tvst

Article

Machine Learning Prediction of Non-Coding Variant Impact in Human Retinal *cis*-Regulatory Elements

Leah S. VandenBosch[1,*], Kelsey Luu[1,*], Andrew E. Timms[1], Shriya Challam[1], Yue Wu[2], Aaron Y. Lee[2,3], and Timothy J. Cherry[1,3,4]

[1] Center for Developmental Biology and Regenerative Medicine, Seattle Children's Research Institute, Seattle, WA, USA
[2] University of Washington Department of Ophthalmology, Seattle, WA, USA
[3] Brotman Baty Institute for Precision Medicine, Seattle, WA, USA
[4] University of Washington Department of Pediatrics, Seattle, WA, USA

Correspondence: Timothy J. Cherry, Center for Developmental Biology and Regenerative Medicine, Seattle Children's Research Institute, 1900 9th Avenue, Seattle, WA 98101, USA. e-mail: timothy.cherry@seattlechildrens.org

Received: October 21, 2021
Accepted: March 25, 2022
Published: April 18, 2022

Keywords: enhancer; machine learning; variant interpretation; *cis*-regulatory variant; retina

Citation: VandenBosch LS, Luu K, Timms AE, Challam S, Wu Y, Lee AY, Cherry TJ. Machine learning prediction of non-coding variant impact in human retinal *cis*-regulatory elements. Transl Vis Sci Technol. 2022;11(4):16, https://doi.org/10.1167/tvst.11.4.16

Purpose: Prior studies have demonstrated the significance of specific *cis*-regulatory variants in retinal disease; however, determining the functional impact of regulatory variants remains a major challenge. In this study, we utilized a machine learning approach, trained on epigenomic data from the adult human retina, to systematically quantify the predicted impact of *cis*-regulatory variants.

Methods: We used human retinal DNA accessibility data (ATAC-seq) to determine a set of 18.9k high-confidence, putative *cis*-regulatory elements. Eighty percent of these elements were used to train a machine learning model utilizing a gapped *k*-mer support vector machine–based approach. In silico saturation mutagenesis and variant scoring was applied to predict the functional impact of all potential single nucleotide variants within *cis*-regulatory elements. Impact scores were tested in a 20% hold-out dataset and compared to allele population frequency, phylogenetic conservation, transcription factor (TF) binding motifs, and existing massively parallel reporter assay data.

Results: We generated a model that distinguishes between human retinal regulatory elements and negative test sequences with 95% accuracy. Among a hold-out test set of 3.7k human retinal CREs, all possible single nucleotide variants were scored. Variants with negative impact scores correlated with higher phylogenetic conservation of the reference allele, disruption of predicted TF binding motifs, and massively parallel reporter expression.

Conclusions: We demonstrated the utility of human retinal epigenomic data to train a machine learning model for the purpose of predicting the impact of non-coding regulatory sequence variants. Our model accurately scored sequences and predicted putative transcription factor binding motifs. This approach has the potential to expedite the characterization of pathogenic non-coding sequence variants in the context of unexplained retinal disease.

Translational Relevance: This workflow and resulting dataset serve as a promising genomic tool to facilitate the clinical prioritization of functionally disruptive non-coding mutations in the retina.

Introduction

Genetic retinal disorders affect over 2 million individuals worldwide and consist of many classes of disease. Over 260 genes have now been associated with retinal disorders[1,2]; however, as many as half of all cases cannot be explained by variants in protein-coding genes alone.[3] This suggests that risk variants located within the non-coding genome may contribute to retinal disease. The comparatively vast non-coding genome harbors *cis*-regulatory elements (CREs), including promoters, enhancers, silencers, and boundary elements, that play a critical role in

gene expression.[4–7] Genome-wide association studies (GWASs) frequently implicate non-coding regions to disease phenotypes.[6,8–11] Additionally, expression quantitative trait locus (eQTL) analyses have associated non-coding variants with changes in retinal gene expression.[12,13] Moreover, individual case studies have identified causal regulatory variants in retinal disorders, including blue cone monochromacy, non-syndromic congenital retinal non-attachment, and aniridia with foveal hypoplasia.[14–16] However, due to the incomplete characterization of the non-coding genome, as well as the current limitations of some GWAS and eQTL analyses, it remains a challenge to systematically interpret the impact of individual variants within CREs.

CRE function is mediated by complex interactions between transcription factors (TFs) and DNA sequences[17,18] to yield the appropriate transcriptional profile for a given cell type.[19] These interactions can be characterized through assays for DNA accessibility (ATAC-seq and DNase-seq) and protein binding (ChIP-seq, CUT&RUN, and CUT&Tag) to identify and characterize candidate CREs in a given tissue or cell type at a single point in time.[20,21] Moreover, changes in DNA accessibility have been quantitatively associated with non-coding variants in some cell types.[22] Despite recent advancements, it remains challenging to understand the functional significance of genetic variants within CREs in complex tissues without further experimental or integrative computational analyses.[23,24] Identifying and investigating all potential regulatory regions and putative variants is a monumental task that requires painstaking efforts.

Recent developments in artificial intelligence have popularized the use of machine learning for the holistic interpretation of multimodal epigenetic sequencing data.[25,26] Many different approaches have been developed to accurately predict the inferred value of genetic sequences, including non-coding regulatory regions.[27–29] Such approaches have demonstrated promise in select cell lines and tissue types and have been used successfully to integrate epigenomic data in the context of the human retina.[30] This supports the premise of a comprehensive, tissue-specific analysis for CRE variant prioritization in the human retina. Although a number of approaches are available to predict sequences and variant impact, it is important to choose a method that is appropriate for the data used in prediction. For the purposes of training a tissue-specific model to predict impacts on longer non-coding sequences, approaches such as a gapped *k*-mer support vector machine (GKM-SVM) can effectively predict the functional impact of single nucleotide variant impacts within CREs (deltaSVM).[31–33] This GKM-SVM approach has been applied successfully to predict sequence values in the context of specific mouse retinal enhancers.[34–37] However, to date, it has not been applied across a wider set of human retinal epigenomic data to perform a comprehensive prediction of CRE variant impact scores.

In this study, we applied GKM-SVM modeling with variant impact score prediction (deltaSVM) in a high-throughput manner to predict the functional impact of variants in human retinal CRE sequences. We generated adult human retina ATAC-seq data to determine a high-confidence set of 18.9k putative CREs.[38] We then used GKM-SVM to train a model that specifically distinguishes retinal CREs versus genomic background sequences while reserving 20% of candidate CRE sequences as a hold-out dataset for model testing. We then performed in silico saturation mutagenesis on this hold-out dataset to generate a database of all possible single nucleotide variants (SNVs) for 3773 test CREs. We compared these variants to the reference sequence via deltaSVM, generating impact scores for each potential variant. The model revealed that predicted impact scores correlate with allele frequencies in human sequences and with phylogenetic conservation within candidate CREs. Additionally, we observed distinct negative prediction scores when a variant disrupted the core sequence of a known retinal TF binding motif, consistent with a putative deleterious effect. As a further demonstration of functional relevance, this model was able to predict the consequences of sequence variations when compared to a mutational scan of the mouse rhodopsin promoter,[39] showing that the model is robust even across species. Using a larger set of putative retinal CREs, we generated a database of variant impact scores in ocular non-coding sequences (VISIONS) available on the University of California, Santa Cruz (UCSC) genome browser.[40] This analysis could be used to identify non-coding variants with higher disease relevance in the retina and prioritize these alleles for functional follow-up. By addressing this diagnostic gap, we aim to contribute to a more robust elucidation of CRE function in the human retina.

Methods

Input Data Sources

For positive training data, we generated ATAC sequencing datasets from eight biological replicates of adult human retinas as previously reported (Fig. 1B).[38] These data and other related datasets have been assembled in a searchable track hub on the UCSC genome browser

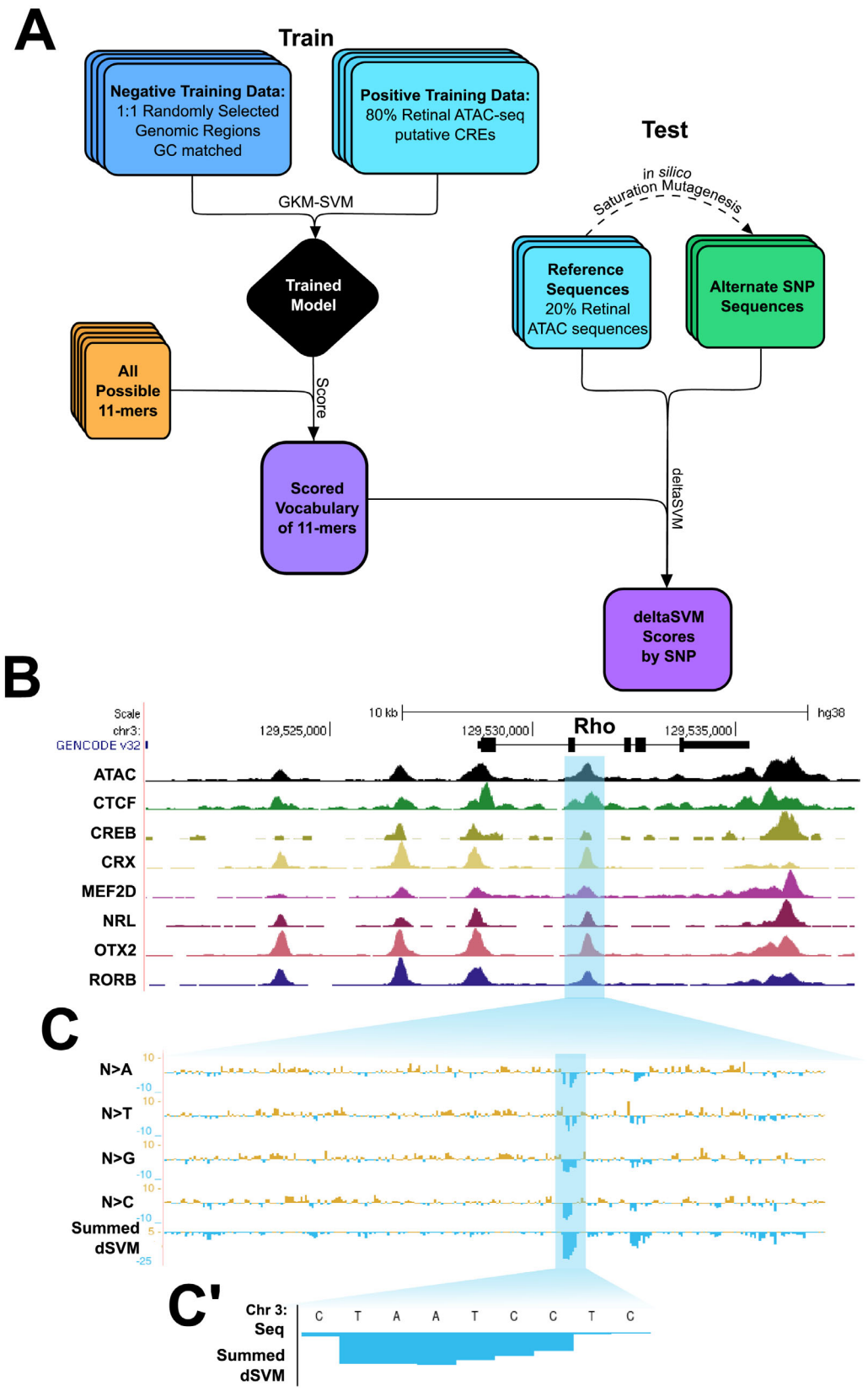


Figure 1. Model overview and training data. (A) Schematic overview of the workflow used in this study to generate a GKM-SVM–based model trained on human retinal ATAC-seq data and randomly selected genomic regions, and deltaSVM variant impact scores through model ranking of in silico saturation mutagenesis of putative retinal CREs. (B) UCSC Genome Browser track positioned at the rhodopsin →

←
(*RHO*) gene visualizing the ATAC-seq dataset used to generate the positive training dataset, schematized in (A), as well as relevant ChIP-seq data for context.[38] One selected region of interest is highlighted in *blue*. (C) Within the highlighted region in (B) are base-pair–resolution deltaSVM variant impact scores, separated by base-pair substitution, and summed negative scores. A region of continuous negative scores is highlighted in *blue*. (C′) In the highlighted region from (C), the summed deltaSVM scores highlight the core TAATC motif of the OTX2 binding site.

(https://tinyurl.com/CherryLab-EyeBrowser).[38] Raw data files were aligned to the hg38 reference genome using Burrows–Wheeler Aligner, and file format conversions were carried out using SAMtools and BEDtools.[41–43] High-confidence peaks across biological replicates of ATAC-seq data were called using the MACS2 algorithm and the ENCODE irreproducible discovery rate (IDR) pipeline with a more stringent *P* value than previously used.[44,45] All eight ATAC replicates were pooled, and two pseudoreplicate files were generated by macs2 randsample with parameter -p 50. Peaks in biological replicate, pooled replicate, and pseudoreplicate samples were called over pooled input samples using MACS2[44] with the parameters –nomodel -g hs -p 1e-2 –extsize 200. This workflow generated a set of 18,866 summits of accessible regions by ATAC-seq. Summits were all extended ±150 bp to generate a set of 18,866 putative CRE regions. For the purposes of training and validating the primary model, 80% of peaks were randomly selected for training, and the remaining 20% were used as hold-out data to test the model (Supplementary Table S1).

For comparisons to non-retinal data, published data from the Gene Expression Omnibus database were used, including ATAC-seq datasets for retinal pigmented epithelium (RPE),[46] primary visual cortex (PVC),[47] and lung fibroblasts.[48] Raw data files were processed as with retinal data, and peaks were called with the same parameters using the MACS2 algorithm and the IDR workflow, and summits were extended ±150 bp. For comparisons to retinal data, all regions overlapping with the retinal ATAC peaks were removed using bedtools intersect. Comparisons of deltaSVM scores to reporter assay expression were made relative to mouse expression data of saturation mutagenesis in RhoCRE3 from Kwasnieski et al.[39]

SVM Model Training and Validation

To train an SVM model in a biologically meaningful way, the positive training data described above must be compared to an appropriate negative training set. To generate a negative training dataset, 1,000,000 regions were randomly selected from the hg38 genome and extended to 301 bp. These regions were filtered against our positive training data, using bedtools intersect to eliminate any overlapping sequences (Supplementary Fig. S1). From here, a negative training set of 301-bp sequences that do not overlap the positive training regions were randomly chosen and GC-matched to the positive set using oPOSSUM.[49] After selecting the training regions, genomic coordinates were converted to fasta format with bedtools and used to train a model using LS-GKM gkmtrain, developed by Lee, Beer, and colleagues.[50] The SVM was trained with the following gkmtrain hyperparameters: L = 11, k = 7, d = 3, C = 1, t = 2, and e = 0.005 (adopted from Shigaki et al.[33]) (Fig. 1A).

To validate the classification of the model, the training data were used in a fivefold cross-validation using the gkmtrain -x 5 -L 11 -k 7 -d 3 -C 1 -t 2 -e 0.005 to generate performance prediction scores of all regions. Model accuracy was visualized using the data in receiver operator characteristic (ROC) and precision-recall curve graphs as calculated using the ROCR package.[51] To assess the parameters and results of this primary model, an additional control model was trained with the same parameters but using training data that were randomly shuffled between positive and negative datasets. To compare genomic region performance in the two models, retinal and non-retinal genomic peaks were scored using LS-GKM gkmpredict. Because GKM-SVM scores in many samples were non-normally distributed, significant differences between retinal and non-retinal data were scored by Kruskal–Wallis χ^2 tests and by pairwise Wilcoxon rank-sum tests for individual comparisons.

Vocabulary and Sequence Scoring

To build a regulatory sequence vocabulary, all possible 2,097,152 non-redundant 11-bp sequences (11-mers) were generated using nrkmers.py from LS-GKM and scored by the trained SVM model using gkmpredict. To validate the biological relevance of the vocabulary scores, scores were sorted by gkmpredict score value. The top 1% of scored 11-mers were subset for validation and validated against known TF binding motifs.

Variant Impact Scoring on In Silico Saturation Mutagenesis

With the previously defined 20% holdout set of 3773 regions, we scored all putative SNVs. To simulate a deep mutational scan, we conducted in silico saturation mutagenesis with a custom-made script, yielding an exhaustive set of 4,542,692 computationally generated sequences that each contained exactly one regulatory SNV. The deltaSVM.pl script was used to quantitatively assess these variant sequences relative to the consensus allele by referencing the regulatory sequence vocabulary, allowing for the calculation of variant impact prediction scores at a single base resolution for the original and shuffled control models.[50]

deltaSVM Variant Impact Validation

Allele Frequency

To assess the biological relevance of deltaSVM scores, first, scores were correlated against human population allele counts for corresponding SNVs. For exact allele correlation, deltaSVM SNV scores were converted to vcf format, and corresponding SNV frequencies in the Genome Aggregation Database (gnomAD v3) were identified via bcftools isec.[52,53] For the comparison of score trends across the 301-bp CRE window, bedtools was used to compute base-wise summary metrics for variant impact scores by summing allele counts per base and negative deltaSVM scores per base. These summary metrics were averaged across CREs to map out the corresponding positional profiles.

Phylogenetic *P* Value Conservation Scores

Next, deltaSVM scores were compared to phylogenetic *P* values (phyloP) of conservation from the PHAST package.[54] The phyloP values across 20 mammalian species were downloaded from the UCSC genome browser. SNVs were binned into representative groups of 2000 alleles from the top, middle, and bottom-most deltaSVM scores for plotting of conservation. Because phyloP scores in deltaSVM bins were non-normally distributed, significant differences between bins was scored by Kruskal–Wallis χ^2 tests and by pairwise Wilcoxon rank-sum tests for individual comparisons.

Transcription Factor Motif Analysis

Positive training data were scored for TF motif enrichment using HOMER findMotifsGenome.pl and findMotifs.pl.[55] Common retinal motifs were selected from the known motif results for analyses of model relevancy. For the scoring of motif prevalence in distinct sequences rather than overall enrichment in a set, sequences were scored against the Homo Sapiens Comprehensive Model Collection (HOCOMOCO) v11 Core database.[56] Motif prevalence in vocabulary and deltaSVM bins against the HOCOMOCO v11 database were scored using Find Individual Motif Occurrences (FIMO) from the MEME suite of tools with a significance threshold of $P \leq 1 \times 10^{-4}$.[57] Significant changes in average deltaSVM within motifs by base pair was scored by analysis of variance (ANOVA) with post hoc Tukey's test.

For the validation of motif interference in deltaSVM scores, known motif positions were obtained from HOMER, and positions were extended by 25 bp using bedtools slop. The bedtools intersect was used to identify motifs that were within regions of interest and collect corresponding average deltaSVM scores.

Results

Human Retinal Epigenomic Data Can Be Used to Train a GKM-SVM Model

To generate impact score predictions for single nucleotide variants within human retinal CREs, we first trained a GKM-SVM model[31] to evaluate putative CRE sequences (Fig. 1A). As input, we started with a set of genomic windows defined by high-confidence ATAC-seq DNA-accessibility peaks (putative CREs). We split this set such that 80% of ATAC regions (~15k candidate CREs) were used as a positive training set, and 20% (~3.7k) were kept as a hold-out set to test the validity of predicted impact scores (Fig. 1B, Supplementary Fig. S1, Supplementary Table S1). Also for input, we generated an equal-sized negative training set of GC-matched non-coding genomic sequences (Supplementary Figs. S1B, S1C). As expected for putative CREs and control regions, we found that both positive and negative datasets were enriched for intronic and intergenic regions. We also found that the negative training dataset was depleted of promoter regions when we removed any overlap with the positive training data (Supplementary Fig. S1C). As a control, we trained a separate model using the same input data but with the positive and negative region labels shuffled randomly to demonstrate the baseline behavior of the model parameters.

To use our trained model to predict the impact of CRE variants, we next generated a scored vocabulary of all possible non-redundant 11-mer sequences. This *k*-mer length was chosen because it is long enough

to encompass most eukaryotic TF binding motifs.[58] We then used the trained model to weigh each 11-mer based on its relative similarity to the positive training set (positive values) versus the negative training set (negative values). This scored vocabulary was subsequently used to evaluate variant sequences in the generation of variant impact (deltaSVM) scores (Fig. 1A).

Finally, to generate individual CRE variant impact scores, we performed base-wise in silico saturation mutagenesis on CREs from the 20% hold-out dataset. We then used the scored 11-mer vocabulary to predict impact scores for every possible single nucleotide variant within these CREs. These predicted impact scores represent the difference between the sum of all 11-mers that scan across a given single nucleotide variant compared to the sum of those that scan across the reference allele (Figs. 1A, 1C).[32] A negative impact score therefore is assigned to a variant when it causes the sequence to become less similar to the positive training dataset compared to the original reference sequence. When deltaSVM scores are combined across a genomic region (Fig. 1B), distinct features of CREs become apparent (Fig. 1C). For example, inspecting for contiguous, highly negative summed deltaSVM scores, it is possible to identify well-characterized TF binding motifs in the reference sequence, such as the TAATCC motif favored by the K50 homeodomain transcription factors orthodenticle homeobox 2 (OTX2) and cone–rod homeobox (CRX) and the CCCTC-binding factor (CTCF) binding motif (Fig. 1C′, Supplementary Fig. S3).

Performance and Biological Relevance of the Trained SVM Model

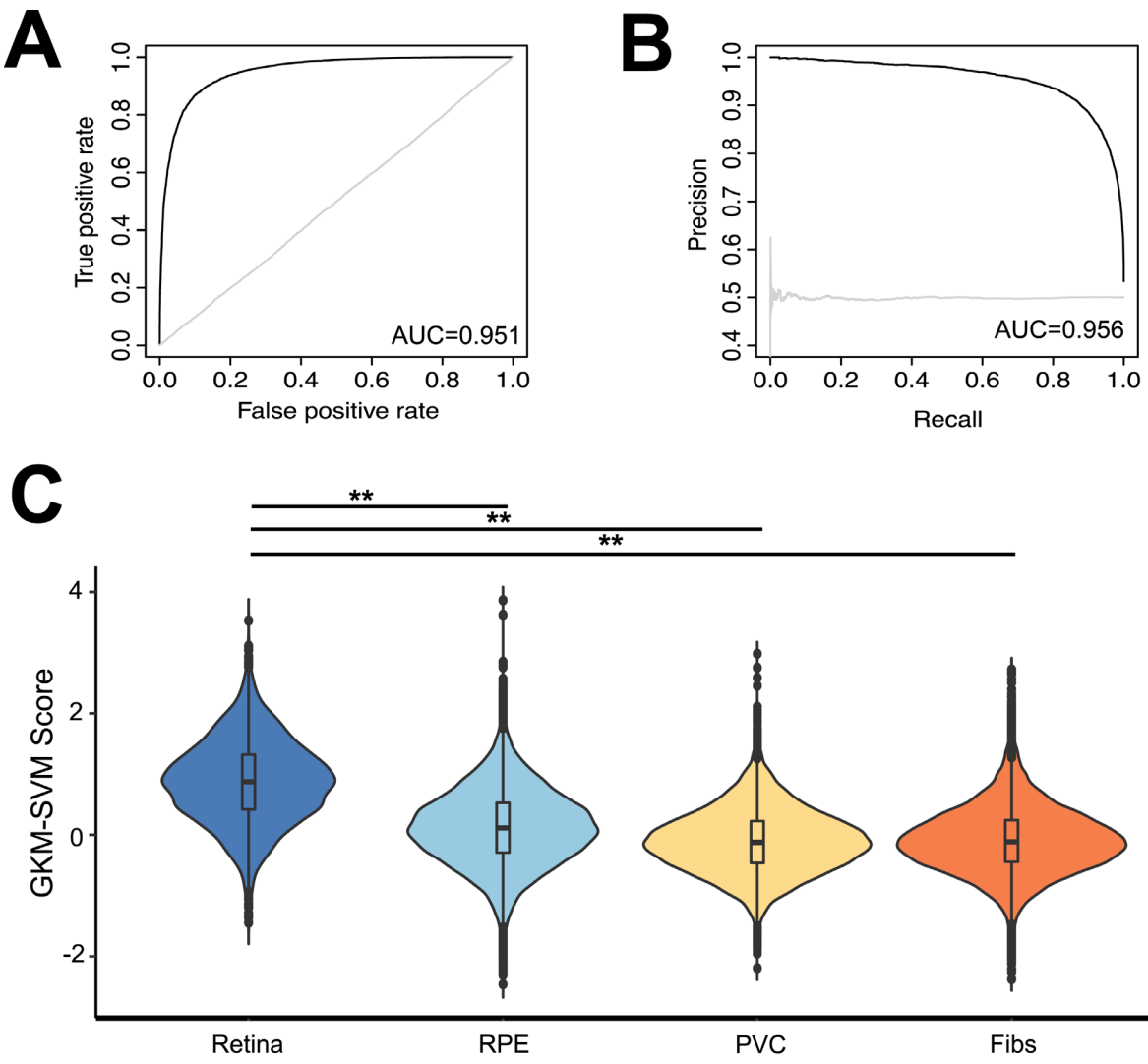
To assess the validity of this approach, we first performed fivefold cross-validation on the original and shuffled models. The training data were randomly assigned to one of five outgroups, and each outgroup was scored against a model trained excluding that outgroup. This cross-validation allows for the specific calculation of false positives and negatives, as well as model precision. These measures of model accuracy can be plotted as a ROC curve (Fig. 2A), or a precision-recall curve (Fig. 2B). For this model's ROC curve, an area under the curve (AUC) of 0.951 was achieved, indicating a highly accurate model with low false positivity. Similarly, the precision-recall curve for this model demonstrated an AUC = 0.956, indicating high precision. In contrast, when positive and negative labels were shuffled for the training data, the ROC and precision-recall curves demonstrated baseline AUCs, showcasing the specificity of model training gained by the true positive and negative datasets.

To determine the tissue-specificity of our trained models, we next used these models to compare retinal versus non-retinal CRE sequences. We found that our original model scored retinal-specific CREs from our hold-out dataset much more highly than non-retinal CRE datasets of equal size (Fig. 2C). Retinal ATAC-seq regions demonstrated a wide variety of scores, averaging at a GKM-SVM score of 0.870. This was significantly higher than all other non-retinal ATAC-seq data ($P < 2e\text{-}16$ in pairwise Wilcoxon rank-sum tests, Bonferroni adjusted). RPE, being developmentally related to the retina, scored most neutrally with an average score of 0.125, as compared to the PVC at –0.104 and fibroblasts at –0.087 (Fig. 2C). These differences were eliminated when we used the shuffled model to score CREs (Supplementary Fig. S2A), demonstrating the specificity of our original model for evaluating human retinal CREs.

To further assess the tissue-specific relevance of our original model, we searched for the enrichment of known transcription factor binding motifs within the top 1% of the scored 11-mer vocabulary. Within this group we found significant enrichment for motifs shared by well-known retinal transcription factors (Fig. 2D, Supplementary Table S2). Photoreceptor-associated motifs such as homeobox domain motifs consistent with CRX and OTX2 binding were most highly enriched, whereas more broadly expressed retinal TF motifs such as basic helix–loop–helix (bHLH) motifs were also highly ranked. This enrichment within the 11-mer vocabulary suggests an additional level of tissue specificity in our model.

Variant Impact Scores Correlate with Conservation of Non-Coding Sequences

Pathological variants within human retinal CREs are relatively rare but can disrupt visual function.[2,14–16] We therefore hypothesized that, if our variant impact scores predict variant pathogenicity, then strongly scored variants may be rare within the normal human population. To test this, we compared our predicted variant impact (deltaSVM) scores to SNV allele frequency in the Genome Aggregation Database (GnomAD).[53] When they were directly plotted against each other, we found that most variants clustered around the neutral deltaSVM scores and that SNVs with higher allele frequencies were preferentially clustered around neutral deltaSVM scores (Fig. 3A). Conversely, alleles with a large predicted impact (negative or positive) were relatively rare. This compar-



D

Motif	Factor	Log p-val
	Homeobox	-4.256e03
	Homeobox	-5.980e02
	bZIP	-5.517e02
	bHLH	-4.907e02
	NFY	-4.336e02

Figure 2. The GKM-SVM model is accurate and retinal specific. (A) ROC curve for fivefold cross-validation of the GKM-SVM model trained on human retinal ATAC-seq data (*black*) and for the model trained on shuffled positive and negative training data (*light gray*). AUC ATAC = 0.951; shuffled = 0.498. (B) Precision-recall curve for fivefold cross-validation of the GKM-SVM model trained on human retinal ATAC-seq data. AUC ATAC = 0.956, shuffled = 0.499. (C) Violin plot demonstrating GKM-SVM model scores for human retinal holdout data (Retinal), retinal pigmented epithelium (RPE) ATAC-seq peaks,[46] primary visual cortex (PVC) ATAC-seq peaks,[47] and human fibroblast (Fibs) ATAC-seq peaks.[48] $P < 2e\text{-}16$ (Kruskal–Wallace). For all pairwise comparisons to retinal scores by the Wilcoxon rank-sum test, $P < 2e\text{-}16$ (**Bonferroni adjusted). (D) Top enriched TF motifs from HOMER in top 1% of scored model 11-mer vocabulary.

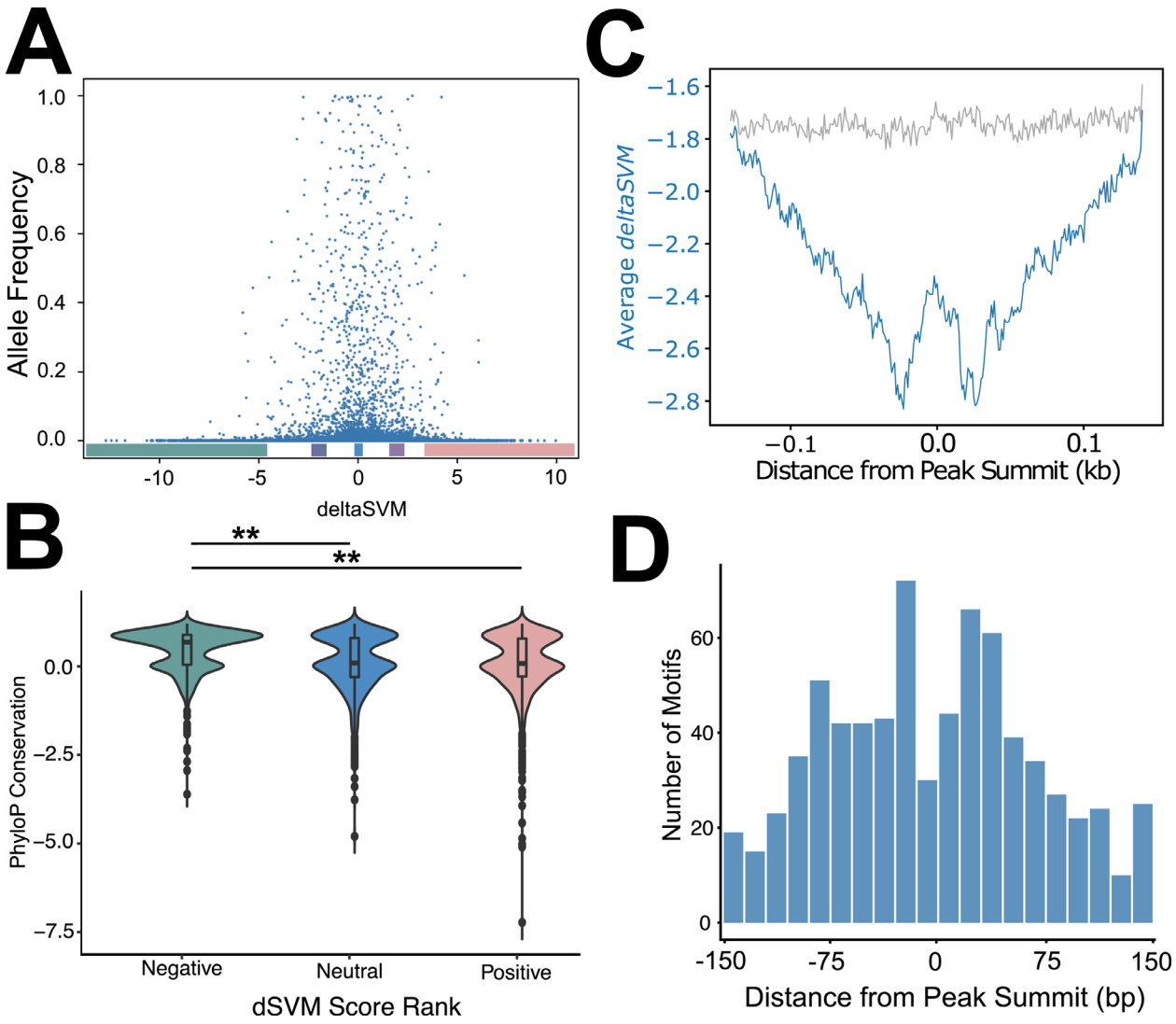


Figure 3. The deltaSVM scores match allele frequencies and conservation. (A) Scatterplot demonstrating the correlation between deltaSVM scores and SNV allele frequencies from the GnomAD database. The deltaSVM bins for the bottom, mid-bottom, mid-top, and top 1% of scores are highlighted along the *x*-axis. (B) Violin plot of phyloP conservation scores in negative, neutral, and positive deltaSVM scores ($P < 2e\text{-}16$, Kruskal–Wallace; $^{**}P < 0.0001$, Wilcoxon rank-sum test, Bonferroni adjusted). (C) Changes in average deltaSVM across the average 301-bp window of 20% outgroup CREs in the model trained on human retina ATAC-seq data and on the shuffled model (*gray*). (D) Sums of retinal motif classes across the 301-bp window of 20% outgroup CREs in 15-bp bins.

ison demonstrates that common alleles in the population are much more likely to have mid-ranging deltaSVM scores, whereas more extreme scoring SNVs are less common in the population. For subsequent analyses, we binned variants by deltaSVM score. The majority of alleles scored with a deltaSVM between –3 and 3, with 1% of all alleles scoring less than –4.9 and another 1% scoring more than 3.4, constituting the top and bottom 1% bins (Fig. 3A). For mid-ranging, more common alleles, we constrained alleles to deltaSVM scores from –2.5 to 2.5. These mid-top and -bottom 1% score bins spanned respectively from –2.5 to –2.3 and 2.26 to 2.5. A fifth bin was additionally defined, spanning the most neutral deltaSVM scores from –0.001 to 0.001 (Fig. 3A, Supplementary Table S3).

Consistent with these trends, we would expect deltaSVM to be negatively correlated with conservation values across species. Evolutionary conservation of specific CRE sequences suggests that those sequences are functionally important. To test this, we binned deltaSVM scores into the most negative, most positive, and neutral deltaSVM categories as defined in Figure 3A and compared these categories to phylogenetic conservation scores from the phyloP database. The most negatively scored SNVs (average –9.98)

corresponded to more conserved sequences (higher phyloP scores around an average of 0.456), whereas more neutral (average 4.9e-06) or positively (7.39) scored SNVs had lower conservation scores (neutral average phyloP score = 0.119; positive average phyloP score = 0.101; negative to neutral/positive P values < 2e-16, Bonferroni corrected) (Fig. 3B, Supplementary Fig. S2B). SNVs with more negative predicted impacts therefore appear to be more highly conserved, indicating their potential regulatory value in a given putative CRE.

Another test of the relevance of the variant scores is the distribution of these scores across the linear sequence of CREs. The center of retinal CREs is typically enriched for transcription factor binding motifs with constrained spacing; therefore, variants closer to the center of CREs are more likely to disrupt TF binding and be pathogenic.[57] We would expect a similar trend for our predicted impact scores where deltaSVM values would be more strongly negative toward the center of CREs, where there are more likely to be functional sequences. We therefore averaged negative deltaSVMs across all 3.7k 301-bp CRE windows in our test dataset and compared them to the same averages from the shuffled model and position within CREs. When plotted, a clear trend emerged, where deltaSVM scores decreased toward the center of CREs, corresponding to the peak of DNA accessibility (Fig. 3C). However, at the center of CREs, the average deltaSVM score increased locally, generating a bimodal distribution of negative impact scores. This may reflect the actual distribution of TF binding within CREs.[57] To test this, we analyzed the distribution of TF motifs and found a similar trend. Motifs consistent with TFs such as cAMP response element-binding protein (CREB), CRX, myocyte enhancer factor 2D (MEF2D), neural retina leucine zipper (NRL), OTX2, and retinoid-related orphan nuclear receptor β (RORB) were most enriched directly adjacent to the center of CREs (Fig. 3D, Supplementary Fig. S4). When the distribution of deltaSVM scores from the shuffled control model was plotted, no such pattern was observed. Altogether, these results demonstrate that deltaSVM scores generated from our original model can accurately predict the enrichment of functional DNA sequences near the center of CREs.

Highly Negative Variant Impact Scores Disrupt TF Binding Motifs

The correspondence of variant impact scores with allele frequency and conservation suggests that deltaSVM value correlates with TF binding motifs. To evaluate this directly, we first determined the counts of specific TF binding motifs in the most negative, most positive, and the neutral deltaSVM categories. In each of these categories (Fig. 3A), the 22 bp around a given SNV were scored for the presence of known motifs in the HOCOMOCO human motif database.[56] In the most negatively scored category, we found many well-characterized retinal TF motifs, such as OTX2 and CREB, represented in the reference sequences (Figs. 4A, 4A″; Supplementary Fig. S5). By contrast, in the SNV sequences, there were far fewer motifs as scored by FIMO, indicating that the variant sequences specifically disrupt the sequence of the motif. This pattern varies in the mid-bottom deltaSVM scoring variants. Although MEF2D motifs show a high number of motif calls, OTX2 calls by contrast are much less frequent (Fig. 4A′). The variants scored in the top bins demonstrate the opposite trend, with few calls for motifs of interest in the reference sequence, with modest increases in the variants likely due to situations where the variant coverts a sequence into an approximation of a TF binding motif (Figs. 4A, 4A′, 4A″; Supplementary Fig. S5).

To gain a better understanding of the relationship between TF motifs within CREs and our predicted variant impact scores, we identified all instances of specific TF motifs in our test dataset and centered these on 60-bp windows. We then plotted the distribution of deltaSVM scores across these windows. We observed that the scores dipped dramatically around canonical motif sequences but the flanking regions were relatively unaffected (Figs. 4B, 4B′, 4B″; Supplementary Fig. S6). This indicated that our scoring strategy is uniquely sensitive to these motifs. TF motif sequences, however, allow for flexibility at specific positions across the motif. We therefore sought to determine how impact scores varied within a motif itself. At a single base-pair resolution, we found that the significance of the core motif of some TFs such as OTX2 is apparent (Fig. 4C, Supplementary Fig. S6). Average deltaSVM scores for SNVs in the core TAATCC sequence are more negative than for SNVs in contextual positions immediately adjacent (Fig. 4C). For other motifs, changes to key nucleotides in a motif sequence become apparent, with larger decreases highlighting the CTA/TAR essential sequences of the MEF2D consensus motif (Fig. 4C′) as well as key bases in the CREB motif (Fig. 4C″). These changes in deltaSVM scores along TF binding motifs demonstrate the specificity of these scores to isolate crucial core sequences in a putative CRE and where alterations to the sequence may have significant impact on function.

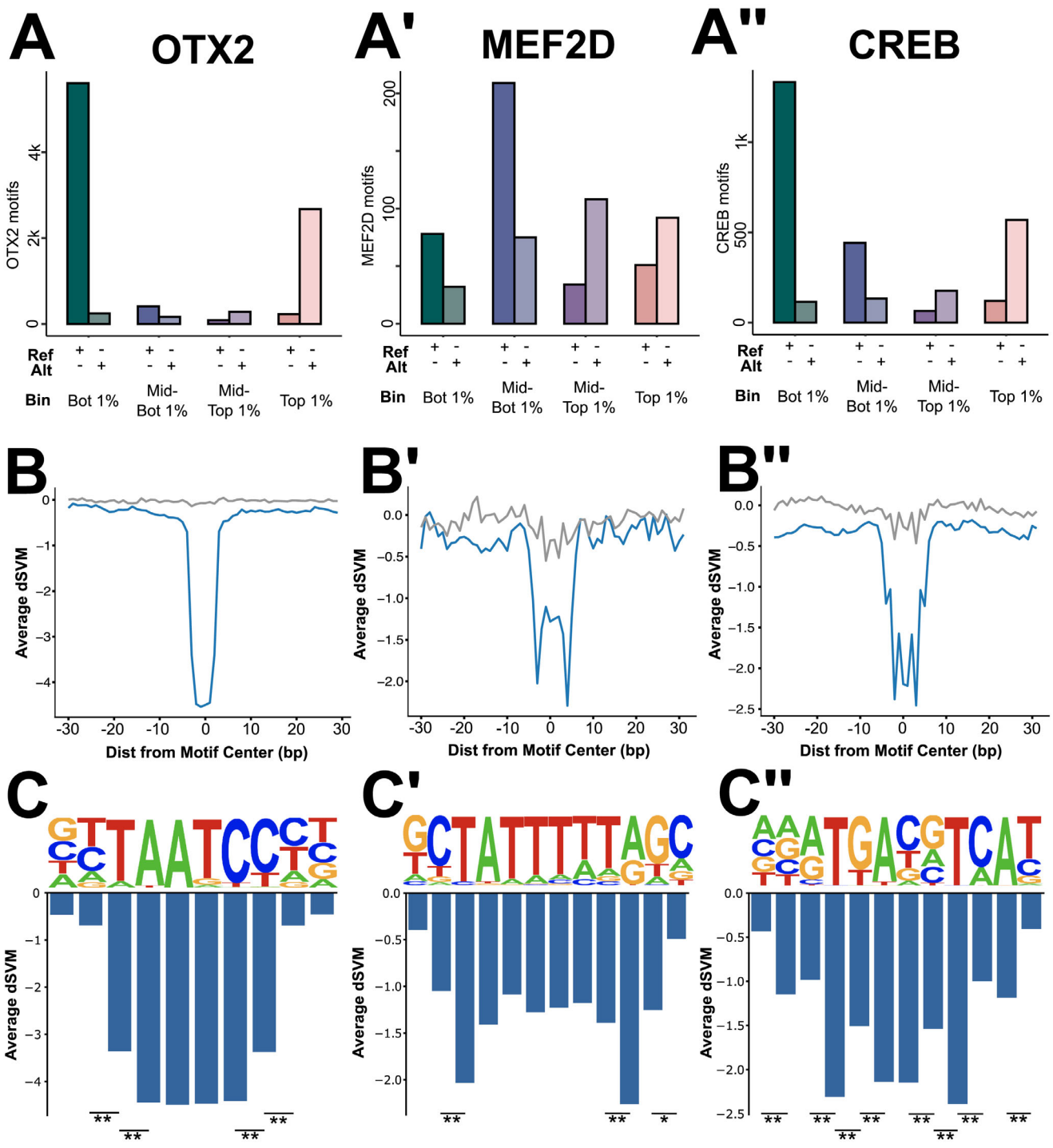


Figure 4. Disruption of retinal TF motifs dramatically reduces deltaSVM scores. (A–A′′) Numbers of motifs as scored by FIMO in reference and variant sequences for bins highlighted in Figure 3A. Motifs shown are OTX2 (A), MEF2D (A′), and CREB (A′′). (B–B′′) Line plots showing the average deltaSVM for SNVs ± 25 bp around the core motifs shown in (A) to (A′′) in *blue*. Scores for the same motifs in the shuffled model are shown in *gray*. (C–C′′) Bar plots showing the average deltaSVM for SNVs on a base-pair resolution within the core motifs of those shown in (A) to (A′′). *$P < 0.02$; **$P < 0.002$ (ANOVA with post hoc Tukey).

Prediction Scores Across a Conserved CRE Match Changes in Reporter Expression

Prior studies have demonstrated the ability to experimentally test the impact of every possible SNV within a retinal CRE using a massively parallel reporter assay (MPRA)-based approach. Kwasnieski et al.[39] used SNV saturation mutagenesis of the mouse rhodopsin promoter to test the impact of every possible variant with base-pair resolution (Fig. 5A). This analysis highlighted the unique importance of CRX and NRL motif sequences within the larger CRE. As a final test of our predicted variant impact scores, we used our human retinal CRE-trained model to assign predicted variant impact scores to every possible SNV within this mouse sequence (Fig. 5B). Although the human ortholog of this CRE was not included in the original 80% training set and despite being tested against reporter data generated in the mouse retina, the model predicted markedly negative deltaSVM scores overlapping the previously identified TF binding sites, highlighted in Figures 5A and 5B, as well as similarities in the region between the CRX(2) and NRL binding motifs (Fig. 5B, Supplementary Table S4). When relative expression from Figure 5A was plotted against deltaSVM scores in Figure 5B, these values were found to be positively correlated, with a Pearson correlation coefficient of 0.506. Altogether, this correlation and consistency across TF motifs suggested to us that our CRE variant prediction strategy is robust.

VISIONS: A Resource for Human Retinal Regulatory Variant Interpretation

The analyses described above suggested that this variant impact scoring strategy using the GKM-SVM/deltaSVM workflow trained on human retinal ATAC-seq data has several biologically relevant features. We therefore extended these scores to include a more inclusive set of 39,437 putative retinal CREs as defined by both our original ATAC-seq regions and the top 10,000 peaks of retina-associated, previously generated[38] TF ChIP-seq data by MACS2 score. This analysis, entitled "Variant Impact Scores in Ocular Non-coding Sequences" (VISIONS), is available on the UCSC genome browser to query and to compare with human retinal DNA-accessibility, transcription factor binding, and histone modifications (http://genome.ucsc.edu/s/CherryLab/VISIONS_TrackHub). It is our hope that these predicted impact scores can assist other researchers in identifying and interpreting variants of interest within non-coding retinal regulatory elements.

Discussion

The identification and characterization of non-coding variants in CREs can be a resource-intensive process. This study demonstrates the value of machine learning to identify highly impactful SNVs and to generate an exhaustive analysis of retinal CRE variant impact scores. These scores were generated through training a GKM-SVM model on adult human retinal ATAC data. By utilizing this machine learning–based approach, large epigenomic sequencing datasets can be analyzed, and, with the GKM-SVM and deltaSVM approaches, sequence variations can be easily screened. This SVM-based method has been previously used to highlight specific sequence features in the mouse retinal epigenome and to predict retinal reporter expression post hoc.[34–37] Together these previous studies and our current work demonstrate the potential of this approach to characterize human retinal CRE sequences for the identification of crucial features and their variants. Although this approach can be applied to many types of sequencing data, the use of general chromatin accessibility via ATAC-seq allows the model to incorporate the sequence features of diverse regulatory elements in a less biased approach than using more specific ChIP-seq data. Ultimately, we hope that the model generated in this study can be used to identify non-coding sequence variants that are likely to disrupt retinal CRE function to guide deeper analyses of non-coding variants. Currently, single nucleotide variant scores for sequences in 39,000 putative retinal CREs can be accessed via our UCSC genome browser track to identify variants with large predicted impacts on retinal CRE function (http://genome.ucsc.edu/s/CherryLab/VISIONS_TrackHub).

This study presents a machine learning model of putative human retinal CREs and the predicted impact of all possible SNVs in a set of tested sequences. This model behaves in a tissue-specific manner and accurately identifies the enrichment of well-characterized TF binding motifs. Through the analysis of related datasets and known motif databases, the model trained on human retinal ATAC data versus genomic background can clearly identify sequences of interest in a biologically relevant manner, specifically scoring retina-associated sequences above non-retinal CRE sequences. Further, in the generation of variant impact deltaSVM scores, the scores for the model follow known conservation, and the distribution of scores as compared to SNV allele frequency implies that variants predicted to have a strong impact on CRE function may be deleterious because they are rare in the human population. These variants

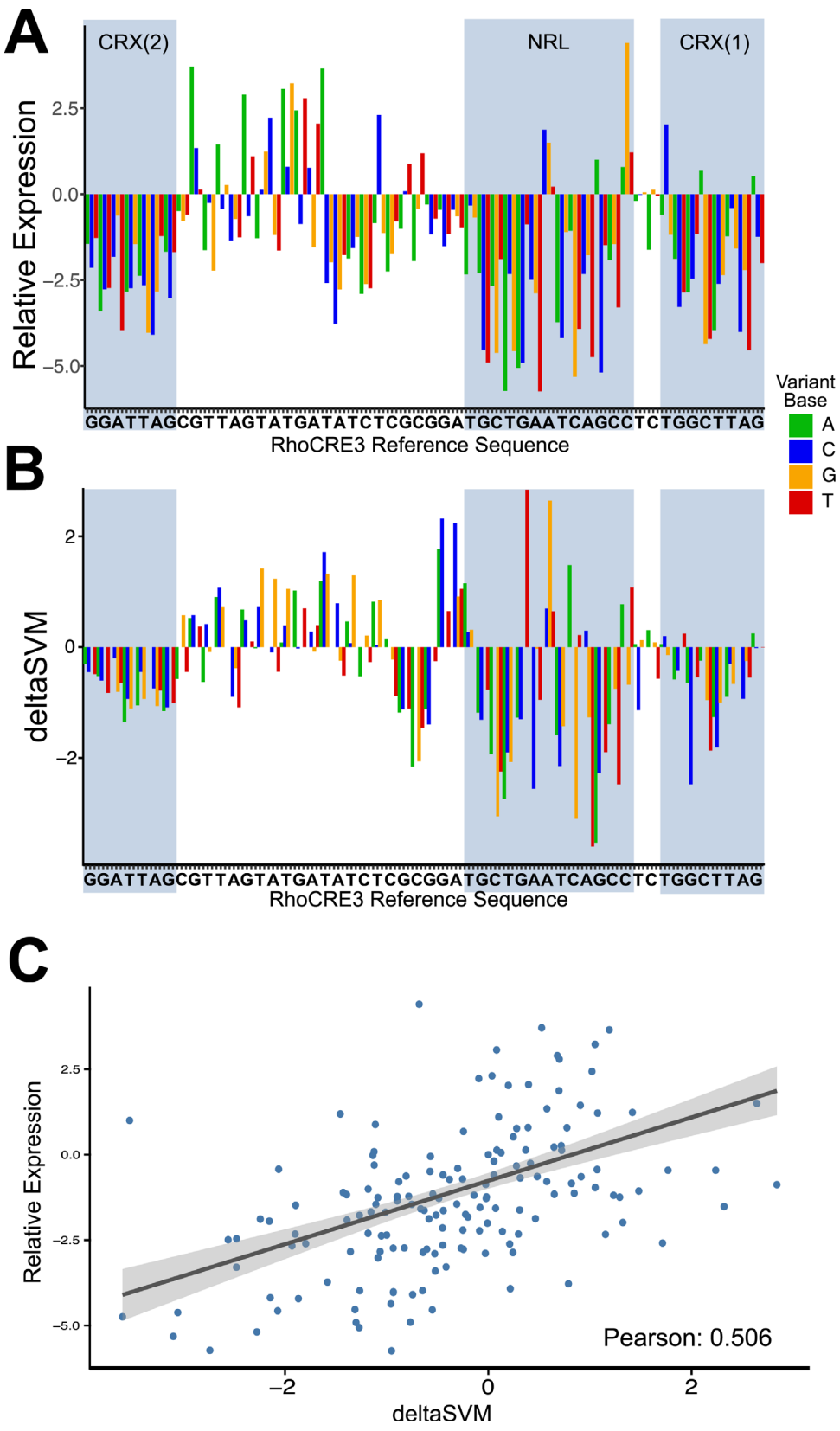


Figure 5 The deltaSVM scores for SNVs in RhoCRE3 correlate to changes in reporter expression. (A) Relative expression (log2[mutant/wild-type]) of fluorescent reporter for variants in RhoCRE3 in mouse retina from Kwasnieski et al.[39] Identified TF binding sites of CRX (1 and 2) and NRL are highlighted. (B) The deltaSVM variant impact scores of the same SNVs as in (A) along the RhoCRE3 locus. Identified TF binding sites of CRX (1 and 2) and NRL are highlighted. (C) Scatterplot of relative expression and deltaSVM scores in (A) and (B) with linear regression and 95% confidence intervals. Pearson = 0.506.

specifically identify where disrupted sequences intersect canonical transcription factor binding motifs to potentially affect CRE activity. The enrichment of negative deltaSVM scores around known motifs specifically highlights well-characterized core sequences and key base positions in motifs and thus the ability of the model to recognize the value of these sequences. Additionally, it becomes apparent that distinct motifs contribute differently to model relevancy. Potentially, the motif disruption and severity of the deltaSVM score may be an indicator of the severity of impact on CRE and therefore retinal function. Those SNVs with the most negative deltaSVM scores were associated with the highest level of conservation, demonstrating that these sequences may have a distinct role in retinal function.

When observing these deltaSVM scores in the general context of the CRE, trends become apparent as to where the most impactful variants are found, confirming known features of CREs. The trend of deltaSVM scores across putative CREs demonstrates both the known density of true TF binding sites near the summit and the depletion at the summit itself. This is consistent with findings from other studies, which show that TF motifs are most enriched around the center of CREs but are somewhat depleted at the direct summit.[55] These data indicate that disruptions to these TF motifs have the most dramatic impact on CRE scoring. Previous studies have performed massively parallel reporter assays to test the function of specific CRE sequences.[39,60,61] The results of these studies emphasize the impact of specific TF motif disruption and also serve as an important resource for the validation of machine learning predictions of variant impact. The characterization of these motifs is highly conserved, as negative deltaSVM scores from this model specifically correlate with MPRA-based approaches.[33,39] These results demonstrate both the ability of this model trained on human retinal epigenomic data to identify variants with notable relevance to changes in gene expression and its ability to operate across species in a conserved manner. This ability of the model to identify sequences of interest, especially variants that correlate to losses in gene expression, demonstrates the ability of this model to predict non-coding variants with relevance to retinal disease.

This model has unique value in the retina, in that it can specifically evaluate sequences associated with CREs, lending itself to a wide variety of applications. The deltaSVM impact scores can be used in the identification of crucial TF binding sites in a high-throughput manner. In particular, the in silico saturation mutagenesis approach to generating a database of deltaSVM scores means that variants can be pre-screened by their predicted change in regulatory function. In the screening of regions identified via GWASs, such data can specifically narrow down regions of interest and locations of variants of functional value to the retina. In more precise applications, variants identified in patients can be quickly ranked by their relevance to this model and prioritized for further functional investigation. This model can be further refined via integration of new epigenomic datasets, in particular single-cell epigenomic datasets, to refine the sensitivity and specificity of these predictions. In the rapidly moving field of artificial intelligence, new machine learning strategies will also likely enable characterization of new and different sequence-based features within CREs.

In sum, this workflow and the resulting prediction scores serve as a promising genomic tool for guiding the interpretation of non-coding sequence variation and for narrowing the search space for potentially pathogenic regulatory variants in visual disorders. Validation of the model demonstrates its capacity for tissue specificity and the identification of crucial CRE features. By applying a deltaSVM approach to putative CRE sequences, it is possible to pre-screen variant sequences of interest for further in vivo analyses. With further model validation, the presented database of SNV scores could be used in the identification of clinically relevant sequence variations and have applications beyond the bench.

Acknowledgments

The authors thank Thomas Vierbuchen, Marty Yang, Eric Thomas, and Brendan McShane for their review and valuable discussions regarding the progression of this manuscript.

Supported by a grant from the National Eye Institute, National Institutes of Health (R01EY028584 to TJC) and by a Catalytic Collaboration Grant from the Brotman Baty Institute (TJC and AYL).

Disclosure: **L.S. VandenBosch,** None; **K. Luu**, None; **A.E. Timms**, None; **S. Challam**, None; **Y. Wu**, None; **A.Y. Lee**, None; **T.J. Cherry**, None

* LSV and KL contributed equally to this work.

References

1. Daiger SP. RetNet: Retinal Information Network. Available at: https://sph.uth.edu/retnet/. Accessed April 6, 2022.

2. Duncan JL, Pierce EA, Laster AM, et al. Inherited retinal degenerations: current landscape and knowledge gaps. *Transl Vis Sci Technol*. 2018;7(4):6.
3. Berger W, Kloeckener-Gruissem B, Neidhardt J. The molecular basis of human retinal and vitreoretinal diseases. *Prog Retin Eye Res*. 2010;29(5):335–375.
4. Perez-Cerevantes C, Smith LA, Nadadur RD, et al. Enhancer transcription identifies *cis*-regulatory elements for photoreceptor cell types. *Development*. 2020;147(3):dev184432.
5. Nord AS, Blow MJ, Attanasio C, et al. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell*. 2013;155(7):1521–1531.
6. Ward LD, Kellis M. Interpreting Non-coding variation in complex disease genetics. *Nat Biotechnol*. 2012;30(11):1095–1106.
7. Howard ML, Davidson EH. *cis*-Regulatory control circuits in development. *Dev Biol*. 2004;271(1):109–118.
8. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337(6099):1190–1195.
9. Fritsche LG, Igl W, Cooke Bailey JN, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet*. 2016;48(2):134–143.
10. Scerri TS, Quaglieri A, Cai C, et al. Genome-wide analyses identify common variants associated with macular telangiectasia type 2. *Nat Genet.* 2017;49(4):559–567.
11. Bonelli R, Jackson VE, Prasad A, et al. Identification of genetic factors influencing metabolic dysregulation and retinal support for MacTel, a retinal disorder. *Commun Biol*. 2021;4(1):274.
12. Ratnapriya R, Sosina O, Starostik M, et al. Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat Genet.* 2019;51(4):606–610.
13. Struntz T, Kiel C, Grassmann F, et al. A mega-analysis of expression quantitative trail loci in retinal tissue. *PLoS Genet.* 2020;16(9):e1008934.
14. Ghiasvand NM, Rudolph DD, Mashayekhi M, Brzezinski JA, Goldman D, Glaser T. Deletion of a remote enhancer near ATOH7 Disrupts retinal neurogenesis, causing NCRNA disease. *Nat Neurosci*. 2011;14(5):578–86.
15. Nathans J, Davenport CM, Maumenee IH, et al. Molecular genetics of human blue cone monochromacy. *Science*. 1989;245(4920):831–838.
16. Bhatia G, Patterson N, Sankaraman S, Price AL. Estimating and interpreting FST: the impact of rare variants. *Genome Res*. 2013;23(9):1514–1521.
17. Jindal GA, Farley EK. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Dev Cell*. 2021;56(5):575–587.
18. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. *Cell*. 2013;152(6):1237–1251.
19. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*. 2012;13(9):613–626.
20. Klein DC, Hainer SJ. Genomic methods in profiling DNA accessibility and factor localization. *Chromosome Res*. 2020;28(1):69–85.
21. DeAngelis JT, Farrington WJ. An overview of epigenetic assays. *Mol Biotechnol*. 2008;38(2):179–183.
22. Degner JF, Pai AA, Pique-Regi R, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*. 2012;482(7385):390–394.
23. French JD, Edwards SL. The role of noncoding variants in heritable disease. *Trends Genet*. 2020;36(11):880–891.
24. Kvon EZ, Waymack R, Elabd MG, Wunderlich Z. Enhancer redundancy in development and disease. *Nat Rev Genet*. 2021;22(5):324–336.
25. Holder LB, Haque MM, Skinner MK. Machine learning for epigenetics and future medical applications. *Epigenetics*. 2017;12(7):505–514.
26. Ernst J, Kellis M. ChromHMM: automating chromatin state discovery and characterization. *Nat Methods*. 2013;9(3):215–216.
27. Chen S, Gan M, Lv H, Jiang R. DeepCAPE: a deep convolutional neural network for the accurate prediction of enhancers. *Genomics Proteomics Bioinformatics*. 2021;19(4):565–577.
28. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods*. 2015;12(10):931–934.
29. Movva R, Greenside P, Marinov GK, Nair S, Shrikumar A, Kundaje A. Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. *PLoS One*. 2019;14(6):e0218073.
30. Aldiri I, Xu B, Wang L, et al. The dynamic epigenetic landscape of the retina during development, reprogramming, and tumorigenesis. *Neuron*. 2017;94(3):550–568.
31. Ghandi M, Lee D, Mohammad-Noori M, Beer MA. Enhanced regulatory sequence prediction

using gapped k-mer features. *PLoS Comput Biol*. 2014;10(7):e1003711.
32. Lee D, Gorkin DU, Baker M, et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet*. 2015;47(8):955–961.
33. Shigaki D, Adato O, Adhikari AN, et al. Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. *Hum Mutat*. 2019;40(9):1280–1291.
34. Mo A, Luo C, Davis FP, Mukamel EA, et al. Epigenomic landscapes of retinal rods and cones. *eLife*. 2016;5:e11613.
35. Hughes AEO, Meyers CA, Corbo JC. A massively parallel reporter assay reveals context-dependent activity of homeodomain binding sites in vivo. *Genome Res*. 2018;28(10):1520–1531.
36. Beer MA. Predicting enhancer activity and variant impact using gkm-SVM. *Hum Mutat*. 2017;38(9):1251–1258.
37. Friedman RZ, Granas DM, Meyers CA, Corbo JC, Cohen BA, White MA. Information content differentiates enhancers from silencers in mouse photoreceptors. *eLife*. 2021;10:e67403.
38. Cherry TJ, Yang MG, Harmin DA, et al. Mapping the *cis*-regulatory architecture of the human retina reveals noncoding genetic variation in disease. *Proc Natl Acad Sci USA*. 2020;117(16):9001–9012.
39. Kwasnieski JC, Mogno I, Meyers CA, Jorbo JC, Cohen BA. Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. *Proc Natl Acad Sci USA*. 2012;109(47):19498–19503.
40. Kent WJ, Sugnet CW, Furey TS, et al. The Human Genome Browser at UCSC. *Genome Res*. 2002;12(6):996–1006.
41. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26(5):589–595.
42. Li H, Handsaker B, Wysoker A, Fennell T, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079.
43. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–842.
44. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137.
45. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57–74.
46. Wang J, Zibetti C, Sripathi SR, et al. ATAC-Seq analysis reveals a widespread decrease of chromatin accessibility in age-related macular degeneration. *Nat Commun*. 2018;9(1):1364.
47. Fullard JF, Hauberg ME, Bendi J, et al. An atlas of chromatin accessibility in the adult human brain. *Genome Res*. 2018;28(8):1243–1252.
48. Mai T, Markov G, Brady JJ, et al. NKX3-1 is required for induced pluripotent stem cell reprogramming and can replace OCT4 in mouse and human iPSC induction. *Nat Cell Biol*. 2018;20(8):900–908.
49. Kwon AT, Arenillas DJ, Worsley Hunt R, Wasserman WW. oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3 (Bethesda)*. 2012;2(9):987–1002.
50. Lee D. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics*. 2016;32(14):2196–2198.
51. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics*. 2005;21(20):3940–3941.
52. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10(2):giab008.
53. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434–443.
54. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 2010;20(1):110–121.
55. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 2010;38(4):576–589.
56. Kulakovksiy IV, Vorontsov IE, Yevshin IS, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res*. 2018;46(D1):D252–D259.
57. Grossman SR, Engreitz J, Ray JP, et al. Positional specificity of different transcription factor classes within enhancers. *Proc Natl Acad Sci USA*. 2018;115(30):E7222–E7230.
58. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. 2011;27(7):1017–1018.
59. Stewart AJ, Hannenhalli S, Plotkin JB. Why transcription factor binding sites are ten nucleotides long. *Genetics*. 2012;192(3):973–985.

60. Oh IY, Chen S. High throughput analysis of retinal *cis*-regulatory networks by massively parallel reporter assays. *Adv Exp Med Biol*. 2019;1185:359–364.

61. White MA, Kwasnieski JC, Myers CA, Shen SQ, Corbo JC, Cohen BA. A simple grammar defines activating and repressing *cis*-regulatory elements in photoreceptors. *Cell Rep*. 2016;17(5):1247–1254.