Supporting Information

White et al. 10.1073/pnas.1307449110

SI Text: Explanation of Data Types in Datasets S1–S3

Dataset S1: Library Sequences and Read Counts. This is a FASTA file containing the complete sequence of each oligonucleotide in the library and *cis*-regulatory element (CRE)-seq reads per million for each RNA and DNA sample. The FASTA header consists of the unique ID for that sequence, followed by the barcode CRE-seq data in reads per million in the following order: plasmid library DNA; RNA sample 1; RNA sample 2; RNA sample 3; RNA sample 4; RNA sample 5; RNA sample 6 (CBRM, CBRs with Cone-rod homeobox (Crx) motifs; CBRNO, CBRs lacking Crx motifs; UBR, unbound regions with Crx motifs; CBRMUT, mutant versions of CBRM sequences; SCRCBRM, scrambled CBRM sequences; SCRCBRNO, scrambled CBRNO sequences; SCRUBR, scrambled UBR sequences).

Each CRE sequence is structured as follows: positions 1–15, priming sequence GTAGCGTCTGTCCGT; positions 16–21, EcoRI site; positions 22–105, 84-bp CRE sequence; positions 106–111, SpeI site; position 112, C (spacer between SpeI and SphI sites); positions 113–118, SphI site; positions 119–127, 9-bp barcode; positions 128–135, NotI site; positions 136–150, priming sequence CAACTACTACTACAG.

Dataset S2: CRE Expression and Sequence Annotations. This is a tab-delimited text file containing mean CRE expression and other data used in the analyses. The data are presented in columns, with one row per CRE.

Explanation of data types. LOCUS: Unique identifier for each CRE, comprised of the original genomic position of each CRE (if a CBR or UBR) or the locus name of the original sequence (if a scrambled sequence), and the CRE class abbreviation. Sequences are taken from the July 2007 mouse genome assembly (NCBI37/mm9) (1).

CLASS: Type of CRE—CBRM, CBRs with Crx motifs; CBRNO, CBRs lacking Crx motifs; UBR, unbound regions with Crx motifs; CBRMUT, mutant versions of CBRM sequences; SCRCBRM, scrambled CBRM sequences; SCRCBRNO, scrambled CBRNO sequences; SCRUBR, scrambled UBR sequences.

MEAN: Mean CRE expression, averaged across the six RNA-seq replicates for each barcode that passed a filter of plasmid DNA >1 read/million. Each RNA measurement (in reads per million) was normalized to the plasmid library DNA measurement (in reads per million) to control for differential representation of CREs in the original plasmid library.

SEM: SE of mean CRE expression, calculated as described in Kwasnieski et al. (2).

CI_95: 95% confidence intervals for mean CRE expression.

DF: Degrees of freedom, obtained via error propagation (2) for statistical testing.

NUM_BC: Number of barcodes that passed a filter of plasmid DNA >1 read/million.

CHIP_READS: Number of ChIP-seq reads assigned to that CBR by Corbo et al. (3); not applicable (NA) to non-CBR CREs.

CRX_MOTIFN: Number of Crx motif occurrences scoring better than a $P = 0.001$ threshold (*Materials and Methods*).

CRX_MOTIF_SUM: Sum of position weight matrix match scores for all Crx motifs scoring above a $P = 0.001$ threshold, as calculated by FIMO (4).

GC_FRAC: Fraction of G and C nucleotides in the CRE sequence.

ORCHID_SUM: Sum of per-base ORChID2 scores (5).

PHOTOR_GENE: CBR associations with photoreceptor genes from Corbo et al. (3). Not applicable (NA) to non-CBR sequences. 0, not associated with photoreceptor gene; 1, associated with gene down-regulated in $Crx^{-/-}$ retina; 2, associated with gene up-regulated in $Crx^{-/-}$ retina; 3, associated with other dysregulated genes in $Crx^{-/-}$ or $Nrl^{-/-}$ retina.

OCCUPANCY_6: Predicted Crx occupancy of CRE sequence at medium Crx concentration ($\mu = 6$), using the binding model of Zhao et al. (6).

OCCUPANCY_9: Predicted Crx occupancy of CRE sequence at high Crx concentration ($\mu = 9$), using the binding model of Zhao et al. (6).

AAAA: Number of AAAA occurrences.

CCCGGGCG_FREQ: Frequency of CC/CG/GG/GC dinucleotides.

NRL_MOTIFN: Number of Nrl motif occurrences scoring better than a $P = 0.001$ threshold (*Materials and Methods*).

NRL_MOTIF_SCORE: Sum of position weight matrix match scores for all Nrl motifs scoring better than a $P = 0.001$ threshold, as calculated by FIMO (4).

Dataset S3: Expression Data and Annotations for CBR WT/Mutant Pairs. This is a tab-delimited text file containing data on matched WT/MUT CBR pairs. Only WT/MUT pairs with equal numbers of barcodes passing a filter of plasmid DNA >1 read/million were considered.

Explanation of data types. LOCUS: Chromosome and start position for the CBR sequence. Sequences are taken from the July 2007 mouse genome assembly (NCBI37/mm9) (1).

WT_MEAN: Mean CRE-seq expression; these data are taken from the MEAN column for CBRM sequences in Dataset S2.

FOLD_CHANGE: Ratio of mean CBRMUT expression/mean CBRM expression.

WELCH_t: Welch's *t* statistic used for significance testing for a difference in mean expression between the WT and mutant CBR pairs, calculated as described in Sokal and Rohlf (7).

DF: Degrees of freedom, obtained via error propagation (2) for statistical testing.

N: Number of replicate measurements for each CBR version in the WT/MUT pair, calculated as (six replicates) × (number of barcodes that passed the DNA filter).

RAW_PVAL: Raw *P* values for difference in mean CRE-seq expression between the WT and mutant versions of a CBR, calculated by Welch's *t* test.

ADJ_PVAL: Adjusted *P* values for difference in mean CRE-seq expression between the WT and mutant versions of a CBR, obtained by correcting raw *P* values using the method of Benjamini and Hochberg (8).

PHOTOR_GENE: CBR associations with photoreceptor genes from Corbo et al. (3). Not applicable (NA) to non-CBR sequences. 0, not associated with photoreceptor gene; 1, associated with gene down-regulated in $Crx^{-/-}$ retina; 2, associated with gene up-regulated in $Crx^{-/-}$ retina; 3, associated with other dysregulated genes in $Crx^{-/-}$ or $Nrl^{-/-}$ retina.

1. Waterston RH, et al.; Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562.

2. 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 USA* 109(47):19498–19503.

3. Corbo JC, et al. (2010) CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Res* 20(11):1512–1525.
4. Grant CE, Bailey TL, Noble WS (2011) FIMO: Scanning for occurrences of a given motif. *Bioinformatics* 27(7):1017–1018.
5. Bishop EP, et al. (2011) A map of minor groove shape and electrostatic potential from hydroxyl radical cleavage patterns of DNA. *ACS Chem Biol* 6(12):1314–1320.
6. Zhao Y, Granas D, Stormo GD (2009) Inferring binding energies from selected binding sites. *PLOS Comput Biol* 5(12):e1000590.
7. Sokal RR, Rohlf FJ (1994) *Biometry* (W. H. Freeman, New York), 3rd Ed, pp 404–405.
8. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B* 57(1):289–300.

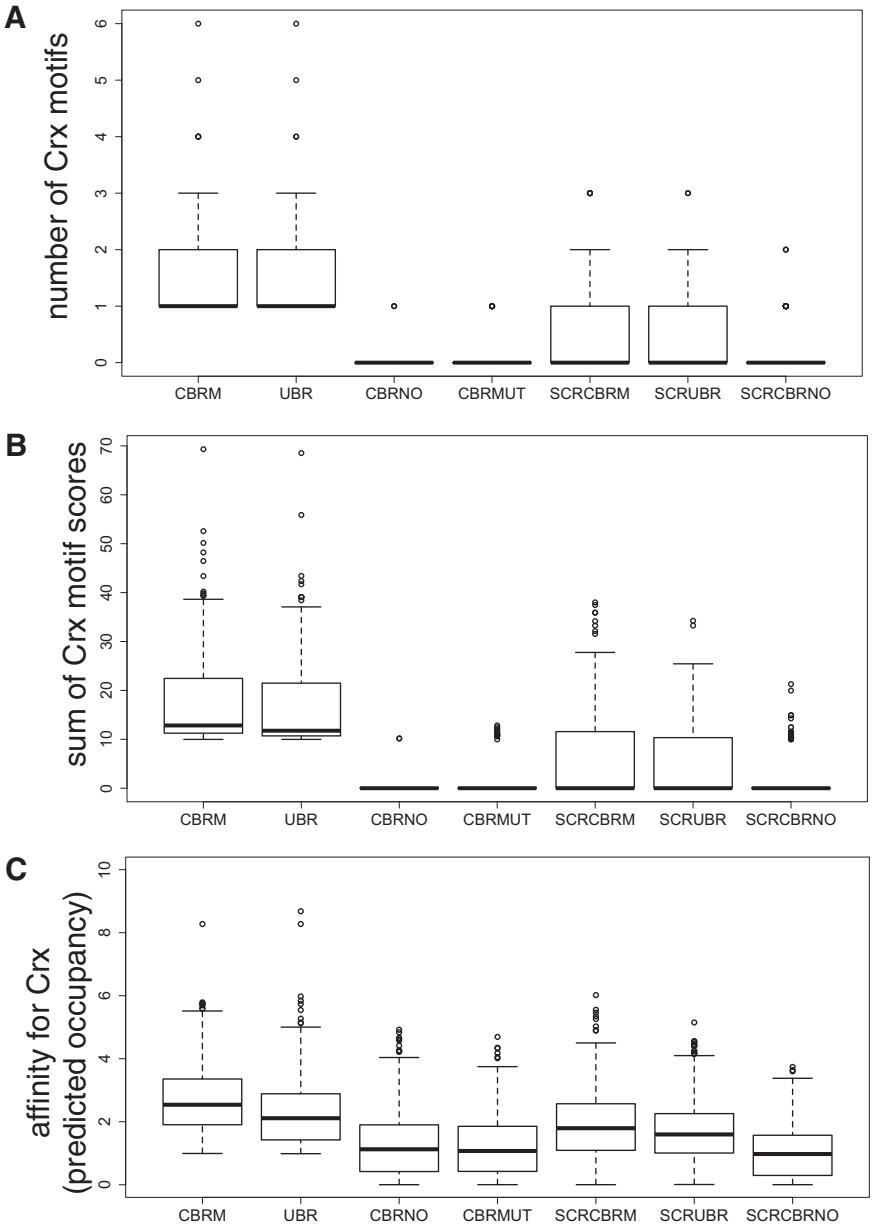

Fig. S1. Crx motif content of synthesized CREs. Box and whisker plots showing (*A*) number of Crx motifs and (*B*) total score of all Crx motifs scoring better than a $P = 0.001$ threshold in CRE sequences, as determined by FIMO (1). (*C*) Total predicted Crx occupancy (determined by a binding model; *Materials and Methods*) of CRE sequences. Distribution medians are indicated by bold horizontal bars. CBRM, CBRs with motifs; CBRMUT, mutCBRs; CBRNO, CBRs lacking motifs; SCRCBRM, scrambled CBRs with motifs; SCRCBRNO, scrambled CBRs lacking motifs; SCRUBR, scrambled UBRs; UBR, unbound regions with motifs.

1. Grant CE, Bailey TL, Noble WS (2011) FIMO: Scanning for occurrences of a given motif. *Bioinformatics* 27(7):1017–1018.

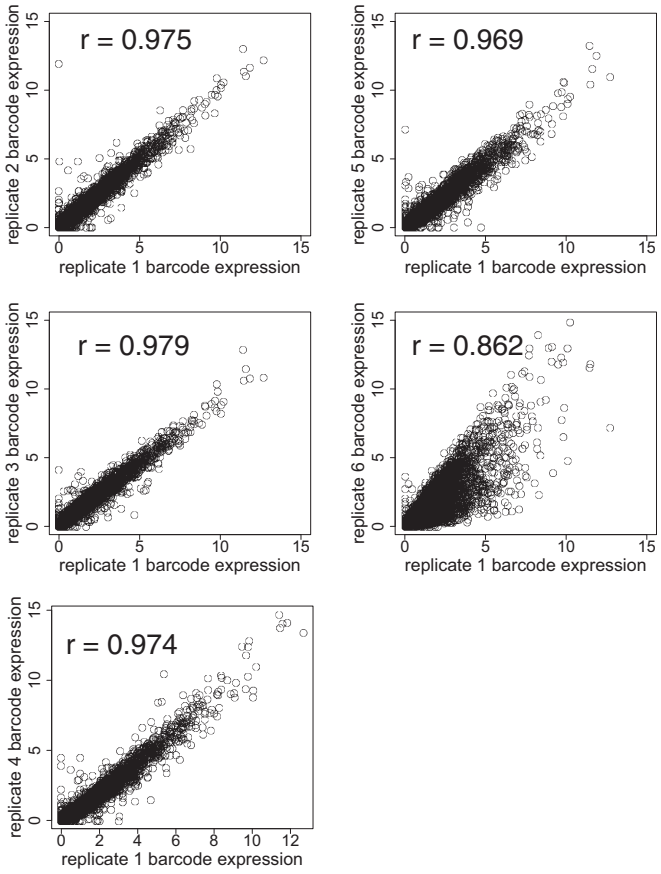

Fig. S2. Correlation between biological replicates of CRE-seq. CRE-seq expression of individual barcodes in biological replicate 1 (*x* axis) and replicates 2–6 (*y* axes). *r*, Pearson's correlation coefficient.

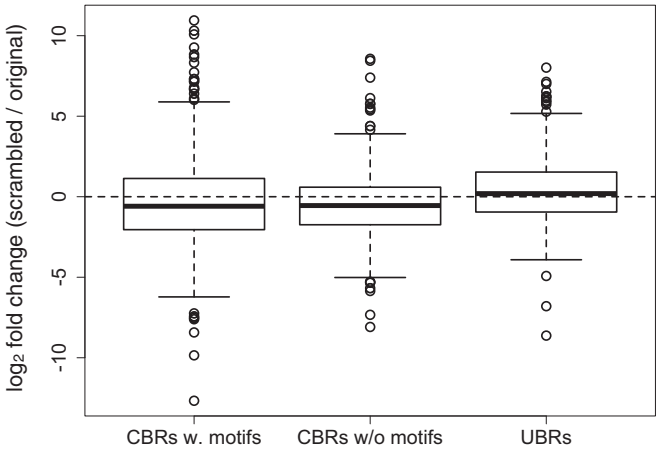

Fig. S3. Pairwise comparisons of CBR and UBR activity with that of corresponding scrambled sequences. Corresponding scrambled sequences for 545 CBRs with motifs, 247 CBRs without (w/o) motifs, and 320 UBRs were included in the CRE-seq library. Box and whisker plots show the distribution of $\log_2$ fold changes of scrambled/original pairs. The dashed horizontal line indicates the point of no change between the scrambled and original sequences.

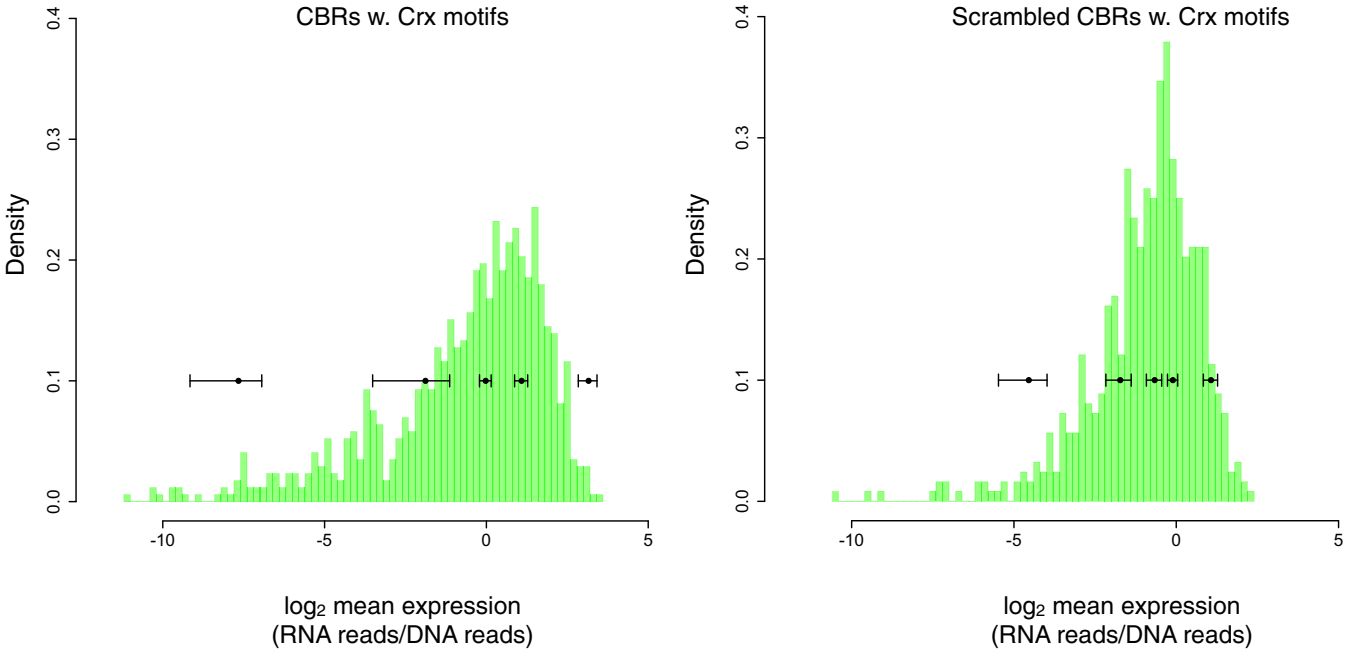

Fig. S4. Most CRE sequences, including scrambled sequences, modify the activity of the *Rho* minimal promoter. Log_2 mean expression of representative individual CREs (black points, ±95% CI) compared with the distribution of log_2 mean expression for all CREs (green), demonstrating that individual CBRs (*Left*) and scrambled DNA sequences (*Right*) drive distinct and reproducible levels of transcription.

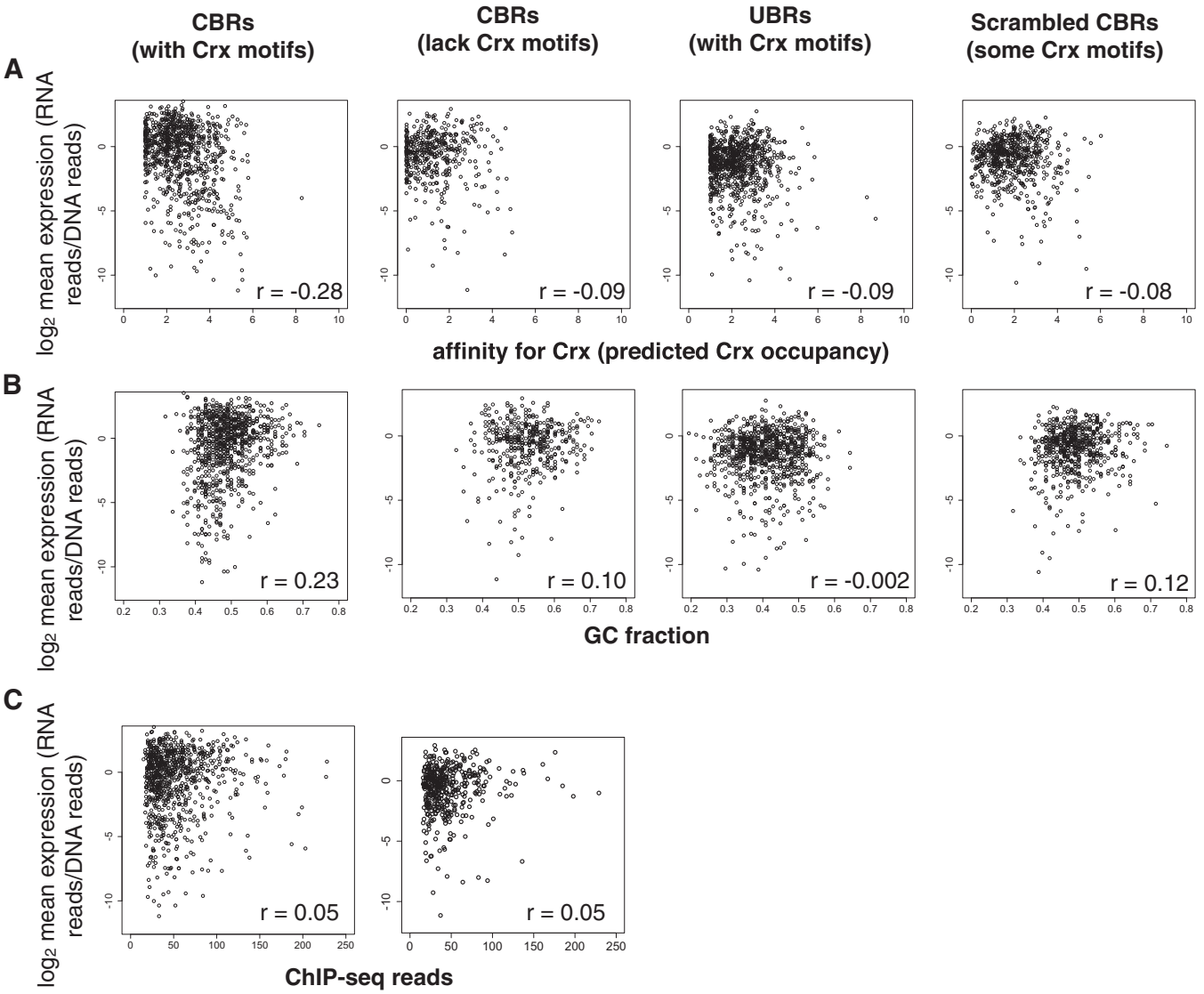

Fig. S5. Quantitative relationships between CRE expression and sequence features. (*A*) Affinity for Crx (given as predicted Crx occupancy), (*B*) GC nucleotide fraction, and (*C*) Crx occupancy measured by number of ChIP-seq reads (18) differ significantly among different classes of CREs, but do not quantitatively predict CRE expression.

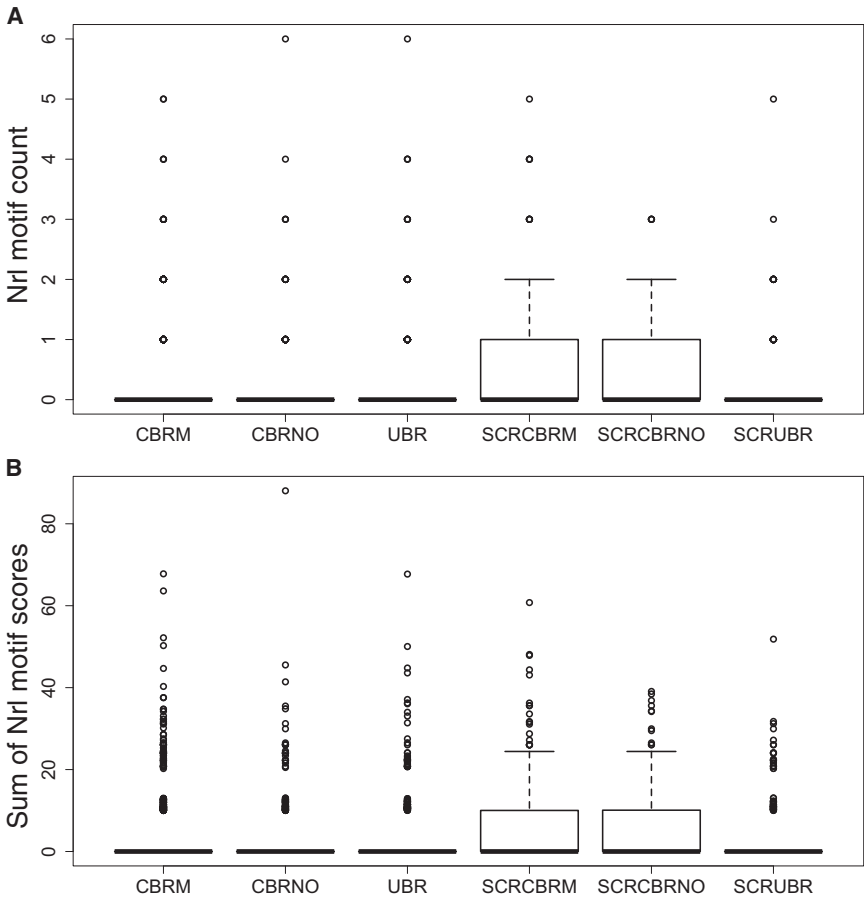

Fig. S6. Nrl motifs are equally rare in CBRs and UBRs. Box and whisker plots showing (*A*) number of Nrl motifs and (*B*) total score of all Nrl motifs, which scored better than a $P = 0.001$ threshold in CRE sequences, as determined by FIMO (1) using the Nrl position-weight matrix (*Materials and Methods*). Medians are indicated by bold horizontal bars. CBRM, CBRs with motifs; CBRMUT, mutCBRs; CBRNO, CBRs lacking motifs; SCRCBRM, scrambled CBRs with motifs; SCRCBRNO, scrambled CBRs lacking motifs; SCRUBR, scrambled UBRs; UBR, unbound regions with motifs.

1. Grant CE, Bailey TL, Noble WS (2011) FIMO: Scanning for occurrences of a given motif. *Bioinformatics* 27(7):1017–1018.

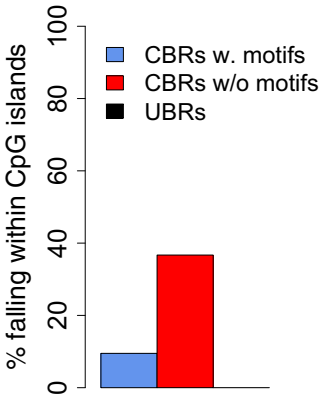

Fig. S7. CBRs lacking Crx motifs are more likely to reside in CpG islands than CBRs with motifs or UBRs. Bars indicate the percentage of each sequence class from the CRE-seq library that overlap annotated CpG islands (1).

1. Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Res* 12(6):996–1006.

Table S1. Primers used in CBR CRE-seq library construction and sequencing

Primer name	Sequence	Purpose
MO563	GTAGCGTCTGTCCGTGAATT	F primer to amplify CRE oligo library
MO564	CTGTAGTAGTAGTTGGCGGC	R primer to amplify CRE oligo library
MO574	CACCTGTTCCTGTAGGCATGC	F primer to amplify barcodes for Illumina sequencing
MO575	TATTACAATTGTAGCCAGAAGTCAGATGCTCAAG	R primer to amplify barcodes for Illumina sequencing

Other Supporting Information Files

Dataset S1 (TXT)
Dataset S2 (TXT)
Dataset S3 (TXT)