268(9), 6677-6688.

O'Neill, L. (2006). How Toll-like receptors signal: What we know and what we don't know. Current Opinion in Immunology, 18(1), 3-9.

Ostuni, R., Piccolo, V., Barozzi, I., Polletti, S., Termanini, A., Bonifacio, S., . . . Natoli, G. (2013). Latent enhancers activated by stimulation in differentiated cells. Cell, 152(1-2), 157-171.

Penvose, A., Keenan, J., Bray, D., Ramlall, V., & Siggers, T. (2019). Comprehensive study of nuclear receptor DNA binding provides a revised framework for understanding receptor specificity. Nature Communications, 10(1).

Pham, T., Benner, C., Lichtinger, M., Schwarzfischer, L., Hu, Y., Andreesen, R., . . . Rehli, M. (2012). Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. Blood, 119(24), e161-e171.

Pham, T.-H., Minderjahn, J., Schmidl, C., Hoffmeister, H., Schmidhofer, S., Chen, W., . . . Rehli, M. (2013). Mechanisms of in vivo binding site selection of the hematopoietic master transcription factor PU. 1. Nucleic acids research, 41(13), 6391-6402.

Puckett, J., Muzikar, K., Tietjen, J., Warren, C., Ansari, A., & Dervan, P. (2007). Quantitative microarray profiling of DNA-binding molecules. Journal of the American Chemical Society, 129(40), 12310-12319.

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

Raisner, R., Kharbanda, S., Jin, L., Jeng, E., Chan, E., Merchant, M., . . . Gascoigne, K. (2018). Enhancer Activity Requires CBP/P300 Bromodomain-Dependent Histone H3K27 Acetylation. Cell Reports, 24(7), 1722-1729.

Ramos, Y., Hestand, M., Verlaan, M., Krabbendam, E., Ariyurek, Y., van Galen, M., . . . 't Hoen, P. (2010). Genome-wide assessment of differential roles for p300 and CBP in transcription regulation. Nucleic acids research, 38(16), 5396-408.

Rehli, M., Poltorak, A., Schwarzfischer, L., Krause, S., Andreesen, R., & Beutler, B. (2000). PU. 1 and interferon consensus sequence-binding protein regulate the myeloid expression of the human Toll-like receptor 4 gene. Journal of Biological Chemistry, 275(13), 9773-9781.

Reiter, F., Wienerroither, S., & Stark, A. (2017). Combinatorial function of transcription factors and cofactors. Current Opinion in Genetics and Development, 43, 73-81. Elsevier Ltd.

Riegel, E., Heimbucher, T., Höfer, T., & Czerny, T. (2017). A sensitive, semi-quantitative mammalian two-hybrid assay. BioTechniques, 62(5), 206-214.

Riley, T., Slattery, M., Abe, N., Rastogi, C., Liu, D., Mann, R., & Bussemaker, H. (2014). SELEX-seq: A method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. Methods in Molecular Biology, 1196, 255-278.

Rojano, E., Seoane, P., Ranea, J., & Perkins, J. (2019). Regulatory variants: From detection to predicting impact. Briefings in Bioinformatics, 20(5), 1639-1654.

Rosenbauer, F., & Tenen, D. (2007). Transcription factors in myeloid development: balancing differentiation with transformation. Nature Reviews Immunology, 7(2), 105-117.

RStudio Inc. (2013). Easy web applications in R.

Sakaguchi, S., Negishi, H., Asagiri, M., Nakajima, C., Mizutani, T., Takaoka, A., . . . Taniguchi, T. (2003). Essential role of IRF-3 in lipopolysaccharide-induced interferon-β gene expression and endotoxin shock. Biochemical and Biophysical Research Communications, 306(4), 860-866.

Sandelin, A. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research, 32(90001), 91D-94.

Schmiedel, B. J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A. G., White, B. M., Zapardiel-Gonzalo, J., ... & Seumois, G. (2018). Impact of genetic polymorphisms on human immune cell gene expression. Cell, 175(6), 1701-1715.

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic acids research, 18(20), 6097–6100. https://doi.org/10.1093/nar/18.20.6097.

Scott, E., Simon, M., Anastasi, J., & Singh, H. (1994). Requirement of transcription factor PU. 1 in the development of multiple hematopoietic lineages. Science, 265(5178), 1573-1577.

Shannon, P., & Richards, M. (2018). MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs.

Sharf, R., Meraro, D., Azriel, A., Thornton, A., Ozato, K., Petricoin, E., . . . Levi, B.-Z. (1997). Phosphorylation events modulate the ability of interferon consensus sequence binding protein to interact with interferon regulatory factors and to bind DNA. Journal of Biological Chemistry, 272(15), 9785-9792.

Shi, W., Fornes, O., Mathelier, A., & Wasserman, W. (2016). Evaluating the impact of single nucleotide variants on transcription factor binding. Nucleic Acids Research, 44(21), 10106-10116.

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: From properties to genome-wide predictions. Nature Reviews Genetics, 15(4), 272-286. Nature Publishing Group.

Siggers, T., & Gordân, R. (2014). Protein–DNA binding: complexities and multi-protein codes. Nucleic acids research, 42(4), 2099-2111.

Siggers, T., Chang, A., Teixeira, A., Wong, D., Williams, K., Ahmed, B., . . . Bulyk, M. (2012). Principles of dimer-specific gene regulation revealed by a comprehensive characterization of NF-κB family DNA binding. Nature Immunology, 13(1), 95-102.

Siggers, T., Duyzend, M., Reddy, J., Khan, S., & Bulyk, M. (2011). Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. Molecular Systems Biology, 7(555), 1-14.

Siggers, T., Gilmore, T., Barron, B., & Penvose, A. (2015). Characterizing the DNA binding site specificity of NF-κB with protein-binding microarrays (PBMs). In T. Siggers, T. Gilmore, B. Barron, & A. Penvose, NF-kappa B (pp. 609-630). Springer.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., . . . et al. (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell, 147(6), 1270-1282.

Slattery, M., Zhou, T., Yang, L., Machado, A., Gordân, R., & Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends in biochemical sciences, 39(9), 381-399.

Smale, S. (2012). Transcriptional regulation in the innate immune system. Current Opinion in Immunology, 24(1), 51-57.

Soccio, R. E., Chen, E. R., Rajapurkar, S. R., Safabakhsh, P., Marinis, J. M., Dispirito, J. R., ... & Lim, H. W. (2015). Genetic variation determines PPARγ function and anti-diabetic drug response in vivo. Cell, 162(1), 33-44.

Stormo, G., & Fields, D. (1998). Specificity, free energy and information content in protein-DNA interactions. Trends in Biochemical Sciences, 23(3), 109-113. Elsevier Ltd.

Sun, W., & Hu, Y. (2013). eQTL Mapping Using RNA-seq Data. Statistics in Biosciences, 5(1), 198-219.

Tamura, T., Yanai, H., Savitsky, D., & Taniguchi, T. (2008). The IRF family transcription factors in immunity and oncogenesis. Annual Review of Immunology, 26, 535-584.

Tan, G., & Lenhard, B. (2016). TFBSTools: an R/bioconductor package for transcription factor binding site analysis. Bioinformatics, 32(10), 1555-1556.

Tewhey, R., Kotliar, D., Park, D., Liu, B., Winnicki, S., Reilly, S., . . . Sabeti, P. (2016). Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell, 165(6), 1519-1529.

Tootle, T., & Rebay, I. (2005). Post-translational modifications influence transcription factor activity: a view from the ETS superfamily. Bioessays, 27(3), 285-298.

Touzet, H., & Varré, J. S. (2007). Efficient and accurate P-value computation for Position Weight Matrices. Algorithms for Molecular Biology, 2(1), 15.

Tremblay, B.-M. (2019). universalmotif: Import, Modify, and Export Motifs with R.

Valouev, A., Johnson, D., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., . . . Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nature Methods, 5(9), 829-834.

Van De Geijn, B., Mcvicker, G., Gilad, Y., & Pritchard, J. (2015). WASP: Allele-specific software for robust molecular quantitative trait locus discovery. Nature Methods, 12(11), 1061-1063. Nature Publishing Group.

Vaquerizas, J., Kummerfeld, S., Teichmann, S., & Luscombe, N. (2009). A census of human transcription factors: Function, expression and evolution. Nature Reviews Genetics, 10(4), 252-263.

Vo, N., & Goodman, R. (2001). CREB-binding Protein and p300 in Transcriptional Regulation. Journal of Biological Chemistry, 276(17), 13505-13508. American Society for Biochemistry and Molecular Biology Inc.

Wagih, O. (2017). ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics, 33(22), 3645-3647.

Warren, C. (2005). Defining the sequence-recognition profile of DNA-binding molecules. Proceedings of the National Academy of Sciences, 103(4), 867-872.

Wei, B., Jolma, A., Sahu, B., Orre, L., Zhong, F., Zhu, F., . . . Taipale, J. (2018). A protein activity assay to measure global transcription factor activity reveals determinants of chromatin accessibility. Nature Biotechnology, 36(6), 521-529.

Wei, G., Badis, G., Berger, M., Kivioja, T., Palin, K., Enge, M., . . . Taipale, J. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO Journal, 29(13), 2147-2160.

Weinert, B., Narita, T., Satpathy, S., Srinivasan, B., Hansen, B., Schölz, C., . . . Choudhary, C. (2018). Time-Resolved Analysis Reveals Rapid Dynamics and Broad Scope of the CBP/p300 Acetylome. Cell, 174(1), 231-244.e12.

Weirauch, M., Yang, A., Albu, M., Cote, A., Montenegro-Montero, A., Drewe, P., . . . Hughes, T. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell, 158(6), 1431-1443.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wilke, C. (2019). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'.

Wingender, E. (2000). TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Research, 28(1), 316-319.

Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Briefings in bioinformatics, 9(4), 326-32.

Xie, H., Ye, M., Feng, R., & Graf, T. (2004). Stepwise reprogramming of B cells into macrophages. Cell, 117(5), 663-676.

Xie, Q., Kashiwabara, Y., & Nathan, C. (1994). Role of transcription factor NF-κB/Rel in induction of nitric oxide synthase. Journal of Biological Chemistry, 269(7), 4705-4708.

Xie, Y., Cheng, J., & Tan, X. (2020). DT: A Wrapper of the JavaScript Library 'DataTables'.

Xing, H., Mo, Y., Liao, W., & Zhang, M. (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).

Yang, C., Shapiro, L., Rivera, M., Kumar, A., & Brindle, P. (1998). A Role for CREB Binding Protein and p300 Transcriptional Coactivators in Ets-1 Transactivation Functions. Molecular and Cellular Biology, 18(4), 2218-2229.

Yang, L., Zhou, T., Dror, I., Mathelier, A., Wasserman, W., Gordân, R., & Rohs, R. (2014). TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic acids research, 42(D1), D148–D155.

Yoon, S., & Eom, G. (2016). HDAC and HDAC Inhibitor: From Cancer to Cardiovascular Diseases. Chonnam Medical Journal, 52(1), 1.

Zabidi, M., & Stark, A. (2016). Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors. Trends in Genetics, 32(12), 801-814. Elsevier Ltd.

Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D., Bernstein, B., . . . Shirley, X. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biology, 9(9).

Zhao, R. W. (2020). Characterizing cell-specific transcriptional recruitomes using a high-throughput microarray platform. Boston University Theses & Dissertations.

Zhou, Q., Liu, M., Xia, X., Gong, T., Feng, J., Liu, W., . . . Qin, J. (2017). A mouse tissue transcription factor atlas. Nature Communications, 8.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, 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.

CURRICULUM VITAE

[redacted]

[redacted]

[redacted]

[redacted]