Title

Bacterial origin of a key innovation in the evolution of the vertebrate eye

Permalink
https://escholarship.org/uc/item/0kw425tn

Journal
Proceedings of the National Academy of Sciences of the United States of America, 120(16)

ISSN
0027-8424

Authors
Kalluraya, Chinmay A
Weitzel, Alexander J
Tsu, Brian V
et al.

Publication Date
2023-04-18

DOI
10.1073/pnas.2214815120

Peer reviewed

PNAS RESEARCH ARTICLE | EVOLUTION OPEN ACCESS

Bacterial origin of a key innovation in the evolution of the vertebrate eye

Chinmay A. Kalluraya[a], Alexander J. Weitzel[a], Brian V. Tsu[a], and Matthew D. Daugherty[a,1]

Edited by Joan Strassmann, Washington University in St. Louis, St. Louis, MO; received August 31, 2022; accepted February 28, 2023

The vertebrate eye was described by Charles Darwin as one of the greatest potential challenges to a theory of natural selection by stepwise evolutionary processes. While numerous evolutionary transitions that led to the vertebrate eye have been explained, some aspects appear to be vertebrate specific with no obvious metazoan precursor. One critical difference between vertebrate and invertebrate vision hinges on interphotoreceptor retinoid-binding protein (IRBP, also known as retinol-binding protein, RBP3), which enables the physical separation and specialization of cells in the vertebrate visual cycle by promoting retinoid shuttling between cell types. While IRBP has been functionally described, its evolutionary origin has remained elusive. Here, we show that IRBP arose via acquisition of novel genetic material from bacteria by interdomain horizontal gene transfer (iHGT). We demonstrate that a gene encoding a bacterial peptidase was acquired prior to the radiation of extant vertebrates >500 Mya and underwent subsequent domain duplication and neofunctionalization to give rise to vertebrate IRBP. Our phylogenomic analyses on >900 high-quality genomes across the tree of life provided the resolution to distinguish contamination in genome assemblies from true instances of horizontal acquisition of IRBP and led us to discover additional independent transfers of the same bacterial peptidase gene family into distinct eukaryotic lineages. Importantly, this work illustrates the evolutionary basis of a key transition that led to the vertebrate visual cycle and highlights the striking impact that acquisition of bacterial genes has had on vertebrate evolution.

horizontal gene transfer | comparative genomics | phylogenetics | vertebrate eye evolution | Lateral gene transfer

Vertebrates have a camera-like eye that allows for precise, spatially resolved vision that has evolved as a result of multiple evolutionary transitions (1). Among these transitions from an invertebrate-type eye to a vertebrate-type eye is physical separation of the cells that allows for light sensing from the cells that enzymatically recycle the light-sensing retinoids (Fig. 1*A*). Light sensing, through light-induced isomerization of rhodopsin-bound 11-cis retinal to all-trans-retinal, occurs in photoreceptor (PR) cells, while the enzymatic regeneration of 11-cis-retinal occurs in physically separated cells that are known as retinal pigment epithelium (RPE) cells (1, 2). (Fig. 1*A*). The separation of light sensing and retinoid regeneration in vertebrates has been suggested as a means by which vertebrates are able to see in low light conditions (3).

The protein that facilitates the physical separation between light sensing and retinoid regeneration in vertebrates is interphotoreceptor retinoid-binding protein (IRBP, also known as retinol-binding protein 3, RBP3). IRBP is the major soluble protein localized within the interphotoreceptor matrix and serves the critical function of shuttling retinoids between PR cells and RPE cells (2, 3, 6). Due to its central role in the visual system, mutations in human IRBP cause a variety of retinal diseases such as retinitis pigmentosa and retinal dystrophy (6). Moreover, IRBP sequence and four repeat domain structure are both highly conserved in vertebrates and have been used to clarify features of vertebrate evolution (7–9). Interestingly, there is no obvious IRBP homolog in most other eukaryotes, leaving the origin of this key innovation in vertebrate eye evolution unresolved.

In this work, we use phylogenetic methods to demonstrate that IRBP arose from an interdomain horizontal gene transfer (iHGT) event from bacteria into the ancestor of all extant vertebrates. We find that vertebrate IRBP protein sequences form a single monophyletic clade that are most similar to bacterial S41 family peptidases based on multiple phylogenetic reconstruction methods. Moreover, we find that the characteristic vertebrate IRBP four-domain protein architecture and three intron-containing gene structure, as well as the loss of protease catalytic residues, evolved soon after acquisition of the single-domain peptidase gene from bacteria. We further identified several additional IRBP homologs in eukaryotic genomes, all of which are found to be evolutionarily distant from vertebrate IRBP in our phylogenetic analyses. Although two of these instances are likely due to

Significance

Since the time of Charles Darwin, explaining the stepwise evolution of the eye has been a challenge. Here, we describe the essential contribution of bacteria to the evolution of the vertebrate eye, via interdomain horizontal gene transfer (iHGT), of a bacterial gene that gave rise to the vertebrate-specific interphotoreceptor retinoid-binding protein (IRBP). We demonstrate that IRBP, a highly conserved and essential retinoid shuttling protein, arose from a bacterial gene that was acquired, duplicated, and neofunctionalized coincident with the development of the vertebrate-type eye >500 Mya. Importantly, our findings provide a path by which complex structures like the vertebrate eye can evolve: not just by tinkering with existing genetic material, but also by acquiring and functionally integrating foreign genes.

Author affiliations: [a]Department of Molecular Biology, University of California, San Diego, La Jolla, CA 92093

Author contributions: C.A.K., A.J.W., and M.D.D. designed research; C.A.K., A.J.W., B.V.T., and M.D.D. performed research; C.A.K., A.J.W., and B.V.T. contributed new reagents/analytic tools; C.A.K., B.V.T., and M.D.D. analyzed data; and C.A.K. and M.D.D. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: mddaugherty@ucsd.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2214815120/-/DCSupplemental.

Published April 10, 2023.

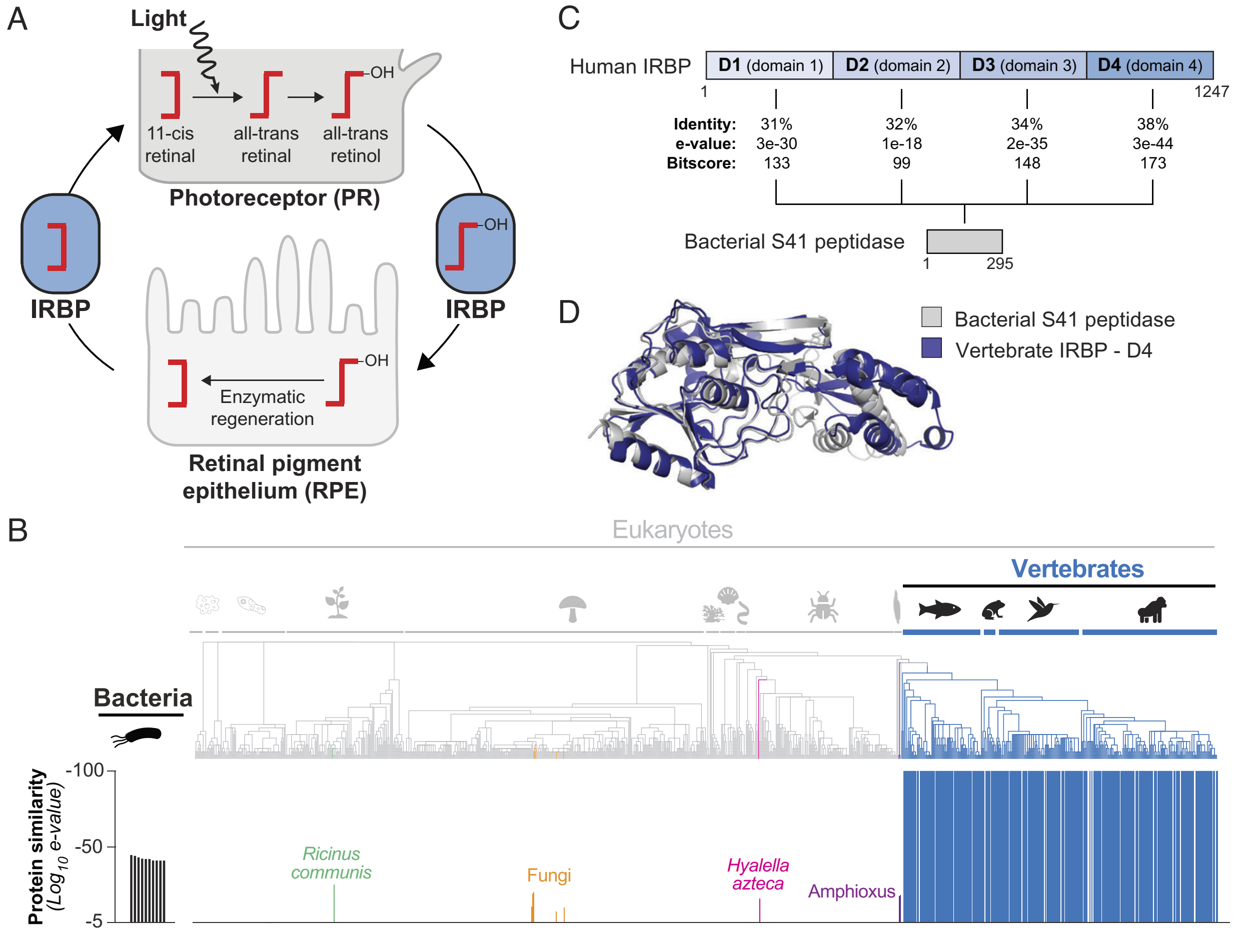

Fig. 1. Discontinuous distribution of IRBP homologs across the tree of life. (*A*) Schematic of the vertebrate visual cycle indicating the physical separation of retinoid-mediated sensing of light in photoreceptor (PR) cells and retinoid regeneration in retinal pigment epithelium (RPE) cells. Interphotoreceptor retinoid-binding protein (IRBP, also known as retinol-binding protein, RBP3) is required to shuttle retinoids between the two cell types. (*B*) Histogram of BLASTp *e*-values obtained after searching the RefSeq protein database for IRBP homologs (Dataset S1 and *SI Appendix, Extended Methods*). Single-domain homologs in nonvertebrate eukaryotes are labeled and colored by species name or species group. Above is a species tree of eukaryotes whose genomes were queried in this study. Species with branches colored gray lack detectable IRBP homologs in their genomes. The *e*-values of the 10 closest bacterial homologs to IRBP are shown to the left. (*C*) Sequence comparison between the individual domains of human IRBP and the top scoring bacterial homolog (complete alignment in *SI Appendix*, Fig. S1). (*D*) Structural comparison of D4 from bovine IRBP (PDB: 7JTI) (4) and a predicted structure of a bacterial homolog that was generated by AlphaFold2 (5).

bacterial contamination in eukaryotic genome assemblies, we also identify two additional independent instances of iHGT of bacterial S41 peptidase genes into eukaryotes, once into fungi and once into amphioxus species (i.e., lancelets). Our work thus identifies recurrent iHGT of bacterial peptidase genes into eukaryotes and establishes the critical impact that acquisition of one such bacterial gene had on punctuating the evolution of the vertebrate eye.

Results

A Bacterial Origin for Vertebrate IRBP. To understand the evolutionary origin of IRBP in vertebrates, we queried the RefSeq database with the human IRBP protein sequence (accession NP_002891.1). Outside of four-domain IRBP homologs in the genomes of most queried vertebrate species, the next most closely related proteins ranked by BLASTp bit score (S) and probability (*e*-value) scores are found in bacterial genomes (Fig. 1*B*, Dataset S1, and *SI Appendix, Extended Methods*). These bacterial homologs, which are annotated as S41 peptidases, are single-domain proteins with sequence and structural similarity to each of the four human IRBP domains (Fig. 1 *C* and *D*, Dataset S1, and *SI Appendix*, Fig. S1). In nonvertebrate eukaryotes, we only identified single-domain IRBP homologs in nine of the 685 genomes we queried (Fig. 1*B* and Dataset S1).

IRBP has previously been noted to have similarity to bacterial proteins (10). In fact, IRBP was among 223 iHGT candidates upon initial sequencing of the human genome (11) and in a subsequent list of 128 putative iHGT candidates in humans (12). However, many of the iHGT claims in both papers were later argued to be due to bacterial contamination in the human genome assembly or, in some cases, misassigned vertical inheritance from the last universal common ancestor (LUCA) of bacteria and eukaryotes followed by gene loss in certain lineages (13–16). In the specific case of IRBP, the presence of a homolog in a plant genome, *Ricinus communis*, was used as evidence that IRBP was present in the LUCA, and that gene loss and sequence divergence led to an incorrect assessment of iHGT (16). These results left the evolutionary origin of vertebrate IRBP, and therefore a critical innovation in the vertebrate visual system, unresolved.

With many more genomes available now than when those studies were published, we turned to phylogenetic reconstruction methods to address the origin of IRBP, as these can be used to distinguish true instances of iHGT from instances of vertical inheritance or bacterial contaminants in eukaryotic genome assemblies (17). Using maximum likelihood phylogenetic reconstruction methods (*SI Appendix, Extended Methods*), we found that vertebrate IRBP proteins form a single monophyletic clade with strong branch support (100%) within the larger phylogeny of bacterial proteins (Fig. 2*A*). Importantly, vertebrate IRBP proteins are phylogenetically distant from IRBP homologs from other eukaryotes, including *R. communis* (Fig. 2*A*). Consistent with iHGT occurring before the radiation of extant vertebrates >500 Mya (18), the vertebrate IRBP sequences follow the known phylogeny of vertebrate species. For instance, we observed that the sea lamprey, *Petromyzon marinus*, a member of the jawless vertebrate lineage that diverged from jawed vertebrates >500 Mya (18), encodes an IRBP sequence that resides in a sister group to IRBP sequences from jawed vertebrates (Fig. 2*B*). We further confirmed that our phylogenetic inferences of a bacterial origin for vertebrate IRBP were robust to choices of phylogenetic analysis software, substitution model, and sequence number (Table 1 and *SI Appendix*, *Extended Methods*, all complete trees with branch support found in Dataset S5). In all cases, the vertebrate IRBP sequences form a clear monophyletic clade nested within bacterial sequences with strong branch support (Fig. 2*A* and Table 1), although the most closely related bacterial sequences vary depending on parameter choice (Dataset S5). Even within a given phylogenetic tree, such as the one shown in Fig. 2*A*, a variety of bacterial orders are represented within the clade nearest to vertebrate IRBP (*SI Appendix*, Fig. S2). This ambiguity surrounding the exact bacterial species or order that contributed to this iHGT event is not surprising given our inference that the transfer event occurred before the radiation of vertebrates >500 Mya, providing ample opportunity for bacterial protein evolution that can complicate phylogenetic inferences. Despite this uncertainty of the exact bacterial species from which the gene was co-opted, our phylogenetic analyses robustly support the inference that vertebrate IRBP arose as the result of an iHGT of a bacterial gene >500 Mya.

Neofunctionalization of IRBP from a Bacterial S41 Peptidase. In contrast to the single-domain S41 peptidases from which it arose, human IRBP is a four-domain tandem repeat protein. Based on our expanded IRBP phylogeny, we infer that IRBP gene structure likely arose early in the evolution of vertebrates soon after the iHGT event from bacteria. Consistent with this inference, the IRBP homolog from sea lamprey, an extant member of the ancient jawless lineage that diverged from jawed vertebrates >500 Mya, consists of a four-domain repeat-like human IRBP, contains three introns in the same position as human IRBP within domain 4 (D4), and is found next to the same gene (*ZNF488*) as human *IRBP* (Fig. 2*C*). Similarly, our analyses of individual vertebrate domains across all vertebrates revealed that all individual domains group together phylogenetically, indicating that the four-domain structure arose once in IRBP evolution, likely soon after iHGT acquisition (Fig. 2*D* and Dataset S5).

One notable exception to the conserved four-domain IRBP protein structure in vertebrates is among members of teleost fishes. Teleosts sporadically encode two-domain, three-domain, or five-domain IRBP proteins in additional to the more common four-domain protein found in other vertebrates (Dataset S1). This apparent discrepancy within otherwise well-conserved vertebrate IRBP has been attributed to a sequence of evolutionary events involving whole-genome duplication at the base of *Teleostei*, gene conversion, and gene loss (7), although the exact origins of multiple IRBP versions in *Teleostei* remain unknown. Regardless, we find that even these domains group within the larger domain tree in a consistent pattern (*SI Appendix*, Fig. S3), indicating that the stereotyped four-domain structure arose via two rounds of gene/domain duplication, followed by lineage-specific gene rearrangements in teleost fish.

Also in contrast to the S41 peptidases from which it arose, human IRBP lacks proteolytic activity (19). Bacterial S41 peptidases, which include C-terminal-processing peptidases, contain a conserved catalytic serine found at the start of an alpha-helix in the active site (20) (Fig. 2*E*). Although the overall structures of vertebrate IRBP and bacterial S41 peptidases are similar (21) (Fig. 1*D*), the catalytic serine is not conserved in the majority of vertebrate IRBP domains, despite well-conserved nearby amino acids (Fig. 2*F*). Ancestral sequence reconstruction using vertebrate IRBP domains, along with IRBP homologs in bacteria and non-vertebrate eukaryotes (*SI Appendix*, Extended Methods), suggests that the single domain ancestor of vertebrate IRBP contained a catalytic serine (*SI Appendix*, Fig. S4). However, after iHGT acquisition and domain duplication, but prior to radiation of extant vertebrates, the serine was replaced by noncatalytic residues in most IRBP domains (*SI Appendix*, Fig. S4), resulting in the current residues that are found in human IRBP domains (Fig. 2*F*). Together, these analyses highlight the neofunctionalization that occurred along the evolutionary path from initial acquisition of a single-domain gene for a bacterial peptidase to a functional vertebrate IRBP.

Independent iHGT of Bacterial S41 Peptidase Genes into Separate Eukaryotic Lineages. The above evolutionary analyses revealed the bacterial origin of vertebrate IRBP. They also revealed the existence of several other IRBP homologs in eukaryotic genomes, all of which are distinct from vertebrate IRBP in our phylogenetic analyses (Fig. 2 *A* and *D*, Table 1, and Dataset S5). We therefore next wished to understand the origin of these other eukaryotic IRBP homologs to determine whether they are best explained by bacterial contamination in eukaryotic genomes, a known issue in many genome assemblies (22), or are independent iHGT events. In the case of the previously published instance of an IRBP homolog in the genome of *R. communis* (16), we found that this protein resides in a region of the phylogenetic tree that contains neither vertebrate IRBP nor any other eukaryotic IRBP homolog (Fig. 2*A*). In addition, the protein is >98% identical to a bacterial protein while being <40% identical to any other eukaryotic protein and is encoded by an intron-less gene that resides on an unplaced <1,600 basepair genome scaffold that is >97% identical to a bacterial chromosome (Fig. 3*A*). All of these characteristics strongly argue that this instance represents either a very recent occurrence of iHGT or, more likely, bacterial contamination in the *R. communis* genome assembly. Similarly, we infer that the IRBP homolog in the genome of the crustacean *H. azteca* (XP_018028457.1) is also the result of bacterial contamination, since the *IRBP* homolog and its flanking genes are all intron less and >90% identical to bacterial proteins (Fig. 3*B*). We also noted that eukaryotic contamination in bacterial genomes can also complicate analyses of iHGT. For instance, although we performed our primary searches against the curated RefSeq database (23), we did find an IRBP homolog from *Bartonella* bacteria (accession AMR68920.1) in the NCBI NR database with 84% sequence identity to human IRBP. However, the *Bartonella* protein is 100% identical to the IRBP homolog from *Mastomys coucha*, the same genus of mouse from which the *Bartonella* was isolated from (24), suggesting that eukaryotic host DNA was misassigned as part of the *Bartonella* genome. These examples underscore how genomic contamination can complicate iHGT inferences and the importance of careful phylogenetic analyses when evaluating any iHGT claims.

Interestingly, IRBP homologs in fungi and amphioxus species appear to be bona fide additional instances of iHGT of bacterial S41 peptidase genes into eukaryotes. The IRBP homologs in fungi have conserved synteny with intron-containing genes found in related fungi (Fig. 3*C*) and form a single monophyletic clade in all of our analyses (Fig. 3*D* and Table 1), both characteristics that

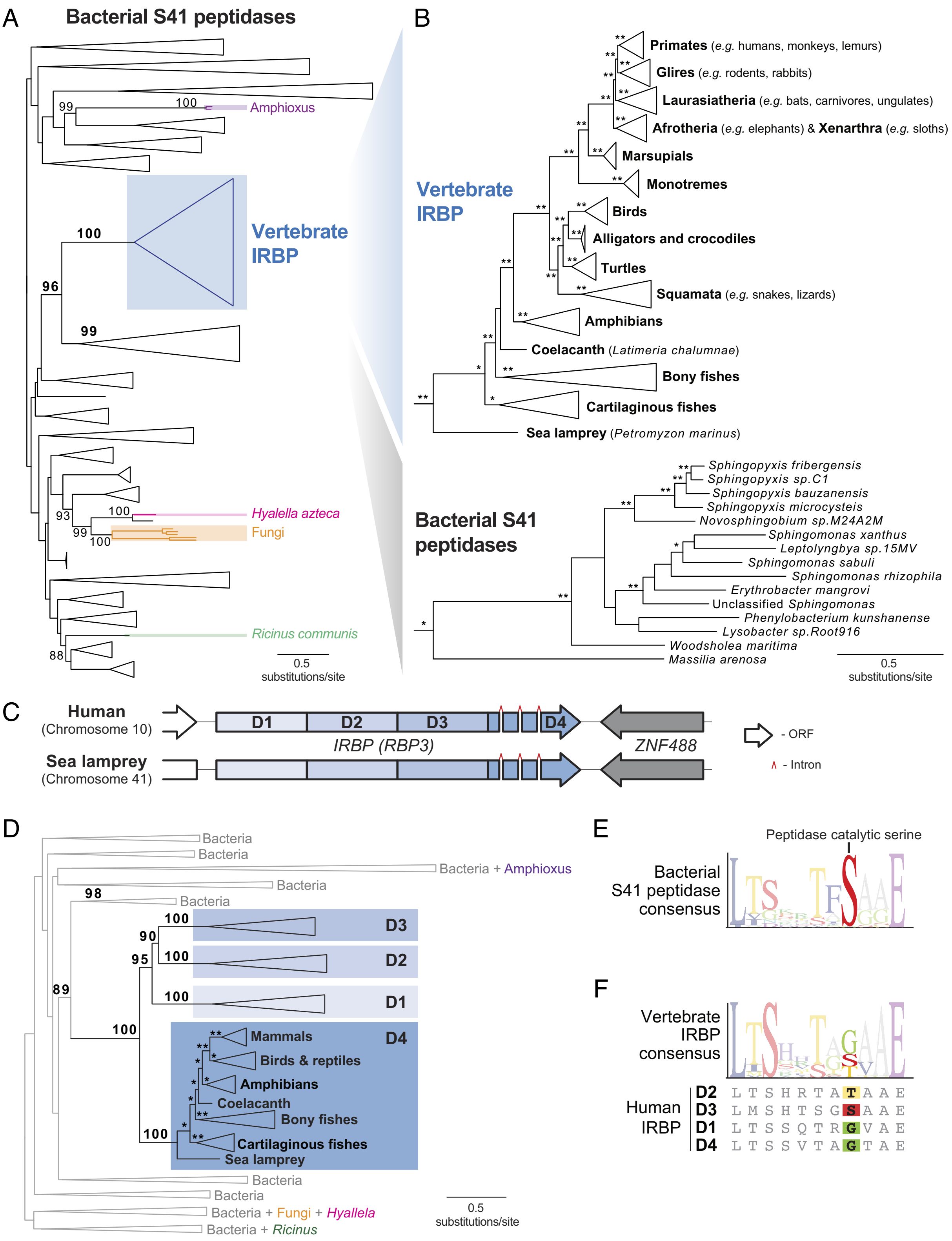

Fig. 2. Vertebrate IRBP originated as a horizontally transferred gene from bacteria. (*A*) Maximum likelihood phylogenetic tree of eukaryotic and bacterial IRBP protein homologs. Relevant bootstrap branch support values are shown. Eukaryotic sequences are indicated by shaded boxes using colors as in Fig. 1*B*. Remaining sequences are bacterial S41 family peptidases. (*B*) Expanded views of the phylogenetic tree shown in Fig. 2*A*. (*Top*) Vertebrate IRBP protein phylogeny, indicating major vertebrate lineages. (*Bottom*) Phylogeny of nearest bacterial IRBP homologs. Asterisks indicate bootstrap branch support (*>80% support, **100% support). (*C*) Schematic comparison of the regions of human chromosome 10 and sea lamprey (*Petromyzon marinus*) chromosome 41 that contain the gene encoding IRBP (gene name RBP3). Positions of IRBP protein domain (D1 through D4) boundaries are shown. The positions of three introns in D4, which are conserved between human and sea lamprey, are shown. Also shown is the presence of ZNF488 as the 3′ neighboring gene that is conserved between human and sea lamprey. (*D*) Phylogenetic tree of individual vertebrate IRBP domains with bacterial and nonvertebrate eukaryotic IRBP homologs. Vertebrate IRBP branches are collapsed by domain or, in the case of D4, major clades of vertebrates. Relevant bootstrap branch support values are shown. (*A*–*D*) A complete list of species and accession numbers is found in Dataset S1, sequence alignments are found in Datasets S2 and S3, and all phylogenetic trees with bootstrap support values are found in Dataset S5. (*E*) Consensus sequence from bacterial IRBP homologs (S41 peptidases) shown in Fig. 2*A*, with the catalytic serine residue indicated. (*F*) Consensus sequence from individual vertebrate IRBP domains, showing the loss of the catalytic serine in most IRBP domains. Below are aligned sequences from the four human IRBP domains.

Table 1. Phylogenetic inferences are robust to different software, sequence number, and substitution models

						Phylogenetic distance from vertebrates			
			Organism-specific clade support			Node count to:		Patristic distance to:	
Software	Sequences	Substitution model	Vertebrates	Amphioxus	Fungi	Bacteria	Amphioxus	Bacteria	Amphioxus
IQ-TREE	972	JTT+I+G4 (*)	100	100	100	3	12	2.36	3.25
IQ-TREE	972	JTT	100	100	100	2	8	2.02	3.02
IQ-TREE	972	LG+I+G4	100	100	100	3	16	2.5	3.89
IQ-TREE	972	LG	100	100	100	4	8	2.21	3.18
IQ-TREE	972	WAG+I+G4	100	100	100	2	22	1.99	3.49
IQ-TREE	972	WAG	100	100	100	4	6	1.9	2.76
IQ-TREE	587	WAG+F+I+G4 (*)	100	100	100	3	14	1.79	3.05
RAxML	587	WAG+Gamma (*)	100	100	100	5	18	2.07	3.49
IQ-TREE	794 (domains)	Q.pfam+G4 (*)	100	100	100	4	9	2.04	3.08
IQ-TREE	524 (domains)	Q.pfam+I+G4 (*)	100	100	100	4	13	1.95	3.15

Phylogenetic trees of eukaryote and bacterial IRBP protein homologs were generated with different indicated parameters (*SI Appendix*, Extended Methods). Asterisks indicating the best-fitting substitution model. For each analysis, branch support values for monophyletic clades are shown, along with the number of nodes and patristic distance (substitutions/site) from the nearest vertebrate IRBP to the nearest bacterial or amphioxus IRBP homolog. The last two rows indicate phylogenetic analyses that were performed using individual vertebrate IRBP domains rather than full-length vertebrate IRBPs. Complete phylogenetic trees with bootstrap support values for all analyses are found in Dataset S5.

support iHGT rather than genomic contamination. These proteins retain the catalytic serine (Datasets S2 and S3 and *SI Appendix*, Fig. S4), suggesting that they may still serve as peptidases in fungi, although further investigation is needed to determine their true function. Similarly, the amphioxus IRBP homologs have characteristics of true eukaryotic genes, as they contain several introns, have conserved synteny across amphioxus species, and form a single monophyletic grouping of proteins (Fig. 3 *E* and *F*). Moreover, gene expression of amphioxus *IRBP* has been observed in pooled larva of *Branchiostoma floridae* (accession GESZ01046803.1) (26), and during the late stage of larval development in *Branchiostoma lanceolatum* (*SI Appendix*, Fig. S5) (27). Because amphioxus species are chordates, like vertebrates and tunicates (28), we considered whether the amphioxus genes represent a separate HGT instance or may instead have derived from the same iHGT event that gave rise to vertebrate IRBP. Several features argue that this was in fact a separate iHGT event. First, despite conservation of the four-domain structure and gene synteny across >500 Mya of vertebrate evolution, neither feature is shared between vertebrate IRBP and amphioxus IRBP (Figs. 2*C* and 3*E*). Second, the positions of introns in vertebrate *IRBP* have been conserved for >500 Mya (Fig. 2*C*), but there is no overlap between vertebrate *IRBP* intron positions and the intron positions in amphioxus *IRBP* (*SI Appendix*, Fig. S6). Third, a shared ancestry of amphioxus and vertebrate IRBP would suggest that we may find IRBP homologs in the genomes of tunicates, as these chordate species share a more recent common ancestor with vertebrates than amphioxus species do (*SI Appendix*, Fig. S7) (28). However, we found no evidence for intact or pseudogenized IRBP homologs in these genomes, despite clear homologs of the retinoid recycling enzyme, retinol dehydrogenase, RDH5 (*SI Appendix*, Fig. S7). Fourth, and most compellingly, we consistently found that vertebrate IRBPs were more phylogenetically related to bacterial IRBP homologs than to amphioxus IRBP homologs, regardless of input alignment or parameter choice for phylogenetic analyses (Table 1). Although it is difficult to completely rule out that a single iHGT event gave rise to vertebrate IRBP and amphioxus IRBP homologs, the above evidence suggests that a separate iHGT event gave rise to IRBP homologs in amphioxus species. Intriguingly, the amphioxus IRBP homologs have undergone protein changes following iHGT from bacteria that parallel vertebrate IRBP. Like vertebrate IRBP, amphioxus IRBP homologs have a predicted N-terminal signal peptide for protein secretion (Fig. 3*G*), reminiscent of another example of independent signal peptide acquisition following several iHGT events into distinct eukaryotic lineages (29). In addition, amphioxus IRBP homologs lack the serine that is required for catalysis in S41 peptidases (Fig. 3*G*). These features suggest that the amphioxus IRBP homologs are secreted proteins that lack proteolytic activity, although the exact function of these proteins will require additional studies in amphioxus species. Taken together, our data indicate that bacterial S41 peptidase genes have been acquired at least three independent times during eukaryotic evolution (Fig. 3*H*), including one instance that gave rise to vertebrate IRBP and its unique role in the vertebrate visual cycle.

Discussion

There is a growing appreciation for the role that iHGT from bacteria to eukaryotes has played in providing novel functionality to specific eukaryotic lineages, including metazoans (30, 31). Here, we describe the striking role of a bacterial gene in one of the long-standing mysteries in evolutionary biology: the evolution of the vertebrate eye. Taking advantage of the large number of available eukaryotic and bacterial genomes, we find unequivocal evidence that a bacterial gene was acquired by a vertebrate ancestor, neofunctionalized, and became the gene that encodes a critical retinoid shuttling protein, IRBP, that allows for the separation of light sensing from retinoid regeneration that is a distinguishing feature of the vertebrate eye. Thus, unlike many other genes in the vertebrate visual cycle that arose from gene duplication and subfunctionalization (3), IRBP required the horizontal acquisition of genetic material from bacteria to facilitate the evolution of novel visual system functionality in vertebrates.

Our analyses also revealed two additional instances of bacterial S41 peptidase genes being acquired by eukaryotes. These instances are phylogenetically distinct from vertebrate IRBP (Fig. 2*A* and Table 1), suggesting convergent acquisition of distinct members of the same bacterial gene family. While the fungal IRBP homologs retain features consistent with the ancestral bacterial S41 peptidase function, amphioxus IRBP homologs have acquired an N-terminal signal peptide and have lost the peptidase catalytic residue similar to vertebrate IRBP. It is tempting to therefore speculate that amphioxus IRBP has been functionally co-opted into the visual cycle of amphioxus similar to

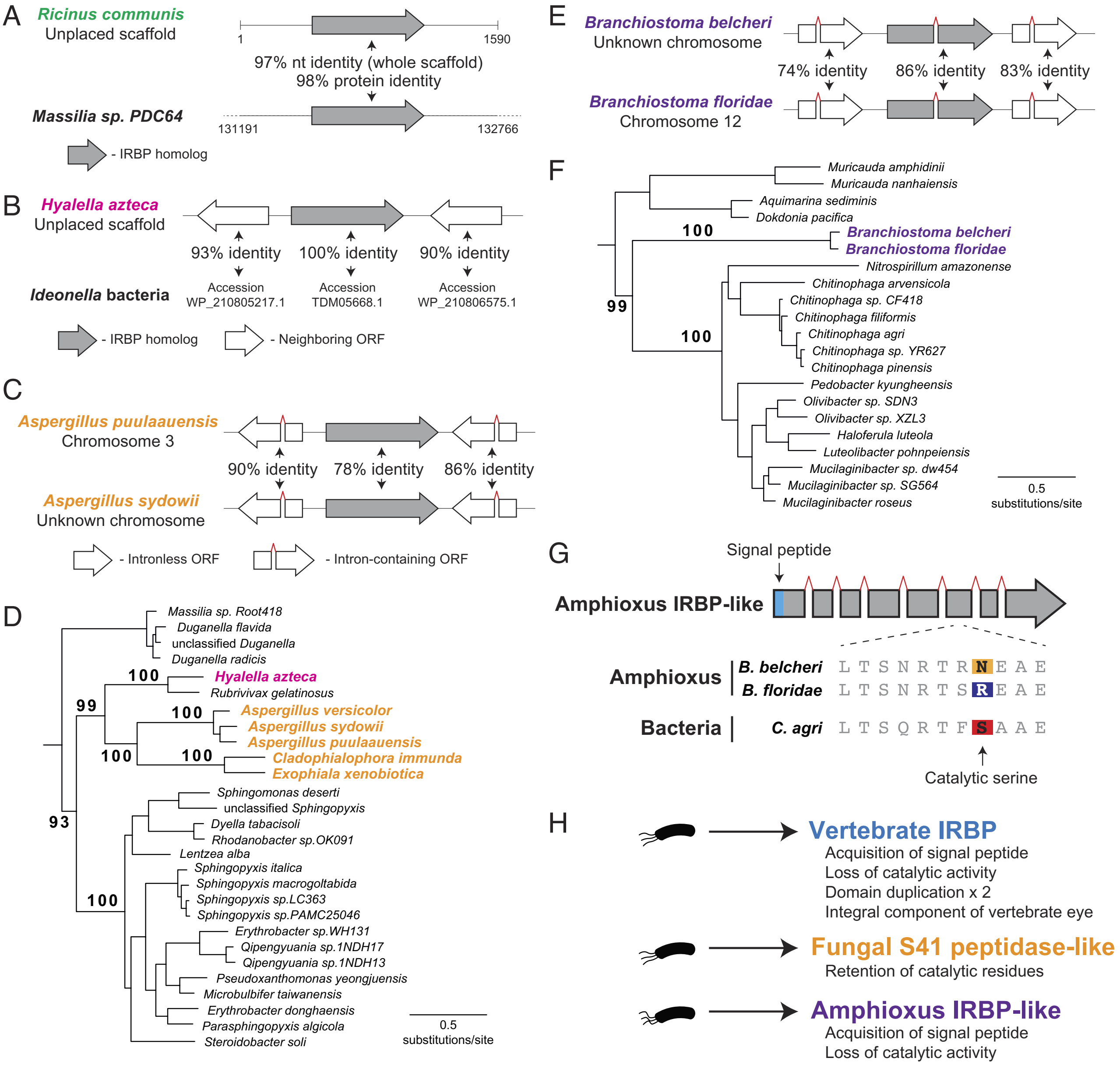

Fig. 3. Independent iHGT of bacterial S41 peptidases in distinct eukaryotic lineages. (*A*) Schematic of the unplaced scaffold in the *R. communis* genome assembly that contains the gene for an IRBP homolog previously used as evidence against a bacterial origin for IRBP (16). No other predicted proteins exist on this scaffold (NW_002999638) and the gene (LOC8272865) contains no introns. The entire *R. communis* DNA scaffold is >97% identical to the indicated region of the *Massilia sp. PDC64* genome and *the R. communis* IRBP homolog is >98% identical to the indicated *M. sp. PDC64* protein. (*B*) Schematic of the unplaced scaffold in the *H. azteca* genome assembly that contains an IRBP homolog (protein accession XP_018028457.1). As in the case of *R. communis*, the gene contains no introns. The two flanking genes also lack introns. Identical or near-identical homologs of all the three proteins are found in *Ideonella* bacteria with protein sequence identity indicated. (*C*) Schematic of the chromosomal region in *A. puulaauensis* that contains an IRBP homolog (protein accession XP_041555295.1). Intron-containing flanking genes are indicated. The closest protein homologs are found in a syntenic region in a related fungal species with protein identity indicated. (*D*) Portion of the phylogenetic tree shown in Fig. 2*A* that contains the IRBP homologs found in fungal genomes. Relevant branch supports are shown. In all analyses (Table 1 and Dataset S5), all fungal homologs group together with 100% branch support, providing further evidence that an additional iHGT event gave rise to IRBP homologs in fungal genomes. (*E*) Schematic of the chromosomal region in *Buccinum belcheri* that contains an IRBP homolog (protein accession XP_019634665.1). The IRBP homolog gene and flanking genes contain introns and are highly similar to a protein in a syntenic region in *B. floridae.* Neither gene that flanks the amphioxus IRBP homolog (encoding proteins most similar to vertebrate BCOR or SULT1A1) is homologous to genes that flank vertebrate IRBPs. (*F*) Portion of the phylogenetic tree shown in Fig. 2*A* that contains the IRBP homologs found in amphioxus genomes. Relevant branch supports are shown. In all analyses (Table 1 and Dataset S5), amphioxus homologs group together and are distant from vertebrate IRBPs, suggestive of a separate iHGT instance that gave rise to amphioxus IRBP homologs. (*G*) Gene structure of the IRBP homolog from *B. belcheri.* Positions of seven introns in the IRBP homolog coding sequence are indicated, none of which overlap with vertebrate IRBP intron positions (*SI Appendix*, Fig. S5). The presence of a predicted N-terminal signal peptide [SignalP 6.0 (25) probability = 0.9996] is indicated. Below are sequences from amphioxus and bacterial proteins that align to the region containing the catalytic serine in bacterial S41 peptidases. (*H*) Summary of three independent iHGT events from bacterial S41 peptidase genes, including the one that gave rise to vertebrate IRBP.

vertebrate IRBP. The amphioxus frontal eye, which develops during the larval stage and persists in adult amphioxus, has been considered to represent the ancestral state of the vertebrate eye structure (32, 33), even though many vertebrate visual cycle proteins do not have obvious reciprocal best-hit orthologs in amphioxus (3). No data exist to resolve whether the amphioxus IRBP homolog is expressed specifically in the frontal eye; however, in *B. lanceolatum*, expression does increase greatly between 15 hpf and 36 hpf (*SI Appendix*, Fig. S5), which mirrors the timing of the development of the amphioxus frontal eye (33).

Functional studies will be required to resolve the role of the IRBP homolog in amphioxus species.

Taken together, our work reveals further evidence that bacterial genes provide a rich source of evolutionary novelty, not just to other bacterial species, but to eukaryotes as well. Unlike evolution of existing genes, or the so-called tinkering, acquisition of foreign genetic material has the potential to punctuate eukaryotic evolution by providing immediate functional novelty. Specifically, the contribution of a bacterial gene to vertebrate eye evolution adds to the complicated and multifaceted evolutionary pathway that has fascinated biologists since Darwin's time. The microbial origin of IRBP, a protein required for a complex vertebrate organ, is reminiscent of the retroviral origin of Syncytin, a protein required for the formation of the mammalian placenta (34). Indeed, a growing list of eukaryotic functions owe their origins to bacterial and viral genes, including components of antiviral and antibacterial immunity, metabolic functions, adaptation to environmental stress, and now an essential component of vertebrate vision (31, 35, 36). As the availability of eukaryotic and prokaryotic genomes continues to grow, it is increasingly likely that we will continue to discover that lineage-specific functions in eukaryotes owe their evolutionary origins to iHGT from bacteria.

Materials and Methods

Human IRBP (also known as RBP3, accession NP_002891.1) was used to query the RefSeq protein database using BLASTp (37) to obtain vertebrate, nonvertebrate eukaryote, and bacterial IRBP homologs. Resulting sequences are shown in Dataset S1. Where indicated, identical or near-identical sequences were removed using CD-HIT (38). Sequences were aligned using Clustal Omega (39), generating alignments found in Datasets S2–S4. Maximum likelihood phylogenetic analyses were performed using IQ-TREE (40) or RAxML (41) as indicated in Table 1 and Dataset S5. Complete phylogenetic trees, including branch support values, are found in Dataset S5. See *SI Appendix, Extended Methods* for additional details of database searching, phylogenetic analyses, structural comparisons, and ancestral sequence reconstruction.

Data, Materials, and Software Availability. All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. We thank members of the Daugherty laboratory, Mark Kunitomi, Michael Perry, Alistair Russell, and Barry Grant for their suggestions and/or manuscript comments. This work was supported by the NIH (R35 GM133633), Pew Biomedical Scholars program, and Burroughs Wellcome Fund Investigators in the Pathogenesis of Infectious Disease program to M.D.D., undergraduate research scholarships from the Halıcıoğlu Data Science Institute and the UC TRELS program to C.A.K., and NIH T32 GM007240 to A.J.W and B.V.T.

1. T. D. Lamb, S. P. Collin, E. N. Pugh, Evolution of the vertebrate eye: Opsins, photoreceptors, retina and eye cup. *Nat. Rev. Neurosci.* **8**, 960–976 (2007).
2. T. G. Kusakabe, N. Takimoto, M. Jin, M. Tsuda, Evolution and the origin of the visual retinoid cycle in vertebrates. *Phil. Trans. R. Soc. B* **364**, 2897–2910 (2009).
3. R. Albalat, Evolution of the genetic machinery of the visual cycle: A novelty of the vertebrate eye? *Mol. Biol. Evol.* **29**, 1461–1469 (2012).
4. A. E. Sears *et al.*, Single particle cryo-EM of the complex between interphotoreceptor retinoid-binding protein and a monoclonal antibody. *FASEB J.* **34**, 13918–13934 (2020).
5. M. Mirdita *et al.*, ColabFold: Making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
6. S. Zeng *et al.*, Interphotoreceptor retinoid-binding protein (IRBP) in retinal health and disease. *Front. Cell. Neurosci.* **14**, 577935 (2020).
7. J. M. Nickerson, R. A. Frey, V. T. Ciavatta, D. L. Stenkamp, Interphotoreceptor retinoid-binding protein gene structure in tetrapods and teleost fish. *Mol. Vis.* **12**, 1565–1585 (2006).
8. C. Poux, E. J. P. Douzery, Primate phylogeny, evolutionary rate variations, and divergence times: A contribution from the nuclear gene IRBP. *Am. J. Phys. Anthropol.* **124**, 1–16 (2004).
9. M. J. Stanhope *et al.*, Mammalian evolution and the interphotoreceptor retinoid binding protein (IRBP) gene: Convincing evidence for several superordinal clades. *J. Mol. Evol.* **43**, 83–92 (1996).
10. P. R. Anbudurai, T. S. Mor, I. Ohad, S. V. Shestakov, H. B. Pakrasi, The ctpA gene encodes the C-terminal processing protease for the D1 protein of the photosystem II reaction center complex. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 8082–8086 (1994).
11. C. International Human Genome Sequencing, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
12. A. Crisp, C. Boschetti, M. Perry, A. Tunnacliffe, G. Micklem, Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biol.* **16**, 50 (2015).
13. S. L. Salzberg, O. White, J. Peterson, J. A. Eisen, Microbial genes in the human genome: Lateral transfer or gene loss? *Science* **292**, 1903–1906 (2001).
14. M. J. Stanhope *et al.*, Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* **411**, 940–944 (2001).
15. D. P. Genereux, J. M. Logsdon Jr., Much ado about bacteria-to-vertebrate lateral gene transfer. *Trends Genet* **19**, 191–195 (2003).
16. S. L. Salzberg, Horizontal gene transfer is not a hallmark of the human genome. *Genome Biol* **18**, 85 (2017).
17. M. Ravenhall, N. Škunca, F. Lassalle, C. Dessimoz, Inferring horizontal gene transfer. *PLoS Comput. Biol.* **11**, e1004095 (2015).
18. S. Kumar, S. B. Hedges, A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).
19. E. A. Gross *et al.*, Prediction of structural and functional relationships of Repeat 1 of human interphotoreceptor retinoid-binding protein (IRBP) with other proteins. *Mol. Vis.* **6**, 30–39 (2000).
20. D. I. Liao, J. Qian, D. A. Chisholm, D. B. Jordan, B. A. Diner, Crystal structures of the photosystem II D1 C-terminal processing protease. *Nat. Struct. Biol.* **7**, 749–753 (2000).
21. A. Loew, F. Gonzalez-Fernandez, Crystal structure of the functional unit of interphotoreceptor retinoid binding protein. *Structure* **10**, 43–49 (2002).
22. V. Lupo *et al.*, Contamination in reference sequence databases: Time for divide-and-rule tactics. *Front Microbiol.* **12**, 755101 (2021).
23. N. A. O'Leary *et al.*, Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–745 (2016).
24. A. Martin-Alonso *et al.*, Bartonella spp. in small mammals, Benin. *Vector Borne Zoonotic Dis.* **16**, 229–237 (2016).
25. F. Teufel *et al.*, SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol* **40**, 1023–1025 (2022).
26. J. X. Yue *et al.*, Conserved noncoding elements in the most distant genera of cephalochordates: The goldilocks principle. *Genome Biol. Evol.* **8**, 2387–2405 (2016).
27. F. Marletaz *et al.*, Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **564**, 64–70 (2018).
28. O. Simakov *et al.*, Hemichordate genomes and deuterostome origins. *Nature* **527**, 459–465 (2015).
29. S. Chou *et al.*, Transferred interbacterial antagonism genes augment eukaryotic innate immune function. *Nature* **518**, 98–101 (2015).
30. J. C. Dunning Hotopp, Horizontal gene transfer between bacteria and animals. *Trends Genet* **27**, 157–163 (2011).
31. F. Husnik, J. P. McCutcheon, Functional horizontal gene transfer from bacteria to eukaryotes. *Nat. Rev. Microbiol.* **16**, 67–79 (2018).
32. T. C. Lacalli, Sensory systems in amphioxus: A window on the ancestral chordate condition. *Brain Behav. Evol.* **64**, 148–162 (2004).
33. P. Vopalensky *et al.*, Molecular analysis of the amphioxus frontal eye unravels the evolutionary origin of the retina and pigment cells of the vertebrate eye. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 15383–15388 (2012).
34. S. Mi *et al.*, Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**, 785–789 (2000).
35. C. Gilbert, C. Feschotte, Horizontal acquisition of transposable elements and viral sequences: Patterns and consequences. *Curr. Opin. Genet Dev.* **49**, 15–24 (2018).
36. T. Wein, R. Sorek, Bacterial origins of human cell-autonomous innate immune mechanisms. *Nat. Rev. Immunol.* **22**, 629–638 (2022), 10.1038/s41577-022-00705-4.
37. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
38. W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
39. F. Sievers *et al.*, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst Biol.* **7**, 539 (2011).
40. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
41. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).