Mechanism for DNA transposons to generate introns on genomic scales

Jason T. Huff1,2, Daniel Zilberman1 & Scott W. Roy3

The discovery of introns four decades ago was one of the most unexpected findings in molecular biology1. Introns are sequences interrupting genes that must be removed as part of messenger RNA production. Genome sequencing projects have shown that most eukaryotic genes contain at least one intron, and frequently many2,3. Comparison of these genomes reveals a history of long evolutionary periods during which few introns were gained, punctuated by episodes of rapid, extensive gain4–8. However, although several detailed mechanisms for such episodic intron generation have been proposed4–8, none has been empirically supported on a genomic scale. Here we show how short, non-autonomous DNA transposons independently generated hundreds to thousands of introns in the prasinophyte Micromonas pusilla and the pelagophyte Aureococcus anophagefferens. Each transposon carries one splice site. The other splice site is co-opted from the gene sequence that is duplicated upon transposon insertion, allowing perfect splicing out of the RNA. The distributions of sequences that can be co-opted are biased with respect to codons, and phasing of transposon-generated introns is similarly biased. These transposons insert between pre-existing nucleosomes, so that multiple nearby insertions generate nucleosome-sized intervening segments. Thus, transposon insertion and sequence co-option may explain the intron phase biases2 and prevalence of nucleosome-sized exons9 observed in eukaryotes. Overall, the two independent examples of proliferating elements illustrate a general DNA transposon mechanism that can plausibly account for episodes of rapid, extensive intron gain during eukaryotic evolution2,3.

We began by examining the clearest case of recent, pervasive intron gain, which is in the prasinophyte alga M. pusilla10. This genome gained thousands of spliceosomal introns with highly similar sequences (named ‘introner’ elements10) by an unresolved mechanism. These introner elements have distinctive lengths and sequences compared to other spliceosomal introns in the genome10,11. We first surveyed the 3,347 RNA-sequencing-validated introner elements that we identified (Supplementary Data 1) using previously generated nucleotide-resolution genomic chromatin maps12. Unexpectedly, we found that most introner elements aligned with nucleosomes and contained one nucleosome each (Fig. 1a, b), with 73% of introner element ends located in nucleosome linker DNA, which is specifically marked by high cytosine methylation in M. pusilla12. Other introns do not show appreciable alignment to nucleosomes (Fig. 1a, b), and the number of introner element ends in linkers is significantly higher than for other intron ends (42%; P < 2.2 × 10−16). One possible explanation for this finding is that introner elements insert into the linker DNA between pre-existing nucleosomes. To assess this possibility, we used the fact that eukaryotic genes generally have phased nucleosome arrays emanating from their starts13, thus providing information about nucleosome positions before introner element insertion. We found that introner element locations aligned with linker DNA in phase with the starts of genes (Fig. 1c), revealing that introner elements do insert into pre-existing linkers between nucleosomes. Identification of introner elements in the 5′ portions of genes, which lack DNA methylation (Fig. 1c), suggests that introner elements insert into nucleosome linkers per se, rather than specifically into regions of methylated DNA. Consistent with this hypothesis, introner elements in unmethylated regions still exhibited alignment with nucleosomes (Extended Data Fig. 1).

Figure 1 | M. pusilla introner elements insert between pre-existing nucleosomes. a, Each introner element (IE) contains a nucleosome with ends in linker DNA, which is specifically marked by methylation in this organism. Validated introns and chromatin data12 are displayed. HEME1 contains two introner elements (green). b, Introner element introns are generally in phase with nucleosome positions, whereas other introns are not. Chromatin maps12 are aligned to 5′ introner element intron ends (dark lines) or other intron ends (light lines). c, Introner elements are in phase with the starts of genes, indicating insertion between pre-existing nucleosomes. Chromatin maps12 and 5′ introner element ends are aligned to gene starts. A kernel density estimate of introner element ends is shown with peaks marked.

1Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA. 2California Institute for Quantitative Biosciences, University of California, Berkeley, California 94720, USA. 3Department of Biology, San Francisco State University, San Francisco, California 94132, USA.
To identify potential examples of mechanistically similar intron gains, we searched for introns with the distinctive lengths (Fig. 2a) and correspondence to nucleosome positions (Fig. 1) of *M. pusilla* intron elements in other species, for which we previously generated validated intron junctions and chromatin maps. We found hundreds of introns with these characteristics in the very distantly related pelagophyte alga *A. anophagefferens*. An unbiased search in *A. anophagefferens* for the hallmarks of intron elements—sequence similarity between introns that also do not have similarity in neighbouring exons (Fig. 2b)—identified 602 candidate intron elements (11% of all RNA-seq-validated introns; Supplementary Data 2).

Like *M. pusilla* intron elements, most *A. anophagefferens* intron elements have distinctive lengths (Fig. 2a). Phylogenetic analysis of these intron elements suggests two large related groups (Fig. 2c), which correspond to the two peak sizes (Fig. 2a). *A. anophagefferens* intron elements also generally contain one positioned nucleosome in phase with the start of the gene (Extended Data Fig. 2), consistent with insertion into pre-existing nucleosome linkers. There is no appreciable sequence similarity in intron elements between the two organisms, suggesting that they evolved independently.

Within each genome, intron elements are characterized by sequence similarity between introns, suggesting that they have spread from one site to another. It has been proposed that intron elements (and other introns) spread through an RNA intermediate involving reverse splicing, resembling the mobility mechanisms of group II intron elements. However, we found that intron element insertion sites exhibit directly duplicated sequences in both genomes (target site duplications, TSDs), which are not expected to result from reverse splicing. These TSDs have characteristic lengths (8 bp in *A. anophagefferens* and 3 bp in *M. pusilla*; see Extended Data Fig. 3 and Supplementary Discussion), but the sequence of each tends to be particular to that intron element insertion site (Fig. 3a). Such TSDs indicate that insertion into double-stranded DNA followed by repair of staggered single-stranded regions has caused direct duplication of a short sequence that differs for each element. Within the intron elements and immediately flanked by the TSDs, inverted repeats are also observed (Fig. 3a and Extended Data Fig. 4). Elements that insert into double-stranded DNA to generate characteristic TSDs immediately adjacent to such terminal inverted repeats (TIRs) are known to be DNA transposons. Intron elements are diminutive and contain no appreciable open reading frames, making them presumably reliant upon transposases encoded elsewhere in the genome. Thus, intron elements are short, non-autonomous DNA transposons (also known as miniature inverted-repeat transposable elements or MITEs).

We found that intron elements in both genomes carried one splice site at the end of one TIR (Fig. 3b): *A. anophagefferens* intron elements carry a 3′ splice site (5′-AG-3′) and *M. pusilla* intron elements carry a 5′ splice site (5′-GT-3′ or 5′-GC-3′). For both types of intron element, the other splice site is constructed from TSD sequence (Fig. 3b), which originates from duplication of exonic sequence during intron element insertion. The sequence remains exonic on one side of the intron element, so that protein encoding is unaltered following intron splicing out of the RNA (Extended Data Fig. 5).

DNA transposons can generally insert into a genome in either of two orientations relative to a gene. Given the presence of TIR sequences at the ends of each intron element, there is the potential to carry the splice site in both orientations. Indeed, many *A. anophagefferens* intron elements carry 3′ splice sites in both orientations (see Supplementary Discussion and first intron element in Fig. 3a) and apparently have generated introns in either orientation. Indeed, many *A. anophagefferens* intron elements carry 3′ splice sites in both orientations (see Supplementary Discussion and first intron element in Fig. 3a) and apparently have generated introns in either orientation. For other *A. anophagefferens* intron elements, the TIRs differ in such a way that they carry a 3′ splice site in only one orientation, and are correspondingly found in that orientation in genes (Fig. 2c and second and third intron elements in Fig. 3a). Likewise, in *M. pusilla* intron elements, branchpoint sequences for splicing are apparent in the dominant orientation found in genes, and the vast majority carry a 5′ splice site in only that orientation (see Supplementary Discussion). Notably, *M. pusilla* intron element sequences are found occasionally in the opposite orientation, in which case they are not spliced as introns. They are also found in intergenic regions, consistent with proliferation as transposons that need not generate introns in every case.

**Figure 2** | Identification of intron elements in *A. anophagefferens*. a. Validated lengths for intron element (IE, green) and other (grey) introns. b. *A. anophagefferens* intron elements share sequence similarity in intronic sequence but not in neighbouring exonic sequence. Six example intron elements contain regions with maximal pairwise identities from 96 to 100%. Bases identical in at least 5 of the 6 sequences are green. c. Most *A. anophagefferens* intron elements can be aligned to form one or more related groups. Nodes present in >50% of 1,000 bootstraps are indicated with black dots on the ML tree. Intron elements are found in either orientation with respect to the intron (orange and blue). Many elements carry 3′ splice sites in both orientations (black lines on right).
Introner elements are DNA transposons that carry one inverted repeat, TIRs). Inverted repeats (underlined) are at introner element ends (terminal inverted repeats, TIRs).

In genes, the presence of introns is not a constant feature, and their absence is not uncommon. However, the presence of intron splicing sites (3′ ss), and the length of these elements, can vary significantly, even within the same species. This diversity in intron length is thought to be driven by the need to accommodate the requirements of the genetic code, which dictates that exons be of a certain minimum length to avoid premature translation termination.

Recent studies have suggested that intron gain may be quite general. Intron gain is often associated with the acquisition of new splice sites, which can be co-opted from the genome or generated de novo. However, the mechanisms by which these new splice sites are selected are not well understood.

Introner elements (IEs) are often found in genes that have undergone recent or ancient genomic rearrangements. These elements can be inserted into genes as a result of DNA transposition, which is a process by which DNA sequences are copied and inserted into different locations in the genome. Introner elements can also be co-opted to construct splice sites, which can then be used to generate new exons.

To explore the dynamics of introner elements, we sequenced the genome of another A. anophagefferens isolate sampled from the environment 11 years after the reference genome isolate. Sequence variation in the newer isolate demonstrates that its genome is diploid (Extended Data Fig. 6). Of the introner elements in the reference genome, 87% appear to be present within both alleles in the newer isolate (Fig. 4a), revealing the relative success of many introner elements in stably colonizing the genome. On the other hand, 42 of the reference introner element loci are present within only one allele of the newer isolate, and 32 of the reference loci lack the introner element in both alleles (Fig. 4a). We also identified 47 introner element insertions in the newer isolate that were not present in the reference genome, 31 of which (Fig. 4a) show the expected sizes of intron gain may be quite general.

Intron element dynamics and genomic implications. a. Presence–absence variation in a newer isolate of A. anophagefferens. b. Sequences that can be co-opted to construct splice sites are biased with respect to codon phasing. For M. pusilla, introner element introns should be biased by availability of AG sequences that can be co-opted as 3′ splice sites (3′ss). For A. anophagefferens, introner element introns should be biased by availability of GY (Y is C or T) sequences that can be co-opted for 5′ splice sites (5′ss). Introner element introns have phase biases similar to the respectively co-opted sequence (bold). c. Nearby introner element insertions generate nucleosome-sized segments. Distances between neighbouring introner element introns (solid) and between other neighbouring introns (broken) are displayed as kernel density estimates. Nucleosome repeat lengths of 206 bp for M. pusilla and 168 bp for A. anophagefferens show the expected sizes of integer numbers of nucleosomes (vertical lines).

Second, the sequences that can be co-opted to construct either 5′ or 3′ splice sites are biased in their phase distributions with respect to codons in genes (Fig. 4b). Therefore, selection for functional introns following introner element insertions in each organism should similarly bias the respectively co-opted splice site sequences, as is observed (Fig. 4b). This mechanism may explain the intron phase biases that are commonly observed in eukaryotes.

Insertion of DNA transposons makes sense of the apparent introner element preference for nucleosome linkers (Fig. 1 and Extended Data Figs 1, 2), as other DNA transposons show a strong preference for inserting between nucleosomes. Biased insertion of introner elements into nucleosome linkers provides a mutational mechanism for chromatin features to instruct the generation of new genetic material, namely introns (Extended Data Fig. 5). This possibility was proposed some time ago and has implications for the structure of eukaryotic genomes. For example, rapid insertion of multiple introner elements in close proximity could generate intervening segments (that is, exons if in the same gene) with sizes corresponding to integer numbers of nucleosomes; this is indeed observed (Figs 1a, 4c). Instruction of intron generation by chromatin features provides a straightforward mutational explanation for the tendency of animal exons to be approximately one nucleosome in length.

The lack of sequence similarity between the introner elements of A. anophagefferens and M. pusilla and the divergence of these organisms more than a billion years ago suggests that introner elements evolved independently. This idea is supported by the difference in TSD lengths (Extended Data Fig. 3), which implicate different transposase superfamilies, and the fact that introner elements carry a 5′ splice site in one organism and a 3′ splice site in the other (Fig. 3b). This independent evolution suggests that a DNA transposon mechanism for intron gain may be quite general.

Non-autonomous DNA transposons probably excel at generating introns for several reasons. First, DNA elements do not need to be
transcribed for transposition, especially if they are non-autonomous, permitting spread between genes that are not highly expressed (Extended Data Fig. 8). Second, whereas the extensive intron–exon base pairing required for group II intron splicing and mobility greatly constrains their genomic insertion sites and strongly reduces host gene expression. DNA transposons carrying one splice site can generate introns that are perfectly spliced out with only minimal requirements of sequence co-option for the second splice site (Extended Data Fig. 5). Third, non-autonomous transposons can be short and noncoding, enabling relatively efficient splicing and freedom from the constraint to encode transposases.

The intron element mechanism described here substantiates long-standing proposals that DNA transposons are a major source of genomic introns. Episodes of rapid intron gain would naturally occur following the chance evolution of intron elements, which are simply short DNA transposons carrying a splice site at an end. The antiquity and near ubiquity of DNA transposons opens up the possibility of an intron element mechanism for most intron gains in eukaryotes, both recent and ancient.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 November 2015; accepted 19 September 2016.
Published online 19 October ; corrected online 26 October 2016 (see full-text HTML version for details).

1. Gilbert, W. Why genes in pieces? Nature 271, 501 (1978).
2. Rogozin, I. B., Carmel, L., Courou, M. & Koonin, E. V. Origin and evolution of spliceosomal introns. Biol. Direct 7, 11 (2012).
3. Irimia, M. & Roy, S. W. Origin of spliceosomal introns and alternative splicing. Cold Spring Harb. Perspect. Biol. 6, a016071 (2014).
4. Cavalier-Smith, T. Selfish DNA and the origin of introns. Nature 315, 283–284 (1985).
5. Purugganan, M. & Wessler, S. The splicing of transposable elements and its role in intron evolution. Genetica 86, 295–303 (1992).
6. Wang, W., Yu, H. & Long, M. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in Drosophila species. Nat. Genet. 36, 523–527 (2004).
7. Li, W., Tucker, A. E., Sung, W., Thomas, W. K. & Lynch, M. Extensive, recent intron gains in Daphnia populations. Science 326, 1260–1262 (2009).
8. Yenarali, P. & Zhou, L. Identifying the mechanisms of intron gain: progress and trends. Biol. Direct 7, 29 (2012).
9. Schwartz, S., Moshor, E. & Ast, G. Chromatin organization marks exon-intron structure. Nat. Struct. Mol. Biol. 16, 990–995 (2009).
10. Warden, A. Z. et al. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Science 324, 268–272 (2009).
11. Verhelst, B., Van de Peer, Y. & Rouzé, P. The complex intron landscape and massive intron invasion in a picoeukaryote provides insights into intron evolution. Genome Biol. Evol. 5, 2393–2401 (2013).
12. Ht, J. T. & Zilberman, D. Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. Cell 156, 1286–1297 (2014).
13. Gobler, C. J. et al. Niche of harmful alga Aureococcus anophagefferens revealed through ecogenomics. Proc. Natl Acad. Sci. USA 108, 4352–4357 (2011).
14. van der Burg, A., Severing, E., de Wit, P. J. G. M. & Collere, J. Birth of new spliceosomal introns in fungi by multiplication of intron-like elements. Curr. Biol. 22, 1260–1265 (2012).
15. Simmons, M. P. et al. Intron invasions trace algal speciation and reveal nearly identical Arctic and Antarctic Micromonas populations. Mol. Biol. Evol. 32, 2219–2233 (2015).
16. Lambowitz, A. M. & Zimmerly, S. Group II introns: mobile ribozymes that invade DNA. Cold Spring Harb. Perspect. Biol. 3, a003616 (2011).
17. Calos, M. P., Johnsrud, L. & Miller, J. H. DNA sequence at the insertion sites of the element IS1. Cell 13, 411–418 (1978).
18. Grindley, N. D. F. IS1 insertion generates duplication of a nine base pair sequence at its target site. Cell 13, 419–426 (1978).
19. Wessler, S. R., lady, T. & White, S. E. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr. Opin. Genet. Dev. 5, 814–821 (1995).
20. van Baren, M. J. et al. Evidence-based green algal genomics reveals marine diversity and ancestral characteristics of land plants. BMC Genomics 17, 267 (2016).
21. Gangadharan, S., Malaron, L., Fin-Thornton, J., Wheelan, S. J. & Craig, N. L. DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo. Proc. Natl Acad. Sci. USA 107, 21966–21972 (2010).
22. Buenrostro, J. D., Gress, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013).
23. Parfrey, L. W., Lah, D. J. G., Knoll, A. H. & Katz, L. A. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc. Natl Acad. Sci. USA 108, 13624–13629 (2011).
24. Fescht, C. & Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. 41, 331–368 (2007).
25. Qi, G. et al. RNA-RNA interactions and pre-mRNA mislocalization as drivers of group II intron loss from nuclear genomes. Proc. Natl Acad. Sci. USA 111, 6612–6617 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Guthrie’s group for discussion of splicing, J. Stabile for discussion of A. anophagefferens isolates, K. Bi for discussion of sequence variation and the Vincent J. Coates Genomics Sequencing Laboratory for sequencing. This was supported by NIH grant T32HG000047 (J.T.H.) and a Beckman Young Investigator Award (D.Z.).

Author Contributions J.T.H. and S.W.R. performed the initial search for intron elements. J.T.H. performed the remaining experiments. J.T.H., D.Z. and S.W.R. designed the project, interpreted data and wrote the manuscript.

Author Information The A. anophagefferens intron element tree is available at TreeBASE (https://treebase.org; study 18167). Newly sequenced A. anophagefferens data are available at the SRA (SRP083781). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.Z. (danielz@berkeley.edu) and S.W.R. (scottroy@gmail.com).

Reviewer Information Nature thanks R. Chalmers, L. Hurst, D. Penny and the other anonymous reviewer(s) for their contribution to the peer review of this work.
METHODS

RNA sequencing analysis. Splice junction calls (Supplementary Data 1 and 2) from the strand-specific RNA sequencing reads of our previous study12 (GEO accession GSE46692) were made using TopHat v2.0.6 (ref. 26) with minimum and maximum intron lengths of 20 and 2,000 bases, respectively. We mapped to the JGI M. pusilla CCMP1545 assembly v3.0 ‘MicpuC3’10 and JGI A. anophagefferens assembly v1.0 ‘Auran1’13, in each case using both the genome sequence and the existing annotations. Intron position data were calculated from these transcriptome annotations. RNA levels of genes (Extended Data Fig. 8) were estimated using Cufflinks v2.0.2 with bias correction27 after duplicates had been removed using Picard v1.79 (http://broadinstitute.github.io/picard/).

General computational analyses. Analyses were performed with custom R and Perl scripts. Alignment of chromatin data (DNA methylation and nucleosome) from our previous study13 (GEO accession GSE46692) to intron positions and annotated gene start positions (as well as alignment of intron element intron positions to gene start) was performed using dslab-tools v1.5.52 (http://dslab-pmb.berkeley.edu/tools/). Intron positions are from the splice junctions we called (Supplementary Data 1 and 2). Gene start positions are from the existing transcriptome annotations (JGI M. pusilla CCMP1545 assembly v3.0 ‘MicpuC3’10 and JGI A. anophagefferens assembly v1.0 ‘Auran1’13). Mean values at each base pair for genes or sets of introns are presented for nuclease centre data in Fig. 1b, c, and for DNA methylation data in Fig. 1b, c and Extended Data Fig. 2. Kernel density estimates in Figs 1c, 4c and Extended Data Fig. 2b were made with the density function in R at each base pair with a Gaussian smoothing bandwidth of 25 bp. Each peak in Fig. 1c and Extended Data Fig. 2b was defined as the base pair position with the local maximum kernel density estimate. Logos (Figs 3b and Extended Data Fig. 7) were made with WebLogo v3.4 (http://weblogo.threeps-lusone.com/). Predicted intron phase distributions from co-opted sequences (Fig. 4b) are displayed if either all existing GY sequences were co-opted for 5′ splice sites or all existing AG sequences were co-opted as 3′ splice sites (Fig. 4b). These phases come from the predicted codons in JGI M. pusilla CCMP1545 assembly v3.0 ‘MicpuC3’10 and JGI A. anophagefferens assembly v1.0 ‘Auran1’13.

Genome-wide search for intron element sequences using sequence similarity between introns. BLASTN28 searches were performed between the intron sequences supported by splice junctions (Supplementary Data 1 and 2). Beginning with a seed sequence (a previously identified intron element10 for M. pusilla CCMP1545 and an intron with a high degree of similarity to many other introns for A. anophagefferens), a recursive greedy BLAST search was performed against all other intronic sequences. Specifically, the seed sequence was BLASTed against other intronic sequences, but not between neighbouring exonic sequences. BLASTN hits were collected and then used as seed sequences for the next round of BLAST. This recursive process terminated after several rounds for both organisms. BLASTN scores were filtered by requiring at least a sequencing depth of 50 and a variant quality score of 100. The resulting allele fractions for alternate alleles are presented in Extended Data Fig. 6a.

Reference intron element loci (without the first 8 bp of their introns to remove TSD sequence) were genotyped from the genomic sequencing reads using T-lex23 v2.2.2 with options ‘-noFilterTE -TSD -250 -v 25 -l 20 -lim 20’. Trimmomatic trimmed reads were mapped to the A. anophagefferens assembly v1.0 ‘Auran1’13 using BWA MEM19 v0.7.13 with default settings.

General genome variation was called using FreeBayes24 v1.2.9-29-g41c1313 using the options ‘-K 0 -Q 20’. Trimmed reads were mapped to the A. anophagefferens assembly v1.0 ‘Auran1’13 using BWA MEM19 v0.7.13 with default settings.

For the test presented in the text, intron ends were categorized as being in either nucleosome cores or linkers. Nucleosome cores were defined by finding the 1/206th genomic positions with highest nucleosome centre values12 as being in either nucleosome cores or linkers. Nucleosome cores were identified by aligning pairwise distances estimated by the maximum composite likelihood. A discrete gamma distribution for evolutionary rate differences among sites (+G; 5 categories with a parameter estimated to be 1.8285) was used. To compare only intron element sequences, we removed the TSD sequence (see Supplementary Discussion for description of TSD identification) present in each intron. Only positions with 75% or more coverage (that is, alignment gaps in less than 25% of the alignment) were used. Those introns that were perfectly aligned after manually removing terminal positions with many gaps and sequences with long wandering branches. This resulted in 168 positions of 398 sequences in the final analysis (Fig. 2c). We also performed 1,000 bootstraps. The final tree was rooted by midpoint, and terminal branches were coloured according to intron element orientation in Fig.Tree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

Internal branch was coloured according to majority rule of its constituent terminal branch orientations. Internal branches with equal numbers of constituent terminal branches in both orientations were coloured grey. A. anophagefferens genome sequencing. The reference genome isolate of A. anophagefferens (CCMP1848) was collected originally on June 6, 1986 in the Great South Bay of Long Island, New York, USA (40.6667°N, 73.25°W). We chose to sequence one of the most divergent isolates available (CCMP1794), collected originally on June 26, 1997 in Barnegat Bay near Ship Bottom, New Jersey USA (39.6475°N, 74.179°W). We obtained genomic DNA from this newer isolate directly from the Provasoli-Guillard National Center for Marine Algae and Microbiota.

Ten micrograms of DNA were sonicated for 2 × 2 min. on power setting 2.5 of a Misonix 3000 water bath sonicator at room temperature. The sheared DNA was cleaned up and size-selected by first binding it to 0.7 volumes of Agencourt AM Pure X beads (Beckon Coulter) and collecting the supernatant. That supernatant was then diluted and mixed with 0.75 volumes of Agencourt Ampure XP beads, this time washing and eluting the bound DNA. Two hundred nanograms of the purified and size-selected DNA was made into a library without PCR by using the Encore Rapid Library System (NuGEN; this kit has since been renamed the Ovation Rapid DR System). Paired-end 100-base sequencing was performed with the Illumina HiSeq 2000 platform.

Sequencing reads were first quality- and adaptor-trimmed using Skewer v2.0.2 (https://sourceforge.net/projects/skewer/) with options ‘-Q 30 -q 20’. Trimmed reads were mapped to the A. anophagefferens assembly v1.0 ‘Auran1’13 using BWA-MEM19 v0.7.13 with default settings.

General genome variation was called using FreeBayes24 v1.2.9-29-g41c1313 using the options ‘-K -F 0.01’, which makes calls without making assumptions about the ploidy as long as the minimum alternate allele fraction is 0.01. The variants were filtered by requiring at least a sequencing depth of 50 and a variant quality score of 100. The resulting allele fractions for alternate alleles are presented in Extended Data Fig. 6a.

Reference intron element loci (without the first 8 bp of their introns to remove TSD sequence) were genotyped from the genomic sequencing reads using T-lex23 v2.2.2 with options ‘-noFilterTE -TSD -250 -v 25 -l 20 -lim 20’. Resulting calls of ‘polymorphic’ and ‘absent’ loci were manually curated by inspecting soft-clipped mapped reads, and the results are presented in Fig. 4a.

Intron element sequences not present in the reference genome were identified using RetroSeq36 v1.41 using the reference intron element intron sequences to discover new candidates. Calls were then made using options ‘-reads 10 -depth 1000’. Candidate calls were further qualified and precisely assembled using the Kidd group’s pipeline37, originally devised for Alu elements. We modified the pipeline to use a custom repeat library composed of the A. anophagefferens intron element reference sequences from the alignment used to make the tree in Fig. 2c. Insertions were filtered for having lengths of between 185 and 205 bp (typical size range of A. anophagefferens intron elements) and putative TSDs of 7–11 bp. We expect an 8-bp TSD, but 7-bp TSDs can also be reported because of frequently erroneous 3-way alignments, which are used to estimate TSD length. Longer TSDs can also be reported because there can be longer 3-way alignments simply by chance. The 48 putative elements were manually curated by inspecting soft-clipped mapped reads, resulting in identification of 47 intron element sequences not present in the reference genome assembly with 31 present within only one of the two alleles (Fig. 4a).

Statistical testing. For the test presented in the text, intron ends were categorized as being in either nucleosome cores or linkers. Nucleosome cores were identified by finding the 1/206th genomic positions with highest nucleosome centre values12 and extending for 73 bp in both directions, which is the size of a nucleosome core. Nucleosome linkers are the regions in between the core regions. The P value of M. pusilla intron element intron ends in nucleosome linkers (73%) versus other intron ends in linkers (42%) was calculated using Fisher’s exact test. This includes both 5′ and 3′ intron ends together, but P values < 2.2 × 10−16 were also obtained using only 5′ intron ends or only 3′ intron ends.
To test whether discrete distributions of values differ for intron element introns versus other introns, the observed Kolmogorov–Smirnov statistic was compared to the statistics obtained in 10⁵ permutations of the data labels. For both genomes, intron element intron length distribution (Fig. 2a) differs significantly from that of other introns ($P < 10^{-5}$). For both genomes, the distribution of splice site distances (Fig. 4c) between intron element introns differs significantly from that of other introns ($P < 10^{-5}$ for A. anophagefferens). To test whether sequence and intron phases differ from unbiased probability (Fig. 4b) we used the multinomial test by enumeration, where possible (‘xmulti’ function from the R XNomial v1.0.4 package; https://cran.r-project.org/web/packages/XNomial/index.html), or otherwise a Monte Carlo multinomial test with $10^6$ trials (‘xmonte’ function from the R XNomial v1.0.4). For intron phases, we tested against unbiased probability of 1/3 each. For 5′ and 3′ sequences to co-opt, we tested against the observed probabilities of any sequences in each phase, which differed only slightly from 1/3 each. Unbiased probabilities of 1/3 in each phase are shown for reference as broken lines in Fig. 4b. For each distribution, observed numbers differ significantly from being unbiased ($P < 2.2 \times 10^{-16}$, except $P < 1.3 \times 10^{-7}$ for A. anophagefferens intron element introns).

To test whether expression of intron element-containing genes differs from that of all genes, Mann–Whitney–Wilcoxon tests were used (Extended Data Fig. 8).

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

**Modified dot plots for TSD and TIR characterization.** We modified the dot plot method²⁸ for pairwise comparison of the ends of each intron sequence. The matrix was altered, so that the vertical axis clearly indicates each of the relative offsets (also known as lags or phases) of one end to the other. These offsets are given relative to alignment of the 5′ and 3′ splice sites. The horizontal axis shows the positions in the pairwise comparison for each offset. The data in the matrix are given as the percentage of the set of introns, each of which has its 5′ and 3′ ends compared pairwise. Colour intensity displays the percentage, either with pairwise end-sequence identity (TSD characterization, Extended Data Fig. 3) or with pairwise end-sequence complementarity (orientation of the 3′ splice site end is also reversed for TIR characterization, Extended Data Fig. 4). Identity and complementarity were called only if they are part of at least a 2-mer of identity (examples in Extended Data Fig. 3b, d) or complementarity (examples in Extended Data Fig. 4b, d), respectively. Further details of TSD and TIR identification are in the Supplementary Discussion.

26. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. **14**, R36 (2013).
27. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. **12**, R22 (2011).
28. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. Genome Res. **14**, 1188–1190 (2004).
29. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. **215**, 403–410 (1990).
30. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. **30**, 772–780 (2013).
31. Nei, M. & Kumar, S. Molecular Evolution and Phylogenetics (Oxford Univ. Press, 2000).
32. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. **30**, 2725–2729 (2013).
33. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
34. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://arxiv.org/abs/1207.3907 (2012).
35. Fiston-Lavieu, A.-S., Barrón, M. G., Petrov, D. A. & González, J. T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. Nucleic Acids Res. **43**, e22 (2015).
36. Keane, T. M., Wong, K. & Adams, D. J. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics **29**, 389–390 (2013).
37. Wildschutte, J. H., Baron, A., Diroff, N. M. & Kidd, J. M. Discovery and characterization of Alu repeat sequences via precise local read assembly. Nucleic Acids Res. **43**, 10292–10307 (2015).
38. Gibbs, A. J. & McIntyre, G. A. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. Eur. J. Biochem. **16**, 1–11 (1970).
Extended Data Figure 1 | *M. pusilla* introner elements are in phase with nucleosome linker DNA, even without methylation. Unmethylated regions (indicated by the line with arrowheads) are defined as containing no base positions with fractional methylation 0.5 or greater in a window starting from 50 bp upstream of the 5′ end of the introner element intron and continuing 234 bp downstream, which is 50 bp beyond the predominant *M. pusilla* introner element intron size of 184 bp (Fig. 2a). Mean values at each base position are shown for chromatin maps12 aligned to the subset (7%) of introner element introns residing in unmethylated regions (dark grey and dark blue for nucleosome centres and DNA methylation, respectively), compared with alignment to all introner element introns (light grey and light blue; same data as in Fig. 1b for introner element introns). On the other hand, to assess whether introner elements could be in phase with methylated regions that are not also nucleosome linkers, we looked for introner elements that had both ends in methylated DNA regions12 but not in nucleosome linkers, which gave 35 potential candidates (1% of introner elements). Manual inspection revealed that 34 of the 35 candidates apparently nonetheless have ends in nucleosome linkers, simply being missed by the filtering criteria we used for calling linkers. This leaves one candidate, indicating little evidence that DNA methylated regions are found at introner element ends, which are not also nucleosome linkers. Thus, unmethylated nucleosome linkers could be the primary determinant of introner element insertion in at least some cases, whereas we find virtually no evidence that methylated regions could be the primary determinant of introner element insertion without also being nucleosome linkers.
Extended Data Figure 2 | *A. anophagefferens* introner elements insert into pre-existing nucleosome linkers. **a.** Introner element (IE) introns are generally in phase with nucleosome positions, whereas other introns are not. DNA methylation\(^1\) was aligned to the 5′ ends of introner element introns (dark blue) or other introns (light blue). We did not generate nucleosome data previously for *A. anophagefferens* but DNA methylation is a reliable indicator of linker locations\(^1\). **b.** Introner elements are in phase with the starts of genes, indicating insertion between pre-existing nucleosomes. The 5′ ends of introner element introns and DNA methylation\(^1\) were aligned to gene starts. A kernel density estimate of introner element ends is displayed with peaks marked by vertical broken lines.
Extended Data Figure 3 | Target site duplications (TSDs) at intron element introns. a, c, Intron sequences contain directly repeated sequences at their ends. Each *A. anophagefferens* (a) and *M. pusilla* (c) intron 5’ and 3’ end is directly aligned in each possible offset from −10 to 10bp apart. Positions relative to the 5’ splice site from 10 bp upstream to 10bp downstream are shown. Introner element (IE) introns are shown on the left and other regular non-introner element introns are in the centre, and the differences obtained by subtracting the identity percentages of other introns from those of introner element introns are on the right. Each panel is separated by a vertical black line and a diagonally stepped black line to delineate different regions: the upper left region represents alignment of upstream exon versus 3’ intron end sequence; the upper right represents 5’ intron end versus 3’ intron end; the lower right represents 5’ intron end versus downstream exon; and the lower left represents upstream exon versus downstream exon. The red arrowheads on the right indicate the offset with maximum average identity (0 in both cases). The red boxes in the right panels highlight the identified TSD length and position (see Supplementary Discussion). b, d, An example of an aligned 5’ (above) and 3’ (below) intron end of an introner element for the offset with maximum identity is shown in b for *A. anophagefferens* and d for *M. pusilla*. Exonic sequence is uppercase and boxed; intronic is lowercase. Vertical lines show identities that are part of at least an identical 2-mer with the red lines corresponding to the boxed regions in a and c.
Extended Data Figure 4 | Terminal inverted repeats (TIRs) in introner element introns. a, c, Intron end sequences contain inverted repeats. Each *A. anophagefferens* (a) and *M. pusilla* (c) intron 5′ and reverse of the 3′ end is aligned in each possible offset from −30 to 30 bp apart. Positions relative to the 5′ splice site from 30 bp upstream to 30 bp downstream are shown. Introner element (IE) introns are shown on the left and other regular non-introner element introns are on the right. In each panel the upper left region represents upstream exon versus downstream exon sequence, the upper right represents 5′ intron end versus downstream exon, the lower right represents 5′ intron end versus 3′ intron end, and the lower left represents upstream exon versus 3′ intron end. The red arrowheads (right) indicate the offset with maximum average complementarity. b, d, An example of an aligned 5′ (top) and 3′ (bottom, reversed so that it is 3′ to 5′) end of an introner element intron for the offset with maximum complementarity is shown in b for *A. anophagefferens* (offset of +8) and d for *M. pusilla* (offset of −5). Exonic sequence is uppercase and boxed; intronic is lowercase. Vertical lines show complementarities that are part of at least an identical 2-mer.
Extended Data Figure 5 | Intron gain templated by nucleosomes and co-opted sequences. Model for intron generation by introner elements acting as short non-autonomous DNA transposons that carry a splice site and insert between nucleosomes with co-option of the other splice site sequence.
Extended Data Figure 6 | Diploid genomic sequence variation in a more recent isolate of A. anophagefferens. a, Calling of sequence variation from genomic sequencing reads without an assumption of ploidy reveals a peak at an alternate allele fraction of approximately 0.5. The most likely scenario is that this A. anophagefferens isolate has a diploid genome. It is not physically plausible for it to have higher ploidy because that amount of chromatin could not fit into its extremely compact nucleus. b, An example reference introner element (IE) is present within one allele and absent from the alternate allele. The locus is displayed as in Fig. 3a. The reference introner element is located in an annotated protein-coding gene with a 200-bp RNA sequencing-validated intron in the reference isolate. The alternate allele is probably exonic without an intron (broken lines), so that it encodes the same amino acid sequence. The TSD within the reference allele is 8 bp, immediately flanking the introner element TIRs. c, An example introner element not found within the reference allele is present within the alternate allele. The locus is displayed as in Fig. 3a. The alternate introner element is within an annotated protein-coding gene with a predicted 200-bp intron (broken lines). If the predicted intron is indeed spliced out of the RNA, then the alternate allele encodes the same amino acid sequence. The TSD within the alternate allele is 8 bp, immediately flanking the introner element TIRs.
**Extended Data Figure 7 | Splice site sequences.** Logos for the 10 bp upstream and downstream of 5′ and 3′ splice sites for intron element and other introns are shown for each organism. The rectangles show exonic positions. The core splice sites are GY (Y is C or T) and AG. Introner elements (IEs) combined with co-opted exonic sequence that is duplicated (Fig. 3) to generate particular sequences that extend beyond the core sites (bracketed). Specifically, this results in a predominance of AG|GY sequences (| denotes the position of splicing that ultimately occurs) at 5′ splice sites in *M. pusilla* intron element introns and 3′ splice sites in *A. anophagefferens* intron element introns. Similar respective sequences are observed in other introns in each organism: G|GT for *M. pusilla* 5′ splice sites and AG|G for *A. anophagefferens* 3′ splice sites. In non-intron element introns, these sequences have been under selection for long periods of time to promote RNA splicing, revealing the sequences extending beyond core sites that probably contribute to optimal splicing in each organism. The similarity of intron element intron splice sites to other intron splice sites thus suggests that intron elements in each organism generate new introns that are spliced reasonably well.
Extended Data Figure 8 | Most intron elements are located in genes expressing low to average RNA levels. Distributions of detectable RNA levels of all transcripts (black) and only those containing at least one intron element (IE-containing, green) are shown as measured by RNA sequencing. Box plots indicate the median, first and third quartiles with whiskers extending up to data 1.5 times the interquartile range away from the box. For *M. pusilla*, intron element-containing gene expression does not differ significantly from that of all genes, $P = 0.59$. For *A. anophagefferens*, intron element-containing gene expression is slightly lower than that of all genes, $P = 0.041$. 

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.