**The Dissertation Committee for David Minh Truong Certifies that this is the approved version of the following dissertation:**

# Mobile group II intron: host factors, directed evolution, and gene targeting in human cells

**Committee:**

Alan M. Lambowitz, Supervisor

Jaquelin P. Dudley

Andrew D. Ellington

Rick Russell

Marvin Whiteley

# Mobile group II intron: host factors, directed evolution, and gene targeting in human cells

**by**

**David Minh Truong, B.S.**

**Dissertation**

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

**Doctor of Philosophy**

**The University of Texas at Austin**

**May 2014**

# Dedication

To my family, who overcame great hardships before coming to this country.

# Acknowledgements

Doctoral work is a very personal experience, because we budding scientists push at barriers of knowledge in which no experts may exist. Nevertheless, I have been helped, supported, and cajoled along the way by many brilliant and inspiring people.

First and foremost, Dr. Alan Lambowitz deserves my humblest gratitude. He has been an exceptionally patient, supportive, and insightful advisor. He has provided world-class resources and the freedom to explore, and this has transformed me into a more complete scientist. I have no doubt that my training in his laboratory will take me far in my career.

I want to offer my sincere thanks to my Ph.D. committee members. Drs. Jackie Dudley, Andy Ellington, Rick Russell, and Marvin Whiteley offered essential guidance, new perspectives, and enthusiasm that helped further my work. I have been honored to have you all through this journey.

I want to acknowledge all members of the Lambowitz laboratory, past and present. Some of you I know personally, others only by your work and ideas, but all of you have inspired me in different ways. I want to give special thanks to Georg Mohr and Jun Yao for always listening to my half-brained ideas. Thanks to Jamie Vernon for broadening my horizons and being my first Lambo-lab mentor. Fellow human targetroners, Joe Hanson and Curtis Hewitt, you understand what this dissertation took. Ben Gilman thanks for being a good buddy. Finally, thanks to all the rest of the ICMB community, all of you graduate students and postdocs made the time almost fly by.

# Mobile group II intron: host factors, directed evolution, and gene targeting in human cells

David Minh Truong, Ph.D.

The University of Texas at Austin, 2014

Supervisor: Alan M. Lambowitz

Mobile group II introns are retroelements that are found in prokaryotes, archaea, and the organelles of plants and fungi, but not in the nuclear genomes of eukaryotes. They consist of a catalytically active RNA and intron-encoded reverse transcriptase, which together promote site-specific integration into DNA sites in a mechanism called retrohoming. The group II intron Ll.LtrB has been developed into a programmable, DNA-targeting agent called "targetron", which is widely used in bacteria and an attractive technology for gene targeting in eukaryotes. However, group II intron genome targeting in human cells has not been equivocally shown. This dissertation focuses on the hypothesis that the low $Mg^{2+}$-concentrations found in higher eukaryotes present a natural barrier to group II introns. First, I studied *E. coli* host proteins that aid group II intron retrohoming and found that synthesis of a second DNA-strand relies on host replication restart proteins. Next, I demonstrated that mutations in the distal stem of the catalytic core domain V (DV) improve Ll.LtrB retrohoming in a low $Mg^{2+}$-concentration *E. coli* mutant and in biochemical assays. These results suggest that DV is involved in an RNA-folding step that becomes rate limiting at low $Mg^{2+}$. Subsequently, I performed directed evolution of the intron RNA by injecting *in vitro* prepared mutant intron libraries into

*Xenopus laevis* oocyte nuclei. The mutations were analyzed using Roche 454 sequencing to generate an intron fitness landscape, which revealed conserved positions and potentially beneficial mutations, enabling enhanced retrohoming in *Xenopus* oocytes. Finally, I used a hybrid Pol II/T7 Ll.LtrB eukaryotic expression system to show that high exogenous $MgCl_2$ in the growth media enables retrohoming into plasmids and genomic DNA in human cells. *In vivo* directed evolution and mutation analyses using PacBio RS circular consensus sequencing indicated that only a few mutations may improve intron activity in human cells. This dissertation provides evidence that efficient group II intron retrohoming in human cells is limited by low $Mg^{2+}$-concentrations and develops new approaches for overcoming this limitation to enable use of group II introns for gene targeting in higher organisms.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1: Introduction

Mobile group II introns are retroelements that have been developed into site-specific DNA-targeting vectors called "targetrons" (Lambowitz and Zimmerly, 2004; Enyeart *et al*., 2014). They consist of a large catalytic RNA that acts in concert with a multi-functional Intron-Encoded Protein (IEP) in a ribonucleoprotein (RNP) complex. Mobile group II intron RNPs promote intron integration into specific DNA target sites at efficiencies up to 100% in a process called "retrohoming" and to ectopic sites that resemble the normal homing site at low frequencies ($10^{-4}$-$10^{-6}$) in a process called "ectopic retrohoming" or "retrotransposition" (Ichiyanagi *et al*., 2002; Lambowitz and Zimmerly, 2011). Generally conferring no added benefit to the host, group II introns are considered genetic parasites or junk DNA. Nevertheless, their ability to relocate to new DNA sites has enabled them to propagate widely into new DNA loci and hosts. The mechanisms used by group II introns for integrating into new sites have been uncovered for a number of examples, including the extensively studied Ll.LtrB intron, which was discovered in the *ltrB* relaxase gene in *Lactococcus lactis* (Mills *et al*., 1996, 1997b; Lambowitz and Zimmerly, 2004). The Ll.LtrB group II intron has been studied at multiple levels including RNP structure/function, life-cycle, mechanism, and interaction with host proteins.

Our detailed understanding of the group II intron retrohoming mechanism made it possible to adapt them for the targetron technology, which is now widely used for genetic engineering of prokaryotes (Enyeart *et al*., 2014). Group II introns are readily programmed for insertional mutagenesis at desired DNA sites and have been used in a

wide range of bacteria for gene disruption and site-specific DNA insertion. Moreover, potential use of group II intron-derived targetrons in eukaryotes could have an even wider-impact. However, although group II introns are thought to be evolutionary ancestors of spliceosomal introns in higher organisms (Martin and Koonin, 2006), group II introns as such are not generally found in eukaryotic nuclear genomes. The lack of functional group II introns in eukaryotic genomes raises the possibility that host defense mechanisms and/or inherent properties of the introns themselves limit their ability to propagate in eukaryotic nuclei, and thus their application in genetic engineering of higher organisms. My work addresses these limitations and culminates with insight into why group II introns function inefficiently in eukaryotes and approaches that might be used to further develop group II introns for gene targeting applications in human cells.

### 1.1 Group II intron distribution and evolution

Group II introns were first discovered in the 1980's from sequencing of organellar genomes (Michel and Ferat, 1995), and shown to have catalytic activity in 1986 by three laboratories (Peebles *et al*., 1986b; Schmelzer and Schweyen, 1986; van der Veen *et al*., 1986). They are primarily found in prokaryotes, the organellar genomes of fungi and plants, and to some extent in archaea, but are not found in the nuclear genomes of eukaryotes (Lambowitz and Zimmerly, 2011). Sequence database searches have identified ~600 different mitochondrial (mt) group II introns (Lang *et al*., 2007), ~400 full-length bacterial group II introns, and 16 archaeal group II introns (Candales *et al*., 2012) (http://webapps2.ucalgary.ca/~groupii/index.html).

The broad distribution of group II introns in prokaryotes and the mitochondria (mt) and chloroplasts (cp) of some eukaryotes, contrasts with the paucity of group II

introns in metazoans. Only two group II introns have been discovered in animals: one within the mitochondria of the simple placozoan *Trichoplax adhaerens* and the other within the mitochondria of a carnivorous worm *Nephtys* sp. (Dellaporta *et al*., 2006; Valles *et al*., 2008). However, both of these introns are truncated, presumably non-functional, and likely acquired by horizontal transfer from a co-habiting bacterium. The apparent paucity of group II introns within metazoans suggests that they may have been eliminated through an as yet unknown mechanism. However, other nucleic acid and protein elements in metazoans suggest a group II intron-like ancestry, notably LINE-1 retrotransposons, nuclear introns, and the spliceosomal machinery, which together make up almost half the human genome (Lander *et al*., 2001).

The spliceosome excises pre-mRNA introns by a mechanism that is chemically identical to that used by group II introns (see "life-cycle" section) (Keating *et al*., 2010). It is now commonly theorized that early group II introns contributed to eukaryotic genome evolution as ancestors to spliceosome machinery and nuclear introns (Cech, 1986; Cavalier-Smith, 1991; Martin and Koonin, 2006; Keating *et al*., 2010; Rogozin *et al*., 2012). Recent evidence supports this theoretical framework. Group II introns share similar core structural motifs with snRNA's found in the spliceosome RNP (Keating *et al*., 2010; Lambowitz and Zimmerly, 2011). The recent crystal structure of the spliceosomal protein Prp8 shows a structure similar to group II intron IEPs including regions that resemble the reverse transcriptase and thumb domains (Galej *et al*., 2013). Finally, new evidence confirms that the U6 snRNA catalyzes both steps of nuclear pre-mRNA splicing as was initially suggested by the similarity to the group II intron splicing mechanism (Fica *et al*., 2013). Thus, while group II introns likely flourished during early

eukaryotic evolution, some intrinsic barriers now exist in higher eukaryotes that caused their degeneration from independent to host-dependent elements.

### 1.2 Group II intron RNA structure and the intron-encoded protein

Although they differ in primary sequence, all group II intron RNAs organize into a conserved secondary structure consisting of six conserved domains (denoted DI-DVI), which interact via tertiary contacts to fold the RNA into a catalytically active three-dimensional structure (Figure 1.1) (Lambowitz and Zimmerly, 2011). DI, the largest domain, provides a structural scaffold for the assembly of the other domains and contains exon-binding sites (EBS) that position the 5'- and 3'-splice sites and ligated-exon junction at the ribozyme active site for RNA splicing and reverse splicing reactions. DII contributes to assembly of the active site core, and DIII functions as a catalytic effector. DIV is the location of the ORF encoding the IEP, and DVI contains the branch-point nucleotide, typically an adenosine, used for lariat formation. DV is a small conserved domain that binds catalytic metal ions and interacts with DI and a single-stranded junction region called J2/3 to form the intron RNA's active site. It is thought to be the cognate of the U2/U6 snRNAs of the spliceosome (Michel *et al*., 2009; Keating *et al*., 2010).

Three major structural classes of group II intron RNAs, denoted IIA, IIB, and IIC, have been identified with differences in both peripheral and active-site elements, and can be further subdivided into IIA1, IIA2, IIB1, and IIB2 structures (Lambowitz and Zimmerly, 2011). The different subclasses can be distinguished by differences in group II intron domains, interdomain tertiary interactions, and interactions used to bind the 5’ and 3’ exon sequences at the active site. For example, the group II intron studied in this

dissertation, Ll.LtrB is a member of the IIA subgroup and uses sequence elements EBS1, EBS2 (exon binding sites 1 and 2), and δ in DI to base-pair with IBS1, IBS2 (intron-binding sites 1 and 2), and δ' located in the 5' and 3' exons. In contrast, group IIB and IIC introns can include EBS3/IBS3 interactions or recognize stem-loops derived from transcription terminators.

Group II introns can also be grouped according to the IEPs that they encode, of which there are eight lineages: bacterial classes A-F, ML (mitochondrial-like), and CL (chloroplast-like) (Zimmerly *et al*., 2001; Simon *et al*., 2009). Nearly all group II intron IEPs are located within DIV and are reverse-transcriptase (RT)-related proteins. The IEP encoded by the Ll.LtrB intron is referred to as LtrA, and it is one of the best-characterized examples. The LtrA protein contains four domains: reverse transcriptase (RT), thumb/maturase (X), DNA binding (D), and DNA endonuclease (En) (Figure 1.2) (Lambowitz and Zimmerly, 2011). The RT domain is characterized by seven conserved sequence blocks (RT1-7), which are found in the fingers and palm regions of retroviral RTs, and an N-terminal extension, RT0, which is conserved in non-LTR retrotransposon RTs (Blocker *et al*., 2005). The RT has additional insertions denoted 2a, 3a, 4a, and 7a, with 2a being conserved amongst non-LTR retrotransposon RTs (Malik *et al*., 1999; Blocker *et al*., 2005). The X domain corresponds to the thumb domain of retroviral RTs and is also called the "maturase" domain because it helps with RNA splicing of the intron (Cui *et al*., 2004). Finally, the D and En domains act during intron mobility by helping to recognize DNA target sites and cleaving the target DNA, respectively (San Filippo and Lambowitz, 2002).

### 1.3 Group II intron life-cycle: RNA splicing and retrohoming to new sites

Group II introns are transcribed by host RNA polymerases and splice out of precursor RNAs in a mechanism chemically identical to that of nuclear pre-mRNAs (Keating *et al*., 2010). The group II intron RNA itself catalyzes splicing into a lariat RNA via two sequential transesterification reactions (Figure 1.3) (Lambowitz and Zimmerly, 2011). During the first step of splicing, the 2'-OH of the bulged adenosine found in DVI acts as a nucleophile to attack the 5' splice site, producing a lariat/3'-exon intermediate. In the second step, the 3'-OH of the free 5'-exon acts as a nucleophile to attack the 3' splice site, which generates the free intron lariat and ligates the exons together. Some group II introns, such as Ll.LtrB, can self-splice under high, non-physiological salt concentrations (e.g., 50 mM $Mg^{2+}$) (Matsuura *et al*., 1997; Saldanha *et al*., 1999). Within cells, however, most group II introns require the IEP or host-proteins to splice out of RNA transcripts, as physiological $Mg^{2+}$ concentrations are below 4 mM (Lusk *et al*., 1968; Snavely *et al*., 1991). The IEP functions in place of $Mg^{2+}$ by stabilizing the catalytically active structure of the intron RNA (Noah and Lambowitz, 2003).

For mobile group II introns such as Ll.LtrB, the RNP comprised of the lariat RNA bound to its IEP constitutes an active mobile genetic element, ready to insert into a new DNA site via reverse-splicing. The initial RNP-catalyzed stages of retrohoming occur through a series of steps called target-primed reverse transcription (TPRT) (Figure 1.3) (Lambowitz and Zimmerly, 2011). During TPRT, the RNP binds DNA non-specifically and scans for the target site (Aizawa *et al*., 2003). Initial recognition occurs with specific contacts between LtrA and the 5' exon, and this leads to localized melting of the DNA, which permits the base pairing of EBS1, EBS2 and δ to the complementary sequences IBS1, IBS2, and δ’ of the DNA target site (Guo *et al*., 1997; Singh and Lambowitz,

2001). The lariat intron RNA then performs reversal of the two transesterification reactions, reverse-splicing, into the top strand of the DNA initiated by nucleophilic attack of the 3'-OH of the terminal intron nucleotide at the splice junction (Aizawa *et al*., 2003). A limited number of LtrA contacts with the 3' exon are made, enabling cleavage by the En domain to produce a staggered single nick on the bottom DNA strand (Singh and Lambowitz, 2001). LtrA then reverse transcribes the intron RNA by using the free 3'-OH of the bottom strand nick as a primer. Finally, host proteins are thought to generate the second top DNA strand for completion of retrohoming (Smith *et al*., 2005; Beauregard *et al*., 2006).

**1.4 Role of host proteins**

Because group II introns are freestanding genetic elements, much like viruses, and require a host to propagate, their life-cycle has evolved to become intimately linked with that of the host in which they reside. For the majority of mobile group II introns, the early steps of retrohoming and retrotransposition rely predominantly upon the IEP and the RNA's catalytic activity (Lambowitz and Zimmerly, 2004). However, some group II introns have evolved to depend upon host-encoded proteins to perform other steps including splicing, late-steps in retrohoming, and RNP localization (Lambowitz and Zimmerly, 2011). Understanding how cellular processes have evolved to affect group II intron function in bacteria and organelles may ultimately provide clues to requirements for activity of group II introns within the nucleus of higher eukaryotes.

Many mitochondrial and chloroplast group II introns lack an IEP and have co-opted host-encoded proteins to promote intron splicing (Lambowitz and Zimmerly, 2011). In maize, CRM family proteins (chloroplast RNA splicing and ribosome

maturation) including CRS1, CRS2, CAF1, and CAF2 act as splicing factors (Barkan *et al*., 2007). Other cp group II introns found in plants utilize PROR (plant RNA recognition) and PPR (pentatricopeptide repeat) proteins for splicing (Stern *et al*., 2010). The DEAD-box helicases Cyt-19 in *Neurospora crassa* and Mss116 in *Saccharomyces cerevisiae* can also promote splicing of mt group II introns *in vitro* and *in vivo* near physiological conditions (Mohr *et al*., 2006; Solem *et al*., 2006; Halls *et al*., 2007). Knockouts of Mss116 acquire a number of mitochondrial defects due to the loss of intron splicing from essential genes, highlighting how intimately group II introns have evolved towards using host proteins. These examples suggest important parallels for how group II introns may have evolved into the eukaryotic spliceosome machinery.

Reliance on host proteins is not limited to promoting splicing of the RNA from transcripts. In *E. coli* and *L. lactis*, host proteins affect Ll.LtrB RNP localization, resulting in the polar localization of the RNP particle (Zhao and Lambowitz, 2005; Beauregard *et al*., 2006; Zhao *et al*., 2008). The polar localization results in biased retrohoming near the replication origins, but stress responses that lead to the overproduction of polyphosphate result in delocalization of LtrA, leading to more uniform retrohoming across the genome (Zhao *et al*., 2008). In *Arabidopsis thaliana* protoplasts, transgenically expressed RmInt group IIC intron IEP, found in *Sinorhizobium meliloti*, localizes to the nucleolus (Nisa-Martinez *et al*., 2013). Previous work from our laboratory shows that Ll.LtrB group II intron RNPs require a nuclear localization signal fused to the IEP to allow nuclear entry, whereas otherwise it remains diffuse throughout the cytoplasm (Cui, 2006).

The late steps of retrohoming, principally the mechanism of top-strand DNA synthesis have remained unresolved for Ll.LtrB and likely involve host proteins. While the early steps catalyzed by group II intron RNPs are common to retrohoming pathways in all organisms, the late host-mediated steps of second-strand DNA synthesis and cDNA integration can occur in different ways. In *S. cerevisiae* mt, cDNA integration occurs largely via recombination in which the nascent intron cDNA at the recipient site invades an intron-containing allele at another DNA site for completion of intron DNA synthesis before switching back to the recipient DNA in the upstream exon (Eskes *et al*., 2000). In bacteria, however, the reverse-spliced intron RNA of Ll.LtrB produces a full-length intron cDNA that is integrated directly into the recipient DNA by non-recombination mechanisms involving host DNA enzymes (Mills *et al*., 1997a; Cousineau *et al*., 1998; Smith *et al*., 2005). Recently, non-lariat, linear forms of the *L. lactis* Ll.LtrB intron RNA were found to retrohome in *Xenopus laevis* and *Drosophila melanogaster* by using host non-homologous end-joining enzymes for cDNA integration (Zhuang *et al*., 2009b; White and Lambowitz, 2012).

In addition to its native host, the Ll.LtrB intron splices and retrohomes efficiently in a wide variety of other bacteria, including *E. coli*, where it has been studied using the facile genetic and biochemical methods available in that organism (Cousineau *et al*., 1998). In previous work, analysis of *E. coli* mutants identified a number of candidate host factors that potentially function in the late steps in retrohoming (Smith *et al*., 2005). The proteins include RNase H1, the 5'→3' exonuclease activity of Pol I, the host replicative polymerase Pol III, and DNA ligase A. Additional mutants that decreased retrohoming included the host exonuclease and endonuclease proteins RecJ, DnaQ (MutD), and SbcD.

On the other hand, increased retrohoming frequencies were found for mutants deficient in RNases I and E and exonuclease III, which likely affect intron RNA levels (Smith *et al*., 2005). More recently, Coros *et al*. (Coros *et al*., 2008, 2009) identified additional host factors potentially involved in retrohoming, including polynucleotide phosphorylase (PNPase), the DNA helicase Rep, and MnmE (TrmE), which functions in tRNA modification, with additional host proteins (CyaA, SpoT, and AtpA) acting by affecting accessibility of chromosomal target sites or energy metabolism. However, host factors that function in late steps in the retrohoming of group II intron lariat RNAs, the major retrohoming pathway used in nature, have not been identified conclusively in any organism. Consequently, the mechanisms used for these steps have remained poorly understood.

### 1.5 Gene targeting in prokaryotes

The Ll.LtrB group II intron functions efficiently in many types of bacteria and can be modified to insert into new DNA sites (Enyeart *et al*., 2014). As retrohoming relies predominantly upon base pairing between the short EBS and δ sequences within the intron and the DNA insertion site, modification of the EBS sequences allows for retargeting to new genomic sites (Guo *et al*., 1997, 2000; Mohr *et al*., 2000). Rules governing base-pairing preferences have been identified (Mohr *et al*., 2000) and codified into a weighted computer algorithm (Perutka *et al*., 2004) to identify best-possible target sites with a suitable site found in most stretches of DNA. This technology is now sold commercially as the TargeTron® kit by Sigma-Aldrich. To date, targetrons have been used successfully in at least twenty-three bacterial genera (Johnson and Fisher, 2013; Enyeart *et al*., 2014). The Ll.LtrB-based targetron has been adapted for use in

*Clostridium* spp. as the ClosTron (Heap *et al*., 2007). *Clostridium* species have generally been intractable to genetic engineering techniques, and the ClosTron has made major inroads towards improved reverse genetics in this organism.

In addition to high targeting frequency and specificity, targetron technology allows for different ways to perform genetic alterations. Direct integration of the intron into a target gene permanently disrupts gene expression, and genetic cargo of up to ~2 kb of DNA can be carried within DIV of the intron sequence, allowing for delivery of short markers such as promoters and antibiotic resistance cassettes (Karberg *et al*., 2001; Frazier *et al*., 2003; Perutka *et al*., 2004; Plante and Cousineau, 2006). The intron can also be used to induce DNA double-strand break mediated homologous recombination, and by co-transforming an additional DNA construct, place in large sections of DNA (Karberg *et al*., 2001). Recently, a new technology involving delivery of Cre/*lox* sites within DIVb permits the positioning of recombinase sites, allowing for genome rearrangements, deletions, inversions, translocations, and delivery of very large DNA cargo at high efficiency in a broad range of bacteria (Enyeart *et al*., 2013).

In addition to Ll.LtrB, several other group II introns have been engineered into targetrons including the IIB introns EcI5 and RmInt1, and recently the thermostable IIB intron TeI3 with the RT from TeI4 (Zhuang *et al*., 2009a; Garcia-Rodriguez *et al*., 2011; Mohr *et al*., 2013). The TeI3/TeI4 "thermotargetron", whose intron RNA and IEP components evolved to function in the thermophilic cyanobacterium *Thermosynechococcus elongatus*, can function at temperatures of 60°C or higher and inserts into targeted genes at frequencies up to 100% (Mohr *et al*., 2010, 2013). The thermotargetron has found recent use in *Clostridium thermocellum* to increase cellulolytic

ethanol production (Mohr *et al*., 2013) and should prove useful in other industrially important organisms.

### 1.6 Gene targeting technologies in higher eukaryotes and potential for targetrons

Programmable gene targeting in higher eukaryotes has the potential to improve human health, products in livestock animals, and genetics research in animals. Traditional gene targeting techniques, such as that used for the generation of transgenic mice, involves homologous recombination of exogenous DNA, although the frequency is $<10^{-6}$ events and requires antibiotic selection (Porteus and Baltimore, 2003). However, studies using I-SceI endonucleases have shown that introduction of a double-strand break can improve the recombination frequency of exogenous DNA constructs >1000-fold in metazoan cells (Segal and Carroll, 1995; Porteus and Baltimore, 2003). Recent technologies use the Zinc finger and TAL DNA-binding domains to deliver the Fok-I endonuclease to specific DNA sites for the introduction of recombinogenic double-strand breaks, thereby greatly increasing the efficiency of homologous recombination (Wu *et al*., 2007; Mussolino and Cathomen, 2012). New binding sites can be engineered or selected for the Zinc finger proteins and engineered using specific modules in the case of TALs (Carroll, 2011; Joung and Sander, 2013). However, these technologies are susceptible to off-targeting effects due to binding at degenerate sequences similar to the target, causing unaccounted for double-strand breaks and cell death (Handel and Cathomen, 2011; Fine *et al*., 2013). In addition, it is relatively difficult to engineer new sites for Zinc Fingers (Gaj *et al*., 2013). While TALs are generally modular, not all motifs work together consistently, and cloning the motifs can require multiple steps (Wei *et al*., 2013).

Another recent technology uses the bacterial CRISPR immune system to target double-stranded breaks via RNA-DNA base-pairing using a RNA-guide sequence bound to the Cas9 endonuclease (Jinek *et al*., 2012). This system works in bacteria, human cells, fruit flies, zebrafish, and monkeys at high efficiencies (Mali *et al*., 2013b; Niu *et al*., 2014), and with the same ease of retargeting as that of targetrons. However, studies have shown high off-targeting frequencies using this system (Fu *et al*., 2013), which may or may not be alleviated by strategies such as using two CRISPR/Cas9 single strand nickases to make a staggered cut (Mali *et al*., 2013a; Ran *et al*., 2013) or by using truncated guide RNAs, which presumably minimizes non-essential base-pairing (Fu *et al*., 2014).

Group II intron-derived targetrons allow for similarly facile re-programming as CRISPR/Cas9 systems, but with potentially higher target specificity. Because mismatches during RNA-DNA binding of the targetron affect the $k_{cat}$ of the reaction, off-target binding is unlikely to lead to targetron insertions via reverse splicing (Xiang *et al*., 1998). In contrast, CRISPR/Cas9 systems potentially lead to off-targeting because mismatches only affect the $K_m$ and binding can lead to cleavage as long as the nuclease is bound even transiently at a DNA site. Additionally, the minimally 18 nucleotide recognition sequence of targetrons means that integration sites are likely unique within a human genome, as there is only a 1 in $6 \times 10^{10}$ chance of repetition compared to $3 \times 10^9$ nucleotides in the genome.

The first studies of targetrons in human cells involved the electroporation of *in vitro* prepared RNPs into human cell culture, showing low efficiency integrations into plasmids containing the CCR5 gene (Guo *et al*., 2000). Subsequent studies using refined

techniques for generating RNPs *in vitro* could show integration in genomic DNA of the 45S rRNA gene in HEK-293 cells (Cui, 2006). However, the above studies required nested PCR and never reported integration frequencies. Recent studies have been unable to detect targeted integrations in K-562, HEK-293, or HeLa cells using a variety of reporter constructs, including GFP, lacZ, and antibiotic cassette based modules (Vernon, 2010; Hanson, 2013).

### 1.7 Magnesium dependence of mobile group II introns

A number of observations suggest efficient retrohoming in higher eukaryotes is limited by the low free magnesium concentrations [$Mg^{2+}$] found in higher eukaryotic cells, which are lower in eukaryotes than in bacteria (Lusk *et al*., 1968; Gunther, 2006). For instance, knockout of the yeast mitochondrial $Mg^{2+}$ transporter Mrs2p inhibits splicing of endogenous group II introns contained within mitochondrial genes (Gregan *et al*., 2001). Moreover, studies of Ll.LtrB retrohoming into plasmids within *Xenopus laevis* oocytes, *Drosophila melanogaster* embryos, and zebrafish (*Danio rerio)* embryos show a clear relationship between intron retrohoming frequency and the co-injection of surplus $Mg^{2+}$ of up to 500 mM, leading to free [$Mg^{2+}$] of ~10 mM (Mastroianni *et al*., 2008). Consistent with these observations, *in vitro* experiments using reconstituted group II intron RNP complexes show that efficient protein-assisted splicing and retrohoming require concentrations of free [$Mg^{2+}$] of 5-10 mM (Matsuura *et al*., 1997; Saldanha *et al*., 1999). EDTA-chelation measurements of soluble free [$Mg^{2+}$] in *E. coli*, an organism in which group II introns function efficiently, suggest free concentrations of ~4 mM [$Mg^{2+}$] (Lusk *et al*., 1968), in line with the *in vitro* requirements for group II intron retrohoming. Mammalian cells contain only 0.2-1 mM free [$Mg^{2+}$] during the majority of the cell cycle,

as determined by NMR and fluorescent-probe based methods (Gunther, 2006; Rubin, 2007). While the total [$Mg^{2+}$] in mammalian cells is around 14 mM, most of these molecules are compartmentalized within organelles and/or chelated to molecules such as nucleic acids, ATP, and fatty acids of the plasma membrane (Gupta *et al*., 1984; Grinstein and Dixon, 1989).

The relationship between $Mg^{2+}$ ions and RNA's such as the group II intron may have developed early in their evolution. $Mg^{2+}$ acts as a cofactor for RNA's by neutralizing electronegative zones generated at and between phosphate backbones during folding of RNAs into more complex tertiary structure, and it can act to accelerate catalytic reactions (DeRose, 2003). Two types of $Mg^{2+}$ binding on nucleic acids may occur: site-bound $Mg^{2+}$ such as that found within RNA folds or ensemble $Mg^{2+}$ ions bound transiently around the RNA surface (Misra *et al*., 2003; Draper, 2004). Interestingly, a group II intron discovered in *Pylaiella litorallis*, self-splices *in vitro* at concentrations as low as 0.5 mM [$Mg^{2+}$] unassisted by an IEP (Costa *et al*., 1997). By comparison, the Ll.LtrB intron requires 50 mM [$Mg^{2+}$] for efficient self-splicing *in vitro*. However, the *P. litorallis* intron has not been shown to be a mobile element (Karberg and Lambowitz, 2006), and the intron and IEP when expressed in mammalian cells does not lead to RNA splicing (Zerbato *et al*., 2013). Since all group II introns have conserved secondary and tertiary structures, the findings for the *P. litorallis* intron suggest that it may be possible to reduce the $Mg^{2+}$ requirements for the Ll.LtrB intron by a process of directed evolution or by rational design.

## 1.8 Overview of dissertation research

This dissertation focuses on improving the ability of mobile group II introns to function in higher eukaryotic cells. Chapter 2 begins with the study of *E. coli* host proteins that help during retrohoming of the Ll.LtrB intron. The rationale for this study was to determine if additional proteins may underlie the high efficiency of targeting in bacteria and might subsequently aid in targeting in higher eukaryotes. Although this aim was not achieved, I did uncover the mechanism through which top strand DNA synthesis occurs during retrohoming. Chapter 3 focuses on improving the activity of Ll.LtrB by mutating catalytic center DV and selecting for improved variants within an *E. coli* mutant with decreased intracellular $Mg^{2+}$ concentrations. In addition to uncovering numerous improved variants, I found that DV might facilitate an RNA-folding step that becomes rate limiting at low $Mg^{2+}$ concentrations. In Chapter 4, I performed directed evolution studies on the whole intron RNA by injecting *in vitro* prepared RNPs directly into *Xenopus laevis* oocyte nuclei. I analyzed the variants using Roche 454 next generation sequencing to generate a mutation fitness landscape and identified mutations that potentially improve activity as well as found conserved regions that are essential for group II intron function. Finally in Chapter 5, I show that a hybrid Pol II/T7 Ll.LtrB eukaryotic expression system in combination with high exogenous $MgCl_2$ leads to site-specific retrohoming directly into plasmids and genomic DNA in human cells. Directed evolution studies within the cells were subsequently analyzed by long read Pacific Biosciences RS circular consensus sequencing indicating regions that may improve group II intron function in human cell culture.

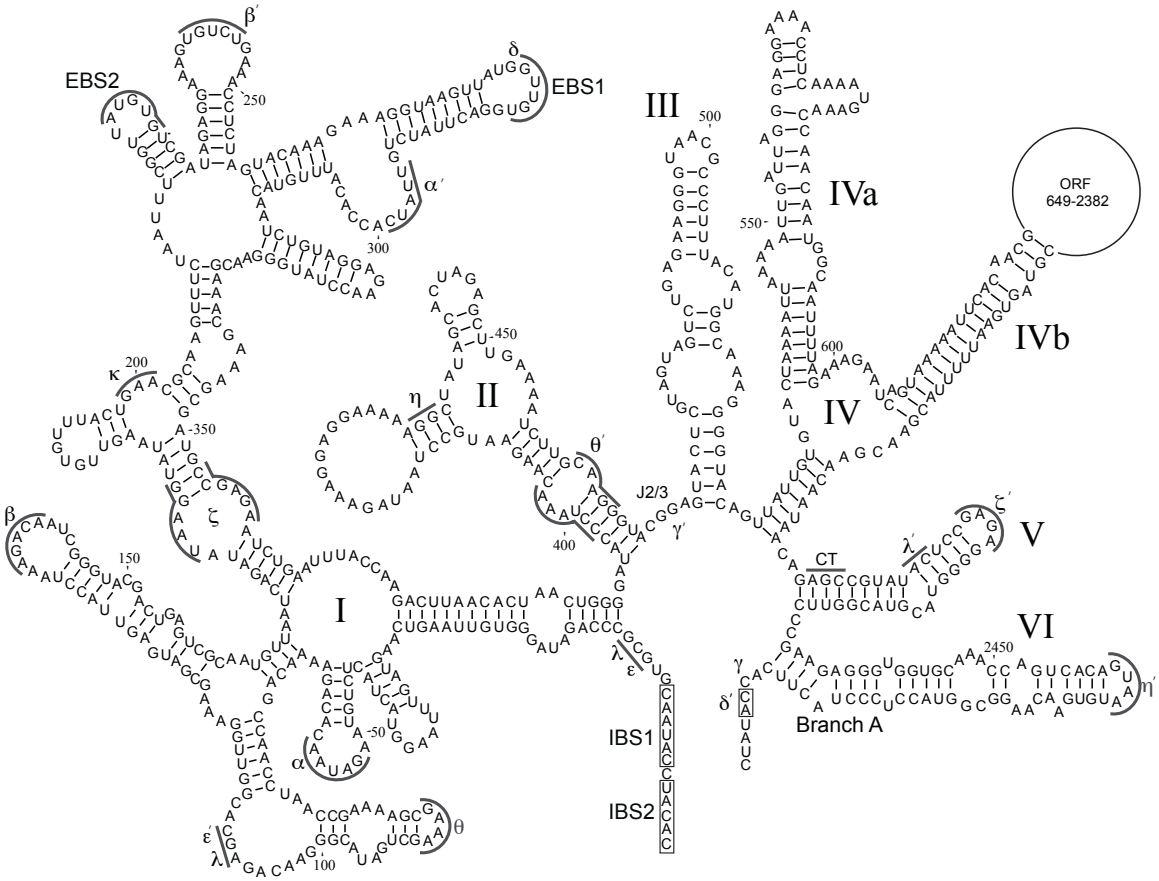



Figure 1.1: Secondary structure of the Ll.LtrB group II intron RNA

The Ll.LtrB group II intron RNA organizes into six conserved helical domains, denoted DI-DVI, emanating from a central hub. Tertiary contacts that mediate RNA folding are indicated by Greek letters. The largest domain, DI, contains sequence elements denoted EBS1, EBS2, and δ which base pair with 5'- and 3'-exon sequences denoted IBS1, IBS2, and δ' for RNA splicing. These same elements are used for DNA target site binding during retrohoming. DI, DV, and J2/3 comprise the catalytic core of the intron, while DII and DIII enhance the rate of folding and catalysis. DIV contains the ORF for the IEP, LtrA, and also contains a binding site for the LtrA protein in subdomain DIVa. The bulged adenosine used for splicing into lariat RNA is found near the end of DVI. The CT in DV refers to the conserved catalytic triad.


Figure 1.2: LtrA protein domain organization

The IEP encoded by the Ll.LtrB group II intron, denoted LtrA, is organized into four domains: reverse transcriptase (RT), thumb/maturase (X), DNA binding (D), and DNA endonuclease (En). The numbers refer to amino acid positions found at the domain boundaries. The RT domain contains seven conserved motifs (RT1-7), and an additional N-terminal extension (RT0) found in non-LTR retroelement RTs (Lambowitz and Zimmerly, 2004). Insertions between the conserved motifs are denoted 2a, 3a, 4a, and 7a, and 2a is conserved in non-LTR-retroelement RTs (Malik *et al*., 1999; Blocker *et al*., 2005). The RT and X domains together bind the intron RNA and stabilize the catalytically active intron RNA structure, and synthesize cDNA from RNA templates, such as that generated from reverse splicing of the intron lariat into DNA (see Figure 1.3). The D and En domains interact with the DNA target site during intron retrohoming (see Figures 1.3 and 1.4). Image adapted from (Lambowitz and Zimmerly, 2011).

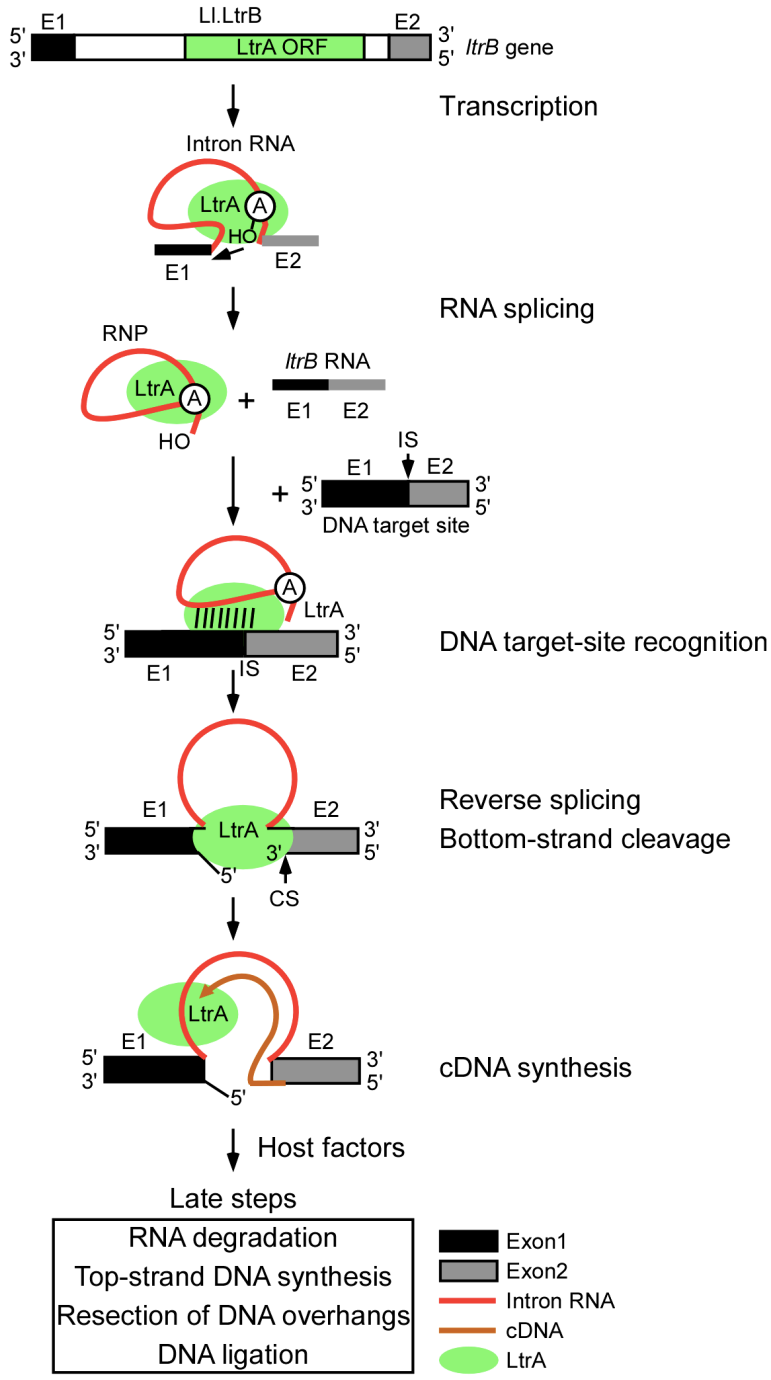



Figure 1.3: Retrohoming pathway of Ll.LtrB intron lariat RNA in bacteria.

The Ll.LtrB intron encodes a multi-functional RT (LtrA protein) with RT, RNA splicing, DNA-binding, and DNA endonuclease activities (Lambowitz and Zimmerly, 2011). Transcription of the *ltrB* gene yields a precursor RNA containing the intron flanked by 5' and 3' exons (E1 and E2, respectively). LtrA is translated from within the intron using its own Shine-Dalgarno sequence and then binds to the intron in the precursor RNA to stabilize the catalytically active RNA structure for RNA splicing. RNA splicing occurs via two sequential RNA-catalyzed transesterification reactions that are initiated by nucleophilic attack of the 2' OH of a branch point A-residue near the 3' end of the intron at the 5'-splice site and result in ligated *ltrB* exons and an excised intron lariat RNA. After splicing, LtrA remains tightly bound to the excised intron lariat RNA in an RNP. RNPs initiate retrohoming by recognizing DNA target sequences (corresponding to the ligated *ltrB* E1-E2 sequence), using both the IEP and base pairing of the intron RNA. The intron RNA then inserts via reversal of the two transesterification reactions used for RNA splicing (referred to as "reverse splicing") into the intron-insertion site (IS) at the ligated-exon junction in the top strand of the DNA target site. LtrA uses its DNA endonuclease activity to cleave the bottom strand between positions +9 and +10 (CS) of E2 and uses the 3' end at the cleavage site as a primer for reverse transcription of the inserted intron RNA. The resulting intron cDNA is then integrated into the genome by host enzymes in late steps that minimally include degradation of the intron RNA template strand, second (top)-strand DNA synthesis, resection of DNA overhangs, and sealing of DNA strand nicks. This figure was adapted from (Yao *et al*., 2013).

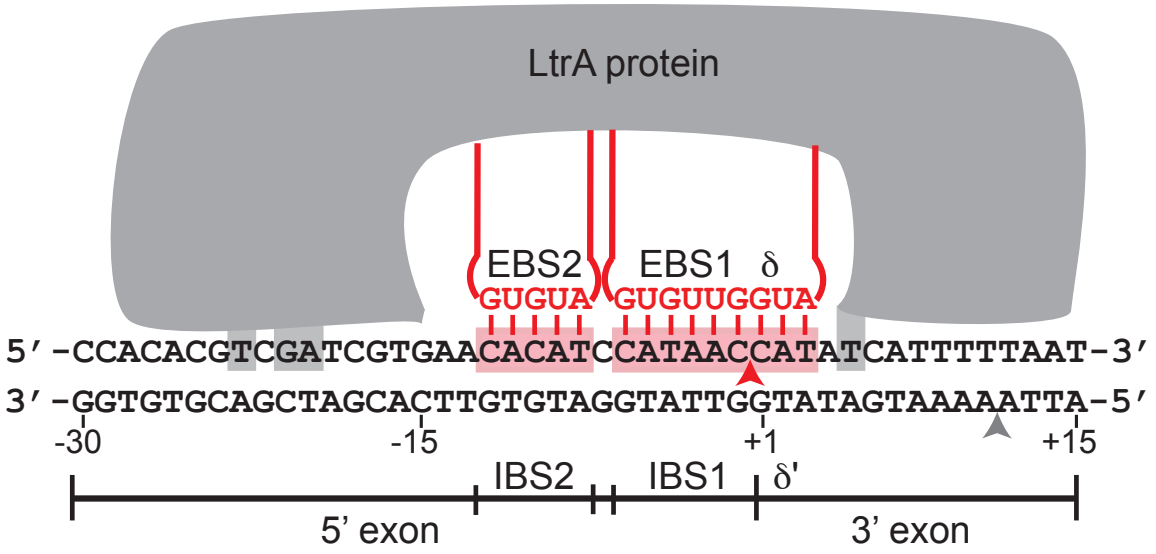



Figure 1.4: DNA target site recognition by the Ll.LtrB RNP.

DNA targeting using Ll.LtrB group II intron targetrons relies on recognition of a target site that spans from -23 to +10 of the RNA insertion site. The figure shows the LtrA protein in gray and the intron RNA sequences EBS1, EBS2, and δ found within DI in red. Initial recognition occurs with specific contacts between LtrA and the 5' exon at positions -23, -21, and -20 relative to the intron-insertion site (highlighted in gray). Localized melting of the DNA permits base pairing of the RNA sequences EBS1, EBS2 and δ to the IBS1, IBS2, and δ’ sequences in the DNA target site (highlighted in red), leading to reverse splicing of the lariat intron RNA into the top strand of the DNA at the red arrowhead. At the same time, LtrA contacts the 3' exon, most critically the T at the +5 position (highlighted in gray), enabling the En domain to make a staggered nick on the bottom DNA between position +9 and +10 at the gray arrowhead. The nick is subsequently used as a priming site for cDNA synthesis by the RT and thumb domains of LtrA.

# Chapter 2: Taqman qPCR genetic assays reveal a key role for replication restart in group II intron retrohoming[1]

## 2.1 Introduction

During retrohoming of the group II intron Ll.LtrB, the fully reverse spliced intron RNA is reverse transcribed to yield a full-length intron cDNA that is integrated directly into the recipient DNA by a RecA-independent mechanism hypothesized to involve host DNA repair enzymes (Mills *et al*., 1997a; Cousineau *et al*., 1998; Smith *et al*., 2005). However, host factors that function in late steps in the retrohoming of group II intron lariat RNAs, the major retrohoming pathway used in nature, have not been identified conclusively in any organism and, consequently, the mechanisms used for these steps have remained poorly understood. In addition to its native host, the Ll.LtrB intron splices and retrohomes efficiently in a wide variety of other bacteria, including *Escherichia coli*, where it has been studied by using the facile genetic and biochemical methods available for that organism (Cousineau *et al*., 1998).

By screening *E. coli* mutants using two different plasmid-based retrohoming assays, our lab in collaboration with the Belfort laboratory previously identified candidate host factors that potentially function in the late steps in retrohoming, including RNase H1 and the 5'→3' exonuclease activity of Pol I, both of which could contribute to degrading the intron RNA template strand; the host replicative polymerase Pol III, which may function in second-strand DNA synthesis; and DNA ligase A, which presumably seals

[1] The following work is adapted from Yao J, Truong DM, and Lambowitz AM, (2013) PLOS Genetics, 9(4):e1003469, and should be considered together with results found therein. All work presented was written and performed by the dissertation author, with writing contributions from Yao J and Lambowitz AM.

strand nicks (Smith *et al*., 2005). Decreased retrohoming frequencies were also found in mutants deficient in host exo- and endonucleases activities [RecJ, DnaQ (MutD), and SbcD], which could function to resect overhangs or resolve intermediates, and increased retrohoming frequencies were found for mutants deficient in RNases I and E and exonuclease III, which in wild-type strains may suppress retrohoming by degrading the intron RNA or nascent cDNA (Smith *et al*., 2005).

In follow-up to these studies, Dr. Jun Yao, a previous student and current post-doctoral researcher in our laboratory, identified a number of genes that may function during retrohoming. He used a plasmid-based group II intron retrohoming assay that controls for indirect effects and screened an *E. coli mariner* transposon-insertion library for mutants that have decreased or increased retrohoming efficiency (Yao, 2008; Yao *et al*., 2013). Dr. Yao identified a set of 67 candidate protein-encoding genes, whose disruption or altered expression due to the proximity of the transposon insertion results in decreased retrohoming efficiency. Six of these candidate genes were sites of multiple transposon insertions, and 12 were genes with nucleic acid-related functions found downstream of transposon-insertion sites within operons. However, transposon insertions can lead to gene polarity effects and genetic screens may not indicate direct interactions. Such indirect effects could result from mutations that impair the propagation or expression of the intron-donor plasmid, decrease the intracellular levels or activity of group II intron RNPs, or impede the accessibility of group II intron RNPs to DNA target sites.

To complement the transposon-insertion screen performed by Jun Yao, I screened individual candidate strains for efficient integration into a chromosomal target site

directly by using Taqman qPCR assays to quantify both the 5'- and 3'-intron-integration junctions, thereby eliminating false positives that arise from mutations affecting expression of a drug-resistance phenotype (Figure 2.1). These assays confirmed that proteins RNase H1, Pol I, and Pol III have direct roles during group II intron retrohoming. In addition, I identified a number of proteins found in the replication restart complex as potential candidates involved during top strand DNA synthesis during retrohoming. All of these proteins were subsequently validated by Jun Yao with a new biochemical assay that uses *E. coli* cellular extracts to recapitulate top strand DNA synthesis during *in vitro* retrohoming of RNPs into a short DNA target.

## 2.2 Results

### 2.2.1 Taqman qPCR assays in Keio deletion strains

Because mobile group II introns recognize DNA target sequences primarily by base pairing of sequence elements within the intron RNA, they can be retargeted to retrohome into different chromosomal DNA sites simply by modifying the intron RNA to base pair to the new target, a gene targeting technology known as "targetron" (Guo *et al*., 2000; Karberg *et al*., 2001; Perutka *et al*., 2004). The Taqman qPCR assay uses an Ll.LtrB-ΔORF intron that was retargeted in this way to retrohome efficiently into a site in the *rhlE* gene, which encodes a non-essential DEAD-box protein whose disruption has no effect on cellular growth rate (Figure 2.1) (Ohmori, 1994; Perutka *et al*., 2004). The intron was expressed from the broad-host range donor plasmid pBL1 and employs an *m*-toluic acid-inducible promoter; the latter functions independently of host factors and is activated by a freely permeable inducer (*m*-toluic acid) that does not require cellular

transporters to enter the cell (Yao and Lambowitz, 2007). Additionally, the screen was carried out in mutant strains from the Keio collection in which deleted genes are replaced with a *kan*$^{R}$ marker, thereby mitigating polarity effects on downstream genes in operons (Baba *et al*., 2006). The Keio strains were supplemented by temperature-sensitive mutants to test the contribution of essential genes.

I used the Taqman qPCR assay to test all 68 candidate host factors identified in the initial transposon-insertion screen, as well as 30 additional candidate proteins that act on nucleic acids, including all 21 such candidates identified in previous mutant screens (Smith *et al*., 2005; Beauregard *et al*., 2006; Coros *et al*., 2008). Table 2.1 shows results of the Taqman qPCR assay for notable mutants, and Tables 2.2 and Table 2.3 shows complete results for the Taqman qPCR assay tested at 37°C and 30°C, respectively. Among the 68 candidates identified in the initial transposon library screen, only ten (*dnaC*, *dnaT*, *gyrB*, *mdoB*, *paoD*, *rpoH*, *rpoN*, *tonB*, *ydcM*, and *yjjB*) had statistically significant decreases in retrohoming efficiency in the Taqman qPCR assay, and only four (*dnaC*, *dnaT*, *gyrB*, and *rpoH*) had substantial decreases (10-67% of wild type retrohoming efficiency; Table 2.1 and Table 2.2). This poor correlation between the initial genetic screen and Taqman qPCR assay highlights the difficulty of distinguishing direct and indirect effects and the necessity of using multiple approaches to identify host factors that function in retrohoming.

Among the candidates identified as potential retrohoming factors in previous screens (Smith *et al*., 2005; Beauregard *et al*., 2006; Coros *et al*., 2008), the Taqman qPCR assay confirmed significant reductions in retrohoming efficiency in the Keio deletions of *rnhA* (RNase H1, the major cellular RNase H (Kanaya and Crouch, 1983));

*seqA* (initiation of chromosomal DNA replication ); *sbcC* (ATP-dependent exonuclease); *hns* (histone-like nucleoid structuring protein ); and *tus* (DNA replication termination site-binding protein); as well as at restrictive temperatures in the temperature-sensitive mutants *polAex*$^{ts}$, which is defective in the 5'→3' exonuclease but not the DNA polymerase activity of Pol I (Uyemura *et al*., 1976); and *dnaE*$^{ts}$ in the catalytic (α) subunit of the host replicative DNA polymerase Pol III (Maki *et al*., 1985). Also in agreement with previous results, I found no strong decrease in retrohoming efficiency in a Keio deletion of *rnhB* (RNase H2 (Rydberg and Game, 2002)).

In contrast to results of the transposon-insertion genetic assays, the Taqman qPCR assays found no decrease in retrohoming efficiency for Keio deletions of *recJ* (single-stranded DNA exonuclease (Lovett and Clark, 1984)); *dnaQ* (Pol III ε subunit, which has the proofreading exonuclease activity (DiFrancesco *et al*., 1984)); *rep* and *recQ* (DNA helicases (Takahashi *et al*., 1979; Umezu *et al*., 1990)); *recF* (RecA-dependent recombination); *stpA* (H-NS-like DNA- and RNA-binding protein with RNA chaperone activity (Cusick and Belfort, 1998)); *mnmE* (*trmE*; tRNA modification); *ligA* and *ligB* (DNA ligases (Olivera and Lehman, 1967; Sriskanda and Shuman, 2001)); and *pnp* (polynucleotide phosphorylase (Kinscherf and Apirion, 1975)). Additionally, the Taqman qPCR assay found no decrease in retrohoming efficiency for Keio deletions of the genes encoding DNA repair polymerases *polB* (Pol II (Kornberg and Gefter, 1971)), *dinB* (Pol IV (Wagner *et al*., 1999)), and *umuC* or *D* (Pol V (Shinagawa *et al*., 1983)), whereas *polB* and *dinB* deletions in a different strain showed moderate decreases in previous genetic assays (Smith *et al*., 2005).

The new candidates that were identified in the transposon-insertion screen and confirmed to have substantial decreases in retrohoming efficiency in the Taqman qPCR assay (10-67% wild type; Table 2.1) were: DnaC and DnaT, which function in replication restart (identified as genes downstream of the transposon insertion in the *yjjB* operon (Masai and Arai, 1988; Heller and Marians, 2005a)); GyrB (DNA gyrase subunit B); and RpoH (RNA polymerase $\sigma^{32}$ factor). For several mutants in which inhibition of retrohoming was found in genetic assays but not in the Taqman qPCR assay, Dr. Yao subsequently found significant effects on top- or bottom-strand DNA synthesis in biochemical assays (*e.g.*, *dinB*, *dnaQ*, *ligA*, *recJ*, *pnp*, *polB,* and *stpA*). The disagreement between the genetic transposon screen and Taqman qPCR assays for these mutants may reflect: (i) that qPCR monitors only short DNA regions at the intron-integration junctions; (ii) the longer time of the Taqman qPCR assay, which may give alternative enzymes a greater chance to act; or (iii) the different genetic backgrounds of the strains used in the two assays. The results emphasize the need to use multiple assays to identify retrohoming factors, with biochemical support for a genetic assay in our view providing the most definitive identification.

### 2.2.2 Taqman qPCR assays of replication restart proteins

The decreased retrohoming efficiency resulting from a transposon insertion in the *yjjB* operon containing *dnaC* and *dnaT* in the transposon genetic screen (see (Yao, 2008; Yao *et al*., 2013)) focused our attention on replication restart proteins as attractive candidates for playing a role in the late steps of retrohoming. I therefore performed systematic Taqman qPCR assays of replication restart mutants and found significant reductions in retrohoming in Keio deletions of *priA*, *priC*, and *dnaT*, and in a

temperature-sensitive mutant of *dnaB* (Table 2.1). PriA and PriC are key proteins that independently recognize stalled or collapsed replication forks in the three major *E. coli* replication restart pathways (denoted the PriA-PriB, PriA-PriC, and PriC-Rep pathways; (Liu *et al*., 1999; Heller and Marians, 2005a)), whereas DnaT interacts with PriA and PriC to load the replicative DNA helicase DnaB (Ueda *et al*., 1978; Allen and Kornberg, 1993). I also found decreased retrohoming efficiencies in temperature-sensitive mutants of several essential genes that function in replication restart, including those encoding DnaC, which interacts with DnaB prior to loading (Wahle *et al*., 1989); DnaG, the DNA primase (Rowen and Kornberg, 1978); and the single-stranded DNA binding protein Ssb, which has been shown to promote the formation of the primosome at the chromosomal replication origin (*oriC*) and interacts with PriA to stimulate the loading of DnaB during replication restart (Meyer and Laine, 1990; Cadman and McGlynn, 2004). However, deletion of the genes encoding PriB, an auxiliary component of the PriA-dependent pathway, and Rep, which functions in conjunction with PriC (Sandler, 2000), showed only small (1-21%) reductions in retrohoming efficiency.

### 2.2.3 Taqman qPCR assays in SOS-deficient strains

I also performed Taqman qPCR assays of retrohoming in a different set of replication restart mutants in the genetic background of *E. coli* SS996, a *recA*$^+$ strain containing a GFP reporter for SOS induction. The results indicated that the decreased retrohoming efficiencies in the affected replication restart mutants are not allele specific and are larger than can be accounted for by the proportion of cells undergoing the SOS response (Figure 2.2; different alleles tested for all genes except *dnaE* and *dnaB*). Additionally, the mutant strain SS4610 (*lexA51::Tn5*), which has a *lexA* null allele and is

constitutively induced for the SOS response in nearly 100% of cells (McCool *et al*., 2004), showed only minimally decreased retrohoming frequencies in Taqman qPCR assays (81-87% wild type; Figure 2.2 and Table 2.1). Collectively, the above findings indicate that the decreased retrohoming efficiency in the replication restart mutants is not a secondary effect of cell cycle arrest during SOS induction and indicate a requirement for replication restart proteins in group II intron retrohoming.

## 2.3 Discussion

In addition to the initial transposon-genetic screens, Dr. Jun Yao developed a biochemical assay in which host factors function together with group II intron RNPs to reconstitute the complete retrohoming reaction *in vitro*. The biochemical assays were generally consistent with the results from my Taqman qPCR assays, confirming many of the same proteins. The combined results confirm that the Ll.LtrB group II intron relies on host DNA polymerases for second-strand DNA synthesis, with a major role for the host replicative polymerase Pol III. A major function for RNase H1 in retrohoming is indicated and, by contrast, mutations in the *rnhB* gene encoding RNase H2 do not significantly inhibit retrohoming (this work and (Smith *et al*., 2005)). A major finding is that replication restart proteins function in retrohoming and are required for second-strand DNA synthesis.

The replication restart components found here to function in retrohoming include PriA and PriC, the host proteins that initiate replication restart by recognizing stalled or collapsed replication forks (Heller and Marians, 2005a); the accessory proteins DnaC and DnaT (Sandler *et al*., 1999); and the replicative helicase DnaB (Allen and Kornberg,

1993). During replication restart, PriA recognizes a branched intermediate in which the 3' OH of the nascent leading strand is close to the replication fork (no gap or a gap of < 3 nts), whereas PriC recognizes an intermediate with a larger gap (> 7 nts) (Heller and Marians, 2005a). These proteins may recognize the branched intermediate formed after RNase H degradation of the intron RNA template strand and extension of intron cDNA synthesis into the 5' exon. They then initiate replisome loading and second-strand DNA synthesis by Pol III by mechanisms similar or identical to those ordinarily used for replication restart at stalled or gapped replication forks. During retrohoming, longer or shorter gaps in the branched intermediate could result from more or less resection of a stalled nascent bottom strand after dissociation of the RT prior to reinitiation of DNA synthesis by a host DNA polymerase. Although the top strand of the retrohoming intermediate contains annealed RNA fragments that result from RNase H digestion and may thus resemble a nascent lagging strand, the location of this strand relative to the branch differs from that at a replication fork, and it is unclear how or if it might also contribute to recognition by PriA or PriC.

In contrast to other replication restart components, I found no contribution to group II intron retrohoming for PriB, an accessory protein in the PriA pathway, or Rep, which ordinarily functions together with PriC on the stalled fork by unwinding the dsDNA, especially when the 5' end of the newly synthesized lagging strand is close to the fork (Sandler, 2000; Heller and Marians, 2005b). The dispensability of these factors is consistent with findings showing that PriC can load DnaB on stalled replication forks independently of either PriA or Rep (Heller and Marians, 2005a).

The combined genetic and biochemical assays also indicate a major role in retrohoming for two other essential proteins that function in conjunction with replication restart machinery, the single-stranded DNA binding protein Ssb and the primase DnaG. Ssb binds ssDNA regions after unwinding by Rep or PriA and has been shown to physically interact with PriA to stimulate the loading of DnaB at stalled forks (Cadman and McGlynn, 2004; Heller and Marians, 2005a). DnaG synthesizes short RNA primers, which are used for initiation of DNA synthesis by Pol III, and triggers the release of DnaC from DnaB (Makowska-Grzyska and Kaguni, 2010).

I also identified a number of additional host proteins in which mutations likely inhibit retrohoming indirectly. A number of these proteins act on chromosomal DNA or in transcription (*e.g.*, GyrB, Hns, RpoH, SbcC, Tus) and could impact group II intron retrohoming *in vivo* by affecting chromosome structure, replication status, or target site accessibility.

Considered together, these results suggest a model for retrohoming of Ll.LtrB intron lariat RNAs in *E. coli* shown in Figure 2.3. In initial previously characterized steps, Ll.LtrB RNPs recognize the double-stranded DNA target site and the intron RNA reverse splices into one DNA strand, while the IEP cuts the opposite DNA strand and uses the cleaved strand as a primer for reverse transcription of the reverse-spliced intron RNA (Matsuura *et al.*, 1997; Cousineau *et al.*, 1998). The major host RNase H, RNase H1 encoded by *rnhA*, degrades the intron RNA template strand during or after cDNA synthesis, leaving residual RNA fragments that could serve as primers for top-strand DNA synthesis. Crucially, after synthesis of a full-length intron cDNA, either the group II intron RT or host DNA polymerase extends bottom-strand synthesis into the 5' exon,

yielding a branched intermediate that is recognized by the replication restart proteins PriA or PriC, which act preferentially on intermediates with short or long gaps between the branch and the 3' end of the nascent strand (Heller and Marians, 2005a). These replication restart proteins then initiate a replisome-loading cascade leading to top-strand DNA synthesis by the host replicative polymerase, Pol III. The 5'→3' exonuclease activity of Pol I is possibly required for second-strand synthesis during retrohoming, presumably to degrade RNA primers attached to newly synthesized DNA, and Pol I DNA polymerase activity could additionally contribute by helping to fill gaps, both functions of Pol I in host cell DNA replication (Okazaki *et al*., 1971; Konrad and Lehman, 1974). Additional proteins including LigA and RecJ were found in the biochemical assays to possibly be involved in sealing nicks and resecting the 5' overhangs resulting from the staggered double-strand break made by the group II intron RNP. The proteins identified here likely represent the majority of host factors involved during the late-steps of group II intron retrohoming.

Our results have implications for other mobile genetic elements. Like mobile group II introns, most non-LTR-retrotransposons do not encode RNase H and presumably rely on a cellular enzyme to degrade the RNA template strand after cDNA synthesis (Eickbush and Malik, 2002). The mechanism used for second-strand DNA synthesis by non-LTR-retrotransposons is unknown, but given that non-LTR-retrotransposons carry out reverse transcription in the nucleus could well involve the use of a host DNA polymerase and replication restart proteins as found here for group II introns. Secondly, these findings resemble recent results for bacteriophage Mu where PriA was found to be required for filling in 5-bp gaps at each end of the Mu insertion in the absence of DNA

replication (Jang *et al*., 2012). Thus, replication restart proteins may play a more general role both in the repair of DNA damage and propagation of mobile elements than was thought previously, including as an integral part of the group II intron retrohoming mechanism.

## 2.4 Methods

### 2.4.1 *E. coli* strains and growth conditions

*E. coli* HMS174(DE3) (Novagen) was used for the transposon library screen and retrohoming assays and DH5α was used for cloning. The construction of the *mariner* transposon library in HMS174(DE3) was described previously (Zhao *et al*., 2008). *E. coli* Keio deletions and their parental wild-type strain BW25113 were obtained from the National BioResource Project (National Institute of Genetics, Japan). Wild-type SS996 and mutant strains in this genetic background were obtained from Dr. Steven Sandler (University of Massachusetts) (Long *et al*., 2010). A complete listing of strains and genotypes is given in Table 2.4. Cells were grown in Luria-Bertani (LB). Antibiotics were added at the following concentrations: ampicillin, 100 μg/ml; chloramphenicol, 25 μg/ml; kanamycin, 40 μg/ml; 10 μg/ml. Thymine was added at 2 μg/ml.

### 2.4.2 Recombinant plasmids

pBL1Cap is a broad host range intron-donor plasmid that uses an *m*-toluic acid-inducible promoter to express the Ll.LtrB-ΔORF intron and flanking exons. It was derived from the broad host range intron-donor plasmid pBL1 (Yao and Lambowitz, 2007) by replacing the $tet^R$ marker with a $cam^R$ marker (1.5-kb NheI/PshAI fragment of

pACD3 blunt ended and cloned in place of *tet*$^{R}$ between the FspI sites of the pBL1). pBL1-rhlE is a derivative of pBL1Cap that expresses an Ll.LtrB-ΔORF intron that was retargeted to insert into a site in the antisense strand of the *rhlE* gene (Perutka *et al*., 2004).

### 2.4.3 FACS analysis

Wild-type and mutant SS996 (P*sulA*-GFP) strains were grown at 30°C to $O.D._{595}$ = 0.2-0.3 and induced with 4 mM *m*-toluic acid for 1 h at 37°C. Cells were collected by centrifugation, resuspended in phosphate-buffered saline (140 mM NaCl, 2.7 mM KCl, 9 mM $Na_2HPO_4$, 1.6 mM $KH_2PO_4$, pH 7.4), and analyzed by using a FACS Caliber (Becton Dickinson), with filter FL1 set at 530 ± 30 nm. FACS data were analyzed with the CELLQuest Pro program (Becton Dickinson).

### 2.4.4 Taqman qPCR assay of retrohoming

Retrohoming frequencies in Keio deletion and temperature-sensitive mutant strains were determined by using a retargeted Ll.LtrB-ΔORF intron (rhlE481a) that inserts at a site in the chromosomal *rhlE* gene (Perutka *et al*., 2004) and quantifying the 5'- and 3'-integration junctions relative to the number of *rhlE* genes by Taqman qPCR. For this assay, *E. coli* strains were transformed with pBL1-rhlE and grown overnight at 37°C in LB medium containing chloramphenicol. The overnight culture was subcultured into fresh medium and incubated at 37°C until $O.D._{595}$ = 0.2-0.3. Then, triplicate 3-ml cultures were induced with 4 mM *m*-toluic acid for 1 h at 37°C. The *priA* Keio deletion strain was grown and induced at 30°C to avoid the accumulation of suppressor mutations. Five temperature-sensitive mutants (*dnaE*$^{ts}$, *gyrB*$^{ts}$, *ligA*$^{ts}$, *rpoH*$^{ts}$, and *ssb*$^{ts}$) that could not be grown at 37°C were grown at 30°C and then shifted to 37°C for 1 h followed by a 1-h

induction with *m*-toluic acid. After induction, cells were lysed immediately and DNA was extracted by using a DNeasy Blood and Tissue Kit (Qiagen).

Taqman qPCR was carried out on 10 ng of total DNA in 384-well plates using a Universal qPCR Master Mix kit (Applied Biosystems), with the following primer/probe sets:

(i) 5'-integration junction.

P5-forward 5'-GGTGCAAACCAGTCACAGTAATG;

reverse 5'-GTCAGCTTCATCGAGGACGAG;

Taqman probe 5'-CAAGGCGGTACCTCC;

(ii) 3'-integration junction.

P3- forward 5'-ATAAAGCCCATGTCGAGCATG;

reverse 5'-TGTAAGATAACACAGAAAACAGCCAA;

Taqman probe 5'-TGCGCCCAGATAGGGTGTTAAGTCAAGTAGT;

(iii) *rhlE* gene.

rhlE-forward 5'-CAGCAACGTCCCGGGG;

reverse 5'-ACGCAGTTTCATCATCTGCG;

Taqman probe 5'-CCACCAGCACATCAACGCCGC.

Taqman qPCR primers and probes with a 5' FAM (6-carboxyfluorescein) label and 3' MGB (dihydrocyclopyrroloindole tripeptide major groove binder) quencher were obtained from Applied Biosystems. Standard curves were generated by serial ten-fold dilutions of a TOPO-2.1 vector carrying a cloned DNA fragment containing the Ll.LtrB-ΔORF intron integrated within the *rhlE* gene. All primer/probe sets had > 90% amplification efficiency over the concentration range of the standard curve. Retrohoming

frequencies were calculated as the numbers of 5'- and 3'-integration junctions relative to the number of *rhlE* genes after subtraction of background signal without *m*-toluic acid induction and are the mean ± standard error of the mean (S.E.M.) for triplicate 1-h inductions.

Table 2.1: Taqman qPCR assays of retrohoming in notable *E. coli* mutants.

| Gene | Function | Retrohoming frequency (% WT) | |
|---|---|---|---|
| | | 5' Junction | 3' Junction |
| *dinB*[†] | DNA polymerase IV | 125 ± 4% | 127 ± 3% |
| *dnaB*[ts] | Replicative DNA helicase | 41 ± 13% | 77 ± 3% |
| *dnaC*[ts,#] | Replication/restart initiation | 21 ± 1% | 10 ± 1% |
| *dnaE*[ts,†] | DNA polymerase III, α-subunit | 27 ± 4% | 47 ± 9% |
| *dnaG*[ts] | DNA primase | 36 ± 5% | 26 ± 7% |
| *dnaQ*[†] | DNA polymerase III, ε-subunit | 135 ± 5% | 128 ± 5% |
| *dnaT*[#] | Primosomal protein I | 39 ± 3% | 67 ± 3% |
| *gyrB*[ts,#] | DNA gyrase, subunit B | 14 ± 8% | 26 ± 10% |
| *hns*[†] | Histone-like nucleoid structuring protein | 20 ± 5% | 16 ± 5% |
| *lexA*[a] | Transcriptional repressor of SOS | 87 ± 13% | 81 ± 13% |
| *ligA*[ts,†] | DNA ligase | 113 ± 1% | 101 ± 3% |
| *ligB* | DNA ligase | 122 ± 4% | 102 ± 4% |
| *mnmE*[†] | tRNA modification | 128 ± 12% | 109 ± 11% |
| *pnp*[†] | Polynucleotide phosphorylase | 88 ± 4% | 102 ± 1% |
| *polAex*[ts,†] | DNA polymerase I | 38 ± 5% | 50 ± 6% |
| *polB*[†] | DNA polymerase II | 98 ± 9% | 84 ± 7% |
| *priA* | Primosomal protein N' | 52 ± 2% | 54 ± 1% |
| *priB* | Primosomal protein N | 92 ± 1% | 79 ± 2% |
| *priC* | Primosomal protein N" | 35 ± 3% | 23 ± 3% |
| *recF*[†] | ss/dsDNA binding protein | 125 ± 6% | 111 ± 4% |
| *recJ*[†] | 5'→3' exonuclease | 112 ± 2% | 104 ± 5% |
| *recQ*[†] | ATP-dependent DNA helicase | 115 ± 7% | 90 ± 2% |
| *rep*[†] | ATP-dependent DNA helicase | 167 ± 5% | 99 ± 5% |
| *rnhA*[†] | RNase HI | 11 ± 1% | 15 ± 2% |
| *rnhB*[†] | RNase HII | 106 ± 3% | 79 ± 2% |
| *rpoH*[#] | RNA polymerase $\sigma^{32}$ ($\sigma^{H}$) factor | 14 ± 4% | 11 ± 2% |
| *sbcC*[†,b] | ATP-dependent dsDNA exonuclease | 36 ± 3% | 31 ± 1% |
| *sbcD*[†] | ATP-dependent dsDNA exonuclease | 101 ± 7% | 103 ± 8% |
| *seqA*[†] | Inhibitor of replication initiation | 51 ± 8% | 44 ± 7% |
| *ssb*[ts] | ssDNA binding protein | 37 ± 3% | 41 ± 1% |
| *stpA*[†] | H-NS-like DNA/RNA-binding protein | 146 ± 3% | 123 ± 5% |
| *tus*[†] | Inhibition of replication at Ter sites | 39 ± 8% | 31 ± 9% |
| *umuC*[†] | DNA polymerase V | 119 ± 3% | 117 ± 7% |
| *umuD*[†] | DNA polymerase V | 142 ± 5% | 121 ± 4% |

The Table summarizes Taqman qPCR assays of Ll.LtrB intron retrohoming into a chromosomal *rhlE* target site in mutant strains. Keio deletion, non-temperature-sensitive mutants, and their parental wild-type strains containing donor plasmid pBL1-rhlE were grown to mid-log phase at 37°C and then intron expression was induced with 4 mM *m*-toluic acid for 1 h at 37°C. Five temperature-sensitive mutants (*dnaE*$^{ts}$, *gyrB*$^{ts}$, *ligA*$^{ts}$, *rpoH*$^{ts}$, and *ssb*$^{ts}$) that could not be grown at 37°C and their parental wild-type strains were grown at 30°C and then shifted to 37°C for 1 h prior to a 1-h induction with 4 mM *m*-toluic acid at 37°C. The *priA* deletion strain was grown and induced at 30°C and was confirmed by sequencing to lack second-site suppressor mutations in *dnaC*, which are known to accumulate in PriA mutants (Sandler *et al*., 1999). Taqman qPCR was carried out on extracted DNA to determine the number of 5'- and 3'-integration junctions relative to the number of *rhlE* genes (see Figure 2.1 and 2.4 Methods). Values are the mean ± S.E.M. for three experimental replicates normalized to the retrohoming frequency of the parental wild-type strain assayed in parallel. Mutants that showed decreased retrohoming frequencies were assayed at least three times. Retrohoming frequencies of parental wild-type control strains for mutants indicated in parentheses expressed as percent of available *rhlE* targets sites were: BW25113 (Keio deletion strains) 34%; AB1157 (*dnaE*$^{ts}$) 20%; N2603 (*ligA*$^{ts}$) 40%; BW30384 (*polAex*$^{ts}$) 25%; DG76 (*dnaB*$^{ts}$, *dnaC*$^{ts}$, and *dnaG*$^{ts}$) 56%; KL921 (*ssb*$^{ts}$) 40%; PR100 (*pnp*$^{ts}$) 31%; SS996 (*priB*, *lexA*) 60%; EJ1261(*gyrB*$^{ts}$) 30%; KY1445(*rpoH*$^{ts}$) 30%.

[ts] Temperature sensitive.

[†] Genes in which mutations decreased retrohoming frequencies in published genetic screens (Smith *et al*., 2005; Beauregard *et al*., 2006; Coros *et al*., 2008).

[#] Genes also identified as contributing to retrohoming in the transposon-insertions screen using the Tp$^{R}$-RAM assay in (Yao, 2008). Retrohoming efficiencies in the Tp$^{R}$-RAM assay relative to the wild-type control were: *gyrB*, 1.1%; *recJ*, 18.9%: *rpoH*, 3.4%; and *yjjB* (upstream gene in operon with *dnaC* and *dnaT*), 0%.

[a] The *lexA* mutant is *lexA51::Tn5* in the SS996 strain background and has a constitutively induced SOS response.

[b] The *sbcC* mutant was not deficient in retrohoming in Taqman qPCR assays of retrohoming at 30°C (Table 2.3).

Table 2.2: Taqman qPCR assays of retrohoming in other *E. coli* Keio deletion mutants analyzed in this work.

| Gene | Function | Retrohoming frequency (% WT) | |
|---|---|---|---|
| | | **5' Junction** | **3' Junction** |
| **(A) Nucleic acid related** | | | |
| *agaR* | Transcriptional repressor | 100 ± 3% | 121 ± 3% |
| *cpdA* | cAMP phosphodiesterase | 84 ± 9% | 92 ± 7% |
| *dtd* | D-Tyr-tRNA$^{Tyr}$ deacylase | 81 ± 11% | 90 ± 13% |
| *helD* | DNA helicase IV | 123 ± 2% | 151 ± 15% |
| *hofB* | Protein involved in plasmid replication | 115 ± 2% | 117 ± 3% |
| *hofC* | Maintains mini-F plasmids | 110 ± 7% | 99 ± 9% |
| *mutM* | Formamidopyrimidine DNA glycosylase | 83 ± 4% | 86 ± 5% |
| *nudF* | ADP-ribose pyrophosphatase | 91 ± 6% | 89 ± 5% |
| *recO* | RecFOR recombinase component | 88 ± 10% | 94 ± 6% |
| *recR* | RecFOR recombinase component | 86 ± 7% | 89 ± 5% |
| *rpoN* | RNA polymerase $\sigma^{54}$ ($\sigma^{N}$) factor | 80 ± 7% | 80 ± 5% |
| *udk* | Uridine/cytidine kinase | 128 ± 4% | 116 ± 3% |
| *xerD* | Site-specific recombinase | 101 ± 3% | 84 ± 4% |
| *ybeB* | Ribosome associated protein | 137 ± 2% | 130 ± 1% |
| **(B) Enzymes** | | | |
| *aldB* | Acetaldehyde dehyrogenase | 89 ± 4% | 84 ± 5 |
| *allB* | Allantoinase | 99 ± 4% | 89 ± 6% |
| *avtA* | Valine-pyruvate aminotransferase | 139 ± 7% | 111 ± 20% |
| *cadA* | Lysine decarboxylase I | 81 ± 1% | 89 ± 2% |
| *csrD* | Regulator of csrB and csrC decay | 137 ± 13% | 137 ± 17% |
| *dsbC* | Protein disulfide isomerase II | 96 ± 8% | 91 ± 9% |
| *ecpD* | Pilin chaperone | 152 ± 5% | 154 ± 2% |
| *glnD* | Uridylyltransferase | 103 ± 2% | 108 ± 0% |
| *gpp* | Guanosine pentaphosphatase | 102 ± 7% | 96 ± 5% |
| *hslV* | Peptidase component of HslUV protease | 98 ± 8% | 91 ± 8% |

Table 2.2 continued.

| | | | |
|---|---|---|---|
| *mdoB* | Phosphoglycerol transferase I | 73 ± 3% | 66 ± 2% |
| *metL* | Aspartate kinase II | 93 ± 4% | 92 ± 3% |
| *mhpE* | 4-hydroxy-2-ketovalerate aldolase | 84 ± 5% | 84 ± 6% |
| *rutB* | Peroxyurieidoacrylate amido hydrolase | 105 ± 8% | 100 ± 7% |
| *sseA* | 3-mercaptopyruvate sulfurtransferase | 107 ± 5% | 110 ± 8% |
| *ycaO* | β-methylthiolation of ribosomal protein S12 | 110 ± 5% | 107 ± 3% |
| *yggF* | Fructose-1,6-biphosphatase | 136 ± 7% | 129 ± 8% |
| *yiaK* | 2,3-diketo-L-gulonate dehydrogenase | 160 ± 6% | 156 ± 7% |
| *yidA* | Sugar phosphatase | 87 ± 1% | 80 ± 3% |
| *yqiI* | Detoxification of methylglyoxal | 127 ± 13% | 117 ± 13% |
| **(C) Membrane proteins** | | | |
| *emrE* | Multidrug efflux transporter | 118 ± 4% | 110 ± 6% |
| *hokC* | Toxic membrane protein | 121 ± 6% | 110 ± 4% |
| *nhaA* | Sodium-proton antiporter | 133 ± 7% | 136 ± 1% |
| *ptsG* | Glucose PTS permease | 123 ± 9% | 130 ± 1% |
| *srlA* | Glucitol/sorbitol PTS permease | 86 ± 3% | 84 ± 1% |
| *tonB* | Membrane spanning protein | 80 ± 4% | 76 ± 5% |
| *yiaM* | 2,3-diketo-L-gulonate-$Na^+$ symporter subunit | 134 ± 16% | 115 ± 5% |
| *yiaN* | 2,3-diketo-L-gulonate-$Na^+$ symporter subunit | 110 ± 8% | 116 ± 10% |
| *yiaO* | 2,3-diketo-L-gulonate-$Na^+$ symporter subunit | 125 ± 1% | 112 ± 2% |
| **(D) Unidentified** | | | |
| *htrE* | Putative outer membrane protein | 86 ± 5% | 79 ± 6% |
| *paoD* | Conserved protein | 73 ± 2% | 68 ± 3% |
| *ppdD* | Putative type IV pilin | 121 ± 10% | 114 ± 6% |
| *yaaH* | Conserved inner membrane protein | 90 ± 5% | 100 ± 2% |
| *yagP* | Predicted transcriptional regulator | 80 ± 5% | 84 ± 4% |
| *yahF* | Predicted acyl-CoA synthetase | 85 ± 1% | 80 ± 1% |
| *ybjX* | Conserved protein | 107 ± 5% | 101 ± 5% |
| *ycjW* | Predicted transcriptional regulator | 109 ± 2% | 99 ± 1% |

Table 2.2 continued.

| | | | |
|---|---|---|---|
| *ycjZ* | Predicted transcriptional regulator | 93 ± 7% | 89 ± 4% |
| *ydcM* | Predicted transposase | 82 ± 2% | 74 ± 1% |
| *yggC* | Conserved protein | 87 ± 7% | 96 ± 9% |
| *yggD* | Predicted transcriptional regulator | 87 ± 6% | 95 ± 6% |
| *yghA* | Predicted glutathionylspermidine synthase | 100 ± 13% | 105 ± 9% |
| *ygiC* | Predicted enzyme | 100 ± 8% | 99 ± 4% |
| *yhdE* | Conserved protein | 94 ± 10% | 88 ± 12% |
| *yidB* | Conserved protein | 94 ± 9% | 78 ± 7% |
| *yidR* | Conserved protein | 108 ± 4% | 96 ± 3% |
| *yifO* | Conserved protein | 113 ± 0% | 112 ± 9% |
| *yiiD* | Predicted acetyltransferase | 118 ± 7% | 125 ± 4% |
| *yjjB* | Conserved inner membrane protein | 89 ± 12% | 86 ± 11% |
| *ysaA* | Predicted hydrogenase | 100 ± 3% | 99 ± 3% |

The Table summarizes retrohoming frequencies for integration into a chromosomal *rhlE* target site in Keio deletion strains based on Taqman qPCR quantitation of the 5'- and 3'-integration junctions relative to the number of *rhlE* genes. Assays were done on total DNA from cells containing donor plasmid pBL1-rhlE. Cells were grown to early mid-log phase at 37°C and then intron-expression induced with 4 mM *m*-toluic acid for 1 h. Values are the mean ± S.E.M. for three experimental replicates normalized to the retrohoming frequency of the wild-type control strain BW25113 assayed in parallel.

Table 2.3: Taqman qPCR assays of retrohoming in *E. coli* Keio deletion mutants at 30°C.

| Gene | Function | Retrohoming frequency (% WT) | |
|---|---|---|---|
| | | 5' Junction | 3' Junction |
| *avtA* | aminotransferase | 107 ± 5% | 110 ± 4% |
| *corA* | magnesium transporter | 64 ± 3% | 57 ± 2% |
| *dinG* | ATP-helicase | 123 ± 1% | 113 ± 1% |
| *dnaB*$^{ts}$ | replication/restart DNA helicase | 71 ± 2% | 77 ± 3% |
| *dnaC*$^{ts}$ | replication/restart initiation | 49 ± 6% | 50 ± 6% |
| *dnaE*$^{ts, \dagger}$ | DNA Polymerase III, α-subunit | 36 ± 6% | 31 ± 5% |
| *dnaT* | restart replication at gaps | 69 ± 5% | 41 ± 2% |
| *dnaQ*$^{\dagger}$ | DNA Polymerase III, ε-subunit | 114 ± 7% | 115 ± 8% |
| *dsbC* | Protein Disulfide Isomerase II | 115 ± 12% | 105 ± 3% |
| *evgA* | transcriptional activator | 95 ± 13% | 87 ± 11% |
| *gppA* | Guanosine pentaphosphatase | 90 ± 12% | 89 ± 2% |
| *gutM* | gutM transcriptional activator | 103 ± 2% | 121 ± 4% |
| *gutQ* | arabinose isomerase | 154 ± 5% | 150 ± 5% |
| *hofB* | conserved protein | 116 ± 5% | 108 ± 5% |
| *hofC* | protein transport protein | 74 ± 2% | 80 ± 2% |
| *ligB* | DNA ligase | 124 ± 6% | 116 ± 2% |
| *lyxK* | L-xyulose kinase | 94 ± 12% | 94 ± 8% |
| *mgtA* | ATPase magnesium transporter | 118 ± 3% | 114 ± 5% |
| *mhpE* | aldolase | 98 ± 4% | 97 ± 5% |
| *mutM* | DNA glycosylase | 102 ± 9% | 99 ± 8% |
| *phoP* | transcriptional dual regulator | 84 ± 5% | 86 ± 6% |
| *polA*$^{ts, \dagger}$ | DNA Polymerase I | 57 ± 5% | 48 ± 4% |
| *prfB* | peptide chain release factor RF2 | 91 ± 1% | 101 ± 6% |
| *ppdD* | putative type IV pilin | 10 ± 1% | 4 ± 1% |
| *priA* | restart of short-gap stalled forks | 78 ± 4% | 61 ± 4% |
| *priB* | restart in priA pathway | 88 ± 1% | 80 ± 3% |
| *priC* | restart of large-gap stalled forks | 118 ± 3% | 103 ± 2% |
| *ptsG* | Glucose-specific PTS enzyme | 125 ± 2% | 119 ± 1% |
| *recF*$^{\dagger}$ | ss/dsDNA binding protein | 2 ± 1% | 2 ± 1% |
| *recJ*$^{\dagger}$ | 5'-->3' exonuclease | 98 ± 13% | 95 ± 11% |
| *recO* | subunit RecFOR complex | 107 ± 7% | 104 ± 9% |

Table 2.3 continued.

| | | | |
|---|---|---|---|
| *recR* | recombination and repair | 99 ± 7% | 97 ± 6% |
| *rep*† | ATP-dependent helicase | 131 ± 5% | 117 ± 2% |
| *rhlB* | ATP-RNA helicase | 81 ± 7% | 89 ± 8% |
| *rnhA*† | degrades DNA-RNA hybrids | 32 ± 4% | 32 ± 2% |
| *rutA* | pyrimidine utilization | 139 ± 9% | 144 ± 14% |
| *rutB* | predicted enzyme | 93 ± 4% | 90 ± 2% |
| *rutC* | predicted endoribonuclease | 114 ± 12% | 134 ± 14% |
| *rutD* | α/β hydrolase | 108 ± 5% | 118 ± 7% |
| *rutE* | predicted nitroreductase | 103 ± 8% | 109 ± 3% |
| *rutG* | predicted xanthine transporter | 92 ± 7% | 88 ± 1% |
| *sbcC*† | dsDNA exonuclease | 105 ± 3% | 113 ± 1% |
| *sbcD*† | dsDNA exonuclease | 102 ± 7% | 96 ± 8% |
| *sgbH* | decarboxylase | 101 ± 2% | 102 ± 2% |
| *sgbU* | xyulose epimerase | 117 ± 10% | 99 ± 6% |
| *sgbE* | ribulose epimerase | 143 ± 25% | 124 ± 10% |
| *ssb*ts | ssDNA binding protein | 53 ± 2% | 51 ± 4% |
| *srlA* | IIC component of PTS | 107 ± 2% | 106 ± 5% |
| *srlB* | subunit PTS permease | 104 ± 5% | 114 ± 3% |
| *srlD* | sorbitol dehydrogenase | 106 ± 10% | 135 ± 14% |
| *srlE* | subunit PTS permease | 139 ± 4% | 155 ± 7% |
| *srlR* | gutR transcriptional repressor | 94 ± 8% | 96 ± 4% |
| *ycbB* | L,D-transpeptidase | 109 ± 15% | 100 ± 13% |
| *ycbK*ts | conserved protein | 78 ± 7% | 73 ± 9% |
| *ycbL* | metal-binding enzyme | 131 ± 11% | 134 ± 8% |
| *yejH* | predicted ATP-helicase | 130 ± 1% | 120 ± 1% |
| *yiaI* | predicted hydrogenase | 108 ± 5% | 106 ± 5% |
| *yiaL* | conserved protein | 115 ± 3% | 113 ± 7% |
| *yiaK* | gulonate reductase | 117 ± 3% | 118 ± 4% |
| *yiaM* | subunit yiaMNO transporter | 52 ± 4% | 53 ± 5% |
| *yiaN* | subunit yiaMNO transporter | 90 ± 3% | 82 ± 6% |
| *yiaO* | DKG transporter | 10 ± 2% | 10 ± 3% |
| *yidR*† | putative membrane protein | 44 ± 2% | 35 ± 1% |
| *yjjB* | inner membrane protein | 103 ± 8% | 102 ± 10% |
| *yoaA* | predicted helicase | 122 ± 5% | 115 ± 5% |
| *uvrB* | helicase in NER | 98 ± 3% | 95 ± 2% |

The Table summarizes Taqman qPCR assays of Ll.LtrB intron retrohoming into a chromosomal *rhlE* target site in mutant strains. Keio deletion, non-temperature-sensitive mutants, and their parental wild-type strains containing donor plasmid pBL1-rhlE were grown to mid-log phase at 30°C and then shifted to 37°C and induced with 4 mM *m*-toluic acid for 1 h. Taqman qPCR was carried out on extracted DNA to determine the number of 5'- and 3'-integration junctions relative to the number of *rhlE* genes (see Figure 2.1B and 2.4 Methods). Values are the mean ± S.E.M. for three experimental replicates normalized to the retrohoming frequency of the parental wild-type strain assayed in parallel. Mutants that showed decreased retrohoming frequencies were assayed at least three times. Retrohoming frequencies of parental wild-type control strains for mutants indicated in parentheses expressed as percent of available *rhlE* targets sites.

[ts] Temperature sensitive.

[†] Genes in which mutations decreased retrohoming frequencies in published genetic screens (Smith *et al*., 2005; Beauregard *et al*., 2006; Coros *et al*., 2008).

Table 2.4: *E. coli* strains used in this work.

| Strain | *ts* | Genotype |
|---|---|---|
| AB1157[a] | N | F$^-$, *thr-1*, *araC14*, *leuB6*(Am), *Δ(gpt-proA)62*, *lacY1*, *tsx-33*, *qsr'-0*, *glnV44*(AS), *galK2*(Oc), l$^-$, *Rac-0*, *isG4*(Oc), *rfbC1*, *mgl-51*, *rpoS396*(Am), *rpsL31*(strR), *kdgK51*, *xylA5*, *mtl-1*, *argE3*(Oc), *thi-1* |
| AB1157*dnaE*$^{ts}$ | Y | F$^-$, *thr-1*, *araC14*, *leuB6*(Am), *Δ(gpt-proA)62*, *lacY1*, *tsx-33*, *qsr'-0*, *glnV44*(AS), *galK2*(Oc), l$^-$, *Rac-0*, *isG4*(Oc), *rfbC1*, m*gl-51*, *rpoS396*(Am), *rpsL31*(strR), *kdgK51*, *xylA5*, *mtl-1*, *argE3*(Oc), *thi-1*,*dnaE486*(ts), *zae502::Tn10* |
| BW25113[b] | N | F$^-$, *Δ(araD-araB)567*, *ΔlacZ4787*(::rrnB-3), *λ*$^-$, *rph-1*, *Δ(rhaD-rhaB)568*, *hsdR514* |
| BW30384[c] | N | F$^-$, *l*$^-$, *IN(rrnD-rrnE)1* |
| DG76[d] | N | F$^-$, *leuB6*(Am), *l*$^-$, *thyA47*, *rpsL153*(strR), *deoC3* |
| DH5a | N | F$^-$, ϕ80*lacZΔ*M15 *Δ*(*lacZYA-argF*), U169, *recA1*, *endA1*, *hsdR17*($r_K^-$, $m_K^+$), *gal*$^-$, *phoA*, *supE44*, *λ*$^-$, *thi-1*, *gyrA96*, *relA1* |
| EJ1261[e] | N | F$^-$, *galK2*(Oc), *λ*$^-$, *IN(rrnD-rrnE)1*, *rpsL200*(strR), *maeA1* |
| HMS174(DE3) | N | F$^-$, *recA1*, *hsdR*($r_{K12}^-$, $m_{K12}^+$), *Rif*$^r$, λDE3 |
| KL921[f] | N | F$^-$, *Δ(gpt-lac)5*, *LAMcI*(ind), *thyA0*, *rpsL*-(strR), *malE145::Tn10*, *deo-* |
| KL922*ssb*$^{ts}$ | Y | F$^-$, *Δ(gpt-lac)5*, *LAMcI*(ind), *thyA0*, *rpsL*-(strR), *malE145::Tn10*, *deo-*,*ssb-1*(ts) |
| KY1429*rpoH*$^{ts}$ | Y | F$^-$, *[araD139]$_{B/r}$*, *Δ(argF-lac)169*, *λ*$^-$, *flhD5301*, *Δ(fruK-yeiR)725(fruA25)*, *relA1*, *rpsL150*(strR), *zhh-50::Tn10*, *rpoH606*(ts), *rbsR22*, *Δ(fimB-fimE)632(::IS1)*, *deoC1* |
| KY1445[g] | N | F$^-$, *[araD139]$_{B/r}$*, *Δ(argF-lac)169*, *λ*$^-$, *flhD5301*, *Δ(fruK-yeiR)725(fruA25)*, *relA1*, *rpsL150*(strR), *zhh-50::Tn10*, *rbsR22*, *Δ(fimB-fimE)632(::IS1)*, *deoC1* |
| N2603[h] | N | F$^-$, *strA*$^r$, *ptsI105*, *r*$^-$*m*$^+$, *gal*$^+$ |
| N2603*ligA*$^{ts}$ | Y | F$^-$, *strA*$^r$, *ptsI105*, *r*$^-$*m*$^+$, *gal*$^+$,*lig7*(ts) |
| N4177*gyrB*$^{ts}$ | Y | F$^-$, *galK2*(Oc), *λ*$^-$, *IN(rrnD-rrnE)1*, *rpsL200*(strR), *gyrB221*(Cou$^R$), *gyrB203*(ts) |
| PC1*dnaC*$^{ts}$ | Y | F$^-$, *leuB6*(Am), *l*$^-$, *thyA47*, *rpsL153*(strR), *deoC3*,*dnaC1*(ts) |
| PC3*dnaG*$^{ts}$ | Y | F$^-$, *leuB6*(Am), *l*$^-$, *thyA47*, *rpsL153*(strR), *deoC3*,*dnaG3*(ts) |
| PC8*dnaB*$^{ts}$ | Y | F$^-$, *leuB6*(Am), *l*$^-$, *thyA47*, *rpsL153*(strR), *deoC3*,*dnaB8*(ts) |
| PR7 | N | F-, *thr-1*, *leuB6*(Am), *lacY1*, *rna-19*, *λ*$^-$, *pnp-7*, *rpsL132*(strR), *malT1*(λ$^R$), *xyl-7*, *mtlA2*, *thiE1* |
| PR100[i] | N | F-, *thr-1*, *leuB6*(Am), *lacY1*, *rna-19*, *λ*$^-$, *rpsL132*(strR), *malT1*(λ$^R$), *xyl-7*, *mtlA2*, *thiE1* |
| RS5064*polAex*$^{ts}$ | Y | F$^-$, *l*$^-$,*trpA33*, *IN(rrnD-rrnE)1*, *polA480*(ts,EX) |

Table 2.4 continued.

| | | |
|---|---|---|
| SS996[j] | N | F$^-$, lacMS286, *argE3*, *his-4*, *thi-1*, *xyl-5*, *mtl-1*, *sulB103*, *Δ* (attB):P*sulA-gfp* |
| SS1091 | N | F$^-$, lacMS286, *argE3*, *his-4*, *thi-1*, *xyl-5*, *mtl-1*, *sulB103*, *Δ* (attB):P*sulA-gfp*, *dnaC809* |
| SS1411 | N | F$^-$, lacMS286, *argE3*, *his-4*, *thi-1*, *xyl-5*, *mtl-1*, *sulB103*, *Δ* (attB):P*sulA-gfp*, *priA2::kan* |
| SS1419 | N | F$^-$, lacMS286, *argE3*, *his-4*, *thi-1*, *xyl-5*, *mtl-1*, *sulB103*, *Δ* (attB):P*sulA-gfp*, *zji-202::Tn10*, *dnaT822* |
| SS1443 | N | F$^-$, lacMS286, *argE3*, *his-4*, *thi-1*, *xyl-5*, *mtl-1*, *sulB103*, *Δ* (attB):P*sulA-gfp*, *D* (*priB*)*302* |
| SS3403 | N | F$^-$, lacMS286, *argE3*, *his-4*, *thi-1*, *xyl-5*, *mtl-1*, *sulB103*, *Δ* (attB):P*sulA-gfp*, *priC303::kan* |
| SS4610 | N | F$^-$, lacMS286, *argE3*, *his-4*, *thi-1*, *xyl-5*, *mtl-1*, *sulB103*, *Δ* (attB):P*sulA-gfp*, *lexA51::Tn5* |
| SS6239 | N | F$^-$, lacMS286, *argE3*, *his-4*, *thi-1*, *xyl-5*, *mtl-1*, *sulB103*, *Δ* (attB):P*sulA-gfp*, *zae-502::Tn10*, *dnaE486* |
| SS6253 | N | F$^-$, lacMS286, *argE3*, *his-4*, *thi-1*, *xyl-5*, *mtl-1*, *sulB103*, *Δ* (attB):P*sulA-gfp*, *malF::Tn10*, *dnaB8*(ts) |
| SS6668 | N | F$^-$, lacMS286, *argE3*, *his-4*, *thi-1*, *xyl-5*, *mtl-1*, *sulB103*, *Δ* (attB):P*sulA-gfp*, *aer-3075::Tn10*, *dnaG2903* |

The table summarizes the genetic backgrounds for each strain used in this work.
[ts] Temperature sensitive.
[a] AB1157 is the parental strain of AB1157*dnaE*$^{ts}$, from Dr. Marlene Belfort (University at Albany, State University of New York).
[b] BW25113 is the parental strain of the Keio collection, from the National BioResource Project (NIG, Japan).
[c]: BW30384 is the parental strain of RS5064*polAex*$^{ts}$, from the Coli Genetic Stock Center (CGCS) at Yale.
[d]: DG76 is the parental strain of PC1*dnaC*$^{ts}$, PC3*dnaG*$^{ts}$, and PC8*dnaB*$^{ts}$, from CGCS.
[e]: EJ1261 is the parental strain of N4177*gyrB*$^{ts}$, from CGCS.
[f]: KL921 is the parental strain of KL922*ssb*$^{ts}$, from CGCS.
[g]: KY1445 is the parental strain of KY1429*rpoH*$^{ts}$, from CGCS.
[h]: N2603 is the parental strain of N2603*ligA*$^{ts}$, from Dr. Marlene Belfort (University at Albany, State University of New York).
[i]: PR100 is the parental strain of PR7, from CGCS.
[j]: SS996 is the parental strain of other SS strains, from Dr. Steven Sandler (University of Massachusetts).

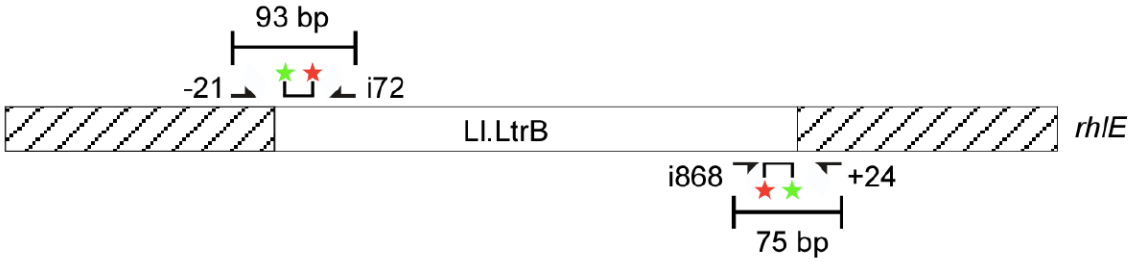
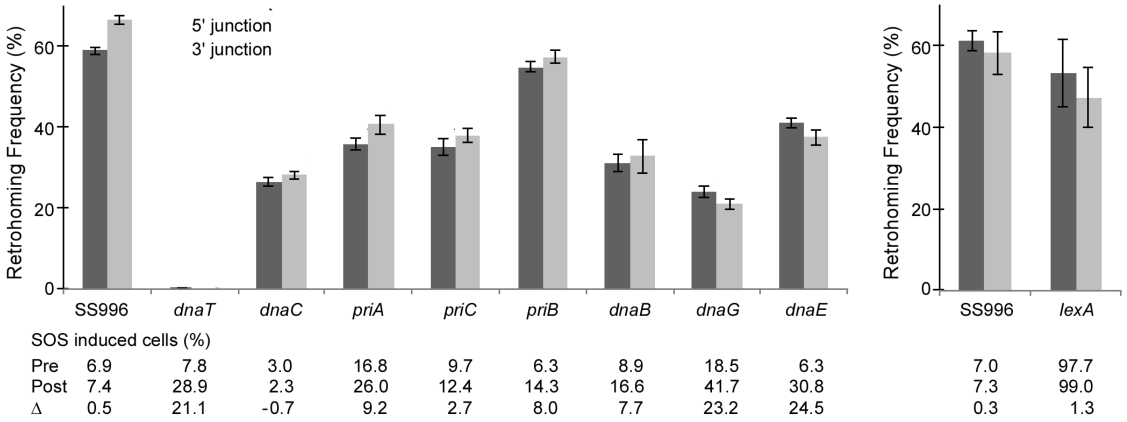



Figure 2.1: Taqman qPCR assay used to identify *E. coli* mutants deficient in retrohoming.

Taqman qPCR assay. The assay quantifies 5'- and 3'-intron integration junctions resulting from retrohoming of a retargeted Ll.LtrB-ΔORF intron into a site in the *rhlE* gene in the *E. coli* chromosome. Retrohoming events are quantified by Taqman qPCR, which utilizes the 5'→3' exonuclease activity of Taq DNA polymerase to cleave a fluorescently labeled DNA probe that base pairs to an internal region of a PCR amplicon. Digestion of the probe by Taq DNA polymerase releases the FAM label (red star) free of the MGB quencher (green star), resulting in a quantifiable fluorescence signal for each amplification event. The numbers of 5'- and 3'-intron integration junctions relative to the number of *rhlE* targets were determined by quantifying the fluorescence signals in three separate PCRs relative to standard curves generated from serial dilutions of a reference plasmid (see 2.4 Methods). Primers for these PCRs are depicted by arrows with numbers indicating the positions of the 5' nucleotide of the upstream primer and 3' nucleotide of the downstream primer relative to the intron-integration junction.




SOS induced cells (%)

| | SS996 | *dnaT* | *dnaC* | *priA* | *priC* | *priB* | *dnaB* | *dnaG* | *dnaE* | SS996 | *lexA* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Pre | 6.9 | 7.8 | 3.0 | 16.8 | 9.7 | 6.3 | 8.9 | 18.5 | 6.3 | 7.0 | 97.7 |
| Post | 7.4 | 28.9 | 2.3 | 26.0 | 12.4 | 14.3 | 16.6 | 41.7 | 30.8 | 7.3 | 99.0 |
| Δ | 0.5 | 21.1 | -0.7 | 9.2 | 2.7 | 8.0 | 7.7 | 23.2 | 24.5 | 0.3 | 1.3 |

Figure 2.2: Decreased retrohoming frequencies in replication restart mutants are not due to the SOS response.

*E. coli* SS996 P*sulA*-GFP strains with mutations in genes encoding replication restart proteins were grown in LB medium at 30°C until $O.D._{595}$ = 0.2-0.4, then shifted to 37°C and induced with 4 mM *m*-toluic acid for 1 h. Retrohoming frequencies were determined by Taqman qPCR assay of retrohoming into a chromosomal target site in the *rhlE* relative to the number of available *rhlE* target sites, and SOS induction in the same cultures was assessed by the difference (Δ) in the percentage of cells showing GFP fluorescence in a FACS assay before (pre) and after (post) the shift to 37°C. The error bars indicate the S.E.M. for three separate *m*-toluic acid-induced cultures.

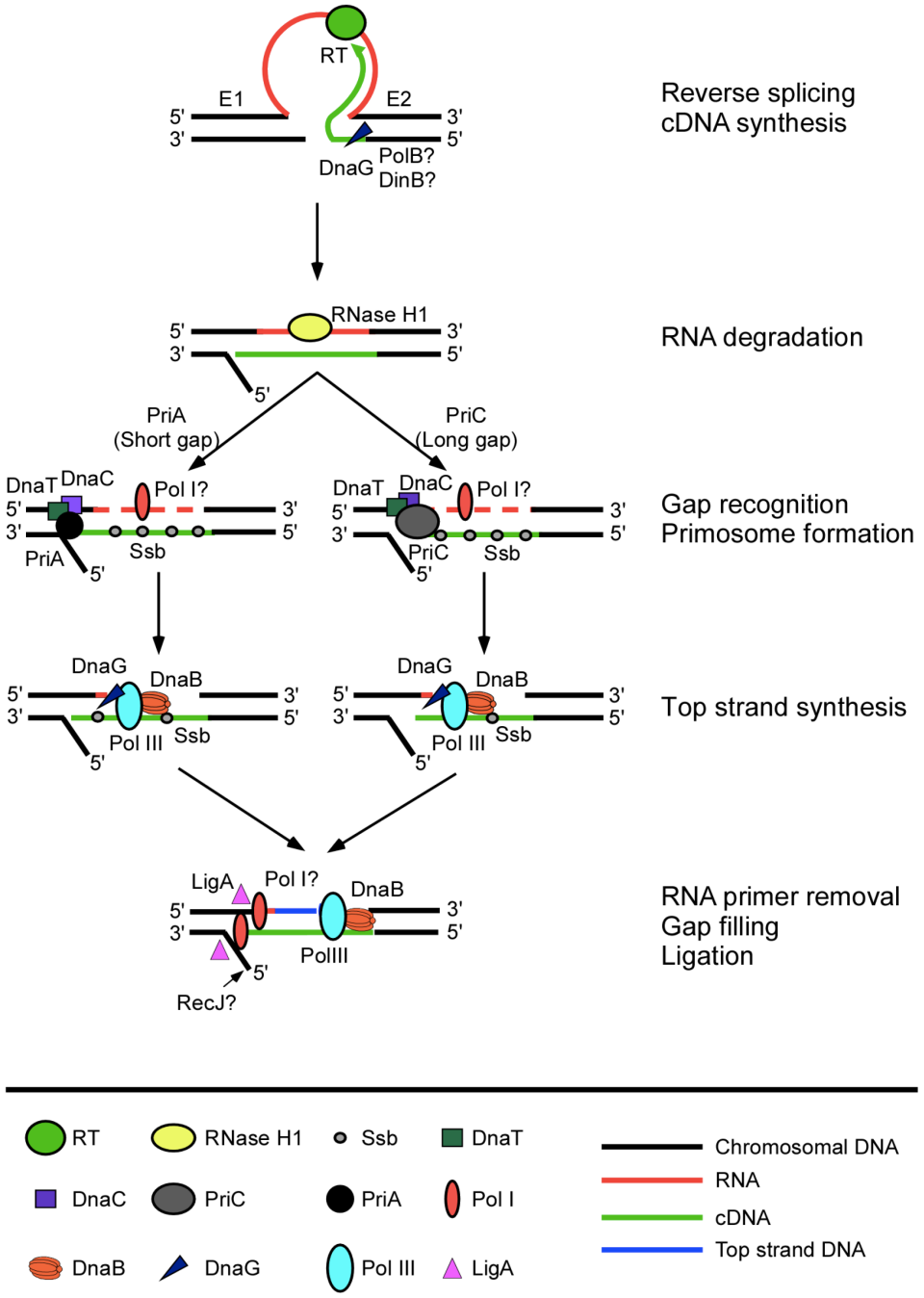



Figure 2.3: Model for function of host factors in group II intron retrohoming in *E. coli*.

In initial steps, the group II intron lariat RNA reverse splices into the top strand of the DNA target site, while the intron-encoded RT cuts the bottom DNA strand and uses the 3' end of the cleaved strand as a primer for target DNA-primed reverse transcription of the intron RNA. During or after cDNA synthesis, a host RNase H (RNase H1) degrades the intron RNA template strand. Extension of the intron cDNA into the 5' exon displaces the bottom-DNA strand resulting in a branched intermediate that is recognized by the replication restart proteins PriA or PriC, with PriA preferentially recognizing intermediates with short gaps in the bottom strand and PriC preferentially recognizing intermediates with long gaps in the bottom strand. PriA and PriC then initiate a replisome loading cascade involving the sequential recruitment of the replicative helicase DnaB, the primase DnaG, and the replicative polymerase Pol III for second-strand DNA synthesis. Ssb stabilizes single-stranded DNA in gapped regions and interacts with PriA to stimulate the loading of DnaB. The 5'→3' exonuclease activity of Pol I contributes to the removal of residual RNA primers and its DNA polymerase activity may contribute to filling in gaps. Biochemical assays indicate that a host DNA ligase (LigA) seals nicks in the top and bottom strands. Although bottom-strand synthesis is completely dependent on group II RT activity, biochemical assays show that it is strongly inhibited in a DNA primase (DnaG) mutant and moderately inhibited in repair DNA polymerase DinB and PolB mutants, suggesting a previously unsuspected role for host factors in initiating bottom-strand (cDNA) synthesis in extracts. Deletion of RecJ moderately inhibits synthesis of full-length bottom strands in biochemical assays, consistent with a role in resection of the 5'-overhang resulting from the staggered cleavage of the DNA substrate by group II intron RNPs (Smith *et al*., 2005).

# Chapter 3: Enhanced group II intron retrohoming in magnesium-deficient *Escherichia coli* via selection of mutations in the ribozyme core[2]

## 3.1 Introduction

Mobile group II introns are retrotransposons that are found in prokaryotes and the mitochondrial and chloroplast DNAs of some eukaryotes and are thought to be evolutionary ancestors of spliceosomal introns, the spliceosome, retrotransposons, and retroviruses in higher organisms (Lambowitz and Zimmerly, 2011). They consist of two components -- an autocatalytic intron RNA ("ribozyme") and an intron-encoded protein (IEP) with reverse transcriptase (RT) activity -- which function together in a ribonucleoprotein (RNP) complex to promote RNA splicing and site-specific integration of the intron into new DNA sites in a process called retrohoming (Lambowitz and Zimmerly, 2011). Like spliceosomal introns, group II introns splice via two sequential transesterification reactions that yield an excised intron lariat RNA (Peebles *et al*., 1986a). For group II introns, the splicing reactions are catalyzed by the intron RNA with the assistance of the IEP, which binds specifically to the intron RNA and stabilizes the catalytically active RNA structure. The IEP then remains bound to the excised intron lariat RNA in an RNP that promotes retrohoming via reverse splicing of the intron RNA directly into DNA sites followed by reverse transcription by the IEP. The resulting intron cDNA is integrated into the genome by host enzymes (Smith *et al*., 2005; Yao *et al*., 2013). The ribozyme-based, site-specific DNA integration mechanism used by group II introns enabled their development into gene targeting vectors ("targetrons"), which

[2] The following chapter appears as written from Truong DM, Sidote DJ, Russell R, and Lambowitz AM, (2013) Proc Natl Acad Sci USA, 110(40):E3800-E3809, with additional unpublished data. All work presented was written and performed by the dissertation author, with writing contributions from Russell R and Lambowitz AM. Sidote DJ contributed to Figures 3.1 and 3.17.

combine high integration efficiency and specificity with readily programmable DNA target specificity (Perutka *et al*., 2004). However, while group II introns retrohome and function efficiently for gene targeting in bacteria, their natural hosts, they do so inefficiently in eukaryotes, at least in part due to lower free $Mg^{2+}$ concentrations (Mastroianni *et al*., 2008), which decrease group II intron ribozyme activity (see below). These lower $Mg^{2+}$ concentrations constitute a natural barrier that impedes group II introns from invading the nuclear genomes of present day eukaryotes and limits their utility as gene targeting vectors for higher organisms.

Recent X-ray crystal structures of a group II intron RNA provide a structural framework for investigating group II intron splicing and retrohoming mechanisms and potentially for improving their function in gene targeting (Toor *et al*., 2008a, 2008b; Marcia and Pyle, 2012). Group II intron RNAs consist of six conserved domains (denoted DI-DVI), which interact via tertiary contacts to fold the RNA into a catalytically active three-dimensional structure (Figure 3.1*A*) (Lambowitz and Zimmerly, 2011). DV is a small conserved domain that binds catalytic metal ions and interacts with DI and J2/3 to form the intron RNA's active site. It is thought to be the cognate of the U2/U6 snRNAs of the spliceosome, and consequently, its architecture and function are central to understanding the mechanism and evolution of RNA splicing in higher organisms (Michel *et al*., 2009; Keating *et al*., 2010). DI, the largest domain, provides a structural scaffold for the assembly of the other domains and contains exon-binding sites that position the 5'- and 3'-splice sites and ligated-exon junction at the ribozyme active site for RNA splicing and reverse splicing reactions. DIII functions as a catalytic effector, DIV is

the location of the ORF encoding the IEP, and DVI contains the branch-point adenosine used for lariat formation.

Three major structural subclasses of group II intron RNAs, denoted IIA, IIB, and IIC, have been identified with differences in both peripheral and active-site elements (Lambowitz and Zimmerly, 2011). X-ray crystal structures of the *Oceanobacillus iheyensis* group IIC intron reveal the folded structure of DI-V, with and without bound exons (Toor *et al*., 2008a, 2008b; Marcia and Pyle, 2012). Although the *O. iheyensis* structures do not include DVI and the intron RNA construct lacks the ability to self-splice via lariat formation, they do show the active-site region formed by DI and DV, which can catalyze splicing via 5'-splice-site hydrolysis rather than branching. Importantly, the structures show that this active site contains a heteronuclear metal ion-binding center consisting of two pairs of $Mg^{2+}$ and $K^{+}$ ions bound site-specifically to DV, one pair (M1/K1) to a catalytic triad (CT) of three highly conserved nucleotides in the proximal stem of DV, and the other pair (M2/K2) to a dinucleotide bulge at the hinge region between the proximal and distal stems (Figure 3.1*B*, *C*). These two pairs of metal ions are brought together at the active site by a sharp bend in DV that moves the distal stem towards the catalytic triad. This sharp bend is stabilized both by bound metal ions bridging different regions of the phosphodiester backbone and by interactions between DV and other regions of the intron RNA (Toor *et al*., 2008a). By contrast, structures of DV of the yeast aI5γ and *Pylaiella littoralis* LSU/2 group II introns in isolation showed an extended helix with many of the same metal ion-binding sites but without the bend (Sigel *et al*., 2004; Seetharaman *et al*., 2006).

Besides the catalytic metal ions, several other $Mg^{2+}$ ions are seen at different sites in DV in the *O. iheyensis* RNA crystal structures. Three putative $Mg^{2+}$ ions lie within the major and minor grooves near the κ' motif in the proximal stem, and three others (denoted here as M3-M5) are bound to the distal stem, one (M3) within a kink adjacent to the G of the GNRA tetraloop, and another (M4) at the R of the GNRA tetraloop. The third $Mg^{2+}$ bound to the distal stem (M5) forms a bridge between the base pair distal to λ' and the third nucleotide of the catalytic triad, potentially stabilizing the sharp bend in DV. These additional $Mg^{2+}$-binding sites in the *O. iheyensis* intron are generally consistent with the locations of terbium-cleavage sites in DV of the aI5γ intron (Sigel *et al*., 2000).

Like other ribozymes, group II introns use $Mg^{2+}$ for both RNA folding and catalysis (DeRose, 2003; Sigel, 2005). However, the $Mg^{2+}$ concentrations required for group II intron function are higher than those for other ribozymes. Thus, mutations in the yeast mitochondrial $Mg^{2+}$ transporter Mrs2 specifically inhibit the splicing of all four yeast mt group II introns, while having minimal effects on the transcription or splicing of group I introns (Gregan *et al*., 2001). Moreover, efficient retrohoming in *Xenopus laevis* oocytes, *Drosophila melanogaster* embryos, and zebrafish (*Danio rerio*) embryos requires the co-injection of additional $Mg^{2+}$ to achieve an intracellular concentration of 5-10 mM (Mastroianni *et al*., 2008). Bacteria, the natural hosts of group II introns, typically have free intracellular $Mg^{2+}$ concentrations of 1-4 mM (Lusk *et al*., 1968), whereas *Xenopus laevis* oocyte nuclei contain 0.3 mM $Mg^{2+}$ (Horowitz and Tluczek, 1989), and mammalian cells contain 0.2-1 mM $Mg^{2+}$ during the majority of the cell cycle (Gunther, 2006). The latter values are well below the optimal $Mg^{2+}$ concentrations for protein-assisted group II intron RNA splicing and reverse splicing *in vitro* (5-10 mM) (Saldanha

*et al*., 1999). For some ribozymes, it has been possible to select new variants that function at lower $Mg^{2+}$ concentrations or utilize different metal ions (Lehman and Joyce, 1993; Bagby *et al*., 2009; Chen *et al*., 2009). However, the extent to which group II introns can be similarly evolved or engineered to decrease their $Mg^{2+}$ dependence is unknown.

Here, we used an *E. coli* mutant with a disruption in the magnesium transporter *mgtA* for *in vivo* selection of group II intron variants with enhanced function at lower $Mg^{2+}$ concentrations from libraries of *Lactococcus lactis* Ll.LtrB introns with mutations in DV. We thus identified 43 improved variants that have > 3-fold increased retrohoming efficiency within the *mgtA* disruptant. We find that all of these improved variants have mutations in the distal stem of DV, with the two most active variants (16- and 22-fold increased) having mutations restricted to the distal stem. Biochemical analysis demonstrates that the enhanced retrohoming of these variants reflects a higher fraction of intron RNAs that folds into an active structure at low $Mg^{2+}$ concentrations, and terbium-cleavage assays on isolated DV indicate that the mutations enhance $Mg^{2+}$ binding to sites in the distal stem. Together, our results suggest a model in which the variants achieve more efficient retrohoming and splicing by enhancing $Mg^{2+}$ binding to the distal stem of DV and promoting a conformational change that is required for formation of an active RNA structure, perhaps the sharp bending of DV needed to form the heteronuclear metal ion center at the active site.

## 3.2 Results

### 3.2.1 *E. coli* mutants with defects in $Mg^{2+}$ transport are deficient in group II intron retrohoming.

Because dysfunction of mitochondrial $Mg^{2+}$ transporters in *S. cerevisiae* inhibits the splicing of endogenous group II introns (Gregan *et al*., 2001), we reasoned that disrupting $Mg^{2+}$ transporters should similarly impair the function of the Ll.LtrB intron in *E. coli*. The resulting $Mg^{2+}$-deficient intracellular environment could then be used to select DV variants with a decreased $Mg^{2+}$ requirement. *E. coli* relies on two major $Mg^{2+}$ transport proteins, CorA and MgtA (Maguire, 2006), and the *Salmonella* homolog of CorA can rescue a group II intron splicing defect in a yeast mutant with a disruption of the mitochondrial $Mg^{2+}$ transporter Mrs2 (Bui *et al*., 1999).

We determined the retrohoming efficiency of the Ll.LtrB intron in *E. coli* mutants with targetron-mediated disruptions of the *corA* and *mgtA* genes (Yao *et al*., 2005) by using a plasmid-based assay (Karberg *et al*., 2001) (Figure 3.2*A*). In this assay, a precursor RNA containing an Ll.LtrB-ΔORF intron (*i.e*., an Ll.LtrB intron deleted for the ORF encoding the IEP) with a phage T7 promoter sequence inserted near its 3' end is expressed in tandem with the IEP (denoted LtrA protein) from a $Cap^R$ donor plasmid. The LtrA protein promotes the splicing and retrohoming of the Ll.LtrB-ΔORF intron carrying the T7 promoter into a target site cloned in an $Amp^R$ recipient plasmid upstream of a promoterless tetracycline-resistance ($tet^R$) gene, thereby activating that gene. After plating on LB medium containing antibiotics, retrohoming efficiencies are quantified as the ratio of ($Tet^R$ + $Amp^R$)/$Amp^R$ colonies.

The retrohoming efficiency of the Ll.LtrB-ΔORF in the wild-type *E. coli* HMS174(DE3) strain was 48% and was decreased ~2-fold in the *corA* disruptant and

200-fold in the *mgtA* disruptant (Figure 3.2*B*). Measurements of $Mg^{2+}$ concentration with the fluorescent probe mag-fura-2 (Froschauer *et al*., 2004) showed that exponentially growing wild-type HMS174(DE3) has an intracellular free $Mg^{2+}$ concentration of 3.1 mM, in the range found previously for *E. coli* (Lusk *et al*., 1968), whereas the *corA* and *mgtA* disruptants have lower but similar $Mg^{2+}$ concentrations (1.6 and 1.7 mM, respectively; Figure 3.2*B*). The much stronger inhibition of retrohoming in the *mgtA* disruptant may reflect that the fluorescent probe does not accurately measure the effective $Mg^{2+}$ concentration during group II intron homing in exponentially growing cells, either because of the non-growth conditions needed to permeabilize the cells to the fluorescent probe or because of differences in the local $Mg^{2+}$ concentrations in regions of the cell in which retrohoming occurs (see Discussion). In the experiments below, we used the *mgtA* disruptant to provide a stringent environment for the selection of group II intron variants with improved function at lower $Mg^{2+}$ concentrations.

### 3.2.2 Selection of functional DV variants from a partially randomized ("doped") library in an *E. coli mgtA* disruptant.

We used the plasmid-based mobility system described above to select variants of the Ll.LtrB-ΔORF intron that are active in retrohoming within the *mgtA* disruptant. The selection was done from a library of ~$10^9$ intron variants in which a 36-nt region encompassing DV was partially randomized ("doped") with 70% of the wild-type nucleotides and 10% of each of the three other nucleotides at each position. After two rounds of selection, we used colony PCR to amplify and sequence DVs from ~ 300 individual $Tet^R$ colonies and identified 106 unique DV sequences that had retrohomed in the *mgtA* disruptant (Figure 3.3). These variants contained 1-9 mutations, with an average of 2.6 mutations over the 36-nt randomized region. From these sequences, we generated a

mutational map displaying regions amenable or refractory to mutagenesis (Figure 3.4*A*). As expected, most of the mutants maintained the DV secondary structure and had mutations outside of regions known to be important for RNA catalysis and tertiary-structure formation. The most conserved regions were those containing the catalytic triad (CT), the κ' motif, and a conserved G involved in a *trans* sugar-edge/Hoogsteen (H) contact to the I(i) loop (Keating *et al*., 2010). Within the catalytic triad, the first two nucleotides (A and G) and the central G-U base pair were invariant, and the third nucleotide was mutated in only a single variant with strongly decreased retrohoming efficiency (DV93; C → G with a compensatory G → U mutation in the opposite nucleotide; Figure 3.3). Only three other nucleotides were invariant: the G immediately preceding the catalytic triad; a G that base pairs with a U or C in the λ motif; and the A in the terminal GNRA tetraloop. The most mutable regions were the three terminal base pairs of the distal stem, a nucleotide between the catalytic triad and sugar-edge/Hoogsteen contact, and the dinucleotide bulge, which is a major $Mg^{2+}$-binding site in the *O. iheyensis* group IIC and aI5γ group IIB introns (Sigel *et al*., 2000; Toor *et al*., 2008a).

### 3.2.3 Characteristics of variants with increased retrohoming efficiency in the *mgtA* disruptant.

To determine which variants have increased retrohoming efficiency in the *mgtA* disruptant, we screened all 106 variants recovered from the selection by using a high-throughput version of the plasmid-based mobility assay (Figure 3.3*B*). This screen was done in 96-well deep culture plates and measures the number of $Tet^R$ cells via $O.D._{595}$ after two-days growth under selective conditions. We thus identified 30 candidates that had mobility efficiencies equal to or higher than the wild-type intron in the *mgtA*

disruptant (candidate map; Figures 3.4*B* and 3.5). We tested all 30 of these candidates by plating assays and found 20 DV variants (denoted "improved variants") that performed significantly (> 3-fold) better than the wild-type DV sequence in the *mgtA* disruptant, with one (DV20) reaching 22-fold enhancement (improved map; Figures 3.4*C* and 3.5).

The 30 candidate variants had relatively few mutations, with a maximum of 7 (Figure 3.5). Most mutations were found in the distal stem of DV, but a small number were found in the proximal stem outside the catalytic triad. Most of the mutant DVs retained a fully wild-type secondary structure, but eleven mutant DVs, with up to 6-fold increased retrohoming efficiency, had U-U or A-C mispairs in the distal stem (DV88, DV54, DV36, DV35, DV49, DV104, DV46, DV44, DV94, DV7, DV32; Figure 3.5). A number of the lower performing improved variants contained mutations in the dinucleotide bulge, which binds $Mg^{2+}$ (DV100, DV99, DV88, DV54; 1-2 fold increased). All 20 improved variants had mutations in the distal stem, and several had only a single mutation in the distal stem (DV49, DV32, DV59, DV14; 4-16 fold increased). Two of the improved variants also had altered tetraloops (GAAA; DV62 and DV37; 5-6-fold). The two most improved variants DV14 and DV20 (16- and 22-fold, respectively) had mutations confined to the distal stem (Figure 3.4*D* and *E*). Notably, both DV14 and DV20 have additional U-G or G-U wobble pairs in the distal stem, and the single base mutation in DV14 results in two tandem U-G pairs. G-U wobble pairs within RNA stems are frequently $Mg^{2+}$-binding sites due to the high electronegativity along their major groove surface, and two tandem G-U pairs has been shown to be a strong $Mg^{2+}$-binding site in RNA stems (Varani and McClain, 2000; Keel *et al*., 2007).

### 3.2.4 Saturation mutagenesis of mutable positions in DV.

We performed two more saturating selections focused on nucleotide residues in DV that were found to be mutable in the initial selection. As before, we used the plasmid-based mobility system to select variants of the Ll.LtrB-ΔORF intron that enhance retrohoming within the *mgtA* disruptant relative to the wild-type intron. For each saturating selection, we limited the number of randomized nucleotides to < 12 to ensure that all possible combinations of variants could be sampled in libraries of reasonable size.

Our first saturating selection focused on the mutable nucleotides found in the 20 improved variants from the initial selection (Figure 3.6). These 11 nucleotides included the G at the R position in the GNRA tetraloop, the three terminal base pairs of the distal stem, and two base pairs in the proximal stem between the catalytic triad and the sugar-edge/Hoogsteen contact (Figure 3.4*C*). After five rounds of selection, we identified 23 unique DV sequences, which were not found among the improved variants in the initial selection (Figure 3.6). All 23 of these variants had mutations in the distal stem of DV, and 18 had an additional U-G base pair in the proximal stem between the catalytic triad and the sugar-edge/Hoogsteen contact. When tested individually by using the plasmid-based plating assay, all 23 of the new variants outperformed the wild-type DV sequence by > 3-fold, with the three best (DV134, DV193 and DV164) having 15- to 16-fold higher retrohoming efficiency than the wild-type intron (Figure 3.6). However, these new variants performed no better than the two best variants, DV14 and DV20, from the initial selection.

The second saturating selection (Figure 3.7) was based on the sequence of the best performing variant from the initial selection, DV20 (22-fold increased), which contains three nucleotide changes in the distal stem of DV. We constructed a library of variants

that contained the three DV20 mutations and randomized eight additional nucleotides that were mutable in the initial selection (Figures 3.4 and 3.5). These included the dinucleotide bulge, a major $Mg^{2+}$-binding site, the first base pair of the λ' motif, and the same two base pairs between the catalytic triad and sugar-edge/Hoogsteen contact that were randomized in the first saturating selection. In the new selection, we compared the retrohoming efficiency of the library and pooled variants from three selection cycles to that of the DV20 variant by using the plasmid-based plating assay. The retrohoming efficiency of the initial pool of variants (cycle 0) was ~750-fold lower than DV20, but rapidly rose to roughly the same level as DV20 (~3% retrohoming efficiency) after selection cycle 1. The retrohoming efficiency of the pools from cycles 2 and 3 remained high but lower than that from cycle 1. Sequence analysis of 24 individual clones from each cycle showed that the DV20 sequence with no other changes increased in abundance to 7% in cycle 1, 25% in cycle 2, and 50% in cycle 3. The remaining 50% at cycle 3 comprised many lower frequency variants, the most abundant of which (7%) was DV176, which was also identified in the first saturating selection (Figure 3.6). The progressive enrichment of the DV20 sequence suggests that it cannot be improved further by changes in the other regions randomized in this selection. Notably, the dinucleotide bulge reverted to wild-type AC sequence in cycle 1 and the λ' base pair reverted to wild-type A-U by cycle 2, indicating that variants with mutations at these positions function suboptimally.

### 3.2.5 Rational Design

We also tried combining the mutations in the distal stems of the two best performing variants DV14 and DV20, but found that the combined variant had lower retrohoming efficiency (2.2-fold wild type) than either single variant (16- and 22-fold

wild type for DV14 and DV20, respectively). Finally, we substituted the DV distal stem from the *Pylaiella littoralis* LSU/2, which self-splices at unusually low magnesium concentrations (Costa *et al*., 1997), and found the retrohoming efficiency to be only 1.3-fold that of the wild-type intron (Figure 3.4*F*).

We also attempted to rationally engineer DV for enhanced retrohoming activity, however, none of the constructed variants performed better than wild-type (Figure 3.8*A*). The four mutable nucleotides of the proximal stem were grouped into four families (A, B, C, D) which covaried with the six nucleotides in the distal stem. The combinations attempted did not perform better than wild-type, and we also did not find any general effect by altering the four nucleotides in the proximal stem. A combination of the high activity single mutation found in variant DV14 with the DV20 distal stem decreased the retrohoming efficiencies in the *mgtA* disruptant to near wild-type DV levels.

The notable enrichment for tandem G-U pairs in the distal stem suggested that by rationally placing G-U pairs into DV, we may also further enhance $Mg^{2+}$-binding and retrohoming. The binding affinities of cobalt (III) hexammine to various G-U pairs within the stem of an RNA hairpin resembling that of DV have been determined by X-ray crystallography, and these data can be used to infer possible $Mg^{2+}$-binding affinity (Varani and McClain, 2000; Keel *et al*., 2007). In this particular study, a sequence containing G-U pairs capped with a GNRA tetraloop resembling the distal stem of DV14 bound cobalt at relatively high affinity, however, two sequences ranked higher (see Figure 3.8*B*). We generated these two DV mutants that contained these G-U mutations as found in (Keel *et al*., 2007) and tested these variants for retrohoming in the *mgtA* disruptants (Figure 3.8*B*). Surprisingly, these higher cobalt binding sequences reduced

retrohoming to below wild-type levels. These results suggest that the selected variants in DV perform the best in the MgtA mutants.

### 3.2.6 Northern hybridization of wild-type and variant Ll.LtrB RNAs *in vivo*.

To investigate how the DV mutations increase retrohoming efficiency, we focused on the two best performing variants from the selections, DV14 and DV20, which have retrohoming efficiencies 16- and 22-fold higher, respectively, than that of the wild-type intron in the *mgtA* disruptant (Figures 3.4 and 3.5). To determine whether the changes in retrohoming efficiency are due to changes in the intracellular levels of the Ll.LtrB intron RNA, we carried out Northern hybridizations of total cellular RNAs run in a 1% agarose gel and hybridized with a $^{32}$P-labeled intron probe (Figure 3.9). The cells were grown under the conditions of the retrohoming assay so that the intron RNA levels could be correlated with the retrohoming efficiencies.

For both the wild-type and mutant introns, the major band detected in the Northern blots corresponds to the excised intron RNA and co-migrates with lariat RNA obtained by self-splicing of the Ll.LtrB-ΔORF intron *in vitro*, consistent with previous findings for the wild-type intron that excised intron lariat RNA accumulates *in vivo* (Matsuura *et al*., 1997). The Northern blots for the wild-type intron show similar levels of the excised intron RNA in both wild-type HMS174(DE3) and *mgtA* mutant cells, indicating that the 200-fold decrease in retrohoming efficiency in the mutant cells is due instead to decreased activity of group II intron RNPs, as expected for lower intracellular $Mg^{2+}$ concentrations. The Northern blots for the DV14 and DV20 variants show that the levels of excised intron RNA in the *mgtA* disruptant are similar to or slightly lower than that of the wild-type intron, indicating that their enhanced retrohoming in the *mgtA*

disruptant is due to higher activity of the intron RNPs at lower $Mg^{2+}$ concentrations. Because the LtrA protein does not bind to DV (Dai *et al*., 2008), there is no expectation that mutations in DV could increase RNP activity by enhancing protein binding at low $Mg^{2+}$ concentrations.

### 3.2.7 Splicing of wild-type and variant intron RNAs at different $Mg^{2+}$ concentrations.

To investigate whether the DV variants have enhanced group II intron ribozyme activity at low $Mg^{2+}$ concentrations, we compared the *in vitro* splicing of the wild-type Ll.LtrB-ΔORF, DV14, and DV20 introns at three different $Mg^{2+}$ concentrations. In these experiments, a $^{32}$P-labeled *in vitro* transcript containing the wild-type or mutant Ll.LtrB-ΔORF intron flanked by short exon sequences was spliced by adding a 10-fold molar excess of LtrA protein. Plots of precursor RNA disappearance and lariat RNA accumulation are shown in Figure 3.10, and the corresponding gels are shown in Figure 3.11.

At 5 mM $Mg^{2+}$, the splicing reaction of the wild-type intron proceeded in multiple phases. A prominent rapid phase gave a rate constant $k_1$ of 1.9 $min^{-1}$ for the disappearance of precursor RNA and a slower phase gave $k_2$ of 0.074 $min^{-1}$, in agreement with previous results (Saldanha *et al*., 1999). The slow phase most likely reflects a population of intron RNAs that folds and assembles with LtrA more slowly, as suggested previously (Saldanha *et al*., 1999). An alternative model postulates that the two phases reflect a rapid internal equilibration of the chemical steps of splicing, followed by a slower, irreversible step. However, we do not favor this model because inefficient folding has been demonstrated for other group II intron RNAs under similar conditions (Russell *et al*., 2013) and because of additional results at lower $Mg^{2+}$ concentrations (see below).

Together these two phases accounted for splicing of 70-80% of the precursor in different assays. The splicing reactions of DV14 and DV20 were also multiphasic, but with smaller amplitudes for the fast phase, resulting in less efficient splicing of the mutant introns at 5 mM $Mg^{2+}$. The decreased amplitudes for the fast phase may reflect that the variants were selected to function optimally at low $Mg^{2+}$ concentrations and now function suboptimally at higher $Mg^{2+}$.

The situation differs at lower $Mg^{2+}$ concentrations. At 2.5 mM $Mg^{2+}$, the splicing of all three introns showed two phases on the time scale of the experiment. For all three introns, the fast phases again gave rate constants $k_1$ of ~1 $min^{-1}$, but with amplitudes that were lower than those at 5 mM $Mg^{2+}$. In addition, substantial fractions of the precursor remained unspliced even after completion of the slower observed phase, suggesting the presence of at least one additional population that folds even more slowly. At 1.5 mM $Mg^{2+}$, the splicing of all three introns was slower and only a single phase with a small amplitude (≤ 5%) was observed on the experimental time scale. The lower amplitudes at 2.5 and 1.5 mM $Mg^{2+}$ provide additional evidence against the alternative model in which the fast phase reflects internal equilibration of the chemical steps, as there is no expectation that lower $Mg^{2+}$ concentrations would favor the precursor or that the equilibrium would be so far from unity ($K_{eq}$ < 0.05 at 1.5 mM $Mg^{2+}$). The decreased rate constant at 1.5 mM $Mg^{2+}$ did not result from incomplete binding of the LtrA protein, as increasing the protein concentration did not increase the rates or extents of splicing of the wild-type or mutant introns (Figure 3.12), but could reflect greater difficulty in folding into the active RNA structure or a decreased rate for the catalytic steps at low $Mg^{2+}$ concentration. Importantly, the amplitudes were larger for the mutants than for the wild-

type intron, such that the mutants splice more efficiently than the wild-type intron at low $Mg^{2+}$ concentrations. The DV20 mutant gave more splicing than DV14, mirroring the relative activity of the two variants *in vivo*.

Together, these results suggest a model in which the initial fast phase of splicing observed at 5 and 2.5 mM $Mg^{2+}$ reflects a population of precursor RNAs that can fold and assemble with LtrA protein rapidly to form an active RNA structure. The slower phases at these $Mg^{2+}$ concentrations most likely reflect populations that fold and assemble with LtrA more slowly. At 1.5 mM $Mg^{2+}$, an initial fast phase was not observed, either because the fraction of the RNA that folds rapidly is too small or because the catalytic steps of splicing are slower at this $Mg^{2+}$ concentration. For all three introns, the fraction of reactive RNPs decreases at lower $Mg^{2+}$ concentrations, but to a lesser extent for DV14 and DV20 than for the wild-type intron, enabling the mutant introns to splice more efficiently than the wild-type intron at 1.5 mM $Mg^{2+}$.

### 3.2.8 Reverse splicing of wild-type and variant intron RNAs at different $Mg^{2+}$ concentrations.

We carried out similar experiments to assess the effect of the DV14 and DV20 mutations on reverse splicing of the intron RNA during target DNA-primed reverse transcription reactions. In these experiments, a 129-bp $^{32}P$-labeled DNA substrate containing the Ll.LtrB intron-insertion site was incubated with a 50-fold molar excess of wild-type and variant Ll.LtrB RNPs (wild-type LtrA protein reconstituted with wild-type or variant Ll.LtrB-ΔORF intron lariat RNA) in reaction medium containing different $Mg^{2+}$ concentrations and dNTPs to enable target DNA-primed reverse transcription of the reverse spliced intron RNA, as would occur *in vivo*. Reverse splicing of the intron lariat into the DNA target sites occurs in two steps: ligation of the 3' end of the lariat RNA to

the 5' end of the downstream exon, referred to as partial reverse splicing, followed by ligation of the 5' end of the lariat RNA to the 3' end of the upstream exon, referred to as full reverse splicing (schematic Figure 3.14). Plots showing time courses for net completion of the first step (partial plus full reverse splicing) and full reverse splicing (insertion of the intron RNA between the 5' and 3' DNA exons) are shown in Figure 3.13, and gels are shown in Figure 3.14.

At 5 mM $Mg^{2+}$, the reactions of wild-type RNPs were monophasic with a *k* of 0.068 $min^{-1}$ for net completion of the first step and 0.030 $min^{-1}$ for full reverse splicing, with the fully reverse spliced product reaching 34% at the end of the time course. The DV14 and DV20 RNPs show similar kinetics with little or no decrease in amplitude and reaction rate. In this case, the reaction starts with RNPs containing lariat RNAs that have already folded into the active RNA structure. In addition, because the reaction is performed with excess RNPs, a fraction of inactive or misfolded RNPs would not necessarily inhibit the reverse splicing reaction, provided that these molecules do not compete with active RNPs for binding to the target DNA. Either or both of these factors may underlie the finding that the mutations do not impair reverse splicing at high magnesium concentration, as they do for RNA splicing.

At 2.5 and 2 mM $Mg^{2+}$, the reaction rate for all three introns decreased 10- to 20-fold, and the amplitudes decreased progressively, indicating a smaller proportion of reactive molecules. For all three introns at low $Mg^{2+}$ concentrations, the rate constants for net completion of the first step were similar to those for full reverse splicing, indicating that processes leading up to or culminating with the first step of reverse splicing are rate limiting. The population then presumably undergoes the first and second steps reversibly,

leading to steady state populations of the partially and fully reverse-spliced products until the reaction is rendered irreversible by cDNA synthesis. Thus, the decreased rates at low $Mg^{2+}$ concentrations may reflect difficulties with a required RNA conformational change or catalytic step, while the decreased amplitudes may reflect a tendency of the intron lariat RNAs or RNPs to revert to inactive conformations. As for RNA splicing, the decrease in the proportion of reactive RNPs at low $Mg^{2+}$ concentrations was less for the mutants than for the wild-type intron, enabling the mutants to function more efficiently at low $Mg^{2+}$ concentration. DV20 again outperformed DV14 at low $Mg^{2+}$ concentrations, in agreement with the relative activity of the two variants *in vivo*.

### 3.2.9 Terbium-cleavage assays.

To investigate how mutations in the distal stem affect $Mg^{2+}$ binding and $Mg^{2+}$-dependent structure formation within DV, we performed terbium ($Tb^{3+}$)-cleavage assays on isolated wild-type and variant DVs. By efficiently deprotonating water and 2'-OH groups in RNA, $Tb^{3+}$ enhances cleavage of RNA's phosphodiester backbone and is therefore useful as a probe of RNA structure (Forconi and Herschlag, 2009). Additionally, because $Tb^{3+}$ is close in size to $Mg^{2+}$ and coordinates similarly to oxygen ligands, it can localize at positions that replace bound $Mg^{2+}$ in folded RNAs, thereby mapping potential $Mg^{2+}$-binding sites (Sigel, 2005). $Tb^{3+}$-cleavage of isolated DV of the yeast aI5γ intron identified a high affinity $Mg^{2+}$-binding site at the dinucleotide bulge and lower affinity sites within the distal helix and tetraloop (Sigel *et al*., 2000), which were later supported by the *O. iheyensis* intron crystal structure (Toor *et al*., 2008a).

Results of $Tb^{3+}$-cleavage titration assays for wild-type and variant Ll.LtrB DV RNAs in reaction media containing 50 mM KCl and 1.5 mM or 5 mM $MgCl_2$ are shown

in Figure 3.15 and Figure 3.16, respectively. The RNA was probed at $Tb^{3+}$ concentrations ranging from 10 μM to 10 mM, with cleavages displaying saturation at lower $Tb^{3+}$ concentrations indicating higher affinity binding (Sigel, 2005). Consistent with the *O. iheyensis* crystal structure and aI5γ intron $Tb^{3+}$ cleavages (Sigel *et al*., 2000; Toor *et al*., 2008a), we observed saturated cleavage at low $Tb^{3+}$ concentrations at the dinucleotide bulge in the wild-type Ll.LtrB DV RNA, primarily 3' to the A-residue and to a lesser extent 3' of the C-residue, indicating high affinity binding. Moderate cleavages, two- to three-fold higher than at weaker sites at some $Tb^{3+}$ concentrations, were observed after the C-residue preceding the tetraloop and a G-residue in the proximal stem, while weaker cleavages occur throughout the distal stem and tetraloop. The cleavages decrease at high concentrations of $Tb^{3+}$, presumably reflecting misfolding due to replacement of $Mg^{2+}$ by $Tb^{3+}$ throughout the RNA (Harris *et al*., 2004). At 5 mM $Mg^{2+}$, several sites show two-fold higher cleavage than at 1.5 mM $Mg^{2+}$, including the C-residue proximal to the tetraloop, all four residues of the tetraloop, the catalytic triad, and the sugar-edge/Hoogsteen contact, while cleavages at other sites have similar intensity to those at 1.5 mM $Mg^{2+}$ (Figure 3.16).

The DV14 and DV20 variants show enhanced cleavages at 1.5 mM $Mg^{2+}$ at sites in the distal stem, in the tetraloop, and at some sites in the proximal stem (Figure 3.15). At low concentrations of $Tb^{3+}$, DV14 shows dramatically increased cleavage at the C-residue preceding the tetraloop and a G-residue in the proximal stem. These cleavages appear to saturate at lower $Tb^{3+}$ concentrations in the mutant, suggesting increased metal-ion affinity. DV20 shows moderately enhanced cleavages at the C-residue preceding the tetraloop and three other sites in the distal stem, but without apparent changes in affinity.

Additional weaker cleavages throughout the distal stem and the tetraloop are enhanced in the mutants relative to the wild-type DV by up to 2-fold. Notably, in reaction medium containing 5 mM $Mg^{2+}$, the cleavage intensities relative to input RNA for the variants remain similar to those at 1.5 mM $Mg^{2+}$, suggesting that variant DV stems fold similarly at both $Mg^{2+}$ concentrations (Figure 3.16). In general, we find that the variants enhance $Tb^{3+}$ cleavage relative to input RNA throughout the distal stem at 1.5 mM $Mg^{2+}$ to levels found for the wild-type DV at 5 mM $Mg^{2+}$. Considered together with the results of the RNA splicing and reverse splicing assays and the appearance of G-U base pairs in the distal stem of DV in the selected introns, the $Tb^{3+}$-cleavage data suggest a model in which enhanced metal ion binding to the distal stem of the variants promotes a conformational transition to a more native structure, which is inefficient for the wild-type DV at low $Mg^{2+}$ concentrations.

### 3.3 Discussion

Here, we used an *E. coli* mutant lacking the $Mg^{2+}$ transporter MgtA to select variants of the Ll.LtrB-ΔORF group II intron with mutations in DV that enable more efficient retrohoming at low $Mg^{2+}$ concentrations. We find that such mutations are clustered in the distal stem of DV. The two most efficient variants, DV14 and DV20, have mutations restricted to the distal stem and increase retrohoming efficiencies in the *mgtA* mutant by 16- and 22-fold, respectively, without substantially affecting intron RNA levels *in vivo*. Biochemical analysis of these variants leads to a model in which the mutations enhance $Mg^{2+}$ binding to the distal stem of DV and this enhanced $Mg^{2+}$ binding facilitates an RNA-folding step, resulting in increased efficiencies of both RNA splicing

and reverse splicing at low $Mg^{2+}$ concentrations *in vitro*. Our findings reveal that DV is involved in a critical $Mg^{2+}$-dependent folding step that contributes to the unusually high $Mg^{2+}$ concentrations required for group II intron function, and they have implications for the evolution of splicing mechanisms in higher organisms and the use of group II introns for gene targeting.

*E. coli* and related bacteria possess two major $Mg^{2+}$-transporters: CorA, which maintains overall $Mg^{2+}$ levels, and MgtA, which has been shown to function as a $Mg^{2+}$ scavenging system at low $Mg^{2+}$ concentrations in *Salmonella typhimurium* (Snavely *et al*., 1991). In setting up our selection system, we found that the retrohoming efficiency of the wild-type Ll.LtrB-ΔORF intron in the *mgtA* disruptant was 100-fold lower than in the *corA* disruptant, even though both mutants appear to have similarly decreased intracellular $Mg^{2+}$ concentrations, using a fluorescent probe assay (Figure 3.2). This finding indicates that the fluorescent probe does not accurately measure the functional $Mg^{2+}$ concentration for group II intron retrohoming *in vivo*, either because of the non-growth conditions required for uptake of the probe or because of localized differences in $Mg^{2+}$ concentration in different regions of the cell. In *E. coli*, Ll.LtrB RNPs localize to the cellular poles, along with plasmid DNAs and chromosomal DNA replication origins and termination sites, which are favored regions for Ll.LtrB retrohoming and retrotransposition (Ichiyanagi *et al*., 2003; Zhao and Lambowitz, 2005; Beauregard *et al*., 2006; Yao *et al*., 2007). Notably, although CorA is uniformly distributed in the membrane throughout the cell, MgtA is localized to the cell membrane near the cellular poles, where it could serve to boost local $Mg^{2+}$ concentrations that support retrohoming (Genobase of the Nara Institute of Science and Technology online resource;

http://ecoli.naist.jp/data/GeneProfiles/GFPimages/34-08-2.jpg) (Gray *et al*., 2011). To our knowledge, localized differences in intracellular $Mg^{2+}$ concentration affecting biological processes have not been reported previously, but could have wide implications.

Our primary selection in the *mgtA* disruptant used a library of Ll.LtrB-ΔORF introns in which all residues in DV were partially randomized with 70% of the wild-type nucleotide and 10% of each mutant nucleotide at each position. In agreement with previous mutational analyses of DV of the yeast aI5γ group II intron (Boulanger *et al*., 1995), the most strongly conserved regions were the catalytic triad, the sugar-edge/Hoogsteen contact, and the κ' and μ' tertiary contacts (invariant or ≤ 2 mutations in 106 active variants). More variation (≥ 3 mutations) was found in three other functionally important regions: the dinucleotide bulge, which is a high-affinity $Mg^{2+}$-binding site; the GNRA tetraloop, which interacts with a tetraloop acceptor in DI (Costa and Michel, 1995); and the λ' motif (Boudvillain *et al*., 2000). However, the variants in these regions functioned suboptimally and reverted to the wild-type sequences in secondary selections (Figures 3.6 and 3.7). The latter finding agrees with previous studies showing that mutations in the dinucleotide bulge of the yeast aI5γ group II intron decrease the efficiency of self-splicing (Schmidt *et al*., 1996). The most mutable regions of DV in our selection were a nucleotide between the catalytic triad and sugar-edge/Hoogsteen contact and the three terminal base pairs of the distal stem between the λ motif and the tetraloop. The distal stem is the most variable region of DV of naturally occurring group II introns and is interchangeable among different group II introns (Boulanger *et al*., 1995).

Strikingly, although the upper region of the distal stem of DV was thought to play no major functional role, we found that it is the major site of mutations that improved

function of the Ll.LtrB intron at low $Mg^{2+}$ concentrations. All 43 improved variants with retrohoming efficiencies > 3-fold higher than that of the wild-type intron in the *mgtA* disruptant had mutations in this part of the distal stem, including 15 variants that contained only mutations in this region. By contrast, we identified only two variants (DV1 and DV103) that enhance retrohoming at low $Mg^{2+}$ concentrations without mutations in the distal stem (Figure 3.5). Both of these variants have the same A → G mutation, which generates a G-U base pair in the proximal stem and increases retrohoming efficiency in the *mgtA* mutant two- to three-fold. The distal stem mutations that improved function at low $Mg^{2+}$ concentrations trend towards weaker base pairings, G-C pairs changed to G-U or A-U pairs, or mispairings, which may enhance dynamics and facilitate bending of DV (Figures 3.4, 3.5, and 3.6). G-U wobble pairs within RNA stems are frequent $Mg^{2+}$-binding sites due to the high electronegativity along their major groove surface, with the flanking base pairs influencing the affinity and position of the bound metal ion (Varani and McClain, 2000; Keel *et al*., 2007). The two most efficient variants, DV14 and DV20, both have additional U-G or G-U wobble pairs in the distal stem, and the single base mutation in DV14 results in two tandem U-G base pairs, a motif identified as a strong $Mg^{2+}$-binding site in RNA stems (Varani and McClain, 2000; Keel *et al*., 2007). The strong clustering of mutations with improved function at low $Mg^{2+}$ in the distal stem may reflect that it is one of the few regions of DV that can be mutated without impairing other functions and/or that $Mg^{2+}$-binding sites in the distal stem of DV are critical for a key RNA folding step.

Biochemical assays of the two most improved variants DV14 and DV20 shows that the distal stem mutants have higher splicing and reverse splicing activity than the

wild-type intron at low $Mg^{2+}$ concentrations, reflecting a greater propensity to fold into the catalytically active RNA structure. The introduction of a sharp bend in DV that is required to form the active site is likely an energetically unfavorable conformational change and is an attractive candidate for a folding step that limits the efficiency of native state formation at low $Mg^{2+}$ ion concentrations, leading to the relatively high $Mg^{2+}$ requirement of group II introns compared to other ribozymes. For DV RNA in isolation or in early folding intermediates of the intron, the conformational change is likely to be rapid and reversible. However, upon global collapse of the intron and encapsulation by DI, the transition may become very slow and divide the RNA into active and inactive populations, which give rise to the multiple phases observed in group II intron splicing reactions at low $Mg^{2+}$ concentrations (Russell *et al*., 2013). By stabilizing the native, bent conformation of DV, the mutations could increase the fraction of the RNA that readily folds to the native state and gives activity. The finding that the mutations also enhance reverse splicing activity of the conformationally constrained intron lariat RNA is consistent with the possibility that only a limited conformational change involving DV or its activation by $Mg^{2+}$ binding may be rate-limiting at low $Mg^{2+}$ concentrations.

To understand how mutations in the distal stem of DV might contribute to native folding of DV, we modeled the distal stem of DV in the Ll.LtrB intron in the extended state based on the NMR structure of the *Pylaiella littoralis* LSU/2 intron (Seetharaman *et al*., 2006) and in the folded state based on the crystal structure of the *O. iheyensis* group II intron (Marcia and Pyle, 2012) (Figure 3.17). The models show that some $Tb^{3+}$ cleavages in the distal stem of DV of Ll.LtrB that are enhanced in variants DV14 and DV20 correlate with nearby metal ion-binding sites seen in the *O. iheyensis* structures,

including at the positions of M3 and M4 near the GNRA tetraloop and M5 near the site of the bend between the proximal and distal stem. Other cleavages that are enhanced in the mutants, such as those within the 5' strand of the distal stem, are not near metal ion-binding sites seen in the *O. iheyensis* structures, but are near U-G or G-U base pairs introduced by the mutations, suggesting an origin for the increased cleavages (*e.g.* the hypothetical $Mg^{2+}$ represented as a blue sphere with a "?" bound in the major groove in Figure 3.17). Although the localization of $Tb^{3+}$ in the major groove of a G-U pair is likely to restrict its access to adjacent 2'-OH groups and therefore limit RNA cleavage at these sites (Keel *et al.*, 2007), it is possible that $Tb^{3+}$ can be sufficiently dynamic at G-U pairs to retain significant cleavage activity. Alternatively or in addition, the metal-bound G-U pairs in the distal stem may increase RNA flexibility in the DV mutants, promoting formation of a bent conformation and leading to enhanced cleavage at multiple sites. In the intact intron, these site-bound metal ions may stabilize a bent and functional conformation of DV directly and/or indirectly by stabilizing tertiary contacts in the folded intron structure that enforce the bend. Delocalized metal ions may also contribute to the stability of the bent conformation of DV because it is more compact than the extended form (Draper, 2004).

Group II introns are evolutionarily related to nuclear spliceosomal introns of eukaryotes (Keating *et al.*, 2010; Lambowitz and Zimmerly, 2011; Rogozin *et al.*, 2012). According to one scenario, group II introns entered eukaryotes with bacterial endosymbionts that gave rise to mitochondria and chloroplasts and invaded the nucleus, where they or their descendants proliferated before degenerating into spliceosomal introns and the spliceosome. The nuclear membrane has been hypothesized to have

evolved in response to this group II intron invasion as a means of separating transcription from translation, thereby preventing translation of unspliced introns (Martin and Koonin, 2006). A further consequence of the nuclear membrane was sequestration of the genome into a separate compartment, where $Mg^{2+}$ is chelated by chromosomal DNA (Strick *et al*., 2001), possibly leading to even lower free $Mg^{2+}$ concentrations than in the cytoplasm. Indeed, such differential $Mg^{2+}$ concentrations between the nucleus and cytoplasm may explain recent findings that an Ll.LtrB-ΔORF intron cannot be spliced by the LtrA protein in the yeast nucleus, but can be spliced by the LtrA protein after export of precursor RNAs to the cytoplasm (Chalamcharla *et al*., 2010). Spliceosomal introns have evolved to function at lower $Mg^{2+}$ concentrations in eukaryotes, perhaps reflecting their disintegration into snRNAs, which may facilitate conformational changes, and their increased reliance on protein cofactors, which can substitute for $Mg^{2+}$ to promote RNA folding. However, the lower $Mg^{2+}$ concentrations now constitute a natural barrier to further group II intron proliferation in eukaryotic nuclear genomes. The ability to select group II introns that function at lower $Mg^{2+}$ concentrations may ultimately enhance their utility in gene targeting in higher organisms, although selections done directly in eukaryotes using libraries with mutations throughout the intron may be required to achieve maximal retrohoming efficiency.

## 3.4 Methods

### 3.4.1 *E. coli* strains and growth conditions.

The wild-type *E. coli* strain used for retrohoming assays was HMS174 (λDE3) (Novagen). The *corA* and *mgtA* disruptants in HMS174 (λDE3) were obtained by

targetron mutagenesis and have the genotypes *corA:Ll.LtrB*$^{640a}$ and *mgtA::Ll.LtrB*$^{2466a}$ (Yao *et al*., 2005). *E. coli* MegaX DH10b electrocompetent cells (Invitrogen) were used for library construction, and Artic Express (Stratagene) was used for LtrA protein expression. Chemically competent cells were generated by using the Inoue method (Sambrook and Russell, 2006b), and highly electrocompetent *mgtA* cells for library selections were generated by combining two previous procedures [see below; (Chuang *et al*., 1995; Shi *et al*., 2003)]. Antibiotics were added at the following concentrations: ampicillin, 50 μg/ml; chloramphenicol, 25 μg/ml; tetracycline, 25 μg/ml; gentamicin, 10 μg/ml.

### 3.4.2 Generation of highly electrocompetent *mgtA* cells for library transformations.

We obtained highly electrocompetent *mgtA* disruptant cells (5 x $10^7$ transformants/μg DNA) by using potassium acetate supplemented medium and low growth temperatures (Chuang *et al*., 1995; Shi *et al*., 2003). A fresh colony growing on an LB agar plate was inoculated into 5 ml of Luria-Bertani (LB) medium and grown at 37°C for 2 h, then 1 ml was inoculated into 50 ml of KOB-25 (SOB + 25 g/l potassium acetate) and grown at 25°C overnight. The following day, the cells were subcultured at 1:100 dilution into 1 l KOB-35 (SOB + 35 g/l potassium acetate) and grown at 18°C until $O.D._{595}$ = 0.35. The cells were then cooled on ice for 20 min, centrifuged at 1,000 x g for 20 min at 4°C, and sequentially washed by centrifugation in 200 mM HEPES (pH 7.5), filtered water, and filtered 10% glycerol. Finally, the cells were concentrated to $O.D._{595}$ = 100 in 10% glycerol and 150 mM trehalose (Sigma) and flash frozen in liquid nitrogen.

### 3.4.3 Recombinant plasmids.

The intron-donor plasmid pACD2 used for retrohoming assays carries a *cap*$^R$ marker and contains a 0.94-kb Ll.LtrB-ΔORF intron with a phage T7 promoter sequence inserted within DIV; an ORF encoding the IEP, denoted LtrA protein, is cloned downstream of the intron and expressed in tandem (Karberg *et al*., 2001). pACD2-ΔDV, which was used as the backbone for intron library construction, contains a "stuffer" fragment in place of DV in order to prevent carryover of incompletely digested wild-type introns into the library. It was derived from pACD2 by deleting and replacing DV (intron positions 844-879) with the stuffer fragment and replacing a SalI site in DIV with an ApaII site (Figure 3.1*A*). The intron-recipient plasmid pBRR3-ltrB carries an *amp*$^R$ marker and contains a promoterless *tet*$^R$ gene cloned downstream of a wild-type Ll.LtrB target site (Karberg *et al*., 2001). Ll.LtrB-ΔORF introns with mutations in DV were generated by site-directed mutagenesis of pACD2.

### 3.4.4 Conventional and high-throughput plasmid-based retrohoming assays.

*E. coli* plasmid-based intron retrohoming assays were performed as described (Guo *et al*., 2000; Karberg *et al*., 2001). The Cap$^R$ intron-donor plasmid pACD2 and Amp$^R$ recipient plasmid pBRR3-ltrB were co-transformed into wild-type HMS174(DE3) or the *corA* and *mgtA* disruptants, and transformants were grown overnight at 37°C to stationary phase. The cells were then subcultured to $O.D._{595}$ = 0.3, diluted 50-fold, and induced with 0.1 mM isopropyl β-D-1 thiogalactopyranoside (IPTG) for 1 h at 37°C. Retrohoming of the intron from the donor to the recipient plasmid introduces a T7 promoter upstream of the *tet*$^R$ gene and enables selection for Tet$^R$ colonies. Cells were plated on LB medium containing tetracycline plus ampicillin or ampicillin alone, and

retrohoming efficiencies were quantified as the ratio of ($Tet^R$ + $Amp^R$)/$Amp^R$ colonies. All assays used fresh tetracycline + ampicillin plates made 24 h prior to use.

A high-throughput version of the retrohoming assay in 96-deep-well plates (2-ml wells) used selection for tetracycline plus ampicillin or ampicillin during growth in LB medium with the plates shaken at 250 rpm at 37°C. For these 96-well plate assays, overnight 1-ml cultures of the *mgtA* disruptant transformed with the donor and recipient plasmids were subcultured 1:100 into 1 ml of LB, grown to $O.D._{595}$ = 0.3, diluted 50-fold into 1 ml of LB containing 0.1 mM IPTG, induced for 1 h at 37°C, and then subcultured 1:50 into 1 ml of LB containing either tetracycline plus ampicillin or ampicillin alone. Growth was measured at 24-h intervals for cells in the presence of tetracycline plus ampicillin and once at 16-h for cells in the presence of ampicillin by transferring 100-µl portions of the culture into 96-well BD Optilux plates and determining $O.D._{595}$ by using a Spectramax M3 plate reader (Molecular Devices). Retrohoming efficiencies were quantified as the ratio of $O.D._{595}$ in the presence of tetracycline + ampicillin to that in ampicillin alone.

**3.4.5 DV mutant libraries and selections.**

Donor plasmid libraries containing mutations in DV were constructed in pACD2-ΔDV by replacing the ApaLI + KpnI fragment containing the "stuffer" substituted for DV (see above) with the corresponding 143-nt segment of the intron containing DV with doped or randomized nucleotides at different positions. The mutagenized DV segment was generated by overlap extension PCR using complementary overlapping oligonucleotides with the desired mutations and Vent DNA polymerase (New England Biolabs). Doped and randomized oligonucleotides were purchased from Integrated DNA

Technologies and confirmed by sequencing individual library clones to have the expected nucleotide frequencies.

For the initial selection using a library of Ll.LtrB-ΔORF introns in which DV was partially randomized ("doped") (Figures 3.4, 3.5, and 3.3), six separate ligation reactions were electroporated into *E. coli* MegaX DH10b electrocompetent cells (Invitrogen). Each electroporation used 1 μg of ligation product after phenol/chloroform extraction and ethanol precipitation in the presence of 0.3 M sodium acetate and linear acrylamide carrier. Transformants were selected by growth in LB containing chloramphenicol. The plasmid library DNAs ($10^9$ variants) was then extracted by an alkaline lysis procedure (Birnboim and Doly, 1979), purified using a midiprep column (Qiagen), and electroporated into *mgtA* disruptant cells that had been pre-transformed with the recipient plasmid for use in the plasmid-targeting assay. Electroporation was done at 1.8 kV using a series of aliquots containing 3 μg of plasmid per 50 μl of electrocompetent cells, yielding 5 x $10^7$ colonies per electroporation, until a library of $10^9$ variants was obtained.

For each selection cycle, cells were plated on fresh LB plates containing ampicillin and tetracycline, and grown for four days. $Tet^R$ colonies were then scraped and pooled to isolate plasmid-products by alkaline lysis, and the region of the intron containing DV was amplified by PCR of the retrohoming products with an upstream primer (p60S 5'-CGTCCAGATATTTATTACGTGGCGACG) in DIV and a downstream primer (p73A 5'-AATGGACGATATCCCGCA) in the 3' exon using Phusion Polymerase (New England Biolabs). This PCR product was digested with MluI and KpnI to generate a 238-nt fragment that was swapped for the corresponding fragment in pACD2-ΔDV to create a library for the next round of selection. To identify individual variants, $Tet^R$

colonies were picked from selection plates and subjected to colony PCR using the DIV and 3'-exon primers described above. The PCR product was then sequenced at the UT Austin Core facility using a primer (p63A 5'-CGTAGAATTAAAAATGATATGGTGAAGTAGGG) that extends across the 3'-integration junction. Finally, unique variants were isolated by nested PCR (primers p60S + p63A) and recloned into the donor plasmid vector to determine retrohoming efficiency.

### 3.4.6 Determination of intracellular free [$Mg^{2+}$].

The intracellular free magnesium concentration ($[Mg^{2+}]_i$) in wild-type HMS174 (λDE3) and the *corA* and *mgtA* disruptants was determined by using the fluorescent $Mg^{2+}$ probe mag-fura-2-AM (acetomethoxyl ester form; Molecular Probes), essentially as described (Froschauer *et al*., 2004). Stocks of mag-fura-2-AM were kept at 1 mM in dimethyl sulfoxide and then dispersed 1:1 in 10% (w/v) pluronic F-127 dispersant (Life Technologies) before use. To load cells with the probe, the mag-fura-2/pluronic F-127 mixture was diluted 200-fold into a 1-ml solution of 0.9% NaCl, 10 mM HEPES (pH 7.5) containing ~2 x $10^9$ mid-log cells ($O.D._{595}$ 0.3-0.4) that had been washed twice with 0.9% NaCl. The cells were incubated in the solution for 60 min at 25°C and then washed three times with 0.9% NaCl. To measure the ratio of $Mg^{2+}$-bound to free mag-fura-2, cellular fluorescence at 509 nm was measured after excitation at 340 nm (bound) or 380 nm (unbound) in BD Optilux 96-well plates using a Spectramax M3 plate reader (Molecular Devices). The cells were then lysed by incubating with lysozyme (0.5 mg/ml for 10 min at room temperature; Sigma) followed by 1% SDS, and minimum and maximum ratios (R) of fluorescence at 340 nm (bound) to 380 nm (unbound) were determined after adding 1 mM EDTA ($R_{min}$) followed by 40 mM $Mg^{2+}$ ($R_{max}$). The $[Mg^{2+}]_i$ was determined

by fitting the 340-nm and 380-nm measurements (background fluorescence subtracted) to the following equation (Grynkiewicz *et al*., 1985; Poenie, 1990):

$$[Mg^{2+}]_i = K_d \times \frac{F_o}{F_s} \times \frac{R - R_{min} V_f}{R_{max} V_f - R}$$

where $K_d$ is the dissociation constant of mag-fura-2 for $Mg^{2+}$ (1.9 mM); $F_o/F_s$ is the ratio of the 380-nm fluorescence for the unbound probe (1 mM EDTA) to that for the saturated probe (40 mM $[Mg^{2+}]$); R is the ratio of the 340/380-nm intracellular fluorescence; $R_{min}$ and $R_{max}$ are the 340/380-nm fluorescence ratios of the lysed cell solutions at 1 mM EDTA and 40 mM $Mg^{2+}$, respectively (Grynkiewicz *et al*., 1985); and $V_f$ is a constant (0.85) that corrects for the reduction in fluorescence intensity due to intracellular viscosity (Poenie, 1990).

**3.4.7 Northern hybridization.**

*E. coli* transformed with the pACD2 intron-donor and pBRR-ltrB recipient plasmids were grown and induced with IPTG, as described above for intron mobility assays, and RNA was extracted by using a lysozyme/freeze-thaw and hot phenol method (Matsuura *et al*., 1997; Slagter-Jager *et al*., 2006). Northern hybridizations were performed as described (Sambrook and Russell, 2006a). RNA (1.5 μg) and a ssRNA ladder (New England Biolabs) were denatured with glyoxal and run in a 1% agarose gel in Bis-Tris/PIPES/EDTA (BPTE) running buffer. The gel was soaked in 0.05 M NaOH and blotted to a Hybond nylon membrane (Amersham) by capillary transfer in 20X SSC (3 M NaCl, 0.3 mM sodium citrate). The blot was then hybridized with a 5'-$^{32}$P-labeled DNA oligonucleotide probe for the Ll.LtrB intron (5'-CCGTGCTCTGTTCCCGTATCA) in RapidHyb buffer (Amersham) overnight at 45°C, and the membrane was washed and scanned with a Typhoon Trio PhosphorImager (Amersham).

### 3.4.8 Preparation of LtrA protein and Ll.LtrB RNPs.

The LtrA protein was expressed with an N-terminal intein-cleavable chitin-binding domain from plasmid pImp-1P (Saldanha *et al*., 1999) in *E. coli* Arctic Express cells (Stratagene) following the manufacturer's recommendations. Briefly, three to five freshly transformed colonies were grown to stationary phase overnight in LB medium containing ampicillin and gentamicin. The following day, the cells were subcultured at 1:100 dilution into LB medium without antibiotics for 3 h at 30°C, followed by induction with 1 mM IPTG for 24 h at 12°C and shaking at 250 rpm. The LtrA protein was then purified via chitin-affinity-column chromatography and intein cleavage, as described (Saldanha *et al*., 1999), except that prior to loading the column, nucleic acids in the cell supernatant were precipitated by incubating with 0.4% polyethylenimine (PEI) under constant stirring at 4°C for 1 h, followed by centrifugation at 16,000 x g for 30 min at 4°C. The supernatant was then concentrated ~4-fold by dialysis overnight against 0.25 M NaCl containing 50% glycerol. The protein was > 95% pure and stored at ~3 mg/mL.

Intron RNAs for reconstituting Ll.LtrB RNPs were transcribed from DNA templates generated by PCR of pACD2 or derivatives containing intron DV mutants using primers that append a phage T3 promoter sequence and yield a precursor RNA containing the Ll.LtrB-ΔORF intron with 5' and 3' exons of 19 and 21 nt, respectively. Transcription was performed overnight using a T3 Megascript kit (Ambion), and the precursor RNA was self-spliced *in vitro* to generate excised intron lariat RNA, as described (Saldanha *et al*., 1999). Ll.LtrB RNPs were reconstituted as described (Mastroianni *et al*., 2008) by incubating 40 nM self-spliced RNA with 160 nM LtrA protein in 450 mM NaCl, 5 mM $MgCl_2$, and 40 mM Tris-HCl (pH 7.5) at 30°C for 30 min, and then concentrating RNPs by ultracentrifugation at 40,000 x g at 4°C overnight.

The RNP pellet was dissolved in 10 mM KCl, 10 mM $MgCl_2$, and 40 mM HEPES (pH 8.0) at a concentration of 2 μg/μl based on $O.D._{260}$.

**3.4.9 Biochemical assays.**

RNA splicing assays were performed by using an internally $^{32}P$-labeled 1.1-kb precursor RNA containing the Ll.LtrB-ΔORF intron with 5' and 3' exons of 66 and 42 nt, respectively. The precursor RNA was transcribed by a mutant T7 RNA polymerase that reads through cryptic termination sites (Lyakhov *et al*., 1997) from DNA templates generated by PCR of pACD2 with primers that append a T7 promoter. Transcription was performed in T7 transcription buffer (New England Biolabs) in the presence of [α-$^{32}P$]-UTP (800 Ci/mmol; Perkin-Elmer) overnight. RNA splicing reactions were carried out under previously described conditions by incubating 20 nM precursor RNA with 200 nM LtrA protein at 37°C in 100 μl of reaction medium containing 450 mM NaCl, 40 mM Tris-HCl (pH 7.5), 5 mM dithiothreitol, and different concentrations of $MgCl_2$ (Saldanha *et al*., 1999). Prior to splicing, the precursor RNA was renatured by heating 50 nM RNA in 40 μl of 10 mM Tris-HCl (pH 7.5) to 90°C for 1 min and then immediately diluting 2-fold into 40 μl of 2x reaction medium at 37°C. Splicing reactions were initiated by adding LtrA protein (20 μl in 1x reaction medium) to the 80 μl sample. For time courses, a 6-μl portion of the reaction was removed at each time point, and the reaction was terminated by mixing with 2 μl of a 4x stop solution of 0.2 M EDTA, 0.4% SDS, and 4 mg/ml proteinase K, incubating for at least 30 min at 37°C, and then mixing with 2x formamide/EDTA RNA loading buffer containing bromophenol blue and xylene cyanol. The reaction products were heated to 85°C for 5 min and analyzed by electrophoresis in a denaturing 4% (w/v) polyacrylamide gel, which was dried and scanned with a Typhoon

Trio PhosphorImager (Amersham Biosciences) and quantified using ImageQuant (Molecular Dynamics). Kinetic data were fitted to single or double exponential equations using Prism 6.0 (GraphPad).

Reverse splicing assays were performed as described (Mastroianni *et al*., 2008) with a 129-bp internally labeled DNA substrate containing the Ll.LtrB intron-insertion site generated by PCR of pLHS-ltrB (Matsuura *et al*., 1997) with Vent DNA polymerase (New England Biolabs) in the presence of [$\alpha$-$^{32}$P]-dTTP (800 Ci/mmol; Perkin-Elmer). The reactions were performed by incubating 1.65 nM DNA substrate with 82.5 nM RNPs (50-fold molar excess) in 100 μl of reaction medium containing 10 mM KCl, 50 mM Tris-HCl (pH 7.5), 5 mM dithiothreitol, and different concentrations of $MgCl_2$ at 37°C in the presence of 0.2 mM each of dATP, dCTP, dGTP, and dTTP complexed with equimolar $MgCl_2$. The reactions were initiated by adding RNPs, incubated for different times, and terminated and analyzed in a denaturing 6% (w/v) polyacrylamide gel, as described above for RNA splicing assays.

**3.4.10 Terbium-cleavage assays.**

Terbium cleavage was performed as described (Sigel *et al*., 2000). DV RNAs (41 nt; intron positions 842-882) were transcribed by using a Megascript SP6 kit (Ambion) from PCR-generated DNA templates that append an SP6 promoter. The RNA was 5' labeled with [$\gamma$-$^{32}$P]-ATP (800 Ci/mmol; Perkin-Elmer), gel-purified, and stored in 10 mM MOPS (pH 6.5), 0.1 mM EDTA at -80°C. For cleavage assays, labeled DV RNAs were diluted with 1.5 μM unlabeled RNA in 25 mM MOPS (pH 7.0) and 50 mM KCl and normalized to 15, 000 cpm/μl. Samples were renatured by heating in a thermocycler at 95°C for 45 s, then cooling to 42°C, adding the desired concentration of $MgCl_2$, and

incubating for 15 min before cooling to 25°C. Cleavage reactions were initiated by adding $TbCl_3$ and incubating at room temperature for 1 h. $TbCl_3$ (99.99% pure; Sigma) was stored in 5 mM cacodylate (pH 5.5) in 200 mM stock solutions. Reactions were quenched in formamide/EDTA RNA loading buffer, run in a denaturing 17% polyacrylamide gel, which was dried and scanned using a Typhoon Trio PhosphorImager (Amersham Biosciences), and quantified by using ImageQuant (Molecular Dynamics).

### 3.4.11 Structure modeling

Ll.LtrB DV models were generated by using PyMol, ModeRNA (Rother *et al*., 2011), and energy minimized using Phenix (Adams *et al*., 2010) with the help of David J. Sidote. The Ll.LtrB DV folded and extended models were created by using the crystal structure of the *O. iheyensis* group II intron (PDB:4E8M) and the NMR structure of the *P. littoralis* large ribosomal RNA intron (PDB:2F88), respectively. The *O. iheyensis* DV metal-ion ensemble was generated from 15 structures aligned in PyMol from the following PDB files: 3BWP, 3EOG, 3EOH, 3IGI, 4E8M, 4E8P, 4E8Q, 4E8R, 4E8V, 4FAQ, 4FAR, 4FAU, 4FAW, 4FAX, 4FB0.

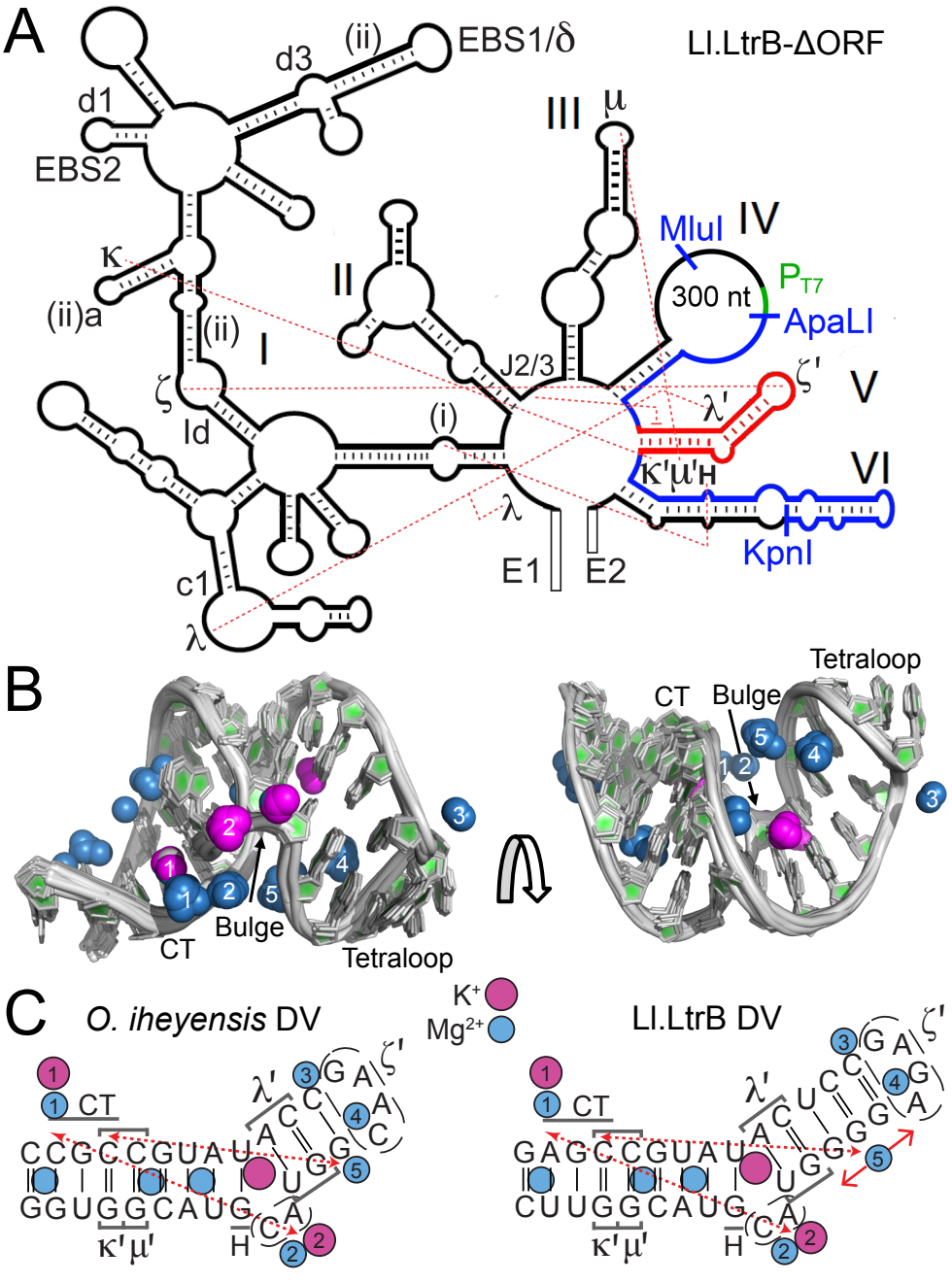
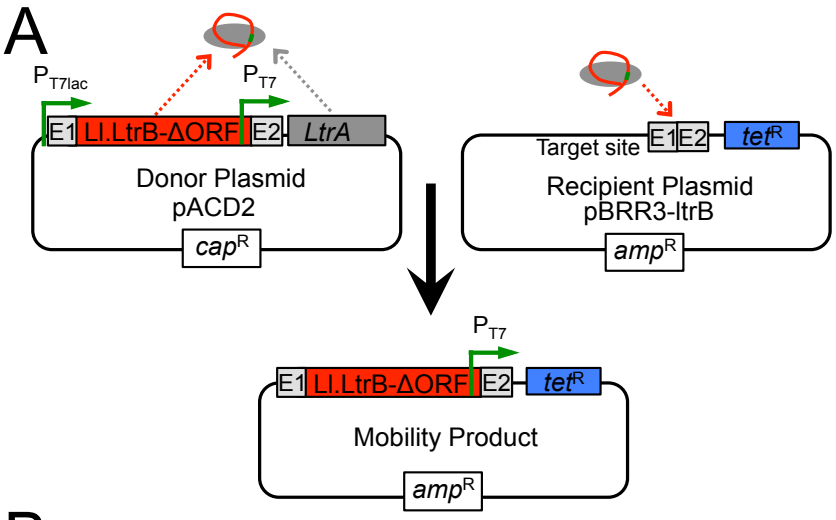



Figure 3.1: Group II intron RNA and DV structures.

(*A*) Secondary structure model of the Ll.LtrB-ΔORF group II intron, with DV highlighted in red. Greek letters indicate sequence elements involved in long-range tertiary interactions (dashed red lines). The locations of the ApaLI, KpnI, and MluI sites used in library construction and the inserted phage T7 promoter sequence used for genetic selections are indicated. (*B*) Metal ions bound to DV in an ensemble of 15 superposed X-ray crystal structures of the *O. iheyensis* group IIC intron (3BWP, 3EOG, 3EOH, 3IGI, 4E8M, 4E8P, 4E8Q, 4E8R, 4E8V, 4FAQ, 4FAR, 4FAU, 4FAW, 4FAX, 4FB0). $Mg^{2+}$ and $K^{+}$ ions are shown as blue and violet spheres, respectively. The $Mg^{2+}$ and $K^{+}$ ions that comprise the heteronuclear metal ion cluster at the active site are numbered M1/K1 and M2/K2, and the three $Mg^{2+}$ ions bound to the distal stem of DV are numbered M3, M4, and M5. (*C*) Secondary structure diagrams of DV of the *O. iheyensis* and *L. lactis* Ll.LtrB group II introns showing the locations of $Mg^{2+}$ and $K^{+}$ ions identified in the ensemble of *O. iheyensis* crystal structures of panel *B*. Ions shown outside the secondary structure are bound to the phosphodiester backbone, and those shown inside are bound within the helix in the *O. iheyensis* intron structures. A range of possible locations (red arrows) is shown for M5 in the Ll.LtrB intron due to the different lengths of the distal stem in subgroup IIA introns (Ll.LtrB; 5 bp) and IIC introns (*O. iheyensis*; 3 bp) (Lambowitz and Zimmerly, 2011). Dashed red lines show regions of the phosphodiester backbone that are bridged by bound $Mg^{2+}$ ions in the *O. iheyensis* structures. Abbreviations: CT, catalytic triad; H, sugar-edge/Hoogsteen contact with I(i) loop in Domain I.




**B**

| Strain | Retrohoming efficiency (%) | $[Mg^{2+}]_i$ (mM) | $[Mg^{2+}]_i$ (mM) range |
|---|---|---|---|
| WT (HMS174) | 48 ± 3 | 3.1 ± 0.4 | 2.7-3.9 |
| *corA* disruptant | 27 ± 2 | 1.6 ± 0.1 | 1.4-1.7 |
| *mgtA* disruptant | 0.24 ± 0.02 | 1.7 ± 0.1 | 1.6-1.8 |

Figure 3.2: *E. coli* selection system for variants of the Ll.LtrB intron with mutations in DV that increase retrohoming efficiency at low $Mg^{2+}$ concentrations.

(*A*) Plasmid-based group II intron retrohoming system used to select active DV variants in the *E. coli mgtA* disruptant. The $Cap^R$ intron-donor plasmid pACD2 uses a T7lac promoter to produce group II intron RNPs consisting of the Ll.LtrB-ΔORF intron lariat RNA and the IEP, denoted LtrA protein. The intron carries a phage T7 promoter sequence within DIV. The $Amp^R$-recipient plasmid contains an Ll.LtrB target site (ligated E1-E2 sequence of the *ltrB* gene from position -30 downstream to +15 upstream of the intron-insertion site) preceding a promoterless $tet^R$ marker. Retrohoming of the Ll.LtrB-ΔORF lariat RNA into the target site introduces the T7 promoter needed for $tet^R$ expression, enabling selection for $Tet^R$ colonies. (*B*) Retrohoming efficiency of the Ll.LtrB-ΔORF intron in *E. coli* wild-type (WT) HMS174(DE3) and mutants with disruptions in the *corA* and *mgtA* genes encoding $Mg^{2+}$-transporters. Retrohoming efficiencies were determined as the ratio of ($Tet^R$ + $Amp^R$)/$Amp^R$ colonies in the plasmid-based retrohoming assay, and intracellular free $Mg^{2+}$ concentrations ($[Mg^{2+}]_i$) were measured by using the fluorescent probe mag-fura-2. The values shown are the mean ± the S.E.M. for three determinations, with the range of $Mg^{2+}$ concentrations in different experiments also indicated.

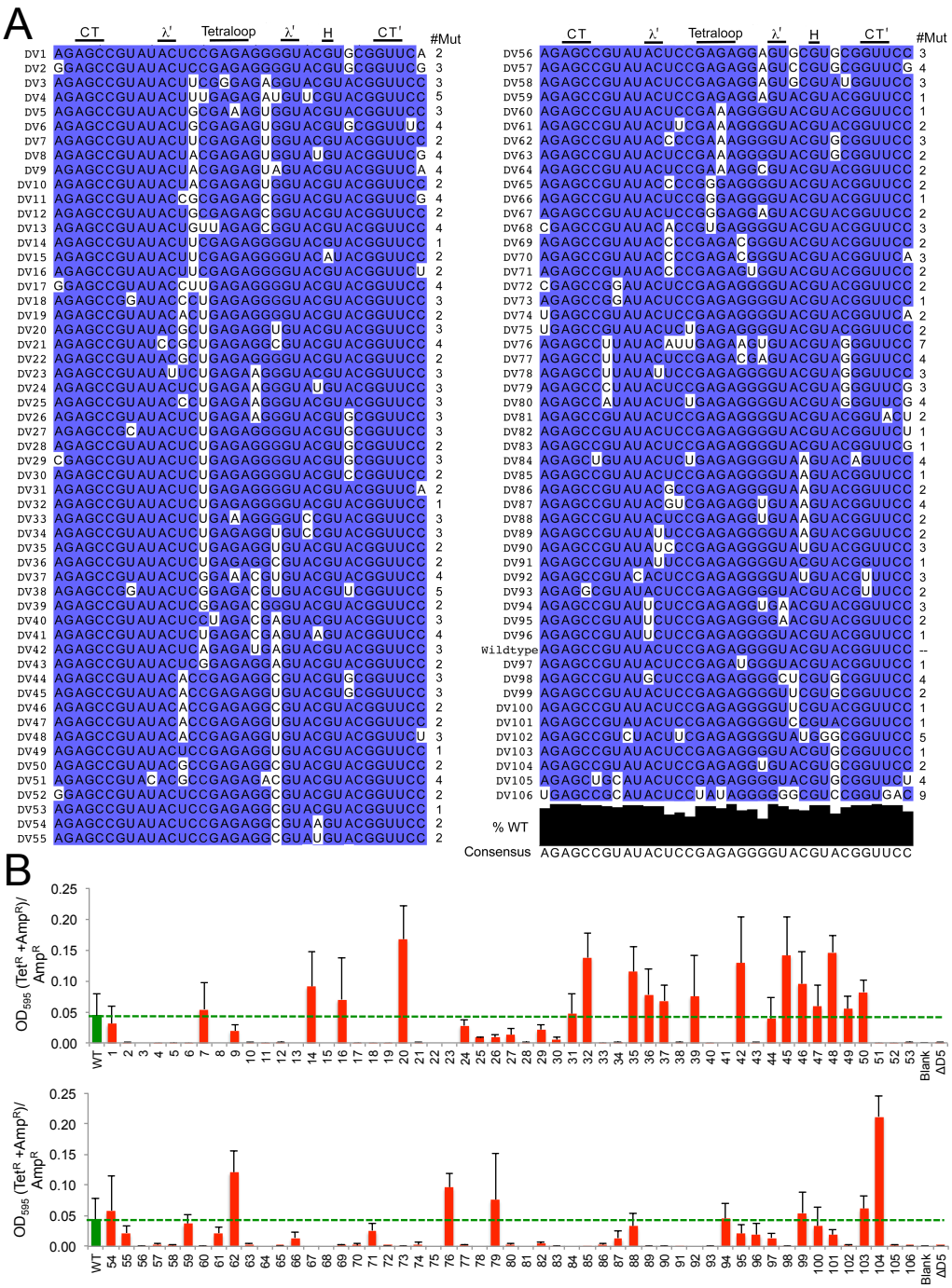



Figure 3.3: Sequences and retrohoming efficiencies of 106 active DV variants identified in a selection for Ll.LtrB-ΔORF introns that retrohome in the *mgtA* disruptant.

The variants were selected by using the plasmid-based retrohoming system of Figure 3.2*A* from a library of Ll.LtrB-ΔORF introns in which DV was partially randomized ("doped") with 70% of the wild-type and 10% of each mutant nucleotide at each position. Each $Tet^R$ + $Amp^R$ colony obtained in the selection contained an intron variant, which was identified by colony PCR and sequencing. Approximately 300 colonies were picked and found to contain 106 unique DV sequences. (*A*) Sequences of the 106 active DV variants organized by sequence similarity. Blue shading indicates a match to the wild-type (WT) sequence. The number of mutations (#Mut) found in each variant is indicated to the right. The percentage of wild-type nucleotides at each position and a consensus sequence for the selected variants are shown at the bottom. (*B*) Retrohoming efficiencies of the wild-type Ll.LtrB-ΔORF and 106 active DV variants determined by using a high-throughput 96-well plate version of the plasmid-based retrohoming assay. The bars show the retrohoming efficiency as measured by ratio of the $O.D._{595}$ in the presence of tetracycline + ampicillin to that with ampicillin alone for each variant. The values are the mean ± the S.E.M. for three determinations. The green line shows the retrohoming efficiency of the wild-type Ll.LtrB-ΔORF assayed in parallel. Thirty variants had retrohoming efficiencies equal to or greater than that of the wild-type intron in this assay.

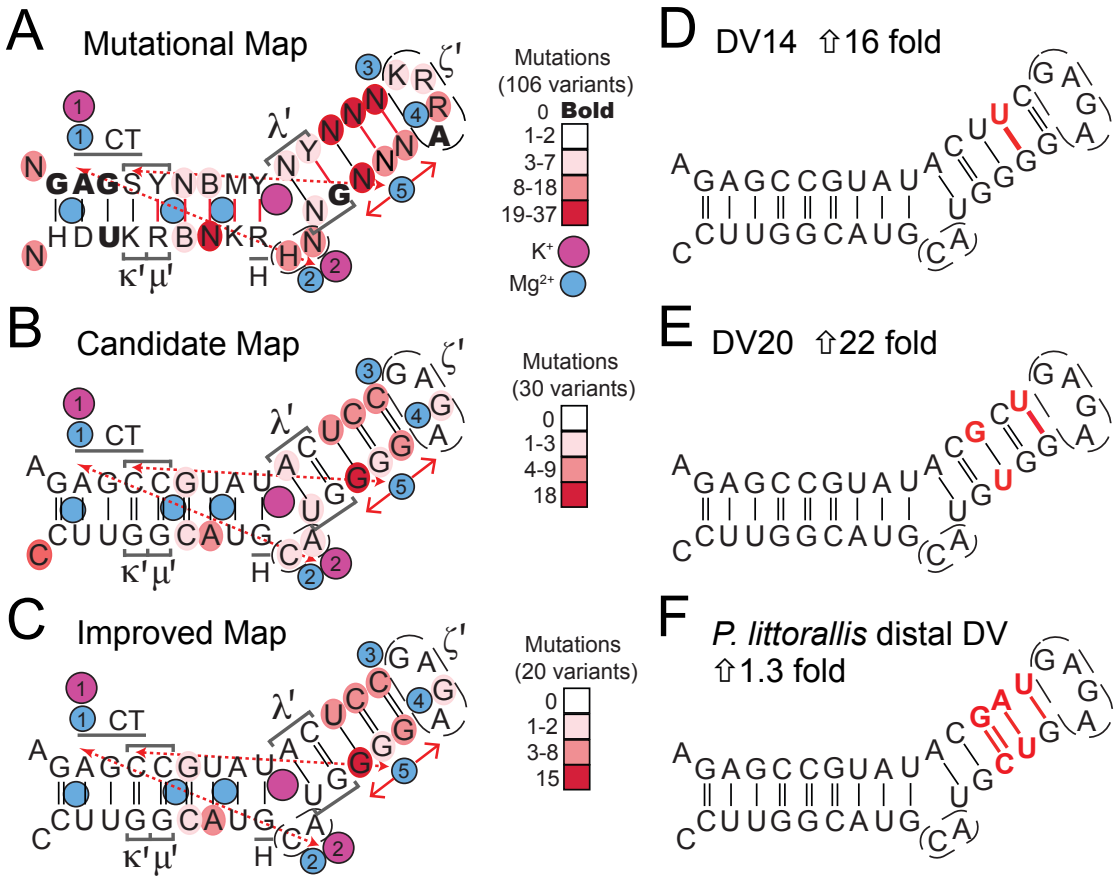



Figure 3.4: DV variants obtained in a selection for increased retrohoming efficiency in the *mgtA* disruptant from a library of Ll.LtrB introns in which DV was partially randomized.

The selection was done by using the plasmid-based retrohoming system (Figure 3.2) with a library of ~$10^9$ Ll.LtrB-ΔORF intron variants in which DV (intron positions 844-879) was partially randomized with 70% of the wild-type nucleotide and 10% of each mutant nucleotide at each position. (*A*) DV mutational map for the 106 active variants obtained after two rounds of selection. Nucleotides that were invariant in the selection are indicated in bold, and those with 1-2 mutations are shown with no shading. Nucleotides with larger numbers of mutations are highlighted according to a color scale shown in the Figure. Nucleotides are indicated as follows: B = C, G, U; D = A, G, U; H = A, C, U; K = U, G; N = all; M = A, C; R = A, G; S = C, G; Y = C, U. Base pairs for which compensatory base changes were found in the selections are shown in red. Nucleotide sequences and retrohoming efficiences for the 106 variants are shown in Figure 3.3. (*B*) and (*C*) Mutational maps of DV for the 30 candidate variants having high retrohoming efficiencies in the high-throughput retrohoming assay (*B*) and the 20 improved variants confirmed in plating assays to have at least three-fold higher retrohoming efficiency than the wild-type intron in the *mgtA* disruptant (*C*). Numbers of mutations at each position are indicated by color highlighting according to the scales shown for each panel. Nucleotide sequences of the 30 candidate mutants including the 20 improved variants are shown in Figure 3.5. (*D*) and (*E*) DVs of the top performing selected variants DV14 and DV20, respectively. (*F*) An additional variant containing nucleotide substitutions found in the distal stem of the *Pylaiella littorallis* LSU/2 intron, which self-splices at lower $Mg^{2+}$ concentrations than other group II introns (Costa *et al*., 1997). Nucleotides that differ from those in the wild-type Ll.LtrB intron are shown in red, and red bars show changed numbers of hydrogen bonds between base pairs.

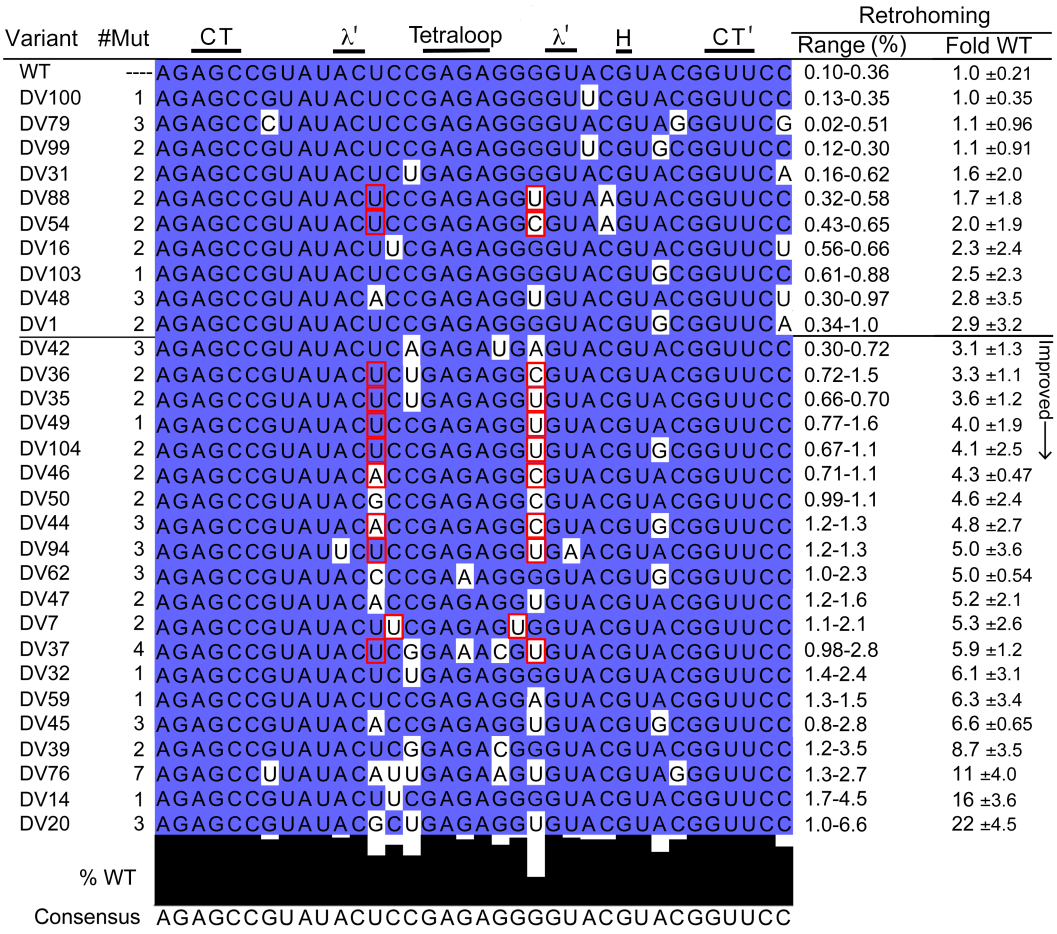



Figure 3.5: DV sequences of variants with increased retrohoming efficiency in the *mgtA* disruptant selected from an Ll.LtrB intron library in which DV was partially randomized.

The Figure shows sequences and retrohoming efficiencies determined by plating assay for the 30 DV variants found by the high-throughput 96-well plate assay to have retrohoming frequencies equal to or greater than that of the wild-type (WT) intron in the *mgtA* disruptant (Figure 3.3). Variants with a retrohoming efficiency > three-fold higher than that of the wild-type intron in the plating assay are denoted as improved variants and are demarcated in the Figure. Blue shading indicates a match to the wild-type sequence, and red boxes indicate mispairings. The number of mutations found in each variant is indicated to the left, and the retrohoming efficiency and fold increase relative to the wild-type intron in the plating assay (mean ± S.E.M. for three determinations) are indicated to the right. The percentage of wild-type nucleotides at each position and a consensus sequence for the selected variants are shown at the bottom.

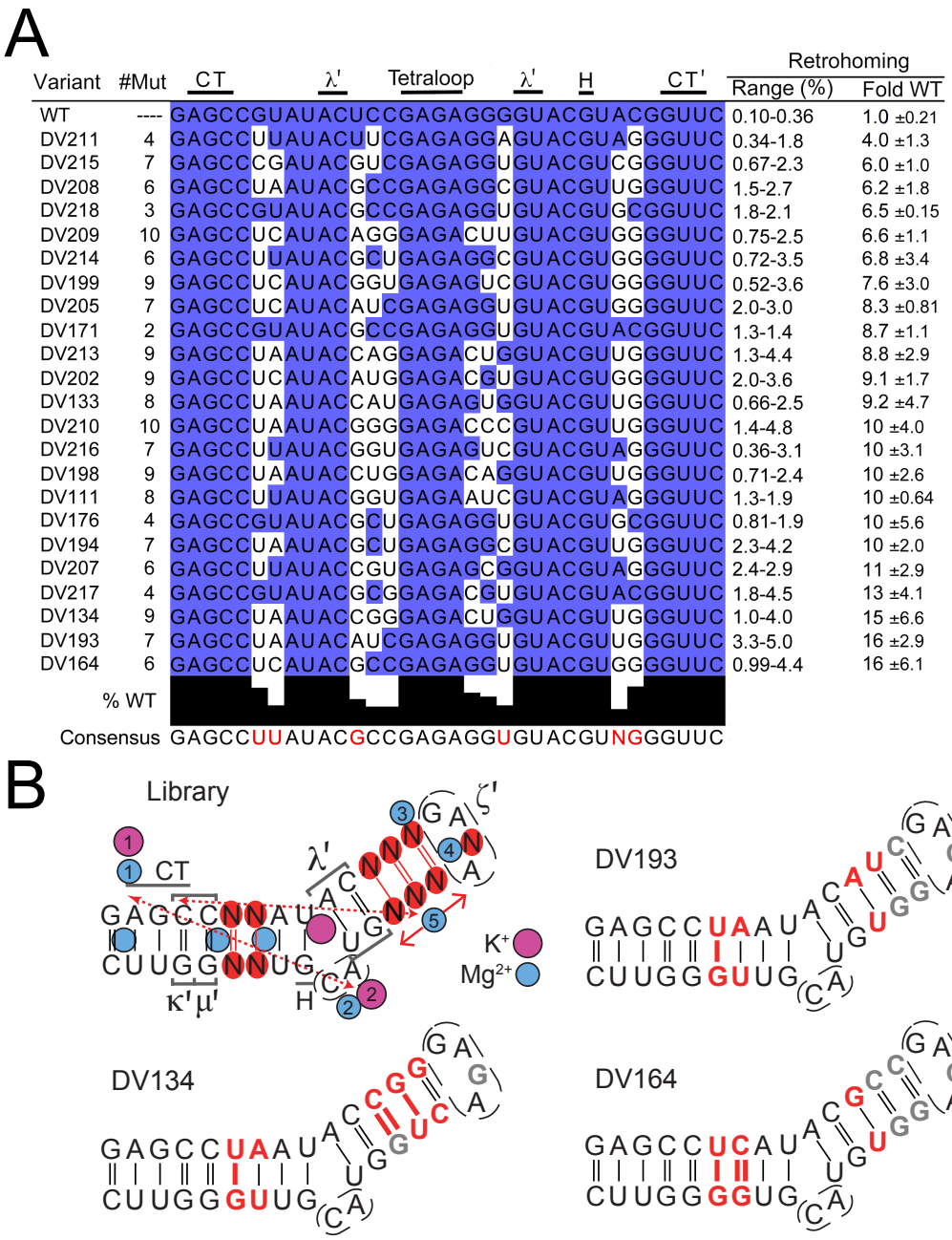



Figure 3.6: Saturating selection of DV variants at 11 nucleotide positions that were sites of mutations in the improved variants from the initial "doped" DV selection.

We performed five rounds of selection using the plasmid-based mobility assay in the *mgtA* disruptants starting from a library of 4.2 x $10^6$ Ll.LtrB-ΔORF intron variants randomized at the 11 sites in DV that were mutated in improved variants in the initial selection (Figures 3.4 and 3.5). After sequencing 96 individual colonies from the fifth round, we identified 23 unique variants and determined their retrohoming efficiencies by the plating assay of Figure 3.2*A*. (*A*) Sequences of the 23 enhanced DV sequences from the selection, ordered by increasing retrohoming efficiency in the *mgtA* disruptant. Retrohoming efficiencies were determined by quantitative plating assays. Blue shading indicates a match to the wild-type sequence. The number of mutations (#Mut) found in each variant is indicated to the left, and the retrohoming efficiency and fold increase relative to the wild-type intron assayed in parallel (mean ± S.E.M. for three determinations) are indicated to the right. The percentage of wild-type nucleotides at each position and a consensus sequence for the selected variants are shown at the bottom, with nucleotides in red indicating those that differ from the wild-type sequence. (*B*) Map of DV showing positions in the library that were randomized (N, red circles) and the DV structures for the top three variants with the highest retrohoming efficiency in the *mgtA* disruptant. In the variants, nucleotides that differ from those in the wild-type Ll.LtrB-ΔORF intron are shown in red, and red bars show changed numbers of hydrogen bonds between base pairs.

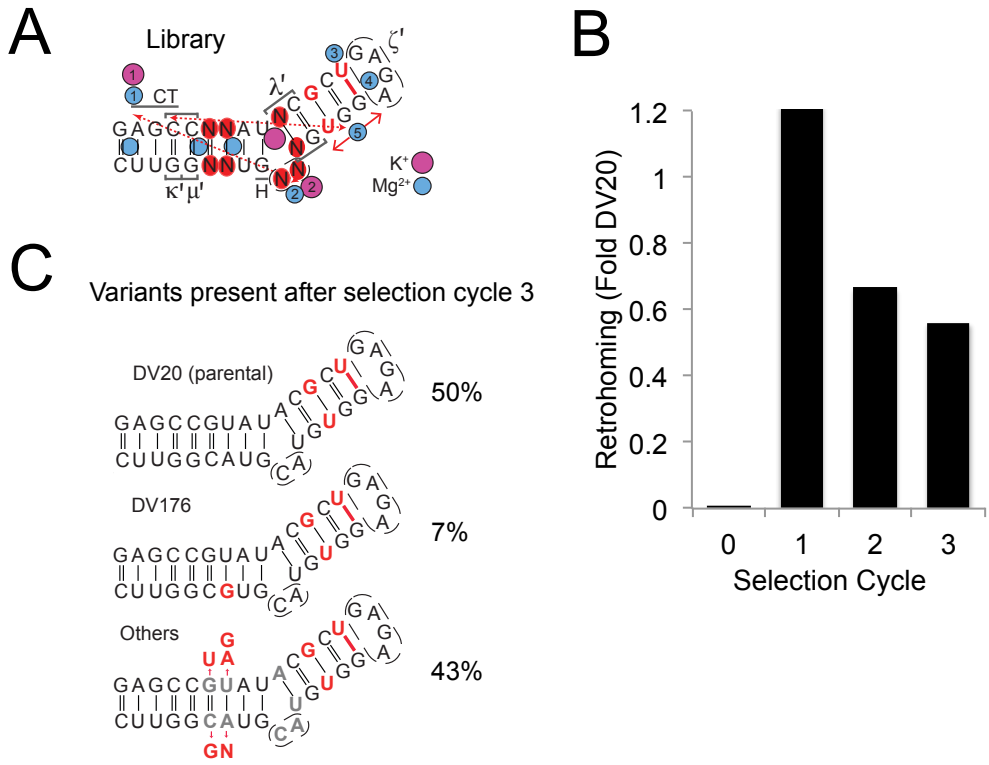



Figure 3.7: Saturating selection of DV variants based on combining mutations found in the distal stem of the highest performing variant, DV20, with modifications at eight other nucleotides near potential $Mg^{2+}$-binding sites.

We performed selections using the plasmid-mobility assay in the *mgtA* disruptant from a library of $10^4$ Ll.LtrB-ΔORF intron variants with the distal stem mutations found in DV20 and randomized nucleotides at the eight other positions near potential $Mg^{2+}$-binding sites. Retrohoming efficiencies relative to the parental DV20 variant were determined for pools obtained by combining plasmids from $Tet^R$ + $Amp^R$ colonies from each round of selection, and individual clones were sequenced periodically. (*A*) Map of DV, with mutated nucleotides in DV20 shown by red letters and additional nucleotides randomized for the selection highlighted by red circles. (*B*) Retrohoming efficiency of the initial library and pooled variants from three rounds of selection for retrohoming in the *mgtA* disruptant. (*C*) Map of DV showing mutations in DV20, DV176, and other variants present in the final pool (mutant nucleotides and changed numbers of hydrogen bonds shown in red) based on colony PCR sequencing of 24 variants. The wild-type dinucleotide bulge sequence AC was selected in all variants sampled in cycle 1, and the λ' base pair reverted to wild-type A-U by cycle 2.

A

| Variant | family | proximal | distal | distal | proximal | Retrohoming (%) | Retrohoming Fold WT |
|---|---|---|---|---|---|---|---|
| **WT** | A | AGAGCCGUAUACUCCGAGAGGGGUACGUACGGUUC | | | | **0.34** | **----** |
| DVR1 | D | AGAGCCUUAUACGCCGAGAGGUGUACGUAGGGUUC | | | | 0.56 | 1.6 |
| DVR2 | B | AGAGCCUUAUACGCCGAGAGGUGUACGUGGGGUUC | | | | 0.34 | 1.0 |
| DVR3 | B | AGAGCCUUAUACGGUGAGAGUUGUACGUGGGGUUC | | | | 0.12 | 0.4 |
| DVR4 | D | AGAGCCUUAUACGGUGAGAGUUGUACGUAGGGUUC | | | | 0.05 | 0.1 |
| DVR5 | C | AGAGCCUAAUACGUCGAGAGGUGUACGUUGGGUUC | | | | 0.51 | 1.5 |
| DVR6 | A | AGAGCCGUAUACGUUGAGAGGUGUACGUACGGUUC | | | | 0.75 | 2.2 |
| DVR7 | C | AGAGCCUAAUACGUUGAGAGGUGUACGUUGGGUUC | | | | 0.21 | 0.6 |
| DVR8 | D | AGAGCCUUAUACGUUGAGAGGUGUACGUAGGGUUC | | | | 0.25 | 0.7 |
| DVR9 | B | AGAGCCUUAUACGUUGAGAGGUGUACGUGGGGUUC | | | | 0.13 | 0.4 |
| DVR10 | A | AGAGCCGUAUACGUGGAGACGUGUACGUACGGUUC | | | | 0.22 | 0.6 |
| DVR11 | C | AGAGCCUAAUACGUGGAGACGUGUACGUUGGGUUC | | | | 0.03 | 0.1 |
| DVR12 | D | AGAGCCUUAUACGUGGAGACGUGUACGUAGGGUUC | | | | 0.27 | 0.8 |
| DVR13 | B | AGAGCCUUAUACGUGGAGACGUGUACGUGGGGUUC | | | | 0.12 | 0.4 |

B

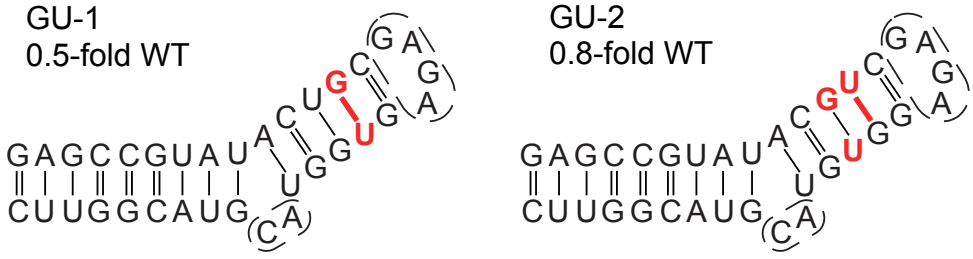



Figure 3.8: Rationally designed sequences for DV.

(A) Based on the variants identified from the selections, we grouped the proximal stem mutations into families (A-D) and combined them with variations in the distal stem of DV. These variants were tested for retrohoming activity in the MgtA mutant using the plasmid-based mobility assay. The proximal stem family is indicated to the left. Blue shading indicates a match to the wild-type sequence. The retrohoming efficiency and fold increase relative to the wild-type intron in the plating assay (mean ± S.E.M. for three determinations) are indicated to the right. (B) We generated two G-U mutation variants in the distal stem of DV based on $Mg^{2+}$-binding affinity studied in the work by (Keel *et al.*, 2007). The variants were tested in the MgtA mutant for retrohoming activity.

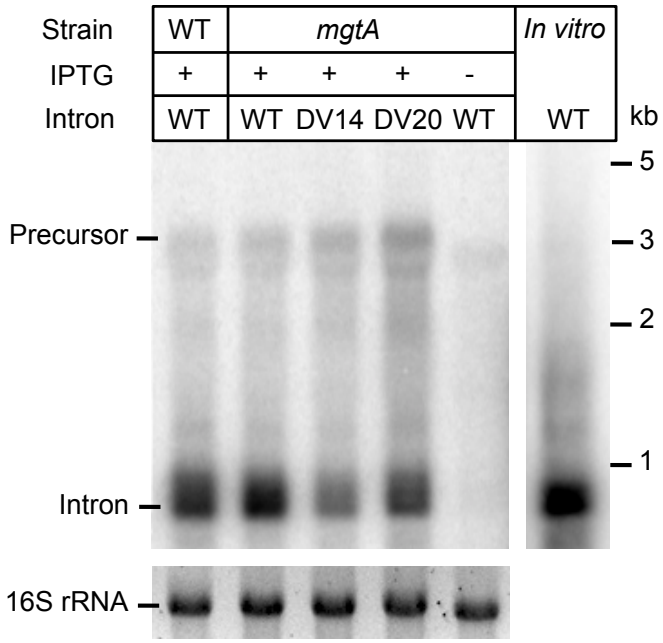



Figure 3.9: Northern hybridization of the wild-type Ll.LtrB-ΔORF intron and variants DV14 and DV20 in wild-type HMS174(DE3) and the *mgtA* disruptant.

Cells containing the intron-donor plasmid pACD2 and recipient plasmid pBRR3-ltrB were grown in 50-ml cultures and induced with IPTG as for plasmid-based retrohoming assays, and then total cellular RNA was extracted, denatured with glyoxal, and run in a 1% agarose gel. The gel was blotted to a nylon membrane and hybridized with a $^{32}$P-labeled intron probe. Gel loads were normalized based on $O.D._{260}$ and levels of 16S rRNA detected by ethidium bromide staining (shown below). Self-spliced Ll.LtrB-ΔORF intron RNA was run in a parallel lane shown to the right. An additional control lane shows that the intron RNAs were not detected without IPTG induction. The numbers to the right of the gel indicate the positions of ssRNA size markers (New England Biolabs).

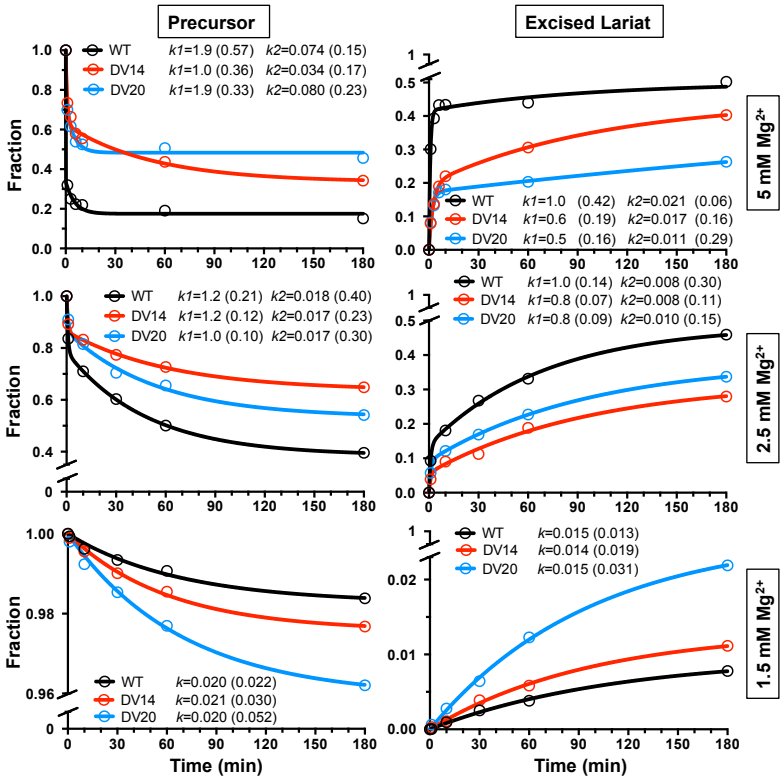



Figure 3.10: Splicing of the wild-type and variant DV14 and DV20 Ll.LtrB-ΔORF introns at different $Mg^{2+}$ concentrations.

Splicing time courses were performed by incubating $^{32}$P-labeled precursor RNA containing the wild-type (WT) or variant introns with a 10-fold molar excess of LtrA protein in reaction media containing different $Mg^{2+}$ concentrations (5, 2.5, or 1.5 mM) at 37°C. Reactions were initiated by adding LtrA protein and incubated for different times up to 3 h. The products were analyzed in a denaturing 4% polyacrylamide gel, which was dried and quantified with a Phosphorimager. The plots show disappearance of precursor RNA (left) and appearance of excised lariat RNA (right) fit to a single or double exponential equation. The rate constants $k_1$ or $k_2$ in $min^{-1}$ and amplitudes [fraction of precursor RNA remaining (left) and fraction of precursor RNA spliced to excised lariat (right)] for each phase are indicated in each plot and are the averages from three experiments. The standard deviations for rate constants and amplitudes were < 20%, except for the small slow phase of splicing of the wild-type intron at 5 mM $Mg^{2+}$, where the standard deviations were ~40%.

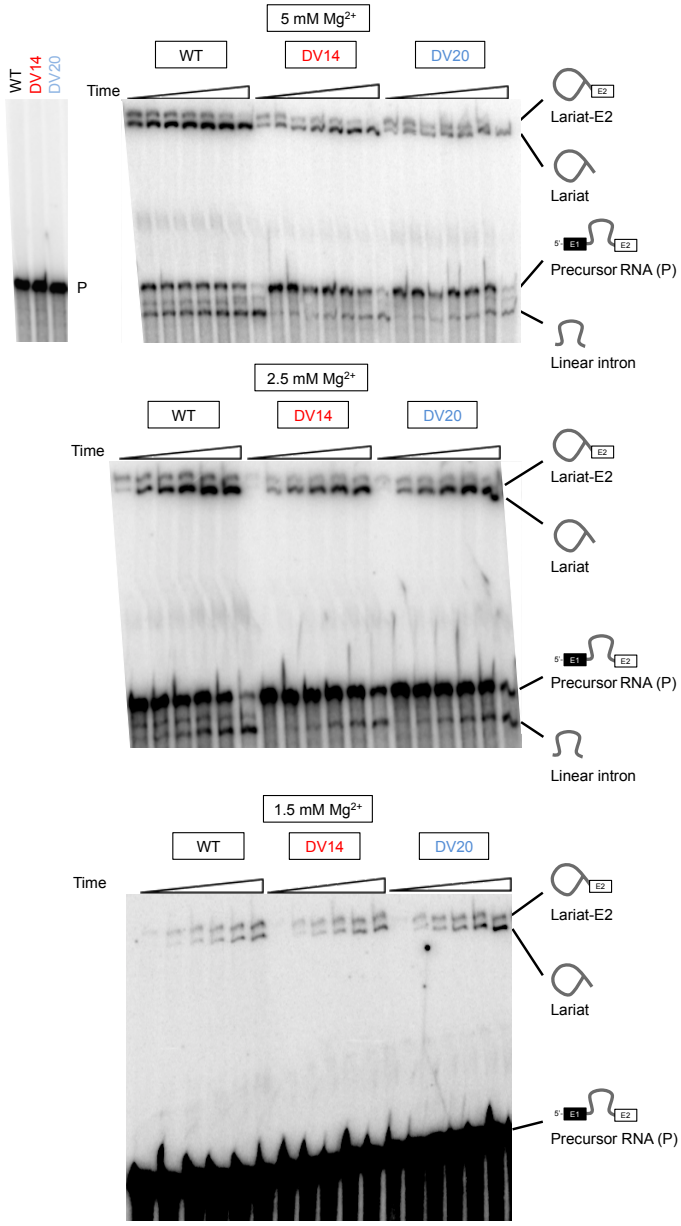



Figure 3.11: Representative gels from time courses of LtrA-promoted splicing of wild-type and variant Ll.LtrB-ΔORF introns at different $Mg^{2+}$ concentrations.

Times ranged from 1 to 180 min. The gels shown are those used for the plots in Figure 3.10. Products are identified schematically to the right of each gel. A section of the same gel showing unreacted wild-type and variant precursor RNAs (P) in splicing medium prior to addition of LtrA is shown at the top left.

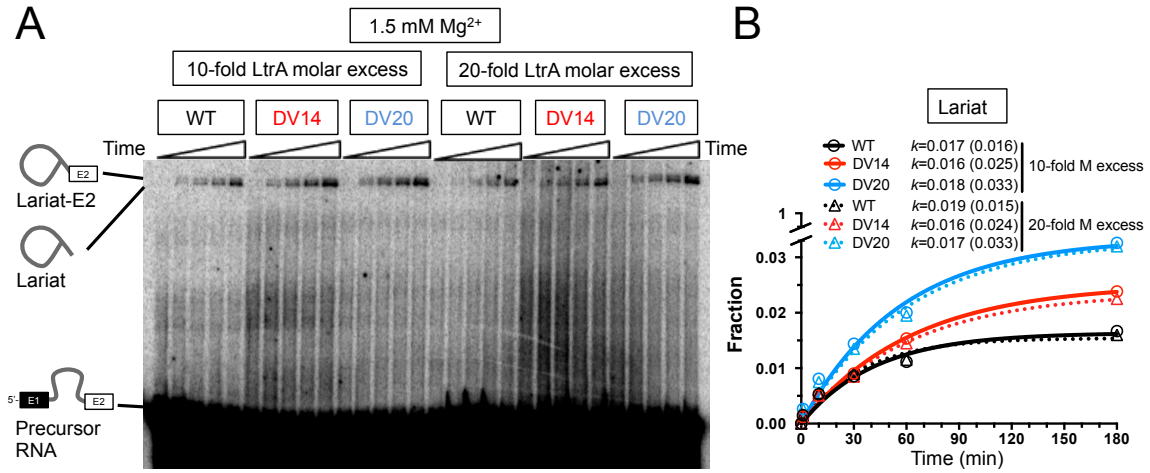



Figure 3.12: Splicing time courses of wild-type and variant DV14 and DV20 Ll.LtrB-ΔORF introns at 1.5 mM $Mg^{2+}$ with a 10- and 20-fold molar excess of LtrA protein.

Splicing reactions were performed as described in Figure 3.10 and Methods. Products are identified schematically to the left of the gel. (*A*) $^{32}$P-labeled precursor RNA (20 nM) containing wild-type (WT) or variant introns were spliced with 200 or 400 nM LtrA protein in 1.5 mM $Mg^{2+}$ for up to 3 h. (*B*) The plot shows appearance of lariat RNA products from (*A*) fit to a single exponential equation with rate constants ($k$, $min^{-1}$) and amplitudes (fraction of precursor RNA spliced to lariat RNA) indicated in the plot.

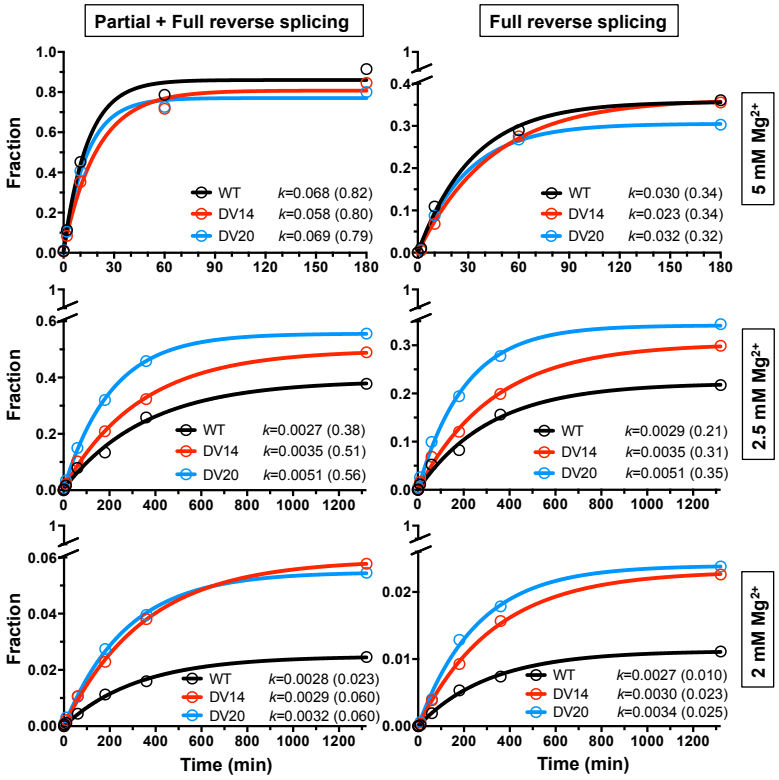



Figure 3.13: Time courses of reverse splicing of group II intron RNPs in target DNA-primed reverse transcription reactions for the wild-type and variant DV14 and DV20 Ll.LtrB-ΔORF introns at different $Mg^{2+}$ concentrations.

Reverse splicing time courses were performed by incubating a 50-fold molar excess of RNPs containing the wild-type (WT) Ll.LtrB-ΔORF and variant DV14 and DV20 intron RNAs with a 129-bp internally labeled DNA substrate containing the Ll.LtrB intron-target site in reaction media containing dNTPs for cDNA synthesis and different $Mg^{2+}$ concentrations (5, 2.5, or 2 mM) at 37°C. Reactions were initiated by adding RNPs and incubated for different times up to 22 h. The products were run in a denaturing 6% polyacrylamide gel, which was dried and quantified with a Phosphorimager. The plots show the accumulation of products resulting from the first and second steps of reverse splicing (lariat RNA joined to the 3' DNA exon and linear intron RNA inserted between the DNA exons, respectively) and of full reverse splicing (linear intron RNA inserted between the two DNA exons) fit to an equation with a single exponential. The rate constants ($k$, $min^{-1}$) and amplitudes (fraction of DNA substrate that has undergone reverse splicing) indicated in the plots are the averages from three experiments with standard deviations of < 20%. Similar results were obtained for reactions with a 100-fold molar excess of RNPs.

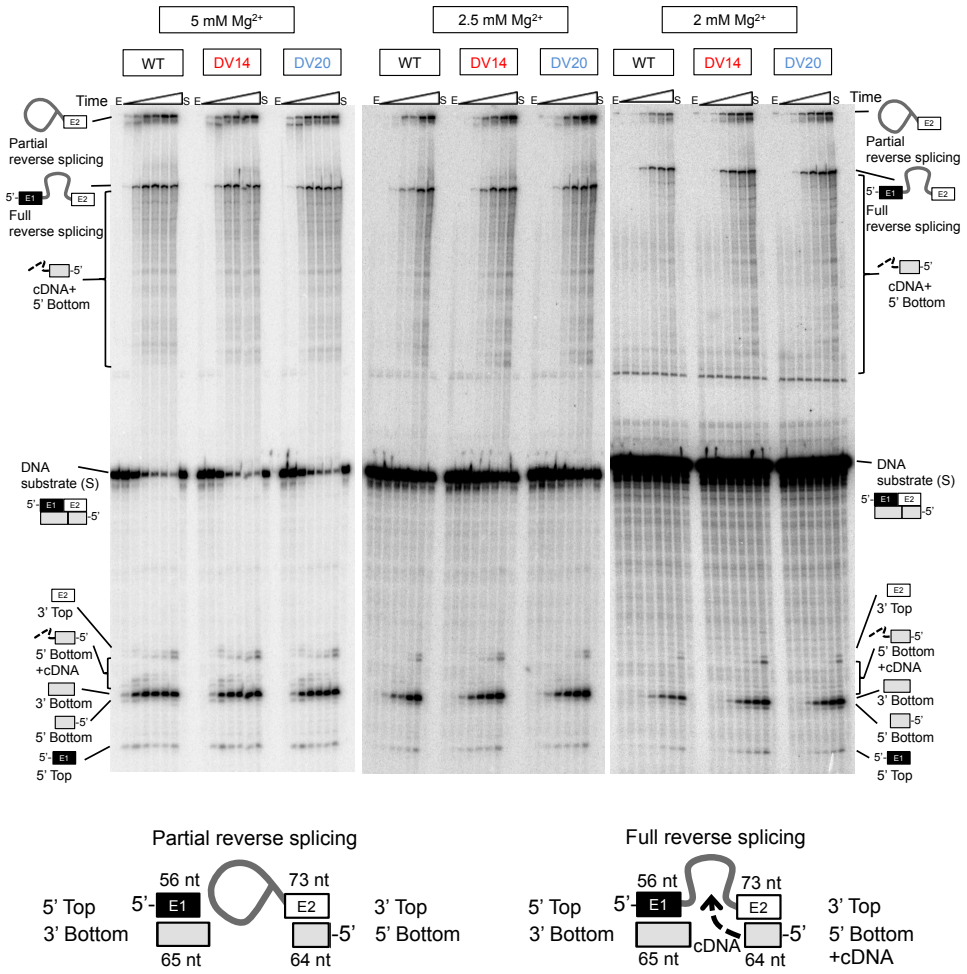



Figure 3.14: Representative gels from time courses of reverse splicing of the Ll.LtrB-ΔORF intron during target DNA-primed reverse transcription at different $Mg^{2+}$ concentrations.

Times ranged from 1 to 1,300 min. The gels shown are those used for the plots in Figure 3.13. Products are identified schematically to the left and right of the gel. For each variant and $Mg^{2+}$ concentration, control lane E shows all reaction components including Ll.LtrB RNPs incubated for 3 h in reaction medium containing 50 mM EDTA and 5 mM $MgCl_2$, and control lane S shows the substrate prior to the reaction. The partial reverse splicing product is resolved as a doublet, possibly due to nicking of the attached lariat RNA (11). Schematics of the products resulting from partial and full reverse splicing and bottom-strand cleavage are shown below the gel. The 5' bottom-strand cleavage product is extended by synthesis of the intron cDNA (dashed line), leading to higher molecular weight products.

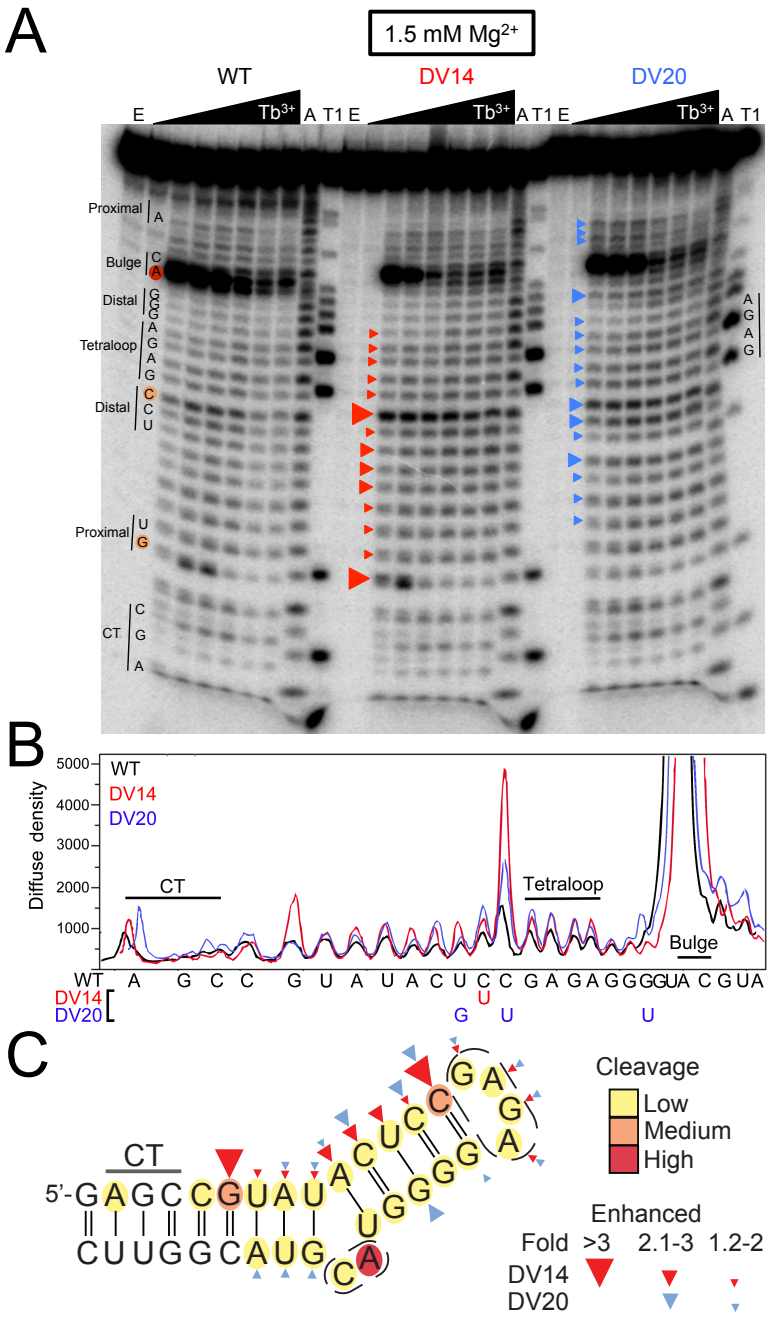



Figure 3.15: Terbium cleavage of isolated DV from the wild-type, DV14, and DV20 Ll.LtrB introns at 1.5 mM $MgCl_2$.

(*A*) Gel assay of terbium ($Tb^{3+}$) cleavage. 5' $^{32}$P-labeled DV RNA in 50 mM KCl, 1.5 mM $MgCl_2$, and 25 mM MOPS (pH 7.0) was incubated with increasing concentrations of $TbCl_3$ (0, 0.01, 0.03, 0.1, 0.3, 1, and 10 mM) for 1 h at room temperature. For each DV RNA, additional lanes show alkaline hydrolysis (A) and RNase T1 (T) ladders and a control in which the RNA was pre-incubated in 50 mM EDTA (E), which chelates both $Mg^{2+}$ and $Tb^{3+}$, prior to adding 10 mM $TbCl_3$. Samples were analyzed in a denaturing 17% polyacrylamide gel, which was dried and scanned with a Phosphorimager. The locations of the $Tb^{3+}$-cleavage sites are shown on a secondary structure map of DV below. Nucleotides that gave high or medium cleavage in the wild-type (WT) intron are shown in red or orange circles, respectively. Nucleotides that displayed enhanced cleavage in the DV14 and DV20 variants, based on the scales described below, are indicated by red and blue triangles, respectively. (*B*) Phosphorimager scan showing quantification of cleavage at 10 μM $TbCl_3$. Bands heights are normalized to input DV RNA (top of gel). (*C*) Secondary structure map of DV showing the location of cleavage sites. Nucleotides in the wild-type DV that gave significant $Tb^{3+}$ cleavage over background are highlighted in yellow (low, 1.5-3 fold); orange (medium, 3.1-10 fold); or red (high, > 10 fold). Residues that showed enhanced cleavage in the DV14 or DV20 variants are indicated as colored triangles, with the size of the triangle indicating the fold-increase relative to the wild-type DV (large, > 3 fold; medium, 2.1-3 fold; small, 1.2-2 fold). Cleavages occur 3' to the nucleotide indicated. The experiment was repeated three times with similar results.

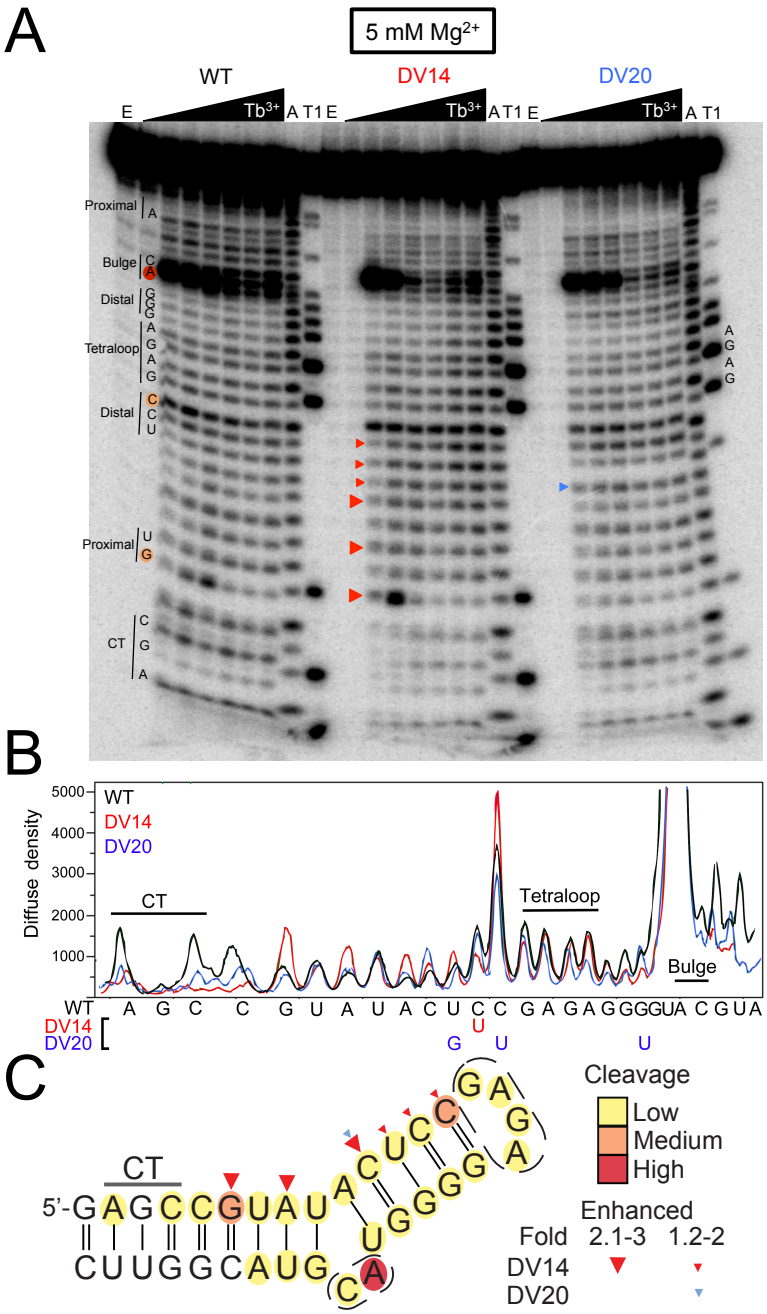



Figure 3.16: Terbium cleavage of isolated DV from the wild-type, DV14, and DV20 Ll.LtrB-ΔORF introns at 5 mM $MgCl_2$.

(*A*) Gel assay of terbium ($Tb^{3+}$) cleavage. 5' $^{32}$P-labeled DV RNA in 50 mM KCl, 5 mM $MgCl_2$, and 25 mM MOPS (pH 7.0) was incubated with increasing concentrations of $TbCl_3$ (0, 0.01, 0.03, 0.1, 0.3, 1, and 10 mM) for 1 h at room temperature. For each DV RNA, additional lanes show alkaline hydrolysis (A) and RNase T1 (T) ladders and a control in which the RNA was pre-incubated in 50 mM EDTA (E), which chelates both $Mg^{2+}$ and $Tb^{3+}$, prior to adding 10 mM $TbCl_3$. Samples were analyzed in a denaturing 17% polyacrylamide gel, which was dried and scanned with a Phosphorimager. Nucleotides that gave high or medium cleavage in the wild-type (WT) intron are shown in red or orange circles, respectively. Nucleotides that displayed enhanced cleavage in the DV14 and DV20 variants, based on the scales described below, are indicated by red and blue triangles, respectively. (*B*) Phosphorimager scan showing quantification of cleavages in wild-type (WT), DV14, and DV20 DVs at 10 μM $TbCl_3$. Band heights for the different DV RNAs are normalized to the input full-length RNA (top). (*C*) Secondary structure map of DV showing the location of cleavage sites. Nucleotides in the wild-type DV that gave significant $Tb^{3+}$ cleavage over background are highlighted in yellow (low, 1.5-3 fold); orange (medium, 3.1-10 fold); or red (high, > 10 fold). Residues that showed enhanced cleavage in the DV14 or DV20 variants are indicated as colored triangles, with the size of the triangle indicating the fold-increase relative to the wild-type DV (medium, 2.1-3 fold; small, 1.2-2 fold). Cleavages occur 3' to the nucleotide indicated. The experiment was repeated three times with similar results.

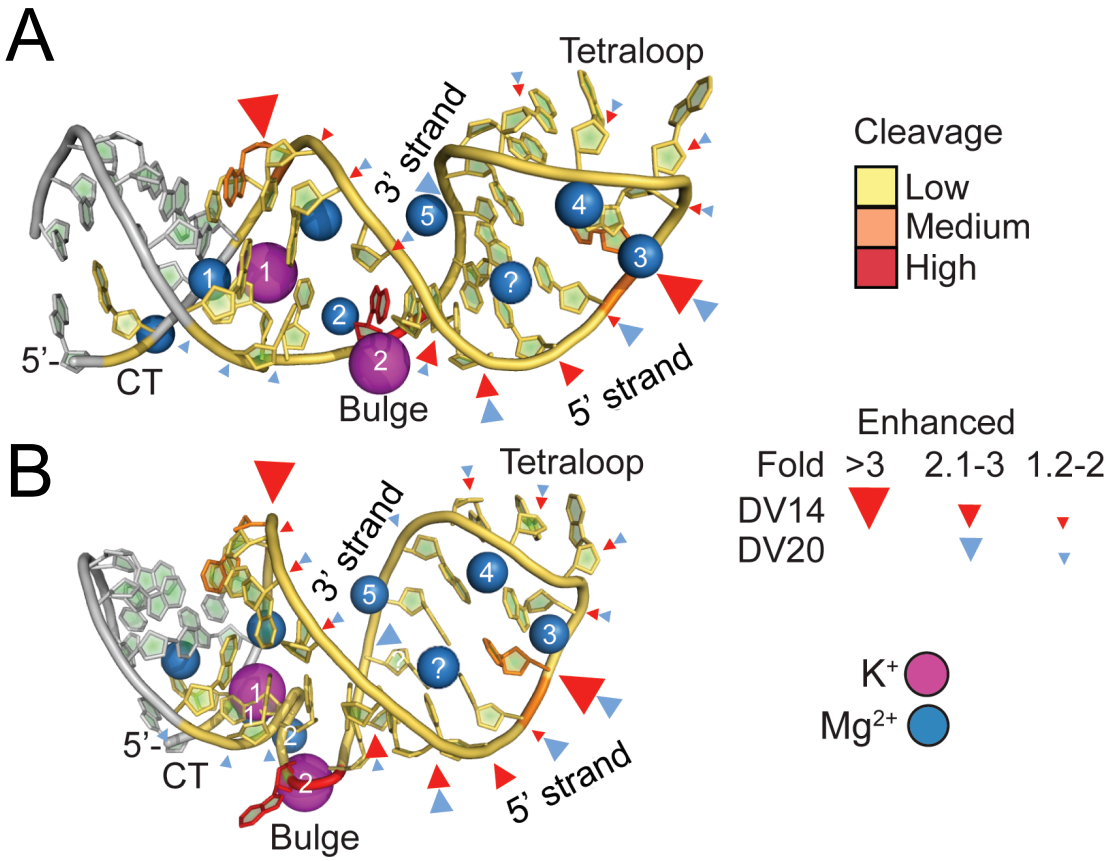



Figure 3.17: Models of $Mg^{2+}$ binding and terbium cleavage on tertiary structures of extended and folded Ll.LtrB intron DV.

Terbium ($Tb^{3+}$) cleavage patterns and $Mg^{2+}$ and $K^{+}$ ions found in the *O. iheyensis* DV crystal structure ensemble were modeled onto tertiary structure models of the Ll.LtrB intron DV. (*A*) and (*B*) three-dimensional models of DV of the Ll.LtrB intron based on (*A*) the NMR structure of the isolated DV of the *P. litorralis* intron (Seetharaman *et al.*, 2006), and (*B*) the folded DV structure in the crystal structures of the *O. iheyensis* intron (Marcia and Pyle, 2012). In both models, $Mg^{2+}$ and $K^{+}$ ions were positioned based on their locations in an ensemble of 15 superposed *O. iheyensis* DV structures shown in Figure 3.1*B* (Toor *et al.*, 2008a, 2008b; Marcia and Pyle, 2012). $Tb^{3+}$ cleavages found in the isolated DVs of wild-type and variant Ll.LtrB introns at 1.5 mM $Mg^{2+}$ are shown on the model, with $Tb^{3+}$ cleavages in the wild-type DV indicated by colors on the phosphodiester backbone and those in the variant introns indicated by different-sized triangles as in Figure 3.15. Cleavages occur 3' to the nucleotide indicated. The model was generated in collaboration with David J. Sidote.

# Chapter 4: Deep sequencing reveals the fitness landscape of a group II intron during directed evolution for improved retrohoming within the *Xenopus laevis* oocyte nucleus

## 4.1 Introduction

The factors limiting group II intron activity within higher eukaryotes remain unclear, although low [$Mg^{2+}$] remains a strong candidate. In order to engineer a group II intron for this environment, directed evolution within a higher eukaryotic cell could mitigate some of this uncertainty. Our laboratory has previously shown that Ll.LtrB RNPs retrohome efficiently into plasmids within the *Xenopus laevis* oocyte nucleus if supplemented with 500 mM $MgCl_2$ (Mastroianni *et al*., 2008). The *Xenopus laevis* oocyte possesses a number of features that make it ideal for group II intron directed evolution. First, the nucleus contains similar [$Mg^{2+}$] to human cells (Horowitz and Tluczek, 1989; Gunther, 2006) and, has historically, been used in molecular biology to elucidate nuclear processes, including recombination, repair, and transcription (Brown, 2004). Second, the large size of the nucleus (~1 mm) permits microinjection of foreign particles that can be generated *in vitro*, which permits the introduction of *in vitro*-prepared RNP libraries at $>10^{11}$ RNPs per oocyte (Mastroianni *et al*., 2008). Thus, the *Xenopus* oocyte combines the advantages of *in vivo* cellular context with the large mutant library screening capabilities for *in vitro* evolution experiments. At the time of these experiments, there were no robust methods for isolating retrohoming events from mammalian cell culture. Thus, the oocyte was an attractive platform for directed evolution in a higher eukaryote.

Here, I demonstrate that the mobile group II intron Ll.LtrB can retrohome and evolve without additional $Mg^{2+}$ within the nucleus of the *Xenopus laevis* oocyte, and I assessed the intron's adaptive fitness landscape at high resolution using deep sequencing.

I first identified a variant of domain V (DV) that improved targeting frequencies within the oocyte. Starting from this variant, I improved a 750-nucleotide version of the Ll.LtrB intron over the course of six cycles of directed evolution with a moderately high mutation rate and subsequently assessed variations using Roche 454 sequencing. In addition to illuminating regions of mutability within the intron, the fitness map identified conserved regions critical for retrohoming and *in vitro* self-splicing, including ~80% of known RNA tertiary contacts. In general, the fitness landscape of the group II intron is rugged, and the 330 nucleotides accumulating mutations are widely distributed. I found five high frequency mutations that were under positive selection and reach from 10% to 60% of the total population. The highest frequency variant was confirmed to improve the retrohoming efficiencies of the group II intron in *Xenopus laevis* oocyte nuclei by 4-10 fold.

## 4.2 Results

### 4.2.1 Ll.LtrB plasmid targeting in the nucleus of *Xenopus laevis* oocytes

It was previously reported that the direct injection of 500 mM $MgCl_2$ into *Xenopus laevis* oocyte nuclei led to Ll.LtrB targeting efficiencies of up to 38% into a recipient plasmid containing the Ll.LtrB target site (Mastroianni *et al*., 2008). The relatively small volume of injected $MgCl_2$ diffuses throughout the oocyte, such that the final approximate concentration approaches 10 mM, the *in vitro* optimum for retrohoming (Mastroianni *et al*., 2008). To assess the targeting frequencies at significantly lower [$Mg^{2+}$], I developed a sensitive Taqman qPCR based assay for detecting both 5'- and 3'-integration junctions of group II intron RNPs that retrohome into plasmid-borne target sites within *Xenopus*

*laevis* oocyte nuclei (Figure 4.1). The assay enables the detection of targeting frequencies as low as $10^{-7}$. In addition to added $MgCl_2$, I provide 17 mM of each dNTP, as this aids with the reverse transcription reaction (Mastroianni *et al*., 2008).

I first assessed retrohoming in oocytes and the $Mg^{2+}$-dependence of a 750-nucleotide wild-type group II intron ΔD4(b1-b3) construct using coinjected $MgCl_2$ concentrations of 500, 200, 100, 50, and 0 mM (Figure 4.2). Targeting efficiencies were robust at 500 mM, resulting in 17% of plasmids containing the 5' junction and 30% of the plasmids containing the 3' junction, similar to retrohoming efficiencies observed in Mastroianni *et. al*. I found that targeting without additional $MgCl_2$ leads to a targeting frequency of $10^{-6}$. Sequencing of the PCR amplified integration product demonstrates that the integrations are precise retrohoming events into the correct target site (data not shown). Furthermore, negative controls using either wild-type RNP injected without plasmid target sites or coinjection of RNPs targeting the human CCR5 gene with the wild-type target site, showed frequencies less than $10^{-7}$ for the wild-type RNA or no qPCR amplification for the CCR5 RNPs (data not shown). Wild-type RNPs injected were routinely used as negative controls.

### 4.2.2 Direct selection for DV variants in *Xenopus laevis* oocyte nuclei

In Chapter 3, I showed that regions within domain V (DV) could be selected to function more efficiently at reduced $Mg^{2+}$ in *E. coli* MgtA mutants (Truong *et al*., 2013). Although I tested the top variants selected in *E. coli* in oocytes (DV14, DV20), the targeting frequencies at all $MgCl_2$ concentrations appeared similar to that of the wild-type intron (data not shown). I reasoned that reselection of DV within the *Xenopus laevis* oocyte nucleus might identify oocyte-specific DV sequences. As done in Chapter 3, I

randomized eleven nucleotides within DV, two base-pairs in the proximal stem between the catalytic triad and Hoogstein sugar/edge, the three terminal base-pairs of the distal stem, and selected this saturation mutagenesis library within oocyte nuclei. The selection uses *in vitro* prepared RNPs (2 x $10^{11}$ RNPs per oocyte), which retrohome into 4 x $10^{9}$ coinjected target plasmids per oocyte without additional $Mg^{2+}$ for 1 h at 37°C. A straightforward PCR isolates mutant RNPs that functioned within the oocyte nucleus, because retrohoming of the RNA into the precise target site is followed by conversion to cDNA via the IEP (Figure 4.1*C*), and primers were designed to amplify the 3'-integration junction and DV region. PCR products generated from the initial selection were sequenced and shown to contain the precise 3'-integration junction as well as unique DV sequences (data not shown). I enriched for high performing variants by performing a total of three selection cycles (Figure 4.3). Cycle 2 showed a diverse set of sequences not observed in Chapter 3 and, by cycle 3, a single sequence DV-XL7 dominated the pool (14 of 17 sequences). Surprisingly, the sequence of DV-XL7 is the same as DV20 found in Chapter 3, except that the second GC pair in the distal stem is flipped (Figure 4.3*B*).

I initially tested the $Mg^{2+}$-dependence of DV-XL7 by coinjecting 500, 200, 100, 50, and 0 mM $MgCl_2$ as done above for wild-type RNPs (Figure 4.4). The targeting frequency at 500 mM $MgCl_2$ reached 46% of plasmids containing the 3' junction, which is higher than that observed for the wild-type RNP (Figure 4.2) (Mastroianni *et al*., 2008). Without additional $MgCl_2$, the targeting frequency approached $10^{-5}$, which is also higher than that observed for the wild-type intron.

To determine the significance of this, I compared DV-XL7 with the wild-type sequence directly on the same day using the same batch of oocytes without additional

$Mg^{2+}$ and found improved targeting frequencies relative to wild-type RNPs. Where the wild-type intron produced targeting frequencies of 2 x $10^{-6}$ 5' junctions and 6 x $10^{-6}$ 3' junctions, DV-XL7 gave frequencies of 3.5 x $10^{-5}$ and 8.5 x $10^{-5}$, respectively. These targeting frequencies give an average 14-fold improvement at both junctions. On the other hand, subsequent repeats of the experiment showed more modest improvements of 2.6- and 4-fold targeting relative to wild-type at the 5' and 3' junctions, possibly due to variability in oocyte batches. The data are presented as two clusters, based on at least two experiments in each cluster, showing a 14-fold increase in cluster 1 and a 4-fold increase in cluster 2 (Figure 4.5*A*). I observed a range in 3' junction targeting frequencies for the wild-type DV of 5 x $10^{-7}$ - 6 x $10^{-6}$ compared to 1 x $10^{-6}$ - 2 x $10^{-4}$ for the DV-XL7 sequence. The variability in the targeting frequencies is likely due to variability of the oocytes, as they are distinct in size and shape and different batches come from genetically different frogs. In addition, oocyte quality is greatly affected by husbandry practice and seasonal variation (Delpire *et al*., 2011). In general, however, I found that DV-XL7 consistently outperforms the wild-type DV RNP.

To confirm the improvement of the DV-XL7 sequence biochemically, I developed an *in vitro* target-primed reverse transcription assay (TPRT) for measuring $Mg^{2+}$-dependent targeting using the plasmid target site and Taqman qPCR probes from the oocyte assay, instead of radiolabelled substrates as done previously (Truong *et al*., 2013). The assay is robust and sensitive to extremely low levels of retrohoming and cDNA synthesis. I tested DV-XL7 relative to the wild-type RNP at 2 mM $MgCl_2$ and at 1.5 mM $MgCl_2$ (Figure 4.5*B*). At 2 mM $MgCl_2$, the rate constants for both RNPs were 0.003 and 0.002 $min^{-1}$, which is in agreement with previous results for wild-type RNPs at

this $Mg^{2+}$-concentration (Truong *et al*., 2013). Although the targeting frequencies for both RNPs are low at ~$10^{-4}$, DV-XL7 increased the amplitude of the reaction by 2-fold relative to the wild-type RNP. At 1.5 mM $MgCl_2$ both RNPs had a *k* values similar to that observed at 2 mM $MgCl_2$, and although the targeting frequencies decreased to ~$10^{-6}$, DV-XL7 had a reaction amplitude 5-fold greater than that of the wild-type RNP. These results are consistent with previous observations showing that mutations in catalytic center DV increase the fraction of active RNP at low [$Mg^{2+}$] (Truong *et al*., 2013).

### 4.2.3 Directed evolution of the full-length Ll.LtrB RNA in *Xenopus laevis* oocytes

To further enhance the activity of the Ll.LtrB intron within *Xenopus laevis* oocyte nuclei, I performed six rounds of directed evolution starting with the variant containing the DV-XL7 catalytic center DV under an average mutation rate of 6 per intron per cycle. To isolate variants that fully retrohomed into the plasmid target-site within the oocyte nucleus, I used a robust PCR amplification strategy where the forward primer sits directly on the 5'-integration junction and the reverse primer sits outside the 3' integration site (Figure 4.1*C*). This amplification strategy showed sensitivity and specificity for targeting frequencies as low as $10^{-6}$. During the selections, I monitored the activity of the pools by qPCR of both junctions. For each cycle, the library size totaled $10^{10}$ unique variants injected in at least 50-fold molar excess to the total available plasmid target sites.

I recovered total pools of $10^5$-$10^6$ mutants at each selection cycle that retrohomed within the oocyte nucleus. However, variability in oocyte batches contributed to a large range in targeting frequencies over the course of the directed evolution cycles (Figure 4.6). Plasmid targeting frequencies at the 3' junction for cycles 1, 2 and 4 averaged 2 x $10^{-4}$, values similar to those of DV-XL7 in cluster 1, which is 14-fold better than that for

the wild-type intron (Figure 4.5*A*). Cycles 3 and 5 showed 3' junction targeting frequencies of 3 x $10^{-6}$, similar to the expected levels for DV-XL7 targeting in cluster 2 (Figure 4.5*A*). By cycle 6, I observed a targeting frequency of $10^{-5}$, a value that did not correlate with either cluster, and as such, I did not proceed further with the selection.

Although cycle 6 has increased targeting frequency relative to that of cycles 3 and 5, it remained lower than cycles 1, 2 and 4. Consequently, I could not infer that the selection cycles were improving the intron, since each selection occurs on different days and with different batches of oocytes. To determine whether the selection cycles led to increased activity, I isolated the introns selected at cycles 2, 4 and 6 and generated RNPs from each post-selected pool. I tested these three pools against the DV-XL7 parental RNP as well as against the wild-type DV RNP without additional $MgCl_2$ (Figure 4.7*A*). In two separate experiments, I found that pool 6 has increased targeting frequencies relative to DV-XL7 by 4-5 fold at the 3' junction. Compared to the wild-type DV RNP, pool 6 shows a 10-30 fold increase in targeting frequency at the 3' junction. Additionally, I tested the activity of the pools compared to DV-XL7 with 50 mM coinjected $MgCl_2$ and although the effect is smaller, I observed increased activity (Figure 4.7*B*). At this $Mg^{2+}$-concentration, DV-XL7 targeting frequency approaches $10^{-5}$ and pool 6 has 3-fold increased 3' junctions. Therefore, based on *Xenopus laevis* targeting frequencies, pool 6 likely contains mutations that have improved the activity of the intron RNP in an environment with low [$Mg^{2+}$].

### 4.2.4 Deep sequencing reveals conserved and mutable intron regions

To determine which nucleotide positions are evolving in the intron, I used Roche 454 deep sequencing to analyze the pools during directed evolution. At the time of these

experiments, 454 provided the longest sequencing read lengths (Mardis, 2008). Although 454 had an average read length of approximately 400 nucleotides, the group II intron has a length of ~750 nucleotides. Therefore, for each selection cycle the intron was sequenced in two separate read fragments of 442 nucleotide (positions 1-442) and 364 nucleotides (positions 386-750). First, I assessed the sequencing error rate of the wild-type intron to determine the feasibility of using the method for assessing low-level mutations found during directed evolution. Substitution errors occurred at a rate of 0.02% per nucleotide, in agreement with the expected rate (Droege and Hill, 2008). However, I also found that insertion and deletion (InDel) errors occurred at frequencies up to 10% near homopolymeric nucleotide regions, although averaged across the whole intron the overall rate is low. Furthermore, this does not complicate the analysis since all InDels are excluded during data processing. Therefore, I found the fidelity of 454 and the long read lengths to be sufficient for analysis of a directed evolution experiment. Based on these results, the threshold for sequence conservation was set at less than 0.3% nucleotide variation at a position, fifteen-fold higher than the substitution error rate. In addition, nucleotide positions equal to or lower than the error rate (0.02%) were considered invariant positions.

I sequenced the selection pools from cycles 2, 4, and 6 and obtained ~15,000 sequencing reads each. I also sequenced the pre-selection mutagenized templates and obtained ~500 reads each (Table 4.1). Notably, I observed a steady increase in the number of average mutations in the pools for cycles 2, 4, and 6. The initial pre-selection template from cycle 2 began with ~6 mutations per intron on average, but the selected cycle 2 population only retains 3 mutations per sequence. This result suggests selection

against the majority of mutations within the intron during early stages of *in vitro* evolution. However, beginning at cycle 4, I observed a pre-selection mutagenic template rate of ~8 mutations that fixate into the pool as 4.8 mutations per intron, and by cycle 6, I observed 9.6 mutations in the template and 7.5 mutations fixating into the population. The steady increase in average mutation number within the selected pools indicates significant new sequence variation, and may indicate increased robustness of the genetic pool to high mutation rates over time.

Based on the sequencing data, I generated a composite fitness heat map of the evolving group II intron through six cycles of evolution without additional $Mg^{2+}$ in the *Xenopus laevis* oocyte nucleus (Figure 4.8). Nucleotides that never rose above a 0.3% mutation rate through six cycles are highlighted in blue shading as conserved nucleotides. These positions may indicate essential structural elements used during splicing and retrohoming. Nucleotides with the highest frequency of mutation in cycle 6 and shaded in red. Many of these positions fixed upon a single nucleotide type, as opposed to a distribution between all three alternative nucleotides, and may indicate positive selection. Therefore, I defined positive selection as those positions in which >80% of the mutation distribution is for one nucleotide type. Those positions undergoing positive selection at high frequency (>2% of the population) are indicated with a green arrow containing the nucleotide to selection at that position.

For nucleotides that remained conserved, the fitness map appears striking. The vast majority of these sites remained conserved (<0.3%) throughout all six cycles, providing evidence for an important role during intron function. In general, I found that 416 nucleotides out of a possible 746 do not change throughout the six cycles of

evolution. Notably, I found that 80% of nucleotides found in known tertiary contacts remain conserved on the fitness map (Table 4.2)(Dai *et al*., 2008). As expected, the base of catalytic core DV remains highly conserved, but the terminal three base-pairs tolerate mutations to a modest degree (0.3-1% variation), which is consistent with my previous findings (Truong *et al*., 2013). A few other regions not known to have a structural role also appear conserved, such as DI(iiia) which is possibly a protein binding region (Dai *et al*., 2008).

Approximately 330 nucleotide positions had mutation frequencies of >0.3%, and those positions showing both high frequency and positive selection are scattered across the intron, with few of these nucleotides concentrating in any particular region (Figure 4.8). Regions with relatively high mutability (1-10%) but no obvious positive selection include: the central hub in DI, the first single stranded loop in DII, DIVb, and sections in DVI. These regions may possibly have relaxed roles during intron function and thus, accumulate mutations at near neutral rates. On the other hand, high frequency positively selected mutations likely contribute to increased group II intron function in the *Xenopus laevis* oocytes. For instance, nucleotide G282 located in EBS1 is positively selected towards an A nucleotide, a change that converts the G-T RNA-DNA non-canonical base-pair to a canonical A-T and has previously been shown to give 50% higher retrohoming *in vitro* (Mohr *et al*., 2000).

Five notable positions were under positive selection and possessed high mutation frequencies, reaching from 10% to 60% of the total population, and were increasing throughout all six cycles (Figure 4.9). These five candidates likely contribute the greatest to the higher activity in the intron pools observed in Figure 4.7*A*. These five mutations

are located near the 5' and 3' ends of the intron, and because of the limitations in 454 read length, not all five mutations could be correlated to each other. The three highest frequency mutations are in the first sequence read: A84G (25%), G106A (20%), and C111U (59%), and the other two mutations G694A (10%) and C745U (11%) are in the second sequence read. First, I determined the percentage frequency in which the combinations of these mutants co-occur throughout the course of the six cycles (Table 4.3). The combination of A84G, G106A, and C111U has been rising 40-fold every 2 cycles to reach 2.8% of the population. Intriguingly, the three mutations are located near the λ and ε contacts, which coordinate this region with the 5' end of the intron and DV. The combination of G694A and C745U remained at a low frequency (0.7%) by cycle 6, and they were not detected in earlier cycles. Although the combination mutants are candidates for increased activity, they may be increasing in the population due to genetic drift.

To determine the rate of covariation between mutation pairs amongst the top five mutations, I performed standard linkage disequilibrium calculations (Table 4.4)(Hayden *et al*., 2011). The combinations of 84G-111U and 106A-111U, both had *D'* values near zero, indicating that these mutations act independently of each other. The combinations of 84G-106A and 694A-745C both had negative *D'* values (-0.14 and -0.46), suggesting negative selection against these combinations. Thus, none of the mutation pairs show evidence for acting synergistically. However, the lack of quantifiable covariation between mutants could be due to a high degree of recombination occurring at some step during sequencing. Recombination can occur during PCR of mixed populations of molecules, generating what are called PCR chimeras, and PCR chimera formation can occur in up to

50% of the total population due to high cycles of PCR (Lahr and Katz, 2009; Shao *et al*., 2013). As Roche 454 sequencing utilizes high cycles of PCR for both sequencing and generating templates, the formation of molecular chimeras may have obscured covariation between mutations.

### 4.2.5 The C111U mutation increases targeting frequencies in *Xenopus laevis* oocytes

In spite of these complications, mutations need not have strong pairwise covariation rates to improve activity. Hayden *et. al*., found four mutations that did not covary in pairs, but together improved the activity of the *Azoarcus* group I intron ribozyme when evolved for splicing in formamide (Hayden *et al*., 2011). To determine the effects of some of the candidate single mutations and combined mutations on intron integration frequencies in oocytes, I generated RNPs of the three most frequent mutants and a number of combinations from Table 4.4 for testing in *Xenopus laevis* oocytes. The results from these experiments are shown in Figure 4.10, where I performed two experiments without coinjected $MgCl_2$ and an additional experiment with 50 mM coinjected $MgCl_2$. Without additional $MgCl_2$ most of the individual mutations and combination mutations performed poorly relative to DV-XL7 in the oocytes. However, both C111U individually and the combination of C111U-G106A performed between 4-10 fold better than DV-XL7 between the two experiments. In contrast, with 50 mM coinjected $MgCl_2$, many mutation combinations performed no better than or had decreased activity relative to DV-XL7, although it should be noted the selections were not performed under this condition. Since C111U is the highest frequency mutation in the population (59%), these data suggest it underlies the increased activity found in pool 6 (Figure 4.7).

## 4.3 Discussion

In this Chapter, I showed that group II intron retrohoming within the *Xenopus laevis* oocyte nucleus occurs without additional $MgCl_2$, although at very low frequencies. Activity rises progressively with the addition of $MgCl_2$, reaching a targeting efficiency of ~20% with 200 mM $MgCl_2$ and ~30% with 500 mM $MgCl_2$. These results suggest that efficient group II intron retrohoming within higher eukaryotic cells relies predominantly upon the free [$Mg^{2+}$]. As mammalian cells and *Xenopus laevis* oocytes contain very low free [$Mg^{2+}$] of ~0.2-1 mM (Horowitz and Tluczek, 1989; Gunther, 2006), I selected a version of DV optimized for the basal [$Mg^{2+}$] found within *Xenopus laevis* oocytes, and this variant had increased targeting frequencies relative to wild-type. Starting with this intron variant, I performed an additional six rounds of directed evolution on the full-length intron until I observed improvement in targeting frequencies. I mapped the entire fitness landscape of mutations found in the Ll.LtrB RNA during the selection cycles in the oocytes using Roche 454 sequencing. The fitness map generally agrees with previous biochemical work on conserved regions of group II introns, and provides an important comprehensive view of intron structure and function. Finally, I show that the highest frequency mutation, C111U, performed better than the parental DV-XL7 variant in *Xenopus laevis* oocyte microinjection assays.

The fitness map identifies regions important for both *in vitro* self-splicing of the intron at high [$Mg^{2+}$] (50 mM) and retrohoming in *Xenopus laevis* oocytes at very low [$Mg^{2+}$]. To determine the relationship between all regions identified, I displayed the mutation frequencies on a three-dimensional model of the Ll.LtrB intron (Figure 4.11)

(Dai *et al*., 2008). Tertiary contacts are generally highly conserved (Table 4.2) with many nucleotide positions possibly invariant as they had mutations less than the observed 454 substitution error rate (0.02%). Conserved tertiary contacts include those important for coordinating DV with DI and the central hub: λ-λ', k-k ', the catalytic triad to J2/3, and the Hoogstein/sugar-edge to DI(i). Conserved intradomain tertiary contacts within DI include α-α' and β-β', and the interdomain contacts of θ-θ' from DI to DII, μ-μ' from DV to DIII, and η-η' from DII-DVI. As expected the EBS1/δ and EBS2 regions were highly conserved as they are important for both splicing and target-recognition during retrohoming. The consistency by which the main tertiary contact remained conserved indicates the utility of using mutagenesis and deep sequencing for identifying important regions in RNA structure. Based on these observations, I also identified new regions potentially important for structure function, such as DI(iii)a, a region which possibly interacts with the LtrA protein (Dai *et al*., 2008).

Although only a handful of nucleotide position variations rose to high frequency, greater than 200 sites had mutation percentages between 0.3 and 10. The most mutable regions of the intron include single-stranded loop regions. The single-stranded loop in DII contains high mutation percentages across every nucleotide of this region (Figure 4.11; orange outline), and based on the Dai model (Dai *et al*., 2008) is a region on the outer surface of the intron that possibly interacts with LtrA. The first major hub in DI, after the I(i) loop, along with the ζ acceptor also had high mutability. As expected DIVb, a region not known to serve a structural or catalytic function, had overall high mutability. Although DIVb is not shown on the structure (Figure 4.11), as it was not included in the (Dai *et al*., 2008) model, it presumably is also an external loop. Interestingly, DVI, which

contains the invariant branch-A nucleotide, had numerous mutable regions situated predominantly at single-stranded loops. The first X-ray crystal structures of the *O. iheyensis* IIC intron lacked DVI, as it did not pack uniformly, possibly due to flexibility (Toor *et al*., 2008a).

Five mutations rose to high frequency towards a single nucleotide each, suggesting positive selection. When I tested their respective targeting frequencies in *Xenopus laevis* oocytes relative to the parental DV-XL7 variant only C111U and C111U plus G106A improved activity, although only without additional coinjected $MgCl_2$. Deep sequencing offered the possibility of identifying combinations of mutations at low frequency that cooperatively act to increase activity. However, none of these combinations identified by my analysis proved better than C111U alone. Although the other mutations did not improve activity, they may serve other functions during directed evolution. For instance, they could increase robustness to mutagenesis, thereby permitting mildly detrimental mutations to accumulate at near neutral levels. In contrast to my results, an earlier study that selected the *Tetrahymena* group I intron ribozyme to function using calcium *in vitro*, showed that, in general, seven high frequency mutations functioned cooperatively (Lehman and Joyce, 1993). Furthermore, a recent study by (Hayden *et al*., 2011) found that the *Azoarcus* group I intron ribozyme had four mutations (*Azo** variant) that worked cooperatively to increase splicing in formamide. It is possible that a technical limitation, such as PCR recombination, may have obscured my ability to identify cooperative sets of mutations.

Although the C111U mutation improved activity in *Xenopus leavis* oocytes, our lab developed a group II intron expression system for human cell culture, which I show in

Chapter 5 permits targeting within that cellular environment. This major advance encouraged me to perform similar directed evolution studies and next generation sequencing analysis as described in this chapter, in the more medically relevant human cell culture system. It is of interest to know whether the conserved and mutable regions identified for retrohoming within *Xenopus laevis* oocytes correspond to those evolved for human cell culture.

## 4.4 Methods

### 4.4.1 Materials and plasmids

The following antibiotics were used: ampicillin (100 μg/ml) and gentamicin (10 μg/ml). All PCRs were performed using GC-rich Phusion polymerase (New England Biolabs) unless otherwise indicated. The recipient plasmid for all group II intron targeting experiments was pLHSQ, a derivative of pLHS (Matsuura *et al*., 1997) containing the wild-type Ll.LtrB target site followed by sequences with optimal $T_m$ for PCR and qPCR.

### 4.4.2 Library construction and selection

The mutagenized DV library was generated using site-randomized oligonucleotides comprising the DV region purchased from Integrated DNA Technologies. The randomized oligonucleotide was used as a primer to amplify the 3' half of the intron, and then subsequently regenerated into the full-length intron by overlap extension PCR with a 5' fragment of the intron. The full-length product was then appended with a T3-promoter for use in *in vitro* transcription.

Libraries for each cycle of directed evolution were generated using Mutazyme II (Stratagene) according to the manufacturer's recommendations for a high mutagenesis rate. Approximately 10 ng of linear DNA template (800 nucleotides) was mutagenized in a 50 µl PCR for 30 cycles. The resultant product (~5 x $10^{10}$ variants; 500 ng) was then re-amplified and appended with a T3-promoter by PCR before use in *in vitro* transcription and subsequent RNP preparation. After each round of selection, the *Xenopus laevis* oocyte selected intron plasmid integrations are amplified using a primer that sits directly on the 5' junction and one that sits downstream of the 3' exon. Primers were 176S, 5'-CATCCATAACGTGCGCC and 73S, 5'-AATGGACGATATCCCGCA. The PCR selected pool was then used as the template for subsequent rounds of mutagenic PCR.

### 4.4.3 Preparation of LtrA protein, *in vitro* transcription, and RNPs

The LtrA protein was expressed from plasmid pIMP-wild-type in Arctic express cells (Stratagene) following the manufacturer's recommendations (Truong *et al*., 2013). Briefly, 3 to 5 freshly transformed colonies were grown to stationary phase overnight in ampicillin and gentamicin antibiotics. The following day, cells were subcultured at 1:100 at 30°C without antibiotics for 3 h, followed by induction with 1 mM IPTG at 12°C for 24 h with shaking at 250 rpm. The LtrA protein was then purified via chitin affinity-column chromatography and subsequently concentrated in 50% glycerol by dialysis as previously described (Saldanha *et al*., 1999), except that prior to loading the column, nucleic acids in the cell supernatant were precipitated using 0.4% polyethylenimine (PEI) under constant stirring for 1 h and then subjected to centrifugation.

All RNA for generating RNPs was transcribed from a linear PCR product appended with a T3 promoter sequence. Transcription was performed overnight using an

Ambion T3 Megascript kit with 100 ng PCR template per 1x (20 µl) reaction. Post transcription, RNA was treated with Turbo DNase for 30 min and then extracted twice with acid phenol-chloroform (pH 4.8) to remove residual DNA and then once with chloroform. RNA was precipitated using 3 volumes of 8 M lithium chloride and washed twice with 70% ethanol. To remove small RNAs, the pellet was resuspended in water, reprecipitated with 1/3 volume 7.5 M ammonium acetate solution, and washed twice with 70% ethanol.

To generate lariat RNA, the RNA was *in vitro* self-spliced as previously described (Saldanha *et al*., 1999; Truong *et al*., 2013). RNP complexes were reconstituted by incubating renatured lariat RNA with a 6-fold molar excess of LtrA protein at 30°C and concentrated by ultracentrifugation at 40,000 x g as previously described (Mastroianni *et al*., 2008; Truong *et al*., 2013). RNPs were then resuspended in 5 mM $MgCl_2$, 10 mM KCl, and 40 mM HEPES (pH 8). Prior to injection, RNPs were heated at 37°C for 10 seconds.

### 4.4.4 Plasmid targeting in *Xenopus Laevis* oocyte nuclei and TPRT assays

*Xenopus laevis* oocyte nuclear targeting was based on previously published assays (Mastroianni *et al*., 2008; Zhuang *et al*., 2009b). *Xenopus laevis* oocytes were freshly isolated from hormone-induced females and then manually defollicated under a dissecting scope. Isolated oocytes were then treated with collagenese (1 mg/ml) in Barth's solution (Truong *et al*., 2008) for 10 min and then stored at 18°C in Barth's solution until use, although for no more than 2 h. Each oocyte was then injected with 18 nl of a solution containing 278 ng/µl pLHSQ plasmid, 17 mM of each dNTP, 17 mM $MgCl_2$, and additional $MgCl_2$ as specified. Following this, the oocytes were then injected with 18 nl

of a solution containing 3-5 μg/μl of RNPs. On average, each oocyte receives 5 ng of plasmid targets (4 x $10^9$ molecules) and 54-90 ng RNP (1-2 x $10^{11}$ molecules). For targeting experiments, batches of 6-8 oocytes for each replicate (typically four), were injected and incubated at 37°C for 1 h in a 1.5-ml snap tube. For the selection cycles 50-60 oocytes were used. To extract the plasmid DNA, oocytes were immediately vortexed in lysis buffer using a Qiagen DNeasy Blood and Tissue Kit and incubated overnight at 55°C. The following day, the sample was treated with 400 μg of RNaseA for 30 min at room temperature, and then purified according to manufacturer's recommendations. The isolated plasmid and oocyte genomic DNA was used for PCR amplification of selected retrohoming or Taqman qPCR.

For TPRT experiments, a 5-fold molar excess of RNPs were incubated with 0.25 ng/μl pLHSQ plasmid target sites in a buffer containing 10 mM KCl, 50 mM Tris-HCl (pH 7.5), and 5 mM diothiothreitol (DTT), along with 0.2 mM of each dNTP and $MgCl_2$ as indicated at 37°C for various times. Aliquots at the indicated time point were quenched in 5 mM EDTA and heated to 95°C for 5 min to stop the reactions. Samples were then diluted 20-fold into water and analyzed using Taqman qPCR for each junction in technical triplicates.

### 4.4.5 Taqman qPCR

Quantitative PCR was performed using the Applied Biosystems Viia7 system in 384-well format using Taqman probes (Yao *et al*., 2013). Reactions were performed in technical triplicate in 10 μl volumes for 40 cycles using Taqman PCR universal mastermix under standard conditions. Standard curves for absolute quantification used four 10-fold dilutions of a TOPO2.1 plasmid containing an Ll.LtrB integration in the

wild-type target site and had >90% efficiency across the range of concentrations used. Dilutions were buffered in carrier DNA of ~5 ng/μl λ-virus DNA. The primer/probe sets were:

(i) 5'-integration junction.

Taqman probe 198S-Q10 5'-FAM-ATCCATAACGTGCGCCCA-MGB;

forward 200S 5'-CCGCTCTAGAACTAGTGGATCCA;

reverse 201A 5'-TCGGTTAGGTTGGCTGTTTTCT;

(ii) 3'-integration junction.

Taqman probe 189S-Q1 5'-FAM-CTACTTCACCATATCATTTT-MGB;

forward 197S 5'-AAGAGGGTGGTGCAAACCAG;

reverse 191A 5'-AATGGACGATATCCCGCAAG;

(iii) ampicillin marker.

Taqman probe P081 5'-FAM-TGTCACGCTCGTCGTTTGGTATGGC-BkFQ;

forward P079 5'GCCATTGCTACAGGCATCGT;

reverse P080 5'-GGGAACCGGAGCTGAATGA.

Taqman probes with a 5'-FAM (6-carboxyfluorescien) and a 3'-MGB (dihydrocyclopyrroloindole tripeptide major groove binder) were obtained from Applied Biosystems and those with a 5'-FAM and 3'-BkFQ (Iowa Black FQ) from Integrated DNA Technologies.

### 4.4.6 Deep sequencing and data analysis

Deep sequencing of the group II intron selections was performed using the Roche 454 GS FLX+ titanium chemistry at the genomic sequencing and analysis facility at UT-Austin. Due to read length limitations, the intron was sequenced in two separate reads of

442 and 364 nucleotides each (positions 1-442 and 386-750, respetively), and both overlap by 40 nucleotides using a 20 nucleotide constant region each. The second fragment gave significantly fewer reads possibly due to greater numbers of RNA hairpins, and this was adjusted for by increasing the concentration of the second fragment by 2-fold. Additionally, nucleotide regions near the end of the intron gave significantly fewer reads and, therefore, have less coverage depth (Table 4.1). The adaptors (A and B) used for sequencing the fragments were generated by PCR of the selection cycles according to manufacturer's recommendations, and for each cycle also included a unique six-nucleotide barcode sequence for computationally extracting sequences from multiplexed samples.

Raw sequence reads in the FastQ file format were aligned to the wild-type reference sequence using Mosaik Aligner 1.0 (https://code.google.com/p/mosaik-aligner/), and text files were extracted using the Tablet browser (Milne *et al*., 2010). Sequence gaps were removed using a Perl script, Gapstreeze, available online at (http://www.hiv.lanl.gov/content/sequence/GAPSTREEZE/gap.html). Aligned sequences were then analyzed for nucleotide variation using a Perl script courtesy of Dr. Scott Hunicke-Smith. All other data analyses, including calculation of nucleotide frequencies and mining of covariation, were performed using Unix shell scripts including: grep, cut, uniq, sort, and awk. Standard linkage disequilibrium was calculated as $D=(P_{AB} \times P_{ab})-(P_{Ab} \times P_{aB})$, where $P_{AB}$ is the frequency in which the mutations occur together, $P_{Ab}$ and $P_{aB}$ are the mutations occurring independently, and $P_{ab}$ the frequency where neither occurred. The normalized linkage disquilibrium ($D'$) was calculated by dividing positive $D$ values by the theoretical maximum co-occurrance and negative $D$ values by a theoretical

minimum co-occurrence based on the observed individual frequencies in the population (Hayden *et al*., 2011).

Table 4.1: Roche 454 sequence read numbers and average mutations per cycle.

| Type | Reads per base, range (ave per base) | Ave mutations per 746 nt (intron) |
|---|---|---|
| Cycle 2 template | 150-530 | 5.9 |
| Cycle 2 selection | 6773-35,701 (~20,000) | 3 |
| Cycle 4 template | 84-603 | 8 |
| Cycle 4 selection | 2023-28,690 (~15,000) | 4.8 |
| Cycle 6 template | 100-802 | 9.6 |
| Cycle 6 selection | 2099-30346 (~15,000) | 7.5 |
| Wildtype | 100-1480 | 0.1 |

The range in number of reads refers to the number of total base calls, which was higher at the 5' intron end than at the 3' intron end.

Table 4.2: Comparison of known tertiary contacts to fitness map conservation.

| Tertiary contact / motifs | Conserved nts (<0.3% mutability) | Positive nts selection |
|---|---|---|
| α-α' | 10/10 | 0 |
| θ-θ' | 10/15 | 0 |
| λ-λ' | 5/5 * | 0 |
| ε-ε' | 2/2 * | 0 |
| β-β' | 6/7 | 0 |
| ζ-ζ' | 8/16 | 0 |
| κ-κ ' | 6/7 | 1 |
| EBS2 | 5/5 | 0 |
| EBS1/δ | 7/8 | 1 |
| η-η' | 6/8 | 2 |
| μ-μ' | 2/2 | 0 |
| ϒ-ϒ' | 2/2 | 0 |
| AGC triad | 3/3 | 0 |
| branch A | 1/1 | 0 |

The table shows major tertiary contacts, and the number of nucleotides that were conserved in the fitness map of cycle 6 from Figure 4.8. The right column shows nucleotides that were under positive selection, as >80% of that position mutated to a single nucleotide type. The tertiary contacts shown here are taken from (Dai *et al*., 2008). * One nucleotide (nt) from the λ and two from ε were not mutagenized as these regions were used as priming sites for the PCR selection step seen in Figure 4.1*C*.

Table 4.3: Population frequency and fold change of mutations and combination of mutations.

| Genotypes | Cycle 2 | | Cycle 4 | | Cycle 6 | |
|---|---|---|---|---|---|---|
| **First 442 nts** | % pop. | fold Δ | % pop. | fold Δ | % pop. | fold Δ |
| A84, G106, C111 (wild-type) | 90.5 | - | 72.4 | 0.8 | 24.1 | 0.3 |
| 84G | 5.3 | - | 9.7 | 1.8 | 7.6 | 0.8 |
| 106A | 1.1 | - | 4.3 | 3.9 | 7 | 1.6 |
| 111U | 3.0 | - | 11.1 | 3.7 | 36.4 | 3.3 |
| 84G, 106A | 0.03 | - | 0.5 | 16.4 | 2.4 | 4.8 |
| 84G, 111U | 0.15 | - | 1.6 | 10.4 | 12.8 | 8.2 |
| 106A, 111U | 0.02 | - | 0.5 | 25.4 | 6.9 | 15.4 |
| 84G, 106A, 111U | 0.0017 | - | 0.1 | 40.9 | 2.8 | 39.8 |
| **Second 384 nts** | | | | | | |
| G694, C745 (wild-type) | 98.7 | - | 95 | 1.0 | 78.1 | 0.8 |
| 694A | 0.7 | - | 1.8 | 2.7 | 10.2 | 5.6 |
| 745U | 0.7 | - | 3.2 | 4.8 | 11 | 3.4 |
| 694A, 745U | 0.00 | - | 0.00 | - | 0.7 | - |

The table shows the five high frequency positively selected mutations either alone or in combination with the others as a percentage of the population at cycles 2, 4 and 6. For all five mutations and their combinations, they rose rapidly through the population, as shown by the fold Δ in the population between the cycles examined. The fold Δ value refers to the fold increase in frequency from the previously examined cycle.

Table 4.4: Standard linkage disequilibrium between mutation pairs.

| Mutation combination | *D* | *D'* |
|---|---|---|
| 84G, 106A | -0.016 | -0.14 |
| 84G, 111U | 0.003 | 0.01 |
| 106A, 111U | 0.002 | 0.02 |
| 694A, 745U | -0.006 | -0.46 |

The table shows calculations for standard linkage disequilibrium (*D*) between the five top variants, and the normalized linkage disequilibrium (*D'*) (see methods for calculations). The value for *D* and *D'* can take on positive or negative values, indicating whether the mutations co-occur frequently or infrequently compared to the expected rate from their individual frequencies. Values close to zero indicate linkage equilibrium between the two mutations.

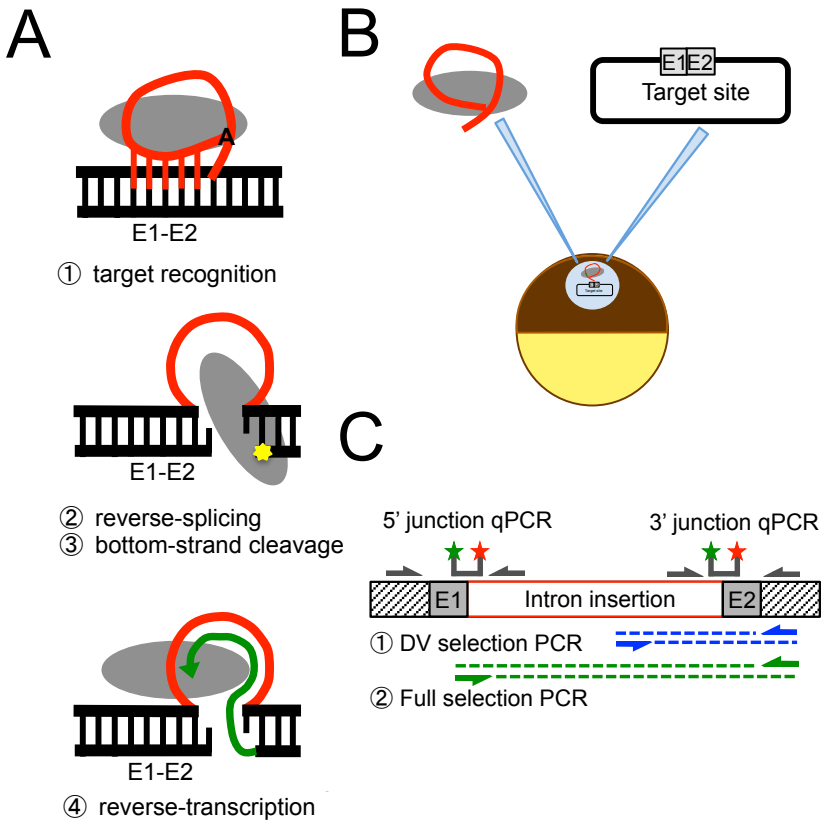



Figure 4.1: Overview of Ll.LtrB group II intron retrohoming within *Xenopus laevis* oocytes, selection, and qPCR analysis.

(*A*) Binding of the Ll.LtrB RNP to the target site via base-pairing of EBS1, EBS2, δ, and limited protein contacts leads to reverse splicing of the RNA directly into the top-DNA strand. At the same time, the protein LtrA generates a single strand nick (yellow star) on the opposing DNA strand, and this break is used as a priming site for reverse transcription of the RNA into a cDNA copy. (*B*) Fresh, manually defollicated *Xenopus laevis* oocytes are microinjected using two separate needles, directly into the nucleus located in the brown, animal pole. The first needle contains 276 ng/µl of plasmid target, 17 mM dNTPs, and additional $MgCl_2$ as required. The second needle contains 54-90 ng/µl RNP, and then the oocytes are incubated at 37°C for 1 h. (*C*) The cDNA copy generated from successful retrohoming (*A*) can be detected using Taqman qPCR probes overlapping the 5' and 3' junctions. (1) The blue arrows and dashed lines indicate the primers used to amplify and select DV mutants during saturation mutagenesis relative to the intron integration site. (2) The green arrows show the location of primers used during selection of the full-length intron.

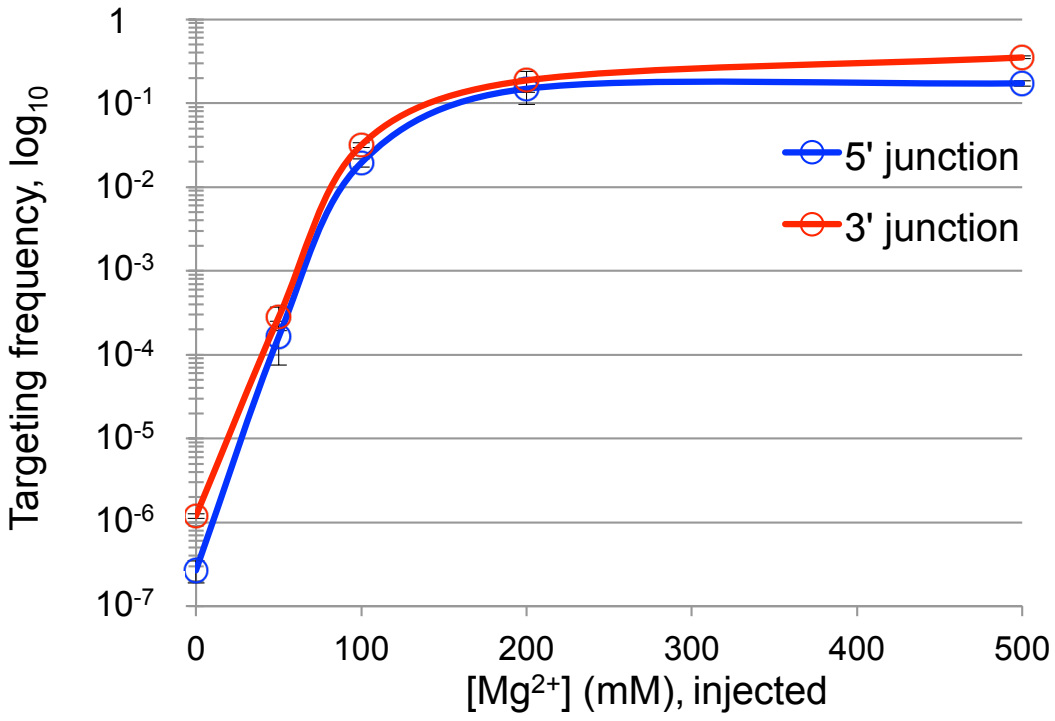



Figure 4.2: $Mg^{2+}$-dependence of wild-type group II intron plasmid targeting within *Xenopus laevis* oocytes analyzed by Taqman qPCR.

Wild-type (WT) Ll.LtrB RNPs were prepared *in vitro*, as previously described (Truong *et al*., 2013) and microinjected into *Xenopus laevis* oocyte nuclei along with plasmid target sites and various $MgCl_2$ concentrations as indicated. After 1 h incubation at 37°C, total DNA was extracted from the oocytes, and targeting frequency at the 5' and 3' junctions were determined by Taqman qPCR. The data points shown are the average of three replicate oocyte batches (8 oocytes each), and the error bars are SEM. The targeting frequencies are shown on a $\log_{10}$ scale. A negative control consisting of injecting RNPs without plasmid target sites gave background level amplification of $<10^{-7}$.

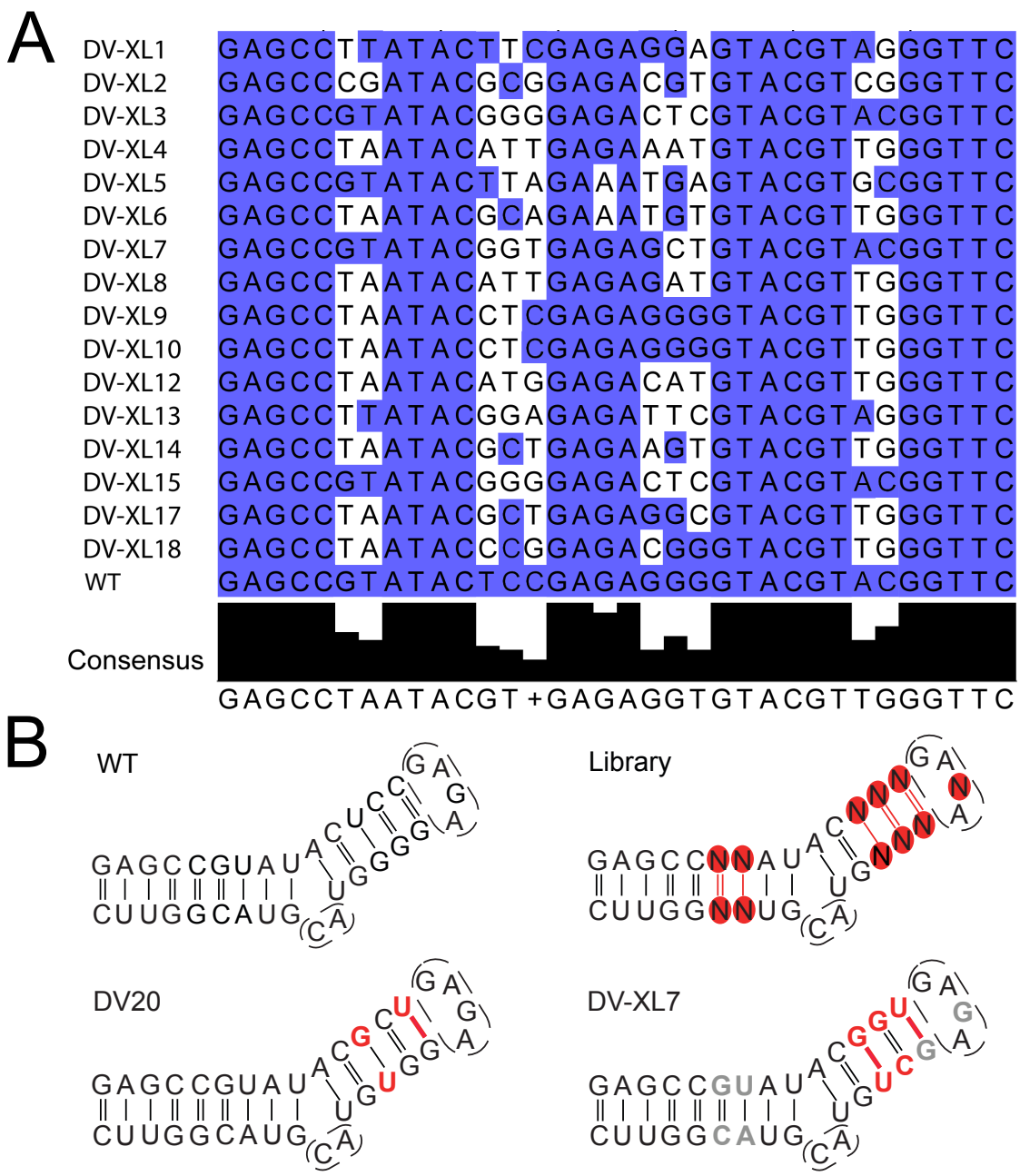



Figure 4.3: Saturation mutagenesis and selection of DV within *Xenopus laevis* oocyte nuclei.

A saturation mutagenesis library was generated by randomizing nucleotides in DV as previously reported (Truong *et al*., 2013), made intro RNPs, and microinjected into *Xenopus laevis* oocytes for selection, and recovered by PCR using the scheme described in Figure 4.1*C*. After an additional round of selection, variants were PCR amplified, cloned into a TOPO-blunt vector, transformed into *E. coli*, and individual colonies were isolated for sequencing. (*A*) The alignment shows sequences identified after two rounds of selection. Nucleotides matching that of the wild-type sequence are highlighted in blue. The percentage identity is shown at the bottom along with the consensus sequence. After the third selection round, sequence 7 (DV-XL7) overtook the pool and present in 14 of 17 clones. (*B*) The secondary structure of wild-type DV is shown along with nucleotides that were randomized, highlighted in red ovals. Secondary structure of DV-XL7 and DV20 are shown, which resemble each other in sequence. DV20 has previously been shown to improve retrohoming in low $Mg^{2+}$ *E. coli* (Truong *et al*., 2013).

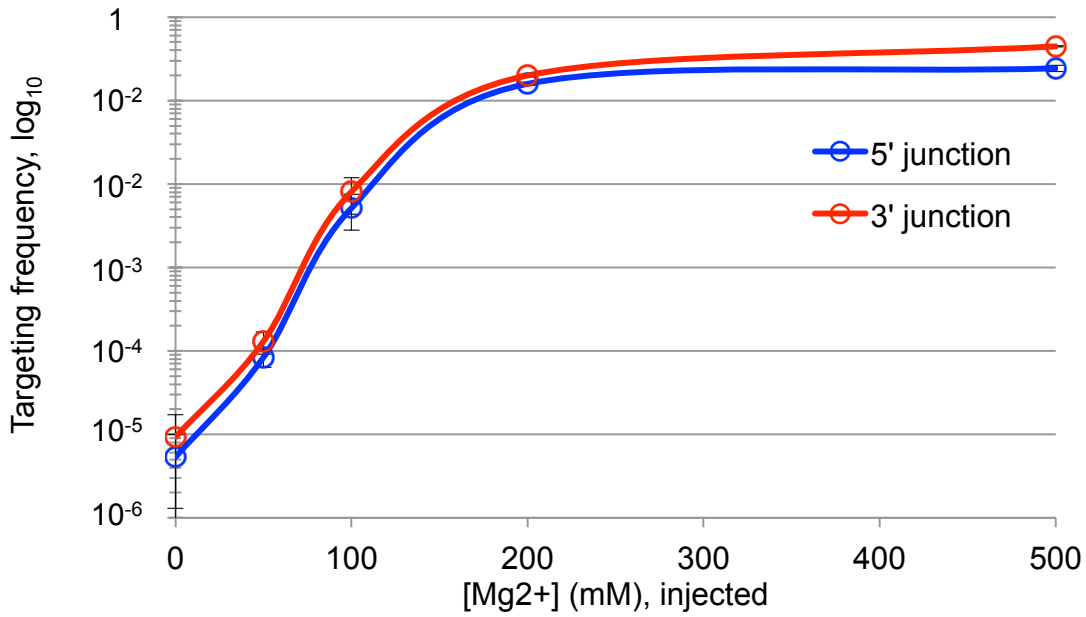



Figure 4.4: $Mg^{2+}$-dependence of DV-XL7 in plasmid targeting within *Xenopus laevis* oocytes analyzed by Taqman qPCR.

DV-XL7 RNPs were prepared *in vitro* and microinjected into *Xenopus laevis* oocyte nuclei along with plasmid target sites and various $MgCl_2$ concentrations as indicated. After 1 h incubation at 37°C, total DNA was extracted from the oocytes, and targeting frequencies at the 5' and 3' junctions were determined by Taqman qPCR. The data points shown are the average of three replicate oocyte batches (8 oocytes each), and error bars are SEM. The targeting frequencies are shown on a $\log_{10}$ scale. A negative control consisting of injecting RNPs without plasmid target sites gave background level amplification.

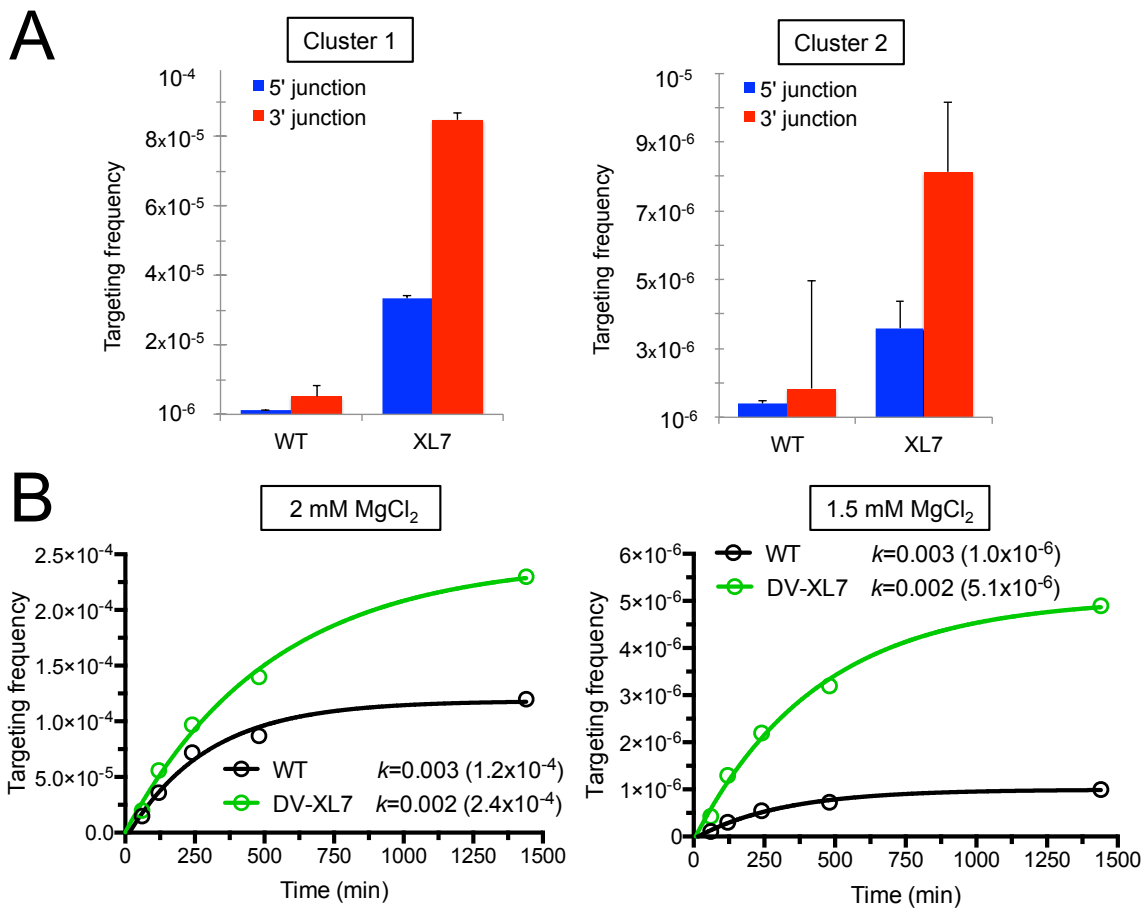



Figure 4.5: DV-XL7 enhances plasmid targeting in *Xenopus laevis* oocytes and under low [$Mg^{2+}$] *in vitro*.

(*A*) 5'- and 3'- targeting frequencies of DV-XL7 RNPs in *Xenopus laevis* oocytes without additional $MgCl_2$ compared to wild-type RNPs. The experiments fell into two separate clusters. Cluster 1 is the average of two experiments in which DV-XL7 performed 14-fold better than wildtype at both junctions. Cluster 2 shows the average of two experiments with 2- and 4-fold enhanced targeting frequencies at the 5' and 3' junctions, respectively. Standard deviations are shown for each data point. For each experimental sample, a negative control was included that showed background amplification levels, and consisted of injecting RNPs without plasmid target sites followed by Taqman qPCR analysis. (*B*) DV-XL7 RNPs were compared to wild-type RNPs in a target-primed reverse transcription assay (TPRT) under *in vitro* conditions with 2- and 1.5-mM $MgCl_2$ in a time course for 24 h. A 5-fold molar excess of RNPs were incubated against plasmid target sites along with dNTPs and $MgCl_2$ in a buffer containing 10 mM KCl, 50 mM Tris-HCl (pH 7.5), and 5 mM DTT at 37°C. Aliquots at the indicated time point were quenched in 5 mM EDTA and heated to 95°C for 5 min. Samples were then diluted 20-fold into water and analyzed using Taqman qPCR for each junction. The rate constant ($k$, $min^{-1}$) is shown with the amplitude in parentheses. RNPs incubated without $MgCl_2$ for 24 h had background amplification equivalent to $10^{-8}$ events.

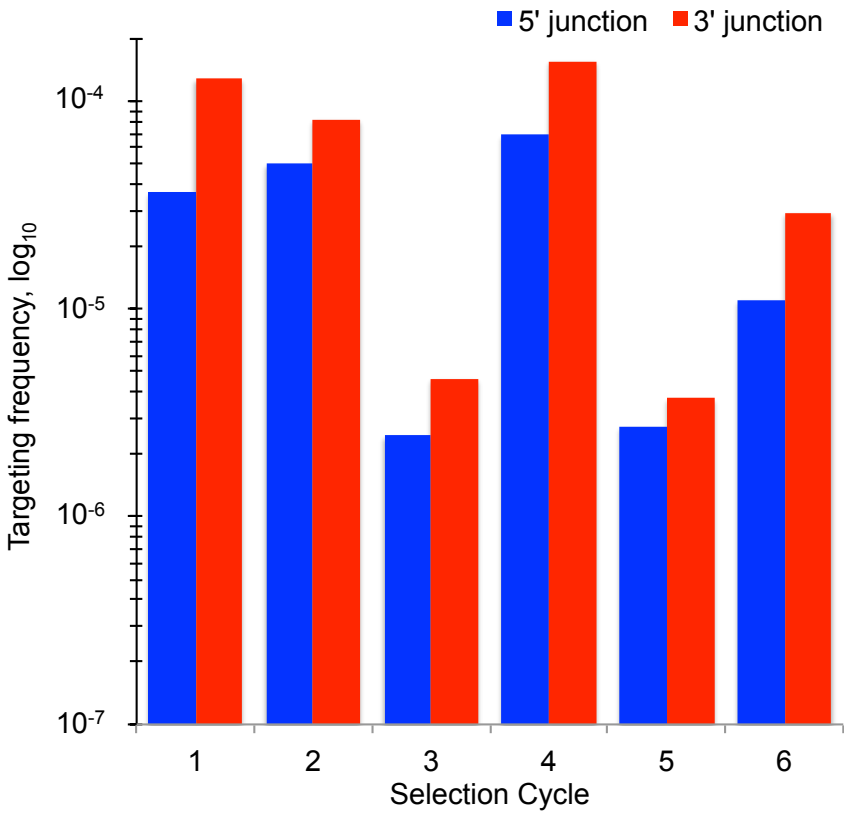



Figure 4.6: Directed evolution of the full-length group II intron over six cycles in *Xenopus laevis* oocytes.

The full-length intron was mutagenized at a rate of ~6 mutations per intron using Mutazyme II PCR, made into RNPs, and selected within oocytes as described in Figure 4.1. Each selection used 50-60 microinjected oocytes. Further mutations, at a rate of 6 per intron, were done between each round of selection. Targeting frequencies for the 5' and 3' junctions were determined by Taqman qPCR, are shown for all six cycles on a $\log_{10}$ scale. Cycles 1, 2 and 4 had high targeting frequencies consistent with that found for DV-XL7 RNPs found in cluster 2 of Figure 4.5*A*. RNPs injected without plasmid target sites showed background amplification levels.

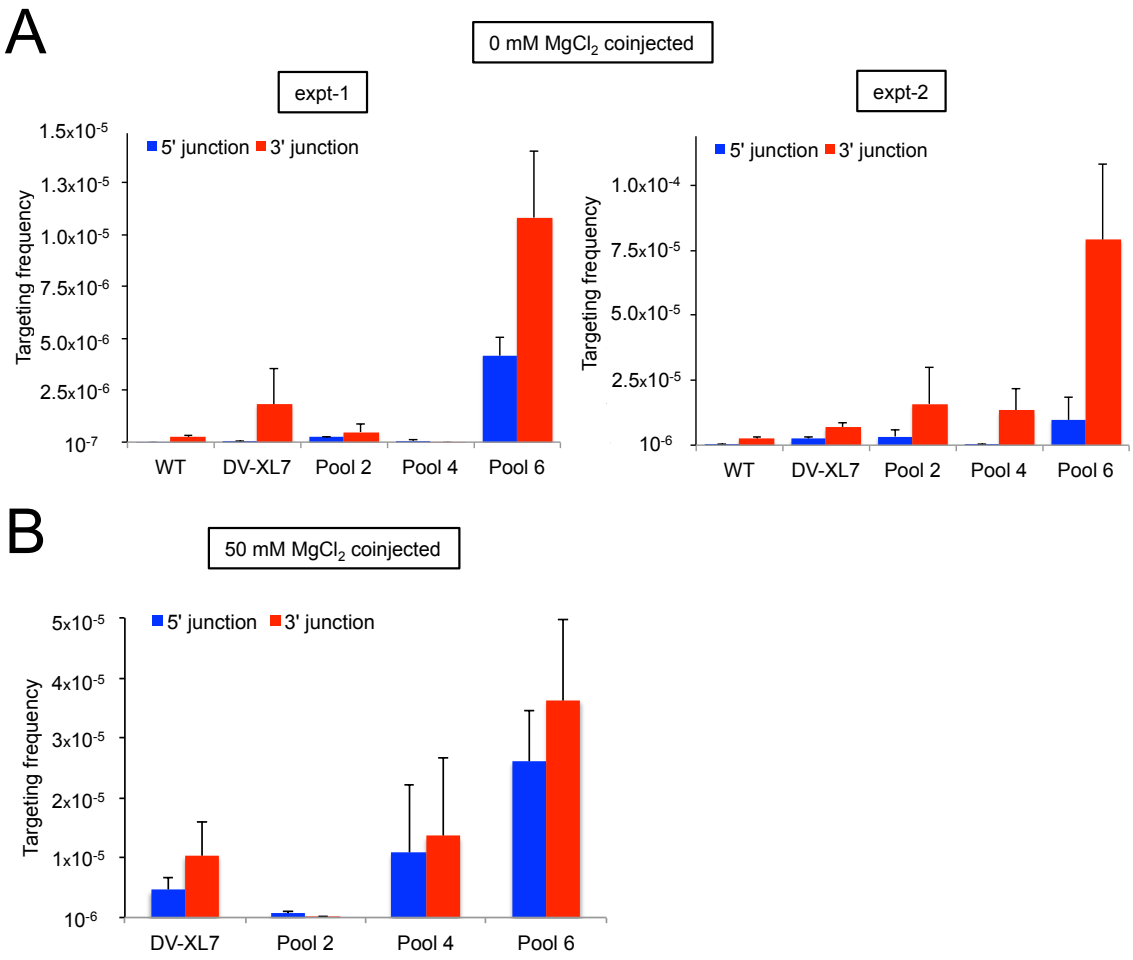



Figure 4.7: Post-selected cycle 6 pooled variants have increased targeting frequencies in *Xenopus laevis* oocytes at low [$Mg^{2+}$].

The post-selection variant pools from cycle 2, 4, and 6 were re-appended with a T3-promoter for transcription and then generated into RNPs. The RNPs from each pool were compared to that of the parental DV-XL7 variant RNP with or without $MgCl_2$ coinjections into *Xenopus laevis* oocytes as indicated, and then analyzed by Taqman qPCR for the 5' and 3' junctions. RNPs injected without plasmid target sites had amplification levels below that of the experimental samples. (*A*) Targeting frequencies from two separate experiments using RNPs from post-selected cycles 2, 4, and 6 without additional $MgCl_2$ compared to DV-XL7 and wild-type. (*B*) Targeting frequencies of post-selected cycles 2, 4, and 6 without additional $MgCl_2$ compared to DV-XL7.

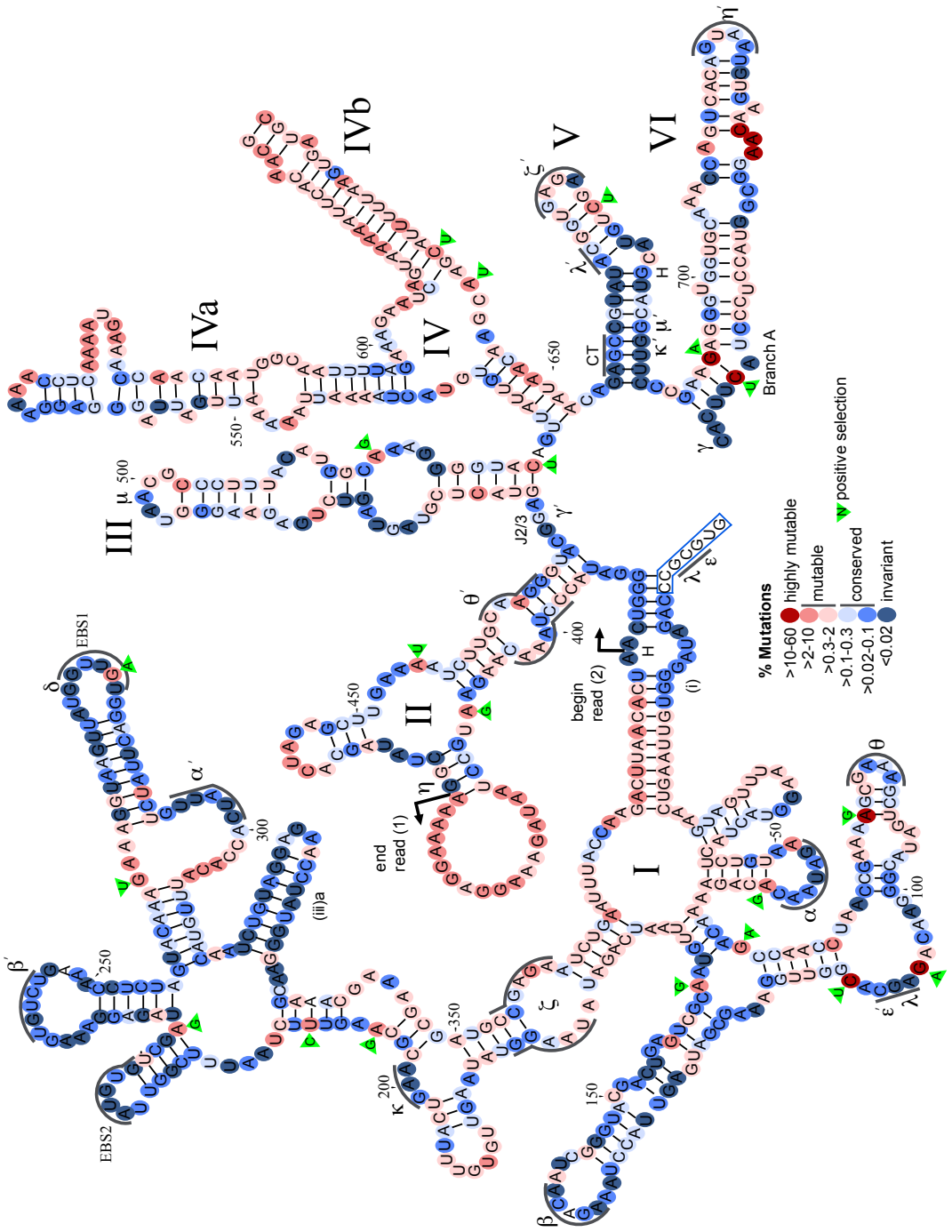



Figure 4.8: Deep sequencing fitness heat map of the six cycles of directed evolution in *Xenopus laevis* oocyte nuclei.

Roche 454 sequencing was used to assess cycles 2, 4, and 6 during directed evolution of the group II intron and generated approximately 15,000 sequencing reads for each cycle analyzed. The introns from each cycle were sequenced in two separate reads of ~400 nucleotides each, marked with black arrows on the figure. The mutations are presented here as a fitness landscape color-coded as a heat map placed on the secondary structure of Ll.LtrB. Blue ovals represent conserved nucleotide positions with mutations present in <0.3% of the reads and did not fluctuate appreciable throughout all six cycles. Red ovals indicate mutation frequencies at the nucleotide position from ≥0.3-60% of the population assessed at cycle 6. Green arrows with nucleotide inscribed, indicate that the nucleotide mutation at that position makes up >80% of the variability, and indicates positive selection. Major tertiary contacts are indicated with black bars and their respective Greek symbol.

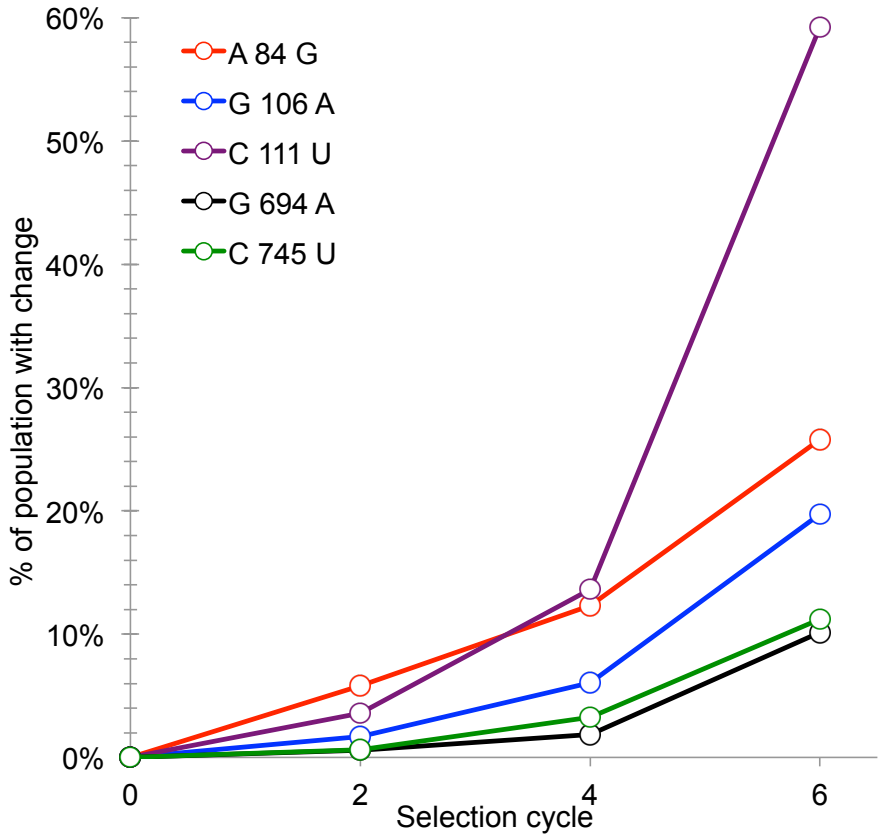



Figure 4.9: Nucleotide positions undergoing positive selection and reaching high frequency.

Five mutations reached 10-60% of the population by cycle 6. The graph shows each mutation's percentage within the population at cycles 2, 4, and 6. The mutation rate conferred during error-prone PCR contributes only a 0.9% mutation rate at each nucleotide per selection cycle, which would give a maximum value for neutral fixation through six cycles as ~5.4%. Therefore, the rise in frequency of these mutations suggests they are not due to neutral drift.

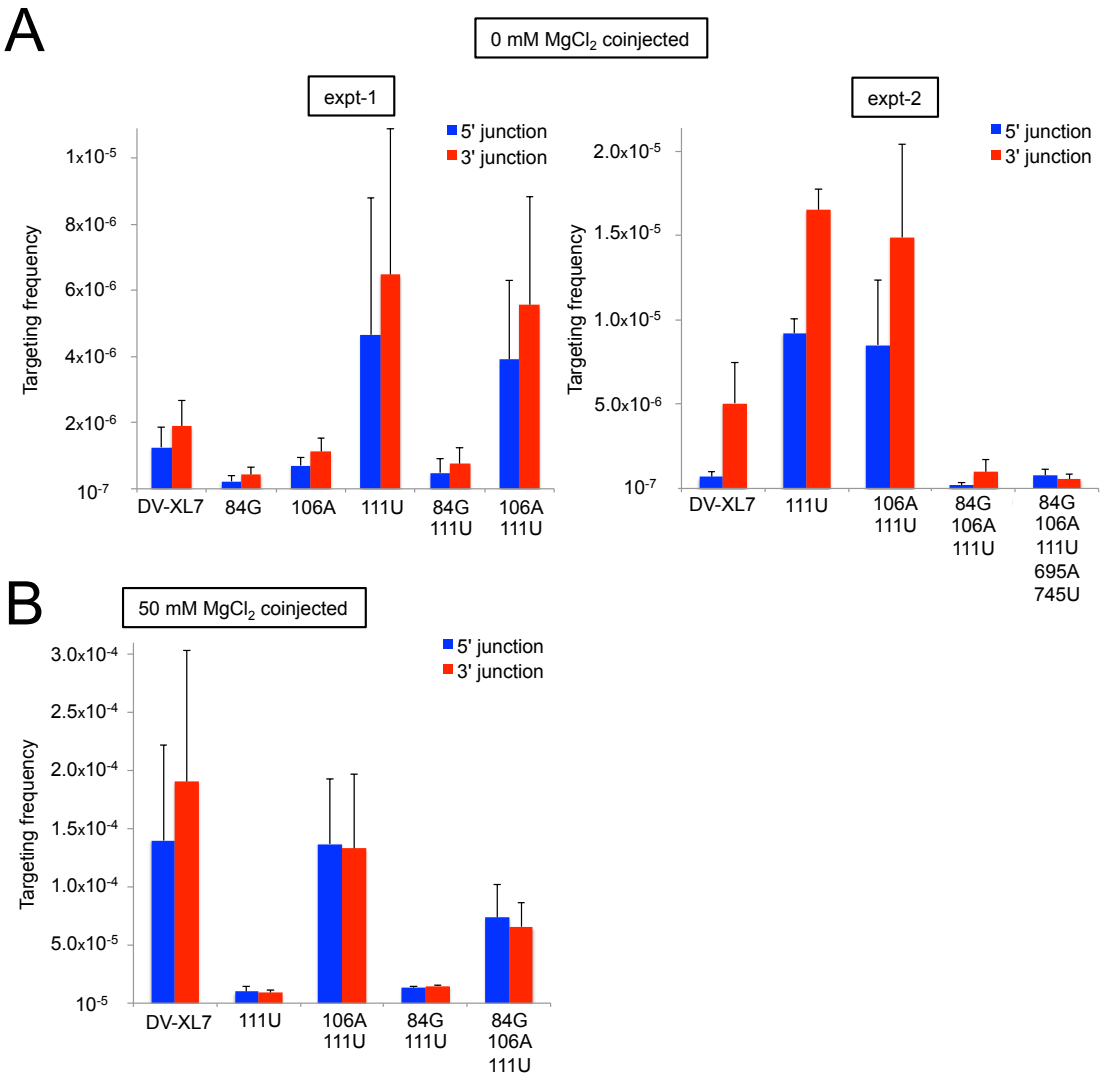



Figure 4.10: Targeting frequencies of mutants and mutant combinations in *Xenopus laevis* oocytes.

Mutants and mutant combinations (see Table 4.3) were generated into RNPs and tested for targeting frequency in *Xenopus laevis* oocytes using Taqman qPCR at the 5' and 3' junctions relative to the parental DV-XL7 RNP. Standard error of the means are reported from four batches of 6-8 oocytes for each data point. RNPs injected without plasmid target sites gave background amplification levels. (*A*) *Xenopus laevis* oocyte targeting without additional $MgCl_2$ for two separate experiments. (*B*) *Xenopus laevis* oocyte targeting with 50 mM $MgCl_2$.

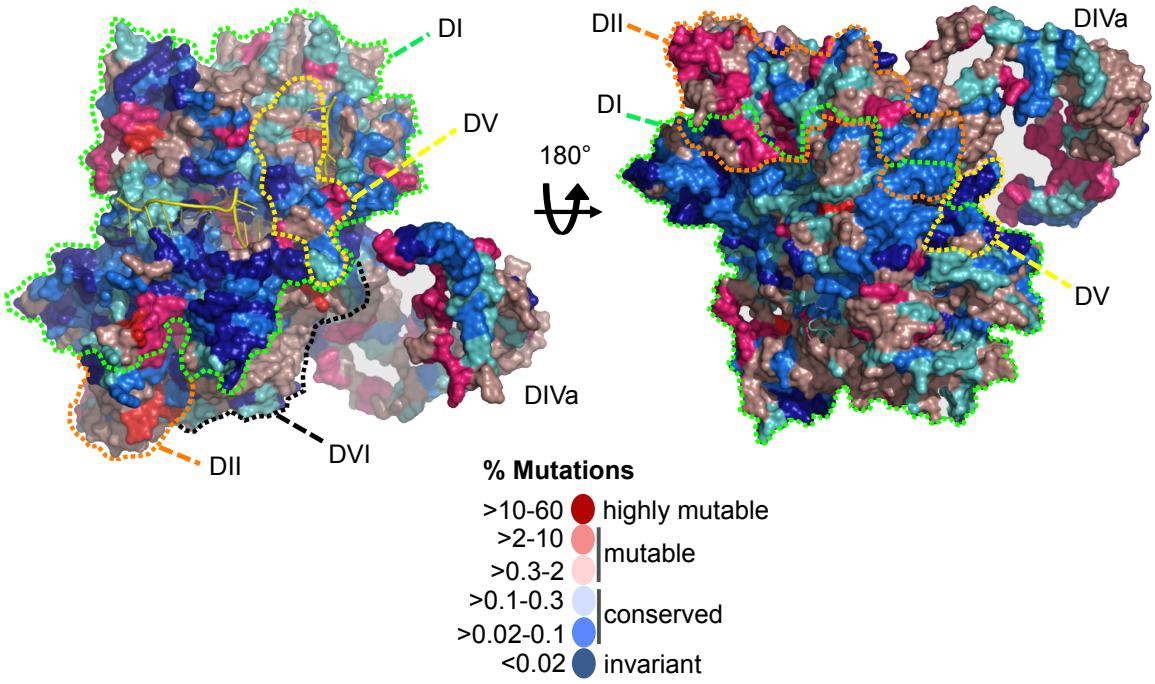



Figure 4.11: Fitness heat map projected onto a three-dimensional model of Ll.LtrB.

An Ll.LtrB model from (Dai *et al*., 2008) was used to map mutation frequencies from the *Xenopus laevis* oocyte selections. As in Figure 4.8, blue colored bases indicate conserved positions (<0.3% mutations), while red colored bases indicate mutation rates ≥0.3-60% at the position. Domains are outlined in broken lines, where visible, with the following colors: DI, green; DII, orange; DV, yellow; DVI, black. DIII is not visible at these views, and DIVa is a stem-loop. The additional region DIVb was not modeled in (Dai *et al*., 2008), and therefore is not shown in our model, but presumably is an additional looping extension. The top five mutants (Figures 4.8 and 4.9) are found within the interior of the Ll.LtrB model and are not visible.

# Chapter 5: DNA targeting and *in vivo* directed evolution of a mobile group II intron within human HEK-293 cells

## 5.1 Introduction

Guo *et al*. performed the first studies on group II intron retrohoming within human cells. Using group II intron RNPs expressed and purified from *E. coli*, they showed that a retargeted group II intron could potentially retrohome into plasmids within human cells bearing the CCR5 gene (Guo *et al*., 2000), which is the coreceptor used by HIV to enter immune cells. However, detection of the CCR5 integrations required the use of nested PCR, suggesting that group II introns function inefficiently in higher eukaryotes. Refined methods for preparing RNPs from *in vitro* prepared RNA and purified LtrA protein failed to yield improved group II intron retrohoming in human cells (Cui, 2006; Vernon, 2010; Hanson, 2013). Nor did the development of plasmid-based eukaryotic expression constructs for Ll.LtrB provide consistent evidence of group II intron retrohoming within human cells (Cui, 2006; Hanson, 2013).

In spite of these inconsistent results, the work described in this chapter builds on the work of three previous members of the laboratory to finally show that group II intron retrohoming in human cells occurs at rates higher than recombination. A previous graduate student, Xioxia Cui generated a human-codon optimized version of LtrA (hLtrA), which she showed could localize to the nucleus when tagged with an SV40 nuclear localization signal (NLS) (Cui, 2006). She further showed that hLtrA could splice Ll.LtrB intron-containing transcripts, and that the addition of exogenous 80 mM $MgCl_2$ led to self-splicing of Ll.LtrB. Another previous graduate student, Joseph Hanson, developed a T7 RNA polymerase (T7 Pol) based expression system for Ll.LtrB, and he showed that this system had minimal cytotoxicity in HEK-293 cells (Hanson, 2013). The

T7 Pol Ll.LtrB expression system was an important development, as it has been shown in *Saccharomyces cerevisae* that Ll.LtrB transcripts expressed by RNA polymerase II (Pol II) are subject to non-sense mediated decay (NMD) (Chalamcharla *et al*., 2010). A previous postdoctoral researcher, Dr. F. Curtis Hewitt, subsequently confirmed that Ll.LtrB transcripts are also subject to NMD in human cells. The results of these studies set the stage for the development of a hybrid group II intron expression system combining T7 Pol driven Ll.LtrB expression with Pol II driven hLtrA expression and its use to study group II intron retrohoming within a human cell.

Here, I have used the hybrid Pol II/T7 Ll.LtrB expression system in human cell culture (HEK-293) to show that group II intron retrohoming occurs when the extracellular $Mg^{2+}$-concentration is raised to 20-80 mM. This environment enables site-specific intron integration into plasmids containing the intron target site at efficiencies up to 0.2% and into a genomically integrated Ll.LtrB intron target site at efficiencies up to 0.02%. I also developed a selection system for the *in vivo* directed evolution of the intron ribozyme in HEK-293 cells and evolved the intron through fifteen cycles in human cells. I analyzed the fitness landscape using Pacific Biosciences single molecule sequencer and identified mutations that rose to high frequency. I found that the *Xenopus leavis* oocyte evolved intron libraries from Chapter 4 have a different fitness trajectory when evolved within HEK-293 cells, which limits their use as a starting point for intron evolution. Finally, I show that combinations of several mutations identified in the selections increase the retrohoming of Ll.LtrB introns in human cells by ~two-fold.

## 5.2 Results

### 5.2.1 Ll.LtrB retrohoming in HEK-293 Flp-In adherent cells requires high $Mg^{2+}$

The hybrid Pol II/T7 group II intron expression system utilizes three plasmids (Figure 5.1*A*) (Hanson, 2013). The T7 Pol plasmid generates T7 RNA polymerase using a nuclear Pol II-driven CMV promoter. The T7 RNA polymerase (T7 RNAP) contains an N-terminal SV40-NLS in order to generate nuclear Ll.LtrB transcription, and it drives the expression of Ll.LtrB RNA from a T7 promoter-driven Ll.LtrB expression plasmid. The human-codon optimized LtrA protein (hLtrA) on a third plasmid is also expressed from a CMV promoter and is fused to an SV40-NLS that localizes it to the nucleus. The hybrid Pol II/T7 group II intron expression system thus provides a way to generate high levels of LtrB RNA unaffected by non-sense mediated decay (Chalamcharla *et al*., 2010), and RNPs that are potentially located in the nucleus.

To test for genomic group II intron insertions, Curtis Hewitt generated a cell line that contains an integrated copy of the wild-type Ll.LtrB target site using the HEK-293 Flp-in cell system from Invitrogen. This cell line contains a single Ll.LtrB target site per cell. Co-transfection of plasmid-based target sites also provides a convenient method of assessing retrohoming levels, as plasmid targets are expected to be present in multiple copies, and the majority of events should occur in the nucleus as both the T7 RNAP and hLtrA proteins have SV40-NLS signals. Xiaoxia Cui's previous observation that exogenous 80 mM $MgCl_2$ placed directly onto cell culture medium stimulated splicing of LtrB suggested that it might similarly stimulate retrohoming within human cells (Cui, 2006).

As indicated for *X. laevis* in Chapter 4, I developed a sensitive and robust Taqman qPCR-based assay to quantify both the 5'- and 3'-integration junctions resulting from the

retrohoming of Ll.LtrB into the wild-type target site in human cells (Yao *et al*., 2013) (Figure 5.1*B*). A typical experiment extended over 2 days, a 24-h period of PEI-based transfection of the expression plasmids, and an additional 24 h in which cells were bathed in growth medium containing 80 mM $MgCl_2$. After $MgCl_2$ treatment, the cells were collected, including both adherent and non-adherent cells, and the DNA was extracted for qPCR analysis.

To test for *bona fide* retrohoming within human cells, I compared four conditions: three negative controls in which cells: (i) received all expression plasmids, but no added $MgCl_2$; (ii) received all expression plasmids except hLtrA and no $MgCl_2$; or (iii) expressed a CCR5 targetron instead of wild-type, and these were compared to a sample that received the three expression plasmids and 80 mM $MgCl_2$ (Figure 5.2*A*). All three negative controls showed similar low-level non-specific background amplification during the qPCR. The samples treated with 80 mM $MgCl_2$ strongly stimulated retrohoming relative to the controls. In three separate genomic targeting experiments using total cells, I detected an average of 0.23% of genomic target sites with the 3'-integration junction and 0.033% of genome target sites with the 5'-integration junction. In contrast, plasmid-targeting experiments led to 1.4% of plasmids containing the 3' junction and 0.056% containing the 5' junction. The difference between numbers of 3' junctions relative to 5' junctions suggests incomplete cDNA synthesis, possibly because hLtrA is less processive in this environment or is impeded by RNA-binding proteins. Therefore, Ll.LtrB RNP targeting efficiencies are most reflected by the numbers of 3'-integration junctions detected by Taqman qPCR. Notably, I could recover the fully retrohomed intron sequence by PCR for the plasmid insertion events due to higher overall copy numbers,

but not for the genomic insertion events, and sequencing of the PCR fragments indicated correct insertion into the wild-type target sequence (data not shown). Unfortunately, I observed a significant increase in cellular blebbing, a hallmark of apoptosis (Charras, 2008), after the addition of 80 mM $MgCl_2$ relative to samples without $MgCl_2$, leading to ~50% of cells losing adherence.

As the addition of 80 mM $MgCl_2$ proved slightly detrimental to cells, I considered that length of time might alleviate this problem. I performed a time course study of genomic targeting using 80 mM $MgCl_2$ at times of 2 to 72 h and analyzed the total cells, both adherent and non-adherent, by Taqman qPCR (Figure 5.2*B*). Genomic targeting was readily detected after 2 h, with 0.026% of target sites containing the 3'-integration junction. At 24-h, I observed a targeting efficiency of 0.5% and, surprisingly, a genomic targeting efficiency of 1.1% at 72-h treatment. However, longer exposure to $MgCl_2$ led to a progressive loss in cellular adherence at the 72 h time point, although the cells that survive do so indefinitely when supplemented with fresh growth media containing 80 mM $MgCl_2$.

The lower genomic targeting efficiency observed at the shorter incubation periods with 80 mM $MgCl_2$ suggested that I try an alternative, such as reduced $Mg^{2+}$-concentrations. I determined genomic targeting rates using 20, 40, 60, and 80 mM $MgCl_2$ for 24 h incubation periods. Concurrently, I determined cellular health by trypan blue staining, and the relative numbers of cells that can re-attach after dissociating the cells and re-seeding (Figure 5.2*C*). As expected, genomic targeting levels increased with rising concentrations of $MgCl_2$, from a low of 0.02% at 20 mM $MgCl_2$ to up to 0.4% at 80 mM $MgCl_2$. Viability remains high between 20-60 mM $MgCl_2$, but it decreases to 60% at 80

mM $MgCl_2$. In contrast to viability, the number of cells that can reattach after dissociating the cells and re-seeding significantly decreased with rising $MgCl_2$ concentrations, dropping to 70% at 20 mM $MgCl_2$ and 15% at 80 mM $MgCl_2$ relative to cells that did not receive $MgCl_2$. Taken together, these data suggested that genomic targeting efficiencies may correlate strongly with loss of cell adherence, possibly because less viable cells take up more $MgCl_2$ from the medium.

To test this possibility, I compared the targeting efficiencies between adherent and non-adherent cells for both genomic and plasmid target sites at 80 mM $MgCl_2$ (Figure 5.2*D*). In this experiment, non-adherent cells were collected separately from those that remain adherent for 24 h in high $MgCl_2$ growth medium. The 3'-integration junctions were detected at efficiencies between 0.7 and 1% for genomic and plasmid targeting, respectively, in non-adherent cells, which is consistent with the idea that they have high Ll.LtrB targeting efficiencies. In contrast, only 0.02% of genomic target sites contained the 3'-integration junction in adherent cells, and 5'-integration junctions were not detectable. The number of 3'-integration junctions detected in plasmid targeting was 10-fold higher than for genomic targeting, and the 5'-integration junctions were around 0.01%. I found that 40 mM $MgCl_2$ could also stimulate Ll.LtrB plasmid targeting at rates 10-fold lower than at 80 mM $MgCl_2$, but could not detect any events for the genomic target site (Figure 5.2*D*). Taken together, these results indicate that high concentrations of $Mg^{2+}$ in cell growth medium stimulate Ll.LtrB retrohoming in living cells, and that reduced cell viability potentially leads to higher targeting rates possibly due to enhanced $Mg^{2+}$ influx. Unfortunately, I could not find conditions in which the non-viable cells could re-attach and grow. Nevertheless, the cells that remain adherent do so indefinitely

within higher $MgCl_2$, as simply replacing the media with fresh high $MgCl_2$ growth medium stimulated cell growth without affecting the adherence of this population of cells. Therefore, these cells may represent a population of robust cells that adapt to high $MgCl_2$ and permit Ll.LtrB retrohoming.

In addition to the above experiments, I tested whether different $Mg^{2+}$-counter ions could increase genomic targeting or the number of viable cells at both 80 and 40 mM concentrations (Figure 5.3). As in the above experiments, genomic targeting efficiency was higher in the non-adherent cells. The alternative $Mg^{2+}$-counter ions generally performed poorly in genomic targeting experiments, and I noticed that improved cell viability led to lower overall targeting. Although MgOAc had ~2-fold higher targeting relative to $MgCl_2$, the cells were generally less viable. Therefore, for further experiments I routinely used $MgCl_2$ for all targeting experiments.

Finally, I tested multiple domain V (DV) mutants that had been previously shown to have decreased $Mg^{2+}$-dependence *in vitro* and/or in a low $Mg^{2+}$ *E. coli* mutant line (Figure 5.4)(Chapter 3 and 4) (Truong *et al*., 2013). For both genomic and plasmid targeting experiments at 80 mM $MgCl_2$, the DV variants performed similarly to the wild-type intron. Thus, these results are consistent with observations in Chapter 4 that different DV variants function optimally in different cellular environments.

### 5.2.2 Directed evolution of Ll.LtrB in HEK-293 Flp-In adherent cells

Although I could consistently detect retrohoming in HEK-293 cells using high $MgCl_2$ in the growth medium, the relatively low efficiency suggested that I consider adapting the intron for this environment by directed evolution as in Chapter 4. My initial strategy for selecting and isolating functional mutagenized group II introns out of human

cells used straightforward direct PCR amplification of integration events as in Chapter 4. Selections utilized the high copy numbers of the plasmid-based system, predominantly because full-length Ll.LtrB integrations into the single genomic target site could not be recovered by PCR. Unfortunately, after three rounds it became apparent that a PCR artifact rapidly overtook and poisoned the selections. In other ribozyme selections, such PCR artifacts have been referred to as *minimonsters* (Breaker and Joyce, 1994).

My alternative strategy utilizes the plasmid-based mobility assay found in Chapter 3 (Truong *et al*., 2013), in which a promoterless tetracycline resistance gene (*tet*$^R$) is activated when a T7 promoter carrying group II intron inserts upstream within a wild-type target site, thereby activating the gene. I found that insertion of the T7 promoter into DIVb did not affect the retrohoming efficiencies of Ll.LtrB into plasmid targets at 80 mM $MgCl_2$ (data not shown). To perform the HEK-293 plasmid-mobility assay, the Ll.LtrB expression system was co-transfected into HEK-293 cells with the *tet*$^R$ recipient plasmid, and after 24 h in $MgCl_2$, plasmids from within the HEK-293 cells were extracted and then electroporated into *E. coli* HMS174(λDE3) cells, which permits selection for Tet$^R$ positive colonies. Tet$^R$ positive colonies were then screened using colony PCR with primers that flank the integration site, distinguishing intron containing from empty target sites. Using 80 mM $MgCl_2$, roughly 40% of colonies contained introns, although the majority of these appeared on the second day of growth, thus forming smaller colonies (Figure 5.5*A*). Importantly, a control lacking additional $MgCl_2$ in the media contained zero introns from forty-eight colonies screened (Figure 5.5*B*). Therefore, this control indicates that targeting does not occur after transformation of the plasmids into *E. coli* by

leaky expression from the ampicillin promoter, and is further evidence that $MgCl_2$ stimulates Ll.LtrB retrohoming in human cells.

The above HEK-293 plasmid-based mobility assay was then used to perform eight rounds of *in vivo* directed evolution in the presence of 80 mM $MgCl_2$ via an adaptive walk (Figure 5.6*A*, *C*). The intron was mutagenized using error-prone PCR at an average mutation frequency of 3 mutations per intron. In parallel, the wild-type intron was tested at every cycle for determining the relative activity of the library for both the 5'- and 3'-integration junctions. As expected, the mutagenized pool initially retrohomed less efficiently than wild-type from cycles one through five by approximately 2-fold at the 3' junction and 10-20% at the 5' junction. Gradually, the mutant intron pool rose in activity beginning at cycle 6, with higher than wild-type 5' junctions detected and near wild-type 3' junctions.

To enrich for mutation combinations that enhance retrohoming in HEK-293 cells, I increased the stringency of the selection by reducing the $MgCl_2$ concentration to 40 mM, and performed four additional selection rounds without the addition of new mutations between cycles (cycles 9-12)(Figure 5.6*B*, *C*). After cycle 12, I selected the pools with additional mutations for two rounds (cycles 13 and 14), and a further selection round without mutations (cycle 15) for a total of seven additional rounds at the new $MgCl_2$ concentration. For the first five selection rounds at 40 mM $MgCl_2$, the pool had higher activity than the wild-type intron compared on the same day (Figure 5.6*C*). Surprisingly, cycles 14 and 15 only had near wild-type activity, because the wild-type intron had relatively higher targeting in these rounds. To determine whether earlier selection rounds were better than the wild-type intron, I retested cycle 12 against the

wild-type intron (Figure 5.7). While the cycle 12 pool had similar targeting as in the initial experiment, the wild-type intron had >2-fold increased activity compared to the last experiment, thus negating the mutant pools gains and suggesting they had near wild-type activity.

### 5.2.3 High-throughput sequencing of Ll.LtrB introns evolved in HEK-293 cells.

Although the mutant pools were not increasing in activity at a sufficiently rapid pace, the possibility remained that specific mutation combinations in the pool had enhanced retrohoming in HEK-293 cells. To identify the mutational diversity of the evolution cycles, I used Pacific Biosciences single-molecule sequencer (PacBio RS), which provides long read lengths (1000-15,000 nt), complemented with circular consensus sequencing (CCS), which compensates for sequencing errors by using rolling-circle amplification to generate concatemer-sequencing reads of the same molecule (Travers *et al*., 2010). To assess the sequencing error-rate using PacBio CCS, I sequenced the wild-type intron and determined the number of substitution, insertion, and deletion errors. My analysis indicates that with a baseline of three rolling-circle sequencing passes of the intron, the substitution error rate is no greater than 0.01%. The insertion and deletion rate was 0.21% and 0.07% respectively, and these occurred predominantly at homopolymeric regions. However, these InDels are not problematic as they can be removed during processing of the sequencing reads. Another advantage of the PacBio RS is that it reads single-molecules directly, and therefore, alleviates problems stemming from formation of molecular hybrids during PCR in other systems, which can over-estimate the number of unique sequences in molecular diversity experiments (Lahr and Katz, 2009; Shao *et al*., 2013). I avoided formation of these PCR hybrids by directly

preparing my sequencing libraries from $Tet^R$ positive recipient plasmids that contained integrated introns.

I initially sequenced cycle 8 and generated a fitness map that displays the degree of conservation of each nucleotide on a secondary structure diagram of the Ll.LtrB intron. The degree of conservation is shown with a heat map scale consisting of conserved sites (0-0.3% mutations; blue) and mutable sites (>0.3-60% mutations; red) (Figure 5.8). On average, the cycle 8 mutant pool contained 4.4 mutations per intron. The majority of nucleotides in the intron remained conserved over the course of eight cycles of directed evolution and regions important for tertiary contacts were highly conserved. The regions that contained the most mutations include DIVb, the single-stranded loop within DII, and DVI. However, many of the nucleotide changes in these regions may be neutral drift, as they do not move towards any specific nucleotide. Position 642 had the highest mutation frequency at 51% of the population. Intriguingly, it is the -2 position relative to the transcription start site of the T7 promoter located within intron DIVb, and also a canonical "TATA-box", a feature used by eukaryotic RNA polymerases, and is also the initiation domain for T7 RNA pol (positions -4 to -1) (Chapman and Burgess, 1987). Sixty-three percent of the mutations were from U to A and the other thirty-seven percent were U to C. I noticed that during the directed evolution cycles, the number of $Tet^R$ positive clones containing the integrated intron had been increasing. Indeed, colony PCR of cycle 8 demonstrated that >85% of $Tet^R$ colonies contained *bona-fide* intron insertions (Figure 5.9), which is in contrast to the frequency found for the wild-type intron (Figure 5.5*A*). Early studies on T7 promoter -2 position T to A or C mutations showed modestly reduced transcription levels *in vitro*, and these mutations were some of

the least detrimental to T7 RNA pol activity (Chapman and Burgess, 1987). Therefore, mutations at the -2 position of the promoter may slightly attenuate transcription and possibly lead to less T7-induced toxicity in the *E. coli* host cell.

A number of mutated sites reached high prevalence in the population (>10%), and many of these were also positively selected, meaning that >80% of the variation was accounted for by a single nucleotide type. Noteworthy among these was the EBS1 G282A mutation identified in the *Xenopus laevis* oocyte selections in Chapter 4, and has been previously been shown to result in fifty percent enhanced retrohoming *in vitro* (Mohr *et al*., 2000).

To determine the effects of different mutation types, I assessed the degree of linkage disequilibrium amongst the most prevalent mutations (Table 5.2) to determine if positive sign epistasis was occurring amongst any intron positions. I note that during these initial eight directed evolution cycles, significant PCR recombination was expected to occur due to high numbers of cycles used for generating mutations. The majority of mutation pairs had *D*' values close to 0, indicating equilibrium between the mutations and no obvious sign epistasis. Three mutations compared in pairwise combinations had *D*' values between 1 and 2.3 suggesting strong linkage. These mutation positions (U642A, G651A, and U652C) are within 10 nucleotides of each other, thus, that they have strong linkage was expected. However, the alternative U642C mutation had *D*' values that were near zero for the 651 and 652 position mutations, arguing that U642A does indeed link because of sign epistasis.

To determine which mutation combinations were increasing during the 40 mM $MgCl_2$ selections (Figure 5.6*B*, *C*), I sequenced cycles 12 and 15 using PacBio CCS.

Figure 5.10 shows the fitness map from cycle 8 with arrows indicating prominent positions in cycle 8 at 80 mM $Mg^{2+}$, whose mutant nucleotide frequencies either decreased or increased further during cycles 12 and 15 at lower $Mg^{2+}$. Surprisingly, half of the high frequency and positively selected mutation nucleotides in cycle 8 decreased to a mutation frequency of less than 0.3% mutation in cycles 12 and 15. Amongst the positively selected mutations in cycles 12 and 15, A84G was of interest because it was also identified in the *Xenopus laevis* oocyte selections. However, most of the positively selected mutations in cycles 12 and 15 were at positions that were previously marginal in cycle 8 (>0.3-2%), but increased by between 2-5 fold in the population at 40 mM $MgCl_2$. These results suggest that different combinations of mutations confer different fitness effects at 80 vs. 40 mM $MgCl_2$. Two positions, however, fixated in the population. The EBS1 mutation G282A reached 93% of the population by cycle 15, and the T7 promoter "TATA-box" mutations at position 642 also fixated in the population at 99.7%. However, the lower frequency mutation in cycle 8, U642C, now made up 96% of the mutations, and the previously prevalent U642A decreased to 3.7% of the population. Because recombination was not expected to occur during these selection cycles, linkage disequilbrium analysis of these data would have limited value.

I determined the top five highest frequency sequencing reads present in cycles 8, 12, and 15 (Table 5.3). Many of these contained similar mutations and are candidates for improving retrohoming in human cells. The highest frequency sequencing read in cycle 8 reached 0.86% of the population, whereas the highest frequency reads in cycles 12 and 15 reached 4.99% and 2.77%, respectively. While cycle 12 and 15 shared common mutation combinations amongst the top five, cycle 8 had a different subset and included the wild-

type sequence, possibly because it was early in the evolution cycles. Of the greater than 2,000 sequencing reads for each cycle, >90% of reads from each cycle occurred only once in the population, with the average population frequency of these reads ranging from ~0.03-0.07%, depending on the number of total number of sequence reads. Finally, I note that the top mutation combination in cycle 12 decreased in cycle 15. This decrease may reflect that additional rounds of mutagenesis during cycles 13 and 14 add mutations and thus generate new unique combinations, thereby depleting the parental genotype.

### 5.2.4 Selection of Ll.LtrB libraries evolved in *Xenopus laevis* oocytes in HEK-293 cells and PacBio sequencing

In Chapter 4, I showed that directed evolution of Ll.LtrB in *Xenopus laevis* ooctyes at low $Mg^{2+}$ led to increased retrohoming into a plasmid target site. To determine whether these preselected mutations offered ready-made fitness advantages in HEK-293 cells, I first cloned in the DIVb-located T7 promoter variants derived from cycle 8 into pool 6 from the *Xenopus* selections. This pool was then selected in HEK-293 cells for four selection rounds at 40 mM $MgCl_2$ in parallel to the wild-type intron for comparison purposes (Figure 5.11*A*, *B*). The *Xenopus* library initially had reduced activity relative to wild-type, suggesting that the *Xenopus* mutant pool did not offer ready-made fitness advantages. After three additional selection rounds, the pool had slightly higher activity than wild-type, but the numbers of detected junctions were within the range of those found for the wild-type intron in other experiments. As performed for the HEK-293 directed evolution experiments, PacBio CCS was used to identify mutations that were increasing or decreasing in frequency. Positions that changed the most are shown with arrows on the original fitness map from *Xenopus* oocytes (Figure 5.11*C*). As found for the transition from 80 to 40 mM $MgCl_2$ in HEK-293 cells, many mutations that were

positively selected in *Xenopus laevis* oocytes decreased to a mutation frequency of <1.0%. The C111U mutation, which increased plasmid targeting in *Xenopus* oocytes, decreased from 60% to 40%, suggesting that this mutation may be slightly detrimental in HEK-293 cells. The EBS1 mutation increased 3-fold in frequency from 5% to 18% of the population. The T7 promoter "TATA-box" mutation also fixated towards U642C as found for the HEK-293 selections. These results suggest that the *Xenopus laevis* mutations did not provide ready-made fitness advantages in HEK-293 cells.

### 5.2.5 Synthetic shuffling of Xenopus evolved libraries in HEK-293 and PacBio sequencing

I also generated three rationally designed mutagenesis libraries, which were either synthetically randomized or synthetically shuffled, using Assembly PCR (Stemmer *et al*., 1995) based on the *Xenopus laevis* oocyte fitness map. Each library had the -2 position of the DIV T7 promoter randomized, and the libraries were marked by inserting an additional nucleotide before the T7 promoter. The first two libraries consisted of 26 and 65 completely randomized nucleotides. The third library consisted of those 28 nucleotides undergoing positive selection towards a single nucleotide in the oocyte map, and these positions were doped fifty/fifty towards either the new nucleotide or the wild-type nucleotide. These three libraries were selected in HEK-293 cells at 40 mM $MgCl_2$ and tested in parallel with the wild-type for retrohoming at this $Mg^{2+}$-concentration. In the very first cycle, library 2 failed to retrohome and produced no colonies containing introns. Both libraries 1 and 3 retrohomed at low frequencies, which were nevertheless high enough to continue the selection for three additional rounds (Figures 5.12*A*, *B* and 5.13*A*, *B*). As for the *Xenopus*-derived library, I performed PacBio CCS to determine mutation types that were positively selected, and these are presented as sequence logos in

Figures 5.12*C* and 5.13*C*. As observed for the Xenopus laevis/HEK-293 hybrid selection, the selection pools for the synthetically generated libraries had higher activity relative to wild-type, but the overall targeting efficiencies stayed within the same range observed for the wild-type intron. The sequencing results of both libraries show that the majority of positions were selected toward the wild-type nucleotide, except for positions that were already identified as being under positive selection in the HEK-293 selections. As found earlier, both mutations G282A and U642C fixated in the population. Thus, the highest frequency mutation positions found for introns evolved in *Xenopus laevis* oocytes were dissimilar to the positions identified for HEK-293 evolved introns. Finally, analysis of mutation combinations in both of the rationally designed libraries showed no specific mutation combinations that were significantly enriching in the population.

### 5.2.6 Testing of clones derived from libraries

To determine whether individual variants had enhanced retrohoming in HEK-293 cells, I tested arbitrarily cloned variants from the libraries of cycle 8, cycle 12, and the hybrid *Xenopus*/HEK-293 selection by performing plasmid targeting at both 40 and 80 mM $MgCl_2$ (Figure 5.14 and Table 5.4). The mutations found amongst the tested variants differed considerably, with greater than four mutations per sequence (Table 5.4). Most contained the EBS1 G282A mutation and the DIV-T7 promoter mutations of U642A or U642C. Between the two experiments, only two variants showed increased activity at 80 mM $MgCl_2$. The variant h12-1 had ~2-fold higher frequencies of 5'- and 3'-integration junctions relative to wild-type, and h12-2 had ~2 fold higher 5'-integration junctions but a only a modest effect on the frequency of 3' junctions. Surprisingly, these variants were derived from the 40 mM $MgCl_2$ selections, but had enhanced activity at 80 mM. Both of

these variants contained an EBS1 G282A mutation and an A or C in position 642. Although the overall increase was modest, each variant contained multiple other mutations that could possibly reduce activity, and it will be of interest to see whether variants with only the highest frequency mutations have better retrohoming in human cells.

### 5.3 Discussion

In this chapter I demonstrate for the first time that mobile group II introns retrohome in human cells if the extracellular growth medium contains high extracellular $Mg^{2+}$ of up to 80 mM. These results are consistent with the idea that the major limitation for group II intron-based gene targeting in higher eukaryotes is the limited intracellular concentration of $Mg^{2+}$. To enhance the levels of *in vivo* retrohoming further, I developed a modified plasmid-based mobility assay and used it for the *in vivo* directed evolution of Ll.LtrB for retrohoming in human cells. In addition to possibly enhancing Ll.LtrB-based DNA targeting, this assay permitted the study of how a large ribozyme evolves within the human cellular environment. I have successfully used the long-reads of the PacBio RS with circular consensus sequencing to map mutation combinations and mutation frequencies of the evolving populations. I found that mobile group II introns are highly tuned to specific intracellular milieus, as subtle changes in $Mg^{2+}$-concentration or host organism significantly altered the fitness trajectories of the evolving introns. Finally, I have shown that some of the selected variants may have enhanced retrohoming into plasmid target sites in human cells relative to the wild-type intron.

One of the highest frequency mutations was the EBS1 mutation of G282A, first identified in Chapter 4, and also tested for *in vitro* retrohoming in (Mohr *et al*., 2000). This mutation converts a G-T base-pair to a Watson-Crick A-T base-pair, and possibly enhances the stability of the RNP during base-pairing to the target-site. The *in vitro* studies by Mohr *et. al*. showed that it enhanced activity by 50% under optimal conditions, and within the alien environment of human cells could have stronger effects, as the two variants with enhanced activity had ~2-fold higher retrohoming. This particular mutation was favored over the wild-type sequence in a previous study of DNA targeting rules, as were changes to G-C base-pairs (Perutka *et al*., 2004). Therefore, base-pairs with better hydrogen-bonding could possibly enhance the retrohoming in low $Mg^{2+}$ environments, which could be tested by generating targetrons with higher GC-content in EBS1/EBS2.

I have shown that the other high frequency mutation in the DIV T7 promoter in position 642 enhances recovery of $Tet^R$ colonies that contain integrated introns, possibly by attenuating transcription (Figure 5.9) (Chapman and Burgess, 1987). It is not yet clear whether the T7 mutations also enhance the activity of the ribozyme, possibly by influencing tertiary structure. The mutations could also affect the retrohoming assay in HEK-293 cells in other ways. For instance, the Ll.LtrB expression plasmid also contains an upstream T7 promoter, which could be negatively influenced by the internal DIV T7 promoter, and Ll.LtrB transcripts in the cell could be mixtures of full-length and truncated RNAs. T7 mutations could influence the balance of cellular Ll.LtrB transcripts by reducing the ability of T7 Pol to bind or initiate at the internal promoter, thereby increasing the relative number of full-length transcripts for mutants. Additionally, the T7 promoter "TATA-box" region has been shown to interact with TFIID and Pol II from

HeLa cell extracts, suggesting that Pol II competes with T7 RNA Pol (Lieber *et al*., 1993; Sandig *et al*., 1993). Thus, TFIID and Pol II could also block read-through from the upstream T7 promoter, and mutations that alter the canonical "TATA-box" could minimize these proteins from binding.

Finally, my observation of changes in the mutation distribution between *Xenopus laevis* oocyte evolved introns compared to HEK-293 evolved introns, and between different $Mg^{2+}$-concentrations argues that group II intron RNAs are finely tuned for their environment. That mutant ribozymes have different fitness for specific environments and require different mutation combinations has been previously shown for the *Azoarcus* group I intron ribozyme, in which mutations were tested in the wild-type conditions, in the presence of formamide, or with a new RNA substrate (Hayden and Wagner, 2012). Although the PacBio high-throughput sequencing identified mutation combinations that could possibly increase activity, testing of every conceivable mutation combination remains an inefficient means of identifying the best variant for human cells. The best strategy would be to use synthetic shuffling of high frequency mutations from the fitness map identified from human cells at the concentration giving the highest activity. This method would permit the screening of many mutation combinations at once, and the optimal combination could then be identified by PacBio sequencing.

## 5.4 Methods

### 5.4.1 Materials and *E. coli* strains

The following antibiotics were used: ampicillin (100 μg/ml), carbenicillin (150 μg/ml), tetracycline (Tet; 15 μg/ml), penicillin (1000 U/ml), streptomycin (1000 μg/ml),

and hygromycin B (50-100 μg/ml). All PCRs were performed using GC-rich Phusion polymerase mastermix (New England Biolabs), unless otherwise indicated. *E. coli* HMS174(λDE3) (Novagen) cells were used for isolation of retrohoming events from human cells. Electrocompetent HMS174(λDE3) were generated as described and had a cfu of >2 x $10^{10}$ measured using pUC19 plasmid (Sambrook and Russell, 2006c; Mastroianni *et al*., 2008).

### 5.4.2 Recombinant plasmids

Plasmid pT7-LtrB contains the Ll.LtrBΔD4(b1-b3) intron RNA (750 nucleotides) cloned downstream of a minimal T7 promoter. Variants of this plasmid include a stuffer variant (pT7LtrB-stuffer), which does not contain the intron and was used for library cloning, and pT7-LtrB(T7), which contains a minimal T7 promoter in DIVb (positions 627-646). Plasmid pCMV-nT7Pol contains the T7 RNA polymerase ORF and an N-terminal appended SV40-NLS under a CMV promoter (Hanson, 2013). Plasmid pIVS-hLtrA-SV contains the human-codon optimized LtrA ORF (hLtrA) and a C-terminal appended SV40-NLS under a CMV promoter. Recipient plasmid pFRT contains a wild-type Ll.LtrB target site inserted into the Flp-In recombinase site and is identical to the region inserted into the HEK-293 Flp-In genome. Recipient plasmid pBRRQ3 contains a wild-type Ll.LtrB target site flanked by sequences with $T_m$ values optimized for qPCR (Table 5.1), and was cloned upstream of a promoter-less *tet*$^R$ ORF. All plasmids used contain an *amp*$^R$ marker.

### 5.4.3 Retrohoming of Ll.LtrB in HEK-293 Flp-In cells

HEK-293 Flp-In cells, which were purchased from Invitrogen (Flp-In™ 293), contain a FRT recombinase site in a decondensed region of the genome. A wild-type

Ll.LtrB insertion site was recombined as a single copy into the FRT containing genomic locus according to manufacturer's recommendations. HEK-293 Flp-In cells were maintained in growth medium consisting of Dulbecco's Modified Eagle Medium (DMEM) supplemented with glutaMAX (Invitrogen), 10% fetal bovine serum (Gemini Biosystems), penicillin, streptomycin, and hygromycin B at 37°C with 5% $CO_2$ unless otherwise stated. For all targeting experiments, HEK-293 cells were seeded 24 h prior to transfection to reach a confluency of 60-80% on the day of transfection in multi-well culture plates (Corning). Cells were dissociated using Stem Pro Accutase (Invitrogen), and cell counting was performing with a hemacytometer or using the Millipore Scepter system.

For genomic targeting experiments plasmids, pT7-LtrB, pCMV-nT7Pol, and pIVS-hLtrA-SV were transfected at 276 ng each, using 2.76 μg branched polyethyleneimine (PEI) (Polysciences, Inc), per well in a 12-well culture plate for 24 h. For plasmid targeting experiments, recipient plasmid pFRT or pBRRQ was included at 276 ng per well in addition to the above three plasmids. After 24 h, the media was removed and replaced with growth medium containing $MgCl_2$ or other $Mg^{2+}$ salts for an additional 24 h unless otherwise specified. The cells were typically 80-90% confluent under ideal conditions. The next day, non-adherent cells were removed by vigorously rinsing with PBS three times, and then collecting adherent cells into a 1.5 ml snap-tube unless otherwise specified. Genomic DNA was extracted from cell pellets using the ZR-genomic miniprep kit (Zymo research) according to manufacturer's recommendations. In plasmid targeting experiments, plasmids were extracted from cells using alkaline lysis with the Wizard SV-miniprep system from Promega. Experiments typically used three

well replicates consisting of independently seeded and transfected wells in parallel used for determination of SEM. Experimental replicates were performed on separate days and report SD.

### 5.4.4 HEK-293 Ll.LtrB selections

Ll.LtrB HEK-293 plasmid retrohoming events were isolated using a modified plasmid-based mobility assay (Mastroianni *et al*., 2008; Truong *et al*., 2013). HEK-293 cells were transfected with plasmids for the hybrid Pol II/T7 expression system, including pT7-LtrB(T7), which contains a minimal T7 promoter in DIV and pBRRQ3, a plasmid which contains a promoter-less $tet^R$ ORF. Retrohoming of LtrB(T7) into pBRRQ brings in a T7-promoter upstream of the $tet^R$ gene, thereby activating it. Plasmids were then isolated by alkaline lysis using the Wizard SV plasmid miniprep kit (Promega). An aliquot was diluted and used for Taqman qPCR, and the remainder was concentrated to 6 μl using a Zymo clean and concentrator column. The concentrated plasmid was electroporated into 100 μl of electrocompetent *E. coli* HMS174(λDE3) cells, and then plated onto Tet (15 μg/ml) LB-agar plates and grown for 2 days. The colonies were then pooled, and the $Tet^R$ plasmids were isolated using alkaline lysis in a Wizard SV miniprep kit (Promega). PCR for 21 cycles using primers that flank the integration site was then used to isolate retrohoming events from the $Tet^R$ plasmids. The product was then band isolated from an agarose gel and used for library generation or mutagenesis.

### 5.4.5 LtrB mutant library generation

pT7-LtrB mutation libraries for each cycle of directed evolution were generated with Mutazyme II (Stratagene) using manufacturer's recommendations for a low mutation rate (3 per kb). Approximately 200 ng DNA template was mutagenized in a 50

μl PCR for 30 cycles, and then re-amplified to obtain a higher yield using Phusion polymerase. The PCR product was agarose gel purified under blue-light illumination using SYBR gold (Invitrogen), and then digested overnight with restriction enzymes AatII and NheI-HF (New England Biolabs). After purification, 750 ng of the insert was ligated to 1 μg of linearized and dephosphorylated pT7-LtrB-stuffer for 2 h at room temperature in a volume of 400 μl using 4000 units of T4 DNA ligase (New England Biolabs). The ligation mix was purified using a Zymo clean and concentrator column to 6 μl and then electroporated into 100 μl MegaXDH10B cells (Invitrogen) with total transformants typically reaching >2 x $10^8$. The resultant library was purified into endotoxin-free supercoiled plasmid using an Endotoxin-free MiniKit II from Omega Biosciences. The plasmid was used for transfection into HEK-293 Flp-In cells for both targeting and selection experiments.

Assembly PCR was used to generate synthetically randomized or synthetically shuffled libraries (Stemmer *et al*., 1995). Briefly, multiple 80-120-mer oligonucleotides spanning the length of the intron containing the randomized or doped positions of interest, and designed to contain respective overlaps with $T_m$ of 55°C were synthesized at the Center for Systems and Synthetic Biology at UT-Austin. For each intron library, a 500-ng equimolar mix of oligonucleotides for each intron was run for 25 cycles in 50 μl of Phusion PCR mastermix. A 5-μl aliquot was placed in 300 μl of Phusion PCR mix with forward and reverse primers that synthesize the full-length intron and run for an additional 25 cycles. The full-length product was extracted from agarose gels and used for library generation as stated above.

### 5.4.6 Taqman qPCR

Quantitative PCR was performed using the Applied Biosystems Viia7 system in 384-well format using Taqman probes (Yao *et al*., 2013). Reactions were performed in technical triplicate in 10-µl volumes for 35-40 cycles using Taqman PCR universal mastermix under standard conditions in 384-well format. Standard curves for absolute quantification used four 10-fold dilutions of either pBRRQ3 or pFRT plasmid containing an Ll.LtrB integration and had >90% efficiency across the range of concentrations used. Plasmids were initially quantified using a Qubit system (Life Technologies). Dilutions were buffered in carrier DNA of 10 ng/µl lambda virus DNA. The primer/probe sets can be found in Table 5.1

### 5.4.7 High-throughput sequencing and computational analysis

Libraries for Pacific Biosciences RS circular consensus sequencing were generated according to manufacturer's recommendations for A-tailed inserts, and sequencing was performed at the John Hopkins deep sequencing and microarray core. Inserts for sequencing were generated directly from $Tet^R$ positive plasmids isolated after directed evolution cycles. Greater than 50 µg of pBRRQ3 $Tet^R$ positive plasmids were digested with AatII and EcoRI-HF, and the band was agarose gel isolated under blue light using SYBR Gold staining.

Sequence reads were filtered to remove reads that did not reach at least three circular passes. Raw sequence reads in the FastQ file format were aligned to the wild-type Ll.LtrB reference sequence using Mosaik Aligner 1.0 (https://code.google.com/p/mosaik-aligner/), and then text files were extracted using the Tablet browser (Milne *et al*., 2010). Insertion gaps were removed using a Perl script,

Gapstreeze, available online at (http://www.hiv.lanl.gov/content/sequence/GAPSTREEZE/gap.html), and reads containing deletion-errors were removed. Aligned sequences were then analyzed for nucleotide variation using a Perl script courtesy of Dr. Scott Hunicke-Smith. All other data analysis, including calculation of nucleotide frequencies and mining of covariation was performed using Unix shell scripts including: grep, cut, uniq, sort, and awk.

Standard linkage disequilibrium was calculated as $D = (P_{AB} \times P_{ab}) - (P_{Ab} \times P_{aB})$, where $P_{AB}$ is the frequency in which the mutations occur together, $P_{Ab}$ and $P_{aB}$ are the mutations occurring independently, and $P_{ab}$ the frequency where neither occurred. The normalized linkage disquilibrium ($D'$) was calculated by dividing positive $D$ values by the theoretical maximum co-occurrence and negative $D$ values by a theoretical minimum co-occurrence based on the observed individual frequencies in the population (Hayden *et al.*, 2011).

Table 5.1: Taqman probes and primers used for detecting retrohoming in HEK-293 cells.

| Target | Name | Type/ orientation | Sequence |
|---|---|---|---|
| 5' junction - pFRT and genomic | 198S-Q10 | Taqman probe | 5'-FAM-ATCCATAACGTGCGCCCA-MGB |
| | 268S | Forward | 5'-CCCCAGCATGCATTACCC |
| | 201A | Reverse | 5'- TCGGTTAGGTTGGCTGTTTTCT |
| 3' junction - pFRT and genomic | 189S-Q1 | Taqman probe | 5'-FAM-CTACTTCACCATATCATTTT-MGB |
| | 197S | Forward | 5'-AAGAGGGTGGTGCAAACCAG |
| | 269A | Reverse | 5'-ACGTAGATAAGTAGCATGGCGGGT |
| *hpt* - pFRT and genomic | 273A | Taqman probe | 5'-FAM-AAGACCTGCCTGAAACCGAACTGCC-BkFQ |
| | 271A | Forward | 5'-CGAGAGCCTGACCTATTGCAT |
| | 272S | Reverse | 5-CGACCGGCTGCAGAACA |
| 5' junction - pBRRQ3 | 198S-Q10 | Taqman probe | 5'-FAM-ATCCATAACGTGCGCCCA-MGB |
| | 200S | Forward | 5'-CCGCTCTAGAACTAGTGGATCCA |
| | 201A | Reverse | 5'-TCGGTTAGGTTGGCTGTTTTCT |
| 3' junction - pBRRQ3 | 189S-Q1 | Taqman probe | 5'-FAM-CTACTTCACCATATCATTTT-MGB |
| | 197S | Forward | 5'-AAGAGGGTGGTGCAAACCAG |
| | 269A | Reverse | 5'-ACGTAGATAAGTAGCATGGCGGGT |
| $tet^R$- pBRRQ3 | 338P | Taqman probe | 5'-FAM-TCGGCACCGTCACCCTGGATG-BkFQ |
| | 336S | Forward | 5'-ACAATGCGCTCATCGTCATC |
| | 337A | Reverse | 5'-CCGGCATAACCAAGCCTATG |

The *hpt* target refers to the gene hygromycin phosphotransferase, which confers hygromycin B resistance in the HEK-293 Flp-In cells. It is located upstream of the wild-type Ll.LtrB target site in the genomic FRT recombinase site. Taqman probes with a 5'-FAM (6-carboxyfluorescien) and a 3'-MGB (dihydrocyclopyrroloindole tripeptide major groove binder) were obtained from Applied Biosystems and those with a 5'-FAM and 3'-BkFQ (Iowa Black FQ) from Integrated DNA Technologies.

Table 5.2: Standard linkage disequilibrium of mutations found in HEK-293 directed evolution cycle 8.

| **Mutations** | ***D*** | ***D'*** |
|---|---|---|
| G282A-A548C | -0.002 | -0.01 |
| G282A-T642A | 0.008 | 0.12 |
| G282A-T642C | 0.0004 | 0.01 |
| G282A-G651A | -0.0001 | -0.001 |
| G282A-T652C | 0.011 | 0.18 |
| T642A-T652C | 0.088 | 1.29 |
| T642A-G651A | 0.036 | 0.99 |
| T642C-T652C | -0.001 | -0.10 |
| A548C-T642A | -0.0005 | -0.01 |
| A548C-T642C | -0.0049 | -0.07 |
| G651A-T652C | 0.073 | 2.74 |
| A548C-T652C | 0.0046 | 0.10 |
| A548C-G651A | -0.0077 | -0.04 |

The table shows calculations for standard linkage disequilibrium (*D*) between the highest frequency variants in the HEK-293 selection at cycle 8, and the normalized linkage disequilibrium (*D'*) (see methods for calculations). The value for *D* and *D'* can be positive or negative, indicating whether the mutations co-occur more frequently or less frequently, respectively, than expected from each mutations individual frequency. Values close to zero indicate linkage equilibrium between the two mutations.

Table 5.3: Top mutation combinations (variants) identified in the HEK-293 evolved selections.

| Library/ number | Mutations | Frequency (%) |
|---|---|---|
| **Cycle 8 mutants** | | |
| hM8-1 | C622U, A643U, G651A | 0.86 |
| hM8-2 | none (wild-type) | 0.79 |
| hM8-3 | C639G | 0.72 |
| hM8-4 | U642A, G651A, U652C | 0.72 |
| hM8-5 | U642C | 0.65 |
| **Cycle 12 mutants** | | |
| hM12-1 | G282A, U642C | 4.99 |
| hM12-2 | U642C | 2.31 |
| hM12-3 | G282A, U642C, U661C | 1.21 |
| hM12-4 | A548C, U642C | 1.14 |
| hM12-5 | G282A, G424A, U642C | 1.07 |
| **Cycle 15 mutants** | | |
| hM15-1 | G282A, U642C | 2.77 |
| hM15-2 | G282A, A548C, U642C | 1.09 |
| hM15-3 | A84G, G282A, U642C | 0.88 |
| hM15-4 | G282A, G423A, U642C | 0.78 |
| hM15-5 | G282A, G422U, U642C | 0.73 |

Mutations were aligned vertically when the mutation was shared amongst other variants. The average frequency of those variants occurring only once was ~0.03-0.07% of the total sequencing reads for each library. Gray shading indicates mutants from the same cycle.

Table 5.4: Randomly cloned variants that were tested in Figure 5.13.

| Name | Library | Mutations |
|---|---|---|
| h12-1 | hM-12 | A58G, G282A, A463U, A574U, U642A, U652C |
| h12-2 | hM-12 | G282A, A417C, G418A, C461A, A463U, U642C |
| h8-1 | hM-8 | G106A, G282A, G419A, A424U, A574U, U642A, G651A, U652C, U689C |
| h8-2 | hM-8 | G334A, C619U, U642A, U652C |
| h8-3 | hM-8 | G282A, A548C, C632U, 600A, 660del, A696G |
| h8-4 | hM-8 | A84G, A131G, G352A, G426A, A548C, A573G, A584G, C590U, C632U, 611AA |
| X-1 | X | G68A, A265U, G282A, U642A, G651A, U652C, DV-XL7 |
| X-2 | X | A84G, C111U, U217C, G282A, A364G, A380U, A430U, A515G, U642C, U652C, DV-XL7 |

Individual clones from the libraries indicated were arbitrarily chosen after cloning of the pooled library, colony PCR, and sequencing. DV-XL7 refers to the mutations in the distal stem of DV identified in Chapter 4. The mutated positions were at 692-694 and 700-701 to the nucleotides GGU and CU, respectively. The hM-8 and hM-12 libraries were those generated from directed evolution directly in HEK-293 cells. The X library refers to the library selected first in *Xenopus laevis* oocytes and then HEK-293 cells.

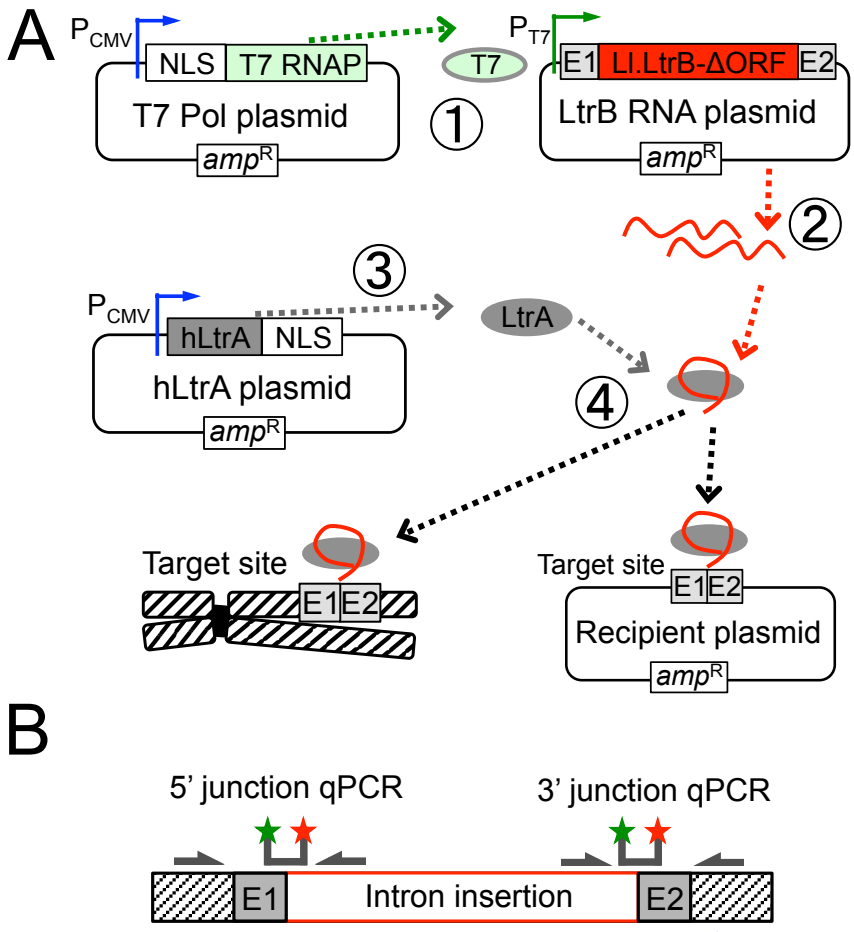
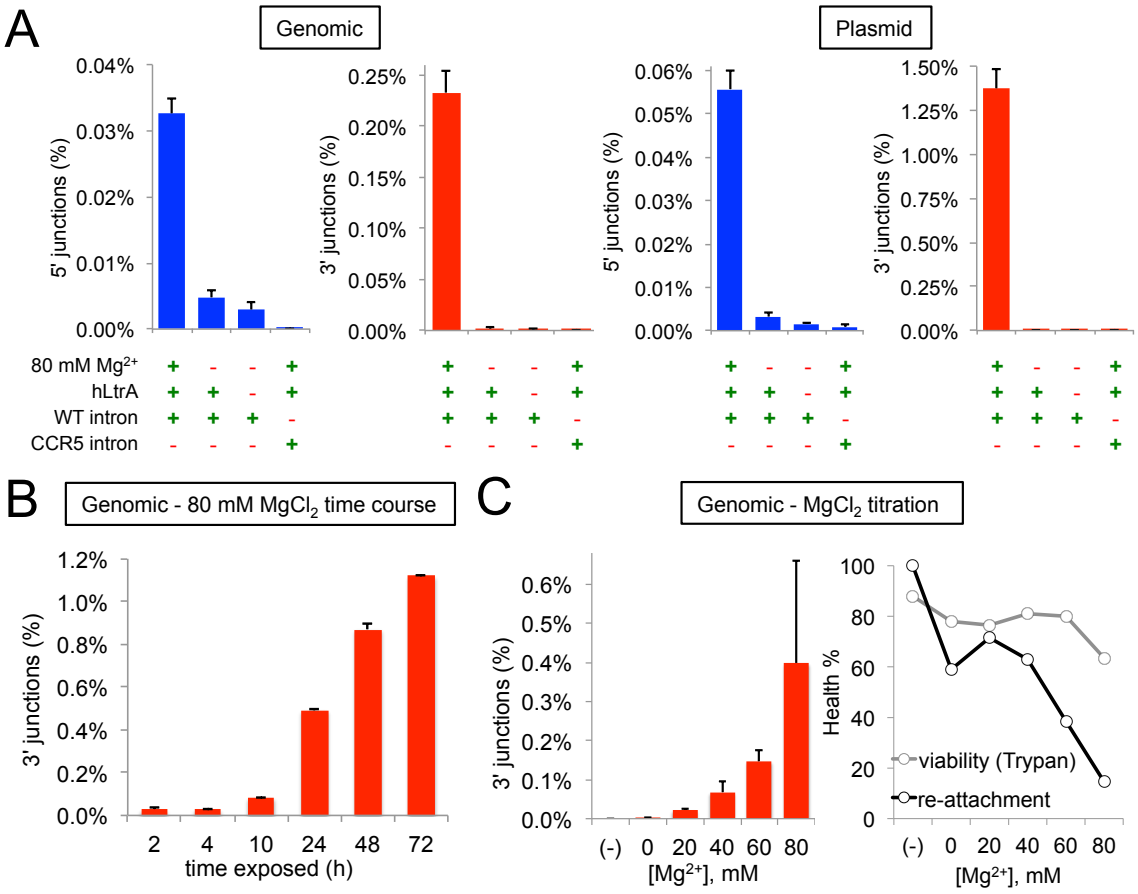



Figure 5.1: Hybrid Pol II/T7 Ll.LtrB human expression system and Taqman qPCR.

The Ll.LtrB eukaryotic expression system consists of three plasmids transiently transfected into HEK-293 cells. After retrohoming of the RNP into the wild-type target site located in either a genomically integrated site or a recipient plasmid, integration events are detected by Taqman qPCR. (*A*) The figure shows the plasmids used in the eukaryotic expression system. The steps leading to Ll.LtrB RNP formation are numbered as follows: 1) T7 RNA Pol is transcribed from the nucleus using Pol II and transported to the cytoplasm; 2) T7 RNA Pol, which contains an NLS, transcribes Ll.LtrB RNA; 3) hLtrA is transcribed in the nucleus via Pol II and splices Ll.LtrB RNA; and 4) after splicing and RNP formation, RNPs can retrohome into recipient plasmids or in a genomically located target site. (*B*) The Taqman probes overlap the 5'- and 3'-integration junctions and bind to amplicons resulting from PCR of the primers indicated in the Figure (see Table 5.1).




D

| | 5' junctions (%) adherent cells | 3' junctions (%) adherent cells | 5' junctions (%) non-adherent cells | 3' junctions (%) non-adherent cells |
|---|---|---|---|---|
| | | **80 mM $MgCl_2$** | | |
| **Genomic target** | not detected | 0.02 ± 0.006 | 0.01 ± 0.001 | 0.71 ± 0.29 |
| **Plasmid target** | 0.01 ± 0.007 | 0.19 ± 0.06 | 0.03 ± 0.02 | 1.05 ± 0.47 |
| | | **40 mM $MgCl_2$** | | |
| **Genomic target** | not detected | not detected | unknown | unknown |
| **Plasmid target** | 0.002 ± 0.0006 | 0.025 ± 0.004 | unknown | unknown |

Figure 5.2: Ll.LtrB retrohomes into plasmid and genomic target sites using high extracellular $Mg^{2+}$ concentrations.

(*A*) HEK-293 Flp-In cells were examined in three experiments for plasmid or genomic targeting in culture medium supplemented with 80 mM $MgCl_2$ for 24 h, following a 24-h transfection period. The three types of negatives controls are indicated. Junctions were detected by Taqman qPCR. Values are for three experiments performed on separate occasions, and the error bars are SD. (*B*) Time course of genomic targeting in the presence of 80 mM $MgCl_2$. Cells were incubated in the presence of $MgCl_2$ for periods of 2, 4, 10, 24, 48, or 72 h and, at the end of each time point total cells (both adherent and non-adherent) were collected. (*C*) The bar graph shows 3' junctions detected by Taqman qPCR for increasing concentrations of $MgCl_2$ (0, 20, 40, 60, 80) after 24 h. The (-) control indicates no addition of hLtrA protein. The second line graph shows viability as determined by trypan blue staining and re-attachment as measured by cells adherent after reseeding compared to the (-) control. Values are for three separate transfections on the same day, and the error bars are SEM. Taqman qPCR was performed on the 3' junctions of extracted genomic DNA. (*D*) The table shows the numbers of junctions detected for genomic or plasmid targeting in adherent vs. non-adherent cells at 80 and 40 mM $MgCl_2$ for 24 h. The values are the average of ≥ 4 separate experiments, and the error bars show SD. Further experiments routinely used only adherent cells.

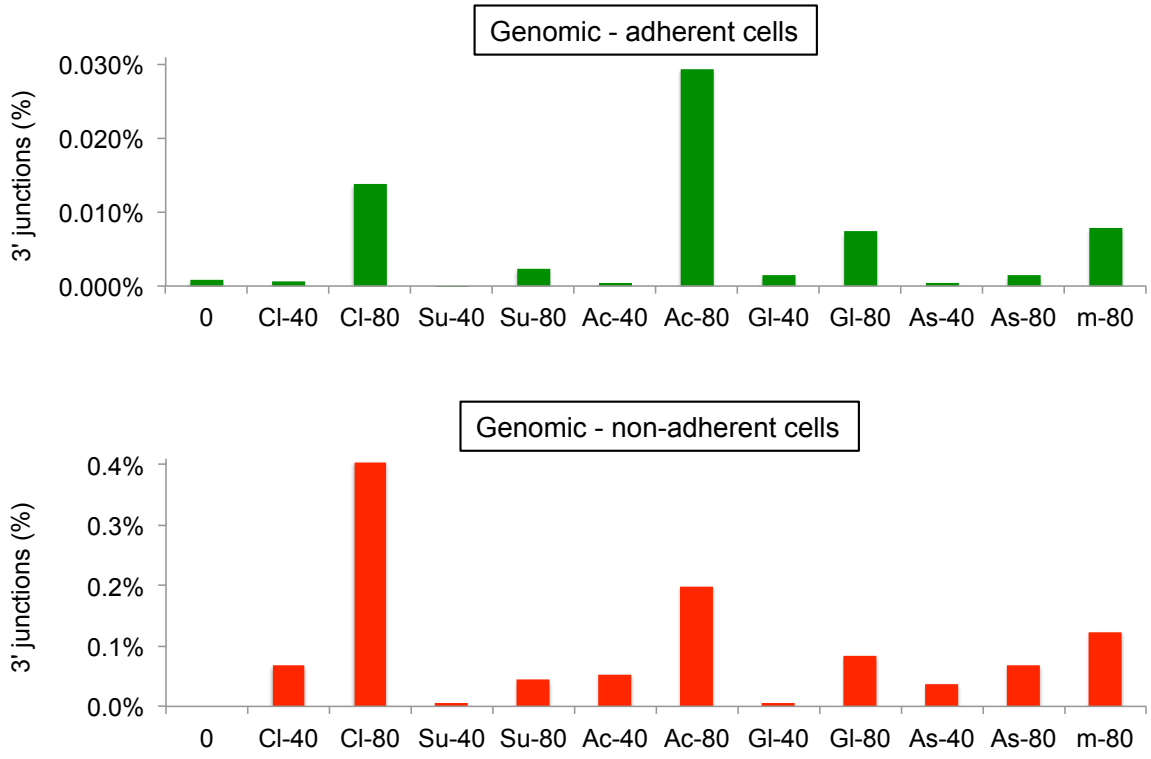



Figure 5.3: Different $Mg^{2+}$-counterions lead to lower levels of targeting than $MgCl_2$.

The two bar graphs show a comparison between genomic targeting rates in 'adherent' vs 'non-adherent' cells using various $Mg^{2+}$-counterions at 40 or 80 mM after 24 h. Detection of Ll.LtrB 3' junctions in the genomic wild-type target site were determined by Taqman qPCR. The different $Mg^{2+}$-counterions used were: Cl, $MgCl_2$; Su, $MgSO_4$; Ac, MgOAc; Gl, Mg-glutamate; As, Mg-aspartate; and m-80, equimolar $MgCl_2$, $MgSO_4$, Mg-glutamate, and Mg-aspartate. The concentration used is indicated next to the counterion abbreviation.

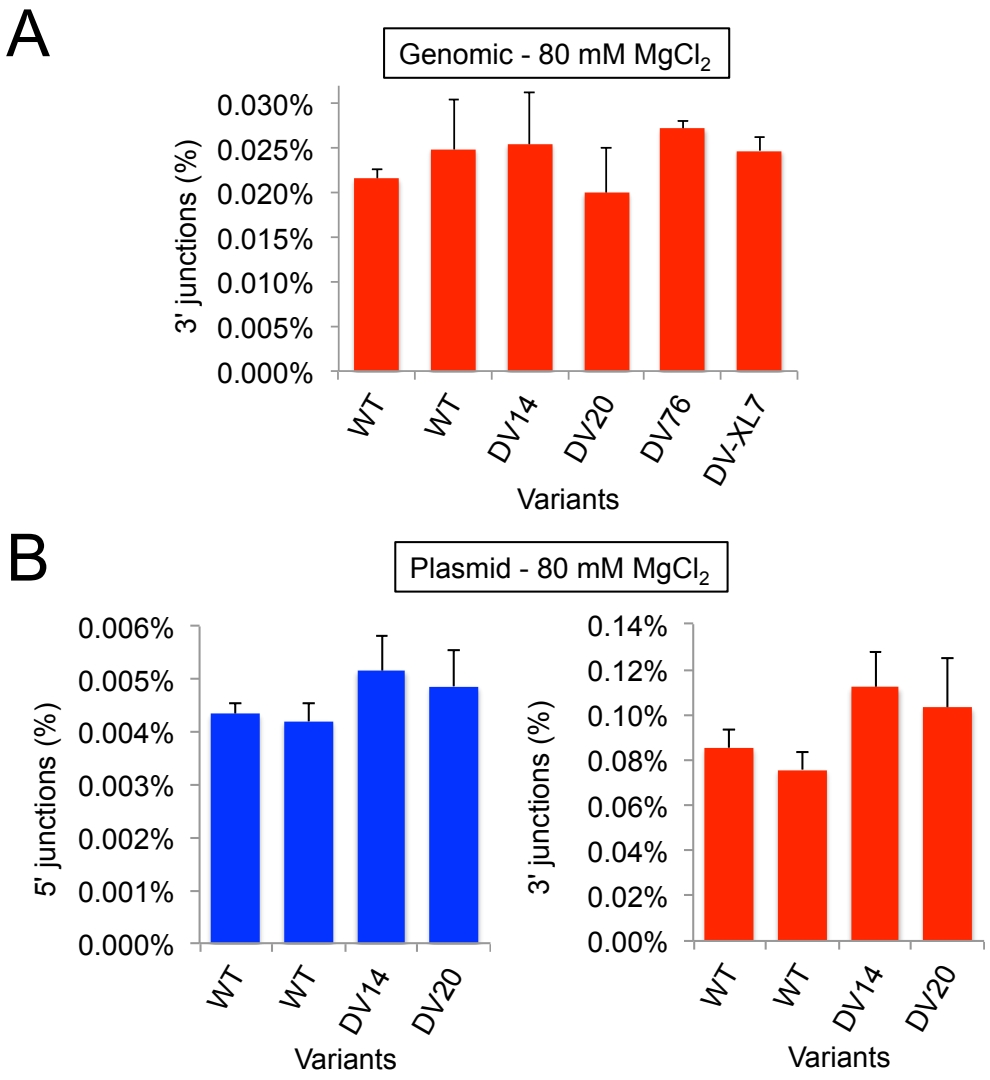



Figure 5.4: DV mutants do not improve genomic or plasmid targeting in HEK-293 cells.

DV mutants from Chapters 3 and 4 were tested in parallel to the wild-type intron for differences in genomic or plasmid targeting at 80 mM $MgCl_2$ in HEK-293 cells. The junctions were determined by Taqman qPCR, and are shown for adherent cells only. Values are for three separate transfections on the same day and error bars are SEM. The wild-type intron was tested twice in triplicate to show the experimental range.

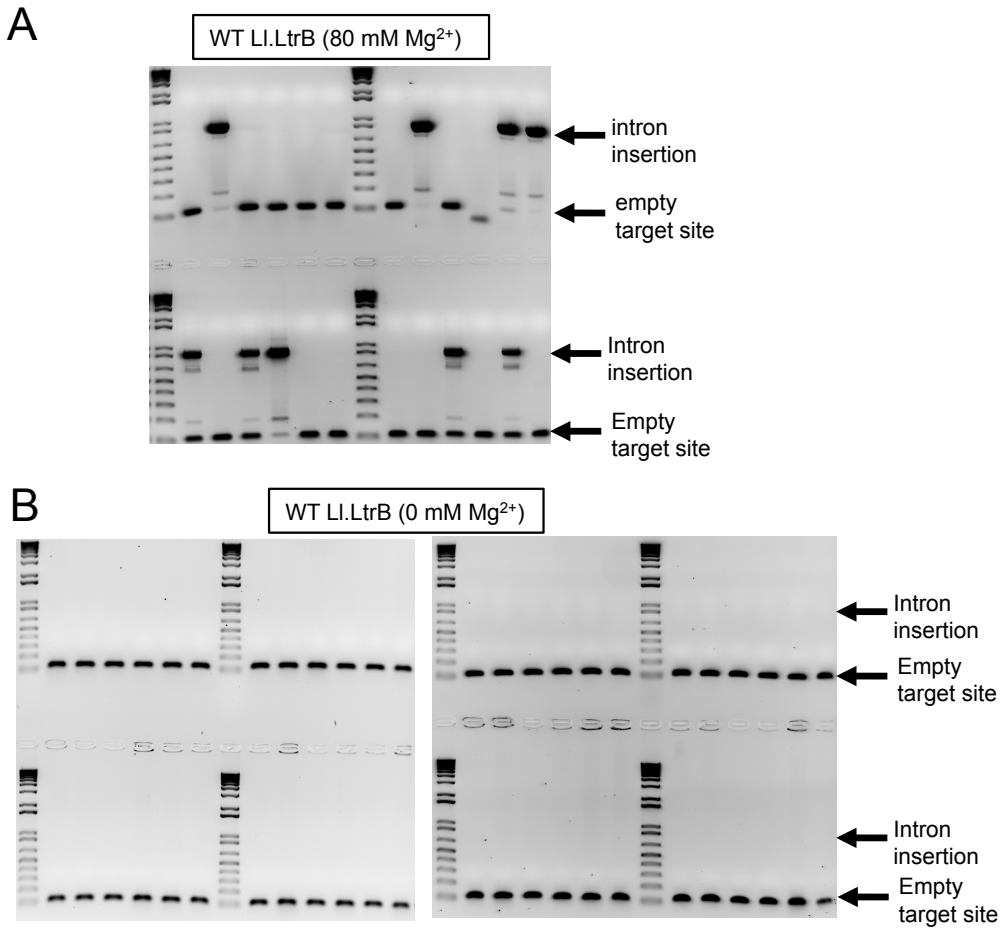



Figure 5.5: Plasmid-mobility assay selection scheme isolates Ll.LtrB retrohoming from HEK-293 cells.

Ll.LtrB retrohoming was isolated from HEK-293 cells by using a modified plasmid-based mobility assay (Truong *et al*., 2013). Retrohoming of a T7 promoter carrying LtrB construct into plasmid pBRRQ3 introduces a T7 promoter upstream of a promoterless *tet*$^R$ gene, thereby activating it (see Chapter 3, Figure 3.2). After performing the targeting experiments in HEK-293 cells, total plasmids were extracted and electroporated into *E. coli* HMS174(λDE3) cells and plated onto Tet LB-agar plates. Randomly selected Tet$^R$ positive colonies were subjected to colony PCR using primers that flank the integration site. (*A*) Colony PCR of Tet$^R$ colonies after retrohoming of wild-type Ll.LtrB in HEK-293 cells in the presence of 80 mM $MgCl_2$ and electroporating total plasmids into *E. coli* HMS174(DE3). (*B*) Colony PCR of Tet$^R$ colonies as in (*A*), but without the addition of $MgCl_2$ during Ll.LtrB expression in HEK-293 cells.

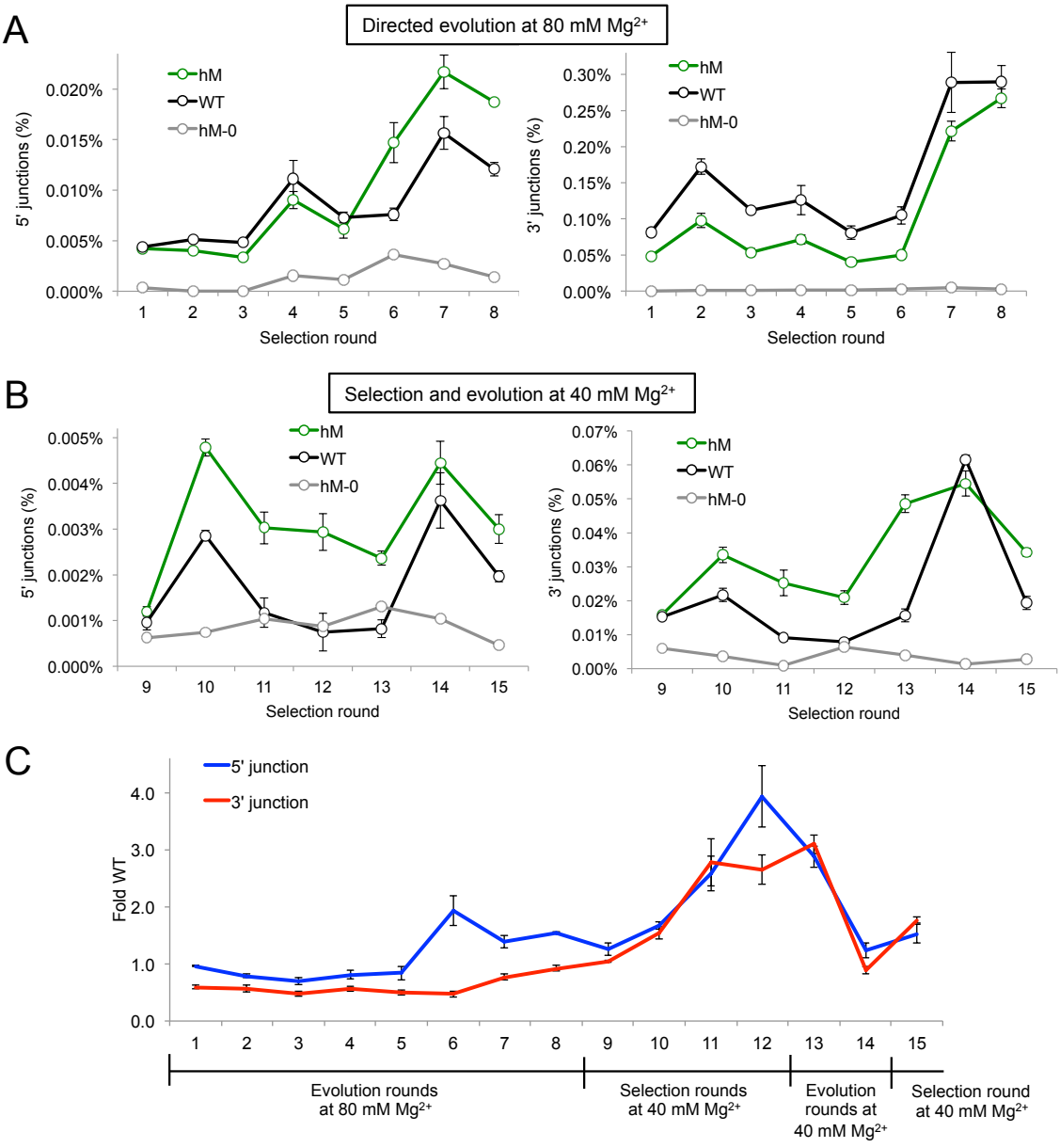



Figure 5.6: Directed evolution and selection of Ll.LtrB within HEK-293 cells at different $MgCl_2$ concentrations.

Mutated Ll.LtrB libraries were selected by using the modified plasmid-based mobility assay as described in Figure 5.5. Total plasmids from the HEK-293 cells were isolated and electroporated into *E. coli* HMS174(λDE3) cells and selected on Tet LB-agar plates. Colonies were scraped, the plasmids isolated, and integration events amplified by PCR. Junctions values are for three separate transfections on the same day determined by Taqman qPCR, and error bars are SEM. The inset in each graph defines symbols for: hM=selection library, WT=wild-type, and hM-0=selection library without additional $MgCl_2$. (*A*) Ll.ltrB was evolved for retrohoming into plasmid targets within HEK-293 cells for eight cycles at 80 mM $MgCl_2$ with addition of three new mutations per kb between each selection cycle. (*B*) Selection cycle 8 from (*A*) was re-selected for an additional four cycles at 40 mM $MgCl_2$ to enrich for variants that enhance retrohoming into plasmids within HEK-293 cells. Additional mutations were added at three per kb during cycles 13 and 14, and cycle 15 was selected without new mutations. (*C*) Fold retrohoming activity of the mutant pool relative to the wild-type intron in assays performed on the same day for pools from all fifteen cycles.

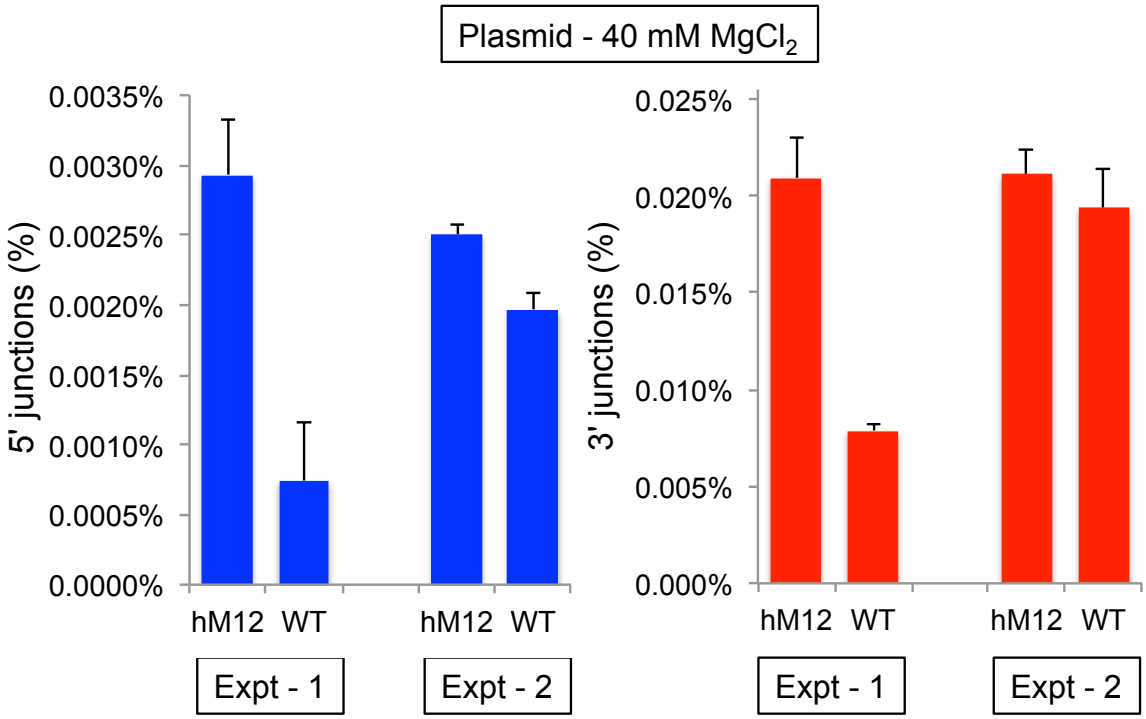



Figure 5.7: Cycle 12 retested against wild type is not significantly enhanced in activity.

The cycle 12 pool was retested against wild-type Ll.LtrB intron (Expt-2) and compared to the experiment performed for the initial selection (Expt-1) in HEK-293 cells. Values are for three separate transfections on the same day determined by Taqman qPCR and the error bars are SEM.

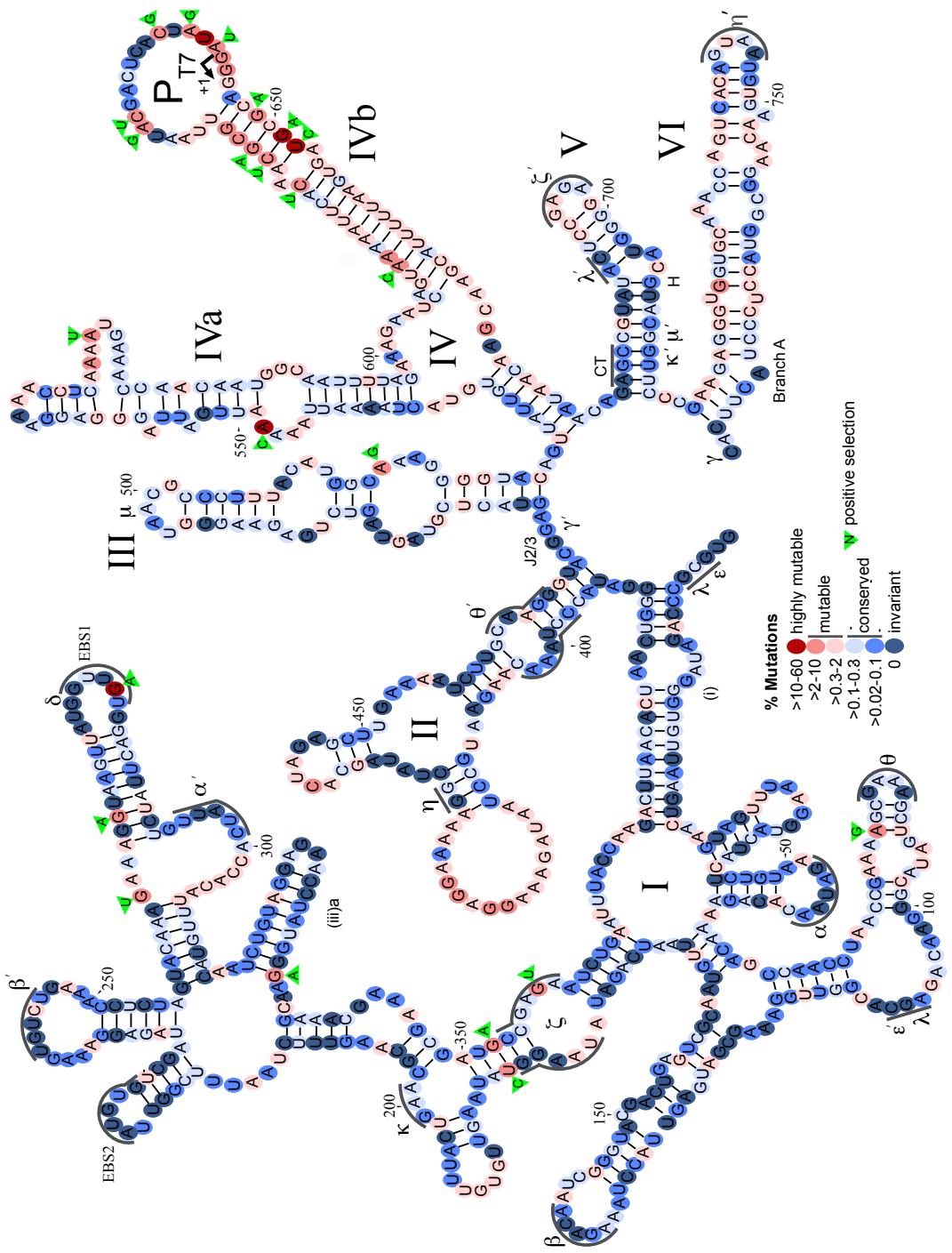



Figure 5.8: Group II intron directed evolution fitness map after eight cycles of directed evolution in HEK-293 cells in medium supplemented with 80 mM $MgCl_2$.

Pacific Biosciences RS circular consensus sequencing was used to assess the number of mutations present in cycle 8 from the directed evolution experiments using 2428 full-length reads. The mutations are presented here as a fitness landscape color-coded as a heat map placed on the secondary structure of the Ll.LtrB intron. Blue ovals represent conserved nucleotide positions with mutations present in 0-0.3% of the reads. Red ovals indicate mutation frequencies at the nucleotide position from >0.3-60% of the population. Green arrowheads with nucleotide inscribed, indicate that the indicated nucleotide comprise >80% of the variants, indicating positive selection. Major tertiary contacts are indicated with black bars and their respective Greek symbol.

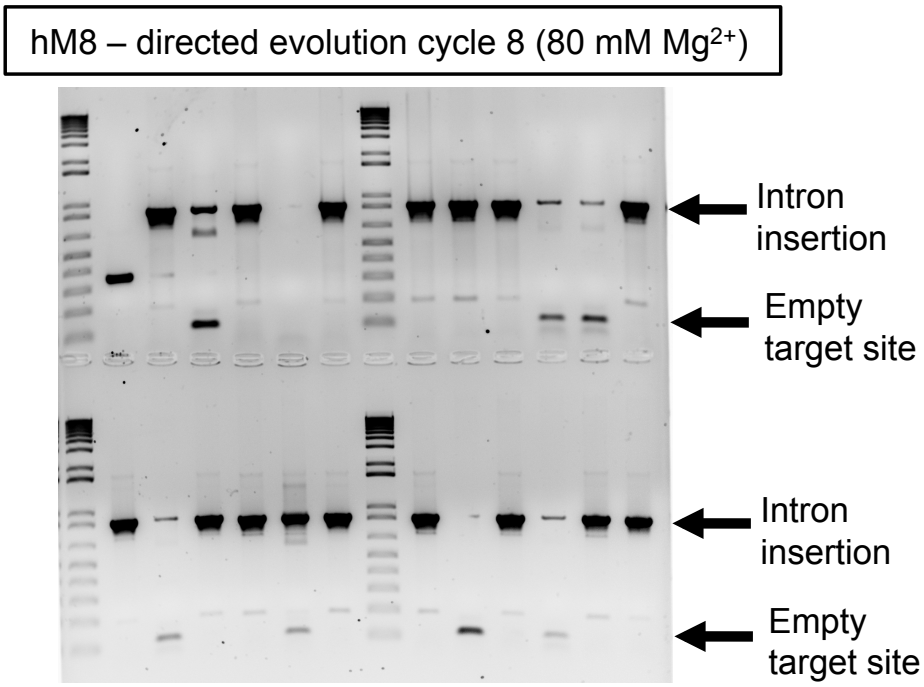



Figure 5.9: Mutations in the DIV T7 promoter enhance the number of *tet*$^R$ colonies containing Ll.LtrB integrations in cycle 8 compared to the wild-type intron.

In cycle 8, sixty-percent of the recovered introns contained an A or C mutation in the -2 position of the DIV T7 promoter of the "TATA-box". To determine if these mutations affect the T7-based Tet$^R$ selection as in Figure 5.6, pBRRQ plasmids recovered from the indicated targeting experiments were electroporated into HMS174(λDE3) and plated onto Tet LB-agar plates. Individual colonies were subjected to colony PCR using primers that flank the integration site. Colony PCR of Tet$^R$ colonies are from the cycle 8 directed evolution library from Figure 5.5*A*.

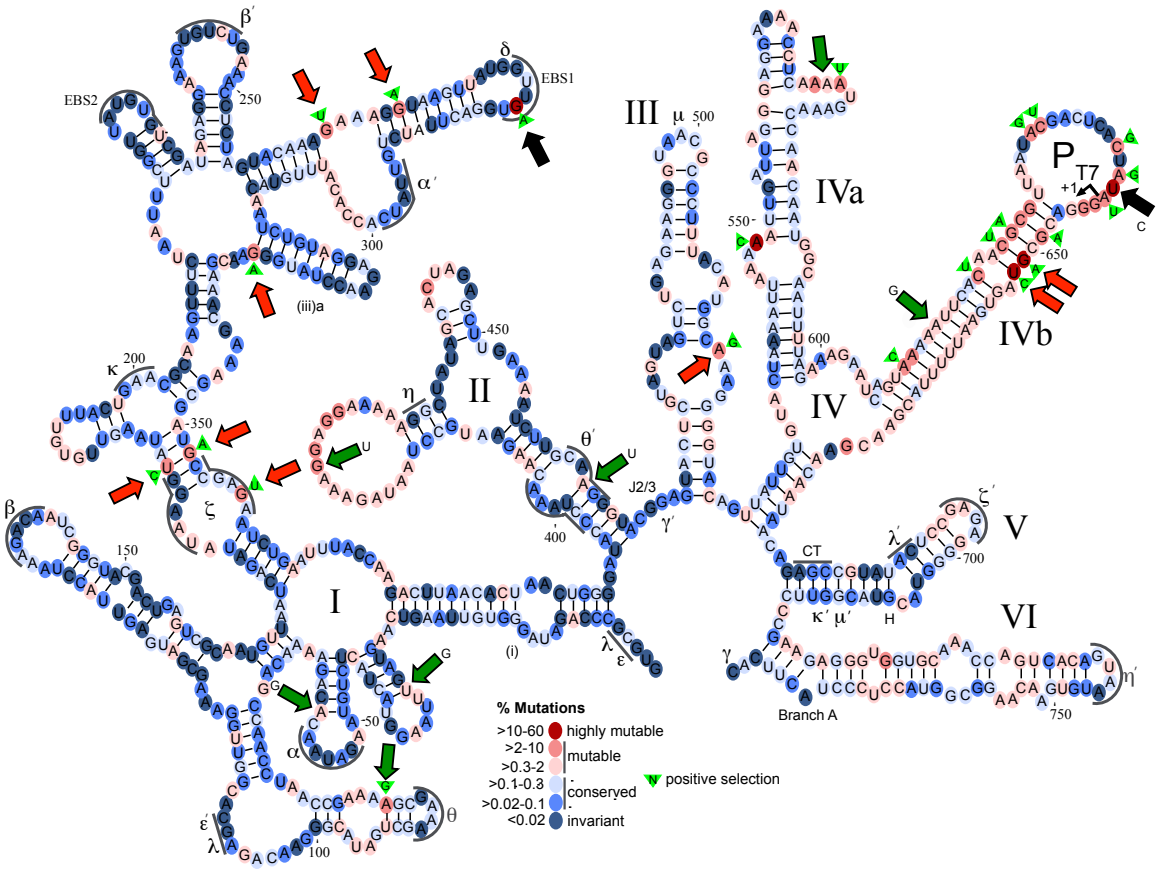



Figure 5.10: Positions undergoing positive selection in cycle 8 at 80 mM $MgCl_2$ whose frequencies increased further or decreased during cycles 12 through 15 at 40 mM $MgCl_2$.

Pacific Biosciences RS circular consensus sequencing was used to identify the mutations present in cycles 12 and 15 from the directed evolution experiments (Figure 5.5*B*) using 3069 and 1913 full-length reads. The figure shows the fitness heat map of cycle 8 (Figure 5.6) and arrows corresponding to positions that changed the most over the course of cycles 12 and 15. Red arrows indicated positions in which the mutations were decreased over the two cycles to conserved (0-0.3%) levels. Green arrows show positions in which mutations were increased towards a positively selected nucleotide. Black arrows show positions in which the mutation indicated fixated to >90% of the population. New nucleotides for previously low mutation frequency positions are indicated next to the arrow.

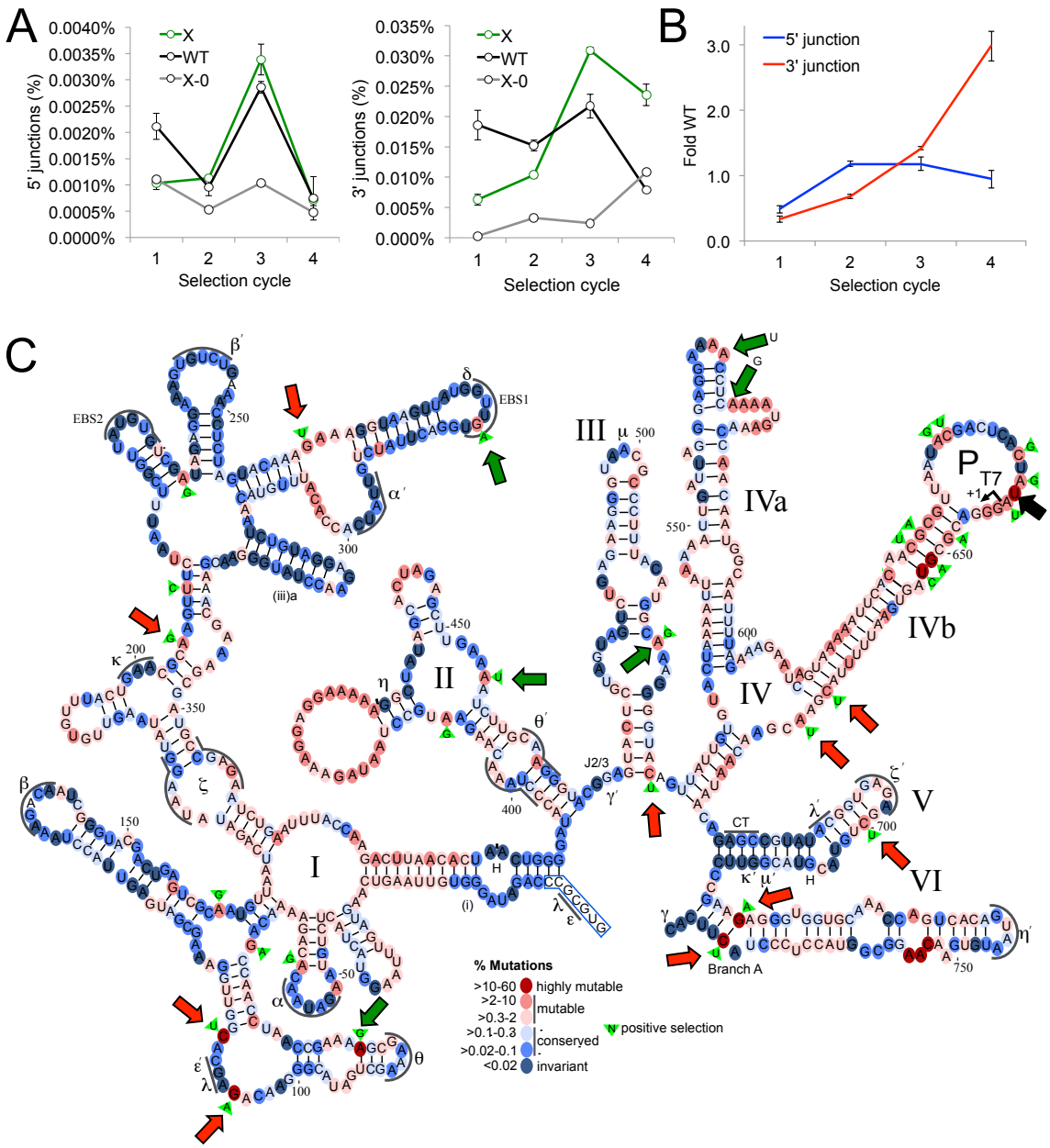



Figure 5.11: *Xenopus laevis*/HEK-293 hybrid selection at 40 mM $MgCl_2$ for four cycles indicates mutations are finely tuned for different environments.

(*A*) The *Xenopus laevis* oocyte pool 6 mutants found in Chapter 4 were modified to add the DIV T7-promoter sequence with the -2 mutation selected in HEK-293 (hM) cycle 8, and re-selected in HEK-293 cells for four cycles. 5' and 3' junctions were quantified by Taqman qPCR during the selection cycles, with the wild-type intron tested in parallel. Values are for three separate transfections on the same day, and the error bars are SEM. The *Xenopus laevis* pool (X) was tested without additional $MgCl_2$ as a negative control. Mutation libraries generally had higher background amplification than when using the wild-type intron. (*B*) Fold activity of the *Xenopus laevis* pool relative to the wild-type intron as performed for experiments on the same day. (*C*) Pacific Biosciences RS circular consensus sequencing (1765 full-length reads) was used to identify the mutations present in the *Xenopus laevis*/HEK293 hybrid selection after four cycles. The figure shows the fitness heat map of the *Xenopus laevis* oocyte pool 6 (Chapter 4, Figure 4.8) and arrows corresponding to positions that changed the most from the *Xenopus* selection to the HEK-293 selection. Red arrows indicated positions in which the mutations decreased over the four cycles. Green arrows show positions in which mutations increased towards a positively selected nucleotide. Nucleotides for positions that increased to high frequency and positively selected in the HEK-293 selections are indicated next to the arrow.

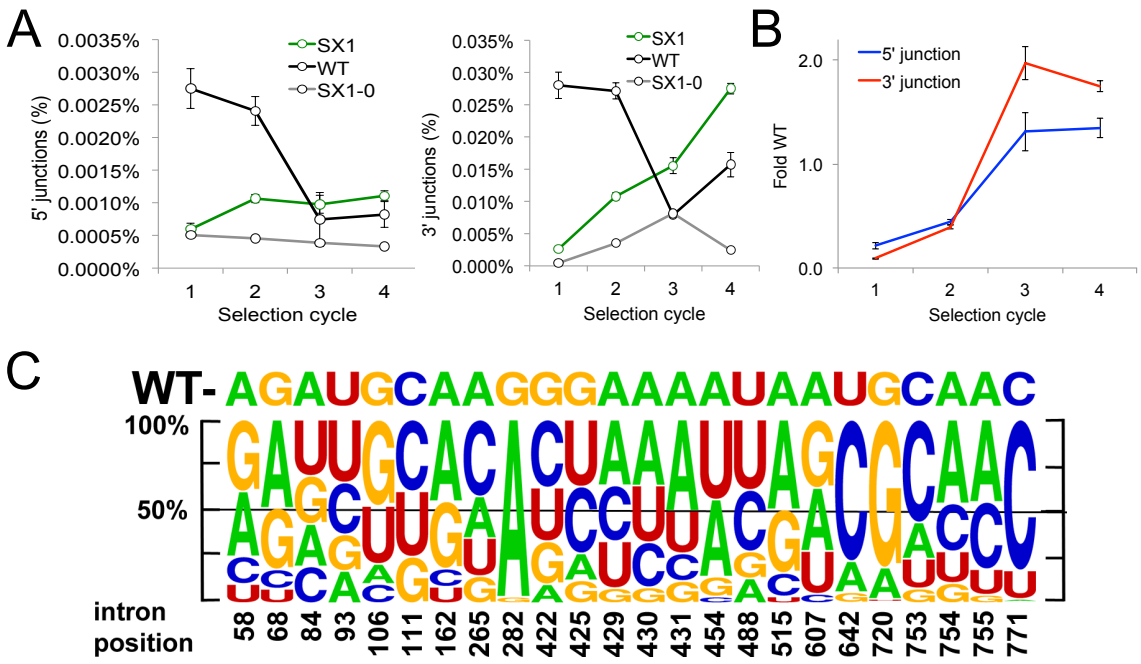



Figure 5.12: Synthetically randomized libraries based on the *Xenopus laevis* oocyte cycle 6 fitness map selected within HEK-293.

(*A*) Nucleotide positions showing a high frequency of mutations in the *Xenopus laevis* oocyte cycle 6 fitness map (Chapter 4, Figure 4.8) were synthetically randomized with 80-120-mer oligonucleotides using Assembly PCR to generate a full-length intron (SX1), and then selected in HEK-293 cells for four cycles. The twenty-four randomized positions are indicated below the sequence logo (intron position) in (*C*). 5' and 3' junctions were quantified by Taqman qPCR during the selection cycles, with the wild-type intron was tested in parallel. Values are for three separate transfections on the same day, and the error bars are SEM. (*B*) Fold activity of the synthetically randomized library relative to the wild-type intron as performed for experiments on the same day. (*C*) Pacific Biosciences RS circular consensus sequencing (4807 full-length reads) was used to assess the number of mutations present after four rounds of selection in HEK-293 cells. The graph shows the percentage of nucleotides present as a sequence logo, with the size of the nucleotide corresponding to its frequency in the population after four rounds of selection. The initial frequency for each nucleotide was ~25%, based on Sanger sequencing. At the top is the wild-type nucleotide relative to the intron position at bottom.

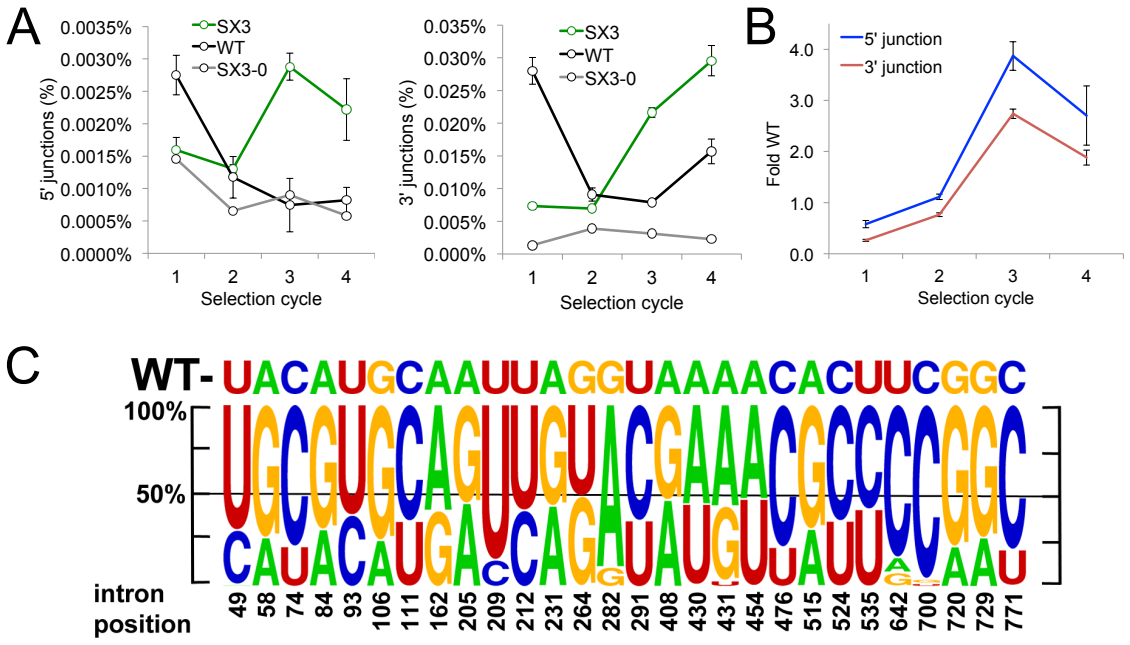



Figure 5.13: Synthetically shuffled libraries based on the *Xenopus laevis* oocyte cycle 6 fitness map selected within HEK-293.

(*A*) Nucleotide positions showing a high frequency of mutation and positive selection in the *Xenopus laevis* oocyte cycle 6 fitness map (Chapter 4, Figure 4.8) were doped fifty percent against the wild-type nucleotide with 80-120-mer oligonucleotides using Assembly PCR to generate a full-length intron (SX3), and then selected in HEK-293 cells for four cycles. The twenty-seven doped positions are indicated below the sequence logo (intron position) in (*C*), except position 642, which was randomized. 5' and 3' junctions were quantified by Taqman qPCR during the selection cycles, with the wild-type intron was tested in parallel. Values are for three separate transfections on the same day, and the error bars are SEM. (*B*) Fold activity of the synthetically shuffled library relative to the wild-type intron as performed for experiments on the same day. (*C*) Pacific Biosciences RS circular consensus sequencing (2307 full-length reads) was used to assess the number of mutations present after four rounds of selection in HEK-293 cells. The graph shows the percentage of nucleotides present as a sequence logo, with the size of the nucleotide corresponding to its frequency in the population after four rounds of selection. The initial frequency for each nucleotide was ~50%, based on Sanger sequencing, and position 642 was ~25%. At the top is the wild-type nucleotide relative to the intron position at bottom.

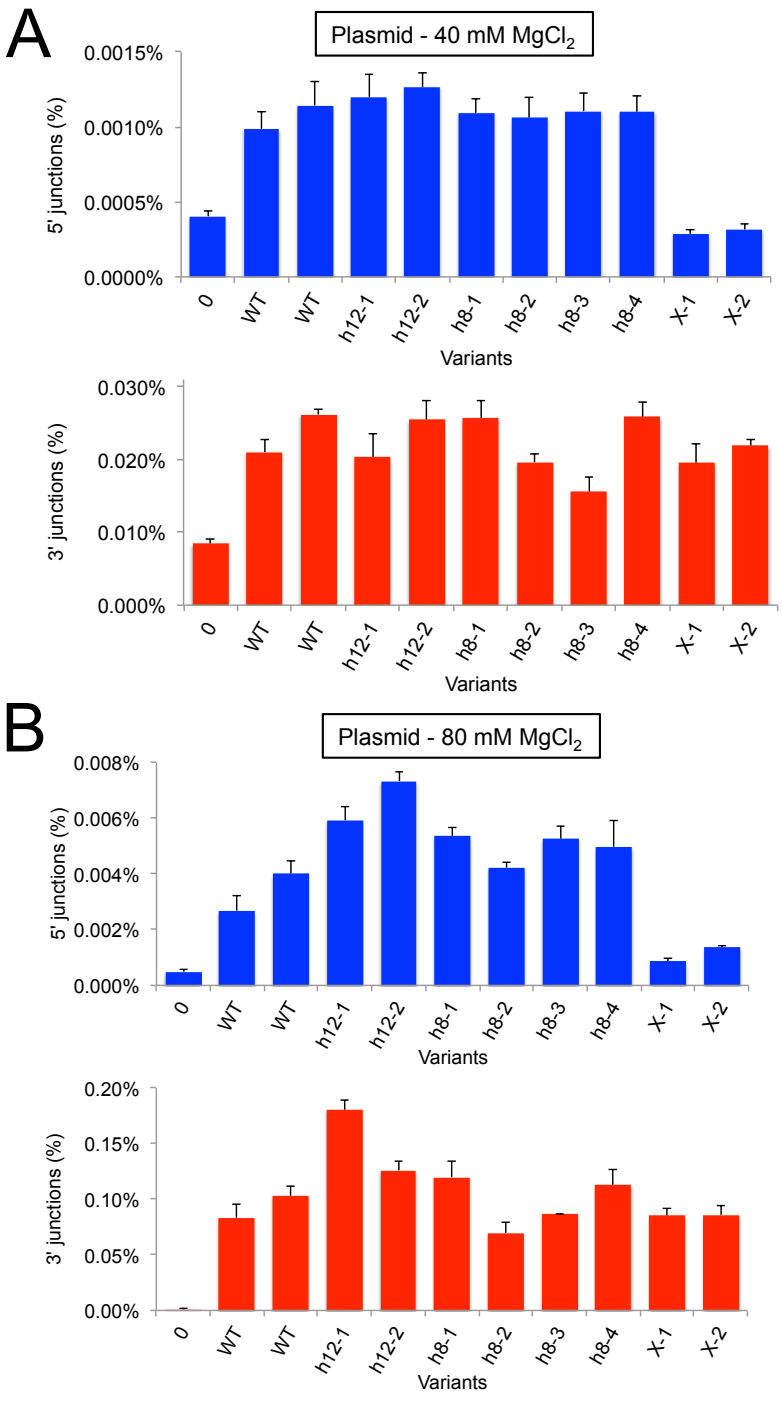



Figure 5.14: Plasmid targeting in HEK-293 cells of randomly cloned Ll.LtrB variants from cycles 8, 12, and the *Xenopus laevis*/HEK-293 hybrid selection.

Variants were randomly selected from the sequence pools, cloned into the Ll.LtrB expression plasmid, and tested for retrohoming in HEK-293 cells with wild-type Ll.LtrB assayed in parallel. Two variants were selected from hM cycle 12, four from hM cycle 8, and two from the *Xenopus laevis*/HEK-293 hybrid selection (X). The mutations tested and the pool in which they originated are summarized in Table 5.3. Plasmid targeting for the variants was tested in HEK-293 cells at both 40 and 80 mM $MgCl_2$ and for both the 5'- and 3'-integration junctions as measured by Taqman qPCR. Values shown are the average of three well replicates and the error bars are SEM. The negative control (0) refers to the wild-type intron tested without additional $MgCl_2$. The wild-type intron was tested twice in triplicate to show the experimental range.

## References


Adams PD, Afonine PV, Bunkoczi G, *et. al*. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr*, 66: 213–221

Aizawa Y, Xiang Q, Lambowitz AM, Pyle AM (2003) The pathway for DNA recognition and RNA integration by a group II intron retrotransposon. *Mol Cell*, 11: 795–805

Allen GCJ, Kornberg A (1993) Assembly of the primosome of DNA replication in Escherichia coli. *J Biol Chem*, 268: 19204–19209

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol*, 2: 2006.0008

Bagby SC, Bergman NH, Shechner DM, Yen C, Bartel DP (2009) A class I ligase ribozyme with reduced $Mg^{2+}$ dependence: Selection, sequence analysis, and identification of functional tertiary interactions. *RNA*, 15: 2129–2146

Barkan A, Klipcan L, Ostersetzer O, Kawamura T, Asakura Y, Watkins KP (2007) The CRM domain: an RNA binding module derived from an ancient ribosome-associated protein. *RNA*, 13: 55–64

Beauregard A, Chalamcharla VR, Piazza CL, Belfort M, Coros CJ (2006) Bipolar localization of the group II intron Ll.LtrB is maintained in Escherichia coli deficient in nucleoid condensation, chromosome partitioning and DNA replication. *Mol Microbiol*, 62: 709–722

Birnboim HC, Doly J (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res*, 7: 1513–1523

Blocker FJ, Mohr G, Conlan LH, Qi L, Belfort M, Lambowitz AM (2005) Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. *RNA*, 11: 14–28

Boudvillain M, de Lencastre A, Pyle AM (2000) A tertiary interaction that links active-site domains to the 5' splice site of a group II intron. *Nature*, 406: 315–318

Boulanger SC, Belcher SM, Schmidt U, Dib-Hajj SD, Schmidt T, Perlman PS (1995) Studies of point mutants define three essential paired nucleotides in the domain 5 substructure of a group II intron. *Mol Cell Biol*, 15: 4479–4488

Breaker RR, Joyce GF (1994) Minimonsters: Evolutionary byproducts of in vitro RNA amplification. *Self-Production of Supramolecular Structures*, 446: 127–135

Brown DD (2004) A tribute to the Xenopus laevis oocyte and egg. *J Biol Chem*, 279: 45291–45299

Bui DM, Gregan J, Jarosch E, Ragnini A, Schweyen RJ (1999) The bacterial magnesium transporter CorA can functionally substitute for its putative homologue Mrs2p in the yeast inner mitochondrial membrane. *J Biol Chem*, 274: 20438–20443

Cadman CJ, McGlynn P (2004) PriA helicase and SSB interact physically and functionally. *Nucleic Acids Res*, 32: 6378–6387

Candales MA, Duong A, Hood KS, Li T, Neufeld RA, Sun R, McNeil BA, Wu L, Jarding AM, Zimmerly S (2012) Database for bacterial group II introns. *Nucleic Acids Res*, 40: D187–90

Carroll D (2011) Genome engineering with zinc-finger nucleases. *Genetics*, 188: 773–782

Cavalier-Smith T (1991) Intron phylogeny: a new hypothesis. *Trends Genet*, 7: 145–148

Cech TR (1986) The generality of self-splicing RNA: relationship to nuclear mRNA splicing. *Cell*, 44: 207–210

Chalamcharla VR, Curcio MJ, Belfort M (2010) Nuclear expression of a group II intron is consistent with spliceosomal intron ancestry. *Genes Dev*, 24: 827–836

Chapman KA, Burgess RR (1987) Construction of bacteriophage T7 late promoters with point mutations and characterization by in vitro transcription properties. *Nucleic Acids Res*, 15: 5413–5432

Charras GT (2008) A short history of blebbing. *J Microsc*, 231: 466–478

Chen X, Denison L, Levy M, Ellington AD (2009) Direct selection for ribozyme cleavage activity in cells. *RNA*, 15: 2035–2045

Chuang SE, Chen AL, Chao CC (1995) Growth of *E. coli* at low temperature dramatically increases the transformation frequency by electroporation. *Nucleic Acids Res*, 23: 1641

Coros CJ, Piazza CL, Chalamcharla VR, Belfort M (2008) A mutant screen reveals RNase E as a silencer of group II intron retromobility in *Escherichia coli*. *RNA*, 14: 2634–2644

Coros CJ, Piazza CL, Chalamcharla VR, Smith D, Belfort M (2009) Global regulators orchestrate group II intron retromobility. *Mol Cell*, 34: 250–256

Costa M, Fontaine JM, Loiseaux-de Goer S, Michel F (1997) A group II self-splicing intron from the brown alga *Pylaiella littoralis* is active at unusually low magnesium concentrations and forms populations of molecules with a uniform conformation. *J Mol Biol*, 274: 353–364

Costa M, Michel F (1995) Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J*, 14: 1276–1285

Cousineau B, Smith D, Lawrence-Cavanagh S, Mueller JE, Yang J, Mills D, Manias D, Dunny G, Lambowitz AM, Belfort M (1998) Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell*, 94: 451–462

Cui X (2006) RNA/protein interactions during group II intron splicing and toward group II intron targeting in mammalian cells. *Ph.D. Dissertation*. The University of Texas at Austin, Austin, TX

Cui X, Matsuura M, Wang Q, Ma H, Lambowitz AM (2004) A group II intron-encoded maturase functions preferentially in cis and requires both the reverse transcriptase and X domains to promote RNA splicing. *J Mol Biol*, 340: 211–231

Cusick ME, Belfort M (1998) Domain structure and RNA annealing activity of the *Escherichia coli* regulatory protein StpA. *Mol Microbiol*, 28: 847–857

Dai L, Chai D, Gu SQ, Gabel J, Noskov SY, Blocker FJ, Lambowitz AM, Zimmerly S (2008) A three-dimensional model of a group II intron RNA and its interaction with the intron-encoded reverse transcriptase. *Mol Cell*, 30: 472–485

Dellaporta SL, Xu A, Sagasser S, Jakob W, Moreno MA, Buss LW, Schierwater B (2006) Mitochondrial genome of Trichoplax adhaerens supports placozoa as the basal lower metazoan phylum. *Proc Natl Acad Sci U S A*, 103: 8751–8756

Delpire E, Gagnon KB, Ledford JJ, Wallace JM (2011) Housing and husbandry of Xenopus laevis affect the quality of oocytes for heterologous expression studies. *J Am Assoc Lab Anim Sci*, 50: 46–53

DeRose VJ (2003) Metal ion binding to catalytic RNA molecules. *Curr Opin Struct Biol*, 13: 317–324

DiFrancesco R, Bhatnagar SK, Brown A, Bessman MJ (1984) The interaction of DNA polymerase III and the product of the *Escherichia coli* mutator gene, mutD. *J Biol Chem*, 259: 5567–5573

Draper DE (2004) A guide to ions and RNA structure. *RNA*, 10: 335–343

Droege M, Hill B (2008) The Genome Sequencer FLX System--longer reads, more applications, straight forward bioinformatics and more complete data sets. *J Biotechnol*, 136: 3–10

Eickbush TH, Malik HS (2002) Origins and evolution of retrotransposons. *Mobile DNA ii*,

Enyeart PJ, Chirieleison SM, Dao MN, Perutka J, Quandt EM, Yao J, Whitt JT, Keatinge-Clay AT, Lambowitz AM, Ellington AD (2013) Generalized bacterial genome editing using mobile group II introns and Cre-lox. *Mol Syst Biol*, 9: 685

Enyeart PJ, Mohr G, Ellington AD, Lambowitz AM (2014) Biotechnological applications of mobile group II introns and their reverse transcriptases: gene targeting, RNA-seq, and non-coding RNA analysis. *Mob DNA*, 5: 2

Eskes R, Liu L, Ma H, Chao MY, Dickson L, Lambowitz AM, Perlman PS (2000) Multiple homing pathways used by yeast mitochondrial group II introns. *Mol Cell Biol*, 20: 8432–8446

Fica SM, Tuttle N, Novak T, Li NS, Lu J, Koodathingal P, Dai Q, Staley JP, Piccirilli JA (2013) RNA catalyses nuclear pre-mRNA splicing. *Nature*, 503: 229–234

Fine EJ, Cradick TJ, Zhao CL, Lin Y, Bao G (2013) An online bioinformatics tool predicts zinc finger and TALE nuclease off-target cleavage. *Nucleic Acids Res*, Epub ahead of print

Forconi M, Herschlag D (2009) Metal ion-based RNA cleavage as a structural probe. *Methods Enzymol*, 468: 91–106

Frazier CL, San Filippo J, Lambowitz AM, Mills DA (2003) Genetic manipulation of Lactococcus lactis by using targeted group II introns: generation of stable insertions without selection. *Appl Environ Microbiol*, 69: 1121–1128

Froschauer EM, Kolisek M, Dieterich F, Schweigel M, Schweyen RJ (2004) Fluorescence measurements of free [$Mg^{2+}$] by use of mag-fura 2 in *Salmonella enterica*. *FEMS Microbiol Lett*, 237: 49–55

Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, Sander JD (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol*, 31: 822–826

Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK (2014) Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol*, Epub ahead of print

Gaj T, Gersbach CA, Barbas CFr (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol*, 31: 397–405

Galej WP, Oubridge C, Newman AJ, Nagai K (2013) Crystal structure of Prp8 reveals active site cavity of the spliceosome. *Nature*, 493: 638–643

Garcia-Rodriguez FM, Barrientos-Duran A, Diaz-Prado V, Fernandez-Lopez M, Toro N (2011) Use of RmInt1, a group IIB intron lacking the intron-encoded protein endonuclease domain, in gene targeting. *Appl Environ Microbiol*, 77: 854–861

Gray AN, Henderson-Frost JM, Boyd D, Sharafi S, Niki H, Goldberg MB (2011) Unbalanced charge distribution as a determinant for dependence of a subset of *Escherichia coli* membrane proteins on the membrane insertase YidC. *MBio*, 2:

Gregan J, Kolisek M, Schweyen RJ (2001) Mitochondrial $Mg^{2+}$ homeostasis is critical for group II intron splicing in vivo. *Genes Dev*, 15: 2229–2237

Grinstein S, Dixon SJ (1989) Ion transport, membrane potential, and cytoplasmic pH in lymphocytes: changes during activation. *Physiol Rev*, 69: 417–481

Grynkiewicz G, Poenie M, Tsien RY (1985) A new generation of Ca2+ indicators with greatly improved fluorescence properties. *Journal of Biological Chemistry*, 260: 3440–3450

Gunther T (2006) Concentration, compartmentation and metabolic function of intracellular free $Mg^{2+}$. *Magnes Res*, 19: 225–236

Guo H, Karberg M, Long M, Jones JPr, Sullenger B, Lambowitz AM (2000) Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science*, 289: 452–457

Guo H, Zimmerly S, Perlman PS, Lambowitz AM (1997) Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA. *EMBO J*, 16: 6835–6848

Gupta RK, Gupta P, Moore RD (1984) NMR studies of intracellular metal ions in intact cells and tissues. *Annu Rev Biophys Bioeng*, 13: 221–246

Halls C, Mohr S, Del Campo M, Yang Q, Jankowsky E, Lambowitz AM (2007) Involvement of DEAD-box proteins in group I and group II intron splicing. Biochemical characterization of Mss116p, ATP hydrolysis-dependent and -independent mechanisms, and general RNA chaperone activity. *J Mol Biol*, 365: 835–855

Handel EM, Cathomen T (2011) Zinc-finger nuclease based genome surgery: it's all about specificity. *Curr Gene Ther*, 11: 28–37

Hanson JH (2013) DNA target site recognition and toward gene targeting in mammalian cells by the Ll. LtrB group II intron RNP. *Ph.D. Dissertation*. The University of Texas at Austin, Austin, TX

Harris DA, Tinsley RA, Walter NG (2004) Terbium-mediated footprinting probes a catalytic conformational switch in the antigenomic hepatitis delta virus ribozyme. *J Mol Biol*, 341: 389–403

Hayden EJ, Ferrada E, Wagner A (2011) Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature*, 474: 92–95

Hayden EJ, Wagner A (2012) Environmental change exposes beneficial epistatic interactions in a catalytic RNA. *Proc Biol Sci*, 279: 3418–3425

Heap JT, Pennington OJ, Cartman ST, Carter GP, Minton NP (2007) The ClosTron: a universal gene knock-out system for the genus Clostridium. *J Microbiol Methods*, 70: 452–464

Heller RC, Marians KJ (2005a) The disposition of nascent strands at stalled replication forks dictates the pathway of replisome loading during restart. *Mol Cell*, 17: 733–743

Heller RC, Marians KJ (2005b) Unwinding of the nascent lagging strand by Rep and PriA enables the direct restart of stalled replication forks. *J Biol Chem*, 280: 34143–34151

Horowitz SB, Tluczek LJ (1989) Gonadotropin stimulates oocyte translation by increasing magnesium activity through intracellular potassium-magnesium exchange. *Proc Natl Acad Sci U S A*, 86: 9652–9656

Ichiyanagi K, Beauregard A, Belfort M (2003) A bacterial group II intron favors retrotransposition into plasmid targets. *Proc Natl Acad Sci U S A*, 100: 15742–15747

Ichiyanagi K, Beauregard A, Lawrence S, Smith D, Cousineau B, Belfort M (2002) Retrotransposition of the Ll.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. *Mol Microbiol*, 46: 1259–1272

Jang S, Sandler SJ, Harshey RM (2012) Mu insertions are repaired by the double-strand break repair pathway of Escherichia coli. *PLOS Genet*, 8: e1002642

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337: 816–821

Johnson CM, Fisher DJ (2013) Site-specific, insertional inactivation of incA in Chlamydia trachomatis using a group II intron. *PLOS One*, 8: e83989

Joung JK, Sander JD (2013) TALENs: a widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol*, 14: 49–55

Kanaya S, Crouch RJ (1983) DNA sequence of the gene coding for Escherichia coli ribonuclease H. *J Biol Chem*, 258: 1276–1281

Karberg M, Guo H, Zhong J, Coon R, Perutka J, Lambowitz AM (2001) Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. *Nat Biotechnol*, 19: 1162–1167

Karberg MS, Lambowitz AM (2006) Group II intron mobility and its applications in biotechnology and gene therapy. *Ph.D. Dissertation*. The University of Texas at Austin, Austin, TX

Keating KS, Toor N, Perlman PS, Pyle AM (2010) A structural analysis of the group II intron active site and implications for the spliceosome. *RNA*, 16: 1–9

Keel AY, Rambo RP, Batey RT, Kieft JS (2007) A general strategy to solve the phase problem in RNA crystallography. *Structure*, 15: 761–772

Kinscherf TG, Apirion D (1975) Polynucleotide phosphorylase can participate in decay of mRNA in *Escherichia coli* in the absence of ribonuclease II. *Mol Gen Genet*, 139: 357–362

Konrad EB, Lehman IR (1974) A conditional lethal mutant of Escherichia coli K12 defective in the 5' leads to 3' exonuclease associated with DNA polymerase I. *Proc Natl Acad Sci U S A*, 71: 2048–2051

Kornberg T, Gefter ML (1971) Purification and DNA synthesis in cell-free extracts: properties of DNA polymerase II. *Proc Natl Acad Sci U S A*, 68: 761–764

Lahr DJ, Katz LA (2009) Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques*, 47: 857–866

Lambowitz AM, Zimmerly S (2004) Mobile group II introns. *Annu Rev Genet*, 38: 1–35

Lambowitz AM, Zimmerly S (2011) Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol*, 3: a003616

Lander ES, Linton LM, Birren B, *et. al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, 409: 860–921

Lang BF, Laforest MJ, Burger G (2007) Mitochondrial introns: a critical view. *Trends Genet*, 23: 119–125

Lehman N, Joyce GF (1993) Evolution in vitro of an RNA enzyme with altered metal dependence. *Nature*, 361: 182–185

Lieber A, Sandig V, Strauss M (1993) A mutant T7 phage promoter is specifically transcribed by T7-RNA polymerase in mammalian cells. *Eur J Biochem*, 217: 387–394

Liu J, Xu L, Sandler SJ, Marians KJ (1999) Replication fork assembly at recombination intermediates is required for bacterial growth. *Proc Natl Acad Sci U S A*, 96: 3552–3555

Long JE, Massoni SC, Sandler SJ (2010) RecA4142 causes SOS constitutive expression by loading onto reversed replication forks in Escherichia coli K-12. *J Bacteriol*, 192: 2575–2582

Lovett ST, Clark AJ (1984) Genetic analysis of the recJ gene of Escherichia coli K-12. *J Bacteriol*, 157: 190–196

Lusk JE, Williams RJ, Kennedy EP (1968) Magnesium and the growth of *Escherichia coli*. *J Biol Chem*, 243: 2618–2624

Lyakhov DL, He B, Zhang X, Studier FW, Dunn JJ, McAllister WT (1997) Mutant bacteriophage T7 RNA polymerases with altered termination properties. *J Mol Biol*, 269: 28–40

Maguire ME (2006) Magnesium transporters: properties, regulation and structure. *Front Biosci*, 11: 3149–3163

Maki H, Horiuchi T, Kornberg A (1985) The polymerase subunit of DNA polymerase III of Escherichia coli. I. Amplification of the dnaE gene product and polymerase activity of the alpha subunit. *J Biol Chem*, 260: 12982–12986

Makowska-Grzyska M, Kaguni JM (2010) Primase directs the release of DnaC from DnaB. *Mol Cell*, 37: 90–101

Mali P, Aach J, Stranges PB, Esvelt KM, Moosburner M, Kosuri S, Yang L, Church GM (2013a) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol*, 31: 833–838

Mali P, Esvelt KM, Church GM (2013b) Cas9 as a versatile tool for engineering biology. *Nat Methods*, 10: 957–963

Malik HS, Burke WD, Eickbush TH (1999) The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol*, 16: 793–805

Marcia M, Pyle AM (2012) Visualizing group II intron catalysis through the stages of splicing. *Cell*, 151: 497–507

Mardis ER (2008) Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, 9: 387–402

Martin W, Koonin EV (2006) Introns and the origin of nucleus-cytosol compartmentalization. *Nature*, 440: 41–45

Masai H, Arai K (1988) Operon structure of dnaT and dnaC genes essential for normal and stable DNA replication of Escherichia coli chromosome. *J Biol Chem*, 263: 15083–15093

Mastroianni M, Watanabe K, White TB, Zhuang F, Vernon J, Matsuura M, Wallingford J, Lambowitz AM (2008) Group II intron-based gene targeting reactions in eukaryotes. *PLOS One*, 3: e3121

Matsuura M, Saldanha R, Ma H, Wank H, Yang J, Mohr G, Cavanagh S, Dunny GM, Belfort M, Lambowitz AM (1997) 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 Dev*, 11: 2910–2924

McCool JD, Long E, Petrosino JF, Sandler HA, Rosenberg SM, Sandler SJ (2004) Measurement of SOS expression in individual Escherichia coli K-12 cells using fluorescence microscopy. *Mol Microbiol*, 53: 1343–1357

Meyer RR, Laine PS (1990) The single-stranded DNA-binding protein of Escherichia coli. *Microbiol Rev*, 54: 342–380

Michel F, Costa M, Westhof E (2009) The ribozyme core of group II introns: a structure in want of partners. *Trends Biochem Sci*, 34: 189–199

Michel F, Ferat JL (1995) Structure and activities of group II introns. *Annu Rev Biochem*, 64: 435–461

Mills DA, Manias DA, McKay LL, Dunny GM (1997a) Homing of a group II intron from Lactococcus lactis subsp. lactis ML3. *J Bacteriol*, 179: 6107–6111

Mills DA, Manias DA, McKay LL, Dunny GM (1997b) Homing of a group II intron from Lactococcus lactis subsp. lactis ML3. *J Bacteriol*, 179: 6107–6111

Mills DA, McKay LL, Dunny GM (1996) Splicing of a group II intron involved in the conjugative transfer of pRS01 in lactococci. *J Bacteriol*, 178: 3531–3538

Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D (2010) Tablet--next generation sequence assembly visualization. *Bioinformatics*, 26: 401–402

Misra VK, Shiman R, Draper DE (2003) A thermodynamic framework for the magnesium-dependent folding of RNA. *Biopolymers*, 69: 118–136

Mohr G, Ghanem E, Lambowitz AM (2010) Mechanisms used for genomic proliferation by thermophilic group II introns. *PLOS Biol*, 8: e1000391

Mohr G, Hong W, Zhang J, Cui GZ, Yang Y, Cui Q, Liu YJ, Lambowitz AM (2013) A targetron system for gene targeting in thermophiles and its application in Clostridium thermocellum. *PLOS One*, 8: e69032

Mohr G, Smith D, Belfort M, Lambowitz AM (2000) Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. *Genes Dev*, 14: 559–573

Mohr S, Matsuura M, Perlman PS, Lambowitz AM (2006) 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

Mussolino C, Cathomen T (2012) TALE nucleases: tailored genome engineering made easy. *Curr Opin Biotechnol*, 23: 644–650

Nisa-Martinez R, Laporte P, Jimenez-Zurdo JI, Frugier F, Crespi M, Toro N (2013) Localization of a Bacterial Group II Intron-Encoded Protein in Eukaryotic Nuclear Splicing-Related Cell Compartments. *PLOS One*, 8: e84056

Niu Y, Shen B, Cui Y, *et. al*. (2014) Generation of Gene-Modified Cynomolgus Monkey via Cas9/RNA-Mediated Gene Targeting in One-Cell Embryos. *Cell*, 156: 836–843

Noah JW, Lambowitz AM (2003) Effects of maturase binding and $Mg^{2+}$ concentration on group II intron RNA folding investigated by UV cross-linking. *Biochemistry*, 42: 12466–12480

Ohmori H (1994) Structural analysis of the rhlE gene of Escherichia coli. *Jpn J Genet*, 69: 1–12

Okazaki R, Arisawa M, Sugino A (1971) Slow joining of newly replicated DNA chains in DNA polymerase I-deficient Escherichia coli mutants. *Proc Natl Acad Sci U S A*, 68: 2954–2957

Olivera BM, Lehman IR (1967) Linkage of polynucleotides through phosphodiester bonds by an enzyme from Escherichia coli. *Proc Natl Acad Sci U S A*, 57: 1426–1433

Peebles CL, Perlman PS, Mecklenburg KL, Petrillo ML, Tabor JH, Jarrell KA, Cheng HL (1986a) A self-splicing RNA excises an intron lariat. *Cell*, 44: 213–223

Peebles CL, Perlman PS, Mecklenburg KL, Petrillo ML, Tabor JH, Jarrell KA, Cheng HL (1986b) A self-splicing RNA excises an intron lariat. *Cell*, 44: 213–223

Perutka J, Wang W, Goerlitz D, Lambowitz AM (2004) Use of computer-designed group II introns to disrupt *Escherichia coli* DExH/D-box protein and DNA helicase genes. *J Mol Biol*, 336: 421–439

Plante I, Cousineau B (2006) Restriction for gene insertion within the Lactococcus lactis Ll.LtrB group II intron. *RNA*, 12: 1980–1992

Poenie M (1990) Alteration of intracellular Fura-2 fluorescence by viscosity: a simple correction. *Cell calcium*, 11: 85–91

Porteus MH, Baltimore D (2003) Chimeric nucleases stimulate gene targeting in human cells. *Science*, 300: 763

Ran FA, Hsu PD, Lin CY, Gootenberg JS, Konermann S, Trevino AE, Scott DA, Inoue A, Matoba S, Zhang Y, Zhang F (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, 154: 1380–1389

Rogozin IB, Carmel L, Csuros M, Koonin EV (2012) Origin and evolution of spliceosomal introns. *Biol Direct*, 7: 11

Rother M, Rother K, Puton T, Bujnicki JM (2011) RNA tertiary structure prediction with ModeRNA. *Brief Bioinform*, 12: 601–613

Rowen L, Kornberg A (1978) Primase, the dnaG protein of Escherichia coli. An enzyme which starts DNA chains. *J Biol Chem*, 253: 758–764

Rubin H (2007) The logic of the Membrane, Magnesium, Mitosis (MMM) model for the regulation of animal cell proliferation. *Arch Biochem Biophys*, 458: 16–23

Russell R, Jarmoskaite I, Lambowitz AM (2013) Toward a molecular understanding of RNA remodeling by DEAD-box proteins. *RNA Biol*, 10: 44–55

Rydberg B, Game J (2002) Excision of misincorporated ribonucleotides in DNA by RNase H (type 2) and FEN-1 in cell-free extracts. *Proc Natl Acad Sci U S A*, 99: 16654–16659

Saldanha R, Chen B, Wank H, Matsuura M, Edwards J, Lambowitz AM (1999) RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. *Biochemistry*, 38: 9069–9083

Sambrook J, Russell DW (2006a) Northern hybridization. *CSH Protoc*, 2006:

Sambrook J, Russell DW (2006b) The inoue method for preparation and transformation of competent *E. coli*: "ultra-competent" cells. *CSH Protoc*, 2006:

Sambrook J, Russell DW (2006c) Transformation of *E. coli* by Electroporation. *CSH Protoc*, 2006:

San Filippo J, Lambowitz AM (2002) Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. *J Mol Biol*, 324: 933–951

Sandig V, Lieber A, Bahring S, Strauss M (1993) A phage T7 class-III promoter functions as a polymerase II promoter in mammalian cells. *Gene*, 131: 255–259

Sandler SJ (2000) Multiple genetic pathways for restarting DNA replication forks in Escherichia coli K-12. *Genetics*, 155: 487–497

Sandler SJ, Marians KJ, Zavitz KH, Coutu J, Parent MA, Clark AJ (1999) dnaC mutations suppress defects in DNA replication- and recombination-associated functions in priB and priC double mutants in Escherichia coli K-12. *Mol Microbiol*, 34: 91–101

Schmelzer C, Schweyen RJ (1986) Self-splicing of group II introns in vitro: mapping of the branch point and mutational inhibition of lariat formation. *Cell*, 46: 557–565

Schmidt U, Podar M, Stahl U, Perlman PS (1996) Mutations of the two-nucleotide bulge of D5 of a group II intron block splicing in vitro and in vivo: phenotypes and suppressor mutations. *RNA*, 2: 1161–1172

Seetharaman M, Eldho NV, Padgett RA, Dayie KT (2006) Structure of a self-splicing group II intron catalytic effector domain 5: parallels with spliceosomal U6 RNA. *RNA*, 12: 235–247

Segal DJ, Carroll D (1995) Endonuclease-induced, targeted homologous extrachromosomal recombination in Xenopus oocytes. *Proc Natl Acad Sci U S A*, 92: 806–810

Shao W, Boltz VF, Spindler JE, Kearney MF, Maldarelli F, Mellors JW, Stewart C, Volfovsky N, Levitsky A, Stephens RM, Coffin JM (2013) Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of Low-frequency drug resistance mutations in HIV-1 DNA. *Retrovirology*, 10: 18

Shi X, Karkut T, Alting-Mees M, Chamankhah M, Hemmingsen SM, Hegedus DD (2003) Enhancing *Escherichia coli* electrotransformation competency by invoking physiological adaptations to stress and modifying membrane integrity. *Anal Biochem*, 320: 152–155

Shinagawa H, Kato T, Ise T, Makino K, Nakata A (1983) Cloning and characterization of the umu operon responsible for inducible mutagenesis in Escherichia coli. *Gene*, 23: 167–174

Sigel RK, Sashital DG, Abramovitz DL, Palmer AG, Butcher SE, Pyle AM (2004) Solution structure of domain 5 of a group II intron ribozyme reveals a new RNA motif. *Nat Struct Mol Biol*, 11: 187–192

Sigel RK, Vaidya A, Pyle AM (2000) Metal ion binding sites in a group II intron core. *Nat Struct Biol*, 7: 1111–1116

Sigel RKO (2005) Group II Intron Ribozymes and Metal Ions A Delicate Relationship. *Eur J Inorg Chem*, 2005: 2281–2292

Simon DM, Kelchner SA, Zimmerly S (2009) A broadscale phylogenetic analysis of group II intron RNAs and intron-encoded reverse transcriptases. *Mol Biol Evol*, 26: 2795–2808

Singh NN, Lambowitz AM (2001) Interaction of a group II intron ribonucleoprotein endonuclease with its DNA target site investigated by DNA footprinting and modification interference. *J Mol Biol*, 309: 361–386

Slagter-Jager JG, Allen GS, Smith D, Hahn IA, Frank J, Belfort M (2006) Visualization of a group II intron in the 23S rRNA of a stable ribosome. *Proc Natl Acad Sci U S A*, 103: 9838–9843

Smith D, Zhong J, Matsuura M, Lambowitz AM, Belfort M (2005) Recruitment of host functions suggests a repair pathway for late steps in group II intron retrohoming. *Genes Dev*, 19: 2477–2487

Snavely MD, Gravina SA, Cheung TT, Miller CG, Maguire ME (1991) Magnesium transport in *Salmonella typhimurium*. Regulation of mgtA and mgtB expression. *J Biol Chem*, 266: 824–829

Solem A, Zingler N, Pyle AM (2006) A DEAD protein that activates intron self-splicing without unwinding RNA. *Mol Cell*, 24: 611–617

Sriskanda V, Shuman S (2001) A second NAD(+)-dependent DNA ligase (LigB) in Escherichia coli. *Nucleic Acids Res*, 29: 4930–4934

Stemmer WP, Crameri A, Ha KD, Brennan TM, Heyneker HL (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene*, 164: 49–53

Stern DB, Goldschmidt-Clermont M, Hanson MR (2010) Chloroplast RNA metabolism. *Annu Rev Plant Biol*, 61: 125–155

Strick R, Strissel PL, Gavrilov K, Levi-Setti R (2001) Cation-chromatin binding as shown by ion microscopy is essential for the structural integrity of chromosomes. *J Cell Biol*, 155: 899–910

Takahashi S, Hours C, Chu A, Denhardt DT (1979) The rep mutation. VI. Purification and properties of the Escherichia coli rep protein, DNA helicase III. *Can J Biochem*, 57: 855–866

Toor N, Keating KS, Taylor SD, Pyle AM (2008a) Crystal structure of a self-spliced group II intron. *Science*, 320: 77–82

Toor N, Rajashankar K, Keating KS, Pyle AM (2008b) Structural basis for exon recognition by a group II intron. *Nat Struct Mol Biol*, 15: 1221–1222

Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res*, 38: e159

Truong DM, Kaler G, Khandelwal A, Swaan PW, Nigam SK (2008) Multi-level analysis of organic anion transporters 1, 3, and 6 reveals major differences in structural determinants of antiviral discrimination. *J Biol Chem*, 283: 8654–8663

Truong DM, Sidote DJ, Russell R, Lambowitz AM (2013) Enhanced group II intron retrohoming in magnesium-deficient Escherichia coli via selection of mutations in the ribozyme core. *Proc Natl Acad Sci U S A*, 110: E3800–9

Ueda K, McMacken R, Kornberg A (1978) dnaB protein of Escherichia coli. Purification and role in the replication of phiX174 DNA. *J Biol Chem*, 253: 261–269

Umezu K, Nakayama K, Nakayama H (1990) Escherichia coli RecQ protein is a DNA helicase. *Proc Natl Acad Sci U S A*, 87: 5363–5367

Uyemura D, Eichler DC, Lehman IR (1976) Biochemical characterization of mutant forms of DNA polymerase I from Escherichia coli. II. The polAex1 mutation. *J Biol Chem*, 251: 4085–4089

Valles Y, Halanych KM, Boore JL (2008) Group II introns break new boundaries: presence in a bilaterian's genome. *PLOS One*, 3: e1488

van der Veen R, Arnberg AC, van der Horst G, Bonen L, Tabak HF, Grivell LA (1986) Excised group II introns in yeast mitochondria are lariats and can be formed by self-splicing in vitro. *Cell*, 44: 225–234

Varani G, McClain WH (2000) The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep*, 1: 18–23

Vernon JL (2010) Toward group II intron-based genome targeting in eukaryotic cells. *Ph.D. Dissertation*. The University of Texas at Austin, Austin, TX

Wagner J, Gruz P, Kim SR, Yamada M, Matsui K, Fuchs RP, Nohmi T (1999) The dinB gene encodes a novel *E. coli* DNA polymerase, DNA pol IV, involved in mutagenesis. *Mol Cell*, 4: 281–286

Wahle E, Lasken RS, Kornberg A (1989) The dnaB-dnaC replication protein complex of Escherichia coli. I. Formation and properties. *J Biol Chem*, 264: 2463–2468

Wei C, Liu J, Yu Z, Zhang B, Gao G, Jiao R (2013) TALEN or Cas9 - rapid, efficient and specific choices for genome modifications. *J Genet Genomics*, 40: 281–289

White TB, Lambowitz AM (2012) The retrohoming of linear group II intron RNAs in Drosophila melanogaster occurs by both DNA ligase 4-dependent and -independent mechanisms. *PLOS Genet*, 8: e1002534

Wu J, Kandavelou K, Chandrasegaran S (2007) Custom-designed zinc finger nucleases: what is next? *Cell Mol Life Sci*, 64: 2933–2944

Xiang Q, Qin PZ, Michels WJ, Freeland K, Pyle AM (1998) Sequence specificity of a group II intron ribozyme: multiple mechanisms for promoting unusually high discrimination against mismatched targets. *Biochemistry*, 37: 3839–3849

Yao J (2008) Bacterial gene targeting using group II intron L1. LtrB splicing and retrohoming. *Ph.D. Dissertation*. The University of Texas at Austin, Austin, TX

Yao J, Lambowitz AM (2007) Gene targeting in gram-negative bacteria by use of a mobile group II intron ("Targetron") expressed from a broad-host-range vector. *Appl Environ Microbiol*, 73: 2735–2743

Yao J, Truong DM, Lambowitz AM (2013) Genetic and biochemical assays reveal a key role for replication restart proteins in group II intron retrohoming. *PLOS Genet*, 9: e1003469

Yao J, Zhong J, Lambowitz AM (2005) Gene targeting using randomly inserted group II introns (targetrons) recovered from an *Escherichia coli* gene disruption library. *Nucleic Acids Res*, 33: 3351–3362

Yao S, Helinski DR, Toukdarian A (2007) Localization of the naturally occurring plasmid ColE1 at the cell pole. *J Bacteriol*, 189: 1946–1953

Zerbato M, Holic N, Moniot-Frin S, Ingrao D, Galy A, Perea J (2013) The brown algae Pl.LSU/2 group II intron-encoded protein has functional reverse transcriptase and maturase activities. *PLOS One*, 8: e58263

Zhao J, Lambowitz AM (2005) A bacterial group II intron-encoded reverse transcriptase localizes to cellular poles. *Proc Natl Acad Sci U S A*, 102: 16133–16140

Zhao J, Niu W, Yao J, Mohr S, Marcotte EM, Lambowitz AM (2008) Group II intron protein localization and insertion sites are affected by polyphosphate. *PLOS Biol*, 6: e150

Zhuang F, Karberg M, Perutka J, Lambowitz AM (2009a) EcI5, a group IIB intron with high retrohoming frequency: DNA target site recognition and use in gene targeting. *RNA*, 15: 432–449

Zhuang F, Mastroianni M, White TB, Lambowitz AM (2009b) Linear group II intron RNAs can retrohome in eukaryotes and may use nonhomologous end-joining for cDNA ligation. *Proc Natl Acad Sci U S A*, 106: 18189–18194

Zimmerly S, Hausner G, Wu X (2001) Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res*, 29: 1238–1250

# Vita

David Minh Truong was born in Santa Barbara, California, and he grew up in North San Diego County, California in the town of Escondido. At Escondido High School, he became interested in genetic engineering from an outstanding teacher, Dale Cornelius, who has inspired many along the way. In 2001, he matriculated at University of California, San Diego at Muir College to study Molecular Biology. As an undergraduate, he worked on generating transgenic mice with Dr. Sanjay K. Nigam, under the guidance of Dr. Wei Wu. After receiving his Bachelor of Science in 2005, he became a research associate in the laboratory of Dr. Sanjay K. Nigam. Between the years 2005-2007, he studied Organic Anion Transporters in mature and embryonic kidneys, proteins important for the excretion of drugs, metabolites, and toxins. In 2007, he entered doctoral studies at The University of Texas at Austin for Cell and Molecular Biology under the guidance of Dr. Alan M. Lambowitz. He was the chairman of the Paul D. Gottlieb Endowed Lecture Series in 2010, and part of the Academy for Future Science Faculty between 2012-2013.

email: dmtruong@utexas.edu

This dissertation was typed by David M. Truong.