GENETICS

Conservation of cis-regulatory codes over half a billion years of evolution

Yohey Ogawa[1]†, Yu Liu[1], Connie A. Myers[1], Ala Morshedian[2], Gordon L. Fain[2], Alapakkam P. Sampath[2], Joseph C. Corbo[1]*

Identifying homologous cell types across species is essential for understanding cell type evolution. The retina is ideal for comparative analysis because its six major cell classes have persisted since the origin of vertebrates more than half a billion years ago. Here, we show that the retina's conserved cellular architecture is mirrored by deep conservation of the cis-regulatory codes that govern gene expression. Through single-cell chromatin accessibility analysis of lamprey, fish, bird, and mammalian retinas, we demonstrate cross-species conservation of cis-regulatory codes in all retinal cell classes despite extensive turnover of cis-regulatory regions. Conservation manifests as clustering of high-affinity transcription factor binding sites in cell class–specific open chromatin regions. Thus, the retina's cellular Bauplan is controlled by cis-regulatory codes, which predate the divergence of extant vertebrates.

INTRODUCTION

The evolutionary origin of cell types is a subject of enduring fascination (*1*–*4*). To infer the existence of specific cell types in the common ancestor of extant vertebrates—a species that lived ~560 million years (Ma) ago—it is necessary to compare homologous cell types between the two most evolutionarily distant vertebrate taxa: the jawless fishes (i.e., lampreys and hagfishes) and the jawed vertebrates (cartilaginous and bony fishes, amphibians, reptiles, birds, and mammals). The vertebrate retina is an ideal system for inferring cellular and molecular features that existed in the common vertebrate ancestor because its basic features—cell classes, connectivity patterns, and function—are remarkably conserved among all vertebrate taxa (*5*–*7*), including between jawed and jawless species (Fig. 1, A and B) (*8*–*10*). Nearly all vertebrate retinas contain six major cell classes (photoreceptors, bipolar cells, horizontal cells, amacrine cells, ganglion cells, and Müller glia) (*7*). Each cell class—with the exception of Müller glia— consists of multiple closely related "sister" cell types—expressing divergent sets of effector genes but retaining many shared transcriptional regulators—which arose via duplication and divergence from a single ancestral cell type (*1*, *2*, *11*). In the course of evolution, individual vertebrate species have expanded or contracted the number of cell types within each cell class to adapt to specific light environments or lifestyles (*6*, *12*, *13*). Thus, vertebrate retinas display remarkable cell type diversity couched within an evolutionarily stable framework of six invariant cell classes.

Cell class– and type–specific transcriptomes are determined by the action of transcriptional regulatory networks, which consist of hierarchical cascades of transcription factors (TFs) that bind to cognate binding sites within cis-regulatory elements (i.e., enhancers and promoters) to regulate gene expression and determine cell type identity (*14*, *15*). A "cis-regulatory code" or "grammar" is the particular combination and arrangement of TF binding sites within cis-regulatory elements that drives expression in a specific cell type or class. In the present study, we sought to determine whether the cis-regulatory codes governing retinal class–specific gene expression are conserved across vertebrates, including between jawed and jawless species. We find that the architectural invariance of the vertebrate retina is mirrored by deep conservation of the underlying cis-regulatory codes and that these codes emerged in the common ancestor of extant vertebrates more than half a billion years ago.

RESULTS

Single-cell chromatin accessibility profiling of vertebrate retinas

To determine the evolutionary antiquity of the cis-regulatory codes that govern gene expression in the vertebrate retina, we carried out a systematic analysis of TF-binding sites in the retinal cell class–enriched open chromatin regions (OCRs) of six diverse vertebrate species (Fig. 1). These species inhabit a wide range of photic environments and have correspondingly evolved divergent retinal cell type inventories (*7*, *16*). Thus, it is impossible to define one-to-one homology relationships for individual cell types across all species. We therefore focused our analysis on cell class–specific cis-regulatory codes. To accomplish this task, we acquired published retinal single-cell gene expression profiling [single-cell RNA sequencing (scRNA-seq)] and single-nucleus chromatin accessibility sequencing (snATAC-seq) data from two teleost fishes (zebrafish and goldfish) and two placental mammals (mouse and human). To broaden our phylogenetic sampling, we additionally conducted single-cell analyses on retinas from chicken (*Gallus gallus*) and sea lamprey (*Petromyzon marinus*), a jawless species. We examined these six datasets—generated by distinct protocols and from diverse sources—using either Signac/Seurat or ArchR, depending on whether the data were generated by multiome (snRNA-seq + snATAC-seq) or snATAC-seq analysis, respectively. After initial preprocessing to remove low-quality cells, we performed dimensionality reduction followed by shared nearest neighbor modularity optimization–based clustering. For each species, we assigned clusters to one of the six retinal cell classes, either based on the expression of known class-specific marker genes in those species for which multiome data were available (lamprey, zebrafish, and human), or based on

[1]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA. [2]Department of Ophthalmology, David Geffen School of Medicine, UCLA, Los Angeles, CA 90095, USA.
*Corresponding author. Email: jcorbo@wustl.edu
†Present address: Molecular Life History Laboratory, Department of Genomics and Evolutionary Biology, National Institute of Genetics, 1111 Yata, Mishima 411-8540, Japan.

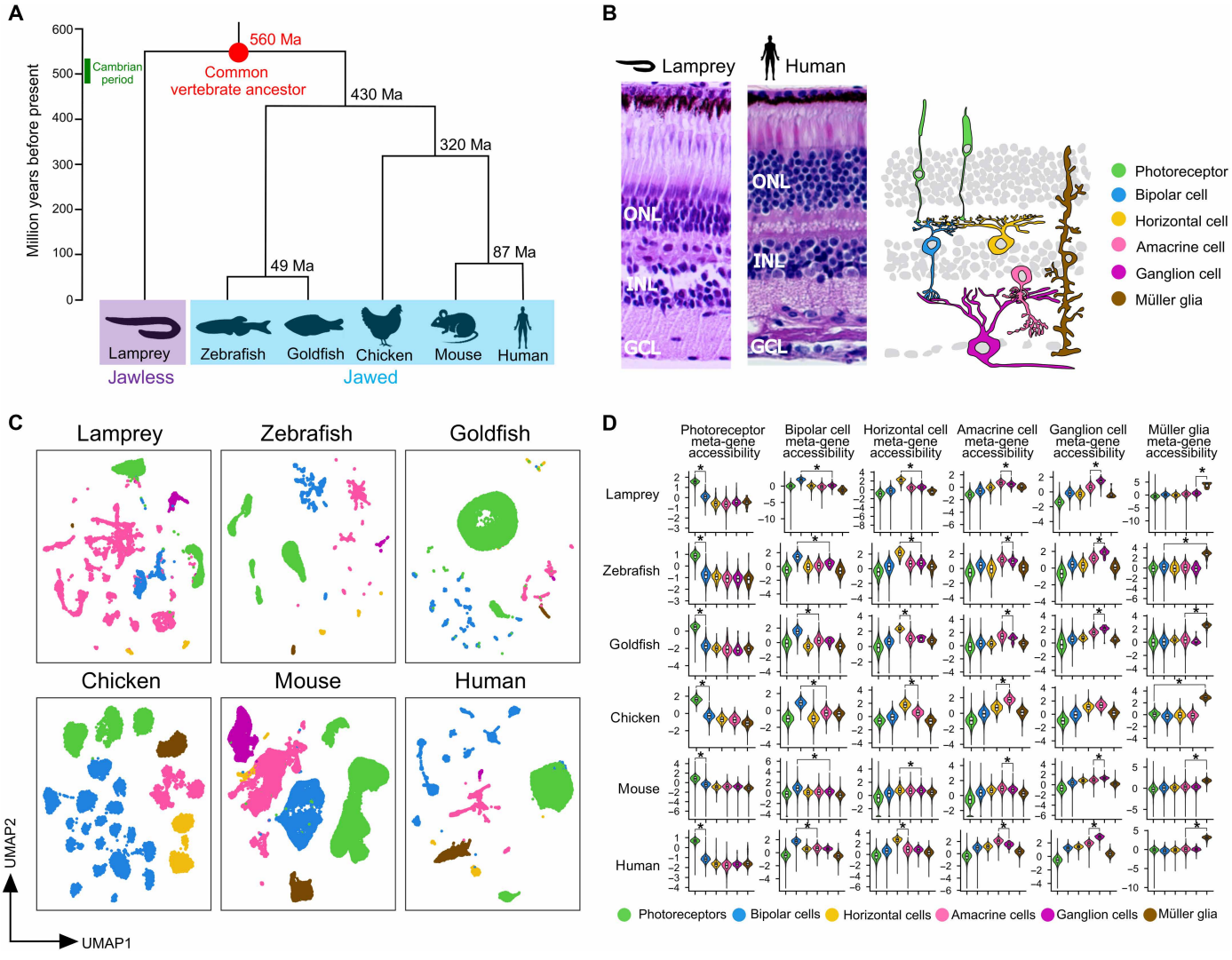

Fig. 1. Single-nucleus chromatin accessibility profiles for six retinal cell classes in six vertebrate species. (**A**) A phylogenetic tree of the vertebrate species used in this study. Divergence times are estimated on the basis of a published database (*66*). (**B**) (Left) Hematoxylin and eosin–stained sections of lamprey and human retina. (Right) Schematic of the six major retinal cell classes in vertebrates. ONL, outer nuclear layer; INL, inner nuclear layer; GCL, ganglion cell layer. (**C**) Single-nucleus ATAC-seq profiles of the six indicated species. Retinal cell classes were identified by clustering analysis and visualized in two-dimensional space using the Uniform Manifold Approximation and Projection (UMAP) method. Cell classes are color-coded as in (B). (**D**) Violin plots showing the average chromatin accessibility of cell class–enriched meta-genes generated from sets of EC marker genes for each cell class in each species (see Materials and Methods). The rows are grouped by species for each snATAC-seq dataset, and the columns are grouped by cell class. Cell-class meta-gene enrichment was determined by comparing the cell class exhibiting the highest score for the meta-gene with the second-highest scoring cell class. An asterisk indicates an adjusted *P* value <0.05 (Wilcoxon rank sum test followed by Bonferroni correction). Ma, million years.

chromatin accessibility at the promoters of class-specific marker genes in those species for which multiome data were not available (goldfish, chicken, and mouse). We removed from the analysis any residual clusters that could not be assigned to one of the major retinal cell classes. In this way, we identified clusters corresponding to each of the six retinal cell classes in all six species, with the exception of chicken ganglion cells, which were absent from the snATAC-seq dataset although present in scRNA-seq data (Fig. 1C and fig. S1).

Next, we sought to determine whether our cluster annotations are reflective of shared patterns of chromatin accessibility at class-enriched gene loci. To achieve this goal, we devised a quantitative measure of class-specific chromatin accessibility for each of the six retinal cell classes. First, we used scRNA-seq data to define evolutionarily conserved (EC) class-specific "meta-genes" consisting of a set of genes that were differentially expressed across cell classes and had a high class-specificity index in lamprey and three or more of the jawed species (see table S1 and Materials and Methods). We then measured chromatin accessibility over the promoter and gene body of each gene in the meta-gene and aggregated the values to create a single "meta-gene accessibility" score for each of the six cell classes in each of the six species (except for chicken ganglion cells; see above). We found that in all comparisons, the meta-gene accessibility score was highest for the expected cell class (Fig. 1D). These findings validate our cluster annotations and confirm that

class-specific signatures of chromatin accessibility are shared across all six species.

Extensive enhancer turnover during vertebrate evolution

Next, we wished to determine whether retinal class–specific cis-regulatory elements show sequence-level conservation across species. cis-regulatory elements typically occur in chromatin regions that are selectively open in the cell type(s) in which the element is active. For example, we previously showed extensive overlap between photoreceptor-enriched OCRs and the location of photoreceptor-specific enhancers and promoters across the mouse genome (*17*, *18*). We therefore decided to use class-enriched OCRs as a surrogate for class-enriched cis-regulatory elements in the present analysis. To quantify the extent of sequence-level conservation of class-enriched OCRs across species, we used single-cell data to create pseudo-bulk ATAC-seq profiles for each of the six retinal cell classes in five species: lamprey, zebrafish, chicken, mouse, and human (except for chicken ganglion cells; see above). We identified class-enriched OCRs for each retinal cell class in each species using a test of differential chromatin accessibility (fig. S2; see Materials and Methods). We then used the UCSC Genome Browser's LiftOver utility to map the union of each species' class-enriched OCRs onto all other vertebrate reference genomes for which precomputed reciprocal best-hit whole-genome alignment files were publicly available. In this way, we quantified "alignability" as the percentage of a species' class-enriched OCRs, which could be aligned with the genome of the target species (see Materials and Methods for details). We found that sequence alignability of retinal class–enriched OCRs progressively decayed with evolutionary distance such that beyond ~400 Ma, fewer than 3% of OCRs were alignable with the target genome (Fig. 2 and table S2). For example, the average alignability at 430 Ma (the distance between ray-finned fishes and amniotes) was 1.35%, while the average alignability at 563 Ma (the distance between jawed and jawless species) was 0.52%. We therefore infer that vertebrate cis-regulatory element turnover is extensive at great evolutionary distances.

To model the decline in OCR alignability over time, we fitted the data with a Gompertz equation, which is often used to model the growth or decay of populations. We observed close agreement with the model at short (i.e., <50 Ma) and long (>280 Ma) evolutionary distances but found major deviations from the model at middle

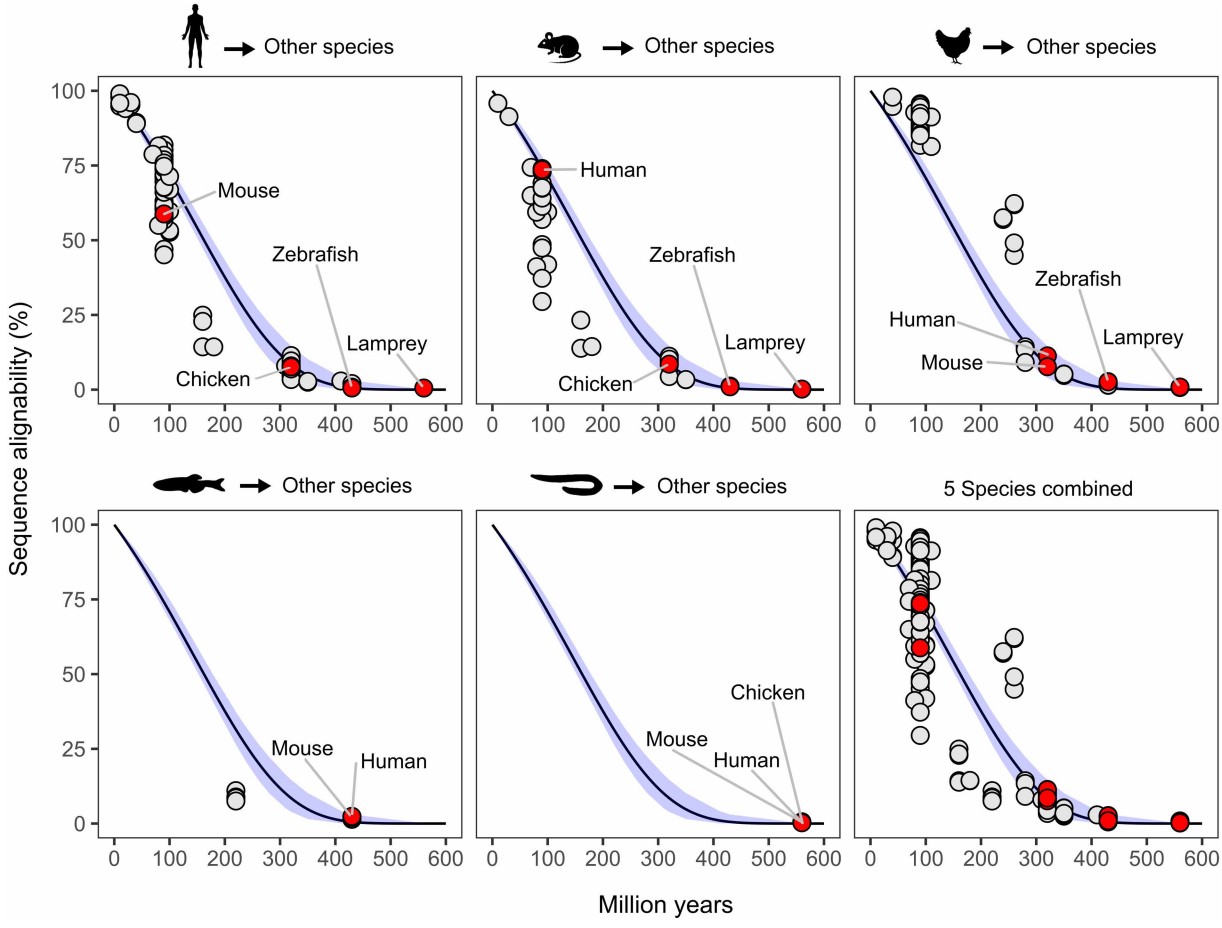

Fig. 2. Nearly complete sequence turnover of retinal cell class–enriched OCRs after 400 Ma of evolution. Retinal cell class–enriched chromatin regions in five query species (human, mouse, chicken, zebrafish, and lamprey) were mapped onto the genomes of diverse vertebrate species (also see table S2). The *x*-axis represents the evolutionary divergence times between query and target species according to a published database (*66*). The *y*-axis indicates the percent of query sequences that can be aligned to the target genome. The decay of sequence alignability over evolutionary time was modeled with the Gompertz equation using divergence time as a variable. The 95% confidence intervals measured by bootstrap resampling (see Materials and Methods) are shown as blue shading. The query species, when used as targets, are labeled and highlighted in red.

distances (50 to 280 Ma) (Fig. 2 and table S2). The greatest downward deviations—indicative of more extensive turnover than predicted by the model—were observed in mouse/human-to-mammal comparisons, particularly at ~90 Ma (i.e., mouse/human-to-placental comparisons), 160 Ma (mouse/human-to-marsupial), and 180 Ma (mouse/human-to-monotreme) (Fig. 2). These deviations are likely attributable to accelerated rates of evolution in certain mammalian clades (*19*, *20*). Conversely, we observed unexpectedly low rates of OCR turnover in chicken-to-bird comparisons (~90 Ma) and chicken-to-alligator/turtle comparisons (240–260 Ma), while chicken-to-lizard/snake comparisons (280 Ma) largely agreed with the model (Fig. 2). Despite wide variation in the rates of alignability at middle evolutionary distances, both the data and the model suggest nearly complete (i.e., ~99.5%) evolutionary turnover of retinal class–specific cis-regulatory elements beyond 500 Ma.

Deep conservation of retinal cis-regulatory codes

We and others observed conserved patterns of expression of class-specific genes despite extensive turnover of the cis-regulatory elements controlling their expression (table S1) (*7*, *9*). We therefore hypothesized that the underlying cis-regulatory codes might be conserved despite an absence of linear sequence conservation. To test this idea, we undertook a detailed comparative analysis of retinal class–specific cis-regulatory codes. The fundamental building blocks of a cis-regulatory code are TF-binding sites. Thus, as a first step toward elucidating the retinal codes, we used HOMER (*21*), a TF-binding site motif discovery algorithm, to comprehensively identify motifs enriched within 201–base pair (bp) regions centered on the summits of class-specific OCRs of six species (the rationale for choosing a 201-bp window is described in Materials and Methods). All existing motif databases derive from in-depth study of a small number of species. Thus, to avoid biases that might be introduced by focusing on "known" motifs, we used HOMER to detect de novo motifs. HOMER identifies enriched motifs by comparing an "experimental" set of target sequences with a set of control sequences. We therefore used class-enriched OCRs as our experimental dataset and a set of OCRs broadly open across multiple cell classes as our controls (Fig. 3A). HOMER generates a list of motifs in the form of a position probability matrix accompanied by an optimal detection threshold to maximize the enrichment of the motif in the target sequences. To ensure detection of relatively low-frequency but functionally important motifs, we retained all motifs that were present in >2% of the cell class–enriched OCRs and whose statistical significance of enrichment was $<10^{-10}$ (calculated using the binomial distribution). We analyzed a total of 35 datasets (i.e., six species × six cell classes, except for chicken ganglion cells), identifying a median of 47 de novo motifs in each dataset, with a minimum of eight motifs observed for lamprey Müller glial cells, likely due to the small number of these cells in our dataset (table S3).

Next, we sought to compare class-specific motifs across species to determine whether diverse vertebrates use a shared set of motifs in each retinal cell class. To accomplish this goal, we used Tomtom (MEME Suite; version 5.4.1) (*22*) to conduct pairwise motif comparisons and then hierarchically clustered motifs based on their similarity scores. HOMER often identifies multiple related motifs; thus, the several dozen de novo motifs found for a given cell class in an individual species may correspond to a smaller set of truly distinct motifs. By including all identified motifs for a given cell class in our hierarchical clustering, we can both delineate intraspecific motif redundancy and identify interspecific similarities in motif inventory.

Visual inspection of motif similarity matrices for each of the six retinal cell classes reveals multiple well-defined clusters of motifs for each cell class (Fig. 3B). Motifs in these clusters typically exhibit some of the highest "motif enrichment" scores (Fig. 3B; see Materials and Methods). In addition, some motif clusters are quite large, encompassing as much as ~35 to 50% of the motifs in a given cell class (e.g., in photoreceptors, bipolar cells, and horizontal cells), underscoring the presence of motif redundancy in the HOMER outputs. Within individual clusters, we typically find motifs from multiple species, indicative of cross-species conservation of motifs. To define discrete motif clusters likely corresponding to individual conserved motifs, we truncated the motif dendrogram (obtained by hierarchical clustering) at a height of 0.9 (equal to one minus the Pearson correlation coefficient of motif similarity values). We then designated a motif cluster as "EC" (if it included motifs from both jawless (i.e., lamprey) and three or more jawed species and if the median of the Pearson correlation coefficient among the motif similarities of the most enriched motifs from each species within a cluster was greater than 0.5 (Fig. 3, B and C; see Materials and Methods). For each EC motif cluster, the motifs with the highest motif enrichment score from each species were aggregated into a single "merged motif" (see Materials and Methods), which we designated as an "EC motif" (Fig. 3C and fig. S3). In this way, we identified a total of 16 EC motifs, with two to four motifs in each retinal cell class (Fig. 3, B and D). We infer that these motifs formed part of the retinal cis-regulatory codes of the most recent common ancestor of extant vertebrates based on their enrichment in both jawed and jawless species.

The close similarity of the species-specific position probability matrices used to create merged EC motifs suggests that these motifs are bound by homologous TFs with very similar DNA binding preferences across species. To nominate TFs likely to bind these EC motifs, we used Tomtom to compare EC motifs to known motifs in HOCOMOCO (*23*), a curated database of mouse and human TF-binding site motifs. We found that all EC motifs showed highly significant matches to one or more motifs in the database (fig. S3 and table S4). For example, the most enriched EC motifs in photoreceptors (PH_13) and bipolar cells (BC_11) are very similar to each other and closely match paired-type "K50" homeodomain–binding sites (K50 denoting the presence of lysine at position 50 of the homeodomain) bound by CRX and/or OTX2 in the HOCOMOCO database. These TFs and their cognate motifs are enriched in mammalian photoreceptor and bipolar cells relative to other retinal cell types and play critical roles in controlling development and gene expression in these cell classes (*24*, *25*). Most of the EC motifs show matches to binding sites of mammalian TFs previously shown to play key roles in regulating gene expression in the cell class in which the EC motif was found to be enriched (table S4) (*26*, *27*).

We postulated that the nonmammalian species in our study likely also express TFs in their respective cell classes with similar binding preferences to those of their mammalian counterparts. To test this idea, we mapped the candidate mammalian TFs onto their closest homologs in the nonmammalian species using OrthoFinder (*28*). We then intersected the resultant TF orthology groups with lists of differentially expressed genes obtained from scRNA-seq profiling of retinas from each of the six species. For 12 of the 16 EC motifs, we were able to identify cognate orthologous TFs whose expression was enriched in the corresponding cell class in jawless and four or more jawed species (fig. S4 and table S4; see Materials and Methods). This finding suggests that cross-species conservation of

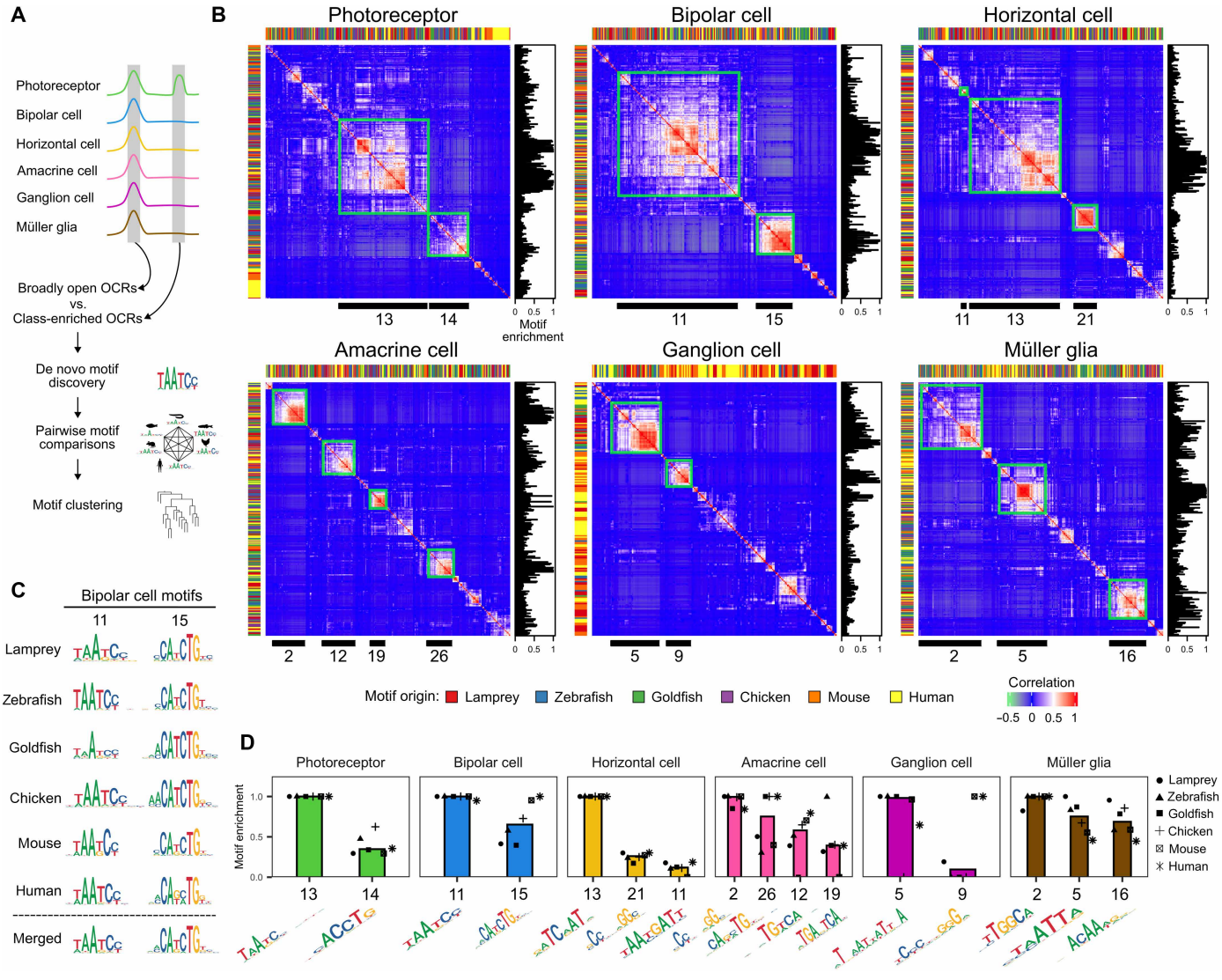

Fig. 3. Multiple cis-regulatory motifs for each retinal cell class are conserved between jawed and jawless species. (**A**) Schematic showing the methodology used for the identification of cell class–enriched OCRs, the discovery of sequence motifs, and motif clustering. (**B**) Heatmaps showing motif-similarity correlation values for all pairs of significantly enriched de novo motifs from six species. The motif-pair values were hierarchically clustered to reveal families of related motifs (see Materials and Methods). EC motif families—as defined in the main text—are enclosed by green boxes and indicated by numbered black bars at the bottom of each heatmap. The species of origin for each motif is indicated by color-coding across the top and left sides of each heatmap. Motif enrichment is the normalized statistical significance of motif enrichment in each species (i.e., the most enriched motif in each species has a motif enrichment = 1). (**C**) Representative examples of two EC motifs in bipolar cells. The sequence logo for the most significantly enriched motif in each species is shown, along with the logos for a merged motif representing the average of the six species motifs. For the full set of EC motifs, see fig. S3. (**D**) The median of the normalized statistical significance of motif enrichment for each of the 16 de novo motifs is shown (see Materials and Methods). The normalized motif enrichment value for each of the individual species is also presented. The sequence logos at the bottom represent the merged motifs.

class-enriched motifs is paralleled by cross-species conservation of cognate TF expression.

Next, we sought to systematically determine which features of class-specific cis-regulatory grammar are conserved across species. cis-regulatory grammar can be subdivided into two components: "vocabulary," consisting of the occurrence, affinity, and location of individual motifs within an OCR, and "syntax," comprising the co-occurrence, spacing, and relative orientation of pairs of motifs in an individual OCR (Fig. 4A). So far, we have identified significant class-specific enrichment of 16 EC motifs (Fig. 3). We next determined the spatial distribution of these motifs within class-enriched OCRs and quantified their position weight matrix (PWM) scores, a surrogate measure of binding affinity. We found that most motifs display a Gaussian-like distribution of enrichment conserved across species with a peak centered on the OCR summit (Fig. 4B). PWM scores also peaked at the OCR summit (Fig. 4B). In most cases, both motif enrichment and PWM scores declined monotonically with distance from the summit, approaching baseline levels around ±100 bp. Fundamentally similar motif distributions were identified in OCRs occurring in promoter regions (i.e., between −2000 and +1000 bp

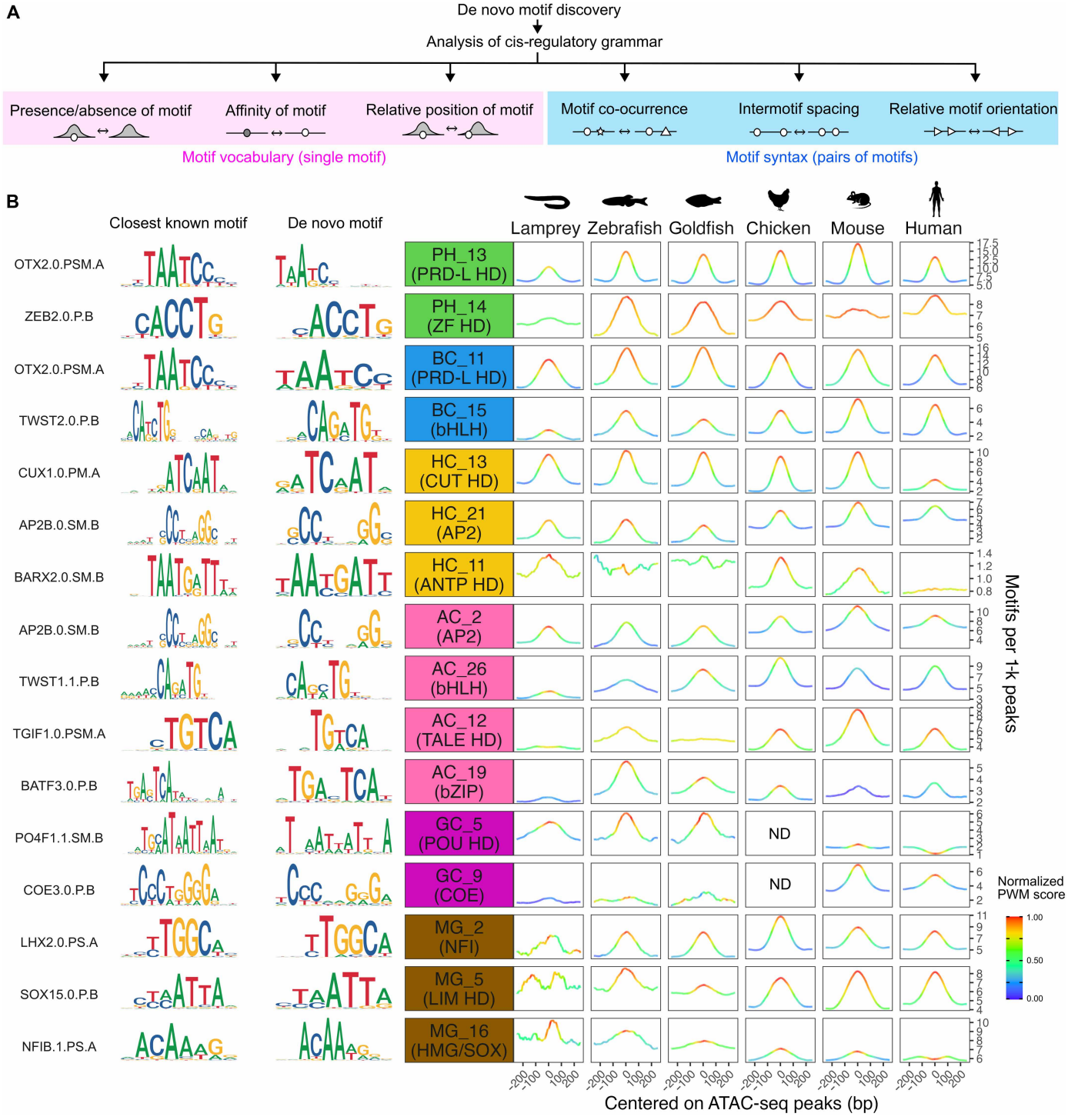

Fig. 4. Evolutionary conservation of retinal cis-regulatory motif vocabulary. (**A**) Schematic representation of the major features of cis-regulatory grammar. (**B**) The spatial distribution and normalized PWM scores of the EC motifs within class-enriched OCRs in the six species are presented. The sequence logos for each EC motif and the closest known motif in the HOCOMOCO database (*23*) are displayed on the left. Analysis of motif syntax is presented in figs. S6 and S7. PH, photoreceptor cell; BC, bipolar cell; HC, horizontal cell; AC, amacrine cell; GC, ganglion cell; MG, Müller glia; ND, not determined.

of the transcription start site) and nonpromoter regions (fig. S5). Together, these results indicate that features of cis-regulatory vocabulary are largely shared across motifs and species.

To ascertain whether syntactic features are conserved across species, we analyzed the co-occurrence, spacing, and relative orientation of all homo- and heterotypic pairs of EC motifs enriched in the same cell class (Fig. 4A). As expected from the central pattern of enrichment of individual motifs (Fig. 4B), we observed enriched co-occurrence of all motif pairs—with the exception of GC_9 + GC_9—across all six species (fig. S6). In contrast, we observed little evidence for conserved patterns of relative motif spacing or orientation (fig. S7). One motif, BC_11, shows enrichment of tandem pairs with an intersite spacing of 8 to 11 bp (i.e., approximately one helical turn), but this pattern is only conserved across jawed species (i.e., not in lamprey). A similar pattern of co-occurrence of monomeric K50 homeodomain–type motifs was previously noted in CRX chromatin immunoprecipitation sequencing peaks of mouse photoreceptors (*29*). The related photoreceptor-enriched K50-type motif identified in the present study (PH_13) is a dimeric motif (*18*). We therefore reanalyzed our photoreceptor-enriched OCRs using a monomeric K50 motif, which revealed helical co-occurrence of motif pairs similar to that observed for BC_11 (fig. S7B). Again, this co-occurrence pattern appears to be restricted to jawed species. We also detected a distinctive pattern of motif co-occurrence for the Müller glia–enriched HMG/SOX-type motif, MG_16, which consisted of pairs of motifs on opposite strands of the double helix, separated by 4 or 5 bp (fig. S7). This pattern of co-occurrence was conserved across all species including lamprey and likely represents a binding site for homo- or heterodimeric SOXE TFs (i.e., SOX8, SOX9, and SOX10), as previously described (*30*). We found that SOXE-type TFs showed Müller glia–enriched expression in all six species (fig. S4 and table S4). We therefore consider this dimeric site to represent a single motif occurrence and not a feature of higher-order syntax. Thus, although almost all motif pairs show higher rates of co-occurrence than in control regions, few other higher-order syntactic features are shared between jawed and jawless species.

Machine learning models of cis-regulatory grammar cluster by cell class

To enable quantitative comparison of class-specific cis-regulatory grammars across species, we trained machine learning models of grammar for each of the six cell classes in each of six species (with the exception of chicken ganglion cells). In light of the findings in the preceding section, we decided to build minimal vocabulary-based models that encompass only two key features of cis-regulatory grammar: the presence/absence of motifs and motif affinity. For this purpose, we constructed gapped *k*-mer support vector machine (gkm-SVM) (*31*) classifiers to distinguish cell class–enriched OCRs (i.e., the positive training set) from broadly OCRs (the negative training set). We constructed a total of 35 gkm-SVM models by randomly partitioning each training dataset into five subsets and performing fivefold cross-validation (see Materials and Methods). We then used the five resultant models for each dataset to score the 35 test sets from each cell class and species. We measured the performance of the models by calculating the receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC) (fig. S8). For same-dataset validation, the mean ROC-AUC value for the 35 models was 0.854 (±0.049 SD), with the best performance observed with the mouse horizontal cell model (0.920 ± 0.004 SD) and the worst performance with the lamprey amacrine cell model (0.733 ± 0.004 SD). Next, we evaluated the ability of the models to classify OCRs in the 34 other datasets. The average cross-species performance of models on other-class OCRs (e.g., lamprey photoreceptor model classifying mouse horizontal cell OCRs) was essentially random (ROC-AUC = 0.520 ± 0.086 SD) and provides an empirical estimate of baseline model performance. In contrast, for photoreceptor, bipolar cell, horizontal cell, and Müller glial models, we observed good performance in classifying same-class OCRs from different species, with all ROC-AUC scores above baseline performance, except for the lamprey Müller glial model whose performance was borderline overall, but consistently better for same-class OCRs than other-class OCRs (0.60 to 0.69 compared to 0.487 ± 0.030 SD) (fig. S8). The relatively poor performance of the lamprey Müller glial model is likely attributable to the small number of Müller glia identified in our snATAC-seq analysis and the corresponding paucity of class-enriched OCRs (151 sequences in total) available for model training (tables S5 and S6). We observed comparable cross-species performance for amacrine and ganglion cell models, except for the zebrafish and goldfish models, which demonstrated excellent performance on each other's datasets (ROC-AUC ≥ 0.80) but worse performance on non-teleost datasets (see fig. S8). Overall, the cross-species classificatory performance of these models confirms the existence of universally shared class-specific grammar features.

To evaluate the ability of these models to predict functionally important TF-binding sites across species, we used them to analyze the mouse *Gnb3* promoter, which drives expression in both photoreceptors and bipolar cells. We previously showed that this promoter contains five phylogenetically conserved K50 homeodomain–binding sites, two of which are required for photoreceptor and bipolar expression (*11*). We used photoreceptor and bipolar cell models trained on lamprey, zebrafish, goldfish, chicken, and human datasets to score the mouse *Gnb3* promoter. To visualize the contribution of individual nucleotides to the overall model scores, we used GkmExplain (*32*), a feature attribution method that displays the predicted relative contribution of individual nucleotides as a sequence logo. A produced highly concordant sequence logos, attributing particular importance to the two K50 motifs required for promoter activity (fig. S9). Mutations of the highest-scoring nucleotides in all models (in motifs #2 and #4) result in a severe reduction of promoter activity (fig. S9). These findings demonstrate that cis-regulatory models from evolutionarily distant species are able to predict functionally important motifs—and even individual nucleotides—with great precision.

To evaluate the similarity and relatedness of models across species, we quantified the pairwise distances between models and used the resultant data to hierarchically cluster them. To accomplish this task, we first used the models to score all possible 11-mers, extracted the top 200 highest-scoring 11-mers for each model, and combined them to define a set of 4276 unique 11-mers. We found that 53.8% (2300 of 4276) of these 11-mers contain EC motifs, indicating that these models capture grammar features identified in the preceding section as well as additional features not detected by our motif-based approach. Next, we measured the pairwise distance between models by calculating the Pearson correlation coefficient between each model's scores for the 4276 11-mers (Fig. 5B). We then hierarchically clustered the models using one minus the Pearson correlation coefficient as a distance metric. The resulting hierarchical clustering revealed that the models group by cell class and not by

species (Fig. 5C). The cell-class clusters aggregate into two superclusters, one comprising photoreceptor and bipolar cell models and the other comprising all other models. This higher-order grouping likely reflects the fundamental distinction between grammars dominated by the presence of K50 homeodomain–binding site motifs (photoreceptor and bipolar cell grammars) and those that are not. Consistent with these results, we calculated a silhouette score—a metric used to evaluate the stability and robustness of clusters—and found that it culminated with the formation of six clusters (fig. S10). In summary, the robust coclustering of retinal cell–class models from both jawed and jawless species highlights the deep evolutionary conservation of retinal cis-regulatory codes.

DISCUSSION

The fundamental cell class architecture of the vertebrate retina has remained largely unchanged over more than 500 Ma of evolution. Using single-cell chromatin accessibility profiles of retina from lamprey, zebrafish, goldfish, chicken, mouse, and human, we investigated cis-regulatory grammar and evaluated cross-species homology at the resolution of the cell class. Cross-species comparison of cis-regulatory grammar features revealed deep conservation of class-specific motif vocabulary but little preservation of higher-order syntax between jawed and jawless species. We identified between two and four EC motifs for each retinal cell class, underscoring the combinatorial nature of eukaryotic cis-regulatory codes (*33*). Although we did not detect consistent patterns of motif spacing or orientation between jawed and jawless species, the central enrichment of motifs within OCRs results in a tendency for co-occurring motifs to cluster near the OCR summit. Pairwise comparison of machine learning models of cis-regulatory grammar demonstrated close similarity of models within retinal cell classes, highlighting the evolutionary antiquity of vertebrate retinal cis-regulatory codes. We also observed higher-order grouping of models, possibly reflective of deeper sister relationships among retinal cell classes as previously demonstrated for photoreceptors and bipolar cells (*11*). Overall, these findings confirm

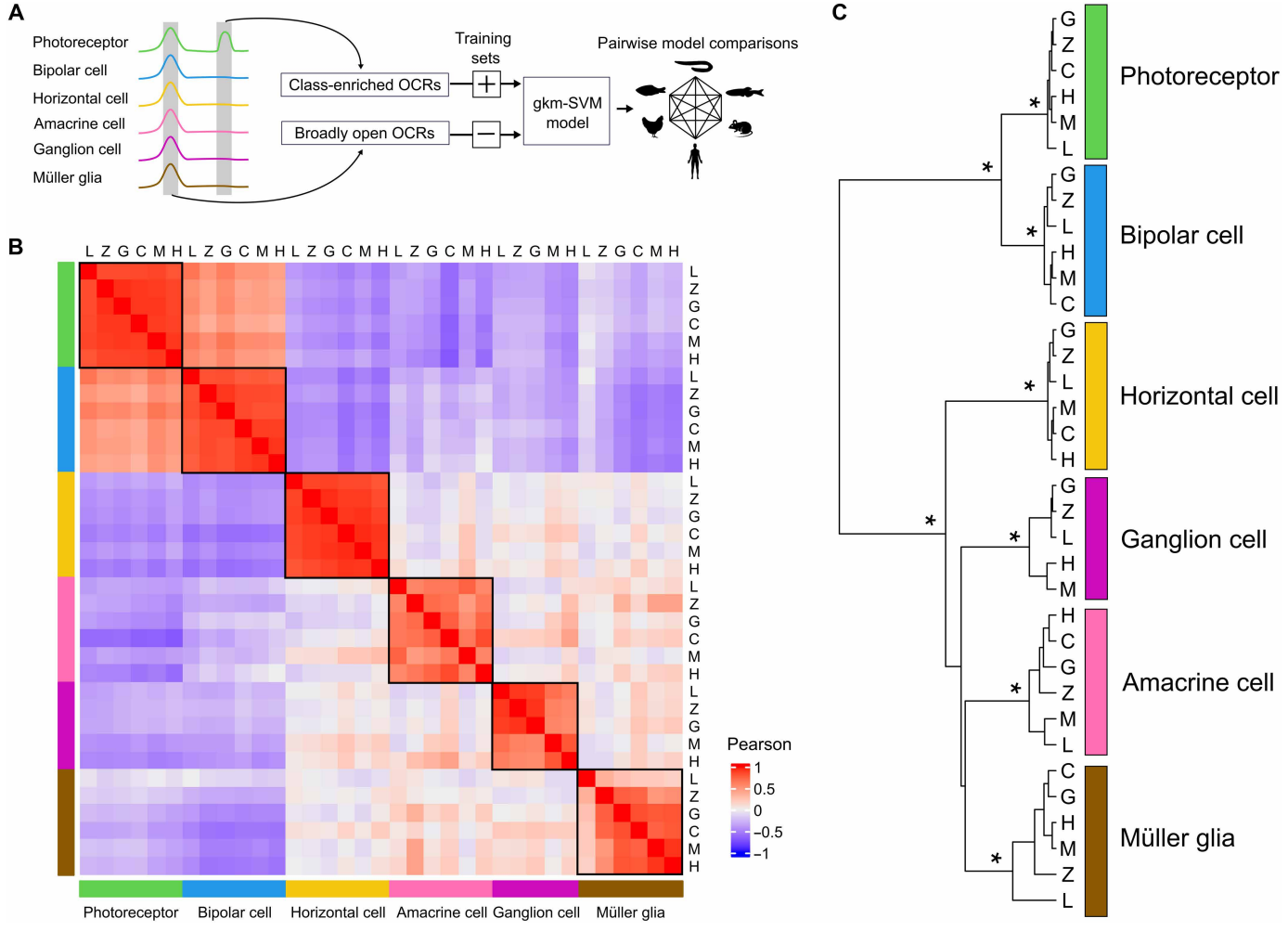

Fig. 5. Retinal cell class–specific cis-regulatory grammars are conserved between jawed and jawless species. (**A**) Schematic showing methodology used to generate and compare gkm-SVM models of cis-regulatory grammar. (**B**) Heatmap showing Pearson correlation coefficients between gkm-SVM models. Models for a given cell class are enclosed by black boxes. (**C**) Hierarchical clustering of the gkm-SVM models. Branch nodes with a statistical significance of $P < 0.01$ are denoted by an asterisk (approximately unbiased P value for selective inference by bootstrap resampling analysis followed by a multiscale resampling). L, lamprey; Z, zebrafish; G, goldfish; C, chicken; M, mouse; and H, human.

that the six class-level cis-regulatory codes controlling vertebrate retinal gene expression arose in the common ancestor of extant vertebrates more than half a billion years ago and persist despite near-total enhancer replacement.

Our study is distinctive in several respects, including the evolutionary depth of the cis-regulatory comparisons (i.e., between jawed and jawless species), the inclusion of all major cell classes of the study tissue, the revelation of a conserved motif-based grammar with central enrichment in OCRs, and the demonstration of perfect coclustering of class-specific cis-regulatory grammar models across all study species. Prior work comparing vertebrate retinal cell types by scRNA-seq demonstrated deep conservation of transcriptomic signatures in the six major retinal cell classes, including between jawed and jawless species (*7*, *9*). However, the present study performed comparative cis-regulatory analysis of retina cell classes across a wide range of vertebrates using open chromatin data. Comparative analyses of cis-regulatory regions in other vertebrate tissues (e.g., brain, heart, and liver) have shown a progressive decay of enhancer sequence alignability with evolutionary distance (*34*–*37*), as also shown here (Fig. 2). Some of those prior studies also found evidence of conserved enhancer function despite sequence turnover (*34*, *37*). One cross-species analysis of cerebellar cell types (*34*) is especially comparable to the present study in several respects: It focused on a morphologically conservative part of the central nervous system; it included all major tissue cell types; it revealed conservation of motif enrichment and co-occurrence patterns within type-specific cis-regulatory regions; and it showed that sequence-based models trained on OCRs can accurately predict the cell-type specificity of accessibility in other species, implying conservation of the underlying cis-regulatory grammars. Thus, the fundamental features of cis-regulatory grammar conservation appear to be similar in diverse parts of the vertebrate central nervous system. Nevertheless, the present work is distinctive, if not unique, in imputing cis-regulatory features for all major retinal cell classes onto the common ancestor of extant vertebrates, while most prior studies were limited to comparisons among mammalian species (*34*, *36*) or amniotes (*35*, *37*). Furthermore, unlike some prior studies using deep-learning models, our work delves into the underlying features of cis-regulatory grammar to reveal a deeply conserved pattern of high-affinity TF-binding site enrichment within the central region of OCRs.

In the course of evolution, novel cell types can arise via "duplication" of a single ancestral cell type into two descendant daughter cell types, which subsequently evolve distinctive cellular features via a process known as "individuation" (*2*). Shared cellular features may arise via evolutionary convergence in cell types derived from remote lineages. Thus, the expression of effector genes controlling cell type–specific features—for example, the expression of opsins in photoreceptor types—is often a poor guide to the underlying evolutionary relationships of cell types. In contrast, the transcriptional regulatory networks that control expression of effector genes often persist for long evolutionary periods and are therefore more stable objects for evolutionary comparison. For this reason, Arendt and colleagues previously proposed that the presence of sets of terminal selector TFs—dubbed "core regulatory complexes (CoRCs)"—should be used to define cell types and trace their evolutionary origins (*2*, *38*). The most common method for evaluating CoRCs is to measure the expression of TFs in individual cell types. However, most cell types express dozens of TFs, and it can therefore be difficult, without a priori knowledge, to prioritize factors for evolutionary comparison based on expression pattern alone. The present study offers a methodologic solution to this problem: By elucidating the cis-regulatory codes of individual cell types or classes, it is possible to nominate the most likely cognate TFs that bind the enriched motifs that comprise the primary feature of cis-regulatory grammar.

Another reason why cis-regulatory codes may be more reliable guides to deep evolutionary relationships than TF expression alone is that differential TF paralog choice may obfuscate shared patterns of TF usage in homologous cell types across species. For example, in mammals, the Maf family TF *Nrl* is required for rod photoreceptor cell fate determination and gene expression (*39*). Yet, while bird retinas contain rod photoreceptors (*40*), avian genomes lack the *NRL* gene (*41*). Instead, avian rods express another Maf TF *MAFA* (*42*, *43*), which is thought to play an analogous regulatory role to that of *Nrl* in mammals (*43*, *44*). Similarly, while *nrl* is required for rod development in zebrafish larvae, it is dispensable for rod cell fate in adult fish (*45*). This differential requirement might be explained by the fact that zebrafish coexpress two Maf family members in rods, *nrl* and *mafba*, and either of these factors could contribute to maintenance of rod gene expression (*46*). Thus, diverse vertebrates appear to have co-opted different Maf family TFs for the same regulatory purpose in rods. These various Maf family members bind similar motifs (*47*), suggesting that shared patterns of motif enrichment within OCRs could inform evolutionary comparison despite differential paralog usage across species. Thus, in the face of both extensive enhancer turnover and differential TF paralog choice, comparison of cis-regulatory codes may be the most reliable method for defining interrelationships among evolutionarily distant cell types.

The remarkable evolutionary stability of cis-regulatory codes is likely attributable to their instantiation in thousands of enhancers across the genome and the lack of an evolutionary mechanism for coordinate changes in multiple genomic regions. If a mutation in a TF changes its DNA binding preference, it could result in a catastrophic failure of binding across the genome and widespread perturbations of gene expression. Conversely, mutations in an individual enhancer motif might render the enhancer inactive unless its cognate TF undergoes simultaneous changes in its binding preference, yet such changes would, in turn, render the TF incapable of binding other enhancers. For these reasons, it appears that cis-regulatory codes and TF-binding preferences are interlocked and exceptionally resistant to evolutionary change.

MATERIALS AND METHODS

Animal husbandry

Lamprey

Downstream- and upstream-migrant sea lamprey (*P. marinus*) were provided by the Hammond Bay Biological Station of the US Geological Survey, Millersburg, MI, USA. Downstream migrant lamprey was captured by drift net in the St. Marys River while they were in the process of migrating to Lake Huron to begin the parasitic stage of the life cycle. Adult lamprey was captured in tributaries of Lake Huron (Ocqueoc River and Cheboygan River) in the process of their upstream spawning migration. Downstream- and upstream-migrant lamprey were kept in well-aerated tanks in cyclic 12L/12D–hour lighting in accordance with the rules and regulations of the National Institutes of Health (NIH) guidelines for research animals, as approved by the University of California Los Angeles Animal Research Committee (Protocol #14-005; animal welfare assurance #A3196-01).

Chicken

Fertilized specific pathogen–free white leghorn chicken (*G. gallus domesticus*) eggs (Charles River Laboratories) were incubated at 38°C until hatching. Chicks were maintained in groups of three to five individuals under constant illumination from a heat lamp and were provided Purina Start & Grow medicated feed (Purina Animal Nutrition LLC) and tap water ad libitum until retinas were harvested. Chicken husbandry and experimental procedures were carried out in accordance with US NIH guidelines (*48*) and approved by the Washington University in St. Louis Animal Care and Use Committee (protocol #22-0430; animal welfare assurance #D16-00245).

Lamprey retina RNA-seq and transcript annotation

To improve retinal gene annotations in lamprey, Iso-Seq was performed on retinas of upstream-migrants. Lampreys were deeply anesthetized with tricaine methanesulfonate (400 mg/liter; MS-222, E10521, Sigma-Aldrich), decapitated, and enucleated. After removing the anterior chamber and the vitreous body from the eye, the retina was isolated from the retinal pigment epithelium in Hepes-buffered Ames' solution at room temperature, flash frozen in liquid nitrogen, and stored at −80°C. Total RNA was extracted using TRIzol and purified using an RNeasy Mini Kit (Qiagen). A SMRTbell Iso-Seq library was prepared according to the manufacturer's protocol (Pacific Biosciences). The library was then sequenced using a single PacBio Sequel II SMRT cell. Demultiplexed PacBio circular consensus sequences–generated HiFi reads with a predicted accuracy ≥Q20 were first processed using lima in (isoseq3; v.3.4.0; http://isoseq.how/) for 5′ and 3′ primer removal with parameters (--iso-seq --peek-guess). PolyA^{+} tails and artificial concatemers were trimmed and removed using refine (isoseq3) with the parameter (--require-polya), resulting in full-length nonconcatemer reads. Clustering was performed using the partial order alignment algorithm using cluster (isoseq3) with the parameter (--use-qvs). For the identification of unidentified isoforms in the lamprey retina, the resultant high-quality consensus sequences were mapped to the genome (kPetMar1) using minimap2 (*49*) with the following parameters: -ax splice -u f. StringTie (v2.2.1) was then used (with the command line option –L) to assemble a transcriptome based on the mapped Iso-Seq reads. StringTie was then run again (with the command line option "merge -F 0 -T 0") to obtain an updated transcriptome annotation. This annotation incorporated modifications to transcript body definitions from the existing National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) reference (GCF_010993605.1) and previously unidentified transcripts supported by the Iso-Seq reads.

Gene orthology inference

To infer orthologous genes among evolutionarily diverse vertebrate species, we used Orthofinder (*28*), which defines an orthogroup (OG) as a set of orthologous and paralogous genes that can be used as an object for comparative analysis. In principle, an OG contains the set of genes that are descended from a single gene in the last common ancestor of all the species being considered. Several representative vertebrate and nonvertebrate species (vase tunicate, clubbed tunicate, hagfish, brook lamprey, skate, white shark, catshark, medaka, reedfish, gar, frog, coelacanth, green anole, platypus, and elephant) along with the six study species (lamprey, zebrafish, goldfish, chicken, mouse, and human) were used to infer OGs. OGs were defined by setting the ascidian node as a base node, and then hierarchical OGs (HOGs) were defined with a jawless-jawed vertebrate divergence node as a base node. For this purpose, protein sequences were obtained from the NCBI RefSeq or Ensembl databases, except for the lamprey and goldfish. For the latter species, the transcript sequences were retrieved from the genome with the transcript annotation using gffread (version 0.12.7), discarding multi-exon mRNAs that had any intron with a noncanonical splice-site consensus (i.e., not GT-AG, GC-AG, or AT-AC). Next, the sense strand of the gene transcripts was used to predict coding sequences via TransDecoder (version 5.5). This process used the TransDecoder.LongOrfs function with the "-S" option. The longest sequence of predicted amino acids was selected for each gene and used for orthology inference. The updated transcript assembly of the lamprey was used, as described in the preceding section. Orthology inference was conducted using *Orthofinder* (version 2.5.5.2) with the following parameters: -M msa -S blast -A mafft -T fasttree. The species tree was subsequently corrected manually using the -s option. In lieu of the default FastTree tree inference program, IQtree2 (version 2.3.2) was used for the generation of all trees. The resultant phylogenetic trees were subjected to a duplication-loss-coalescence analysis with the rooted gene trees to resolve speciation and gene duplication events. Last, HOGs were identified by setting the jawless-jawed vertebrate divergence node as the root node. The nomenclature of genes in lamprey and goldfish was updated on the basis of the information provided by HOGs. For example, two genes, "NTNG1-NTNG2-1" and "NTNG1-NTNG2-2," in lamprey are paralogs, and both are orthologs of the human genes NTNG1 and NTNG2. The accession numbers of gene annotations are provided in table S7.

Sample collection and library preparation for single-cell sequencing

Lamprey

The retinas of downstream-migrant lamprey were dissected and snap-frozen as described above. Frozen nuclei were extracted using the Chromium Nuclei Isolation Kit (10x Genomics) according to the manufacturer's instructions. In brief, three retinas were dissociated in lysis buffer with a plastic pestle until a homogeneous solution was obtained. Residual tissue debris was removed with the nuclei isolation column followed by centrifugation in debris removal buffer. The supernatant was discarded, and the nuclei pellet was resuspended in wash buffer. The nuclei were pelleted by centrifugation and resuspended in 50 μl of resuspension buffer. A sample of nuclei was stained with propidium iodide and quantified using a hemocytometer. The remaining resuspended nuclei (~18,000 nuclei per library) were used for transposition and loaded into the 10x Genomics Chromium Single Cell system. From a single-nuclei suspension, two replicates of the multiome library were constructed using the Chromium Next GEM Single-Cell Multiome Reagent Kit version 1 (10x Genomics), according to the manufacturer's instructions. The libraries were then subjected to sequencing on the Illumina NovaSeq platform.

Chicken

Newly hatched chicks were killed by carbon dioxide inhalation followed by manual cervical dislocation. Retinas were removed from the eye by dissection, transferred to calcium and magnesium-free Hanks' balanced salt solution (HBSS, Thermo Fisher Scientific), and dissociated into single cells by papain digestion (*11*). A single chick retina was incubated in 800 μl of HBSS with 1.3 mg of papain (Worthington Biochemical Corporation) at 37°C for 10 min. The retina was then dissociated by gentle trituration with a pipette. The

dissociated cells were washed with Dulbecco's modified Eagle's media (Thermo Fisher Scientific) containing 10% fetal bovine serum (Thermo Fisher Scientific) and deoxyribonuclease I (Roche) for 5 min at 37°C. The cells were then pelleted by centrifugation at 1000*g* for 1 min. The cells were resuspended in Dulbecco's phosphate-buffered saline with 1% bovine serum albumin (BSA) and filtered. To isolate nuclei, dissociated cells were subjected to centrifugation at 300*g* for 5 min at 4°C. The cell pellet was then resuspended in 100 μl of lysis buffer {0.1% IGEPAL CA-630, 0.1% Tween-20, and 0.01% digitonin in lysis dilution buffer [10 mM tris-HCl, 10 mM NaCl, 3 mM $MgCl_2$, and 1% BSA (pH 7.4)]}, mixed three times by pipetting, and incubated on ice for 3 min. The nuclei were then added to 1 ml of wash buffer (0.1% Tween-20 in the lysis dilution buffer) and centrifuged at 500*g* for 5 min at 4°C. Last, the nuclei pellet was resuspended in 50 μl of 1× diluted nuclei buffer (10x Genomics) and filtered through a 40-μm Flowmi cell strainer. The nuclei were quantified using a hemocytometer. Nuclei (~22,000) were then resuspended and used for transposition and subsequently loaded into the 10x Genomics Chromium Single Cell system. The scATAC-seq library was constructed using the Chromium-Single Cell ATAC Reagent Kits v1.1 (10x Genomics), according to the manufacturer's instructions. The libraries were subjected to sequencing on the Illumina NovaSeq platform.

Sequencing data preprocessing

Publicly available data were retrieved from previous studies, including snATAC-seq datasets for goldfish (*50*) and mouse (*51*) as well as multiome (snRNA-seq + snATAC-seq) datasets for zebrafish (*52*) and human (*53*). The processed sequencing data were retrieved from the original study in zebrafish, while for the other five species, the raw sequence data were initially processed by the Cell Ranger ATAC version 2.0.0 pipeline or the Cell Ranger ARC version 2.0.0 pipeline (10x Genomics) for read filtering, alignment against the genome, and barcode counting. The genome assemblies used were kPetMar1 (lamprey), GCA_014332655.1 (goldfish), galGal6 (chicken), mm10 (mouse), and hg38 (human). The processed data were loaded into ArchR (version 1.0.2) (*54*) for goldfish, chicken, and mouse, while the lamprey and human multiome data were loaded into Seurat (version 4.0.3) (*55*) and Signac (version 1.3.0) (*56*). Transcriptome annotations used in this study are as follows: lamprey, custom annotation as described in the previous section; goldfish, the previously described custom annotation (*50*); chicken, NCBI *G. gallus* Annotation Release 104; mouse, mm10-2020-A-2.0.0 provided by 10x Genomics; and human, GRCh38-2020-A-2.0.0 provided by 10x Genomics.

snATAC-seq analysis

Lamprey

Multiome (snRNA-seq + snATAC-seq) sequence reads from upstream-migrant lamprey retina were used as input data for analysis by Seurat and Signac, respectively. Low-quality cells were removed if the cell contained <300 expressed genes, if >0.3% of the cell's total gene expression derived from mitochondrial genes, if <1000 fragments were detected, or if the cell had a transcription start site enrichment score <3. Putative doublets were removed on the basis of the presence of >40,000 fragments or >10,000 expressed genes.

For gene expression analysis, the negative binomial regression normalization method implemented in SCTransform (Seurat) was used to standardize the gene expression matrix and reduce gene expression noise. Dimensionality reduction of gene expression was then conducted with RunPCA. The significance of each principal component in the RNA analysis was evaluated manually using elbow plots. One to 30 dimensions were chosen for the subsequent analysis.

A cell-by-region count matrix of ATAC-seq reads overlapping OCRs defined by Cell Ranger ATAC was generated using FeatureMatrix and CreateChromatinAssay (Signac). The count matrix was subjected to normalization and dimensionality reduction using FindTopFeatures (Signac) with a min.cutoff of "q0," RunTFIDF (Signac), and RunSVD (Signac) with default settings. The correlation between total counts and each dimension in the ATAC analysis was visualized with DepthCor (Signac), and the first latent semantic indexing (LSI) component, which captured sequencing depth (technical variation), was filtered.

The nearest neighbors for each cell were identified on the basis of a weighted combination of two modalities (RNA and ATAC) by constructing a weighted nearest neighbor graph using FindMultiModalNeighbors with 1 to 30 dimensions [principal components analysis (PCA)] and 2 to 30 dimensions (LSI). The clusters were determined by modularity optimization using FindClusters with a resolution of 0.2. Cell clusters were assigned to retinal cell classes based on the expression of the following cell-class marker genes (*9*): *RHO* and *GNAT2* (photoreceptor); *SLC17A6_1*, *SLC17A6_2*, *GRIK2_1*, *GRIK2_2*, and *PRKCA* (bipolar cell); *ONECUT1* (horizontal cell); *SLC6A9*, *SLC6A11_1*, *SLC6A11_2*, and *SLC32A1* (amacrine cell); and *RBPMS* (ganglion cell). The Müller glia cluster was manually annotated using CellSelector (Seurat) based on the expression of *GLUL*.

Two technical replicates were independently analyzed. Following cell annotation in each of the two technical replicates, the data were merged into a single dataset. Gene expression data were then normalized, scaled, and subjected to PCA analysis as described above. The clusters were identified by constructing a *k*-nearest neighbor graph with FindNeighbors (Seurat) with 1 to 40 dimensions (PCA), followed by modularity optimization using FindClusters (Seurat) with a resolution of 1.5. The cell embedding was obtained with RunUMAP (Seurat). The cell annotations were transferred from the multimodal clustering analysis conducted in each sample. Last, putative mixed clusters (i.e., those assigned annotations for more than one cell class) were removed from the analysis.

Zebrafish

Multiome (snRNA-seq + snATAC-seq) sequence reads from adult zebrafish retina were retrieved from a repository [see the original study (*52*)] and used as input data for analysis by Seurat and Signac, respectively. Low-quality cells were removed if >5% of the cell's total gene expression derived from mitochondrial genes, if <1000 fragments were detected, or if the cell had a transcription start site enrichment score <2. Putative doublets were removed on the basis of the presence of >6000 fragments or >4000 expressed genes. Cell annotations presented in the original study (*52*) were used.

Goldfish

Barcoded and aligned fragments were used as input data for ATAC analysis by ArchR. A genome-wide tile matrix with insertion counts was calculated on 500-bp nonoverlapping windows using createArrowFiles. Low-quality cells were removed if they had a transcription start site enrichment score of <15 or <1000 fragments. Putative doublets were removed using filterDoublets with a filterRatio of 2. Nuclei with >50,000 fragments were also excluded.

Dimensionality reduction was implemented with addIterativeLSI using the LSI method. Cell clustering was then performed with

addClusters using a shared nearest neighbor (SNN) modularity optimization–based clustering algorithm. A total of eight clusters were identified, comprising rod photoreceptors, cone photoreceptors, bipolar cells, horizontal cells, amacrine cells/ganglion cells, Müller glial cells, microglia, and oligodendrocytes. Cluster annotation was performed on the basis of marker genes identified in the original study (*50*). To distinguish between amacrine and ganglion cells, cell annotations from scRNA-seq data were projected onto cells in the snATAC-seq dataset. First, 501-bp OCRs were generated with addReproduciblePeakSet using the pseudo-bulk ATAC replicates for each cluster. Next, peak sets from each cell class were merged to create a nonredundant union set, which was then subjected to analysis in Signac for normalization and dimensionality reduction, as described above for lamprey. Clusters were identified using FindClusters with a resolution of 2.0. The processed data were then integrated with the scRNA-seq data, where the gene expression profile was processed using a standard protocol in Seurat, and the cell clusters were annotated according to the expression of marker genes identified in the original study (*50*). Cell annotations for the scRNA-seq data were then projected onto the cells of the snATAC-seq dataset using FindTransferAnchors with the cca reduction method, followed by TransferData with the LSI weight reduction method, using 2 to 30 dimensions. Transferred labels were retained on the basis of the most abundant cells within each cluster. Only the five classes of retinal neurons and Müller glia were retained for subsequent analysis.

Chicken

Barcoded and aligned fragments were used as input data for ATAC analysis by ArchR. Low-quality nuclei were removed if they had a transcription start site enrichment score of <15 or <3000 fragments. Putative doublets were removed using the filterDoublets function with a filterRatio of 3. Nuclei with >25,000 fragments were also excluded.

Dimensionality reduction was implemented by the LSI method using addIterativeLSI with two iterations and cluster parameters of resolution of 0.1. Cell clustering was then performed using an SNN modularity optimization–based clustering algorithm, using addClusters with a resolution of 0.4. Retinal cell classes were identified by enrichment of the following marker genes (*57*): *GNAT2* and *RHO* (photoreceptors), *VSX1* and *VSX2* (bipolar cells), *ONECUT3* (horizontal cells), *SLC32A1* (amacrine cells), *RLBP1* (Müller glia), and *OLIG2* (oligodendrocytes). To verify that ganglion cells were absent from our dataset, we projected cell annotations for scRNA-seq data from postnatal day 0 (P0) chicken retina (described below in scRNA-seq analysis) onto the cells in the snATAC-seq data, as described above for goldfish. We failed to detect any ganglion cell clusters. Only the four other classes of retinal neurons and Müller glia were retained for subsequent analysis.

Mouse

Barcoded and aligned fragments from multiple developmental stages [embryonic day 11 (E11), E12, E14, E16, E18, P0, P5, P8, P11, and P14] were pooled and used as input data for analysis by ArchR. Low-quality cells were removed if they had a transcription start site enrichment score <10. Putative doublets were removed using the filterDoublets function with a filterRatio of 2. Nuclei with >30,000 fragments were also excluded.

Dimensionality reduction and cell clustering were performed using addClusters with a resolution of 0.7. A total of fourteen clusters were identified, comprising rod photoreceptors, cone photoreceptors, bipolar cells, horizontal cells/amacrine cells, ganglion cells, Müller glial cells, early cone photoreceptors, early rod photoreceptors, early neurogenic cells, late neurogenic cells, early retinal progenitor cells, and three stages of retinal progenitor cells. Cluster annotation was performed on the basis of marker genes identified in the original study (*51*). To distinguish between horizontal cells and amacrine cells, we projected cell annotations for developmental stage–matched scRNA-seq data (*58*) onto the cells in the snATAC-seq data, as described above for goldfish, except for using FindClusters with a resolution of 0.4. Only mature retinal neurons and Müller glia were retained for subsequent analysis.

Human

Multiome (snRNA-seq + snATAC-seq) data were processed in a manner analogous to that described in the original study (*53*). In brief, for each sample, count matrices for both RNA-seq and ATAC-seq were loaded into ArchR. Low-quality cells were removed if they had <200 RNA transcripts, >0.8% mitochondrial gene transcripts, <3000 ATAC fragments, or a transcription start site enrichment score <7. Putative doublets were removed using the filterDoublets function with a filter ratio of 5. Additional doublets were removed if cells expressed >25,000 RNA transcripts, >7000 genes, or >70,000 ATAC fragments. The remaining nuclei in all preprocessed samples were subsequently merged into a single Seurat object. Gene expression counts were normalized using NormalizeData, scaled using ScaleData, and batch-corrected using Harmony (*59*). Graph-based clustering was then performed on the Harmony-corrected data using the top 20 principal components at a resolution of 0.5. Cluster annotation was performed on the basis of marker genes identified in the original study (*53*). Clusters coexpressing marker genes from different cell classes were excluded; clusters devoid of any marker genes were also excluded. Only the five classes of retinal neurons and Müller glia were retained for subsequent analysis.

scRNA-seq analysis

scRNA-seq data for lamprey, zebrafish, and human derive from multiome datasets described in the preceding section. For mouse, scRNA-seq data and corresponding cell cluster annotations were retrieved from a single-cell expression atlas of adult retina (*60*). For goldfish, the raw sequence data were obtained from a published study (*50*) and processed using Cell Ranger (version 7.1.0; 10x Genomics) for read filtering, alignment against the genome, and barcode counting. Goldfish retinal cell–class annotations were also retrieved from the paper (*50*), and cells that were both retained in our analysis and included in their original cell annotations were used for subsequent analysis. For chicken, scRNA-seq data were generated in the present study from newly hatched (P0) chicken retinas. The raw sequence data were processed using Cell Ranger (version 7.0.0), and the aligned gene expression reads were loaded into Seurat. Low-quality cells were removed if the cell contained <1000 expressed genes or if >10% of the cell's total gene expression derived from mitochondrial genes. Putative doublets were removed on the basis of the presence of >15,000 expressed genes. Count data were normalized using SCTransform (Seurat), and dimensionality reduction was performed using RunPCA (Seurat). The significance of each principal component in the RNA analysis was evaluated manually using elbow plots. One to 40 dimensions (PCA) were used for identifying the clusters with FindNeighbors, followed by modularity optimization with FindClusters using a resolution of 1.2. Retinal cell classes were identified on the basis of the expression of cell-class marker genes (*57*). Raw data acquisition and processing are described in another publication (*61*).

Cell-class meta-gene analysis

To devise a metric for measuring EC cell-class meta-gene activity, differentially expressed genes were first identified using the Wilcoxon rank sum test. scRNA-seq datasets from the six study species (described above) were analyzed using FindAllMarkers (Seurat) with the following parameters: logic.threshold = 0, min.pct = 0.05, and return.thresh = 1. Differentially expressed genes were defined as those which exhibited an average $\log_2$-fold change >0.1, an adjusted P value <0.01, and a percentage of cells expressing the gene >10%. Differentially expressed genes were further filtered for cell class–enriched expression as follows. Pseudo-bulk gene expression was quantified for each cell class using AverageExpression (Seurat), and then tau (τ), a measure of cell-class specificity of gene expression, was calculated (*62*). Tau $(\tau) = \frac{\sum_{i=1}^{N}(1-x_i)}{N-1}$, where N is the number of cell classes, and x_i is the expression profile component normalized by the maximal expression value. Genes with $\tau > 0.6$ were retained for subsequent analysis.

The top 1000 genes with the highest tau index were identified for each cell class in each species and then intersected across the six study species to identify EC differentially expressed genes. To accomplish this task, HOGs were defined using Orthofinder described above in Gene orthology inference. HOGs that contained differentially expressed genes from lamprey and three or more jawed species were retained, and for each species, the cell class–enriched genes included in the HOGs were defined as EC differentially expressed (ecDE) genes. We then retrieved GO annotations (i.e., molecular function terms) from Ensembl BioMart using the R package biomaRt (version 2.64.0).

Next, chromatin accessibility over the promoter and gene body of ecDE genes was quantified using GeneActivity (Signac). Individual gene "activity" (hereafter referred to as "accessibility") scores were aggregated into a single accessibility score, referred to as "cell-class meta-gene accessibility." Meta-gene accessibility was quantified in individual cells for each cell class, and the gene accessibilities of both the meta-genes and the remaining genes were normalized and scaled in accordance with the standard protocol in Signac. The statistical significance of cell-class enrichment of the meta-genes was determined by comparing the cell class exhibiting the highest score for the meta-gene with the second-highest scoring cell class using the Wilcoxon rank sum test, followed by Bonferroni correction, performed using Wilcox.exact in the exactRankTests package in R.

Identifying differentially accessible OCRs

Two replicates of pseudo-bulk ATAC data were generated from subsets of snATAC-seq data for each retinal cell class using addGroupCoverages (ArchR). Consensus OCRs (201-bp peaks) were generated using addReproduciblePeakSet for the replicates for each cell class with the following parameters: extendSummits = 100 and cutOff = 0.1. The peak sets from each cell class were then combined, and the resultant nonredundant union peak set was used for the following analysis. Single-cell differential accessibility tests were performed using FindAllMarkers (Seurat) with the following parameters: only.pos = TRUE, min.pct = 0.01, test.use = LR, latent.vars = nCount_ATAC, and slot = counts. Peaks with an adjusted P value <0.01 and an average $\log_2$-fold change >0.01 were retained. In addition, peaks were retained if their average accessibility in pseudo-bulk ATAC-seq data—calculated with AverageExpression (Seurat)—was four times greater than the average of the average accessibilities from the other cell classes. Furthermore, cell class–enriched peaks were filtered if they did not overlap with peaks called in the corresponding cell type. The resultant peak sets were defined as differentially accessible peaks. Broadly open peaks were defined as all peaks not contained in the union of differentially accessible peaks for each species. Chromatin accessibility of cell class–enriched OCRs was visualized using deepTools (version 3.5.4) (*63*). The number of OCRs used for the following analysis is provided in table S6.

Determining cross-species alignability of cell class–enriched OCRs

The union of all differentially accessible OCRs for each of five study species (lamprey, zebrafish, chicken, mouse, and human) was mapped onto the reference genomes of various vertebrate species. Regions that overlapped with exons were excluded before mapping. The UCSC Genome Browser's LiftOver utility (version 377 in bioconda) (*64*) was used for mapping using a parameter minMatch = 0.5. For this analysis, precomputed reciprocal best-hit whole-genome alignment (rbest.chain) files were downloaded from the UCSC Genome Browser website. Differentially accessible OCRs in zebrafish were mapped from danRer11 onto danRer7 and then mapped onto the reference genomes of other species using the publicly available reciprocal best chain files. Mapping between conspecific reference genomes was conducted using the LiftOver utility with a parameter minMatch = 0.95 and the "over.chain" file. Similarly, lamprey OCRs were mapped from kPetMar1 onto petMar3 with the custom chain file generated using flo (*65*), a UCSC Genome Browser command line wrapper. The evolutionary divergence times of all internal branches were obtained from timetree (*66*).

The decay of sequence mappability (i.e., alignability) was modeled using a fitted Gompertz equation $f(x) = 100e^{\left(\frac{a}{b}\right)[1-\exp(bx)]}$, with divergence time as a variable. Equation fitting was performed using the Gauss-Newton algorithm with starting estimates ($a = 0.001$, $b = 0.01$) and a maximum of 1000 iterations allowed in the nls function in the stats package (R; version 4.1.0). The resultant fitted parameter values were $a = 0.002508772$ and $b = 0.006060751$. The confidence intervals were determined by 10,000 bootstrap samples using the boot_nls function in nlraa (version 0.89, https://github.com/femiguez/nlraa).

Discovery of EC motifs

De novo motif discovery

De novo motifs (position probability matrices) were identified for each species using findMotifsGenome.pl (HOMER v4.11), with parameter -size 201. The choice of this window size (i.e., 201 bp) was based on a study that evaluated the performance of three machine learning models for cis-regulatory sequence detection using sequences from 20 to 600 bp as input (*67*). The authors found that all three classifiers achieved maximal or near-maximal performance at ~200 bp and concluded that typical cis-regulatory elements are approximately the length of a nucleosome footprint and flanking linker sequences. A 201-bp (instead of 200 bp) window size was chosen to have an equal number of bases on either side of the summit to facilitate downstream analysis. Cell class–enriched OCRs were used as target sequences, and the broadly open regions were used as background sequences for each species. Identified motifs were retained for subsequent analysis if the statistical significance of the motif was $<10^{-10}$ and if the motif was present in >2% of the cell class–enriched OCRs.

Motif clustering

The motif clustering approach was adapted from a prior study (*68*). For each cell class, de novo motifs of all study species were subjected to pairwise comparison using Tomtom, with the following parameters: -dist kullback, -motif-pseudo 0.1, and -min-overlap 1. The pairwise comparison similarity values (*E* values) were calculated by multiplying each Tomtom-reported *P* value by the number of target motifs, and a group of $-\log_{10}$(*E* values) for each motif was then used to measure the Pearson correlation coefficient between two motifs. $-\text{Log}_{10}$(*E* values) that exceeded 100 were capped at a maximum value of 100. Hierarchical clustering was performed using the average linkage comparison method with one minus the Pearson correlation coefficient as the distance metric. hclust in the stats package (R; version 4.1.0) was used for clustering. Motif clusters were identified by cutting the resultant dendrogram at a height of 0.9. To eliminate clusters with dissimilar motifs, the median value of the Pearson correlation coefficient among the most significantly enriched motifs from each species within a cluster was calculated, and those clusters with median values ≤0.5 were removed. Motif clusters were defined as EC if they included motifs from lamprey and three or more jawed species. Given the absence of chicken ganglion cells in our dataset, EC motif clusters in ganglion cells were only required to contain motifs from lamprey and two or more jawed species.

To create consensus merged motifs, the most significantly enriched motifs from each species in each retained cluster were subjected to sequence alignment, and for each position, the mean value of the position probability matrix was calculated using mergeMotifs [motifStack (*69*), version 1.38.0]. Undefined flanking regions for each motif were assigned a probability of zero and included in the mean calculation. The merged motifs were converted from a position probability matrix to an information content matrix with a uniform background distribution of nucleotides, and then the motif flanking region was trimmed from either side if the information content for each motif position was <0.05. The resultant merged motifs were then compared with mammalian TF binding motifs in the HOCOMOCO database (v12) (*23*) using Tomtom.

Analysis of cis-regulatory grammar

Motif scanning

To identify occurrences of motifs, OCRs were scanned for motifs using scan_sequences (universalmotif; version 1.16.0; https://github.com/bjmt/universalmotif), retaining the highest-scoring motif if two motifs overlapped. The cutoff thresholds for calling motifs were calculated as the median of thresholds identified by HOMER across species. The background nucleotide frequencies were set to those observed in the corresponding cell class–enriched 201-bp OCRs for each of the six species.

Motif distribution and affinity

To graph motif distributions and affinities within OCRs, the median PWM score and motif density per nucleotide were calculated. The per-nucleotide values were further smoothed using a 101-bp sliding window centered on the nucleotide in question.

Motif co-occurrence

For each pair of motifs, the number of OCRs with at least one pair was counted. Co-occurrences involving two overlapping motifs, regardless of the strand, were excluded. The enrichment of co-occurrence was then calculated as the ratio of the frequency of motif pairs in cell class–enriched OCRs to the frequency of motif pairs in background OCRs. A total of 50,000 background sequences were semirandomly selected from the set of broadly open regions using homer2 bg (HOMER; version 5.1) with parameters: -ikmer 2 -N 50000 -NN 100000000. In this way, the distribution of dinucleotide frequencies in background sequences was matched to that in the target cell class–enriched OCRs.

Motif spacing

To identify preferential spacing between motifs, the distance between the primary and secondary motifs in cell class–enriched OCRs was quantified. Two relative distances to the primary motif were measured: the distance between the 5′ nucleotide of the primary motif and the 3′ nucleotide of the secondary motif, and the distance between the 3′ nucleotide of the primary motif and the 5′ nucleotide of the secondary motif. The distance between the two motifs was determined by selecting the value closest to zero. If the secondary motif was found to be 3′ downstream of the primary sequence, the length of the primary motif was added to this distance, ensuring that the distance was always measured from the 5′ nucleotide of the primary motif. The number of instances of the secondary motif at all nucleotide positions on both strands between ±40 bp relative to the primary motif was tallied.

Nominating cognate TFs for EC motifs

To identify cognate TFs that might bind EC motifs, merged motifs were compared with mammalian TF-binding motifs in the HOCOMOCO database (v12) (*23*) using Tomtom (*22*). To nominate cognate TFs across the six study species, we assumed that homologous TFs belonging to the same HOG (see Gene orthology inference) exhibit similar binding preferences. Zinc finger proteins belonging to the largest OG were excluded from this analysis on account of the extensive divergence of binding preferences among zinc finger proteins, which is well-attested (*70*).

Next, scRNA-seq data were used to identify TFs exhibiting cell class–enriched expression in each of the six study species. Differential expression was determined using the Wilcoxon rank sum test in FindAllMarkers (Seurat) with the following parameters: logfc.threshold = 0, only.pos = FALSE, min.pct = 0.0001, return.thresh = 1, assay = "RNA." Candidate cognate TFs were retained if the corresponding motif similarity value (*q* value in the Tomtom output) was $<10^{-1}$, and the significance of the cell-class enrichment (adjusted *P* value in the output of the differential expression test) was $<10^{-1.5}$.

Training and comparison of gkm-SVM models

Model training and validation

Two hundred–base pair OCRs identified from snATAC-seq data in the six study species were used for training gkm-SVM models (*31*). Differentially accessible OCRs in each of the six retinal cell classes were used as the positive training set, and broadly open OCRs from the same species were used as the negative training set. For each species, the same number of positive and negative OCRs were used for training and testing. The datasets were randomly divided into five equal groups, ensuring that each group was nonoverlapping, and a fivefold cross-validation was performed. Models were trained using LS-GKM (version 0.1.0) (*71*) with the following parameters: a linear gkm kernel without center weight (kernel 2), $L = 11$, $K = 7$, and $C = 1$.

Trained models were used to score all test sets within each cell class of each species. ROC-AUC scores were calculated on the basis of the scores of the corresponding positive and negative test sets using ROCR (version 1.0-11) (*72*). The overall performance of the models was determined by measuring the mean and SD of ROC-AUC

scores from the five independent replicates of the fivefold cross-validation models.

Model interpretation

The contribution score of each nucleotide in the OCRs to the classification was computed using gkmexplain (release version 1.0.0) (*32*). The importance scores for the mouse Gnb3 promoter were calculated for photoreceptor and bipolar cell models of all study species except mouse. Both importance and hypothetical importance scores were calculated using the gkmexplain command, and the importance scores were normalized to the hypothetical importance score per developer's recommendation. The resultant normalized importance scores were averaged among five model replicates obtained by fivefold cross-validation (described above).

Model clustering

All possible 11-mer sequences were scored with the gkm-SVM models using the gkmpredict command. The resultant scores were averaged among five model replicates obtained by fivefold cross-validation for each cell class of each species. The 200 highest-scoring 11-mers were selected for each model and merged to create a union of the best-scoring 11-mers (4276 11-mers in total). Agglomerative hierarchical clustering was performed using one minus the Pearson correlation coefficient as a distance metric and Ward's minimum variance method, using hclust in the stats package (R; version 4.1.0). The statistical significance of a branching node was calculated using the pvclust package (version 2.2-0, https://github.com/shimo-lab/pvclust), where the approximately unbiased *P* value for selective inference (*P* value) was calculated by bootstrap resampling analysis followed by a multiscale resampling implemented in pvclust. The tree was visualized using the ggtree package (version 3.10.1). The silhouette score of each gkm-SVM model was measured using silhouette in the cluster package (version 2.1.2). The resulting scores were averaged across models within a cluster and subsequently averaged across clusters to evaluate cluster robustness.

Supplementary Materials

The PDF file includes:

Figs. S1 to S10

Legends for tables S1 to S7

Other Supplementary Material for this manuscript includes the following:

Tables S1 to S7

REFERENCES AND NOTES

1. D. Arendt, P. Bertucci, K. Achim, J. M. Musser, Evolution of neuronal types and families. *Curr. Opin. Neurobiol.* **56**, 144–152 (2019).
2. D. Arendt, J. M. Musser, C. V. H. Baker, A. Bergman, C. Cepko, D. H. Erwin, M. Pavlicev, G. Schlosser, S. Widder, M. D. Laubichler, G. P. Wagner, The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757 (2016).
3. M. A. Tosches, From cell types to an integrated understanding of brain evolution: The case of the cerebral cortex. *Annu. Rev. Cell Dev. Biol.* **37**, 495–517 (2021).
4. M. B. Pomaville, S. M. Sattler, P. B. Abitua, A new dawn for the study of cell type evolution. *Development* **151**, dev200884 (2024).
5. T. D. Lamb, S. P. Collin, E. N. Jr, Evolution of the vertebrate eye: Opsins, photoreceptors, retina and eye cup. *Nat. Rev. Neurosci.* **8**, 960–976 (2007).
6. T. Baden, The vertebrate retina: A window into the evolution of computation in the brain. *Curr. Opin. Behav. Sci.* **57**, 101391 (2024).
7. J. Hahn, A. Monavarfeshani, M. Qiao, A. H. Kao, Y. Kölsch, A. Kumar, V. P. Kunze, A. M. Rasys, R. Richardson, J. B. Wekselblatt, H. Baier, R. J. Lucas, W. Li, M. Meister, J. T. Trachtenberg, W. Yan, Y.-R. Peng, J. R. Sanes, K. Shekhar, Evolution of neuronal cell classes and types in the vertebrate retina. *Nature* **624**, 415–424 (2023).
8. G. L. Fain, Lamprey vision: Photoreceptors and organization of the retina. *Semin. Cell Dev. Biol.* **106**, 5–11 (2020).
9. J. Wang, L. Zhang, M. Cavallini, A. Pahlevan, J. Sun, A. Morshedian, G. L. Fain, A. P. Sampath, Y.-R. Peng, Molecular characterization of the sea lamprey retina illuminates the evolutionary origin of retinal cell types. *Nat. Commun.* **15**, 10761 (2024).
10. D. G. Suzuki, S. Grillner, The stepwise development of the lamprey visual system and its evolutionary implications. *Biol. Rev. Camb. Philos. Soc.* **93**, 1461–1477 (2018).
11. D. P. Murphy, A. E. Hughes, K. A. Lawrence, C. A. Myers, J. C. Corbo, Cis-regulatory basis of sister cell type divergence in the vertebrate retina. *eLife* **8**, e48216 (2019).
12. T. Baden, Ancestral photoreceptor diversity as the basis of visual behaviour. *Nat. Ecol. Evol.* **8**, 374–386 (2024).
13. T. Baden, From water to land: Evolution of photoreceptor circuits for vision in air. *PLOS Biol.* **22**, e3002422 (2024).
14. S. Kim, J. Wysocka, Deciphering the multi-scale, quantitative cis-regulatory code. *Mol. Cell* **83**, 373–392 (2023).
15. M. Levine, Transcriptional enhancers in animal development and evolution. *Curr. Biol.* **20**, R754–R763 (2010).
16. T. Baden, T. Euler, P. Berens, Understanding the retinal basis of vision across species. *Nat. Rev. Neurosci.* **21**, 5–20 (2020).
17. A. E. Hughes, J. M. Enright, C. A. Myers, S. Q. Shen, J. C. Corbo, Cell type-specific epigenomic analysis reveals a uniquely closed chromatin architecture in mouse rod photoreceptors. *Sci. Rep.* **7**, 43184 (2017).
18. A. E. O. Hughes, C. A. Myers, J. C. Corbo, A massively parallel reporter assay reveals context-dependent activity of homeodomain binding sites in vivo. *Genome Res.* **28**, 1520–1531 (2018).
19. O. R. Bininda-Emonds, Fast genes and slow clades: Comparative rates of molecular evolution in mammals. *Evol. Bioinform. Online* **3**, 59–85 (2007).
20. R. A. Gibbs, G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren, S. Scherer, G. Scott, D. Steffen, K. C. Worley, P. E. Burch, G. Okwuonu, S. Hines, L. Lewis, C. DeRamo, O. Delgado, S. Dugan-Rocha, G. Miner, M. Morgan, A. Hawes, R. Gill, Celera, R. A. Holt, M. D. Adams, P. G. Amanatides, H. Baden-Tillson, M. Barnstead, S. Chin, C. A. Evans, S. Ferriera, C. Fosler, A. Glodek, Z. Gu, D. Jennings, C. L. Kraft, T. Nguyen, C. M. Pfannkoch, C. Sitter, G. G. Sutton, J. C. Venter, T. Woodage, D. Smith, H. M. Lee, E. Gustafson, P. Cahill, A. Kana, L. Doucette-Stamm, K. Weinstock, K. Fechtel, R. B. Weiss, D. M. Dunn, E. D. Green, R. W. Blakesley, G. G. Bouffard, P. J. De Jong, K. Osoegawa, B. Zhu, M. Marra, J. Schein, I. Bosdet, C. Fjell, S. Jones, M. Krzywinski, C. Mathewson, A. Siddiqui, N. Wye, J. McPherson, S. Zhao, C. M. Fraser, J. Shetty, S. Shatsman, K. Geer, Y. Chen, S. Abramzon, W. C. Nierman, P. H. Havlak, R. Chen, K. J. Durbin, R. Simons, Y. Ren, X. Z. Song, B. Li, Y. Liu, X. Qin, S. Cawley, K. C. Worley, A. J. Cooney, L. M. D'Souza, K. Martin, J. Q. Wu, M. L. Gonzalez-Garay, A. R. Jackson, K. J. Kalafus, M. P. McLeod, A. Milosavljevic, D. Virk, A. Volkov, D. A. Wheeler, Z. Zhang, J. A. Bailey, E. E. Eichler, E. Tuzun, E. Birney, E. Mongin, A. Ureta-Vidal, C. Woodwark, E. Zdobnov, P. Bork, M. Suyama, D. Torrents, M. Alexandersson, B. J. Trask, J. M. Young, H. Huang, H. Wang, H. Xing, S. Daniels, D. Gietzen, J. Schmidt, K. Stevens, U. Vitt, J. Wingrove, F. Camara, M. M. Albà, J. F. Abril, R. Guigo, A. Smit, I. Dubchak, E. M. Rubin, O. Couronne, A. Poliakov, N. Hübner, D. Ganten, C. Goesele, O. Hummel, T. Kreitler, Y. A. Lee, J. Monti, H. Schulz, H. Zimdahl, H. Himmelbauer, H. Lehrach, H. J. Jacob, S. Bromberg, J. Gullings-Handley, M. I. Jensen-Seaman, A. E. Kwitek, J. Lazar, D. Pasko, P. J. Tonellato, S. Twigger, C. P. Ponting, J. M. Duarte, S. Rice, L. Goodstadt, S. A. Beatson, R. D. Emes, E. E. Winter, C. Webber, P. Brandt, G. Nyakatura, M. Adetobi, F. Chiaromonte, L. Elnitski, P. Eswara, R. C. Hardison, M. Hou, D. Kolbe, K. Makova, W. Miller, A. Nekrutenko, C. Riemer, S. Schwartz, J. Taylor, S. Yang, Y. Zhang, K. Lindpaintner, T. D. Andrews, M. Caccamo, M. Clamp, L. Clarke, V. Curwen, R. Durbin, E. Eyras, S. M. Searle, G. M. Cooper, S. Batzoglou, M. Brudno, A. Sidow, E. A. Stone, J. C. Venter, B. A. Payseur, G. Bourque, C. López-Otín, X. S. Puente, K. Chakrabarti, S. Chatterji, C. Dewey, L. Pachter, N. Bray, V. B. Yap, A. Caspi, G. Tesler, P. A. Pevzner, D. Haussler, K. M. Roskin, R. Baertsch, H. Clawson, T. S. Furey, A. S. Hinrichs, D. Karolchik, W. J. Kent, K. R. Rosenbloom, H. Trumbower, M. Weirauch, D. N. Cooper, P. D. Stenson, B. Ma, M. Brent, M. Arumugam, D. Shteynberg, R. R. Copley, M. S. Taylor, H. Riethman, U. Mudunuri, J. Peterson, M. Guyer, A. Felsenfeld, S. Old, S. Mockrin, F. Collins, Genome sequence of the brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
21. S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, C. K. Glass, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
22. S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, W. S. Noble, Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).
23. I. V. Kulakovskiy, I. E. Vorontsov, I. S. Yevshin, R. N. Sharipov, A. D. Fedorova, E. I. Rumynskiy, Y. A. Medvedeva, A. Magana-Mora, V. B. Bajic, D. A. Papatsenko, F. A. Kolpakov, V. J. Makeev, 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 (2018).
24. C. L. Freund, C. Y. Gregory-Evans, T. Furukawa, M. Papaioannou, J. Looser, L. Ploder, J. Bellingham, D. Ng, J. A. Herbrick, A. Duncan, S. W. Scherer, L. C. Tsui,

A. Loutradis-Anagnostou, S. G. Jacobson, C. L. Cepko, S. S. Bhattacharya, R. R. McInnes, Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor. *Cell* **91**, 543–553 (1997).
25. T. Furukawa, E. M. Morrow, C. L. Cepko, *Crx*, a novel *otx*-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. *Cell* **91**, 531–541 (1997).
26. E. A. Bassett, V. A. Wallace, Cell fate determination in the vertebrate retina. *Trends Neurosci.* **35**, 565–573 (2012).
27. E. Petridou, L. Godinho, Cellular and molecular determinants of retinal cell fate. *Annu. Rev. Vis. Sci.* **8**, 79–99 (2022).
28. D. M. Emms, S. Kelly, OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
29. J. C. Corbo, K. A. Lawrence, M. Karlstetter, C. A. Myers, M. Abdelaziz, W. Dirkes, K. Weigelt, M. Seifert, V. Benes, L. G. Fritsche, B. H. Weber, T. Langmann, CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Res.* **20**, 1512–1525 (2010).
30. Y.-H. Huang, A. Jankowski, K. S. E. Cheah, S. Prabhakar, R. Jauch, SOXE transcription factors form selective dimers on non-compact DNA motifs through multifaceted interactions between dimerization and high-mobility group domains. *Sci. Rep.* **5**, 10398 (2015).
31. M. Ghandi, D. Lee, M. Mohammad-Noori, M. A. Beer, Enhanced regulatory sequence prediction using gapped k-mer features. *PLOS Comput. Biol.* **10**, e1003711 (2014).
32. A. Shrikumar, E. Prakash, A. Kundaje, GkmExplain: Fast and accurate interpretation of nonlinear gapped k-mer SVMs. *Bioinformatics* **35**, i173–i182 (2019).
33. Z. Wunderlich, L. A. Mirny, Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet.* **25**, 434–440 (2009).
34. I. Sarropoulos, M. Sepp, T. Yamada, P. S. L. Schäfer, N. Trost, J. Schmidt, C. Schneider, C. Drummer, S. Mißbach, I. I. Taskiran, N. Hecker, C. B. González-Blas, N. Kempynck, R. Frömel, P. Joshi, E. Leushkin, F. Arnskötter, K. Leiss, K. Okonechnikov, S. Lisgo, M. Palkovits, S. Pääbo, M. Cardoso-Moreira, L. M. Kutscher, R. Behr, S. M. Pfister, S. Aerts, H. Kaessmann, The evolution of gene regulation in mammalian cerebellum development. bioRxiv 2025.2014.643248 [Preprint] (2025); https://doi.org/10.1101/2025.03.14.643248.
35. N. Hecker, N. Kempynck, D. Mauduit, D. Abaffyová, R. Vandepoel, S. Dieltiens, L. Borm, I. Sarropoulos, C. B. González-Blas, J. D. Man, K. Davie, E. Leysen, J. Vandensteen, R. Moors, G. Hulselmans, L. Lim, J. D. Wit, V. Christiaens, S. Poovathingal, S. Aerts, Enhancer-driven cell type comparison reveals similarities between the mammalian and bird pallium. *Science* **387**, eadp3957 (2025).
36. D. Villar, C. Berthelot, S. Aldridge, T. F. Rayner, M. Lukk, M. Pignatelli, T. J. Park, R. Deaville, J. T. Erichsen, A. J. Jasinska, J. M. Turner, M. F. Bertelsen, E. P. Murchison, P. Flicek, D. T. Odom, Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
37. M. H. Q. Phan, T. Zehnder, F. Puntieri, A. Magg, B. Majchrzycka, M. Antonović, H. Wieler, B.-W. Lo, D. Baranasic, B. Lenhard, F. Müller, M. Vingron, D. M. Ibrahim, Conservation of regulatory elements with highly diverged sequences across large evolutionary distances. *Nat. Genet.* **57**, 1524–1534 (2025).
38. O. Hobert, Regulatory logic of neuronal diversity: Terminal selector genes and selector motifs. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 20067–20071 (2008).
39. A. J. Mears, M. Kondo, P. K. Swain, Y. Takada, R. A. Bush, T. L. Saunders, P. A. Sieving, A. Swaroop, Nrl is required for rod photoreceptor development. *Nat. Genet.* **29**, 447–452 (2001).
40. Y. A. Kram, S. Mantey, J. C. Corbo, Avian cone photoreceptors tile the retina as five independent, self-organizing mosaics. *PLOS ONE* **5**, e8992 (2010).
41. M. Coolen, K. Sii-Felice, O. Bronchain, A. Mazabraud, F. Bourrat, S. Rétaux, M. P. Felder-Schmittbuhl, S. Mazan, J. L. Plouhinec, Phylogenomic analysis and expression patterns of large Maf genes in Xenopus tropicalis provide new insights into the functional evolution of the gene family in osteichthyans. *Dev. Genes Evol.* **215**, 327–339 (2005).
42. H. Ochi, K. Sakagami, A. Ishii, N. Morita, M. Nishiuchi, H. Ogino, K. Yasuda, Temporal expression of L-Maf and RaxL in developing chicken retina are arranged into mosaic pattern. *Gene. Expr. Patterns.* **4**, 489–494 (2004).
43. J. M. Enright, K. A. Lawrence, T. Hadzic, J. C. Corbo, Transcriptome profiling of developing photoreceptor subtypes reveals candidate genes involved in avian photoreceptor diversification. *J. Comp. Neurol.* **523**, 649–668 (2015).
44. J.-W. Kim, H.-J. Yang, A. Oel, M. Brooks, L. Jia, D. Plachetzki, W. Li, W. Allison, A. Swaroop, Recruitment of Rod photoreceptors from short-wavelength-sensitive cones during the evolution of nocturnal vision in mammals. *Dev. Cell* **37**, 520–532 (2016).
45. A. P. Oel, G. J. Neil, E. M. Dong, S. D. Balay, K. Collett, W. T. Allison, Nrl is dispensable for specification of rod photoreceptors in adult zebrafish despite its deeply conserved requirement earlier in ontogeny. *iScience* **23**, 101805 (2020).
46. F. Liu, Y. Qin, Y. Huang, P. Gao, J. Li, S. Yu, D. Jia, X. Chen, Y. Lv, J. Tu, K. Sun, Y. Han, J. Reilly, X. Shu, Q. Lu, Z. Tang, C. Xu, D. Luo, M. Liu, Rod genesis driven by mafba in an nrl knockout zebrafish model with altered photoreceptor composition and progressive retinal degeneration. *PLOS Genet.* **18**, e1009841 (2022).
47. T. K. Kerppola, T. Curran, A conserved region adjacent to the basic domain is required for recognition of an extended DNA binding site by Maf/Nrl family proteins. *Oncogene* **9**, 3149–3158 (1994).
48. *Guide for the Care and Use of Laboratory Animals* (The National Academies Press, ed. 8, 2011).
49. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
50. T. Kon, K. Fukuta, Z. Chen, K. Kon-Nanjo, K. Suzuki, M. Ishikawa, H. Tanaka, S. M. Burgess, H. Noguchi, A. Toyoda, Y. Omori, Single-cell transcriptomics of the goldfish retina reveals genetic divergence in the asymmetrically evolved subgenomes after allotetraploidization. *Commun. Biol.* **5**, 1404 (2022).
51. P. Lyu, T. Hoang, C. P. Santiago, E. D. Thomas, A. E. Timms, H. Appel, M. Gimmen, N. Le, L. Jiang, D. W. Kim, S. Chen, D. F. Espinoza, A. E. Telger, K. Weir, B. S. Clark, T. J. Cherry, J. Qian, S. Blackshaw, Gene regulatory networks controlling temporal patterning, neurogenesis, and cell-fate specification in mammalian retina. *Cell. Rep.* **37**, 109994 (2021).
52. P. Lyu, M. Iribarne, D. Serjanov, Y. Zhai, T. Hoang, L. J. Campbell, P. Boyd, I. Palazzo, M. Nagashima, N. J. Silva, P. F. Hitchcock, J. Qian, D. R. Hyde, S. Blackshaw, Common and divergent gene regulatory networks control injury-induced and developmental neurogenesis in zebrafish retina. *Nat. Commun.* **14**, 8477 (2023).
53. S. K. Wang, S. Nair, R. Li, K. Kraft, A. Pampari, A. Patel, J. B. Kang, C. Luong, A. Kundaje, H. Y. Chang, Single-cell multiome of the human retina and deep learning nominate causal variants in complex eye diseases. *Cell Genom.* **2**, 100164 (2022).
54. J. M. Granja, M. R. Corces, S. E. Pierce, S. T. Bagdatli, H. Choudhry, H. Y. Chang, W. J. Greenleaf, ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
55. Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck III, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, R. Satija, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
56. T. Stuart, A. Srivastava, S. Madad, C. A. Lareau, R. Satija, Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
57. M. Yamagata, W. Yan, J. R. Sanes, A cell atlas of the chick retina based on single-cell transcriptomics. *eLife* **10**, e63907 (2021).
58. B. S. Clark, G. L. Stein-O'Brien, F. Shiau, G. H. Cannon, E. Davis-Marcisak, T. Sherman, C. P. Santiago, T. V. Hoang, F. Rajaii, R. E. James-Esposito, R. M. Gronostajski, E. J. Fertig, L. A. Goff, S. Blackshaw, Single-cell RNA-seq analysis of retinal development identifies NFI factors as regulating mitotic exit and late-born cell specification. *Neuron* **102**, 1111–1126. e5 (2019).
59. I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P. R. Loh, S. Raychaudhuri, Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods* **16**, 1289–1296 (2019).
60. J. Li, J. Choi, X. Cheng, J. Ma, S. Pema, J. R. Sanes, G. Mardon, B. J. Frankfort, N. M. Tran, Y. Li, R. Chen, Comprehensive single-cell atlas of the mouse retina. *iScience* **27**, 109916 (2024).
61. Y. Liu, E. C. Hurley, Y. Ogawa, M. Gause, M. B. Toomey, C. A. Myers, J. C. Corbo, Avian photoreceptor homologies and the origin of double cones. *Curr. Biol.* **35**, 2474 (2025).
62. I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, O. Shmueli, Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
63. F. Ramírez, D. P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dündar, T. Manke, deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
64. L. R. Nassar, G. P. Barber, A. Benet-Pagès, J. Casper, H. Clawson, M. Diekhans, C. Fischer, J. N. Gonzalez, A. S. Hinrichs, B. T. Lee, C. M. Lee, P. Muthuraman, B. Nguy, T. Pereira, P. Nejad, G. Perez, B. J. Raney, D. Schmelter, M. L. Speir, B. D. Wick, A. S. Zweig, D. Haussler, R. M. Kuhn, M. Haeussler, W. J. Kent, The UCSC genome browser database: 2023 update. *Nucleic Acids Res.* **51**, D1188–D1195 (2023).
65. R. Pracana, A. Priyam, I. Levantis, R. A. Nichols, Y. Wurm, The fire ant social chromosome supergene variant Sb shows low diversity but high divergence from SB. *Mol. Ecol.* **26**, 2864–2879 (2017).
66. S. B. Hedges, J. Marin, M. Suleski, M. Paymer, S. Kumar, Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.* **32**, 835–845 (2015).
67. Z. M. Patel, T. R. Hughes, Global properties of regulatory sequences are predicted by transcription factor recognition mechanisms. *Genome Biol.* **22**, 285 (2021).
68. J. Vierstra, J. Lazar, R. Sandstrom, J. Halow, K. Lee, D. Bates, M. Diegel, D. Dunn, F. Neri, E. Haugen, E. Rynes, A. Reynolds, J. Nelson, A. Johnson, M. Frerker, M. Buckley, R. Kaul, W. Meuleman, J. A. Stamatoyannopoulos, Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).
69. J. Ou, S. A. Wolfe, M. H. Brodsky, L. J. Zhu, motifStack for the analysis of transcription factor binding site evolution. *Nat. Methods* **15**, 8–9 (2018).
70. A. Rosanova, A. Colliva, M. Osella, M. Caselle, Modelling the evolution of transcription factor binding preferences in complex eukaryotes. *Sci. Rep.* **7**, 7596 (2017).
71. D. Lee, LS-GKM: A new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198 (2016).

72. T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, ROCR: Visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).

Acknowledgments: We thank M. White, M. Toomey, and L. Volkov for providing insightful comments on the manuscript. We also thank the Genome Technology Access Core (GTAC) in the Department of Genetics at Washington University in St. Louis for preparing the Iso-Seq library and performing next-generation sequencing. We are grateful to N. Johnson and the Hammond Bay Biological Station staff of the US Geological Survey for supplying lampreys. **Funding:** This work was supported by the National Institutes of Health (EY030075 to J.C.C.). **Author contributions:** Conceptualization: Y.O. and J.C.C. Methodology: Y.O., Y.L., A.M., G.L.F., and A.P.S. Investigation: Y.O., Y.L., and C.A.M. Supervision: G.L.F., A.P.S., and J.C.C. Writing—original draft: Y.O. and J.C.C. Writing—review and editing: Y.O., Y.L., C.A.M., A.M., G.L.F., A.P.S., and J.C.C. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All raw single-cell sequencing data and relevant processed data generated in this study have been deposited in the Gene Expression Omnibus (accession #GSE284136). Cell metadata for scRNA-seq and snATAC-seq, hierarchical orthogroups, genomic coordinates of open chromatin regions, evolutionarily conserved motif PWMs, and gkm-SVM models have been deposited in figshare: https://doi.org/10.6084/m9.figshare.28014575.v1. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 14 February 2025
Accepted 12 November 2025
Published 12 December 2025
10.1126/sciadv.adw7681