Introner Elements drive ongoing intron gain in *Oikopleura dioica*

Preet Kaur and Landen Gozasthi
Advised by Dr. Russell Corbett-Detig
Department of Biomolecular Engineering and Bioinformatics
University of California, Santa Cruz

April 2, 2020

Abstract

Eukaryotic genes are characterized by the presence of spliceosomal introns, which interrupt genes and are removed from mRNA transcripts by a complex protein-RNA machinery called the spliceosome. Spliceosomal introns perform various functions and play a critical role in genome structure evolution, but are unequally distributed across species due to gain and loss events[3]. Until recently, intron gain was considered to be a relatively rare event, and the fundamental drivers of intron gain are unclear. Nonetheless, transposition has long been argued to be a primary driver of intron gain based on both the possibility of a single element creating many introns and the observed concentration of intron gains in a subset of lineages, and an increasing number of studies demonstrate that several species experienced recent intron gain likely through a mechanism involving transposition[11, 19, 10]. A recent study from our group illustrated that transposable introns, known as introner elements (or IEs) are pervasive across diverse eukaryotic species and may act as the principal drivers for intron gain[10]. Here, we implemented molecular and computational methods to investigate how IEs can generate variation in intron positions between individuals in the same species, using the pelagic tunicate, *Oikopleura dioica*, as our model. We sequenced an *O. dioica* isolate from the San Francisco Bay and detected 27 novel candidate IEs in its genome. In addition, we found widespread variation in intron positions between Genbank's reference isolate and our own, suggesting that intron gain is ongoing in *O. dioica*. We also demonstrate that IEs in *O. dioica* were inherited ancestrally as opposed to acquisition via horizontal gene transfer (HGT). We report evidence of heterozygous intron positions in our isolate, suggesting that introns are polymorphic within the San Francisco Bay *O. dioica* population. Overall, our work prefigures population genetic studies that will explore the functional impacts and fitness effects of intron gain, and provides an imperative model with which they can be conducted.

Introduction

Introns are nucleotide sequences within protein coding regions in DNA that are removed via RNA splicing after transcription and before translation. During transcription, RNA polymerase produces a pre-mRNA transcript which contains introns and protein coding exons. The pre-mRNA undergoes further processing in which a large RNA-protein machinery called the spliceosome removes introns and ligates exons to construct a mature mRNA transcript for protein expression. Despite their excision from protein coding transcripts, introns play important roles in the regulation of gene expression, nonsense-mediated decay, translation yield, cytoplasmic localization, nuclear export and in bilaterian animals, the widespread diversification of the proteome through alternative splicing[4]. Although the importance of introns is well understood, their origins are still debated[12]. Comparative genomics has revealed that early eukaryotic genes were broken into many pieces through intron insertion, and that this process is ongoing in many lineages[14]. Transposition has been argued to be the major contributor to intron gain, based on the possibility of a single element creating many introns, but examples of widespread intron gain events remain relatively scarce[24].

DNA transposons are selfish mobile genetic elements. Transposons are capable of duplicating themselves in large quantities and account for a large proportion of genomic material in many species. The marine alga, *Micromonas pusilla*, is most notable for novel intron gain events[18, 22, 20]. Transposable introns, called introner elements (or IEs), invaded its genome in astounding quantities, causing the number of introns to double[22, 19]. IEs are unique from other introns in that they exhibit similar sequences and lengths[24, 11]. A recent study from our group demonstrated that IEs are widespread across diverse eukaryotes and may function as fundamental drivers of widespread intron gain[10].

Here we exercise molecular and computational methods to investigate how IEs can generate polymorphic intron positions in a population. The candidate we chose for our study is the pelagic appendicularian tunicate *Oikopleura Dioica*. Tunicates are the closest living relatives of vertebrates[8].They fall under the chordata phylum, indicating that they posses a notochord, a dorsal nerve cord, pharyngeal slits, an endostyle, and post-anal tail.

Most intron positions are highly conserved across higher eukaryotes[7]. However, studies have shown that *O. dioica* experienced recent intron gain and that a remarkable 73% of its introns exist at nonconserved positions. Many introns in the species exhibit exceptionally high rates of sequence similarity among them, implicating recent and perhaps-ongoing transposition in their origins.[9, 8]. In addition, our lab previously identified active IEs in *O. dioica* using a systematic IE detection pipeline[10]. The confident identification of a species with ongoing intron gain would present a powerful resource for understanding the proximate evolutionary and molecular causes of intron gain via transposition. *O. dioica* is therefore a particularly appealing candidate species for further study of intron gain and transposition.

To determine if intron gain via transposition is ongoing in this species, we collected an *O. dioica* individual in the San Francisco Bay and sequenced its genome. We identified candidate IEs in our sequenced isolate that lack homology to those we identified in the geographically divergent reference isolate, suggesting that IEs are a property shared with the ancestor of these populations. Moreover, we showed in our previous work that the distantly related ascidian tunicate, C*iona intestinalis*, does not harbor IEs, suggesting that IEs either invaded following the divergence between acidean and appendicularian tunicates or evolved beforehand but were filtered by negative selection. Additionally, we identify widespread variation between intron positions in our sequenced isolate and Genbank's reference, indicating that transposition presently drives intron gain in higher eukaryotes. We also

observe evidence of heterozygous intron positions within our sequenced *O. dioica* isolate, suggesting that intron positions are polymorphic within the San Francisco Bay population.

Methods

Collection of *O. dioica*

With our collaborator, Dr. Sarah Cohen, who is an expert on tunicates and zooplankton, we collected water from the San Francisco Bay using a zooplankton net. We dropped a net, from the boat, to a depth of four meters and pulled in samples which were then quickly transferred to large jars for sorting. Sorting entailed isolating *O. dioica* under the microscope. Once the organism was located, it was transferred to a microcentrifuge tube pre-loaded with $100\mu L$ of milliQ water. The samples were then flash frozen using dry ice and ethanol because liquid nitrogen was unavailable at the time.

Lysis of *O. dioica*

An isolate was thawed and subsequently added to $40\mu L$ of Lysis Buffer in a microcentrifuge tube. To the same tube, $3\mu L$ of Proteinase K was added. The tube was then placed on the heat block at 55°C for 3 hours. Then the sample was vortexed and used for DNA purification via SPRI beads.

Tn5 library prep for NGS

Next generation sequencing requires library prep of DNA samples to sequence on flowcells. It entails fragmentation of genomic DNA followed by ligations of adapters and index primers. It begins by adding equal amounts of oligos into two separate microcentrifuge tubes where one tube contains oligo A + oligo R and the other contains oligo B + oligo R. The oligos are annealed in their separate tubes at 95°C for 5 minutes. Next, Tn5 transposase enzyme is added to each tube containing annealed primers. After a 1-hour incubation at room temperature, the volumes of the two tubes are combined and mixed.

From there, we proceed with tagmentation of the genomic DNA. We began by adding $5\mu L$ of the mixed Tn5-oligo, $4\mu L$ of TAPS-PEG buffer, $1\mu L$ of H2O and $10\mu L$ of genomic DNA for a total volume of $20\mu L$. This mixture is then set to 55°C for 10 min. After the 10 mins, $5\mu L$ of 0.2% SDS is added to denature the Tn5 to prevent overtagmentation of the DNA.

The tagmented DNA is now ready for PCR amplification. We used $11.75\mu L$ of H2O, $5\mu L$ of 5x KAPA HiFi buffer, $0.75\mu L$ of dNTPs, $0.5\mu L$ of HiFi polymerase, $5\mu L$ of tagemented DNA, $1\mu L$ of i5 index and $1\mu L$ of i7 index for a total reaction volume. The reaction then runs with the following parameters: 1 .72°C for 5 min, 2. 95°C for 3 min, 3. 98°C for 20 sec, 4. 65°C for 15 sec, 5. 72°C for 30 sec, 6. Repeat 11x from step 3 , and 7. 72°C for 5 min for a final extension phase.

Sequence alignment of *O. dioica* isolates to its reference genome

The DNA libraries were sequenced at Fulgent Genetics using their Illumina HiSeq 4000 which uses Illumina's sequence by synthesis technology to produce paired-end data. After receiving the data, we began by trimming the adapters from the reads. We then aligned the trimmed reads to the annotated *O. dioica* reference genome, GCA_000209555.1, using the Burrows-Wheeler alignment tool. We indexed the reference genome with the command:

```
$ bwa index -a bwtsw GCA_000209555.1_ASM20955v1_genomic.fna
```

This was followed by aligning our paired-end sequencing results, formatted as FASTQ files, to the indexed reference genome to create BAM files:

```
$ bwa mem GCA_000209535.1_ASM20955v1_genomic.fna Oik_1A_S3_R1_001.fastq.gz /
Oik_1A_S3_R2_001.fastq.gz > OikA.bam
```

Then we removed the PCR duplicates from the data using the Samtools rmdup tool with the following commands:

```
$ samtools rmdup OikA.bam FinOikA.bam
```

To find the percentage of aligned reads, we divided the number of aligned reads by half of the number of paired-end reads and multiplied by one hundred. We had 11.43% aligned reads thus indicating our isolate was divergent from the reference or a relatively large contribution of contaminating DNA perhaps derived from the microbial communities associated with the mucosal structures that O. dioica produces. Nonetheless, owing to the relatively small size of the reference genome, we obtained sufficient data for our planned analyses.

Recovering NGS data and finding structural variants

We used the bioinformatic tool, *Pilon*, to process the paired-end sequencing data to fix single nucleotide polymorphisms[21]. These reads were then realigned to the genome using BWA. Subsequently, we used *Pindel* to find structural variants, specifically insertions and deletions, across our reads[23].

Filtering for variants supported by multiple PCR reactions

We used a custom python script to filter our *Pindel* output for deletions and insertions supported by reads from multiple PCR reactions. By doing so, we ensured that predicted variants were not artefacts produced by a single PCR reaction. PCR preferentially amplifies short sequences, causing them to rise to high frequency in a genomic library. Such sequences can deceivingly generate signatures of inaccurate structural variants. Filtering for insertions and deletions which are supported by multiple PCR reactions vastly reduces the probability of this phenomenon.

Identifying candidate IEs

Filtering for insertions in genes

Like other transposons, IEs have been shown to insert in both coding and noncoding regions[11]. However, IEs often occur at high frequencies in genes due to their ability to be spliced after insertion, which limits their effects on gene expression[11]. Thus, we filtered our data for insertions that occur in genes with respect to Genbank's GCA_000209535.1 reference genome.

Clustering

IEs that have recently transposed exhibit strong sequence and length similarity. In light of this, we implemented a pipeline previously developed by our group to cluster insertions in genes based on sequence and length similarity[10]. The pipeline implements *all-vs-all* BLAST to generate pairwise

overlaps between insertions, then employs a Girvan-Newman algorithm to cluster insertions based on homogeneity and associate putative introner element families.

Recognizing candidate intron deletions relative to the reference

To identify candidate intron deletions in our isolate (or intron insertions in the reference), we first filtered for deletions that exist in genes with respect to Genbank's GCA_000209535.1 reference. Then, we used a custom script to pinpoint deletions in our isolate that spanned or contained full length introns with respect to the reference, accomodating for a standard error of 10 base pairs (bp).

Detecting evidence of heterozygosity

We used our *Pindel* output in combination with *samtools tview* and the *Integrative Genomics Viewer* (IGV) to visually identify possible cases of heterozygous intron positions in our *O. dioica* isolate. We searched for cases in which a fraction of reads mapped to predicted deletions. We also probed for cases in which insertions were supported by a small fraction of reads relative to the coverage of surrounding regions.

Results and Discussion

Candidate novel IEs in a sequenced *O. dioica* isolate

We sequenced an *O. dioica* individual from the San Francisco bay population and identified novel candidate introner element sequences in its genome. Although some prior studies have discovered novel IEs using annotated genome assembly data[11, 22, 10], ours is the first to identify candidate IE sequences using raw genomic data from a single individual. Candidate introners in our isolate share strong sequence and length similarity, which we expect to observe in IEs that have recently transposed[22, 11, 18]. In addition, sequences end with a canonical "AG" 3' splice site and are recognized as insertions when aligned to Genbank's GCA_000209555.1 *O. dioica* reference genome(Figure 1). Previous studies have shown that IEs in many species carry one splice site, and

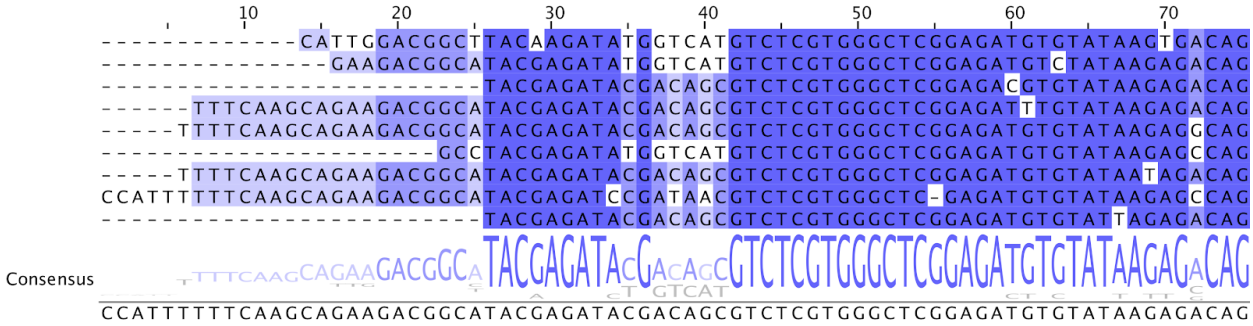
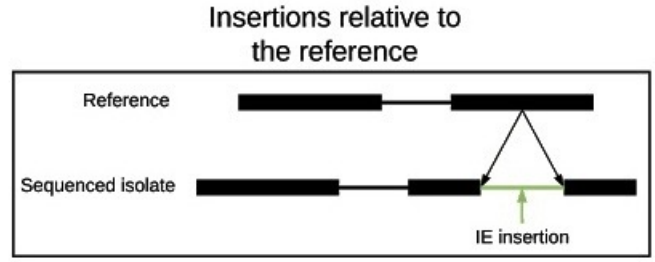


Figure 1: Candidate introner elements from our isolate aligned to previously identified introner elements in the *O. dioica* reference. Note that the candidate introners end with the "AG" 3' splice site and share similarities in sequence and length.

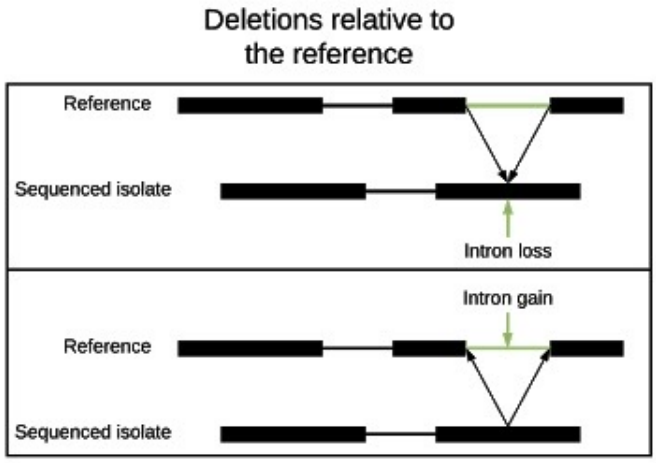
co-opt the other from their insertion site, which likely limits their effects on fitness and promotes their persistence in a population[11]. Most transposons are highly deleterious, since they can impair essential genes upon insertion. However, IEs which carry or co-opt splice sites are likely much less harmful, since they are mostly spliced out during transcription, and their effects on gene expression depend predominantly on their proliferation mechanism[10].

O. dioica is experiencing ongoing intron gain

Recent intron loss and gain has occurred in a diverse subset of eukaryotic lineages[16, 15]. Nevertheless, intron positions are generally conserved across higher eukaryotic genomes, and intron gain in metazoans is considered to be rare[2, 1]. Coincidingly, no study has reported widespread variation in intron positions between individuals from the same metazoan species.



(a)

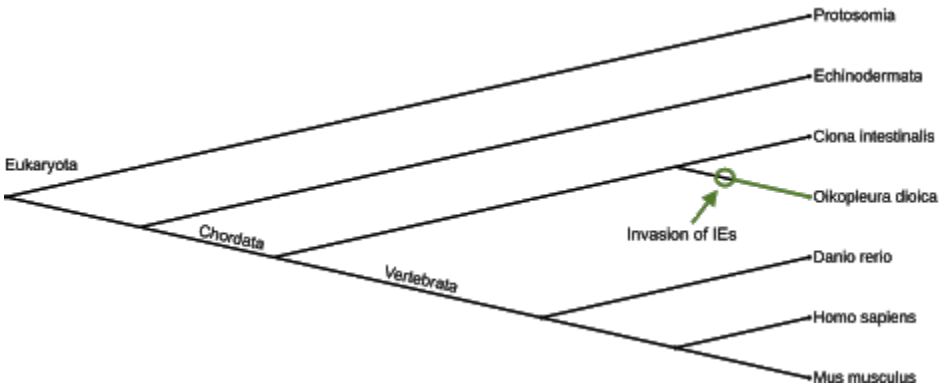


(b)

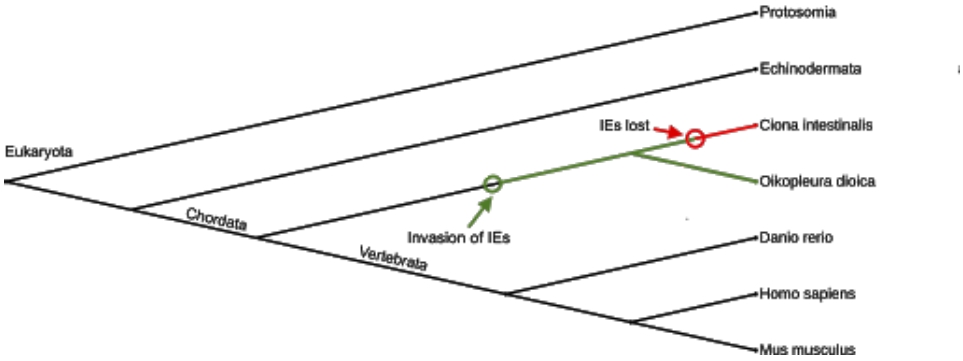
Figure 2: Subfigure 2a provides a visualization of an introner element insertion of a sequenced isolate. When compared to the reference, these specific introns do not exist at those insertion sites. This could be attributed to intron loss in the isolate or intron gain in the reference, as depicted by subfigure 2b.

Previous works have suggested that *O. dioica* may accentuate unexplored variation amidst metazoan genomes, partly because it appears to have experienced recent intron gain, possessing an unprecedented proportion of introns at non-conserved positions[8]. We observe 27 insertions within gene-coding regions in our sequenced *O. dioica* isolate which we attribute to IEs, due to the

relatively low probability that several insertions in genes would exhibit highly similar sequences and the architecture required for splicing by chance. Moreover, introns do not exist at putative IE insertion positions in Genbank's GCA_000209555.1 reference, suggesting intron gain in our isolate(Figure 2a). In addition, we observe 29 positions in our sequenced *O. dioica* isolate at which introns appear to be deleted with respect to Genbank's reference. These observations could result from either intron loss in our isolate or intron gain in the reference (Figure 2b). Regardless, such cases contribute to the total variation of intron positions between two individuals from the same metazoan species, further highlighting the underappreciated plasticity of metazoan genomes.



(a)



(b)

Figure 3: Subfigure 3a depicts the hypothesis suggesting introner elements invaded appendicularian genomes after the divergence between appendicularians and acideans. Subfigure 2b highlights the possibility of the invasion of introner elements prior to the divergence between appendicularians and acideans and the loss of IEs due to negative selection in acideans.

Inroner Elements in *O. dioica* were inherited ancestrally

A study from our group recently identified several IE families in Genbank's GCA_000209555.1 *O. dioica* reference genome[10, 17]. The same study also suggests that IEs move between species and individuals within species via horizontal gene transfer (HGT), since IE containing species are characterized by germline accessibility. However, the GCA_000209555.1 reference isolate was sampled from the fjords of Norway, approximately 5,100 miles from our sampling site. Since the probability

of HGT decreases with decreased proximity, such geographic severance limits the possibility that IEs evolved independently in one *O. dioica* population and invaded the other via HGT[6]. Therefore, we propose that IEs in modern *O. dioica* populations were inherited from a common ancestor. Furthermore, we demonstrated in our previous work that the phylogenetically divergent ascidian tunicate, *Ciona intestinalis*, does not harbor introner elements. Thus, we hypothesize that IEs either invaded ancestral appendicularian genomes following the divergence between acideans and appendicularians or evolved beforehand but were filtered by negative selection (Figure 3). Indeed, a delicate equilibrium exists between a transposon and fitness cost to its host[13]. It is entirely possible that IEs existed in all tunicates but became too deleterious, and were thus filtered from the vast majority of species by selection.

Evidence of polymorphic intron positions within the San Francisco Bay *O. dioica* population

O. dioica are diploid, meaning that each individual possesses two complete sets of chromosomes, one from each parent. Thus, heterozygous intron positions in an individual suggest that one of its parents possessed introns that the other parent did not, indicating that intron positions are polymorphic within its respective population (Figure 4). We observe evidence of such cases within our San Francisco Bay *O. dioica* isolate.

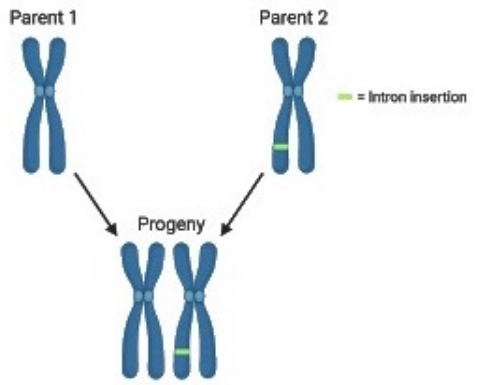


Figure 4: Diploid organisms inherit one chromosme from each parent, resulting in a complete set of two chromosomes. The presence of heterozygous intron positions is an indication that at least one parent possessed introns while the other parent did not. The illustration above depicts this polymorphic heterozygosity of a diploid organism.

In an ideal scenario, in which we assume constant read coverage across both alleles, we expect to observe approximately half of that coverage at a given polymorphic intron position. Moreover, under such circumstances, a homozygous insertion should exhibit relatively the same coverage as its flanking regions, and a homozygous deletion should exhibit a coverage of zero. Our data is far from ideal. However, we observe putative IE insertion sites at which very few reads map relative to the coverage of surrounding regions. We also observe cases in which reads map to predicted intron deletions in our isolate with respect to the reference (Figure 5). Both of these phenomena provide

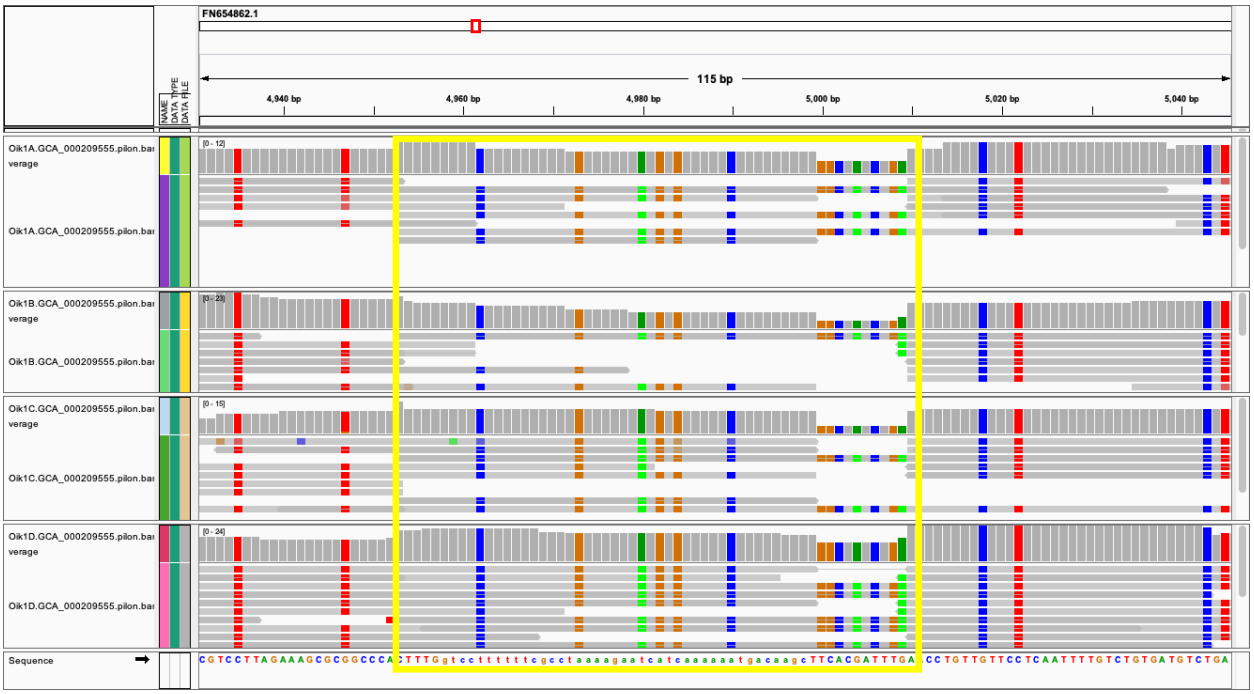


Figure 5: Using the *Integrated Genome Viewer*, we aligned our reads to the reference *O. dioica* genome. Using the deletion coordinates we received from running *Pindel* on our data, we located a deletion. This is the range of bases depicted within the yellow box. Here we see that only a fraction of our reads mapped to this range which suggests a possible heterozygous intron position.

compelling evidence of heterozygous intron positions within our *O. dioica* isolate and subsequently suggest that intron positions are polymorphic within the San Francisco Bay population.

Conclusion

Introner elements drive widespread intron gain in diverse eukaryotes[5, 10, 11, 19]. In our previous work, we employed a systematic IE detection pipeline on all annotated genomes in NCBI's Genbank database and identified multiple IE families in the appendicularian tunicate, *O. dioica*[10, 17]. Here we sequenced another *O. dioica* isolate and discovered 27 novel candidate IEs that are non-homologous to those found in Genbank's reference. Prospective IEs in our isolate exist in genes at nonconserved positions with respect to the reference, exhibit strong sequence and length similarity, and harbor components required for splicing, suggesting that they recently transposed and generated new introns as a result. We also observe 29 cases in which entire introns appear as deletions in our isolate with respect to Genbank's reference, indicating either intron loss in our isolate or intron gain in the reference.

Our previous work demonstrates that IE containing species are characterized by accessible germlines, suggesting that IEs invade new species and individuals via horizontal gene transfer (HGT)[10]. However, circumstantial evidence suggests that HGT did not occur between our isolate and the reference, and we propose that IEs in O. dioica were inherited ancestrally. In addition, we previously demonstrated that IEs do not exist in the divergent ascidian tunicate, *Ciona in-*

testinalis[10]. In light of this, we hypothesize that IEs either invaded after the divergence between ascidian and appendicularian tunicates or evolved beforehand but were filtered by negative selection.

Finally, we provide compelling evidence of heterozygous intron positions in our isolate, indicating that introns are polymorphic within the San Francisco Bay *O. dioica* population. Our data reveal that intron gain is ongoing in *O. dioica*, illuminating the plasticity of metazoan genomes. In addition, our study is the first to unearth widespread variation in intron positions between two individuals of the same metazoan species. It is also the first to identify evidence of polymorphic intron positions within a population. Nevertheless, important queries remain regarding the frequency, function and fitness effects of new intron insertions[10]. Here, we recognize *O. dioica* as a quintessential model for studying intron evolution. However, further analyses require population data. By comparing *O. dioica* individuals within the same population, we can better understand the functional impacts of intron gain, which may illuminate the evolutionary origins of introns and their possible roles in adaptation. Thus, our work opens up a range of opportunities to interrogate the functional and fitness effects of intron gain.

Acknowledgments

We thank Dr. Sarah Cohen at the San Francisco State University Department of Biology for her expertise on tunicate species and for helping us collect and isolate our samples. We thank Evan Pepper for his expertise on molecular biology techniques, specifically those pertaining to NGS. Finally, we extend our sincerest gratitude to Dr. Russell Corbett-Detig of the University of California, Santa Cruz Biomolecular Engineering and Bioinformatics Department and the University of California, Santa Cruz Genomics Institute. He lended us his extensive knowledge and skills within the field of computational biology and genomics in turn contributing to our continued passion and success in research.

References

[1] R. Assis, A. S. Kondrashov, E. V. Koonin, and F. A. Kondrashov. Nested genes and increasing organizational complexity of metazoan genomes. *Trends in genetics*, 24(10):475–478, 2008.

[2] L. Carmel, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. Patterns of intron gain and conservation in eukaryotic genes. *BMC evolutionary biology*, 7(1):192, 2007.

[3] F. Catania and M. Lynch. Where do introns come from? *PLoS biology*, 6(11), 2008.

[4] C. Cenik, H. N. Chua, H. Zhang, S. P. Tarnawsky, A. Akef, A. Derti, M. Tasan, M. J. Moore, A. F. Palazzo, and F. P. Roth. Genome analysis reveals interplay between 5 utr introns and nuclear mrna export for secretory and mitochondrial genes. *PLoS genetics*, 7(4), 2011.

[5] J. Collemare, H. G. Beenen, P. W. Crous, P. J. de Wit, and A. van der Burgt. Novel intronerlike elements in fungi are involved in parallel gains of spliceosomal introns. *PLoS One*, 10(6), 2015.

[6] V. Daubin and G. J. Szöllősi. Horizontal gene transfer and the history of life. *Cold Spring Harbor Perspectives in Biology*, 8(4):a018036, 2016.

[7] A. D. De Roos. Conserved intron positions in ancient protein modules. *Biology direct*, 2(1):7, 2007.

[8] F. Denoeud, S. Henriet, S. Mungpakdee, J.-M. Aury, C. Da Silva, H. Brinkmann, J. Mikhaleva, L. C. Olsen, C. Jubin, C. Cañestro, et al. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science*, 330(6009):1381–1385, 2010.

[9] R. B. Edvardsen, E. Lerat, A. D. Maeland, M. Flåt, R. Tewari, M. F. Jensen, H. Lehrach, R. Reinhardt, H.-C. Seo, and D. Chourrout. Hypervariable and highly divergent intron–exon organizations in the chordate oikopleura dioica. *Journal of molecular evolution*, 59(4):448–457, 2004.

[10] L. Gozashti, B. Thornlow, M. Ares Jr., and R. Corbett-Detig. De novo creation of spliceosomal introns by different transposition mechanisms in diverse eukaryotes. in prep.

[11] J. T. Huff, D. Zilberman, and S. W. Roy. Mechanism for dna transposons to generate introns on genomic scales. *Nature*, 538(7626):533–536, 2016.

[12] M. Irimia and S. W. Roy. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harbor perspectives in biology*, 6(6):a016071, 2014.

[13] E. S. Kelleher, R. B. Azevedo, and Y. Zheng. The evolution of small-rna-mediated silencing of an invading transposable element. *Genome biology and evolution*, 10(11):3038–3057, 2018.

[14] I. B. Rogozin, L. Carmel, M. Csuros, and E. V. Koonin. Origin and evolution of spliceosomal introns. *Biology direct*, 7(1):11, 2012.

[15] I. B. Rogozin, Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, and E. V. Koonin. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Current Biology*, 13(17):1512–1517, 2003.

[16] S. W. Roy and W. Gilbert. Rates of intron loss and gain: implications for early eukaryotic evolution. *Proceedings of the National Academy of Sciences*, 102(16):5773–5778, 2005.

[17] E. W. Sayers, M. Cavanaugh, K. Clark, J. Ostell, K. D. Pruitt, and I. Karsch-Mizrachi. Genbank. *Nucleic acids research*, 47(D1):D94–D99, 2019.

[18] M. P. Simmons, C. Bachy, S. Sudek, M. J. Van Baren, L. Sudek, M. Ares Jr, and A. Z. Worden. Intron invasions trace algal speciation and reveal nearly identical arctic and antarctic micromonas populations. *Molecular Biology and Evolution*, 32(9):2219–2235, 2015.

[19] A. van der Burgt, E. Severing, P. J. de Wit, and J. Collemare. Birth of new spliceosomal introns in fungi by multiplication of introner-like elements. *Current Biology*, 22(13):1260–1265, 2012.

[20] B. Verhelst, Y. Van de Peer, and P. Rouze. The complex intron landscape and massive intron invasion in a picoeukaryote provides insights into intron evolution. *Genome biology and evolution*, 5(12):2393–2401, 2013.

[21] B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*, 9(11), 2014.

[22] A. Z. Worden and F. Not. Ecology and diversity of picoeukaryotes. *Microbial ecology of the Oceans*, 2:159–205, 2008.

[23] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, 2009.

[24] P. Yenerall and L. Zhou. Identifying the mechanisms of intron gain: progress and trends. *Biology direct*, 7(1):29, 2012.