Finished product to be submitted to the journal of Molecular Biology and Evolution

Introner Elements Inhabit Algal and Protist Genomes

Landen Gozashti
Advised by Russell Corbett-Detig
University of California Santa Cruz Department of Biomolecular Engineering

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 (Simmons et al. 2015). Many species exhibit major intron loss events, which probably occur through RNA mediated homologous recombination of cDNA (Lee and Stevens 2016). 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 (Huff et al. 2016). Contrary to canonical introns, introner elements exhibit conserved sequences and lengths. Similar phenomena are known to exist in fungi (van der Burgt et al 2012; Wu et al. 2017). We developed a computational pipeline for introner element detection and employed it on 1113 algal and protist assemblies. We report novel IE discoveries in 6 species, suggesting that IE invasions are more widespread than previously thought. IE families in closely related species do not share sequence similarity, indicating that they may have evolved independently. Perhaps intron gain is a fundamental driver of genome structure evolution.

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 thank Evan Pepper for assisting me in fine tuning my molecular biology skills. I also thank Bryan Thornlow and Christopher Vollmers for assisting me in producing my figures. I leave my deepest 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.

Introduction

Introns are stretches of noncoding DNA found between exons (coding regions). During transcription, RNA polymerase produces pre-mRNA containing both introns and exons from template DNA. This pre-mRNA then undergoes a process known as splicing, catalyzed by the spliceosome. The spliceosome excises introns from the pre-mRNA transcript and ligates exons together to construct a mature mRNA transcript (Figure 1). Introns are a defining characteristic of eukaryotes, existing in all eukaryotic genomes with one possible exception (Lane et al. 2007).

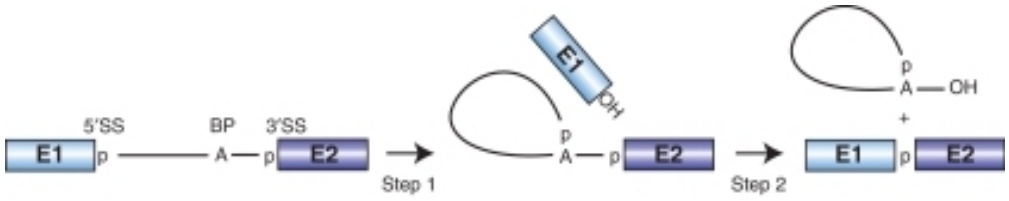

Figure 1: A schematic representation of the 2 step splicing mechanism. This process results in 2 products: a mature mRNA transcript containing E1 and E2, and an intron lariat. Source: Will et al. (2013). Cold Spring Harbor Perspectives in Biology 3

Despite their removal from mature mRNA transcripts, introns play several roles within eukaryotic organisms. Introns regulate gene expression levels (Vasil et al. 1989; Buchman and Berg 1988; Beaulieu et al. 2011; Gaunitz et al. 2004, 2005; Juneau et al. 2006; Palmiter et al. 1991; Zhang et al. 2011) . They also enable alternative splicing to take place, allowing many eukaryotic organisms to exhibit a proteome diversity far exceeding the number of genes in their genomes (Nilson and Graveley 2010). Moreover, introns play notable roles in regulation of nonsense mediated decay, translation yield, cytoplasmic localization, and nuclear export (Chorev and Carmel 2012; Jo and Choi 2015).

The origin of introns is still heavily debated. Two competing theories exist. Early intron theory states that ancestral prokaryotes possessed introns but lost them due to genomic streaming (Roy and Gilbert 2006; Bhattachan 2017). This theory implies that intron loss is the primary evolutionary force at hand. In contrast, late intron theory suggests that prokaryotes never harbored introns, and that eukaryotes evolved to acquire them, insinuating that intron gain is the principal force (Irimia and Roy 2014; Bhattachan 2017). Studies investigated intron loss and gain events in different species. Several found profound evidence of widespread intron loss in fungi (Derr et al. 1993; Roy and Gilbert 2005, 2006). Intron gain has been identified in a limited number of species and was considered until recently to be a relatively rare event (Verhelst et al. 2013; Simmons et al. 2015; Huff et al. 2016).

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 (Simmons et al. 2015; Worden et al. 2006; Verhelst et al. 2013). Contrary to canonical introns, these introns, deemed introner elements (IEs), exhibit conserved sequences and lengths (Verhelst et al. 2013). Massive introner element invasion caused the number of introns in strain CCMP1545 to double (Verhelst et al. 2013). A recent study also identified IEs in the distantly related pelagophyte alga, *Aureococcus anophagefferens* (Huff et al. 2016).

Although introners are known to exist, no study has proactively looked for them. We are profoundly interested in how intron gain events on large scales may affect the genomic landscape of

species in which they occur, as well as the mechanisms by which they propagate. IEs were previously thought to proliferate through a mechanism involving reverse splicing, since most insertions were observed in coding regions(Tseng and Chen 2008; Yenerall and Zhou 2012; Verhelst et al. 2013; Simmons et al. 2015). However, a recent study identifies the presence of target site duplications (TSDs) in a significant number of insertion events in both *Micromonas pusilla* and *Aureococcus anophagefferens*, a trademark of DNA transposition (Huff et al. 2016). Target site duplications result after the insertion of DNA transposons, when sticky ends created by staggered cuts in the target DNA are repaired.

In light of these discoveries, we hypothesize that IEs likely inhabit many more species, possibly those closely related to M. pusilla, if they were passed down from an ancestral species prior to divergence. Moreover, IEs may exist in more distantly related species as a result of either independent evolution or lateral gene transfer. Thus, we developed a pipeline for introner element detection from standardized genome assembly data available through NCBI. We ran this pipeline on 1113 algal and protist assemblies and discovered novel IEs in six species: *Volvox carteri f. nagariensis*, *Salpingoeca rosetta*, *Tetrahymena thermophila*, *Emiliania huxleyi*, *Aphanomyces astaci*, and *Micractinium conductrix*. Four of these species (*Volvox carteri f. nagariensis*, *Salpingoeca rosetta*, *Emiliania huxleyi*, and *Micractinium conductrix*) are relatively closely related to *M. pusilla* and each other. *Tetrahymena thermophila* and *Aphanomyces astaci* are much more distant relatives of all other considered species and each other.

Methods

We constructed a pipeline for introner element detection from any genome assembly available on NCBI, displayed in Figure 2 (Geer et al. 2010). It merely requires a genome fasta file and annotation file, which defines the positions of genetic elements in the genome, as input, and generates a variety of files as output. These include a fasta file containing all introns in the genome, a PAF file containing all vs all alignment results, a fasta file containing putative introner elements organized by family, a BLAST results text file, and a fasta file containing recovered introner elements produced using BLASTn results (Altschul et al. 1990).

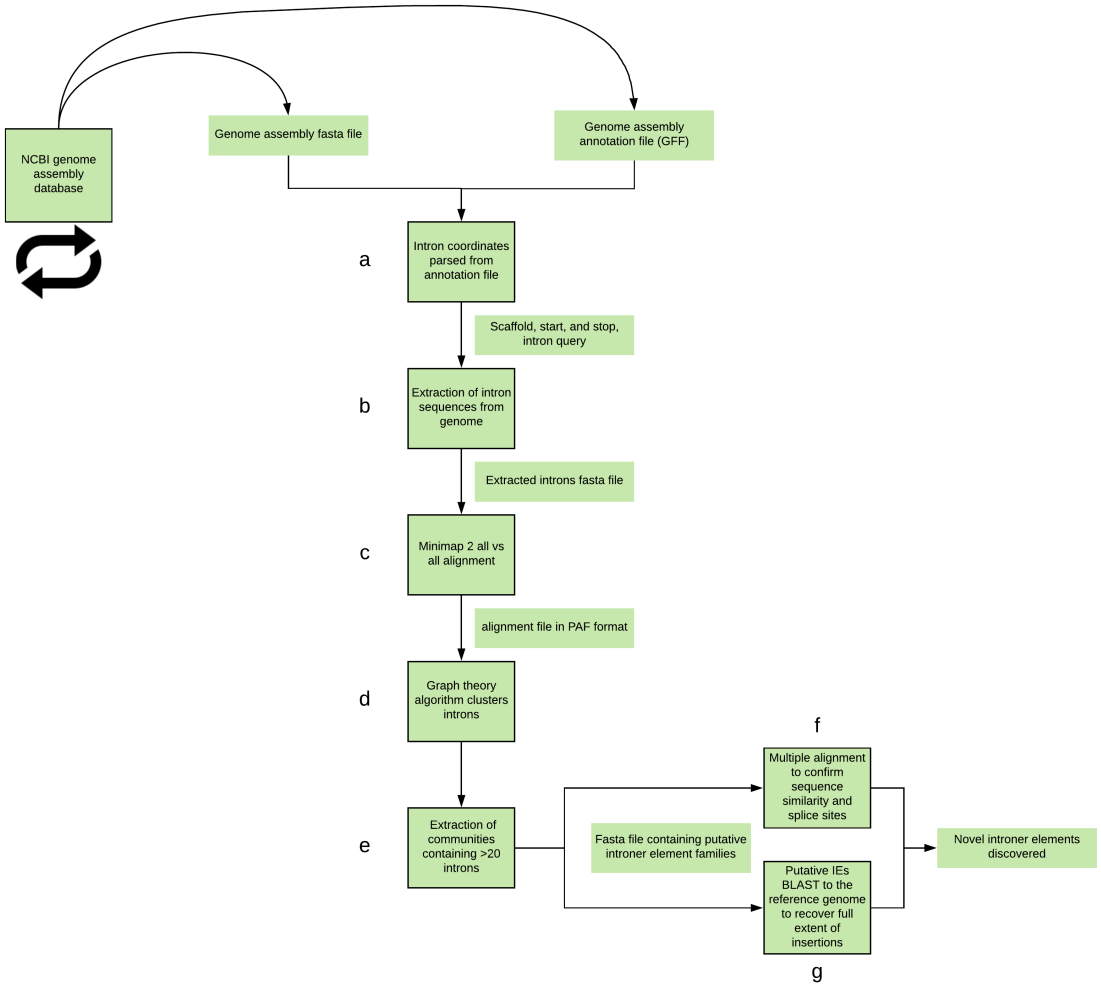

Figure 2: IE recognition pipeline, implemented on the standardized genbank database available through NCBI. (a) Introns coordinates are identified using a genome annotation file. (b) Intron sequences are extracted from the genome and deposited in a fasta file. (c) Minimap 2 performs an all vs. all alignment on extracted introns and outputs a Pairwise Mapping Format (PAF) alignment file. (d) A graph-theory-based algorithm clusters introns based on sequence similarity in which each intron is one node, and edges are quantified using similarity of each intron to all others in the genome. (e) Communities are formed from intron clusters using networkx, and families containing 20 individuals are identified as putative IEs. IE families are annotated and deposited in a fasta file (Hagberg et al. 2008). (f) A multiple alignment is performed using MAFFT, to confirm that putative IEs in each family exhibit sequence and length similarity, and to generate a consensus sequence for each family (Katoh et al. 2012). (g) IEs are BLASTed to the reference genome with a minimum percent identity requirement of 99%, to recover the full extent of insertions (Altschul et al. 1990).

Intron Extraction

We parsed exon coordinates from a genome annotation file and use them to identify intron coordinates (Figure 3). We then extracted intron sequences from the genome fasta file. Since all previously identified introner elements exhibit a length of less than 300bp, we filtered out all sequences longer than 300bp, and deposited the rest in a fasta file.

Figure 3: Since intron coordinates are not annotated in genbank genome annotation files, we must use exon coordinates to retrieve them. For each gene, the start of Intron 1 is simply 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

All vs. All Alignment and Clustering

We performed an all vs. all alignment on extracted introns using Minimap 2 and restricted output to overlaps with more than 99 residue matches. We used the PAF output file to generate a graph, in which introns are nodes positioned based on their sequence similarity to every other intron in the genome. We then identified clusters of introns and quantified introner elements families using a community cutoff of 20 individuals (Figure 4). After clustering, annotated introner element families and deposited them in a fasta file.

Multiple Alignment and Insertion Recovery

We performed a multiple alignment using all putative introners for each IE family to confirm sequence and length similarity and generated a consensus sequence for each family using MAFFT through Jalview (Waterhouse et al. 2008). We performed multiple alignments between consensus sequences of families from different species to detect intra species IE sequence similarity. We also BLASTed the consensus of each family to its corresponding reference genome to recover any insertions outside of coding regions (Altschul et al. 1990).

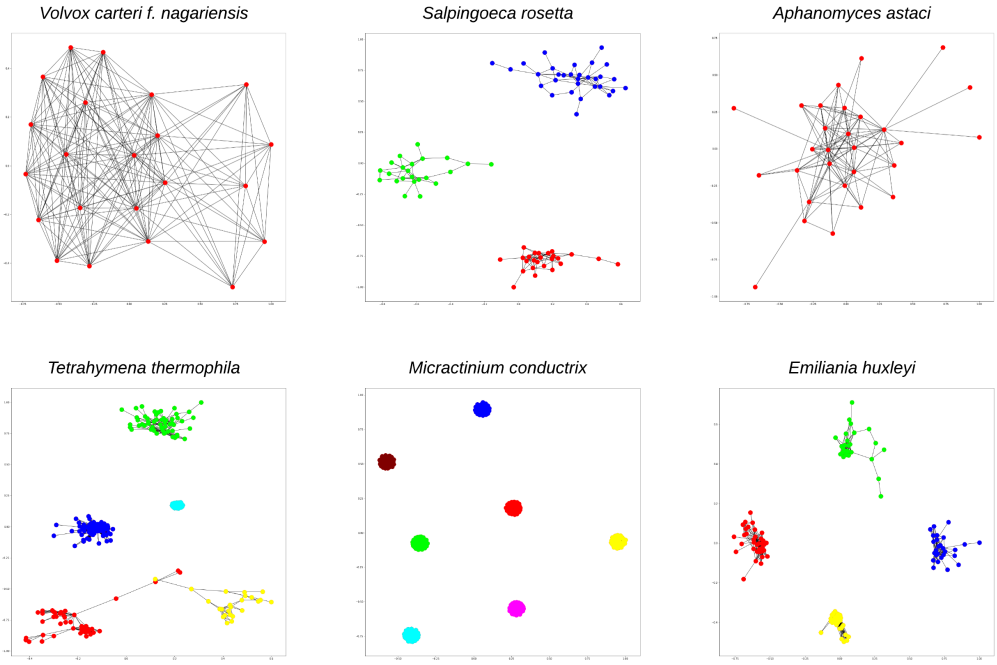

Figure 4: Here we display intron clusters for each of the species in which we identified novel introner elements. Each node represents a specific putative introner element. Edges connect introns that share over 99% identity. Introns with high sequence similarity form clusters which are divided into color coded families.

Results and Discussion

Phylogenetic Distribution

We found introner elements in six algal species, four of which are relatively closely related: *Volvox carteri f. nagariensis*, *Salpingoeca rosetta*, *Emiliania huxleyi*, and *Micractinium conductrix*. These species are also closely related to *Micromonas pusilla*, which exhibits perhaps the most notable instance of introner element invasion (Verhelst et al. 2013). *Tetrahymena thermophila* and *Aphanomyces astaci* are much more distant relatives both from the species mentioned above, and they are distant from each other (Figure 5).

Their phylogenetic relationships suggest that introner element families in *Volvox carteri f. nagariensis*, *Salpingoeca rosetta*, *Emiliania huxleyi*, *Micractinium conductrix*, and *Micromonas pusilla* may have evolved in an ancient ancestral species, although IE families from different species do not exhibit significant sequence similarity. Therefore, two mechanisms might explain the present phylogenetic distribution of IE's across these species. First, this could be the result of mutation over time, which is expected to erode a signal of shared ancestry after a sufficient number of mutations accumulate within active IE families. Alternatively, IE's might arise independently in each lineage. The fact that the host species appear reasonably closely related could then suggest that there is a mechanism of "pre-adaptation" that allows IE's to proliferate in the genome. For example, because these are non-autonomous transposable elements, if each host species expresses a transposase at relatively high abundances, this might facilitate the proliferation of IE's across their genomes. Further work will be necessary to determine if these elements are shared via common descent or independently evolved in each lineage.

IE families in Tetrahymena thermophila stand out the most when compared to families from other species. These families are all extremely "AT" rich and lack any degree of sequence similarity to those found in other species, indicating that they probably evolved independently, or possibly invaded the Tetrahymena thermophila genome from another species via in which we have yet to discover introner elements via horizontal gene transfer (Jain et al. 1999; Keeling Palmer 2008).

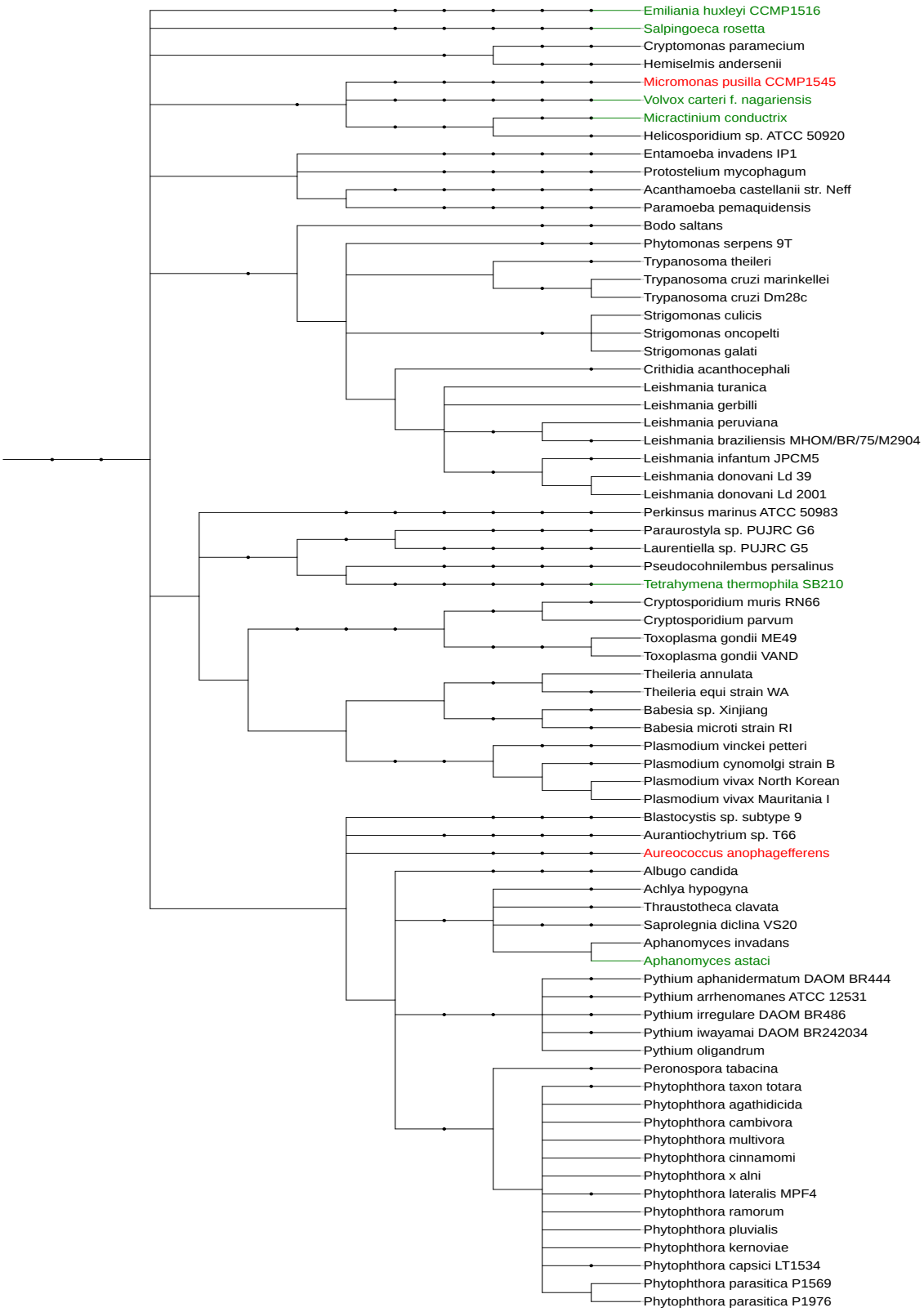

Figure 5: *Volvox carteri f. nagariensis*, *Salpingoeca rosetta*, *Emiliania huxleyi*, and *Micractinium conductrix* are all relatively closely related to *Micromonas pusilla*. Genomes in which we found novel introner elements are labeled in green. Positive controls are labeled in red.

Proliferation Mechanism

In accordance with Huff et al. (2016)'s observations in *Micromonas pusilla* and *Aureococcus anophagefferens*, a significant number of IE insertions in 3 of the 6 considered species exhibit target site duplication. The presence of TSDs is a strong indication of a proliferation mechanism involving DNA transposition (Gafner Philippsen 1980). We observe insertions outside of coding regions in all considered species. This phenomenon also supports DNA transposition as the primary proliferation mechanism for these elements. Any mechanism involving reverse splicing would result in almost every insertion existing in coding regions (Yenerall Zhou 2012; Verhelst et al. 2013; Simmons et al. 2015)

We performed a permutation test to assess whether we observe more TSD events in corresponding insertion coordinates than we would in the genome by chance. To do this, we selected 3bp sequences at random from each genome and compared them to see how many were the same n_i number of times where n_i is the number of insertions inside genes, and n_o number of times, where n_o is the number of insertions outside of genes. That is, our permutation test controls for the abundance of elements within and outside of genes. We then calculated the weighted average for the theoretical number of TSDs per insertion and compared it to our actual number of TSD events per insertion. We repeated this procedure 1000 times, and divided the number of times our theoretical value was greater than our actual value by the total number of samples to find a p value for each genome. We observe a significant excess of 3bp TSD events in IE insertions across all families of *Micractinium conductrix*, *Tetrahymena thermophila*, and *Volvox carteri f. Nagariensis* (Table 1, Figure 6).

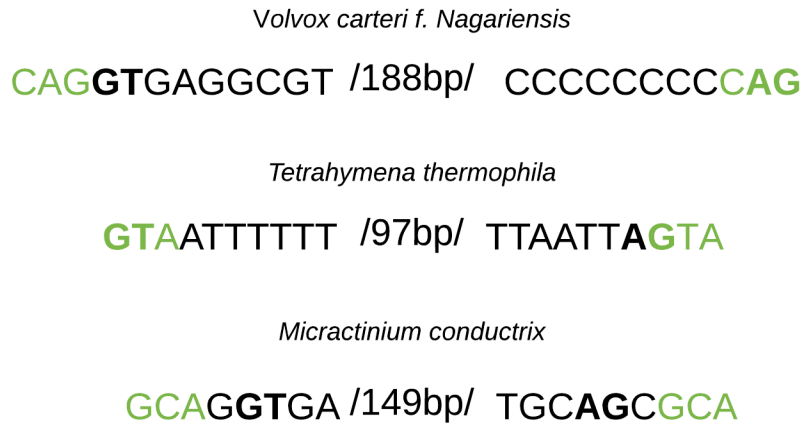

Figure 6: Here we demonstrate the locations of TSDs among introner element insertions in *Volvox carteri f. Nagariensis*, *Tetrahymena thermophila*, and *Micractinium conductrix*. We display an example introner element from family 1 of *Volvox carteri f. Nagariensis* and *Micractinium conductrix*, and family 5 of *Tetrahymena thermophila*. All sequences are shown 5' to 3'. TSDs are shown in green, and highly conserved GT/C 5' and AG 3' splice sites are shown in bold. We observe 3bp TSDs in among a significant number of IE insertions in all of these species.

In *Volvox carteri f. Nagariensis*, TSD sites contain the G of the 5' splice site and the 2 nucleotides adjacent to it in thee 5' direction, and the entire 3' splice site (Figure 6). Therefore,

Family	Insertions in Coding Regions	Insertions not in Coding Regions	TSDs per Insertion	P Value	TSD Length
Aphamomyces astaci					
1	29	44	31%	0.964	3bp
Emiliani huxleyi					
1	36	12	15.1%	0.992	3bp
2	26	10			
3	27	9			
4	30	1067			
Micractinium conductrix					
1	35	154	19.5%	<0.001	3bp
2	42	110			
3	31	85			
4	30	84			
5	36	92			
6	37	94			
7	35	93			
Volvox carteri f. nagariensis					
1	22	9	54.5%	0.005	3bp
Salpingoeca rosetta					
1	25	12	10.1%	0.984	3bp
2	23	13			
3	31	5			
Tetrahymena thermophila					
1	46	1908	16.2%	<0.001	3bp
2	54	*			
3	127	6			
4	24	*			
5	39	*			

Table 1: A summary of the number of insertions in coding regions, number of insertions outside of coding regions, TSDs per insertion, p value acquired from a permutation test, and TSD length for each IE family in *Aphanomyces astaci*, *Emiliania huxleyi*, *Micractinium conductrix*, *Volvox carteri f. nagariensis*, *Salpingoeca rosetta*, and *Tetrahymena thermophila*.

introner elements must co-opt their 3' splice site from their insertion site. A multiple alignment suggests that introners in *Volvox carteri f. Nagariensis* are more flexible. They have the ability to choose alternative splice sites, just as any intron does. We believe that they can also co-opt their 5' splice site after insertion, causing them to vary a bit more in sequence length and more noticeably, sequence similarity, just inside their 5' splice sites (Figure 7).

IEs in *Tetrahymena thermophila* exhibit an even more remarkable feature. Introners cause a frameshift mutation upon insertion within protein-coding sequences due to the two base pairs generated from the TSD just downstream of the 3' splice site (Figure 6). This likely accounts for why we observe so few insertions in coding region relative to noncoding regions. Frameshifts are extremely detrimental to genes. Nucleotides are translated in groups of three, referred to as codons. A frameshift occurs when a number of nucleotides that is not a multiple of three is inserted or deleted from a sequence, causing the entire frame of reference to shift. This almost always results in a completely different translation of the original gene (Roth 1974). The insertion of IEs inside coding regions in *Tetrahymena thermophila* is thus highly deleterious, and most insertions inside genes are probably filtered by selection. Also due to the nature of their TSDs, IEs in *Tetrahymena thermophila* must co-opt both their 5' and 3' splice site, further limittings their ability to insert inside of genes without extremely deleterious effects (Figure 6).

IEs in *Micractinium conductrix* carry both splice sites, presumably rendering them an advantage over the other IE families observed in this study (Figure 6). They are not restricted by the need to co-opt one splice site, or both, and do not cause frameshifts upon insertion. These likely account

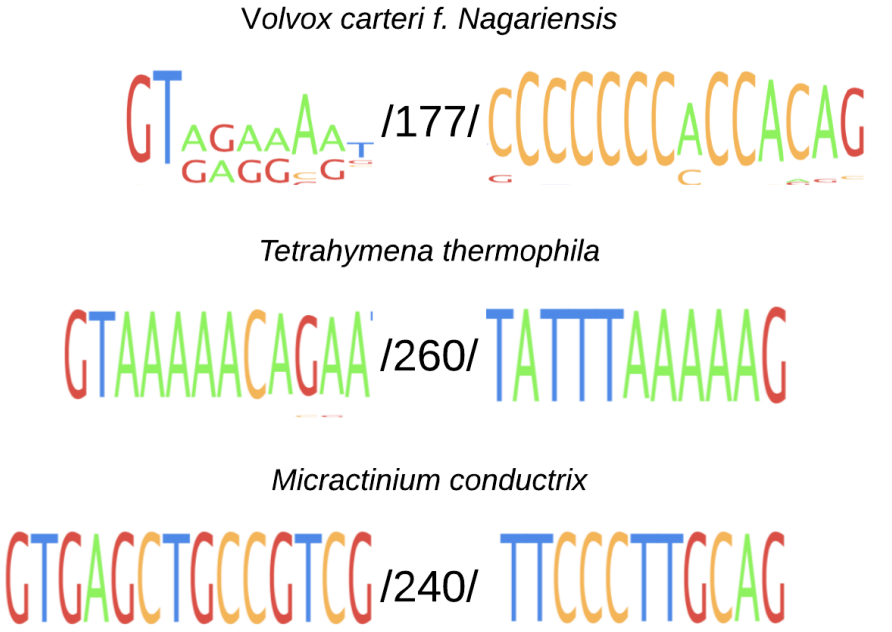

Figure 7: Consensuses for multiple alignments on all insertions for IE family 1 of *Volvox carteri f. Nagariensis*, *Tetrahymena thermophila*, and *Micractinium conductrix* produced using MAFFT through Jalview (Katoh et al. 2002; Waterhouse et al. 2008). The variation in sequence similarity just downstream of the 5' splice site in the consensus sequence for family 1 of *Volvox carteri f. Nagariensis* is peculiar. We expect this phenomenon if the 5' splice site is co-opted. However, the 5' TSD site lies outside of the 5' splice site in this species (Figure 6), indicating that IEs carry this splice site with them. Notice that introner element sequences in each family of each species are highly similar if not identical. Larger letters in the motif plot represent more highly conserved nucleotides at specific positions among all introners in a specific family. The size of each letter directly corresponds to the number of a specific nucleotides found at that position in the multiple alignment.

for why we observe such an abundance of IE families in *Micractinium conductrix*, as well as an abundance of insertions both inside and outside of coding regions.

TSDs degrade with time until unrecognizable due to mutation that accumulate either through neutral processes or due to selection to restabilize the protein sequence. This is likely why we do not ever observe perfect TSDs in all insertions of a given family. Older insertions should exhibit less definitive TSDs. It is possible, that this accounts for why we observe an insignificant amount of TSD events among insertions in *Emiliania huxleyi*, *Salpingoeca rosetta* and *Aphanomyces astaci*. It is also possible, however, that introner elements in *Emiliania huxleyi*, *Salpingoeca rosetta* and *Aphanomyces astaci* proliferate primarily through a mechanism other than DNA transposition; perhaps through reverse splicing, or double stranded break repair (Yenerall and Zhou 2012; van der Burgt et al. 2013; Verhelst et al. 2014; Simmons et al. 2015).

Effect on the Genomic Landscape

Introner elements express a profound impact on the genomic landscape of their hosts, causing a genome wide increase in total intron count, and in recent invasion events, large spikes at specific intron lengths. We can observe this phenomenon by comparing the distribution of intron lengths in genomes containing IEs with closely related ones that do not contain IEs (Figure 8). Intron counts at lengths between 100bp and 300bp in Micractinium conductrix are striking when compared to its close relative, Helicosporidium sp. All IE families in Micractinium conductrix exhibit lengths within these limits. Introner elements may have driven this excessive concentration of introns of relatively short lengths in the genome. We would expect this abundance, since IEs in Micractinium conductrix carry both splice sites and can thus insert more freely into any genomic region (Figure 6). Older invasion events likely evade our identification pipeline, since they lack sequence similarity as a result of mutation, which may be why we are unable to recover the massive number of introner elements harbored within the Micractinium conductrix genome. However, we observe in this case how their lengths remain relatively conserved (Figure 8).

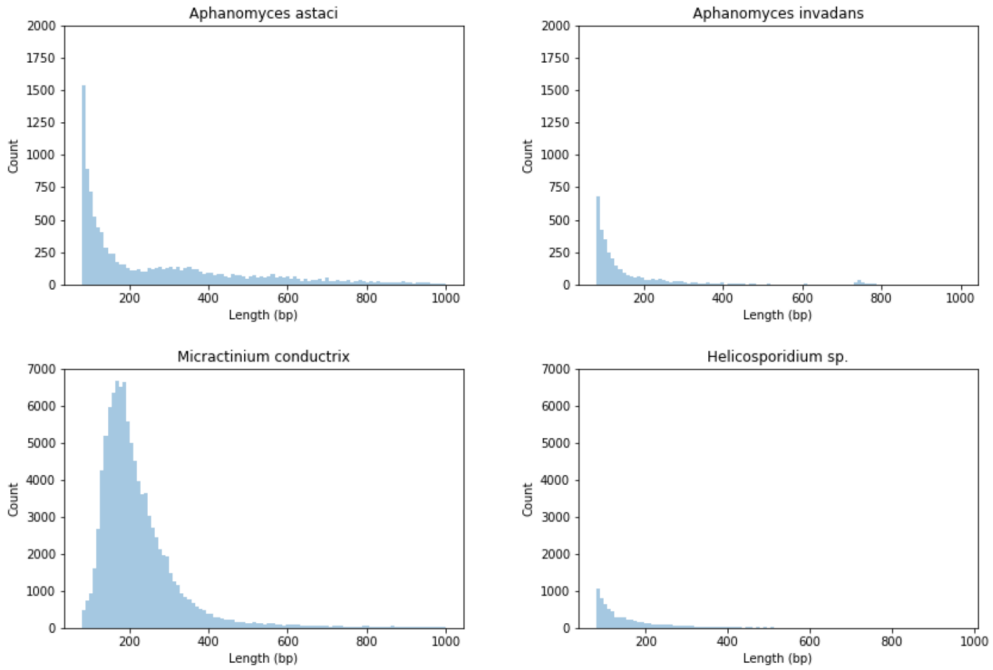

Figure 8: Distributions of the number of introns at lengths ranging from 80bp to 1000bp for *Aphanomyces astaci*, *Aphanomyces invadans*, *Micractinium conductrix*, and *Helicosporidium sp.* *Aphanomyces astaci* and *Aphanomyces invadans* are both taxonomically in the same genus, and are thus very closely related phylogenetically (Figure 4). *Aphanomyces astaci* contains active introner elements and *Aphanomyces invadans* does not. *Micractinium conductrix*, and *Helicosporidium sp.* both stem from the family, chlorellaceae, and are slightly more distantly related than *Aphanomyces astaci* and *Aphanomyces invadans* but still relatively closely related. The *Micractinium conductrix* genome contains introner elements and the *Helicosporidium sp.* genome does not.

When we compare the genomic landscape of the IE containing *Aphanomyces astaci* to the

IE lacking *Aphanomyces invadans*, we observe a much less extravagant difference. The number of introns in the *Aphanomyces astaci* genome nearly doubles that of *Aphanomyces invadans*, but we do not observe an acute peak at 200bp-300bp, the length range for introner elements in *Aphanomyces astaci* (Figure 8). We postulate that either *Aphanomyces astaci* acquired IEs relatively recently, or invasion events occur on a much smaller scale, possibly because they operate through a different mechanism than in *Micractinium conductrix*, although we have yet to test this.

Conclusion

Although introners are known from the genomes of a few representative clades, no study has previously conducted a systematic search for them more broadly. We developed a pipeline for novel IE detection using sequencing data available on NCBI and found introner elements in six algal species, four of which are relatively closely related: *Volvox carteri f. nagariensis*, *Salpingoeca rosetta*, *Emiliania huxleyi*, and *Micractinium conductrix*. However, we do not observe intra species sequence similarity when comparing IE families. Perhaps IEs in these species were passed down from an ancient ancestor and no longer exhibit sequence similarity due to mutation. On the contrary, these species could possess a mechanism of "pre-adaptation" that allows IEs to proliferate in their genomes.

A significant number of IE insertions resulting from all families of *Micractinium conductrix*, *Tetrahymena thermophila*, and *Volvox carteri f. Nagariensis* exhibit 3bp TSDs indicative proliferation mechanism involving DNA transposition. However, the position of TSD sites vary between species. IEs in *Tetrahymena thermophila* must co-opt their 3' splice sites, and IEs in *Volvox carteri f. Nagariensis* must co-opt both splice sites from their target sequence. IEs in *Tetrahymena thermophila* cause frameshifts upon insertion, limiting their abundance in coding regions. In contrast, IEs in *Micractinium conductrix* carry both splice sites.

What we observe from introner elements in *Aphanomyces astaci*, *Emiliania huxleyi*, *Micractinium conductrix*, *Volvox carteri f. nagariensis*, *Salpingoeca rosetta*, and *Tetrahymena thermophila* and what others have observed in *M. pusilla* and *A. anophagefferens* provide significant support for the argument that introns are under positive selection (Yenerall and Zhou 2012; van der Burgt et al. 2013; Verhelst et al. 2014; Simmons et al. 2015; Huff et al. 2016). Introners are transposons that possess a niche which allows them to insert into genes without being as strongly affected by selection as unspliced tranpsonsons. They persist in genomes even where they must co-opt splice sites or harm genes in which they insert, and possess the ability to exist in noncoding regions in between invasion events if they are filtered out of genes by selection.

Perhaps introners that must co-opt splice sites or cause frameshifts upon insertion have yet to reach their full potential. Indeed, we observe in *Micractinium conductrix*, how introners that carry their own splice sites and have relatively little effect on the translation of genes in which they insert can affect the distribution of introns in a genome. Maybe we perceive less of an effect in the other genomes we consider because IEs in these genomes have yet to master the art of intron gain. Introners are more widespread than previously thought and use a diverse array of splicing mechanisms. Therefore, our work opens up a range of opportunities to study the mechanistic basis and consequences of transposition.

References

Altschul, Stephen F., Gish, Warren, Miller, Webb, Myers, Eugene W., and Lipman, David J. (1990). Basic local alignment search tool. J. Mol. Biol. 215; 403-410

Beaulieu, J., Tank, J., Hamilton, S., Wollheim, W., Hall, R., J Mulholland, P., J Peterson, B., Ashkenas, L., Cooper, L., Dahm, C., et al. (2011). Nitrous oxide emission from denitrication in stream and river networks. Proc. Natl. Acad. Sci. U.S.A. 108, 214-219.

Bhattachan, P., and Dong, B. (2017). Origin and evolutionary implications of introns from analysis of cellulose synthase gene: Cellulose synthase intron shows eukaryotic invention. Journal of Systematics and Evolution 55, 142–148.

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

Chorev, M., and Carmel, L. (2012). The Function of Introns. Front Genet 3.

Derr, L.K., and J.N. Strathern. (1993). A Role for Reverse Transcripts in Gene Conversion. Nature 361, 170–173.

Gafner, J., and Philippsen, P. (1980). The yeast transposon Ty1 generates duplications of target DNA on insertion. Nature 286, 414–418.

Gaunitz, F., Heise, K., and Gebhardt, R. (2004). A Silencer Element in the First Intron of the Glutamine Synthetase Gene Represses Induction by Glucocorticoids. Molecular Endocrinology (Baltimore, Md.) 18, 63–69.

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

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

Hagberg, A.A., Schult, D.A., and Swart, P.J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. 5.

Huff, J.T., Zilberman, D., and Roy, S.W. (2016). Mechanism for DNA transposons to generate introns on genomic scales. Nature 538, 533-536.

Irimia, M., and Roy, S.W. (2014). Origin of Spliceosomal Introns and Alternative Splicing. Cold Spring Harb Perspect Biol 6.

Jain, R., Rivera, M.C., and Lake, J.A. (1999). Horizontal gene transfer among genomes: The complexity hypothesis. PNAS 96, 3801–3806.

Jo, B.-S., and Choi, S.S. (2015). Introns: The Functional Benefits of Introns in Genomes. Genomics Informatics 13, 112.

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

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

Keeling, P.J., and Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics 9, 605–618.

Lane, C.E., van den Heuvel, K., Kozera, C., Curtis, B.A., Parsons, B.J., Bowman, S., and Archibald, J.M. (2007). 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.

Lee, S., and Stevens, S.W. (2016). Spliceosomal intronogenesis. Proc. Natl. Acad. Sci. U.S.A. 113, 6514–6519.

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

Roth, J.R. (1974). Frameshift Mutations. Annual Review of Genetics 8, 319–346.
Roy, S.W., and Gilbert, W. (2005). Complex early genes. Proc Natl Acad Sci U S A 102, 1986–1991.

Roy, S.W. and Gilbert, W. (2006). The Evolution of Spliceosomal Introns: Patterns, Puzzles and Progress. Nat. Rev. Genet. 7, 211-221.

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

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

van der Burgt, A., Severing, E., de Wit, P.J.G.M., and Collemare, J. (2012). Birth of New Spliceosomal Introns in Fungi by Multiplication of Introner-like Elements. Current Biology 22, 1260–1265.

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

Verhelst, B., Van de Peer, Y., and Rouzé, P. (2013). The Complex Intron Landscape and Massive Intron Invasion in a Picoeukaryote Provides Insights into Intron Evolution. Genome Biol Evol 5, 2393–2401.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.

Worden A.Z., Lee J.H., Mock T., Rouzé P., Simmons M.P., Aerts A.L., Allen A.E., Cuvelier M.L., Derelle E., Everett M.V. (2009) Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes Micromonas. Science. 324, 268-272.

Wu, B., Macielog, A.I., and Hao, W. (2017). Origin and Spread of Spliceosomal Introns: Insights from the Fungal Clade Zymoseptoria. Genome Biology and Evolution 9, 2658–2667. Yenerall, P., and Zhou, L. (2012). Identifying the mechanisms of intron gain: progress and trends. Biology Direct 7, 29.

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