2019

Characterizing mechanisms of regulatory specificity in the nuclear receptors and general transcriptional cofactors

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES AND COLLEGE OF ENGINEERING

Dissertation

CHARACTERIZING MECHANISMS OF REGULATORY SPECIFICITY IN THE NUCLEAR RECEPTORS AND GENERAL TRANSCRIPTIONAL COFACTORS

by

JESSICA KEENAN

B.S., Massachusetts Institute of Technology, 2009
M.S., Boston University, 2014

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2019

Approved by

First Reader _____

Trevor Siggers, Ph.D.
Associate Professor of Biology

Second Reader _____

Juan Fuxman Bass, Ph.D.
Professor of Biology

ACKNOWLEDGMENTS

I wish to sincerely thank the many people in my life who made this work possible:

- *My loving family for their continuous love and support, without whom I would not have been able to follow the path that led me here.*
- *Trevor Siggers, for his constant support, mentorship, and passion for research.*
- *The members of the Siggers Lab, past and present, who have always been collaborative, creative, and the best colleagues I could have asked for during my PhD work.*
- *The many fantastic mentors I have had throughout my scientific journey, who used their time and experience to help guide my path.*
- *The staff of the Bioinformatics Program, who were always there to lend a helping hand when I needed it.*
- *The members of the Bioinformatics Program, with whom I shared this journey, and all of its ups and downs. I am fortunate to have been surrounded with such kind, creative, and passionate people.*
- *My friends outside of lab, for their encouragement, support, and reminding me to enjoy life outside of lab.*

CHARACTERIZING MECHANISMS OF REGULATORY SPECIFICITY IN THE NUCLEAR RECEPTORS AND GENERAL TRANSCRIPTIONAL COFACTORS

Jessica Keenan

Boston University Graduate School of Arts and Sciences and College of Engineering, 2019

Major Professor: Trevor Siggers, Associate Professor of Biology

ABSTRACT

Gene regulation, at its most basic level, is controlled by transcription factors (TFs) binding to genomic regulatory elements and recruiting regulatory cofactors (CoFs). Therefore, to understand specificity in gene regulation, we must address how TF-DNA binding translates into target gene specification in the genome and how TF-CoF interactions are regulated within the cell. Towards this goal, we describe a comprehensive study of the DNA binding specificity of the type II nuclear receptor (NR) family of TFs, and introduce a novel high-throughput technique for assaying the many TF-CoF complexes functioning in a cell.

The type II nuclear receptors function as heterodimeric TFs with the retinoid X receptor (RXR) to regulate diverse biological processes. DNA-binding specificity has been proposed as a primary mechanism for NR gene regulatory specificity. We use protein-binding microarrays (PBMs) to comprehensively analyze the DNA binding of 12 NR:RXRα heterodimers. We find more promiscuous NR-DNA binding than has been

reported, challenging the view that NR binding specificity is defined by half-site spacing. We show that NRs bind DNA using two distinct modes, explaining widespread NR binding to half-sites *in vivo*. Finally, we show that the current models of NR specificity better reflect binding-site activity rather than binding-site affinity. Our rich dataset and revised NR binding models provide a framework for understanding NR regulatory specificity and will facilitate more accurate analyses of genomic datasets.

Central to gene regulation is the recruitment of CoFs (e.g., co-activators and co-repressors) to DNA by site-specific TFs. There are currently no high-throughput approaches to identify and characterize the many TF-cofactor complexes simultaneously operating in a cell. To this end, we have developed the CoRec (Cofactor Recruitment) approach to monitor cofactor recruitment by hundreds of TFs from nuclear lysates. We have used CoRec to examine CoF recruitment in resting and LPS-activated human macrophages, as well as resting and T cell receptor-stimulated human T cells. We demonstrate CoF recruitment to both known and novel regulatory elements and compare regulatory strategies between these two cell types. We anticipate CoRec will be a powerful tool to study the assembly and regulation of nuclear TF-cofactor complexes in a cellular context.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUC ..... Area Under the Curve

bp..... Base Pair

bZIP..... Basic Leucine Zipper

CBP ..... CREB-Binding Protein

ChIP .....Chromatin Immunoprecipitation

CoF.....Regulatory Cofactor

CRE.....cis-Regulatory Element (generally) or cAMP Response Element (as defined in the relevant text)

DBD .....DNA-Binding Domain

DR ..... Direct Repeat

EMSA ..... Electrophoretic Mobility Shift Assay

HDAC .....Histone Deacetylase

HT .....High Throughput

IP .....Immunoprecipitation

ISRE.....Interferon-stimulated Response Element

kb..... kilo base pairs

LPS.....Lipopolysaccharide

NR.....Type II Nuclear Receptor

PBM .....Protein-Binding Microarray

PMA.....Phorbol 12-Myristate 13-Acetate

PTM .....Post-Tanslational Modification

PWM .....Position Weight Matrix

ROC ..... Receiver Operating Characteristic

SNP ..... Single Nucleotide Polymorphism

SNV.....Single Nucleotide Variant

TCR..... T Cell Receptor

TF .....Transcription Factor

THP-1..... A human monocyte cell line

TLR .....Toll-Like Receptor

CHAPTER ONE: Introduction

1.1 Introduction

Gene regulation is fundamental to all biological processes, and cells utilize a wide variety of strategies to ensure that genes are expressed in the correct tissue type, amount, and for the appropriate amount of time. At the most basic level, gene expression is controlled by the binding of transcription factors (TFs) to specific cis-regulatory elements (CREs) and the TF-dependent recruitment of regulatory cofactor (CoF) proteins to the locus. In this way, CREs, such as enhancers and promoters, integrate information about cell-state (e.g., cell type or metabolic state) and the extracellular environment (e.g., signaling molecules), and translate these signals into gene expression changes. Thus, CREs represent a genomic encoding of information required for the proper regulation of gene expression. Among CREs, enhancers play a key role in mediating cell-specific gene expression patterns, often operating over large genomic distances (i.e., from several kilobases to megabases) to alter the transcription of target genes (Calo and Wysocka, 2013). The realization that certain chromatin features can be used to identify and annotate enhancers has fueled genomic-profiling efforts aimed at cataloging the enhancer repertoire in cells. The ENCODE project, the largest of these efforts (reviewed in (Buecker and Wysocka, 2012)), has identified ~400,000 enhancer elements in the examined cell lines. Driven by advances in genomic profiling technologies, research efforts to identify enhancers and other CREs have greatly outpaced our efforts to describe the molecular mechanisms by which they function to alter gene expression. Therefore, to work toward a global understanding of how genes are regulated in cells, there is a need to investigate the

molecular composition and function of the hundreds of thousands of annotated enhancers and promoters. Critical to this effort will be methods to identify the regulatory proteins that assemble at CREs to mediate their function. Toward this goal, in this thesis, we first focus on characterizing the sequence determinants of DNA binding for the type II nuclear receptors (NRs), an important class of mammalian TFs that regulates diverse physiological responses to ligands. We then go on to demonstrate a novel high-throughput approach for characterizing the many DNA-TF-CoF complexes operating in a cell at a given time.

1.2 Elucidating DNA Determinants of Nuclear Receptor Binding to CREs

1.2.1 Nuclear Receptor Structure, Function, and Regulation

The human nuclear receptor superfamily comprises 48 TFs, many of which function in a ligand-dependent manner to regulate gene expression programs in development, metabolism, homeostasis, and inflammation (Evans and Mangelsdorf, 2014) (Perissi and Rosenfeld, 2005). Most NRs exhibit a common architecture involving a central DNA-binding domain (DBD) and a carboxy-terminal ligand-binding domain. The type II NRs can bind DNA as dimers with RXR partners (i.e., RXRα, RXRβ, and RXRγ) (Evans and Mangelsdorf, 2014). Despite their common RXR partners, the type II NRs, hereafter simply ‘NRs’, mediate signals from a diverse array of lipophilic ligands (Table 1).

Nuclear Receptor	Ligands
Liver X Receptor (LXR)	oxysterols
Farnesoid X Receptor (FXR)	bile acids
Peroxisome Proliferator Activated Receptor (PPAR)	fatty acids
Retinoic Acid Receptor (RAR)	retinoids
Retinoid X Receptor (RXR)	retinoids
Pregnane X Receptor (PXR)	xenobiotics
Constitutive Androstane Receptor (CAR)	xenobiotics

Table 1: Examples of the type II NRs and their ligands
Adapted from Mangelsdorf 2014.

As an example of the complex processes regulated by the NRs, we briefly review the roles of the NRs LXR, FXR, and the α, δ, and γ isotypes of PPAR in metabolic homeostasis (Evans and Mangelsdorf, 2014). Upon ingestion of food, LXR is activated by binding to cholesterol, leading to the activation of gene networks that facilitate the transport of cholesterol from the periphery to the liver, and its subsequent catabolism into bile acids. FXR is activated by binding to bile acids, stimulating diverse responses; FXR activation leads to the upregulation of nutrient transporters, activation of liver metabolism, and control of intestinal inflammation. In conjunction with the LXR and FXR responses to cholesterol and bile acids, increased fatty acid levels activate the three known isotypes of PPAR: PPARα, PPARδ, and PPARγ. Upon binding fatty acids, PPARα and PPARδ stimulate fatty acid catabolism. As a complement to these processes, PPARγ binding to fatty acids initiates genetic programs that promote the storage of excess energy in adipose tissue. Interestingly, PPARα also plays an important role in the

response to starvation. Upon sensing decreased fatty acid levels, PPARα triggers conversion of fatty acids into useable energy sources and sends stress signals to other parts of the body, facilitating a coordinated response to the lack of nutrients (Evans and Mangelsdorf, 2014).

In contrast to the NRs that regulate metabolic homeostasis, other NRs, such as PXR and CAR are responsible for the clearance of toxic small molecules. The binding domains of these NRs are highly promiscuous, allowing them to be activated by hundreds of different ligands, many of which are dietary and drug metabolites. Upon binding to a ligand, VDR and CAR activate genetic programs that facilitate clearance of xenobiotics by increasing expression of detoxification enzymes (Evans and Mangelsdorf, 2014). Thus, the NRs play key roles in the regulation of diverse and complex genetic programs.

In the canonical model of type II NR signaling, NRs are bound to DNA in complex with transcriptional corepressors in the absence of ligand. Upon ligand binding, coactivators take the place of the corepressors, leading to gene expression (McKenna and O'Malley, 2002). The NRs VDR and CAR function in a different manner, residing in the cytosol in the absence of ligand. Upon ligand binding they translocate to the nucleus to bind DNA, recruit coactivators, and regulate target genes. In both forms of activation, ligand-activated NRs bind to DNA cis-regulatory elements (CREs) in promoters or enhancers of target genes (Khorasanizadeh and Rastinejad, 2001). Many NRs are expressed in the same tissues and at the same times (Bookout et al., 2006). Consequently, the differential

binding of RXR dimers to different CREs in the genome is the primary mechanism of specificity in NR signaling (Claessens and Gewirth, 2004). The NRs represent a mammalian signaling family in which ligands bind directly to transcriptional regulators to effect gene expression; subsequently, fewer (or even single) TFs are activated in response to ligand, simplifying the regulatory network. Therefore, the NRs represent an ideal model system in mammalian signaling in which to test our ability to generate predictive models of target gene specificity.

1.2.2 Current Models of Nuclear Receptor Regulatory Specificity Are Incomplete

DNA-binding differences between RXR dimers is the primary determinant in NR signaling specificity, dictating the genomic loci to which NRs bind to regulate target genes (Rastinejad et al., 2013). Consensus NR binding sites consist of direct repeats (DRs) of the 5'-AGGTCA-3' hexamer separated by a variable length spacer (Rastinejad et al., 2013). For example, a site comprising direct repeats with a 1-base pair (bp) spacer – 5'-AGGTCANAGGTCA-3' – is referred to as a DR1 site, and binds well to RAR:RXR and PPAR:RXR dimers. A distinguishing feature for different RXR dimers is the spacer length between the half-sites in their target binding sites**.** For example, PPAR:RXR dimers prefer binding to DR1 elements, whereas the LXR:RXR dimers prefer DR4 elements (Khorasanizadeh and Rastinejad, 2001). Genome-wide studies, using chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq), have confirmed these basic dimer binding preferences, and reinforced the connection between *in vitro* and *in vivo* binding (Rastinejad et al., 2013) (Boergesen et al., 2012) (Lefterova et al., 2010). However, these studies have also revealed limitations to our current models of

NR specificity. First, the current position weight matrix (PWM) models for DNA-binding of RXR dimers are either too restrictive, as they only identify NR motifs in a minority of genomic binding loci (e.g., 4-10% for PPARγ and LXR (Boergesen et al., 2012)), or they are too degenerate and do not accurately capture the impact of DNA polymorphisms. A second limitation to our models is that RXR dimers are described using a single PWM representing a single DR-type element. However, NR binding is more flexible than this and NRs can recognize different DR elements. For example, a study of PPAR:RXR and LXR:RXR dimers in mouse revealed that these dimers share many binding regions (71-88%), and bind to degenerate response elements in a mutually exclusive manner, highlighting that the simple PPAR-DR1/LXR-DR4 model described above is insufficient. Similarly, a ChIP-seq study of LXR in THP-1 macrophages found DR4 sites as the most abundant DR spacing, but it was only found in 7-40% of sites (depending on threshold). However, other NR-type sites were found, suggesting alternative recognition sequences (Pehkonen et al., 2012). In Chapter 2, we address this deficiency by characterizing the DNA binding profiles of 12 RXR dimers in a comprehensive manner to all DR spacings (DR0-DR5).

1.2.3 High-throughput Methods for Characterizing Protein Interactions with DNA

Many techniques have been developed to characterize the interactions between transcription factors and DNA sequences. Each of these methods facilitate a deeper understanding of the role of regulatory elements in gene regulation, while also having drawbacks. Furthermore, integrating information gleaned from multiple methods can

provide an even more robust understanding of DNA regulatory elements. Here, we review several of these methods.

1.2.3.1 Chromatin immunoprecipitation followed by sequencing

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a widely-used method for inferring DNA sequences that either directly or indirectly recruit specific regulatory proteins. In a ChIP-seq experiment, a tissue of interest is treated with a fixing agent to chemically crosslink proteins with each other and with DNA to which they may be bound. An antibody specific to the protein of interest is used to immunoprecipitate the protein along with any DNA to which it has been crosslinked, providing a set of DNA sequences co-immunoprecipitated with the protein of interest. These sequences are then aligned to the appropriate genome, and "peaks" of sequences that align to the same region are identified, representing likely binding sites. However, ChIP-seq peaks are generally large (> 500 bp) compared to regulatory elements (generally 5-15 bp). Thus, computational methods, such as *de novo* motif analysis, or scanning with known PWMs must then be used to infer which DNA sequences were responsible for recruiting the protein of interest. Thus, ChIP-seq facilitates the inference of genomic regulatory elements responsible for recruiting specific proteins in a high-throughput manner. However, this inference is often inefficient; many peaks will not contain identifiable binding sites for the ChIP-ed TF. In contrast, multiple potential binding sites may be identified within a given peak, and the contribution of each site to protein recruitment cannot be elucidated without additional experimental analyses. Additionally, ChIP-seq is limited to genomic DNA sequences in their native context; thus elucidating the impact of specific variants is limited by the ability to produce samples with these variations.

1.2.3.2 Systematic evolution of ligands by exponential enrichment followed by sequencing

Systematic evolution of ligands by exponential enrichment followed by sequencing (SELEX-seq) is another technique that has greatly expanded insight into DNA-protein interactions (Riley et al., 2014). SELEX-seq utilizes a library of DNA probes, which is then incubated with protein of interest. The DNA probes are then co-immunoprecipitated with the protein of interest. The DNA can be amplified, and additional rounds of selection performed, after which the DNA library is sequenced. This method facilitates the identification of TF-specific DNA motifs, and unlike ChIP-seq is not limited to genomic sequences.

1.2.3.3 Protein-binding Microarrays

PBMs are arrays of double stranded DNA oligonucleotides that allow for the simultaneous measurement of the relative affinities for a protein to tens of thousands of DNA sequences (Berger et al., 2006). In a PBM experiment, the protein of interest is incubated on the array and subsequently probed with a protein-specific primary antibody and fluorescently-labeled secondary antibody. The fluorescence intensity for a given probe is proportional to the protein's affinity for the DNA probe sequence. To facilitate the characterization of DNA determinants of protein recruitment at a single nucleotide resolution, our lab routinely uses a 'seed and SNV' approach (Andrilenas et al., 2018; Mohaghegh et al., 2019; Penvose et al., 2019). For a given sequence of interest, for example, a consensus response element for a protein of interest, probes containing all single nucleotide variants of the sequence are generated, allowing the direct measurement of the effect of each nucleotide on protein binding. Using this approach, binding logos

can be generated that represent the binding energy contribution of each nucleotide within the binding site.

1.2.4 PBM Approach for Characterizing NR-DNA Binding Preferences

In Chapter 2 of this thesis, we utilize PBMs to comprehensively analyze the DNA-binding preferences of 12 type II NRs (Penvose et al., 2019). We use custom PBMs that contain 24 different seeds and all SNV variants for each DR with spacers from 0 – 5 nt, allowing us to examine the DNA binding preferences of NRs to a wide range of possible binding sites with single nucleotide resolution. We find that NRs bind DNA with greater promiscuity than previously reported, and that these non-canonical binding sites are consistent with published ChIP-seq data and can regulate gene expression *in vivo*. These analyses provide a revised framework for understanding NR-DNA binding.

1.3 A High-throughput Approach to Examining Recruitment of Regulatory Complexes to DNA

1.3.1 Elucidation of Cofactor Recruitment to Cis-regulatory Elements is Fundamental to Understanding Gene Regulation

While TF-DNA binding is key to gene regulation, it represents only one part of complex set of well-choreographed events required for proper gene regulation. Central to TF function is the recruitment of regulatory cofactors (CoFs) to DNA (Roeder, 2005; Weake and Workman, 2010). CoFs (i.e., co-activators or co-repressors), which often lack sequence-specific DNA binding ability, are recruited to DNA by TFs to facilitate transcriptional control of genes. CoFs mediate diverse processes required for the gene

regulation, including histone modification (e.g., histone acetylation or de-acetylation), ATP-dependent chromatin remodeling, or facilitating the formation and function of the preinitiation complex (Roeder, 2005). Thus, TFs act as biological adaptors, responsible for recruiting additional regulatory proteins to specific loci. Consequently, TF-CoF complexes are critical to diverse steps in transcriptional control, and delineating the TF-CoF complexes functioning in a cell is critical to understanding the control of gene expression in healthy and disease cell contexts. However, TF-CoF complexes are not routinely analyzed at a high-throughput level, leaving this central aspect of gene regulation greatly understudied. As TF-CoF complexes can be regulated at the protein level by nuclear localization, PTMs (Filtz et al., 2014; Tootle and Rebay, 2005), and auxiliary CoFs (Siggers et al., 2011), they cannot be reliably inferred using transcript-level analyses. Therefore, there is a need for HT protein-based techniques to characterize the potentially hundreds of TF-CoF complexes operating in a cell. Here, we review the role of select CoFs, their mechanisms of transcriptional control, and their recruitment by TFs. We then describe approaches for the HT characterization of TF-CoF complexes in the cell.

1.3.2 Cofactors and their Roles in Gene Regulation

1.3.2.1 P300/CBP

The transcriptional activator p300, and its homolog CREB binding protein (CBP) are recruited to DNA and promote transcription by acetylating histones and TFs, and mediating interactions with general transcriptional machinery (Lin et al., 2001). p300 and CBP contain an acetyltransferase domain and several conserved regions through which

they interact with a diverse array TFs and other cellular proteins (Lin et al., 2001). Enhancer regions can be identified and annotated genome-wide using global maps of specific histone marks: histone 3 lysine 4 mono-methylation (H3K4me) and H3K27 acetylation (H3K27ac) define poised (H3K4me only) and active (H3K4me and H3K27ac) enhancer states (Creyghton et al., 2010; Heintzman et al., 2007). Genome-wide studies have found that p300 binding is highly enriched at active enhancers in cells, and that p300 ChIP-seq can be used to annotate active enhancers in resting and stimulated cells(Ghisletti et al., 2010; Heintzman et al., 2007; 2009). Furthermore, a primary substrate acetylated by p300/CBP in vivo is H3K27, and deletion of p300/CBP drastically reduces genome-wide H3K27ac levels (Jin et al., 2011; Pasini et al., 2010; Tie et al., 2009). Therefore, a picture emerges in which a key step in enhancer activation is the recruitment of p300/CBP to chromatin by sequence-specific TFs, leading to the deposition of the activation mark H3K27ac. However, the ability of recruited p300 to mark enhancers appears also to depend on other factors (Rada-Iglesias et al., 2011). Genomic analysis in human embryonic stem cells revealed two classes of enhancers marked by distinct chromatin features. 5,118 genomic loci resembled putative active enhancers and were marked by high p300, H3K4me1 and H3K27ac levels, but low (or no) H3K4me3 and H3K27me3. A second class of 2,287 p300-bound regions were marked by H3K4me1 but, unexpectedly, lacked H3K27ac, despite the presence of p300, and were enriched for H3K27me3 (a modification associated with polycomb silencing). Therefore, competition for modification of H3K27 residue likely exists between p300 and other recruited enzymes, and can affect the functional status of the enhancer.

Due to the role of p300/CBP as global activators that are recruited by diverse TFs to enhancer regions, we can view TF-p300/CBP complexes as representing key regulators of any cellular response. This idea is supported by genomic studies of the p300 binding landscape in LPS-stimulated mouse macrophages (Ghisletti et al., 2010). LPS-stimulation of mouse macrophages led to the induced recruitment of p300 to 2742 loci (i.e., 'peaks'). The majority (~85%) of these induced p300 peaks were at enhancers (i.e., > 2.5 kb from transcription start site of any gene), and motif analysis of these peaks revealed enrichment of binding motifs for the known key TF activators of the LPS response, such as NF-kB and interferon regulatory factors (IRFs). Therefore, TF-p300/CBP complexes appear to define key activator complexes of cellular responses that function to establish the active enhancer landscape.

1.3.2.2 NCOR/SMRT

The histone deacetylase (HDAC)-containing transcriptional repressor complexes NCOR (nuclear receptor co-repressor, also known as NCOR1) and SMRT (silencing mediator of retinoic acid and thyroid hormone receptor, also known as NCOR2) are repressor complexes recruited to DNA by a wide range of TFs (Alland et al., 1997; Heinzel et al., 1997; Nagy et al., 1997; Perissi et al., 2010). Histone acetylation is generally associated with relaxation of chromatin structure and increased transcriptional activity; therefore, HDAC-dependent removal of acetyl groups is associated with chromatin compaction and reduced gene expression (Perissi et al., 2010). The family of human HDACs contains 18 members that can be divided into four subclasses based on homology to yeast enzymes (Finkel et al., 2009; Martin et al., 2009; Verdin et al., 2003). The Class I HDACs

(HDAC1-3, HDAC8) are catalytic subunits of various multi-protein complexes responsible for transcriptional repression. For example, HDAC1 and HDAC2 have been associated with the SIN3A co-repressor complexes (Hassig et al., 1997; Zhang et al., 1997), the CoREST/RCOR1 complex (Hakimi et al., 2002; Humphrey et al., 2001; You et al., 2001), and the NURD complex (Tong et al., 1998) (Xue et al., 1998; Zhang et al., 1998). HDAC3 is a component of the homologous NCOR and SMRT repressor complexes. Along with HDAC3, core components of NCOR/SMRT complexes include TBL1/TBL1X (transducing β-like 1), TBLR1/TBL1XR1 (TBL-related 1), and GPS2 (G-protein-pathway suppressor 2) (Guenther et al., 2000; Li et al., 2000; Yoon et al., 2003).

Genome-wide assays of HDAC binding and chromatin modifications have revealed that HDACs, and by extension the associated co-repressor complexes, bind broadly on gene promoters and gene bodies (Wang et al., 2009). Unexpectedly, despite the role of de-acetylation in gene repression, HDACs are enriched not on silenced genes, but at poised (enriched for H3K4 methylation) and active promoters (Wang et al., 2009). Characterizing a diverse array of HDACs and chromatin features in human CD4+ T cells, Wang et al. showed that class I HDACs (HDAC1-3) and the class II HDAC6 were enriched at promoters of poised or active genes, and not silent genes. HDAC1 and HDAC3 were mainly detected in promoter regions, while HDAC2 and HDAC6 were elevated in promoters and gene bodies. HDAC binding was also correlated with mRNA expression levels and Pol II binding. These observations suggest that HDAC activity is actually associated with gene activation. One proposed function of HDACs in regulating

global gene expression is to remove the acetyl groups added by histone acetyltransferase (e.g., p300/CBP, PCAF), to allow a resetting of chromatin modifications after a round of gene activation, as has been suggested for the Hos2 gene (Wang et al., 2009). Despite these genome-wide trends, HDAC-containing complexes exhibit selective transcriptional control of target genes. While all HDACs were enriched at poised/active promoters, their binding does not completely overlap, suggesting that gene-specific recruitment of HDACs provides target gene specificity (Wang et al., 2009). Furthermore, HDAC inhibitors were found to effect the expression of only a minority of genes (Hanigan et al., 2018; Richon et al., 2000), and the gene-specific effects of HDAC inhibitors were cell-type dependent (Chambers et al., 2003; Hanigan et al., 2018).

To activate a gene in the presence of competing HAT and HDAC activities at promoters, there is a need for de-repression (i.e., removal of the HDAC co-repressor complexes). De-repression of the NCOR/SMRT complex can be achieved by diverse mechanisms, providing a range of regulatory options for signal and gene-specific activation. As has been described for the nuclear receptors (NRs), de-repression can occur via ligand-dependent switching of NR-bound NCOR/SMRT with a co-activator (reviewed in (Perissi and Rosenfeld, 2005)). Activation signals can also alter post-translation modifications (PTMs) on co-repressors (e.g., phosphorylation or ubiquitination), which can lead to their active removal by enzymes or translocation out of the nucleus (Perissi and Rosenfeld, 2005). Finally, NCOR/SMRT levels can be affected by altering the TF binding at promoters that recruit NCOR/SMRT.

Recruitment of NCOR/SMRT has been observed at inflammation-associated genes regulated by the TFs NF-kB, AP-1 and the ETS factor TEL/ETV6, demonstrating that TF-NCOR/SMRT complexes indicate key response regulators. Systems-level analysis of TF-CoF interactions using a mammalian two hybrid (M2H) assay have identified interactions of the NCOR and SMRT proteins with a diverse array of regulators, including NRs (e.g., ESR1, PPARG1, RXRA, THRB), inflammatory regulators (e.g., RELA, NFKB1, JUN), and early response regulators (e.g., FOS) (Ravasi et al., 2010). Thus, TF-NCOR/SMRT complexes include a diverse set of regulators, are critical for the global control of gene expression, and can be found at promoters of active and poised genes.

1.3.2.32 SWI/SNF

SWI/SNF enzymes are ATP-dependent remodelers of chromatin structure, making chromatin more accessible for TF binding (Imbalzano et al., 1994; Kwon et al., 1994) (Wang et al., 1996). The SWI/SNF enzymes are heterogeneous multi-subunit complexes; they contain two catalytic ATPase subunits, BRM/SMARCA2 and BRG1/SMARCA4, that are mutually exclusive, and additional accessory proteins known as BAFs (BRG-associated factors) (Ramirez-Carrozzi et al., 2009; Wang et al., 1996). SWI/SNF-mediated chromatin alterations can affect expression of constitutively active genes, but appear most closely linked to gene expression in response to developmental, environmental, or other extracellular signals (Wu et al., 2017).

The role of SWI/SNF as a stimulus-specific gene regulatory complex is illustrated by the variable requirement for SWI/SNF complexes observed for primary response genes in LPS-stimulated macrophages (Ramirez-Carrozzi et al., 2009; Smale et al., 2014). One class of primary response genes in macrophages is characterized by CpG-island promoters, which facilitate expression without a requirement for SWI/SNF-mediated chromatin remodeling (Ramirez-Carrozzi et al., 2009). SWI/SNF independence was observed at promoters of rapidly induce genes, such as *FOS, JUN, EGR1 and NFKBIA,* that are known to be activated in response to a broad range of stimuli (Hargreaves et al., 2009; Ramirez-Carrozzi et al., 2009). A second class of response genes is characterized by non-CpG-island promoters, which assemble into stable nucleosomes and exhibit an inhibitory chromatin environment. This second class of genes requires SWI/SNF-dependent chromatin remodeling and the binding of TFs that recruit SWI/SNF for inducible gene expression. Consistent with this additional recruitment for SWI/SNF recruitment, specific SWI/SNF-dependent genes, such as *IL12B, IL6* and *NOS2*, are known to be induced in macrophages by a selective set of stimuli (Hargreaves et al., 2009; Ramirez-Carrozzi et al., 2009). Some cellular stimuli (e.g., serum or TNFα) bias toward activation of SWI/SNF-independent CpG-island genes, whereas other stimuli (e.g., IFNβ) bias toward SWI/SNF-dependent non-CpG-island genes. This stimulus selectivity for activation SWI/SNF-dependent genes is likely mediated by differential activation of TFs that can recruit SWI/SNF complex to DNA, but may also involve differential PTMs of TFs (or SWI/SNF components) that can alter the TF-SWI/SNF interactions. Therefore, TF-SWI/SNF complexes that are present in different cell types,

and that are activated in response to different stimuli, define critical transcriptional regulatory complexes for specific classes of genes that require chromatin remodeling for their activation.

1.3.2.4 Set1/MLL COMPASS-like complexes

Genomic studies have identified H3K4 methylation status as a faithful predictor of enhancer and promoter status. In particular, H3K4me and H3K4me3 are predictors of enhancer (H3K4me high, H3K4me3 low) and promoter (H3K4me3 high) elements, respectively. H3K4 methylation activity is conserved from yeast to humans. The yeast *Saccharomyces cerevisiae* Set1/COMPASS complex was the first H3K4 methylase identified. In humans, SET1A, SET1B, and MLL1-MLL4 (mixed lineage leukemia) are the homologs of the yeast SET1, and are the core catalytic subunits of the methylase complexes (reviewed in (Shilatifard, 2012)). The human homologs have all been found in COMPASS-like complexes and are capable of establishing mono-, di- and tri-methylation marks on H3K4. In addition to the catalytic subunits, all the COMPASS complexes from yeast to humans share a number of common subunits, including ASH2, RBBP5, WDR5 and DPY30. Therefore, in vertebrates, recruitment of the six COMPASS-like complexes to enhancers and promoters to regulate H3K4 methylation is a key regulatory event in establishment and function of enhancers and promoters.

The SET1A/B complexes are the major regulators of global H3K4me3 state (Hallson et al., 2012; Wu et al., 2008; Yoon et al., 2003). Furthermore, the trimethylation activity is mediated by the WDR82 subunit of the SET1A/B complexes, which is absent from the

MLL1-4 complexes (Wu et al., 2008). Studies suggest that SET1A/B complexes are recruited directly by the Poll II machinery via the Ser5-Phospho CTD (Clouaire et al., 2012; Yoon et al., 2003), and to non-methylated CpG dinucleotides, which are present at many promoters, via the CFP1 subunit (Lee and Skalnik, 2007). In contrast, MLL complexes appear to have more varied effects, and play a more prominent role in H3K4 methylation at distal enhancers. For example, deletion of MLL4 disrupts levels of many enhancer features, such as H3K4me, H3K27ac, Mediator binding, RNA pol II binding and enhancer RNAs, although this does not require the MLL methylase catalytic activity (Dorighi et al., 2017; Lee et al., 2013), and MLL3/MLL4 knockout disrupts p300/CBP binding at enhancers (Wang et al., 2016) (Lai et al., 2017). MLL3/4 complexes appear to play a key role in the establishment and maintenance of enhancers. Much remains unclear about the loci-specific recruitment of MLL complexes to enhancers; however, MLL proteins have been reported to interact with cell-type-specific and signal-dependent TFs, including NRs, Pax proteins, and beta-catenin, suggesting that TF-MLL complexes may play important roles in shaping the enhancer landscape of the cell (Crump and Milne, 2019).

1.3.3 HT Methods to Assay TF-CoF Complexes

As reviewed above, TF-CoF complexes are critical to diverse aspects of the transcriptional control of genes. Furthermore, for a particular cellular state, the TFs present in the cell and the CoFs that they can interact with define the active transcriptional regulators in the cell. Therefore, the ability to characterize not just the TF

proteins present in the nucleus, but the TF-CoF complexes present in the nucleus would provide a powerful approach to understand how gene regulation is controlled in response to stimuli and changes in cell state.

A number of medium- to high-throughput methods exists to monitor TF levels in a cell. Commercial products based on enzyme-linked immunosorbent assays (ELISAs) are available, such as TransAM™ or Luminex®200™ (Active Motif), and employ a combined ELISA-type detection with DNA oligonucleotide-binding selection strategy for the detection of TFs from cell lysates; however, currently these approaches only scale up to allow detection of ~10-15 TFs. The 'TF Activation Profiling Plate Array I' assay available from Signosis does not rely on an antibody-based detection step and can monitor a panel of 48 known TFs. This higher-throughput method from Signosis involves incubating cell lysates with a panel of DNA oligonucleotides containing consensus binding sites for each TF. Binding to these select DNA sequences is monitored and used to identify the TFs present in the sample. Alternate methods for detecting TFs also involve enrichment of proteins binding to select DNA oligonucleotides and detection by mass spectrometry (Simicevic et al., 2013) (Mittler et al., 2008) or by immunofluorescence (Arbab et al., 2013). More recently, the ATI (active TF identification) approach has been developed that can screen the activity (interpreted as the ability of TFs to bind DNA) of hundreds of TFs present in a cell nucleus using a HT electrophoretic mobility shift assay (EMSA)-based method followed by next-generation

sequencing (Wei et al., 2018). However, none of these available methods is designed to characterize the diversity of TF-CoF complexes in the cell.

The two current high-throughput (HT) approaches to identify TF-CoF interactions are (1) yeast or mammalian two-hybrid (Y2H (Fields and Song, 1989) or M2H (Ravasi et al., 2010)) assays to identify direct protein-protein interactions between protein pairs, and (2) immunoprecipitation (IP) followed by mass spectrometry to identify co-precipitating proteins (Wierer and Mann, 2016). The Y2H approach has been used to map pairwise protein interactions of human proteins (Rolland et al., 2014; Rual et al., 2005; Yu et al., 2008), but is limited to binary interactions and does not capture cell-specific regulation of these interactions. The M2H approach has been used to examine TF-TF interactions in a more biologically relevant mammalian cell context (Ravasi et al., 2010), but the M2H approach is similarly limited to binary interactions, and remains labor-intensive to conduct at a HT level in different cell conditions. As the name implies, IP-mass spectrometry approaches use mass spectrometry to identify proteins co-precipitating with a target protein or oligonucleotide from cell extracts. The IP-mass spectroscopy approach has been used to study protein-protein associations, chromatin-protein associations, and DNA-protein associations (Wierer and Mann, 2016). DNA oligomer-based precipitation followed by protein identification provides a way to identify TF and CoFs that associate with a particular DNA sequence (Hubner et al., 2015). CoF-based precipitation followed by protein identification provides a way to identify interacting TFs, but it does not explicitly assay DNA-bound complexes (Wierer and Mann, 2016). The approaches

described above all provide large-scale methods to monitor complexes in a native context, but generally require high cell numbers and large amounts of starting material, although new approaches are reducing the material needed (Hubner et al., 2015). Nevertheless, none of these currently available approaches provides a rapid, HT approach to identify the many TFs that recruit particular CoFs to DNA in a cellular condition of interest.

1.3.4 Nuclear Extract Protein-Binding Microarrays (nextPBMs)

Our lab has recently developed the nextPBM technology to characterize DNA-bound TF complexes from cell nuclear extracts (Mohaghegh et al., 2019). 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 traditional PBM methodology by using endogenous proteins from nuclear extracts in place of bacterially expressed or HEK 293-cell over-expressed proteins. Therefore, nextPBM enables the examination of endogenous levels of natural DNA-binding complexes (e.g., TF-CoFs) present in cells. In nextPBM experiments, DNA-bound TFs are probed for and identified using primary antibodies, followed by a fluorescently labeled secondary antibody. Critically, the nextPBM approach is performed using total nuclear lysates, which includes all TF and CoF protein present in the cell nucleus, which presented the possibility of measuring TF-CoF complexes using the same platform.

1.3.5 CoRec (Cofactor Recruitment) Approach to Monitor TF-Cofactor Complexes in Cells

In this thesis, we describe the CoRec approach for the HT characterization of TF-CoF complexes present in a cell nucleus (Chapter 3). CoRec is an extension of the nextPBM assay (Mohaghegh et al., 2019) in which CoF recruitment to DNA is assayed, rather than direct TF-DNA binding. Using antibodies against general CoF proteins that interact with diverse TFs (i.e., p300/CBP, NCOR/SMRT subunits, MLL complex subunits, etc.) we can monitor the binding of many possible TF-CoF complexes in a single assay. We demonstrate the ability of the CoRec assay to characterize TF-CoF complexes in resting and stimulated human macrophages and T cells, demonstrating the tremendous utility of this approach for the HT characterization of key transcriptional regulatory complexes in cells.

CHAPTER TWO: Comprehensive Study of Nuclear Receptor DNA Binding Provides a Revised Framework for Understanding Receptor Specificity

Note: Portions of this chapter were published previously in (Penvose et al., 2019), for which Ashely Penvose (AP) and Jessica Keenan (JLK) were co-first authors. The concept for the custom PBM was developed by JLK and AP with input from Trevor Siggers (TS). JLK created the probe sequences for the array (see Methods for more details). Computational assessments of preliminary array designs were performed by JLK (see Discussion). All genomic analyses were performed by JLK. Experiments (including PBM experiments, reporter assays, and EMSAs) were performed by AP. Individual contributions for the work underlying each figure are noted at the end of each figure legend. The manuscript was co-written by JLK, AP, and TS. All supplementary data can be found in the online version of this article.

2.1 Abstract

The type II nuclear receptors (NRs) function as heterodimeric transcription factors with the retinoid X receptor (RXR) to regulate diverse biological processes in response to endogenous ligands and therapeutic drugs. DNA-binding specificity has been proposed as a primary mechanism for NR gene regulatory specificity. Here we use protein-binding microarrays (PBMs) to comprehensively analyze the DNA binding of 12 NR:RXRα dimers. We find more promiscuous NR-DNA binding than has been reported, challenging the view that NR binding specificity is defined by half-site spacing. We show that NRs bind DNA using two distinct modes, explaining widespread NR binding to half-

sites in vivo. Finally, we show that the current models of NR specificity better reflect binding-site activity rather than binding-site affinity. Our rich dataset and revised NR binding models provide a framework for understanding NR regulatory specificity and will facilitate more accurate analyses of genomic datasets.

2.2 Introduction

The type II nuclear receptors (hereafter simply NRs) are ligand-activated transcription factors (TFs) that control diverse cellular processes including development, metabolism, and inflammation (de Aguiar Vallim et al., 2013; Evans and Mangelsdorf, 2014). NRs include peroxisome-proliferator activated receptor (PPAR), liver x receptor (LXR), retinoic acid receptor (RAR), farnesoid x receptor (FXR), pregnane x receptor (PXR), thyroid hormone receptor (THR), and vitamin D receptor (VDR) (Evans and Mangelsdorf, 2014; Kliewer et al., 1999). NRs function as heterodimers with the common partner, the retinoid x receptor (RXR). Individual NR heterodimers can regulate distinct gene programs (Calkin and Tontonoz, 2012); however, the current models of NR-DNA binding specificity are insufficient to explain NR-specific gene regulation.

NRs bind the sequence 5'-RGKTCA-3' organized as direct repeats with a variable length spacer of 0–5 base pairs (bp) (DR0-DR5, Fig. 2.1a) (Claessens and Gewirth, 2004; Cotnoir-White et al., 2011; Weikum et al., 2018). Current models propose that DR spacer length is a key determinant of DNA-binding specificity for NRs (Cotnoir-White et al., 2011; Evans and Mangelsdorf, 2014; Kurokawa et al., 1993; Mader et al., 1993;

Perlmann et al., 1993). For example, PPAR:RXR dimers prefer binding to DR1 elements, whereas LXR:RXR dimers prefer DR4 elements (Fig. 2.1b). However, there are more NRs than available spacer lengths; therefore, either DRs are bound by multiple NRs, which presents a problem for achieving NR-specific gene activation, or there are additional determinants of NR-binding specificity beyond DR spacer length.

Differences in DNA-binding specificity for each NR would provide a mechanism for NRs to regulate distinct target genes in vivo. Genome-wide chromatin immunoprecipitation followed by sequencing (ChIP-seq) studies have confirmed known NR preferences for particular DR spacer lengths, and have reinforced the connection between in vitro and in vivo binding (Boergesen et al., 2012; Chatagnon et al., 2015; Lefterova et al., 2010; Rastinejad et al., 2013; Savic et al., 2016; Soccio et al., 2015; Zhan et al., 2014). However, these studies have also revealed limitations to current models of NR-DNA binding. For example, PPARγ and LXRα regulate distinct yet overlapping gene programs but do not share a DR element to explain their many common genomic targets (Boergesen et al., 2012; Savic et al., 2016). Additionally, many genomic regions that are bound in vivo lack an identifiable binding site for the NR being investigated (e.g., 90–96% for PPARγ and LXR) (Boergesen et al., 2012). Together, these observations suggest that current models of NR-DNA-binding specificity are incomplete.

To address the need for revised models of NR binding, we use protein-binding microarrays (PBMs) to compare the binding of 12 NR:RXRα dimers to thousands of

DNA sequences. To examine DR spacer preferences, we assay NR binding at all spacer lengths (DR0-DR5). We identify both NR-shared and NR-specific binding features in our dataset, and discuss implications for NR-signaling specificity. By integrating PBM and ChIP-seq datasets, we examine the relationship between in vitro and in vivo binding. We address the role of activity versus affinity in current models of NR specificity by integrating PBM data with reporter gene experiments. Our results demonstrate the limitations of DR spacer length for defining NR specificity and of DNA binding affinity for predicting functional binding events.

2.3 Results

2.3.1 Characterizing NR Heterodimer Binding with PBMs

We used PBMs to characterize the DNA binding of 12 distinct RXR heterodimers (hereafter NRs). PBMs are double-stranded DNA microarrays that enable the high-throughput study of protein-DNA binding (Berger et al., 2006). To characterize both DNA-base and DR-spacing preferences, we measured NR binding to over 1600 unique sequences at each of six DR spacer lengths (DR0-DR5). For each DR spacer length, we measured NR binding to 24 starting sequences, which we refer to as seed sequences (Fig. 2.1c). Seed sequences were generated by combining different half-site sequences exhibiting a range of degeneracy from the consensus 5'-RGKTCA-3'. Most seed sequences contain two distinct half-site sequences. To assay NR binding specificity for each seed sequence we also measured binding to all possible single-nucleotide variants (SNVs), with each SNV included as a separate probe on the PBM (Fig. 2.1c). This SNV-

based approach allows us to generate a binding logo (i.e., energy matrix or position-weight matrix (PWM)) for each individual seed sequence by measuring the impact on binding caused by perturbation at each base position (Fig. 2.1c, Methods). To capture binding preferences for DR spacer and flank sequences, we included SNVs across the spacer sequence and for the five nucleotides upstream and downstream of the DR. Using this comprehensive SNV-type PBM design, we characterized the DNA-base and DR-spacing preferences of the NRs.

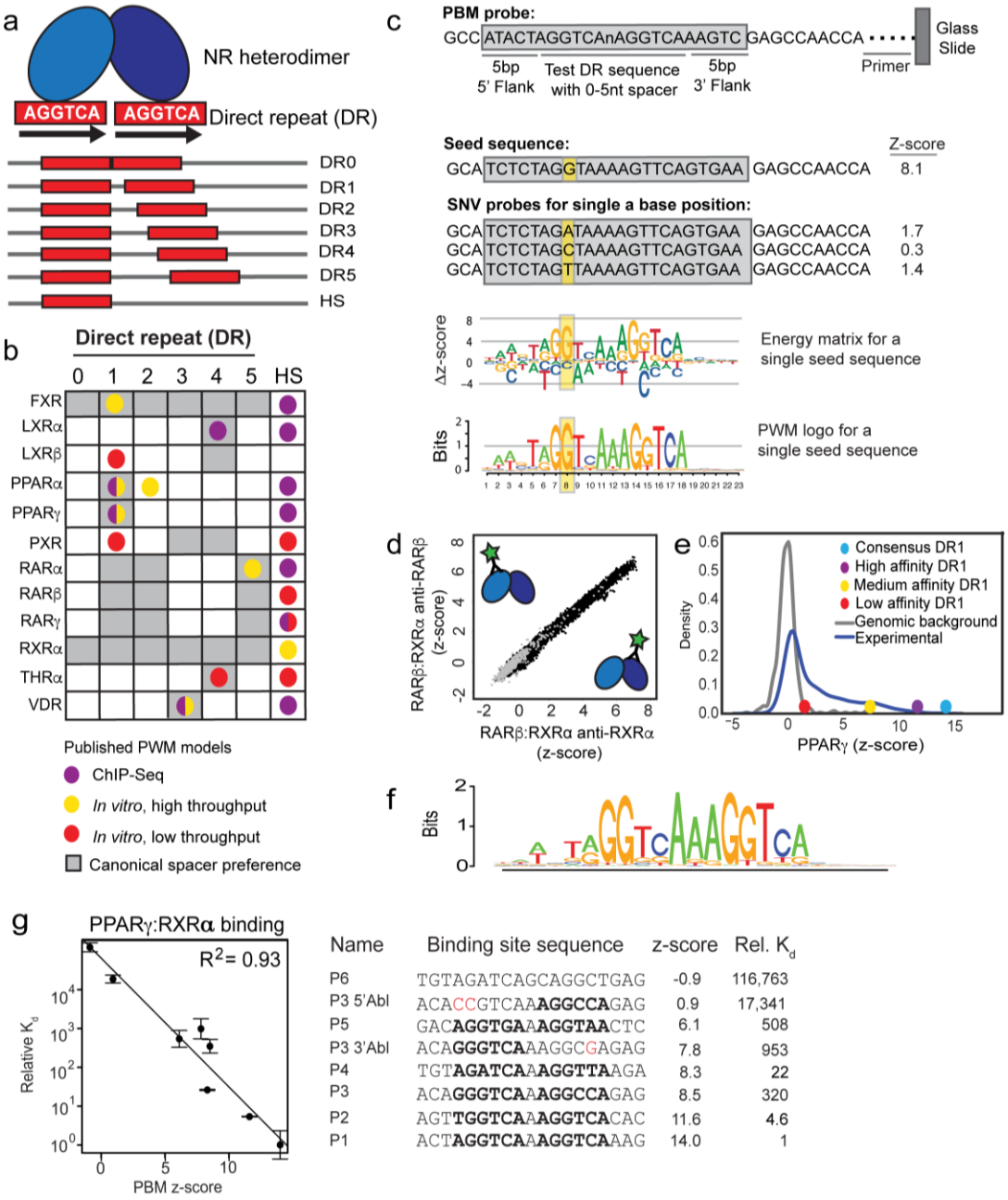


Figure 2.1: Characterizing NR-DNA binding with PBMs
a Schematic of spacer preferences for NRs to direct repeats (DRs) and half-sites (HS). **b** Canonical spacer preferences of NRs indicate preferred spacer lengths from the literature (Supplementary Data 2 and 3). Published PWM models are shown in colored dots that indicate the methodology used to derive the model (Supplementary Data 3). **c** Schematic of PBM probes, SNV probe organization and SNV-based motif generation for a single seed sequence. **d** Scatter plot of z-scores for RARβ:RXRα experiments detected with antibodies against each heterodimer partner. Dots represent average over ~5 replicates for all 10,728 unique SNV probes (black dots) and 500 background probes (gray dots) **e** PBM replicate averaged z-score

distributions for PPARγ:RXRα to all SNV probes. Z-scores for consensus DR1 and reported functional binding sites are highlighted (Supplementary Data 2) (Juge-Aubry et al., 1997). **f** DR1 DNA-binding logo for PPARγ:RXRα generated from all DR1 full-site models from the PBM experiments. **g** Comparison of PPARγ:RXRα PBM z-scores and competition EMSA-determined relative K_d measurements for binding sites spanning a wide affinity range. Relative K_d values are normalized to the highest affinity sequence (P1) and represent mean over two independent experiments (error bars = STDEV). Identifiable DR half-sites in each binding sequence are shown in bold. Mutations introduced to ablate the 5' half-site of P3 (P3 5'Abl) or the 3' half-site of P3 (P3 3'Abl) are shown in red.
Contributions: a,c-f the PBM was design by JLK and AP as described in the note at the beginning of the chapter. PBM experiments were performed by AP **b** was created by AP with help from JLK. **g** EMSA experiments were performed by AP.

PBM experiments for NR heterodimers (NR:RXRα) were performed by combining purified RXRα with purified samples of each partner NR. Hereafter, we refer to NR:RXRα heterodimers simply by the NR partner, and RXRα:RXRα homodimers as RXRα, unless otherwise stated. Most NRs do not bind DNA with high affinity as homodimers; therefore, proteins were combined at a 3:1 NR:RXRα ratio to force RXRα heterodimer formation (exceptions indicated in Supplementary Data 1). To ensure heterodimer binding, we required that the binding results agreed when performed using antibodies for both RXRα and the non-RXRα partner. Binding profiles using separate antibodies showed strong correlation, demonstrating that both protein partners were bound to each DNA probe at similar levels (Fig. 2.1d, R^2 of antibody replicates in Supplementary Data 1). Binding of homodimers were not correlated with each other, nor with the heterodimers (Supplementary Fig. 2.1), further demonstrating heterodimer binding. To quantify binding specificity, PBM fluorescence values were converted into z-scores using a set of 500 random genomic background sequences (Fig. 2.1e). Validated PPARγ binding sites score significantly above background, down to a z-score of 1.5 (Fig. 2.1e). We set a more stringent z-score cutoff of 3.0 to define the affinity cutoff for functional binding sites. A DR1 DNA binding logo generated for PPARγ agrees well

with known logos from ChIP-seq (Fig. 2.1f), demonstrating the sensitivity of our assay. To validate our PBM results with an orthogonal approach, we used competition electrophoretic-mobility shift assays (EMSAs) to measure the relative binding affinity of PPARγ:RXRα to DNA sequences bound over a wide range of PBM z-scores (Fig. 2.1g, Supplementary Fig. 2.2). We find strong agreement between the relative binding affinities derived using both approaches ($R^2 = 0.93$). Our protein samples were produced in bacterial or insect cells; however, our ability to capture known NR-binding specificity suggests our data reflect native mammalian dimer-binding specificity. These results demonstrate that our PBMs accurately capture sequence-specific binding of NR heterodimers.

2.3.2 NRs Bind Promiscuously to Most DR Spacings

To understand NR-signaling specificity, studies have examined the DNA-binding differences between NRs (summarized in Fig. 2.1b, Supplementary Data 2 and Supplementary 3 from (Penvose et al., 2019)) (Cotnoir-White et al., 2011; Evans and Mangelsdorf, 2014; Mader et al., 1993; Näär et al., 1991). A prevailing view is that NRs are distinguished by their preference for DR sites with specific half-site spacing) (Cotnoir-White et al., 2011; Evans and Mangelsdorf, 2014; Mader et al., 1993; Näär et al., 1991; Perlmann et al., 1993; Zechel et al., 1994); however, individual NRs are functional on DR sites with various spacings (Katz et al., 1995; Kurokawa et al., 1993). Therefore, for each NR we examined which DR spacings were bound with sufficient affinity such that they might be functional in vivo.

To visualize the NR-binding landscape, we generated a DNA-binding logo from high-scoring seeds at each DR spacing (Fig. 2.2). Strikingly, for all NRs we were able generate DNA-binding logos at nearly every DR spacing, demonstrating broader binding preferences than previously reported. Comparing our logos with published DR binding preferences (Fig. 2.1b), we find high-affinity binding for many NRs at new DR spacings. The binding logos for all NRs exhibit the canonical 5'-RGKTCA-3' sequence preferences in each half-site and agree with base preferences reported by other methods (Isakova et al., 2017). The logo similarity demonstrates broad conservation in NR-binding specificity; however, NR-specific preferences are also present. For example, PPARγ prefers an AT-rich sequence 5' of the first half-site of a DR (Juge-Aubry et al., 1997) and our PPARγ logo shows this extended footprint (Figs. 2.1f, 2.2). Overall, our data reveal that all NR heterodimers can bind to sites with variable DR spacings and with highly overlapping base specificities.

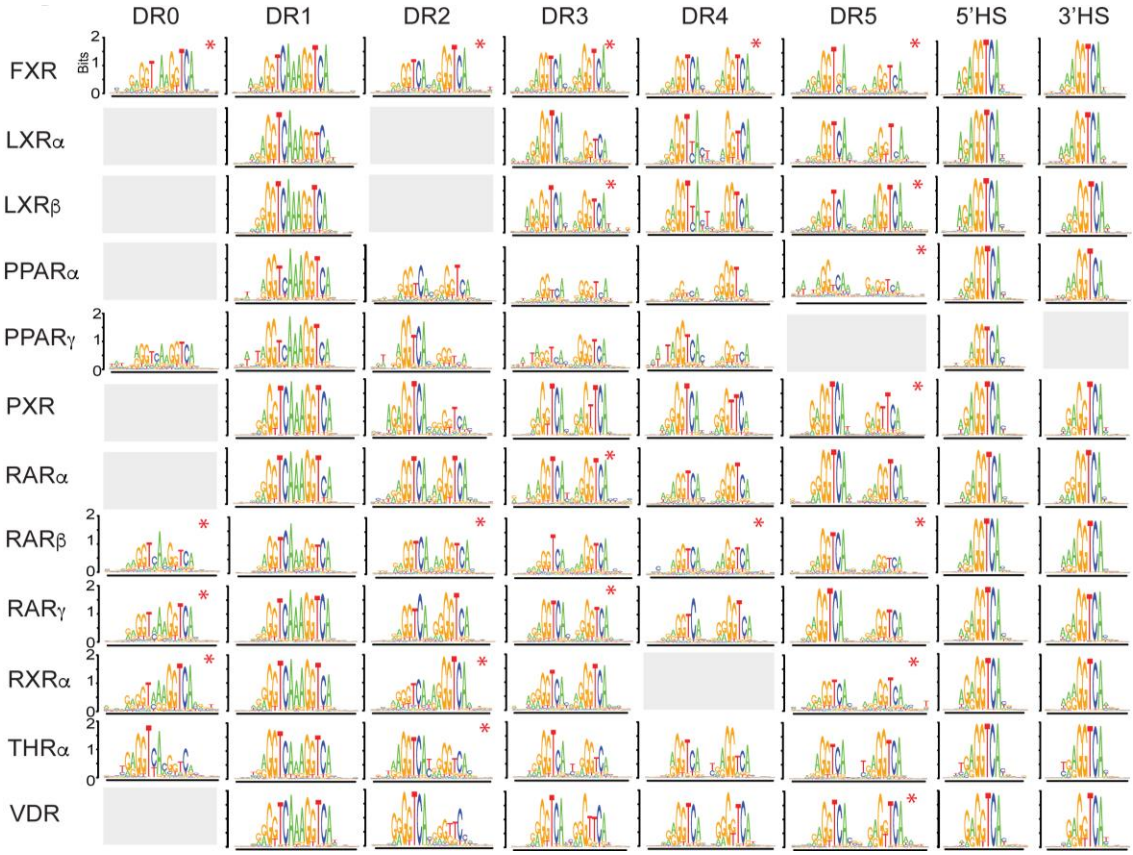


Figure 2.2: NR-binding specificity and DR preferences
PBM-derived DNA-binding logos for 12 NRs at all examined DR spacer lengths. Half-site logos identified for each NR on either the 5' half-site (5'HS) or 3' half-site (3'HS) are shown. Logos based upon a single significant (z-score > 3.0) seed sequence are indicated (∗).
Contributions: analyses for this figure were performed by AP and TS with help from JLK.

2.3.3 All Type II NRs Can Bind DNA Using a Half-Site Mode

We find that all NRs can bind with high affinity to half-sites (Fig. 2.2, final two columns). For all NRs, we obtain both 5'- and 3'-half-site logos, with the exception of PPARγ, for which we only find clear 5'-half-site binding (Fig. 2.2). Half-site logos indicate that NR binding is only perturbed by SNVs in one half-site of a DR. To illustrate, we show the impact of SNVs on LXRα dimer binding to seed sequences with different binding modes (Fig. 2.3a-c, Supplementary Fig. 2.3). Critically, our data agree

for PBMs probed with antibodies against either dimer member; therefore, half-sites are bound by NR heterodimers and are not a result of monomer binding. The presence of both full-site and half-site logos suggests that NRs can engage with DNA in two binding modes: (1) full-site mode where the NR engages with both half-sites and (2) half-site mode where the NR engages with a single half-site (either 5' or 3') (Fig. 2.3a).

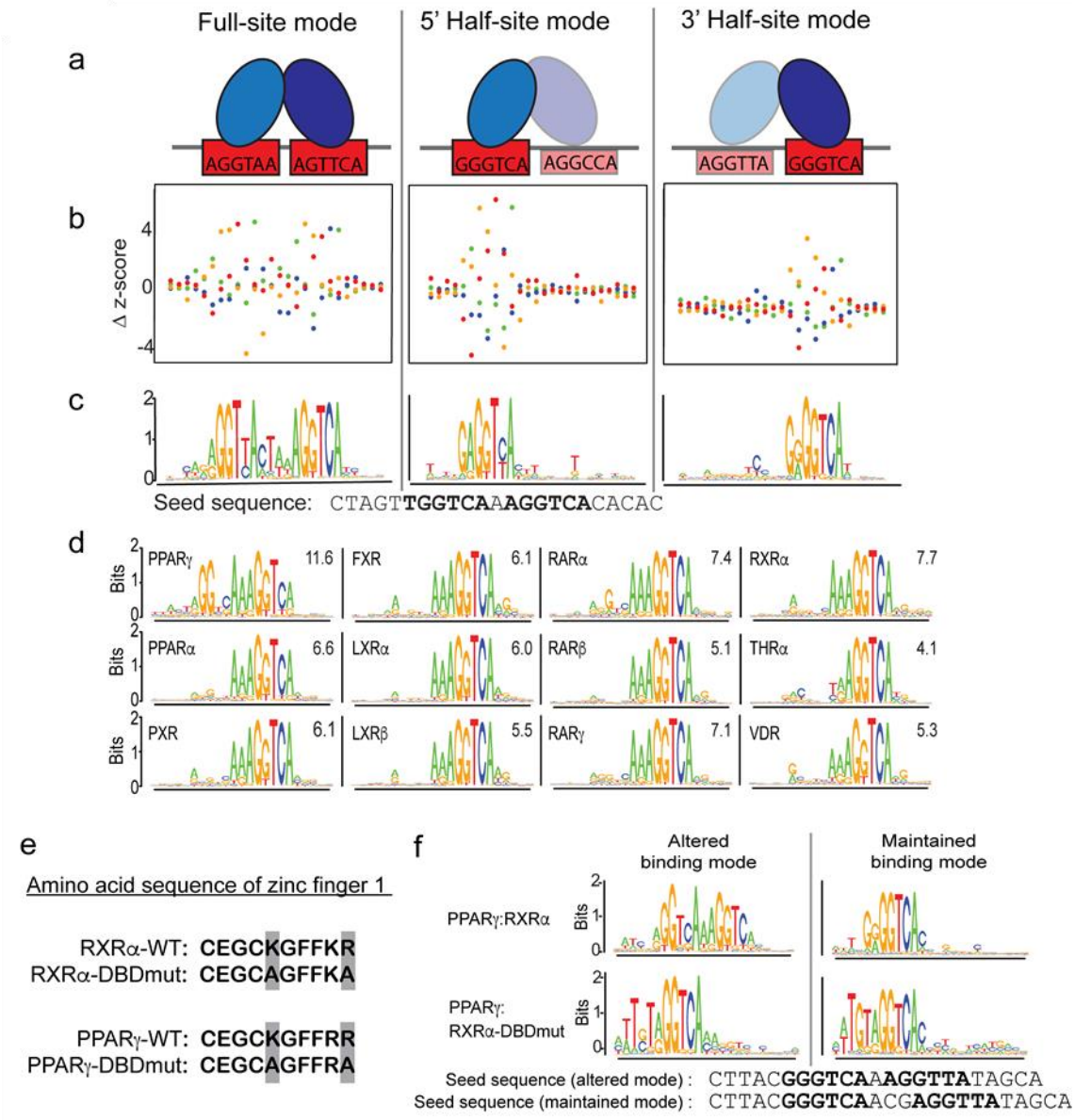


Figure 2.3: NR half-site binding mode
a Schematic of NR full-site or half-site binding modes. **b**, **c** For three seed sequences bound with different modes, the impact of SNVs on LXRα heterodimer binding and the corresponding DNA-binding logos are shown. Binding perturbation for each SNV is shown as a Δz-score from the median z-score of all four base variants at each position. Colors correspond to base identity indicated in logos below. **d** DNA-binding logos for all 12 NRs generated for the single DR1 seed sequence shown. **e** Amino acid sequence of zinc finger 1 for the wild-type RXRα, RXRα DNA-binding domain mutant, wild-type PPARγ, and the PPARγ DNA-binding domain mutant. Altered amino acids are highlighted in gray. **f** DNA-binding logos for individual seed sequences (shown) for which the binding mode was either altered (left) or maintained (right) for the PPARγ:RXRα-DNA binding domain mutant.
Contributions: analyses for this figure were performed by AP and TS. DBD mutants were made by AP.

To ensure that the widespread half-site binding was not a result of our methodology, we performed several analyses. First, we tested whether half-site binding was due to the orientation of the NR-binding site within the PBM probe with respect to the microarray slide. We find that regardless of orientation of the probe, binding mode is maintained (Supplementary Fig. 2.4). Second, we performed PBMs at successively lower concentrations to test whether half-site binding is affected by protein concentration and find nearly identical DNA binding logos at all concentrations (Supplementary Fig. 2.5). Finally, we used EMSA experiments to test the impact of base mutations on a DNA site bound in half-site mode (Fig. 2.1g, sequences P3, P3 5'-Abl, P3 3'-Abl). Critically, the 5' half-site mode of PPARγ:RXRα determined by PBM is corroborated by EMSA experiments (i.e., 5' half-site ablation greatly reduced binding whereas 3' half-site ablation only modestly affected binding) (Fig. 2.1g, Supplementary Fig. 2.2). These results demonstrate that PBM-derived binding modes accurately represent native NR-binding modes.

NRs are known to bind half-sites (Fig. 2.1b), though half-sites have primarily been identified in ChIP-seq data and not through direct binding assays. Our analysis clarifies

that NR heterodimers can bind half-sites, and can engage in a half-site mode even on canonical DR sites composed of two good half-sites (i.e., both half-sites score well using PWMs). For example, logos generated for a near-consensus DR1 seed sequence that scores highly by DR1 PWMs reveal both full- and half-site binding modes (Fig. 2.3d). While all NRs bind this site with high-affinity (z-scores are shown), only PPARγ binds in a full-site mode, while other NRs bind with nearly identical half-site modes. This shows that binding mode can vary for different NRs on the same DNA site, and that throughout the genome NR-binding to DR sites may in fact be mediated through a half-site binding mode.

2.3.4 Role of Monomers in Half-Site Binding

To examine the contribution of each protein within an NR heterodimer to DNA binding, we created DNA-binding domain mutants (DBDmut) of RXRα and PPARγ. Two residues within zinc finger 1 of RXRα and PPARγ that make base-specific contact with DNA were mutated to alanines (K156A and R161A; and K132A and R137A, respectively, Fig. 2.3e) (Chandra et al., 2008). Binding of PPARγ:RXRα-DBDmut is highly correlated using either anti-RXRα or anti-PPARγ antibodies (R^2 of antibody replicates given in Supplementary Data 1), showing that all DNA sites are bound by the mutant as a heterodimer. For PPARγ-DBDmut:RXRα, PBMs performed using an anti-RXR antibody are dominated by RXR homodimer signal, therefore binding of PPARγ-DBDmut:RXRα was determined using only the anti-PPARγ antibody. RXRα homodimer binding was not observed in wild-type heterodimer experiments (see above). All DBD mutant proteins were produced by IVT and PBM data for IVT-produced wild-type dimers

agree with experiments using purified proteins, demonstrating that IVT proteins form heterodimers and function in DNA-binding assays similarly to purified proteins (model curations can be found in Supplementary Data 4 of (Penvose et al., 2019).

To confirm that these mutations abrogated DNA interactions, we examined the binding of mutant homodimers using PBMs. The mutant RXRα (RXRα-DBDmut) bound no sequences with z-score > 3.0 (as compared to a max z-score of 7.0 for PPARγ:RXRα-DBDmut described below), demonstrating an abrogation of sequence-specific DNA binding. The mutant PPARγ (PPARγ-DBDmut) showed binding with z-score > 3.0 to only five seed sequences. Previous experiments have shown residual DNA-binding activity for PPARγ DBD mutants (Temple et al., 2005); therefore, we chose to disregard these five sequences from further analysis of the PPARγ-DBDmut:RXRα heterodimer experiments.

We first examined mutant heterodimer binding to sequences that PPARγ:RXRα binds in full-site mode. As expected, binding in full-site mode was almost completely abrogated for the PPARγ:RXRα-DBDmut (38/39 full-sites were lost). Of these sites, 40% (15/38) were now bound in the 5' half-site mode (e.g., Fig. 2.3f , curation of modes can also be found in Supplementary Data 5 of (Penvose et al., 2019), demonstrating an altered binding mode for the PPARγ:RXRα-DBDmut heterodimer. The remaining 60% (23/38) of these sites were bound with low affinity by PPARγ:RXRα-DBDmut, and scored below our z-score threshold for modeling interactions. The reciprocal mutant experiment with

PPARγ-DBDmut:RXRα showed a complete loss of binding (i.e., z-score < 3.0) to nearly all of the full-sites (35/36, note that we have disregarded three sequences in this category as described above, model curations can also be found in Supplementary Data 5 of (Penvose et al., 2019). These results demonstrate that DNA must be engaged by both dimer partners in order for PPARγ:RXRα to utilize a full-site binding mode, and shows that half-site binding can occur when only one partner can bind DNA.

Next, we examined which partner of the wild-type PPARγ:RXRα dimer engages with DNA when binding in a half-site mode. Of the 34 sequences that PPARγ:RXRα bound in a half-site mode, 53% (18/34) remained bound in half-site mode by PPARγ:RXRα-DBDmut, demonstrating that for these sequences PPARγ is making base-specific contacts with the DNA and can tolerate loss of base-specific DNA contacts mediated by RXRα (Fig. 2.3f). For the remaining 47% (16/34) of sequences bound by PPARγ:RXRα in a half-site mode, the mutant dimer binding was too low affinity to model (i.e., z-score < 3.0). Interestingly, PPARγ-DBDmut:RXRα, showed a loss of binding to 82% (29/32, note two sequences in this category were disregarded as above) of the half-site sequences. These results demonstrate that a single partner of an NR heterodimer can mediate half-site binding; however, for other sites, mutation of either NR partner can lead to loss of heterodimer binding. The strong impact of mutations to either member of the heterodimer may be attributable to the ability of either partner to engage with the half-site, or to a contribution in binding energy through non-specific interactions from the non-engaged partner, which were abrogated by the mutations we made.

2.3.5 NR Spacer Preferences Do Not Define High-affinity Binding

Previous studies have examined the impact of DR spacer length on NR binding (Cotnoir-White et al., 2011; Evans and Mangelsdorf, 2014; Kurokawa et al., 1993; Perlmann et al., 1993); however, our results show that NRs can bind in a half-site mode even on DR sites, which complicates the interpretation of these experiments. SNV binding models are advantageous as they allow examination of NR-binding mode on each sequence, thus facilitating a more rigorous assessment of NR spacer preferences. We analyzed the NR-binding landscape to all 24 seed sequences at each DR spacing and used the resulting binding logos to annotate whether each sequence was bound in a full-site or half-site mode (Fig. 2.4).

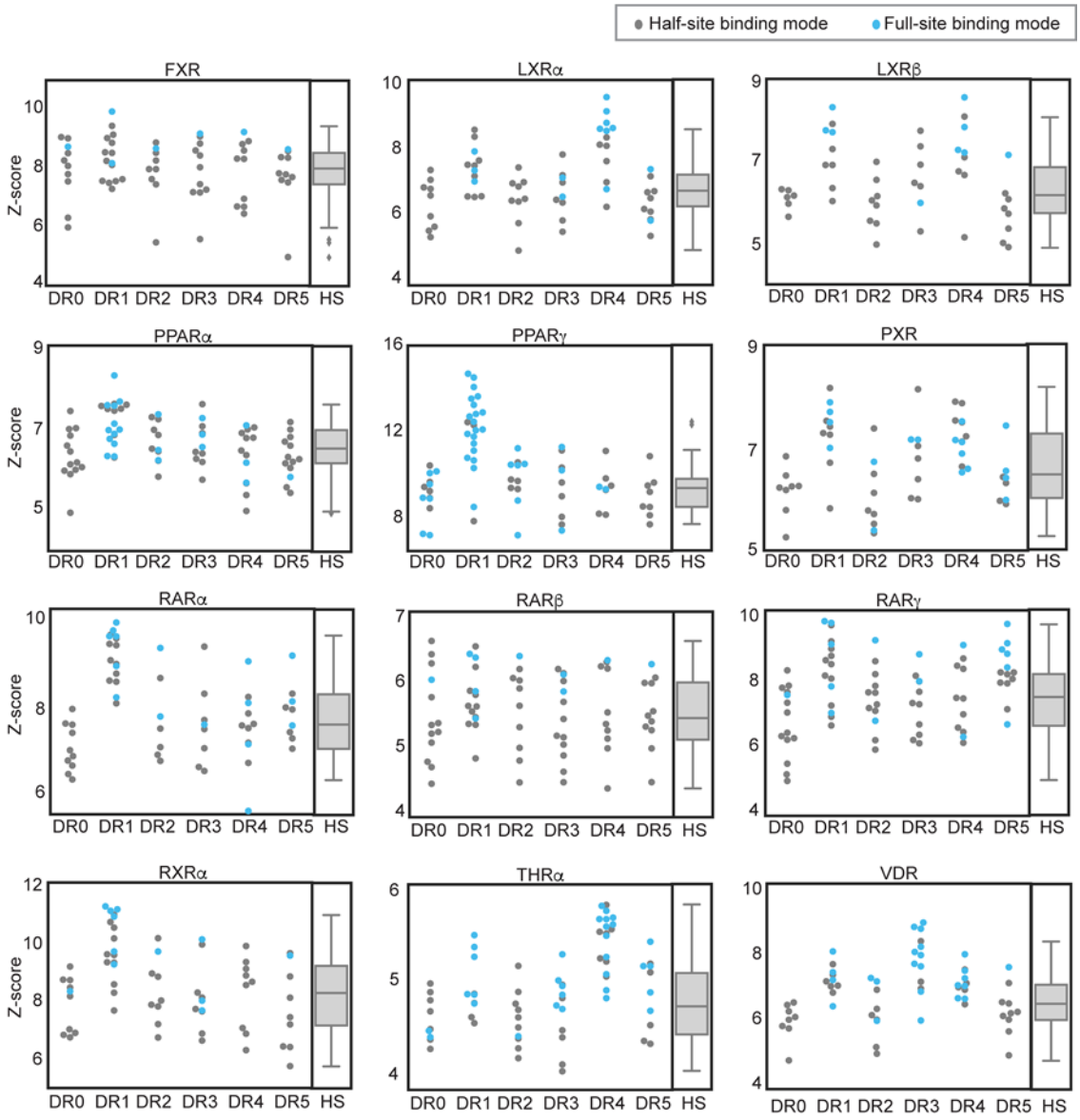


Figure 2.4: NR-binding affinity and mode for sequences at each DR spacer length
At each spacer length, the replicate averaged z-score of the highest scoring SNV for each seed sequence is shown; seed sequences with z-score < 3 are not represented. Colors indicate binding mode for each seed sequence. For each NR, box plots show the z-score distributions for all sites that are bound in half-site modes across all direct repeat spacer lengths (the aggregate of all gray dots). Center line: median; box limits: upper and lower quartiles; whiskers: last datum within 1.5x interquartile range.
Contributions: analyses for this figure were performed by AP.

In contrast to the prevailing view of NR spacer preferences (Cotnoir-White et al., 2011; Evans and Mangelsdorf, 2014; Näär et al., 1991; Rastinejad et al., 1995), we observed

that NRs can bind with high affinity to DRs at all spacer lengths (Fig. 2.4). For most NRs, high-affinity binding to many DR spacer lengths is predominantly mediated via a half-site binding mode (Fig. 2.4 gray dots). Despite this promiscuous NR binding, our results recapitulate literature-reported NR spacer preferences, which are demonstrated by an enrichment of full-site binding mode and higher z-scores for specific DR spacer lengths (Fig. 2.4 blue dots). For example, PPARγ engages with DR1 sequences almost entirely via a full-site binding mode. Similar observations corroborate previously described DR-spacing preferences, for example LXRs (DR1 & DR4), THRα (DR4), and VDR (DR3) (see Fig. 2.1b). However, for most NRs the increase in binding affinity to certain DR spacers is more modest than observed for PPARγ, suggesting that spacer preferences do not define the DNA binding landscape of each NR. In fact, for some NRs the canonical DR-spacing preferences appear primarily as enrichment in full-site binding mode, but not a large increase in binding affinity. For example, PPARα preferentially engages with DR1 sites in a full-site binding mode but only binds with moderately higher z-scores to these sites. Our results reveal a complicated NR-DNA binding landscape in which DR spacer preferences contribute to altered NR-binding modes and binding affinity, but which do not strongly define the landscape of all possible high-affinity binding.

2.3.6 Diverse Mechanisms Contribute to NR-DNA Binding

Despite broad similarities seen in binding logos (Fig.2.2), our dataset reveals that NR-binding differences result from multiple mechanisms: DR-spacing preferences, DNA-base preferences, and DNA-binding-mode differences. To illustrate the roles of spacing

preferences and binding modes, we compared the binding of PPARγ and LXRα to DR1 and DR4 sites and observe both NR-shared and NR-specific binding sites (Fig. 2.5a). The LXRα preference for DR4 sites and PPARγ preference for DR1 sites are demonstrated as biases in the z-score distributions. However, as we see high-affinity binding of PPARγ to DR4 sites and LXRα to DR1 sites, the aforementioned preferences do not explain all high-affinity binding. To explicitly test the impact of DR spacing, we examined binding to pairs of seed sequences that differ only in their spacer length (e.g., Fig. 2.5a, sequences DR1.1 and DR4.1). Critically, we examined the DNA-binding mode for each interaction using the DNA-binding logos generated for each seed sequence (Fig. 2.5b). For PPARγ, DR4 sites are bound with lower affinity than corresponding DR1 variants; however, DR4.1 is still bound with high affinity via a half-site binding mode (Fig. 2.5a). In contrast, when LXRα binds via a full-site mode the DR4 variant is bound with higher affinity (DR1.1 vs DR4.1), but when binding via a half-site mode the DR4 variant is bound with lower affinity (DR1.2 vs DR4.2) (Fig. 2.5a). Therefore, both NRs can bind the same sequence with high affinity, but may utilize distinct binding modes. Taken together, these results demonstrate that both spacer preference and binding mode contribute to binding specificity.

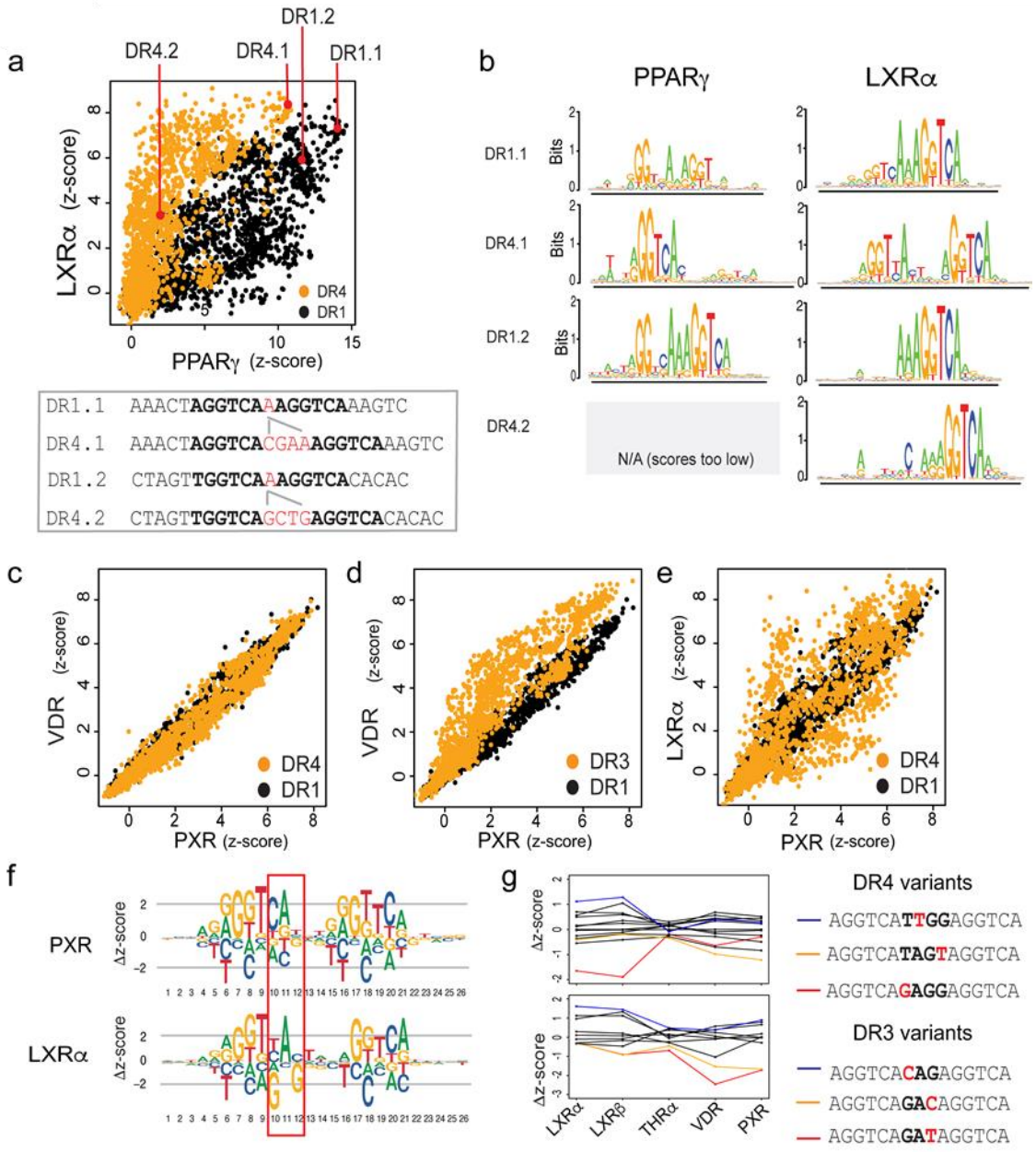


Figure 2.5: NR specificity differences
a Scatter plots of LXRα and PPARγ binding to DR1 and DR4 sites. Each spot is the average of ~5 replicates for each unique DNA sequence (~1600 at each spacer length) on the PBM. DR1 and DR4 spacer-variant sequences are shown in the box below panel. **b** Binding logos generated for LXRα and PPARγ for the spacer-variant seed sequences from **a** are shown. **c** Scatter plots as in **5a** of VDR and PXR binding to DR1 and DR4 sites. **d** Scatter plots as in **5a** of VDR and PXR binding to DR1 and DR3 sites. **e** Scatter plots as in **5a** of LXRα and PXR binding to DR1 and DR4 sites. **f** DR4 z-score logos, directly representing Δz-scores of SNV binding, are shown for LXRα and PXR. Δz-scores are calculated (separately for each NR) as the difference from the median of all SNV variants. **g** Differential binding of NRs to spacer-sequence variants. (Top panel) Binding is shown for five NRs to the DR4 seed sequence 5'-

AGGTCATAGGAGGTCA-3' and all 12 SNVs of the spacer region (spacer region in bold). Δz-scores are calculated as in **2**.**5f**. (Bottom panel) Binding is shown for five NRs to the DR3 sequence 5'-AGGTCAGAGAGGTCA-3' and all nine corresponding SNVs of the spacer region (spacer region in bold). Examples of highly variant spacer sequences are indicated.
Contributions: analyses for this figure were performed by TS, AP, and JLK.

To investigate the plasticity of DR spacer preferences, we compared PXR and VDR, which exhibit broadly similar binding to DR1 and DR4 sites but differ for DR3 binding. PXR and VDR bind with nearly identical specificity to DR1 and DR4 sites (Fig. 2.5c, $R^2 = 0.98$ for both); however, the VDR preference for DR3 sites is seen as an increase in z-score for most DR3 sequences (Fig. 2.5d). This example illustrates that NRs can bind similarly on one DR spacing while having distinct binding preferences for another DR spacing.

Next, we asked whether shared spacer preferences might constrain DNA-base preferences. PXR and LXRα both exhibit preferences for DR1 and DR4 sites (Fig. 2.4); their binding profiles are highly correlated for DR1 sites ($R^2 = 0.95$), but show lower correlation on DR4 sites ($R^2 = 0.83$) (Fig. 2.5e). Analysis of the standard DNA-binding logos did not reveal a strong basis for differential DNA-base preferences. However, by directly examining the impact of SNVs on binding via visualization as an energy matrix (which indicates both favorable and unfavorable interactions), we see strong differences between PXR and LXRα at positions 10 and 12 (Fig. 2.5f). The majority of the PXR-specific binding sites are explained by the existence of a guanine base at position 10 that is highly disfavored by LXRα (G10 carries a z-score penalty of −3.21 for LXRα compared to −0.47 for PXR). We note that the highly unfavorable G10 preference for

DR4 sites (Δz-score = −3.21) is not observed for DR1 sites (Δz-score = −0.65), demonstrating that this NR-specific preference is not shared across all spacer lengths (Supplementary Fig. 2.6). These results highlight the advantages of visualization of energy logos over traditional DNA binding logos, and demonstrate that novel base preferences can arise on DR sites of different lengths.

In NR-binding logos, we observe base preferences in the spacer sequence between DR half-sites (e.g., Figs. 2.2, 2.5f, positions 12–15). We note a strong preference for an adenine in the spacer of DR1 sites, which has been demonstrated for PPAR and other NRs (Bolotin et al., 2010); however, such a distinct base preference is absent at longer spacer lengths (DR2–DR5). To investigate the contribution of the spacer sequence to NR specificity, we examined how spacer variants modulate NR-DNA binding (Fig. 2.5g). We focused our analyses on NRs that exhibit preferences for DR3 and DR4 sites. Examining the binding affinity distribution for SNVs within the spacer of a single seed sequence, we find that the spacer sequence can have considerable impact on binding affinity in an NR-specific manner (Fig. 2.5g), consistent with reports that NRs make DNA contacts with the spacer sequence (Lou et al., 2014). Given the established role for DNA shape in TF binding specificity (Yang et al., 2014; Zhou et al., 2015), we investigated whether DNA shape features in the spacer sequence might also contribute to the selectivity for different binding sites. We examined DNA shape features for spacer variants of DR3 and DR4 sites that enhance or diminish the binding of LXRα and VDR (Supplementary Fig. 2.7). The DNA shape features (i.e., major groove width, helix twist, propeller twist, and roll)

examined are nearly identical for all comparisons. However, we observed a significant difference in the major groove width and roll parameters for VDR binding to DR3 sites. Our results suggest that DNA shape may also play a role in NR-binding specificity. Future studies that more exhaustively sample spacer sequences may enable identification of more subtle differences.

2.3.7 Genomic Binding Agrees with In Vitro Binding Preferences

Our NR-binding landscape (Fig. 2.2) indicates DNA binding to DR sites with many spacer lengths. To determine whether NRs use these diverse sites in vivo, we evaluated the ability of our PBM-derived models to explain in vivo-bound regions from published ChIP-seq datasets (Methods). Examining published PPARγ binding data in HT29 colorectal cancer cells (GSE77039) (Savic et al., 2016), we find that all PPARγ models (DRs and half-sites) can discriminate bound regions from unbound. However, the DR1 model best describes the data (area under the curve (AUC) = 0.70, Fig. 2.6a), in agreement with established PPARγ binding preferences and our PBM data (Fig. 2.4). Testing other published DR1 models (Methods and Supplementary Data 3 from (Penvose et al., 2019) (Isakova et al., 2017; Matys et al., 2006), we find the HOCOMOCO-f1 DR1 model performs best (AUC = 0.67) and with similar accuracy to our DR1 model. These results suggest that binding to DR1 sites is an important determinant of in vivo PPARγ binding. In contrast, all models for LXRα yield similar AUCs (Fig. 2.6b), with the canonically preferred DR4 model performing similarly to the half-site models. Testing other published DR4 models we find JASPAR MA0494.1 (DR4) performs the best (AUC = 0.63), and performs similarly to PBM-derived half-site models (AUCs = 0.64).

These in vivo binding results are consistent with our in vitro binding data, which show a strong DR1 preference for PPARγ and broader binding preferences for LXRα.

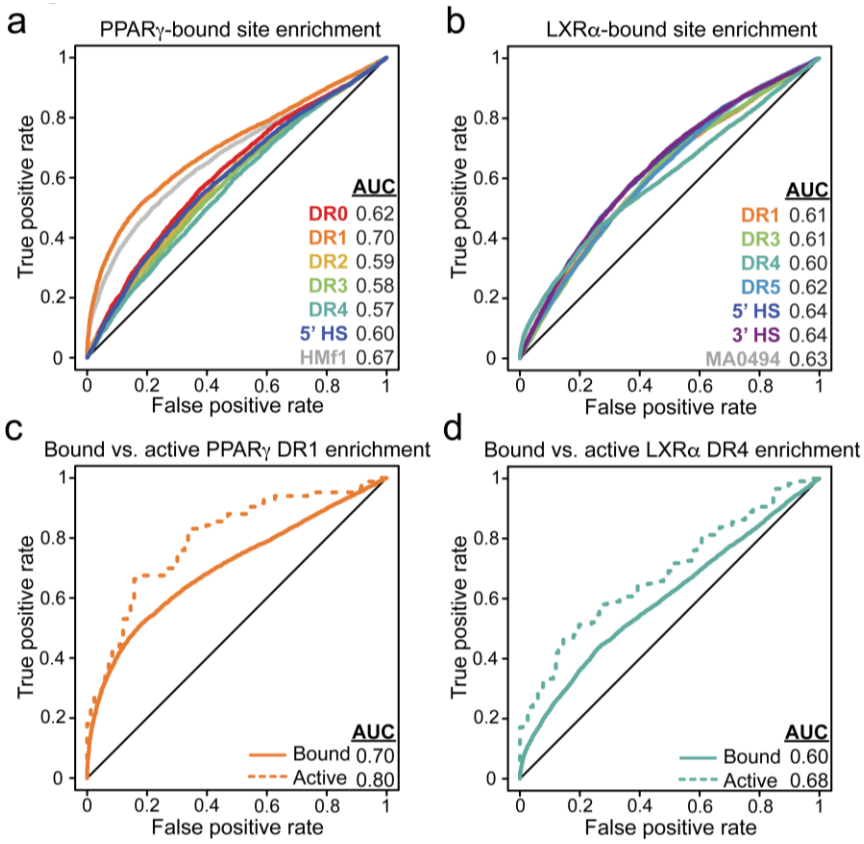


Figure 2.6: Genomic enrichment of NR-binding motifs
a Receiver-operating characteristic (ROC) curves for PPARγ motif enrichment in ChIP-seq data. ROC curves and area under the curve (AUC) values for different PBM-derived NR-binding models are shown, along with the results for best-performing published PPARγ DR1 motif (HOCOMOCO-f1, HMf1(Kulakovskiy et al., 2017)). Motif enrichment for all models had p-values < 10^{-46}, using a Wilcoxon rank sum test with continuity correction and Bonferroni corrected for multiple hypotheses. **b** ROC curves for LXRα motif enrichment in ChIP-seq data. ROC curves and AUC values for different LXRα binding models are shown. Results for best-performing published LXRα DR4 motif (JASPAR MA0494.1(Khan et al., 2018)) are shown. **c** ROC curves and AUC values are shown for PPARγ DR1 motif enrichment in reproducibly-bound PPARγ ChIP-seq peaks (solid lines, Methods), and for those peaks occurring within 10 kb upstream of differentially expressed genes (i.e., active peaks). **d** ROC curves and AUC values are shown for LXRα DR4 motif enrichment in reproducibly-bound LXRα ChIP-seq peaks (solid lines, Methods), and for those peaks occurring within 10-kb upstream of differentially expressed genes (i.e., active peaks).
Contributions: Analyses for this figure were performed by JLK.

2.3.8 Functional Sites Agree with Canonical NR Preferences

We hypothesized that functional binding sites that regulate gene expression may have a different motif composition than the full set of genomic binding sites. Binding sites were annotated as 'functional' if they were located within 10 kb upstream of the transcription start site of genes whose expression changed >2-fold upon agonist treatment (GSE77039 (Savic et al., 2016), Methods). We then performed motif enrichment analysis for these functional PPARγ or LXRα binding sites. Strikingly, we observe an increase in the enrichment of the PPARγ DR1 and the LXRα DR4 models for their respective functional sites (Fig. 2.6c,d). These same trends are observed when we use alternate genomic constraints to define functional sites (i.e., 10 kb up- and downstream, or 50 kb upstream) (Supplementary Fig. 8). These results are consistent with a model wherein NRs preferentially utilize DR full-sites at a canonical spacing for activating transcription, while genome-wide binding is determined by a broader set of DR and half-site sequences, consistent with our in vitro binding data.

2.3.9 NRs Binding via a Half-Site Mode Can Drive Gene Expression

Our analyses reveal widespread binding of NRs to half-site sequences both in vitro and in vivo. Furthermore, we show that half-site mode is utilized by NRs to bind not only to half-sites, but also to canonical DR sites. To determine whether NR half-site mode binding is functional and can drive gene expression, we assayed the ability of LXRα to activate a reporter gene from a binding site bound in a half-site mode on our PBM. Expression of luciferase reporter genes was monitored in HEK293T cells in the presence of over-expressed LXRα:RXRα and ligand or vehicle (Methods). We find that LXRα

strongly induces gene expression, in a ligand-dependent manner, from a DR1 site (DR1.7) that is bound in a half-site mode by PBM (Fig. 2.7a, b, logo illustrates the 5'-half-site binding mode). Ablating the 5' half-site sequence (DR1.18) abrogates binding and drastically reduced reporter gene expression. Ablating the 3' half-site (DR1.17) does not affect binding affinity; however, unexpectedly, it strongly affected reporter gene expression, demonstrating that in vitro affinity does not necessarily predict binding-site activity. Therefore, NRs binding via a half-site mode in vitro can drive gene expression, but DNA bases that do not affect binding affinity in vitro can affect function in vivo.

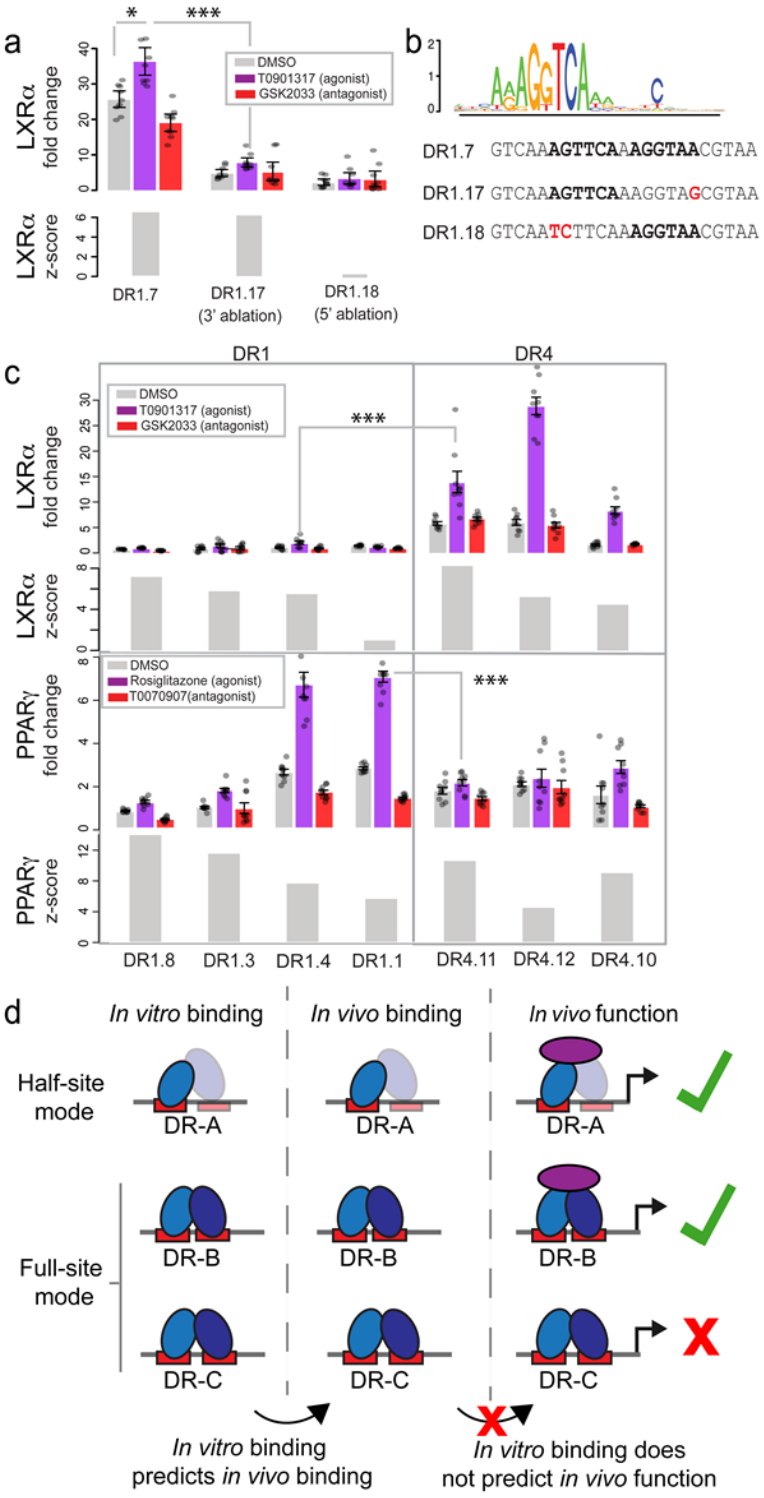

a
LXRα fold change
LXRα z-score
DMSO
T0901317 (agonist)
GSK2033 (antagonist)
DR1.7
DR1.17 (3' ablation)
DR1.18 (5' ablation)
b
DR1.7 GTCAAAGTTCAAAGGTAACGTAA
DR1.17 GTCAAAGTTCAAAGGTAGCGTAA
DR1.18 GTCAATCTTCAAAGGTAACGTAA
c
DR1
DR4
PPARγ fold change
PPARγ z-score
Rosiglitazone (agonist)
T0070907(antagonist)
DR1.8
DR1.3
DR1.4
DR1.1
DR4.11
DR4.12
DR4.10
d
In vitro binding
In vivo binding
In vivo function
Half-site mode
Full-site mode
DR-A
DR-B
DR-C
In vitro binding predicts in vivo binding
In vitro binding does not predict in vivo function

Figure 2.7: Activity versus affinity for distinct classes of NR-binding sites
a LXRα-dependent activity and binding affinity of a sequence bound in a half-site mode. Luciferase reporter gene activation, and corresponding z-scores, are shown for the DR1.7 sequence, which is bound in a half-site mode on PBM, and sequences with each half-site ablated (DR1.17 and DR1.18), sequences shown in **b**. Fold-change reporter expression indicates luciferase activity in HEK293T cells over-expressing LXRα and RXRα normalized to cells not over-expressing these proteins. Fold-change expression is shown for cells treated with DMSO (vehicle), agonist (T0901317), or antagonist (GSK2033), and values represent mean over nine replicate measurements (error bars = SEM). Reporter gene *p*-values: * < 0.01, *** < 0.0001 (calculated using Student's two-tailed *t*-test). **b** Logo for LXRα heterodimer binding to DR1.7, and sequences for DR1.7, DR1.17, and DR1.18 discussed in **a**. **c** LXRα- and PPARγ-dependent activity and PBM-derived binding scores to select DR1 and DR4 sites. Fold-change expression for LXRα is as described in **a**. Fold-change for PPARγ is shown for cells treated with DMSO (vehicle), agonist (rosiglitazone), or antagonist (T0070907), and values represent the mean over nine replicate measurements (error bars = SEM). **d** Overview of relation between NR in vitro binding, in vivo binding, and function. **Contributions: a-c** reporter assays were performed by AP. The concept for the model in **d** was developed by AP, JLK, and TS.

2.3.10 NR Spacing Preferences Are Defined by Function Not Affinity

We next examined the ability of LXRα and PPARγ to promote gene expression from a range of DR1 and DR4 binding sites (Fig. 2.7c). In general, PPARγ drives higher levels of gene expression from DR1 sites, and LXRα functions better on DR4 sites, in agreement with their canonical spacer preferences. However, we see exceptions to these simple rules. First, LXRα can promote expression from the DR1.7 site (Fig. 2.7a) at a comparable or higher level than from the three DR4 sites (Fig. 2.7c). Second, for PPARγ, several high-affinity DR1 sites (DR1.8, DR1.3) show comparable or lower activity than the three DR4 sites, which are all bound with comparable or lower affinity. Complicating the interpretation, without NR overexpression, the DR4 sites exhibit lower reporter gene activity than DR1 sites (Supplementary Fig. 2.9). This low basal activity may exaggerate the NR-dependent activation determined for these sites, which is calculated as the fold-change between basal and NR-over-expressed conditions. Despite these complications, it is clear that affinity does not strongly predict activity of different NRs.

2.4 Discussion

Here we report the most comprehensive DNA binding dataset to date for the type II NRs, and provide a revised framework for interpreting NR-binding and regulatory specificity. We demonstrate more promiscuous DNA binding for NRs than has been previously reported, challenging the view that NR-binding specificity is defined solely by distinct DR spacer preferences. Our findings agree with other PBM-based studies of NR homodimers that demonstrated nearly identical binding for RXRα and COUP-TF2 (Fang et al., 2012), and found that NR specificity does not solely depend on DR-spacing rules (Bolotin et al., 2010; Fang et al., 2012). We demonstrate that NR-binding-site activity does not follow binding affinity, and that the canonical NR DR spacer-length preferences better reflect activity rather than DNA-binding-site affinity. Our revised framework for NR-binding and function shows that NRs bind DNA via two binding modes to a broad set of DR and half-site sequences; this binding corresponds with in vivo binding, but does not correspond to in vivo function, which may involve additional layers of specificity (e.g., allostery) (Fig. 2.7d). Future studies that focus on refining the rules for NR-binding-site activity will clarify this general framework and improve genomic analyses aimed at predicting NR-dependent gene regulation, or the impact of SNPs on NR function, as in a recent analysis of PPARγ function (Soccio et al., 2015).

Our study challenges the prevailing view that each NR heterodimer prefers binding to DR sites of specific spacer lengths (Cotnoir-White et al., 2011; Evans and Mangelsdorf, 2014). We show that all NRs can bind with high affinity to many DR spacer lengths in a

full-site binding mode. Previous studies that sought to identify DR spacer preferences did not explicitly account for multiple NR-binding modes, potentially complicating their interpretations (Cotnoir-White et al., 2011; Evans and Mangelsdorf, 2014; Kurokawa et al., 1993; Perlmann et al., 1993). While we observe previously described DR spacer preferences, our study suggests a distinct biophysical interpretation for these preferences. We propose that DR preferences of NRs are not based on a large increase in binding affinity, but arise from a preference to bind in a full-site mode over a half-site mode, coupled with a moderate increase in affinity (i.e., LXRα and PPARα, Fig. 2.4). The implication that NR spacer preferences are primarily about binding mode, rather than affinity, may provide a biophysical interpretation of NR preferences that links binding mode to in vivo function.

The disagreement between the promiscuous NR binding seen in our study and the canonical DR spacer preferences reported in the literature may be explained by differences in the approaches utilized. DR spacer preferences were initially characterized on a small number of DNA sequences obtained from promoter regions of genes that were upregulated upon ligand treatment, naturally biasing towards functional genomic binding sites (Cotnoir-White et al., 2011; Evans and Mangelsdorf, 2014; Kurokawa et al., 1993; Mader et al., 1993). Other high-throughput methodologies used to examine NR heterodimer binding preferences bias towards high-affinity binding sites and thus do not capture the full landscape of NR-binding specificity (Isakova et al., 2017). Our PBM

approach, which queried the binding across a broad range of affinities and DR spacer lengths, reveals a more promiscuous NR-binding landscape.

We note that we tried other methods for learning NR binding models before deciding the seed-SNV approach was the most effective for this study. Earlier PBM designs utilized a combination of binding sites identified in ChIP-seq data sets and probes containing synthetic DRs, created by combinatorically pairing half-site sequences with varying degrees of degeneracy. We then used support vector regression to generate a binding model for PPARγ:RXRα on DR1s. Briefly, the identities of the nucleotides at each position along the DR were used as features, and the PBM score for each probe was used as the response variable. The PBM data was split into training and testing tests. A parameter search with cross-validation was performed on the training set, and the best set of parameters was used to train the model. The model was then used to predict the PBM scores of the test set. We found using a Gaussian kernel resulted in better predictions than a linear kernel (e.g., R^2 between actual and predicted values was ~0.6-0.7 for the Gaussian kernel vs ~0.5 for the linear kernel), however use of a linear kernel is preferable, since it is easily interpretable and can be represented as a standard DNA logo. As a relatively large number of probes was required to generate this model for a single spacer, we instead decided to try the seed-SNV approach described above. We ultimately found that the seed-SNV approach was preferable, as it enabled direct measurement of the impact of changes in DNA sequence to NR binding without requiring any prediction, and required a smaller sequence space.

Our NR-binding data are consistent with in vivo binding, and provide an updated framework for interpreting genome-wide binding data. For example, PPARγ ChIP-seq peaks are best modeled by a DR1 motif, consistent with the high-affinity binding observed for DR1 sites. In contrast, LXRα ChIP-seq peaks are modeled equally well by most DR models and half-sites (Fig. 2.6), consistent with broader in vitro specificity for LXRα. We note that a DR4 motif was identified by de novo motif analysis using this LXRα ChIP-seq dataset (Savic et al., 2016), but only when restricting the analysis to the highest scoring ChIP-seq peaks; when motif finding is performed on the full dataset, a half-site motif is identified. This example illustrates a source of confusion in the field: reinforcement of established NR-binding preferences by conclusions supported by only a small fraction of the genome-wide binding data (Boergesen et al., 2012; Everett and Lazar, 2013; Savic et al., 2016). Re-interpreting the genomic data in light of our dataset, we find that the broader specificity found in vitro is consistent with in vivo binding.

Unexpectedly, we found that all type II NR heterodimers have the ability to bind DNA via a half-site mode on both full-sites and half-sites. This is a clear example of DNA-based allostery, in which interactions with DNA alter the structure of DNA-bound TFs. Allostery has been reported for the NRs (Gronemeyer and Bourguet, 2009; Meijsing et al., 2009; Schöne et al., 2016; Watson et al., 2013), and provides a mechanism to decouple affinity and activity. A provocative idea is that NR-binding mode may predict activity and explain NR functional preferences. Supporting this idea, a recent study of the

glucocorticoid receptor (GR), a steroid hormone nuclear receptor, showed that GR homodimers can bind to half-site sequences in vivo to repress gene expression (Hudson et al., 2018). Our data on the preference of PPARγ and LXRα to bind in a full-site mode and drive gene expression from DR1 and DR4 sites, respectively, offer additional support for this idea. Other work has demonstrated that NR binding can be altered by cofactor proteins (Issa et al., 2001; Lefebvre et al., 1998), raising the possibility that NR binding modes may be altered in the presence of endogenous cofactors. Future studies that assess NR-DNA binding and binding modes in the presence of cofactors will help clarify the relationship between NR-binding mode, affinity, and activity. Our PBM dataset provides a valuable resource for these future studies aimed at elucidating the mechanisms of NR specificity in gene regulation.

2.5 Materials and Methods

2.5.1 Protein Expression and Purification

Full-length, wild-type human RXRα and PPARγ isoform 1 constructs were cloned into the Gateway vector pDEST17 (LifeTech) for propagation, mutagenesis, and expression. A TEV-protease recognition sequence was included between the coding sequence of the His-tag and RXRα and used to cleave the His-tag after purification. His-tagged RXRα and PPARγ were expressed using the BL21(DE3) *E. coli* strain (NEB). Transformed bacteria were propagated on Luria-Bertani broth (LB) plates supplemented with 100 μg/mL of carbenicillin. Protein expression was carried out in LB supplemented with 100 μg/mL of carbenicillin, with an initial outgrowth at 37 °C up to an OD of 0.4,

transferred to ~20 °C until they reached an OD of 0.6–0.7 and then induced with 1 mM IPTG. Protein was expressed at room temperature (~20 °C) for 3 h. Cells were pelleted and stored at −80 °C until purification. Purification was carried out using HisTrapFF columns (GE Healthcare). The binding buffer was composed of 20 mM Tris HCl, 300 mM NaCl, 25 mM Imidazole, and 1 mM DTT and the elution buffer was composed of 20 mM Tris HCl, 300 mM NaCl, 250 mM Imidazole, and 1 mM DTT. Buffers were supplemented with cOmplete Mini protease inhibitor tablets according to the manufacturer's instructions (Roche). Eluted fractions were analyzed by SDS-PAGE and fractions containing protein were combined. For PPARγ, the combined elution fractions were buffer exchanged into phosphate buffered saline pH 7.4 with 1 mM PMSF and 10% glycerol using an Amicon Ultra centrifugal filter (30k MWCO). Elution fractions of RXRα were dialyzed against three changes of binding buffer. Next, the His-tag was cleaved from RXRα by overnight incubation at 4 °C with TEV protease (Sigma–Aldrich). After cleavage, the RXRα sample was re-purified as described above; however, this time the flow-through fraction from the column loading was collected and used in all PBM experiments, as this fraction contained the RXRα from which the His-tag was successfully cleaved. The combined flow-through fractions were buffer exchanged into phosphate buffered saline pH 7.4 with 1 mM PMSF and 10% glycerol using an Amicon Ultra centrifugal filter (30k MWCO).

The RXRα and PPARγ DNA binding domain mutants were made by site-directed mutagenesis using the NEB Q5 site-directed mutagenesis kit (New England Biolabs)

following the manufacturer's instructions. Primers used for the mutagenesis were: RXRα: Forward = 5'-CTTCTTCTTCAAGGCGACGGTGCGCAAGGACCTG, Reverse = 5'-CCCGCGCACCCCTCGCAGCTGTACACTCCATCAGC; PPARγ: Forward = 5'-CTTCCGGGCAACAATCAGATTGAAGCTTATCTATGACAG, Reverse = 5'-AAACCCGCGCATCCTTCACAAGCATGAACTCCATAGTG. For DNA binding domain mutant experiments, both wild-type and mutant RXRα were expressed using the PURExpress IVT kit (NEB) according to manufacturer instructions. The concentration of all IVT-produced proteins was estimated by western blot by comparison to purified proteins. All other purified proteins used were purchased (see Supplementary Data 1 for details).

2.5.2 PBM Custom Design

PBM experiments were performed using custom-designed microarrays (Agilent Technologies Inc. AMADID 084387, 4 × 180 K format). PBM probes contain a 24 nt constant primer region, a 34 nt variable region, and a 5' GC dinucleotide cap (probe sequences can be found in Supplementary Data 4 of (Penvose et al., 2019). For each unique SNV probe sequence, five replicate probes were included in each orientation (10 probes per unique sequence). For all other probe sequences four replicate probes were included with the 34 nt variable region in each orientation (8 probes per unique sequence).

SNV probes: DR seed sequences, defined by two 6-bp half-sites and a variable spacer (0–5 bp), were aligned within in the 34 nt variable region of each PBM probe, or shifted 1bp away from the center towards the free end of the DNA in the case of spacers that contain

an odd number of nucleotides. Flanking regions around the DR in the seed were randomly generated within the constraints that they create no spurious binding sites (as described in "*Minimizing spurious half-sites in probe design*") and that they do not create repeats of a single nucleotide longer than 3 nt. For each seed sequence, SNV probes were created that had a single-nucleotide variant at each position of the DR half-sites, the spacer sequence between the DR half-sites, and in the 5 bp flanks of each site. Therefore, for a single 13 bp DR1 site (i.e., 6 + 6 + 1 = 13), including 5 bp flanks on either side, there would be 69 (i.e., 23 × 3) unique SNV probe sequences.

Minimizing spurious half-sites in probe design: Care was taken to minimize the presence of half-sites in probe sequences outside of the intended DR, including those introduced by concatenation of the 34 bp variable region with the GC cap or primer region. We define a spurious half-sites as any region that yields a score greater than 0.1 with the following PWM:

A	0.4905	-1.6556	-1.658	-0.3546	-1.6556	0.3858
C	-1.6556	-0.6556	-1.658	-1.6556	0.4905	-0.3546
G	-0.1785	0.5485	0.4559	-0.1785	-0.1785	-1.6556
T	-0.6556	-0.6556	0.041	0.4583	-0.6556	0.0434

The primer was designed to contain no spurious half-sites and to minimize the creation of spurious half-sites upon concatenation with the variable region. In cases where concatenation of the 34 nt variable region with the GC cap or primer region introduced a spurious half-site, these spurious binding sites were ablated (modified to score below 0.1

with the PWM described above) with a minimal number of changes. Changes were applied to regions farthest from the DR, and were applied to all SNV probes from the same seed. If we were unable to identify a set of changes that ablated the spurious binding site without introducing new spurious BS in SNV probes from the same seed, the original sequence was kept.

Half-site ablation probes: For each seed sequence, probe variants were created with each half-site ablated. We define an ablation as reducing the score below 0.1 when scored with the PWM described in "*Minimizing spurious half-sites in probe design*." Ablations were performed by identifying the position in the spurious half-site that contributes most to the score and replacing it randomly with a lower scoring base. If the alteration did not introduce any new half-sites with a score greater than 0.1, the new sequence was kept. If the score was less than 0.1, this sequence was used as the ablation for that half-sites. If the score remained above 0.1, the process was repeated with the next highest scoring position until the half-site score was below 0.1. For DR0 seeds, 4 nt around the desired BS were protected from changes. For DR1 – DR5 seeds, 5 nt around the desired BS were protected from changes.

2.5.3 PBM Experiments and Analysis

Microarrays were double-stranded as previously described (PBM double-stranding primer 5'-CCTTCATTCTACGCTGTCAATCGC-3') (Berger and Bulyk, 2009; Berger et al., 2006). All washes were performed in coplin jars on an orbital shaker at 125 rpm. Double-stranded microarrays were first pre-wetted in PBS containing 0.01% Triton X-100 for 5 min, rinsed in a PBS bath, and then blocked with 2% milk in PBS for 1 h. After

blocking, arrays were washed in PBS containing 0.1% Tween-20 for 5 min, then in PBS containing 0.01% Triton X-100 for 2 min and then rinsed in a PBS bath. Proteins were then incubated on the array for 1 h in a binding reaction containing: PBS pH 7.4 with 2% milk, 0.02% Triton X-100, 1 mM DTT, 0.2 mg/mL bovine serum albumin, and 0.4 mg/ml salmon testes DNA (Sigma D7656). See Supplementary Data 1 for protein concentrations. Preliminary PBM experiments for PPARγ:RXRα and RXRα were performed with and without the ligands rosiglitazone and 9-*cis* retinoic acid, respectively, and we found no change in NR binding; therefore, all experiments were performed in the absence of ligand. Following the protein incubation, microarrays were washed with PBS containing 0.5% Tween-20 for 3 min, then in PBS containing 0.01% Triton X-100 for 2 min followed by a brief PBS rinse. Microarrays were then incubated with 20 μg/ml of primary antibody in 2% milk in PBS for 20 min. For heterodimers, separate experiments were performed using an antibody against each protein within the heterodimer. In all experiments, anti-RXRα antibody (Active Motif 61029) was used to detect RXRα and anti-His antibody (Sigma H1029) was used to detect the NR partner with the following exceptions: anti-PPARγ antibody (Abcam 41928) was used in all experiments with PPARγ, and Alexa488-conjugated anti-GST antibody (Life Tech A11131) was used for all PPARα experiments. Excess primary antibody was removed by washing with PBS containing 0.05% Tween-20 for 3 min and then in PBS containing 0.01% Triton X-100 for 2 min. Arrays were next incubated with 20 μg/ml of Alexa488-conjugated secondary antibody (anti-mouse A488, Life Tech A11001) in 2% milk in PBS for 20 min (PPARα was probed with an Alexa488-conjugated anti-GST primary antibody as described above

and did not require a secondary antibody). Excess antibody was removed by washing 2x with PBS containing 0.05% Tween-20 for 3 min and then in PBS for 2 min. Microarrays were scanned with a GenePix 4400 A 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 and fluorescence values from each experiment are provided (See Supplementary Data 4 of (Penvose et al., 2019). NR dimers exhibit an orientation-specific bias in our PBM experiments; therefore, data from probes in a single orientation (i.e., '_o1′ probes in Supplementary Data 4 of (Penvose et al., 2019) was used in our final analysis. However, all results were observed for probes in both orientations and models from each orientation showed good agreement.

Position frequency matrices (PFMs) and DNA-binding logos were generated for each seed sequence with z-score >3.0 using the previously described SNV-based approach(Andrilenas et al., 2018), with β set to 15/maximum z-score. Briefly, logos for single seed sequences are generated using the binding data to each seed sequence and all the single-nucleotide variant (SNV) sequences for that seed sequence. For a binding site of length L there will be 3xL SNV sequences. Logos for an NR binding to a specific DR spacer length are determined by averaging over the individual seed sequence logos. To generate logos for a specific DR spacer length (Fig. 2.2), PFMs for all seed sequences at that spacer length were clustered into full-site, 5'-half-site or 3'-half-site PFMs. Average PFMs of each type (i.e., full, 5'-half-site or 3'-half-site) were then generated by directly averaging over the individual PFMs (i.e., averaging individual matrix elements and

normalizing each column to 1). As the half-site PFMs are the same length regardless of the starting DR seed length, the final 5'-half-site and 3'-half-site PFMs were further averaged over PFMs generated at all spacer lengths. The z-score energy matrix (Fig. 2.5f) was generated in the same manner, without the initial transformation from z-score to frequency (Andrilenas et al., 2018).

2.5.4 Reporter Gene Assays

PPARγ, LXRα, and RXRα were cloned into the N-terminal His-tagged protein mammalian expression plasmids (pDEST26, LifeTech). Reporter constructs for test sequences were ordered synthesized (Twist Bioscience) and were flanked by two BsaI cut sites, which were used to clone the sequences into pNL3.1-minP/Nluc (Promega). All sequences tested can be found in Supplemental Data 2. HEK293T (ATCC) cells were cultured in DMEM (Gibco 11965-092) supplemented with 10% FBS (Gibco 26140079). Cells were plated in tissue culture treated 96-well plates seeded at a density of 12,500 cells per well and allowed to adhere overnight. PEI:DNA complexation reactions were prepared at a ratio of 3:1 (PEI:DNA) in 500 µl of Opti-MEM (Gibco 51985-034) and allowed to complex for 20 min at room temperature. Each 96-well plate well received 20 µl of transfection mixture containing 16 ng of total plasmid: 1 ng of transfection normalization plasmid (pGL4.54-Luc2/TK); 10 ng of reporter plasmid (pNL3.1-minP/Nluc); and either 5 ng of empty pDEST26 for the no overexpression conditions (NoOE); or 2.5 ng of RXRα in pDEST26 combined with either 2.5 ng of PPARγ or LXRα in pDEST26 for protein overexpression condition (OE); Twenty-four hours after transfection, 80 µl of media was removed from each well and replaced with 80 µl of fresh

media containing the appropriate ligand treatment. PPARγ ligands were 1 μM rosiglitazone (Sigma–Aldrich) and 1 μM T0070907 (Sigma–Aldrich). LXRα ligands used were 1 μM GSK2033 (Sigma–Aldrich) and 500 nM T0901317 (Sigma–Aldrich). Luciferase activity was assessed 18 h after addition of the ligand using the Nano-Glo Dual Luciferase reporter assay system (Promega). Dual luciferase signal was quantified using a VICTOR-3 plate reader (PerkinElmer). To control for transfection efficiency, the Nluc reporter plasmid signal was normalized to the constitutive luciferase signal (i.e., signal from pGL4.54 plasmid) (Nluc/Luc2). Normalized signal for all test DNA elements were then further normalized to empty vector (pNL3.1-Nluc with an insert of equal length to test sequences but lacking any half-site or direct repeat sequences). Fold-induction values for each protein + reporter combination were calculated relative to the background activity of each reporter plasmid in the absence of protein overexpression: (protein + reporter)/(control + reporter) = OE/NoOE (Supplementary Fig. 2.9). Reporter assays were performed as three biological replicates with three technical replicates per biological replicate.

2.5.5 EMSA Experiments

Complementary DNA oligonucleotides (sequences in Supplementary Data 2) were ordered from IDT and annealed in a thermocycler by raising the temperature to 98 °C and reducing the temperature by 0.1 °C/sec until a temperature of 4 °C was reached. All DNA sequences are provided in Supplementary Data 2. EMSA buffer formulation for all reactions was 1x PBS with 0.2% BSA, 5 mM DTT, 10% glycerol, and 0.02% Triton-X100. For the direct binding experiment, 1 nM of IR700-labeled P1 probe was incubated

with varying concentrations of PPARγ:RXRα in a 20 μL reaction. For competition experiments, 2 nM of IR700-labeled P1 probe was incubated with PPARγ:RXRα (12 nM:4 nM) in a 20 μL reaction with various concentrations of unlabeled competitor sequences (0, 0.2, 0.63, 2, 6.3, 20, 63, 200, 630, and 2000 nM). Reactions were incubated for 1 h at room temperature and then run in 0.5x TBE on a 6% TBE-acrylamide gel at 50 V for 3 h. Gels were scanned on the Odyssey CL-X (LI-COR) at 84 μM resolution. Fluorescence of the shifted band was quantified using ImageStudioLite software. All K_d calculations were done with DynaFit 4 software (Kuzmic, 1996) using a previously described competition protocol (Golden et al., 2013). Percent competition was calculated by the formula:

% inhibition = $(F_0 - F_c)/F_0 * 100$

F_0: fluorescence of shifted band with no competitor DNA

F_c: fluorescence of shifted band at given concentration of competitor DNA.

2.5.6 Enrichment of NR-binding Sites in ChIP-seq Data

Receiver-operating characteristic (ROC) curve analyses were performed to quantify the extent to which NR-bound (true positive) regions scored more highly than unbound (true negative) regions with PWM models. True-positive regions for LXRα and PPARγ were derived from ChIP-seq data from HT29 colorectal cancer cells (GSE77039) (Savic et al., 2016). ChIP-seq was available for two biological replicates of HT29 cells treated with agonist (GW3965 for LXRα or rosiglitazone for PPARγ) for 2 h and 48 h. For each NR, ChIP peaks with 50% reciprocal overlap within time points and between time points were considered true-positive regions. True-negative regions were derived from DNase-seq of

HT29 cells (GSE90403) (Consortium et al., 2013). Regions with 50% reciprocal overlap between the two available DNase-seq biological replicates were identified, and all ChIP peaks from the corresponding NR ChIP datasets were then subtracted from the DNase-seq regions. Regions matched in size to each ChIP-derived true-positive region were randomly chosen from ChIP-subtracted DNase-seq regions to create the true negative regions. Background nucleotide frequencies for calculating PWMs from PFMs were taken from the nucleotide distribution of the DNase-seq regions with 50% reciprocal overlap between the two replicates. To score sequences, the following formalism was used:

$$p_{i,j=}\frac{f_{i,j}+sb_i}{\sum_i f_{i,j}+s}$$

Probability of an A,C,G or T (i = 0,1,2,3 respectively) occurring at position j of the sequence being evaluated.

$f_{i,j}$: frequency defining the position frequency matrix

b_i: nucleotide background frequencies: A: 0.24; T: 0.24; C: 0.26; G: 0.26

s: pseudo-count to deal with zeros (s = 0.001)

The PWM score is the sum over all base positions (j) of the corresponding $S_{i,j}$ values for a particular sequence:

$$s_{i,j} = log_2(\frac{p_{i,j}}{b_i})$$

Area under the ROC curve (AUC) values are reported to quantify the enrichment, and a Wilcox-Mann-Whitney (WMW) U test was applied to calculate the significance of each AUC value. AUC and WMW U test values were calculated in the R statistical package using the wilcox.test function. All manipulations of genomic regions (identification of overlapping regions, region subtractions, etc.) were performed with BEDTools 2.26.0 (Quinlan and Hall, 2010).

To examine the motif enrichment of currently available models, we performed the ROC analyses described above with publicly available PFMs. Each PFM was normalized such that the nucleotide frequencies at each position sum to 1. The following models were used: LXRα (MA0494.1(Khan et al.); HOCOMOCO f1(Kulakovskiy et al., 2017)), PPARγ(Isakova et al., 2017); M00512, M00515, M00528(Matys et al., 2006); MA0065.1, MA0065.2, MA0066.1(Khan et al.); HOCOMOCO f1, HOCOMOCO s1 (Kulakovskiy et al., 2017)).

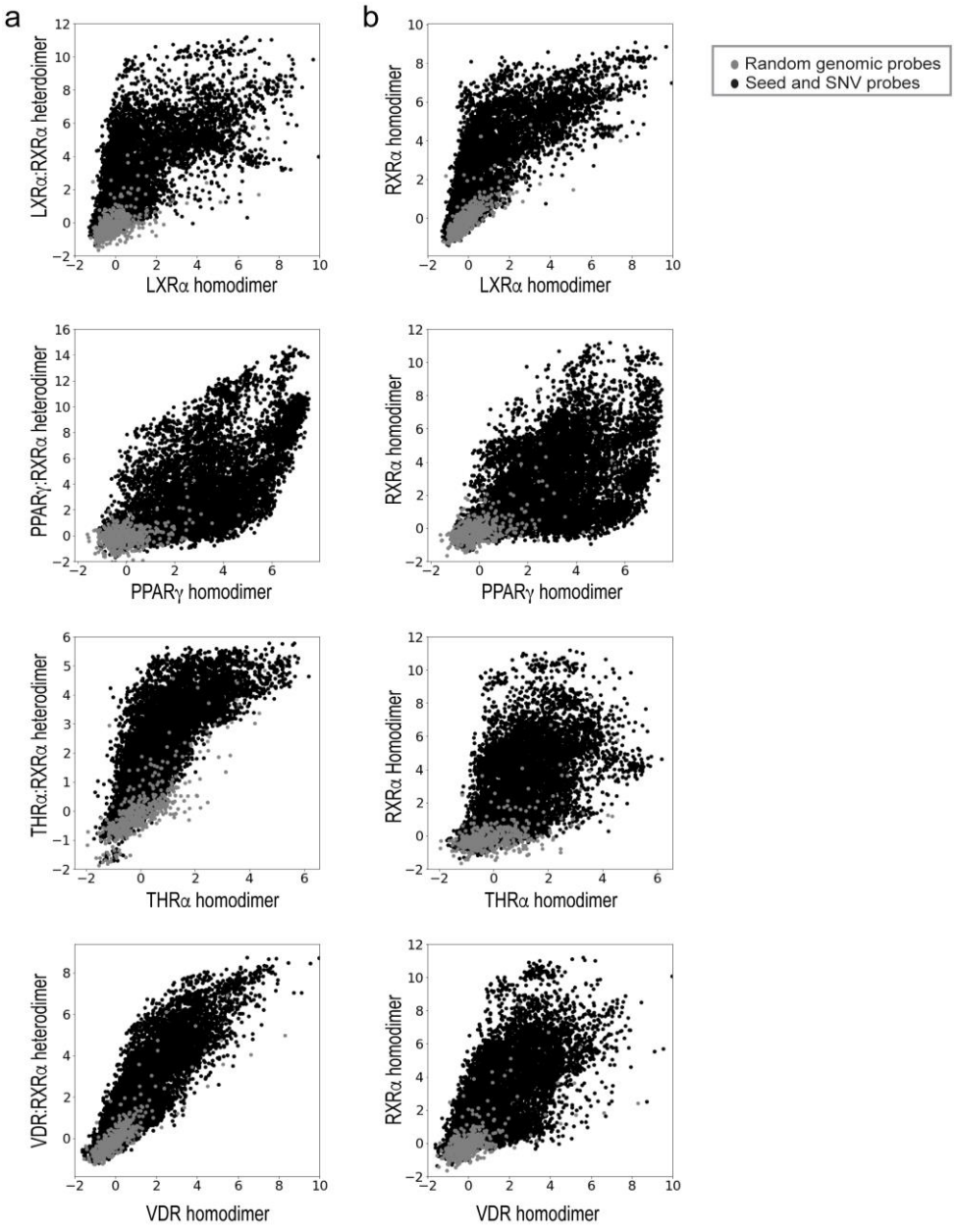
To examine motif enrichment for putative 'active' sites near differentially expressed genes, RNA-seq data from HT29 cells (Savic et al., 2016) were used to identify regions that are likely to be actively controlling transcription. We re-analyzed the published RNA-seq data using DESeq2 (Love et al., 2014) to identify genes upregulated upon agonist treatment compared to vehicle only (DMSO). Transcripts with a fold-change greater than 2 and adjusted *p*-values less than 0.01 were considered upregulated. For PPARγ, transcripts upregulated after both 24 and 48 h of rosiglitazone treatment were

considered for further analysis. For LXRα, transcripts upregulated after 48 h of GW3965 and T0901317 treatment were considered for further analysis. For each NR, ChIP regions with 50% reciprocal overlap between replicates and time points and within the indicated regions associated with upregulated genes were considered active true positives for enrichment analysis. Regions matched in size to each active region were randomly chosen from the true-negative regions described above to create the true negative regions. ROC analyses were performed as described above.

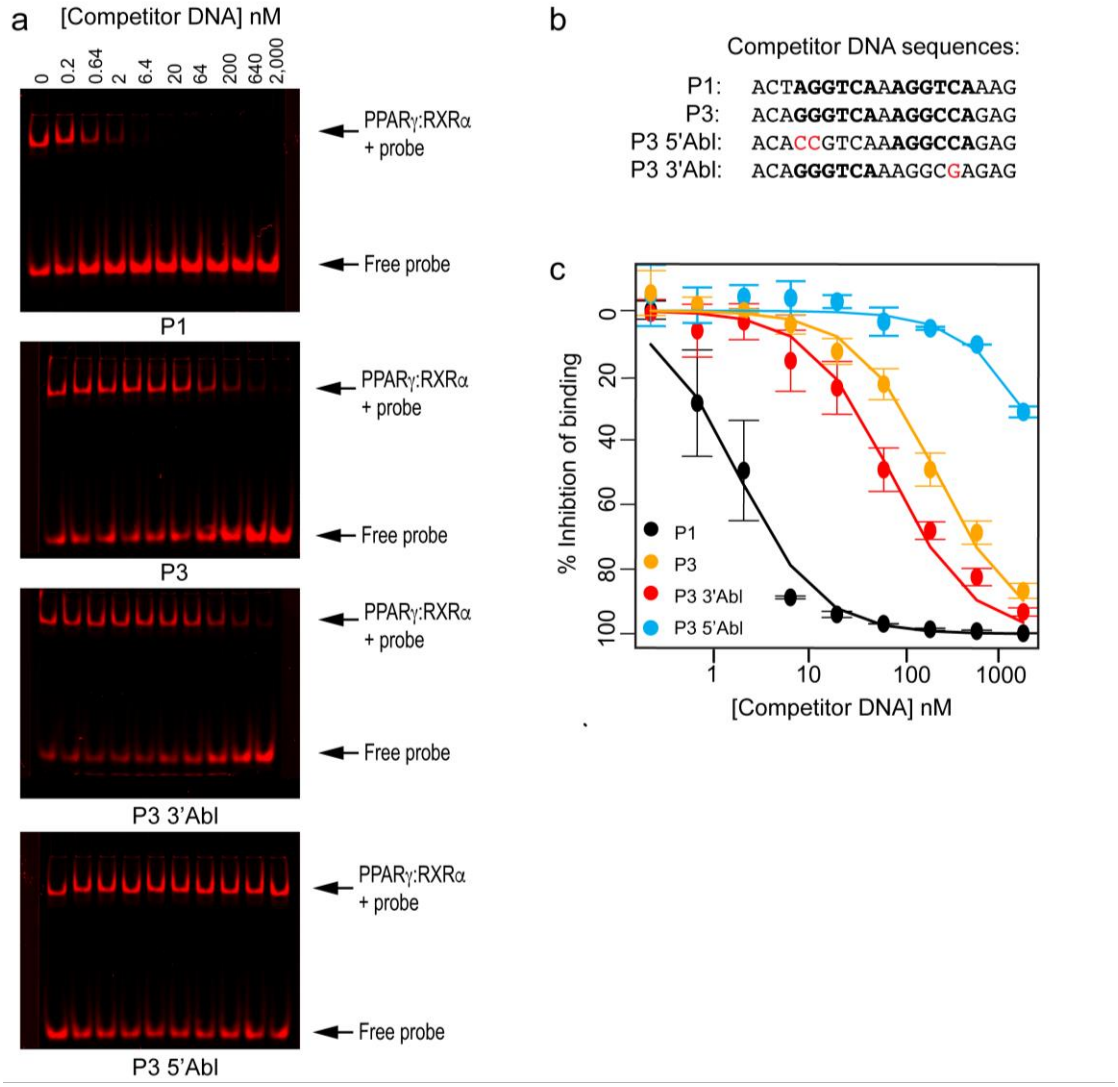
2.5.7 DNA Shape Analysis

Binding to spacer-sequence variants of five DR3 and five DR4 seed sequences was analyzed (Supplementary Fig. 2.7). For each DR3 seed sequence, the PBM z-scores of the seed sequence and corresponding 9 SNV sequences (i.e., sequences with base variants at positions B1, B2, or B3) were analyzed to identify the two highest affinity and the two lowest affinity sites for each of the five seeds, resulting in a total of ten high and ten low-affinity spacer variants. The same procedure was performed for the DR4 sequences and the corresponding 12 SNVs at positions B1, B2, B3, and B4. For each of the 10 spacer variants, the following DNA shape parameters were calculated at each base position using the TFBSshape server (Yang et al., 2014): major groove width (MWG), helix twist (HelT), propeller twist (ProT), and roll. The distribution of the DNA shape parameters associated with high and low-affinity sequences were compared at each base position using a two-tailed *t*-test.

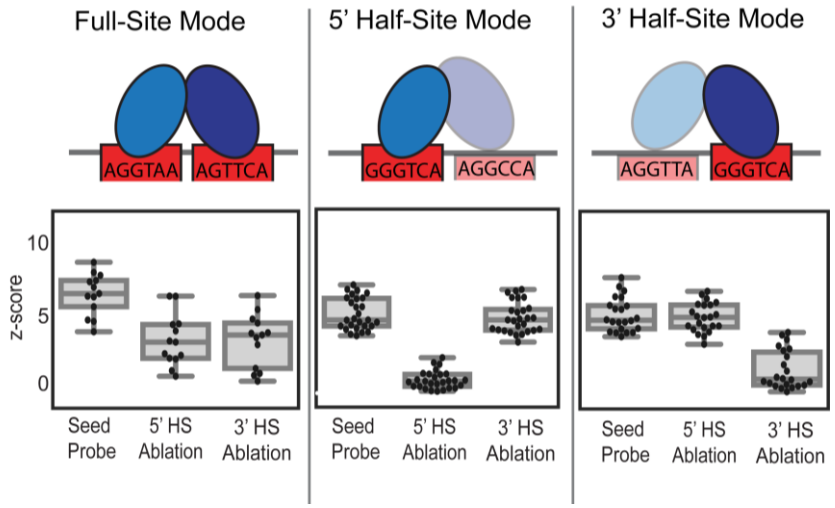
2.6 Supplementary Figures



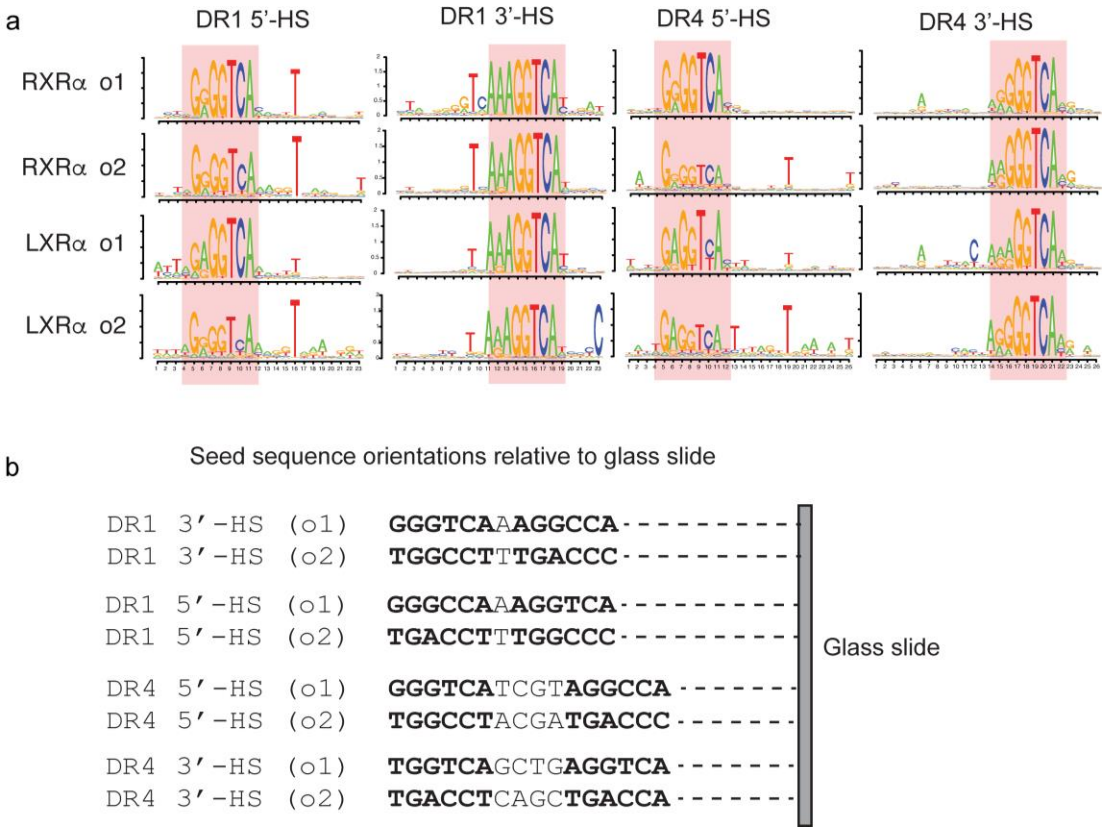
Supplementary figure 2.1: Comparison of NR homodimer and heterodimer binding
Z-scores for **a** NR as a heterodimer with RXR against the corresponding NR homodimers or **b** for RXR homodimer against NR homodimer. Dots represent average over ~5 replicates for all 10,728 unique SNV probes and 500 background probes.
Contributions: PBM experiments were performed by AP**.**



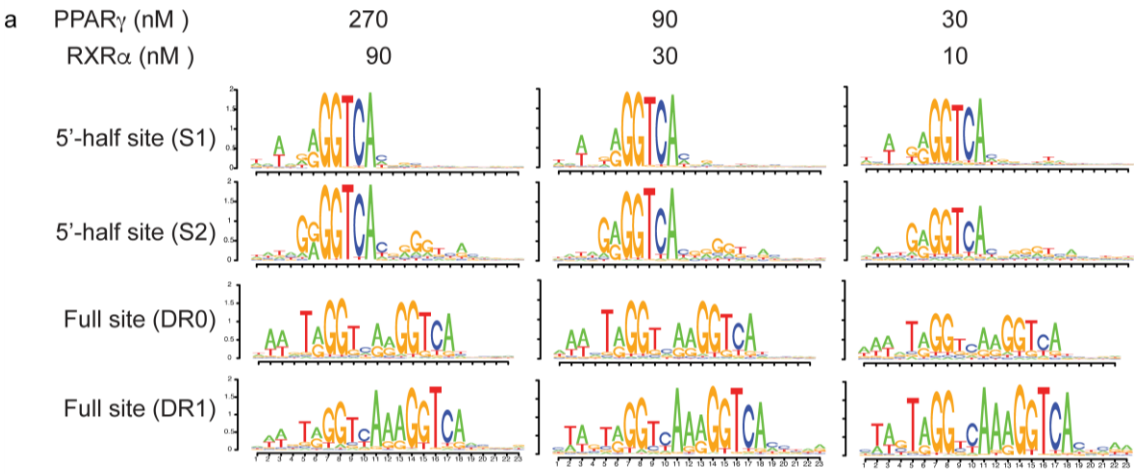
Supplementary figure 2.2: Competition EMSA experiments for PPARγ:RXRα
a Representative EMSA gels of competition for binding by PPARγ:RXRα to labeled DNA probe (P1, as described in Fig. 1g) and four unlabeled competitor DNA sequences whose sequence are shown in **b**. **c** Inhibition curves determined by quantifying the intensity of the bound probe band at different competitor concentrations for the different competitor experiments (error=STDEV, n=2).
Contributions: EMSA experiments were performed by AP**.**



Supplementary figure 2.3: Impact of half-site ablation on LXRα binding
Z-score distribution of LXRα binding to seed probes bound in the full-site mode or half-site mode, and the z-score distributions for binding to corresponding sequences with 5' or 3' half-site ablations.
Contributions: PBM experiments were performed by AP**.** HS probes were designed by JLK.



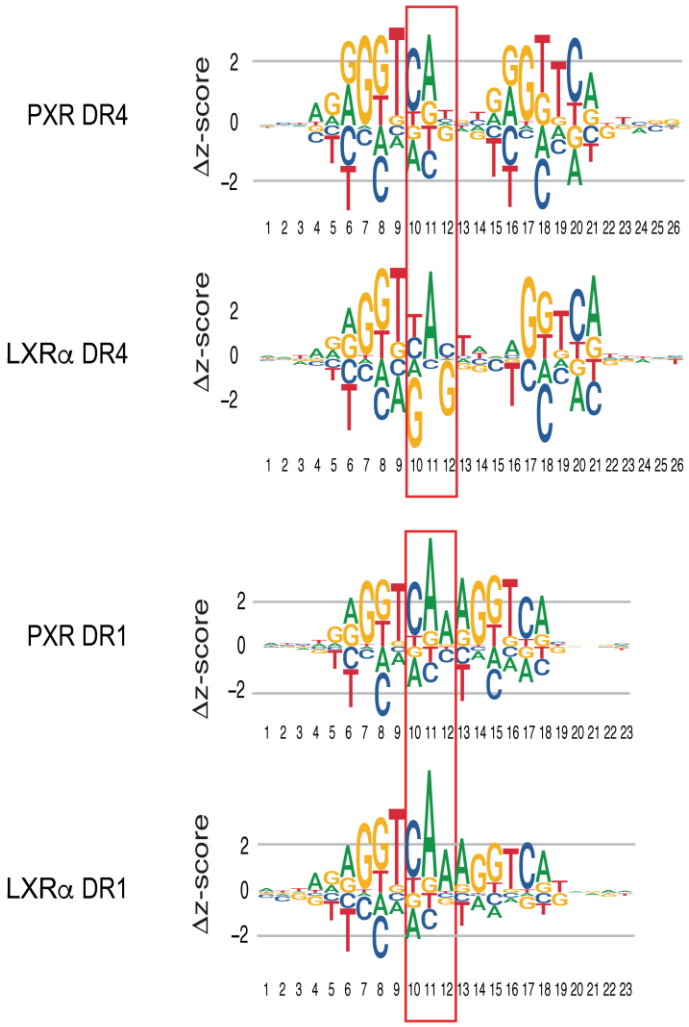
Supplementary figure 2.4: Impact of PBM probe orientation on NR binding logos
a DNA binding logos for RXRα homodimers and LXRα:RXRα are shown for DNA sequences bound in either a 5' half-site or 3' half-site binding mode. DNA binding logos were determined separately from PBM probes in which the binding site (and all SNVs used in the logo determination) are oriented in either the o1 or o2 orientation with respect to the glass slide (schematized in **b**). Bases indicating the binding mode preference are highlighted with the red overlay box**. b** Schematic of DNA seed sequences used to generate the logos showing the orientation relative to the microarray glass slide.
Contributions: PBM experiments were performed by AP**.**



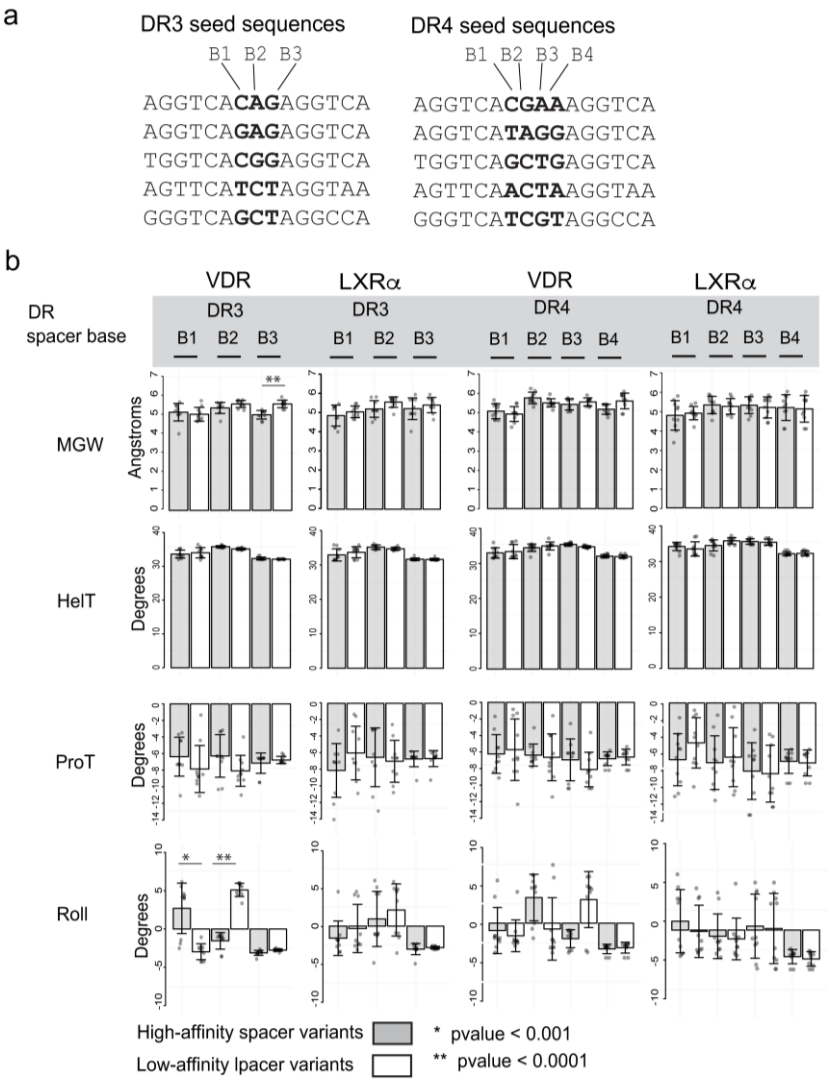
b Seed Sequences for Motif Models

5'-half site (S1)	GAACT**AGGTCA**CC**AGGACA**GTGAA
5'-half site (S2)	ATACA**GGGTCA**CT**AGGCCA**GAGTA
Full site (DR0)	TATGT**AGGTTAGGGTCA**TTCAG
Full site (DR1)	CTAGC**AGGCCA**A**GGGTCA**CTCAA

Supplementary figure 2.5: Impact of protein concentration on NR binding logos
a PPARγ:RXRα DNA binding logos for DNA seed sequences bound in full or half-site binding modes are shown for PBM experiments performed at three different concentrations. The concentration of each monomer used in each PBM experiment is indicated. **b** The seed sequences for which the logos in **a** were generated. Identifiable DR half-sites in each binding sequence are shown in bold.
Contributions: PBM experiments were performed by AP**.**



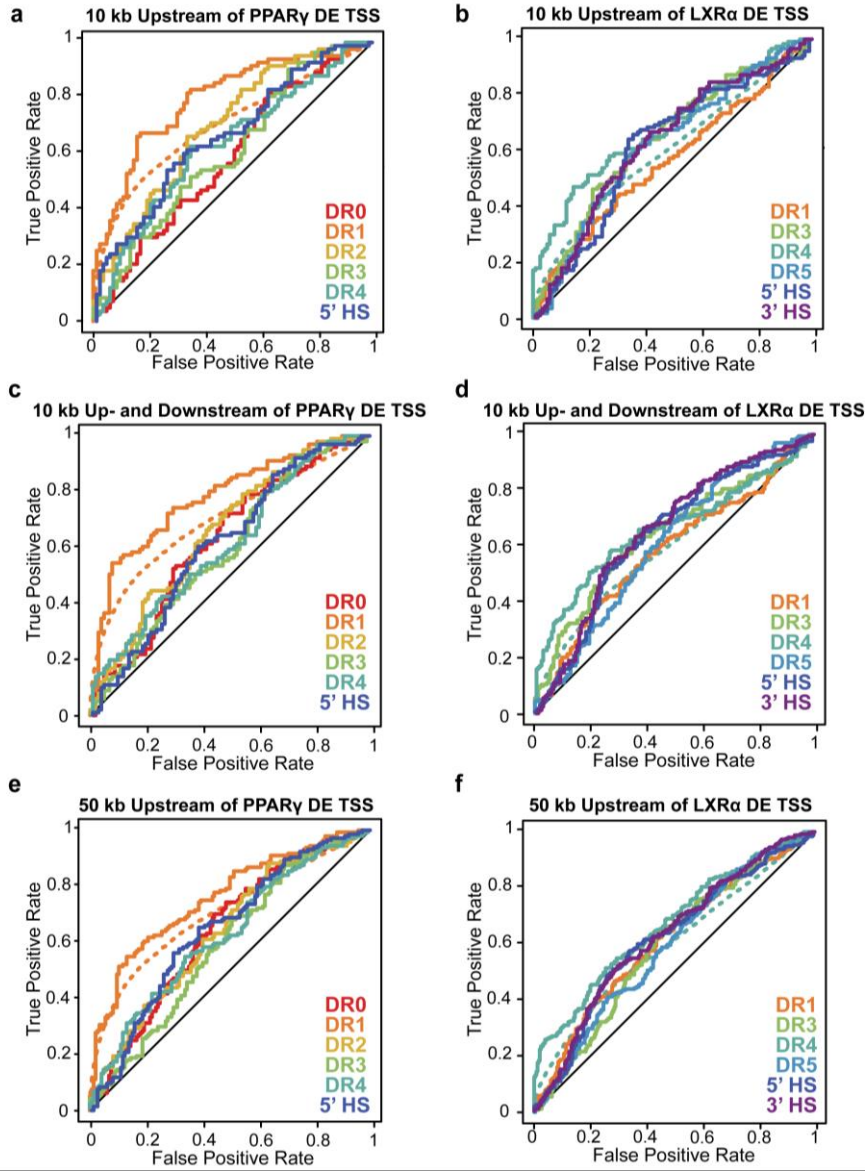
Supplementary figure 2.6: DNA energy matrix logos for LXRα and PXR
DR1 and DR4 logos, directly representing Δ z-scores of SNV binding, are shown for LXRα and PXR. DR4 logos are derived from the same experiments as those in Fig. 5 and are shown for comparison. Positive Δ z-scores indicate z-scores higher than the median z-score for all base variants at that position
Contributions: PBM experiments were performed by AP**.** PBM design was performed by JLK and AP as described in the note at the beginning of this chapter. TS performed analyses for the related Figure 3.5.



Supplementary figure 2.7: DNA-shape parameters of spacer sequences for high and low-affinity NR binding sites

a Schematic of DNA seed sequences used to analyze DNA shape features (shown in **b**). Base positions in the spacer sequence between the DR half-sites are indicated in bold and referred to as B1,B2,B3 (DR3 site) and B1,B2,B3,B4 (DR4 site). Seed sequences were selected to represent diverse spacer sequences. **b** Distribution of DNA shape features for spacer sequences in either high-affinity sites (grey bars) or low-affinity sites (white bars). Data is shown for VDR and LXRα heterodimer binding experiments. For each of the 5 seed sequences (at each spacer length), we identified the two highest affinity and the two lowest affinity spacer sequence variants. Therefore, there are 10 (i.e., 5x2) high-affinity and 10 low-affinity spacer sequences considered for each bar plot. For each of the 10 spacer variants, DNA shape parameters were calculated at each base position using the TFBSshape server(Yang et al., 2014) – major groove width (MWG), helix twist (HelT), propeller twist (ProT), and roll. Shown at each base position is the mean parameter over 10 sequences (error = STDEV). Distributions that were significantly different between the high and low-affinity sequences are shown (p-value calculated using a two-tailed t-test). Source data are provided as a Source Data file.

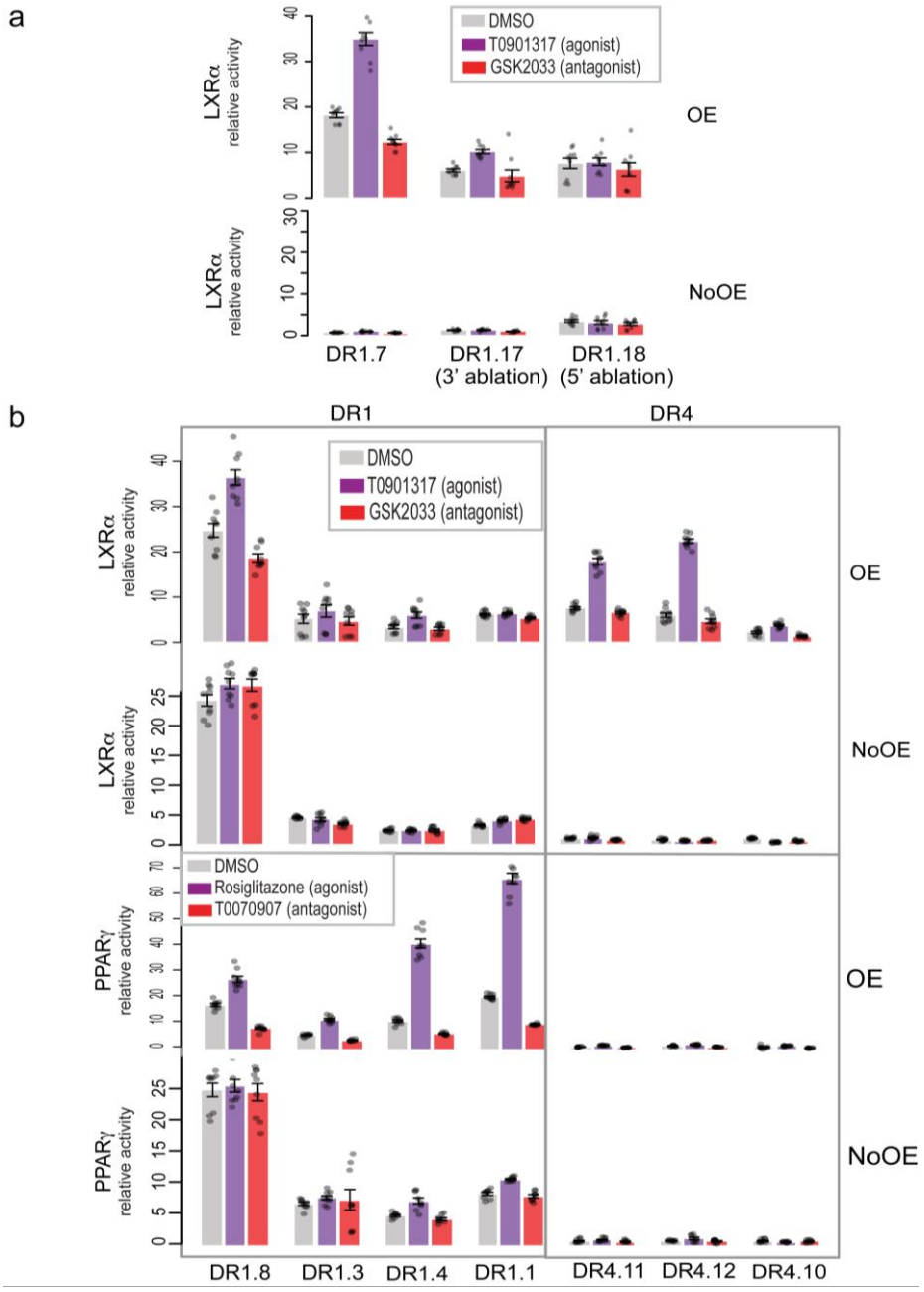
Contributions: TS performed the analyses for this figure.



Supplementary figure 2.8: Receiver operating characteristic (ROC) curves for PPARγ and LXRα motif enrichment in ChIP-seq data

ROC curves for motif enrichment of PBM-derived PPARγ-binding models are shown for all reproducibly-bound PPARγ ChIP-seq peaks (dotted lines, **a,c,e**) and reproducibly-bound PPARγ ChIP-seq peaks occurring within **a** 10 kb upstream, **c** 10 kb upstream or downstream, and **e** 50 kb upstream of the transcription start site of differentially expressed genes (solid lines, Methods). ROC curves for motif enrichment of PBM-derived LXRα binding models are shown for all reproducibly-bound LXRα ChIP-seq peaks (dotted lines, **b,d,e**) and reproducibly-bound LXRα ChIP-seq peaks occurring within **b** 10 kb upstream, **d** 10 kb upstream or downstream, **f** and 50 kb upstream of the transcription start site of differentially expressed genes (solid lines, Methods). ROC curves determined using different PWMs for different DR and half-site (HS) modes are indicated.

Contributions: JLK performed the analyses for this figure.



Supplementary figure 2.9: Impact of NR over-expression on reporter gene activity
a,b LXRα- and PPARγ-dependent activity for the sequences described in Fig. 7 in the same treatment conditions. Shown separately are the luciferase activity values for the cells in which the NR:RXRα proteins were overexpressed (OE) and the values in which the proteins were not overexpressed (NoOE), each normalized to empty vector. Fold-change values in Fig. 7 are the ratio of these sets of values (i.e., OE/NoOE). Values represent the mean over nine replicate measurements (error bars = SEM)
Contributions: reporter experiments and analysis were performed by AP.

CHAPTER THREE: A High-throughput Approach for Elucidating CoF Recruitment to CREs

3.1 Abstract

Central to gene regulation is the recruitment of cofactors (CoFs, e.g., co-activators and co-repressors) to DNA by site-specific TFs. There are currently no high-throughput approaches to identify and characterize the many TF-CoF complexes simultaneously operating in a cell. To this end, we have developed the CoRec (Cofactor Recruitment) approach to monitor CoF recruitment by potentially hundreds of TFs from nuclear lysates, and to infer the identity of the DNA-bound TF. By using CoRec to assay TF-CoF complexes in resting and LPS-stimulated THP-1 macrophages, we recapitulated known complexes involved in macrophage development and activation, demonstrating the fidelity of the method. We compared TF-CoF complexes in resting and LPS-stimulated macrophages to TF-CoF complexes in resting and TCR-stimulated T cells, identifying complexes unique to each cell type, and complexes common to all cell types. Thus, we demonstrate that CoRec is a powerful approach to study the assembly and regulation of nuclear TF-CoF complexes in a cellular context.

3.2 Introduction

3.2.1 Motivation

Systems-level methods for monitoring changes in cell state have revolutionized analysis of cellular function and disease. Changes in cell state are catalyzed by changes in gene

expression driven by activation of transcription factors (TFs) and their subsequent recruitment of regulatory cofactors (CoFs) to specific genomic loci. These CoFs perform a wide range of functions, including histone modification, chromatin remodeling, and recruitment of general transcriptional machinery. Thus, delineating the TF-CoF complexes functioning in a cell is critical to understanding the control of gene expression in healthy and disease contexts. However, TF-CoF complexes are not routinely analyzed at a multiplexed level, leaving this central aspect of gene regulation understudied.

3.2.2 Current Methods for Examining TF-CoF Complex Formation and Recruitment to CREs

Traditional methods to monitor DNA-bound TF-CoF complex formation, such as gel-shift assays and protein-DNA pull down assays can only be performed in low-to-moderate throughput (~tens of factors per experiment). ChIP-seq provides a high-throughput method for identifying genome-wide TF and CoF recruitment; however, binding events are generally attributed to wide genomic regions on the order of 100s of bp. Thus, co-occupancy of TFs and CoFs at these regions does not identify TF-CoF complexes and further analysis must be performed to determine the interacting TFs and CoFs.

Yeast or mammalian two-hybrid (Y2H (Fields and Song, 1989) or M2H (Ravasi et al., 2010)) assays provide more direct approaches to identify interactions between protein pairs. The 2H approaches have been invaluable for mapping protein-protein interactions, but are limited to binary interactions and labor-intensive to conduct at a HT level in

different cell conditions. Immunoprecipitation (IP) followed by mass spectrometry has also been a useful strategy for identifying co-precipitating proteins (Wierer and Mann, 2016). The IP-mass spectroscopy approach with CoF-based precipitation followed by protein identification provides a way to identify TF-CoF complexes, but it does not explicitly assay DNA-bound complexes (Wierer and Mann, 2016). Moreover, current implementations of these approaches generally require high cell numbers and large amounts of starting material. Several HT methods have been described to characterize the presence of TFs in the nucleus (Mittler et al., 2008; Simicevic et al., 2013; Wei et al., 2018), providing protein-based characterization of the TFs functioning in cells. However, these assays are not designed to characterize TF-CoF complexes; therefore, there remains a need for HT methods that can characterize TF-CoF complexes in a cellular context.

3.3 The CoRec Approach

3.3.1 Overview of the CoRec Approach

In the CoRec assay, we monitor CoF recruitment to thousands of DNA sequences on protein-binding microarrays (PBMs), and infer the identity of the underlying TF-CoF complexes (Figure 3.1). PBMs are a high-throughput, microarray-based platform for measuring protein-DNA interactions (Berger et al., 2006). CoRec is an extension of our recently developed nextPBM platform for monitoring protein-DNA complexes from cell nuclear extracts (Mohaghegh et al., 2019); however, instead of monitoring the binding of TFs to DNA, we monitor the recruitment of CoFs to DNA by TFs. As nuclear extracts are

used in our assay, all CoFs and TFs present in the cell nucleus are available to bind to the DNA microarray probes.

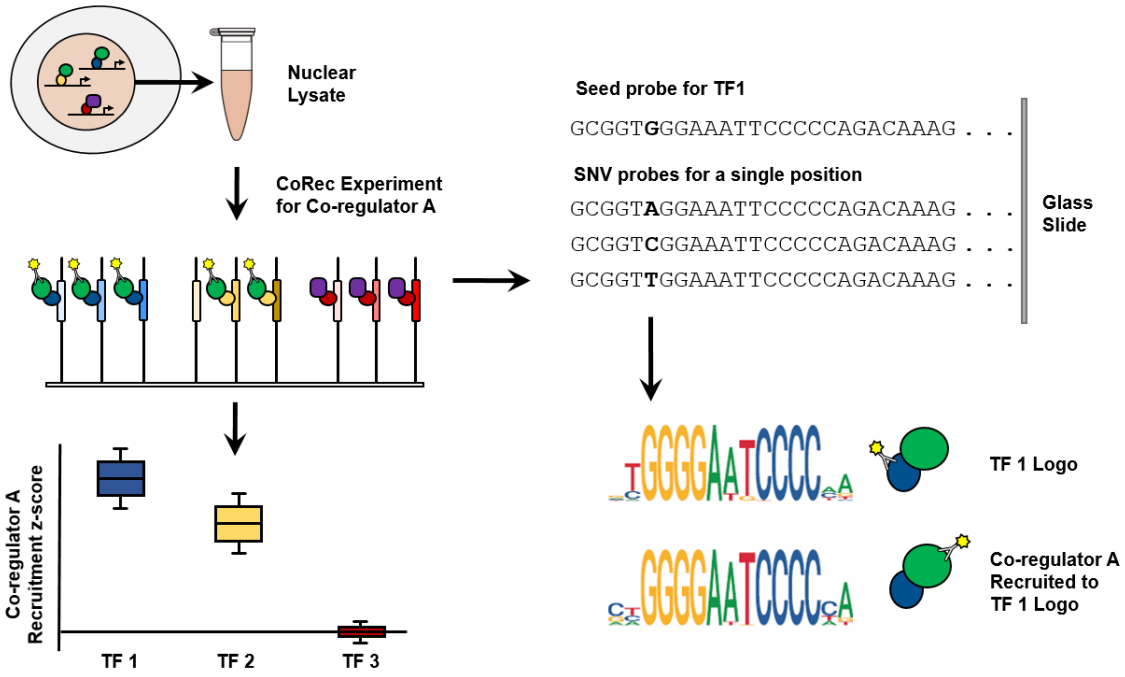


Figure 3.1: Schematic overview of the CoRec approach
Nuclear lysates are harvested from cells and incubated on a protein-binding microarray. Fluorescently labeled antibodies are used to detect indirect recruitment of CoFs to DNA probes containing response elements for a panel of transcription factors. Single nucleotide variants of these probes facilitate the generation of logos representing DNA determinants of CoF recruitment.

We use two approaches to infer the identity of DNA-bound TF-CoF complexes in our CoRec assay. First, we measure the CoF recruitment to high-affinity consensus binding sites chosen to identify different TFs (Figure 3.1). For example, recruitment of the CoF p300 to a consensus site for the TF NF-kB suggests the presence of a p300-NF-kB complex. Second, we determine CoF recruitment motifs that reveal the DNA-binding motif of the underlying TF, which can be matched to motif databases to infer the TF identity. CoF recruitment motifs are determined using a single-nucleotide variant (SNV)

approach (Andrilenas et al., 2018; Mohaghegh et al., 2019; Penvose et al., 2019) in which CoF recruitment is monitored to each TF consensus binding sequence, as well as to all SNVs of that sequence (each SNV sequence is a separate probe on the microarray), allowing a binding motif to be directly determined (Figure 3.1). Using this approach, a binding motif is determined that quantifies the binding specificity of the proteins bound to that *single seed* sequence. We have previously used this SNV approach to define TF binding motifs (i.e., logos), and have found that it can be used to define CoF recruitment motifs that match the motif of the underlying TF. Combining these two approaches, we infer the identity of a TF-CoF complex when the CoF is (1) recruited to the cognate TF binding sites and (2) the CoF recruitment motif matches to cognate TF motif. This approach, based on inferring the TF identity using DNA-binding specificity, is not able to discriminate between TFs that share similar DNA-binding specificity, but has the advantages that (1) many different CoF-TF complexes can be assayed using the same sequences, increasing the throughput of the assay, and (2) we can also establish the DNA binding specificity of the CoF-TF complexes, and how this may change across conditions.

To define the critical TF-CoF complexes that are coordinating the transcriptional response of a cell, we sought to characterize the TF-CoF complexes that involve broadly acting CoFs. We selected CoFs and CoF complexes that are known to interact with many different TFs and that play key regulatory functions in the transcriptional control of genes, such as the acetyltransferase and general activator p300, the deacetylase

NCOR/SMRT complexes, the MLL/COMPASS methytransferase complexes, and the SWI/SNF chromatin remodeling complex. By monitoring the recruitment of these broadly acting CoFs, we sought to define the profile of active TF-CoF complexes functioning in a cell.

3.3.2 Cell Types Used for CoRec Proof of Concept Experiments

As a proof of concept for the CoRec approach, we focused on the TF-CoF complex landscape of two key cell types of the immune system: macrophages and T cells. These cells are of particular interest both because of the roles they play in immunity, and because of their innate ability to respond to environmental stimuli. Both of these cell types are derived from hematopoietic stem cells, but diverge early in lineage commitment; macrophages are members of the myeloid lineage, whereas T cells derive from the lymphoid lineage (Chaplin, 2010). By utilizing two separate but related cell types, we can gain insight into conserved and cell-type specific regulatory strategies.

3.3.2.1 Macrophage Biology and Response to TLR4 Stimulation

Macrophages are cells of the innate immune system, responsible for detection and phagocytosis of pathogens. They are derived from monocytes that continuously circulate in the periphery, monitoring the presence of pathogen-associated signals including bacterial cell wall components, viral nucleic acids, and cytokines secreted by other pathogen-responsive cells. Upon encountering one of these signals, monocytes differentiate into macrophages and intravasate into the local tissue. Here, they continue to surveil the environment for a wide array of activating and inhibitory signals,

phagocytose pathogens, and secrete cytokines to coordinate the activities of local cells (Chaplin, 2010).

The complex set of genetic programs that underlie these behaviors is controlled by DNA-TF-CoF complexes, the composition and function of which are determined by the integration of a heterogeneous set of signals encountered by the cell. Here, we focus on the activation of surface receptor TLR4, which binds to the bacterial cell wall component lipopolysaccharide (LPS). Binding of LPS to TLR4 triggers a signaling cascade that modulates the ability of several key transcription factors to bind DNA response elements and recruit CoFs, including interferon regulatory factors (IRFs), activating protein 1 (AP-1) family members, and NF-kB (Tugal et al., 2013) (Figure 3.2).

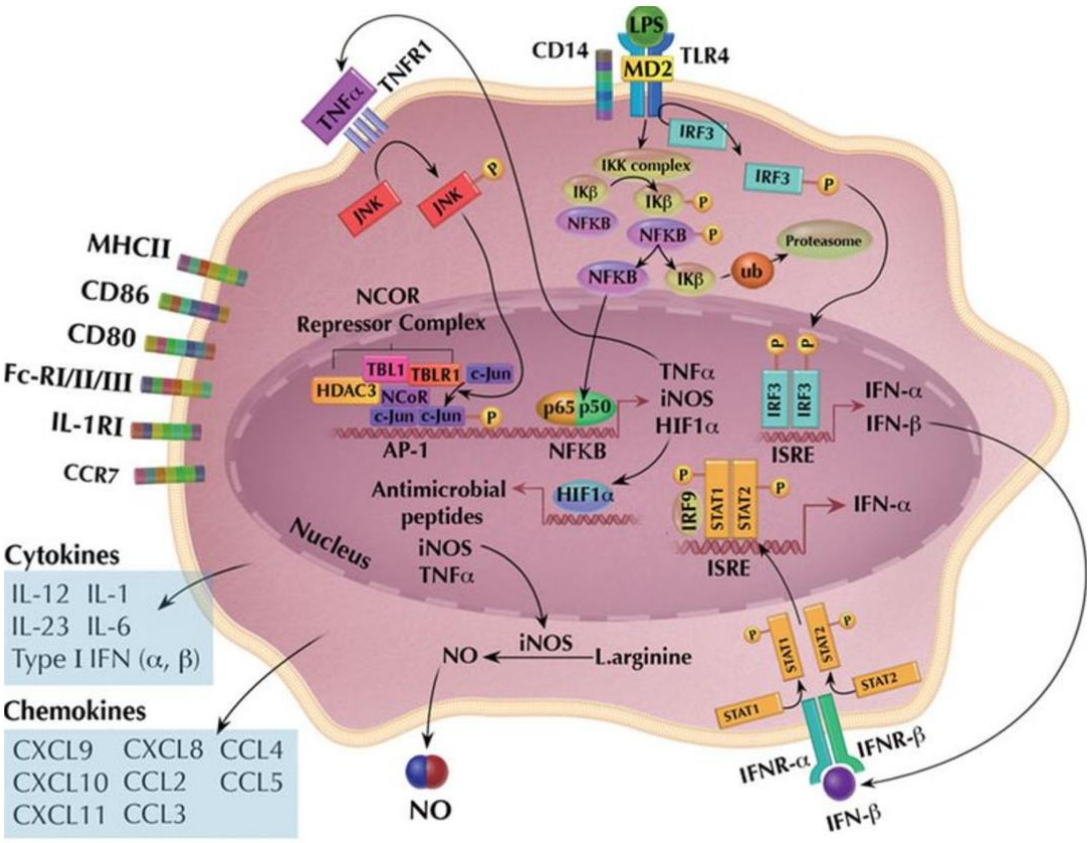


Figure 3.2: Schematic of LPS activation of macrophages
LPS binds to the TLR4 receptor on macrophages, triggering signaling cascades that activate transcription factors including NF-kB, IRF3, and AP-1. These factors then regulate the expression of chemokines and cytokines. This figure is taken from Tugal (2013).

IRF3 activation is induced upon TLR4 stimulation; prior to LPS activation of TLR4, IRF3 is sequestered in the cytoplasm, unable to coordinate gene expression. Activation of TLR4 leads to IRF3 phosphorylation and dimerization, as well as its translocation to the nucleus and recruitment of p300 to interferon-stimulated regulatory elements (ISREs), facilitating expression of IRF3 target genes (Tugal et al., 2013).

NF-kB is a dimeric transcription factor composed of different Rel family members including RelA (p65), RelB, C-Rel, p50, and p52. Complexes composed of different Rel proteins control distinct yet overlapping gene regulatory programs that drive many functions including inflammatory and stress responses (Kawai and Akira, 2007). The p65:p50 heterodimer is particularly important in bacterial response, and prior to TLR4 activation remains sequestered in the cytoplasm by inhibitory proteins (IkBs). LPS binding to TLR4 triggers a signaling cascade leading to the phosphorylation, ubiquitination, and degradation of IkBs, releasing NF-kB from its inhibition. NF-kB then translocates to the nucleus where it binds its cognate regulatory elements to effect gene regulation. In addition to the regulatory specificity achieved via specific NF-kB partners, PTMs further define NF-kB's ability to interact with CoFs proteins (Kawai and Akira, 2007).

TLR4 activation also stimulates activating protein 1 (AP-1) (Tugal et al., 2013). Like NF-kB, AP-1 is dimeric transcription factor than can be composed of many different subunits, most notably Jun and Fos. Canonically, prior to stimulation, Jun dimers bind AP-1 response elements bound to corepressors. TLR4 activation leads to Jun phosphorylation and a switch between Jun homodimers and Jun:Fos heterodimers. These heterodimers then recruit coactivators to initiate gene expression (Tugal et al., 2013).

In addition to these LPS-stimulated factors, many TFs play important roles in maintaining macrophage homeostasis under basal conditions. Members of the ETS

family, most notably PU.1/SPI1, as well as RUNX and C/EBP family members, play key roles in macrophage development and defining the basal chromatin landscape (Zhu et al., 2016). PU.1 is required for monocyte development. It can function as a monomer or dimerize with other TFs, including IRF8, which is also required for monocyte development. RUNX1 is another TF critical for hematopoiesis. It regulates PU.1 expression, and is dysregulated in acute myeloid leukemia (Zhu et al., 2016). Finally, C/EBP family members are also critical for monocyte development, and monocyte enhancers tend to be co-bound by PU.1 and C/EBPB (Zhu et al., 2016).

3.3.2.2 T cell Biology and Response to T Cell Receptor Stimulation

Like macrophages, T cells implement complex genetic regulatory programs in response to environmental stimuli to facilitate pathogen clearance (Smith-Garvin et al., 2009). Prior to activation, T cells remain in a quiescent state, monitoring their environment for activating signals. Unlike macrophages, T cell activation requires multiple signals to prevent spurious activation events. T cells express the T cell receptor (TCR), an integral membrane protein complex that binds a specific antigen. As the first signal of T cell activation, an antigen-presenting cell (APC) must present an antigen to which the TCR binds. Second, the APC must have encountered a pathogen itself, leading to the expression of surface proteins such as B7 on the APC's surface. These surface markers bind cognate receptors on the T cell surface, such as CD28 and CD2. These two events; interaction of the TCR with the APC-bound antigen and a costimulatory interaction, such as CD28 binding B7, are both necessary and sufficient for T cell activation (Smith-Garvin et al., 2009).

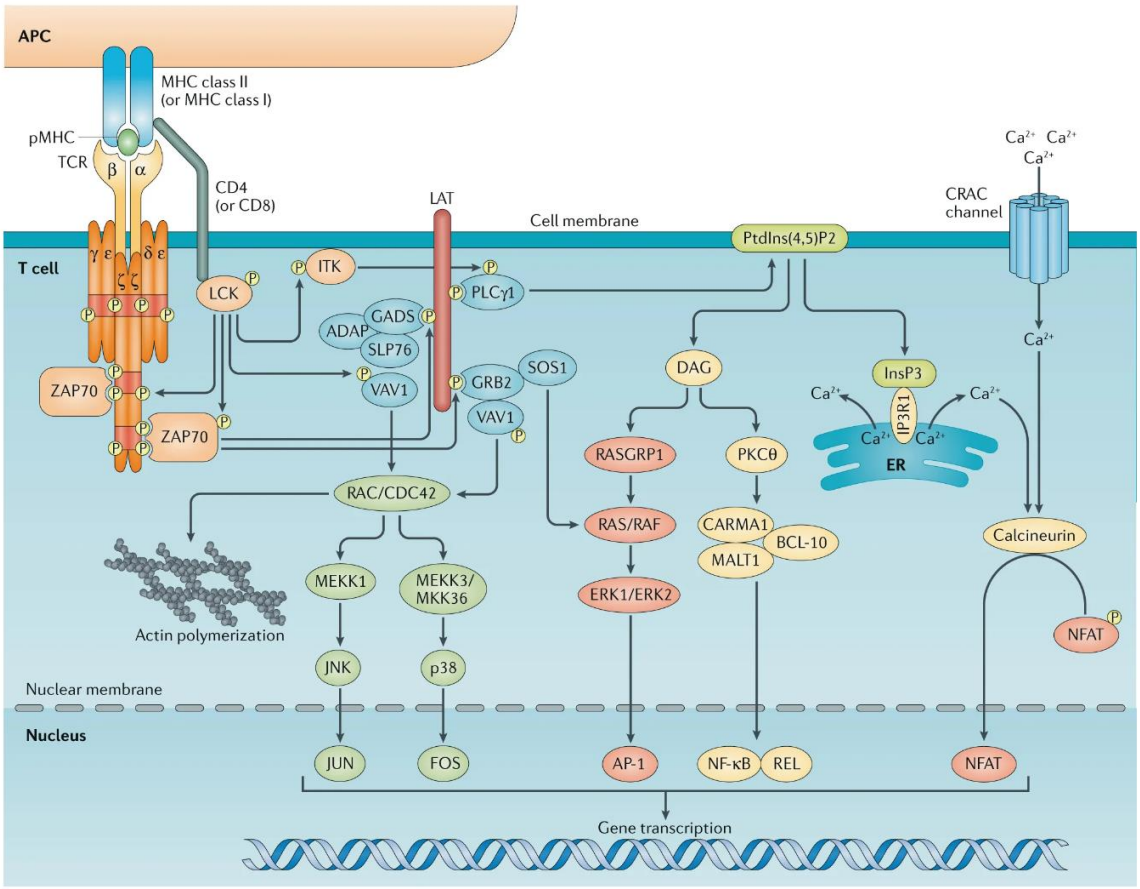


Figure 3.3: Schematic of TCR activation and signaling
TCR binding to antigen along with a costimulatory interaction (not shown) leads to signaling cascades that activate transcription factors including NF-kB and AP-1. This figure is taken from Gaud (2018).

TCR activation results in the stimulation of many of the same TFs as TLR4 activation in macrophages, including NF-kB and AP-1 (Gaud et al., 2018) (Figure 3.3). Similarly, T cells utilize many of the same families of basal TFs as macrophages, including ETS, RUNX, and C/EBP. As in macrophages, PU.1 is required for T cell development, however it is silenced in fully committed T cells (Mak et al., 2011). Other ETS family members play roles in T cell development, including SAP-1 and ETS-1 (Sharrocks, 2001). RUNX family members, required for hematopoiesis, and thus both macrophage

and T cell development, are also responsible for committing T cells to the cytotoxic T cell lineage (Naito et al., 2011). In contrast to the role C/EBP family members play in monocyte development, repression of C/EBPα is critical to for commitment to the T cell lineage (De Obaldia et al., 2013).

3.3.2.3 THP-1 and Jurkat Cell Lines as Models for Macrophages and T cells

In order to examine the CoF landscape of macrophages in basal and TLR4-stimulated conditions, we utilized the THP-1 human cell line, which is derived from peripheral monocytes of a patient with acute monocytic leukemia. We chose to examine a cell line rather than primary cells as it is a more time-efficient and cost-effective strategy for producing the sample quantity necessary for assay development and optimization; however, we anticipate this approach will work with primary cells as well. THP-1s are a standard model for monocytes and macrophages. THP-1 monocytes can be grown and expanded in suspension culture, and differentiated into macrophage-like cells by treatment with phorbol 12-myristate 13-acetate (PMA). Upon PMA treatment, THP-1s become adherent, take on a macrophage-like morphology, express macrophage surface markers, and display macrophage-like expression profiles (Park et al., 2007). Thus the ease of use and similarity of THP-1s to primary cells makes them an ideal cell line for this study.

Similarly, use of primary T cells for these preliminary studies would have been prohibitively expensive. Instead, we used the human Jurkat T cell line, derived from a patient with acute T cell leukemia. They are commonly used to study T cell signaling (Abraham and Weiss, 2004) and are easily expanded in culture, hence we use them for a

model of T cells in this study. To stimulate T cell activation in a reproducible manner, we utilize a mixture of monoclonal antibodies that bind to CD3 (a member of the TCR complex) to simulate antigen binding, and to CD28 and CD2 to simulate co-stimulation.

3.4 Results

3.4.1 Characterizing Regulatory Complexes in THP-1 Macrophages with CoRec

To survey the TF-CoF complexes active in human macrophages, we performed a CoRec experiment using nuclear lysates with unstimulated and LPS-stimulated THP-1 macrophages. We profiled recruitment of seven general CoFs representing a wide range of biological functions: P300 (acetyltransferase and general activator); SMARCA4/BRG1 (catalytic subunit of SWI/SNF remodeling complex); RBBP5 (conserved subunit of the SET1/MLL histone methyltransferase complexes); GPS2, NCOR1, and TBLR1 (three subunits of the NCOR repressor complex); and HDAC1 (catalytic subunit of various histone deacetylase complexes). As a control for stimulation, and to test the CoRec platform's ability to survey indirect TF recruitment, we also profiled binding of the NF-kB family member RelA/p65, which is strongly active in macrophages upon LPS stimulation. Binding of CoFs and RelA/p65 was monitored to 91 consensus sequences representing ~20 TF specificity groups.

For unstimulated THP-1 macrophages, we identified 18 unique DNA logos, 17 of which matched known TF motifs with q-values < 0.01. Strikingly, for all CoFs examined, we identified recruitment motifs that match known TF motifs, demonstrating that the CoRec

approach is likely generalizable to include an even wider panel of CoF proteins. In general, CoF motifs generated from the same consensus TF and SNV sequences are similar, suggesting that these different CoFs are recruited to these DNA sequences by common TFs. To simplify our illustration of measured TF-CoF interactions (Figure 3.4) we have selected a single representative motif for groups of similar motifs, and have used this representative in comparisons against TF motif databases (see Methods for details). In some cases, CoF motifs generated for the same DNA sequences differ, indicating that different TFs are recruiting the separate CoFs to the same sequences (discussed more below). Summarizing our results, we found 18 motifs representing 58 possible TF-CoF complexes in unstimulated macrophages. It is possible that similar motifs may actually represent the same TF-CoF complex; for example, there are several motifs that match Interferon Regulatory Factors (IRFs). Furthermore, a single motif may represent TF-CoF complexes with several TFs sharing a common DNA-binding specificity; for example, the ETS-type motifs could be bound by many ETS factors.

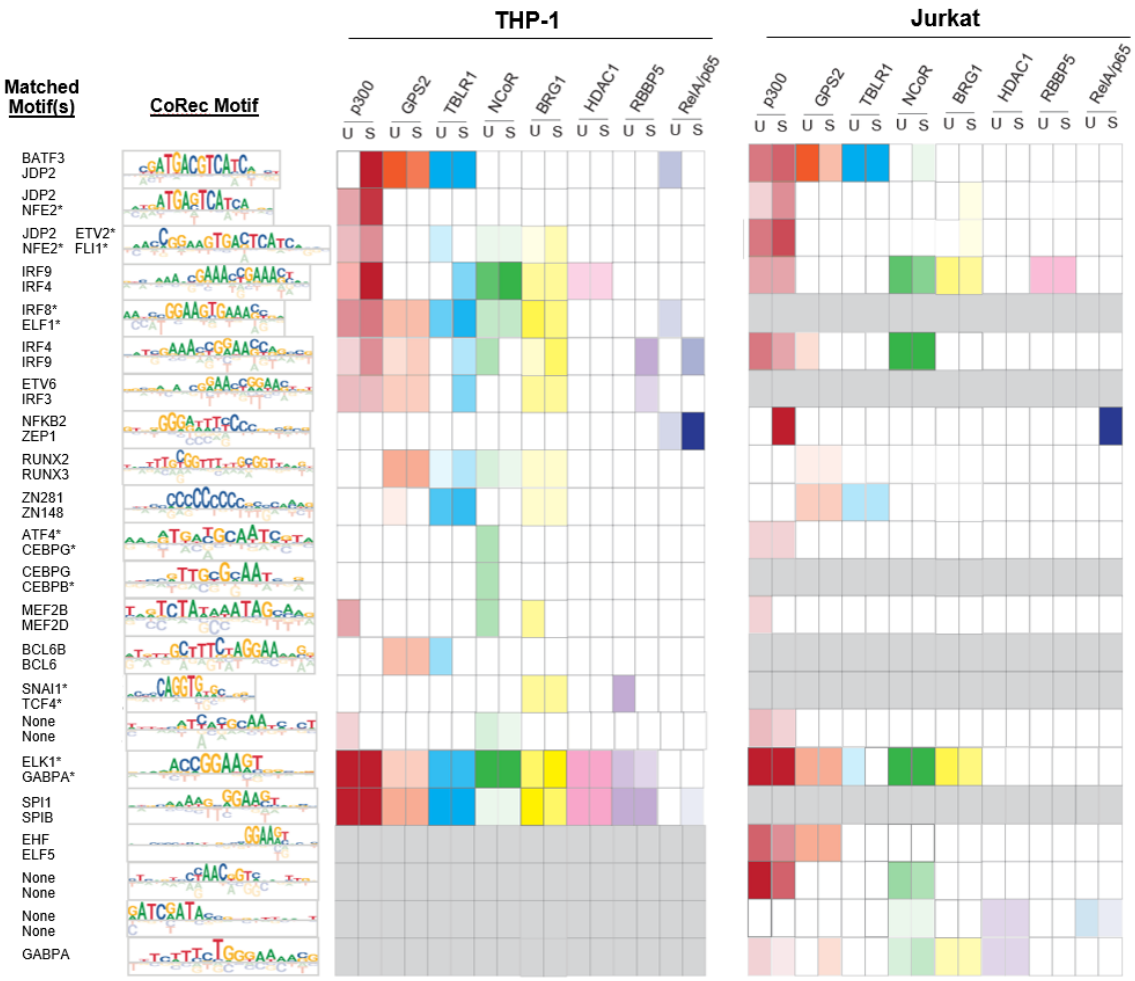


Figure 3.4: Summary of CoFs recruited to regulatory motifs in resting and stimulated THP-1 and Jurkat cells
CoRec motifs identified in THP-1 cells (left grid) and Jurkat cells (right grid) in unstimulated (U) and stimulated (S) conditions. Representative motifs are shown on the left with the two best TF matches from PWM databases (see Methods for details). All reported matches have q-values less than 10^{-3}, unless indicated with *, in which case the q-value is less than 10^{-2}. If a motif was identified for a given CoF, the corresponding box was shaded in a color specific to that CoF. The opacity of the box indicates the z-score of the seed sequence; a z-score of 1 corresponds to 10% opacity, a z-score of 2 corresponds to 20% opacity, etc. Z-scores greater than 10 are shown with 100% opacity.

Our CoF motifs identified many of the TF families that we expect to be functional in resting human macrophages, including ETS, RUNX, and C/EBP factors. We also identified motifs representing TF families more commonly associated with activated

macrophages, such as the IRFs, AP-1, and NF-kB (Figure 3.4). Two well-studied lineage factors that function to establish the enhancer chromatin landscape in macrophages are PU.1/SPI1 and C/EBPα (Garber et al., 2012; Natoli, 2010). Strikingly, we found all seven CoFs (p300, GPS2, BRG1, NCOR1, TBLR1, HDAC1, and RBBP5) are recruited to the ETS motif that matches the PU.1/SPI1, suggesting that PU.1 can form DNA-bound complexes with all of these diverse CoFs. In contrast, we found that only NCOR1 is recruited to the two motifs that match C/EBP family members, suggesting a possible distinction between these two lineage factors in terms of regulatory complexes. The RUNX factors have also been implicated in establishing the myeloid lineage identity, and we found that the RUNX factors recruit the NCOR1, GPS2, TBLR1, and BRG1 proteins. The lineage factor PU.1 also forms cooperative complexes with IRF8 in monocytes/macrophages, and binds to a composite 5'-GGAAnnGAAA-3' PU.1:IRF8 element (Mohaghegh et al., 2019). We found this composite element (labeled as IRF8/ELF1 motif in Figure 3.4) recruits p300, GPS2, BRG1, NCOR1, and TBLR1 (i.e., many of the same CoFs PU.1 itself recruited). This result demonstrates the sensitivity of the CoRec approach to assay more complicated arrangements of multiple, cooperatively binding TFs – that all these CoF motifs resemble the known composite element (and not the PU.1 site alone) indicates that their recruitment to these DNA sequences is sensitive to the cooperative interaction between PU.1 and IRF8. Finally, the other broadly recruiting motif in our analysis was a consensus ETS site (labeled as the ELK1/GABPA motif in Figure 3.4), which also recruits all 7 of the CoFs (p300, GPS2, BRG1, NCOR1, TBLR1, HDAC1, and RBBP5). This ETS motif is distinctly different from the

PU.1/SPI1-type motif based primarily on the CC-dinucleotide preference immediately 5-prime to the 5'-GGAA-3' ETS core, as has been previously described (Wei et al., 2010); therefore, this likely indicates robust CoF recruitment by ETS factors distinct from PU.1/SPI1.

3.4.2 CoF Complexes Change Upon LPS-stimulation of THP-1 Macrophages

To analyze how LPS-stimulation alters the landscape of TF-CoF complexes in macrophages, we performed CoRec using macrophages stimulated with LPS for 45 min. We found many of the TF-CoF relationships are maintained, for example the ETS motifs (PU.1 and ELK-type) remain broadly recruiting; however, there are also clear differences. A number of TF-CoF complexes identified in resting macrophages are no longer present in our assay for the stimulated macrophages, such as complexes involving MEF2 and C/EBP family TFs. In contrast, other TF-CoF relationships appear in stimulated macrophages. For example, with LPS stimulation Rbbp5 was recruited to motifs for IRF4/IRF9 and ETV6/IRF3.

TF-CoF changes between experimental conditions can also be examined using the magnitude of our CoF recruitment signal for individual seeds, which would result from either more TF or CoF available to form complexes, or an enhancement of the TF-CoF interaction strength (perhaps due to PTMs). In Figure 3.5 (left-hand panels), we illustrate the magnitude of p300 recruitment signal to each binding site (consensus TF seed and

SNV sequences) grouped according to cognate TF family in THP-1s. In unstimulated THP-1s, p300 was highly recruited to response elements for the ETS factors PU.1 and ELK, consistent with their known roles in defining macrophage cell fate. We also observed low levels of p300 recruitment to ISREs, which likely resulted from the PMA treatment used to differentiate THP-1 monocytes into macrophages. Upon LPS stimulation, we observed similar recruitment levels to ETS response elements. In contrast, p300 was more highly recruited to ISREs and bZIP dimer response elements, including AP-1 response elements and CREs, consistent with the role of these LPS-inducible factors in inflammatory response.

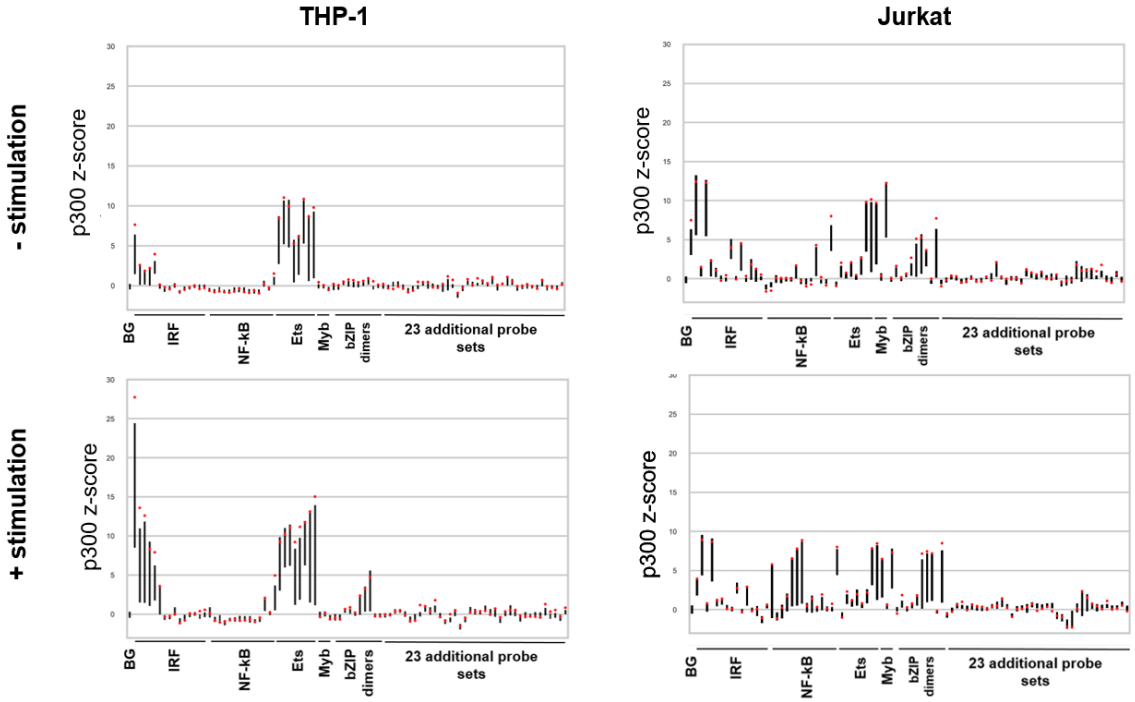


Figure 3.5: P300 recruitment in resting and stimulated THP-1 and Jurkat cells
Z-scores are shown for seeds (red dots) and the Q1-Q3 range for the SNVs for each seed (black lines). Seed-SNV sets are grouped by the type of TF they were designed to bind. The additional probe sets include probes designed for TFs including RUNX factors, STATs, nuclear receptors, NFAT, BCL6, MEFs, and HIF.

3.4.3 Indirect Recruitment of TFs to CREs

In addition to CoF recruitment, TFs themselves can also be indirectly recruited by other TFs to DNA. This has been specifically described for the NF-kB family member RelA/p65 in macrophages, where it can be recruited by IRF dimers to ISREs (Ogawa et al., 2005). CoRec provides a natural platform to investigate the ability of TFs to indirectly recruit other TFs to DNA. To test this, we examined RelA/p65 binding and recruitment in macrophages and found direct DNA binding to NF-kB sites, but also indirect recruitment to the PU.1/SPI1 in stimulated conditions, IRF-type motifs in both conditions, and to BATF3/JDP2 motifs in unstimulated conditions. These results confirm the observation the RelA/p65 can be recruited indirectly to IRF-type sites in macrophages, and identified several new recruiting TFs.

3.4.4 Comparison of Regulatory Complexes in Macrophages and T cells

To examine the cell-type dependence of TF-CoF complexes, we performed CoRec using resting and TCR-stimulated Jurkat T cells (Figure 3.4, right grid). We identified 16 CoF recruitment motifs, 13 of which matched known PWMs (q-value < 0.01). These motifs are consistent with response elements for ETS, RUNX, C/EBP, NF-kB, IRF, CREB, and AP-1 family members. The remaining three unmatched logos may represent previously unidentified response elements (Figure 3.4, right grid).

Similar to THP-1s, p300 is recruited to ELK, MEF, AP-1, and IRF response elements in unstimulated Jurkats. However, we also observed recruitment to response elements for CREB, C/EBP, GABPA, and EHF. Upon TCR stimulation, we observed a large increase

in p300 recruitment to NF-kB response elements, as well as to AP-1 response elements (Figure 3.5, right hand panels, AP-1 elements are contained in the bZIP dimer group), which is consistent with known TCR-activation induced TFs. Furthermore, we did not observe the highly increased p300 recruitment to the IRF family sites that we do for the macrophages, consistent with IRF factors not being activated in response to TCR activation. In Jurkat cells, we also saw broad CoF recruitment to ETS sites with a 5' CC flank. ETS factors are widely expressed in all cell types, and ETS TFs that bind to the more canonical 5'-CCGGAA-3' sites are the largest class of ETS TFs (Wei et al., 2010), many of which are known to bind to promoters and regulate house-keeping genes (Curina et al., 2017). Therefore, our results are consistent with ETS factors that bind to the canonical ETS site being potent CoF recruiters in both cell types. In contrast, we saw no recruitment to the SPI1/PU.1 type site in Jurkat cells, consistent with the role this factor plays in macrophage-specific development. We did however, observe p300 and GPS recruitment to a 5'-GGAA-3' lacking the 5' CC flank in Jurkats under both conditions. This site also lacks the 5' A/G flank characteristic of SPI1/PU.1, suggesting another ETS factor is recruiting these CoFs in Jurkats. Overall, these results demonstrate that the CoRec approach can identify both cell- and stimulus-type specific TF-CoF complexes.

3.4.5 CoF Recruitment to Interferon-Stimulated Regulatory Elements in THP-1s

IRFs play key roles in macrophage development and response to environmental stimuli. For example, IRF8 expression is highly specific to hematopoietic cells and is necessary for monocyte and macrophage development (Zhu et al., 2016). Other IRF family members, notably IRF3, are also critical for regulating macrophage response to

environmental stimuli, including bacteria and viruses (Tugal et al., 2013). Our initial CoRec analysis identified a number of CoF recruitment motifs that resembled a number of similar, yet distinguishable, IRF-family motifs. These results suggest that many diverse TF-CoF complexes may bind to single DNA sequences, complicating our understanding of how TF binding sites are used in vivo. To examine this complexity, we analyzed the CoF recruitment motifs generated for a single consensus interferon-stimulated response element (ISRE) seed-SNV set that can be bound with high affinity by a range of IRF complexes.

In unstimulated macrophages, recruitment of NCOR, RBBP5, P300, BRG1, and GPS2 to the ISRE site resulted in two classes of binding motifs (Figure 3.6). NCOR and RBBP5 motifs showed the canonical GAAAnnGAAA-type ISRE that matches binding motifs for many IRF family members (IRF9, 4, 8, 3, 5, 7, 1, 2, q-values < 1 x 10^{-3}), suggesting IRFs are responsible for their recruitment to this seed sequence. In contrast, for p300 and BRG1 we observed a variant motif that differs within the half-sites: G<u>G</u>AACCG<u>G</u>AA. GGAA is the consensus motif for the ETS family of factors, which plays crucial roles in macrophage development and can heterodimerize with IRFs (Zhu et al., 2016). Thus, under unstimulated conditions, p300 and BRG1 are preferentially recruited to IRF/ETS composite element variants of this ISRE, while NCOR and RBBP5 are preferentially recruited to the consensus ISRE.

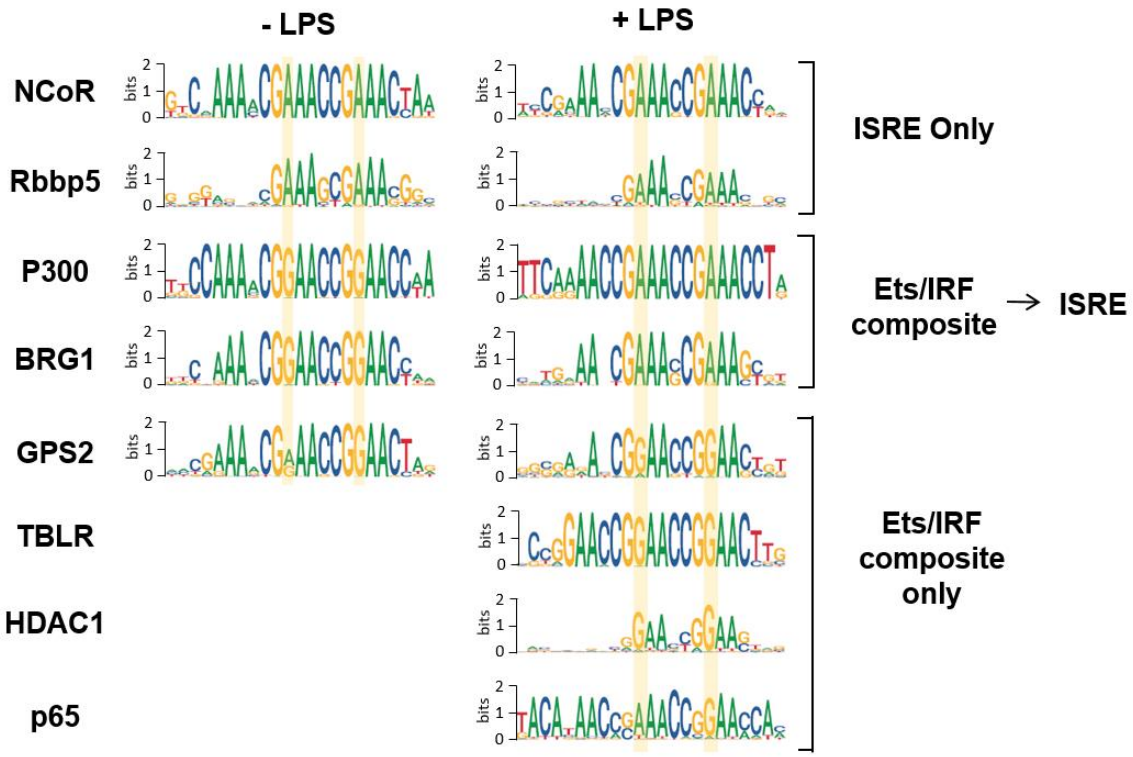


Figure 3.6: CoF logos for a single ISRE seed and SNV set
Logos are shown for THP-1 cells with and without LPS stimulation. Positions of interest are highlighted in yellow.

Upon LPS stimulation, NCOR and RBBP5 maintain this ISRE preference, though the strength of NCOR binding increases (z-score 11.8 in unstimulated cells compared to 14.7 in stimulated cells) and the strength of the RBBP5 interaction decreases (z-score 14.8 in unstimulated cells compared to 12.7 in stimulated cells). In addition to RBBP5 and NCOR, p300 and BRG1 have a preference for the ISRE upon LPS stimulation and bind with high affinity (z-score 27.7 for p300 (maximum 31.1) and z-score 8.4 for BRG1 (maximum 16.5)). For both p300 and BRG1, this represents a large increase in recruitment to the seed sequence compared to unstimulated cells (unstimulated z-score

7.7 for p300 and 4.4 for BRG1). These results are consistent with studies that have shown many IRF family members (including IRF1, 2, 3, 5, and 7) utilize p300 as a transcriptional activator, and that BRG1 is required for the induction of many LPS-inducible genes in macrophages (Ramirez-Carrozzi et al., 2009).

Finally, several recruited factors did not show a preference for the ISRE compared to the composite variants under either condition. We observed a preference for GPS2 recruitment to ETS/IRF composite sites under both conditions and a preference for TBLR and HDAC1 for composite sites upon stimulation only. We also observed a composite logo for p65 upon stimulation, suggesting this TF is indirectly recruited to DNA via the ISRE or its ETS/IRF composite variant: GAAACCGGAA.

These results suggest cell-state and CoF-specific regulatory strategies. For example, NCOR and RBBP5 were preferentially recruited to this ISRE in both unstimulated and stimulated cells, whereas BRG1 and p300 binding switched from favoring ETS/IRF composite variants without LPS stimulation to favoring the ISRE upon stimulation. Thus, we identify at least three different regulatory models: a constitutive preference for the ISRE, a switch from composite sites to the ISRE, and a preference for composite sites only.

3.4.6 CoF Recruitment to cAMP Response Elements and AP-1 Regulatory Elements in THP-1 and Jurkat Cells

Next, we examined CoF recruitment to AP-1 response elements in THP-1 macrophages and Jurkat T cells. AP-1 is a dimeric transcription factor that can be composed of many different partners, including Jun, Fos, ATF, and JDP family members. The canonical AP-1 consensus binding site is the palindromic sequence TGA[C|G]TCA; however, dimers composed of different AP-1 family members can have different DNA-binding preferences (Rodríguez-Martínez et al., 2017). Moreover, AP-1 proteins are part of the larger group of basic leucine zipper (bZIP) transcription factors, and can dimerize with many other bZIP proteins, such as C/EBP, to achieve additional regulatory complexity. Dimers composed of different bZIP family members may have DNA-binding preferences that reflect the preferences of each partner, or may have entirely different, emergent preferences (Rodríguez-Martínez et al., 2017). Within the AP-1 family, different partners can also have different functionalities; for example, ATF3:Jun dimers promote transcriptional activation, whereas ATF3 homodimers act as repressors (Hsu et al., 1992). Thus, DNA-binding specificity provides a mechanism to impart functional specificity for these factors.

To better understand AP-1 regulation in T cells and macrophages, we utilized seed-SNV set with a seed containing the consensus AP-1 binding site. Of the eight CoFs assayed, we found that only p300 yields logos consistent with the AP-1 motif. P300 recruitment to this site is induced by both LPS stimulation in THP-1 cells (3.1-fold induction) and TCR activation in Jurkats (1.4-fold induction). By examining the binding logos

generated from this seed sequence and its SNV variants, we gained further insight into the DNA-sequence characteristics that contributed to its regulation (Figure 3.7a). In both conditions and both cell types, p300 showed a strong preference for the consensus core element. In stimulated THP-1 cells and Jurkats, p300 prefers an additional T downstream of the core binding site. Interestingly, however, p300 also shows a preference for a 3' G rather than T in unstimulated Jurkats, and a preference for either a 3' G or T in unstimulated THP-1s. The highest scoring matches for this extended TGACTCAG site are MAF family members (q-value < 1 x 10^{-4}), which can heterodimerize with other AP-1 subunits. Thus, a MAF factor may be responsible for recruiting p300 under unstimulated conditions. Overall, this demonstrates that p300 is preferentially recruited to AP-1 sites with a 3' T flank in stimulated THP-1s and Jurkats, compared to a 3' G flank under unstimulated conditions.

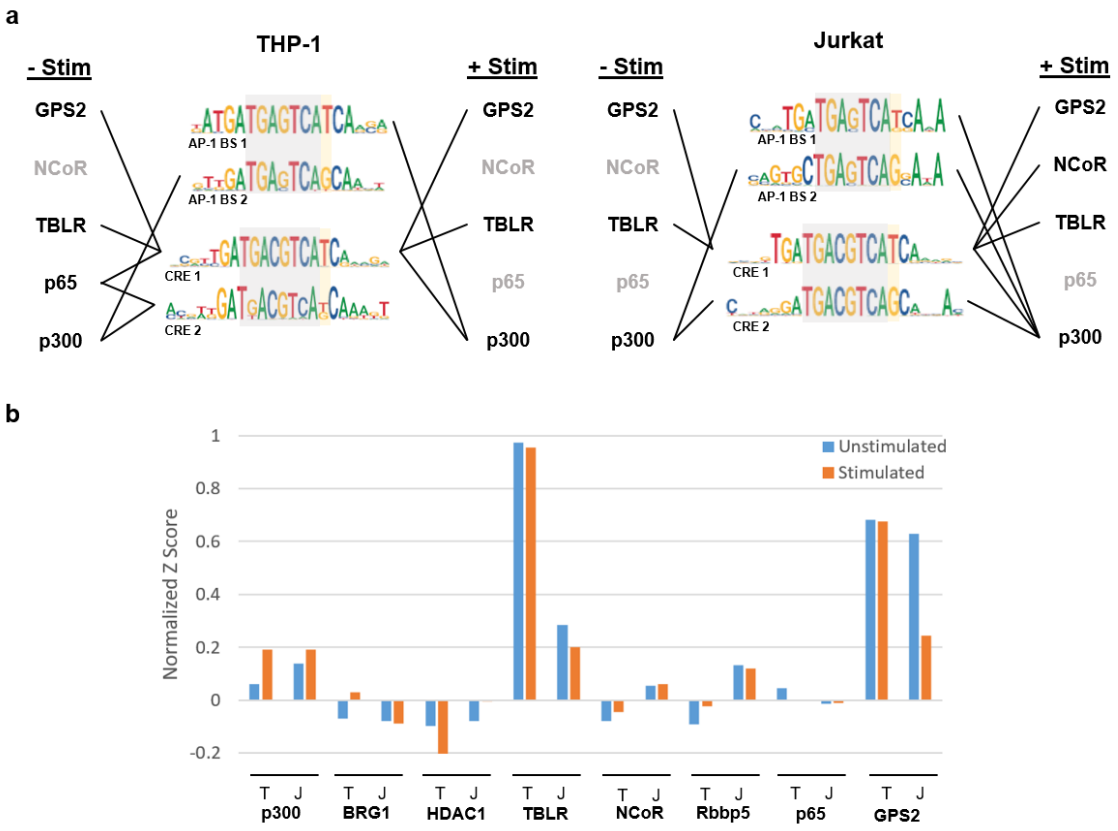


Figure 3.7: CoF recruitment to CRE and AP-1 response elements
a Diagram of representative logos generated from AP-1 and CRE seeds and CoFs that generate these logos in THP-1 and Jurkat cells with and without stimulation. Core response elements are shaded in gray and positions of interest are highlighted in yellow. **b** Normalized z-scores of CoFs recruited to the same CRE seed for unstimulated and stimulated THP-1 cells (T) and Jurkats (J). Z-scores were normalized to the maximum value measured for each CoF.

In addition to binding the canonical AP-1 response element, many AP-1 family members can also bind the cAMP response element (CRE), 5'-TGACGTCA-3', which utilizes adjacent (rather than overlapping, as for the AP-1 binding site) inverted repeats of the half site 5'-TGAC-3'. To better understand the regulatory mechanisms of these related elements, we examined CoF recruitment to CREs in unstimulated and stimulated THP-1 cells and Jurkats.

In both THP-1 cells and Jurkat cells, we found that a SNV seed containing the consensus CRE site produced CRE logos for TBLR, NCOR, GPS2, and p300 in both stimulated and unstimulated conditions, representing an expanded repertoire of CoF interactions compared to the AP-1 site. In both cell types, p300 recruitment increased upon stimulation (Figure 3.7b). TBLR was highly recruited to this seed in both cell types and under both conditions; this seed is one of the highest TBLR-recruiting sequences on the array. GPS2 was also strongly recruited to this sequence in both cell types and conditions and it recruitment increased slightly upon stimulation in THP-1 cells, but decreases upon stimulation in Jurkats. Although GPS2 and TBLR can form a complex with NCOR, NCOR levels are much lower and are only greater than the mean of the background distribution in Jurkats.

Examining the SNV logos, we found that TBLR, NCOR, GPS2, and p300 all showed similar preferences for the consensus sequence within the core CRE site (Figure 3.7a). All of these factors also showed some preference for an additional TGA (or GA) upstream of the core sequence, and in stimulated conditions in both cell types, these factors showed a preference for an additional TCA downstream of the core sequence. Interestingly, in unstimulated conditions, p300 preferred a 3' G over a 3' T, as was observed for the AP-1 response element. In contrast, TBLR, NCOR, and GPS2, preferred a 3' T unstimulated conditions. Thus, p300 has a regulatory strategy unique from all other CoFs assayed; it was preferentially recruited to TGACGTCAG prior to

stimulation in both Jurkats and THP-1s, but switches to TGACGTCAT sites upon stimulation, while NCOR, TBLR, and GPS2 bind TGACGTCAT in both conditions.

RBBP5 is moderately recruited to this seed in Jurkats, however the logo it produces has low information content and does not match any known TF motifs. Thus, it seems likely that this recruitment is actually nonspecific, perhaps via electrostatic or other interactions with the DNA. This example highlights the benefit of using an approach that directly measures binding to each SNV; if we had measured recruitment to only the seed sequence (as would commonly be done in EMSAs or DNA pulldown assays), we may have incorrectly drawn the conclusion that RBBP5 is recruited to this CRE site in a sequence-specific manner.

While CoF recruitment patterns to this sequence are largely similar between Jurkats and THP-1s, we observe several differences. First, NCOR yielded a CRE logo upon stimulation in Jurkat cells, however it did not in THP-1s. In both cell types, NCOR signals were relatively low for this site, so it may be that the signal is close to the limit of detection of this assay, allowing small changes in concentration to affect which logos are observed.

Unexpectedly, we observed indirect recruitment of p65 to CRE sites in unstimulated THP-1s. This result is particularly surprising because nuclear p65 levels are much higher in LPS-stimulated cells than untreated cells, suggesting this interaction results from

additional regulation beyond the presence of p65 (such as PTMs). We were unable to find evidence of this interaction in the literature, suggesting we have identified a novel interaction that is both cell-type and stimulus specific.

3.5 Discussion

Here, we describe CoRec, a novel approach for the HT characterization of DNA-bound TF-CoF complexes operating in a cell. By assaying TF-CoF complexes in resting and LPS-stimulated THP-1 macrophages, we recapitulate known complexes involved in macrophage development and activation, demonstrating the fidelity of the method. We compare TF-CoF complexes in resting and LPS-stimulated macrophages to TF-CoF complexes in resting and TCR-stimulated T cells, identifying complexes unique to each cell type, and complexes common to all examined cell types. Thus, we show that CoRec characterizes cell-type and stimulus-specific TF-CoF complexes in a HT manner.

While we recapitulate many known TF-CoF interactions, we also identify novel interactions, which merit further study. For example, we find a RUNX-GPS2 complex forms in macrophages and T cells under all conditions assayed; however, we were unable to find evidence of this interaction in the literature. Thus, CoRec may be used for discovering protein complexes, and it is likely that a more comprehensive comparison between our CoRec results and known interactions will yield more candidates for further study.

We also demonstrate the ability of CoRec to identify the indirect recruitment of TFs to CREs. Although p65 is well-known to directly bind DNA via NF-kB response elements, it can also be indirectly recruited to ISREs via IRF3 to facilitate transcription (Ogawa et al., 2005). In this study, we also find that p65 can be recruited to motifs that match IRF motifs. In addition, we find indirect recruitment of p65 to motifs representative of other TFs, including PU.1 and AP-1 family members. We were unable to find examples of p65 binding to PU.1 in the literature; however, p65 has been shown to directly interact with Fos and Jun. Previous work has shown that Fos:Jun is indirectly recruited to p65:p50 at NF-kB response elements in the long terminal repeat (LTR) that controls HIV-1 expression (Yang et al., 1999). Our results demonstrate the converse; that p65 can be indirectly recruited to AP-1 sites, suggesting a greater crosstalk between NF-kB and AP-1 signaling pathways than previously known. Similarly, indirect recruitment of p65 to PU.1 may be of biological significance, and merits further study.

Our results suggest that CoRec can also be used to identify novel regulatory elements. Most of the identified CoF motifs can be matched to known TF motifs, suggesting our assay accurately reflects CoF-TF interactions. However, three motifs match no known TF motif, suggesting these may represent novel CREs and merit further study. DNA pulldown assays using probes with these consensus sequences followed by mass spectrometry could be used to identify the DNA-interacting factor. Alternatively, it may simply be that similar motifs have been observed previously, but are not present in the databases used for our comparisons. While we used several well-established and

extensive motif databases (see Methods), comparison to additional databases may elucidate the DNA-interacting TF.

Although our results recapitulate many expected TF-CoF interactions, we note that we do not observe p300 motifs consistent with recruitment to NF-kB binding sites in LPS-stimulated macrophages, despite that NF-kB is highly upregulated in this condition (as assessed by western blot and PBM-binding; data not shown) and the NF-kB-p300 interaction is well-documented (Mukherjee et al., 2013). We propose two explanations for this discrepancy. First, the lack of observed p65-p300 interaction may accurately reflect the TF-CoF complexes utilized in these cells. NF-kB activation is relatively fast, occurring on the order of 10s of minutes (Hoberg et al., 2005). Preparation of nuclear lysates similarly takes place on the order of 10s of minutes, thus it may be that unintentional inconsistencies in the time it takes to prepare the lysates result in observing a later phase of LPS stimulation, in which p300 is preferentially recruited to AP-1 rather than NF-kB.

Alternatively, it may be that the p65-p300 interaction is below the limit of detection of our assay. We note that we have observed p300 motifs that closely match NF-kB binding sites in preliminary CoRec experiments not described here. However, these experiments were performed with a higher concentration of nuclear lysate. In preliminary experiments that used lower lysate concentrations, the p65-p300 interaction was not observed, whereas stronger interactions (such as IRF-p300), were retained. The

concentration of lysate used in a CoRec experiment is currently limited by the NaCl concentration in the sample. Nuclear lysate extraction requires high NaCl concentration (420 mM), which is not conducive to mimicking the NaCl concentration in the nucleus. Thus, we dilute the lysates so that they do not exceed 110 mM NaCl in the PBM, which also limits the lysate concentration that can be used for the PBM experiment. To increase the limit of detection of this assay, our lab is working on CoRec experiments that utilize in vitro expressed and purified peptides from specific p300 domains. These purified peptides allow us to increase the concentration of CoF domains of interest to elucidate interactions below the current limit of detection.

CoRec uses a seed and SNV approach to derive models for CoF recruitment to DNA, allowing us to compare the resulting motif to known TF motifs. We have found this approach to be invaluable for accurately identifying the DNA-bound TFs compared to methods that rely on a consensus DNA sequence alone. For example, for several seeds containing NF-kB response elements, if we examined CoF recruitment only to the seed in Jurkat T cells, we would have concluded that these response elements recruit p65-p300 upon TCR stimulation and p65-TBLR1 in both stimulated and unstimulated cells. However, by examining the CoF motifs, we observe that the TBLR1 interaction depends only on a G repeat that is part of the NF-kB consensus sequence. This G repeat is consistent with motifs for ZNF281 and ZNF148, suggesting that TBLR1 recruitment to these seeds is not actually dependent on NF-kB. Had we assumed CoF recruitment was

mediated by the TF for which the sequence was intended, we would have mischaracterized this interaction.

This study characterizes TF-CoF complexes in two cell types and two treatment conditions per cell type, but we anticipate this approach will be widely extensible to other cell types and stimuli. Moreover, coupling CoRec with other methods could provide a deeper understanding of CoF complex formation in many cell types and conditions. For example, performing mass spectrometry or phosphoproteomic analysis of CREs characterized by CoRec could identify additional members of regulatory complexes and assess the role of phosphorylation state on complex formation. By coupling CoRec with HT methods that assay the ability of response elements to activate expression, such as massively parallel reporter assays (Melnikov et al., 2012), we could develop a pipeline for understanding the relationship between DNA-binding, TF-CoF recruitment, and gene expression. Overall, we anticipate CoRec will be an invaluable platform for characterizing the many TF-CoF complexes operating in a cell and further understanding the molecular mechanisms that regulate gene expression.

3.6 Methods

3.6.1 Tissue Culture and Stimulation

THP-1 human monocytes (ATCC TIB-202) were cultured in RPMI-1640 with 10% FBS and 1mM sodium pyruvate in a 37C incubator with 5% CO_2. For each treatment condition, three 30 mL cultures at ~8 x 10^5 cells/mL were differentiated into

macrophages by incubation with 25 ng/mL phorbol 12-myristate 13-acetate for 96 hours. After 96 hours, cells were washed twice with PBS and fresh media was added. Cells were allowed to rest for an additional 48 hr. For LPS treatment, 1 μg/mL lipopolysaccharide (Sigma L3024) was added to the PMA-differentiated THP-1s for 45 min prior to preparation of nuclear lysates.

Jurkat human T cells (ATCC TIB-152) were cultured in RPMI-1640 with 10% FBS in a 37C incubator with 5% CO_2. For each treatment condition, four 30 mL cultures at 1 x 10^6 cells/mL were used. TCR stimulation was performed by adding ImmunoCult Human CD3/CD28/CD2 T cell Activator (STEMCELL Technologies) to cultures according to the manufacturer's instructions (25 μL/mL culture) for 45 min prior to nuclear lysate preparation.

Preliminary experiments were performed to optimize incubation times with LPS or ImmunoCult. Based on western blot analyses of p65 in nuclear lysates prepared as described above, nuclear p65 concentration was highest at 45 min in both cell types.

3.6.2 Nuclear Lysate Preparation

Jurkat and THP-1 cells were placed on ice immediately after the 45 min stimulation period to minimize additional changes due to stimulation. Jurkats were washed by centrifuging at 500xg for 5 min, aspirating the supernatant, adding 25 mL ice-cold PBS, and repeating the centrifugation and aspiration. THP-1 cells were washed twice with ice-cold PBS, scraped, pelleted at 500xg for 5 min, and the supernatant was aspirated. 1 mL

of Buffer A (10 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl, 1:1000 protease inhibitor cocktail (Sigma P8340), and 0.5 mM DTT) was added to the cell pellets, and they were incubated on ice for 10 min. 20 μL of 10% Igepal was added to the solution, and vortexed for 10 s. Trypan blue staining was performed to check for successful cell lysis and intact nuclei. Nuclei were pelleted at 4C for 5 min at 500xg. The supernatant containing the cytosolic fraction was aspirated, and the pellet was washed with 500 μL Buffer A to remove remaining cytosolic components, centrifuged again as above, and the supernatant removed. To make Buffer C, Buffer B (20 mM HEPES pH 7.9, 25% glycerol, 1.5 mM $MgCl_2$, 10 mM KCl, 1:1000 protease inhibitor cocktail (Sigma P8340), and 0.5 mM DTT) was diluted to a final concentration of 420 mM NaCl with a 3 M NaCl stock solution. 100 μL of Buffer C was added to the pelleted nuclei and the suspension was vortexed for 30 s prior to incubating for 60 min at 4C in a Hula mixer with settings: orbital: 25/off, reciprocal: 90/30, vibro: 5/5. Insoluble components were pelleted at 21,130xg for 20 min. The supernatant was removed, aliquotted into single use samples, flash frozen, and stored at -80C.

3.6.3 PBM Probe Design

PBM experiments were performed using a custom-designed microarray (Agilent Technologies, Inc., Design ID 086002, 8 × 60k format). Starting from the slide surface, PBM probes contain a 24 nt constant primer region, a 34 nt variable region, and a 5’ GC dinucleotide cap. For each probe sequence, five replicate probes were included with the 34 nt variable region in each orientation with respect to the slide surface (10 probes per unique sequence).

SNV probes: Seed sequences were chosen based on sequences that bound well in previously-performed PBM experiments, response elements identified from the literature, and previously-reported PWMs. For most probes, an ACGT tetramer was placed upstream of the binding site to provide extra space between the binding site and free end of the probe. If extra 3' sequence was required to fill out the 34 nt variable region, the additional sequence was taken from the 5' end of a probe that bound poorly in preliminary CoRec experiments. For each seed sequence, SNV probes were created that had a single-nucleotide variant at each position of the intended binding site as well as several nt, usually 4, upstream and downstream of the binding site.

Random genomic probes: 34 nt regions were randomly chosen from the UCSC hg19 build of human genome. Sequences were removed that contained Ns or single-nucleotide repeats longer than three nucleotides.

3.6.4 PBM Experiments and Analysis

Microarrays were double-stranded as previously described (PBM double-stranding primer 5'- CAGCAGCGTCAAGCGAATCAAGAC-3') (Berger et al., 2006). All washes were performed in coplin jars on an orbital shaker at 125 rpm. Double-stranded microarrays were first pre-wetted by washing in HBS (20 mM HEPES, 137 mM NaCl, 1.5 mM $MgCl_2$) containing 0.01% Triton X-100 for 5 min. They were then rinsed in an HBS bath and blocked with 2% milk in HBS for 1 h. All milk stock solutions were prepared from nonfat dehydrated milk (Fisher NC9121673) and centrifuged for 10 min at 20,000xg. The supernatants were then filtered through a 0.2 um filter. After blocking,

arrays were washed in HBS containing 0.1% Tween-20 for 5 min, then in HBS containing 0.01% Triton X-100 for 2 min, and finally rinsed in an HBS bath. Nuclear lysates were then incubated on the array for 1 h in a binding reaction containing: 20 mM HEPES with 0.3% milk, 0.02% Triton X-100, 1 mM DTT, 0.2 mg/mL bovine serum albumin, 0.4 mg/ml salmon testes DNA (Sigma D7656), and 0.4 mM $MgCl_2$. Jurkat lysates were incubated at a final concentration of 3.2 mg/mL and THP-1 lysates at a final concentration of 2.9 mg/mL, resulting in 110 mM NaCl in the binding reactions. After lysate incubation and each of the following antibody incubation steps, microarrays were briefly rinsed with HBS containing 0.05% Tween-20, then dewet in HBS. After lysate incubation, microarrays were incubated with 20 μg/ml of primary antibody in 2% milk in HBS for 20 min, followed by 20 μg/ml of fluorescently-labeled secondary antibody in 2% milk in PBS for 20 min. The following antibodies were used: anti-P300 (Abcam ab14984), anti-TBLR1 (Santa Cruz sc-100908), anti-BRG1 (Santa Cruz sc-11796), anti-p65 (Santa Cruz sc-8008), anti-HDAC1 (Abcam ab7028), anti-GPS2 (ABclonal A3901), anti-RBBP5 (Bethyl Laboratories, A300-109A), anti-NCOR1 (Bethyl Laboratories, A301-145A), anti-mouse IgG-Alexa488 (Invitrogen A11001), anti-rabbit IgG-Alexa488 (Invitrogen A27034), anti-mouse IgG-Alexa647 (Invitrogen A21236), anti-rabbit IgG-Alexa647 (Invitrogen A21245). 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).

Position frequency matrices (PFMs) and DNA-binding logos were generated for each seed using the previously described SNV-based approach (Andrilenas et al., 2018), with

β set to 2, except for NF-kB motifs, for which β was set to 1. Some proteins exhibit an orientation-specific bias in our PBM experiments; therefore, data from each orientation was considered separately for these analyses.

Motif matching was performed with tomtom, part of the MEME suite (version 5.0.3). The following parameters were used: minimum overlap: 3, distance metric: Euclidean, incomplete scoring. The following PWM databases from MEME were used: JASPAR2018_CORE_vertebrates_non-redundant.meme, uniprobe_mouse.meme, jolma2010.meme, jolma2013.meme, HOCOMOCOv11_full_HUMAN_mono_meme_format.meme, HOCOMOCOv11_full_MOUSE_mono_meme_format.meme, wei2010_human_mws.meme.

CHAPTER FOUR: Discussion

In this thesis, we examine regulatory protein interactions with cis-regulatory elements (CREs) from two perspectives. First, we characterize the DNA-binding landscapes of the NR family of TFs, then we go on to describe a novel approach for assessing DNA-TF-CoF complexes more generally.

In Chapter 2, we provide the most comprehensive characterization of the DNA binding landscape the type II nuclear receptors to date. We find all examined NRs have more promiscuous DNA binding preferences than previously reported, challenging the view that NR-DNA binding specificity is defined by half-site spacing. Intriguingly, we find

that all DRs tested can be bound by multiple NRs; for example, DR1 and DR3 were bound by all tested NRs, and all NRs can bind a single half-site. These results suggest that that additional mechanisms beyond DR spacer length must be utilized to specify NR gene regulation. For these experiments we used purified proteins, allowing us to characterize the inherent DNA-binding preferences of these TFs in a highly controlled manner; however, a cellular context may provide additional determinants for NR regulatory specificity. For example, NRs interact with a wide array of CoFs, and allosteric interactions with CoFs could modulate NR-DNA binding preferences, suggesting a potential mechanism for refining NR regulatory specificity. Moreover, PTMs are also known to modulate NR-CoF interactions (Becares et al., 2016), suggesting any additional specificity provided by CoF interactions may be cell-state specific and modified by PTMs. Thus, future studies that assess NR-DNA binding in a cellular context are necessary to clarify the relationship between NR-DNA binding preferences and activity.

To this end, our lab is working toward performing PBM experiments with nuclear lysates from 3T3-L1 adipocytes, allowing us to evaluate NR-DNA binding in a cellular context. 3T3-L1 cells are multi-potential fibroblasts that can be differentiated into adipocyte-like cells, and are a common model for adipocytes. PPARγ is an important factor in the differentiation of these cells into adipocytes, and responsible for increasing the rate of free fatty acid uptake and intracellular triglyceride content characteristic of adipocytes (Tamori et al., 2002). Thus, these cells serve as an ideal model for examining PPARγ

regulation in a more native context. For these experiments, we utilize the NR-specific PBM described in Chapter 2, containing 24 different seed sequences for each DR with spacers of 0 – 5 nt. This data will allow us to directly assess the differences in DNA binding between purified PPARγ and PPARγ in a cellular context, complete with native CoFs and PTMs. By applying the CoRec approach of examining cell lysates and probing the array with CoF-specific antibodies, we will also be able to assess the relationship between PPARγ-DNA binding in a cellular context and CoF recruitment.

Our analysis also reveals that canonical models of NR specificity better reflect NR activity rather than DNA binding. For a limited set of sequences, we show that PPARγ tends to drive gene expression more strongly from its canonically preferred DR1 site, as does LXRa from its canonically preferred DR4 site. Higher throughput strategies, such as massively parallel reporter assays (MPRAs) (Melnikov et al., 2012), must be used to more generally assess the relationship between regulatory element sequence and gene regulation. To bridge the gap between NR-DNA binding and activity, an MPRA utilizing the seed and SNV sequences we assessed by PBM as response elements would enable us to build models of NR activity. In such an approach, the PBM probe sequences would be cloned upstream of a minimal promoter and reporter gene to create a reporter library that would be transfected into 3T3-L1 adipocytes, and the expression of the reporter gene from each unique regulatory element measured. Differential expression driven by different regulatory element variants could be transformed into models representing activity. By comparing these activity models to models derived from CoF and TF PBMs

from nuclear lysates, we would be able to develop a deeper understanding of the relationship between DNA sequence, NR binding, CoF recruitment, and gene regulation.

While we focus on NRs here, the question of how specificity is achieved between TFs that show similar DNA sequence preferences is a problem common to many families of TFs. For example, the members of the ETS family of TFs play both overlapping and distinct roles in many biological processes, though these factors display similar DNA binding preferences (Andrilenas et al., 2015). PBMs have been used to characterize their binding preferences in vitro, elucidating TF-specific preferences (reviewed in (Andrilenas et al., 2015)). Our data suggest that additional specificity may be achieved in a cellular context for the NRs, suggesting additional study of ETS family members in a cellular context may reveal additional insight into how these TFs achieve regulatory specificity. High-throughput strategies such as nextPBM and CoRec will be invaluable to resolving the strategies TFs with similar binding preferences implement to obtain regulatory specificity.

In Chapter 3, we demonstrate a novel strategy for elucidating the many TF-CoF complexes operating in a cell and the sequence determinants of their recruitment to DNA. For example, in both THP-1 and Jurkat T cells, p300 regulates AP-1 sites, whereas p300, NCOR, TBLR1, and GPS2 are all recruited to CREs. Our analyses reveal cell-type and stimulus specific interactions, such as the recruitment of p65 to unstimulated THP-1 cells, but not stimulated THP-1 cells or either stimulated or unstimulated Jurkats. We

recapitulate known interactions, such as LPS-inducible recruitment of p300 to AP-1 response elements and CREs, suggesting this approach accurately represents CoF interactions. We also identify new interactions, such as the indirect recruitment of p65 in unstimulated THP-1s, and identify new regulatory elements.

In this chapter, we only the scratch the surface of the information that can be obtained with CoRec. Similar studies could be performed with additional cells types to gain a better understanding of cell-type specific and conserved regulatory strategies. Similarly, we anticipate this approach would work with many stimuli; for example, we could observe the change in CoF landscape upon treatment with different cytokines, or at different points of lineage commitment. In this work, we utilize cell lines, however, preliminary work with mouse liver lysates suggests the approach is extensible to primary cells. Thus, CoRec could be used both to examine CoF recruitment in primary cells and to compare CoF recruitment in primary cells to corresponding cell lines used as models to assess the validity of these models.

We anticipate this platform will be widely useful in expanding our understanding of the role CoFs play in gene regulation. For example, as described in Chapter 1, CoFs are often part of large multi- subunit complexes, and different CoFs can be recruited to different complexes. Additional studies that probe for differential subunit composition of these larger complexes will further elucidate their roles in regulation. For example, TBLR1 and TBL1X have both been shown to interact with the NCOR complex either

individually or together, but their unique roles in this complex are yet to be fully elucidated (Perissi et al., 2008).

The CoRec approach could also be a valuable tool for understanding the impact of drugs on gene regulation, as many drugs target TFs or CoFs. For example, glucocorticoids are a family of drugs that target the glucocorticoid receptor, a member of the type I nuclear receptors (Newton and Holden, 2007). Upon glucocorticoid binding, cytosolic GR translocates to the nucleus where it represses the expression of pro-inflammatory genes to achieve its desired therapeutic effect. However, it also activates other genes, and this transactivation has been linked to undesired side effects. Thus, there has been a significant effort to find GR agonists that preferentially lead to transrepression rather than transactivation (Newton and Holden, 2007). Testing the effects of different GR agonists on CoF complex formation and recruitment could yield insight into the molecular mechanisms underlying these observations, and may suggest strategies for the development of more specific drugs.

CoRec provides a high-throughput approach for elucidating TF-CoF recruitment to DNA elements in a particular cellular state. By coupling CoRec with MPRAs, the relationship between these regulatory complexes and gene expression can be examined. For example, a CoRec-probe based MPRA library could be transfected into THP-1 and Jurkat cells, and expression could be assessed with and without TLR4 or TCR stimulation. Comparison of the CoF models described in Chapter 3 to expression models derived from

the MPRA would allow us to directly assess the relationship between CoF recruitment and gene expression. Similarly, CoRec could be coupled with mass spectrometry or phosphoproteomic data to identify additional members of regulatory complexes and query the role of phosphorylation state on complex formation. Thus the CoRec approach, either alone, or paired with other techniques, will be a powerful tool for developing a deeper understanding of gene regulation.

BIBLIOGRAPHY

Abraham, R.T., and Weiss, A. (2004). Jurkat T cells and development of the T-cell receptor signalling paradigm. Nature Reviews. Immunology *4*, 301–308.

Alland, L., Muhle, R., Hou, H., Potes, J., Chin, L., Schreiber-Agus, N., and DePinho, R.A. (1997). Role for N-CoR and histone deacetylase in Sin3-mediated transcriptional repression. Nature *387*, 49–55.

Andrilenas, K.K., Penvose, A., and Siggers, T. (2015). Using protein-binding microarrays to study transcription factor specificity: homologs, isoforms and complexes. Briefings in Functional Genomics *14*, 17–29.

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

Arbab, M., Mahony, S., Cho, H., Chick, J.M., Rolfe, P.A., van Hoff, J.P., Morris, V.W.S., Gygi, S.P., Maas, R.L., Gifford, D.K., et al. (2013). A multi-parametric flow cytometric assay to analyze DNA-protein interactions. Nucleic Acids Research. *41*, e38–e38.

Becares, N., Gage, M.C., and Pineda-Torra, I. (2016). Posttranslational Modifications of Lipid-Activated Nuclear Receptors: Focus on Metabolism. Endocrinology *158*, 213–225.

Berger, M.F., and Bulyk, M.L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nature Protocols *4*, 393–411.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nature Biotechnology *24*, 1429–1435.

Boergesen, M., Pedersen, T.A., Gross, B., van Heeringen, S.J., Hagenbeek, D., Bindesboll, C., Caron, S., Lalloyer, F., Steffensen, K.R., Nebb, H.I., et al. (2012). Genome-Wide Profiling of Liver X Receptor, Retinoid X Receptor, and Peroxisome Proliferator-Activated Receptor in Mouse Liver Reveals Extensive Sharing of Binding Sites. Molecular and Cellular Biology *32*, 852–867.

Bolotin, E., Liao, H., Ta, T.C., Yang, C., Hwang-Verslues, W., Evans, J.R., Jiang, T., and Sladek, F.M. (2010). Integrated approach for the identification of human hepatocyte nuclear factor 4alpha target genes using protein binding microarrays. Hepatology *51*, 642–653.

Bookout, A.L., Jeong, Y., Downes, M., Yu, R.T., Evans, R.M., and Mangelsdorf, D.J. (2006). Anatomical profiling of nuclear receptor expression reveals a hierarchical

transcriptional network. Cell *126*, 789–799.

Buecker, C., and Wysocka, J. (2012). Enhancers as information integration hubs in development: lessons from genomics. Trends in Genetics *28*, 276–284.

Calkin, A.C., and Tontonoz, P. (2012). Transcriptional integration of metabolism by the nuclear sterol-activated receptors LXR and FXR. Nature Reviews Mol Cell Biol *13*, 213–24.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Molecular Cell *49*, 825–837.

Chambers, A.E., Banerjee, S., Chaplin, T., Dunne, J., Debernardi, S., Joel, S.P., and Young, B.D. (2003). Histone acetylation-mediated regulation of genes in leukaemic cells. European Journal of Cancer *39*, 1165–1175.

Chandra, V., Huang, P., Hamuro, Y., Raghuram, S., Wang, Y., Burris, T.P., and Rastinejad, F. (2008). Structure of the intact PPAR-gamma-RXR- nuclear receptor complex on DNA. Nature *456*, 350–356.

Chaplin, D.D. (2010). Overview of the immune response. Journal of Allergy and Clinical Immunology *125*, S3–S23.

Chatagnon, A., Veber, P., Morin, V., Bedo, J., Triqueneaux, G., Sémon, M., Laudet, V., d'Alché-Buc, F., and Benoit, G. (2015). RAR/RXR binding dynamics distinguish pluripotency from differentiation associated cis-regulatory elements. Nucleic Acids Research *43*, 4833–4854.

Claessens, F., and Gewirth, D.T. (2004). DNA recognition by nuclear receptors. Essays in Biochemistry *40*, 59–72.

Clouaire, T., Webb, S., Skene, P., Illingworth, R., Kerr, A., Andrews, R., Lee, J.-H., Skalnik, D., and Bird, A. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. Genes & Development *26*, 1714–1728.

Consortium, T.E.P., data analysis coordination, O.C., data production, D.P.L., data analysis, L.A., group, W., scientific management, N.P.M., steering committee, P.I., Boise State University and University of North Carolina at Chapel Hill Proteomics groups (data production and analysis), Broad Institute Group (data production and analysis), Cold Spring Harbor, University of Geneva, Center for Genomic Regulation, Barcelona, RIKEN, Sanger Institute, University of Lausanne, Genome Institute of Singapore group (data production and analysis), et al. (2013). An integrated encyclopedia of DNA elements in the human genome. Nature *488*, 57–74.

Cotnoir-White, D., Laperrière, D., and Mader, S. (2011). Evolution of the repertoire of

nuclear receptor binding sites in genomes. Molecular and Cellular Endocrinology *334*, 76–82.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America *107*, 21931–21936.

Crump, N.T., and Milne, T.A. (2019). Why are so many MLL lysine methyltransferases required for normal mammalian development? Cellular and Molecular Life Sciences (CMLS) *76*, 2885–2898.

Curina, A., Termanini, A., Barozzi, I., Prosperini, E., Simonatto, M., Polletti, S., Silvola, A., Soldi, M., Austenaa, L., Bonaldi, T., et al. (2017). High constitutive activity of a broad panel of housekeeping and tissue-specific cis-regulatory elements depends on a subset of ETS proteins. Genes & Development *31*, 399–412.

de Aguiar Vallim, T.Q., Tarling, E.J., and Edwards, P.A. (2013). Pleiotropic roles of bile acids in metabolism. Cell Metabolism *17*, 657–669.

De Obaldia, M.E., Bell, J.J., Wang, X., Harly, C., Yashiro-Ohtani, Y., DeLong, J.H., Zlotoff, D.A., Sultana, D.A., Pear, W.S., and Bhandoola, A. (2013). T cell development requires constraint of the myeloid regulator C/EBP-α by the Notch target and transcriptional repressor Hes1. Nature Immunology *14*, 1277–1284.

Dorighi, K.M., Swigut, T., Henriques, T., Bhanu, N.V., Scruggs, B.S., Nady, N., Still, C.D., Garcia, B.A., Adelman, K., and Wysocka, J. (2017). Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. Molecular Cell *66*, 568–576.e4.

Evans, R.M., and Mangelsdorf, D.J. (2014). Nuclear Receptors, RXR, and the Big Bang. Cell *157*, 255–266.

Everett, L.J., and Lazar, M.A. (2013). Cell-specific integration of nuclear receptor function at the genome. Wiley Interdisciplinary Reviews. Systems Biology & Med *5*, 615–629.

Fang, B., Mane-Padros, D., Bolotin, E., Jiang, T., and Sladek, F.M. (2012). Identification of a binding motif specific to HNF4 by comparative analysis of multiple nuclear receptors. Nucleic Acids Research *40*, 5343–5356.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. Nature *340*, 245–246.

Filtz, T.M., Vogel, W.K., and Leid, M. (2014). Regulation of transcription factor activity by interconnected post-translational modifications. Trends in Pharmacol. Sci. *35*, 76–85.

Finkel, T., Deng, C.-X., and Mostoslavsky, R. (2009). Recent progress in the biology and physiology of sirtuins. Nature *460*, 587–591.

Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z., et al. (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Molecular Cell *47*, 810–822.

Gaud, G., Lesourne, R., and Love, P.E. (2018). Regulatory mechanisms in T cell receptor signalling. Nature Reviews. Immunology *18*, 485–497.

Ghisletti, S., Barozzi, I., Mietton, F., Polletti, S., De Santa, F., Venturini, E., Gregory, L., Lonie, L., Chew, A., Wei, C.-L., et al. (2010). Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. Immunity *32*, 317–328.

Golden, M.S., Cote, S.M., Sayeg, M., Zerbe, B.S., Villar, E.A., Beglov, D., Sazinsky, S.L., Georgiadis, R.M., Vajda, S., Kozakov, D., et al. (2013). Comprehensive experimental and computational analysis of binding energy hot spots at the NF-κB essential modulator/IKKβ protein-protein interface. J. Am. Chem. Soc. *135*, 6242–6256.

Gronemeyer, H., and Bourguet, W. (2009). Allosteric Effects Govern Nuclear Receptor Action: DNA Appears as a Player. Science Signaling *2*, pe34–pe34.

Guenther, M.G., Lane, W.S., Fischle, W., Verdin, E., Lazar, M.A., and Shiekhattar, R. (2000). A core SMRT corepressor complex containing HDAC3 and TBL1, a WD40-repeat protein linked to deafness. Genes & Development *14*, 1048–1057.

Hakimi, M.-A., Bochar, D.A., Chenoweth, J., Lane, W.S., Mandel, G., and Shiekhattar, R. (2002). A core-BRAF35 complex containing histone deacetylase mediates repression of neuronal-specific genes. Proceedings of the National Academy of Sciences *99*, 7420–25.

Hallson, G., Hollebakken, R.E., Li, T., Syrzycka, M., Kim, I., Cotsworth, S., Fitzpatrick, K.A., Sinclair, D.A.R., and Honda, B.M. (2012). dSet1 is the main H3K4 di- and tri-methyltransferase throughout Drosophila development. Genetics *190*, 91–100.

Hanigan, T.W., Danes, J.M., Taha, T.Y., Frasor, J., and Petukhov, P.A. (2018). Histone deacetylase inhibitor-based chromatin precipitation for identification of targeted genomic loci. Journal of Biological Methods *5*, 88.

Hargreaves, D.C., Horng, T., and Medzhitov, R. (2009). Control of inducible gene expression by signal-dependent transcriptional elongation. Cell *138*, 129–145.

Hassig, C.A., Fleischer, T.C., Billin, A.N., Schreiber, S.L., and Ayer, D.E. (1997). Histone deacetylase activity is required for full transcriptional repression by mSin3A.

Cell *89*, 341–347.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics *39*, 311–318.

Heinzel, T., Lavinsky, R.M., Mullen, T.M., Söderstrom, M., Laherty, C.D., Torchia, J., Yang, W.M., Brard, G., Ngo, S.D., Davie, J.R., et al. (1997). A complex containing N-CoR, mSin3 and histone deacetylase mediates transcriptional repression. Nature *387*, 43–48.

Hoberg, J.E., Popko, A.E., Ramsey, C.S., and Mayo, M.W. (2005). I B Kinase-Mediated Derepression of SMRT Potentiates Acetylation of RelA/p65 by p300. Molecular and Cellular Biology *26*, 457–471.

Hsu, J.C., Bravo, R., and Taub, R. (1992). Interactions among LRF-1, JunB, c-Jun, and c-Fos define a regulatory program in the G1 phase of liver regeneration. Molecular and Cellular Biology *12*, 4654–4665.

Hubner, N.C., Nguyen, L.N., Hornig, N.C., and Stunnenberg, H.G. (2015). A quantitative proteomics tool to identify DNA-protein interactions in primary cells or blood. Journal of Proteome Research *14*, 1315–1329.

Hudson, W.H., de Vera, I.M.S., Nwachukwu, J.C., Weikum, E.R., Herbst, A.G., Yang, Q., Bain, D.L., Nettles, K.W., Kojetin, D.J., and Ortlund, E.A. (2018). Cryptic glucocorticoid receptor-binding sites pervade genomic NF-κB response elements. Nature Communications *9*, 1337.

Humphrey, G.W., Wang, Y., Russanova, V.R., Hirai, T., Qin, J., Nakatani, Y., and Howard, B.H. (2001). Stable histone deacetylase complexes distinguished by the presence of SANT domain proteins CoREST/kiaa0071 and Mta-L1. Journal of Biological Chemistry *276*, 6817–6824.

Imbalzano, A.N., Kwon, H., Green, M.R., and Kingston, R.E. (1994). Facilitated binding of TATA-binding protein to nucleosomal DNA. Nature *370*, 481–485.

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

Issa, L.L., Leong, G.M., Barry, J.B., Sutherland, R.L., and Eisman, J.A. (2001). Glucocorticoid receptor-interacting protein-1 and receptor-associated coactivator-3 differentially interact with the vitamin D receptor (VDR) and regulate VDR-retinoid X receptor transcriptional cross-talk. Endocrinology *142*, 1606–1615.

Jin, Q., Yu, L.-R., Wang, L., Zhang, Z., Kasper, L.H., Lee, J.-E., Wang, C., Brindle, P.K., Dent, S.Y.R., and Ge, K. (2011). Distinct roles of GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear receptor transactivation. The EMBO Journal *30*, 249–262.

Juge-Aubry, C., Pernin, A., Favez, T., Burger, A.G., Wahli, W., Meier, C.A., and Desvergne, B. (1997). DNA binding properties of peroxisome proliferator-activated receptor subtypes on various natural peroxisome proliferator response elements. Importance of the 5'-flanking region. Journal of Biological Chemistry *272*, 25252–259.

Katz, R.W., Subauste, J.S., and Koenig, R.J. (1995). The interplay of half-site sequence and spacing on the activity of direct repeat thyroid hormone response elements. Journal of Biological Chemistry *270*, 5238–5242.

Kawai, T., and Akira, S. (2007). Signaling to NF-κB by Toll-like receptors. Trends in Molecular Medicine *13*, 460–469.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G., et al. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Research *46*, D260–D266.

Khorasanizadeh, S., and Rastinejad, F. (2001). Nuclear-receptor interactions on DNA-response elements. Trends in Biochemical Sciences *26*, 384–390.

Kliewer, S.A., Lehmann, J.M., and Willson, T.M. (1999). Orphan nuclear receptors: shifting endocrinology into reverse. Science *284*, 757–760.

Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., et al. (2017). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Research *46*, D252–D259.

Kurokawa, R., Yu, V.C., Näär, A., Kyakumoto, S., Han, Z., Silverman, S., Rosenfeld, M.G., and Glass, C.K. (1993). Differential orientations of the DNA-binding domain and carboxy-terminal dimerization interface regulate binding site selection by nuclear receptor heterodimers. Genes & Development *7*, 1423–1435.

Kuzmic, P. (1996). Program DYNAFIT for the analysis of enzyme kinetic data:

application to HIV proteinase. Analytical Biochemistry *237*, 260–273.

Kwon, H., Imbalzano, A.N., Khavari, P.A., Kingston, R.E., and Green, M.R. (1994). Nucleosome disruption and enhancement of activator binding by a human SW1/SNF complex. Nature *370*, 477–481.

Lai, B., Lee, J.-E., Jang, Y., Wang, L., Peng, W., and Ge, K. (2017). MLL3/MLL4 are required for CBP/p300 binding on enhancers and super-enhancer formation in brown adipogenesis. Nucleic Acids Research *45*, 6388–6403.

Lee, J.-H., and Skalnik, D.G. (2007). Wdr82 is a C-terminal domain-binding protein that recruits the Setd1A Histone H3-Lys4 methyltransferase complex to transcription start sites of transcribed human genes. Molecular and Cellular Biology *28*, 609–618.

Lee, J.-E., Wang, C., Xu, S., Cho, Y.-W., Wang, L., Feng, X., Baldridge, A., Sartorelli, V., Zhuang, L., Peng, W., et al. (2013). H3K4 mono- and di-methyltransferase MLL4 is required for enhancer activation during cell differentiation. Elife *2*, e01503.

Lefebvre, P., Mouchon, A., Lefebvre, B., and Formstecher, P. (1998). Binding of retinoic acid receptor heterodimers to DNA. A role for histones NH2 termini. Journal of Biological Chemistry *273,* 12288–12295.

Lefterova, M.I., Steger, D.J., Zhuo, D., Qatanani, M., Mullican, S.E., Tuteja, G., Manduchi, E., Grant, G.R., and Lazar, M.A. (2010). Cell-specific determinants of peroxisome proliferator-activated receptor gamma function in adipocytes and macrophages. Molecular and Cellular Biology *30*, 2078–2089.

Li, J., Wang, J., Nawaz, Z., Liu, J.M., Qin, J., and Wong, J. (2000). Both corepressor proteins SMRT and N-CoR exist in large protein complexes containing HDAC3. The EMBO Journal *19*, 4342–4350.

Lin, C.H., Hare, B.J., Wagner, G., Harrison, S.C., Maniatis, T., and Fraenkel, E. (2001). A Small Domain of CBP/p300 Binds Diverse Proteins. Molecular Cell *8*, 581–590.

Lou, X., Toresson, G., Benod, C., Suh, J.H., Philips, K.J., Webb, P., and Gustafsson, J.-A. (2014). Structure of the retinoid X receptor α-liver X receptor β (RXRα-LXRβ) heterodimer on DNA. Nature Structural & Molecular Biology *21*, 277–281.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550.

Mader, S., Chen, J.Y., Chen, Z., White, J., Chambon, P., and Gronemeyer, H. (1993). The patterns of binding of RAR, RXR and TR homo- and heterodimers to direct repeats are dictated by the binding specificites of the DNA binding domains. The EMBO Journal *12*, 5029–5041.

Mak, K.S., Funnell, A.P.W., Pearson, R.C.M., and Crossley, M. (2011). PU.1 and Haematopoietic Cell Fate: Dosage Matters. Int J Cell Biol *2011*, 1–6.

Martin, M., Kettmann, R., and Dequiedt, F. (2009). Class IIa histone deacetylases: Conducting development and differentiation. Int. J. of Developmental Biology *53*, 291–301.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Research *34*, D108–D110.

McKenna, N.J., and O'Malley, B.W. (2002). Combinatorial control of gene expression by nuclear receptors and coregulators. Cell *108*, 465–474.

Meijsing, S.H., Pufall, M.A., So, A.Y., Bates, D.L., Chen, L., and Yamamoto, K.R. (2009). DNA binding site sequence directs glucocorticoid receptor structure and activity. Science *324*, 407–410.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Kinney, J.B., et al. (2012). systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nature Biotechnology *30*, 271–277.

Mittler, G., Butter, F., and Mann, M. (2008). A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. Genome Research *19*, 284–293.

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

Mukherjee, S.P., Behar, M., Birnbaum, H.A., Hoffmann, A., Wright, P.E., and Ghosh, G. (2013). Analysis of the RelA:CBP/p300 Interaction Reveals Its Involvement in NF-κB-Driven Transcription. PLoS Biology *11*, e1001647.

Nagy, L., Kao, H.Y., Chakravarti, D., Lin, R.J., Hassig, C.A., Ayer, D.E., Schreiber, S.L., and Evans, R.M. (1997). Nuclear receptor repression mediated by a complex containing SMRT, mSin3A, and histone deacetylase. Cell *89*, 373–380.

Naito, T., Tanaka, H., Naoe, Y., and Taniuchi, I. (2011). Transcriptional control of T-cell development. International Immunology *23*, 661–668.

Natoli, G. (2010). Maintaining Cell Identitythrough Global Control of Genomic Organization. Immunity *33*, 12–24.

Näär, A.M., Boutin, J.M., Lipkin, S.M., Yu, V.C., Holloway, J.M., Glass, C.K., and Rosenfeld, M.G. (1991). The orientation and spacing of core DNA-binding motifs dictate selective transcriptional responses to three nuclear receptors. Cell *65*, 1267–1279.

Newton, R., and Holden, N.S. (2007). Separating Transrepression and Transactivation: A Distressing Divorce for the Glucocorticoid Receptor? Molecular Pharmacol. *72*, 799–809.

Ogawa, S., Lozach, J., Benner, C., Pascual, G., Tangirala, R.K., Westin, S., Hoffmann, A., Subramaniam, S., David, M., Rosenfeld, M.G., et al. (2005). Molecular Determinants of Crosstalk between Nuclear Receptors and Toll-like Receptors. Cell *122*, 707–721.

Park, E.K., Jung, H.S., Yang, H.I., Yoo, M.C., Kim, C., and Kim, K.S. (2007). Optimized THP-1 differentiation is required for the detection of responses to weak stimuli. Inflammation Research. *56*, 45–50.

Pasini, D., Malatesta, M., Jung, H.R., Walfridsson, J., Willer, A., Olsson, L., Skotte, J., Wutz, A., Porse, B., Jensen, O.N., et al. (2010). Characterization of an antagonistic switch between histone H3 lysine 27 methylation and acetylation in the transcriptional regulation of Polycomb group target genes. Nucleic Acids Research *38*, 4958–4969.

Pehkonen, P., Welter-Stahl, L., Diwo, J., Ryynänen, J., Wienecke-Baldacchino, A., Heikkinen, S., Treuter, E., Steffensen, K.R., and Carlberg, C. (2012). Genome-wide landscape of liver X receptor chromatin binding and gene regulation in human macrophages. BMC Genomics *13*, 50.

Penvose, A., Keenan, J.L., Bray, D., Ramlall, V., and Siggers, T. (2019). Comprehensive study of nuclear receptor DNA binding provides a revised framework for understanding receptor specificity. Nature Communications *10*, 2514.

Perissi, V., and Rosenfeld, M.G. (2005). Controlling nuclear receptors: the circular logic of cofactor cycles. Nature Reviews. Molecular Cell Biology *6*, 542–554.

Perissi, V., Jepsen, K., Glass, C.K., and Rosenfeld, M.G. (2010). Deconstructing repression: evolving models of co-repressor action. Nature Reviews Genetics *11*, 109–123.

Perissi, V., Scafoglio, C., Zhang, J., Ohgi, K.A., Rose, D.W., Glass, C.K., and Rosenfeld, M.G. (2008). TBL1 and TBLR1 Phosphorylation on Regulated Gene Promoters Overcomes Dual CtBP and NCoR/SMRT Transcriptional Repression Checkpoints. Molecular Cell *29*, 755–766.

Perlmann, T., Rangarajan, P.N., Umesono, K., and Evans, R.M. (1993). Determinants for selective RAR and TR recognition of direct repeat HREs. Genes & Development *7*, 1411–1422.

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

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114–128.

Rastinejad, F., Perlmann, T., Evans, R.M., and Sigler, P.B. (1995). Structural determinants of nuclear receptor assembly on DNA direct repeats. Nature *375*, 203–211.

Rastinejad, F., Huang, P., Chandra, V., and Khorasanizadeh, S. (2013). Understanding nuclear receptor form and function using structural biology. Journal of Molecular Endocrinology *51,* T1-T21.

Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., et al. (2010). An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. Cell *140*, 744–752.

Richon, V.M., Sandhoff, T.W., Rifkind, R.A., and Marks, P.A. (2000). Histone deacetylase inhibitor selectively induces p21WAF1 expression and gene-associated histone acetylation. Proceedings of the National Academy of Sciences, USA *97*, 10014–10019.

Riley, T.R., Slattery, M., Abe, N., Rastogi, C., Liu, D., Mann, R.S., and Bussemaker, H.J. (2014). SELEX-seq: A Method for Characterizing the Complete Repertoire of Binding Site Preferences for Transcription Factor Complexes. In Hox Genes, (New York, NY: Springer New York), pp. 255–278.

Rodríguez-Martínez, J.A., Reinke, A.W., Bhimsaria, D., Keating, A.E., and Ansari, A.Z. (2017). Combinatorial bZIP dimers display complex DNA-binding specificity landscapes. Elife *6*.

Roeder, R.G. (2005). Transcriptional regulation and the role of diverse coactivators in animal cells. FEBS Letters *579*, 909–915.

Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A Proteome-Scale Map of the Human Interactome Network. Cell *159*, 1212–1226.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., et al. (2005). Towards a

proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173–1178.

Savic, D., Ramaker, R.C., Roberts, B.S., Dean, E.C., Burwell, T.C., Meadows, S.K., Cooper, S.J., Garabedian, M.J., Gertz, J., and Myers, R.M. (2016). Distinct gene regulatory programs define the inhibitory effects of liver X receptors and PPARG on cancer cell proliferation. Genome Medicine *8*, 74.

Schöne, S., Jurk, M., Helabad, M.B., Dror, I., Lebars, I., Kieffer, B., Imhof, P., Rohs, R., Vingron, M., Thomas-Chollier, M., et al. (2016). Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity. Nature Commun *7*, 12621.

Sharrocks, A.D. (2001). The ETS-domain transcription factor family. Nature Reviews. Molecular Cell Biology *2*, 827–837.

Shilatifard, A. (2012). The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. Annual Rev. Biochem. *81*, 65–95.

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

Simicevic, J., Schmid, A.W., Gilardoni, P.A., Zoller, B., Raghav, S.K., Krier, I., Gubelmann, C., Lisacek, F., Naef, F., Moniatte, M., et al. (2013). Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics. Nature Methods *10*, 570–576.

Smale, S.T., Tarakhovsky, A., and Natoli, G. (2014). Chromatin contributions to the regulation of innate immunity. Annual Reviews of Immunology *32*, 489–511.

Smith-Garvin, J.E., Koretzky, G.A., and Jordan, M.S. (2009). T cell activation. Annual Review of Immunology *27*, 591–619.

Soccio, R.E., Chen, E.R., Rajapurkar, S.R., Safabakhsh, P., Marinis, J.M., Dispirito, J.R., Emmett, M.J., Briggs, E.R., Fang, B., Everett, L.J., et al. (2015). Genetic Variation Determines PPARγ Function and Anti-diabetic Drug Response In Vivo. Cell *162*, 33–44.

Tamori, Y., Masugi, J., Nishino, N., and Kasuga, M. (2002). Role of Peroxisome Proliferator-Activated Receptor- in Maintenance of the Characteristics of Mature 3T3-L1 Adipocytes. Diabetes *51*, 2045–2055.

Temple, K.A., Cohen, R.N., Wondisford, S.R., Yu, C., Deplewski, D., and Wondisford, F.E. (2005). An intact DNA-binding domain is not required for peroxisome proliferator-activated receptor gamma (PPARgamma) binding and activation on some PPAR response elements. Journal of Biological Chemistry *280*, 3529–3540.

Tie, F., Banerjee, R., Stratton, C.A., Prasad-Sinha, J., Stepanik, V., Zlobin, A., Diaz, M.O., Scacheri, P.C., and Harte, P.J. (2009). CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. Development *136*, 3131–3141.

Tong, J.K., Hassig, C.A., Schnitzler, G.R., Kingston, R.E., and Schreiber, S.L. (1998). Chromatin deacetylation by an ATP-dependent nucleosome remodelling complex. Nature *395*, 917–921.

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

Tugal, D., Liao, X., and Jain, M.K. (2013). Transcriptional control of macrophage polarization. Arteriosclerosis, Thrombosis, and Vascular Biology *33*, 1135–1144.

Verdin, E., Dequiedt, F., and Kasler, H.G. (2003). Class II histone deacetylases: versatile regulators. Trends in Genetics *19*, 286–293.

Wang, C., Lee, J.-E., Lai, B., Macfarlan, T.S., Xu, S., Zhuang, L., Liu, C., Peng, W., and Ge, K. (2016). Enhancer priming by H3K4 methyltransferase MLL4 controls cell fate transition. Proceedings of the National Academy of Sciences *113*, 11871–11876.

Wang, W., Côté, J., Xue, Y., Zhou, S., Khavari, P.A., Biggar, S.R., Muchardt, C., Kalpana, G.V., Goff, S.P., Yaniv, M., et al. (1996). Purification and biochemical heterogeneity of the mammalian SWI-SNF complex. The EMBO Journal *15*, 5370–5382.

Wang, Z., Zang, C., Cui, K., Schones, D.E., Barski, A., Peng, W., and Zhao, K. (2009). Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. Cell *138*, 1019–1031.

Watson, L.C., Kuchenbecker, K.M., Schiller, B.J., Gross, J.D., Pufall, M.A., and Yamamoto, K.R. (2013). The glucocorticoid receptor dimer interface allosterically transmits sequence-specific DNA signals. Nature Structural & Molecular Biology *20*, 876–883.

Weake, V.M., and Workman, J.L. (2010). Inducible gene expression: diverse regulatory mechanisms. Nature Reviews. Genetics *11*, 426–437.

Wei, B., Jolma, A., Sahu, B., Orre, L.M., Zhong, F., Zhu, F., Kivioja, T., Sur, I., Lehtiö, J., Taipale, M., et al. (2018). A protein activity assay to measure global transcription factor activity reveals determinants of chromatin accessibility. Nature Biotechnology *15*, 351.

Wei, G.-H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R., et al. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. The EMBO Journal *29*, 2147–2160.

Weikum, E.R., Liu, X., and Ortlund, E.A. (2018). The nuclear receptor superfamily: A

structural perspective. Protein Science *27*(11), 1876-1892.

Wierer, M., and Mann, M. (2016). Proteomics to study DNA-bound and chromatin-associated gene regulatory complexes. Human Molecular Genetics *25*, R106–R114.

Wu, M., Wang, P.F., Lee, J.S., Martin-Brown, S., Florens, L., Washburn, M., and Shilatifard, A. (2008). Molecular regulation of H3K4 trimethylation by Wdr82, a component of human Set1/COMPASS. Molecular and Cellular Biology *28*, 7337–7344.

Wu, Q., Lian, J.B., Stein, J.L., Stein, G.S., Nickerson, J.A., and Imbalzano, A.N. (2017). The BRG1 ATPase of human SWI/SNF chromatin remodeling enzymes as a driver of cancer. Epigenomics *9*, 919–931.

Xue, Y., Wong, J., Moreno, G.T., Young, M.K., Côté, J., and Wang, W. (1998). NURD, a novel complex with both ATP-dependent chromatin-remodeling and histone deacetylase activities. Molecular Cell *2*, 851–861.

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

Yang, X., Chen, Y., and Gabuzda, D. (1999). ERK MAP Kinase Links Cytokine Signals to Activation of Latent HIV-1 Infection by Stimulating a Cooperative Interaction of AP-1 and NF-κB. Journal of Biological Chemistry *274*, 27981–27988.

Yoon, H.-G., Chan, D.W., Huang, Z.-Q., Li, J., Fondell, J.D., Qin, J., and Wong, J. (2003). Purification and functional characterization of the human N-CoR complex: the roles of HDAC3, TBL1 and TBLR1. The EMBO Journal *22*, 1336–1346.

You, A., Tong, J.K., Grozinger, C.M., and Schreiber, S.L. (2001). CoREST is an integral component of the CoREST- human histone deacetylase complex. Proceedings of the National Academy of Sciences of the United States of America *98*, 1454–1458.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., et al. (2008). High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. Science *322*, 104–110.

Zechel, C., Shen, X.Q., Chambon, P., and Gronemeyer, H. (1994). Dimerization interfaces formed between the DNA binding domains determine the cooperative binding of RXR/RAR and RXR/TR heterodimers to DR5 and DR4 elements. The EMBO Journal *13*, 1414–1424.

Zhan, L., Liu, H.-X., Fang, Y., Kong, B., He, Y., Zhong, X.-B., Fang, J., Wan, Y.-J.Y., and Guo, G.L. (2014). Genome-wide binding and transcriptome analysis of human farnesoid X receptor in primary human hepatocytes. PLoS ONE *9*, e105930.

Zhang, Y., Iratni, R., Erdjument-Bromage, H., Tempst, P., and Reinberg, D. (1997). Histone deacetylases and SAP18, a novel polypeptide, are components of a human Sin3 complex. Cell *89*, 357–364.

Zhang, Y., LeRoy, G., Seelig, H.P., Lane, W.S., and Reinberg, D. (1998). The dermatomyositis-specific autoantigen Mi2 is a component of a complex containing histone deacetylase and nucleosome remodeling activities. Cell *95*, 279–289.

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

Zhu, Y.P., Thomas, G.D., and Hedrick, C.C. (2016). 2014 Jeffrey M. Hoeg Award Lecture: Transcriptional Control of Monocyte Development. Arteriosclerosis, Thrombosis, and Vascular Biology *36*, 1722–1733.

Vita