RESEARCH ARTICLE SUMMARY

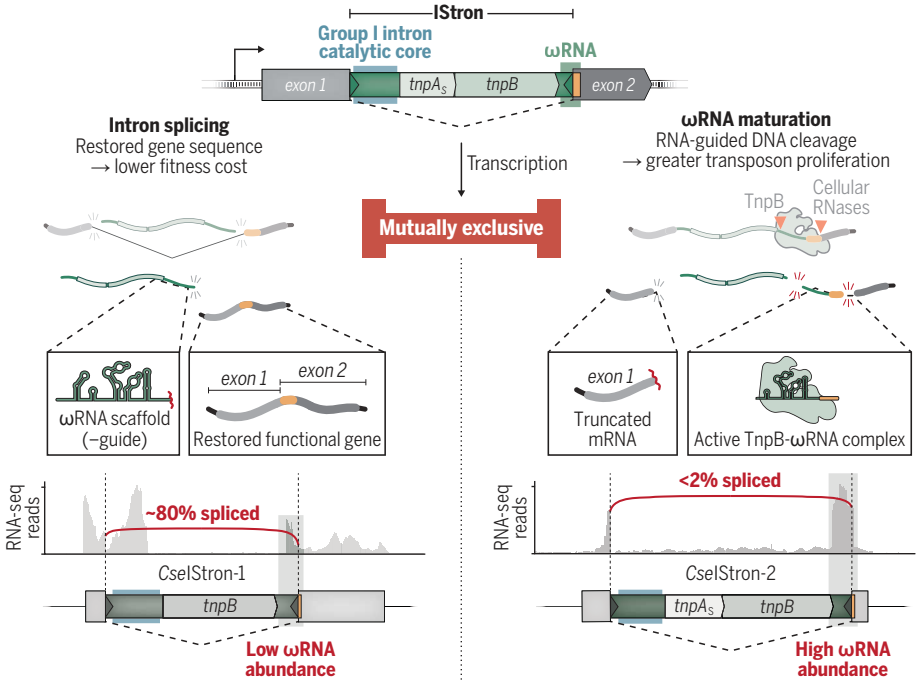
MOLECULAR BIOLOGY

Antagonistic conflict between transposon-encoded introns and guide RNAs

Rimantė Žedaveinytė, Chance Meers, Hoang C. Le, Edan E. Mortman, Stephen Tang, George D. Lampe, Sanjana R. Pesari, Diego R. Gelsinger, Tanner Wiegand, Samuel H. Sternberg*

INTRODUCTION: TnpB nucleases represent the evolutionary precursors to CRISPR-Cas12 and are widespread in all domains of life, presumably owing to their key role in transposon proliferation. We recently demonstrated that IS605-family TnpB homologs function as programmable homing endonucleases by exploiting transposon-encoded guide RNAs (ωRNA) to cleave genomic DNA, thereby driving transposon maintenance through DNA double-strand break–stimulated homologous recombination. Whether this pathway is conserved in other genetic contexts, and in association with other transposases, is unknown. In this study, we focused on IS607-family transposons, whose transposition lifestyles are poorly understood and which frequently encode self-splicing group I introns alongside TnpB and ωRNA.

RATIONALE: IStrons—complex transposable elements encoding group I introns—feature boundary sequences that must be faithfully recognized by a serine-family $TnpA_S$ transposase, and they also encode catalytic intron RNAs and TnpB-associated ωRNAs. Splicing potentially renders the transposon DNA insertion phenotypically silent by rejoining the interrupted gene segments at the RNA level. However, this reaction conflicts with normal ωRNA maturation by severing the scaffold and guide sequences at the 3′ splice site. We hypothesized that IStron sequences converged to satisfy these mutually exclusive molecular requirements and thereby balance the needs for transposon retention and spread while limiting their fitness cost.



Conflicting nature of IStron-encoded noncoding RNAs. TnpB nucleases play a critical role in transposon proliferation. Alongside their presence in simple transposons, *tnpB* genes and associated guide RNAs (ωRNA) are mobilized by complex IStrons that encode a transposase and self-splicing intron. We found that ωRNA maturation (right) is mutually exclusive with RNA splicing (left), revealing an intricate balance between enzymatic activities that are necessary for both promoting transposon maintenance and mitigating host fitness costs. RNase, ribonuclease; RNA-seq, RNA sequencing.

RESULTS: After applying a systematic informatics approach to analyze diverse bacterial IStrons, we selected a candidate element from *Clostridium botulinum* (*Cbo*IStron) and reconstituted DNA transposition, RNA splicing, and RNA-guided DNA cleavage in a heterologous *Escherichia coli* host. Our data revealed that the $TnpA_S$ transposase and group I intron mediate scarless rejoining of the same nucleotide sequences at the DNA and RNA level, respectively, highlighting convergent recognition of the left- and right-end sequences. Transposition occurs through a circular intermediate and targets GG dinucleotide target sites for DNA integration, which matches the transposon-adjacent motif that is recognized by TnpB during RNA-guided DNA cleavage. When we monitored transposon copy-number changes in the presence of TnpA and TnpB, we found that IS607-encoded TnpB nucleases, like their counterparts in IS605 elements, target the scarless donor joint for DNA cleavage and drive transposon retention through *recA*-dependent homologous recombination. Targeted genetic perturbations identified regions that are critical for both ωRNA and group I intron function while further demonstrating that splicing and TnpB activity are in conflict with each other. In particular, structural features of the ωRNA alone suppress splicing, and this effect is enhanced in the presence of TnpB. These antagonistic features were exemplified by two native IStrons from *Clostridium senegalense* (*Cse*IStron-1 and *Cse*IStron-2), which were alternately highly active for either splicing or guide RNA maturation, but not both.

CONCLUSION: Our work reveals that IStrons have evolved a sensitive equilibrium to balance competing and mutually exclusive activities within the same transposon-encoded transcript. Self-splicing activity conferred by the group I intron has the potential to mitigate one of the most deleterious traits of transposons—insertional mutagenesis events that impose a severe fitness cost—by restoring the function of the interrupted gene. However, this property comes at the cost of truncating and inactivating a potential guide RNA molecule, thus depriving TnpB of its retention mechanism and placing the transposon at greater risk of extinction. Taken together, these findings about IStrons highlight the diverse enzymatic activities that emerged during selfish transposon spread and the multifunctional potential of transposon-encoded noncoding RNAs. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: shsternberg@gmail.com
Cite this article as R. Žedaveinytė *et al.*, *Science* **385**, eadm8189 (2024). DOI: 10.1126/science.adm8189

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.adm8189

RESEARCH ARTICLE

MOLECULAR BIOLOGY

Antagonistic conflict between transposon-encoded introns and guide RNAs

Rimantė Žedaveinytė[1], Chance Meers[1], Hoang C. Le[1], Edan E. Mortman[2], Stephen Tang[1], George D. Lampe[1], Sanjana R. Pesari[1]†, Diego R. Gelsinger[1], Tanner Wiegand[1], Samuel H. Sternberg[1]*

TnpB nucleases represent the evolutionary precursors to CRISPR-Cas12 and are widespread in all domains of life. IS605-family TnpB homologs function as programmable RNA-guided homing endonucleases in bacteria, driving transposon maintenance through DNA double-strand break–stimulated homologous recombination. In this work, we uncovered molecular mechanisms of the transposition life cycle of IS607-family elements that, notably, also encode group I introns. We identified specific features for a candidate "IStron" from *Clostridium botulinum* that allow the element to carefully control the relative levels of spliced products versus functional guide RNAs. Our results suggest that IStron transcripts evolved an ability to balance competing and mutually exclusive activities that promote selfish transposon spread while limiting adverse fitness costs on the host. Collectively, this work highlights molecular innovation in the multifunctional utility of transposon-encoded noncoding RNAs.

Transposase genes are among the most abundant genes in nature, owing to their high copy number and cross-kingdom host diversity (*1*). These enzymes transpose DNA through chemically disparate mechanisms and can be classified based on the presence of conserved domains that fall into one of several major protein families, including DD(E/D), tyrosine transposase, and serine transposase (*2*, *3*). In turn, the DNA elements recognized by transposases bear hallmark sequence features at their left end (LE) and right end (RE), ensuring a tight molecular coupling between transposon substrates and the enzymes that mobilize them (*4*). Two large bacterial transposon families typified by Insertion Sequence 605 (IS605) and Insertion Sequence 607 (IS607) encode distinct transposase machineries but the same accessory factor, TnpB (Fig. 1A) (*3*, *5*–*9*). TnpB is an RNA-guided DNA endonuclease that forms a ribonucleoprotein complex with a transposon RE-derived guide RNA, known as ωRNA (*9*), and mediates double-stranded DNA (dsDNA) cleavage using its RuvC domain (*8*, *9*). Eukaryotic TnpB homologs known as Fanzors are associated with an even greater diversity of transposases, highlighting the pervasive and recurrent co-option events involving this ubiquitous gene, as well as the useful properties it apparently provided to these mobile elements over evolutionary timescales (*10*–*13*). TnpB domestication also notably led to the evolution of type V CRISPR-Cas adaptive immunity, because Cas12-family RNA-guided nucleases can be phylogenetically linked to these transposon-encoded enzymes (*5*, *6*, *8*, *9*, *14*).

We recently showed that TnpB and guide RNAs play a key role in the maintenance and spread of IS605-family transposons through a peel-and-paste, cut-and-copy transposition mechanism that involves RNA-guided DNA cleavage (*7*). Two key insights resulted from our observations: The tyrosine-family TnpA recombinase responsible for IS605 mobilization (hereafter $TnpA_Y$) catalyzes transposon excision more frequently than transposon integration, and scarless excision leads to permanent loss of the element at the donor site, foreboding eventual extinction of the mobile element. In the presence of TnpB, though, targeted DNA double-strand breaks (DSBs) trigger efficient recombination with a sister chromosome still harboring the transposon, thereby promoting retention. We demonstrated this biological function using representative transposable elements from *Geobacillus stearothermophilus*, but the extent to which this pathway applies to other transposon families that also encode TnpB nucleases, such as IS607, remained unknown (*3*, *15*).

IS607 transposons are among the most widespread transposable elements in nature, spanning bacteria, archaea, and eukaryotes; they are even found in viruses that infect these hosts (*16*–*18*). These transposons are mobilized by a serine-family recombinase also named TnpA (hereafter $TnpA_S$) via a recombination reaction akin to small and large serine recombinases such as Tn3 resolvase and Bxb1 integrase, respectively (*19*–*21*). Despite the common naming scheme, $TnpA_Y$ and $TnpA_S$ do not share a common ancestor, and they use distinct catalytic mechanisms and substrate preferences. Whereas $TnpA_Y$ transposes IS605 elements through a circular single-stranded DNA (ssDNA) intermediate using a 5′-phospho-tyrosine linkage (*22*, *23*), $TnpA_S$ transposes IS607-family elements through a circular dsDNA intermediate, using a 5′-phospho-serine linkage. (*4*, *18*, *24*) (Fig. 1A). Yet a common feature of both pathways is the role that the transposase plays in chemically resealing the sequences that flank the transposon at the donor site during the excision step to produce a scarless "donor joint" that regenerates the original genomic sequence (Fig. 1A). Thus, IS607 transposons, similar to IS605 transposons, are at risk of copy-number loss and eventual extinction, under circumstances where the excision frequency outpaces the integration frequency at new target sites (*7*). We therefore hypothesized that TnpB might satisfy the same evolutionary function for both types of elements to promote their maintenance and spread through a programmable cut-and-copy mechanism.

A subset of IS605 and IS607 transposons feature an additional evolutionary adaptation—the presence of self-splicing group I introns—and are thus termed "IStrons" (Fig. 1, B and C) (*25*–*30*). Group I introns are large RNA enzymes, or ribozymes, that require only Mg^{2+} and guanosine triphosphate (GTP) for self-splicing (*31*) (Fig. 1B), and they possess the exceptional ability to persist silently in genomes by removing themselves from interrupted transcripts, thereby restoring contiguous structural RNA or protein-coding genes (*32*). Unlike group II introns, which use protein cofactors (maturases) for splicing and reverse transcriptases for mobilization (*33*), group I introns either use reverse splicing for mobilization (*34*, *35*) or co-opt alternative enzymatic pathways that involve homing endonucleases or transposases (*36*–*38*). The intricate genetic architecture of IStrons, which incorporates both TnpA and TnpB enzymes alongside the presence of transposon end sequences that overlap with intron splice sites (Fig. 1C), suggests a sophisticated spread-stealth mechanism that enables DNA transposition while simultaneously reducing fitness costs by facilitating splicing. However, initial studies that described IStrons were unaware of the presence of TnpB-associated guide RNAs (ωRNAs) (*25*–*30*), which would likely compete with the process of splicing itself. We therefore wondered how mutually exclusive activities involving the same IStron transcript could satisfy the molecular requirements for RNA splicing and RNA-guided DNA targeting.

In this work, focusing on IS607-family elements for which the role of TnpB has not been explored, we reconstituted DNA transposition, RNA self-splicing, and RNA-guided DNA cleavage activities of a representative IStron from *Clostridium botulinum*, the causative

[1]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA. [2]Department of Genetics and Development, Columbia University, New York, NY 10032, USA.
*Corresponding author. Email: shsternberg@gmail.com
†Present address: Biochemistry and Molecular Biophysics Program, University of California, San Diego, La Jolla, CA 92093, USA.

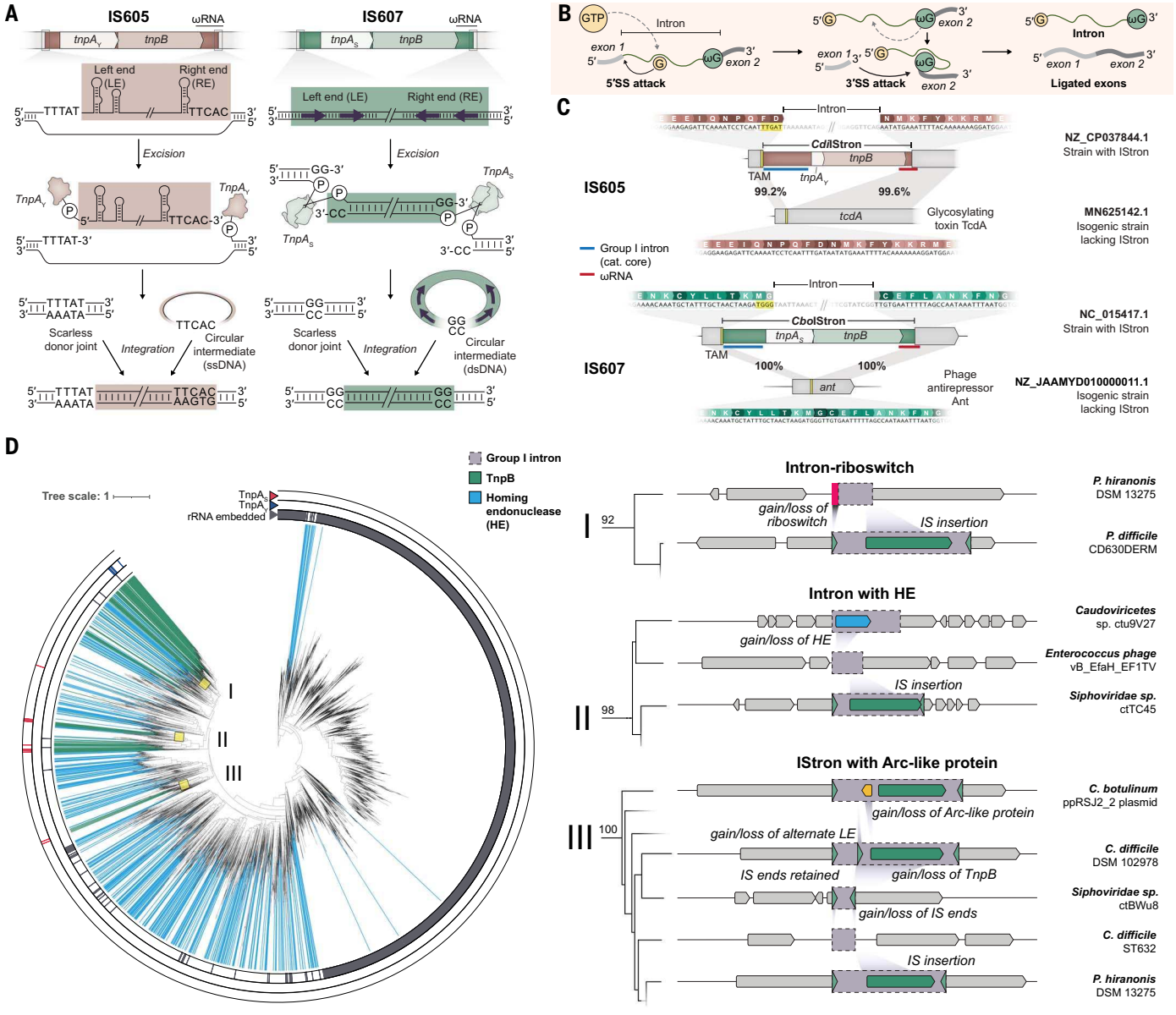


Fig. 1. Genomic architecture and phylogenetic analysis of TnpB-encoding IStrons. (**A**) IS605 transposons mobilize through a ssDNA intermediate using a tyrosine-family recombinase ($TnpA_Y$, left); by contrast, IS607 transposons mobilize through a dsDNA intermediate using a serine-family transposase ($TnpA_S$, right). Both transposon families are bounded by LE and RE sequences, encode *tnpB* and ωRNA, excise as circular intermediates, and generate scarless donor joints that regenerate the original (uninterrupted) genomic sequence. (**B**) Schematic of the group I intron self-splicing pathway. Splicing is initiated by binding of exogenous GTP (yellow) and attack of the 5′SS, followed by attack of the 3′SS to yield ligated exons and the excised intron RNA. ωG (green) refers to the terminal nucleotide of the intron and is unrelated to ωRNA. (**C**) Genetic architecture of representative IS605- and IS607-family IStrons from Clostridia strains, alongside homologous sites that lack the transposon insertion. The percent sequence identity between shaded regions (gray) is shown, as are the genomic accession IDs; TAMs are highlighted in yellow. Flanking exons are labeled, and the inset (bottom) highlights the predicted mRNA splicing products that regenerate a functional open reading frame (ORF). (**D**) Phylogenetic tree of group I introns related to *Cbo*IStron (left), with genetic architectures of selected clades (yellow boxes) schematized on the right. The outer tree rings indicate associations with $TnpA_S$ or $TnpA_Y$, as well as whether the group I intron is embedded within a ribosomal RNA locus. The green and blue colors indicate associations with TnpB nucleases or homing endonucleases (HEs). Bootstrap values are indicated for select nodes. *P. hiranonis*, *Peptacetobacter hiranonis*.

agent of botulism (*39*–*41*). Our experiments revealed the overlapping and contingent molecular determinants of three distinct nucleic acid scission-joining reactions, highlighting the notable ability for a single polynucleotide sequence to direct orthogonal and multifunctional chemical reactions. In particular, we found that the full-length IStron transcript serves parallel, but distinct, roles as an intron capable of protein-independent self-splicing and as a noncoding guide RNA that specifies genomic target sites for DNA cleavage. The relevant balance between these activities is controlled both by intrinsic structural features of the RNA itself and by the TnpB nuclease, highlighting the central regulatory role of TnpB in the IStron life cycle. Collectively, our work unveils an ingenious and unprecedented convergence of molecular activities that exemplifies the capacity for molecular innovation by transposable elements.

Results

Bioinformatic analyses of IStron elements

IStrons feature unusually long intergenic regions upstream of *tnpA*, which are clearly contained within the transposon boundaries and encode a structured noncoding RNA that matches the self-splicing group I intron covariance model (CM) from the Rfam database (*42*) (Fig. 1C). After systematically analyzing the genomic neighborhood of more than 95,000 *tnpB* genes in our dataset, we identified numerous clades that were associated with introns, with clear evidence of independent ancestral co-option events between catalytic ribozymes and distinct IS*605* (*tnpA*$_Y$) or IS*607* (*tnpA*$_S$) transposons (figs. S1 to S3). Motivated by this observation, we constructed an improved group I intron CM and used it to identify all related group I introns in the National Center for Biotechnology Information (NCBI) nucleotide database. This analysis highlighted diverse genomic contexts and associations with an eclectic mix of auxiliary genes, including TnpA transposases, TnpB nucleases, homing endonucleases, and uncharacterized accessory factors (Fig. 1D and fig. S4; see supplementary text).

We focused our attention on IS*607* elements, for which TnpB and transposition have not been well studied, and identified an IStron in *C. botulinum* strain BKT015925 (hereafter *Cbo*IStron) that encodes full-length TnpB and a serine recombinase (TnpA$_S$). The *Cbo*IStron is located on a large extrachromosomal plasmid alongside an IStron that lacks *tnpA*, multiple additional IS*607* elements, and the botulinum neurotoxin (fig. S3A and table S1) (*40*, *41*). IS*607* transposons lack inverted repeats or imperfect palindromic ends, and transposition does not generate a target-site duplication (*15*, *18*, *24*), rendering unbiased transposon boundary detection with sequence- or structure-based models challenging. However, we were able to provisionally assign LE and RE boundaries for these elements by comparing IStron-containing and IStron-lacking strains (Fig. 1C and fig. S3B), which, together with the analysis of homologous *Cbo*IStrons, revealed a guanine-rich transposon-adjacent motif (TAM) with consensus 5′-TGGG-3′ sequence (fig. S3C). Based on analysis of the flanking exonic gene segments, the putative splice sites perfectly matched the transposon DNA boundaries (Fig. 1C).

We were especially interested to examine the putative *Cbo*TnpB ωRNA. In the case of IS*605* transposons, conserved ωRNA stem-loops bound by TnpB coincide with palindromic DNA stem-loops in the RE (fig. S2, F and G), which are specifically recognized by TnpA$_Y$ during transposon excision and integration (Fig. 1A) (*22*). Conversely, IS*607* transposons are recognized as dsDNA by TnpA$_S$ through poorly understood sequence determinants (*15*, *18*, *24*), suggesting an alternative evolution of RE sequences that would satisfy requirements of both TnpA$_S$ (DNA) and TnpB (RNA). After identifying the likely ωRNA through a *Cbo*TnpB-anchored CM (fig. S3E), we found that ωRNAs exhibit universal predicted structural features despite their association with distinct IS*605* and IS*607* elements, characterized by three consecutive stem-loops upstream of the guide sequence that make up the ωRNA "scaffold" (fig. S3F). The first SL is predicted to base pair with complementary sequences at the scaffold 3′ end through a pseudoknot (PK) interaction (fig. S3F), and this feature is conserved across all characterized TnpB-ωRNA systems and numerous CRISPR-Cas12 systems (*43*–*47*). These commonalities attest to the evolutionary relatedness of TnpB and Cas12 proteins and their RNA substrates and demonstrate that essential RNA structural motifs can be accommodated by specific DNA requirements across diverse transposable elements.

We next set out to reconstitute the enzymatic activities of IStron-encoded TnpA$_S$, TnpB, and the group I intron as a way to probe the overlapping and potentially antagonistic processes of RNA-guided DNA cleavage and RNA splicing.

TnpA$_S$ catalyzes efficient IStron excision and integration

We started by developing a heterologous transposon excision assay in *Escherichia coli*, in which cells were transformed with a *Cbo*TnpA$_S$ expression plasmid (pTnpA$_S$) and a transposon donor plasmid (pDonor) encoding a minimal *Cbo*IStron (Mini-IS) that lacked *tnpA*$_S$ and *tnpB* genes. Transposon excision is expected to generate a circular DNA transposon intermediate and regenerated donor site containing the adjoined TAM and target site (donor joint), which we detected using either endpoint or quantitative polymerase chain reaction (PCR) (Fig. 2A and fig. S5A). TnpA$_S$ catalyzed robust transposon excision, an activity that was abolished in the presence of protein mutations to the TnpA$_S$ active site or DNA mutations to the core dinucleotides or transposon ends (Fig. 2B). Sanger sequencing revealed that plasmid excision products exhibited precise donor joints (Fig. 2C), and we also readily detected the presence of circular excision intermediates (fig. S5B). DNA excision efficiencies were ~10% after culturing overnight (fig. S5, C and D), and thus orders of magnitude higher than previously reported excision efficiencies of TnpA$_Y$ from IS*605* elements (*7*), suggesting that excisional transposon loss might be an even greater risk for IS*607* elements. In addition, we quantified the effects of serial LE and RE deletions on excision, which revealed that only 40 and 60 base pairs (bp) of the LE and RE, respectively, were critical for TnpA$_S$ recognition (fig. S5, C to E).

Having confirmed that TnpA$_S$ transposition proceeds via a circular DNA intermediate, we analyzed genomic integration using a modified transposition assay. We expressed TnpA$_S$ in the presence of nonreplicative suicide vectors containing abutted transposon ends, thus mimicking the circularized intermediate (pDonor$_{CI}$), such that integration would lead to a selectable antibiotic resistance phenotype (Fig. 2D). After first confirming that cellular survival was dependent on catalytically active TnpA$_S$ (Fig. 2E), we applied a next-generation sequencing (NGS) approach to interrogate genome-wide integration specificity (fig. S5F). These experiments revealed hundreds of distinct genomic integration sites that were broadly distributed and shared a strict GG dinucleotide requirement (Fig. 2, F and G, and fig. S5, G and H), with additional preferences upstream to yield a 5′-TGGG-3′ target site motif, in excellent agreement with the bioinformatically predicted TAM (fig. S3C). Target-site selectivity for large serine recombinases can be modified by mutating core dinucleotides within the recombination sequence (*20*, *48*, *49*), so we wondered whether small TnpA$_S$ recombinases would similarly switch their target-site selectivity if the GG dinucleotide were altered. Intriguingly, TnpA$_S$ still targeted 5′-TGGG-3′ sites when pDonor$_{CI}$ contained a GC dinucleotide, but when we carefully inspected the actual genomic integration products from NGS data, we found that the LE and RE harbored nonmatching GC and GG dinucleotides, respectively (Fig. 2G). Further testing of other modified dinucleotides on pDonor$_{CI}$ revealed an unexpected diversity of preferred integration sites and LE and RE genomic integration products (fig. S5H), highlighting the need for additional biochemical experiments to resolve the detailed mechanism of TnpA$_S$-mediated recombination.

These experiments demonstrated that IS*607* donor-site transposon loss is a direct consequence of TnpA$_S$ activity, just as IS*605* donor-site transposon loss is a consequence of TnpA$_Y$ activity (*7*, *23*). Thus, we next set out to investigate *Cbo*TnpB and its impacts on *Cbo*IStron transposon maintenance.

IStrons encode active TnpB nucleases that promote transposon maintenance and spread

We first identified the *Cbo*TnpB-associated guide RNA (ωRNA) by performing RNA immunoprecipitation sequencing (RIP-seq; Fig. 3A). TnpB strongly enriched a 197-nucleotide (nt) ωRNA substrate, consisting of a 179-nt scaffold derived from the transposon RE and an 18-nt guide derived from the 3′ flanking sequence. We were surprised to detect a strong shoulder in the read coverage located 42 nt downstream of the start site for the ωRNA CM, suggesting the possibility of intramolecular processing (Fig. 3B). When we repeated the experiment using a TnpB variant containing an inactivating mutation in the RuvC nuclease domain (dTnpB), the shoulder was entirely abolished,

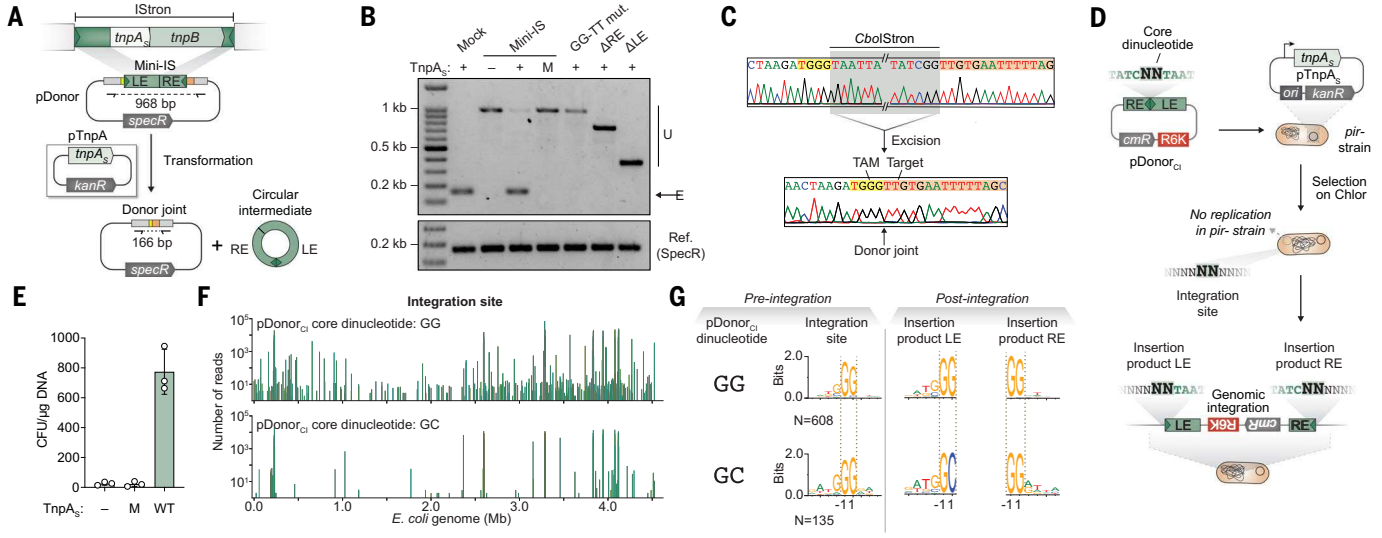


Fig. 2. *Cbo*TnpA$_S$ catalyzes efficient IStron excision and integration, with distinctive dinucleotide requirements. (**A**) Schematic of transposon excision assay using a TnpA$_S$ expression plasmid (pTnpA$_S$) and IStron donor plasmid harboring a minimal transposon (Mini-IS) with LE and RE boundaries (pDonor). Expected substrates and products of transposon excision generated by PCR are indicated, as are the primer binding sites. (**B** and **C**) Gel electrophoresis (B) and Sanger sequencing (C) of PCR products from (A), demonstrating that TnpA$_S$ is active in excising the IStron. Cell lysates were tested with the indicated substrates, which included mutants with mismatched dinucleotides (LE: 5′-GG-3′; RE: 5′-TT-3′) and RE or LE deletions. Mock denotes a positive excision control; U and E refer to unexcised and excised products, respectively; and M denotes an S67A (Ser^{67}→Ala) TnpA$_S$ mutant. In (C), the rejoined TAM and ωRNA-matching target are highlighted in yellow and orange, respectively. (**D**) Schematic of transposon integration assay using pTnpA$_S$ and an IStron circularized intermediate donor plasmid harboring abutted LE and RE sequences (pDonor$_{CI}$). With this suicide vector, transposon integration events are enriched by chloramphenicol selection and deep-sequenced using TagTn-seq (see Materials and methods). (**E**) Cell viability data from experiments in (D), plotted as CFU per μg DNA, when cells contained either empty vector (–), mutant S67A (M), or WT TnpA$_S$. Data are means ± SD (n = 3). (**F**) Genome-wide distribution of TagTn-seq reads from experiments in (D) using WT TnpA$_S$, mapped to the *E. coli* genome. Data are shown for pDonor$_{CI}$ substrates containing either a GG (top) or GC (bottom) dinucleotide. (**G**) Meta-analyses of target-site preferences and integration product dinucleotides at the LE and RE junctions for the genome-wide insertion data with GG and GC dinucleotide substrates shown in (F); the number of unique integration sites is indicated. The preferred genomic target motif is GG for both substrates, but NGS across the LE and RE junctions clearly reveals that noncanonical dinucleotides in the pDonor$_{CI}$ template correspond to noncanonical dinucleotides at the LE junction upon recombinational integration.

implicating TnpB in enzymatic processing of the precursor transcript comprising its own mRNA fused directly to the ωRNA (Fig. 3B). Similar results were recently described for other TnpB homologs, suggesting a generalizable role of transposon-encoded nucleases in ωRNA biogenesis (*50*). Whether a similar feature is also true for related bacterial IscB and eukaryotic Fanzor nucleases remains to be determined.

Based on prior work with IS605-encoded TnpB homologs (*7*, *8*), we hypothesized that IS607-encoded *Cbo*TnpB would recognize and cleave the scarless donor joint generated upon TnpA$_S$-mediated IStron excision, in which the TAM and guide-matching target region are directly adjoined (Fig. 3C). Using a plasmid interference assay, in which successful targeting results in cell lethality under antibiotic selection, we found that TnpB was highly active for RNA-guided DNA cleavage of the scarless donor joint, causing a ~10^5-fold decrease in colony-forming units (CFUs) (Fig. 3D). This reaction strictly required guide RNA–target DNA complementarity, a cognate TAM, and an intact RuvC active site and was similarly effective whether we used a synthetic expression vector with separate promoters for *tnpB* and ωRNA or a native-like *Cbo*IStron (lacking *tnpA*) with single transcript–derived *tnpB*-ωRNA (Fig. 3D).

Transposition assays revealed that *Cbo*TnpA$_S$ specifically targets 5′-TGGG-3′ sequences during DNA integration (Fig. 2G), and we hypothesized that *Cbo*TnpB would recognize a similar TAM motif during RNA-guided DNA cleavage. We tested this using a degenerate TAM library target plasmid in a cellular DNA cleavage assay, in which cleaved molecules are strongly depleted under selective growth conditions (fig. S6A). These experiments revealed a robust 5′-TGGG-3′ TAM (Fig. 3E) and precisely delineated the ωRNA scaffold-guide boundary (fig. S6, B to G), confirming that self-cleavage of the transposon RE harboring a 5′-TCGG-3′ motif is only avoided by stringent discrimination of cytidine at the –3 position underlined. Along with previous work with IS605 elements (*7*–*9*), these results highlight another notable example wherein two distinct enzymes, the transposase (TnpA$_S$) and the nuclease (TnpB), have converged to recognize the same DNA motif.

We next set out to investigate whether *Cbo*TnpB prevents permanent transposon loss that is otherwise inevitable owing to the role of TnpA$_S$ in catalyzing scarless transposon excision, using our previously developed blue-white assays (*7*). First, we inserted the native *Cbo*IStron into *lacZ* in the antisense orientation, to avoid confounding effects of splicing, and monitored transposon excision through regeneration of a *lacZ*$^+$ phenotype (white-to-blue color change) in the presence of wild-type (WT) or mutant TnpA$_S$ or TnpB (Fig. 3F). TnpA$_S$ drove extensive and rapid excisional transposon loss, leading to 47% of colonies regaining a blue phenotype, and this frequency was suppressed 10-fold in the presence of WT, but not nuclease-dead, TnpB (Fig. 3G); these results were further corroborated by diagnostic PCR (fig. S7A). To prove that transposon maintenance was the consequence of donor-joint cleavage followed by homologous recombination, we monitored the ability of TnpB-ωRNA to specifically mobilize the *Cbo*IStron into empty donor joints present in the *E. coli lacZ* locus (fig. S7B). The presence of WT TnpB nuclease strongly reduced the overall transformation efficiency, as expected, and a substantial proportion (>80%) of surviving colonies indeed represented faithful recombination products by their white-colony

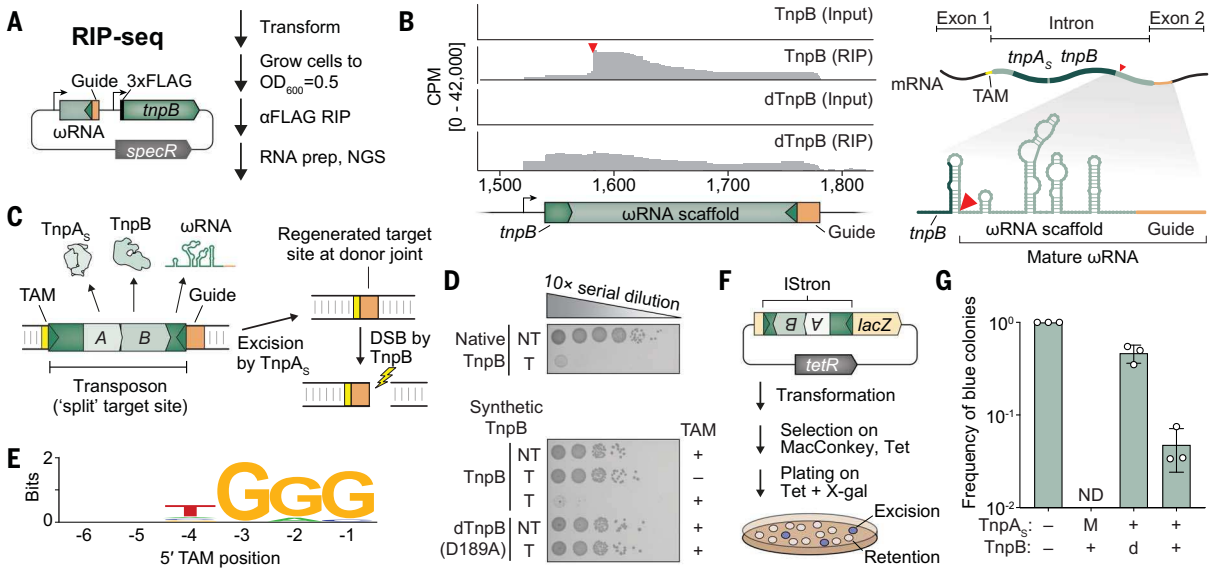


Fig. 3. *Cbo*TnpB is a potent RNA-guided nuclease that prevents *Cbo*TnpA$_S$-mediated transposon extinction. (**A**) Schematic of RIP-seq workflow to uncover RNA binding partners of *Cbo*TnpB using the pEffector shown. (**B**) RIP-seq read coverage for experiments with WT TnpB and RuvC-inactivated dTnpB (D189A; Asp^{189}→Ala) mapped to pEffector (left), alongside associated input controls. The pre-ωRNA processing site is indicated with a red triangle, in both the graph and the RNA schematic shown at the right. The green region labeled "*tnpB*" corresponds to the 3′ end of the ORF. (**C**) Schematic showing the regenerated target site that is produced upon transposon excision, with abutted TAM and target site. This donor joint is predicted to be recognized and cleaved by TnpB. (**D**) Bacterial spot assays demonstrate that TnpB is highly active for RNA-guided DNA cleavage of the donor joint, as assessed by plasmid interference assays. Cells expressing TnpB from a native IStron or synthetic expression plasmid context were transformed with a target-containing (T) or non–target-containing (NT) plasmid, transformants were serially diluted (10×), plated on selective media, and cultured at 37°C for 24 hours. Additional controls included a mutant TAM (where "–" is 5′-ACCC-3′) or RuvC-inactive (D189A) dTnpB. (**E**) Results from a TAM library cleavage assay using a WT ωRNA, revealing that *Cbo*TnpB requires a consensus 5′-TGGG-3′ TAM for efficient DNA cleavage. The WebLogo was generated using the 20-most-depleted sequences after deep sequencing pTarget from surviving colonies (see fig. S6). (**F** and **G**) Schematic of the assay to measure transposon fate in *E. coli* with TnpA$_S$ and TnpB-ωRNA (F) and bar graph showing the frequency of transposon excision or retention for each condition (G), quantified by blue-white colony screening. *Cbo*IStron variants [TnpA mutant S67A (M), nuclease dead TnpB (d), or WT] were inserted at a compatible TAM in a plasmid-encoded *lacZ* and used for transformation of a Δ*lacZ* strain. White or blue colonies indicate transposon retention or excision, respectively. Data are means ± SD (n = 3); "–" indicates and empty vector lacking IStron, TnpA, and TnpB that encodes uninterrupted *lacZ*; and ND is not detected.

phenotype (fig. S7C). This effect was completely ablated by nuclease-dead TnpB, confirming the critical role of TnpB-induced DSBs in the process. We also performed recombination assays in *E. coli* knockout strains that lacked specific DNA repair factor genes to test the hypothesis that transposon cut-and-copy maintenance would depend on classical DSB recombination proteins. We observed that *recA* and *recB* were essential to obtain the white-colony phenotype, confirming the important role of the host homologous recombination machinery in transposon preservation at the donor site (fig. S7D).

Collectively, these experiments powerfully expand our paradigm for the essential role of TnpB in promoting transposon survival to one of the most pervasive transposon families found in nature—IS607.

Group I intron splicing levels are regulated by TnpB and ωRNA structure

Compared with simpler IS605 and IS607 transposons, IStrons are distinctive in their predicted ability to self-splice at the RNA level and rejoin flanking exonic segments, which would presumably render their presence—at the DNA level—phenotypically silent. The *Cbo*IStron displays all of the expected secondary structural elements for group I introns (*51*, *52*), and our analyses tentatively predicted the 5′ splice site (5′SS) and 3′ splice site (3′SS) as perfectly aligning with the transposon LE and RE boundaries, such that the DNA donor joint would match the spliced mRNA exon-exon joint (Fig. 1C). We set out to empirically investigate splicing by developing a cellular assay in which unspliced and spliced products are directly detected by reverse transcription PCR (RT-PCR) (Fig. 4A). We detected the dual presence of both species using a minimal IStron substrate that lacked both *tnpA* and *tnpB*, and the spliced product exhibited scarless joining of the two flanking exons (Fig. 4B), with the exact same nucleotide connectivity as the TnpA$_S$-mediated DNA donor joint (Figs. 2C and 4C). Splicing was abolished either by removing the P7-P9 loop region, which contains the intron catalytic site (Fig. 4B and fig. S3D), or by introducing point mutations within the TAM upstream of the 5′SS (Fig. 4B). This result demonstrates another critical importance of this short motif, underscoring its role in TnpA-mediated DNA transposition, TnpB-mediated DNA cleavage, and intron-mediated RNA splicing. Importantly, we observed the same splicing products from in vitro biochemical reactions, indicating that no protein factors were required (fig. S8A).

Unspliced transcripts produced from native *Cbo*IStrons presumably satisfy numerous functions, of which some are mutually compatible and others are mutually exclusive. Both unspliced and spliced introns can serve as protein-coding mRNAs for TnpA$_S$ and TnpB expression, but spliced introns can no longer act as functional ωRNAs because the splicing reaction severs the ωRNA scaffold and guide (Fig. 4, A and D). Additionally, TnpB-mediated ωRNA processing (Fig. 3B) would segregate the 5′SS and 3′SS onto two distinct molecules, rendering only trans-splicing possible, and TnpB-ωRNA binding would also likely obstruct physical access to the 3′SS. We therefore hypothesized that the structure of the intron RNA 3′ end, the presence of TnpB, or both would regulate the relative efficiencies of splicing and RNA-guided DNA cleavage and set out to

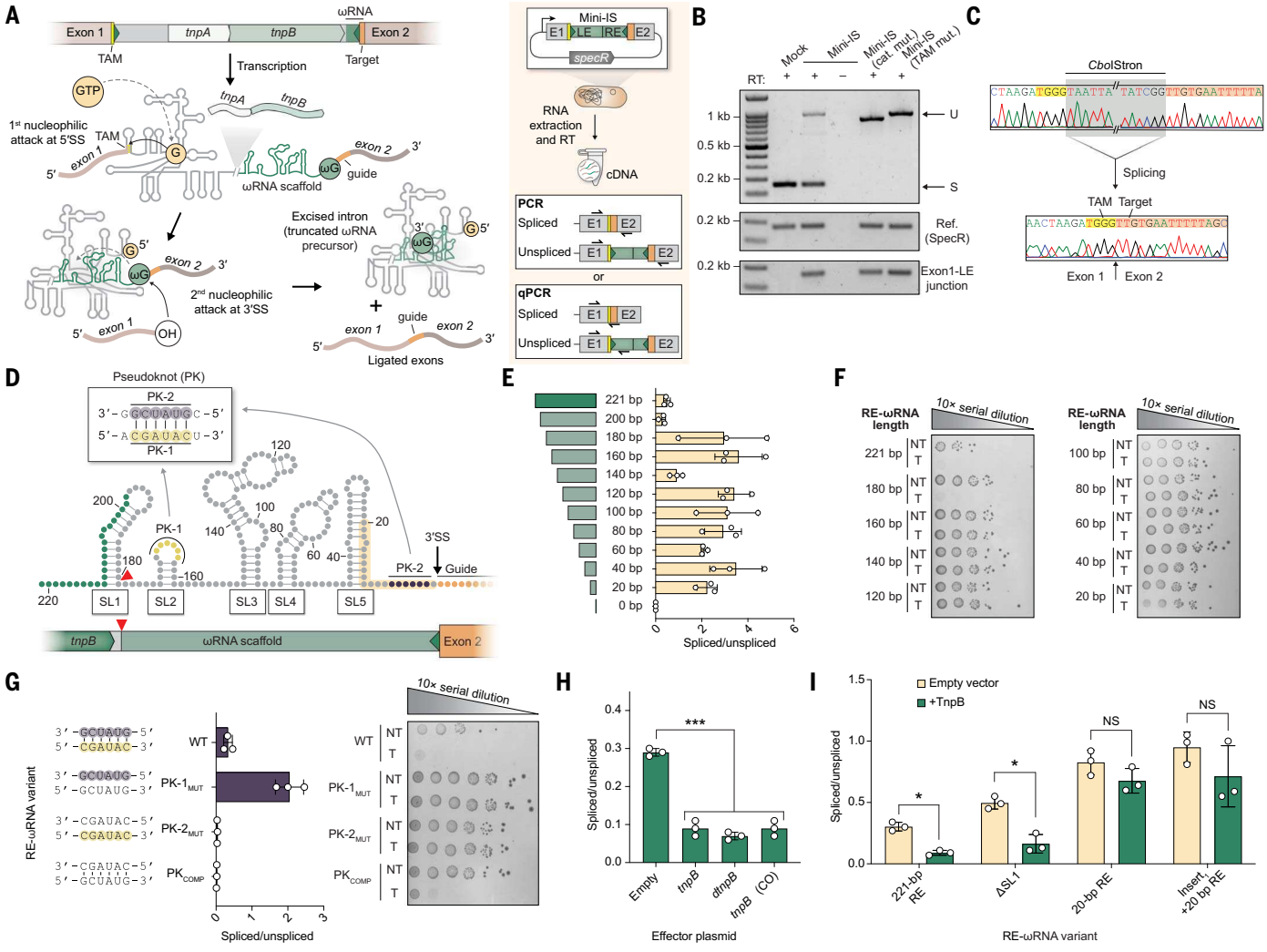


Fig. 4. *Cbo*IStron-encoded introns regenerate transposon-free transcripts while competing with TnpB-ωRNA activity. (**A**) Schematic of general IStron splicing mechanism (left) and *E. coli*–based cellular splicing assay (right). Spliced and unspliced products were detected and/or quantified by RT-PCR and RT-qPCR, respectively, using the primer-pair strategies indicated at the bottom. (**B**) Agarose gel electrophoresis of RT-PCR products from splicing assays in (A) with the indicated constructs, showing unspliced (U) and spliced (S) products (top) relative to reference amplicons for a SpecR drug marker (middle) and exon1-LE junction (bottom). RT indicates reverse transcriptase, "Mock" denotes a positive splicing control, IStron (cat. mut.) contains a P7-P9 loop deletion in the intron catalytic core, and IStron (TAM mut.) contains a 5′-TGTA-3′ TAM that disrupts base pairing required for 5′SS recognition. (**C**) Sanger sequencing of RT-PCR products from (B), for both the unspliced (top) and spliced products (bottom). These sequences are identical to the nucleotide sequences of unexcised and excised DNA sequences shown in Fig. 2C. (**D**) Schematic of predicted *Cbo*IStron ωRNA secondary structure encoded within the transposon RE, with stem-loops, truncation coordinates, and PK motifs labeled. PK-2 overlaps with nucleotides important for 3′SS recognition. The red arrowheads indicate TnpB processing sites, green dots indicate the *tnpB* ORF, and yellow shading indicates the minimal sequence required for efficient splicing. (**E**) RT-qPCR analysis of splicing efficiency for IStron variants in which the RE-ωRNA region was systematically truncated relative to the full-length construct (221 bp). The large splicing change with the 180-bp construct suggests sequence and/or structural features around this position that repress splicing in the full-length design. Data are means ± SD (n = 3). (**F**) Bacterial spot assays for the same RE-ωRNA deletion constructs as in (E), in which RNA-guided DNA cleavage leads to cell death. Cells expressing TnpB were transformed with either a target-containing (T) or non–target-containing (NT) plasmid, and transformants were serially diluted (10×), plated on selective media, and cultured at 37°C for 24 hours. (**G**) RT-qPCR analysis of splicing efficiency (middle) and spot assays to monitor RNA-guided DNA cleavage activity (right) for the indicated PK mutations (left), plotted as in (E) and (F). $PK\text{-}1_{MUT}$ and $PK\text{-}2_{MUT}$ contain mutations to either motif, whereas PK_{COMP} contains compensatory mutations in both motifs. Data are means ± SD (n = 3). (**H**) RT-qPCR analysis of splicing efficiency in the presence of a second effector plasmid harboring *tnpB*, *dtnpB*, or a codon-optimized (CO) *tnpB* gene, revealing the repressive role of TnpB. Empty refers to an empty vector control. Statistical significance was determined using Welch's t test; ***p < 0.001. Data are means ± SD (n = 3). (**I**) RT-qPCR analysis of splicing efficiency in the absence or presence of TnpB for the indicated RE-ωRNA variants. The repressive effect of TnpB on splicing is largely ablated when the ωRNA scaffold is missing (20-bp RE) or replaced with an unrelated sequence ($Insert_1$+20-bp RE). Statistical significance was determined using Welch's t test; *p < 0.05, and NS is not significant. Data are means ± SD (n = 3).

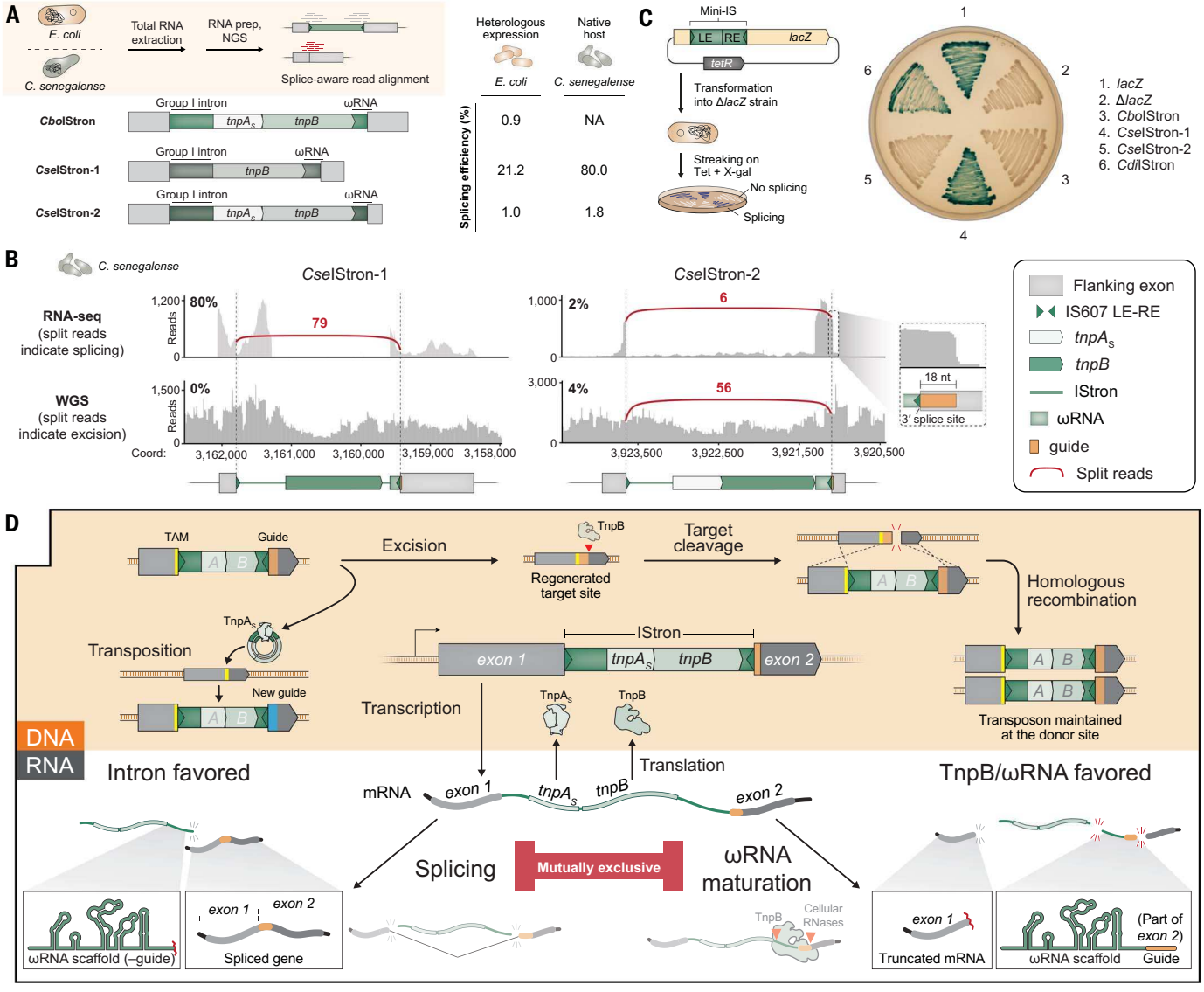


Fig. 5. Competition between intron splicing and TnpB-ωRNA activity establishes a balance between transposon stealth and maintenance. (**A**) Schematic of total RNA-seq (top left) to quantify splicing in *E. coli* (heterologous) or *C. senegalense* (native) for the indicated IS607-family IStrons (bottom left). The observed splicing efficiencies are shown on the right. NA, not available. (**B**) RNA-seq and WGS data for *Csel*IStron-1 and *Csel*IStron-2 in their native context. The number and connectivity of spliced exon-exon junction reads are indicated over the red line, and quantitative comparison of exon-intron and exon-exon reads yields an apparent splicing percentage at the RNA level (RNA-seq graphs, top left) and apparent excision percentage at the DNA level (WGS graphs, top left). These analyses indicate the *Csel*IStron-1 undergoes highly efficient splicing, whereas the inefficient *Csel*IStron-2 splicing (2%) can be fully explained by low-level DNA excision within the bacterial culture, as reported by WGS. The inset at the right shows overlap between the ωRNA and 3′SS, highlighting that most reads extend 18 nt past the 3′SS boundary, which indicates efficient ωRNA maturation. (**C**) Functional splicing assay using blue-white screening. Mini-IS elements from four distinct IStrons were inserted in *lacZ* (sense orientation), such that splicing restores WT *lacZ* expression and a blue-colony phenotype. *lacZ* and *ΔlacZ* controls encoded WT or frame-shifted *lacZ*, respectively. *Cdi*IStron refers to an IS605-family IStron (figs. S2 and S11). (**D**) Overall model for the balanced effects of intron splicing, TnpB-ωRNA, and $TnpA_S$ transposition activity in the maintenance and spread of IS607-family IStron elements. Scarless DNA excision by $TnpA_S$ leads to donor-site transposon loss and risks eventual transposon extinction, without the crucial function provided by TnpB-ωRNA to generate targeted DSBs and trigger homologous recombination to maintain presence of the transposon (top). Unlike canonical transposons, group I intron–containing IStrons mitigate fitness costs on the host by splicing themselves out of interrupted transcripts at the RNA level, thereby restoring functional gene expression (bottom). Splicing and ωRNA maturation are mutually exclusive because splicing severs the ωRNA scaffold and guide sequences and TnpB represses splicing through competitive binding of the 3′SS. The competition between intron splicing and TnpB-ωRNA activity thus serves to regulate the dual objectives of maintaining transposon stealth and promoting transposon proliferation for IStron elements. A similar mechanism is hypothesized for IS605-family IStrons.

test this hypothesis by means of targeted perturbations.

Using a quantitative RT-PCR (RT-qPCR) approach, we first established that deletions to the intron structural core within the LE were deleterious for splicing (fig. S8B). We then introduced RE truncations, starting just upstream of the ωRNA 5′ end, and observed a surprising increase in RNA splicing efficiency for deletions up to 20 bp from the 3′SS (Fig. 4E and fig. S8, B and C). A different pattern emerged when we tested the same truncations

for TnpB-mediated DNA cleavage activity using a plasmid interference assay, because all ωRNA deletions beyond the first stem-loop (SL1) were nonfunctional (Fig. 4F). When we removed specific stem-loops, either individually or in combination, we found that although most ωRNA perturbations notably impaired TnpB cleavage activity, the same perturbations again had a positive effect on splicing efficiency, except for SL5, supporting the conclusion that the terminal 20 bp of the intron are essential for splicing (fig. S8, D to F). The observation that increased splicing efficiencies occurred with RE-ωRNA truncations prompted us to investigate whether this was due to the removal of specific structural features or merely a preference for shorter intron length. We controlled for length by inserting three *lacZ*-derived sequences upstream of the minimal 20-bp RE and found that these constructs were again more active for splicing than the WT *Cbo*IStron, similarly to the minimal splicing variant (fig. S8G). These results support the interpretation that splicing and TnpB-ωRNA activity are inversely correlated, and that RE-ωRNA truncations stimulate splicing in a TnpB-independent manner, suggesting a role for RNA sequence and structural properties in modulating splicing activity.

We were particularly intrigued by a predicted ωRNA PK that is observed in structures of diverse TnpB and Cas12 homologs (*43*–*47*) (fig. S9A), which would sterically occlude nucleotides required for 3′SS recognition by the intron (Fig. 4D). We found that PK formation was essential for TnpB activity because DNA cleavage was eliminated with mutations to either PK-1 or PK-2 but was restored with compensatory mutations to both motifs (Fig. 4G). By contrast, selective mutations to the upstream PK-1 motif substantially increased splicing ratios by sevenfold, presumably by preventing the formation of splicing-incompetent RNA structures. Conversely, mutations to the downstream PK-2 motif, or compensatory mutations to both PK-1 and PK-2, eliminated splicing, likely because of the overlap between these mutations and the sequences required for 3′SS recognition (Fig. 4G). Next, to specifically assess the impact of TnpB on splicing, we compared various expression strategies and established an in trans splicing assay, in which a Mini-IS that encodes the intron and ωRNA, but lacks *tnpA* and *tnpB*, is encoded separately from a TnpB expression vector (fig. S9, B to D). We observed that the expression of either active or catalytically dead TnpB potently inhibited splicing (Fig. 4H), and this inhibition was abolished when the ωRNA scaffold region that is important for TnpB binding was removed (Fig. 4I).

Overall, these findings indicate that binding of TnpB to the ωRNA region of an IStron transcript acts in direct conflict to intron splicing. When considered together with the observation that RE-ωRNA sequence alone depresses splicing efficiency, these results show that both TnpB and RNA-intrinsic properties regulate the efficiency of intron splicing.

Diverse IStrons exhibit various levels of splicing

Finally, we set out to monitor the splicing of native IStrons that encode full-length $tnpA_S$ and/or *tnpB*, focusing on *Cbo*IStron and two related IS*607*-family IStrons from a *Clostridium senegalense* strain that we cultured in-house. *Cse*IStron-1 is a nonautonomous element that encodes only *tnpB*, whereas *Cse*IStron-2 is an autonomous element that shares ~80% sequence identity with *Cbo*IStron (Fig. 5A); the *C. senegalense* strain also harbors nine other IS*607* elements that lack introns (fig. S10A and table S1). When we expressed these IStron elements heterologously in *E. coli* and analyzed splicing by total RNA-sequencing (RNA-seq), we observed low splicing efficiencies (~1%) for both *Cbo*IStron and *Cse*IStron-2, but the splicing efficiency of *Cse*IStron-1 exceeded 20% (Fig. 5A and fig. S10B). When we examined native splicing in *C. senegalense*, we observed even more pronounced differences, with *Cse*IStron-1 exhibiting 80% splicing efficiency compared with just ~2% for *Cse*IStron-2 (Fig. 5B and fig. S10, C and D). Further investigation by whole-genome sequencing (WGS) indicated that the apparent splicing of *Cse*IStron-2 could be entirely attributed to transcription from heterogeneous genomic loci that lost the IS element through DNA excision, a conclusion that was corroborated by additional targeted deep sequencing (Fig. 5B and fig. S10E). By contrast, no evidence for DNA-level loss of *Cse*IStron-1 was observed.

The large discrepancy between native splicing efficiencies of *Cse*IStron-1 and CseIStron-2 allowed us to retrospectively test our model for the antagonistic relationship between intron splicing and ωRNA activity. The *Cse*IStron-1 locus exhibited a paucity of read coverage mapping to the ωRNA region, whereas prominent coverage was observed for the *Cse*IStron-2 ωRNA, with the expected ~18-nt guide sequence extending beyond the RE boundary that matched heterologous RIP-seq experiments in *E. coli* (Figs. 3B and 5B). Intriguingly, a large proportion of *Cse*IStron-2 RNA-seq read pairs terminated at the 5′SS, indicative of prematurely aborted splicing reaction products that only progressed through the first nucleophilic reaction at the upstream exon-intron junction (fig. S10D). These data are fully consistent with efficient ωRNA maturation and/or TnpB binding acting in competition with splicing and, in particular, with recognition of the 3′SS.

Finally, we wanted to directly test the functional impact of splicing on protein restoration by cloning IStrons into *lacZ* and monitoring the splicing-dependent regeneration of WT *lacZ* mRNA sequence and resulting β-galactosidase activity. In excellent agreement with RT-qPCR– and RNA-seq–based splicing measurements, we observed a robust blue-colony phenotype with *Cse*IStron-1 and a candidate IStron from *Clostridioides difficile*, a representative of the IS*605*-family IStrons (Fig. 5C, figs. S2 and S11, and supplementary text). By contrast, cells encoding *Cbo*IStron and *Cse*IStron-2 remained white (Fig. 5C), reflecting splicing efficiencies that were not sufficiently high to facilitate detectable production of functional β-galactosidase. Collectively, these analyses reveal the complex dynamics of IStron splicing and their impact on gene expression and protein function.

Discussion

This work reveals an intricate balancing act that IStrons evolved to satisfy mutually exclusive functions of transposon-encoded noncoding RNAs (Fig. 5D). Efficient self-splicing of the group I intron is crucial for these elements to access a larger targetable space in the genome, because transposon insertions within coding genes, even those essential to the host, can occur without prohibitively adverse phenotypic consequences; this conclusion is readily borne out by the genetic context of diverse IStrons found in Clostridial genomes (*26*–*30*) (Fig. 1C and figs. S2C and S3B). Yet the intron 3′SS occurs precisely between the ωRNA scaffold and targeting sequence, such that splicing effectively severs the head (guide) from the body (scaffold) of the guide RNA (Fig. 5D). Without a functional ωRNA, TnpB is unable to serve its role in transposon maintenance through targeted cleavage of empty donor joints formed by TnpA. Thus, IStron transcripts must exhibit the ability to exist in both splicing-competent and ωRNA-competent states, with the relative proportion dictating the balance between fitness cost maintenance (RNA-driven intron splicing) and selfish transposon preservation (RNA-guided DNA cleavage).

Our results highlight the separate roles of TnpB and RNA structure in establishing this sensitive equilibrium. TnpB-ωRNA binding acts to physically obstruct the second step of splicing (Fig. 4, H and I), preventing the upstream exon from accessing the scissile phosphate at the 3′SS through steric occlusion. The RNA alone also regulates its own fate through a critical PK interaction within the ωRNA, which competes for direct binding to the 3′SS and thereby represses splicing (Fig. 4, D and G). By systematically mutagenizing the *Cbo*IStron RE-ωRNA region and measuring the effects on splicing and TnpB activity, we were able to provide evidence for the existence of conflicting conformational states within IStron transcripts. In vivo data from native IStron elements in *C. senegalense* further corroborated the inverse relationship between intron splicing and ωRNA

biogenesis (Fig. 5B). The substantial differences in the behaviors of two *Cse*IStrons suggest that each element exists in a delicately adjusted balance favoring either ωRNA maturation or splicing. This adaptation may preserve the IStron in a specific insertion site by producing a functional TnpB-ωRNA complex or lower the burden on the host by efficiently restoring the gene disrupted by IStron insertion. Such fine-tuning could result from mutations in the element or transcriptional regulation at the site of insertion, highlighting the complex interplay between IStrons and their host genomes in ensuring gene functionality. Some ancestral group I introns, which do not co-occur with transposable elements (Fig. 1D), encode a riboswitch at the 5′ end of the intron that allosterically regulates its splicing activity, highlighting the diverse regulatory mechanisms that govern splicing (*53*). IStron splicing will be similarly governed by dynamic features that depend on sequence and context, though a complete understanding will require more comprehensive structural, mutational, and chemical probing approaches.

On the surface, co-occurrence of group I introns with IS*605*- and IS*607*-family transposons appears ill-suited to satisfy their selfish needs, because autonomous functions of the intron RNA are subverted by the conflict with ωRNA. The well-studied co-option of homing endonucleases also provides a seemingly effective dispersion strategy for group I introns (*36*–*38*) because these DNA endonucleases can be encoded and expressed from within the intron without disrupting access to the 5′SS or 3′SS. But in fact, IStrons provide far greater flexibility in mobilization, because the association with diverse tyrosine-family ($TnpA_Y$) and serine-family ($TnpA_S$) recombinases ensures even greater opportunities for broad intron dissemination (Fig. 1A), without being limited to homologous sites targeted by homing endonucleases. Moreover, the programmable nature of TnpB nucleases not only enables homing to each new insertion site but also addicts the cell to maintaining the intron at its donor site (Figs. 3 and 5D). This capability is distinctively enabled by the elegant overlap between the transposon LE-RE boundaries at the DNA level and the intron 5′SS-3′SS boundaries at the RNA level, ensuring a perfect correspondence between the nucleotide sequence inserted during DNA transposition and removed again during RNA splicing (Figs. 2C and 4C). Our comprehensive bioinformatics survey of group I introns provides a compelling view of the diverse strategies that these ancient ribozymes have adopted over evolutionary timescales (Fig. 1D) and may reveal previously unknown genetic associations with additional enzymatic modules in the future. Interestingly, transposons are also capable of mediating the spread of spliceosomal introns in eukaryotes (*54*, *55*), suggesting that the synergy of transposition and splicing spans the cross-kingdom boundary.

Another key finding of our study relates to the central role of TnpB in IS*607* transposon preservation. Earlier work from the Zhang and Siksnys labs identified TnpB as an RNA-guided DNA endonuclease (*8*, *9*), and our recent study proved that this biochemical activity promotes selfish spread of IS*605* transposons by triggering a composite transposition pathway, together with the $TnpA_Y$ transposase, that we termed peel-and-paste, cut-and-copy (*7*). In this work, we demonstrate that TnpB plays an equally crucial function in promoting the spread of IS*607*-family transposons (Fig. 3, F and G, and fig. S7), one of nature's most ubiquitous selfish genetic elements found in all three domains of life (*10*, *17*, *56*). Moreover, we demonstrate that this process is dependent on host DNA recombination proteins, namely RecA and RecB. IS*605* and IS*607* transposons encode phylogenetically distinct transposases and mobilize through different intermediates, but they both suffer the same risk of permanent loss owing to their scarless mechanism of excision that is uncoupled from integration (Fig. 1A). Notably, both transposon families encode structurally similar ωRNA guides that overlap the transposon RE, despite requiring vastly different DNA sequences and structures to satisfy transposase-transposon recognition constraints (Fig. 1A and fig. S3, E and F). This flexibility, together with the convergent specificity for overlapping target and transposon-adjacent motifs during DNA integration and DNA cleavage by TnpB and TnpA, respectively (Figs. 2 and 3), has ensured functionally intertwined coupling of enzymatic machineries that catalyze transposon spread and transposon retention. It is noteworthy that TnpB evolved disparate preferences for T/A-rich and G-rich TAMs during its evolution with IS*605* and IS*607* transposons, respectively, likely predating the later PAM diversification that arose for Cas12 nucleases during the evolution of antiviral adaptive immunity (*57*–*61*).

Insertion sequence (IS) transposons were initially defined as short DNA segments that encode only the enzymes necessary for their transposition (*3*). IStrons provide a notable counterexample, showcasing the multiple overlapping biochemical reactions that all contribute to the complex behavior of a transposable element, with as much attention focused on DNA preservation and phenotypic silencing as on transposition itself. In turn, IStrons—and IS*605*- and IS*607*-family transposons, more generally—highlight the specific molecular functionalities that have arisen during co-evolution of distinct machineries encoded within mobile genetic elements. It is through IS*607*-family transposons that TnpB nucleases likely gained access to the eukaryotic domain, giving rise to homologous proteins called Fanzors, with viruses likely serving as a potent mode of transmission (*10*–*13*). In addition, in the ensuing diversification of eukaryotic Fanzors, coevolutionary relationships have yielded fascinating associations with new molecular actors, including *piggyBac*, *Tc1/mariner*, and *Helitron* transposases, alongside other as-yet-uncharacterized genes (*10*, *13*). It appears likely that new biochemical activities and biological functions will continue to be uncovered within this fascinating context.

Materials and methods

Protein detection and database curation

TnpB

The TnpB database was previously curated (*7*). As described therein, homologs were comprehensively detected using the *Helicobacter pylori* TnpB (*Hpy*TnpB) amino acid sequence (NCBI accession no. WP_078217163.1) and the *G. stearothermophilus* TnpB amino acid sequence (NCBI accession no. WP_047817673.1) as seed queries for two independent iterative JackHMMER (HMMER suite v3.3.2) searches against the NCBI nonredundant (NR) database (retrieved on 11 June 2021), with an inclusion and reporting threshold of 1×10^{-30}. The union of the two searches was taken, and proteins smaller than 250 amino acids long were removed to trim partial or fragmented sequences, resulting in a database of 95,731 nonredundant TnpB homologs. Contigs of all putative *tnpB* loci were retrieved from NCBI for downstream analysis using the Bio.Entrez package.

$TnpA_Y$ and $TnpA_S$

For *tnpB*-associated contigs, $tnpA_Y$ was detected using the Pfam Y1_Tnp (PF01797) model for a HMMsearch from the HMMR suite (v3.3.2), with an E-value threshold of 1×10^{-4}. This search was performed on the curated coding sequences of each contig from NCBI. IS elements that encoded *tnpB* within 1 kb of a detected $tnpA_Y$ were defined as autonomous. Analysis of $tnpA_S$ association with *tnpB* was performed with the same methodology mentioned above but using the serine resolvase Pfam model (PF00239).

Arc-like ORF

A manually identified Arc-like protein (NCBI accession no. WP_003367503.1) was used as the seed query in a two-round PSI-BLAST search against the NR database (retrieved on 17 August 2023). A neighborhood analysis was conducted on open reading frames (ORFs) within 10 kb of all detected Arc-like ORF loci using HMMscan from the HMMR suite (v3.3.2) with the Pfam database of HMMs (retrieved on 29 June 2021), and TnpB homologs were specifically identified using the TnpB-specific models produced from the JackHMMER performed in Meers *et al*. (*7*). High-frequency associations with

Arc-like ORFs were manually inspected, and putative functional associations were manually annotated.

Noncoding RNA covariation analyses

Group I intron

The initial search for group I introns associated with *tnpB* was performed using models of available subclasses from the Group I Intron Sequence and Structure Database (GISSD), refined by Zhou *et al.* (*51*) and Nawrocki *et al.* (*62*). The 14 group I intron subclass models were searched against all identified *tnpB*-associated contigs with cmscan (Infernal v1.1.4). A liberal minimum bit score of 15 was used in an attempt to capture distant or degraded introns, and the identification of a putative IStron was supported by its proximity, orientation, and relative location to the nearest identified *tnpB* ORF. The remaining intron hits were considered associated with *tnpB* if they were upstream, on the same strand, and within 1 kb of a *tnpB* ORF. After manually inspecting the database of models and the boundaries of hits, only the catalytic subdomains of the intron were captured, resulting in the poor identification of other substructures both upstream and downstream of the hit. To address this, the boundaries of the group I intron found to be associated with *tnpB* genes were refined and used to generate a more accurate, comprehensive covariance model. A new model was built using a set of 103 *tnpB* loci, including the *C. botulinum* IStron experimentally tested in this study, and closely related loci. To build the new model, 1.5 kb of sequence upstream of the *tnpB* gene was extracted and clustered by 99% length coverage and 99% alignment coverage using CD-HIT (*63*) (v4.8.1) to remove identical sequences. The resulting sequences were aligned using MAFFT (*64*) (v7.508) with the E-INS-I method for eight iterations. The 5′ boundary of the intron (and LE of the IStron) was manually identified as the position of notable drop-off of sequence identity in the alignment; sequences were subsequently trimmed to that boundary. A structure-based multiple alignment was then performed using mLocARNA (*65*) (v1.9.1), with the following parameters:

–max-diff-am 25 –max-diff 60 –min-prob 0.01 –indel –50 –indel-open –750 –plfold-span 100 –alifold-consensus-dp

The resulting alignment with structural information was used to generate a new group I intron covariance model with the Infernal suite and was verified and refined with R-scape at an E-value threshold of 1×10^{-5}. The resulting covariance model was used with cmsearch to discover new group I introns within our curated *tnpB*-associated contig database. The resulting sequences were aligned to generate a new CM model that was used to again search our *tnpB*-associated contig database. After refinement, the final group I intron CM model was searched against the entire nucleotide database from NCBI (retrieved on 29 August 2023) with a higher bit-score of 40.

ωRNA

The initial boundaries of the ωRNA associated with IStron *tnpB* genes were identified in Meers *et al.* (*7*). To refine these models so that structures were more representative of both IS605- and IS607-family IStrons, we extracted sequences 200 bp downstream and 50 bp upstream of the last nucleotide of *tnpB* ORFs to define the RE and transposon boundaries. The ~250-bp sequences were clustered by 99% length coverage and 99% alignment coverage using CD-HIT to remove duplicates. The remaining sequences were then clustered again by 95% length coverage and 95% alignment coverage using CD-HIT. This was done to identify clusters of sequences that were closely related but not identical, which is expected of IS elements that have recently mobilized to new locations. For the 100 largest clusters, which all had a minimum of 10 sequences, MUSCLE (*66*) (v3.8.1551) was used to align each cluster of sequences with default parameters. Then, each cluster alignment was manually inspected for the boundary between high and low conservation, or where there was a stark drop-off in mean pairwise identity over all sequences. This coordinate was annotated for each cluster as the putative 3′ end of the IS element. If there was no conservation boundary, sequences in these clusters were expanded by another 150 bp, to capture the transposon boundaries, and realigned. The consensus sequence of each alignment (defined by a 50% identity threshold up until the putative 3′ end) was extracted, and rare insertions that introduced gaps in the consensus were manually removed. With the 3′ boundary of the IS element, and thus the 3′ boundary of the *tnpB* ωRNA properly defined, a covariance model of the ωRNA for any *tnpB* clade of interest could now be built.

A 200-bp window of sequence upstream of the 3′ end for elements of the *Cbo*IStron clade and the *Cdi*IStron clade was extracted. A structure-based multiple alignment was then performed using CMfinder (*67*) (v0.4.1.9) with the following parameters:

-skipClustalw -minCandScoreInFinal 10 -combine -fragmentary -commaSepEmFlags x–filter-non-frag,–max-degen-per-hit,2,–max-degen-flanking-nucs,7,–degen-keep,–amaa

The resulting structural motifs were then used to generate either a *Cbo*IStron- or *Cdi*IStron-specific ωRNA covariance model with Infernal and were refined and verified with R-scape (*68*) at an E-value threshold of 1×10^{-5}. Each model was used to search against our *tnpB*-associated contig database with cmsearch at a bit-score of 40 to identify all structurally related ωRNA. ωRNA hits were considered associated with *tnpB* if they were downstream, on the same strand, and within 200 bp of a *tnpB* ORF. The alignment of hits from this search were refined, verified, and visualized with R-scape.

Phylogenetic analyses

TnpB and Arc-like ORF

For *tnpB* genes found in putative IStron elements, protein sequences were clustered at 95% length coverage and 95% alignment coverage using CD-HIT. The clustered representatives were taken and aligned using MAFFT with the E-INS-I method for 16 rounds. Postalignment cleaning consisted of using trimAl (*69*) (v1.4.rev15) to remove columns containing more than 99% of gaps and manual inspection. The phylogenetic tree was created using IQ-Tree 2 (*70*) (v2.2.3) with a model of substitution identified using ModelFinder (*71*) (JTTDCMut+F+R10) and optimized trees with nearest neighbor interchange to minimize model violations. Branch support was evaluated with 1000 replicates of SH-aLRT, aBayes, and ultrafast bootstrap support (*72*) from the IQ-Tree 2 package. The tree with the highest maximum likelihood was used as the reconstruction of the IStron *tnpB* phylogeny.

All Arc-like ORF hits were aligned using MAFFT with the E-INS-I method for eight rounds. The rest of the analysis was identically performed as above. ModelFinder identified LG+I+R5 as the best-fit model.

Group I introns

After the search of the NT database using an updated covariation model, hits smaller than 300 bp were removed. The remaining sequences were clustered at 90% length coverage and 90% alignment coverage using CD-HIT. The clustered representatives were taken and aligned using MAFFT with the E-INS-I method for two rounds. Postalignment cleaning consisted of using trimAl (v1.4.rev15) to remove columns containing more than 99% of gaps and manual inspection. The phylogenetic tree was created using IQ-Tree 2 (v2.1.4) with a model of substitution identified using ModelFinder (GTR+F+R10), and optimized trees with nearest neighbor interchange to minimize model violations. Branch support was evaluated with 1000 replicates of SH-aLRT, aBayes, and ultrafast bootstrap support from the IQTREE package. The tree with the highest maximum likelihood was used as the reconstruction of the group I intron phylogeny. Neighborhood analysis was performed similarly to how the Arc-like ORFs were analyzed. Annotation of rRNA context was extracted from GenBank annotations.

Generating structures of group I intron RNA structures

The identification of typical group I intron domains was achieved through sequence

alignment to the previously characterized *Cdi*IStron (CdISt1) (*27*). Predicted secondary RNA structures were generated using the Mfold web server and visualized with the use of RNA canvas (*73*).

Culturing of C. senegalense

A *Clostridium* strain encoding IStrons with ~80% similarity to the *Cbo*IStron experimentally characterized in this study was obtained from ATCC (strain 25772), where it was defined as belonging to an unknown species classification. Internal rRNA phylogenetic analysis led to the assignment of this strain as a member of species *senegalense*. *C. senegalense* was cultured from a lyophilized ATCC pellet in 5 ml of gifu anaerobic medium broth, modified (mGAM; HyServe) in an anaerobic chamber (5% H_2, 10% CO_2, and 85% N_2). All media was prereduced for ~24 hours before use in culturing. *C. senegalense* was then banked as a glycerol stock (final concentration 20%) and subcultured into 100 ml cultures of mGAM. The growth of these cultures was monitored with a spectrophotometer over ~6 hours until a final optical density at 600 nm (OD_{600}) of 0.4 to 0.6 was reached (exponential phase), at which point cultures were poured into two 50-ml conical tubes and cooled on ice for 10 min. The cultures were then centrifuged at 4000*g* for 10 min at 4°C, the supernatant was decanted, and cell pellets were flash frozen in liquid nitrogen. Pellets were stored at −80°C until RNA extraction and processing.

RNA extraction

RNA from *C. senegalense* cell pellets were extracted in 96-well format using a silica-bead beating-based protocol adapted from a prior study (*74*). Briefly, 200 μl of 0.1-mm zirconia silica beads (Biospec,) were added to each well of 96-well deep-well plates (Thermo Fisher Scientific). Next, cell pellets were resuspended in 500 μl of DNA/RNA shield buffer (Zymo Research) and transferred to separate wells, and the plates were affixed with a sealing mat and centrifuged for 1 min at 4500*g*. To avoid overheating, the plates were vortexed for 5 s and incubated at −20°C for 10 min before beating. Then, plates were fixed on a bead beater (Biospec) and subjected to bead beating for 5 min, followed by a 10-min cooling period. The bead-beating cycle was repeated three times, and plates were then centrifuged at 4500*g* for 5 min to remove cell debris. Next, 60% of the bead-beating volume was transferred to the Quick-RNA Miniprep Plus kit (Zymo Research), and RNA was purified using the manufacturer's protocol for Gram-positive bacteria. RNA quality was assessed using the 260/280 nm ratio (~2.0) as measured by Nanodrop, and concentration was measured by the Qubit RNA High Sensitivity Assay Kit (Thermo Fisher Scientific) using the manufacturer's protocol. RNA was stored at −80°C until library preparation.

Total RNA-seq

For total RNA-seq library preparation, 10 μg of purified RNA was treated with Turbo DNase I (Thermo Fisher Scientific) for 1 hour at 37°C using the manufacturer's protocol. A 2× volume of Mag-Bind TotalPure NGS magnetic beads (Omega) was added to each sample, and the RNA was purified using the manufacturer's protocol. The RNA was then diluted in NEBuffer2 [New England Biolabs (NEB)] and fragmented by incubating at 92°C for 1.5 min. To generate RNA with 5′-monophosphate and 3′-hydroxyl ends, samples were treated with RppH (NEB) supplemented with SUPERase•In RNase Inhibitor (Thermo Fisher Scientific) for 30 min at 37°C, followed by T4 PNK (NEB) in 1× T4 DNA ligase buffer (NEB) for 30 min at 37°C. Samples were column-purified using RNA Clean and Concentrator-5 (Zymo Research), and the concentration was determined using the DeNovix RNA Assay (DeNovix). Illumina adapter ligation and cDNA synthesis were performed using the NEBNext Small RNA Library Prep kit. Dual index barcodes were added by PCR amplification (12 cycles), and the cDNA libraries were purified using the Monarch PCR and DNA Cleanup Kit (NEB). High-throughput sequencing was performed on an Illumina NextSeq 550 in paired-end mode with 150 cycles per end.

WGS of C. senegalense

Genomic DNA from *C. senegalense* was extracted using the Promega Wizard Genomic DNA purification kit following the manufacturer's protocol for Gram-positive bacteria. DNA was measured by fluorescent quantification. TnY, a homolog of Tn5, was purified in-house following previous methods (*75*). Ten nanograms of purified genomic DNA (gDNA) was tagmented with TnY preloaded with Nextera Read 1 and Read 2 oligos (table S4) following previous methods (*75*), followed by proteinase K treatment (NEB, final concentration 16 U ml^{-1}) and column purification. PCR amplification and Illumina barcoding was done for 13 cycles with KAPA HiFi Hotstart ReadyMix (table S4), with an annealing temperature of 63°C and an extension time of 1 min. The PCR reaction was then resolved on a gel, and a smear from 400 to 800 bp was extracted for sequencing on a paired end, 150×150 NextSeq kit. Downstream analysis was performed as described in the subsequent section, RNA-seq analyses. De novo genome assembly was also performed by Plasmidsaurus, and the assembled genome was in agreement with the 4-Mb genome provided for ATCC strain 25772.

Targeted tagmentation-based detection of IS excision events

One hundred nanograms of purified gDNA of *C. senegalense* was tagmented with TnY preloaded with full-length Nextera Read 2/Indexed oligos (table S4). An initial PCR amplification was done with a forward oligo that anneals in the upstream genomic sequence flanking the IStron and an oligo that anneals to the P7 sequence (table S4) using KAPA HiFi Hotstart, with an annealing temperature of 55°C and 1-min extension time. After bead cleanup using Omega Mag-Bind TotalPure magnetic beads (Omega Bio-Tek) at a ratio of 0.9×, a second PCR was done with an oligo that annealed to the initial PCR amplicon within ~40 bp of the genome-IStron junction; this forward oligo had all necessary sequences for Illumina sequencing (table S4). After 15 cycles of PCR under the same conditions, the reaction was resolved on a gel, and a smear from 350 to 800 bp was extracted for sequencing with at least 75 Read 1 cycles. After adapter trimming, the relative abundance of reads that contained a 20-bp sequence of the IStron end or contained a 20-bp sequence of the downstream genomic sequence were tallied using BBDuk from the BBTools suite (v.38.00; https://sourceforge.net/projects/bbmap), with a Hamming distance of 2 and an average Q-score greater than 20.

RNA-seq analyses

RNA-seq datasets were either generated as described above (*C. senegalense* RNA-seq) or downloaded from NCBI (*C. difficile* total RNA-seq: SRR14415999; *C. difficile* small RNA-seq: SRR12329944). RNA-seq data were processed using cutadapt (*76*) (v4.2) to remove adapter sequences, trim low-quality ends from reads, and exclude reads shorter than 18 bp. Reads were mapped to the reference genome (*Cdi*: NZ_CP010905.2; *Cse*: ATCC 25772) using the splice-aware aligner STAR (*77*) (v2.7.10), with –outFilterMultimapNmax 10. Mapped reads were sorted and indexed using SAMtools (*78*) v1.17. Splice junctions inferred by STAR flanking loci of interest were used to create a custom genome annotation file for a second round of STAR alignment to refine splice junction quantitation. Sashimi plots showing read coverage and splice events at specific loci were generated with ggsashimi (*79*) (v1.1.5) in strand-specific mode. Additional read coverage plots were generated by converting alignment files to bigwig files with bamCoverage (*80*) (v3.5.1) using a bin size of 1 and extending reads to fragment size. Coverage over selected genomic regions was visualized in Integrative Genomics Viewer (IGV). For analysis of 3′ boundaries of read pairs, uniquely mapping reads were filtered using SAMtools to only include read 2, and mapping coordinates were extracted using the bamtobed utility from

bedtools (*81*) (v2.31.0). The 3′ boundary of each read pair was determined as the start coordinate of read 2, for transcripts on the "–" strand, or the end coordinate of read 2, for transcripts on the "+" strand. To quantify splicing activity at each intron locus, reads were mapped to a mock reference sequence spanning either the 5′ exon1-intron junction, intron-exon2 junction, or the exon-exon junction, using bwa-mem2 (*82*) (v2.2.1). Reads mapping uniquely to each reference sequence were quantified using featureCounts (*83*) (v2.0.2), with a minimum overlap of 3 bp on either end of the junction. Splicing activity was calculated with following equation: splicing activity (%) = N(reads, exon1-exon2)/[N(reads, exon1-exon2) + N_{Av}(reads, exon-intron)] *100%, where N_{Av} = [N(reads, exon1-intron) + N(reads, intron-exon2)]/2.

RIP-seq

E. coli strain K-12 substrain MG1655 (sSL0810) was transformed with plasmids encoding ωRNA and 3×FLAG-CboTnpB (pSL5412) or 3×FLAG-CboTnpB(D189A) (pSL5413). Single colonies were inoculated in liquid LB with spectinomycin (100 μg ml^{-1}) and grown overnight. The next day, the culture was used to inoculated 50 ml of liquid LB with spectinomycin (100 μg ml^{-1}) at 100× dilution and grown until OD_{600} reached 0.5. Ten milliliters of culture was centrifuged at 4000*g* for 10 min at 4°C, and the supernatant was removed. The pellet was washed once with 1 ml of cold tris-buffered saline (TBS) (20 mM Tris-HCl (pH 7.5 at 25°C), 150 mM NaCl) and centrifuged at 10,000*g* for 5 min at 4°C, the supernatant was removed, and the resulting pellet was flash-frozen in liquid nitrogen. Pellets were stored at −80°C. Antibodies for immunoprecipitation were conjugated to magnetic beads as follows: For each sample, 30 μl of Dynabeads Protein G (Thermo Fisher Scientific) were washed 3× in 1 ml RIP lysis buffer [20 mM Tris-HCl (pH 7.5 at 25°C), 150 mM KCl, 1 mM $MgCl_2$, 0.2% Triton X-100], resuspended in 1 ml of RIP lysis buffer, combined with 10 μl of anti-FLAG M2 antibody (Sigma-Aldrich, F3165), and rotated for >3 hours at 4°C. Antibody-bead complexes were washed three times to remove unconjugated antibodies and resuspended in 30 μl of RIP lysis buffer per sample.

To generate cell lysates, flash-frozen pellets were first resuspended in 1.2 ml of RIP lysis buffer supplemented with cOmplete Protease Inhibitor Cocktail (Roche) and SUPERase•In RNase Inhibitor (Thermo Fisher Scientific). Cells were then sonicated for 1.5 min total (2 s on, 5 s off) at 20% amplitude. To clear cell debris and insoluble material, lysates were centrifuged for 15 min at 4°C at 21,000*g*, and the supernatant was transferred to a new tube. At this point, a small volume of each sample (24 μl, or 2%) was set aside as the "input" starting material and stored at −80°C.

For immunoprecipitation, each sample was combined with 30 μl of antibody-bead complex and rotated overnight at 4°C. The next day, each sample was washed three times with ice-cold RIP wash buffer [20 mM Tris-HCl (pH 7.5 at 25°C), 150 mM KCl, 1 mM $MgCl_2$]. After the last wash, beads were resuspended in 1 ml of TRIzol (Thermo Fisher Scientific) and incubated at room temperature for 5 min to allow separation of RNA from the beads. A magnetic rack was used to isolate the supernatant, which was transferred to a new tube and combined with 200 μl of chloroform. Each sample was mixed vigorously by inversion, incubated at room temperature for 3 min, and centrifuged for 15 min at 4°C at 12,000*g*. RNA was isolated from the upper aqueous phase using the RNA Clean and Concentrator-5 kit (Zymo Research) and eluted in 15 μl of RNase-free water. RNA from input samples was isolated in the same manner using TRIzol and column purification.

For RIP-seq library preparation (input and RIP eluates), 6 μl of RNA was diluted in FastAP Buffer (Thermo Fisher Scientific) supplemented with SUPERase•In RNase Inhibitor (Thermo Fisher Scientific) to a total volume of 18 μl and fragmented by heating to 92°C for 1.5 min. Each sample was treated with 2 μl of TURBO DNase (Thermo Fisher Scientific) for 30 min at 37°C and column-purified using the RNA Clean and Concentrator-5 kit (Zymo Research), eluting in 12.5 μl of RNase-free water. RNA concentration was quantified using the DeNovix RNA Assay. Illumina sequencing libraries were prepared using the NEBNext Small RNA Library Prep kit, and libraries were sequenced on an Illumina NextSeq 500 in paired-end mode with 75 cycles per end.

Adapter trimming, quality trimming, and read-length filtering of RIP-seq reads was performed as described above for total RNA-seq experiments. Trimmed and filtered reads were mapped to a reference containing both the MG1655 genome (NC_000913.3) and plasmid (pSL5412) sequences using bwa-mem2 (*84*) (v2.2.1) with default parameters. SAMtools (v1.17) was used to filter uniquely mapped reads (MAPQ > 1) as well as used to sort and index the uniquely mapped reads. Coverage tracks were generated using bamCoverage (v3.5.1) with a bin size of 1, read extension to fragment size, and normalization by counts per million mapped reads (CPM) with exact scaling. Coverage was visualized in IGV.

Sequence identity matrices

Pairwise sequence identity matrices were generated in Geneious from MAFFT alignments of intron or ωRNA nucleic acid sequences, or TnpA or TnpB amino acid sequences, using default settings. For introns, the identity matrices used the conserved structured portion of each IStron, as determined by multiple sequence alignment, and the group I intron boundaries determined by splicing analysis of RNA-seq data or predicted by the intron covariance model. For ωRNAs, the identity matrices used the ωRNA boundaries determined by RNA-seq (for *C. senegalense*), or the 117 nt downstream of TnpB (for *C. difficile*, owing to poor resolution of ωRNA boundaries in the public RNA-seq datasets analyzed in this study). Accession numbers for IStron, TnpA, and TnpB sequences are listed in table S1.

Plasmid and E. coli *strain construction*

All strains and plasmids used in this study are described in tables S2 and S3, respectively, and a subset is available from Addgene. Briefly, genes encoding *Cbo*TnpA and native *Cbo*IStron sequence were synthesized by Twist Bioscience. *E. coli* codon-optimized *Cbo*TnpB, and bioinformatically predicted ωRNA were synthesized and cloned into a single pCDF-Duet vector by Genscript, with two separate J-23 series promoters driving their expression. Transposase expression plasmids were generated using Gibson assembly, by inserting the *tnpA* gene downstream of pLac or T7 promoters in a minimal pCOLADuet-1 vector (*7*). Native IStron, IStron with TnpB only and Mini-IS sequence (581 bp from the LE and 221 bp of the RE) were cloned using Gibson assembly by inserting them into a pCDF-Duet vector downstream of T7 promoter. pTarget plasmids were generated by around-the-horn PCR, by inserting a 44-bp target sequence into a minimal pCOLADuet vector. A transposition circular intermediate ($pDonor_{CI}$) was generated by Gibson assembly of the *Cbo*IStron LE (581 bp), RE (221 bp), R6K ori, and chloramphenicol resistance gene. The cloning mix was used to transform a pir^+ strain (sSL0281, CGSC), to allow for the propagation of the R6K ori-bearing plasmid. Derivatives of these plasmids were cloned using a combination of methods, including Gibson assembly, restriction digestion-ligation, ligation of hybridized oligonucleotides, Golden Gate Assembly, and around-the-horn PCR. Plasmids were cloned and propagated in NEB Turbo cells (NEB)—except for $pDonor_{CI}$ derivatives, which were propagated in strain sSL0281—purified using Miniprep Kit (Qiagen), and verified by Sanger sequencing (GENEWIZ).

E. coli strain K-12 substrain MG1655 (sSL0810) derivatives with removed DNA repair factors were generated using *lambda-red* recombineering strategy, where the gene of interest was replaced with a chloramphenicol-resistance gene cassette. Briefly, the chloramphenicol-resistance gene and its promoter were amplified with overhangs specific to the locus to be modified, PCR purified, and electroporated into sSL0810 cells. After recovery, clonal isolates were selected on chloramphenicol-containing media, and recombination was confirmed by PCR

amplification of the modified region in the genome.

DNA cleavage assays with TnpB

Plasmid interference assays were performed in *E. coli* strain K-12 substrain MG1655 (sSL0810) when a synthetic *Cbo*TnpB expression construct was used, and in *E. coli* strain BL21 (DE3) for all other experiments. When *Cbo*TnpB was coexpressed with ωRNA from the same plasmid, BL21(DE3) cells were transformed with a pEffector plasmid, and single colony isolates were selected to prepare chemically competent cells. Two hundred nanograms of pTarget plasmid were then delivered by transformation. After 2 hours, cells were spun down at 6000 rpm for 5 min and resuspended in 30 μl of LB. Cells were then serially diluted (10×) and plated on LB-agar media containing spectinomycin (100 μg ml^{-1}) and kanamycin (50 μg ml^{-1}), and grown for 24 hours at 37°C. Plates were imaged in an Amersham Imager 600. For experiments when a Mini-IS was used as a guide for *Cbo*TnpB, BL21(DE3) cells were cotransformed with Mini-IS and TnpB expression plasmids, and single colony isolates were selected to prepare chemically competent cells. A second transformation was performed as indicated previously, and cells were plated on LB-agar media containing spectinomycin (100 μg ml^{-1}), chloramphenicol (25 μg ml^{-1}), kanamycin (50 μg ml^{-1}), and IPTG (0.1 mM) and grown for 24 hours at 37°C. Plates were imaged in an Amersham Imager 600.

TAM library experiments and analyses

TAM library experiments were prepared for sequencing, as previously described (*7*). Analyses were also performed as previously described. In brief, reads were filtered on containing the correct sequence both upstream and downstream of the TAM region. TAM sequences were then extracted and tallied, and depletion values were calculated as the relative abundance of the library member in the output library divided by the relative abundance of the library member in the input. Sequence logos were generated with the library members that were depleted more than fivefold (depletion value greater than 32) using WebLogo (*85*) (v2.8), and the top 5% of depleted library members were used to generate TAM wheels (depletion scores were inverted to properly visualize TAM wheels) (*86*).

Transposon excision assays with TnpA$_S$

All transposition experiments were performed in *E. coli* MG1655 strains. Chemically competent cells were first transformed with a plasmid encoding *tnpA* under an inducible *lac* promoter, and transformants were isolated by plating on LB-agar plates with antibiotic (50 μg ml^{-1} kanamycin). Liquid cultures were then inoculated from single colonies, and the resulting strains were made chemically competent using standard methods, aliquoted, and snap-frozen. This strain was then subsequently transformed with 100 ng of a separate pDonor plasmid containing a Mini-IS variant. Cultures were grown overnight at 37°C on LB-agar plates under antibiotic selection (100 μg ml^{-1} spectinomycin, 50 μg ml^{-1} kanamycin) and IPTG induction (0.5 mM). The next day, all colonies were scraped and resuspended in LB medium. To prepare cell lysates, approximately 3.2×10^8 cells (equivalent to 200 μl of culture at OD_{600} = 2.0) were transferred to a 96-well plate. Cells were pelleted by centrifugation at 4000*g* for 5 min and resuspended in 80 μl of H_2O. Next, cells were lysed by incubating at 95°C for 10 min in a thermal cycler. The cell debris was pelleted by centrifugation at 4000*g* for 5 min, and 10 μl of lysate supernatant was removed and serially diluted in H_2O to generate 10- and 100-fold lysate dilutions for PCR and quantitative PCR (qPCR) analyses, respectively.

PCR reactions were performed using primers annealing in 5′ and 3′ IStron-flanking sequences, which can be used to amplify both unexcised loci (longer amplicon) and excised loci (shorter amplicon). Each 20-μl PCR reaction contained 1× OneTaq Master Mix (NEB), 0.2 μM of each primer, and 1 μl of 10-fold-diluted lysate. Thermal cycling conditions included DNA denaturation (94°C for 30 s), 30 cycles of amplification (denaturation: 95°C for 15 s; annealing: 46°C for 15 s; extension: 68°C for 15 s), followed by a final extension (68°C for 5 min). Products were resolved by 1.5% agarose gel electrophoresis and visualized by staining with SYBR Safe (Thermo Fisher Scientific). qPCR was performed in a 10 μl reaction that contained 5 μl of SsoAdvanced Universal SYBR Green Supermix (BioRad), 1 μl of H_2O, 2 μl of primer pair at 2.5 μM concentration, and 2 μl of 100-fold-diluted lysate. Two primer pairs were used: (i) excision products were specifically captured using a forward primer annealing to 5′ IStron flanking sequence (oSL12715) and reverse primer spanning the donor joint (oSL12719), and (ii) a reference gene (*specR*) was amplified with primers that annealed to the same pDonor plasmid (oSL13566 and oSL13567). Excision efficiency (%) was calculated as a ratio between the experimental sample and a lysate from cells transformed with a mock plasmid (mimicking donor joint), using the following formula: excision efficiency (%) = $2^{-\Delta\Delta Cq}$*100%, where ΔΔCq = ΔCq[Cq_{sample}(donor joint) – (reference gene)] – ΔCq[Cq_{mock}(donor joint) – (reference gene)]. Primer sequences are provided in table S4.

TnpA$_S$ transposition intermediate (minicircle) capture using PCR

A transposition assay was performed in *E. coli* strain BL21(DE3) transformed with pDonor plasmid encoding Mini-IS with *cmR* gene as a cargo (pSL5192). Cells were made chemically competent and transformed with plasmids encoding *tnpA* (pSL5006), mutant *tnpA* (pSL5094), or an empty vector (pSL4032). Transformants were plated on LB-agar plates with antibiotic (100 μg ml^{-1} carbenicillin, 50 μg ml^{-1} kanamycin). The next day, several colonies were used to inoculate liquid cultures in LB with 100 μg ml^{-1} carbenicillin and 50 μg ml^{-1} kanamycin and grown overnight. An aliquot of turbid culture equivalent of 200 μl of culture at OD_{600} = 2.0 was taken for lysis. Lysis was performed as described previously, and minicircle junction was amplified using a forward primer annealing on transposon RE, and a reverse primer annealing on the transposon LE (see table S4 for primer sequences). Twenty microliters of PCR reaction contained 1× OneTaq Master Mix (NEB), 0.2 μM of each primer, and 1 μl of 10-fold diluted lysate. Thermal cycling: DNA denaturation (94°C for 30 s), 30 cycles of amplification (denaturation: 95°C for 15 s, annealing: 51°C for 15 s, extension: 68°C for 20 s), followed by a final extension (68°C for 5 min). Products were resolved by 1.5% agarose gel electrophoresis and visualized by staining with SYBR Safe (Thermo Fisher Scientific). The expected junction was confirmed by Sanger sequencing.

Transposon integration assays with TnpA$_S$

Plasmids with an R6K ori, a *CmR* marker, and inverted IStron ends (pDonor$_{CI}$) were cloned in *pir*$^+$ strain (sSL0281). *E. coli* strain K-12 substrain MG1655 transformed with either an empty vector or a TnpA or mutant TnpA expression plasmid driven by pLac was grown in liquid LB with IPTG (0.5 mM), and cells were first made chemically competent. These cells were transformed with a pDonor$_{CI}$ with a canonical GG core dinucleotide. After recovery for 7 hours, cells were plated on LB-agar plates with chloramphenicol (25 μg ml^{-1}) and grown for ~24 hours. The number of surviving colonies were counted across replicates. Surviving colonies from WT TnpA-expressing cells were then scraped for downstream sequencing. To investigate the impact of core dinucleotide modifications, MG1655 cells with a WT TnpA expression plasmid driven by pLac were grown in liquid LB with IPTG (0.5 mM) and made electrocompetent at OD_{600} = 0.6. These cells were then electroporated with pDonor$_{CI}$ variants, recovered for 7 hours, plated on LB-agar plates with chloramphenicol (25 μg ml^{-1}), and grown for ~24 hours. Surviving colonies were pooled, and genomic DNA was extracted and quantified with Qubit. Approximately 100 ng of gDNA was tagmented with TnY pre-loaded with Read 2 Nextera oligos (table S4). Two rounds of PCR were performed as described for targeted detection of IS excision events with oligos that annealed to either the IStron

LE or RE (table S4). Paired-end, 76×76 cycle sequencing was performed on a NextSeq platform. Using BBDuk (*87*), reads were then filtered for containing the proper IStron end sequence, and the flanking genomic sequence was extracted. Reads that contained the parental $pDonor_{CI}$ sequence were removed during this process. Flanking genomic sequences were then aligned to the *E. coli* genome using Bowtie2 (*88*). WebLogo (*85*) representations were generated using the input sequence on both the IStron LE and RE, as well as the mapped genomic insertion sites.

Transposon maintenance experiments with $TnpA_S$ and TnpB

sSL3391, a derivative of *E. coli* strain K-12 substrain MG1655 with *lacZ* deletion replaced by a chloramphenicol-resistance cassette, was transformed with 400 ng of plasmid encoding an intact *lacZ* gene (pSL4825, empty vector) or a CboIStron-interrupted *lacZ* gene (pSL5948, pSL5949, pSL5950; see table S3 for descriptions). After transformation, colonies were plated on MacConkey agar media containing tetracycline (10 μg ml^{-1}) to enrich for IStron excision events. Cells were grown at 37°C for 36 hours, then harvested, serially diluted, plated onto LB-agar containing tetracycline (10 μg ml^{-1}) and X-gal (200 μg ml^{-1}), and grown for 18 hours at 37°C. The total number of colonies were counted, along with the number of blue colonies, to determine the frequency of excision and reintegration events. Colony counts are represented as CFUs per μg of DNA, with experiments performed in biological triplicates. In addition, genomic lysates were harvested from cells, as described in the assay for excision detection, for PCR analysis; a representative gel from a single replicate is shown in fig. S7A.

Transposon recombination assay with $TnpA_S$ and TnpB

E. coli strain K-12 substrain MG1655 (sSL0810) containing an intact *lacZ* loci was chemically transformed with 400 ng of plasmid encoding an intact *lacZ* gene (pSL4825, empty vector) or *Cbo*IStron-interrupted *lacZ* gene (pSL5948, pSL5949, pSL5950; see table S3 for descriptions), recovered for 1 hour at 37°C in liquid LB, and serially diluted on LB-agar plates with tetracycline (10 μg ml^{-1}). The next day, colonies were counted, and the tetracycline plates were replica-plated to LB-agar plates containing both tetracycline (10 μg ml^{-1}) and X-gal (200 μg ml^{-1}) for blue-white colony screening. White colonies were counted to determine the frequency of recombination events at the genomic *lacZ* locus. All colony counts are represented as CFUs per μg of DNA.

To assess the importance of certain DNA repair proteins for recombination, strains with a single gene knockout (sSL4268, sSL4269, sSL4270, sSL3640 see table S2 for descriptions) were chemically transformed with 400 ng of plasmid encoding *lacZ* interrupted by either *Cbo* Mini-IS (pSL6919) or *Cbo*IStron with mutant TnpA (pSL5950). The rest of the assay was performed as described for a WT *E. coli* strain. All recombination experiments were performed in biological triplicates.

In vitro splicing assays

Templates for in vitro splicing reactions were obtained by PCR amplification of mock excised (pSL5516), splicing mutant (pSL5026), and Mini-IS (pSL5515) containing plasmids. All templates had a T7 promoter encoded within the amplicon, which is required for transcription (see table S4 for primer sequences). PCR products were extracted after gel electrophoresis, and 1 μg of each construct was used in 50 μl of in vitro transcription reaction. Reactions comprised 30 mM Tris (pH 8.0 at 25°C), 10 mM DTT, 0.1% Triton X-100, 0.1% spermidine, 60 mM $MgCl_2$, 0.2 μl SUPERase•In (Thermo Fisher Scientific), 6 mM each NTP and 0.02 mg ml^{-1} of T7 polymerase. Reactions were incubated overnight at 37°C. The next day, pyrophosphate precipitate was removed by centrifugation, and the DNA template was digested by adding 1 μl of TURBO DNase (2 U $μl^{-1}$) (Thermo Fisher Scientific) and incubating for 30 min at 37°C. The resulting RNA was purified using the Monarch RNA Cleanup Kit (NEB) and stored at −80°C.

In vivo splicing assays

In vivo splicing assays were performed in *E. coli* strain BL21 (DE3) transformed with a Mini-IS variant–encoding plasmid or cotransformed with Mini-IS and TnpB expression plasmids. For single-plasmid transformations, single colonies were picked from a plate and used to inoculate overnight cultures in LB with spectinomycin (100 μg ml^{-1}). In the morning, the cultures were reinoculated at 40× dilution in LB supplemented with spectinomycin (100 μg ml^{-1}) and IPTG (0.1 mM) and grown until the OD_{600} reached 0.5 to 0.7. Then, an aliquot equivalent to 250 μl of cell suspension at OD_{600} = 0.5 was taken from each culture and centrifuged at 6000 rpm for 5 min, and the cell pellet was resuspended in 750 μl of Trizol (Thermo Fisher Scientific). After incubating 10 min at room temperature, 150 μl of chloroform was added, and tubes were shaken and centrifuged at 12,000*g* for 15 min at 4°C. The aqueous phase was transferred to a new tube and mixed with an equal volume of absolute ethanol (>96%), followed by RNA purification using the Monarch RNA Cleanup Kit (NEB). Purified RNA was stored at −80°C. For splicing assays with TnpB co-expressed in trans, single colonies were used to inoculate overnight cultures in LB with spectinomycin (100 μg ml^{-1}) and chloramphenicol (25 μg ml^{-1}). In the morning, the cultures were reinoculated at 40× dilution in LB supplemented with spectinomycin (100 μg ml^{-1}), chloramphenicol (25 μg ml^{-1}), and IPTG (0.5 mM) and grown until the OD_{600} reached 0.5 to 0.7. All downstream steps were performed as described above.

RT-PCR and RT-qPCR analyses

Two hundred nanograms of the purified total RNA was used as an input for reverse transcription reactions. First, total RNA was treated with 1 μl of dsDNase (Thermo Fisher Scientific) in 1× dsDNase reaction buffer in the final volume of 10 μl, incubating at 37°C for 20 min. Then 1 μl of 10 mM dNTP, 1 μl of 2 μM oSL12027, and 1 μl of 2 μM oSL13568 were added for gene-specific priming, and samples were heated at 65°C for 5 min. Tubes were then placed directly on ice, followed by the addition of 4 μl of SSIV buffer, 1 μl of 100 mM dithiothreitol (DTT), 1 μl of SUPERase•In (Thermo Fisher Scientific), and 1 μl of SuperScript IV Reverse Transcriptase (200 U $μl^{-1}$, Thermo Fisher Scientific); incubation at 53°C for 10 min; and then incubation at 80°C for 10 min. The resulting cDNA was diluted and used for endpoint or qPCR.

Endpoint PCR was performed in a 20-μl reaction volume containing 1× OneTaq Master Mix (NEB), 0.2 μM of each primer, and 1 μl of 100-fold diluted cDNA. Thermal cycling conditions were as follows: DNA denaturation (94°C for 30 s), 30 cycles of amplification (denaturation: 95°C for 15 s; annealing: 46°C for 15 s; extension: 68°C for 15 s), and a final extension (68°C for 5 min). Products were resolved by 1.5% agarose gel electrophoresis and visualized by staining with SYBR Safe (Thermo Fisher Scientific). We verified that splicing was not a result of residual DNA excision products being formed in cells (e.g., from trans-acting *E. coli* transposases) by carefully verifying the absence of any junction products under $\Delta TnpA_S$ conditions in transposon excision assays (Fig. 2B).

For quantitative measurements, experiments were carried out in biological triplicates, unless indicated otherwise. qPCR was performed in 10-μl reactions containing 5 μl of SsoAdvanced Universal SYBR Green Supermix (BioRad), 1 μl of H_2O, 2 μl of primer pair at 2.5 μM concentration, and 2 μl of 100-fold diluted cDNA (10-fold when intron was expressed from a J23114 promoter). Two primer pairs were used: (i) spliced RNAs were captured using a forward primer annealing to exon 1 (oSL12715) and reverse primer spanning the splice-junction (oSL12719), and (ii) unspliced products were amplified using the same forward primer annealing to exon 1 (oSL12715) and reverse primer annealing to the IStron left end (oSL13922). Reactions were prepared in 384-well clear/white PCR plates (BioRad), and measurements

were performed on a CFX384 RealTime PCR Detection System (BioRad) using the following thermal cycling parameters: polymerase activation and DNA denaturation (98°C for 2.5 min), 40 cycles of amplification (98°C for 10 s, 62°C for 20 s), and terminal melt-curve analysis (decrease from 95°C to 65°C in 0.5°C per 5 s increments). For each sample, the ratio of spliced to unspliced was obtained by calculating spliced/unspliced = $2^{-\Delta Cq}$, where ΔCq = Cq (spliced) – Cq(unspliced). All primer sequences are provided in table S4.

Phenotypic splicing assay

E. coli strain K-12 substrain MG1655 derivative with *lacZ* deletion replaced by a chloramphenicol-resistance cassette (sSL3391) was transformed with plasmids encoding IStron-interrupted *lacZ* gene, intact *lacZ*, or frameshift containing *lacZ* controls. Single colonies were isolated on LB agar plates supplemented with tetracycline (10 μg ml^{-1}). The obtained colonies were streaked on LB agar with tetracycline (10 μg ml^{-1}) and X-gal (200 μg ml^{-1}), and a blue phenotype was expected if a functional *lacZ* transcript was formed as a result of splicing.

REFERENCES AND NOTES

1. R. K. Aziz, M. Breitbart, R. A. Edwards, Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* **38**, 4207–4217 (2010). doi: 10.1093/nar/gkq140; pmid: 20215432
2. A. B. Hickman, F. Dyda, Mechanisms of DNA transposition. *Microbiol. Spectr.* **3**, MDNA3-0034-2014 (2015). doi: 10.1128/microbiolspec.MDNA3-0034-2014; pmid: 26104718
3. P. Siguier, E. Gourbeyre, A. Varani, B. Ton-Hoang, M. Chandler, Everyman's guide to bacterial insertion sequences. *Microbiol. Spectr.* **3**, MDNA3-0030-2014 (2015). doi: 10.1128/microbiolspec.MDNA3-0030-2014; pmid: 26104715
4. N. D. F. Grindley, K. L. Whiteson, P. A. Rice, Mechanisms of site-specific recombination. *Annu. Rev. Biochem.* **75**, 567–605 (2006). doi: 10.1146/annurev.biochem.73.011303.073908; pmid: 16756503
5. V. V. Kapitonov, K. S. Makarova, E. V. Koonin, ISC, a novel group of bacterial and archaeal DNA transposons that encode Cas9 homologs. *J. Bacteriol.* **198**, 797–807 (2015). doi: 10.1128/JB.00783-15; pmid: 26712934
6. E. V. Koonin, K. S. Makarova, Mobile genetic elements and evolution of CRISPR-Cas systems: All the way there and back. *Genome Biol. Evol.* **9**, 2812–2825 (2017). doi: 10.1093/gbe/evx192; pmid: 28985291
7. C. Meers *et al.*, Transposon-encoded nucleases use guide RNAs to promote their selfish spread. *Nature* **622**, 863–871 (2023). doi: 10.1038/s41586-023-06597-1; pmid: 37758954
8. T. Karvelis *et al.*, Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. *Nature* **599**, 692–696 (2021). doi: 10.1038/s41586-021-04058-1; pmid: 34619744
9. H. Altae-Tran *et al.*, The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. *Science* **374**, 57–65 (2021). doi: 10.1126/science.abj6856; pmid: 34591643
10. W. Bao, J. Jurka, Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. *Mob. DNA* **4**, 12 (2013). doi: 10.1186/1759-8753-4-12; pmid: 23548000
11. M. Saito *et al.*, Fanzor is a eukaryotic programmable RNA-guided endonuclease. *Nature* **620**, 660–668 (2023). doi: 10.1038/s41586-023-06356-2; pmid: 37380027
12. P. H. Yoon *et al.*, Eukaryotic RNA-guided endonucleases evolved from a unique clade of bacterial enzymes. *Nucleic Acids Res.* **51**, 12414–12427 (2023). doi: 10.1093/nar/gkad1053; pmid: 37971304
13. K. Jiang *et al.*, Programmable RNA-guided DNA endonucleases are widespread in eukaryotes and their viruses. *Sci. Adv.* **9**, eadk0171 (2023). doi: 10.1126/sciadv.adk0171; pmid: 37756409
14. H. Altae-Tran *et al.*, Diversity, evolution, and classification of the RNA-guided nucleases TnpB and Cas12. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2308224120 (2023). doi: 10.1073/pnas.2308224120; pmid: 37983496
15. D. Kersulyte, A. K. Mukhopadhyay, M. Shirai, T. Nakazawa, D. E. Berg, Functional organization and insertion specificity of IS*607*, a chimeric element of *Helicobacter pylori*. *J. Bacteriol.* **182**, 5300–5308 (2000). doi: 10.1128/JB.182.19.5300-5308.2000; pmid: 10986230
16. J. Filée, P. Siguier, M. Chandler, I am what I eat and I eat what I am: Acquisition of bacterial genes by giant viruses. *Trends Genet.* **23**, 10–15 (2007). doi: 10.1016/j.tig.2006.11.002; pmid: 17109990
17. C. Gilbert, R. Cordaux, Horizontal transfer and evolution of prokaryote transposable elements in eukaryotes. *Genome Biol. Evol.* **5**, 822–832 (2013). doi: 10.1093/gbe/evt057; pmid: 23563966
18. W. Chen *et al.*, Multiple serine transposase dimers assemble the transposon-end synaptic complex during IS*607*-family transposition. *eLife* **7**, e39611 (2018). doi: 10.7554/eLife.39611; pmid: 30289389
19. P. Ghosh, N. R. Pannunzio, G. F. Hatfull, Synapsis in phage Bxb1 integration: Selection mechanism for the correct pair of recombination sites. *J. Mol. Biol.* **349**, 331–348 (2005). doi: 10.1016/j.jmb.2005.03.043; pmid: 15890199
20. P. Ghosh, A. I. Kim, G. F. Hatfull, The orientation of mycobacteriophage Bxb1 integration is solely dependent on the central dinucleotide of attP and attB. *Mol. Cell* **12**, 1101–1111 (2003). doi: 10.1016/S1097-2765(03)00444-1; pmid: 14636570
21. E. Nicolas *et al.*, The Tn3-family of replicative transposons. *Microbiol. Spectr.* **3**, 3.4.14 (2015). doi: 10.1128/microbiolspec.MDNA3-0060-2014; pmid: 26350313
22. O. Barabas *et al.*, Mechanism of IS200/IS605 family DNA transposases: Activation and transposon-directed target site selection. *Cell* **132**, 208–220 (2008). doi: 10.1016/j.cell.2007.12.029; pmid: 18243097
23. C. Pasternak *et al.*, ISDra2 transposition in *Deinococcus radiodurans* is downregulated by TnpB. *Mol. Microbiol.* **88**, 443–455 (2013). doi: 10.1111/mmi.12194; pmid: 23461641
24. M. R. Boocock, P. A. Rice, A proposed mechanism for IS607-family serine transposases. *Mob. DNA* **4**, 24 (2013). doi: 10.1186/1759-8753-4-24; pmid: 24195768
25. S. He *et al.*, The IS200/IS605 family and "peel and paste" single-strand transposition mechanism. *Microbiol. Spectr.* **3**, 3.4.02 (2015). doi: 10.1128/microbiolspec.MDNA3-0039-2014; pmid: 26350330
26. V. Braun *et al.*, A chimeric ribozyme in *Clostridium difficile* combines features of group I introns and insertion elements. *Mol. Microbiol.* **36**, 1447–1459 (2000). doi: 10.1046/j.1365-2958.2000.01965.x; pmid: 10931294
27. O. Hasselmayer, C. Nitsche, V. Braun, C. von Eichel-Streiber, The IStron Cd*ISt1* of *Clostridium difficile*: Molecular symbiosis of a group I intron and an insertion element. *Anaerobe* **10**, 85–92 (2004). doi: 10.1016/j.anaerobe.2003.12.003; pmid: 16701504
28. O. Hasselmayer *et al.*, *Clostridium difficile* IStron Cd*ISt1*: Discovery of a variant encoding two complete transposase-like proteins. *J. Bacteriol.* **186**, 2508–2510 (2004). doi: 10.1128/JB.186.8.2508-2510.2004; pmid: 15060058
29. N. J. Tourasse, E. Helgason, O. A. Økstad, I. K. Hegna, A.-B. Kolstø, The *Bacillus cereus* group: Novel aspects of population structure and genome dynamics. *J. Appl. Microbiol.* **101**, 579–593 (2006). doi: 10.1111/j.1365-2672.2006.03087.x; pmid: 16907808
30. N. J. Tourasse, F. B. Stabell, A.-B. Kolstø, Survey of chimeric IStron elements in bacterial genomes: Multiple molecular symbioses between group I intron ribozymes and DNA transposons. *Nucleic Acids Res.* **42**, 12333–12351 (2014). doi: 10.1093/nar/gku939; pmid: 25324310
31. T. R. Cech, Self-splicing of group I introns. *Annu. Rev. Biochem.* **59**, 543–568 (1990). doi: 10.1146/annurev.bi.59.070190.002551; pmid: 2197983
32. K. Kruger *et al.*, Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31**, 147–157 (1982). doi: 10.1016/0092-8674(82)90414-7; pmid: 6297745
33. A. M. Pyle, A. M. Lambowitz, 17 group II introns: Ribozymes that splice RNA and invade DNA. *Cold Spring Harbor Monograph Archive* **43**, 469–505 (2006).
34. J. Roman, S. A. Woodson, Reverse splicing of the Tetrahymena IVS: Evidence for multiple reaction sites in the 23S rRNA. *RNA* **1**, 478–490 (1995). pmid: 7489509
35. D. Bhattacharya, V. Reeb, D. M. Simon, F. Lutzoni, Phylogenetic analyses suggest reverse splicing spread of group I introns in fungal ribosomal DNA. *BMC Evol. Biol.* **5**, 68 (2005). doi: 10.1186/1471-2148-5-68; pmid: 16300679
36. P. Haugen, V. Reeb, F. Lutzoni, D. Bhattacharya, The evolution of homing endonuclease genes and group I introns in nuclear rDNA. *Mol. Biol. Evol.* **21**, 129–140 (2004). doi: 10.1093/molbev/msh005; pmid: 14595099
37. M. Belfort, R. J. Roberts, Homing endonucleases: Keeping the house in order. *Nucleic Acids Res.* **25**, 3379–3388 (1997). doi: 10.1093/nar/25.17.3379; pmid: 9254693
38. B. Dujon, Group I introns as mobile genetic elements: Facts and mechanistic speculations—a review. *Gene* **82**, 91–114 (1989). doi: 10.1016/0378-1119(89)90034-6; pmid: 2555264
39. D. M. Gill, Bacterial toxins: A table of lethal amounts. *Microbiol. Rev.* **46**, 86–94 (1982). doi: 10.1128/mr.46.1.86-94.1982; pmid: 6806598
40. Y. Sakaguchi *et al.*, The genome sequence of *Clostridium botulinum* type C neurotoxin-converting phage and the molecular mechanisms of unstable lysogeny. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 17472–17477 (2005). doi: 10.1073/pnas.0505503102; pmid: 16287978
41. H. Skarin, T. Håfström, J. Westerberg, B. Segerman, *Clostridium botulinum* group III: A group with dual identity shaped by plasmids, phages and mobile elements. *BMC Genomics* **12**, 185 (2011). doi: 10.1186/1471-2164-12-185; pmid: 21486474
42. I. Kalvari *et al.*, Rfam 14: Expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, D192–D200 (2021). doi: 10.1093/nar/gkaa1047; pmid: 33211869
43. T. Yamano *et al.*, Crystal structure of Cpf1 in complex with guide RNA and target DNA. *Cell* **165**, 949–962 (2016). doi: 10.1016/j.cell.2016.04.003; pmid: 27114038
44. S. N. Takeda *et al.*, Structure of the miniature type V-F CRISPR-Cas effector enzyme. *Mol. Cell* **81**, 558–570.e3 (2021). doi: 10.1016/j.molcel.2020.11.035; pmid: 33333018
45. W. Y. Wu *et al.*, The miniature CRISPR-Cas12m effector binds DNA to block transcription. *Mol. Cell* **82**, 4487–4502.e7 (2022). doi: 10.1016/j.molcel.2022.11.003; pmid: 36427491
46. R. Nakagawa *et al.*, Cryo-EM structure of the transposon-associated TnpB enzyme. *Nature* **616**, 390–397 (2023). doi: 10.1038/s41586-023-05933-9; pmid: 37020030
47. G. Sasnauskas *et al.*, TnpB structure reveals minimal functional core of Cas12 nuclease family. *Nature* **616**, 384–389 (2023). doi: 10.1038/s41586-023-05826-x; pmid: 37020015
48. M. C. A. Smith, R. Till, M. C. M. Smith, Switching the polarity of a bacteriophage integration system. *Mol. Microbiol.* **51**, 1719–1728 (2004). doi: 10.1111/j.1365-2958.2003.03942.x; pmid: 15009897
49. P. Ghosh, L. A. Bibb, G. F. Hatfull, Two-step site selection for serine-integrase-mediated excision: DNA-directed integrase conformation and central dinucleotide proofreading. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 3238–3243 (2008). doi: 10.1073/pnas.0711649105; pmid: 18299577
50. S. P. Nety *et al.*, The transposon-encoded protein TnpB processes its own mRNA into ωRNA for guided nuclease activity. *CRISPR J.* **6**, 232–242 (2023). doi: 10.1089/crispr.2023.0015; pmid: 37272862
51. Y. Zhou *et al.*, GISSD: Group I Intron Sequence and Structure Database. *Nucleic Acids Res.* **36**, D31–D37 (2008). doi: 10.1093/nar/gkm766; pmid: 17942415
52. G. Hausner, M. Hafez, D. R. Edgell, Bacterial group I introns: Mobile RNA catalysts. *Mob. DNA* **5**, 8 (2014). doi: 10.1186/1759-8753-5-8; pmid: 24612670
53. E. R. Lee, J. L. Baker, Z. Weinberg, N. Sudarsan, R. R. Breaker, An allosteric self-splicing ribozyme triggered by a bacterial second messenger. *Science* **329**, 845–848 (2010). doi: 10.1126/science.1190713; pmid: 20705859
54. J. T. Huff, D. Zilberman, S. W. Roy, Mechanism for DNA transposons to generate introns on genomic scales. *Nature* **538**, 533–536 (2016). doi: 10.1038/nature20110; pmid: 27760113
55. L. Gozashti *et al.*, Transposable elements drive intron gain in diverse eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2209766119 (2022). doi: 10.1073/pnas.2209766119; pmid: 36417430
56. T. Rolland, C. Neuvéglise, C. Sacerdot, B. Dujon, Insertion of horizontally transferred genes within conserved syntenic

regions of yeast genomes. *PLOS ONE* **4**, e6515 (2009). doi: 10.1371/journal.pone.0006515; pmid: 19654869
57. B. Zetsche *et al*., Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759–771 (2015). doi: 10.1016/j.cell.2015.09.038; pmid: 26422227
58. W. X. Yan *et al*., Functionally diverse type V CRISPR-Cas systems. *Science* **363**, 88–91 (2019). doi: 10.1126/science.aav7271; pmid: 30523077
59. J. Strecker *et al*., Engineering of CRISPR-Cas12b for human genome editing. *Nat. Commun.* **10**, 212 (2019). doi: 10.1038/s41467-018-08224-4; pmid: 30670702
60. J. Strecker *et al*., RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48–53 (2019). doi: 10.1126/science.aax9181; pmid: 31171706
61. T. Karvelis *et al*., PAM recognition by miniature CRISPR-Cas12f nucleases triggers programmable double-stranded DNA target cleavage. *Nucleic Acids Res.* **48**, 5016–5023 (2020). doi: 10.1093/nar/gkaa208; pmid: 32246713
62. E. P. Nawrocki, T. A. Jones, S. R. Eddy, Group I introns are widespread in archaea. *Nucleic Acids Res.* **46**, 7970–7976 (2018). doi: 10.1093/nar/gky414; pmid: 29788499
63. W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006). doi: 10.1093/bioinformatics/btl158; pmid: 16731699
64. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). doi: 10.1093/molbev/mst010; pmid: 23329690
65. S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, R. Backofen, LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. *RNA* **18**, 900–914 (2012). doi: 10.1261/rna.029041.111; pmid: 22450757
66. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004). doi: 10.1093/nar/gkh340; pmid: 15034147
67. Z. Yao, Z. Weinberg, W. L. Ruzzo, CMfinder—A covariance model based RNA motif finding algorithm. *Bioinformatics* **22**, 445–452 (2006). doi: 10.1093/bioinformatics/btk008; pmid: 16357030
68. E. Rivas, RNA structure prediction using positive and negative evolutionary information. *PLOS Comput. Biol.* **16**, e1008387 (2020). doi: 10.1371/journal.pcbi.1008387; pmid: 33125376
69. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009). doi: 10.1093/bioinformatics/btp348; pmid: 19505945
70. B. Q. Minh *et al*., IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020). doi: 10.1093/molbev/msaa015; pmid: 32011700
71. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017); doi: 10.1038/nmeth.4285; pmid: 28481363
72. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018). doi: 10.1093/molbev/msx281; pmid: 29077904
73. P. Z. Johnson, A. E. Simon, RNAcanvas: Interactive drawing and exploration of nucleic acid structures. *Nucleic Acids Res.* **51**, W501–W508 (2023). doi: 10.1093/nar/gkad302; pmid: 37094080
74. B. W. Ji *et al*., Quantifying spatiotemporal variability and noise in absolute microbiota abundances using replicate sampling. *Nat. Methods* **16**, 731–736 (2019). doi: 10.1038/s41592-019-0467-y; pmid: 31308552
75. N. Liscovitch-Brauer *et al*., Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens. *Nat. Biotechnol.* **39**, 1270–1277 (2021). doi: 10.1038/s41587-021-00902-x; pmid: 33927415
76. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011). doi: 10.14806/ej.17.1.200
77. A. Dobin *et al*., STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013). doi: 10.1093/bioinformatics/bts635; pmid: 23104886
78. H. Li *et al*., The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi: 10.1093/bioinformatics/btp352; pmid: 19505943
79. D. Garrido-Martín, E. Palumbo, R. Guigó, A. Breschi, ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. *PLOS Comput. Biol.* **14**, e1006360 (2018). doi: 10.1371/journal.pcbi.1006360; pmid: 30118475
80. F. Ramírez *et al*., deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016). doi: 10.1093/nar/gkw257; pmid: 27079975
81. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). doi: 10.1093/bioinformatics/btq033; pmid: 20110278
82. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). doi: 10.1093/bioinformatics/btp324; pmid: 19451168
83. Y. Liao, G. K. Smyth, W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014). doi: 10.1093/bioinformatics/btt656; pmid: 24227677
84. M. Vasimuddin, S. Misra, H. Li, S. Aluru, "Efficient architecture-aware acceleration of BWA-MEM for multicore systems" in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (IEEE, 2019), pp. 314–324.
85. G. E. Crooks, G. Hon, J.-M. Chandonia, S. E. Brenner, WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004). doi: 10.1101/gr.849004; pmid: 15173120
86. B. D. Ondov, N. H. Bergman, A. M. Phillippy, Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**, 385 (2011). doi: 10.1186/1471-2105-12-385; pmid: 21961884
87. SourceForge, BBMap (2023); https://sourceforge.net/projects/bbmap/.
88. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). doi: 10.1038/nmeth.1923; pmid: 22388286

ACKNOWLEDGMENTS

We thank Z. Akhtar for laboratory support, N. Sanjana for help with TnY-based tagmentation assays, A.-L. Steckelberg for help with in vitro transcription reactions, A. Bernheim and C. Le Maréchal for helpful discussions about *C. botulinum*, H. H. Wang for anaerobic chamber access for *C. senegalense* culturing, L. F. Landweber for qPCR instrument access, and the JP Sulzberger Columbia Genome Center for NGS support. **Funding:** C.M. was supported by an NIH F32 Postdoctoral Fellowship (GM143924). S.T. was supported by an NIH Medical Scientist Training Program grant (5T32GM145440-02). D.R.G. is supported by the Burroughs Welcome Fund Postdoctoral Diversity Enrichment Program. This research was supported by an NSF Faculty Early Career Development Program (CAREER) Award (2239685), a Pew Biomedical Scholarship, an Irma T. Hirschl Career Scientist Award, and a generous start-up package from the Columbia University Irving Medical Center Dean's Office and the Vagelos Precision Medicine Fund (to S.H.S.). **Author contributions:** R.Ž., C.M., and S.H.S. conceived of and designed the project. R.Ž. developed excision assays and performed all TnpB and splicing assays. H.C.L. performed most bioinformatics analyses, with support from C.M. and S.T. S.T. performed and analyzed RIP-seq and RNA-seq experiments. C.M. selected experimental IStron systems and performed transposon retention and recombination assays. G.D.L. performed and analyzed tagmentation, TAM library, and dinucleotide swapping experiments. E.E.M. performed and analyzed transposon excision assays and group I intron secondary structures. S.R.P. developed and performed initial splicing assays. D.R.G. cultured *C. senegalense* and prepared RNA samples. T.W. assisted with bioinformatics and figure generation. R.Ž., C.M., and S.H.S. discussed the data and wrote the manuscript, with input from all authors. **Competing interests:** Columbia University has filed a patent application related to this work. S.H.S. is a co-founder and scientific advisor to Dahlia Biosciences, a scientific advisor to CrisprBits and Prime Medicine, and an equity holder in Dahlia Biosciences and CrisprBits. **Data and materials availability:** NGS datasets generated and analyzed in this study are available in the NCBI Gene Expression Omnibus (GSE261344, BioProject PRJNA1086522) and Sequence Read Archive (PRJNA1107998).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adm8189
Supplementary Text
Figs. S1 to S11
Tables S1 to S4
References (*89*–*99*)
MDAR Reproducibility Checklist

Submitted 7 November 2023; accepted 8 May 2024
10.1126/science.adm8189