# Complex effects of nucleotide variants in a mammalian *cis*-regulatory element

Jamie C. Kwasnieski[a,1], Ilaria Mogno[a,1], Connie A. Myers[b], Joseph C. Corbo[b,2], and Barak A. Cohen[a,2]

[a]Center for Genome Sciences and Systems Biology, Department of Genetics, Washington University School of Medicine in St. Louis, St. Louis, MO 63108; and [b]Department of Pathology and Immunology, Washington University School of Medicine in St. Louis, St. Louis, MO 63110

Edited* by Jeffrey I. Gordon, Washington University School of Medicine in St. Louis, St. Louis, MO, and approved October 7, 2012 (received for review June 22, 2012)

***Cis*-regulatory elements (CREs) control gene expression by recruiting transcription factors (TFs) and other DNA binding proteins. We aim to understand how individual nucleotides contribute to the function of CREs. Here we introduce CRE analysis by sequencing (CRE-seq), a high-throughput method for producing and testing large numbers of reporter genes in mammalian cells. We used CRE-seq to assay >1,000 single and double nucleotide mutations in a 52-bp CRE in the *Rhodopsin* promoter that drives strong and specific expression in mammalian photoreceptors. We find that this particular CRE is remarkably complex. The majority (86%) of single nucleotide substitutions in this sequence exert significant effects on regulatory activity. Although changes in the affinity of known TF binding sites explain some of these expression changes, we present evidence for complex phenomena, including binding site turnover and TF competition. Analysis of double mutants revealed complex, nucleotide-specific interactions between residues in different TF binding sites. We conclude that some mammalian CREs are finely tuned by evolution and function through complex, nonadditive interactions between bound TFs. CRE-seq will be an important tool to uncover the rules that govern these interactions.**

gene regulation | systems biology | genomics | retina | CRX

Mutations in *cis*-regulatory elements (CREs) often have unexpected effects on gene regulation. We lack models with the predictive power to accurately interpret the functional consequences of noncoding polymorphisms. More generally, we do not understand the nucleotide-level architecture that distinguishes true CREs from nonfunctional groupings of transcription factor (TF) binding sites (TFBS). Although consortium-driven efforts continue to predict that large numbers of mammalian sequences are CREs (1, 2), we lack a corresponding high-throughput method for functionally analyzing the consequences of variants in these elements. Addressing these problems requires fine structure mutational analysis of mammalian CREs on a large scale—experiments that are difficult to perform using traditional assays. To facilitate such experiments, we developed CRE analysis by sequencing (CRE-seq), a high-throughput reporter gene assay for mammalian cells.

CRE-seq leverages recent advances in oligonucleotide (oligo) synthesis (3) and high-throughput sequencing (4). Using array-based oligo synthesis, we construct large numbers of reporter genes with unique sequence barcodes in their 3′ UTRs. These libraries of barcoded reporter genes are then transfected, *en masse*, into mammalian cells and quantified by performing RNA sequencing (RNA-Seq) (5) on the sequence barcodes. Here we present a study using CRE-seq to dissect a CRE in mouse *Rhodopsin* (*Rho*), a gene that is expressed strongly and specifically in the mammalian retina.

Tight control of *Rho* expression is critical for the function of mammalian retinas (6, 7). *Rho* expression is regulated in mice by multiple CREs located at varying distances from the transcription start site (TSS) (8, 9). These elements are occupied in vivo by CRX, a retinal homeodomain TF (8). One of these CREs, RhoCRE3, is located immediately upstream of the TSS and is sufficient to drive high levels of expression in rod photoreceptors (10, 11). This element contains binding sites for CRX and for NRL, a retina-specific basic-leucine-zipper protein (12, 13). How individual nucleotides within RhoCRE3 contribute to its function is not clear. To elucidate the functional architecture of RhoCRE3, we used CRE-seq to analyze the effects of >1,000 variants of this element.

## Results

We developed CRE-seq, a method that parallelizes the construction and measurement of mammalian reporter genes. Using high-throughput oligo synthesis (3), we created 1,040 variants of RhoCRE3, including all possible single nucleotide substitutions (156 mutations) and a large number of double substitutions (819 mutations). We also synthesized the wild-type RhoCRE3 sequence 65 times to include as controls. Each mutant RhoCRE3 was synthesized adjacent to a unique 9-bp sequence barcode. To provide redundancy in the experiment, each single mutant was attached to 10 unique barcodes, and each double mutant was attached to 5 unique barcodes. We then cloned the native *Rho* minimal promoter driving the fluorescent protein DsRed between the RhoCRE3 variants and their identifying barcodes. In the final plasmid library, RhoCRE3 variants were located in a position immediately upstream of the TSS, surrounded by the same context sequence as in the endogenous *Rho* gene. The library contains 5,720 distinct reporter genes, each with a unique sequence barcode in the 3′ UTR of the DsRed gene (Fig. 1*A*). We electroporated this library into explanted newborn mouse retinas (9). Using DsRed to monitor the progression of the experiment, we confirmed that library expression was specific to rod photoreceptors (Fig. 1*B*), the cell type in which *Rho* is expressed.

We measured the expression of all library members using high-throughput sequencing. After growing the electroporated retinal explants in culture, we extracted both RNA and DNA and sequenced the barcodes from both samples. The expression level of each barcoded reporter gene was calculated as the cDNA barcode sequence counts normalized to the DNA barcode counts (Dataset S1). To test the reliability of CRE-seq, we compared the expression of 12 RhoCRE3 variants measured by both CRE-seq and a standard fluorescence reporter assay (14). We observed strong correlation between both measurements ($R^2$ = 0.95; Fig. 1*C*), providing evidence that CRE-seq accurately quantifies gene expression in this system.

In a previously reported in vitro multiplex reporter gene assay (16), sequence barcodes were placed adjacent to the TSS and had large effects on reporter gene expression. To ameliorate this problem, we placed our barcodes in the 3′ UTR of DsRed, far from the start site of transcription. To demonstrate that the barcodes did not interfere with our assay, we performed an

Author contributions: J.C.K., I.M., C.A.M., J.C.C., and B.A.C. designed research; J.C.K., I.M., and C.A.M. performed research; J.C.K., I.M., and J.C.C. analyzed data; and J.C.K., I.M., J.C.C., and B.A.C. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]J.C.K. and I.M. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: cohen@genetics.wustl.edu or jcorbo@wustl.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1210678109/-/DCSupplemental.

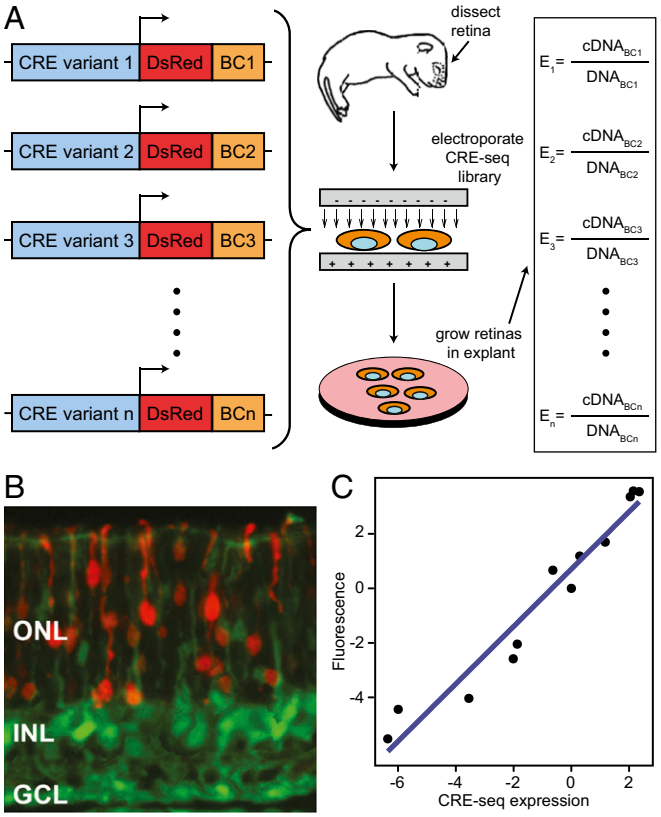


**Fig. 1.** (*A*) Schematic of the CRE-seq method. Each CRE variant in the library is fused to a reporter gene marked by a unique DNA barcode in the 3′ UTR of DsRed. The library is electroporated into newborn mouse retinas, which are cultured as explants for 8 d. Barcodes are then sequenced from harvested mRNA and DNA. The cDNA/DNA ratio of each barcode is a quantitative measure of the expression levels driven by each CRE variant in the library. (*B*) Cell type specificity of the CRE-seq library. A cross section of an electroporated retina shows DsRed expression in rod photoreceptors residing in the outer nuclear layer (ONL). Green fluorescence driven by a ubiquitously expressing CAG-GFP construct is observed in all layers including the inner nuclear layer (INL) and the ganglion cell layer (GCL). (*C*) Correlation between CRE-seq and fluorescent reporter genes. Twelve RhoCRE3 variants were quantified by using a standard dual-color fluorescent reporter gene assay (14, 15) and also by using CRE-seq. *x* axis, CRE-seq expression, $\log_2$(variant RhoCRE3/wild-type RhoCRE3); *y* axis, fluorescent reporter gene expression, $\log_2$(fluorescence of variant RhoCRE3/fluorescence of wild-type RhoCRE3).

experiment in which we fused more than 100,000 different barcodes to a single promoter, the wild-type RhoCRE3, and measured the expression of this control library in electroporated retinas. If barcodes exerted a consistent effect on expression, we would expect to observe a correlation between replicates of this control library, because barcodes with an activating effect would reproducibly increase expression, whereas barcodes with a repressive effect would reproducibly decrease expression. No such correlation was found ($R^2 = 0.04$; Fig. S1*A*), which suggests that the variation between members of the control library resulted from experimental noise rather than the barcode sequences. In contrast, the correlation between replicates of our library of RhoCRE3 mutants was high ($R^2 = 0.95$; Fig. S1*B*, Dataset S2). Together, our results suggest that the 3′ UTR barcodes have little effect in our assay and that there are strong and reproducible differences between RhoCRE3 variants in our library.

**Analysis of Single Mutants.** We first analyzed expression data from single nucleotide substitutions in RhoCRE3. We found that 86% of single nucleotide substitutions significantly altered expression (Welch's *t* test, $P < 0.05$), with many substitutions causing large changes (>30-fold) in expression (Fig. 2*A*). For 98% (51/52) of the positions in RhoCRE3, at least one of the three possible substitutions had a significant effect on reporter gene expression. At 20% of positions, different substitutions showed opposite effects on expression depending on the identity (A, G, C, or T) of the substituted base. Our results suggest that RhoCRE3 is an element whose function is finely tuned and that the majority of single nucleotide substitutions alter this function.

This finding differs dramatically from those of Melnikov et al. (18) and Patwardhan et al. (19), two recent studies that use methods similar to CRE-seq to analyze mammalian enhancers. In these studies, very few single nucleotide substitutions resulted in significant changes in expression. All single nucleotide changes in Melnikov et al. (18) and Patwardhan et al. (19) showed changes of <3.5-fold, with the vast majority of effects at <1.5-fold. In contrast, 49% of the single nucleotide substitutions we measured had effects between 3.5- and 30-fold. This sharp difference in results may reflect functional differences between the CREs in each study, or it may be due to differences in the technologies used by the different groups.

Single mutants with significant effects delineate five distinct regions across RhoCRE3 (Fig. 2*A*). Regions A and E contain known CRX sites, and region D contains a known NRL site (12, 13). Regions B and C do not correspond to previously identified regulatory sequences in RhoCRE3. Regions of RhoCRE3 that show evolutionary conservation in mammalian lineages coincide with regions that have strong effects on expression (Fig. 2*B*). However, the modest correlation between effect size and conservation ($R^2 = 0.31$) shows that the degree of conservation does not quantitatively predict the effect size of individual mutations, although it does delineate functional regions.

**Binding Site Mutations.** We examined the behavior of mutations in regions A and E, the known CRX binding sites, as well as region D, the region that contains the known NRL binding site. We generated activity logos (20) for these regions using the single nucleotide substitution data (Fig. S2 *A*–*C*). Using the average log-likelihood ratio (ALLR) test (21), we found that the activity logos for regions A and D are similar ($P < 0.05$) to position weight matrices (PWMs) generated for these TFs in previous studies (14, 22). Although there is a validated CRX site in region E, the activity logo made from region E does not show statistically significant similarity to the CRX PWM ($P = 0.22$), but it does show qualitative similarity. This finding suggests that the effect of mutations on expression does not depend solely on their effects on TF binding affinity, an issue we addressed in more detail.

We examined the quantitative relationship between the predicted affinity of sequences for CRX or NRL and gene expression levels. Using a PWM model of CRX specificity derived from quantitative in vitro gel shift assays (14), we computed the predicted change in affinity for each mutation in regions A and E for CRX. We then compared the predicted effect of each mutation on CRX affinity to its observed effect on expression (Fig. 2*C*). We performed a similar analysis for region D using a PWM that describes NRL binding specificity (22). For all three sites, the wild-type sequence was predicted to have high relative affinity and drove high expression. However, although mutations that decrease predicted affinity tended to have lower expression, overall the correlation between predicted affinity and observed expression was modest [$R^2 = 0.22$ CRX(1); $R^2 = 0.10$ CRX(2); $R^2 = 0.41$ NRL]. In all three regions, several mutations that created sequences with predicted affinity as high as the wild-type sequence exhibited markedly decreased expression. Our data suggest that variables besides the affinities of CRX and NRL also help determine the in vivo activity of RhoCRE3 variants. Mutations in the binding sites for CRX and NRL may have effects on the binding of other TFs.

Regions A, D, and E of RhoCRE3 contain binding sites for known transcriptional activators (CRX, NRL), and accordingly these regions contain dense clusters of single nucleotide

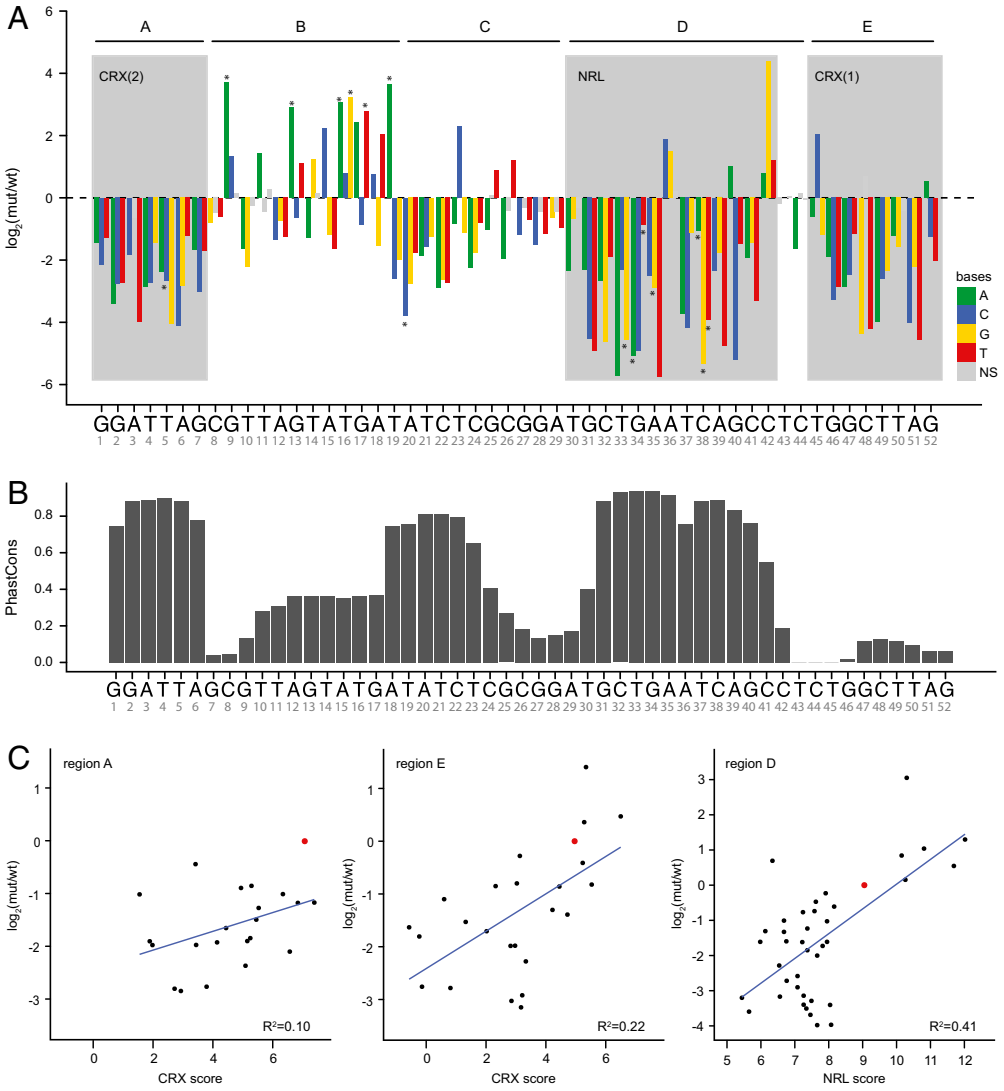


**Fig. 2.** (*A*) Effects of single nucleotide mutations on reporter expression. *x* axis, nucleotide position in RhoCRE3; *y* axis, relative expression by CRE-seq quantified as the $\log_2$(variant RhoCRE3/wild-type RhoCRE3). Colored bars represent substitutions whose expression was significantly different from wild-type (Welch's *t* test, $P < 0.05$). Gray bars represent substitutions that are not significantly different from wild-type. Each position has three bars representing the three possible substitutions at that position. For analysis we divided RhoCRE3 into five regions (A–E). Experimentally validated binding sites for CRX and NRL are shaded in gray. Asterisks mark the locations of substitutions that create new CRX binding sites (≥2% predicted affinity relative to consensus). (*B*) Phylogenetic conservation of nucleotides in RhoCRE3. *x* axis, position in RhoCRE3; *y* axis, PhastCons score derived from a multiple alignment of 28 mammalian sequences (17). (*C*) Relationship between the effects of mutations on the predicted binding affinity of TFs versus the mutations' effects on gene expression. *x* axis, predicted affinity of RhoCRE3 mutations for CRX or NRL; *y* axis, observed effects of mutations on gene expression. In each graph the red dot represents the wild-type sequence of RhoCRE3, and the blue line represents the linear regression model.

substitutions that decrease expression (Fig. 2*A*). Because region C shows a similar pattern of variants with decreased expression, we hypothesized that it also represents a TFBS. Using the single mutant expression data, we created an activity logo of this sequence (Fig. S2*D*). The activity logo generated for region C does not match any of the sequence logos in the Transfac database (23), which suggests that if these mutants effect the binding of a TF, it is a TF whose binding specificity has yet to be determined. However, the activity logo for the region overlapping with region B does show similarity to the CRX PWM ($P < 0.05$, ALLR test), which suggests that many mutations in region B exert their effects by creating a CRX site (Fig. S2*E*).

Region B is qualitatively different from the other regions of RhoCRE3 in that substitutions in region B often increase, rather than decrease expression. This finding suggests that mutations in region B either disrupt a binding site for a repressor or create new binding sites for activators. Because the density of mutations with large effects in this region is less than the density in regions with known TFBS, we favor the hypothesis that these mutations create binding sites for an activator. An obvious candidate for this activator is CRX.

To identify mutations that create CRX binding sites, we searched the entire library of single nucleotide mutants for matches to the CRX PWM. Mutations that create sequences with low predicted affinity to CRX are marked with an asterisk in Fig. 2*A*. Surprisingly, mutations that create putative CRX sites, as defined by a PWM score threshold, are not distributed randomly throughout RhoCRE3; rather, these mutations cluster in regions B and D. This finding suggests that both regions B and D contain sequences that are very close to CRX binding sites. We analyzed these sequences in more detail.

Region B corresponds to a region of high evolutionary conservation in vertebrates. We scanned the orthologous regions from multiple vertebrate species with our CRX PWM and found

that in most other mammals, region B contains a sequence with a low-affinity match to CRX PWM (Fig. S3). The presence of this putative low-affinity CRX site in most mammals accounts for the strong evolutionary conservation of region B. However, neither the rat nor the mouse genome has a potential CRX site in region B as defined by the same PWM score threshold. By using a neutral rate of 0.233 substitutions/site (24), the three observed differences between mouse and rat in region B were not significantly different from the 2.33 that are expected under neutrality. This finding suggests that these two species are accumulating substitutions in region B at a rate consistent with neutral evolution and that there has been recent turnover of CRX sites in region B in the rodent lineage. Together, these results suggest that mutations in region B that create CRX sites increase expression by re-creating sequences that resemble the ancestral form of RhoCRE3.

In contrast to region B, all mutations that create sequences matching CRX sites in region D decrease expression. In region D, these new hypothetical CRX sites disrupt the known NRL site. These overlapping CRX and NRL binding sites create complex expression patterns that suggest competition between the two factors. For example, mutations at position 36 suggest competition between NRL and CRX binding. All three substitutions at position 36 increase the predicted affinity of NRL, but only the two substitutions that also decrease the predicted affinity of CRX increase expression (Fig. 3). Only some mutations in region D support the competition model, suggesting that other factors play a role in regulation of region D (Fig. S4). Further experiments are needed to determine whether CRX and NRL actually compete for binding to the sequence in region D.

Other position-specific effects on expression cannot be explained by CRX and NRL site affinity. For example, mutations at position 42 have a dramatic effect on expression but a relatively minor effect on predicted NRL or CRX affinity. Mutations at position 42 may create a new TFBS (Fig. S2*F*) or alter the DNA helix structure (Fig. S5) (25, 26).

**Analysis of Double-Mutant CRE Variants.** Our library also contained all possible double mutations between the NRL site in region D and the CRX site in region E. On average, the double-mutant RhoCRE3 variants have lower expression than the single mutants ($P < 0.05$, Wilcoxon). In addition, many (58%) double mutants have effect sizes that are larger than either of their component single mutations. We used linear modeling (*Materials and Methods*) to identify significant nonadditive interactions between mutations and, if present, to determine the strength and direction of such interactions. Using ANOVA, we found that 82% of double mutants had a significant interaction between positions that cannot be explained by an additive model of single-mutant expression values ($P < 0.01$, F test). Our results contrast with those of Melnikov et al. (18) and Patwardhan et al. (19), who found very few interactions between mutations in enhancers. Our data show that some positions have interactions with many other positions, whereas other positions do not participate in any interactions (Fig. 4*A*). For some positions, a particular substitution can have an interaction that is either more or less than additive, depending on the position of the second mutation (Fig. 4*B*).

In addition, we observed that interactions between pairs of mutations are usually nucleotide-specific, rather than position-specific. Most models are improved with the addition of some, but not all, of the nine possible interaction terms ($P < 0.01$, F test). Mutations at two particular residues can combine to be either more or less than additive, depending on the specific identities of the substituted nucleotides (Fig. 4*C*). Combinations of mutations give rise to expression levels that cannot easily be predicted given the expression levels of single mutants. Physical interactions between NRL and CRX (27) likely underlie some of the complexity we observe, but more detailed physical models will be required to unravel the molecular basis of these nonlinear genetic interactions.

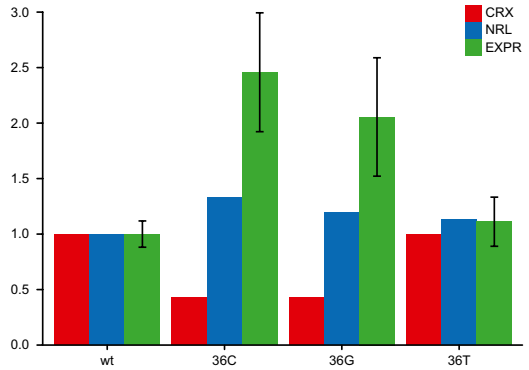


**Fig. 3.** Complex interactions between CRX and NRL in region D. The effects of mutations at position 36 support a model of competition between NRL and CRX. The *x* axis shows the identity of the base at position 36. Red bars, relative affinity of CRX normalized to the wild-type sequence; blue bars, relative affinity of NRL normalized to the wild-type sequence; green bars, expression normalized to the wild-type sequence. Errors bars represent the SEM.

## Discussion

Using CRE-seq, we found that single-nucleotide substitutions in RhoCRE3 often result in large increases or decreases in expression, suggesting that this element may be highly constrained as to which mutations it tolerates through evolution—a hypothesis supported by the strong evolutionary conservation of this region. Changes in the predicted affinity of TFBS to known TFs partially explain the effects of these mutations, but they cannot explain the full in vivo activity of substitutions. More complex phenomena, such as recent binding site turnover as well as competition between CRX and NRL, may help explain the effects of some substitutions. The fact that the majority of double mutants have expression levels consistent with a nonadditive model between positions suggests that interactions between CRX and NRL underlie some of the complex behaviors of RhoCRE3 mutants. Consistent with the many other studies, our results suggest that the interpretation of noncoding polymorphisms would require consideration of both low-affinity sites and the creation of novel sites. Overall, our results paint a picture of CREs as complex regulatory elements whose function can easily be altered by subtle changes to their nucleotide sequences.

In contrast, Melnikov et al. (18) and Patwardhan et al. (19) conclude that CREs are largely redundant at the nucleotide level, such that few mutations create large changes in expression. Our results showing large effects of single-nucleotide substitutions agree with many detailed studies of single reporter genes (for example, refs. 28–36). We also show evidence for extensive nonlinear interactions between CRE mutations, whereas both Melnikov et al. (18) and Patwardhan et al. (19) conclude that mutations in enhancers act independently of each other. What might account for these dramatic differences in conclusions?

Some differences between these studies and our study are likely due to the CREs each group chose to examine. We chose to study an extensively validated CRE (8, 9) that drives the expression of *Rho*, the most highly expressed gene in the mammalian retina. This element shows strong evolutionary conservation throughout the vertebrate lineage and lies proximal to the start site of transcription. These features may make RhoCRE3 activity in CRE-seq susceptible to substitutions in ways that other CREs are not.

Technical differences between protocols probably contribute to some of the differences between these studies. Template switching during the PCR cycles used for library preparation can result in high rates of chimerism (37), which can disrupt the unique correspondence between CRE variants and barcodes. Chimerism decreases both the accuracy and dynamic range of the assay. In Melnikov et al. (18) and Patwardhan et al. (19), limited dynamic range due to the extensive use of PCR may have made it appear that enhancers are robust to mutation. In contrast, our method

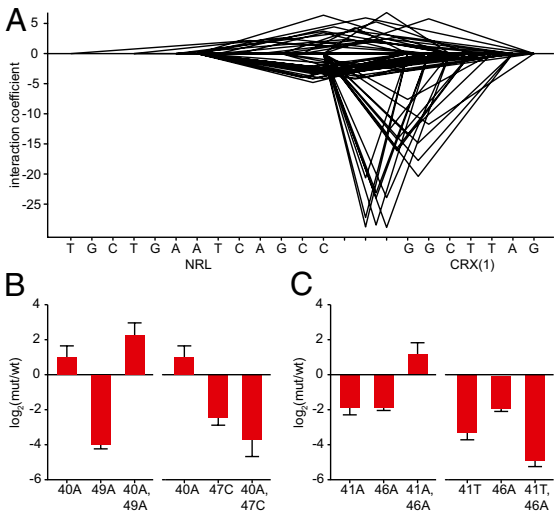


**Fig. 4.** Interactions between mutations in TFBS within regions D and E. (*A*) Pattern of interactions. Each arc connects two nucleotides that have a significant interaction term by ANOVA (*Materials and Methods*). The height of the arc represents the magnitude of the interaction term. Only interactions greater than 3 or less than −3 are shown. (*B*) The magnitude and direction of interactions are position-specific. The interaction between mutations 40A (CRX) and 49A (NRL) is significantly more than additive and greater than the effect of either single mutant, whereas the interaction between mutations 40A and 47C (NRL) is less than additive and greater than the effect of either single mutant. (*C*) Interactions between the same positions are base-specific. The interaction between 41A (CRX) and 46A (NRL) is significantly more than additive, whereas the interaction between 41T and 46A is less than additive.

makes limited use of PCR amplification (*Materials and Methods*) to avoid the creation of chimeras. Finally, we chose to focus on only 1,040 mutations in a single experiment, which allowed us to obtain very high sequence coverage of our library. Both Melnikov et al. (18) and Patwardhan et al. (19) assayed much larger numbers of mutant promoters, making low sequence coverage and the resulting loss of statistical power a significant issue in these studies. In Patwardhan et al. (19), low sequence coverage of their library necessitated the use of an indirect measure of barcode counts as the metric of expression.

We have demonstrated that CRE-seq is a robust technology for assaying large numbers of CRE mutations. CRE-seq has a large dynamic range, nucleotide level resolution, and excellent reproducibility. Our analyses revealed a surprising amount of previously unknown complexity in RhoCRE3, an element that has been extensively studied. Our results demonstrate that nucleotides within CREs interact in complex and often nonintuitive ways to produce regulated patterns of gene expression. We anticipate that CRE-seq will be an important tool for unraveling the rules that govern the function of CREs.

## Materials and Methods

**CRE-seq Library Construction.** To create the CRE-seq library plasmid backbone, we replaced the NotI site in plasmid Rho_minprox-DsRed (8, 9) with an EagI–XhoI–ClaI polylinker, creating plasmid pJK01. We then engineered sites for MfeI at position 25 and KasI at position 102 (following the numbering in refs. 8 and 9), creating pJK02.

The wild-type RhoCRE3 sequence spans positions 115,881,830–115,881,881 on mouse chromosome 6 (NCBI37/mm9). This region corresponds to the sequence between positions 68 and 120.

A pool of 5,720 unique 150-mer oligos was ordered through a limited licensing agreement with Agilent Technologies. Oligos were designed with the following structure: JKP1F, MfeI site, RhoCRE3 variant, KasI site, EcoRI site, EagI site, 9 bp barcode, ClaI site, and JKP1R (Table S1). RhoCRE3 variants were designed between positions 68 and 120.

We amplified the oligo pool using four cycles of PCR with Phusion High-Fidelity polymerase (New England Biolabs) and primers JKP1F and JKP1R. We cloned the amplicon into pJK01 using MfeI and ClaI. We prepared DNA from 48,000 colonies to generate library PL5_1. We then cloned the minimal *Rho* promoter driving DsRed into PL5_1. A cassette containing the *Rho* minimal promoter fused to DsRed was amplified from pJK01 with primers JKP2F and JKP2R (Table S1). The PCR amplicon was cloned into library PL5_1 by using KasI and EagI, creating library PL5_2. To select for library members with full-length RhoCRE3 variants, we linearized PL5_2 with HindIII, gel purified the library, and recircularized to create library PL5_3.

We created library (BL_1) to determine the effect of barcode sequences on reporter gene expression. Barcode sequence inserts JKP5F and JKP5R (Table S1) were annealed and cloned into pJK01 by using EagI and ClaI. We prepared library DNA from 100,000 colonies.

**Retinal Electroporation.** Electroporations and explant cultures were performed as described (9) by using 0.5 μg/mL library PL5_3 as well as 0.5 μg/mL *Rho* minimal promoter driving GFP for visualization of electroporation efficiency (14). After 8 d in culture, retinas were washed twice in sterile HBSS (Gibco, Life Technologies), and total RNA and DNA were extracted by using TRIzol according to manufacturer's instructions (Ambion, Life Technologies).

**Comparison Between a Fluorescence Assay and CRE-seq** Twelve previously characterized RhoCRE3 variants (14, 15) were excised from their respective vectors with XbaI and KpnI and cloned into the barcode library BL_1. For each variant, we isolated 10 uniquely barcoded constructs. We mixed all 120 constructs together for CRE-seq analysis.

**Preparing Samples for RNA-Seq.** Isolated RNA was treated with DNaseI (Ambion) to eliminate potential genomic DNA contamination. The first strand of cDNA was synthesized with Invitrogen SuperScript II reverse transcriptase by using oligo-dT primers. After first strand synthesis, the reaction was treated with RNase H (NEB) to remove RNA. The 3′ UTR of the DsRed gene, including the barcode sequence, was amplified from both cDNA and isolated plasmid DNA samples. Amplification (98 °C for 1 min, 21 cycles: 98 °C for 10 s, 58 °C for 30 s, 72 °C for 30 s, and 72 °C for 5 min; NEB HF Phusion MM) of the cDNA and plasmid DNA samples isolated by PCR using primers JKP3F and JKP3R (Table S1) yielded barcode sequences that were flanked with EagI and EcoRI restriction enzyme sites. The products were purified with Qiagen QIAquick PCR Purification kit and digested with EagI and EcoRI. Illumina adapter sequences were ligated onto these overhangs. This product was amplified (98 °C for 1 min, 21 cycles: 98 °C for 30 s, 65 °C for 30 s, 72 °C for 30 s, and 72 °C for 5 min) with HF Phusion MM by using primers JKP4F and JKP4R (Table S1) to enrich for molecules that contain both adapter sequences.

To measure expression of the barcoded library, electroporation replicates were multiplexed and run on two lanes of an Illumina HiSeq machine, which generated 48.2 million sequence reads corresponding to cDNA and 48.8 million reads corresponding to DNA. Sequencing reads that matched the first 20 nucleotides of designed sequence were counted, regardless of quality score. Only barcodes with >10 reads in either cDNA or DNA pools were used for analysis.

**Determination of Expression Levels of *Rho* Variants.** We determined the expression levels of each RhoCRE3 variant by computing the average cDNA/DNA ratio for all barcodes that marked the same RhoCRE3 variant (Dataset S2). We also computed the SEM for each average in each replicate. We then averaged the averages from the three replicates and propagated the SE using the following formula:

$$SE\left(\frac{\sum E_n}{n}\right) = \sqrt{\sum\left(\frac{SE(E_n)}{n}\right)^2}.$$

The propagation of SE when computing the ratio of mutant to wild-type expression levels was calculated as:

$$SE\left(\frac{E_{mut}}{E_{wt}}\right) = \sqrt{\frac{E_{mut}^2}{E_{wt}^2}\left[\left(\frac{SE(E_{mut})^2}{E_{mut}^2}\right) + \left(\frac{SE(E_{wt})^2}{E_{wt}^2}\right)\right]}.$$

**Analysis of Binding Site Mutations.** The CRX PWM was derived from quantitative binding affinity data (14) according to ref. 38. Similarly, a NRL PWM was derived from sequence data from ref. 22. The two TF matrices were scored against every position of each CRE variant by using *patser* (39).

**Nonadditive Model for Variants with Two Substitutions.** Linear regression was used to determine the extent of nonadditive expression in the variants with

two substitutions. First, the following two models were generated for each combination of positions and compared by using ANOVA ($P < 0.01$):

$$E_{1,2} = WT + \beta_1 mut_1 + \beta_2 mut_2$$

$$E_{1,2} = WT + \beta_1 mut_1 + \beta_2 mut_2 + \beta_3 mut_{1,2} \ .$$

Models that were significantly improved (82%, 75/91) by the addition of the interaction term were further analyzed for base-specific interactions. After selecting positions with significant nonadditive expression, we used ANOVA ($P < 0.01$) to test each of the nine possible base-specific interactions between positions.

All experiments were conducted in accordance with the Guide for the Care and Use of Laboratory Animals and the Animal Welfare Act and were approved by the Washington University in St. Louis Institutional Care and Use Committee.

**ACKNOWLEDGMENTS.** We thank Heather Lawson and Jim Cheverud for advice on statistical analyses; Sung Chun and Justin Fay for advice on models of molecular evolution; Aaron Spivak for assistance with TF binding analysis; Chris Fiore for discussions of CRE-seq technical issues; and Tim Tullius and Dana Zafiropoulos for help with DNA structure analysis. This work was supported by National Institutes of Health Grants GM092910 (to B.A.C.), EY018826 (to J.C.C.), and HG006346 (to B.A.C. and J.C.C.).

1. Bernstein BE, et al. (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 28(10):1045–1048.
2. Birney E, et al.; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146):799–816.
3. LeProust EM, et al. (2010) Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res* 38(8): 2522–2540.
4. Mardis ER (2008) Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 9:387–402.
5. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):621–628.
6. Humphries MM, et al. (1997) Retinopathy induced in mice by targeted disruption of the rhodopsin gene. *Nat Genet* 15(2):216–219.
7. Olsson JE, et al. (1992) Transgenic mice with a rhodopsin mutation (Pro23His): A mouse model of autosomal dominant retinitis pigmentosa. *Neuron* 9(5):815–830.
8. Corbo JC, et al. (2010) CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Res* 20(11):1512–1525.
9. Hsiau TH, et al. (2007) The cis-regulatory logic of the mammalian photoreceptor transcriptional network. *PLoS ONE* 2(7):e643.
10. Lem J, Applebury ML, Falk JD, Flannery JG, Simon MI (1991) Tissue-specific and developmental regulation of rod opsin chimeric genes in transgenic mice. *Neuron* 6(2): 201–210.
11. Zack DJ, et al. (1991) Unusual topography of bovine rhodopsin promoter-lacZ fusion gene expression in transgenic mouse retinas. *Neuron* 6(2):187–199.
12. Chen S, Zack DJ (1996) Ret 4, a positive acting rhodopsin regulatory element identified using a bovine retina in vitro transcription system. *J Biol Chem* 271(45):28549–28557.
13. Rehemtulla A, et al. (1996) The basic motif-leucine zipper transcription factor Nrl can positively regulate rhodopsin gene expression. *Proc Natl Acad Sci USA* 93(1):191–195.
14. 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 Ther* 17(11):1390–1399.
15. Montana CL, Myers CA, Corbo JC (2011) Quantifying the activity of cis-regulatory elements in the mouse retina by explant electroporation. *J Vis Exp*(52).
16. Patwardhan RP, et al. (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* 27(12):1173–1175.
17. Miller W, et al. (2007) 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* 17(12):1797–1808.
18. Melnikov A, et al. (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* 30(3): 271–277.
19. Patwardhan RP, et al. (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* 30(3):265–270.
20. Shin I, Kim J, Cantor CR, Kang C (2000) Effects of saturation mutagenesis of the phage SP6 promoter on transcription activity, presented by activity logos. *Proc Natl Acad Sci USA* 97(8):3890–3895.
21. Wang T, Stormo GD (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* 19(18):2369–2380.
22. Kataoka K, Noda M, Nishizawa M (1994) Maf nuclear oncoprotein recognizes sequences related to an AP-1 site and forms heterodimers with both Fos and Jun. *Mol Cell Biol* 14(1):700–712.
23. Wingender E, Dietze P, Karas H, Knüppel R (1996) TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 24(1):238–241.
24. Chun S, Fay JC (2009) Identification of deleterious mutations within three human genomes. *Genome Res* 19(9):1553–1561.
25. 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.
26. Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH (2009) Local DNA topography correlates with functional noncoding regions of the human genome. *Science* 324(5925):389–392.
27. Mitton KP, et al. (2000) The leucine zipper of NRL interacts with the CRX homeodomain. A possible mechanism of transcriptional synergy in rhodopsin regulation. *J Biol Chem* 275(38):29794–29799.
28. Goodbourn S, Maniatis T (1988) Overlapping positive and negative regulatory domains of the human beta-interferon gene. *Proc Natl Acad Sci USA* 85(5):1447–1451.
29. Gurnett CA, et al. (2007) Two novel point mutations in the long-range SHH enhancer in three families with triphalangeal thumb and preaxial polydactyly. *Am J Med Genet A* 143(1):27–32.
30. Lettice LA, et al. (2012) Opposing functions of the ETS factor family define Shh spatial expression in limb buds and underlie polydactyly. *Dev Cell* 22(2):459–467.
31. Myers RM, Tilly K, Maniatis T (1986) Fine structure genetic analysis of a beta-globin promoter. *Science* 232(4750):613–618.
32. Shimell MJ, Simon J, Bender W, O'Connor MB (1994) Enhancer point mutation results in a homeotic transformation in Drosophila. *Science* 264(5161):968–971.
33. Singh K, Tokuhisa JG, Dennis ES, Peacock WJ (1989) Saturation mutagenesis of the octopine synthase enhancer: Correlation of mutant phenotypes with binding of a nuclear protein factor. *Proc Natl Acad Sci USA* 86(10):3733–3737.
34. Smith JR, Osborne TF, Goldstein JL, Brown MS (1990) Identification of nucleotides responsible for enhancer activity of sterol regulatory element in low density lipoprotein receptor gene. *J Biol Chem* 265(4):2306–2310.
35. Weiher H, König M, Gruss P (1983) Multiple point mutations affecting the simian virus 40 enhancer. *Science* 219(4585):626–631.
36. Yun Y, Adesanya TM, Mitra RD (2012) A systematic study of gene expression variation at single-nucleotide resolution reveals widespread regulatory roles for uAUGs. *Genome Res* 22(6):1089–1097.
37. Stemmer WP (1994) DNA shuffling by random fragmentation and reassembly: In vitro recombination for molecular evolution. *Proc Natl Acad Sci USA* 91(22):10747–10751.
38. Stormo GD, Fields DS (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci* 23(3):109–113.
39. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res* 10(9): 2997–3011.