Supplementary Discussion

TSD and TIR identification

For *A. anophagefferens*, identified IE introns (and their neighboring exons) have a clear pattern of direct sequence identity (Fig. 3a and Extended Data Fig. 3a) of 8 bp when each 5′ and 3′ end is aligned with an offset of 0. Although some identity is found in other non-IE introns in the same offset (indicating general preference for exonic sequence next to all intronic splice sites), the amount in IE introns is much greater, as seen by subtracting the identity values of other introns from the identity values of IE introns (Extended Data Fig. 3a). These direct sequence identities (for an element that has presumably inserted) are known as TSDs. The TSD of each IE tends to be particular to that IE, which is seen comparing individual examples (Fig. 3a). More generally, there is almost no similarity of TSDs between different elements, which is seen when comparing many elements in a sequence logo (Fig. 3b). The exception is the 5′-GT-3′ and 5′-GC-3′ sequences, which are presumably selected to constitute 5′ splice sites. Comparison of IE intron ends shows reverse complementarity when each 5′ and (reversed) 3′ end is aligned with an offset of +8, which demonstrates an imperfect (occasionally perfect, example in Fig. 3a) TIR over approximately 14 to 18 bp; no such reverse complementarity is apparent in other non-IE introns (Extended Data Fig. 4a). TSDs of 8 bp and TIRs with an offset of +8 indicate the simple scenario in which a DNA sequence with TIRs has cleanly inserted, making 8 bp TSDs to generate each intron.

For *M. pusilla* CCMP1545, identified IE introns (and their neighboring exons) have a pattern of direct sequence identity (Fig. 3a and Extended Data Fig. 3c) apparently of 4 bp when each 5′ and 3′ end is aligned with an offset of 0. Although some identity is found in other non-IE introns in the same offset (indicating general preference for exonic sequence next to all intronic splice sites), the amount in IE introns is greater, as seen by subtracting the identity values of other introns from the identity values of IE introns (Extended Data Fig. 3c). The TSD of each element tends to be particular to that IE, which is seen comparing individual examples (Fig. 3a). Comparison of IE intron ends shows imperfect reverse complementarity when each 5′ and (reversed) 3′ end is aligned with an offset of -5, which demonstrates an imperfect TIR over approximately 10 to 16 bp (Fig. 3a and Extended Data Fig. 4c); no such reverse complementarity is apparent in other non-IE introns (Extended Data Fig. 4c).

The differences between identical sequences comparing *M. pusilla* IE and other non-IE introns suggest that after accounting for general intron sequence preferences, the specific TSD may only be 3 bp (Extended Data Fig. 3c). TSDs of either 3 or 4 bp and TIRs with an offset of -5 indicate a more complex scenario than in *A. anophagefferens*. First, a TSD length of 3 bp must be correct, because a length of 4 bp for TSDs would necessitate that 1 bp of one (but not the other) TSD and TIR overlap; specifically, overlap of the last G of the exon immediately 5′ of the intron (Fig. 3a,b). It is unclear mechanistically how such overlap or TIR end asymmetry could be produced. For example, as far as we know, no such asymmetric overlap has ever been found in DNA transposons. Instead, TSDs of 3 bp would suggest that the G 5′ of the intron is simply the most terminal position of the TIR—then, no overlap is required. When the 5′ and 3′ splice sites are aligned (Extended Data Fig. 3c,d), the G would be immediately downstream of the 3 bp TSD, making it appear as 4 bp of identity, because the G matches the invariant G necessary to constitute the 3′ splice site. Thus, the TIR G base apparently supplants an existing G (presumably

selected from the original exonic sequence) to construct the last base of the upstream exon, whereas the original G becomes the last base of the intron to construct the 3′ splice site (Fig. 3b). Thus, the A base of the 5′-AG-3′ immediately 5′ of the intron is the last base of 3 bp TSDs and is presumably selected to constitute the 3′ splice site after duplication (Fig. 3b). Even with selection constraint on 1 of 3 of the TSD base pairs, there is still only modest similarity between different element TSDs, which is seen when comparing elements (Fig. 3).

For both organisms an offset of 0 for TSDs (Extended Data Fig. 3) corresponds to perfect excision of introns—combining the TSD sequences in silico yields an intronless sequence encoding the same amino acid sequence, which is what occurs in effect by the process of RNA splicing. Put another way, one of the duplicated sequences is spliced out of transcribed RNA, leaving a single copy of the sequence in the RNA that presumably corresponds to the original exonic sequence into which the IE inserted (Extended Data Fig. 5).

Splice site orientation in IEs

A. anophagefferens IE introns necessarily have a 3′ splice site (5′-AG-3′ for the spliceosome) in the TIR at the 3′ end with respect to their host gene (“right TIR”, Fig. 3 and Extended Data Fig. 7). The additional presence of the reverse complement of the 3′ splice site sequence (5′-CT-3′) in the alternate orientation of each IE (Fig. 2c) was assessed at the equivalent position in the other IE TIR at the 5′ end with respect to the host gene (“left TIR”; example in Fig. 3a). By this criterion 107 of 398 aligned sequences (27%; black lines in Fig. 2c) have 3′ splice sites in both orientations. For comparison, the genomic background frequency of the sequence 5′-CT-3′ is 4.5%. The first *A. anophagefferens* example in Fig. 3a carries 3′ splice sites in both orientations (5′-AG-3′ at the end of the right TIR and also a 5′-CT-3′ at the end of the left TIR); the next two *A. anophagefferens* examples in Fig. 3a represent alternate orientations carrying a 3′ splice site (5′-AG-3′ at the end of only the right TIR) in only one orientation. See the previous section for details of TIR identification.

M. pusilla IE introns apparently have a 5′ splice site that is 5′-GY-3′ (either 5′-GT-3′ or 5′-GC-3′), which is in the TIR at the 5′ end with respect to their host gene (“left TIR”, Fig. 3 and Extended Data Fig. 7). We found that 7.0% (235 of 3347) of *M. pusilla* IEs have the reverse complement of a 5′ splice site (either 5′-AC-3′ or 5′-GC-3′) in the alternate orientation at the equivalent position in the other IE TIR at the 3′ end with respect to the host gene (“right TIR”). For comparison, the genomic background frequency of the sequences 5′-GT-3′ or 5′-GC-3′ is 17.5%. See the previous section for details of TIR identification.