Spring 5-15-2019

Mapping and Functional Analysis of cis-Regulatory Elements in Mouse Photoreceptors

Andrew Hughes
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Molecular Genetics and Genomics

Dissertation Examination Committee:
Joseph C. Corbo, Chair
Shiming Chen
Donald F. Conrad
Kristen L. Kroll
Gary D. Stormo

Mapping and Functional Analysis of *cis*-Regulatory Elements in Mouse Photoreceptors

by

Andrew Everett Oliver Hughes

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2019
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

First, I would like to acknowledge my advisor, Joe Corbo, for his contributions to this work. Joe has been an outstanding mentor—patient, rigorous, and engaged. His enthusiasm for science is infectious, and he fosters a positive and productive training environment. In addition, Joe dedicates an extraordinary amount of time and mental energy to his students. From extended in-person meetings to rapid-fire emails at odd hours, I never doubted that Joe sincerely cared about my success as a graduate student. I am incredibly grateful for his mentorship.

I would also like to thank my thesis committee—Shiming Chen, Don Conrad, Kirsten Kroll, and Gary Stormo—for their encouragement, constructive criticism, and generosity with their time.

In addition, I would like to acknowledge the members of the Corbo lab. I am especially grateful to Jenny Enright for sharing the ATAC-seq project after she got the protocol up and running and generated the initial rod and cone datasets, providing the foundation of the project. In addition, Connie Myers has been an outstanding scientific mentor and molecular biology teacher (from cloning to cell culture), as well as a general knower of things. Furthermore, Connie made major contributions to the experiments presented in this thesis, including the construction of (and initial experiments with) the human CRE-seq library. I am also grateful to members of the Corbo lab past and present—including Jeongsook Kim-Han, Cindy Montana, Dan Murphy, Susan Shen, Matt Toomey, and Leo Volkov—for their encouragement, advice, and camaraderie.

I should also mention that several Washington University core facilities played essential roles in this project. In particular, I am grateful to Toni Sinwell and the Genome Technology Access Center for technical advice and sequencing services. Similarly, I would like to thank Jess Hoisington-Lopez from the DNA Sequencing Innovation Lab at the Edison Family Center for

Genome Sciences and Systems Biology for her sequencing expertise. Finally, I would like to thank Eric Martin and Brian Koebbe for running the High Throughput Computing Facility and always offering advice and help when I was troubleshooting computational problems.

Of course, I also need to thank the Washington University MSTP, including Wayne Yokoyama, Brian Sullivan, Christy Durbin, Liz Bayer, and Linda Perniciaro for all of their efforts in guiding me through the program. I am also grateful for my MSTP classmates—including Marina Avetisyan, David Cotter, Susan Shen, Ben Solomon, Xiaodi Wu, Christine Yokoyama, and Andrew Young—for their friendship as well as mapping out paths to successful PhDs that I could follow. Eventually.

Importantly, financial support for this work came from the National Institutes of Health, specifically the Washington University MSTP Training Grant (T32GM007200), the Vision Sciences Training Grant (T32EY013360), and grants to Joe Corbo (EY025196, EY026672, and EY024958).

Finally, I would like to acknowledge my family. My parents, Dave and Krista, taught me the value of persistence (among other things). I would also like to thank my brothers and sister-in-law—Doug, Dave, and Shannon—for their encouragement over the years. Last of all, I am incredibly grateful to my wife, Britta, for her constant love and support.

Andrew Hughes

Washington University in St. Louis

May 2019

ABSTRACT OF THE DISSERTATION

Mapping and Functional Analysis of *cis*-Regulatory Elements in Mouse Photoreceptors

by

Andrew Everett Oliver Hughes

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics and Genomics

Washington University in St. Louis, 2019

Professor Joseph C. Corbo, Chair

Photoreceptors are light-sensitive neurons that mediate vision, and they are the most commonly affected cell type in genetic forms of blindness. In mice, there are two basic types of photoreceptors, rods and cones, which mediate vision in dim and bright environments, respectively. The transcription factors (TFs) that control rod and cone development have been studied in detail, but the *cis*-regulatory elements (CREs) through which these TFs act are less well understood. To comprehensively identify photoreceptor CREs in mice and to understand their relationship with gene expression, we performed open chromatin (ATAC-seq) and transcriptome (RNA-seq) profiling of FACS-purified rods and cones. We find that rods have significantly fewer regions of open chromatin than cones (as well as >60 additional cell types and tissues), and we demonstrate that this uniquely closed chromatin architecture depends on the rod master regulator *Nrl*. Finally, we find that regions of rod- and cone-specific open chromatin are enriched for distinct sets of TF binding sites, providing insight into the *cis*-regulatory grammar of these cell types.

We also sought to understand how the regulatory activity of rod and cone open chromatin regions is encoded in DNA sequence. Cone-rod homeobox (CRX) is a paired-like homeodomain TF and master regulator of both rod and cone development, and CRX binding sites are by far the most enriched TF binding sites in photoreceptor CREs. The *in vitro* DNA binding preferences of

CRX have been extensively characterized, but how well *in vitro* models of TF binding site affinity predict *in vivo* regulatory activity is not known. In addition, paired-class homeodomain TFs bind DNA as both monomers and dimers, but whether monomeric and dimeric CRX binding sites have distinct regulatory activities is not known. To address these questions, we used a massively parallel reporter assay to quantify the activity of thousands native and mutant CRX binding sites in explanted mouse retinas. These data reveal that dimeric CRX binding sites encode stronger enhancers than monomeric CRX binding sites. Moreover, the activity of half-sites within dimeric CRX binding sites is cooperative and spacing-dependent. In addition, saturating mutagenesis of 195 CRX binding sites reveals that, while TF binding site affinity and activity are moderately correlated across mutations within individual CREs, they are poorly correlated across mutations from distinct CREs. Accordingly, we show that accounting for baseline CRE activity improves the prediction of the effects of mutations in regulatory DNA from sequence-based models. Taken together, these data demonstrate that the activity of CRX binding sites depends on multiple layers of sequence context, providing insight into photoreceptor gene regulation and illustrating functional principles of homeodomain TF binding sites.

Chapter 1: Introduction to gene regulation in mouse photoreceptors

1.1 Mechanisms of transcriptional regulation in mammals

In multicellular organisms, the development and maintenance of specialized cell types and complex tissues require the precise control of gene expression. In individual cell types, gene regulation is mediated by both *cis*- and *trans*-acting factors—i.e., *cis*-regulatory elements (CREs) and the transcription factors (TFs) that bind them. As a result, genetic variation in both protein-coding and noncoding regions of the genome play important roles in phenotypic diversity and disease susceptibility [1-3]. In humans and mice, the locations of protein-coding genes are largely known, and mutations in coding sequence result in predictable changes in amino acids based on the universal genetic code. In contrast, efforts to comprehensively map CREs in individual cell types are ongoing [4-6], and a comprehensive 'genetic code' for regulatory DNA has yet to be defined. Ultimately, understanding how CREs encode information requires a detailed understanding of mechanisms of transcriptional regulation—from the three-dimensional organization of the nucleus to the arrangement of TF binding sites within individual CREs.

1.1.1 Chromatin structure and genome organization

Within the nucleus, DNA associates with proteins to form chromatin, which provides the structural basis for hierarchical genome organization. The fundamental unit of chromatin is the nucleosome—146 bp of DNA wrapped around a histone octamer (two copies each of histones H2A, H2B, H3, and H4) [7]. Nucleosome core particles are bound by the linker histone, H1, which modulates the packing of chromatin fibers [8]. *In vitro*, these fibers have been visualized by cryo-electron microscopy as a double helix of tetra-nucleosomal subunits [9], but evidence suggests that variable numbers of nucleosomes cluster along chromatin fibers *in vivo* [10]. Importantly,

chromatin packing influences gene expression—tightly packed chromatin (heterochromatin) is largely untranscribed, while accessible chromatin (euchromatin) is available for transcription.

During interphase, the distinction between heterochromatin and euchromatin can be visualized directly by light microscopy [11-15]. In most cell types, heterochromatin localizes to the nuclear periphery and around nucleoli, with centric and pericentric regions forming dense foci of constitutive heterochromatin known as chromocenters [16]. Genomic regions that physically associate with the nuclear lamina have been mapped in both human and mouse cell lines, revealing that they are depleted of genes, transcriptionally repressed, AT-rich, and enriched for LINE repeats [17-19]. Regions of the genome that associate with nucleoli have similar properties, although they are enriched for satellite repeats as opposed to LINEs [20, 21]. Taken together, these data indicate that subnuclear localization reflects a functional partitioning of the genome into transcriptionally active and inactive compartments.

Recently, chromosome confirmation capture (3C) and related techniques (4C, 5C, and Hi-C) have revealed principles of genome organization at high resolution [22-27]. 3C-based methods quantify the frequency of physical interactions between pairs of genomic loci, yielding genome-wide maps of regional proximity. These maps demonstrate that intrachromosomal interactions are significantly more common than interchromosomal interactions (even over hundreds of megabases), validating microscopic studies suggesting that individual chromosomes segregate in discrete territories [22]. Furthermore, 3C-based methods show that the genome is partitioned into cell-type-specific A and B compartments, which alternate along chromosomes at megabase scale and correlate with gene density, GC content, repeat content, and transcriptional activity (similar to heterochromatin and euchromatin) [22, 26]. Within these compartments, 3C-based methods reveal the presence of highly self-interacting domains (so-called topologically associating domains, or TADs) [23-25, 27]. TADs are typically a few hundred kilobases in size and have well-defined

boundaries—i.e., interactions within TADs are much more frequent than interactions between TADs. These boundaries are highly enriched for architectural proteins (including CTCF, cohesin, and mediator) and are thought to constrain interactions between enhancers and gene promoters [25, 27-30].

1.1.2 CRE structure and function

Within TADs, the transcription of individual genes is coordinated by CREs—i.e., promoters, enhancers, and repressors. In current models of transcription initiation, TATA-binding protein (TBP) interacts with TBP-associated factors (TAFs) to recognize specific sequences within gene promoters, and this complex (TFIID) then recruits RNA polymerase II (Pol II) and general transcription factors (TFIIA, TFIIB, TFIIE, TFIIF, and TFIIH) to form the pre-initiation complex (PIC) [31, 32]. Historically, this process has been described in the context of recognizing a TATA box (TATAWAWR) ~30 bp upstream the transcription start site (TSS). However, genome-wide mapping of TSSs in humans reveals that the majority of core promoters (i.e., sequences within ~50 bp of TSSs that are sufficient to recruit Pol II) lack a TATA box [33]. Instead, mammalian core promoters harbor a diverse set of recognition sequences (including Inr, BRE, DCE, DPE, and MTE) and often multiple functional TSSs distributed over 50 to 100 bp [31, 32].

By themselves, core promoters drive transcription at very low levels, but interactions with enhancers can increase expression >100-fold. Enhancers are short stretches of DNA (typically 100 bp to 1 kb) that harbor clusters of TF binding sites (typically 6-13 bp) and encode cell-type-specific regulatory activity. The TFs that bind enhancers modulate expression by recruiting components of the basal transcriptional machinery (i.e., the PIC) as well as transcriptional coregulators (including chromatin modifiers, chromatin remodeling complexes, DNA methyltransferases, and architectural proteins).

Frequently, transcriptional coregulators directly alter chromatin structure. Chromatin modifiers (histone acetyltransferases, deacetylases, methyltransferases, and demethylases) catalyze the addition or removal of acetyl or methyl groups to lysine and arginine residues on exposed histone tails. These covalent modifications are strongly correlated with regulatory state: acetylation promotes chromatin relaxation and is associated with activation, while the effect of methylation depends on the specific lysine or arginine targeted and the number of methyl groups present [34]. Similarly, ATP-dependent chromatin remodeling complexes (e.g., the BAF, NuRD, and SWR1 complexes) dynamically reposition or eject nucleosomes and/or alter their histone composition [35]. Finally, DNA methyltransferases (DNMTs) catalyze the addition of methyl groups at cytosines in CpG dinucleotides, which promotes the formation of heterochromatin and gene silencing [36]. Thus, in addition to directly recruiting components of the PIC, CREs mediate targeted changes in chromatin structure as a means of regulating the expression of individual genes.

1.1.3 Methods to map CREs genome-wide

Over the past 10 years, advances in high-throughput sequencing have enabled CRE-associated chromatin features to be mapped genome-wide (e.g., TF occupancy, covalent histone modifications, and chromatin accessibility), facilitating systematic identification of CREs in diverse cell types and tissues (Fig. 1.1). In ChIP-seq, proteins are crosslinked to DNA, and chromatin is sheared to generate 100-300 bp DNA fragments [37]. Sequences bound by a protein of interest are then enriched by antibody selection, after which crosslinking is reversed and fragments are sequenced. This strategy can be applied to diverse classes of proteins, provided suitable antibodies, including TFs, covalent histone modifications, regulatory complexes, and Pol II.

In parallel, multiple strategies have been developed for mapping chromatin accessibility, which identify CREs independent of a specific protein of interest. In DNase-seq, native chromatin is gently digested with DNase I, which preferentially cleaves nucleosome-free regions of the genome, and fragments are then purified and sequenced [38]. More recently, mapping transposase-accessible chromatin (ATAC-seq) has proven to be as effective as DNase-seq, but with the advantage that it can be performed on as few as 500 cells, enabling open chromatin mapping of rare cell types purified by FACS [39].

Through consortium efforts, including the ENCODE Project and Roadmap Epigenomics Project, thousands of epigenomic profiles have now been generated in diverse cell types and tissues from humans as well as model organisms [4, 6]. These data provide a powerful resource for understanding how specific loci are regulated as well as general principles of transcriptional regulation. Nevertheless, the majority of these initial studies were conducted in bulk tissues, merging the epigenomic profiles of multiple cell types and precluding the analysis of cell-type-specific *cis*-regulatory architecture. Furthermore, while regions identified by epigenomic profiling are highly enriched for CREs, determining which regions are genuinely active CREs requires functional assays.

1.1.4 Methods to quantify CRE activity

Historically, the regulatory activities of CREs of interest have been quantified by reporter assays using colorimetric, luminescent, or fluorescent readouts. One limitation of these approaches is that they are not readily multiplexed, meaning CREs have to be analyzed one-at-a-time. Recently, the development of massively parallel reporter assays (MPRAs) has overcome this limitation by leveraging DNA barcodes to quantify the activity of large numbers of CREs simultaneously [40]. The experimental details of individual MPRAs vary, but general principles are illustrated by *cis*-regulatory element analysis by sequencing (CRE-seq) [41-46]. In CRE-seq, a library of test

sequences is generated by custom oligo synthesis or targeted capture, and these oligos are cloned upstream of a reporter gene harboring a CRE-specific barcode in the 3' UTR. This reporter library is then introduced into cells by electroporation, chemical transfection, or viral transduction, and CRE activity is quantified by harvesting RNA and DNA and sequencing barcodes. MPRAs, such as CRE-seq, have now been used in diverse cell types and tissues from both humans and model organisms to study principles of *cis*-regulatory grammar [41-50], to identify and engineer cell-type-specific enhancers [51], and to screen human variants for functional effects [52, 53].

MPRAs typically assess CREs in an episomal context (i.e., on a non-integrating plasmid), but assaying CREs in their native chromosomal context is necessary to definitively link specific regulatory variants to changes in gene expression. To address this, several studies have recently demonstrated the effectiveness of forward genetic screens using pooled libraries of CRISPR-Cas9 guide RNAs targeting noncoding regions [54-57]. Coupled with a single-cell readout of target gene activity (e.g., antibody staining or an in-frame fluorescent reporter fusion), this strategy enables high-resolution functional mapping of CREs *in situ*.

1.2 Regulation of mouse photoreceptor development

The retina is an extension of the central nervous system (CNS) composed of highly specialized cells types that receive and process visual information. In vertebrates, there are seven major classes of retinal cells: rod and cone photoreceptors, bipolar cells, horizontal cells, amacrine cells, retinal ganglion cells and Müller glia. These cells have a precise laminar architecture consisting of three nuclear layers (the ganglion cell layer and the inner and outer nuclear layers) and two synaptic layers (the inner and outer plexiform layers). Photoreceptors are located in the outer nuclear layer and express light-sensitive opsins that initiate phototransduction. In the inner retina, bipolar, horizontal and amacrine cells process and relay information from photoreceptors to retinal ganglion cells, which project to the brain.

Despite its complexity, the retina is among the best understood regions of the CNS. The mouse retina, in particular, has emerged as a powerful system for studying the genetics and physiology of highly specialized cell types and intricate neural circuits—from single-cell transcriptome profiling of tens of thousands of cells [58, 59] to high-resolution reconstruction of inner retinal circuitry by serial electron microscopy [60]. Furthermore, mouse photoreceptors have proven valuable as a model for understanding the transcriptional networks that regulate neural development (Fig. 1.2).

1.2.1 TFs regulating mouse photoreceptor development

Early in development (E11.5), expression of the homeodomain TF OTX2 restricts retinal progenitor cells to photoreceptor or bipolar cell fates, both of which fail to develop in retina-specific *Otx2* knockout mice [61-63]. Shortly after (E12.5), *Otx2* activates *Crx*, a second homeodomain TF, which plays a key role in photoreceptor maturation [64-66]. In *Crx* knockout mice, photoreceptors lack outer segments and have no detectable responses to light [67]. Furthermore, CRX ChIP-seq shows that CRX binds thousands of CREs flanking photoreceptor genes [68], many of which are down-regulated in *Crx* knockout retinas [69, 70]. Taken together, these data establish *Otx2* and *Crx* as central TFs within photoreceptor gene regulatory networks.

Downstream of *Otx2* and *Crx*, additional TFs are required for photoreceptor maturation and survival, but these TFs generally regulate more restricted subsets of genes: *Neurod1* [71, 72], *Mef2d* [73, 74], and *Rax* [75]. Furthermore, several TFs are required for the expression of cone-specific genes, especially the patterning of cone opsins (*Opn1mw* and *Opn1sw*): *Nr2f1* and *Nr2f2* [76], *Onecut1* and *Onecut2* [77], *Rora* [78], *Rxrg* [79], *Sall3* [80], and *Thrb* [81]. In rods, *Rorb* directly activates *Nrl* [82-85], which orchestrates extensive rod-specific transcriptional changes [82, 86-88], including the activation of the rod-specific TFs *Esrrb* [89], *Mef2c* [90], and *Nr2e3*

[91-93]. Thus, detailed genetic studies demonstrate the power of the mouse retina as a model for understanding the transcriptional mechanisms regulating the development of complex cell types.

1.2.2 Genetic diseases of photoreceptors

Photoreceptors play a critical role in human vision, and they are the most commonly affected cell type in inherited blindness [94]. The genetic architecture of blindness is complex: coding mutations in over 120 genes have been implicated in just three forms of retinal disease (retinitis pigmentosa, cone or cone-rod dystrophy, and Leber congenital amaurosis), and distinct mutations within the same gene can produce different disease pathologies and clinical presentations [95, 96]. Many of the affected genes have well-characterized roles in photoreceptor gene regulation (e.g., *Crx*, *Nrl*, *Nr2e3*, *Neurod1*, *Otx2*, and *Rax2*) and phototransduction (e.g., *Cnga3*, *Cngb3*, *Gnat2*, *Guca1b*, *Gucy2d*, *Pde6a*, *Pde6b*, *Pde6c*, *Pde6g* and *Rho*) [94, 96]. Nevertheless, the mechanisms by which mutations in these genes produce specific disease phenotypes are not completely understood.

1.3 Aims and scope of thesis

As described above, rod and cone photoreceptors are highly specialized sensory neurons that mediate vision and are the most commonly affected cell types in genetic forms of blindness. Detailed molecular and genetic studies have defined the transcriptional networks underlying rod and cone identity, but how rod- and cone-specific gene expression is encoded in regulatory DNA is less well understood. In particular, the locations of these CREs have not been systematically identified, and the sequence features that encode rod- and cone-specific regulatory activity are not known. Therefore, **the aims of this thesis are (1) to comprehensively map and characterize rod- and cone-specific CREs, and (2) to determine how photoreceptor CREs encode regulatory information in primary sequence.**

In Chapter 2, I use ATAC-seq and RNA-seq to profile chromatin accessibility and gene expression in FACS-purified rods and cones. This approach yields highly reproducible cell-type-specific maps of candidate CREs, which are moderately correlated with transcriptional activity genome-wide. Furthermore, these data reveal that rods are depleted of open chromatin compared to cones (as well as >60 control cell types and tissues), i.e., that they have a uniquely closed chromatin architecture. This phenotype likely reflects the unusual distribution of chromatin in rods—suggesting a correlation between global changes in nuclear organization and chromatin structure at the resolution of individual CREs. Finally, I identify sequence features (mono- and di-nucleotide content as well as TF binding sites) enriched in photoreceptor open chromatin, providing insight into the *cis*-regulatory grammar of mouse rods and cones.

In Chapter 3, I use CRE-seq to perform detailed functional analyses of sequence features that influence the activity of CRX binding sites. First, I identify sequence features and orthogonal epigenomic measurements that are correlated with the wild-type activity of >1200 CRX-bound regions in the mouse genome. I then assay the effect of inactivating mutations in >1700 CRX binding sites within these elements, which reveals that dimeric CRX binding sites encode stronger enhancers than monomeric CRX binding sites. Furthermore, mutating pairs of CRX sites within the same CRE individually and in combination reveals that CRX sites frequently act cooperatively. Finally, I examine the effect of all possible single nucleotide substitutions in a 13-bp window overlapping 195 CRX binding sites, which reveals that changes in TF binding site affinity are moderately correlated with changes in activity within individual CREs, but poorly correlated across mutations from different CREs.

In the Appendix, I describe my initial efforts to adapt these approaches to study the functional impact of genetic variation on human retinal CREs, including the prioritization of candidate pathogenic variants in cases of inherited blindness.

Taken together, these studies provide insight into the mechanisms by which photoreceptor CREs encode regulatory information in DNA sequence. In addition, they provide valuable reference datasets for studying the regulation of individual rod and cone genes, and they reveal principles of *cis*-regulatory grammar that have implications for developing quantitative models of CRE activity. Therefore, these results have the potential to enhance the functional interpretation of noncoding variation and accelerate the identification of regulatory sequences involved in complex traits and human disease.

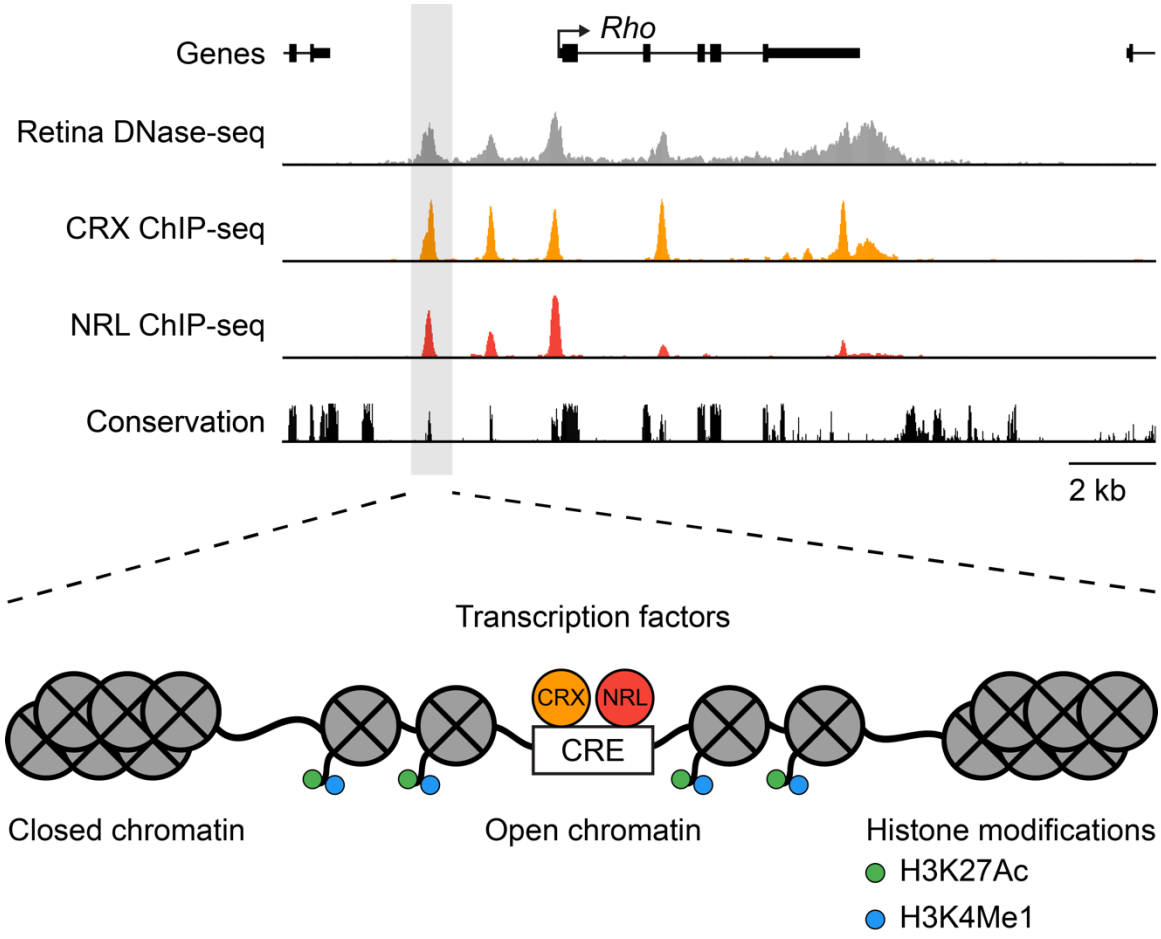


Figure 1.1. Identification of photoreceptor CREs by epigenomic profiling. Upper panel: tracks from the UCSC genome browser showing a 20 kb window centered on the mouse rhodopsin (*Rho*) locus, including open chromatin (DNase-seq), TF occupancy (CRX and NRL ChIP-seq), and conservation data (phastCons 60-way). Chromatin accessibility, TF occupancy, and conservation profiles align with the *Rho* promoter, as well as several upstream and downstream enhancers. Lower panel: schematic illustrating the chromatin structure of a single upstream cis-regulatory element (CRE). As shown in both panels, this CRE overlaps a region of open chromatin (is nucleosome-free) and is bound by two TFs (CRX and NRL). Also, this CRE is flanked by adjacent nucleosomes (closed chromatin) with covalent histone modifications typical of active enhancers (H3K27Ac and H3K4Me1).

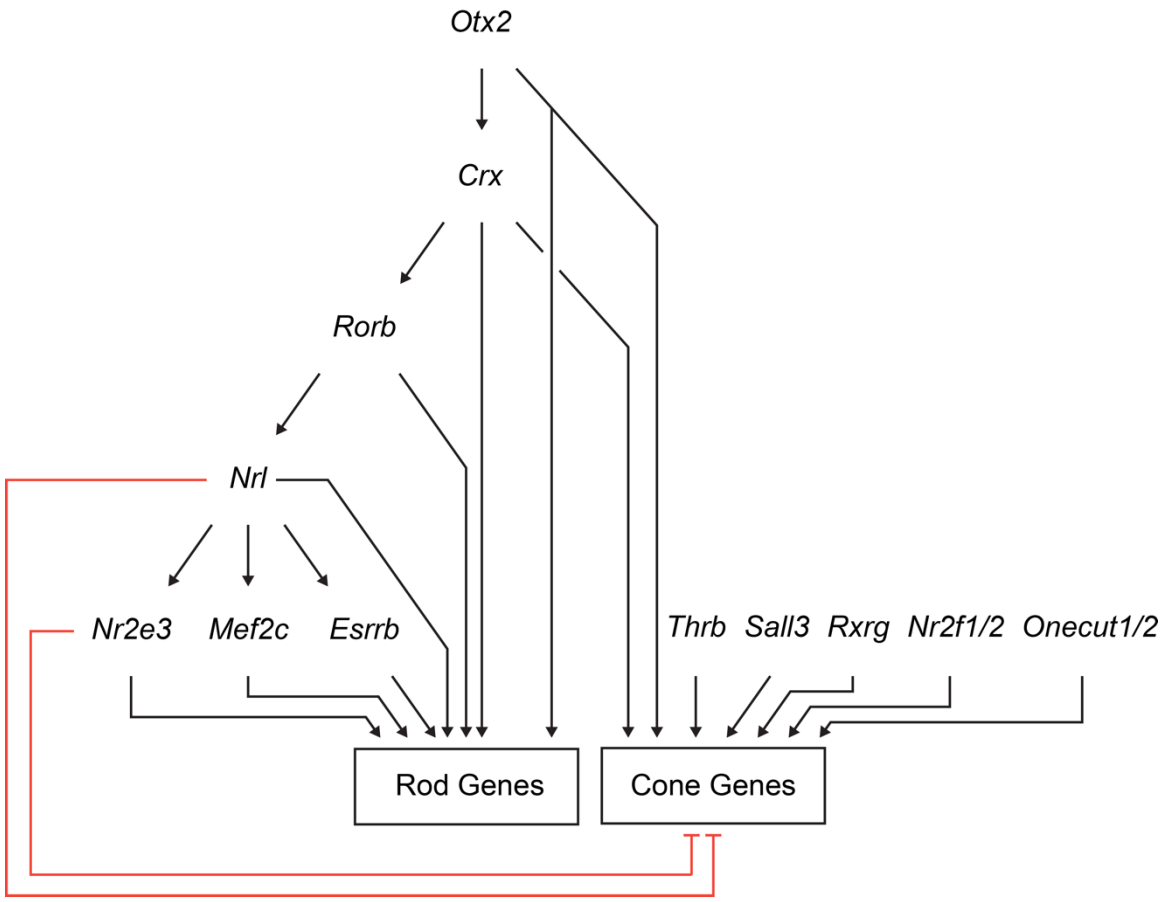


Figure 1.2. Transcriptional regulation of mouse photoreceptor development. *Otx2* and *Crx* are master regulators of photoreceptor development, controlling the expression of additional photoreceptor TFs as well as many rod and cone genes. Downstream of *Crx*, *Nrl* is the master regulator of rod differentiation, activating rod-specific TFs (and other rod genes) and repressing cone genes.

Chapter 2: Cell-type-specific epigenomic analysis reveals a uniquely closed chromatin architecture in mouse rod photoreceptors

Rod photoreceptors are specialized neurons that mediate vision in dim light and are the predominant photoreceptor type in nocturnal mammals. The rods of nocturnal mammals are unique among vertebrate cell types in having an 'inverted' nuclear architecture, with a dense mass of heterochromatin in the center of the nucleus rather than dispersed clumps at the periphery. To test if this unique nuclear architecture is correlated with a unique epigenomic landscape, we performed ATAC-seq on mouse rods and their most closely related cell type, cone photoreceptors. We find that thousands of loci are selectively closed in rods relative to cones as well as >60 additional cell types. Furthermore, we find that the open chromatin profile of photoreceptors lacking the rod master regulator *Nrl* is nearly indistinguishable from that of native cones, indicating that *Nrl* is required for selective chromatin closure in rods. Finally, we identified distinct enrichments of transcription factor (TF) binding sites in rods and cones, revealing key differences in the *cis*-regulatory grammar of these cell types. Taken together, these data provide insight into the development and maintenance of photoreceptor identity, and highlight rods as an attractive system for studying the relationship between nuclear organization and local changes in gene regulation.

2.1 Introduction

Photoreceptors are light-sensitive neurons that express opsins tuned to specific wavelengths of light. In the mouse retina >95% of photoreceptors are rods, which express rhodopsin (*Rho*) and mediate vision in dim light [97]. In contrast, cone photoreceptors express both short- and/or medium-wavelength opsins (*Opn1sw* and *Opn1mw*, respectively) and mediate bright light vision as well as color vision [98, 99] (Fig. 2.1A,B). The rods of nocturnal mammals have evolved a unique nuclear architecture to enhance visual sensitivity in low-light environments [100]. Whereas

the nuclei of nearly every vertebrate cell type harbors multiple discrete clusters of heterochromatin largely localized to the nuclear periphery, the rods of nocturnal mammals harbor a dense central core of heterochromatin surrounded by a ring of euchromatin (Fig. 2.1C).

Emerging evidence indicates that the three-dimensional organization of chromatin within the nucleus is essential for regulating gene expression [101, 102]. This organization is fundamentally hierarchical: within the nucleus, chromosomes segregate into discrete territories [103, 104], and they are partitioned into megabase-scale compartments (similar to euchromatin and heterochromatin) [22], which are further organized into topologically associating domains (TADs) [25]. Within TADs, architectural proteins (including CTCF, mediator, and cohesin) direct chromatin looping, mediating local interactions between transcriptional enhancers (i.e., *cis*-regulatory elements, or CREs) and their target genes [27, 105]. While the unique localization of chromatin in rods has been described extensively at an ultrastructural level [100], whether or not this higher-order reorganization is correlated with changes in the regulatory landscape of rods (and ultimately gene expression) is not known.

At the molecular level, the inverted nuclear organization of rods depends on the silencing of key nuclear envelope proteins, including lamin B receptor (*Lbr*) and lamin A/C (*Lmna*) [106]. Although it is not clear how these specific loci are regulated in rods, a tremendous amount is known about the transcriptional networks underlying rod and cone identity [107, 108]. Commitment to a photoreceptor fate is determined early in development by a pair of 'K50' homeodomain (HD) transcription factors (TFs)—*Otx2* and *Crx*—whose binding specificity is determined, in part, by a lysine (K) in position 50 of the homeodomain [109]. Beginning at E11.5, *Otx2* provides a necessary signal for photoreceptor specification [63, 110]. *Otx2* then directly activates *Crx* (first detectable at E12.5), which is also required for photoreceptor maturation [64-66]. Comparative expression profiling of wild-type and $Crx^{-/-}$ retinas has shown that *Crx* regulates

additional photoreceptor TFs as well as components of the phototransduction machinery [69, 70]. Furthermore, *Crx*$^{-/-}$ photoreceptors lack outer segments and fail to respond to light [67]. Thus, detailed molecular studies have placed *Crx* at the core of the photoreceptor transcriptional network.

The decision for a photoreceptor progenitor to differentiate into a rod versus a cone photoreceptor is driven by the TF *Nrl*, which is both necessary and sufficient for rod differentiation [86, 111]. Expression profiling has demonstrated that *Nrl* regulates a large set of rod-specific genes [87, 88, 112], and the rods of *Nrl*$^{-/-}$ mice are transfated into cells with many of the features of native blue cones, including a conventional nuclear architecture with multiple heterochromatin clusters at the periphery [86, 113, 114]. In addition to regulating genes that contribute directly to phototransduction, *Nrl* activates additional TFs required for subsets of rod gene expression, including the nuclear receptors (NRs) *Nr2e3* [93, 115] and *Esrrb* [89], as well as the myocyte enhancer factor 2 family member, *Mef2c* [90].

Expression profiling of wild-type and TF mutant mouse retinas has rapidly advanced the understanding of how photoreceptor TFs are organized into regulatory networks [70, 88, 112, 116, 117]. In parallel, significant advances have been made in the discovery of retinal CREs across the mouse genome. Specifically, DNase-seq has been used to map regions of open chromatin (containing candidate CREs) in whole retina [118], and ChIP-seq has been used to elucidate the genome-wide occupancy of individual TFs—CRX [68], NRL [119], and MEF2D [73]—also in whole retina. While these studies have proven highly informative, profiling whole retina is limited in that the ascertainment of CREs is biased towards highly abundant cell types, and the specific cell types underlying individual signals is ambiguous.

To begin to elucidate the epigenomic landscape of individual photoreceptor subtypes, we utilized the assay for transposase accessible chromatin using sequencing (ATAC-seq) [39] to

profile purified populations of rods, cones, and *Nrl*$^{-/-}$ photoreceptors. In addition, we performed RNA-seq on flow-sorted rods and *Nrl*$^{-/-}$ photoreceptors, in order to correlate changes in chromatin accessibility with changes in gene expression. Strikingly, we find that mouse rods display a global reduction in open chromatin (relative to native cones as well as >60 additional mouse cell types and tissues profiled by the ENCODE project), and that this reduction depends on *Nrl*. Through comparative analysis of these datasets, we define thousands of photoreceptor class- and subtype-specific candidate CREs, and we find that distinct subsets of these elements are enriched for different sets of TF binding sites. Taken together, these data reveal that *Nrl* mediates a uniquely closed chromatin architecture in rods, and they provide a framework for elucidating the cell-type-specific *cis*-regulatory grammar of rods and cones.

2.2 Photoreceptor ATAC-seq yields cell-type-specific maps of open chromatin

To isolate purified populations of adult mouse photoreceptors, we performed fluorescence-activated cell sorting (FACS) on dissociated retinas harvested from 8-week-old mice harboring transgenic reporter constructs (Fig. 2.1, Supplemental Table S2.1). Specifically, we obtained rods from *Nrl-eGFP* mice (Fig. 2.1D,G) [88], native 'green' cones from *Opn1mw-eGFP* mice (Fig. 2.1E,H) [120], and *Nrl*$^{-/-}$ photoreceptors (putatively blue cones) from *Nrl*$^{-/-}$;*Nrl-eGFP* mice (Fig. 2.1F,I) [88]. We performed ATAC-seq on all three sorted cell types and RNA-seq on sorted rods and *Nrl*$^{-/-}$ photoreceptors, yielding reproducible chromatin accessibility and expression profiles (Pearson correlation coefficients between biological replicates of 0.86-0.99 for ATAC-seq and 0.95-1.00 for RNA-seq) (Supplemental Fig. S2.1).

ATAC-seq of purified photoreceptors revealed cell-type-specific patterns of open chromatin flanking canonical rod- and cone-specific genes: *Rho* (rod-specific), *Opn1mw* (cone-specific), and *Opn1sw* (cone-specific) (Fig. 2.2A-C). In general, rod ATAC-seq exhibited high

concordance with whole-retina DNase-seq (as well as CRX ChIP-seq and NRL ChIP-seq), especially near rod-specific genes (Fig. 2.2A) [68, 118, 119]. This result is expected, given that rods constitute >75% of cells in the mouse retina [97]. In contrast, cones represent only 2% of cells in the mouse retina. Therefore, we hypothesized that cell-type-specific ATAC-seq would be more sensitive than whole-retina DNase-seq for detecting cone-specific regulatory elements. Indeed, ATAC-seq of green cones and *Nrl*$^{-/-}$ photoreceptors revealed many regions of open chromatin that were not previously detected by epigenomic profiling of whole retina, especially near cone-specific genes (Fig. 2.2B,C). While we observed robust cell-type-specific ATAC-seq peaks flanking known rod- and cone-specific genes, we noted that even highly cell-type-specific loci frequently harbored peaks open in all three photoreceptor types (Supplemental Fig. S2.2). This multiplicity of both cell-type-specific and shared open chromatin elements in photoreceptors is similar to the 'locus complexity' of cell-type-specific enhancers recently described by Gonzalez *et al.* [121].

Although the majority of mouse retinal cells are photoreceptors, we hypothesized that many previously identified whole-retina DNase-seq peaks derive from non-photoreceptor cell types. To compare the chromatin accessibility profiles of flow-sorted photoreceptors and whole retina, we first merged peak calls from cell-type-specific ATAC-seq (Supplemental Table S2.2) and whole-retina DNase-seq to produce a reference set of 60,414 candidate regulatory elements. We then plotted the chromatin accessibility profile of each cell or tissue over these intervals (Fig. 2.2D). We found that there were many whole-retina DNase-seq peaks that did not correspond to photoreceptor ATAC-seq peaks (Fig. 2.2D, yellow box). For example, whereas >95% of rod ATAC-seq peaks overlapped whole-retina DNase-seq peaks, <50% of whole-retina DNase-seq peaks overlapped rod ATAC-seq peaks (Supplemental Table S2.3). To validate that this reduced overlap was due to enhanced specificity (vs. reduced sensitivity), we also examined the overlap of

ATAC-seq peaks with photoreceptor-specific ChIP-seq data. We found that rod and cone ATAC-seq peaks overlapped >90% of whole-retina CRX ChIP-seq peaks (i.e., photoreceptor-specific regulatory elements). Furthermore, rod ATAC-seq peaks overlapped >90% of NRL ChIP-seq peaks (i.e., rod-specific regulatory elements). Taken together, these data suggest that photoreceptor ATAC-seq revealed the rod- and cone-specific subsets of open chromatin elements identified by whole-retina DNase-seq, and that a substantial number of candidate CREs identified by whole-retina DNase-seq belong to non-photoreceptor cell types.

We also found that the open chromatin profiles of green cones and *Nrl*$^{-/-}$ photoreceptors were nearly indistinguishable, indicating that the well-described cone-like features of *Nrl*$^{-/-}$ photoreceptors [86, 113, 114] are encoded at the level of chromatin accessibility. This finding was supported by principal component analysis (PCA) as well as hierarchical clustering of the genome-wide open chromatin profiles of photoreceptors and control tissues (Fig. 2.2E, Supplemental Fig. S2.3). As *Nrl*$^{-/-}$ photoreceptors preferentially express the short-wavelength opsin (*Opn1sw*), we will henceforth refer to these cells as 'blue cones'.

2.3 *Nrl* is required for global chromatin closure in rods

In our global comparison of rod and cone open chromatin profiles, we noted a significant excess of cone-specific open chromatin relative to rods (Fig. 2.2D, red box). When we reviewed the genome-wide distribution of these peaks, we found that many occurred in long runs (hundreds of kilobases to tens of megabases), overlapping regions that frequently harbored open chromatin peaks in additional (non-photoreceptor) cell types and tissues (Fig. 2.3A,B). Therefore, many of the apparently cone-specific regions of open chromatin are more accurately described as regions that are selectively closed in rods. This finding suggested that rods have a more closed chromatin landscape than other cell types and tissues.

To quantify this observation, we examined the distribution of open chromatin signal (ATAC-seq or DNase-seq reads) across the mouse genome in 64 additional cell types and tissues (Fig. 2.3C, Supplemental Table S2.4). We divided the genome into fixed 50 kb windows, scored each window for normalized open chromatin signal (normalized ATAC-seq or DNase-seq reads), and then plotted the cumulative distribution (proportion of windows with coverage less than or equal to each observed value) for each cell type or tissue. This analysis revealed that the genome-wide distribution of open chromatin signal in mouse rods was shifted relative to every other cell type and tissue we examined in a pattern consistent with a more closed chromatin landscape. In *Nrl*$^{-/-}$ photoreceptors, in contrast, these rod-closed regions are open to a similar extent as in other tissues, demonstrating that *Nrl* is necessary for the global chromatin closure phenotype in rods.

We next asked if global differences between the accessible chromatin landscapes of rods and cones were correlated with changes in gene expression. To address this, we mapped individual rod and blue cone ATAC-seq peaks to the nearest transcription start site (TSS) and examined the expression of the corresponding gene (Fig. 2.3D, Supplemental Tables S2.5-S2.6). In general, changes in chromatin accessibility were directionally correlated with changes in gene expression, i.e., genes near rod-specific peaks tended to have higher expression in rods, while genes near cone-specific peaks tended to have higher expression in cones (Fig. 2.3D, Supplemental Fig. S2.4A). We noted, however, that genes near cone-specific peaks had lower expression in both rods and cones relative to genes near rod-specific peaks or shared peaks (Supplemental Fig. S2.4A), indicating that cone-specific peaks are located in regions that have lower average transcriptional activity in both photoreceptor types. Furthermore, cone-specific open chromatin elements were located significantly farther from annotated genes compared to rod-specific or shared open chromatin elements (median distance 50 kb vs. 18 kb) (Supplemental Fig. S2.4B). Finally, rod-specific regions of open chromatin were highly enriched for gene ontology (GO) terms related

specifically to photoreceptor biology, while cone-specific regions were enriched more generally for terms related to neurodevelopment (Supplemental Fig. S2.5). Thus, regions of open chromatin that are selectively closed in rods appear to be depleted of genes, especially photoreceptor genes, and have lower transcriptional activity. These findings suggest that rod-specific chromatin closure may not reflect targeted gene silencing, and may instead be secondary to large-scale alterations in nuclear organization (Fig. 2.3E).

2.4 *Lmna* expression is selectively downregulated in rods

A prior study showed that expression of either lamin B receptor (encoded by *Lbr*) or lamin A/C (encoded by *Lmna*) is required to maintain a conventional nuclear architecture in non-rod cell types [106]. To determine whether expression of these two genes is associated with selective chromatin closure in rods, we examined the open chromatin and transcriptional profiles of these loci in detail (Fig. 2.4). The *Lbr* locus harbors a single open chromatin peak overlapping the TSS, with comparable ATAC-seq signal in rods, green cones and blue cones (Fig. 2.4A). It was previously shown by antibody staining that that lamin B receptor is downregulated in both rods and cones as they differentiate [106]. Nevertheless, we detected modest levels of *Lbr* transcript in both rods and blue cones (Fig. 2.4C), suggesting that either the level of lamin B receptor in adult photoreceptors is regulated post-transcriptionally, or that it is below the limit of detection by antibody staining. While the open chromatin profile surrounding *Lbr* is similar in rods and cones, two peaks at the *Lmna* locus are selectively closed in rods—one overlapping the gene promoter and another ~6.5 kb upstream (Fig. 2.4B, red boxes). The rod-specific closure of these peaks is correlated with a marked reduction in the level of *Lmna* transcript in rods (Fig. 2.4D), consistent with the rod-specific reduction in Lamin A/C protein levels reported previously [106]. Taken together, these data indicate that rods selectively downregulate *Lmna* at the transcript level, and this downregulation may be mediated by the selective closure of two upstream open chromatin regions.

Furthermore, these findings indicate that NRL mediates chromatin closure at the *Lmna* locus, either directly or indirectly, offering a mechanistic link between the expression of a key rod cell fate determinant and the cell's inverted nuclear architecture.

2.5 Photoreceptor open chromatin is enriched for binding sites for photoreceptor TFs

Having characterized photoreceptor open chromatin globally, we next sought to identify local sequence features that mediate the regulatory activity of individual CREs. Combining ATAC-seq peaks from both rods and cones, we detected a total of 55,161 regions of open chromatin in photoreceptors. For each cell type, we partitioned peaks into "promoters" (<1 kb upstream and <100 bp downstream of the nearest TSS) and "enhancers" (>1 kb upstream or >100 bp downstream of the nearest TSS), in light of work demonstrating that cell-type-specific regulatory elements are preferentially enriched among TSS-distal elements [122]. Supporting the idea that these regions constitute distinct functional elements, we found that promoter elements in all three photoreceptor types were centered on stronger and broader enrichments of ATAC-seq signal, phylogenetic conservation, and GC content, and that promoter peaks were more strongly correlated with gene expression (Supplemental Fig. S2.6).

To identify TF binding sites enriched in photoreceptor ATAC-seq peaks, we tested 319 known sequence motifs curated by the HOMER suite of sequence analysis tools for overrepresentation (Fig. 2.5A, Supplemental Fig. S2.7, Supplemental Tables S2.7-S2.8) [123]. Whereas promoter peaks were strongly enriched for motifs corresponding to ubiquitous transcriptional regulators (Supplemental Fig. S2.7), enhancer peaks were enriched for motifs bound by known photoreceptor TFs (Fig. 2.5A,B). Among enhancer elements in both rods and cones, the most strongly enriched motif corresponded to the zinc-finger (ZF) architectural protein CTCF. CTCF is frequently among the most enriched motifs in open chromatin from most cell

types [122] and is thought to mediate interactions between regulatory elements via chromatin looping [27, 124]. The second most enriched motif, 'CTAATCC', represented the binding of a K50 HD TF, most likely the photoreceptor master regulator CRX [68]. In addition, this motif is also bound by OTX2 (highly expressed during photoreceptor development but at reduced levels in adulthood) [125] and can be bound by SIX6 as well [126].

While CTCF and CRX were by far the most enriched motifs in photoreceptor open chromatin, we also observed modest enrichments of additional motifs that again corresponded to binding sites recognized by well-characterized photoreceptor TFs (Fig. 2.5A, B). These included Q50 HD (TAATTA), basic helix-loop-helix (bHLH) (CATATG), MADS ($CC[A/T]_8GG$), NR (AGGTCA), GATA (GATA), and bZIP (TGANTCA) families. The enrichment of non-CRX binding sites is of considerable interest, as it suggests that specific TFs may cooperate with CRX to shape the regulatory activity of photoreceptor-specific CREs. RAX, for example, is the most highly expressed Q50 HD TF (which has glutamine [Q] in position 50 of the homeodomain) in both rods and cones. It has recently been shown that RAX plays an important role in photoreceptor development and is essential for the survival of mature cones [75]. bHLH motifs are likely bound by NEUROD1, which is expressed in developing and mature photoreceptors (both rods and cones) [127]. The MADS family members MEF2D (rods and cones) and MEF2C (rod-specific) are additional essential photoreceptor TFs, and MEF2D in particular has been shown to be recruited by CRX to photoreceptor-specific binding sites [73]. The specific role of individual NR motifs is more complex, as rods and cones express distinct sets of NR TFs—RORB, ESRRB, and NR2E3 are significantly upregulated in rods, while RXRG and THRB are expressed in cones (Fig. 2.5B). Finally, we observed relative enrichment in cones of a bZIP-type motif (TGANTCA), which may be bound by NRL, a bZIP TF in the MAF subfamily (though MAF family members typically prefer an extended consensus sequence, TGCTGANTCAGCA).

Motifs overrepresented in photoreceptor ATAC-seq peaks were enriched for co-occurrence (overrepresentation of pairs of motifs within individual ATAC-seq peaks) (Supplemental Fig. S2.8), suggesting they may have cooperative roles in mediating the activity of individual regulatory elements. In particular, we identified a candidate regulatory hub consisting of K50 HD (CRX), NR (RORB), MADS (MEF2D), MAF (NRL), and bZIP (NRL?) motifs that were co-enriched in photoreceptor open chromatin (Supplemental Fig. S2.8). Notably, while we did not observe a robust enrichment of MAF motifs by themselves, we did observe an enrichment of motif pairs involving MAF motifs. This suggests that NRL may be recruited to rod-specific enhancers cooperatively with additional cell-type-specific TFs, analogous to the mechanism of photoreceptor-specific MEF2D occupancy described recently [73]. While overrepresented enhancer TF binding site motifs were typically enriched for co-occurrence, peaks harboring CTCF motifs were depleted of other photoreceptor-specific motifs (Supplemental Fig. S2.8).

Consistent with previous work, we observed modest preferences in spacing and orientation between pairs of enriched motifs [68] (Supplemental Fig. S2.9). Nevertheless, specific configurations account for a minority of enriched motif pairs, supporting the idea that photoreceptor regulatory activity is encoded by a flexible sequence grammar, as suggested by previous studies [42, 68]. Finally, we noted that while the majority of rod and cone ATAC-seq peaks were shared across cell types, many enriched motifs are likely bound by TFs that are differentially expressed in rods and cones *in vivo* (Fig. 2.5B). This raises the possibility that CREs open in both rods and cones encode cell-type-specific regulatory activity via the recruitment of distinct arrays of TFs and/or by competition among differentially expressed TFs (e.g., NRs) for common binding sites. For example, although NRL is a rod-specific TF, >80% (1603/1935) of NRL ChIP-seq peaks overlap elements that are accessible in both rods and cones (Fig. 2.2D). Thus, differential expression of *trans* factors may be an important mechanism by which photoreceptor

regulatory elements encode cell subtype-specific functions within shared *cis*-regulatory elements.

In addition to identifying enrichments of known TF binding site motifs, we performed *de novo* motif discovery (Supplemental Tables S2.9-S2.10). In general, *de novo* motifs were highly concordant with known motifs that were found to be enriched. A key exception was the *de novo* motif corresponding to a K50 HD family member (likely bound by CRX) (Fig. 2.6A). While previous *in vitro* studies have shown that CRX exhibits a strong preference for the consensus CTAATCCC [109], photoreceptor ATAC-seq peaks showed a preference for CRX motifs in a paired configuration on opposite strands separated by exactly three nucleotides (TAAT[N]$_3$ATTA), a well-known dimer configuration of HD TFs [128, 129]. Furthermore, TAAG is highly enriched in photoreceptor open chromatin, especially in a paired configuration (TAAG[N]$_3$CTTA or TAAT[N]$_3$CTTA). This finding indicates that, in certain contexts, the fourth position of the TAAT homeodomain core tolerates a 'G' better than would be expected from quantitative gel shift assays [109]. This conclusion was also reached by a previous study that analyzed hundreds of variants within a single photoreceptor promoter [41]. When we examined individual k-mers underlying the paired K50 HD *de novo* motif, they were highly enriched for the canonical CRX monomer motif (CTAATCC), the 3' G variant (CTAAGCC), as well as homotypic (TAAT[N]$_3$ATTA or TAAG[N]$_3$CTTA) and especially heterotypic (TAAT[N]$_3$CTTA) dimer configurations (Fig. 2.6A).

We next asked if distinct configurations of CRX monomeric and dimeric motifs had functional significance with respect to the *cis*-regulatory activity of individual photoreceptor CREs (Fig. 2.6B). Previously, a library of 84-bp CRX-bound sequences (including 865 WT sequences and 865 versions with CRX binding sites eliminated by point mutation) were assayed for activity in explanted mouse retina via CRE-seq, a massively parallel reporter assay [42]. We re-analyzed these data with respect to CRX motif configuration and did not find strong differences in

expression depending on the presence of monomeric (TAAT or TAAG cores), homotypic 3' G dimeric (TAAG[N]$_3$CTTA), or heterotypic dimeric (TAAT[N]$_3$CTTA) k-mers. However, we did observe a significant repressive effect for the homotypic TAAT dimer configuration (TAAT[N]$_3$ATTA) (Fig. 2.6B). Furthermore, examining the corresponding control sequences with mutated CRX binding sites, we found that this repressive effect depended on the presence of CRX binding sites (Fig. 2.6B). Accordingly, these data demonstrate that distinct arrangements of CRX motifs prevalent in endogenous CREs can yield significant differences in *cis*-regulatory activity.

Having identified a repressive effect of TAAT[N]$_3$ATTA CRX motifs, we asked if additional motifs enriched in photoreceptor open chromatin contributed to photoreceptor CRE activity. Consistent with the original analysis by White *et al*., we did not observe significant differences in reporter activity among endogenous CRX-bound sequences due to the presence or absence of CRX sites, suggesting the importance of additional sequence features and context (Fig. 2.6C) [42]. In contrast, we found that constructs harboring bHLH (NEUROD1), MADS (MEF2C and MEF2D), or NR (RORB, NR2E3, and ESRRB) motifs had significantly higher expression compared to those without (Fig. 2.6C). This activation was retained (though reduced) in control sequences in which CRX sites were eliminated by point mutations, suggesting that CRX may act cooperatively with these TFs to drive enhancer activity. Thus, high-throughput analysis of photoreceptor CREs suggests that specific TF binding sites enriched in photoreceptor open chromatin act as potent transcriptional activators *in vivo*.

2.6 Rod- and cone-specific regions of open chromatin are enriched for distinct TF binding sites

Finally, we asked if rod- and cone-specific ATAC-seq peaks were enriched for distinct sequence features. To define rod- and cone-specific regions, we used DESeq2 [130] to test for differences in accessibility between rods and cones (Methods). This analysis yielded 48,143 shared peaks,

6,324 cone-specific peaks and 693 rod-specific peaks (Supplemental Table S2.5). To focus on elements likely to contain cell-type-specific enhancers, we removed peaks overlapping promoters, peaks shared with non-photoreceptor cell types, and peaks with >70% overlap with repeat sequences. Filtering reduced the totals to 17,485 photoreceptor-specific peaks, 3,606 cone-specific peaks, and 394 rod-specific peaks (Fig. 2.7A). Thus, we found that the vast majority of photoreceptor-specific peaks were shared between rods and cones, and that there were approximately 10-fold more cone-specific peaks than rod-specific peaks.

As expected, peaks present in non-photoreceptor cell types and tissue showed little enrichment for motifs corresponding to photoreceptor TFs, while shared (rod and cone) photoreceptor-specific peaks were enriched for the factors discussed above (Fig. 2.7B). Interestingly, rod- and cone-specific peaks were enriched for distinct sets of TF motifs. Specifically, cone-specific open chromatin was enriched for Q50 (RAX), bHLH (NEUROD1), paired NR (THRB and/or RXRG), and a bZIP motif (whose cognate TF is unknown). In contrast, rod-specific open chromatin was specifically enriched for a MAF motif (NRL) and a distinct NR motif (RORB, ESRRB, and/or NR2E3), i.e., motifs bound by rod-specific TFs. Thus, comparative analysis of cell-type-specific open chromatin identified specific sequence features that may play important roles in determining rod- vs. cone-specific gene regulation.

2.7 Discussion

In this study, we presented genome-wide maps of open chromatin and gene expression in individual photoreceptor subtypes. These data revealed a striking depletion of accessible chromatin in rods relative to cones and other cell types, which we hypothesize is related to the unique nuclear organization of rods. Furthermore, we leveraged cell-type-specific open chromatin maps to identify sequence features that define shared photoreceptor-specific, rod-specific, and cone-specific regulatory elements.

The inverted nuclear architecture of the rods of nocturnal mammals is thought to enhance visual sensitivity in dim light environments [100]. Computational simulations predict that this inverted organization of rod nuclei produces less light scattering than the conventionally organized nuclei of other cell types. Thus, the inverted nuclear architecture confers desirable optical properties to photoreceptor nuclei. In addition, our data suggest that inverted nuclear organization has important implications for the regulatory landscape of mouse rods. Specifically, inverted architecture is correlated with the closure of thousands of regions of open chromatin, which are often clustered within large (10^5-10^6 bp) domains and biased towards gene-sparse regions of the genome with lower transcriptional activity in photoreceptors. Thus, while inverted nuclear organization may reflect selection for certain optical properties, precisely how the nucleus is re-packaged may be additionally constrained by the gene expression programs essential for photoreceptor function.

An important outstanding question with respect to inverted nuclear architecture is how it is regulated. As described above, previous work has shown that rod nuclear organization depends on silencing the nuclear envelope proteins *Lbr* and *Lmna [106]*. At the protein level, it was shown that LBR is present in both rods and cones in early postnatal life, and downregulated in both cell types as they mature. Furthermore, as LBR is downregulated, Lamin A/C protein is selectively upregulated in cones (and other retinal cell types) but not rods. In the current study, we find low levels of *Lmna* transcript present in adult cones but not rods, suggesting that rods selectively inhibit *Lmna* transcription. In addition, we find two open chromatin peaks (candidate regulatory elements) at the *Lmna* locus that are selectively closed in rods. We note that these peaks are open in both endogenous cones and $Nrl^{-/-}$ photoreceptors, indicating that *Nrl* mediates selective closure of these elements, though additional experiments are needed to determine if this is a direct or indirect effect. Nevertheless, these observations constitute a first step towards localizing changes in the expression

of key nuclear envelope genes within rod-specific gene regulatory networks.

We previously showed that acute knockout of *Nrl* results in a partial conversion of rods into cones and can delay photoreceptor degeneration in a mouse model of retinitis pigmentosa [131]. The present study suggests that widespread chromatin closure in adult rods may represent an epigenetic barrier to complete rod-to-cone reprogramming after acute *Nrl* knockout. It will be interesting to determine whether the dynamics of chromatin closure in developing rods temporally correlates with the acquisition of resistance to reprogramming observed in our prior study. Furthermore, if rod-specific chromatin closure is indeed related to the cell's inverted nuclear architecture, adult human rods, which have a conventional chromatin architecture, may prove more responsive to direct reprogramming strategies than mouse rods.

Recently, Mo *et al.* also evaluated the cell-type-specific epigenomic profile of mouse rods [132]. In particular, Mo *et al.* identified differential DNA methylation in rods and cones, which revealed a striking rod-specific enrichment of hypomethylated DNA in closed chromatin. Mo *et al.* suggest that many of these regions represent 'vestigial enhancers'—regulatory elements active earlier in development that retain hypomethylation marks in adult cell types despite loss of activity [133]. This is consistent with our observations that these regions appear to be less transcriptionally active in both adult rods and cones, and that cone-specific open chromatin is more strongly enriched for GO terms related to neural development than photoreceptor physiology. Finally, while we show that *Nrl* is required for chromatin closure in rods, Mo *et al*. also profiled *Nr2e3*$^{-/-}$ retinas, which demonstrate an intermediate epigenetic phenotype between that of *Nrl*$^{-/-}$ and WT retinas. This finding suggests that rod-specific nuclear organization depends at least in part on TFs downstream of *Nrl.*

In addition to defining the genome-wide chromatin accessibility profile of individual

photoreceptor subtypes, our analysis suggests a simple taxonomy of photoreceptor regulatory elements based on TF motif content and their proximity to genes. Consistent with previous work, we found that promoters and enhancers are distinct with respect to size, sequence content and specificity. In particular, promoters tend to be shared across cell types and are enriched for binding sites corresponding to ubiquitous transcriptional regulators, whereas enhancers exhibit greater cell-type-specificity and are enriched for binding sites corresponding to photoreceptor TFs. Furthermore, enhancers are highly enriched for CTCF binding sites, most of which are shared with non-photoreceptor cell types (likely reflecting ubiquitous TADs), though some are photoreceptor-specific (potentially mediating cell-type-specific contact domains). Enhancers containing CTCF motifs were largely devoid of motifs for photoreceptor TFs, whereas enhancers without CTCF motifs are highly enriched for the latter. Taken together, these findings suggest that CREs can be divided into three classes: promoters, CTCF-bound enhancers, and non-CTCF-bound enhancers. We suggest that this tripartite classification is applicable to other cell types.

With respect to non-CTCF-bound enhancers, our motif enrichment analysis identified a set of candidate TF binding sites that extend the 'vocabulary' of photoreceptor regulatory elements beyond K50 HD (CRX) binding sites, suggesting key roles for Q50 HD, bHLH, NR, MADS, MAF and bZIP TF family members. In parallel, our expression analysis identified TFs that were robustly expressed in rods and/or cones that likely bind these motifs. With respect to K50 HD TFs, we noted that *Six6* is expressed at markedly higher levels in adult cones than in rods. While *Six6* is essential for retinal development in vertebrates [134-136], a photoreceptor-specific role for *Six6* has not been described in mammals. Nevertheless, *Six6* has been shown to mediate photoreceptor differentiation in medaka [137], and a closely related homolog, *Six7*, has recently been shown to regulate cone-specific gene expression and survival in zebrafish [138, 139]. Additional work is needed to determine if *Six6* plays an analogous role in mouse cones.

In addition to identifying binding sites for TFs that likely cooperate with CRX, our analysis suggests that CRX binding sites themselves harbor considerable complexity, including both monomeric and dimeric configurations with distinct nucleotide preferences. That key photoreceptor enhancers contain multiple K50 HD sites capable of binding CRX dimers has been shown previously [69, 140]. Here, we show that this paired configuration is widespread among photoreceptor regulatory elements. This is particularly intriguing in light of recent work showing that individual nucleotides within paired K50 and Q50 homeodomain motifs (employing the same spacing and orientation preferences found in mouse photoreceptors) dictate photoreceptor subtype-specific gene expression in *Drosophila* [141]. Comprehensive functional analysis will be required to determine if mammalian photoreceptors exploit a similar *cis*-regulatory logic to control photoreceptor subtype-specific expression. Such analyses in primary photoreceptors (both in culture and *in vivo*) are now possible given recent innovations in sequencing-based multiplex reporter assays [41-43, 45].

We note that while nearly all motifs enriched in photoreceptor enhancers corresponded to known photoreceptor TFs, we were unable to unambiguously assign the cone-specific enrichment of non-MAF bZIP motifs (TGANTCA) to a candidate regulator. Nevertheless, we suggest that NRL might mediate selective closure of these elements in rods, given that bZIP TFs, including NRL, bind DNA as flexible homotypic or heterotypic dimers that tolerate a range of binding sites. NRL in particular has been shown to bind non-MAF bZIP sites as a heterodimer with either Fos or Jun [142]. This raises the intriguing possibility that NRL has opposing activities (activation vs. repression) at distinct classes of bZIP sites *in vivo*—either autonomously or in combination with non-cell-type-specific bZIP TFs.

Overall, the patterns of motif enrichment we describe yield comprehensive, single-nucleotide resolution maps of regulatory sequence features that can inform functional studies of

photoreceptor-specific enhancer activity. These data suggest that—even when comparing highly related cell types—cell-type-specific regulatory activity is encoded by complex interactions among shared and cell-type-specific TFs acting cooperatively and/or competitively to bind target sequences. Thus, a complete understanding of cell-type-specific regulatory logic will likely require an integrated analysis of both *cis*- and *trans*-acting factors.

2.8 Materials and methods

Mice. Mouse husbandry and all procedures (including euthanasia by CO_2 inhalation and cervical dislocation) were conducted in accordance with the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health, and were approved by the Washington University in St. Louis Institutional Animal Care and Use Committee.

Histology. Retinas were harvested from adult (8-week-old) mice, fixed in 4% paraformaldehyde in phosphate buffered saline (PBS) overnight at 4˚C, equilibrated in 30% sucrose in PBS, embedded in Tissue-Tek O.C.T (Sakura), and stored at -80˚C until sectioning. 14 µm cryosections were prepared at -20˚C. Imaging was performed with a Zeiss LSM 700 confocal microscope.

Retinal dissociation and FACS. Retinas were harvested from adult (8-week-old) male mice unless otherwise noted (Supplemental Table S2.1) and dissociated with trypsin as described previously [143]. Dissociated cells were resuspended in sorting buffer (1% fetal bovine serum [FBS], 0.1 mM ethylenediaminetetraacetic acid [EDTA] in calcium- and magnesium-free Hank's buffered salt solution [HBSS]). Cells were sorted using a FACSAria II (BD Biosciences). Single, viable cells were isolated by gating on forward scatter, side scatter, and pulse width. GFP-negative retinas were included in each sort as negative controls. GFP-positive cells were collected in 10% FBS in PBS (for ATAC-seq) or Buffer RLT (RNeasy mini kit, Qiagen) (for RNA-seq). For green cones, we found that double-sorting was necessary to achieve optimal purity. For these sorts, cells

were first collected in sort buffer and then re-sorted and collected in 10% FBS in PBS.

ATAC-seq. ATAC-seq was performed as described previously [39]. Briefly, 10,000-50,000 sorted cells were pelleted by centrifugation (500 g for 5 min at 4°C), washed twice with PBS, and nuclei were harvested by cold NP-40 lysis. Nuclei were incubated with 2.5 µl of Tn5 transposase (Nextera, Illumina) in a 50 µl reaction volume at 37°C for 30 minutes. Tagmented DNA was then purified using a MinElute PCR Purification kit (Qiagen). Library fragments were amplified with Phusion High-Fidelity DNA Polymerase (NEB), with the total number of PCR cycles calibrated by parallel qPCR reactions. Libraries were purified using PureLink PCR Purification kits (Invitrogen). Library quality was assessed by gel electrophoresis, and final libraries were quantified using KAPA Library quantification kits (Kapa Biosystems). Libraries were pooled in equimolar ratios and run on an Illumina HiSeq 2500 (paired-end 50 bp reads) (Supplemental Table S2.1).

RNA-seq. Library preparation was performed using 5 ng of total RNA. RNA integrity was assessed using an Agilent Bioanalyzer. cDNA was prepared using the SMARTer Ultra Low RNA kit for Illumina Sequencing-HV (Clontech) per manufacturer's instructions. cDNA was fragmented using a Covaris E210 sonicator using duty cycle 10, intensity 5, cycles/burst 200, time 180 seconds. cDNA was blunt-ended, had an 'A' base added to the 3' ends, and then had Illumina sequencing adapters ligated to the ends. Ligated fragments were amplified for 12 cycles using primers incorporating unique index tags. Replicate libraries from both cell types were pooled in equimolar ratios and sequenced on an Illumina HiSeq 2500 (single-end 50 bp reads) (Supplemental Table S2.1).

ATAC-seq data processing. Paired-end ATAC-seq reads were aligned to the GRCm38/mm10 mouse genome assembly using Bowtie2 (v2.2.3) in end-to-end mode with a maximum fragment

size of 2000 [144]. Alignments were filtered to remove reads with mapping quality <30, discordant read pairs, reads aligning to the mitochondrial genome, and reads aligning to ENCODE blacklisted regions [4] using SAMtools (v1.3) [145]. PCR duplicates were removed using Picard (v1.121) (http://picard.sourceforge.net). Finally, we removed alignments with an insertion size greater than 100 bp to enrich for nucleosome-free reads (NFR).

For visualization in the UCSC genome browser, bedgraph files were generated from NFR alignments for pooled replicates using HOMER (v4.8), specifying a maximum file size of 500 Mb [123]. Reproducible peaks were called for each cell type using MACS2 (v2.1.0) (using a -100 bp shift and a 200 bp extension, and calling subpeaks) [146] with an irreproducible discovery rate (IDR) threshold of 0.01 [147]. 'ATAC-seq peaks' described in subsequent analysis refer to 200-bp elements centered on peak summits. To assess reproducibility between biological replicates, read counts were normalized using the median-of-ratios method [148], and we then plotted $\log_2$(normalized read counts+1) for each peak for each replicate (Supplemental Fig. S2.1) and calculated the Pearson correlation coefficient (PCC) between replicates. This workflow was used for ATAC-seq data generated in the current study, as well as previously generated ATAC-seq data presented in Fig. 2.3C: pre-B cells (GSE63302) [149], activated B cells (GSE71698) [150], and purified neurons (GSE63137) [151].

RNA-seq data processing. Single-end RNA-seq reads were aligned to GRCm38/mm10 STAR (v2.4.2a), using an index prepared for 50 bp reads and the RefSeq gene model [152]. Read counts per gene were calculated using HTSeq [153]. To assess reproducibility between biological replicates, read counts were normalized using the median-of-ratios method [148], and we then plotted $\log_2$(normalized read counts+1) for each gene for each replicate (Supplemental Fig. S2.1) and calculated the Pearson correlation coefficient (PCC) between each pair of replicates. For visualization in the UCSC genome browser, bedgraph files were generated for pooled replicates

using HOMER (v4.8), specifying a maximum file size of 500 Mb [123].

ENCODE DNase-seq data processing. FASTQ files from DNase-seq datasets generated by ENCODE were downloaded from the ENCODE data portal (https://www.encodeproject.org/) and processed identically to ATAC-seq data except that single-end reads were used when paired-end reads were not available and alignments were not filtered for NFR reads (insert size <100 bp). Individual datasets and accessions are listed in Supplemental Table S2.4.

ChIP-seq data processing. FASTQ files from previously generated ChIP-seq datasets were downloaded from GEO and processed identically to ATAC-seq data except that single-end reads were used when paired-end reads were not available, alignments were not filtered for NFR reads (insert size <100 bp), and peaks were called on ChIP-seq replicates (vs. input control) with MACS2 (v2.1.0) using default parameters and an FDR of 0.01. These data are presented in Fig. 2.2A-D and Supplemental Table S2.3: CRX ChIP-seq (GSE20012) [68] and NRL ChIP-seq (https://datashare.nei.nih.gov/nnrlMain.jsp) [119]. Two biological replicates were pooled for both WT and *Nrl*$^{-/-}$ CRX ChIP-seq. A single biological replicate of NRL ChIP-seq was run on an Illumina platform (Genome Analyzer). Only this replicate was included in our analysis of NRL ChIP-seq data.

Quantification of data over photoreceptor ATAC-seq peaks. ATAC-seq, DNase-seq, and ChIP-seq read depths as well as phylogenetic conservation (phastCons 60-way vertebrate conservation) [154], and GC content were quantified over peak sets using HOMER (v4.8) [123]. For per-feature count tables (used to generate the heatmap in Fig. 2.2D), photoreceptor ATAC-seq and adult whole-retina DNase-seq peaks were combined using BEDOPS (v2.4.14) [155], and the indicated datasets were scored over these intervals in 5-bp bins distributed over a 3 kb window centered on peak summits. For average signal histograms (Supplemental Fig. S2.6), data were

scored separately for rod, green cone, and *Nrl*$^{-/-}$ photoreceptor ATAC-seq peaks in 5-bp bins distributed over a 2 kb window centered on the nearest TSS (promoter peaks) or peak summits (enhancers).

Quantification of sample relatedness. For PCA of brain DNase-seq and photoreceptor ATAC-seq, peaks from each cell type were merged into a master list with BEDOPS (v2.4.14) [155]. For each replicate, reads were counted within elements of the master list using bedtools (v2.24.0) [156]. PCA was then performed on regularized logarithm-transformed values [130]. For pairwise analysis between replicates of photoreceptor ATAC-seq and whole retina, brain, lung, liver and B cell DNase-seq, peaks were again combined using BEDOPS (v2.4.14) [155]. Elements in the union were scored for coverage in each replicate using bedtools (v2.24.0) [156]. We then calculated the pairwise Spearman correlation coefficient (ρ) matrix, clustering rows and columns by 1-ρ using average linkage.

Comparative analysis of global chromatin closure. To quantify the global distribution of chromatin accessibility, we partitioned the genome into 50 kb fixed windows and calculated read coverage over these windows for each cell type and tissue using bedtools (v2.24.0) [156]. For each sample, we normalized counts to total reads, and plotted the empirical cumulative distribution (proportion of 50 kb windows with less than or equal to each observed coverage value).

Identification of differentially accessible peaks. ATAC-seq peaks from each photoreceptor type were merged into a master list with BEDOPS (v2.4.14) [155]. For each ATAC-seq replicate, reads were counted within elements of the master list using bedtools (v2.24.0) [156], and differentially accessible peaks were identified with DEseq2 testing for a $\log_2$(fold change) greater than 1 at an FDR of 0.1 [130]. Counts from rods, double-sorted green cones, and *Nrl*$^{-/-}$ photoreceptors were used for differential accessibility analysis, with cone subtypes collapsed to a single level. The

results of this analysis were used to partition photoreceptor ATAC-seq peaks into three subsets: "rod-specific" (significantly more open in rods), "cone-specific" (significantly more open in cones), and "shared" (differential accessibility not statistically significant).

Identification of differentially expressed genes. We used DEseq2 to test for differential expression between rods and cones using per gene read counts for rod and *Nrl*$^{-/-}$ photoreceptor RNA-seq data [130]. Statistical testing was performed using a $\log_2$(fold change) threshold of 1 and an FDR of 0.05.

Integrated analysis of chromatin accessibility and gene expression. To assess the relationship between the accessibility of individual peaks and the expression of nearby genes, peaks were assigned to genes by nearest TSS. In this way, differential expression of individual genes or the distribution of expression of groups of genes could be putatively associated with ATAC-seq peaks.

GO enrichment analysis. Enrichment of GO Biological Process term within rod- or cone-specific ATAC seq peaks was performed using GREAT (v3.0.0) [157].

Motif enrichment analysis. Known motif enrichment and *de novo* motif discovery were performed for photoreceptor ATAC-seq peaks as well as brain, liver, lung and B cell DNase-seq peaks using HOMER (v4.8) [123]. Target sequences consisted of 200-bp elements centered on peak summits. Background sequences consisted of approximately 50,000 randomly selected 200-bp intervals from the mouse genome normalized for mono- and di-nucleotide content relative to each target set. Repeat sequences were masked from the genome, and targets with >70% of bases masked were dropped from enrichment analysis. Motif enrichment was performed separately for promoter peaks (peaks less than 1000 bp upstream and 100 bp downstream of an annotated TSS) and enhancer peaks. Sequence logos presented in Fig. 2.5 and Supplemental Fig. S2.7 were trimmed to remove flanking positions with low information content (<1 bit). Complete known and

de novo motif enrichment results are presented in Supplemental Tables S2.7-S2.10. To assess motif co-occurrence, we first collapsed the database of 319 known motifs curated by HOMER to a set of 66 non-redundant motifs (by aligning motifs, calculating the PCC across nucleotide frequencies, and selecting the highest scoring motif among pairs with a PCC > 0.6). For each motif, we then counted the number of peaks with ≥1 occurrence, and for each pair, we counted the number of peaks with ≥1 pair. The enrichment of co-occurrence was then calculated as the $\log_2$(observed pairs/expected pairs), where the number of expected pairs was estimated from the frequency of individual motifs: expected = (number of peaks with motif 1) x (number of peaks with motif 2) / (total peaks). The pairwise co-occurrence enrichment matrix was plotted as a heatmap, with rows and columns clustered by Euclidean distance and average linkage. Preferential spacing between highly enriched motifs (K50 HD, MAF, bZIP, MADS, and NR motifs) was assessed by first centering shared photoreceptor ATAC-seq peaks on individual motifs, and then plotting the density of secondary motifs (using a relaxed log odds threshold of 5) on either strand upstream and downstream of the primary motif.

Analysis of CRE-seq data. Previously generated CRE-seq data were downloaded from the supporting information included in White *et al.* [42]. Constructs were classified as having or not having k-mers of interest based on exact matches to TAAT (or ATTA), TAAG (or CTTA), TAAT…ATTA, TAAG…CTTA, or TAAT…CTTA (or TAAG…ATTA). We then examined the distribution of CRE-seq expression (a measure of the enhancer activity) based on the presence or absence of individual k-mers. In addition, CRE-seq constructs were determined to contain instances of motifs enriched in photoreceptor ATAC-seq peaks by scanning sequences for individual motifs (CRX, NEUROD1, MEF2D, and RORB) using HOMER (v4.8) [123]. We again examined the distribution of CRE-seq expression based on the presence or absence of individual motifs.

Statistics and data visualization. Statistical analyses were implemented in R (v3.3.0) [158] using base packages, as well as coin [159] and DEseq2 [130]. Data visualization was implemented with ggplot2 [160] and gplots [161].

Data access. ATAC-seq data (raw data, signal tracks, peak calls, and count tables) and RNA-seq data (raw data, signal tracks, and count tables) have been deposited in the NCBI Gene Expression Omnibus (GEO) under accession number GSE83312.

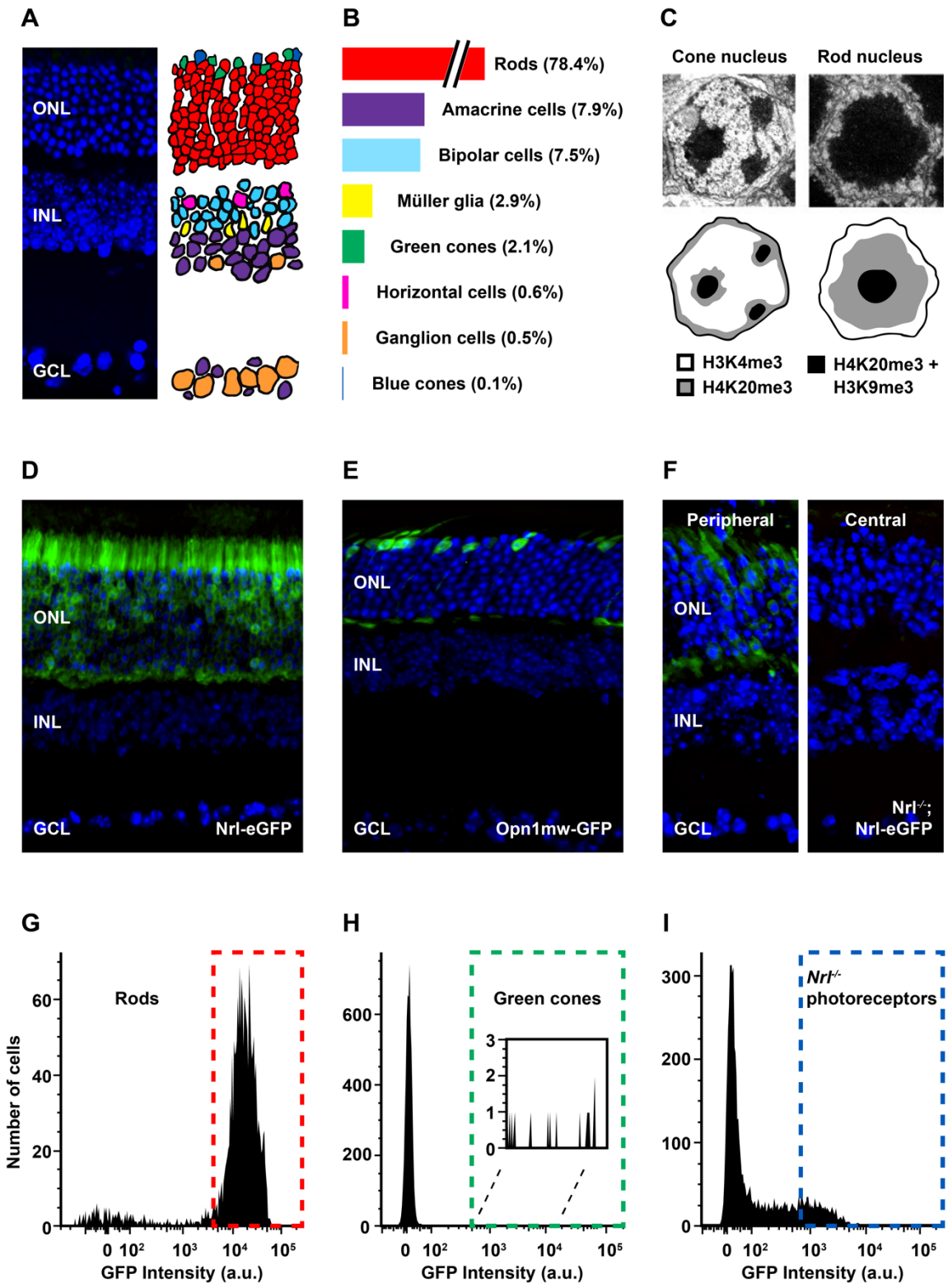


Figure 2.1. Epigenomic analysis of photoreceptor subtypes. (A) The mouse retina is composed of three layers: the outer nuclear layer, the inner nuclear layer, and the ganglion cell layer (ONL,

INL, and GCL, respectively). (B) There are seven major classes of cells in the retina: rod and cone photoreceptors, bipolar cells, horizontal cells, amacrine cells, ganglion cells, and Müller glia. (C) Compared to the nuclei of other cell types (e.g., cones) rod nuclei have an inverted architecture with inactive heterochromatin (H3K9me3, H4K20me3) localized to the center and active euchromatin (H3K4me3) localized to the periphery (image adapted from [68]). (D-F) Reporter lines used to purify individual photoreceptor types: *Nrl-eGFP* (rod), *Opn1mw-GFP* (green cones), *Nrl*$^{-/-}$*; Nrl-eGFP* (*Nrl*$^{-/-}$ photoreceptors, or blue cones). In the adult, *Nrl*$^{-/-}$*;Nrl-eGFP* retinas exhibit a peripheral (high) to central (low) gradient of GFP expression (F). (G-I) Representative FACS plots from individual sorts of dissociated retinal cells from each reporter line.

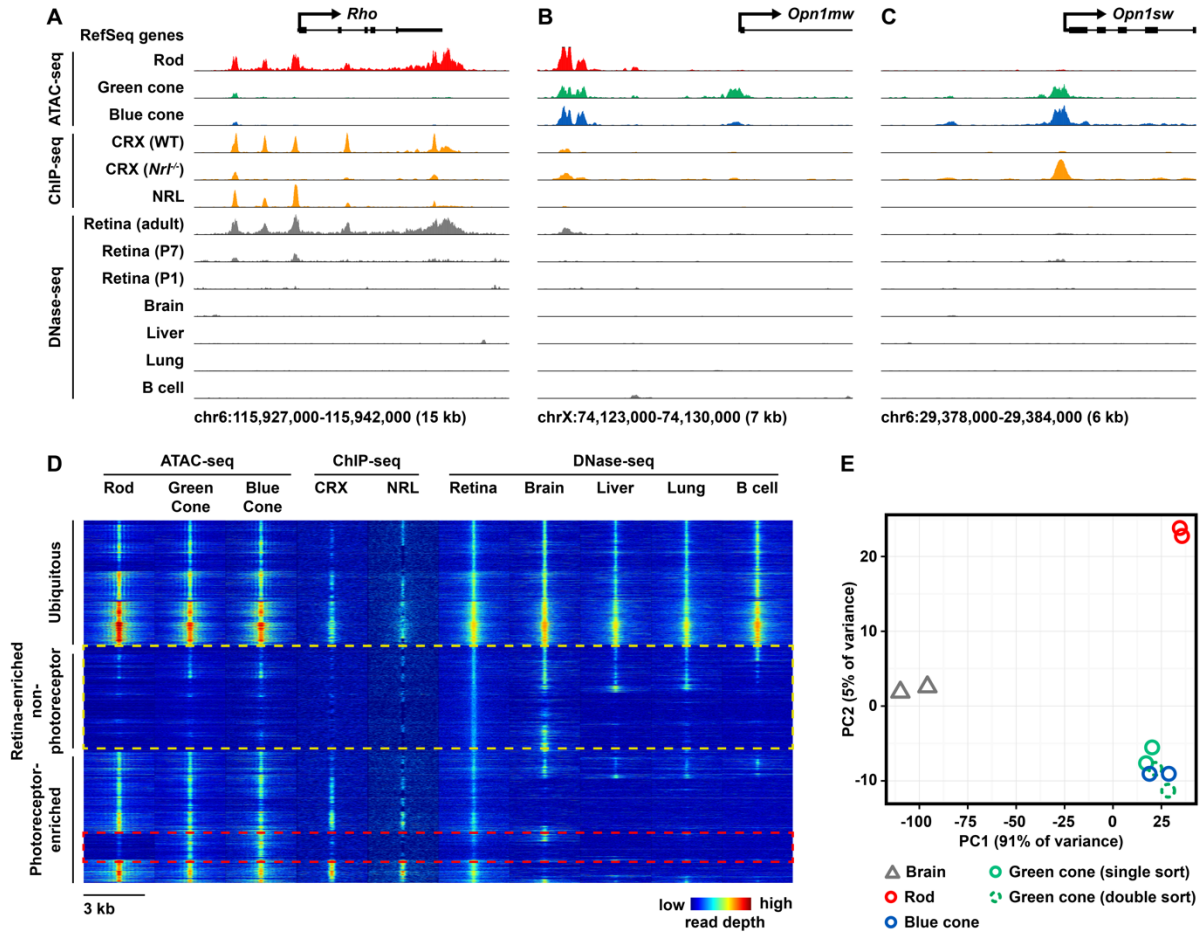


Figure 2.2. ATAC-seq of flow-sorted photoreceptors yields cell-type-specific maps of open chromatin. (A-C) ATAC-seq profile at rod- (A) and cone-specific (B-C) loci. Previously generated data sets also shown: CRX ChIP-seq (photoreceptor-specific TF) in WT (rod-dominant) and *Nrl*$^{-/-}$ (all-cone) retinas, NRL ChIP-seq (rod-specific TF) in WT retina, as well as DNase from whole retina (P1, P7, and adult), brain, liver, lung, and B cells derived from the ENCODE project. (D) Genome-wide profile of epigenomic datasets shown in (C) across photoreceptor ATAC-seq and adult whole-retina DNase-seq peaks. Rows show the read depth (denoted by pseudo-colored intensity) of the indicated epigenomic dataset in 3 kb windows centered on photoreceptor ATAC-seq and whole-retina DNase-seq peak summits (60,414 peaks randomly down-sampled to 10,000 for plotting). Rows are ordered by hierarchical clustering, revealing that approximately one-third of peaks are ubiquitously accessible (top), one third are non-photoreceptor peaks (middle, yellow

box), and one-third are photoreceptor-enriched peaks (bottom). Rod and cone profiles are highly similar, but a subset of cone-enriched peaks appear to be selectively closed in rods (red box). (E) Principal component analysis (PCA) of brain DNase-seq and photoreceptor ATAC-seq (two replicates per cell type) shows that the open chromatin profile of rods and cones are highly similar but distinct, whereas cone subtypes are indistinguishable.

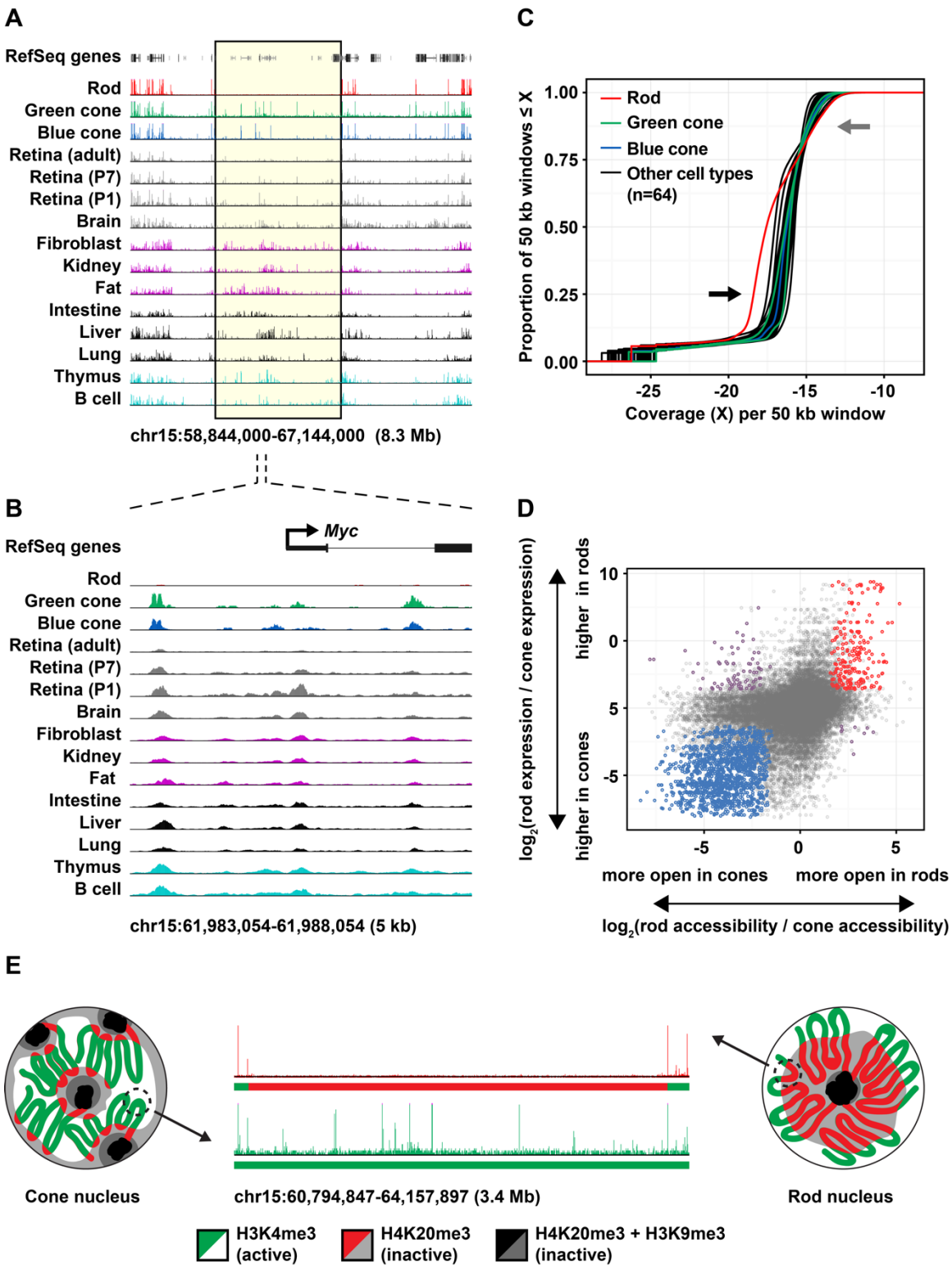


Figure 2.3. Mouse rods have a uniquely closed epigenomic landscape. (A) Representative genomic interval showing an extended run of open chromatin elements selectively closed in rods

(yellow box). Tracks show ATAC-seq profiles for rods, green cones, and blue cones and DNase-seq profiles for ten additional tissues (including three developmental time points for whole retina). (B) Representative window from (A) at higher resolution illustrating rod-specific closure of individual peaks. (C) Empirical cumulative distribution functions for genome-wide chromatin accessibility (normalized ATAC-seq or DNase-seq reads in fixed 50 kb windows, see Methods). Rods have a uniquely closed epigenomic landscape relative to blue and green cones as well as 64 additional mouse tissues and cell types (black arrow). Regions that were open in rods tended to have especially high read counts (gray arrow). (D) Change in accessibility vs. change in the gene expression in rods and blue cones. For each photoreceptor ATAC-seq peak (n=55,161), the log of the ratio of normalized ATAC-seq reads (rods/blue cones) is plotted on the x-axis, and the log of the ratio of normalized RNA-seq reads corresponding to the nearest gene (rods/blue cones) is plotted on the y-axis. Peaks that are both significantly differentially accessible (FDR < 0.1) and significantly differentially expressed (FDR < 0.1) are colored red (more open in rods, higher expression in rods), blue (more open in cones, higher expression in cones), or purple (more open in rods, higher expression in cones, or more open in cones, higher expression in rods). Changes in accessibility and expression are directionally correlated, but many peaks near differentially expressed genes are not differentially accessible, and many differentially accessible peaks are not correlated with differential gene expression (especially in cones, left half of plot). (E) Schematic proposing how global differences in nuclear organization in rods and cones are correlated with local differences in chromatin accessibility (figure design adapted from [100]).

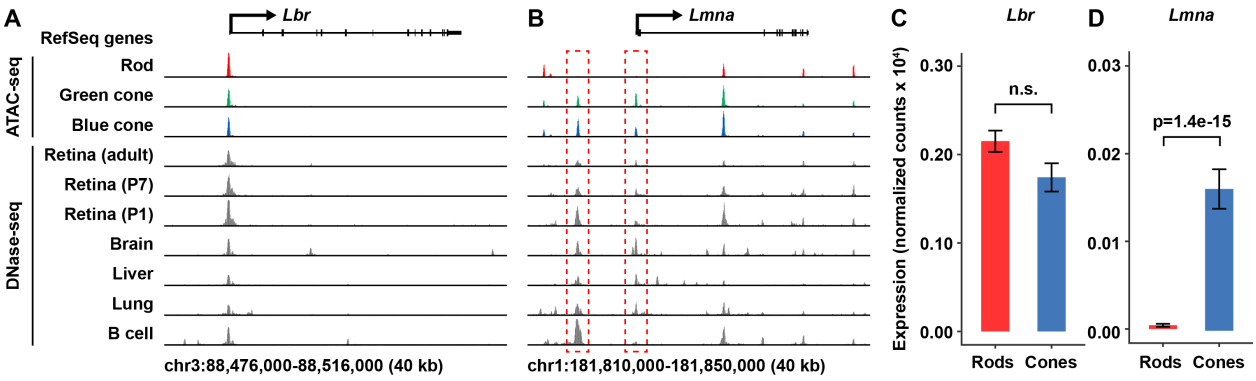


Figure 2.4. Rod-specific chromatin closure and reduced expression of *Lmna* but not *Lbr*. (A) ATAC-seq profiles from rods, green cones, blue cones, and DNase-seq profiles from whole retina (P1, P7, and adult), brain, liver, lung, and B cells at the *Lbr* locus. (B) Same datasets as in (A) shown for the *Lmna* locus. One open chromatin peak overlapping the promoter and one peak ~6.5 kb upstream are selectively closed in rods (red boxes). (C) *Lbr* expression in rods and blue cones as measured by RNA-seq. (D) *Lmna* expression in rods and blue cones as measured by RNA-seq. For (C) and (D), bar height corresponds to expression mean (normalized read counts x 10^4); error bars correspond to ±1 standard deviation. Note that the data in (C) and (D) are plotted on different scales.

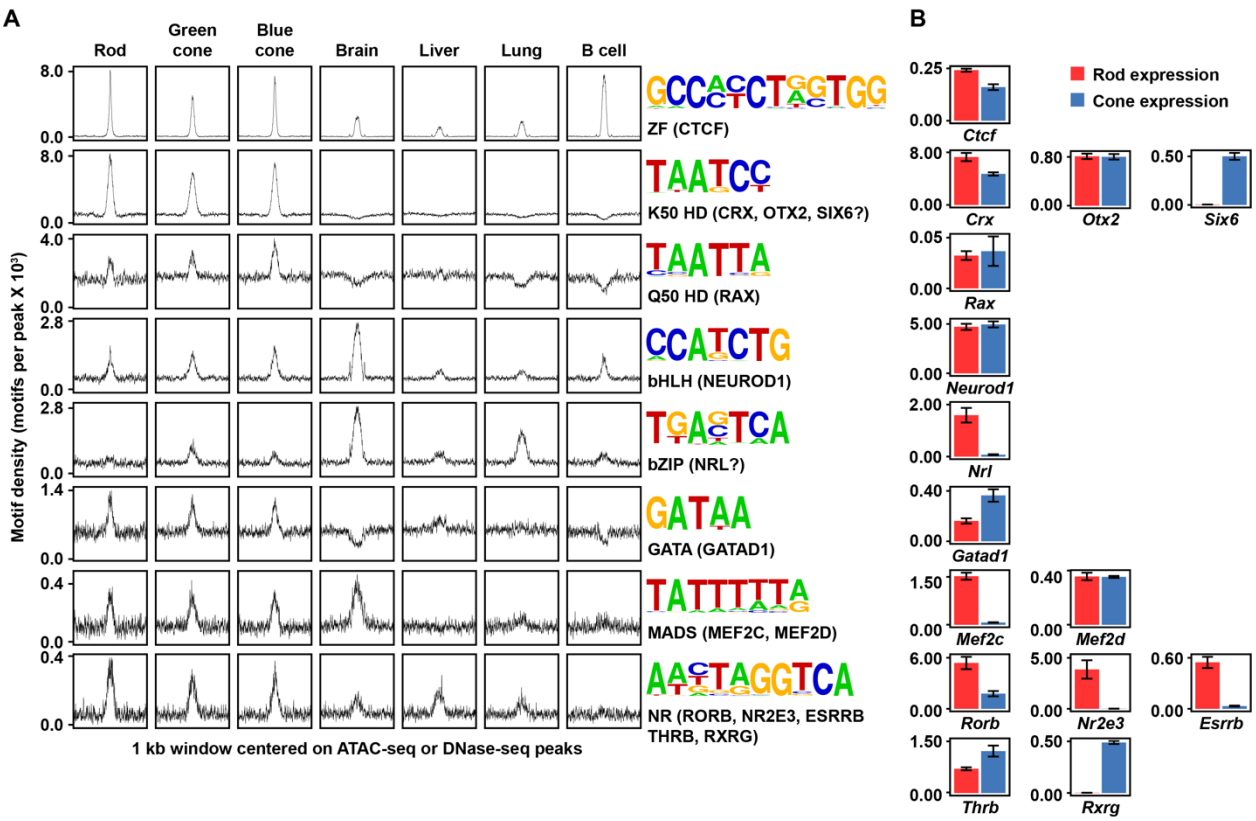


Figure 2.5. TF binding site motif enrichment in photoreceptor enhancers (A) Enrichment of known TF binding site motifs in enhancer (TSS-distal) peaks in rods, green cones, blue cones, brain, liver, lung, and B cells. For each panel, distance from peak summit (-500 bp to 500 bp) is plotted on the x-axis and motif density (motifs per peak at each position) is plotted on the y-axis, illustrating central enrichments of the motifs presented. Motifs are labeled by TF class, and candidate photoreceptor TFs that may bind each motif are indicated in parentheses. (B) Expression of candidate TFs for each motif in (A) in rods and cones as measured by RNA-seq. Bar height corresponds to expression mean (normalized read counts x 10^4); error bars correspond to ±1 standard deviation.

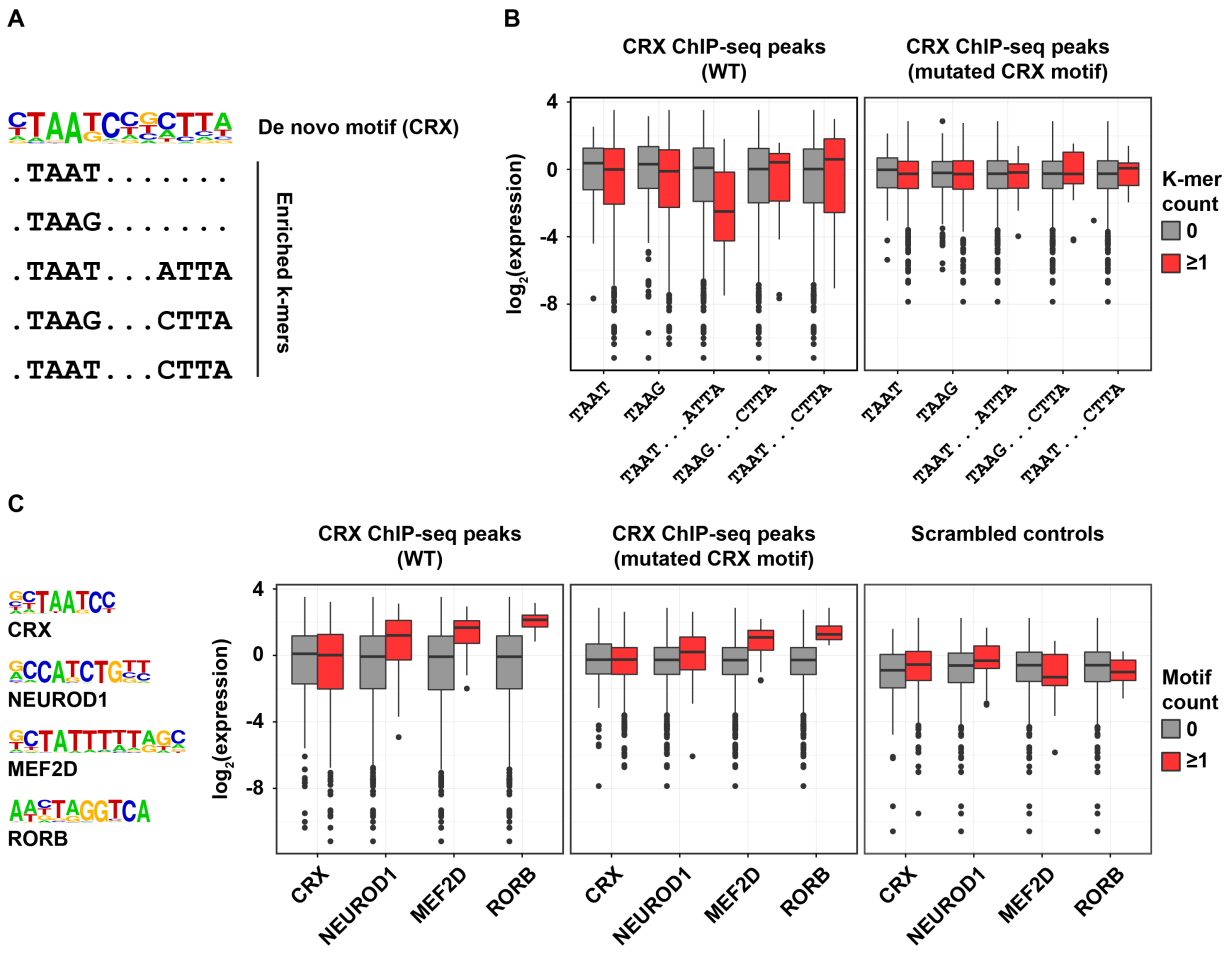


Figure 2.6. Functional effects of enriched motifs on photoreceptor enhancer activity. (A) Dimeric K50 HD motif identified in photoreceptor enhancers (TSS-distal ATAC-seq peaks). The motif logo shows nucleotide preferences scaled to observed frequencies; highly enriched k-mers that compose this motif are listed beneath. (B) Expression of CRE-seq constructs from White *et al.* harboring distinct TAAT and TAAG monomeric and dimeric motifs (gray: indicated k-mer not in tested sequence, red: ≥1 instance of indicated k-mer in tested sequence) [42]. The left panel shows the expression of native constructs (84-bp sequences centered on CRX ChIP-seq peaks), and the right plot shows the expression of constructs with CRX motifs eliminated by point mutation (CTAATCC to CTACTCC). (C) the expression of constructs assayed by White *et al.* stratified by the presence or absence of four motifs found to be highly enriched in photoreceptor

open chromatin: CRX, NEUROD1, MEF2D, and RORB. Box plots show the distribution of expression for constructs with or without the indicated motif in endogenous elements (left panel), constructs with mutated CRX sites (center panel), and scrambled controls (right panel).

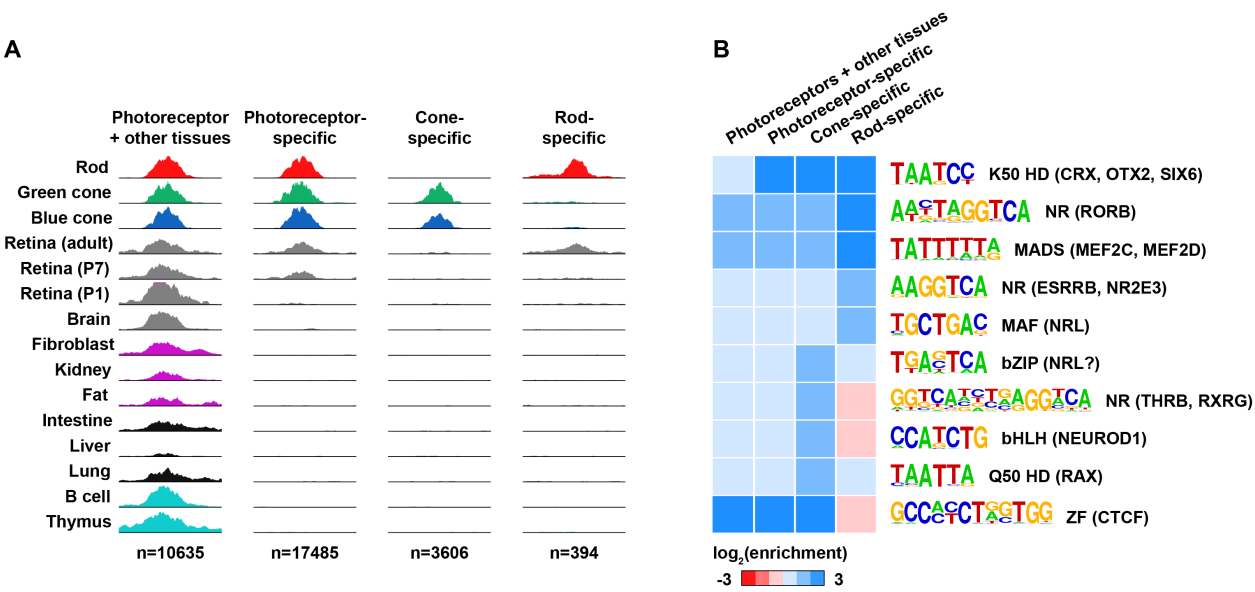


Figure 2.7. Rods and cones show distinct patterns of TF binding site enrichment. (A) Examples of peak sets used for motif enrichment analysis: (1) peaks open in photoreceptors and other cell types; (2) photoreceptor-specific peaks (open in rods and cones but not other cell types); (3) cone-specific (open in cones but not rods or other cell types); and (4) rod-specific (open in rods but not cones or other cell types). (B) Motif enrichment (ratio of motifs in target vs. background sequences) for selected motifs in the indicated peak sets. Motif logos are labeled by TF family, and cognate photoreceptor TFs are listed in parentheses.

2.9 Supplemental materials

2.9.1 Supplemental tables

Supplemental Table S2.1. Technical covariates and sequencing metrics for photoreceptor ATAC-seq and RNA-seq. RIN: RNA integrity number. Raw sequencing reads: number of forward and reverse reads for each sample. Processed sequencing reads: number of forward and reverse reads after filtering out improperly paired reads, reads with mapping quality <30, reads aligning to the mitochondrial genome, reads aligning to unplaced or unlocalized contigs, ENCODE blacklist regions, and PCR duplicates. URL:

https://images.nature.com/original/nature-assets/srep/2017/170228/srep43184/extref/srep43184-s2.xls

Supplemental Table S2.2. Annotated ATAC-seq peaks. Consensus peak calls for each cell type are presented on separate worksheets. URL:

https://images.nature.com/original/nature-assets/srep/2017/170228/srep43184/extref/srep43184-s3.xls

Supplemental Table S2.3. Overlap between photoreceptor ATAC-seq and additional whole-retina epigenomic datasets. URL:

https://images.nature.com/original/nature-assets/srep/2017/170228/srep43184/extref/srep43184-s4.xls

Supplemental Table S2.4. Datasets and accessions. URL:

https://images.nature.com/original/nature-assets/srep/2017/170228/srep43184/extref/srep43184-s5.xls

Supplemental Table S2.5. Differentially accessible peaks (rods vs. cones). Rod 1, rod 2, green cone 1, green cone 2, blue cone 1, blue cone 2: normalized ATAC-seq reads for each cell type. Base mean: average across all cell types. Log_2(fold change): fold change is calculated as rod ATAC-seq reads / cone ATAC-seq reads. Green cones and blue cones are collapsed into a single level for statistical analysis. URL:

https://images.nature.com/original/nature-assets/srep/2017/170228/srep43184/extref/srep43184-s6.xls

Supplemental Table S2.6. Differentially expressed genes (rods vs. blue cones). Rod 1, rod 2, rod 3, blue cone 1, blue cone 2, blue cone 3: normalized RNA-seq reads for each cell type. Base mean: average across all cell types. Log_2(fold change): fold change is calculated as rod RNA-seq reads / blue cone RNA-seq reads. URL:

https://images.nature.com/original/nature-assets/srep/2017/170228/srep43184/extref/srep43184-s7.xls

Supplemental Table S2.7. Known motif enrichment: photoreceptor promoters. A complete list of sequence logos and position weight matrices (PWMs) for individual motifs is available online in the HOMER database (http://homer.salk.edu/homer/motif/HomerMotifDB/homerResults.html). URL:

https://images.nature.com/original/nature-assets/srep/2017/170228/srep43184/extref/srep43184-s8.xls

Supplemental Table S2.8. Known motif enrichment: photoreceptor enhancers. A complete list of sequence logos and position weight matrices (PWMs) for individual motifs is available online in the HOMER database: http://homer.salk.edu/homer/motif/HomerMotifDB/homerResults.html.

URL:

https://images.nature.com/original/nature-assets/srep/2017/170228/srep43184/extref/srep43184-s9.xls

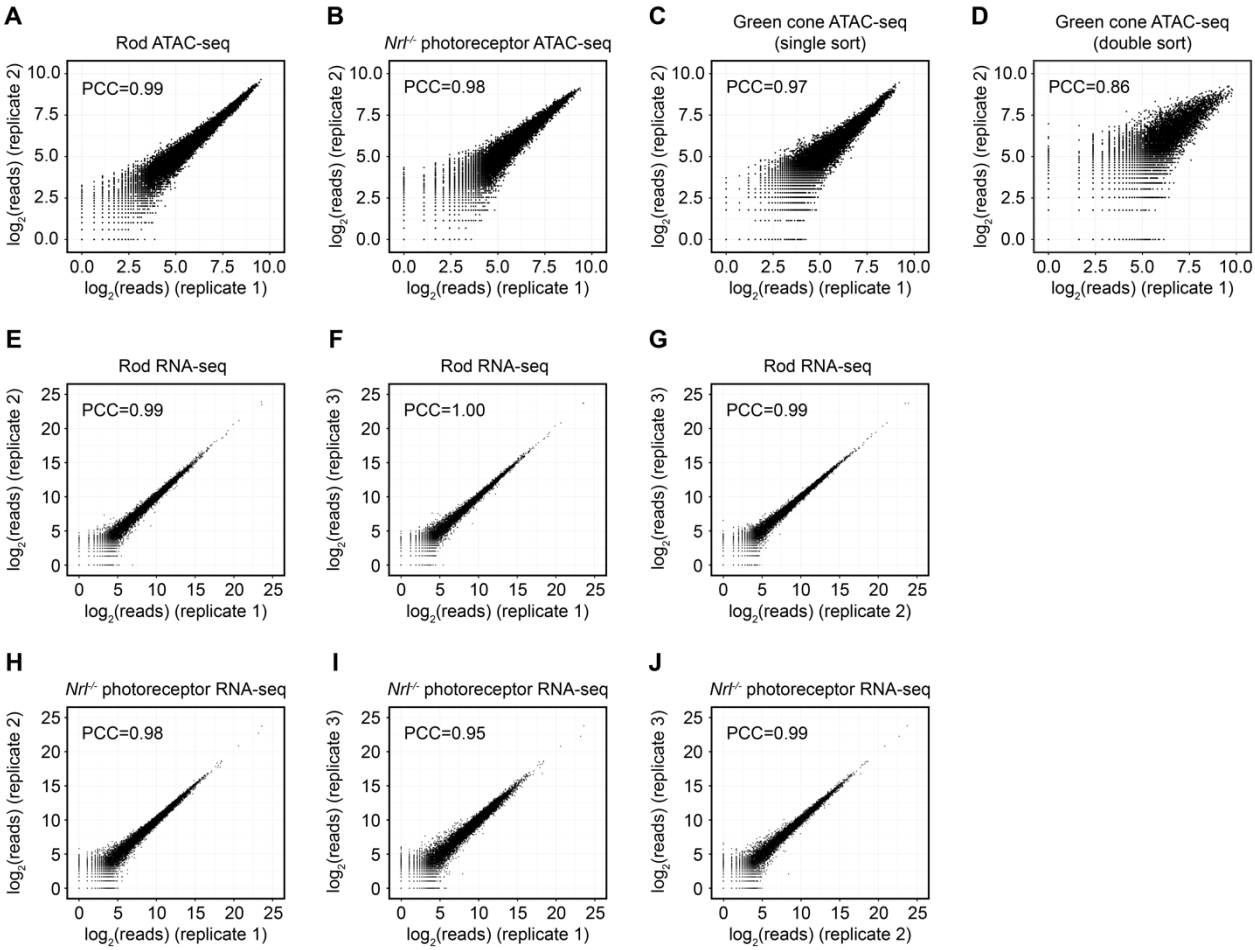
Supplemental Table S2.9. *De novo* motif enrichment: photoreceptor promoters. PWMs for *de novo motifs* are included as a separate worksheet. URL:

https://images.nature.com/original/nature-assets/srep/2017/170228/srep43184/extref/srep43184-s10.xls

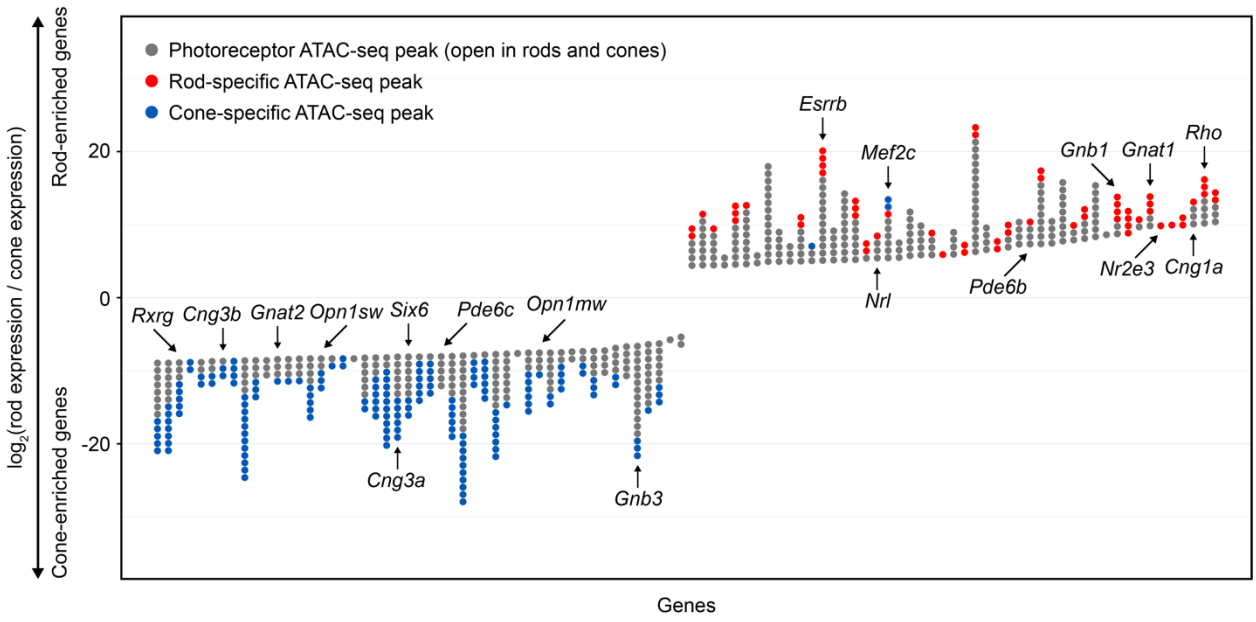
Supplemental Table S2.10. *De novo* motif enrichment: photoreceptor enhancers. PWMs for *de novo motifs* are included as a separate worksheet. URL:

https://images.nature.com/original/nature-assets/srep/2017/170228/srep43184/extref/srep43184-s11.xls

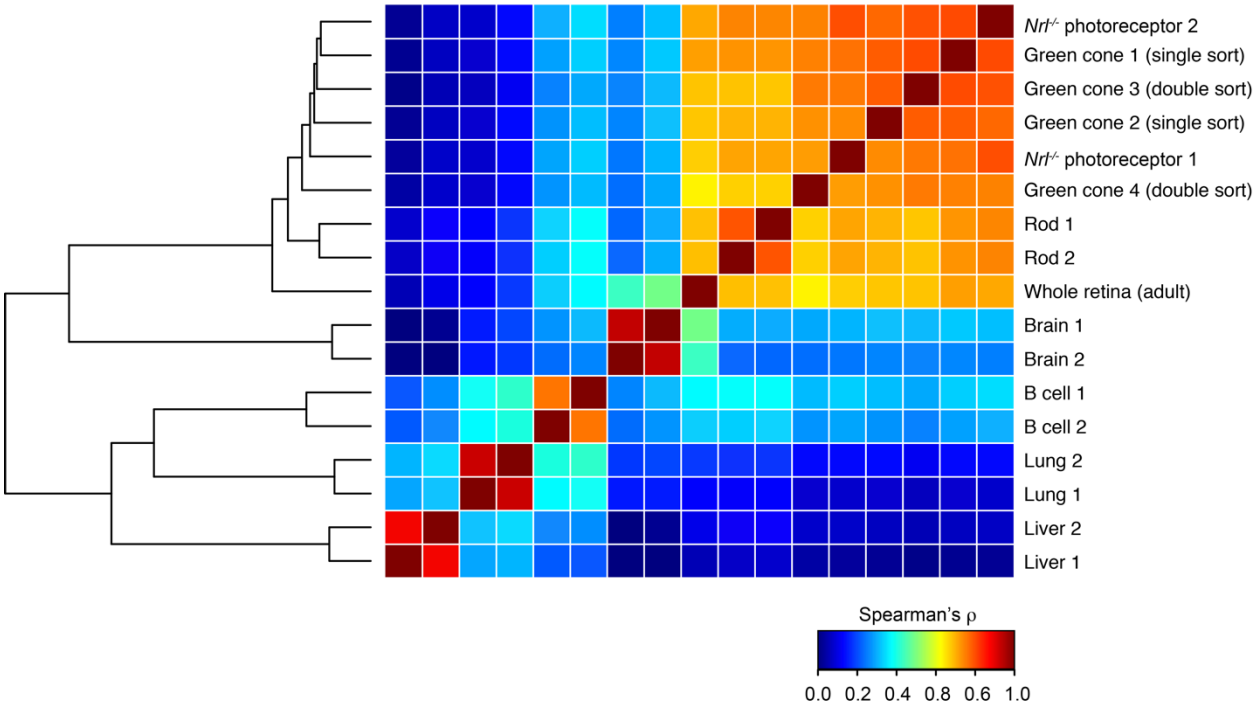
2.9.2 Supplemental figures



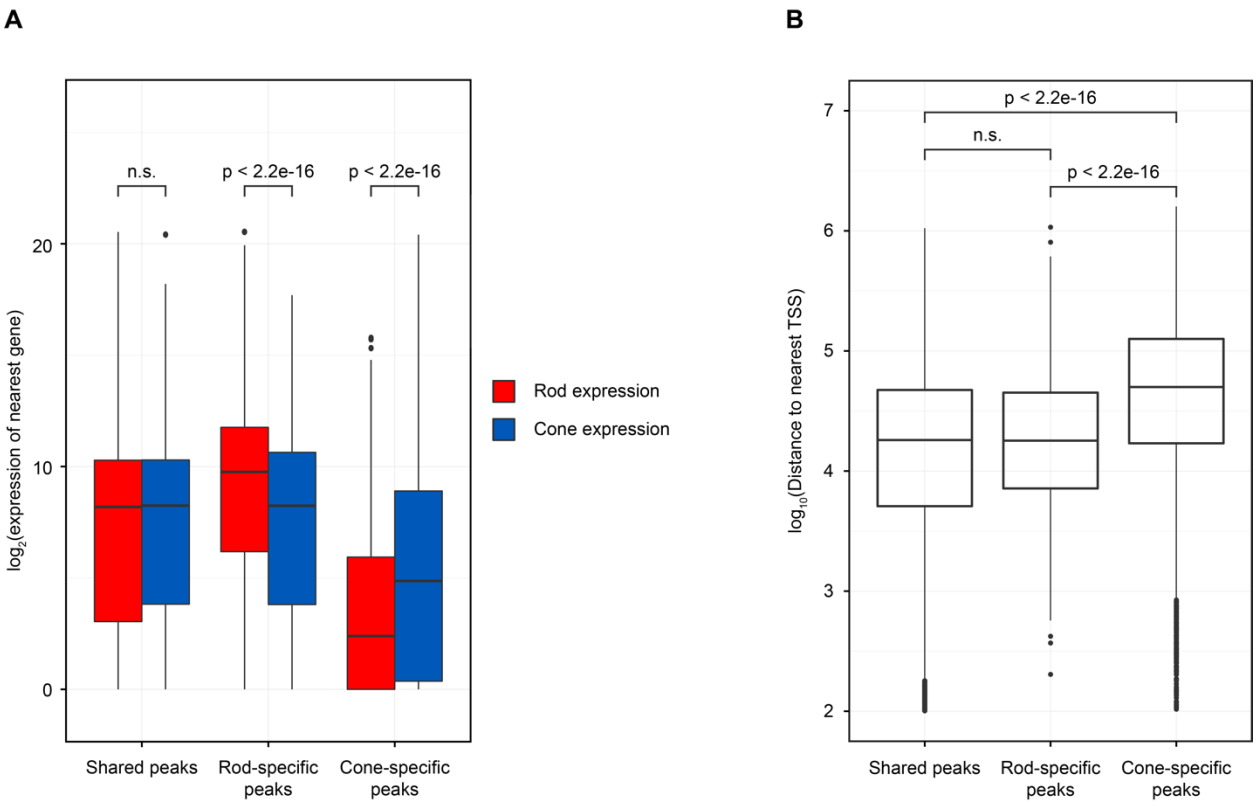
Supplemental Figure S2.1. Photoreceptor ATAC-seq and RNA-seq reproducibility. (A-D) Reproducibility between pairs of biological replicates for rod, *Nrl*$^{-/-}$ photoreceptor (blue cone), single-sorted green cone, and double-sorted green cone ATAC-seq. For each peak, the read count for biological replicate 1 (x-axis) is plotted against the read count for biological replicate 2 (y-axis). (E-J) Reproducibility between pairs of biological replicates for rod and *Nrl*$^{-/-}$ photoreceptor (blue cone) RNA-seq. For each gene, the read count for biological replicate 1 (x-axis) is plotted against the read count for biological replicate 2 (y-axis) (three pairs of replicates per cell type). PCC: Pearson correlation coefficient.



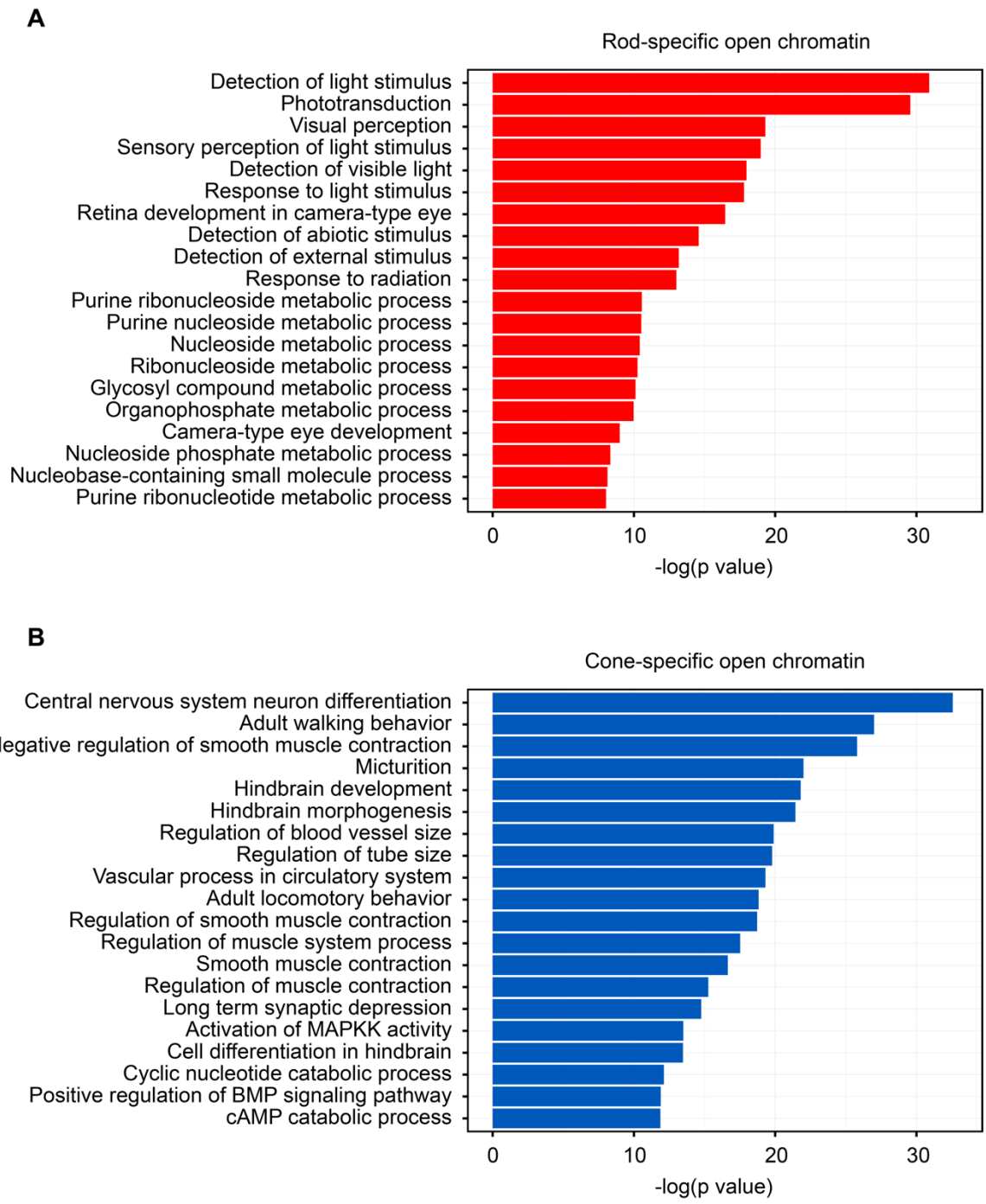
Supplemental Figure S2.2. Locus complexity of rod- and cone-specific genes. The 50 most differentially expressed genes between rods and blue cones (as assessed by RNA-seq) are ordered on the x-axis by fold-change (rod expression/cone expression). For each gene (column), each point represents an ATAC-seq peak that mapped to that gene (based on nearest TSS). Peaks shared by rods and cones are colored gray, rod-specific peaks are red, cone-specific peaks are blue. For rods, the bottom point for each gene is positioned on the y-axis to correspond to the estimated log of the fold change (rods/blue cones) in the expression of that gene; additional points (if present) are stacked on top. For cones, this value corresponds to the top point, and additional points (if present) are included beneath. Genes that are more highly expressed in rods are typically flanked by rod-specific peaks, whereas genes that are more highly expressed in cones are flanked by cone-specific peaks. Both rod- and cone-enriched genes are frequently flanked by shared ATAC-seq peaks. This style of data presentation was adapted from [121].



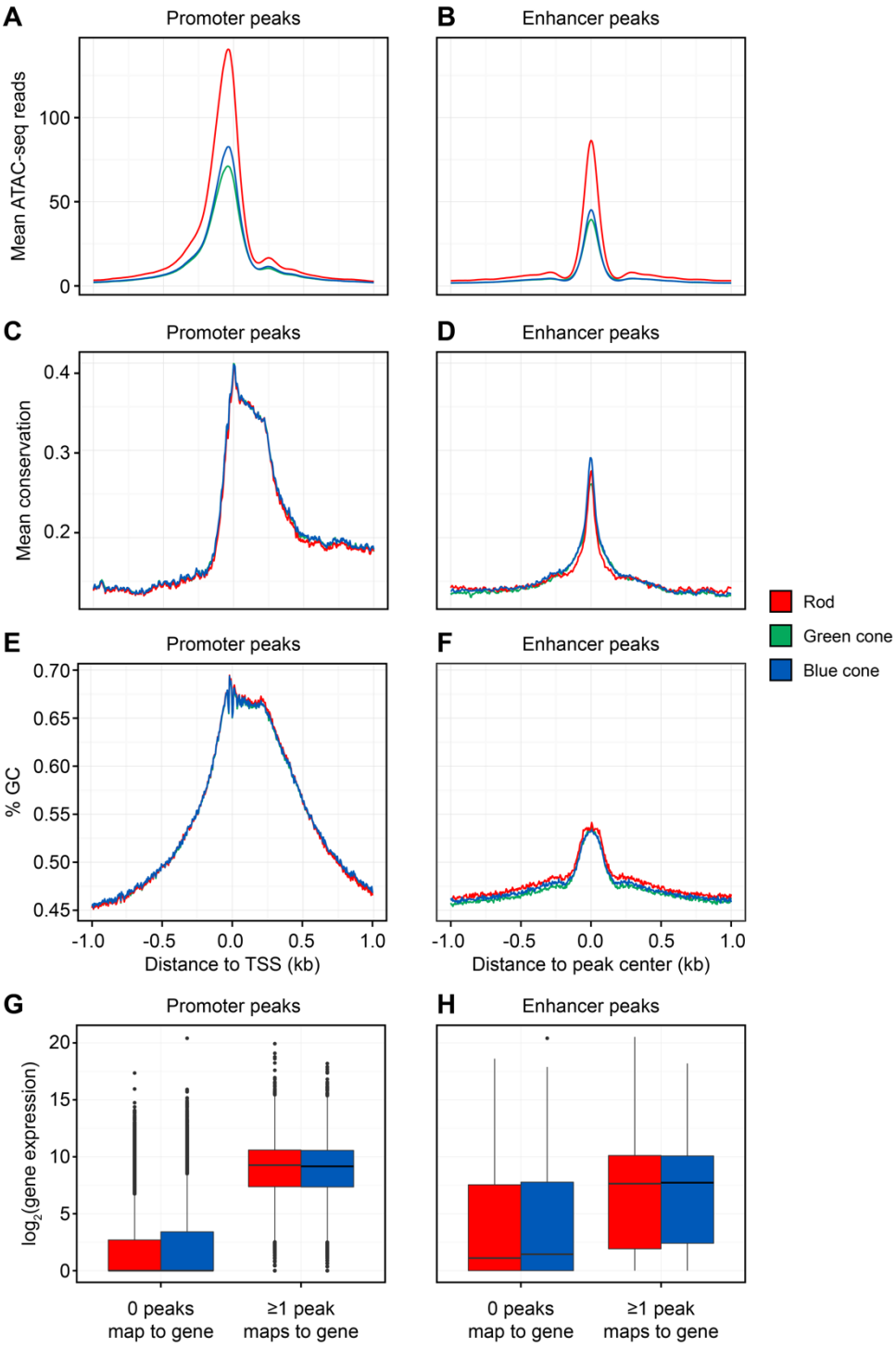
Supplemental Figure S2.3. Pairwise correlations between open chromatin datasets. Peaks from each cell or tissue type were merged into a common set of 173,219 regulatory elements. Reads in each feature were counted in each cell type, and samples were clustered across features using Spearman's ρ with average linkage. Control tissues cluster distinctly from photoreceptors. Whole retina falls between brain and photoreceptors. Rods cluster distinctly from cones, and cone subtypes cluster together.



Supplemental Figure S2.4. Gene expression near shared and cell-type-specific CREs. (A) Each photoreceptor enhancer ATAC-seq peak (n=35,173), was mapped to a gene based on nearest TSS. The distribution of expression for genes surrounding shared peaks was not significantly different in rods vs. cones. Gene expression was significantly higher in rods near rod-specific peaks (permutation test, 10,000 permutations). Gene expression was significantly higher in cones near cone-specific peaks (permutation test, 10,000 permutations). I.e., rod- and cone-specific ATAC-seq peaks are associated with elevated expression in rods and cones, respectively, but genes near cone-specific ATAC-seq peaks have lower expression in both cell types relative to genes near shared or rod-specific peaks. (B) Shared and rod-specific peaks were located approximately the same distance from genes, whereas cone-specific open chromatin was located significantly farther from genes (permutation test 10,000 permutations). In other words, regions selectively closed in rods are located farther from genes than typical open chromatin elements.

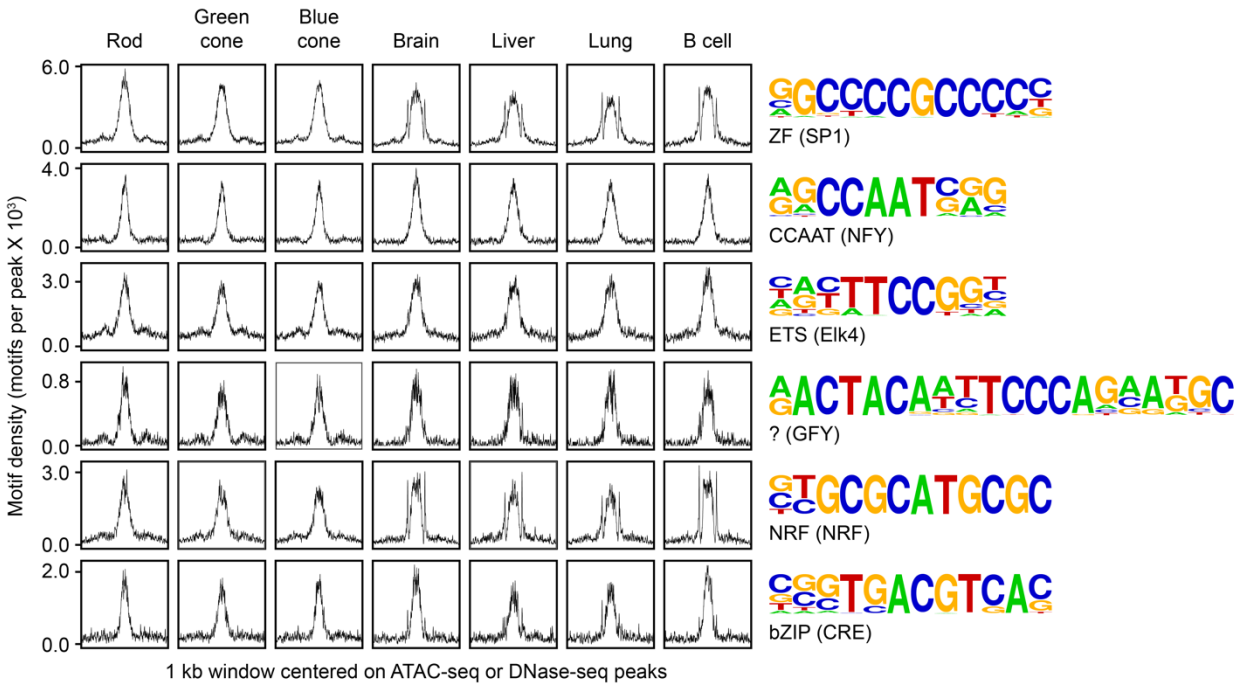


Supplemental Figure S2.5. GO enrichment analysis for cell-type-specific ATAC-seq peaks. (A) Rod-specific peaks were highly enriched for terms specifically related to vision and photoreceptor biology. (B) Cone-specific peaks were highly enriched for terms related to neurodevelopment and smooth muscle biology.

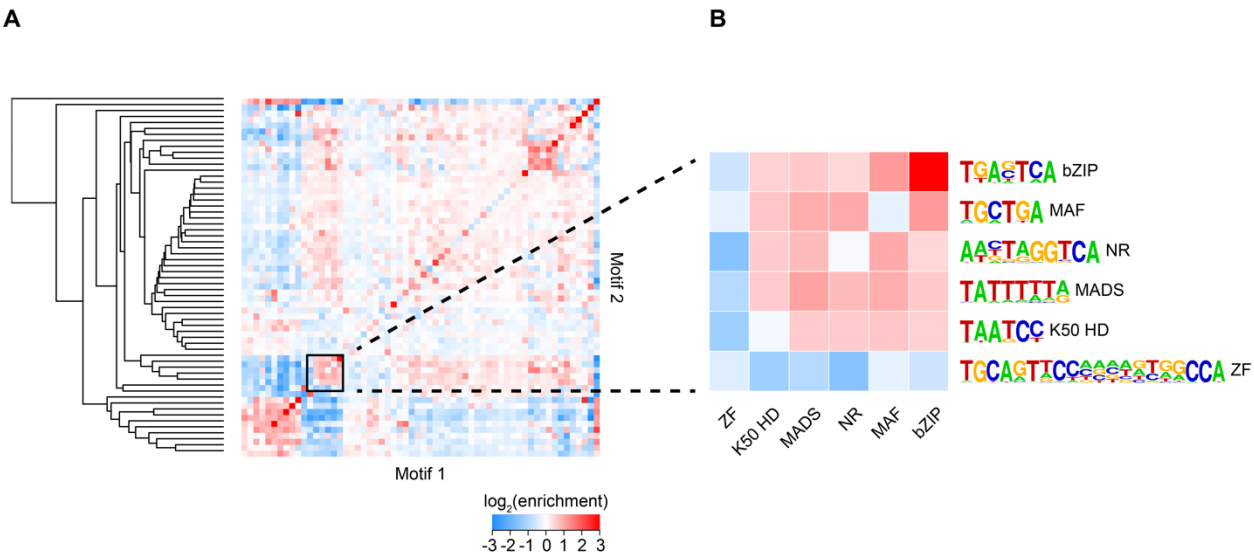


Supplemental Figure S2.6. Distinct features of photoreceptor promoters and enhancers. ATAC-seq peaks within -1 kb to +100 bp of a TSS were classified as "promoter" peaks, whereas peaks outside this interval were classified as "enhancer" elements. (A-B) The average ATAC-seq

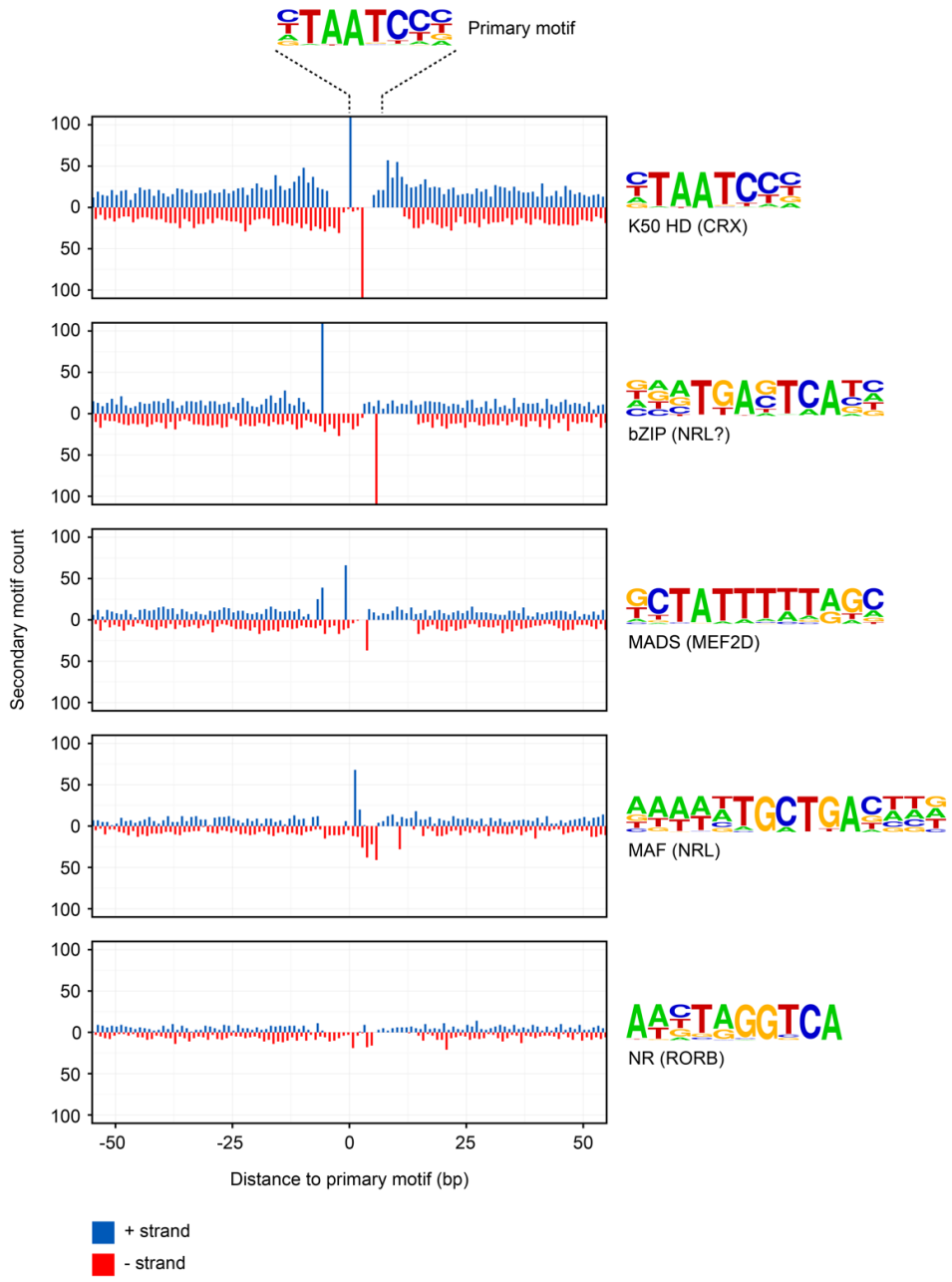
signal (normalized reads per base pair) was higher in promoters compared to enhancers, and the enrichment over baseline was broader. (C-D) Mean conservation (phastCons 60-way vertebrate conservation) as well as GC content (E-F) were also higher and broader surrounding promoter peaks vs. enhancer peaks. (G-H) Promoter and enhancer peaks were mapped to genes by nearest TSS. In both rods and cones, the presence or absence of a promoter peak was more strongly correlated with gene expression than the presence or absence of an enhancer peak, but the presence of either was associated with increased expression.



Supplemental Figure S2.7. Enrichment of known TF binding sites in TSS-proximal (promoter) peaks. For each panel, distance from peak summit (-500 bp to 500 bp) is plotted on the x-axis and motif density (motifs per peak at each position) is plotted on the y-axis, illustrating central enrichments of the motifs presented. Motifs are labeled by TF class, and the specific factor from which the sequence logo shown was derived is indicated in parenthesis. Compared to enhancers (Fig. 2.4A), motif enrichment in promoters is highly similar across cell and tissue types.



Supplemental Figure S2.8. Motif co-occurrence in photoreceptor ATAC-seq peaks. (A) Enrichment of motif co-occurrence was calculated for pairs of 60 motifs based on the observed counts of pairs in peaks compared to the expected counts based on the known frequency of individual motifs in peaks (Methods). Rows and columns are hierarchically clustered by Euclidean distance using average linkage. (B) The motifs from the boxed region in (A). bZIP, NR, MAF, MADS, and K50 HD TFs cluster together and are mutually co-enriched. These motifs are depleted of CTCF motifs (ZF, blue row and column).



Supplemental Figure S2.9. Preferential motif spacing flanking K50 HD sites. Each panel shows the strand-specific per nucleotide density of the indicated secondary motif upstream and downstream of K50 HD motifs in photoreceptor ATAC-seq peaks. Secondary motifs on the positive strand are shown in blue, above the midline. Secondary motifs on the negative strand are shown in red, below the midline.

Chapter 3: A massively parallel reporter assay reveals context-dependent activity of homeodomain binding sites *in vivo*

Cone-rod homeobox (CRX) is a paired-like homeodomain transcription factor (TF) and a master regulator of photoreceptor development in vertebrates. The *in vitro* DNA binding preferences of CRX have been described in detail, but the degree to which *in vitro* binding affinity is correlated with *in vivo* enhancer activity is not known. In addition, paired-class homeodomain TFs can bind DNA cooperatively as both homodimers and heterodimers at inverted TAAT half-sites separated by two or three nucleotides. This dimeric configuration is thought to mediate target specificity, but whether monomeric and dimeric sites encode distinct levels of activity is not known. Here, we use a massively parallel reporter assay, CRE-seq, to determine how local sequence context shapes the regulatory activity of CRX binding sites in mouse photoreceptors. We assay inactivating mutations in >1700 TF binding sites, and we find that dimeric CRX binding sites act as stronger enhancers than monomeric CRX binding sites. Furthermore, the activity of dimeric half-sites is cooperative, dependent on a strict 3-bp spacing, and tuned by the identity of the spacer nucleotides. Saturating single-nucleotide mutagenesis of 195 CRX binding sites shows that, on average, changes in TF binding site affinity are correlated with changes in regulatory activity, but this relationship is obscured when considering mutations across multiple CREs. Taken together, these results demonstrate that the activity of CRX binding sites is highly dependent on sequence context, providing insight into photoreceptor gene regulation and illustrating functional principles of homeodomain binding sites that may be conserved in other cell types.

3.1 Introduction

Advances in high-throughput sequencing have enabled genome-wide mapping of *cis*-regulatory elements (CREs) in diverse cell types and tissues, providing a powerful resource for studying the

role of noncoding genetic variation in health and disease [4]. Nevertheless, predicting the functional impact of regulatory variants requires a detailed understanding of the sequence constraints mediating interactions between CREs and transcription factors (TFs). The DNA binding preferences of thousands of TFs have been characterized *in vitro* [162-164], but how accurately these models predict the activity of TF binding sites *in vivo* is not known.

Cone-rod homeobox (CRX) is a paired-like homeodomain TF and a master regulator of photoreceptor gene expression in vertebrates [64, 69, 70]. The DNA binding preferences of CRX and the closely related homologs OTX1 and OTX2 have been defined by quantitative gel shift, high-throughput SELEX, and protein binding microarray (PBM) [3, 109, 162, 165]. All three methods yield similar models of TF binding site affinity, with the high-affinity consensus 5'-TAATCC-3'. In addition, detailed structural studies have shown that paired-class homeodomains can bind DNA cooperatively as both homodimers and heterodimers at inverted TAAT repeats [128, 129, 166]. Paired-class TFs with a lysine (K) or a glutamine (Q) in position 50 of the homeodomain (K50 or Q50 TFs) bind dimeric half-sites with a 3-bp spacing, while those with a serine (S) in position 50 (S50) bind with a 2-bp spacing. Furthermore, K50 homeodomains, including CRX, prefer cytosines 3' of each TAAT half-site (5'-TAATCNGATTA-3'). In an earlier study, we mapped CRX occupancy in adult mouse photoreceptors by ChIP-seq, showing that CRX-bound regions *in vivo* are enriched for both monomeric and dimeric CRX binding sites [68]. However, whether monomeric and dimeric TF binding sites encode distinct levels of activity is not known.

Massively parallel reporter assays (MPRAs) leverage barcoded reporter constructs to quantify the activity of large numbers of CREs simultaneously. We and others have used MPRAs to characterize sequence features mediating tissue- and cell-type-specific regulatory activity in cell lines [50, 51, 167], explanted tissues [41, 42, 46], and *in vivo* [45, 48]. MPRAs

can be designed to measure both the promoter activity of individual CREs (the level of reporter expression they drive autonomously) as well as their enhancer or repressor activity (the level of reporter expression they drive above or below a basal promoter). Previously, we used an MPRA, CRE-seq, to quantify the effects of all possible single-nucleotide substitutions in a 52-bp segment of the *Rhodopsin* promoter in mouse retina [41]. We found that changes in the affinity of two targeted CRX binding sites were moderately correlated with changes in CRE activity, but the extent to which these results generalize to other loci is not known. We subsequently quantified the effects of mutating CRX binding sites in hundreds of CRX-bound promoters and enhancers, but we only assayed a single mutation in each CRE, precluding a comprehensive analysis of the relationship between TF binding site affinity and activity [42].

In the current study, we aim to build upon these results to further understand how sequence context shapes the regulatory activity of CRX binding sites. First, we use CRE-seq to quantify the enhancer activity of >1200 CRX-bound regions in explanted, intact mouse retinas, and we identify primary sequence features and orthogonal epigenomic measurements that are correlated with native CRE activity. We then assay inactivating mutations in >1700 CRX binding sites within these CREs to compare the regulatory activity of monomeric and dimeric CRX binding sites. Furthermore, when multiple CRX binding sites are present within individual CREs, we mutate them individually and in combination to assess TF binding site cooperativity. Next, to examine the relationship between CRX binding site affinity and activity in detail, we select 195 CRX binding sites and quantify the effect of all possible single-nucleotide substitutions in a 13-bp window overlapping each TF binding site. Finally, to determine if the spacing between dimeric half-sites is functionally constrained, we assay the effect of small insertions and deletions on CRE activity.

Taken together, our results demonstrate that both the precise configuration of individual CRX binding sites as well as the broader sequence context in which they occur play key roles in determining their activity. These observations provide specific insight into how photoreceptor CREs encode information, and they have general implications for the quantitative modeling and interpretation of *cis*-regulatory variation. Furthermore, as homeodomains constitute one of the largest classes of metazoan TFs, defining functional principles of CRX binding sites may provide general insights applicable to the *cis*-regulatory mechanisms of additional cell types and tissues.

3.2 A simple model combining dinucleotide frequencies and TF binding sites accurately predicts CRX occupancy *in vivo*

We previously used ChIP-seq to profile CRX occupancy in adult mouse photoreceptors, showing that CRX-bound regions are phylogenetically conserved, have elevated GC content, and are enriched for K50 homeodomain binding sites, as well as binding sites for several other TFs [68] (Fig. 3.1A-C, Supplemental Fig. S3.1-S3.2). Strikingly, we subsequently reported that none of these features alone is an accurate predictor of CRX occupancy genome-wide [42]. Here, we revisit these data to determine if a model integrating multiple predictors can accurately classify CRX-bound vs. CRX-unbound regions and provide insight into the sequence features that determine CRX occupancy *in vivo*.

For this analysis, we selected 5250 200-bp sequences centered on the summits of CRX ChIP-seq peaks, focusing on distal enhancers (i.e., peaks occurring >1 kb upstream and >100 bp downstream of any TSS) (Fig. 3.1A). We then selected 52500 200-bp CRX-unbound sequences sampled randomly from the mouse genome, controlling for GC and repeat content [168]. We scored each CRX-bound and CRX-unbound sequence for two classes of features—all ten non-redundant dinucleotide frequencies and occurrences of 206 TF binding sites [162, 169]. We then

used these features to train logistic regression classifiers to differentiate CRX-bound from CRX-unbound sequences. We used lasso regularization to control model complexity [170], and we evaluated model performance with repeated ten-fold cross-validation (see Methods).

To develop an intuition for the information contained in specific classes of features, we measured the performance of models using increasingly complex subsets of variables, quantified by area under the receiver operating characteristic (AUC-ROC) and area under the precision recall curve (AUC-PR) (Supplemental Fig. 3.3, Supplemental Table S3.1). First, we considered a model using only dinucleotide frequencies. We found that despite the limited resolution of these predictors (average dinucleotide frequencies over a 200-bp window), this model performs nearly as well as using counts of individual CRX binding sites (AUC-ROC=0.75 vs. AUC-ROC=0.77, respectively). Next, we noted that identifying CRX binding sites by scoring sequences with a position weight matrix (PWM) and simply counting matches above a defined threshold fails to account for the quality of the PWM match (and, implicitly, TF binding site affinity). To address this issue, we binned counts of CRX sites into four 'affinity' classes (high, medium, low, and very low) based on the PWM match p-value (see Methods), thereby partitioning one variable (the number of CRX sites) into four. This modification significantly improves model performance (AUC-ROC=0.84), highlighting the value of incorporating graded TF binding site affinity into models of TF occupancy. Finally, we trained a model using counts of 206 TF binding sites (using a single, fixed threshold) selected from a database of PWMs for mouse and human TFs [162, 169]. This approach again yielded a significant improvement in performance (AUC-ROC=0.92), illustrating that multiple TF binding site models capture non-redundant information in CRX ChIP-seq peaks.

Next, we combined dinucleotide frequencies and counts of 206 TF binding sites binned by motif score in a single model, yielding the best performance among the logistic regression

classifiers we tested (AUC-ROC of 0.95) (Fig. 3.1D). Two key advantages of regularized logistic regression are that (1) the fitted coefficients have an accessible interpretation (in this case, the relative contribution of individual dinucleotide frequencies or TF binding sites to the likelihood [log odds] of CRX binding), and (2) lasso regularization shrinks the coefficients of irrelevant or redundant variables towards zero. Strikingly, from an initial set of 834 predictors, only 18 have non-zero coefficients in the final model predicting CRX occupancy: GC, AG, and CG dinucleotide frequencies, as well as six TF binding sites, the four strongest of which correspond to homeodomain binding sites (Supplemental Table S3.2). These results suggest that, genome-wide, a substantial fraction of CRX occupancy can be explained by the presence of one (or more) of a limited set of homeodomain TF binding sites in a favorable dinucleotide context. Furthermore, these results illustrate that multiple PWMs (even ostensibly similar homeodomain PWMs) capture non-redundant information about TF binding preferences, predicting TF occupancy more accurately in combination than any one individually.

Recently, several groups have developed methods for predicting TF occupancy from k-mer content (DNA words of length k, typically six to ten base pairs), and these have generally yielded highly accurate models [168, 171-174]. We used one of these tools, gkm-SVM, to classify CRX-bound vs. CRX-unbound regions, and find that, indeed, this model outperforms classifiers built on dinucleotide frequencies and PWMs (AUC-ROC=0.99). Nevertheless, k-mer based models are challenging to interpret. For example, one can use a gkm-SVM model to score all non-redundant k-mers and then cross-reference strong predictors with databases of TF binding sites, but it may not be clear how many discrete TF binding site signals are represented by these lists. Accordingly, we believe that the interpretability of regularized logistic regression using experimentally-derived TF binding sites complements the high prediction accuracy of k-mer based methods.

Finally, given that multiple models of primary sequence features accurately predict CRX occupancy *in vivo*, we asked if the location of informative features was spatially constrained relative to the center of CRX ChIP-seq peaks. To address this question, we trained new versions of our logistic regression classifier (using dinucleotide frequencies and counts of TF binding sites binned by motif p-value), extracting features from windows ranging from 20 bp up to 200 bp (Fig. 3.1E). We find that the maximum AUC-ROC and AUC-PR values are obtained by restricting features to the central ~140 bp relative to the summit, suggesting that most of the information mediating CRX occupancy is contained within the footprints of individual nucleosomes (i.e., ~146 bp) [7].

3.3 Dinucleotide frequencies and TF binding site content are correlated with the enhancer activity of CRX-bound regions *in vivo*

Having identified primary sequence features that predict CRX occupancy genome-wide, we asked if these same features were correlated with regulatory activity *in vivo*. To quantify the activity of many CREs simultaneously, we used CRE-seq, as described previously (Fig. 3.2A) [41, 42, 45, 46]. Briefly, we used array-based oligonucleotide synthesis to generate a library of 1230 100-bp DNA fragments centered on native enhancers identified by CRX ChIP-seq [68]. We cloned this library upstream of a photoreceptor promoter (either *pRho* or *pCrx*) driving DsRed, and we included a CRE-specific DNA barcode in the 3' UTR of each construct. We electroporated this library into newborn (P0) mouse retinas, which were cultured for eight days. Finally, we harvested RNA and DNA and PCR amplified and sequenced barcodes to measure copy number-adjusted enhancer activity (the ratio of normalized cDNA and DNA counts for each barcode, see Methods).

In previous studies [42, 46], we used CRE-seq to assay the activity of CRX-bound regions cloned upstream of a 205-bp segment of the *Rho* promoter (*pRho*), which drives high expression

in rod photoreceptors [83, 175]. In the current study, we assayed our library on *pRho* as well as on the 206-bp *Crx* promoter (*pCrx*), which drives moderate expression in both rod and cone photoreceptors, to assess the effect of distinct promoters on CRE activity. CRE-seq generates reproducible estimates of enhancer activity on both *pRho* and *pCrx* (Spearman correlation coefficients between biological replicates of 0.98-0.99 and 0.93-0.94, respectively) (Supplemental Fig. S3.4), and we observe similar distributions of relative CRE activity on both promoters (Fig. 3.2B). However, this distribution is skewed toward higher expression on *pRho* compared to *pCrx*, which may indicate that the linear range of CRE-seq is sensitive to the absolute activity of the promoter.

To identify primary sequence features that influence CRE activity, we examined the correlation between dinucleotide frequencies or counts of TF binding sites and CRE-seq expression (FDR<0.05) (Fig. 3.2C-D). Consistent with previous work [42, 46], we find that CRE activity is negatively correlated with the number of CRX binding sites when assayed on *pRho* (PCC=-0.088) (Fig. 3.2C). In contrast, CRE activity is positively correlated with the number of CRX binding sites when assayed on *pCrx* (PCC=0.062). This inversion may reflect complex promoter-enhancer interactions, or it may be a technical effect related to the absolute activity of each promoter.

Overall, we identified 24 sequence features that were significantly correlated with CRE activity on either *pRho* or *pCrx* (Fig. 3.2D). These correlations were directionally consistent between *pRho* and *pCrx*, except for CRX binding sites (noted above) and AT-rich dinucleotide classes (AA, AT, and TA) as well as CA and AC dinucleotides. In contrast to the sequence features that predict CRX occupancy (Supplemental Table S3.2), we find that a more complex set of sequence features are correlated with enhancer activity, including multiple non-K50 TF binding

sites. This suggests that, while homeodomain binding sites are necessary for CRX binding, they are not sufficient to determine activity.

TF binding sites that are correlated with CRE-seq activity can be assigned to five distinct TF families (Fig. 3.2D). TF binding sites for nuclear receptors, basic helix-loop-helix, and zinc finger TFs are positively correlated with activity, whereas TF binding sites for Q50 homeodomain and T-box TFs are negatively correlated with activity (see discussion of K50 motifs above). Most of these families correspond to one or more TFs that play a well-characterized role in mouse photoreceptor development (Supplemental Fig. S3.5): *Esrrb*, *Nr2e3*, *Rorb*, *Rxrg*, and *Thrb* (nuclear receptors); *Neurod1* (a basic helix-loop-helix TF); *Sp1*, *Sp3*, and *Sp4*, among others (zinc finger TFs); *Crx* and *Otx2* (K50 homeodomain TFs); and *Rax* (a Q50 homeodomain TF). In contrast, a role for T-box TFs in mammalian photoreceptors has not yet been established, but recent transcriptome profiling of mouse photoreceptors shows that, among all 13 *Tbx* family members in mouse, only *Tbx2* and, to a lesser degree, *Tbx3* are expressed in the early postnatal period (Supplemental Fig. S3.5) [176, 177].

In addition to assessing sequence features correlated with CRE activity, we compared our CRE-seq data to epigenomic profiling data from retina and other tissues (Fig. 3.2E, Supplemental Table S3.3) [4, 118, 119, 132, 178]. Specifically, we examined the correlation between CRE activity and the signal strength (read depth) of open chromatin (ATAC-seq or DNase-seq), TF ChIP-seq, and histone modification ChIP-seq data (Supplemental Fig. S3.6, Supplemental Table S3.4). Datasets generated in retina have the strongest absolute correlations, suggesting that CRE-seq captures tissue- and cell-type-specific regulatory activity. Activating histone marks, ChIP-seq for the photoreceptor-specific TFs CRX and NRL, and retina and photoreceptor chromatin accessibility are positively correlated with CRE activity, while repressing histone marks and non-retina chromatin accessibility are negatively correlated with CRE activity. Interestingly, a number

of these correlations are stronger than those observed between activity and primary sequence features (Supplemental Table S3.4).

To test how well combinations of primary sequence features or epigenomic data ('chromatin features') predict the enhancer activity of CRX-bound regions, we used regularized logistic regression to classify CREs with high (>3-fold above the median) vs. low (within 1.2-fold of the median) activity (see Methods). For both *pRho* and *pCrx* CRE-seq data, chromatin features classify CREs more accurately (AUC-ROC=0.79, and AUC-ROC=0.78, respectively) than sequence features (AUC-ROC=0.75, and AUC-ROC=0.69, respectively) (Fig. 3.2E, Supplemental Table S3.5). Nevertheless, the performance of models predicting CRE activity using either feature set is modest compared to the performance of models classifying CRX occupancy (Fig. 3.1D). This may be due in part to working with smaller numbers of elements when predicting activity, but it may also reflect a need to include additional features and/or non-additive interactions to predict CRE activity—i.e., fundamental differences in the *cis*-regulatory grammar underlying TF occupancy and CRE activity.

3.4 Dimeric CRX binding sites encode stronger enhancers than monomeric CRX binding sites

As discussed above, homeodomain TFs bind DNA as both monomers and dimers, and CRX ChIP-seq peaks are enriched for both monomeric and dimeric CRX binding sites (Fig. 3.1C). Furthermore, monomeric and dimeric CRX binding sites within CRX ChIP-seq peaks have distinct profiles of evolutionary conservation (Fig. 3.3A). Specifically, both TAAT cores of dimeric binding sites have elevated average phyloP scores [179], suggesting both are functional. In addition, the most highly conserved position within the TAAT core (in both monomeric and dimeric CRX binding sites) is the second adenine (TA<u>A</u>T). We previously showed that

substitutions at this position effectively eliminate CRX binding *in vitro* [109], a finding consistent with the key role of this position in mediating homeodomain-DNA interaction [180].

To quantify the regulatory activity of individual TF binding sites, we used CRE-seq to assay the effect of inactivating mutations in 1756 CRX binding sites within a subset of the 1230 CRX-bound regions described above (Fig. 3.3B). In general, CRX binding sites act as enhancers—74% of mutations decrease activity, and 25% decrease activity by more than two-fold (vs. 4% that increase activity by more than two-fold) (Fig. 3.3C). Nearly half (49%) of these changes are statistically significant (FDR<0.05), and 85% of statistically significant differences are decreases in activity (Fig. 3.3D).

We next asked if the predicted affinity of CRX binding sites is correlated with their activity. We calculated the motif score for each targeted CRX site using both a monomeric and a dimeric homeodomain PWM, and we binned sites based on the match p-value into very low-, low-, medium-, and high-affinity classes. To account for the fact that CRX binding sites can act as either enhancers or repressors, we considered the absolute log fold change in activity (instead of the signed log fold change). We were surprised that monomeric CRX binding site scores are not significantly correlated with activity. In contrast, dimeric CRX binding site scores are significantly (though modestly) correlated with activity (PCC=0.23). Similarly, the activity of targeted sites does not vary significantly across affinity classes defined by the monomeric homeodomain PWM, whereas all pairwise comparisons (except medium vs. high) are significantly different when comparing affinity classes defined by the dimeric homeodomain PWM (FDR<0.05) (Fig. 3.3E, Supplemental Table S3.6). Taken together, these data show that medium- and high-affinity dimeric CRX binding sites encode stronger enhancer activity than monomeric CRX binding sites.

3.5 Pairs of CRX binding sites act cooperatively

To characterize the interactions between CRX binding sites, we selected 225 CREs with exactly two monomeric CRX binding sites (with unconstrained intersite spacing and orientation) and mutated them individually and in combination. For each pair, we defined the binding site with the higher monomeric homeodomain PWM score to be "site 1." Consistent with the above results, we find that mutating either binding site has a similar effect on wild-type activity, independent of their relative affinities (Fig. 3.3F). Furthermore, the effects of mutating distinct CRX binding sites within a single CRE are highly correlated (PCC=0.81) (Supplemental Fig. S3.7), which suggests that flanking sequence plays a key role in determining CRX binding site activity. Finally, mutating both CRX binding sites has only a slightly greater effect than mutating either site individually (Fig. 3F, Supplemental Fig. S3.7). This result suggests that at least some pairs of CRX binding sites act synergistically—i.e., they increase (or decrease) CRE activity more in combination than we would predict based on the marginal activity of either site alone. To test this, we modeled the activity of each CRE as a linear function of the presence or absence of both targeted CRX binding sites and an interaction term (see Methods). For nearly half (47%) of the 225 CREs we analyzed, the interaction term was significant (FDR<0.05) (Supplemental Table S3.7), suggesting that non-additive interactions between CRX binding sites are common among photoreceptor CREs.

We used this same approach (mutating sites individually and in combination) to dissect the activity of half-sites in 130 dimeric CRX binding sites. Similar to the above, we defined "half-site 1" and "half-site 2" based on the orientation of the highest scoring match to a dimeric homeodomain PWM. We again find that mutating either half-site alone significantly decreases activity, but that mutating them together has minimal additional effect (Fig. 3.3F), suggesting that both half-sites play a key role in determining activity. Indeed, the effects of mutating either half-site within dimeric binding sites are highly correlated (PCC=0.85) (Supplemental Fig. S3.7),

indicating that both half-sites contribute to wild-type activity. Finally, we again used linear models to test for interactions between half-sites, which revealed statistically significant interactions in 62% of cases (FDR<0.05) (Supplemental Table S3.8), indicating that half-sites frequently act cooperatively.

3.6 The correlation between CRX binding site affinity and activity is CRE-dependent

We were surprised that the effects of inactivating mutations in CRX binding sites (n=1756) are not strongly correlated with TF binding site affinity (Fig. 3.3E). However, this analysis only considered a single mutation (TAAT to TACT). To characterize the relationship between CRX binding site affinity and activity in greater detail, we used CRE-seq to quantify the effect of all possible single nucleotide substitutions in a 13-bp window overlapping 97 monomeric and 98 dimeric CRX binding sites (Fig. 3.4A). We again find that mutations in dimeric CRX binding sites are enriched for strong effects compared to mutations in monomeric CRX binding sites, suggesting that dimeric homeodomain motifs identify functional CRX binding sites more specifically (Fig. 3.4B). Furthermore, these data define the key functional positions within CRX binding sites, with mutations in the TAAT core and the first 3' nucleotide (typically a cytosine) having the strongest effects on activity, including both half-sites within dimeric CRX binding sites (Fig. 3.4B-D; Supplemental Fig. S3.8-S3.9). In general, these results are consistent with the *in vitro* DNA binding preferences of CRX that we determined previously by quantitative gel shift (Fig. 3.4C) [109], as well as the DNA binding preferences of closely related homeodomain TFs that have been estimated by high-throughput SELEX (Supplemental Fig. S3.10) [162]. Strikingly, the median effects of specific substitutions at each position within CRX binding sites are highly correlated with the associated changes in PWM score (PCC=0.82 for both monomeric and dimeric CRX binding sites) (Supplemental Fig. S3.10).

While the median effects of specific substitutions are strongly correlated with their impact on CRX binding, we find substantial variation in the effects of specific mutations across CREs (Fig. 3.4D). Accordingly, when we consider the relationship between changes in TF binding site affinity and activity across all mutations (without aggregating by position or nucleotide identities), this correlation is significantly lower (PCC=0.32 for monomeric CRX binding sites, and PCC=0.35 for dimeric CRX binding sites) (Supplemental Fig. S3.10). Nevertheless, within individual CREs, these correlations are often stronger (median absolute PCC=0.54 for monomeric CRX binding sites, and median absolute PCC=0.58 for dimeric CRX binding sites). This apparent contradiction is explained by examining the relationship between TF binding site affinity and activity within individual CREs (Supplemental Fig. S3.11). For a given CRE, changes in activity are typically well-described by a linear function of changes in affinity, but the slope of the fit is CRE-dependent. In other words, sequence context appears to re-scale the relationship between changes in CRX binding site affinity and activity across CREs—preserving the relative effects of different mutations, but altering their absolute effects (Supplemental Fig. S3.11). These results suggest that, while models of TF binding site affinity may accurately predict the relative effects of mutations within a single CRE, predicting the relative effects of mutations across multiple CREs may prove challenging.

In addition to examining the relationship between TF binding site affinity and activity, we asked if phylogenetic conservation is correlated with the effects of mutations in CRX binding sites. We find that the average conservation (phyloP scores) of specific positions within CRX binding sites are correlated with the median effects of mutations at those positions (compare Fig. 3.3A and Fig. 4D, see also Supplemental Fig. S3.12) (PCC=0.89-0.95). Nevertheless, conservation is poorly correlated with mutation effects across CREs (PCC=0.07-0.10), suggesting that simple

conservation scores should be used cautiously when prioritizing candidate regulatory variants from multiple loci.

3.7 The activity of dimeric CRX binding sites depends on half-site spacing

In addition to analyzing single-nucleotide substitutions, we quantified the effect of small insertions and deletions in monomeric and dimeric CRX binding sites on CRE activity (Fig. 3.5). In particular, we tested if the activity of dimeric CRX binding sites was dependent on half-site spacing. First, we made all 1-, 2-, and 3-bp deletions (n=7 per targeted site) of the three nucleotides 3' of the TAAT core in the 97 monomeric and 98 dimeric CRX binding sites described above. On average, deletions significantly decrease activity ($p<1.4x10^{-4}$) (Fig. 3.5A), but deletions of different sizes do not have significantly different effects. Interpreting these results with respect to spacing is challenging since deletions impact the affinity of individual half-sites as well as potential TF interactions, though we note that deletions in dimeric CRX binding sites have stronger effects than deletions in monomeric CRX binding sites ($p<8.5x10^{-4}$).

To test the effect of alterations in half-site spacing while minimizing effects on the affinity of individual half-sites, we made the following insertions (Fig. 3.5B). For 1-bp insertions, we doubled the center nucleotide (e.g., 5'-TAAT CAG ATTA-3' to 5'-TAAT CA<u>A</u>G ATTA-3'). For 2-bp insertions, we tripled the center nucleotide (e.g., 5'-TAAT CAG ATTA-3' to 5'-TAAT CA<u>AA</u>G ATTA-3'). And for 3-bp insertions, we doubled the entire triplet (e.g., 5'-TAAT CAG ATTA-3' to 5'-TAAT CAG<u>CAG</u> ATTA-3'). All three mutations significantly decrease the activity of dimeric, but not monomeric, CRX binding sites ($p<4.9x10^{-7}$). Taken together, these data indicate that the activity of dimeric CRX binding sites depends on a strict three-nucleotide spacing, consistent with structural studies of paired-class homeodomain TF-DNA complexes [129, 166].

Finally, our analysis of single-nucleotide substitutions suggests that the activity of dimeric CRX binding sites is optimized by the spacer triplet CCG, which creates the highest-affinity K50 homeodomain binding sites on each strand given the spacing constraints described above. However, TAAT cores (separated by three base pairs on opposite strands) with alternative spacer nucleotides are also enriched in CRX ChIP-seq peaks. To determine the activity of these different spacer sequences, we substituted the six most enriched triplets (CCG, CAG, AGG, AAG, CTC, and CCA) into each of the 98 dimeric CRX binding sites tested above (in both orientations). First, we note that the effect of individual substitutions is similar regardless of their orientation (Supplemental Fig. S3.13), suggesting that dimeric CRX binding sites are not oriented. In addition, we find that specific spacer sequences have distinct effects on CRE activity (Fig. 3.5C). These effects are positively correlated with the resulting affinity of each half-site for K50 homeodomain TFs, and negatively correlated with the resulting affinity for Q50 homeodomain TFs. In particular, the spacer CCA has the lowest activity, and this sequence forms a high-affinity K50 binding site on the forward strand, and a high-affinity Q50 binding site on the reverse strand. Previously, K50 and Q50 paired-class TFs have been shown to antagonize one another at specific dimeric homeodomain binding sites [166], suggesting that the reduced activity of dimeric CRX binding sites with a CCA spacer may reflect such antagonism.

3.8 Accounting for baseline CRE activity improves the prediction of variant effects

We next asked how accurately primary sequence features predict the effects of mutations in CRX binding sites. We used simple linear regression to model the effects of changes in CRX binding sites as a function of changes in TF binding site affinity for each of the 195 CREs in our dense substitution analysis. As described above, gkm-SVM learns a flexible k-mer representation of CRX binding preferences that classifies CRX-bound vs. CRX-unbound regions more accurately

than PWMs [181]. In addition, deltaSVM is an extension of gkm-SVM that attempts to predict the effects of regulatory variants by calculating the difference between mutant and wild-type scores under a specific gkm-SVM model [182]. Accordingly, to predict the effects of mutations in CRX binding sites on activity, we compared representing changes in TF binding site affinity as the difference between mutant and wild-type PWM scores vs. deltaSVM scores (Supplemental Tables S3.9-S3.11).

Fitting models for each CRE individually, we find that deltaSVM scores derived from CRX ChIP-seq data predict the effects of mutations in CRX binding sites more accurately than changes in PWM scores (R^2=0.41 vs. R^2=0.32) (Fig. 3.6A). However, both approaches perform significantly worse when we fit models for all mutations simultaneously (R^2=0.14 for deltaSVM scores and R^2=0.12 for changes in PWM scores) (Fig. 3.6B). We hypothesized that the performance of these models might be limited by neglecting the contribution of sequence features outside the targeted binding sites. In addition, we asked whether models derived from multiple epigenomic assays could be combined to yield a more complete representation of sequence features relevant to photoreceptor CRE activity. Accordingly, we trained gkm-SVM models on 15 additional datasets (DNase-seq, ATAC-seq, TF ChIP-seq, and/or histone ChIP-seq from retina and six non-retinal tissues), and quantified how well combinations of gkm-SVM (wild-type) and/or deltaSVM scores (mutant-specific) predict the activity of mutations in CRX binding sites with or without interactions (see Methods) [4, 118, 119, 132, 178]. We found that none of these multi-feature models substantially improves upon CRX ChIP-seq deltaSVM scores (R^2=0.14-0.16) (Fig. 3.6C). These results suggest that, while additive models of primary sequence features accurately predict TF occupancy, more complex models (e.g., ones incorporating interactions between TF binding sites) may be necessary to accurately predict regulatory activity.

As an alternative to predicting the effects of mutations (i.e., change in CRE activity) from primary sequence features alone, we asked how accurately the combination of wild-type (baseline) CRE activity and sequence features could predict the activity of mutant CREs. Strikingly, wild-type activity alone explains 66% of the variation in mutant activity (Fig. 3.6D). Furthermore, combining wild-type expression with deltaSVM scores from multiple datasets significantly improves performance (R^2=0.73), and incorporating interactions between wild-type activity and deltaSVM scores yields additional improvement (R^2=0.76) (Fig. 3.6D). These data demonstrate that the effects of mutations in CRX binding sites are highly dependent on wild-type CRE activity. Accordingly, while modeling the effects of CRX binding site mutations from primary sequence alone remains challenging, knowledge of baseline CRE activity significantly improves the prediction of the functional effects of *cis*-regulatory variants.

3.9 Discussion

Understanding the impact of *cis*-regulatory variation on phenotypic diversity and disease requires a detailed understanding of how CREs encode information in TF binding sites. Here, we assay thousands wild-type and mutant CREs to identify sequence features that modulate the activity of CRX binding sites, and we find that the activity of CRX binding sites depends on multiple layers of sequence context. Both the affinity of individual CRX binding sites as well as their configuration (i.e., monomeric vs. dimeric) have modest but significant effects on their activity. Moreover, the broader sequence context in which CRX binding sites occur, including the number and affinity of additional K50 as well as non-K50 TF binding sites, has even stronger effects on their activity. We also find that multiple instances of CRX binding sites within individual CREs frequently act cooperatively, and interactions between CRX (and potentially other) TF binding sites may explain why flanking sequence plays a stronger role than TF binding site affinity in determining CRX binding site activity. Thus, while CRX occupancy can largely be explained by an additive model

of homeodomain TF binding sites and dinucleotide content, the activity of CRX-binding sites appears to be determined by a richer vocabulary of sequence features, including significant non-additive interactions. These results are broadly consistent with a recent study of PPARG-bound regions in a mouse adipocyte cell line [50], suggesting that they highlight general mechanisms by which mammalian enhancers encode *cis*-regulatory activity.

As described above, our analysis of TF binding sites correlated with native CRE activity confirms a role for several families of TFs known to regulate photoreceptor development (i.e., K50 and Q50 homeodomain, nuclear receptor, basic helix-loop-helix, and zinc finger TFs) (Fig. 3.2D). One surprising result is that T-box motifs are negatively correlated with photoreceptor CRE activity (Fig. 3.2D). Although T-box TFs have no established role in mammalian photoreceptor development, *Tbx2b* mutant zebrafish show a transfating of ultraviolet cones into rods [183]. In addition, a recent expression profiling study in chicken shows that *Tbx2* is expressed in the embryonic retina and may play a role in violet cone development [143]. In mouse, *Tbx2* is the most highly expressed T-box TF in developing photoreceptors (Supplemental Fig. S3.5) [176, 177], and TBX2 has generally been shown to act as a repressor [184]. Together with CRE-seq data, these observations suggest that TBX2 acts as a transcriptional repressor in the developing mouse retina, which would establish this TF as an ancient and highly conserved regulator of vertebrate photoreceptor identity.

In addition to T-box motifs, we find that Q50 TF binding sites are negatively correlated with CRE activity (Fig. 3.2D), suggesting they may also act as repressors. Among Q50 TFs in mouse, *Rax* is the most highly expressed in photoreceptors [176, 177] and has been shown to play a role in photoreceptor maturation and survival [75]. However, reporter assays in cultured cells suggest that RAX is a weak activator of photoreceptor CREs, not a repressor [75]. Accordingly, additional functional studies are needed to determine if individual Q50 TF binding sites are

activators or repressors in photoreceptors. Interestingly, other retinal cell types express additional Q50 TFs (e.g., VSX2 in bipolar cells), raising the possibility that retinal CREs harboring Q50 TF binding sites may encode distinct activities in different cell types.

Recently, Inoue *et al.* reported a systematic comparison of MPRAs using integrating vs. non-integrating lentiviral reporter constructs in a human liver cell line [49]. The authors found that activity estimates from integrating constructs were somewhat more reproducible (Spearman correlation of 0.94 vs. 0.91 between biological replicates) and more strongly correlated with orthogonal genomic annotations (Spearman correlation of 0.52 vs. 0.39), raising concerns about the generalizability of conclusions drawn from episomal reporter assays (such as CRE-seq). In the current study, we find that the activity of native CRX-bound regions (as estimated by CRE-seq) is correlated with the presence of binding sites for photoreceptor-specific TFs as well as orthogonal retina- and photoreceptor-specific epigenomic datasets. Furthermore, saturating mutagenesis of CRX binding sites demonstrates that the average position-specific effects of different substitutions are correlated with their effects on CRX binding *in vitro*. These results suggest that CRE-seq captures relevant features of cell-type-specific *cis*-regulatory grammar even though elements are not assayed in their native context. Nevertheless, assaying CREs in their native genomic context is necessary to validate predictions from MPRAs.

Currently, there is intense interest in developing quantitative models to predict the effects of noncoding variants on CRE activity to identify causal variants underlying disease association signals [174, 182, 185]. Several groups have recently evaluated the accuracy of specific methods for predicting the activity of native CREs (as estimated by MPRA), finding that current models explain 4-38% of the variation in regulatory activity [49, 50, 182]. Here, we report that linear models using multiple deltaSVM scores explain 16% of the variation in the effects of mutations in CRX binding sites (i.e., changes in activity). Thus, multiple studies suggest that additional work

is needed to predict regulatory activity from primary sequence, and we hypothesize that models incorporating higher order interactions between features are likely to prove valuable. Nevertheless, we show that combining estimates of wild-type CRE activity with deltaSVM scores explains most of the variation in mutant CRE activity (R^2=0.76) (note that R^2=0.91 between biological replicates). Additional experiments are necessary to determine if this approach can predict the effects of mutations in non-CRX binding sites or in other cell types and tissues. If it can, combining reference sets of wild-type cell-type-specific CRE activity (e.g., ascertained via CRE-seq or similar methods) with cell-type-specific models of sequence grammar (e.g., deltaSVM models) could be a powerful strategy for predicting the cell-type-specific effects of novel *cis*-regulatory variants.

3.10 Materials and methods

CRE-seq library design. 1270 target regions were selected from CRX ChIP-seq peaks (100-bp elements centered on peak summits). Candidate monomeric CRX binding sites were identified with FIMO (v4.11.2) [186] using the OTX2_DBD_1 PWM [162] and a p-value threshold of $p < 10^{-3}$. Candidate dimeric CRX binding sites were identified by matches to the pattern 5'-TAAKNNNMTTN-3' or 5'-NAAKNNNMTTA-3' (two TAAT or TAAG cores on opposite strands, separated by three base pairs, allowing up to one mismatch in the first thymidine of either TAAT core). For monomeric CRX binding sites, we mutated each TAAT core to TACT (one mutant construct per site). For dimeric CRX binding sites, we mutated each TAAT half-site to TACT individually as well as both in combination (three mutant constructs per site). In addition, within each CRE, we mutated each CRX binding site individually as well as all in combination (n+1 mutant constructs for n sites). For saturating mutagenesis, we selected 100 monomeric and 100 dimeric CRX binding sites and made all possible single-nucleotide substitutions in a 13-bp window overlapping each CRX binding site as well as selected insertions, deletions, and

substitutions of the spacer nucleotides. Each of the 14987 wild-type and mutant target sequences were paired with six or seven unique 13-bp barcodes (pairwise edit distance >2), yielding a final library of 100,000 targets.

CRE-seq library construction. 170-bp oligos were generated by array-based oligonucleotide synthesis through a limited licensing agreement with Agilent Technologies. Oligos were amplified and cloned into (Rho-prox)-DsRed [83] as described previously [41, 42]. The resulting plasmid library was sequenced to assess CRE representation (99.7% of 100,000 targeted oligos were detected, 98.8% at >1 barcode per million barcodes). A promoter-DsRed reporter construct (either pRho-DsRed or pCrx-DsRed) was then cloned between each CRE and barcode to generate the final CRE-seq library. See Supplemental Methods for details.

CRE-seq assay. Retinas were isolated from P0 CD-1 mice, and electroporated in a solution containing 30 μg of CRE-seq library and 30 μg of CAG-GFP as described previously [41, 42, 45, 187]. Electroporated retinas were cultured for eight days, at which point they were harvested, washed three times with HBSS (Gibco), and stored in TRIzol (Invitrogen) at -80˚C. Five retinas were pooled for each biological replicate, and three replicates were performed for each CRE-seq library. RNA and DNA were extracted from TRIzol according to manufacturer instructions, and RNA samples were treated with TURBO DNase (Invitrogen) as described previously [45]. cDNA synthesis was performed with SuperScript IV (Invitrogen) using an oligo(dT) primer according to manufacturer instructions. 197-bp fragments (including unique 13-bp barcodes) were amplified from cDNA and DNA samples by PCR (21 cycles), and indexed sequencing adapters were added with an additional round of PCR (eight cycles). Libraries were sequenced on an Illumina HiSeq 3000 loaded at 150 pM with 20% PhiX DNA.

CRE-seq data processing. Sequencing reads were demultiplexed with ea-utils (v1.02) (github.com/ExpressionAnalysis/ea-utils), and 13-bp barcodes were identified in sequencing reads based on exact matches to designed barcodes, including 22 bp of fixed sequence context (18 bp upstream and 4 bp downstream). A pseudocount of 1 was added to raw cDNA and DNA counts, which were then scaled across biological replicates using the median ratio method [148]. Only barcodes with >5 scaled counts in each DNA library were included in the analysis. For each target sequence, activity was estimated as the log2 ratio of cDNA counts (summed over barcodes) over DNA counts (summed over barcodes). These values were then quantile normalized across replicates [188].

Models of TF occupancy and CRE activity. Logistic regression models predicting TF occupancy or CRE activity and linear regression models predicting mutation effects were implemented in R (v3.3) [158] using the packages glmnet [189] and caret (Kuhn 2017). For each target sequence, the ten non-redundant dinucleotide frequencies were calculated on both strands, and instances of TF binding sites were identified with FIMO (v4.11.2) [186] using a database of 206 human and mouse PWMs derived from high-throughput SELEX and ChIP-seq [162, 169]. gkm-SVM models predicting TF occupancy, chromatin accessibility, and histone modifications were trained with LS-GKM (word length 11, 7 informative columns) [173] using background sequence sets generated with the gkm-SVM R package [168, 181]. See Supplemental Methods for details.

Data visualization. Plots were generated in R (v3.3) [158] using the ggplot2 package [160].

Animals. Mouse husbandry and all procedures were conducted in accordance with the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health, and were approved by the Washington University in St. Louis Institutional Animal Care and Use Committee.

Data access. Sequencing reads, barcode counts, and normalized expression values for annotated CREs have been deposited in the GEO (GSE106243). See Supplemental Table S3.3 for an overview of previously generated datasets also referenced in this study.

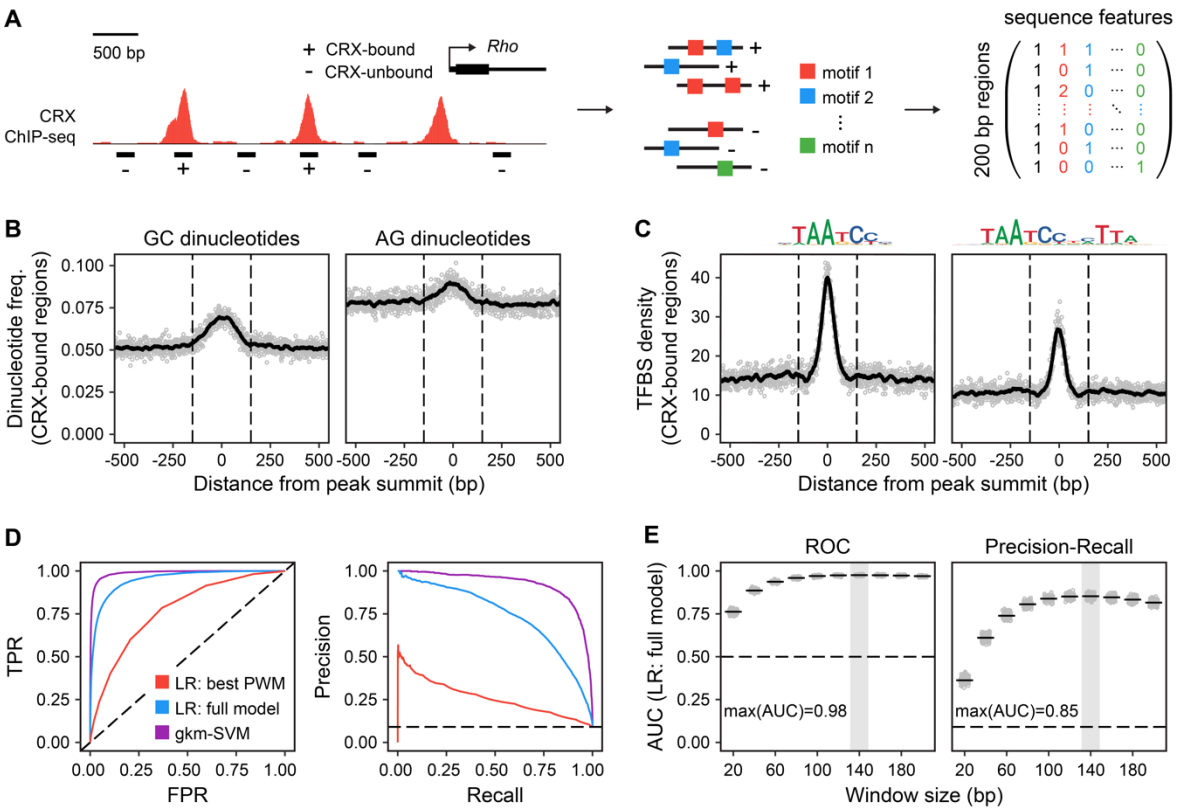


Figure 3.1. Primary sequence features predict CRX occupancy in vivo. (A) Schematic of analytical approach. 5250 CRX-bound regions and 52500 CRX-unbound regions were selected based on CRX ChIP-seq data (200-bp elements centered on peak summits). Feature vectors composed of average dinucleotide frequencies and/or counts of specific TF binding sites (up to 206) were defined for each sequence. (B) CRX ChIP-seq peaks are centered on local enrichments of specific dinucleotide classes, including elevated GC and AG dinucleotide content. (C) CRX ChIP-seq peaks are centered on local enrichments of specific TF binding sites, including monomeric and dimeric homeodomain (CRX) binding sites. (D) Performance of specific models classifying CRX-bound vs. CRX-unbound sequences visualized with ROC (FPR vs. TPR) and PR (recall vs. precision) curves. TPR: true positive rate. FPR: false positive rate. Dashed lines: performance of random classifiers. LR: logistic regression. LR: best PWM—counts of dimeric CRX binding sites (single PWM) (AUC-ROC=0.77, AUC-PR=0.26). LR: full model—dinucleotide frequencies and counts of 206 TF binding sites (binned by PWM score) (AUC-

ROC=0.95, AUC-PR=0.74). See Supplemental Table S3.2 for feature weights. gkm-SVM: 11-mers with 7 informative positions (AUC-ROC=0.99, AUC-PR=0.92). (E) Performance of LR: full model with features extracted from windows of different sizes (20 bp to 200 bp).

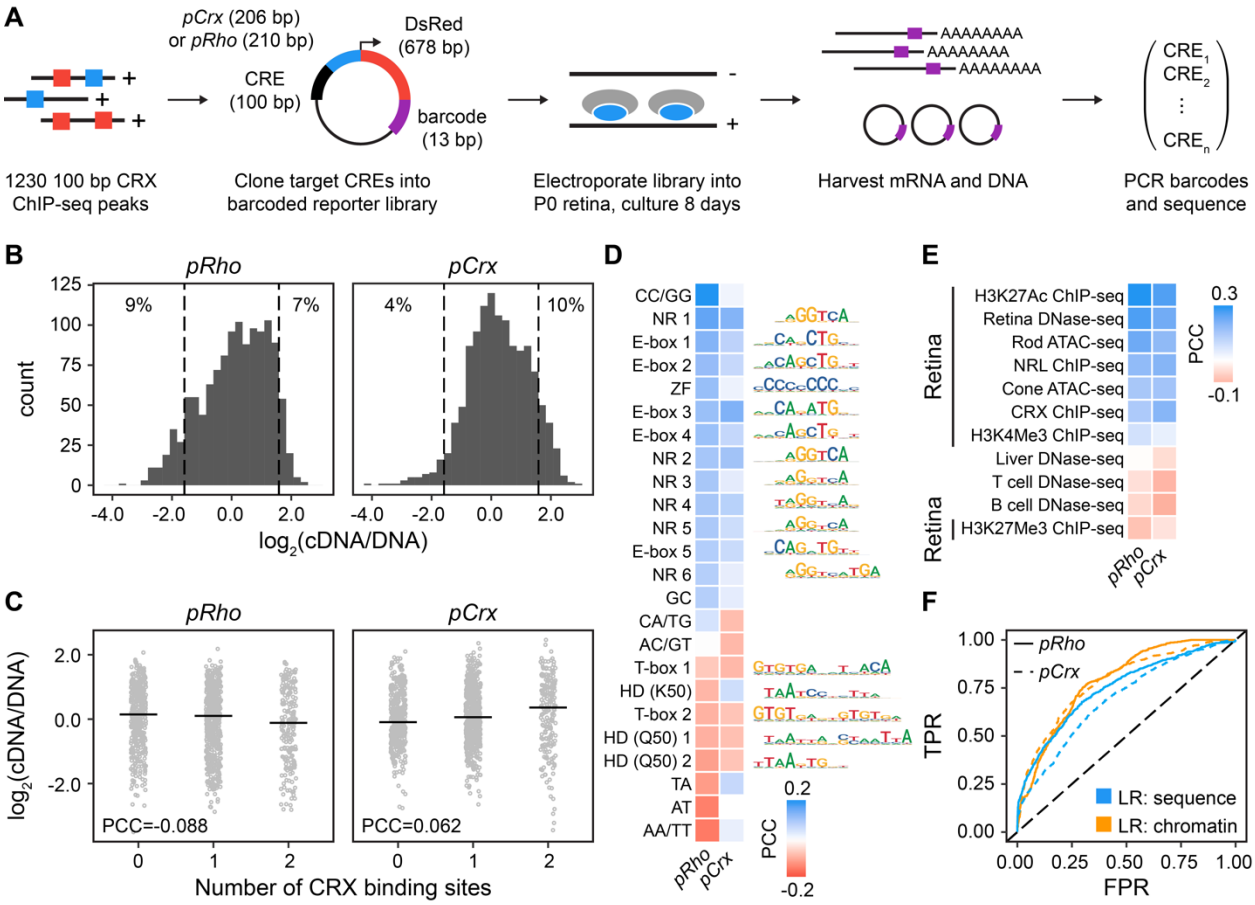


Figure 3.2. Primary sequence features are correlated with CRE activity in vivo. (A) Schematic of experimental approach. 100-bp elements centered on CRX ChIP-seq peaks were cloned upstream of a photoreceptor promoter driving DsRed with CRE-specific barcodes. Constructs were electroporated into P0 mouse retina and cultured for eight days, at which point RNA and DNA were harvested and barcodes were amplified and sequenced to quantify activity. (B) Distribution of activity of elements assayed on either pRho or pCrx. Data are median-centered. Dashed lines: three-fold decrease or increase relative to median. The percentage of constructs with activity above or below this threshold is indicated. (C) Correlation between number of dimeric CRX binding sites and activity on pRho and pCrx. (D) Heatmap of Pearson correlations between specific dinucleotide frequencies or counts of TF binding sites and activity on pRho and pCrx. Included features were significantly correlated with activity on at least one promoter. (E) Heatmap

of Pearson correlations between epigenomic datasets and CRE activity. (F) Performance of specific models classifying elements with low (within 1.2-fold of the median) vs. high (>3-fold above the median) activity. LR: sequence—logistic regression classifier using dinucleotide frequencies and counts of 206 TF binding sites (binned by PWM score) (pRho: AUC-ROC=0.75; pCrx: AUC-ROC=0.69). LR: chromatin—logistic regression classifier using signal strength from multiple epigenomic datasets (pRho: AUC-ROC=0.79; pCrx: AUC-ROC=0.78). Dashed line: performance of a random classifier.

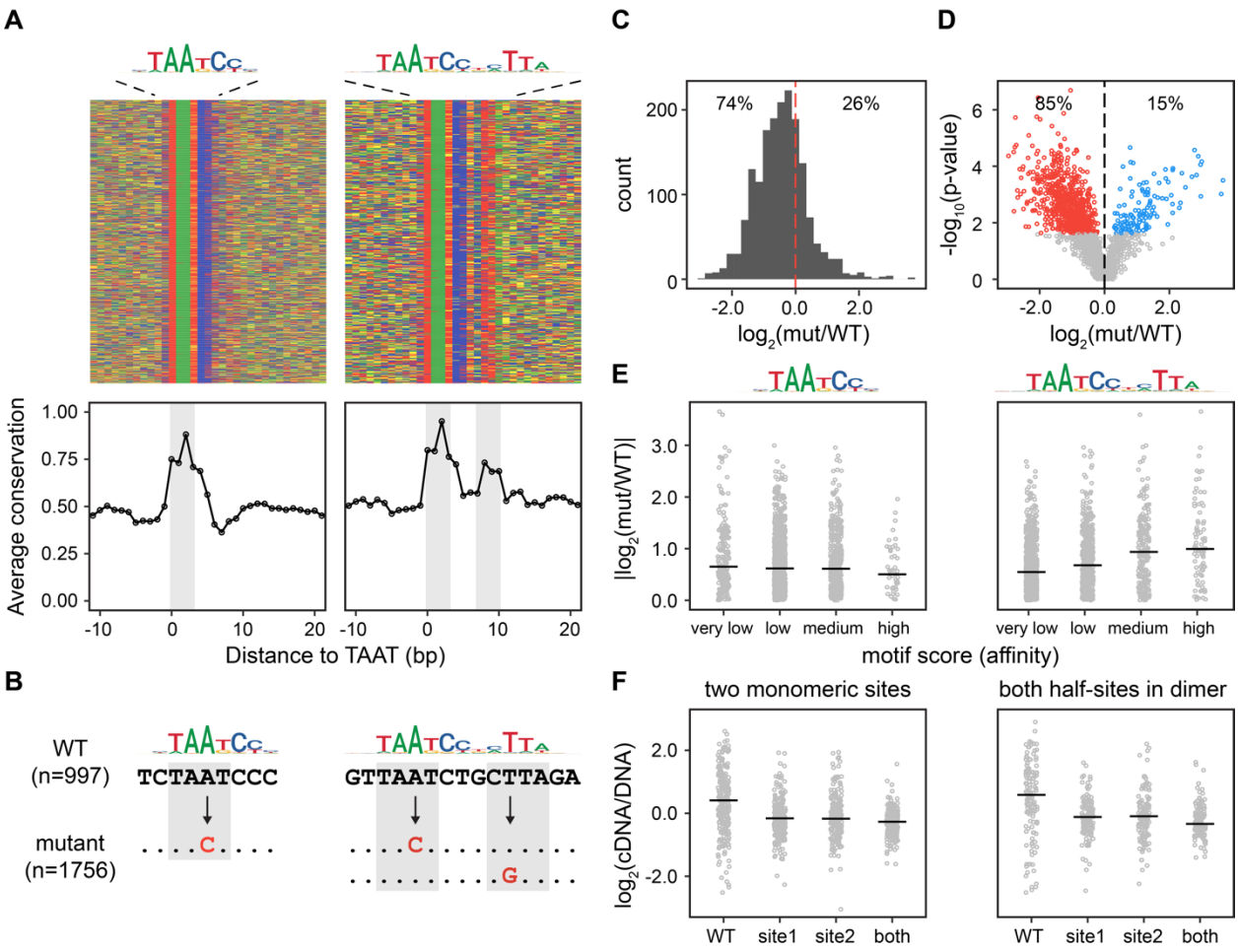


Figure 3.3. Dimeric CRX sites have higher activity than monomeric CRX sites. (A) Upper panels: heatmaps of nucleotide content in a 30-bp window centered on monomeric or dimeric CRX binding sites. Rows corresponds to distinct TF binding sites, columns correspond to distinct positions, and tiles are colored by nucleotide identity. Lower panels: average conservation (60-way vertebrate phyloP scores) at each position. Positions 0-3 (and 7-10 for dimeric TF binding sites) correspond to TAAT cores (gray boxes). (B) Schematic of experimental approach. The effects of 1-bp substitutions (TAAT to TACT) in 1756 CRX binding sites within CRX ChIP-seq peaks were quantified by CRE-seq. (C) Distribution of mutation effects (log2 fold change). (D) Volcano plot of mutation effects. Among mutations that significantly change activity (FDR<0.05), 85% decrease activity and 15% increase activity. Red: significant decrease in activity. Blue:

significant increase in activity. Gray: change in activity not significant. (E) Absolute effect size vs. monomeric or dimeric PWM score (binned by match p-value). (F) Left panel: activity distributions of CREs with two monomeric CRX binding sites when neither, one, or both are mutated. Right panel: activity distributions of CREs with dimeric CRX binding sites when neither, one, or both half-sites are mutated.

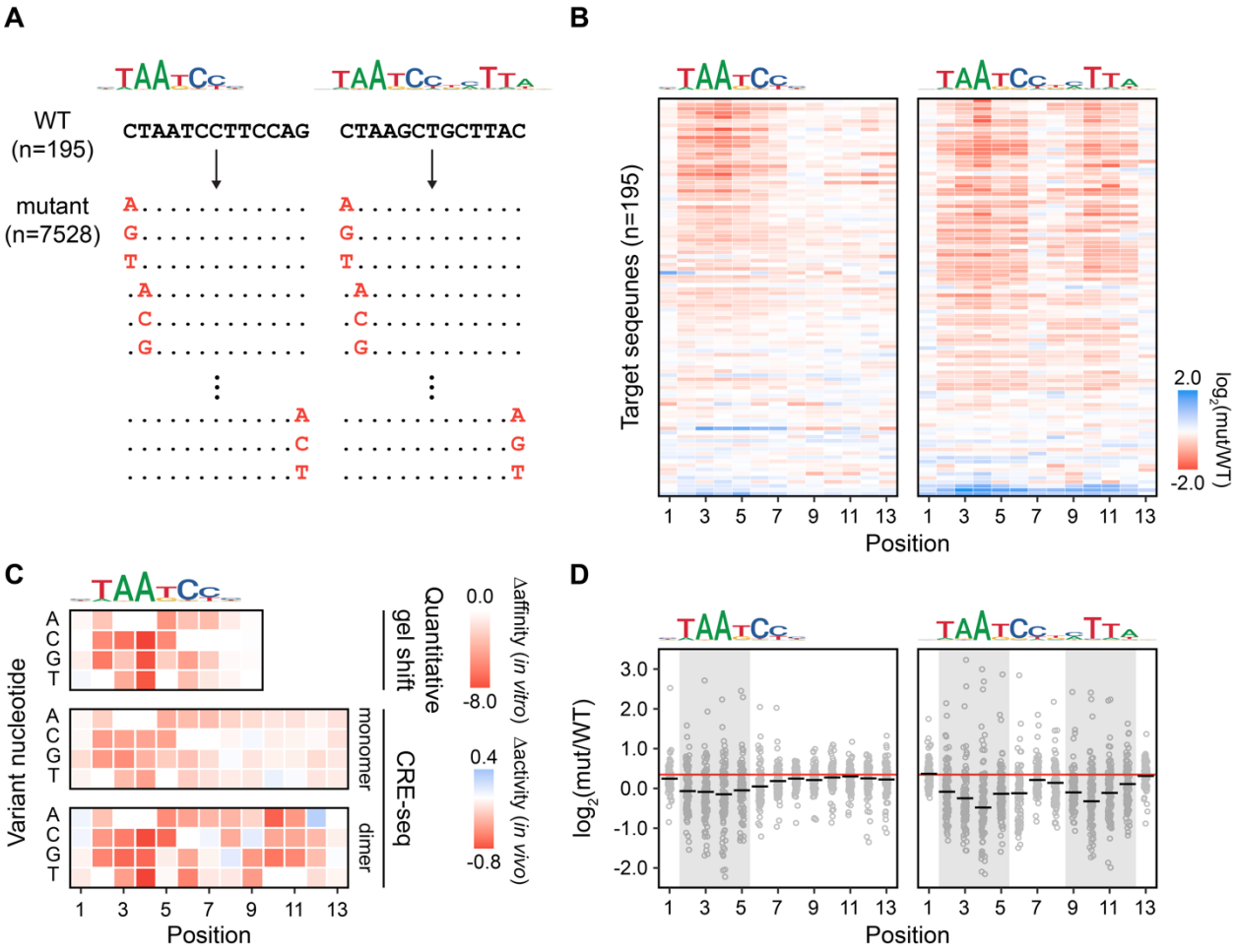


Figure 3.4. Dense mutagenesis of monomeric and dimeric CRX binding sites. (A) Schematic of experimental approach. All single-nucleotide substitutions in a 13-bp window overlapping 97 monomeric and 98 dimeric CRX binding sites were quantified by CRE-seq (n=39 mutations per TF binding site). (B) Heatmaps of median effects (across all three substitutions) at each position (columns) in each targeted CRX binding site (rows). Each heatmap represents 97 or 98 distinct elements, and rows are sorted by wild-type activity (high to low). (C) Heatmaps of median effects (across all target sites) at each position (columns) for specific substitutions (rows). Top panel: change in CRX binding site affinity determined by quantitative gel shift for all possible substitutions in a single target sequence [109]. Middle panel: change in activity determined by CRE-seq for substitutions in 97 monomeric CRX binding sites. Bottom panel: change in activity

determined by CRE-seq for substitutions in 98 dimeric CRX binding sites. (D) Scatter plot of median effects (for all three substitutions) (y-axis) at each position (x-axis) in each targeted CRX binding site. Points represent different targeted CRX binding sites, and horizontal bars represent the median across all targets.

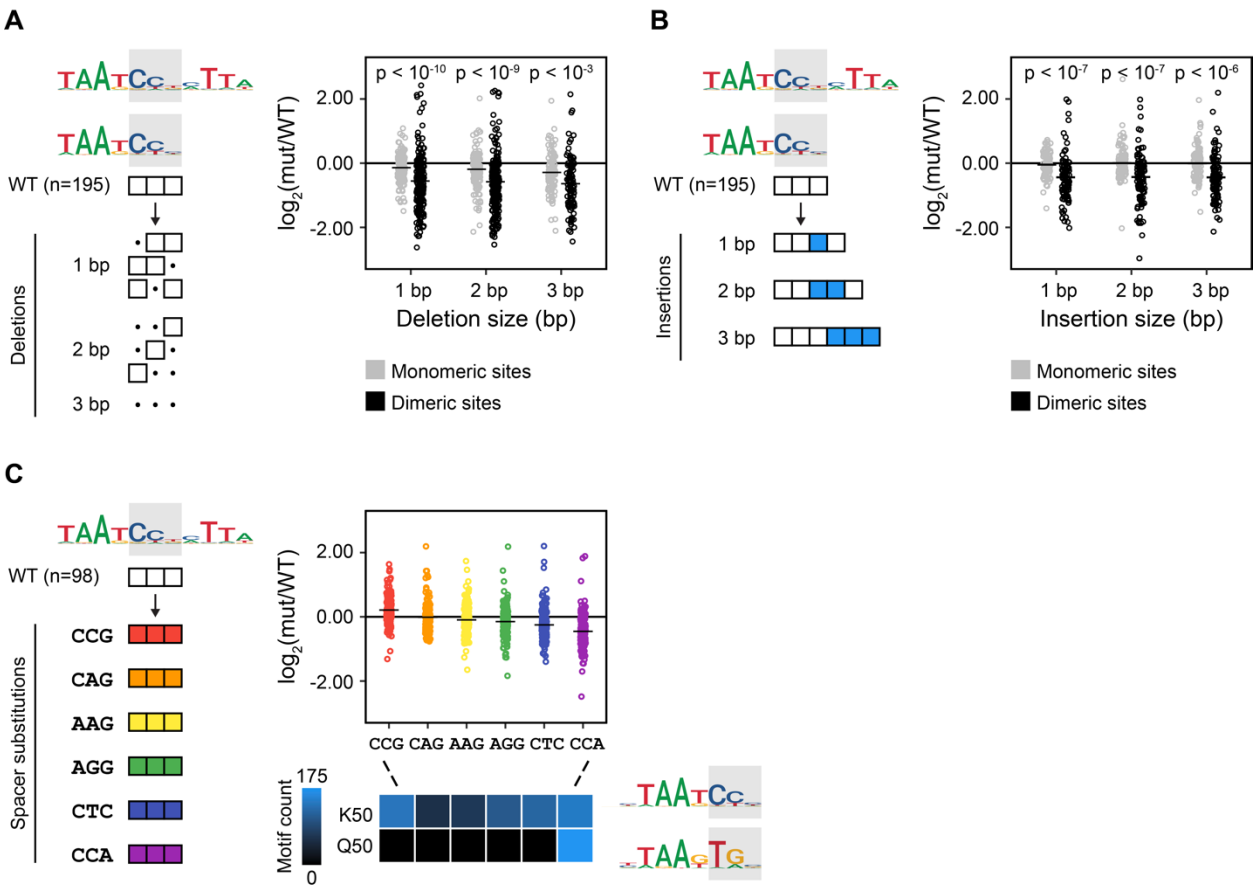


Figure 3.5. The activity of Dimeric CRX binding sites depends on half-site spacing. (A) Left: schematic of experimental approach. The effect of all 1-, 2-, and 3-bp spacer deletions in 195 CRX binding sites were quantified by CRE-seq. Right: scatter plot of mutation effects. Points represent individual mutations and horizontal bars represent the median across all targets for deletions of the indicated size. (B) Left: schematic of experimental approach. The effect of specific one-, two-, and 3-bp spacer insertions in 195 CRX binding sites were quantified by CRE-seq. Right: scatter plot of mutation effects. Points represent individual mutations and horizontal bars represent the median across all targets for insertions of the indicated size. In (A) and (B), p-values are reported for Mann-Whitney-Wilcoxon tests comparing the distributions of effects between mutations in monomeric vs. dimeric CRX binding sites. (C) Left: schematic of experimental approach. The effects of selected 3-bp spacer substitutions in 98 dimeric CRX binding sites were quantified by

CRE-seq. Right: scatter plot of mutation effects. Points represent individual mutations and horizontal bars represent the median across all targets for the indicated substitution. The included heatmap shows counts of the indicated K50 and Q50 motifs among binding sites with each spacer substitution.

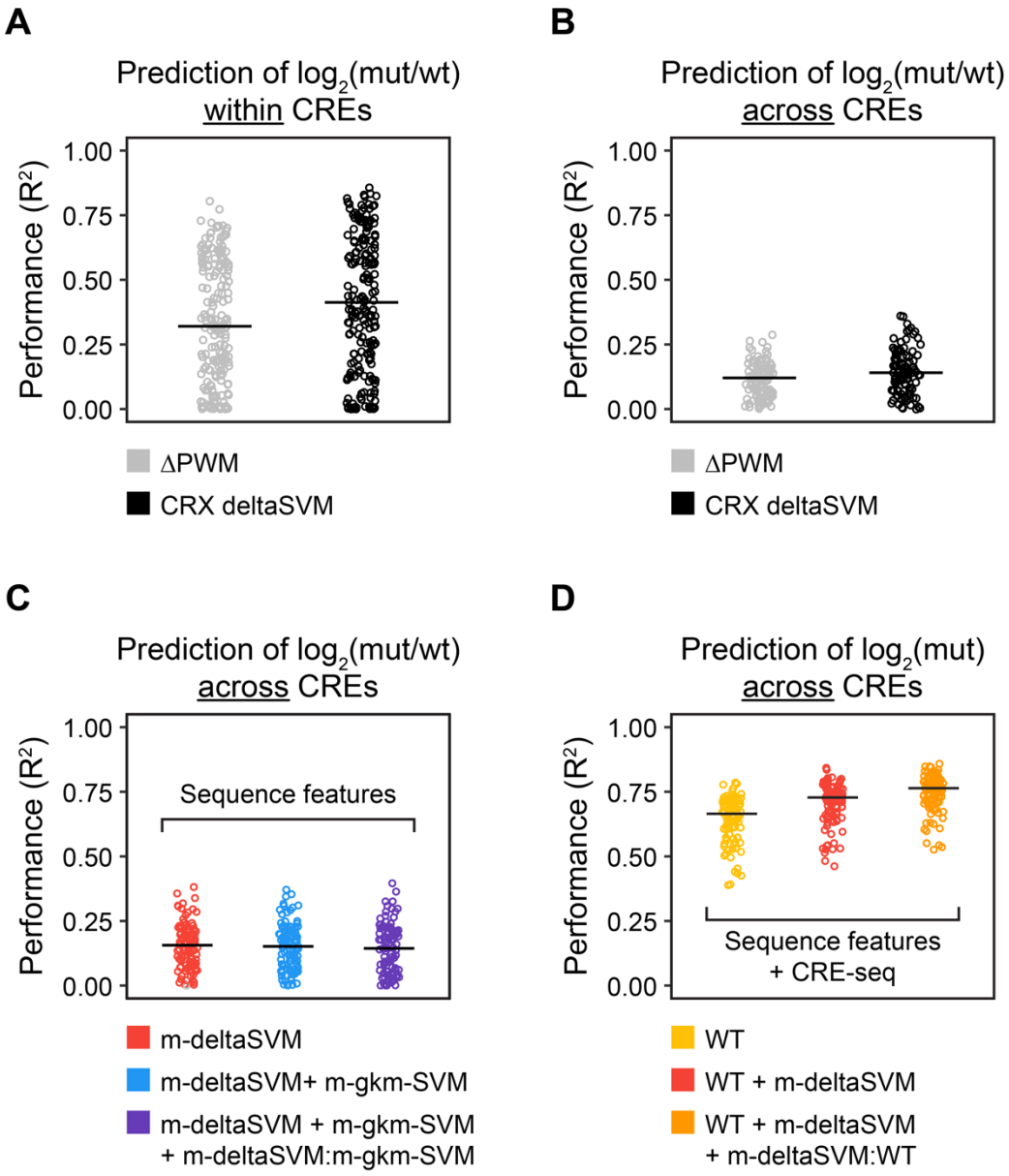


Figure 3.6. Accounting for baseline CRE activity improves the prediction of variant effects. (A) Performance (R^2) of simple linear regression predicting the effect of individual substitutions from changes in PWM scores or CRX ChIP-seq deltaSVM scores, fitting separate models for each CRE. (B) Same as in (A), except fitting a single model for all CREs. (C) Performance of multiple linear regression predicting the effect of individual substitutions using deltaSVM scores from multiple datasets (m-deltaSVM), m-deltaSVM and the corresponding gkm-SVM scores from multiple datasets (m-gkm-SVM), or m-deltaSVM scores and m-gkm-SVM scores including all

pairwise interactions. (D) Performance of multiple linear regression predicting mutant expression using wild-type (WT) expression, WT expression and m-deltaSVM scores, or WT expression, m-deltaSVM scores, and interactions between WT expression and deltaSVM scores. In (A), individual points represent the performance of models fit for different CREs (n=195). In (B-D), individual points represent the performance of models estimated from different folds of repeated ten-fold cross validation (n=100).

3.11 Supplemental materials

3.11.1 Supplemental methods

CRE-seq library construction

Oligo library structure. 100,000 170-bp oligos were ordered from Agilent with the following structure:

```
PCR_primer_1                 CRE_100bp          Barcode_13bp       PCR_primer_2
******************     *                 *************    ********************
GTAGCGTCTGTCCGTGTCGAC-X-ACTAGTCGGTACCNNNNNNNNNNNNNNGCGGCCGCCAACTACTACTACAG
                ******       ******  ******                                ********
                SalI     SpeI   KpnI               NotI
```

Oligo library amplification. 170-bp oligos were supplied at ~10 pmol and reconstituted in 100 µl TE (yielding a stock of 100nM or 11 ng/µl). Library oligos were amplified in 25 µl PCR reactions as follows: 1 µl 100nM library oligos (11 ng), 12.5 µl Phusion Hot Start Flex 2X Master Mix (NEB), 1.25 µl 10 mM PCR_primer_1, 1.25 µl 10 mM PCR_primer_2, 0.75 µl DMSO (Agilent), and 8.25 µl H_2O. PCR conditions were as follows: 98˚C for 30 seconds, 6 cycles of (98˚C for 10 seconds, 59˚C for 30 seconds, 72˚C for 30 seconds), and 72˚C for 5 minutes. PCR reactions were purified with a MinElute PCR Purification Kit (Qiagen) and eluted in 10 µl EB.

Oligo library digest. PCR-amplified oligos were digested as follows: 10 µl PCR products, 3 µl CutSmart buffer (NEB), 0.3 µl SalI-HF (NEB), 0.3 µl NotI-HF (NEB), and 16.4 µl H_2O. Reactions were incubated at 37˚C for 3 hours and then run on a 10% TBE gel (Bio-Rad) at 50V for 2 hours. The gel was stained with SYBR gold (Invitrogen) for 30 minutes, and a ~140 bp band was excised and minced with a clean razor blade. Gel fragments were transferred to a 1.5 ml microcentrifuge tube, combined with an equal volume of elution buffer (0.5M NH_4OAc and 1 mM EDTA), and incubated at 37˚C overnight. Gel fragments were centrifuged at 10,000 x g for 10 minutes, and the supernatant was transferred to a fresh microcentrifuge tube. Gel fragments were resuspended in a

half volume of elution buffer, vortexed, centrifuged at 10,000 x g for 10 minutes, and the supernatants were combined. DNA was then purified by ethanol precipitation and eluted in 15 µl EB (Qiagen).

Vector backbone preparation. The vector backbone was digested as follows: 1 µg (Rho-prox)-DsRed [83], 3 µl CutSmart buffer (NEB), 1 µl SalI-HF (NEB), 1 µl NotI-HF (NEB), final volume to 30 µl with H_2O. Reactions were incubated at 37˚C for 3 hours, and 1 µl alkaline phosphatase (Roche) was added after 2 hours. Restriction digests were run on a 1% agarose gel at 100V for 90 minutes, purified with a QIAquick Gel Extraction Kit (Qiagen), and eluted in 30 µl TE.

Oligo library transformation. Digested oligos and vector were ligated with Mighty Mix (Takara Bio) at room temperature for 30 minutes at a 1:1 molar ratio with 1 ng of insert per reaction. A total of 15 ligations were transformed into NEB 5-alpha Competent E. coli (High Efficiency) according to manufacturer instructions. After 1 hour outgrowth at 37˚C, transformations were pooled and split into three 5 ml aliquots, each of which was added to 150 ml of LB/ampicillin and cultured overnight in a 37˚C shaker. Aliquots of the pooled transformations were plated to estimate transformation efficiency (~$0.8x10^6$ CFUs). After overnight culture, plasmid DNA was harvested with a PureLink HiPure Maxiprep Kit (Thermo Fisher).

Sequencing library preparation (barcoded CRE plasmid library). Four PCR reactions amplifying a 212-bp fragment from the barcoded CRE plasmid library (prior to the insertion of promoter-DsRed constructs) were prepared as follows: 1 ng plasmid library, 25 µl Phusion Hot Start Flex 2X Master Mix (NEB), 2.5 µl 10 mM CRE-bc_F, 2.5 µl 10 mM CRE-bc_R, final volume to 50 µl with H_2O. PCR conditions were as follows: 30 seconds 98˚C, 14 cycles of (10 seconds 98˚C, 30 seconds 72˚C), and 5 minutes 72˚C. PCR reactions were purified with a MinElute PCR Purification Kit (Qiagen) and eluted in 10 µl of EB. An A-tailing reaction was prepared as follows:

300 ng purified PCR product, 5 µl NEBuffer 2 (NEB), 1 µl 10 mM dATP, 3 µl Klenow Fragment (3'→5' exo-) (NEB), final volume to 50 µl with H_2O. The A-tailing reaction was incubated at 37˚C for 30 minutes, then purified with Agencourt AMPure XP (Beckman Coulter) and eluted in 12 µl H_2O. Illumina adapters (annealed) were ligated to A-tailed PCR products in the following reaction: 10 µl A-tailed PCR products, 3.1 µl T4 DNA Ligase Reaction Buffer (NEB), 2 µl 25 µM Illumina adapters (annealed), 1 µl T4 DNA ligase (NEB), 13.9 µl H_2O. The ligation was incubated at 20˚C for 30 minutes, then purified with Agencourt AMPure XP (Beckman Coulter) and eluted in 32 µl H_2O. PCR reactions to enrich adapter-ligated fragments were prepared as follows: 20 µl adapter-ligated DNA, 25 µl Phusion Hot Start Flex 2X Master Mix (NEB), 2.5 µl 10 mM Multiplex_PCR_primer_1.0, 2.5 µl 10 mM SIC_index_NNNN. PCR conditions were as follows: 30 seconds 98˚C, 16 cycles of (10 seconds 98˚C, 30 seconds 72˚C), and 5 minutes 72˚C. The sequencing library was purified with Agencourt AMPure XP (Beckman Coulter), eluted in 50 µl H_2O, and sequenced with 2x250 bp reads on an Illumina MiSeq (~$11x10^6$ reads, ~100X coverage).

Subcloning of promoter-DsRed constructs. (Rho-prox)-DsRed [83] was digested with KpnI-HF (NEB), blunted, and ligated to eliminate a KpnI site between pRho and the coding sequence of DsRed. The modified pRho-DsRed sequence was amplified using the primers SpeI_pRho and KpnI-DsRed and cloned into pBlueScript with SpeI and KpnI. Similarly, pCrx-DsRed was amplified using the primers SpeI_pCrx and KpnI-DsRed and cloned into pBlueScript with SpeI and KpnI.

Insertion of promoter-DsRed constructs into barcoded CRE library. Restriction digests were prepared for the barcoded CRE library, pRho-DsRed in pBlueScript, and pCrx-DsRed in pBlueScript as follows: 1 µg plasmid DNA, 3 µl CutSmart buffer (NEB), 1 µl SpeI-HF (NEB), 1 µl KpnI-HF (NEB), final volume to 30 µl with H_2O. Digests were incubated at 37˚C for 3 hours, and 1 µl alkaline phosphatase (Roche) was added to the barcoded CRE library after 2 hours.

Digested fragments were run on an agarose gel at 100V for 90 minutes and gel purified using a QIAquick Gel Extraction Kit (Qiagen). Purified products were ligated and transformed into NEB 5-alpha Competent E. coli (High Efficiency) as described above (see Oligo library transformation), with an estimated transformation efficiency of $1.2x10^6$ CFUs.

CRE-seq sequencing library preparation

Sequencing library preparation (CRE-seq cDNA and DNA). PCR reactions were prepared for purified cDNA and DNA as follows: 2 µl purified DNA (or 3µl cDNA), 25 µl Phusion Hot Start Flex 2X Master Mix (NEB), 2.5 µl 10 mM read1_bc, 2.5 µl 10 mM read2_DsRed, final volume to 50 µl with H_2O. PCR conditions were as follows: 30 seconds 98˚C, 21 cycles of (10 seconds 98˚C, 30 seconds 72˚C), and 5 minutes 72˚C. PCR reactions were purified with a MinElute PCR Purification Kit (Qiagen) and eluted in 10 µl of EB. PCR products were amplified and indexed with an additional round of PCR as follows: 10 ng PCR products, 25 µl Phusion Hot Start Flex 2X Master Mix (NEB), 2.5 µl 10 mM Multiplex_PCR_primer_1.0, 2.5 µl 10 mM SIC_index_NNNN, final volume to 50 µl with H_2O. PCR conditions were as follows: 30 seconds 98˚C, 8 cycles of (10 seconds 98˚C, 30 seconds 72˚C), and 5 minutes 72˚C. PCR reactions were purified with a PureLink PCR Purification Kit (Thermo Fisher) and eluted in 30 µl of EB. cDNA and DNA libraries from biological replicates were pooled and sequenced 1x50 bp reads on an Illumina HiSeq 3000 (29-$43x10^6$ reads per library).

Bioinformatic analysis of barcoded CRE plasmid library

Bioinformatic analysis of barcoded CRE plasmid library. Prior to the insertion of promoter-DsRed constructs, 212-bp fragments (spanning target CREs and barcodes) were amplified from the barcoded CRE plasmid library and sequenced with 2x250 bp reads to assess target representation and proper CRE-barcode pairing (see Sequencing library preparation [barcoded CRE plasmid library]). Paired-end reads were merged with FLASH (v1.2) [190], and barcodes for

which >10% of reads could not be merged were removed from subsequent analysis (n=1286). Merged reads were aligned to designed oligos (including 100 bp of vector sequence on either side) with bowtie2 (v2.3.0) [144]. Barcodes for which >10% of reads were not aligned to the proper CRE were removed from subsequent analysis (n=4344). The mismatch rate at each position within library oligos was calculated with pysamstats (v0.24) (https://github.com/alimanfoo/pysamstats), and barcodes for oligos with >50% mismatch rate at any individual position were removed from subsequent analysis (n=1207). Taken together, these quality control steps flagged 6313 barcodes that were removed from analysis.

Pre-processing of previously generated datasets

DNase-seq, TF ChIP-seq, and Histone ChIP-seq data processing. See Supplemental Table 3.3 for a complete list of datasets used in the current study, including accessions and references. Single-end sequencing reads from DNase-seq of various mouse tissues were downloaded from ENCODE [4]. Single-end sequencing reads from CRX ChIP-seq (wild-type and *Nrl*$^{-/-}$ whole retina) were downloaded from the GEO (GSE20012) [68]. Single-end sequencing reads from NRL ChIP-seq (wild-type whole retina) were downloaded from the NEI [119]. Single-end sequencing reads from H3K27Ac, H3K4Me1, H3K4Me3, and H3K27Me3 ChIP-seq were downloaded from the GEO (GSE72550) [132]. Reads were aligned to mm10 with bowtie2 (v2.3.0) [144]. Alignments with mapping quality <30 or overlapping ENCODE blacklist regions [4] were removed with SAMtools (v1.5) [145]. Alignments were sorted and deduplicated with Picard (broadinstitute.github.io/picard/). Peaks were called with MACS2 (v2.1.1) [146] and annotated with HOMER (v4.9) [123]. For wild-type CRX ChIP-seq, dinucleotide frequencies, known motif enrichment, and motif densities were calculated with HOMER.

ATAC-seq data processing. Paired-end sequencing reads from rod and cone ATAC-seq were downloaded from the GEO (GSE83312) [178]. Reads were aligned to mm10 with bowtie2 (v2.3.0) [144], allowing fragment lengths up to 2 kb. Alignments with mapping quality <30 or overlapping ENCODE blacklist regions [4] were removed with SAMtools (v1.5) [145]. Alignments were sorted and deduplicated with Picard (broadinstitute.github.io/picard/), and alignments were filtered for nucleosome-free reads (read pairs with fragment length <150). Peaks were called with MACS2 (v2.1.1) [146] and annotated with HOMER (v4.9) [123].

RNA-seq data processing. Sequencing reads from transcriptome profiling of developing mouse rods and cones were downloaded from the GEO (GSE74660) [176, 177]. Transcript abundance (transcripts per million, or TPM) was estimated with kallisto (v0.43) [191] using the Ensembl gene model (GRCm38 assembly, release 79).

Models of TF occupancy and CRE-seq activity

Prediction of CRX-bound regions with logistic regression. 200-bp regions centered on TSS-distal (>1000 bp upstream and >100 bp downstream) CRX ChIP-seq peaks annotated as intergenic or intronic were lifted over to mm9 with HOMER (n=5250). FASTA files for these regions as well as a 10X set of background sequences were generated with the gkmSVM R package [168]. Features for each positive (CRX-bound) and negative (CRX-unbound) sequence were extracted as follows. The ten non-redundant dinucleotide frequencies were calculated over both strands for each sequence: 1) AA or TT, 2) AC or GT, 3) AG or CT, 4) AT, 5) CA or TG, 6) CC or GG, 7) CG, 8) GA or TC, 9) GC, and 10) TA. The AC or GT dinucleotide class was arbitrarily removed to eliminate linear dependency. In addition, instances of 843 TF binding sites [162] in each sequence were identified with FIMO (v4.11.2) [186] using the mononucleotide frequencies of negative sequences as a background model and a p-value threshold of $p < 10^{-2}$. The number of distinct TF binding sites was reduced by collapsing TF binding sites belonging to the same cluster

as defined by a recent analysis of 9650 PWMs [169], retaining the most prevalent TF binding site in each cluster as a representative member. Models were fit using TF binding site counts above a single threshold ($p < 10^{-2}$) as well as counts binned by match p-value (as a proxy for TF binding site affinity): high ($p < 10^{-5}$), medium ($10^{-5} < p < 10^{-4}$), low ($10^{-4} < p < 10^{-3}$), and very low ($10^{-3} < p < 10^{-2}$). Logistic regression models were fit with the R packages glmnet [189] and caret [192], and lasso regularization was used to control the complexity of models that included more than one feature. To promote sparse solutions, the largest regularization parameter (lambda) that yielded an AUC-ROC within 2% of the maximum AUC-ROC was selected. Model performance was evaluated by repeated 10-fold cross-validation (10 repeats), and ROC and PR curves were generated with the package PRROC [193, 194].

Prediction of CRX-bound regions with gkm-SVM. Positive (CRX-bound) and negative (CRX-unbound) regions were defined as described above. gkm-SVM models were trained with LS-GKM [173] using a word length of 11 with 7 informative positions. Model performance was evaluated by 10-fold cross-validation, and ROC and PR curves were generated with the PRROC package [193, 194].

Correlation between primary sequence features or chromatin features and CRE-seq activity. Wild-type CRE-seq constructs were scored for dinucleotide frequencies and TF binding sites as described above (using a single p-value threshold of $p < 10^{-3}$). TF binding sites present in fewer than 30 CREs were removed. For ATAC-seq, DNase-seq, and TF ChIP-seq datasets, normalized read depths in 100-bp windows centered on each CRE (calculated with HOMER (v4.9) [123] were used as feature scores. For histone ChIP-seq datasets, normalized read depths in 100-bp windows centered 180 bp downstream of each CRE (calculated with HOMER (v4.9) [123] were used as feature scores. For each CRE, the median expression across biological replicates was used as the

response variable. All variables were standardized, and separate linear models were fit for each feature (dinucleotide frequency class, TF binding site count, or chromatin feature).

Prediction of CRX-bound regions with high vs. low activity by logistic regression. The wild-type expression of each CRE was classified by its activity relative to the median. "Low" was defined as being within 1.2-fold of the median (197 elements on pRho, and 245 elements on pCrx). "High" was defined as being >3-fold over the median (81 elements on pRho, and 121 elements on pCrx). For each CRE, feature vectors were defined using either primary sequence features (dinucleotide frequencies and TF binding sites counts binned by match p-value) or chromatin features (normalized read depth from various epigenomic datasets as described above). Logistic regression models were fit with the R packages glmnet [189] and caret [192], and lasso regularization was used to control model complexity. Model performance was evaluated by repeated 10-fold cross-validation (10 repeats), and ROC and PR curves were generated with the PRROC package [193, 194].

Prediction of mutation effects from primary sequence features. Wild-type and mutant PWM scores were calculated for each CRE in the saturating mutagenesis analysis. Monomeric CRX binding sites (97) were scored with a monomeric PWM (PITX1_DBD), and dimeric CRX binding sites (98) were scored with a dimeric PWM (OTX2_DBD_1) [162]. For each mutation, deltaSVM scores were calculated using various gkm-SVM modes. For CRX (wild-type or *Nrl*$^{-/-}$ retina) and NRL (wild-type retina) ChIP-seq, gkm-SVM models were trained on all TSS-distal peaks annotated as intergenic or intronic [173]. For ATAC-seq and DNase-seq data, gkm-SVM models were trained using cell- and tissue-type-specific peaks identified by DESeq2 [130], again restricting peaks to TSS-distal elements annotated as intergenic or intronic. For H3K27Ac, H3K4Me1, H3K4Me3, and H3K27Me3 ChIP-seq data, photoreceptor ATAC-seq peaks were ranked by the normalized level of each histone mark in 200-bp windows 180 bp downstream of

peak summits, and gkm-SVM models were trained on the top 2000 peaks identified in this way for each histone mark. Linear regression models were trained using different combinations of features (changes in PWM score, deltaSVM scores, and gkm-SVM scores) using the R packages glmnet [189] and caret [192]. Lasso regularization was used to control the model complexity, and model performance was evaluated by repeated 10-fold cross-validation (10 repeats). Training and testing folds were partitioned such that CREs used for training were excluded from testing.

Conservation analysis

Conservation profiles centered on CRX binding sites. CRX binding sites were identified with FIMO (v4.11.2) [186] using the PWMs PITX1_DBD (monomeric) and OTX2_DBD_1 (dimeric) [162] at a p-value threshold of $p < 10^{-3}$. Peaks were centered on either monomeric or dimeric binding sites, and average conservation (phyloP scores derived from a multiple alignment of 100 vertebrate genomes to the mouse mm10 assembly) was calculated over these intervals using bedtools (v2.26) [156].

PCR primers

Oligo library amplification (170 bp):

```
>PCR_primer_1
GTAGCGTCTGTCCGTGTC

>PCR_primer_2
CTGTAGTAGTAGTTGGCGGC
```

Barcoded CRE library sequencing amplicon (212 bp):

```
>CRE-bc_F
TAAACAAATAGGGGTTCCGCGCACA

>CRE-bc_R
GATAGGCAGCCTGCACCTGAGGAGT
```

Illumina adapters (annealed):

```
>adapter_oligo1
/5Phos/GATCGGAAGAGCACACGTCT
```

```
>adapter_oligo2
ACACTCTTTCCCTACACGACGCTCTTCCGATC*T
```

Illumina library amplification and indexing (adapters added by ligation):

```
>Multiplex_PCR_primer_1.0
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

>SIC_index_NNNN
CAAGCAGAAGACGGCATACGAGATNNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGA
                         *********
                         9_bp_index
```

pRho-DsRed and *pCrx-DsRed* subcloning:

```
>SpeI_pRho
TAGCTACTAGTCTAGAATGTCACCTTGGCCCCTCT

>SpeI_pCrx
CTGACTAGTCCTGGTTGCAGGCAGGAGTTGGGCTT

>KpnI_DsRed
ATTAGGTACCCTACAGGAACAGGTGGTGGCGG
```

CRE-seq sequencing amplicon (197 bp):

```
>read1_bc
ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATAGGCAGCCTGCACCTGAGGAGT

>read2_DsRed
AGACGTGTGCTCTTCCGATCTGTCCATCTACATGGCCAAGAAGCCC
```

3.11.2 Supplemental tables

Supplemental Table S3.1. Performance of models predicting CRX occupancy in mouse photoreceptors. URL:

https://genome.cshlp.org/content/suppl/2018/09/14/gr.231886.117.DC1/supplemental_table_1.xlsx

Supplemental Table S3.2. Feature weights for models predicting CRX occupancy in mouse photoreceptors. URL:

https://genome.cshlp.org/content/suppl/2018/09/14/gr.231886.117.DC1/supplemental_table_2.xlsx

Supplemental Table S3.3. Summary of datasets used in the current study. URL:

https://genome.cshlp.org/content/suppl/2018/09/14/gr.231886.117.DC1/supplemental_table_3.xlsx

Supplemental Table S3.4. Primary sequence features and chromatin features significantly correlated with wild-type CRE activity. URL:

https://genome.cshlp.org/content/suppl/2018/09/14/gr.231886.117.DC1/supplemental_table_4.xlsx

Supplemental Table S3.5. Feature weights for logistic regression models predicting wild-type CRE activity. URL:

https://genome.cshlp.org/content/suppl/2018/09/14/gr.231886.117.DC1/supplemental_table_5.xlsx

Supplemental Table S3.6. Monomeric and dimeric PWM scores and mutant CRE activity for 1756 inactivating mutations in CRX binding site. URL:

https://genome.cshlp.org/content/suppl/2018/09/14/gr.231886.117.DC1/supplemental_table_6.xlsx

Supplemental Table S3.7. Linear modeling of interactions between pairs of CRX binding sites. URL:

https://genome.cshlp.org/content/suppl/2018/09/14/gr.231886.117.DC1/supplemental_table_7.xlsx

Supplemental Table S3.8. Linear modeling of interactions between half-sites within dimeric CRX binding sites. URL:

https://genome.cshlp.org/content/suppl/2018/09/14/gr.231886.117.DC1/supplemental_table_8.xlsx

Supplemental Table S3.9. ΔPWM scores, gkm-SVM scores, deltaSVM scores, and CRE-seq expression values for saturating mutagenesis of 195 CRX binding sites. URL:

https://genome.cshlp.org/content/suppl/2018/09/14/gr.231886.117.DC1/supplemental_table_9.xlsx

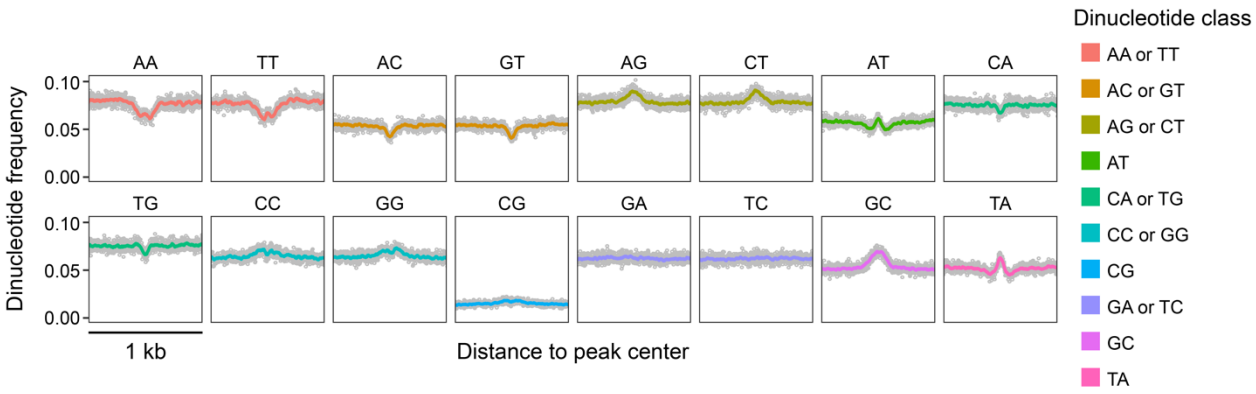
Supplemental Table S3.10. Performance of linear regression models predicting the effects of mutations in CRX binding sites. URL:

https://genome.cshlp.org/content/suppl/2018/09/14/gr.231886.117.DC1/supplemental_table_10.xlsx

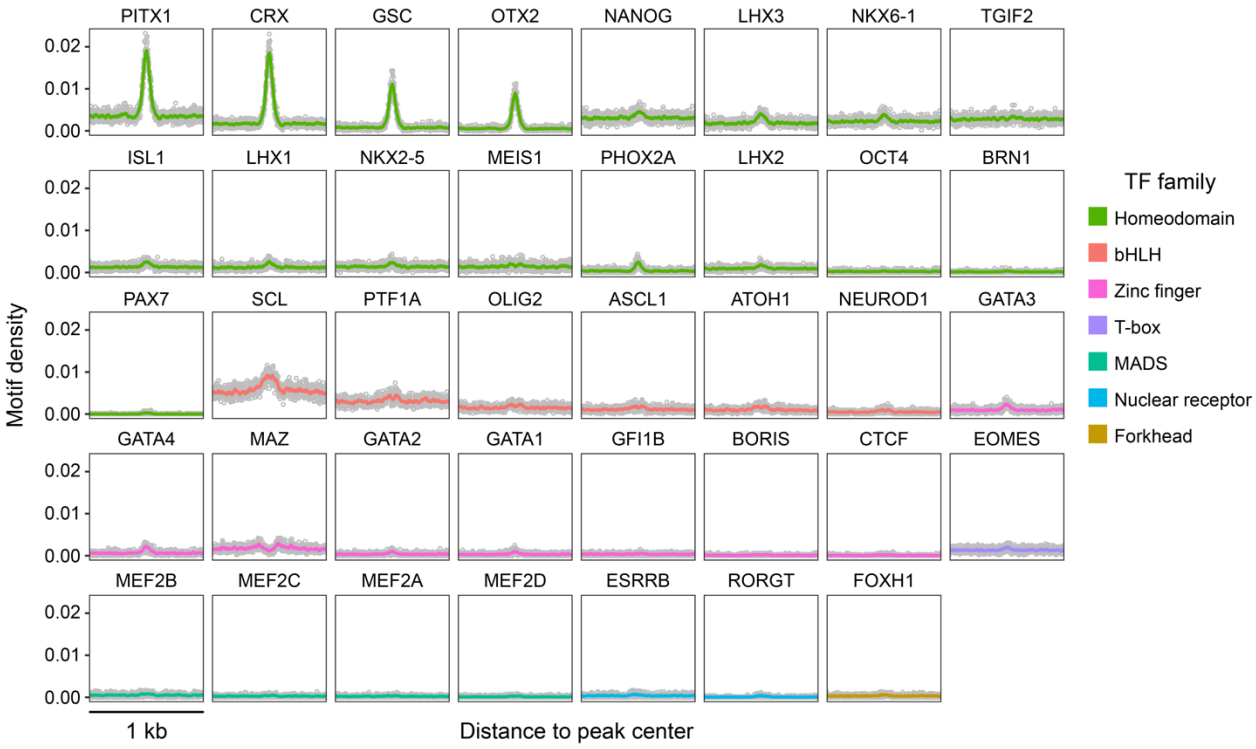
Supplemental Table S3.11. Feature weights for linear regression models predicting the effects of mutations in CRX binding sites. URL:

https://genome.cshlp.org/content/suppl/2018/09/14/gr.231886.117.DC1/supplemental_table_11.xlsx

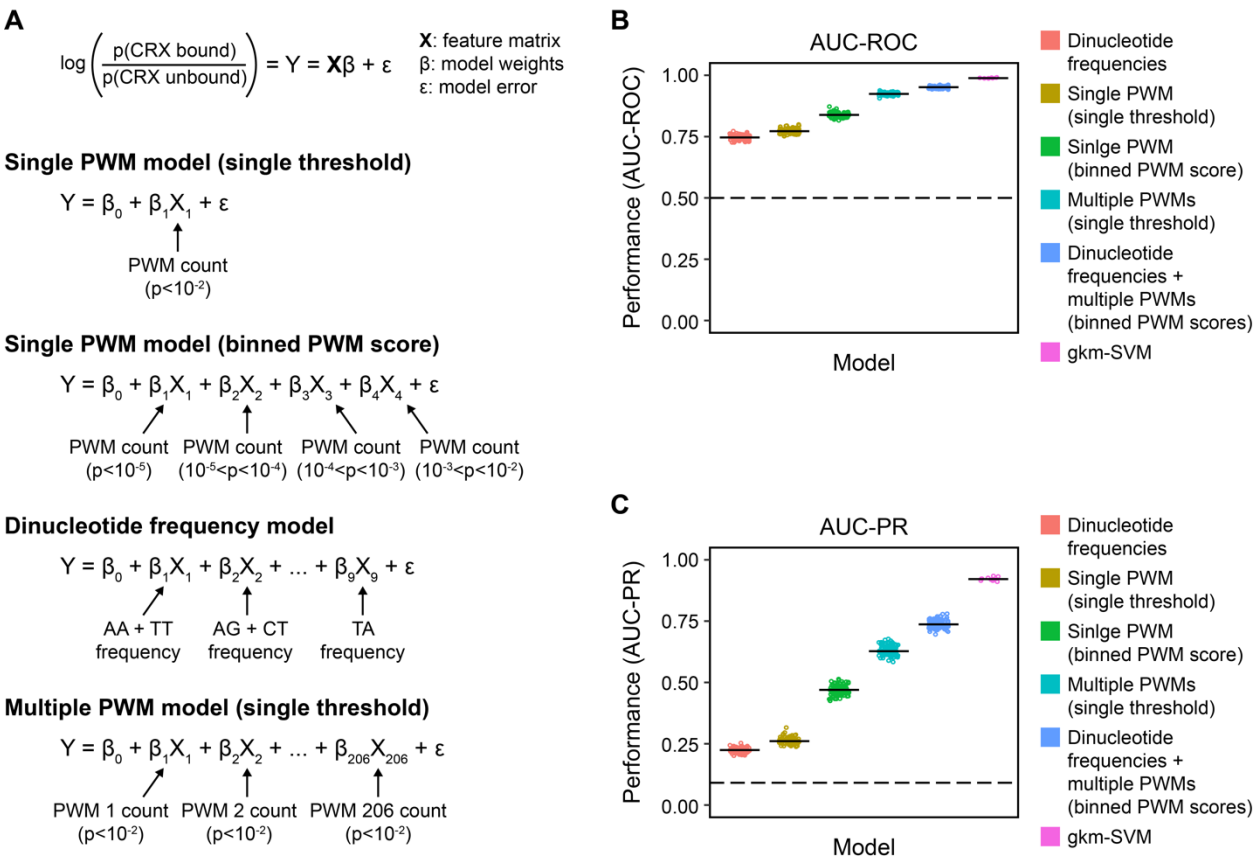
3.11.3 Supplemental figures



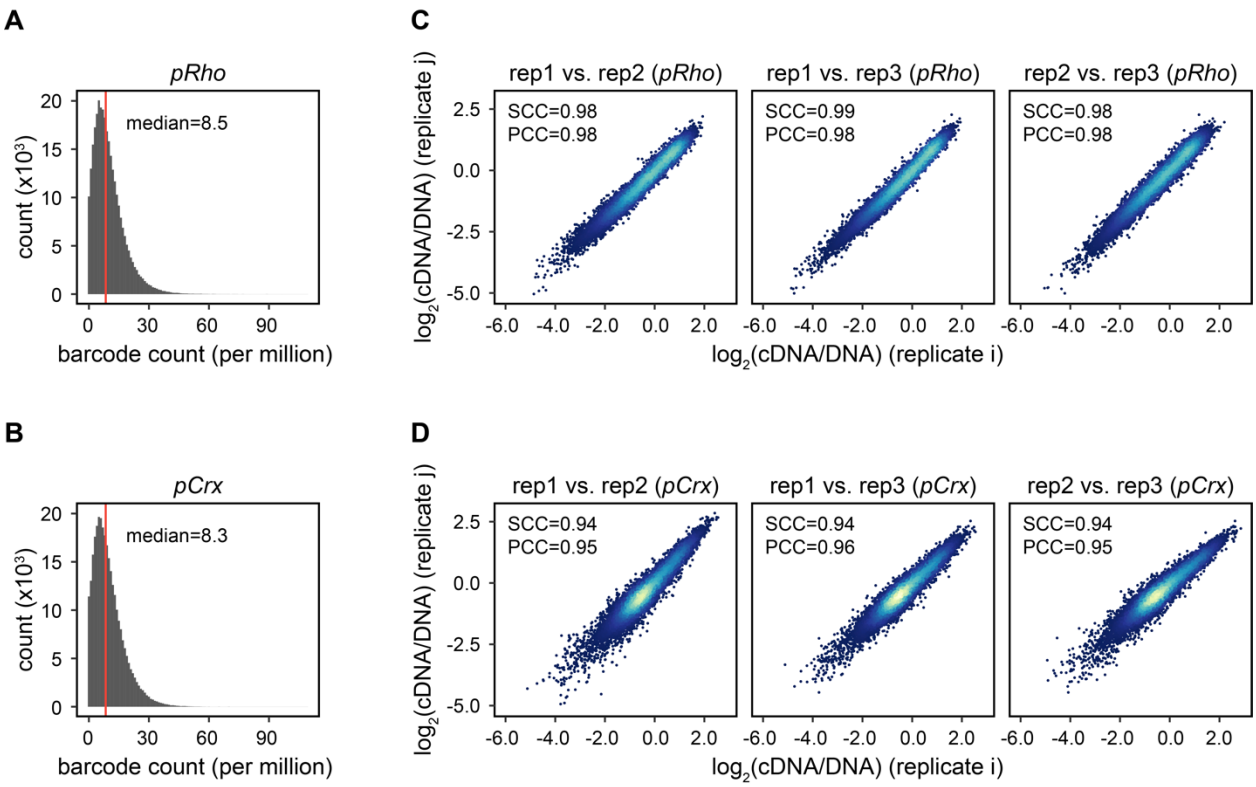
Supplemental Figure S3.1. Dinucleotide profiles of CRX bound regions. Average dinucleotide frequencies per base pair per peak (gray dots) with a 25-bp rolling mean (colored lines) in a 1 kb window centered on TSS-distal (>1 kb upstream or >100 bp downstream of an annotated TSS) CRX ChIP-seq peaks (n=5250).



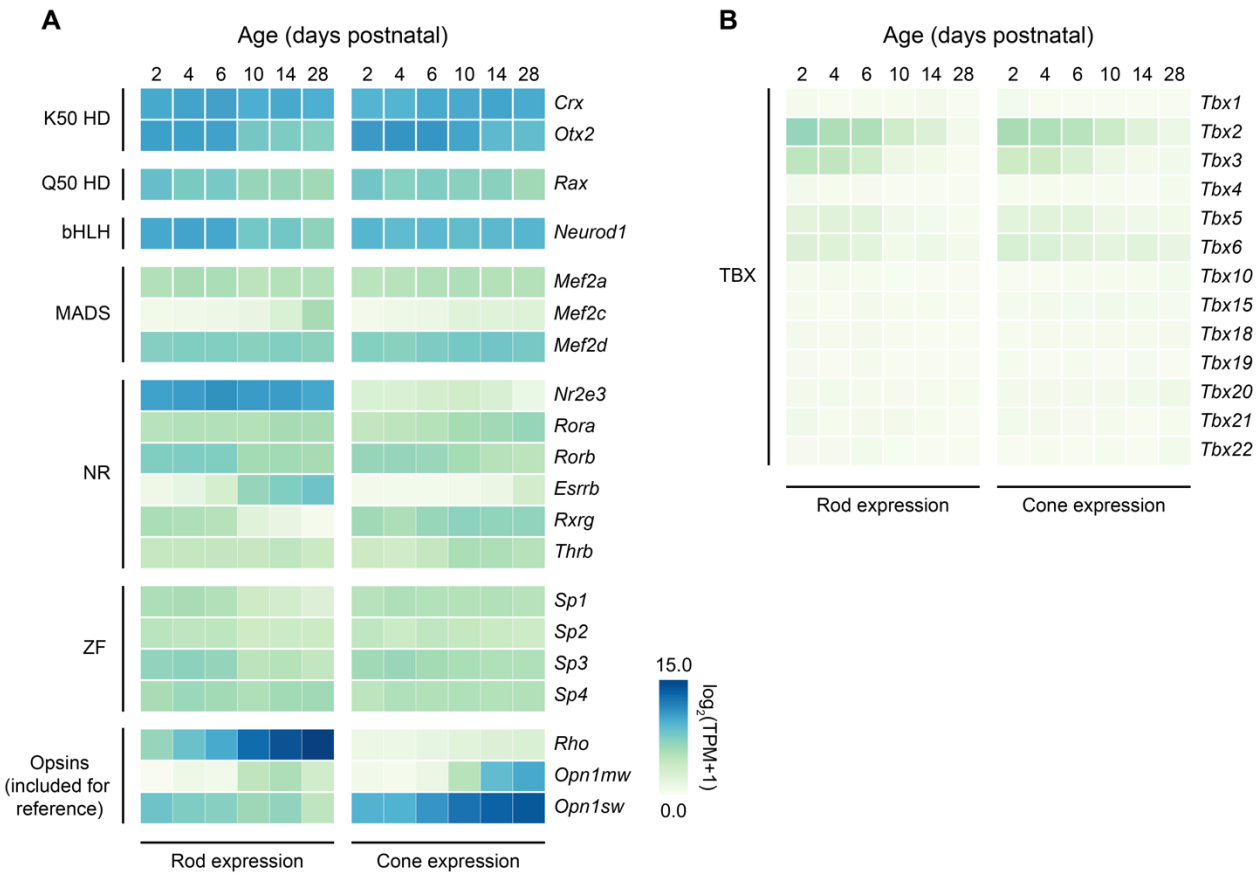
Supplemental Figure S3.2. TF binding site density of CRX bound regions. Average number of TF binding sites per base pair per peak (gray dots) with a 25-bp rolling mean (colored lines) in a 1 kb window centered on TSS-distal (>1 kb upstream or >100 bp downstream of an annotated TSS) CRX ChIP-seq peaks (n=5250). Motif enrichment was calculated for 319 known motifs curated by the HOMER suite of sequence analysis tools [123]. TF binding site densities are shown for motifs with enrichment p-values less than 10^{-10}.



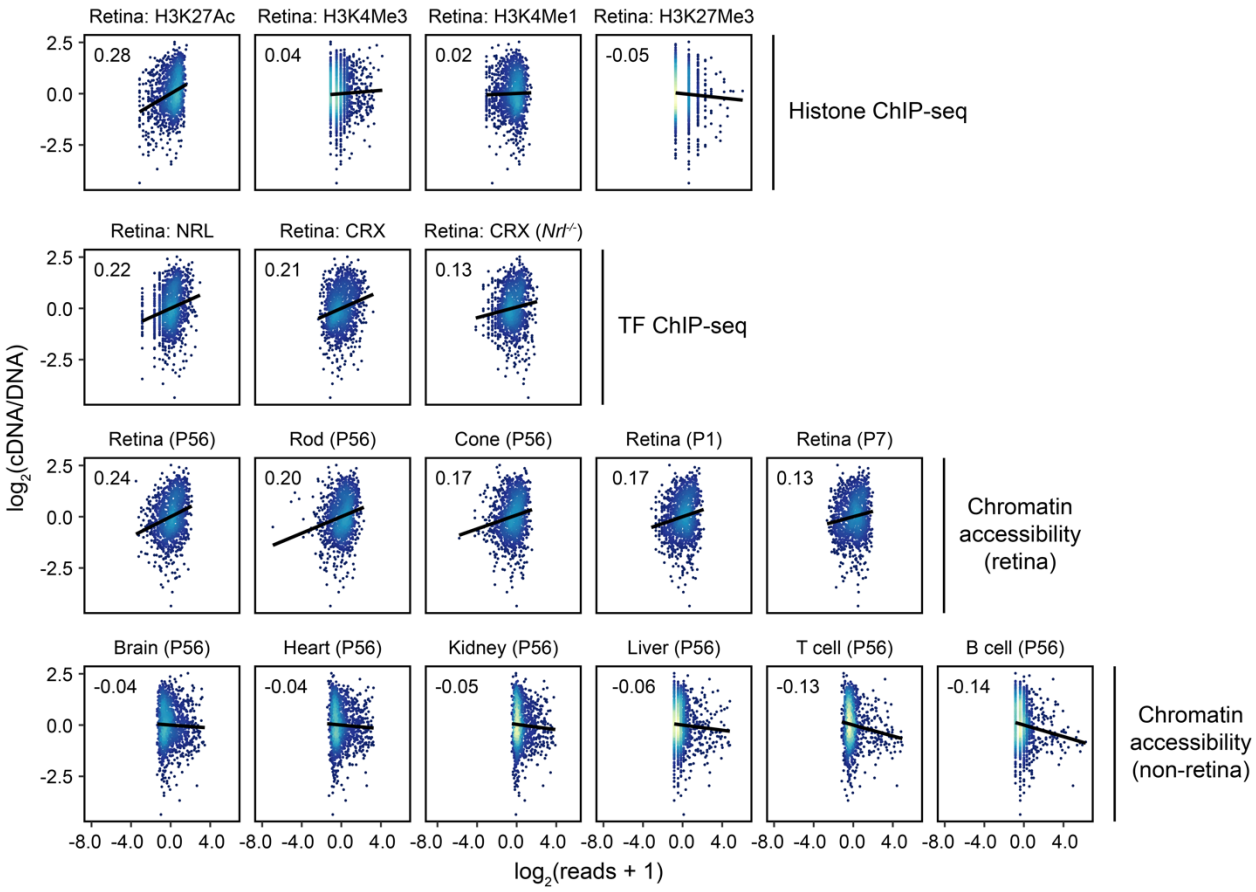
Supplemental Figure S3.3. Prediction of CRX-bound regions from primary sequence features. (A) Overview of logistic regression models using primary sequence features to classify CRX-bound (n=5250) vs. CRX-unbound (n=52500) regions. (B) Performance of specific models as measured by area under the receiver operating characteristic curve (AUC-ROC). Dashed line: performance of a random classifier (ROC-AUC=0.50) (C) Performance of specific models as measured by area under the precision recall curve (AUC-PR). Dashed line: performance of a random classifier (AUC-PR=0.09, the positive class rate). In (B-C), individual points indicate the performance of models estimated from different folds of repeated 10-fold cross-validation (logistic regression models) or 10-fold cross-validation (gkm-SVM), and horizontal bars represent the median cross-validated performance of each model.



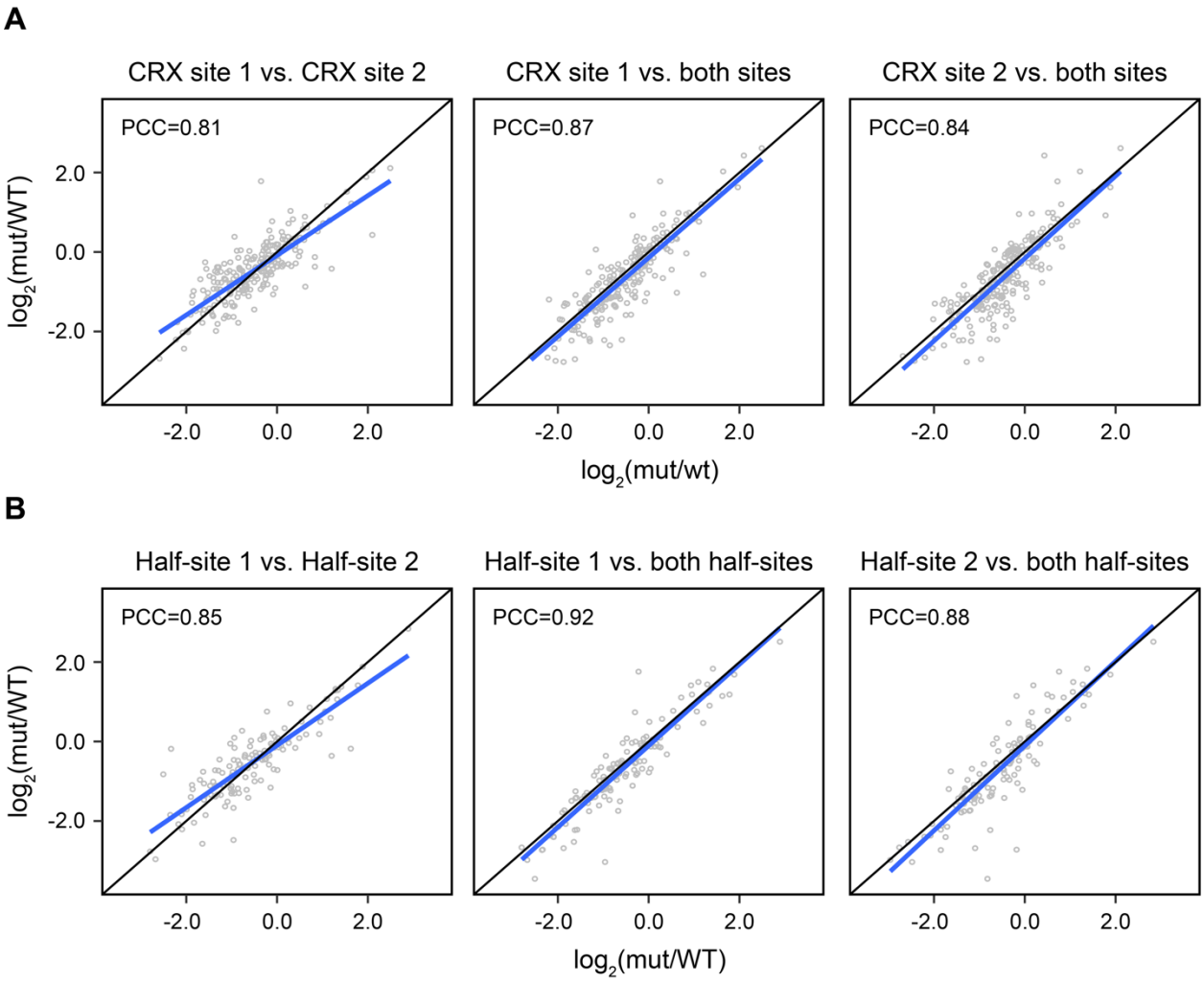
Supplemental Figure S3.4. CRE-seq library complexity and reproducibility. (A-B) Histogram of average barcode counts (per million barcodes) from CRE-seq DNA libraries (three replicates each for *pRho* and *pCrx*). (C-D) Scatter plots of CRE-seq activity for pairs of biological replicates. Points are colored by density. PCC: Pearson correlation coefficient. SCC: Spearman correlation coefficient.



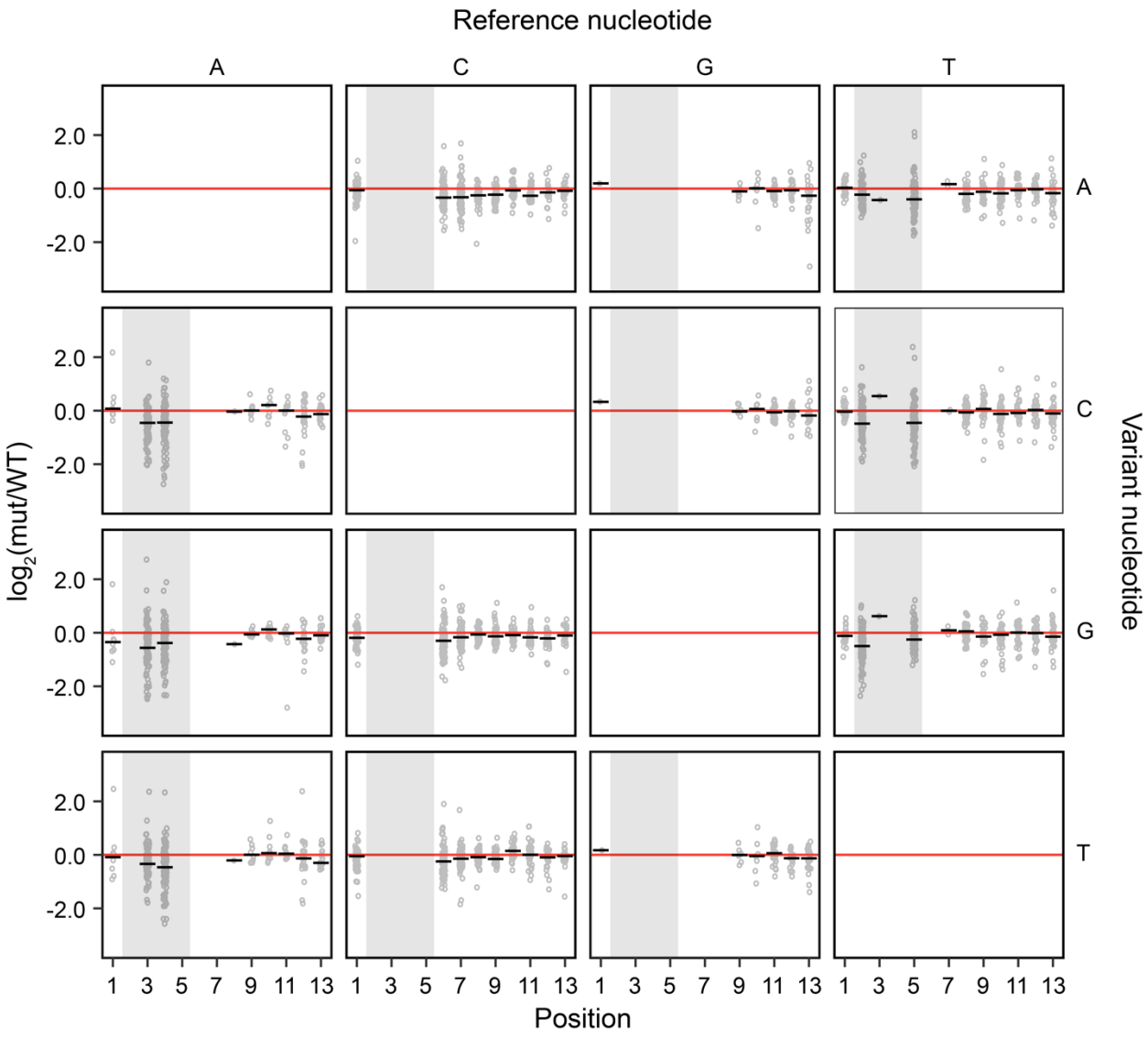
Supplemental Figure S3.5. Expression of selected TFs during photoreceptor development. (A) Heatmap of developmental expression for selected transcription factors corresponding to TF binding site families correlated with activity (see Fig. 3.3D) [176, 177]. Rows correspond to specific TFs, and columns correspond to developmental age. Separate panels are included for rod and cone photoreceptor expression. Opsins are highly expressed rod-specific (*Rho*) and cone-specific (*Opn1mw* and *Opn1sw*) genes (not transcription factors) included for reference. K50 HD: K50 homeodomain. Q50 HD: Q50 homeodomain. bHLH: basic helix-loop-helix. NR: nuclear receptor. ZF: Zinc finger. (B) Developmental expression profiles of all T-box family members in mouse.



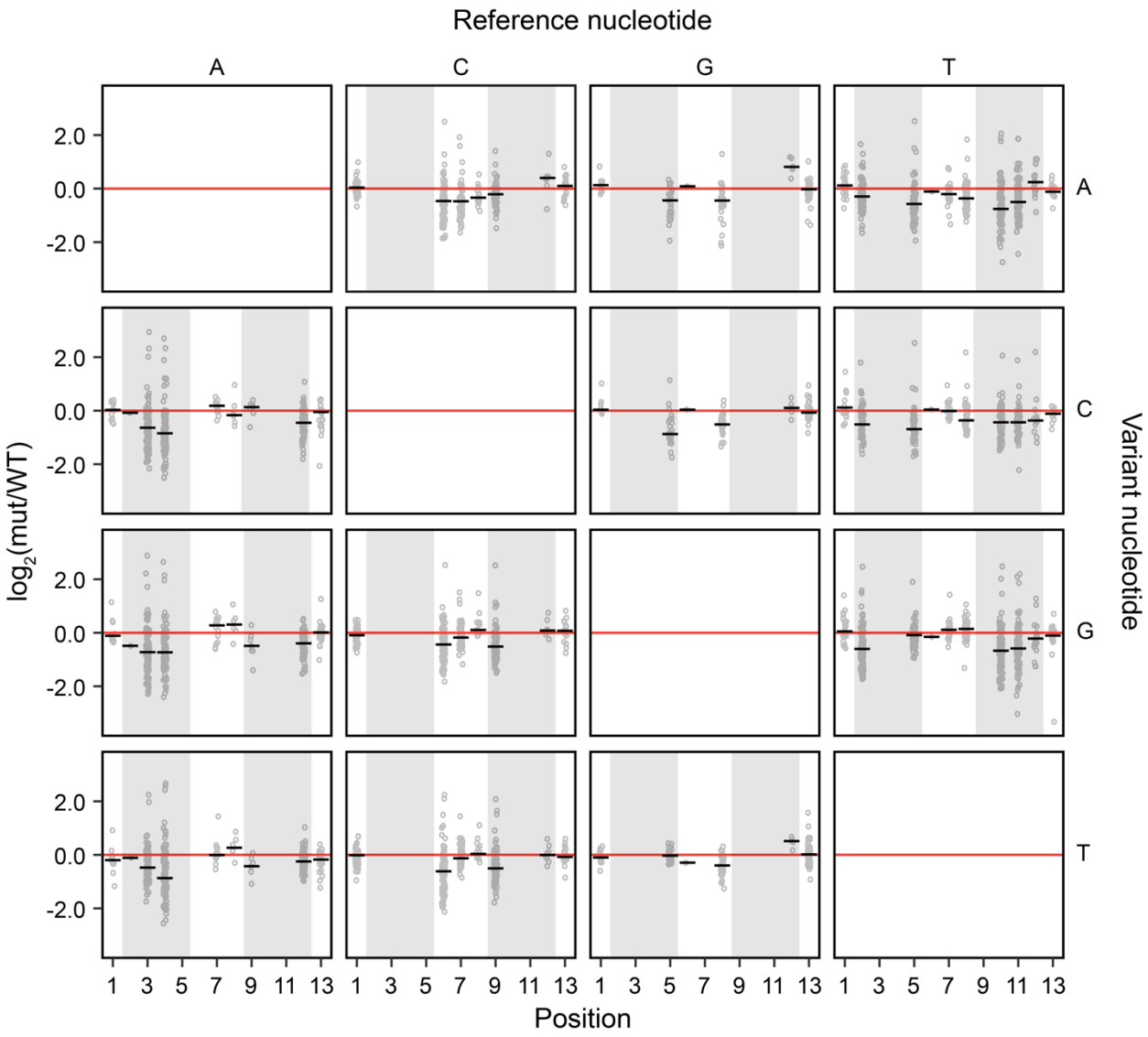
Supplemental Figure S3.6. Correlation between individual chromatin features and CRE-seq activity. Scatter plots showing signal strength [$\log_2$(reads+1)] vs. CRE-seq activity (assayed on *pCrx*) for various epigenomic datasets. Lines show a linear fit for each scatter plot, and Pearson correlation coefficients are included in the upper left of each panel.



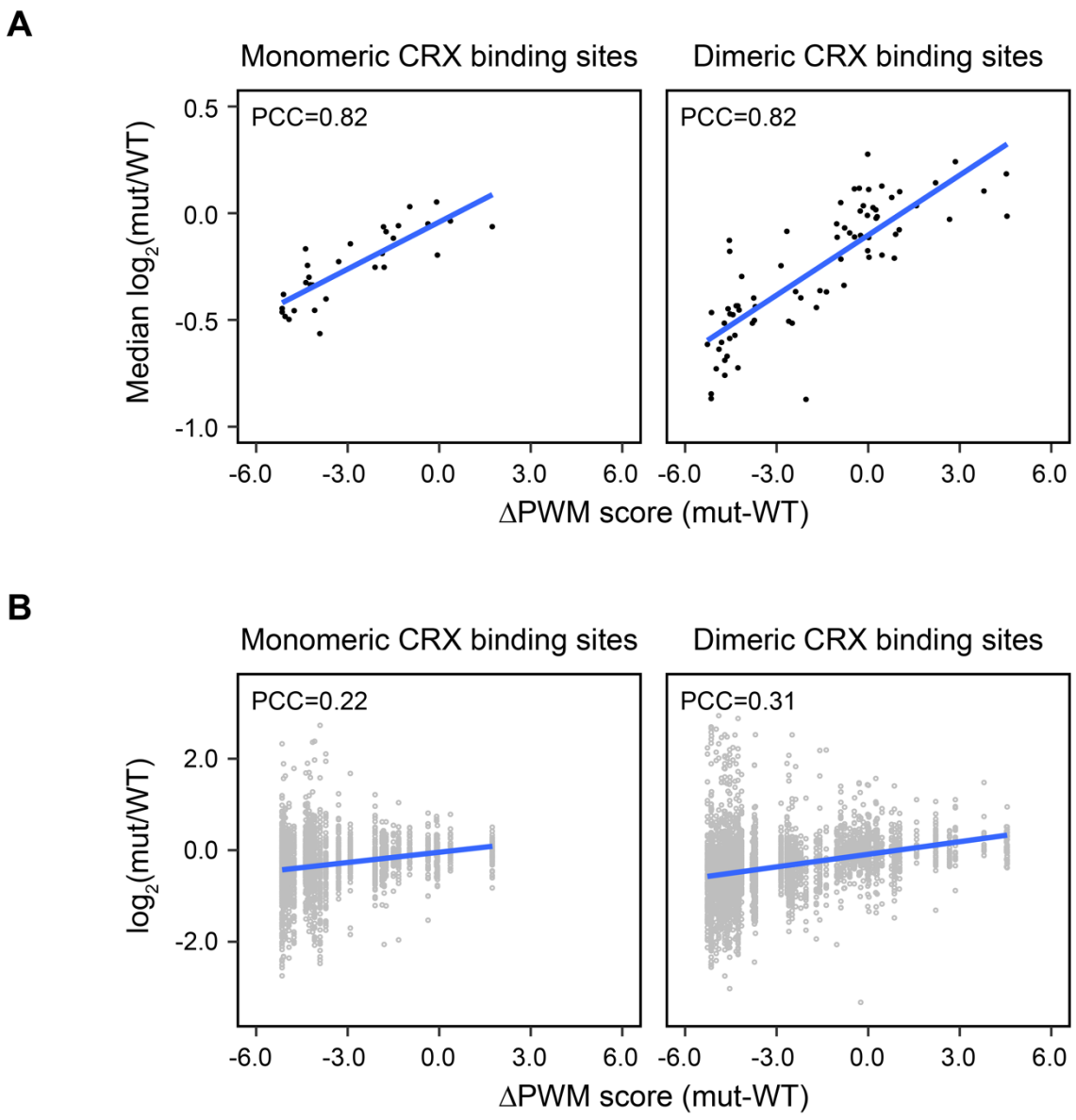
Supplemental Figure S3.7. Effect of single- and double-mutants within the same CRE. (A) Scatter plots showing the effect of mutating pairs of CRX binding sites individually or in combination for CREs with two predicted CRX binding sites. (B) Scatter plots showing the effect of mutating either half site of dimeric CRX binding sites individually or in combination. Black lines: identity lines. Blue lines: linear fits. PCC: Pearson correlation coefficient.



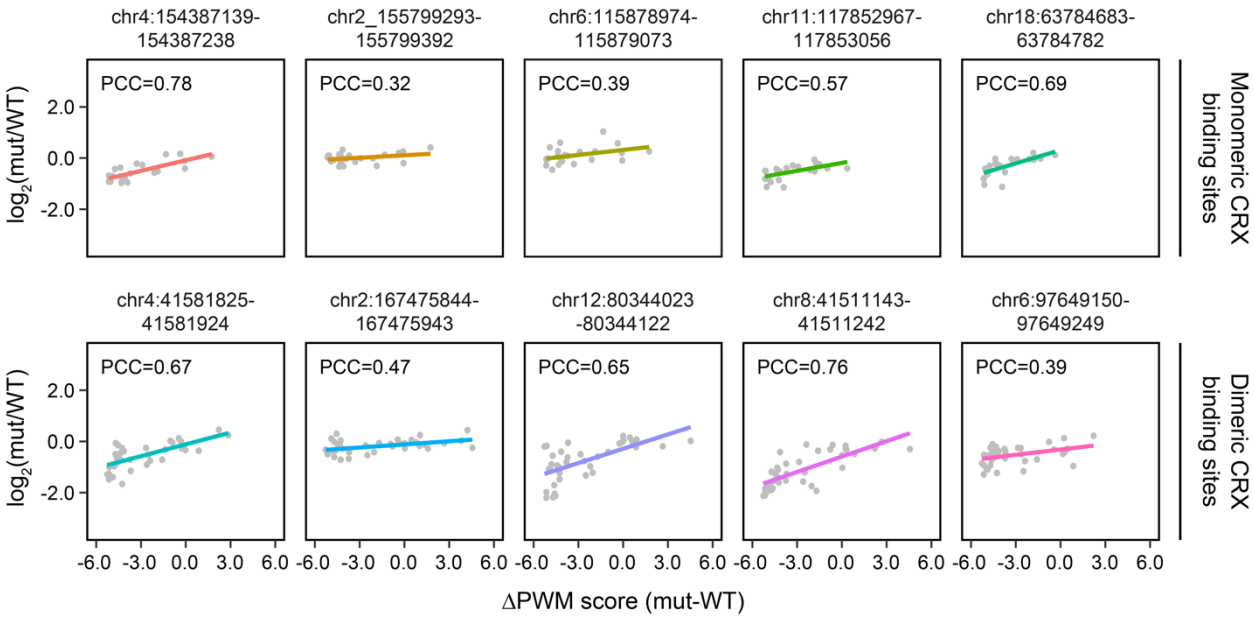
Supplemental Figure S3.8. Dense substitution analysis of monomeric CRX binding sites. Scatter plots of the effects of specific substitutions at specific positions (gray points) within monomeric CRX binding sites. Separate panels are included for each of the 12 possible reference to variant substitutions at each position. For example, the first column of panels shows the effects of mutating a reference A at each position to a C (row 2), G (row 3), or T (row 4). Horizontal bars: the median effects of each substitution at each position. Red lines: no effect (log fold change equals zero). Gray boxes: TAAT core positions.



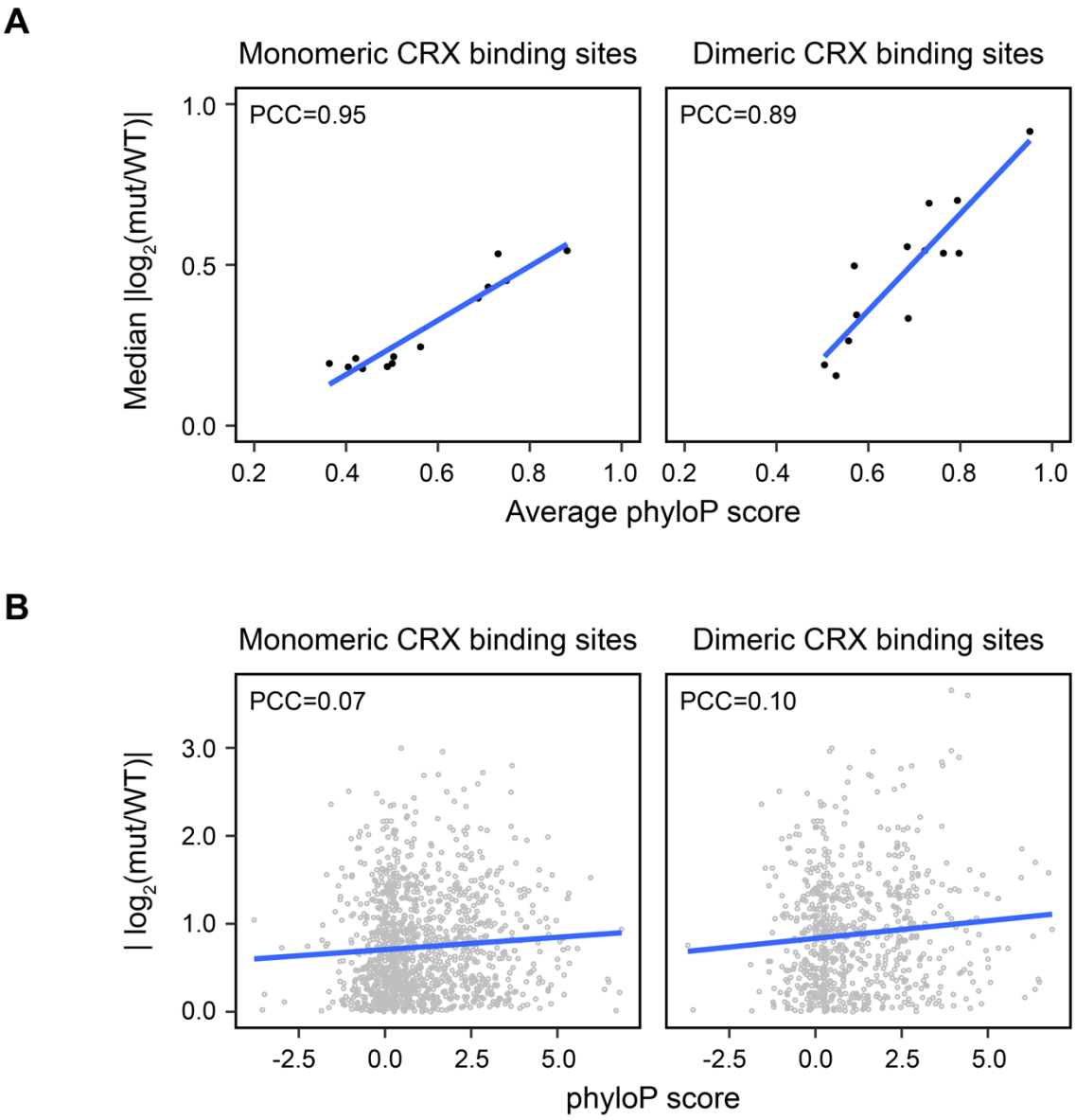
Supplemental Figure S3.9. Dense substitution analysis of dimeric CRX binding sites. Scatter plots of the effects of specific substitutions at specific positions (gray points) within dimeric CRX binding sites. Separate panels are included for each of the 12 possible reference to variant substitutions at each position. For example, the first column of panels shows the effects of mutating a reference A at each position to a C (row 2), G (row 3), or T (row 4). Horizontal bars: the median effects of each substitution at each position. Red lines: no effect (log fold change equals zero). Gray boxes: TAAT core positions.



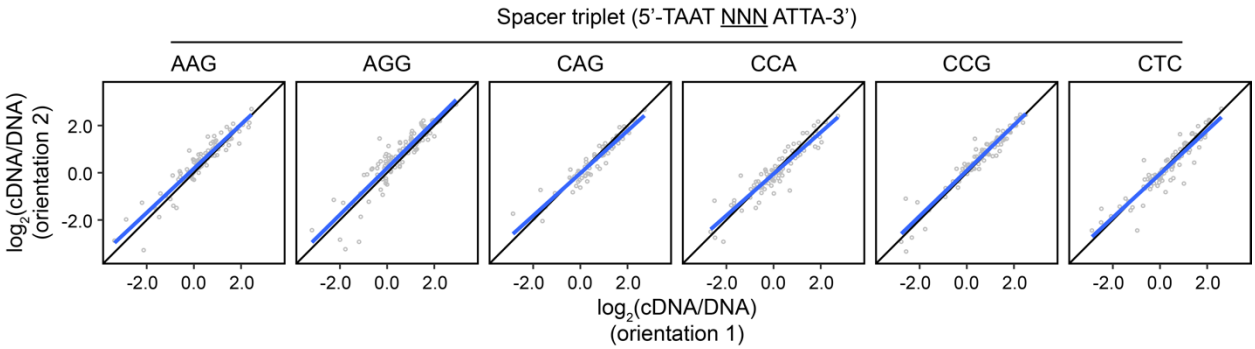
Supplemental Figure S3.10. Aggregate correlation between change in affinity and change in CRE-seq activity. (A) Scatter plot of change in PWM score vs. median change in activity as measured by CRE-seq for monomeric (left panel) and dimeric (right panel) CRX binding sites. Aggregating across CREs, changes in affinity are strongly correlated with changes in activity. Blue lines: linear fits. (B) Scatter plot of change in PWM score vs. change in activity as measured by CRE-seq for monomeric (left panel) and dimeric (right panel) CRX binding sites. Each point represents an individual mutation. Blue lines: linear fits. PCC: Pearson correlation coefficient.



Supplemental Figure S3.11. CRE-level correlation between change in affinity and change in CRE-seq activity. Scatter plots of change in PWM score (TF binding site affinity) vs. change in activity (as measured by CRE-seq) for a random sample of five monomeric and five dimeric CRX binding sites. Whereas the correlation between affinity and activity is low across CREs (see Supplemental Fig. S3.10), correlations are much higher when considering mutations within individual CREs. PCC: Pearson correlation coefficient.



Supplemental Figure S3.12. Correlation between phylogenetic conservation and change in CRE-seq activity. (A) Scatter plot of average conservation (60-way vertebrate phyloP scores) vs. median absolute change in CRE-seq activity (aggregated by position) for monomeric (left panel) and dimeric (right panel) CRX binding sites. (B) Scatter plot of conservation (60-way vertebrate phyloP scores) vs. absolute change in CRE-seq activity for individual mutations. PCC: Pearson correlation coefficient.



Supplemental Figure S3.13. Effect of spacer orientation on CRE-seq activity. Scatter plots comparing the activity of CREs with selected spacer triplets in the forward (x-axis) vs. reverse (y-axis) orientation. Panel labels indicate the forward spacer orientation (with the reverse complement being the reverse spacer orientation). The orientation of each motif was determined by the orientation with the highest-scoring match to a dimeric homeodomain PWM (OTX2_DBD_1) (Jolma et al. 2013).

Chapter 4: Summary and future directions

4.1 Summary

Understanding how regulatory variation impacts the evolution of complex traits and human disease requires a detailed understanding of how *cis*-regulatory elements (CREs) encode information in primary sequence. In this thesis, I define the *cis*-regulatory architecture of individual mouse retinal cell types (rod and cone photoreceptors), and I systematically identify sequence features that influence the activity of mouse photoreceptor CREs *in vivo*. In this chapter, I summarize the main results from these studies, and I discuss potential directions for future research.

4.1.1 Genome-wide identification and characterization of rod- and cone-specific CREs

In Chapter 2, I present genome-wide open chromatin maps from FACS-sorted rods, blue cones (*Nrl*$^{-/-}$ photoreceptors), and green cones, along with transcriptome profiles of rods and blue cones. Globally, the open chromatin profiles of rods and cones are highly similar, and those of blue and green cones are nearly indistinguishable. Nevertheless, thousands of loci are selectively closed in rods compared to both cone types (as well as >60 control cell types and tissues), and I show that this uniquely closed chromatin architecture depends on the rod-specific transcription factor (TF) *Nrl*. Furthermore, regions that are selectively closed in rods frequently cluster in megabase-scale domains, suggesting that their closure reflects changes in higher-order genome organization. In addition, I show that rod- and cone-specific genes are typically flanked by complex arrangements of shared and cell-type-specific open chromatin peaks, and that changes in the accessibility of individual CREs are only moderately correlated with changes in the expression of nearby genes. Finally, I identify enrichments of TF binding sites in rod- and cone-specific open chromatin, providing a basis for functional studies of cell-type-specific *cis*-regulatory grammar.

Prior to this study, retinal CREs had generally been mapped by whole-retina DNase-seq [118] or TF ChIP-seq [68, 73, 90], which are either not specific or not comprehensive with respect to photoreceptors. Maps of cell-type-specific CREs offer a blueprint for studying the regulation of individual genes, selecting and optimizing CREs for gene therapy, and implementing strategies for directed differentiation and cellular reprogramming. In addition, my results demonstrate that the so-called 'inverted' nuclear architecture of mouse rods is correlated with changes in chromatin structure at the level of individual CREs. This observation suggests that the mouse rod photoreceptor represents an informative model for studying the relationship between genome organization, chromatin state, and gene regulation.

4.1.2 Functional analysis of photoreceptor CREs

In Chapter 3, I quantify the activity of thousands of wild-type and mutant CREs to determine how sequence context shapes the activity of binding sites for CRX, a homeodomain TF and a master regulator of photoreceptor gene expression. My results show that dimeric CRX binding sites encode stronger enhancers than monomeric CRX binding sites, and that multiple instances of CRX binding sites within individual CREs act cooperatively. In addition, the activity of half-sites within dimeric TF binding sites is cooperative, depends on a three-nucleotide spacing, and is sensitive to the identity of the spacer nucleotides. Within individual CREs, I find that changes in CRX binding site affinity are generally correlated with changes in CRE activity. Nevertheless, the strength of this relationship is CRE-dependent, and the correlation between TF binding site affinity and activity is obscured when considering mutations across multiple CREs. Finally, CRE activity is more strongly correlated with the number and affinity of E-box, nuclear receptor, Q50 homeodomain, and T-box binding sites than with the number and affinity of CRX binding sites. Taken together, these results demonstrate that activity of CRX binding sites depends on multiple layers of sequence context. Moreover, they show that quantitative models of enhancer activity

need to account for the affinity of individual TF binding sites as well as interactions between binding sites.

4.2 Implications for future research

4.2.1 Molecular characterization of rod chromatin architecture and its regulation

ATAC-seq reveals dramatic alterations in rod chromatin accessibility, but additional experiments are needed to understand these changes in the context of higher-order genome organization. For example, Hi-C could be used to determine if regions that are selectively closed in rods correspond to well-defined genomic compartments, as well as the degree to which rod chromatin closure preserves or disrupts the genome organization of cones [27]. In addition, ChIP-seq for covalent histone modifications (e.g., H3K27Ac, H3K27Me3, H3K4Me3, H3K4Me1, and H3K9Me3) would reveal the extent to which changes in chromatin accessibility are associated with changes in regulatory state. Furthermore, at the level of individual cells, fluorescent *in situ* hybridization (FISH) could be used to compare how specific regions of the genome are localized within rod and cone nuclei. Taken together, these approaches would provide insight into how rod chromatin closure relates to three-dimensional nuclear organization.

In addition to further characterization of rod nuclear architecture, more work is needed to understand how rod-specific chromatin changes are regulated. I show that rod chromatin closure depends on *Nrl*, and previous work has shown that downregulation of both *Lmna* and *Lbr* is sufficient for inverted nuclear organization [106]. However, it remains to be shown if *Nrl* directly represses *Lmna* and *Lbr*, or if intermediate regulators are involved. In addition, *Lmna* and *Lbr* mediate interactions with the nuclear periphery, but factors that directly modulate rod chromatin accessibility have not been identified (e.g., histone modifiers, histone remodeling complexes, and histone variants). Furthermore, the developmental timing of rod chromatin closure has not been

established, but could be determined by an ATAC-seq time series. Coupled with gene expression profiling, these data would have the potential to identify additional regulators that mediate rod-specific changes in chromatin accessibility. In addition, this analysis would establish the temporal relationship between rod chromatin closure and the development of inverted nuclear architecture (which is gradual and not complete until adult stages) [100].

4.2.2 Integrated analysis of *cis*- and *trans*-acting factors in photoreceptor CREs

In addition to revealing global changes in chromatin architecture, my results show that photoreceptor ATAC-seq is a powerful method for mapping rod- and cone-specific CREs. Nevertheless, identifying the TFs that regulate individual CREs remains challenging. ChIP-seq maps TF occupancy genome-wide, but demonstrating that TFs bind a particular CRE is not sufficient to demonstrate that they contribute to CRE function. Thus, while ChIP-seq experiments for additional photoreceptor TFs would be useful (e.g., RAX, RORB, NR2E3, ESRRB, THRB, and RXRG), complementary approaches are needed.

A functional approach to identify TFs that regulate individual CREs would be to combine CRE-seq with targeted knockdown (or deletion) of photoreceptor TFs. For example, the same CRE-seq library described in Chapter 3 could be assayed in retinas from $Nrl^{flox/flox}$ mice co-electroporated with Cre recombinase (or control) to identify CREs that are enhanced or repressed by NRL. Alternatively, these experiments could be performed with RNAi-mediated TF knockdown, or CRISPR/Cas9-mediated TF knockout, to target a broader set of TFs than those for which floxed alleles are available. In addition, by assaying native CREs as well as CREs with mutations in TF binding sites, this approach could identify the sequences that mediate the effects of TFs at single-nucleotide resolution. Importantly, as in most RNA-seq analysis, CRE-seq normalization relies on the assumption that the total level of transcriptional activity is constant across experimental conditions. This assumption may not be valid when assaying photoreceptor

CREs while simultaneously knocking down photoreceptor TFs. Therefore, orthogonal standards (e.g., a diverse set of ubiquitous CREs with a broad range of activity) will be needed to accurately quantify CRE activity across experimental conditions.

4.2.3 Developing more complete models of photoreceptor *cis*-regulatory grammar

CRE-seq has facilitated detailed functional analyses of mutations in CRX binding sites, but additional studies are needed to quantify the contribution of non-CRX binding sites. In particular, my results suggest that E-box, nuclear receptor, Q50 homeodomain, and T-box TF binding sites play significant roles in determining photoreceptor CRE activity, but validating these predictions requires comprehensive analysis of TF binding site mutations (e.g., by CRE-seq). Alternatively (or in parallel), an analysis of sparsely mutated photoreceptor CREs by CRE-seq has the potential to identify functional TF binding sites in an efficient and unbiased manner. For example, a CRE-seq library could be synthesized consisting of ~500 100-bp CREs in which a 10-bp sliding window with a 5-bp step was scrambled (yielding 20 constructs per target sequence). Assaying this library in photoreceptors could then identify a comprehensive set of functional TF binding sites and yield quantitative estimates of their relative activities.

As described above, CRE-seq demonstrates that pairs of CRX binding sites often act cooperatively. Therefore, once TF binding sites that modulate photoreceptor CRE activity have been comprehensively defined, mutating them individually and in combination will be necessary to determine if TF binding site cooperativity is a general phenomenon, or if it is restricted to specific pairs of binding sites. Furthermore, comparison of TF binding site instances that do and do not show evidence of cooperativity could be analyzed to identify patterns of relative spacing and orientation that are associated with non-additive interactions. Taken together, these approaches

have the potential to provide a more complete picture of how specific combinations of TFs encode regulatory activity in photoreceptor CREs.

4.3 Conclusion

In this thesis, I present detailed molecular and computational studies of the transcriptional mechanisms underlying the development and maintenance of mouse photoreceptors. In particular, I demonstrate that integrated epigenomic and transcriptomic profiling of mouse rods and cones is a powerful approach for defining the *cis*-regulatory architecture of individual retinal cell types. In addition, I perform comprehensive functional studies, which reveal that the regulatory activity of homeodomain binding sites is highly dependent on the broader sequence context in which these motifs occur. Finally, I suggest how these approaches can be adapted to study the functional impact of regulatory variation in humans (see Appendix). Thus, these studies provide detailed insights into the mechanisms of photoreceptor CRE function as well as a framework for studying the *cis*-regulatory architecture of additional retinal cell types.

Works Cited

1. Shlyueva, D., G. Stampfel, and A. Stark, *Transcriptional enhancers: from properties to genome-wide predictions.* Nat Rev Genet, 2014. **15**(4): p. 272-86.

2. Albert, F.W. and L. Kruglyak, *The role of regulatory variation in complex traits and disease.* Nat Rev Genet, 2015. **16**(4): p. 197-212.

3. Barrera, L.A., et al., *Survey of variation in human transcription factors reveals prevalent DNA binding changes.* Science, 2016. **351**(6280): p. 1450-1454.

4. Encode Project Consortium, *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

5. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues.* Nature, 2014. **507**(7493): p. 455-461.

6. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.

7. Luger, K., et al., *Crystal structure of the nucleosome core particle at 2.8 A resolution.* Nature, 1997. **389**(6648): p. 251-60.

8. Fyodorov, D.V., et al., *Emerging roles of linker histones in regulating chromatin structure and function.* Nat Rev Mol Cell Biol, 2017.

9. Song, F., et al., *Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units.* Science, 2014. **344**(6182): p. 376-80.

10. Ricci, M.A., et al., *Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo.* Cell, 2015. **160**(6): p. 1145-58.

11. Heitz, E., *Das Heterochromatin der Moose.* Jahrb Wiss Bot, 1928. **69**: p. 762-818.

12. Heitz, E., *Heterochromatin, Chromocentren, Chromomeren.* Ber Botan Ges, 1929. **47**: p. 274-284.

13. Heitz, E., *Die somatische Heteropyknose bei Drosophila melanogaster und ihre genetische Bedeutung.* Z Zellforsch Mikrosk Anat, 1933. **20**: p. 237-287.

14. Heitz, E., *Über totale und partielle somatische Heteropyknose, sowie strukturelle Geschlechtschromosomen bei Drosophila funebris.* Z Zellforsch Mikrosk Anat, 1933. **19**: p. 720-742.

15. Heitz, E., *Chromosomenstruktur und Gene.* Z Indukt Vereb, 1935. **70**: p. 402-447.

16. Guenatri, M., et al., *Mouse centric and pericentric satellite repeats form distinct functional heterochromatin.* J Cell Biol, 2004. **166**(4): p. 493-505.

17. Guelen, L., et al., *Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions.* Nature, 2008. **453**(7197): p. 948-51.

18. Peric-Hupkes, D., et al., *Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation.* Mol Cell, 2010. **38**(4): p. 603-13.

19. Meuleman, W., et al., *Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence.* Genome Res, 2013. **23**(2): p. 270-80.

20. van Koningsbruggen, S., et al., *High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli.* Mol Biol Cell, 2010. **21**(21): p. 3735-48.

21. Nemeth, A., et al., *Initial genomics of the human nucleolus.* PLoS Genet, 2010. **6**(3): p. e1000889.

22. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, 2009. **326**(5950): p. 289-93.

23. Nora, E.P., et al., *Spatial partitioning of the regulatory landscape of the X-inactivation centre.* Nature, 2012. **485**(7398): p. 381-5.

24. Sexton, T., et al., *Three-dimensional folding and functional organization principles of the Drosophila genome.* Cell, 2012. **148**(3): p. 458-72.

25. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-80.

26. Zhang, Y., et al., *Spatial organization of the mouse genome and its role in recurrent chromosomal translocations.* Cell, 2012. **148**(5): p. 908-21.

27. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, 2014. **159**(7): p. 1665-80.

28. Kagey, M.H., et al., *Mediator and cohesin connect gene expression and chromatin architecture.* Nature, 2010. **467**(7314): p. 430-5.

29. de Wit, E., et al., *CTCF Binding Polarity Determines Chromatin Looping.* Mol Cell, 2015. **60**(4): p. 676-84.

30. Rao, S.S.P., et al., *Cohesin Loss Eliminates All Loop Domains.* Cell, 2017. **171**(2): p. 305-320 e24.

31. Sainsbury, S., C. Bernecky, and P. Cramer, *Structural basis of transcription initiation by RNA polymerase II.* Nat Rev Mol Cell Biol, 2015. **16**(3): p. 129-43.

32. Sandelin, A., et al., *Mammalian RNA polymerase II core promoters: insights from genome-wide studies.* Nat Rev Genet, 2007. **8**(6): p. 424-36.

33. Yang, C., et al., *Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters.* Gene, 2007. **389**(1): p. 52-65.

34. Calo, E. and J. Wysocka, *Modification of enhancer chromatin: what, how, and why?* Mol Cell, 2013. **49**(5): p. 825-37.

35. Becker, P.B. and J.L. Workman, *Nucleosome remodeling and epigenetics.* Cold Spring Harb Perspect Biol, 2013. **5**(9).

36. Jones, P.A., *Functions of DNA methylation: islands, start sites, gene bodies and beyond.* Nat Rev Genet, 2012. **13**(7): p. 484-92.

37. Landt, S.G., et al., *ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.* Genome Res, 2012. **22**(9): p. 1813-31.

38. Song, L. and G.E. Crawford, *DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells.* Cold Spring Harb Protoc, 2010. **2010**(2): p. pdb prot5384.

39. Buenrostro, J.D., et al., *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.* Nat Methods, 2013. **10**(12): p. 1213-8.

40. Elkon, R. and R. Agami, *Characterization of noncoding regulatory DNA in the human genome.* Nat Biotechnol, 2017. **35**(8): p. 732-746.

41. Kwasnieski, J.C., et al., *Complex effects of nucleotide variants in a mammalian cis-regulatory element.* Proc Natl Acad Sci U S A, 2012. **109**(47): p. 19498-503.

42. White, M.A., et al., *Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks.* Proc Natl Acad Sci U S A, 2013. **110**(29): p. 11952-7.

43. Mogno, I., J.C. Kwasnieski, and B.A. Cohen, *Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants.* Genome Res, 2013. **23**(11): p. 1908-15.

44. Kwasnieski, J.C., et al., *High-throughput functional testing of ENCODE segmentation predictions.* Genome Res, 2014. **24**(10): p. 1595-602.

45. Shen, S.Q., et al., *Massively parallel cis-regulatory analysis in the mammalian central nervous system.* Genome Res, 2016. **26**(2): p. 238-55.

46. White, M.A., et al., *A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors.* Cell Rep, 2016. **17**(5): p. 1247-1254.

47. Patwardhan, R.P., et al., *High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis.* Nat Biotechnol, 2009. **27**(12): p. 1173-5.

48. Patwardhan, R.P., et al., *Massively parallel functional dissection of mammalian enhancers in vivo.* Nat Biotechnol, 2012. **30**(3): p. 265-70.

49. Inoue, F., et al., *A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity.* Genome Res, 2017. **27**(1): p. 38-52.

50. Grossman, S.R., et al., *Systematic dissection of genomic features determining transcription factor binding and enhancer function.* Proc Natl Acad Sci U S A, 2017. **114**(7): p. E1291-E1300.

51. Melnikov, A., et al., *Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay.* Nat Biotechnol, 2012. **30**(3): p. 271-7.

52. Tewhey, R., et al., *Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay.* Cell, 2016. **165**(6): p. 1519-1529.

53. Ulirsch, J.C., et al., *Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits.* Cell, 2016. **165**(6): p. 1530-1545.

54. Canver, M.C., et al., *BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis.* Nature, 2015. **527**(7577): p. 192-7.

55. Diao, Y., et al., *A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening.* Genome Res, 2016. **26**(3): p. 397-405.

56. Sanjana, N.E., et al., *High-resolution interrogation of functional elements in the noncoding genome.* Science, 2016. **353**(6307): p. 1545-1549.

57. Diao, Y., et al., *A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells.* Nat Methods, 2017. **14**(6): p. 629-635.

58. Macosko, E.Z., et al., *Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.* Cell, 2015. **161**(5): p. 1202-1214.

59. Shekhar, K., et al., *Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics.* Cell, 2016. **166**(5): p. 1308-1323 e30.

60. Helmstaedter, M., et al., *Connectomic reconstruction of the inner plexiform layer in the mouse retina.* Nature, 2013. **500**(7461): p. 168-74.

61. Muranishi, Y., et al., *An essential role for RAX homeoprotein and NOTCH-HES signaling in Otx2 expression in embryonic retinal photoreceptor cell fate determination.* J Neurosci, 2011. **31**(46): p. 16792-807.

62. Omori, Y., et al., *Analysis of transcriptional regulatory pathways of photoreceptor genes by expression profiling of the Otx2-deficient retina.* PLoS One, 2011. **6**(5): p. e19685.

63. Nishida, A., et al., *Otx2 homeobox gene controls retinal photoreceptor cell fate and pineal gland development.* Nat Neurosci, 2003. **6**(12): p. 1255-63.

64. Chen, S., et al., *Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes.* Neuron, 1997. **19**(5): p. 1017-30.

65. Freund, C.L., et al., *Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor.* Cell, 1997. **91**(4): p. 543-53.

66. Furukawa, T., E.M. Morrow, and C.L. Cepko, *Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation.* Cell, 1997. **91**(4): p. 531-41.

67. Furukawa, T., et al., *Retinopathy and attenuated circadian entrainment in Crx-deficient mice.* Nat Genet, 1999. **23**(4): p. 466-70.

68. Corbo, J.C., et al., *CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors.* Genome Res, 2010. **20**(11): p. 1512-25.

69. Livesey, F.J., et al., *Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene Crx.* Curr Biol, 2000. **10**(6): p. 301-10.

70. Hsiau, T.H., et al., *The cis-regulatory logic of the mammalian photoreceptor transcriptional network.* PLoS One, 2007. **2**(7): p. e643.

71. Pennesi, M.E., et al., *BETA2/NeuroD1 null mice: a new model for transcription factor-dependent photoreceptor degeneration.* J Neurosci, 2003. **23**(2): p. 453-61.

72. Ochocinska, M.J., et al., *NeuroD1 is required for survival of photoreceptors but not pinealocytes: results from targeted gene deletion studies.* J Neurochem, 2012. **123**(1): p. 44-59.

73. Andzelm, M.M., et al., *MEF2D drives photoreceptor development through a genome-wide competition for tissue-specific enhancers.* Neuron, 2015. **86**(1): p. 247-63.

74. Omori, Y., et al., *Mef2d is essential for the maturation and integrity of retinal photoreceptor and bipolar cells.* Genes Cells, 2015. **20**(5): p. 408-26.

75. Irie, S., et al., *Rax Homeoprotein Regulates Photoreceptor Cell Maturation and Survival in Association with Crx in the Postnatal Mouse Retina.* Mol Cell Biol, 2015. **35**(15): p. 2583-96.

76. Satoh, S., et al., *The spatial patterning of mouse cone opsin expression is regulated by bone morphogenetic protein signaling through downstream effector COUP-TF nuclear receptors.* J Neurosci, 2009. **29**(40): p. 12401-11.

77. Sapkota, D., et al., *Onecut1 and Onecut2 redundantly regulate early retinal cell fates during development.* Proc Natl Acad Sci U S A, 2014. **111**(39): p. E4086-95.

78. Fujieda, H., et al., *Retinoic acid receptor-related orphan receptor alpha regulates a subset of cone genes during mouse retinal development.* J Neurochem, 2009. **108**(1): p. 91-101.

79. Roberts, M.R., et al., *Retinoid X receptor (gamma) is necessary to establish the S-opsin gradient in cone photoreceptors of the developing mouse retina.* Invest Ophthalmol Vis Sci, 2005. **46**(8): p. 2897-904.

80. de Melo, J., et al., *The Spalt family transcription factor Sall3 regulates the development of cone photoreceptors and retinal horizontal interneurons.* Development, 2011. **138**(11): p. 2325-36.

81. Ng, L., et al., *A thyroid hormone receptor that is required for the development of green cone photoreceptors.* Nat Genet, 2001. **27**(1): p. 94-8.

82. Jia, L., et al., *Retinoid-related orphan nuclear receptor RORbeta is an early-acting factor in rod photoreceptor development.* Proc Natl Acad Sci U S A, 2009. **106**(41): p. 17534-9.

83. Montana, C.L., et al., *Transcriptional regulation of neural retina leucine zipper (Nrl), a photoreceptor cell fate determinant.* J Biol Chem, 2011. **286**(42): p. 36921-31.

84. Kautzmann, M.A., et al., *Combinatorial regulation of photoreceptor differentiation factor, neural retina leucine zipper gene NRL, revealed by in vivo promoter analysis.* J Biol Chem, 2011. **286**(32): p. 28247-55.

85. Fu, Y., et al., *Feedback induction of a photoreceptor-specific isoform of retinoid-related orphan nuclear receptor beta by the rod transcription factor NRL.* J Biol Chem, 2014. **289**(47): p. 32469-80.

86. Mears, A.J., et al., *Nrl is required for rod photoreceptor development.* Nat Genet, 2001. **29**(4): p. 447-52.

87. Yoshida, S., et al., *Expression profiling of the developing and mature Nrl-/- mouse retina: identification of retinal disease candidates and transcriptional regulatory targets of Nrl.* Hum Mol Genet, 2004. **13**(14): p. 1487-503.

88. Akimoto, M., et al., *Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors.* Proc Natl Acad Sci U S A, 2006. **103**(10): p. 3890-5.

89. Onishi, A., et al., *The orphan nuclear hormone receptor ERRbeta controls rod photoreceptor survival.* Proc Natl Acad Sci U S A, 2010. **107**(25): p. 11579-84.

90. Hao, H., et al., *The transcription factor neural retina leucine zipper (NRL) controls photoreceptor-specific expression of myocyte enhancer factor Mef2c from an alternative promoter.* J Biol Chem, 2011. **286**(40): p. 34893-902.

91. Milam, A.H., et al., *The nuclear receptor NR2E3 plays a role in human retinal photoreceptor differentiation and degeneration.* Proc Natl Acad Sci U S A, 2002. **99**(1): p. 473-8.

92. Chen, J., A. Rattner, and J. Nathans, *The rod photoreceptor-specific nuclear receptor Nr2e3 represses transcription of multiple cone-specific genes.* J Neurosci, 2005. **25**(1): p. 118-29.

93. Oh, E.C., et al., *Rod differentiation factor NRL activates the expression of nuclear receptor NR2E3 to suppress the development of cone photoreceptors.* Brain Res, 2008. **1236**: p. 16-29.

94. Wright, A.F., et al., *Photoreceptor degeneration: genetic and mechanistic dissection of a complex trait.* Nat Rev Genet, 2010. **11**(4): p. 273-84.

95. Sohocki, M.M., et al., *A range of clinical phenotypes associated with mutations in CRX, a photoreceptor transcription-factor gene.* Am J Hum Genet, 1998. **63**(5): p. 1307-15.

96. *RetNet*. Available from: http://www.sph.uth.tmc.edu/RetNet/.

97. Jeon, C.J., E. Strettoi, and R.H. Masland, *The major cell populations of the mouse retina.* J Neurosci, 1998. **18**(21): p. 8936-46.

98. Applebury, M.L., et al., *The murine cone photoreceptor: a single cone type expresses both S and M opsins with retinal spatial patterning.* Neuron, 2000. **27**(3): p. 513-23.

99. Haverkamp, S., et al., *The primordial, blue-cone color system of the mouse retina.* J Neurosci, 2005. **25**(22): p. 5438-45.

100. Solovei, I., et al., *Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution.* Cell, 2009. **137**(2): p. 356-68.

101. Solovei, I., K. Thanisch, and Y. Feodorova, *How to rule the nucleus: divide et impera.* Curr Opin Cell Biol, 2016. **40**: p. 47-59.

102. Pueschel, R., F. Coraggio, and P. Meister, *From single genes to entire genomes: the search for a function of nuclear organization.* Development, 2016. **143**(6): p. 910-23.

103. Cremer, T., et al., *Rabl's model of the interphase chromosome arrangement tested in Chinese hamster cells by premature chromosome condensation and laser-UV-microbeam experiments.* Hum Genet, 1982. **60**(1): p. 46-56.

104. Pinkel, D., T. Straume, and J.W. Gray, *Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization.* Proc Natl Acad Sci U S A, 1986. **83**(9): p. 2934-8.

105. Phillips-Cremins, J.E., et al., *Architectural protein subclasses shape 3D organization of genomes during lineage commitment.* Cell, 2013. **153**(6): p. 1281-95.

106. Solovei, I., et al., *LBR and lamin A/C sequentially tether peripheral heterochromatin and inversely regulate differentiation.* Cell, 2013. **152**(3): p. 584-98.

107. Swaroop, A., D. Kim, and D. Forrest, *Transcriptional regulation of photoreceptor development and homeostasis in the mammalian retina.* Nat Rev Neurosci, 2010. **11**(8): p. 563-76.

108. Brzezinski, J.A. and T.A. Reh, *Photoreceptor cell fate specification in vertebrates.* Development, 2015. **142**(19): p. 3263-73.

109. Lee, J., et al., *Quantitative fine-tuning of photoreceptor cis-regulatory elements through affinity modulation of transcription factor binding sites.* Gene Ther, 2010. **17**(11): p. 1390-9.

110. Koike, C., et al., *Functional roles of Otx2 transcription factor in postnatal mouse retinal development.* Mol Cell Biol, 2007. **27**(23): p. 8318-29.

111. Oh, E.C., et al., *Transformation of cone precursors to functional rod photoreceptors by bZIP transcription factor NRL.* Proc Natl Acad Sci U S A, 2007. **104**(5): p. 1679-84.

112. Corbo, J.C., et al., *A typology of photoreceptor gene expression patterns in the mouse.* Proc Natl Acad Sci U S A, 2007. **104**(29): p. 12069-74.

113. Nikonov, S.S., et al., *Photoreceptors of Nrl -/- mice coexpress functional S- and M-cone opsins having distinct inactivation mechanisms.* J Gen Physiol, 2005. **125**(3): p. 287-304.

114. Daniele, L.L., et al., *Cone-like morphological, molecular, and electrophysiological features of the photoreceptors of the Nrl knockout mouse.* Invest Ophthalmol Vis Sci, 2005. **46**(6): p. 2156-67.

115. Peng, G.H., et al., *The photoreceptor-specific nuclear receptor Nr2e3 interacts with Crx and exerts opposing effects on the transcription of rod versus cone genes.* Hum Mol Genet, 2005. **14**(6): p. 747-64.

116. Brooks, M.J., et al., *Next-generation sequencing facilitates quantitative analysis of wild-type and Nrl(-/-) retinal transcriptomes.* Mol Vis, 2011. **17**: p. 3034-54.

117. Siegert, S., et al., *Transcriptional code and disease map for adult retinal cell types.* Nat Neurosci, 2012. **15**(3): p. 487-95, S1-2.

118. Wilken, M.S., et al., *DNase I hypersensitivity analysis of the mouse brain and retina identifies region-specific regulatory elements.* Epigenetics Chromatin, 2015. **8**: p. 8.

119. Hao, H., et al., *Transcriptional regulation of rod photoreceptor homeostasis revealed by in vivo NRL targetome analysis.* PLoS Genet, 2012. **8**(4): p. e1002649.

120. Fei, Y. and T.E. Hughes, *Transgenic expression of the jellyfish green fluorescent protein in the cone photoreceptors of the mouse.* Vis Neurosci, 2001. **18**(4): p. 615-23.

121. Gonzalez, A.J., M. Setty, and C.S. Leslie, *Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation.* Nat Genet, 2015. **47**(11): p. 1249-59.

122. Song, L., et al., *Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity.* Genome Res, 2011. **21**(10): p. 1757-67.

123. Heinz, S., et al., *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.* Mol Cell, 2010. **38**(4): p. 576-89.

124. Splinter, E., et al., *CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus.* Genes Dev, 2006. **20**(17): p. 2349-54.

125. Briata, P., et al., *Binding properties of the human homeodomain protein OTX2 to a DNA target sequence.* FEBS Lett, 1999. **445**(1): p. 160-4.

126. Hu, S., A. Mamedova, and R.S. Hegde, *DNA-binding and regulation mechanisms of the SIX family of retinal determination proteins.* Biochemistry, 2008. **47**(11): p. 3586-94.

127. Morrow, E.M., et al., *NeuroD regulates multiple functions in the developing neural retina in rodent.* Development, 1999. **126**(1): p. 23-36.

128. Wilson, D., et al., *Cooperative dimerization of paired class homeo domains on DNA.* Genes Dev, 1993. **7**(11): p. 2120-34.

129. Wilson, D.S., et al., *High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA.* Cell, 1995. **82**(5): p. 709-19.

130. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

131. Montana, C.L., et al., *Reprogramming of adult rod photoreceptors prevents retinal degeneration.* Proc Natl Acad Sci U S A, 2013. **110**(5): p. 1732-7.

132. Mo, A., et al., *Epigenomic landscapes of retinal rods and cones.* Elife, 2016. **5**.

133. Hon, G.C., et al., *Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues.* Nat Genet, 2013. **45**(10): p. 1198-206.

134. Li, X., et al., *Tissue-specific regulation of retinal and pituitary precursor cell proliferation.* Science, 2002. **297**(5584): p. 1180-3.

135. Bernier, G., et al., *Expanded retina territory by midbrain transformation upon overexpression of Six6 (Optx2) in Xenopus embryos.* Mech Dev, 2000. **93**(1-2): p. 59-69.

136. Jean, D., G. Bernier, and P. Gruss, *Six6 (Optx2) is a novel murine Six3-related homeobox gene that demarcates the presumptive pituitary/hypothalamic axis and the ventral optic stalk.* Mech Dev, 1999. **84**(1-2): p. 31-40.

137. Conte, I., et al., *Proper differentiation of photoreceptors and amacrine cells depends on a regulatory loop between NeuroD and Six6.* Development, 2010. **137**(14): p. 2307-17.

138. Ogawa, Y., et al., *Homeobox transcription factor Six7 governs expression of green opsin genes in zebrafish.* Proc Biol Sci, 2015. **282**(1812): p. 20150659.

139. Sotolongo-Lopez, M., et al., *Genetic Dissection of Dual Roles for the Transcription Factor six7 in Photoreceptor Development and Patterning in Zebrafish.* PLoS Genet, 2016. **12**(4): p. e1005968.

140. Chen, S., et al., *Functional analysis of cone-rod homeobox (CRX) mutations associated with retinal dystrophy.* Hum Mol Genet, 2002. **11**(8): p. 873-84.

141. Rister, J., et al., *Single-base pair differences in a shared motif determine differential Rhodopsin expression.* Science, 2015. **350**(6265): p. 1258-61.

142. Kerppola, T.K. and T. Curran, *Maf and Nrl can bind to AP-1 sites and form heterodimers with Fos and Jun.* Oncogene, 1994. **9**(3): p. 675-84.

143. Enright, J.M., et al., *Transcriptome profiling of developing photoreceptor subtypes reveals candidate genes involved in avian photoreceptor diversification.* J Comp Neurol, 2015. **523**(4): p. 649-68.

144. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

145. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

146. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.

147. Li, Q.H., et al., *Measuring Reproducibility of High-Throughput Experiments.* Annals of Applied Statistics, 2011. **5**(3): p. 1752-1779.

148. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.

149. Mandal, M., et al., *Histone reader BRWD1 targets and restricts recombination to the Igk locus.* Nat Immunol, 2015. **16**(10): p. 1094-103.

150. Minnich, M., et al., *Multifunctional role of the transcription factor Blimp-1 in coordinating plasma cell differentiation.* Nat Immunol, 2016. **17**(3): p. 331-43.

151. Mo, A., et al., *Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain.* Neuron, 2015. **86**(6): p. 1369-84.

152. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

153. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.

154. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.

155. Neph, S., et al., *BEDOPS: high-performance genomic feature operations.* Bioinformatics, 2012. **28**(14): p. 1919-20.

156. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.

157. McLean, C.Y., et al., *GREAT improves functional interpretation of cis-regulatory regions.* Nat Biotechnol, 2010. **28**(5): p. 495-501.

158. R Core Team, *R: A Language and Environment for Statistical Computing*. 2016, R Foundation for Statistical Computing: Vienna, Austria.

159. Hothorn, T., et al., *Implementing a Class of Permutation Tests: The coin Package.* Journal of Statistical Software, 2008. **28**(8): p. 1-23.

160. Wickham, H., *ggplot2 : elegant graphics for data analysis*. Use R! 2009, New York: Springer. viii, 212 p.

161. Warnes, G.R., et al., *gplots: Various R Programming Tools for Plotting Data.* 2016.

162. Jolma, A., et al., *DNA-binding specificities of human transcription factors.* Cell, 2013. **152**(1-2): p. 327-39.

163. Badis, G., et al., *Diversity and complexity in DNA recognition by transcription factors.* Science, 2009. **324**(5935): p. 1720-3.

164. Weirauch, M.T., et al., *Determination and inference of eukaryotic transcription factor sequence specificity.* Cell, 2014. **158**(6): p. 1431-1443.

165. Chatelain, G., et al., *Molecular dissection reveals decreased activity and not dominant negative effect in human OTX2 mutants.* J Mol Med (Berl), 2006. **84**(7): p. 604-15.

166. Tucker, S.C. and R. Wisdom, *Site-specific heterodimerization by paired class homeodomain proteins mediates selective transcriptional responses.* J Biol Chem, 1999. **274**(45): p. 32325-32.

167. Arnold, C.D., et al., *Genome-wide quantitative enhancer activity maps identified by STARR-seq.* Science, 2013. **339**(6123): p. 1074-7.

168. Ghandi, M., et al., *gkmSVM: an R package for gapped-kmer SVM.* Bioinformatics, 2016. **32**(14): p. 2205-7.

169. Castro-Mondragon, J.A., et al., *RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections.* Nucleic Acids Res, 2017. **45**(13): p. e119.

170. Tibshirani, R., *Regression shrinkage and selection via the Lasso.* Journal of the Royal Statistical Society Series B-Methodological, 1996. **58**(1): p. 267-288.

171. Setty, M. and C.S. Leslie, *SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps.* PLoS Comput Biol, 2015. **11**(5): p. e1004271.

172. Fletez-Brant, C., et al., *kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W544-56.

173. Lee, D., *LS-GKM: a new gkm-SVM for large-scale datasets.* Bioinformatics, 2016. **32**(14): p. 2196-8.

174. Kelley, D.R., J. Snoek, and J.L. Rinn, *Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks.* Genome Res, 2016. **26**(7): p. 990-9.

175. Zack, D.J., et al., *Unusual topography of bovine rhodopsin promoter-lacZ fusion gene expression in transgenic mouse retinas.* Neuron, 1991. **6**(2): p. 187-99.

176. Kim, J.W., et al., *Recruitment of Rod Photoreceptors from Short-Wavelength-Sensitive Cones during the Evolution of Nocturnal Vision in Mammals.* Dev Cell, 2016. **37**(6): p. 520-32.

177. Kim, J.W., et al., *NRL-Regulated Transcriptome Dynamics of Developing Rod Photoreceptors.* Cell Rep, 2016. **17**(9): p. 2460-2473.

178. Hughes, A.E., et al., *Cell Type-Specific Epigenomic Analysis Reveals a Uniquely Closed Chromatin Architecture in Mouse Rod Photoreceptors.* Sci Rep, 2017. **7**: p. 43184.

179. Pollard, K.S., et al., *Detection of nonneutral substitution rates on mammalian phylogenies.* Genome Res, 2010. **20**(1): p. 110-21.

180. Chaney, B.A., et al., *Solution structure of the K50 class homeodomain PITX2 bound to DNA and implications for mutations that cause Rieger syndrome.* Biochemistry, 2005. **44**(20): p. 7497-511.

181. Ghandi, M., et al., *Enhanced regulatory sequence prediction using gapped k-mer features.* PLoS Comput Biol, 2014. **10**(7): p. e1003711.

182. Lee, D., et al., *A method to predict the impact of regulatory variants from DNA sequence.* Nat Genet, 2015. **47**(8): p. 955-61.

183. Alvarez-Delfin, K., et al., *Tbx2b is required for ultraviolet photoreceptor cell specification during zebrafish retinal development.* Proc Natl Acad Sci U S A, 2009. **106**(6): p. 2023-8.

184. Abrahams, A., M.I. Parker, and S. Prince, *The T-box transcription factor Tbx2: its role in development and possible implication in cancer.* IUBMB Life, 2010. **62**(2): p. 92-102.

185. Zhou, J. and O.G. Troyanskaya, *Predicting effects of noncoding variants with deep learning-based sequence model.* Nat Methods, 2015. **12**(10): p. 931-4.

186. Grant, C.E., T.L. Bailey, and W.S. Noble, *FIMO: scanning for occurrences of a given motif.* Bioinformatics, 2011. **27**(7): p. 1017-8.

187. Montana, C.L., C.A. Myers, and J.C. Corbo, *Quantifying the activity of cis-regulatory elements in the mouse retina by explant electroporation.* J Vis Exp, 2011(52).

188. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.

189. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* J Stat Softw, 2010. **33**(1): p. 1-22.

190. Magoc, T. and S.L. Salzberg, *FLASH: fast length adjustment of short reads to improve genome assemblies.* Bioinformatics, 2011. **27**(21): p. 2957-63.

191. Bray, N.L., et al., *Near-optimal probabilistic RNA-seq quantification.* Nat Biotechnol, 2016. **34**(5): p. 525-7.

192. Kuhn, M., *Building Predictive Models in R Using the caret Package.* Journal of Statistical Software, 2008. **28**(5): p. 1-26.

193. Grau, J., I. Grosse, and J. Keilwagen, *PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R.* Bioinformatics, 2015. **31**(15): p. 2595-7.

194. Keilwagen, J., I. Grosse, and J. Grau, *Area under precision-recall curves for weighted and unweighted data.* PLoS One, 2014. **9**(3): p. e92209.

195. Carss, K.J., et al., *Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease.* Am J Hum Genet, 2017. **100**(1): p. 75-90.

196. Consortium, U.K., et al., *The UK10K project identifies rare variants in health and disease.* Nature, 2015. **526**(7571): p. 82-90.

197. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome.* Nature, 2012. **489**(7414): p. 75-82.

198. Boyle, A.P., et al., *High-resolution mapping and characterization of open chromatin across the genome.* Cell, 2008. **132**(2): p. 311-22.

199. *Picard*. Available from: broadinstitute.github.io/picard/.

200. Smallwood, P.M., Y. Wang, and J. Nathans, *Role of a locus control region in the mutually exclusive expression of human red and green cone pigment genes.* Proc Natl Acad Sci U S A, 2002. **99**(2): p. 1008-11.

201. Wang, Y., et al., *A locus control region adjacent to the human red and green visual pigment genes.* Neuron, 1992. **9**(3): p. 429-40.

202. Nathans, J., et al., *Molecular genetics of human blue cone monochromacy.* Science, 1989. **245**(4920): p. 831-8.

Appendix: Functional regulatory variation in the human retina

5.1 Introduction

Inherited retinal disease (IRD) is a heterogeneous group of genetic diseases characterized by retinal dysfunction and vision loss. The genetic architecture of IRD is complex. More than 250 different genes have been implicated in IRD [96], and different variants within the same gene can lead to distinct pathologies [95]. Furthermore, patients harboring the same causal variant can present with divergent phenotypes, suggesting that genetic modifiers play an important role. As a result, high-throughput sequencing is playing an increasingly important role in the diagnosis of IRD. Recently, a major study reported whole-genome sequencing of 650 IRD patients, which identified pathogenic protein-coding variants in ~57% of cases [195]. These results raise the possibility that noncoding (regulatory) variation contributes to a subset of the remaining ('unsolved') cases. Nevertheless, assessing the contribution of regulatory variation is challenging, because human retinal *cis*-regulatory elements (CREs) have not been comprehensively defined, and how regulatory variation impacts the activity of human retinal CREs is not known.

Here, we compare open chromatin maps from human retina, heart, kidney, and lung to identify >26,000 retina-specific CREs. These CREs are phylogenetically conserved, flank photoreceptor-specific genes, and are enriched for transcription factor (TF) binding sites corresponding to photoreceptor-specific TFs. To determine which of these regions are functional, we have begun to undertake a massively parallel reporter assay (MPRA) to quantify the activity of >4,600 CREs in explanted human and mouse retinas. In addition to native CREs, we will assay rare and common variants from the UK10K project to determine the general impact of genetic variation on CRE function [196]. Finally, we will assay rare variants from the IRD 650 cohort to

identify regulatory variants with pathogenic potential. Thus, this study will constitute the first high-throughput functional characterization of human retinal CREs *in vivo*. Furthermore, this study will assess the utility of MPRAs for interpreting clinical genomic data, and potentially identify regulatory variants that contribute to IRD.

5.2 Methods and preliminary results

5.2.1 Identification of retina-specific open chromatin in humans

Open chromatin profiling (including DNase-seq and ATAC-seq) is a powerful approach for mapping candidate CREs at high resolution genome-wide [39, 197, 198]. Recently, we obtained ATAC-seq data from the Qian and Blackshaw groups from paired macula and peripheral retina samples from five adult human donors. In addition, we downloaded DNase-seq data from adult human brain, heart, kidney, and lung as well as fetal retina from the ENCODE project [4]. ATAC-seq and DNase-seq data were processed as follows. Paired-end reads were aligned to hg19 with bowtie2 (v2.3.0), allowing a maximum fragment length of 2 kb [144]. Alignments with improper pairing, mapping quality <30, or overlapping ENCODE blacklist regions [4] were removed with SAMtools (v1.5) [145]. Alignments were then sorted, merged, and deduplicated with Picard [199]. To enrich for nucleosome-free reads, read pairs with fragment sizes greater than 150 bp were discarded. Finally, peaks were called with MACS2 (v2.1.1) [146].

ATAC-seq generated reproducible profiles of genome-wide chromatin accessibility across all ten retinal samples (pairwise correlation 0.81-0.97), and we identified a total of 195,782 retinal ATAC-seq peaks. Comparison of open chromatin profiles between human and mouse indicates that the *cis*-regulatory architecture of individual retinal genes is often conserved [118, 132, 178]. For example, at the rod-specific rhodopsin locus (Fig. 5.1A-B), both human and mouse ATAC-seq identify strong open chromatin peaks at the promoter, two upstream enhancers, an intronic enhancer 3' of the second exon, and several additional peaks towards the 3' UTR. Furthermore,

ATAC-seq detects well-characterized human retinal CREs, including a cluster of retina-specific peaks ~3.5 kb upstream of the long-wavelength opsin gene (*OPN1LW*) (Fig. 5.1C). These peaks overlap the so-called 'locus control region' (LCR) which regulates the expression of both *OPN1LW* and *OPN1MW* (~20 kb downstream) [200, 201]. Previously, a deletion series ascertained from 12 families with blue cone monochromacy demonstrated that a 579-bp region precisely overlapping one of the LCR ATAC-seq peaks is critical for the expression of both *OPN1LW* and *OPN1MW* [202]. Finally, ATAC-seq from human retina allows us to study the regulation of photoreceptor genes present in humans but not mice, such as *RAX2* (Fig. 5.3D). Thus, ATAC-seq detects functional retinal CREs, including CREs that play a role in human retinal disease and CREs that cannot be studied in mice.

To define retina-specific CREs, we first merged peak calls from adult retina ATAC-seq and control tissue DNase-seq to generate a master list of 508,060 regions. We then used DESeq2 (v1.14.1) to identify peaks with significantly more reads in adult retina vs. control tissues (excluding fetal retina), yielding a set 26,156 adult retina-specific peaks (FDR<1%). We used GREAT (v3.0.0) to analyze functional annotations associated with these regions, and we found that they were highly enriched for gene ontology (GO) terms associated with photoreceptor physiology, and human and mouse phenotypes associated with vision and retinal disease (Table 5.1) [157]. In addition, we used HOMER (v4.9) to identify TF binding sites enrichment in tissue-specific CREs from retina and control tissues [123]. This analysis revealed that human retina-specific CREs are centered on focal enrichments in TF binding sites corresponding to known photoreceptor TFs (Fig. 5.2): K50 homeodomain (*OTX2* and *CRX*), Q50 homeodomain (*RAX*), basic helix-loop-helix (*NEUROD1*), nuclear receptor (*RORA*, *RORB*, NR2E3, *ESRRB*, and *THRB*), and MADS TFs (*MEF2C*, *MEF2D*). Furthermore, we find that the TF binding sites enriched in

human retina are highly similar to those we have previously characterized in mouse rods and cones, suggesting that the *cis*-regulatory grammar of these cell types is largely conserved.

5.2.2 Quantifying the effect of regulatory variation in human retinal CREs by MPRA

Previously, we and others have described CRE-seq, an MPRA for assaying the activity of thousands of CREs in a single experiment [41-46]. For CRE-seq, oligos of interest (typically 80-150 bp) are cloned upstream of a promoter-DsRed construct harboring CRE-specific barcodes in the 3' UTR. This library is then delivered to cells of interest, and activity is later quantified by harvesting RNA and DNA and counting barcodes by high-throughput sequencing.

To quantify the activity of human retinal CREs, and to test the effect of single nucleotide variants (SNVs) on CRE activity, we have designed a CRE-seq library as follows. First, we obtained variant calls from the Raymond group for whole-genome sequencing of 650 patients with IRD [195]. Across all 650 patients, we identified 77,819 SNVs and indels that overlap retina ATAC-seq peaks. Furthermore, 55,205 of these variants were identified in patients where no pathogenic variants were identified ('unsolved', n=269) or in patients where a single pathogenic allele was detected in a recessive disease gene ('partially solved', n=13). To screen for pathogenic regulatory variants, we selected rare SNVs from the 282 unsolved or partially solved cases overlapping retina-specific ATAC-seq peaks near known retinal disease genes and included these in our CRE-seq library. Specifically, we selected all SNVs with a minor allele frequency (MAF) <0.01 located within 75 bp of retina-specific ATAC-seq peak summits that were within 200 kb of genes in the RetNet database [96]. These criteria identified 1,952 variants in 1,111 CREs.

We then selected additional variants to broadly survey the functional impact of regulatory variation across a range of population frequencies. From all 650 IRD patients, we identified 8,423 SNVs overlapping conserved positions (GERP>2) within 75 bp of retinal ATAC-seq peaks [179].

From these variants, we selected all common SNVs (MAF >0.01) (n=1,420), all SNVs overlapping K50 homeodomain, Q50 homeodomain, basic helix-loop-helix, or nuclear receptor TF binding sites (n=872), and a random sample of the remaining (rare) SNVs (n=2,500). These criteria identified 3,989 SNVs in 3,279 retinal ATAC-seq peaks.

Finally, in addition to the 5,816 unique SNVs described above (and corresponding reference sequences), we included all retina-specific ATAC-seq peaks near RetNet genes (n=1,650). Thus, the final CRE-seq library consisted of 150-bp elements centered on retinal ATAC-seq peaks—4,698 reference sequences and 5,618 sequences harboring SNVs for a total of 10,514 distinct oligos. We paired each oligo with 9-10 unique 15-bp barcodes (pairwise edit distance > 2), yielding a set of 100,000 library oligos. These oligos were cloned upstream of a pCrx-DsRed reporter as described previously (see Chapter 3), and will be assayed by electroporation into explanted mouse retina as well as viral transduction into postmortem adult human retinas in culture in future experiments.

5.3 Summary

IRD is a clinically important disease that is incompletely explained by genetic variation in protein-coding genes. Here, we present an initial identification and characterization of retina-specific CREs, which provides a valuable resource for studying the role of noncoding variation in retinal biology. Furthermore, we have designed a functional assay to quantify the effects of common and rare variants within these regions—including an extensive screen for pathogenic regulatory variants from a large cohort of patients with IRD. Thus, this study will define and quantify the functional activity of a reference set of human retinal CREs, assess the contribution of a relatively under-studied class of genetic variation to retinal disease, and provide a framework for using MRPAs to characterize the pathogenic potential of individual regulatory variants.

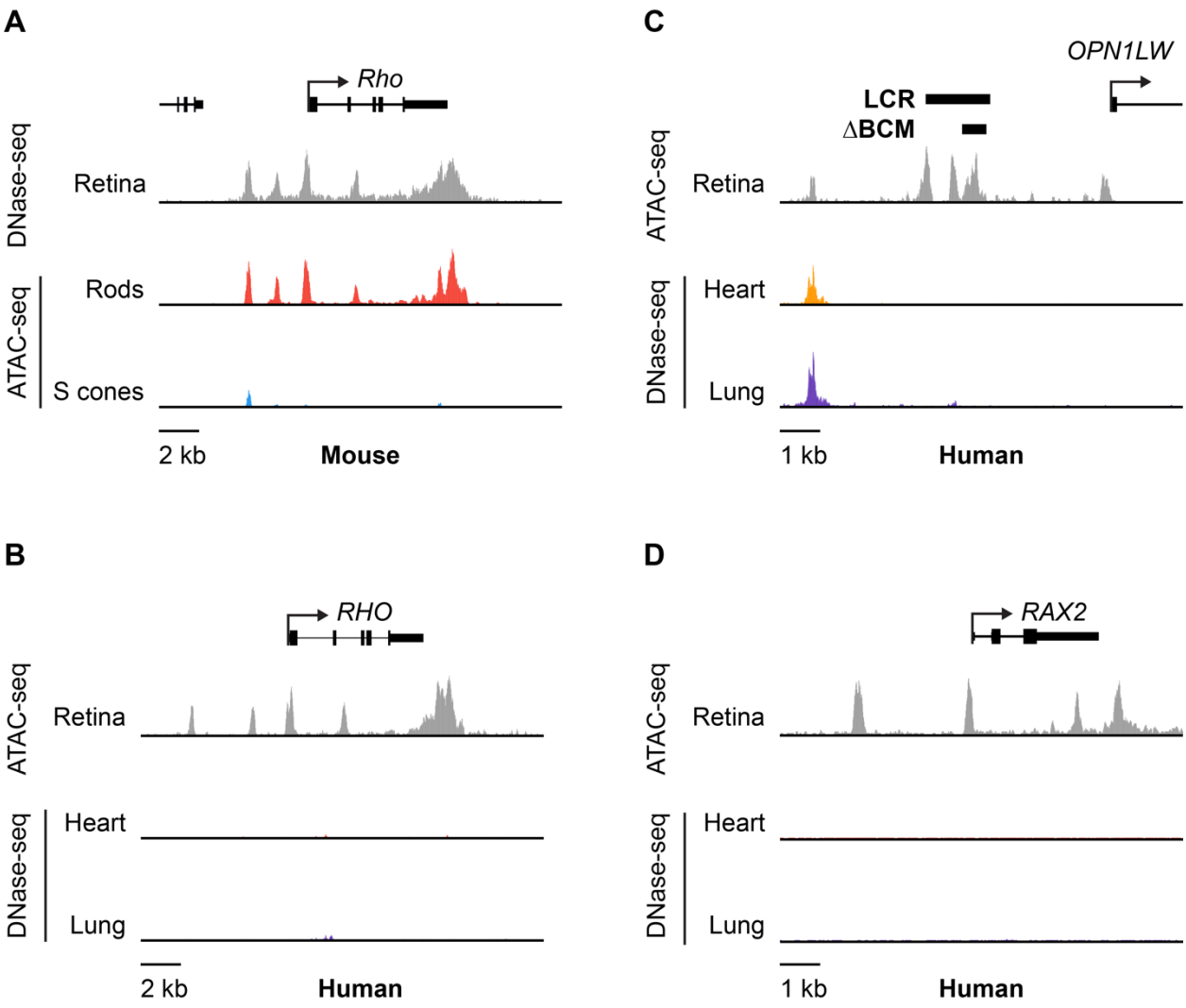


Figure 5.1. Open chromatin profiles of human and mouse retina. (A) DNase-seq from adult mouse retina and ATAC-seq from adult mouse rods and cones at the rod-specific *Rho* locus. (B) ATAC-seq from adult human retina and DNase-seq from adult human heart and lung at the rod-specific *RHO* locus. Human and mouse retinas have highly similar open chromatin profiles, illustrating conservation of regulatory architecture. (C) As in (B) at the cone-specific *OPN1LW* locus. LCR: locus control region. ΔBCM: a 579-bp region critical for *OPN1LW* and *OPN1MW* expression, previously shown to be deleted in patients with blue cone monochromacy. (D) as in (B) at the human-specific *RAX2* locus.

GO Collection	Term	Adjusted P-value	Fold Enrichment
GO Biological Process	Rhodopsin mediated signaling pathway	1.54E-60	3.46
	Regulation of rhodopsin mediated signaling pathway	4.32E-60	3.55
	Detection of light stimulus	4.33E-53	2.08
	Detection of visible light	6.92E-51	2.14
	Photoreceptor cell differentiation	8.50E-49	2.27
	Phototransduction	1.69E-48	2.07
	Retina homeostasis	7.96E-48	2.97
	Photoreceptor cell maintenance	6.80E-44	2.99
	Phototransduction, visible light	2.48E-43	2.09
	Eye photoreceptor cell differentiation	2.55E-41	2.20
	Cellular response to light stimulus	1.20E-37	2.09
	Photoreceptor cell development	1.36E-31	2.19
	Neural retina development	7.68E-28	2.02
	Eye photoreceptor cell development	2.04E-24	2.09
GO Human Phenotype	Night blindness	2.90E-121	2.85
	Abnormal electroretinogram	2.21E-96	2.45
	Abnormality of corneal thickness	1.16E-92	2.82
	Decreased corneal thickness	2.34E-90	2.83
	Hyperinsulinemia	4.13E-85	2.76
	Photophobia	6.73E-83	2.33
	Type II diabetes mellitus	7.09E-71	2.49
GO Mouse Phenotype	Retinal photoreceptor degeneration	9.75E-86	2.72
	Abnormal retinal photoreceptor morphology	2.22E-85	2.13
	Abnormal rod electrophysiology	3.77E-83	2.49
	Abnormal retinal photoreceptor layer morphology	1.04E-79	2.02
	Abnormal cone electrophysiology	2.70E-73	2.56
	Abnormal retinal rod cell outer segment morphology	4.39E-60	4.36
	Abnormal retinal outer nuclear layer morphology	5.41E-60	2.05
	Abnormal retinal rod cell morphology	5.08E-58	2.67
	Abnormal photoreceptor outer segment morphology	9.49E-58	2.20
	Retinal degeneration	3.62E-57	2.26
	Decreased retinal photoreceptor cell number	9.42E-40	2.22
	Abnormal retinal cone cell morphology	2.24E-36	2.27
	Short photoreceptor outer segment	7.67E-33	2.58
	Retinal rod cell degeneration	5.05E-29	3.57
	Abnormal retinal rod bipolar cell morphology	1.11E-23	2.31
	Absent branchial arches	1.81E-19	3.00
	Absent startle reflex	1.11E-16	2.06
	Retinal cone cell degeneration	1.67E-15	2.32

Table 5.1. GO enrichment analysis of human retina-specific ATAC-seq peaks.

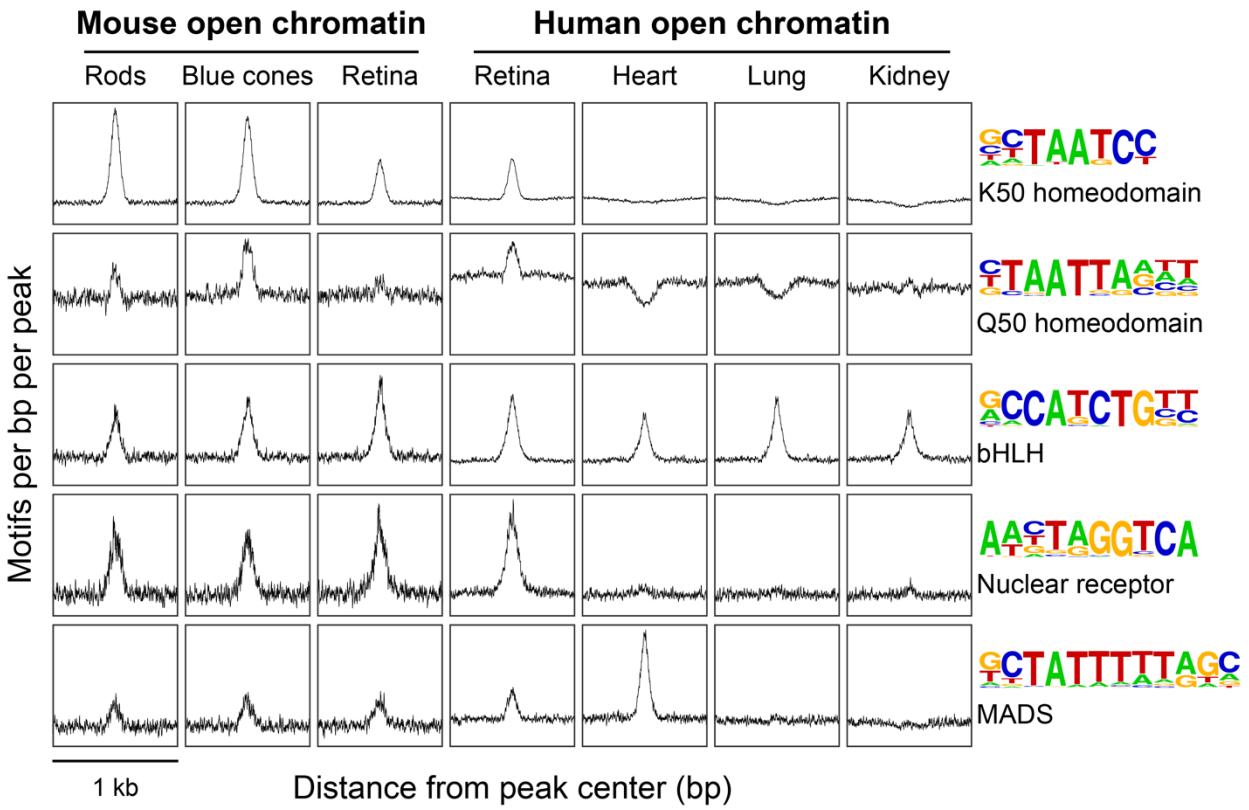


Figure 5.2. Motif enrichment in human and mouse open chromatin. Histograms showing average motif density in open chromatin from adult mouse retina and photoreceptors (left) as well as adult human retina, heart, lung, and kidney (right). Plots show normalized motif counts in a 1-kb window centered on ATAC-seq or DNase-seq from each tissue—smoothed with a 20-bp rolling average. Retina-specific TF binding site enrichment is similar between mice and humans, highlighting conservation of cis-regulatory grammar across species.