Information Content Differentiates Enhancers From Silencers in Mouse Photoreceptors

Ryan Z. Friedman[1], David M. Granas[1], Connie A. Myers[2], Joseph C. Corbo[2], Barak A. Cohen[1], Michael A. White[1]*

[1]Edison Family Center for Genome Sciences and Systems Biology, and Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA
[2]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA
*Corresponding author (mawhite@wustl.edu)

ORCID IDs: 0000-0001-9013-8676 (RZF), 0000-0002-9323-7140 (JCC), 0000-0002-3350-2715 (BAC), 0000-0001-8511-6026 (MAW)

Abstract

Enhancers and silencers often depend on the same transcription factors (TFs) and are conflated in genomic assays of TF binding or chromatin state. To identify sequence features that distinguish enhancers and silencers, we assayed massively parallel reporter libraries of genomic sequences targeted by the photoreceptor TF CRX in mouse retinas. Both enhancers and silencers contain more TF motifs than inactive sequences, but relative to silencers, enhancers contain motifs from a more diverse collection of TFs. We developed a measure of information content that describes the number and diversity of motifs in a sequence and found that, while both enhancers and silencers depend on CRX motifs, enhancers have higher information content. The ability of information content to distinguish enhancers and silencers targeted by the same TF illustrates how motif context determines the activity of *cis*-regulatory sequences.

Introduction

Active *cis*-regulatory sequences in the genome are characterized by accessible chromatin and specific histone modifications, which reflect the action of DNA-binding transcription factors (TFs) that recognize specific sequence motifs and recruit chromatin modifying enzymes (Klemm et al., 2019). These epigenetic hallmarks of active chromatin are routinely used to train machine learning models that predict *cis*-regulatory sequences, based on the assumption that such epigenetic marks are reliable predictors of genuine *cis*-regulatory sequences (Ernst and Kellis, 2012; Ghandi et al., 2014; Hoffman et al., 2012; Kelley et al., 2016; Lee et al., 2011; Sethi et al., 2020; Zhou and Troyanskaya, 2015). However, results from functional assays show that many predicted *cis*-regulatory sequences exhibit little or no *cis*-regulatory activity. Typically, 50% or more of predicted *cis*-regulatory sequences fail to drive expression in Massively Parallel Reporter Assays (MPRAs) (ENCODE Project Consortium et al., 2020; Kwasnieski et al., 2014), indicating that an active chromatin state is not sufficient to reliably identify *cis*-regulatory sequences.

Another challenge is that both enhancers and silencers often respond to the same TFs (Alexandre and Vincent, 2003; Iype et al., 2004; Liu et al., 2014; Martínez-Montañés et al., 2013; Peng et al., 2005; Stampfel et al., 2015; White et al., 2016). Many TF-bound sequences act as silencers (Clyde et al., 2003; Doni Jayavelu et al., 2020; Gisselbrecht et al., 2020; Grossman et al., 2017; Kwasnieski et al., 2014; Truskina et al., 2020; White et al., 2013), but predictions of *cis*-regulatory sequences based on TF binding or chromatin state often do not make a distinction between these two classes of sequence. Since enhancers in one cell type are often silencers in other cell types (Barolo and Posakony, 2002; Clyde et al., 2003; Gisselbrecht et al., 2020; Junion et al., 2012; Parker et al., 2011; Rachmin et al., 2015; Swanson et al., 2010), enhancers and silencers likely share similar sequence features. Understanding how enhancers and silencers are distinguished in a particular cell type remains an important challenge.

Using MPRAs in explanted, intact mouse retinas, we assayed the activity of thousands of candidate *cis*-regulatory sequences targeted by cone-rod homeobox (CRX), a K50 homeodomain TF required for the specification and maintenance of rod and cone photoreceptors (Chen et al., 1997; Freund et al., 1997; Furukawa et al., 1997). We show that while regions of accessible chromatin and CRX binding exhibit a range of *cis*-regulatory activity, enhancers and silencers contain more TF motifs than inactive sequences, and that enhancers are distinguished from silencers by a higher diversity of TF motifs. We capture the differences between these sequence classes with a new metric, sequence information content, that considers only the number and diversity of TF motifs in a candidate *cis*-regulatory sequence. Our results suggest that CRX-targeted enhancers are defined by a flexible regulatory grammar and demonstrate how DNA sequence encodes functional differences between genomic loci with similar chromatin states.

Results

We tested the activities of 4,844 putative CRX-targeted *cis*-regulatory sequences (CRX-targeted sequences) by MPRA in live retinas. The MPRA libraries consist of 164 bp genomic sequences centered on the best match to the CRX position weight matrix (PWM) (Lee et al., 2010) whenever possible and matched sequences in which all CRX motifs were abolished by point mutation (Methods). The MPRA libraries include 3,299 CRX-bound sequences identified by ChIP-seq in the adult retina (Corbo et al., 2010) and 1,545 sequences that do not have measurable CRX binding in the adult retina but reside in accessible chromatin in adult photoreceptors (Hughes et al., 2017) and have the H3K27ac enhancer mark in postnatal day 14 (P14) retina (Ruzycki et al., 2018) ("ATAC-seq peaks"). We split the sequences across two plasmid libraries, each of which contained the same 150 scrambled sequences as internal controls (**Supplementary Datasets 1 and 2**). We cloned sequences upstream of the *Rhodopsin* (*Rho*) promoter and a *DsRed* reporter gene, electroporated libraries into explanted mouse retinas at P0 in triplicate, harvested the retinas at P8, and then sequenced the RNA and input DNA plasmid pool. The data is highly reproducible across replicates ($R^2 > 0.96$, **Supplementary Figure 1**). After activity scores were calculated and normalized to the basal *Rho* promoter, the two libraries were well calibrated and merged together (two-sample Kolmogorov-Smirnov test $p = 0.09$, **Supplementary Figure 2, Supplementary Dataset 3**, and Methods).

Strong enhancers and silencers have high CRX motif content

The *cis*-regulatory activities of CRX-targeted sequences vary widely (**Figure 1a**). We defined enhancers and silencers as those sequences that have statistically significant activity that is at least two-fold above or below the activity of the basal *Rho* promoter (Welch's t-test, Benjamini-Hochberg false discovery rate (FDR) $q < 0.05$). We defined inactive sequences as those whose activity is both within a two-fold change of basal activity and not significantly different from the basal *Rho* promoter. We further stratified enhancers into strong and weak enhancers based on whether or not they fell above the 95th percentile of scrambled sequences. Using these criteria, 22% of CRX-targeted sequences are strong enhancers, 28% are weak enhancers, 19% are inactive, and 17% are silencers; the remaining 13% were considered ambiguous and removed from further analysis. We sought to identify features that would accurately classify these different classes of sequences.

Neither CRX ChIP-seq binding status nor DNA accessibility as measured by ATAC-seq strongly differentiates between these four classes (**Figure 1b**). Compared to CRX ChIP-seq peaks, ATAC-seq peaks that lack CRX binding in the adult retina are slightly enriched for inactive sequences (Fisher's exact test $p = 2 \times 10^{-7}$, odds ratio = 1.5) and slightly depleted for strong enhancers (Fisher's exact test $p = 1 \times 10^{-21}$, odds ratio = 2.2). However, sequences with ChIP-seq or ATAC-seq peaks span all four activity categories, demonstrating that DNA accessibility and TF binding data are not sufficient to identify functional enhancers and silencers.

We examined whether the number and affinity of CRX motifs differentiate enhancers, silencers, and inactive sequences by computing the predicted CRX occupancy (i.e. expected number of bound molecules) for each sequence (White et al., 2013). Consistent with our previous work (White et al., 2016), both strong enhancers and silencers have higher predicted CRX occupancy than inactive sequences (Mann-Whitney U test, $p = 6 \times 10^{-10}$ and 6×10^{-17} respectively, **Figure 1c**), suggesting that total CRX motif content helps distinguish silencers and strong enhancers from inactive sequences. However, predicted CRX occupancy does not distinguish strong enhancers from silencers: a logistic regression classifier trained with five-fold cross-validation only achieves an area under the receiver operating characteristic (AUROC) curve of 0.548±0.023 and an area under the precision recall (AUPR) curve of 0.571±0.020 (**Figure 2a** and **Supplementary Figure 3**). We thus sought to identify sequence features that distinguish strong enhancers from silencers.

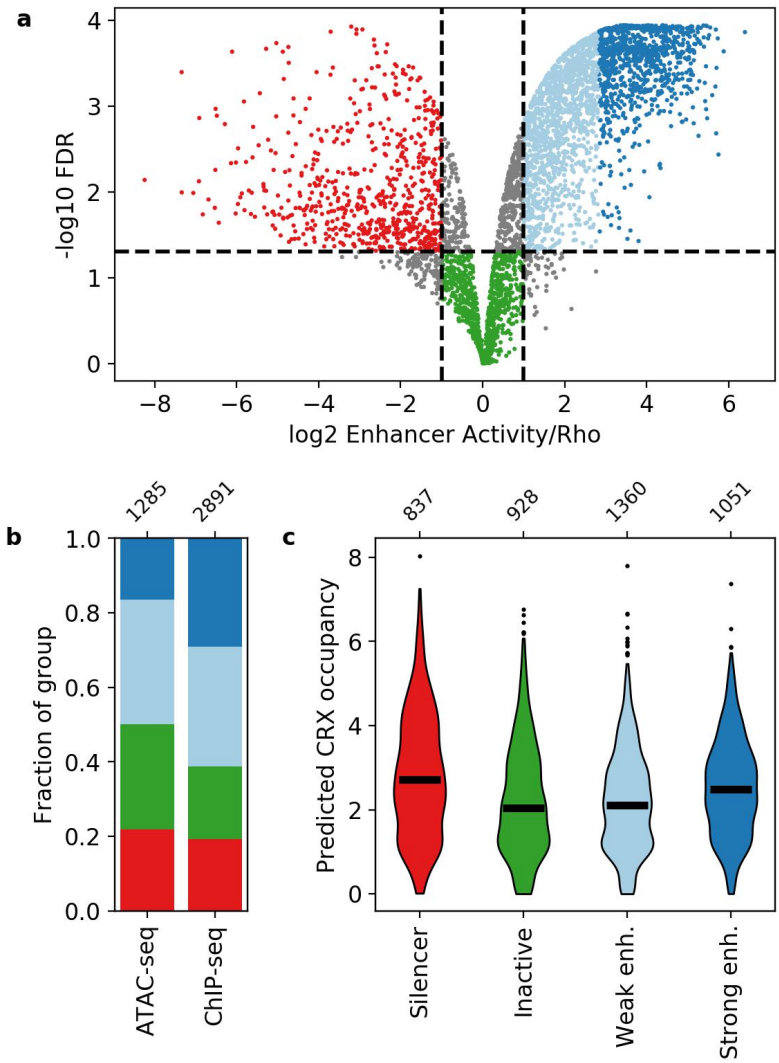


Figure 1: Activity of putative *cis*-regulatory sequences with CRX motifs. a) Volcano plot of activity scores relative to the *Rho* promoter alone. Sequences are grouped as strong enhancers (dark blue), weak enhancers (light blue), inactive (green), silencers (red), or ambiguous (grey). Horizontal line, FDR q = 0.05. Vertical lines, 2-fold above and below *Rho*. **b)** Fraction of ChIP-seq and ATAC-seq peaks that belong to each activity group. **c)** Predicted CRX occupancy of each activity group. Horizontal lines, medians; enh., enhancer. Numbers at top of **(b and c)** indicate n for groups.

Lineage-defining TF motifs differentiate strong enhancers from silencers

We performed a *de novo* motif enrichment analysis to identify motifs that distinguish strong enhancers from silencers and found several motifs matching known TFs that are differentially enriched between strong enhancers and silencers. For motifs that matched multiple TFs, we selected one representative TF for downstream analysis (**Supplementary Figure 4**, Methods). Strong enhancers are enriched for several motif families that include TFs that interact with CRX or are important for photoreceptor development: NeuroD1/NDF1 (E-box binding bHLH) (Morrow et al., 1999), RORB (nuclear receptor) (Jia et al., 2009; Srinivas et al., 2006), MAZ or Sp4

(C2H2 zinc finger) (Lerner et al., 2005), and NRL (bZIP) (Mears et al., 2001; Mitton et al., 2000). Sp4 physically interacts with CRX in the retina (Lerner et al., 2005), but we chose to represent the zinc finger motif with MAZ based on the PWM qualities. Silencers were enriched for a motif that resembles a partial K50 homeodomain motif but instead matches the zinc finger TF GFI1, a member of the Snail repressor family (Chiang and Ayyanathan, 2013) expressed in developing retinal ganglion cells (Yang et al., 2003). Therefore, while strong enhancers and silencers are not distinguished by their CRX motif content, strong enhancers are uniquely enriched for several lineage-defining TFs.

To quantify how well these TF motifs differentiate strong enhancers from silencers, we trained two different classification models with five-fold cross-validation. First, we trained a 6-mer support vector machine (SVM) (Ghandi et al., 2014) and achieved an AUROC of 0.781±0.013 and AUPR of 0.812±0.020 (**Figure 2a** and **Supplementary Figure 3**). The SVM considers all 2,080 non-redundant 6-mers and provides an upper bound to the predictive power of models that do not consider the exact arrangement or spacing of sequence features. We next trained a logistic regression model on the predicted occupancy for eight lineage-defining TFs and compared it to the upper bound established by the SVM. In this model, we considered CRX, the five TFs identified in our motif enrichment analysis, and two additional TFs enriched in photoreceptor ATAC-seq peaks (Hughes et al., 2017): RAX, a Q50 homeodomain TF that contrasts with CRX, a K50 homeodomain TF (Irie et al., 2015); and MEF2D, a MADS box TF which co-binds with CRX (Andzelm et al., 2015). The logistic regression model performs nearly as well as the SVM (AUROC 0.698±0.036, AUPR 0.745±0.032, **Figure 2a** and **Supplementary Figure 3**) despite a 260-fold reduction from 2,080 to eight features. To determine whether the logistic regression model depends specifically on the eight lineage-defining TFs, we established a null distribution by fitting 100 logistic regression models with randomly selected TFs (Methods). Our logistic regression model outperforms the null distribution (one-tailed Z-test for AUROC and AUPR, $p < 0.0008$, **Supplementary Figure 5**), indicating that the performance of the model specifically requires the eight lineage-defining TFs. To determine whether the SVM identified any additional motifs that could be added to the logistic regression model, we generated *de novo* motifs using the SVM *k*-mer scores and found no additional motifs predictive of strong enhancers. Finally, we found that our two models perform similarly on an independent test set of CRX-targeted sequences (White et al., 2013) (**Supplementary Figure 5**). Since the logistic regression model performs near the upper bound established by the SVM and depends specifically on the eight selected motifs, we conclude that these motifs comprise nearly all of the sequence features captured by the SVM that distinguish strong enhancers from silencers among CRX-targeted sequences.

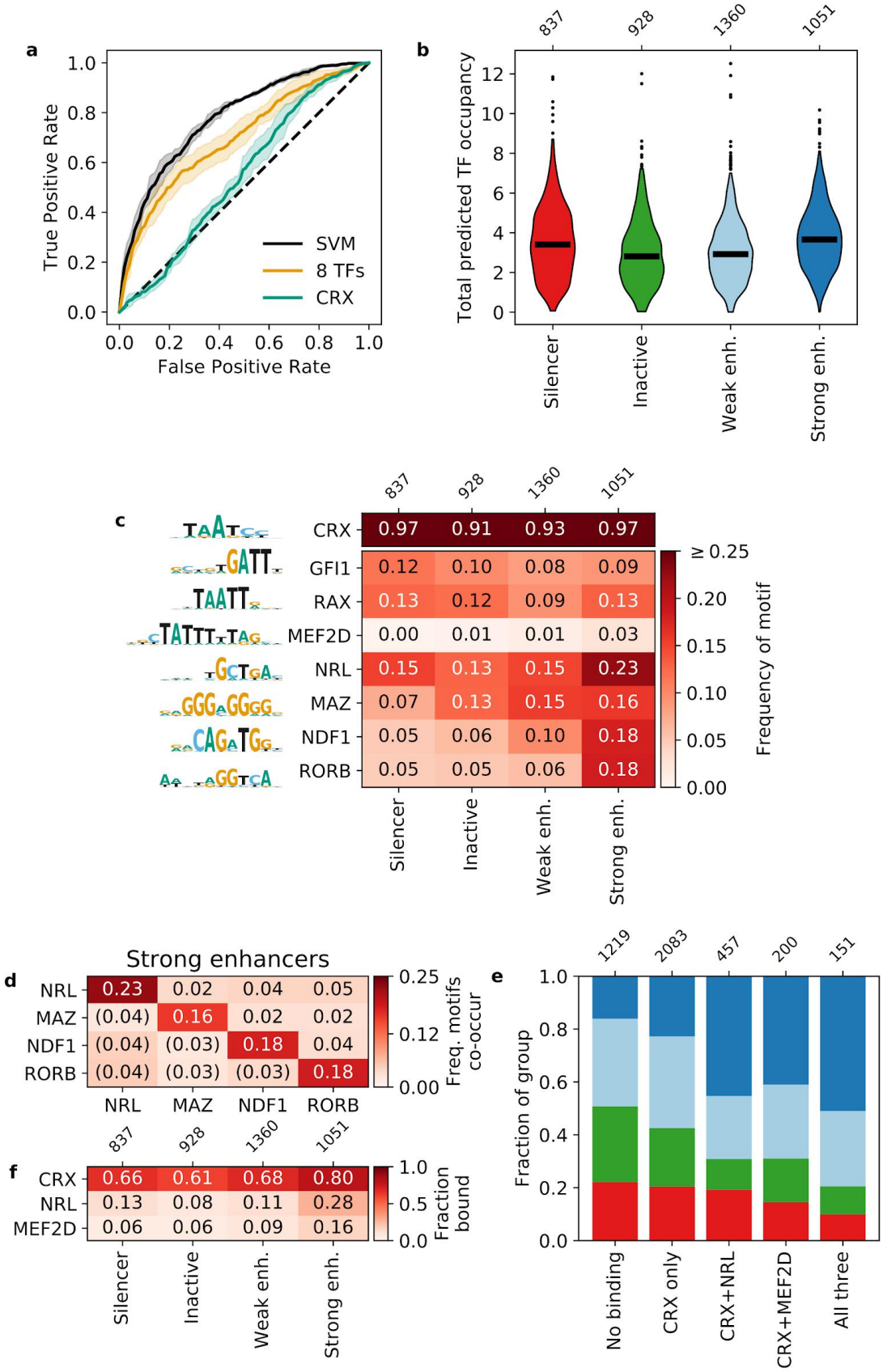

a
True Positive Rate
False Positive Rate
SVM
8 TFs
CRX
b
Total predicted TF occupancy
837
928
1360
1051
Silencer
Inactive
Weak enh.
Strong enh.
c
CRX
GFI1
RAX
MEF2D
NRL
MAZ
NDF1
RORB
Frequency of motif
d
Strong enhancers
Freq. motifs co-occur
e
Fraction of group
1219
2083
457
200
151
No binding
CRX only
CRX+NRL
CRX+MEF2D
All three
f
Fraction bound

Figure 2: Strong enhancers contain a diverse array of motifs. a) Receiver operating characteristic for classifying strong enhancers from silencers. Solid black, 6-mer SVM; orange, 8 TF predicted occupancy logistic regression; aqua, predicted CRX occupancy logistic regression; dashed black, chance; shaded area, 1 standard deviation based on five-fold cross-validation. **b and c)** total predicted TF occupancy **(b)** and frequency of TF motifs **(c)** in each activity class. **d)** Frequency of co-occurring TF motifs in strong enhancers. Lower triangle is expected co-occurrence if motifs are independent. **e)** Frequency of activity classes, colored as in **(b)**, for sequences in CRX, NRL, and/or MEF2D ChIP-seq peaks. **f)** Frequency of TF ChIP-seq peaks in activity classes. TFs in **(c)** are sorted by feature importance of the logistic regression model in **(a)**.

Strong enhancers are characterized by diverse total motif content

To understand how these eight TF motifs differentiate strong enhancers from silencers, we first calculated the total predicted occupancy of each sequence by all eight lineage-defining TFs and compared the different activity classes. Strong enhancers and silencers both have higher total predicted occupancies than inactive sequences, but total predicted occupancies do not distinguish strong enhancers and silencers from each other (**Figure 2b, Supplementary Dataset 5**). Since strong enhancers are enriched for several motifs relative to silencers, this suggests that strong enhancers are distinguished from silencers by the diversity of their motifs, rather than the total number.

We considered two hypotheses for how the more diverse collection of motifs function in strong enhancers: either strong enhancers depend on specific combinations of TF motifs ("TF identity hypothesis"), or they instead must be co-occupied by multiple lineage-defining TFs, regardless of TF identity ("TF diversity hypothesis"). To distinguish between these hypotheses, we examined which specific motifs contribute to the total motif content of strong enhancers and silencers. We considered motifs for a TF present in a sequence if the TF predicted occupancy was above 0.5 molecules, which generally corresponds to at least one motif with a relative K_D above 3%. This threshold captures the effect of low affinity motifs that are often biologically relevant (Crocker et al., 2015; Farley et al., 2016, 2015; Parker et al., 2011). As expected, 97% of strong enhancers and silencers contain CRX motifs since the sequences were selected based on CRX binding or significant matches to the CRX PWM within open chromatin (**Figure 2c**). Compared to silencers, strong enhancers contain a broader diversity of motifs for the eight lineage-defining TFs (**Figure 2c**). However, while strong enhancers contain a broader range of motifs, no single motif occurs in a majority of strong enhancers: NRL motifs are present in 23% of strong enhancers, NeuroD1 and RORB in 18% each, and MAZ in 16%. Additionally, none of the motifs tend to co-occur as pairs in strong enhancers: no specific pair occurred in more than 5% of sequences (**Figure 2d**). Similarly, no single motif occurs in more than 15% of silencers (**Figure 2c**). These results suggest that strong enhancers are defined by the diversity of their motifs, and not by specific motif combinations.

The results above predict that strong enhancers are more likely to be bound by a diverse but degenerate collection of TFs, compared with silencers or inactive sequences. We tested this prediction by examining *in vivo* TF binding using published ChIP-seq data for NRL (Hao et al.,

2012) and MEF2D (Andzelm et al., 2015). Consistent with the prediction, sequences bound by CRX and either NRL or MEF2D are approximately twice as likely to be strong enhancers compared to sequences only bound by CRX (**Figure 2e**). Sequences bound by all three TFs are the most likely to be strong or weak enhancers rather than silencers or inactive sequences. However, most strong enhancers are not bound by either NRL or MEF2D (**Figure 2f**), indicating that binding of these TFs is not required for strong enhancers. Our results support the TF diversity hypothesis: CRX-targeted enhancers are co-occupied by multiple TFs, without a requirement for specific combinations of lineage-defining TFs.

Strong enhancers have higher motif information content than silencers

Our results indicate that both strong enhancers and silencers have a higher total motif content than inactive sequences, while strong enhancers contain a more diverse collection of motifs than silencers. To quantify these differences in the number and diversity of motifs, we computed the information content of CRX-targeted sequences using Boltzmann entropy. The Boltzmann entropy of a system is related to the number of ways the system's molecules can be arranged, which increases with either the number or diversity of molecules (Phillips et al., 2012, chap. 5). In our case, each TF is a different type of molecule and the number of each TF is represented by its predicted occupancy for a *cis*-regulatory sequence. The number of molecular arrangements is thus W, the number of distinguishable permutations that the TFs can be ordered on the sequence, and the information content of a sequence is then $\log_2 W$ (Methods).

We found that on average, strong enhancers have higher information content than both silencers and inactive sequences (Mann-Whitney U test, $p = 1 \times 10^{-23}$ and 7×10^{-34} respectively, **Figure 3a, Supplementary Dataset 5**), confirming that information content captures the effect of both the number and diversity of motifs. Quantitatively, the average silencer and inactive sequence contains 1.6 and 1.4 bits, respectively, which represents approximately three total motifs for two TFs. Strong enhancers contain on average 2.4 bits, representing approximately three total motifs for three TFs or four total motifs for two TFs. To compare the predictive value of our information content metric to the model based on all eight motifs, we trained a logistic regression model and found that information content classifies strong enhancers from silencers with an AUROC of 0.634±0.008 and an AUPR of 0.663±0.014 (**Figure 3b** and **Supplementary Figure 6**). This is only slightly worse than the model trained on eight TF occupancies despite an eight-fold reduction in the number of features, which is itself comparable to the SVM with 2,080 features. The difference between the two logistic regression models suggests that the specific identities of TF motifs make some contribution to the eight TF model, but that most of the signal captured by the SVM can be described with a single metric that does not assign weights to specific motifs. Information content also distinguishes strong enhancers from inactive sequences (AUROC 0.658±0.012, AUPR 0.675±0.019, **Figure 3b** and **Supplementary Figure 6**). These results indicate that strong enhancers are characterized by higher information content, which reflects both the total number and diversity of motifs.

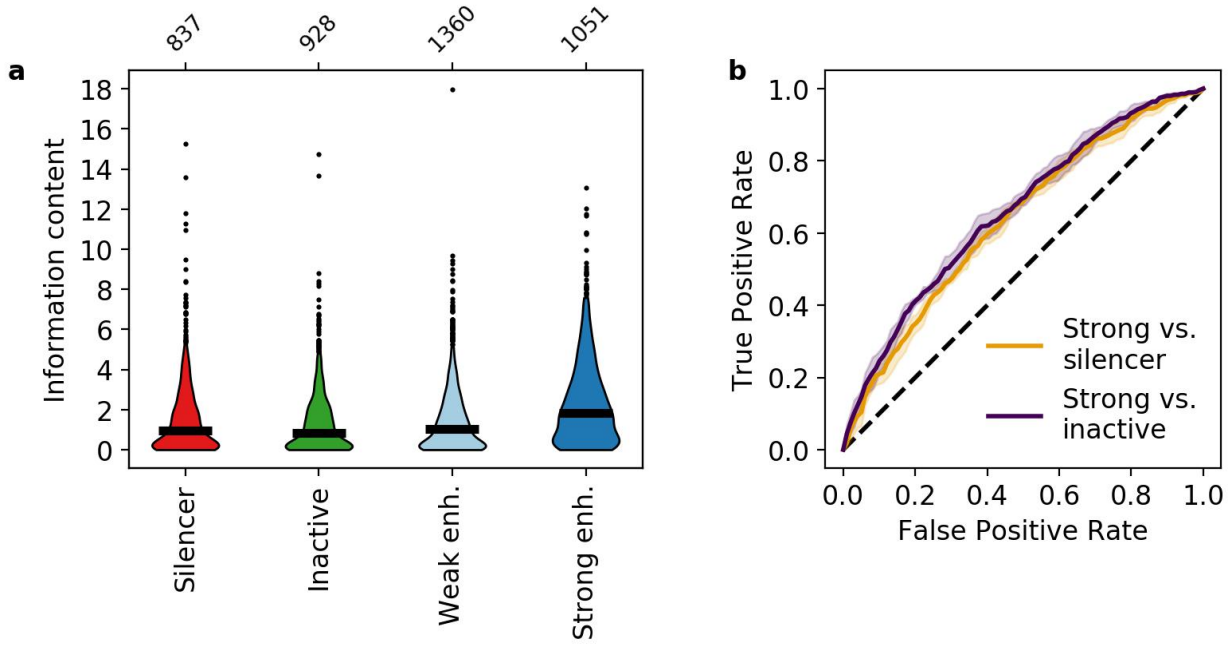


Figure 3: Information content classifies strong enhancers. a) Information content for different activity classes. **b)** Receiver operating characteristic of information content to classify strong enhancers from silencers (orange) or inactive sequences (indigo).

Strong enhancers require high information content but not NRL motifs

Our results show that except for CRX, none of the lineage-defining motifs occur in a majority of strong enhancers. However, all sequences were tested in reporter constructs with the *Rho* promoter, which contains an NRL motif and three CRX motifs (Corbo et al., 2010; Kwasnieski et al., 2012). Since NRL is a key co-regulator with CRX in rod photoreceptors, we tested whether strong enhancers generally require NRL, which would be inconsistent with our TF diversity hypothesis. We removed the NRL motif by recloning our MPRA library without the basal *Rho* promoter. If strong enhancers require an NRL motif for high activity, then only CRX-targeted sequences with NRL motifs will drive reporter expression. If information content (i.e. total motif content and diversity) is the primary determinant of strong enhancers, only CRX-targeted sequences with sufficient motif diversity, measured by information content, will drive reporter expression regardless of whether or not NRL motifs are present.

We replaced the *Rho* promoter with a minimal 23 bp polylinker sequence between our libraries and *DsRed*, and repeated the MPRA (**Supplementary Figure 1, Supplementary Dataset 3**). CRX-targeted sequences were designated as “autonomous” if they retained activity in the absence of the *Rho* promoter ($\log_2$(RNA/DNA) > 0, Methods). We found that 90% of autonomous sequences are enhancers in the presence of the *Rho* promoter (**Figure 4a**). However, while most autonomous sequences are enhancers, only 39% of strong enhancers and 9% of weak enhancers act autonomously. Consistent with a role for information content, autonomous strong enhancers have higher information content (Mann-Whitney U test $p = 4 \times 10^{-8}$, **Figure 4b**) and higher predicted CRX occupancy (Mann-Whitney U test $p = 9 \times 10^{-12}$,

Figure 4c) than non-autonomous strong enhancers. We found no evidence that specific lineage-defining motifs are required for autonomous activity, including NRL, which is present in only 25% of autonomous strong enhancers (**Figure 4d**). Similarly, NRL ChIP-seq binding (Hao et al., 2012) occurs more often among autonomous strong enhancers (41% vs. 19%, Fisher's exact test $p = 2 \times 10^{-14}$, odds ratio = 3.0), yet NRL binding still only accounts for a minority of these sequences. We thus conclude that strong enhancers require high information content, rather than any specific lineage-defining motifs.

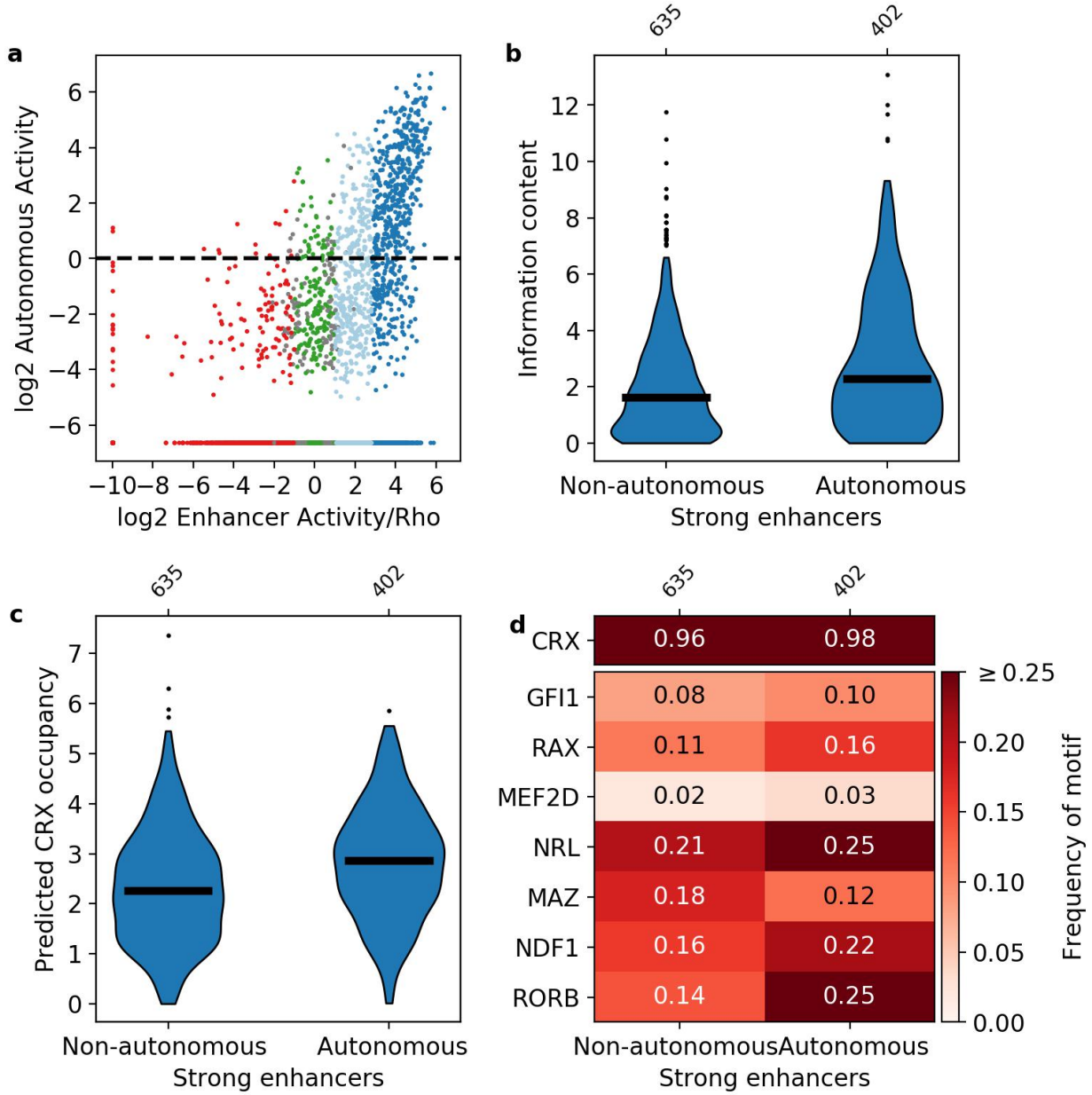


Figure 4: Sequence features of autonomous and non-autonomous strong enhancers. a) Activity of library in the presence (x-axis) or absence (y-axis) of the *Rho* promoter. Dark blue, strong enhancers; light blue, weak enhancers; green, inactive; red, silencers; grey, ambiguous; horizontal line, cutoff for autonomous activity. Points on the far left and/or very bottom are sequences that were present in the plasmid pool but not detected in the RNA. **b-d)** Comparison

of autonomous and non-autonomous strong enhancers for information content **(b)**, predicted CRX occupancy **(c)**, and frequency of TF motifs **(d)**.

TF motifs contribute independently to strong enhancers

Our results indicate that information content distinguishes strong enhancers from silencers and inactive sequences. Information content only takes into account the total number and diversity of motifs in a sequence and not any potential interactions between them. The classification success of information content thus suggests that each TF motif will contribute independently to enhancer activity. We tested this prediction with CRX-targeted sequences where all CRX motifs were abolished by point mutation (**Supplementary Dataset 3**). Consistent with our previous work (White et al., 2013), mutating CRX motifs causes the activities of both enhancers and silencers to regress towards basal levels (Pearson's r = 0.608, **Figure 5a**), indicating that most enhancers and silencers show some dependence on CRX. However, 40% of wild-type strong enhancers show low CRX dependence and remain strong enhancers with their CRX motifs abolished. Although strong enhancers with high and low CRX dependence have similar wild-type information content (**Figure 5b**), strong enhancers with low CRX dependence have lower predicted CRX occupancy than those with high CRX dependence (Mann-Whitney U test $p = 2 \times 10^{-9}$, **Figure 5c**), and also have higher "residual" information content (i.e. information content without CRX motifs, Mann-Whitney U test $p = 1 \times 10^{-7}$, **Figure 5d**). Low CRX dependence sequences have an average of 1.5 residual bits, which corresponds to three motifs for two TFs, while high CRX dependence sequences have an average of 1.0 residual bits, which corresponds to two motifs for two TFs (**Figure 5e**).

Strong enhancers with low and high CRX dependence have similar wild-type information content and similar total predicted occupancy (**Figures 5b and e**). As a result, sequences with more CRX motifs have fewer motifs for other TFs, suggesting that there is no evolutionary pressure for enhancers to contain additional motifs beyond the minimum amount of information content required to be active. To test this idea, we calculated the minimum number and diversity of motifs necessary to specify a relatively unique location in the genome (Wunderlich and Mirny, 2009) and found that a 164 bp sequence only requires five motifs for three TFs (Methods). These motif requirements can be achieved in two ways with similar information content that differ only in the quantitative number of motifs for each TF. In other words, the number of motifs for any particular TF is not important so long as there is sufficient information content. Taken together, we conclude that each TF motif provides an independent contribution towards specifying strong enhancers.

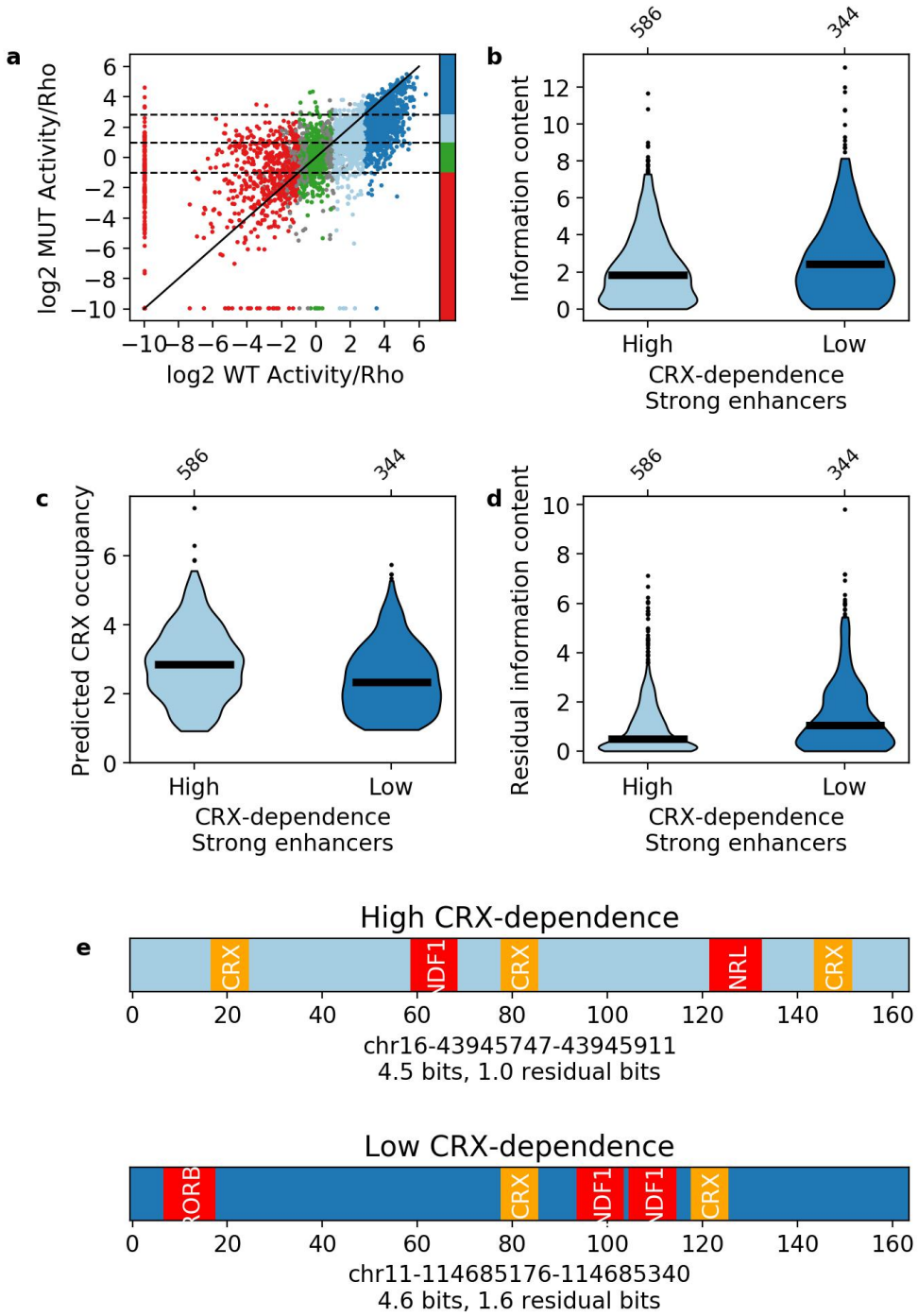


Figure 5: Independence of TF motifs in strong enhancers. a) Activity of sequences with and without CRX motifs. Points are colored by the activity group with CRX motifs intact: dark blue, strong enhancers; light blue, weak enhancers; green, inactive; red, silencers; grey, ambiguous; horizontal dotted lines and color bar represent the cutoffs for the same groups when CRX motifs are mutated. Solid black line is the y = x line. **b-d)** Comparison of strong enhancers with high and low CRX dependence for information content **(b)**, predicted CRX occupancy **(c)**, and residual information content **(d)**. **(e)** Representative strong enhancers with high (top) or low (bottom) CRX-dependence.

Discussion

Many regions in the genome are bound by TFs and bear the epigenetic hallmarks of active *cis*-regulatory sequences, yet fail to exhibit *cis*-regulatory activity when tested directly. The discrepancy between measured epigenomic state and *cis*-regulatory activity indicates that enhancers and silencers consist of more than the minimal sequence features necessary to recruit TFs and chromatin-modifying factors. Our results show that enhancers, silencers, and inactive sequences in developing photoreceptors can be distinguished by their motif content, even though they are indistinguishable by CRX binding or chromatin accessibility. We show that both enhancers and silencers contain more TF motifs than inactive sequences, and that enhancers also contain more diverse sets of motifs for lineage-defining TFs. These differences are captured by our measure of information content. Information content, as a single metric, identifies strong enhancers nearly as well as an unbiased set of 2,080 non-redundant 6-mers used for an SVM, indicating that a simple measure of motif number and diversity can capture the key sequence features that distinguish enhancers from other sequences that lie in open chromatin.

The results of our information content classifier are consistent with the TF collective model of enhancers (Junion et al., 2012; Spitz and Furlong, 2012): globally, active enhancers are specified by the combinatorial action of lineage-defining TFs with little constraint on which motifs must co-occur. We show that CRX-targeted enhancers are distinguished from inactive CRX-targeted sequences by a larger, more diverse collection of TF motifs, and not any specific combination of motifs. This indicates that enhancers are active because they have acquired the necessary number of TF binding motifs, and not because they are defined by a strict regulatory grammar. Sequences with fewer motifs may be bound by CRX and reside within open chromatin, but they lack sufficient TF binding for activity. Such loose constraints would facilitate the *de novo* emergence of tissue-specific enhancers and silencers over evolution and explain why critical cell type-specific TF interactions, such as CRX with NRL in rod photoreceptors, occur at only a minority of the active enhancers in that cell type (Hsiau et al., 2007; Hughes et al., 2018; White et al., 2013).

Like enhancers, silencers require higher motif content and are dependent on CRX, but they lack the TF diversity of enhancers. The lack of TF diversity in silencers parallels the architecture of signal-responsive *cis*-regulatory sequences, which are silencers in the absence of a signal and require multiple activators for induction (Barolo and Posakony, 2002). Consistent with this, we previously showed with synthetic sequences that high occupancy of CRX alone is sufficient to encode silencers while the addition of a single NRL motif converts synthetic silencers to enhancers (White et al., 2016). The contrast in motif diversity between enhancers and silencers in the current study could prevent a common TF from activating all its *cis*-regulatory targets at the same time to achieve differences in spatiotemporal gene expression. In support of this model, CRX-targeted sequences become active at different times in photoreceptor development (Ruzycki et al., 2018). We speculate that silencers may contain sufficient information to act as enhancers, but are repressive because certain activators are not expressed. This hypothesis would be consistent with the finding that many silencers are enhancers in other cell types (Gisselbrecht et al., 2020), and could explain why mouse photoreceptor and bipolar cells utilize

similar combinations of motifs but exhibit divergent transcriptomes (Murphy et al., 2019): the presence or absence of a single TF could cause the same *cis*-regulatory sequence to be an enhancer in one cell type and a silencer in the other. Our work suggests that characterizing sequences by their motif information content offers a way to identify these different classes of *cis*-regulatory sequences in the genome.

Methods

Library Design

CRX ChIP-seq peaks re-processed by Ruzycki *et al.* (Ruzycki et al., 2018) were intersected with previously published CRX MPRA libraries (Hughes et al., 2018; White et al., 2013) and one unpublished library to select sequences that had not been previously tested by MPRA. These sequences were scanned for instances of CRX motifs using FIMO version 4.11.2 (Bailey et al., 2009), a p-value cutoff of 2.3×10^{-3}, and a CRX PWM derived from an electrophoretic mobility shift assay (Lee et al., 2010). We centered 2622 sequences on the highest scoring CRX motif. For 677 sequences without a CRX motif, we instead centered them using the Gibbs sampler from ShapeMF (Github commit abe8421) (Samee et al., 2019) and a motif size of 10.

For sequences unbound in CRX ChIP-seq but in open chromatin, we took ATAC-seq peaks collected in 8 week FACS-purified rods, green cones, and *Nrl*$^{-/-}$ blue cones (Hughes et al., 2017) and removed sequences that overlapped with CRX ChIP-seq peaks. The remaining sequences were scanned for instances of CRX motifs using FIMO with a p-value cutoff of 2.5×10^{-3} and the CRX PWM. Sequences with a CRX motif were kept and the three ATAC-seq data sets were merged together, intersected with H3K27ac and H3K4me3 ChIP-seq peaks collected in P14 retinas (Ruzycki et al., 2018), and centered on the highest scoring CRX motifs. We randomly selected 1004 H3K27ac^{+}H3K4me3^{-} sequences and 541 H3K27ac^{+}H3K4me3^{+} to reflect the fact that ~35% of CRX ChIP-seq peaks are H3K4me3^{+}. After synthesis of our library, we discovered 11% of these sequences do not actually overlap H3K27ac ChIP-seq peaks (110/1004 of the H3K4me3^{-} group and 60/541 of the H3K4me3^{+} group), but we still included them in the analysis because they contain CRX motifs in ATAC-seq peaks.

All data was converted to mm10 coordinates using the UCSC liftOver tool (Haeussler et al., 2019) and processed using Bedtools version 2.27.1 (Quinlan and Hall, 2010). All sequences in our library design were adjusted to 164 bp and screened for instances of EcoRI, SpeI, SphI, and NotI sites. In total, our library contains 4844 genomic sequences (2622 CRX ChIP-seq peaks with motifs, 677 CRX ChIP-seq peaks without motifs, 1004 CRX^{-}ATAC^{+}H3K27ac^{+}H3K4me3^{-} CRX motifs, and 541 CRX^{-}ATAC^{+}H3K27ac^{+}H3K4me3^{+} CRX motifs), a variant of each sequence with all CRX motifs mutated, 150 scrambled sequences, and a construct for cloning the basal promoter alone.

For sequences centered on CRX motifs, all CRX motifs with a p-value of 2.5×10^{-3} or less were mutated by changing the core TAAT to TACT (Lee et al., 2010) on the appropriate strand, as described previously (Hughes et al., 2018; White et al., 2013). We then re-scanned sequences and mutated any additional motifs inadvertently created.

To generate scrambled sequences, we randomly selected 150 CRX ChIP-seq peaks spanning the entire range of GC content in the library. We then scrambled each sequence while preserving dinucleotide content as previously described (White et al., 2013). We used FIMO to confirm that none of the scrambled sequences contain CRX motifs.

Plasmid Library Construction

We generated two 15,000 libraries of 230 bp oligonucleotides (oligos) from Agilent Technologies through a limited licensing agreement. Our library was split across the two oligo pools, ensuring that both the genomic and mutant forms of each sequence were placed in the same oligo pool (**Supplementary Datasets 1 and 2**). Both oligo pools contain all 150 scrambled sequences as an internal control. All sequences were assigned three unique barcodes as previously described (White et al., 2013). In each oligo pool, the basal promoter alone was assigned 18 unique barcodes. Oligos were synthesized as follows: 5' priming sequence (GTAGCGTCTGTCCGT)/EcoRI site/Library sequence/SpeI site/C/SphI site/Barcode sequence/NotI site/3' priming sequence (CAACTACTACTACAG). To clone the basal promoter into barcoded oligos without any upstream *cis*-regulatory sequence, we placed the SpeI site next to the EcoRI site, which allowed us to place the promoter between the EcoRI site and the 3' barcode.

We cloned the synthesized oligos as previously described by our group (Kwasnieski et al., 2012; White et al., 2016, 2013). Specifically, for each oligo pool we used 50 femtomoles of template and 4 cycles of PCR in each of multiple 50 microliter reactions (NEB Phusion) using primers MO563 and MO564 (**Supplementary Dataset 6**), 2% DMSO, and an annealing temperature of 57C. PCR amplicons were purified from a 2% agarose gel (NEB), digested with EcoRI-HF and NotI-HF (NEB), and then cloned into the EagI and EcoRI sites of the plasmid pJK03 with multiple 20 microliter ligation reactions (NEB T4 ligase). The libraries were transformed into 5-alpha electrocompetent cells (NEB) and grown in liquid culture. Next, 2 micrograms of each library was digested with SphI-HF and SpeI-HF (NEB) and then treated with Antarctic phosphatase (NEB).

The *Rho* basal promoter and *DsRed* reporter gene was amplified from the plasmid pJK01 using primers MO566 and MO567 (**Supplementary Dataset 6**). The Polylinker and *DsRed* reporter gene was amplified from the plasmid pJK03 using primers MO610 and MO567 (**Supplementary Dataset 6**). The Polylinker is a short 23 bp multiple cloning site with no known core promoter motifs. Inserts were purified from a 1% agarose gel (NEB), digested with NheI-HF and SphI-HF (NEB), and cloned into the libraries using multiple 20 microliter ligations (NEB T4 ligase). The libraries were transformed into 5-alpha electrocompetent cells (NEB) and grown in liquid culture.

Retinal Explant Electroporation

Animal procedures were performed in accordance with a Washington University in St. Louis Institutional Animal Care and Use Committee approved vertebrate animals protocol. Electroporation into retinal explants and RNA extraction was performed as described previously (Hsiau et al., 2007; Hughes et al., 2018; Kwasnieski et al., 2012; White et al., 2016, 2013).

Briefly, retinas were isolated from P0 newborn CD-1 mice and electroporated in a solution with 30 micrograms library and 30 micrograms *Rho*-GFP. Electroporated retinas were cultured for eight days, at which point they were harvested, washed three times with HBSS (Gibco), and stored in TRIzol (Invitrogen) at -80C. Five retinas were pooled for each biological replicate and three replicates were performed for each library. RNA was extracted from TRIzol according to manufacturer instructions and treated with TURBO DNase (Invitrogen). cDNA was prepared using SuperScript RT III (Invitrogen) with oligo dT primers. Barcodes from both the cDNA and the plasmid DNA pool were amplified for sequencing (described below). The resulting products were mixed at equal concentration and sequenced on the Illumina NextSeq platform. We obtained greater than 1300x coverage across all samples.

Rho libraries were amplified using primers MO574 and MO575 (**Supplementary Dataset 6**) for 6 cycles at an annealing temperature of 66C followed by 18 cycles with no annealing step (NEB Phusion) and then purified with the Monarch PCR kit (NEB). PCR amplicons were digested using MfeI-HF and SphI-HF (NEB) and ligated to custom annealed adaptors with PE2 indexing barcodes and phased P1 barcodes (**Supplementary Dataset 6**). The final enrichment PCR used primers MO588 and MO589 (**Supplementary Dataset 6**) for 20 cycles at an annealing temperature of 66C (NEB Phusion), followed by purification with the Monarch PCR kit. Polylinker libraries were amplified using primers BC_CRX_Nested_F and BC_CRX_R (**Supplementary Dataset 6**) for 30 cycles (NEB Q5) at an annealing temperature of 67C and then purified with the Monarch PCR kit. Illumina adaptors were then added via two further rounds of PCR. First, P1 indexing barcodes were added using forward primers P1_inner_A through P1_inner_D and reverse primer P1_inner_nested_rev (**Supplementary Dataset 6**) for 5 cycles at an annealing temperature of 55C followed by 5 cycles with no annealing step (NEB Q5). PE2 indexing barcodes were then added by amplifying 2 microliters of the previous reaction with forward primer P1_outer and reverse primers PE2_outer_SIC69 and PE2_outer_SIC70 (**Supplementary Dataset 6**) for 5 cycles at an annealing temperature of 66C followed by 5 cycles with no annealing step (NEB Q5) and then purified with the Monarch PCR kit.

Data Processing

All data processing, statistical analysis, and downstream analyses were performed in Python version 3.6.5 using Numpy version 1.15.4 (Harris et al., 2020), Scipy version 1.1.0 (Virtanen et al., 2020), and Pandas version 0.23.4 (McKinney, 2010); and visualized using Matplotlib version 3.0.2 (Hunter, 2007) and Logomaker version 0.8 (Tareen and Kinney, 2020). All statistical analysis used two-sided tests unless noted otherwise.

Sequencing reads were filtered to ensure that the barcode sequence perfectly matched the expected sequence (>93% reads in a sample for the *Rho* libraries, >86% reads for the Polylinker libraries). For the *Rho* libraries, barcodes that had less than 10 raw counts in the DNA sample were considered missing and removed from downstream analysis. Barcodes that had less than 5 raw counts in any cDNA sample were considered present in the input plasmid pool but below the detection limit and thus set to zero in all samples. Barcode counts were normalized by reads per million (RPM) for each sample. Barcode expression was calculated by

dividing the cDNA RPM by the DNA RPM. Replicate-specific expression was calculated by averaging the barcodes corresponding to each library sequence. After performing statistical analysis (see below), expression levels were normalized by replicate-specific basal mean expression and then averaged across biological replicates.

For the Polylinker assay, the expected lack of expression of many constructs required different processing. Barcodes that had less than 50 raw counts in the DNA sample were removed from downstream analysis. Barcodes were normalized by RPM for each replicate. Barcodes that had less than 8 RPM in any cDNA sample were set to zero in all samples. cDNA RPM were then divided by DNA RPM as above. Within each biological replicate, barcodes were averaged as above but were not normalized to basal expression because there is no basal construct. Expression values were then averaged across biological replicates. Due to the low expression of scrambled sequences and the lack of a basal construct, we were unable to assess data calibration with the same rigor as above.

Assignment of Activity Classes

Activity classes were assigned by comparing expression levels to basal promoter expression levels across replicates. The null hypothesis is that the expression of a sequence is the same as basal levels. Expression levels were approximately log-normally distributed, so we computed the log-normal parameters for each sequence and then performed Welch's t-test. We corrected for multiple hypotheses using the Benjamini-Hochberg False Discovery Rate (FDR) procedure. We corrected for multiple hypotheses in each library separately to account for any potential batch effects between libraries. The $\log_2$ expression was calculated after adding a pseudocount of 1×10^{-3} to every sequence.

Sequences were classified as enhancers if they were 2-fold above basal and the q-value was below 0.05. Silencers were similarly defined as 2-fold below basal and q-value less than 0.05. Inactive sequences were defined as within a 2-fold change and q-value greater than or equal to 0.05. All other sequences were classified as ambiguous and removed from further analysis. We used scrambled sequences to further stratify enhancers into strong and weak enhancers, using the rationale that scrambled sequences give an empirical distribution for the activity of random sequences. We defined strong enhancers as enhancers that are above the 95th percentile of scrambled sequences and all other enhancers as weak enhancers.

For the Polylinker assay, we did not have a basal construct as a reference point. Instead, we defined a sequence to have autonomous activity if the average cDNA barcode counts were higher than average DNA barcode counts, and non-autonomous otherwise. The $\log_2$ expression was calculated after adding a pseudocount of 1×10^{-2} to every sequence.

Motif Analysis

We performed motif enrichment analysis using the MEME Suite version 5.0.4 (Bailey et al., 2009). We searched for motifs that were enriched in one group of sequences relative to another group using DREME-py3 with the parameters -mink 6 -maxk 12 -e 0.05 and compared the de novo motifs to known motifs using TOMTOM on default parameters. We ran DREME using

strong enhancers as positives and silencers and negatives, and vice versa. For TOMTOM, we used version 11 of the full mouse HOCOMOCO database (Kulakovskiy et al., 2018) with the following additions from the JASPAR human database (Khan et al., 2018): NRL (accession MA0842.1), RORB (accession MA1150.1), and RAX (accession MA0718.1). We added these PWMs because they have known roles in the retina, but the mouse PWMs were not in the HOCOMOCO database. We also used the CRX PWM that we used to design the library. Motifs were selected for downstream analysis based on their matches to the de novo motifs, whether the TF had a known role in retinal development, and the quality of the PWM. We used one TF for each DREME motif. We did not use any PWMs with a quality of "D". We excluded DREME motifs without a match to the database from further analysis; most of these resemble dinucleotides.

Predicted Occupancy

We computed predicted occupancy as previously described (White et al., 2013; Zhao et al., 2009). Briefly, we normalized each letter probability matrix by the most probable letter at each position. We took the negative log of this matrix and multiplied by 2.5, which corresponds to the ideal gas constant times 300 Kelvin, to obtain an energy weight matrix. We used a chemical potential μ of 9 for all TFs. At this value, the probability of a site being bound is at least 0.5 if the relative K_D is at least 0.03 of the optimal binding site. We computed the predicted occupancy for every site on the forward and reverse strands and summed them together to get a single value for each TF.

Information Content

To capture the effects of TF predicted occupancy and diversity in a single metric, we calculated the sequence information content using Boltzmann entropy. Boltzmann's equation states that the entropy of a system S is related to the number of ways the molecules can be arranged (microstates) W via the equation $S = k_B \log W$, where k_B is Boltzmann's constant (Phillips et al., 2012, chap. 5). The number of microstates is defined as $W = \frac{N!}{\prod_i N_i!}$ where N is the total number of particles and N_i are the number of the i-th type of particles. In our case, the system is the collection of predicted binding motifs for different TFs in a *cis*-regulatory sequence. We assume each TF is a different type of molecule because the DNA binding domain of each TF belongs to a different subfamily. The number of molecular arrangements W represents the number of distinguishable ways that the TFs can be ordered on the sequence. Thus, N_i is the predicted occupancy of the i-th TF and N is the total predicted occupancy of all TFs on the *cis*-regulatory sequence. Because the predicted occupancies are continuous values, we exploit the definition of the Gamma function, $\Gamma(N+1) = N!$, to rewrite $W = \frac{\Gamma(N+1)}{\prod_i \Gamma(N_i+1)}$.

If we assume that each arrangement of motifs is equally likely, then we can write the probability of arrangement $w = 1, \ldots, W$ as $p_w = \frac{1}{W}$ and rewrite the entropy as $S = -\log(\frac{1}{W}) = -\log(p_w)$, where we have dropped Boltzmann's constant since the connection between molecular arrangements and temperature is not important. Because each arrangement is equally likely, then $\frac{1}{W}$ is also the expected value of p_w and we can write the entropy as $S = -E[\log(p_w)] = -\sum_w p_w \log(p_w)$, which is Shannon entropy. By definition, Shannon entropy

is also the expected value of the information content: $E[I] = -\sum_w p_w \log(p_w) = \sum_w p_w I(w)$ where the information content I of a particular state is $I(w) = -\log(p_w)$. Since we assumed each arrangement is equally likely, then the expected value of the information content is also the information content of each arrangement. Therefore, the information content of a *cis*-regulatory sequence can be written as $I = -\log(p_w) = \log W$. We use log base 2 to express the information content in bits.

With this metric, *cis*-regulatory sequences with higher predicted TF occupancies generally have higher information content. Sequences with higher TF diversity have higher information content than lower diversity sequences with the same predicted occupancy. Thus our metric captures the effects of both TF diversity and total TF occupancy. For example, consider hypothetical TFs A, B, and C. If motifs for only one TF are in a sequence, then W is always 1 and the information content is always zero (regardless of total occupancy). The simplest case for non-zero information content is one motif for A, one motif for B, and zero motifs for C (1-1-0). Then $W = \frac{2!}{1!1!} = 2$ and $I = 1$ bit. If we increase predicted occupancy by adding a motif for A (2-1-0), then $W = \frac{3!}{2!1!} = 3$ and $I = 1.6$ bits, which is approximately the information content of silencers and inactive sequences. If we increase predicted occupancy again and add a second motif for B (2-2-0), then $W = \frac{4!}{2!2!} = 6$ and $I = 2.6$ bits, which is approximately the information content of strong enhancers. If instead of increasing predicted occupancy, we instead increase diversity by replacing a motif for A with a motif for C (1-1-1), then $W = \frac{3!}{1!1!1!} = 6$ and once again $I = 2.6$ bits, which is higher than the lower diversity case (2-1-0).

According to Wunderlich and Mirny (Wunderlich and Mirny, 2009), the probability of observing k total motifs for m different TFs in a w bp window is $P(k) \sim Poisson(k; \lambda)$, where $\lambda = pmw$ and p is the probability of finding a spurious motif in the genome. The expected number of windows with k total motifs in a genome of length N is thus $E(k) = P(k) \cdot N$. In mammals, $N \approx 10^9$ and Wunderlich and Mirny find that $p = 0.00025$ for multicellular eukaryotes. For $m = 3$ TFs and a $w = 164$ bp window (which is the size of our sequences), $\lambda = 0.123$ and $E(5) = 1.6$, meaning that five total motifs for three different TFs specifies an approximately unique 164 bp location in a mammalian genome. Five total motifs for three different TFs can be achieved in two ways: three motifs for A, one for B, and one for C (3-1-1), or two motifs for A, two for B, and one for C (2-2-1). In the case of 3-1-1, $W = \frac{5!}{3!1!1!} = 20$ and $I = 4.3$ bits. In the case of 2-2-1, $W = \frac{5!}{2!2!1!} = 30$ and $I = 4.9$ bits.

Machine Learning

The *k*-mer SVM was fit using gkmSVM (Ghandi et al., 2014). All other machine learning, including cross-validation, logistic regression, and computing ROC and PR curves, was performed using scikit-learn version 0.19.1 (Pedregosa et al., 2011). We wrote custom Python wrappers for gkmSVM to allow for interfacing between the C++ binaries and the rest of our workflow. We ran gkmSVM with the parameters -l 6 -k 6 -m 1. To estimate model performance, all models were fit with stratified five-fold cross-validation after shuffling the order of sequences. For the TF occupancy logistic regression model, we used L2 regularization. We selected the

regularization parameter C by performing grid search with five-fold cross-validation on the values 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}, 10^{4} and selecting the value that maximized the F1 score. The optimal value of C was 0.01, which we used as the regularization strength when assessing the performance of the model with other feature sets.

To assess the performance of the logistic regression model, we randomly sampled 8 PWMs from the HOCOMOCO database and computed the predicted occupancy of each TF on each sequence. We then fit a new logistic regression model with these features and repeated this procedure 100 times to generate a background distribution of model performances.

To generate de novo motifs from the SVM, we generated all 6-mers and scored them against the SVM. We then ran the svmw_emalign.py script from gkmSVM on the *k*-mer scores with the parameters -n 10 -f 2 -m 4 and a PWM length of 6, and then used TOMTOM to compare them to the database from our motif analysis.

Other Data Sources

We used our previously published library (White et al., 2013) as an independent test set for our machine learning models. We defined strong enhancers as ChIP-seq peaks that were above the 95th percentile of all scrambled sequences. There was no basal promoter construct in this library, so instead we defined silencers as ChIP-seq peaks that were at least two-fold below the $\log_2$ mean of all scrambled sequences.

Previously published ChIP-seq data for NRL (Hao et al., 2012) that was re-processed by Hughes et al. (Hughes et al., 2017) and MEF2D (Andzelm et al., 2015) was used to annotate sequences for *in vivo* TF binding. We converted peaks to mm10 coordinates using the UCSC liftOver tool and then used Bedtools to intersect peaks with our library.

Code and Data Availability

The pJK01 and pJK03 plasmids will be deposited with AddGene. Raw sequencing data and barcode counts have been uploaded to the NCBI GEO database under accession GSE165812. All processed activity data, predicted occupancy, and information content values are available in the supplementary material. All code for data processing, analysis, and visualization is available on Github at https://github.com/barakcohenlab/CRX-Information-Content.

Acknowledgements

We thank Gary Stormo and members of the Cohen Lab for critically reading the manuscript and helpful discussions; Philip A. Ruzycki, Andrew E.O. Hughes, and Timothy J. Cherry for providing processed ChIP-seq data; and Jessica Hoisington-Lopez and MariaLynn Crosby from the DNA Sequencing Innovation Lab for assistance with high-throughput sequencing. RZF was partially supported by a Kirschstein National Research Service Award from the National Institutes for Health, F31HG011431. This work was supported by grants from the National Institutes for Health, R01GM121755 (to MAW), R01EY027784 (to BAC), and R01 EY025196 and R01 EY03075 (to JCC).

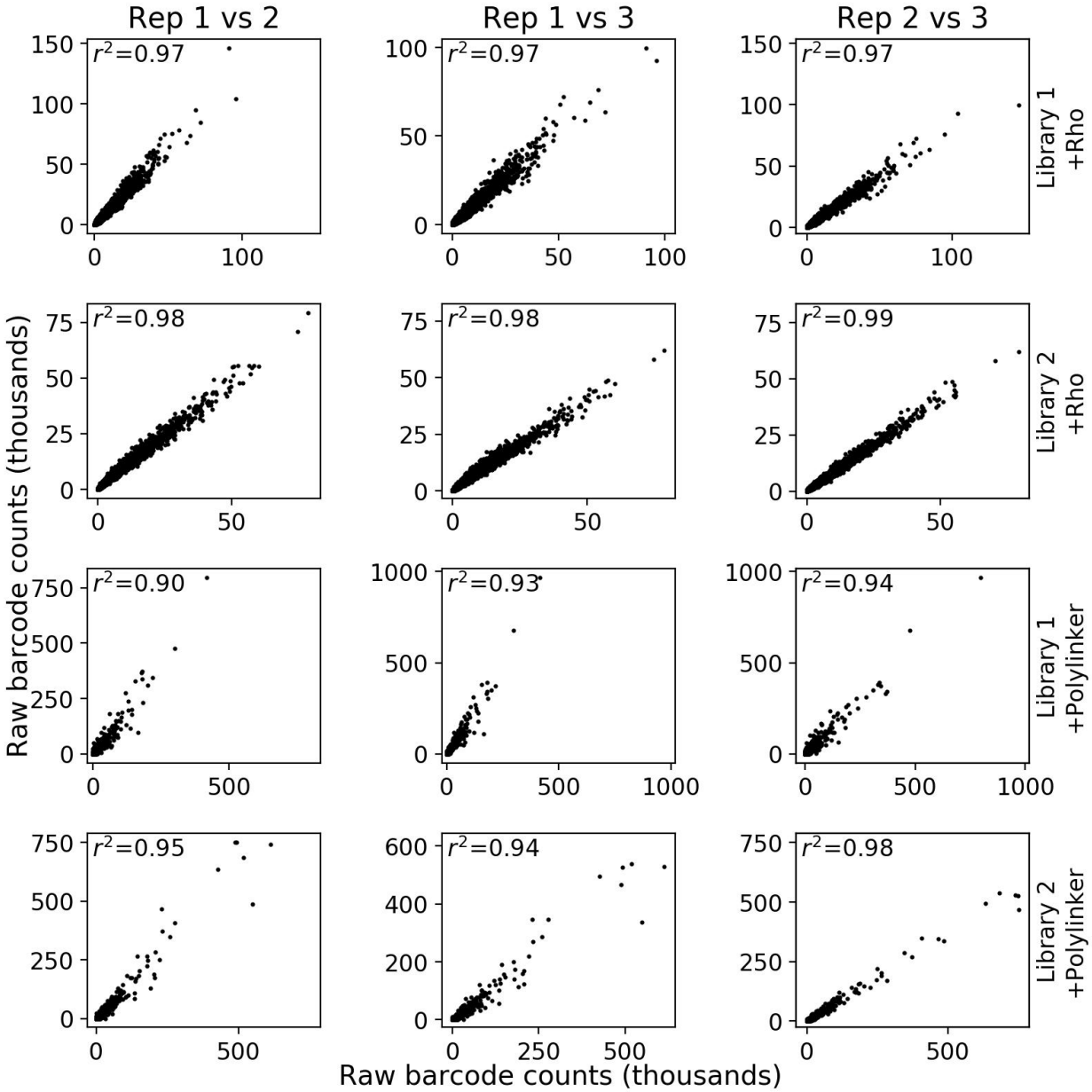
Competing interests

The authors declare no competing interests.

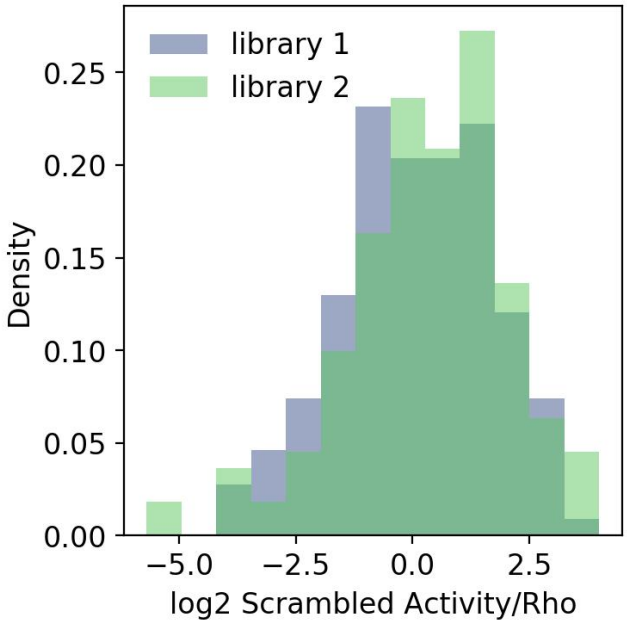
Author Contributions

RZF, BAC, and MAW designed the overall research project. JCC, BAC, and MAW supervised the research. RZF and DMG cloned MPRA libraries. CAM performed electroporations and cultured retina. RZF performed all analysis. RZF, BAC, and MAW developed the information content metric. RZF, JCC, BAC, and MAW wrote the manuscript.

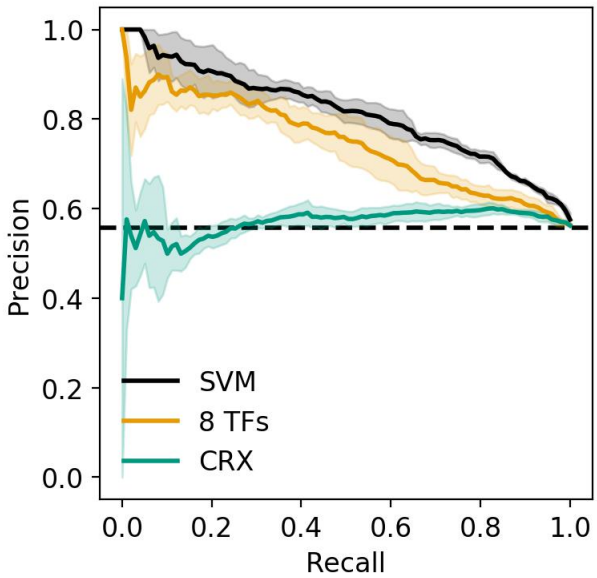
Supplementary Figures



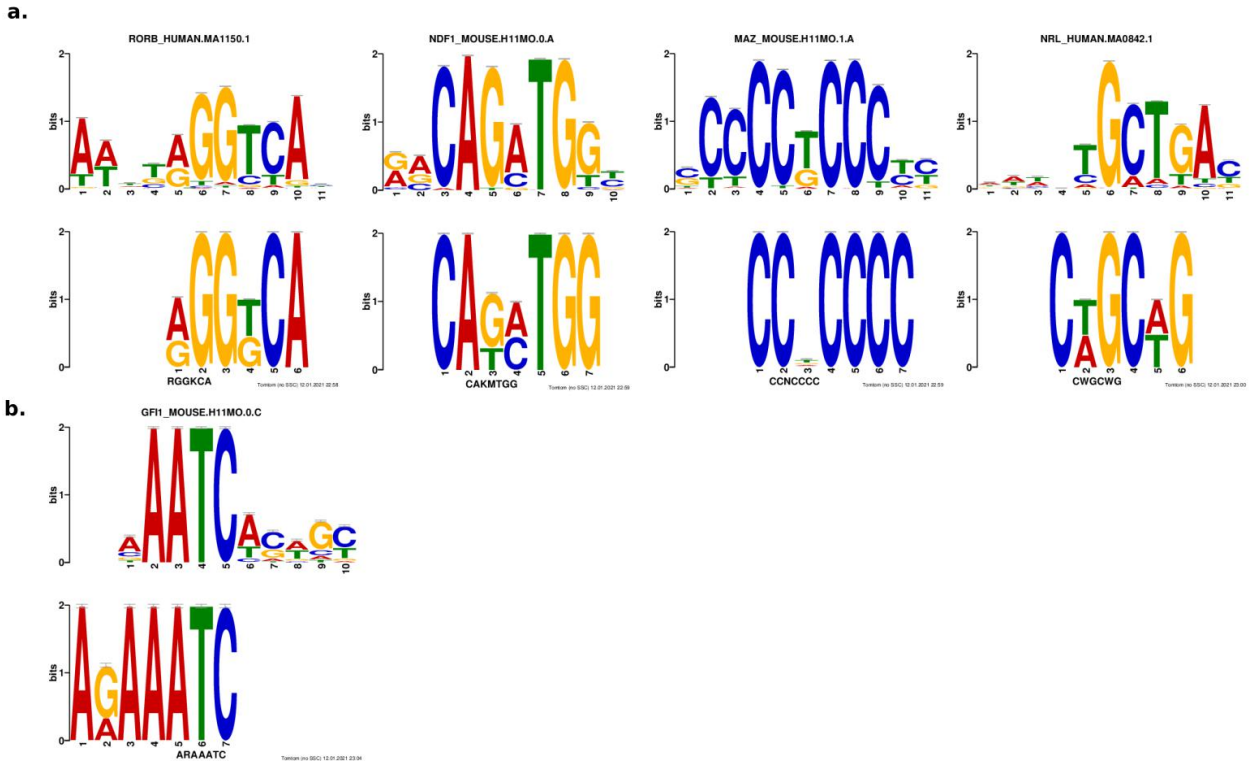
Supplementary Figure 1: Reproducibility of MPRA measurements. Each row represents a different library and experiment. For each column, the first replicate in the title is the x-axis and the second replicate is the y-axis.



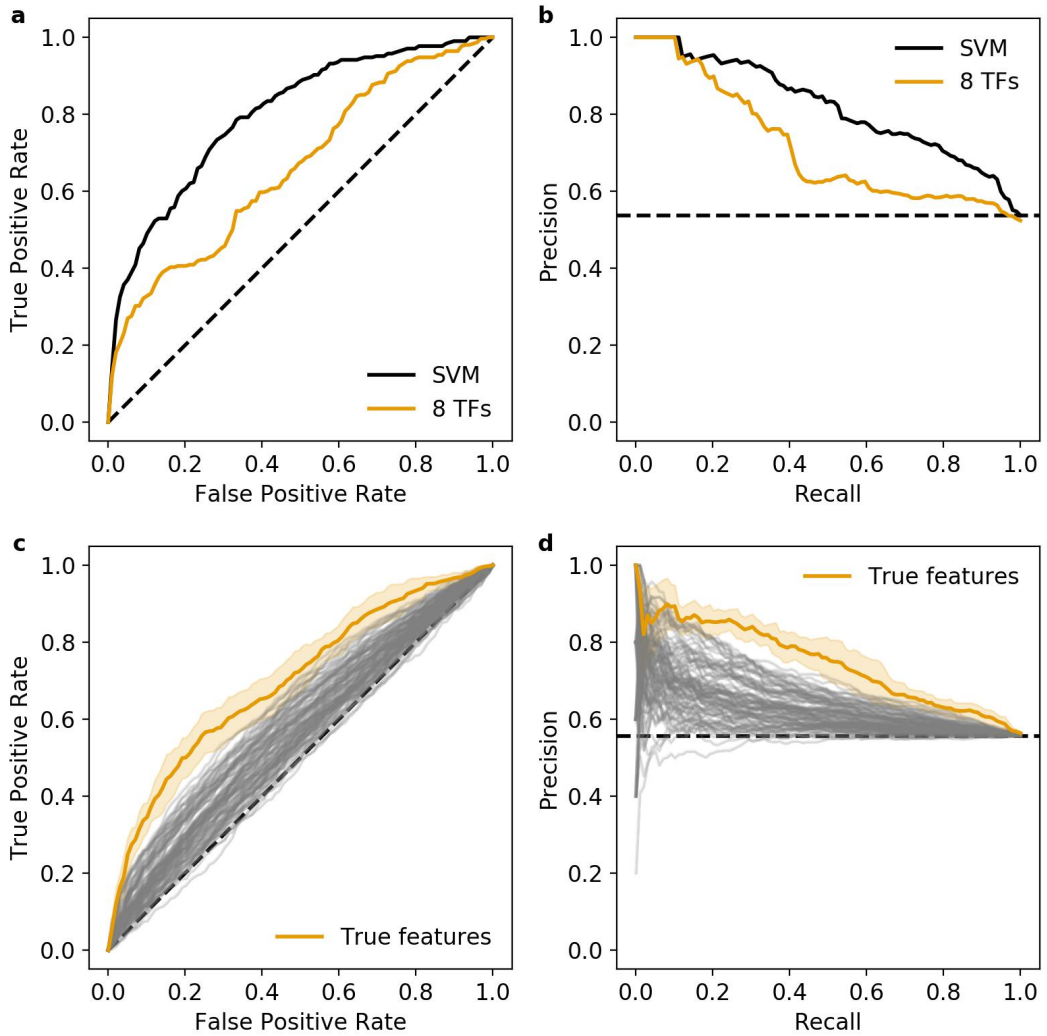
Supplementary Figure 2: Calibration of MPRA libraries with the *Rho* promoter. Probability density histogram of the same 150 scrambled sequences in two libraries after normalizing to the basal *Rho* promoter.



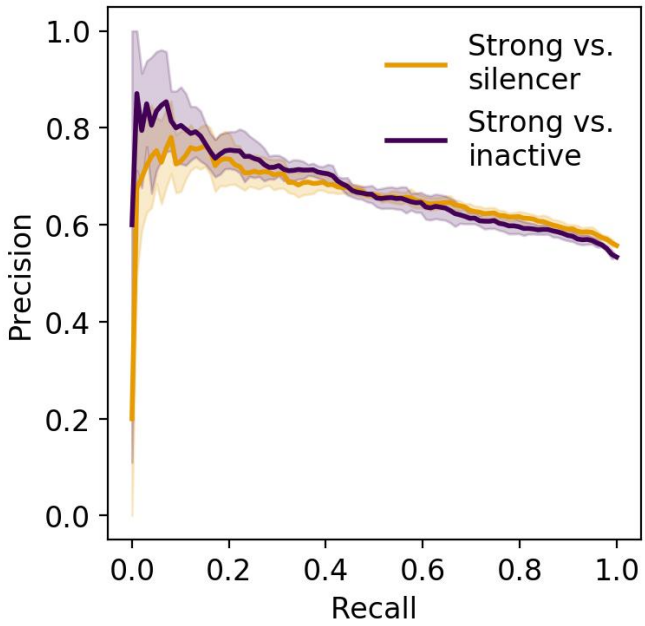
Supplementary Figure 3: Precision recall curve for strong enhancer vs. silencer classifiers. Solid black, 6-mer SVM; orange, 8 TF predicted occupancy logistic regression; aqua, predicted CRX occupancy logistic regression; dashed black, chance; shaded area, 1 standard deviation based on five-fold cross-validation.



Supplementary Figure 4: Results from *de novo* motif analysis. Motifs enriched in strong enhancers **(a)** and silencers **(b)**. Bottom, *de novo* motif identified with DREME; top, matched known motif identified with TOMTOM.



Supplementary Figure 5: Additional validation of the 8 TF predicted occupancy logistic regression model. a and b) Predictions of the 6-mer SVM (black) and 8 TF predicted occupancy logistic regression model (orange) on an independent test set. **c and d)** Null distribution of 100 logistic regression models trained using randomly selected motifs (grey) compared to the true features (orange). Shaded area, 1 standard deviation based on five-fold cross-validation. **a and c)** Receiver operating characteristic, **b and d)** precision recall curve. Dashed black line represents chance in all panels.



Supplementary Figure 6: Precision recall curve of logistic regression classifier using information content. Orange, strong enhancer vs. silencer; indigo, strong enhancer vs. inactive; shaded area, 1 standard deviation based on five-fold cross-validation.

Supplementary Datasets

Dataset 1: FASTA file of all sequences in library 1. Sequences were named using the following nomenclature: "chrom-start-stop_annotations_variant". "Chrom", "start", and "stop" correspond to the mm10 genomic coordinates of the sequences in BED format. "Annotations" is a four letter string where the first position indicates CRX binding status (**C**hIP-seq peak or **U**nbound), the second position indicates CRX motif status (**P**WM hit, **S**hape motif, or **B**oth PWM and shape motif), the third position indicates ATAC-seq status (peak in **R**ods but not cones, peak in **C**ones but not rods, peak in both rod and cone **P**hotoreceptors, or peak in **N**one of the above), and the fourth position indicates histone ChIP-seq status ("**E**nhancer marked" with H3K27Ac^{+}H3K4me3^{-}, "**P**romoter marked" with H3K27Ac^{+}H3K4me3^{+}, **Q** for H3K27Ac^{-}H3K4me3^{+}, or **N**either mark). "Variant" indicates whether the sequence is genomic ("WT"), mutated CRX motifs ("MUT-allCrxSites"), scrambled shape motif ("MUT-shape"), or a scrambled control ("scrambled").

Dataset 2: FASTA file of all sequences in library 2. Sequences were named as in **Dataset 1**.

Dataset 3: Expression measurements and annotations of all sequences. Values are tab-delimited. Rows are named based on the sequence name from **Datasets 1 and 2** without the "variant" information. Columns ending in "_WT" indicate the wild-type sequence with the *Rho* promoter, "_MUT" as the CRX motif mutant sequence with the *Rho* promoter, and "_POLY" as the wild-type sequence with the Polylinker. Sequences with the scrambled shape motif were excluded from the "_MUT" columns. Columns are named as follows: label, the sequence name from **Dataset 1 and 2** without the "variant" information; expression, average activity of the

sequence, NaN indicates sequence was missing from the plasmid pool; expression_std, standard deviation of activity; expression_reps, number of replicates in which the sequence was measured; expression_pvalue, p-value from Welch's t-test of log-normal data for the null hypothesis that the activity of the sequence with *Rho* is no different than the *Rho* promoter alone; expression_qvalue, FDR-correction of the p-values; library, which library contains the sequence; expression_log2, log2 average activity of the sequence; group_name, activity classification of the sequence with the *Rho* promoter; plot_color, hex code for visualization; variant, the "variant" portion of the sequence identifier; wt_vs_mut_log2, log2 fold change between the wild-type and mutant version of the sequence, NaN indicates the wild-type and/or mutant version was not measured; wt_vs_mut_pvalue, p-value from Welch's t-test for the null hypothesis that the wild-type and mutant sequences have the same activity; wt_vs_mut_qvalue, FDR-correction of the p-values; autonomous_activity, boolean value for if the wild-type sequence is autonomous with the Polylinker; crx_bound, nrl_bound, and mef2d_bound, boolean values for if the sequence overlaps a ChIP-seq peak for the corresponding TF; binding_group, string denoting each of the eight possible combinations of CRX, NRL, and MEF2D binding.

Dataset 4: Predicted occupancy scores for each TF and each sequence. Values are tab-delimited. Rows are named based on the sequence name from **Datasets 1 and 2** including the "variant" information. Columns are the predicted occupancy scores for the denoted TF.

Dataset 5: Information content and related metrics for each sequence. Values are tab-delimited. Rows are named based on the sequence name from **Datasets 1 and 2**, including the "variant" information. Columns are named as follows: total_occupancy, total predicted occupancy of all 8 TFs; diversity, number of TFs with predicted occupancy above 0.5; entropy, information content (which is also entropy).

Dataset 6: Primers used in this study.

References

Alexandre, C., Vincent, J.-P., 2003. Requirements for transcriptional repression and activation by Engrailed in Drosophila embryos. Development 130, 729–739.

Andzelm, M.M., Cherry, T.J., Harmin, D.A., Boeke, A.C., Lee, C., Hemberg, M., Pawlyk, B., Malik, A.N., Flavell, S.W., Sandberg, M.A., Raviola, E., Greenberg, M.E., 2015. MEF2D drives photoreceptor development through a genome-wide competition for tissue-specific enhancers. Neuron 86, 247–263.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S., 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37, W202–8.

Barolo, S., Posakony, J.W., 2002. Three habits of highly effective signaling pathways: Principles of transcriptional control by developmental cell signaling. Genes and Development 16, 1167–1181.

Chen, S., Wang, Q.-L., Nie, Z., Sun, H., Lennon, G., Copeland, N.G., Gilbert, D.J., Jenkins, N.A., Zack, D.J., 1997. Crx, a Novel Otx-like Paired-Homeodomain Protein, Binds to and Transactivates Photoreceptor Cell-Specific Genes. Neuron 19, 1017–1030.

Chiang, C., Ayyanathan, K., 2013. Snail/Gfi-1 (SNAG) family zinc finger proteins in transcription

regulation, chromatin dynamics, cell signaling, development, and disease. Cytokine Growth Factor Rev. 24, 123–131.
Clyde, D.E., Corado, M.S.G., Wu, X., Paré, A., Papatsenko, D., Small, S., 2003. A self-organizing system of repressor gradients establishes segmental complexity in Drosophila. Nature 426, 849–853.
Corbo, J.C., Lawrence, K.A., Karlstetter, M., Myers, C.A., Abdelaziz, M., Dirkes, W., Weigelt, K., Seifert, M., Benes, V., Fritsche, L.G., Weber, B.H.F., Langmann, T., 2010. CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. Genome Res. 20, 1512–1525.
Crocker, J., Abe, N., Rinaldi, L., McGregor, A.P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., Mann, R.S., Stern, D.L., 2015. Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness. Cell 160, 191–203.
Doni Jayavelu, N., Jajodia, A., Mishra, A., Hawkins, R.D., 2020. Candidate silencer elements for the human and mouse genomes. Nat. Commun. 11, 1061.
ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., Kaul, R., Halow, J., Van Nostrand, E.L., Freese, P., Gorkin, D.U., Shen, Y., He, Y., Mackiewicz, M., Pauli-Behn, F., Williams, B.A., Mortazavi, A., Keller, C.A., Zhang, X.-O., Elhajjajy, S.I., Huey, J., Dickel, D.E., Snetkova, V., Wei, X., Wang, X., Rivera-Mulia, J.C., Rozowsky, J., Zhang, J., Chhetri, S.B., Zhang, J., Victorsen, A., White, K.P., Visel, A., Yeo, G.W., Burge, C.B., Lécuyer, E., Gilbert, D.M., Dekker, J., Rinn, J., Mendenhall, E.M., Ecker, J.R., Kellis, M., Klein, R.J., Noble, W.S., Kundaje, A., Guigó, R., Farnham, P.J., Cherry, J.M., Myers, R.M., Ren, B., Graveley, B.R., Gerstein, M.B., Pennacchio, L.A., Snyder, M.P., Bernstein, B.E., Wold, B., Hardison, R.C., Gingeras, T.R., Stamatoyannopoulos, J.A., Weng, Z., 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699–710.
Ernst, J., Kellis, M., 2012. ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods 9, 215–216.
Farley, E.K., Olson, K.M., Zhang, W., Brandt, A.J., Rokhsar, D.S., Levine, M.S., 2015. Suboptimization of developmental enhancers. Science 350, 325–328.
Farley, E.K., Olson, K.M., Zhang, W., Rokhsar, D.S., Levine, M.S., 2016. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. Proc. Natl. Acad. Sci. U. S. A. 113, 6508–6513.
Freund, C.L., Gregory-Evans, C.Y., Furukawa, T., Papaioannou, M., Looser, J., Ploder, L., Bellingham, J., Ng, D., Herbrick, J.A.S., Duncan, A., Scherer, S.W., Tsui, L.C., Loutradis-Anagnostou, A., Jacobson, S.G., Cepko, C.L., Bhattacharya, S.S., McInnes, R.R., 1997. Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor. Cell 91, 543–553.
Furukawa, T., Morrow, E.M., Cepko, C.L., 1997. Crx, a Novel otx-like Homeobox Gene, Shows Photoreceptor-Specific Expression and Regulates Photoreceptor Differentiation. Cell 91, 531–541.
Ghandi, M., Lee, D., Mohammad-Noori, M., Beer, M.A., 2014. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. PLoS Comput. Biol. 10, e1003711.
Gisselbrecht, S.S., Palagi, A., Kurland, J.V., Rogers, J.M., Ozadam, H., Zhan, Y., Dekker, J., Bulyk, M.L., 2020. Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts. Mol. Cell 77, 324–337.
Grossman, S.R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B.E., Mikkelsen, T.S., Lander, E.S., 2017. Systematic dissection of genomic features determining transcription factor binding and enhancer function. Proc. Natl. Acad. Sci. U. S. A. 114, E1291–E1300.
Haeussler, M., Zweig, A.S., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Hinrichs, A.S., Gonzalez, J.N., Gibson, D., Diekhans, M., Clawson, H., Casper, J.,

Barber, G.P., Haussler, D., Kuhn, R.M., Kent, W.J., 2019. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. 47, D853–D858.
Hao, H., Kim, D.S., Klocke, B., Johnson, K.R., Cui, K., Gotoh, N., Zang, C., Gregorski, J., Gieser, L., Peng, W., Fann, Y., Seifert, M., Zhao, K., Swaroop, A., 2012. Transcriptional regulation of rod photoreceptor homeostasis revealed by in vivo NRL targetome analysis. PLoS Genet. 8, e1002649.
Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., Del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362.
Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., Noble, W.S., 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods 9, 473–476.
Hsiau, T.H.-C., Diaconu, C., Myers, C.A., Lee, J., Cepko, C.L., Corbo, J.C., 2007. The Cis-regulatory logic of the mammalian photoreceptor transcriptional network. PLoS One 2, e643.
Hughes, A.E.O., Enright, J.M., Myers, C.A., Shen, S.Q., Corbo, J.C., 2017. Cell Type-Specific Epigenomic Analysis Reveals a Uniquely Closed Chromatin Architecture in Mouse Rod Photoreceptors. Sci. Rep. 7, 43184.
Hughes, A.E.O., Myers, C.A., Corbo, J.C., 2018. A massively parallel reporter assay reveals context-dependent activity of homeodomain binding sites in vivo. Genome Res. 28, 1520–1531.
Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9, 90–95.
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.
Iype, T., Taylor, D.G., Ziesmann, S.M., Garmey, J.C., Watada, H., Mirmira, R.G., 2004. The transcriptional repressor Nkx6.1 also functions as a deoxyribonucleic acid context-dependent transcriptional activator during pancreatic beta-cell differentiation: evidence for feedback activation of the nkx6.1 gene by Nkx6.1. Mol. Endocrinol. 18, 1363–1375.
Jia, L., Oh, E.C.T., Ng, L., Srinivas, M., Brooks, M., Swaroop, A., Forrest, D., 2009. Retinoid-related orphan nuclear receptor RORbeta is an early-acting factor in rod photoreceptor development. Proc. Natl. Acad. Sci. U. S. A. 106, 17534–17539.
Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E.H., Birney, E., Furlong, E.E.M., 2012. A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell 148, 473–486.
Kelley, D.R., Snoek, J., Rinn, J.L., 2016. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 26, 990–999.
Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G., Baranasic, D., Arenillas, D.J., Sandelin, A., Vandepoele, K., Lenhard, B., Ballester, B., Wasserman, W.W., Parcy, F., Mathelier, A., 2018. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 46, D1284.
Klemm, S.L., Shipony, Z., Greenleaf, W.J., 2019. Chromatin accessibility and the regulatory epigenome. Nat. Rev. Genet. 20, 207–220.
Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., Kolpakov, F.A., Makeev, V.J., 2018. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 46, D252–D259.
Kwasnieski, J.C., Fiore, C., Chaudhari, H.G., Cohen, B.A., 2014. High-throughput functional

testing of ENCODE segmentation predictions. Genome Res. 24, 1595–1602.
Kwasnieski, J.C., Mogno, I., Myers, C.A., Corbo, J.C., Cohen, B.A., 2012. Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc. Natl. Acad. Sci. U. S. A. 109, 19498–19503.
Lee, D., Karchin, R., Beer, M.A., 2011. Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res. 21, 2167–2180.
Lee, J., Myers, C.A., Williams, N., Abdelaziz, M., Corbo, J.C., 2010. Quantitative fine-tuning of photoreceptor cis-regulatory elements through affinity modulation of transcription factor binding sites. Gene Ther. 17, 1390–1399.
Lerner, L.E., Peng, G.H., Gribanova, Y.E., Chen, S., Farber, D.B., 2005. Sp4 is expressed in retinal neurons, activates transcription of photoreceptor-specific genes, and synergizes with Crx. J. Biol. Chem. 280, 20642–20650.
Liu, Y.-R., Laghari, Z.A., Novoa, C.A., Hughes, J., Webster, J.R.M., Goodwin, P.E., Wheatley, S.P., Scotting, P.J., 2014. Sox2 acts as a transcriptional repressor in neural stem cells. BMC Neurosci. 15, 95.
Martínez-Montañés, F., Rienzo, A., Poveda-Huertes, D., Pascual-Ahuir, A., Proft, M., 2013. Activator and repressor functions of the Mot3 transcription factor in the osmostress response of Saccharomyces cerevisiae. Eukaryot. Cell 12, 636–647.
McKinney, W., 2010. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. Austin, TX, pp. 51–56.
Mears, A.J., Kondo, M., Swain, P.K., Takada, Y., Bush, R.A., Saunders, T.L., Sieving, P.A., Swaroop, A., 2001. Nrl is required for rod photoreceptor development. Nat. Genet. 29, 447–452.
Mitton, K.P., Swain, P.K., Chen, S., Xu, S., Zack, D.J., Swaroop, A., 2000. The leucine zipper of NRL interacts with the CRX homeodomain. A possible mechanism of transcriptional synergy in rhodopsin regulation. J. Biol. Chem. 275, 29794–29799.
Morrow, E.M., Furukawa, T., Lee, J.E., Cepko, C.L., 1999. NeuroD regulates multiple functions in the developing neural retina in rodent. Development 126, 23–36.
Murphy, D.P., Hughes, A.E., Lawrence, K.A., Myers, C.A., Corbo, J.C., 2019. Cis-regulatory basis of sister cell type divergence in the vertebrate retina. Elife 8, e48216.
Parker, D.S., White, M.A., Ramos, A.I., Cohen, B.A., Barolo, S., 2011. The cis-regulatory logic of Hedgehog gradient responses: key roles for gli binding affinity, competition, and cooperativity. Sci. Signal. 4, ra38.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Others, 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research 12, 2825–2830.
Peng, G.H., Ahmad, O., Ahmad, F., Liu, J., Chen, S., 2005. The photoreceptor-specific nuclear receptor Nr2e3 interacts with Crx and exerts opposing effects on the transcription of rod versus cone genes. Hum. Mol. Genet. 14, 747–764.
Phillips, R., Kondev, J., Theriot, J., Garcia, H., 2012. Physical Biology of the Cell. Garland Science.
Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.
Rachmin, I., Amsalem, E., Golomb, E., Beeri, R., Gilon, D., Fang, P., Nechushtan, H., Kay, G., Guo, M., Yiqing, P.L., Foo, R.S.-Y., Fisher, D.E., Razin, E., Tshori, S., 2015. FHL2 switches MITF from activator to repressor of Erbin expression during cardiac hypertrophy. Int. J. Cardiol. 195, 85–94.
Ruzycki, P.A., Zhang, X., Chen, S., 2018. CRX directs photoreceptor differentiation by accelerating chromatin remodeling at specific target sites. Epigenetics Chromatin 11, 42.
Samee, M.A.H., Bruneau, B.G., Pollard, K.S., 2019. A De Novo Shape Motif Discovery Algorithm Reveals Preferences of Transcription Factors for DNA Shape Beyond Sequence

Motifs. Cell Syst 8, 27–42.e6.
Sethi, A., Gu, M., Gumusgoz, E., Chan, L., Yan, K.-K., Rozowsky, J., Barozzi, I., Afzal, V., Akiyama, J.A., Plajzer-Frick, I., Yan, C., Novak, C.S., Kato, M., Garvin, T.H., Pham, Q., Harrington, A., Mannion, B.J., Lee, E.A., Fukuda-Yuzawa, Y., Visel, A., Dickel, D.E., Yip, K.Y., Sutton, R., Pennacchio, L.A., Gerstein, M., 2020. Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. Nat. Methods 17, 807–814.
Spitz, F., Furlong, E.E.M., 2012. Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. 13, 613–626.
Srinivas, M., Ng, L., Liu, H., Jia, L., Forrest, D., 2006. Activation of the blue opsin gene in cone photoreceptor development by retinoid-related orphan receptor beta. Mol. Endocrinol. 20, 1728–1741.
Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., Stark, A., 2015. Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature 528, 147–151.
Swanson, C.I., Evans, N.C., Barolo, S., 2010. Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. Dev. Cell 18, 359–370.
Tareen, A., Kinney, J.B., 2020. Logomaker: beautiful sequence logos in Python. Bioinformatics 36, 2272–2274.
Truskina, J., Han, J., Chrysanthou, E., Galvan-Ampudia, C.S., Lainé, S., Brunoud, G., Macé, J., Bellows, S., Legrand, J., Bågman, A.-M., Smit, M.E., Smetana, O., Stigliani, A., Porco, S., Bennett, M.J., Mähönen, A.P., Parcy, F., Farcot, E., Roudier, F., Brady, S.M., Bishopp, A., Vernoux, T., 2020. A network of transcriptional repressors modulates auxin responses. Nature.
Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272.
White, M.A., Kwasnieski, J.C., Myers, C.A., Shen, S.Q., Corbo, J.C., Cohen, B.A., 2016. A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors. Cell Rep. 17, 1247–1254.
White, M.A., Myers, C.A., Corbo, J.C., Cohen, B.A., 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.
Wunderlich, Z., Mirny, L.A., 2009. Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet. 25, 434–440.
Yang, Z., Ding, K., Pan, L., Deng, M., Gan, L., 2003. Math5 determines the competence state of retinal ganglion cell progenitors. Dev. Biol. 264, 240–254.
Zhao, Y., Granas, D., Stormo, G.D., 2009. Inferring binding energies from selected binding sites. PLoS Comput. Biol. 5, e1000590.
Zhou, J., Troyanskaya, O.G., 2015. Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods 12, 931–934.