PNAS RESEARCH ARTICLE | EVOLUTION

Horizontal transmission of functionally diverse transposons is a major source of new introns

Landen Gozashti[a,b,c,1], Anne Nakamoto[d,e], Shelbi Russell[d,e], and Russell Corbett-Detig[d,e,1]

Edited by W. Doolittle, Dalhousie University, Halifax, NS, Canada; received July 24, 2024; accepted March 28, 2025

Since the discovery of spliceosomal introns in eukaryotic genomes, the proximate molecular and evolutionary processes that generate new introns have remained a critical mystery. Specialized transposable elements (TEs), introners, are thought to be one of the major drivers of intron gain in diverse eukaryotes. However, the molecular mechanism(s) and evolutionary processes driving introner propagation within and between lineages remain elusive. Here, we analyze 8,716 genomes, revealing 1,093 introner families in 201 species spanning 1.7 billion years of evolution. Introners are derived from functionally diverse TEs including families of terminal-inverted-repeat DNA TEs, retrotransposons, cryptons, and helitrons as well as mobile elements with unknown molecular mechanisms. We identify eight cases where introners recently transferred between divergent host species and show that giant viruses that integrate into genomes may facilitate introner transfer across lineages. We propose that ongoing intron gain is primarily a consequence of TE activity in eukaryotes, thereby resolving a key mystery of genome structure evolution.

genome evolution | introner | transposable elements | intron gain

Significance

Introns are major components of eukaryotic genomes with poorly understood origins. Introners, transposable elements (TEs) which can generate introns upon insertion, are thought to be major drivers of intron gain in diverse lineages. However, the molecular mechanisms of introner mobilization and the evolutionary processes shaping introner distributions across species remain elusive. Here, we show that introners can arise from highly diverse TEs with ancient origins. We find evidence that these elements can move between divergent species through horizontal gene transfer and that giant viruses may contribute to their transmission. Together, our results suggest that ongoing intron gain is largely a consequence of TE activity and genomic conflict in eukaryotes.

Spliceosomal introns are a fundamental component of eukaryotic genomes with poorly understood origins. Introns are noncoding regions within genes that must be removed from transcripts before translation. They are nearly ubiquitous across eukaryotic genomes and contribute to a range of molecular processes including regulation of transcription and protection against transcription-associated genome instability (1, 2). Introns also enable alternative splicing, vastly expanding the molecular and functional diversity encoded by eukaryotic genes (3). Many introns are essential to gene function, and intron deletion can be costly and even lethal (4). Although introns now serve important molecular functions, these roles must have evolved after the emergence of introns (5). Therefore, the proximal origins of spliceosomal introns remain poorly understood, representing a longstanding question in biology.

Transposable elements (TEs)—selfish genes that propage within genomes by copying themselves—may be the source of most new introns. The very first eukaryotic introns and spliceosomal small nuclear RNAs, likely evolved from TEs (3, 6). The last eukaryotic common ancestor was likely intron-rich, and many eukaryotic lineages have primarily experienced intron loss (7, 8). Nonetheless, the number of introns per gene varies tremendously across species (range ~0.003 to 20 per gene) (9–11), and comparisons of orthologous intron positions across lineages suggest that while some lineages have primarily undergone intron loss, others have experienced rapid episodic intron gain (8, 10). Several mechanisms of intron gain have been proposed (reviewed in ref. 12); however, de novo intron creation by specialized intron-generating transposable elements, termed introners, is the only mechanism that could explain the "bursts" of intron gains observed across lineages (10).

Introners are transposable elements which can be correctly spliced out of exons upon insertion (either by encoding or co-opting spliceosomal recognition sequences) and have the unique ability to create thousands of introns within a single genome (13, 14). Recently active introners can be identified by comparing sequence similarity among introns from diverse locations across the genome, have been reported in diverse lineages and may explain the vast majority of ongoing intron gain (15–22). However, the mechanisms driving introner proliferation remain almost entirely unexplored, and it is unclear whether introners arise from diverse transposable elements or are restricted to a specific mechanism of mobilization. Furthermore, introners show patchy taxonomic distributions and are enriched in species that experience frequent horizontal gene transfer (HGT), such as aquatic unicellulars and fungi, suggesting that HGT may play an important role in shaping introner distributions across the tree of life (15). Nonetheless, direct evidence for HGT has not been discovered, and the precise molecular mechanisms of transposition are almost completely unknown. Thus, the biological processes underlying de novo intron creation remain poorly understood.

Author affiliations: [a]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138; [b]Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138; [c]HHMI, Harvard University, Cambridge, MA 02138; [d]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064; and [e]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064

Preprint servers: bioRxiv: 10.1101/2024.06.04.597373.

Author contributions: L.G. and R.C.-D. designed research; L.G. performed research; L.G. and A.N. contributed new reagents/analytic tools; L.G., A.N., S.R., and R.C.-D. analyzed data; and L.G., A.N., S.R., and R.C.-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: lgozashti@g.harvard.edu or rucorbet@ucsc.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2414761122/-/DCSupplemental.

Published May 22, 2025.

Results

Transposable Elements Generate Introns across the Eukaryotic Tree. We systematically searched for introners in 8,716 annotated eukaryotic genome assemblies (*Materials and Methods*), revealing diverse species whose genomes contain introns derived from recent transposition. Introners are present in an exceptionally broad range of eukaryotic species (Fig. 1 and Dataset S1). In line with prior work, we identified an abundance of introners in aquatic organisms, unicellular species, and fungi (>98% of introner-containing species are aquatic, unicellular, or fungi; Dataset S1) (15). However, we also find introner families in an expanded range of taxa (*Materials and Methods* and *SI Appendix*, Figs. S1–S7). In particular, we observe recently active introners in land plants including the grasses, (*Panicum virgatum*) and eudicots (e.g., *Salvia splendens*), as well as an echinoderm, the purple sea urchin (*Strongylocentrotus purpuratus*). Our search also revealed introners in basidiomycete fungi as well as a myriad of diverse protist lineages (Fig. 1 and Dataset S1). Thus, introner-derived introns are widespread across diverse eukaryotic taxa, highlighting the universality of this mechanism of intron gain.

Intron-Generating TEs Span Exceptional Mechanistic Diversity. Mirroring the broad host taxonomic variety, TEs capable of generating introns are also exceptionally mechanistically and evolutionarily diverse. We used a combination of systematic and manual approaches to identify transposition mechanisms and identified introners as they relate to known transposable elements. Our approach classified TEs using homology to known TEs, expected functional domains and structural features consistent with specific mechanisms, and conserved terminal motifs and target site duplications (TSDs) observed for many elements (*Materials and Methods*, *SI Appendix*, Fig. S8, and Dataset S2). We find that introners arise from TEs spanning ~80% of orders and at least 50% of superfamilies of known mobile genetic elements as defined by ref. 23 [noting that other classification systems have also been proposed (24, 25) (Fig. 2*A*). These include diverse terminal inverted repeat (TIR) DNA transposons, long terminal repeat (LTR) retrotransposons, non-LTR retrotransposons, rolling circle helitrons and tyrosine recombinase (Crypton) elements (each comprising ~82%, ~12%, 5%, 1%, and <1% of confidently categorized examples; Fig. 2*B* and Dataset S2; reviewed in refs. 26 and 27). Notably, the ancient origins of these diverse TEs suggests that introners have likely been generating introns throughout eukaryotic evolution (28, 29). The vast majority of introners are nonautonomous, likely reflecting strong purifying selection against new introns that encode the functional machinery required for autonomous transposition. Additionally, ~72% of elements cannot be confidently categorized. We suspect that these unknown elements will provide insight into as-yet unexplored mobile elements (Dataset S2). This breadth emphasizes the remarkable functional diversity of transposons that are capable of generating introns on genomic scales.

We further identified elements where both autonomous and nonautonomous elements contribute concurrently to intron gain (Fig. 2*A* and Dataset S2). For example, an introner family in *P. virgatum* displays homology to known Copia LTR elements and several copies display functional domains required for autonomous Copia mobilization. However, other copies exist as solo-LTRs, products of ectopic recombination resulting in the removal of the internal TE sequence (30) (Fig. 2*A*). Both fully autonomous elements and solo-LTRs generate functional introns. TIR DNA transposons display similar patterns in several species (Fig. 2*A*). Some types of transposable elements, such as cryptons and helitrons, are particularly challenging to classify in the absence of autonomous elements due to their lack of conserved structural features. Thus, we required multiple lines of evidence supporting these introner classifications. For example, *Suillus subalutaceus* family 5 not only shows homology to known helitrons but also inserts between TT and TC nucleotides and fails to generate TSDs, as expected for helitron elements (31) (*SI Appendix*, Fig. S8).

In some species, a single introner family contributes to the vast majority of introner-mediated intron gain, whereas other species display multiple abundant introners with divergent origins. For

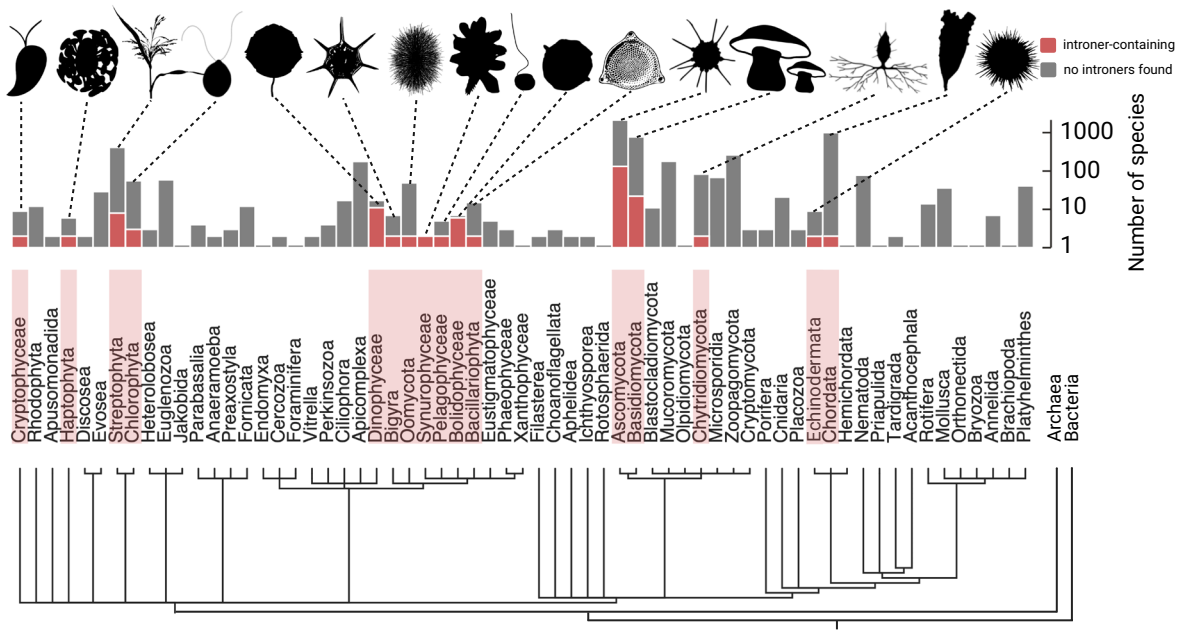

Fig. 1. Distribution of introners across the eukaryotic tree of life. Relationships between phyla for considered genomes are shown as a cladogram. Considered genomes are arranged into phyla according to NCBI taxonomy with outgroup lineages that do not contain a spliceosome: archaea and bacteria. Phyla with introner-containing species are highlighted in red. Gray and red bars display the number of species evaluated and the number of introner-containing species for each phylum, respectively. Silhouettes show examples of introner-containing species for each phylum and were retrieved from PhyloPic.

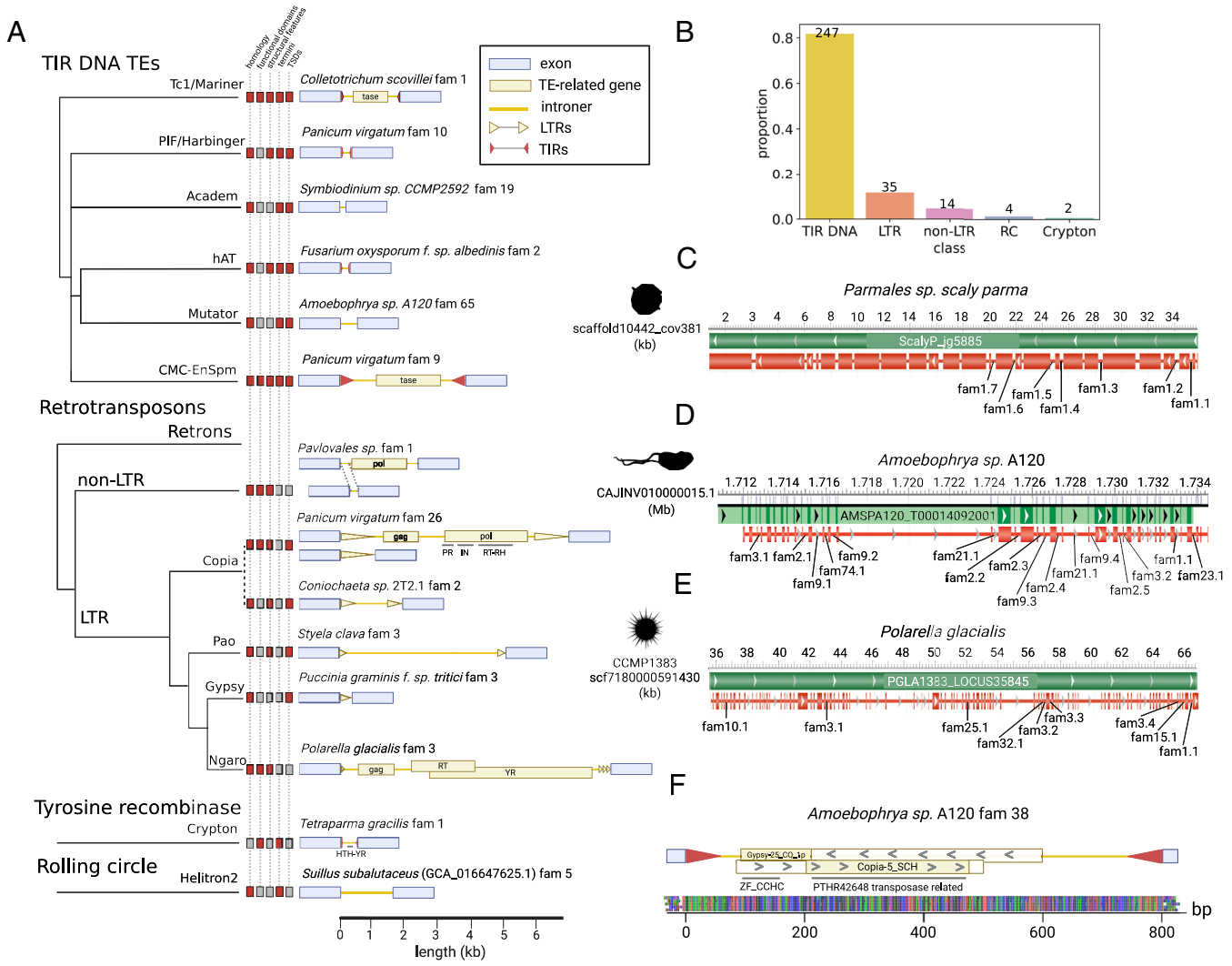

Fig. 2. Functional diversity and exemplary introners for major TE families. (*A*) Cladogram displaying relationships among introner transposable elements families (following ref. 26) and examples of the introner structure for each introner family identified in this work. Cladogram leaf tips denote lines of evidence supporting introner classifications, including homology to known TEs, expected functional domains, structural features such as LTRs or TIRs, conserved termini, and expected TSDs (or lack thereof in the case of cryptons and helitrons) (*Materials and Methods*). For introner structures, blue boxes denote flanking host exons. Introner features are shown in gold, with the exception of TIRs, which are plotted in red for visibility. Open reading frames (ORFs) with homology to known transposable element-associated genes are labeled using the following notation: tase = transposase, gag = capsid gene, pol = polymerase, RT = reverse transcriptase, YR = tyrosine recombinase. Functional domains are annotated using black lines beneath introner models and include the following abbreviations: PR = protease, IN = integrase, RT = reverse transcriptase, RH = ribonuclease, HTH-YR = helix-turn-helix domain associated with tyrosine recombinase. (*B*) Proportion of introner families represented by each major TE subclass for subset of introners which could be confidently classified. Numbers on bars display total counts for each subclass. (*C*–*E*) Genome browser screen shots for selected introner-containing genes, generated using NCBI's Genome Data Viewer. Introner positions are annotated for different introner families in each species (Dataset S3). Species silhouettes were retrieved from PhyloPic. (*F*) Model of *Amoebophrya* sp. A120 family 38. Introner ORFs are shown as boxes with regions showing homology to known TEs shaded in gold. Functional domains spanning these regions of homology are labeled below, followed by a multiple sequence alignment of all intron-generating copies from this introner family.

example, in the marine diatom *Parmales sp. scaly parma*, one introner family contributes to >91% of recognizable intron gains via an unknown mechanism (Fig. 2*C*). In contrast, the parasitic dinoflagellate *Amoebophrya* sp. A120 harbors introners attributed to non-LTR retrotransposons, LTR retrotransposons, and diverse TIR DNA transposons (Dataset S2). These diverse introners contribute to massive intron gain in *Amoebophrya*, in some cases generating tens of introns in a single gene (Fig. 2*D*). Similarly to *Amoebophrya*, the free living dinoflagellate, *Polarella glacialis*, harbors a tremendous diversity of introners contributing to ongoing intron gain (Fig. 2*E* and Dataset S2).

Beyond transposon families that we can confidently identify, some introners are derived from unknown transposition mechanisms. In particular, in *Amoebophrya* sp. A120, one intron-generating transposon family displays hallmarks of DNA transposons but shares homology to LTR retrotransposons. This element contains open reading frames with homology to *Copia* and *Gypsy* LTR elements but is flanked by DNA element-like terminal inverted repeats and exhibits several near-identical copies (Fig. 2*F* and *SI Appendix*, Fig. S9). Interestingly, this homologous region within *Copia-5_SCH* spans part of a putative transposase as annotated by Panther (PTHR42648) but does not overlap with any other specific functional domain associated with DNA transposons (*SI Appendix*, Fig. S10). This observation is reminiscent of Ginger DNA transposons, which employ transposases related to the integrases of LTR retrotransposons (32). However, the element identified in *Amoebophrya* sp. A120 does not show homology to previously identified Gingers and also lacks the termini and TSDs expected for Ginger elements (24). Several other introners in *Amoebophrya* sp. A120 show similar patterns. This observation might provide clues to the ancient relationships among LTR retrotransposons and DNA transposons (33, 34) and further highlights the still-unknown

diversity of transposable elements that are apparently capable of generating new introns across genomes.

Horizontal Gene Transfer Shapes Introner Distributions. Evidence of recent homology strongly implicates horizontal transmission as one of the major drivers of intron propagation within and among eukaryotic lineages. Transposable elements frequently transfer horizontally between lineages (35, 36) and HGT could be an important phenomenon shaping introner evolution (15). By performing an extensive blast search of all introner families identified in this work against the NCBI nucleotide database (*Materials and Methods* and Dataset S4), we found 8 unambiguous examples of recent horizontal transmission of introner-generating transposons (4% of all introner-containing species; Dataset S5). Horizontal transmission of transposable elements is therefore a critical biological process driving intron propagation between divergent populations.

We identified two examples of HGT within ascomycete fungi where the same elements have recently generated introns among highly divergent lineages (Fig. 3*A*). *Xylaria* and *Lasiodiplodia* last shared a common ancestor approximately 350 Mya (37, 38). Nonetheless, each species' genome contains an introner family with 83.8% sequence similarity across the complete 134 bp sequence between elements found in either lineage (Fig. 3*B* and Dataset S4). Furthermore, molecular dating estimates suggest that introners in these distant lineages last diverged a mere ~7 Mya (Fig. 3*B* and *SI Appendix*, Fig. S11). Similarly, *Alternaria* and *Parastagonospora* last shared a common ancestor approximately 134 Mya (Fig. 3*A*), but each contains a recently active introner family of length 64 bp with ~78.5% sequence similarity (Dataset S4) (37, 39). Estimates for *Alternaria* and *Parastagonospora* introners suggest they last diverged ~10 Mya (Fig. 3*C* and *SI Appendix*, Fig. S12). Notably, we identified introners in multiple genomes from different strains or species within these two genera, and introners found in each genome are most closely related to others found in genomes of the same genus. This implies that introner observations are not an idiosyncratic assembly artifact and strongly implicates recent transposition in highly divergent lineages rather than the transfer of an intron that did not continue transposing in the recipient population (Fig. 3*C*).

We also found cases of HGT between lineages in which introners in one species' genome may propagate as ordinary transposable elements in another. Introners in the dinoflagellate *P. glacialis* show strong homology to multiple loci in the glass sponge *Aphrocallistes beatrix* (up to 72% sequence similarity and 87% coverage; *SI Appendix*, Fig. S13 and Datasets S4, S5, S8, and S9). However, we find no evidence that these introner homologs contribute to intron gain in *A. beatrix*. The common ancestor between dinoflagellates and sponges is also the last eukaryotic common ancestor [>1.6 Bya (37, 40)] (Fig. 3*D*). Nonetheless, estimates suggest that *P. glacialis* introners and their *A. beatrix* homologs diverged less than 1 Mya (Fig. 3*E* and *SI Appendix*, Fig. S14). *P. glacialis* and *A. beatrix* share overlapping ranges in both Arctic and Antarctic regions (41–43). Furthermore, dinoflagellates and sponges exhibit frequent horizontal gene transfer, including interkingdom and interdomain transfer (44–46) (*SI Appendix*, Fig. S15 and *Supplementary Text*), implying that introner gain through HGT may be driven by these species' ecology (47, 48). These results imply that HGT of introners can occur between even the most divergent eukaryotic lineages.

Giant Viruses May Facilitate Horizontal Gene Transfer of Introners across Lineages. Viruses are frequent vectors of horizontal transfer and may underlie transmission of TEs between diverse taxa (49–55). A bioinformatic search across our dataset revealed several cases of introners within viral genes integrated within host genomes (Dataset S10 and *Materials and Methods*). An introner-containing gene in *Alternaria burnsii* (GeneID:62205305) is homologous to a klosneuvirus protein (MK072332.1; ~50% Sequence ID, e-value ~ 0), and the upstream protein also shares high levels of homology to a klosneuvirus virus (MK072078.1; e-value = 6.28E-17). Klosneuviruses are a clade of giant viruses that infect diverse eukaryotic hosts including fungi and protists, acquire host genes at high rates, and contribute to HGT (55, 56). Other examples of introner-containing genes homologous to klosneuviruses include the divergent dinoflagellates *Effrenium voratum* (gene EVOC421_LOCUS2994 and viral accession MT418680.1, ~54% SID, e-value ~ 0) and *P. glacialis* (gene PGLA1383_LOCUS8179 and viral accession KY684104.1.1, 91% SID, e-value = 2.42E-117). Overall, introner-containing genes are enriched for viral proteins (tblastn e-values < 1E-5; MWU, $P < 0.0001$ compared to expectations from random resampling). Among these, klosneuviruses are strongly overrepresented (9.4% of introner-containing genes that are viral in origin, whereas klosneuviruses comprise <0.0001 of sequences in RVDB; binomial test $P < 0.000001$). Furthermore, we observe no difference in splice site frequencies between introners in viral genes and other introners in the genome, suggesting purifying selection for correct splicing in viral genes postintegration. These findings suggest that viruses infecting diverse hosts may provide a mechanistic explanation for HGT of introners across divergent lineages.

Discussion

Our comprehensive survey of the tree of life provides direct evidence that HGT of transposable elements contributes to intron gain across eukaryotic lineages. Although other factors must contribute, we propose that HGT explains several foundational patterns in intron evolution: 1) the genome of the last eukaryotic common ancestor, which was a single-celled aquatic organism, contained thousands of introns, possibly due to an abundance of HGT of diverse mobile genetic elements (29, 57, 58); 2) introners and intron gain are disproportionately abundant in the genomes of aquatic and fungal taxa—both are known to experience high rates of HGT (15, 35, 48, 59, 60); and 3) conversely, the apparent lack of recent intron gain in lineages such as vertebrates (61, 62) may in part reflect the paucity of HGT and relative lack of TE diversity for many vertebrate lineages (59, 63, 64). Therefore, widespread HGT of intron-generating transposons resolves a fundamental question about why new introns evolve and what determines their abundances across species.

The abundance of evolutionarily and mechanistically diverse transposable elements that generate introns across the eukaryotic tree of life illuminates a longstanding mystery of evolution. Our results demonstrate that seemingly any variety of transposable element can and does generate introns within eukaryotic genomes. Efficient splicing after insertion into an exon should drastically reduce a TE's deleterious effect on host fitness and provides an adaptation both from the perspective of a transposable element and its host. The vast majority of introners are nonautonomous elements. While this pattern likely reflects stronger purifying selection on autonomous element insertions in genes relative to nonautonomous insertions, it could also result from the de novo evolution of introner function in nonautonomous families, which is consistent with the complete absence of autonomous introners observed for some elements. Selective pressure for this innovation should be most pronounced when a TE is highly active, such as after the introduction of a "new" TE unrecognizable to host suppression machinery via HGT (36). Thus, HGT may not only facilitate introner transmission between divergent lineages but also

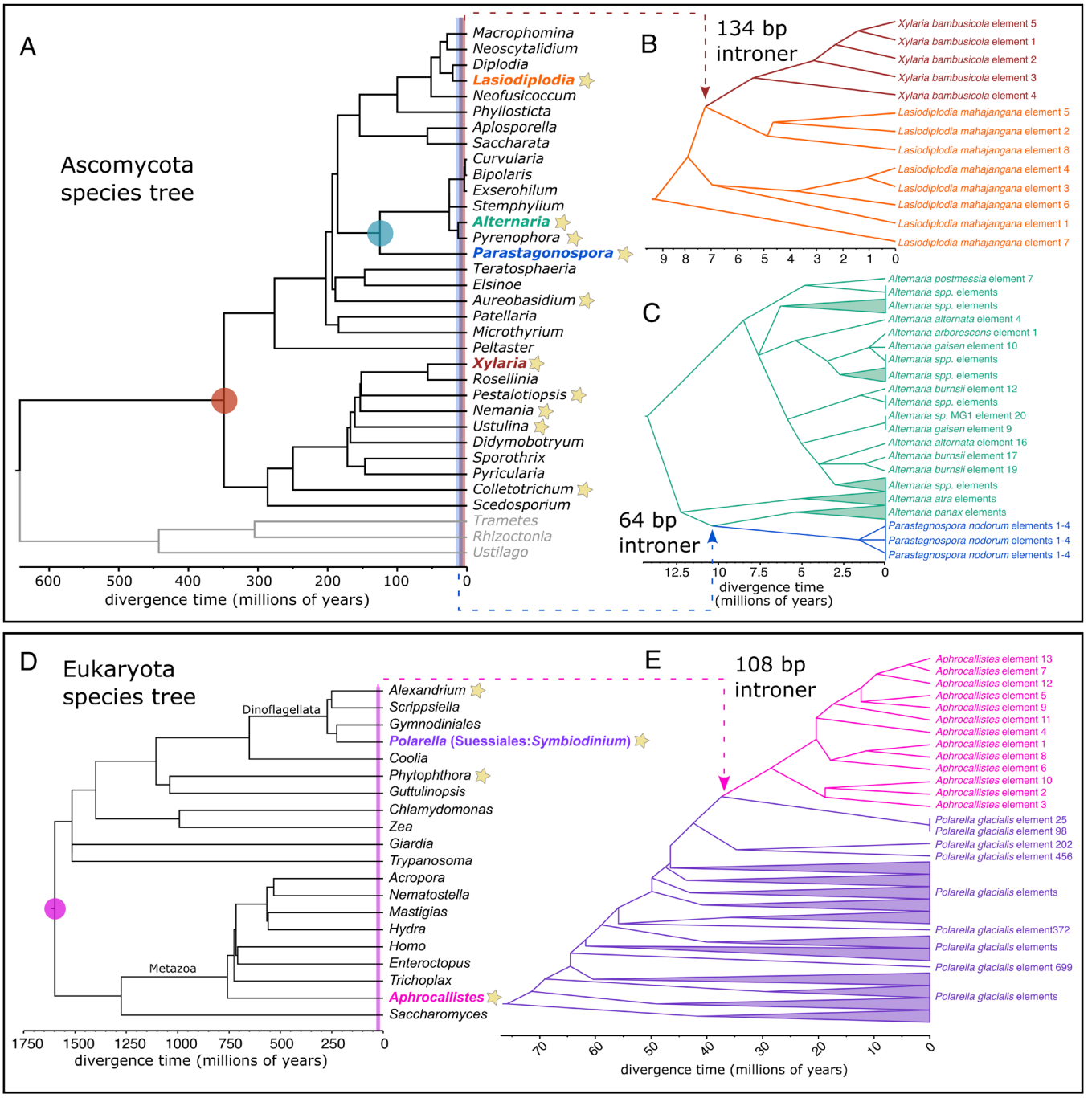

Fig. 3. Horizontal gene transfer of introners between divergent taxa. (*A*) Phylogeny of ascomycete fungi retrieved from TimeTree (37). Lineages showing evidence of HGT for introners are highlighted with colored font. Shaded circles denote the last common ancestor for species showing introner HGT and vertical lines show estimated divergence times for homologous introners between lineages with brown and blue corresponding to *Xylaria+Lasiodiplodia* and *Parastagonospora+Alternaria*, respectively. Stars indicate other genera with known introners (Dataset S6). (*B* and *C*) Phylogenies for prospective horizontally transferred introner families in (*B*) *Xylaria* and *Lasiodiplodia*, and (*C*) *Parastagonospora* and *Alternaria*. Colors in (*B*) and (*C*) correspond to leaf labels in (*A*). (*D*) Phylogeny of eukaryotes retrieved from TimeTree (Dataset S7) (37). A shaded circle denotes the last common ancestor between the introner-containing dinoflagellate lineage *Polarella* (purple) and the sponge lineage *Aphrocallistes* (pink), which contains homologous sequences but no evidence of introner-mediated intron gain. A shaded line denotes the estimated divergence between *Polarella* introners and *Aphrocallistes* homologs. Stars indicate other genera where we identified introners (Dataset S7). (*E*) Phylogeny of *Polarella* introners and *Aphrocallistes* homologs.

favor their evolution de novo from diverse TEs. The fundamental implication is that introns will accumulate as a likely consequence of the ubiquitous genetic conflict between eukaryotic genomes and transposable elements.

Materials and Methods

Retrieving Relevant Genomic Data. We retrieved all annotated genomes available from the GenBank and RefSeq databases using NCBI's command line datasets tool (retrieved 08/28/2023). Metadata for all genomes accessed in this way are presented in Dataset S1.

Systematic Candidate Introner Identification. We used a previously developed pipeline for systematic introner identification (https://github.com/lgozasht/Introner-elements). A detailed description of this pipeline can be found in ref. 15. Briefly, for each annotated genome, we extracted all introns and then performed an all vs. all blast (65) to search for highly similar introns. These introns were clustered based on sequence similarity (e-value < 1E−5), consistent with

signatures expected for recently active transposable elements. This clustering produced families of candidate introners.

Since several alternative reasons could account for sequence similarity between introns, we then performed several filtering steps. First, we filtered candidate introners which displayed sequence similarity due to secondary insertion of transposable elements into preexisting introns. To do this, we required that sequence similarity extended to the exon–intron boundaries, but not into the exons for each intron in the expected orientation (5′ regions aligned with 5′ regions and 3′ regions aligned with 3′ regions across candidate introners in a given family). Second, we removed introns which displayed sequence similarity as a result of whole gene duplication. To do this, we filtered candidate introners which displayed sequence similarity extending into exons, assuming that duplication of an entire intron without any flanking exonic sequence is unlikely. While this filter removed the vast majority of false positives due to paralogous gene duplication, our previous work suggests that rapid exonic evolution (or simple chance substitution) can lead to paralogous sequences being retained in some cases (such as for the large, fast-evolving var gene families of *Plasmodium* species) (15). To filter the remaining false positives caused by introns in paralogous genes, we translated all introner-containing genes and used diamond (66) (version 0.9.24) (e-value < 1E-20) to identify cases of sequence similarity between encoded proteins (paralogy groups) in each species. We retained all introner families with at least 4 sequences from genes in different paralogy groups that passed these filters. All introners considered in this study are publically available (67).

Manual Curation of Candidate Introners. Manual curation is essential for validating candidate introner and transposable element models more generally (15, 68). We performed several steps of manual inspection to validate candidate introners. First, we generated multiple sequence alignments for each candidate introner family using MAFFT (69) and viewed alignments using Aliview (70). We manually checked multiple sequence alignments to confirm that introner homology extends to near to the intron edges but not far into flanking exons and that homology boundaries are consistent across all introners in a given family. In doing so, we manually removed spurious introners resulting from gene duplications, which may have inadvertently passed our systematic filters (above). We also inspected introner sequences for low-complexity regions and removed introners for which sequence homology was primarily driven by simple repeats or satellites. Finally, we manually inspected introner sequences for signatures of spurious intron annotations. We checked that 5′ and 3′ splice sites were largely consistent for all introners in a given subfamily and further interrogated introner families which primarily showed inconsistent or unfamiliar splice sites. Canonical and noncanonical 5′ splice sites for major and minor introns generally include GT, GC, GA, and AT, whereas 3′ splice sites include AG and AC (71, 72). Due to the possibility of annotation errors, for introner families with primarily noncanonical splice sites, we manually inspected splice junctions in NCBI's genome data viewer to ensure correct splicing of introners and expression of introner-containing genes based on transcript or RNA-seq alignments. We discarded introner families which did not meet these requirements. We constructed consensus sequences and calculated Kimura divergence from the consensus for introner copies using RepeatModeler's utility tool, Refiner (73).

Introner TE Functional Classification. We classified introners as they relate to transposable elements using a combination of systematic and manual approaches and relied on five main sources of evidence for classification. These included homology to known TEs [based on RepeatClassifier (73) results and other homology searches using TE-AID (74) and MCHelper], the presence of expected protein domains [based on scans for known functional domains from interproscan (75, 76) and MCHelper (77)], structural features (e.g., LTRs and TIRs), element termini [based on scans using DFAM (78) models for known TE termini as well as manual inspection], and expected target site duplication lengths. First, we ran RepeatClassifier (73) on introner consensus sequences using all curated TE models available from the DFAM (78) and RepBase (79) databases. RepeatClassifier is a homology-based approach that prioritizes accuracy rather than sensitivity. Since many species in our data are divergent from those with curated TE models, we also classified introners de novo by annotating structural features and functional domains that might be consistent with transposition. Indeed, various transposable elements display structural features such as long terminal repeats or terminal inverted repeats, and the presence of certain protein domains, such as a reverse transcriptase, can be used to classify transposable elements.

We used TE-AID (74) to annotate and visualize introner sequence structures. TE-AID produces four different types of plots which aid in TE classification (see *SI Appendix*, Fig. S9 for example). First, TE-AID retrieves all prospective copies for a given TE family by blasting the consensus sequence at the respective reference genome and plots divergence from the consensus for all fragmented and full-length copies. Then, it plots coverage with respect to the consensus. This especially aids with classifying LTR retrotransposons since LTRs at element edges often exhibit much higher copy numbers than interior regions due to frequent ectopic recombination between LTRs resulting in the interior region's removal (30). TE-AID also produced self-v-self dotplots for TE-consensus sequences, allowing us to assess low-complexity regions as well as hallmark features of different TE mechanisms such as LTRs and TIRs. Finally, TE-AID produces a plot showing the locations of open reading frames (ORFs) within each TE-consensus as well as their homology to known TEs. We supplemented these homology searches with additional homology searches to TE-models in RepBase (79) and RepetDB (80).

We also sought to identify functional machinery and protein features within introner sequences. To do this, we first predicted and translated all possible ORFs in introner sequences using orfm (81). Then, we used interproscan (75, 76) (version 5.65-97.0) to identify functional protein-coding domains and features in all introner sequences. We also separately ran hmmer (82) on ORFs from introner consensus sequences to scan for TE-related functional domains using all available models from the PFAM and Gypsy databases (83, 84). Since specific protein machinery is associated with different known TE mechanisms, these results provided evidence for introner TE classifications.

For introners with no clear classification based on these methods, we ran MCHelper (77) to further extend and refine introner consensus sequences. MCHelper also automatically identifies structural features and functional domains in refined consensus sequences. Due to the challenges associated with automated transposable element extension and refinement (74), we hypothesized that some introner consensus sequences produced by Refiner may have been overextended or may exhibit other problems that inhibit automated feature discovery and classification. Thus, we ran MCHelper using three different inputs for each introner: 1) using the entire consensus, 2) using the first half of the consensus as input (and again using the second half if the first half yielded no results) and 3) using the middle 100 bp as input (to account for consensus overextension). This allowed us to classify additional introners with previously ambiguous classifications (Dataset S2). Since many introners are nonautonomous and are challenging to classify due to their lack of functional transposition machinery, we also retrieved HMMs for conserved termini of known transposable elements available from DFAM. We used nhmmer (85) to search for these conserved termini in introner families, filtering for hits with E < 0.001. We classified additional introners based on the presence of these conserved termini at expected positions in introner consensus sequences. Finally, we used probability weight matrices to systematically identify TSDs flanking each introner and manually inspected introners with pre-existing classifications to ensure that TSD lengths matched expectations from previously observed elements (Dataset S2).

Validating Introner-Containing Taxa. We performed additional steps to validate the presence and correct assembly of introners in several divergent taxa where evidence for introners was previously limited, which include *S. purpuratus* (purple sea urchin), *P. virgatum* (switchgrass), and *Styela clava* (tunicate). While this list is not exhaustive, it captures the expanded diversity of introner-containing species. Generally, we checked for splicing of the introners based on mapped RNA-Seq reads and looked at orthologous genes in a closely related species and a more distant relative to determine whether the insertion caused intron gain. For *S. purpuratus* and *S. clava*, we utilized the genome browser available on NCBI for RefSeq genomes to view mapped RNA-Seq reads and orthologous genes, and used Phytozome (86) to do this for *P. virgatum*. For *S. purpuratus*, we additionally utilized available PacBio datasets to check that mapped long reads spanned the introner regions and did not indicate a deletion of the putative introner (*SI Appendix*, Figs. S1–S7). *S. purpuratus* long read accessions used for mapping and the number of reads that mapped from each are listed in Dataset S11. In all cases, we were able to successfully validate the presence of intron-generating introners through long read coverage that extended well into the flanking portions of the genome from introner insertions.

Searching for Horizontal Gene Transfer. To search for possible HGT of introners, we performed blastn searches (65) for each introner consensus sequence against the NCBI nucleotide and RefSeq reference genome databases (87) (accessed 05/15/2024). Then, we filtered for hits with e-values < 0.0001, percent sequence ID > 70 and coverage > 60% in genera other than the source species'. These results are displayed in Dataset S4. We emphasize that these cutoffs are conservative and we expect to only find the most recently horizontally transferred elements between highly genetically divergent host species about which we can be the most confident. Furthermore, many genera contain exceptionally diverse species, and this simple approach may exclude HGT that occurs between distantly related species that nonetheless share a single genus. Then, we manually inspected results to filter out false positives resulting from conserved regions of transposons [e.g., regions of the RT domain in retrotransposons (88)] as well as low-quality subject sequences.

These efforts identified 8 candidate instances of HGT of recently active introners (Dataset S5). Together, introner HGT events involve 11/201 introner-containing lineages (5.5% of surveyed species). Some HGT events (4/9) involve introners actively generating introns in both lineages, whereas for other HGT events (5/9), introners are only actively generating introns in one lineage and not the other. For example, the ascomycete fungi, *Xylaria* and *Lasiodiplodia* (diverged 350 Mya) and *Alternaria* and *Parastagonospora* (diverged 134 Mya) both display highly similar active introner families driving ongoing intron gain (37–39). In contrast, introners in the lichen *Amylostereum chailletii* display high sequence similarity to sequences in *Stereum hirsutum* (112 MY diverged; Datasets S4 and S5). However, we do not observe introner-mediated intron gain in *Stereum hirsutum* (Dataset S2). When evaluating evidence of HGT, we retrieved homologous sequences from distantly related species iteratively using NCBI's entrez tool through BioPython (89).

We performed additional confirmation of HGT between the dinoflagellate, *P. glacialis*, and the glass sponge, *A. beatrix*, as this represents the HGT event between the most distantly related eukaryotic lineages that we detected using this approach. The correct assembly of introners in *P. glacialis* was determined by mapping long reads, as described above for confirming bacterial insertions. We also utilized this method to check for the correct assembly of regions in *A. beatrix* with homology to *P. glacialis* introner sequences, as found by the blastn searches. This allowed us to successfully validate the presence of introners in *P. glacialis*, as well as their non-intron-generating counterparts in *A. beatrix* (*SI Appendix*, Fig. S13 and Datasets S8 and S9). We used this same approach to validate HGT of introners and non-intron-generating homologs between the diatom, *Thalassiosira oceanica,* and the leech, *Piscicola geometra* (*SI Appendix*, Fig. S16 and Dataset S12).

Phylogenetic Inference. For instances of HGT, we attempted phylogenetic reconstruction to estimate the approximate time of transfer and to investigate patterns of transmission within and across the host genome. To do this, we performed multiple sequence alignment of introner families (or regions homologous to introners in the source species) using MAFFT with the --adjustdirectionaccurately flag. Then, we calculated mean percent identity for all interspecies introner comparisons.

Maximum likelihood phylogenies were inferred from the ascomycete introner alignments with W-IQ-Tree (90), using the ModelFinder option, 0.5 perturbation strength, 100 unsuccessful iterations till stop, generalized midpoint root optimization, and 100 standard bootstrap replicates.

To isolate the informative sites in the highly polymorphic dinoflagellate and sponge introner sequence alignments, we used ClipKIT (91) to trim regions with gaps in greater than 90% of sequences and retain only parsimony-informative and constant sites (mode: kpic-gappy). Next, we inferred maximum likelihood phylogenies from the dinoflagellate and sponge introner elements with IQ-Tree (92) using the ModelFinder option and 100 standard bootstrap replicates.

We estimated time-scaled phylogenies with TreeTime (93), specifying sampling dates from NCBI and substitution rates from the literature. For ascomycete fungi, neutral rates range from 1e−8 to 1e−9 substitutions per site per year (94, 95). Consistent with this value, the substitution rate for full-length LTR retrotransposons in fungi has been estimated to be 1.3e−8 substitutions per site per year (96). We used a substitution rate of 1.3e−8 for our ascomycete introner divergence time estimates. While substitution rates have not yet been measured for any sponges, they have been measured for a wide range of invertebrate taxa. The estimated genome-wide neutral substitution rate for *Alpheus* snapping shrimp is 2.64e−9 substitutions per site year (97) and the average rate across insects and mollusks is estimated to be 4.40e-9 per site per year (98). In diatoms, the base substitution mutation rate per site per generation has been measured to be 4.77e−10 and approximately 1.32 generations occur per day (99). This short generation time equates to 2.30e-7 substitutions per site per year (4.77e−10 subs/site/generation * 365 d/y * 1.32 generations/day). From these rates, we used the lowest (2.64e-9) for our introner divergence date estimates because it is the most conservative.

We caution that due to the exceptionally short lengths of sequences considered that there is substantial phylogenetic uncertainty in the resulting trees (*SI Appendix*, Figs. S11–S13 contain bootstrapped phylogenies). Nonetheless, the overarching patterns are readily apparent and support our interpretation.

To compare introner divergence time estimates to organismal divergence times, we obtained dated species trees from the TimeTree database (37) for each of the focal taxonomic groups containing introner HGT events. The *Polarella* clade was not represented in TimeTree. So, we identified *Symbiodinium* as another dinoflagellate in the same order (*Suessiales*), which was available. Given the deep split between *Symbiodinium* and its sister group, the Gymnodiniales order, the topology of the tree is likely the same for *Symbiodinium* as it would be for *Polarella*, if it had been sampled. Multiple sequence alignments underlying introner phylogenies in Fig. 3 *B*, *C*, and *E* are available at https://github.com/lgozasht/Introner-elements/tree/main/msa.

Searching for Introners in Genes of Viral Origin. To identify introners in genes of viral origin, we retrieved all introner-containing proteins from NCBI and performed tblastn (65) searches against the Reference Viral DataBase (100) (RVDB v28.0; accessed May 2, 2024). We then filtered for hits with e-values < 0.0001. This revealed several high-confidence instances of introner-containing genes of viral origin (Dataset S10). To test for overrepresentation of introner-containing genes among genes with homology to viral proteins, we randomly resampled genes from each introner-containing genome (N = number of introner-containing genes). We performed tblastn searches against RVDB with these randomly sampled genes to derive an expected distribution of genes showing homology to viral proteins in each species. Then, we performed two statistical tests for enrichment. First, we compared the observed distribution of candidate introner-containing viral-derived genes to an expected distribution aggregated across species using a nonparametric MWU test ($P < 0.00001$). Second, we performed a two-way binomial test comparing the number of species with a greater number of introner-containing genes represented among viral proteins than expected ($P < 0.0001$). To search for possible overrepresentation of introner-containing genes among specific viral lineages, we first retrieved NCBI lineage taxonomic data for each viral accession in RVDB using ete3 (101, 102). This revealed that ~10% of introner-containing genes with homology to viral proteins mapped to proteins in klosneuviral genomes. To test for overrepresentation of introner-containing proteins among klosneuviral sequences, we performed a two-way binomial test where the null expectation was the proportion of nucleotides in RVDB represented by klosneuviral sequences.

Fig. 1 Cladogram. To generate a cladogram of eukaryotic lineages, we first retrieved NCBI taxonomy IDs for the phylum and order corresponding to each considered species. Then, we assembled a list of nonredundant phyla from these results. For eukaryotic lineages with no taxonomic ID for phylum, we used order instead. Then, we submitted this list of nonredundant taxonomy IDs to Interactive Tree of Life (103), which produced a cladogram in newick format. The resulting tree was manipulated using ete3 (102) and visualized using toytree (104).

Data, Materials, and Software Availability. Our introner identification pipeline and sequences for all introner families are available at https://github.com/lgozasht/Introner-elements (67). Previously published data were used for this work (All genomic data are available from NCBI. Specific accession numbers are listed Datasets S1, S4, S8, S9, and S11–S13.).

ACKNOWLEDGMENTS. L.G. thanks Hopi E. Hoekstra for her support and advice throughout this work. The computations for this work were run on the Cannon cluster supported by the Faculty of Arts and Sciences Division of Science Research Computing Group at Harvard University. We thank Erik Enbody, Jennifer Chen, Peter Sudmant, and David Haussler for helpful feedback on this manuscript. This work was supported in part by NIH/NIGMS R35GM128932 to R.C.-D. and R00GM135583 to S.R. A.N. was supported by an NSF GRFP and by NIH/NHGRI T32HG012344.

1. A. Bonnet *et al.*, Introns protect eukaryotic genomes from transcription-associated genetic instability. *Mol. Cell* **67**, 608–621.e6 (2017).
2. O. Shaul, How introns enhance gene expression. *Int. J. Biochem. Cell Biol.* **91**, 145–155 (2017).
3. M. Irimia, S. W. Roy, Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb. Perspect. Biol.* **6**, a016071 (2014).
4. J. Parenteau *et al.*, Deletion of many yeast introns reveals a minority of genes that require splicing for function. *Mol. Biol. Cell* **19**, 1932–1941 (2008).
5. S. J. Gould, R. C. Lewontin, The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proc. R. Soc. Lond. B Biol. Sci.* **205**, 581–598 (1979).
6. W. Martin, E. V. Koonin, Introns and the origin of nucleus-cytosol compartmentalization. *Nature* **440**, 41–45 (2006).
7. M. Csuros, I. B. Rogozin, E. V. Koonin, A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput. Biol.* **7**, e1002150 (2011).
8. I. B. Rogozin, Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, E. V. Koonin, Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* **13**, 1512–1517 (2003).
9. E. Shoguchi *et al.*, Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Curr. Biol.* **23**, 1399–1408 (2013).
10. S. W. Roy, W. Gilbert, The evolution of spliceosomal introns: Patterns, puzzles and progress. *Nat. Rev. Genet.* **7**, 211–221 (2006).
11. H. G. Morrison *et al.*, Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. *Science* **317**, 1921–1926 (2007).
12. P. Yenerall, L. Zhou, Identifying the mechanisms of intron gain: Progress and trends. *Biol. Direct* **7**, 29 (2012).
13. J. T. Huff, D. Zilberman, S. W. Roy, Mechanism for DNA transposons to generate introns on genomic scales. *Nature* **538**, 533–536 (2016).
14. A. Z. Worden *et al.*, Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324**, 268–272 (2009).
15. L. Gozashti *et al.*, Transposable elements drive intron gain in diverse eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2209766119 (2022).
16. F. Denoeud *et al.*, Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* **330**, 1381–1385 (2010).
17. S. Farhat *et al.*, Rapid protein evolution, organellar reductions, and invasive intronic elements in the marine aerobic parasite dinoflagellate *Amoebophrya* spp. *BMC Biol.* **19**, 1 (2021).
18. S. W. Roy *et al.*, Intron-rich dinoflagellate genomes driven by Introner transposable elements of unprecedented diversity. *Curr. Biol.* **33**, 189–196.e4 (2023).
19. A. van der Burgt, E. Severing, P. J. G. M. de Wit, J. Collemare, Birth of new spliceosomal introns in fungi by multiplication of introner-like elements. *Curr. Biol.* **22**, 1260–1265 (2012).
20. S. F. F. Torriani, E. H. Stukenbrock, P. C. Brunner, B. A. McDonald, D. Croll, Evidence for extensive recent intron transposition in closely related fungi. *Curr. Biol.* **21**, 2017–2022 (2011).
21. J. Collemare, A. van der Burgt, P. J. G. M. de Wit, At the origin of spliceosomal introns: Is multiplication of introner-like elements the main mechanism of intron gain in fungi? *Commun. Integr. Biol.* **6**, e23147 (2013).
22. B. Verhelst, Y. Van de Peer, P. Rouzé, The complex intron landscape and massive intron invasion in a picoeukaryote provides insights into intron evolution. *Genome Biol. Evol.* **5**, 2393–2401 (2013).
23. T. Wicker *et al.*, A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
24. K. K. Kojima, Structural and sequence diversity of eukaryotic transposable elements. *Genes Genet. Syst.* **94**, 233–252 (2020).
25. I. R. Arkhipova, Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob. DNA* **8**, 19 (2017).
26. J. N. Wells, C. Feschotte, A field guide to eukaryotic transposable elements. *Annu. Rev. Genet.* **54**, 539–561 (2020).
27. M. J. Curcio, K. M. Derbyshire, The outs and ins of transposition: From mu to kangaroo. *Nat. Rev. Mol. Cell Biol.* **4**, 865–877 (2003).
28. M. Lynch, B. Walsh, *The Origins of Genome Architecture* (Sinauer Associates, Sunderland, MA, 2007).
29. M. Krupovic, V. V. Dolja, E. V. Koonin, The virome of the last eukaryotic common ancestor and eukaryogenesis. *Nat. Microbiol.* **8**, 1008–1017 (2023).
30. K. M. Devos, J. K. M. Brown, J. L. Bennetzen, Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12**, 1075–1079 (2002).
31. M. Chandler, N. Craig, "Helitrons, the eukaryotic rolling-circle transposable elements" in *Mobile DNA III*, M. Chandler, N. Craig, Eds. (American Society of Microbiology, 2015), pp. 893–926.
32. W. Bao, V. V. Kapitonov, J. Jurka, Ginger DNA transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons. *Mob. DNA* **1**, 3 (2010).
33. P. Capy, R. Vitalis, T. Langin, D. Higuet, C. Bazin, Relationships between transposable elements based upon the integrase-transposase domains: Is there a common ancestor? *J. Mol. Evol.* **42**, 359–368 (1996).
34. P. Capy, T. Langin, D. Higuet, P. Maurer, C. Bazin, "Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor?" in *Evolution and Impact of Transposable Elements*, P. Capy, Ed. (Springer Netherlands, 1997), pp. 63–72.
35. H. Qiu, G. Cai, J. Luo, D. Bhattacharya, N. Zhang, Extensive horizontal gene transfers between plant pathogenic fungi. *BMC Biol.* **14**, 41 (2016).
36. S. Schaack, C. Gilbert, C. Feschotte, Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol. Evol.* **25**, 537–546 (2010).
37. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, TimeTree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
38. M. Prieto, M. Wedin, Dating the diversification of the major lineages of Ascomycota (Fungi). *PLoS One* **8**, e65576 (2013).
39. M. A. Van der Nest *et al.*, Saprophytic and pathogenic fungi in the Ceratocystidaceae differ in their ability to metabolize plant-derived sucrose. *BMC Evol. Biol.* **15**, 273 (2015).
40. S. Blair Hedges, S. Kumar, *The Timetree of Life* (OUP Oxford, 2009).
41. J. Copley *et al.*, Megafauna from sublittoral to abyssal depths along the Mid-Atlantic Ridge south of Iceland. *Oceanol. Acta* **19**, 549–559 (1996).
42. J. Stephens, Atlantic sponges collected by the Scottish National Antarctic Expedition. *Earth Environ. Sci. Trans. R. Soc. Edinb.* **50**, 423–467 (1915).
43. M. Montresor, C. Lovejoy, L. Orsini, G. Procaccini, S. Roy, Bipolar distribution of the cyst-forming dinoflagellate *Polarella glacialis*. *Polar Biol.* **26**, 186–194 (2003).
44. C. Rot, I. Goldfarb, M. Ilan, D. Huchon, Putative cross-kingdom horizontal gene transfer in sponge (Porifera) mitochondria. *BMC Evol. Biol.* **6**, 71 (2006).
45. C. Conaco *et al.*, Detection of prokaryotic genes in the *Amphimedon queenslandica* genome. *PLoS One* **11**, e0151092 (2016).
46. J. H. Wisecaver, M. L. Brosnahan, J. D. Hackett, Horizontal gene transfer is a significant driver of gene innovation in dinoflagellates. *Genome Biol. Evol.* **5**, 2368–2381 (2013).
47. X. Wang, X. Liu, Close ecological relationship among species facilitated horizontal transfer of retrotransposons. *BMC Evol. Biol.* **16**, 201 (2016).
48. L. D. McDaniel *et al.*, High frequency of horizontal gene transfer in the oceans. *Science* **330**, 50 (2010).
49. S. A. Widen *et al.*, Virus-like transposons cross the species barrier and drive the evolution of genetic incompatibilities. *Science* **380**, eade0705 (2023).
50. V. Loiseau *et al.*, Monitoring insect transposable elements in large double-stranded DNA viruses reveals host-to-virus and virus-to-virus transposition. *Mol. Biol. Evol.* **38**, 3512–3530 (2021).
51. C. Gilbert, R. Cordaux, Viruses as vectors of horizontal transfer of genetic material in eukaryotes. *Curr. Opin. Virol.* **25**, 16–22 (2017).
52. C. Gilbert, C. Feschotte, Horizontal acquisition of transposable elements and viral sequences: Patterns and consequences. *Curr. Opin. Genet. Dev.* **49**, 15–24 (2018).
53. N. A. T. Irwin, A. A. Pittis, T. A. Richards, P. J. Keeling, Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat. Microbiol.* **7**, 327–336 (2022).
54. T.-W. Sun *et al.*, Host range and coding potential of eukaryotic giant viruses. *Viruses* **12**, 1337 (2020).
55. F. Schulz *et al.*, Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).
56. F. Schulz *et al.*, Giant viruses with an expanded complement of translation system components. *Science* **356**, 82–85 (2017).
57. L. Carmel, I. B. Rogozin, Y. I. Wolf, E. V. Koonin, Patterns of intron gain and conservation in eukaryotic genes. *BMC Evol. Biol.* **7**, 192 (2007).
58. M. Csűrös, "Likely scenarios of intron evolution" in *Comparative Genomics*, A. McLysaght, D. H. Huson, Eds. (Springer Berlin Heidelberg, 2005), pp. 47–60.
59. H.-H. Zhang, J. Peccoud, M.-R.-X. Xu, X.-G. Zhang, C. Gilbert, Horizontal transfer and evolution of transposable elements in vertebrates. *Nat. Commun.* **11**, 1362 (2020).
60. J. C. Slot, A. Rokas, Horizontal transfer of a large and highly toxic secondary metabolic gene cluster between fungi. *Curr. Biol.* **21**, 134–139 (2011).
61. S. W. Roy, A. Fedorov, W. Gilbert, Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 7158–7162 (2003).
62. I. B. Rogozin, L. Carmel, M. Csuros, E. V. Koonin, Origin and evolution of spliceosomal introns. *Biol. Direct* **7**, 11 (2012).
63. C. G. Sotero-Caio, R. N. Platt II, A. Suh, D. A. Ray, Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol. Evol.* **9**, 161–177 (2017).
64. J. O. Andersson, Lateral gene transfer in eukaryotes. *Cell. Mol. Life Sci.* **62**, 1182–1197 (2005).
65. C. Camacho *et al.*, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
66. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
67. L. Gozashti, A. Nakamoto, S. Russell, R. Corbett-Detig, Data from "Horizontal transmission of functionally diverse transposons is a major source of new introns." GitHub. https://github.com/lgozasht/Introner-elements. Deposited 7 September 2024.
68. J. M. Storer, R. Hubley, J. Rosen, A. F. A. Smit, Curation guidelines for de novo generated transposable element families. *Curr. Protoc.* **1**, e154 (2021).
69. K. Katoh, D. M. Standley, MAFFT Multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
70. A. Larsson, AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).
71. M. Burset, I. A. Seledtsov, V. V. Solovyev, Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28**, 4364–4375 (2000).
72. B. Pucker, S. F. Brockington, Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. *BMC Genomics* **19**, 980 (2018).
73. J. M. Flynn *et al.*, RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9451–9457 (2020).
74. C. Goubert *et al.*, A beginner's guide to manual curation of transposable elements. *Mob. DNA* **13**, 7 (2022).
75. P. Jones *et al.*, InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
76. M. Blum *et al.*, The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).
77. S. Orozco-Arias, P. Sierra, R. Durbin, J. González, MCHelper automatically curates transposable element libraries across species. *Genome Res.* **34**, 2256–2268 (2024).
78. R. Hubley *et al.*, The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).
79. J. Jurka, Repbase update: A database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
80. J. Amselem *et al.*, RepetDB: A unified resource for transposable element references. *Mob. DNA* **10**, 6 (2019).
81. B. J. Woodcroft, J. A. Boyd, G. W. Tyson, OrfM: A fast open reading frame predictor for metagenomic data. *Bioinformatics* **32**, 2702–2703 (2016).
82. S. R. Eddy, Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
83. J. Mistry *et al.*, Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
84. C. Llorens *et al.*, The Gypsy Database (GyDB) of mobile genetic elements: Release 2.0. *Nucleic Acids Res.* **39**, D70–D74 (2011).
85. T. J. Wheeler, S. R. Eddy, nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489 (2013).
86. D. M. Goodstein *et al.*, Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).
87. N. A. O'Leary *et al.*, Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).

88. Y. Xiong, T. H. Eickbush, Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**, 3353–3362 (1990).
89. P. J. A. Cock *et al.*, Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
90. J. Trifinopoulos, L.-T. Nguyen, A. von Haeseler, B. Q. Minh, W-IQ-TREE: A fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* **44**, W232–W235 (2016).
91. J. L. Steenwyk, T. J. Buida III, Y. Li, X.-X. Shen, A. Rokas, ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol.* **18**, e3001007 (2020).
92. B. Q. Minh *et al.*, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
93. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).
94. T. Kasuga, T. J. White, J. W. Taylor, Estimation of nucleotide substitution rates in Eurotiomycete fungi. *Mol. Biol. Evol.* **19**, 2318–2324 (2002).
95. M. L. Berbee, J. W. Taylor, Dating the molecular clock in fungi–How close are we? *Fungal Biol. Rev.* **24**, 1–16 (2010).
96. J. Labbé *et al.*, Characterization of transposable elements in the ectomycorrhizal fungus *Laccaria bicolor*. *PLoS One* **7**, e40197 (2012).
97. K. Silliman, J. L. Indorf, N. Knowlton, W. E. Browne, C. Hurt, Base-substitution mutation rate across the nuclear genome of *Alpheus* snapping shrimp and the timing of isolation by the Isthmus of Panama. *BMC Ecol. Evol.* **21**, 104 (2021).
98. R. Allio, S. Donega, N. Galtier, B. Nabholz, Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: Implications for genetic diversity and the use of mitochondrial DNA as a molecular marker. *Mol. Biol. Evol.* **34**, 2762–2772 (2017).
99. M. Krasovec, S. Sanchez-Brosseau, G. Piganeau, First estimation of the spontaneous mutation rate in diatom. *Genome Biol. Evol.* **11**, 1829–1837 (2019).
100. N. Goodacre, A. Aljanahi, S. Nandakumar, M. Mikailov, A. S. Khan, A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. *mSphere* **3**, e00069-18 (2018).
101. C. L. Schoch *et al.*, NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database (Oxford)* **2020**, baaa062 (2020).
102. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
103. I. Letunic, P. Bork, Interactive Tree of Life (iTOL) v6: Recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024), 10.1093/nar/gkae268.
104. D. A. R. Eaton, Toytree: A minimalist tree visualization and manipulation library for Python. *Methods Ecol. Evol.* **11**, 187–191 (2020).