Towards a Systematic Approach for Characterizing Regulatory Variation

Towards a Systematic Approach for Characterizing Regulatory Variation

A DISSERTATION PRESENTED
BY
LUIS ALBERTO BARRERA
TO
THE COMMITTEE ON HIGHER DEGREES IN BIOPHYSICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOPHYSICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
SEPTEMBER 2015

Dissertation advisors: Professors Martha Bulyk & Manolis Kellis Luis Alberto Barrera

Towards a Systematic Approach for Characterizing Regulatory Variation

Abstract

A growing body of evidence suggests that genetic variants that alter gene expression are responsible for many phenotypic differences across individuals, particularly for the risk of developing common diseases. However, the molecular mechanisms that underlie the vast majority of associations between genetic variants and their phenotypes remain unknown. An important limiting factor is that genetic variants remain difficult to interpret, particularly in noncoding sequences. Developing truly systematic approaches for characterizing regulatory variants will require: (a) improved annotations for the genomic sequences that control gene expression, (b) a more complete understanding of the molecular mechanisms through which genetic variants, both coding and noncoding, can affect gene expression, and (c) better experimental tools for testing hypotheses about regulatory variants.

In this dissertation, I present conceptual and methodological advances that directly contribute to each of these goals. A recurring theme in all of these developments is the statistical modeling of protein-DNA interactions and its integration with other data types. First, I describe enhancer-FACS-Seq, a high-throughput experimental approach for screening candidate enhancer sequences to test for *in vivo*, tissue-specific activity. Second, I present an integrative computational analysis of the *in vivo* binding of NF-κB, a key regulator of the immune system, yielding new insights into how genetic variants can affect NF-κB binding. Next, I describe the first comprehensive survey of coding variation in human transcription factors and what it reveals about additional sources of genetic variation that can affect gene expression. Finally, I present SIFTED, a statistical framework and web tool for the optimal design of TAL effectors, which have been used successfully in genome editing and can thus be used to test hypotheses about regulatory variants. Together, these developments help fulfill key needs in the quest to understand the molecular basis of human phenotypic variation.

Contents

List of Figures

To Marsella, my grandmother, who wished she could have celebrated the end of this journey with me.

Acknowledgments

Writing the acknowledgments section for a dissertation is the moment when one finally realizes that graduate school was an experience largely shaped not by the research you did, but by the people that influenced you while you were doing it.

The first two people I would like to thank are perhaps obvious choices: my advisors, Martha and Manolis. I was truly lucky to have the opportunity to be part of both of your labs. Through years of observing the contrasts in your ways of approaching questions and solving problems, I gained many key insights about being an effective scientist. Five years ago, I was a physics student with an interest in biology. Thanks to your guidance, I now feel comfortable calling myself a computational biologist. Working with each of you presented me with unique opportunities and challenges that were key drivers of that successful transition and of the personal growth that accompanied it.

In a similar fashion, I would like to sincerely thank my labmates at both MIT and Harvard. The knowledge that I gained through conversations with you, both casual and serious, cannot be underestimated. Your contributions to my personal and scientific development throughout the years were equally important. In particular, I would like to thank Alex, Steve, Luca, Jesse, Leila, Anton, Trevor, and Raluca in the Bulyk lab; and Wouter, Matt, Pouya, Andreas, Anshul, Ah-Ram, Bob, Luke, Jianrong, Richard, and Jason in the Kellis lab. At some point in my grad school experience, each of you were there as a mentor, a friend, an inspiration, or sometimes all of the above.

I would also like to express my sincere gratitude to my external collaborators: Ben Gewurz, Bo Zhao, Elliott Kieff, Brad Nelms, Keith Joung, Deepak Reyon, Jeffrey Sanders, Marc Vidal, and David

Hill. Working with you expanded my scientific interests to many new topics, some of which proved to be instrumental in shaping my career decisions.

Many other faculty members provided extremely helpful advice throughout this process. In particular, I would like to express my gratitude towards my dissertation advisory committee: Shamil Sunyaev, Richard Maas and Peter Kharchenko, for their patience, understanding and advice. Likewise, I sincerely thank my PQE committee: Ben Gewurz, Shirley Liu and Leonid Mirny, for their time and advice. Ben deserves a special mention for the amount of guidance, both scientific and personal, that he gave me throughout the years.

My decision to become a grad student at Harvard/MIT, and many fantastic experiences since, can be traced to my participation in the i2b2/HST Summer Institute. I cannot thank Susanne Churchill enough for making the decision to invite me to the program in the first place and for granting me the opportunity to continue being involved throughout grad school. Along the same lines, I want to thank Zak Kohane and Barbara Mawn for making it possible. Through the program, I had the opportunity to supervise many truly remarkable students: Jenn Ge, Maddy McDonald, Jeff Gerold, Carles Boix, Tony Wang, Kevin Mei, and Emily Levenson. Working with such talented students was one of the highlights of my grad school years. I sincerely thank you for the many things I learned from you.

I was also fortunate to share my grad school experience with incredibly talented and inspiring classmates. It all started with us taking classes together and culminated in the Biophysics Breakfast Club and our famous Wednesday lunches, which were always one of the highlights of my week. I really appreciated your company and support through both the good times and the darkest moments, George, James, Luvena, Stephanie, Hanlin, Vikram, Thomas and Drago. I would also like to thank my former housemates, friends, and sometimes classmates at the 10 Martin International Yet-To-Be-Named Cooking Co-op: Mike, Mingjie, Genya, Keisuke, Toni, Aaron, Oren, Danny, Shay and Adrian. Sharing a house and so many great meals with you was a defining feature of my early years as a grad student

and I will always look fondly at my time at the co-op.

A grad student's life provides plenty of challenges, even when everything goes right. I was extremely lucky to have a team of academic advisors and staff to guide me through the process and deal with any issues. I am sincerely grateful for the work of Jim Hogle and Michele Jakoulov in the Biophysics program, and Julie Greenberg, Laurie Ward and Traci Anderson in the HST program. Similarly, I thank Karen Barry and Derek Aylward for their significant help in navigating the BWH and MIT bureaucracies, respectively.

I would also like to thank the people without whom I would never have reached the finish line: my family and friends from times past. Without the overwhelming support from my family, I would probably have never come to the U.S., let alone have the opportunity to study at Harvard and MIT. Having such a supportive group of people with whom to celebrate my successes and be comforted during the difficult moments was essential for surviving graduate school.

Last, but certainly not least, I would like to thank my wonderful girlfriend, Abby. Meeting you was the best thing that happened during grad school. Sharing this time with you has helped turn the last few years into the best of my life. I cannot be more grateful for your unconditional love and support. I am incredibly happy and excited to be able to share the end of this journey with you.

The laws of genetics had never depended upon knowing what the genes were chemically and would hold true even if they were made of green cheese.

Ed Lewis

1

Introduction

The fundamental goal of genetics is to elucidate the mechanisms that underlie heredity. For most of the field's history, progress was made in spite of almost complete ignorance of the molecular mechanisms responsible for linking genotypes to phenotypes. The discovery that DNA was the macromolecule that carried genetic information[1] and the subsequent determination of its double-helical structure[2] provided fundamental clues about the molecular basis of genotype-phenotype relationships. Several decades later, genetic linkage analysis[3] enabled the identification of the molecular basis of disorders caused by mutations in a single gene, such as cystic fibrosis[4] and Huntington's disease[5].

Yet, it was only after the DNA-sequencing revolution that began with the publication of the human genome sequence in 2001[6] that it became possible to study the genetic basis of human traits on a large scale. This technological revolution has created unprecedented scientific opportunities, but also deliv-

ered many surprises. While a rapidly growing number of genetic variants have now been statistically associated with phenotypes[7], knowledge about the mechanisms through which such variants exert their phenotypic effects has lagged behind. One of the most surprising insights was the realization that over 80% of variants associated with complex phenotypes in humans affect noncoding DNA[8,9]. Developing improved capabilities for interpreting both coding and noncoding genetic variants is one of the key challenges in genetics and genomics today.

In this section, I first provide an overview of the state of the field and describe the process through which we have established that genetic variants that alter gene regulation play an important role in explaining phenotypic variation in humans. Next, I review key concepts related to transcription factors and regulatory sequences. Finally, I explain how the individual projects described in this dissertation represent important steps towards the ultimate goal of characterizing regulatory variation in humans and understanding its contribution to phenotypic variation.

Genetic variation and and association studies

Most genetic variation between human individuals exists in the forms of single nucleotide polymorphisms (SNPs)[10], which involve a change in the identity of a single DNA base pair (bp) in the genome. Each haploid human genome is composed of ~3 x 10^9 bp[6]. Due to a combination of chemical damage and spontaneous errors in DNA replication and repair, each of these base pairs is subject to a mutation rate of ~10^{-8} per generation[11,12]. As a consequence, each individual is born with 40-80 *de novo* single base pair mutations, with most of the variance being explained by the age of the father[13]. This mutational process, compounded over thousands of human generations, has created an extraordinary amount of genetic diversity. Sequencing projects have now identified over 112 million unique SNPs[14], a number that will continue to grow as more individuals are sequenced.

Genome-wide association studies (GWASs) have greatly facilitated the process of linking genetic variants to common disease phenotypes. The typical study design for a GWAS involves selecting a

group of individuals afflicted by a particular disease (cases) and a group of unaffected individuals (controls)[10]. Ideally, the populations are matched as closely as possible in terms of ethnicity, sex, and age distribution in order to minimize the influence of confounders. In its simplest form, a GWAS involves applying a chi-squared (χ^2) test for each genotyped or imputed SNP to determine if one allele exists at higher frequencies in the cases than in the controls. The underlying assumption is that an allele occurring at higher frequencies in the cases than in the controls is more likely to be a risk factor for a particular disease. The SNPs that pass a stringent threshold of genome-wide statistical significance (typically, a P-value $< 10^{-8}$) are then considered to be associated with the phenotype of interest. In practice, a logistic regression is often the preferred method to test for SNP-disease associations, as it is conceptually equivalent but allows covariates to be modeled. Additionally, the case-control study design can be modified to study quantitative phenotypes, in which case the logistic regression is replaced with a generalized linear model[10]. Regardless of the study design, the main output of a GWAS is a list of SNPs associated with the phenotype of interest and a corresponding level of statistical significance for each SNP.

However, even a high degree of statistical significance does not imply that a particular SNP is causal; *i.e.*, that a SNP is actually involved in the molecular mechanism responsible for the phenotype. This is partly due to linkage disequilibrium (LD), which describes the phenomenon whereby alleles in physical proximity along a chromosome co-occur in non-random patterns[15]. The existence of LD is a result of the positional bias of genetic recombination: if two SNPs are in sufficient proximity that essentially no recombination events take place between them, the corresponding alleles will remain correlated throughout subsequent generations. Therefore, without further evidence, SNPs identified in GWAS can only be considered "tag SNPs" that indicate the presence of a genetic association with a putative causal variant in the vicinity of the tag SNP. On average, a common SNP will have a surrounding region of LD (or "LD block") extending for ~60 kb, although the exact number varies across loci and between human populations[15]. Therefore, while the tag SNP itself can in principle be the causal vari-

ant, this can rarely be established without further analysis.

In limited cases, strong candidates for being causal SNPs can be identified within the LD block of a tag SNP. The most common case involves the presence of a non-synonymous SNP (nsSNP) in a gene of relevant biological function for the phenotype of interest[16]. For example, a nsSNP causing a T300A substitution in ATG16L1 was suggested as a candidate by a GWAS and later validated as increasing the risk of developing Crohn's disease[17]. Another possible scenario is that the tag SNP is forming a "synthetic association" by being in LD with a rare coding variant that was not genotyped in the same study[18]. While simulations have shown that synthetic associations are theoretically possible[18], later studies have failed to find evidence that such associations between common and rare variants are a widespread phenomenon[19].

Increasing evidence suggests that coding variants are unlikely to be responsible for the majority of GWAS signals. Approximately 93% of reported GWAS tag SNPs are located in noncoding sequences[9]. While this number includes intronic variants, only 11% of intronic GWAS SNPs are in strong LD with coding SNPs. Instead, 76.6% of GWAS variants were found to be either within, or in complete LD with SNPs in genomic regions of open chromatin, which are predominantly noncoding and often associated with regulatory elements[9]. In a few cases, regulatory SNPs have been directly implicated in disease mechanisms. For example, the rs12740374 variant creates a binding site for the transcription factor C/EBP, which in turn alters the hepatic expression of the SORT1 gene[20]. This SNP was found to explain GWAS signals associated with higher low-density lipoprotein cholesterol and increased risk of myocardial infarction. However, the vast majority of noncoding GWAS variants have not been associated with such mechanisms.

These observations have highlighted a pressing need to develop better approaches for studying noncoding sequences. While nsSNPs often provide an intuitive hypothesis about the biological mechanism behind the association, the process of generating and testing hypotheses about the effects of noncoding variants remains significantly more challenging. Although intergenic sequences harbor di-

verse elements that could be affected by genetic variation, such as many classes of noncoding RNAs[21], alterations in regulatory elements are likely to account for a significant number of functional associations[22,23]. Before fully describing the arguments for why regulatory variants are of particular interest, I will review several key concepts about the molecular biology and computational modeling of transcriptional regulation.

Transcription factors and models of binding specificity

The ability of multicellular organisms to respond to external stimuli and orchestrate complex developmental programs relies largely on exercising precise control of gene expression. Transcription factors (TFs) represent a broad category of proteins that have evolved to regulate the specific portions of an organism's genome that are transcribed at a given time and cellular state. In broad terms, TFs can be classified into two groups: general and sequence-specific. General TFs encompass the proteins that are required for forming the pre-initiation complex at eukaryotic promoters, which enables the subsequent recruitment of RNA polymerase and the initiation of transcription[24]. Meanwhile, sequence-specific TFs bind to regulatory sequences in the genome to exert control over gene expression, most commonly through domains that participate in protein-protein interactions, leading to trans-activation or trans-repression of their target genes. Throughout this dissertation, I employ the "TF" designation to refer to sequence-specific TFs, unless otherwise specified.

The most widely used scheme for classifying transcription factors is based on the structure of their DNA-binding domains (DBDs)[25]. DBDs are largely responsible for enabling TFs to recognize their target DNA sequences with high specificity. Classifying TFs by the structural classes of their DBDs enables inference about homology relationships, likely modes of protein-DNA recognition, and in some cases, the types of target sequences bound by a given TF. In addition, homologous proteins within the same DBD class but with different binding preferences can be compared to identify the amino acids involved in determining sequence specificity[26–29]. Finally, DBDs from a given class tend to

have a high level of amino acid sequence similarity that allows their position within a protein sequence to be identified computationally[24].

DBDs' ability to bind only certain DNA sequences with high affinity is derived from several biophysical factors. First and foremost, energetically favorable interactions between bases and residues, particularly in cases where hydrogen bonds are formed, contribute significantly to the selectivity for specific sequences[30]. For example, alanine side chains in the major groove can form two hydrogen bonds with guanine[31]. However, base-residue interactions do not happen in isolation: the protein backbone may not be able to adopt a conformation that enables all favorable residue-base interactions to happen simultaneously. In some cases, well-positioned water molecules can also mediate contacts between bases and residues[32]. The added contribution of these effects makes many TFs able to bind a fairly degenerate set of sequences, the identity of which is not always easy to predict[33]. Therefore, even in this simplified model, DBD-DNA interactions can occur within a relatively complex energetic landscape.

A significant amount of effort has been dedicated to developing models of the energetic landscapes, or approximations thereof, of interactions between TFs and their binding sites. The core of most models is the representation of TF-DNA binding as a kinetic process, in which a TF and a DNA molecule, both in solution, form a complex at a rate K_{on} and dissociate into free components at a rate K_{off} (Eq. 1.1).

$$\text{TF} + \text{DNA} \underset{K_{\text{off}}}{\overset{K_{\text{on}}}{\rightleftharpoons}} \text{TF} \cdot \text{DNA} \tag{1.1}$$

Often, it is sufficient to know the fraction of DNA sequences bound by TFs at equilibrium. In such cases, the rate constants K_{off} and K_{on} can be combined to form an equilibrium constant, $K_d = K_{\text{off}} / K_{\text{on}}$, which is commonly referred to as a dissociation constant. The value of K_d can be obtained directly by measuring the concentrations of reactants and products at equilibrium. In addition, at

constant temperature, the value of K_d is directly related to the Gibbs free energy (ΔG) of the TF-DNA interaction (Eq. 1.2).

$$K_d = \frac{K_{\text{off}}}{K_{\text{on}}} = \frac{[\text{TF}]_{\text{eq}}[\text{DNA}]_{\text{eq}}}{[\text{TF} \cdot \text{DNA}]_{\text{eq}}} = e^{\frac{\Delta G}{RT}} \tag{1.2}$$

Although measuring K_d is conceptually simple, in practice it is often a laborious process, particularly when many sequences are involved. Most TFs recognize DNA sequences in the 6-12 bp range[34]. Therefore, for a typical TF, it would be necessary to measure K_d values corresponding to ~4^6 - 4^{12} sequences. Traditional approaches for measuring K_d values, such as electrophoretic mobility shift assays (EMSAs)[35] or surface plasmon resonance[36] are too laborious to employ in such large scales. Microfluidic approaches, such as mechanically induced trapping of molecular interactions (MITOMI), enable the simultaneous measurement of K_d values for thousands of DNA sequences[37]. However, to date, MITOMI has only been used to measure the binding sites of few TFs.

A more feasible approach is to avoid measuring K_d values for all sequences and instead model the relative change in K_d value for a particular sequence relative to the optimal binding site (*i.e.*, the sequence with the smallest K_d value). For each sequence S, a relative $K'_d(S)$ value defines its relative affinity compared to the dissociation constant of the optimal binding site, K_d^{opt}. Each value of $K'_d(S)$ is in turn associated with a change in free energy $\Delta\Delta G(S)$ (Eq. 1.3). In this transformation, information about TFs' DNA-binding affinity (*i.e.*, absolute binding preference) is lost. However, information about DNA-binding specificity (*i.e.*, relative preference between different sequences) is preserved, which suffices for many applications.

$$K'_d(S) = K_d^{\text{opt}} e^{\frac{\Delta\Delta G(S)}{RT}} \tag{1.3}$$

In practice, most models of TF binding preferences focus on describing these $\Delta\Delta G(S)$ values, or equivalent representations thereof. However, instead of assigning a $\Delta\Delta G(S)$ value to each possible

sequence S, most models rely on assumptions that reduce the number of model parameters. For example, the widely used position weight matrix (PWM) model assumes that the free energy terms associated with the contacts between the TF and each possible base in the binding site can be broken down into additive contributions. Then, the free energy term can be calculated as in Eq. 1.4, where s_j represents the identity of the nucleotide at position j in the binding site, $\Delta\Delta G_i(s_j)$ is the free energy term associated with the substitution of s_j at position i in the binding site, and L is the total length of the binding site. In other words, the free energy contributions are modeled by a 4 x L matrix, each describing the additive effect of substituting a given nucleotide at each position in the binding site.

$$\Delta\Delta G(s) = \sum_{i=1}^{L} \Delta\Delta G_i(s_j) \tag{1.4}$$

This biophysical definition of PWMs has a corresponding statistical interpretation, which is how PWMs are frequently reported in the literature. Through the Boltzmann distribution, the $\Delta\Delta G$ values can be converted into probabilities for observing each base at each position in the binding site. Intuitively, these values describe the probability of observing a specific sequence S among the total set of sequences bound by the TF, under the assumption that the library is randomly generated and fully represents all possible sequences. In practice, this probabilistic representation of a PWM can be estimated empirically by aggregating the sequences bound by the TF and counting the frequency with which a given base appears at each position in the binding site. As a consequence, the probabilistic representation is sometimes described as a position frequency matrix (PFM). A comparison between the two is shown in Figure 1.1. Throughout this dissertation, I will use the term PFM to refer exclusively to this frequency-based representation and energy matrix (EM) to refer to the free energy representation described above. In most contexts, these mathematical representations are equivalent, and I will simply refer to these models as PWMs.

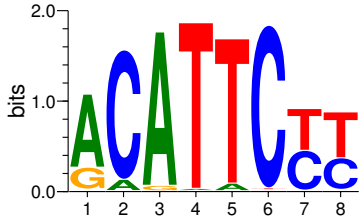
Several concepts related to PWMs are used frequently in the literature, but unfortunately, are rarely

$$\begin{pmatrix} A & 0.00 & 2.44 & 0.00 & 4.27 & 2.95 & 7.96 & 5.78 & 4.32 \\ C & 4.33 & 0.00 & 7.29 & 6.93 & 7.7 & 0.00 & 0.05 & 0.37 \\ G & 1.21 & 5.91 & 3.32 & 7.09 & 6.3 & 4.88 & 4.14 & 3.81 \\ T & 4.25 & 6.92 & 6.46 & 0.00 & 0.00 & 4.46 & 0.00 & 0.00 \end{pmatrix}$$

Energy matrix (kT)

$$\begin{pmatrix} A & 0.754 & 0.080 & 0.963 & 0.014 & 0.049 & 0.000 & 0.002 & 0.008 \\ C & 0.010 & 0.917 & 0.001 & 0.001 & 0.000 & 0.981 & 0.482 & 0.399 \\ G & 0.225 & 0.002 & 0.035 & 0.001 & 0.002 & 0.007 & 0.008 & 0.013 \\ T & 0.011 & 0.001 & 0.002 & 0.984 & 0.948 & 0.011 & 0.508 & 0.580 \end{pmatrix}$$

Position frequency matrix



Sequence logo

Figure 1.1: Position weight matrices and sequence logos. This figure displays the energy matrix (top), position frequency matrix (middle) and sequence logo (bottom) of the yeast TF Tec1 as measured by a protein-binding microarray (data obtained from UniPROBE [38]).

used consistently. Here, I use TF binding site "motifs" and PWMs as interchangeable, while using "motif instance" or sometimes "motif match" to refer specifically to a genomic sequence that is predicted to be bound by a TF using a PWM model. In contrast, the "consensus" sequence corresponds to the optimal binding site as predicted by the PWM. Another useful concept is that of a sequence logo, or "logo" for brevity, which displays the parameters of a PWM in an information-theoretic manner. In a sequence logo, the height of the letter stack at each position in the binding site corresponds to the relative entropy, or information content (IC), associated with the multinomial probability distribution for that position. Intuitively, taller stacks indicate larger deviations from uniform (random) distributions. The relative heights of the letters within each stack are in turn proportional to the frequencies for each base in the corresponding position in the PFM. An example logo is shown in Figure 1.1.

Despite the apparent intricacies of protein-DNA interaction energy landscapes, PWM models have been remarkably successful at predicting both *in vitro* and *in vivo* TF binding[39]. However, there are cases where PWMs are not the most appropriate model choice. For example, some TFs have energetic interdependencies for contacting adjacent nucleotides[40], which contradicts the main assumption of the PWM model. To address such limitations, higher order models, such as those that consider all possible di-nucleotide interactions, have been developed and employed successfully to improve binding site predictions[39].

At the other end of the spectrum, TF specificities can be modeled without making strong assumptions by assigning a score to every possible DNA k-mer (*i.e.*, a sequence of length k). A disadvantage of full k-mer models is that they have a significantly higher number of free parameters compared to PWM or di-nucleotide models. However, if appropriate data are available to fit a more complex model, such an approach can be advantageous. For example, protein-binding microarrays (PBMs) enable the measurement of the relative sequence preferences of a TF to all DNA 8-mers[41]. In statistical terms, the choice of TF specificity model represents the traditional bias-variance tradeoff[42]. Throughout this dis-

sertation, I will employ both PWM and full *k*-mer models, depending on the details of the application and on the availability of sufficient training data. When models are properly selected, such decisions can greatly facilitate computational analyses.

Transcription factors and regulatory elements

In order to modulate gene expression, TFs bind a wide range of sequences with regulatory functions, including promoters, enhancers, silencers and insulators. Broadly, promoters are divided into two parts: the core promoter, which is ~40 bp upstream of the transcription start site (TSS) and contains the sequence elements necessary for the formation of the pre-initiation complex, and the proximal promoter, which comprises a region of ~1-2 kb upstream of the TSS[43]. Proximal promoters often contain binding sites for TFs, which can exert activating or repressive effects on transcription from the nearby TSS upon binding.

Although promoters are an essential component of the regulatory machinery, spatiotemporally precise control of gene expression is achieved primarily through the action of enhancers. Enhancer sequences are typically hundreds of base pairs in length and located more distally from TSSs than proximal promoters[34]. A recurring feature in enhancer sequences is the presence of TF binding sites clusters[44], often for multiple distinct TFs. This architecture allows enhancers to integrate inputs from multiple signaling pathways, which in turn facilitates the creation of precisely defined expression patterns[45]. For example, stripe patterns created by gap enhancers in early *Drosophila* development are achieved through a combination of binding by widely expressed activator TFs and spatially restricted repressor TFs[46].

While enhancer sequences show considerable heterogeneity, they also possess some important recurring features. An important consideration is that, due to the presence of nucleosomes, TFs only occupy a small fraction of genomic instances of their consensus sequence[47,48]. Therefore, most enhancer sequences must typically become accessible to binding by TFs before robust transcriptional activation

can be achieved. At its core, the underlying process is a competition between TFs and nucleosomes for occupying the same DNA sequence[49]. Under the right conditions, such as when sufficient nuclear concentrations of specific TFs are reached, the equilibrium can shift in favor of TF binding over nucleosome occupancy. This process is facilitated by a variety of mechanisms, such as favorable energetic interactions between TFs that lead to cooperative binding and the recruitment of chromatin remodelers through protein-protein interactions with TFs[44]. A class of TFs known as pioneer factors is able to directly bind nucleosomal DNA, which can in turn facilitate the recruitment of other TFs to the same sequence[50]. Another important feature of enhancer sequences is that, once bound, TFs can form protein-protein interfaces that recruit co-activator proteins, as is the case in the well known IFN-β enhancer[51].

Identifying the genomic sequences that function as enhancers in different cell types and conditions has been one of the major goals of the genomic era. At first, these efforts were primarily driven by the desire to understand enhancers' role in development and their contribution to morphological changes throughout evolution. However, in light of observations about the abundance of GWAS variants in noncoding sequences, identifying enhancers is increasingly seen as an essential task for identifying putative causal variants and understanding their mechanistic consequences. As described earlier, LD blocks around tag SNPs identified through GWAS often extend for tens of kilobases[15]. Genomic annotations that delineate enhancer sequences can allow specific variants within LD blocks to be prioritized for further study. If the tissue or cell-type specificity of those enhancers is known, such information can provide further insights about potential mechanisms underlying the phenotypic association. To provide a context for the advances described in Chapter 2, I will review the evolution of experimental and computational methods for identifying enhancers and their applications in prioritizing noncoding variants.

Goal #1: Developing better tools to identify tissue-specific enhancers

Unlike protein-coding genes, transcriptional enhancers do not possess sequence properties that allow them to be easily identified by computational methods. Even before the full sequence of the human genome was published, several algorithms achieved satisfactory performance at identifying likely protein-coding sequences in mammals[52,53]. These algorithms typically relied on hidden Markov models (HMM) that were able to identify sequences depleted for stop codons while accounting for the presence of introns. However, transcriptional enhancers lacked such prominent, consistent features that would enable their computational identification. Therefore, while the protein-coding sequences in the human genome were mostly annotated by the time the human genome sequence was published, enhancer sequences have proven significantly more difficult to characterize.

Early algorithms for the prediction of transcriptional enhancers relied on identifying putative TF binding sites in genomic sequences. Analyses of known enhancers revealed that clusters of TF motif instances were a recurring feature of enhancer sequences[54]. These observations were used as the basis of early methods to predict enhancer sequences, which achieved moderate success in predicting developmental enhancers in *Drosophila*[55]. Later computational methods incorporated other features, such as the evolutionary conservation of motif instances[56] and the orientation and relative position of motif matches[57].

However, several important limitations restricted the ability of computational approaches to be applied on a large scale. First, such methods depend on knowing the combination of TFs that are relevant for transcriptional activation in a particular cell or tissue type, along with the DNA-binding specificity of those TFs. While such information was available for a subset of well studied systems, such as the *Drosophila* blastoderm or a handful of mammalian tissues, the lack of comprehensive knowledge of *cis*-regulatory codes limited the applicability of computational methods.

Another difficulty lies in identifying whether genomic sequences require a specific "grammar" in

order to become active enhancers. Generally speaking, there are two opposing models of enhancer function: the enhanceosome model[58], which posits that most enhancers have precise requirements in terms of the orientation, order and identity of TF binding sites; and the billboard model[59], in which most enhancer sequences have a flexible architecture that functions regardless of the specific properties of their TF binding sites. In practice, it is likely that most enhancer sequences fall on a spectrum between these two extremes[34]. However, not knowing whether a particular sequence is sensitive to grammar features makes predictions more difficult. Theses difficulties are compounded by the difficulty in predicting TF binding sites in genomic sequences. As described before, most motif instances are not actually bound *in vivo*[60]. Although evolutionary conservation can improve the ability to predict *in vivo* binding sites, conserved motif instances may not actually be bound in the tissue of interest[61,62]. In practice, each of these limitations can lead to false positives and negatives and has limited the applicability of methods to computationally predict enhancers.

Enhancer identification was greatly aided by the development of experimental methods to identify TFs' *in vivo* binding sites genome-wide. An essential milestone was the development of chromatin immunoprecipitation (ChIP) followed by microarray quantification of the precipitated DNA fragments (ChIP-on-chip)[63]. The typical protocol of a ChIP-on-chip experiment can be summarized as follows: (1) bound proteins are cross-linked to genomic DNA within the nucleus; (2) the nuclei are lysed and genomic DNA is fragmented by sonication; (3) an antibody that is specific for the protein of interest is used to select for bound DNA fragments; (4) cross-linking is reversed; and (5) the abundance of bound sequences is determined by fluorescently labeling the fragments and quantifying them through hybridization in a single-stranded DNA microarray.

In one prominent study, the ChIP-on-chip technique was applied successfully to identify enhancers regulating the development of the *D. melanogaster* mesoderm[64]. Using ChIP-on-chip, Zinzen et al. profiled the *in vivo* binding patterns of five mesodermal TFs (Twi, Tin, Mef2, Bap and Bin) at key developmental timepoints. Sequences bound jointly by various combinations of these five TFs were

tested for enhancer activity using *in situ* hybridization of a *lacZ* reporter driven by each sequence. The genomic sequences with clustered TF binding sites often drove mesodermal expression in well-defined spatiotemporal patterns, highlighting the value of this approach for enhancer prediction.

Although such approaches, based on TF ChIP-on-chip, have led to significant insights, they also have important limitations when it comes to identifying enhancers. First, such studies require knowledge of the TFs involved in the regulation of the tissue or cell-type of interest. Even more importantly, a successful experiment requires the availability of an antibody with sufficient specificity and avidity for each TF of interest. In many cases, these are not available, significantly hampering the ability to identify enhancers in many tissues. Additionally, ChIP-on-chip methods rely on the ability to quantify sequences from most of the genome in a microarray with a limited number of probe sequences. While DNA probes tiling most of the fly genome could be easily designed to fit on a microarray slide, much larger mammalian genomes were tedious and expensive to assay in this manner, limiting the applicability of ChIP-on-chip to identify mammalian enhancers. Finally, a significant fraction of sequences bound by TFs *in vivo* did not drive transcriptional activity in reporter assays[65,66]. Therefore, such methods are expected to create a significant number of false positive enhancer predictions

The ability to use ChIP-based approaches to assay TF binding in larger genomes was enabled by the development of ChIP-Sequencing (ChIP-Seq). Although most of the experimental protocol is similar to ChIP-on-chip, instead of using microarray hybridization to quantify bound DNA sequences, ChIP-Seq uses next-generation sequencing (NGS) to directly measure the abundance of bound fragments[67]. The transition to sequencing-based assays enabled the genome-wide profiling of TF binding in mammalian genomes at a higher resolution than is practically feasible by ChIP-on-chip. In many cases, the genome-wide binding profiles of TFs have proven useful for enhancer prediction. For example, genomic loci bound jointly by Oct4, Sox2 and Nanog, as identified by ChIP-Seq in mouse embryonic stem cells, were found to often drive expression in embryonic development when tested in reporter assays[68].

However, a key breakthrough was achieved by using ChIP-Seq to assay not just TFs' binding sites, but the genomic positions of other, more general indicators of enhancer activity. For example, p300 plays important roles as a transcriptional coactivator and as a histone acetyltransferase[69]. When ChIP-Seq was used to assay p300 binding in mouse embryonic tissues, it was discovered that p300-bound sequences often drove expression in the corresponding tissue during embryonic development[69]. In addition, the presence of certain histone modifications in the nucleosomes surrounding enhancer elements was found to be predictive of enhancer activity[70]. For instance, the presence of a mono-methylated lysine 4 in histone 3 (shortened as H3K4me1) was indicative of increased transcriptional output of the associated sequence in reporter assays. Other histone modifications were later found to be predictive of enhancer activity, such as H3K27ac and H3K9ac[71]. With the development of highly specific antibodies for a large set of histone modifications, genome-wide identification of putative enhancer sequences became feasible, although contingent on the ability to perform a successful ChIP-Seq experiment on the biological sample of interest.

Another important milestone was the development of sequencing-based methods to assay chromatin accessibility. The binding of TFs to DNA, as in enhancer elements, will lead to the displacement of nucleosomes[72]. The enzyme Deoxyribonuclease I (DNase I) is an endonuclease that preferentially cleaves phosphodiester bonds in DNA[73]. Importantly, DNase I's cleavage rates are increased in DNA sequences with lower nucleosome occupancy. Based on these principles, DNase-seq[74] was developed as a massively parallel, sequencing-based method to assay genomic regions of high chromatin accessibility, here called DNase hypersensitive sites (DHSs). Because a significant fraction of TF-bound regions are thought to act as enhancers, DNase-seq has proven useful in creating large scale maps of candidate enhancer regions across a wide range of tissues[23,72]. However, an important limitation is that not all DHSs function as active enhancers: insulator sequences bound by CTCF are DNase I-accessible but do not drive enhancer activity[75]. Similarly, repressed enhancers also display the characteristic features of DHS regions but are transcriptionally inactive[76].

The data obtained through ChIP-Seq and DNase-seq have been used to train more complex models that can be used to predict enhancer activity. For example, ChromHMM[71] and Segway[77] enable the unsupervised discovery of recurrent patterns of histone modifications and DHSs across the genome. Certain patterns, or chromatin states, identified by such methods have been found to be associated with enhancer activity. For example, the joint presence of the histone modifications H3K27ac, H3K4me1 and H3K9ac in a genomic region is predictive of activity in luciferase assays[71]. Unsupervised approaches based on profiling histone modifications and DHSs have been applied to predict candidate enhancers in a wide range of tissues, leading to identification of over 1 million putative enhancer sequences[23].

However, while these methods have proven extremely useful for generating genome-wide maps of putative regulatory elements, they have several important limitations. First of all, these techniques measure features that are known to be correlated with enhancer elements, but do not directly assay whether a sequence is capable of driving expression from a core promoter. In a sizeable proportion of cases, genomic regions with histone marks characteristic of enhancer activity do not actually drive expression in reporter assays[60,78]. Additionally, these methods typically require large numbers of cells[79], which has limited their applications beyond cell lines and a subset of tissue types for which large samples can be obtained.

The method described in Chapter 2, enhancer-FACS-Seq (eFS), takes advantage of recent methodological developments in enhancer prediction while addressing their weaknesses. Chiefly, eFS enables hundreds of candidate sequences to be assayed directly for tissue-specific enhancer activity *in vivo*. The direct confirmation of transcriptional enhancement by a candidate sequence remains the "gold standard" test for whether a sequence can be considered an enhancer[69,80]. Therefore, eFS occupies an essential niche in the enhancer discovery and validation pipeline, where enhancers predicted by genome-wide methods can be further screened for transcriptional activity and tissue or cell-type specificity.

The eFS approach possesses important advantages over other competing methodologies. For example, STARR-seq[76], MPRA[81] and CRE-seq[60] are all alternative approaches for enhancer screening. However, uniquely among these methods, eFS allows screening experiments to be carried out with relatively low (~10,000) numbers of cells. This feature is particularly useful for enhancer screening in rare cell types, for which obtaining large enough samples to study by other methods may not be feasible. In addition, eFs experiments are carried out in a more natural chromatin context (*i.e.*, an actual chromosome) whereas other methods rely on measuring expression from plasmids.

The advantages of eFs are likely to be particularly useful for the prioritization of causal noncoding variants. For example, a significant number of tissue-specific GWAS enrichments have been reported in samples of complex tissues[23], such as brain, lung, liver and pancreas, An approach like eFS can potentially enable the screening of the enhancers in a more restricted subset of cells, allowing high confidence associations between enhancer sequences and the cell-types or tissues in which they drive expression. Importantly, having control over the exact sequences tested can allow genetic variants to be assayed to determine if they cause changes in enhancer activity. In Chapter 6, I discuss both the potential and the challenges of extending eFS to mammalian systems in order to assay candidate enhancers and their genetic variants.

However, finding enhancers that may harbor causal variants for common disease is only the first step. Several genetic variants may be present within an enhancer sequence, any of which could be responsible for the regulatory changes that lead to phenotypic differences. The most natural hypothesis is that at least one variant is disrupting a TF binding site that is necessary for proper enhancer function, or potentially creating a new binding site that causes aberrant activity. Indeed, several studies have followed this line of reasoning to successfully link changes in TF binding sites with increases in risk for developing type 2 diabetes[82], prostate cancer[83], obesity[84] and systemic lupus erythematosus (SLE)[85].

However, despite the intuitive appeal of the "enhancer SNP disrupts TF binding site" model, func-

tional studies have have cast doubt on whether such variants can account for most differences in TF binding across individuals. Kasowski *et al.* employed ChIP-Seq to compare the patterns of NF-κB binding in 10 individuals and found that only 2-3% of differentially bound regions were associated with changes in NF-κB motif instances[86]. In a separate study, Reddy *et al.* looked at instances of allelic imbalance in the ChIP-Seq signal of 24 TFs (*i.e.*, differential binding between maternal and paternal chromosomes)[87]. While differentially bound regions were enriched for SNPs that disrupted motif instances, only ~12% of differences could be explained by changes in motif sequences. Yet, allelic imbalances in binding have been shown to be heritable and frequently transmitted from parents to children[88], implying that the mechanisms underlying such binding variation are likely to have a substantial genetic component.

Goal #2: To better understand the *in vivo* determinants of TF binding, using NF-κB as a model system

As previously discussed, a significant gap exists between the binding patterns of TFs *in vivo* and what would be predicted purely on the basis of motif matches in genomic sequences. Most prominently, the vast majority of binding sites predicted purely on the basis of sequence are not occupied *in vivo*[60]. At the same time, widespread binding has been observed at genomic loci that do not harbor motif instances for the bound TF[89]. A better understanding of the mechanisms that influence TF binding *in vivo* is essential for bridging this gap. The discovery of additional features that can predict *in vivo* binding could be used to fine tune predictions of genetic variants with functional consequences.

In this section, I will argue that studying the genomic binding patterns of NF-κB is of significant value towards understanding both the *in vivo* determinants of TF binding and the transcriptional regulation of immune responses. NF-κB is the name given to homo- and hetero-dimers of certain proteins with Rel homology domains. In humans, there are five subunits of the NF-κB family, each encoded by a different gene: RelA/p65, RelB, cRel, p50 and p52. NF-κB dimers play central roles

in signaling pathways for many key biological processes, including immune responses, lymph node development, synaptic plasticity, and cell-fate determination[90].

NF-κB is an ideal target for the study of the determinants of *in vivo* TF binding patterns and how they vary across individuals. First, the binding of NF-κB subunits containing RelA has already been shown to vary significantly across individuals, with over 7.5% of bound loci being variable[86]. Second, different types of NF-κB dimers possess distinct binding specificities *in vitro*[91], but how these specificities translate into *in vivo* binding differences remains poorly understood. Furthermore, because NF-κB binds DNA in dimeric form, this creates an additional layer of complexity: certain binding sites may be occupied by some dimers but not others. Because certain NF-κB dimers act primarily as activators (*e.g.*, RelA:p50) whereas others act as repressors (*e.g.*, p50:p50)[92], mechanisms that alter the preferences of dimers for specific binding sites could lead to transcriptional changes.

In addition, the exact sequence of NF-κB binding sites has been shown to affect transcriptional output through allosteric mechanisms. The identity of the central nucleotide in the traditional κB binding site can modulate co-factor recruitment, and therefore the extent of transcriptional activation, without significantly changing binding affinity[93]. For example, p52 homodimers in complex with Bcl3 generally recruit coactivators when occupying sites with G/C central bases, but recruit corepressors when binding sequences with central A/T bases[94]. Whether other properties of binding site sequences influence dimer recruitment and transcriptional output remains mostly unknown.

There are also abundant reasons for studying NF-κB binding for biomedical reasons. First, NF-κB activation has been associated with a range of autoimmune and inflammatory diseases, including rheumatoid arthritis, atherosclerosis, asthma and inflammatory bowel disease[95]. Genetic variants that disrupt an NF-κB binding site in an enhancer element regulating TNFAIP3 have been associated with increased risk of developing SLE[85]. Furthermore, NF-κB activation has increasingly been proposed as a mechanism linking inflammation to cancer through the induction of tumor promoting cytokines, such as IL-6 or TNF-α[96]. Yet, despite considerable progress in understanding the signal transduction

pathways that lead to NF-κB activation, little is known about how those signals are integrated into NF-κB's functions in the nucleus. Identifying the subunits that regulate specific pro-inflammatory genes is of significant therapeutic interest.

Another key question about NF-κB relates to the roles played by each of the subunits and the different dimers they can form. The five NF-κB subunits share an N-terminal Rel-homology domain, which functions as both a DBD and a dimerization domain[90]. RelA, RelB and c-Rel all possess C-terminal transactivation domains. In contrast, p50 and p52 have C-terminal ankyrin repeats, which typically have trans-repressive effects[97]. Yet, despite the similarities in domain structure across subunits, each of them exhibits a distinct knockout phenotypes[98]. For example, *rela*$^{-/-}$ mice die during development as a consequence of TNF-α induced cell death, while *relb*$^{-/-}$ mice are viable but have deficiencies in lymphoid organs, dendritic cells and T-cells[98].

The signaling mechanisms that result in NF-κB activation have been traditionally divided into the canonical and noncanonical pathways. The canonical pathway is associated with rapid immune responses (particularly inflammation)[99], while the noncanonical pathway is associated with secondary lymphoid organogenesis, B-cell maturation and survival, and dendritic cell maturation[100]. The canonical pathway is primarily regulated through the phosphorylation and subsequent degradation of several inhibitors (IκBα, IκBβ and IκBε) that sequester NF-κB dimers in the cytoplasm. The noncanonical pathway, in contrast, is primarily controlled by the expression and processing of the p100 protein, which is the precursor for p52[100]. Whereas p105 is constitutively processed into p50[101], the processing of p100 into p52 depends on post-translational modifications triggered by specific signaling events[102]. Canonical pathway activation is associated with the formation of RelA:p50, c-Rel:p50, and c-Rel:c-Rel and RelA:RelA dimers, while noncanonical activation leads primarily to the formation of RelB:p52 and p52:p52[99]. Although this scheme is conceptually clear and appealing, there is evidence of crosstalk between the two pathways, primarily through the action of IκBδ, which is noncanonically regulated, upon RelA dimers[99]. Furthermore, the extent to which the nuclear functions of NF-κB reflect the

two pathway paradigm remains unclear.

For these reasons, it is important to study NF-κB binding in a cell type where both the canonical and noncanonical pathways are active, and therefore, all subunits are present in the nucleus. Immortalized lymphoblastoid cell lines (LCLs) can be created by infecting primary B cells with Epstein-Barr Virus (EBV)[103]. In LCLs, the EBV-encoded protein LMP1 mimics the action of CD40 and activates the canonical and noncanonical NF-κB pathways[104]. In addition, studying the features that direct *in vivo* NF-κB binding can be facilitated by performing experiments in a cell line for which complementary functional data are already available. The GM12878 LCL was selected by the ENCODE project for extensive profiling[22], including TF and histone modification ChIP-Seq and RNA-sequencing. Therefore, GM12878 represents an ideal system in which to study the complexities of NF-κB genomic binding and its connection to signaling pathways.

In the work described in Chapter 3, we examined the patterns of NF-κB binding by generating and analyzing high-quality ChIP-Seq data for all five subunits in GM12878. This dataset is the first reported instance where the binding sites of all subunits has been simultaneously mapped genome-wide. Analyzing these data provided several new insights into the determinants of NF-κB binding, including newly appreciated dependencies on flanking sequences and putative modes of indirect binding through recruitment by other TFs. This expanded model of binding determinants is likely to be useful in identifying additional genetic variants that may affect NF-κB binding. More generally, these results highlight the value of detailed *in vivo* profiling for TFs that are potentially relevant to specific cell types or phenotypes. In Chapter 6, I discuss the implications of these observations and describe how these data have already been used in other studies to identify regulatory variants that affect the risk of developing allergies.

Goal #3: To better understand *trans* regulatory variation in humans by systematically characterizing coding variants in transcription factors

Thus far, I have proceeded under the implicit assumption that differences in TF binding across individuals are primarily caused by changes in the sequences within or close to the binding sites. The notion that the local sequence context of TF binding sites is important in distinguishing bound vs. unbound motif instances is supported by functional evidence. For example, White et al. found that 84-bp sequences containing *Crx* motif instances were significantly more likely to drive expression in reporter assays when they were bound in retinal cells than when they were not[60]. However, other lines of evidence suggest that many, if not most, differences in TF binding across individuals cannot be explained by local sequence variation. For instance, ~2/3 of NF-κB binding sites that are variable across individuals lack any genetic variants within 200 bp[86].

Studies analyzing the inheritance patterns of gene expression levels have reached similar conclusions. The expression levels of most human genes have been found to be heritable across a wide range of tissues[105]. The proportion of variance that can be explained by additive genetic effects, also called the narrow-sense heritability (h^2), has been estimated at 15-35%[106]. However, variants detected through statistical approaches (described below) have been unable to explain the majority of the observed heritability[107,108].

These observations imply the presence of genetic variants that affect TF binding and gene expression across individuals, but have remain undetected thus far. Before explaining why such variants may have been missed by current methods, I will first introduce some useful concepts. A commonly used dichotomy for regulatory variants distinguishes those that act in *cis* and those that act in *trans*. There are a few, mostly overlapping definitions for these two classes. For example, Gaffney defines *cis* variants as those that exclusively modulate the expression of one allele, while *trans* variants affect both alleles[106]. Alternatively, *cis* variants can be defined as those affecting a gene whose TSS is located

within a certain distance of the associated variant (typically < 1 Mb), whereas *trans* variants encompass all other regulatory variants. Throughout this dissertation, I employ Gaffney's definition, but in practice the distinction will be irrelevant for the vast majority of cases.

The predominant method for finding genetic variants that alter transcript levels is called expression quantitative trait loci (eQTL) mapping. At its core, eQTL mapping is a form of GWAS, where the expression level of each transcript is modeled as a quantitative phenotype. Similarly to GWAS, the typical outcome of an eQTL mapping study is a list of tag SNPs are deemed to affect expression of at least one transcript at a level of confidence that exceeds a genome-wide statistical significance threshold[109]. As a consequence of mapping studies in multiple populations and tissue types, over 40,000 eQTLs have now been identified in humans (NCBI eQTL browser).

However, because of considerations related to statistical power, eQTL mapping studies have been strongly biased in terms of the types of variants they identify. In theory, any of millions of genetic variants observed in humans could have an effect on the expression of each of the ~20,000 human genes. If all such pairwise combinations were tested for statistical associations, the multiple hypothesis testing burden would be enormous[110]. Therefore, in practice, most eQTL studies have focused on genetic variants located in relative proximity (< 1 Mb) to the gene whose expression they modulate. This approach can significantly boost statistical power at the expense of strongly favoring the detection of *cis* over *trans* eQTLs variants.

Other factors contribute to the challenges of finding *trans* variants through eQTLs mapping. For example, computational approaches to correct for batch effects may sometimes incorrectly remove signals for eQTLs that affect large numbers of loci[111]. Similarly, eQTL mapping assumes that the total amount of RNA is comparable across samples. Therefore, eQTL mapping is not suited for identifying variants that cause widespread changes in absolute mRNA levels, an effect that has been observed, for example, when the concentration of c-Myc is varied[112]. Finally, there is evidence that the effects of *trans* eQTLs that affect signaling pathways may remain hidden when assaying steady-state expression

levels, as is the case in most studies[106].

Yet, the inability of *cis* variants to explain most of the observed heritability suggests that finding *trans* variants is likely to be important for the goal understanding the landscape of heritable regulatory variation. Because of the limitations of statistical methods described above, new approaches are needed in order to ascertain the prevalence and effects of *trans* variants. One potential approach is to develop methods to predict likely *trans* variants computationally and then validate them experimentally. However, such a strategy depends on the ability to identify genetic variants that are likely to have regulatory consequences.

Genetic variants that alter TF binding sites are regarded as a common molecular mechanism underlying *cis*-regulatory variation. An intuitive *trans* counterpart involves genetic variants that change TFs' DNA-binding preferences. For example, a coding mutations in a TF's DBD could rewire transcriptional networks by altering the genomic sequences that are bound at high occupancy. The relative importance of such *cis* and *trans* regulatory mutations has been extensively debated in evolutionary and developmental biology[113], without any consensus being reached.

In humans, many nsSNPs that alter TFs have been identified through large-scale sequencing studies. However, studying the consequences of coding variants in a systematic manner poses significant challenges. As previously discussed, eQTL mapping is statistically underpowered for detecting associations in *trans*. In the case of nsSNPs, this problem is further compounded by typically low minor allele frequencies[114], which further reduces statistical power. In theory, techniques such as ChIP-Seq could be used to study whether specific mutations alter the *in vivo* binding patterns of TFs. However, the throughput of such experimental approaches is low. There are thousands of variants that could potentially be tested and it remains difficult to predict which variants are likely to have an effect. As such, detailed functional studies have only been performed for a handful of TF mutations with known Mendelian disease associations[115].

In Chapter 4, I describe a combined experimental and computational approach to study coding

variants that affect the DBDs of human TFs. Because DBDs tend to be highly conserved both in terms of structure and amino acid sequence, co-crystal structures of protein-DNA interfaces and metrics of evolutionary conservation can be used to prioritize specific variants for experimental testing. In the aforementioned study, we show that protein-binding microarrays can be used to compare the binding properties of reference and alternative alleles across many TFs. We compare the DNA-binding perturbations caused by known Mendelian disease mutations and nsSNPs found in predominantly healthy individuals. This approach enables us to identify variants with potential regulatory and phenotypic consequences based on *in vitro* binding data. In addition, we develop a framework to identify additional, untested variants that are likely to affect DNA-binding. This work represents the first systematic study of coding variation in human TFs. The methodological advances and results derived from it are likely to be of significant value for future efforts to understand the contributions of TF nsSNPs to regulatory variation. In Chapter 6, I discuss the potential implications of widespread DBD variation in humans for understanding the heritability of gene expression.

Goal #4: Develop better tools to test hypotheses about the causality and phenotypic effects of regulatory variants

Thus far, I have described several approaches for identifying putative regulatory variants, such as eQTL mapping, enhancer assays and — in the case of coding variants — PBMs. Ultimately, the gold standard test for the causality of regulatory variants is to experimentally show that a specific sequence change alone can account for expression differences.

Traditionally, reporter assays have played a prominent role in testing the effects of putative *cis* regulatory variants. In a typical reporter assay, the enhancer sequence is placed upstream of a reporter gene whose expression levels can be quantified optically, either through enzymatic reactions that create visible products (*e.g.*, *lacZ*) or by fluorescence intensity (*e.g.*, GFP). However, determining that a genetic variant alters the expression of a reporter gene driven by a core promoter is only the first step in un-

derstanding the function of a variant. In order to obtain insights into the etiology of disease-causing variants, it is often necessary to link the presence of a regulatory variant to a molecular phenotype.

Furthermore, multiple lines of evidence support a model where polygenic effects on gene regulation are prevalent. These effects encompass the additive influence of multiple eQTLs with small effects acting on a given gene[106] and the influence of *trans*-acting variants, as described in Chapter 4. Because reporter assays typically assay contiguous DNA fragments up to a few kilobases in size, testing the combined regulatory effects of all genetic variants that affect a given gene is not always feasible.

Clearly, the optimal scenario in which to test the effects of regulatory variants involves measuring the *in vivo* expression of genes when specific combinations of regulatory variants occur in their natural, chromosomal context. Because regulatory variants are sometimes tissue-specific[116], it is ideal to perform such experiments in a cell- or tissue- type that is relevant to the organismal phenotype of interest.

The ability to systematically assay the molecular phenotypes of regulatory variants *in vivo* will depend on being able to edit genomic sequences efficiently and specifically. In the last few years, significant progress has been made towards that goal. Here, I briefly review the technological developments that have contributed towards making genome editing a viable approach for studying regulatory variation.

Methods for editing genomic sequences almost universally rely on the DNA replication and repair machinery that is naturally present in organisms. Several biological processes that take place during cell replication or DNA-damage repair can be used, under the right conditions, to replace genomic sequences with the sequence found in a template DNA molecule. Under many physiologic conditions, template sequences with sufficient homology to a target locus will be incorporated into genomic DNA at a small rate as part of the process of homologous recombination. In organisms with highly efficient homologous recombination, such as yeast, this approach can be used to replace genomic sequences after transformation or transfection of template DNA into dividing cells[117]. A comparable approach

was used to selectively edit genes in mouse embryonic stem cells[118], ultimately enabling the creation of the first knockout mice. However, the natural efficiency of homologous recombination in most organisms and cell types is too low to be practical.

The frequency of homologous recombination can be greatly increased by creating a double strand break (DSB) in DNA. In the aftermath of a DSB, genomic DNA will typically be repaired by one of two pathways: nonhomologous end-joining (NHEJ) or homology-directed repair (HDR)[119]. Repairs performed by the NHEJ pathway have a relatively high probability of causing nucleotide insertions and deletions (indels) at the site of the DSB, which is particularly useful for disrupting translational reading frames. Meanwhile, HDR relies on the presence of a donor template: a sequence with sufficient homology to the locus where the DSB occurred that it can be used as a template to repair the damaged sequence. By using an exogenous donor template harboring specific mutations, the HDR pathway can be used to introduce targeted changes into genomic DNA. As a consequence, the ability to create DSBs at specific genomic loci is a key step in being able to edit genomes precisely.

DSBs can be reliably created by the action of certain endonucleases. However, many endonucleases only possess limited sequence specificity. If used by themselves to cleave genomic DNA, DSBs would be created in many loci, which limits their ability to be used for site-specific genome editing. A subclass of endonucleases known as "meganucleases" are able to recognize fairly long DNA sequences (>14 bp) specifically[120] and induce targeted DSBs. However, engineering meganucleases with programmable DNA-binding specificity has remained challenging, limiting their widespread application[121].

A successful strategy for creating targeted DSBs is to engineer chimeric proteins composed of a sequence-specific DNA-binding domain fused to a sequence-agnostic nuclease, such as FokI[122]. The two most commonly used types of DBDs have been C_2H_2 zinc fingers (ZNFs) and TAL effectors (TALEs). ZNF nucleases were developed first and have been used for gene editing applications in several organisms, including human cell lines, fruit flies, rats, tobacco and maize[123]. However, as with meganucleases, engineering ZNF nucleases with custom binding preferences remains a laborious pro-

cess, limiting their widespread use[123]. Furthermore, targeting many A/T rich sequences with ZNF nucleases has proven difficult[120].

TALEs were first discovered as DNA-binding proteins expressed by *Xanthomonas* bacteria during infection of various plant species[124]. TALEs posses an array of nearly identical repeat domains, each spanning ~34 amino acids.[125]. The repeat domains in TALEs were demonstrated to be essential for sequence-specific DNA-binding[126] activity. It was also discovered that TALE proteins contain a constant N-terminal region that preferentially engages in contacts with a thymine base[127], validating observations that TALEs' optimal binding sites often contained a 5' T.

The first insights into the molecular basis of sequence-specific DNA recognition by TALEs were obtained by comparing the protein sequences of naturally occurring TALEs and the promoter sequences of genes that were differentially regulated during infection. Through separate computational approaches, a one-to-one correspondence between repeat sequences and the prefered nucleotides in the target site was identified[128,129]. In particular, residues 12 and 13 in the TALE repeat domains were identified as determinants of binding preferences to specific bases[128,129]. This pair of amino acids was named the "repeat variable disresidues" (RVDs)[129].

These observations led to the notion of the "TALE code," in which the identity of RVDs in consecutive repeat domains defined the protein's DNA-binding site. For example, the presence of an "HD" RVD indicated that particular repeat would preferentially make contacts with a cytosine, while the presence of "NI" would favor adenosine[128]. The discovery of the TALE code and the identification of RVDs with preferences for each of the four nucleotides greatly facilitated the development of programmable DNA-binding proteins. In principle, as long as the 5' base in the binding site was a thymine, any genomic sequence could be targeted.

The discovery of the TALE code was swiftly followed by the development of methods to assemble TALE proteins with custom DNA-binding specificities[130]. These TALE constructs have been fused to a wide variety of domains, including nuclease[131], trans-activation[132], DNA demethylase[133], and histone

demethylase[134] domains. The programmable nature of TALE DNA-binding domains enabled the effects of each of these domains to be targeted to specific genomic loci. As such, monomeric TALEs could be used as programmable TFs or to selectively edit histone modifications.

TALE nucleases (TALENs), in particular, have been widely used for genome editing. Some of the the earliest applications included the creation of specific gene knockouts in human cell lines[135] and zebrafish[131], which were achieved by using TALENs to create indels at target exons. Pairs of TALENs, each carrying a FokI domain and targeting genomic sequences in close proximity, were used successfully to introduce targeted modifications in human embryonic stem cells[136]. TALENs were found to have comparable cleavage efficiency to ZNF nucleases, which made them preferable due to the greater ease of designing TALEs to target specific sequences[131,136].

However, while engineered TALENs were generally successful at editing their intended target sequence, multiple studies reported significant rates of off-target activity[137–139]. Often, these off-targets could not be easily predicted based on genomic sequences. In other words, sites with several mismatches relative to the consensus binding site predicted by the TALE code were still being edited at relatively high rates. While methods were developed to assay the specificity of TALENs[140] experimentally, these approaches did not translate to better off-target prediction for TALENs without the need for laborious experiments, limiting their widespread applicability.

In Chapter 5, I describe the development of an improved model of TALE specificity, which we refer to as SIFTED (Specificity Inference For TAL Effector Design). The first step in creating SIFTED was the high-throughput measurement of TALE-DNA binding preferences using custom PBMs. To analyze these data, I developed a novel statistical model to infer free energy parameters from PBM data. Next, I developed a regression model that was trained on the PBM-derived measurements to predict the free energy parameters of proteins that had not been assayed.

Several tools for predicting TALE specificity had been released prior to the creation of SIFTED. In Chapter 5, I describe how the predictive accuracy of SIFTED was benchmarked against these other

tools in a variety of usage scenarios, ranging from TALEN-mediated genome editing to transcriptional activation by monomeric TALEs. The SIFTED model consistently performed better than other existing tools when compared across a range of applications. This improved predictive performance was achieved through a machine learning approach that was able to infer interdependencies between the RVDs in adjacent TALE repeats.

The development of SIFTED has provided insights that increase the precision for designing TALEs that are optimized to maximize on-target and minimize off-target effects. This improved understanding of TALE DNA-binding is likely to facilitate a wide range of applications, including genome editing, targeted epigenetic modification, and transcriptional activation or repression. All of these applications are likely to be of use in the process of characterizing regulatory variation in humans. For example, programmable transcriptional repressors can be used to further support hypotheses about the effects of regulatory variants that disrupt motif instances. If a TALE repressor binds the enhancer where an activator motif is disrupted by a SNP, one would expect to see a consistent effect on target gene expression.

However, no discussion about the applicability of SIFTED would be complete without mentioning a major development in programmable DNA targeting: the discovery of CRISPR-Cas9 and its subsequent use for many of the same applications as TALEs. In Chapter 6, I discuss the implications of this development and how it may affect the criteria for choosing a platform for genome editing and related applications.

In summary, I have presented a case for why studying regulatory variation is a key problem in human genetics. I have also described how each of the developments presented in this dissertation is likely to help. In the following chapters, each of these developments is described in full detail. In Chapter 6, I discuss the significance of the findings derived from each project, individually and in aggregate, as well as their limitations and potential future directions.

All of life is trade-offs.

Steve Gisselbrecht

2

Highly parallel assays of tissue-specific enhancers

Abstract

Transcriptional enhancers are a primary mechanism by which tissue-specific gene expression is achieved. Despite the importance of these regulatory elements in development, responses to environmental stresses, and disease, testing enhancer activity in animals remains tedious, with a minority of enhancers having been characterized. Here, we have developed 'enhancer-FACS-Seq' (eFS) technology for highly parallel identification of active, tissue-specific enhancers in *Drosophila* embryos. Analysis of enhancers identified by eFS to be active in mesodermal tissues revealed enriched DNA binding site motifs of known and putative, novel mesodermal transcription factors (TFs). Naïve Bayes classifiers using TF binding site motifs accurately predicted mesodermal enhancer activity. Application of eFS to other cell types and organisms should accelerate the cataloging of enhancers and understanding how transcriptional regulation is encoded within them.

2.1 Background

In metazoans, gene expression is regulated in a tissue-specific manner predominantly via noncoding genomic regions referred to as *cis* regulatory modules (CRMs) that typically regulate the expression of nearby gene(s)[141]. CRMs contain one or more DNA binding sites for one or more sequence-specific transcription factors (TFs) that activate or repress gene expression. CRMs that activate gene expression are frequently referred to as transcriptional enhancers[142].

The fruit fly *Drosophila melanogaster* has served as a powerful model organism for studies of transcriptional enhancers[142]. It has been estimated that there are ~50,000 enhancers in the *D. melanogaster* genome[143], yet to date only ~1,800 are known[144]. Technology for identifying active enhancers in particular cell types would aid in defining functional *cis* regulatory elements and would facilitate computational identification of key regulatory elements important for cell-type specific enhancer activity. Currently, TF-occupied regions identified by chromatin immunoprecipitation (ChIP) are tested

by low-throughput, traditional reporter assays[64,145]. Automated image analysis of reporter assays in *Drosophila* embryos[143,146] requires vast infrastructure and resources. Although other highly parallel technologies for testing the activity of *cis* regulatory elements have been developed recently[76,81,147–150], none of those approaches directly identify enhancer activity in a genomic context (*i.e.*, integrated into the genome) in particular cell types of interest in a whole animal.

We have developed a new technology, termed 'enhancer-FACS-Seq' (eFS), for highly parallel identification of active, tissue-specific transcriptional enhancers in the context of whole *Drosophila* embryos (Figure 2.1a). As with traditional enhancer assays, each candidate CRM (cCRM) is cloned upstream of a reporter gene. Our key innovation is that we replace the use of microscopy to screen for tissue-specific enhancers with fluorescence activated cell sorting (FACS) of dissociated cells. This approach utilizes a two-marker system: in each fly, one marker (here, the rat CD2 cell surface protein[151]) is used to label cells of a specific tissue so that they can be sorted by FACS, and the other marker (here, green fluorescent protein (GFP)) is used as a reporter of CRM activity. Cells are sorted by tissue type and then by GFP fluorescence. Thus, we are able to screen hundreds of cCRMs in a time- and cost-efficient manner.

2.2 Results

2.2.1 Library of candidate *cis* regulatory modules (cCRMs)

We focused on embryonic mesoderm as our model system because: (a) it comprises a variety of cell types, (b) the major regulatory factors governing mesoderm development are conserved between vertebrates and *Drosophila*[152], and (c) numerous data sets are available for genomic features associated with active enhancers. We created a plasmid library of hundreds of reporter constructs for ~1 kb cCRMs (Methods) located next to mesodermally expressed genes and comprising: ChIP-CRMs[64] bound by at least one of the somatic mesoderm TFs Twist (Twi), Tinman (Tin), or Myocyte enhancing factor 2

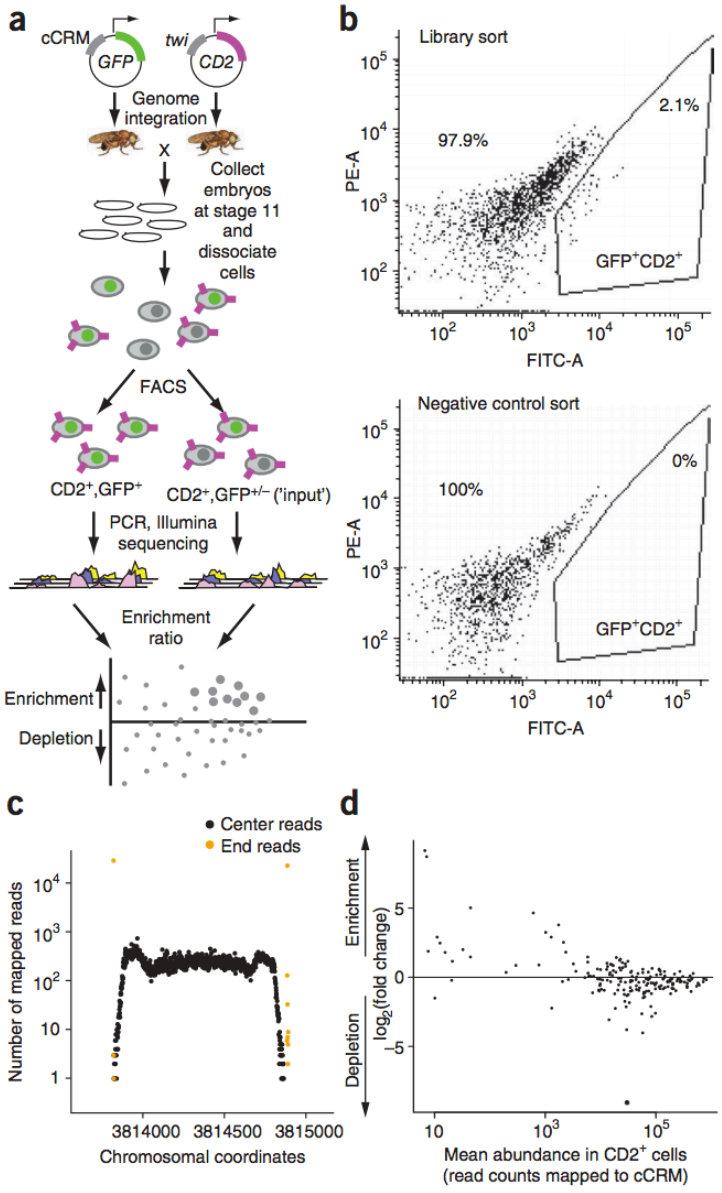


Figure 2.1: Enhancer-FACS-Seq overview. **(a)** Overall design of enhancer-FACS-Seq (eFS). **(b)** FACS purification of GFP+CD2+ cells prepared from embryos resulting from a cross of *Mef2-I-E*$_{D5}$:CD2 females to *(upper panel)* cCRM library transgenic males, and *(lower panel)* wild type (GFP-negative) males. In each panel, the plot shows yellow (``PE-A'') versus green (``FITC-A'') fluorescence for cells that pass the CD2+ gate out of 10^6 cells prepared from embryos. **(c)** Representative example of a cCRM, surrounding by native genomic flanking sequence, detected by eFS. **(d)** Enrichment ratios for cCRMs in *twi*:CD2- cells, as compared to *twi*:CD2+ cells. *Large points:* significantly enriched (P_{adj} < 0.1), *small points:* P_{adj} > 0.1.

(Mef2); regions bound by the transcriptional coactivator CREB binding protein (CBP)[153,154]; regions containing DNase I hypersensitive sites (DHS)[155]; dense clusters of evolutionarily conserved binding site motif occurrences for mesodermal TFs[56]; and additional regions surrounding known mesodermal genes (Methods).

2.2.2 Enhancer-FACS-Seq (eFS) experiments

Our cCRM plasmid library was injected into two different batches of embryos. In the first batch, we injected ~3,500 embryos, and crossed transformant males to females from two different CD2 lines to identify enhancers active in distinct tissues: *twi*:CD2 for whole mesoderm, and *Mef2-I-E*$_{D5}$:CD2 for a subset[156] of largely fusion-competent myoblasts (FCMs). In the second batch, ~4,500 embryos were injected, from which transformant males were crossed to *duf* :CD2 females to identify activity in somatic mesoderm founder cells (FCs). Each resulting embryo has one GFP reporter under the control of one cCRM integrated at the same genomic site by the ϕC31 integrase[157]. Use of a site-specific integrase avoids artifacts that would result if more than one cCRM were present in a cell and also avoids potential positional effects on enhancer activity.

At developmental stages 11-12, embryos were dissociated and purified by FACS. From the *twi*:CD2 embryos, we collected ~315,000 GFP+CD2+ cells, ~198,000 GFP+CD2- cells and 1×10^6 mock-sorted cells (*i.e.*, 'input') (see Methods; Figure 2.1b). We collected fewer GFP+CD2+ cells from the *Mef2-I-E*$_{D5}$:CD2 and *duf* :CD2 embryos since the *Mef2-I-E*$_{D5}$ enhancer is active in approximately 50-fold fewer cells than is the *twi* enhancer, which is active in one-quarter to one-third of all cells at this developmental stage, while the *duf* enhancer is active in the vast majority of the 660 FCs per embryo, nearly an order of magnitude fewer cells than for the *Mef2-I-E*$_{D5}$ enhancer.

To analyze cCRM integration into embryos, we extracted genomic DNA from the collected cells, amplified the cCRMs by PCR, and sequenced the resulting amplicons on the Illumina platform (Methods). We mapped the sequencing reads (Figure 2.1c) to the *D. melanogaster* genome using the segemehl

analysis package[158]. 213 and 400 cCRMs were detected (false discovery rate (FDR) < $5x10^{-5}$; see Methods) as having integrated into the fly genome from the first and second batches of injections, respectively. The greater number of cCRMs detected from the second batch was likely due to collection of transformant progeny from a larger number of injected embryos.

To evaluate the enhancer activity of the detected cCRMs, we calculated each cCRM's enrichment in a particular cell population as compared to the corresponding 'input' sample (Figure 2.1a) using the DESeq R package[159]. The input sample provides information on the baseline read counts due to cCRM representation within the embryo populations. In control experiments CD2+ and CD2- cells exhibited no significant differences in their cCRM content (Figure 2.1d). Therefore, CD2+ cells were used as the input sample for *twi*:CD2+GFP+, while for the rarer FCM and FC cell types CD2- cells were used as the input sample.

In total, 150 of the detected cCRMs were identified by eFS as being active enhancers (adjusted P-value (P_{adj}) < 0.1) in at least one cell population. Of these, we identified 57 as being active mesodermal enhancers: 34 in whole mesoderm (Figure 2.2a), 18 in FCMs (7 of which were also identified in whole mesoderm), and 20 in FCs (3 and 8 of which were also identified in FCMs or whole mesoderm, respectively). 12 of these 57 active mesodermal cCRMs overlap by at least 100 bp with a known mesodermal enhancer at an overlapping developmental timepoint in the REDfly database of curated CRMs[160], while the remaining 45 represent putative novel mesodermal enhancers, including 16 in FCMs and 14 in FCs. Analysis of GFP+CD2- cells collected from *twi*:CD2, *Mef2-I-E$_{D5}$*:CD2, and *duf*:CD2 embryos revealed 93 putative non-mesodermal enhancers (Figure 2.2b). A recent study screened a genomic DNA library for enhancer activity in the S2 cell line and in cultured ovarian somatic cells[76]; only 13 of the 57 total nonredundant enhancers active in mesoderm by eFS overlap by at least 100 bp with enhancers found in that study, although their screen does not provide information about enhancer activity in the same mesodermal cells. This comparison highlights the value of eFS for identifying enhancers active in particular cell types of interest within whole embryos.

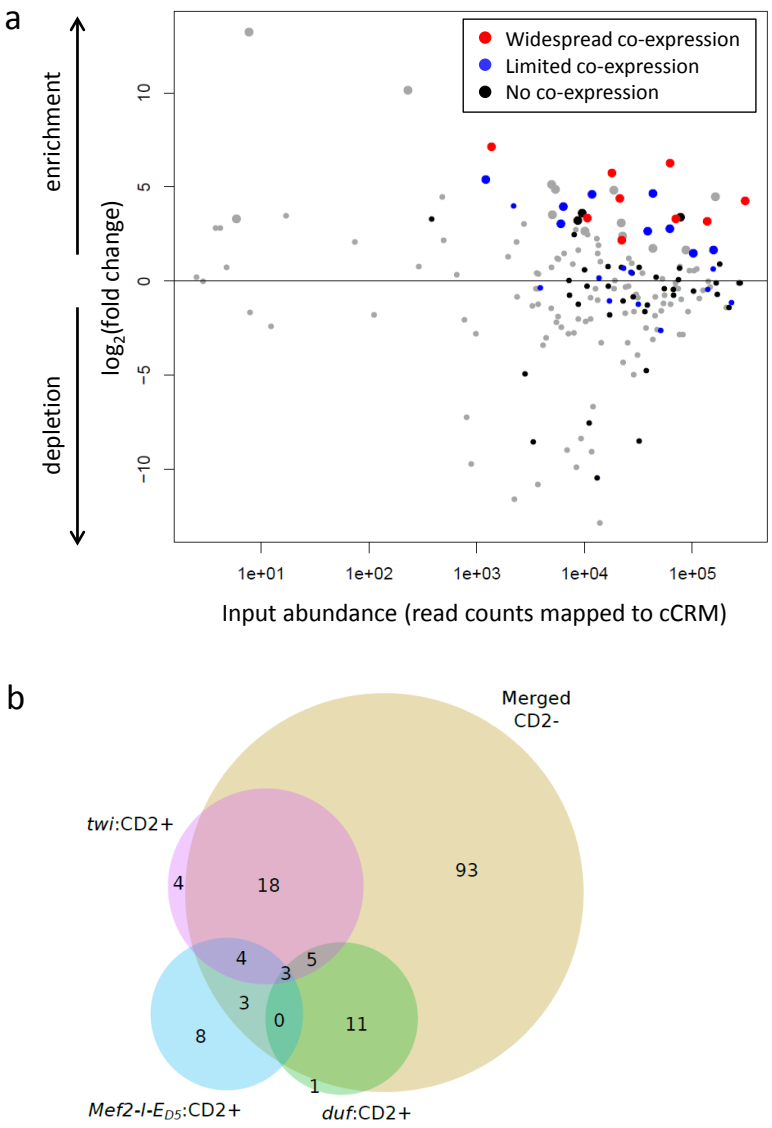


Figure 2.2: Overview of CRMs detected by eFS. (a) Enrichment ratios for cCRMs in *twi*:CD2+GFP+ cells, as compared to *twi*:CD2 input cells. *Large points:* significantly enriched (P_{adj} < 0.1), *small points:* P_{adj} > 0.1. Results from traditional reporter assays revealed cCRMs whose GFP expression shows widespread *(red)*, limited *(blue)*, or no *(black)* co-expression with *twi*:CD2 expression. **(b)** Venn diagram of active enhancers (P_{adj} < 0.1) identified from different cell populations: *twi*:CD2+; *Mef2-I-E*$_{D5}$:CD2+; *duf*:CD2+; nonredundant union of *twi*:CD2-, *Mef2-I-E*$_{D5}$:CD2-, and *duf*:CD2- (``Merged CD2-'').

2.2.3 Validation of eFS results

To validate our eFS results, we performed traditional reporter assays in whole *Drosophila* embryos (see Methods). For the *twi*:CD2+ eFS data, we tested 69 of the cCRMs analyzed by eFS, including: 21 putative active mesodermal enhancers ($P_{adj} < 0.1$) and 48 putative inactive cCRMs ($P_{adj} > 0.1$). The specificity of eFS was excellent among significantly enriched cCRMs: 18 of the 21 tested putative mesodermal enhancers drove expression in mesoderm at stage 11-12 (Figure 2.3). eFS exhibited moderate sensitivity for significantly enriched enhancers that were active in relatively few mesodermal cells: 9 gave expression patterns that were manually assessed as 'widespread co-expression' (*i.e.*, expression in a majority of strongly *twi*:CD2+ cells), while the other 9 drove 'limited co-expression' in smaller subsets of *twi*:CD2+ cells. 12 of the 48 putative inactive cCRMs drove 'limited co-expression.' Some of these eFS false negatives drove expression in cells that express low levels of CD2 and might have been missed by our use of a relatively stringent FACS gate for collecting *twi*:CD2+ cells. Although the data are slightly noisier for FCM and FC enhancers (6 out of 9 tested putative FCM enhancers, and 9 out of 11 tested putative FC enhancers, drove mesodermal expression) likely because roughly 20-fold fewer CD2+GFP+ cells were collected from the more specific *Mef2-I-E*$_{D5}$:CD2 and *duf*:CD2 lines, the results nevertheless demonstrate that eFS can successfully identify enhancers active in rarer cell types within whole embryos. We also evaluated the activity of 47 cCRMs identified by eFS as active in any of the three CD2-GFP+ cell collections, and found that the majority (35) were indeed active at this developmental stage.

2.2.4 Comparisons of eFS data to other genomic data types

We examined the eFS-identified enhancers for enrichment of previously described enhancer-associated chromatin marks, which were not used in the selection of cCRMs tested by eFS. Comparison to data from batch isolation of tissue-specific chromatin for immunoprecipitation (BiTS-ChIP) for mesoder-

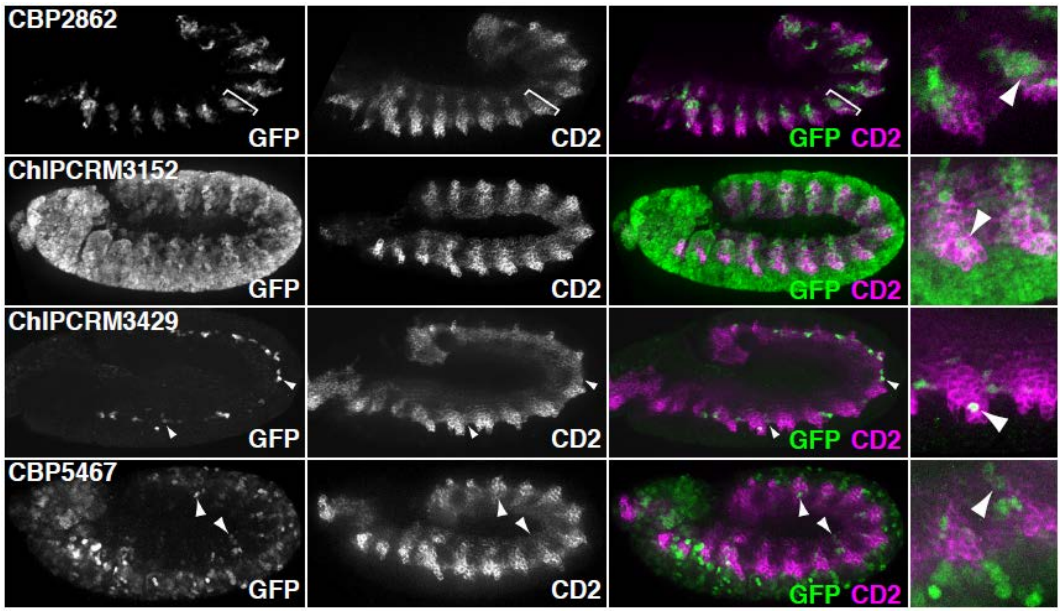


Figure 2.3: Validation of eFS predictions. (a) Sample validations of eFS predictions of enhancer activity. Constructs scored as driving ``widespread co-expression'' drove GFP specifically in a large fraction of the mesoderm (*e.g.*, somatic mesoderm, bracketed, in CBP2862) or in mesodermal plus non-mesodermal cells (ChIPCRM3152). ``Limited co-expression'' generally described expression in isolated mesodermal cells (arrowheads, in ChIPCRM3429 and CBP5467) or in a specific mesodermal structure. Co-expression is observed as green and purple in the same cells, since the GFP in these embryos is nuclear, while CD2 is expressed on the cell surface. Assessment of co-expression was performed with the annotator being blind to the predicted activity of the cCRMs.

mal cells from stage 10-11 embryos[161] showed that acetylation of histone H3 on lysine 27 (H3K27ac), monomethylation of histone H3 on lysine 4 (H3K4me1), H3K4 trimethylation (H3K4me3), H3K79 trimethylation (H3K79me3), and RNA Pol II, all of which previously have been found to be associated with active enhancers[70,80,161,162], are enriched (area under the receiver operating characteristic curve (AUC) $\geq$ 0.6, $P < 0.05$ by Wilcoxon-Mann-Whitney U-test) among the enhancers found to be active in mesoderm by eFS (Figure 2.4). However, in contrast to a prior report that H3K27 trimethylation (H3K27me3) was depleted among active mesodermal enhancers[161], we found H3K27me3 to be enriched among mesodermal enhancers. We also observed enrichment of H3K27Ac, H3K4me1 and H3K9Ac in comparisons of modENCODE data for whole embryos at 4-8 hr[153] to active enhancers identified by eFS in *duf*:CD2- cells, which approximate whole embryo samples (Figure 2.4). While H3K9Ac has been described as a mark of active transcription start sites[163], our observed enrichment of H3K9Ac among active enhancers supports an earlier observation of H3K9Ac in the 'strong enhancer' chromatin state in human cells[71]. As expected, overall there is a greater enrichment of DHSs identified from stage 5, 9, or 10 embryos, as compared to DHSs from later stages of development, among our enhancers, which were generated from stage 11-12 embryos. CBP occupancy was least enriched among these various enhancer-associated genomic features; these results suggest that CBP does not play a primary role in enhancer activity at this stage of embryonic development, and are consistent with a recent report of p300-independent enhancers in human cells[164].

Our collections of active enhancers allowed us to investigate which genomic data types[64,153–155] provide the greatest utility in identifying likely enhancers. Occupancy by sequence-specific TFs (Twi, Tin, Mef2, Bagpipe (Bap), Biniou (Bin)) expressed specifically in the mesoderm was by far most enriched among active mesodermal enhancers (Figure 2.4). DHSs[155] were nearly as enriched as enhancer-associated histone modifications (Figure 2.4).

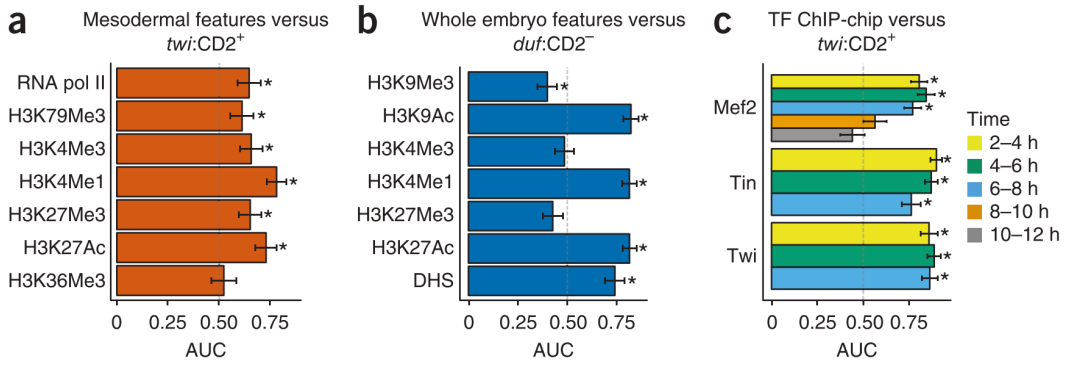


Figure 2.4: Enrichment of genomic features in eFS positives. Enrichment of **(a)** DHS, **(b)** histone modifications, and **(c)** TF ChIP-binding) associated with active enhancers in mesoderm (*twi*:CD2+) or in approximately whole embryos (*duf*:CD2-). AUC: area under receiver operator characteristic curve. * indicates $P < 0.05$ by Wilcoxon-Mann-Whitney U-test.

2.2.5 Analysis for over-represented transcription factor binding site motifs and combinations

Next, we separately analyzed each of the three sets of eFS-identified mesodermal enhancers (*i.e.*, in whole mesoderm, FCMs, or FCs) for over-represented TF binding site motifs and pair-wise motif combinations that might be required for enhancer activity. Briefly, we used the PhylCRM and Lever algorithms[56] to determine enrichment of matches, scored according to their evolutionary conservation, to 567 publicly available *Drosophila* TF binding site motifs[64,165–168] (see Methods). Numerous motifs were significantly enriched (AUC $\geq$ 0.65, FDR $\leq$ 0.1) either individually or in pair-wise combination (Figure 2.5a) for the whole-mesoderm and FCM enhancers.

For each of these two sets of eFS-positive cCRMs, we observed strong enrichment of the primary, known master regulator of that cell population: Twi for whole mesoderm[169], and Lmd for FCMs[156,170]. Motifs for other known mesodermal regulators were found in enriched combinations, including Bap, Lola-PC, and Mef2 in whole mesoderm, and Twi and Mef2 in FCMs. We also saw strong enrichment of motifs for several sequence-specific DNA-binding proteins – z, grh, and Trl (also known as GAGA Fac-

tor) – known to participate in recruitment of chromatin-modifying PcG and trxG proteins[171]; these results support prior findings of the enrichment of the z and/or Trl motifs among regions bound by Mef2, Twi, or Tin in ChIP-chip[172]. For the eFS-positive FC enhancers, no individual motifs or combinations thereof exceeded our statistical significance criteria.

cCRMs that appear to be active in FCMs show enrichment for a variety of conserved motifs (*e.g.*, Twi and Trl) in combination with a conserved Lmd motif, supporting the previously observed enrichment of these motifs in Lmd ChIP-Seq peaks[165]. We also observed numerous significantly enriched motif combinations (*e.g.*, many involving the uncharacterized zinc finger protein CG7928) not found in the Lmd ChIP-Seq study[165]. Since eFS data are not constrained by occupancy by a particular TF, they allow for a more unbiased identification of *cis* regulatory motifs. We also observed enrichment of numerous motif combinations comprising a master regulator and a factor with either ubiquitous or mesoderm-specific expression at the appropriate stage but no previously characterized role in mesoderm development (*e.g.*, schlank, Lola-PK) , suggesting novel regulators of mesodermal expression (square nodes in Figure 2.5a).

2.2.6 Machine learning classifier to predict mesodermal enhancer activity

We developed a machine learning approach to model whether cCRMs will be active or inactive in mesoderm or specifically in FCMs. We selected the mesodermal TF binding site motifs[64,168], using forward feature selection within 10-fold cross-validation, that were most discriminatory in distinguishing active versus inactive cCRMs (see Methods). We then used only the eFS data to train a Naïve Bayes classifier[173] (Figure 2.5b) based on the number and quality of matches to the discriminatory motifs. We independently trained models for whole mesoderm, FCMs, and FCs. Since the motifs contribute independently to these classifiers (*i.e.*, strict combinations of motifs are not required), this approach can potentially capture flexible, partially overlapping *cis* regulatory codes.

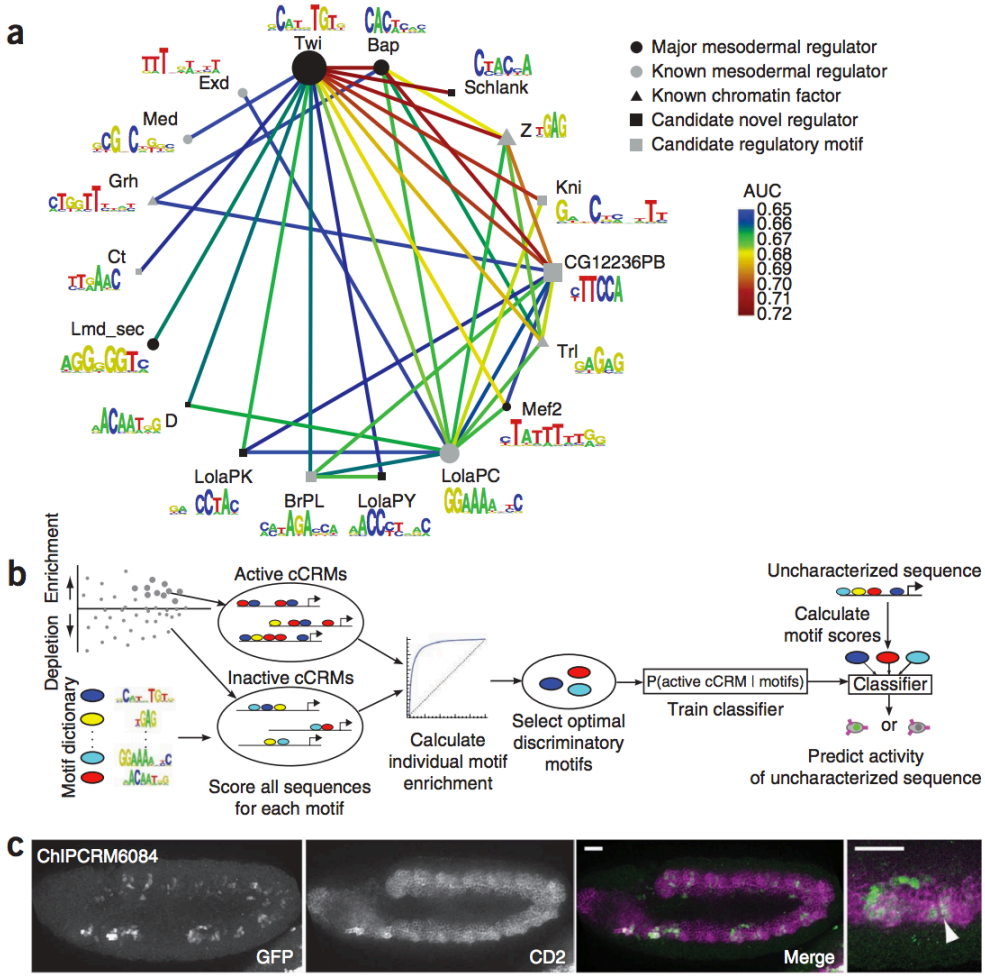


Figure 2.5: Motif enrichment and classification analysis of eFS CRMs. (a) TF binding site motifs or motif combinations significantly enriched (AUC $\geq$ 0.65, FDR $\leq$ 0.1) among eFS-identified active enhancers in *twi*:CD2+ cells. Nodes represent motifs for sequence-specific DNA-binding proteins that target chromatin-modifying PcG and trxG complexes to DNA *(triangles)*, major mesodermal regulators *(black circles)*, other factors known to have a role in mesodermal gene expression *(gray circles)*, putative novel regulators *(black squares)*; putative regulatory motifs for which the representative factors shown are not expressed in the embryonic mesoderm at the appropriate time *(gray squares)* and may be recognized by other trans-acting factors. Edges represent significant pair-wise AND combinations. Node diameter is proportional to (AUC-0.5)^2 considering the Lever AUC for the individual motif. **(b)** Schema of classifier analysis. **(c)** Maximum Intensity Projection of GFP expression driven by ChIPCRM6084, correctly predicted to drive co-expression with *twi*:CD2. Co-expression is observed and was assessed as described in Figure 2.3.

We assessed the accuracy of our models in predicting enhancer activity by 10-fold cross-validation. The whole mesoderm model achieved an AUC of 0.74 (P = 3.9 x 10^{-4}, Wilcoxon-Mann-Whitney U test) using 12 discriminatory motifs, while the FCM-specific model performed even better, with an AUC of 0.93 (P = 1.2 x 10^{-6}, Wilcoxon-Mann-Whitney U test) using 3 motifs. Importantly, these models outperformed ones based solely on previously known *cis* regulatory motifs for mesoderm and FCMs (AUC of 0.59 and 0.72, respectively; see Methods). A classifier was trained for FCs using the same procedure, but it did not achieve classification performance that was statistically significant (P-value > 0.05, Wilcoxon-Mann-Whitney U test).

To further demonstrate the practical utility of our computational models, we tested whether they could predict the activity of cCRMs whose activity had not been measured by eFS. We tested 39 classifier predictions by traditional reporter assays: 10 were predicted to be active, and 29 were predicted to be inactive, in mesoderm. Six out of 10 cCRMs predicted to be active enhancers in mesoderm drove co-expression of GFP with CD2 (Figure 2.5c); 19 out of 29 cCRMs predicted to be inactive drove no expression in CD2+ cells, while 9 of the 10 remaining predicted negative cCRMs drove limited co-expression at stages 11-12. Thus, our models based on motif content quite accurately predict cCRM activity in mesodermal tissue. Indeed, consistent with many of the *twi*:CD2+ eFS-positive enhancers in the training set exhibiting 'widespread co-expression' with CD2 and fewer exhibiting 'limited co-expression', our classifier appears to perform better in predicting the activity of cCRMs with 'widespread co-expression'.

2.3 Discussion

Our results demonstrate the utility of eFS technology for highly parallel testing of cCRMs for tissue-specific enhancer activity. Considering the various genomic features enriched among active mesodermal enhancers, no single data type (sequence-specific TF binding, histone modifications, or DHS) was most enriched across all three tissues. Moreover, none of the different classes of genomic features that

we used to prioritize cCRMs for testing by eFS (*i.e.*, ChIP-CRMs, CBP-bound regions, DHS) was significantly enriched ($p < 0.1$) among active versus inactive cCRMs considering either each of the three mesodermal CD2+ cell populations separately or their nonredundant union. It is perhaps not surprising that these regions were not enriched in either the *Mef2-I-E*$_{D5}$:CD2+GFP+ or *duf* :CD2+GFP+ data, since both FCMs and FCs correspond to relatively rare cell types and also since many of the putative regulatory regions might drive expression in other cell types as the adjacent genes are often expressed in more than just FCMs or FCs and not necessarily at this developmental stage.

Computational motif analysis of the cCRMs enriched for activity in whole mesoderm or in FCMs led to the discovery of novel *cis* regulatory motifs for these tissues and allowed us to train classifiers that accurately predict the activity of cCRMs in these tissues. Future studies will be needed to determine the regulatory functions of the putative mesodermal TFs suggested by the motif analysis results. Our observed enrichment of binding sites for PcG and trxG recruitment factors, and combinations thereof with ubiquitously expressed and mesoderm-specific TFs, in active enhancers suggests a model in which regulatory competence of a noncoding region requires the confluence of binding sites for chromatin factors with those for tissue-specific TFs.

The results of our classifier analysis demonstrate the utility of eFS data in learning *cis* regulatory sequences and indicate that *cis* regulation in FCMs is specified by a smaller set of TFs than those used in regulation of a broader class of mesodermal genes expressed in a wider range of cell types, each of which might utilize different *cis* regulatory codes, consistent with prior studies suggesting plasticity in mesodermal *cis* regulatory codes[64,174]. Likewise, the lack of a statistically significant classifier for FCs is likely due to the heterogeneity of the FC population and their associated enhancers[167,174]; identification of enhancers by eFS using CD2 driver lines specific to subsets or even unique FCs should aid in the elucidation of FC-specific *cis* regulatory codes. In addition, our results on enrichment of various histone modifications within the sets of active mesodermal enhancers are consistent with the model that there exist different classes of active enhancers that show enrichment for different sets of histone

modifications[161].

Here, we applied the eFS technology to the discovery of muscle enhancers. However, eFS can be used to test cCRMs in any other cell type that has at least one known enhancer, by constructing CD2 driver lines using known enhancers active in those cell types. Importantly, eFS can be used to screen cCRMs without any prior functional evidence (*e.g.*, ChIP data). Moreover, eFS can be adapted for use in other organisms, including vertebrates; the ϕC31 integrase system has been employed successfully in other species, including zebrafish[175], human and mouse cells[176], and mice[177]. In addition, the eFS technology could be implemented using a different site-specific recombinase or other transformation method. Broader application of eFS should greatly expand the repertoire of well-defined CRMs and facilitate the development of a more comprehensive picture of the landscape and organization of CRMs across genomes.

2.4 Methods

2.4.1 Cloning and preparation of candidate CRMs (cCRMs) library

Selection of candidate CRMs (cCRMs)

We designed our cCRM library to comprise ChIP-CRMs[64] bound by a somatic mesoderm (SM) TF or by the transcriptional coactivator CBP[153,154], DNase I hypersensitive sites (DHSs)[155] near or not near a mesodermally expressed gene, regions tiling the noncoding DNA around 4 genes, and a number of control regions. Specifically, we selected 491 cCRMs, which fall into 5 categories: (1) 288 ChIP-CRMs[64] bound by a somatic mesoderm (SM) TF (Twi, Tin, or Mef2); (2) 38 regions bound by the transcriptional coactivator CBP[153,154] and located next to mesodermally expressed genes lacking adjacent ChIP-CRMs (note: 45 of the 288 selected ChIP-CRMs overlap CBP-bound regions); (3) 58 regions containing DHSs[155] located next to mesodermally expressed genes lacking adjacent ChIP-CRMs; (4) 41 cCRMs predicted by the PhylCRM algorithm on the basis of dense clusters of evolutionarily con-

served TF binding site motif occurrences[56] adjacent to genes expressed in either somatic or cardiac mesoderm; and (5) 64 genomic windows tiling the noncoding DNA surrounding 10 genes (*cib, mud, nau, jumu, rgr, slou, CG3303, CG11202, CG13794, CG15319*) that we selected on the basis of biological interest and/or compact noncoding DNA. For each tiled gene, the intergenic sequences upstream and downstream of the gene and any introns longer than 1 kb were divided into overlapping segments of ~1 kb (*e.g.*, nt 1–1000, 501–1500, 1001–2000, etc.). Introns 200–1,000 bp long were included with additional coding DNA to extend them to a size of ~1 kb. All cCRMs were chosen to be 900-1,100 bp long to avoid potential PCR bias in later steps after cell sorting.

473 of the 491 cCRMs (96.5%) were successfully cloned into the eFS vector, as determined by Illumina sequencing of the resulting pooled, cCRM plasmid library (see below). Of the 473 injected cCRMs (see below), we detected (false discovery rate < 5×10^{-5}; see Section 2.4.7 below) 189 cCRMs in the *twi*:CD2 'input' samples, 213 in the *Mef2-I-E*$_{D5}$:CD2 input samples, and 411 in the *duf* :CD2 input samples; in total, 431 cCRMs were detected in at least one input sample. The remaining 42 cCRMs may have been missed due to underrepresentation in the injected cCRM plasmid library, stochastic lack of integration at the phiC31 genomic integration site, or stochastic underrepresentation among germline stem cells that led to the collected embryos.

PCR amplification of cCRMs

A two-step PCR amplification was used to include Gateway attB sites, and specific forward and reverse sequencing primers. We performed these PCRs in five 96-well plates, with a sixth plate to reattempt initially poor or failed PCRs; >95% (474/494) of the selected cCRMs were successfully amplified by the two-step procedure. First round PCR was performed in 25 μL reactions with Phusion enzyme (2X Master Mix with HF buffer) (New England Biolabs) using 50 ng *Drosophila melanogaster* OreR genomic DNA as template and 2.5 pmol of each PCR primer. Primer pairs were designed using the MacVector (v. 11.1) (MacVector, Inc.) primer design function, starting with default parame-

ters and relaxing them stepwise until a suitable primer pair was found. Common PCR primer sequences were engineered onto the 5' ends of each forward (SEQ1: CAAGACGAGGCTATGCTCTAGC) and reverse (SEQ2: TAGAGTTGGCTTGCCATAGACC) PCR primer. Cycling conditions were as follows: 1 x 30" @ 98°C; 30 x [5" @ 98°C, 10" @ 60°C, 30" @ 72°C]; 1 x 5' @ 72°C; hold @ 4°C. Second round PCR was performed under identical conditions, using 1 μL of a 1:500 dilution of first round PCR (in water) as template and a common primer pair (attB1-SEQ1: GGGGACAAGTTTGTACAAAAAAGCAGGCTCAAGACGAGGCTATGCTCTAGC and attB2-SEQ2: GGGGACCACTTTGTACAAGAAAGCTGGGTAGCTAGAGTTGGCTTGCCATAGACC).

Design of reporter vector for eFS.

We created the vector for enhancer-FACS-Seq, pEFS-Dest, by blunt-end cloning the 1.8 kb HindIII-SpeI fragment of pPelican[178] (containing a nuclear-localized GFP reporter construct with a *gypsy* insulator element upstream of the MCS and minimal promoter) into pWattB, then replacing the Multiple Cloning Site with a cassette providing attR1 and attR2 sites for Gateway cloning.

A Gateway LR cloning reaction between pEFS-Dest and a donor plasmid containing a cCRM flanked by attL1 and attL2 sites results in the insertion of the cCRM proximal to the minimal promoter driving expression of the reporter gene (see below). pWattB was made by inserting (1) the ϕC31 attB site from *Streptomyces lividans*[157] and (2) the mini-white gene into the small cloning vector pSP73 (Promega). The ϕC31 attB site supports highly efficient, site-specific integration of plasmids derived from this vector into the genome of flies that carry an attP transgene when ϕC31 integrase is expressed; the site-specificity of the target means that a library of reporter plasmids can be injected and only a single plasmid per haploid genome will integrate. Additionally, site-specific integration permits the selection of an ideal insertion site conferring negligible position effects on reporter gene expression[179]. The mini-white gene permits selection of transformant adult flies by eye color. The reporter cassette com-

prises the Hsp70 minimal promoter driving expression of a nuclear localization signal-tagged EGFP gene with an SV40 polyadenylation sequence[178]. We have tested pEFS by inserting the previously characterized mesodermal enhancers for *even skipped*[45], *Ndg* [167], and the *Lbl*-expressing FC[167]. In each case, GFP expression recapitulated the characterized activity of the enhancer.

Purification, normalization and cloning of cCRM library into eFS reporter vector.

Aliquots of all PCR reactions were run on ethidium bromide agarose gels along with High DNA Mass Ladder (Invitrogen) and photographed and roughly quantified using Quantity One software (BioRad). Equal masses of each 900-1,100 bp band were calculated from this quantification, pooled, precipitated, and gel-purified, then cloned as a pool using Gateway BP Clonase II (Invitrogen) into pDONR221 (Invitrogen) according to the manufacturer's protocols. Cloning reactions were transformed into commercial competent *E. coli* Top10 cells (Invitrogen) and plated on LB+kanamycin agar, yielding ~30,000 colonies. These colonies were scraped up and a plasmid pool purified from them, from which the combined inserts were cloned using Gateway LR Clonase II (Invitrogen) into pEFS-Dest. Transformed cells were plated on LB+ampicillin agar, yielding ~30,000 colonies, from which the final library plasmid pool was prepared for embryo injection.

In a preliminary test of PCR insert retrieval and Illumina library preparation (described in detail in Section 2.4.4 below), libraries prepared from dilutions of this plasmid pool (here, diluted to a final estimated concentration of 50,000 plasmid molecules in the plasmid sample prior to PCR) were sequenced. According to the sequencing data and our read mapping algorithm and criteria described below, 473 of the 474 successfully amplified cCRMs were successfully cloned into the eFS vector and were detected in 4 out of 4 Illumina sequencing libraries.

2.4.2 Generation of CD2 reporter vectors

A minimal promoter was fused to rat CD2 and subsequently cloned into P-element transformation vectors by PCR-amplifying the TATA box from pUAST-NTAP with PCR primer pair ATGGCTAG-CTAGCGAGCGCCGGAGTATAA and GGTGTCAATTCCCAATTCCCTATTCAG, and CD2 from twi-CD2[151] with primer pair CTGAATAGGGAATTGGGAATTGACACCATGAGATGT-AAATTCCTAGGGAGTTTCTTT and TAGGCTAGCTTAATTAGGGGGTGGC. These PCR products served as templates for an assembly PCR reaction using the primer pair ATGGCTAGCTA-GCGAGCGCCGGAGTATAA and TAGGCTAGCTTAATTAGGGGGTGG. This PCR product was subcloned into pCR (Invitrogen), sequence-verified, digested with NheI and cloned into XbaI-digested pETWN[45], resulting in our CD2 vector pETWCD2.

The enhancer region for *Mef2-I-E_{D5}* was synthesized in vitro (Integrated DNA Technologies, Coralville, IA, USA) and subcloned into pETWCD2, while the enhancer region for *duf* was PCR-amplified from genomic DNA isolated from a BAC clone (Children's Hospital and Research Center at Oakland, BACPAC Resources, Oakland, CA, USA) using the following primer pair (AATTCCCTT-CCACATTGGTCCTCTC, GATCCAAGTGTGATATCGTGTTTGGCC), subcloned into pETW-CD2 and sequence-verified.

2.4.3 Fly embryo injections and husbandry

We begin with a description of our strategy; experimental details follow. The resulting pooled plasmid cCRM clone library was injected into embryos. We are using a strain of flies containing an attP site that gives very low basal expression but very high induced expression in the embryonic mesoderm when an inducible promoter construct is integrated there[179]. This strain additionally expresses a nuclear-localized ϕC31 integrase under the control of the *nanos* promoter, which causes mRNA to be produced during oogenesis and deposited in the egg before fertilization. Injection of a plasmid containing an attB

site into fertilized eggs of this line at the posterior pole, where the primordia of the germ line will form, results in the transmissible integration of the plasmid DNA into the genome. The recombination of an attP and an attB site, mediated by the integrase enzyme, produces an attL and an attR site (distinct from and not cross-reacting with those used in the Gateway system, which are used by a different bacteriophage) which are not themselves substrates for the integrase; thus, integration is non-reversible and one integration event destroys the attP site used, preventing any further events at that genomic locus.

In principle, then, each primordial germ cell in an injected embryo can integrate two molecules of plasmid, one into each copy of the haploid genome. When gametes arising from these cells contribute to the progeny of these flies, each will receive only a single integrated transgene. As each embryo produces ~20–30 primordial germ cells before mitotic expansion, the gametes produced by an injected fly will carry a population of different library insertions. Progeny of injected flies that receive an integrated transgene can be recognized by the inclusion of the mini-white gene in the reporter vector, which causes a partial reversion of the adult eye color toward wildtype. *w+* (transformant) males are collected and crossed to virgin females of a strain homozygous for a transgene which drives expression of rat CD2 in the cell type of interest, producing the desired population of embryos.

To investigate whether competition for stem cell niches in the gonads of injected flies could limit the number of independent insertion events transmitted by each parent[151], we examined the transformant progeny of individual injected flies. For this investigation, we injected an earlier library of 275 cCRMs into embryos and reared survivors to adulthood. Males (to avoid confounding effects from mated females) were crossed to *y w* virgin females individually in vials. Up to 24 *w+* male progeny from each of several vials were isolated and processed for single-fly, insertion site targeted PCR, and the resulting PCR products sequenced to identify which transformed progeny of an injected fly had common inserts and which contained unique inserts. At one extreme, 24 flies from a cross with 47% transformant progeny had identical inserts; at the other, 10 flies from a vial with 22% transformant

progeny had among them 7 different inserts. The average number of insertions per injected parent over all sequenced transformants was 3.5.

For our full-scale eFS experiments, after purification and normalization, our pooled plasmid cCRM clone library was diluted to 0.75 mg/mL in standard injection buffer (5 mM KCl, 0.1 mM PO_4, pH 7.8) and injected posteriorly into syncytial embryos carrying the *nos-*ϕC31\int.NLS transgene[180] on the X chromosome and the attP40 insertion[179] on the 2nd chromosome. Injections were performed either in-house (used for the *twi*:CD2 and *Mef2-I-*E_{D5}:CD2 sorts) or by Rainbow Transgenic Flies, Inc. (Camarillo, CA) (used for the *duf* :CD2 sorts). Surviving males were crossed to excess *y w* virgin females. Transformant male progeny were selected on the basis of eye color. In pilot experiments, roughly 10% of all progeny of injected flies were transformant; based on the uniformity of the resulting eye color (and variation expected due to differences in insertion site), we estimate that the vast majority (~99%) of insertions were into the intended target location.

We collected several thousand transformant males and, separately, several thousand virgin females from each tissue-specific CD2 line of interest (see main text). These flies were combined in population cages ~36 hours before the beginning of embryo collections. Additional control cages containing only wild type (*y w*) flies, or wild type males with CD2 virgin females, were generated to provide control cells for assessment of FACS parameters, as discussed in the next Section.

Population cages were collected from twice "pre-lays" to minimize the presence of older embryos due to retention of fertilized eggs by females, then two collections of 2 hours (for *twi*:CD2 sorting) or 2.5 hours (for *Mef2-I-*E_{D5}:CD2 and *duf* :CD2 sorting) were performed. These plates were aged 10–11 hours at 18°C, after which embryos were collected and dechorionated, and single cell suspensions were prepared for FACS.

2.4.4 Fluorescence activated cell sorting (FACS)

We previously described and used extensively the isolation of single cells for FACS from live *Drosophila*

embryos at stage 11, when the SM FC and FCM populations are being specified[181]. We have modified the existing protocol by incorporating a step in which dissociated cells are resuspended in *Drosophila* cell culture medium and incubated on ice (10 minutes with moderate shaking) with a commercially available Alexa647-conjugated anti(rat CD2) antibody (AbD-Serotec, cat. #MCA154A647). The antibody is diluted 1:400 in Schneider medium + 8% FBS and preadsorbed by incubation with cells prepared from non-CD2-expressing embryos, then after removal of cells by centrifugation, samples were filtered with Nytex mesh and supplemented with 2 μg/mL DAPI to permit the detection and removal of dead cells and yolk granules. After brief washing (i.e. two cycles of centrifugation and resuspension in Schneider + 8% FBS), the cells are analyzed and separated by FACS.

Cells were concentrated by centrifugation of collection tubes (15 min at ~400g) and aspiration of all but ~0.5 mL of culture medium, then resuspended in the remaining culture medium and transferred to 0.6 mL PCR tubes and centrifuged 5 minutes @ 5,000g. This two-step collection process proved best for collecting very small numbers of cells in extensive pilot experiments. After complete removal of culture medium, the cell pellet is vortexed in 5 μL extraction buffer (10 mM Tris pH 7.5, 1 mM EDTA, 25 mM NaCl, 200 μg/mL proteinase K) and incubated 30 minutes at 37°, then 10 minutes at 95° to inactivate proteinase K and stored at -20°.

2.4.5 cCRM insert amplifications from collected cells, and Illumina library preparation including indexing

We designed our cCRM library to be compatible with Illumina sequencing. Each insert, as initially synthesized, has a common upstream end comprising, in 5' to 3' order, a Gateway attB1 site, and a common forward sequencing primer. The downstream end of each synthesized library insert comprises, in 5' to 3' order, a Gateway attB2 site, and a common reverse sequencing primer. The attB1 and attB2 sites facilitate the cloning of the inserts *en masse* first into a Gateway DONR vector and then, after amplification of the library, into the pEFS-Dest vector. After a population of cells containing inserts

of interest is isolated, PCR using the common sequencing primers allows recovery of a very pure insert population for quantification by high-throughput sequencing. We designed these common primers using our UniPROBE database of TF DNA binding specificities[38]; we generated synthetic sequence that is not predicted to be bound by any known TF (considering hundreds of TFs from multiple organisms) to minimize chances of it modulating the activity of any inserted CRM, and then tested primers derived from it in pairs, first for their suitability as PCR primers and then in control reporter experiments with known enhancers, to ensure that they showed the expected pattern of embryonic reporter expression.

We recovered library inserts by nested PCR from sorted cell genomic DNA. Crude cell extracts were pooled according to sample where necessary to achieve sufficient numbers for accurate quantification of insert abundance (Figure 1c), then split five-fold before PCR amplification. This ensures that ample PCR product will be produced for Illumina library preparation.

Each first-round PCR was performed in a 25 μL reaction using KAPA Hi-Fi HotStart ReadyMix (Kapa Biosystems), supplemented with 6% ethylene glycol and $MgCl_2$ to offset the EDTA introduced with the crude cell extract (typically 0.5 μL of 25 mM $MgCl_2$ when ~6 μL of template is added to each reaction). Primers for the first-round (outer) PCR are specific to our pEFS reporter vector (see Section 2.4.1) and were used at 1 μM each: nestF = GAATTGAATTGTCGCTCCGTAGAC; nestR = CAAGTATTTCCCCTTCGAGCTTG. Cycling conditions (optimized for sensitivity of detection and amplification) were: 1 x 2' @ 98°C; 1 x [20" @ 98°C, 3' @ 63°C, 30" @ 72°C]; 1 x [20" @ 98°C, 2:30 @ 63°C, 30" @ 72°C]; 1 x [20" @ 98°C, 2' @ 63°C, 30" @ 72°C]; 1 x [20" @ 98°C, 1:30 @ 63°C, 30" @ 72°C]; 13 x [20" @ 98°C, 1:15 @ 65°C, 30" @ 72°C]; 1 x 1' @ 72°C; hold @ 4°C (17 cycles in all).

Second-round (inner) PCRs were performed in 50 μL reactions, again using KAPA Hi-Fi HotStart ReadyMix supplemented with 6% ethylene glycol. The template for each reaction was 2 μL of first-round PCR, and the reaction mixture was heated to 80°C before template was added in order to ensure a true hot start despite the addition of active enzyme with the template. Primers for the second-

round PCR are the SEQ1 and SEQ2 sequences introduced at the ends of each cCRM during library construction (see Section 2.4.1) and were used at 2 μM each. Cycling conditions (optimized for yield of full-length product and accurate representation of spiked-in controls) were: 1 x 2' @ 98°C; 21 x [20" @ 98°C, 1' @ 65°C, 30" @ 72°C]; 1 x [20" @ 98°C, 1' @ 65°C, 45" @ 72°C]; 1 x [20" @ 98°C, 1' @ 65°C, 1' @ 72°C]; 1 x [20" @ 98°C, 1' @ 65°C, 1:15 @ 72°C]; 1 x [20" @ 98°C, 1' @ 65°C, 1:30 @ 72°C]; 1 x [20" @ 98°C, 1' @ 65°C, 1:45 @ 72°C]; 1 x [20" @ 98°C, 1' @ 65°C, 2' @ 72°C]; 1 x [20" @ 98°C, 1' @ 65°C, 2:15 @ 72°C]; hold @ 4°C (28 cycles in all).

PCR products (900-1,100 bp) were agarose gel-purified and quantified by NanoDrop. 1 μg each of the 5 PCR product pools from a given cell pool were combined to make the starting material for Illumina library preparation.

Spike-in and replicate control experiments.

We employed 'spike-in' controls to assess bias and replicate lanes to assess noise in amplifying cCRMs from a complex library; the known, qPCR-validated dilution of each control enabled us to accurately assess the representation of each control in the resulting sequencing lanes. Specifically, we amplified 900-1,100 bp regions of the *E. coli* chromosome (using the same 2-step PCR protocol described in Section 2.4.1), and cloned them individually into pEFS-Dest. Purified control plasmids were spiked into pooled library plasmid at various known dilutions, and the dilutions of each control plasmid in this mixture were confirmed by qPCR with insert-specific primer pairs; we performed duplicate measurements of two pairs per insert. An aliquot of ~50,000 plasmids (1 μL of a 500 fg/μL dilution) was subjected to our library insert recovery PCR workflow, as described above, followed by Illumina library preparation and sequencing (see below). The following table shows the comparison of the frequency with which each control *E. coli* region was detected by PCR from a complex mixture followed by high-throughput sequencing, compared with the expected frequency assuming the qPCR results accurately reflect the composition of the library pool:

Table 2.1: Summary of data obtained from spike-in controls

spike-in control	nominal dilution factor	measured dilution factor (qPCR)	expected counts / million reads	observed counts / million reads
E_coli_control_3	1:100	1:80	12,500	10,256.85
E_coli_control_5	1:100	1:87	11,494	12,243.19
E_coli_control_7	1:1,000	1:838	1,193	343.17
E_coli_control_11	1:1,000	1:829	1,206	77.08
E_coli_control_12	1:10,000	1:2,752	363	18.14
E_coli_control_16	1:10,000	1:2,726	367	21.23
E_coli_control_21	1:100,000	1:88,551	11	0.11
E_coli_control_22	1:100,000	1:968	1,033	0.77

These experiments show that, for very common inserts, our workflow is fairly accurate in recovering the abundance of insert in a complex mixture, but that accuracy suffers at very high dilutions. Since only ~50,000 plasmids were sampled, a 1:100,000 dilution would be expected to present a difficulty for accurate representation. To test whether the observed variation from expected frequencies represented bias in PCR amplification of different inserts based on their sequence (which would be expected to affect alternate samples equally, so that a comparison of observed abundance in different samples would still be informative) or noise introduced during the PCR amplification step, we performed the entire workflow on four different aliquots of a library pool and compared the resulting sequence counts. The entire sample preparation and sequencing process is highly reproducible: the number of reads mapped to each cCRM or control region was consistent across sequencing libraries prepared from four different aliquots of the plasmid library pool (mean Pearson $R^2 = 0.81$, ranging from 0.78 to 0.83 over six pairwise comparisons).

Illumina library preparation, including indexing

Illumina sequencing libraries were prepared using very minor modifications of standard protocols and the Multiplexing Sample Preparation Oligonucleotide Kit (Illumina). Pooled, purified PCR product was sonicated by Covaris S2. Samples were then end-repaired with the End-IT DNA End-Repair Kit (EpiCentre Biotechnologies) and A-tailed with Klenow exo- (New England Biolabs). Standard adapters (Index PE Adapter Oligo Mix) were ligated using Quick T4 DNA Ligase (New England Biolabs). The products of adapter ligation were run on 2% agarose gels and size-selected.

Purified products were quantified and checked for concentration and size distribution by Agilent 2200 TapeStation. Enrichment PCRs were performed in 50 μL reactions using Phusion thermostable polymerase (New England Biolabs). Each reaction contained 25 ng of template DNA and 1 μL each of primer InPE2.0, InPE1.0, and a numbered index primer. Indices were chosen for the 36 total samples in order to fill 7 lanes of a flow cell, and so that each statistical comparison (*e.g.*, 3 *twi*:CD2+GFP+ samples versus 3 *twi*:CD2+ 'input' samples) would be between samples run in different lanes with the same index. Cycling conditions, chosen to balance adequate yield for sequencing with minimal variance introduced at the enrichment PCR step, were: 1 x 30" @ 98°C; 10 x [10" @ 98°C, 30" @ 65°C, 30" @ 72°C]; 1 x 5' @ 72°C; hold @ 4°C.

2.4.6 Illumina DNA sequencing

Purified enrichment PCR products were again assessed by Agilent 2200 TapeStation and submitted to the Partners Center for Personalized Genetic Medicine for concentration measurement by PicoGreen fluorescence and qPCR, followed by equimolor index pooling and sequencing (50 base single-end read) on the Illumina HiSeq 2000.

2.4.7 Mapping Illumina sequencing reads

In principle, it would be faster (in terms of total CPU run-time) and potentially less noisy to map high-throughput sequencing reads directly to a reduced 'genome' constructed from our library of cCRMs. However, we anticipate that tiling large swaths of noncoding sequence to assess their regulatory potential will be a major application of eFS going forward, and as this requires the testing of overlapping sequences, it would create huge numbers of ambiguous mappings to repetitive regions. We have therefore instead chosen to pursue a two-step strategy, in which sequencing reads are first mapped to the *D. melanogaster* genome, and then mapped positions are assigned to members of the cCRM library, as described in detail below.

2.4.8 Mapping of sequencing reads to cCRM library

We used the segemehl algorithm[158] (version 0.0.9.4) to map reads to the *dm3* version of the *D. melanogaster* genome, using the parameters shown below.

- A maximum of two mismatch or indels in the seed area (*-D*)
- A maximum of an E-value (introduced in BLAST[182]) of 5 in the seed alignment (*-E*)
- At most 100 occurrences of the seed in the genome (*-M*)
- 80% or higher matching of characters after extending the seed to the entire read (*-A*).

We refer to reads that contained SEQ1 or SEQ2 primers (see Section 2.4.1 for a description of the SEQ1 and SEQ2 primers) as 'border' (or 'end') reads, since they derive from the 5' or 3' ends of the cCRMs; reads between the end reads are referred to as 'center' reads.

Pipeline for mapping sequencing reads.

The raw reads are processed in five steps: (1) preprocessing; (2) mapping; (3) filtering; (4) stacking and (5) assignment to cCRMs.

The preprocessing step (1) includes the identification and removal of PCR primer sequences (*i.e.*, SEQ1 and SEQ2, described in Section 2.4.1) from the 5' ends of the reads. We look only for complete primer sequences starting at the first sequenced position, since the detection of the other cases would bring only minor improvement to the overall results but would take significantly longer in terms of analysis run time. Due to the origin of the two data sets with reads that had a primer and those that did not, we refer to them as 'border' (or 'end') and 'center' reads, respectively (see description above).

Segemehl can report multiple locations for one read (parameter *-H 0*, which is also the default setting). If a read maps to exactly one 'expected' position, *i.e.*, in the range of a cCRM for center reads or within a 5-nt window at a cCRM start or end for border reads, we accept this location as the source of the read. If a read maps to no reasonable position or fits ambiguously, it is discarded. These operations represent the filtering step (3). In the stacking step (4), similar reads, *i.e.*, reads starting at the same chromosomal position and with the same length, are collected.

In the last step (5), the pipeline assigns the reads to the cCRM sequences. One potential problem is that overlapping cCRM windows contribute indistinguishable reads to the same genomic regions. We found that this can be addressed in a relatively straightforward manner by using the unambiguous border reads as weights for dividing the reads that map to overlapping cCRM windows, since we found that in isolated cCRMs the number of border reads and the total number of reads are highly correlated (Pearson $R > 0.95$).

The final output of the read mapping pipeline includes, for each cCRM: its name and chromosomal location, the number of border reads that mapped to its start and end, the resulting weight, the number of center positions covered by the beginning of a mapped center read and the total count of

reads mapped to the cCRM. From these data, one can then readily compute whether a cCRM has been detected and can compare the results quantitatively to those from replicate experiments.

2.4.9 Statistical analysis of enhancer-FACS-Seq (eFS) data

In order to detect and quantify enrichment of cCRMs in GFP+ cell populations, we collected the number of reads mapped to each cCRM for each replicate population and control 'input' population, and then filtered out cCRMs not detected in any input sample replicate.

The comparison of the frequency with which high-throughput sequencing reads mapping to a given cCRM are observed in a selected population with the frequency observed in an input population is fundamentally similar to the problem of detecting differential expression in an RNA-Seq experiment, or enrichment in a ChIP-Seq experiment if genomic occupancy regions are pre-defined. Thus, we utilized the DESeq package[159], which is intended for use with these data types. Specifically, enrichment and statistical significance were calculated using DESeq with standard parameters and size factor estimation. We considered an adjusted p-value < 0.1 as evidence of statistically significant enrichment or depletion. Importantly, the DESeq software package permitted the comparison of unreplicated data using variance estimated from replicate controls, and it produced results that agree quite well with those obtained from validation by traditional reporter assays (see Section 2.4.12 below). The distribution of results is depicted by plotting the $\log_2$(fold enrichment or depletion) against the abundance of reads in the input population (Figure 2.2a).

For abundant populations (*twi*:CD2+, *twi*:CD2-, *Mef2-I-E*$_{D5}$:CD2- and *duf* :CD2-), three or more GFP+ replicates were compared to three or more input replicates (*e.g.*, GFP+CD2+ cell replicates were compared to CD2+ cells collected from the same preparation). CD2 expression driven by *Mef2-I-E*$_{D5}$:CD2 or *duf* :CD2 is in a much more restricted population of cells (*i.e.,* 0.8% of viable cells on average for *Mef2-I-E*$_{D5}$:CD2); since a much smaller number of GFP+CD2+ cells could be collected, we were concerned that the noise introduced by sampling such small numbers of cells would reduce the

sensitivity of our analysis. Therefore, we pooled all three or six days' collections of double-positive cells from these experiments into a single sample. To conservatively estimate variance and sampling noise, we collected input populations of twice the size of the double-positive cell population (50% of input cells are expected to carry inserts, as the male parents of the dissociated embryos were heterozygous for reporter transgenes, while in principle 100% of GFP+ cells should carry inserts) or fewer. As collection of relatively rare CD2+ cells from this experiment would have been costly in terms of sorting time and wasted opportunity to collect GFP+CD2+ cells, we used CD2- cells as input controls. In principle, the abundance of each cCRM in CD2+ and CD2- cell populations should be the same, as these are derived from the same embryos. Nevertheless, to test this assumption, we simultaneously assessed the false positive rate at which cCRMs would be called significantly enriched or depleted in identical populations, by performing an identical comparison of *twi*:CD2+ (triplicate) to *twi*:CD2- (triplicate) cell samples. The results, shown in Figure 1d, support that these populations are indeed essentially identical.

2.4.10 DNA sequence motif over-representation analysis

One of the goals of this study was to identify *cis* regulatory motifs and the corresponding transcription factors that act through transcriptional enhancers active in the *Drosophila* embryonic mesoderm. One approach for identifying putative transcriptional regulators is to look for enrichment or depletion of DNA sequence motifs associated with known TFs, within sets of transcriptional enhancers active in the same tissue. The DNA-binding specificities of hundreds of *D. melanogaster* TFs have been determined by a variety of experimental methods [64,165–168]. Therefore, we used the Lever algorithm[56] to identify significant over-representation of matches, scored according to their evolutionary conservation, to 567 publicly available *Drosophila* TF binding site motifs[64,165–168] as compared to matched, randomly selected noncoding sequence, as described in more detail below.

A dictionary of *Drosophila* TF binding site motifs was compiled from the Fly Factor Survey database[168]

as well as from the literature [64,165–167]. Uninformative positions flanking the motifs were trimmed from both sides until either one position with information content (IC) $\geq$ 1 or two consecutive positions with IC $\geq$ 0.5 were reached.

To remove redundancy within the motif dictionary, motifs were clustered following the approach of Kheradpour et al. [61], using centroid-linkage hierarchical clustering and a Pearson correlation coefficient cutoff of >0.77 for cluster merging. Then, for each of the resulting 139 clusters, a motif representative ('exemplar') was chosen as the motif that was closest to the cluster average and whose TF had evidence of mesodermal expression [181,183,184]. In a few cases where a well known mesodermal regulator was part of a cluster, its motif was chosen as the exemplar (*e.g.*, Bap).

To identify the putative regulatory motifs of the *twi*:CD2+ and *Mef2-I-E*$_{D5}$:CD2+ foreground (FG) sequence sets, we used the previously described algorithm Lever [56]. Briefly, this analysis calculates the over-representation of individual motifs or combinations of motifs, according to their density and evolutionary conservation as quantified by the PhylCRM scoring scheme [56], in each FG sequence set as compared to a matched random set of background (BG) sequences. Rejection sampling was used to build a BG set for each FG sequence set; the BG set was ~20 times the size of the FG set and matched its length, GC content, and repeat content (as defined by RepeatMasker) by successive rejection sampling steps. The settings used for Lever and PhylCRM were the same as previously described [185].

We first ran Lever to score each of the 139 exemplar motifs individually. To reduce the loss of statistical significance due to a large number of hypotheses tested, any motif that did not have occurrences in at least one-quarter of the FG sequences was considered unlikely to contribute significantly to the overall FG set and thus was removed from the 139-exemplar motif dictionary, resulting in an 86-exemplar motif dictionary. We then used Lever to inspect the over-representation of all single and pairwise combinations of the resulting reduced dictionary of 86 motif exemplars.

2.4.11 Classifier analysis

We sought to determine whether cCRMs could be classified as active or inactive in a given cell population based on the number and quality of motif matches in the cCRM sequence. For each cCRM, we generated a feature vector of scores that quantify the presence of motif matches for each PWM in the motif exemplar dictionary. The score for a particular PWM and a particular cCRM was defined as the sum of the log-odds ratios of PWM matches in the cCRM sequence that exceeded a permissive match threshold (log-odds ratio > 3.0). This scoring method allows weaker motif instances to be captured and avoids threshold effects. The results did not change noticeably when this threshold was varied as long as it remained permissive (log-odds ratio < 6.0).

For classification, we used the Gaussian Naïve Bayes implementation in the *scikit-learn* package[186] for Python. A Naïve Bayes (NB) classifier implements a simple probabilistic model where each feature (here, the score for a given motif) contributes independently to the posterior probability of belonging to a class. NB classifiers previously have been shown to outperform more complex models at predicting expression based on sequence in other systems[173]. Here, we compared the performance of the NB model to more complex classifiers (Random Forest and Support Vector Machine classifiers with Gaussian kernels), which can model interactions between features, and found that the more complex models did not significantly improve classification performance.

For each motif, the NB classifier separately fits a Gaussian distribution to the scores for that PWM in the active cCRM set and in the inactive cCRM set (see below). These distributions are then used to generate posterior class probabilities for unlabeled data. The level of background motif matches is modeled by a Gaussian distribution inferred from the inactive cCRMs, which compensates for the permissive motif match threshold that was used to generate the scores.

We assigned class labels to specific cCRMs as follows: as for the motif over-representation analysis, positive cCRMs are those with DESeq *P_adj* < 0.1; here, negative cCRMs are those from an equally

sized set chosen from the bottom of the ranked list. To evaluate classification accuracy, we split the labeled cCRM feature vectors into training and test sets using stratified 10-fold cross-validation. Forward feature selection was performed independently for each of the folds: in each, the k motifs with the highest individual AUC values in the training set were selected. The classifier was then trained using features corresponding only to those k motifs. We evaluated the classifier's performance by calculating the AUC statistic over all of the cross-validated predictions. We compared performance with different values of k, and found that the classifier for *Mef2-I-E*$_{D5}$:CD2+ cCRMs performed better with lower values of k relative to the *twi*:CD2+ classifier, which is consistent with expectations based on tissue heterogeneity.

We compared the performance of our feature selection method (*i.e.*, in which we identified discriminatory motifs) relative to manually picking motifs of known regulators (see below) for that tissue type. Classifier models trained with our feature selection method outperformed those that used the motifs of known regulators for both *twi*:CD2+ (AUC = 0.74 using feature selection method versus 0.59 using known regulatory motifs) and *Mef2-I-E*$_{D5}$:CD2+ (AUC: 0.93 using feature selection method versus 0.72 using known regulatory motifs) eFS data.

The known regulatory motifs used for *twi*:CD2+ eFS data (*i.e.*, whole mesoderm) are:

1258_twi_Zinzen

241_Tin_Cell_FBgn0004110

1256_mef2_Zinzen

1254_bap_Zinzen

1246_Lmd_Busser_Sec (called "lmd_sec" in Figure 4a

573_lmd_SOLEXA_5_FBgn0039039

The known regulatory motifs used for *Mef2-I-E*$_{D5}$:CD2+ eFS are:

1258_twi_Zinzen

1256_mef2_Zinzen

1254_bap_Zinzen

1246_Lmd_Busser_Sec (called "lmd_sec" in Figure 4a)

573_lmd_SOLEXA_5_FBgn0039039}

338_SuH_FlyReg_FBgn0004837

The discriminatory motifs learned in at least 50% of the 'folds' in the 10-fold cross-validation for *twi*:CD2+ eFS data (*i.e.*, whole mesoderm) are:

716_lola-PC_SANGER_5_FBgn0005630

659_CG8765_SANGER_5_FBgn0036900

658_CG8281_SANGER_5_FBgn0035824

652_CG3919_SANGER_5_FBgn0036423

623_kay_Jra_SANGER_5_FBgn0001297_SANGER_5_FBgn0001291

583_CG7928_SOLEXA_5_FBgn0039740

269_D_NAR_FBgn0000411

1258_twi_Zinzen

297_brk_FlyReg_FBgn0024250

700_lola-PO_SANGER_5_FBgn0005630

600_vfl_SOLEXA_5_FBgn0259789 4

The discriminatory motifs learned in at least 50% of the 'folds' in the 10-fold cross-validation for *Mef2-I-E*$_{D5}$:CD2+ eFS (*i.e.*, FCMs) are:

573_lmd_SOLEXA_5_FBgn0039039

702_lola-PW_SANGER_5_FBgn0005630

583_CG7928_SOLEXA_5_FBgn0039740

Discriminatory motifs learned in 10-fold cross-validation for *duf*:CD2+ eFS (*i.e.*, FCs) are not reported here, since the results of the classifier analysis were not statistically significant, as determined above.

2.4.12 Traditional reporter assays

104 cCRMs assayed by eFS in any cell population were individually validated by more traditional means. 37 of these had been included in the library on the basis of earlier pilot experiments and sequence-validated clones in pEFS-Dest were available for them, while the remaining 67 were recovered from single transformant male flies (and identified by PCR and subsequent sequencing of their library inserts) after library injections. These latter 67 lines could in principle contain PCR-induced mutations which might affect the pattern in which they drive GFP expression, since they derived from library clones which were not sequence-verified. For 25 of these lines, we were able to ascertain the complete sequence of the library insert by sequencing two independent PCR reactions using vector-specific primers and the complete sequence of the corresponding genomic regions in the wild type flies (OreR) from which our library clones were initially amplified. In 5 of the 25 cases, the library insert matched the published *D. melanogaster* sequence (dm3) at every position, while in the remaining 20 cases all differences from the published sequence were found to be present in our wild type population. We thus conclude that we can neglect PCR-induced mutations as a significant source of error in subsequent library generation using this PCR protocol.

Homozygous (or, where unavailable, balanced heterozygous) transformant males were crossed to homozygous *twi*:CD2 females in small population cages, and broad collections (~2-17 hours after egg deposition) of embryos were fixed and stained for immunofluorescence by standard protocols[45]. Antibodies used were mouse anti-rat CD2 Alexa647 conjugate (MCA154A647, AbD Serotec) at a final dilution of 1:200, and rabbit anti-GFP, goat anti-rabbit Alexa488 conjugate, and goat anti-mouse

Alexa647 conjugate (A-11122, A-11034, and A-21236, Molecular Probes), all at 1:500, and were preadsorbed to wild type embryos before use. Stained embryos were resuspended in Vectashield (Vector Labs) and, for those lines in which GFP was expressed at stage 11–12, imaged with a Zeiss Imager Z1 with Apotome in optical sectioning mode. Coexpression of GFP with CD2 was evaluated in individual optical sections with the annotator being blind to the predicted activity of the cCRMs. Coexpression is observed as GFP and CD2 being present in the same cells, since the GFP in these embryos is nuclear, while CD2 is expressed on the cell surface. Maximum Image Projections were constructed for display (Figure 3a, Figure 4c). Coexpression was additionally subjectively assessed as "widespread" (roughly, a majority of high-CD2-expressing cells express GFP), "limited," or "no coexpression".

This chapter is a modified version of a published article describing this work:

Gisselbrecht SS, Barrera LA, Porsch M, Aboukhalil A, Estep PW 3rd, Vedenko A, Palagi A, Kim Y, Zhu X, Busser BW, Gamble CE, Iagovitina A, Singhania A, Michelson AM, Bulyk ML. Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. *Nature Methods* (2013) 10(8):774-780.

Acknowledgments

This project was supported in part by a US National Science Foundation Graduate Research Fellowship to L.A.B. and by grant R01 HG005287 from the US National Institutes of Health to M.L.B. We thank G. Losyev and C. Durkin for technical assistance, K.G. Guruharsha and K. Vijay Raghavan for sharing coordinates of the duf enhancer before its publication, R.P. McCord, M. Markstein and O. Iartchouk for helpful discussion, and R. Gordân, M. Markstein and T. Siggers for critical reading of the manuscript.

Author contributions

M.L.B. designed the study; S.S.G., P.W.E., A.M.M. and M.L.B. developed the eFS technology; S.S.G. and A.V. sorted flies; S.S.G., L.A.B., M.P. and A.A. performed computational data analysis; S.S.G., P.W.E., A.V., Y.K. and X.Z. performed PCRs; B.W.B., X.Z., A.S. and C.E.G. generated CD2 fly lines; S.S.G., A.V., A.P. and A.I. performed validation assays; S.S.G., L.A.B., M.P., A.A. and M.L.B. prepared figures and tables; and S.S.G. and M.L.B. wrote the manuscript.

My contributions

- Developed the initial computational pipeline for analyzing eFS data. Being the product of a novel experimental method, these data required unique considerations compared to other types of sequencing-based experiments.
- Compared the enhancers identified by eFS to TF ChIP-Seq, histone modification and DNase-seq data to evaluate the predictive power of different data types for identifying sequences that drive tissue-specific expression.
- Developed a rejection sampling approach for creating a set of genomic background sequences that match the foreground set in length, nucleotide composition and repeat content.
- Performed motif enrichment analysis that identified putative novel mesodermal regulators as targets for further experimental validation. (Pairwise combination analysis by Anton Abhoukhalil)
- Developed a Naive Bayes classifier that could accurately predict which sequences would drive tissue-specific expression based on their motif composition. This classifier was validated by imaging the expression patterns from candidate enhancers predicted by the classifier but never included as part of the training set.

I find it astonishing that the immune system embodies a degree of complexity which suggests some more or less superficial though striking analogies with human language, and that this cognitive system has evolved and functions without assistance of the brain.

Niels K. Jern

3

The genomic landscape of NF-*κ*B binding

Abstract

The nuclear factor κB (NF-κB) subunits RelA, RelB, cRel, p50 and p52 are each critical for B-cell development and function. To systematically characterize their responses to canonical and non-canonical NF-κB pathway activity, we performed ChIP-Seq analysis in lymphoblastoid B-cells (LCLs). We found a surprisingly complex NF-κB binding landscape, which did not readily reflect the two NF-κB pathway paradigm. Instead, ten subunit binding patterns were observed at promoters and eleven at enhancers. Nearly one-third of NF-κB binding sites lacked κB motifs and were instead enriched for alternative motifs. The oncogenic forkhead box protein FOXM1 co-occupied nearly half of NF-κB binding sites, and was identified in protein complexes with NF-κB on DNA. FOXM1 knockdown decreased LCL NF-κB target gene expression and ultimately induced apoptosis, highlighting FOXM1 as a synthetic lethal target in B-cell malignancy. These studies provide a resource for understanding mechanisms that underlie NF-κB nuclear activity and highlight opportunities for selective NF-κB blockade.

3.1 Background

NF-κB is a family of dimeric transcription factors (TFs) that mediate differentiation, development, proliferation and survival[90] . NF-κB is a principal component of the body's defense against infection, and is critically important for most immune and inflammatory responses. Yet, NF-κB hyperactivation contributes to inflammatory disorders and cancer, in particular B-cell malignancies[187,188]. Despite progress in understanding cytosolic pathways that activate NF-κB TFs, comparatively little is known about the mechanisms that govern nuclear NF-κB function[189–191].

Microbes nonetheless use NF-κB to enable their replication and spread. Oncogenic viruses encode factors that constitutively activate NF-κB, including Epstein-Barr virus (EBV), Kaposi's sarcoma-associated herpesvirus, human T-cell leukemia virus, hepatitis B and hepatitis C[192]. Constitutive NF-κB activation also contributes to the pathogenesis of numerous human cancers, in particular B-cell

lymphomas[187,188]. However, the genome-wide effects of constitutive NF-κB activation on NF-κB transcription factor binding have not been defined.

Mammalian genomes encode five NF-κB subunits: p105/p50, p100/p52, RelA (p65), RelB and cRel. Each has an N-terminal Rel homology domain that mediates sequence-specific binding to κB sites[90] on DNA. RelA, RelB and cRel also have C-terminal transcription activation domains. NF-κB dimers can further induce or suppress target gene expression through cofactor recruitment. Inhibitor of NF-κB (IκB) proteins retain NF-κB dimers in the cytosol, with the exception of p50 homodimers, which are constitutively nuclear[90].

Two NF-κB pathways trigger NF-κB activity by inducing IκB degradation and NF-κB nuclear translocation[193]. The canonical pathway responds to pro-inflammatory signals and is essential for rapid immune responses. The canonical pathway triggers IκBα degradation, which enables RelA and cRel-containing complexes to translocate to the nucleus, including RelA:p50, cRel:p50, RelA:RelA and cRel:cRel dimers. The non-canonical NF-κB pathway promotes secondary lymphoid organogenesis, B-cell development and survival[194]. The non-canonical pathway triggers processing of p100 to p52, which enables the p52-containing complexes RelB:p52, p52:p52, and p50:p52 to enter the nucleus.

When both pathways are active in B-cells, up to 14 distinct NF-κB dimers form, including canonical/non-canonical hybrids such as RelA:p52[99]. Murine genetic studies indicate that each NF-κB subunit, and perhaps each dimer, has unique functions in B-cell development and activation[98]. The generation and maintenance of mature B-cells requires both canonical and non-canonical NF-κB pathway activity[195]. CD40-mediated activation of both pathways are required for B-cell responses such as homotypic aggregation, which requires both cRel and p52[196]. Yet, the extent of intrinsic NF-κB dimer binding preference for its target sites *in vivo*, and the mechanisms that establish dimer-specific binding, are not well understood. Likewise, little is known about the extent to which target genes are regulated independently, or jointly, by the canonical and non-canonical pathways.

κB sites in mammalian genomes vary widely from the consensus sequence 5'-GGGRNYYYCC-3'

(where R is a purine, Y is a pyrimidine, N is any nucleotide). Moreover, a single base pair difference in a κB site can induce distinct NF-κB dimer conformations and affect coactivator requirements[93]. The extent to which NF-κB family members differentially recruit TFs to κB sites remains to be examined *in vivo*. Likewise, NF-κB recruitment by other sequence-specific TFs to non-κB DNA sites has not been extensively investigated.

To date, genome-scale analyses of NF-κB binding by chromatin immunoprecipitation (ChIP)-based methods have generally been limited to RelA[86,197–200]. Where multiple subunits were studied, cells were stimulated by Toll-like receptor agonists that preferentially activate the canonical NF-κB pathway[201,202]. In B-cells, only RelA has been studied systematically[86].

To systematically investigate how NF-κB TFs recognize *in vivo* targets, we performed ChIP-Seq analysis of all five NF-κB subunits in the EBV-transformed lymphoblastoid B-cell line (LCL) GM12878, where the EBV-encoded membrane protein LMP1 mimics CD40 to constitutively activate the canonical and non-canonical pathways. GM12878 has a relatively normal karyotype, is one of three ENCODE project Tier 1 cell lines, and is an original HapMap cell line used in many genetic studies. We identified a complex NF-κB binding landscape, with distinct NF-κB subunit binding patterns at LCL promoters and enhancers, and with frequent recruitment of NF-κB to DNA sites that lack κB motifs. Numerous B-cell transcription factors co-occupied LCL NF-κB sites, including the Forkhead box protein FOXM1. FOXM1 was present at nearly half of all LCL NF-κB sites and was recruited to NF-κB complexes on DNA. Collectively, our results provide new insights into B-cell nuclear NF-κB regulation, including CD40-stimulated germinal center B-cells and lymphomas with constitutive NF-κB activity.

3.2 Results

3.2.1 Genome-wide NF-κB subunit DNA binding in lymphoblastoid B-cells

Chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq) was used to assess NF-κB subunit DNA binding in GM12878 cells. Validated anti-RelA, RelB, cRel, and p50 antibodies were used, each of which have been shown to be specific by western blot and ChIP [200,203–210], and by a ChIP-microarray analysis of NF-κB promoter occupancy in lipopolysaccharide-stimulated monocytes [202]. Anti-p52 antibody specificity was validated by immuno-precipitation.

We identified 20,067 RelA, 16,617 RelB, 6,765 cRel, 4,298 p50, and 10,814 p52 peaks at an irreproducible discovery rate (IDR) < 0.01 [211], significantly expanding the known number of B-cell NF-κB binding sites. Datasets for each NF-κB subunit exceeded ENCODE project quality control standards [211], implying that the sequencing depth was adequate to capture biologically meaningful binding (Methods). In performing comparisons of binding between subunits, we took multiple steps to mitigate differences that arose from sequence depth effects, including normalization of ChIP signals across experiments [212] (Methods). Where a NF-κB binding site was identified for any subunit, we cross-compared raw signals for all five subunits at that site, in order to minimize threshold effects caused by binary peak calls. Robust peaks for all NF-κB subunits were evident at κB sites at many well-characterized B-cell κB target genes, including the BCL2 locus (Figure 3.1A).

Using GM12878 chromatin state annotations based on histone modifications [71], we found NF-κB binding predominantly (73%) at predicted active enhancers, as characterized by H3K4me1 and H3K27ac marks (Figure 3.1B). For example, the dominant BCL2 NF-κB peaks localized to an enhancer (Figure 3.1A). Nonetheless, in comparison with other NF-κB subunits, a higher proportion of cRel peaks occurred at active promoters (~40% of all cRel peaks), as defined by H3K4me3 and H3K9ac marks. By contrast, only ~15% of mapped RelB peaks localized to active promoters. NF-κB binding site motifs derived *de novo* from the ChIP-Seq data were similar to each other, with the cRel motif

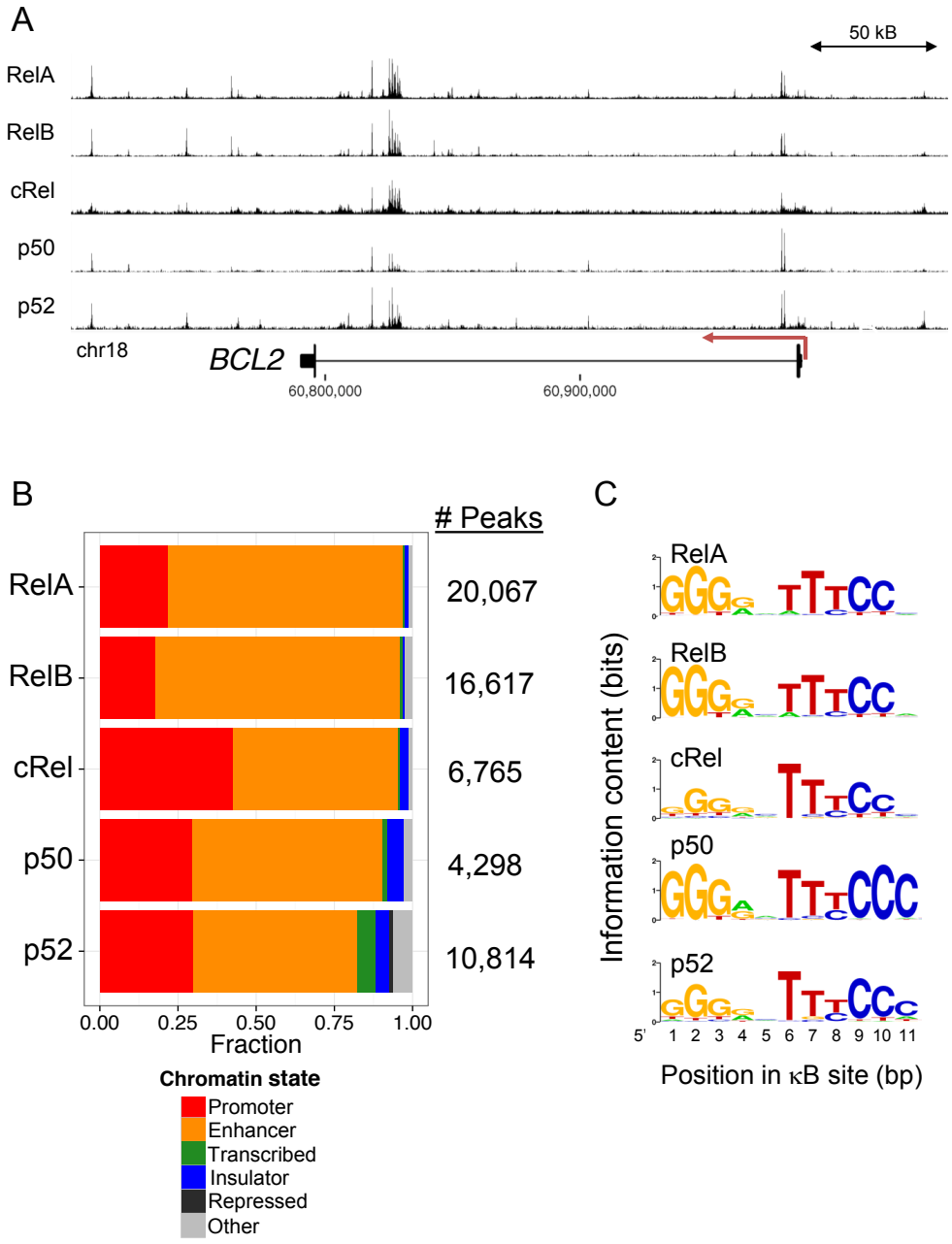


Figure 3.1: NF-κB subunit genome-wide distribution and consensus motif. (A) NF-κB subunit ChIP-Seq signals at the BCL2 locus. The y-axis is scaled between zero and twenty times the median signal value of the surrounding 100 kB. (B) Genome-wide NF-κB subunit distribution across chromatin states, as defined by GM12878 histone modifications, CTCF and Pol2 occupancy. Each horizontal bar shows the fraction of NF-κB subunit peaks that were assigned to each chromatin state. The total number of NF-κB subunit peaks is shown to the right of each bar. (C) Consensus *de novo* motif for each NF-κB subunit.

showing increased degeneracy in its 5' half-site and p50, and to a lesser extent p52, exhibiting an extended motif at the 3' end (Figure 3.1C).

3.2.2 Patterns of NF-κB subunit co-binding

A rich NF-κB dimer milieu was present in LCL nuclei (Figure 3.2). For instance, RelA and cRel bound to similar amounts of p50 and p52, and at a lower level, to each other. By contrast, RelB preferentially associated with p52, to a lower level with p50, and to a substantially lower level with RelA. Both p50 and p52 associate with all NF-κB subunits (Figure 3.2). κB sites in theory could be bound by a single NF-κB dimer, or could be accessed by distinct NF-κB dimer combinations in equilibrium with one another.

To identify NF-κB subunit binding patterns (SBPs), we applied *k*-means clustering to the ChIP-Seq data (Methods). We found 10 distinct SBPs at LCL promoters and 11 at enhancers (Figures 3.3A and 3.3B). SBPs with binding by multiple NF-κB subunits likely reflected NF-κB dimer exchange at these sites, rather than simultaneous binding by distinct NF-κB dimers to a single site. In support of the specificity of the antibodies used and despite the RelA dataset having the highest number of peaks, clusters with predominant binding by each of the NF-κB subunits were observed at promoters and enhancers, except for RelA. To minimize the possibility that SBPs arose from differences in peak number alone, we generated *k*-means clustered heat maps using only the top scoring 4000 peaks for each subunit. Even when using an equal number of peaks for each subunit, very similar SBPs were again observed, suggesting that SBPs do not arise from differences in antibody sensitivity alone.

Intriguingly, some SBPs were evident at both promoters and enhancers, while others were unique to either. For example, cluster P10 promoters, but no enhancer clusters, were occupied by all NF-κB subunits except cRel. Combinations of distinct SBPs were observed at several key NF-κB target genes, such as at the seven NF-κB ChIP-Seq peaks near the highly LMP1-induced target gene TRAF1 (Figure 3.3C).

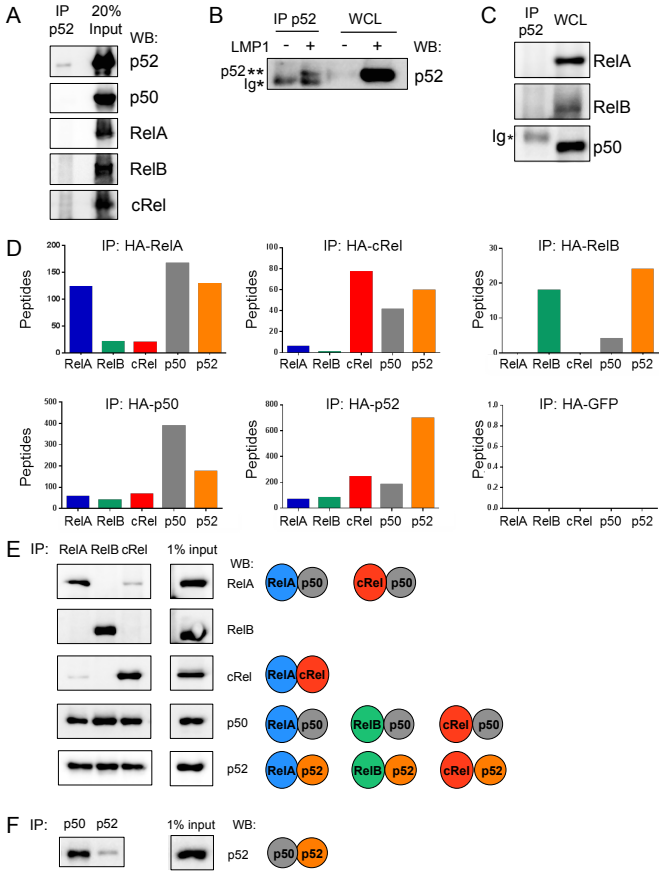


Figure 3.2: Anti-p52 antibody validation and characterization of NF-κB dimers in LCLs. Anti-p52 antibody validation, and analysis of LCL NF-κB dimers. (A) The anti-p52 antibody used for ChIP-Seq was tested at 4 ug/ml for ability to immuno-precipitate (IP) each NF-κB subunit, prepared by *in vitro* translation to avoid formation of p52- containing NF-κB heterodimers. Western blot (WB) analysis, performed for each NF-κB subunit, demonstrated that the antibody only pulled down p52. (B) The anti-p52 antibody immuno-precipitated p52 from LMP1+ HEK-293 cells. Extracts were prepared from 293 cells without LMP1 expression, which lacked p100 processing to p52, or following LMP1 induction, which stimulated p100 processing. Extracts were subject to immuno-precipitation with the anti-p52 antibody. *Asterisk indicates the immunoglobulin heavy chain (Ig). (C) The anti- p52 antibody did not immune-precipitate RelA, RelB, or p50 from 293 cells (293 cells do not produce sufficient cRel for analysis). 293 cell extracts were prepared from cells without LMP1 expression, to avoid p100 processing and p52 heterodimer formation. (D) Proteomic analysis of NF-κB dimers present in LCLs. Using nuclear extracts from GM12878 LCLs that express an HA-tagged NF-κB subunit (either RelA, RelB, cRel or p50), HA-immuno-precipitation was followed by HA-peptide elution. Tandem mass spectrometry was used to identify the abundance of purified NF-κB subunits. The average number from biological replicate samples of peptides retrieved for each NF-κB subunit are shown. Results from a control purification from GM12878 that express HA-GFP are shown for comparison. (E-F) Analysis of endogenous NF-κB heterodimers present in untagged GM12878. Endogenous RelA, RelB, cRel (E) p50 or p52 (F) were immuno-precipitated from GM12878 extracts, and western blotted to analyze NF-κB heterodimer formation, as shown.

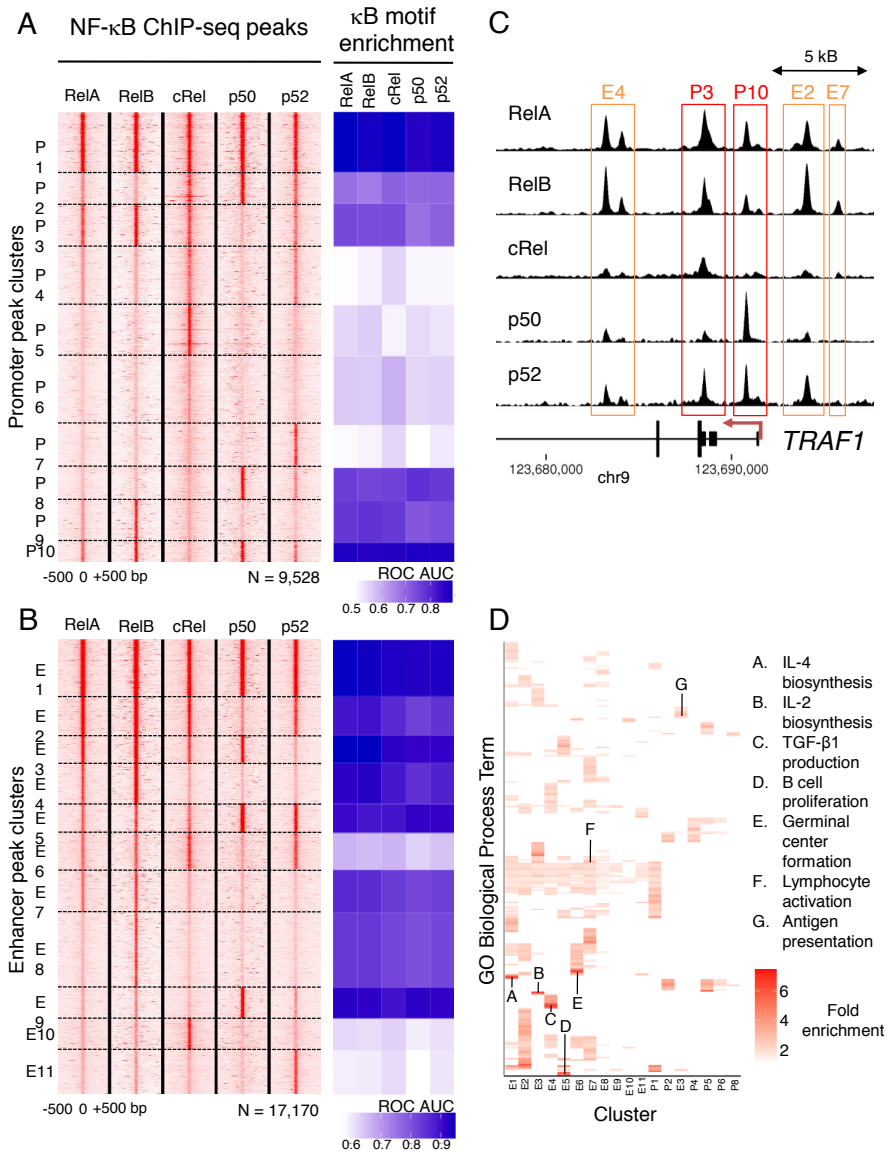


Figure 3.3: NF-κB subunit binding profiles. (A) 10 promoter and (B) 11 enhancer peak clusters with distinct NF-κB subunit binding profiles were identified by *k*-means clustering of ChIP-Seq signals at regions bound by at least one subunit. Red values indicate higher ChIP-Seq signal intensity. The total number of promoter vs. enhancer NF-κB binding sites is shown at the lower right. The heat map to the right of the peak clusters displays the extent of consensus *de novo* NF-κB subunit motif enrichment in each cluster. (C) NF-κB subunit ChIP-Seq signals at the TRAF1 locus illustrate the co-occurrence of multiple SBPs at an NF-κB target gene. Red boxes enclose promoter-associated peaks and orange boxes enclose enhancer-associated peaks. (D) Gene set enrichment analysis of NF-κB clusters using GO Biological Process (BP) terms, as determined by GREAT analysis. Each row corresponds to a unique GO BP term with a false discovery rate (FDR) < 0.01. A subset of highly enriched terms is highlighted.

We reasoned that ChIP-Seq analysis of the five NF-κB subunits in GM12878 might identify target genes unique to either pathway. Indeed, we found SBPs with predominant cRel (clusters P5 and E10) or p52 (clusters P7 and E11) binding, indicative of canonical versus non-canonical activity, respectively (Figures 2A and 2B). Strikingly, most observed SBPs were not readily explained by subunits that are activated by just one pathway. Rather, they were hybrids that resulted from activation of both pathways. For example, cluster E1 and P1 genomic regions were highly occupied by all five NF-κB subunits and were therefore targeted by subunits activated by both NF-κB pathways. Similarly, RelA, RelB, cRel and p52 were present at clusters P3, E2, and E6. The abundance of RelA, RelB and cRel heterodimers with p50 and with p52 in LCL nuclei (Figure 3.2), as well as NF-κB homodimers, likely contributed to these patterns. Although p50- and p52-containing heterodimers are prototypical canonical and non-canonical pathway dimers, respectively, RelA, RelB and/or cRel predominated at clusters P5, P9, E4, E7, E8, and E10. These results indicate that both NF-κB pathways contribute to the activation of many LMP1 target genes in LCLs.

3.2.3 SBPs are associated with unique biological processes

To investigate whether NF-κB binding at promoters versus enhancers might correspond to different LCL biological functions, we evaluated each cluster for enrichment of Gene Ontology (GO) annotation terms. We used GREAT [213] analysis to assign ChIP-Seq peaks to their putative target genes, mainly by proximity. Most clusters were enriched for distinct GO Biological Process terms and mouse knockout phenotypes (Figure 3.3D), consistent with the hypothesis that different SBPs have distinct roles in NF-κB responses. Since the formation of many SBPs requires both NF-κB pathways to be active, this result has relevance for the observation that CD40-mediated canonical and non-canonical pathway activation engenders phenotypes that activation of either pathway alone does not produce [196]. Intriguingly, we obtained a larger number of significantly enriched GO terms at enhancers than at promoters, and observed that promoter clusters were typically enriched for 'housekeeping' functions,

while enhancer clusters were often enriched for B-cell specific functions.

3.2.4 An 11 bp κB motif with 3' cytosine is enriched in all clusters with p50 occupancy

The canonical κB motif is 10 bp long, although *in vitro* studies have found that different NF-κB dimers prefer sites that are 9 to 12 bp[91]. The p50 homodimer recognizes an 11 bp κB motif and makes base-specific contacts with cytosine at position 11[214,215]. LCL ChIP-bound p50 binding sites were highly enriched for an 11 bp κB motif ending in cytosine, providing the first genome-wide confirmation of the importance of this p50 recruitment motif (Figure 3A). The extent of 11-bp enrichment at *in vivo* p50 binding sites was unexpected, since p50 homodimers, and to a lesser extent p50 heterodimers, exhibited moderate preference for this motif *in vitro*[91]. Since the 11bp κB motif was highly enriched in all clusters with high p50 ChIP signal, and since LCLs contain abundant p50 heterodimers (Figure 3.2), our results suggest that p50 heterodimers also prefer the 11-bp site (Figures 3.4A and S3). The 3' cytosine in p50-bound sites was evolutionarily conserved across 33 mammalian species (Figures 3.4C), supporting its importance. The fifth, largely degenerate, base pair in κB motif instances was also frequently conserved, consistent with this position influencing cofactor recruitment[93].

3.2.5 Effect of κB motifs on subunit binding

We investigated the extent to which specific κB motif sequences determined each SBP. We compared κB sites at ChIP peaks in each cluster with protein binding microarray (PBM) data, which provide binding preferences for specific NF-κB dimers[91]. We calculated the area under the receiver operating characteristic (AUC) curve as a measure of motif or 12-mer enrichment in each cluster. The enrichment values obtained using ChIP-derived *de novo* motifs (Figures 3.3A and 3.3B) were generally similar to those obtained using PBM-derived 12-mer data. Aside from the discriminatory power afforded by the 3' cytosine in the 11-bp κB motif, the κB motif did not vary greatly across clusters, suggesting that other

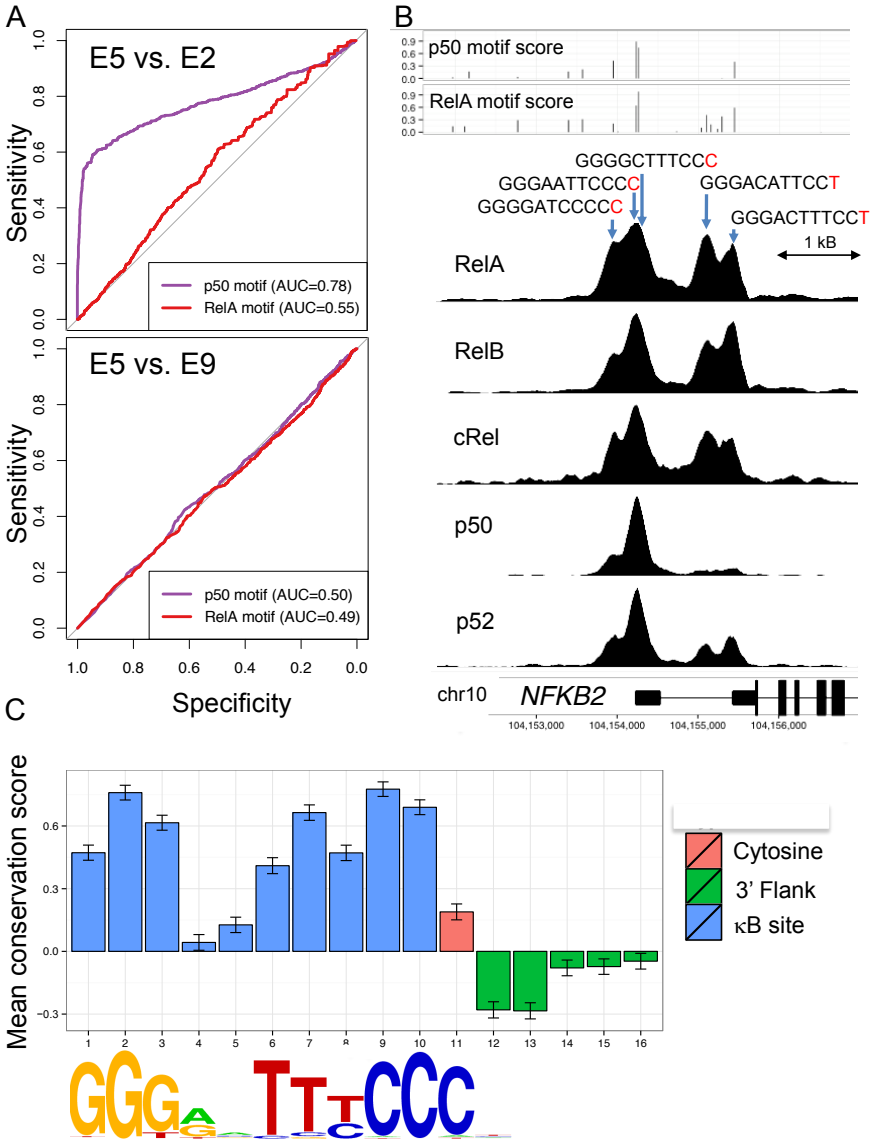


Figure 3.4: Effect of an 11 bp motif with a 3' cytosine on p50 recruitment to NF-$\varkappa$B sites. (A) ROC curves show that the p50 motif 11th base pair cytosine predicts p50 ChIP occupancy with high specificity. In the comparisons shown, peaks from clusters E5 and E2 (top; differential p50 binding) and E5 and E9 (bottom; both bound by p50) are compared. The maximum p50, but not RelA motif match score for each sequence, predicts whether peaks belong to E5 (defined as a positive) or to E2 (negative). Meanwhile, neither motif has predictive power when comparing the p50-bound clusters E5 and E9. See also Figures S3B-D. (B) p50 signals are significantly higher at NF$\varkappa$B2 locus binding sites that contain the 11 bp motif. Other NF-$\varkappa$B subunits bind more uniformly. Motif match scores are scaled from a minimum match threshold of 0.0 to a maximum of 1.0. (C) The functional importance of the p50-preferred 11 bp $\varkappa$B site 3' cytosine is supported by its evolutionary conservation across 33 mammalian species. Each bar shows the mean GERP++ score (higher values indicate more evolutionary constraint) at various positions of $\varkappa$B motif instances. Error bars indicate 1 s.e. of the mean.

mechanisms were responsible for establishing specific NF-κB binding patterns.

3.2.6 NF-κB recruitment to DNA sites that lack a κB motif

Nearly one-third of LCL NF-κB subunit-bound active promoters and enhancers did not harbor instances of κB motifs. Interestingly, four promoter (clusters P4-P7) and three enhancer clusters (clusters E6, E10 and E11) were not highly enriched for a κB motif, suggesting alternative mechanisms for NF-κB recruitment to these sites (Figures 3.3A, 3.3B). While NF-κB recruitment to DNA regions that lack κB sites has been observed previously, alternative motifs that directly or indirectly recruit NF-κB to these sites had not been identified. Using *de novo* motif discovery, we identified four alternative motifs that were associated with specific combinations of NF-κB subunit binding in LCLs and that may participate in NF-κB recruitment to these sites. Each of these discovered motifs was further supported by overlapping ChIP-Seq signals and centralized enrichment of motif instances within NF-κB peaks.

First, we asked how cRel might selectively be recruited to promoters in the absence of κB motifs (cluster P5), and found significant enrichment of E-box motifs (AUC = 0.61, $p = 1.66 \times 10^{-18}$) (Figure 3.5). Indeed, the basic helix-loop-helix (bHLH) transcription factors USF1 and USF2, which recognize E-box motifs, co-occupied 40.2% and 39.3% of GM12878 cluster P5 regions, respectively (Figures 3.5 and 3.6). Our results support a model in which bHLH factors recruit cRel homodimers to LCL E-box sites.

p52 was selectively recruited to genomic regions belonging to clusters P7 and E11. *De novo* motif analyses identified a composite ETS/ISRE-consensus element (EICE) in cluster E11 (AUC = 0.66, $p = 1.87 \times 10^{-45}$), rather than a peak-centered κB motif. EICE motifs recruit PU.1 and IRF4 heterodimers and are essential for lymphocyte development and activation[216]. Indeed, ENCODE GM12878 data confirmed PU.1 and IRF4 co-occupancy at many E11 sites (Figure 3.5). PU.1 may also function as a pioneer factor at these sites by creating areas of nucleosome-free DNA that are accessible to p52[89,190,217]. However, selective p52 recruitment to EICE sites, in the absence enrichment for a κB motif or other

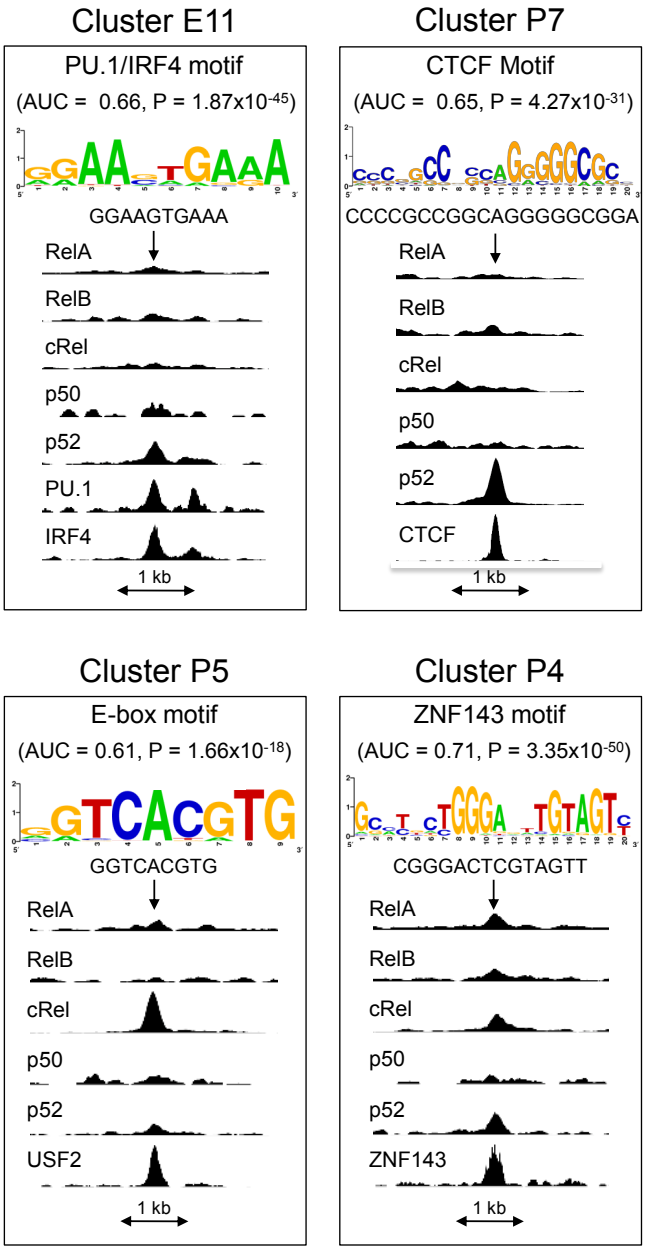


Figure 3.5: Alternative motifs enriched in clusters that lack κB sites. De novo motif discovery revealed enrichment for E-box, CTCF, composite PU.1/IRF4, and ZNF143 motifs at several clusters with low κB site enrichment. The discovered motif, AUC and p-value motif enrichment values, and ChIP-Seq signals at representative loci are shown.

identifiable motifs, is consistent with a direct role for PU.1 and IRF4 in p52 recruitment. Notably, PU.1 motifs were not identified as being enriched by *de novo* analysis in other clusters that lacked κB motif enrichment. p52 recruitment to EICE sites may thereby enable cross-talk between the non-canonical NF-κB, PU.1 and IRF4 pathways, each of which are important for B-cell development and activation [216].

An alternative mechanism may selectively recruit p52 to P7 promoters, in the absence of κB motifs. *De novo* analysis instead identified the CTCF motif to be enriched within P7 ChIP-Seq peaks (AUC = 0.65, p = 4.27 x 10^{-31}), while the EICE motif was not significantly enriched in P7 (AUC = 0.52) (Figure 3.5). In support of a possible CTCF-dependent recruitment mechanism, ENCODE data demonstrated CTCF occupancy at many P7 sites (Figure 3.6). Since CTCF coordinates long-range interactions between DNA regulatory elements together with cohesin [218], we examined whether other cohesin complex members co-occupied P7 sites. Interestingly, 13.7% of P7 sites were co-occupied by SMC3 and RAD21, and 24.7% of P7 peaks were co-occupied by either SMC3 or RAD21 (Figures 3.6). Long-range enhancer-promoter looping interactions involving RelA have been shown to arise as a result of TNFα stimulation in human fibroblasts [198].

All NF-κB subunits except p50 were recruited to cluster P4 promoters, which were enriched for the ZNF143 motif (AUC = 0.71, p = 3.35 x 10^{-50}). High ZNF143 ChIP signals were detected near the centers of cluster P4 promoters (Figures 3.5 and 3.6). How NF-κB is selectively recruited by ZNF143 to P4 promoters, but not other promoters bound by ZNF143, requires further investigation. Collectively, our data suggest that NF-κB recruitment to DNA in the absence of κB motifs significantly expands the range of NF-κB genomic targets, and enables subunits to perform unique functions.

3.2.7 NF-κB predominant versus highly co-occupied LCL sites

Comparison of our datasets with ENCODE ChIP-Seq data, obtained for 65 other TFs in GM12878 cells, identified two classes of NF-κB occupied promoters and enhancers. One class was bound either

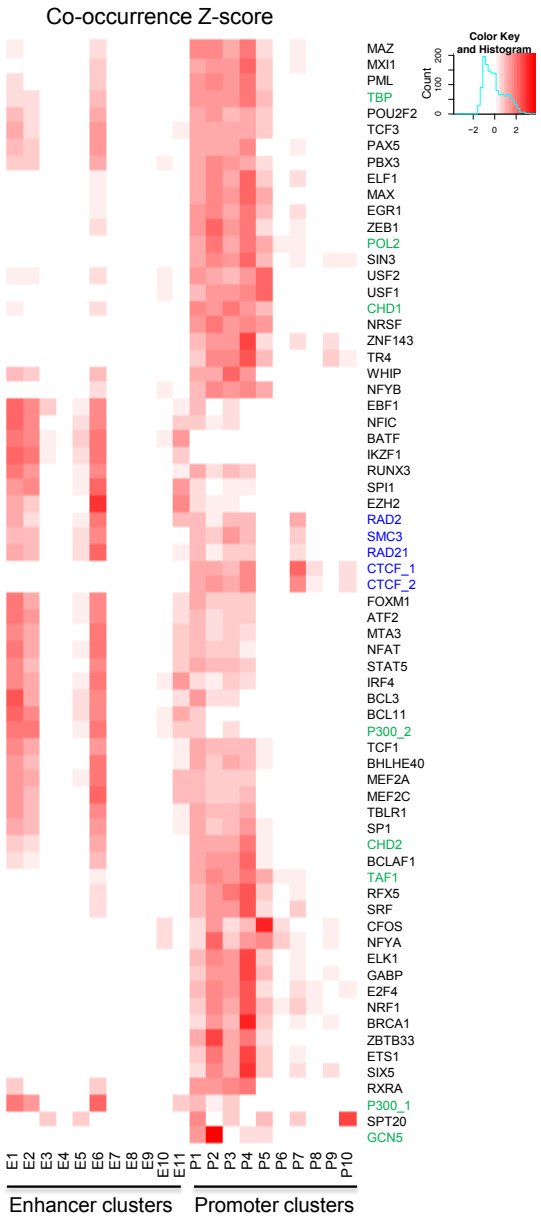


Figure 3.6: Co-occurrence of NF-$\varkappa$B subunits and GM12878 TF ChIP-Seq peaks. Red hues indicate the extent of NF-$\varkappa$B co-occupancy with indicated TFs in GM12878 at promoter and enhancer clusters, normalized for the number of cluster elements and the number of total peaks for each TF. Basal TFs names are indicated in green, DNA looping factors are in blue.

exclusively by NF-κB (clusters E4, E7-9), or by NF-κB in combination with a small number of other TFs (clusters E3, E10, P3, P6, P8-10) (Figure 3.6). The second class was bound by NF-κB together with many TFs, which occurred in different combinations across the cluster (clusters E1-2, E5-6, E11, P1-5, P7). Distinct TF profiles were generally apparent at enhancers versus promoter clusters (Figure 3.6).

3.2.8 NF-κB and FOXM1 are present together in DNA-bound complexes at κB sites

Incorporation of ENCODE GM12878 ChIP-Seq data revealed that multiple TFs co-occurred with NF-κB, including both well-characterized and novel putative NF-κB cofactors. The oncogenic forkhead TF FOXM1 was present at nearly 59% of enhancers occupied by NF-κB, and at 50% of all NF-κB occupied LCL sites. NF-κB co-occupied sites comprised nearly half of FOXM1 genome-wide binding in LCLs. Intriguingly, a κB motif, but not a forkhead recognition motif, was enriched at these sites (Figure 3.7A). At strong enhancers, as defined by chromatin states[71], FOXM1 and NF-κB subunit ChIP peak summit heights were correlated (Spearman R = 0.5 for p52). Moreover, FOXM1 co-occupied κB sites at many well-established NF-κB target genes, such as TNFAIP3 (which encodes A20), NFKBIA (which encodes IκBα), BIRC3 (which encodes cIAP2) and CXCR4 (which encodes CXCR4) (Figures 3.7B and S4). These results suggest that NF-κB, or another factor that interacts with NF-κB, recruited FOXM1 to these LCL sites, particularly in clusters E1, E2, E5, E6 and E11.

NF-κB and FOXM1 are hyper-activated in many of the same malignancies[187,219]. Despite also having numerous overlapping biological roles, FOXM1 and NF-κB are not known to be cofactors. We therefore assessed whether FOXM1 and NF-κB are present together in DNA-bound protein complexes in LCLs. First, we used sequential GM12878 ChIP (ChIP re-ChIP), in which anti-FOXM1 ChIP was followed by ChIP using either anti-RelA antibody, anti-FOXM1 antibody (positive control), or no antibody (negative control). Quantitative real-time PCR data showed that PLK1 or BCL2 target loci were significantly enriched in the RelA ChIP versus the control (Figure 3.7C), suggesting that NF-κB

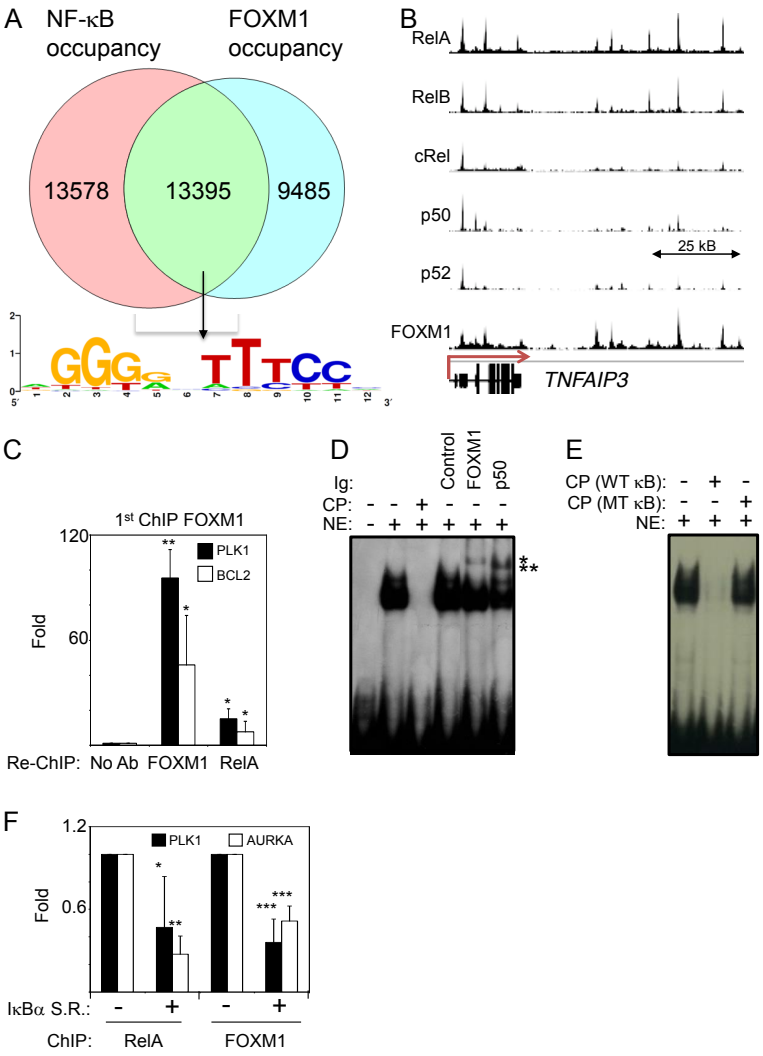


Figure 3.7: NF-κB and FOXM1 are present in DNA-bound protein complexes at κB sites. (A) Venn diagram showing the extent to which NF-κB and FOXM1 ChIP-Seq peaks overlap in GM12878. The number of sites occupied by NF-κB, FOXM1, or co-occupied by both, and the consensus *de novo* motif for sites co-occupied by NF-κB and FOXM1, is shown. (B) FOXM1 ChIP-Seq signals at the TNFAIP3 locus show a high degree of co-occupancy with NF-κB. (C) ChIP-re-ChIP identified FOXM1 and RelA as present together in DNA-bound protein complexes at the PLK1 and BCL2 loci. FOXM1 ChIP was followed by RelA ChIP. Mean fold enrichment -/+ SD for replica experiments is shown. *p<0.05, **p<0.01. (D) EMSA identified that both FOXM1 and NF-κB from GM12878 LCL nuclear extract (N.E.) bind a PLK1 promoter DNA probe that contains a κB site, but no forkhead recognition site. Incubation with a cold probe (C.P.) is indicated. *FOXM1 super-shift. **p50 super-shifted band. A representative EMSA of three independent experiments is shown. (E) A PLK1 promoter probe with a central κB motif (WT κB), but not a probe with a mutant κB motif (MT κB), competed for binding in EMSA (see Supplemental Experimental Procedures for full details). (F) Inducible expression of an IκBα super-repressor (IκBα S.R.) in IB4 LCLs diminished RelA and FOXM1 ChIP signals at the PLK1 and AURKA promoters. Mean + SD for replica experiments is shown. *p<0.05, **p<0.01, ***p<0.001

and FOXM1 are part of a DNA-bound protein complex. Second, electrophoretic mobility shift assays (EMSA) using GM12878 nuclear extract and DNA probes representing the PLK1 and AURKA regions further validated FOXM1 recruitment to κB sites. Supershift assays, in which antibodies against p50 or FOXM1 were added to the binding reaction, produced a slower migrating band, consistent with recruitment of both NF-κB and FOXM1 to the probe (Figure 3.7D). Notably, the probe contained a central κB site, but not a forkhead recognition motif, and had minimal flanking DNA, supporting FOXM1 recruitment by a NF-κB-dependent mechanism. A DNA probe with mutant κB site failed to compete for binding (Figure 3.7E). Finally, induced expression of a non-degradable IκBα super-repressor in IB4 LCLs significantly reduced both RelA and FOXM1 occupancy at the PLK1 and AURKA loci (Figure 3.7F). Our results suggest that NF-κB and FOXM1 are present together in DNA-bound complexes at NF-κB sites, and that recruitment to NF-κB sites is dependent on NF-κB DNA binding.

To investigate the functional consequences of FOXM1 recruitment to κB sites, we tested the effects of FOXM1 depletion on NF-κB target gene expression. By 48 hours after short hairpin RNA (shRNA) lentiviral delivery, each of three different anti-FOXM1 shRNAs strongly reduced FOXM1 expression, and also markedly impaired expression of loci co-occupied by NF-κB and FOXM1, including TNFAIP3, BIRC3, CXCR4, NFKBIA, and MAP3K7 (Figure 3.8A). By contrast, FOXM1 depletion did not impair expression of control LCL target genes, whose promoters and proximal enhancers were not occupied by either NF-κB or FOXM1 (Figure 3.8B). FOXM1 depletion did not affect cell viability at 48 hours post-transduction (Figures 3.8C). However, all three FOXM1 shRNAs reduced the number of cells in S-phase and triggered apoptosis at 120 hours post-transduction (Figures 3.8D and S5B-D). While NF-κB-independent FOXM1 cell cycle roles may have strongly contributed to this phenotype, it nonetheless underscores FOXM1 as a novel LCL synthetic lethal target.

FOXM1 is a master regulator of germinal center B-cell proliferation [220] and is expressed in diffuse large B-cell lymphoma (DLBCL) [221,222]. Impelled by these and our results, we investigated whether

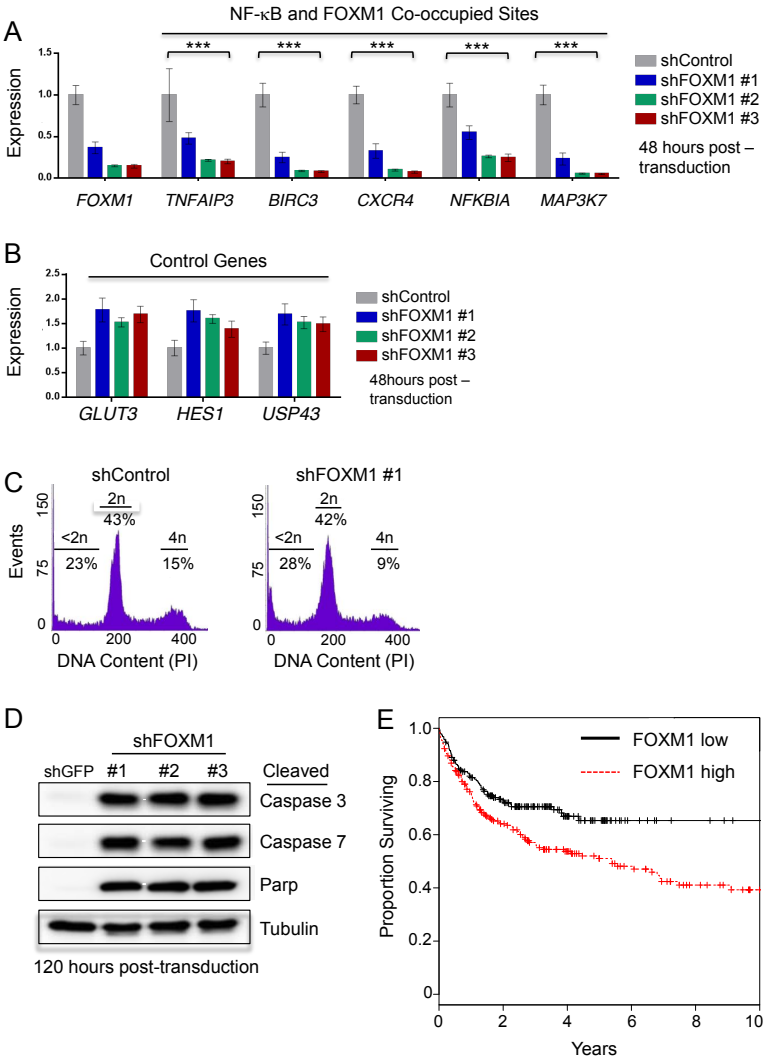


Figure 3.8: FOXM1 cooperates with NF-κB to regulate GM12878 target gene expression. (A) Three independent shRNAs against FOXM1 reduced expression of key NF-κB target gene mRNAs by 48 hours after delivery. Mean -/+ SEM effects from three independent experiments are shown. $p < 0.05$ for TNFAIP3, $p<0.01$ for all other comparisons between control and FOXM1 shRNAs. (B) FOXM1 shRNAs did not reduce expression control genes (which were not identified as NF-κB or FOXM1 targets in GM12878) at 48 hours post-shRNA delivery. (C) FOXM1 depletion inhibited LCL proliferation and caused accumulation of cells with <2n DNA content by Propidium Iodide (PI) analysis. (D) At 120 hours after shRNA delivery, FOXM1 depletion triggered cleavage of caspase 3, 7, and their substrate PARP. Tubulin load control is shown. Representative western blot of three independent experiments is shown. (E) FOXM1 expression in DLBCL tumor samples correlated with worse clinical outcome. Retrospective analysis of FOXM1 expression in microarray datasets obtained from 414 DLBCL tumor samples, and its relationship with patient survival (Wald test P-value = 0.0037).

FOXM1 expression correlates with clinical outcome in DLBCL. We retrospectively analyzed microarray datasets from 414 DLBCL tumor samples[223], and found that elevated FOXM1 expression levels were significantly correlated with worse overall survival, even controlling for tumor stage and subtype (P = 0.0037, Wald test) (Figure 3.8E). While FOXM1 roles independent of NF-κB may underlie this observation, our analysis nonetheless suggests that FOXM1 levels may be a valuable prognostic indicator in DLBCL.

3.3 Discussion

In the classical model of NF-κB activation, stimuli trigger IκB degradation, NF-κB dimer nuclear translocation and κB site binding. However, this model does not adequately consider the complexities that further shape NF-κB nuclear function[191,224,225]. Likewise, most genomic studies of NF-κB binding have focused on immediate events following canonical pathway stimulation by TNF-α or lipopolysaccharide, and have not fully addressed why both pathways are needed to activate particular target genes. To our knowledge, our results provide the first genomic survey of NF-κB subunit binding when both the canonical and non-canonical NF-κB signaling pathways are persistently active. Consequently, new insights into the extent of cross-talk between the canonical and non-canonical pathways emerged.

The NF-κB binding landscape in LCLs was complex, but largely describable in terms of a small number of SBPs, suggestive of both common and unique NF-κB subunit roles. Frequently, subunits activated by both the canonical and non-canonical pathway each contributed to SBPs. These results provide novel insights into how NF-κB may function during physiologic B-cell activation, where CD40-mediated persistent activation of both the canonical and non-canonical pathways is central to the generation of germinal centers and humoral immunity[195,196]. Our datasets provide a resource for studies of EBV oncoprotein-mediated NF-κB activation, constitutive NF-κB activity in tumors, and comparative genomics, since many TF families similarly evolved by gene duplication and diversifica-

tion.

Numerous NF-κB cofactors, for which GM12878 data are not yet available, might contribute to SBP formation. For instance, ENCODE ChIP-Seq data are not yet available for RPS3, which binds to RelA and promotes RelA:p50 and RelA:RelA dimer binding to select κB sites[191]. An important future area of investigation will be the identification of B-cell cofactors that may similarly affect dimer binding properties and thereby contribute to shaping the observed SBPs.

The extent to which NF-κB binding activates transcription remains an open question. Studies in TNFα-stimulated LCLs and LPS-stimulated THP-1 monocytes suggested that only a minority of RelA binding events induce transcription[199,200]. However, limitations in the assignment of enhancers to their target genes may have resulted in underestimates of regulatory binding events. In contrast, we found that nearly all NF-κB binding in LCLs, including by p50 and p52, occurred at active enhancers or promoters. Indeed, NF-κB promoter occupancy highly correlates with induction of transcription in LPS-stimulated monocytes[202], and the vast majority of NF-κB binding events occurs at active promoters and enhancers in LPS-stimulated murine dendritic cells[201].

Nearly one-third of LCL NF-κB binding events occurred at DNA sites lacking κB motifs. cRel and p52 may more frequently be recruited to these sites (clusters P5, P7, E10 and E11). Consistent with our findings, a prior ChIP-ChIP study of RelA chromosome 22 binding events in TNFα-stimulated HeLa cells reported that 44% of identified RelA sites did not have a κB motif[200]. Similarly, ChIP-Pet analysis of LPS-stimulated THP-1 cells found the RelA motif to be absent at 57% of RelA binding sites[199]. However, alternative NF-κB recruitment motifs were not identified. We report four motifs that are highly enriched at LCL NF-κB binding sites that lack κB motifs: E-boxes at cRel-occupied promoters, ZNF143 motifs at promoters occupied by all NF-κB subunits except p50, CTCF sites at p52-occupied promoters, and Ets/ISRE elements p52-occupied enhancers. Indirect recruitment to sites lacking κB motifs may provide an important mechanism through which NF-κB subunits perform specific functions and cross-talk with other pathways. Our analysis offers insights into why each NF-κB

subunit has non-redundant roles in B-cells[98].

NF-κB requires additional transcription co-factors to fully activate target gene expression[191,224,226]. Certain promoter and enhancer clusters were co-occupied by at least ten additional TFs, though it is likely that fewer bind to an individual site at the same time. Binding at these loci is unlikely to be an artifact of the ChIP experimental procedures, as SBPs with the highest co-occupancy (*e.g.*, P1, E1, E2) were enriched for κB binding site sequences. NF-κB was also found to bind frequently at highly co-occupied sites in LPS-stimulated murine dendritic cells[201]. High TF co-occupancy may be due to a more accessible chromatin state at these genomic regions.

Our data highlight FOXM1 as an important coactivator of NF-κB target gene transcription in LCLs, present at 50% of all NF-κB peaks. Since a FOXM1 DNA motif was not enriched at these sites, our data are consistent with a model in which NF-κB recruits FOXM1 to κB sites, either directly or indirectly through additional cofactors. In support the former possibility, the MMB activator complex directly recruits FOXM1 to coactivate transcription[214]. MuvB and B-Myb also interact with FOXM1 to regulate gene expression during the G2 phase of cell cycle[227]. Curiously, we did not find evidence for NF-κB recruitment to FOXM1-bound forkhead box recognition sites; DNA allostery may induce conformation changes in the bound TF, leading to differences in protein-protein interactions[93]. FOXM1 depletion impaired transcription of key NF-κB target genes, and ultimately induced LCL apoptosis, reminiscent of the phenotype of NF-κB inhibition on these cells[228].

To our knowledge, FOXM1 has not previously been reported to function jointly with NF-κB in target gene regulation. However, cross-talk between NF-κB and FOXM1 may underlie published studies. Both FOXM1 and NF-κB are implicated in the pathogenesis of K-Ras-induced non-small cell lung cancer[229]. Moreover, conditional FOXM1 deletion impairs K-Ras-mediated expression of multiple NF-κB pathway components, including IKK-β, RelA, p105/p50, and p100/p52. Intriguingly, NF-κB and FOXM1 have each been implicated as drivers of B-cell lymphomagenesis[188,221,222,230], although a joint role of FOXM1 and NF-κB in driving B-cell malignancy has not yet been proposed.

Despite the promise of therapies that block pathogenic NF-κB hyperactivity in cancer and autoimmune diseases, side-effects have largely precluded the use of broadly-acting NF-κB inhibitors, such as IKK-β kinase antagonists. Uncovering subunit-specific transcriptional mechanisms may facilitate approaches to selectively alter NF-κB target gene expression. Targets that require both canonical and non-canonical NF-κB pathway activation may be particularly sensitive to disruption. Since FOXM1 is not expressed in most adult tissues, our data highlight the FOXM1 pathway as a potential therapeutic target in B-cell malignancy. An increasingly sophisticated understanding of NF-κB nuclear function promises to highlight novel therapeutic strategies for selective NF-κB inhibition.

3.4 Methods

3.4.1 Cell lines and antibodies

GM12878 cells were cultured in RPMI 1640 medium supplemented with 10% Fetalplex (Gemini), L-glutamine, streptomycin, and penicillin. IB4 cells expressing tetracycline-regulated IκBα deleted for the N-terminal 36 amino acids (dN)[228] were cultured in RPMI 1640 medium supplemented with 0.2ug/ml hygromycin, 0.25 ug/ml G418 and 1 ug/ml tetracycline. The following antibodies were used for ChIP, ChIP-Seq and ChIP-re-ChIP: p50, sc-1190; p65, sc-372; RelB, sc-226; c-Rel, sc-71; FOXM1, sc-502, (all from Santa Cruz Biotechnology); and p52, A300-BL7039 (Bethyl Laboratories). Murine stem cell leukemia virus vectors were used to derive GM12878 cell lines with HA-epitope tagged NF-κB subunit for our LCL dimer analysis, as previously described[231]. HEK-293 cells with inducible LMP1 expression have been previously described[232].

3.4.2 ChIP-Seq experimental procedures

GM12878 cells were cross-linked with 1% formaldehyde. Chromatin was sonicated to an average size of 500 bp. Biological replicates were obtained using cells grown on separate days. Antibodies against

RelA, RelB, cRel, p50 and p52 were used to immunoprecipitate protein-DNA complexes. Captured protein-DNA complexes were eluted from protein A beads and reverse-crosslinked. Eluted DNA was purified using PCR purification columns (Qiagen). ChIP-Seq libraries were prepared using NEBNefxt library preparation kits (New England Biolabs) and sequenced by HiSeq 2000 (Illumina).

3.4.3 ChIP-Seq data analysis

To analyze our ChIP-Seq data, we followed the set of standards and guidelines established by the ENCODE and modENCODE project consortia[211]. We used strand cross-correlation analysis as the primary metric for success of a ChIP-Seq experiment[233]. Landt et al. introduced two main metrics to evaluate the signal-to-noise ratio of a ChIP-Seq experiment: the normalized strand coefficient (NSC), which quantifies the fragment-length cross- correlation over the background cross-correlation rate, and the relative strand correlation (RSC), which computes the ratio of cross-correlation observed at the predicted fragment size against the artifactual cross-correlation observed at read length[211]. All the NF-κB ChIP-Seq datasets we generated had NSC and RSC values that exceeded the ENCODE criteria for experiment success (NSC > 1.05 and RSC > 0.8).

Reads from all ChIP-Seq experiments were mapped to the hg19/GRCh37 build of the human genome using bowtie v0.12.8[233]. Aligned reads that had more than 2 mismatches or that did not map to a unique position in the genome were discarded. The "–best" and "–strata" flags were used to enable the breaking of ties using read quality scores. Samtools v0.1.19 was used to generate compressed BAM files from the bowtie output.

We eliminated all reads that were mapped to genomic regions that were blacklisted by the ENCODE consortium because of their potential to cause high-signal artifacts[22]. The blacklisted regions include centromeric and telomeric repeats, satellites and high mappability islands. For more information about how the blacklisted regions were generated, see:

https://sites.google.com/site/anshulkundaje/projects/blacklists

We used SPP v1.10 [234] (http://compbio.med.harvard.edu/Supplements/ChIP-Seq/) to identify regions with high enrichment of ChIP-Seq tags ("peaks"). SPP was first used to call peaks at a relaxed threshold to obtain ~300,000 peaks for each ChIP-Seq replicate. To determine a threshold at which peaks were deemed to be reproducible across replicates in a statistically significant manner, we used the Irreproducible Discovery Rate (IDR) framework [235]. For each subunit, we directly compared the peaks obtained in replicate experiments and set the peak calling threshold to yield an IDR of 1%.

3.4.4 Generation of continuous ChIP signal tracks

To generate ChIP signal tracks for visualization, we first used the fragment length estimates from SPP to computationally extend mapped reads to their predicted fragment size. Afterwards, we used Wiggler [236] with a smoothing window (-w) parameter of 300 to generate bigWig files for all of the experiments.

The y-axes in all of the ChIP signal tracks displayed in the figures are scaled from 0 (no ChIP fragments observed) to whichever value was larger: (a) 20 times the median for the 100-kb region centered on the locus depicted in the track, or (b) the maximum signal observed within that track region.

3.4.5 Comparison with ENCODE ChIP-Seq data

For the peak regions in each promoter or enhancer cluster, we computed the fraction of NF-κB peak regions that overlapped with ENCODE ChIP-Seq peaks for all of the experiments that passed QC. ENCODE ChIP-Seq peaks were weighted by both their height and the number of base pairs that overlapped NF-κB peaks to compute a raw co-occurrence score. However, the ENCODE data vary widely in terms of the number of peaks reported. Therefore, to normalize for the number of peaks per ChIP-Seq experiment and their potentially varying width, the overlap was represented as a fraction of the total number of bases covered by peaks in each ENCODE experiment. Then, this co-occurrence score was divided by the number of peaks occurring in each cluster, to control for the different numbers of

elements in different promoter and enhancer clusters. Finally, these scores were z-score normalized to show co-occurrence values in a consistent scale.

All of the ENCODE GM12878 TF and co-activator ChIP-Seq data sets that were used in our comparisons were downloaded from the hg19 version of "Transcription Factor ChIP- seq Uniform Peaks from ENCODE/Analysis" datasets at the UCSC ENCODE Project portal. The GM12878 ChIP-Seq data for histone modifications were also downloaded from the UCSC ENCODE portal, from "Uniform Signals tracks."

3.4.6 Comparison with HOT and LOT regions

Yip et al. used ENCODE GM12878 ChIP- seq data to identify genomic regions with particularly high and low degrees of TF co- binding (HOT and LOT regions, respectively)[237]. We computed the fractional overlap of all NF-κB peaks from each cluster with GM12878 HOT and LOT regions. A value of 1 indicates that all peaks in a given cluster overlapped with HOT or LOT regions, while a value of 0 indicates that none did. We used the HOT and LOT prediction tracks obtained from:

http://metatracks.encodenets.gersteinlab.org/HOT_Gm12878_merged.bed.gz

http://metatracks.encodenets.gersteinlab.org/LOT_Gm12878_merged.bed.gz

3.4.7 Chromatin state distribution

We used GM12878 ChromHMM chromatin state annotations[71] to annotate the overlap between NF-κB subunits and chromatin states. Briefly, ChromHMM bins the genome into 200-bp regions and classifies each of these regions into various chromatin states based on the co-occurrence patterns of the histone marks H3K27me3, H3K36me3, H4K20me1, H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K9ac, and the binding of CTCF. To simplify the display of the chromatin state information, chromatin states from the paper were merged into more general categories, as shown in the following table. Because of its small number of occurrences, the "poised promoter" state was excluded.

Table 3.1: Schema for merging ChromHMM states

Merged state	Original ChromHMM states
Promoter	Active promoter, Weak promoter
Enhancer	Strong enhancer, Weak enhancer
Insulator	Insulator
Transcribed	Transcriptional transition, Transcriptional elongation, Weakly transcribed
Repressed	Polycomb repressed
Other	Heterochromatic/low signal, Repetitive/CNV

3.4.8 Clustering of subunit peaks

All loci that were identified as a ChIP peak location for at least one of the NF-κB subunits and that overlapped with a promoter or enhancer ChromHMM state were clustered. NF-κB ChIP signal intensities were used as the numerical values for calculating distances for clustering. Importantly, this approach minimizes threshold effects in peak calling between the five subunits (*i.e.*, loci where two subunits bound, but one barely missed the peak-calling threshold), which could have resulted from variable antibody and experimental quality. We used seqMiner v1.2[212] to find clusters. Briefly, the rank-normalized ChIP signal intensities for each subunit relative to peak centers are used to perform *k*-means clustering. Because each experiment can potentially have a different signal-to-noise ratio, rank-normalization was used to make the peaks comparable across experiments. To determine the number of clusters, we started at a high number (20) and decreased it until clusters were qualitatively unique. Clustering was performed 10 times with different random seeds to verify stability of the clusters.

3.4.9 *De novo* motif discovery

We employed ChIPMunk[238] to discover potential regulatory motifs in the NF-κB and ENCODE ChIP-Seq data. In a recent comparison of methods for modeling TF specificity, ChIPMunk was the top performer among methods for deriving motifs from both ChIP-Seq data and PBM data[39]. To

derive a consensus motif for each subunit (Figure 3.1C), ChIPMunk was run separately on the peaks obtained from each of the five subunit ChIP-Seq experiments. In each case, the continuous signal track was used in addition to the peaks to further enhance the signal-to-noise ratio (using the "-p" option), searching for motifs of length 8 to 12, and performing local GC-content matching.

To find motifs for potential co-factors or TFs that tend to co-occur with different combinations of NF-κB subunits, we combined *de novo* motif finding and enrichment analysis using a dictionary of known motifs. Briefly, to perform *de novo* motif finding, we ran the MEME-ChIP suite[239] independently for the peaks in each cluster. We searched for up to five motifs of length 6 to 20, using both the MEME and DREME algorithms. In addition, the central enrichment of motifs found by the above methods was determined using CentriMo and reported as a P-value. The actual enrichment of all of the motifs found by the MEME-ChIP suite relative to genomic background sequences was computed as described below. To determine enrichment for known motifs, we used HOMER v3.0[89] and its associated dictionary of vertebrate motifs, which includes position weight matrices (PWMs) from several databases, such as TRANSFAC and JASPAR. Q-values were obtained from a false discovery rate (FDR) corrected hypergeometric test. The background sequence set was normalized for dinucleotide composition using the "autonormalize" option in HOMER.

3.4.10 Motif enrichment determination

We used the area under the receiver operating characteristic curve (AUC) statistic to quantify motif enrichment. The entire region that was defined as a peak by SPP was used for motif analysis. Briefly, if the presence or absence of a motif in a genomic sequence influences binding, then the motif score of a given sequence should be able to predict whether a TF will bind or not more accurately than expected by chance. In our comparisons, we compared the motif scores for sequences in a foreground (FG) set, which usually contains loci bound by NF-κB or other TFs, to those in a background set (BG), which are meant to reflect the overall sequence biases of the non-coding genome. The methods used

to obtain an appropriate BG set are described in the following section.

The PWM of a motif can be used to scan any DNA sequence and assign it a motif match score. For each sequence in the FG and BG sets, we computed the maximum PWM log-odds score across all possible alignments and orientations and used it as a motif-match score for the sequence. If most sequences in the FG set have a higher motif match score than the BG sequences, then there should be a threshold motif match score that can predict to which set a sequence belongs. We obtained ROC curves for each PWM by calculating the sensitivity and specificity at different motif match thresholds. The area under the curve (AUC) statistic summarizes the discriminatory power (or equivalently, enrichment) of a motif at different threshold values. AUC values close to 1 indicate that NF-κB subunits bind to sequences containing the particular motif at a much higher rate than to unbound genomic sequences; a value of 0.5 is expected by chance alone (Figures 3.4A). Simultaneously, we used the Wilcoxon-Mann-Whitney U test to assign P-values to each enrichment calculation, as implemented in the R function wilcox.test. To test for central enrichment of the motif relative to the center of ChIP-Seq peaks, we used the CentriMo tool[239].

In some cases, we compared two sets of bound regions (*e.g.*, Figure 3.4A). In such cases, the first set of sequences listed in the comparison is considered the FG set, while the second set is the BG set. Therefore, true positives represent the case where the presence of the motif accurately predicts the sequence as belonging to the FG set. When an AUC score > 0.5 is observed in such comparisons, it implies that the motif can discriminate between the two sets of sequences.

3.4.11 Comparisons with PBM 12-mer data

We performed the analysis similarly to the above case (where PWMs were used), but instead of taking the maximum PWM log- odds score for each sequence, we used the maximum predicted z-score for 12-bp sequences, as determined previously from PBM experiments on NF-κB dimers using custom-designed oligonucleotide arrays[91], to score the sequences bound *in vivo*. Non-traditional sites were

defined as 12-mers that had a z-score > 4.0 as determined from PBM data, but a PWM match score < 4.0. This was similar to the approach of Siggers et al.[91] but used the PWMs derived by ChIPMunk instead.

3.4.12 Generation of background sets

The rate of occurrence of random motif instances in the genome can be highly influenced by local G/C content, particularly in the case of G/C-rich motifs like those of NF-κB proteins. In addition, the propensity of some NF-κB dimers to bind guanine-rich half-sites[91] could artificially inflate enrichment at repetitive regions. Therefore, the BG set should be similar to the FG in both G/C content and the proportion of repetitive sequences to avoid reporting false enrichments.

We used rejection sampling to obtain a background set with an equal distribution of G/C content and proportion of repetitive sequence as the foreground set. For each distinct FG set, we used the "shuffleBed" program (part of the BedTools suite v2.16.2[240]) to obtain genomic intervals that matched the length distribution of the foreground set but that were randomly distributed throughout the genome. All coding exons and blacklisted regions (as defined above) were excluded from the list of allowable positions for the randomly positioned intervals. First, the ranges for observed values of both G/C and repeat percentages were separately discretized into 10 equally sized bins. Then, for each length-matched interval that was randomly placed in the genome, we perform successive rejection sampling steps for G/C content and for repeat content. The sequences that are selected after rejection sampling are then used to form the BG set. This method is identical to that used by Gisselbrecht et al.[241]

3.4.13 Cluster-specific motifs

We used the PWM derived from the RelA ChIP-Seq experiment (which corresponds to the 10-bp κB site) to scan peak sequences from each cluster for the highest scoring motif instance. For each cluster,

we computed the frequencies at which each nucleotide occurred in each position in the sequence of the best scoring motif match. Then, we used the nucleotide frequencies at each position to create a new logo that represented the sequence biases of κB sites specifically in that cluster.

3.4.14 Evolutionary conservation analysis

We used the GERP++ score, as calculated over the evolutionary tree of 33 mammals[242], as a measure of evolutionary constraint in the human genome. Briefly, the GERP++ score for a particular base pair in the genome represents the number of estimated "rejected substitutions" that have occurred at that position throughout mammalian evolution. Higher scores indicate that the identity of that base pair has changed less than would be expected if the sequence were undergoing mutations at the neutral rate and therefore suggest the action of purifying selection.

To generate the plots in Figures 3.4C, all NF-κB peaks in enhancer sites at clusters bound by p50 (E1, E3, E5 and E9) were scanned to find the top scoring motif instance using a 10-bp motif (the RelA PWM shown in Figure 3.1). For each position in the motif, the GERP++ score was averaged across all motif matches.

3.4.15 Gene set enrichment analysis

We used GREAT[213] to predict potential biological functions of different clusters of NF-κB ChIP-Seq bound regions. For each cluster, GREAT assigns peaks to nearby genes, predominantly within -5 kb to +1 kb relative to a gene's transcriptional start site. However, when no other genes are present nearby, or in a few loci with experimentally determined regulatory interactions, GREAT considers interactions up to 1 Mb away. We used GO Biological Process terms to link clusters and SBPs to putative biological functions. All reported terms had a Benjamini-Hochberg FDR q-value < 0.01 in both the hypergeometric and binomial tests used by GREAT. This is the setting recommended by McLean et al. for obtaining high confidence associations[213]. All parameters were set to the default values and the

whole genome was used as the background. For the heat map in Figure 3.3D, we included only those GO Biological Process terms that had 2-fold or greater fold-enrichment.

3.4.16 Analysis of DLBCL datasets

We used a Cox proportional hazards model to determine whether FOXM1 expression was predictive of survival in DLBCL patients. We used the R package Survival to plot survival curves and perform statistical tests. The processed microarray expression data and the survival time statistics were both downloaded from GEO series GSE10846. The hgu133plus2.db library in the Bioconductor package was used to map microarray probes to gene symbols.

The P-value was calculated by performing a Wald test on the FOXM1 expression coefficient after fitting a Cox proportional hazards model to the data using the statistical model formula:

Survival_Time ~ FOXM1_Expression + Tumor_Stage + Microarray_Diagnosis

3.4.17 *In vitro* translation

The TnT® T7 Quick Coupled Transcription/Translation System was used to translate the five NF-κB subunits from sequence-verified expression vectors with each cDNA downstream of the T7 promoter, according to the manufacturer's instructions.

3.4.18 Proteomic analysis of LCL NF-κB dimers

GM12878 cells that stably express either RelA, RelB, cRel, p50 or p52 were established. Nuclei were isolated from 100 million GM12878 cells by hypotonic lysis and dounce homogenization, and HA-epitope tagged NF-κB subunits were then immuno-purified under isotonic conditions, washed extensively, and eluted by HA-peptide competition, as previously described [231]. Eluted material was run on a 10% SDS-PAGE gel, and subjected to liquid chromatography followed by tandem mass spectrometry analysis at the Harvard Taplin proteomics core, as previously described [243]. Alternatively, endogenous

NF-κB subunits were immuno-precipitated from 10 million GM12878 LCLs, using validated NF-κB subunit-specific antibodies and protein A beads, washed extensively, and then subject to western blot analysis, as indicated.

3.4.19 Electrophoresis Mobility Shift Assays

EMSAs were performed using an oligonucleotide probe encompassing the κB site at the PLK1 promoter (Wild-type: CCG GGT CCG TGT CAA TCA GGT TTT CCC CGG CTG GGT CCG GGT T; Mutant: CCG GGT CCG TGT CAA TCA GGT TTT AAA AGG CTG GGT CCG GGT T). 32P-dCTP labeled probes were incubated with nuclear extracts prepared from GM12878 cells, either untreated or treated with PMA (50 ug/ml) and ionomycin (500 ng/ml, 2 hours, 370C). 100x excess of unlabeled probe and NF-κB site mutant probe were added to determine the specificity of NF-κB binding. Antibodies against FOXM1, p50 and control antibody were added the reaction. The protein-DNA complexes were separated by non-denaturing PAGE and visualized by exposure to X-ray film.

3.4.20 ChIP, ChIP-re-ChIP and qPCR

To induce the dN IκBα super-repressor expression, 10^7 IB4 cells were washed three times with fresh media supplemented with Tet-free fetal calf serum. Cells were then grown in the presence or absence of Tet for 3 days. ChIP assays were performed with antibodies against RelA, FOXM1 and pre-immune sera as a control. qPCR using primers for sites occupied by NF-κB and FOXM1 in PLK1 and AURKA promoters was used to quantify the binding to these sites. Fold-enrichment over control was first determined, and the fold enrichment of the uninduced IκBα super- repressor condition was then set to 1. Re-ChIP-IT kit (Active Motif) was used for ChIP- re-ChIP experiments following the manufacturer's protocol. Anti-FOXM1 antibody was used for the first ChIP from GM12878 cells. Re-ChIP was performed using antibodies against FOXM1 or RelA, or as a control with no antibody added. qPCR was used to quantify binding to the PLK1 and BCL2 promoters as described above. Fold enrichment over

the no antibody control was determined with controls set to 1.

3.4.21 shRNA, qRT-PCR, cell-cycle and apoptosis assays

pCMV-dR8.91 and pMD2.G were transfected into 293T cells to produce lentiviruses. Two rounds of lentivirus transduction of GM12878 were performed, 24 hours apart. Lentivirus-transduced GM12878 LCLs were allowed to recover for 24 hours, and then selected in puromycin for 24 hours. Total RNA was extracted using RNeasy kit (Qiagen). qRT-PCR was performed using the RNA-to-Ct 1-step kit (Life Technologies).

For cell cycle analyses, cells were fixed with 70% ethanol, stained with propidium iodide and analyzed on a FACSCalibur instrument. Caspase activity was determined using Caspase-Glo assay systems (Promega). Cell Signaling cleaved caspase assay starter kit was used.

This chapter is a modified version of a published article describing this work:

Zhao B*, Barrera LA*, Ersing I, Willox B, Schmidt S, Greenfeld H, Zhou H, Mollo S, Shi T, Takasaki K, Cahir-McFarland E, Kellis M, Bulyk ML**, Kieff E**, Gewurz B**. The NF-kappaB genomic landscape in lymphoblastoid B-cells. *Cell Reports* (2014) 8(5):1595-606.

Acknowledgments

This project was supported by a Burroughs Wellcome Medical Scientist career award (to B.E.G.), NIH K08 CA140780 (to B.E.G.), NIH RO1 CA085180, CA170023, and CA047006 (to E.K.), a National Science Foundation Graduate Research Fellowship (to L.A.B.), and grant R01 HG003985 from NIH/NHGRI (to M.L.B.). We thank Trevor Siggers and Suzanne Gaudet for critical reading of the manuscript.

Author contributions

B.Z., L.A.B., E.C.-M., E.K., and B.E.G. designed the study. B.Z., M.K., M.L.B., E.K., and B.E.G. supervised research. B.Z., I.E., B.W., S.C.S.S., T.T.S., H.G., S.B.M., K.T., and B.E.G. performed experiments. L.A.B. performed the computational data analysis. L.A.B., B.Z., B.E.G., E.K, and M.L.B. analyzed data. B.Z., S.J., and H.Z. provided analytic tools. L.A.B., B.Z., and B.E.G. prepared figures and/or tables. B.E.G., L.A.B., B.Z., E.K., and M.L.B. wrote the manuscript.

My contributions

- Implemented a computational pipeline to analyze raw ChIP-Seq data, which calculated various metrics to evaluate the quality of experiments and generated robust peak calls by comparing replicate experiments.
- Used this pipeline to aid in the screening and selection of antibodies and in the optimization of ChIP-Seq protocols. This process culminated with the first collection of high quality ChIP-Seq datasets that profiled all NF-κB subunits in same condition and cell type.
- Performed clustering of ChIP-Seq signals to discover that NF-κB proteins bind regulatory elements in a limited number of combinations, which we refer to as subunit binding patterns (SBPs).
- Employed a combination of *de novo* motif finding, motif enrichment analysis and ChIP-Seq data to show that specific SBPs were associated with indirect binding of NF-κB through four other TFs.
- Demonstrated the importance of the 3' flanking nucleotide of the traditional 10 bp kB motif in determining whether NF-κB dimers containing the p50 subunit were present or absent at a

particular genomic region. Showed that this hypothesis was supported by evolutionary conservation.

- Analyzed >50 ENCODE ChIP-Seq datasets in GM12878 cells to identify FOXM1 as a frequently co-localizing factor with NF-κB that lacked enrichment for its own motif.
- Participated in the design of validation experiments that showed FOXM1 was not binding its own motifs but was being recruited to DNA by a complex that contained NF-κB (*e.g.*, selected sequences and mutations to be used for EMSA probes)
- Performed survival analysis on expression datasets from diffuse large B-cell lymphoma patients to show that higher FOXM1 expression was associated with a worse prognosis.

In biology, nothing is clear, everything is too complicated, everything is a mess, and just when you think you understand something, you peel off a layer and find deeper complications beneath. Nature is anything but simple.

Richard Preston

4

Systematic characterization of coding variation in human transcription factors

Abstract

Regulatory variation is an important driver of phenotypic differences across individuals and species. However, identifying genetic variants that modulate gene expression in humans remains challenging, particularly for rare *trans*-acting alleles. We developed a computational, structure-based approach to identify coding variants that are likely to alter the DNA-binding preferences of human transcription factors. Using protein-binding microarrays, we assayed the DNA-binding properties of 115 transcription factor alleles found in large-scale population sequencing studies of healthy individuals and in families with Mendelian disease. We identified 74 missense variants that affect the DNA-binding affinity or specificity of transcription factors. By comparing the DNA-binding effects of non-synonymous SNPs and known disease-causing variants, we identified putative disease risk alleles and characterized the molecular perturbations underlying several Mendelian disease phenotypes. We estimate that thousands of rare alleles that alter DNA-binding affinity or specificity exist in human populations and that a typical individual carries several genetic variants that alter TF binding preferences.

4.1 Background

Mutations that alter gene expression have been established as key drivers of phenotypic variation across species and individuals. In particular, genetic variants in *cis*-regulatory elements have been shown to drive evolutionary adaptation in morphology, physiology and behavior[244]. Generally, *cis*-regulatory mutations are considered more likely to alter gene expression in a modular fashion, therefore reducing the potential for deleterious pleiotropic effects[244].

However, there has been considerable debate about the contribution of *trans* regulatory variation to evolutionary adaptation and phenotypic variation. A key unanswered question is the extent to which transcriptional networks can be rewired through amino acid substitutions in transcription factors without causing major detrimental effects on fitness[113]. There exist well-known examples where

amino acid substitutions that caused substantial changes in DNA-binding specificity became fixed in certain lineages. For example, a lysine to glutamine substitution in the Bcd homeodomain protein substantially altered its DNA-binding specificity in the Drosophila lineage[245]. In practice, such substitutions are likely to be facilitated by compensatory mechanisms, such as buffering by paralogous proteins[113]. However, whether these mechanisms are prevalent enough to allow TF alleles with alternative binding preferences to segregate in predominantly healthy populations remains an open question.

Variation in gene expression across human individuals has increasingly been associated with common disease risk. In particular, disease-associated variants have a tendency to alter gene expression in tissues or cell types that are relevant to the associated pathogenic process[116]. Therefore, being able to link changes in expression to the disease phenotype can provide key biological insights about the underlying etiology. In humans, both *cis* and *trans* mechanisms have been shown to contribute to gene expression variation[246]. However, finding genetic variants that affect expression in *trans* remains a significant challenge. In a typical *cis*-eQTL study, only variants that are proximal (usually < 1Mb) to a gene body are tested for associations with expression differences. When testing for *trans*-eQTL associations, this constraint is no longer applicable, greatly increasing the multiple-hypothesis testing burden and reducing statistical power to detect true associations.[110].

Several lines of evidence suggest that *trans* regulatory variation contributes to disease risk and is likely to be prevalent in human populations. A large *trans*-eQTL meta-analysis in whole blood discovered 233 significant SNPs by testing only for *trans* associations with SNPs previously implicated in common disease risk[110]. Furthermore, the effects on expression of variants associated with type 2 diabetes risk showed a prevalence of *trans*, but not *cis*, effects on expression[247]. These observations are further supported by reports of differential TF binding in the same locus across individuals. In a study comparing the binding patterns of NF-κB in 10 individuals, 7.5% of binding sites were found to vary across individuals[86]. Variable binding sites were preferentially associated with differential ex-

pression, but only ~1/3 of the binding variability across individuals could be explained by local genetic variation. These observations imply that there is a pressing need to evaluate the contribution of *trans* regulatory variants to human phenotypic variation. Although eQTL analyses have proven useful for this purpose, they remain limited to variants with sufficiently high allele frequencies. Furthermore, eQTL studies depend on the availability of the relevant tissue from a sufficiently large number of individuals.

An alternative is to use a genotype-first approach, in which genetic variants found in human populations are selected and then experimentally tested to determine their effects. Mutations that alter the coding sequence of TFs are prime candidates for having regulatory consequences. In particular, mutations that affect TFs' DNA-binding domains (DBDs) have the potential to cause changes in the expression of target genes by altering TF binding patterns at regulatory elements. Such missense variants have been extensively reported in families with Mendelian disease[248,249], but their prevalence in the population, their relative effect sizes, and their consequences remain largely unknown.

Various studies have reported that loss-of-function (LoF) variants are surprisingly prevalent in human genomes. Typically, these studies have been limited to frameshift or nonsense mutations, as these features are highly predictive of loss of protein function. LoF variants have been shown to cause substantial differences in transcript levels across individuals[250]. Non-synonymous SNPs (nsSNPs), variants which change the identity of a single amino acid in a protein, are significantly more common than frameshift or nonsense mutations. However, the functional consequences of nsSNPs are much harder to predict. Thus, the extent to which missense variants in TFs contribute to regulatory variation remains unclear.

Therefore, an essential task is to ascertain the prevalence and effects of nsSNPs in DNA-binding domains, hereafter referred to as DBD polymorphisms (DBDPs), in the human population. We describe a computational approach to analyze genotype data from exome and whole-genome sequencing projects to identify the variants most likely to affect DNA-binding. Using protein-binding microar-

rays, we compared the DNA-binding properties of reference and alternative TF alleles. In addition, we assayed a large set of TF alleles reported to cause various Mendelian diseases. Determining the binding preferences of Mendelian alleles revealed novel insights about the molecular basis underlying several phenotypes, particularly in cases where different mutations in the same TF are associated with distinct disease presentations. In addition, the experimental testing of disease-causing TF alleles enabled us to compare to the effect sizes of DBDPs in the population. Based on these observations, we estimate the prevalence of functional DBDPs in humans and discuss its implications for interpreting noncoding variation and association studies.

4.2 Results

4.2.1 Widespread variation in DNA-binding domains

The human genome encodes ~1400 sequence-specific TFs, which together harbor DBDs spanning >20 structural classes[24]. Multiple exome and whole-genome sequencing projects have now identified >3 million coding variants in tens of thousands of individuals. We first determined how many of these reported nsSNPs altered the amino acid sequences of DNA-binding domains in a set of 1,364 manually curated, high-confidence TFs[24]. Analyzing genotype data from predominantly healthy individuals surveyed by the 1000 Genomes Project, the Exome Sequencing Project (ESP6500) and the Exome Aggregation Consortium (ExAC), we identified 52,956 unique DBDPs in 64,706 individuals (Figure 4.1a). The majority of such variants are rare: beyond a set of 12,011 DBDPs discovered independently by at least two projects, 40,945 were reported in only one study. These results imply that many new DBDPs are likely to be identified as the number of sequenced exomes and genomes continues to grow. In addition, in the same set of individuals, we identified 4,533 nonsense variants that result in partial or full truncation of DBDs in their respective genes. These nonsense variants are likely to result in a loss of DNA-binding activity, whereas the consequences of DBDPs may span a range of

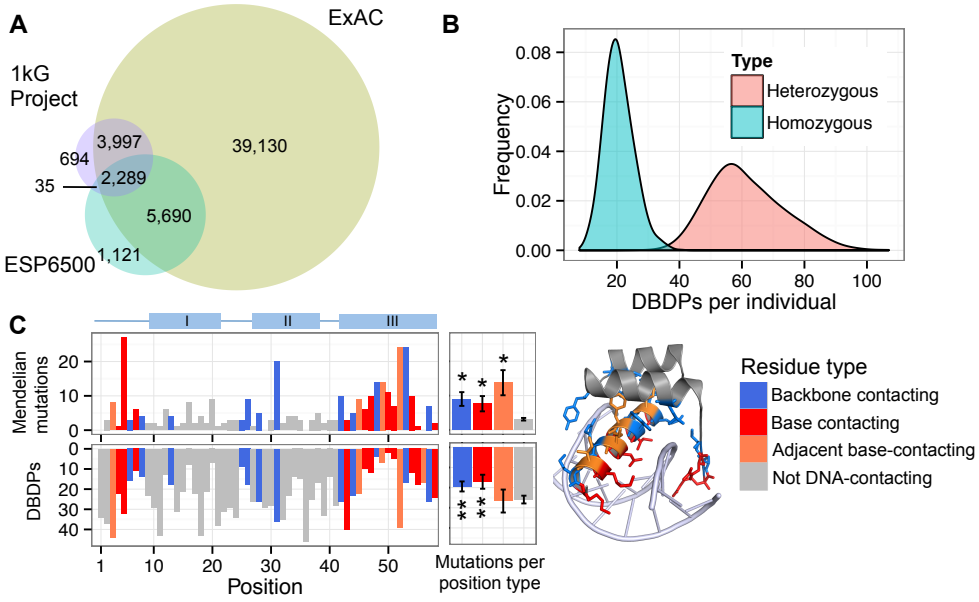


Figure 4.1: Patterns of variation in DNA-binding domains. (**A**) Number of unique DBDPs found in individuals genotyped by the 1000 Genomes Project (Phase III, 2,504 individuals), the Exome Sequencing Project (ESP6500, 6,503 individuals) and the Exome Aggregation Consortium (ExAC v0.2, 61,486 individuals). (**B**) Number of unique DBDPs found per individual in either homozygous or heterozygous states. (**C**) Number of Mendelian mutations (*upper left*) and nsSNPs (*lower left*) found in ExAC v0.2 across all homeodomain TFs as a function of their position in the domain and the type of DNA contact associated with residues in that position. ``I'', ``II'', ``III'' refer to α-helices, where helix III is the DNA recognition helix. Adjacent bar graphs (*middle*) depict the mean number of variants for each type of position; * or ** denote enrichment or depletion, respectively, relative to non-DNA-contacting residues ($P < 0.05$, permutation test), error bars (1 standard error of the mean). Representative co-crystal structure (*right*) of the Engrailed homeodomain (PDB: 1HDD) depicting annotation of DNA-contacting residues using the same scheme as on the left.

effect sizes and types.

As a complementary way of quantifying the prevalence of DBDPs, we determined how many DBDPs with potential effects on protein function are present in a typical individual's genome. We determined that a human genome harbors a median of 60 heterozygous and 20 homozygous DBDPs across the set of 1,364 TFs described above (Figure 4.1B). DBDPs are found at significantly reduced frequencies in TFs with known Mendelian phenotypes: a median of 6 heterozygous and 2 homozygous variants per individual, which corresponds to a significant depletion relative in the overall number of DBDPs relative to the 229 TF genes associated with Mendelian disease phenotypes (odds-ratio = 3.7

and P-value = 0.005, Fisher's exact test). This observation suggests that DBDPs are less likely to be tolerated in genes where mutations can have direct phenotypic consequences. These results suggest an abundance of TFs with potentially altered function per individual and a need for in-depth analysis of the functional consequences of these variants.

4.2.2 Enrichment of disease-causing mutations at protein-DNA interfaces

To better understand the potential consequences of DBDPs, we examined how frequently mutations associated with Mendelian disease affected different residues of DNA-binding domains. For this purpose, we focused on homeodomain proteins and the DBDPs and disease-causing variants that alter their amino acid sequences. Homeodomains are the second most common DBD class in humans[24] and are associated with a wide range of disease phenotypes[251]. In addition, the residues within the homeodomain that mediate protein-DNA recognition have been relatively well characterized[26,28]. This makes homeodomains an ideal structural class in which to study patterns of sequence variation as they relate to DNA-contacting residues.

We extracted data on DNA-contacting residues from publicly available homeodomain-DNA co-crystal structures in the Protein Data Bank (PDB) and aligned the corresponding homeodomain protein and bound DNA sequences to assemble a composite protein-DNA "contact map" for homeodomains. We then mapped disease associated mutations in homeodomains to the canonical amino acid numbering scheme for homeodomains. Residues that participated in base contacts, backbone contacts, or that were adjacent to residues that participate in base contacts, were significantly more likely to be reported as disease-causing than other homeodomain residues (Figure 4.1C, all P-values < 0.005, permutation test). In the exomes of healthy individuals, we observed a roughly inverse pattern, with fewer variants affecting backbone (P-value = 0.03121, permutation test) and base-contacting positions (P-value = 0.0134, permutation test), particularly in the recognition helix (III). However, the corresponding depletion in DBDPs was comparatively small (Figure 4.1C), suggesting that many DB-

DPs alter residues in the same homeodomain positions as those altered by Mendelian mutations in homologous proteins.

4.2.3 Prioritization of DBD variants for experimental testing

Based on the results obtained for homeodomains, we reasoned that structural information about protein-DNA contacts could be used to identify DBDPs with potential regulatory effects and undiscovered phenotypes. We devised a scheme to identify DBDPs of interest based on multiple criteria, such as (a) the position of the residue relative to the protein-DNA interface in co-crystal structures, (b) literature-derived annotations for specific DBD classes, (c) scores from several tools designed to predict the pathogenicity of mutations, (d) minor allele frequencies, and (e) known phenotypic associations, including those derived from linkage studies in families and through GWAS (Methods).

Using this prioritization scheme, we selected 35 DBDPs for experimental testing. Yet, hundreds of mutations in DBDs have already been linked to Mendelian phenotypes, largely through genetic linkage studies. We reasoned that when Mendelian variants are found in the same genes as DBDPs, side-by-side experimental testing would enable effect-size comparisons between pathogenic alleles and variants of unknown significance. Therefore, we identified a set of TFs where several Mendelian mutations altered the same DBD and selected a subset of those variants to assay experimentally. In some cases, different mutations in the same DBD had already been associated with distinct phenotypes. This suggested that experimental testing could provide further insights into the molecular perturbations associated with particular disease presentations. In other cases, we selected Mendelian disease mutations occurring in the same genes where DBDPs had been selected for experimental testing, which could enable comparisons of effect sizes and the identification of DBDPs with similar effects to known disease-causing mutations. In aggregate, we selected a set of 79 Mendelian disease alleles to assay experimentally.

In total, our set of selected proteins comprises allelic series corresponding to the DBDs of 41 TFs,

each of which includes the reference allele and at least one variant allele. The 114 variant DBD alleles span six major structural classes: ZF-C4 (also known as nuclear hormone receptor, NHR), ZF-C2H2, POU, PAX, homeodomain, and forkhead (Figure 4.1C). We focused our analysis primarily on ZF-C2H2 and homeodomain DBDs since those are the two largest structural classes of human TFs and their DNA recognition has been studied extensively[26,252]. For each of these 155 alleles, we created N-terminal glutathione S-transferase (GST) fusion constructs that included all DBDs encoded by each gene and their flanking sequences (Methods).

4.2.4 Characterization of variants using protein-binding microarrays

We employed universal protein-binding microarrays (PBMs) to comprehensively assay the DNA-binding preferences of the selected TF alleles. Briefly, a universal PBM is a double-stranded DNA microarray whose probe sequences have been designed to contain at least 32 independent instances of all DNA 8-mers. When a fluorescently-labeled TF binds the probes on the array, the resulting fluorescent signal intensities can be used to quantify the TF's relative preference for binding different sequences. The DNA-binding preference of a TF allele to a given DNA 8-mer is summarized by the E-score, a rank-based statistic that quantifies the extent to which the protein preferentially binds that 8-mer sequence relative to others[41,253]. E-scores are highly reproducible across replicate PBM experiments[41] and can be used to accurately predict *in vitro* and *in vivo* binding[39] . In addition, E-score comparisons have been used successfully to identify subtle differences in binding preferences across closely related TFs[26,253,254].

We used two distinct approaches based on E-score comparisons to identify affinity and specificity differences between TF alleles. We determined that affinity changes could be detected by comparing the distributions of top E-scores obtained from experiments for different alleles (Methods). We used a Mann-Whitney U-test on the top 50 E-scores for each allele and corrected the resulting P-values using the Benjamini-Hochberg procedure, identifying alleles with altered affinity as those having a q-value

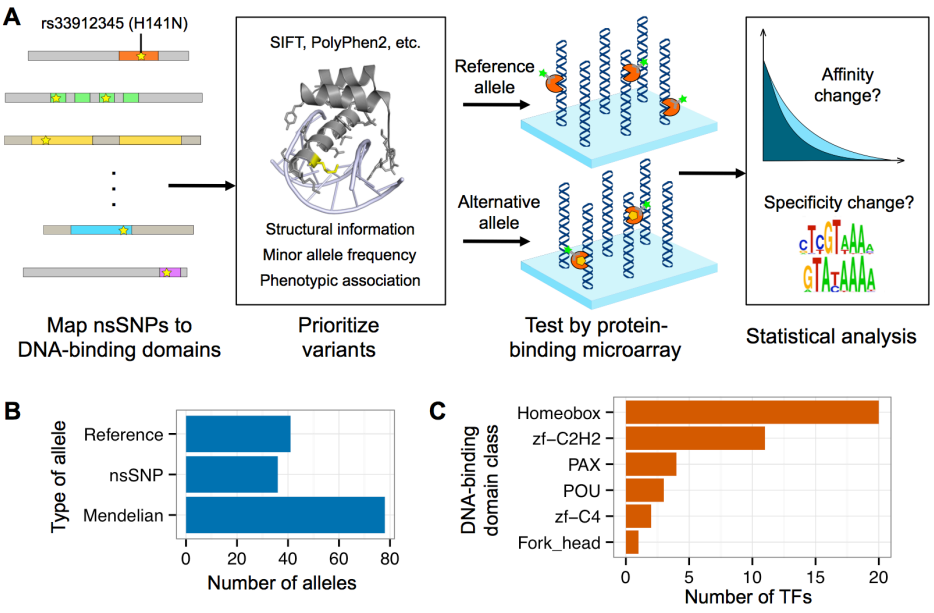


Figure 4.2: Experimental schema and study design. (**A**) Graphical summary of experimental schema. (**B**) Categories of alleles assayed by protein-binding microarrays. (**C**) DNA-binding domain structural classes corresponding to TFs selected for this study.

<0.05. This method was highly reproducible across replicate experiments, with 97% agreement in alleles identified as having altered specificity. In addition, the results obtained with this approach largely agreed with results previously derived by biochemical testing of the same Mendelian disease alleles using techniques such as electrophoretic mobility shift assays (Methods). Our PBM-based method achieved perfect specificity and 71% sensitivity for identifying affinity changes that were previously identified through low-throughput assays. The decrease in sensitivity was caused by not detecting relatively small changes in binding affinity for certain alleles, suggesting that our approach is conservative, but highly specific. To identify specificity changes, we used a previously described method for determining whether sets of 6-mers were preferentially bound by one TF over another[254]. Here, we employed this method to make pairwise comparisons between TF alleles. We identified alleles with 6-mers having a q-value < 0.05 relative to the reference allele and imposed additional filtering criteria to obtain results that were reproducible across replicates in 86% of cases (Methods).

We classified all the variant TF alleles based on whether they altered DNA-binding specificity, affinity, or both. Both prioritized DBDPs and Mendelian mutations were associated with a range of perturbations on DNA-binding, changing specificity, affinity, and sometimes both (Figure 4.3c). Although the frequency of DNA-binding changes was higher for Mendelian mutations than for DBDPs prioritized for their likely effects on binding (72% vs. 60%), this difference was not statistically significant (P-value = 0.32, Fisher's exact test). This observation implies that our approach for prioritizing DBDPs can identify variants with a high success rate. Intriguingly, ~28% of Mendelian mutations did not cause a detectable change in DNA-binding affinity or specificity, suggesting that, despite the pathogenicity of these mutations, the changes in DNA-binding they cause might be very subtle. TF alleles with Mendelian mutations typically lost their ability to bind a larger fraction of their binding sites than those with DBDPs (P-value = 0.0047, Wilcoxon rank-sum test), but there was no statistically significant difference in terms of the number of binding sites gained (P-value = 0.48, Wilcoxon rank-sum test) (Figure 4.3E).

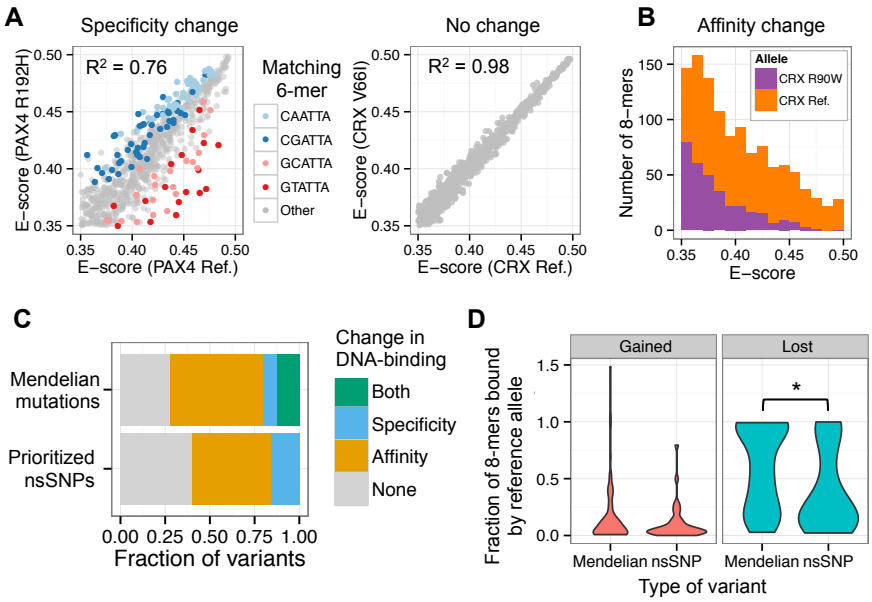


Figure 4.3: DNA-binding perturbations caused by DBD variants. (**A**) Comparison of the specificity changes observed in the R192H PAX4 allele (*left*) vs. the lack of specificity changes in the V66I CRX allele (*right*). Colored 6-mers were identified as being preferred either allele at q < 0.05 (Methods). (**B**) Example of histograms depicting top E-score distributions observed for a variant (CRX R90W) with altered DNA binding affinity as compared to the reference allele. (**C**) Bar graphs depicting the fraction of alleles with observed changes in DNA-binding affinity, specificity, neither, or both as determined from PBM 8-mer binding profiles. The DBDPs included in this figure were selected for this study as likely to change DNA-binding (*i.e.*, ``DNA-contacting'' or ``predicted damaging'' categories in Table S5) (**D**) Violin plots depicting the fraction of 8-mer binding sites gained or lost by variants relative to the number of 8-mers bound by the reference allele. Binding sites are considered gained or lost if an 8-mer has E $\geq$ 0.4 for one allele and E $<$ 0.4 for the other allele. * denotes P = 0.0047, Wilcoxon rank-sum test.

Given the heterogeneity of observed DNA-binding changes, we analyzed the effects of mutations in specific genes in more detail. PAX4 is a paired homeobox protein that is essential for the formation of beta-cells during pancreatic islet development[255]. Mutations in *PAX4* have been associated with both Mendelian and non-Mendelian forms of diabetes. We identified four *PAX4* nsSNPs in the population that were predicted as likely to have an effect on DNA binding. The R192H variant was reported as a risk factor for maturity onset diabetes of the young (MODY) and early onset type 2 diabetes (T2D) in the Thai population[256]. Meanwhile, the R133W allele had been described in association with autosomal recessive ketosis-prone diabetes (KPD) in West Africans[255].

Conversely, we identified two alleles (R192S and R183C) that have not been associated with a phenotype. R192S is particularly common in individuals of East Asian descent (ExAC MAF = 0.035), while R183C exists at low frequencies in both East Asian (1kG ASN MAF = 0.005) and African American populations (ESP AA MAF = 0.0002) populations. Hierarchical clustering of PBM binding profiles (Figure 4.4A) revealed that both mutations lead to DNA-binding perturbations that resemble those caused by the previously described mutant alleles. Both the R192S and R192H mutations altered the specificity of PAX4, while the R133W and R183C substitutions significantly changed its binding affinity, as defined by the criteria described above. Based on the similarity of their molecular phenotypes, we propose that the R192S and R183C are likely to have similar pathogenic potential for causing early onset T2D and KPD, respectively.

In other cases, PBM profiling confirmed that certain alleles are likely to be benign, but provided insights into the molecular basis for clinical heterogeneity of disease mutations in the same genes. CRX is a homeodomain protein that is essential for the proper function of photoreceptor cells[262]. Mutations in *CRX* have been linked to a range of Mendelian phenotypes, including retinitis pigmentosa (RP), cone-rod dystrophy 2 (CORD2) and Leber congenital amaurosis 7 (LCA7). While all three phenotypes involve retinal degeneration and ultimately loss of vision, they span a spectrum of severity: RP is the mildest and has the slowest progression, CORD2 is more severe and presents earlier than

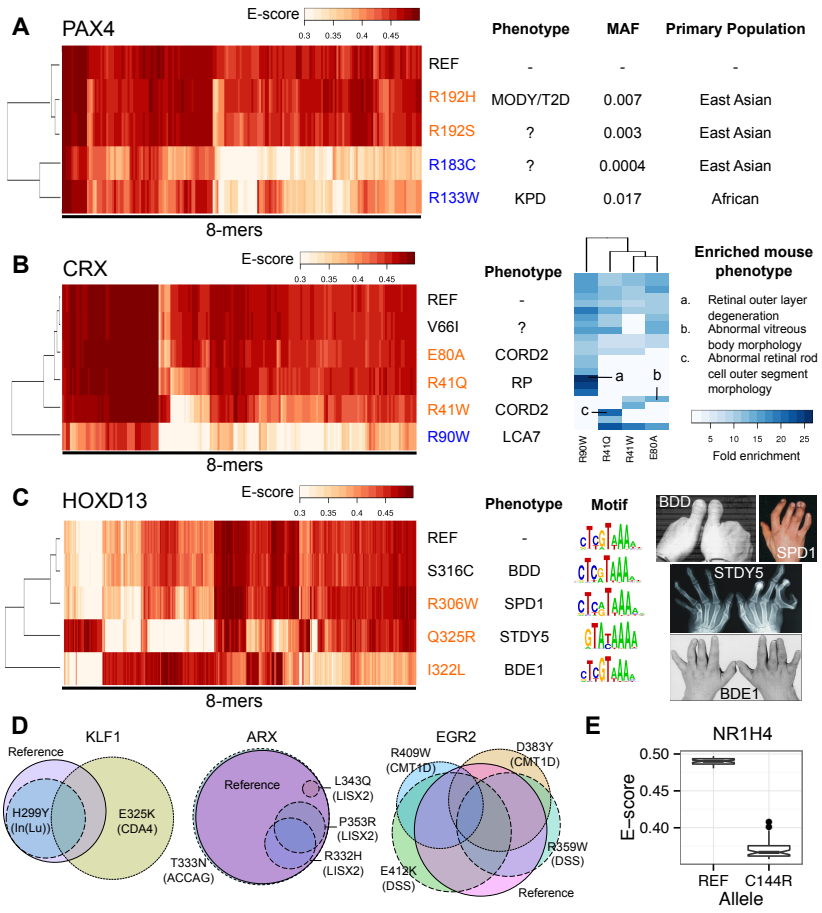


Figure 4.4: PBM profiling of 8-mer binding preferences in several allelic series. (**A-C**) Heatmaps (*left*) depicting PBM E-scores (columns) of different DBD alleles (rows) for all 8-mers bound strongly ($E > 0.45$) by at least one allele within each allelic series. Both rows and columns are clustered hierarchically. Variants indicated in blue or orange font to the right of the heatmaps exhibited altered DNA binding affinity or specificity, respectively. (A) Binding profiles of PAX4 allelic series with corresponding phenotypes (``?'' if no phenotype is known), ExAC minor allele frequencies (MAF) and population where each allele is most prevalent. (B) Binding profiles of CRX allelic series and corresponding phenotypes. Heatmap (*right*) depicting enriched mouse phenotype ontology terms ($Q < 0.05$, hypergeometric test) for genes near CRX binding sites in mouse retinal cells predicted to be most disrupted by each allele. (C) Binding profiles of HOXD13 allelic series with corresponding phenotypes and primary motifs derived using the Seed-and-Wobble algorithm [257] (*middle*). Images (*right*) depict typical manifestations of each phenotype. Images from [258--261]. (D) Proportional Venn diagrams depicting number of 8-mers shared across alleles within the KLF1, ARX, and EGR2 allelic series. Within each allelic series the size of the circles is proportional to the number of bound 8-mers ($E > 0.4$) by each allele. Reference alleles are indicated by purple circles with solid outlines. Disease associated with each variant is abbreviated in parentheses. (E) Comparison of top 50 E-scores between the NH1R4 reference allele and the C144R allele, which was identified as a LoF allele (P-value $< 2.2 \times 10^{-6}$, Mann-Whitney U test).

RP, and LCA7 involves severe loss of vision early in life[263–265].

We profiled CRX alleles reported in individuals with RP (R41Q), CORD2 (R41W and E80A) and LCA7 (R90W) and a DBDP predicted to be benign (V66I / rs61748438), which is found in all ExAC populations (overall MAF = 0.003). The V66I allele exhibited no significant change in DNA-binding affinity or specificity relative to the reference (Figure 4.3B and 4.4B). In contrast, the R90W allele lost the ability to bind the vast majority of 8-mers bound by other alleles, in accordance with its associated phenotypic severity. Meanwhile, the R41Q, R41W and E80A alleles showed a change in binding specificity, but not affinity. These observations are consistent with differences in mode of inheritance: the R90W allele was associated with an autosomal recessive inheritance pattern, whereas the R41Q and E80A mutations were autosomal dominant[265]. To determine whether these changes could help explain the associated phenotypes, we used the PBM data to predict which *in vivo* binding sites were more likely to be disrupted in CRX ChIP-Seq data derived from mouse retinal cells[266] (Methods). We determined if certain phenotypes were associated with genes in proximity to the 100 *in vivo* binding sites predicted to have the largest change in binding based on PBM data (Q <0.05, hypergeometric test). We found that the predicted affected categories clustered according to their phenotype, supporting the idea that the organismal phenotypes are related to the underlying perturbations in DNA binding *in vivo*. Furthermore, the enriched categories were consistent with murine ocular phenotypes: the "abnormal retinal rod cell outer segment morphology" category, enriched for the R41Q allele, was consistent with the lack of outer rod segments observed in a mouse model of RP[267].

In contrast, Mendelian mutations in other TFs were associated with more complex patterns of binding site gains and losses. HOXD13 is a homeodomain TF with key roles in limb development[268]. Frameshift mutations and poly-alanine expansions in *HOXD13* have generally been associated with the presence of extra digits and increased webbing between them (synpolydactyly), while missense mutations have been linked to shortening of fingers and toes (brachydactyly)[269]. These observations suggest that loss-of-function mutations are the likely cause of synpolydactyly symptoms, while gain-

of-function mutations are likely to be responsible for the features observed in patients with brachydactyly[115].

We assayed four HOXD13 alleles with distinct phenotypic associations: R306W (Synpolydactyly 1), S316C (Brachydactyly D), I322L (Brachydactyly E1), and Q325R (Syndactyly 5). Surprisingly, the DNA-binding profiles for each allele did not cluster according to phenotype (Figure 4.4C). These results suggest that the consequences of DNA-binding alterations on the occupancy of specific binding sites, rather than a broad gain-of-function vs. loss-of-function classification, is likely to underlie the differences in phenotypic effects of *HOXD13* mutations. This hypothesis is supported by evidence that other *HOXD13* mutations associated with Mendelian phenotypes uniquely alter HOXD13's *in vivo* binding patterns in mesenchymal stem cells[115].

Overall, different TFs exhibited unique mutational landscapes, with varying degrees to which the binding alterations caused by mutations could be linked to specific phenotypes. We compared the fractional overlap between 8-mers bound at high affinity (E-score > 0.45) by different alleles of the same TF and depicted the results as proportional Venn diagrams (Figure 4.4D). The measured binding profiles for KLF1 mutants were consistent with a loss-of-function (H299Y) vs. gain-of-function (E325K) dichotomy, as previously reported[270,271]. Similarly, the three mutations in ARX associated with LISX2 (L343Q, P353R, R332H)[272,273] were identifiable as loss-of-function mutations, whereas the T333N mutation, associated with ACCAG[272], involved a subtle change in specificity. In contrast, all four mutations in EGR2 caused unique changes in binding specificity, with only partial overlap between the sites bound by alleles linked to the same phenotype.

PBM profiling also identified DBDPs that cause a loss of detectable DNA-binding by HOXB7, NR1H4, PHOX2B, VENTX and ZNF200. For example, the C144R mutation in the nuclear hormone receptor NR1H4 alters a zinc-coordinating residue essential for maintaining the proper fold of the C4-ZF DBD, likely creating an unstable protein that lacks the ability to bind DNA specifically (Figure 4.4E). Finding mutations that abrogate DNA-binding suggests that individuals are able to

tolerate a heterozygous loss-of-function state in these genes without major phenotypic consequences. Out of these genes, only HOXB7 has been previously reported as being LoF tolerant[274], highlighting the value of PBMs as a tool to identify additional genes that do not exhibit haploinsufficiency.

4.2.5 Computational prediction of variants with DNA-binding effects

The number of DBDPs discovered in humans is likely to grow at a rate that exceeds the capacity to test individual variants experimentally. Therefore, efforts to understand the functional effects of DBDPs would be greatly aided by being able to predict whether a particular DBDP is likely to have an effect on DNA-binding. Using our experimental results as a benchmarking set, we evaluated whether information about DNA-contacting residues would be valuable for predicting whether a mutation would cause DNA-binding alterations. We observed that mutations that affected residues in the protein-DNA interface were significantly more likely to cause changes in binding affinity or specificity (odd-ratio = 3.3, P-value = 0.014, Fisher's exact test; Figure 4.5A), suggesting that information about DNA-contacting residues has predictive power for identifying changes in binding.

We explored whether information about protein-DNA contacts and the output from tools used to predict the consequences of mutations (*e.g.*, Polyphen2, SIFT) could be used to distinguish variants that caused DNA-binding changes and those that did not. We calculated the precision and sensitivity for predicting that amino acid substitutions in DNA-contacting residues (base and backbone) would alter DNA-binding (Figure 4.5B). We compared these values to those obtained using PolyPhen2 and SIFT and predicting that variants identified as "probably damaging" and "damaging" (respectively) would alter DNA-binding (Figure 4.5B). Both approaches performed similarly, with predictions based on DNA-contacting residues exhibiting higher sensitivity (0.79 vs. 0.71) but slightly reduced precision (0.76 vs. 0.79). Predicting changes in binding by the combination of the two methods (*i.e.*, considering only variants predicted to change binding by both) achieved to similar performance (Figure 4.5B).

Based on these results, we sought to estimate the prevalence of DBD variants that are likely to alter

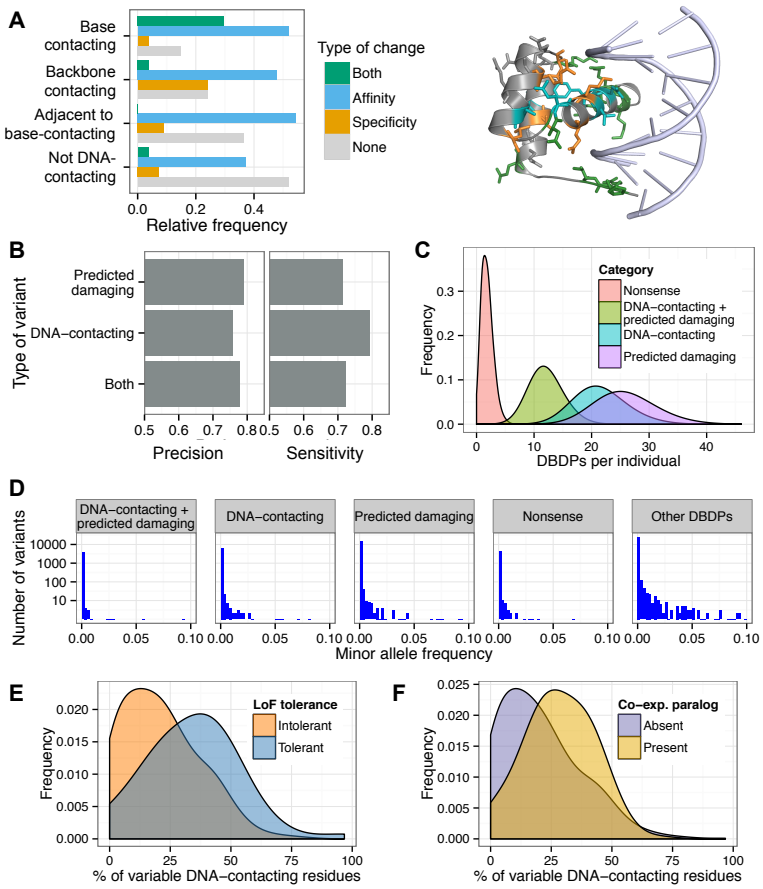


Figure 4.5: Functional associations of DBDPs predicted to have altered DNA-binding.(**A**) Relative frequency of DNA-binding changes observed for variants at different categories of DNA-contacting residues (*left*). Representative co-crystal structure (*right*) of Engrailed homeodomain (PDB: 1HDD) depicting residues color-coded according to the changes in DNA-binding specificity, affinity or both caused by variation in those positions. The ``both'' category in the co-crystal structure includes residues at which variants changed both DNA binding affinity and specificity either simultaneously in one protein or separately across multiple different proteins. Side chains are shown for residues that were tested experimentally by PBMs. (**B**) Precision and sensitivity metrics for identifying variants that alter DNA-binding affinity or specificity. The ``predicted damaging'' category includes variants that were predicted as damaging by both PolyPhen2 and SIFT. The ``DNA-contacting'' category includes base- and backbone-contacting residues. (**C**) Number of DBD variants per individual (1000 Genomes Project Phase III), including the categories of DBDPs depicted in panel B and DBD-truncating nonsense variants. (**D**) Minor allele frequencies (ExAC v0.2) of DBD variants in the same categories from panel C. (**E**) Comparison of the percentage of DNA-contacting amino acids (base or backbone) per gene altered by at least one nsSNP, for genes that have been found to be tolerant of heterozygous LoF mutations and for genes for which LoF tolerance has not been observed [274] (P-value = 6.519×10^{-8}, permutation test). (**F**) Equivalent to panel E, but comparing the percentage of variable DNA-contacting residues at TF genes with at least one co-expressed paralog and those without a co-expressed paralog (Methods) (P-value = 5.674×10^{-8}).

DNA-binding. We calculated the number of DBDPs per diploid genome predicted to be damaging and/or affect DNA-contacting residues. Different individuals harbored a broad range of DBD variation (Figure 4.5C), with a median of 16 DBDPs in DNA-contacting residues and 21 that are predicted as damaging by both PolyPhen2 and SIFT, with 9 predicted in common. In addition, a median of 2 DBD-truncating nonsense variants were observed per genome. Approximately 3/4 of the variants identified by each of these approaches are found in a heterozygous state. The variants in DNA-contacting residues and predicted as damaging by PolyPhen2 and SIFT were only partially overlapping. Given the similar predictive performance of either approach (Figure 4.5B), these results suggest that tools such as PolyPhen2 and SIFT are likely to correctly identify additional structurally damaging DBDPs, but may be too conservative in predicting whether mutations in DNA-contacting residues will alter binding.

Variants in each of the categories described above had significantly lower MAFs (all P-values < 0.05, permutation test; based on ExAC v0.2 MAFs) than DBDPs that were not included in any category (Figure 4.5D), suggesting they are more likely to alter TF function and potentially have deleterious effects. We investigated whether certain features of TFs harboring DBDPs predicted to alter DNA-binding could explain their tolerance for damaging variants. Because the number of DNA-contacting residues can vary significantly across DBD structural classes, we computed the fraction of DNA-contacting residues per TF that were altered by any of the 52,956 DBDPs we identified (Figure 4.1A). Intriguingly, TF genes reported to tolerate homozygous LoF mutations[274] had a significantly higher fraction of variable DNA-contacting residues (P-value = 6.519×10^{-8}, permutation test) (Figure 4.1D). These results suggest LoF-tolerant TFs are under reduced selection to maintain their DNA-binding preferences and may thus be more likely to tolerate DBDPs that alter their genomic binding. We observed a similar pattern when comparing the fraction of variable DNA-contacting residues in TFs that have a co-expressed paralog and those that do not (Figure 4.1F), with the former having a higher fraction of variable residues (P-value = 5.674×10^{-8}, permutation test). These results suggest

that the presence of co-expressed paralogous TFs may also reduce the selective pressure for TFs to maintain their binding preferences, potentially allowing greater regulatory diversity. These two enrichments were found to be independently significant through a generalized linear model considering both LoF tolerance and the presence of a co-expressed paralog ($P < 0.005$, t-test on regression coefficients). Additional compensation could arise from higher expression levels of the wildtype TF allele: epistatic selection between deleterious coding variants and *cis*-regulatory variants has been suggested in a study of eQTLs and allele-specific expression [275].

4.3 Discussion

Here, we described an integrative experimental and computational approach that has enabled the discovery of genetic variants that alter the DNA-binding preferences of human TFs. These results support the existence of a complex landscape of DBD variation, where variants that cause different types of DNA-binding alterations across a broad range of effect sizes are found in many TFs. The degree to which DBDPs with predicted effects are tolerated in human genomes is surprising, particularly when considering the evolutionary arguments for the likely deleteriousness of *trans*-regulatory variants. However, the unexpected prevalence of DBDPs parallels the discovery of LoF variants in many human genes [114,274]. These observations suggest that transcriptional networks in human tissues are sometimes robust to genetic perturbations. This robustness may be accomplished through diverse mechanisms, such as buffering of expression changes by paralogous TFs or by ability of transcriptional networks to tolerate losses of TF function. These compensatory mechanisms may allow a TF allele with altered binding preferences to adopt new target genes without widespread misregulation.

In many cases, single-amino acid changes in DBDs were associated with subtle changes in DNA-binding preferences. The ability of coding variants to selectively re-wire transcriptional networks represents another mechanism through which *trans*-regulatory variation may be tolerated in the population. Intriguingly, many Mendelian mutations caused only subtle changes in DNA-binding pref-

erences, suggesting that their phenotypic effects may be associated with specific gains and losses of binding sites. These results are consistent with recent studies showing that a significant fraction of disease-causing mutations selectively disrupt protein-protein or TF-enhancer interactions[276,277]. If DBDPs are generally able to modulate gene expression while causing minimal pleiotropic effects, they are likely to have more potential for causing phenotypic diversity than has commonly been assumed.

PBM profiling enables the comprehensive measurement of the DNA-binding preferences of TFs *in vitro*. This approach permits the screening of hundreds of alleles and the identification of DNA-binding changes of diverse effect sizes. We have showed how PBM measurements can be used directly to identify putative disease risk variants by leveraging the results from association studies in one population and finding variants in other populations that cause similar DNA-binding changes. There are hundreds of additional DBD variants occurring in genes with known phenotypic effects that were not tested in this study. Characterizing additional DBDPs that affect the same genes is likely to reveal many additional variants with potential phenotypic consequences. In addition, PBM data provide detailed insights into the molecular mechanisms underlying Mendelian phenotypes caused by mutations in TFs. These results imply significant heterogeneity in terms of the relationship between DNA-binding effects and observed phenotypes. It is likely that the exact perturbations caused by mutations, the genetic background of the individual, and variable expressivity play unique roles in explaining the phenotypic variability associated with Mendelian mutations across different TF genes.

However, this approach is likely to underestimate the prevalence of variants with *trans*-regulatory potential. Coding variants affect TF sequences in regions besides DBDs, such as trans-activating, trans-repressive, and dimerization domains. Such variants could have regulatory consequences that are similar to those of DBDPs. For example, a variant that lowers trans-activation efficiency may have comparable *in vivo* effects to a DBDP that reduces binding affinity of an activator TF. In addition, even Mendelian mutations in DBDs can have phenotypic effects without causing changes in DNA-binding. For instance, Mendelian mutations in the HESX1 homeodomain have been shown to im-

pair its dimerization ability without changing its monomeric binding preferences[278]. In cases where Mendelian disease mutations had no detectable effect on DNA-binding, the phenotypic consequences may be caused by similar perturbations or by affinity or specificity changes that are too subtle to detect with PBMs. In addition, some of the mutants without detectable changes may be incorrectly associated with disease. Further work will be needed to understand the relative contribution of non-DBD variants and the diversity of mechanisms through which DBD mutations can have phenotypic effects without directly altering DNA-binding.

Nevertheless, the abundance of DBDPs with predicted binding effects has significant implications for the interpretation of noncoding variation. Our results imply that unrelated individuals typically have a unique repertoire of TF alleles with altered binding preferences. These data also support the increasingly accepted hypothesis that human genomes harbor a significant number of loss-of-function alleles[114,274], but provide additional evidence that this is also the case for TFs. Therefore, different individuals are likely to have unique *trans*-regulatory landscapes, where certain TFs have altered binding preferences, some are present at lower levels and some are absent altogether. Understanding how DBD variants interact with *cis*-regulatory variants is likely to be essential in developing a more complete picture of regulatory variation in humans and to determine if such genetic interactions may play a role in explaining missing heritability.

So far, most regulatory variants have been identified through eQTL mapping. However, DBDPs are typically rare and likely to act in *trans*, which are both circumstances that significantly reduce the statistical power of eQTL analysis. Characterizing the effects of rare DBDPs on expression is likely to require novel approaches, including the functional testing of alternative alleles *in vivo* or the stratification of individuals to include specific variants of interest in eQTL studies. The methods described in this study represent an important step towards the development of approaches that account for the prevalence of coding variation in human TFs. These insights are likely to be essential in developing a more complete understanding of the molecular basis of regulatory variation and its role in

determining phenotypic differences across individuals.

4.4 Methods

Identification of DNA-binding domains of sequence-specific transcription factors

We identified genes for sequence-specific TFs based on a previously published, manually curated census of human TFs[24]. Only TF genes from the two highest confidence categories, requiring direct functional evidence or the presence of domains never found in non-TF genes (encompassing 1,364 genes), were considered in this study. Protein sequences for the selected TF genes were retrieved from the Ensembl database (version 67)[279]. To identify matches to DNA-binding domains (DBDs), we retrieved hidden markov models (HMMs) from the Pfam database corresponding to DBD structural classes that have been identified in human genomes[280]. The hmmscan tool, which is part of the HMMER 3.0 package[281], was used to scan human protein sequences for DBD instances. We used the default hmmscan parameters, except for a more stringent domain match threshold (E-value < 0.0001).

We used variant annotations obtained from dbNSFP v2.0b[282] to link nucleotide changes to amino acid substitutions and identify nsSNPs. For each SNP, its effect on all overlapping Ensembl transcript models was considered. DBDPs were identified as missense SNPs that altered the sequences of the DBDs identified by HMMER, as described above. In rare instances, DBD matches differed across transcripts due to alternative splicing. In such cases, the transcript with the best match score (lowest E-value) to the Pfam HMM was selected to represent the DBD for that gene.

Sources of SNPs and disease mutations

The nsSNPs selected for experimental testing were drawn from either the 1000 Genomes Project Phase II release[11] (1700 individuals) or the Exome Sequencing Project 6500 release (February, 2013)[283]. A

later release of the 1000 Genomes Project data (Phase III, September 2014 release, with 2,504 individuals) was used for statistical analyses. Variants from the Exome Aggregation Consortium (ExAC) v0.2 release were also used for statistical analyses and determination of allele frequencies, but were not considered for experimental testing. For analyses involving the number of variants found per individual, only the 1000 Genomes Project Phase III data were used because other datasets did not provide full genotype data. In all cases, only variants that passed the most stringent level of quality control filters ("PASS" value in the VCF file), were used for statistical analyses or selected for experimental testing. Mendelian variants were retrieved from the curated set of Online Mendelian Inheritance in Man (OMIM) mutations in the UniProtKB[284] (release 2013_05) database. For all genes, the coordinates of amino acid substitutions were mapped to the canonical splice isoform selected by UniProtKB. The domain position affected by mutations was determined from the optimal alignment between the Pfam HMM and the protein sequence, as determined by the hmmscan tool in the HMMER 3.0 package[281].

Annotation of DNA-contacting residues in select Pfam domains

Four DBD structural classes were selected for detailed annotation of residues likely to engage in DNA contacts: C_2H_2 zinc-fingers (Pfam: zf-C_2H_2), homeodomains (Pfam: Homeobox), forkhead (Pfam: Fork_head), and basic helix-loop-helix domains (Pfam: HLH). These classes were prioritized based on their occurrences in significant numbers of human TFs and the availability of prior knowledge about the amino acid residues that are involved in DNA-contacts. For all classes except homeodomains, backbone- and base-contacting domain positions were identified based on published studies[27,29,285–291]. For each class, the positions of amino acids that had been described explicitly as base- or backbone- contacting in the literature were manually linked to the corresponding positions in the Pfam domain. If a residue at a given position in the domain was reported as making both base and backbone contacts, it was annotated as base-contacting. Residues at positions adjacent to base-contacting residues that were

not identified as making backbone contacts were annotated as "adjacent to a base-contacting residue."

In the case of homeodomains, we analyzed structural data to comprehensively identify residues that may play a role in protein-DNA contacts. Ten homeodomain co-crystal structures (PDB IDs: 1IG7, 2H1K, 3LNQ, 3HDD, 9ANT, 1JGG, 1DU0, 2HDD, 2HOS, and 1APL) were chosen to sample a wide range of sequence diversity within the homeodomain family while excluding complexes that exhibited cooperative dimerization or included co-factors. When multiple identical proteins were contained within the same unit cell, a single instance was selected for analysis. Coordinates were extracted from PDB files using the "pdbread" function from the MATLAB Bioinformatics Toolbox, which was also used to calculate distances between amino acid residues and non-hydrogen atoms in DNA. We separately considered contacts between amino acid residues and DNA bases and amino acid residues and the backbone. The minimal distance between amino acid residues and DNA was used to define contact strength: contacts within 3.5 Å were assigned a score of 2, while contacts between between 3.5 Å and 5 Å were assigned a score of 1. Contact maps for separate proteins were aligned using ClustalW v2.1 with default settings to perform a multiple sequence alignment of the corresponding protein sequences. DNA sequences were aligned by visual inspection. For each position in the domain and each position in the binding site, we calculated the mean contact score over all structures, creating an average contact map that summarizes the likelihood that a residue participates in DNA contacts. The average score obtained for each domain position was used for subsequent prioritization, as described below. All homeodomain positions with non-zero average scores for backbone or base contacts were annotated as putatively DNA-contacting.

Prioritization of variants for experimental testing

We used several criteria to filter DBDPs found in population sequencing studies and identify variants that were likely to alter DNA-binding. These criteria can be summarized as (a) the prevalence of the variant in the population, (b) inferred proximity of affected residues to DNA based on structural

data, (c) deleteriousness of the mutation as predicted by published tools, and (d) known phenotypic associations of the affected gene, and are described in more detail below. To minimize the number of selected variants that may be due to sequencing errors, variants found in heterozygous form in only one individual were excluded. Otherwise, DBDPs that had certain combinations of features of interest were manually evaluated and curated. Ideally, we sought to find variants that were present in many individuals, affected genes with known phenotypes, and were predicted to have a significant potential to alter DNA-binding properties or disrupt protein stability. In practice, variants were considered for testing if they met the criteria for at least two categories. If multiple DBDPs were found in the same gene, variants that met just one criterion were sometimes tested alongside variants that met multiple criteria to allow comparisons of effect sizes between prioritized and non-prioritized variants. Similarly, we selected a few variants that were not predicted to alter DNA-binding but occurred in genes for which Mendelian disease mutations had been chosen for experimental testing.

Structural information was used to prioritize variants by determining if the affected residue was in an annotated DNA-contacting position. For the four DBD classes for which Pfam domains were annotated, the per-position annotations were used to evaluate whether residue changes were likely to affect protein-DNA contacts. This was done by finding the optimal match to the Pfam HMM for a given protein sequence and determining if the domain position in which the amino acid substitution occurred was annotated as DNA-contacting. A small subset of 8 DBDPs was prioritized by manual evaluation of the consequences of the amino acid substitution on homologous co-crystal structures.

Several tools designed to predict whether coding mutations are likely to be biochemically damaging were used to aid in the prioritization of variants: SIFT[292], PolyPhen2[293], LRT[294], MutationTaster[295], and MutationAssessor[296]. Studies comparing the agreement between predictions made by different tools have reported significant discrepancies, but have also shown that combining predictions from different tools improves overall accuracy[297]. Based on these observations, DBDPs that were predicted to be damaging by at least three of the five tools were assigned the highest priority. However, variants

were also considered in cases where at least one predictor tool considered the variant as damaging and the effect of the substitution was deemed of high likelihood to impact DNA-binding through the methods described above.

In addition to residue-specific considerations, we integrated information about gene-level phenotypic associations into our prioritization scheme. DBDPs affecting genes with at least one associated OMIM code, as annotated in UniProt KB[284], were assigned a higher priority, as these variants are *a priori* more likely to have phenotypic consequences. We also considered whether genes harboring DBDPs were associated with variants found in genome-wide association studies (GWAS) in the NHGRI GWAS catalog[7]. DBDPs in genes that were directly reported in association to traits (*i.e.*, in the "Reported Gene(s)" column in the GWAS catalog) were given higher priority. In addition, we considered whether DBDPs were in linkage disequilibrium (LD) with GWAS tag SNPs. We retrieved LD tables derived from the AFR (African), AMR (Admixed American), EUR (European) and ASN (Asian) populations in the 1000 Genomes Phase I data from the HaploReg tool[298]. DBDPs in LD with GWAS SNPs from the NHGRI catalog at a threshold of $R^2 > 0.5$ in any population were assigned a high priority for experimental testing.

Finally, we selected a set of DBDPs that were considered as unlikely to affect DNA-binding but were deemed to be interesting for other reasons. These included DBDPs that occurred in genes that were being assayed for the effect of other variants or that occurred at high minor allele frequencies in genes that had known Mendelian phenotypes.

We selected Mendelian disease mutations under two general categories: (a) mutations affecting genes for which DBDPs were prioritized for experimental testing, (b) mutations occurring in genes in which several Mendelian disease variants were known to affect the same DNA-binding domain. Whenever a gene harbored a DBDP that was prioritized for experimental testing and the same gene had known Mendelian disease mutations, at least one mutation was selected for experimental testing. Mendelian disease mutations were also chosen for testing in cases where different mutations within

the DBD were associated with distinct OMIM codes (*i.e.*, phenotypes), particularly when certain mutations affected DNA-contacting residues.

4.4.1 Selection of TF subsequences for cloning

We identified TF amino acid sequences corresponding to the DBDs, as defined by Pfam HMM matches, plus 15 amino acid (a.a.) flanks extending towards both the N-terminal and C-terminal ends. Previous studies have successfully used GST-tagged constructs comprising the DBD and 15 a.a. flanks in PBM experiments[26,253]. Here, we employed the same strategy. In cases where multiple DBDs were present in the same protein (*e.g.*, PAX TFs, or proteins with multiple C2H2 zinc-finger domains), we created constructs that encompassed all DBDs plus 15 flanking amino acids of the DBDs located closest to the protein termini.

4.4.2 Generation of TF Entry clones and LR transfer into pDEST15 vector

Entry clones carrying the selected TF subsequences were generated by PCR-based Gateway recombinational cloning. For PCR amplification, all the forward and reverse primers contained attB1 and attB2 sites, respectively, at their 5' ends. PCR reactions were performed using KOD Hot Start DNA polymerase according to the manufacturer (Novagen), and using TF reference clones from human ORFeome version 7.1 (http://horfdb.dfci.harvard.edu/hv7/) as template. The resulting PCR products were then cloned into pDONR223 vector by Gateway BP reactions, yielding desired TF Entry clones. After bacterial transformation, miniprep plasmid DNA of all Entry clones was extracted, and then transferred individually by in vitro Gateway LR cloning into pDEST15 expression vector, deriving N-terminal GST-tagged TF fusions. All these expression clones were sequence-verified in two directions using universal primers pGEXfw and T7-Terminator, and no mutations were found. The primer sequences are as follows:

- pGEXfw: 5'-GGCAAGCCACGTTTGGTG-3'

- T7-Terminator: 5'-GCTAGTTATTGCTCAGCG-3'

4.4.3 Generation of mutant clones

To generate mutant TF clones, we used an enhanced, two-stage, site-directed mutagenesis pipeline[277]. Briefly, for a given TF mutation, the mutagenesis platform consisted of two "primary PCRs" to generate TF fragments, and one "fusion PCR" to obtain the mutated TF. For the primary PCRs, vector-specific universal primers were used in combination with the respective two TF-specific internal forward and reverse primers to generate overlapping fragments containing the desired nucleotide substitution. The universal primers allowed the Gateway recombination sites to be preserved on both ends of the TFs. The mutation-specific primers, MutF and MutR, harboring the desired nucleotide changes, were designed to be complementary to each other. Site-directed PCRs were performed on either TF domains already cloned into the Destination vector pDEST15 or on TF domain Entry clones in pDONR223. For TF domains in pDEST15, the two TF fragments flanking a given mutation were amplified using the primer pair Tag1-pGEXfw and MutR, and the primer pair Tag2-T7-Term and MutF, respectively. In the subsequent fusion PCR, the two primary fragments were fused together using the primer pair Tag1 and Tag2 to generate the mutated TFs, and the mutant TF PCR products were then introduced into pDONR223 by a BP reaction followed by bacterial transformation. For TF domains in pDONR223, the two TF fragments flanking a given mutation were amplified using the primer pair M13G-FOR and MutR, and the primer pair M13G-REV and MutF, respectively. In the subsequent fusion PCR, the two primary fragments were fused together using the primer pair M13G-FOR and M13G-REV to generate the mutated TFs, and the mutant TF PCR products were then introduced into pDEST15 by an LR reaction followed by bacterial transformation. At least two independent colonies per mutant TF were isolated. Following sequence confirmation by Sanger sequencing, the clones that had only the desired mutations (no additional mutations) were selected and consolidated. Mutant TFs in pDONR223 were transferred to pDEST15 by Gateway LR reactions.

Primer sequences used are as follows:

- M13G-FOR: 5'-CCCAGTCACGACGTTGTAAAACG-3'
- M13G-REV: 5'-GTGTCTCAAAATCTCTGATGTTAC-3'
- Tag1-pGEXfw: 5'-GGCAGACGTGCCTCACTACTGGCAAGCCACGTTTGGTG-3'
- Tag2-T7-Term: 5'-CTGAGCTTGACGCATTGCTAGCTAGTTATTGCTCAGCG-3'
- Tag1: 5'-GGCAGACGTGCCTCACTACT-3'
- Tag2: 5'-CTGAGCTTGACGCATTGCTA-3'

4.4.4 Protein expression and quantification

In vitro transcription and translation (IVT) reaction were performed according to the manufacturer's protocol (NEB PURExpress IVT Kit). Western blots were used to estimate molar concentrations of all in vitro translated proteins by utilizing a dilution series of recombinant GST (Sigma) essentially as described previously[26]. Equal volumes of IVT samples and known concentrations of GST were suspended in 4x XT Sample Buffer (BioRad), heated to 95 °C for 5 minutes, and loaded on a precast 4-12% Bis-Tris Criterion gel (Bio-Rad). Samples were subject to electrophoresis at 190 V for 35 minutes and then transferred to a nitrocellulose membrane (Sigma) at 100-115 mA for 2 hours. Membranes were visualized using the SuperSignal West Femto Maximum Sensitivity Substrate kit (Pierce) according to the manufacturer's protocols. Primary antibody was added to achieve a final concentration of 20 ng/ml (rabbit anti-GST antibody; Sigma cat #097K4767). Secondary antibody was added at a final concentration of 5 ng/ml (goat anti-rabbit secondary Ab; ThermoScientific #31460). Films were scanned and concentrations of full-length proteins were determined using Quantity One software version 4.5.0 (BioRad), in accordance with the GST standard curve. All reference and alternative allele proteins were expressed in the same IVT batch.

4.4.5 Protein-binding microarray experiments

Oligonucleotide arrays were double-stranded and PBM experiments were performed following previously described experimental protocols [41,257]. The array design employed was an "all 10-mer" universal array in 8 x 60K format (Agilent Technologies; AMADID #030236). To minimize potential batch effects, reference and mutant alleles for the same TF were assayed on separate chambers in the same PBM slide. All experiments comparing reference and alternative alleles used proteins expressed in the same batch and diluted to achieve equal TF concentrations across an allelic series.

4.4.6 Protein-binding microarray data processing

PBM scan images were obtained using a GenePix 4000A Microarray Scanner (Molecular Devices). The resulting image data were processed using GenePix Pro v7.2 to obtain signal intensity data for each spot. The data were then further processed by using Masliner software (v1.02) [299] to combine scans from different intensity settings, increasing the effective dynamic range of the signal intensity values. If a dataset had any negative background-subtracted intensity (BSI) values (which can occur if the region surrounding a spot is brighter than the spot itself), consistent pseudocounts were added to all BSI values such that they all became nonnegative. All BSI values were normalized using the software for spatial de-trending providing in the Universal PBM Analysis Suite [41], as previously described [41,257].

4.4.7 PBM-based evaluation of DNA-binding changes

For each PBM experiment, we used the Seed-and-Wobble algorithm [257], which is part of the Universal PBM Analysis Suite [41], to calculate an enrichment score (E-score) for each DNA 8-mer. The E-score is a rank-based statistic that is closely related to the area under the receiver operating characteristic (ROC) curve. Larger E-score values reflect higher specificity for binding a particular 8-mer. Z-scores for each 8-mer and position weight matrices (PWMs) were also derived using the Universal PBM Analysis Suite

and Seed-and-Wobble algorithm, respectively. Throughout this text, only E-scores and Z-scores scores for ungapped 8-mers were used. Sequence logos for each allele were created by using the Seed-and-Wobble PWM as input for WebLogo v2.8.2[300] with default parameters.

The presence of E-scores ≥ 0.45 has been reported as a viable quality control metric to identify successful PBM experiments[26,301]. Here, we deemed a PBM experiment to be of acceptable quality under a more stringent criterion of yielding $\geq$ five 8-mers with an E-score ≥ 0.45. Because some mutant TF alleles are expected to lose their ability to bind DNA specifically, we considered such experiments acceptable for publication as long as the reference allele protein expressed and tested in the same batch yielded $\geq$ five 8-mers with E-scores ≥ 0.45.

4.4.8 Identifying affinity differences

To determine if two alleles exhibited a difference in binding affinity, we compared the distribution of E-scores obtained for each allele. A high E-score value indicates a strong deviation from the null distribution for the ranks of probes containing instances of a particular 8-mer. As the affinity of a TF allele increases while the concentration is constant, more binding sites will be occupied at high frequencies. Therefore, with all other parameters remaining constant, a higher affinity allele should yield a PBM dataset with a larger number of high-scoring 8-mers.

We used the Wilcoxon rank-sum test to determine whether a pair of experiments showed differences in their top E-scores. We calculated the Wilcoxon rank-sum test P-value when comparing the highest 50 E-scores in each experiment. We corrected the P-values derived from comparing reference and alternative alleles using the Benjamini-Hochberg correction[302], which was calculated over all pairwise comparisons between reference and alternative alleles. Mutations were classified as changing affinity when $Q < 0.05$. The direction of the affinity change (*i.e.*, increase or decrease) was determined by comparing the median value among the top 50 E-scores for each allele and selecting the allele with the larger median value as the one with the predicted higher affinity.

4.4.9 Identifying specificity differences

To detect specificity differences between alleles, we used a previously described method[254] for identifying statistically significant differences among 8-mer E-scores between two PBM datasets. Briefly, DNA 8-mers are placed into overlapping groups composed of all 8-mers that contain matches to a given DNA 6-mer. The E-scores corresponding to 8-mers in each these groups are then compared across alleles using an intersection-union test[254], followed by the adjustment of P-values using the Benjamini-Hochberg correction[302]. The result is a set of 6-mers that are bound preferentially by one TF allele over the other.

Here, we developed a stringent set of criteria for determining whether a mutant TF allele bound DNA with altered specificity relative to the reference allele. First, we excluded any experiments where the alternative allele significantly lost sequence-specific binding activity, as these cases might lead to confounded affinity and specificity changes. Therefore, only datasets from alternative alleles that met the same quality control criterion used for reference alleles (at least five 8-mers with E-scores ≥ 0.45) were tested for specificity differences. In addition, we excluded pairs of alleles where the number of 8-mers bound by the alternative allele at an E-score ≥ 0.45 was at least 2-fold less than the number bound by the reference allele. For the remaining pairs, we used the method described above to find preferred 6-mers with a q-value < 0.05. We found that pairwise comparisons between alleles where at least ten 6-mers were bound preferentially by either allele were highly reproducible across replicates (see below). Therefore, we considered alleles that matched all criteria described in this paragraph and for which pairwise comparisons with the reference allele yielded ≥ 10 preferred 6-mers to have altered specificity.

4.4.10 Reproducibility of affinity and specificity differences

E-scores have been previously shown to be highly reproducible across replicate PBM experiments[41]. We verified that alleles identified as having altered affinity or specificity were consistently labeled as such in a set of 58 duplicate PBM experiments. The dataset for each replicate experiment was independently scored using the criteria described above. Affinity calls were found to be consistent across replicate experiments in 97% of replicate pairs, while specificity calls were consistent across 86% of replicates. In discordant cases, the replicate experiments with the largest total number of E-scores $\geq$ 0.45 were used to determine whether a particular allele had altered affinity or specificity.

4.4.11 Concordance with experimental data from other studies

We searched the literature to identify cases where the same mutations selected for this study had been previously tested experimentally to determine their biochemical effects. Through manual curation, we collected a set of 20 experiments that directly or indirectly measured the binding affinities of mutant alleles.In most of these cases, only qualitative data were provided, such as gel images derived from non-quantitative electrophoretic mobility shift assays. Therefore, to enable systematic comparisons, we manually curated each reported experiment and assigned the mutant allele to one of three categories: (a) no effect on DNA binding (o), (b) partial loss of binding (-), and (c) complete loss of binding (--).

4.4.12 Predictions of ChIP-Seq peak disruptions

We used previously identified Crx ChIP-Seq peaks in mouse retinal cells[266] to predict likely *in vivo* effects of CRX mutations. For each Crx peak and for each CRX allele, we determined the DNA 8-mer with the highest E-score for the sequence within the peak boundaries. For each non-reference allele, we calculated the z-score difference for the top 8-mer in each peak relative to the top z-score for the reference allele. Based on these z-score differences, we predicted the top 100 peaks with the highest

predicted change in binding by each non-reference allele. Each set of 100 peaks was then used as input to GREAT[213], which determines if genes in the proximity of peaks (within 1 Mb) are enriched for certain ontology terms. We used the default GREAT settings and tested for association with terms in the "Mouse Phenotype" ontology. Terms that had a q-value < 0.05 for the hypergeometric test, as reported by GREAT, were considered to be enriched for each allele.

4.4.13 Identification of co-expressed paralogs and LoF-tolerant genes

Paralogous gene pairs were identified using annotations from the Duplicated Genes Database (DGD) (February 25, 2015 release)[303]. Any pair of human genes belonging to the same homology group, as defined by DGD, was considered to be paralogous. Co-expression was determined using the Hsa.v13 dataset obtained from COXPRESdb[304]. COXPRESdb provides a matrix of Pearson correlation coefficients quantifying the similarity of expression pattern of gene pairs across a wide range of tissues. We identified gene pairs as being co-expressed when one of the genes was among the 25 genes with the highest correlation coefficient for the other gene. The results related to co-expressed paralogs were essentially unchanged when the threshold was varied to include the top 50 or top 100 most correlated genes as being co-expressed. Genes that were tolerant of LoF mutations were defined based on the results of Sulem et al.[274]. Briefly, a gene was considered LoF-tolerant if at least one of the individuals studied (which are putatively healthy) was reported as being homozygous for a frameshift or nonsense variant.

4.4.14 Statistical testing of DBDP enrichment in TF subsets

For each gene, we calculated the number of predicted base- or backbone-contacting residues that were altered by at least one genetic variant. To account for the fact that TFs can have different numbers of DNA-contacting residues, we normalized the number of residues affected by genetic variation by dividing by the number of DNA-contacting residues in each TF. We used a two-sample permutation

test to determine whether certain subsets of TFs had a higher fraction of variable residues than others (*e.g.*, genes with co-expressed paralogs vs. those without). For all permutation tests, we used the 'permTS' function in the *perm* R package with standard parameter values.

To determine that the enrichments observed were statistically independent (*e.g.*, not due to genes with co-expressed paralogs often being tolerant of LoF mutations), we fitted a standard linear regression model ('lm' function in R) with the fraction of variable DNA-contacting residues as the dependent variable, and binary values representing LoF-tolerance, the presence of a co-expressed paralog, and their interaction as dependent variables (see statistical formula below). Both LoF-tolerance and paralog presence were highly significant predictive features independently ($P < 10^{-5}$, t-test), while the interaction term was not significant ($P = 0.296$, t-test).

```
  Fraction.Variable.Residues ~  LoF.Tolerant + Paralog.Present +
LoF.Tolerant:Paralog.Present
```

This chapter is derived from a draft manuscript, which is being submitted for publication as:

Luis A. Barrera, Jesse V. Kurland, Anastasia Vedenko, Stephen S. Gisselbrecht, Julia M. Rogers, Elizabeth J. Rossin, Jaie Woodard, Trevor Siggers, Leila Shokri, Raluca Gordân, Nidhi Sahni, Chris Cotsapas, Tong Hao, Song Yi, Manolis Kellis, Mark J. Daly, Marc Vidal, David E. Hill, Martha L. Bulyk. Survey of variation in human transcription factors reveals prevalent DNA binding changes.

Acknowledgments

We thank Max Hume, Yu-Han Hsu, Yun Shen and Dawit Balcha for technical assistance, and Sasha Gimelbrant for helpful discussions. The authors would like to thank the Exome Aggregation Consortium and the groups that provided exome variant data for comparison. A full list of contributing

groups can be found at http://exac.broadinstitute.org/about. This work was supported by the National Institutes of Health (NHGRI R01 HG003985 to M.L.B.), and by National Science Foundation Graduate Research Fellowships to L.A.B. and J.M.R. All PBM data (probe signal intensities, 8-mer scores, and position weight matrices) have been deposited into the UniPROBE database (publication dataset accession BAR15A).

Author contributions

J.V.K. and A.V. prepared proteins and performed PBM experiments, J.V.K., A.V., N.S., S.Y., D.B., T.H. and J.M.R. performed cloning, L.A.B., E.J.R., C.C., T.S., S.S.G., L.S., R.G. and J.M.R. curated variants for PBM analysis, L.A.B. performed computational data analysis, M.K., M.J.D., M.V., D.E.H. and M.L.B. supervised research, L.A.B. and M.L.B. designed the study and wrote the manuscript, L.A.B., N.S., D.H. and M.L.B. wrote the Materials and Methods. M.L.B. is a co-inventor on U.S. patent #8,530,638 on universal PBM technology.

My contributions

- Expanded the original project idea to encompass a much greater number of variants, include Mendelian mutations, and use new sources of information for variant prioritization.
- Analyzed genotype data from the 1000 Genomes Project and ESP6500 to identify non-synonymous SNPs that affect DNA-binding domains of human TFs.
- Performed linkage disequilibrium analysis, using genotypes from the 1000 Genomes Project data, to identify nsSNPs in LD with GWAS and eQTL tag SNPs.
- Developed a computational pipeline for prioritizing nsSNPS by mapping them to annotated DNA-contacting residues

- Led the selection of a set of 144 mutant alleles for testing by PBM.
- Coordinated with collaborators at the DFCI Center for Cancer Systems Biology for the generation of DBD subclones, both for reference and mutant alleles. This involved developing an automated pipeline to screen available clones from the library, design the target mutation, check the primers used for mutagenesis, and verify Sanger sequencing results.
- Played an active role in experimental design, working with technicians to evaluate data quality, tweak protocols and re-do experiments as necessary.
- Developed a pipeline for the automated analysis of PBM data, running several algorithms to quantify binding preferences and calculating various metrics for QC.
- Developed a statistical framework to reproducibly identify significant differences in binding affinity between alleles in PBM experiments.
- Performed PBM data analysis, comparing the general properties of nsSNPs and Mendelian mutations, and contrasting the effects of mutations in a single TF.
- Compared our results with predictions made by commonly used variant prioritization tools, such as PolyPhen 2, SIFT, etc.
- Developed an approach that can be used to prioritize mutations in TFs that are likely to affect DNA-binding affinity or specificity.

Essentially, all models are wrong, but some are useful.

George E.P. Box

5

Improved tools for TAL effector design

Abstract

Transcription Activator-Like Effector (TALE) proteins recognize DNA using a seemingly simple code, which makes them attractive for use in genome engineering technologies that require precise targeting. While this code has been used successfully to design TALEs to target specific sequences, off-target binding has been observed and proven difficult to predict. Here, we explore TALE-DNA interactions comprehensively by quantitatively assaying the DNA binding specificities of 21 representative TALEs to ~5,000-20,000 unique DNA sequences per protein, using custom-designed protein binding microarrays (PBMs). We find that protein context features exert significant influences on binding specificity. Thus, the canonical recognition code does not fully capture the complexity of TALE-DNA binding. We used the PBM data to develop a computational model, Specificity Inference for TAL-Effector Design (SIFTED), to predict the DNA-binding specificity of any TALE. We provide SIFTED as a publicly available web tool that predicts potential genomic off-target sites for improved TALE design.

5.1 Background

The discovery of Transcription Activator-Like Effector (TALE) proteins has enabled the development of a host of genome and epigenome editing technologies[132–134,305–309]. Naturally occurring as bacterial virulence factors, TALE proteins harbor an array of repeats, each 33 or 34 amino acids in length[128,129]. The sequence of the repeats is highly conserved except at the hypervariable positions 12 and 13, termed the repeat variable diresidues (RVDs). The amino acids at the RVD positions determine which DNA base is preferred, and each repeat in the TALE contacts one base in the target site. This led to a simple one-to-one "TALE code" that uniquely predicts the optimal DNA target from the sequence of RVDs within the repeat array[128,129]. The most commonly used RVDs are NI, HD, NN, and NG, which are used to target A, C, G, and T, respectively. Co-crystal structures have shown the mechanism of this

one-to-one code, in which the TALE protein wraps around the DNA in a helical structure with each repeat contacting a single base[310,311]. Additionally, contacts between the N-terminal region (NTR) of the TALE protein and DNA specify a preference for a thymine base at the 5' end of the DNA target site[127].

This simple TALE recognition code allows for any DNA site preceded by a T to be targeted by a TALE protein designed with the corresponding repeat sequence. Because of the relative simplicity of this approach, the TALE DNA binding domain has been adapted for use in many technologies that require precise targeting of genomic loci. For example, dimeric TALE nucleases (TALENs) have been used in various organisms and cell lines to knock out genes by creating base pair insertions or deletions (indels) or to create specific nucleotide substitutions[307]. Fusions of TALE monomers to transcriptional activation or repression domains can create artificial transcription factors, which have been shown to strongly and cooperatively modulate gene expression[132,306,309]. Monomeric TALE fusions to chromatin-modifying enzymes can introduce specific DNA or histone modifications at target loci, resulting in changes in expression of the associated genes[133,134]. TALEs can also be used to pull down specific genomic regions to identify bound proteins[305]. Additionally, TALEs fused to fluorescent proteins can be used to visualize chromatin dynamics in live cells[305,308]. While other technologies, (*e.g.*, CRISPR-Cas9) have also been developed for some of these targeting applications[312], TALE versus dCas9 fusions might be more effective in different applications and having both technologies in the toolkit for genome engineering is likely optimal.

Despite these successes in genome editing, off-target activities of TALE fusions have been described but have proven difficult to predict[136–140,313,314]. Experimental approaches have identified off-target TALEN effects[140], but no technology has directly measured off-target binding for monomeric TALE fusions[137–139,315]. Here, we define TALE protein specificity as the relative binding energies of the protein to different DNA sequences. Computational tools that use the specificities of the individual repeats to predict the specificity of the whole protein have been developed to predict off-target binding

sites[316,317]; these approaches assume that each repeat independently contributes to the specificity of the whole protein and that each instance of a given repeat RVD type has the same preference for its intended base. However, a quantitative analysis of TALE-DNA binding affinity indicated that repeat position within the repeat array affects RVD specificity, indicating a potential role for repeat context in predicting specificity[318]. Other studies have also found that total protein length affects specificity[140]. Additionally, particular repeat types may contribute differentially to overall protein specificity. One study showed that some repeats are more active when assembled into a TALE activator, leading to the distinction between strong (NN and HD) and weak (NI and NG) repeats, although the relationship between RVD strength and specificity is unclear[319]. Altogether, these findings suggest that TALE-DNA binding specificity may be more complex than previously thought, but these effects have yet to be assayed comprehensively and quantitatively.

Tools used to predict TALE specificity and to identify likely genomic targets have not kept pace with these increasing, albeit qualitative, reports on TALE-DNA recognition. Some computational tools, such as PROGNOS and Talvez, have incorporated context effects qualitatively in predicting TALEN pair off-target sites, but assume all repeat types are affected identically by context[320,321]. A recently described approach used a selection-based cleavage assay to characterize a TALEN pair's specificity profile in order to identify potential TALEN off-target sites; however, that study did not provide a predictive model, but instead required that the specificity of each TALEN pair be determined experimentally[140]. As such, there remains a need for a purely computational tool that quantitatively incorporates these context effects in predicting TALE specificity, and thus, off-target binding sites.

In this study, we perform a quantitative, in-depth examination of context effects on RVD specificity in order to infer general rules for highly accurate prediction of the DNA sequence-specificity of any TALE protein. We designed custom protein binding microarrays (PBMs) to investigate the DNA binding specificities of 21 TALE proteins that comprise all possible pairs of repeat types. The custom PBMs contain probes in which all possible mono- and di-nucleotide substitutions within the

TALE target sites are represented. The resulting quantitative binding data for the TALE proteins to ~20,000 unique DNA sequences allow us to quantify the effects of TALE repeat array length, repeat position, and neighboring repeat types on the specificity of each RVD, henceforth referred to as RVD specificity. We use the PBM-derived quantitative binding data to develop a computational model (Specificity Inference for TAL-Effector Design, or SIFTED) that incorporates these context effects to predict both the DNA binding specificity and the potential off-target sites of any TALE protein without requiring any additional PBM experiments. We implement this model in a publicly available, user-friendly suite of web tools at http://thebrain.bwh.harvard.edu/sifted.html.

5.2 Results

5.2.1 Custom-designed PBMs to assay TALE DNA-binding specificity

In order to develop a more in-depth, quantitative understanding of TALE-DNA recognition, we determined the DNA-binding specificities of 21 representative TALE proteins using custom-designed PBMs[41,257,322] (Figure 5.1 a, Figure 5.1b). We selected these proteins to allow us to examine the effects of different protein features on specificity. In particular, these proteins represent all possible consecutive repeat pairs and thus allow us to assay all possible direct neighbor effects on RVD specificity (Figure 5.1a)[91]. In addition, this set spans protein lengths from 8.5 to 18.5 repeats (targeting sites 10 to 20 base pairs in length); these lengths typically have been used in the design of monomeric TALE fusion proteins for genomic applications[132].

PBMs are double-stranded DNA microarrays that permit rapid, highly parallel measurement of the binding of a protein of interest to tens of thousands of unique DNA sequences in replicate, allowing for a much richer picture of TALE-DNA recognition than has resulted from prior studies. Since the vast majority of our selected TALE proteins were designed to recognize sequences longer than those on the previously designed 'all 10-mer' universal PBM design[257], we designed custom TALE-PBMs

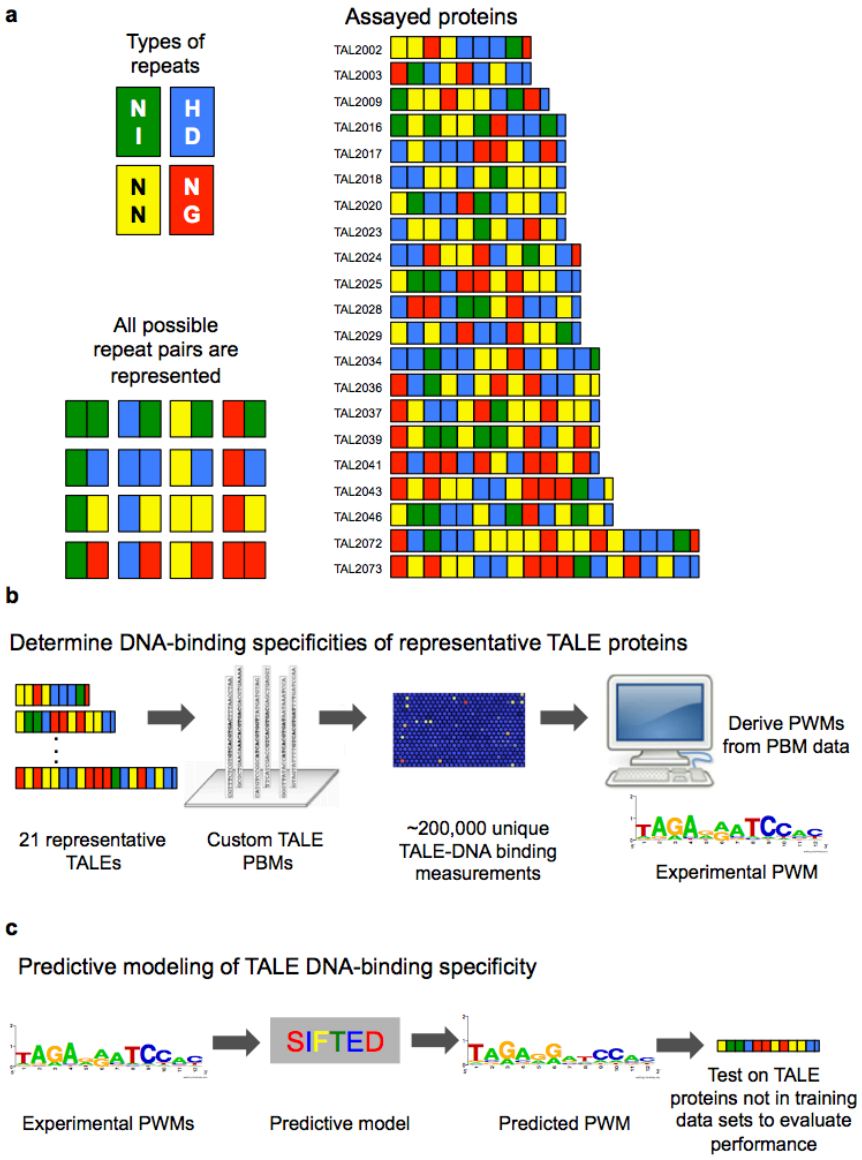


Figure 5.1: Overall experimental design and analysis scheme. (**a**) 21 Representative TALE proteins used in this study. Repeats are indicated by colored rectangles, and C-terminal half-repeats are indicated by smaller rectangles. RVD identities are indicated by letters. The set was chosen to include all possible repeat pairs and to cover a range of repeat lengths from 8.5 to 18.5 repeats. **(b)** Custom-designed PBMs were used to determine the specificity of representative TALE proteins. **(c)** These specificity profiles were used to learn features of TALE-DNA recognition and to train a predictive TALE specificity model, SIFTED (Specificity Inference for TAL-Effector Design).

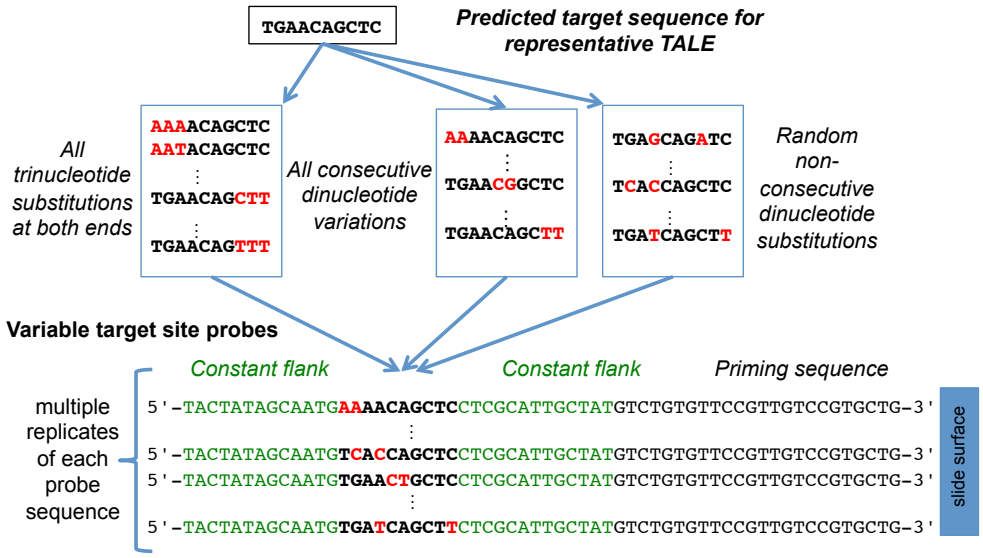


Figure 5.2: Design of probes on custom arrays. Schematic representation of custom arrays design for a representative TALE protein. Red font indicates variable sequences, while green font indicates constant sequence.

for this study. Each probe sequence was represented on at least eight replicate spots on the arrays. The initial custom array was designed to broadly assay the binding preferences of our representative set of TALE proteins. Subsequently, additional arrays were designed to validate particular observations about TALE specificity, as described below and depicted in Figure 5.2.

Types of probe sequences included in custom PBMs

All consecutive dinucleotide substitutions within the target site. For each protein, the target site is predicted using the TALE code, where the NI RVD targets A, HD targets C, NN targets G, and NG targets T. All target sites are preceded on the 5' end by a T. Sequences with all consecutive dinucleotide substitutions are generated. These target sites are positioned within constant flanking sequence.

Additional target site substitutions. The target site is predicted using the TALE code, as above, and random sets of up to five substitutions are made. These target sites are positioned within constant flanking sequence.

Clusters of substitutions at 5' and 3' end of binding site. The target site is predicted using the TALE code, as above, and clusters of three substitutions are introduced in the first three positions of the target site or in the last three positions. These target sites are positioned within constant flanking sequence.

Determination of TALE specificity using PBMs

We determined the DNA binding specificities of each TALE protein using probe sets that contain each protein's target site as predicted by the canonical TALE code[129], as well as variants thereof, flanked by constant DNA sequence and situated at a fixed position within the probe relative to the slide surface (Figure 5.2). The constant flanking sequence was designed to not be bound by any of the TALEs tested in this study by containing no matches to the binding sites predicted by the TALE code, considering up to 5 mismatches. For each protein, we measured binding to between 160 and 320 variant target sites that cover all possible adjacent dinucleotide substitutions. Although the absolute K_d of a protein-DNA interaction cannot be determined from a single PBM experiment[91], by measuring how much each substitution changes protein binding to the DNA probe, we can infer changes in binding free energy ($\Delta\Delta G$ values) for each possible substitution within the target site (Figure 5.2).

From these $\Delta\Delta G$ values, we derived a position weight matrix (PWM) for the protein (Figure 5.3 2a). The inferred PWMs were consistent across experimental replicates and across PBM experiments performed at different concentrations of TALE proteins. Our PWMs accurately predict the 60-base-pair probe signal intensities, with a median R^2 of 0.959 (Figure 5.3b), indicating that they perform well as accurate models of TALE DNA-binding specificity.

The fact that our PWMs explain binding well suggests that an additive binding model with independence between the nucleotides in the TALE target site is quite accurate. To test if this nucleotide independence extends beyond two adjacent mismatches, we designed a probe set that contains up to five nonadjacent mismatches in the target site (5.2). The PWM models derived from the dinucleotide

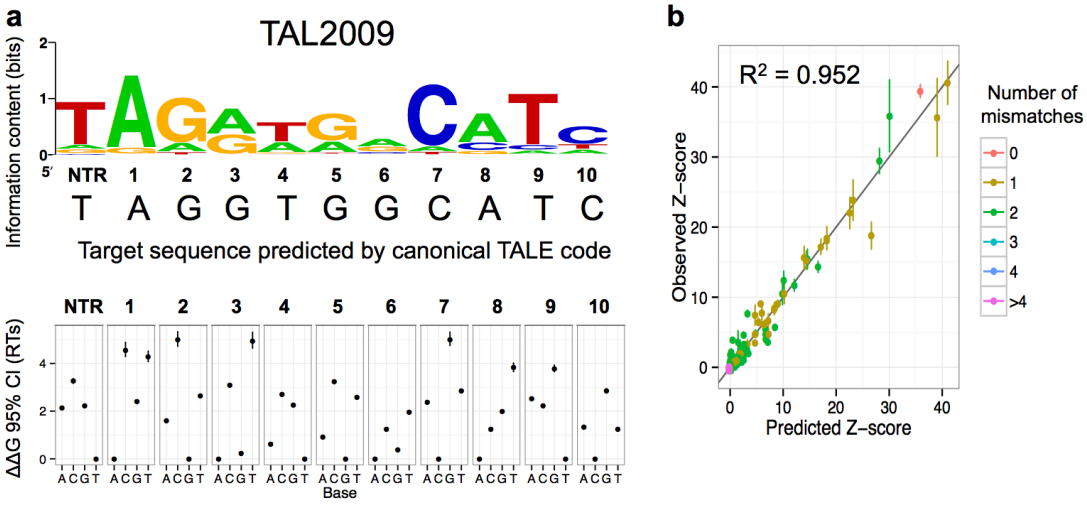


Figure 5.3: Determining PWMs from custom-designed PBMs. (a) Representative logo and $\Delta\Delta G$ estimates. The vertical bars represent the 95% credible interval (CI) and the points show the mean of the posterior distribution, in units of RT. The base predicted for each position by the TALE code is indicated below the logo. **(b)** Representative comparison between the probe z-scores measured in PBMs and the z-scores predicted by the derived PWM. Points represent the mean and vertical bars show its 95% confidence interval. Points are colored by the number of mismatches between the sequence in the probe and the consensus sequence predicted from RVD identities using the canonical TALE code.

substitution probes accurately predicted binding to these sequences with additional mismatches (median R^2 greater than 0.9 for all numbers of substitutions tested), indicating that the simple PWM models with mononucleotide independence perform well in modeling TALE DNA-binding specificity. These results are roughly consistent with a recent study of TALEN pair specificity determined by a selection-based cleavage assay, in which general independence in DNA recognition was observed; however, our data support a fully independent model of TALE-DNA binding, rather than a model with slightly increased tolerance for adjacent mismatches[140].

5.2.2 Modeling repeat context improves specificity prediction

Although we observed mononucleotide independence within TALE target sites, we found that the protein-DNA interactions of a given repeat are influenced by its context. In other words, the energetic parameters of a given TALE-DNA contact are not affected by neighboring nucleotide changes, but they are affected by the repeat context. Intriguingly, even within a single TALE protein, different occurrences of the same repeat type can exhibit very different specificities. For example, in TAL2009, repeats 7 and 10 were both designed with the HD RVD to target C, but within the context of the TAL2009 protein each exhibits substantially different relative preferences for C as compared to other nucleotides (Figure 5.3a). Typically, the highest scoring probe corresponded to the target sequence predicted by the canonical TALE code; however, we observed multiple cases (*e.g.*, TAL2024) where a TALE protein bound mismatched sequences with comparable binding strength, hereafter referred to as affinity. Moreover, some TALEs (*e.g.*, TAL2009) even preferred a mismatched sequence to the predicted optimal target sequence; this most frequently involved an NN RVD, which can target both a G and an A in different contexts (for example, see repeats 3 and 6 in Figure 5.3a)[128]. Altogether, these results highlight that the simple one-to-one TALE code is not sufficient to accurately predict DNA binding specificity.

Since our results suggested that interactions between repeats modulate their individual RVD speci-

ficities, we modeled the PBM data to predict TALE specificity considering the context of each repeat in a TALE protein (Figure 5.1c). We named our model and its associated software tools SIFTED (Specificity Inference For TAL-Effector Design). In addition to modeling the intrinsic specificity of each RVD, SIFTED considers a variety of repeat context features, including the number of repeats in the protein, each repeat's position within the repeat array, and the immediately adjacent N- and C-terminal neighboring repeat types. The NTR, which specifies the preference for the 5' T in the binding site, was also included in the model and was treated equivalently to a repeat, except for the omission of its position and length features.

We trained the SIFTED model by performing a linear regression with Elastic Net regularization, using the $\Delta\Delta G$ values inferred for each protein as the input data[323]. To prevent overfitting and to assess performance, we used a nested leave-one-out cross-validation strategy. Briefly, one protein was held out from the dataset in an iterative fashion. The remaining proteins were divided into training and test sets, which were used to derive parameter values and to control the complexity of the model. The predicted PWM for each of the 21 TALE proteins was obtained from the model trained on data from the remaining 20 proteins in our dataset (Figure 5.1a). For specificity predictions of proteins not in our dataset (*e.g.*, TALEN pairs), the regression was performed on the full dataset (no proteins excluded) and the resulting model was used to make PWM predictions.

To assess how well our model explains binding, we used the PWMs obtained from the cross-validated SIFTED model to predict PBM probe signal intensities. The SIFTED PWMs accurately predict the probe-level PBM binding data (median $R^2 = 0.877$). Additionally, SIFTED outperformed the specificity models from other available computational tools designed to predict off-target sites in explaining the PBM data ($P < 10^{-6}$, Wilcoxon signed-rank test) (Figure 5.4a). Two of these tools, TALE-NT 2.0[316] and TALgetter[317], do not consider any context effects. Others, such as PROGNOS[320] and Talvez[321], include context effects on an RVD's specificity only as discrete penalties. In contrast, SIFTED models context effects quantitatively and also allows each repeat type (*i.e.*, NI, HD, NN, and NG) to be influ-

enced differently by its context. These detailed context parameters in our model are keys to its success; the full model predictions from SIFTED are more accurate ($P < 10^{-6}$, Wilcoxon signed-rank test) than those of an RVD-only model that represents the canonical, one-to-one TALE-DNA recognition code (median $R^2 = 0.798$) (Figure 5.5).

We validated that our SIFTED model can predict off-array binding affinity measurements (K_d values) more accurately than other published tools[324] (Figure 5.4b). While PWMs cannot be used to predict absolute dissociation constants, they are able to predict the affinity of a sequence relative to that of the optimal binding site (i.e., relative K_d values)[325]. The full SIFTED model performed significantly better than PROGNOS, TALE-NT 2.0, TALgetter, Talvez, or a reduced SIFTED model with no context effects in predicting relative K_d values for one protein and 18 DNA sequences[324].

5.2.3 Quantitative modeling of context effects on RVD specificity

Since context effects contributed significantly to the predictive power of our model, we investigated in greater depth how length, position, and neighboring repeats each affect specificity. While our baseline RVD specificities (Figure 5.5a) largely agree with previous studies[129] (*e.g.*, NN is the least specific RVD and can target both G and A), in the SIFTED model these specificities are modulated by the protein context of each instance of the repeat.

Our data are consistent with previous reports that longer proteins tolerate more mismatches in their target sites[140] (Figure 5.6b). Our comprehensive profiling also revealed that NN and NG repeats are affected more strongly by protein length than are either NI or HD. Additionally, our set of proteins included two proteins of different lengths designed to target overlapping sites. The longer protein (TAL2073) is less specific overall (*i.e.*, lower total information content) than the shorter protein (TAL2043), directly supporting our overall finding that increased TALE protein length diminishes RVD specificity.

Repeat position within the repeat array also affects the specificity of C-terminal repeats that tar-

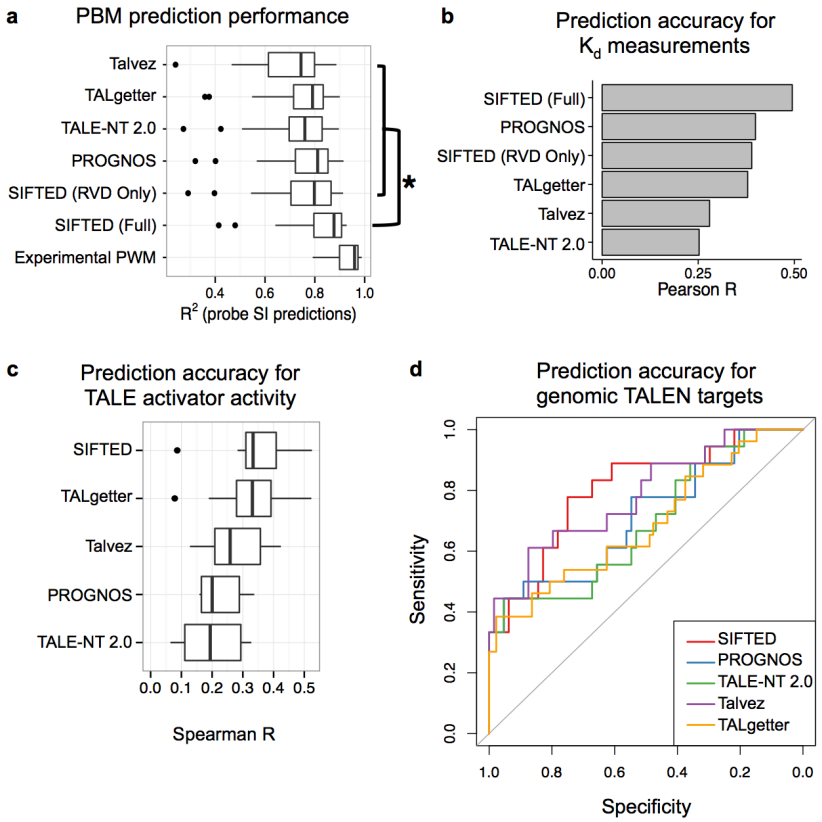


Figure 5.4: SIFTED predictive model performance. (a) Comparison of prediction accuracy of PWMs derived by different methods. The box plot shows how well the PBM probe intensities for each protein are predicted by the PWMs generated by SIFTED and other methods. Two versions of SIFTED are shown: one that only models repeats independently (``SIFTED (RVDs Only)'') and one that considers all repeat context features (``SIFTED (Full)''). Experimental PWMs are those derived from the PBM data. (*) The brackets highlight a subset of statistically significant differences ($P < 10^{-6}$, Wilcoxon signed-rank test). The box plots shows the median and the first and third quartiles. Whiskers extend to data points not considered outliers, while outliers are shown as individual points. Data are considered outliers when they are 1.5 times the interquartile range (IQR) higher than the third quartile, or 1.5 * IQR lower than the first quartile. **(b)** Prediction accuracy for relative binding affinity. PWMs derived from existing tools or from SIFTED (as in (a)) were used to predict relative Kd values for a single TALE protein [320,324]. The bars display the Pearson correlation coefficient between observed and predicted log(K_d) values. **(c)** Validation of TALE activator binding specificity predictions by comparison to TALE activator activity data reported in Mali *et al* [315]. The five predictive methods were used to score all reported binding sites up to three mismatches away from the predicted target. These scores were compared to a expression score associated with that binding site using Spearman correlation. **(d)** Validation of TALEN binding specificity predictions by comparison to cell-based TALEN activity data, reported in Guilinger *et al* [140]. The five methods shown were used to predict the binding of TALEN pairs to genomic target sites. The ROC curves show the sensitivity and specificity of each method for distinguishing genomic sites that showed nuclease activity (i.e., indels) and those that did not.

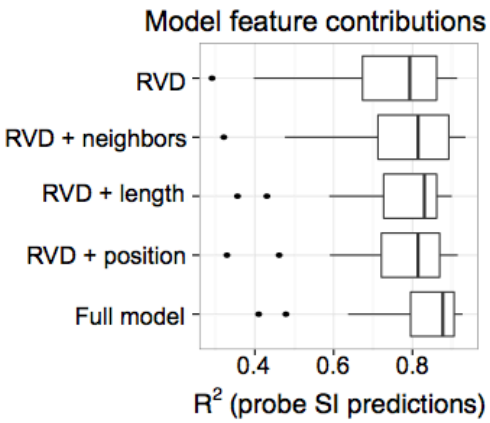


Figure 5.5: Contribution of model features (a) The plot shows the accuracy at predicting PBM probe intensities of a PWM predicted with no context features (top), with one single context feature added (middle) or with all context features included (bottom). Box plots are formatted as in Figure 5.4a.

get the 3' end of the DNA binding site, resulting in their being more tolerant to substitutions than N-terminal RVDs. To test this modeling result, we designed a custom PBM that included probes containing clusters of three nucleotide substitutions located at either the 5' or 3' end of the target site (5.2). In general, substitutions at the 5' end impaired binding more than substitutions at the 3' end ($P < 0.05$, Wilcoxon signed-rank test), supporting prior observations from reporter assays[318,326] . Talvez and PROGNOS model this polarity effect discretely as a constant decrease in specificity after a certain position in the repeat array for all repeat types[320,321]. In contrast, SIFTED continuously models the decrease in specificity over the length of the protein and allows different repeat types to be affected to different extents (Figure 5.6b).

Lastly, we observed that a repeat's specificity is impacted by the identity of the immediately adjacent N- or C- terminal repeat (Figure 5.6c). Such local context effects previously have been observed only for the 5' T preference, which is more important for binding when the first repeat is an HD[327]. We also observed the influence of HD in the first position, but found an even stronger effect when the first repeat is an NN. Additionally, we observed neighbor context effects between repeats within the

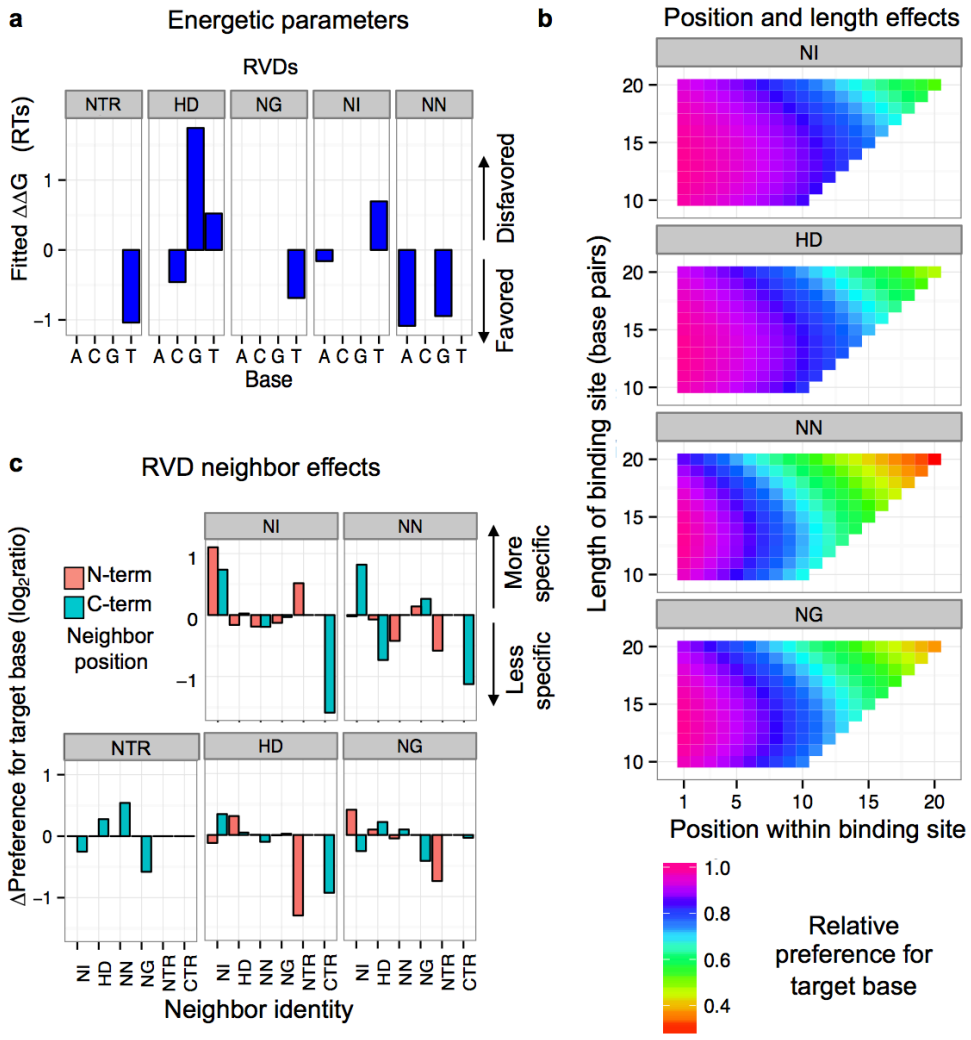


Figure 5.6: Protein features that affect repeat specificity. (a) RVD identity. $\Delta\Delta$G values from the model are indicated for each repeat type with each base. Additionally, the $\Delta\Delta$Gs for the four bases at the 5' T position, which are contacted by the NTR, are shown. **(b)** Length and position. The effects of protein length and repeat position on the specificity of each repeat type are shown. **(c)** Effect of neighboring repeats or terminal regions on specificity. For each repeat type and the NTR, the bar heights display the effect on specificity for different neighbors in the N- or C-terminal direction (orange and teal, respectively). The quantity shown is the $\log_2$ ratio between the PWM frequency predicted with and without the presence of a given neighbor in the model. NTR refers to the N-terminal region of the protein. CTR refers to the C-terminal region; repeats with the CTR as the C-terminal neighbor are the half-repeats in the final repeat position.

protein. For example, the NN repeat is more specific for targeting a G when the NI repeat is either N- or C-terminal to it; however, it is much less specific for G when it is positioned at the C-terminal end of a TALE repeat array.

We found that a particular repeat type can exert different effects as an N- or C- terminal neighbor (Figure 5.6c). PROGNOS includes a parameter to reduce an RVD's specificity when it is next to a strong RVD (NN or HD), positing that a stronger neighboring interaction may allow for greater mismatch tolerance[319,320]; however, it does not distinguish between N- and C- terminal neighbors. The neighbor effects we found are more complex, and in fact, the strong RVDs do not always decrease specificity. The complexities of the neighboring effects are captured quantitatively in SIFTED; each of the four RVDs as well as the 5' T preference are modeled as being affected differently by its N- and C- terminal neighboring repeats.

These observations of context effects can be condensed into some simple guidelines for TALE design. Certain repeat combinations (*e.g.* NI-NI) are predicted to have increased specificity, while others (*e.g.* NG as the N-terminal repeat) can make an RVD more tolerant to mismatch and therefore should be avoided. However, when designing TALE proteins, one must ultimately consider all the context effects in the protein, as well as the prevalence of potential off-target sites in the genome. As such, we tested if the SIFTED model could accurately predict genomic off-target sites, and therefore could be used to guide TALE protein design.

5.2.4 Predicting TALE off-target sites using SIFTED

To assess whether SIFTED can predict genomic off-target sites for TALE proteins that have not been assayed by PBMs, we examined a dataset of *in vivo* TALE reporter activity[315]. SIFTED had the highest median performance of the five tools tested (Figure 5.4c).

Although SIFTED was designed to predict TALE monomer specificity, we also tested its ability to predict TALEN binding by examining a large dataset of TALEN activity in cells[140]. We derived the

specificities of TALEN pairs from the specificities of the component monomers predicted by SIFTED. The PWMs from SIFTED resulted in better sensitivity and specificity than those from any of the other models in distinguishing genomic target sites that showed nuclease activity from those that did not (Figure 5.4d). The area under the receiver operating characteristic (AUROC) curve statistic was used to quantify the ability of the five tools to distinguish target from non-target sites across all possible score thresholds. SIFTED demonstrated superior sensitivity and specificity across most thresholds.

Additionally, we considered that a typical TALE user might investigate about 20 off-target sites when analyzing the specificity of their designed protein in their genome of interest. To provide a performance comparison for this typical use case, we investigated how many of the top 20 off-target sites predicted by these tools have been identified as TALEN pair off-targets *in vivo*, or were among the 20 off-targets with the highest measured *in vivo* activity. Again, SIFTED performed better than the other tools, demonstrating higher sensitivity by predicting more of the true off-targets than the other tools.

5.2.5 Prediction of genomic off-targets with SIFTED web tool

SIFTED was the top-performing model overall, highlighting the value of incorporating repeat context effects in predicting specificity. While other tools may perform comparably to SIFTED in a specific application, SIFTED was the only tool that was consistently a top performer across the wide range of benchmarks of predictive performance (Figure 5.4). Given the success of SIFTED in predicting off-target binding, we developed it into a web-based suite of tools to aid in TALE design implemented on the Galaxy platform[328–330] at http://thebrain.bwh.harvard.edu/sifted.html. We provide stand-alone tools for individual tasks, such as predicting the specificity and genomic binding sites of a user-specified TALE, as well as a pipeline that combines various tools to automate the process of designing a TALE to target a particular genomic region. The complete pipeline takes a user-defined genomic target region as input, and then (1) identifies candidate TALEs to target that input region, (2) predicts the candidates' specificities, (3) finds instances of off-target sites in a user-specified genome and (4) outputs a list of

candidate TALE proteins ranked by their off-target binding potential, thus allowing the user to select the best candidate protein.

5.3 Discussion

By analyzing TALE proteins of different lengths and containing all possible consecutive pairs of repeats, we were able to identify the influence of repeat context on DNA binding specificity. In contrast to other studies that used cell-based TALEN activity as a measurement of TALE specificity[324], our experimental design allowed us to directly assay the intrinsic binding properties of TALE monomers. We measured a total of ~200,000 binding interactions between 21 TALE proteins and ~5,000-20,000 unique DNA sequences per protein using custom-designed PBMs. Importantly, the resulting dataset allowed us to develop a model to predict TALE specificity for any candidate TALE protein without requiring any additional experimental analysis.

Our results highlight that RVD specificity is not determined simply by what base a particular RVD will bind, but also which bases it strongly disfavors. This information could be useful in designing TALEs for allele-specific applications, such as rapid, spatially resolved genotyping of patient samples through binding of fluorescently tagged, allele-specific TALEs. The HD RVD has the greatest power to discriminate between two alleles: it prefers binding to a C and strongly disfavors binding to a G. Therefore, targeting an allele where there is a C/G SNP may lead to stronger discrimination between the two alleles.

We found that longer TALEs are generally less specific than shorter TALEs. This effect could be due to excess DNA binding energy in TALE proteins with many repeats[140]. The mechanism of the context effects on RVD specificity remains to be determined. An ability to tolerate some binding site mismatches may allow a TALE protein from xanthomonad pathogens to overcome mutations in host genomic target sites, as the plant host may be under selection to escape xanthomonad infection. How-

ever, TALEs with very low specificity may lead to potential negative effects on virulence due to additional binding in the host genome[124]. Thus, the specificity of TALE proteins may have been strongly shaped by the complex interactions between host and pathogen.

SIFTED predicts that some DNA sequences should be targeted with greater specificity, which could be interpreted as guidelines for TALE design. Interestingly, some of these guidelines would contradict published guidelines that were developed as part of the SAPTA tool for designing more active TALEN pairs[331]. For example, while we predict that A-runs can be targeted with high specificity by TALE monomers, SAPTA predicts that TALENs targeting A-runs will have lower nuclease activity. The discrepancies in these guidelines and results might reflect different rules affecting the binding of monomeric TALEs versus dimeric TALENs. Alternatively, it is possible that a trade-off exists between optimizing activity and specificity in designing TALENs. Previous reports have found no correlation between activity and affinity[324]. This lack of correlation between *in vitro* binding and different cell-based activity measurements might be due to other genomic features in cells, such as the chromatin state and competition with other transcription factors at the target and off-target sites. Ultimately, in designing TALEs, the intrinsic specificity of the protein must be considered in light of its potential off-target binding sequences in the genome. For example, the decreasing specificity of longer TALEs may be compensated by longer target sites being more rare in the genome, thus increasing the effective specificity of a protein[140]. SIFTED can both model protein specificity as well as identify genomic off-target sites, revealing the effective specificity of a TALE, so users can choose the most specific TALE protein for their particular application.

Future studies will be required to identify chromatin features that might modulate binding specificity *in vivo*. Additionally, the specificities of other alternative RVDs (*e.g.*, NH to target G) could be studied to enable design of TALE proteins with higher sequence specificity. An improved understanding of TALE-DNA binding should allow for development of more precise genome engineering tools.

5.4 Methods

5.4.1 Cloning of TALE proteins

TALEN expression vectors[130] were digested with SacII and BamHI to obtain the DNA-binding domain comprising the Δ152 N-terminal domain (NTD), the RVD repeats, and the +63 C-terminal domain (CTD). This fragment was ligated into a modified pDONR221 vector (Invitrogen), with SacII and BamHI restriction sites internal to attL recombination sites, to create Gateway-compatible TALE Entry clones. The TALE constructs were then transferred by Gateway recombinational cloning into the pDEST15 expression vector, which adds an N-terminal glutathione S-transferase (GST) tag (Invitrogen), by an LR reaction. All clones were full-length sequence-verified.

5.4.2 Protein binding microarray experiments

Proteins were expressed using the PURExpress *In Vitro* Transcription and Translation Kit (New England Biolabs). Protein concentrations were determined by anti-GST Western blots with a dilution series of recombinant GST (Sigma). Proteins were stored at +4 °C until being used in PBM assays. PBMs were performed as described[41], with a 30-minute incubation with an Alexa488-conjugated anti-GST antibody (Invitrogen A-11131). The final concentration of TALE protein in the PBM binding reactions was 200 nM, unless otherwise indicated.

5.4.3 Custom PBM Design

Target sites for each TALE protein were determined using the canonical TALE code (NI: A, HD: C, NN: G, NG: T), and are preceded by the 5' T to create the full target site. The constant flanking region was the same as that used in a prior custom PBM design and does not contain binding sites for any of the TALE proteins in this study[332].

5.4.4 PBM data quantification

Raw data files generated by GenePix Pro (Molecular Devices) were processed using the same general approach as used for universal PBMs [41]. Briefly, masliner software [299] was used to combine Alexa488 scans at different laser power levels and resolve the signal intensity in spots that are saturated at high laser power settings. The adjusted BSI data were then normalized by the corresponding double-stranded DNA content of the spots and their position on the array using the same approach as described for universal PBMs [41]. We calculated the median normalized BSI over all replicate probes on the same array. For each TALE protein we defined a background set of probes that comprises all the probes on the array designed to represent binding sites for other TALE proteins (not the one being assayed in a given experiment). The z-score for each probe was calculated relative to the median and standard deviation of its corresponding background probes. For more detail on normalization procedure, see Methods.

5.4.5 Position weight matrix model fitting

We developed a Bayesian Markov chain Monte Carlo (MCMC) method to infer free energy parameters of TALE-DNA interactions from PBM data. We relied on the theoretical framework developed for the BEEML-PBM algorithm [325], which can accurately derive $\Delta\Delta G$ values for protein-DNA contacts from universal PBM experiments. The BEEML-PBM framework estimates $\Delta\Delta G$ values for each possible nucleotide substitution in a protein's DNA binding site motif. These values can be assembled to construct an energy matrix (EM), in which each column represents a position within the binding site and each row represents a nucleotide. The EM values can be converted to probabilities using the Boltzmann distribution, creating a position weight matrix (PWM). The statistical model is described in full detail in Eq. 5.1, where Y_j represents z-score transformed PBM signal intensity values for each experiment.

$$Y_j = a + \frac{b}{1 + e^{\sum_i \Delta\Delta G_{i,j} - M}} + \varepsilon_j$$

$$\begin{aligned} \Delta\Delta G_{i,j} &\sim \text{Exp}(\beta) \\ a &\sim U(-100.0, 100.0) \\ b &\sim U(0, 1000.0) \\ M &\sim U(-20.0, 20.0) \\ \varepsilon_j &\sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{5.1}$$

5.4.6 SIFTED predictive model for ΔΔG values

The ΔΔG values inferred from the TALE PBM experiments were used to train a predictive model using an Elastic Net regression[323]. We used the Elastic Net implementation in the glmnet v1.9-5 R package to train our model. Each ΔΔG value in the dataset is paired with a vector of predictive features to create the feature matrix, in which each row is an independent observation, and each column is a different feature. The features include repeat identity, position, neighboring repeat identity, and total length of the target site. To allow for nonlinear position and length effects, we also included the natural logarithm of each as a feature.

To prevent and to accurately assess the model's performance, we used a cross-validation scheme consisting of two nested levels. On the outer level, we used leave-one-out cross-validation to form a validation set by excluding a single protein in each iteration. Once a protein is excluded, the inner level performs 5-fold cross-validation on the remaining proteins. This entire process is repeated for each protein, leading to cross-validated predictions for the entire dataset. These predictions were then used for all model evaluation purposes.

5.4.7 Predicting probe signal intensities and K_d values from PWMs

The predictions of probe signal intensities were obtained using the same mathematical framework as for fitting PWMs (Eq. 5.1). However, in this case the ΔΔG parameters are known and the only parameters that need to be fitted to predict probe intensities are the chemical potential μ and the scaling terms a and b. To determine these parameters, we used the implementation of the Levenberg-Marquardt algorithm in the SciPy v0.12 package with default convergence parameters. The model parameters were initialized as follows: a = minimum z-score in input data, b = maximum z-score in input data, μ = -1.0. After these parameters were fitted from the observed z-scores, the predicted z-scores were obtained by using the total ΔΔG for the binding site in each probe and the fitted variables as input.

In order to validate SIFTED predictions with measured K_d values[324], relative K_d values for target and off-target sites were predicted from SIFTED PWMs. Relative K_d values were predicted by setting the K_d of the optimal site to 1. The predicted K_d for off-target sequences were obtained through the equation $e^{\Delta\Delta G/RT}$, where ΔΔG represents the difference in total free energy between the optimal binding site sequence and the sequence of the off-target site. The measured relative K_d values were similarly adjusted so that the optimal site had a K_d of 1. Because K_d values span many orders of magnitude, the correlation coefficient was computed after taking the natural logarithm of the K_d values, which prevents the calculation from being dominated by the extreme values.

5.4.8 Comparison using PWMs from other tools

PROGNOS, TALgetter, Talvez, and TALE-NT 2.0, the publicly available tools against which we compared SIFTED, do not explicitly provide the user with predicted PWMs[316,317,320,321]. However, with the exception of TALgetter, each tool uses an internal scoring scheme that is mathematically equivalent to a PWM (i.e., the score for a site represents the sum of an independent score for each nucleotide

position). Therefore, in the comparisons with PROGNOS, Talvez and TALE-NT 2.0, we predicted PWMs based on the scheme described by each paper and the associated parameters[316,320,321]. To predict TALgetter scores, we instead used the downloadable TALgetter software tool to compute log-odds values for all binding site sequences in a given experiment[317]. These binding scores can then be compared directly to PWM log-odds scores, even if the underlying scoring scheme is distinct. For comparisons using TALEN activity data, we combined the values predicted by PWMs for each TALE in a TALEN pair using the same scoring scheme as PROGNOS[320]. The scoring scheme is shown in Eq. 5.2, where x is y and a corresponds to b .

$$\text{Pair Score} = \left(\frac{S_{\text{left}}(\text{optimal site})}{S_{\text{left}}(\text{target site})}\right)^{0.6} + \left(\frac{S_{\text{right}}(\text{optimal site})}{S_{\text{right}}(\text{target site})}\right)^{0.6} \tag{5.2}$$

We analyzed the TALEN target sites reported by Guillinger *et al.*[140]. We scored each reported target site that contained only NN, NI, HD, and NG RVDs using the TALEN Pair Score derived from the PWMs obtained from SIFTED, PROGNOS and TALE-NT 2.0. We summarized the performance of each tool as a receiver operating characteristic (ROC) curve, which shows the sensitivity and specificity values achieved by each tool when predicting sites that were targeted by the TALEN pairs. The different sensitivity and specificity values represent different Pair Score thresholds, above which a locus is predicted to show evidence of nuclease activity (indels).

We also compared against the TALE activator reported by Mali et al.[315] All of the reported binding sites up to three mismatches away from the predicted site were scored as described above. These scores were then compared to a normalized expression score (the ratio of barcode tags for that binding site relative to a control experiment) associated with that binding site-TALE combination. Since we expect the relationship between TALE occupancy and expression to be nonlinear, we compared the results using Spearman correlation.

5.4.9 Algorithmic approach of SIFTED web tool

The overall approach of the entire pipeline to identify and score candidate TALEs to target a genomic region is as follows:

1. Find candidate TALE binding sites within the user-input DNA sequence.
2. For each site found in (1), determine the protein that targets that sequence using the TALE code, and predict its PWM.
3. For each protein, use the PWM to enumerate all putative binding site sequences (both target and off-target sequences) within a relative K_d threshold (by default, set to 10), using a bounded breadth-first search.
4. Find all genomic instances of the putative binding site sequences from (3) using a short read aligner (*bowtie*).
5. Calculate a summary score for each protein that describes the overall number and strength of genomic target sequences.

Under default parameter settings (*e.g.*, 13.5 repeat TALE, 1-kb region), the SIFTED pipeline typically identifies optimal TALE candidates within minutes. Additionally, a user can input a TALE with a defined RVD sequence, and SIFTED will predict its specificity and identify potential genomic off-target sites. Tutorials are hosted on the SIFTED website for designing TALEs to target a region, and for predicting the specificity of a pre-designed TALE, and include additional guidelines for setting parameters and troubleshooting.

This chapter is a modified version of a published article describing this work:

Rogers JM*, Barrera LA*, Reyon D, Sander JD, Kellis M, Joung JK, Bulyk ML. Context influences on TALE-DNA binding revealed by quantitative profiling. *Nature Communications* (2015) 6:7440.

Acknowledgements

This project was supported in part by National Science Foundation Graduate Research Fellowships to J.M.R. and L.A.B., grant R21 HG007573 from NIH/NHGRI to M.L.B., an NIH Director's Pioneer Award (DP1 GM105378) to J.K.J., and the Jim and Ann Orr MGH Research Scholar Award to J.K.J. We thank Alexandre Palagi and Kian Hong Kock for helpful discussion.

Author contributions

M.L.B., J.M.R. and L.A.B. designed the study, D.R. and J.D.S. assembled TALEs, J.M.R. and L.A.B. designed custom oligonucleotide arrays, J.M.R. cloned and expressed TALEs and performed PBM experiments, J.M.R. and L.A.B. performed data analysis, L.A.B. developed and fitted models, performed statistical analyses, and created the SIFTED web tool, M.K., J.K.J. and M.L.B. supervised research, J.M.R., L.A.B. and M.L.B. wrote the manuscript.

My contributions

- Developed, jointly with Julia Rogers, graduate student in Bulyk lab, three custom PBM probe sequence designs to address various questions about the DNA binding properties of TALEs.
- Conceived and implemented a software pipeline for the normalization and analysis of custom PBM experiments.
- Developed a Bayesian approach, using Hamiltonian Markov-Chain Monte Carlo sampling, to accurately infer free energy parameters from custom PBM datasets.
- Conceived and developed a predictive model for the specificity of TALE proteins with any RVD configuration by training a regularized regression model on the inferred free energy parameters for TALE-DNA interactions of 21 proteins.

- Through modeling, discovered that TALE-DNA binding has additional complexity beyond the simple RVD code. Neighboring RVDs influence the binding preferences of a given repeat within the array, implying that predicting binding specificity requires knowledge of the protein context in which a TALE repeat occurs.

- Demonstrated that the predictive model can accurately predict *in vitro* TALE binding data in cross-validation of PBM data and can also make accurate predictions about the indel rate achieved by TALE nucleases and the induction achieved by TALE transcriptional activators.

- Benchmarked the SIFTED model and showed that its predictions are more accurate for predicting various kinds of TALE activity than existing tools.

- Created a publicly available web tool, which uses the Galaxy platform to provide a user-friendly implementation of the predictive SIFTED model. This web tool uses the model that was fitted from the PBM data to aid in common design tasks for TALE proteins, such as selecting a protein that targets a particular sequence effectively but minimizes the off-target binding to other genomic sequences.

I almost wish I hadn't gone down that rabbit-hole—and yet—and yet—it's rather curious, you know, this sort of life!

Alice's Adventures in Wonderland, by Lewis Carroll

6

Conclusions

Developing a truly systematic approach to characterize regulatory variation is a task that will keep the scientific community occupied for many years, if not decades. In this dissertation, I have presented my case for why such efforts will be essential towards understanding the molecular mechanisms that underlie transcriptional regulation and human phenotypic variation. I have also described my own contributions and how they relate to the broader goals of the field. In this section, I present a more detailed discussion about the significance, limitations, and future areas of work that relate to the advances reported in the previous chapters.

The development and application of eFS have provided novel biological insights and demonstrated the feasibility of a new high-throughput approach for enhancer screening. The enhancers identified by eFS have already provided valuable information about transcriptional regulation in the *D.*

melanogaster mesoderm. Using motif enrichment analysis, we identified several TFs as having a potential role in driving the expression programs of mesodermal development, several of which had not been previously reported. One of the factors identified by this analysis of eFS data, Trithorax-like (Trl), was subsequently reported as an enhancer-promoter specificity factor[333], which mediates interactions between developmental enhancers and the promoters of their target genes. Other predicted mesodermal regulators are the subject of ongoing study in the Bulyk lab. Overall, the enhancers identified by eFS were associated with a large number of TF motif combinations that extend beyond the set of known master regulators, such as Twi, Tin, Bap, Bin and Mef2[64]. These observations suggest that, even in the extensively studied *Drosophila* mesoderm, much remains to be understood about the complexity and flexibility of *cis* regulatory codes.

An important advantage of the eFS methodology is that it can readily be applied to study other developmental stages and tissues. In principle, the experimental and computational framework described here can be used to study any subset of cells in which an enhancer is known to drive expression at sufficiently high levels. Recent developments in enhancer assays have facilitated the identification of candidate enhancers that drive expression in a desired cell type. For example, a high-throughput, imaging based approach was used to identify the spatiotemporal expression patterns associated with 3,557 developmental enhancers[334]. Such data can be analyzed directly, or integrated with gene expression patterns identified through *in situ* hybridization[183], to greatly aid the identification of enhancer sequences that can provide cell-type and tissue selectivity in future eFS experiments.

The development of eFS has added a novel tool to a rapidly growing repertoire of methods for identifying enhancer sequences. Each of these methods has unique strengths and weaknesses, ranging from their throughput, whether they function *in vivo* or *in vitro*, and whether enhancers are integrated into the genome or drive expression in plasmids. For example, STARR-seq allows million of candidate enhancers to be sequenced, but can only be feasibly carried out in cell lines and relies on plasmid expression[76]. Meanwhile, CRE-seq, as described by Mogno et al., allows the screening of thousands

of regulatory elements in a constant genomic context, but has only been used in yeast[335]. In contrast, eFS has a lower throughput, but is able to assay specific cell-types in a developing organism and detect enhancers using comparatively small numbers of cells. In this regard, it fills a unique and important niche in the toolbox of experimental methods to systematically study enhancers.

Having an enhancer assay that directly measures transcriptional activation can enable benchmarking of indirect methods for enhancer prediction. As methods based on DNase I hypersensitivity and TF and histone mark ChIP-Seq have become widely used, there is a pressing need to compare their predictive performance and understand why certain sequences do not drive expression despite possessing certain features that are typical of enhancers. When comparing between eFS results and indirect predictions for the same sequences, we observed similar predictive performance between DNase I hypersensitivity and histone modifications. However, we found that the binding of TFs known to regulate the tissue of interest offered the highest predictive performance. These results are consistent with recent observations about the predictive power of TAL1 binding for identifying active enhancers in mouse erythroid cells[336]. Additional work is needed to understand how well certain combinations of indirect features can predict enhancer activity and whether such observations are universal among tissue types, developmental stages, and species. Having a high-confidence set of enhancers identified through techniques such as eFS will greatly aid in performing such studies, and thus increase the interpretability of enhancer predictions from genome-wide indirect methods.

The issues encountered by studies of regulatory variation in humans, such as the inability of discovered eQTLs to explain most of the observed heritability of gene expression, have prompted similar studies in model organisms. Bloom et al. measured several quantitative traits across genetically heterogeneous yeast cells and found that, in contrast to studies carried out in humans, the vast majority of phenotypic variance could be explained by the identified QTLs[337]. Additionally, despite the relative simplicity of transcriptional regulation in yeast, many genes were shown to be affected by multiple eQTLs[338]. These observations suggest that model organisms may be an appealing option for under-

standing the mechanistic basis of regulatory variation.

Transcriptional regulation in *D. melanogaster* is significantly more complex than in yeast, making it a more appropriate model for vertebrate gene regulation. Studies of gene expression variability in flies have identified *cis* eQTLs for over 2,000 genes[339]. Combining high-throughput enhancer assays with eQTL mapping in flies is likely to facilitate the fine mapping of causal *cis* regulatory variants. For instance, an eFS library could be constructed from enhancer sequences harboring candidate regulatory SNPs. Because eFS allows the screening of relatively long sequences in a cell-type specific way, the results would be more directly interpretable than those based on screening enhancer fragments in plasmids transfected into cell lines, such as MPRA and CRE-seq[60,81].

The version of eFS described here has two main limitations: (a) the absence of a quantitative readout of enhancer activity (*i.e.*, an indication of how strongly a particular sequence drives expression), and (b) a comparatively lower throughput than other methods. The first limitation can likely be addressed by engineering a sequence tag into the GFP transcript that uniquely identifies each candidate enhancer. After FACS sorting, CD2+ cells (*i.e.*, those from the tissue of interest) can be separated into a subset that is sequenced to quantify the abundance of candidate enhancer integrations and a subset used for RNA-sequencing. After normalizing for the abundance of cells with a particular insert, the counts for a particular mRNA barcode should provide a quantitative activity readout.

The comparatively lower throughput of eFS could be improved in two main ways: (a) by increasing the number of embryos that are injected with plasmids harboring the candidate enhancer library, and (b) by eliminating the requirement of a single integration event per haploid genome. Solving the first challenge is largely a matter of increasing the scale at which embryos are injected and the number of cells that are collected. With additional resources, the eFS approach could be used to screen a few thousand sequences per experimental run. A larger improvement could be achieved by allowing multiple random integration events in a genome, which can be achieved through the widely used P element transposon[340]. This approach would function similarly to enhancer trapping[341], but con-

tain the candidate enhancer sequence in addition to the reporter gene. While the eFS methodology described in Chapter 2 depends on single integration events (otherwise, it is unclear which enhancer is driving GFP expression), the inclusion of unique enhancer barcodes would allow the output from multiple independent insertion events to be resolved. Comparing the results from single and multiple integration methods may also provide useful insights into the dependence of enhancer activity on genomic context.

These improvements could also facilitate the development of eFS-based approaches for use in mammals. In principle, the current experimental approach could be used in transfectable human cell lines that express the φC31 integrase transgene and harbor a genomic attP site. However, one of the main advantages of eFS is its tissue specificity in a fully *in vivo* setting. Therefore, a more significant improvement would be to develop an extension of eFS that functions in mice. While there are significant challenges, it should in principle be possible to deliver tagged eFS constructs into mouse embryonic stem cells using viral vectors such as lentiviruses[342]. These embryonic stem cells could then be used to generate embryonic chimeras, which would contain random integration events[343]. The eFS approach could then be readily applied to sort specific kinds of circulating cells (*e.g.*, subsets of immune cells) based on tissue-specific markers and GFP activity, which could then be subjected to RNA-sequencing for enhancer barcodes. Mouse embryos have already proven to be a useful system for determining the tissue specificity of human enhancers[344], implying the presence of conserved *trans* regulatory architecture across mouse and human development. Developing higher throughput approaches could facilitate the screening of human enhancers as well as the sequence variants found within them.

While all such developments will require solving significant experimental challenges, the computational framework already developed for eFS is likely to remain useful in future iterations. Two fundamental ideas are likely to remain particularly relevant: how to process sequencing data corresponding to the amplified insertion events and how to determine the statistical significance of differences in read counts for candidate enhancers between GFP+/CD2+ and input cells. While incorporating an RNA-

based readout would present new challenges, the same idea of using a statistical framework developed for differential expression analysis is likely to remain relevant. Furthermore, the statistical framework for the enrichment of motifs, combinations of motifs, histone modifications and other functional features is likely to be directly transferrable to new experimental datasets. As increasing numbers of enhancer sequences are identified, it will become possible to evaluate whether more complex predictive models of enhancer activity can improve the accuracy of computational enhancer prediction and provide new biological insights into combinatorial gene regulation.

Within the overall goal of characterizing regulatory variation in humans, developing eFS has demonstrated the feasibility of a new class of experimental approaches for testing enhancers. In turn, this should facilitate advances in enhancer assays that function in mammalian cell lines and tissues. Ultimately, the ability to identify enhancers with high confidence across all cell types and tissues will be a key step towards understanding *cis* regulatory variation. As more complete maps of enhancers are developed, the next goal will be to determine whether specific genetic variants disrupt their activity in tissues that are relevant to various disease processes. Both the methods developed for eFS and the lessons learned during the process are likely to contribute towards achieving these goals.

The work described in Chapter 3 has provided a more complete understanding of the complexity of NF-κB binding and yielded useful insights about the study of *in vivo* TF binding. Importantly, the binding of all five NF-κB subunits was profiled simultaneously for the first time. Despite the similar domain organization and nearly identical DNA-binding domain sequences across subunits, each displayed a unique binding pattern. In aggregate, the subunits formed a complex binding landscape, where a given κB site can be bound by dimers encompassing combinations of anywhere from one to five subunits. Intriguingly, different subunit binding patterns often possessed unique signatures of association with specific biological processes. These observations represent important steps in understanding how different subunits and dimers are used to orchestrate specific transcriptional responses in response to cytosolic signaling events. In addition, they provide a potentially useful framework

for interpreting disparate observations about NF-κB biology, such as the different mouse knockout phenotypes that arise from disruptions in each of the five subunits[98].

Through careful comparisons of NF-κB bound sequences and co-occupying TFs, we identified mechanisms that influence the presence or absence of subunits at certain binding sites. In some cases, these mechanisms depended on the immediate sequence context of the κB site, as in the case of the cytosine preference in the 3' nucleotide. While the same position had been described as changing p50 binding affinity *in vitro*[91], the extent to which the identity of the 3' nucleotide affected RelA:p50 binding *in vivo* was surprising. In addition, we identified several mechanisms that led to the recruitment of specific subunit combinations, but did not involve binding to a κB site. In each of these cases, four TFs and their corresponding motifs were identified as potentially driving the recruitment of specific subunit combinations to DNA. However, additional work is needed to validate these interactions and determine whether recruitment happens through direct interactions or depends on the formation of higher-order complexes.

An equally significant observation is that several differences in binding patterns cannot easily be explained by the sequence of the κB site. The cytosine 3' from the traditional κB site had a significant effect on the recruitment of p50, and to a lesser extent, p52. However, the mechanisms that determine whether RelA, RelB or c-Rel bind a particular motif instance remain less clear. *In vitro* studies of NF-κB binding have not revealed intrinsic differences in the binding specificity of dimers containing these subunits. These results are consistent with the lack of differences in *k*-mer frequencies for sites bound by different subunits *in vivo*. Therefore, it is likely that mechanisms other than sequence specificity play a role in determining NF-κB binding patterns. These may include interactions with other TFs, or effects related to dimer competition at different concentrations and levels of chromatin accessibility. Now that suitable antibodies for all subunits have been identified, further ChIP-Seq experiments are likely to be useful in identifying the mechanisms that control the formation of specific subunit binding patterns. In particular, generating time course data following stimulation of the different pathways

and comparing results across cell lines is likely to be of significant value.

Performing ChIP-Seq experiments for all subunits also provided important insights about TFs that tend to co-occupy regulatory elements with NF-κB. In particular, we found that the cell-cycle regulator FOXM1 was frequently localized with NF-κB at enhancer regions, even in the absence of its preferred motifs. Through further experimental validation, we showed that FOXM1 is recruited to DNA as part of a complex with NF-κB and that the binding of this complex relies on the presence of a κB motif. Furthermore, we were able to show that in diffuse large B-cell lymphomas (DLBCLs), which are often driven by aberrant NF-κB signaling, higher FOXM1 expression was associated with a worse prognosis. Our results suggest that FOXM1 is a putative therapeutic target in DLBCL. In a recent study, this conclusion was reached independently by comparing expression profiles of DLBCL-like tumors in mice[221]. Another study has identified FOXM1 as a therapeutic target in B-cell acute lymphoblastic leukemia[345], suggesting that the joint regulatory role of FOXM1 and NF-κB should be explored further in the context of other cell types and lymphomas.

An essential task will be to re-evaluate how much the newly identified mechanisms contribute to explaining inter-individual variability in NF-κB binding, gene expression and disease risk. A recent study has profiled chromatin state variation in LCLs across 47 individuals, and discovered correlations in TF binding and chromatin state that extend across large regions (~100 kb)[346]. Variants affecting the binding sites of PU.1, one of the TFs identified in association with NF-κB binding in Chapter 3, were found to influence these large scale chromatin states. Fully understanding the mechanisms that create variable NF-κB binding across individuals is likely to require integrating such large-scale factors with local effects mediated by sequence differences. In addition, other TFs are likely to influence NF-κB binding across different cell types. In a recent study, our data were used to identify how a SNP associated with increased risk of developing allergies (rs2370615) affects a RelA binding site in an enhancer regulating the *PAG1* gene[347]. Interestingly, the SNP did not affect a κB motif instance, but rather a predicted binding site for a forkhead TF.

In general, the work described in Chapter 3 has highlighted the value of performing ChIP-Seq studies of human TFs with known roles in gene expression variation and disease risk. We have shown how such data can contribute to the identification of biological mechanisms of medical relevance and how it can be a useful resource for the identification of causal regulatory variants. In addition, we have developed a framework for studying the binding patterns of dimeric TFs and elucidating the mechanisms that contribute to binding differences. As more data about *in vivo* TF binding is generated, such approaches are likely to contribute to identifying additional mechanisms that contribute to regulatory variation.

The combined experimental and computational approach described in Chapter 4 has revealed previously unappreciated genetic diversity in human TFs. Traditionally, mutations that alter the binding properties of TFs have been hypothesized as being highly pleiotropic, and therefore likely to have deleterious effects[113]. However, we found that a typical human genome harbors multiple alleles—often very rare—that encode TFs with likely alterations in their DNA-binding preferences. The associated variants include SNPs that alter base-contacting residues, predicted damaging variants in non-contacting residues, and DBD-truncating nonsense variants. These observations suggest that different individuals may harbor unique *trans* regulatory landscapes that affect gene expression in distinct ways.

The approach described here was focused on DBD variants for two main reasons: (1) the relative ease of testing the effects of DBD mutations *in vitro* and (2) the ability to prioritize variants based on existing structural data. However, such an approach is certain to underestimate the potential for regulatory variation caused by changes in TF amino acid sequences. As discussed in Chapter 1, the ability of TFs to modulate gene expression depends on forming protein-protein interactions (PPIs) with other TFs and components of the transcriptional machinery. In a recent study, certain common nsSNPs were shown to selectively alter the PPI networks of TFs[277]. Such results highlight the different ways in which TF coding variants can affect gene regulation. Integrating the results of different *in vitro* methods, such as PBMs and yeast two-hybrid screens[277], is likely to be helpful in understanding the

full spectrum of coding regulatory variation. In addition, both nsSNPs that impair proper folding of TFs and DBD-truncating nonsense mutations are likely to contribute to regulatory diversity across individuals.

An important area of future work will be to improve the computational methods used to detect affinity and specificity differences based on PBM data. While the current methods were adequate for identifying large changes in DNA binding, comparisons with data generated by other experimental approaches revealed moderate sensitivity for detecting affinity differences. Furthermore, the approach used to detect specificity differences required that the two alleles bind DNA at roughly comparable affinities. Some of these limitations have a biochemical basis: if a mutant allele has lower affinity, the ability to detect specificity changes at sequences bound weakly by the reference allele may be impaired by a greater signal-to-noise ratio. Furthermore, the use of the E-score, a rank-based statistic, was necessary because of its high reproducibility across experiments relative to probe signal intensities and 8-mer z-scores [41]. However, the rank transformation leads to a loss of information about the dynamic range of fluorescence signal intensities, potentially reducing sensitivity for detecting affinity differences and making the readout less quantitative. Future efforts are likely to be aided by a combination of experimental work to increase the reproducibility of z-scores across PBM experiments and computational methods to incorporate the dynamic range information to detect affinity differences. After developing a more quantitative model for detecting affinity differences, it may be possible to jointly model affinity and specificity changes to determine if specificity changes are significant given the predicted affinity difference instead of using a binary cutoff.

Regardless of how TF coding variants are identified, a critical step in understanding how they contribute to gene expression variability will be to assay their *in vivo* effects. Because of the low allele frequencies of many DBDPs that are predicted to have functional consequences (Chapter 4), assaying their effects through eQTL mapping will remain difficult unless cohort sizes are drastically increased. However, various experimental approaches can be used to study the functional consequences of vari-

ants of particular interest. For example, cell lines can be transfected with tagged constructs of different TF alleles, after which ChIP-Seq and RNA-seq can be performed to identify changes in binding and expression, respectively. This approach was used to study the effects of several Mendelian mutations in HOXD13[115]. However, one limitation of this approach is that the concentrations of tagged TFs may not be physiological, and thus any changes in binding and/or expression may be exaggerated. In addition, the endogenous TF genes may need to be knocked down or knocked out in order to assay the effects of the tagged TF without any confounding effects from a functional reference allele.

With the development of increasingly efficient and precise genome editing methods, allelic replacement is likely to become the preferred method for assaying the effects of regulatory variants. When variants are introduced through genome editing, the expression of the TF gene is controlled by the same regulatory elements, mitigating some of the problems of transfecting tagged TFs. However, and perhaps more importantly, genome editing allows multiple DBDPs to be tested at once and enables the study of genetic interactions between regulatory variants. This scenario would be of particular interest when DBDPs occur in multiple genes that are expressed in the same cell type or tissue. In addition, such an approach would allow the high-throughput testing of interactions of *cis* and *trans* variants. For instance, one could perform a high-throughput enhancer assay to identify how SNPs in regulatory sequences change expression and compare the results across cell lines with different genetic backgrounds in terms of TF alleles.

Despite the development of these methods, deciphering potentially complex networks of genetic interactions with regulatory consequences is likely to remain challenging. The large number of variants with predicted functional effects implies that the combinatorial space for the variants present in any individual's genome is enormous. In addition, predicting which noncoding variants may interact genetically with coding variants is likely to be difficult in the absence of functional data, such as ChIP-Seq. The development of improved, deep-learning-based methods for predicting the effects of noncoding variants[348,349] is likely to aid in the prioritization of variants that may be candidates for

cis-trans interactions, particularly since they can be trained on PBM data[348]. Another important challenge will be in determining the downstream phenotypic effects, if any, that are caused by regulatory variants.

Although much remains to be done in understanding the effects of coding variation in TFs, the work described in Chapter 4 represents an important step in establishing the existence of variants that alter TF binding that are segregating in human populations. The framework for prioritizing variants described here is likely to be useful for identifying additional variants of interest for both further experimental testing and for inclusion into eQTL models as potential sources of genetic interactions. In addition, the power of eQTL studies could potentially be boosted by selecting cohorts that harbor certain variants. For instance, the PAX4 variants described in Chapter 4 would be promising targets, given their relatively high minor allele frequencies in Asian populations and known clinical relevance.

In addition, the computational framework developed for this project is likely to be useful for other applications involving the prioritization of genetic variants. One obvious application is in the prediction of potential driver mutations in sequencing datasets from cancer patients. In particular, mutations that alter the specificity of TFs are good candidates for causing gain-of-function effects. Similarly, being able to identify TF mutations with functional effects is likely to be useful in exome sequencing studies performed in trios with an affected child, as it may facilitate the identification of potentially pathogenic *de novo* mutations.

The development and increasing refinement of genome editing technologies has important implications for biology and biomedicine. Large-scale experimental approaches that would have been infeasible just a few years ago, such as the creation of genome-wide gene knockout libraries[350], are increasingly becoming routine. These developments will also facilitate high-throughput screening of genetic variants, such as the ones described in Chapter 4. However, whenever a study design calls for the use of genome editing, one of several technologies must be chosen. Deciding which genome editing approach to use can be difficult, as the pros and cons of various approaches have yet to be fully

evaluated for all applications.

However, one question is likely to remain key in making such decisions: how efficiently and specifically can a particular genome editing technology target a sequence of interest? In order to make such decisions, it is essential to have methods that can predict the on- and off-target effects of each technology as accurately as possible. Developing predictive models for the activity of various genome editing approaches would enable users to select an approach guided by comparisons of optimal performance between such methods at a particular locus of interest.

The development of SIFTED, as described in Chapter 5, has led to the most significant refinement of the "TALE code" to date. By incorporating higher-order effects into the SIFTED model, the specificity of TALE proteins can now often be predicted with improved accuracy over the previous generation of tools. These observations are significant for both for practical applications involving TALEs and for the study of protein-DNA interactions. Because the RVDs in neighboring repeats are not in direct contact with each other, their interdependency is likely to involve more complex allosteric mechanisms. Computational modeling of TALE-DNA interactions has revealed that TALEs undergo significant conformational changes as they compress along the helical axis to contact the bases in their target site[351]. Further work will be useful in revealing whether such conformational effects may contribute to explaining the dependencies between neighboring RVDs.

Nonetheless, the current version of SIFTED model does have certain limitations that should be kept in mind. Most importantly, SIFTED is trained on measurements obtained for TALEs harboring only four types of RVDs: NI, HD, NN, and NG. While this RVD set has been used successfully to target each of the four DNA bases, the inclusion of alternative RVDs may provide optimal specificity in certain contexts. For example, the NH RVD has been reported as being more specific at targeting G than NN[319], which was used in this study. A recent study has significantly expanded the set of known sequence-specific RVDs by comprehensively assaying RVDs not found in naturally occurring TALEs[352]. Incorporating more RVDs into the SIFTED model would enable additional options for

optimizing targeting to a given sequence. In addition, the context effects of RVDs beyond the four already incorporated into SIFTED have not been characterized. Potentially novel context effects could be discovered and used to further optimize the binding of TALEs to a loci of interest and minimize the number of off-targets.

Another current limitation of SIFTED is that its predictive model was trained primarily on shorter TALE proteins. This limitation was primarily the result of experimental considerations, as longer TALE proteins proved more difficult to express by *in vitro* translation and assay in PBMs. While TALEs spanning a range of 8.5 to 18.5 repeats were tested, only two of these were longer than 13.5 RVDs. The SIFTED model was formulated in a way that enables interpolation, and even extrapolation, to predict the specificity of TALEs of lengths that were not assayed. However, generating additional training data corresponding to the higher end of the TALE length spectrum is likely to be of value both for increasing accuracy and for more rigorously validating the predictions made by SIFTED.

Fortunately, the SIFTED model was developed in a way that allows new measurements, both for longer proteins and for additional RVDs, to be easily incorporated into its statistical framework. Efforts to test an expanded set of proteins, incorporating both longer lengths and additional RVDs, are already underway. An updated SIFTED model, trained with this expanded set of proteins, can be readily integrated into the web tool that was released along with the published version of Chapter 5. In turn, this will allow researchers to evaluate the viability and optimality of using TALEs for their particular application with increasing precision.

Despite the performance improvements achieved by SIFTED over previous tools, the predictions of the model do not fully explain the specificity of TALE proteins; *i.e.*, the PBM-derived PWMs still explain probe signal intensity data more accurately than those predicted by SIFTED. With a larger set of PBM data for additional TALEs, it should be possible to explore a larger space of possible context effects, such as models incorporating the contributions of non-adjacent TALE repeats. In addition, if some of the observed context effects are truly allosteric in nature, the additive free energy model em-

ployed by SIFTED may not be a fully adequate approximation of the energy landscape of TALE-DNA interactions. If that is the case, it may be beneficial to perform structural simulations of TALE-DNA interfaces and attempt to isolate additional energetic contributions that arise when TALE repeats with certain combinations of RVDs make contacts with specific adjacent nucleotides. After developing a mechanistic hypothesis based on structural analyses, it may be possible to construct features that incorporate such effects into the statistical framework used by SIFTED.

While the work described in Chapter 5 was underway, a major development in genome editing technology was announced: the discovery and application of the CRISPR-Cas9 system[312]. Cas9 (CRISPR associated protein 9) proteins are RNA-guided endonucleases, which have evolved in bacteria as a mechanism to cleave the DNA of invading viruses. Cas9 can be programmed to introduce DSBs at targeted genomic loci using a "guide RNA" molecule that is complementary to the target DNA sequence[312]. By eliminating the cloning steps required to assemble TALENs, CRISPR-Cas9 has greatly reduced the time and effort required to achieve programmable DNA targeting. Furthermore, CRISPR-Cas9 can easily be multiplexed to facilitate the editing of multiple genomic loci at once.

Admittedly, the discovery of CRISPR-Cas9 has created a formidable competitor for TALEs. Cas9 and its catalytically inactive analog, dCas9, have been used successfully for many of the same applications as TALEs, including mammalian genome editing[353,354] and targeted transcriptional activation[355]. Similarly, fusions between dCas9 and FokI can be used in a manner that is analogous to TALEN pairs[356]. There is little doubt that CRISPR-Cas9 and its variants provide a simple and effective platform for most applications that require programmable DNA-binding proteins. However, TALEs do possess some unique properties over CRISPR-Cas9 that may favor their use in specialized applications. For example, while Cas9 binds to methylated DNA[357], certain TALE RVDs can bind to methylated cytosines (NG) while others (HD) do not[358]. This property is likely to facilitate methylation-dependent targeting of TALE proteins that is not currently feasible with Cas9.

In addition, a major concern for Cas9-based technologies is the presence of widespread off-target

effects. The binding of dCas9 was profiled using ChIP-Seq, identifying hundreds, or sometimes thousands, of off-target binding sites for each guide RNA that was tested[359]. A novel, sequence-based approach to detect DSBs, called GUIDE-seq, was recently used to identify the genomic positions where Cas9 created DSBs[360]. Again, large numbers of off-target sites were observed for multiple guide RNAs, although the numbers varied widely. Interestingly, many off-target sites did not overlap with the regions identified by ChIP-Seq and were not among the highest scoring off-target sites predicted by computational tools. These observations suggest that Cas9, at least in its present form, may not always be the most desirable method for applications that require high specificity. Nevertheless, TALE-based approaches have also been reported to cause off-target effects[137–139]. The work described in Chapter 5 has provided an important new data point about the specificity of TALEs, which should facilitate predictions and subsequent comparisons of off-target activity.

However, the methodology developed along with SIFTED is not limited to TALEs. In principle, the DNA-binding specificity of Cas9 could be similarly assayed *in vitro* using custom PBMs. Such an approach could use either a tagged version of dCas9 or measure the cleavage rates of fluorescently labeled DNA by Cas9. The same statistical approach could then be used to train a model for Cas9 binding that incorporates information about the specificity achieved at different positions in its binding site and how specificity depends on the identity of the adjacent nucleotides in the guide RNA. In practice, it may be necessary to overcome some experimental hurdles. For example, the large size of the commonly used *S. pyogenes* Cas9 may impair binding due to steric effects caused by the proximity of the glass slide and neighboring probes on the array slide. Fortunately, smaller but equally efficient versions of Cas9 are already being identified, such as the one found in *S. aureus*, which is ~350 amino acids shorter than *S. pyogenes* Cas9[361].

In summary, the work described in this dissertation involves multiple significant advances that contribute to the ultimate goal of fully characterizing regulatory variation in humans. In the long run, the work of human geneticists would be greatly aided by a comprehensive, systems-level understanding

of regulatory networks in the human body and how they can be perturbed by genetic variation. Such models may one day allow researchers to input a combination of coding and noncoding variants and trace their effects to specific regulatory sequences, alterations in TF binding sites, and downstream effects on gene expression in specific cell types and developmental stages. With the availability of such tools, it would be significantly easier to generate mechanistic hypotheses about likely causal variants underlying phenotypic associations.

For now, that prospect remains only a dream. Despite significant progress, our maps of regulatory sequences remain incomplete and coarse-grained. There is still much that we do not understand about what causes TFs to bind some genomic sequences but not others—or to bind a given locus in only some individuals. Understanding how the presence of individual coding variants in TFs, let alone possible combinations thereof, may contribute to regulatory variation is likely to require substantial experimental and computational efforts. Even with the increasing feasibility of large-scale genome editing approaches, designing informative experiments and using the resulting data to train more accurate models of gene regulation is likely to require substantial time and ingenuity.

Yet, I believe such efforts will be incredibly worthwhile, given their potential contribution towards understanding both the basis of common disease and the principles of gene regulation. Given the magnitude of the task at hand, one individual's contribution may seem small. However, I find Kenneth Burke's "unending conversation" metaphor to be appropriate here.

> Imagine that you enter a parlor. You come late. When you arrive, others have long preceded you, and they are engaged in a heated discussion, a discussion too heated for them to pause and tell you exactly what it is about. In fact, the discussion had already begun long before any of them got there, so that no one present is qualified to retrace for you all the steps that had gone before. You listen for a while, until you decide that you have caught the tenor of the argument; then you put in your oar. Someone answers; you answer him; another comes to your defense; another aligns himself against you, to either the embarrassment or gratification of your opponent, depending upon the quality of your ally's assistance. However, the discussion is interminable. The hour grows late, you must depart. And you do depart, with the discussion still vigorously in progress.[362]

I am fortunate to have had the opportunity to "put in my oar" at such an exciting time in both human genetics and regulatory genomics. My hope is that the methods, concepts and biological insights that have resulted from the efforts described in this dissertation will prove to be of value in the long road ahead.

References

[1] O. T. Avery, C. M. Macleod, and M. McCarty. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *The Journal of Experimental Medicine*, 79(2):137–158, February 1944.

[2] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, April 1953.

[3] D Botstein, R L White, M Skolnick, and R W Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3):314–331, May 1980.

[4] J. R. Riordan, J. M. Rommens, B. Kerem, N. Alon, R. Rozmahel, Z. Grzelczak, J. Zielenski, S. Lok, N. Plavsic, J. L. Chou, and Et Al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*, 245(4922):1066–1073, September 1989.

[5] Marcy E. MacDonald, Christine M. Ambrose, Mabel P. Duyao, Richard H. Myers, Carol Lin, Lakshmi Srinidhi, Glenn Barnes, Sherryl A. Taylor, Marianne James, Nicolet Groot, Heather MacFarlane, Barbara Jenkins, Mary Anne Anderson, Nancy S. Wexler, James F. Gusella, Gillian P. Bates, Sarah Baxendale, Holger Hummerich, Susan Kirby, Mike North, Sandra Youngman, Richard Mott, Gunther Zehetner, Zdenek Sedlacek, Annemarie Poustka, Anna-Maria Frischauf, Hans Lehrach, Alan J. Buckler, Deanna Church, Lynn Doucette-Stamm, Michael C. O'Donovan, Laura Riba-Ramirez, Manish Shah, Vincent P. Stanton, Scott A. Strobel, Karen M. Draths, Jennifer L. Wales, Peter Dervan, David E. Housman, Michael Altherr, Rita Shiang, Leslie Thompson, Thomas Fielder, John J. Wasmuth, Danilo Tagle, John Valdes, Lawrence Elmer, Marc Allard, Lucio Castilla, Manju Swaroop, Kris Blanchard, Francis S. Collins, Russell Snell, Tracey Holloway, Kathleen Gillespie, Nicole Datson, Duncan Shaw, and Peter S. Harper. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72(6):971–983, March 1993.

[6] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, LaDeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa de la Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R. Cop-

ley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G. R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F. A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, and Michael J. Morgan. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[7] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(Database issue):D1001–D1006, January 2014.

[8] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.

[9] Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099):1190–1195, September 2012.

[10] William S. Bush and Jason H. Moore. Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol*, 8(12):e1002822, December 2012.

[11] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010.

[12] M. W. Nachman and S. L. Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304, September 2000.

[13] Augustine Kong, Michael L. Frigge, Gisli Masson, Soren Besenbacher, Patrick Sulem, Gisli Magnusson, Sigurjon A. Gudjonsson, Asgeir Sigurdsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Wendy S. W. Wong, Gunnar Sigurdsson, G. Bragi Walters, Stacy Steinberg, Hannes Helgason, Gudmar Thorleifsson, Daniel F. Gudbjartsson, Agnar Helgason, Olafur Th Magnusson, Unnur Thorsteinsdottir, and Kari Stefansson. Rate of de novo mutations and the importance of father/'s age to disease risk. *Nature*, 488(7412):471–475, August 2012.

[14] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, January 2001.

[15] David E. Reich, Michele Cargill, Stacey Bolk, James Ireland, Pardis C. Sabeti, Daniel J. Richter, Thomas Lavery, Rose Kouyoumjian, Shelli F. Farhadian, Ryk Ward, and Eric S. Lander. Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204, May 2001.

[16] Mark I. McCarthy, Gonçalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, May 2008.

[17] Jochen Hampe, Andre Franke, Philip Rosenstiel, Andreas Till, Markus Teuber, Klaus Huse, Mario Albrecht, Gabriele Mayr, Francisco M. De La Vega, Jason Briggs, Simone Günther, Natalie J. Prescott, Clive M. Onnie, Robert Häsler, Bence Sipos, Ulrich R. Fölsch, Thomas Lengauer, Matthias Platzer, Christopher G. Mathew, Michael Krawczak, and Stefan Schreiber. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16l1. *Nature Genetics*, 39(2):207–211, February 2007.

[18] Samuel P. Dickson, Kai Wang, Ian Krantz, Hakon Hakonarson, and David B. Goldstein. Rare Variants Create Synthetic Genome-Wide Associations. *PLoS Biol*, 8(1):e1000294, January 2010.

[19] Karen A. Hunt, Vanisha Mistry, Nicholas A. Bockett, Tariq Ahmad, Maria Ban, Jonathan N. Barker, Jeffrey C. Barrett, Hannah Blackburn, Oliver Brand, Oliver Burren, Francesca Capon, Alastair Compston, Stephen C. L. Gough, Luke Jostins, Yong Kong, James C. Lee, Monkol Lek, Daniel G. MacArthur, John C. Mansfield, Christopher G. Mathew, Charles A. Mein, Muddassar Mirza, Sarah Nutland, Suna Onengut-Gumuscu, Efterpi Papouli, Miles Parkes, Stephen S. Rich, Steven Sawcer, Jack Satsangi, Matthew J. Simmonds, Richard C. Trembath, Neil M. Walker, Eva Wozniak, John A. Todd, Michael A. Simpson, Vincent Plagnol, and David A. van Heel. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature*, 498(7453):232–235, June 2013.

[20] Kiran Musunuru, Alanna Strong, Maria Frank-Kamenetsky, Noemi E. Lee, Tim Ahfeldt, Katherine V. Sachs, Xiaoyu Li, Hui Li, Nicolas Kuperwasser, Vera M. Ruda, James P. Pirruccello, Brian Muchmore, Ludmila Prokunina-Olsson, Jennifer L. Hall, Eric E. Schadt, Carlos R. Morales, Sissel Lund-Katz, Michael C. Phillips, Jamie Wong, William Cantley, Timothy Racie, Kenechi G. Ejebe, Marju Orho-Melander, Olle Melander, Victor Koteliansky, Kevin Fitzgerald, Ronald M. Krauss, Chad A. Cowan, Sekar Kathiresan, and Daniel J. Rader. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, 466(7307):714–719, August 2010.

[21] Thomas R. Cech and Joan A. Steitz. The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones. *Cell*, 157(1):77–94, March 2014.

[22] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[23] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal,

Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, Manolis Kellis, and Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, February 2015.

[24] Juan M. Vaquerizas, Sarah K. Kummerfeld, Sarah A. Teichmann, and Nicholas M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263, April 2009.

[25] N. M. Luscombe, S. E. Austin, H. M. Berman, and J. M. Thornton. An overview of the structures of protein-DNA complexes. *Genome Biology*, 1(1):REVIEWS001, 2000.

[26] Michael F. Berger, Gwenael Badis, Andrew R. Gehrke, Shaheynoor Talukder, Anthony A. Philippakis, Lourdes Peña-Castillo, Trevis M. Alleyne, Sanie Mnaimneh, Olga B. Botvinnik, Esther T. Chan, Faiqua Khalid, Wen Zhang, Daniel Newburger, Savina A. Jaeger, Quaid D. Morris, Martha L. Bulyk, and Timothy R. Hughes. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, 133(7):1266–1276, June 2008.

[27] Federico De Masi, Christian A. Grove, Anastasia Vedenko, Andreu Alibés, Stephen S. Gisselbrecht, Luis Serrano, Martha L. Bulyk, and Albertha J. M. Walhout. Using a structural and logics systems approach to infer bHLH-DNA binding specificity determinants. *Nucleic Acids Research*, 39(11):4553–4563, June 2011.

[28] Marcus B. Noyes, Ryan G. Christensen, Atsuya Wakabayashi, Gary D. Stormo, Michael H. Brodsky, and Scot A. Wolfe. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, 133(7):1277–1289, June 2008.

[29] S. A. Wolfe, L. Nekludova, and C. O. Pabo. DNA recognition by Cys2his2 zinc finger proteins. *Annual Review of Biophysics and Biomolecular Structure*, 29:183–212, 2000.

[30] Carl O. Pabo and Lena Nekludova. Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition?1. *Journal of Molecular Biology*, 301(3):597–624, August 2000.

[31] N. C. Seeman, J. M. Rosenberg, and A. Rich. Sequence-specific recognition of double helical nucleic acids by proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 73(3):804–808, March 1976.

[32] J. C. Miller and C. O. Pabo. Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. *Journal of Molecular Biology*, 313(2):309–315, October 2001.

[33] Trevor Siggers and Raluca Gordân. Protein–DNA binding: complexities and multi-protein codes. *Nucleic Acids Research*, 42(4):2099–2111, February 2014.

[34] François Spitz and Eileen E. M. Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, September 2012.

[35] Mark M. Garner and Arnold Revzin. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Research*, 9(13):3047–3060, July 1981.

[36] P. Schuck. Use of surface plasmon resonance to probe the equilibrium and dynamic aspects of interactions between biological macromolecules. *Annual Review of Biophysics and Biomolecular Structure*, 26:541–566, 1997.

[37] Sebastian J. Maerkl and Stephen R. Quake. A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. *Science*, 315(5809):233–237, January 2007.

[38] Daniel E. Newburger and Martha L. Bulyk. UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. *Nucleic Acids Research*, 37(Database issue):D77–D82, January 2009.

[39] Matthew T. Weirauch, Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R. Riley, Julio Saez-Rodriguez, Thomas Cokelaer, Anastasia Vedenko, Shaheynoor Talukder, Dream5 Consortium, Harmen J. Bussemaker, Quaid D. Morris, Martha L. Bulyk, Gustavo Stolovitzky, and Timothy R. Hughes. Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31(2):126–134, February 2013.

[40] Panayiotis V. Benos, Martha L. Bulyk, and Gary D. Stormo. Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Research*, 30(20):4442–4451, October 2002.

[41] Michael F. Berger and Martha L. Bulyk. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature Protocols*, 4(3):393–411, March 2009.

[42] Stuart Geman, Elie Bienenstock, and René Doursat. Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4(1):1–58, January 1992.

[43] Boris Lenhard, Albin Sandelin, and Piero Carninci. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, 13(4):233–245, April 2012.

[44] Daria Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286, April 2014.

[45] M.S. Halfon. Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors. *Cell*, 103:63–74, 2000.

[46] D. Stanojevic, S. Small, and M. Levine. Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science*, 254(5036):1385–1387, November 1991.

[47] Roy Joseph, Yuriy L Orlov, Mikael Huss, Wenjie Sun, Say Li Kong, Leena Ukil, You Fu Pan, Guoliang Li, Michael Lim, Jane S Thomsen, Yijun Ruan, Neil D Clarke, Shyam Prabhakar, Edwin Cheung, and Edison T Liu. Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. *Molecular Systems Biology*, 6, December 2010.

[48] Tommy Kaplan, Xiao-Yong Li, Peter J. Sabo, Sean Thomas, John A. Stamatoyannopoulos, Mark D. Biggin, and Michael B. Eisen. Quantitative Models of the Mechanisms That Control Genome-Wide Patterns of Transcription Factor Binding during Early Drosophila Development. *PLoS Genet*, 7(2):e1001290, February 2011.

[49] Leonid A. Mirny. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences of the United States of America*, 107(52):22534–22539, December 2010.

[50] K.S. Zaret and J.S. Carroll. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.*, 25:2227–2241, 2011.

[51] Daniel Panne, Tom Maniatis, and Stephen C. Harrison. An atomic model of the interferon-beta enhanceosome. *Cell*, 129(6):1111–1123, June 2007.

[52] Alexander V. Lukashin and Mark Borodovsky. GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Research*, 26(4):1107–1115, February 1998.

[53] Sanja Rogic, Alan K. Mackworth, and Francis B. F. Ouellette. Evaluation of Gene-Finding Programs on Mammalian Sequences. *Genome Research*, 11(5):817–832, May 2001.

[54] M. I. Arnone and E. H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–1864, May 1997.

[55] Benjamin P. Berman, Yutaka Nibu, Barret D. Pfeiffer, Pavel Tomancak, Susan E. Celniker, Michael Levine, Gerald M. Rubin, and Michael B. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):757–762, January 2002.

[56] J. Warner. Systematic identification of mammalian regulatory motifs' target genes and functions. *Nat. Methods*, 5:347–353, 2008.

[57] Outi Hallikas, Kimmo Palin, Natalia Sinjushina, Reetta Rautiainen, Juha Partanen, Esko Ukkonen, and Jussi Taipale. Genome-wide Prediction of Mammalian Enhancers Based on Analysis of Transcription-Factor Binding Affinity. *Cell*, 124(1):47–59, January 2006.

[58] Menie Merika and Dimitris Thanos. Enhanceosomes. *Current Opinion in Genetics & Development*, 11(2):205–208, April 2001.

[59] David N. Arnosti and Meghana M. Kulkarni. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry*, 94(5):890–898, April 2005.

[60] Michael A. White, Connie A. Myers, Joseph C. Corbo, and Barak A. Cohen. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proceedings of the National Academy of Sciences of the United States of America*, 110(29):11952–11957, July 2013.

[61] Pouya Kheradpour, Alexander Stark, Sushmita Roy, and Manolis Kellis. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Research*, 17(12):1919–1931, December 2007.

[62] J. Omar Yanez-Cuna, Huy Q. Dinh, Evgeny Z. Kvon, Daria Shlyueva, and Alexander Stark. Uncovering cis-regulatory sequence requirements for context specific transcription factor binding. *Genome Research*, page gr.132811.111, April 2012.

[63] Bing Ren, François Robert, John J. Wyrick, Oscar Aparicio, Ezra G. Jennings, Itamar Simon, Julia Zeitlinger, Jörg Schreiber, Nancy Hannett, Elenita Kanin, Thomas L. Volkert, Christopher J. Wilson, Stephen P. Bell, and Richard A. Young. Genome-Wide Location and Function of DNA Binding Proteins. *Science*, 290(5500):2306–2309, December 2000.

[64] R.P. Zinzen, C. Girardot, J. Gagneur, M. Braun, and E.E. Furlong. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462:65–70, 2009.

[65] William W. Fisher, Jingyi Jessica Li, Ann S. Hammonds, James B. Brown, Barret D. Pfeiffer, Richard Weiszmann, Stewart MacArthur, Sean Thomas, John A. Stamatoyannopoulos, Michael B. Eisen, Peter J. Bickel, Mark D. Biggin, and Susan E. Celniker. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 109(52):21330–21335, December 2012.

[66] Xiao-yong Li, Stewart MacArthur, Richard Bourgon, David Nix, Daniel A Pollard, Venky N Iyer, Aaron Hechmer, Lisa Simirenko, Mark Stapleton, Cris L. Luengo Hendriks, Hou Cheng Chu, Nobuo Ogawa, William Inwood, Victor Sementchenko, Amy Beaton, Richard Weiszmann, Susan E Celniker, David W Knowles, Tom Gingeras, Terence P Speed, Michael B Eisen, and Mark D Biggin. Transcription Factors Bind Thousands of Active and Inactive Regions in the Drosophila Blastoderm. *PLoS Biol*, 6(2):e27, February 2008.

[67] David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, 316(5830):1497–1502, June 2007.

[68] Jonathan Göke, Marc Jung, Sarah Behrens, Lukas Chavez, Sean O'Keeffe, Bernd Timmermann, Hans Lehrach, James Adjaye, and Martin Vingron. Combinatorial Binding in Human and Mouse Embryonic Stem Cells Identifies Conserved Enhancers Active in Early Embryonic Development. *PLoS Computational Biology*, 7(12), December 2011.

[69] Axel Visel, Matthew J. Blow, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Feng Chen, Veena Afzal, Bing Ren, Edward M. Rubin, and Len A. Pennacchio. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858, February 2009.

[70] N.D. Heintzman. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459:108–112, 2009.

[71] Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Noam Shoresh, Lucas D. Ward, Charles B. Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis, and Bradley E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, May 2011.

[72] Robert E. Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T. Maurano, Eric Haugen, Nathan C. Sheffield, Andrew B. Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K. Canfield, Morgan Diegel, Douglas Dunn, Abigail K. Ebersol, Tristan Frum, Erika Giste, Audra K. Johnson, Ericka M. Johnson, Tanya Kutyavin, Bryan Lajoie, Bum-Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D. Nguyen, Hongzhu Qu, Alex P. Reynolds, Vaughn Roach, Alexias Safi, Minerva E. Sanchez, Amartya Sanyal, Anthony Shafer, Jeremy M. Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari, Michael O. Dorschner, R. Scott Hansen, Patrick A. Navas, George Stamatoyannopoulos, Vishwanath R. Iyer, Jason D. Lieb, Shamil R. Sunyaev, Joshua M. Akey, Peter J. Sabo, Rajinder Kaul, Terrence S. Furey, Job Dekker, Gregory E. Crawford, and John A. Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, September 2012.

[73] Masato Enari, Hideki Sakahira, Hideki Yokoyama, Katsuya Okawa, Akihiro Iwamatsu, and Shigekazu Nagata. A caspase-activated DNase that degrades DNA during apoptosis, and its inhibitor ICAD. *Nature*, 391(6662):43–50, January 1998.

[74] Alan P. Boyle, Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*, 132(2):311–322, January 2008.

[75] Hualin Xi, Hennady P Shulha, Jane M Lin, Teresa R Vales, Yutao Fu, David M Bodine, Ronald D. G McKay, Josh G Chenoweth, Paul J Tesar, Terrence S Furey, Bing Ren, Zhiping Weng, and Gregory E Crawford. Identification and Characterization of Cell Type–Specific and Ubiquitous Chromatin Regulatory Structures in the Human Genome. *PLoS Genet*, 3(8):e136, August 2007.

[76] C.D. Arnold. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339:1074–1077, 2013.

[77] Michael M. Hoffman, Orion J. Buske, Jie Wang, Zhiping Weng, Jeff A. Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–476, May 2012.

[78] Pouya Kheradpour, Jason Ernst, Alexandre Melnikov, Peter Rogov, Li Wang, Xiaolan Zhang, Jessica Alston, Tarjei S. Mikkelsen, and Manolis Kellis. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, 23(5):800–811, May 2013.

[79] Gregor D. Gilfillan, Timothy Hughes, Ying Sheng, Hanne S. Hjorthaug, Tobias Straub, Kristina Gervin, Jennifer R. Harris, Dag E. Undlien, and Robert Lyle. Limitations and possibilities of low cell number ChIP-seq. *BMC genomics*, 13:645, 2012.

[80] N.D. Heintzman. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, 39:311–318, 2007.

[81] Alexandre Melnikov, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, Andreas Gnirke, Curtis G. Callan, Justin B. Kinney, Manolis Kellis, Eric S. Lander, and Tarjei S. Mikkelsen. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, 30(3):271–277, March 2012.

[82] Melina Claussnitzer, Simon N. Dankel, Bernward Klocke, Harald Grallert, Viktoria Glunk, Tea Berulava, Heekyoung Lee, Nikolay Oskolkov, Joao Fadista, Kerstin Ehlers, Simone Wahl, Christoph Hoffmann, Kun Qian, Tina Rönn, Helene Riess, Martina Müller-Nurasyid, Nancy Bretschneider, Timm Schroeder, Thomas Skurk, Bernhard Horsthemke, Derek Spieler, Martin Klingenspor, Martin Seifert, Michael J. Kern, Niklas Mejhert, Ingrid Dahlman, Ola Hansson, Stefanie M. Hauck, Matthias Blüher, Peter Arner, Leif Groop, Thomas Illig, Karsten Suhre, Yi-Hsiang Hsu, Gunnar Mellgren, Hans Hauner, Helmut Laumen, Benjamin F. Voight, Laura J. Scott, Valgerdur Steinthorsdottir, Andrew P. Morris, Christian Dina, Ryan P. Welch, Eleftheria Zeggini, Cornelia Huth, Yurii S. Aulchenko, Gudmar Thorleifsson, Laura J. McCulloch, Teresa Ferreira, Harald Grallert, Najaf Amin, Guanming Wu, Cristen J. Willer, Soumya Raychaudhuri, Steve A. McCarroll, Claudia Langenberg, Oliver M. Hofmann, Josée Dupuis, Lu Qi, Ayellet V. Segrè, Mandy van Hoek, Pau Navarro, Kristin Ardlie, Beverley

Balkau, Rafn Benediktsson, Amanda J. Bennett, Roza Blagieva, Eric Boerwinkle, Lori L. Bonnycastle, Kristina Bengtsson Boström, Bert Bravenboer, Suzannah Bumpstead, Noël P. Burtt, Guillaume Charpentier, Peter S. Chines, Marilyn Cornelis, David J. Couper, Gabe Crawford, Alex S. F. Doney, Katherine S. Elliott, Amanda L. Elliott, Michael R. Erdos, Caroline S. Fox, Christopher S. Franklin, Martha Ganser, Christian Gieger, Niels Grarup, Todd Green, Simon Griffin, Christopher J. Groves, Candace Guiducci, Samy Hadjadj, Neelam Hassanali, Christian Herder, Bo Isomaa, Anne U. Jackson, Paul R. V. Johnson, Torben Jørgensen, Wen H. L. Kao, Norman Klopp, Augustine Kong, Peter Kraft, Johanna Kuusisto, Torsten Lauritzen, Man Li, Aloysius Lieverse, Cecilia M. Lindgren, Valeriya Lyssenko, Michel Marre, Thomas Meitinger, Kristian Midthjell, Mario A. Morken, Narisu Narisu, Peter Nilsson, Katharine R. Owen, Felicity Payne, John R. B. Perry, Ann-Kristin Petersen, Carl Platou, Christine Proença, Inga Prokopenko, Wolfgang Rathmann, N. William Rayner, Neil R. Robertson, Ghislain Rocheleau, Michael Roden, Michael J. Sampson, Richa Saxena, Beverley M. Shields, Peter Shrader, Gunnar Sigurdsson, Thomas Sparsø, Klaus Strassburger, Heather M. Stringham, Qi Sun, Amy J. Swift, Barbara Thorand, Jean Tichet, Tiinamaija Tuomi, Rob M. van Dam, Timon W. van Haeften, Thijs van Herpt, Jana V. van Vliet-Ostaptchouk, G. Bragi Walters, Michael N. Weedon, Cisca Wijmenga, Jacqueline Witteman, Richard N. Bergman, Stephane Cauchi, Francis S. Collins, Anna L. Gloyn, Ulf Gyllensten, Torben Hansen, Winston A. Hide, Graham A. Hitman, Albert Hofman, David J. Hunter, Kristian Hveem, Markku Laakso, Karen L. Mohlke, Andrew D. Morris, Colin N. A. Palmer, Peter P. Pramstaller, Igor Rudan, Eric Sijbrands, Lincoln D. Stein, Jaakko Tuomilehto, Andre Uitterlinden, Mark Walker, Nicholas J. Wareham, Richard M. Watanabe, Goncalo R. Abecasis, Bernhard O. Boehm, Harry Campbell, Mark J. Daly, Andrew T. Hattersley, Frank B. Hu, James B. Meigs, James S. Pankow, Oluf Pedersen, H.-Erich Wichmann, Inês Barroso, Jose C. Florez, Timothy M. Frayling, Leif Groop, Rob Sladek, Unnur Thorsteinsdottir, James F. Wilson, Thomas Illig, Philippe Froguel, Cornelia M. van Duijn, Kari Stefansson, David Altshuler, Michael Boehnke, and Mark I. McCarthy. Leveraging Cross-Species Transcription Factor Binding Site Patterns: From Diabetes Risk Loci to Disease Mechanisms. *Cell*, 156(1):343–358, January 2014.

[83] Qilai Huang, Thomas Whitington, Ping Gao, Johan F. Lindberg, Yuehong Yang, Jielin Sun, Marja-Riitta Väisänen, Robert Szulkin, Matti Annala, Jian Yan, Lars A. Egevad, Kai Zhang, Ruizhu Lin, Arttu Jolma, Matti Nykter, Aki Manninen, Fredrik Wiklund, Markku H. Vaarala, Tapio Visakorpi, Jianfeng Xu, Jussi Taipale, and Gong-Hong Wei. A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding.

Nature Genetics, 46(2):126–135, February 2014.

[84] Melina Claussnitzer, Simon N. Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S. Sousa, Jacqueline L. Beaudry, Vijitha Puviindran, Nezar A. Abdennur, Jannel Liu, Per-Arne Svensson, Yi-Hsiang Hsu, Daniel J. Drucker, Gunnar Mellgren, Chi-Chung Hui, Hans Hauner, and Manolis Kellis. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *New England Journal of Medicine*, 0(0):null, August 2015.

[85] Shaofeng Wang, Feng Wen, Graham B. Wiley, Michael T. Kinter, and Patrick M. Gaffney. An Enhancer Element Harboring Variants Associated with Systemic Lupus Erythematosus Engages the TNFAIP3 Promoter to Influence A20 Expression. *PLoS Genet*, 9(9):e1003750, September 2013.

[86] Maya Kasowski, Fabian Grubert, Christopher Heffelfinger, Manoj Hariharan, Akwasi Asabere, Sebastian M. Waszak, Lukas Habegger, Joel Rozowsky, Minyi Shi, Alexander E. Urban, Mi-Young Hong, Konrad J. Karczewski, Wolfgang Huber, Sherman M. Weissman, Mark B. Gerstein, Jan O. Korbel, and Michael Snyder. Variation in transcription factor binding among humans. *Science (New York, N.Y.)*, 328(5975):232–235, April 2010.

[87] Timothy E. Reddy, Jason Gertz, Florencia Pauli, Katerina S. Kucera, Katherine E. Varley, Kimberly M. Newberry, Georgi K. Marinov, Ali Mortazavi, Brian A. Williams, Lingyun Song, Gregory E. Crawford, Barbara Wold, Huntington F. Willard, and Richard M. Myers. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Research*, 22(5):860–869, May 2012.

[88] Ryan McDaniell, Bum-Kyu Lee, Lingyun Song, Zheng Liu, Alan P. Boyle, Michael R. Erdos, Laura J. Scott, Mario A. Morken, Katerina S. Kucera, Anna Battenhouse, Damian Keefe, Francis S. Collins, Huntington F. Willard, Jason D. Lieb, Terrence S. Furey, Gregory E. Crawford, Vishwanath R. Iyer, and Ewan Birney. Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans. *Science*, 328(5975):235–239, April 2010.

[89] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, 38(4):576–589, May 2010.

[90] Matthew S. Hayden and Sankar Ghosh. NF-κB, the first quarter-century: remarkable progress and outstanding questions. *Genes & Development*, 26(3):203–234, February 2012.

[91] Trevor Siggers, Abraham B. Chang, Ana Teixeira, Daniel Wong, Kevin J. Williams, Bilal Ahmed, Jiannis Ragoussis, Irina A. Udalova, Stephen T. Smale, and Martha L. Bulyk. Principles of dimer-specific gene regulation revealed by a comprehensive characterization of NF-κB family DNA binding. *Nature Immunology*, 13(1):95–102, January 2012.

[92] Robin E. C. Lee, Sarah R. Walker, Kate Savery, David A. Frank, and Suzanne Gaudet. Fold Change of Nuclear NF-κB Determines TNF-Induced Transcription in Single Cells. *Molecular Cell*, 53(6):867–879, March 2014.

[93] Thomas H. Leung, Alexander Hoffmann, and David Baltimore. One nucleotide in a kappaB site can determine cofactor specificity for NF-kappaB dimers. *Cell*, 118(4):453–464, August 2004.

[94] Vivien Ya-Fan Wang, Wendy Huang, Masataka Asagiri, Nathanael Spann, Alexander Hoffmann, Christopher Glass, and Gourisankar Ghosh. The Transcriptional Specificity of NF-κB Dimers Is Coded within the κB DNA Response Elements. *Cell Reports*, 2(4):824–839, October 2012.

[95] Paul P. Tak and Gary S. Firestein. NF-κB: a key role in inflammatory diseases. *Journal of Clinical Investigation*, 107(1):7–11, January 2001.

[96] Michael Karin. NF-kappaB as a critical link between inflammation and cancer. *Cold Spring Harbor Perspectives in Biology*, 1(5):a000141, November 2009.

[97] T. D. Gilmore. Introduction to NF-kappaB: players, pathways, perspectives. *Oncogene*, 25(51):6680–6684, October 2006.

[98] S. Gerondakis, R. Grumont, R. Gugasyan, L. Wong, I. Isomura, W. Ho, and A. Banerjee. Unravelling the complexities of the NF-κB signalling pathway using mouse knockout and transgenic models. *Oncogene*, 25(51):6781–6799, 2006.

[99] Vincent Feng-Sheng Shih, Rachel Tsui, Andrew Caldwell, and Alexander Hoffmann. A single NFκB system for both canonical and non-canonical signaling. *Cell Research*, 21(1):86–102, January 2011.

[100] Shao-Cong Sun. The noncanonical NF-κB pathway. *Immunological Reviews*, 246(1):125–140, March 2012.

[101] L. Lin, G. N. DeMartino, and W. C. Greene. Cotranslational biogenesis of NF-kappaB p50 by the 26s proteasome. *Cell*, 92(6):819–828, March 1998.

[102] G. Xiao, E. W. Harhaj, and S. C. Sun. NF-kappaB-inducing kinase regulates the processing of NF-kappaB2 p100. *Molecular Cell*, 7(2):401–409, February 2001.

[103] Heidemarie Neitzel. A routine method for the establishment of permanent growing lymphoblastoid cell lines. *Human Genetics*, 73(4):320–326, August 1986.

[104] Hector Ardila-Osorio, Bernard Clausse, Zohair Mishal, Joëlle Wiels, Thomas Tursz, and Pierre Busson. Evidence of LMP1–TRAF3 interactions in glycosphingolipid-rich complexes of lymphoblastoid and nasopharyngeal carcinoma cells. *International Journal of Cancer*, 81(4):645–649, May 1999.

[105] Alkes L. Price, Agnar Helgason, Gudmar Thorleifsson, Steven A. McCarroll, Augustine Kong, and Kari Stefansson. Single-Tissue and Cross-Tissue Heritability of Gene Expression Via Identity-by-Descent in Related or Unrelated Individuals. *PLoS Genet*, 7(2):e1001317, February 2011.

[106] Daniel J. Gaffney. Global Properties and Functional Complexity of Human Gene Regulatory Variation. *PLoS Genet*, 9(5):e1003501, May 2013.

[107] Elin Grundberg, Kerrin S. Small, Åsa K. Hedman, Alexandra C. Nica, Alfonso Buil, Sarah Keildson, Jordana T. Bell, Tsun-Po Yang, Eshwar Meduri, Amy Barrett, James Nisbett, Magdalena Sekowska, Alicja Wilk, So-Youn Shin, Daniel Glass, Mary Travers, Josine L. Min, Sue Ring, Karen Ho, Gudmar Thorleifsson, Augustine Kong, Unnur Thorsteindottir, Chrysanthi Ainali, Antigone S. Dimas, Neelam Hassanali, Catherine Ingle, David Knowles, Maria Krestyaninova, Christopher E. Lowe, Paola Di Meglio, Stephen B. Montgomery, Leopold Parts, Simon Potter, Gabriela Surdulescu, Loukia Tsaprouni, Sophia Tsoka, Veronique Bataille, Richard Durbin, Frank O. Nestle, Stephen O'Rahilly, Nicole Soranzo, Cecilia M. Lindgren, Krina T. Zondervan, Kourosh R. Ahmadi, Eric E. Schadt, Kari Stefansson, George Davey Smith, Mark I. McCarthy, Panos Deloukas, Emmanouil T. Dermitzakis, Tim D. Spector, and The Multiple Tissue Human Expression Resource (MuTHER) Consortium. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, 44(10):1084–1089, October 2012.

[108] Alkes L. Price, Nick Patterson, Dustin C. Hancks, Simon Myers, David Reich, Vivian G. Cheung, and Richard S. Spielman. Effects of cis and trans genetic ancestry on gene expression in African Americans. *PLoS genetics*, 4(12):e1000294, December 2008.

[109] Daniel A. Skelly, James Ronald, and Joshua M. Akey. Inherited Variation in Gene Expression. *Annual Review of Genomics and Human Genetics*, 10(1):313–332, 2009.

[110] Harm-Jan Westra, Marjolein J. Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W. Christiansen, Benjamin P. Fairfax, Katharina Schramm, Joseph E. Powell, Alexandra Zhernakova, Daria V. Zhernakova, Jan H. Veldink, Leonard H. Van den Berg, Juha Karjalainen, Sebo Withoff, André G. Uitterlinden, Albert Hofman, Fernando Rivadeneira, Peter A. C. 't Hoen, Eva Reinmaa, Krista Fischer, Mari Nelis, Lili Milani, David Melzer, Luigi Ferrucci, Andrew B. Singleton, Dena G. Hernandez, Michael A. Nalls, Georg Homuth, Matthias Nauck, Dörte Radke, Uwe Völker, Markus Perola, Veikko Salomaa, Jennifer Brody, Astrid Suchy-Dicey, Sina A. Gharib, Daniel A. Enquobahrie, Thomas Lumley, Grant W. Montgomery, Seiko Makino, Holger Prokisch, Christian Herder, Michael Roden, Harald Grallert, Thomas Meitinger, Konstantin Strauch, Yang Li, Ritsert C. Jansen, Peter M. Visscher, Julian C. Knight, Bruce M. Psaty, Samuli Ripatti, Alexander Teumer, Timothy M. Frayling, Andres Metspalu, Joyce B. J. van Meurs, and Lude Franke. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243, October 2013.

[111] Oliver Stegle, Leopold Parts, Richard Durbin, and John Winn. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. *PLoS Comput Biol*, 6(5):e1000770, May 2010.

[112] Zuqin Nie, Gangqing Hu, Gang Wei, Kairong Cui, Arito Yamane, Wolfgang Resch, Ruoning Wang, Douglas R. Green, Lino Tessarollo, Rafael Casellas, Keji Zhao, and David Levens. c-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells. *Cell*, 151(1):68–79, September 2012.

[113] Hopi E. Hoekstra and Jerry A. Coyne. The locus of evolution: evo devo and the genetics of adaptation. *Evolution; International Journal of Organic Evolution*, 61(5):995–1016, May 2007.

[114] Daniel G. MacArthur, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, Lukas Habegger, Joseph K. Pickrell, Stephen B. Montgomery, Cornelis A. Albers, Zhengdong D. Zhang, Donald F. Conrad, Gerton Lunter, Hancheng

Zheng, Qasim Ayub, Mark A. DePristo, Eric Banks, Min Hu, Robert E. Handsaker, Jeffrey A. Rosenfeld, Menachem Fromer, Mike Jin, Xinmeng Jasmine Mu, Ekta Khurana, Kai Ye, Mike Kay, Gary Ian Saunders, Marie-Marthe Suner, Toby Hunt, If H. A. Barnes, Clara Amid, Denise R. Carvalho-Silva, Alexandra H. Bignell, Catherine Snow, Bryndis Yngvadottir, Suzannah Bumpstead, David N. Cooper, Yali Xue, Irene Gallego Romero, Jun Wang, Yingrui Li, Richard A. Gibbs, Steven A. McCarroll, Emmanouil T. Dermitzakis, Jonathan K. Pritchard, Jeffrey C. Barrett, Jennifer Harrow, Matthew E. Hurles, Mark B. Gerstein, and Chris Tyler-Smith. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science*, 335(6070):823–828, February 2012.

[115] Daniel M. Ibrahim, Peter Hansen, Christian Rödelsperger, Asita C. Stiege, Sandra C. Doelken, Denise Horn, Marten Jäger, Catrin Janetzki, Peter Krawitz, Gundula Leschik, Florian Wagner, Till Scheuer, Mareen Schmidt-von Kegler, Petra Seemann, Bernd Timmermann, Peter N. Robinson, Stefan Mundlos, and Jochen Hecht. Distinct global shifts in genomic binding profiles of limb malformation-associated HOXD13 mutations. *Genome Research*, 23(12):2091–2102, December 2013.

[116] GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (New York, N.Y.)*, 348(6235):648–660, May 2015.

[117] R P Moerschell, S Tsunasawa, and F Sherman. Transformation of yeast with synthetic oligonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, 85(2):524–528, January 1988.

[118] Kirk R. Thomas and Mario R. Capecchi. Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. *Cell*, 51(3):503–512, November 1987.

[119] Jeffry D. Sander and J. Keith Joung. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology*, 32(4):347–355, April 2014.

[120] David Benjamin Turitz Cox, Randall Jeffrey Platt, and Feng Zhang. Therapeutic genome editing: prospects and challenges. *Nature Medicine*, 21(2):121–131, February 2015.

[121] Julianne Smith, Sylvestre Grizot, Sylvain Arnould, Aymeric Duclert, Jean-Charles Epinat, Patrick Chames, Jesús Prieto, Pilar Redondo, Francisco J. Blanco, Jerónimo Bravo, Guillermo

Montoya, Frédéric Pâques, and Philippe Duchateau. A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. *Nucleic Acids Research*, 34(22):e149, 2006.

[122] Y. G. Kim, J. Cha, and S. Chandrasegaran. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proceedings of the National Academy of Sciences of the United States of America*, 93(3):1156–1160, February 1996.

[123] Fyodor D. Urnov, Edward J. Rebar, Michael C. Holmes, H. Steve Zhang, and Philip D. Gregory. Genome editing with engineered zinc finger nucleases. *Nature Reviews Genetics*, 11(9):636–646, September 2010.

[124] Keyu Gu, Bing Yang, Dongsheng Tian, Lifang Wu, Dongjiang Wang, Chellamma Sreekala, Fan Yang, Zhaoqing Chu, Guo-Liang Wang, Frank F. White, and Zhongchao Yin. R gene expression induced by a type-III effector triggers disease resistance in rice. *Nature*, 435(7045):1122–1125, June 2005.

[125] Sebastian Schornack, Annett Meyer, Patrick Römer, Tina Jordan, and Thomas Lahaye. Gene-for-gene-mediated recognition of nuclear-targeted AvrBs3-like bacterial effector proteins. *Journal of Plant Physiology*, 163(3):256–272, February 2006.

[126] Sabine Kay, Simone Hahn, Eric Marois, Gerd Hause, and Ulla Bonas. A bacterial effector acts as a plant transcription factor and induces a cell size regulator. *Science (New York, N.Y.)*, 318(5850):648–651, October 2007.

[127] Stefano Stella, Rafael Molina, Igor Yefimenko, Jesús Prieto, George Silva, Claudia Bertonati, Alexandre Juillerat, Phillippe Duchateau, and Guillermo Montoya. Structure of the AvrBs3–DNA complex provides new insights into the initial thymine-recognition mechanism. *Acta Crystallographica Section D Biological Crystallography*, 69(9):1707–1716, September 2013.

[128] Jens Boch, Heidi Scholze, Sebastian Schornack, Angelika Landgraf, Simone Hahn, Sabine Kay, Thomas Lahaye, Anja Nickstadt, and Ulla Bonas. Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. *Science*, 326(5959):1509–1512, December 2009.

[129] Matthew J. Moscou and Adam J. Bogdanove. A Simple Cipher Governs DNA Recognition by TAL Effectors. *Science*, 326(5959):1501–1501, December 2009.

[130] Deepak Reyon, Shengdar Q. Tsai, Cyd Khayter, Jennifer A. Foden, Jeffry D. Sander, and J. Keith Joung. FLASH assembly of TALENs for high-throughput genome editing. *Nature Biotechnology*, 30(5):460–465, May 2012.

[131] Jeffry D. Sander, Lindsay Cade, Cyd Khayter, Deepak Reyon, Randall T. Peterson, J. Keith Joung, and Jing-Ruey J. Yeh. Targeted gene disruption in somatic zebrafish cells using engineered TALENs. *Nature Biotechnology*, 29(8):697–698, August 2011.

[132] Morgan L. Maeder, Samantha J. Linder, Deepak Reyon, James F. Angstman, Yanfang Fu, Jeffry D. Sander, and J. Keith Joung. Robust, synergistic regulation of human gene expression using TALE activators. *Nature Methods*, 10(3):243–245, March 2013.

[133] Morgan L. Maeder, James F. Angstman, Marcy E. Richardson, Samantha J. Linder, Vincent M. Cascio, Shengdar Q. Tsai, Quan H. Ho, Jeffry D. Sander, Deepak Reyon, Bradley E. Bernstein, Joseph F. Costello, Miles F. Wilkinson, and J. Keith Joung. Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nature Biotechnology*, 31(12):1137–1142, December 2013.

[134] Eric M. Mendenhall, Kaylyn E. Williamson, Deepak Reyon, James Y. Zou, Oren Ram, J. Keith Joung, and Bradley E. Bernstein. Locus-specific editing of histone modifications at endogenous enhancers. *Nature Biotechnology*, 31(12):1133–1136, December 2013.

[135] Jeffrey C. Miller, Siyuan Tan, Guijuan Qiao, Kyle A. Barlow, Jianbin Wang, Danny F. Xia, Xiangdong Meng, David E. Paschon, Elo Leung, Sarah J. Hinkley, Gladys P. Dulay, Kevin L. Hua, Irina Ankoudinova, Gregory J. Cost, Fyodor D. Urnov, H. Steve Zhang, Michael C. Holmes, Lei Zhang, Philip D. Gregory, and Edward J. Rebar. A TALE nuclease architecture for efficient genome editing. *Nature Biotechnology*, 29(2):143–148, February 2011.

[136] Dirk Hockemeyer, Haoyi Wang, Samira Kiani, Christine S. Lai, Qing Gao, John P. Cassady, Gregory J. Cost, Lei Zhang, Yolanda Santiago, Jeffrey C. Miller, Bryan Zeitler, Jennifer M. Cherone, Xiangdong Meng, Sarah J. Hinkley, Edward J. Rebar, Philip D. Gregory, Fyodor D. Urnov, and Rudolf Jaenisch. Genetic engineering of human pluripotent cells using TALE nucleases. *Nature Biotechnology*, 29(8):731–734, August 2011.

[137] Qiurong Ding, Youn-Kyoung Lee, Esperance A. K. Schaefer, Derek T. Peters, Adrian Veres, Kevin Kim, Nicolas Kuperwasser, Daniel L. Motola, Torsten B. Meissner, William T. Hendriks, Marta Trevisan, Rajat M. Gupta, Annie Moisan, Eric Banks, Max Friesen, Robert T.

Schinzel, Fang Xia, Alexander Tang, Yulei Xia, Emmanuel Figueroa, Amy Wann, Tim Ahfeldt, Laurence Daheron, Feng Zhang, Lee L. Rubin, Lee F. Peng, Raymond T. Chung, Kiran Musunuru, and Chad A. Cowan. A TALEN Genome-Editing System for Generating Human Stem Cell-Based Disease Models. *Cell Stem Cell*, 12(2):238–251, July 2013.

[138] Mark J. Osborn, Colby G. Starker, Amber N. McElroy, Beau R. Webber, Megan J. Riddle, Lily Xia, Anthony P. DeFeo, Richard Gabriel, Manfred Schmidt, Christof Von Kalle, Daniel F. Carlson, Morgan L. Maeder, J. Keith Joung, John E. Wagner, Daniel F. Voytas, Bruce R. Blazar, and Jakub Tolar. TALEN-based Gene Correction for Epidermolysis Bullosa. *Molecular Therapy*, 21(6):1151–1159, June 2013.

[139] Laurent Tesson, Claire Usal, Séverine Ménoret, Elo Leung, Brett J. Niles, Séverine Remy, Yolanda Santiago, Anna I. Vincent, Xiangdong Meng, Lei Zhang, Philip D. Gregory, Ignacio Anegon, and Gregory J. Cost. Knockout rats generated by embryo microinjection of TALENs. *Nature Biotechnology*, 29(8):695–696, August 2011.

[140] John P. Guilinger, Vikram Pattanayak, Deepak Reyon, Shengdar Q. Tsai, Jeffry D. Sander, J. Keith Joung, and David R. Liu. Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity. *Nature Methods*, 11(4):429–435, April 2014.

[141] M.L. Bulyk. Computational prediction of transcription-factor binding site locations. *Genome Biol.*, 5:201, 2003.

[142] E. Davidson. Inside the cis-regulatory module: control logic, and how regulatory environment is transduced into spatial patterns of gene expression. *Genomic Regulatory Systems: Development and Evolution*, pages 25–62, 2001.

[143] B.D. Pfeiffer. Tools for neuroanatomy and neurogenetics in Drosophila. *Proc. Natl. Acad. Sci. USA*, 105:9715–9720, 2008.

[144] M.S. Halfon, S.M. Gallo, and C.M. Bergman. REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. *Nucleic Acids Res.*, 36:D594–D598, 2008.

[145] T. Sandmann. A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes Dev.*, 21:436–449, 2007.

[146] M. Jagalur, C. Pal, E. Learned-Miller, R.T. Zoeller, and D. Kulp. Analyzing in situ gene expression in the mouse brain with image registration, feature extraction and block clustering. *BMC Bioinformatics*, 8((suppl. 10)):S5, 2007.

[147] J. Gertz, E.D. Siggia, and B.A. Cohen. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, 457:215–218, 2009.

[148] J. Nam, P. Dong, R. Tarpine, S. Istrail, and E.H. Davidson. Functional cis-regulatory genomics for systems biology. *Proc. Natl. Acad. Sci. USA*, 107:3930–3935, 2010.

[149] Rupali P. Patwardhan, Joseph B. Hiatt, Daniela M. Witten, Mee J. Kim, Robin P. Smith, Dalit May, Choli Lee, Jennifer M. Andrie, Su-In Lee, Gregory M. Cooper, Nadav Ahituv, Len A. Pennacchio, and Jay Shendure. Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology*, 30(3):265–270, March 2012.

[150] E. Sharon. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.*, 30:521–530, 2012.

[151] O.M. Dunin-Borkowski and N.H. Brown. Mammalian CD2 is an effective heterologous marker of the cell surface in Drosophila. *Dev. Biol.*, 168:689–693, 1995.

[152] M. Bate. The mesoderm and its derivatives. *The development of Drosophila melanogaster*, pages 1013–1090, 1993.

[153] S. Contrino. modMine: flexible access to modENCODE data. *Nucleic Acids Res.*, 40:D1082–D1088, 2012.

[154] S. Roy. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, 330:1787–1797, 2010.

[155] S. Thomas. Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biol.*, 12:R43, 2011.

[156] H. Duan, J.B. Skeath, and H.T. Nguyen. Drosophila Lame duck, a novel member of the Gli superfamily, acts as a key regulator of myogenesis by controlling fusion-competent myoblast development. *Development*, 128:4489–4500, 2001.

[157] A.C. Groth, M. Fish, R. Nusse, and M.P. Calos. Construction of transgenic Drosophila by using the site-specific integrase from phage phiC31. *Genetics*, 166:1775–1782, 2004.

[158] S. Hoffmann. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.*, 5:e1000502, 2009.

[159] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol.*, 11:R106, 2010.

[160] S.M. Gallo. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res.*, 39:D118–D123, 2011.

[161] S. Bonn. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.*, 44:148–156, 2012.

[162] A. Pekowska. H3k4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J.*, 30:4198–4210, 2011.

[163] P.V. Kharchenko. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, 471:480–485, 2011.

[164] Arnaud R. Krebs, Krishanpal Karmodiya, Marianne Lindahl-Allen, Kevin Struhl, and Làszlò Tora. SAGA and ATAC histone acetyl transferase complexes regulate distinct sets of genes and ATAC defines a class of p300-independent enhancers. *Molecular Cell*, 44(3):410–423, November 2011.

[165] B.W. Busser. Integrative analysis of the zinc finger transcription factor Lame duck in the Drosophila myogenic gene regulatory network. *Proc. Natl. Acad. Sci. USA*, 109:20768–20773, 2012.

[166] B.W. Busser. Molecular mechanism underlying the regulatory specificity of a Drosophila homeodomain protein that specifies myoblast identity. *Development*, 139:1164–1174, 2012.

[167] A.A. Philippakis. Expression-guided in silico evaluation of candidate cis regulatory codes for Drosophila muscle founder cells. *PLoS Comput. Biol.*, 2:e53, 2006.

[168] L.J. Zhu. FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.*, 39:D111–D117, 2011.

[169] B. Thisse, M. el Messal, and F. Perrin-Schmitt. The twist gene: isolation of a Drosophila zygotic gene necessary for the establishment of dorsoventral pattern. *Nucleic Acids Res.*, 15:3439–3453, 1987.

[170] E.E. Furlong, E.C. Andersen, B. Null, K.P. White, and M.P. Scott. Patterns of gene expression during Drosophila mesoderm development. *Science*, 293:1629–1633, 2001.

[171] C. Grimaud, N. Negre, and G. Cavalli. From genetics to epigenetics: the tale of Polycomb group and trithorax group genes. *Chromosome Res.*, 14:363–375, 2006.

[172] C. Herrmann, B. Van de Sande, D. Potier, and S. Aerts. i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Res.*, 40:e114, 2012.

[173] Y. Yuan, L. Guo, L. Shen, and J.S. Liu. Predicting gene expression from sequence: a reexamination. *PLoS Comput. Biol.*, 3:e243, 2007.

[174] B.W. Busser. A machine learning approach for identifying novel cell type-specific transcriptional regulators of myogenesis. *PLoS Genet.*, 8:e1002531, 2012.

[175] J.A. Lister. Transgene excision in zebrafish using the phiC31 integrase. *Genesis*, 48:137–143, 2010.

[176] B. Thyagarajan, E.C. Olivares, R.P. Hollis, D.S. Ginsburg, and M.P. Calos. Site-specific genomic integration in mammalian cells mediated by phage phiC31 integrase. *Mol. Cell Biol.*, 21:3926–3934, 2001.

[177] R.P. Hollis. Phage integrases for the construction and manipulation of transgenic mammals. *Reprod. Biol. Endocrinol.*, 1:79, 2003.

[178] S. Barolo, L.A. Carver, and J.W. Posakony. GFP and beta-galactosidase transformation vectors for promoter/enhancer analysis in Drosophila. *Biotechniques*, 29:726–732, 2000.

[179] M. Markstein, C. Pitsouli, C. Villalta, S.E. Celniker, and N. Perrimon. Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. *Nat. Genet.*, 40:476–483, 2008.

[180] J. Bischof, R.K. Maeda, M. Hediger, F. Karch, and K. Basler. An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. *Proc. Natl. Acad. Sci. USA*, 104:3312–3317, 2007.

[181] B. Estrada. An integrated strategy for analyzing the unique developmental programs of different myoblast subtypes. *PLoS Genet.*, 2:e16, 2006.

[182] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.

[183] Pavel Tomancak, Benjamin P. Berman, Amy Beaton, Richard Weiszmann, Elaine Kwan, Volker Hartenstein, Susan E. Celniker, and Gerald M. Rubin. Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biology*, 8(7):R145, July 2007.

[184] Pavel Tomancak, Amy Beaton, Richard Weiszmann, Elaine Kwan, ShengQiang Shu, Suzanna E. Lewis, Stephen Richards, Michael Ashburner, Volker Hartenstein, Susan E. Celniker, and Gerald M. Rubin. Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biology*, 3(12):research0088, December 2002.

[185] A. Aboukhalil and M.L. Bulyk. LOESS correction for length variation in gene set-based genomic sequence analysis. *Bioinformatics*, 28:1446–1454, 2012.

[186] F. Pedregosa. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[187] Yinon Ben-Neriah and Michael Karin. Inflammation meets cancer, with NF-κB as the matchmaker. *Nature Immunology*, 12(8):715–723, August 2011.

[188] Kian-Huat Lim, Yibin Yang, and Louis M. Staudt. Pathogenetic importance and therapeutic implications of NF-κB in lymphoid malignancies. *Immunological Reviews*, 246(1):359–378, March 2012.

[189] Gioacchino Natoli. Control of NF-κB-dependent Transcriptional Responses by Chromatin Organization. *Cold Spring Harbor Perspectives in Biology*, 1(4):a000224, October 2009.

[190] Stephen T. Smale. Hierarchies of NF-κB target-gene regulation. *Nature Immunology*, 12(8):689–694, August 2011.

[191] Fengyi Wan and Michael J. Lenardo. The nuclear signaling of NF-kappaB: current knowledge, new insights, and future perspectives. *Cell Research*, 20(1):24–33, January 2010.

[192] Masmudur M. Rahman and Grant McFadden. Modulation of NF-κB signalling by microbial pathogens. *Nature Reviews. Microbiology*, 9(4):291–306, April 2011.

[193] Giuseppina Bonizzi and Michael Karin. The two NF-kappaB activation pathways and their role in innate and adaptive immunity. *Trends in Immunology*, 25(6):280–288, June 2004.

[194] Shao-Cong Sun. Non-canonical NF-κB signaling pathway. *Cell Research*, 21(1):71–85, January 2011.

[195] Mary Kaileh and Ranjan Sen. NF-κB function in B lymphocytes. *Immunological Reviews*, 246(1):254–271, March 2012.

[196] Brian Zarnegar, Jeannie Q. He, Gagik Oganesyan, Alexander Hoffmann, David Baltimore, and Genhong Cheng. Unique CD40-mediated biological program in B cell activation requires both type 1 and type 2 NF-kappaB activation pathways. *Proceedings of the National Academy of Sciences of the United States of America*, 101(21):8108–8113, May 2004.

[197] S. Heinz, C. E. Romanoski, C. Benner, K. A. Allison, M. U. Kaikkonen, L. D. Orozco, and C. K. Glass. Effect of natural genetic variation on enhancer selection and function. *Nature*, 503(7477):487–492, November 2013.

[198] Fulai Jin, Yan Li, Jesse R. Dixon, Siddarth Selvaraj, Zhen Ye, Ah Young Lee, Chia-An Yen, Anthony D. Schmitt, Celso A. Espinoza, and Bing Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290–294, November 2013.

[199] Ching-Aeng Lim, Fei Yao, Joyce Jing-Yi Wong, Joshy George, Han Xu, Kuo Ping Chiu, Wing-Kin Sung, Leonard Lipovich, Vinsensius B. Vega, Joanne Chen, Atif Shahab, Xiao Dong Zhao, Martin Hibberd, Chia-Lin Wei, Bing Lim, Huck-Hui Ng, Yijun Ruan, and Keh-Chuang Chin. Genome-wide mapping of RELA(p65) binding identifies E2f1 as a transcriptional activator recruited by NF-kappaB upon TLR4 activation. *Molecular Cell*, 27(4):622–635, August 2007.

[200] Rebecca Martone, Ghia Euskirchen, Paul Bertone, Stephen Hartman, Thomas E. Royce, Nicholas M. Luscombe, John L. Rinn, F. Kenneth Nelson, Perry Miller, Mark Gerstein, Sherman Weissman, and Michael Snyder. Distribution of NF-kappaB-binding sites across human chromosome 22. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21):12247–12252, October 2003.

[201] Manuel Garber, Nir Yosef, Alon Goren, Raktima Raychowdhury, Anne Thielke, Mitchell Guttman, James Robinson, Brian Minie, Nicolas Chevrier, Zohar Itzhaki, Ronnie Blecher-Gonen, Chamutal Bornstein, Daniela Amann-Zalcenstein, Assaf Weiner, Dennis Friedrich, James Meldrim, Oren Ram, Christine Cheng, Andreas Gnirke, Sheila Fisher, Nir Friedman, Bang Wong, Bradley E. Bernstein, Chad Nusbaum, Nir Hacohen, Aviv Regev, and Ido Amit.

A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Molecular Cell*, 47(5):810–822, September 2012.

[202] Joerg Schreiber, Richard G. Jenner, Heather L. Murray, Georg K. Gerber, David K. Gifford, and Richard A. Young. Coordinated binding of NF-kappaB family members in the response of human cells to lipopolysaccharide. *Proceedings of the National Academy of Sciences of the United States of America*, 103(15):5899–5904, April 2006.

[203] Sung Hee Baek, Kenneth A. Ohgi, David W. Rose, Edward H. Koo, Christopher K. Glass, and Michael G. Rosenfeld. Exchange of N-CoR corepressor and Tip60 coactivator complexes links gene expression by NF-kappaB and beta-amyloid precursor protein. *Cell*, 110(1):55–67, July 2002.

[204] T. T. Huang, N. Kudo, M. Yoshida, and S. Miyamoto. A nuclear export signal in the N-terminal regulatory domain of IkappaBalpha controls cytoplasmic localization of inactive NF-kappaB/IkappaBalpha complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 97(3):1014–1019, February 2000.

[205] Jiajian Liu and David I. Beller. Distinct pathways for NF-kappa B regulation are associated with aberrant macrophage IL-12 production in lupus- and diabetes-prone mouse strains. *Journal of Immunology (Baltimore, Md.: 1950)*, 170(9):4489–4496, May 2003.

[206] R. M. Nissen and K. R. Yamamoto. The glucocorticoid receptor inhibits NFkappaB by interfering with serine-2 phosphorylation of the RNA polymerase II carboxy-terminal domain. *Genes & Development*, 14(18):2314–2329, September 2000.

[207] M. S. Rodriguez, J. Thompson, R. T. Hay, and C. Dargemont. Nuclear retention of IkappaBalpha protects it from signal-induced degradation and inhibits nuclear factor kappaB transcriptional activation. *The Journal of Biological Chemistry*, 274(13):9108–9115, March 1999.

[208] Simona Saccani, Serafino Pantano, and Gioacchino Natoli. Modulation of NF-kappaB activity by exchange of dimers. *Molecular Cell*, 11(6):1563–1574, June 2003.

[209] W. Wang, W. F. Tam, C. C. Hughes, S. Rath, and R. Sen. c-Rel is a target of pentoxifylline-mediated inhibition of T lymphocyte activation. *Immunity*, 6(2):165–174, February 1997.

[210] Tetsuo Yamazaki and Tomohiro Kurosaki. Contribution of BCAP to maintenance of mature B cells through c-Rel. *Nature Immunology*, 4(8):780–786, August 2003.

[211] Stephen G. Landt, Georgi K. Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E. Bernstein, Peter Bickel, James B. Brown, Philip Cayting, Yiwen Chen, Gilberto DeSalvo, Charles Epstein, Katherine I. Fisher-Aylor, Ghia Euskirchen, Mark Gerstein, Jason Gertz, Alexander J. Hartemink, Michael M. Hoffman, Vishwanath R. Iyer, Youngsook L. Jung, Subhradip Karmakar, Manolis Kellis, Peter V. Kharchenko, Qunhua Li, Tao Liu, X. Shirley Liu, Lijia Ma, Aleksandar Milosavljevic, Richard M. Myers, Peter J. Park, Michael J. Pazin, Marc D. Perry, Debasish Raha, Timothy E. Reddy, Joel Rozowsky, Noam Shoresh, Arend Sidow, Matthew Slattery, John A. Stamatoyannopoulos, Michael Y. Tolstorukov, Kevin P. White, Simon Xi, Peggy J. Farnham, Jason D. Lieb, Barbara J. Wold, and Michael Snyder. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22(9):1813–1831, September 2012.

[212] Tao Ye, Arnaud R. Krebs, Mohamed-Amin Choukrallah, Celine Keime, Frederic Plewniak, Irwin Davidson, and Laszlo Tora. seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Research*, 39(6):e35, March 2011.

[213] Cory Y. McLean, Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5):495–501, May 2010.

[214] F. E. Chen and G. Ghosh. Regulation of DNA binding by Rel/NF-kappaB transcription factors: structural views. *Oncogene*, 18(49):6845–6852, November 1999.

[215] C. W. Müller, F. A. Rey, M. Sodeoka, G. L. Verdine, and S. C. Harrison. Structure of the NF-kappa B p50 homodimer bound to DNA. *Nature*, 373(6512):311–317, January 1995.

[216] Kyoko Ochiai, Mark Maienschein-Cline, Giorgia Simonetti, Jianjun Chen, Rebecca Rosenthal, Robert Brink, Anita S. Chong, Ulf Klein, Aaron R. Dinner, Harinder Singh, and Roger Sciammas. Transcriptional regulation of germinal center B and plasma cell fates by dynamical control of IRF4. *Immunity*, 38(5):918–929, May 2013.

[217] Serena Ghisletti, Iros Barozzi, Flore Mietton, Sara Polletti, Francesca De Santa, Elisa Venturini, Lorna Gregory, Lorne Lonie, Adeline Chew, Chia-Lin Wei, Jiannis Ragoussis, and Gioacchino Natoli. Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity*, 32(3):317–328, March 2010.

[218] Matthias Merkenschlager and Duncan T. Odom. CTCF and cohesin: linking gene regulatory elements with their targets. *Cell*, 152(6):1285–1297, March 2013.

[219] Marianna Halasi and Andrei L. Gartel. FOX(M1) news–it is cancer. *Molecular Cancer Therapeutics*, 12(3):245–254, March 2013.

[220] Celine Lefebvre, Presha Rajbhandari, Mariano J. Alvarez, Pradeep Bandaru, Wei Keat Lim, Mai Sato, Kai Wang, Pavel Sumazin, Manjunath Kustagi, Brygida C. Bisikirska, Katia Basso, Pedro Beltrao, Nevan Krogan, Jean Gautier, Riccardo Dalla-Favera, and Andrea Califano. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular Systems Biology*, 6:377, June 2010.

[221] Van S. Tompkins, Seong-Su Han, Alicia Olivier, Sergei Syrbu, Thomas Bair, Anna Button, Laura Jacobus, Zebin Wang, Samuel Lifton, Pradip Raychaudhuri, Herbert C. Morse, George Weiner, Brian Link, Brian J. Smith, and Siegfried Janz. Identification of candidate B-lymphoma genes by cross-species gene expression profiling. *PloS One*, 8(10):e76889, 2013.

[222] Shahab Uddin, Azhar R. Hussain, Maqbool Ahmed, Khawar Siddiqui, Fouad Al-Dayel, Prashant Bavi, and Khawla S. Al-Kuraya. Overexpression of FoxM1 offers a promising therapeutic target in diffuse large B-cell lymphoma. *Haematologica*, 97(7):1092–1100, July 2012.

[223] G. Lenz, G. Wright, S. S. Dave, W. Xiao, J. Powell, H. Zhao, W. Xu, B. Tan, N. Goldschmidt, J. Iqbal, J. Vose, M. Bast, K. Fu, D. D. Weisenburger, T. C. Greiner, J. O. Armitage, A. Kyle, L. May, R. D. Gascoyne, J. M. Connors, G. Troen, H. Holte, S. Kvaloy, D. Dierickx, G. Verhoef, J. Delabie, E. B. Smeland, P. Jares, A. Martinez, A. Lopez-Guillermo, E. Montserrat, E. Campo, R. M. Braziel, T. P. Miller, L. M. Rimsza, J. R. Cook, B. Pohlman, J. Sweetenham, R. R. Tubbs, R. I. Fisher, E. Hartmann, A. Rosenwald, G. Ott, H.-K. Muller-Hermelink, D. Wrench, T. A. Lister, E. S. Jaffe, W. H. Wilson, W. C. Chan, L. M. Staudt, and Lymphoma/Leukemia Molecular Profiling Project. Stromal gene signatures in large-B-cell lymphomas. *The New England Journal of Medicine*, 359(22):2313–2323, November 2008.

[224] Andrea Oeckinghaus, Matthew S. Hayden, and Sankar Ghosh. Crosstalk in NF-κB signaling pathways. *Nature Immunology*, 12(8):695–708, August 2011.

[225] Ranjan Sen and Stephen T. Smale. Selectivity of the NF-κB Response. *Cold Spring Harbor Perspectives in Biology*, 2(4):a000257, April 2010.

[226] Gioacchino Natoli, Simona Saccani, Daniela Bosisio, and Ivan Marazzi. Interactions of NF-kappaB with chromatin: the art of being at the right place at the right time. *Nature Immunology*, 6(5):439–445, May 2005.

[227] Subhashini Sadasivam, Shenghua Duan, and James A. DeCaprio. The MuvB complex sequentially recruits B-Myb and FoxM1 to promote mitotic gene expression. *Genes & Development*, 26(5):474–489, March 2012.

[228] E. D. Cahir-McFarland, D. M. Davidson, S. L. Schauer, J. Duong, and E. Kieff. NF-kappa B inhibition causes spontaneous apoptosis in Epstein-Barr virus-transformed lymphoblastoid cells. *Proceedings of the National Academy of Sciences of the United States of America*, 97(11):6055–6060, May 2000.

[229] I.-C. Wang, V. Ustiyan, Y. Zhang, Y. Cai, T. V. Kalin, and V. V. Kalinichenko. Foxm1 transcription factor is required for the initiation of lung tumorigenesis by oncogenic Kras(G12d.). *Oncogene*, 33(46):5391–5396, November 2014.

[230] Michael R. Green, Carlos Aya-Bonilla, Maher K. Gandhi, Rod A. Lea, Jeremy Wellwood, Peter Wood, Paula Marlton, and Lyn R. Griffiths. Integrative genomic profiling reveals conserved genetic mechanisms for tumorigenesis in common entities of non-Hodgkin's lymphoma. *Genes, Chromosomes & Cancer*, 50(5):313–326, May 2011.

[231] Mathew E. Sowa, Eric J. Bennett, Steven P. Gygi, and J. Wade Harper. Defining the human deubiquitinating enzyme interaction landscape. *Cell*, 138(2):389–403, July 2009.

[232] Xiaohui Zhou, Benjamin E. Gewurz, Jennifer M. Ritchie, Kaoru Takasaki, Hannah Greenfeld, Elliott Kieff, Brigid M. Davis, and Matthew K. Waldor. A Vibrio parahaemolyticus T3ss effector mediates pathogenesis by independently enabling intestinal colonization and inhibiting TAK1 activation. *Cell Reports*, 3(5):1690–1702, May 2013.

[233] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[234] Peter V. Kharchenko, Michael Y. Tolstorukov, and Peter J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology*, 26(12):1351–1359, December 2008.

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

[236] Michael M. Hoffman, Jason Ernst, Steven P. Wilder, Anshul Kundaje, Robert S. Harris, Max Libbrecht, Belinda Giardine, Paul M. Ellenbogen, Jeffrey A. Bilmes, Ewan Birney, Ross C. Hardison, Ian Dunham, Manolis Kellis, and William Stafford Noble. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research*, page gks1284, December 2012.

[237] Kevin Y. Yip, Chao Cheng, Nitin Bhardwaj, James B. Brown, Jing Leng, Anshul Kundaje, Joel Rozowsky, Ewan Birney, Peter Bickel, Michael Snyder, and Mark Gerstein. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biology*, 13(9):R48, 2012.

[238] I. V. Kulakovskiy, V. A. Boeva, A. V. Favorov, and V. J. Makeev. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics (Oxford, England)*, 26(20):2622–2623, October 2010.

[239] Timothy L. Bailey, Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37(Web Server issue):W202–208, July 2009.

[240] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–842, March 2010.

[241] Stephen S. Gisselbrecht, Luis A. Barrera, Martin Porsch, Anton Aboukhalil, Preston W. Estep, Anastasia Vedenko, Alexandre Palagi, Yongsok Kim, Xianmin Zhu, Brian W. Busser, Caitlin E. Gamble, Antonina Iagovitina, Aditi Singhania, Alan M. Michelson, and Martha L. Bulyk. Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. *Nature Methods*, 10(8):774–780, August 2013.

[242] Eugene V. Davydov, David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput Biol*, 6(12):e1001025, December 2010.

[243] Shitao Li, Lingyan Wang, Michael Berman, Young-Yun Kong, and Martin E. Dorf. Mapping a dynamic innate immunity protein interaction network regulating type I interferon production. *Immunity*, 35(3):426–440, September 2011.

[244] Gregory A. Wray. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, 8(3):206–216, March 2007.

[245] Cheryl C Hsia and William McGinnis. Evolution of transcription factor function. *Current Opinion in Genetics & Development*, 13(2):199–206, April 2003.

[246] Alexis Battle, Zia Khan, Sidney H. Wang, Amy Mitrano, Michael J. Ford, Jonathan K. Pritchard, and Yoav Gilad. Impact of regulatory variation from RNA to protein. *Science*, 347(6222):664–667, February 2015.

[247] Steven C. Elbein, Eric R. Gamazon, Swapan K. Das, Neda Rasouli, Philip A. Kern, and Nancy J. Cox. Genetic Risk Factors for Type 2 Diabetes: A Trans-Regulatory Genetic Architecture? *The American Journal of Human Genetics*, 91(3):466–477, September 2012.

[248] David S. Latchman. Transcription-Factor Mutations and Disease. *New England Journal of Medicine*, 334(1):28–33, January 1996.

[249] Jean Villard. Transcription regulation and human diseases. *Swiss Medical Weekly*, 134(39-40):571–579, October 2004.

[250] Manuel A. Rivas, Matti Pirinen, Donald F. Conrad, Monkol Lek, Emily K. Tsang, Konrad J. Karczewski, Julian B. Maller, Kimberly R. Kukurba, David S. DeLuca, Menachem Fromer, Pedro G. Ferreira, Kevin S. Smith, Rui Zhang, Fengmei Zhao, Eric Banks, Ryan Poplin, Douglas M. Ruderfer, Shaun M. Purcell, Taru Tukiainen, Eric V. Minikel, Peter D. Stenson, David N. Cooper, Katharine H. Huang, Timothy J. Sullivan, Jared Nedzel, GTEx Consortium, Geuvadis Consortium, Carlos D. Bustamante, Jin Billy Li, Mark J. Daly, Roderic Guigo, Peter Donnelly, Kristin Ardlie, Michael Sammeth, Emmanouil T. Dermitzakis, Mark I. McCarthy, Stephen B. Montgomery, Tuuli Lappalainen, and Daniel G. MacArthur. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science (New York, N.Y.)*, 348(6235):666–669, May 2015.

[251] Young-In Chi. Homeodomain revisited: a lesson from disease-causing mutations. *Human Genetics*, 116(6):433–444, February 2005.

[252] Anton V. Persikov, Joshua L. Wetzel, Elizabeth F. Rowland, Benjamin L. Oakes, Denise J. Xu, Mona Singh, and Marcus B. Noyes. A systematic survey of the Cys2his2 zinc finger DNA-binding landscape. *Nucleic Acids Research*, 43(3):1965–1984, February 2015.

[253] Gwenael Badis, Michael F. Berger, Anthony A. Philippakis, Shaheynoor Talukder, Andrew R. Gehrke, Savina A. Jaeger, Esther T. Chan, Genita Metzler, Anastasia Vedenko, Xiaoyu Chen, Hanna Kuznetsov, Chi-Fong Wang, David Coburn, Daniel E. Newburger, Quaid Morris,

Timothy R. Hughes, and Martha L. Bulyk. Diversity and Complexity in DNA Recognition by Transcription Factors. *Science*, 324(5935):1720–1723, June 2009.

[254] Bo Jiang, Jun S. Liu, and Martha L. Bulyk. Bayesian hierarchical model of protein-binding microarray k-mer data reduces noise and identifies transcription factor subclasses and preferred k-mers. *Bioinformatics*, 29(11):1390–1398, June 2013.

[255] Franck Mauvais-Jarvis, Stuart B. Smith, Cédric Le May, Suzanne M. Leal, Jean-François Gautier, Mariam Molokhia, Jean-Pierre Riveline, Arun S. Rajan, Jean-Philippe Kevorkian, Sumei Zhang, Patrick Vexiau, Michael S. German, and Christian Vaisse. PAX4 gene variations predispose to ketosis-prone diabetes. *Human Molecular Genetics*, 13(24):3151–3159, December 2004.

[256] Nattachet Plengvidhya, Suwattanee Kooptiwut, Napat Songtawee, Asako Doi, Hiroto Furuta, Masahiro Nishi, Kishio Nanjo, Wiwit Tantibhedhyangkul, Watip Boonyasrisawat, Pa-thai Yenchitsomanus, Alessandro Doria, and Napatawn Banchuin. *PAX4* Mutations in Thais with Maturity Onset Diabetes of the Young. *The Journal of Clinical Endocrinology & Metabolism*, 92(7):2821–2826, July 2007.

[257] Michael F. Berger, Anthony A. Philippakis, Aaron M. Qureshi, Fangxue S. He, Preston W. Estep, and Martha L. Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology*, 24(11):1429–1435, November 2006.

[258] David Johnson, Shih-hsin Kan, Michael Oldridge, Richard C. Trembath, Philippe Roche, Robert M. Esnouf, Henk Giele, and O. M. Andrew Wilkie. Missense Mutations in the Homeodomain of HOXD13 Are Associated with Brachydactyly Types D and E. *The American Journal of Human Genetics*, 72(4):984–997, April 2003.

[259] Yasuteru Muragaki, Stefan Mundlos, Joseph Upton, and Bjorn R. Olsen. Altered Growth and Branching Patterns in Synpolydactyly Caused by Mutations in HOXD 13. *Science*, 272(5261):548–551, April 1996.

[260] Nathaniel H. Robin, Jennifer Hurvitz, Matthew L. Warman, and Stuart Morrison. Clinical and molecular studies of brachydactyly type D. *American Journal of Medical Genetics*, 85(4):413–418, August 1999.

[261] Xiuli Zhao, Miao Sun, Jin Zhao, J. Alfonso Leyva, Hongwen Zhu, Wei Yang, Xuan Zeng, Yang Ao, Qing Liu, Guoyang Liu, Wilson H. Y. Lo, Ethylin Wang Jabs, L. Mario Amzel, Xiang-

nian Shan, and Xue Zhang. Mutations in HOXD13 Underlie Syndactyly Type V and a Novel Brachydactyly-Syndactyly Syndrome. *The American Journal of Human Genetics*, 80(2):361–371, February 2007.

[262] Takahisa Furukawa, Eric M. Morrow, Tiansen Li, Fred C. Davis, and Constance L. Cepko. Retinopathy and attenuated circadian entrainment in Crx-deficient mice. *Nature Genetics*, 23(4):466–470, December 1999.

[263] Carol L. Freund, Cheryl Y. Gregory-Evans, Takahisa Furukawa, Myrto Papaioannou, Jens Looser, Lynda Ploder, James Bellingham, David Ng, Jo-Anne S. Herbrick, Alessandra Duncan, Stephen W. Scherer, Lap-Chee Tsui, Aphrodite Loutradis-Anagnostou, Samuel G. Jacobson, Constance L. Cepko, Shomi S. Bhattacharya, and Roderick R. McInnes. Cone-Rod Dystrophy Due to Mutations in a Novel Photoreceptor-Specific Homeobox Gene (CRX) Essential for Maintenance of the Photoreceptor. *Cell*, 91(4):543–553, November 1997.

[264] Kenneth P. Mitton, Prabodh K. Swain, Shiming Chen, Siqun Xu, Donald J. Zack, and Anand Swaroop. The Leucine Zipper of NRL Interacts with the CRX Homeodomain A POSSIBLE MECHANISM OF TRANSCRIPTIONAL SYNERGY IN RHODOPSIN REGULATION. *Journal of Biological Chemistry*, 275(38):29794–29799, September 2000.

[265] Prabodha K. Swain, Shiming Chen, Qing-Liang Wang, Louisa M. Affatigato, Caraline L. Coats, Kevin D. Brady, Gerald A. Fishman, Samuel G. Jacobson, Anand Swaroop, Edwin Stone, Paul A. Sieving, and Donald J. Zack. Mutations in the Cone-Rod Homeobox Gene Are Associated with the Cone-Rod Dystrophy Photoreceptor Degeneration. *Neuron*, 19(6):1329–1336, January 1997.

[266] Joseph C. Corbo, Karen A. Lawrence, Marcus Karlstetter, Connie A. Myers, Musa Abdelaziz, William Dirkes, Karin Weigelt, Martin Seifert, Vladimir Benes, Lars G. Fritsche, Bernhard H. F. Weber, and Thomas Langmann. CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Research*, 20(11):1512–1525, November 2010.

[267] M. M. Humphries, D. Rancourt, G. J. Farrar, P. Kenna, M. Hazel, R. A. Bush, P. A. Sieving, D. M. Sheils, N. McNally, P. Creighton, A. Erven, A. Boros, K. Gulya, M. R. Capecchi, and P. Humphries. Retinopathy induced in mice by targeted disruption of the rhodopsin gene. *Nature Genetics*, 15(2):216–219, February 1997.

[268] C. Fromental-Ramain, X. Warot, N. Messadecq, M. LeMeur, P. Dollé, and P. Chambon. Hoxa-13 and Hoxd-13 play a crucial role in the patterning of the limb autopod. *Development (Cambridge, England)*, 122(10):2997–3011, October 1996.

[269] Frances R. Goodman. Limb malformations and the human HOX genes. *American Journal of Medical Genetics*, 112(3):256–265, October 2002.

[270] Lionel Arnaud, Carole Saison, Virginie Helias, Nicole Lucien, Dominique Steschenko, Marie-Catherine Giarratana, Claude Prehu, Bernard Foliguet, Lory Montout, Alexandre G. de Brevern, Alain Francina, Pierre Ripoche, Odile Fenneteau, Lydie Da Costa, Thierry Peyrard, Gail Coghlan, Niels Illum, Henrik Birgens, Hannah Tamary, Achille Iolascon, Jean Delaunay, Gil Tchernia, and Jean-Pierre Cartron. A Dominant Mutation in the Gene Encoding the Erythroid Transcription Factor KLF1 Causes a Congenital Dyserythropoietic Anemia. *The American Journal of Human Genetics*, 87(5):721–727, December 2010.

[271] Belinda K. Singleton, Nicholas M. Burton, Carole Green, R. Leo Brady, and David J. Anstee. Mutations in EKLF/KLF1 form the molecular basis of the rare blood group In(Lu) phenotype. *Blood*, 112(5):2081–2088, September 2008.

[272] Mitsuhiro Kato, Soma Das, Kristin Petras, Kunio Kitamura, Ken-ichirou Morohashi, Diane N. Abuelo, Mason Barr, Dominique Bonneau, Angela F. Brady, Nancy J. Carpenter, Karen L. Cipero, Francesco Frisone, Takayuki Fukuda, Renzo Guerrini, Eri Iida, Masayuki Itoh, Amy Feldman Lewanda, Yukiko Nanba, Akira Oka, Virginia K. Proud, Pascale Saugier-Veber, Susan L. Schelley, Angelo Selicorni, Rachel Shaner, Margherita Silengo, Fiona Stewart, Noriyuki Sugiyama, Jun Toyama, Annick Toutain, Ana Lía Vargas, Masako Yanazawa, Elaine H. Zackai, and William B. Dobyns. Mutations of ARX are associated with striking pleiotropy and consistent genotype–phenotype correlation. *Human Mutation*, 23(2):147–159, February 2004.

[273] Kunio Kitamura, Masako Yanazawa, Noriyuki Sugiyama, Hirohito Miura, Akiko Iizuka-Kogo, Masatomo Kusaka, Kayo Omichi, Rika Suzuki, Yuko Kato-Fukui, Kyoko Kamiirisa, Mina Matsuo, Shin-ichi Kamijo, Megumi Kasahara, Hidefumi Yoshioka, Tsutomu Ogata, Takayuki Fukuda, Ikuko Kondo, Mitsuhiro Kato, William B. Dobyns, Minesuke Yokoyama, and Ken-ichirou Morohashi. Mutation of ARX causes abnormal development of forebrain and testes in mice and X-linked lissencephaly with abnormal genitalia in humans. *Nature Genetics*, 32(3):359–369, November 2002.

[274] Patrick Sulem, Hannes Helgason, Asmundur Oddson, Hreinn Stefansson, Sigurjon A. Gudjonsson, Florian Zink, Eirikur Hjartarson, Gunnar Th Sigurdsson, Adalbjorg Jonasdottir, Aslaug Jonasdottir, Asgeir Sigurdsson, Olafur Th Magnusson, Augustine Kong, Agnar Helgason, Hilma Holm, Unnur Thorsteinsdottir, Gisli Masson, Daniel F. Gudbjartsson, and Kari Stefansson. Identification of a large set of rare complete human knockouts. *Nature Genetics*, advance online publication, March 2015.

[275] Tuuli Lappalainen, Stephen B. Montgomery, Alexandra C. Nica, and Emmanouil T. Dermitzakis. Epistatic selection between coding and regulatory variation in human evolution and disease. *American Journal of Human Genetics*, 89(3):459–463, September 2011.

[276] Juan I. Fuxman Bass, Nidhi Sahni, Shaleen Shrestha, Aurian Garcia-Gonzalez, Akihiro Mori, Numana Bhat, Song Yi, David E. Hill, Marc Vidal, and Albertha J. M. Walhout. Human gene-centered transcription factor networks for enhancers and disease variants. *Cell*, 161(3):661–673, April 2015.

[277] Nidhi Sahni, Song Yi, Mikko Taipale, Juan I. Fuxman Bass, Jasmin Coulombe-Huntington, Fan Yang, Jian Peng, Jochen Weile, Georgios I. Karras, Yang Wang, István A. Kovács, Atanas Kamburov, Irina Krykbaeva, Mandy H. Lam, George Tucker, Vikram Khurana, Amitabh Sharma, Yang-Yu Liu, Nozomu Yachie, Quan Zhong, Yun Shen, Alexandre Palagi, Adriana San-Miguel, Changyu Fan, Dawit Balcha, Amelie Dricot, Daniel M. Jordan, Jennifer M. Walsh, Akash A. Shah, Xinping Yang, Ani K. Stoyanova, Alex Leighton, Michael A. Calderwood, Yves Jacob, Michael E. Cusick, Kourosh Salehi-Ashtiani, Luke J. Whitesell, Shamil Sunyaev, Bonnie Berger, Albert-László Barabási, Benoit Charloteaux, David E. Hill, Tong Hao, Frederick P. Roth, Yu Xia, Albertha J. M. Walhout, Susan Lindquist, and Marc Vidal. Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. *Cell*, 161(3):647–660, April 2015.

[278] David E. G. McNay, James P. Turton, Daniel Kelberman, Kathryn S. Woods, Raja Brauner, Anastasios Papadimitriou, Eberhard Keller, Alexandra Keller, Nele Haufs, Heiko Krude, Stephen M. Shalet, and Mehul T. Dattani. HESX1 Mutations Are an Uncommon Cause of Septooptic Dysplasia and Hypopituitarism. *The Journal of Clinical Endocrinology & Metabolism*, 92(2):691–697, February 2007.

[279] Fiona Cunningham, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent

Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juettemann, Andreas K. Kähäri, Stephen Keenan, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Bert Overduin, Anne Parker, Mateus Patricio, Emily Perry, Miguel Pignatelli, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Bronwen L. Aken, Ewan Birney, Jennifer Harrow, Rhoda Kinsella, Matthieu Muffato, Magali Ruffier, Stephen M. J. Searle, Giulietta Spudich, Stephen J. Trevanion, Andy Yates, Daniel R. Zerbino, and Paul Flicek. Ensembl 2015. *Nucleic Acids Research*, 43(D1):D662–D669, January 2015.

[280] Xin Guo, Martha L. Bulyk, and Alexander J. Hartemink. Intrinsic disorder within and flanking the DNA-binding domains of human transcription factors. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 104–115, 2012.

[281] S. R. Eddy. Profile hidden Markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763, 1998.

[282] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation*, 32(8):894–899, August 2011.

[283] Jacob A. Tennessen, Abigail W. Bigham, Timothy D. O'Connor, Wenqing Fu, Eimear E. Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, Goo Jun, Hyun Min Kang, Daniel Jordan, Suzanne M. Leal, Stacey Gabriel, Mark J. Rieder, Goncalo Abecasis, David Altshuler, Deborah A. Nickerson, Eric Boerwinkle, Shamil Sunyaev, Carlos D. Bustamante, Michael J. Bamshad, Joshua M. Akey, Broad GO, Seattle GO, and NHLBI Exome Sequencing Project. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, N.Y.)*, 337(6090):64–69, July 2012.

[284] The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 36(suppl 1):D190–D195, January 2008.

[285] Evzen Boura, Lenka Rezabkova, Jiri Brynda, Veronika Obsilova, and Tomas Obsil. Structure of the human FOXO4-DBD-DNA complex at 1.9 Å resolution reveals new details of FOXO binding to the DNA. *Acta Crystallographica. Section D, Biological Crystallography*, 66(Pt 12):1351–1357, December 2010.

[286] Michael M. Brent, Ruchi Anand, and Ronen Marmorstein. Structural basis for DNA recognition by FoxO1 and its regulation by posttranslational modification. *Structure (London, England: 1993)*, 16(9):1407–1416, September 2008.

[287] K. L. Clark, E. D. Halay, E. Lai, and S. K. Burley. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature*, 364(6436):412–420, July 1993.

[288] D. R. Littler, M. Alvarez-Fernández, A. Stein, R. G. Hibbert, T. Heidebrecht, P. Aloy, R. H. Medema, and A. Perrakis. Structure of the FoxM1 DNA-recognition domain bound to a promoter sequence. *Nucleic Acids Research*, 38(13):4527–4538, July 2010.

[289] James C. Stroud, Yongqing Wu, Darren L. Bates, Aidong Han, Katja Nowick, Svante Paabo, Harry Tong, and Lin Chen. Structure of the forkhead domain of FOXP2 bound to DNA. *Structure (London, England: 1993)*, 14(1):159–166, January 2006.

[290] Kuang-Lei Tsai, Cheng-Yang Huang, Chia-Hao Chang, Yuh-Ju Sun, Woei-Jer Chuang, and Chwan-Deng Hsiao. Crystal structure of the human FOXK1a-DNA complex and its implications on the diverse binding specificity of winged helix/forkhead proteins. *The Journal of Biological Chemistry*, 281(25):17400–17409, June 2006.

[291] Kuang-Lei Tsai, Yuh-Ju Sun, Cheng-Yang Huang, Jer-Yen Yang, Mien-Chie Hung, and Chwan-Deng Hsiao. Crystal structure of the human FOXO3a-DBD/DNA complex suggests the effects of post-translational modification. *Nucleic Acids Research*, 35(20):6984–6994, 2007.

[292] Ngak-Leng Sim, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C. Ng. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, 40(Web Server issue):W452–457, July 2012.

[293] Ivan A. Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, April 2010.

[294] Sung Chun and Justin C. Fay. Identification of deleterious mutations within three human genomes. *Genome Research*, 19(9):1553–1561, September 2009.

[295] Jana Marie Schwarz, Christian Rödelsperger, Markus Schuelke, and Dominik Seelow. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, 7(8):575–576, August 2010.

[296] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research*, 39(17):e118, September 2011.

[297] Ayodeji Olatubosun, Jouni Väliaho, Jani Härkönen, Janita Thusberg, and Mauno Vihinen. PON-P: integrated predictor for pathogenicity of missense variants. *Human Mutation*, 33(8):1166–1174, August 2012.

[298] Lucas D. Ward and Manolis Kellis. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*, 40(D1):D930–D934, January 2012.

[299] Aimée M. Dudley, John Aach, Martin A. Steffen, and George M. Church. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proceedings of the National Academy of Sciences*, 99(11):7554–7559, May 2002.

[300] Gavin E. Crooks, Gary Hon, John-Marc Chandonia, and Steven E. Brenner. WebLogo: a sequence logo generator. *Genome Research*, 14(6):1188–1190, June 2004.

[301] Matthew T. Weirauch, Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate Cook, Hong Zheng, Alejandra Goity, Harm van Bakel, Jean-Claude Lozano, Mary Galli, Mathew G. Lewsey, Eryong Huang, Tuhin Mukherjee, Xiaoting Chen, John S. Reece-Hoyes, Sridhar Govindarajan, Gad Shaulsky, Albertha J. M. Walhout, François-Yves Bouget, Gunnar Ratsch, Luis F. Larrondo, Joseph R. Ecker, and Timothy R. Hughes. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–1443, September 2014.

[302] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, January 1995.

[303] Marion Ouedraogo, Charles Bettembourg, Anthony Bretaudeau, Olivier Sallou, Christian Diot, Olivier Demeure, and Frédéric Lecerf. The duplicated genes database: identification and functional annotation of co-localised duplicated genes across genomes. *PloS One*, 7(11):e50653, 2012.

[304] Takeshi Obayashi, Yasunobu Okamura, Satoshi Ito, Shu Tadaka, Ikuko N. Motoike, and Kengo Kinoshita. COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. *Nucleic Acids Research*, 41(Database issue):D1014–1020, January 2013.

[305] Stephanie D. Byrum, Sean D. Taverna, and Alan J. Tackett. Purification of a specific native genomic locus for proteomic analysis. *Nucleic Acids Research*, 41(20):e195–e195, November 2013.

[306] Le Cong, Ruhong Zhou, Yu-chi Kuo, Margaret Cunniff, and Feng Zhang. Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nature Communications*, 3:968, July 2012.

[307] J. Keith Joung and Jeffry D. Sander. TALENs: a widely applicable technology for targeted genome editing. *Nature Reviews Molecular Cell Biology*, 14(1):49–55, January 2013.

[308] Yusuke Miyanari, Céline Ziegler-Birling, and Maria-Elena Torres-Padilla. Live visualization of chromatin dynamics with fluorescent TALEs. *Nature Structural & Molecular Biology*, 20(11):1321–1324, November 2013.

[309] Pablo Perez-Pinera, David G. Ousterout, Jonathan M. Brunger, Alicia M. Farin, Katherine A. Glass, Farshid Guilak, Gregory E. Crawford, Alexander J. Hartemink, and Charles A. Gersbach. Synergistic and tunable human gene activation by combinations of synthetic transcription factors. *Nature Methods*, 10(3):239–242, March 2013.

[310] Dong Deng, Chuangye Yan, Xiaojing Pan, Magdy Mahfouz, Jiawei Wang, Jian-Kang Zhu, Yigong Shi, and Nieng Yan. Structural Basis for Sequence-Specific Recognition of DNA by TAL Effectors. *Science*, 335(6069):720–723, February 2012.

[311] Amanda Nga-Sze Mak, Philip Bradley, Raul A. Cernadas, Adam J. Bogdanove, and Barry L. Stoddard. The Crystal Structure of TAL Effector PthXo1 Bound to Its DNA Target. *Science*, 335(6069):716–719, February 2012.

[312] Jennifer A. Doudna and Emmanuelle Charpentier. The new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213):1258096, November 2014.

[313] Robert Morbitzer, Patrick Römer, Jens Boch, and Thomas Lahaye. Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type tran-

scription factors. *Proceedings of the National Academy of Sciences*, 107(50):21617–21622, December 2010.

[314] Claudio Mussolino, Robert Morbitzer, Fabienne Lütge, Nadine Dannemann, Thomas Lahaye, and Toni Cathomen. A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Research*, 39(21):9283–9293, November 2011.

[315] Prashant Mali, John Aach, P. Benjamin Stranges, Kevin M. Esvelt, Mark Moosburner, Sriram Kosuri, Luhan Yang, and George M. Church. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature Biotechnology*, 31(9):833–838, September 2013.

[316] Erin L. Doyle, Nicholas J. Booher, Daniel S. Standage, Daniel F. Voytas, Volker P. Brendel, John K. VanDyk, and Adam J. Bogdanove. TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. *Nucleic Acids Research*, 40(W1):W117–W122, July 2012.

[317] Jan Grau, Annett Wolf, Maik Reschke, Ulla Bonas, Stefan Posch, and Jens Boch. Computational Predictions Provide Insights into the Biology of TAL Effector Target Sites. *PLoS Comput Biol*, 9(3):e1002962, March 2013.

[318] Joshua F. Meckler, Mital S. Bhakta, Moon-Soo Kim, Robert Ovadia, Chris H. Habrian, Artem Zykovich, Abigail Yu, Sarah H. Lockwood, Robert Morbitzer, Janett Elsäesser, Thomas Lahaye, David J. Segal, and Enoch P. Baldwin. Quantitative analysis of TALE–DNA interactions suggests polarity effects. *Nucleic Acids Research*, 41(7):4118–4128, April 2013.

[319] Jana Streubel, Christina Blücher, Angelika Landgraf, and Jens Boch. TAL effector RVD specificities and efficiencies. *Nature Biotechnology*, 30(7):593–595, July 2012.

[320] Eli J. Fine, Thomas J. Cradick, Charles L. Zhao, Yanni Lin, and Gang Bao. An online bioinformatics tool predicts zinc finger and TALE nuclease off-target cleavage. *Nucleic Acids Research*, 42(6):e42–e42, April 2014.

[321] Alvaro L. Pérez-Quintero, Luis M. Rodriguez-R, Alexis Dereeper, Camilo López, Ralf Koebnik, Boris Szurek, and Sebastien Cunnac. An Improved Method for TAL Effectors DNA-Binding Sites Prediction Reveals Functional Convergence in TAL Repertoires of Xanthomonas oryzae Strains. *PLoS ONE*, 8(7):e68464, July 2013.

[322] Sonali Mukherjee, Michael F. Berger, Ghil Jona, Xun S. Wang, Dale Muzzey, Michael Snyder, Richard A. Young, and Martha L. Bulyk. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genetics*, 36(12):1331–1339, December 2004.

[323] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April 2005.

[324] Alexandre Juillerat, Gwendoline Dubois, Julien Valton, Séverine Thomas, Stefano Stella, Alan Maréchal, Stéphanie Langevin, Nassima Benomari, Claudia Bertonati, George H. Silva, Fayza Daboussi, Jean-Charles Epinat, Guillermo Montoya, Aymeric Duclert, and Philippe Duchateau. Comprehensive analysis of the specificity of transcription activator-like effector nucleases. *Nucleic Acids Research*, 42(8):5390–5402, April 2014.

[325] Yue Zhao and Gary D. Stormo. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature Biotechnology*, 29(6):480–483, June 2011.

[326] Abhishek Garg, Jason J. Lohmueller, Pamela A. Silver, and Thomas Z. Armel. Engineering synthetic TAL effectors with orthogonal target sites. *Nucleic Acids Research*, 40(15):7584–7595, August 2012.

[327] Tom Schreiber and Ulla Bonas. Repeat 1 of TAL effectors affects target specificity for the base at position zero. *Nucleic Acids Research*, 42(11):7160–7169, June 2014.

[328] Daniel Blankenberg, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, Chapter 19:Unit 19.10.1–21, January 2010.

[329] Belinda Giardine, Cathy Riemer, Ross C. Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, Daniel Blankenberg, Istvan Albert, James Taylor, Webb Miller, W. James Kent, and Anton Nekrutenko. Galaxy: A platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451–1455, October 2005.

[330] Jeremy Goecks, Anton Nekrutenko, James Taylor, and $author firstName $author.lastName. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, August 2010.

[331] Yanni Lin, Eli J. Fine, Zhilan Zheng, Christopher J. Antico, Richard A. Voit, Matthew H. Porteus, Thomas J. Cradick, and Gang Bao. SAPTA: a new design tool for improving TALE nuclease activity. *Nucleic Acids Research*, 42(6):e47–e47, April 2014.

[332] T. Siggers, M. H. Duyzend, J. Reddy, S. Khan, and M. L. Bulyk. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Molecular Systems Biology*, 7(1):555–555, April 2014.

[333] Muhammad A. Zabidi, Cosmas D. Arnold, Katharina Schernhuber, Michaela Pagani, Martina Rath, Olga Frank, and Alexander Stark. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, 518(7540):556–559, February 2015.

[334] Evgeny Z. Kvon, Tomas Kazmar, Gerald Stampfel, J. Omar Yáñez-Cuna, Michaela Pagani, Katharina Schernhuber, Barry J. Dickson, and Alexander Stark. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature*, advance online publication, June 2014.

[335] Ilaria Mogno, Jamie C. Kwasnieski, and Barak A. Cohen. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Research*, 23(11):1908–1915, November 2013.

[336] Nergiz Dogan, Weisheng Wu, Christapher S. Morrissey, Kuan-Bei Chen, Aaron Stonestrom, Maria Long, Cheryl A. Keller, Yong Cheng, Deepti Jain, Axel Visel, Len A. Pennacchio, Mitchell J. Weiss, Gerd A. Blobel, and Ross C. Hardison. Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics & Chromatin*, 8:16, 2015.

[337] Joshua S. Bloom, Ian M. Ehrenreich, Wesley T. Loo, Thúy-Lan Võ Lite, and Leonid Kruglyak. Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436):234–237, February 2013.

[338] Frank W. Albert, Sebastian Treusch, Arthur H. Shockley, Joshua S. Bloom, and Leonid Kruglyak. Genetics of single-cell protein abundance variation in large yeast populations. *Nature*, 506(7489):494–497, February 2014.

[339] Andreas Massouras, Sebastian M. Waszak, Monica Albarca-Aguilera, Korneel Hens, Wiebke Holcombe, Julien F. Ayroles, Emmanouil T. Dermitzakis, Eric A. Stone, Jeffrey D. Jensen,

Trudy F. C. Mackay, and Bart Deplancke. Genomic Variation and Its Impact on Gene Expression in Drosophila melanogaster. *PLoS Genet*, 8(11):e1003055, November 2012.

[340] P. M. Bingham, M. G. Kidwell, and G. M. Rubin. The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. *Cell*, 29(3):995–1004, July 1982.

[341] A. H. Brand and N. Perrimon. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development (Cambridge, England)*, 118(2):401–415, June 1993.

[342] Adam S. Cockrell and Tal Kafri. Gene delivery by lentivirus vectors. *Molecular Biotechnology*, 36(3):184–204, July 2007.

[343] Patrick P. L. Tam and Janet Rossant. Mouse embryonic chimeras: tools for studying mammalian. *Development*, 130(25):6155–6163, December 2003.

[344] Dalit May, Matthew J. Blow, Tommy Kaplan, David J. McCulley, Brian C. Jensen, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Veena Afzal, Paul C. Simpson, Edward M. Rubin, Brian L. Black, James Bristow, Len A. Pennacchio, and Axel Visel. Large-scale discovery of enhancers from human heart tissue. *Nature Genetics*, 44(1):89–93, January 2012.

[345] Maike Buchner, Eugene Park, Huimin Geng, Lars Klemm, Johanna Flach, Emmanuelle Passegué, Hilde Schjerven, Ari Melnick, Elisabeth Paietta, Dragana Kopanja, Pradip Raychaudhuri, and Markus Müschen. Identification of FOXM1 as a therapeutic target in B-cell lineage acute lymphoblastic leukaemia. *Nature Communications*, 6, March 2015.

[346] Sebastian M. Waszak, Olivier Delaneau, Andreas R. Gschwind, Helena Kilpinen, Sunil K. Raghav, Robert M. Witwicki, Andrea Orioli, Michael Wiederkehr, Nikolaos I. Panousis, Alisa Yurovsky, Luciana Romano-Palumbo, Alexandra Planchon, Deborah Bielser, Ismael Padioleau, Gilles Udin, Sarah Thurnheer, David Hacker, Nouria Hernandez, Alexandre Reymond, Bart Deplancke, and Emmanouil T. Dermitzakis. Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell*, 162(5):1039–1050, August 2015.

[347] Cristina T. Vicente, Stacey L. Edwards, Kristine M. Hillman, Susanne Kaufmann, Hayley Mitchell, Lisa Bain, Dylan M. Glubb, Jason S. Lee, Juliet D. French, and Manuel A. R. Fer-

reira. Long-Range Modulation of PAG1 Expression by 8q21 Allergy Risk Variants. *American Journal of Human Genetics*, 97(2):329–336, August 2015.

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

[349] Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, advance online publication, August 2015.

[350] Ophir Shalem, Neville E. Sanjana, Ella Hartenian, Xi Shi, David A. Scott, Tarjei Mikkelson, Dirk Heckl, Benjamin L. Ebert, David E. Root, John G. Doench, and Feng Zhang. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science (New York, N.Y.)*, 343(6166):84–87, January 2014.

[351] Holger Flechsig. TALEs from a spring–superelasticity of Tal effector protein structures. *PloS One*, 9(10):e109919, 2014.

[352] Jeffrey C. Miller, Lei Zhang, Danny F. Xia, John J. Campo, Irina V. Ankoudinova, Dmitry Y. Guschin, Joshua E. Babiarz, Xiangdong Meng, Sarah J. Hinkley, Stephen C. Lam, David E. Paschon, Anna I. Vincent, Gladys P. Dulay, Kyle A. Barlow, David A. Shivak, Elo Leung, Jinwon D. Kim, Rainier Amora, Fyodor D. Urnov, Philip D. Gregory, and Edward J. Rebar. Improved specificity of TALE-based genome editing using an expanded RVD repertoire. *Nature Methods*, 12(5):465–471, May 2015.

[353] Prashant Mali, Luhan Yang, Kevin M. Esvelt, John Aach, Marc Guell, James E. DiCarlo, Julie E. Norville, and George M. Church. RNA-Guided Human Genome Engineering via Cas9. *Science*, 339(6121):823–826, February 2013.

[354] Seung Woo Cho, Sojung Kim, Jong Min Kim, and Jin-Soo Kim. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature Biotechnology*, 31(3):230–232, March 2013.

[355] Silvana Konermann, Mark D. Brigham, Alexandro E. Trevino, Julia Joung, Omar O. Abudayyeh, Clea Barcena, Patrick D. Hsu, Naomi Habib, Jonathan S. Gootenberg, Hiroshi Nishimasu, Osamu Nureki, and Feng Zhang. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, 517(7536):583–588, January 2015.

[356] Shengdar Q. Tsai, Nicolas Wyvekens, Cyd Khayter, Jennifer A. Foden, Vishal Thapar, Deepak Reyon, Mathew J. Goodwin, Martin J. Aryee, and J. Keith Joung. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nature Biotechnology*, 32(6):569–576, June 2014.

[357] Patrick D. Hsu, David A. Scott, Joshua A. Weinstein, F. Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J. Fine, Xuebing Wu, Ophir Shalem, Thomas J. Cradick, Luciano A. Marraffini, Gang Bao, and Feng Zhang. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology*, 31(9):827–832, September 2013.

[358] Julien Valton, Aurélie Dupuy, Fayza Daboussi, Séverine Thomas, Alan Maréchal, Rachel Macmaster, Kevin Melliand, Alexandre Juillerat, and Philippe Duchateau. Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation. *The Journal of Biological Chemistry*, 287(46):38427–38432, November 2012.

[359] Xuebing Wu, David A. Scott, Andrea J. Kriz, Anthony C. Chiu, Patrick D. Hsu, Daniel B. Dadon, Albert W. Cheng, Alexandro E. Trevino, Silvana Konermann, Sidi Chen, Rudolf Jaenisch, Feng Zhang, and Phillip A. Sharp. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nature Biotechnology*, 32(7):670–676, July 2014.

[360] Shengdar Q. Tsai, Zongli Zheng, Nhu T. Nguyen, Matthew Liebers, Ved V. Topkar, Vishal Thapar, Nicolas Wyvekens, Cyd Khayter, A. John Iafrate, Long P. Le, Martin J. Aryee, and J. Keith Joung. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology*, 33(2):187–197, February 2015.

[361] F. Ann Ran, Le Cong, Winston X. Yan, David A. Scott, Jonathan S. Gootenberg, Andrea J. Kriz, Bernd Zetsche, Ophir Shalem, Xuebing Wu, Kira S. Makarova, Eugene V. Koonin, Phillip A. Sharp, and Feng Zhang. In vivo genome editing using Staphylococcus aureus Cas9. *Nature*, 520(7546):186–191, April 2015.

[362] Kenneth Burke. *Philosophy of literary form: studies in symbolic action*. [Baton Rouge]: Louisiana state university press, 1941.