Manuscript resulting from this work in preparation for submission to *Science* (Gozashti et al. *in prep*)

De novo creation of spliceosomal introns by different transposition mechanisms across diverse eukaryotes

Landen Gozashti[1]

Advised by Russell Corbett-Detig[1], Scott W. Roy[2], and Manuel Ares Jr.[3]

[1]University of California Santa Cruz Department of Biomolecular Engineering,
UCSC Genomics Institute
[2]San Francisco State University Department of Biology
[3]University of California Santa Cruz Department of Molecular, Cell and Developmental Biology

Contents

Abstract

Introns are sequences interrupting genes that must be removed from mRNA before translation, and are a hallmark of eukaryotic genomes. They likely play important roles in genome evolution, but have poorly understood origins. Many species exhibit major intron loss events, which probably occur through RNA mediated homologous recombination of cDNA. In contrast, some species exhibit prolific intron gain. Micromonas pusilla, an aquatic picophytoplankton, probably exhibits the most notable recent case of intron gain. Intronic sequences known as introner elements (IEs) colonized the M. pusilla genome in astounding quantities, likely through a mechanism involving DNA transposition. Contrary to canonical introns, introner elements exhibit conserved sequences and lengths. Similar phenomena are known to exist in fungi. Although introner elements are known to exist in some species, no study has conducted a systematic search for them. I developed a computational pipeline for introner element detection and implemented it on all annotated assemblies on the Genbank database available through NCBI. I report putative novel IE discoveries in 38 species across 11 phyla, a major expansion upon the handful of previously known cases, which spanned only two phyla. The functional impacts of introner insertions predict evolutionary outcomes and suggest that many are deleterious while others do not display obvious functional effects. Additionally, the presence of IEs is strongly correlated with the accessibility of an organism's germline, indicating that IEs move between species via horizontal gene transfer. My results reveal that transposons may operate as a fundamental driver of intron gain in diverse lineages.

Acknowledgements

I express my deepest thanks to Manuel Ares Jr. of the University of California Santa Cruz, Department of Molecular, Cell, and Developmental Biology for being the first to give me the opportunity to conduct research in a lab as an undergraduate and for believing in my abilities. I also thank Scott W. Roy of the San Francisco State University Department of Biology for collaborating on this project and assisting in mentoring my research. I leave immeasurable gratitude for my mentor, Russell Corbett-Detig of the University of California Department of Biomolecular Engineering and the University of California Santa Cruz Genomics Institute, for inspiring my desire to become an expert in the fields of computational biology and genomics. Although I believe he sometimes overestimates my capabilities and knowledge base, his undisguised brilliance continuously inspires me to strive for greatness in my scientific career. I also thank Bryan Thornlow (PhD candidate in the Corbett-Detig Lab) for his contributions to this project and my development as an independent scientist. Lastly, I thank Paloma Medina (PhD candidate in the Corbett-Detig Lab) for her guidance and support throughout the past few years along with the rest of the Corbett-Detig Lab, Roy Lab, and Ares Lab.

1 Problem Statement

Introns play important roles in eukaryotic genome structure evolution; however, the fundamental drivers of intron gain remain elusive (1). My goal is to interrogate transposition as a primary driver of intron gain on genomic scales in diverse eukaryotic lineages.

Spliceosomal introns are sequences embedded in genes that must be removed from mRNA before translation and are a hallmark of eukaryotic genomes. They regulate gene expression levels, enable alternative splicing, and play notable roles in regulation of nonsense-mediated decay, translation yield, cytoplasmic localization, and nuclear export (2-5). Despite their numerous functions and presence in all sequenced eukaryotic genomes, the origin of introns is still debated and two competing theories exist. Intron early theory states that ancestral prokaryotes possessed introns but purged them over time (6-7). This theory implies that intron loss is the primary evolutionary force at hand. In contrast, intron late theory suggests that prokaryotes never harbored introns, and that eukaryotes evolved to acquire them, suggesting that intron gain is the principal force (7-8).

Several studies have investigated intron loss and gain events in different species. Many have found profound evidence of widespread intron loss in fungi (9-10). In contrast, intron gain has been identified in a limited number of species and is considered to be a relatively rare event. Additionally, the primary drivers of intron gain remain unclear (5, 11). Sequencing efforts on the prasinophyte alga, Micromonas pusilla, revealed one of the most notable events of intron gain ever detected, in which specific introns inserted at novel positions in genes across the genome (12). Contrary to canonical introns, these introns (introner elements or IEs), exhibit conserved sequences and lengths and proliferate through a transposition mechanism (11). Similar events have been observed in the pelagophyte alga, Aureococcus anophagefferens, and in some fungi (13).

In light of these discoveries, I hypothesized that transposition is an important driver of dramatic intron gain in diverse lineages. I developed and applied a pipeline for the systematic discovery of such introns across all annotated genomes in the GenBank database available through NCBI (14). The pipeline employs a Girvan-Newman algorithm to cluster introns based on sequence similarity and associate putative introner element families, then compares homologous intron positions among closely related species to identify intron gain events. I report novel IE discoveries in 38 species across 11 phyla, a major expansion upon the handful of previously known cases, which spanned only three phyla. Overall, my results reveal that transposons may operate as a fundamental driver of intron gain in a subset of lineages. Nonetheless, important questions remain about the frequency, function and fitness effects of new intron insertions. Downstream analyses demonstrate that IEs proliferate through various transposition mechanisms and highlight their multifarious effects on genomic architecture. The functional impacts of introner insertions predict evolutionary outcomes and suggest that some are deleterious while others do not display obvious fitness effects.

2 Background

2.1 What are spliceosomal introns?

Introns are stretches of noncoding DNA found between exons that must be removed from RNA before translation. Four classes of introns exist: group I introns, group II introns, pre-tRNA introns, and spliceosomal introns. Spliceosomal introns are a defining characteristic of eukaryotes and do not exist in prokaryotes (15). During transcription, RNA polymerase produces pre-mRNA that contains both spliceosomal introns and exons from template DNA. This pre-mRNA then undergoes a process known as splicing, catalyzed by the spliceosome. The spliceosome is a complex molecular machine composed of numerous small nuclear RNAs (snRNAs) and proteins. It excises introns

from a pre-mRNA transcript and ligates exons together to construct a mature mRNA transcript that can be correctly translated into protein via a two step mechanism (Figure 1). Excised intron lariats are discarded as a result.

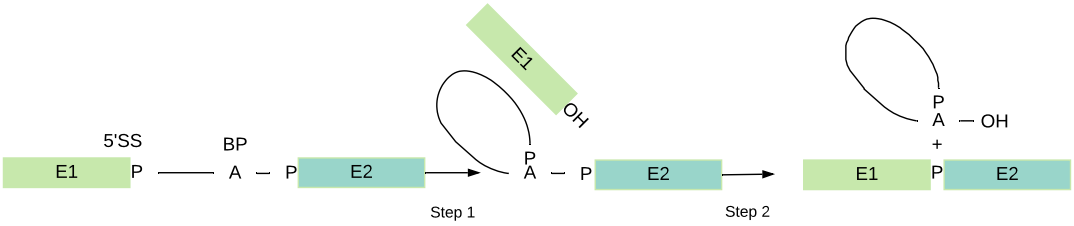

Figure 1: A schematic representation of the 2 step splicing mechanism. This process results in 2 products: a mature mRNA transcript containing E1 and E2, and an intron lariat.

2.2 The roles of spliceosomal introns

Despite their removal from mature mRNA transcripts, introns play several roles within eukaryotic organisms. These include regulation of gene expression, alternative splicing, and regulation of nonsense mediated decay. Introns also play notable roles in translation yield, cytoplasmic localization, and nuclear export (4).

Regulation of gene expression

Introns modify gene expression levels. Studies show that constructs containing introns exhibit higher expression levels than those lacking them (2, 19-21). Introns can contain elements that regulate transcription initiation, such as enhancers and silencers (23-26). They can also enhance 3'-end processing in a splicing independent-manner, which significantly increases the efficiency of transcription termination (4). Increased expression of advantageous genes can warrant increased protein production and can be tremendously beneficial to a cell.

Alternative splicing

Introns play a crucial role in alternative splicing. Alternative splicing allows for the production of multiple different proteins from a single gene and occurs when the spliceosome recognizes certain splice sites at different times and under different conditions. As a result, many eukaryotic organisms exhibit a proteome diversity far exceeding the number of genes in their genome (27). Introns enable alternative splicing to take place through their mere existence. However, they also host cis regulatory elements that regulate alternative splicing (28). These regulatory elements can either initiate or repress spliceosome assembly at a given splice site.

Regulation of Nonsense Mediated Decay (NMD)

Nonsense mediated decay is a surveillance pathway in all eukaryotes. Its primary function is to eliminate mRNAs containing premature stop codons. Therefore, NMD reduces deleterious errors in gene expression. The most common form of NMD is splicing dependent (4). NMD classifies an exon-exon junction complex 50-55 base pairs downstream of an authentic termination codon as a premature transcript. Thus, introns probably play an important role in premature mRNA target

recognition. Studies show that introns in 3' and 5' untranslated regions control NMD transcript sensitivity (5).

2.3 Intron theory

The discovery of introns shook the scientific community and immediately prompted many questions (29). At what point did introns form on the evolutionary timeline? How did they form? What roles do they play in molecular biology? Extensive research initiatives provide answers to many of these questions. However, the origin of introns is still heavily debated and two competing theories exist (Figure 2). Intron early theory suggests that ancestral prokaryotes possessed introns but lost them due to genomic streaming (6-7). This first theory implies that intron loss is the primary evolutionary force. In contrast, intron late theory suggests that prokaryotes never harbored introns, and that eukaryotes evolved to acquire them. This second theory points towards intron gain as the main force at hand and is now widely supported (7-8).

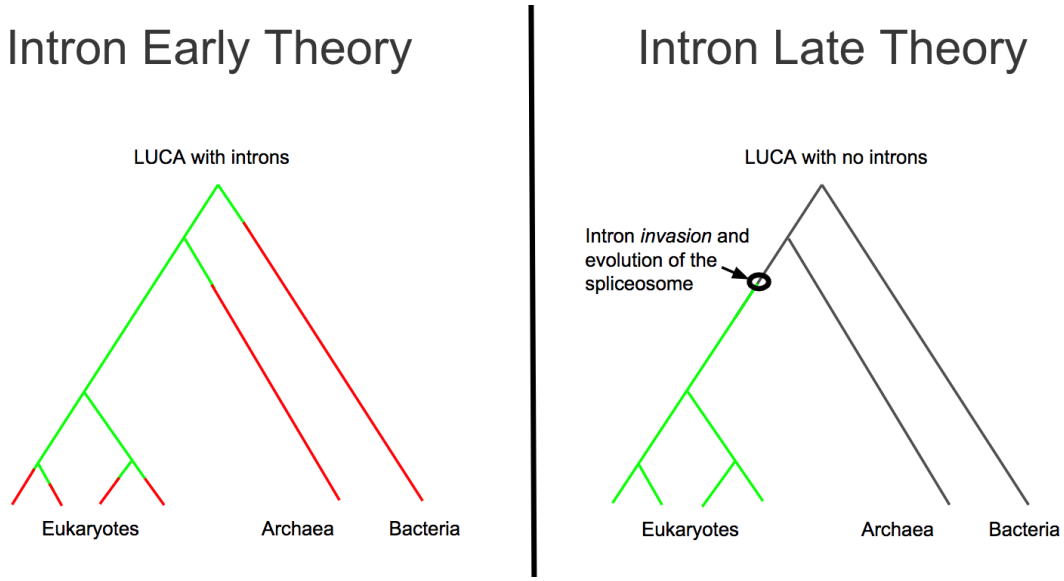

Figure 2: Intron early theory versus intron late theory. Red represents evolutionary intron loss and green represents gain. Intron early theory suggests that introns existed in early prokaryotic genomes but were purged over time. In contrast, intron late theory states that introns invaded eukaryotic genomes after the divergence between prokaryotes and eukaryotes.

2.4 Proposed mechanisms of intron loss and gain

Studies have investigated intron loss and gain events in different species. Several found profound evidence of widespread intron loss in fungi (6, 10). On the contrary, intron gain has been identified in a limited number of species and was considered until recently to be a relatively rare event (1, 11, 13). Several models for intron loss and gain exist (1, 9, 13, 30-31). Intron loss has been experimentally shown to occur through RNA mediated homologous recombination of cDNA and genomic deletion (30, 32-33). Proposed mechanisms of intron gain include transposition, intron gain during double stranded break repair, transfer of an intron from a paralogous gene, intronization and tandem genomic duplication (30-31). However, none have been experimentally validated.

Moreover, transposition is the only mechanism that has been shown to cause widespread intron gain on genomic scales (13).

2.5 Introner elements -mobile introns

The prasinophyte, Micromonas pusilla, is a unicellular marine algae found worldwide. Sequencing efforts on M. pusilla strain CCMP1545 reveal one of the most notable events of intron gain ever detected, in which specific introns have inserted at novel positions in genes across the genome (11). Contrary to canonical introns, these introns, deemed introner elements (IEs), exhibit conserved sequences and lengths (11). Massive introner element invasion caused the number of introns in strain CCMP1545 to double (11). Similar events have been observed in the pelagophyte alga, Aureococcus anophagefferens, and in some fungi (13).

Several studies originally proposed that IEs propagate through a transposition mechanism involving reverse splicing because introners are commonly found in coding regions and in sense orientation (1, 11, 31, 34-35). Simmons et al. illustrates an intriguing mechanism for reverse splicing of IEs into ssDNA produced during R-loops (1) (Figure 3). R-loops are abnormal structures that form behind a stalled RNA polymerase (1). They form when the elongating RNA strand binds to the complementary unwound DNA template strand behind the polymerase (1). R-loops exhibit higher mutation rates and genomic instability, making them an optimal target for IE insertion via reverse splicing (1, 36). In Simmons et al.'s proposed mechanism, the 3' OH of the IE lariat attacks a phosphate in the displaced ssDNA, then acts as the leaving group. The DNA 3' OH then attacks the phosphate of the IE lariat branch. The branch point adenosine 2' OH acts as the leaving group, and the IE inserts into the template ssDNA. Enzymes repair the R-loop and either reverse transcriptase or a DNA repair polymerase copies the inserted RNA to DNA on the complementary strand. DNA polymerase I converts the IE RNA into DNA, resulting in a new intron (1).

A recent study completed by Huff et al. suggests that IEs in M. pusilla proliferate through a DNA transposition mechanism involving a 3bp target site duplication (TSD) (13) (Figure 4). TSDs are a hallmark of direct DNA transposition, and do not result from reverse splicing (13). TSDs form when a DNA transposase makes staggered cuts in target genomic DNA, producing sticky ends that must be repaired after transposon insertion. Huff et al. observes TSDs in a significant number of IE insertions in M. pusilla and A. anophagefferens (13) (Figure 4). Huff et al. also demonstrates that IEs in M. pusilla and A. anophagefferens insert into linker DNA between nucleosomes (13). Chromatin structures like nucleosomes form complex structures of tightly knit DNA, limiting its accessibility and suppressing transposon insertion (37). Thus, insertion into linker DNA is also indicative of DNA transposition (13). Although introners are known from the genomes of a few representative clades, no study has previously conducted a systematic search for them more broadly. Furthermore, while several studies have proposed and analyzed potential mechanisms by which IEs propagate, the functional and fitness effects of intron insertions remain unexplored.

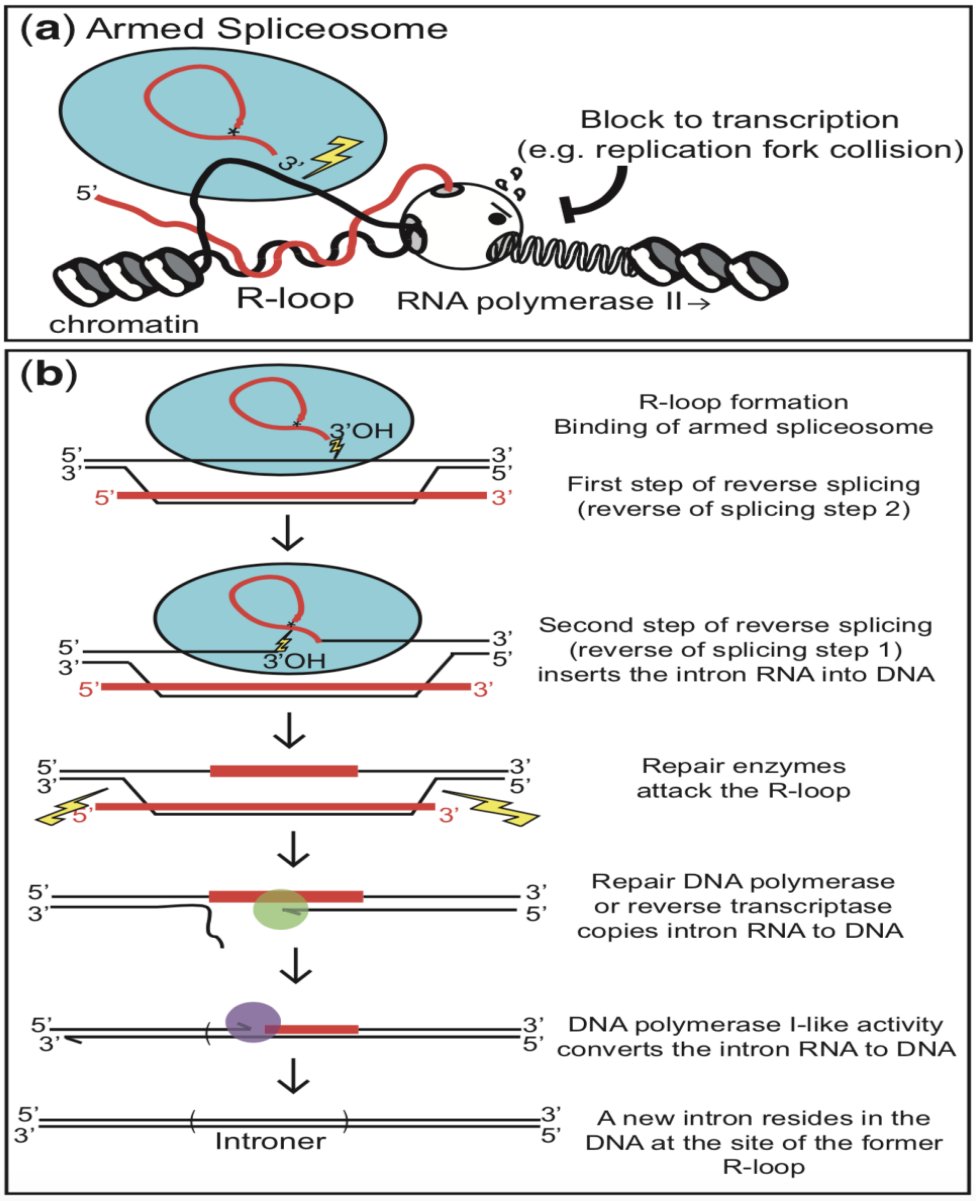

Figure 3: An illustration of Simmons et al.'s proposed model for IE gain in M. pusilla (1). (a) Illustration of a stalled RNA polymerase and consequent R-loop formation. The armed spliceosome carries a recently spliced IE lariat and binds to the displaced DNA template strand. (b) A step-by-step demonstration of Simmons et al.'s mechanism (23). The 3' OH of the IE lariat attacks a phosphate in the displaced ssDNA, then acts as the leaving group in this transesterification. The DNA 3' OH then attacks the phosphate of the IE lariat branch. The branch point adenosine 2' OH acts as the leaving group, and the IE inserts into the template ssDNA. Enzymes repair the R-loop and either reverse transcriptase or a DNA repair polymerase copies the inserted RNA to DNA on the complementary strand. DNA polymerase I converts the IE RNA into DNA, and a new IE insertion prevails (1).

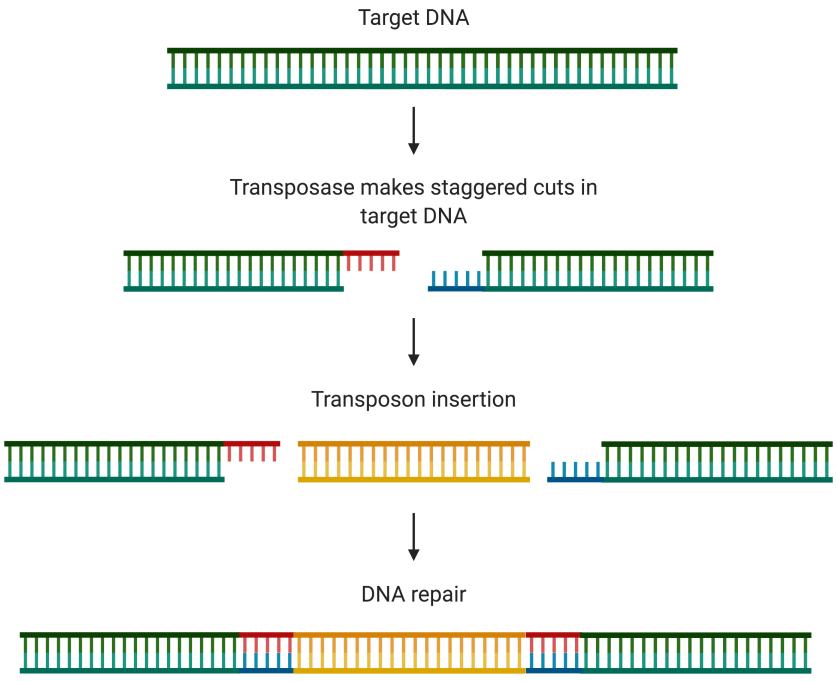

Figure 4: A diagram summarizing target site duplication formation. A DNA transposase makes a staggered cut in target genomic DNA. This staggered cut results in sticky ends on each strand. After transposon insertion, a DNA polymerase repairs these sticky ends, resulting in target site duplication.

3 Work completed

3.1 Intron element detection pipeline

In light of their prevalence in *Micromonas pusilla*, I hypothesized that intron elements likely inhabit other eukaryotic genomes. I hence constructed a pipeline for intron element detection from any genome assembly available in NCBI's Genbank database (Figure 5, 38). It merely requires a genomic DNA fasta file and annotation gff file (that provides gene and exon positions) as input. The pipeline consists of five major steps and generates several transitional files. It ultimately produces a fasta file containing confirmed intron elements and an ex-int file illustrating where intron elements have caused intron gain events.

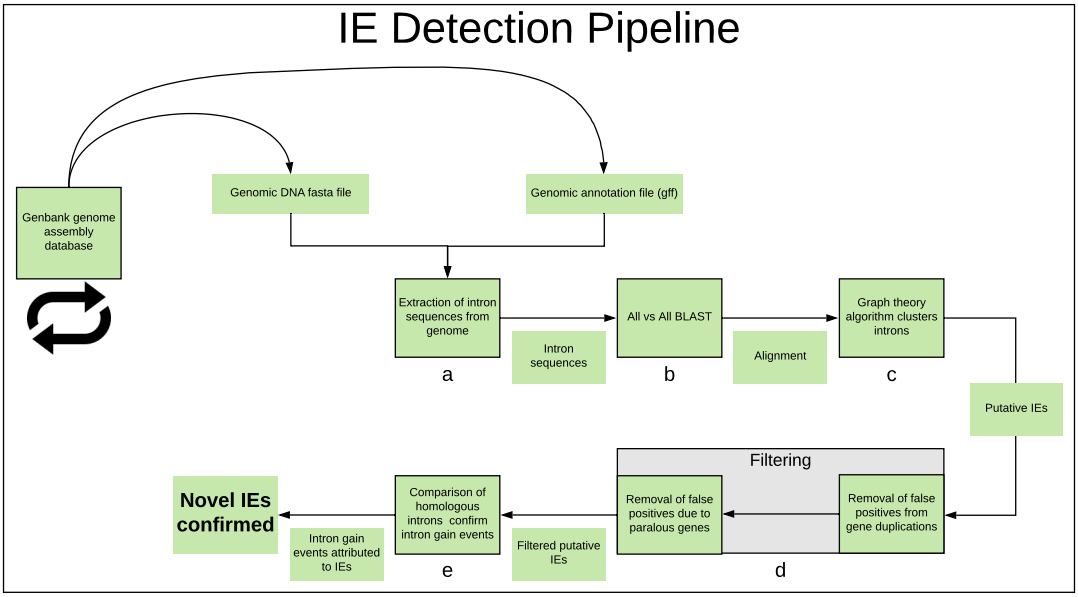

Figure 5: IE detection pipeline. I developed a multistep pipeline to detect intron elements in all annotated genomes on the Genbank database available through NCBI. I ran the pipeline in series on each genome independently (a) Intron sequences are extracted from a genome. (b) All vs. all BLAST generates pairwise overlaps between all introns. (c) Graph theory based algorithm clusters introns that are similar in sequence and length, generating a fasta file containing putative IEs. (d) False positives caused by gene duplications and paralogs are filtered. (e) Comparison of homologous intron positions confirms intron gain events, establishing high confidence IEs.

Extracting intron sequences

Intron coordinates are not explicitly provided in Genbank genomic annotation files. Consequently, I used exon coordinates to retrieve them. For each gene, the start of Intron 1 is the stop of exon 1, and the stop for intron 1 is the start of exon 2. The start of intron 2 is the stop of exon 2, and the stop of exon 2 is the start of exon 3, and so on. I then used these coordinates to extract intron sequences from the genomic DNA fasta file and deposited them into a fasta file.

All vs. all alignment

I performed an all vs. all alignment to generate pairwise comparisons between all extracted introns in each species. I started by generating a custom BLAST database from all introns in a particular genome. I implemented BLAST+'s *makeblastdb* command with options *-dbtype nucl*, *-in (fasta file containing intron sequences)*, *-title introns* and *-out (introns database file)* (39). Next, I blasted intron sequences to the custom intron blast database. I implemented blastn with parameters *-db (intron database)*, *-query (intron fasta file)*, *-outfmt 6*, *-perc_identity 80* and *-out (blast output file)*. The parameter, *-outfmt 6*, ensures that results are deposited in a tab separated file. The parameter, *-perc_identity 80*, restricts sequence overlaps to require a minimum 80 percent identity. I then parsed the newly generated all vs. all alignment file, trashed all duplicate and self hits, and removed overlaps with less than 30 residue matches. In addition, I require that each overlap exhibits an alignment length that exceeds 80% of each respective intron length.

Clustering

I applied a Girvan-Newman algorithm available through the python3 module, *NetworkX* (40, see https://networkx.github.io/documentation/stable/index.html), to cluster introns based on sequence similarity. The algorithm identifies communities by progressively removing edges from an initial network. It generally removes the edge with the highest betweenness centrality at each step, eventually exposing the most tightly knit communities in that network. I generate a network for each species, in which each node represents an intron and each link between two nodes represents the sequence similarity between those introns. The algorithm effectively reveals putative intron element families with high sequence similarity relative to other introns. These families form clusters which I can visually interpret on a graph (Figure 6). I deposited putative families with more than four IEs in a fasta file for downstream filtering

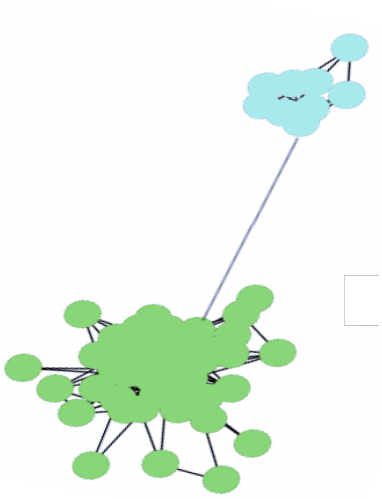

Figure 6: Visualization of a network generated from IEs in *Alternaria alternata*. Each node represents an intron and each edge represents the sequence similarity between two introns. Prospective IEs that exhibit high sequence similarity cluster into tightly knit networks.

Downstream filtering

Gene duplications

Gene duplication could result in the presence of highly similar introns. I filtered out these prospective false positives by comparing the flanking regions of each putative IE to those of all other IEs in its respective family. First, I extracted the 100 base pairs (bp) flanking each IE and concatenated them. Next, I blasted the concatenated edges for each IE against all other members of its respective family with a minimum 80 percent identity requirement (39). I also limited output to only include overlaps with an alignment block longer than 140 bp (70% of the two concatenated 100bp flanking sequences). I then parsed the remaining overlaps and removed its comprising IEs.

Paralogous genes

I filtered out false positives caused by paralogous genes by comparing protein sequences of all IE containing genes for each IE family. First, I parsed the protein sequence for each IE containing gene from the respective assembly's protein fasta file. Then I used blastp to perform an all vs. all alignment between each IE containing gene's corresponding protein and all other proteins corresponding to genes containing IEs in its respective family (39). I required overlaps to exhibit over 50% identity and a minimum alignment block length of 100 amino acids. I labeled any IEs corresponding to the remaining overlaps as false positives due to paralogous genes and discarded them. I deposited the remaining IEs into a final fasta file.

Homologous intron comparison

I employed a modified Perl script (contributed by Scott W. Roy, Professor of Biology at San Francisco State University), to compare homologous intron positions for each putative intron element and subsequently recognize intron gain events. The script generates an ex-int file which annotates where intron positions differ between species. I compared homologous intron positions between each IE containing species and the four most closely related species taxonomically with annotated genomes in the Genbank database. I also included one outgroup species to control for intron loss events. When parsing our output, I only considered genes for which I could identify high confidence orthologs. I then examined our output and identified cases in which putative IEs existed in positions at which more than one compared species (plus our outgroup) lacked an intron. I acknowledged such events as intron gains, and filtered out any species in which I did not observe at least four intron gains attributed to putative IEs. I was confident that IEs in the remaining species were real since they possessed high sequence similarity and caused intron gain events, and thus began conducting downstream analyses.

3.2 IEs exist in a wide range of eukaryotic lineages and propagate via multiple transposition mechanisms

I report novel IE discoveries in 38 species across 11 phyla, a major expansion upon the handful of previously known cases, which spanned only three phyla (Figure 7). IE insertion frequency varies substantially between species. Some genomes contain as few as five elements while others harbour thousands (Figure 8). Such variance in IE success hints at the possibility of multiple acting mechanisms. I examined insertional biases and IE architecture independently in each species in search of species (or clade) specific patterns.

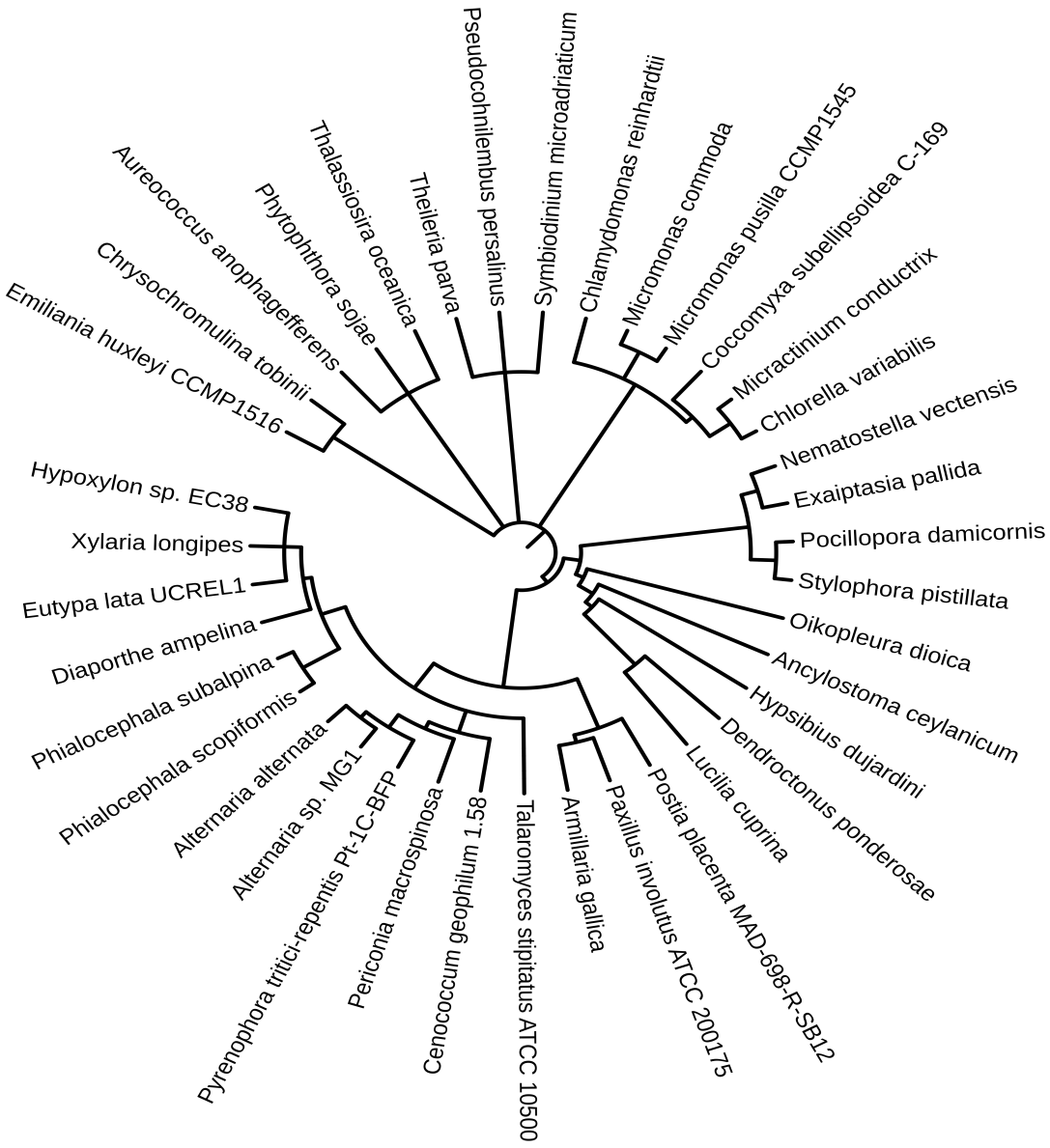

Figure 7: Circular phylogenetic tree of IE containing species rooted at the Last Universal Common Ancestor (LUCA).

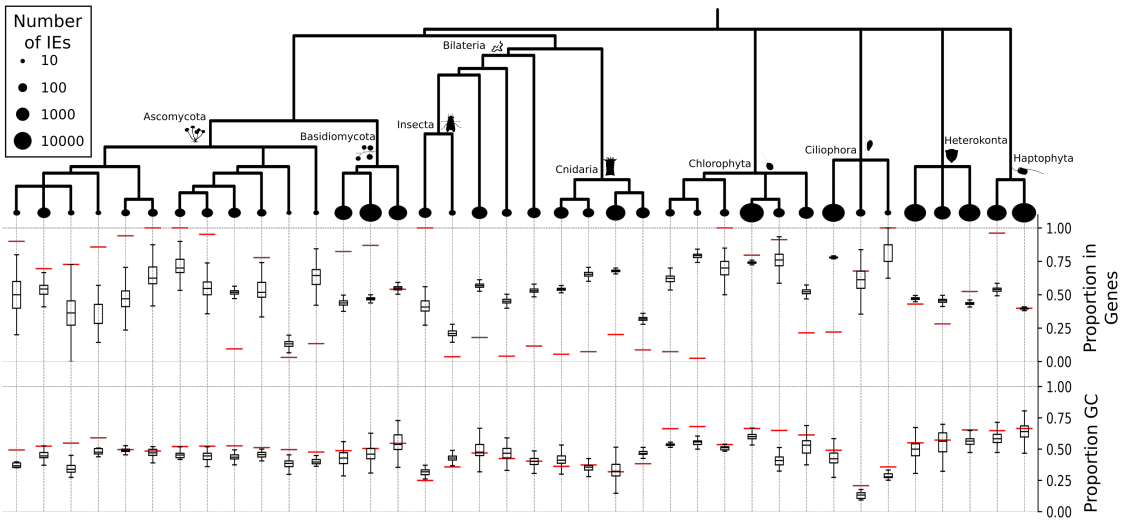

Figure 8: Phylogenetic tree displaying the diverse array of clades containing IEs. circle sizes at the end of species nodes are proportional to the number of IEs in respective species. Box plots adjacent to species nodes represent 1000 permutations for the expected proportion of IEs in genes. Red lines denote observed values. The second set of box plots illustrate the expected proportion GC at IE insertion sites. Again, red lines denote actual values.

Proportion in genes

I was especially curious about the proportion of IE insertions in genes, G, where

$$G = \frac{insertions\,in\,genes}{insertions\,in\,intergenic\,regions}.$$

Huff et al. provided substantial evidence that IEs in *Micromonas pusilla* and *Aureococcus anophagefferens* proliferate via DNA transposition. DNA transposons insert semi-randomly (41). Thus, IEs driven by DNA transposition should also exist in intergenic regions. DNA transposons also lack orientational preference. Spliceosomal introns can only be spliced in the 5' to 3' direction. Even if an IE contains both splice sites in one direction, approximately 50% of its insertions should be in antisense orientation, subsequently generating unspliceable introns. Such insertions will have adverse effects on gene expression and should be filtered by selection (13). Thus, I expect to observe relatively low G values.

Since introners in intergenic regions are unannotated, we developed a systematic method to recover them. I employed multiple alignment using fast Fourier transform (MAFFT, 42) to conduct multiple sequence alignments for each introner element family in each species. Next, I generated a consensus sequence for each introner element family using a custom positional nucleotide frequency matrix. Then, I BLASTed each consensus to its corresponding reference genome and filtered duplicate and self hits. I parsed the remaining hits from the reference and deposited them in a fasta file.

At first glance, G seemed strangely large in many species. However, G is directly correlated to gene density. In other words, if a genome is more gene dense, a transposon is more likely to land in a gene, resulting in a larger G. To correct for this, I generated 1000 permutations for the probability that a particular IE will insert into a gene by chance. I represented each probability with a binomial distribution, such that n = the number of total insertions and p = gene density. I compared these with our actual values to test for insertional enrichment in genic regions ($\alpha = 0.05$, Figure 8). To my surprise, I found that IEs in several species seemingly target genes (Supplementary Table). Shockingly, in some species, IEs only exist in genes, suggesting possible alternative mechanisms to DNA transposition.

GC bias

Perplexed by my observation of insertional bias in genes, I sought possible alternative explanations. Genes are generally GC rich, and transposable elements have been shown to display preference for GC rich regions (43). To test whether IEs that are enriched in genes also demonstrate a bias for GC rich regions, I again employed a permutation test. I calculated the GC content for the concatenated 10 base pairs (bp) upstream and downstream of each insertion (20bp total), noting the insertion's location as either genic or intergenic. I then generated 1000 permutations for the GC content of randomly resampled 20 bp regions (n = total number of insertions). I compared our observed GC proportions to randomly sampled ones in both genic and intergenic regions and found that IEs in many species are also enriched in GC rich regions ($\alpha = 0.05$, Figure 7). I found a significant correlation between insertional enrichment in genes and insertional enrichment in GC rich regions across all species, suggesting that IEs home on GC rich regions rather than genes (one-tail sign test P = 0.0168, Figure 8).

Nucleosome linker DNA

Huff et al. notes that IEs in *Micromonas pusilla* insert into linker DNA between nucleosomes, the basic structural unit of eukaryotic chromatin, and form new nucleosomes as a result. They observed this phenomenon when surveying previously generated chromatin maps for the *Micromonas pusilla* genome. However, the vast majority of genomes in our study lack chromatin maps or ATAC-seq data (data used for generating chromatin maps). Thus, I was forced to use an alternative method. I employed a Hidden Markov Model (HMM) based algorithm to predict nucleosome profiles for the regions surrounding IE insertions and benchmarked it using the *Micromonas pusilla* genome (44).

The model performed reasonably well for species in which I observed a high number of insertions (when sample size was large), but failed to produce convincing signatures for others. Consistent with Huff et al., I observe a strong preference for IE insertions in nucleosome linker regions in *Aureococcus anophagefferens* (Figure 9). However, in the Haptophyte, *Chrysochromulina tobinii*, I observe an enrichment of IE insertions in predicted nucleosomes, a phenomenon characteristic of some retroelements (45, Figure 9). I plotted p-occupancy, the predicted probability that a nucleosome exists at a particular position, for regions surrounding IE insertion sites in *Aureococcus anophagefferens* and Chrysochromulina tobinii (Figure 9). I marked insertion positions with vertical dashed lines. In *Aureococcus anophagefferens*, I observe a trough at the insertion position, indicating that IEs often insert into linker DNA. In contrast, in *Chrysochromulina tobinii*, I observe a peak at the same position, suggesting that IEs might preferentially insert into nucleosomes in some species, and hinting at the presence of alternative mechanisms.

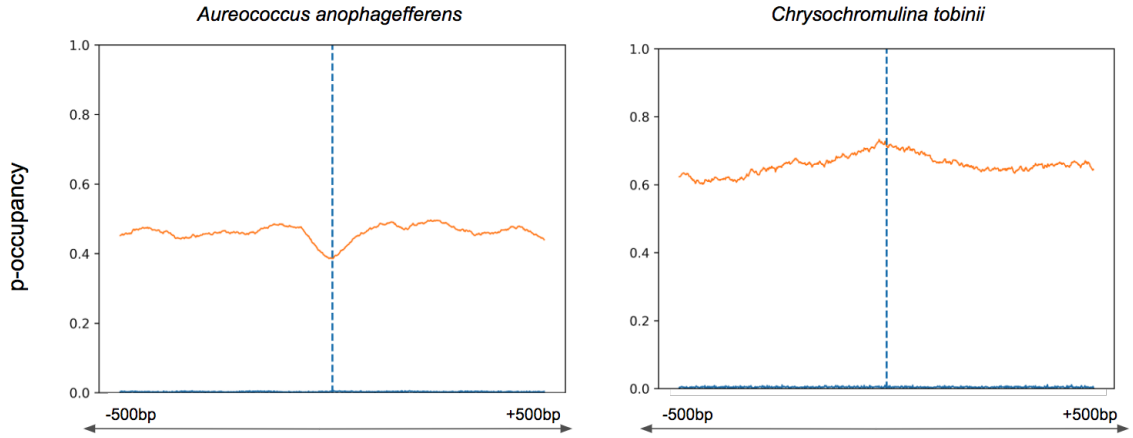

Figure 9: Nucleosome p-occupancy predictions produced using an HMM (44) for regions surrounding IE insertions in *Aureococcus anophagefferens* and *Chrysochromulina tobinii*. Insertion sites are noted by a vertical blue dashed line. Nucleosome p-occupancy refers to the probability that a nucleosome exists at a particular site. In *Aureococcus anophagefferens*, p-occupancy is low at insertion sites relative to surrounding regions, suggesting that IEs insert into linker DNA between nucleosomes. In contrast, in *Chrysochromulina tobinii*, p-occupancy peaks near IE insertion, suggesting that IEs insert into nucleosomes.

Dissatisfied with the limits of our HMM, we searched for other possible evidence that IEs generate nucleosomes upon insertion. Nucleosomes are generally 146-206 bp in length. We calculated the mean length of IEs in each species and found that IEs that exhibit lengths within this range are more abundant (Supplementary Table, Figure 10, MWU P = 0.0123). These results suggest that IEs that might append chromatin structures are less deleterious and highlight the possibility

of adaptation.

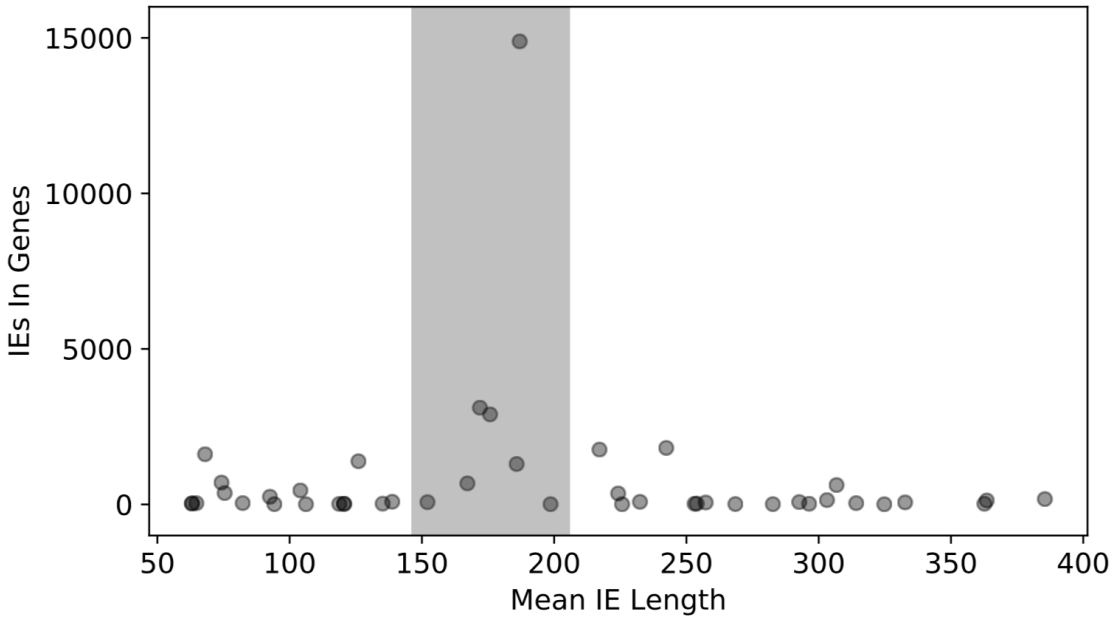

Figure 10: Scatter plot of genic insertion count relative to mean IE length. The range of possible nucleosome lengths is shaded in grey. IEs with the greatest genic insertion frequencies tend to exhibit a mean length within the range of possible nucleosome lengths (MWU P = 0.0123).

IEs propagate via multiple mechanisms

Evidence of DNA transposition or retrotransposition in most eukaryotes

Introners in different species likely proliferate through different mechanisms which directly predict their frequency and insertion biases (Figure 8, Figure 11). I find evidence supporting a DNA transposition or retrotransposition mechanism in several species. In these species, I observe IE insertions in noncoding regions at high frequencies and detect the presence of target site duplications (TSDs) at IE insertion sites (Figure 11). Target site duplications are a hallmark of DNA transposition and some retrotransposition mechanisms.

To recognize TSDs, I parsed the 10bp upstream and downstream of each splice site for all introns in each species. Next, I generated a matrix for each IE, encoding the frequency at which nucleotides were identical at positions from -10 to 10 relative to each splice site. I did the same for all other introns. Then, I subtracted our intron matrix from our IE matrix to remove possible artefacts caused by conserved features of introns (such as respective GT and AG 5' and 3' splice sites), revealing the positions specific to IEs at which nucleotides were most often identical relative to 5' and 3' splice sites. I plotted each matrix on a heat map and was able to visually deduct possible TSD locations and lengths (Figure 11). Once I possessed approximate TSD locations and lengths, I conducted a permutation test to ensure that TSD signatures did not appear by chance (N = 1000). I randomly sampled l bp at n positions from each genome, such that n = total number of IE insertions and l = TSD length, and checked to see if the l bp downstream = l bp upstream at each position. I calculated the probability that I would observe TSDs by chance in each species and compared it to the respective actual value (Figure 11: *Aureococcus anophagefferens* $P \leq 0.001$). In Figure 11, I compare identical base frequency matrices in two species. IEs in *Aureococcus*

anophagefferens produce TSDs upon insertion ($P \leq 0.001$) and those in *Hypsibius dujardini* do not. I annotate the diagonal string of relatively high frequency identical bases indicative of a prospective TSD location with a red box.

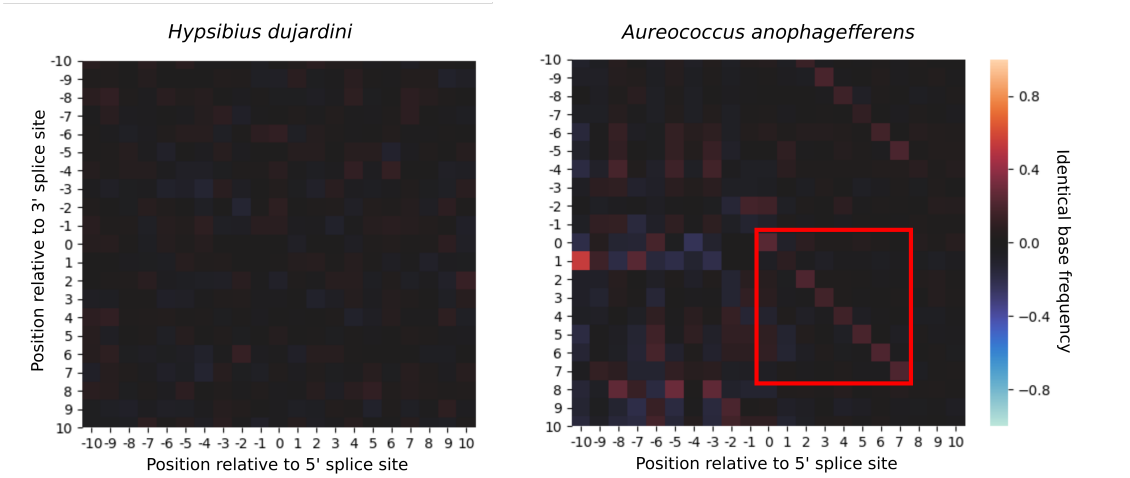

Figure 11: Heat maps demonstrating identical base frequencies between all positions from -10 to +10 bp relative to the 5' and 3' splice sites of IEs in *Hypsibius dujardini* and *Aureococcus anophagefferens*. IEs in *Hypsibius dujardini* display no evidence of TSDs, since base frequencies are essentially neutral between all positions. In contrast, IEs in *Aureococcus anophagefferens*, exhibit diagonal strings of identical bases at high frequencies indicative of possible TSDs.

I provide substantial evidence that DNA transposition or retrotransposition drives IE proliferation in the vast majority of considered species (Supplementary Table). These species are characterized by high insertion frequencies, the presence of insertions in intergenic regions, and often the presence of high confidence TSD signatures. However, IEs in the fungal clade, Ascomycota, do not exhibit these features, suggesting the possible contribution of alternative (and possibly clade specific) mechanisms.

Evidence for distinct reverse splicing

I observe a second distinct class of IEs that fails to exhibit the expected characteristics for DNA transposition or retrotransposition. Consistent with an RNA-based reverse splicing mechanism: these IEs exhibit more pronounced sequence and length similarity; primarily (and in some cases exclusively) exist in genes in sense orientation; occur at significantly lower frequencies; and do not produce TSDs (Figure 8, Figure 12). In Figure 12, I display a multiple sequence alignment of IEs in the Ascomycete, *Alternaria alternata*, produced with MAFFT (42) and visualized with Jalview (46). Introners in *Alternaria alternata* exhibit highly similar sequences and lengths. Remarkably, among our kingdom-wide set of IEs, these features were specific to Ascomycetes, with 2 possible exceptions (*Theileria parva* and *Coccomyxa subellipsoidea*). Thus, our search has both revealed evidence that DNA transposition or retrotransposition is the primary driver of IE proliferation across eukaryotes while also strengthening the evidence that fungal IEs represent a profoundly different phenomenon (Figure 14A-B).

Mechanistic consequences

IE proliferation mechanisms predict their frequencies and insertional distribution. In the case of IEs driven by DNA transposition or retrotransposition, TSD locations relative to 5' and 3' splice

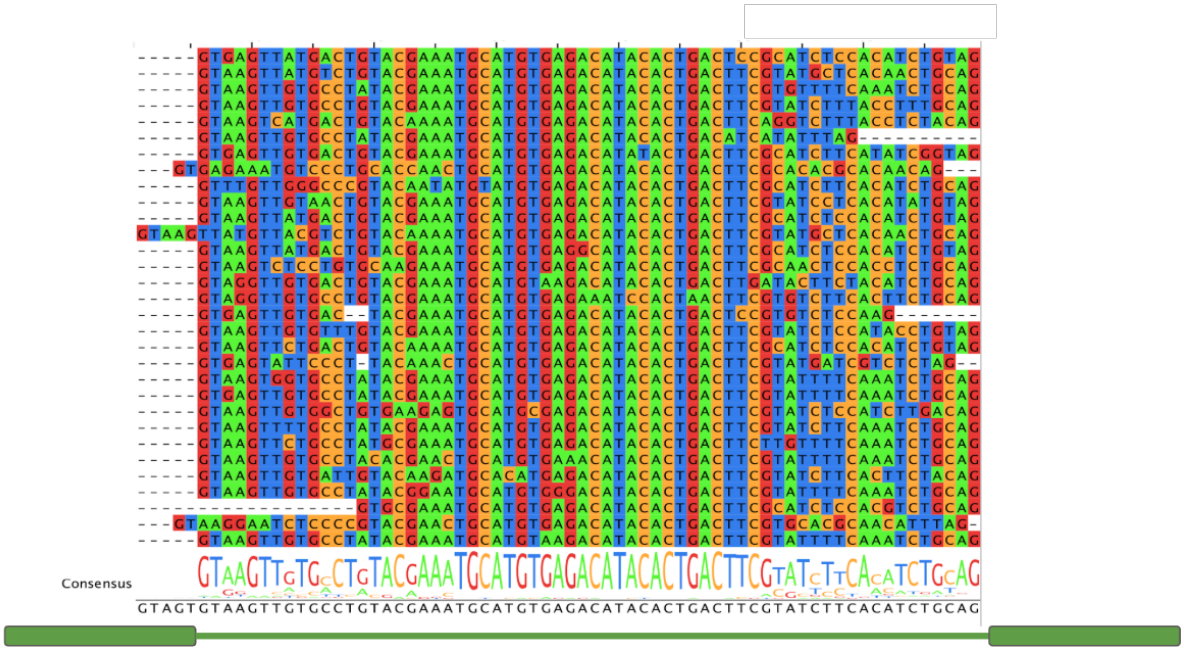

Figure 12: Multiple sequence alignment between IEs in *Alternaria alternata* (visualized with Jalview (46)). IEs in this species exhibit nearly identical full length sequence similarity and lack any variable regions, suggesting a propagation mechanism involving reverse splicing.

sites provide insight about relative fitness costs that reflect insertion frequencies (Figure 13). Some introners carry one splice site and co-opt the other from their insertion site in coincidence with Huff et al. 's observations in *Micromonas pusilla* and *Aureococcus anophagefferens* (13). These IEs exactly replace co-opted bases with TSDs and thus generate a new intron without changing surrounding coding sequences (Figure 13). As a result, I predict that such IEs are mildly deleterious. Consistent with this prediction, these IEs exist at the highest quantities in genes (*Symbiodinium microadriaticum* N = 1762, *Micromonas pusilla* N = 2895).

IEs in other species carry both splice sites, or in some cases, co-opt both from their insertion sites. I only observe IEs that carry both splice sites in Ascomycetes, which likely proliferate via a reverse splicing mechanism and do not produce TSDs (Figure 13). I expect IEs that co-opt both splice sites from their insertion site to be more deleterious, since they boast a much lower probability (approximately $\frac{1}{2}$ the probability for an IE that only co-opts one splice site) of generating an intron upon insertion (Figure 13). Indeed, I observe such IEs at lower frequencies in genes (*Phytophthora sojae* N = 247).

3.3 IE architecture

Secondary structure

(Bryan Thornlow, PhD candidate in the Corbett-Detig Lab, also contributed to this work)
We examined IE architecture in each species and found that several produce stable secondary structures. We hypothesized that these structures pose some functional value since they have been preserved by selection. To determine whether or not these structures were unique to IEs, we compared the stability of IEs and non-IE introns based on minimum free energy (MFE). We used *RNAfold* from *ViennaRNA* (48) to calculate the minimum free energy for IEs and non-IE introns

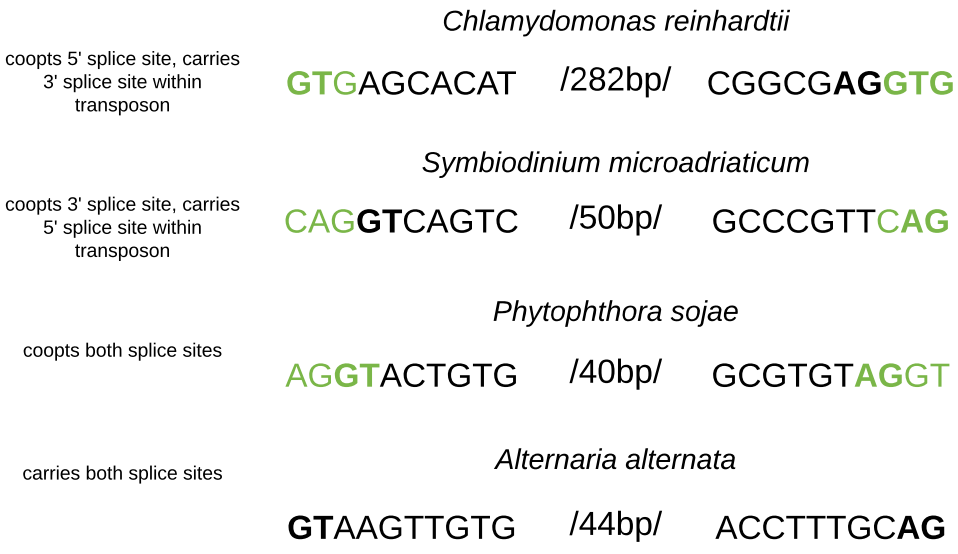

Figure 13: Examples of IEs TSD locations relative to splice site positions. TSDs are colored green. Splice sites are bolded. IEs in *Chlamydomonas reinhardtii* generate 3bp TSDs, co-opt their 5' splice site and their carry 3' splice site. IEs in *Symbiodinium microadriaticum* also generate 3bp TSDs, but instead co-opt their 3' splice site and carry their 5' splice site. IEs in *Phytophora sojae* generate 4bp TSDs and co-opt both splice sites from their insertion site. IEs in *Alternaria alternata* do not generate TSDs and presumably carry both splice sites.

using the *–noGU* argument to approximate DNA folding. We also controlled for intron length. We calculated the mean length of IEs in each species. Then, we trained a generalized linear model (GLM) with the formula,

$$glm(formula = MFE \sim ile \times length),$$

using the Gaussian family (identity link function). We used the model to predict the MFE for instances with the mean minimum free energy across IEs with the "ile" variable set to 0 and to 1. We recorded both of these predictions in Supplementary Table. If the MFE prediction with "ile" = 1 was lower, we considered IEs to be more stable. We also recorded the number of paired and unpaired nucleotides based on each intron's secondary structure.

Our results differed between clades, further underlining the possibility of multiple clade specific mechanisms. We compared our results between the two clades in which observe the most IE containing species: ascomycetes and green algae (Figure 14). Recall that we previously provided evidence that IEs in these two clades propagate via different mechanisms. In coincidence with our prior observations, IEs in ascomycetes form significantly stronger secondary structures than other introns, alluding to an RNA mediated mechanism (Figure 14C-D). RNA is more unstable and more prone to degradation than DNA, and RNAs of functional importance often form secondary structures that increase stability. In contrast, IEs in green algae exhibit less stable secondary structures than other introns, suggesting that their proliferation mechanism lacks an RNA intermediate (Figure 14C-D).

We visualized IE secondary structures in each species independently using *RNAfold* (48) and found that some form extremely pronounced motifs, suggesting possible functional roles. The most conspicuous of these exist in the pelagic tunicate, *Oikopleura dioca* (Figure 15). IEs in this species

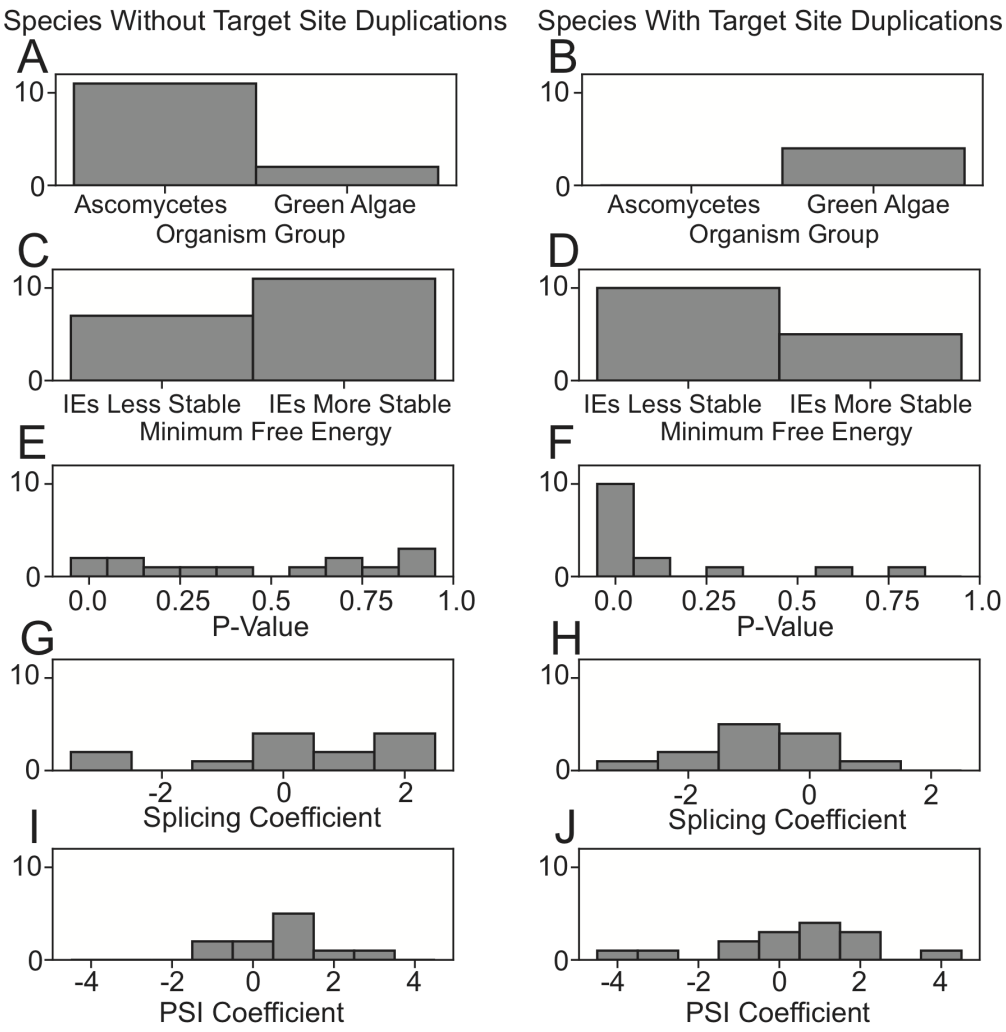

Figure 14: IEs demonstrate distinct splicing behavior dependent upon their mechanism of insertion. Species whose IEs do not have target site duplications (TSDs) are more frequently Ascomycetes (A), generally have IEs that fold more stably than other introns (C), are not enriched in weakly or highly expressed genes (E), more frequently use annotated splice sites than other introns (G) and are less frequently retained (I) than other introns. Conversely, species whose IEs do have TSDs are more often green algae (B), have IEs that fold less stably than other introns (D), are most often found in lowly expressed genes (F), mis-splice more frequently than other introns (H) and are spliced in roughly as frequently as other introns (J). Species containing IEs that are not ascomycetes or green algae are omitted from panels A and B. For panels E and F, p-value refers to results from our permutation test based on RNA-seq data. For G-J, coefficients refer to results from generalized linear models where a positive splicing coefficient (G,H) indicates that IEs splice using annotated splice junctions more frequently than other introns, and a positive PSI coefficient (I,J) indicates that IEs are retained or spliced in less frequently than other introns..

exhibit a highly conserved (within species) 250bp sequence that is a reverse complement to itself (Figure 15). As a result, the sequence binds to itself, forming a stable stem-loop (Figure 15).

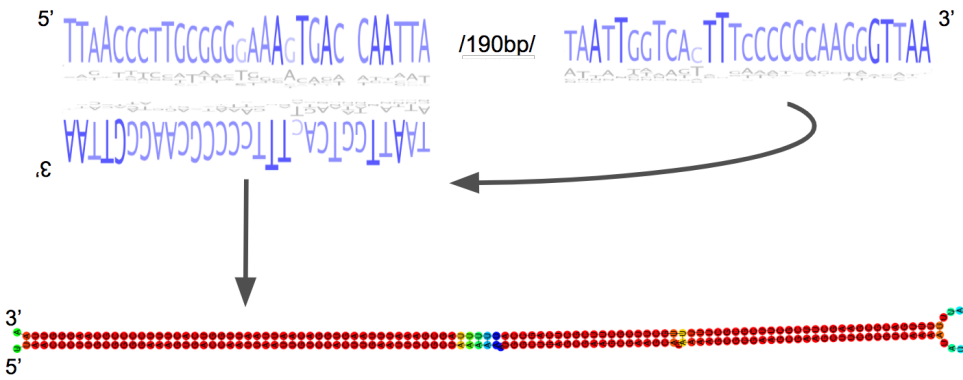

Figure 15: Consensus sequence motif for IEs in *Oikopleura dioca* and predicted secondary structure for that sequence (produced with RNAfold (43)). IEs in *Oikopleura dioca* are reverse complements of themselves, forming a highly stable stem secondary structure.

Non-canonical splice sites

All introns considered in this study are spliced by the U2 spliceosome, a complex molecular machine composed of RNA and proteins. Remarkably, canonical 5' "GT" (and in some species "GC") and 3' "AG" intron splice sites are conserved across nearly all eukaryotes (48). However, noncanonical "GA," and "GG" 5' splice sites have been observed at relatively low frequencies in many species. Additionally, noncanonical splice sites have been shown to occur at higher frequencies in species that have experienced recent intron gain (49). I hypothesized that newly inserted introns might exhibit noncanonical splice sites at higher frequencies than other introns. I used a chi squared test to identify possible enrichment of noncanonical splice sites in IEs and found such cases in *Symbiodinium microadriaticum* and *Oikopleura dioica* (chi squared $P < 0.0001$ for both species). *Symbiodinium microadriaticum* IEs demonstrate enrichment in noncanonical "GA" splice sites relative to other introns. Alternatively, *Oikopleura dioica* IEs exhibit enrichment in "GG" splice sites.

3.4 The functional and possible fitness effects of IEs

Insertions exist in unannotated predicted genes

DNA transposons insert semi randomly and do not usually display preferences for sense or antisense orientation. Thus, IEs that proliferate via DNA transposition should insert in antisense orientation approximately 50% of the time. Such insertions should be highly deleterious since introns can only be spliced in the 5' to 3' direction. I hypothesized that these insertions occasionally destroy genes and looked for evidence of open reading frames surrounding IEs in unannotated intergenic regions (Figure 16).

I used *orfm* to predict possible open reading frames (ORFs) in the regions flanking IEs in intergenic regions. I applied *orfm* with the parameter *-m 288* to filter predicted ORFs shorter than 288 nucleotides and found that IEs tend to exist in opposite orientation to predicted ORFs that flank them (Figure 16). Then, I BLASTed the ORFs surrounding each IE to the respective reference genome. I found several cases in which both predicted ORFs flanking a particular IE BLASTed to the same gene, suggesting that the IE impaired the gene upon insertion. I provide an example of one such case in the green alga, *Chlamydomonas reinhardtii* (Figure 17). An IE appears to have inserted into a predicted Kinesin Associated Protein (KAP) gene, causing it to no longer be transcribed. KAP is essential for the assembly and maintenance of flagella and cilia in various

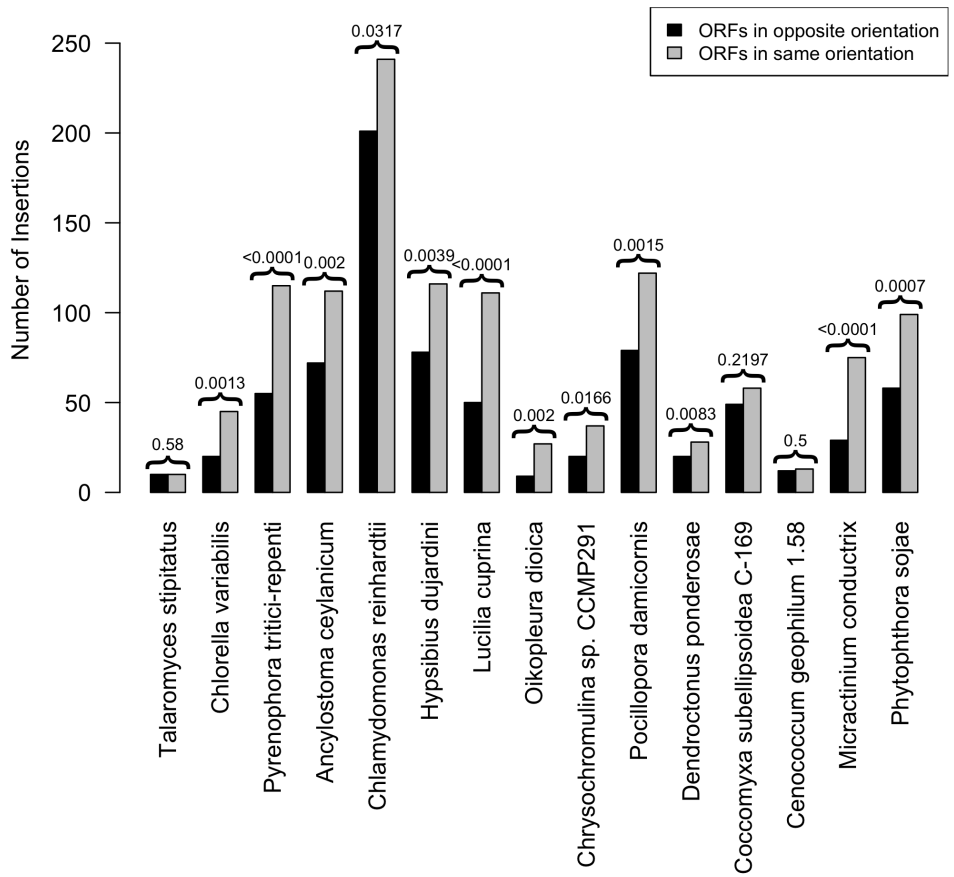

Figure 16: A bar plot comparing the number of insertions in which flanking ORFs are in opposite orientation to the number of insertions in which flanking ORFs are in the same orientation for IEs in a randomly selected subset of species. P values were generated for each species using Fisher's Exact Test to interrogate enrichment in cases in which insertions are flanked by ORFs in the same orientation ($\alpha = 0.05$). A P value resides above each set of box plots for the corresponding species. In most species, I observe a significant number of insertions flanked by ORFs in the same orientation.

cell types. Thus, this insertion was likely deleterious, but probably tolerable due to the presence of other KAP genes in the *Chlamydomonas reinhardtii* reference.

Effect on gene expression

(Bryan Thornlow, PhD candidate in the Corbett-Detig Lab, also contributed to this work)
IE insertions in coding regions should mostly be deleterious, although the extent of their functional and fitness effects likely varies between species. In order to examine the functional effects of introner elements on splicing and gene expression, we mined RNA-seq data from the NCBI SRA database for each species with introner elements. Accession numbers for each isolate used are available in (Supplementary Table). We analyzed untreated, wild-type isolates whenever available, prioritizing

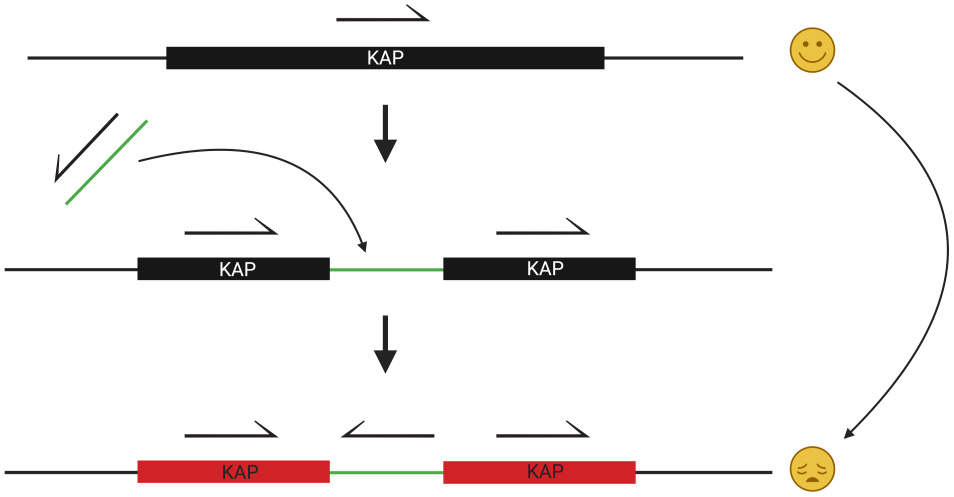

Figure 17: A cartoon depicting how an IE insertion may have destroyed a Kinesin Associated Protein (KAP) gene in *Chlamydomonas reinhardtii*. The orientation of each transcript is labeled with an arrow. An intron element (green) inserts into an exonic region of the KAP gene in opposite orientation and is thus unable to be properly spliced. The KAP gene is "dead" and no longer transcribed as a result.

RNA-seq performed on the same individual from which the genomic assembly was constructed. We aligned RNA reads to each genome assembly and quantified the reads corresponding to each gene using *STAR* (50). We then used *samtools* (51) to filter out reads with mapping quality scores below 20, followed by *leafcutter* (51) to extract splicing junctions. We used *samtools* depth to find the read depth surrounding each splice junction and within each intron. Following this, we used custom python scripts to collect read depth and splicing information for each intron, and to create generalized linear models comparing IEs and non-IE introns across species. We used the R package *lme4* (53) and fit to mixed models using the function *glmer()*, treating differences across RNA isolates as random effects. When only one isolate was available for a given species, we used the function *glm()*.

We hypothesized that IEs likely exist in more lowly expressed genes than other introns, since most insertions are probably deleterious. To test our hypothesis, we conducted a permutation test using our mined RNA-seq data. After aligning genes, we determined the normalized expression value for each gene using the *alignReads* and *quantMode* options in *STAR*, followed by division of the reads corresponding to each gene by the gene's length, as reads are more likely to cover longer genes. For each species, we sampled a set of genes equal to the number of IEs in that species and calculated the mean, normalized expression across this sample. This sampling was random, but weighted by the length of the gene, such that we were more likely to sample longer genes. We repeated this random sampling 10,000 times for each species, and calculated the proportion of these 10,000 samples whose mean normalized expression was lower than that of the mean normalized expression of all IE-containing genes. This value, divided by 10,000, served as our p-value for

expression data (Supplementary Table).

We compared p-values for species in which IEs produce TSDs and those that do not. We found that IEs that produce TSDs are generally enriched in more lowly expressed genes (Figure 14F). Conversely, IEs that lack TSDs do not display any enrichment (Figure 14E). This result further highlights the presence of multiple mechanisms that may predict fitness effects. Any IE that reduces expression levels of genes in which it inserts should be at least marginally deleterious. We previously suggested that IEs that do not produce TSDs might operate through a reverse splicing mechanism. Since these IEs appear to have little effect on gene expression, we propose that they are far less deleterious than IEs that proliferate via DNA transposition.

Splicing aberrancies

(Bryan Thornlow, PhD candidate in the Corbett-Detig Lab, also contributed to this work)
In light of our observation that mechanisms of IE proliferation predict relative expression of IE containing genes, we conducted an analysis on IE splicing relative to other introns. To compare the splicing of IEs and non-IE introns, we used generalized linear mixed-effect models, including data from different SRA submissions as random effects. To determine whether IEs or non-IE introns more frequently used annotated splice sites, our model formulae were,

$$cbind(correctSplices, misSplices) \sim ie + depth + (1|replicate)$$

and

$$cbind(correctSplices, misSplices) \sim depth + (1|replicate),$$

where "correctSplices" refers to splicing events called by *leafcutter* corresponding to splice sites annotated by GenBank, "misSplices" refers to splicing events called by *leafcutter* in which an unannotated splice site within 50 nucleotides of the annotated splice site was used, and "ie" is a boolean variable, 1 for IEs and 0 for non-IE introns. "Depth" refers to the maximum depth within 50 nucleotides of the introner element on either side, as calculated by *samtools depth*. We recorded in Supplementary Table the Akaike information criterion (AIC) for both models and recorded the relative likelihood to ensure that the model including the IE identities performed better. We also recorded the estimate of the coefficient of the "ie" variable in the model along with its standard error. We call this variable a "splicing coefficient." A larger splicing coefficient suggests that IEs are more often spliced at their annotated splice site than are canonical introns. A smaller splicing coefficient signifies the opposite case.

We also compared rates of intron retention (in which introns are not spliced out, or "spliced in") between IEs and other introns. We performed a similar analysis as we did for splicing accurately, using the model formulae,

$$cbind(correctSplices + misSplices, splicedIn) \sim ie + depth + (1|replicate)$$

and

$$cbind(correctSplices + misSplices, splicedIn) \sim depth + (1|replicate),$$

where "splicedIn" refers to the minimum depth within the boundaries of the introner element. As before, we present recorded the AIC values, relative likelihood, coefficient estimates and standard errors in Supplementary Table. We call the estimate for the coefficient of the "ie" variable the "PSI Coefficient". A larger PSI Coefficient suggests that IEs are less often retained than are canonical introns. Conversely, a smaller PSI Coefficient suggests that IEs are more often retained.

We used the splicing coefficient to estimate how accurately IEs are spliced relative to other introns in each species. We found that IEs without TSDs are generally spliced more accurately

than other introns (Figure 14G). In contrast, IEs that produce TSDs are spliced less accurately (Figure 14H). We performed the same comparison using PSI coefficients and found that IEs without TSDs are also more often retained (Figure 14I). Again, we observed the opposite case in IEs with TSDs (Figure 14J). This observation agrees with our conjecture that at least 2 distinct mechanisms drive IE propagation, and substantiates our hypothesis that IEs that originate through reverse splicing are less deleterious. Splicing aberrancies and intron retention can result in reduced expression of important transcripts and the production of adverse isoforms.

3.5 Phylogenetic distribution of IEs

IEs are abundant, existing in approximately 2.5% of annotated eukaryotic genomes on the Genbank database. However, they do appear to be relatively more abundant in fungi, cnidarians, algae and protists than in other clades. Indeed, when I account for ascertainment bias caused by selective sampling, I observe a significant enrichment of IEs in protists (those that are not apicomplexans), green algae, cnidarians, and other bilaterians (those not classified under any other clades shown in Figure 18) (Figure 18, Fisher's Exact $\alpha = 0.05$). The fungal clades, ascomycota and basidiomycota also harbor many IE containing species relative to other clades. IEs are especially enriched in cnidarians: four out of the five total sampled species contain IEs. These observations motivated several downstream phylogenetically corrected analyses. I was curious as to why these particular clades contained IEs, and not others.

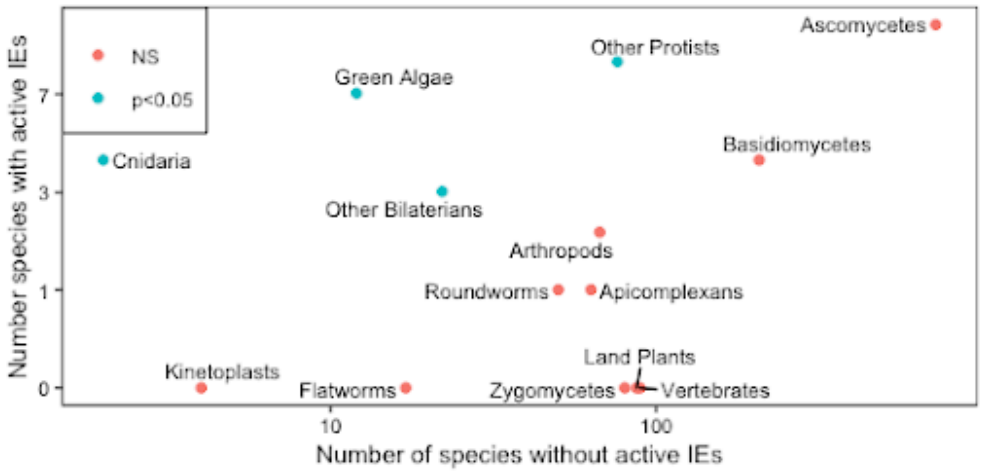

Figure 18: Scatter plot comparing the number of IE containing species relative to the number of species without IEs in each clade. Axes are logscaled. Treating each clade separately accounts for ascertainment bias caused by selective sampling. I observe enrichment of IEs in cnidarians, green algae, "other bilaterians" (those that do not fall under any other presented clades) and "other protists" (those that are not apicomplexans).

Constructing a phylogeny

(Bryan Thornlow, PhD candidate in the Corbett-Detig Lab, also contributed to this work)
To conduct phylogenetic correlations between traits related to IEs, we required a large phylogeny and trait data for each species. We used *OrthoDB* (54) to identify orthologous gene families (OGs)

for which all species containing introner elements had at least one orthologous gene. For these 401 OGs, we used *MAFFT* (44) to create amino acid sequence alignments, only including one gene per species for each OG, and only including species with a representative gene for at least 375 of these OGs. Following this, we concatenated the amino acid sequence alignments to produce one large alignment. To reduce run time, we removed uninformative sites in which at least 30% of species had gaps in the alignment. We then used *FastTree 2* (55) to construct an approximate maximum likelihood phylogeny based on the multiple alignment. We then rerooted the phylogeny using the clade containing all species belonging to the Excavata phylum (55) and pruned species without genome annotations, for which we could not conduct a search for IEs. To assess quality, we ensured that species belonging to the same organism group according to GenBank (e.g. Mammals, Ascomycetes, Land Plants) formed a clade within the phylogeny.

Horizontal transmission and germline accessibility might shape the distributions of IEs among species

(Russ's absolutely brilliant theory; our analysis)
After some simple categorizations, one shared attribute among green algae and protists (two clades that are enriched in IEs) was quickly revealed. Nearly all green algae and protists are single celled and therefore possess an accessible germline. We performed a literature search on the other IE containing species and found that the vast majority of them also possess accessible germlines. In light of this, we hypothesized that IEs infect new species via horizontal gene transfer (HGT). HGT occurs when genetic information passes between distantly related species. We searched for homology between IEs in different species in hopes of identifying recent HGT events but found none, which was unsurprising due to the degenerative nature of DNA. Genetic material passed through HGT does not transcend generations unless it integrates into a species' germline. Species that sequester their germlines are relatively immune to HGT since the chances of genetic material integrating with their germline are much lower (Figure 19). In contrast, single cell organisms and other organisms with accessible germlines are much more susceptible (Figure 19).

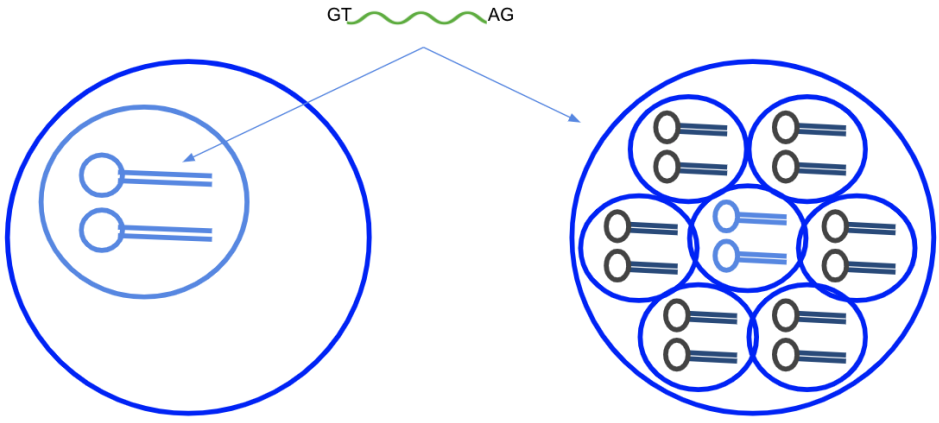

Figure 19: A cartoon demonstrating the difference between species that sequester their germline and species that do not. An organism with an accessible germline is pictured on the left and one with a sequestered germline is pictured on the right. An IE is represented by a green line with "GT" and "AG" splice sites. Species with sequestered germlines are less susceptible to HGT because it is much more likely that genetic material inserts into somatic DNA that does not transcend generations. In contrast species with an accessible germline are far more susceptible.

We used our phylogeny to test our hypothesis that species with accessible germline are more likely to gain IEs. We used the *binaryPGLMM()* command in the *ape* package through R (56), which estimates regression coefficients for binary phylogenetic data. We used the formula $ie \sim sq$ and found that accessibility of the germline is a significant predictor of the presence of IEs (Figure 20, $P < 0.005$). In Figure 20, we annotate our phylogeny for two binary traits: the presence of IEs (left) and the accessibility of the germline (right). For unicellular species, we consider the germline accessible even if the species does not explicitly sequester its germline. Overall, our results demonstrate that the presence of IEs is strongly correlated with the accessibility of an organism's germline, indicating that IEs likely move between species via horizontal gene transfer.

3.6 Transposition drives intron gain

IE containing species have more introns

IEs cause widespread intron gain on genomic scales. Thus, one would expect that IE containing species also have more introns. I initially compared the number of introns in species that have active IEs with those that do not and found that IE containing species have significantly more introns (Figure 22, MWU P = 3.43E-5, *no phylogenetic correction*). However, this P value is not phylogenetically corrected and could be influenced by ascertainment bias in my data set.

Therefore, I reassessed my hypothesis that the presence of IEs significantly increases the rate intron gain using my previously constructed phylogeny. As before, I pruned species whose assembly IDs in *OrthoDB* did not match any listed in our spreadsheet, but also further pruned species for which I could not reliably count the total number of introns. I used *RevBayes* (58) to fit models assessing the correlations between continuous characters (i.e. total number of introns) and the presence of IEs. I fit a Reversible-Jump Markov Chain Monte Carlo (rjMCMC) model, including the presence of IEs as a discrete character, and number of introns as a continuous character (55). This method attempts to fit one model wherein states depend on the discrete character and one in which states do not depend on the discrete character, and returns a p-value corresponding to the proportion of iterations in which the independent model was a better fit. My sample size for this analysis was 2,251 iterations, and all returned a better fit for the state-dependent model (Figure 22, P < 1E-3).

IE shape intron length distribution

I also hypothesized that IEs shape the distribution of intron lengths in genomes. I compared the mean IEs lengths with mean intron lengths for each species and found that they are significantly correlated (Figure 22, Spearman's rank correlation P < 4.82e-5). This correlation suggests that IEs do govern the distribution of intron lengths in genomes they inhabit to some extent.

Figure 20: A tree depicting our phylogeny. Nodes on the left are colored green for IE containing species. Nodes on the right are colored green for species with inaccessible germlines. Silhouettes exemplify different clades on the tree.

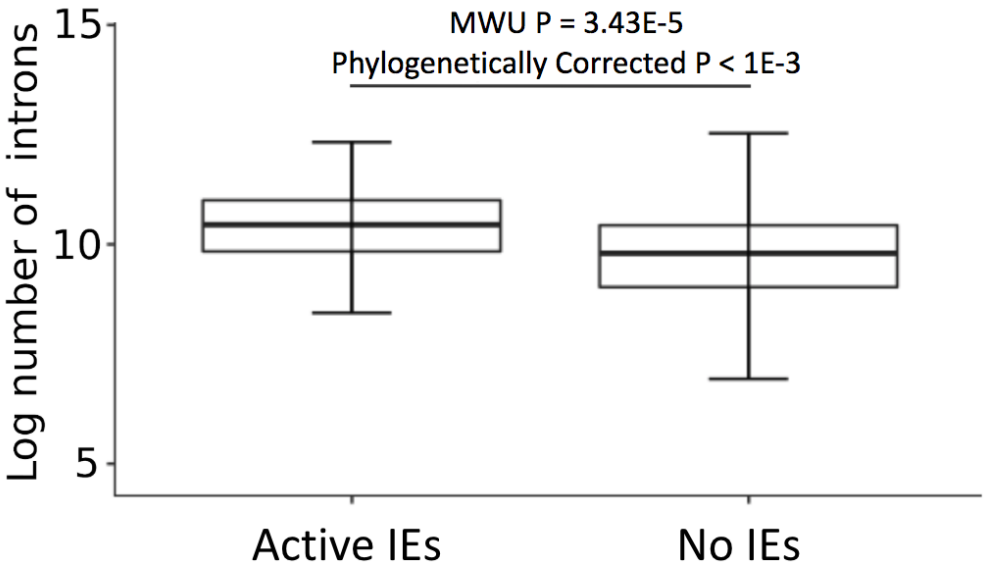

Figure 21: Box plots displaying distributions of the number of introns in species with active IEs and species without IEs. IE containing species have significantly more introns than species without IEs, suggesting that IEs are a fundamental driver of mass intron gain.

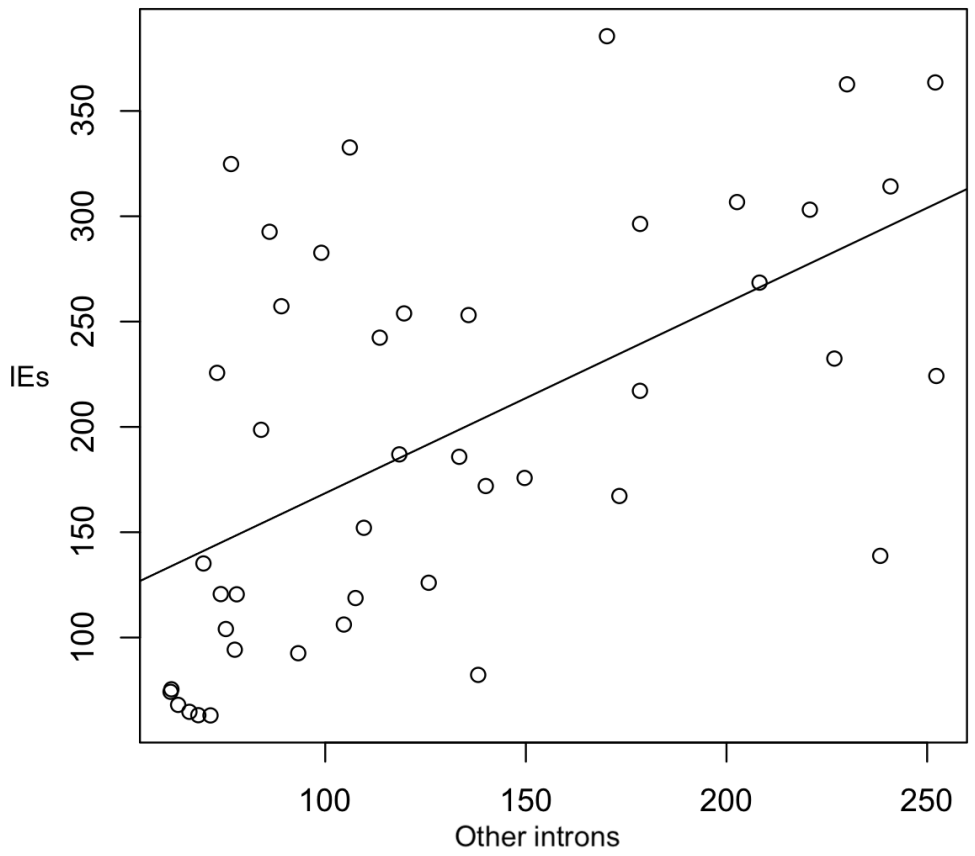

Figure 22: A scatter plot comparing mean intron lengths to mean IE lengths in each IE containing species. The diagonal line represents a linear regression modeling the relationship between mean IE length and mean intron length. This is what one would expect the line to look like in the case that the two variables are correlated. Indeed, they are correlated (Spearman's rank correlation P $<$ 4.82e-5).

4 Conclusion and future work

4.1 Conclusion

The most remarkable distinction of eukaryotic genes from prokaryotic ones is the presence of spliceosomal introns, which interrupt genes and are removed from RNA transcripts by a large RNA-protein machinery called the spliceosome. Comparative genomics has revealed that early eukaryotes underwent dramatic fracturing of their genes into many pieces through intron insertion, and that this process is ongoing in many eukaryotic lineages. In addition to fracturing eukaryotic genes, intron presence provides the basis for additional levels of transcriptional and post-transcriptional regulation, transcript surveillance, and, in bilaterian animals, widespread diversification of the proteome through alternative splicing. Introns are also thought to play important roles in genome structure evolution. However, their origins are poorly understood and the primary drivers of de novo intron creation remain elusive.

Transposition has been argued to be a major contributor of intron gain, based both on the possibility of a single element creating many introns and on its prediction of the observed concentration of intron gains in a subset of lineages. However, intron-creating transposons (collectively termed 'Introner Elements', or IEs) are known from only two lineages. Here, I performed a sensitive, systematic search for IEs across 1894 species for which annotated genomes exist in NCBI's Genbank database. I report 38 IE containing species spanning 11 phyla, suggesting that IEs are abundant in diverse eukaryotic lineages, existing in 2.5% of annotated genomes on Genbank.

My data reveal an unappreciated diversity of IE proliferation mechanisms, some of which are clade specific. The functional impacts of IE insertions predict evolutionary outcomes and suggest that many are deleterious, since they exist in lowly expressed genes and are poorly spliced relative to other introns. Moreover, I provide strong evidence that negative selection shapes the distribution of IEs within genomes. In addition, IEs are especially copious in unicellular species and multicellular species with accessible germlines, which are more susceptible to horizontal transfer of transposable elements. Although I did not find strong sequence homology among elements in different species, circumstantial evidence suggests germline accessibility is an important component of this process and the presence of IEs is strongly correlated by whether a species has an accessible germline. Thus, horizontal transmission and germline accessibility might shape the distributions of IEs among species. Furthermore, IE containing species comprehensively harbor more introns than species that lack IEs.

In light of my discoveries, I postulate that transposition functions as a principal driver of intron gain in diverse eukaryotes. I conclude by remarking that my data likely represent a conservative portion of IE containing species because my methods rely on sequence similarity resulting from recent transposition, and that IEs likely populate additional genomes or once did but have since been suppressed by selection.

4.2 Future work

Interrogating the functional and fitness effects of intron gain

Landen Gozashti[1,2], Preet Kaur[1], Sarah Cohen[3,4], Scott W. Roy[3] and Russell Corbett-Detig[1]

[1]University of California Santa Cruz Department of Biomolecular Engineering, UCSC Genomics Institute
[2]Harvard University Department of Organismic and Evolutionary Biology, Harvard University Museum of Comparative Zoology, Howard Hughes Medical Institute

[3]San Francisco State University Estuary and Ocean Science Center
[4]San Francisco State University Department of Biology

I previously demonstrated that transposition is likely a fundamental driver of intron gain in diverse eukaryotes. Nonetheless, important questions remain about the frequency, function and fitness effects of new intron insertions. The distribution of fitness effects is a key parameter in determining the course of evolution. Comparing individuals within a population that is currently experiencing intron gain would allow us to better understand its fitness effects and impact on gene expression. Understanding the frequency, fitness and functional impacts of intron gain would further illuminate its possible evolutionary origins and roles in adaptation.

Currently, I am working to infer the fitness effects of intron gain events in a particular population. To do this, I am comparing different isolates of a species that possess active IEs and probing for polymorphisms. I am using the pelagic tunicate, *Oikopleura dioica*, as a model for this. O. dioica exhibits an unprecedented number of introns at nonconserved positions as a result of recent intron loss and gain. I have identified IEs in *O. dioica* that exhibit highly similar sequence and length, indicative of recent transposition. In addition, O. dioica is relatively easy to collect in the San Francisco Bay. I collected several *O. dioica* individuals within the San Francisco Bay population with Dr. Sarah Cohen's assistance. Preet Kaur then extracted DNA from each individual and made sequencing libraries. We shipped off 4 of our libraries for sequencing and I am currently awaiting the raw data for each individual. I will align each genome to the reference transcriptome to identify polymorphic introns and infer the effects on fitness using population genetic theory.

I will also use tunicate population genetic data to interrogate the functional effects of intron gain. I previously provided robust evidence that intron gain events likely induce the translation of new isoforms and result in splicing aberrations. They also exist in more lowly expressed genes relative to other introns. However, introns are also known to bear regulatory elements, which can either reduce or increase gene expression and often enhance gene expression through unknown mechanisms merely by their presence. To further investigate how intron gain affects gene expression, I will sequence the transcriptomes of *O. dioica* individuals. By comparing expression patterns of genes with recent intron gains in different individuals in the same population, I will evaluate the functional consequences of intron gain for regulating transcription.

Horizontal transmission of IEs between a symbiont and its host

Landen Gozashti[1,2] and Russell Corbett-Detig[1]

[1]University of California Santa Cruz Department of Biomolecular Engineering, UCSC Genomics Institute
[2]Harvard University Department of Organismic and Evolutionary Biology, Harvard University Museum of Comparative Zoology, Howard Hughes Medical Institute

I previously provided evidence suggesting that IEs may move horizontally between species with accessible germlines. In light of this, I sought a model system in which I could investigate horizontal transfer of IEs between species in wild populations. I formerly demonstrated that IEs are extremely enriched in cnidarian species, existing in 80% of annotated cnidarian genomes on Genbank. Conveniently, IEs also exist in cnidarian endosymbionts within the genus, *Symbiodinium*. I postulate that close proximity may facilitate horizontal gene transfer (HGT) events between *Symbiodinium* and its host.

I sampled several individuals of an unknown sea anemone species off the coast of Davenport,

California. Then, I extracted DNA from each individual, made sequencing libraries, and sent them for sequencing. I am currently awaiting my sequencing data. Howbeit, the fact that my sampled species is *unknown* poses several problems, the foremost being the fact that both the anemone and its *Symbiodinium* symbiont lack reference genomes. Thus, I cannot align my raw genomic data and must either assemble each genome independently or devise an algorithm to detect putative IEs from raw genomic data.

We (Landen Gozashti and Russell Corbett-Detig) chose the former, due to the large financial expense of genome assembly. I am currently developing a pipeline for IE discovery from raw read data (Figure 23). I start by mining predicted transposable elements (TEs) from my raw data using publicly available tools. Next, I use k-mers to reconstruct TE flanking regions and align them to the reference genome of closely related species. Then I filter for TEs that appear to exist in genes. I simultaneously predict possible ORFs within TE flanking regions in an effort to recover TEs that exist in genes but might not align well to the respective reference. Finally, I filter TE families that are enriched in genes relative to others. Most TEs are relatively scarce in genes and primarily exist in noncoding regions since most genic insertions are highly deleterious and are thus filtered by selection. However, my prior data suggests that this does not hold true in the case of IEs, which are relatively abundant in genes. Thus, I shall consider any transposable element family that is enriched in genes relative to other TEs to be a putative IE family. I will run this pipeline on my raw WGS data once using a cnidarian reference and once using a *Symbiodinium reference*. I will then look for homology between putative IEs in each species, the presence of which would provide compelling evidence of recent HGT.

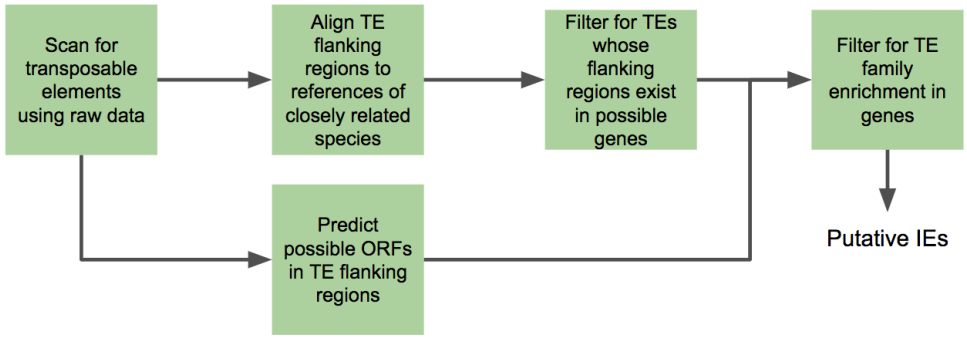

Figure 23: Flow chart depicting the steps in my pipeline for prospective IE detection from raw whole genome sequencing data. I begin by scanning for transposable elements in raw read data. Next, I align TE flanking regions to the reference of a closely related species while simultaneously predicting possible open reading frames (ORFs). Then, I filter for TEs that exist in possible genes revealing TE families that are highly enriched in genes.

References

1. M. P. Simmons, C. Bachy, S. Sudek, M. J. van Baren, L. Sudek, M. Ares, A. Z. Worden, Intron Invasions Trace Algal Speciation and Reveal Nearly Identical Arctic and Antarctic Micromonas Populations. Mol Biol Evol. 32, 2219–2235 (2015).

2. V. Vasil, M. Clancy, R. J. Ferl, I. K. Vasil, L. C. Hannah, Increased Gene Expression by the First Intron of Maize Shrunken-1 Locus in Grass Species 1. Plant Physiol. 91, 1575–1579 (1989).

3. D. C. Jeffares, T. Mourier, D. Penny, The biology of intron gain and loss. Trends in Genetics. 22, 16–22 (2006).

4. M. Chorev, L. Carmel, The Function of Introns. Front Genet. 3 (2012), doi:10.3389/fgene.2012.00055.
5. B.-S. Jo, S. S. Choi, Introns: The Functional Benefits of Introns in Genomes. Genomics & Informatics. 13, 112 (2015).

6. S. Roy, W. Gilbert, The evolution of spliceosomal introns: patterns, puzzles and progress. Nat. Rev. Genet. 7, 211-221 (2006), vol. 7.

7. P. Bhattachan, B. Dong, Origin and evolutionary implications of introns from analysis of cellulose synthase gene. Journal of Systematics and Evolution. 55, 142–148 (2017).

8. M. Irimia, S. W. Roy, Origin of Spliceosomal Introns and Alternative Splicing. Cold Spring Harb Perspect Biol. 6 (2014), doi:10.1101/cshperspect.a016071.

9. L. K. Derr, J. N. Strathern, D. J. Garfinkel, RNA-mediated recombination in S. cerevisiae. Cell. 67, 355–364 (1991).

10. S. W. Roy, W. Gilbert, Rates of intron loss and gain: Implications for early eukaryotic evolution. Proceedings of the National Academy of Sciences. 102, 5773–5778 (2005).

11. 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).

12. A. Z. Worden et al., Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes Micromonas. Science. 324, 268–272 (2009).

13. J. T. Huff, D. Zilberman, S. W. Roy, Mechanism for DNA transposons to generate introns on genomic scales. Nature. 538, 533-+ (2016).

14. L. Y. Geer, A. Marchler-Bauer, R. C. Geer, L. Han, J. He, S. He, C. Liu, W. Shi, S. H. Bryant, The NCBI BioSystems database. Nucleic Acids Res. 38, D492–D496 (2010).

15. C. E. Lane, K. van den Heuvel, C. Kozera, B. A. Curtis, B. J. Parsons, S. Bowman, J. M. Archibald, Nucleomorph genome of Hemiselmis andersenii reveals complete intron loss and compaction as a driver of protein structure and function. Proc. Natl. Acad. Sci. U.S.A. 104,

19908–19913 (2007).

16. M. C. Wahl, C. L. Will, R. Lührmann, The spliceosome: design principles of a dynamic RNP machine. Cell. 136, 701–718 (2009).

17. N. Lane, W. Martin, The energetics of genome complexity. Nature. 467, 929–934 (2010).

18. J. Singh, R. A. Padgett, Rates of in situ transcription and splicing in large human genes. Nat. Struct. Mol. Biol. 16, 1128–1133 (2009).

19. A. R. Buchman, P. Berg, Comparison of intron-dependent and intron-independent gene expression. Mol. Cell. Biol. 8, 4395–4405 (1988).

20. K. Juneau, M. Miranda, M. E. Hillenmeyer, C. Nislow, R. W. Davis, Introns regulate RNA and protein abundance in yeast. Genetics. 174, 511–518 (2006).

21. R. D. Palmiter, E. P. Sandgren, M. R. Avarbock, D. D. Allen, R. L. Brinster, Heterologous introns can enhance expression of transgenes in mice. Proc Natl Acad Sci U S A. 88, 478–482 (1991).

23. J. Beaulieu, J. Tank, S. Hamilton, W. Wollheim, R. Hall, P. J Mulholland, B. J Peterson, L. Ashkenas, L. Cooper, C. Dahm, W. Dodds, N. Grimm, S. Johnson, W. McDowell, G. C Poole, H. Valett, C. Arango, M. Bernot, A. Burgin, S. Thomas, Beaulieu et al. 2011 PNAS (2013).

24. F. Gaunitz, K. Heise, R. Gebhardt, A Silencer Element in the First Intron of the Glutamine Synthetase Gene Represses Induction by Glucocorticoids. Molecular endocrinology (Baltimore, Md.). 18, 63–9 (2004).

25. F. Gaunitz, D. Deichsel, K. Heise, M. Werth, U. Anderegg, R. Gebhardt, An intronic silencer element is responsible for specific zonal expression of glutamine synthetase in the rat liver. Hepatology. 41, 1225–1232 (2005).

26. G. Zhang, X. Li, H. Cao, H. Zhao, A. I. Geller, The vesicular glutamate transporter-1 upstream promoter and first intron each support glutamatergic-specific expression in rat postrhinal cortex. Brain Res. 1377, 1–12 (2011).

27. T. W. Nilsen, B. R. Graveley, Expansion of the eukaryotic proteome by alternative splicing. Nature. 463, 457–463 (2010).

28. B. Hartmann, J. Valcárcel, Decrypting the genome's alternative messages. Curr. Opin. Cell Biol. 21, 377–386 (2009).

29. W. Gilbert, Why genes in pieces? Nature. 271, 501 (1978).

30. S. Lee, S. W. Stevens, Spliceosomal intronogenesis. PNAS. 113, 6514–6519 (2016).

31. P. Yenerall, L. Zhou, Identifying the mechanisms of intron gain: progress and trends. Biology Direct. 7, 29 (2012).

32. G. R. Fink, Pseudogenes in yeast? Cell. 49, 5–6 (1987).

33. T. Mourier, D. C. Jeffares, Eukaryotic intron loss. Science. 300, 1393 (2003).

34. C.-K. Tseng, S.-C. Cheng, Both Catalytic Steps of Nuclear Pre-mRNA Splicing Are Reversible. Science. 320, 1782–1784 (2008).

35. 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. Current Biology. 22, 1260–1265 (2012).

36. Y. A. Chan, P. Hieter, P. C. Stirling, Mechanisms of genome instability induced by RNA-processing defects. Trends Genet. 30, 245–253 (2014).

37. A. Molaro, H. S. Malik, Hide and seek: how chromatin-based pathways silence retroelements in the mammalian germline. Curr Opin Genet Dev. 37, 51–58 (2016).

38. K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, E. W. Sayers, GenBank. Nucleic Acids Res. 44, D67–D72 (2016).

39. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).

40. A. Hagberg, D. Schult, P. Swart, Exploring Network Structure, Dynamics, and Function using NetworkX. Proceedings of the 7th Python in Science conference (SciPy 2008),11-15 (2008).

41. D. J. Finnegan, Transposable elements: How non-LTR retrotransposons do it. Current Biology. 7, R245–R248 (1997).

42. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066 (2002).

43. X. Pan, Y. Li, L. Stein, Site Preferences of Insertional Mutagenesis Agents in Arabidopsis. Plant Physiol. 137, 168–175 (2005).

44. L. Xi, Y. Fondufe-Mittendorf, L. Xia, J. Flatow, J. Widom, J.-P. Wang, Predicting nucleosome positioning using a duration Hidden Markov Model. BMC Bioinformatics. 11, 346 (2010).

45. E. R. Morris, H. Grey, G. McKenzie, A. C. Jones, J. M. Richardson, A bend, flip and trap mechanism for transposon integration. eLife. 5, e15537 (2016).

46. A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, G. J. Barton, Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics. 25, 1189–1191 (2009).

47. R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker, ViennaRNA Package 2.0. Algorithms Mol. Biol. 6, 26 (2011).

48. 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).

49. H. M. Robertson, Noncanonical GA and GG 5 Intron Donor Splice Sites Are Common in the Copepod Eurytemora affinis. G3: Genes, Genomes, Genetics. 7, 3967–3969 (2017).

50. A. Dobin, T. R. Gingeras, Mapping RNA-seq Reads with STAR. Current Protocols in Bioinformatics (2015), pp. 11.14.1–11.14.19.

51. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25 (2009), pp. 2078–2079.

52. Y. I. Li, D. A. Knowles, J. Humphrey, A. N. Barbeira, S. P. Dickinson, H. K. Im, J. K. Pritchard, Annotation-free quantification of RNA splicing using LeafCutter. Nat. Genet. 50, 151–158 (2018).

53. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting Linear Mixed-Effects Models Usinglme4. Journal of Statistical Software. 67 (2015), , doi:10.18637/jss.v067.i01.

54. E. V. Kriventseva, D. Kuznetsov, F. Tegenfeldt, M. Manni, R. Dias, F. A. Simão, E. M. Zdobnov, OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 47, D807–D811 (2019).

55. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE. 5 (2010), p. e9490.

56. T. A. Williams, Evolution: rooting the eukaryotic tree of life. Curr. Biol. 24 (2014), pp. R151–2.

57. E. Paradis, K. Schliep, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35, 526–528 (2019).

58. S. Höhna, M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, F. Ronquist, RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. Syst. Biol. 65, 726–736 (2016).