A MASSIVELY PARALLEL REPORTER ASSAY REVEALS CONTEXT-DEPENDENT ACTIVITY OF HOMEODOMAIN BINDING SITES *IN VIVO*

Andrew E. O. Hughes[1], Connie A. Myers[1], and Joseph C. Corbo[1*]

[1]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, Missouri 63110, USA

*To whom correspondence should be addressed. Tel: +1 314 362 6254; Fax: +1 314 362 4096; Email: jcorbo@pathology.wustl.edu

Running title: Context-dependent activity of CRX binding sites

Key words: homeodomain, cis-regulatory element, massively parallel reporter assay, transcription factor binding site, retina

ABSTRACT

Cone-rod homeobox (CRX) is a paired-like homeodomain transcription factor (TF) and a master regulator of photoreceptor development in vertebrates. The *in vitro* DNA binding preferences of CRX have been described in detail, but the degree to which *in vitro* binding affinity is correlated with *in vivo* enhancer activity is not known. In addition, paired-class homeodomain TFs can bind DNA cooperatively as both homodimers and heterodimers at inverted TAAT half-sites separated by two or three nucleotides. This dimeric configuration is thought to mediate target specificity, but whether monomeric and dimeric sites encode distinct levels of activity is not known. Here, we used a massively parallel reporter assay to determine how local sequence context shapes the regulatory activity of CRX binding sites in mouse photoreceptors. We assayed inactivating mutations in >1,700 TF binding sites and found that dimeric CRX binding sites act as stronger enhancers than monomeric CRX binding sites. Furthermore, the activity of dimeric half-sites is cooperative, dependent on a strict three-base-pair spacing, and tuned by the identity of the spacer nucleotides. Saturating single-nucleotide mutagenesis of 195 CRX binding sites showed that, on average, changes in TF binding site affinity are correlated with changes in regulatory activity, but this relationship is obscured when considering mutations across multiple *cis*-regulatory elements (CREs). Taken together, these results demonstrate that the activity of CRX binding sites is highly dependent on sequence context, providing insight into photoreceptor gene regulation and illustrating functional principles of homeodomain binding sites that may be conserved in other cell types.

INTRODUCTION

Advances in high-throughput sequencing have enabled genome-wide mapping of *cis*-regulatory elements (CREs) in diverse cell types and tissues, providing powerful resources for studying the role of non-coding genetic variation in health and disease (The ENCODE Project Consortium 2012). Nevertheless, predicting the functional impact of regulatory variants requires understanding the sequence constraints mediating interactions between CREs and transcription factors (TFs). The DNA binding preferences of thousands of TFs have been characterized *in vitro* (Badis et al. 2009; Jolma et al. 2013; Weirauch et al. 2014), but how accurately these models predict the regulatory activity of TF binding sites *in vivo* is not known.

Cone-rod homeobox (CRX) is a paired-like homeodomain TF and a master regulator of photoreceptor gene expression in vertebrates (Chen et al. 1997; Livesey et al. 2000; Hsiau et al. 2007). The DNA binding preferences of CRX and the closely related homologs OTX1 and OTX2 have been defined by quantitative gel shift, high-throughput SELEX, and protein binding microarray, all of which identify the high-affinity consensus sequence 5'-TAATCC-3' (Chatelain et al. 2006; Lee et al. 2010; Jolma et al. 2013; Barrera et al. 2016). In addition, structural studies have shown that paired-class homeodomains can bind DNA cooperatively as both homodimers and heterodimers at inverted TAAT repeats (Wilson et al. 1993; Wilson et al. 1995; Tucker and Wisdom 1999). Paired-class TFs with a lysine (K) or a glutamine (Q) in position 50 of the homeodomain (K50 or Q50) bind dimeric half-sites with a three-base-pair spacing, while those with a serine (S) in position 50 (S50) bind with a two-base-pair spacing. Furthermore, K50 homeodomains, including CRX, prefer cytosines 3' of each TAAT half-site (5'-TAATCNGATTA-3'). In an earlier study, we mapped CRX occupancy in mouse photoreceptors by ChIP-seq, which showed that CRX-bound regions *in vivo* are enriched for both monomeric and dimeric CRX binding sites (Corbo et al. 2010). However, whether monomeric and dimeric CRX binding sites promote distinct levels of transcriptional activation is not known.

Recently, massively parallel reporter assays (MPRAs) have emerged as powerful tools for quantifying the regulatory activity of many CREs simultaneously. These assays work by introducing libraries of barcoded reporter constructs into cells of interest followed by harvesting RNA and counting barcodes by sequencing to quantify activity. MPRAs can be designed to measure either the promoter activity of CREs (the level of expression they drive autonomously) or their enhancer or repressor activity (the level of expression they drive above or below that of a basal promoter). To date, MPRAs have been used to quantify tissue- and cell-type-specific CRE activity in cell lines (Melnikov et al. 2012; Arnold et al. 2013; Grossman et al. 2017), explanted tissues (Kwasnieski et al. 2012; White et al. 2013; White et al. 2016), and *in vivo* (Patwardhan et al. 2012; Shen et al. 2016).

Previously, we used MPRAs to begin to elucidate how photoreceptor CRE activity is encoded in CRX binding sites. In one study, we quantified the effect of all possible single-nucleotide substitutions in a 52-bp segment of the *Rhodopsin* promoter (*pRho*) on its autonomous activity in mouse retina (Kwasnieski et al. 2012). We found that changes in the affinity of CRX binding sites within *pRho* are moderately correlated with changes in its activity. In addition, we observed interactions between pairs of mutations in binding sites for CRX and another photoreceptor TF, NRL. However, as we only analyzed a single element, the extent to which these results generalize to other photoreceptor CREs is not known.

In a subsequent study, we assayed the enhancer activity of thousands of 84-bp sequences corresponding to CRX-bound regions, CRX-unbound regions harboring high-affinity CRX binding sites, and scrambled controls (White et al. 2013). We found that CRX-bound, but not CRX-unbound regions, drive higher expression than scrambled controls, despite controlling for CRX binding site content. In addition, we showed that the activity of CRX-bound regions depends on CRX binding sites. These results indicate that individual CRX binding sites within CRX-bound regions are necessary, but not sufficient,

for enhancer activity. They also suggest that sequence context outside of primary binding sites distinguishes functional CRX binding sites from non-functional ones *in vivo*. Nevertheless, the sequence features that quantitatively predict photoreceptor CRE activity have not been clearly defined.

In the current study, we set out to build upon these results to better understand how multiple levels of sequence context influence the regulatory activity of CRX binding sites in mouse photoreceptors. First, we identified sequence features that predict CRX occupancy *in vivo* (as determined by ChIP-seq), and we compared these to sequence features that are correlated with enhancer activity (as measured by MPRA). In addition, we assayed the effect of inactivating mutations in monomeric vs. dimeric CRX binding sites to quantify their relative activity. Finally, we performed a dense mutagenesis of 195 CRX binding sites to examine the relationship between TF binding site configuration and regulatory activity at single-nucleotide resolution.

RESULTS

Combining dinucleotide frequencies and TF binding site content accurately predicts CRX occupancy *in vivo*

We previously used ChIP-seq to profile CRX occupancy in adult mouse photoreceptors, which showed that CRX-bound regions are phylogenetically conserved, have elevated GC content, and are enriched for K50 homeodomain binding sites (Corbo et al. 2010). We subsequently reported that none of these features alone accurately predicts CRX occupancy genome-wide (White et al. 2013). Here, we revisited these data to determine if models incorporating multiple predictors could accurately classify CRX-bound vs. CRX-unbound regions and provide insight into the sequence features that determine CRX occupancy *in vivo*.

For this analysis, we selected 5,250 200-bp sequences centered on CRX ChIP-seq peaks, focusing on distal enhancers (>1 kb upstream and >100 bp downstream of a TSS) (Fig. 1A). We then selected 52,500 200-bp CRX-unbound sequences sampled randomly from the mouse genome, controlling for GC and repeat content (Ghandi et al. 2016). We scored each CRX-bound and CRX-unbound sequence for dinucleotide frequencies as well as occurrences of 206 TF binding sites (Jolma et al. 2013). We found that CRX-bound regions are centered on significant enrichments in specific dinucleotide classes (e.g., GC and AG) as well as TF binding sites (e.g., monomeric and dimeric K50 binding sites) (Fig. 1B-C, Supplemental Fig. 1-2). Next, we used these features to train logistic regression classifiers to differentiate CRX-bound from CRX-unbound sequences, and we used lasso regularization to control model complexity (Tibshirani 1996).

To develop an intuition for the information contained in specific classes of features, we measured the performance of models using increasingly complex subsets of variables, quantified by area under the receiver operating characteristic (AUC-ROC) and area under the precision-recall curve (AUC-PR) (Supplemental Fig. 3, Supplemental Table 1). First, we considered a model using only dinucleotide frequencies, and we found that this model performs nearly as well as one using counts of CRX binding sites (AUC-ROC=0.75 vs. AUC-ROC=0.77, respectively). Next, we noted that modeling CRX binding site content by simply scoring sequences with a position weight matrix (PWM) and counting matches above a single threshold fails to account for TF binding site affinity. To address this, we binned counts of CRX sites into four 'affinity' classes (high, medium, low, and very low) based on the match p-value. This modification improves model performance (AUC-ROC=0.84), highlighting the value of incorporating graded TF binding site affinity into models of TF occupancy. Finally, to account for binding of multiple transcription factors, we trained a model using counts of 206 mouse and human TF binding sites (Jolma et

al. 2013). This approach further improves model performance (AUC-ROC=0.92), illustrating that multiple TF binding site models (even ostensibly similar homeodomain PWMs) capture non-redundant information in CRX ChIP-seq peaks.

Next, we combined dinucleotide frequencies and counts of TF binding sites in a single model, yielding the best performance among the logistic regression classifiers we tested (AUC-ROC of 0.95) (Fig. 1D). Of note, two advantages of regularized logistic regression are that (1) the model coefficients have an accessible interpretation (the contribution of each variable to the likelihood of CRX binding), and (2) lasso regularization shrinks the coefficients of irrelevant or redundant variables towards zero. From an initial set of 834 predictors, only 18 have non-zero coefficients in the final model: GC, AG, and CG dinucleotide frequencies, as well as six TF binding sites, nearly all of which correspond to monomeric or dimeric K50 binding sites (Supplemental Table 2). These results suggest that a substantial fraction of CRX occupancy can be explained by the presence of a limited set of homeodomain TF binding sites in a favorable dinucleotide context.

Recently, several groups have developed methods for predicting TF occupancy from *k*-mer content (DNA words of length k, typically six to ten base pairs), which have generally yielded highly accurate models (Fletez-Brant et al. 2013; Setty and Leslie 2015; Ghandi et al. 2016; Kelley et al. 2016). We used one of these tools, gkm-SVM, to classify CRX-bound vs. CRX-unbound regions using 'gapped' 11-mers (11-bp sequences with 7 informative positions), and we found that this model outperforms logistic regression (AUC-ROC=0.99) (Fig. 1D) (Ghandi et al. 2014). Of note, gkm-SVM learns weights for all non-redundant 11-mers, corresponding to the impact of specific sequences on the likelihood of CRX binding. Analogous to the model coefficients in logistic regression, these weights provide insight into the DNA binding preferences learned by gkm-SVM. For example, top-weighted 11-mers are highly enriched for K50 binding sites, especially dimeric K50 binding sites (Supplemental Table 2).

In addition to manually inspecting highly-weighted 11-mers, we calculated the median change in gkm-SVM score (deltaSVM) (Lee et al. 2015) due to mutations overlapping TF binding sites defined by experimentally-derived PWMs (Supplemental Fig. 4) (Jolma et al. 2013). This approach allowed us to systematically identify and quantify the relative importance of TF binding sites implicitly detected by gkm-SVM. Consistent with our logistic regression models, the majority of high-scoring PWMs correspond to homeodomain TFs (Supplemental Fig. 4). Furthermore, gkm-SVM detects binding sites for additional TF families (including MADS, zinc finger, basic helix-loop-helix, and basic leucine zipper TFs), although the scores associated with these motifs are much lower than those associated with homeodomain binding sites. Thus, while both logistic regression and gkm-SVM identify homeodomain binding sites as key sequence features mediating CRX occupancy, only gkm-SVM captures the contribution of additional TF families.

Finally, given that multiple models of primary sequence features accurately predict CRX occupancy *in vivo*, we asked if the location of informative features was spatially constrained relative to the center of CRX ChIP-seq peaks. To address this question, we re-trained our logistic regression classifier extracting features from windows ranging from 20 bp up to 200 bp (Fig. 1E). We found that the maximum AUC-ROC and AUC-PR values are obtained by restricting features to the central ~140 bp relative to the summit, suggesting that most of the information mediating CRX occupancy is contained within the footprints of individual nucleosomes (i.e., ~146 bp) (Luger et al. 1997).

Models that accurately predict CRX occupancy modestly predict CRE activity

Having identified sequence features that predict CRX occupancy, we asked if these same features could be used to predict the regulatory activity of CRX-bound regions *in vivo*. To quantify the activity of many CREs simultaneously, we performed CRE-seq as described previously (Fig. 2A) (Kwasnieski et al. 2012; White et al. 2013; Shen et al. 2016; White et al. 2016). Briefly, we used custom oligonucleotide synthesis to generate a library of 1,230 100-bp DNA fragments centered on native enhancers identified by CRX ChIP-seq (Corbo et al. 2010). We cloned this library upstream of a photoreceptor promoter driving DsRed, and we included a CRE-specific DNA barcode in the 3' UTR of each construct. We electroporated this library into newborn mouse retinas, which were then cultured for eight days. Finally, we harvested RNA and DNA and PCR amplified and sequenced barcodes to measure copy number-adjusted regulatory activity.

In a previous study, we found that up to 50% of putative photoreceptor CREs have little or no autonomous activity (Shen et al. 2016). We have also shown that CREs lacking autonomous activity can nevertheless act as potent enhancers when cloned upstream of a cell-type-specific promoter (Corbo et al. 2010). Therefore, in the current study, we assayed CREs in an enhancer or repressor context (i.e., upstream of a cell-type-specific promoter) to increase our sensitivity to detect elements with measurable activity. Specifically, we assayed the activity of CRX-bound regions cloned upstream of a 205-bp segment of the *Rho* promoter (*pRho*), which drives high expression in rod photoreceptors (Zack et al. 1991; Montana et al. 2011a). In addition, we assayed our library on a 206-bp segment of the *Crx* promoter (*pCrx*), which drives moderate expression in both rod and cone photoreceptors, to assess the effect of distinct promoters on CRE activity. We found that CRE-seq generates reproducible estimates of enhancer activity on both *pRho* and *pCrx* (Spearman's correlation coefficients between biological replicates of 0.98-0.99 and 0.93-0.94, respectively) (Supplemental Fig. 5), and we observed similar distributions of CRE activity on both promoters (Fig. 2B).

To identify sequence features associated with CRE activity, we examined the correlation between dinucleotide frequencies or counts of individual TF binding sites and CRE-seq expression. To give a specific example, we found that the number of E-box binding sites is positively correlated with CRE activity on both *pRho* and *pCrx* (Pearson correlation coefficient 0.15 and 0.09, respectively) (Fig. 2C). Overall, we identified 24 sequence features that were significantly correlated with CRE activity on either *pRho* or *pCrx* (FDR<0.05) (Fig. 2D). In contrast to the sequence features that predict CRX occupancy by logistic regression (GC, AG, and CG dinucleotides as well as K50 homeodomain binding sites), we found that a more complex set of dinucleotide frequencies and TF binding sites are correlated with regulatory activity. In addition, many of the features that are most strongly correlated with CRE activity are associated with modest decreases in gkm-SVM scores when mutated (e.g., E-box and nuclear receptor binding sites). These results suggest that the sequence features that determine the activity of CRX-bound regions differ in important ways from those that mediate CRX occupancy.

Overall, TF binding sites that are correlated with CRE activity belong to five distinct TF families (Fig. 2D). Binding sites for nuclear receptors, basic helix-loop-helix, and zinc finger TFs are positively correlated with activity, whereas binding sites for Q50 homeodomain and T-box TFs are negatively correlated with activity (K50 binding sites are discussed below). Almost all of these families correspond to TFs that play well-characterized roles in mouse photoreceptor development (Supplemental Fig. 6): *Esrrb*, *Nr2e3*, *Rorb*, *Rxrg*, and *Thrb* (nuclear receptors); *Neurod1* (a basic helix-loop-helix TF); *Sp1*, *Sp3*, and *Sp4* (zinc finger TFs); *Crx* and *Otx2* (K50 homeodomain TFs); and *Rax* (a Q50 homeodomain TF). In contrast, a role for T-box TFs in mammalian photoreceptors has not been established. Additionally, recent transcriptome profiling of mouse photoreceptors has shown that *Tbx2* is robustly expressed during the early postnatal period (Supplemental Fig. 6) (Kim et al. 2016a; Kim et al. 2016b). *Tbx3*, *Tbx5*, and *Tbx6* are also expressed, but at lower levels. Of note, all of these factors are downregulated by P14 and are not expressed in adult photoreceptors (Sowden et al. 2001), suggesting that TBX TFs may regulate photoreceptor gene expression specifically at early postnatal stages.

In addition to identifying sequence features that are correlated with CRE activity, we compared our CRE-seq data to previously generated genomic and epigenomic profiling data (Fig. 2E, Supplemental Table 3) (The ENCODE Project Consortium 2012; Hao et al. 2012; Wilken et al. 2015; Mo et al. 2016; Hughes et al. 2017). In particular, we examined the correlation between CRE activity and open chromatin data (ATAC-seq or DNase-seq) from whole retina, rod and cone photoreceptors, brain, heart, liver, B cells, and T cells. In addition, we examined the correlation between CRE activity and ChIP-seq data for photoreceptor-specific TFs (CRX and NRL) as well as histone modifications (H3K27ac, H3K4me3, H3K4me1, and H3K27me3) in whole retina (Supplemental Fig. 7, Supplemental Table 4). In general, we found that datasets generated in whole retina or photoreceptors have the strongest correlations with CRE-seq expression (Pearson correlation coefficients up to 0.30 on *pRho* and 0.26 on *pCrx*), suggesting that CRE-seq captures meaningful cell-type-specific regulatory activity. Specifically, we found that activating histone marks (H3K27ac and H3K4me1), ChIP-seq for CRX and NRL, and whole-retina and photoreceptor chromatin accessibility are positively correlated with CRE activity, whereas the repressing histone mark H3K27me3 and non-retina chromatin accessibility are negatively correlated with CRE activity.

Finally, we asked if the same models that accurately classify CRX-bound vs. CRX-unbound regions (Fig. 1D) could also be used to predict the regulatory activity of these regions. To evaluate this, we defined CREs with expression >3-fold above the median as having high activity and CREs with expression within 1.2-fold of the median as having low activity. Compared to their accuracy in classifying CRX-bound vs. unbound regions (Fig. 1D), both regularized logistic regression and gkm-SVM perform modestly in predicting high vs. low CRE activity (AUC-ROC=0.61 on *pRho* and 0.71 on *pCrx* for regularized logistic regression, and AUC-ROC=0.61 on *pRho* and 0.74 on *pCrx* for gkm-SVM) (Fig. 2F). In parallel, given that we observed relatively strong correlations between orthogonal genomic and epigenomic datasets and CRE activity (Supplemental Fig. 7), we asked how accurately these chromatin features could classify CREs with high vs. low activity. Individually, these features perform comparably to the sequence-based models derived from CRX ChIP-seq data (AUC-ROC up to 0.75 on *pRho* and AUC-ROC up to 0.71 on *pCrx*). In addition, we found that combining chromatin features classifies CREs more accurately than using any single feature alone (AUC-ROC=0.79 on *pRho* and AUC-ROC=0.78 on *pCrx*) (Fig. 2F). Similarly, we trained separate gkm-SVM models on each genomic and epigenomic dataset and found that a model combining these scores outperforms any individual gkm-SVM model (AUC-ROC=0.75 on *pRho* and AUC-ROC=0.80 on *pCrx*) (Fig. 2F). These results suggest that gkm-SVM models trained on multiple functional genomic datasets from a particular cell type capture non-redundant aspects of the underlying *cis*-regulatory grammar—and that these models can be combined to more accurately predict *cis*-regulatory activity from primary sequence.

Choice of CRE-seq promoter influences estimates of CRE activity

As described above, we assayed our CRE-seq library on both *pRho* and *pCrx* to assess the extent to which CRE activity is promoter-dependent. We observed similar distributions of activity on both promoters (Fig. 2B) and similar correlations with specific sequence features and orthogonal genomic and epigenomic datasets (Fig. 2C-E). Nevertheless, we identified several exceptions, including K50 homeodomain binding sites (i.e., CRX binding sites) and AA, AC, AT, CA, and TA dinucleotides (Fig. 2D). Consistent with previous work, we found that CRE activity is negatively correlated with the number of K50 binding sites when assayed on *pRho* (Pearson correlation coefficient -0.09) (White et al. 2013; White et al. 2016), but positively correlated with the number of K50 binding sites when assayed on *pCrx* (Pearson correlation coefficient 0.06). This inversion may reflect complex promoter-enhancer interactions—i.e., promoter-dependent CRE activity. Alternatively, given that *pRho* has significantly higher baseline activity than *pCrx*, the dynamic range of the CRE-seq assay may differ on *pRho* vs. *pCrx*.

To examine this in greater detail, we compared the activity of each CRE on each promoter, which suggests that CRE activity indeed saturates on *pRho* (Supplemental Fig. 8). Thus, the observed differences between *pRho* and *pCrx* may be, at least in part, technical in nature. Accordingly, we restricted subsequent analyses to data generated on *pCrx*, which appears to have a broader dynamic range.

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

As discussed above, homeodomain TFs bind DNA as both monomers and dimers, and CRX ChIP-seq peaks are enriched for both monomeric and dimeric CRX binding sites (Fig. 1C). Both monomeric and dimeric CRX binding sites within CRX ChIP-seq peaks are evolutionary conserved, including both half-sites within dimeric CRX binding sites (Fig. 3A). In addition, the most highly conserved position within the TAAT core (in both monomeric and dimeric contexts) is the second adenine (TAAT). We previously showed that substitutions at this position effectively eliminate CRX binding *in vitro* (Lee et al. 2010), consistent with the key role of this position in mediating homeodomain-DNA interactions (Chaney et al. 2005).

To quantify the regulatory activity of individual TF binding sites, we used CRE-seq to assay the effect of inactivating mutations (TAAT to TACT) in 1,756 CRX binding sites within the CRX-bound regions described above (Fig. 3B). We found that, in general, CRX binding sites act as enhancers—74% of mutations decrease activity, and 25% decrease activity by more than two-fold (compared to 4% that increase activity by more than two-fold) (Fig. 3C). Nearly half (49%) of these changes are statistically significant (FDR<0.05), and 85% of statistically significant differences are decreases in activity (Fig. 3D).

We next asked if the predicted affinity of CRX binding sites is correlated with their activity. We calculated the motif score for each targeted CRX site, and we binned sites based on the match p-value into four 'affinity' classes (very low, low, medium, and high). To account for the fact that CRX binding sites can act as either enhancers or repressors, we considered the absolute log fold change in activity. We found that monomeric CRX binding site motif scores are not significantly correlated with activity. In contrast, dimeric CRX binding site scores are significantly (though modestly) correlated with activity (Pearson correlation coefficient 0.23, $p<2.2x10^{-16}$). Similarly, the activity of targeted sites does not vary significantly across affinity classes defined by a monomeric homeodomain PWM, whereas all pairwise comparisons (except medium vs. high) are significantly different when comparing affinity classes defined by a dimeric homeodomain PWM (FDR<0.05) (Fig. 3E, Supplemental Table 5). Taken together, these data show that medium- and high-affinity dimeric CRX binding sites encode stronger enhancer (or repressor) activity than monomeric CRX binding sites.

Pairs of CRX binding sites act cooperatively

To characterize interactions between CRX binding sites, we selected 225 CREs with two monomeric CRX binding sites (with unconstrained intersite spacing and orientation) and mutated them individually and in combination. For each pair, we defined the binding site with the higher monomeric homeodomain PWM score to be “site 1.” Consistent with the above results, we found that mutating either binding site has a similar effect on wild-type activity, independent of their relative affinities (Fig. 3F, Supplemental Fig. 9). Furthermore, mutating both CRX binding sites has only a slightly greater effect than mutating either site individually (Fig. 3F, Supplemental Fig. 9). This result suggests that at least some pairs of CRX binding sites act synergistically—i.e., they increase or decrease CRE activity more in combination than we would predict based on the activity of either site alone. To test this, we modeled the

activity of each CRE as a linear function of the presence of both targeted CRX binding sites and an interaction term. Nearly half (47%) of the 225 CREs we analyzed had significant interaction terms (FDR<0.05) (Supplemental Table 6), suggesting that non-additive interactions between pairs of CRX binding sites are common among photoreceptor CREs.

We used this same approach (mutating sites individually and in combination) to dissect the activity of half-sites in 130 dimeric CRX binding sites. Similar to the above, we defined "half-site 1" and "half-site 2" based on the orientation of the highest scoring match to a dimeric homeodomain PWM. We again found that mutating either half-site alone significantly decreases activity, but that mutating them together has minimal additional effect, suggesting that at least some half-sites act cooperatively (Fig. 3F, Supplemental Fig. 9). Similar to our analysis of monomeric sites, we used linear models to test for interactions between half-sites, which revealed statistically significant interactions in 62% of cases (FDR<0.05) (Supplemental Table 7), indicating that half-sites within dimeric CRX binding sites often act cooperatively.

The frequency of interactions between CRX binding sites in determining regulatory activity is notable considering that additive models of TF binding site content (i.e., models without interactions) predict CRX occupancy with reasonable accuracy (Fig. 1D). Indeed, the role of TF cooperativity may represent a significant difference between the sequence grammar that determines CRX occupancy and that which determines the activity of CRX-bound regions. Accordingly, we set out to validate these results using an orthogonal assay. We selected two CREs from our analysis (one with two monomeric CRX binding sites, and one with a dimeric CRX binding site) and tested the effects of the same mutations described above via conventional fluorescent reporter assays (Supplemental Fig. 10, Supplemental Table 8) (Corbo et al. 2010). We found that, indeed, mutating either monomeric CRX binding site (or both) had roughly equivalent effects on CRE activity, i.e., significant down-regulation of fluorescent reporter expression. Similarly, mutating either half-site within the dimeric CRX binding site (or mutating both) essentially eliminated reporter expression. Thus, the results from traditional fluorescent reporter assays are consistent with the results from CRE-seq.

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

In our analysis of CRX binding site mutations described above, we were surprised that the effects of inactivating mutations are not strongly correlated with TF binding site affinity (Fig. 3E). However, this analysis only considered a single mutation (TAAT to TACT). To characterize the relationship between CRX binding site affinity and activity in greater detail, we used CRE-seq to quantify the effect of all possible single nucleotide substitutions in a 13-bp window overlapping 97 monomeric and 98 dimeric CRX binding sites (Fig. 4A). We again found that mutations in dimeric CRX binding sites are enriched for strong effects compared to mutations in monomeric CRX binding sites, consistent with our interpretation that dimeric binding CRX sites encode stronger enhancers (Fig. 4B). Furthermore, these data define the key positions within CRX binding sites, with mutations in the TAAT core and the first 3' nucleotide having the strongest effects (including both half-sites within dimeric CRX binding sites) (Fig. 4B-D; Supplemental Fig. 11-12). In general, these results are consistent with the *in vitro* DNA binding preferences of CRX determined by quantitative gel shift (Fig. 4C) (Lee et al. 2010), as well as the DNA binding preferences of closely related homeodomain TFs that have been estimated by high-throughput SELEX (Supplemental Fig. 13) (Jolma et al. 2013). In addition, we found that the median effects of specific substitutions at each position within CRX binding sites are highly correlated with the associated changes in PWM score (Pearson correlation coefficient 0.82 for both monomeric and dimeric CRX binding sites) (Supplemental Fig. 13).

While the median effects of specific substitutions are strongly correlated with their predicted impact on CRX binding, we found substantial variation in the effects of specific mutations across CREs (Fig. 4D). Accordingly, when we consider the relationship between changes in TF binding site affinity and activity across all mutations (without aggregating by position or nucleotide identity), this correlation is significantly lower (Pearson correlation coefficient 0.32 for monomeric sites and 0.35 for dimeric sites) (Supplemental Fig. 13). Nevertheless, within individual CREs, these correlations are often stronger (median absolute Pearson correlation coefficient 0.54 for monomeric CRX binding sites, and median absolute Pearson correlation coefficient 0.58 for dimeric CRX binding sites). This apparent contradiction is explained by examining the relationship between TF binding site affinity and activity within individual CREs (Supplemental Fig. 14). For a given CRE, changes in activity are typically well-described by a linear function of changes in affinity, but the slope is CRE-dependent. In other words, sequence context appears to re-scale the relationship between changes in CRX binding site affinity and activity across CREs (Supplemental Fig. 14). These results suggest that, while models of TF binding site affinity may accurately predict the relative effects of mutations within a single CRE, predicting the relative effects of mutations across multiple CREs presents an additional challenge.

In addition to examining the relationship between TF binding site affinity and activity, we asked if phylogenetic conservation is correlated with the effects of mutations in CRX binding sites. We found that the average conservation (phyloP scores) of specific positions within CRX binding sites are correlated with the median effects of mutations at those positions (Fig. 4D, see also Supplemental Fig. 15) (Pearson correlation coefficient 0.89-0.95). Nevertheless, conservation is poorly correlated with mutation effects across CREs (Pearson correlation coefficient 0.07-0.10), suggesting that conservation scores should be used cautiously when prioritizing candidate regulatory variants from multiple loci.

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

In addition to analyzing single-nucleotide substitutions, we quantified the effect of small insertions and deletions in dimeric CRX binding sites on CRE activity to determine if their activity depends on half-site spacing (Fig. 5). First, we made all one-, two-, and three-base-pair deletions (n=7 per targeted site) of the three nucleotides 3' of the TAAT core in the 97 monomeric and 98 dimeric CRX binding sites described above. On average, deletions significantly decrease activity ($p<1.4x10^{-4}$) (Fig. 5A), but deletions of different sizes do not have significantly different effects. Interpreting these results with respect to spacing is challenging since deletions impact the affinity of individual half-sites as well as potential TF interactions, though we note that deletions in dimeric CRX binding sites have stronger effects than analogous deletions in monomeric CRX binding sites ($p<8.5x10^{-4}$).

To test the effect of alterations in half-site spacing while minimizing the impact on the affinity of individual half-sites, we made the following insertions (Fig. 5B). For one-base-pair insertions, we doubled the center nucleotide (e.g., 5'- CAG-3' to 5'-CAAG-3'). For two-base-pair insertions, we tripled the center nucleotide (e.g., 5'-CAG-3' to 5'- CAAAG-3'). And for three base-pair-insertions, we doubled the entire triplet (e.g., 5'-CAG-3' to 5'-CAGCAG-3'). All three mutations significantly decrease the activity of dimeric, but not monomeric, CRX binding sites ($p<4.9x10^{-7}$). Taken together, these data indicate that the activity of dimeric CRX binding sites depends on a strict three-nucleotide spacing, consistent with structural studies of paired-class homeodomain TF-DNA complexes (Wilson et al. 1995; Tucker and Wisdom 1999).

Finally, our analysis of single-nucleotide substitutions suggests that the activity of dimeric CRX binding sites is optimized by the spacer triplet CCG, which creates the highest-affinity K50 homeodomain binding sites on each strand given the spacing constraints described above. However, additional spacer

sequences are also enriched in CRX ChIP-seq peaks. To determine the activity of these different spacer sequences, we substituted the six most enriched triplets (CCG, CAG, AGG, AAG, CTC, and CCA) into each of the CRX binding sites tested above in both orientations. We found that the effect of individual substitutions is similar regardless of their orientation (Supplemental Fig. 15). In addition, we found that specific spacer sequences have distinct effects on CRE activity (Fig. 5C). These effects are positively correlated with the resulting affinity of each half-site for K50 homeodomain TFs and negatively correlated with the resulting affinity for Q50 homeodomain TFs. In particular, the spacer CCA has the lowest activity, and this sequence forms a high-affinity K50 binding site on the forward strand, and a high-affinity Q50 binding site on the reverse strand. Previously, K50 and Q50 paired-class TFs have been shown to antagonize one another at specific dimeric homeodomain binding sites (Tucker and Wisdom 1999), suggesting that the reduced activity of dimeric CRX binding sites with CCA spacer sequences may reflect TF competition.

Accounting for baseline CRE activity improves the prediction of variant effects

We next asked how accurately primary sequence features could predict the effects of mutations in CRX binding sites. We used linear regression to model the effects of changes in CRE activity as a function of changes in TF binding site affinity for each of the CREs in our dense substitution analysis. We tested representing changes in TF binding site affinity as both the difference between wild-type and mutant PWM scores as well as the associated change in CRX ChIP-seq gkm-SVM scores (i.e., deltaSVM scores) (Supplemental Tables 9-11) (Ghandi et al. 2014; Lee et al. 2015). Fitting models for each CRE individually, we found that deltaSVM scores predict the effects of mutations in CRX binding sites more accurately than changes in PWM scores (R^2=0.41 vs. R^2=0.32) (Fig. 6A). However, both approaches perform significantly worse when we fit models for all mutations simultaneously (R^2=0.14 for deltaSVM scores and R^2=0.12 for changes in PWM scores) (Fig. 6B). We hypothesized that the performance of these models might be limited by omitting the contribution of sequence features outside the targeted binding sites. In addition, we asked whether models derived from orthogonal genomic and epigenomic assays could be combined to yield a more complete representation of sequence features relevant to photoreceptor CRE activity. Accordingly, we trained gkm-SVM models on 15 additional datasets (DNase-seq, ATAC-seq, TF ChIP-seq, and/or histone ChIP-seq from retina and six non-retinal tissues), and quantified how well combinations of gkm-SVM (wild-type) and/or deltaSVM scores (mutant-specific) predict the effect of mutations in CRX binding sites (The ENCODE Project Consortium 2012; Hao et al. 2012; Wilken et al. 2015; Mo et al. 2016; Hughes et al. 2017). We found that none of these multi-score models substantially improves upon CRX ChIP-seq deltaSVM scores (R^2=0.14-0.16) (Fig. 6C). These results suggest that, while additive models of primary sequence features accurately predict TF occupancy, more complex models (e.g., ones incorporating spacing- and orientation-dependent interactions between TF binding sites) may be necessary to accurately predict the effects of specific mutations on regulatory activity.

As an alternative to predicting the effects of mutations (i.e., change in CRE activity) from primary sequence alone, we asked how accurately the combination of wild-type (baseline) CRE activity and sequence-based models could predict the activity of mutant CREs. We found that wild-type activity alone explains 66% of the variation in mutant activity (Fig. 6D). Furthermore, combining wild-type expression with deltaSVM scores from multiple datasets significantly improves performance (R^2=0.73), and incorporating interactions between wild-type activity and deltaSVM scores yields additional improvement (R^2=0.76) (Fig. 6D). These data demonstrate that the effects of mutations in CRX binding sites depend on wild-type CRE activity. Accordingly, while modeling the effects of CRX binding site mutations from primary sequence alone remains challenging, knowledge of baseline CRE activity significantly improves the prediction of the functional effects of *cis*-regulatory variants.

DISCUSSION

In this study, we assayed thousands of wild-type and mutant CREs to identify sequence features that modulate the activity of CRX binding sites, and we found that the activity of CRX binding sites depends on multiple layers of sequence context. Both the affinity of individual CRX binding sites as well as their configuration (i.e., monomeric vs. dimeric) have modest but significant effects on their activity. Moreover, the broader sequence context in which CRX binding sites occur, including the number and affinity of additional K50 as well as non-K50 binding sites, has even stronger effects on their activity. We also found that multiple instances of CRX binding sites within individual CREs often act cooperatively. Of note, with the exception of half-sites within dimeric CRX binding sites, we did not find evidence that interactions between TF binding sites are constrained with respect to spacing or orientation. This may be because we only measured the effects of pairs of mutations in a few hundred CREs and lacked the power to identify these constraints. Alternatively, the interactions we observed may reflect indirect cooperativity (e.g., nucleosome-mediated cooperativity) with limited constraints on relative spacing and orientation. Regardless, while CRX occupancy can largely be explained by additive models of homeodomain TF binding sites and dinucleotide content, the activity of CRX-binding sites appears to be determined by a richer vocabulary of sequence features as well as interactions between them. These results are broadly consistent with a recent study of PPARG-bound regions in mouse adipocytes (Grossman et al. 2017), suggesting that they may highlight general mechanisms by which mammalian enhancers encode *cis*-regulatory activity.

As described above, our analysis of the correlation between TF binding sites and native CRE activity confirms a role for several families of TFs known to regulate photoreceptor development (Fig. 2D). One unexpected result was that T-box motifs are negatively correlated with photoreceptor CRE activity (Fig. 2D). Although T-box TFs have no established role in mammalian photoreceptor development, *Tbx2b* mutant zebrafish show a conversion of ultraviolet cones into rods (Alvarez-Delfin et al. 2009). In addition, recent expression profiling in chicken has suggested that *Tbx2* plays a role in violet cone development (Enright et al. 2015). In mouse, *Tbx2* is the most highly expressed T-box TF in developing photoreceptors, and it is down-regulated as they mature (Supplemental Fig. 6) (Sowden et al. 2001; Kim et al. 2016a; Kim et al. 2016b). In addition, TBX2 has been shown to act as a repressor (Abrahams et al. 2010). Together with CRE-seq data, these observations suggest that TBX2 may act as a transcriptional repressor in the developing mouse retina, which would establish TBX2 as an ancient and highly conserved regulator of vertebrate photoreceptor identity.

In addition to T-box motifs, we found that Q50 binding sites are negatively correlated with CRE activity (Fig. 2D), suggesting they may also act as repressors. Among Q50 TFs in mouse, *Rax* is the most highly expressed in photoreceptors (Kim et al. 2016a; Kim et al. 2016b) and has been shown to play a role in photoreceptor maturation and survival (Irie et al. 2015). However, traditional reporter assays suggest that RAX is a weak activator of photoreceptor CREs, not a repressor (Irie et al. 2015). Accordingly, additional functional studies are needed to determine if Q50 binding sites generally act as activators or repressors in photoreceptors. Interestingly, other retinal cell types express distinct Q50 TFs (e.g., VSX2 in bipolar cells), raising the possibility that Q50 binding sites could mediate repression of photoreceptor genes in other retinal cell types.

Of note, *Nrl* is a member of the Maf subfamily of basic leucine zipper TFs and a master regulator of rod photoreceptor identify, but we did not detect a significant correlation between the number of NRL binding sites and CRE activity. Importantly, NRL binding sites are not strongly enriched in CRX ChIP-seq peaks (Supplemental Fig. 2). Therefore, it is possible that the elements we assayed by CRE-seq simply lack a sufficient number of NRL-bound sequences to detect the effects of NRL on CRE activity. However, based on NRL ChIP-seq, we estimate that 25% of CRX-bound regions are co-bound by NRL (Hao et al. 2012). Furthermore, we found that CRE-seq activity was significantly correlated with NRL

occupancy as estimated by ChIP-seq. These results suggest that the interactions between NRL and DNA may not be adequately modeled by the conventional NRL PWM and that more complex models (perhaps involving co-binding TFs) may be needed to precisely define the role of NRL in rod-specific gene regulation.

Currently, there is intense interest in developing quantitative models to predict the effects of non-coding variants on CRE activity to identify causal variants underlying disease association signals (Lee et al. 2015; Zhou and Troyanskaya 2015; Kelley et al. 2016). Several groups have recently evaluated the accuracy of specific methods for predicting the activity of native CREs as measured by MPRA, finding that current models explain 4-38% of the variation in regulatory activity (Lee et al. 2015; Grossman et al. 2017; Inoue et al. 2017). Here, we report that linear models using multiple deltaSVM scores explain up to 16% of the variation in the effects of mutations in CRX binding sites (i.e., changes in activity). Thus, additional work is needed to predict regulatory activity from primary sequence, and we hypothesize that models incorporating higher order interactions are likely to prove valuable. Nevertheless, we show that combining estimates of wild-type CRE activity with deltaSVM scores explains most of the variation in mutant CRE activity (R^2=0.76). This result suggests that combining reference sets of wild-type cell-type-specific CRE activity (ascertained by MPRA) with cell-type-specific models of sequence grammar (e.g., deltaSVM models) could be a powerful strategy for predicting the cell-type-specific effects of novel *cis*-regulatory variants.

METHODS

CRE-seq library design

1,270 target regions were selected from CRX ChIP-seq peaks (100-bp elements centered on peak summits). Candidate monomeric CRX binding sites were identified with FIMO (v4.11.2) (Grant et al. 2011) using the OTX2_DBD_1 PWM (Jolma et al. 2013) and a p-value threshold of $p<10^{-3}$. Candidate dimeric CRX binding sites were identified by matches to the pattern 5'-TAAKNNNMTTN-3' or 5'-NAAKNNNMTTA-3'. For monomeric CRX binding sites, we mutated each TAAT core to TACT. For dimeric CRX binding sites, we mutated each TAAT half-site to TACT individually as well as both in combination. In addition, within each CRE, we mutated each CRX binding site individually as well as all in combination. For saturating mutagenesis, we selected 100 monomeric and 100 dimeric CRX binding sites and made all possible single-nucleotide substitutions in a 13-bp window overlapping each CRX binding site as well as selected insertions, deletions, and substitutions of the spacer nucleotides. Each of the 14,987 wild-type and mutant target sequences were paired with six or seven unique 13-bp barcodes yielding a final library of 100,000 oligos.

CRE-seq library construction

170-bp oligos were generated by array-based oligonucleotide synthesis through a limited licensing agreement with Agilent Technologies. Oligos were amplified and cloned into (Rho-prox)-DsRed (Montana et al. 2011a) as described previously (Kwasnieski et al. 2012; White et al. 2013). The resulting plasmid library was sequenced to assess CRE representation (99.7% of 100,000 targeted oligos were detected, 98.8% at >1 barcode per million barcodes). A promoter-DsRed reporter construct (either pRho-DsRed or pCrx-DsRed) was then cloned between each CRE and barcode to generate the final CRE-seq library. See Supplemental Materials for details.

CRE-seq assay

Mouse husbandry and all procedures were conducted in accordance with the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health, and were approved by the Washington University in St. Louis Institutional Animal Care and Use Committee. Retinal electroporation and CRE-seq were performed as described previously (Montana et al. 2011b; Kwasnieski et al. 2012; White et al. 2013; Shen et al. 2016). See Supplemental Materials for details.

Models of TF occupancy and CRE activity

Logistic regression models predicting TF occupancy or CRE activity and linear regression models predicting mutation effects were implemented in R (v3.3) (R Core Team 2016). For each target sequence, non-redundant dinucleotide frequencies were calculated on both strands, and instances of TF binding sites were identified with FIMO (v4.11.2) (Grant et al. 2011) using a database of 206 human and mouse PWMs derived from high-throughput SELEX and ChIP-seq experiments (Jolma et al. 2013). gkm-SVM models predicting TF occupancy, chromatin accessibility, and histone modifications were trained with LS-GKM (l=11, k= 7) (Lee 2016) using background sequence sets generated with the gkm-SVM R package (Ghandi et al. 2014; Ghandi et al. 2016). See Supplemental Materials for details.

DATA ACCESS

The sequence data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www. ncbi.nlm.nih.gov/geo/) under accession number GSE106243.

ACKNOWLEDGEMENTS

The authors would like to thank Daniel Murphy and Susan Shen for providing comments on the manuscript, the Genome Technology Access Center at Washington University in St. Louis for sequencing services, and Jessica Hoisington-Lopez from the DNA Sequencing Innovation Lab at The Edison Family Center for Genome Sciences and Systems Biology for her sequencing expertise. This work was supported by the National Institutes of Health (EY025196, EY026672, and EY024958).

DISCLOSURE DECLARATION

The authors have no competing financial interests to declare.

REFERENCES

Abrahams A, Parker MI, Prince S. 2010. The T-box transcription factor Tbx2: its role in development and possible implication in cancer. *IUBMB Life* **62**: 92-102.

Alvarez-Delfin K, Morris AC, Snelson CD, Gamse JT, Gupta T, Marlow FL, Mullins MC, Burgess HA, Granato M, Fadool JM. 2009. Tbx2b is required for ultraviolet photoreceptor cell specification during zebrafish retinal development. *Proc Natl Acad Sci U S A* **106**: 2023-2028.

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**: 1074-1077.

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X et al. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science* **324**: 1720-1723.

Barrera LA, Vedenko A, Kurland JV, Rogers JM, Gisselbrecht SS, Rossin EJ, Woodard J, Mariani L, Kock KH, Inukai S et al. 2016. Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science* **351**: 1450-1454.

Chaney BA, Clark-Baldwin K, Dave V, Ma J, Rance M. 2005. Solution structure of the K50 class homeodomain PITX2 bound to DNA and implications for mutations that cause Rieger syndrome. *Biochemistry* **44**: 7497-7511.

Chatelain G, Fossat N, Brun G, Lamonerie T. 2006. Molecular dissection reveals decreased activity and not dominant negative effect in human OTX2 mutants. *J Mol Med (Berl)* **84**: 604-615.

Chen S, Wang QL, Nie Z, Sun H, Lennon G, Copeland NG, Gilbert DJ, Jenkins NA, Zack DJ. 1997. Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. *Neuron* **19**: 1017-1030.

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.

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

Enright JM, Lawrence KA, Hadzic T, Corbo JC. 2015. Transcriptome profiling of developing photoreceptor subtypes reveals candidate genes involved in avian photoreceptor diversification. *J Comp Neurol* **523**: 649-668.

Fletez-Brant C, Lee D, McCallion AS, Beer MA. 2013. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res* **41**: W544-556.

Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* **10**: e1003711.

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.

Grossman SR, Zhang X, Wang L, Engreitz J, Melnikov A, Rogov P, Tewhey R, Isakova A, Deplancke B, Bernstein BE et al. 2017. Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc Natl Acad Sci U S A* **114**: E1291-E1300.

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.

Hsiau TH, Diaconu C, Myers CA, Lee J, Cepko CL, Corbo JC. 2007. The cis-regulatory logic of the mammalian photoreceptor transcriptional network. *PLoS One* **2**: e643.

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.

Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, Shendure J. 2017. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res* **27**: 38-52.

Irie S, Sanuki R, Muranishi Y, Kato K, Chaya T, Furukawa T. 2015. Rax Homeoprotein Regulates Photoreceptor Cell Maturation and Survival in Association with Crx in the Postnatal Mouse Retina. *Mol Cell Biol* **35**: 2583-2596.

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.

Kelley DR, Snoek J, Rinn JL. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* **26**: 990-999.

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.

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

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* **47**: 955-961.

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**: 1390-1399.

Livesey FJ, Furukawa T, Steffen MA, Church GM, Cepko CL. 2000. Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene Crx. *Curr Biol* **10**: 301-310.

Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**: 251-260.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Jr., Kinney JB et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**: 271-277.

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.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**: 265-270.

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

Setty M, Leslie CS. 2015. SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLoS Comput Biol* **11**: e1004271.

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.

Sowden JC, Holt JK, Meins M, Smith HK, Bhattacharya SS. 2001. Expression of Drosophila omb-related T-box genes in the developing human and mouse neural retina. *Invest Ophthalmol Vis Sci* **42**: 3095-3102.

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

Tucker SC, Wisdom R. 1999. Site-specific heterodimerization by paired class homeodomain proteins mediates selective transcriptional responses. *J Biol Chem* **274**: 32325-32332.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K et al. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**: 1431-1443.

White MA, Kwasnieski JC, Myers CA, Shen SQ, Corbo JC, Cohen BA. 2016. A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors. *Cell Rep* **17**: 1247-1254.

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.

Wilken MS, Brzezinski JA, La Torre A, Siebenthall K, Thurman R, Sabo P, Sandstrom RS, Vierstra J, Canfield TK, Hansen RS et al. 2015. DNase I hypersensitivity analysis of the mouse brain and retina identifies region-specific regulatory elements. *Epigenetics Chromatin* **8**: 8.

Wilson D, Sheng G, Lecuit T, Dostatni N, Desplan C. 1993. Cooperative dimerization of paired class homeo domains on DNA. *Genes Dev* **7**: 2120-2134.

Wilson DS, Guenther B, Desplan C, Kuriyan J. 1995. High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. *Cell* **82**: 709-719.

Zack DJ, Bennett J, Wang Y, Davenport C, Klaunberg B, Gearhart J, Nathans J. 1991. Unusual topography of bovine rhodopsin promoter-lacZ fusion gene expression in transgenic mouse retinas. *Neuron* **6**: 187-199.

Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**: 931-934.

FIGURE LEGENDS

Fig. 1. Primary sequence features predict CRX occupancy *in vivo*. A) Schematic of analytical approach. 5,250 CRX-bound regions and 52,500 CRX-unbound regions were selected based on CRX ChIP-seq data (200-bp elements centered on peak summits). Feature vectors composed of average dinucleotide frequencies and/or counts of specific TF binding sites (up to 206) were defined for each sequence. **B)** CRX ChIP-seq peaks are centered on local enrichments of specific dinucleotide classes, including elevated GC and AG dinucleotide content. CRX-unbound regions are also modestly enriched for specific dinucleotide classes, likely due to selecting regions with GC content matching that of CRX-bound regions. **C)** CRX ChIP-seq peaks are centered on local enrichments of specific TF binding sites, including monomeric and dimeric CRX binding sites. **D)** Performance of specific models classifying CRX-bound vs. CRX-unbound sequences visualized with ROC (FPR vs. TPR) and PR (recall vs. precision) curves. TPR: true positive rate. FPR: false positive rate. Dashed lines: performance of random classifiers. LR: logistic regression. LR: best PWM—counts of dimeric CRX binding sites (single PWM) (AUC-ROC=0.77, AUC-PR=0.26). LR: full model—dinucleotide frequencies and counts of 206 TF binding sites (binned by PWM score) (AUC-ROC=0.95, AUC-PR=0.74). See Supplemental Table 2 for feature weights. gkm-SVM: 11-mers with seven informative positions (AUC-ROC=0.99, AUC-PR=0.92). **E)** Performance of LR: full model with features extracted from windows of different sizes (20 bp to 200 bp). Gray box: maximum AUC.

Fig. 2. Primary sequence features are correlated with CRE activity *in vivo*. A) Schematic of experimental approach. 100-bp elements centered on CRX ChIP-seq peaks were cloned upstream of a photoreceptor promoter driving DsRed with CRE-specific barcodes. Constructs were electroporated into P0 mouse retina and cultured for eight days, at which point RNA and DNA were harvested and barcodes were amplified and sequenced to quantify activity. **B)** Distribution of activity of elements assayed on either *pRho* or *pCrx*. Data are median-centered. Dashed lines: three-fold decrease or increase relative to median. The percentage of constructs with activity above or below this threshold is indicated. **C)** Correlation between number of E-Box binding sites and activity on *pRho* and *pCrx*. **D)** Heatmap of Pearson correlation coefficients (PCCs) between specific dinucleotide frequencies or counts of TF binding sites and activity. Included features were significantly correlated with activity on at least one promoter. **E)** Heatmap of Pearson correlations between genomic and epigenomic datasets and CRE activity. **F)** Performance of specific models classifying elements with low (within 1.2-fold of the median) vs. high (>3-fold above the median) activity on *pCrx*. LogR (CRX ChIP): logistic regression classifier using scores from logistic regression model trained on CRX ChIP-seq data (full model in Fig. 1D) (AUC-ROC=0.71). SVM (CRX ChIP): logistic regression classifier using scores from gkm-SVM classifier trained on CRX ChIP-seq data (AUC-ROC=0.74). SVM (combined): logistic regression classifier using scores from gkm-SVM models trained on genomic and epigenomic datasets listed in Supplemental Table 3 (AUC-ROC=0.80). Dashed line: performance of a random classifier.

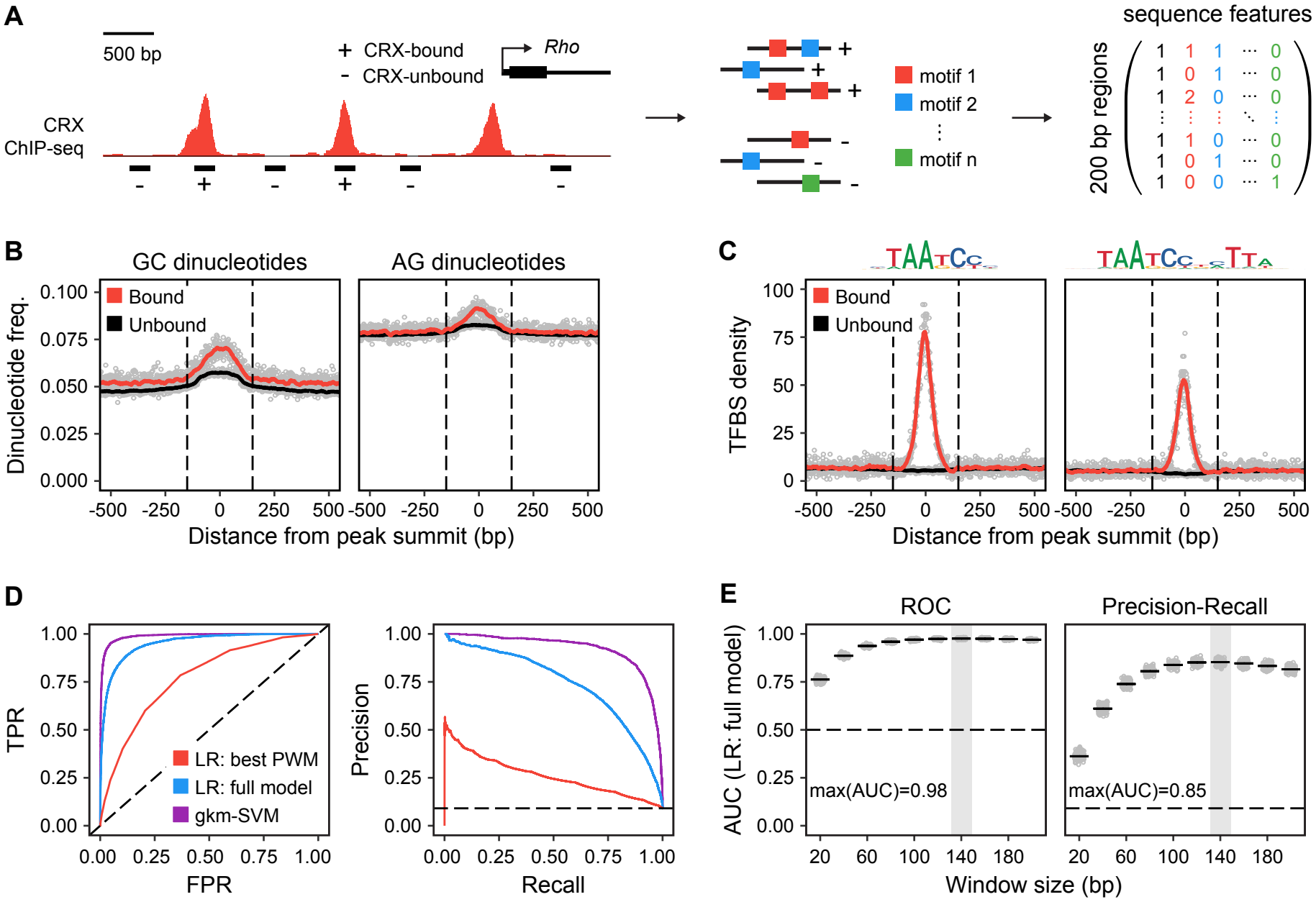
Fig. 3. Dimeric CRX sites have higher activity than monomeric CRX sites. A) Upper panels: heatmaps of nucleotide content in a 30-bp window centered on monomeric or dimeric CRX binding sites. Rows corresponds to distinct TF binding sites, columns correspond to distinct positions, and tiles are colored by nucleotide identity. Lower panels: average conservation (100-way vertebrate phyloP scores) at each position. Positions 0-3 (and 7-10 for dimeric TF binding sites) correspond to TAAT cores (gray boxes). **B)** Schematic of experimental approach. The effects of single-base-pair substitutions (TAAT to TACT) in 1,756 CRX binding sites within CRX ChIP-seq peaks were quantified by CRE-seq. **C)** Distribution of mutation effects ($\log_2$ fold change). **D)** Volcano plot of mutation effects. Among mutations that significantly change activity (FDR<0.05), 85% decrease activity and 15% increase activity. Red: significant decrease in activity. Blue: significant increase in activity. Gray: change in activity not significant. **E)** Absolute effect size vs. monomeric or dimeric PWM score (binned by match p-value). **F)**

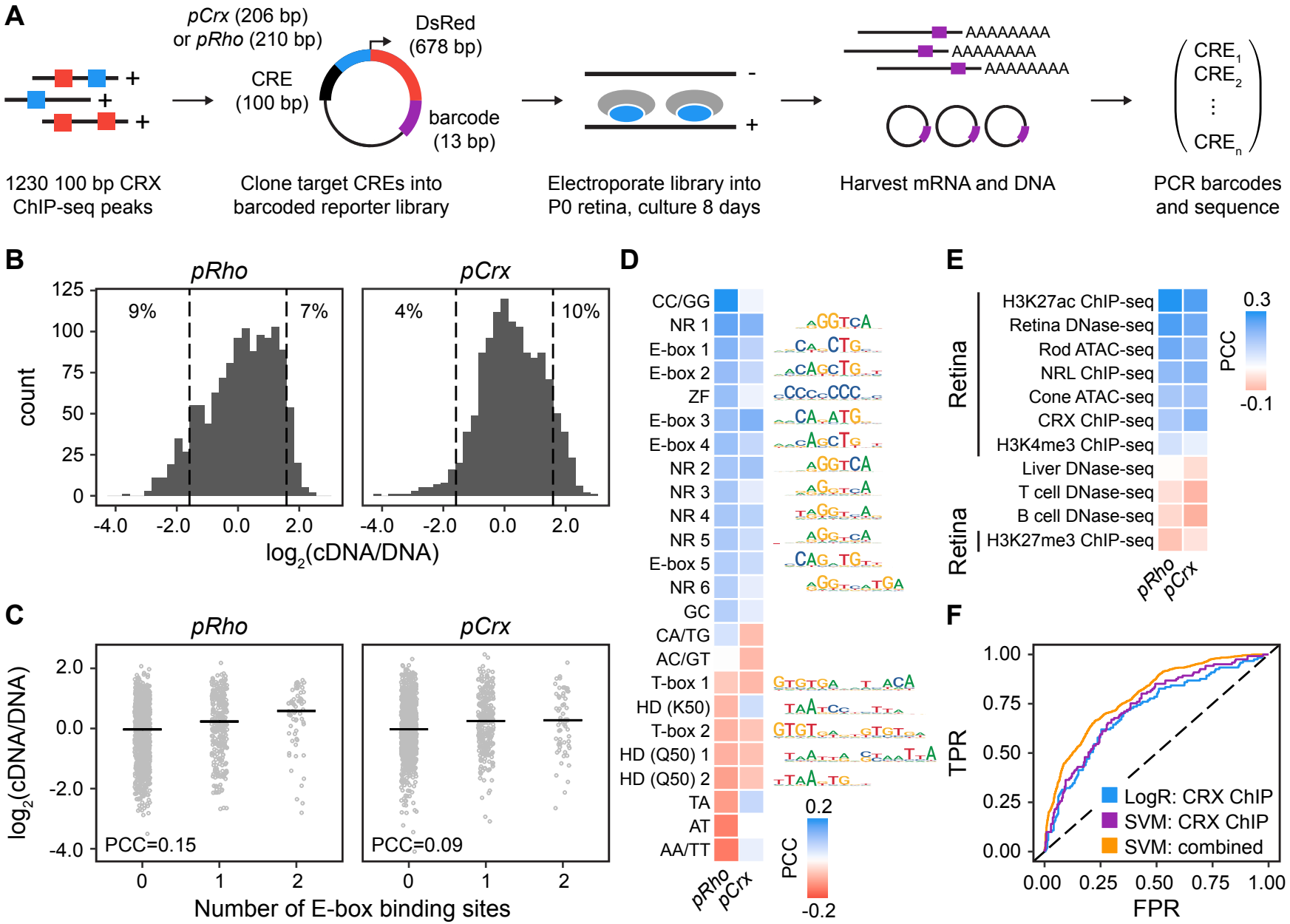
Left panel: activity distributions of CREs with two monomeric CRX binding sites when neither, one, or both are mutated. Right panel: activity distributions of CREs with dimeric CRX binding sites when neither, one, or both half-sites are mutated.

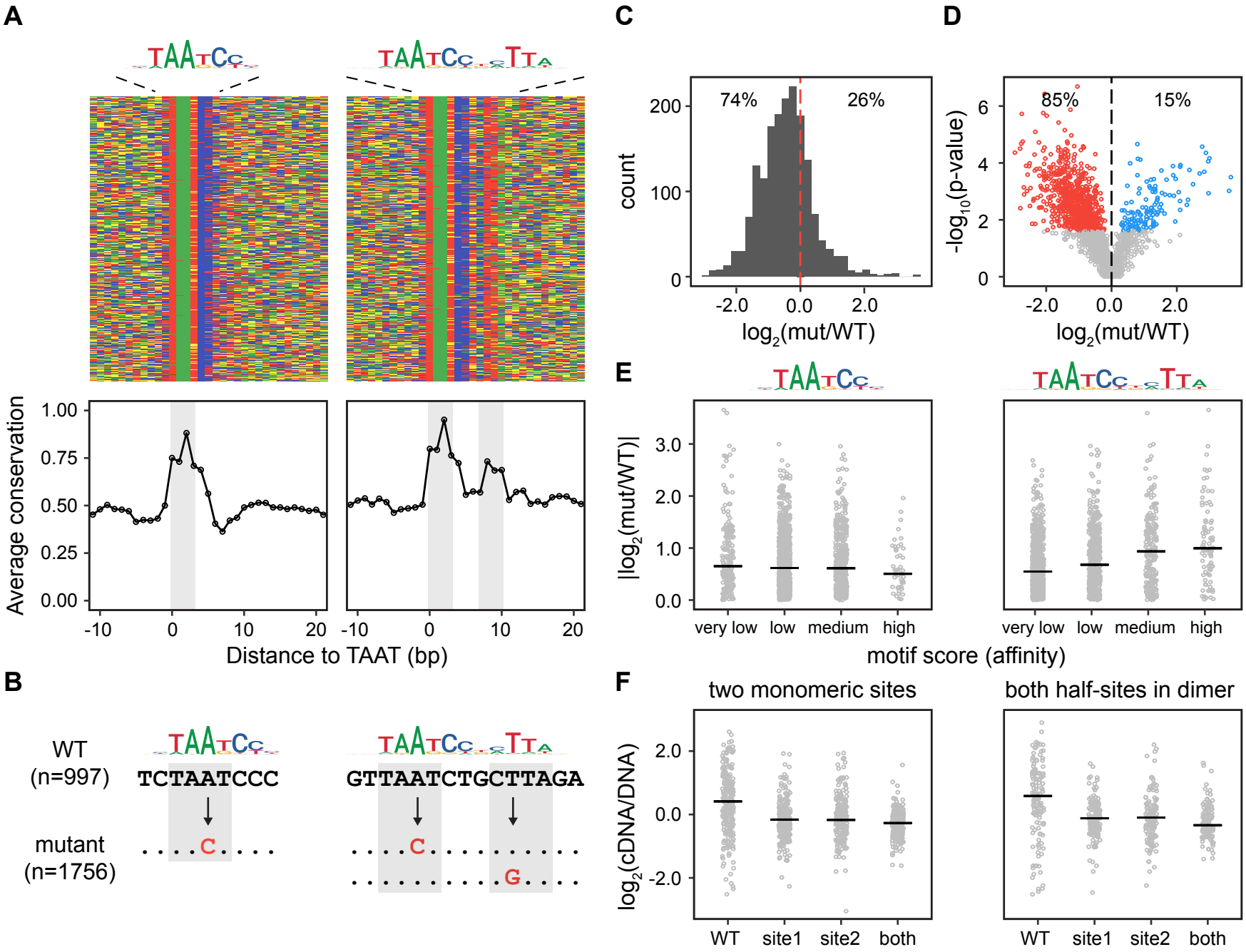
Fig. 4. Dense mutagenesis of monomeric and dimeric CRX binding sites. A) Schematic of experimental approach. All single-nucleotide substitutions in a 13-bp window overlapping 97 monomeric and 98 dimeric CRX binding sites were quantified by CRE-seq (n=39 mutations per TF binding site). **B)** Heatmaps of median effects (across all three substitutions) at each position (columns) in each targeted CRX binding site (rows). Each heatmap represents 97 or 98 distinct elements, and rows are sorted by wild-type activity (high to low). **C)** Heatmaps of median effects (across all target sites) at each position (columns) for specific substitutions (rows). Top panel: change in CRX binding site affinity determined by quantitative gel shift for all possible substitutions in a single target sequence (Lee et al. 2010). Middle panel: change in activity determined by CRE-seq for substitutions in 97 monomeric CRX binding sites. Bottom panel: change in activity determined by CRE-seq for substitutions in 98 dimeric CRX binding sites. **D)** Upper panel: scatter plot of median effects (for all three substitutions) (y-axis) at each position (x-axis) in each targeted CRX binding site. Points represent different targeted CRX binding sites, and horizontal bars represent the median across all targets. Lower panel: average conservation scores (phyloP) at each position (same data as in Fig. 3A).

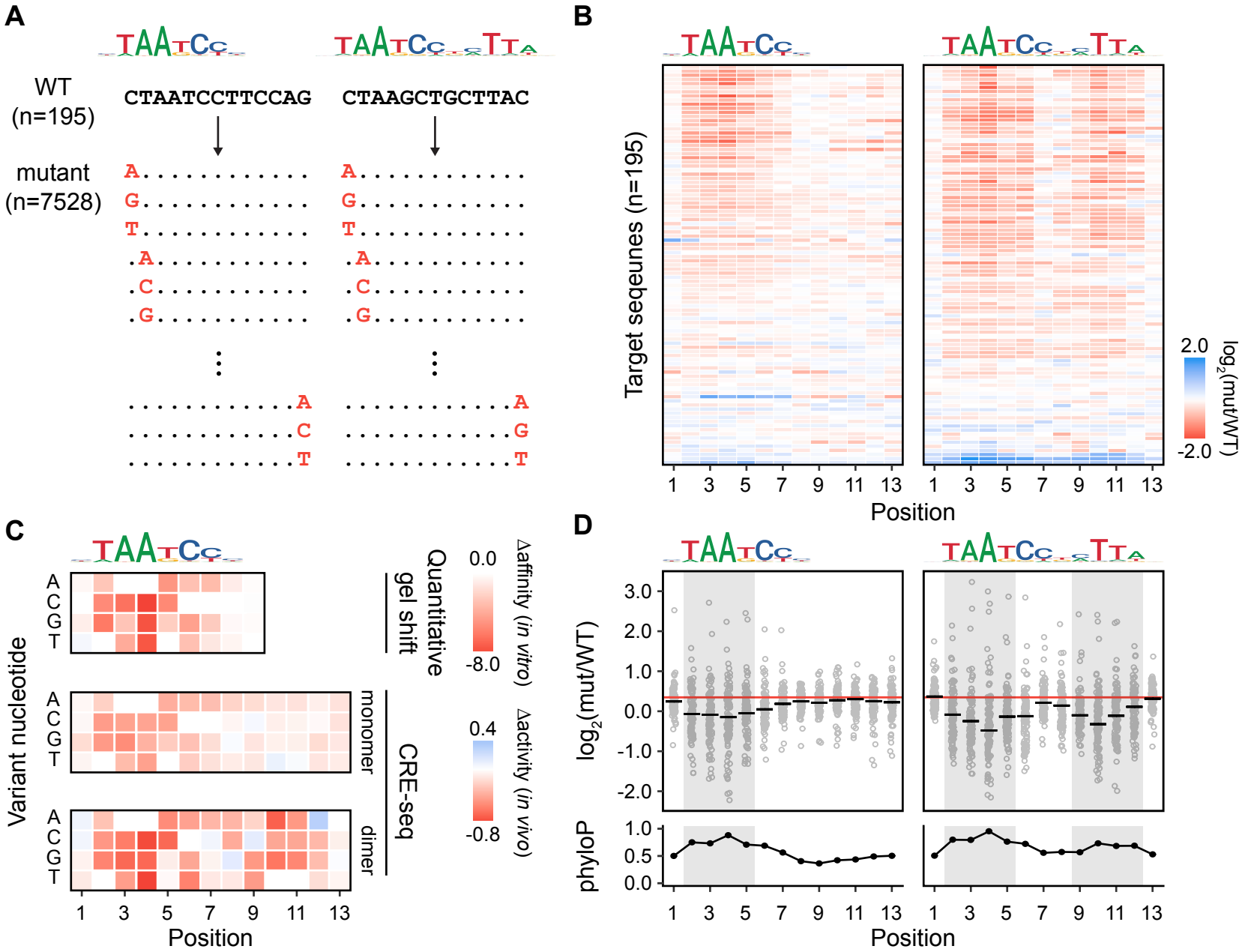
Fig. 5. The activity of dimeric CRX binding sites depends on half-site spacing. A) Left: schematic of experimental approach. The effect of all one-, two-, and three-base-pair spacer deletions in 195 CRX binding sites were quantified by CRE-seq. Right: scatter plot of mutation effects. Points represent individual mutations and horizontal bars represent the median across all targets for deletions of the indicated size. **B)** Left: schematic of experimental approach. The effect of specific one-, two-, and three-base-pair spacer insertions in 195 CRX binding sites were quantified by CRE-seq. Right: scatter plot of mutation effects. Points represent individual mutations and horizontal bars represent the median across all targets for insertions of the indicated size. In A and B, p-values are reported for Mann-Whitney *U* tests comparing the distributions of effects between mutations in monomeric vs. dimeric CRX binding sites. **C)** Left: schematic of experimental approach. The effects of selected three-base-pair spacer substitutions in 98 dimeric CRX binding sites were quantified by CRE-seq. Right: scatter plot of mutation effects. Points represent individual mutations and horizontal bars represent the median across all targets for the indicated substitution. The included heatmap shows counts of the indicated K50 and Q50 motifs among binding sites with each spacer substitution.

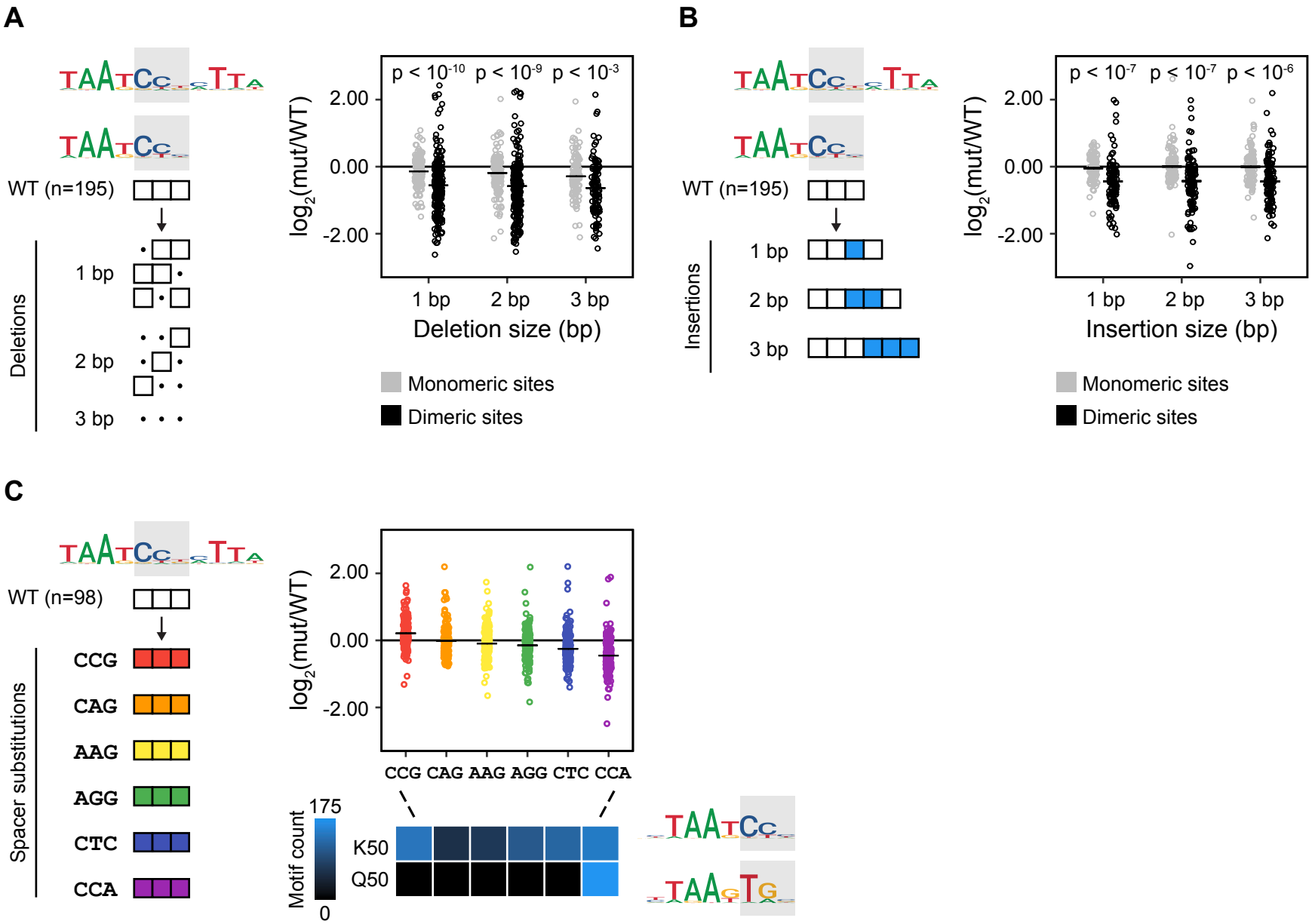
Fig. 6. Accounting for baseline CRE activity improves the prediction of variant effects. A) Performance (R^2) of simple linear regression predicting the effect of individual substitutions from changes in PWM scores or CRX ChIP-seq deltaSVM scores, fitting separate models for each CRE. **B)** Same as in A, except fitting a single model for all CREs. **C)** Performance of multiple linear regression predicting the effect of individual substitutions using deltaSVM scores from multiple datasets (m-deltaSVM), m-deltaSVM and the corresponding gkm-SVM scores from multiple datasets (m-gkm-SVM), or m-deltaSVM scores and m-gkm-SVM scores including all pairwise interactions. **D)** Performance of multiple linear regression predicting mutant expression using wild-type (WT) expression, WT expression and m-deltaSVM scores, or WT expression, m-deltaSVM scores, and interactions between WT expression and deltaSVM scores. In A, individual points represent the performance of models fit for different CREs (n=195). In B-D, individual points represent the performance of models estimated from different folds of repeated ten-fold cross validation (n=100).

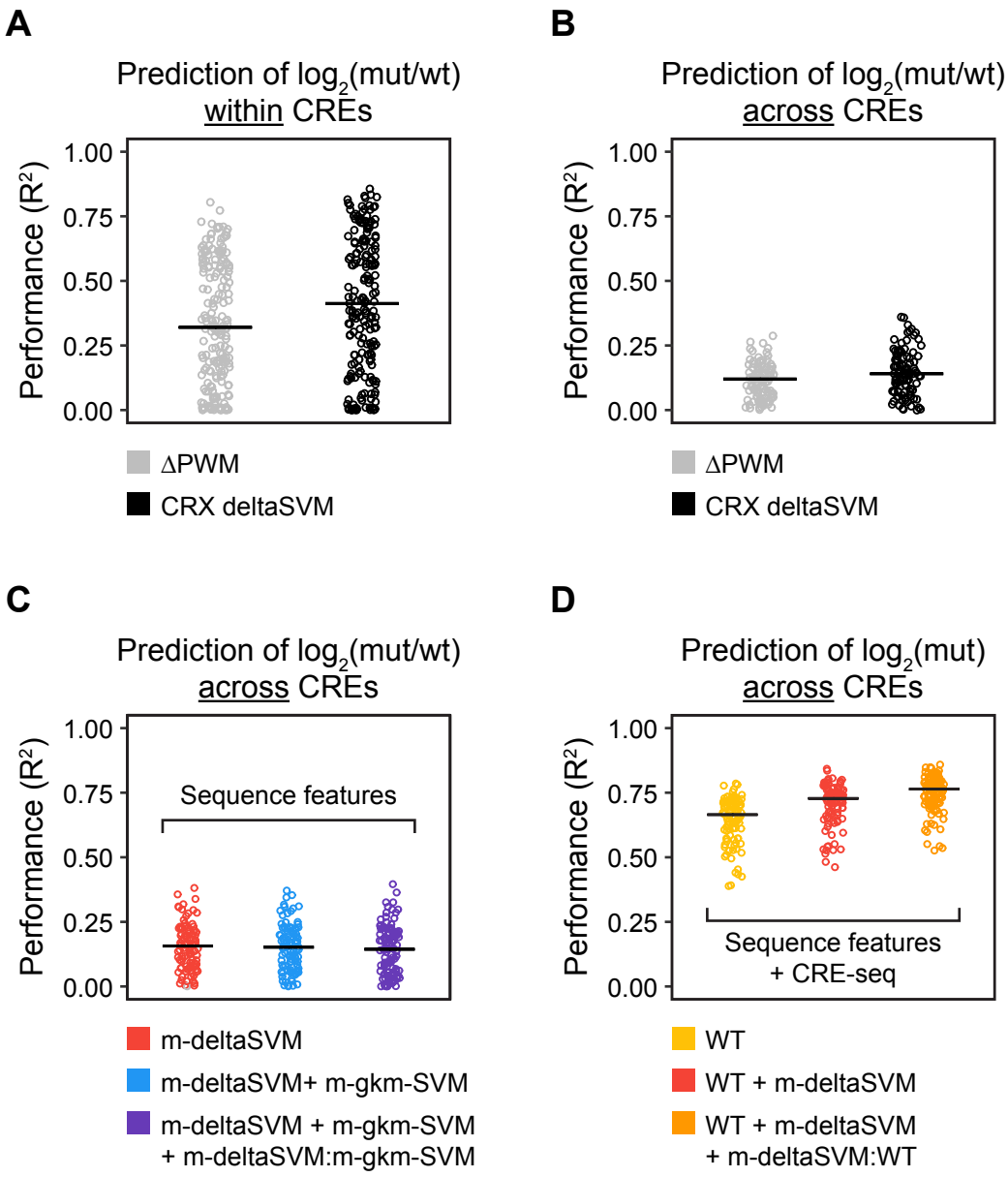

A
500 bp
+ CRX-bound
- CRX-unbound
Rho
CRX
ChIP-seq
motif 1
motif 2
motif n
sequence features
200 bp regions
B
GC dinucleotides
AG dinucleotides
Dinucleotide freq.
Bound
Unbound
Distance from peak summit (bp)
C
TFBS density
Bound
Unbound
Distance from peak summit (bp)
D
TPR
FPR
LR: best PWM
LR: full model
gkm-SVM
Precision
Recall
E
ROC
Precision-Recall
AUC (LR: full model)
max(AUC)=0.98
max(AUC)=0.85
Window size (bp)


A
pCrx (206 bp)
or pRho (210 bp)
CRE
(100 bp)
DsRed
(678 bp)
barcode
(13 bp)
AAAAAAAA
CRE1
CRE2
CREn
1230 100 bp CRX ChIP-seq peaks
Clone target CREs into barcoded reporter library
Electroporate library into P0 retina, culture 8 days
Harvest mRNA and DNA
PCR barcodes and sequence
B
pRho
pCrx
9%
7%
4%
10%
count
log2(cDNA/DNA)
C
PCC=0.15
PCC=0.09
log2(cDNA/DNA)
Number of E-box binding sites
D
CC/GG
NR 1
E-box 1
E-box 2
ZF
E-box 3
E-box 4
NR 2
NR 3
NR 4
NR 5
E-box 5
NR 6
GC
CA/TG
AC/GT
T-box 1
HD (K50)
T-box 2
HD (Q50) 1
HD (Q50) 2
TA
AT
AA/TT
PCC
E
Retina
H3K27ac ChIP-seq
Retina DNase-seq
Rod ATAC-seq
NRL ChIP-seq
Cone ATAC-seq
CRX ChIP-seq
H3K4me3 ChIP-seq
Liver DNase-seq
T cell DNase-seq
B cell DNase-seq
H3K27me3 ChIP-seq
F
TPR
FPR
LogR: CRX ChIP
SVM: CRX ChIP
SVM: combined


A
Average conservation
1.00
0.75
0.50
0.25
0.00
-10
0
10
20
Distance to TAAT (bp)
B
WT
(n=997)
TCTAATCCC
GTTAATCTGCTTAGA
mutant
(n=1756)
C
count
200
100
0
74%
26%
-2.0
0.0
2.0
log2(mut/WT)
D
-log10(p-value)
6
4
2
0
85%
15%
log2(mut/WT)
E
|log2(mut/WT)|
3.0
2.0
1.0
0.0
very low
low
medium
high
motif score (affinity)
F
two monomeric sites
both half-sites in dimer
log2(cDNA/DNA)
2.0
0.0
-2.0
WT
site1
site2
both


A
WT
(n=195)
mutant
(n=7528)
CTAATCCTTCCAG
CTAAGCTGCTTAC
B
Target seqeunes (n=195)
Position
log2(mut/WT)
2.0
-2.0
C
Variant nucleotide
Quantitative gel shift
CRE-seq
monomer
dimer
Δaffinity (in vitro)
Δactivity (in vivo)
0.0
-8.0
0.4
-0.8
Position
D
log2(mut/WT)
phyloP
Position


A
WT (n=195)
Deletions
1 bp
2 bp
3 bp
p < 10-10 p < 10-9 p < 10-3
log2(mut/WT)
2.00
0.00
-2.00
1 bp 2 bp 3 bp
Deletion size (bp)
Monomeric sites
Dimeric sites
B
WT (n=195)
Insertions
1 bp
2 bp
3 bp
p < 10-7 p < 10-7 p < 10-6
log2(mut/WT)
2.00
0.00
-2.00
1 bp 2 bp 3 bp
Insertion size (bp)
Monomeric sites
Dimeric sites
C
WT (n=98)
Spacer substitutions
CCG
CAG
AAG
AGG
CTC
CCA
log2(mut/WT)
2.00
0.00
-2.00
CCG CAG AAG AGG CTC CCA
Motif count
175
0
K50
Q50


A
Prediction of log2(mut/wt)
within CREs
Performance (R2)
1.00
0.75
0.50
0.25
0.00
ΔPWM
CRX deltaSVM
B
Prediction of log2(mut/wt)
across CREs
Performance (R2)
1.00
0.75
0.50
0.25
0.00
ΔPWM
CRX deltaSVM
C
Prediction of log2(mut/wt)
across CREs
Performance (R2)
1.00
0.75
0.50
0.25
0.00
Sequence features
m-deltaSVM
m-deltaSVM+ m-gkm-SVM
m-deltaSVM + m-gkm-SVM
+ m-deltaSVM:m-gkm-SVM
D
Prediction of log2(mut)
across CREs
Performance (R2)
1.00
0.75
0.50
0.25
0.00
Sequence features
+ CRE-seq
WT
WT + m-deltaSVM
WT + m-deltaSVM
+ m-deltaSVM:WT

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

Andrew E.O. Hughes, Connie A. Myers and Joseph C. Corbo

Genome Res. published online August 29, 2018
Access the most recent version at doi:10.1101/gr.231886.117