SUPPLEMENTAL MATERIAL

Aberrant homeodomain-DNA cooperative dimerization underlies distinct developmental defects in two dominant *CRX* retinopathy models

Authors
Yiqiao Zheng[1,2,§], Gary D. Stormo[4*], Shiming Chen[2,3*]

Affiliations
[1] Molecular Genetics and Genomics Graduate Program, Division of Biological and Biomedical Sciences [2] Department of Ophthalmology and Visual Sciences [3] Department of Developmental Biology [4] Department of Genetics, Washington University in St Louis, Saint Louis, Missouri, 63110, USA [§] Current address: Department of Biology, Massachusetts Institute of Technology

TABLE OF CONTENTS

SUPPLEMENTAL METHODS

Animal study and sample collection

ATAC-seq sample collection and library preparation

The assay for transposase-accessible chromatin with sequencing was performed as previously published (Buenrostro et al. 2015). Briefly, for each genotype, three biological replicates, two retinas per replicate from one male and one female were pooled. The pooled retinas were washed twice with cold PBS before being dissociation in 250µl TESCA buffer (50 mM TES, 0.36 mM calcium chloride, pH 7.4 at 37°C) containing 2% collagenase (Sigma) for 10mins at 37°C. The dissociation mixture was subjected to DNase I treatment (10ul DNaseI at 2U/µl, New England Biolabs) for an additional 3mins at 37°C. The DNase I activity was quenched by adding 500ul DMEM (Gibco™) with 10% HI-FBS (Gibco™) and incubating for 5mins at room temperature. The reaction mixture was passed through a 40µm strainer (Falcon® Corning) twice and centrifugated at 400rcf for 5mins at 4°C to collect cells. The cells were washed twice with cold PBS, centrifugated at 500rcf for 5mins at 4°C, and gently resuspended in 400µl cold lysis buffer (ATAC-RSB, 0.1%Tween20, 0.1%NP-40). The lysis reaction was carried out on ice for 5mins. The nuclei were separated from the lysis mixture by centrifugation at 500rcf for 5mins at 4°C, washed once with cold nucleus resuspension buffer (ATAC-RSB, 0.1%Tween20, 0.01% Digitonin), and collected by centrifugation at 500rcf for 10mins at 4°C. The nuclei were gently resuspended in 20µl 2× TD (Illumina Cat #20034211) buffer and counted with a hemocytometer by staining the intact nuclei with SYBR® Green I Nucleic Acid Stain (Invitrogen™). 10k nuclei were aliquoted and subjected to tagmentation (50 µl reaction with 2.5µl Illumina Tagment DNA TDE1 Enzyme) for 30mins at 37°C following a standard protocol. The recipe for buffer ATAC-RSB is 10mM Tris-HCl, 10mM NaCl, 3mM $MgCl_2$.

The tagmented DNAs were cleaned up by the MinElute® PCR purification kit (QIAGEN), eluted in 23µl Nuclease-Free Water (Invitrogen™ Ambion®) and stored at -20°C if not immediately used for library preparation. The ATAC-seq libraries were constructed by PCR amplification of the tagmented DNAs for 10 cycles using the NEBNext® High-Fidelity 2× PCR Master Mix (New England Biolabs). The PCR reactions were directly subjected to double size selection (0.55× and 1.55×) with the AMPure XP Beads (Beckman Coulter) following the manufacturer's protocol. The cleaned-up ATAC-seq fragments were eluted with 12µl Nuclease-Free Water (Invitrogen™). The quantity and quality of the ATAC-seq libraries was assayed using the Qubit™ 3 Fluorometer (Invitrogen™) and the Bioanalyzer (Agilent, Santa Clara, CA) prior to sequencing.

MPRA plasmid library construction

The MPRA plasmid library was constructed following published protocols (Hughes et al. 2018). Briefly, the library of 200mer oligonucleotides was ordered directly from Twist Bioscience (South San Francisco, CA). Each oligo contains a 134bp testing cis-regulatory element (CRE) sequence and a unique 10bp barcode flanked by sequences required for cloning (Supplemental Fig. S6A). The plasmid library was amplified with NEB Phusion® High-Fidelity PCR Master Mix with HF Buffer (New England Biolabs). The amplified fragments were cloned into the EcoRI-EagI sites of the vector pJK03 (Addgene ID173490). Lastly, the *Crx* promoter-DsRed fragment was amplified from the pCrx-DsRed plasmid (Hughes et al. 2018) and cloned into the SpeI-SphI sites located between the CRE sequence and the barcode. The constructed MPRA plasmid library was amplified by transformation into the NEB® 5-alpha Competent E. coli (High Efficiency) DH5α (New England Biolabs) according to manufacturer instructions. Prior to retinal electroporation, the fragments containing the 10bp unique barcode were amplified from the plasmid library and subjected to Sanger sequencing and 2x150bp Nova-seq to ensure successful library construction. The complete CRE sequences and annotations can be found in the GitHub repository for this paper.

MPRA plasmid preparation for electroporation

For each electroporation, 30µg of MPRA plasmid library DNA per retina (3 retinas per replicate, a total of 3-4 replicates per genotype) was aliquoted and brought to 150µl with water. 15µl NaOAc (pH5.2) was added to the DNA followed by vortex. 450µl absolute ethanol was added to the mixture followed by vortex. The DNA was precipitated by centrifugation at 17115g (13500rpm) for 30mins at 4°C followed by a second wash with 400µl 70% ethanol and centrifugation 17115g (13500rpm) for 15mins at 4°C. The DNA pellet was then air-dry until semi-transparent and resuspended thoroughly in appropriate amount of PBS. The concentration of the prepared DNA was measured with the Qubit™ 3 Fluorometer (Invitrogen™) and adjusted with PBS to a final concentration of 0.5µg/µl.

Retinal ex plant electroporation

Retinas were dissected in DMEM/F12 1:1 buffer (Gibco™) from newborn (P0) WT and *Crx* mutant mice. For each replicate, three retinas were transferred into the electroporation chamber (CUY520P5, NEPA GENE Co. Ltd) filled with the prepared MPRA plasmid DNA solution. The retinas were electroporated with the Electro Square Porator™ (ECM®830, BTX) with settings: LV, V: 30 Volts, Pulse Length: 50 msec, # Pulses: 5, Interval: 950 msec. The electroporated retinas were washed once in the retinal *ex plant* culture medium, carefully transferred onto a 0.2µm 25mm Nuclepore™ Track-Etch Membrane (Whatman®) and cultured for 8 days in the incubator (37°C, 5% CO_2) with the retinal *ex plant* culture medium (DMEM/F12 1:1, 10% HI-FBS, 1xPenicillin-Streptomycin). The day-8 *ex plant* retinas were collected, washed once in cold PBS, and stored in 30µl Nuclease-Free Water (Invitrogen™) at -80°C before ready for TRIzol extraction.

Retinal ex plant RNA/DNA Trizol extraction and purification

The RNA and DNA was extracted from *ex plant* cultured retinas using TRIzol™ Reagent (Invitrogen™) following a modified protocol based on the manufacturer's protocol. The retinas were thawed on ice, washed once with ice-cold Molecular Biology Grade water (Corning), and resuspended in 400µl TRIzol™ Reagent. The retinas were homogenized with a PRO250 Homogenizer (PRO Scientific Inc.) and a 5mm×75mm Flat Bottom Generator Probe (PRO Scientific Inc.) at intensity level 3 for 15sec. The homogenized tissue was incubated for 5mins at room temperature before 80µl chloroform was added followed by vigorous mixing by hand for 15sec. The TRIzol-chloroform mixture was incubated at room temperature for 2-3 mins and centrifuged at 12000g at 4°C for 15mins. The top aqueous phase containing RNA was transferred to a new 1.5ml Eppendorf tube. The RNA was purified and concentrated with TURBO DNA-free™ Kit (Invitrogen™) and Zymo RNA Clean & Concentrator™-5 kit (Zymo Research) following manufacturers' protocols and eluted twice with 15µl Nuclease-Free Water (Invitrogen™).

To extract DNA, any remaining aqueous layer containing RNA was removed from the TRIzol mixture. 200µl DNA back extraction buffer (4 M guanidine thiocyanate, 50 mM sodium citrate and 1 M Tris (free base), 0.5 vol of starting TRIzol Reagent used) was added to the remaining TRIzol mixture, mixed by inversion, and incubated at room temperature for 10mins. The mixture was centrifuged at 12000g at room temperature for 30mins. The top aqueous phase containing DNA was transferred to a new 1.5 ml Eppendorf tube. 160µl (0.4 vol of starting TRIzol Reagent used) of absolute isopropanol was added to the aqueous layer, mixed by inversion, and incubated at room temperature for 5mins before centrifugation at 12000g at 4°C for 15mins to pellet the DNA. The DNA pellet was washed twice with 200µl 70% ethanol and centrifuged at 12000g at 4°C for 15mins. The DNA pellet was then air-dry until semi-transparent and resuspended thoroughly in 30µl of Nuclease-Free Water (Invitrogen™). The samples were left at 4°C overnight for the DNA to fully dissolve. The quantity and quality of the prepared RNA and DNA was determined with the Qubit™ 3 Fluorometer (Invitrogen™).

MPRA sequencing library preparation

For each sample, 1µg of purified total RNA was used for cDNA synthesis in 20µl volume using the iScript™ cDNA Synthesis Kit (Bio-Rad Laboratories) following the manufacturer's protocol. Estimation of the average MPRA library activity was determined by qRT-PCR quantification against the DsRed sequence in MPRA library oligos and the endogenous *Crx* and *Rho* genes (Supplemental Table S1). Three rounds of PCR amplifications were performed to add internal phasing indexes and unique-dual-indexes (UDIs) for Illumina sequencing. PCR1: MPRA library fragments containing the 10bp barcodes were amplified from the cDNA mixture with the Q5® High-Fidelity 2× Master Mix (New England Biolabs). For each sample, 4-8 PCR1 replicates were pooled and purified with Monarch® PCR & DNA Cleanup Kit (New England Biolabs), eluted in 20µl of Nuclease-Free Water (Invitrogen™), and quantified with the Qubit™ 3 Fluorometer (Invitrogen™). PCR2: P1 adapters containing the inner phasing index were added to the amplified fragments from PCR1. PCR3: For each sample, 2µl of PCR2 reaction was used for PCR3 to add the Illumina sequencing adapters and indexes. The final reactions were cleaned up with Monarch® PCR & DNA Cleanup Kit (New England Biolabs), eluted in 30µl of Nuclease-Free Water (Invitrogen™), and quantified with the Qubit™ 3 Fluorometer (Invitrogen™). The sequences of the internal phasing indexes, PCR primers and PCR protocols for MPRA sequencing library preparation can be found in the GitHub repository of this paper.

Biochemistry

Protein expression for EMSA

GST-WT, E80A, K88N and R90W HD peptides were expressed in *E. coli* BL-21 (DE3) cells and purified with GST Spintrap™ columns (Cytiva, Marlborough, MA) as published previously (Chen et al. 2002; Zheng et al. 2023). The peptides were eluted following the manufacturer's protocol and buffer exchanged into 1×CRX binding buffer (60mM KCl, 25mM HEPES, 5% glycerol, 1mM DTT) using Amicon centrifugal filters (MilliporeSigma, Burlington, MA). The protein stock was supplemented with 10% glycerol before aliquoted and stored at -80°C.

EMSA BAT-1 probe and Coop-seq library synthesis and purification

The *BAT-1* templates (Supplemental Table S1) and IRDye700-labeled reverse complements were ordered directly from Integrated DNA Technologies (IDT). Equal amounts of two oligo strands were mixed in 20µl EB buffer (QIAGEN) and heated to 94°C for 2mins followed by gradual cooling. The Coop-seq libraries were prepared following the Spec-seq library preparation protocol as previously described (Zheng et al. 2023). Briefly, single-stranded Coop-seq library templates (Supplemental Table S1) and IRDye800-labeled reverse complement primers were ordered directly from Integrated DNA Technologies (IDT). 100pmol of template oligos and 125pmol of reverse complement primers F1 were mixed in Phusion® High-Fidelity PCR Master Mix (New England Biolabs). A 15s denaturing at 95°C following a 10-minute extension at 52°C afforded duplex DNAs. The mixture was treated with 1µl Exonuclease I (New England Biolabs) at 37°C for 30mins to remove residual ssDNA. The probes were purified with the MinElute® PCR purification kit (QIAGEN), eluted in an appropriate amount of Nuclease-Free Water (Invitrogen™), and stored at -20°C.

EMSA and sample preparation for sequencing

The HD-DNA binding reactions and EMSAs were performed as previously described (Zheng et al. 2023). Briefly, the protein-DNA binding reactions was performed in 1x CRX binding buffer (60mM KCl, 25mM HEPES, 5% glycerol, 1mM DTT). A fixed amount (Supplemental Table S1) of IRDye-labelled BAT-1 probes or Coop-seq libraries were incubated on ice for 30 minutes with varying concentrations of CRX HD peptides in 20µl reaction volume. The reaction mixtures were then run at 4°C in native 12% Tris-Glycine PAGE

gel (Invitrogen™) at 160V for 40min. The IRDye-labeled DNA fragments were visualized by Odyssey® CLx and Fc Imaging Systems (LI-COR, Inc.). In the HD-*BAT-1* binding reactions, for visualization purpose, GST tags were cleaved by treating the GST-HD recombinant peptides with 1μl Thrombin protease (1:10 dilution in PBS, Millipore Sigma) for 15mins at room temperature. The GST-HD digestion reactions were used directly in HD-DNA binding assays.

For Coop-seq libraries, the visible bands were excised from the gels. The DNAs were extracted with acrylamide extraction buffer (100mM NH_4OAc, 10mM $Mg(OAc)_2$, 0.1% SDS) and purified with MinElute® PCR Purification Kit (QIAGEN). The extracted Coop-seq DNAs were amplified, barcoded by indexed Illumina primers, then pooled and sequenced on a single 1×50bp Miseq run at DNA Sequencing Innovation Lab at the Center for Genome Sciences & Systems Biology (CGS&SB, WashU).

qRT-PCR

Whole retina RNA extraction, purification and cDNA synthesis was performed as previously described (Zheng et al. 2023). Primers used in this study are listed in Supplemental Table S1. qRT-PCR reactions were assembled using the SsoFast™ EvaGreen® Supermix with Low ROX (Bio-Rad Laboratories) following manufacturer's protocol. Data was obtained from Bio-Rad CFX96 Thermal Cycler (Bio-Rad Laboratories) following a three-step protocol: 1 cycle of 95°C 3 min, 40 cycles of 95°C 10 sec and 60°C 30 sec. Data was exported and further processed with customized Python script.

BAT-1 variant luciferase reporter assay vector construction

For each *BAT-1* variant, three pairs of oligos (Supplemental Table S1) were designed with each pair containing a four base pair overlap with either the backbone vector or to an adjacent annealed oligo pair. For instance, F1-R1 pairs overlap with the vector and F2-R2 on either end; F2-R2 pairs overlap with F1-R1 and F3-R3 on either end; F3-R3 pairs overlap with F2-R2 and the vector on either end. The oligos were ordered directly from Integrated DNA Technologies (IDT). Each pair of oligos were annealed and the annealed oligo pairs were then ligated and cloned into the pGL2-3xBAT1-luc backbone vector between the restriction sites NheI and XhoI.

Cell line transient transfection luciferase reporter assays

HEK293T cell luciferase reporter assays were performed as previously described (Zheng et al. 2023). Briefly, cells were transfected following the calcium phosphate transfection protocol in 6-well plates. Experimental plasmids and usage amounts are described in Supplemental Table 1. 48 hours after transfection, cells were harvested, digested, and assayed for luciferase activity using the Dual-Luciferase Reporter Assay System (Promega) following the manufacturer's protocol. Data were collected with a TD-20/20 Luminometer (Turner Designs) and processed with customized Python scripts.

The HEK293T cells (CRL-3216™) were obtained directly from ATCC (American Type Culture Collection) and used within one year of purchase. The cells were tested negative for mycoplasma contamination.

Data analysis

ATAC-seq data analysis

2×150bp reads from Illumina NovaSeq2000 were obtained for all samples with an average depth of 54M reads at Genome Technology Access Center at the McDonnell Genome Institute (GTAC@MGI, WashU). For each sample, reads were first run through Trim Galore (v0.6.1) (Felix Krueger 2023) to remove adapter sequences and then QC by FastQC (v0.11.7) (Andrews 2010). The trimmed reads were then mapped to the mm10 genome using Bowtie2 (v2.3.5) (Langmead and Salzberg 2012) with parameters -X 2000 --very-sensitive. Only uniquely mapped and properly paired reads were retained with samtools (v1.12) (Li et al. 2009) with parameters -f 0x2 -q 30. Mitochondria reads were removed with samtools (v1.12) (Li et al. 2009). Duplicated reads were marked and removed with Picard (v2.21.4) (2019). Reads mapped to the mm10 blacklist regions were removed by bedtools (v2.27.1) (Quinlan and Hall 2010) by intersect -v. Last, filtered reads were sieved to account for Tn5 insertion by deeptools (v3.5.3) (Ramírez et al. 2016) with command alignmentSieve –ATACshift. bigWig files were generated with deeptools (v3.5.3) (Ramírez et al. 2016) with command bamCoverage --binSize 10 -e --normalizeUsing CPM. For each genotype, an average binding intensity bigWig file from three replicates was generated with deeptools (v3.5.3) (Ramírez et al. 2016) command bigwigAverage --binSize 10.

Peak-calling was performed with MACS2 (v2.2.7.1) (Zhang et al. 2008) on individual replicate with command macs2 callpeak --nomodel --keep-dup all. For each genotype, we generated a genotype-specific high confidence peakset by intersection of peaks called in at least two replicates. R package DiffBind (v3.0.15) (Stark and Brown 2012) and DESeq2 (v1.30.1) (Love et al. 2014) were then used to re-center peaks to ±200bp regions surrounding the summit, generate normalized binding matrix, and differential binding matrix. We defined differentially bound peaks between each mutant and WT sample if the absolute $\log_2$FC was more than 2.0, corresponding to four-fold, and the FDR smaller than 1e-3.

To associate ATAC-seq peaks to genes, we used Genomic Regions Enrichment of Annotations Tool (GREAT v4.0.4) through the R package rGREAT (v1.19.2) (Gu and Hübschmann 2022). Each peak was assigned to the closest transcription start site (TSS) within 50kb. Each gene can have one or more associated ATAC-seq defined cis-regulatory element (CRE). Overlap of ATAC-seq peaks and CRX ChIP-seq peaks were identified by pybedtools (v0.9.1) (Dale et al. 2011) with BedTool intersect function. The compiled ATAC-seq intensity matrices can be found in the GitHub repository of this paper.

De novo motif searching

The mm10 FASTA sequences for each genotype specific peaks were obtained using R package BSgenome (v1.66.3) (Pagès 2020). *De novo* motif enrichment analysis for each set of sequences were then performed with MEME-ChIP in MEME Suite (v5.5.2) (Bailey et al. 2015) using order 1 Markov background model and default parameters. Since homeodomain motifs are relatively short and can be repetitive (e.g. K88N motif), we reported STREME found motifs, which is more sensitive than MEME to find short, repetitive motifs. In Fig. 4E, PWMs for selected retinal basic helix-loop-helix (bHLH) factors were obtained directly from the JASPAR2024 website (Rauluseviciute et al. 2023). The accession numbers are: NEUROD1 (MA1109.1), ASCL1 (MA1100.1), ATOH7 (MA1468.1). Note the ASCL1 PWM was padded at the first position for pattern alignment. All PWMs plotted can be found in Supplemental Table S4.

Dimeric K_{50} HD motif half-site affinity prediction

For the analysis presented in Supplemental Fig. S3A, we first identified dimeric K_{50} HD motifs under $Crx^{E80A/A}$-reduced ATAC-seq peak sequences by FIMO (Grant et al. 2011) searching (--threshold 1.0E-3) using the PWM model identified in *de novo* motif searching (section above). The identified dimeric K_{50} HD motif

instances were further filtered by constraining the HD half-site core motif to match 5'-NAAN-3'. For each dimeric motif, the relative binding affinity of each 6mer half-site (Fig. 1 bottom), normalized to the monomeric consensus 5'-TAATCC-3', was calculated using a published CRX PWM model assuming position independence. The PWM models used in FIMO search and relative binding affinity calculation can be found in Supplemental Table S5.

HD motif type annotation for CRX bound CREs adjacent to CRX-DAGs

The list of CRX dependent-activated genes (CRX-DAGs) and their annotations were taken directly from our previous publication (Zheng et al. 2023). CRX-DAG associated CRX ChIP-seq peaks were scanned for instances of K_{50} and Q_{50} HD monomeric and dimeric motifs with FIMO in MEME Suite (v5.5.2) using order 1 Markov background model and --thresh 1.0E-3. PWMs used for FIMO search can be found in Supplemental Table S5. A published CRX PWM model (Corbo et al. 2010) was used to scan for K_{50} HD monomeric motifs, and a MEME found pattern under the $Crx^{E80A/A}$-reduced ATAC-seq peaks was used to scan for K_{50} HD dimeric motifs. The RAX2 JASPAR motif MA0717.1 was used to scan for Q_{50} HD monomeric motifs, and a MEME found pattern under the $Crx^{K88N/N}$-increased ATAC-seq peaks was used to scan for Q_{50} HD dimeric motifs. FIMO found dimeric HD motif instances were further filtered by constraining the HD half-site core motif to match 5'-NAAN-3'. To distinguish standalone monomeric motifs and high-affinity half-sites within dimeric motifs, we first searched for dimeric motif instances, masked to N, and re-run motif search for monomeric motif instances with the masked sequences. Each CRX bound CRE associated with a CRX-DAG was then annotated to have presence (1) or absence (0) of a K_{50} or Q_{50} HD motif. The full HD motif type annotation for CRX-DAG-adjacent CREs can be found in Supplemental Table S6.

Gene ontology analysis

Gene ontology analysis in Fig. 3D and 4F were performed using R package clusterProfiler (v3.18.1) (Yu et al. 2012; Wu et al. 2021) with the genome wide annotation package org.Mm.eg.db (v3.15.0) (Carlson 2019). For the enrichment analysis in Fig. 3D, a $\log_2FC<-1$ and fdr<0.05 cutoff was used to identify genes that display concordant reduction in expression and in chromatin accessibility at their nearby cis-regulatory regions. Enrichment p-values were adjusted by Benjamini-Hochberg procedure. Redundantly enriched GO terms were removed using simplify() function with parameters cutoff=0.7, by=p.adjust. The enrichment analysis results were then exported in table format and further processed for plotting with Python.

Re-analysis of published data

The CRX ChIP-seq and bulk RNA-seq data for the WT, Crx^{E80A}, Crx^{K88N}, Crx^{R90W} mouse retinas from our previous publication (Zheng et al. 2023) were used without alteration. To generate the Fig. 3E heatmap, the differential gene expression matrix was ordered by hierarchical clustering using the Python package SciPy (v1.11.2) (Virtanen et al. 2020) with linkage method "complete" and distance metric "euclidean".

The developmental time-course bulk ATAC-seq and RNA-seq datasets from the Aldiri et al. study (Aldiri et al. 2017) were obtained from GEO under accession number GSE87064. Replicates are not available for the ATAC-seq data. Otherwise, the ATAC-seq reads were processed similarly to ATAC-seq data generated in this study. The Crx^{E80A} or Crx^{K88N} consensus ATAC-seq peakset was used directly to score signal enrichment of the Aldiri ATAC-seq data (Aldiri et al. 2017). For Fig. 3C, 4C, 5D,E, accessibility z-scores were calculated using only the post-natal ages data (P0, 3, 7, 10, 14) with sklearn.preprocessing.StandardScaler from the scikit-learn package (v1.3.0) (Pedregosa et al. 2011). The Aldiri RNA-seq data (Supplemental Fig. S4A,B) re-processed in our previous publication (Zheng et al. 2023) were used without alteration.

The VSX2 ChIP-seq data from the Bian et al. study (Bian et al. 2022) was obtained from GEO under accession number GSE196106. The reads were processed similarly as the CRX ChIP-seq data described in our

previous study (Zheng et al. 2023). To identify embryonic day 14.5 (E14.5) enriched, adult enriched, and shared VSX2 binding regions used in Fig. 4D, a consensus peakset was first generated and subjected to non-supervised hierarchical clustering using the Python package fastcluster (v1.2.6) (Müllner 2013) with parameters method=single, metric=euclidean. Overlap of Crx^{K88N} differential ATAC-seq peaks and VSX2 ChIP-seq peaks were identified by pybedtools (v0.9.1) (Dale et al. 2011) with BedTool intersect function.

MPRA data preprocessing

2×150bp reads from Illumina NovaSeq2000 were obtained for all samples at the Genome Technology Access Center at the McDonnell Genome Institute (GTAC@MGI, WashU). Libraries prepared from un-electroporated plasmid DNAs were sequenced to 50M reads in two technical replicates (approx. 2500×). Libraries prepared from retinal *ex plant* extracted RNA (n=3/4 per genotype) and DNA (n=1 per genotype) were sequenced to an average depth of 30M reads (approx. 1600× depth). The pre-processing of MPRA reads followed the pipeline previously described (Friedman et al. 2021). Briefly, the sequencing reads were demultiplexed using the inner phasing index within the P1 adapter sequences. Next, the 10bp unique CRE identifying barcodes were extracted and counted with customized Python scripts. The compiled MPRA CRE barcode count matrix and annotation table can be found in the GitHub repository for this paper.

MPRA regulatory activity calculation

By design, each testing CRE was represented by four unique barcodes (Supplemental Fig. S6B). One or more of these barcoded CREs can drop out in the process of cloning or due to low activity. To ensure quality measurement, we dropped 1) barcodes with less than 50 reads in both plasmid libraries; 2) CREs represented by less than 3 unique barcodes after step 1). Then, we normalized the reads from the plasmid libraries by the counts-per-million paradigm without adding pseudocount. Reads from the retinal *ex plant* extracted RNA and DNA libraries were normalized similarly by adding 1 pseudocount to each barcode. Counts of individual RNA libraries were normalized by the average counts of the plasmid libraries to obtain "raw activity score" for individual barcodes. Average raw activity scores by unique barcodes for individual genotypes were calculated by averaging the genotype replicates. Then, average raw activity scores by unique testing CREs for individual genotypes were calculated by averaging the barcode raw activity scores for each CRE. A coefficient of variation threshold of 1.0 was used to filter out CREs whose barcode activity varies greatly among genotype replicates. Last, to calculate the "regulatory activity score", the raw activity score of each CRE was normalized to the average raw activity score of all scrambled control CREs.

MPRA HD motif activity analysis

In the MPRA library design, each testing genomic CRE sequence was scanned for the presence of monomeric and dimeric K_{50} and Q_{50} HD motifs following the same pipeline as HD motif type annotation for CRX-bound CREs (section above). A less stringent --thresh 2.5E-3 for FIMO search was used to capture more non-consensus motifs. Both FIMO-found monomeric and dimeric HD motif instances were further filtered by constraining the HD monomeric site or half-site core motif to match 5'-TAA-3' at the first three positions. If a testing genomic CRE sequence (WT) contains HD motif(s), one or more mutated CRE versions were generated: 1) mutM: all and only the monomeric HD motifs mutated; 2) mutD: all and only the dimeric HD motifs mutated; 3) mutDM: all monomeric and dimeric HD motifs mutated if both are present. To minimize perturbations of nearby/overlapping TF motifs, a single substitution 5'-TAA-3' to 5'-TAC-3' was introduced to disrupt HD motifs. This substitution has been found in EMSA to abolish CRX HD binding to DNA (Lee et al. 2010). K_{50} or Q_{50} type HD motifs were not differentiated due to the complexity of library design and the differential binding of CRX WT vs K88N proteins on these two types of motifs. The "HD motif activity score" (Fig. 7B,C) was calculated as the "regulatory activity score" difference between a mutant CRE version and its matched genomic

CRE sequence. The HD motif position information and compiled HD motif activity score matrix can be found in the GitHub repository for this paper.

Statistical analysis

One-way ANOVA with Turkey honestly significant difference (HSD) tests in Fig. 1E and Supplemental Fig. S1D were performed with Python packages SciPy (v1.11.2) (Virtanen et al. 2020) and scikit_posthocs (v0.7.0) (Terpilowski 2019). Fisher's exact test in Fig. 4D was performed with Python package SciPy (v1.11.2) (Virtanen et al. 2020). Mann-Whitney-Wilcoxon tests in Fig. 7B,C were performed with SciPy through Python package statannotations (v0.4.4) (Charlier et al. 2022).

KEY RESOURCES TABLE

Reagent or Resource	Source	Identifier
Gene		
CRX (Homo sapiens)	HGNC	HGNC: 2383
Crx (Mus musculus)	MGI	MGI: 1194883
Experimental models: Organisms/strains		
BL21 (DE3) / *Escherichia coli*	MilliporeSigma	CMC0016
DH5α / *Escherichia coli*	New England Biolabs	C2987H
WT (C57BL/6J) / *Mus musculus*	The Jackson Laboratory	Cat #000664
Crx^{E80A} (C57BL/6J) / *Mus musculus*	Zheng et al. 2023 (Zheng et al. 2023)	
Crx^{K88N} (C57BL/6J) / *Mus musculus* J	Zheng et al. 2023 (Zheng et al. 2023)	
Crx^{R90W} (C57BL/6J) / *Mus musculus*	Tran et al. 2014 (Tran et al. 2014)	
HEK293T / *Homo sapiens*	ATCC	CRL-3216
Chemicals		
Chloroform	Fisher Scientific	C298-500
Collagenase from *Clostridium histolyticum*	Millipore Sigma	C0130
Digitonin	Promega	G9441
Dithiothreitol (DTT)	Bio-Rad Laboratories	1610611
DNase I (RNase-free)	New England Biolabs	M0303S
Exonuclease I (E. coli)	New England Biolabs	M0293S
Gibco™ Fetal Bovine Serum (FBS), Premium	ThermoFisher Scientific	A5670701
Gibco™ Dulbecco's Modified Eagle Medium	ThermoFisher Scientific	11965084
Gibco™ DMEM/F-12	ThermoFisher Scientific	11320033
Glutathione Sepharose 4B resin	Cytiva	17075601
Isopropyl-β-D-thiogalactopyranoside (IPTG)	ThermoFisher Scientific	BP1755-10

Molecular Biology Grade Water	Corning®	46-000-CM
Nonidet-P40	USBiology	CAS: 9036-19-5
Ambion® Nuclease-Free Water	Invitrogen	AM9937
Penicillin-Streptomycin	ThermoFisher Scientific	15140122
Phosphate Buffered Saline	Corning®	46-013-CM
Roche cOmplete™, Mini Protease Inhibitor Cocktail	Millipore Sigma	11836153001
SYBR™ Green I Nucleic Acid Gel Stain	Invitrogen	S7563
Thrombin protease	Millipore Sigma	GE27-0846-01
Triton X-100	Sigma-Aldrich	T9284
TRIzol™ Reagent	Invitrogen	15596026
TWEEN® 20	Millipore Sigma	P9416
Critical commercial assays		
Agilent DNA 1000 KitT	Agilent	5067-1504
Amicon Ultra-0.5 Centrifugal Filter Unit	Millipore	UFC500324
AMPure XP beads	Beckman Coulter	A63880
Dual-Luciferase Reporter Assay System	Promega	E1910
GST SpinTrap™	Cytiva	28952359
Illumina Tagment DNA TDE1 Enzyme and Buffer Kits	Illumina	Cat #20034211
iScript™ Reverse Transcription Supermix	Bio-Rad Laboratories	1708841
MinElute PCR Purification Kit	QIAGEN	28006
Monarch® PCR & DNA Cleanup Kit (5 µg)	New England Biolabs	T1030S
NEBNext® High-Fidelity 2X PCR Master Mix	New England Biolabs	M0541S
Novex™ WedgeWell™ 12%	Invitrogen	XP00122BOX

Tris-Glycine Mini Protein Gels		
Phusion® High-Fidelity PCR Master Mix with HF Buffer	New England Biolabs	M0531S
Q5® High-Fidelity 2X Master Mix	New England Biolabs	M0492S
Qubit™ dsDNA Quantification Assay Kits	Invitrogen	Q32851
RNA Clean & Concentrator-5	ZYMO Research	R1013
SsoFast™ EvaGreen® Supermix with Low ROX	Bio-Rad Laboratories	1725211
TURBO DNA-free™ Kit	Invitrogen	AM1907
Deposited data		
Coop-seq - in vitro	This study	GEO: GSE256213
ATAC-seq - mouse retina	This study	GEO: GSE256212
MPRA – mouse retina ex plant	This study	GEO: GSE256214
Spec-seq – in vitro	Zheng et al. 2023 (Zheng et al. 2023)	GEO: GSE223658
CRX ChIP-seq - mouse retina	Zheng et al. 2023 (Zheng et al. 2023)	GEO: GSE223657
RNA-seq - mouse retina	Zheng et al. 2023 (Zheng et al. 2023)	GEO: GSE223439
ATAC-seq - mouse retina	Aldiri et al. 2017 (Aldiri et al. 2017)	GEO: GSE87064
RNA-seq - mouse retina	Aldiri et al. 2017 (Aldiri et al. 2017)	GEO: GSE87064
VSX2 ChIP-seq – mouse retina	Bian et al. 2022 (Bian et al. 2022)	GEO: GSE196106
Equipment		
Bioanalyzer	Agilent	2100
Electroporation chamber	NEPA GENE Co. Ltd	CUY520P5
Electro Square Porator™	BTX	ECM 830
Nuclepore™ Track-Etch Membrane	Cytiva	10417006

Odyssey® CLx and Fc Imaging Systems	LI-COR Biosciences	CLX 1894
PRO250 Homogenizer	PRO Scientific Inc.	01-01250
5mm X 75mm Flat Bottom Generator Probe	PRO Scientific Inc.	02-05075
Software and algorithms		
bedtools (v2.27.1)	Quinlan and Hall, 2010 (Quinlan and Hall 2010)	https://bedtools.readthedocs.io/en/latest/
Bowtie2 (v2.3.5)	Langmead and Salzberg, 2012 (Langmead and Salzberg 2012)	https://bowtie-bio.sourceforge.net/bowtie2/index.shtml
BSgenome (v1.66.3)	Pagès, 2020 (Pagès 2020)	https://bioconductor.org/packages/BSgenome
clusterProfiler (v3.18.1)	Yu et al., 2012 (Yu et al. 2012); Wu et al., 2021 (Wu et al. 2021)	https://bioconductor.org/packages/clusterProfiler
DAVID (v6.8)	Sherman et al., 2022 (Sherman et al. 2022)	https://david.ncifcrf.gov/
deeptools (3.5.3)	Ramírez et al., 2016 (Ramírez et al. 2016)	https://deeptools.readthedocs.io/en/develop/
DESeq2 (v1.30.1)	Love et al., 2014 (Love et al. 2014)	https://bioconductor.org/packages/DESeq2
DiffBind (v3.0.15)	Stark and Brown, 2011 (Stark and Brown 2012); Ross-Innes et al., 2012 (Ross-Innes et al. 2012)	https://bioconductor.org/packages/DiffBind
fastcluster (v1.2.6)	Müllner, 2013	http://danifold.net/fastcluster.html
FastQC (v0.11.7)	Andrews (Andrews 2010)	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
GREAT (v4.0.4)	McLean et al., 2010 (McLean et al. 2010)	http://great.stanford.edu/public/html/
logomaker (v0.8)	Tareen and Kinney, 2020 (Tareen and Kinney 2019)	https://logomaker.readthedocs.io/en/latest/
MACS2 (v2.2.7.1)	Zhang et al., 2008 (Zhang et al. 2008)	https://github.com/macs3-project/MACS
matplotlib (v3.7.2)	Hunter, 2007 (Hunter 2007)	https://matplotlib.org/
MEME Suite (v5.5.2)	Bailey et al., 2015 (Bailey et al. 2015)	https://meme-suite.org/meme/index.html
numpy (v1.23.5)	Harris, Millman, van der Walt et al., (2020) (Harris et al. 2020)	https://numpy.org/
org.Mm.eg.db (v3.15.0)	Carlson, 2019 (Carlson 2019)	https://bioconductor.org/packages/org.Mm.eg.db
pandas (v2.1.1)	Reback, 2022 (Jeff Reback 2022)	https://pandas.pydata.org/

Picard (v2.21.4)	Picard toolkit (2019)	http://broadinstitute.github.io/picard/
pybedtools (v0.9.1)	Dale et al., 2011 (Dale et al. 2011)	https://daler.github.io/pybedtools/
Python (v3.9.12)	Rossum and Drake, 1995 (Van Rossum 1995)	https://docs.python.org/3/reference/
R (v4.0.3) for DiffBind	R Core Team 2020 (RCoreTeam 2020)	https://www.R-project.org
R (v4.2.2) for MPRA analysis	R Core Team 2022 (RCoreTeam 2022)	https://www.R-project.org
rGREAT (v1.19.2)	Gu and Huebschmann, 2022 (Gu and Hübschmann 2022)	https://bioconductor.org/packages/rGREAT
samtools (v1.12)	Li et al., 2009 (Li et al. 2009)	http://www.htslib.org/
scikit-learn (v1.3.0)	Pedregosa et al., 2011 (Pedregosa et al. 2011)	https://scikit-learn.org/
scikit_posthocs (v0.7.0)	Terpilowski, 2019 (Terpilowski 2019)	https://scikit-posthocs.readthedocs.io/en/latest/
SciPy (v1.11.2)	Virtanen et al., 2020 (Virtanen et al. 2020)	https://scipy.org/
seaborn (v0.12.2)	Waskom, 2021 (Waskom 2021)	https://seaborn.pydata.org/
statannotations (v0.4.4)	Charlier et al., 2022 (Charlier et al. 2022)	https://github.com/trevismd/statannotations
Trim Galore (v0.6.1)	Krueger et al., 2023 (Felix Krueger 2023)	https://github.com/FelixKrueger/TrimGalore

REFERENCES

2019. Picard toolkit. *Broad Institute, GitHub repository*.

Aldiri I, Xu B, Wang L, Chen X, Hiler D, Griffiths L, Valentine M, Shirinifard A, Thiagarajan S, Sablauer A et al. 2017. The Dynamic Epigenetic Landscape of the Retina During Development, Reprogramming, and Tumorigenesis. *Neuron* **94**: 550-568 e510.

Andrews S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].

Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME Suite. *Nucleic Acids Research* **43**: W39-W49.

Bian F, Daghsni M, Lu F, Liu S, Gross JM, Aldiri I. 2022. Functional analysis of the Vsx2 super-enhancer uncovers distinct cis-regulatory circuits controlling Vsx2 expression during retinogenesis. *Development* **149**.

Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109**: 21.29.21-21.29.29.

Carlson M. 2019. org.Mm.eg.db: Genome wide annotation for Mouse.

Charlier F, Weber M, Izak D, Harkin E, Magnus M, Lalli J, Fresnais L, Chan M, Markov N, Amsalem O et al. 2022. Statannotations. doi:10.5281/zenodo.7213391. Zenodo.

Chen S, Wang Q-L, Xu S, Liu I, Li LY, Wang Y, Zack DJ. 2002. Functional analysis of cone–rod homeobox (CRX) mutations associated with retinal dystrophy. *Human Molecular Genetics* **11**: 873-884.

Corbo JC, Lawrence KA, Karlstetter M, Myers CA, Abdelaziz M, Dirkes W, Weigelt K, Seifert M, Benes V, Fritsche LG et al. 2010. CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Res* **20**: 1512-1525.

Dale RK, Pedersen BS, Quinlan AR. 2011. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**: 3423-3424.

Felix Krueger FJ, Phil Ewels, Ebrahim Afyounian, Michael Weinstein, Benjamin Schuster-Boeckler, Gert Hulselmans, sclamons. 2023. FelixKrueger/TrimGalore: v0.6.9 - fix declaration bug (0.6.9). *Zenodo* doi:10.5281/zenodo.7581188.

Friedman RZ, Granas DM, Myers CA, Corbo JC, Cohen BA, White MA. 2021. Information content differentiates enhancers from silencers in mouse photoreceptors. *eLife* **10**: e67403.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017-1018.

Gu Z, Hübschmann D. 2022. rGREAT: an R/bioconductor package for functional enrichment on genomic regions. *Bioinformatics* **39**.

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ et al. 2020. Array programming with NumPy. *Nature* **585**: 357-362.

Hughes AEO, Myers CA, Corbo JC. 2018. A massively parallel reporter assay reveals context-dependent activity of homeodomain binding sites in vivo. *Genome Res* **28**: 1520-1531.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**: 90-95.

Jeff Reback j, Wes McKinney, Joris Van den Bossche, Tom Augspurger, Matthew Roeschke, Simon Hawkins, Phillip Cloud, gfyoung, Sinhrks, Patrick Hoefler, Adam Klein, Terji Petersen, Jeff Tratner, Chang She, William Ayd, Shahar Naveh, JHM Darbyshire, Marc Garcia, Richard Shadrach, Jeremy Schendel, Andy Hayden, Daniel Saxton, Marco Edward Gorelli, Fangchen Li, Matthew Zeitlin, Vytautas Jancauskas, Ali McMaster, Torsten Wörtwein, Pietro Battiston. 2022. pandas-dev/pandas: Pandas 1.4.2. doi:10.5281/ZENODO.6408044.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357-359.

Lee J, Myers CA, Williams N, Abdelaziz M, Corbo JC. 2010. Quantitative fine-tuning of photoreceptor cis-regulatory elements through affinity modulation of transcription factor binding sites. *Gene Therapy* **17**: 1390-1399.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

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

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**: 495-501.

Müllner D. 2013. fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *Journal of Statistical Software* **53**: 1 - 18.

Pagès H. 2020. BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**: 2825−2830.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44**: W160-W165.

Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R, Castro-Mondragon Jaime A, Ferenc K, Kumar V, Lemma RB, Lucas J, Chèneby J, Baranasic D et al. 2023. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **52**: D174-D182.

RCoreTeam. 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

RCoreTeam. 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, Brown GD, Gojis O, Ellis IO, Green AR et al. 2012. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**: 389-393.

Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T, Chang W. 2022. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res* **50**: W216-221.

Stark R, Brown G. 2012. DiffBind: Differential binding analysis of ChIP-Seq peak data.

Tareen A, Kinney JB. 2019. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**: 2272-2274.

Terpilowski MA. 2019. scikit-posthocs: Pairwise multiple comparison tests in Python. *Journal of Open Source Software* **4**: 1169.

Tran NM, Zhang A, Zhang X, Huecker JB, Hennig AK, Chen S. 2014. Mechanistically Distinct Mouse Models for CRX-Associated Retinopathy. *PLOS Genetics* **10**: e1004111.

Van Rossum GaDJ, Fred L. 1995. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Virtanen P Gommers R Oliphant TE Haberland M Reddy T Cournapeau D Burovski E Peterson P Weckesser W Bright J et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**: 261-272.

Waskom ML. 2021. seaborn: statistical data visualization. *Journal of Open Source Software* **6**: 3021.

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L et al. 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**: 100141.

Yu G, Wang LG, Han Y, He QY. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* **16**: 284-287.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**: R137.

Zheng Y, Sun C, Zhang X, Ruzycki PA, Chen S. 2023. Missense mutations in CRX homeodomain cause dominant retinopathies through two distinct mechanisms. *eLife* **12**: RP87147.