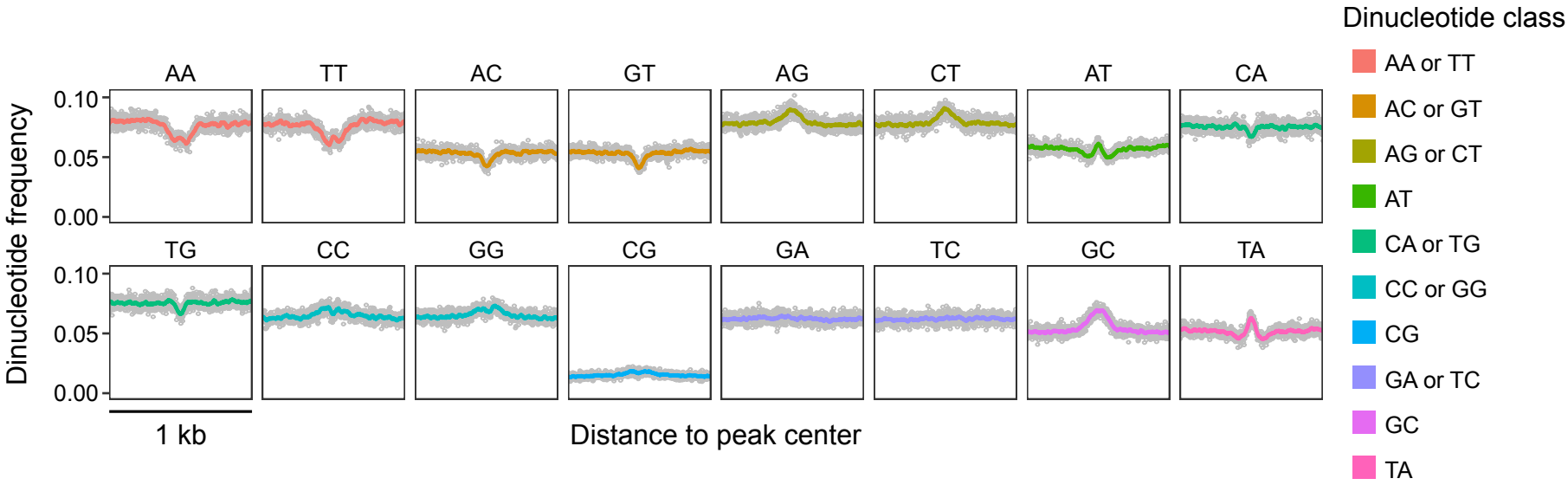
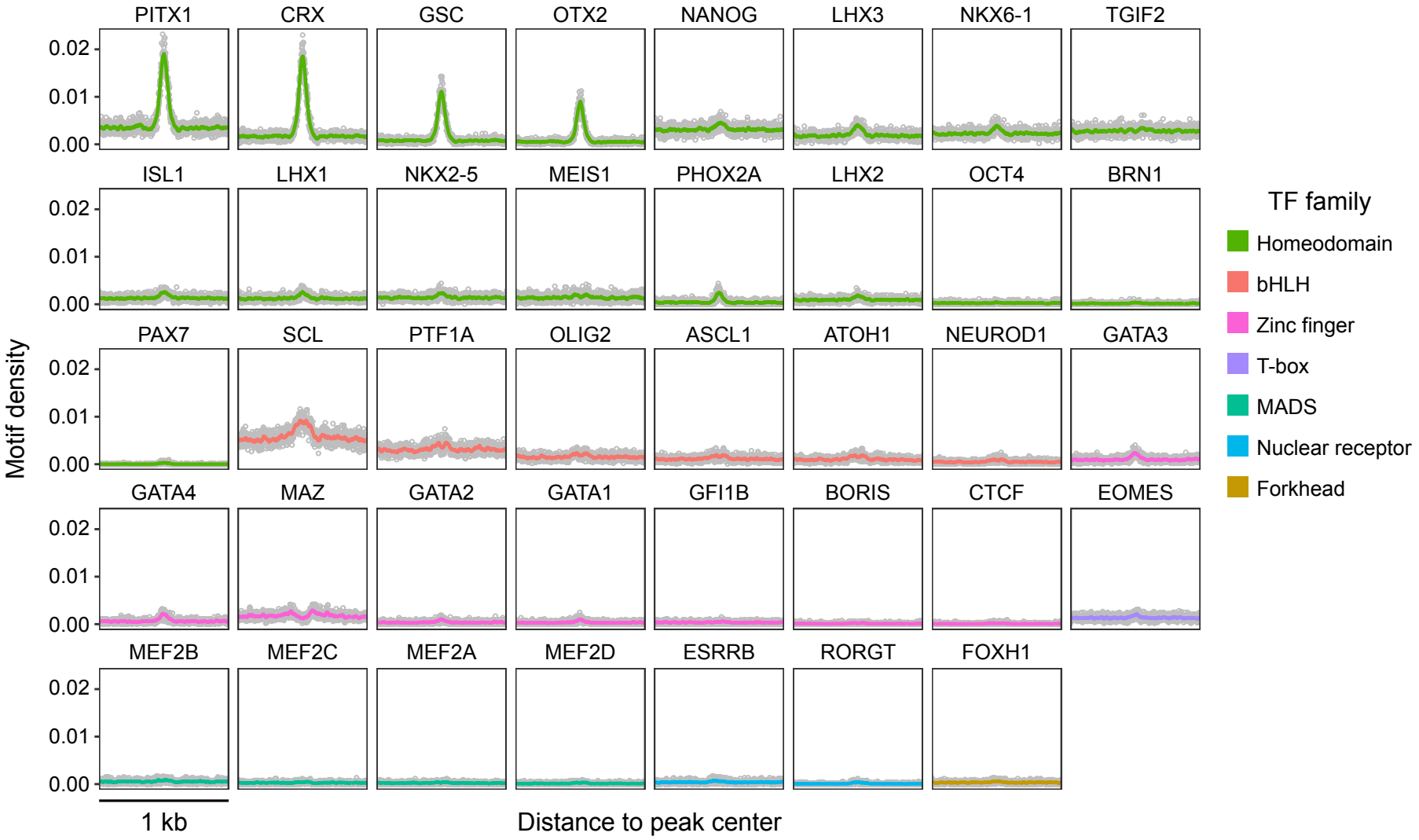
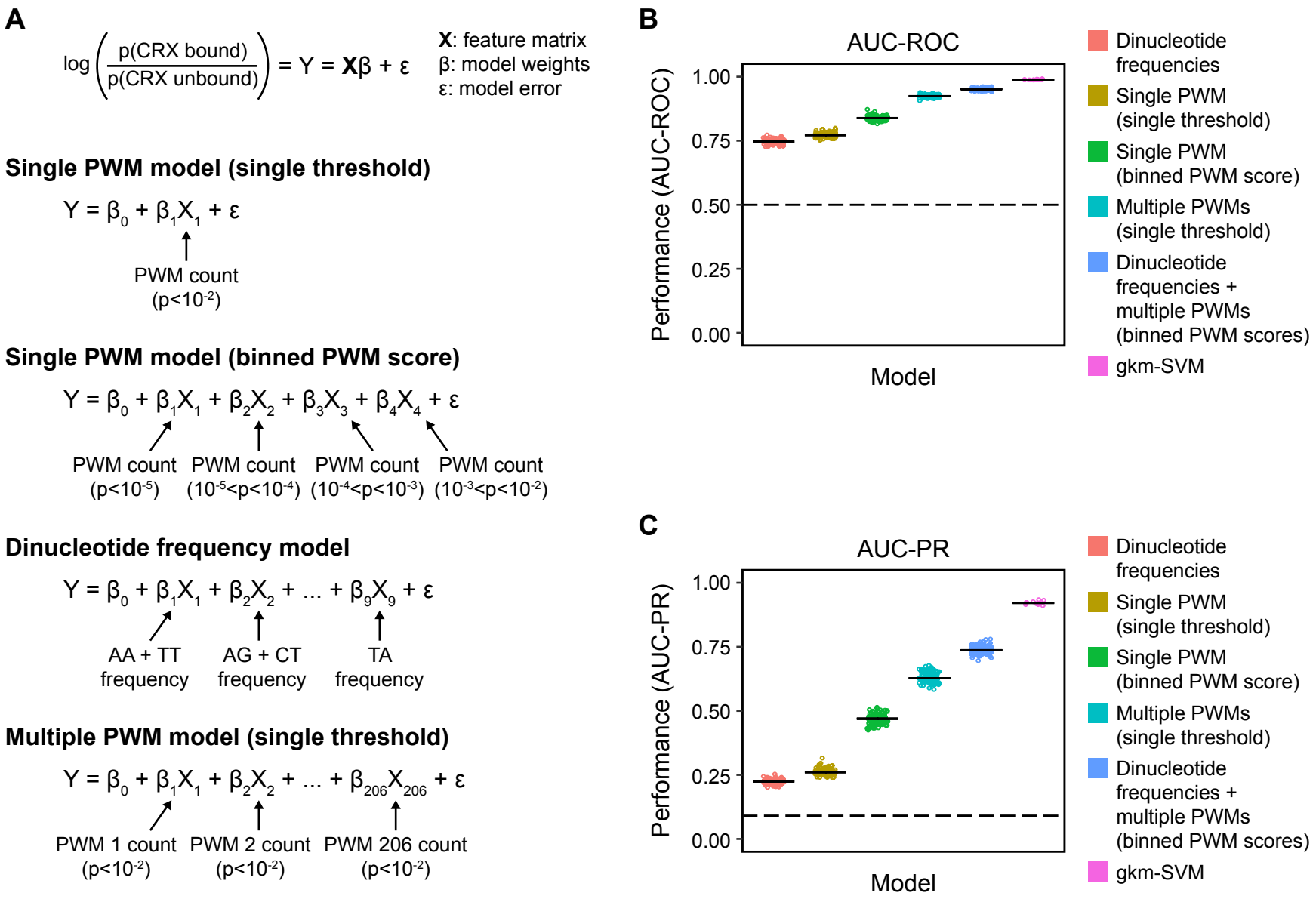
SUPPLEMENTAL MATERIAL



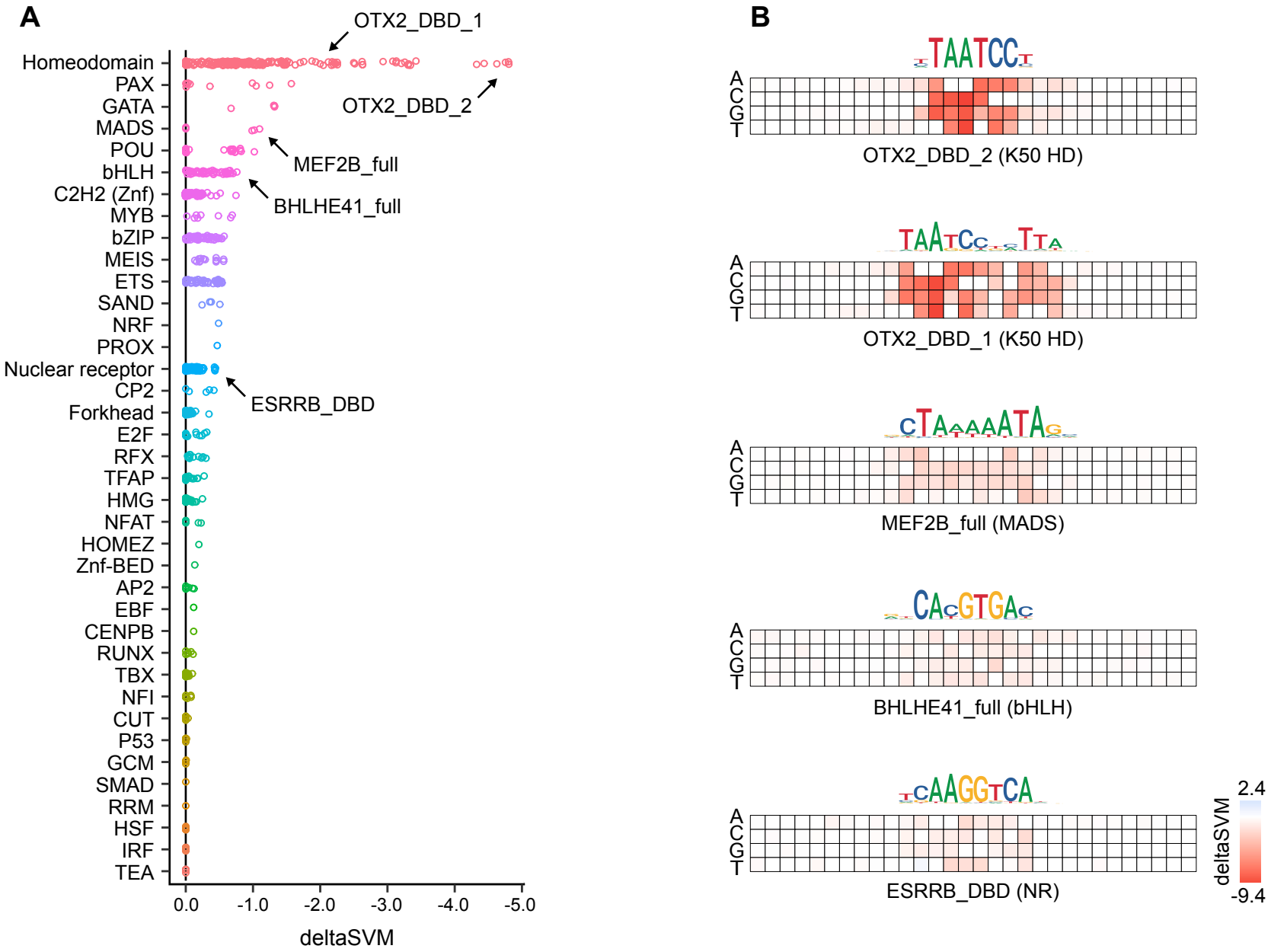
Supplemental Fig. 1. Dinucleotide profiles of CRX bound regions. Average dinucleotide frequencies per base pair per peak (gray dots) with a 25 bp rolling mean (colored lines) in a 1 kb window centered on TSS-distal (>1 kb upstream or >100 bp downstream of an annotated TSS) CRX ChIP-seq peaks (n=5,250).



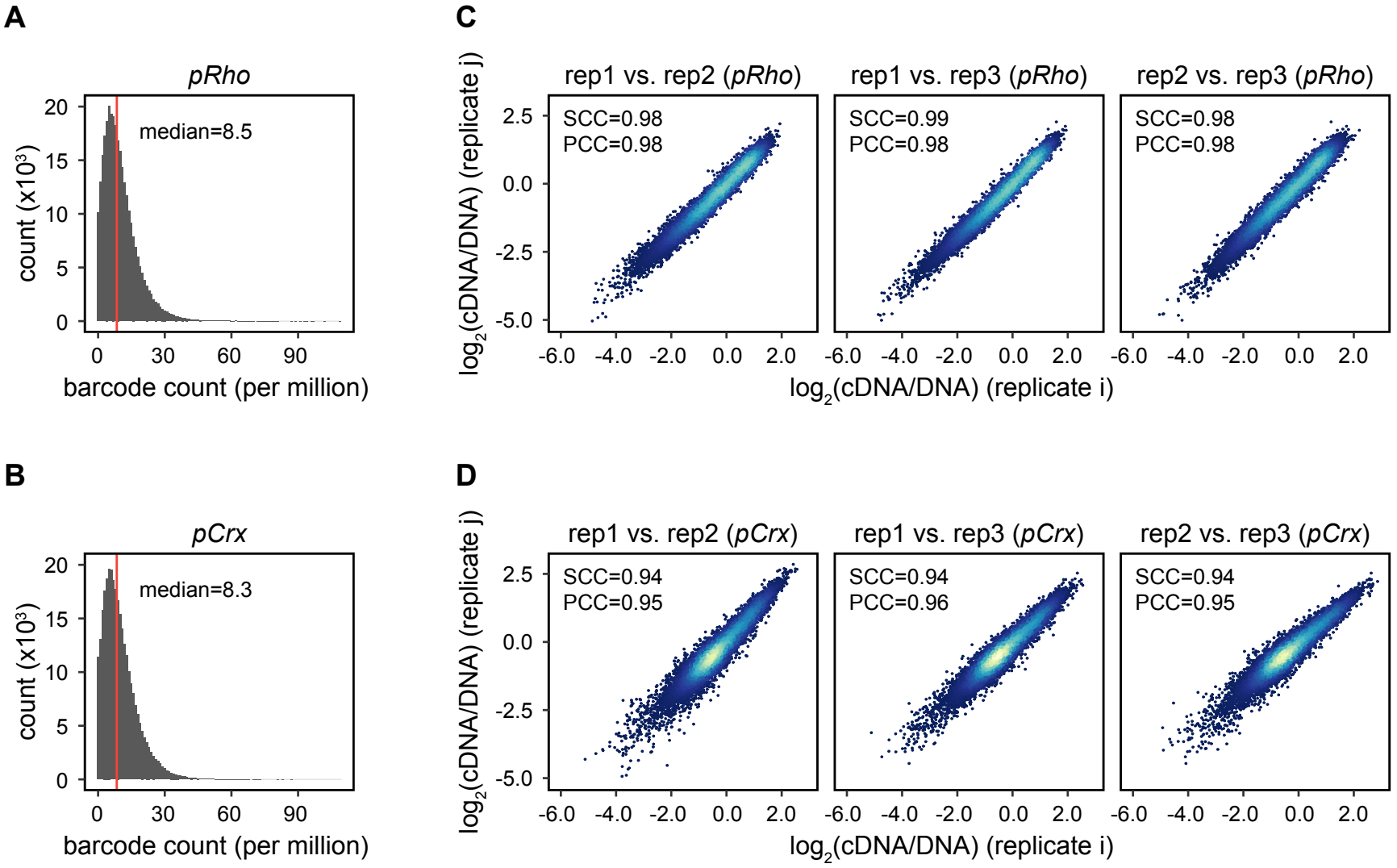
Supplemental Fig. 2. TF binding site density of CRX bound regions. Average number of TF binding sites per base pair per peak (gray dots) with a 25 bp rolling mean (colored lines) in a 1 kb window centered on TSS-distal (>1 kb upstream or >100 bp downstream of an annotated TSS) CRX ChIP-seq peaks (n=5,250). Motif enrichment was calculated for 319 known motifs curated by the HOMER suite of sequence analysis tools (Heinz et al. 2010). TF binding site densities are shown for motifs with enrichment p-values less than 10^{-10}.



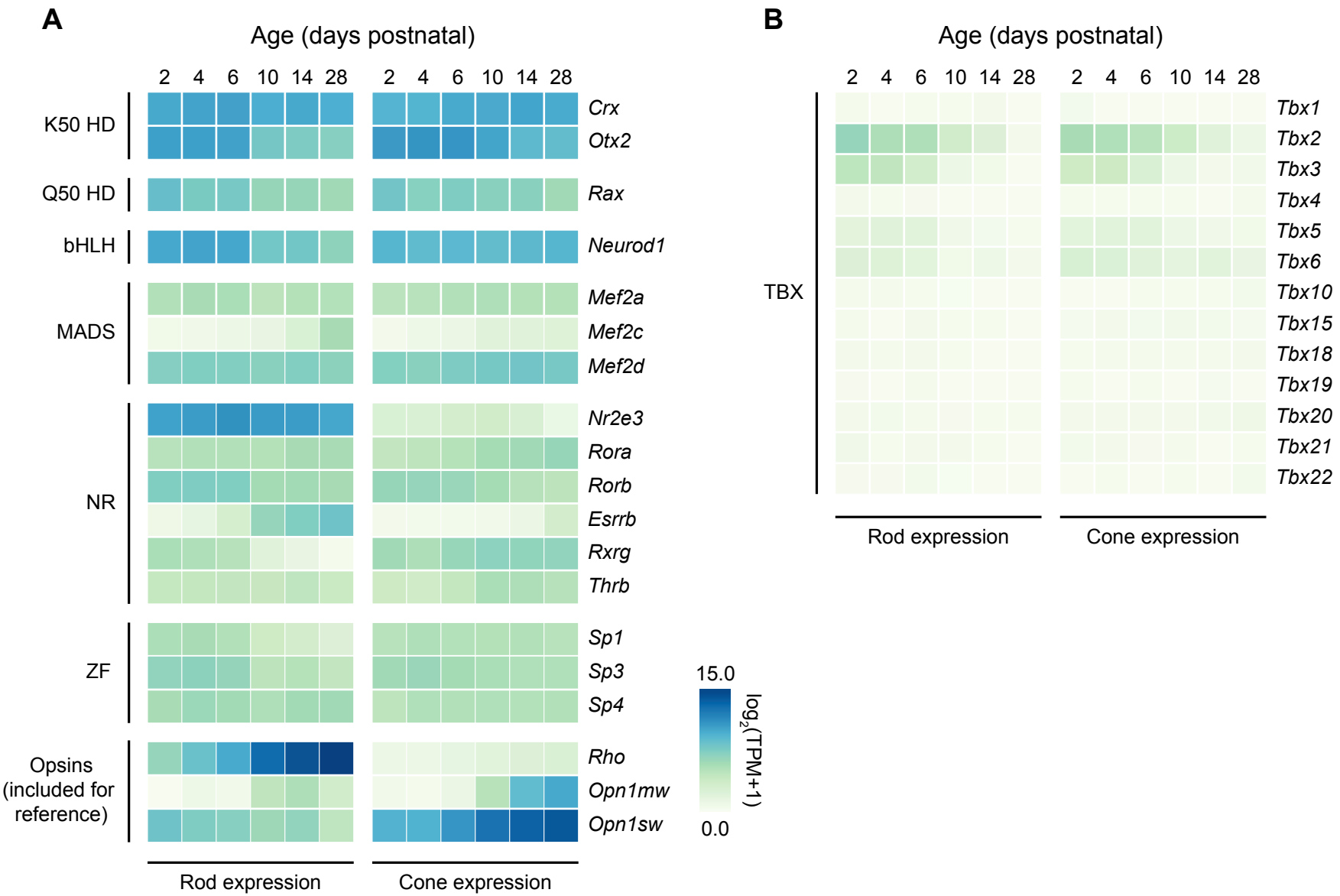
Supplemental Fig. 3. Prediction of CRX-bound regions from primary sequence features. A) Overview of logistic regression models using primary sequence features to classify CRX-bound (n=5,250) vs. CRX-unbound (n=52,500) regions. **B)** Performance of specific models as measured by area under the receiver 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.



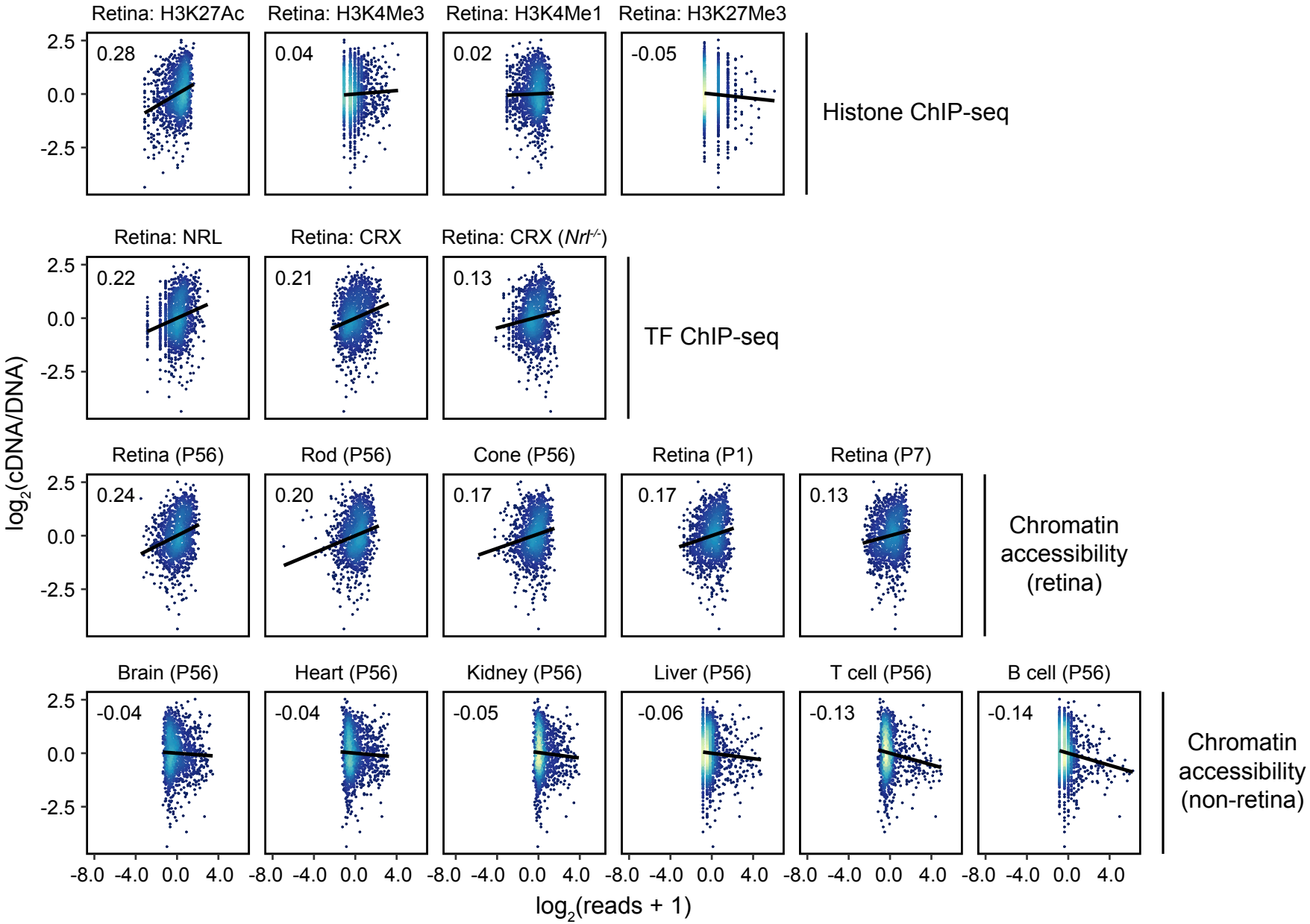
Supplemental Fig. 4. deltaSVM scores overlapping TF binding sites. **A)** Median deltaSVM score for substitutions overlapping instances of TF binding sites within CRX ChIP-seq peaks identified by 843 distinct PWMs (Jolma *et al.* 2013). Predicted substitution effects (deltaSVM scores) were calculated using the gkm-SVM model trained on CRX ChIP-seq data. Each point represents a specific PWM, and PWMs are grouped by TF family along the y-axis. More negative deltaSVM scores correspond to larger decreases in predicted CRX occupancy. **B)** Heat maps illustrating the median predicted effects (deltaSVM scores) of specific substitutions at positions centered selected TF binding sites (corresponding to labeled points in A). X-axis: nucleotide position relative to the motif indicated by the aligned logo. Y-axis: each possible nucleotide substitution. Tile color indicates median deltaSVM score for the indicated substitution at each position.



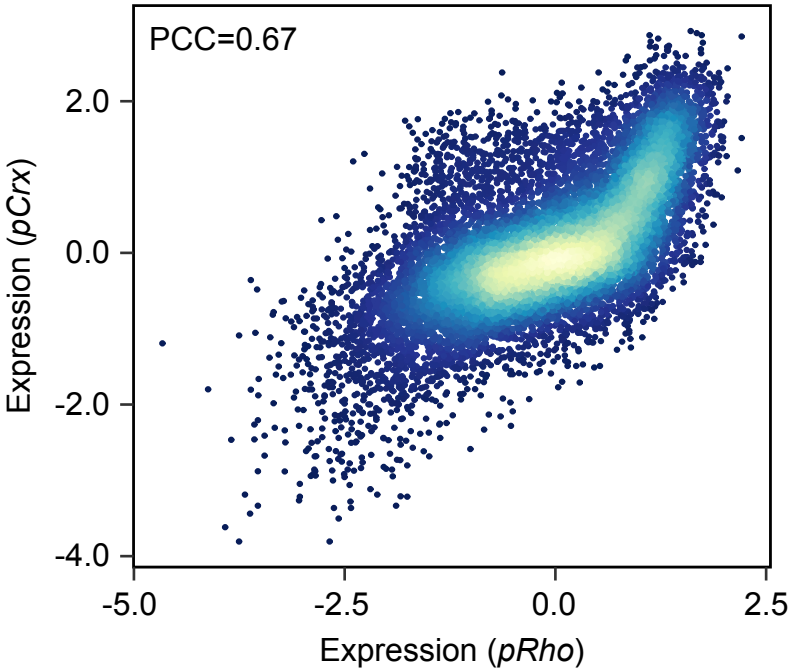
Supplemental Fig. 5. CRE-seq library complexity and reproducibility. A-B) Histogram of average barcode counts (per million barcodes) from CRE-seq DNA libraries (3 replicates each for pRho and pCrx). **C-D)** Scatter plots of CRE-seq activity for pairs of biological replicates. Points are colored by density. PCC: Pearson correlation coefficient. SCC: Spearman correlation coefficient.



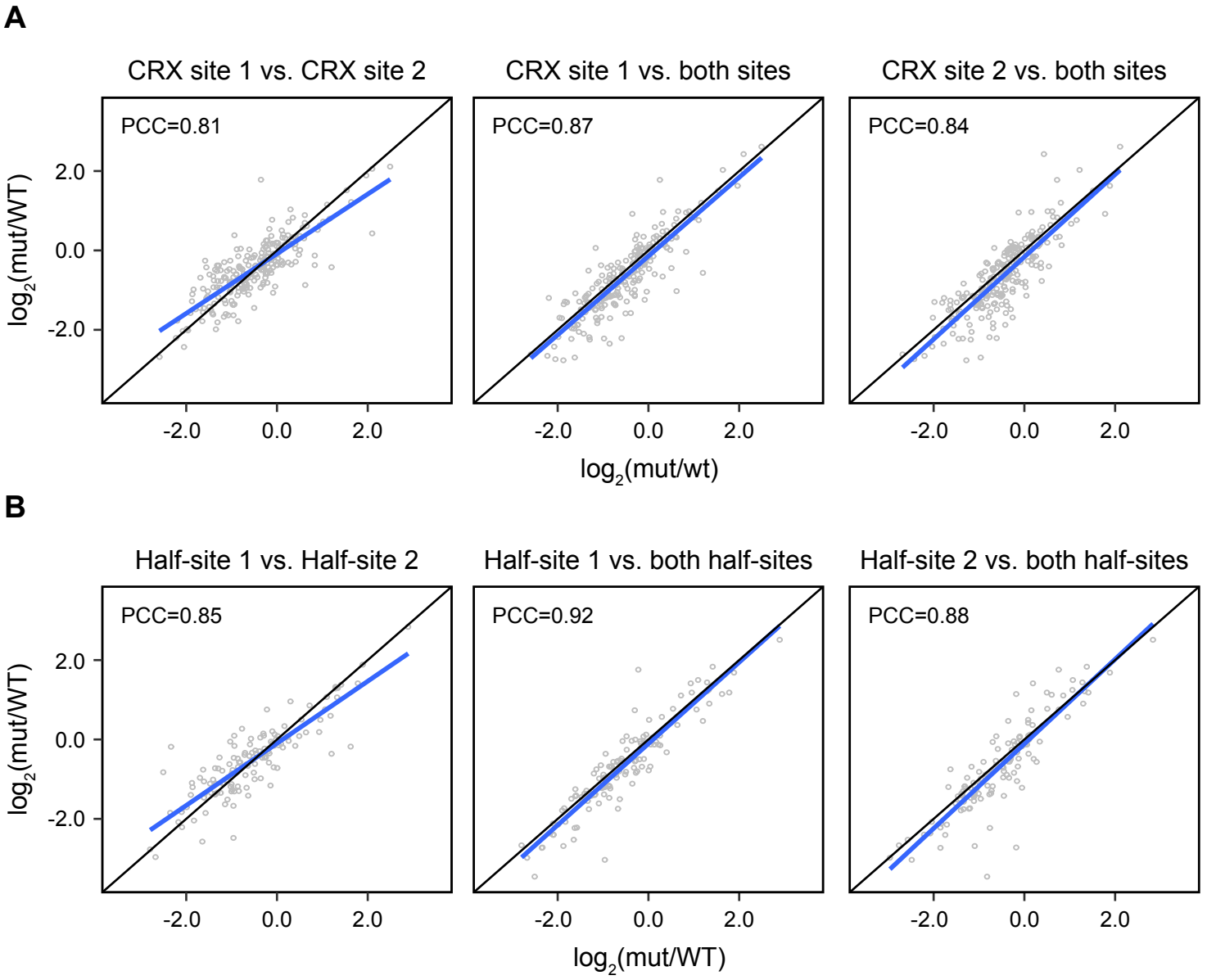
Supplemental Fig. 6. Expression of selected TFs during photoreceptor development. A) Heatmap of developmental expression for selected transcription factors corresponding to TF binding site families correlated with activity (see Fig. 3D) (Kim et al. 2016a; Kim et al. 2016b). Rows correspond to specific 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 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 the mouse genome.



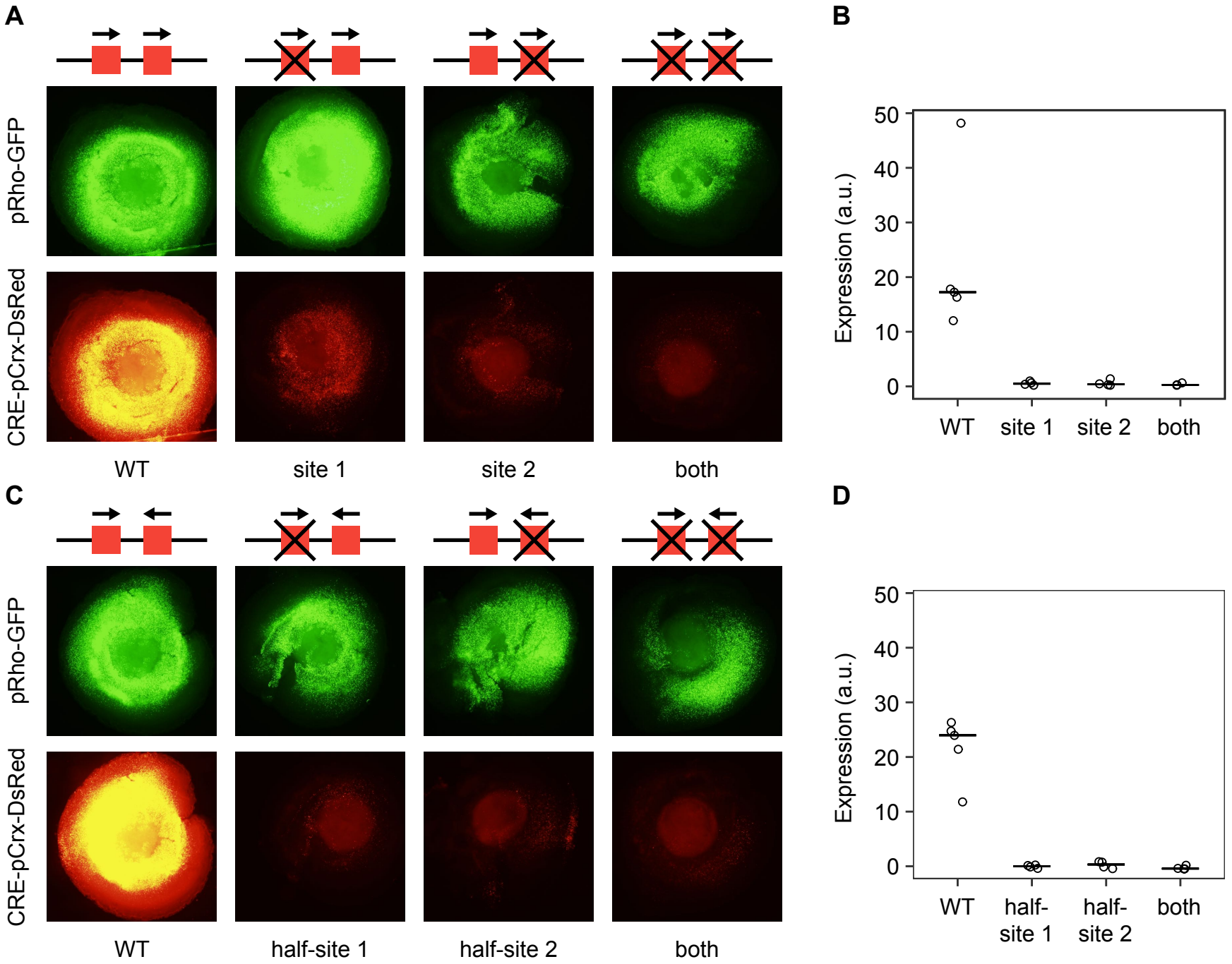
Supplemental Fig. 7. Correlation between individual chromatin features and CRE-seq activity. Scatter plots showing signal strength [$\log_2$(reads+1)] vs. CRE-seq activity (assayed on pCrx) for various epigenomic datasets. Lines show a linear fit for each scatter plot, and Pearson correlation coefficients are included in the upper left of each panel.



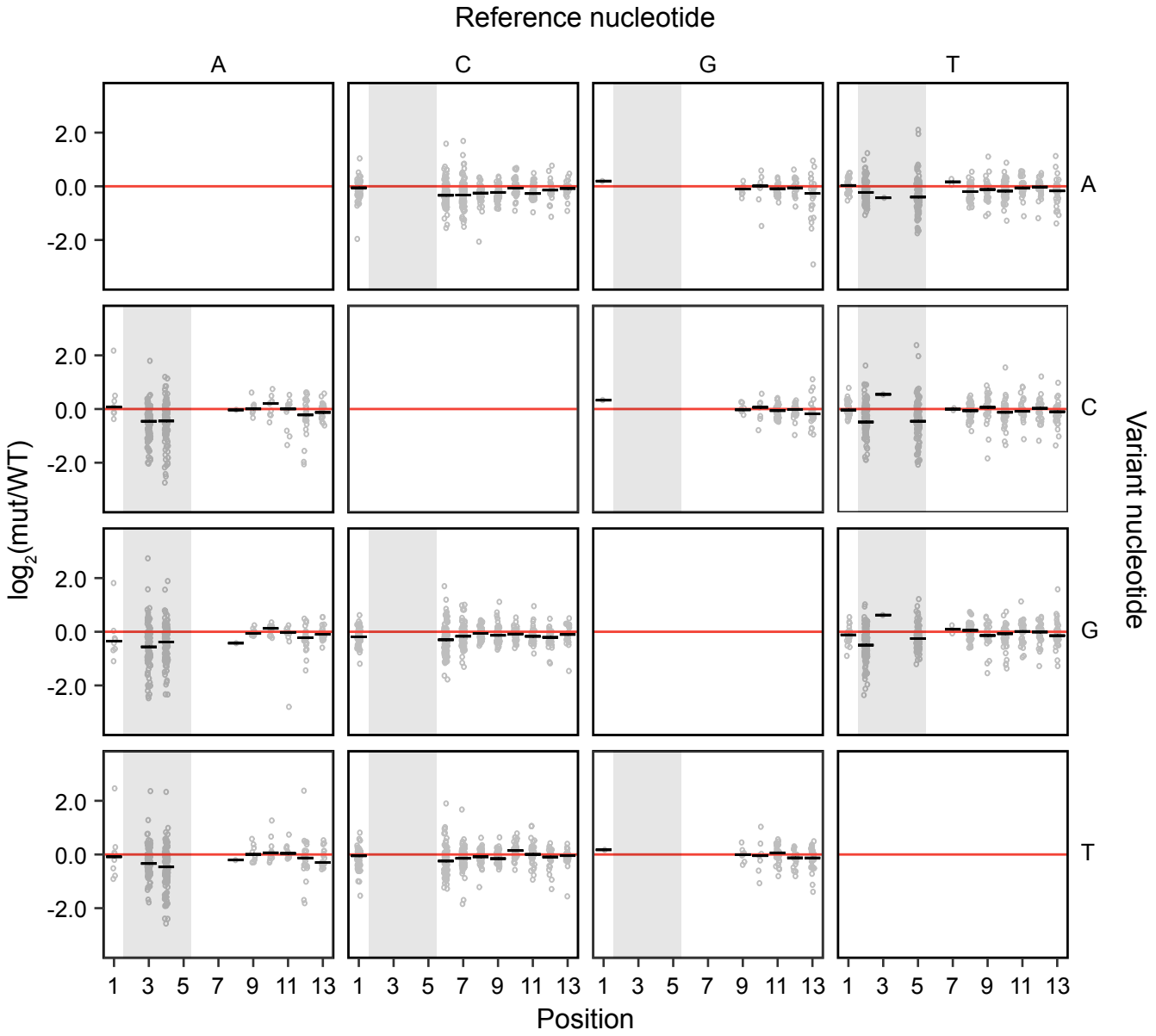
Supplemental Fig. 8. CRE activity estimated on *pRho* vs. *pCrx*. CRE-seq expression for 14,585 constructs assayed on *pRho* vs. *pCrx*. Each point corresponds to an individual CRE. Color indicates point density. Estimates of activity on each promoter are moderately correlated, but there is an inflection point at higher levels of activity. PCC: Pearson correlation coefficient.



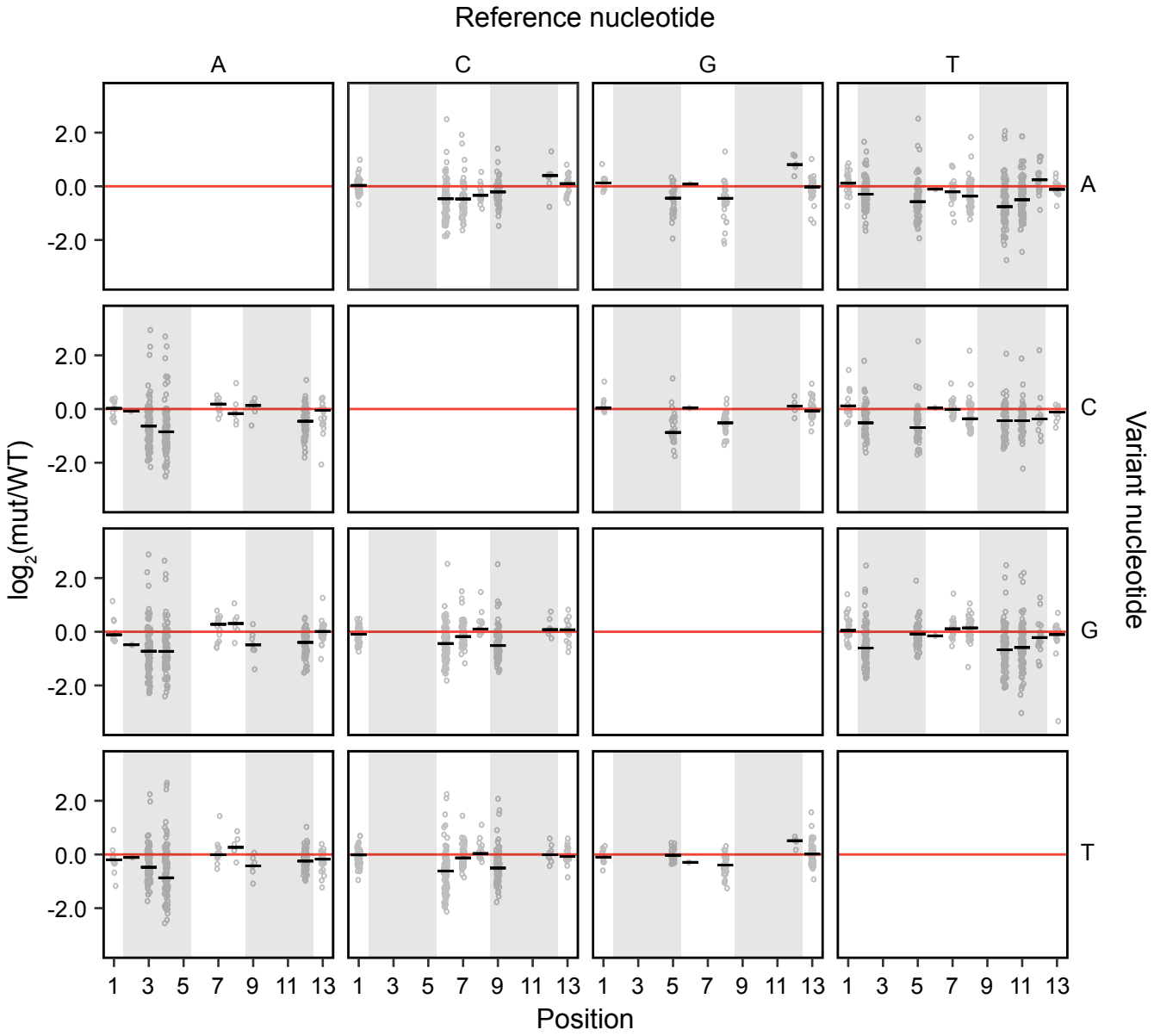
Supplemental Fig. 9. Effect of single- and double-mutants within the same CRE. A) Scatter plots showing the effect of mutating pairs of CRX binding sites individually or in combination for CREs with two predicted CRX binding sites. **B)** Scatter plots showing the effect of mutating either half site of dimeric CRX binding sites individually or in combination. Black lines: identity lines. Blue lines: linear fits. PCC: Pearson correlation coefficient.



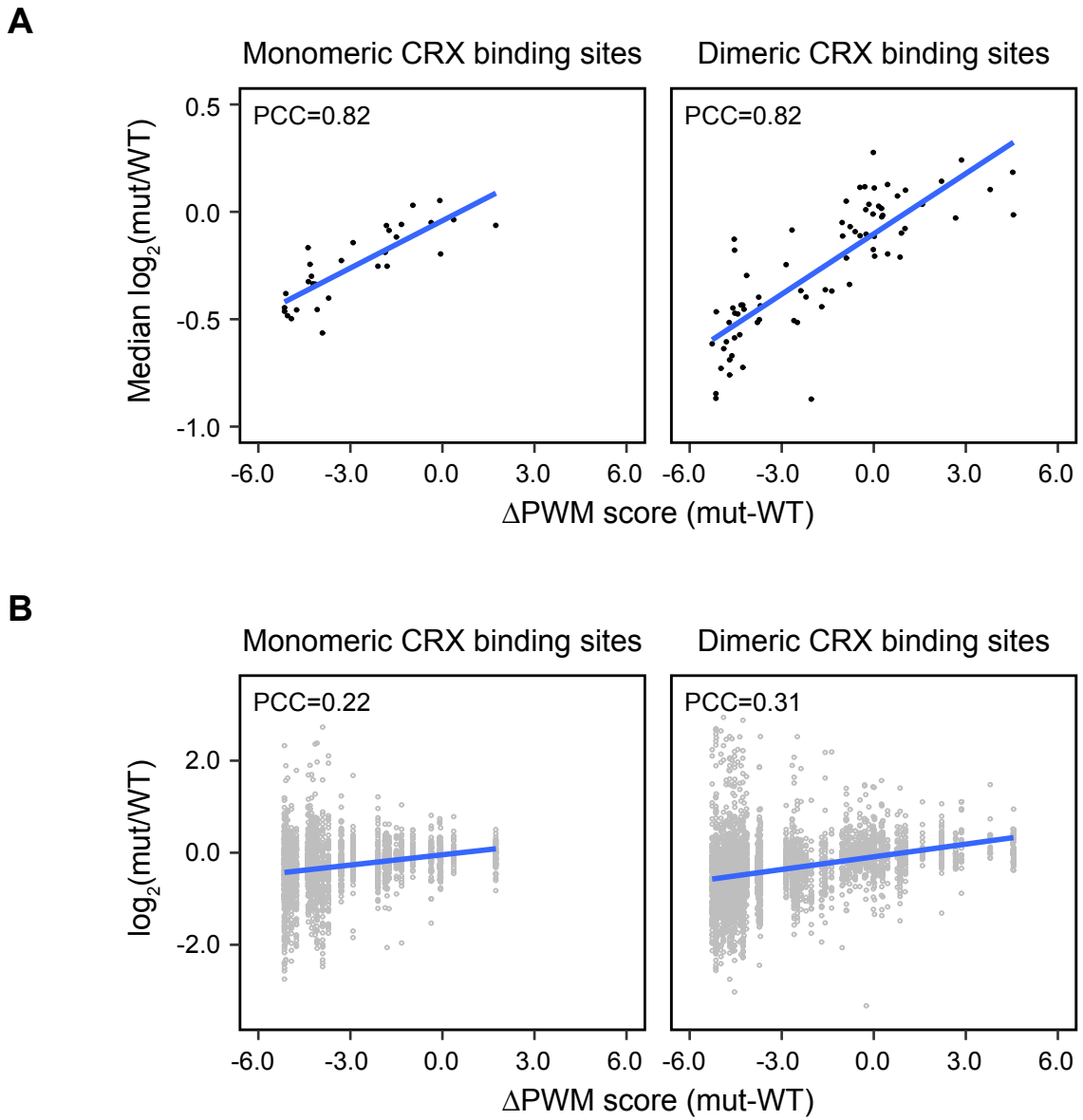
Supplemental Fig. 10. Validation of CRE-seq by fluorescent reporter assay. A) Whole-mount images of retinas electroporated with CRE-pCrx-DsRed test constructs and pRho-GFP controls. The selected CRE includes two monomeric CRX binding sites, which were inactivated by point mutation (TAAT to TACT) individually and in combination. **B)** Quantification of fluorescence for constructs shown in A normalized to pCrx-DsRed/pRho-GFP (see Supplemental Table 8). **C)** As in A**,** except the selected CRE has a dimeric CRX binding site, for which half-sites were inactivated by point mutation (TAAT to TACT) individually and in combination. **D).** Quantification of fluorescence as in B for constructs shown in C**.**



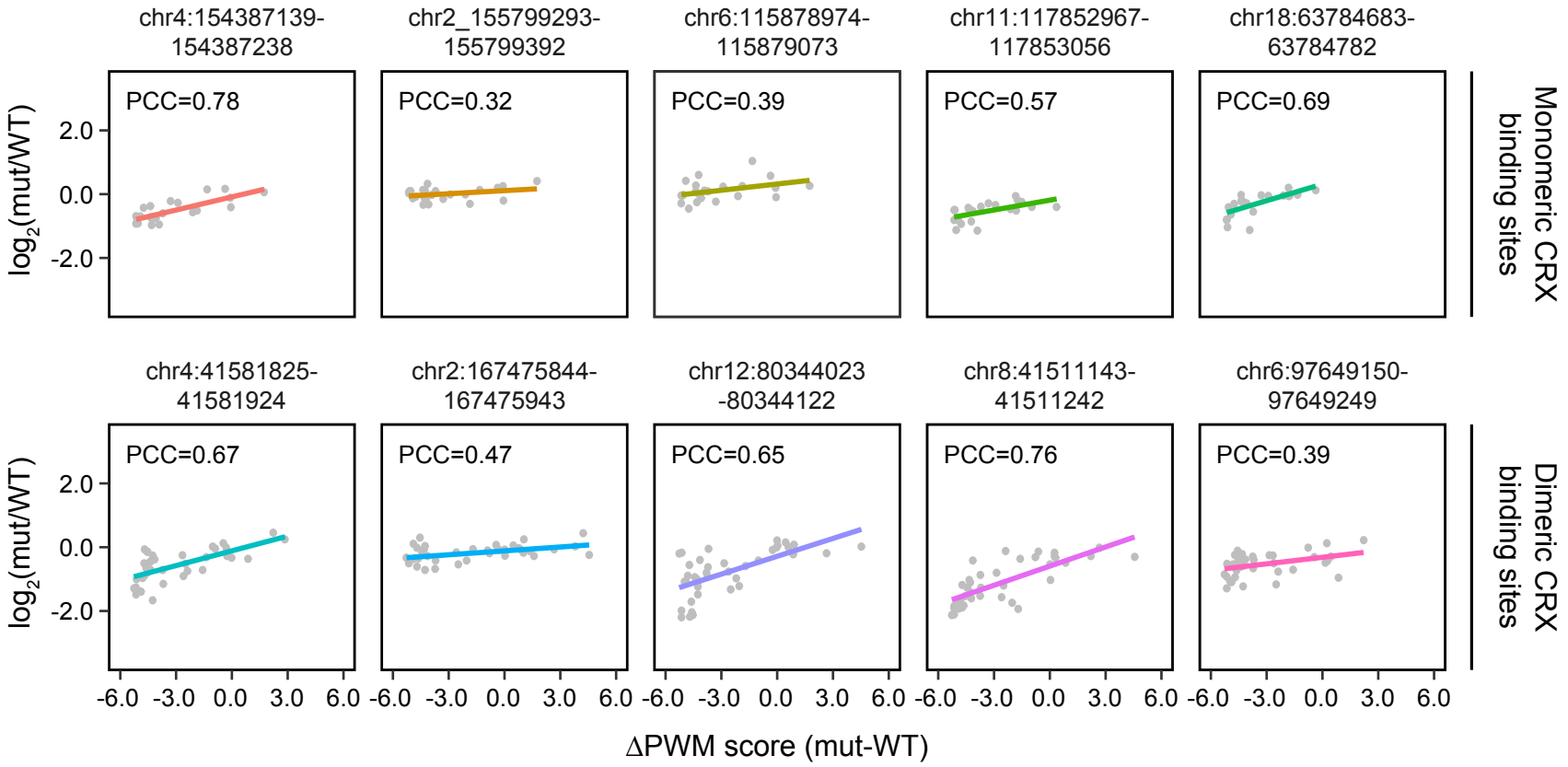
Supplemental Fig. 11. Dense substitution analysis of monomeric CRX binding sites. Scatter plots of the effects of specific substitutions at specific positions (gray points) within monomeric CRX binding sites. Separate panels are included for each of the 12 possible reference-to-variant substitutions 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.



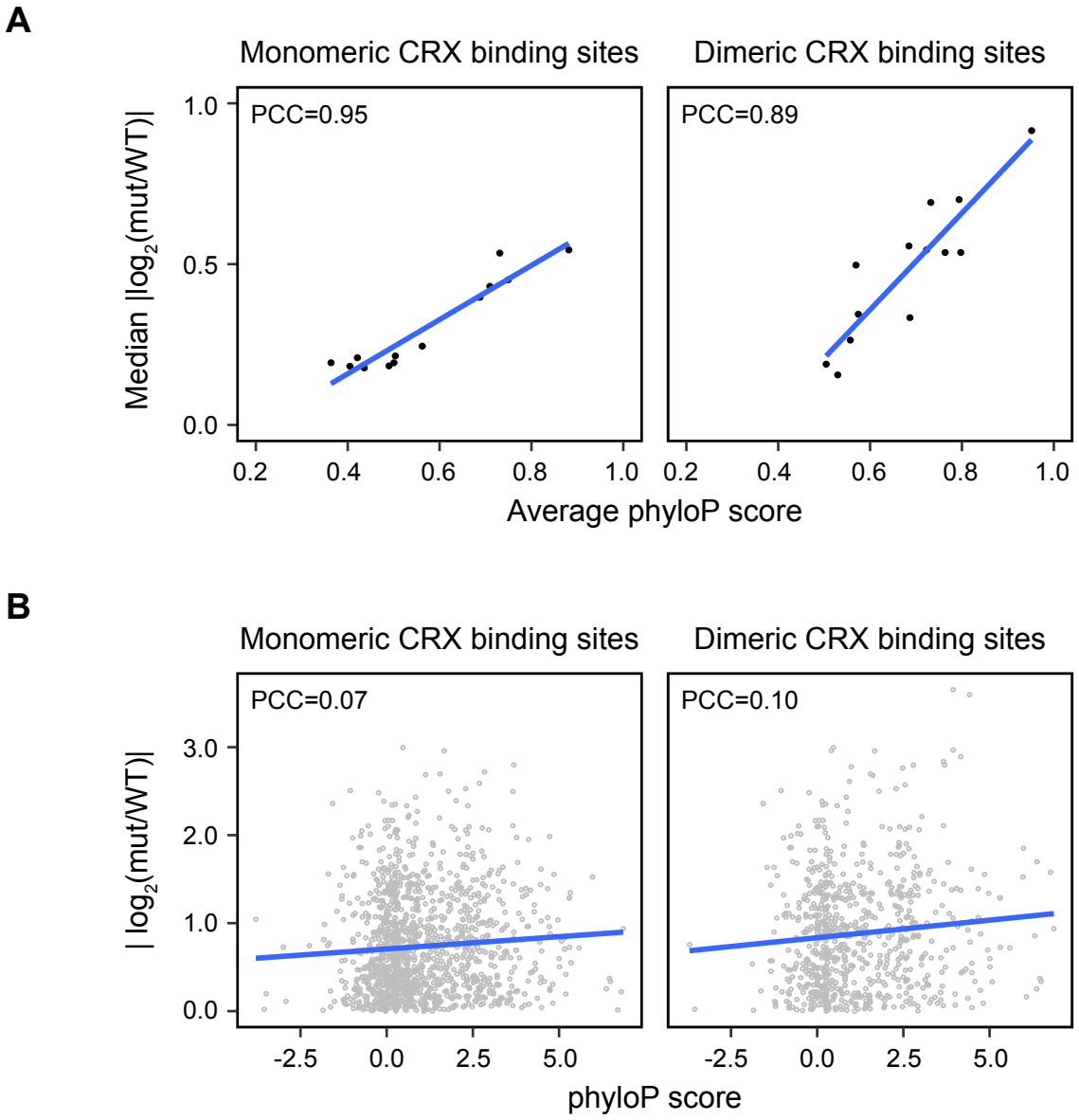
Supplemental Fig. 12. Dense substitution analysis of dimeric CRX binding sites. Scatter plots of the effects of specific substitutions at specific positions (gray points) within dimeric CRX binding sites. Separate panels are included for each of the 12 possible reference-to-variant substitutions at each 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.



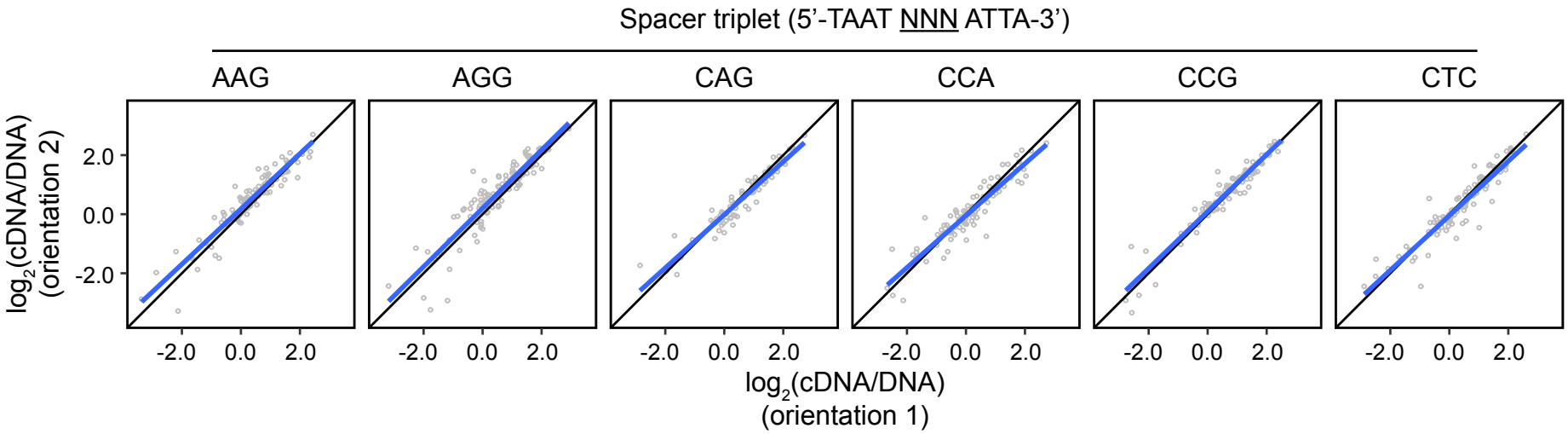
Supplemental Fig. 13. Aggregate correlation between change in affinity and change in CRE-seq activity. A) Scatter plot of change in PWM score vs. median change in activity as measured by CRE-seq for monomeric (left panel) and dimeric (right panel) CRX binding sites. Aggregating across CREs, changes in 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 Fig. 14. 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. 13), correlations are much higher when considering mutations within individual CREs. PCC: Pearson correlation coefficient.



Supplemental Fig. 15. Correlation between phylogenetic conservation and change in CRE-seq activity. A) Scatter plot of average conservation (100-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 (100-way vertebrate phyloP scores) vs. absolute change in CRE-seq activity for individual mutations. PCC: Pearson correlation coefficient.



Supplemental Fig. 16. 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).

2. List of Supplemental Tables

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

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

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

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

Supplemental Table 5. Monomeric and dimeric PWM scores and mutant CRE activity for 1756 inactivating mutations in CRX binding site.

Supplemental Table 6. Linear modeling of interactions between pairs of CRX binding sites.

Supplemental Table 7. Linear modeling of interactions between half-sites within dimeric CRX binding sites.

Supplemental Table 8. Quantification of CRE activity by fluorescent reporter assay.

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

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

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

3. List of Supplemental Datasets

Raw and processed CRE-seq data have been deposited in the GEO (GSE106243).

Supplemental Dataset 1. CRE-seq library oligos. Summary of 100,000 oligos in CRE-seq library. bc: unique 13-bp barcode. id.oligo: unique oligo identifier, including 200-bp parent ChIP-seq peak coordinates (mm9) (CRE sequence corresponds to central 100 bp). seq.alt: CRE sequence. seq.wt: corresponding wild-type sequence. pos.alt: position(s) of any mutated nucleotides. ref: reference nucleotide(s) at corresponding position(s). alt: variant nucleotide(s) at corresponding position(s).

Supplemental Dataset 2. CRE-seq raw counts. Table of raw barcode counts. bc: unique 13-bp barcode. id.oligo: unique oligo identifier, including 200-bp parent ChIP-seq peak coordinates (mm9) (CRE sequence corresponds to central 100 bp). Additional columns: raw counts for each barcode from cDNA and DNA libraries from CRE-seq replicates assayed on *pRho* or *pCrx*.

Supplemental Dataset 3. Size factors. Size factors used for library normalization (estimated with median ratio method) (Anders and Huber 2010). Size factors were estimated using additional samples that were ultimately omitted from the current study. Therefore, we report size factors directly to facilitate reproducible analysis.

Supplemental Dataset 4. CRE-seq normalized expression. Normalized expression values for target CREs. id.cre: unique CRE identifier, including 200-bp parent ChIP-seq peak coordinates (mm9) (CRE sequence corresponds to central 100 bp). Additional columns: normalized expression

values from CRE-seq replicates assayed on *pRho* or *pCrx*. Values are calculated as the log2 of cDNA counts (summed over barcodes) over DNA counts (summed over barcodes).

Supplemental Dataset 5. CRE-seq differential expression. Differential expression table for mutated CREs. id.cre: unique CRE identifier. condition: promoter used for assay (*pRho* or *pCrx*). median.wt: median wild-type activity across three biological replicates. median.alt: median variant activity across three biological replicates. lfc: log2 fold change (median.alt-median.wt). p.value: p-value for Welch's t-test comparing wild-type and mutant CRE activity. p.adj: adjusted p-value (Benjamini-Hochberg). Additional columns: normalized wild-type and mutant expression values from CRE-seq replicates assayed on *pRho* or *pCrx*.

4. Supplemental Methods

4.1 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-ACTAGTCGGTACCNNNNNNNNNNNNNGCGGCCGCCAACTACTACTACAG
               ******   ****** ******              ********
               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 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 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 (Montana et al. 2011a), 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 (Montana et al. 2011a) 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.

4.2 Validation 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) (Magoc and Salzberg 2011), and barcodes for which >10% of reads could not be merged were removed from subsequent analysis (n=1,286). Merged reads were aligned to designed oligos (including 100 bp of vector sequence on either side) with bowtie2 (v2.3.0) (Langmead and Salzberg 2012). Barcodes for which >10% of reads were not aligned to the proper CRE were removed from subsequent analysis (n=4,344). 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=1,207). In aggregate, these quality control steps flagged 6,313 barcodes that were removed from analysis.

4.3 CRE-seq assay and sequencing library preparation

Retinal electroporation, RNA and DNA isolation, and cDNA synthesis. 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 (Montana et al. 2011b; Kwasnieski et al. 2012; White et al. 2013; Shen et al. 2016). 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 (Shen et al. 2016). cDNA synthesis was performed with SuperScript IV (Invitrogen) using an oligo(dT) primer according to manufacturer instructions.

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 loaded at 150 pM with 20% PhiX DNA (29-43x10^6 reads per library).

4.4 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 (Anders and Huber 2010). 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 (Bolstad et al. 2003).

4.5 Data processing of previously generated datasets

DNase-seq, TF ChIP-seq, and Histone ChIP-seq data processing. See Supplemental Table 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 (ENCODE Project Consortium 2012). Single-end sequencing reads from CRX ChIP-seq (wild-type and *Nrl*$^{-/-}$ whole retina) were downloaded from the GEO (GSE20012) (Corbo et al. 2010). Single-end sequencing reads from NRL ChIP-seq (wild-type whole retina) were downloaded from the NEI (Hao et al. 2012). Single-end sequencing reads from H3K27ac,

H3K4me1, H3K4me3, and H3K27me3 ChIP-seq were downloaded from the GEO (GSE72550) (Mo et al. 2016). Reads were aligned to mm10 with bowtie2 (v2.3.0) (Langmead and Salzberg 2012). Alignments with mapping quality <30 or overlapping ENCODE blacklist regions (ENCODE Project Consortium 2012) were removed with SAMtools (v1.5) (Li et al. 2009). Alignments were sorted and deduplicated with Picard (broadinstitute.github.io/picard/). Peaks were called with MACS2 (v2.1.1) (Zhang et al. 2008) and annotated with HOMER (v4.9) (Heinz et al. 2010). 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) (Hughes et al. 2017). Reads were aligned to mm10 with bowtie2 (v2.3.0) (Langmead and Salzberg 2012), allowing fragment lengths up to 2 kb. Alignments with mapping quality <30 or overlapping ENCODE blacklist regions (ENCODE Project Consortium 2012) were removed with SAMtools (v1.5) (Li et al. 2009). 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) (Zhang et al. 2008) and annotated with HOMER (v4.9) (Heinz et al. 2010).

RNA-seq data processing. Sequencing reads from transcriptome profiling of developing mouse rods and cones were downloaded from the GEO (GSE74660) (Kim et al. 2016a; Kim et al. 2016b). Transcript abundance (transcripts per million, or TPM) was estimated with kallisto (v0.43) (Bray et al. 2016) using the Ensembl gene model (GRCm38 assembly, release 79).

4.6 Models of TF occupancy and CRE-seq activity

Prediction of CRX-bound regions with logistic regression. 200 bp regions centered on TSS-distal (>1,000 bp upstream and >100 bp downstream) CRX ChIP-seq peaks annotated as intergenic or intronic were lifted over to mm9 with HOMER (n=5,250). FASTA files for these regions as well as a 10X set of background sequences were generated with the gkmSVM R package (Ghandi et al. 2016). 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 (Jolma et al. 2013) in each sequence were identified with FIMO (v4.11.2) (Grant et al. 2011) 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 site motifs was reduced by collapsing motifs belonging to the same cluster as defined by a recent analysis of 9650 PWMs, yielding a non-redundant set of 206 motifs (Castro-Mondragon et al. 2017), retaining the most prevalent motif in each cluster as a representative member. Models were fit using motif 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 (Friedman et al. 2010) and caret (Kuhn 2008), 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 (Keilwagen et al. 2014; Grau et al. 2015).

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 (Lee 2016) 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 (Keilwagen et al. 2014; Grau et al. 2015).

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 from analysis. 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) (Heinz et al. 2010) 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) (Heinz et al. 2010) 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, motif count, and 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 counts of TF binding sites 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 (Friedman et al. 2010) and caret (Kuhn 2008), 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 (Keilwagen et al. 2014; Grau et al. 2015).

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) (Jolma et al. 2013). 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 (Lee 2016). For ATAC-seq and DNase-seq data, gkm-SVM models were trained using cell- and tissue-type specific peaks identified by DESeq2 (Love et al. 2014), 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 (Friedman et al. 2010) and caret (Kuhn 2008). 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.

4.7 Conservation analysis

CRX binding sites within CRX ChIP-seq peaks were identified with FIMO (v4.11.2) (Grant et al. 2011) using the PWMs PITX1_DBD (monomeric) and OTX2_DBD_1 (dimeric) (Jolma et al. 2013) 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) (Quinlan and Hall 2010).

4.8 Data visualization

Plots were generated in R (v3.3) (R Core Team 2016) using the ggplot2 package (Wickham 2009).

4.9 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
```

5. References

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

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**: 185-193.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525-527.

Castro-Mondragon JA, Jaeger S, Thieffry D, Thomas-Chollier M, van Helden J. 2017. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res* **45**: e119.

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.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

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

Ghandi M, Mohammad-Noori M, Ghareghani N, Lee D, Garraway L, Beer MA. 2016. gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* **32**: 2205-2207.

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

Grau J, Grosse I, Keilwagen J. 2015. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**: 2595-2597.

Hao H, Kim DS, Klocke B, Johnson KR, Cui K, Gotoh N, Zang C, Gregorski J, Gieser L, Peng W et al. 2012. Transcriptional regulation of rod photoreceptor homeostasis revealed by in vivo NRL targetome analysis. *PLoS Genet* **8**: e1002649.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**: 576-589.

Hughes AE, Enright JM, Myers CA, Shen SQ, Corbo JC. 2017. Cell Type-Specific Epigenomic Analysis Reveals a Uniquely Closed Chromatin Architecture in Mouse Rod Photoreceptors. *Sci Rep* **7**: 43184.

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152**: 327-339.

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

Kim JW, Yang HJ, Brooks MJ, Zelinger L, Karakulah G, Gotoh N, Boleda A, Gieser L, Giuste F, Whitaker DT et al. 2016a. NRL-Regulated Transcriptome Dynamics of Developing Rod Photoreceptors. *Cell Rep* **17**: 2460-2473.

Kim JW, Yang HJ, Oel AP, Brooks MJ, Jia L, Plachetzki DC, Li W, Allison WT, Swaroop A. 2016b. Recruitment of Rod Photoreceptors from Short-Wavelength-Sensitive Cones during the Evolution of Nocturnal Vision in Mammals. *Dev Cell* **37**: 520-532.

Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. 2012. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc Natl Acad Sci U S A* **109**: 19498-19503.

Kuhn M. 2008. Building Predictive Models in R Using the caret Package. *J Stat Softw* **28**: 1-26.

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

Lee D. 2016. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**: 2196-2198.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 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 Biol* **15**: 550.

Magoc T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957-2963.

Mo A, Luo C, Davis FP, Mukamel EA, Henry GL, Nery JR, Urich MA, Picard S, Lister R, Eddy SR et al. 2016. Epigenomic landscapes of retinal rods and cones. *Elife* **5**.

Montana CL, Lawrence KA, Williams NL, Tran NM, Peng GH, Chen S, Corbo JC. 2011a. Transcriptional regulation of neural retina leucine zipper (Nrl), a photoreceptor cell fate determinant. *J Biol Chem* **286**: 36921-36931.

Montana CL, Myers CA, Corbo JC. 2011b. Quantifying the activity of cis-regulatory elements in the mouse retina by explant electroporation. *J Vis Exp* doi:10.3791/2821.

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

Shen SQ, Myers CA, Hughes AE, Byrne LC, Flannery JG, Corbo JC. 2016. Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res* **26**: 238-255.

White MA, Myers CA, Corbo JC, Cohen BA. 2013. 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* **110**: 11952-11957.

Wickham H. 2009. *ggplot2 : elegant graphics for data analysis*. Springer, New York.

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 Biol* **9**: R137.