Title

Transposable elements drive intron gain in diverse eukaryotes

Permalink

https://escholarship.org/uc/item/8rc239p7

Journal

Proceedings of the National Academy of Sciences of the United States of America, 119(48)

ISSN

0027-8424

Authors

Gozashti, Landen
Roy, Scott W
Thornlow, Bryan
et al.

Publication Date

2022-11-29

DOI

10.1073/pnas.2209766119

Peer reviewed

PNAS RESEARCH ARTICLE | EVOLUTION OPEN ACCESS

Transposable elements drive intron gain in diverse eukaryotes

Landen Gozashti[a,b,1,2,3], Scott W. Roy[c,1,4], Bryan Thornlow[a,b], Alexander Kramer[a,b], Manuel Ares Jr.[d], and Russell Corbett-Detig[a,b,4]

Edited by W. Doolittle, Dalhousie University, Halifax, Canada; received June 10, 2022; accepted October 19, 2022

There is massive variation in intron numbers across eukaryotic genomes, yet the major drivers of intron content during evolution remain elusive. Rapid intron loss and gain in some lineages contrast with long-term evolutionary stasis in others. Episodic intron gain could be explained by recently discovered specialized transposons called Introners, but so far Introners are only known from a handful of species. Here, we performed a systematic search across 3,325 eukaryotic genomes and identified 27,563 Introner-derived introns in 175 genomes (5.2%). Species with Introners span remarkable phylogenetic diversity, from animals to basal protists, representing lineages whose last common ancestor dates to over 1.7 billion years ago. Aquatic organisms were 6.5 times more likely to contain Introners than terrestrial organisms. Introners exhibit mechanistic diversity but most are consistent with DNA transposition, indicating that Introners have evolved convergently hundreds of times from nonautonomous transposable elements. Transposable elements and aquatic taxa are associated with high rates of horizontal gene transfer, suggesting that this combination of factors may explain the punctuated and biased diversity of species containing Introners. More generally, our data suggest that Introners may explain the episodic nature of intron gain across the eukaryotic tree of life. These results illuminate the major source of ongoing intron creation in eukaryotic genomes.

intron | splicing | genome structure | evolution | comparative genomics

Significance

Introns are a crucial part of eukaryotic genomes, but their origins are poorly understood. Some lineages exhibit large-scale gains in introns extremely rapidly. This pattern might be explained by a type of genetic element, Introners, that creates copies of itself that insert into many genes across the genome. We searched thousands of eukaryotic genomes for Introners and found them in 5% of all species. Introners evolved convergently from many distinct genetic elements, most are consistent with DNA-based transposable elements, and they are disproportionately common in the genomes of aquatic organisms. We propose that horizontal transfer of transposons in aquatic taxa contributes to the biased and highly punctate evolution of intron gains across eukaryotes.

The forces shaping intron–exon structures of eukaryotic genes remain among the longest-standing mysteries of molecular biology. Eukaryotic genomes contain from zero to hundreds of thousands of spliceosomal introns (1). Given the diverse roles of introns in gene expression and genome stability, from transcription enhancement to transcript surveillance to alternative splicing to R-loop avoidance (2–5), these differences may have important functional implications. Intron numbers per gene and per genome exhibit complex phylogenetic patterns, indicating massive recurrent changes in intron numbers through evolution, and comparative analyses attest to important roles for both intron deletion (loss) and creation (gain) (1, 6, 7). Despite decades of debate, no consensus has emerged as to either the proximal or ultimate explanations for these patterns.

Diverse molecular mechanisms of de novo intron creation are known, but their relative contributions to genome evolution across the tree of life remain poorly understood. Proposed mechanisms of de novo intron creation include inexact double strand break repair (8), mitochondrial DNA insertion (9), internal gene duplication (10), and "intronization" of exonic sequence (11). In addition to these ad hoc intron creation mechanisms, the intron-generating transposable elements (TEs) known as Introners represent a mechanism that could explain the high and episodic frequency and genome-wide scale of intron gains observed across eukaryotic lineages. These poorly understood TEs create introns de novo through insertion into exons (12–17). Introners have only been described in five eukaryotic lineages, and even among these cases, the precise molecular mechanisms remain obscure. Some Introner families show clear signatures of DNA TEs (12, 15), while others may be novel RNA-propagated elements (14, 16). More importantly, determining the extent to which Introners are a primary source of ongoing intron gain is essential for interpreting the evolution of genome structure and function and requires a broad survey that spans the eukaryotic tree of life.

By performing a systematic search and in-depth analysis of intron gain across all available eukaryotic genomes, we identified primary shared drivers of intron gain in diverse eukaryotic lineages. Our search identified 27,563 Introner-derived introns from 548 distinct families, with Introners found in 175/3,325 (5.2%) of studied genomes. Introner-containing species span remarkable phylogenetic diversity, from copepods to poorly understood basal protists, representing lineages whose last common ancestor dates back to ~1.7 billion years ago (18). Unexpectedly, aquatic organisms were 6.5 times more likely to contain Introners, and 74% of Introner-containing aquatic genomes harbored multiple distinct Introner families. Overrepresentation in aquatic organisms could reflect higher rates of lateral gene transfer. While we find that Introners are efficiently spliced,

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

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] L.G. and S.W.R. contributed equally to this work.

[2] Present address: Department of Organismic and Evolutionary Biology & Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138

[3] Present address: HHMI, Harvard University, Cambridge, MA 02138

[4] To whom correspondence may be addressed. Email: scottwroy@gmail.com or russcd@gmail.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2209766119/-/DCSupplemental.

Published November 23, 2022.

preferential presence in lowly expressed genes suggests that new insertions are costly. Most Introner families exhibit one or more signatures of DNA-based propagation. Our study indicates that susceptibility to acquire weakly deleterious Introners by lateral gene transfer might play the central role in a taxon's tendency to gain introns.

Results and Discussion

Introners Are Widespread Across Eukaryotes. Our survey across all available eukaryotic genomes indicates that Introners are abundant in diverse lineages. To search for Introners, we developed a pipeline to systematically identify groups of introns with similar sequences, for which the region of sequence similarity extends to near the splice boundary at both ends. This approach allows flexibility in identifying introns created by TE insertions through potentially complex mechanisms while excluding most cases where inter-intron similarities reflect secondary insertion of TEs or evolution of microsatellites within preexisting introns. We then applied this pipeline to 2,805 genomes representing 1,700 species with available genome annotations in Genbank (*SI Appendix*, Table 2). After extensive quality control (see *Methods*), our search revealed sets of Introners in 48 species grouping into eight distinct taxonomic groups representing six major eukaryotic groups. Although here we refer to each of these elements as "Introners," we do not mean to imply any direct evidence for shared homology among the various intron-generating TE families we describe here.

Introners Are Disproportionately Common in Aquatic Lineages. Introners are disproportionately common in aquatic lineages, suggesting an important relationship between external environment and rates of Introner evolution. To our surprise, 7/8 Introner-containing taxonomic groups (all except pezizomycotina fungi) inhabit aquatic environments. To further explore aquatic diversity, we analyzed 520 partial genomes from aquatic organisms, representing 71 distinct genera (19) (*SI Appendix*, Table 3). This revealed 25 additional Introner-containing taxonomic groups, yielding a total of 32 separate taxonomic groups. Each presumably represents independent acquisition/evolution of Introners (Fig. 1), suggesting a highly punctuated pattern of Introner presence across the eukaryotic tree with little phylogenetic signal ($P < 0.001$, abouheif's C_{mean} permutation test). Within this combined dataset, among 1,597 species for which aquatic/non-aquatic status was assignable, 17.0% of aquatic species (39/230) but only 2.6% of non-aquatic species (35/1367) exhibited at least one Introner family, confirming that aquatic habitat is significantly correlated with Introner presence ($P < 10^{-5}$, two-sided Fisher's exact test). Our results imply that aquatic environments are correlated with the evolution of Introners and that aquatic environments may be an important driver of intron gain.

A test of environment association that accounts for phylogenetic relationship strengthens the conclusion that the genomes of organisms in aquatic environments are disproportionately likely to harbor Introners. Because some Introner-containing lineages are closely related, the apparent marginal correlation between aquatic environments and Introner presence might be an idiosyncratic result of shared ancestry rather than an independently associated factor. We therefore retrieved a phylogeny of all eukaryotic species in our study, and we found that a model where the rate at which a lineage evolves Introners depends on the environment is a significantly better fit than a model where the rate of Introner gain is independent of the environment ($P < 4.1 \times 10^{-4}$ Pagel's test, *SI Appendix*, Table 4 and *SI Appendix*, Figs. S1 and S2, See *Methods*). This analysis therefore indicates that the strong correlation between environment and rates of Introner gain is not strictly a product of shared ancestry, but rather may reflect an important determinant of rates of intron gain.

Frequent Convergent Evolution of Introners from DNA Transposons. Introner abundance varies substantially across Introner-containing lineages and even between extremely closely related organisms. We identified 27,563 Introner-derived introns within 175 Introner-containing genomes, representing 548 putatively separate Introner families defined based on sequence similarity (*SI Appendix*, Table 6). Introner families exhibited substantial diversity in copy number per genome (5 to more than 2,000), length (median length 30–654 nucleotides), GC content (20.9 to 83.3%), and percent of rare GC and GA 5′ splice sites (0 to 50.0% and 0 to 40.0%, respectively) (*SI Appendix*, Table 6). Introners themselves generally exhibit lower GC content than exons in host genomes. This may be consistent with the observation that DNA transposons are GC poor relative to host eukaryotic protein-coding regions (20, 21). Genomes differed in the number of predicted Introner families, from one to 43 families (Fig. 1 and *SI Appendix*, Table 5). We also found large differences between closely related organisms. For example, among three *Micromonas* species, an unknown species (isolate TARA_MED_95_MAG_00390) had no detectable Introners, *M. commoda* had two families (44 total Introners), and *M. pusilla* had seven families (3,566 total Introners) with no sequence similarity to the *M. commoda* Introners. In *Florenciella*, some Introner families were found in all five isolates and others in only a subset. We found similar patterns in the highest quality genomes, suggesting that data quality issues do not drive these results. Our findings suggest that Introner presence and content are extremely evolutionarily labile, consistent with rapid changes in intron abundance observed across eukaryotic lineages.

Diverse molecular functions for intron gain suggest that many nonautonomous transposons have convergently evolved into Introners. Some previously characterized Introner families exhibit signatures of DNA cut-and-paste TEs, such as 3–9 bp repeats flanking their insertion site (target site duplications or TSDs) and/or inverted repeats at their 5′ and 3′ ends (terminal inverted repeats or TIRs), while others lack such signatures (12–14, 22). Among Introner families for which TSD and/or TIR presence/absence could be ascertained (both characteristics of DNA TEs), 49/130 show evidence for TSDs and 72/177 show evidence for TIRs. Conversely, other families lack these features, including cases where Introners have 100% end-to-end sequence identity. For example, in *M. pusilla* both TSD/TIR families and non-TSD/non-TIR families are present (*SI Appendix*, Table 6). Introners also show remarkable diversity in mechanisms of splice-site recruitment. Among families with TSDs, we found a variety of orientations of splicing boundaries with respect to the TSDs. These included cases where the Introner carries the 3′ splice site and the 5′ splice site is recruited from the TSD (Fig. 2*B*), cases in which the reverse is true (Fig. 2*C*), or more complex cases in which either both splice sites are entirely or partially recruited from the TSD or from neighboring exonic sequence (Fig. 2*E*). While most families followed previous reports in which insertion and splicing do not lead to a change in mRNA sequence length (12, 13), we also found cases where Introner insertion is associated with insertion or deletion of one or more neighboring codons (Fig. 2*F*). This exceptional range of molecular mechanisms suggests that Introners have independently evolved from diverse autonomous transposable elements and may explain recurrent bursts of intron gain across diverse eukaryotic lineages (7, 23).

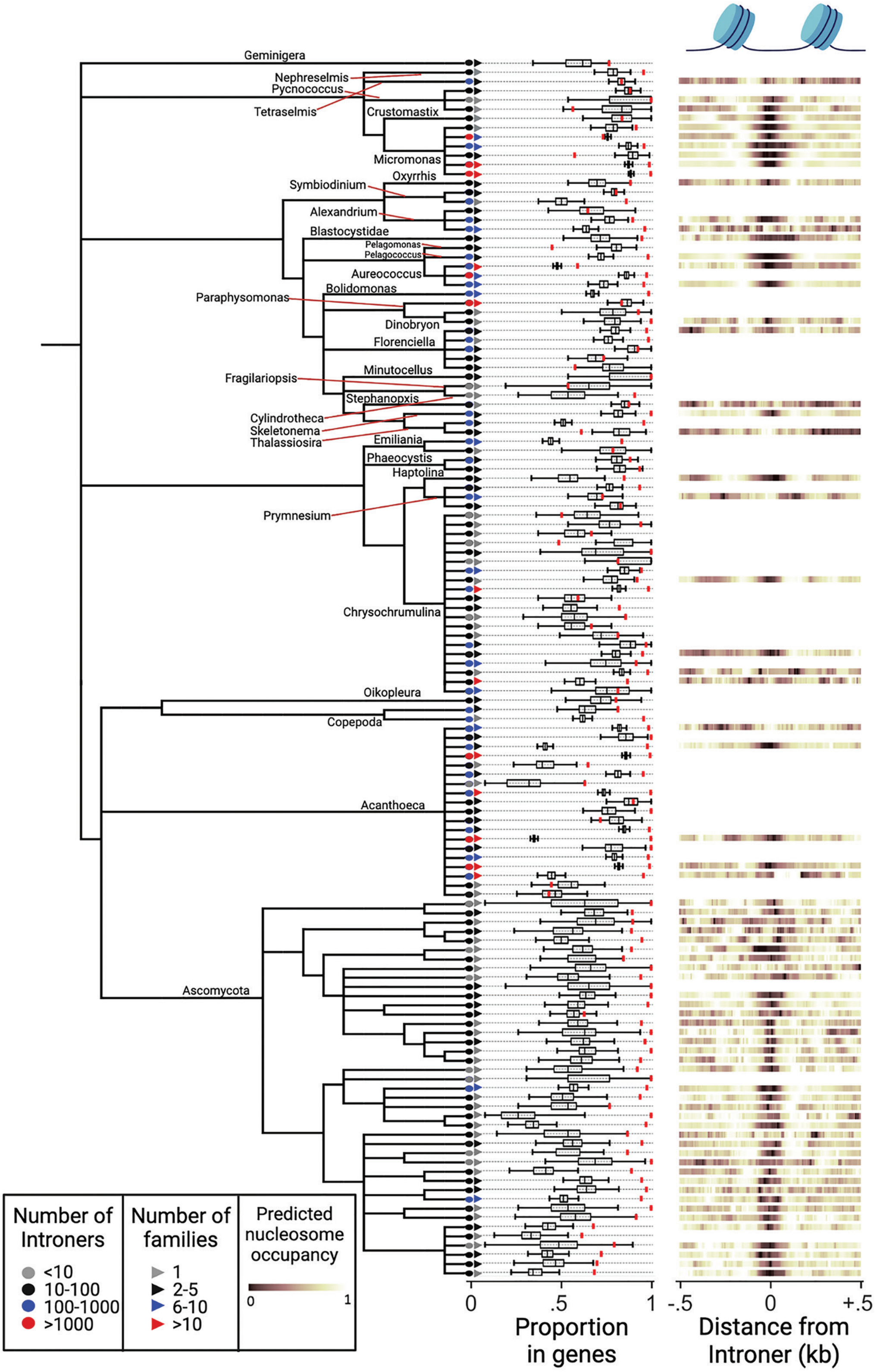

Fig. 1. Diversity and characteristics of Introners across eukaryotes. Results are shown from 130 genomes representing 32 lineages with putatively independent acquisitions of Introners (different colors). Leaf tip colors indicate the total number of predicted Introners for each genome. Proportion in genes is shown by the red mark, which consistently exceeds the expected values as determined by randomization within each genome (black box plots; center line denotes median; box limits denote upper and lower quartiles; whiskers denote 1.5x interquartile range). Heat maps represent predicted nucleosome occupancy for Introner insertion sites and surrounding genomic regions for genomes in which accurate nucleosome occupancy prediction was possible (see *Methods*). Introner insertion sites consistently show reduced histone occupancy (dark) relative to surrounding regions. For genomes within the same genus, multiple genomes are shown only if the genomes have different complements of Introner families (see *Methods*).

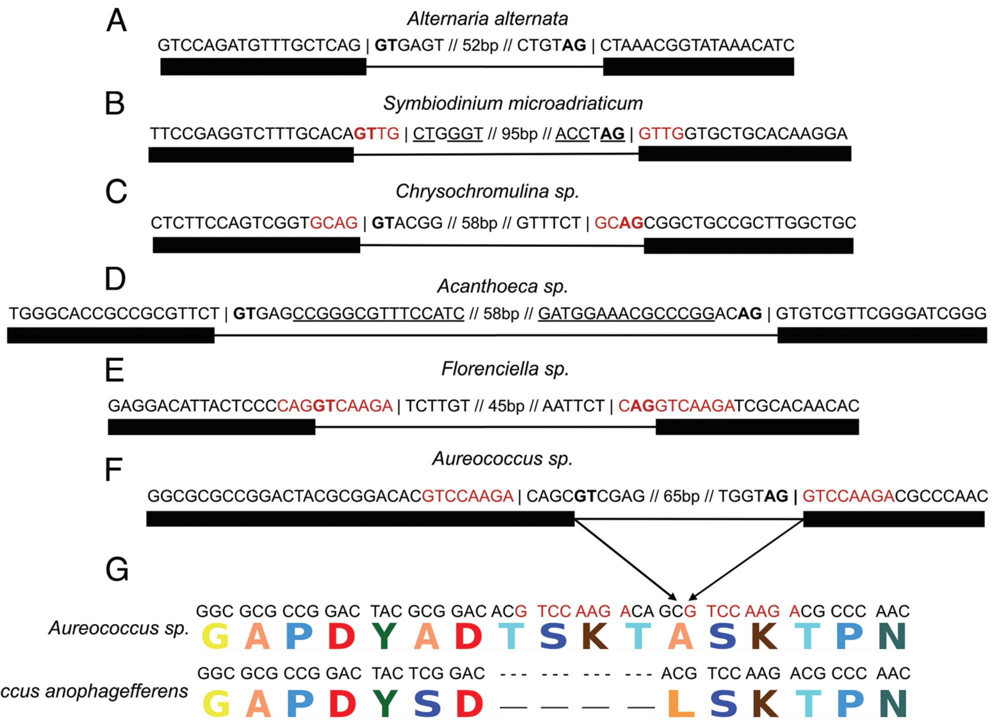
A
Alternaria alternata
GTCCAGATGTTTGCTCAG | GTGAGT // 52bp // CTGTAG | CTAAACGGTATAAACATC
B
Symbiodinium microadriaticum
TTCCGAGGTCTTTGCACAGTTG | CTGGGT // 95bp // ACCTAG | GTTGGTGCTGCACAAGGA
C
Chrysochromulina sp.
CTCTTCCAGTCGGTGCAG | GTACGG // 58bp // GTTTCT | GCAGCGGCTGCCGCTTGGCTGC
D
Acanthoeca sp.
TGGGCACCGCCGCGTTCT | GTGAGCCGGGCGTTTCCATC // 58bp // GATGGAAACGCCCGGACAG | GTGTCGTTCGGGATCGGG
E
Florenciella sp.
GAGGACATTACTCCCCAGGTCAAGA | TCTTGT // 45bp // AATTCT | CAGGTCAAGATCGCACAACAC
F
Aureococcus sp.
GGCGCGCCGGACTACGCGGACACGTCCAAGA | CAGCGTCGAG // 65bp // TGGTAG | GTCCAAGACGCCCAAC
G
Aureococcus sp.
GGC GCG CCG GAC TAC GCG GAC ACG TCC AAG ACA GCG TCC AAG ACG CCC AAC
G A P D Y A D T S K T A S K T P N
Aureococcus anophagefferens
GGC GCG CCG GAC TAC TCG GAC --- --- --- --- ACG TCC AAG ACG CCC AAC
G A P D Y S D — — — — L S K T P N

Fig. 2. Examples of diverse Introner intron creation mechanisms. Splice sites are shown in bold, Introner boundaries are denoted by a vertical bar, and introns and exons are represented by lines and boxes, respectively. (*A*) Introners in *Alternaria alternata* do not exhibit specific sequence features associated with known DNA transposition mechanisms and appear to replicate via direct insertion. (*B*) Introners in *Symbiodinium microadriaticum* show clear evidence of 4 bp TSDs (shown in green) and TIRs (underlined), consistent with many known DNA transposons, carry their 3′ splice site, and co-opt their 5′ splice site from their TSDs upon insertion. (*C*) Introners in *Chrysochromulina sp.* show evidence of 4 bp TSDs but no evidence of TIRs, carry their 5′ splice site, and co-opt their 3′ splice site. (*D*) Introners in *Acanthoeca sp.* show clear evidence of TIRs but no TSDs. (*E*) Introners in *Florenciella sp.* do not carry either splice site and instead co-opt both from their insertion site. (*F*) Introners in *Aureococcus sp.* carry both splice sites but add an extra 12 bp into the transcript upon insertion (4 bp from the Introner + 8 bp from the TSD), (*G*) resulting in the addition of four amino acids to the respective protein when compared to an ortholog from a different isolate which lacks an Introner at that position.

The Majority of Introners May Propagate via DNA-Based Mechanisms. Previous studies proposed different mechanisms for Introner mobilization. Some algal Introners appear to be miniature inverted-repeat transposable elements based on observation of TSDs, TIRs, and biased insertion into nucleosome linker regions (12). In contrast, fungal Introners lack TSDs and TIRs and are highly biased toward gene regions (insertion into nucleosome linkers was not studied) and have been interpreted as novel RNA-based elements that propagate through reverse-splicing of RNA copies of spliced Introners (14, 16, 17).

Although we observe exceptional molecular diversity (above), most Introner families exhibit at least one signature consistent with DNA transposition and ascomycetes are an outlier. Among 130 families for which both TSD and TIR presence/absence could confidently be determined, 59.2% have either TSDs, TIRs, or both (78.5% when we exclude ascomycetes fungi, see below). Presence of separate DNA- and RNA-based families predicts that putative DNA-based signatures are positively associated with each other across families and are negatively associated with the putative RNA-based signature of bias toward genes. However, TSDs are present in equal fractions of TIR-containing and non-TIR-containing non-ascomycetes families (51.7% (30/58) and 48.7% (19/40), respectively; $P = 0.99$ two-sided Fisher's exact test), and we did not find an association between TSD or TIR presence and nucleosome linker bias (*SI Appendix*, Table 7). Furthermore, there is little association between TSD or TIR presence and bias toward insertions in genes (*SI Appendix*, Table 7, see *Methods*). We also found no difference when comparing families that differed in TSD/TIR presence when accounting for species, suggesting that correlations were not obscured by unaccounted for interspecific differences (*SI Appendix*, Table 7). Ascomycete Introner families are an outlier, with no TIRs or TSDs. Together, our results suggest that most Introners propagate via DNA transposition.

Introners Show Insertional Preferences at Various Genomic Scales. We next investigated the signatures of Introner insertion by studying Introner insertion positions at the level of genome region, nucleotide content, and chromatin structure. Introners are enriched in genes in 161/175 Introner-containing genomes (Fig. 1 and *SI Appendix*, Table 5). This pattern echoes some DNA elements, for example *piggyBac* elements in *Drosophila,* which also preferentially insert into coding regions (24). Overrepresentation of Introners within genes could also reflect an insertional bias toward GC-rich regions, since genes are typically more GC-rich than intergenic regions. GC-bias has also been previously reported for DNA TEs (25, 26). We find that Introner families enriched in genes also tend to show biased insertion into GC-rich motifs ($P < 0.0001$ binomial test, *SI Appendix*, Figs. S3–S5, *SI Appendix*, Table 5). One possibility is that TEs that exhibited a preexisting bias for insertion in GC-rich genic regions experience greater selection for efficient splicing to reduce gene disruption thereby creating new Introners. Finally, some putatively DNA-based Introners in *M. pusilla* and *Aureococcus anophagefferens* preferentially insert in nucleosomes linker regions (i.e., between nucleosomes ref. 12), similarly to many DNA transposons (27). Computationally predicted nucleosome occupancy profiles showed a bias toward linker region insertion for 78.8% (104/132) of Introner-containing genomes for which prediction was possible (Fig. 1 and *SI Appendix*, Figs. S6 and S7, *SI Appendix*, Table 5 and 6).

Negative Selection Shapes the Distribution of Introner Insertions. Introners exhibit a range of molecular phenotypes indicative of negative selection on the majority of new insertions. To evaluate splicing efficiency, we estimated the percent-spliced-in (PSI) of Introners and other introns by comparing the relative read depths within adjacent exons and across each intron. Here, if an intron is very efficiently spliced, we should find few or no reads mapping across the intron. We find that observed Introners are generally more efficiently spliced than are other introns ($P = 5.9 \times 10^{-3}$, binomial test, Fig. 3 *A* and *B* and *SI Appendix*, Fig. S8). Because observed Introners may not reflect splicing of all new insertions, it is likely that more deleterious insertions could be removed by selection and thus be absent from sequenced genomes. This pattern suggests that high-frequency Introners may have limited mis-splicing-related fitness costs due to negative selection purging Introners that are frequently mis-spliced. Similarly, we find that Introner insertions are biased toward lowly expressed genes relative to other introns ($P = 2.4 \times 10^{-2}$ binomial test, Fig. 3*C* and *SI Appendix*, Fig. S9), as could be expected if some Introner insertions impose transcription- or splicing-associated costs. This bias is unlikely to result from an insertion site preference given preferential Introner insertion into GC-rich genes, which are typically more highly expressed (28). These data therefore suggest that negative selection impacts the range of Introner insertions we observe.

A New Model for the Evolutionary Forces Governing Intron Gain. Our results suggest that a range of previously proposed models for the major forces governing intron gain requires reconsideration. We find no clear support for previous influential proposals that organismal complexity or small population size promotes intron gain (29). Most strikingly, among animals and land plants, two organismally complex groups with typically small effective population sizes, we find Introners in only two taxonomic groups (both animals), despite accounting for one-quarter (835/3,325) of studied genomes ($P < 0.00001$, two-sided Fisher's exact test). This dearth of Introners is all the more striking given that these lineages are intron-rich, are generally more slowly evolving at the sequence level which facilitates Introner discovery, and widely use introns in gene regulation (1). The large number of observed introns in these lineages is more likely the result of low rates of ancestral Intron loss rather than more recent intron gains (30). Furthermore, we find no evidence for a pattern of Introner gain in species whose biology predicts small population size, such as parasites or vertebrates.

We propose that the distribution of Introners reflects the propensity of lineages to acquire new genetic elements via horizontal gene transfer (HGT) (*SI Appendix*, Fig. S10). All Introner-containing lineages except Ascomycetes and one species of blastocyst are aquatic organisms with aquatic organisms mostly represented by a remarkably diverse array of unicellular organisms. We propose that this pattern reflects greater rates of HGT for these species. Indeed, aquatic environments generally favor HGT (31–34), and aquatic unicellular in particular have been shown to exhibit high rates of HGT possibly because they often live in dense microbial communities and require interactions with other species for their ecology (35). A variety of studies have detailed large-scale lateral gene transfer in aquatic protists, including several that have Introners (35–38). The only two non-aquatic Introner-containing lineages, ascomycetes and blastocysts, have also been reported to have large amounts of HGT (39, 40). We do not find evidence of HGT of Introners identified in this study. However, DNA transposons more generally are well adapted for HGT and have frequently made jumps between highly diverse lineages (41).

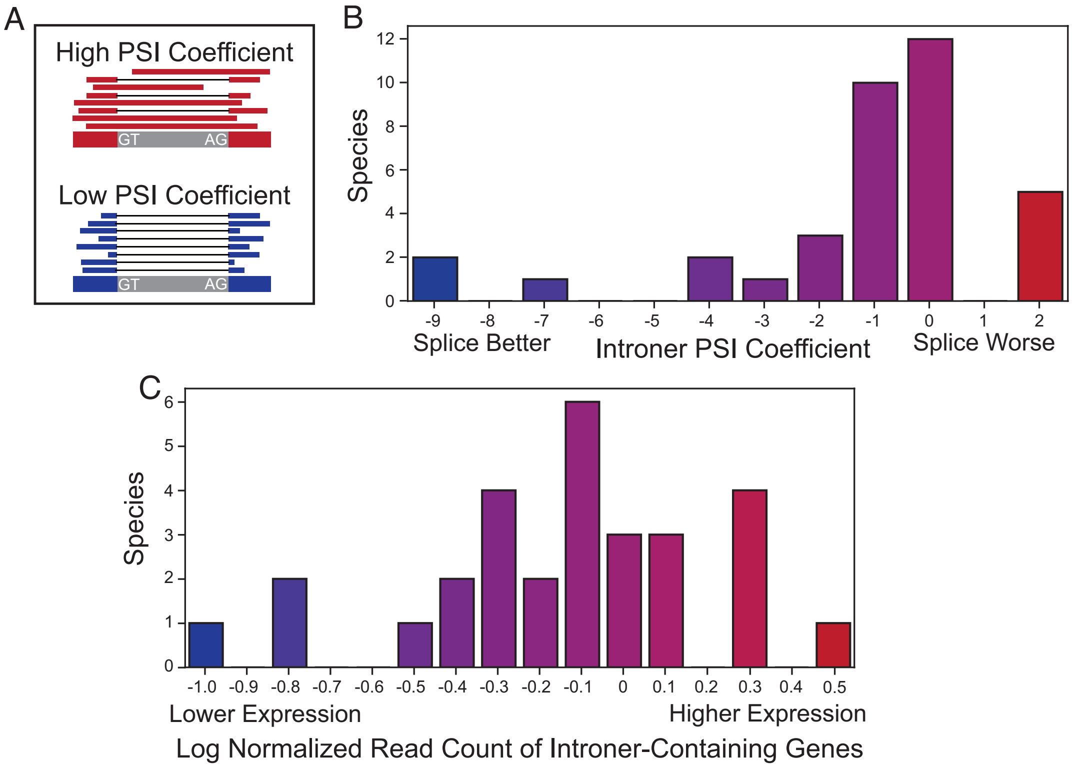

Fig. 3. Relative to other introns, Introners are more efficiently spliced but less frequently found in highly expressed genes. (*A*) Explanation of PSI. Species with greater PSI coefficients (above 0, red) have Introners spliced in more frequently than other introns in the same genome. (*B*) Introners are more efficiently spliced than other introns in most species, as indicated by PSI coefficients less or equal to 0 for 31/36 species. (*C*) Introners are overrepresented in lowly expressed genes for most species. Relative log-normalized gene expression values for Introner-containing genes relative to other Intron-containing genes are shown.

The introduction of a new DNA transposon unfamiliar to a host might also intensify selection for the independent evolution of Introners. Newly acquired TEs often evade host-mediated TE silencing mechanisms and thereby have the freedom to mobilize at high frequencies, representing a major cost to host fitness (41). The ability of a TE to be spliced could alleviate some of these costs. Thus, HGT could not only explain the highly punctated pattern of Introner presence across distantly related taxa, but also favors the independent evolution of Introners in new lineages.

Conclusion

The proximate and ultimate origins of introns remain a fundamental question in biology. Here, we demonstrate that Introners generate new introns on genomic scales in a remarkable diversity of eukaryotic lineages. Despite many similarities, the extensive molecular diversity that underlies Introner transposition reveals that a vast range of transposon families has independently evolved into Introners. In light of these findings, frequent horizontal transfer of TEs and the extreme aquatic biased distribution of species harboring Introners, we propose that a crucial factor governing lineages' tendency to gain introns over time is exposure to transfer of TEs from diverse unrelated eukaryotic organisms.

Methods

Accessing Genomic Data. We performed our systematic search for Introners on all annotated genomes in Genbank (last accessed 9:24 AM April 10, 2020). We used FTP links available through a CSV file downloadable from NCBI (*SI Appendix*, Table 1) to systematically access and download genomic data for our analyses. We filtered out genomes that lacked annotation files or for which annotations were dubious based on low gene number (*SI Appendix*, Table 2). Fasta files and genome annotations were downloaded for the Tara Oceans Eukaryotic Genomes project, from https://www.genoscope.cns.fr/tara/. Genomes without GFF annotation files were filtered out, yielding a total of 520 genome assemblies.

Identifying Highly Similar Introns. To find candidate Introners, for each genome, we first extracted all annotated intron–exon structures, that is, genomic sequences corresponding to annotated protein-coding sequences spanning from translation start to stop codons, and with exon and intron sequences indicated by upper and lower case. We then extracted each intron along with up to 20 nucleotides of flanking sequence (requiring a minimum of ten flanking exonic nucleotides). We then searched for introns with sequence similarity to each other. Candidate similar pairs were identified by an all-against-all blast (42) of introns-plus-flanking sequences within each species, with a minimum e- value of 10^{-5}. Previous results and our preliminary findings reveal that there exist several reasons that two introns can have extensive sequence similarity other than creation by the same Introner family; therefore, we performed several filtering steps to eliminate false positives.

We filtered for two alternative reasons for intron–intron sequence similarity. First, many introns with extensive sequence similarity to each other owe this similarity to secondary insertion of transposable elements or to microsatellites within the intron interior. Conversely, introns from paralogous genes or gene regions can be similar due to duplication of a longer region. These two cases share a frequent signature, namely that the region of intron–intron sequence does not correspond to (roughly) the whole of both introns, but instead is either a subportion of the intron (in the first case) or extends beyond the intron (in the second case). Therefore, we required that the region sequence similarity between the two introns extends to near the exon–intron boundary. After iterative manual scrutiny, we chose to require that sequence similarity begins within a 15 base region spanning five exonic bases or ten intronic bases, and that this be the case for both introns for both 5′ and 3′ intronic boundaries.

After initially requiring end-to-end blast hits, we learned that secondary indels including secondary TE insertion led to many false negatives. Consequently, we applied a different strategy where we judged similarity between each pair of introns based on pairwise similarity of the two ends, either similarity between corresponding ends (5′ with 5′, 3′ with 3′, as expected from same-orientation insertion) or opposing ends (5′ with 3′, 3′ with 5′, as in opposite-orientation insertion). Up to 100 intronic nucleotides were assessed for each end. We required that the similarities were in the expected orientation (i.e., the interiors of the two introns lining up together).

Recent paralogous gene duplications can result in sequence similarity between intron sequences. While the requirement that nucleotide-level sequence similarity begins an end near the intron–exon boundary removes most such cases, we also observed cases where rapid exonic evolution (or simple chance substitution) led to paralogous sequences being retained (the clearest case involved the introns of the huge gene families fast-evolving *var* genes of *Plasmodium* species). To filter remaining false positives by introns in paralogous genes, we first translated all Introner-containing genes from DNA to protein sequence. Then we used diamond (version 0.9.24) (43), with default options except minimum e- value of 10^{-20}, to identify and remove from the list cases of sequence similarity between encoded proteins for intron pairs with similar sequences.

Within each genome with remaining similar intron pairs, we then used pairwise similarities and a greedy algorithm to group introns with at least one remaining pairwise similarity into Introner families. Families with at least four introns were retained. We acknowledge that since we used a homology-based approach to identify Introners, we are more likely to identify Introners in lineages with slower evolutionary rates. Nonetheless, this possible source of bias apparently had very little effect on our results since we did not find any evidence for Introners in lineages which are generally slowly evolving in sequence such as land plants and mammals and instead primarily found Introners in lineages which evolve relatively rapidly such as green algae.

Filtering Assembly Errors. We also filtered out putative Introner families identified as a result of genome assembly problems. For example, sequencing adapters included in genome assemblies might sometimes be annotated as introns, in which case genome assemblies including many sequencing adaptors could have multiple similar annotated intronic sequences. We used blast searches to examine putative Introner for the presence of Illumina adaptors and removed families which contained them. We also performed an Internet search on each putative Introner family consensus sequence and subsections of each consensus sequence to ensure that Introner families did not embody or contain any known sequences associated with genome assembly or sequencing methods.

Filtering Introners with Low Complexity. We filtered putative Introners families with low sequence complexity since sequence similarity between these potential Introners could have resulted from alternative mechanisms than transposition, e.g., microsatellite expansion. To do this, we manually examined putative Introner family consensus sequences and looked for an abundance of short repetitive sequence motifs.

Finalized Introner Sequences. After filtering, we possessed a set of finalized Introner sequences sorted in fasta files by species and family within species. Fasta files for each Introner family in each species can be found at https://github.com/lgozasht/Introner-elements.

A Representative Set of Genomes for Downstream Analyses Based On Introner Content. We next scrutinized patterns of Introner family presence/absence across all Introner-containing genomes. In particular, the presence of clusters of closely related organisms represented within the TARA Oceans genomes led to cases of multiple genomes with very similar Introner complements—both at the level of families and of specific insertions. At the same time, as mentioned in the main text, closely related genomes sometimes exhibit overlapping but non-identical sets of Introner families. Genomes containing identical sets of Introner families were grouped, leading to 16 groups, mostly including two genomes (13 groups), but ranging up to nine genomes.

Proportion of Introners Inside of Genes. Since Introners in intergenic regions cannot be annotated as introns, we developed a systematic method to re-cover them conditional on a known Introner family detected as described above. We employed multiple alignment using fast Fourier transform (MAFFT) (44) to conduct multiple sequence alignments for each Introner family in each species. Next, we generated a consensus sequence for each Introner family using a positional nucleotide frequency matrix. We required that greater than 50% of Introners possess the same nucleotide at a particular position for that base to be included in our consensus sequence. We BLASTed each consensus to its corresponding reference genome and filtered duplicate and self-hits.

We used a permutation test to interrogate possible enrichment for Introners in genes. If a genome is more gene dense, a transposon is more likely to land in a gene. To correct for this, we generated 1,000 permutations for the probability that a particular Introner will insert into a gene by chance by randomizing the Introner positions across the genome such that n = the number of total insertions and p = gene density. We compared these with our actual values to test for insertional enrichment in genic regions.

GC Content Analysis. Genes are generally GC-rich relative to intergenic regions (45), and transposable elements have been shown to display preference for GC-rich regions (46, 47). To test whether Introners that are enriched in genes also demonstrate a bias for GC-rich regions, we employed a permutation test. We calculated the GC content for the concatenated ten base pairs (bp) upstream and downstream of each insertion (20 bp total). We then generated 10,000 permutations for the GC content of randomly resampled 20 bp regions from the same gene in which each respective Introner was found. By using a relatively small window size of 20 bp, we hoped to more accurately capture specific insertion site biases by limiting the noise introduced by surrounding sequences, although we also used a window size of 100 bp and obtained similar results. We compared our observed GC proportions to the randomly sampled distribution and found that Introners in many species are also enriched within GC-rich regions. Across all species, we found a significant correlation between insertional enrichment in genes and insertional enrichment in GC-rich regions, suggesting that Introners may favor GC-rich regions rather than simply an insertion preference for genic regions per se ($P < 0.0001$; binomial test). Here, we constructed this comparison as a one-sided test where we asked if the proportion of species whose insertion preferences exceeded background genic GC content differed from random expectations. Note also that conditioning on the specific genes into which Introners insert is very conservative because any strong GC content skew within a gene would be reflected in the null distribution.

Nucleosome Occupancy Prediction and Analysis. Since the vast majority of genomes in our survey lack available epigenetic data and most cannot be reliably cultured, we applied an in silico predictive approach to interrogate whether Introners insert into nucleosome linker DNA in other lineages. We used the program, *NuPoP* (48), to predict the nucleosome occupancy for the 10 kb spanning and surrounding the insertion sites of all Introners in each species for which we possessed at least four Introners with contiguous sequence assembled 5 kb upstream and 5 kb downstream of insertion sites. We required at least 5 kb around each Introner to ensure that we maximize the accuracy of predictions. The reason is that this approach is based on a hidden Markov model and therefore requires moderate sequence lengths to produce reliable results. We ran *NuPoP* with flags *species=0* and *model=4* first with Introners and then again with Introners computationally removed from the gene sequence as a control (*SI Appendix*, Fig. S6). We observe a pattern across most genomes of decreased nucleosome occupancy for the 100 bp surrounding the 5′ splice sites of Introners relative to background regions ($P = 8.26e{-}07$; binomial test) and are able to reproduce the pattern previously reported for Introners in Micromonas and Aerococcus using nucleosome profile data (*SI Appendix*, Figs. S6 and S7). We observe a decreased nucleosome occupancy within Introner sequences relative to background regions even more often ($P = 4.00e{-}10$; binomial test), suggesting that Introners insert into and inhabit nucleosome linker regions. When we perform the same comparison with Introners removed to replicate surrounding regions, we observe the opposite pattern, in which the nucleosome occupancy is higher ($P = 9.13e{-}12$; binomial test), suggesting that nucleosomes often flank Introner sequences (*SI Appendix*, Fig. S7). We used a Gaussian generalized linear model (GLM) through R to look for an association between the number of Introners and predicted nucleosome occupancy within Introners relative to background nucleosome occupancy of each Introner family in each species with formula *delta_nuc_occup ~ number_of_Introners*. We did the same association for Introner length and species with formulae: *delta_nuc_occup ~ Introner_length* and *delta_nuc_occup ~ species*. We find that the number of Introners in each family and Introner length are both good predictors of nucleosome occupancy within Introners relative to background ($P < 0.0001$ and $P = 0.04$), with smaller families with shorter sequences having lower *delta_nuc_occup*. We postulate that this association may be due to reduced accuracy of predictions on small sample sizes. This may explain why in smaller, shorter Introner families we sometimes do not observe as clear of a pattern of low nucleosome occupancy in Introners. We note, however, that species is an even better predictor than the aforementioned variables ($P < 0.0001$) and that our ability to predict nucleosome profiles accurately with a sequence motif-based hidden Markov model (HMM) could also be limited in highly divergent species relative to those used for training the HMM initially or in low-quality genomes. Indeed, when we filter for species in which we observed *delta_nuc_occup* < 0 and use a GLM with the formula: *delta_nuc_occup ~ GC_content*, we find that GC content explains better than species ($P = 0.0015$).

Identifying TSDs and TIRs. We searched for evidence of TSDs in and around each Introner family in each species for which there were at least 15 genic Introners in the family. Most clearly, positions that are part of a TSD manifest as similarity between 5′ and 3′ ends within individual Introner from a single insertion site, but not globally between introns (i.e., corresponding nucleotides 5′ and 3′ ends match for intron A, but do not necessarily match between introns A and B). However, TSDs often include core GY/AG splice-site nucleotides, in which case nucleotides will match across introns as well. Thus to assess TSD presence/absence, we calculated both absolute match between corresponding 5′ and 3′ nucleotides (i.e., nucleotides the same distance from the splice site, for instance the first nucleotide of the intron and the first nucleotide of the downstream exon), as well as relative match, calculated as the fraction of matches within individual introns divided by the expected value calculated from 1,000 random 5′/3′ pairs. For each Introner family within each species, these values were assessed to look for TSD patterns including at least one nucleotide position with locus-specific match (e.g., NAGgy…nag, where N/n show significant match within but not between insertion sites). For patterns that included only across-intron matches (e.g., families where all or nearly all introns are preceded by an AG (AGgy…xyag), it is not possible to distinguish between the AG representing a TSD (Introner sequence = gy..xyag, TSD = AG), or the AG representing an insertion site without a TSD (Introner sequence = gy..xyag or AGgy…xy, no TSD). In some instances, one of the two duplicate motifs was part of an extended TIR (see below). Otherwise, TSD presence/absence was called as ambiguous.

TIRs were searched for manually by searching the consensus sequence within a region extending from −20 to 20 nucleotides of the intron. This generally yielded either a clear extended TIR (≥6 nucleotides) or no evidence of a TIR. The few cases with partial or short TIRs were called as ambiguous. These calls are available at https://github.com/lgozasht/Introner-elements

Examples used in Fig. 2 *A–F*. For Fig. 2*A*, we used an Introner in family 1 of *Alternaria alternata* (Genbank acc. GCA_001572055.1) on scaffold LPVP01000001.1, position 1639687–1639750. For Fig. 2*B*, we used an Introner in family 1 of *Symbiodinium microadriaticum* (Genbank acc. GCA_001939145.1) on scaffold LSRX01000224.1 position 86725–86839. For Fig. 2*C*, we used an Introner in family 5 of *Chrysochromulina sp.* (TARA_PON_109_MAG_00232) on scaffold 000000000114 position 1755–1824. For Fig. 2*D*, we used an Introner in family 4 of *Acanthoeca sp.* (TARA_AON_82_MAG_00310) on scaffold 000000000270.1.1.3 at position 451–559. For Fig. 2*E*, we used an Introner in family 2 of *Florenciella sp.* (TARA_MED_95_MAG_00409) on scaffold 000000000986.1.2.2 at position 852–918. For Fig. 2*F*, we used an Introner in family 4 of *Aureococcus sp.* (TARA_AOS_82_MAG_00129) on scaffold 000000001056 at position 3289–3365.

Comparing Orthologs Between Aureococcus Isolates. We used BLAST to identify homology between Introner-containing genes in *Aureococcus sp.* (isolate TARA_AOS_82_MAG_00129) and *Aureococcus anophagefferens* (Genbank Acc. GCA_000186865.1). We used MAFFT (44) to perform a multiple sequence alignment (MSA) between each Introner-containing gene in *Aureococcus sp.* and its match in *Aureococcus anophagefferens* with the lowest e-value given the e- value < 0.01. We also performed an msa between translated proteins corresponding to these genes. The example shown in Fig. 2 stems from an alignment between an Introner-containing gene in *Aureococcus sp.* and *Aureococcus anophagefferens* Phosphoenolpyruvate carboxylase kinase 1 (NCBI Acc. XM_009038642.1) at both the nucleotide and protein level. In this example, *Aureococcus sp.* exhibits an Introner insertion at position 545 in this gene, which resulted in the addition of four amino acids relative to *Aureococcus anophagefferens*.

Identifying Potential Mobilizing Elements. We searched for transposase-encoding autonomous elements that may mobilize Introners with similar terminal sequences. For each Introner family in each species, we constructed position weight matrices of length 22 bp at the 5′ and 3′ ends of the elements. We then searched each respective species' genome for matches to the 5′ probability weight matrix (PWM) using PoSSuMsearch (49) and searched the downstream 10,000 bp of each match using the 3′ PWM. We also searched for matches among predicted repetitive elements found using RepeatModeler (50). Open reading frames were found between the 5′ and 3′ pairs of PWM matches, and their translated amino acid sequences were used to search a database of transposases (a subset of UniProt) using BLASTP (*SI Appendix*, Table 8).

Assessing Homology Between Introners in Different Species. We used BLAST to search for homology between consensus sequences of Introners in different species. We performed an all vs. all BLAST of Introner consensi and found no evidence of homology except between relatively recently diverged species (within the same genus). However, we did observe cases for TARA metagenomes (for which only the genus is reported) in which isolates within the same genus possess different Introner families. We treated those isolates as separate species throughout our study.

Checking for Associations Between TSD and TIR Presence and Introner Architecture and Distribution. We tested for associations between TSD and TIR presence and other Introner statistics both on the subset of genomes that possessed Introner families with and without TSDs/TIRs and across all species. To do this, we used generalized linear models through R (*SI Appendix*, Table 7). To test for an association between TSD and TIR presence and insertional preference in genes, we fit a GLM with the following format:

$$\textit{(Introners\_in\_genes, Introners\_outside\_genes)} \sim \textit{TSDs} \text{ and } \textit{cbind(Introners\_in\_genes, Introners\_outside\_genes)} \sim \textit{TIRs}$$

using a binomial family link function. As a control, performed the same analysis with species instead of TSD or TIR presence using a GLM of the format:

$$\textit{(Introners\_in\_genes, Introners\_outside\_genes)} \sim \textit{species}.$$

We found that the term species better explains our data than TSD or TIR presence using Akaike Information Criterion (AIC).

To test for an association between TSDs and TIR and canonical splice-site usage, we fit a GLM of the form:

$$\textit{canonical\_splice\_site\_usage} \sim \textit{TSDs} \text{ and } \textit{canonical\_splice\_site\_usage} \sim \textit{TIRs}$$

under the Gaussian family link function. Again we performed the same association for species and found that species better explains our data than TSD or TIR presence. To test for an association between TSD and TIR presence and number of Introners, we used a GLM of the form:

$$\textit{number\_of\_Introners} \sim \textit{TSDs}$$

$$\textit{number\_of\_Introners} \sim \textit{TIRs}$$

under a Gaussian link function. We find that TSDs and TIRs explain our data poorly in this case. To test for an association between TSD and TIR presence and delta nucleosome occupancy, we a GLM of the format:

$$\textit{background\_nuc\_occup-nuc\_occup\_of\_Introner} \sim \textit{TSDs}$$

$$\textit{background\_nuc\_occup-nuc\_occup\_of\_Introner} \sim \textit{TIRs}$$

again using the Gaussian family link function. We performed the same association with species and again found that species better explained our data.

RNA-seq Analysis. For each species with identified Introners (genus level TARA metagenomes excluded), we searched the Sequence Read Archive (SRA) database for RNA-seq data, prioritizing sequencing runs conducted on the same individual from which the reference genome was assembled (*SI Appendix*, Table 9). We aligned the RNA reads to the reference genome using *STAR* (51), calculated the depth at each site using *samtools* (52), and identified splice junctions using leafcutter (53). We then used custom Python scripts to identify, for each intron, the number of splicing events that used the annotated splice junctions, as well as the number of splicing events that used non-canonical junctions within 50 nucleotides on either side of the annotated junction. We then used the R package *lme4* (54) to construct generalized linear model of the form:

$$\textit{(proper\_splices,missplices)} \sim \textit{Introner} + \textit{depth} + \textit{length}$$

$$\textit{(proper\_splices,missplices)} \sim \textit{depth} + \textit{length}$$

to correct for the depth and length of each intron. To ensure that the i.e., variable (whether or not the intron is an Introner) was significantly correlated with splicing behavior, we calculated the likelihood ratio of the two models using the AIC (55). If the model containing the i.e., variable was a better fit, and if coefficient for the i.e., variable was positive, Introners in this species exhibit more canonical splicing than non-Introner intron.

We used a similar approach for percent spliced in (PSI; Fig. 3), using GLMs of the form:

$$\textit{(spliced\_in,proper\_splices} + \textit{missplices)} \sim \textit{Introner} + \textit{depth} + \textit{length}$$

$$\textit{(spliced\_in,proper\_splices+missplices)} \sim \textit{depth} + \textit{length},$$

ensuring that whether an intron is an Introner is significantly correlated with PSI using AIC likelihood ratios, and considering cases with negative coefficients for the Introner variable as cases where Introners are more likely to be spliced in than non-Introner introns (Fig. 3).

To identify whether Introner-containing genes were more lowly expressed than other genes in each species, we used two methods. First, we used a permutation test. We calculated the average read count for Introner-containing genes and then selected 10,000 sets of randomly sampled genes, where the probability that a given gene was sampled was proportional to its length. The number of random samples in which the average read count of the random sample was greater than the average read count of the Introner-containing genes acted as our *P*-value (*SI Appendix*, Table 9). Additionally, we also performed independent Mann–Whitney U tests comparing reads per kilobase of exon per million distributions for Introner-containing genes and other intron-containing genes. For this analysis, we also removed transcripts with less than 10 mapping reads.

Correlating Aquatic Lifestyle and Introner Presence. For phylogenetic tests, we downloaded the global tree from open tree of life (56) and pruned the tree to retain only species that we considered in this analysis. In the case of TARA metagenomes, for which we only possessed the genera, we randomly selected one species within each genus to represent all isolates. Our tree can be found at https://github.com/lgozasht/Introner-elements/blob/main/pruned_tree2.nwk.gz. The open tree of life maintains arbitrary branch lengths (branch lengths of 1). We recognize that arbitrary branch lengths can imply that more total evolution has taken place between the root and the tips of the tree for those species with more ancestors (57). Thus, to test for whether or not phylogenetic signal could explain the distribution of Introners across species, we used abouheif's C_{mean} test through the R package *adephylo* (58, 59). This method only considers topological relationships among species and is thus robust to branch length ambiguities (60). We obtained a P value < 0.001 across 1,000 permutations, suggesting that phylogenetic signal in itself poorly explains the observed distribution of Introners across the eukaryotic tree. To evaluate a correlation between aquatic lifestyle and the presence of Introners, we used Pagel's test through the R package *phytools* (57, 61) using the aforementioned global phylogeny as an input (*SI Appendix*, Fig. S2). Pagel's test estimates rates of evolution for two given traits, constructs models in which the two traits evolve independently, co-dependently, or with one trait dependent upon the other, and model fits are compared using AIC (*SI Appendix*, Fig. S1). To prepare the input data for these tests, we manually annotated each species considered as aquatic or not using the following criterion: To be considered an aquatic species, a species must spend most of its life surrounded by water (*SI Appendix*, Table 4). A species also cannot be an obligate parasite of a terrestrial organism even though in such a case the species may live primarily within an aqueous solution.

Data, Materials, and Software Availability. Previously published data were used for this work (NCBI).

ACKNOWLEDGMENTS. We thank J. Collemare, A. van der Burght, and J. Logsdon for helpful discussions, as well as several community members who participated in a conversation about the nomenclature for Introners. We also thank Cedric Feschotte for helpful feedback on this manuscript as well as Scott Edwards and Dean Adams for helpful discussions regarding phylogenetic comparative methods. Funding: This work was supported in part by R35GM128932 and by an Alfred P. Sloan Foundation Fellowship to R.C.-D. and by NSF award 1616878 to S.W.R. B.T. was supported by T32HG008345.

Author affiliations: [a]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064; [b]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064; [c]Department of Biology, San Francisco State University, San Francisco, CA 94117; and [d]Department of Molecular, Cell, and Developmental Biology, University of California Santa Cruz, Santa Cruz, CA 95064

1. M. Irimia, S. W. Roy, Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb. Perspect. Biol.* **6**, a016071 (2014).
2. B. R. Graveley, Alternative splicing: Increasing diversity in the proteomic world. *Trends Genet.* **17**, 100–107 (2001).
3. O. Jaillon *et al.*, Translational control of intron splicing in eukaryotes. *Nature* **451**, 359–362 (2008).
4. D. G. Scofield, X. Hong, M. Lynch, Position of the final intron in full-length transcripts: Determined by NMD? *Mol. Biol. Evol.* **24**, 896–899 (2007).
5. D.-K. Niu, Protecting exons from deleterious R-loops: A potential advantage of having introns. *Biol. Direct.* **2**, 11 (2007).
6. M. Csurös, I. B. Rogozin, E. V. Koonin, Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum-likelihood approach. *Mol. Biol. Evol.* **25**, 903–911 (2008).
7. L. Carmel, Y. I. Wolf, I. B. Rogozin, E. V. Koonin, Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res.* **17**, 1034–1044 (2007).
8. W. Li, A. E. Tucker, W. Sung, W. K. Thomas, M. Lynch, Extensive, recent intron gains in Daphnia populations. *Science* **326**, 1260–1262 (2009).
9. B. A. Curtis, J. M. Archibald, A spliceosomal intron of mitochondrial DNA origin. *Curr. Biol.* **20**, R919–R920 (2010).
10. J. H. Rogers, How were introns inserted into nuclear genes? *Trends Genet.* **5**, 213–216 (1989).
11. M. Irimia *et al.*, Origin of introns by "intronization" of exonic sequences. *Trends Genet.* **24**, 378–381 (2008).
12. J. T. Huff, D. Zilberman, S. W. Roy, Mechanism for DNA transposons to generate introns on genomic scales. *Nature* **538**, 533–536 (2016).
13. A. Z. Worden *et al.*, Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. *Science* **324**, 268–272 (2009).
14. 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).
15. 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).
16. E. Fekete *et al.*, Internally symmetrical stwintrons and related canonical introns in hypoxylaceae species. *J. Fungi (Basel)* **7**, 710 (2021).
17. M. P. Simmons *et al.*, Intron invasions trace algal speciation and reveal nearly identical arctic and antarctic micromonas populations. *Mol. Biol. Evol.* **32**, 2219–2235 (2015).
18. 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).
19. T. O. Delmont, Environmental genomics points to non-diazotrophic Trichodesmium species abundant and widespread in the open ocean. *bioRxiv* (2021). 10.1101/2021.03.24.436785. Accessed 16 November 2021.
20. R. Symonová, A. Suh, Nucleotide composition of transposable elements likely contributes to AT/GC compositional homogeneity of teleost fish genomes. *Mob. DNA* **10**, 49 (2019).
21. M. Osanai-Futahashi, Y. Suetsugu, K. Mita, H. Fujiwara, Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, Bombyx mori. *Insect Biochem. Mol. Biol.* **38**, 1046–1057 (2008).
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. 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).
24. S. T. Thibault *et al.*, A complementary transposon tool kit for Drosophila melanogaster using P and piggyBac. *Nat. Genet.* **36**, 283–287 (2004).
25. R. S. Linheiro, C. M. Bergman, Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. *PLoS One* **7**, e30008 (2012).
26. C.-L. Tsai, M. Chatterji, D. G. Schatz, DNA mismatches and GC-rich motifs target transposition by the RAG1/RAG2 transposase. *Nucleic Acids Res.* **31**, 6180–6190 (2003).
27. S. Gangadharan, L. Mularoni, J. Fain-Thornton, S. J. Wheelan, N. L. Craig, DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 21966–21972 (2010).
28. G. Kudla, L. Lipinski, F. Caffin, A. Helwak, M. Zylicz, High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.* **4**, e180 (2006).
29. M. Lynch, Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 6118–6123 (2002).
30. 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).
31. L. D. McDaniel *et al.*, High frequency of horizontal gene transfer in the oceans. *Science* **330**, 50 (2010).
32. A. Goyal, D. Gelbwaser-Klimovsky, J. Gore, Horizontal gene transfer becomes disadvantageous in rapidly fluctuating environments. *bioRxiv* (2020), 2020.08.07.241406.
33. X. Wang, X. Liu, Close ecological relationship among species facilitated horizontal transfer of retrotransposons. *BMC Evol. Biol.* **16**, 201 (2016).
34. S. Venner *et al.*, Ecological networks to unravel the routes to horizontal transposon transfers. *PLoS Biol.* **15**, e2001536 (2017).
35. D. A. Caron *et al.*, Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nat. Rev. Microbiol.* **15**, 6–20 (2017).
36. 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).
37. R. G. Dorrell *et al.*, Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2009974118 (2021).
38. A. Monier *et al.*, Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome Res.* **19**, 1441–1449 (2009).
39. L. Eme, E. Gentekaki, B. Curtis, J. M. Archibald, A. J. Roger, Lateral gene transfer in the adaptation of the anaerobic parasite blastocystis to the gut. *Curr. Biol.* **27**, 807–820 (2017).
40. T. A. Richards, Genome evolution: Horizontal movements in the fungi. *Curr. Biol.* **21**, R166–R168 (2011).
41. 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).
42. C. Camacho *et al.*, BLAST+: Architecture and applications. *BMC Bioinf.* **10**, 421 (2009).
43. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Met.* **12**, 59–60 (2015).
44. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
45. A. E. Vinogradov, DNA helix: The importance of being GC-rich. *Nucleic Acids Res.* **31**, 1838–1844 (2003).
46. K. Hiom, M. Melek, M. Gellert, DNA transposition by the RAG1 and RAG2 proteins: A possible source of oncogenic translocations. *Cell* **94**, 463–470 (1998).
47. S. Cerbin, C. M. Wai, R. VanBuren, N. Jiang, GingerRoot: A novel dna transposon encoding integrase-related transposase in plants and animals. *Genome Biol. Evol.* **11**, 3181–3193 (2019).
48. L. Xi *et al.*, Predicting nucleosome positioning using a duration Hidden Markov model. *BMC Bioinf.* **11**, 346 (2010).
49. J.-I. Ito, K. Ikeda, K. Yamada, K. Mizuguchi, K. Tomii, PoSSuM vol 2.0: Data update and a new function for investigating ligand analogs and target proteins of small-molecule drugs. *Nucleic Acids Res.* **43**, D392–D398 (2015).
50. 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).
51. A. Dobin *et al.*, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
52. P. Danecek *et al.*, Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
53. Y. I. Li *et al.*, Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
54. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4. *arXiv [stat.CO]* (2014). https://arxiv.org/pdf/1406.5823.pdf (Accessed 11 May 2021).
55. C. M. Hurvich, J. S. Simonoff, C.-L. Tsai, Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Stat. Soc. Series B Stat. Methodol.* **60**, 271–293 (1998).
56. Open Tree of Life (2019), 10.5281/zenodo.3937742. Accessed 1 November 2021.
57. M.Pagel, Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B Biol. Sci.***255**, 37–45 (1994).
58. T. Jombart, F. Balloux, S. Dray, adephylo: New tools for investigating the phylogenetic signal in biological traits. *Bioinformatics* **26**, 1907–1909 (2010).
59. J. Thioulouse, D. Chessel, S. Champely, Multivariate analysis of spatial patterns: A unified approach to local and global structures. *Environ. Ecol. Stat.* **2**, 1–14 (1995).
60. S. Pavoine, C. Ricotta, Testing for phylogenetic signal in biological traits: The ubiquity of cross-product statistics. *Evolution* **67**, 828–840 (2013).
61. L. J. Revell, phytools: An R package for phylogenetic comparative biology (and other things). *Met. Ecol. Evol.* **3**, 217–223 (2012).