**The Thesis Committee for Sean Yoon-Seo Kang**
**Certifies that this is the approved version of the following thesis:**

# Mechanism of DNA Target Site Recognition by Group II Introns TeI3c and GsI-IIC and Splicing Activity of GsI-IIC Reverse Transcriptase

**APPROVED BY**

**SUPERVISING COMMITTEE:**

**Supervisor:**

______________________________
Alan Lambowitz

______________________________
Claus Wilke

# Mechanism of DNA Target Site Recognition by Group II Introns TeI3c and GsI-IIC and Splicing Activity of GsI-IIC Reverse Transcriptase

**by**

**Sean Yoon-Seo Kang, BS**

**Thesis**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Master of Arts**

**The University of Texas at Austin**

**December 2016**

## Abstract

### Mechanism of DNA Target Site Recognition by Group II Introns TeI3c and GsI-IIC and Splicing Activity of GsI-IIC Reverse Transcriptase

Sean Yoon-Seo Kang, MA
The University of Texas at Austin, 2016

Supervisor: Alan M. Lambowitz

Mobile group II introns are self-catalytic ribozymes found in bacteria and eukaryotic organelles. They can mobilize within the genomes by retrohoming, which involves RNA-catalyzed splicing followed by the excised intron reverse splicing into a target site. Both RNA splicing and retrohoming are facilitated by an intron-encoded reverse transcriptase (RT). Mobile group II introns are of interest as evolutionary ancestors of spliceosomal introns in higher organisms, for their use as bacterial gene targeting vectors known as targetrons, and as a source of thermostable group II intron reverse transcriptases (TGIRTs) for RNA-seq. The focus of this master's thesis is on two thermophilic group II introns found in bacterial thermophiles: the subgroup IIB intron TeI3c and the subgroup IIC intron GsI-IIC. The TeI3c intron is known to rely on base pairing interaction between exon-binding site sequences 1/2 (EBS1/2), within the intron RNA, and intron-binding site sequences 1/2 (IBS1/2) in the 5' exon of its target DNA, but it is not clear what targeting rules dictate one target sequence to be better or worse than others. I studied the targeting rules of TeI3c during retrohoming by using randomized libraries and next-generation sequencing followed by computational analysis of the sequence data. Understanding the targeting rules of TeI3c can be the important step in the development of thermostable targetron, which can be useful for metabolic engineering in the biofuel industry. Unlike

TeI3c, which relies primarily on base pairing for DNA target recognition, the GsI-IIC intron recognizes a 5'-exon hairpin secondary structure of the target DNA. However, the secondary structure requirements of good targets have not been studied. I studied the secondary structure requirements during GsI-IIC retrohoming by using doped target libraries and next-generation sequencing to find conserved positions within a hairpin target site followed by mobility assays on different target sites with mutated conserved positions. Finally, I studied the forward splicing of GsI-IIC intron by comparing different hairpin target sites including the same mutated target sites tested for their mobility efficiency. These experiments address whether the 5'-exon hairpin structure is recognized similarly for RNA splicing and intron mobility.

## Table of Contents

# List of Tables

## List of Figures

# Chapter 1: Introduction

## MOBILE GROUP II INTRONS

Mobile group II introns are widespread self-catalytic ribozymes found in bacteria, archaea, and the mitochondrial and chloroplast DNAs of some eukaryotes[1]. They are thought to be evolutionary ancestors of spliceosomal introns, non-LTR-retrotransposons, and telomerase, which constitute more than half of the human genome[2]. The group II intron RNA splicing and mobility mechanisms have provided insight into the evolution of introns and splicing mechanisms in eukaryotes. Additionally, mobile group II introns have biotechnological applications as bacterial gene targeting vectors known as targetrons and as source of thermostable group II intron reverse transcriptases (TGIRTs), which are used for RNA-seq[1,3].

Group II introns have a conserved three-dimensional structure with six distinct helical domains[1]. Domains I and V are minimally required for the catalytic activity, while domain VI contains the branch point nucleotide necessary for intron splicing. Mobile group II introns encode an intron-encoded protein (IEP) with both reverse transcriptase (RT) activity, important for intron mobility, and maturase activity which promotes intron splicing by stabilizing the catalytically active ribozyme structure[1]. The open reading frame (ORF) encoding for the IEP is found in domain IV[4,5]. The catalytically active group II intron forms a ribonucleoprotein (RNP) complex with the IEP, whose cooperative action promotes intron mobilization within genomes. The structural features and mobility and splicing mechanisms of mobile group II introns strongly suggest that they are evolutionary ancestors of spliceosomal introns and retrotransposons in eukaryotes[1,2].

## GROUP II INTRON SPLICING MECHANISM

As shown in Figure 1, group II intron splicing begins when the 2' OH of a bulged adenosine within the intron attacks the 5' splice site, creating an intermediate lariat still attached to the 3' exon. The 3' OH of the newly cleaved upstream exon is now exposed and attacks the 3' splice site, ligating the exons and excising the lariat intron[6]. The two sequential transesterification reactions catalyzed by group II intron RNAs are identical to those of spliceosomal introns, which also result in spliced exons and an excised intron lariat RNA[7-10].

To catalyze splicing, the group II intron RNA must fold into a conserved three-dimensional structure consisting of six secondary structure domains (DI-DVI), which form an RNA active site with catalytic $Mg^{2+}$ ions[11-14]. The formation of active ribozyme structure is facilitated by the group II RT[15-18]. The group II intron domains forming the active site are structurally and functionally analogous to U2, U5, and U6 snRNAs that form the active site of the spliceosome. Moreover, the group II intron RT is homologous to the spliceosomal protein Prp8, which is crucial for the assembly of the spliceosomal catalytic core[19]. Recent cryo-EM structures of the spliceosome and a group II intron RNA show virtually identical RNA catalytic cores comprised of snRNAs or group II intron domains bound similarly by Prp8 or the group II intron RT, respectively, providing conclusive evidence for an evolutionary relationship[19,20].

## MECHANISM OF MOBILE GROUP II INTRON MOBILITY: RETROHOMING

Group II introns proliferate within genomes via a retrotransposition mechanism called retrohoming (Figure 2)[1]. During retrohoming the catalytically active group II intron forms an RNP complex with an IEP, whose RT and maturase activities are important for retrohoming. After promoting RNA splicing, the IEP remains tightly bound to the excised lariat intron RNA and aids in DNA target site recognition. Once the target is recognized, the excised intron RNA is then reverse spliced in the sense strand of the target DNA. If the IEP belongs to a class that has DNA endonuclease activity, it subsequently cleaves the antisense strand downstream of the insertion site. This exposes the 3'-end of the antisense strand, which is used as a primer for reverse transcription carried out by the IEP. The resulting intron cDNA is integrated into the genome after several additional steps including RNA degradation, second strand DNA synthesis, and repair (Figure 2)[21,22]. This RT-mediated mobility of group II introns is the basis of proliferation of group II introns in the genomes[21].

## DIFFERENT CLASSES OF MOBILE GROUP II INTRONS

Group II introns are classified as three subclasses IIA, IIB, and IIC, which are characterized by subclass-specific structural features[1]. Group IIA and IIB introns are larger than group IIC introns and typically encode RTs with a C-terminal endonuclease (En) domain, whose endonuclease activity cleaves a target DNA to produce the primer to be used for reverse transcription during retrohoming. On the other hand, the smaller group IIC introns encode RTs without an En domain, and they use the nascent strands at DNA replication forks to prime reverse transcription during retrohoming[1]. Although the mobility mechanism of the group IIC introns seems inherently less efficient than that of group IIA and IIB introns, some of them have been able to successfully proliferate within bacterial genomes[1,23].

Surprisingly, the three classes of mobile group II introns are also differentiated from one another by their target DNA recognition mechanisms during retrohoming[1]. Group IIA and IIB intron rely on internal exon binding site (EBS) sequences that base pair with the complementary DNA target sequences known as intron-binding site (IBS) sequences. Such exon-binding sequences are denoted EBS1, EBS2, and δ for group IIA introns and EBS1, EBS2, and EBS3 for group IIB introns. The RT of group IIA and IIB recognizes additional target DNA sequences near the IBS sequences and helps promote local DNA melting, which enables base pairing between EBS and IBS sequences[1]. Although group IIC introns also contain an EBS1 and an EBS3, they lack an EBS2 and utilize a different targeting mechanism. Instead of relying on base pairing at the EBS2 position, group IIC introns recognize a hairpin structure of the target DNA, such as a bacterial transcription terminator that is located at the IBS2 position of the DNA target site[24].

## OVERVIEW OF THESIS RESEARCH

My research examines the DNA target recognition mechanism of two thermophilic group II introns (TeI3c and GsI-IIC) and splicing activity of GsI-IIC reverse transcriptase. TeI3c is a group IIB intron found in the cyanobacterium *Thermosynechococcus elongatu*s, while GsI-IIC is a group IIC intron found in *Geobacillus stearothermophilus*. First, the

TeI3c targeting mechanism was studied by using deep sequencing with libraries having randomized EBS1/IBS1 or EBS2/IBS2 sequences. EBS/IBS combinations enabling intron mobilization were selected by mobility assays to provide sequences of efficient EBS1/IBS1 and EBS2/IBS2 pairs. This provided information about the nucleotide and base-pairing preference of EBS1/IBS1 and EBS2/IBS2, which are critical for TeI3c targeting. Similarly, a library of target site sequences and deep sequencing were used to study the GsI-IIC target recognition mechanism. Since GsI-IIC recognizes the secondary structure of the target site without heavily relying on base-pairing interactions, a doped library was used to avoid totally disrupting the secondary structures of the target sites. The doped library was then selected through mobility assay to obtain efficient targets that enabled successful retrohoming. This analysis elucidated conserved features and nucleotide positions, which allowed me to further study differential mobility efficiencies of mutated target sites. Finally, splicing activity of GsI-IIC reverse transcriptase was examined by using RNA transcripts with different target sites. Three WT target sites (TS7, TS22, and TS34) found in the genome and mutated versions of TS34 were tested for their effects on forward splicing. I found that recognition of a 5'-exon hairpin structure is required for retrohoming but not for RNA splicing.

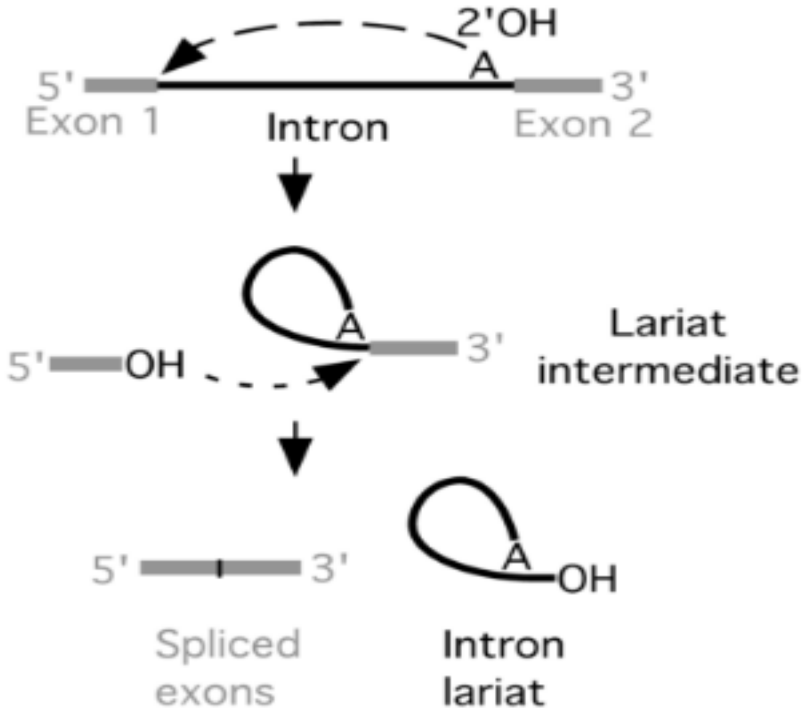


**Figure 1. Group II intron splicing mechanism.** Group II intron splicing begins with 2' OH of a bulged adenosine within the intron attacking the 5' splice site and creating an intermediate lariat still attached to the 3' exon. The 3' OH of the newly cleaved upstream exon is now exposed and attacks the 3' splice site, ligating the exons and excising the lariat intron. Adapted from Lambowitz and Zimmerly 2004[25].

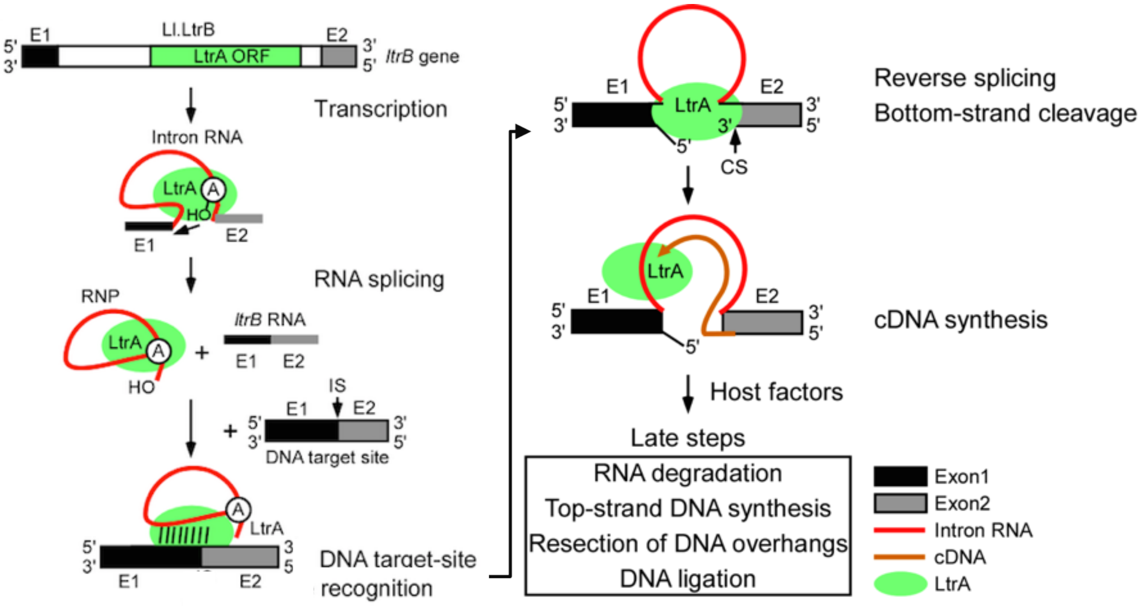


**Figure 2. Group II intron retrohoming mechanism.** During retrohoming the catalytically active group II intron forms an RNP complex with an IEP. The splicing of group II introns involves two sequential transesterification reactions. In the first reaction, the 2'-hydroxyl group of the branch point A nucleotide makes a nucleophilic attack at the 5'-splice site, forming an intron lariat still bound to the 3'-exon. In the second transesterification, the 3'-hydroxyl of the cleaved 5'-exon acts as a nucleophile and attacks the 3'-splice site, resulting in ligated exons and a lariat of intron RNA. The IEP is still tightly bound to the lariat intron RNA and aids in target site recognition. Once the target is recognized, the excised intron RNA is then reverse spliced in the sense strand of the target DNA. The IEP subsequently cleaves the antisense strand downstream of the insertion site. This exposes a 3'-DNA end at the cleavage site of the antisense strand, which is used as a primer for reverse transcription carried out by the IEP. The resulting intron cDNA is integrated into the genome after several additional steps, including RNA degradation, sense-strand DNA synthesis, and DNA repair. This RT-mediated mobility of group II introns is known as retrohoming and is the basis of proliferation of group II introns in the genomes[22].

## Chapter 2: Characterization of the DNA Targeting Site Recognition Rules of the TeI3c intron

TeI3c is a group IIB intron from the thermophilic cyanobacterium, *Thermosynechococcus elongatus*. It has been of interest because of its potential for use as a thermotargetron, which enables gene targeting in thermophiles[3,26]. As discussed previously (see Introduction), group IIA and IIB introns recognize their DNA target sites for retrohoming primarily through base-pairing interactions between exon-binding sequences (EBS1, EBS2 and δ or EBS3) with complementary intron-binding sequences (IBS1 and IBS2 in the 5' exon and δ' or IBS3 in the 3' exon) (Figure 3). The group II intron RT recognizes additional target sequences flanking the IBS sequences and facilitates local melting of DNA, which allows base-pairing interactions between EBS sequences of the intron and IBS sequences of the target DNA. Because the DNA target site is recognized primarily through base pairing of the intron RNA, it is possible to retarget group IIA and IIB introns to insert into different sites by modifying the base-pairing sequences in the intron RNA. This feature has enabled group II introns to be used as bacterial gene targeting vectors known as targetrons[27,28]. Two widely used targetrons are derived from mesophilic group IIA and IIB introns, the *Lactococcus lactis* Ll.LtrB intron and the *E. coli* EcI5 intron, respectively[29,30]. However, these mesophile-derived targetrons do not function in bacterial thermophiles, which include many biologically and industrially important organisms, including those used for bioethanol production[26]. Determining the detailed base-pairing rules of TeI3c is fundamental in the further development and use of the thermophilic targetron derived from the TeI3c group II intron.

## RESULTS AND DISCUSSION

### Experimental strategy

For TeI3c and other group IIB introns, the EBS1/IBS1 and EBS2/IBS2 base-pairing interactions are major determinants of DNA target site recognition during retrohoming. To deduce nucleotide preferences for these interactions that could be used to develop targeting rules for TeI3c, I prepared two sets of randomized libraries: 1) Library 1 includes randomized EBS1 and IBS1 of intron donor plasmid and randomized IBS1 of recipient plasmid; 2) Library 2 includes randomized EBS2 and IBS2 of intron donor plasmid and randomized IBS2 of recipient plasmid. I used the intron mobility assay outlined in Figure 4 to select for successful retrohoming events. In this assay, an intron-donor plasmid expresses a derivative of the intron that carries a T7 promoter sequence near it's 3' end an integrates into a target site cloned in a recipient plasmid upstream of a promoterless tetracycline-resistance (*tet*$^R$) gene, thereby activating that gene. For the library selections, the recipient plasmid of each library was electroporated first then followed by electroporating of the donor into *E. coli* HMS174 (DE3), which expresses an IPTG-inducible T7 RNA polymerase. The transformants carrying both donor and recipient plasmids were then selected overnight and used for intron mobility assays, which selected for successful retrohoming products in Tet$^R$ colonies. The homing products were then deep sequenced and analyzed to elucidate the targeting rules utilized by TeI3c intron.

### The EBS1/IBS1 interaction

For TeI3c, both EBS1 and IBS1 are 6 nt long and their entire length was randomized for the selection. After selection of Tet$^R$ colonies, DNA was isolated and sequenced on Illumina HiSeq 4000 to obtain 2,310,942 paired-end reads of 150 bps. Each sequence contains the intron EBS sequence and the exon IBS sequence targeted during retrohoming, which makes it possible to match EBS sequences with their interacting partner IBS sequences.

The nucleotide preference was calculated at each position using WebLogo (Figure 5A). In the retrohoming products IBS1 is immediately upstream of the inserted intron and IBS1 positions are denoted -1 to -6 from the intron insertion site. EBS1 positions are numbered by the IBS1 position with which they base pair. The EBS1 position -6 shows strong preference for C while its pairing partner strongly prefers G, which suggests this position has a preference for strong base-pairing interaction rather than UA found in the wild-type intron. Interestingly, the -6 position does not equally prefer a reciprocal CG (Figure 5A). The unselected libraries of donor EBSI (11,282,026 sequences) and recipient IBS1 (7,269,323 sequences) were sequenced to make sure the libraries are randomized. Figure 6 shows that the libraries are well randomized since the proportion of each nucleotide is close to 25% per position.

Figure 7 shows that the most preferred EBS1/IBS1 pairs were CG, CG, TA, AT, GC, and CG at -6, -5, -4, -3, -2, and -1 positions, respectively. Watson-Crick (WC) and wobble base pairs are strongly preferred over non-base-pairing nucleotide combinations (Figure 8), which suggests that strong binding between the EBS1 and IBS1 is beneficial for efficient target recognition. No position is particularly tolerable for mismatches and all positions similarly exhibited strong preference towards base pairing.

Figure 9 shows that EBS1/IBS1 interactions with 5 or 6 base pairs are strongly preferred for retrohoming. The 2,310,942 paired EBS1/IBS1 sequences were grouped into two categories: one in which the same sequences appeared once or twice and the other in which the same sequences appeared more than twice. The former group represents the less effective EBS1/IBS1, whereas the latter represents the more effective pairs resulting in successful retrohoming events. The distribution of sequences appearing only once or twice closely matches that of a set of randomized sequences (Figure 10).

Together, my results show that there is strong selection for base pairing throughout the EBS1/IBS1 interaction. EBS1/IBS1 interactions with no more than one mismatch are strongly selected, and there is a strong preference for a CG (nucleotides in the order of EBS1 and IBS1 nucleotides) pair at position -6 and some preference for a CG pair at position -5 and -1. Position -1 also seems to show tolerance for a GT wobble pair.

## The EBS2/IBS2 Interaction

For TeI3c, both EBS2 and IBS2 are 5-nt long and their randomization was extended to neighboring nucleotides as shown in Figure 3. The randomization of EBS2 was confined to the non-stem region so that it does not disrupt the secondary structure, which may be important for intron mobilization. The randomization of IBS2 was also extended to the two adjacent nucleotides between IBS1 and IBS2 (Figure 3).

After selection of $Tet^R$ colonies, DNA was isolated and sequenced on Illumina HiSeq 4000 to obtain 822,341 paired-end (2 x 150 nt) sequences. The nucleotide frequency at each position within EBS2 and IBS2 was plotted using WebLogo (Figure 11A). The randomization of EBS2 was extended to cover two more 3' and three more 5' nucleotide positions (Figure 11B). Unlike EBS1/IBS1 whose mobility efficiency was abolished by the initially attempted extension of the randomized region, EBS2/IBS2 with extended randomization still produced successful homing products. This suggests that the neighboring sequences of EBS2/IBS2 are more tolerant to randomization than the sequences flanking EBS1/IBS1. The most preferred nucleotides of IBS2 are C, A, T, C, and T from -13 to -6 positions, whereas the most preferred nucleotides of EBS2 are G, T, A, G, and A, which can Watson-Crick base pair with the IBS2's most preferred nucleotides. The three nucleotides upstream of EBS2 exhibit preference for A, T, and T from 3' to 5', while the two nucleotides downstream of IBS2 prefer T. The two nucleotide positions upstream of EBS2 show selection against T. This may be because their potential base-pairing interactions with the two A residues at positions -14 and -15 upstream of EBS1 can be detrimental to intron mobility (Figure 11A). A previous study showed that the two A residues at position -14 and 015 are recognized by IEP during retrohoming[26]. The unselected libraries of donor EBS2 (822341 sequences) and recipient IBS2 (1022247 Sequences) were sequenced to make sure the libraries are randomized. Figure 12 shows that the proportion of each nucleotide is closed to 25% per position.

Each set of EBS2 and IBS2 was paired to calculate the frequencies of different base pairs at each position in the EBS2/IBS2 interaction (Figure 13). The most preferred

EBS2/IBS2 pairs were GC, TA, AT, GC, and AT at -13, -12, -11, -10, and -9 positions, respectively. The AT pair at position -9 appears to be strongly favored.

As for the EBS2/IBS2 interaction, Watson-Crick and wobble base pairs are strongly preferred over non-base-pairing nucleotide combinations at each position of the EBS2/IBS2 interaction (Figure 14). However, positions -11 and 12 appear to be more tolerant of mispairings than do any of the other positions in either the EBS2/IBS2 or EBS1/IBS1 interactions.

Similar to the analysis of EBS1/IBS1, the 833,431 paired EBS2/IBS2 sequences were grouped into two categories: one in which the same sequences appeared once or twice and the other in which the same sequences appeared more than twice (Figure 15). The first group includes the less effective EBS2/IBS2 interactions and is skewed toward those with a higher number of mismatches. On the other hand, EBS2/IBS2 interactions with five base pairs are strongly preferred over those with smaller number of base pairs (Figure 16).

Together, my results show that as for the EBS1/IBS1 interaction, base pairing throughout the EBS2/IBS2 interaction is beneficial for efficient target recognition. In the case of EBS2/IBS2, there appears to selection for specific base pairs at some positions, including GC at position -13, TA at position -12, GC or CG at position -10, and AT at position -9. Also, strong selection is present for T at -7 and against C at position -8 between IBS1 and IBS2. Within intron, there is selection against Ts at positions -14 and -15 to prevent formation of TA base pairs that would interfere with protein recognition at these positions.

## EXPERIMENTAL DESIGN AND METHODS

### Library Preparation

The insert fragments with randomized regions were cloned into donor (pACD2X) and recipient (pBRR3T2) plasmid backbones, as described in Mohr et al., 2010[31]. The donor insert with the randomized EBS1 and IBS1 was produced by PCR with primers 3c2aEBS1aLib3 and 3cIBSLib3(s). The donor insert with the randomized EBS2 and

IBS2 was produced by PCR stitching of two DNA fragments. The first PCR fragment was produced from the WT with primers 3cIBS12Lib(s) and 3cUniva, and the second PCR fragment was produced from the WT with primers 3c2aEBS2 and 3cBsibot. The two fragments sharing an overlapping region were then fused by the final round of PCR with primers 3cIBS12Lib(s) and 3cBsibot. The donor inserts and backbone (pACD2X) were digested with SpeI and BsiWI and ligated. The recipient inserts were produced by Klenow fill-up reaction. Oligonucleotides containing randomized IBS1 or IBS2 were filled up by 1hr Klenow (New England Biolabs) reaction at 37ºC (3cIBS1LibTop2 contains the randomized IBS1, 3cIBS2LibTop2 contains the randomized IBS2, and 4cLibrev us as a common reverse primer). The filled-in fragments were purified using MiniElute (Qiagen) then digested with EcoRI and PstI. The digested insert was purified again with MiniElute and ligated into a recipient plasmid (pBRR3T2) already digested with EcoRI and PstI.

### Mobility Assay

Library selections were performed using the *E. coli*-based two plasmid intron mobility assay described in Mohr *et al.* (Figure 4)[26,31]. The ampicillin-resistant ($Amp^R$) recipient plasmids were first electroporated into HMS174 (DE3) strain of *E. coli* cells then the transformed cells were treated to be electrocompetent. Chloramphenicol-resistant ($Cap^R$) donor plasmids were then electroporated into the electrocompetent HMS174 (DE3) cells carrying the recipient plasmids. The order of the sequential electroporation seemed important since electroporating the donor before the recipient did not work. Also, electroporating both libraries at the same time did not work. Cells carrying both the donor and recipient plasmids were selected overnight in LB media with ampicillin (100 mg/L) and chloramphenicol (25 mg/L). The donor plasmid uses a $T7_{lac}$ promoter (PT7lac) to express a group II intron RNA flanked by two exons (E1 and E2) and the group II reverse transcriptase (RT) cloned downstream of the E2. Intron mobilization was induced with IPTG (0.5 mM) at 48°C for 1 h. The mobilized intron inserts into a target site (the ligated E1-E2) cloned in the $Amp^R$ recipient plasmids. The target site is followed by promoterless tetracycline-resistance (*tet*$^R$) gene, so that successful integration of the intron carrying a PT7lac promoter activates the *tet*$^R$ gene which confers tetracycline resistance to the homing product plasmid (Figure 4). The homing product was selected overnight in LB media with tetracycline (25 mg/L).

### Next-Generation Sequencing and Computational Analysis

The homing products were extracted from the overnight-selected culture using Qiagen HiSpeed Plasmid Maxi Kit. Illumina adaptors (IlluminaTOP and Illumina BOT) were added to the tetracycline-selected homing product through PCR. The PCR product was then purified with Agencourt AMpure XP and Illumina tails were added to the cleaned sample via PCR (6 cycles). The PCR product was cleaned with Agencourt AMpure XP again submitted for Illumina HiSeq 4000 paired-end (2 x 150 bases) sequencing at Genomic Sequencing and Analysis Facility (GSAF). Galaxy was used to convert the raw NextGen sequencing data to FASTA format and further trimming of the sequences. Python scripts were used to obtain nucleotide and base-pair frequencies of the doped region. The resulting nucleotide and pair frequencies were plotted using RStudio and Excel.

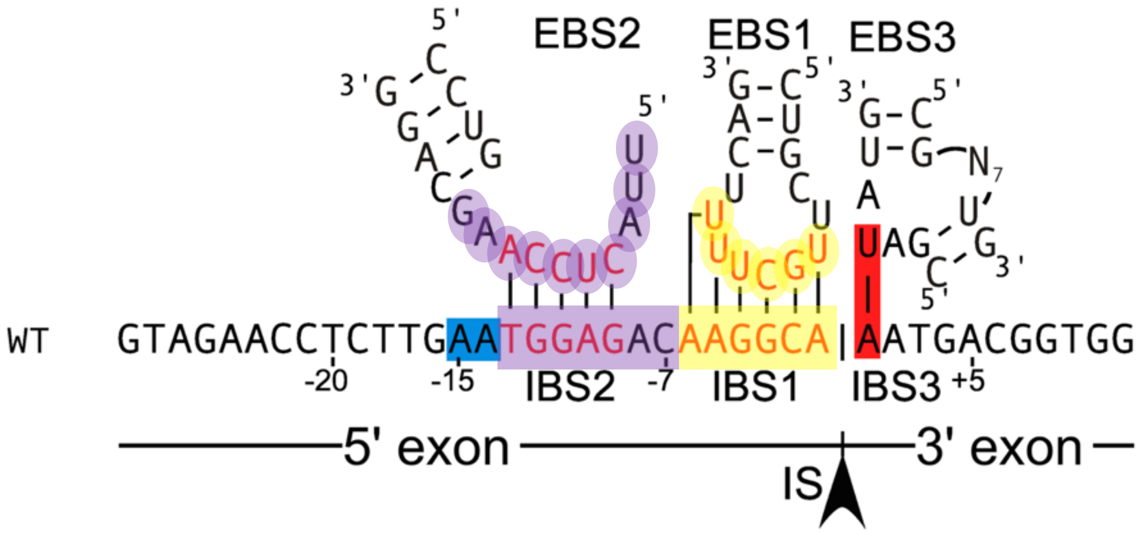


**Figure 3. Schematic diagram of the TeI3c group IIB intron EBS sequences interacting with exon IBS sequences.** DNA target site for group II intron TeI3c showing positions recognized by the IEP (shaded in blue) and intron RNA base pairing (nucleotides shown in red for EBS1/2 and IBS1/2 and shaded in red EBS3 and IBS3)[26]. The regions shaded in yellow were randomized for the library 1 and the ones in purple were randomized for the library 2. Adapted from Mohr *et al*. 2013[26].

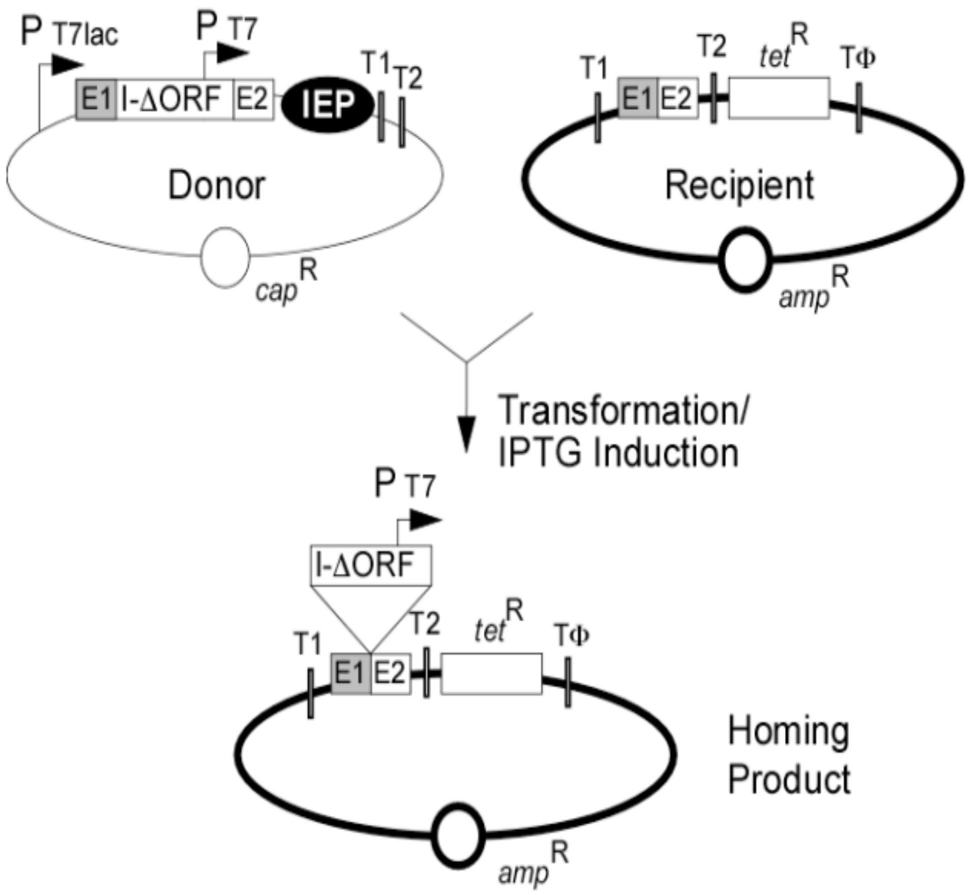


**Figure 4. Schematic diagram of two-plasmid E. coli mobility assay.** Ampicillin-resistant (Amp$^R$) recipient plasmids were first electroporated into HMS174 (DE3) strain of *E. coli* cells and then the transformed cells were treated to be electrocompetent. Chloramphenicol-resistant (Cap$^R$) donor plasmids were then electroporated into the electrocompetent HMS174 (DE3) cells carrying the recipient plasmids. Cells carrying both the donor and recipient plasmids were selected overnight in LB media with ampicillin (100 mg/L) and chloramphenicol (25 mg/L). The donor plasmid uses a T7$_{lac}$ promoter (PT7lac) to express a group II intron RNA flanked between two exons (E1 and E2) and the group II reverse transcriptase (RT) cloned downstream of the E2. Intron mobilization was induced with IPTG (0.5 mM) at 48°C for 1 hour. The mobilized intron inserts into a target site (the ligated E1-E2) cloned in the Amp$^R$ recipient plasmids

upstream of a promoter-less tetracycline-resistance (*tet*$^R$) gene. The successful integration of the intron carrying a PT7lac activates the *tet*$^R$ gene, which confers tetracycline resistance to cells carrying the homing product plasmid. The homing product was selected overnight in LB media with tetracycline (25 mg/L). T1 and T2 are ribosomal RNA transcription terminators that block background transcription of *tet*$^R$ gene by *E. coli* RNA polymerase. Tϕ terminates T7 transcription. Adapted from Mohr *et al*. 2013[26].

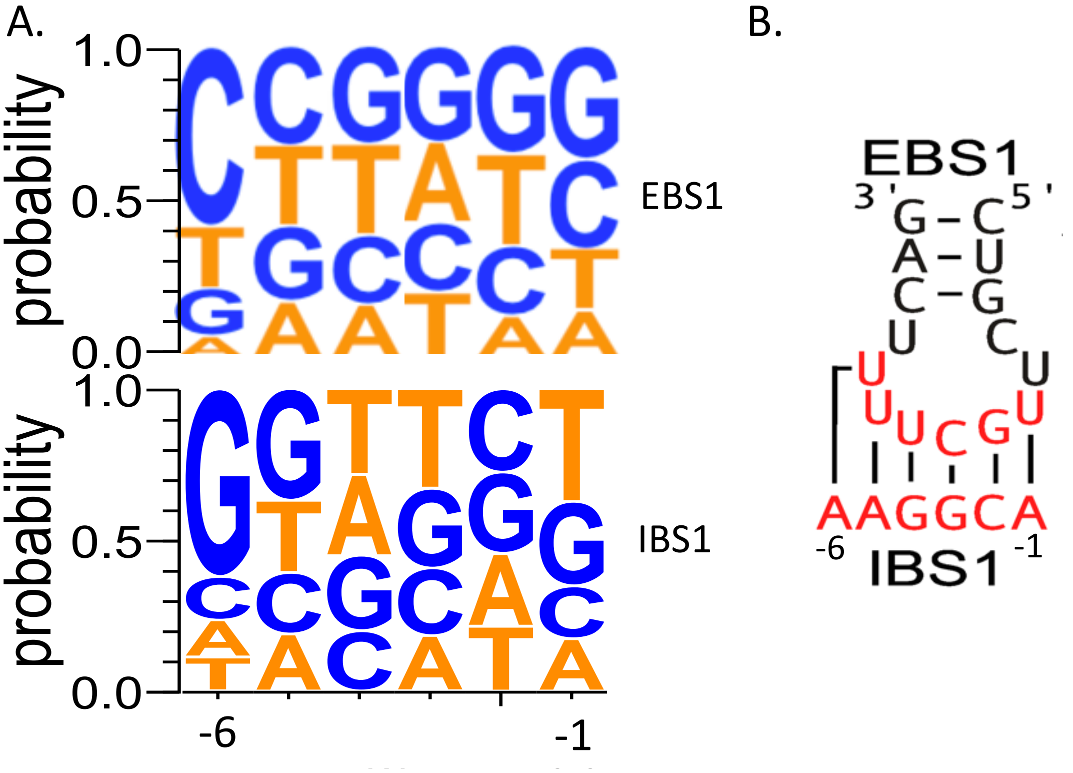


**Figure 5. Nucleotide frequency of each position of EBS1 and IBS1 in the library 1 homing products.** (A) WebLogo of homing product sequences shows different proportion of nucleotides at each position. The size of each letter represents the corresponding nucleotide's proportion. (B) The EBS1 and IBS1 sequences are in red. For library 1, the randomization is confined to the sequences of EBS1 and IBS1.

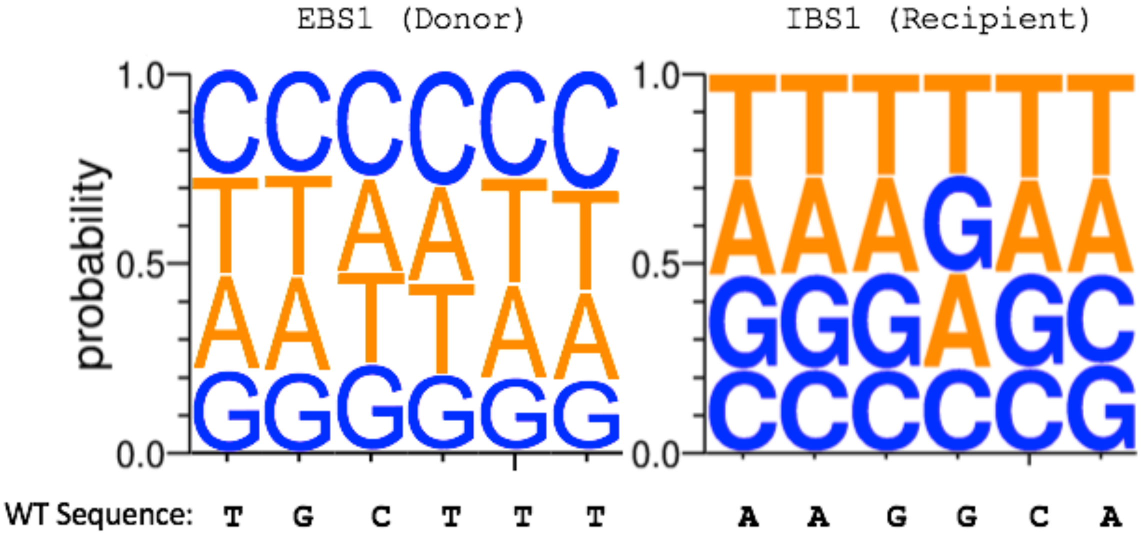


**Figure 6. The unselected libraries of donor EBSI and recipient IBS1.** The unselected libraries of donor EBSI (11,282,026 sequences) and recipient IBS1 (7,269,323 sequences) were sequenced to make sure the libraries are randomized. The size of each letter represents the percentage of the corresponding nucleotide at each position. It is evident that the proportion of each nucleotide is close to 25% per position.

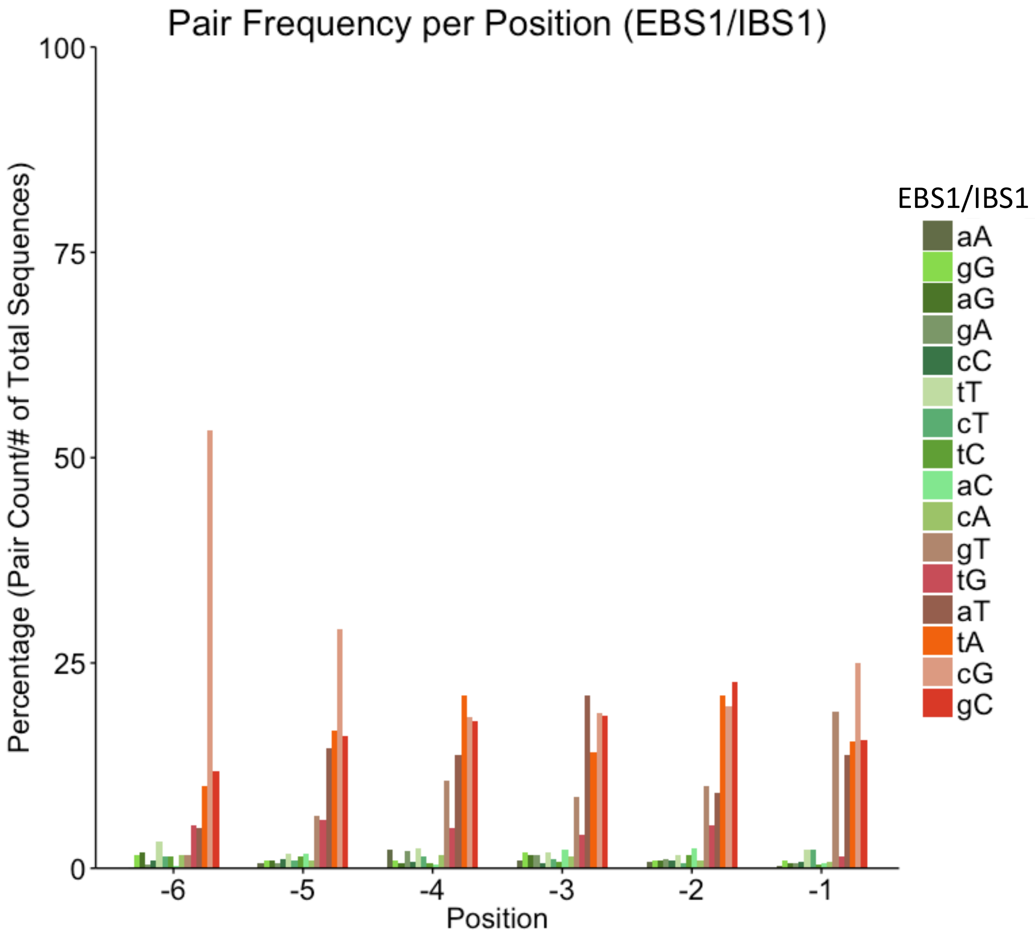


**Figure 7. EBS1/IBS1 pair frequency of each position of the library 1 homing products.** This plot shows preference towards different EBS1/IBS1 pairs per position. The pairs are in the order of non-pairs, wobble pairs, and Watson-Crick (WC) pairs from left to right. The most preferred EBS1/IBS1 pairs were CG, CG, TA, AT, GC, and CG at -6, -5, -4, -3, -2, and -1 positions, respectively. WC base pairs are preferred over non-canonical or wobble base pairs, which suggests that strong binding between the EBS1 and IBS1 is beneficial for efficient target recognition. In the color chart to the right, EBS1 nucleotides are in lower case letters, and IBS1 nucleotides are in uppercase letters.

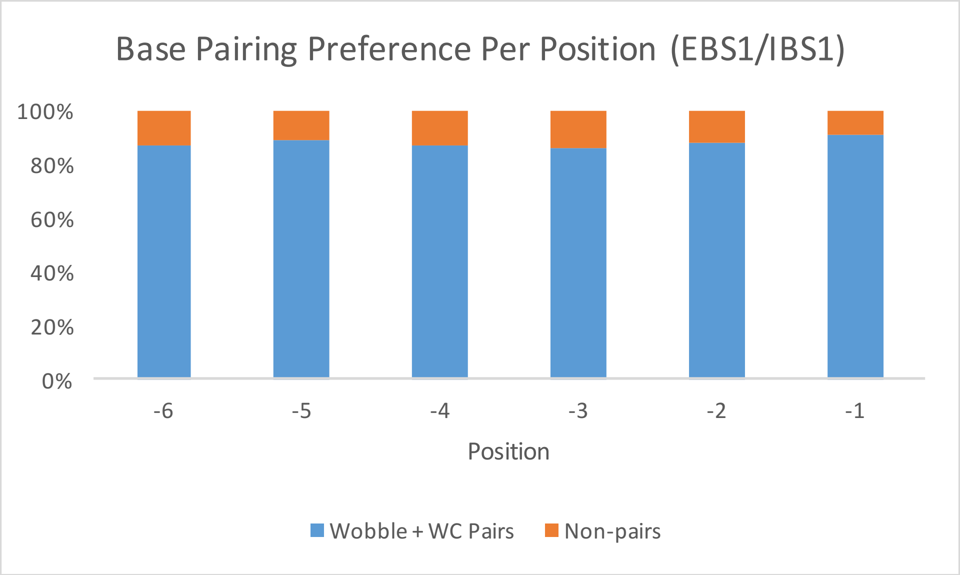


**Figure 8. The percentage of wobble and WC base pairs versus non-pairs at each position in the EBS1/IBS1 interactions in the homing products of the library 1.** The proportion of wobble and WC pairs (blue) is compared to that of non-pairs (orange) for each position of EBS1/IBS1. It is evident that WC and wobble base pairs are strongly preferred (>80%) over non-base-pairing nucleotide pairs.

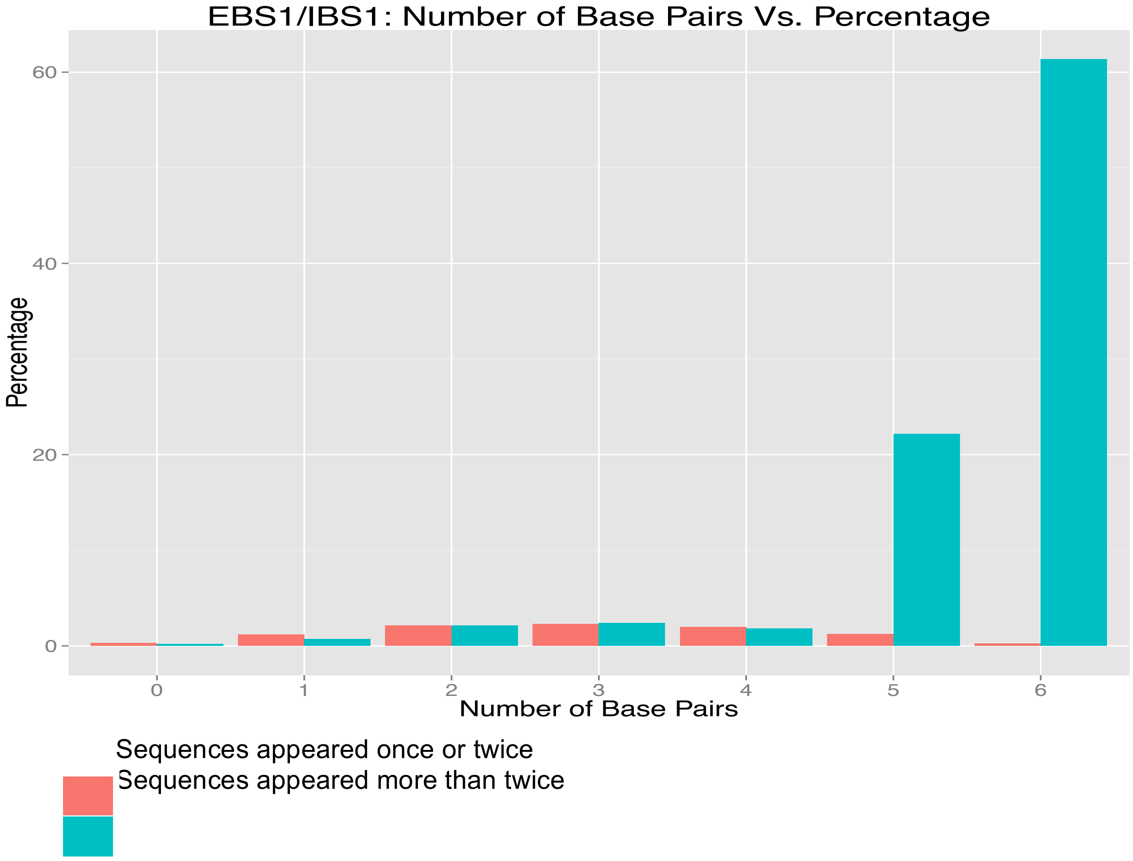


**Figure 9. The total number of wobble and WC base pairs in the EBS1/IBS1 interaction for homing products of the library 1.** EBS1/IBS1 interactions with 5 or 6 base pairs are strongly preferred for retrohoming. Paired EBS1/IBS1 sequences were grouped into two categories: one in which the same sequences appeared once or twice (Red) and the other in which the same sequences appeared more than twice (Green). The former group represents the less effective EBS1/IBS1, whereas the latter represents the more effective pairs resulting in successful retrohoming events.

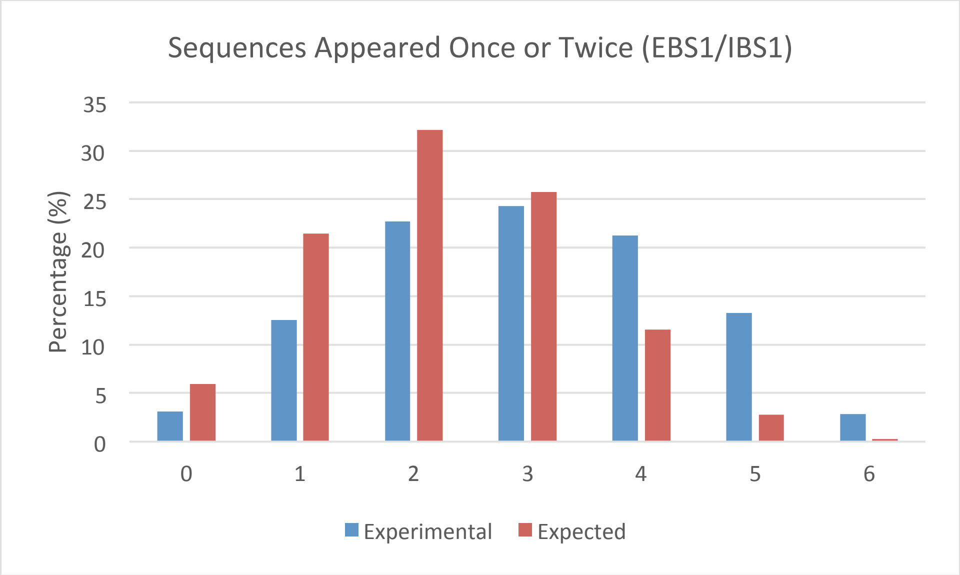


**Figure 10. The total number of wobble and WC base pairs in the EBS1/IBS1 interaction for homing product sequences occurred once or twice.** 214,416 sequences that appeared once or twice were further grouped by the number of WC and wobble pairs. Each blue bar represents the percentage of sequences with the indicated number of base pairs observed experimentally. The red bars represent the expected percentages calculated based on nucleotide frequencies.

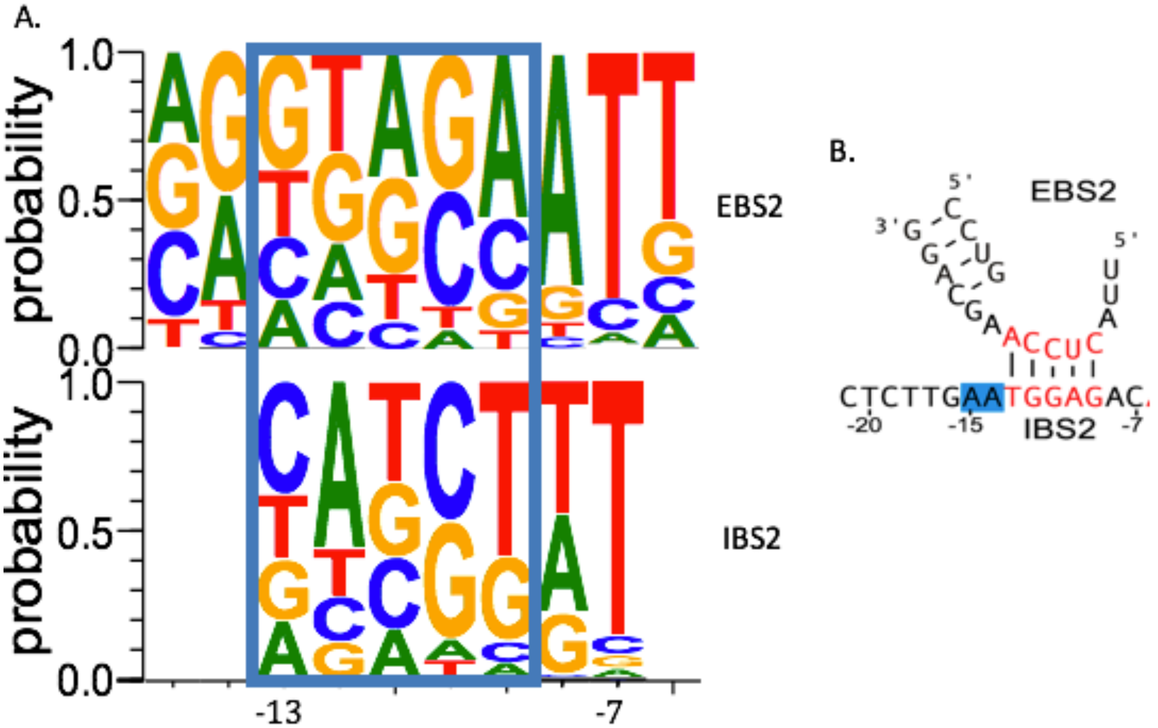


**Figure 11. Nucleotide frequency of each position of EBS2 and IBS2 in the library 2 homing products.** (A) WebLogo of homing product sequences shows different proportion of nucleotides at each position. The size of each letter represents the corresponding nucleotide's proportion. The randomization of EBS2 and IBS2 is extended to neighboring sequences and the blue box indicates the EBS2 and IBS2 regions. (B) The WT EBS1 and IBS1 sequences are in red. For the library 2, the randomization of EBS2 and IBS2 is extended to the adjacent positions. The stem-forming positions of EBS2 are not randomized to avoid disrupting the catalytically important secondary structure of the intron.

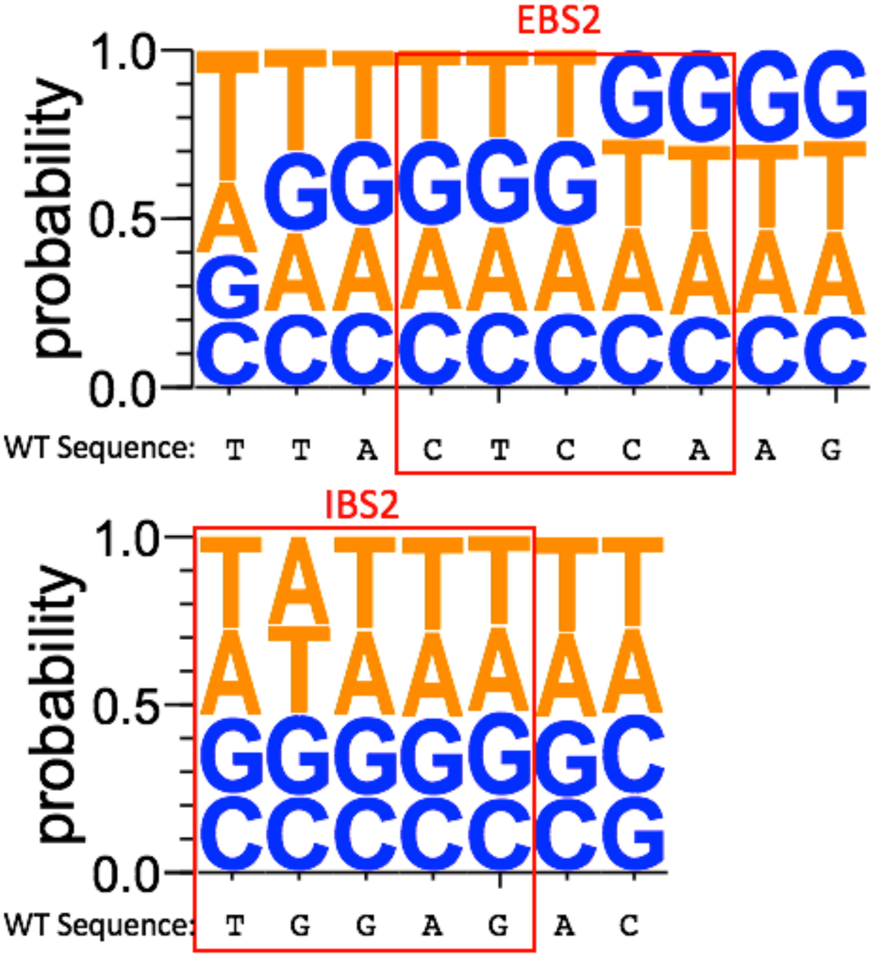


**Figure 12. Unselected libraries of donor EBS2 and recipient IBS2.** Unselected libraries of donor EBS2 (822,341 sequences) and recipient IBS2 (1,022,247 sequences) were sequenced to make sure the libraries are randomized. The size of each letter represents the percentage of the corresponding nucleotide at each position. It is evident that the proportion of each nucleotide is closed to 25% per position in most cases.

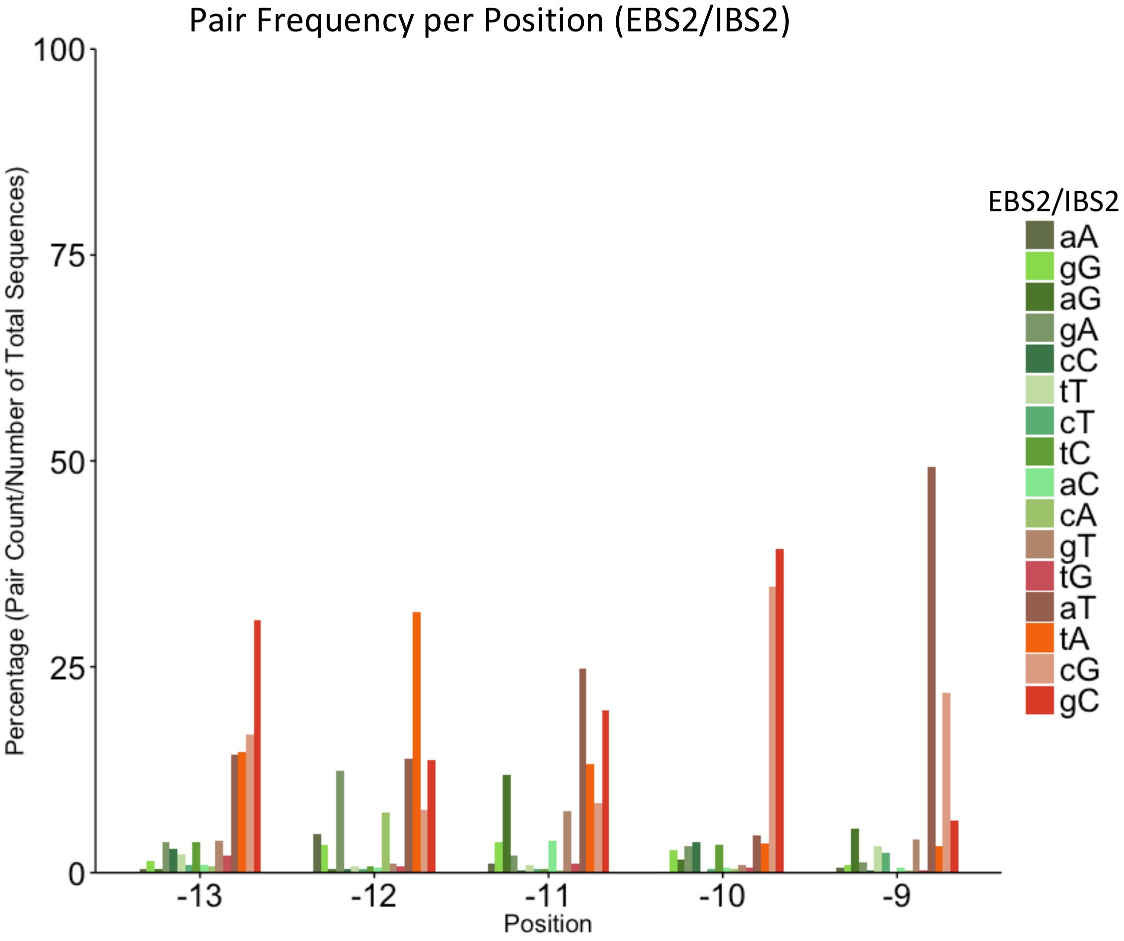


**Figure 13. EBS2/IBS2 pair frequency of each position of the library 2 homing products.** This plot shows preference towards different EBS1/IBS1 pairs per position. The pairs are in the order of non-pairs, wobble pairs, and Watson-Crick (WC) pairs from left to right. The most preferred EBS2/IBS2 pairs were GC, TA, GC, GC, and AT at -13, -12, -11, -10, and -9 positions, respectively. WC base pairs are preferred over non-canonical or wobble base pairs, which suggests that strong binding between the EBS2 and IBS2 is beneficial for efficient target recognition. In the color chart to the right, EBS2 nucleotides are in lower case letters and IBS2 nucleotides are in uppercase letters.

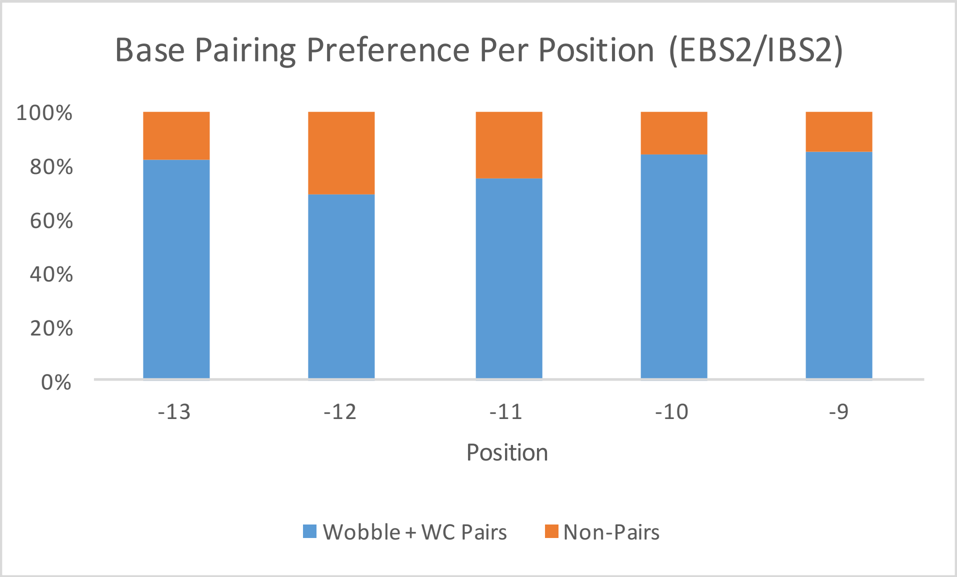


**Figure 14. The percentage of wobble and WC base pairs versus non-pairs at each position in the EBS2/IBS2 interactions in the homing products of the library 2.** The proportion of wobble and WC pairs (blue) is compared to that of non-pairs (orange) for each position of EBS2/IBS2. WC and wobble base pairs are strongly preferred over non-base-pairing nucleotide pairs. Although position -12 exhibits relatively weaker preference towards WC and wobble pairs, the proportion of WC and wobble pairs is still around 70%.

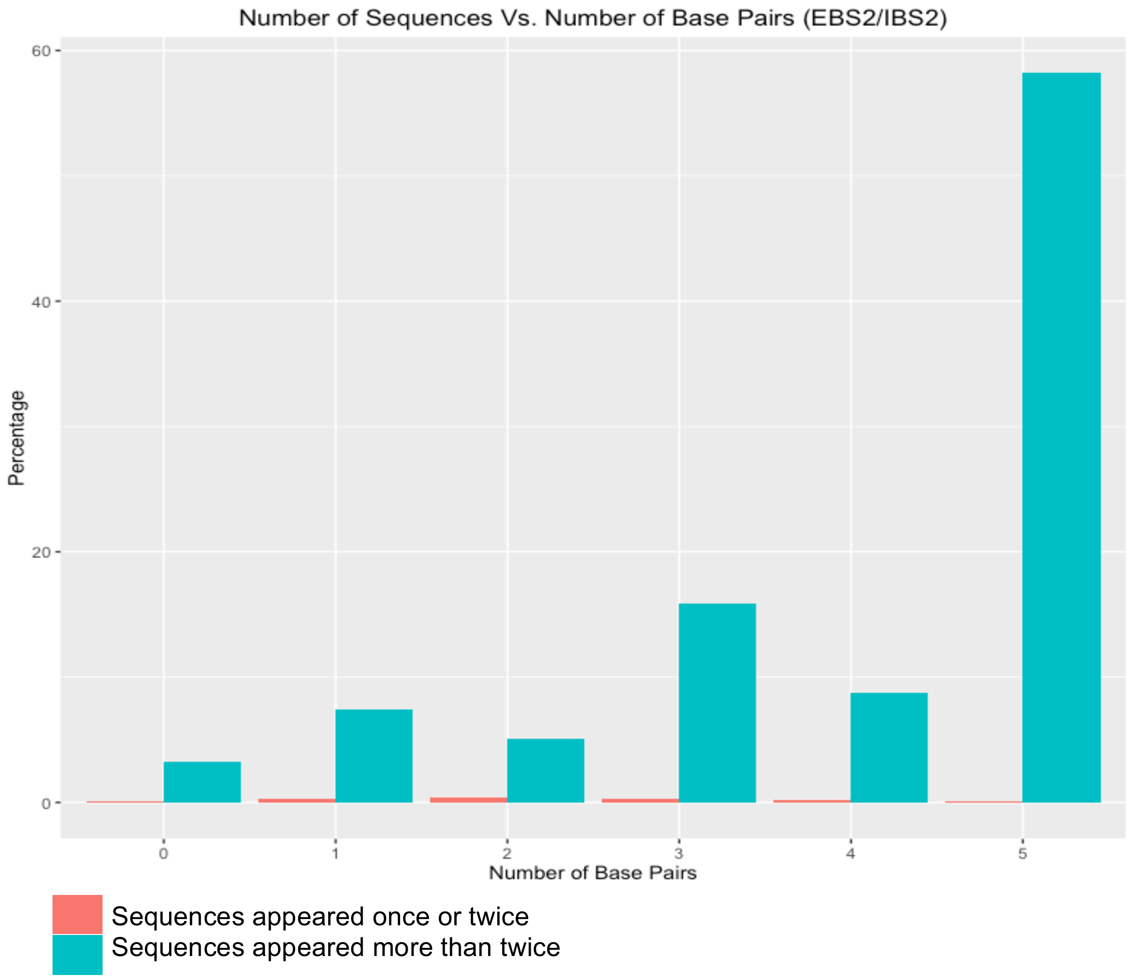


**Figure 15. The total number of wobble and WC base pairs in the EBS2/IBS2 interaction for homing products of the library 2.** EBS2/IBS2 interactions with five base pairs are strongly preferred for retrohoming. Paired EBS2/IBS2 sequences were grouped into two categories: one in which the same sequences appeared once or twice (red) and the other in which the same sequences appeared more than twice (green). The former group represents the less effective EBS2/IBS2, whereas the latter represents the more effective pairs resulting in successful retrohoming events.

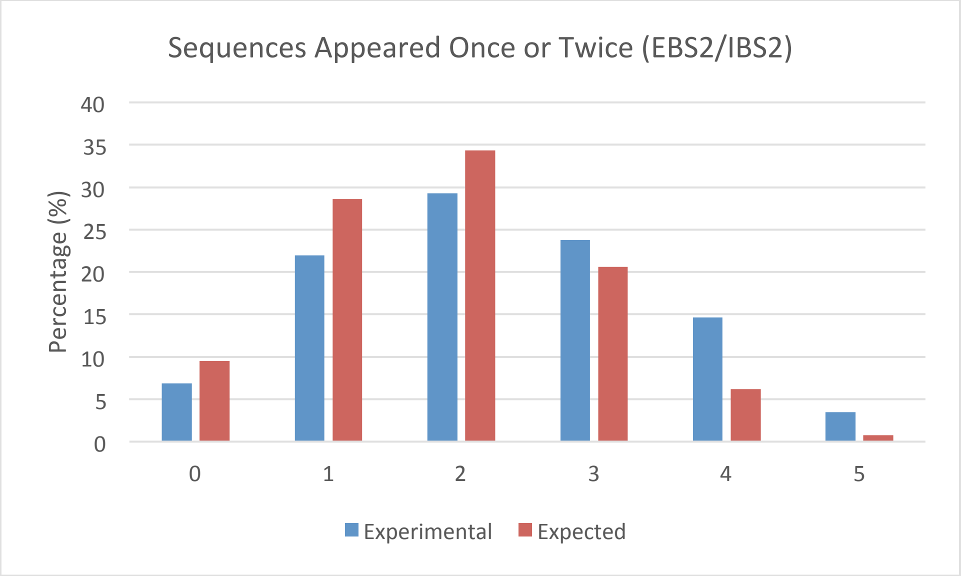


**Figure 16. The total number of wobble and WC base pairs in the EBS1/IBS1 interaction for homing product sequences occurred once or twice.** 11,844 sequences that appeared once or twice were further grouped by the number of WC and wobble pairs. Each blue bar represents the percentage of sequences with the indicated number of base pairs observed experimentally. The orange bars represent the expected percentages calculated based on nucleotide frequency.

## Table 1. Oligonucleotides

| Name | Sequence |
|---|---|
| 3cIBS12Lib(s) | AAAACTAGTAA(N:25252525)(N)(N)(N)(N)(N)(N)AAGGCAGTGCGACGCGAAAGCTAG |
| 3cUniva | TAACGAGGCTTCTAGCG |
| 3c2aEBS2 | CGCTAGAAGCCTCG(N:25252525)(N)(N)(N)(N)(N)(N)(N)(N)(N)CAGGCCAAAGATGCTG |
| 3cBsibot | CCCCGTACGCTGAAAAGCAAGCAGCGTATCCAATCCGCTT |
| 3cIBS1Lib3(s) | AAAACTAGTAATGGAG(N:25252525)(N)(N)(N)(N)(N)(N)(N)GTGCGACGCGAAAGCTAG |
| 3cIBS1LibTop2 | AAACTGCAGCTGTAGAACCTCTTGAATGGAG(N:25252525)(N)(N)(N)(N)(N)(N)(N)AATGACGGTGGACCAGAATTCGACAACCCAACAG |
| 3cIBS2LibTop2 | AAACTGCAGCTGTAGAACCTCTTGAA(N:25252525)(N)(N)(N)(N)(N)(N)AAGGCAAATGACGGTGGACCAGAATTCGACAACCCAACAG |
| 4cLibrev | CTGTTGGGTTGTCGAATT |
| IlluminaTOP | AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGACGATCA |
| Illumina BOT | CAAGCAGAAGACGGCATACGAGATATTCCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| TargetSeq | ATGCGAGAGTAGGGAACTGC |

# Chapter 3: DNA target site recognition and splicing of the group IIC intron *Geobacillus stearothermophilus* GsI-IIC

## INTRODUCTION

Unlike the group IIB intron TeI3c, which utilizes three EBS-IBS base-pairing interactions for DNA target site recognition, group IIC introns, such as GsI-IIC, recognize a 5'-exon hairpin structure in place of the EBS2/IBS2 interaction[1]. However, it remains unclear what structural features of the 5'-exon hairpin region are important for DNA target site recognition by group IIC introns and whether this region is recognized by the intron RNA or the intron-encoded RT. To investigate these issues, I characterized DNA target site recognition and splicing activity of the thermostable GsI-IIC intron, both to identify critical feature of the 5'-exon hairpin region and to determine whether this region is recognized similarly for intron mobility and RNA splicing. My results indicate that the 5'-exon hairpin and at least one nucleotide base upstream of IBS1 are recognized by the intron-encoded RT for DNA integration, but are not required for protein-dependent or self-splicing of the GsI-IIC intron. My results also provide new insights into mechanisms used by GsI-IIC to proliferate to relatively high copy number within its host genome. The work in this chapter was done in collaboration with Dr. Georg Mohr.

## RESULTS

### *G. stearothermophilus* group IIC introns

Figure 17A shows the predicted secondary structure of the GsI-IIC3 intron from *Geobacillus stearothermophilus* strain 10 (GenBank: NZ_CP008934). The structure consists of six secondary structure domains (DI-DVI), which are conserved in all group II introns and potentially interact via tertiary structure contacts (Greek letters; Figure 17). Like other group IIC introns, GsI-IIC contains EBS1 and EBS3 sequences, which can potentially base pair to IBS1 and IBS3 sequences in the 5'- and 3'-exons, respectively (Figure 17).

A previous study identified 17 copies of the GsI-IIC intron in what was then the partial genome sequence of *G. stearothermophilus* strain 10[32]. The completed draft genome sequence revealed 45 copies of the intron, denoted GsI-IIC1-45, which comprise ~2.3% of the genome (Table 2). 44 of these introns are intact and range in size from 1,881 to 1,894 nt. The remaining intron, GsI-IIC41, has a 3,130-nt transposon inserted after amino acid residue 133 of the intron ORF. The different copies of the GsI-IIC intron have >95% sequence identity to each other and are closely related to the *Oceanobacillus ihiensis* group IIC intron, whose X-ray crystal structure has been determined[14], and the *Bacillus halodurans* group IIC intron B.h.I1, whose splicing and mobility mechanisms have been studied previously (~50% and ~60% identity to GsI-IIC over 480-nt of the ribozyme core)[24]. All 45 copies of the GsI-IIC intron are inserted downstream of predicted hairpin structures, and introns inserted in the top and bottom strand are largely segregated on opposite sides of the genome (Table 2 and Figure 18). These features are as expected for group IIC introns, which typically insert preferentially at DNA hairpins in the lagging template-strand and would thus be segregated on opposite sides of a genome undergoing bidirectional replication[1,24]. However, while the requirement for a hairpin upstream of the intron-insertion site appears to be absolute, the genomic distribution (Figure 18) suggests that insertion in the leading template-strand can occur at lower frequency, as borne out by intron mobility assays described below.

The 45 intron copies fall into two secondary structure classes, which differ in the length of DIIb (Figure 17; Table 2). Most of the nucleotide sequence differences between different copies of the GsI-IIC intron correspond to single nucleotide changes or small insertions/deletions in or adjacent to loops or bulges within stems (intron positions 33, 165, 213, and 257 in DI, 288 in DII, 347 in DIII, and 430 in DIV) and are not expected to affect intron function. Exceptions are GsI-IIC2, which has a potentially disruptive single-nucleotide change at the base of DIII, and GsI-IIC21, which has the branch-point A residue in DVI changed to G and a 3-nt deletion in the loop of DIII (Figure 17; Table 2).

## Group II intron-encoded RTs

All 45 copies of the GsI-IIC intron encode an RT of 420 aa with >99% identity to each other. Amino acid substitutions in different copies of the intron are found at only seven positions (Y40D, I41V, R49H, E66G, S105P, M137T, N379K). Except in the intron with a transposon insertion, all the RTs are full-length and have the conserved YADD motif at the RT active site along with other conserved motifs found in active RTs (Figure 19). The purified GsI-IIC34 protein has been shown to have high RT activity[24]. GsI-IIC21, which has a mutation in the bulged A in DVI and deletion in the loop of DIII, also has the most changes in the RT (Y40D, I41V, R49H, S105P, N379K), including three (Y40D, I41V, N379K) that are not found in other copies of the intron (Figure 19). The divergence of the GsI-IIC21 intron may reflect that it was an early insertion that was rendered non-functional by the branch-point mutation and/or that it is inserted at a site that makes it less susceptible to purifying selection than other copies of the intron.

## Genomic insertion sites of the GsI-IIC intron

Previously, Moretz and Lampson[32] described 17 insertion sites of the GsI-IIC intron in *G. stearothermophilus* strain 10 all of which have an upstream hairpin and predicted IBS1 and IBS3 sequences on either side of the intron-insertion site. This holds true for all 45 introns in the completed genome sequence (Figure 20 and Table 2). In 43 cases, the intron is inserted downstream of a gene (22 to 1,625 bp from the stop codon), and the predicted hairpin corresponds to a putative transcription terminator. The two remaining introns (GsI-IIC39 and GsI-IIC24) are inserted within genes. GsI-IIC39 is found within a predicted hairpin in the coding sequence of a two-component sensor histidine kinase (WP_013144786.1), and GsI-IIC24 is inserted within a hairpin in the coding sequence of a short hypothetical protein (WP_053414214.1), but complementary to a putative transcription terminator of the GT50_RS08745 gene on the opposite strand. A web logo shows no strong sequence conservation within the hairpin, but does show strong conservation of the T residue at position -5, which is located in the single-strand region downstream of the hairpin immediately adjacent to the IBS1 sequence (Figure 20).

All copies of the *Geobacillus* GsI-IIC intron in the genome have the same EBS1 (UGGA) and EBS3 (G) sequences, but the IBS1 and IBS3 sequences at the intron-insertion sites differ slightly, with the IBS1 sequences containing as many as two mismatches (including dG:rU and dT:rG pairs) out of the four possible base pairs with EBS1. IBS3 corresponds to a single nucleotide that can base pair with the G residue at EBS (T in 36 introns and C in 8 introns), with one site (GsI-IIC11) having a non-complementary A at the IBS3 position (Figures 17 and 20). The limited EBS1/IBS1 and EBS3/IBS3 interactions suggest that the insertion site specificity of the GsI-IIC intron is dictated primarily by recognition of the hairpin.

Notably, except for introns GsI-IIC41 (transposon insertion) and GsI-IIC21 (branch point A deletion and other mutations), all other copies of GsI-IIC appear to be intact and likely functional, indicating either recent insertion or selection against inactive copies of the intron, despite being inserted downstream of transcription terminators in most cases. BLASTN searches found several hundred related DNA sequences in the Genebank NR database, indicating wide distribution of this intron mainly in *Bacilli* and *Clostridia* and to a lesser degree in *Bacteroidetes* and *Deltaproteobacteria* (G. Mohr, personal communication).

**GsI-IIC mobility assays and effect of deletions in DIV on intron mobility**

To study the GsI-IIC mobility mechanism in detail, we adapted an *E. coli* two-plasmid mobility assay used previously for the *Lactococcus lactis* Ll.LtrB and *Thermosynechococcus elongatus* group II introns (Figure 21A)[24,31]. In this assay, a Cap$^R$ intron-donor plasmid (pADC2X-GsI-IICΔORF+T7) uses a T7$_{lac}$ promoter (PT7lac) to express a GsI-IIC-ΔORF intron carrying a T7 promoter (PT7) with short flanking 5' and 3' exons (E1 and E2, respectively) along with the group II RT, which is cloned downstream of E2. The Amp$^R$-recipient plasmid (pBRR3-GeoTS) contains a DNA target site (the ligated E1–E2 sequence, TS) cloned upstream of a promoterless *tet*$^R$ gene, such that retrohoming of the intron carrying the T7 promoter into the site activates that gene. Two versions of the recipient plasmid were made that differ in the orientation of the target site and Tet$^{R}$ cassette relative to the replication origin to test whether the nascent leading or lagging strand is used preferentially to prime reverse transcription (denoted LEAD and LAG, respectively)[24]. The assays were done in *E. coli* HMS174 (DE3), which contains an IPTG-inducible T7 RNA polymerase, and intron expression was induced with IPTG for 1 h at 48ºC, the highest temperature that could be used without affecting cell viability[31]. Mobility efficiencies were determined in plating assays as the ratio of (Tet$^{R}$+Amp$^{R}$)/Amp$^{R}$ colonies.

### Recognition of the upstream hairpin is required for GsI-IIC mobility

The comparison of the different genomic insertion sites of the 45 copies of the GsI-IIC intron suggested that DNA target specificity is dictated largely by sequence elements in the 5' exon with only the single IBS3 nucleotide in the 3' exon contributing to DNA target site recognition (Figure 20). Supporting this inference, we found no significant difference in mobility efficiency for DNA target sites containing the TS34 5' exon in combination with TS34, TS23, or TS12 3' exons, all of which contain a canonical T residue at the IBS3 position, but have no other common features (Georg Mohr, unpublished data). We therefore focused of recognition elements in the 5' exon, particularly the 5' hairpin structure found upstream of all genomic GsI-IIC insertion sites.

To examine the contribution of the 5'-exon hairpin structure to DNA target site recognition, we selected 5'-exon sequences from three target sites with somewhat different hairpin features (TS7, TS22, and TS34) and tested each in combination with the 3'-exon sequence from a fourth target site (TS12), thereby equalizing any contribution from the 3' exon (Figure 21B). The TS7, TS22, and TS34 sites differ in the length of the hairpin (10 to 14 nt), the presence of unpaired nucleotides in the hairpin, the number of potential base pairs in the IBS1/EBS1 interactions (2 to 4), and the potential to form extra base pairs with additional sequences in the EBS1 loop (nucleotide residues indicated in blue. All three DNA target sites also contain the T-5 residue upstream of IBS1, which was strongly conserved in GsI-IIC genomic insertion sites (Figure 21).

For all three target sites, the insertion frequency into the LAG recipient plasmid is higher (9- to 11-fold) than in the LEAD recipient, confirming preferential use of a lagging strand primer for initiation of reverse transcription as found for other group IIC introns. Comparing the three target sites, mobility efficiencies were highest for TS34 and lowest for TS22, which has the fewest potential base pairs between EBS1 and IBS1, as well as a bulged nucleotide in the middle of the hairpin stem (Figure 21B).

Deleting the hairpin from the TS34 target site, changing one side of the hairpin to match the other side so that it could no longer base pair with the other side, or replacing the hairpin with a sequence of equal length derived from the *tet*$^{R}$ gene decreased the mobility efficiency in the preferred LAG orientation by ~250-fold, with most if not all of the residual mobility found by colony PCR and sequencing to reflect ectopic insertion into a T2 transcription terminator, which is present in the recipient plasmid upstream the *tet*$^{\mathrm{R}}$ marker to prevent read through by *E. coli* RNA polymerase[24]. These findings indicate that the 5'-hairpin structure contributes strongly to DNA target site recognition and identify some features of the hairpin that may be important for optimal recognition.

## Identification of critical features of the 5'-exon hairpin by *in vivo* selection and high-throughput sequencing

To further characterize features of the 5'-exon hairpin region that are important for intron mobility, I carried out an *in vivo* selection experiment using the two-plasmid assay with a recipient plasmid in which the 5'-exon sequence encompassing the TS34 hairpin (positions -5 to -37) was doped at 70% of the wild-type nucleotide residue and 10% of each of the other three nucleotide residues. After induction of donor plasmid expression with IPTG, Tet$^{\mathrm{R}}$ colonies were selected and homing products were amplified by PCR and sequenced on an Illumina HiSeq4000 to obtain ~6 million paired-end (2 X 150) reads.

The data show that the hairpin can be divided into conserved upper and lower stems, which contain strongly selected GC base pairs, separated by a TG elbow at which Watson-Crick base pairing is counterselected (Figure 22A and B). Other strongly conserved nucleotides are two G-residues in the hairpin loop and the T and position -5, which lies in the unpaired region below the hairpin and is also conserved in all 45 GsI-IIC target sites (see Figure 20). Analysis of mutant target sites showed that replacement of the GT elbow and adjacent AT pair with GC base pairs to make a stable continuous helix decreased the mobility efficiency about five-fold. By comparison, mutation of T-5, which lies in the unpaired region below the hairpin just upstream of IBS1, to any other nucleotide residue decreased the mobility efficiency by >250-fold, whereas mutating the adjacent nucleotide G-6 to a T residue had much less effect on intron mobility (decreased four-fold). The very strong effecting of mutating T-5 is surprising and identifies it as a nucleotide residue potentially recognized by the IEP. This would mirror the situation for the thermostable TeI4c group IIB intron, where a single/small number of nucleotide residues immediately upstream of IBS2 appear to be recognized by the IEP and may contribute to local DNA melting[31].

## Protein-dependent and self-splicing of GsI-IIC

To investigate the splicing activity of the GsI-IIC RT, we used an *in vitro* assay in which the purified protein was incubated with a $^{32}$P-labeled precursor RNA containing the 656-nt GsI-IIC-ΔORF intron and short flanking 5’ and 3’ exons. Figure 23A shows that in reaction medium containing 5 mM $Mg^{2+}$ at 50 ºC, the GsI-IIC-ΔORF intron could be spliced efficiently by GsI-IIC RT to produce ligated exons and excised intron lariat RNA. The intron was unable to self-splice under these conditions, but could self-splice hydrolytically to produce linear intron RNA and ligated exons in reaction medium containing 100 mM $Mg^{2+}$ (Figure 23A). Both protein-dependent and self-splicing were more efficient with a construct containing the 5’ exon from the TS34 insertion site than with 5’ exons from the TS7 and TS22 insertion sites, and consequently, the construct with the 5’ exon from the TS34 insertion site was used for all subsequent experiments.

Time-course experiments with 40 nM GsI-IIC precursor RNA and different concentrations of GsI-IIC RT (20 to 200 nM) showed that protein-dependent splicing at 5 mM $Mg^{2+}$ and 50 ºC occurred at a rate of ~6 $min^{-1}$ and was maximally efficient at a molar ratio of 1:2, with ~75% of the precursor RNA spliced under the conditions (Figure 23B). Addition of fresh GsI-IIC protein after 10 min did not result in the splicing of additional precursor RNA, indicating that the unreactive fraction likely reflects precursor RNA molecules that had folded into an inactive conformation. When the amplitude of the reaction was corrected for percentage of active precursor RNA molecules, splicing is seen to have occurred with a stoichiometry of 2:1 protein:RNA throughout the concentration range tested. Together, these findings suggest that two molecules of GsI-IIC, possibly functioning as a dimer, are required to splice one molecule of intron RNA, consistent with previous findings for the Ll.LtrB group IIA intron[18,24].

## Effect of hairpin mutations on RNA splicing

Next, we tested the effect of mutations in the TS34 5' exon on both protein-dependent and self-splicing (Figure 24). The results showed that various mutations including deletion of the hairpin leaving only the IBS1 sequence (Δ34HP), replacing the hairpin with a sequence of equal length derived from the *tet*$^R$ gene, changing T-5 to any other nucleotide, or G-6 to T did not strongly inhibit either protein-dependent or self-splicing. In fact, some of the mutations (*e.g.*, replacing the 5'-exon hairpin with a sequence of equal length derived from the *tet*$^R$ gene, changing T-5 to any other nucleotide, or changing G-6 to T) appeared to enhance RNA splicing. These findings differ from those for intron mobility assays, where the same mutations in the 5' exon of the DNA target site strongly inhibited intron mobility compared to the WT TS34 site (see Figure 22C). Interestingly, changing one side of the hairpin match the other side appeared to significantly decrease the RNA splicing efficiency. This mutation was the only one that leaves no predicted secondary structure in the 5' exon, and it is possible that this feature or the repeat of part of the hairpin sequence have a detrimental effect on RNA splicing. Considered together, these findings indicate that recognition of 5'-exon hairpin is required for intron mobility but not for RNA splicing.

## CONCLUSIONS

In this chapter, I characterized DNA target site recognition and RNA splicing of the *G. stearothermophilus* GsI-IIC, a group IIC intron that has proliferated to high copy number within its host genome. I found that GsI-IIC, like other group IIC introns, inserts downstream of DNA hairpins at sites having short IBS1 and IBS3 sequences recognized by EBS1 and EBS3 sequences within the intron. Both the distribution of genomic insertion sites and *in vivo* mobility assays indicate that insertions occur preferentially into the lagging templates strand, presumably in single-stranded DNA regions at a DNA replication fork, which facilitate formation of the DNA hairpin and enable direct use of lagging strand primers. Both the genomic target site distribution and mobility assays also show that insertions can occur into the leading template strand at lower frequency.

Analysis of genomic insertion sites suggests that insertion specificity is determined primarily by the recognition of the hairpin rather than IBS1 and IBS3 interactions. All of the genomic insertion sites have an upstream hairpin, but some form only 2 of 4 EBS1/IBS1 base pairs and have mismatches at IBS3. This could reflect that retrohoming in *G. stearothermophilus* ordinarily occurs at high temperatures, which promote DNA melting, thereby facilitating the formation of hairpins on the separated strands, while decreasing the contribution of the base-pairing interactions.

*In vivo* selection and mutagenesis experiments identify features of the DNA hairpin and adjoining regions that contribute to DNA target site recognition. Thus, the most efficiently recognized hairpins consist of two stable stems separated by an elbow region, which presumably enables some bending. However, a bulged nucleotide in the elbow may be detrimental because HP22, the worst of the three genomic target sites tested, has a bulged G-residue in this region. The *in vivo* selections with the TS-34 hairpin show selection against base pairing at the GT elbow, and replacement of the elbow and adjoining AT base pair with two GC pairs to make a stable continuous stem inhibits mobility by 5-fold. Surprisingly, I also found that T-5, which is located downstream of the hairpin adjacent to IBS1, contributes strongly to DNA target site recognition. T-5 is conserved in all 34 GsI-IIC target sites, and mutations of T-5 to any other residue decrease mobility efficiencies by >250-fold.

Although various mutations in the 5'-exon hairpin or T-5 nucleotide of the TS34 target site significantly decreased intron mobility efficiencies, all but one of the same mutations had no strong negative effect on either protein-dependent or self-splicing. Only the 5'-exon mutation created by changing one side of the hairpin to match the other side significantly hindered RNA splicing, possibly reflecting an idiosyncratic effect of the mutation on the folding of the precursor RNA. Overall, my results suggest that the 5'-exon hairpin and T-5 are not recognized in the same way for intron mobility and RNA splicing. A likely possibility is that these features in DNA target sites are recognized by the intron-encoded RT, similar to subgroup IIA and IIB introns where distal 5'-exon features upstream of IBS2 are recognized by the intron-encoded protein for intron mobility but not for RNA splicing[1].

Bacterial group II introns that have inserted outside of essential genes frequently degenerate, presumably reflecting that intron mobility is deleterious to the host, so that strains carrying active introns are lost by purifying selection[33]. Further, in previously studied cases where bacterial group II introns have proliferated to high copy number, a significant proportion of the copies are present as twintrons in which one copy of the intron has inserted into another[34]. Surprisingly, despite being inserted outside of genes, nearly all the GsI-IIC introns in the *G. stearothermophilus* genome are potentially active, as judged by retention of conserved structural features of both the intron RNA and intron-encoded RT. Further, there are no GsI-IIC twintrons. The lack of GsI-IIC twintrons is readily explained by the lack of suitable hairpins targets within the intron. The finding that most copies of the GsI-IIC intron appear to be functional despite being inserted downstream of transcription terminators could reflect either recent insertion or purifying selection against inactive copies of the intron. The latter could in turn reflect that the non-coding regions in which the intron has inserted are ordinarily transcribed at a low level and have important functions, which requires retention of RNA splicing activity to reconstitute functional transcripts. Finally, the ability of GsI-IIC to proliferate to relatively high copy number in a bacterial genome is relatively rare for group II introns, which are typically found at one or two copies per genome[35]. The high copy number of GsI-IIC could be due to the large number of transcriptional terminator hairpin sequences that provide suitable target sites combined with mobility at high temperatures, which makes the intron less dependent on base-pairing interactions that would provide more specificity for DNA insertion.

## EXPERIMENTAL DESIGN AND METHODS

### Recombinant plasmids

pGsI2C_35/32 is used for *in vitro* transcription of GsI-IIC RNAs for RNA splicing experiments and contains a 656-nt GsI-IIC-ΔORF intron (nucleotides 551-1791 encoding the intron ORF replaced by CGC) with short flanking 5' and 3' exons (35 and 32 nts, respectively) cloned downstream of a phage T3 promoter between the HindIII and BamHI sites in pUC19 (New England BioLabs). The 5' exon is derived from target site 34, and the 3' exon is derived from target site 23. Derivatives of GsI2C-35/32 with 5' exons from different target sites or mutations in the TS34 hairpin constructed by PCR stitching of two overlapping PCR products with outside primers Stch_HindIII_T3_For and Stch_BamHI, which appended HindIII and BamHI sites, respectively. 5' fragments with different mutations were produced by hybridizing top- and bottom-strand primers, while the common 3' fragment was created via PCR with a set of primers (GsI2Cintron_F and Stch_BamH1) were The PCR amplicons were purified with Wizard PCR Clean-Up System (Promega) then digested with HindIII (NEB) and BamHI (NEB). The digested inserts were purified again and ligated into pGsI2C_35/32 digested with HindIII and BamHI.

The GsI-IIC RT, used for RNA splicing assays, was expressed from pGsI-IIC-MRF and purified as described[26]. The construct expresses the GsI-II RT with an N-terminal maltose-binding protein rigid fusion required to maintain solubility of the protein.

pADC2X-GsI-IICΔORF+T7, the intron-donor plasmid used for mobility assays, contains a cassette consisting of a shortened GsI-IIC intron with most of the RT ORF removed and the RT ORF cloned downstream of a $T7_{lac}$ promoter in a pADC184-derived vector. It was constructed in two steps. First, the ORF encoding the IEP was amplified from pETGsI-IIC DNA[31] using Phusion PCR mix with 5' primer GeoI2ORF5+SDPst (AAACTGCAGGAAGGAGATATACATATGGCTTTGTT), which appends PstI and NdeI sites and a Shine-Dalgarno sequence from pET3, and 3' primer GeoI2ORF3Xho (GGACTCGAGTCAACCTTGACGGAGTTCGA), which appends a XhoI site. The PCR product was digested with PstI and XhoI, band isolated, and cloned into pACD2X[31] cut with the same enzymes thereby replacing the LtrA sequence and producing pADC2XGeoRT. The GsI-IIC intron was amplified by two PCRs that separately amplify 5' and 3' segments of the intron, while introducing a ~1.4-kb deletion in the ORF coding sequence (see above). These PCRs used outside primers that append short flanking exons (57 nts 5' exon derived from target site 34 and 32 nts 3' exon derived from target site 23) and unique cloning sites (XbaI (5') and PstI (3')) together with overlapping internal primers that replace the intron ORF in DIVb with a T7 promoter sequence and an MluI site. The two PCR products were then cloned between XbaI and PstI sites of pADC2XgeoRT resulting in pADC2X-GsI-IICΔORF+T7.

Intron-recipient plasmids used in mobility assays contain GsI-IIC intron insertion sites (positions -40 to +20) cloned into pBRR3A and pBRR3B, which differ in the orientation of the target site and *tet*$^{R}$ gene relative to the replication origin[31]. They were constructed by replacing the Ll.LtrB target site in plasmids pBRR3A-ltrB (LEAD) and pBRR3B-ltrB (LAG) with target sites form *G. stearothermophilus* strain 10. pBRR DNA was digested with AatII and EcoRI and the backbone gel isolated and purified. Target sites were made from annealed top- and bottom-strand oligonucleotides that already contain the AatII and EcoRI overhangs. The annealed oligonucleotides were directly ligated into cut pBRR DNA resulting in pBRR-GeoTS34-LEAD and pBRR-GeoTS34-LAG. Recipient plasmids containing the TS7 and TS22 5' exon or mutated version of the

TS34 5’ exon were made similarly by starting with annealed top and bottom strand oligonucleotides containing the modifications.

**Intron mobility assays**

Mobility assays were done in *E. coli* HMS174(DE3) (Novagen, Madison, WI) grown in LB medium with antibiotics added as required at the following concentrations: ampicillin, 100 mg/ml; chloramphenicol, 25 mg/ml; tetracycline, 25 mg/ml. Cells that had been co-transformed with the Cap$^R$ donor (pADC2X-GsI-IIC-ΔORF+T7) and Amp$^R$ recipient plasmids (pBRR-GeoTS-LEAD or pBRR-GeoTS-LAG) were inoculated into 5 ml of LB medium containing chloramphenicol and ampicillin and grown with shaking (200 rpm) overnight at 37ºC. A small portion (50 μl) of the overnight culture was inoculated into 5 ml of fresh LB medium containing the same antibiotics and grown for 1 h as above. The cells were then induced by adding 1 ml of fresh LB medium containing the same antibiotics and 3 mM IPTG (500 μM final) and incubating for 1 h at 48ºC. The cultures were then placed on ice, diluted with ice-cold LB, and plated at different dilutions onto LB agar containing ampicillin or ampicillin plus tetracycline. After incubating the plates overnight at 37ºC, the mobility efficiency was calculated as the ratio of (Tet$^R$+Amp$^R$) / Amp$^R$ colonies.

**Hairpin selection**

For selection experiments, the target site in pBRR-TS34-LAG was replaced with one in which positions -5 to -38 were doped with 70% of the wild-type nucleotide and 10% of each other nucleotide. To construct this plasmid, the top strand oligonucleotide (Geo34TOP) was doped such that each of the doped nucleotide positions has 70% of the wild-type nucleotide residue and 10% of each of the three mutant nucleotide residues. The complementary strand was made by annealing a short primer (Geo34Bot) to the fixed 3’ end of the doped oligonucleotide and filling in the bottom strand across the doped region using the DNA-dependent DNA polymerase activity of the GsI-IIC RT. (Reaction conditions: 98 ºC 2 min -> 50 ºC 2 min -> 72 ºC 5 min for a single cycle). GsI-

IIC RT was used because it was able to efficiently synthesize the complementary strand through the hairpin DNA while other DNA polymerases (Phusion (NEB), Klenow (NEB), and Taq (NEB) failed to complete bottom strand synthesis likely due to stable secondary structure. The product was then cleaned with a MinElute PCR Purification Kit (Qiagen), digested with AatII and EcoRI-HF (NEB), and cloned between the corresponding sites of pBRR3ltrbLAG.

For *in vivo* selection, the recipient plasmid containing the doped insert was electroporated into *E. coli* HMS174 (DE3) and transformants were selected for the Amp$^R$ marker on the plasmid. The transformants were grown and made electrocompetent for introduction of the donor plasmid pACD2-GsI-IIC-ΔORF+T7. The HMS174 (DE3) cells carrying both donor and recipient plasmids were used for mobility assay as described above. Illumina adaptors (NG_GsI-IIC_HomingProd_3_For and NG_GsI-IIC_HomingProd_3_Rev) were appended to the tetracycline-selected homing product through PCR. The PCR product was then purified with Agencourt AMpure XP and Illumina tails were added to the cleaned sample via PCR (6 cycles). The PCR product was cleaned with Agencourt AMpure XP again ands sequenced on an Illumina HiSeq 4000 paired-end to obtain 150-nt paired-end reads at the UT Austin Genomic Sequencing and Analysis Facility (GSAF). The selected library had 6,597,196 raw sequences and the unselected target library had 26,745,370 raw sequences. The raw data were then filtered by their 16-nt barcode so that the filtered sequences will only contain unique barcodes. Ambiguous sequences with Ns in the doped region and/or barcode were removed. The expected length of the doped region is 33-nt, but negligible fractions of the data had doped regions with shorter or longer than 33 nucleotides (<1% of the selected and ~1% of the unselected library). These sequences were also removed from further analysis. Galaxy was used to convert the raw NextGen sequencing data to FASTA format and for further trimming of the sequences. Python scripts were used to obtain nucleotide and pair frequencies of the doped region. The resulting nucleotide and pair frequencies were plotted using Excel.

**RNA splicing assays**

Splicing assays were performed by using an internally $^{32}$P-labeled *in vitro* transcript containing a 656-nt GsI-IIC-ΔORF intron flanked by a 35-nt 5' exon (positions -35 to -1 of TS34) and 32-nt 3' exon (position +1 to +32 of TS23). This precursor RNA was transcribed from an amplicon generated by PCR from template plasmids constructed using pGsI2C_35/32 as described above, which contains the previously described intron and flanking exons cloned in a pUC19-based vector, using a 5' primer that adds a phage T3 RNA polymerase promoter sequence upstream of the 5' exon. *In vitro* transcription was done by using phage T3 polymerase (60 U per 100 μL reaction; ThermoFisher Scientific) with 2 mM of each NTP and 123.8 nM of [α-$^{32}$P] UTP (3,000 Ci/mmol; Perkin-Elmer) for 2.5 h at 37°C. 2 mM dTTP was included in the reaction mixture to sequester excess free $Mg^{2+}$ ions, which increase hydrolytic splicing during transcription. The transcription reaction (200 μL) was treated with 8 μL RNase-free DNase I (Thermo Scientific) for 15 min at 37°C and cleaned up by using a MEGAclear Transcription Clean-Up Kit (Thermo Scientific).

Splicing reactions were carried out by incubating the $^{32}$P-labeled precursor RNA (20 and 40 nM) and purified GsI-IIC-MRF RT[36] at various molar ratios of 1:0.5, 1:1, 1:2, 1:3, 1:5 in 20 μL of 450 mM KCl, 5 mM $MgCl_2$, and 20 mM Tris-HCl pH 7.5 at 50 ºC. The reactions were initiated by adding protein, which had been pre-warmed to 50 ºC for 30 s (confirmed to result in no loss of activity), and terminated by adding a mixture of 30-μL ice-cold phenol-chloroform-isoamyl alcohol and 0.5-μL 500 mM EDTA. Time-course reactions were stopped at various time points, and end point splicing reactions were stopped after 10 min with phenol CIA (phenol:chloroform:isoamylalcohol, 25:24:1). After centrifugation, the phenol-extracted aqueous phase was mixed with an equal volume of Gel Loading Buffer II (Ambion). The splicing products were analyzed in a denaturing 4% polyacrylamide gel, which was dried and scanned using a phosphorimager (Typhoon FLA 9500; GE Healthcare Life Sciences). Band intensities were quantified by using ImageQuant TL (GE Healthcare Life Sciences). Data were normalized and fitted to a two exponential function ($Y$ = plateau + $a*e^{-K1t}$ + $b*e^{-K2t}$) with Prism6 (GraphPad Software)

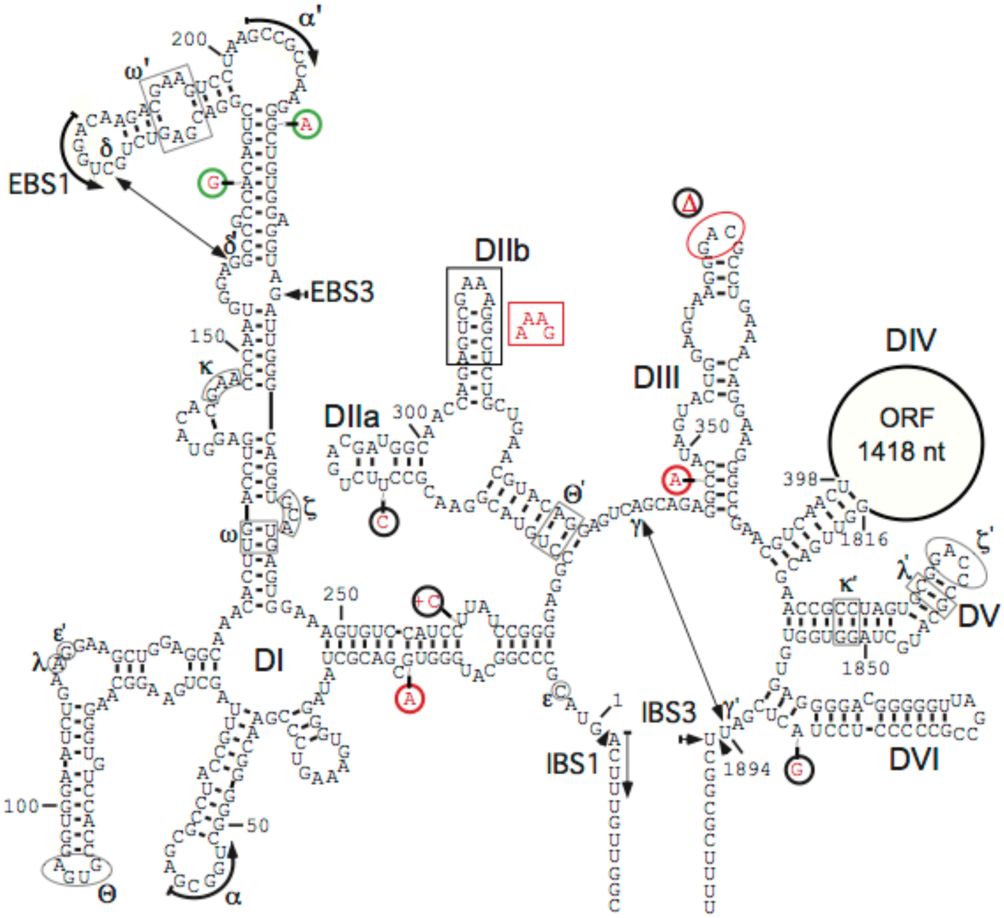


**Figure 17. GsI-IIC intron RNA Secondary structure of the *Geobacillus stearothermophilus* strain 10 group IIC intron.** The intron RNA is comprised of six secondary structure domains (DI-VI). The position of the RT ORF in DIV is indicated by a loop. Sequence variations between the different copies of the intron in the strain 10 genome are indicated in red. Single base changes are circled (green circle, compensatory change in helix; red circle, non-compensatory change in helix; black circle, insertion, deletion, or change in an unpaired position). EBS1 and 3, exon-binding sites 1 and 3; IBS1 and 3, intron binding sites 1 and 3. Greek letters indicate sequence motifs involved in tertiary structure interactions. Figure prepared by Dr. Georg Mohr.

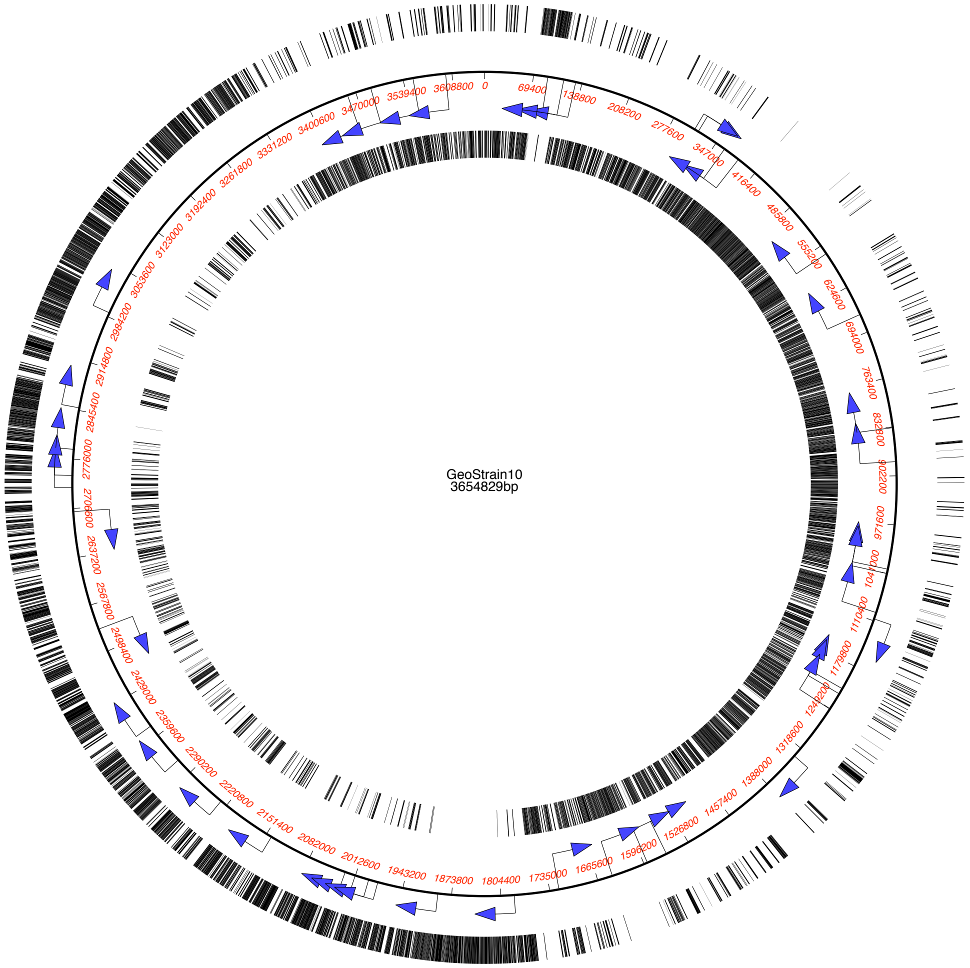


**Figure 18. Position of GsI-II2C introns inserted in genome of *G. stearothermophilus* strain 10.** Introns are indicated by arrows, and insertion site is indicated by vertical line. Top and bottom strand insertions are shown above and below the line, respectively. Vertical dashes on the outside and inside show coding regions on top and bottom strands, respectively. Figure prepared by Dr. Georg Mohr.

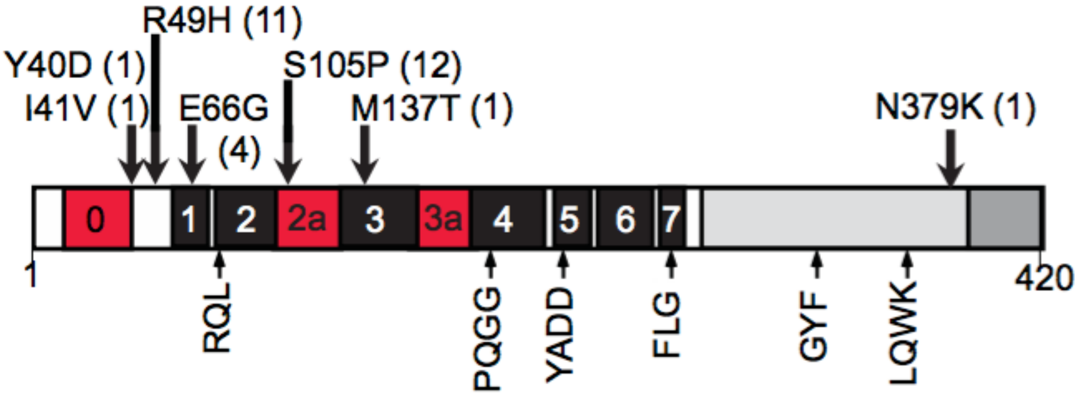


**Figure 19. Schematic of the *Geobacillus stearothermophilus* group IIC intron RT protein.** Conserved sequence blocks found in all RTs are denoted RT1-7 and indicated by black boxes[37]. RT-0, 2a, and 3a (red) indicate an extra N-terminal region and insertions between conserved RT sequence blocks found in group II intron and non-LTR-retrotransposon RTs. Sequence variations in the 45 copies in the genome are indicated on top with the number of occurrences in parenthesis. Figure prepared by Dr. Georg Mohr.

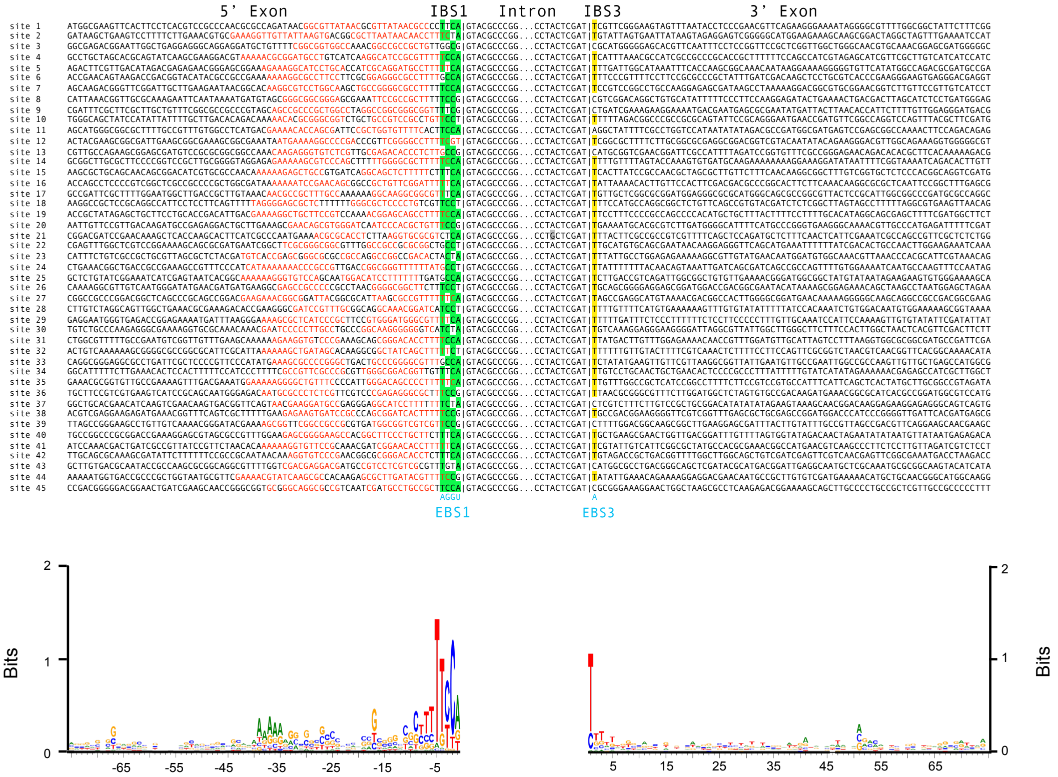


**Figure 20. Sequence alignment of exon-intron junction sequences in the *G. stearothermophilus* strain 10 genome with a sequence logo at the bottom showing sequence conservation in the 5’ and 3’ exons.** Red letters indicate nucleotides that potentially base pair to form the hairpin structure upstream of the intron-insertion site. Green highlighting shows IBS1 sequences that potentially base pair to the EBS1 sequences in the intron. Yellow highlights indicate expected C or T residues at the IBS3 position. Grey highlight indicates mutation of bulged A residue in DVI. The vertical line indicates the exon-intron boundary. The logo was generated with WebLogo3 using standard parameters[31]. Figure prepared by Dr. Georg Mohr.

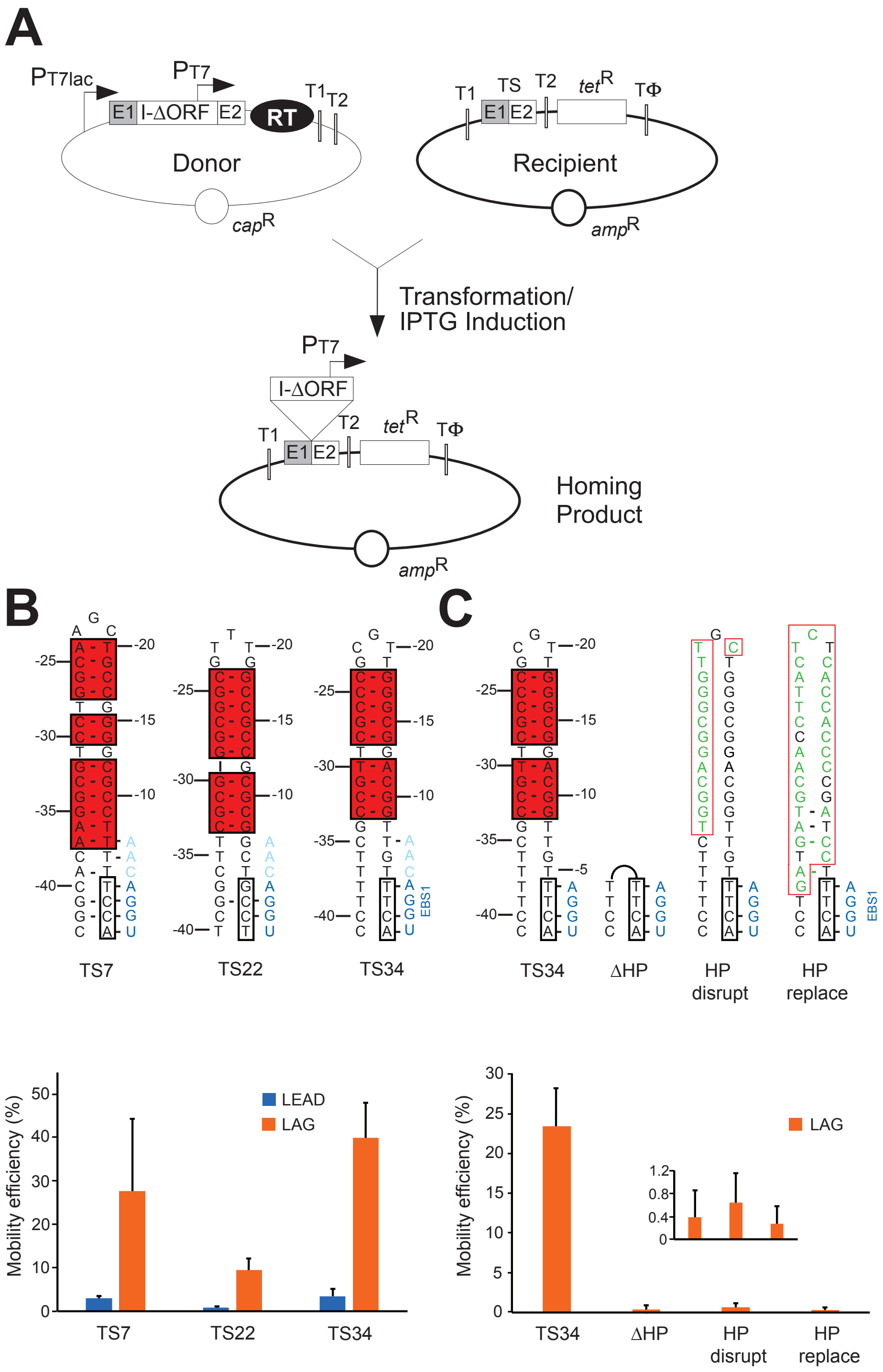

A
PT7lac
PT7
E1
I-ΔORF
E2
RT
T1
T2
Donor
capR
T1
TS
T2
tetR
TΦ
E1
E2
Recipient
ampR
Transformation/
IPTG Induction
PT7
I-ΔORF
T1
E1
E2
T2
tetR
TΦ
Homing
Product
ampR
B
TS7
TS22
TS34
-20
-25
-15
-30
-10
-35
-40
EBS1
C
TS34
ΔHP
HP
disrupt
HP
replace
-5
EBS1
Mobility efficiency (%)
LEAD
LAG
0
10
20
30
40
50
TS7
TS22
TS34
Mobility efficiency (%)
LAG
0
5
10
15
20
25
30
0
0.4
0.8
1.2
TS34
ΔHP
HP
disrupt
HP
replace

**Figure 21.** ***In vivo* mobility assay and requirement for the 5' hairpin structure.** (*A*) Schematic of the assay. The Cap$^R$ intron-donor plasmid (pADC2X-GsI-IIC-ΔORF+T7) uses a T7$_{lac}$ promoter (PT7lac) to express a group II intron RNA with short flanking 5' and 3' exons (E1 and E2, respectively) and the group II RT cloned downstream of E2. The group II intron, which has a T7 promoter sequence (PT7) inserted near its 3' end, integrates into a target site (TS; the ligated E1–E2 sequence) cloned in a compatible Amp$^R$ recipient plasmid (pBRR-TS34-LEAD or pBRR-TS34-LAG) upstream of a promoterless *tet*$^R$ gene, thereby introducing the T7 promoter and activating that gene. The assays are done in *E. coli* HMS174 (DE3), which contains an IPTG-inducible T7 RNA polymerase. Intron expression is induced with IPTG for 1 h at 48ºC, and mobility efficiencies are calculated as the ratio of (Tet$^R$+Amp$^R$)/Amp$^R$ colonies. T1 and T2 are *E. coli* rRNA transcription terminators and TΦ is a T7 transcription terminator. (B) Mobility assays with three different target sites cloned in either the LEAD or LAG recipient plasmids. The top shows 5'-exon sequences of three target sites (TS7, TS22, and TS34 5' exons with TS12 3' exons, denoted TS7, TS22, and TS34, respectively). Base-paired regions of the 5'-exon hairpin are highlighted by red shading, and the IBS1 sequence is boxed. The complementary EBS1 sequence in the intron and three adjoining intron bases that could potentially base pair with 5'-exon sequence in some cases are shown in dark and light blue, respectively. Base pairs are indicated by dashes. The bar graphs below show the mobility efficiency of three target sites cloned in either the LEAD (blue) or LAG (orange) recipient plasmids. The data are the mean for three experiments with the error bars indicating the standard deviation. (C) Mutations in the TS34 target site and their effect on mobility efficiency. The top shows 5'-exon sequences of the TS34 target sites and three mutant target sites, with nucleotides that differ from the TS34 target site shown in green. Features of the target sites are depicted as in panel B. The bar graph below shows the mobility efficiencies of the three mutant target sites cloned in the LAG recipient plasmid compared to that of the TS34 target assayed in parallel. The inset in the plot at the bottom right shows an expanded scale for mutations that have very low

mobility efficiencies. The data are the mean for three experiments with the error bars indicating the standard deviation.

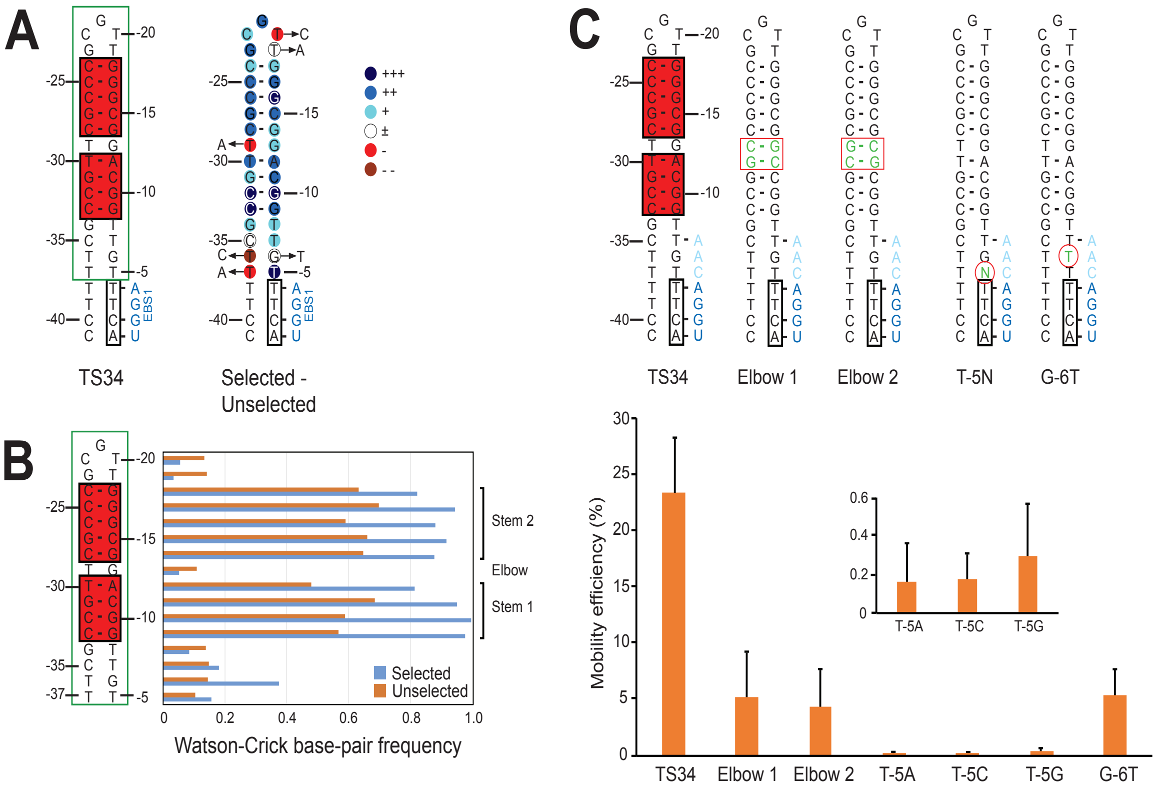


**Figure 22.** ***In vivo* selection of 5' exon hairpin region.** The selection was done using the WT donor plasmid (pACD2-GsI-IIC-ΔORF+T7) and a recipient plasmid (pBRR-GeoTS34-LAG) in which the portion of the target site containing the hairpin structure and flanking regions was partially randomized (doped at 70% of the wild-type nucleotide residue and 10% of each of the other nucleotide residues). Cells were grown in LB medium containing tetracycline overnight then plasmid DNA was isolated. PCR reactions using a primer in the recipient (TargetSeq) and one in the intron (GeoI2D3LoopUp) were done to amplify homing events from the plasmid pool. Subsequently, Illumina tags were added by a second PCR reaction and sequenced by an llumina HiSeq 4000 to obtain 150-nt paired-end reads. The selected library had 6,597,196 raw sequences and the unselected target library had 26,745,370 raw sequences. The unselected library was sequenced to assess differences in the doping frequency at different positions. (*A*) Sequence and predicted structure of the wild-type TS34 target site and degree of selection at different nucleotide positions. The 5'-exon region of TS34 is shown to the left, with the base-

paired region of the hairpin highlighted in red and the region partially randomized for the selection enclosed in a green box. The results of the selection are summarized to the right: +++, nucleotide present in >99% of selected sequences; ++, nucleotide present at >15% higher frequency in selected than in the unselected sequences but <99% of selected sequences; +, nucleotide present at >5-14% higher frequency in the selected than in unselected sequences; ±, nucleotide present at similar frequencies (±4%) in selected and unselected sequences; -, nucleotide present at 5-14% lower frequency in selected than in unselected sequences; --, nucleotide present at >15% lower frequency in selected than unselected sequences. (*B*) Selection for or against base-pairing within the hairpin and flanking regions. The bar graphs show the frequency of base-paired nucleotides at each position in the 5' exon region of the TS34 hairpin in the selected and unselected sequences (blue and red, respectively). Brackets on the right delineate the upper and lower stems in which base pairing is selected for in active target sites, and the TG elbow in which base pairing is selected against. (*C*) Mobility assays with different mutants. The top shows 5'-exon sequences of wild-type TS34 and mutant target sites depicted schematically as in panel A (or Figure 3). Mutations in the hairpin are shown in green within red circles or boxes. The bar graphs below show mobility efficiencies for wild-type and mutant target sites. The inset shows an expanded scale for the T-5 mutations, which have very low mobility efficiencies. The data are the mean for three-independent experiments with the error bars indicating that standard deviation.

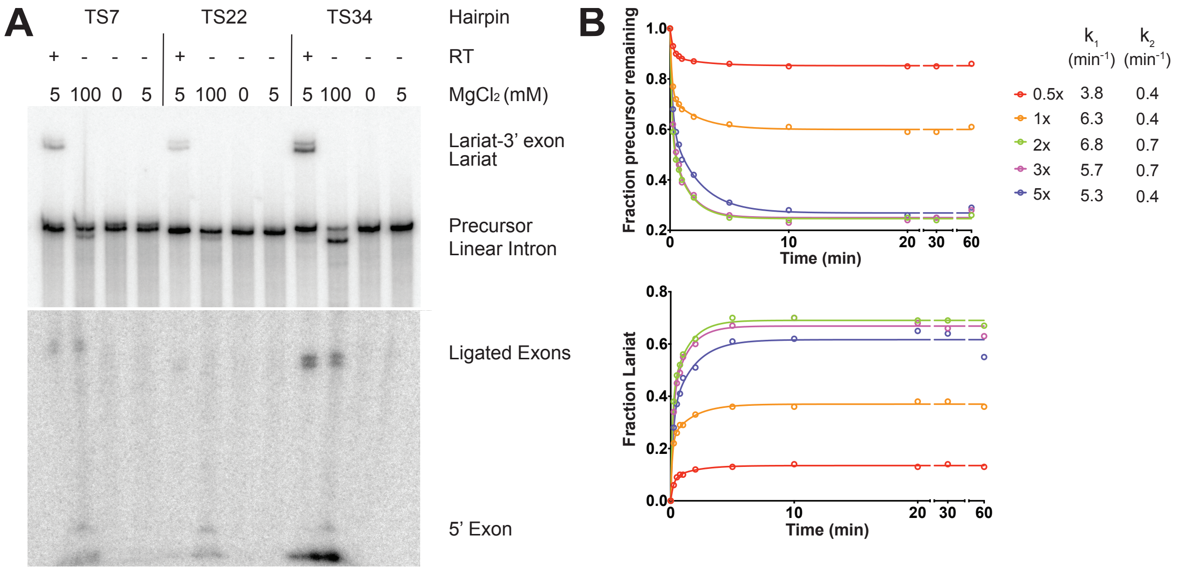


**Figure 23. Protein-dependent and self-splicing of the GsI-IIC intron.** (*A*) Protein-dependent and self-splicing of the GsI-IIC intron with different hairpins. Each RNA was incubated for 10 min at 50 ºC with the GsI-IIC RT (20 nM) in reaction medium containing 5 mM Mg$^{2+}$ or without protein in reaction medium containing high Mg$^{2+}$ (100 mM, self-splicing conditions), no Mg$^{2+}$ (non-splicing control), or in reaction medium containing 5 mM Mg$^{2+}$ (5 mM; control for self-splicing under protein-dependent splicing conditions). Bands are identified to the right of the gel. The gel is split to show top and bottom regions. (*B*) Time courses. 40 nM GsI-IIC RNA was incubated with various amounts of purified GsI-IIC RT (20 nM, 40 nM, 80 nM, 120 nM, and 200 nM) giving ratios of RNA to protein of 0.5x, 1x, 2x, 3x, and 5x. Samples were taken at 0, 0.25, 0.5, 0.75, 1, 2, 5, 10, 20, 30, and 60 min and run on a 4% denaturing polyacrylamide gel. The gel was dried and scanned with a PhosphorImager. Rate constants were obtained by fitting the data to a two exponential equation. The curves at the top shows disappearence of precursor RNA, and the curves at the bottom show appearance of lariat RNA.

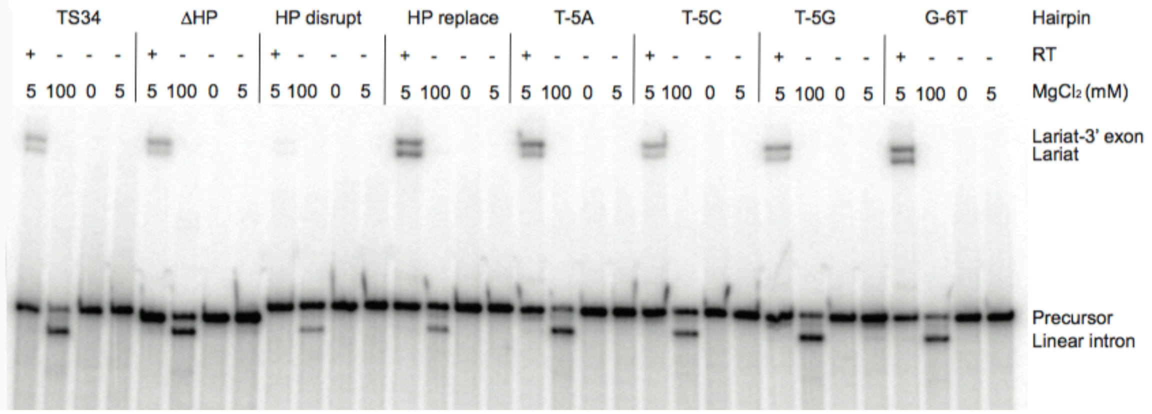
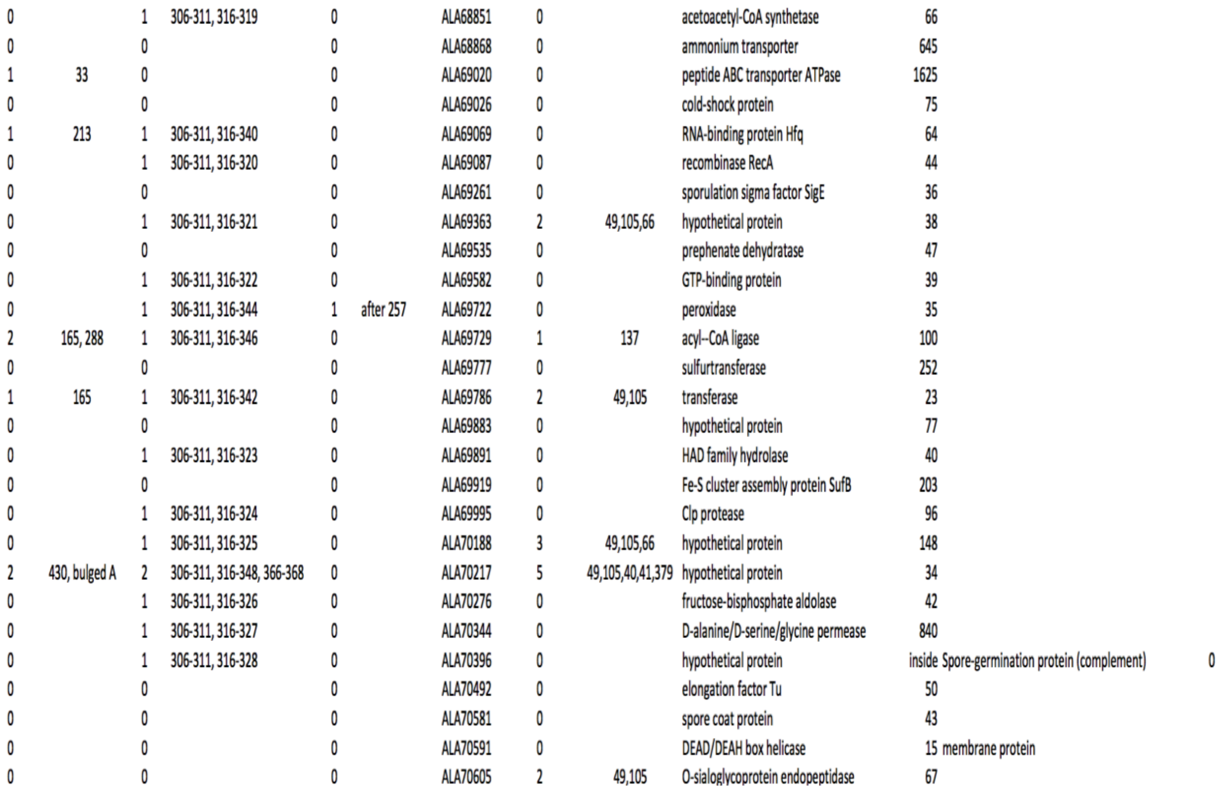


**Figure 24.** Effect of mutations in the 5'-exon hairpin region on protein-dependent and self-splicing of GsI-IIC intron RNA. Splicing reactions for the indicated mutants were carried out for 10 min at 50 °C, as described in Figure 23. Bands are identified to the right of the gel.

**Table 2. GsI-IIC introns in *Geobacillus stearothermophilus* strain 10 (accession number: CP008934).**



| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | 1 | 306-311, 316-319 | 0 | | ALA68851 | 0 | | acetoacetyl-CoA synthetase | 66 | | |
| 0 | | 0 | | 0 | | ALA68868 | 0 | | ammonium transporter | 645 | | |
| 1 | 33 | 0 | | 0 | | ALA69020 | 0 | | peptide ABC transporter ATPase | 1625 | | |
| 0 | | 0 | | 0 | | ALA69026 | 0 | | cold-shock protein | 75 | | |
| 1 | 213 | 1 | 306-311, 316-340 | 0 | | ALA69069 | 0 | | RNA-binding protein Hfq | 64 | | |
| 0 | | 1 | 306-311, 316-320 | 0 | | ALA69087 | 0 | | recombinase RecA | 44 | | |
| 0 | | 0 | | 0 | | ALA69261 | 0 | | sporulation sigma factor SigE | 36 | | |
| 0 | | 1 | 306-311, 316-321 | 0 | | ALA69363 | 2 | 49,105,66 | hypothetical protein | 38 | | |
| 0 | | 0 | | 0 | | ALA69535 | 0 | | prephenate dehydratase | 47 | | |
| 0 | | 1 | 306-311, 316-322 | 0 | | ALA69582 | 0 | | GTP-binding protein | 39 | | |
| 0 | | 1 | 306-311, 316-344 | 1 | after 257 | ALA69722 | 0 | | peroxidase | 35 | | |
| 2 | 165, 288 | 1 | 306-311, 316-346 | 0 | | ALA69729 | 1 | 137 | acyl--CoA ligase | 100 | | |
| 0 | | 0 | | 0 | | ALA69777 | 0 | | sulfurtransferase | 252 | | |
| 1 | 165 | 1 | 306-311, 316-342 | 0 | | ALA69786 | 2 | 49,105 | transferase | 23 | | |
| 0 | | 0 | | 0 | | ALA69883 | 0 | | hypothetical protein | 77 | | |
| 0 | | 1 | 306-311, 316-323 | 0 | | ALA69891 | 0 | | HAD family hydrolase | 40 | | |
| 0 | | 0 | | 0 | | ALA69919 | 0 | | Fe-S cluster assembly protein SufB | 203 | | |
| 0 | | 1 | 306-311, 316-324 | 0 | | ALA69995 | 0 | | Clp protease | 96 | | |
| 0 | | 1 | 306-311, 316-325 | 0 | | ALA70188 | 3 | 49,105,66 | hypothetical protein | 148 | | |
| 2 | 430, bulged A | 2 | 306-311, 316-348, 366-368 | 0 | | ALA70217 | 5 | 49,105,40,41,379 | hypothetical protein | 34 | | |
| 0 | | 1 | 306-311, 316-326 | 0 | | ALA70276 | 0 | | fructose-bisphosphate aldolase | 42 | | |
| 0 | | 1 | 306-311, 316-327 | 0 | | ALA70344 | 0 | | D-alanine/D-serine/glycine permease | 840 | | |
| 0 | | 1 | 306-311, 316-328 | 0 | | ALA70396 | 0 | | hypothetical protein | inside | Spore-germination protein (complement) | 0 |
| 0 | | 0 | | 0 | | ALA70492 | 0 | | elongation factor Tu | 50 | | |
| 0 | | 0 | | 0 | | ALA70581 | 0 | | spore coat protein | 43 | | |
| 0 | | 0 | | 0 | | ALA70591 | 0 | | DEAD/DEAH box helicase | 15 | membrane protein | |
| 0 | | 0 | | 0 | | ALA70605 | 2 | 49,105 | O-sialoglycoprotein endopeptidase | 67 | | |

Introns are numbers by insertion position in the *G*. *stearothermophilus* genome. Intron 3 is set as the standard for comparisons. The numbers of mismatches, deletions and insertions compared to intron 3 along with the position of each change in the intron nucleotide sequence are shown in columns 3 to 8. Columns 9 to 11 show the accession number for each RT protein, the number of mismatches against the RT of intron 3, and the position of each mismatch in the RT ORF. Columns 12 to 15 show the genes flanking each intron insertion, with numbers indicating the nucleotide distance from the 5' or 3' end of each intron insertion. Inside indicates that the intron is inserted in the annotated gene. Table prepared by Dr. Georg Mohr.

## Table 3. Oligonucleotides

| Name | Sequence |
|---|---|
| GEOI2ORF5+SD PST | AAACTGCAGGAAGGAGATATACATATGGCTTTGTT |
| GEOI2ORF3XHO | GGACTCGAGTCAACCTTGACGGAGTTCGA |
| GSI2#34TOP | CCTTTTCGCCGTTCGCCCGCGTTGGGCGGACGGTTGTTTCATCGGCGCTTTTCTTGCGGCGG |
| GSI2#34BOT | AATTCCGCCGCAAGAAAAGCGCCGATGAAACAACCGTCCGCCCAACGCGGGCGAACGGCGAAAAGGACGT |
| GSI2#22TOP | CTCGGCTTCGCGGGCGGCGTTTGGCCGCCGCGCGGCTGCCTTTTATTGCCTGGAGAGAAAAG |
| GSI2#22BOT | AATTCTTTTCTCTCCAGGCAATAAAAGGCAGCCGCGCGGCGGCCAAACGCCGCCCGCGAAGCCGAGACGT |
| GSI2#7TOP | CCACAAGGCGTCCTGGCAAGCTGCCGGGGCGCCTTTTTCCATGTTGCTCGGCGCGGATGGAG |
| GSI2#7BOT | AATTCTCCATCCGCGCCGAGCAACATGGAAAAAGGCGCCCCGGCAGCTTGCCAGGACGCCTTGTGGACGT |
| primer A | /Cy5/CATACAACGCCTTTTTCTCTCCAGG |
| NG_GsI-IIC_HomingProd_3_For | TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNNNGAAGCAACGGCCCGACGTC |
| NG_GsI-IIC_HomingProd_3_Rev | GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNGCACCCATGCCGGGCGTAC |
| Geo#3TOP | CCCGACCTTDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDTTCATCGGCGCTTTTCTTGCGGCGAATTCTAA (D for doped) |
| Geo#3Bot | TTAGAATTCGCCGCAAGAAAAGCG |
| TargetSeq | ATGCGAGAGTAGGGAACTGC |
| GeoI2D3LoopUp | CTCCTGTTTCCAGGCCTCCCCGGATAAGG |
| T3GSIc-F and | GGAGAATTAACCCTCACTAAAGGCCGTTCGCCCGCGTT |
| GsI2c35 3-EX | CATACAACGCCTTTTTCTCTCCAGG |
| HP34_dsrt_mirror | CACCCATGCCGGGCGTAC TGAAACAACCGTCCGCCCAAC |
| HP34_minus5TtoC | CACCCATGCCGGGCGTAC TGAAGCAACCGTCCGCCC |
| HP34_dsrt_random | CACCCATGCCGGGCGTAC TGAAAGGATCGGGGTGGTG |
| 5AND6_universal | AATTAACCCTCACTAAA G GCCGTTCGCCCGCGTTGG |
| elbow_CGGC_F | AATTAACCCTCACTAAA G GCCGCGCGCCCGCGTTGG |
| elbow_GCCG_F | AATTAACCCTCACTAAA G GCCGGCCGCCCGCGTTGG |
| -5TtoC_F | AATTAACCCTCACTAAAGGCCGTTCGCCCGCGTTGGGCGGACGGTTGCTTCA |
| HP_dsrp_mirror_F | AATTAACCCTCACTAAA G TGGCAGGCGGGTCGTTGGGCGGACGGTTGTTTCA |
| HP_dsrp_rndm_F | AATTAACCCTCACTAAA G TGATGCAACCTTACTCACCACCCCGATCCTTTCA |

## References

1 Lambowitz, A. M. & Zimmerly, S. Group II Introns: Mobile Ribozymes that Invade DNA. *Cold Spring Harbor Perspectives in Biology* **3**, doi:10.1101/cshperspect.a003616 (2011).

2 Lambowitz, A. M. & Belfort, M. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiology spectrum* **3** (2015).

3 Enyeart, P. J., Mohr, G., Ellington, A. D. & Lambowitz, A. M. Biotechnological applications of mobile group II introns and their reverse transcriptases: gene targeting, RNA-seq, and non-coding RNA analysis. *Mobile DNA* **5**, 2, doi:10.1186/1759-8753-5-2 (2014).

4 Blocker, F. J. *et al*. Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. *Rna* **11**, 14-28, doi:10.1261/rna.7181105 (2005).

5 Matsuura, M. *et al*. A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes & Development* **11**, 2910-2924 (1997).

6 Sashital, D. G., Cornilescu, G. & Butcher, S. E. U2-U6 RNA folding reveals a group II intron-like domain and a four-helix junction. *Nat Struct Mol Biol* **11**, 1237-1242, doi:http://www.nature.com/nsmb/journal/v11/n12/suppinfo/nsmb863_S1.html (2004).

7 Fica, S. M. Evidence for a group II intron-like catalytic triplex in the spliceosome. **21**, 464-471, doi:10.1038/nsmb.2815 (2014).

8 Fica, S. M. *et al*. RNA catalyses nuclear pre-mRNA splicing. *Nature* **503**, 229-234, doi:10.1038/nature12734 (2013).

9 Gordon, P. M., Sontheimer, E. J. & Piccirilli, J. A. Kinetic characterization of the second step of group II intron splicing: role of metal ions and the cleavage site 2'-OH in catalysis. *Biochemistry* **39**, 12939-12952 (2000).

10 Keating, K. S., Toor, N., Perlman, P. S. & Pyle, A. M. A structural analysis of the group II intron active site and implications for the spliceosome. *Rna* **16**, 1-9, doi:10.1261/rna.1791310 (2010).

11 Gordon, P. M. & Piccirilli, J. A. Metal ion coordination by the AGC triad in domain 5 contributes to group II intron catalysis. *Nature structural biology* **8**, 893-898, doi:10.1038/nsb1001-893 (2001).

12 Marcia, M. & Pyle, A. M. Visualizing group II intron catalysis through the stages of splicing. *Cell* **151**, 497-507, doi:10.1016/j.cell.2012.09.033 (2012).

13 Robart, A. R., Chan, R. T., Peters, J. K., Rajashankar, K. R. & Toor, N. Crystal structure of a eukaryotic group II intron lariat. *Nature* **514**, 193-197, doi:10.1038/nature13790 (2014).

14 Toor, N., Keating, K. S., Taylor, S. D. & Pyle, A. M. Crystal structure of a self-spliced group II intron. *Science* **320**, 77-82, doi:10.1126/science.1153803 (2008).

15 Carignani, G. *et al*. An mRNA maturase is encoded by the first intron of the mitochondrial gene for the subunit I of cytochrome oxidase in S. cerevisiae. *Cell* **35**, 733-742 (1983).

16 Kennell, J. C., Moran, J. V., Perlman, P. S., Butow, R. A. & Lambowitz, A. M. Reverse transcriptase activity associated with maturase-encoding group II introns in yeast mitochondria. *Cell* **73**, 133-146, doi:http://dx.doi.org/10.1016/0092-8674(93)90166-N (1993).

17 Mohr, S., Matsuura, M., Perlman, P. S. & Lambowitz, A. M. A DEAD-box protein alone promotes group II intron splicing and reverse splicing by acting as an RNA chaperone. *Proc Natl Acad Sci U S A* **103**, 3569-3574, doi:0600332103 [pii]10.1073/pnas.0600332103 (2006).

18 Saldanha, R. *et al*. RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. *Biochemistry* **38**, 9069-9083, doi:10.1021/bi982799l (1999).

19 Dlakić, M. & Mushegian, A. Prp8, the pivotal protein of the spliceosomal catalytic center, evolved from a retroelement-encoded reverse transcriptase. *Rna* **17**, 799-808, doi:10.1261/rna.2396011 (2011).

20 Qu, G. *et al*. Structure of a Group II Intron Complexed with its Reverse Transcriptase. *Nat Struct Mol Biol* **23**, 549-557, doi:10.1038/nsmb.3220 (2016).

21 Smith, D., Zhong, J., Matsuura, M., Lambowitz, A. M. & Belfort, M. Recruitment of host functions suggests a repair pathway for late steps in group II intron retrohoming. *Genes & Development* **19**, 2477-2487, doi:10.1101/gad.1345105 (2005).

22 Yao, J., Truong, D. M. & Lambowitz, A. M. Genetic and Biochemical Assays Reveal a Key Role for Replication Restart Proteins in Group II Intron Retrohoming. *PLoS Genet* **9**, e1003469, doi:10.1371/journal.pgen.1003469 (2013).

23 Michel, F. & Ferat, J. L. Structure and activities of group II introns. *Annual review of biochemistry* **64**, 435-461, doi:10.1146/annurev.bi.64.070195.002251 (1995).

24 Robart, A. R., Seo, W. & Zimmerly, S. Insertion of group II intron retroelements after intrinsic transcriptional terminators. *Proc Natl Acad Sci U S A* **104**, 6620-6625, doi:10.1073/pnas.0700561104 (2007).

25 Lambowitz, A. M. & Zimmerly, S. Mobile group II introns. *Annual review of genetics* **38**, 1-35, doi:10.1146/annurev.genet.38.072902.091600 (2004).

26 Mohr, G. *et al*. A Targetron System for Gene Targeting in Thermophiles and Its Application in <italic>Clostridium thermocellum</italic>. *PLoS ONE* **8**, e69032, doi:10.1371/journal.pone.0069032 (2013).

27 Guo, H. *et al*. Group II Introns Designed to Insert into Therapeutically Relevant DNA Target Sites in Human Cells. *Science* **289**, 452-457, doi:10.1126/science.289.5478.452 (2000).

28 Karberg, M. *et al*. Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. *Nature biotechnology* **19**, 1162-1167, doi:10.1038/nbt1201-1162 (2001).

29 Perutka, J., Wang, W., Goerlitz, D. & Lambowitz, A. M. Use of computer-designed group II introns to disrupt Escherichia coli DExH/D-box protein and DNA helicase genes. *Journal of molecular biology* **336**, 421-439 (2004).

30 Zhuang, F., Karberg, M., Perutka, J. & Lambowitz, A. M. EcI5, a group IIB intron with high retrohoming frequency: DNA target site recognition and use in gene targeting. *RNA* **15**, 432-449, doi:10.1261/rna.1378909 (2009).

31 Mohr, G., Ghanem, E. & Lambowitz, A. Mechanisms used for genomic proliferation by thermophilic group II introns. *PLoS Biol.* **8**, e1000391, doi:10.1371/journal.pbio.1000391 (2010).

32 Moretz, S. E. & Lampson, B. C. A group IIC-type intron interrupts the rRNA methylase gene of Geobacillus stearothermophilus strain 10. *J Bacteriol* **192**, 5245-5248, doi:10.1128/JB.00633-10 (2010).

33 Leclercq, S. & Cordaux, R. Selection-driven extinction dynamics for group II introns in Enterobacteriales. *PLoS One* **7**, e52268, doi:10.1371/journal.pone.0052268 (2012).

34 Copertino, D. W. & Hallick, R. B. Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns. *Trends in biochemical sciences* **18**, 467-471 (1993).

35 Dai, L. & Zimmerly, S. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Research* **30**, 1091-1102 (2002).

36 Mohr, S. *et al*. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* **19**, 958-970, doi:10.1261/rna.039743.113 (2013).

37 Xiong, Y. & Eickbush, T. H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**, 3353-3362 (1990).