PNAS RESEARCH ARTICLE | EVOLUTION

Retrotransposition facilitated the establishment of a primary plastid in the thecate amoeba *Paulinella*

Victoria Calatrava[a,1,2], Timothy G. Stephens[b,2], Arwa Gabr[c], Devaki Bhaya[a], Debashish Bhattacharya[b], and Arthur R. Grossman[a]

Edited by Joan Strassmann, Washington University in St. Louis, St. Louis, MO; received November 22, 2021; accepted April 1, 2022

The evolution of eukaryotic life was predicated on the development of organelles such as mitochondria and plastids. During this complex process of organellogenesis, the host cell and the engulfed prokaryote became genetically codependent, with the integration of genes from the endosymbiont into the host nuclear genome and subsequent gene loss from the endosymbiont. This process required that horizontally transferred genes become active and properly regulated despite inherent differences in genetic features between donor (endosymbiont) and recipient (host). Although this genetic reorganization is considered critical for early stages of organellogenesis, we have little knowledge about the mechanisms governing this process. The photosynthetic amoeba *Paulinella micropora* offers a unique opportunity to study early evolutionary events associated with organellogenesis and primary endosymbiosis. This amoeba harbors a "chromatophore," a nascent photosynthetic organelle derived from a relatively recent cyanobacterial association (~120 million years ago) that is independent of the evolution of primary plastids in plants (initiated ~1.5 billion years ago). Analysis of the genome and transcriptome of *Paulinella* revealed that retrotransposition of endosymbiont-derived nuclear genes was critical for their domestication in the host. These retrocopied genes involved in photoprotection in cyanobacteria became expanded gene families and were "rewired," acquiring light-responsive regulatory elements that function in the host. The establishment of host control of endosymbiont-derived genes likely enabled the cell to withstand photooxidative stress generated by oxygenic photosynthesis in the nascent organelle. These results provide insights into the genetic mechanisms and evolutionary pressures that facilitated the metabolic integration of the host–endosymbiont association and sustained the evolution of a photosynthetic organelle.

primary endosymbiosis | endosymbiotic gene transfer | organellogenesis | high light–inducible | gene domestication

Significance

Primary endosymbiosis allowed the evolution of complex life on Earth. In this process, a prokaryote was engulfed and retained in the cytoplasm of another microbe, where it developed into a new organelle (mitochondria and plastids). During organelle evolution, genes from the endosymbiont are transferred to the host nuclear genome, where they must become active despite differences in the genetic nature of the "partner" organisms. Here, we show that in the amoeba *Paulinella micropora*, which harbors a nascent photosynthetic organelle, the "copy-paste" mechanism of retrotransposition allowed domestication of endosymbiont-derived genes in the host nuclear genome. This duplication mechanism is widespread in eukaryotes and may be a major facilitator for host–endosymbiont integration and the evolution of organelles.

The evolution of organelles such as plastids and mitochondria had a profound impact on the history of life on Earth. These events originated through primary endosymbiotic associations in which a phagotrophic cell engulfed and retained a prokaryote. During this process, genes are horizontally transferred from the endosymbiont to the host nuclear genome (1–3). Because of inherent differences between prokaryotic and eukaryotic gene structure (e.g., presence/absence of introns, differences in *cis*- and *trans*-acting transcriptional elements), horizontally transferred genes usually remain inactive, and their function and structure erode over time (4). However, some endosymbiont-derived genes, which can encode functions required for endosymbiont/organelle processes, become transcriptionally active and generate proteins that can be routed into the evolving organelle (5). These "replacement" genes in the host nuclear genome encode functionally redundant proteins that allow for the loss of the endosymbiont gene copy. Although the adaptations of these endosymbiont-derived genes to become functional in their new genomic context of the host nuclear genome are key for the maintenance of an evolving organelle, the genetic mechanisms and selective pressures necessary remain largely unknown. This lack of knowledge is primarily a consequence of the ancient origin of the most well-studied primary organelles, plastids, and mitochondria (ca. 1.5 and 2 billion years ago, respectively), which obscures the nature of the early events in their evolution.

In this work, we analyzed the genome and transcriptome of a genetic model for primary endosymbiosis, *Paulinella micropora* KR01 (designated KR01) (6). This amoeba harbors a nascent photosynthetic organelle (chromatophore) that evolved from an α-cyanobacterium that was acquired ~120 million years ago (Mya) through an endosymbiotic event. This event was independent of the primary endosymbiosis that led to the establishment of the canonical Archaeplastida plastid (i.e., in green algae, land plants, red algae, and glaucophyte algae). Endosymbiont-derived genes involved in photoacclimation and photosynthesis, including those encoding high light (HL)–inducible

Author contributions: V.C., T.G.S., D. Bhattacharya, and A.R.G. designed research; V.C., T.G.S., and A.G. performed research; V.C., T.G.S., and A.G. analyzed data; and V.C., T.G.S., A.G., D. Bhattacharya, D. Bhaya, and A.R.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: vcalatrava@carnegiescience.edu.

[2]V.C. and T.G.S. contributed equally to this work.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2121241119/-/DCSupplemental.

Published May 31, 2022.

proteins (*HLI* genes; encoded HLIPs) and photosystem I (PSI) subunits, were transferred to the host nuclear genome, where they were duplicated and acquired transcriptional responses to HL stress that are similar to those in the cyanobacterial donor (6–8). These genes were lost in the endosymbiont genome but have likely retained their functions in alleviating the harmful consequences associated with the absorption of excess light energy by the photosynthetic apparatus. The light-associated stress experienced by the host–endosymbiont partnership is exacerbated by the still incomplete metabolic integration of their metabolisms, which leads to the production of reactive oxygen species (ROS) even when the amoeba is exposed to moderate levels of light (7). This constraint likely resulted in evolutionary pressure for extensive expansion, reorganization, and acquisition of light regulatory features associated with the nuclear established *HLI* gene family in KR01. The mechanism of gene expansion and how these genes acquired transcriptional regulation in response to light stress in the host nuclear genome is not understood.

Here, we analyzed the features associated with the expansion of the endosymbiotic gene transfer (EGT)–derived *PSAE*, *PSAI*, and *HLI* gene families in the nuclear genome of KR01 and the expression patterns of individual genes in these families (6) to elucidate events required for the establishment and control of their physiological functions. Our findings show that RNA-mediated duplications (i.e., retrotransposition) were key for gene family expansion and the acquisition of transcriptional regulatory elements that enable these genes to respond to light stress. Moreover, extensive DNA-based duplications of the *HLI* retrocopied genes facilitated their domestication in the eukaryotic host and enabled them to evolve different expression patterns and optimize their physiological functions, processes critical for the efficient integration of host-organelle physiologies and energetics.

Results

The assembled genome and RNA sequencing (RNA-seq) data (6) available for KR01 were analyzed for the presence of endosymbiont-derived genes in the host genome that were lost from the genome of the endosymbiont and were regulated by HL. Of the ∼50 EGT-derived genes (6), we identified three gene families that met these criteria; these families encode HLIPs and the PSI subunits PSAE and PSAI. These gene families function in photoacclimation and photosynthesis and are regulated by light conditions in cyanobacteria (9–13). We hypothesized that these host-controlled genes are likely functional and may be critical for establishing physiological integration of the photosynthetic ROS-producing endosymbiont with the host (7). The small size of these proteins (<100 amino acid residues) likely allows them to enter the chromatophore, where they perform their function, without the need for a presequence (14, 15). To investigate how these genes were accommodated after being transferred to the host nuclear genome, we analyzed their structure and arrangement on the nuclear genome and their phylogenetic relationships with orthologous genes in various cyanobacteria.

Retrotransposition of Genes Encoding PSI Subunits PSAE and PSAI. The three nuclear copies of *PSAE* in KR01 (*PSAE1*, *PSAE2*, and *PSAE3*) encode proteins that are 53%, 46%, and 72% identical, respectively, to PsaE of *Synechococcus* sp. WH 5701, the cyanobacterium most closely related to the chromatophore donor (*SI Appendix*, Fig. S1). The *PSAE1* and *PSAE2* sequences are incomplete (their 3′ and 5′ termini are missing from the genome assembly, respectively), and they harbor a deletion of seven amino acid residues relative to the orthologous proteins in *Synechococcus* sp. WH 5701 and *Paulinella chromatophora,* a sister lineage to KR01 (*SI Appendix*, Fig. S1), which suggests that these gene copies are not under selection and may be pseudogenes. In contrast, PSAE3 has a high (72%) sequence identity with the *Synechococcus* sp. WH 5701 protein, is full length, and is identical to its ortholog in *P. chromatophora*. *PSAE3* lacks introns and is associated with a downstream poly(A) tract and retrotransposon domains (long terminal repeat [LTR] reverse transcriptase, ribonuclease [RNase] H, and integrase). These features, which are characteristic of gene retrocopies (i.e., genes that were duplicated through retrotransposition) (16), are not observed for either *PSAE1* or *PSAE2* [i.e., they contain introns and lack a downstream poly(A) tract (*SI Appendix*, Fig. S1 *A* and *C*)]. Our observations strongly suggest that *PSAE3* was retrocopied, possibly from *PSAE1*/*PSAE2* or another EGT-derived ancestral copy of the gene that was not retained.

There is a single *PSAI* gene in the KR01 nuclear genome with multiple exons. However, the region encoding the conserved domain of the protein is limited to a single exon. This gene also has a downstream "degraded" poly(A) tract (53% adenine in a 100-nucleotide region; the KR01 genome is, overall, 28.16% adenine) and a retrotransposon-related domain (non-LTR reverse transcriptase). Despite our inability to detect a parental *PSAI* gene in the genome, these observations suggest that the KR01 *PSAI* also originated as a consequence of retrotransposition. The establishment of this retrocopy could have had a selective advantage over the parental gene, enabling the loss of the parental gene. This scenario is in accord with previous findings that showed that more than 90% of retrogenes in green algae and dinoflagellates are orphans (i.e., their parental genes are no longer present in the genome) (17, 18). In summary, it appears that the light-regulated *PSAI* and *PSAE* genes in KR01, exclusively present in the nuclear genome, originated via retrotransposition

Three Phylogenetic Clades Dominate the KR01 HLI Repertoire. In the nuclear genome of photosynthetic *Paulinella* spp., the *HLI* genes comprise a large family with dozens of copies (6–8). Here, we identified 50 *HLI* copies in KR01, of which four are pseudogenes (see *SI Appendix* for further details). Most *HLI* genes formed three clades (Fig. 1 and *SI Appendix*, Fig. S2). For each of these major clades, cyanobacterial *hli* members were found basal to the *Paulinella* members, which suggests that each of the *Paulinella* clades has a distinct polyphyletic origin (i.e., they each evolved from a different cyanobacterial ancestral gene acquired via independent EGT events). However, clades 2 and 3 show low node support (bootstrap [BS] <95%), likely due to the high divergence and small size of these proteins—particularly those included in clade 2 (∼46-aa residues)—and thus, these results must be interpreted cautiously. Nevertheless, we identified conserved amino acid motifs that are specific for each of the three major clades and are also present in cyanobacterial sequences. This result supports the idea that the KR01 HLIs in clades 1, 2, and 3 originated from distinct *hlis* (*SI Appendix*, Fig. S3) that were all likely derived from a close relative of *Synechococcus* sp. WH 5701 (which, among all cyanobacteria, encodes proteins most similar to those in KR01). Clade 1– and 2–specific motifs were conserved across various cyanobacterial species and, in some cases, in phylogenetically distantly related members, including *Synechocystis* sp. PCC 6803 and the HL-adapted ecotypes of *Prochlorococcus marinus* MIT 9312 (*SI Appendix*, Table S1). The

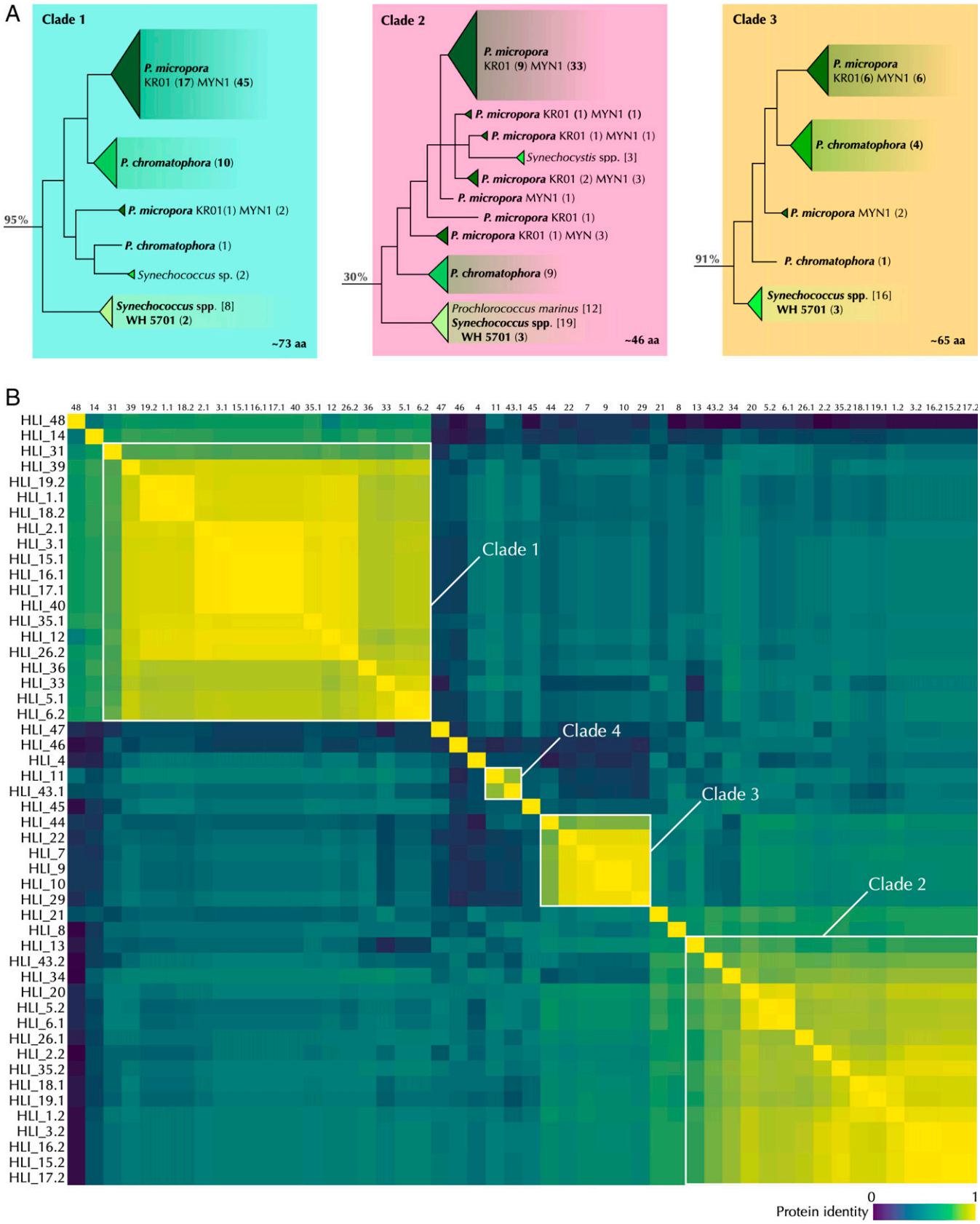

Fig. 1. (*A*) Simplified trees for each of the main *HLI* phylogenetic clades. The number of sequences from each isolate in a collapsed clade is shown in parentheses; if sequences are from multiple species, it is shown in brackets. The percent values are the calculated node support (using 2,000 ultrafast BSs; see *SI Appendix*, Fig. S2 for full maximum likelihood tree and Materials and Methods for full maximum likelihood tree and further details). (*B*) Nucleotide identity matrix of *HLI* gene family in KR01 calculated using Clustal-Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/).

clade 3–specific motif was only found in two *hli* copies in *Synechococcus* sp. WH 5701. Therefore, HLIP proteins associated with clades 1 and 2 are widespread and often represented as gene families in cyanobacteria, whereas clade 3 HLIPs appear to be much less prevalent and have only been identified in the ancestral lineage of the chromatophore. It is not known how important these motifs are to HLIP function or structure; however, they are in some cases present in significantly diverged sequences, suggesting that they have been subject to selection. For instance, the four members of *Synechocystis* sp. PCC 6803 (HliA-D), which are required for growth in HL (9, 19), are highly diverged from the KR01 HLIPs, and yet they harbor the clade 1– (HliA/B) and clade 2– (HliC/D) specific motifs. None of these signatures were found in any of their eukaryotic analogs (LHC superfamily of proteins) analyzed, but surprisingly, they are present in the HLIP-related protein SEPx.2 of the glaucophyte *Cyanophora paradoxa* (*SI Appendix*).

For the three major clades, the HLIP sequences clustered together independently for the two different *Paulinella* species (*P. chromatophora* and *P. micropora* KR01 and MYN1), suggesting that the *HLI* genes underwent duplications after species divergence. Additionally, multiple HLI members clustered into a minor clade (clade 4) that shows some evidence of duplication, but to a lesser extent compared to clades 1 to 3; each *Paulinella* isolate has only two or three gene members in this clade (*SI Appendix*, Fig. S2). Additional clades composed of *Paulinella* and cyanobacterial sequences were identified, but they do not show evidence of gene duplication.

The *HLI* Gene Family Shows Evidence of Retrotransposition in KR01. To understand how the *HLI* gene family expanded in *Paulinella*, we analyzed their structure and arrangement in the KR01 genome. Most *HLI* gene copies contain a single exon with only 7/50 containing multiple exons (Fig. 2). Additionally, nearly one-half of them (26 genes) are arranged in pairs in a divergent or "head-to-head" orientation, as confirmed by both PCR and Sanger sequencing (see *Materials and Methods* for further details). In some cases, unpaired genes were found to be paired head-to-head with pseudogenes (Fig. 2 and *SI Appendix*, Fig. S4; mostly genes from clade 3). We also identified *HLI* gene pairs in *P. micropora* MYN1 (8) and *P. chromatophora* (20) (*SI Appendix*); 74 of the *HLI* genes in MYN1 (62% of the 119 total) are in head-to-head orientation (37 *HLI* pairs), whereas at least eight *HLI* genes in *P. chromatophora* (22.86% of the 35 genome-based gene models) are in head-to-head orientation (four *HLI* pairs). The lower number of *HLI* genes identified in the latter species is probably a consequence of the *P. chromatophora* assembly being highly fragmented.

Most of the identified head-to-head pairs are flanked by conserved repeats; the 3′- end of each *HLI* member of a pair is followed by a downstream poly(A) (>20 bp; *SI Appendix*, Table S2) and long homopyrimidine tract (~300 to 3,000 bp), resulting in homopyrimidine/purine and poly(A/T) tracts flanking each pair, as depicted in Fig. 2 *A*–*C*. These repeats have been associated with retrotransposition (21). The presence of these repeats in the genome was strongly supported by aligned short and long sequence reads that showed no evidence of misassembles and by their presence in ~46% of the identified pairs in MYN1. The long homopyrimidine tracts could not be confirmed in the genome of *P. chromatophora* because of the fragmented assembly, but flanking poly(A) tracts were found to be associated with some of the *HLI* genes.

Curiously, the two *HLI* genes comprising each of the pairs that we identified were never from the same phylogenetic clade. The different pairs included (i) 12 clade 1/clade 2 pairs; (ii) a single clade 2/clade 4 pair; and (iii) four "pseudo-pairs" containing a clade 3 member and a pseudogenized *HLI* gene that most likely originated as a clade 1 member (Fig. 2*E* and *SI Appendix*, Fig. S4). Most pairs (~81%), including the pseudo-pairs, were flanked by the full conserved repeat pattern associated with retrogenes (21) (Fig. 2*B*). This pattern was also observed in some instances where both putative *HLI* genes had been pseudogenized, although in these cases the flanking repeats showed evidence of degradation (*SI Appendix*, Fig. S5*C*). In addition, several *HLI* genes were contiguous on the genome with downstream transposon and retrotransposon sequences, including those encoding reverse transcriptase, RNase H, and integrase domains (*SI Appendix*, Table S2). These observations, together with the single-exon nature of these genes, strongly suggest that these paired *HLI* genes are retrocopies. Almost all of the head-to-head pairs in KR01 showed associated homopyrimidine and poly(A) tracts downstream of at least one member of the pairs, and this genetic arrangement seems to be linked to their retrotransposition. Some of the nonpaired clade 1, 2, and 3 *HLI* genes share high sequence similarity with genes located in pairs and contain downstream (and sometimes also upstream) conserved repeats. This suggests that these unpaired genes were most likely retrocopied; that is, these genes may have originally been in head-to-head pairs that were later fragmented, resulting in the loss of the partner genes. Thus, all *HLI* genes belonging to clades 1 to 4 were likely generated as head-to-head pairs through retrotransposition (i.e., they are putative retrocopies). In contrast, *HLI* gene copies not in head-to-head pairs and lacking downstream poly(A) and homopyrimidine tracts, which also sometimes contain introns, are likely nonretrocopies that are potentially parental to the retrocopies. Alternatively, these are old retrocopies in which the flanking repeats were lost and introns were gained.

Several pairs in KR01, including their repeat patterns, were highly similar and, in some cases, found on the same scaffold, suggesting that they were recently duplicated through a DNA-mediated mechanism. We identified *HLI* genes from *P. chromatophora* in all of the major clades, which suggests that these sequences were acquired before the divergence of the two *Paulinella* lineages. However, the *Paulinella HLI* genes in each clade group with sequences from the same species (i.e., the *HLI* genes within each clade from *P. micropora* KR01 and MYN1 group together, and the *HLI* genes from *P. chromatophora* group together) (Fig. 1*A*). This suggests that the *HLI* pairs expanded independently after speciation; however, in both lineages the clade 1/clade 2 pair represents the most expanded grouping, suggesting they are under similar selective pressures. Within these pairs, the clade 1 partner is more conserved than the clade 2 partner (Fig. 1*B*) despite the greater length of the former (73-aa residues versus 46-aa residues). This suggests that the genes in these pairs are evolving asymmetrically, likely due to more relaxed selective constraints on one of the partners (in this case, the clade 2 member).

EGT-Derived Retrogenes Are Light Responsive. Generally, retrotransposed genes acquire new regulatory elements from sequences adjacent to the genomic insertion site (22). This process provides additional opportunities for gene adaptation through reshaping of their regulatory responses (23). To determine the relationship between the EGT-derived retrocopies identified here and their transcriptional regulation, we analyzed

Fig. 2. The structure and arrangements of *HLI* genes in the KR01 genome. (*A*) Organization of a cluster of *HLI* genes arranged as three pairs in head-to-head orientations and flanked by poly(A) (>20 bp) and homopyrimidine tracts (~300 to 3,000 bp). (*B* and *C*) A closeup view of one of the head-to-head pairs (*B*) showing the size and orientation of the genes and repeat regions along with the (*C*) bases that compose the poly(A) and homopyrimidine tracts. (*D*) A diagram depicting the structure of the different putative *HLI* retrocopies: The head-to-head pairs are composed of genes from (*i*) clade 1 and clade 2 (11 pairs identified in KR01), (*ii*) clade 2 and clade 4 (one pair), and (*iii*) clade 3 with a putative pseudogenized *HLI* gene (four pseudo-pairs). (*E*) Representation of KR01 scaffolds that contain *HLI* genes. Discontinuous arrows (in the tracks at the bottom of the panel) indicate the presence of introns (the lighter shaded parts of the arrows). The gene colors correspond to their phylogenetic clade (shown in the legend in the bottom left corner); poly(A) tails and homopyrimidine tracts are shown as green and yellow boxes, respectively. The images were generated using Geneious Prime 2020.1.1 (https://www.geneious.com) and then assembled using Keynote (11.0.1).

available RNA-seq data for the responses of KR01 to light stress (6). In KR01, the putative pseudogenes *PSAE1-2* are not regulated by HL stress, whereas the identified retrogenes *PSAE3* and *PSAI* respond to HL stress. This suggests that there might be a link between retrotransposition and transcriptional regulation (e.g., selective advantage) of these retrocopied genes in response to light-induced stress.

In KR01, 40/50 of the *HLI* genes are putative retrocopies; there is no evidence of retrotransposition associated with 10 remaining copies (Fig. 2*E*). Of the 40 putative *HLI* retrocopies, 28 (70%) are regulated by HL (Fig. 3 and *SI Appendix*, Fig. S6). In contrast, 9 of the 10 putative nonretrocopies are transcriptionally active, with only one that shows differential regulation in response to HL stress (HLI_45) (Fig. 3 and *SI Appendix*, Fig. S6). These observations suggest that *HLI* genes acquired the capacity for differential transcription in response to HL stress through retrotransposition. Moreover, the ratio of substitution rates at nonsynonymous and synonymous sites (dN/dS ratios) of the two major clades of *HLI* genes (clade 1 and 2) are lower for the putative retrocopies than for nonretrocopies, suggesting that the retrocopies are under stronger purifying selection (*SI Appendix*, Fig. S7). This hypothesis is also supported by branch lengths in the phylogenetic trees built from the HLIPs.

Among the HL stress-regulated *HLI* genes, we identified three different patterns of transcriptional regulation based on differential accumulation of the transcripts as the cells transition from dark to HL relative to their transition from dark to control/low light (6) (Fig. 3). *HLI* genes with the "early" expression pattern show a peak after ∼0.5 h of HL exposure, which decreases after 6 h; those with the "early-persistent" pattern peak after ∼0.5 h of HL and maintain this expression level for at least 6 h; those with the "late" pattern peak after ∼6 h of HL with little or no increase after 0.5 h. Whereas none of the expression patterns are restricted to a specific *HLI* clade, early-persistent is the most prevalent across the four clades. This is the only pattern present for members in all *HLI* clades and, in some cases, for both members of the pair (Fig. 3). These results suggest that this expression reflects the acquisition of a bidirectional promoter that may be ancestral to the two additional patterns. We identified two adjacent, nearly identical quasi-palindromic sequences in the intergenic region of the clade 1/clade 2 *HLI* gene pairs that may act as a bidirectional promoter (*SI Appendix*, Fig. S8). Differences in the DNA sequence near this putative bidirectional promoter were observed that could explain the distinct expression patterns identified; however, the impact of these differences will require further experimental confirmation. Notably, the *HLI* genes that exhibit an early expression pattern are primarily paired with members of the family that show a late expression pattern, which might be caused by transcriptional interference that potentially disrupts the simultaneous expression of both *HLI* copies within these pairs.

Discussion

Establishment of a novel organelle is critically dependent on the early steps of host–endosymbiont integration, yet we have little knowledge about the mechanisms that govern this process. Analyzing chromatophore and nuclear genes in *P. micropora*, we provide key insights into the paths and mechanisms by which endosymbiont/chromatophore genes that have been transferred to the host nuclear genome are rearranged and expressed to enable the evolution of an organelle. Here, we

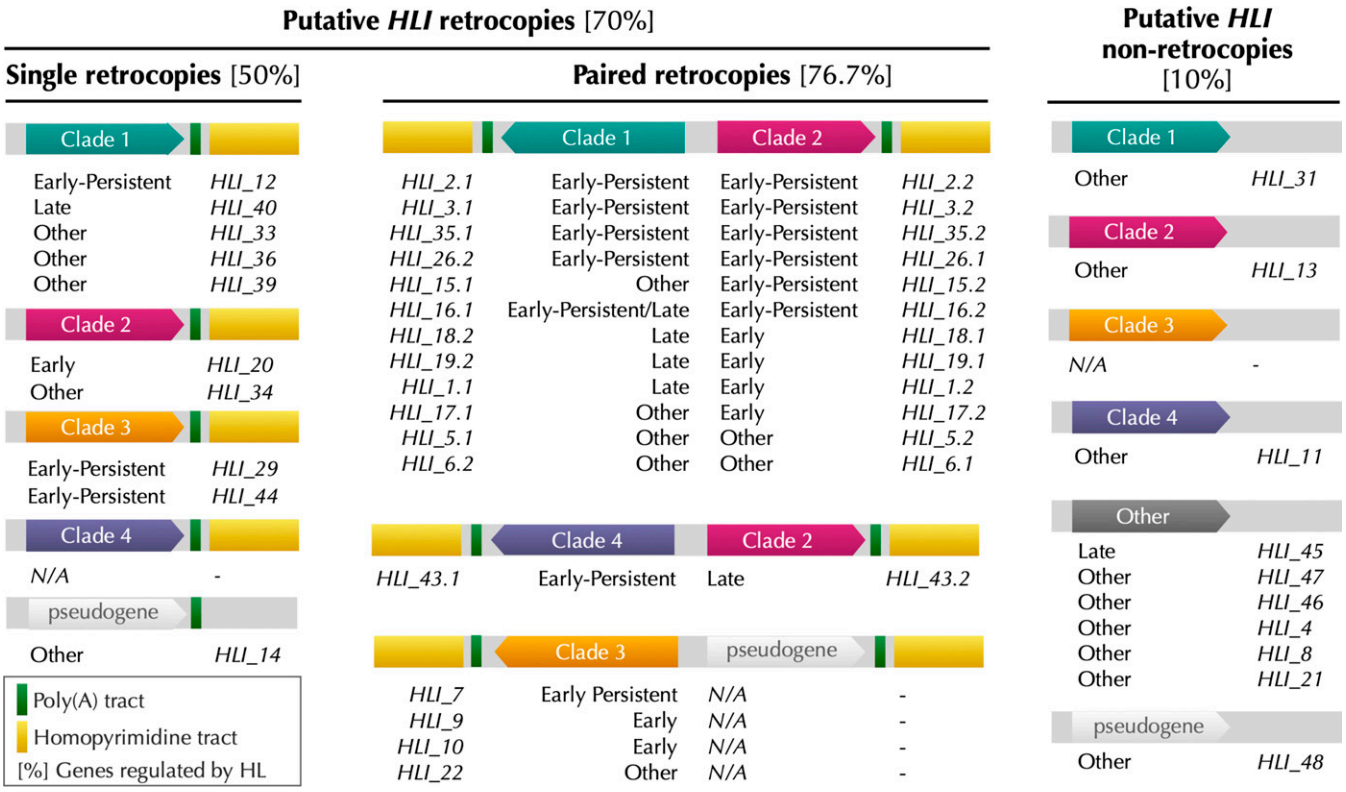

Fig. 3. Expression patterns of *HLI* genes in KR01 as the cells transitioned from dark to HL conditions, compared to their transition from dark to control light conditions; patterns based on RNA-seq data from Lhee et al. (6). The transcription categories are Early, peaks at 0.5 h of HL exposure but decreases after 6 h in the light; Early-Persistent, peaks at 0.5 h of and maintains this transcript level for at least an additional 6 h in HL; Late, little or no increase after 0.5 h of HL with high levels of transcript by 6 h of HL; and Other, no differential expression between HL and control light (or no expression).

posit that retrotransposition is a major facilitator of these processes, contributing to the expansion, reorganization, and acquisition of light regulation by the endosymbiont-derived genes in the host.

Phylogenetic analyses revealed that the *HLI* gene family in *Paulinella* has polyphyletic origins. In the common ancestor of photosynthetic *Paulinella* spp., multiple *hli* genes were likely transferred from the endosymbiont genome to the nuclear genome of the host (Fig. 4*A*). Most *HLI* genes in KR01 are retrocopies and are arranged in pairs positioned in a head-to-head orientation. The *HLI* genes in each head-to-head retrogene pair, despite being homologs, never belong to the same clade (i.e., they are not recent duplicates of each other), which raises the question of how the clusters were formed. This head-to-head gene configuration is uncommon for *HLI* genes in cyanobacteria and cyanophage genomes, where they are often found in tandem head-to-tail arrangements (24, 25). Therefore, it seems unlikely that they were transferred to the host in this arrangement; rather, they were transferred separately and later rearranged as head-to-head pairs in the host nuclear genome. In other eukaryotic genomes, retrogenes are sometimes found in a head-to-head orientation, possibly because inherently bidirectional promoters that are transcriptionally active facilitate the insertion of transposable elements flanking these active promoters (26, 27). Thus, it is plausible that the *HLI* gene pairs in *Paulinella* were generated through their retrotransposition into the same genomic position in a divergent orientation, either simultaneously or stepwise (i.e., the insertion of one gene might have facilitated the insertion of the other). Alternatively, the conserved repeats (homopyrimidine tracts) that flank these gene pairs may have promoted recombination events that led to this genomic arrangement (28). Regardless of the mechanism

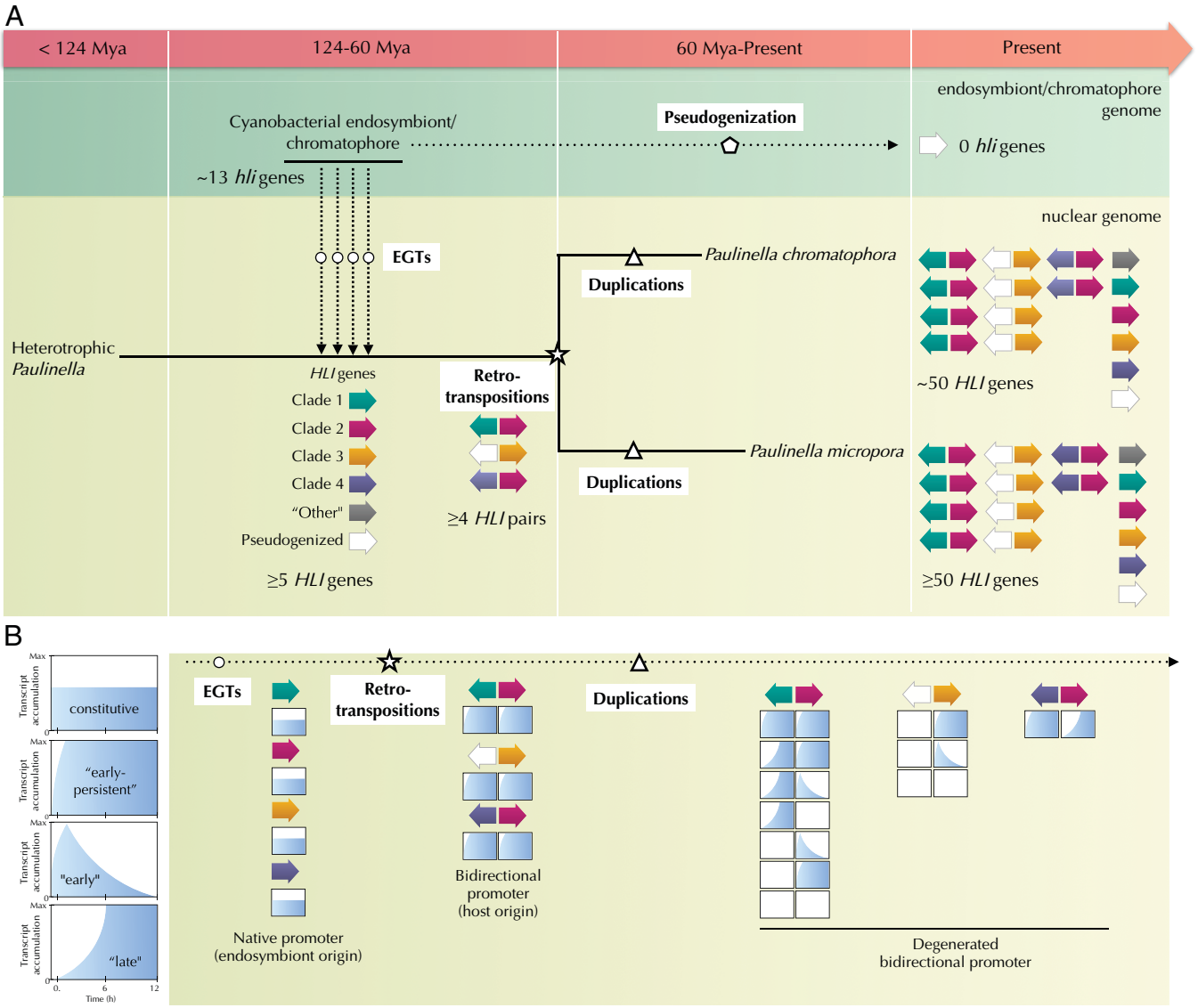

Fig. 4. Hypothetical model for the evolution of the *HLI* gene family in *Paulinella*. (*A*) At the early stages of primary endosymbiosis, before the divergence of *P. chromatophora* and *P. micropora* species, different *HLI* genes were transferred from the cyanobacterial endosymbiont/chromatophore into the host nuclear genome via independent EGT events. These genes were duplicated and retrotransposed, forming at least three kinds of head-to-head gene pairs with different *HLI* gene compositions. After *P. chromatophora* and *P. micropora* diverged, the *HLI* gene pairs were highly duplicated, independently in the two species (via a DNA-mediated mechanism), resulting in expansion of this gene family. Genes were also lost after transfer to the nuclear genome as a consequence of pseudogenization because they were not effectively expressed and/or not functionally effective. (*B*) The ancestral EGT-derived *HLI* genes were most likely associated with their original promoter and not regulated by stress in the host's nuclear genome. Different *HLI* paired retrocopies may have associated with or acquired a bidirectional promoter that is inducible by stress, leading to an early-persistent expression pattern through retrotransposition. Extensive duplications of the *HLI* pairs and degeneration of this ancestral promoter could have resulted in the evolution of two new expression patterns that are complementary, partitioning the early-persistent pattern into the more specialized early and late expression patterns.

underlying the generation of the gene pairs, the finding that the conserved retrotransposon domains located contiguous to the retrocopied genes are of eukaryotic origin suggests that the organization and fixation of these genes were driven by the host. Furthermore, because this *HLI* arrangement is found in both *P. micropora* and *P. chromatophora*, it most likely evolved during the early stages of primary endosymbiosis, predating the divergence of these species (~60 Mya) and generated at least three different *HLI* pairs that contain genes from four different *HLI* clades (clades 1 to 4) (Fig. 4*A*). In addition, we found that most of these retrogenes are differentially regulated in response to HL, whereas the nonretrogenes are mostly not HL responsive, despite being transcriptionally active. This observation leads us to hypothesize that the *HLI* genes acquired HL-responsive regulatory elements through retrotransposition. We also found that these *HLI* retrogene pairs were later expanded through DNA-mediated duplications, independently in the genomes of these two *Paulinella* species. This demonstrates a convergent evolutionary path in which the HL-regulated retrocopies are preferentially retained.

In addition to the *HLI* retrogenes, we identified active retrocopies of the EGT-derived *PSAE* and *PSAI* genes, supporting our hypothesis that retrotransposition has facilitated adaptation of the endosymbiont-derived genes in the amoeba genome. Gene duplications facilitate adaptation by providing new genetic material for mutation, drift, and selection to act upon (29). Moreover, the "copy-paste" mechanism of retrotransposition provides additional opportunities for the acquisition of host gene promoters with adaptive regulatory features (30). Thus, we hypothesize that retrotransposition plays a key role in "rewiring" the expression of prokaryotic genes transferred to eukaryotic genomes by replacing their original promoters, which may not be suitable for expression in the eukaryotic host, with those of host origin, which may confer appropriate regulatory features. Indeed, we found that most of the HL-induced *HLI* genes were generated through retrotransposition, with most of the nonretrocopies being transcriptionally active but not HL regulated. Furthermore, the HL-regulated copies of *PSAE3* and *PSAI* are retrocopies. Consistent with and broadening this hypothesis are the findings that genes transferred from bacteria to worm and insect eukaryotic genomes are embedded in a region of DNA enriched in transposon/retrotransposon elements that may mediate gene duplication events in the host (4, 31, 32). Moreover, retrotransposition played a role in the expansion of genes involved in acclimation to stress, which included bacteria-derived genes, in the cold-adapted green microalga *Chlamydomonas antarctica* (33). Overall, our data strongly suggest that retrotransposition of *Paulinella* genes can generate multiple gene copies that enable the rewiring of genes and their regulatory features to sustain host–endosymbiont interactions. This process provides the foundation for the evolution of additional molecular changes that favor a more efficient and resilient primary endosymbiotic association.

Because of its potential to generate deleterious effects, retrotransposition is suppressed under normal conditions. However, when organisms face extreme stress, this mechanism can become active and induce extensive genetic rearrangements that can lead to more "adaptive" functions and/or the evolution of novel functions that enhance cell survival (23). For instance, large-scale retrotransposition events appear to be triggered by dramatic changes in climate/conditions, resulting in expanded gene families that facilitate stress acclimation. This process is also associated with photosynthesis and the establishment of the symbiosis between cnidarians and their Symbiodiniaceae endosymbionts (18, 34). Similarly, retrotransposition has been associated with the adaptation of *C. antarctica* to cold temperatures (33). Our results show that the *HLI*, *PSAE*, and *PSAI* retrogene copies acquired the ability to respond to HL-induced stress in KR01 during the period in which the endosymbiont was transitioning from a free-living organism to a nascent organelle inside the *Paulinella* host. Because the photosynthetic endosymbiont in *Paulinella* produces ROS even under relatively low-light conditions (7) and retrotransposons are induced by oxidative stress in plants, fungi, and mammals (35–38), we hypothesize that during the early stages of the integration of host–endosymbiont metabolisms, the discord between the metabolisms of two organisms elicited high and sustained levels of ROS production that induced retrotransposon-dependent gene duplications. These events could have elevated *Paulinella* resilience by increasing gene copy number and facilitating altered regulation of endosymbiont-derived genes critical for acclimation of the cyanobacterial ancestor of the chromatophore to oxidative stress.

The finding that *HLI* gene pairs generated through retrotransposition in the common *Paulinella* ancestor were later expanded independently in each of the two *Paulinella* species suggests that the retrocopies conferred an advantage to the host and were subject to selection. For the evolution of the photosynthetic *Paulinella* spp., an increase in *HLI* gene dosage could lessen the potential damaging impact caused by the absorption of excessive excitation energy/light stress. An increase in copy number could have also allowed mutations in these genes that enhance or optimize transcriptional regulation in the host genome to become fixed, tuning HLI levels to the physiological conditions. Here, we identified three *HLI* expression patterns in response to HL stress: early, late, and early-persistent. Because all clades of *HLI* retrocopies show the early-persistent pattern, this expression pattern was likely ancestral to the other two that may have evolved later. Interestingly, the combined expression features of the two putative newly evolved patterns are complementary to the putative ancestral pattern (i.e., early + late = early-persistent). The duplication, degeneration, and complementation model for the preservation of gene duplicates predicts that mutations in regulatory elements can increase the probability of gene preservation, usually by creating more specific gene functions and patterns of expression by partitioning ancestral functions rather than evolving new ones (39). Partitioning of gene expression has been shown in humans and yeast, where it leads to tissue- or subcellular localization–specific expression (40–42). Here, we hypothesize that the two newly evolved expression patterns in KR01 resulted from modification of the ancestral promoter to allow the "subfunctionalization" of different *HLI* genes—not in space, but in time—allowing the organism to more effectively cope with different features of light stress (e.g., photosystem damage during initial light stress versus ROS accumulation after prolonged HL exposure) (Fig. 4*B*). These optimization steps could reduce the oxidative stress generated by the nascent organelle and accelerate host–endosymbiont integration.

By uncovering the genetic mechanisms necessary for domestication of endosymbiont-derived genes in the host nuclear genome of *Paulinella* spp., we provide insights into primary endosymbiosis. The optimization of transcriptional control of these genes by the host via extensive retrotransposition (potentially promoted by light stress) has allowed the loss of the ancestral gene copies in the endosymbiont genome. This led to a genetic dependency on the host that was likely key for stabilizing the integration of the partner organisms and enabling the evolution of a new photosynthetic organelle. Finally, even though this study reveals a major role for retrotransposition in facilitating the evolution of a photosynthetic organelle in *Paulinella* spp., this duplication mechanism is

ubiquitous in many eukaryotes. Thus, the role of retrotransposition in gene domestication might extend to other eukaryotic genomes, with horizontally transferred genes that become established and functionally active to evolve adaptive or novel functions in their recipient organisms (Fig. 5).

Materials and Methods

Processing of Sequencing Reads. DNA and RNA sequences from *P. micropora* KR01 [hereinafter KR01; BioProject PRJNA568118 from (6); *SI Appendix*, Table S3], *P. micropora* MYN1 [hereinafter MYN1; BioProject DRA003106 from (8); *SI Appendix*, Table S4], and *P. chromatophora* CCAC0185 [hereinafter *P. chromatophora*; BioProject PRJNA311736 from (20); *SI Appendix*, Table S5] were retrieved from the National Center for Biotechnology Information Sequencing Read Archive. The genome assemblies and associated predicted genes (or transcriptome assembly in the case of *P. chromatophora*) for the three isolates were retrieved from their respective repositories (6, 8, 20). Illumina short read sequencing data from both DNA and RNA were trimmed using Trimmomatic v0.38 (43) (ILLUMINACLIP:adapters.fa:2:30:10 SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25); only reads from pairs where both mates survived trimming were used in downstream analysis. The trimmed Illumina RNA-seq libraries from KR01 and MYN1 were independently mapped against their associated reference genomes using HISAT2 (v2.1.0; -q –phred33 –no-unal –dta –rf) (44), and transcripts were constructed for each library using StringTie2 (v2.0.6; –rf) (45). The resulting transcript gtf files for each of the isolates were merged into a combined set using StringTie2 (–merge), and transcript sequences were extracted using gffread (v0.11.6) (46). The trimmed Illumina DNA sequence libraries from the three isolates were aligned against their respective reference genomes using Bowtie2 (v2.3.5.1; –very-sensitive –no-unal) (47), and the resulting bam files (one from each aligned library) were combined using samtools merge (v1.8). Minimap2 (v2.17-r941) (48) was used to align the PacBio RNA (–secondary = no -ax splice) and DNA (–secondary=no -ax map-pb) sequence reads against their respective reference genomes.

Identification of repeat regions. The reference genomes of all three species were searched for four types of simple or low-complexity repeats [poly(A), poly(T) (thymine), poly(T/C) (thymine or cytosine), and poly(A/G) (adenine or guanine)] using a sliding window (size 20 bp, step size 1 bp) and a composition threshold of ≥80%. For example, if ≥80% of the bases within a given window are thymine or cytosine residues, then it is classified as a poly(T/C) repeat region. Overlapping windows of the same repeat type were merged into larger regions using bedtools merge (v2.25.0) (49). This method does not require a specific ratio between two residues when identifying poly(T/C) or poly(A/G) repeat regions, only that both residues combined pass the filtering threshold; thus, the regions identified as poly(T/C) or poly(A/G) will overlap with the poly(T) and poly(A) regions, respectively. Because these regions were used to focus our visualization and manual analysis, this is not viewed as a problem.

Identification of *HLI* Genes in *P. micropora* KR01. Genes of *Synechococcus* sp. WH5701 (ASM15304v1) annotated as HL-inducible were extracted, and their protein sequences were used to identify putative *HLI* genes in the three analyzed *Paulinella* isolates. The Hli proteins from *Synechococcus* sp. WH5701 were used as the query for tBLASTn, BLASTp, and exonerate searches in all cases below. Putative *HLI* gene models were predicted in the KR01 genome using a strategy that combined five sources of evidence: 1) existing KR01 proteins annotated as HL-inducible [annotations from Lhee et al. (6), Supplementary Data]; 2) existing KR01 proteins with BLASTp hits (*e*-value $<1e^{-5}$ and query coverage >50%) from *Synechococcus* sp. WH5701 Hli proteins; 3) regions of the KR01 genome with tBLASTn hits (*e*-value $<1e^{-5}$ and query coverage >50%) from *Synechococcus* sp. WH5701 Hli proteins; 4) regions of the KR01 genome with exonerate alignments (v2.2.0; –showquerygff –showtargetgff –model protein2genome) (50) with *Synechococcus* sp. WH5701 Hli proteins; and 5) StringTie2 constructed transcripts with tBLASTn hits (*e*-value $<1e^{-5}$ and query coverage >50%) from *Synechococcus* sp. WH5701 Hli proteins. The genome coordinates of all features from the aforementioned evidence sources were visualized using Integrative Genomics Viewer (IGV) v2.8.2 (51) and a final set of KR01 *HLI* genes constructed via manual annotation (*SI Appendix*, Table S6) with, where possible, open reading frames that extend to in-frame start and stop codons. Multiexon *HLI* gene models were only constructed if they were strongly supported by the RNA-seq reads aligned against the genome using HISAT2. If no appropriate stop or start codons could be identified, because they were not in the correct reading frame as the putative HLI protein, they were not covered by aligned RNA-seq reads, or they would require overextending the protein compared to its current length, then the protein was left partial, including only the positions that were covered by alignments to the query HLI proteins. This approach was adopted to prevent overextending the short HLI proteins into part of the genome that are not transcribed or are part of the untranslated regions (UTRs) of the gene. The identified *HLI* genes were considered pseudogenes if they had stop codons within the conserved region of the gene (i.e., the region with homology to *Synechococcus* sp. WH5701 Hli proteins), were missing the conserved "NGR" amino acid motif that is common to all *Synechococcus* sp. WH5701 Hli proteins, or had apparent frameshift mutations that would prevent the proper translation of the protein. Genes missing start or stop codons were not considered pseudogenes, even if an upstream stop codon prevented the formation of a complete protein. The protein sequences of the new KR01 *HLI* gene models were extracted and compared against the Hli proteins from *Synechococcus* sp. WH5701 using BLASTp (*e*-value $<1e^{-5}$ and query coverage >50%), and proteins without hits to the *Synechococcus* sp. WH5701 Hli proteins were considered pseudogenes. *HLI* gene UTRs were identified by visualizing and manually examining the RNA-seq reads mapped using HISAT against the genome. If no clear translation start or stop codon could be identified from the mapped reads or if no reads were found to have mapped to a putative *HLI* gene, then the UTRs were assumed to extend 100 bp up and downstream of the encoded HLI protein. Full messenger RNA sequences (including UTRs) were extracted for each *HLI* gene and used for downstream analysis.

Identification of *HLI* Genes in *P. micropora* MYN1. Putative *HLI* gene models were predicted in the MYN1 genome using a strategy that combines four sources of evidence: 1) existing MYN1 Coding DNA Sequence with tBLASTn hits (*e*-value $<1e^{-5}$ and query coverage >50%) from *Synechococcus* sp. WH5701 Hli proteins; 2) regions of the MYN1 genome with tBLASTn hits (*e*-value $<1e^{-5}$ and query coverage >50%) from *Synechococcus* sp. WH5701 Hli proteins; 3) regions of the MYN1 genome with exonerate alignments (v2.2.0; –showquerygff –showtargetgff –model protein2genome) from *Synechococcus* sp. WH5701 Hli proteins; and 4) StringTie2-constructed transcripts with tBLASTn hits (*e*-value $<1e^{-5}$ and query coverage >50%) from *Synechococcus* sp. WH5701 Hli proteins. The genome coordinates of all features detected in the aforementioned analyses were manually scrutinized and extracted as described for KR01 (*SI Appendix*, Table S7).

Identification of HLI Genes in *P. chromatophora*. Putative *HLI* gene models were predicted in *P. chromatophora* using the available genome and transcriptome assemblies. First, assembled transcripts with tBLASTn hits (*e*-value $<1e^{-5}$ and query coverage >50%) from *Synechococcus* sp. WH5701 Hli proteins were extracted. These transcripts were manually examined using IGV, and putative coding regions were identified using an approach similar to that used for KR01 but without the need to consider multiexon genes, flanking repeats, or UTRs. The identified coding regions of these transcripts were extracted and aligned against the *P. chromatophora* genome assembly using exonerate (v2.2.0; –showquerygff –showtargetgff –model coding2genome –percent 95), retaining only those alignments that covered >90% of the query sequence; these aligned transcripts were used as evidence for the construction of genome-based *HLI* gene models. HLI transcripts that were not found to be represented in the genome assembly (i.e., those that did not have exonerate alignments above the specified thresholds) were extracted and used for downstream analysis. *HLI* genes were identified in the *P. chromatophora* genome using information from 1) the exonerate aligned *P. chromatophora* HLI transcripts; 2) tBLASTn hits (*e*-value $<1e^{-5}$ and query coverage >50%) from *Synechococcus* sp. WH5701 Hli proteins against the genome; and 3) exonerate alignments (v2.2.0; –showquerygff –showtargetgff –model protein2genome) from *Synechococcus* sp. WH5701 Hli proteins against the genome. The genome coordinates of all features detected in the aforementioned analyses were manually scrutinized and extracted as described for KR01 (*SI Appendix*, Table S8). The combined set of *P. chromatophora* genome-based *HLI* gene models and unaligned HLI transcripts was used for downstream analysis.

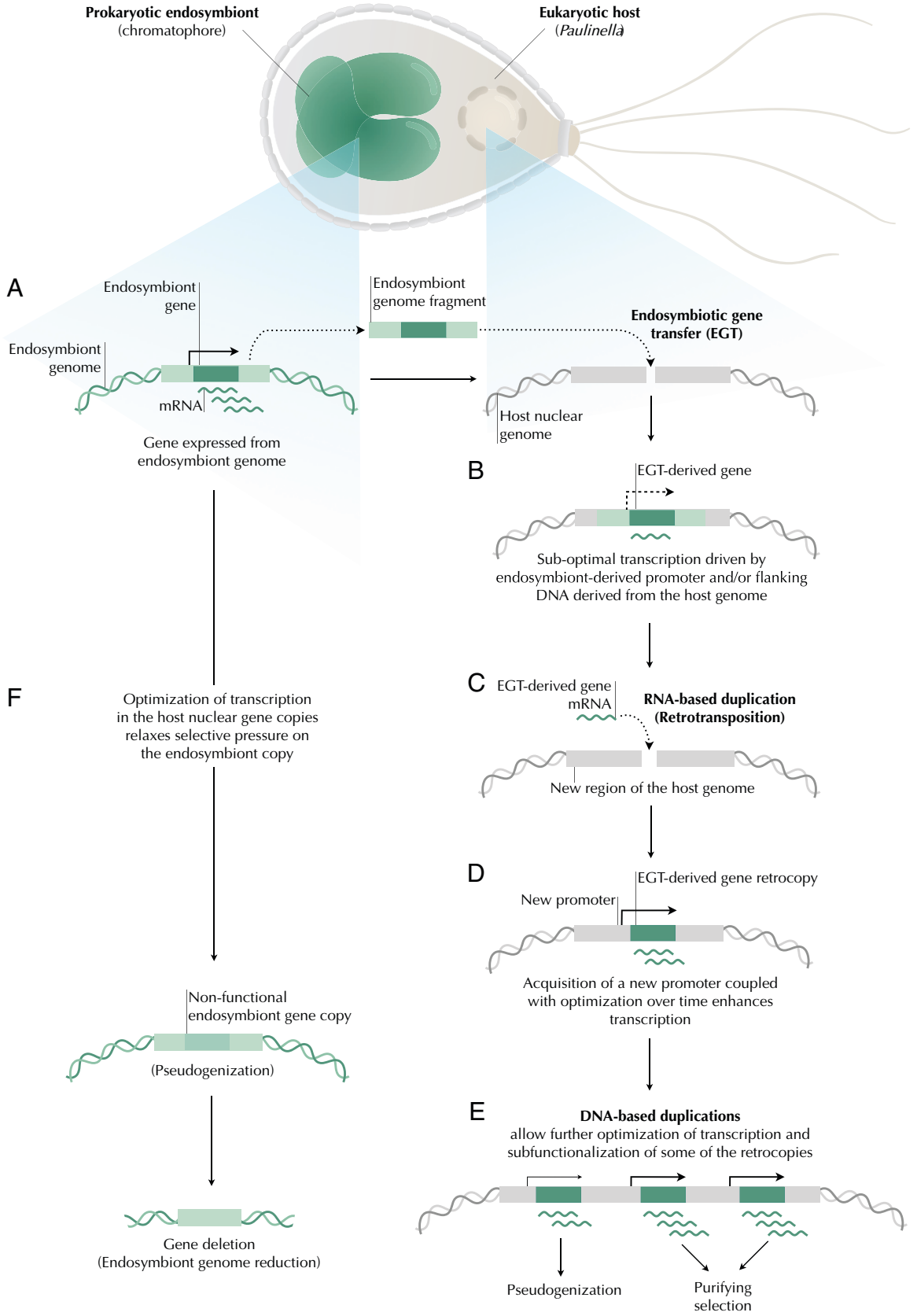

Fig. 5. Hypothetical role of retrotransposition in the domestication of prokaryotic genes in eukaryotic genomes based on the *Paulinella* model. Genome fragments containing genes are transferred from the endosymbiont into the host nuclear genome (*A*). The recently transferred endosymbiont genes of prokaryotic origin are likely not optimally expressed in the eukaryotic genome (*B*). Retrotransposition events cause duplication and relocation of the genes in different locations in the nuclear genome (*C*) that allow the acquisition of new promoters (*D*). These new promoters, in addition to optimization over time, improve transcription of the EGT-derived gene retrocopies. Additional DNA-based duplications allow further optimization of transcription and subfunctionalization of some of the retrocopies, while the suboptimally regulated copies become pseudogenized (*E*). The optimized transcriptional control of the endosymbiont-derived genes in the host nucleus relaxes selective pressure on the endosymbiont copies, resulting in their pseudogenization and loss over time. Consequently, endosymbiont genes become under host control in the nucleus while the endosymbiont genome is reduced. mRNA, messenger RNA.

Differential Expressed Genes. Expression of the KR01 *HLI* genes not on duplicated scaffolds (*SI Appendix*) over time under HL and control light conditions was quantified using RSEM v1.3.3 (52) (–paired-end –bowtie2 –strandedness reverse –estimate-rspd –sort-bam-by-coordinate; using bowtie2 v2.3.5.1). Expression values were normalized using the "median-of-ratios" method (implemented in DESeq2 v1.30.1) (53).

Phylogenetic Analysis of HLIPs. All updated HLI proteins from the three *Paulinella* isolates, excluding those that are putative pseudogenes, were extracted and combined with Hli proteins from six *P. marinus* species, seven *Synechococcus* species (including WH5701), three *Synechocystis* species, and one *Gloeomargarita lithophora* species (*SI Appendix*, Table S9). The combined set of all HLI/Hli sequences was aligned using MAFFT (v7.453; –localpair –maxiterate 1000), and a maximum-likelihood phylogenetic tree was inferred using IQ-TREE (v1.6.12; -m MFP -bb 2000 -alrt 2000 -bnni) (54), allowing the program to choose the best evolutionary model for the alignment (55). Node support was calculated for the inferred consensus tree using 2,000 ultrafast BSs (56). The 20 KR01 HLI proteins encoded on duplicate scaffolds were pruned from the final tree shown in *SI Appendix*, Fig. S2.

Calculation of dN/dS Ratios for HLIPs. Proteins from KR01 clade 1 *HLI* genes were aligned with the clade 1 *Hli* from *Synechococcus* sp. WH 5701 and HliA and HliB from *Synechocystis* sp. PCC 6803 using MAFFT (v7.453; –localpair –maxiterate 1000), and a maximum-likelihood tree was inferred from the alignment using IQ-TREE (v1.6.12; -m MFP -bb 2000 -alrt 2000 -bnni) (54). Proteins from the KR01 clade 2 *HLI* genes were aligned with the clade 2 *Hli* from *Synechococcus* sp. WH 5701 and HliC from *Synechocystis* sp. PCC 6803 using MAFFT (v7.453; –localpair –maxiterate 1000), and a maximum-likelihood phylogenetic tree was inferred from the alignment using IQ-TREE (v1.6.12; -m MFP -bb 2000 -alrt 2000 -bnni). The pal2nal.pl script (v14; -output paml -nogap) (57) was used to prepare a codon-based alignment from the clade 1 and clade 2 protein alignments using the nucleotide sequences of the *HLI* genes. Codeml (runmode = −2) from the PAML package (v4.9j) (58) was used to calculate the dS, dN, and dS/dN values for each pair of sequences in the alignments. The trees were visualized using iTOL (https://itol.embl.de/upload.cgi) and rooted by the *Synechocystis* sp. PCC 6803 sequences in each.

DNA Amplification of KR01 *HLI* Gene Pairs. *P. micropora* KR01 cells (59) were grown in DYV medium (Bigelow National Center for Marine Algae and Microbiota) at 23 °C for 3 wk under light/dark cycles (12 h/12 h) (∼50 μmol photons m^2 s^{-1}). Cells were harvested by centrifugation, and DNA was extracted using the PowerSoil(R) Kit (Cat. #12888–100) (Mo Bio Laboratories, Inc., Qiagen). Primers flanking the *HLI* gene pairs were designed using Primer-BLAST tool (https://www.ncbi.nlm.nih.gov/tools/primer-blast/ and *SI Appendix*, Table S10) and used for PCR with GoTaq Green Master Mix (Cat. #M7122) (Promega) as follows: 94 °C for 5 min; 35 cycles of 94 °C for 30 s, 60 °C for 30 s, and 72 °C for 90 s; and a final extension at 72 °C for 2 min. The amplified DNA products were run into 1% agarose gels, and the resulting bands with the expected sizes were extracted from the gel using GeneJET Gel Extraction Kit #K0692 (Thermo Fisher Scientific Baltics UAB) and sequenced by ELIM Biopharmaceuticals (https://www.elimbio.com/).

Data Availability. The phylogenetic tree shown in *SI Appendix*, Fig. S2 and its associated alignment are available from https://doi.org/10.5281/zenodo.5684911 (60); the *Paulinella* HLI genes reported in this study are available from https://doi.org/10.5281/zenodo.5684817 (61). Previously published sequencing data were used in this work and are available from the Sequence Read Archive (Accession Nos. PRJNA568118, DRA003106, and PRJNA311736); the genomes and predicted genes are available from http://cyanophora.rutgers.edu/P_micropora/, http://cyanophora.rutgers.edu/paulinella/, and the DNA Data Bank of Japan (BJOX01000001–BJOX01008276). All other study data are included in the article and/or supporting information.

ACKNOWLEDGMENTS. A.G., T.G.S., V.C. and D. Bhattacharya were supported by a grant from the National Aeronautics and Space Administration (80NSSC19K0462). A.R.G., D. Bhaya, and V.C. were supported by the Carnegie Institution for Science. D. Bhattacharya was supported by a National Institute of Food and Agriculture–US Department of Agriculture Hatch Grant (NJ01180).

Author affiliations: [a]Department of Plant Biology, The Carnegie Institution for Science, Stanford, CA 94305; [b]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08901; and [c]Graduate Program in Molecular Biosciences, Program in Microbiology and Molecular Genetics, Rutgers University, Piscataway, NJ 08854

1. E. C. M. Nowack *et al.*, Endosymbiotic gene transfer and transcriptional regulation of transferred genes in *Paulinella chromatophora*. *Mol. Biol. Evol.* **28**, 407–422 (2011).
2. J. N. Timmis, M. A. Ayliffe, C. Y. Huang, W. Martin, Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* **5**, 123–135 (2004).
3. R. I. Ponce-Toledo, P. López-García, D. Moreira, Horizontal and endosymbiotic gene transfer in early plastid evolution. *New Phytol.* **224**, 618–624 (2019).
4. F. Husnik, J. P. McCutcheon, Functional horizontal gene transfer from bacteria to eukaryotes. *Nat. Rev. Microbiol.* **16**, 67–79 (2018).
5. J. L. Blanchard, M. Lynch, Organellar genes: Why do they end up in the nucleus? *Trends Genet.* **16**, 315–320 (2000).
6. D. Lhee *et al.*, Amoeba genome reveals dominant host contribution to plastid endosymbiosis. *Mol. Biol. Evol.* **38**, 344–357 (2021).
7. R. Zhang, E. C. M. Nowack, D. C. Price, D. Bhattacharya, A. R. Grossman, Impact of light intensity and quality on chromatophore and nuclear gene expression in *Paulinella chromatophora*, an amoeba with nascent photosynthetic organelles. *Plant J.* **90**, 221–234 (2017).
8. M. Matsuo *et al.*, Large DNA virus promoted the endosymbiotic evolution to make a photosynthetic eukaryote. *bioRxiv*. [Preprint] https://doi.org/10.1101/809541 (2019). Accessed 29 September 2020.
9. Q. He, N. Dolganov, O. Bjorkman, A. R. Grossman, The high light-inducible polypeptides in *Synechocystis* PCC6803. Expression and function in high light. *J. Biol. Chem.* **276**, 306–314 (2001).
10. L. Yu, J. Zhao, U. Muhlenhoff, D. A. Bryant, J. H. Golbeck, PsaE is required for in vivo cyclic electron flow around photosystem I in the *Cyanobacterium Synechococcus* sp. PCC 7002. *Plant Physiol.* **103**, 171–180 (1993).
11. K. El Bissati, D. Kirilovsky, Regulation of *psbA* and *psaE* expression by light quality in *Synechocystis* species PCC 6803. A redox control mechanism. *Plant Physiol.* **125**, 1988–2000 (2001).
12. W. M. Schluchter, G. Shen, J. Zhao, D. A. Bryant, Characterization of *psaI* and *psaL* mutants of *Synechococcus* sp. strain PCC 7002: A new model for state transitions in cyanobacteria. *Photochem. Photobiol.* **64**, 53–66 (1996).
13. Q. Xu *et al.*, Mutational analysis of photosystem I polypeptides in the cyanobacterium *Synechocystis* sp. PCC 6803. Targeted inactivation of *psaI* reveals the function of psaI in the structural organization of psaL. *J. Biol. Chem.* **270**, 16243–16250 (1995).
14. E. C. M. Nowack, A. R. Grossman, Trafficking of protein into the recently established photosynthetic organelles of *Paulinella chromatophora*. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 5340–5345 (2012).
15. A. Singer *et al.*, Massive protein import into the early-evolutionary-stage photosynthetic organelle of the amoeba *Paulinella chromatophora*. *Curr. Biol.* **27**, 2763–2773.e5 (2017).
16. H. Kaessmann, N. Vinckenbosch, M. Long, RNA-based gene duplication: Mechanistic and evolutionary insights. *Nat. Rev. Genet.* **10**, 19–31 (2009).
17. M. Ąkalski *et al.*, Comparative genomic analysis of retrogene repertoire in two green algae *Volvox carteri* and *Chlamydomonas reinhardtii*. *Biol. Direct* **11**, 35 (2016).
18. B. Song *et al.*, Comparative genomics reveals two major bouts of gene retroposition coinciding with crucial periods of *Symbiodinium* evolution. *Genome Biol. Evol.* **9**, 2037–2047 (2017).
19. S. Jantaro, Q. Ali, S. Lone, Q. He, Suppression of the lethality of high light to a quadruple *HLI* mutant by the inactivation of the regulatory protein PfsR in *Synechocystis* PCC 6803. *J. Biol. Chem.* **281**, 30865–30874 (2006).
20. E. C. M. Nowack *et al.*, Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of *Paulinella chromatophora*. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12214–12219 (2016).
21. Q.-R. Liu, P. K. Chan, Identification of a long stretch of homopurine.homopyrimidine sequence in a cluster of retroposons in the human genome. *J. Mol. Biol.* **212**, 453–459 (1990).
22. K. Okamura, K. Nakai, Retrotransposition as a source of new promoters. *Mol. Biol. Evol.* **25**, 1231–1238 (2008).
23. P. Mita, J. D. Boeke, How retrotransposons shape genome regulation. *Curr. Opin. Genet. Dev.* **37**, 90–100 (2016).
24. D. Lindell *et al.*, Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11013–11018 (2004).
25. D. Bhaya, A. Dufresne, D. Vaulot, A. Grossman, Analysis of the *hli* gene family in marine and freshwater cyanobacteria. *FEMS Microbiol. Lett.* **215**, 209–219 (2002).
26. M. Fablet, M. Bueno, L. Potrzebowski, H. Kaessmann, Evolutionary origin and functions of retrogene introns. *Mol. Biol. Evol.* **26**, 2147–2156 (2009).
27. P. Kalitsis, R. Saffery, Inherent promoter bidirectionality facilitates maintenance of sequence integrity and transcription of parasitic DNA in mammalian genomes. *BMC Genomics* **10**, 498 (2009).
28. A. Bacolla *et al.*, Long homopurine*homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. *Nucleic Acids Res.* **34**, 2663–2675 (2006).
29. E. Kuzmin, J. S. Taylor, C. Boone, Retention of duplicated genes in evolution. *Trends Genet.* **38**, 59–72. (2021). Correction in: Trends Genet. 10.1016/j.tig.2022.03.014 (2022).
30. C. Casola, E. Betrán, The genomic impact of gene retrocopies: What have we learned from comparative genomics, population genomics, and transcriptomic analyses? *Genome Biol. Evol.* **9**, 1351–1373 (2017).
31. Y. Pauchet, D. G. Heckel, The genome of the mustard leaf beetle encodes two active xylanases originally acquired from bacteria through horizontal gene transfer. *Proc. Biol. Sci.* **280**, 20131021 (2013).
32. S. N. McNulty *et al.*, Endosymbiont DNA in endobacteria-free filarial nematodes indicates ancient horizontal genetic transfer. *PLoS One* **5**, e11029 (2010).
33. X. Zhang, M. Cvetkovska, R. Morgan-Kiss, N. P. A. Hüner, D. R. Smith, Draft genome sequence of the Antarctic green alga *Chlamydomonas* sp. UWO241. *iScience* **24**, 102084 (2021).
34. S. Lin *et al.*, The *Symbiodinium kawagutii* genome illuminates dinoflagellate gene expression and coral symbiosis. *Science* **350**, 691–694 (2015).

35. C. Mhiri *et al.*, The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress. *Plant Mol. Biol.* **33**, 257–266 (1997).
36. K. Ikeda, H. Nakayashiki, M. Takagi, Y. Tosa, S. Mayama, Heat shock, copper sulfate and oxidative stress activate the retrotransposon MAGGY resident in the plant pathogenic fungus *Magnaporthe grisea*. *Mol. Genet. Genomics* **266**, 318–325 (2001).
37. V. Stribinskis, K. S. Ramos, Activation of human long interspersed nuclear element 1 retrotransposition by benzo(a)pyrene, an ubiquitous environmental carcinogen. *Cancer Res.* **66**, 2616–2620 (2006).
38. T. Stoycheva, M. Pesheva, P. Venkov, The role of reactive oxygen species in the induction of Ty1 retrotransposition in *Saccharomyces cerevisiae*. *Yeast* **27**, 259–267 (2010).
39. A. Force *et al.*, Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
40. M. Lynch, A. Force, The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459–473 (2000).
41. G. C. Conant, K. H. Wolfe, Functional partitioning of yeast co-expression networks after genome duplication. *PLoS Biol.* **4**, e109 (2006).
42. A. C. Marques, N. Vinckenbosch, D. Brawand, H. Kaessmann, Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. *Genome Biol.* **9**, R54 (2008).
43. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
44. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
45. S. Kovaka *et al.*, Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
46. G. Pertea, M. Pertea, GFF utilities: GffRead and GffCompare. *F1000 Res.* **9**, 304 (2020).
47. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
48. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
49. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
50. G. S. C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
51. J. T. Robinson *et al.*, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
52. B. Li, C. N. Dewey, RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
53. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
54. 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).
55. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
56. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
57. M. Suyama, D. Torrents, P. Bork, PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609-12 (2006).
58. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
59. D. Lhee *et al.*, Diversity of the photosynthetic *Paulinella* species, with the description of *Paulinella micropora* sp. nov. and the chromatophore genome sequence for strain KR01. *Protist* **168**, 155–170 (2017).
60. T. Stephens, High light inducible (HLI) protein tree. Zenodo. https://doi.org/10.5281/zenodo.5684911. Deposited 12 November 2021.
61. T. Stephens, High light-inducible (HLI) proteins identified in three Paulinella isolates. Zenodo. https://doi.org/10.5281/zenodo.5684817. Deposited 12 November 2021.