2013-12-04

Alternative splicing of an ORF-less group II intron in Clostridium tetani

McNeil, Bonnie

UNIVERSITY OF CALGARY

Alternative splicing of an ORF-less group II intron in *Clostridium tetani*

by

Bonnie Ashley McNeil

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOLOGICAL SCIENCES

CALGARY, ALBERTA

NOVEMBER, 2013

Abstract

Group II introns are ribozymes that are encoded within all domains of life. They are also capable of mobility through an RNA intermediate. Due to similarities in RNA structure and splicing mechanisms, group II introns are thought to have been the ancestors of nuclear pre-mRNA introns and snRNAs.

In this dissertation I report the discovery of a unique ORF-less group II intron, *C.te*.I1, in the human pathogen *Clostridium tetani*. The intron is encoded within a surface layer protein region of the *C. tetani* chromosome and possesses an unusual genomic organization such that a full-length copy of the intron is followed downstream by three copies of the RNA structural domains 5 and 6 (D5/6). This arrangement led to the hypothesis that *C.te.*I1 is capable of alternative splicing utilizing the downstream copies of D5/6 as alternate 3′ splice sites. RNA extractions and RT-PCR support the hypothesis and revealed that the splicing reaction of *C.te*.I1 links a surface layer protein ORF (CTC00465) in the upstream exon to one of four downstream ORFs that encode transglutaminase-related or protease-related reading frames (CTC00467-CTC00470). Including unspliced transcript, five mRNAs are produced.

Sequencing of the exon junctions showed that the 5′ splice site utilized by *C.te*.I1 is shifted 8 nt upstream both *in vivo* and *in vitro*. Use of this splice site is critical to alternative splicing as it results in the elimination of the stop codon at the end of CTC00465 and results in the correct ligation of 5′ and 3′ exon sequences. Site-directed mutagenesis and self-splicing assays for *C.te.*I1 revealed that the shifted splice site is due to a novel EBS1-IBS1 pairing. Although *C.te.*I1 is thought to be derived from a mobile Class B (IIB) intron that lost its ORF, the intron was found to have evolved to utilize a

IIA-like mechanism of 3′ splice site recognition. These changes represent structural adaptations of the intron to its role in alternative splicing. The structural adaptations and splicing of *C.te.*I1 illustrate the plasticity of group II introns in that they can adapt new RNA structural and catalytic properties which can be utilized to affect gene expression.

Table of Contents

List of Tables

List of Figures

List of Symbols, Abbreviations and Nomenclature

Å	Angstrom
A	adenine
ACN	acetonitrile
AS	anti-sense
ATCC	American Type Culture Collection
BHI	Brain Heart Infusion media
bp	base pair
C	cytosine
cDNA	complementary DNA
CIA	chloroform isoamyl alcohol
CL	chloroplast-like
CTC	*Clostridium tetani* chromosome
CWP	cell wall protein
D	DNA binding domain
DNA	deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
Dscam	*Down syndrome cell adhesion molecule*
DTT	dithiothreitol
EBS	exon binding sequence
EDTA	ethylenediaminetetraacetic acid
En	endonuclease domain
EtOH	ethanol
G	guanine
GMP	guanosine monophosphate
HDV	hepatitis delta virus
hnRNP	heterogeneous nuclear RNP
IBS	intron binding sequence
IEP	intron encoded protein
ISL	intramolecular stem loop
LB	Luria Bertani broth
LC-MS/MS	liquid chromatography mass spectrometry (tandem MS)
LECA	last eukaryotic common ancestor
LTR	long terminal repeat
MALDI-ToF	matrix assisted laser desorption ionization time of flight
mRNA	messenger RNA
miRNA	micro RNA
mtDNA	mitochondrial DNA
ML	mitochondrial-like
MS	mass spectrometry
N	any nucleotide
NAIM	nucleotide analogue interference mapping
NaOAc	sodium acetate
NCBI	National Centre for Biotechnology Information

ncRNA	non-coding RNA
NHEJ	non-homologous end joining
NH_4OAC	ammonium acetate
NH_4HCO_3	ammonium bicarbonate
NMR	nuclear magnetic resonance
nt	nucleotide
NTP	nucleoside triphosphate
OD	optical density
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate buffered saline
PCR	polymerase chain reaction
pKS+	pBluescript KS II +
qRT-PCR	quantitative real-time reverse transcriptase PCR
R	any purine (A or G)
RNA	ribonucleic acid
RNAi	RNA interference
RNP	ribonucleoprotein
rRNA	ribosomal RNA
RT	reverse transcriptase
RT-PCR	reverse transcriptase polymerase chain reaction
SDS	sodium dodecyl sulfate
SELEX	systematic evolution of ligands by exponential enrichment
SER	spliced exon reopening
SF	splice form
SLH	surface layer homology
SLP	surface layer protein
SR	serine arginine
sRNA	small RNA
siRNA	small interfering RNA
snRNA	small nuclear RNA
SSC	self-splicing construct
T	thymine
TFA	trifluoroacetic acid
TPRT	target-primed reverse transcription
tRNA	transfer RNA
U	urasil
UTP	uridine triphosphate
UTR	untranslated region
VS	Varkud satellite
WGS	whole genome shotgun
WT	wild type
X	maturase
Y	any pyrimidine (C,T,U)
ZnOAc	zinc acetate

Chapter One: **INTRODUCTION**

1.1 Our RNA World

The independent discovery of two unique catalytic RNAs in the early 1980s resulted in the Nobel Prize in Chemistry being jointly awarded to Thomas Cech, for the discovery of group I introns, and to Sydney Altman, for his work on Ribonuclease P (RNase P). Specifically, the Cech lab discovered that the intervening sequence in the *Tetrahymena thermophila* 26S rRNA was capable of auto-excision in the presence of monovalent and divalent cations and a guanosine nucleotide cofactor (Cech et al. 1981; Kruger et al. 1982). Altman's group noted that the RNA moiety of RNase P, from both *Escherichia coli* and *Bacillus subtilis,* was able to cleave tRNA precursors in the absence of protein under elevated levels of magnesium (Guerrier-Takada et al. 1983). These two discoveries revealed the catalytic potential of RNA and marked the beginning of an explosion in knowledge and understanding of the complex nature of RNA.

Since their initial discovery, many new types of naturally occurring ribozymes (catalytic RNAs) have been identified. These include one additional class of large ribozymes, group II introns (Schmelzer and Schweyen 1986), and a handful of small ribozymes including the hammerhead (Forster and Symons 1987), hairpin (Buzayan et al. 1986), hepatitis delta virus (HDV) (Chen et al. 1986; Kuo et al. 1988), Varkud satellite (VS) (Saville and Collins 1990) and the *glmS* (Winkler et al. 2004) ribozymes. In addition to these ribozymes, many complex cellular machines have been found to be RNA catalyzed. The ribosome (Noller 1993; Steitz and Moore 2003) and spliceosome (Valadkhan et al. 2009) are RNA catalyzed and, as such, are ribozymes at their core, even though they cannot function without their associated proteins.

The discovery of catalytic RNAs re-invigorated the notion that RNA may have been the primordial biological molecule (Rich 1962) and led to the birth of the RNA world hypothesis (Gilbert 1986). The RNA world hypothesis states that a single type of molecule, RNA, carried out both critical functions of information storage and catalysis in the first organisms. From this RNA-only organism, RNA developed the ability to synthesize proteins, and proteins were subsequently used to synthesize the more chemically stable molecule, DNA. The fact that the ribosome is RNA catalyzed in all known organisms supports this hypothesis, as even in modern cells, RNA catalysis is at the heart of protein synthesis. In addition, RNA SELEX (Systematic Evolution of Ligands by Exponential Enrichment) experiments have shown that RNA can direct its own replication and therefore RNA-catalyzed RNA polymerization would be plausible in a primordial world (Johnston et al. 2001; Zaher and Unrau 2007; Wochner et al. 2011). Lending further credence to the hypothesis, activated pyrimidine ribonucleotides have been synthesized from plausible prebiotic conditions (Powner et al. 2009), suggesting RNA may have been the primordial molecule from which all life arose. Despite the simple nature of this explanation and the support that has accrued for the hypothesis, the view still has its critics. Whether the first life form was RNA-based will likely never be definitively known.

In recent years the known cellular functions of RNA have expanded well beyond its role as an intermediary genetic molecule. In addition to its catalytic potential, RNA has been shown to be capable of other functions and roles within the cell and is used extensively to regulate gene expression. One of the main mechanisms of RNA based regulation (or riboregulation) in prokaryotes is the use of riboswitches. These structured

RNAs are typically located in the 5′ UTRs of genes and they bind and sense levels of various metabolites within the cell. Conformational changes occur upon binding of a metabolite to an aptamer domain which leads to regulation of gene expression either through transcriptional or translational control [for reviews see (Tucker and Breaker 2005; Coppins et al. 2007; Serganov and Nudler 2013)]. Small, non-coding RNAs (sRNAs) have also been found to be encoded by most if not all bacterial species. These sRNAs often act in concert with the RNA binding protein Hfq to promote or inhibit translation of their target RNAs and are responsible for regulation and control of numerous cellular processes [reviewed in (Gottesman 2004; Toledo-Arana et al. 2007; Storz et al. 2011)]. In higher eukaryotes, small interfering RNAs (siRNAs) and micro RNAs (miRNAs) act through the RNA inference (RNAi) pathway to mediate post-transcriptional gene silencing (Carthew and Sontheimer 2009).

The possibility of RNA mediated gene regulation and the ability of RNA to sense and respond to various metabolites is no longer limited to the natural world either as RNA synthetic biology is developing as part of the emerging synthetic biology field [reviewed in (Isaacs et al. 2006; Liang et al. 2011)]. The goal of synthetic biology is to develop engineered biological systems to be used for a variety of functions, such as bioremediation, biofuels, manufacturing and medicine. While synthetic biology has traditionally focused on DNA-protein interactions to control transcription, the malleable nature of RNA has made it an attractive molecule to utilize to achieve programmable modular control. Another attractive feature of RNA in synthetic biology is that RNA-based post-transcriptional control generally occurs on faster time scales than transcriptional control strategies.

Components of synthetic RNA circuits can be harvested from nature and refined (Chen et al. 2009; Wieland et al. 2009), computationally designed (Salis et al. 2009), or evolved through *in vitro* SELEX (Ellington and Szostak 1990; Bartel and Szostak 1993; Jenison et al. 1994) and subsequent *in vivo* selection (Weigand et al. 2008; Sinha et al. 2010). RNA-based control elements have been designed that respond to temperature, nucleic acids, small molecules and proteins. An example with environmental applications that has been developed is an RNA control system that responds to the small molecule, atrazine, a widely used herbicide and a toxic environmental pollutant (Sinha et al. 2010). The atrazine aptamer was linked to the *cheZ* gene to control mobility in *E. coli*. This allowed cells to move along atrazine gradients and destroy the compound by converting it to the less harmful compound, hydroxyatrazine. RNA elements are also being developed to sense levels of various compounds within the cell and are being used for health care applications (Khalil and Collins 2010).

Regardless of whether the initial life forms were born into a strictly RNA world, it is becoming more and more clear that the world we currently live in is intensely regulated and influenced by the diversity and complexity of RNA and its biological functions. This thesis focuses on one such example. It addresses the identification and subsequent characterization and of a unique group II intron sequence located within a surface layer protein gene region of the pathogen *Clostridium tetani*. The intron's organization and novel structural features appear as adaptations to its role in alternative splicing, a process previously unknown to occur within prokaryotes, and represents a fascinating example of the domestication of a mobile-element by a bacterium to regulate gene expression.

1.2 Group II Introns

1.2.1 Introduction

Group II introns are mobile genetic elements that are found within bacterial and archaeal genomes, as well as the organellar genomes of some eukaryotes, including protists, plants and fungi (Copertino and Hallick 1993; Michel and Ferat 1995; Bonen and Vogel 2001; Dai and Zimmerly 2003; Lambowitz and Zimmerly 2004; Toro et al. 2007). In addition, a few occurrences of group II introns in animal mtDNAs have been reported (Dellaporta et al. 2006; Sinniger et al. 2007; Vallès et al. 2008). Although not expressed in nuclear genomes, some group II introns have been found embedded within silent mtDNA derived segments of nuclear chromosomes in *Arabidopsis thaliana* (Lin et al. 1999).

In organelles, most group II introns are found within essential housekeeping genes (Bonen and Vogel 2001). These introns are therefore true introns and are subject to selective pressures to maintain splicing functions. As such, inactive group II introns are rare in organelles, but degeneration of RNA structure and their associated intron encoded protein (IEP) are in fact common (Michel and Ferat 1995; Zimmerly et al. 2001; Barkan 2004). Therefore the degenerate group II introns of organelles often rely either on *trans*-acting group II intron IEPs encoded elsewhere in the genome or on other host nuclear encoded factors for splicing (Bonen and Vogel 2001; Bonen 2008).

Approximately a quarter of eubacterial genomes contain group II introns (Candales et al. 2012) and group II introns are distributed within a number of eubacterial phyla, including Acidobacteria, Actinobacteria, Bacteroidetes/Chlorobi, Cyanobacteria, Firmicutes and Proteobacteria (Simon et al. 2008) (Figure 1). However in bacteria, the

title “intron” is often a misnomer as these elements are frequently located outside of genes and it is mobility that is thought to be their dominant function (Toor et al. 2001; Dai and Zimmerly 2002; Klein and Dunny 2002). In fact, more than half of known bacterial group II introns are encoded within predicted intergenic regions with the remainder of introns found mainly in mobile DNAs or within hypothetical ORFs. Only 8% of known bacterial group II introns are found to be encoded within housekeeping genes (Simon et al. 2008).

Bacterial group II introns typically consist of a catalytic RNA and an intron encoded protein (IEP). These “introns” are capable of self-splicing (ie. excising themselves from an RNA transcript) *in vitro* at elevated monovalent and divalent ion concentrations (Peebles et al. 1986; Schmelzer and Schweyen 1986; Van der Veen et al. 1986). However, they require the aid of the IEP to splice *in vivo* and at physiologically relevant conditions *in vitro.* As such, group II introns that lack the IEP are rare in bacteria and all known examples occur in genomes that harbour other group II introns with their associated protein (Meng et al. 2005; Van der Auwera et al. 2005; Simon et al. 2008). Introns containing degenerate IEPs in bacteria are also rare, with only one example clearly identified (Simon et al. 2008). Despite the lack of introns with degenerate IEPs, there are a very large number of fragmented and inactive group II introns found within bacterial genomes (Candales et al. 2012). This reflects their retroelement nature and intergenic insertion as there is little or no selective pressure to maintain splicing function.

Although group II introns share little sequence conservation at the primary level, they possess a conserved secondary structure that consists of six helical domains (D1-D6), which emanate from a central single-stranded hub (Figure 2). Many long range

tertiary interactions (such as α-α′, β-β′, κ-κ′, λ-λ′, and ζ-ζ′) allow the RNA to fold and form the catalytically active tertiary structure, and are noted on the secondary structure diagram in Figure 2. Exon binding sequences (EBS1, EBS2, and EBS3) are also present within the structure and play important roles in exon recognition (Figures 2 and 3). How these regions contribute to exon recognition will be discussed in depth in Chapter 3 of this thesis.

The prototypical group II intron IEP possesses a start codon located within domain 4 of the intron and consists of up to four functional domains: a reverse transcriptase (RT), a maturase (X), a DNA binding domain (D) and an optional endonuclease (En) domain (Figure 2). The IEP aids in splicing *in vivo* and contributes mobility properties to the intron. Specific features of the IEP will be discussed in the subsequent IEP Structure subsection of this chapter.

Group II introns share mechanistic features with spliceosomal introns and non-LTR retroelements and have been proposed to be the evolutionary ancestors of both elements. The structure, function, and evolution of these large ribozymes will be discussed in greater depth in the subsequent sections of this chapter.

1.2.2 Group II Intron RNA Structural and IEP Classes

Group II introns are divided into three RNA structural families based on overarching similarities in secondary structure: IIA, IIB and IIC (Figure 3) (Michel et al. 1989; Toor et al. 2001). These families share both RNA structural features and their associated biochemical properties, as such prediction of RNA structure can subsequently be used to predict ribozyme properties. Of these families, IIC introns are the most recently discovered and have the smallest, most compact RNA structure. IIC introns have

Figure 1. Distribution of group II introns in bacteria.

The intron encoded ORF phylogeny is presented as a radial phylogram with the groups of host bacteria coded by colour. Black dots indicate nodes with greater than 70% bootstrap support. Figure taken from Simon *et al*. (2008).

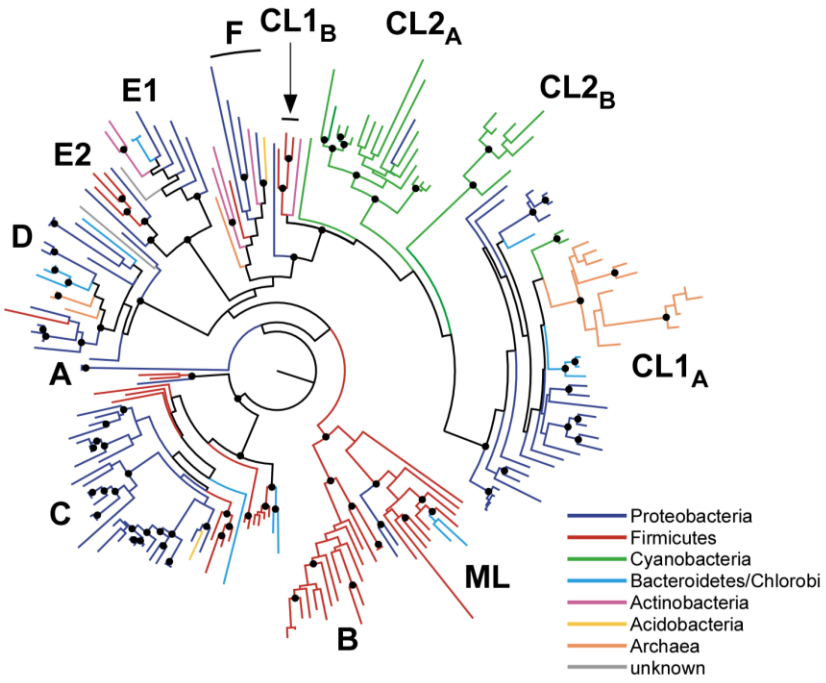


Figure 2. General structure of a group II intron**.**

(A) RNA secondary structure consisting of six domains (1-6) highlighting many important tertiary interactions. (B) Standard structure of a group II intron IEP, including the RT domain, X (maturase) domain, D (DNA binding) domain and optional En (endonuclease) domain. Figure adapted from Simon *et al*. (2008).

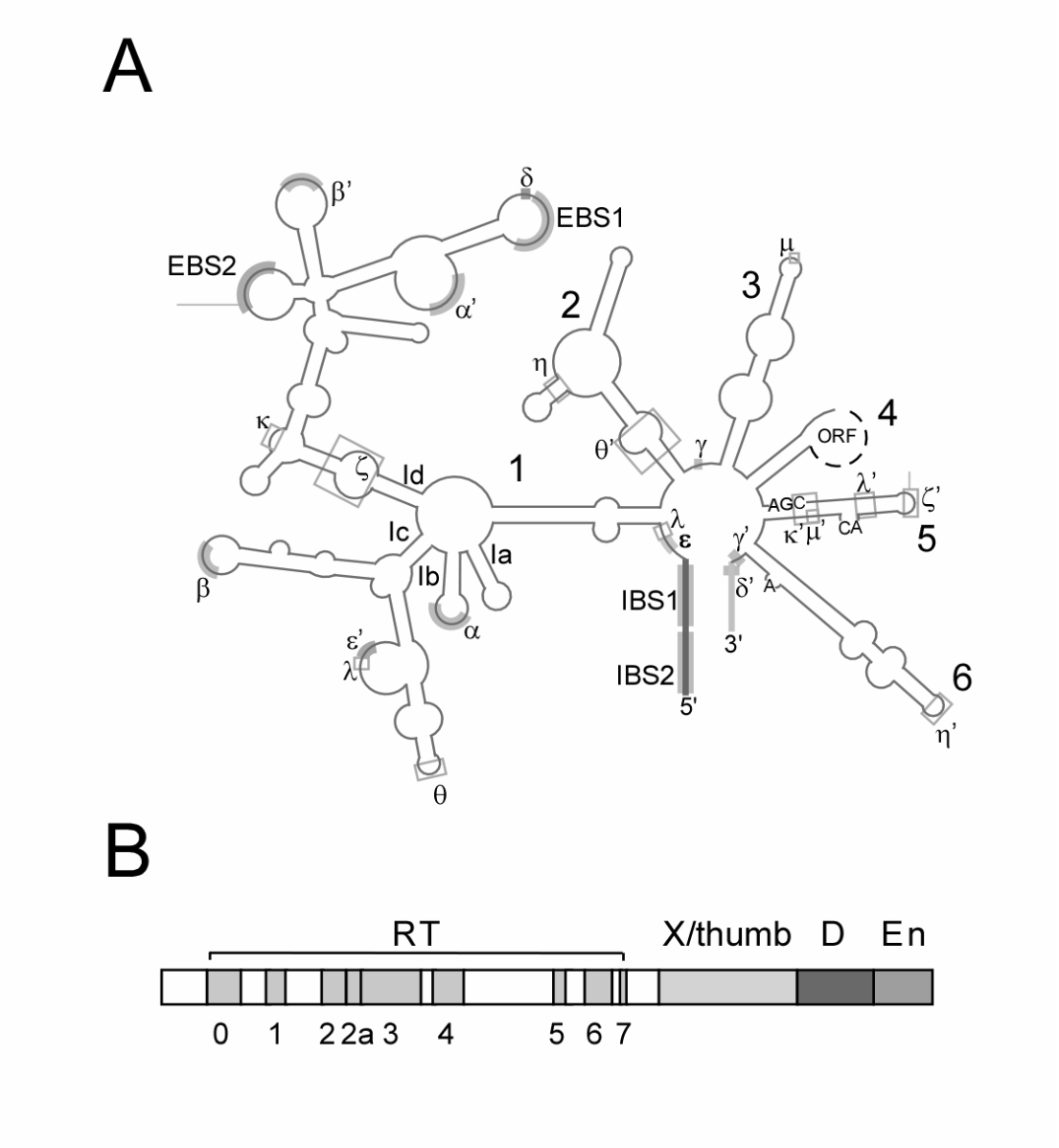
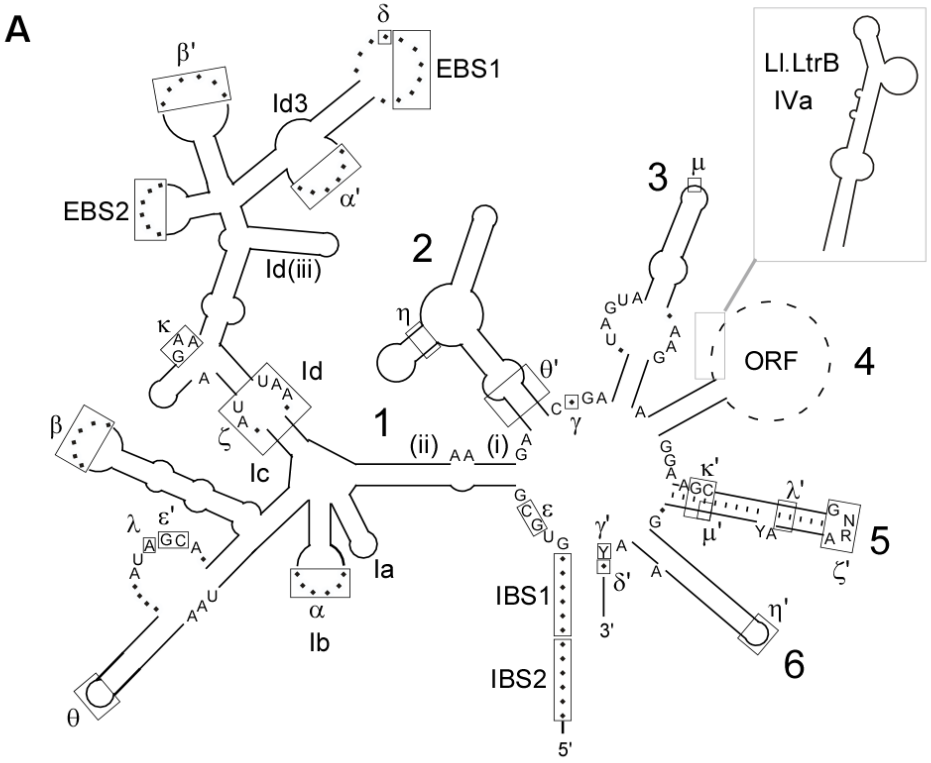
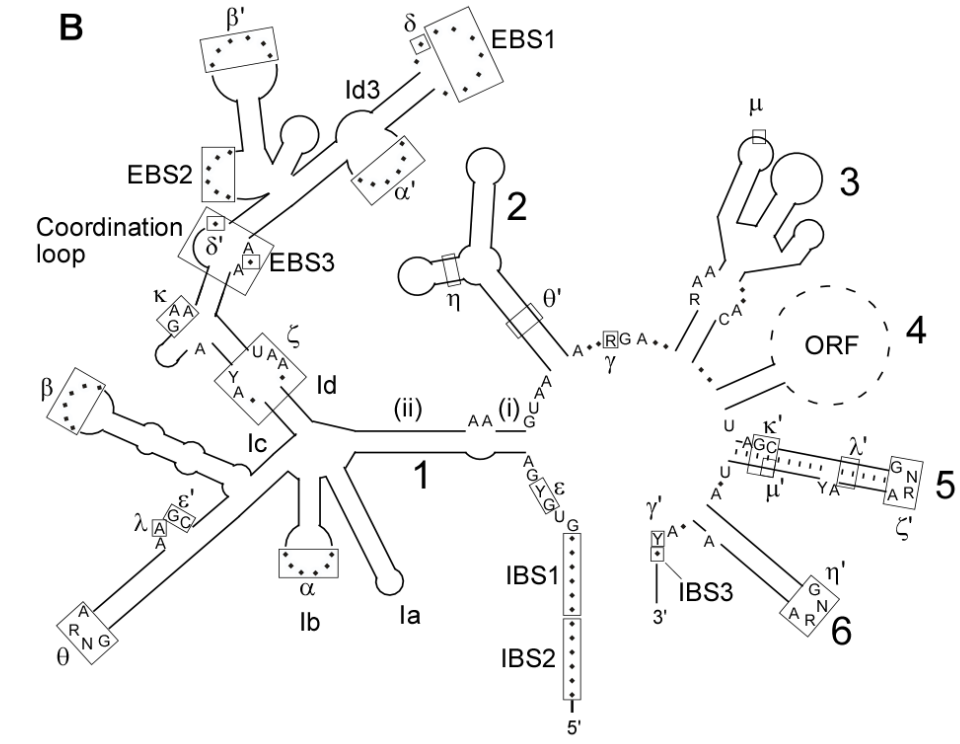
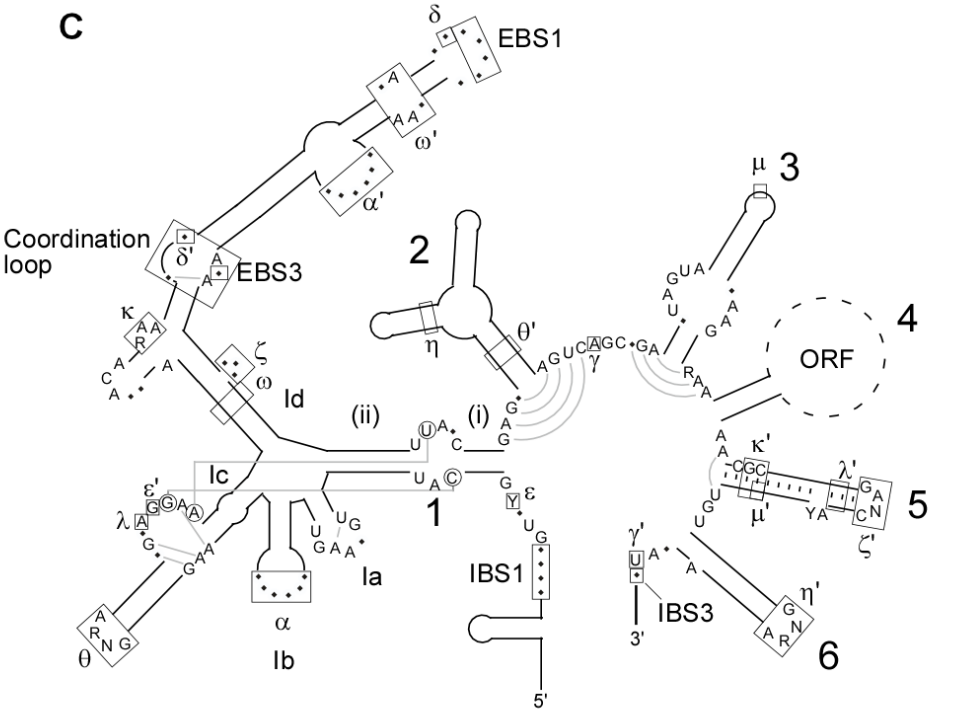


Figure 3. RNA structural families of group II introns.

Secondary structures are shown highlighting stem-loop structures and bulges shared by the various structural families. (A) shows a IIA RNA secondary structure. The location of the high affinity binding site for LtrA is shown for the Ll.LtrB intron in an inset panel. (B) shows a IIB RNA secondary structure and (C) shows a IIC secondary structure. The locations of tertiary interactions are indicated by boxes and Greek letters. Highly conserved nucleotide identities are also shown. Regions involved in exon recognition (EBS and IBS sequences) are also indicated. The coordination loop structure is highlighted in the IIB and IIC structural types that possess it. Figure courtesy of Dr. Steven Zimmerly.




B
β'
δ
EBS1
Id3
EBS2
α'
Coordination loop
δ'
EBS3
κ
ζ
Id
β
(ii)
(i)
Ic
1
2
η
θ'
γ
3
μ
ORF
4
κ'
λ'
μ'
5
ζ'
ε'
λ
ε
γ'
IBS3
η'
6
α
Ib
Ia
IBS1
IBS2
θ
3'
5'


C
δ
EBS1
ω'
α'
Coordination loop
δ'
EBS3
κ
ζ
ω
Id
(ii)
(i)
Ic
ε'
λ
1
2
η
θ'
γ
3
μ
ORF
4
κ'
λ'
μ'
5
ζ'
ε
γ'
IBS3
η'
6
α
Ib
Ia
IBS1
θ
3'
5'

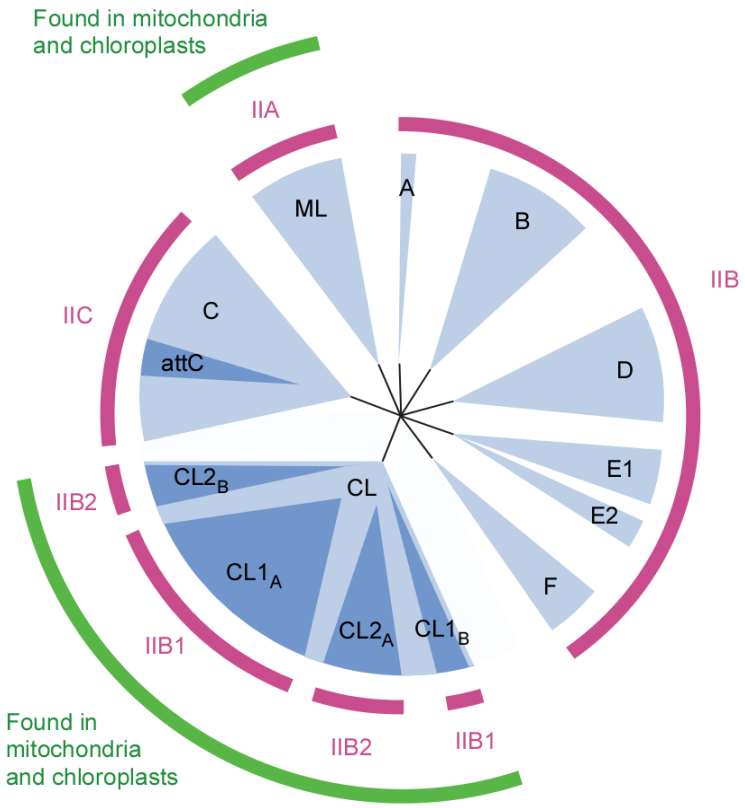
been suggested to be the ancestral group II intron family (Rest and Mindell 2003; Simon et al. 2009). In the case of the IIA and IIB RNA structural families, they can be further subdivided into subtypes: IIA1, IIA2, IIB1 and IIB2 (Michel et al. 1989).

An alternate classification system based on IEP phylogeny is also commonly used that categorizes group II introns into 8 classes: ML (mitochondrial-like), CL (chloroplast-like) Bacterial A, B, C, D, E and F (Zimmerly et al. 2001; Toro et al. 2002; Simon et al. 2008). Classes E and CL can be further subdivided into E1, E2 (Simon et al. 2008), $CL1_A$, $CL1_B$, $CL2_A$ and $CL2_B$ (Zimmerly et al. 2001; Simon et al. 2008; Simon et al. 2009). Of particular note to this thesis, bacterial Class B has also been subdivided into two separate lineages, α and β (Stabell et al. 2009). As group II intron RNA structures have co-evolved with their IEP (Fontaine et al. 1997; Toor et al. 2001) each of these IEP classes has an associated consensus RNA structure. Of these IEP classes, only ML and CL are associated with organellar introns (Figure 4); however, representatives of all 8 classes are encoded within bacterial genomes (Figure 1) (Simon et al. 2009).

When correlating the large RNA structural families to IEP classes we find that the IIA family is composed of exclusively ML introns and the IIC family contains exclusively bacterial Class C introns. All other intron subtypes belong to the IIB RNA structural family. The CL1 and CL2 IEPs associate with standard IIB1 and IIB2 RNA structures, while bacterial Classes A, B, D, E and F possess less typical IIB structures (Figure 4) (Simon et al. 2009).

Figure 4. Group II intron lineages.

The major lineages of group II intron IEPS are shown in light blue: ML (mitochondrial-like), CL (chloroplast like), and Bacterial A-F. Notable sublineages are shown in darker blue. The correspondence of RNA structural class to IEP lineages is shown in pink. As all intron lineages are found in bacteria, only the ones that are also found in organelles are delineated in green. Figure taken from (Lambowitz and Zimmerly 2011).



In addition to the group II introns with standard structures, special lineages containing either 5′ or 3′ terminal extensions have been identified. The introns possessing 3′ terminal extensions consist of a small subgroup of Class B introns containing two additional stem loops between D6 and the 3′ splice site, referred to as Domain 7 (Stabell et al. 2007; Stabell et al. 2009). The introns containing the 5′ terminal extensions are also a subgroup of IIB introns consisting of 10 known mitochondrial introns located within rRNA genes. These introns possess a shortened D6 as well as other structural anomalies and are found to have 5′ insertions of 1-33 nucleotides (Li et al. 2011b). Another lineage that possesses large (~300-600 nt) insertions near the 5′ end exists for a subgroup of $CL1_A$ introns (Adamidi et al. 2003; Ferat et al. 2003; Michel et al. 2007). The existence of many different RNA structural forms of group II introns suggest that group II introns have been subjected to multiple different evolutionary trajectories and the malleable nature of the RNA structure has allowed for adaptations of these introns to their various niche environments.

1.2.3 Group II Intron Splicing

Most group II introns splice through a branching (lariat) pathway both *in vitro* and *in vivo* (Figure 5). This pathway is mechanistically identical to that utilized by spliceosomal introns, which is one of the reasons group II introns have been proposed to be the evolutionary ancestors of spliceosomal introns. The splicing reaction proceeds through two sequential transesterifications and is initiated by a bulged adenosine residue in D6 (Peebles et al. 1987; Jarrell et al. 1988). The 2′ hydroxyl of the adenosine initiates a nucleophilic attack on the activated phosphodiester linkage at the 5′ splice site (Padgett et al. 1994). This results in the formation of a 2′-5′ phosphodiester bond between the bulged

adenosine residue and the first nucleotide of the intron. Structurally this is aided by a kinking of the phosphate backbone at the 5′ exon-intron boundary mediated by the EBS sites of the intron (De Lencastre et al. 2005; Chan et al. 2012). This first step is the rate-limiting step in the splicing reaction (Daniels et al. 1996). A conformational change then occurs in the ribozyme structure that positions the 3′ splice site in the active site of the intron and allows for the second transesterification (Chanfreau and Jacquier 1996). The 3′ hydroxyl of the 5′ exon then attacks the 3′ splice site resulting in the production of ligated exons and free intron lariat. It is likely that the 3′ exon-intron boundary is also kinked to reveal the scissile phosphate (Chan et al. 2012). During the splicing reaction, an inversion of the configuration is observed at the phosphorous indicating that the reaction proceeds through an S_N2 mechanism (Podar et al. 1995).

In some conditions *in vitro* and for some natural intron variants, a hydrolysis pathway is preferred (Jarrell et al. 1988; Vogel and Börner 2002; Li-Pook-Than and Bonen 2006; Toor et al. 2006). This has been noted using potassium containing buffers *in vitro* (Jarrell et al. 1988), for the splicing of IIC introns (Toor et al. 2006), and for a subset of IIB introns containing 5′ terminal extensions (Li et al. 2011b). The reaction mechanism is the same as for the lariat pathway except that the first nucleophilic attack is initiated by a water molecule. This results in ligated exons and linear intron products (Figure 5).

A third pathway which has been reported for some introns is the formation of circles (Jarrell 1993; Murray et al. 2001; Li-Pook-Than and Bonen 2006; Molina-Sánchez et al. 2006). The formation of intron circles is thought to be a biproduct of spliced-exon reopening (SER) and occurs as a result of *trans*-splicing *in vivo*. During SER ligated

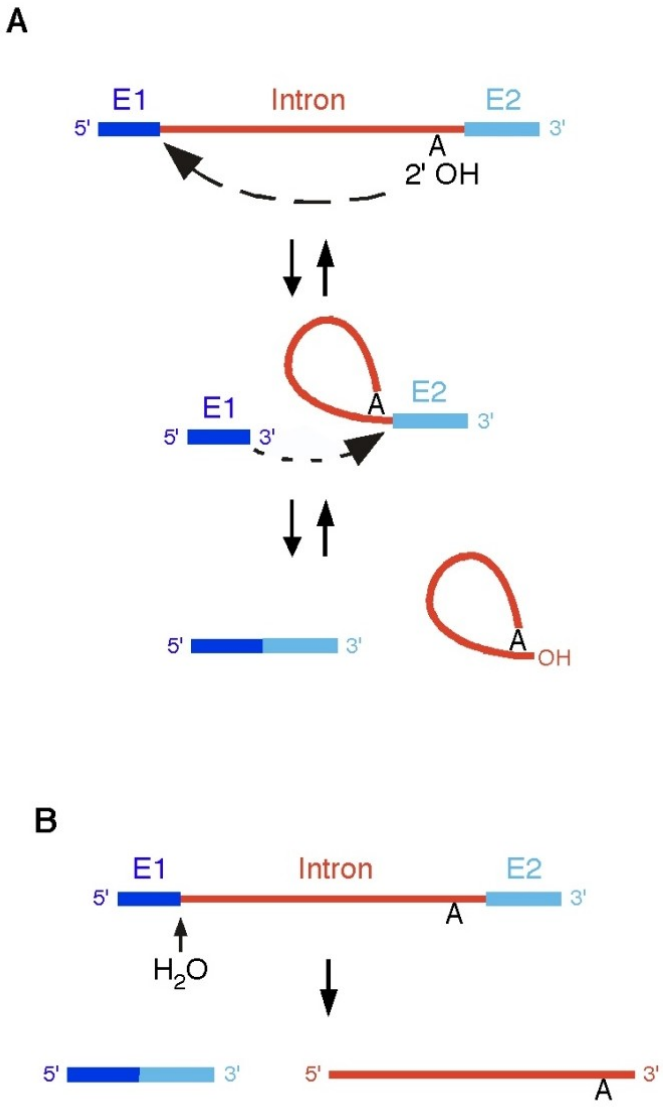
exons are hydrolytically cleaved into free 5′ and 3′ exons. The 3′ hydroxyl of the 5′ exon is thought to attack the 3′ splice site of another unspliced precursor RNA in *trans*, releasing the 3′ intron terminus while re-ligating the exons. In the second step of circle formation then, the 3′ OH of the second intron transcript attacks the 5′ splice site releasing intron circles and regenerating the free 5′ exon [reviewed in (Pyle 2010)].

While most group II intron splicing occurs in *cis*, group II introns have also been shown to be capable of *trans*-splicing. *Trans*-splicing occurs naturally in a variety of different organellar introns (Bonen 1993; Malek and Knoop 1998; Bonen and Vogel 2001; 2008) but no natural *trans*-splicing bacterial group II introns have been reported. In the case of *trans*-splicing, different segments of the intron are encoded on separate RNA transcripts. For naturally occurring *trans*-splicing organellar introns, fragmentation most frequently occurs within D4 of the intron. As such D1, D2, D3 and fragments of D4 are encoded on a single transcript, while the other half of D4, D5 and D6 are encoded on a separate transcript. The intron fragments then associate together using long-range tertiary interactions and form the active ribozyme.

Despite the fact that there are no known natural bacterial *trans*-splicing introns, the Ll.LtrB intron from *Lactococcus lactis* has been shown to be capable of *trans*-splicing in *E. coli* (Belhocine et al. 2007; 2008). A Tn5 based genetic screen was used to fragment the intron and subsequently it was tested for *trans*-splicing. This screen revealed that a greater range of fragmentation sites can be tolerated by group II introns than those observed naturally (Belhocine et al. 2008). This work highlights that although group II introns have not naturally been found to *trans*-splice in bacteria, they are highly versatile and are capable of accurately *trans*-splicing their flanking exons *in vivo*.

Figure 5. Group II intron splicing pathways**.**

(A) Branching pathway utilized by most group II introns. Double arrows indicate that both steps of the reaction are reversible. (B) Hydrolysis pathway. Intermediate products and reaction are not shown but would be similar to the second step of the branching pathway except that the intron-3′ exon intermediate product would be linear. Adapted from (Lambowitz and Zimmerly 2011).



1.2.4 Group II Intron RNA Structure

Since the discovery of group II introns, much work has been directed towards determining how the structure confers catalytic activity to the intron RNA. Biochemical and phylogenetic studies have determined that an extensive network of tertiary interactions between nucleotides contribute to the structure and catalysis of the ribozyme [reviewed in (Michel and Ferat 1995; Lambowitz and Zimmerly 2004; Michel et al. 2009; Pyle 2010; Lambowitz and Zimmerly 2011)]. In this section, I will first describe the long range tertiary interactions present within the various domains of the intron. The reader is encouraged to refer back to Figures 2 and 3 for the locations of interactions and to compare tertiary interactions between RNA structural families. I will then briefly discuss features of the active catalytic structure (Figure 6).

1.2.4.1 Domain 1-The Structural Scaffold

The overall architecture of the active ribozyme structure is a result of complex interactions both within and between domains. *In vitro* studies suggest that D1, the largest group II intron domain, is the first domain to fold and subsequently acts as a scaffold for intron assembly (Qin and Pyle 1998; Fedorova and Zingler 2007; Pyle et al. 2007). Together D1 and D5 can be considered the minimal catalytic core of the intron (Koch et al. 1992; Michels and Pyle 1995). This large domain has a bipartite structure in which two lobes, with relatively separate functions, are linked by the Id stem (see Figures 2 and 3 for stem nomenclature). The upper region is mainly involved in exon recognition (discussed in Chapter 3) while the lower region is actively involved in the creation of the catalytic core. Two intra-domain base-pairing interactions (α-α′ and β-β′) in D1 serve to fold the domain back upon itself bringing the region involved in exon recognition in

contact with the active site of the intron. α-α′ is a kissing loop interaction that is essential for efficient catalysis; however, its role is purely structural as the exact nucleotide sequence is not conserved (Harris-Kerr et al. 1993). The β-β′ pairing is a similar kissing loop type interaction that is found in some IIA and IIB introns, however this interaction is non-essential and is missing from a number of group II introns.

The ε-ε′ interaction is also internal to D1. It involves a conserved pairing between a GY (where Y is any pyrimidine) sequence in an internal helical bulge of the Ic1 stem and the GY present at positions 3 and 4 of the intron within the canonical 5′ GUGYG intron boundary sequence (Jacquier and Michel 1990). The asymmetric bulge that contributes the ε′ sequence from the Ic1 stem differs between RNA structural types (Figure 3) (Michel et al. 1989; Granlund et al. 2001; Toor et al. 2001). The ε-ε′ interaction has long been known to assist in anchoring the 5′ splice site within the catalytic core of the intron (Jacquier and Michel 1990) and it has been shown that ε may be involved in chemical catalysis at the active site (Boudvillain et al. 2000).

Additionally, D1 contains many inter-domain contacts (κ-κ′, ζ-ζ′, and λ-λ′) that co-ordinate the overall structure of the intron. The ζ-ζ′ interaction is a tetraloop-tetraloop receptor interaction between the distal loop of D5 and the tetraloop receptor found in the Id1 stem (Michel and Ferat 1995). The κ-κ′ interaction is involved in docking the base of D5 using a conserved GAA present within the Id stem (Boudvillain and Pyle 1998) and the λ-λ′ interactions anchors the 5′ splice site within the catalytic core of the intron (Boudvillain et al. 2000). These contacts play critical roles in directing folding, anchoring D5, and stabilizing the catalytic core of the intron.

1.2.4.2 Domains 2, 3 and 4

Domains 2 and 3 are considered non-essential for catalysis but enhance catalytic efficiency and stabilize the intron structure (Chanfreau and Jacquier 1996; Costa et al. 1997a; Fedorova et al. 2003; Fedorova and Pyle 2005). Deletion of D2 has been found to have little effect on splicing (Koch et al. 1992); however, the domain does contain structural elements that are important for the formation of two tetraloop-tetraloop receptor interactions (θ-θ′ and η-η′) and contributes to the structural stability of the ribozyme (Chanfreau and Jacquier 1996; Costa et al. 1997a). Although deletion of Domain 3 does not abolish splicing activity, constructs containing the domain have a greatly enhanced rate of catalysis (Griffin Jr et al. 1995; Xiang et al. 1998; Su et al. 2001; Fedorova et al. 2003). A single tertiary contact (μ-μ′) involving the domain has been published (Fedorova and Pyle 2005).

Domain 4 typically encodes the approximately 1.5 kb IEP and is non-essential for RNA catalysis *in vitro*. Deletion of D4 facilitates splicing of the intron *in vitro* as it eliminates a large section of RNA that can interfere in intron folding. The domain is required for maturase-assisted splicing *in vivo* and mobility and regions of the domain have been implicated in the high affinity binding of the IEP (Wank et al. 1999).

1.2.4.3 Domain 5 - The Active Site

Domain 5 is the most highly conserved domain of the intron and together with nucleotides from the single stranded linker region between D2 and D3 (J2/3) forms the catalytic active site. D5 consists of upper and lower helices separated by a 2 nt asymmetric bulge and capped by a GNRA tetraloop. The bulge consists of a phylogenetically conserved AC (Costa et al. 1998). The lower helix contains a conserved

AGC or CGC motif that has been implicated in catalysis and is referred to as the "catalytic triad" (Chanfreau and Jacquier 1994; Michel and Ferat 1995). Both the AC bulge and the catalytic triad have been implicated in the binding of divalent metal ions (Sigel et al. 2000; Gordon and Piccirilli 2001). Two divalent metal ions have been found to bind in this region at 4.1 Å apart and as such the metal ions are appropriately positioned for the two metal ion mechanism of catalysis (Steitz and Steitz 1993). In addition to these interactions which all occur on the "catalytic face" of D5, the previously mentioned interactions (λ-λ′, κ-κ′, and ζ-ζ′) that dock D5 to D1 occur on the opposite side of the D5 helix, known as the "binding face".

1.2.4.4 Domain 6

The last of the intron domains is domain 6. D6 is poorly conserved between subgroups of introns with the exception of a bulged adenosine residue near the base of the helix that initiates the branching reaction (Chu et al. 2001; Michel et al. 2009). The stem loop is capped by the tetraloop sequence that is involved in the η-η′ interaction thought to be responsible for the conformational change between the first and second steps of splicing (Chanfreau and Jacquier 1996). An internal loop within D6 has been proposed to be part of a novel interaction, termed ι-ι′, which is thought to be responsible for branching in group II introns (Li et al. 2011a).

1.2.4.5 Tertiary Structure

In 2008, the 3.1 Å x-ray crystal structure was solved for a IIC intron from *Oceanobacillus iheyenis* (Toor et al. 2008). The RNA construct used was an abbreviated construct of 412 nt and contained a deletion of the IEP sequence in D4. In addition, the distal stems of D2, D3 and D6 were deleted. D6 was unresolved in the structure and was

later found to be degraded The fact that D6 is not static was employed to explain why D6 was not visualized in the crystal structure (Toor et al. 2008; Pyle 2010). The structure used for the initial crystallization of the intron was a post-catalysis structure however, the pre-catalysis and many other intermediate structures of the *O. iheyenis* intron have since been reported (Chan et al. 2012; Marcia and Pyle 2012).

At the same time as the initial crystal structure, a complete three-dimensional model was created for the well-studied IIA intron from *Lactococcus lactis*, Ll.LtrB (Dai et al. 2008). This model was based on extensive crosslinking data that compared the ribozyme structures both pre- and post-catalysis. Although partial models of the active site had previously been published (De Lencastre et al. 2005; Hamill and Pyle 2006), the Ll.LtrB model was the first model published that included the entire intron structure. Both the model and the crystal structure have contributed substantially to our knowledge of the structure and catalysis of group II introns and have confirmed much of the previous biochemical data.

The crystal structure of the IIC intron from *O. iheyenis* and the model of Ll.LtrB revealed that coaxial stacking is highly prevalent and dictates the overall structure. Within D1, helices I(i) and I(ii) are coaxially stacked, as are stems IA and IB. Additionally, the D3 and D4 helices are stacked coaxially (Toor et al. 2008). A secondary structure illustrating the coaxial stacking and major base pairing interactions is depicted in Figure 6. In addition to coaxial stacking, a variety of long range tertiary interactions help stabilize the ribozyme structure including tetraloop-tetraloop receptor interactions, kissing loops, and ribose zipper interactions among other less standard RNA motifs.

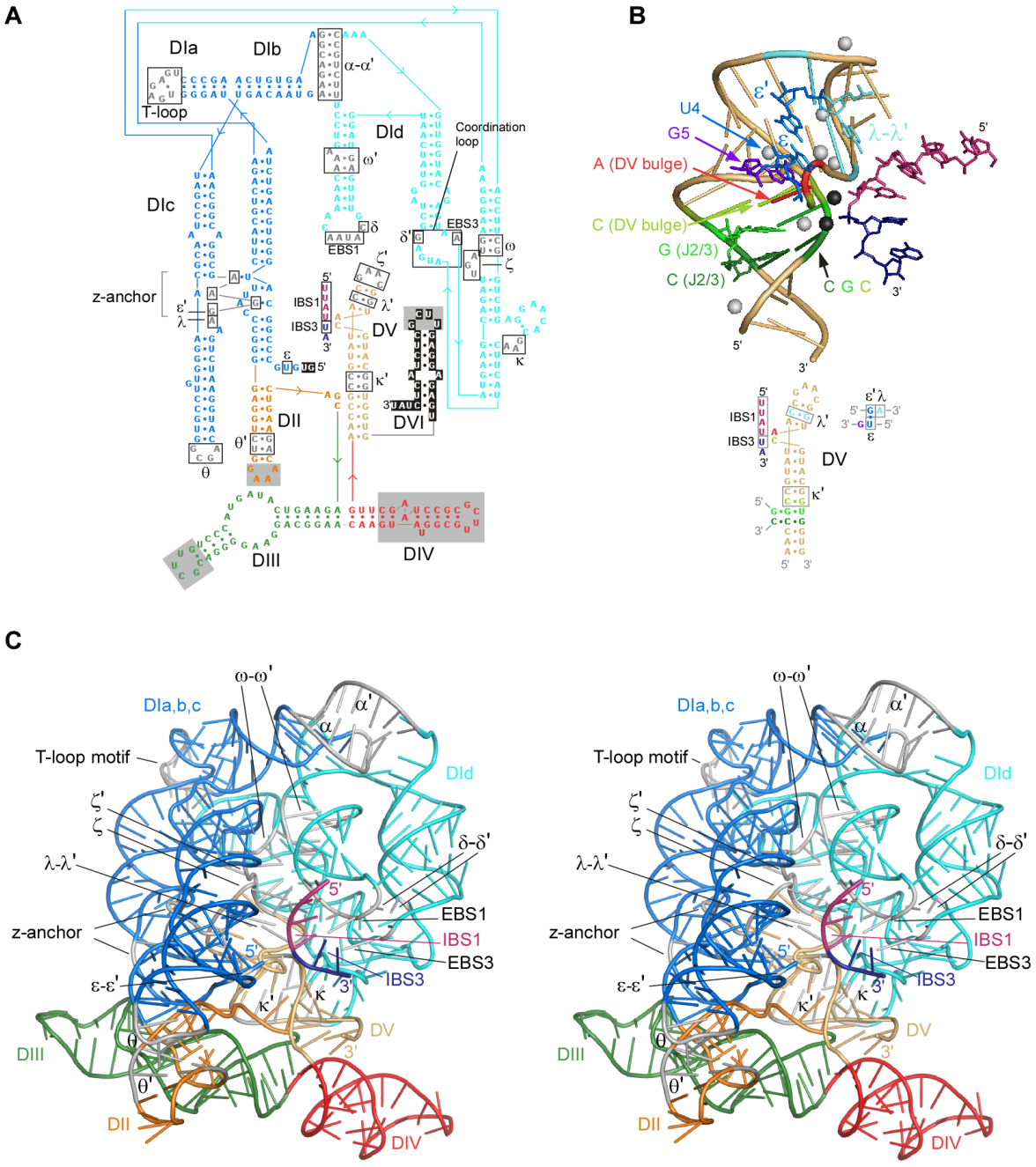
While the initial compaction of most large ribozyme structures has been reported to be fast (Fang et al. 2002; Russell et al. 2002), the initial compaction of group II ribozymes and the formation of the D1 scaffold is slow and represents the rate limiting step in the formation of the active ribozyme structure (Su et al. 2005). The scaffold structure of D1 itself is dictated by the formation of a 5-way junction formed between the IA, IB, IC, ID and I(i/ii) helices and as well as by the interactions within the κ-ζ region (Pyle et al. 2007; Toor et al. 2008; Waldsich and Pyle 2008). The subsequent docking of the other intron domains to the scaffold occurs rapidly (Su et al. 2005). NAIM studies and non-denaturing gel electrophoresis revealed that the crucial elements for compaction are found within the κ-ζ region which has been deemed the "folding control element" (Waldsich and Pyle 2006). Interestingly, this region serves as the active site receptor in the later stages of intron folding (Waldsich and Pyle 2006; Toor et al. 2008). Similar to other RNAs, Mg^{2+} binding is a requirement for molecular collapse (Russell et al. 2002; Das et al. 2003; Su et al. 2005; Woodson 2005).

During folding, a substructure forming a scaffold of the active site is created. This structure, termed the "z-anchor", was initially observed in the crystal structure (Toor et al. 2008). Kinking of the phosphate backbone causes the bases of the Ic stem (noted in Figure 6A) to alternate from side to side forming base pairs, triples and stacking interactions with various regions of the ribozyme. This structure brings together the 5′ splice site, D5 and J2/3. The previously described interactions ε-ε′ and λ-λ′ are now known to be part of this larger "z-anchor" substructure.

Figure 6. Crystal structure of the *Oceanobacillus iheyenis* group IIC intron.

(A) Sequence and structure of crystallized RNA. Black outlined boxes highlight tertiary interactions. Grey shaded boxes show regions deleted in the crystallized construct. Color code of the domains is carried throughout the subsequent parts of the figure. (B) Structure of the active site region. Corresponding color coded secondary structure region is shown below. DV is a beige tube helix, with a bound RNA modeled as the 5′ and 3′ exons [pink and indigo, respectively; (Toor et al. 2008; Toor et al. 2010)]. The interactions in the catalytic triple helix between the CGC triad, the CG of J2/3, and the C of the AC bulge are shown in dark green, green and yellow-green. The three-base stack consisting of the A of the AC bulge, G5, and U4 of the ε-ε′ interaction are in red, purple and blue, respectively, while the λ-λ′ interaction is cyan. Metal ions bound to D5 in the crystal are indicated by spheres, with black spheres representing the proposed active-site Mg^{2+} ions.

(C) X-ray crystal structure. Stereoviews are shown with the domains colored as in part (A) and regions involved in tertiary interactions shown in grey (Toor et al. 2008). Figure taken from (Lambowitz and Zimmerly 2011).


A
DIa
DIb
α-α'
T-loop
DIc
DId
Coordination loop
ω'
δ
δ'
EBS3
EBS1
ω
ζ
ζ'
z-anchor
ε'
λ
IBS1
IBS3
λ'
DV
ε
κ'
DII
DVI
θ'
θ
DIII
DIV
B
U4
G5
A (DV bulge)
C (DV bulge)
G (J2/3)
C (J2/3)
ε'
λ-λ'
C G C
5'
3'
IBS1
IBS3
DV
ε'λ
ε
κ'
C
ω-ω'
DIa,b,c
α'
α
T-loop motif
DId
δ-δ'
EBS1
IBS1
EBS3
IBS3
ε-ε'
κ'
κ
DV
DIII
θ
θ'
DII
DIV

As D5 has long been known to be a key feature in the catalytic core, a crystal structure of D5 alone was solved before the complete group II intron crystal structure was obtained (Sigel et al. 2004; Toor et al. 2008); however, the *O. iheyenis* x-ray crystal structure differs significantly from the previously published free structures of D5. Instead of the upper and lower helices being coaxially stacked as they had appeared to be in the free crystal structure and NMR structures, the domain twists through the asymmetric AC bulge to bend back upon itself, bringing together the AC bulge and the catalytic triad (Toor et al. 2008). This concentrates the phosphate backbone and creates a region of extreme negative electrostatic potential. The crystal structure revealed that 2 bases just upstream of γ in J2/3 form major groove base triples with the CG of the catalytic triad. The C of the AC bulge stacks upon the base triples and forms a triple with the final C of the triad. As such, the catalytic triad is essentially a catalytic triple helix in its active form.

1.2.5 IEP Structure

As previously stated, the prototypical IEP consists of 4 functional domains: RT, X, D and En (Figure 2B). The RT domain contains sub-domains 0-7 (Xiong and Eickbush 1990; Zimmerly et al. 2001). Domains 1-7 are conserved and alignable across all known RTs and act as the palm and fingers of the polymerase protein (Xiong and Eickbush 1990). RT domain 5 contains the highly conserved YADD sequence that is part of the active site of the protein. The maturase (X) domain component is the thumb domain of the polymerase. It associates with intron RNA and promotes folding of the intron into a catalytically active structure (Matsuura et al. 1997). This maturase domain was identified due to mutations in the region that affect the splicing ability of introns

(Mohr et al. 1993; Moran et al. 1994). Both the RT domain and the X domain are necessary for splicing *in vivo* (Cui et al. 2004).

The sequence of the D region is not conserved across group II introns but is thought to be a zinc-finger-like domain and has been functionally identified as a DNA binding region for the L.l.LtrB intron (San Filippo and Lambowitz 2002). Bacterial classes C, D, E and F contain IEPs that lack the En domain and, as such, the phylogeny of group II introns is consistent with the recent acquisition of the En domain in Bacterial B, ML and CL introns (Simon et al. 2009). The endonuclease encoded by group II introns is a Mg^{2+} dependent DNA endonuclease of the HNH family of endonucleases that cleaves the target DNA strand and provides a primer for reverse transcription (San Filippo and Lambowitz 2002).

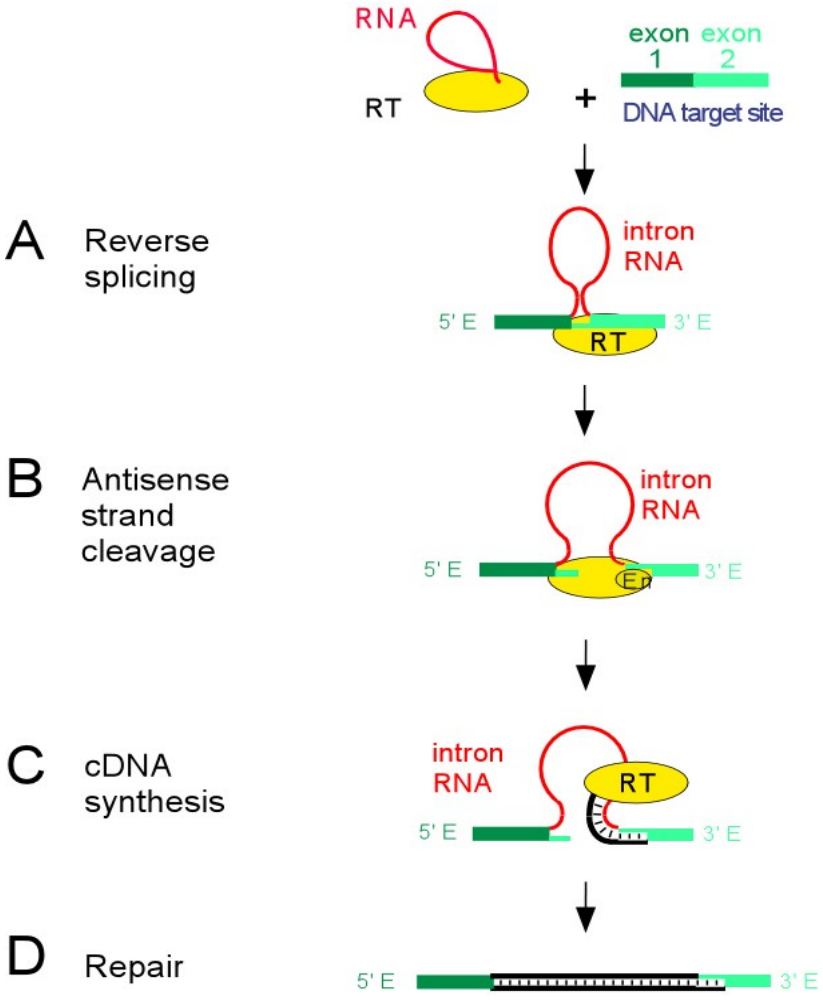
A small group of fungal mitochondrial group II introns encodes non-RT IEPs that belong to a family of DNA endonucleases characterized by LAGLIDADG motif (Toor and Zimmerly 2002). These ORFs have also been noted in a handful of the IIB1 introns containing 5′ terminal insertions (Li et al. 2011b). These LAGLIDADG endonucleases are typical of group I introns where they function to promote double strand break induced homologous recombination (Lambowitz and Belfort 1993; Chevalier and Stoddard 2001). While maturase function is associated with some LAGLIDADG ORFs in group I introns (Ho et al. 1997; Bassi et al. 2002), this has not been reported for group II introns (Mullineux et al. 2010; Li et al. 2011b). As such these LAGLIDADG ORFs do not contribute to splicing or to RNA-based mobility typical of group II introns.

1.2.6 Group II Intron Mobility

Both steps of the branching pathway splicing reaction are reversible (Augustin et al. 1990; Mörl and Schmelzer 1990). As such, group II introns are capable of reverse splicing into intron-less copies of both RNA and DNA. *In vivo*, once the intron lariat is excised it remains associated with its IEP dimer, forming an active RNP complex (Zimmerly et al. 1995; Lambowitz and Zimmerly 2004). This complex can then undergo mobility. Group II introns are capable of both retrohoming (inserting into a site-specific position in the genome) and retrotransposition (inserting into a novel or ectopic site within the genome), although retrotransposition occurs at much lower frequencies (Cousineau et al. 2000; Martínez-Abarca and Toro 2000; Ichiyanagi et al. 2002). The group II intron RNP reverse splicing into the DNA using a process known as target primed reverse transcription (TPRT) (Figure 7). The intron recognizes the homing site via base-pairing interactions with IBS1 and IBS2. Reverse splicing results in the insertion of the intron into the top strand of the DNA helix (Guo et al. 1997). The endonuclease domain of the IEP then cleaves the bottom strand of the DNA 9 or 10 nt downstream. The RT domain utilizes the nick as a primer to synthesize the complementary DNA. Host repair systems are then recruited to complete the insertion process (Zimmerly et al. 1995). In the case of IIC introns, insertion targets intrinsic transcriptional terminators (Granlund et al. 2001; Dai and Zimmerly 2002; Robart et al. 2007), presumably as a mechanism evolved to minimize damage to the host while still allowing the spread of introns. A subset of group II introns lack an En domain and, as such, rely on primers provided during the passage of the replication fork to complete insertion (Ichiyanagi et al. 2002; Zhong and Lambowitz 2003).

Figure 7. Mechanism of group II intron mobility.

Once the RNA is transcribed the RT is translated and then the RNA splices creating the intron lariat which associates with the RT to form an RNP complex. This complex recognizes a target site within the DNA sequence and reverse splices into the target strand (A). The En domain cleaves the antisense strand (B) and then the RT domain reverse transcribes the intron into DNA (C).



Although many group II introns are observed to splice through hydrolysis resulting in the production of a linear intron, the linear introns are generally believed to be inefficient or unable to carry out reverse splicing (Dème et al. 1999; Mohr et al. 2006; Gordon et al. 2007). The well-studied mitochondrial intron from *Saccharomyces cerevisiae*, ai5γ, has been shown to be capable of reversal of the second step of splicing resulting in the ligation of the 3′ end of the intron to the 5′ end of the downstream exon (Roitzsch and Pyle 2009). It has also been shown that *in vivo* mobility of linear intron products can occur in eukaryotes by relying on the host systems, such as non-homologous end joining (NHEJ) to ligate the 5′ end of the intron, but that the resultant retrohoming frequencies are lower than those observed for lariat introns (Zhuang et al. 2009). As such, it is still believed that the intron lariat is the preferred substrate for mobility and that lariat intron is required for efficient mobility *in vivo*.

1.2.7 Group II Intron Evolution

The group II intron retroelement ancestor hypothesis states that the ancestral group II intron was likely a mobile group II intron, consisting of both catalytic RNA and an associated compact IEP (Toor et al. 2001). This early bacterial group II intron was possibly similar to IIC introns, having a compact RNA structure and encoding an IEP lacking an endonuclease domain (Simon et al. 2008). Group II introns subsequently were subjected to many unique evolutionary trajectories in which RNA structures and IEPs co-evolved to create the unique lineages of group II introns that are known today.

The presence of a few introns and at least nearly complete spliceosomes in all eukaryotes with sequenced genomes (Collins and Penny 2005; Martin and Koonin 2006; Roy and Gilbert 2006), suggest that introns were present within the last eukaryotic

common ancestor (LECA). Similarities in splicing mechanism and other structural similarities have led to the hypothesis that group II introns are the evolutionary ancestors of snRNAs and the spliceosome, as well as spliceosomal introns (Lambowitz and Zimmerly 2004; Martin and Koonin 2006). It has been proposed that group II introns were present in the α-proteobacteria endosymbiont ancestor of the mitochondrion and this hypothesis is supported by the presence of group II introns in many modern organelles and by the abundance of group II introns found within α-proteobacterial genomes. Occasional lysis of the endosymbiont would have released DNA and RNA into the cytosol and allowed for the spread of introns into the host DNA through recombination or intron-mediated homing reactions. The spread of introns would have created an intron-rich genome and the subsequent decay of both the IEP and catalytic RNA structure would have subsequently led to the evolution of alternate splicing machinery (Sharp 1991; Martin and Koonin 2006).

In addition to a plausible evolutionary scenario in which the spread of group II introns resulted in the formation of spliceosomal introns and the spliceosome, there are tangible lines of evidence that link the fates of these introns. Group II introns and spliceosomal introns utilize the same chemical mechanism of splicing; however, the mechanism is not so complex that it could not be envisioned to have arisen independently on multiple occasions. The discovery of bipartite and tripartite *trans*-splicing in organelles led Phil Sharp to speculate that "during evolution of the progenitor cell for eukaryotes, group II introns could have been fragmented into small *trans*-acting RNAs. Division of a group II intron into "five easy pieces" could have generated the precursors for the five snRNAs that form the spliceosome in the nuclei of eukaryotic cells (1991)."

As a caveat to his famous hypothesis he noted that if the above statement held true that identity in sequence and structure between group II introns and snRNAs should be expected but had yet to be identified.

Today, several structural similarities have been noted. These include a base pairing interaction between an ACA sequence in the U6 snRNA and nucleotide positions 4-6 of the intron thought to be analogous to the group II intron ε-ε′ interaction, an interaction between the U5 snRNA and the 5′ splice site thought to be analogous to the EBS1-IBS1 interaction of group II introns (Sontheimer and Steitz 1993), and a pairing between the U2 snRNA and the UACUACC box that creates a bulged adenosine analogous to D6 (Valadkhan and Manley 2001). Most striking, D5 is also structurally analogous to the U6 and $U6_{atac}$ intramolecular stem loop (ISL) and can substitute for the $U6_{atac}$ ISL in the spliceosome and allow for efficient catalysis (Shukla and Padgett 2002).

Additionally, Prp8, a major protein in the catalytic core of the spliceosome shows similarity to group II intron IEPs suggesting a possible evolutionary relationship (Dlakić and Mushegian 2011). Recently, the relationship between group II and spliceosomal introns was strongly supported by the X-ray crystal structure of the spliceosomal protein Prp8. The structure depicts a reverse transcriptase-derived domain located in the heart of the spliceosome, within direct cross-linking distance to the intron and exon substrates and to the snRNAs (Galej et al. 2013). The structure is consistent with the scenario that the spliceosome was derived from a mobile group II intron, and contains remnants of both a self-splicing ribozyme and a reverse transcriptase IEP.

An evolutionary relationship has also been suggested for group II introns and non-LTR retroelements as both insert through the same mechanism of target primed reverse transcription (TPRT) and share similarities in their RTs (Malik et al. 1999).

1.3 Alternative Splicing of Eukaryotic Introns

Alternative splicing is the combinatorial processing of exon and intron sequences to create multiple unique mRNA transcripts from a single gene. It is one of the hallmarks of spliceosomal introns in higher eukaryotes. It was first described over 30 years ago when it was noted that both membrane-bound and secreted antibodies were produced from the same gene (Alt et al. 1980; Early et al. 1980). Current estimates from high throughput sequencing studies suggest that more than 90% of the genes in the human genome are alternatively spliced (Pan et al. 2008; Wang et al. 2008). As such, alternative splicing is the main mechanism responsible for proteomic diversity in higher eukaryotes and it has even been suggested that alternative splicing-mediated transcriptome diversification can lead to speciation (Ast 2004). In perhaps the most spectacular example of alternative splicing, the *Drosophila melanogaster* gene *Down syndrome cell adhesion molecule* (*dscam1*) can generate 38,016 distinct mRNA isoforms (Schmucker et al. 2000). Although alternative splicing is generally assumed to occur in *cis*, there are eukaryotic examples of alternative splicing that occur in *trans* (Nilsen and Graveley 2010). For example, the *modifier of mdg4* gene *(mod(mdg4))* in *D. melanogaster* produces 28 mRNAs through a reaction in which a common 5′ exon is spliced to one of 28 variable 3′ exons. At least seven of the reactions occur in *trans* (Dorn et al. 2001).

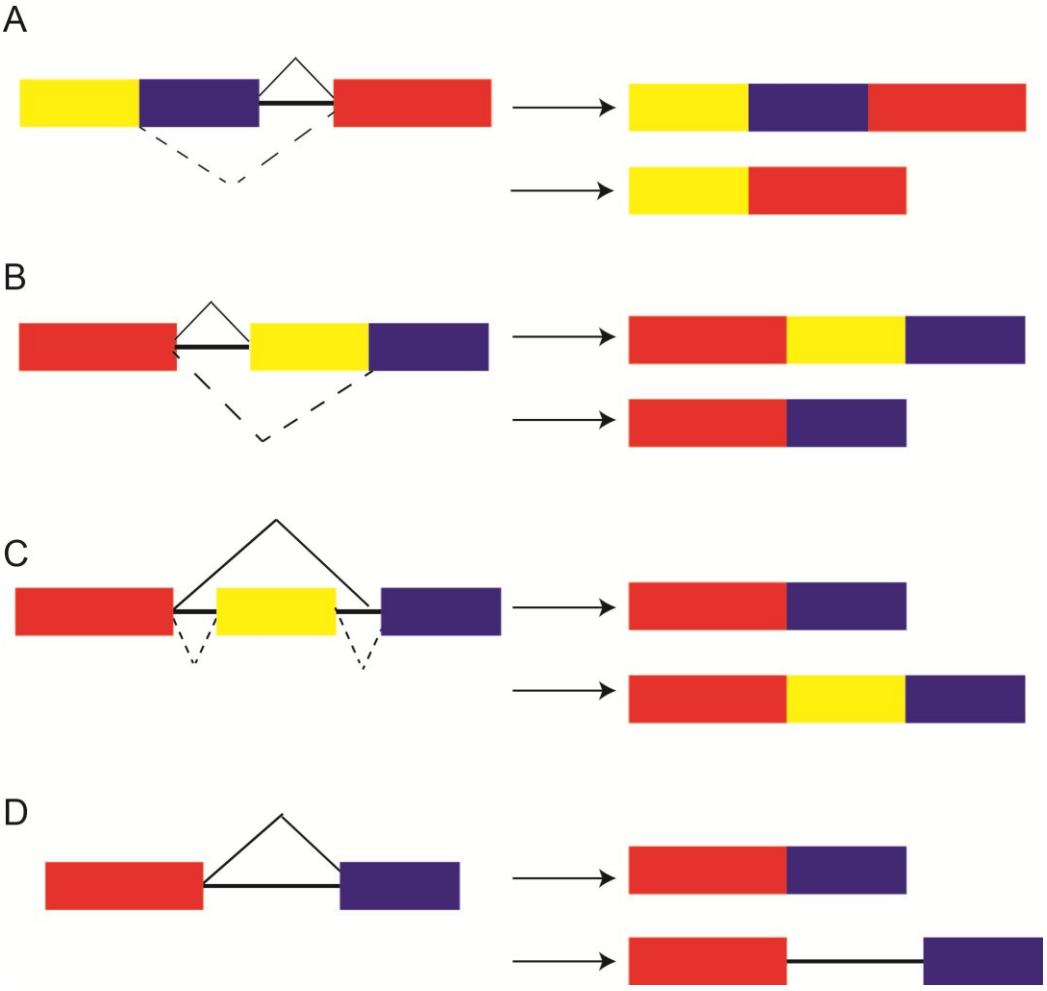
As only regulated splicing events are well-studied, it remains unclear how much of alternative splicing is constitutive or rather to what extent the multiple mRNA

isoforms are produced at the same ratios in all cell types or under all conditions. Well-studied alternative splicing events are often tissue specific, dictated by developmental or differentiation-specific signals, or in response to external stimuli (Xie and Black 2001; Shin and Manley 2004; Boutz et al. 2007; Makeyev et al. 2007; Sánchez 2008).

Nearly all cases of alternative splicing result from one or more of four basic schemes: alternative 5′ splice site choice; alternative 3′ splice site choice; exon cassette inclusion or skipping; and intron retention [Figure 8; for reviews see (Breitbart et al. 1987; Lopez 1998; Graveley 2001; Keren et al. 2010; Nilsen and Graveley 2010; De Bock et al. 2011)]. Although both sequence-based *cis*-elements and *trans*-factor based mechanisms have been implicated in the complex process of alternative splicing, one of the main forms of regulation identified is a form of *trans*-factor based regulation from SR proteins (so named because they are rich in serine and arginine) and hnRNPs (heterogeneous nuclear RNPs) (Lin and Fu 2007; Martinez-Contreras et al. 2007). These proteins are ubiquitously expressed but their relative abundance can fluctuate in different cell types or under different environmental conditions. SR proteins are RNA binding proteins that are generally thought to activate splicing by binding to exons and recruiting the spliceosome; hnRNPs on the other hand are generally thought to repress splicing by interfering with the ability of the core splicing machinery to engage at a particular splice site. Despite this general rule there are SR proteins that are known to suppress splicing and hnRNPs that are known to activate splicing and the function of these proteins are far from completely understood.

Figure 8. The four most common patterns of alternative splicing.

(A) Alternative 5′ splice site selection (B) Alternative 3′ splice site selection (C) Exon cassette inclusion or skipping and (D) Intron retention. Possibilities outlined by the top solid black lines are shown as the top option on the right. The possibilities outlined by the dashed lines are shown as the bottom option of the pair on the right. Intron is shown as a thin grey line, while exons are depicted as coloured boxes. In the case of (D) the retained intron would be translated.



Many additional factors can have large effects on splicing patterns including: transcription rate (Kornblihtt 2007), core splicing machinery levels (Park et al. 2004; Pleiss et al. 2007), intron size (Fox-Walsh et al. 2005), RNA structure, competition between splice sites (Yu et al. 2008) and chromatin structure. Both non-coding RNAs (ncRNAs) and siRNAs have been shown to contribute to splice site selection through regulation of the levels of splicing factors at various developmental stages, through influencing heterochromatin formation and through sequence specific interaction with the target pre-mRNA.

1.4 The Host Organism, *Clostridium tetani*

Clostridium tetani is a Gram positive, anaerobic, spore-forming bacterium that is commonly found in soils and within the intestines of various animals. The organism is regarded as a true representative of the genus *Clostridium* (ie. is a member of *Clostridium sensu stricto*) (Gupta and Gao 2009). *C. tetani* is the causative agent of the disease tetanus and infections are generally established through anaerobic puncture wounds that allows entry of spores. Spores then must germinate and the vegetative cells must interact with host tissues. Tetanus is subsequently caused by the release of a neurotoxic substance, TeTx (Tetanus toxin), that is transported to the spinal cord where it blocks exocytosis of inhibitory neurotransmitters (Schiavo et al. 1992). This leads to continuous muscle contractions and eventual death. During the Second World War a potent toxoid-based vaccine was developed and subsequent immunization has virtually eliminated the disease in industrialized nations; however, *C. tetani* is still responsible for numerous neo-natal deaths in the third world. Despite a wealth of knowledge about the mechanism of action of the tetanus toxin itself (Schiavo et al. 1992; Montecucco and Schiavo 1995;

Pellizzari et al. 1999; Brunger and Rummel 2009), little is known about the cellular mechanisms *C. tetani* uses to establish infection.

In 2003, the genome sequence of *Clostridium tetani E88* was published (Brüggemann et al. 2003). It was found to contain a single circular chromosome of 2,799, 250 bp in length with a low GC content of 28.6%. In addition to the chromosome, a 74 kb plasmid, pE88 is also found within the bacteria. The plasmid contains an even lower GC content of 24.5%. The plasmid encodes the TeTx gene and its transcriptional activator, TetR. As such, it is the plasmid which is responsible for toxin production and the tetanus disease associated with the organism.

The chromosome of *C. tetani* encodes 32 gene products that show homologies to known cell surface proteins. These include nineteen homologues (including CTC00462) of Cwp66, an adhesin in *C. difficile* (Waligora et al. 2001), two proteins with multiple leucine-rich repeats domains with identity to *Listeria monocytogenes* internalins that mediate binding to host epithelial cells (Mengaud et al. 1996), and eleven additional SLP-related ORFs that have no homologues in other sequenced clostridial genomes but contain characterized domains of SLPs (identified in this group is CTC00465) (Brüggemann et al. 2003). While the highest level of gene order conservation among sequenced clostridial genomes is between *C. tetani* and *C. botulinum*, it is relatively low compared with enterobacterial genomes (Brüggemann and Gottschalk 2004).

1.5 Bacterial Surface Layers

Surface layers (or S-layers) are crystalline-like, two-dimensional arrays of proteinaceous subunits found on the outermost surface of most eubacteria and virtually all archaeabacteria (Sleytr et al. 1996; Sleytr and Beveridge 1999). In Gram-positive

bacteria, the S-layer is non-covalently bound on the outside of peptidogylcan, pseudomurein and other secondary cell wall polymers; in Gram negative organisms the S-layer is attached via the lipidpolysaccharide component of the outer membrane (Sleytr and Beveridge 1999; Claus et al. 2005; Pavkov-Keller et al. 2011). S-layer proteins (SLPs) typically possess regions of three tandem repeats of approximately 55 amino acids that are known as surface layer homology (SLH) domains. These proteins are mainly weakly acidic and typically contain 40-60% hydrophobic amino acids and few or no sulfur containing amino acids such as cysteine or methionine (Sleytr and Beveridge 1999). In Gram positive organisms, the interactions between the S-layer and the cell is typically mediated either by direct binding of the SLH domains to the peptidoglycan (Zhao et al. 2006) or indirectly through non-covalent interactions between the SLH domains and pyruvylated carbohydrates that are covalently linked to the peptidoglycan (Chauvaux et al. 1999; Ilk et al. 1999; Mesnage et al. 2000). For SLPs that lack the SLH domain, association to secondary cell wall polymers occurs either through the C- or N-terminal protein domains (Egelseer et al. 1998; Sára et al. 1998; Smit et al. 2001). In addition to the domains responsible for attachment to the cell, SLPs possess domains for self-assembly (crystallization domains) that cause SLPs to assemble into their characteristic regular lattices (Pavkov-Keller et al. 2011). Most S-layer proteins (SLPs) are secreted via the general secretory pathway, however secretation via alternate pathways have been described in some organisms (Sleytr and Beveridge 1999). Post-translational modifications are known to occur in many SLPs including cleavage of the amino- or carboxy-terminal fragments, phosphorylation and glycosylation (Bahl et al. 1997).

Surface layer proteins are typically the most abundant proteins produced by the cell, and hence there is a high cost for the bacterium to maintain an S-layer. Because of this it has been assumed that S-layers have a common underlying function, but to date no single function has been assigned. Upon repeated culturing in the laboratory S-layers can be lost (Baldermann et al. 1998; Engelhardt 2007), suggesting that if there is a evolutionarily common function for S-layers that the benefits must be greatest in a natural environment.

Individual S-layers from diverse bacterial species have been implicated in a variety of cellular roles. It is thought that one of these roles may be acting as a molecular sieve or barrier to the external environment. Typical S-layer arrangements create pores of approximately 2-8 nm between subunits (Pavkov-Keller et al. 2011). This small pore size might prevent molecules such as lytic enzymes, complement, antibodies, and biocides from entering the cell. In the case of Gram positive organisms, large molecules leaving the cell may be trapped beneath the S-layer, creating a space functionally equivalent to the periplasm in Gram negative bacteria. In addition, the S-layer can serve as scaffolding for extracellular enzymes and provide overall stability to the shape of the cell (Engelhardt 2007). S-layers can also provide resistance to bacteriophage (Koval and Hynes 1991). In certain Gram positive pathogens, the S-layer has been established as a virulence factor. SLPs can also help establish infection by providing resistance to the host immune response through antigenic variability, as is the case for the surface layer from *C. difficile* (Takeoka et al. 1991) or may influence the ability of the organism to adhere to the basal membrane as is the case in *Bacillus cereus* (Kotiranta et al. 1998).

1.6 Project Hypothesis

During a bioinformatic search for ORF-less group II introns (Simon et al. 2008), a unique genomic arrangement containing a group II intron was identified in a region of surface layer proteins (SLPs) in the *Clostridium tetani* genome. The region consisted of a previously reported SLP (CTC00465) followed by a full-length group II intron-like sequence. This sequence was immediately followed by a 3′ exon ORF sequence. Three intron fragments consisting of D5/6 were also present downstream and flanked immediately by alternate 3′ exon sequences. As such, it was hypothesized that the intron could utilize the downstream fragments of D5/6 to provide a series of 3′ splice sites allowing for alternative splicing of the intron. As the intron was located in an SLP region, alternative splicing would likely be beneficial to the organism and may increase the overall fitness. This thesis explores the alternative splicing hypothesis focusing on the characterization of the splicing reaction *in vivo* and on SLP expression in *C. tetani*. Additionally, it looks at the characterization of the intron ribozyme structure and how it is adapted to its genomic environment.

Chapter Two: **CHARACTERIZATION OF ALTERNATIVE SPLICING *IN VIVO***

2.1 Introduction

Although discovered in the chloroplasts and mitochondria of some eukaryotes in the mid 1980's (Peebles et al. 1986; Schmelzer and Schweyen 1986; Van der Veen et al. 1986), group II introns were not identified in bacteria until 1993 when they were identified through PCR screening (Ferat and Michel 1993). Since their initial discovery the number of known group II introns in bacterial genomes has increased substantially, mainly due to the increase in genome sequencing projects. Now over 600 unique full-length bacterial group II introns have been identified and it is known that approximately one quarter of all bacterial genomes encode group II introns [(Candales et al. 2012) and S. Zimmerly unpublished data].

As group II introns share little conservation at the level of primary sequence, bacterial group II introns are typically identified by the presence of the RT-based ORF sequence in D4. Subsequently the 5′ and 3′ intron ends are located and the intron RNA structure is folded, using MFOLD (Zuker 2003) and manual constraints guided by consensus structures, to verify the presence of the intron. While this approach is generally successful in identifying most bacterial group II introns it specifically does not identify ORF-less introns or introns that encoded non-RT ORFs such as the LAGLIDADG homing endonucleases (Toor and Zimmerly 2002; Simon et al. 2008). Because bacterial group II introns are often located in intergenic regions (Dai and Zimmerly 2002; Klein and Dunny 2002), rather than in housekeeping genes, atypical (ORF-less or LAGLIDADG) or previously unknown forms of these introns may be overlooked.

An alternate method of identifying group II introns is to search for conserved RNA structural motifs (ie. D5) using programs such as RNAmotif (Macke et al. 2001), independently of the ORF sequence. Using this approach (Simon et al. 2008), Dr. Dawn Simon — a former postdoc in the Zimmerly lab, identified a unique group II intron located within a surface layer protein region of the human pathogen *Clostridium tetani* E88. Four D5 sequences were encoded within an approximately 5 kb region. Each D5 sequence was followed immediately by a D6 sequence. The first D5/6 sequence was part of a full-length group II intron-like RNA structure. The intron, named according to our lab's standard intron naming conventions — *C.te*.I1, does not encode its own IEP and no other group II introns or group II intron RT-related proteins are encoded elsewhere in the genome.

The full copy of the intron, *C.te*.I1, is flanked by ORF sequences on the 5′ and 3′ sides and each of the downstream D5/6 sequences are also flanked on the 3′ side by ORF sequences. This unique arrangement led to the hypothesis that *C.te*.I1 may be capable of alternative spicing by associating with the downstream D5/6-3′ exon sequences and utilizing alternate 3′ splice sites.

In this chapter, I describe the initial characterization of the intron and its *in vivo* splicing reaction. As well, I investigate potential regulation of the alternative splicing reaction. This portion of the chapter has been published (McNeil et al. 2013). Finally, I attempt to elucidate physiological relevance of the alternative splicing reaction at both the RNA and protein levels in a variety of *C. tetani* strains.

2.2 Materials and Methods

2.2.1 Strains and Growth Conditions

Clostridium tetani strain ATCC10779 (Designation 43415 – Harvard Strain) was obtained from the American Type Culture Collection. CN655 (NCTC 2918), CTHCM19, CTHCM22, and CTHCM 25 were supplied by Dr. Neil Fairweather, Imperial College London. CN655 was obtained from the National Collection of Type Cultures, Colindale, UK. Strains CTHCM19, CTHCM22 and CTHCM 25 were isolated from patients with clinical tetanus in the Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam (Qazi et al., 2007). Two additional strains, designated NML98A045 and NML070850, were acquired from the Public Health Agency of Canada, National Microbiology Laboratory, and represent clinical isolates of *C. tetani* from patients in Canada isolated in 1998 and 2007, respectively. Cultures were grown in Brain Heart Infusion (BHI) medium (Oxoid CM1135) at 37°C under anaerobic conditions using the GasPak EZ anaerobic container system (BD Biosciences). Colonies were streaked on BHI agar plates (1-4% agar). Strains were also grown in an anaerobic hood on blood agar plates for comparison to the above system, but no differences in colony morphology or growth were observed and therefore studies were conducted with the EZ GasPak System (BD Biosciences).

For stress conditions, ATCC10779 was grown under anaerobic conditions to mid-logarithmic phase before the introduction of the stressor.

Oxidative Stress:

Cultures were grown up to mid-logarithmic phase (OD_{600} of 0.5 to 0.8) under standard conditions. Once the desired OD was reached a sample of cells were removed from the culture as a no stress control. Samples were moved to a 37 °C incubator and

shaken to introduce oxygen. Samples of the culture were then removed at various time points including 15, 45, 90 minutes, 3 hours and 5 hours following the introduction of oxygen through aeration.

Temperature Stress:

Cultures were grown under standard conditions except the incubator temperature was decreased to 25°C.

Nutrient Stress:

Cultures were grown in various dilutions and concentrations of BHI media. 2X, 4X, 0.75X and 0.5X concentrated BHI were used.

Osmotic Stress:

Cultures were grown up to mid-logarithmic phase (an OD_{600} of 0.5 to 0.6) under standard conditions. Once the desired OD was reached a sample of cells were removed from the culture as a no stress control. NaCl (1 g) was dissolved in a 25 mL culture to a concentration of 4% NaCl. Cultures were again made anaerobic using the EZ GasPak anaerobic container system (BD) and then allowed to grow for an additional 2 hours under stress before cells were harvested.

Metal Ion Stress:

Cultures were grown up to an OD_{600} of 0.7 under standard conditions. Once the desired OD was reached a sample of cells were removed from the culture as a no stress control. Zinc acetate was dissolved in a 25 mL culture to a concentration of 300 μM. Cultures were again made anaerobic using the EZ GasPak anaerobic container system (BD) and then allowed to grow for an additional 2 hours under stress before cells were harvested.

Alcohol Stresses:

Ethanol:

Cultures were grown up to mid-logarithmic phase (OD_{600} of 0.5 to 0.8) under standard conditions. Once the desired OD was reached a sample of cells were removed from the culture as a no stress control. EtOH was added to a 25 mL culture to a final concentration of either 4% or 10%. Cultures were again made anaerobic using the EZ GasPak anaerobic container system (BD) and then allowed to grow for an additional 2 hours under stress before cells were harvested.

Butanol:

Cultures were grown up to mid-logarithmic phase (OD_{600} of 0.5 to 0.8) under standard conditions. Once the desired OD was reached a sample of cells were removed from the culture as a no stress control. Butanol was added to a 25 mL culture to a final concentration of 0.25%, 0.5% or 1%. Cultures were again made anaerobic using the EZ GasPak anaerobic container system (BD) and then allowed to grow for an additional 2 hours under stress before cells were harvested.

Repeated culturing of single colonies:

Expression of RNA was tracked in single colonies over generations through repeated culturing. The initial culture was streaked on BHI plates containing 4% agar to isolate single colonies (referred to as the first generation). Single colonies were picked and streaked on new 4% agar BHI plates (2^{nd} generation). From these plates single colonies were once again picked and streaked on new 4% agar BHI plates (3^{rd} generation). Liquid cultures of each generation were made up and cells were harvested and RNA extracted once the cultures had reached mid-logarithmic phase (OD_{600} of 0.5 to

0.8). This was done in two separate trials. For the first trial, data were collected from the 1st, 2nd and 3rd generations while for the second trial, data were collected from the 1st generation and for two separate colonies from the 3rd generation.

2.2.2 RNA Extraction

RNA extraction was performed using the RNeasy Mini Kit (QIAGEN). Two volumes RNA Protect Bacterial Reagent (QIAGEN) was added to 2 mL of *C. tetani* culture (OD_{600}= ~0.6; unless otherwise indicated) and was allowed to permeate cells for 5 minutes. Cells were pelleted by centrifugation, supernatant was removed, and then the cells were re-suspended in 200 µL TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA) with 20 mg/mL lysozyme. Twenty µL Proteinase K (20 mg/mL) was also added. Cells were incubated in this lysis solution for 15 minutes at room temperature while vortexing occasionally. One mL hot (65°C) QIAzol lysis reagent (QIAGEN), a guanidine thiocyanate phenol solution, was added and was mixed by vortexing for 3 minutes; this was followed by a 5 minute incubation at room temperature. Two hundred µL of chloroform was then added, mixed vigorously by shaking, and the mixture was incubated at room temperature for 3 minutes. Centrifugation was performed at 4°C for 15 minutes and the aqueous phase was transferred to a new tube. Five hundred µL anhydrous ethanol was then added to the aqueous phase and mixed by pipetting. From this point Protocol 7 of the RNAprotect Bacteria Reagent Handbook (Second Edition, December 2005) was followed either with or without optional on-column DNase digestion, producing pure RNA or a DNA/RNA mixture. RNA was eluted in 40 µL RNase-free H_2O or TE (pH 8.0). Integrity of RNA preparations were visualized on a 1.2% agarose gel and purity and

concentration were assessed using a NanoDrop-1000 Spectrophotometer (Thermo Scientific).

2.2.3 DNA Extraction

Genomic DNA was prepared from a 15 ml broth culture. Pelleted cells were washed and resuspended in 0.45 ml of TE [10 mM Tris–HCl (pH 7.4), 1 mM EDTA]. Lysis was by the addition of 0.1 mg proteinase K and SDS to a final concentration of 1% with incubation at 37°C for 45 min. The sample was repeatedly extracted with an equal volume of phenol-CIA (25:24:1 of phenol:chloroform:isoamyl alcohol) until there was a clear interface upon centrifugation. This was followed by ethanol precipitation in the presence of 0.3 M NaOAc (pH 5.2). The pellet was washed with cold 75% ethanol, air-dried and dissolved in TE [10 mM Tris–HCl (pH 7.4) and 1 mM EDTA].

2.2.4 Region Amplification and PCR

A generalized PCR reaction was used to amplify DNA fragments for cloning and other experiments. Oligos utilized are listed in Table 1. Oligo sequences were synthesized either by Alpha DNA or by Integrated DNA Technologies. The reaction composition was 2.5 µl 10X *Pwo* buffer (100 mM Tris-HCl (pH 8.8), 25 mM $MgCl_2$, 500 mM KCl, 1% Triton X-100), 2.5 µl 2 mM dNTP (ACGT) solution, 1 µl Forward/Sense primer (10 pmol/µl), 1 µl Reverse/AntiSense primer (10 pmol/µl), 1 µl *Pwo* DNA polymerase (prepared by our laboratory; unknown concentration), 1 µl DNA template (1-250 ng) and 16 µl ddH_2O for a total volume of 25 µl. Products were amplified in a Veriti 96 well thermocyler (Applied Biosystems) by heating to 94°C for 2 minutes to denature the template DNA followed by 35 cycles of 94°C for 30 seconds, 55°C for 30 seconds and 72°C for 2 minutes (1 minute per 0.5 kb of target DNA). Although a typical annealing

temperature of 55°C is shown in the description above, annealing temperatures were variable based on Tm of primers and/or was experimentally determined through gradient PCR. Following completion of the PCR reaction, the desired DNA products were agarose gel extracted using the Zymoclean Gel DNA Recovery Kit (ZymoResearch, Irvine, CA). The PCR products were subsequently cloned into a pBluescript vector, pKS+ (Stratagene), and confirmed via sequencing (University of Calgary Core DNA Sequencing). Plasmids were transformed using DH5α chemically competent cells (Inoue et al. 1990).

2.2.5 16S rRNA Verification of Bacterial Species

Following gDNA extraction from each of the *C. tetani* strains, PCR reactions were performed as above using the primers 16S-F (5′TGG CTC AGA TTG AAC GCT GGC GGC) and 16S-R (5′TAC CTT GTT ACG ACT TCA CCA CA). PCR products were run on a 1% agarose gel and gel extracted using the Zymoclean Gel DNA Recovery Kit (ZymoResearch, Irvine CA). PCR products were then sent for sequencing (University of Calgary Core DNA Services). Blastn searches were performed with the sequencing results against the NCBI Nucleotide Collection (nr/nt) Database to identify the bacterial species.

2.2.6 RT-PCR and qPCR

cDNA synthesis was performed in 20 µl using 10 pmole gene-specific primer (either O1R, I1R2, O2R4, O3R3, O4R2, or O5R3) and 200 U Superscript II reverse transcriptase (Invitrogen) with 1 µg total RNA as template, according to the manufacturer's protocol. No-RT controls were performed in which no enzyme was added. Quantitative PCR reactions were done in a total volume of 12.5 µL containing 10

pmol of each forward (O1F or O1F4) and reverse primers (O1R, I1R2, O2R4, O3R3, O4R2, or O5R3), 6.25 µL 2X iQ SYBR Green Supermix (Bio-Rad) and 2 µL of the previous RT-reaction as template. A three-cycle amplification was performed (95°C for 10 seconds, 57°C for 20 seconds, 72°C for 20 seconds) with melt curve (ramping from 72°C to 95°C) using the Rotor-Gene Q, real-time PCR cycler (QIAGEN). No-RT and no-template controls were included in each run. All unknowns were run in technical triplicates. Standard curves were made from serial dilutions of plasmids with cloned exon junctions of identical sequence to allow for quantification of unknown samples. All primer pairs amplified with 95-107% efficiency. Total RNA samples obtained were assessed for quality, integrity and purity prior to qRT-PCR. For non-quantitative RT-PCR, PCR reactions were performed as above with *Pwo* DNA polymerase in 10 mM Tris-HCl (pH 8.8), 2.5 mM $MgCl_2$, 50 mM KCl, and 0.1% Triton X-100.

2.2.7 Extraction of Surface Layer Proteins

Extraction of surface layers was essentially performed as described (Qazi et al. 2007). Ten mL of *C. tetani* cells were pelleted by centrifugation and cells were re-suspended in 500 µL PBS. Cells were then pelleted again and re-suspended in 1 mL of the extraction reagent [either 4 M urea or 0.2 M glycine (pH 2.2)]. Cells in 0.2 M glycine were incubated for 20 minutes at room temperature while 4 M urea extractions were performed at 37°C for 30 minutes. After the extraction period, cells were pelleted by centrifugation for 10 minutes (17 900 g) and the supernatant containing the SLPs removed. SLPs were stored at 4°C or -80°C. Before storage, glycine extracts were neutralized with 2M Tris-HCl (pH 8.0).

2.2.8 SDS-PAGE

SDS-PAGE was performed according to the method of Laemmli (1970), using an 8% (w/v) acrylamide resolving gel. Protein bands were visualized by staining with Coomassie Brilliant Blue or by silver staining. Apparent molecular masses were determined by comparison to Novex high molecular weight pre-stained protein standards (Life Technologies).

2.2.9 MALDI-ToF MS

Protein bands were excised from gels and destained in 25 mM ammonium bicarbonate (NH_4HCO_3)/50% acetonitrile (ACN) at room temperature. Proteins were reduced by the addition of 50 µL of 10 mM dithiothreitol (DTT) in 100 mM NH_4HCO_3 for 1 hour at 56°C. Proteins were alkylated by the addition of 50 µL of 50 mM iodoacetamide in 100 mM NH_4HCO_3 for 30 min at room temperature in the dark, followed by washing twice with 200 µL of 100 mM NH_4HCO_3 for 15 min at room temperature. Samples were dehydrated using 100 µL of 100% ACN and vacuum dried. Proteins were digested overnight with 12.5 ng/µL trypsin (Promega) in 25 mM NH_4HCO_3 at 37°C. Peptides were extracted twice with 50 µL 50% ACN/1% formic acid. Supernatants were pooled and vacuum dried.

Samples were diluted in 50% ACN/0.05% Trifluoroacetic acid (TFA) immediately prior to analysis by MS using a matrix assisted laser desorption ionization-time of flight (MALDI-ToF) instrument (AB Sciex TOF/TOF 5800). One half of a µL of sample was spotted onto target plates together with 0.5 µL of matrix (2 mg/mL α-Cyano-4-hydroxycinnamic acid in 50% ACN/0.05% TFA). Once dry, the plate was loaded into the MALDI-ToF instrument. Mass lists were created using the TOF/TOF series explorer

software and peptides were searched against bacteria using Mascot/Matrix Science server. The following parameters were used: peptide mass tolerance was 100 ppm, carbamidomethylation was set as a fixed modification, the oxidation of methionine as a variable modification and a maximum of one missed cleavage was allowed. The significance threshold for search results was set at $p < 0.05$, which indicates identity or extensive homology.

2.2.10 LC MS/MS

Following separation on an 8% SDS-PAGE and Coomassie staining, the high molecular weight band from the CN655 strain was cut from the gel with a clean razor blade and the gel slice was sent for LC MS/MS analysis (The Southern Alberta Mass Spectrometry Centre for Proteomics, Calgary).

2.2.11 Glycan Staining

Glycan staining was performed using the Pro Q Emerald 300 glycan stain kit (Life Technologies) according to manufacturer's protocol.

2.2.12 Bioinformatic Predictions of Promoters and Terminators

Promoter sequences were predicted by submitting the region surrounding the *C.te*.I1 sequence to BPROM (Gautheret and Lambert 2001). Subsequently the results were evaluated and only promoter sequences found outside of the known ORF sequences were considered to be true hits. Terminator sequences were similarly evaluated by submitting sequence for the full *C.te*.I1 containing region to the ARNold server (Macke et al. 2001; Solovyev and Salamov 2011).

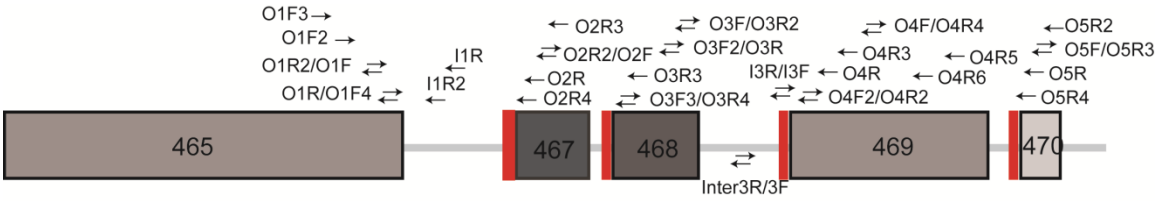
Table 1. Primer sequences corresponding to the *C. tetani* region of interest

Primer Name	5′→3′ Sequence
O1F4*	ACAGCAGAAGATGGAACAACAGC
O1F3	TGTAGCAGGACCAGTAGTAG
O1F2	GCAGGAATAGAAGTAGAAGT
O1F*	GGCTACTGTAGAAATACTTGATGC
O2F	ATGAAGAAACAGGAGATACTGG
O3F3	GCTGATAATCCAGAAGAT
O3F2	CTTAGGAGCTACGAAGACG
O3F	AAGCATTCCAATTACTTGCAGAG
Inter3F	CACTCAGCGCCATTAATCC
I3F	GTGGACCAAGCGAAATC
O4F2	GCTGATAATCAGTCAGTGC
O4F	CAGCTTACGGAGTATTGGTGG
O5F	GCAGACAGTGGTGAGCATAAA
O1R	GTTCCATCTTCTGCTGTTA
O1RL*	GTTCCATCTTCTGCTGTTACCTGTATTACCATTGTTGC
O1R2	GCATCAAGTATTTCTACAGTAGCC
I1R	CCAGTTAAGTACTTCATCTT
I1R2	CTTACTTATAGGATAACGTTTCGCAC
I1R2L*	CTTACTTATAGGATAACGTTTCGCACTTTTATTGTATGG
O2R	CTGCTGTTTCTGTTACATCAC
O2R2	CCAGTATCTCCTGTTTCTTCAT
O2R3	GTCTATACTCACTGCCGTA
O2R4*	GTTTTGTTATTTCACTTATTAGTTCC
O3R	CGTCTTCGTAGCTCCTAAG
O3R2	CTCTGCAAGTAATTGGAATGCTT
O3R3*	CTATAGTATTTGGATAATTTTCC
O3R4	ATCTTCTGGATTATCAGC
Inter3R	GGATTAATGGCGCTGAGTG
I3R	GATTTCGCTTGGTCCAC
O4R	TGAAACTCCGTCAATATCC
O4R2*	GCACTGACTGATTATCAGC
O4R3	ACCAAGAGCCGTATCTACAAG
O4R4	CCACCAATACTCCGTAAGCTG
O4R5	GTAATCATCTGGTGCACTTGTA
O4R6	CTTAACCATGTTCCAAGCA
O5R	ATTTCTGGAGTATCTTGTTC
O5R2	CATACAAGTCTTCATCATA
O5R3*	TTTATGCTCACCACTGTCTGC
O5R4	CTCACCCACTGTCTGCTGATTCACTGC
O6R	ATCTATAACATCTACATCTGCTGC
O6R2	TGGTCTATAACCACCTTGCACTAC

*primer sequences marked above were used as primers for qRT-PCR

Figure 9. Primer locations in the region of interest.

An ‘O’ as the first letter of the primer name indicates the sequence is located within an ORF. ‘I’ indicates the primer sequence corresponds to intron sequence. ‘Inter’ indicates that the primer sequence is intergenic and corresponds to neither intron nor ORF sequence. The second place in the primer name corresponds to the ORF number (ie. O1F binds in the first ORF, 465, while O2F binds in the second ORF, 467) or intron number (ie. I3R corresponds to the third intron fragment). F or R indicate forward or reverse primers. Figure is not to scale.



2.3 Results

2.3.1 Bioinformatic Identification of C.te.*I1*

As stated previously, a search for ORF-less group II introns in bacterial genomes based on the conserved D5 motif (Simon et al. 2008) revealed the presence of a novel group II intron-like sequence in the genome of *Clostridium tetani* E88. The search located four group II intron D5 motifs within approximately 5 kb. Each D5 hit possessed nearly identical sequences and was followed immediately by a plausible D6 secondary structure (Figure 10). For the initial full-length sequence, referred to as *C.te.*I1, an entire secondary structure can be modeled. The structure, however, possesses several unconventional features especially within D1. These unusual features include an insertion in the EBS1 loop, and the lack of an EBS2 motif structure and IBS2-EBS2 pairing (Figure 10B) and are investigated in further detail in Chapter 3 of this thesis. The intron also does not encode an IEP, nor is there a group II intron-related RT sequence anywhere in the *C. tetani* genome.

Based on the presence of a CGC catalytic triad sequence within D5 and the proposed secondary structure model of the intron, *C.te.*I1 appears to be most similar to introns of Class B. Blast results using our local group II intron database (Altschul et al. 1997; Candales et al. 2012) also identify 8 out of 10 of the closest relatives of the intron as belonging to bacterial Class B (Table 2). Thus it appears that *C.te.*I1 is derived from a conventional Class B mobile group II intron that subsequently lost its IEP and acquired novel RNA structural features. Supporting this inference, Class B introns are the most common intron type in *Clostridium* and *Bacillus* genera, accounting for over half of group II introns encoded by those organisms (Simon et al, 2008; and unpublished data).

Figure 10. Genomic arrangement and intron secondary structure.

(A) The *C.te.I1* intron (red) is located between annotated ORFs CTC00465 (blue) and CTC00467 (dark grey), and is downstream of the major SLP gene of *Clostridium tetani* (*slpA*, CTC00462). Three downstream D5/6 motifs (outlined red boxes; D5/6-B, C, D) are followed by ORFs CTC00468, CTC00469 and CTC00470 (medium grey), together forming a series of four alternative 3′ splice sites. Dotted grey lines indicate potential alternative splicing reactions, and thin black lines are intergenic sequences. The diagram is drawn to scale. (B) Secondary structure model of *C.te.I1*. The structure is typical of Class B, but has significant structural variations near the EBS1 motif (grey boxes with dotted outline). The sequence of the 311 nt loop in domain 4 is not shown. The 5′ splice site predicted by the group II boundary motif is indicated with a black diamond, while the actual 5′ splice site is shown with a black star, and the 3′ intron boundary by a black triangle. The UAA stop codon of the upstream ORF CTC00465 is located between the predicted and actual 5′ splice sites, and is indicated with red shading. The IBS1-EBS1 pairing sequences are labelled with light blue shading, and other predicted tertiary interactions are indicated by Greek letters and grey shading. Yellow boxed nucleotides in domains 4b, 5 and 6 indicate polymorphisms among the four D5/6 motifs. The annotated start codon for the downstream exon is circled in green. (C) Secondary structures of the downstream domain 5 and 6 motifs, highlighting the sequence polymorphisms (yellow boxes) and annotated start codons (green circles) (compare with Panel B).

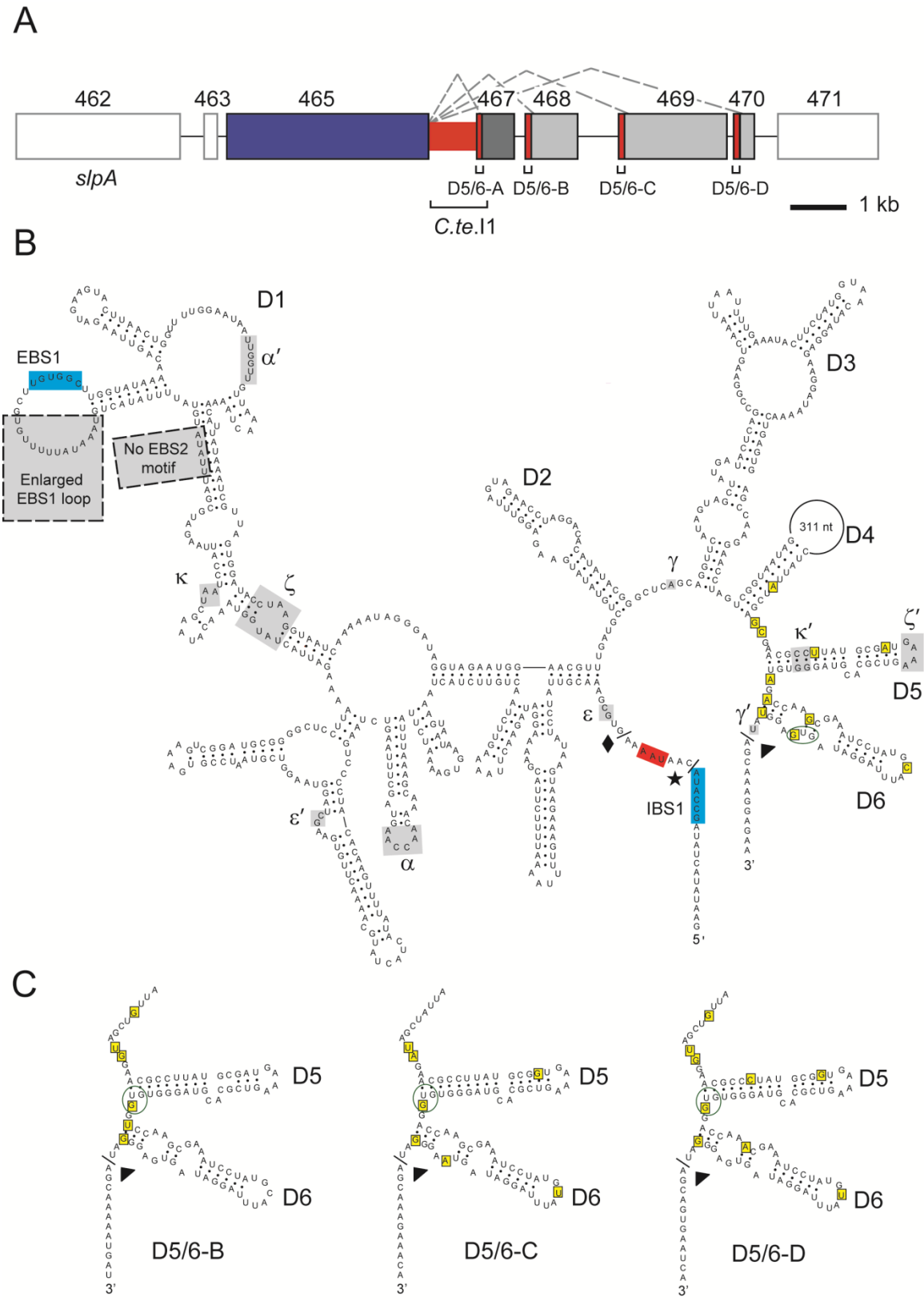

A
462
463
465
467
468
469
470
471
slpA
D5/6-A
D5/6-B
D5/6-C
D5/6-D
C.te.I1
1 kb
B
D1
EBS1
α′
Enlarged EBS1 loop
No EBS2 motif
D2
D3
311 nt
D4
κ
ζ
γ
κ′
ζ′
D5
ε
γ′
IBS1
D6
ε′
α
3′
5′
C
D5
D6
D5/6-B
D5/6-C
D5/6-D
3′

Table 2. Closest intron relatives of *C.te.*I1 as determined by a local, bacterial group II intron database blastn search

Sequences producing significant alignments	**Class**[a]	**Score**[b]	**E-value**[c]
*B.th.*I7	Bacterial B	55.4	3e-08
*Cl.pe.*I1	Bacterial B	55.4	3e-08
*Cl.pe.*I2	ORF-less	53.6	1e-07
*E.fm.*I4	Bacterial B	53.6	1e-07
*B.th.*I11	Bacterial B	48.2	5e-06
*C.ce.*I1	Bacterial B	48.2	5e-06
*Ly.sc.*I1	Bacterial B	48.2	5e-06
*B.th.*I5	Bacterial B	48.2	5e-06
*B.c.*I6	Bacterial B	48.2	5e-06
*Le.pn.*I1	CL1	46.4	2e-05

[a] IEP class corresponding to the intron identified as a High-Scoring Pair (HSP) noted in the above table as "sequence producing significant alignments".
[b] Score of HSP. Higher scores indicate better matches to the query sequence
[c] E-value corresponds to the E-value of the final gapped local alignment and is used to evaluate the significance of a HSP. E-values closer to zero indicate higher significance of the result.

The intron, *C.te*.I,1 is encoded within the *C. tetani* chromosome (CTC) in a region containing an array of surface layer associated genes (Brüggemann et al. 2003). The intron, *C.te*.I1, is encoded approximately 4 kb downstream of the predominant SLP expressed in *C. tetani, slpA* (CTC00462) (Qazi et al. 2007) and lies immediately downstream of an annotated SLP (CTC00465) that has been shown to be expressed in some *C. tetani* strains (Qazi et al. 2007). Downstream of *C.te*.I1 is the hypothetical ORF CTC00467 (Figure 10A). The three downstream D5/6 motifs are each directly followed by an ORF, CTC00468-CTC00470, with start codons annotated in the preceding D5/6 sequence (Figure 10 B, C).

Based on the predicted secondary structure, splicing of *C.te.*I1 would join CTC00465 and CTC00467 into one mRNA; but the ORFs would not be fused because the 5′ boundary of the intron (5′GUGCG) lies 2 nucleotides downstream of the stop codon of CTC00465 (Figure 10B). In addition, splicing would remove the annotated start codon of CTC00467 in D5/6, and presumably disrupt translation of the downstream 467 ORF. Potential alternative splicing reactions involving domains 1-4a of *C.te.*I1 and any of the three downstream D5/6 motifs would similarly fail to link the reading frames and would remove the start codons of the downstream ORFs (CTC00468-CTC00470). Based on the bioinformatic analysis alone, the functionality of *C.te.*I1 was unclear as the unusual secondary structure features, the absence of an IEP, and the predicted failure to ligate the exon reading frames would perhaps suggest the intron is non-functional.

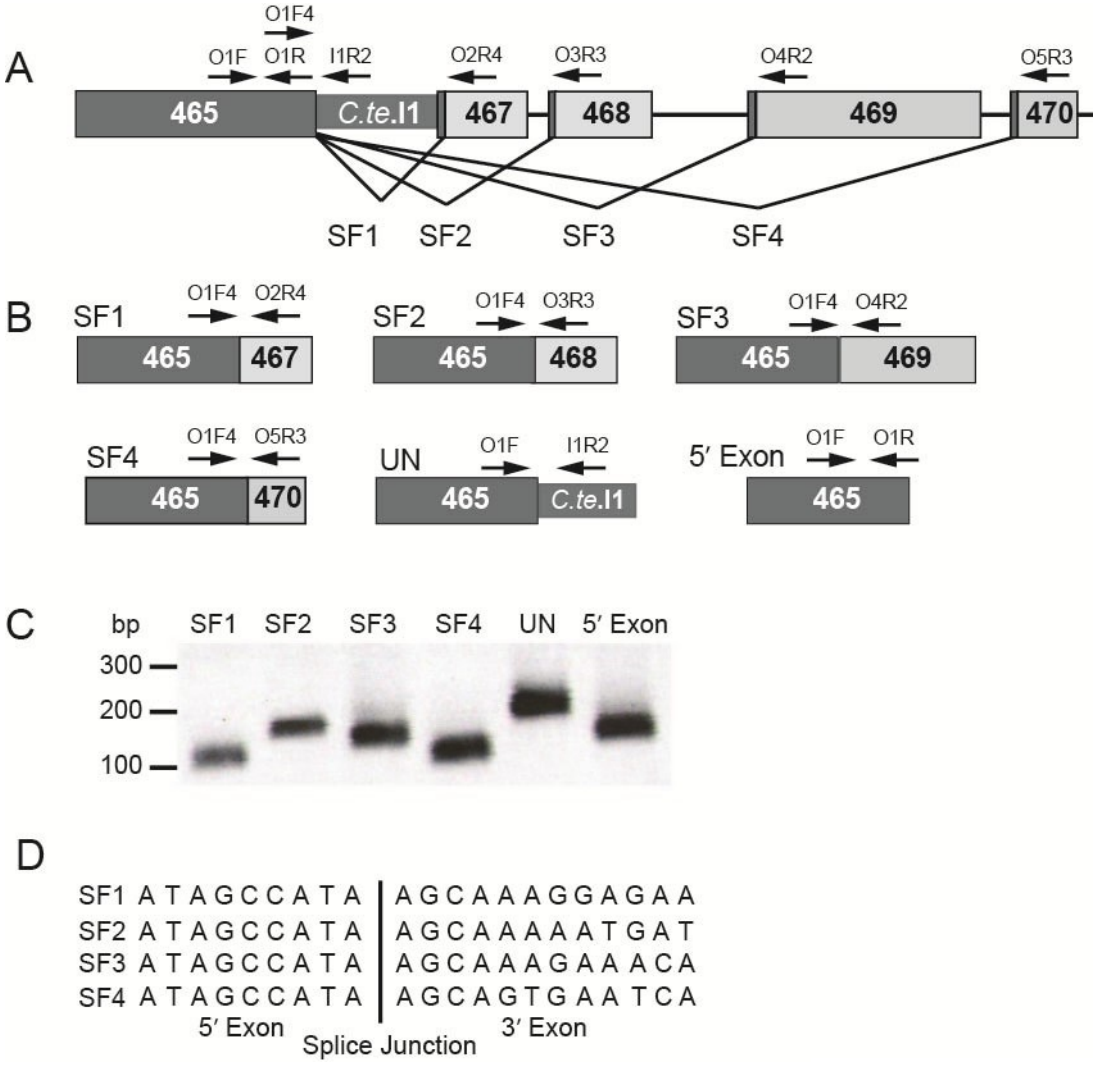
2.3.2 C.te.*I1 Splices* in vivo *to Produce Five Distinct Coding Sequences*

To analyze *C.te*.I1 splicing *in vivo* and determine if the intron was indeed functional and therefore potentially capable of alternative splicing, the *C. tetani* strain ATCC 10779 was acquired. Both the sequenced strain, Massachusetts substrain E88, and ATCC 10779 are derivatives of the Harvard strain (Mueller and Miller 1945) and ATCC10779 is thought to be the parent strain of E88 (Johnston et al. 2010). As such, both strains are likely to be highly similar in genomic sequence. Over the course of this study, a region of ~6.5 kb containing the intron *C.te*.I1 was PCR amplified and sequenced. The region was confirmed to be identical to the sequenced strain E88. A single gap of 559 bp was not amplified that corresponds to the intergenic region between CTC00468 and CTC00469.

Both *in vivo* splicing and alternative splicing of the *C.te*.I1 intron were investigated by RT-PCR using total cellular RNA from strain ATCC10779. The RT-PCR assays amplified all four candidate exon junctions, as well as unspliced transcript (Figure 11C). The amplified DNAs were cloned and sequenced (Figure 11D). Sequencing revealed that the splice junctions do not conform to standard group II intron 5′ boundaries. For each of the four sequenced spliced products, the 5′ exon-intron boundary was found to be shifted eight nucleotides upstream of the canonically predicted intron start (5′GUGYG). The use of this unusual splice site eliminates the stop codon of the upstream ORF (CTC00465) and results in the in-frame ligation of the flanking exons. These data show that *C.te*.I1 is indeed functional for splicing *in vivo*, and moreover is capable of four alternative splicing reactions, which along with unspliced transcript produces five distinct RNAs.

Figure 11. Alternative splicing *in vivo*.

(A) The four alternative splicing reactions are indicated by black lines joining CTC00465 with downstream exon ORFs (splicing forms SF1-4). Primers for PCR reactions are shown by arrows. (B) Diagram of spliced exons and the RNAs amplified by RT-PCR. (C) Agarose gel of RT-PCR amplification products. RNA preparations were digested with DNase I prior to RT-PCR reactions, and RT-PCR products were dependent on reverse transcriptase (data not shown). (D) Sequences of cloned splice junctions.



2.3.3 *Quantification of Alternative Spliced RNAs Produced* in vivo

In eukaryotes, most of the well-studied examples of alternative splicing occur in a highly regulated manner, such that specific alternatively spliced isoforms are produced in various tissue types, at various developmental stages, or in response to external stimuli [reviewed in (Nilsen and Graveley 2010)]. It remains unclear how much this strong representation of regulated alternative splicing events in the literature is due to research bias and how often alternative splicing events resulting in constitutive ratios of alternatively spliced isoform expression are simply overlooked.

To determine if *C.te*.I1 undergoes regulated alternative splicing, quantitative real-time reverse-transcriptase PCR (qRT-PCR) experiments were performed using total RNA from *C. tetani* ATTC10779 cells. The integrity, purity and concentration of the RNA preparations were assessed prior to use in the qRT-PCR assays. Integrity of RNA was accessed visually by running an aliquot of the RNA sample on 1.2% agarose gel, while purity and concentration were assessed via 260/280 ratios on a NanoDrop-1000 Spectrophotometer (ThermoScientific). Six RNA segments were amplified: a 5′ exon segment, the unspliced exon-intron junction, and the four spliced exon junctions (Figure 11) and absolute amounts of the alternatively spliced isoforms were determined from standard curves prepared from cloned versions of the exon junctions. Efficiency of amplification for each sequence was between 95- 107% and amplification was specific as single products are observed both in melt curves and on a gel (Appendix A; Figure 11). The unspliced RNA isoform was determined to be predominant with significantly lower but detectable levels of all other splice forms present (Figure 12A).

Splicing levels were then analyzed from a variety of different growth and stress conditions. RNA was extracted from *C. tetani* ATCC10779 cultures subjected to stress conditions, including some of those known to influence gene expression of virulence factors in pathogenic bacteria (Mekalanos 1992). Treatments included stresses of osmotic strength, temperature, oxygen exposure and metal ions. *C. tetani* cells were also collected throughout growth phases, and over multiple experimental preparations, as well as different strengths of the nutrient broth (see Materials and Methods). Under all conditions tested, unspliced RNA accounted for the majority of transcripts comprising ~75-95% of total 5′ exon containing mRNAs (Figure 12B, Tables 3 and 4).

Although some variation in splice forms was observed, the differences were not dramatic. In general, only slight variation was seen under any condition, but variation also existed between replicate trials of the conditions, suggesting a degree of stochastic variation in expression of splicing from the locus. As the list of conditions tested is not exhaustive and it is difficult to simulate natural conditions in the laboratory, it is possible that significant regulation may occur in response to conditions other than those tested or upon infection.

In addition to regulated alternative splicing and stochastic variability, scenarios that can be envisioned for the locus include alternative splicing that may occur constitutively and function to produce a low level of four variant proteins in addition to the predominant isoform of CTC00465 or alternative splicing may be cell-specific such that a certain percentage of cells within a population express each of the different variants in a manner similar to the expression of the *C. difficile* cell wall protein, CwpV (Emerson et al. 2009).

Although random variation is observed between different trials of biological conditions and no significant trends in splicing are apparent, a few points in regard to the data can be noted. In general, it appears that splicing occurs more frequently up until cells reach an OD_{600} of approximately 0.6 with 15-20% of 5′ exon containing RNAs representing splice forms at the earlier points of growth (Figure 13A). In particular SF2, SF3, and SF4 represent a larger percentage of transcripts at this stage than they do later at higher ODs. In the later stages of growth less than 10% of 5′ exon containing RNAs correspond to splice forms in the majority of trials (exceptions can be noted for the OD=0.7(1), OD=0.8(2X BHI) and OD=1.0 data points).

Following oxygen exposure, only low levels of splicing are observed. However, for the OD=0.7 trial series very low levels of splicing were observed before the exposure to oxygen as well (Figure 13B). For the OD=0.54 trial this shift is perhaps due to a response to oxygen exposure itself and/or induced sporulation but may simply reflect increased growth , following induction of the stressor, that would seemingly also lead to a decrease in splicing, based on the above data that splicing decreases with increase culture OD.

For the other stress conditions tested, stochastic variation rather than any form of trend in the data is observed. Somewhat notable is that production of SF4 seems to increase upon ZnOAc exposure (Figure 14A), however as there is variation in splicing observed between all conditions it cannot be concluded that this variation was caused by ZnOAc exposure. One trend that appears is that subculturing and re-streaking subsequent generations of individual colonies leads to increased production of splice forms over the

Figure 12. Quantification of splice forms by qRT-PCR.

(A) Quantification of RNAs present in 1 μg of total RNA isolated from *C. tetani* cultures grown to an OD_{600} of 0.125. The four splice forms, unspliced RNA and 5′ exon were quantified, and molar amounts were calculated based on calibrations with DNA standards of identical sequence and known quantity (see Materials and Methods). Means are based on three technical replicates and errors bars indicate standard deviation.

(B) qPCR quantification of RNA forms present for various growth conditions. Values are expressed as % of the total molar amounts of the five RNAs. Each trial shown is the mean of three technical replicates. For a detailed description of each of the conditions shown, see Materials and Methods. A full table of values with standard deviations is presented in Table 3 and Table 4.

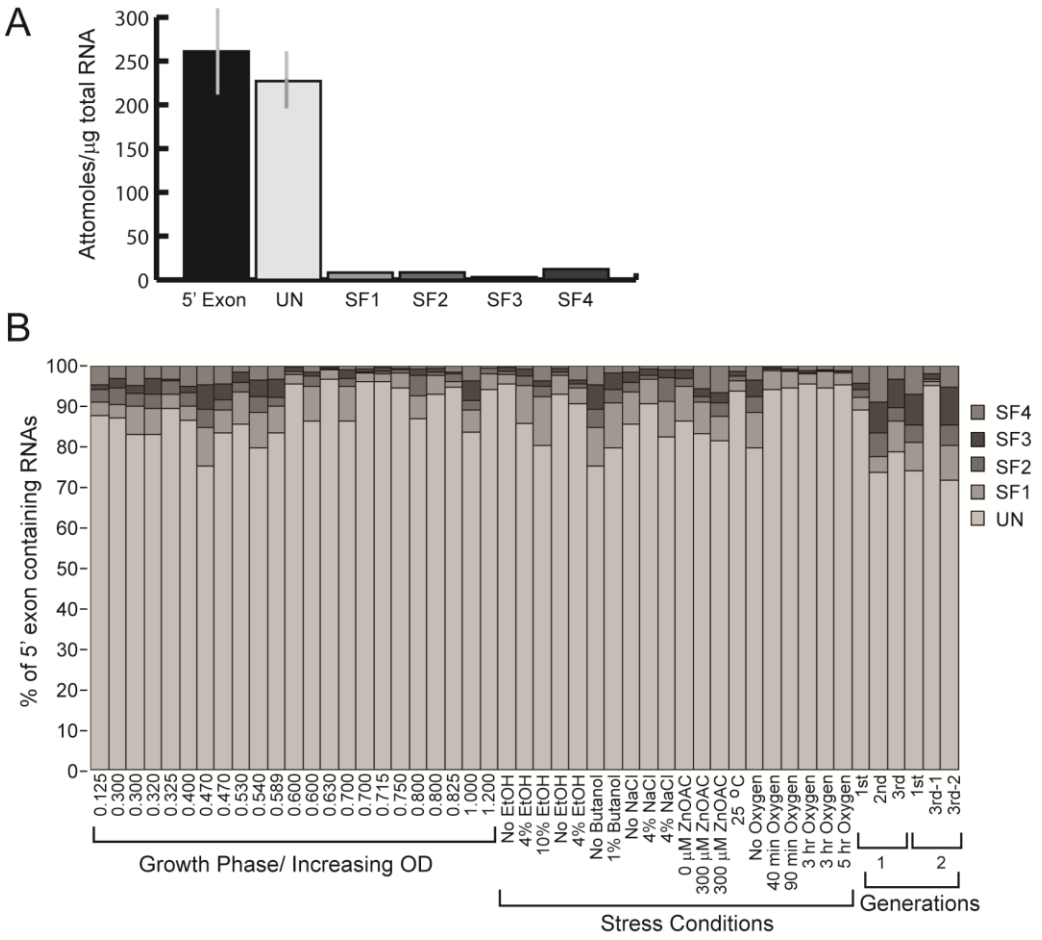


Figure 13. qRT-PCR data for OD and oxygen exposure.

(A) Expression of alternatively splice forms throughout the growth curve. Numbers in brackets indicate separate trials of the OD [ie. (1) or (2)] or indicate if alternate concentrations of the media were used for that trial [ie. (0.75X BHI)] (B) Variation in splice form expression following exposure to oxygen in two separate trials. Numbers in the x-axis labels correspond to culture ODs prior to exposure in the various trials. The legend shown in (B) corresponds to (A) as well. Values presented in this figure are derived from the data presented in Tables 3 and 4.

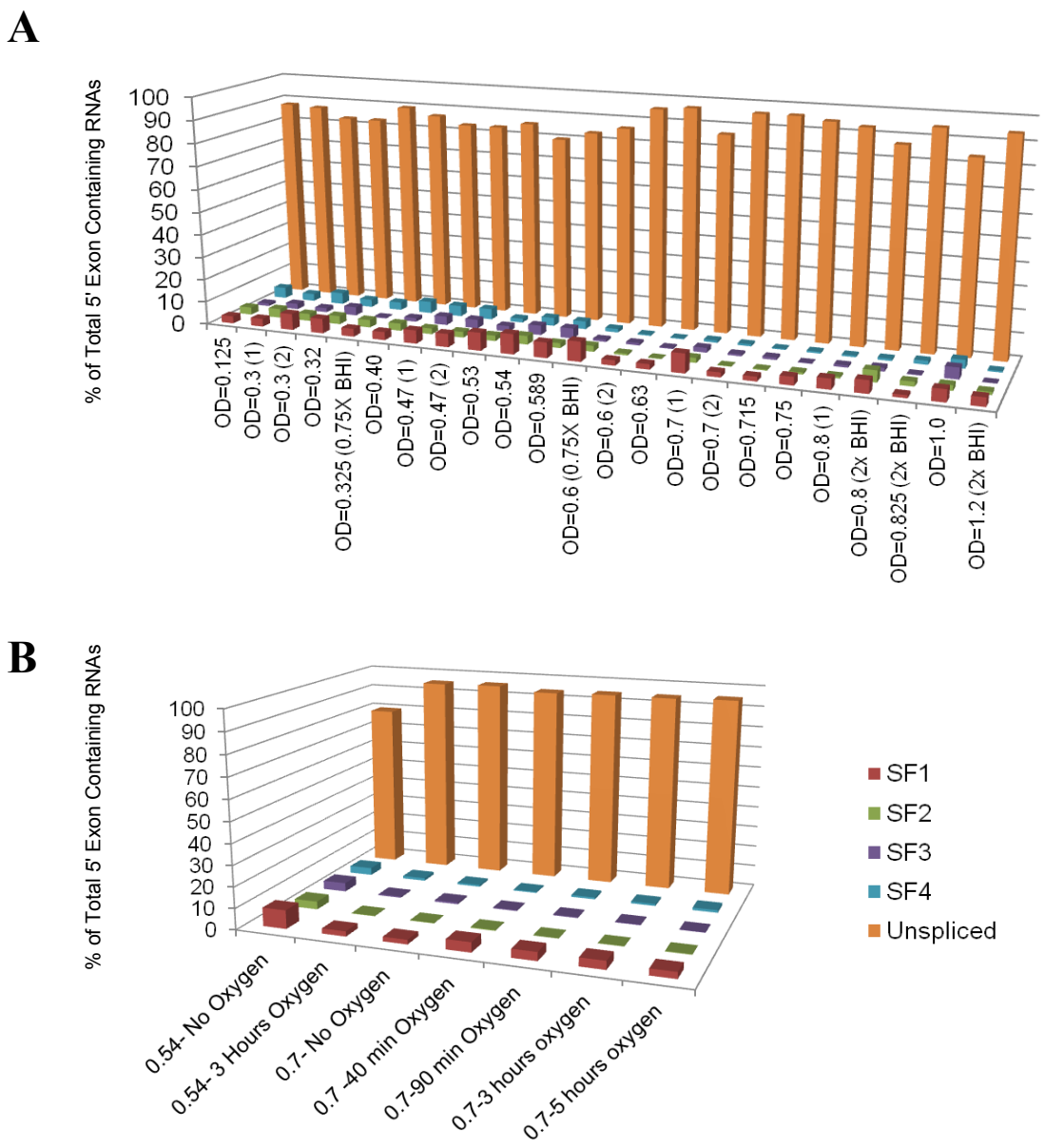
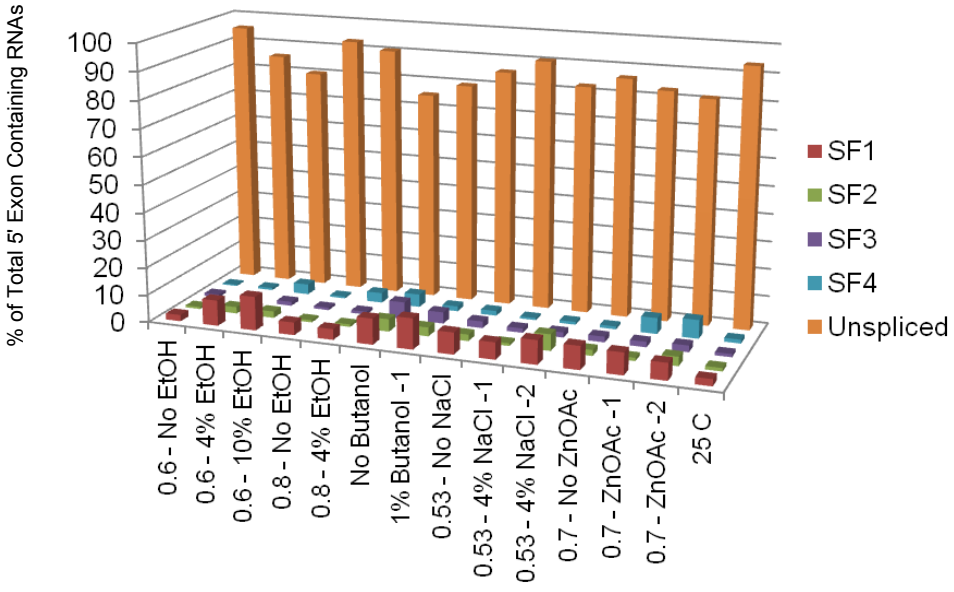


Figure 14. qRT-PCR data for stress conditions and culture generations.

(A) Effect of various stress conditions on the production of different alternative splice forms. "-1" and "-2" indicate data from separate trials. Numbers in the x-axis labels correspond to culture ODs prior to exposure in the various trials. (B) Variation in splice forms observed between individual colonies over generations. Values presented in this figure are derived from the data presented in Table 4.

A



B

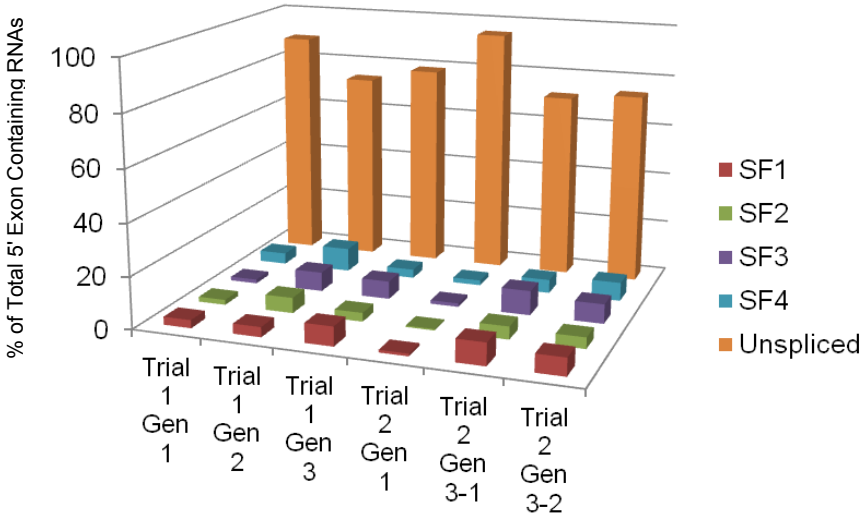


Table 3. qRT-PCR data throughout the growth cycle for *C . tetani* ATCC10779[1].

	SF1			**SF2**			**SF3**			**SF4**			**Unspliced**			**5′ exon (Control)**		
Conditions	$\bar{x}$ [2]	σ[3]	No RT[4]	$\bar{x}$	σ	No RT	$\bar{x}$	σ	No RT	$\bar{x}$	σ	No RT	$\bar{x}$	σ	No RT	$\bar{x}$	σ	No RT
OD=0.125	0.826	0.072	0.00078	0.839	0.1781	0.00055	0.306	0.0862	0.00102	1.220	0.3149	0.0007	22.7	3.4117	0.00207	26.1	4.6765	0.0057
OD=0.3 (1)[5]	1.012	0.182	0.00098	1.151	0.3225	0.000325	0.703	0.2047	0.00158	0.922	0.1843	0.0092	25.3	3.5384	0.00423	22.467	3.6937	0.0021
OD= 0.3 (2)	1.101	0.830	0.00069	0.488	0.0384	0.00186	0.285	0.0670	0.00167	0.769	0.1398	0.00226	12.833	0.9074	0.00267	13.617	8.1232	0.0024
OD=0.32	0.517	0.043	0.00006	0.294	0.0735	0.000851	0.300	0.0822	0.00324	0.262	0.0244	0.00286	6.653	0.7060	0.00203	1.863	0.4500	0.0004
OD=0.325 (0.75XBHI)[6]	0.115	0.024	0.00092	0.111	0.0098	0.00114	0.015	0.0077	0.000284	0.113	0.0207	0.00529	2.983	0.2479	0.00291	2.82	0.3751	0.0014
OD=0.47(1)	0.448	0.094	0.00028	0.192	0.0301	0.000734	0.292	0.1194	0.00169	0.359	0.0195	0.00325	6.427	0.9468	0.00234	7.033	2.2712	0.0006
OD=0.47 (2)	0.349	0.021	0.00126	0.168	0.007	0.00654	0.230	0.0052	0.00505	0.174	0.0085	0.00094	2.797	0.7629	0.0833	7.227	0.4271	0.0838
OD=0.53	0.423	0.065	0.0016	0.127	0.0108	0.0032	0.132	0.0489	0.00509	0.092	0.0051	0.0011	4.603	0.3761	0.0961	8.647	1.2189	0.1
OD=0.54	0.311	0.016	0.0040	0.133	0.0105	0.00082	0.148	0.0232	0.00559	0.126	0.0085	0.0011	2.817	0.4607	0.077	5.987	1.1851	0.0573
OD=0.55 (1)	0.024	0.020	0.00082	0.024	0.0046	0.000661	0.021	0.0055	0.000908	0.042	0.0164	0.00312	3.55	0.6065	0.00304	2.853	0.9084	0.0021
OD=0.55 (2)	0.034	0.019	0.00122	0.041	0.0084	0.000975	0.022	0.0078	0.00139	0.029	0.0109	0.00689	5.133	2.2893	0.00348	5.087	1.4500	0.0012
OD=0.589	0.554	0.078	0.0012	0.174	0.0092	0.0011	0.362	0.1414	0.00169	0.281	0.0331	0.00397	6.847	0.6667	0.00239	7.483	3.1948	0.0266
OD=0.6 (0.75X BHI)	1.733	0.104	0.0019	0.505	0.0476	0.00199	0.198	0.0807	0.00145	0.310	0.0354	0.00111	17.2	1.8520	0.00872	16.073	12.280	0.0209
OD=0.6 (2)	0.315	0.018	0.0029	0.105	0.0085	0.0011	0.139	0.0176	0.00524	0.062	0.0008	0.00169	12.933	2.4685	0.057	13.633	0.6110	0.033
OD=0.63	0.334	0.065	0.00015	0.030	0.0298	0.000731	0.064	0.0471	0.000009	0.055	0.0052	0.00301	14.1	4.2	0.00172	10.027	3.2617	0.0005
OD=0.7	1.006	0.091	0.00379	0.253	0.0220	0.00849	0.246	0.0289	0.00601	0.123	0.0075	0.00088	10.32	2.6045	0.06	16.7	1.3077	0.0944
OD=0.7	0.160	0.020	0.00299	0.017	0.0015	0.00611	0.060	0.0055	0.000009	0.066	0.0052	0.0139	7.199	1.301	0.0107	9.523	1.1417	0.0030
OD=0.715	0.483	0.061	0.00023	0.102	0.0341	0.000089	0.253	0.1208	0.000042	0.102	0.0085	0.0032	22.767	2.6727	n/d	21.133	5.2548	0.0165
OD=0.75 (1)	1.617	0.090	0.00107	0.394	0.0816	0.000119	0.115	0.0506	0.000905	0.312	0.1083	0.00641	42.133	0.7572	0.0802	25.267	5.1003	0.1
OD=0.75 (2)	1.012	0.005	0.00714	0.119	0.0086	0.0025	0.183	0.0301	0.00951	0.097	0.0043	0.00002	0.113	0.0162	0.137	0.110	0.0213	0.115
OD=0.8 (1)	0.966	0.011	0.01434	0.168	0.0131	0.00169	0.193	0.0135	0.00744	0.146	0.0021	n/d[7]	19.567	2.9143	0.425	24.8	1.1790	0.137
OD=0.8 (2X BHI)	0.819	0.184	0.00072	0.739	0.0709	0.000838	0.226	0.0923	0.00141	0.136	0.0321	0.00605	12.633	0.4726	0.00408	17.5	2.5865	0.0294
OD=0.825 (2X BHI)	0.058	0.005	0.0007	0.093	0.0123	0.000478	0.013	0.0066	0.000891	0.077	0.0129	0.00547	4.3	0.5069	0.0024	3.343	1.1719	0.0002
OD=1.0	0.054	0.007	0.00084	0.023	0.0108	0.00108	0.045	0.0181	0.00148	0.037	0.0044	0.00377	0.832	0.0797	0.00222	1.453	0.3729	0.0083
OD=1.2 (2X BHI)	2.37	0.135	0.00074	0.821	0.2449	0.000035	0.076	0.0551	0.00146	0.371	0.1424	0.00843	57.767	5.1936	0.00315	51.3	11.241	0.0043

[1]Values are in attomoles of DNA per 100 ng total RNA – refer to materials and methods for exact conditions used.
[2]Mean value of three technical triplicates [3]Standard deviation [4] "No RT" is the negative control value for RNA samples with no reverse transcriptase added
[5] (1) or (2) indicate separate trials of identical conditions [6] BHI= brain heart infusion media [7] n/d indicates the value was not determined.

Table 4. qRT-PCR data for *C. tetani* stress conditions[1]

		SF1			SF2			SF3			SF4			Unspliced			5' exon (Control)		
Conditions		$\bar{x}$ [2]	σ[3]	No RT[4]	$\bar{x}$	σ	No RT	$\bar{x}$	σ	No RT	$\bar{x}$	σ	No RT	$\bar{x}$	σ	No RT	$\bar{x}$	σ	No RT
OD=0.325 (0.75XBHI)[6]		0.115	0.024	0.00092	0.111	0.0098	0.00114	0.015	0.0077	0.000284	0.113	0.0207	0.00529	2.983	0.2479	0.00291	2.82	0.3751	0.0014
OD=0.6 (0.75X BHI)		1.733	0.104	0.0019	0.505	0.0476	0.00199	0.198	0.0807	0.00145	0.310	0.0354	0.00111	17.2	1.8520	0.00872	16.073	12.280	0.0209
OD=0.8 (2X BHI)		0.819	0.184	0.00072	0.739	0.0709	0.000838	0.226	0.0923	0.00141	0.136	0.0321	0.00605	12.633	0.4726	0.00408	17.5	2.5865	0.0294
OD=0.825 (2X BHI)		0.058	0.005	0.0007	0.093	0.0123	0.000478	0.013	0.0066	0.000891	0.077	0.0129	0.00547	4.3	0.5069	0.0024	3.343	1.1719	0.0002
OD=1.2 (2X BHI)		2.37	0.135	0.00074	0.821	0.2449	0.000035	0.076	0.0551	0.00146	0.371	0.1424	0.00843	57.767	5.1936	0.00315	51.3	11.241	0.0043
OD=0.6	No EtOH	0.315	0.018	0.0029	0.105	0.0085	0.0011	0.139	0.0176	0.00524	0.062	0.0008	0.00169	12.933	2.4685	0.057	13.633	0.6110	0.033
	4% EtOH	0.089	0.005	0.00067	0.023	0.0126	0.000225	0.018	0.0014	0.00356	0.008	0.0002	0.00005	0.8243	0.0159	0.0166	2.637	0.3372	0.0432
	10% EtOH	2.589	0.103	0.0006	0.496	0.0423	0.000529	0.314	0.0137	0.00867	0.788	0.0717	0.00009	16.9	1.7521	0.179	28.267	1.7953	0.0244
OD=0.8	No EtOH	0.966	0.011	0.01434	0.168	0.0131	0.00169	0.193	0.0135	0.00744	0.146	0.0021	n/d	19.567	2.9143	0.425	24.8	1.1790	0.137
	4% EtOH	0.036	0.001	0.00039	0.009	0.0030	0.031	0.010	0.0005	0.006039	0.033	0.0049	0.01209	0.838	0.016	0.00948	0.882	0.0214	n/d
OD=0.47	No Butanol	0.349	0.021	0.00126	0.168	0.007	0.00654	0.230	0.0052	0.00505	0.174	0.0085	0.00094	2.797	0.7629	0.0833	7.227	0.4271	0.0838
	1% BuOH	0.775	0.038	0.00118	0.224	0.0246	0.000232	0.286	0.0266	0.00696	0.124	0.0017	0.00136	5.523	0.0611	0.0113	8.401	0.4136	0.062
OD=0.75	No Butanol	1.012	0.005	0.00714	0.119	0.0086	0.0025	0.183	0.0301	0.00951	0.097	0.0043	0.00002	0.113	0.0162	0.137	0.110	0.0213	0.115
	1% BuOH	0.081	0.003	0.00034	0.092	0.0280	0.00914	0.109	0.0149	0.004618	0.092	0.0152	0.013	10.096	1.007	0.01181	10.938	1.8463	n/d[7]
OD=0.53	No NaCl	0.423	0.065	0.0016	0.127	0.0108	0.0032	0.132	0.0489	0.00509	0.092	0.0051	0.0011	4.603	0.3761	0.0961	8.647	1.2189	0.1
	4% NaCl 1	0.231	0.014	0.00046	0.037	0.0094	0.000147	0.060	0.0021	0.00608	0.029	0.0007	0.00004	3.4	0.0721	0.0219	6.453	0.2723	0.0527
	4% NaCl 2	0.044	0.001	0.00427	0.030	0.0062	0.0288	0.009	0.0006	0.00687	0.005	0.0010	0.00066	0.415	0.020	0.00881	0.642	0.0362	0.006632
OD=0.7	No ZnOAc	1.006	0.091	0.00379	0.253	0.0220	0.00849	0.246	0.0289	0.00601	0.123	0.0075	0.00088	10.32	2.6045	0.06	16.7	1.3077	0.0944
	ZnOAc (1)[5]	0.038	0.001	0.00082	0.006	0.0006	0.000243	0.009	0.0011	0.00464	0.028	0.0019	0.00026	0.392	0.0195	0.0171	0.887	0.0420	0.0508
	ZnOAc (2)	0.028	0.003	0.00039	0.015	0.0003	0.0174	0.011	0.0008	0.00747	0.030	0.0046	0.0065	0.363	0.020	0.00712	0.318	0.1713	0.0058
25 °C		0.313	0.022	0.00054	0.153	0.0138	n/d	0.143	0.0130	0.006198	0.177	0.0091	0.0101	11.863	0.601	0.01025	10.86	0.14	n/d
OD=0.54	No O_2	0.311	0.016	0.0040	0.133	0.0105	0.00082	0.148	0.0232	0.00559	0.126	0.0085	0.0011	2.817	0.4607	0.077	5.987	1.1851	0.0573
	3 Hours O_2	0.167	0.013	0.0008	0.017	0.0016	0.000252	0.037	0.0015	0.00617	0.088	0.0068	0.00004	6.53	0.6428	0.0225	9.57	0.5462	0.0338
OD=0.7	0 min O_2	0.160	0.020	0.00299	0.017	0.0015	0.00611	0.060	0.0055	0.000009	0.066	0.0052	0.0139	7.199	1.301	0.0107	9.523	1.1417	0.0030
	40 min O_2	0.746	0.038	0.00336	0.064	0.0063	0.00434	0.064	0.0050	0.007589	0.075	0.0545	0.0194	14.857	1.050	0.0102	10.213	4.9364	0.006
	90 min O_2	0.353	0.020	0.00298	0.010	0.0017	0.00391	0.028	0.0025	0.005439	0.079	0.0053	0.0125	8.155	1.121	0.00416	13.933	0.8387	0.0056
	3 hours O_2	0.172	0.017	0.00195	0.006	0.0031	0.00903	0.015	0.0015	0.009055	0.040	0.0012	0.00730	4.036	0.061	0.00596	0.007	0.0001	0.0081
	5 hours O_2	0.062	0.004	0.00343	0.005	0.0019	0.0201	0.005	0.0026	0.0048	0.025	0.0062	0.00314	1.918	0.583	0.00119	4.400	0.1703	0.0103
Trial 1	1st Gen	0.257	0.002	0.00207	0.161	0.0040	0.0241	0.126	0.0210	0.005161	0.359	0.0907	0.00362	7.3543	0.2180	0.01246	9.126	0.5293	0.0042
	2nd Gen	0.198	0.008	0.00232	0.307	0.0514	0.00687	0.383	0.0290	0.002937	0.468	0.0717	0.0192	3.772	0.2268	0.00077	2.667	0.3747	0.0050
	3rd Gen	0.468	0.037	0.00047	0.212	0.0055	0.00186	0.432	0.1455	0.000425	0.210	0.0081	0.00335	4.903	0.465	0.007	10.833	0.0577	0.0046
Trial 2	1st Gen	0.067	0.003	0.00023	0.036	0.0067	0.00686	0.091	0.0166	0.0043	0.130	0.0031	0.00672	6.323	0.376	0.0477	7.707	0.5299	0.078
	3rd Gen-1	1.243	0.083	0.00020	0.710	0.1080	0.000023	1.357	0.0379	0.000122	0.751	0.0254	0.00027	10.263	0.782	0.00938	18.633	1.7243	0.0018
	3rd Gen -2	0.112	0.011	0.00082	0.071	0.0039	0.0016	0.123	0.0244	0.00105	0.117	0.0085	0.00049	1.207	0.106	0.0143	2.08	0.3869	0.193

[1]Values are in attomoles of DNA per 100 ng total RNA

[2]Mean value of three technical triplicates [3]Standard deviation [4] "No RT" is the negative control value for RNA samples with no reverse transcriptase added

[5] (1) or (2) indicate separate trials of identical conditions [6] BHI= brain heart infusion media [7] n/d indicates the value was not determined.

initial culturing (Figure 14B). Although trial 1 generation 3 shows a decrease in splicing over generation 2, it is still an increase over the initial generation, Gen 1, for that trial. It should also be noted that in almost every trial the most abundant RNA form other than the unspliced transcript is SF1, in which the 5′ exon (CTC00465) is ligated to the 3′ exon immediately following the full copy of the intron (CTC00467).

2.3.4 Potential Functions of ORFs Encoded by 5′ and 3′ Exons

Each of the five alternatively spliced mRNAs is predicted to produce a different protein isoform with expected molecular weights of 145, 179, 178, 218 and 155 kDa corresponding to unspliced transcript and splice forms 1-4 respectively. The specific functions for the various alternatively spliced ORFs are not known, however the 5′ exon ORF (CTC00465) has been identified as a minor constituent in SLP preparations from the *C. tetani* strain CN655 (Qazi et al. 2007). While it is known that CTC00465 is expressed as a component of the surface layer, CTC00465 appears to be a non-essential component of the surface layer. MALDI-ToF MS of the non-pathogenic strain the Fairweather group tested did not identify the CTC00465 ORF and similar sized bands were not observed for any of the clinical isolates tested (Qazi et al. 2007). Of the four downstream exon ORFs that become appended to CTC00465 by the alternative splicing reaction, three show similarities to transglutaminases or proteases (Table 5).

Transglutaminases are a large family of enzymes that catalyze post-translational modifications via transamidation of glutamine residues, resulting in the formation of ε-(γ-glutamyl) lysine crosslinks (Folk 1980). Creation of these crosslinks increases resistance to proteolytic degradation, which in turn increases the strength of the S-layer. In addition,

Table 5. Inferred functions of exon-encoded proteins

Annotated ORF	Exon Location	GenBank Annotation[a]	BLASTP Matches[b]	Pfam Matches[c]
CTC00465	5′ exon	Putative S-layer protein	Hypothetical protein, Cell wall binding protein, Extracellular nuclease	No matches
CTC00467	3′ exon	Hypothetical protein	Hypothetical protein	No significant matches
CTC00468	3′ exon	Predicted transglutaminase/ protease	Transglutaminase/protease, Cell wall binding protease, Surface layer protein	Transglutaminase -like superfamily
CTC00469	3′ exon	Predicted transglutaminase/ protease	Cell wall binding protease, Transglutaminase/protease, Surface layer protein	Transglutaminase -like superfamily
CTC00470	3′ exon	Hypothetical protein	Surface layer protein, Transglutaminase/protease, Hypothetical protein	No significant matches

[a]Protein function listed in GenBank annotation.
[b]The three most common functional descriptions of protein matches. (e < 2e-07; except for CTC00470 where matches were between 4.4 and 0.008) (www.ncbi.nlm.nih.gov/BLAST/).
[c]Protein family domains identified by Pfam 3.0 (www.pfam.sanger.ac.uk).

proteases are known to be required for assembly of some surface layers. For example, SlpA in *Clostridium difficile* is proteolytically processed into high- and low-molecular weight forms (Kirby et al. 2009). In considering the biological function of *C.te.*I1 alternative splicing, it is plausible that cross-linking and/or protease activities encoded by downstream exons could alter the S-layer properties, either directly or through the assembly process, potentially proving advantageous for the cell.

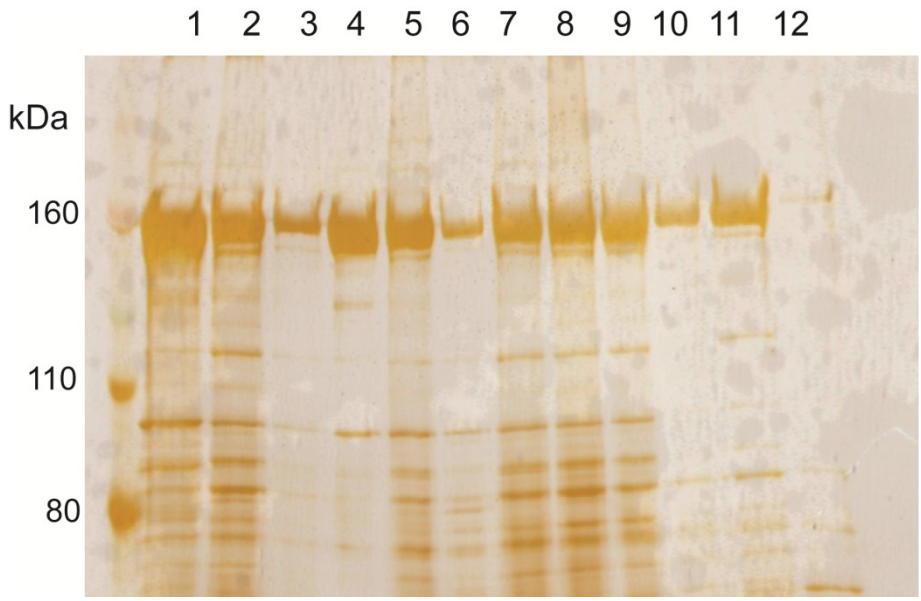
2.3.5 Analysis of Surface Layer Proteins Expressed by C. tetani *ATCC10779*

Although it was determined that ATCC10779 shows alternative splicing at the RNA level, it was unknown whether the effects of alternative splicing would be detectable at the protein level. To examine whether the observed mRNA splice forms were subsequently both translated and secreted as SLPs in the ATCC10779 strain, SLPs were extracted using 4M urea to disrupt the non-covalent interactions between the S-layer and the rest of the cell envelope. A caveat to these extraction conditions, however, is that secreted extracellular proteins and other proteins that complex with the S-layer will also be represented within the supernatants. Consistent with previous work (Qazi et al. 2007), the major band was ~160 kDa (Figure 15) and was assumed to be the major surface layer protein, SlpA. No other bands of higher molecular weight were observed. The identity of the single SLP band was later confirmed by Matrix Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF-MS) and peptide fingerprinting to be SlpA (CTC00462) (Table 6).

Although some strains of Gram positive bacteria have been shown to change the SLPs they express in response to growth conditions such as elevated temperature and

Figure 15. Extraction of surface layer proteins from ATCC10779.

Extractions were performed with 4 M urea from cell grown under a variety of different growth and stress conditions. It should be noted that these extraction conditions simply disrupt the non-covalent interactions that attach the S-layer and will result in various secreted proteins and proteins interacting with the S-layer also being represented in the SLP preparations. Lane 1 is an extraction from fresh liquid broth culture, lanes 2-5 are extractions from various individual colonies, lane 6 represents an extraction from frozen cells, lane 7 is following oxygen stress, lane 8 is following osmotic stress, lane 9 is exposure to 4% ethanol, lane 10 is exposure to 10% ethanol, lane 11 is 1% butanol exposure, and lane 12 is ZnOAc metal ion stress. Ten µL of each SLP preparation was loaded on 10 % SDS-PAGE and the gel was silver stained.



dissolved oxygen content of growth media (Egelseer et al. 2001; Scholz et al. 2001), when SLPs were extracted from a selection of the conditions used for qRT-PCR no variation in expression was observed with visualization by silver staining (Figure 15). A few slower migrating bands are faintly visible in a few of the preparations but as these bands are not visible by Coomassie staining their identities could not be confirmed via MALDI-ToF MS. Therefore it can be concluded that although the mRNA for CTC00465 and its various splice forms are expressed in the ATCC10779 strain, the corresponding proteins are either not translated or not secreted in this strain or are translated and secreted in minute quantities not detectable by silver staining.

2.3.6 Analysis of SLP Expression in Additional C. tetani *Strains*

C. tetani ATCC10779 is used in toxoid production for vaccine purposes and therefore has been passaged extensively in the laboratory. It is this repeated passaging that has been suggested to be responsible for the differences in lipid expression it exhibits compared to other *C. tetani* stains (Johnston et al. 2010). It was suggested that perhaps through the extensive passaging of the strain, ATCC10779 may have lost the ability to establish infection (Johnston et al. 2010). As passaging may alter S-layer expression compared to environmental or clinical isolates and has been shown to result in complete loss of S-layer expression in some cases (Sleytr et al. 1996), other strains of *C. tetani* were obtained to further investigate the effects of alternative splicing.

Previous work with *C. tetani* strains showed that protein profiles of surface layer proteins vary among strains, with SlpA being the major protein band seen for all isolates, but with variability for other protein bands (Qazi et al. 2007). Notably, strain CN655 showed a slow migrating isoform that was identified by mass spectrometry as

CTC00465. Therefore, the ORF encoded in the exon upstream of *C.te*.I1 is translated at appreciable levels in CN655. The isolate CN655 was acquired and tested for SLP levels and splicing, as were clinical isolates CTHCM19, CTHCM22, and CTHCM25. Two additional patient isolated strains were acquired from the Canadian National Microbiology Laboratories, NML98A045 and NML070850. All strains were grown both in an anaerobic hood and in the EZ GasPak chamber system from BD. Use of the chamber system appeared to not compromise growth and as such growth in the chamber was continued for all subsequent studies.

Once the newly acquired strains were grown up on BHI agar plates it became apparent that the strains exhibited markedly different phenotypes (Figure 16). The patient isolated strains showed higher motility and swarming and as such only thin films were observed on the plates for CTHCM19, CTHCM 22, NML98A045, and NML070850. Strain CN655, like ATCC10779, formed discrete colonies with irregular edges. CTHCM25 however showed a unique colony morphology unlike any of the other strains which caused the identity of the strain to be in question. As such primers corresponding to the 16S rRNA were obtained and part of the 16S rRNA gene sequence was amplified from each of the strains (Figure 17). Unfortunately, of the newly acquired strains, only CTHCM19, CTHCM 22, and CN655 were determined to be *C. tetani* isolates, with the others having 16S sequences corresponding most closely to *Clostridium sporogenes*. Sequencing revealed that the strains present were not identical in 16S rRNA sequences and as such were unlikely to represent a single, common, introduced contaminant. Because identity of three of the strains was determined not to be *C. tetani*, the majority of

Figure 16. Streak plates of presumed *C. tetani* strains.

Strains were grown on 1% BHI agar plates. Although not shown, ATCC10779 exhibits similar colony morphology to CN655 (top left) and forms discrete colonies. The other strains (CTHCM 19, CTHCM 22, NML98A045, and NML070850) display swarming on 1% plates and result in a thin film of bacterial growth across the surface of the plate. Unusual colony morphology is displayed by CTCHM25 (bottom right).

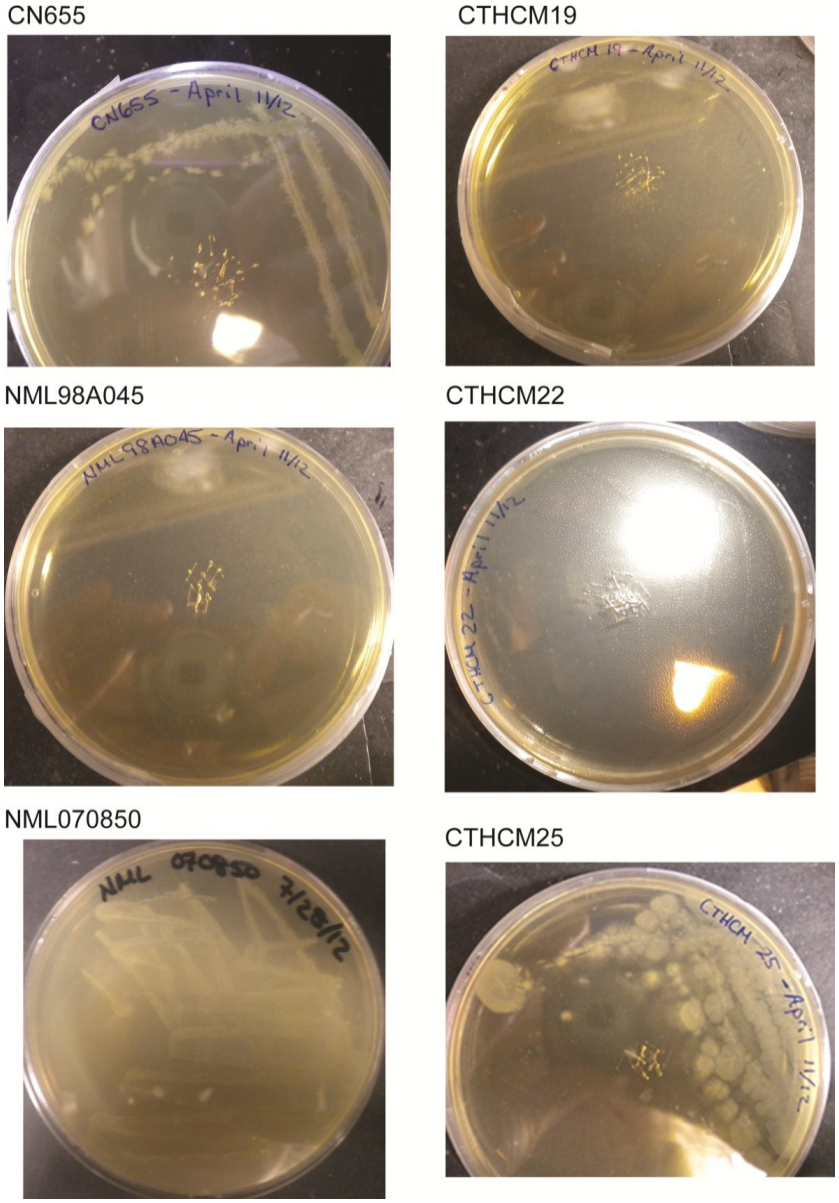
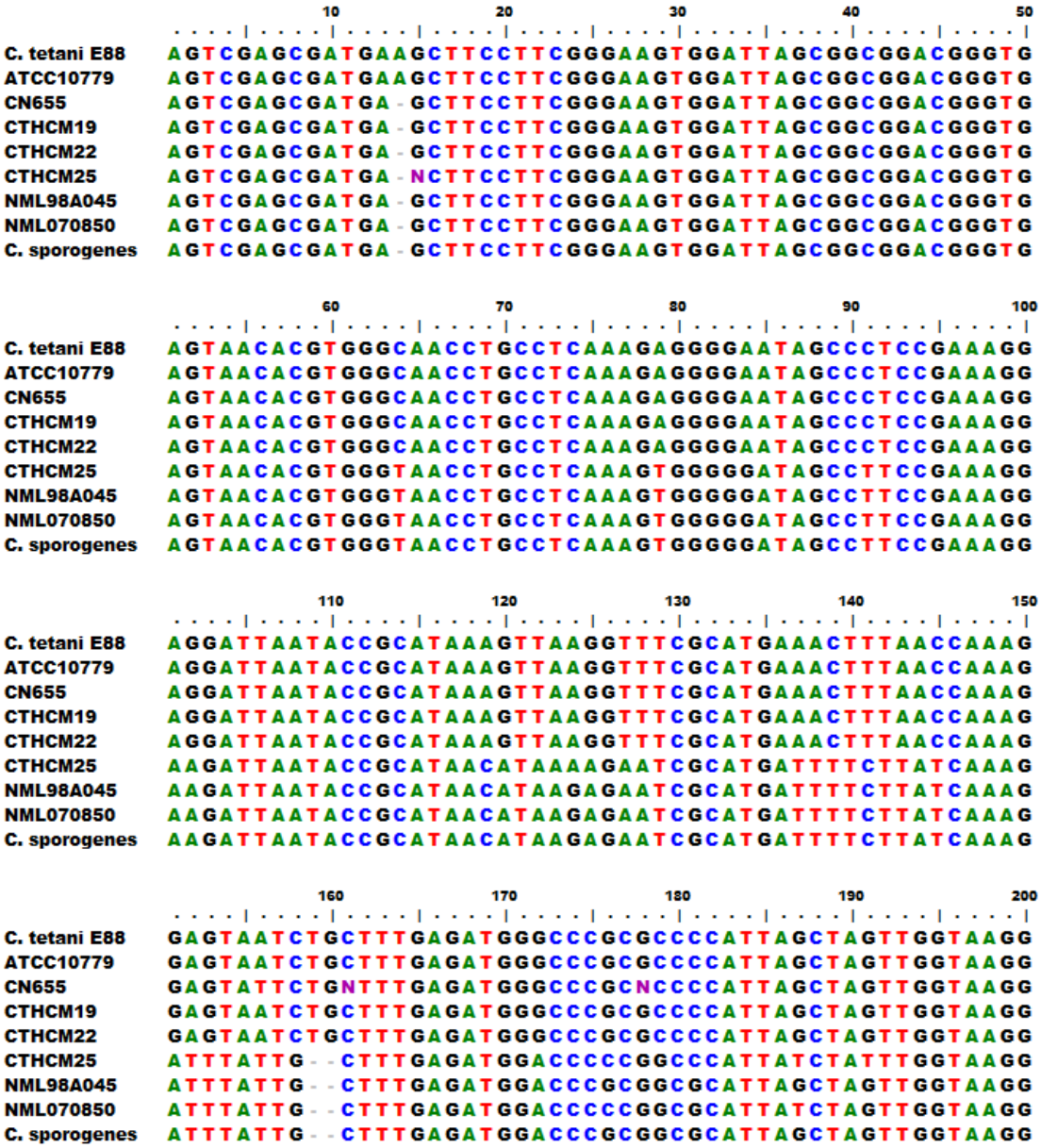


Figure 17. Partial 16S rRNA sequence alignment of cDNAs from *Clostridium* strains.

C. tetani strain CN655 shows slight variation in 16S sequence compared to the other the reference strain C. *tetani* E88 (shown at the top of the alignment). Strains CTHCM25, NML98A045, and NML070850 show conserved differences in sequences which correspond most strongly to the 16S rRNA sequence of *C. sporogenes* (shown in the last line of the alignment for reference).



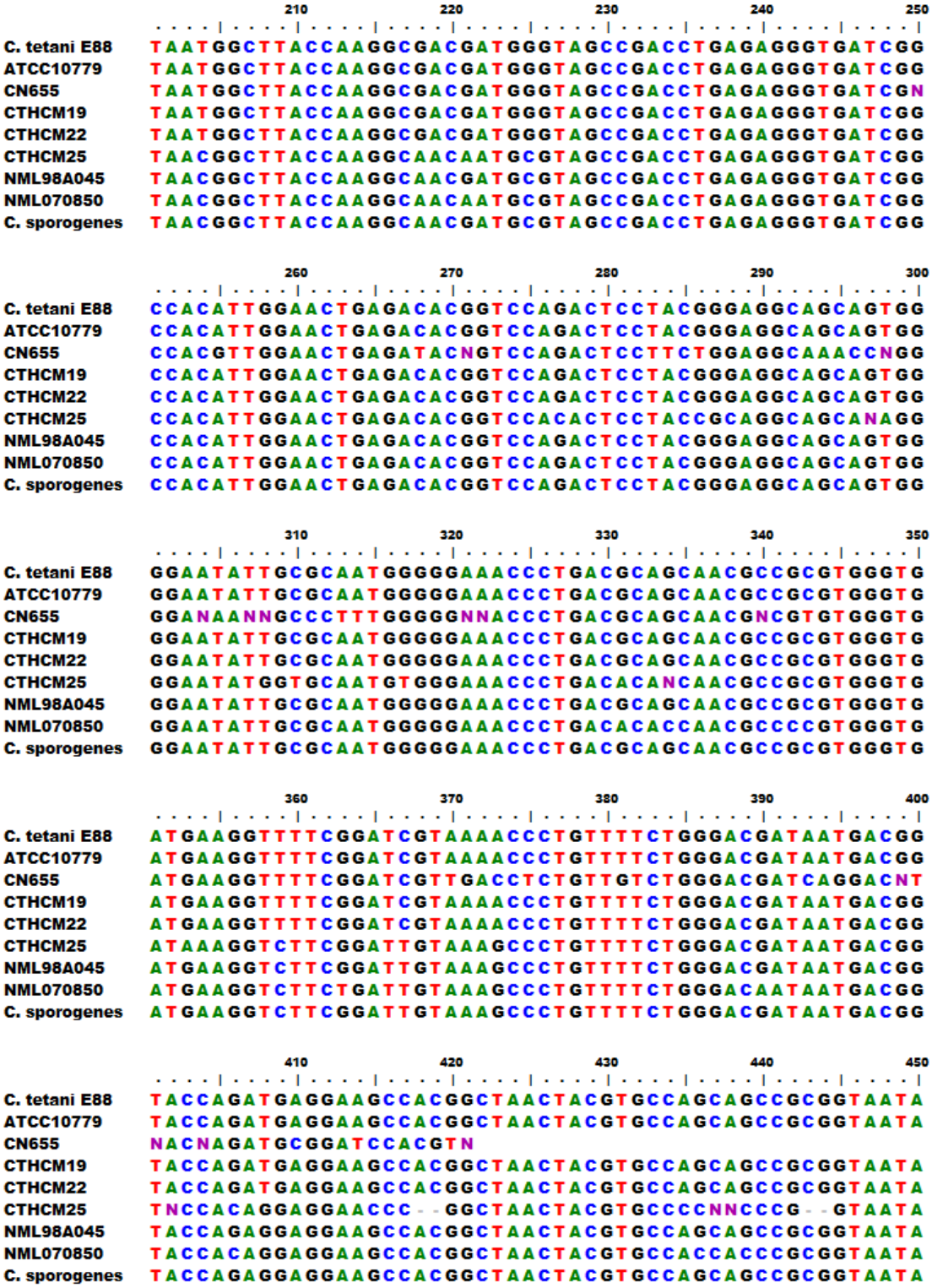
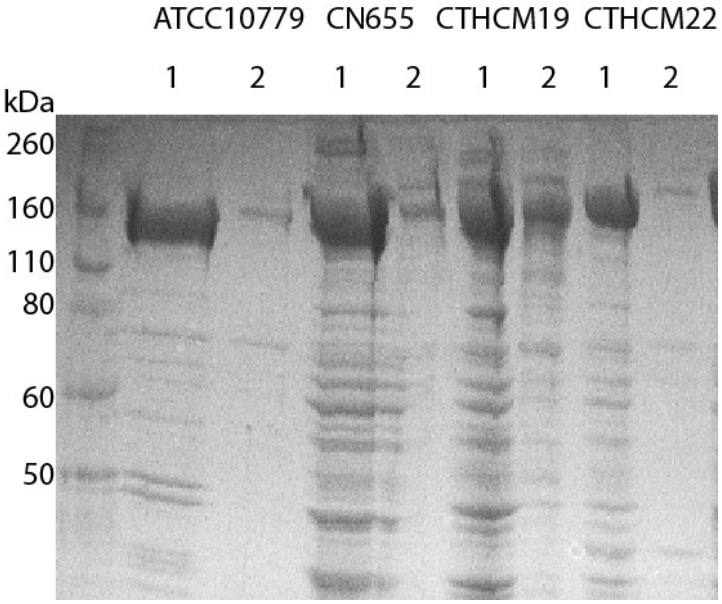

210 220 230 240 250
C. tetani E88 TAATGGCTTACCAAGGCGACGATGGGTAGCCGACCTGAGAGGGTGATCGG
ATCC10779 TAATGGCTTACCAAGGCGACGATGGGTAGCCGACCTGAGAGGGTGATCGG
CN655 TAATGGCTTACCAAGGCGACGATGGGTAGCCGACCTGAGAGGGTGATCGN
CTHCM19 TAATGGCTTACCAAGGCGACGATGGGTAGCCGACCTGAGAGGGTGATCGG
CTHCM22 TAATGGCTTACCAAGGCGACGATGGGTAGCCGACCTGAGAGGGTGATCGG
CTHCM25 TAACGGCTTACCAAGGCAACAATGCGTAGCCGACCTGAGAGGGTGATCGG
NML98A045 TAACGGCTTACCAAGGCAACGATGCGTAGCCGACCTGAGAGGGTGATCGG
NML070850 TAACGGCTTACCAAGGCAACAATGCGTAGCCGACCTGAGAGGGTGATCGG
C. sporogenes TAACGGCTTACCAAGGCAACGATGCGTAGCCGACCTGAGAGGGTGATCGG
260 270 280 290 300
C. tetani E88 CCACATTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGG
ATCC10779 CCACATTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGG
CN655 CCACGTTGGAACTGAGATACNGTCCAGACTCCTTCTGGAGGCAAACCNGG
CTHCM19 CCACATTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGG
CTHCM22 CCACATTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGG
CTHCM25 CCACATTGGAACTGAGACACGGTCCACACTCCTACCGCAGGCAGCANAGG
NML98A045 CCACATTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGG
NML070850 CCACATTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGG
C. sporogenes CCACATTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGG
310 320 330 340 350
C. tetani E88 GGAATATTGCGCAATGGGGGAAACCCTGACGCAGCAACGCCGCGTGGGTG
ATCC10779 GGAATATTGCGCAATGGGGGAAACCCTGACGCAGCAACGCCGCGTGGGTG
CN655 GGANAANNGCCCTTTGGGGGNNACCCTGACGCAGCAACGNCGTGTGGGTG
CTHCM19 GGAATATTGCGCAATGGGGGAAACCCTGACGCAGCAACGCCGCGTGGGTG
CTHCM22 GGAATATTGCGCAATGGGGGAAACCCTGACGCAGCAACGCCGCGTGGGTG
CTHCM25 GGAATATGGTGCAATGTGGGAAACCCTGACACANCAACGCCGCGTGGGTG
NML98A045 GGAATATTGCGCAATGGGGGAAACCCTGACGCAGCAACGCCGCGTGGGTG
NML070850 GGAATATTGCGCAATGGGGGAAACCCTGACACACCAACGCCCCGTGGGTG
C. sporogenes GGAATATTGCGCAATGGGGGAAACCCTGACGCAGCAACGCCGCGTGGGTG
360 370 380 390 400
C. tetani E88 ATGAAGGTTTTCGGATCGTAAAACCCTGTTTTCTGGGACGATAATGACGG
ATCC10779 ATGAAGGTTTTCGGATCGTAAAACCCTGTTTTCTGGGACGATAATGACGG
CN655 ATGAAGGTTTTCGGATCGTTGACCTCTGTTGTCTGGGACGATCAGGACNT
CTHCM19 ATGAAGGTTTTCGGATCGTAAAACCCTGTTTTCTGGGACGATAATGACGG
CTHCM22 ATGAAGGTTTTCGGATCGTAAAACCCTGTTTTCTGGGACGATAATGACGG
CTHCM25 ATAAAGGTCTTCGGATTGTAAAGCCCTGTTTTCTGGGACGATAATGACGG
NML98A045 ATGAAGGTCTTCGGATTGTAAAGCCCTGTTTTCTGGGACGATAATGACGG
NML070850 ATGAAGGTCTTCTGATTGTAAAGCCCTGTTTTCTGGGACAATAATGACGG
C. sporogenes ATGAAGGTCTTCGGATTGTAAAGCCCTGTTTTCTGGGACGATAATGACGG
410 420 430 440 450
C. tetani E88 TACCAGATGAGGAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATA
ATCC10779 TACCAGATGAGGAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATA
CN655 NACNAGATGCGGATCCACGTN
CTHCM19 TACCAGATGAGGAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATA
CTHCM22 TACCAGATGAGGAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATA
CTHCM25 TNCCACAGGAGGAACCC--GGCTAACTACGTGCCCCNNCCCG--GTAATA
NML98A045 TACCAGAGGAGGAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATA
NML070850 TACCACAGGAGGAAGCCACGGCTAACTACGTGCCACCACCCGCGGTAATA
C. sporogenes TACCAGAGGAGGAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATA

Figure 18. Comparison of SLP extraction using 4 M urea and 0.2 M glycine (pH 2.2).

SLP extractions were performed for all *C. tetani* strains. Identical amounts of each extraction were run on a 10% SDS-PAGE and visualized through Coomassie staining. Lanes marked with a 1 are extraction with 4 M urea. Lanes marked with a 2 were extracted using 0.2 M glycine (pH 2.2).



experiments were conducted with only the four strains ATCC10779, CN655, CTHCM19 and CTHCM22.

Using these strains, SLPs were prepared using both extraction with 4 M urea and 0.2 M glycine pH 2.2 (Figure 18). The comparison of these two extraction conditions determined that 4 M urea was the most effective method at removing the SLP, consistent with what was previously determined by Qazi et al. (2007). Only 4 M urea was used for extraction of the surface layer in further experiments. Surface layer protein preparations for ATCC10779 and CTHCM22 showed a 160 kDa band as the largest protein species, whereas CN655 and CTHCM19 showed other higher molecular weight protein bands that could potentially correspond to alternative splicing forms of the SLPs. The bands were excised from Coomassie stained gels (Figure 19) and subjected to MALDI-ToF MS and peptide fingerprinting, which revealed that CN655 produces a SLP migrating at approximately 260 kDa that has identity to CTC00465. In contrast, the higher molecular weight protein band in CTHCM19 showed identity only to CTC00462 (Table 6). This result was not completely unexpected as SLPs often have unusual migrations in SDS-PAGE gels. This could be the result of post-translational modifications such as protein crosslinking or other modifications such as glycosylation. Glycosylation was analyzed through glycan staining and mass spectrometry by Qazi et al. (2007), but no evidence for glycosylation was observed.

As the protease functions of the appended ORFs could produce lower molecular weight products other than those described above, as is the case for the *C. difficile* SlpA protein (Kirby et al. 2009), the 2 most intense of the smaller bands observed were also

Figure 19. Bands extracted for MALDI-ToF MS.

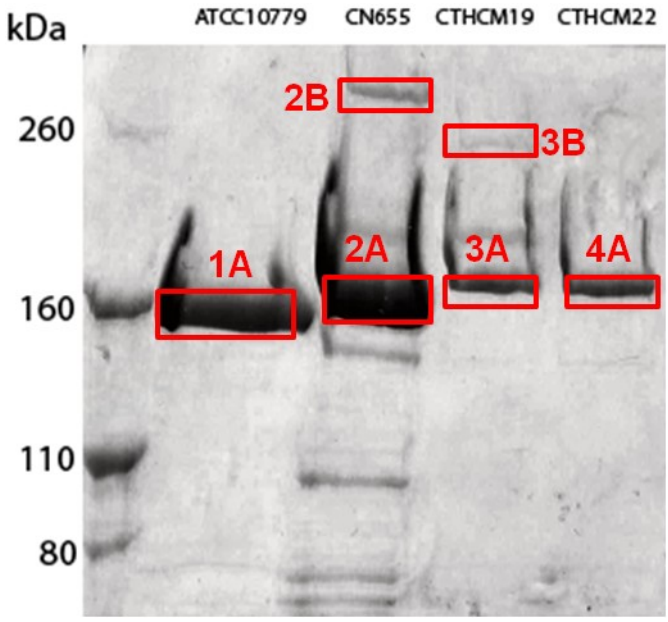


Table 6. Results of MALDI-ToF MS and peptide fingerprinting

Band[a]	**Strain**[b]	**Peptide Fingerprint ID**[c]	**SCORE**[d]	**Coverage (%)**
1A	ATCC10779	CTC00462	113	50
2A	CN655	CTC00462	113	50
2B	CN655	CTC00465	108	56
3A	CTHCM19	CTC00462	83	23
3B	CTHCM19	CTC00462	100	69
4A	CTCHM22	CTC00462	143	56

[a]Band labelled in Figure 19
[b]Strain of isolate of *Clostridium tetani*
[c]Identified ORF sequence based on searches of the NCBI databases
[d]SCORE determined using
http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=PMF
Score thresholds are calculated based on the individual searches with $p<0.05$. Score threshold for these searches is 77 and as such, scores above 77 are significant.

subjected to MALDI-ToF MS. These intense lower bands (<160kDa) were found not to correspond to surface layer proteins but instead to other proteins secreted from the cell. The higher of these bands corresponded to a pyruvate-flavodoxin oxidoreductase and the lower of the bands to elongation factor G (data not shown).

Due to differences in the methodology, LC-MS/MS based peptide sequencing has higher sensitivity and reliability than MADLI-ToF MS for protein identification. This is especially true if a gel slice contains two or more proteins. Although database searches will still generate hits, the reliability of protein identification based on MALDI-ToF MS peptide fingerprinting will become problematic as identification of various proteins is based solely on the mass-to-charge (*m/z*) ratio of the peptides. In contrast, LC-MS/MS is able to analyze very complex protein mixtures since each peptide is independently sequenced through an additional fragmentation step [reviewed in (Cho 2007)]. From the perspective of the NCBI database, the alternatively spliced proteins in question would be viewed as multiple proteins within the same sample and identification of 3′ exon sequences therefore might be more successful using LC-MS/MS. As such the high molecular weight band from CN655 (Figure 19; 2B) was submitted for LC-MS/MS (SAMS institute; University of Calgary). LC-MS/MS identified two potential proteins present in the upper band; however the additional protein identified did not correspond to one of the four potential 3′ exon sequences (Table 7). CTC00465 (our 5′ exon of interest) was identified along with the major SLP band, SlpA (CTC00462). While both MALDI-ToF MS and LC-MS/MS failed to detected alternatively spliced forms of CTC00465 there still may be trace amounts of the alternatively spliced forms expressed by the

Table 7. Results of LC MS-MS protein ID for CN655 high molecular weight band

Accession #	Score[a]	Matches[b]	Sequences[c]	Description
gi|28210215	642	31	15	S-layer protein (CTC00465) [*Clostridium tetani* E88]
gi|28210213	530	20	12	S-layer protein/N-acetylmuramoyl-L-alanine amidase (CTC00462) [*Clostridium tetani* E88]

[a] SCORE determined using
http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=PMF
Score thresholds are calculated based on the individual searches with $p<0.05$.
Scores above 49 are significant
[b]Number of significant peptide matches
[c]Number of significant distinct sequences identified

various *C. tetani* strains. These may be expressed at levels that are not detectable by Coomassie staining or are not secreted by the cell. Alternatively they may be cleaved by proteases and may be represented by one of the more minor bands found below the major 160 kDa band corresponding to SlpA.

As CTC00465 is not detectably alternatively spliced in the CN655 strain that expresses it, it remained curious that the protein ran at such an unusually high position on a denaturing protein gel (>260 kDa) when the expected size of the unspliced protein would run at approximately 145 kDa. As previously stated, this could be the result of post-translational modifications such as glycosylation or could be due to protein crosslinking. Although glycosylation was analyzed through glycan staining and mass spectrometry by Qazi et al. (2007), no evidence for glycosylation was observed. The production of glycosylated surface layer proteins is, however, known to be important in a number of Gram positive pathogens including *Mycobacterium* and *Streptococcus* species (Erickson and Herzberg 1993; Stimson et al. 1995; Dobos et al. 1996). In addition, within the domain Eubacteria glycosylated SLPs have only been demonstrated for low G+C Gram positive organisms (Schäffer and Messner 2001); thus, *Clostridium tetani* would seem a good candidate for such modifications. Therefore, I decided to independently investigate whether or not glycosylation of the CTC00465 SLP, rather than alternative splicing, could explain its unusual position on denaturing gels.

Standard methods to investigate glycosylation involve both chemical and enzymatic methods of deglycosylation, as well as the use of glycan stains. Enzymatic deglycosylation ideally requires knowledge of the type of glycosylation present, as enzymes are specific to the glycan linkage type. While PNGaseF is effective for removal

of virtually all N-linked glycans, most bacterial S-layers contain *O*-glycosidic linkages (Schäffer and Messner 2001). In the case of O-linked glycans, monosaccharides must be removed by a series of exoglycosidases until only the glycan core remains linked to the serine or threonine moieties. *O*-Glycosidase can then remove the core structures only if there are no modifications. Modifications of the core structures require the use of additional enzymes, and many of the enzymes only work on common O-linked glycans. Trying each enzyme would prove costly and inefficient — perhaps ineffective — particularly since previous studies have not resulted in evidence that the surface layer proteins of *C. tetani* are glycosylated and negative data could be expected.

On the other hand, chemical methods of deglycosylation are non-specific, and capable of removing the entire glycan complement. While certain methods of chemical deglycosylation result in the complete degradation of the protein an alternate procedure using trifluoromethanesulfonic acid (TFMS) results in minimal degradation of the protein (Edge et al. 1981; Edge 2003). The reaction, however, requires lyophilization of the protein prior to deglycosylation and for the deglycosylation reaction to be performed anhydrously. While this method of glycan removal was attempted, re-solubilizing the proteins following TMFS treatment proved problematic and interpreting the resultant SDS-PAGE was impossible.

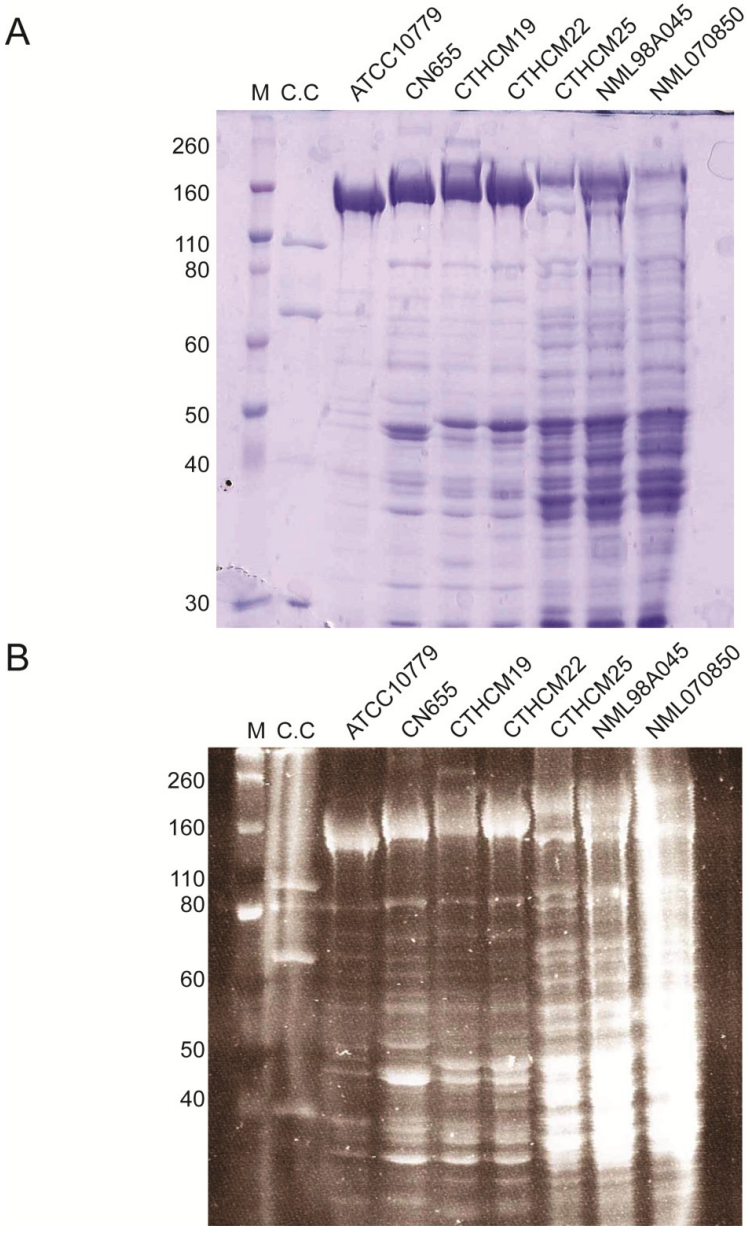
As an alternative to the removal of glycans, a glycan stain was utilized. The most common method of detection of glycoproteins in polyacrylamide gels or electroblots is by periodic acids/Schiff (PAS) staining (Packer et al. 1997). The Qazi et al. (2007) paper utilized a digoxigenin glycan-labelling kit. I chose to utilize the Pro-Q Emerald 300 Glycoprotein Gel and Blot Stain Kit (Life Technologies) which is touted as having

increased sensitivity compared to PAS staining. This approach utilizes a fluorescent hydrazide dye that is linked to the glycoprotein using a standard PAS conjugation mechanism. Periodic acid is utilized to oxidize the glycols to aldehydes and then the hydrazide dye reacts with the aldehydes to form a fluorescent conjugate (Steinberg et al. 2001). The Pro Q Emerald 300 kit is supplied with Candy Cane molecular weight standards (Life Technologies), which consist of a mixture of glycosylated and non-glycosylated proteins intended to act as both positive and negative controls for the glycan stain. As the dye has an excitation maximum at 280 nm and an emission maximum at approximately 530 nm, the gels can be visualized using a 300 nm transilluminator and glycoproteins should then be indicated by green fluorescence. The Pro-Q Emerald 300 stain, however, showed fluorescence for all proteins including the negative controls within the candy cane ladder (Figure 20).

It is known that the aromatic rings of tryptophan (Trp), tyrosine (Tyr) and phenylalanine (Phe) can act as intrinsic fluorophores and that this property can be exploited to visualize protein bands following SDS-PAGE (Roegener et al. 2003); however, this intrinsic protein fluorescence has also been found to interfere with protein detection by extrinsic fluorescent dyes. A study looking at the detection of glycoproteins in tear samples using the Pro-Q Emerald 300 glycoprotein stain, noted that lysozyme showed fluorescence both in the tear samples, as well as in the candy cane ladder. This was unexpected as lysozyme is not a glycoprotein (Zhao et al. 2007). The authors show that incubation in fix solution alone (50% methanol, 5% acetic acid) according to manufacturer's instructions (Invitrogen, Eugene, OR) was enough to induce fluorescence. I experienced similar problems with the detection of glycoproteins using this stain.

Figure 20. Pro Q Emerald 300 glycosylation analysis.

(A) Coomassie stain following staining with Pro Q Emerald 300 to show total protein (B) glycan staining with Pro Q Emerald 300. "C.C" indicates the presence of the candy cane ladder. Lane M contains the Novex pre-stained high molecular weight marker. Sizes corresponding to positions in lane M are shown to the left of the gel in kDa.



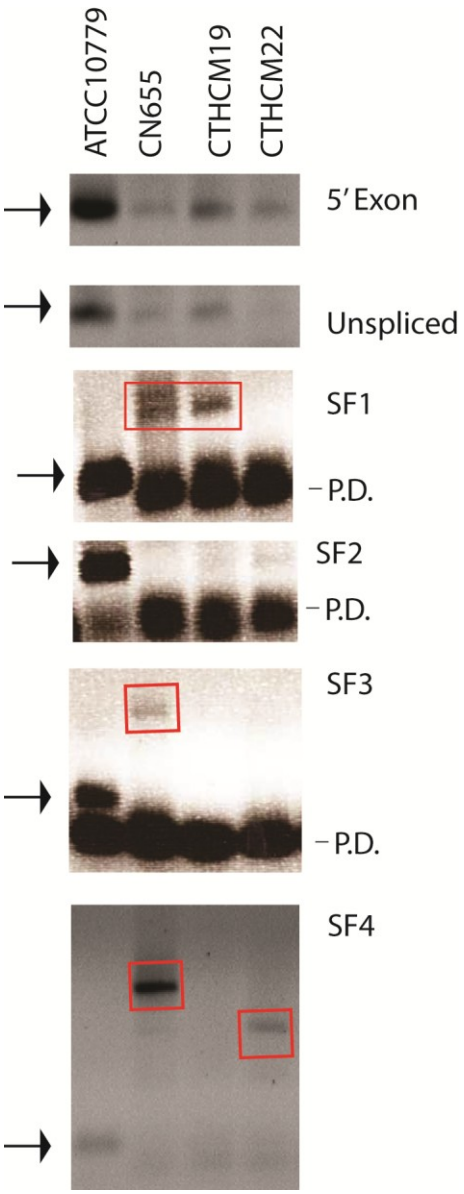
Looking at the Candy Cane ladder in Figure 20A (Lane C.C) definitive bands can be seen at 97, 66, 42 and 29 kDa in the Coomassie stained gel. Following staining with Pro Q Emerald 300 (Figure 20B) all of the above mentioned bands plus an 82 kDa band are identifiable despite only the 82 kDa (glucose oxidase) and the 42 kDa (α_1- acid glycoprotein) band being glycoproteins. As such, positive and negative controls were not established and I was unable to determine the glycosylation state of the SLPs extracted from *C. tetani*.

2.3.7 Strain-Specific RT-PCR

Although SLP extractions failed to show evidence at the protein level for alternative splicing in the new strains of *C. tetani*, RNA was extracted from each of the various strains under standard growth conditions to look for possible variation in expression of splice forms between stains. The 5′ exon control and the unspliced intron were detectable via qRT-PCR, although at low levels in the additional strains. The products were visualized on an agarose gel (Figure 21), the bands were subsequently gel extracted and the PCR product was sent directly for sequencing. Sequencing confirmed that a small 167 nt region at the CTC00465-*C.te.*I1 junction, corresponding to 132 nt of exon CTC00465 and the first 35 nt of *C.te.*I1, is present in each of the confirmed *C. tetani* stains (ATCC10779, CN655, CTHCM19 and CTHCM22). RT-PCR for the other splice forms either did not result in amplification or did not show the expected amplification products (Figure 21). The PCR products indicated by the boxed bands in Figure 21 were gel extracted and sent for sequencing. The results were uninformative as reactions consistently failed to produce sequence and when sequencing results were

Figure 21. Strain specific RT-PCR.

Panels are shown for amplification of 5′ Exon, Unspliced intron, and splice forms 1-4 (SF1-4). The location of the expected products based on E88 sequence and that are observed in ATCC10779 are indicated with arrows. Unexpected products are shown with red boxes. Bands lower than the expected product seen in some lanes are primer dimers resulting from inefficient strain specific amplification and are indicated by the letters P.D.



available they failed to produce HSPs via blastn searches. Notably, the sequence was reliably determined for the ATCC10779 positive control sequences.

The failure of the strain-specific RT-PCR to detect the various splice-forms in each of the strains may indicate that alternative splicing does not occur in these strains or that the genomic organization of the region differs between the sequenced strain E88 and the strains CN655, CTHCM19 and CTHCM22. Sequencing of the alternative splicing locus will be the most informative way to determine if genomic differences exist between strains in this potentially variable region. However, PCR amplifications in the region that were successful for ATCC10779 DNA have not been successful using genomic DNA from the other *C. tetani* strains.

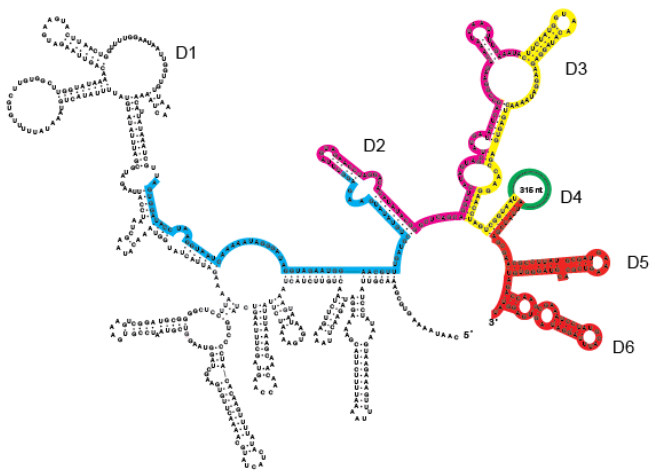
2.3.8 Evidence for the Formation of the Locus in C. tetani *and Other Prokaryotes*

Formation of the *C.te*.I1 intron genomic locus is most likely the result of a conventional mobile, bacterial Class B group II intron, whose homing site would have been in the vicinity of the surface layer genes. The location of the predicted homing site is supported by the bioinformatically identified remnants of *C.te*.I1-related sequences found among SLP-related genes of other firmicutes (Figure 22). A Blastn search using the nucleotide sequence of the *C.te*.I1 intron, including D4, was performed both with standard databases and using the Whole Genome Shotgun (WGS) database through NCBI. Top hits from this blast search revealed remnants of the intron sequence flanked by various surface layer proteins or surface layer protein related sequences. As D5 is highly conserved among all group II introns, hits corresponding to D5 sequence were

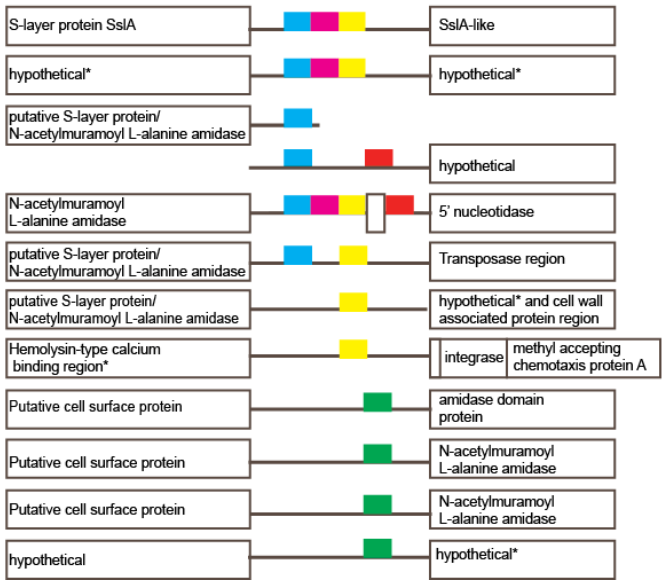
Figure 22. Ribozyme-derived sequences among SLP genes in various Firmicutes.

(A) Secondary structure of *C.te.*I1, highlighting the regions corresponding to BLASTN matches in other species. (B) Colored boxes indicate BLASTN matches between the corresponding region of *C.te*.I1 in Panel A and the indicated genomic sequence (matches are not shown to scale).Hits corresponding to D5/6 were included only if hits of D1-4 were also observed. The match length, E-value, and % identity for each high-scoring pair (HSP) is shown to the right. The apparent sequence remnants of group II introns in intergenic regions of surface layer proteins of Firmicutes suggest that the ancestor of *C.te.*I1 had a homing site in a surface layer-coding region of the genome.

A



B



Species	Match Length (nt)	E-value	Identity
Sporosarcina uraeae ATCC13881	217	1.00 e -14	70%
Exiguobacterium sp. AT1b	213	1.00 e -8	69%
Bacillus spaericus 2363 (partial genome)	70	2.00 e -5	81%
Brevibacillus sp. BC25 (WGS contig)	70, 37	1.00 e -1, 8.00 e -3	76%, 92%
Bacillus sp. B14905 (WGS contig)	219, 49	7.00 e -4,1.00 e -1	66%, 83%
Lysinibacillus spaericus C3-41	70, 63	2.00 e -5, 4.20 e -1	81%, 77%
Lysinibacillus spaericus C3-41	63	5.1 e 0	77%
Lysinibacillus spaericus C3-41	49	1.5 e -1	82%
Clostiridium botulinum B1 strain Okra	66	3.00 e -3	79%
Clostiridium botulinum F strain 230613	90	2.00 e -5	74%
Clostiridium botulinum F strain Langeland	90	2.00 e -5	74%
Clostiridium botulinum D strain 1873 (WGS contig)	74	3.00 e -9	83%

omitted from analysis unless also accompanied by sequence matches within D1-4. In these sequences in other genomes, the ancestral mobile intron of *C.te.*I1 presumably inserted and degenerated, leaving behind only partial segments of the intron structure. In contrast, when the intron inserted in *C. tetani* E88 selective pressures acted on the intron resulting in the loss of its IEP, RNA structural adaptations, and duplication events that led to the presence of 3 alternate D5/6 segments downstream. It is also possible that the additional domain 5/6 copy corresponded to group II intron insertions rather than duplication events. As E88 is the only sequenced *C. tetani* strain it is unclear how widespread the *C.te.*I1 alternative splicing locus is even within *C. tetani* strains.

2.4 Discussion

The discovery of the *C.te*.I1 intron marks the first known case of alternative splicing in Eubacteria that truly resembles eukaryotic alternative splicing, in that multiple unique RNA isoforms are generated from a selection of alternate exons. Five variant surface layer RNAs are produced from the annotated SLP gene CTC00465 with the most abundant RNA corresponding to the annotated gene and a fraction of products (5-25%) corresponding to the upstream ORF (CTC00465) fused correctly in-frame with one of four downstream-encoded domains (CTC00467, CTC00468, CTC00469 or CTC00470). The location of the *C.te*.I1 intron in a surface layer protein region suggests that although physiological relevance has yet to be established, alternative splicing of this locus is likely to be beneficial to the cell.

2.4.1 Alternative Splicing and Intron-Based Gene Regulation

The only other example of alternative splicing of a group II intron in bacteria is for the *B.a.*I2 intron of *Bacillus anthracis* that utilizes two 3′ splice sites located 4 nt apart, thus producing two protein products. The use of the upstream splice site produces a protein comprised of the 5′ ORF alone, and accounts for the majority of transcript and splicing products. The downstream splice site is used during 4% of splicing reactions and it fuses upstream and downstream ORFs to form a two-domain protein (Robart et al. 2004); however, the functions of both of the exons involved in this splicing event are unknown.

While *B.a.*I2 represents the only case of alternative splicing known in bacteria prior to *C.te.*I1, other group II and group III introns (streamlined versions of group II introns) found in *Euglena gracilis* chloroplasts have been reported to alternatively splice (Copertino and Hallick 1993; Jenkins et al. 1995). The use of alternate 5′ and 3′ splice sites occurred in a small proportion of transcripts and resulted in the production of truncated ORFs or small 2-4 amino acid insertions. Translated proteins corresponding to these products were not identified and these alternative splicing reactions are thought to represent aberrant splicing tolerated by the cell rather than beneficial reactions. *C.te.*I1 alternative splicing differs from all previously reported cases of group II intron alternative splicing as its location in a SLP region and the production of 5 distinct RNA forms likely indicate that this case of alternative splicing is beneficial to the organism.

An interesting example of bacterial genetic regulation through splicing has been found in *C. difficile* that involves a group I intron whose splicing is controlled by a 84 nt, cyclic-di-GMP-sensing, riboswitch (Lee et al. 2010). In the absence of c-di-GMP the

formation of an anti-5′splice-site stem is favored and GTP attack — the first step of group I intron splicing — occurs at an alternate site. Alternatively, binding of c-di-GMP results in splicing at the expected 5′ splice site and inclusion of the ribosome binding site in the 5′ UTR of the transcript. As this allosteric ribozyme is located in a putative virulence gene in *C. difficile*, c-di-GMP mediated splicing results in the subsequent translation of the virulence factor. As such, the splicing of this group I intron provides translational control.

While the group I intron in *C. difficile* does not result in alternative splicing, it shares evolutionary parallels with *C.te*.I1. First, both introns have been domesticated by *Clostridium* species for regulation of gene expression in regions that are linked to virulence. Secondly, both of the introns have lost their respective mobility-associated proteins. Loss of mobility is likely key in the domestication of both "selfish" genetic elements.

2.4.2 Regulation of Alternative Splicing

Although unspliced transcript was the most common transcript found corresponding to the CTC00465 locus and regulated alternative splicing was not detected from *C.te.*I1, it remains possible that regulated splicing may occur in some strains, in response to conditions other than those tested (such as iron levels or pH), or upon infection. In addition to the possibility that splicing is regulated, the possibilities of constitutive or stochastic expression must also be considered.

Stochastic variation in splice site selection has been proposed as the mechanism to generate the large variety of *dscam1* molecules that provide self-identity to neurons in *Drosophila melanogaster* (Neves et al. 2004). Although expression of isoforms differ

between cell types, each individual cell expresses 14-50 mRNA isoforms of the *dscam1* gene. This provides a diverse pool of molecules that can be further regulated through translation and degradation mechanisms. Although on a smaller scale, such a scenario could be envisioned for the *C. tetani* alternative splicing locus as well. Random variation in the usage of splice sites, perhaps caused by variation in the concentrations of splicing factors or RNAs (if the region is expressed monocistronically – see section below on *trans*-splicing), could produce a pool of mRNAs which may be subsequently regulated through processes such as RNA degradation and translation.

Stochastic variability or cell-specific constitutive splicing patterns, could be envisioned to result in a fraction of cells within a population expressing each of the different variants in a manner similar to the phase variable expression of the *C. difficile* cell wall protein, CwpV (Emerson et al. 2009). Expression of different surface layer proteins by different cells within a clonal population allows for diversity and would confer a selective advantage to the population upon infection.

2.4.3 *Alternative Splicing of* C.te.I1 *Most Likely Occurs in* trans

Mechanistically it is unknown whether the alternative splicing shown for the *C. tetani* locus occurs through *cis*- or *trans*-splicing. In eukaryotes, alternative splicing is made possible because of relatively short recognition sequences for both 5′ and 3′ splice sites, which are acted upon by the *trans*-acting spliceosome. Alternative splicing is guided by splicing factors that either enhance or inhibit usage of the different 5′ and 3′ recognition sequences [reviewed in (Breitbart et al. 1987; Matlin et al. 2005)]. The use of the *trans*-acting machinery of the spliceosome allows the splice sites used to be far away from each other. Typically group II introns cannot readily undergo alternative splicing,

because the exon junctions are located directly adjacent to a complex RNA structure that catalyzes splicing.

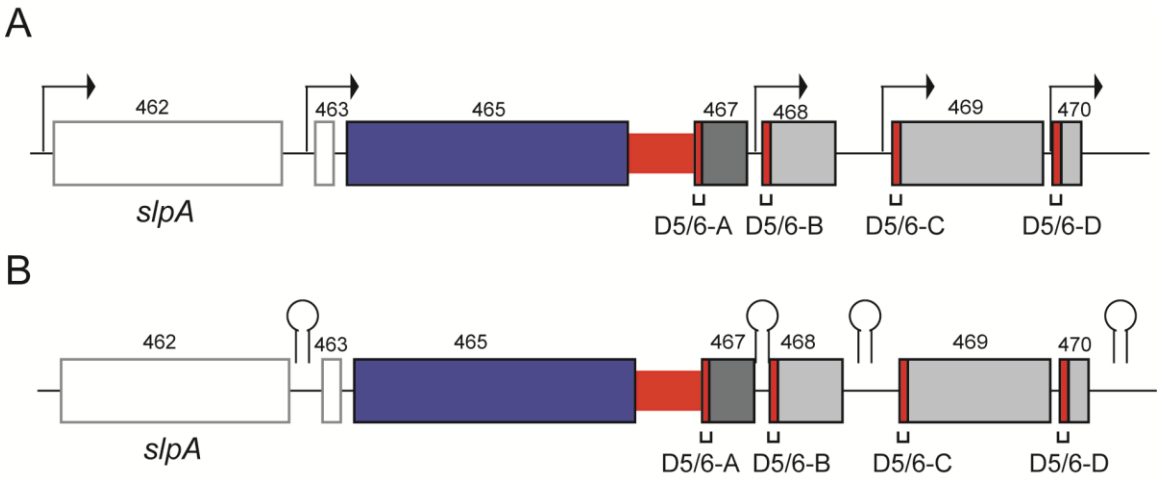
A *cis*-splicing mechanism is difficult to imagine but not theoretically impossible. In order for *cis*-splicing to occur using the downstream splice sites the additional sequence must either be looped out, similar to D4 sequence, or be incorporated into the ribozyme structure. As the alternative D5/6 sequences are kilobases apart, only the potential looping out of sequence through D4 seems even remotely plausible; however, it seems likely that large regions of additional sequence would interfere with folding and potentially inhibit splicing.

It is more likely that alternative splicing of *C.te.I1* is based on a *trans*-splicing mechanism. The reasoning for this is two-fold. First, the additional D5/6 sequences can be encoded more than 5 kb away from the 5′ splice site of the *C.te.I1* intron and folding and splicing these introns as a single transcript is questionable in its feasibility. Secondly, bioinformatic prediction of bacterial promoter sequences (using BPROM) and transcriptional terminators (using ARNold) suggest that these exons are encoded monocistronically rather than as a polycistronic transcript (Gautheret and Lambert 2001; Macke et al. 2001; Solovyev and Salamov 2011) (Figure 23).

In mitochondrial and chloroplast genomes, there are many examples of *trans*-splicing group II introns, in which the intron structure is encoded in two sites, and two transcripts of RNA assemble to splice *in trans* (Bonen 1993; Bonen 2008). In nearly all

Figure 23. Predicted promoter and transcriptional terminator sequences.

Approximate locations of (A) BPROM predicted promoter sequences and (B) transcriptional terminators predicted by ARNold are shown relative to the exon and intron sequences present in the region. These data suggest that the region is transcribed as individual ORFs with the exception of 463, 465 and 467 which could be transcribed together.



cases the fragmentation point in the *trans*-splicing introns is in domain 4 (Kohchi et al. 1988; Goldschmidt-Clermont et al. 1991). In the case of the *C.te.*I1 intron, fragmentation would also occur in D4. Although natural *trans*-splicing has not been reported, bacterial group II introns are capable of *trans*-splicing (Belhocine et al. 2007; 2008) and it seems most probable that *trans*-splicing is the mechanism utilized by *C.te*.I1 to carry out alternative splicing.

2.4.4 Variation in SLP Expression

As the 5′ exon sequence for *C.te*.I1 is clearly annotated as a surface layer protein and has been experimentally validated as such (Qazi et al. 2007), the possibility of alternative splicing of this intron was especially intriguing. Surface layers are found as the outermost layer on the surface of virtually all archaeabacteria and most eubacteria (Sleytr et al. 1996; Sleytr and Beveridge 1999). As such, these proteins, or other associated proteins, are actively involved in establishing infection and avoiding the host immune response in pathogens (Pavkov-Keller et al. 2011). Variation in these proteins is often essential and as such many mechanisms have been described that function to create diversity within SLPs (Dworkin and Blaser 1997; Egelseer et al. 2001; Scholz et al. 2001; Emerson et al. 2009). In pathogenic bacteria, S-layer variation is a strategy that may be employed to escape an effective immune response (Munn et al. 1982; Blaser et al. 1988; Wei-Mei et al. 1992). Alternative splicing in this region therefore has potential to represent a novel mechanism of gene regulation and a novel mechanism to generate diversity of surface layer proteins, potentially increasing the fitness of the microorganism.

Typically, bacteria encode a variety of SLP genes within their genomes, but only a single major protein or glycoprotein is produced and secreted; however, some organisms are known to produce complex S-layers consisting of more than one repetitive subunit. An example is the *C. difficile* S-layer, which is composed of a High Molecular Weight (HMW) and a Low Molecular Weight (LMW) component. These components are produced by the catalytic cleavage of the precursor protein SlpA by the cysteine protease Cwp84 (Kirby et al. 2009). Interestingly the SlpA protein of *C. difficile* shows identity to the main S-layer component of the *C. tetani* S-layer, CTC00462 (Qazi et al. 2007). This is interesting as the composition of S-layers and the amino acid sequences of SLPs are highly variable, both across and within species. As such, CTC00462 has been referred to as SlpA for consistency in naming purposes. The *C. tetani* protein, however, is not known to undergo equivalent cleavage reactions.

S-layer production can also be altered by various strains within a species or in response to different environmental conditions, resulting is variation and providing selective advantages to certains strains or subpopulations within a culture. The Gram positive micro-organism, *Geobacillus stearothermophilus* displays strain specific expression of a variety of surface layer proteins. In addition to the strain specific expression [reviewed in (Pavkov-Keller et al. 2011)], certain changes in SLP expression can be observed due to changes in growth conditions. The wild-type strain PV72/p6 expresses SbsA (Kuen et al. 1994) but growth under elevated oxygen conditions creates a strain variant PV72/p2 that expresses SbsB (Scholz et al. 2001). Interestingly, the oxygen-induced SLP switch is due to a DNA rearrangement between chromosomal and plasmid loci that encode the two SLP genes (Scholz et al. 2001). When *sbsB* is encoded

on the plasmid it is transcriptionally inactive, but becomes active upon integration into the chromosome as the upstream regulatory elements of *sbsA* are not involved in the rearrangement. The strain ATCC 12980 expresses another SLP, SbsC, however growth at elevated temperatures (67°C) causes the switch to expressing a glycosylated SLP, SbsD (Egelseer et al. 2001).

DNA inversions have similarly been shown to control SLP or cell wall gene expression in *Campylobacter fetus* and *C. difficile* (Dworkin and Blaser 1997; Emerson et al. 2009). In some of these cases the antigenic variability has been demonstrated. Overall, the composition of S-layers and the amino acid sequences of SLPs are highly variable, both across and within species. In addition to expressing an S-layer that comprises two SLPs, *C. difficile* also creates diversity in a secondary cell wall protein, CwpV. An inversion that results in the production of CwpV occurs in approximately 5% of cells in a population (Emerson et al. 2009) and results in the phase variable expression of that protein. The antigenic variation of the S-layer displayed by *Campylobacter fetus* also involves a DNA inversion of SLP gene sequences with expression controlled by a single promoter (Dworkin and Blaser 1997).

Although evidence of S-layer variation due to alternative splicing was not observed at the protein level, it does not indicate that variation does not occur. It seems unlikely that the development of this mechanism is simply coincidence and more likely that the specific strains exhibiting these switches or conditions that result in these switches simply have yet to be found. Thus, alternative splicing in the surface layer protein region of *C. tetani* ultimately appears to represent yet another mechanism utilized to generate variability in S-layer and cell wall proteins.

2.4.5 Evolutionary Significance

The origin of alternative splicing is not certain; in some cases, alternatively spliced introns appear to have arisen recently, e.g., in the primate lineage (Zhuo et al. 2007), whereas other alternative splicing organizations are inferred to be ancient, dating back to the unicellular ancestor of plants, animals and fungi (Irimia et al. 2007). Although alternative splicing events are quite rare in lower eukaryotes and consist primarily of intron retention events, alternative splicing has been shown in lower metazoans, fungi, and the protozoan *Dictyostelium discoideum* (Yatzkan and Yarden 1999; Okazaki and Niwa 2000; Vilardell et al. 2000; Escalante et al. 2003; Ebbole et al. 2004). This, combined with the fact that basal branching eukaryotes possess degenerate splice signals (Schwartz et al. 2008) which are known to promote alternative splicing when present in higher eukaryotes (Ast 2004), is consistent with an early eukaryotic origin for alternative splicing.

In the eukaryotic ancestor where the spliceosome emerged, it has been hypothesized that the genome was replete with group II introns, and that group II introns were a driving force in the development of eukaryotic cells (Martin and Koonin 2006; Koonin 2009; Rogozin et al. 2012). Such an abundance of group II introns in a genome would provide ample opportunities for the formation of alternative splicing organizations such as *C.te*.I1. The discovery of this alternative spliced group II intron in *C. tetani* illustrates that group II introns are capable of producing alternative splicing arrangements adding support to the possibility that the emergence of alternative splicing may have occurred early in the evolutionary history of eukaryotes. Finally, the properties of *C.te*.I1 suggest that alternative splicing could have occurred prior to genesis of the spliceosome.

2.4.6 Summary

At first glance, both the loss of the IEP and the RNA structural anomalies might have suggested degeneration of the intron, but instead the intron was shown to be capable of alternative splicing creating five distinct RNAs from the CTC00465 exon. The 5′ splice site utilized by the intron is shifted 8 nt upstream of the standard 5′ GUGYG group II intron consensus sequence and functions to eliminate the stop codon of ORF CTC00465 and allow for in-frame fusion of upstream and downstream reading frames.

In addition to the alternative splicing properties, *C.te*.I1 is unique in the fact that it is the first known bacterial group II intron that has been found to splice without a maturase encoded either by the intron or elsewhere in the bacterial genome. As such, the intron may be capable of pure self-splicing *in vivo* or may rely on host encoded splicing factors similar to organellar group II introns.

Although the *C.te*.I1 intron was identified initially through bioinformatics and other similar organizations were looked for, it is likely that other instances of alternative splicing in bacteria have been overlooked. Due to their mobile nature, group II introns in bacteria are very often fragmented, with truncated sequences well outnumbering full-length intron copies. Given the known ability of introns to use 5′ and 3′ splice sites that are not adjacent to the ribozyme structure, it seems probable that there are additional alternative and/or *trans*-splicing reactions that have been overlooked.

Chapter Three: **THE RIBOZYME STRUCTURE OF *C.TE.*I1**

3.1 Introduction

In bacteria there is a clear co-evolution between intron RNA and IEP structures (Toor et al. 2001). Thus variations in RNA structure correspond to specific IEP classes. Bacterial group II introns can be divided into 8 structural classes based on IEP phylogeny: ML (mitochondrial-like), CL (Chloroplast-like; CL1 and CL2) Bacterial A, B, C, D, E and F (Zimmerly et al. 2001; Simon et al. 2008). Correlating these IEP classes back to the three main RNA structural families we find that ML class introns possess IIA structures, Class C introns belong to the IIC RNA structural group and all other intron classes fall under the more broad umbrella of IIB introns. As RNA structural families share biochemical and mechanistic properties, one can predict the expected properties of an intron either from its apparent RNA structure or from its ORF sequence. Although the *C.te.*I1 intron from *C. tetani* lacks an IEP, it bears greatest structural similarity to IIB introns, more specifically to bacterial Class B introns, and therefore can be expected to share mechanistic properties of IIB introns.

One of the crucial aspects of group II intron splicing and mobility is accurate splice site recognition. Precise selection of both 5′ and 3′ splice sites is achieved through base-pairing interactions between sequences present in D1 of the ribozyme and exon sequences adjacent to the splice sites. The mechanisms utilized for 5′ and 3′ splice site recognition differ between intron structural types. IIA and IIB introns share a common mechanism of 5′ splice site recognition, while IIB and IIC share a common mechanism of 3′ splice site recognition.

In IIA and IIB introns the 5′ splice site is specified by two sets of Watson-Crick base pairings: IBS1-EBS1 and IBS2-EBS2 (Figure 24). IBS1-EBS1 is formed between the 6 nt exon binding sequence 1 (EBS1) and the 6 nt sequence immediately upstream of the 5′ intron splice site, intron binding sequence 1 (IBS1) (Jacquier and Michel 1987). A second base-pairing interaction, IBS2-EBS2, is formed between an additional sequence in D1 (EBS2) and the sequence immediately upstream of IBS1, known as IBS2. Together these interactions provide an extended region of complementary, ~ 12 base pairs in length, which confers a high degree of specificity and allows for highly site-specific targeting during the mobility reaction (Guo et al. 1997). In IIC introns, the IBS2-EBS2 pairing is not present and a shorter 4 nucleotide IBS1-EBS1 pairing is used. As IIC introns insert downstream of intrinsic transcriptional terminators (Granlund et al. 2001; Dai and Zimmerly 2002; Robart et al. 2007) and *attC* site inverted repeat sequences (Centrón and Roy 2002; Quiroga et al. 2008), it has been suggested that IIC introns may interact with the upstream stem-loop to help recognize the 5′ splice site rather than utilizing the EBS2 pairing (Figure 24). As *C.te*.I1 possesses a IIB-type RNA structure it would be expected to utilize both the IBS1-EBS1 and IBS2-EBS2 pairings during 5′ exon recognition.

Recognition of the 3′ splice site is also conferred by sequences within D1. The IIB and IIC intron families share a common mechanism of 3′ exon recognition; however, the mechanism utilized by IIA introns differs and would have evolved independently in these introns. Common to all group II introns is a single nucleotide pairing between a nucleotide in the single stranded region between domains 2 and 3 (J2/3) and the final

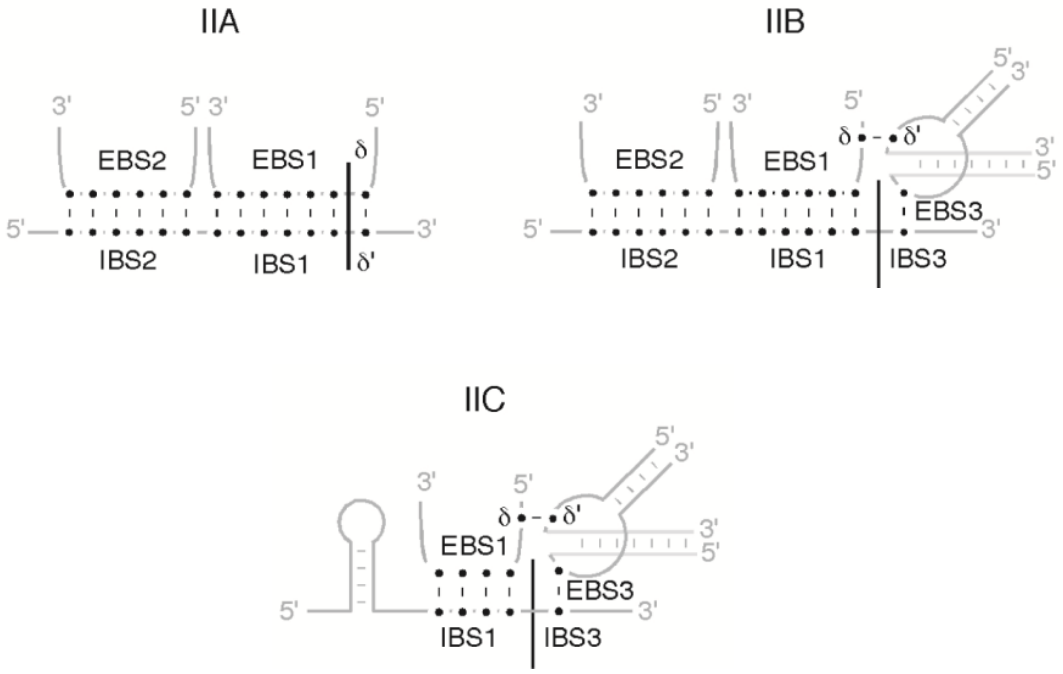
nucleotide of the intron sequence. This interaction, known as γ-γ′, serves to position the 3′ splice site at the active site (Jacquier and Michel 1990; Robart et al. 2004).

For IIA introns, the 3′ exon is positioned by a single nucleotide pairing between δ (the nucleotide adjacent to EBS1) and the first nucleotide of the 3′ exon (δ′) (Jacquier and Michel 1987). The standard mechanism of 3′ exon recognition for IIB and IIC introns involves a single nucleotide pairing between the nucleotide just upstream of the EBS1 sequence (δ) and a nucleotide (δ′) on the 5′ side of a bulge in the Id stem known as the "coordination loop" (Michel and Jacquier 1987). This pairing appropriately positions a nucleotide on the opposite side of the coordination loop (EBS3) and allows the formation of a Watson-Crick base pair with the first nucleotide of the 3′ exon (IBS3; the same nucleotide as δ′ in IIA nomenclature) (Costa et al. 2000). As a IIB intron, *C.te.*I1 would be presumed to utilize both δ-δ′ and EBS3-IBS3 pairings.

As mentioned in the previous chapter, a novel property of *C.te*.I1 is that the intron uses a 5′ splice site located eight nucleotides upstream of the canonically predicted 5′ intron-exon boundary *in vivo*. Use of this splice site is critical for alternative splicing as it eliminates a stop codon and allows correct fusion of the exon reading frames, and as it was identified *in vivo* could be specified by either the ribozyme itself or by other cellular factors; however, *C.te*.I1 possesses a unique secondary structure and it was hypothesised that features present within the novel RNA structure contributed to the recognition of the shifted 5′ splice site. In this chapter I investigate the unique structure of the *C.te*.I1 ribozyme and determine how the unique 5′ splice site is specified. Through this process I identified a potential extended pairing at the 3′ splice site which led to the investigation of whether a novel mechanism of 3′ splice site recognition is employed by *C.te*.I1.

Figure 24. Mechanisms of exon recognition.

IIA and IIB introns share similar mechanisms of 5′ splice site recognition, employing 6 nt EBS1-IBS1 and EBS2-IBS2 pairings. IIC introns use a shortened 4 nt EBS1-IBS1 and lack EBS2. IIA introns differ from IIB and IIC introns in their mechanism of 3′ splice site recognition. As these introns lack a "coordination loop" a single nucleotide pairing between δ and δ′ is utilized. In both IIB and IIC introns, pairing between δ and δ′ in the coordination loop allows base pairing between EBS3 and IBS3. Figure adapted from Lambowitz and Zimmerly (2011).



3.2 Materials and Methods

3.2.1 Strains, Growth, and gDNA Extraction

Clostridium tetani strain ATCC10779 (Designation 43415 – Harvard Strain) was obtained from the American Type Culture Collection. Cultures were grown in Brain Heart Infusion (BHI) media (Oxoid CM1135) at 37°C under anaerobic conditions using the GasPak EZ anaerobic container system (BD Biosciences). Genomic DNA was prepared from a 15 ml culture. Pelleted cells were washed and resuspended in 0.45 ml of TE [10 mM Tris–HCl (pH 7.4), 1 mM EDTA]. Lysis was by the addition of 0.1 mg proteinase K and SDS to a final concentration of 1% with incubation at 37°C for 45 min. The sample was repeatedly extracted with an equal volume of phenol-CIA (25:24:1 of phenol:chloroform:isoamyl alcohol) until there was a clear interface upon centrifugation. This was followed by ethanol precipitation in the presence of 0.3 M NaOAc (pH 5.2). The pellet was washed with cold 75% ethanol, air-dried and dissolved in TE [10 mM Tris–HCl (pH 7.4) and 1 mM EDTA].

Escherichia coli strains DH5α and BL21 were used for transformations and expression of pBluescript KS+ (Stratagene) generated constructs. Plasmid containing strains were grown in LB broth or on LB agar plates containing 100 μg/mL ampicillin at 37°C. Broth cultures were grown with aeration.

3.2.2 Cloning and Mutagenesis

The *C.te*.I1 intron was PCR amplified from genomic DNA in two pieces to delete 311 bp of additional sequence in domain 4 replacing it with 26 nt of loop sequence, using the primers S-5SSC-Bam and AS-5SSC-Eco, and S-3SSC-Eco and AS-3SSC-Cla (Table 8). The two pieces were ligated by recombinant PCR, and the DNA was cloned into the

BamHI (Invitrogen) and ClaI (Invitrogen) sites of pBluescript KS+ (Stratagene) in the orientation for T7 transcription. Similarly, a full-length control was made through amplification of *C. tetani* genomic DNA with the S-5SSC-Bam and AS-3SSC-Cla primers and cloned into the BamHI and ClaI sites of pKS+.

Mutant constructs were made using either end-to-end PCR or recombinant PCR for site-directed mutagenesis. End-to-end site-directed mutagenesis was performed as previously described (Byrappa et al. 1995) using 5′ phosphorylated oligos and *Pwo* DNA polymerase. All oligos used were synthesized by either Alpha DNA or Integrated DNA Technologies. Phosphorylated oligos were created using T4 polynucleotide kinase (Invitrogen; Carlsbad, CA) according to manufacturer's instructions. PCR products were extracted from a 1% agarose gel using the Zymoclean Gel DNA Recovery Kit (ZymoResearch) and ligated overnight using T4 DNA ligase (Invitrogen or New England Biolabs). Resultant plasmid constructs were transformed into chemically competent *E. coli* DH5α cells (Inoue et al. 1990). Plasmids were isolated using the GeneJET Plasmid Miniprep Kit (Fermentas) according to manufacturer's instructions. Oligonucleotide sequences used in mutagenesis are provided in Table 8. All plasmid constructs were confirmed by sequencing (U of C DNA sequencing lab) and are listed in Table 9.

3.2.3 RT-PCR

cDNA synthesis was performed using 10 pmole reverse primer (O1RL or AS-3SSC-Cla; Table 8) and 200 U Superscript II Reverse Transcriptase (Invitrogen) with 1 µg RNA as template, according to manufacturer's protocol. No RT controls in which no enzyme was added were also performed. One tenth of each RT-Reaction was used as template for PCR.

3.2.4 In vitro *Transcription and Self-Splicing*

The templates for *in vitro* transcription were linearized with XhoI (Invitrogen) for 90 minutes and then purified using the Zymoclean DNA Clean and Concentrate kit (ZymoResearch). Transcription reactions were performed in a volume of 20 µL of 40 mM Tris–HCl (pH 8.0), 4 mM $MgCl_2$, 50 mM NaCl, 1 mM NTPs, 5 mM DTT, 0.05% Triton X-100, 500 ng template and 2 µL T7 RNA polymerase (noncommercial source, undetermined activity). The reaction was incubated at 37°C for 30 minutes followed by phenol-CIA (25:24:1) extraction and ethanol precipitation in the presence of 2.5 M NH_4OAc. For radiolabeled transcripts, reactions contained 1 µL of [α-^{32}P] UTP (10 mCi/ml; 3000 Ci/mmol; MP Biomedical). Transcripts were re-suspended in TE. For self-splicing reactions, ^{32}P-labelled (100, 000 c.p.m.) or cold transcript (200 ng) was first subjected to a pre-folding program in a Veriti 96 well thermocycler (Applied Biosystems) with incubations at 90°C for 1 min, 75°C for 5 min and slow cooling to 53°C over 15 min. Buffer was added to a final volume of 50 µL of 100 mM $MgCl_2$, 0.5 M NH_4Cl and 40 mM Tris–HCl (pH 7.5), and the reaction was incubated at 53°C for 5 minutes unless otherwise specified. The splicing reaction was ethanol precipitated with 0.3 M NaOAc (pH 5.2) and resuspended in formamide dye [95% formamide, 5 mM EDTA (pH 8.0), 0.1% of xylene cyanol and 0.1% of bromophenol blue], heated to 85°C for 2 min and resolved on a 4% polyacrylamide (19:1 acrylamide:bisacrylamide ratio)/8 M urea gel. For all unspliced negative controls, reactions were done in parallel without $MgCl_2$ or monovalent salts. Alternate splicing conditions include variations in temperature from 38-63°C, varying concentrations of monovalent salts including NaCl, KCl and NH_4Cl and varying $MgCl_2$ concentrations from 0 to 100 mM.

Gels were exposed over night to imaging plates and were then imaged using a Storm 860 phospho-imager (Amersham). Quantification was performed using the ImageQuant TL 2005 software from Amersham Sciences.

3.2.5 Branchpoint PCR

For sequencing through the unusual products created in the self-splicing reactions of the EBS1 and IBS1 mutants, cold transcripts were self-spliced and ethanol precipitated. An RT reaction was performed using the I1R oligo (Table 8) according to manufacturer's instructions. Subsequent PCR was performed according to the generalized PCR protocol using a gradient of annealing temperatures and the I1R and 3′-SSC-S oligo. Resultant PCR products were gel extracted and sent for sequencing.

Table 8. Oligos used for construction of wild type and mutant self-splicing constructs

Primer/Oligo Name	Oligo Sequence (5'→3')
S-5SSC-Bam	CGCGGATCCACAGCAGAAGATGGAACAACAGC
AS-5SSC-Eco	CCGGAATTCTTCCACTTTTCTATTACCGACTAGG
S-3SSC-Eco	CCGGAATTCTATACACTTCCTATTATCGAGCG
AS-3SSC-Cla	CCATCGATGTTTTGTTATTTCACTTATTAGTTCC
mB′-S	AAAACAGCTTGTGGCTTGGTATAAAACAG
mB′-AS	ATATTTAACAGTATAAAATACATATAAATCGC
mB-S	TGTTTTGTGCGAAACGTTATCC
mB-AS	TGTATGGCTATAGTATATTCAGC
mIBS1-S	CGGTATCAATAAAAGTGCGAAACG
mIBS1-AS	TATAGTATATTCAGCTGTTGTTCC
mEBS1-S	ATACCGTTGGTATAAAACAG
mEBS1-AS	AGCACAAAAATATTTAACAGTATAAAATAC
ΔIBS1-fs-S (Δ8 nt)	GCCATAGTGCGAAACGTTATCC
mEBS1loop-S	ATTTATTTTTTGTGCTTGTGGC
mEBS1loop-AS	CAGTATAAAATACATATAAATCGCTAC
ΔEBS1loop-S	CTTGTGGCTTGGTATAAAACAGTTAAGATG
mEBS2-S	ATATCATATTTATACTGTTAAATATTTTTGTGC
mEBS2-AS	AAATCGCATCTTAATGGATTAGCTTATGTTTACC
mIBS2-S	TATGTATAGCCATACAATAAAAGTGCG
mIBS2-AS	ATTCAGCTGTTGTTCCATCTTCTGC
mUGC-S	TTTTGACGTTGTGGCTTGGTATAAAACAG
m3'-Exon-S	ACGTAAGGAGAAAATACAGATAATTTAGGTGTGG
m3'-Exon-AS	ATACCTCACTATCCTAAATGCATAGGATTTCGC
mδ-S	CGCAAAGGAGAAAATACAGATAATTTAGGTGTGG
mδ′-S	ATTAAGATCCGATTTATATG
mδ′-AS	GGATTAGCTTATGTTTACCATAGTAATC
mα′-S	TTTGGAATAAAACCAAGTTAAACTAAAACATTATAAATCGTTAGTGG
mα-S	GCAAACTTGGTTGTAGCTTTAAGTCTAAGTCCCCTACACAAG
mα-AS	TTTAAATAAGATTTTCACTTATTTACTTTAGTAGAACAGTTAGAAC
mEBS3-S	TGAGTGGATACCTAAGG
mEBS3-AS	CGATTTATAATGTTTTAGTTTAAC
mIBS3-S	TGCAAAGGAGAAAATACAGATAATTTAGGTGTGG
Stem1-m-S	CTAAAGTAATAAGTGAAAATCTTATTT
Stem1-m-AS	GATCTTCTGTTAGAACTTTTAAGA
Stem1-m′-S	GGATCTTTGGAACGTTTGTAG
Stem1-m′-AS	TATCCCTATTTTTGATTACCTTAGG
Stem2-m-S	TAAATATTTTTGTGCTTGTGGC
Stem2-m-AS	CAGGCGCAAAATACATATAAATCGCATC
Stem2-m′-AS	GCCACAAGCACAAAAATATTTAACAG
Stem2-m′-S	TTGGGCGCAAACAGTTAAGATGAAGTAC
Stem3-F	GATGAAGTACTTAACTGGTTTTGGAATAATTG
Stem3-Fmut	GATGAAGTACTTTTGTGGTTTTGGAATAATTG
Stem3-R	GTACTTCATCTTAACTGTTTTATACCAAGCC
Stem3-Rmut	GTACTTCATCTTTTGTGTTTTATACCAAGCC
Stem4-m-S	GGTTGTTAAACTAAAACATTA
Stem4-m-AS	AATTATTCCTTTTCCAGTTAAGTACTTC
Stem4-m′-S	GGTTGTTAAACTTTTTCATTATAAATCGTTAGTGG
Stem4-m′-AS	AAATTATTCCAAAACCAGTTAAGTACTTG
G-3'Exon-C-S	ACCAAAGGAGAAAATACAGATAATTTAGGTGTGG
C-3'Exon-G-S	AGGAAAGGAGAAAATACAGATAATTTAGGTGTGG
A-3'Exon-T-S	AGCTAAGGAGAAAATACAGATAATTTAGGTGTGG
GC-3'Exon-CG-S	ACGAAAGGAGAAAATACAGATAATTTAGGTGTGG
G-EBS1loop-C	TTTTGTCCTTGTGGCTTGGTATAAAACAG
C-EBS1loop-G	TTTTGTGGTTGTGGCTTGGTATAAAACAG

Table 9. List of self-splicing constructs

Construct name	Description	Template	Primers Used
WT; *C.te*.I1-SSC	WT construct containing a 311 nt deletion of D4, 90 nt of 5′ exon (CTC00465) and 81 nt of 3′ exon (CTC00467)	*C. tetani* ATCC10779 gDNA	5′ SSC S
			5′ SSC AS
			3′ SSC S
			3′ SSC AS
mEBS1	TGTGGC→ATACCG mutation in EBS1	*C.te*.I1-SSC	mEBS1-S
			mEBS1-AS
mIBS1	GCCATA→CGGTAT mutation in IBS1	*C.te*.I1-SSC	mIBS1-S
			mIBS1-AS
mEBS1/mIBS1	Combination of mEBS1 and mIBS1 mutations	mEBS1	mIBS1-S
			mIBS1-AS
mB′	TTTTGT→AAAATA Mutation in B′	*C.te*.I1-SSC	mB′-S
			mB′-AS
mB	ATAAAA→TGTTTT Mutation in B	*C.te*.I1-SSC	mB-S
			mB-AS
mB/mB′	Combination of mB and mB′	mB′	mB-S
			mB-AS
Δ8nt IBS1	Deletion of the 8 nt between IBS1 & the 5′ GUGYG	*C.te*.I1-SSC	mIBS1-AS
			ΔIBS1-fs-S
ΔEBS1 loop	EBS1 loop deleted to Standard Class B sized loop	*C.te*.I1-SSC	ΔEBS1loop-S
			mEBS1loop-AS
mEBS1 loop	TAAATA→ATTTAT at 5′ end of EBS1 loop	*C.te*.I1-SSC	mEBS1loop-AS
			mEBS1loop-S
mEBS2	GTAT→CATA at possible location of EBS2	*C.te*.I1-SSC	mEBS2-S
			mEBS2-AS
mIBS2	ATAC→TATG at possible location of IBS2	*C.te*.I1-SSC	mIBS2-S
			mIBS2-AS
mUGC	TGC→ACG immediately upstream of δ	*C.te*.I1-SSC	mUGC-S
			mB'-AS
m3'exon	GCA→CGT immediately downstream of IBS3	*C.te*.I1-SSC	m3′-Exon-S
			m3′-Exon-AS
mUGC/m3'exon	Combined mutations of mUGC and m3′exon	mUGC	m3′-Exon-S
			m3′-Exon-AS
mα	AACCAA→TTGGTT mutation at end of IB stem	*C.te*.I1-SSC	mα-S
			mα-AS
mα′	TTGGTT→AACCAA mutation in α′	*C.te*.I1-SSC	mα′-S
			I1R
mα/mα′	Combination of both mα and mα′	mα	mα′-S
			I1R
mδ	T→A mutation of the nucleotide 5′ of EBS1	*C.te*.I1-SSC	mδ-S
			mB'-AS

Construct Name	Description	Template	Primers Used
mδ′	G→C mutation on the 5′ side of the coordination loop	*C.te*.I1-SSC	mδ′-S
			mδ′-AS
mδ/mIBS3	Combination of mδ and mIBS3	mIBS3	mδ-S
			mB'-AS
mIBS3	A→T mutation at the first position of the 3′ exon	*C.te*.I1-SSC	mIBS3-S
			m3′-Exon-AS
mEBS3	T→A mutation on the 3′ side of the coordination loop	*C.te*.I1-SSC	mEBS3-S
			mEBS3-AS
mEBS3/mIBS3	Combination of both EBS3 and IBS3 mutations	mIBS3	mEBS3-S
			mEBS3-AS
Stem1 m	ATCTT→TAGAA mutation of the 5′ side of the I(ii) stem	*C.te*.I1-SSC	Stem1-m-S
			Stem1-m-As
Stem1 m′	TAGAA→ATCTT mutation of the 3′ side of the I(ii) stem	*C.te*.I1-SSC	Stem1-m′-S
			Stem1-m′-AS
Stem1 m/m′	Compensatory mutation of both sides of the I(ii) stem	Stem1-m	Stem1-m′-S
			Stem1-m′-AS
Stem2 m	TATA→GCGC mutation in the 5′ side of the EBS1stem	*C.te*.I1-SSC	Stem2-m-S
			Stem2-m-AS
Stem2 m′	TATA→GCGC mutation in the 3′ side of the EBS1stem	*C.te*.I1-SSC	Stem2-m′-S
			Stem2-m′-AS
Stem2 m/m′	Compensatory mutation of both 5′ and 3′ sides	Stem2-m′	Stem2-m-S
			Stem2-m-AS
Stem3 m	GTT→CAA on the 5′ side of the stem following EBS1 through recombinant PCR	ATCC10779 genomic DNA	5′ SSC S
			Stem3Rmut
			Stem3F
			3′ SSC AS
Stem3 m′	AAC→TTG on the 3′ side of the stem following EBS1 through recombinant PCR	ATCC10779 genomic DNA	5′ SSC S
			Stem3R
			Stem3Fmut
			3′ SSC AS
Stem3 m/m′	Both 5′ and 3′ mutations on the stem following EBS1 through recombinant PCR	ATCC10779 genomic DNA	5′ SSC S
			Stem3Rmut
			Stem3Fmut
			3′ SSC AS
Stem4 m	TTTT→AAAA mutation 5′ of the α′ sequence	*C.te*.I1-SSC	Stem4-m-S
			Stem4-m-AS
Stem4 m′	AAAA→TTTT mutation at the 3′ side of the α′ sequence	*C.te*.I1-SSC	Stem4-m′-S
			Stem4-m′-AS
Stem4 m/m′	Compensatory mutation of Stem4 m and Stem4 m′	Stem4 m′	Stem4-m-S
			Stem4-m-AS

GloopC	G→C mutation in the UGC sequence of the EBS1 loop	*C.te*.I1-SSC	G-EBS1loop-C
			mB′-AS
CloopG	C→G mutation in the UGC sequence of the EBS1 loop	*C.te*.I1-SSC	C-EBS1loop-G
			mB′-AS
G3'ExonC	Second conserved 3′ Exon position mutated G→C	*C.te*.I1-SSC	G-3'Exon-C-S
			3′-Exon-AS
C3'exonG	Third conserved 3′ Exon position mutated C→G	*C.te*.I1-SSC	C-3'Exon-G-S
			3′-Exon-AS
GC3'ExonCG	2^{nd} & 3^{rd} conserved 3′ Exon position mutated GC→CG	*C.te*.I1-SSC	GC-3'Exon-CG-S
			3′-Exon-AS
A3'exonT	Final (4^{th}) conserved 3′ Exon position mutated A→T	*C.te*.I1-SSC	A-3'Exon-T-S
			3′-Exon-AS

3.3 Results

3.3.1 Secondary Structure and Unique Intron Features

As discussed briefly in Chapter 2 of this thesis, the secondary structure of *C.te.*I1 (Figure 25) is most closely related to intron RNAs of Class B. Consistent with this relationship, the secondary structure of *C.te.*I1 has features shared by nearly all Class B introns (Figure 25), including a CGC "catalytic triad" rather than AGC, and a double stem insertion between the I(i) and I(ii) stems near the start of the intron. In addition, there are features shared by most but not all B introns, including a short stem 3′ to the α′ sequence and the lack of a β-β′ interaction [(Toor et al. 2001); and Zimmerly lab, unpublished]. More specifically, *C.te.*I1 has features of the β lineage of Class B introns (Stabell et al. 2009), including the presence of a domain IA motif and an additional stem 5′ adjacent to the α′ sequence (Figures 25 and 26). Although the intron does not encode an IEP, it is apparent that *C.te.*I1 is derived from a mobile intron of Class B that lost its IEP and acquired structural variations. Interestingly, the 311 nt loop in D4 bears no resemblance to the IEP sequence that, presumably, originally resided there.

Most of the unique structural features possessed by *C.te*.I1 are located in D1 near the EBS1 motif, which is responsible for 5′ exon recognition. Most notably, potential EBS1 sequences are located within a loop of 24 nt, whereas EBS1 is located in a smaller loop of 8-11 nt in other class B introns. An alignment of the Class B introns belonging to the β-lineage shows that this loop is tightly conserved in size in this lineage with only 8 or 9 nt located in the loop (Figure 27). In most Class B introns, the stem beside EBS1 possesses an internal bulge motif. This stem in *C.te*.I1, however, is perfectly complementary and does not contain a bulge. Another unusual feature of *C.te.*I1 is that

Figure 25. Secondary structure of *C.te.*I1.

Locations of major tertiary interactions are noted (α-α′, γ-γ′, ε-ε′, ζ-ζ′, κ-κ′, and λ-λ′). Interactions in parentheses are elements not supported by mutagenesis experiments (δ′, IBS2-EBS2, EBS3). Locations of both the expected IBS1 (based on the 5′ consensus of group II introns) and actual IBS1 (experimentally supported) are noted and the corresponding intron boundaries are shown with a diamond and star, respectively. The 3′ intron junction is shown with a triangle. Boxes with dotted outlines and light grey shading indicate unexpected structural features, unfilled boxes with dotted lines indicate features typical of Class B introns, and the boxes with asterisks denote features found in the β sublineage. The inset box shows a partial consensus secondary structure of same sublineage of Class B introns, highlighting the irregularities of the *C.te.*I1 EBS1/EBS2 region.

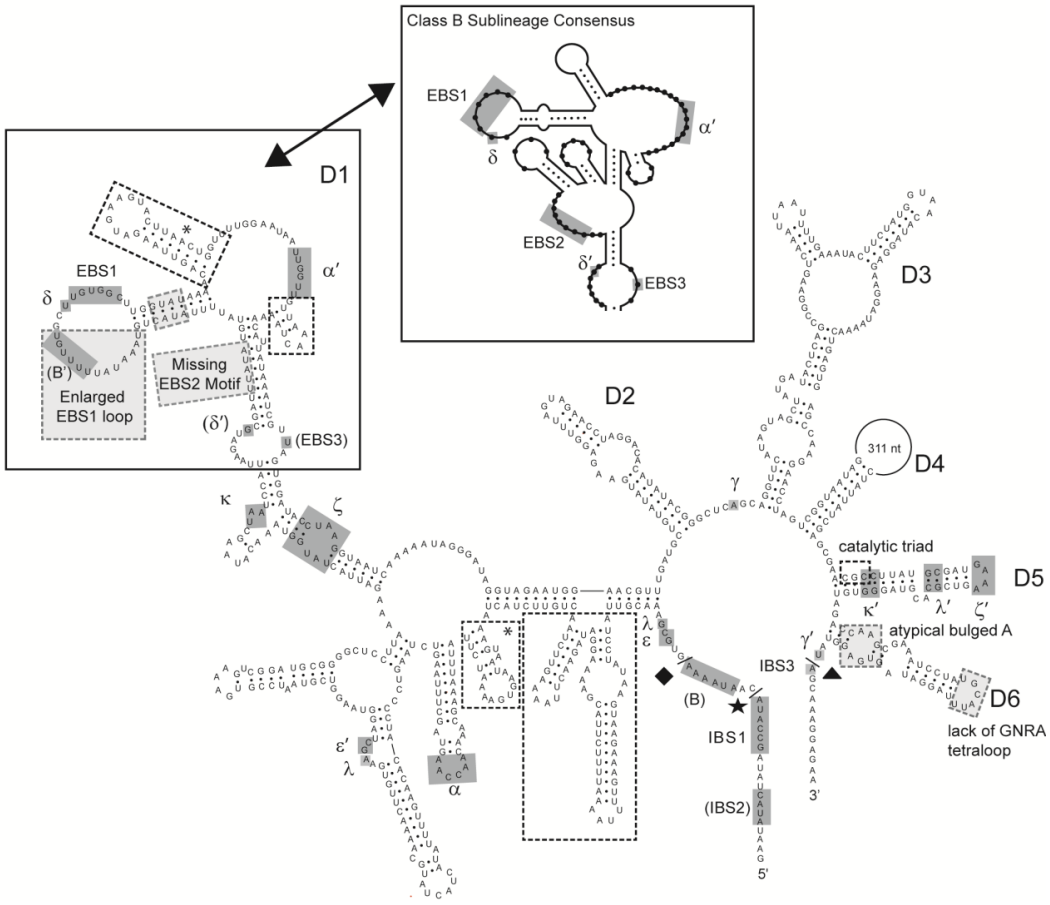


Figure 26. Schematic drawings illustrating the differences in the secondary structures between the two subgroups (α and β) within Class B group II introns.

Boxed areas show the features unique to each subgroup. For the β lineage the α′ stem can form either the structure shown for the α lineage or be folded such that α′ is present in a loop sequence at the end of a stem. Introns from Bacillus cereus, *B.c.*I4 and *B.c.*I5, were taken as representatives of subgroup α and β, respectively. Image taken from (Stabell et al. 2009).

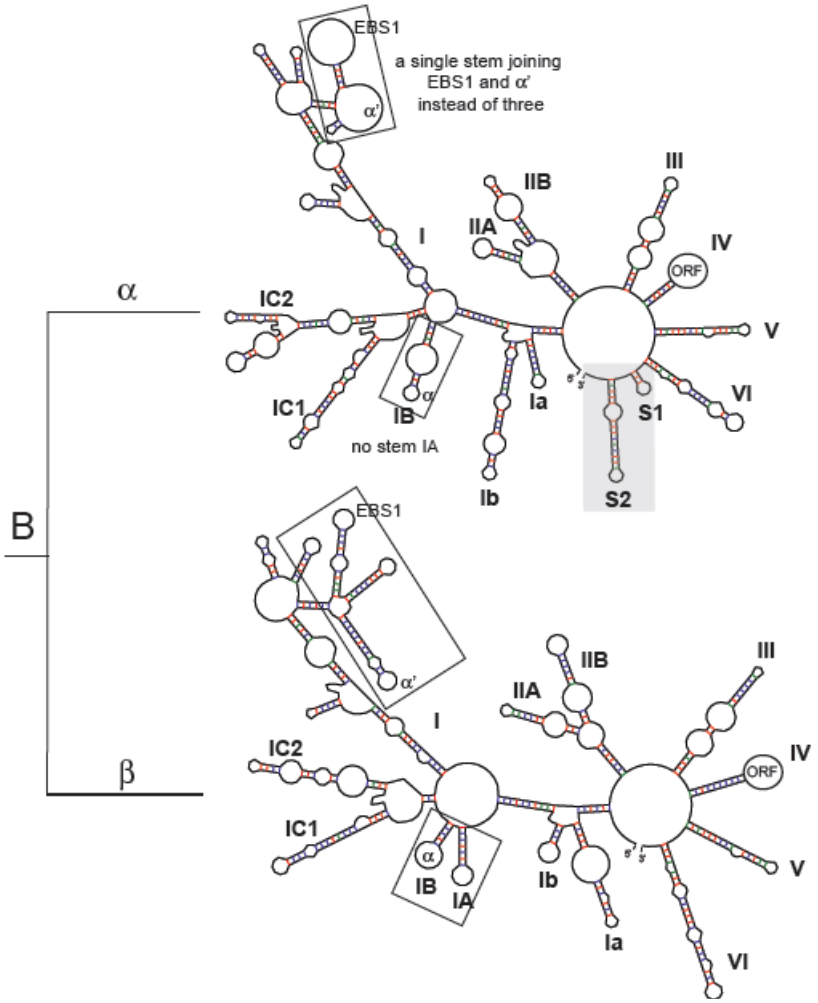


Figure 27. RNA sequence alignment for the EBS1 stem loop of the β-Lineage of Class B Nucleotides pairing in the stem are indicated by brackets underneath the sequence. The single nucleotide at position 35 in this alignment is delta (d) and the EBS1 sequences are aligned at positions 37-42.

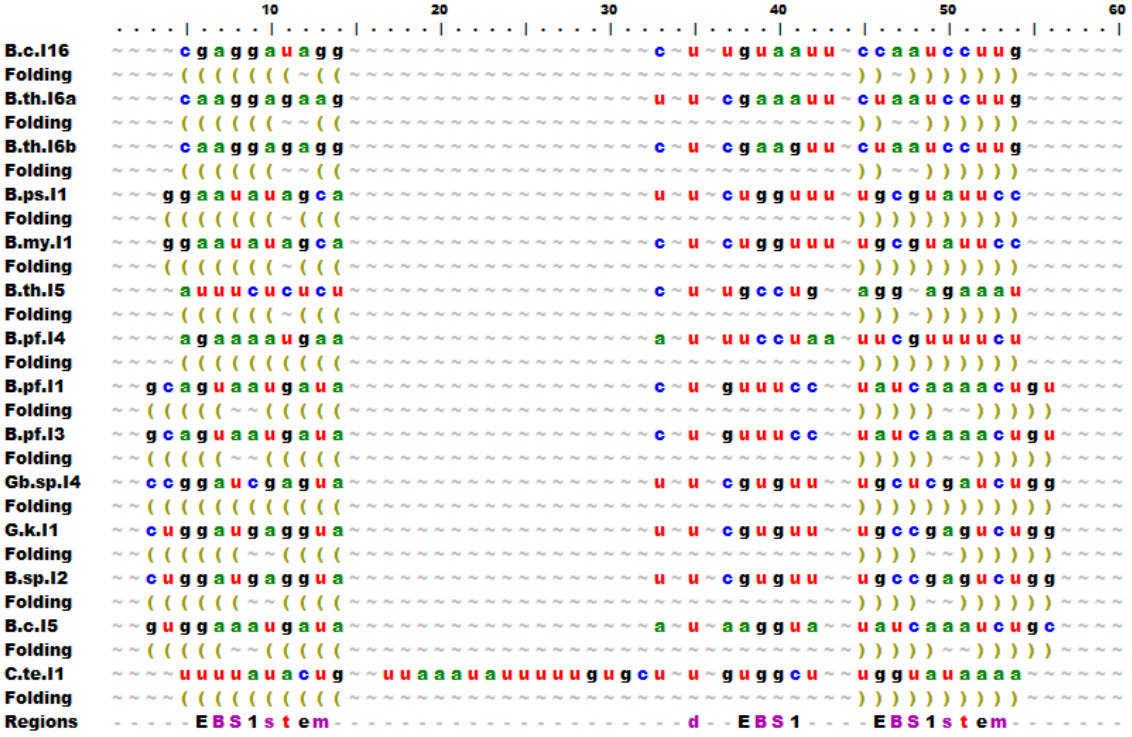
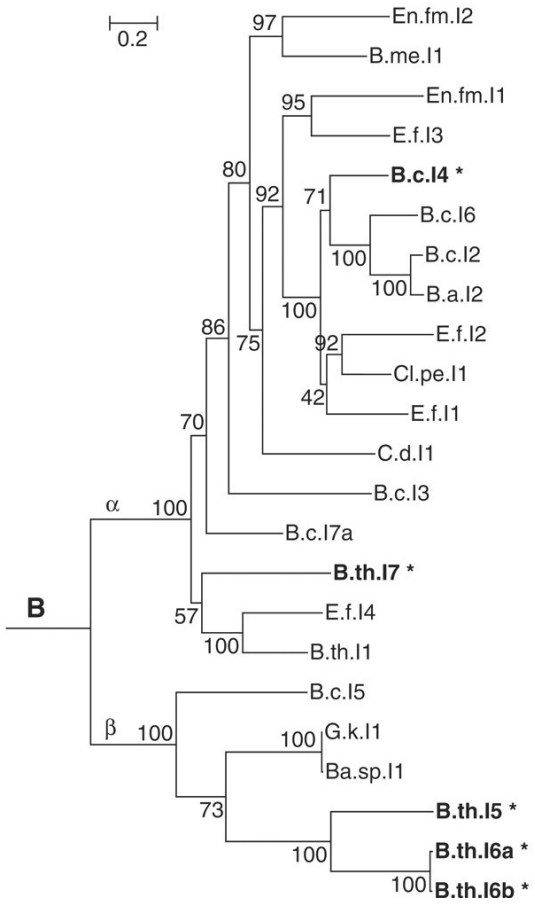


Figure 28. Detailed rooted phylogenetic tree of the B class group II introns

The tree was constructed using the maximum-likelihood method (RAxML program) and was based on amino acid sequences covering the full length of the intron encoded ORF. Asterisks indicate introns with 3′ terminal extensions known as D7. Sequences and secondary structure models of introns can be found in the Bacterial Group II intron Database [http://webapps2.ucalgary.ca/~groupii/index.html#]. The numbers next to branch nodes indicate bootstrap support values (in percentage out of 1000 replicates). Scale bars are in average numbers of amino acid substitutions per site. Sublineages within the Class B are labeled α and β. Figure adapted from (Stabell et al. 2009).



the EBS2 motif, which normally contributes to 5′ exon recognition, appears to be missing. It is of note, however, that two alternate foldings of the secondary structure allow for the formation of potential, modest EBS2 sequences. In other parts of the RNA structure, the D6 stem lacks the GNRA tetraloop that participates in the η-η′ interaction (Chanfreau and Jacquier 1996), and the branch A motif does not have a clearly bulged A residue. Interestingly the structural variations in D6 are maintained in all of the downstream copies of D5/6 that are utilized in alternative splicing.

The structural changes between the α and β lineages of class B introns also correspond to changes in IEP amino acid sequence and as such, the sub-lineages form separate clades in a phylogenetical tree (Figure 28). Even though *C.te*.I1 lacks an IEP it possesses features of the β lineage and would be expected to group with these introns in RNA-based phylogenetical analyses. Several, but not all, of the known group II introns possessing domain 7 structures also belong to this sub-lineage.

3.3.2 *Establishing a Self-Splicing Assay for* C.te.I1

In order to assess the contributions of the various structural features to the ribozyme, a self-splicing assay was developed for the intron and the standard self-splicing properties of the intron assessed *in vitro*. *C.te*.I1 was cloned into a pBluescript vector (pKS+) and 311 nt of sequence in domain 4 were deleted and replaced by 26 nt of loop sequence, as it is known that this region can interfere with the ability of the intron to fold and self-splice *in vitro*. This construct containing the deletion of D4 will be referred to as the wild type self-splicing construct (WT-SSC). A full-length construct containing the full D4 sequence was also made as a self-splicing comparison. Both constructs were found to be capable of self-splicing *in vitro* (Figure 29). The intron splices through a

hydrolysis pathway, with no detectable lariat product formed. This indicates that the first step of the self-splicing reaction is initiated by a water molecule rather than by the bulged adenosine residue in D6 that typically initiates the splicing reaction for most group II introns. As an atypical bulged adenosine sequence is present within D6, this observation is consistent with the expected behaviour of the intron.

The wild-type (WT) intron construct spliced so robustly that self-splicing occurred during the transcription reaction used to generate precursor for the self-splicing assays. This is unusual, as the conditions during transcription (37°C and 20 mM $MgCl_2$) are closer to physiological than typically necessary for *in vitro* self-splicing [100 mM $MgCl_2$, 0.5 M NH_4Cl and 40 mM Tris–HCl (pH 7.5]), however the *O. iheyenis* IIC intron has also been reported to undergo self-splicing during transcription (Toor et al. 2008). As such, the buffer used for transcription was optimized in order to eliminate self-splicing of the intron construct while still allowing the activity of the T7 RNA polymerase (Figure 30). It was found that the intron spliced in buffers containing as little as 10 mM $MgCl_2$ without spermidine and in 5 mM $MgCl_2$ buffers with spermidine. Subsequently a buffer containing a working concentration of 4 mM $MgCl_2$ and 0 mM spermidine was used. The elimination of self-splicing during transcription resulted in the production of a single detectable RNA species, the intron precursor molecule, which allowed transcript to be used directly in self-splicing assays without the need for gel purification.

The WT construct was tested for self-splicing efficiency through a time course assay of *in vitro* self-splicing (Figure 31). Under these conditions, the intron self-spliced very rapidly with approximately 40% fully spliced after 30 seconds and 73% splicing in just 2 minutes (for quantification details see Materials and Methods). After 5 minutes, the

Figure 29. Self-splicing assay of wild-type *C.te*.I1 constructs

Polyacrylamide gels of self-splicing reactions for wild-type *C.te*.I1 constructs. Regions above 600 nt are shown. The ΔD4 construct is the standard WT construct used throughout this thesis and contains a 311 nt deletion of the sequence present in D4. The full-length (FL) construct contains the full D4 sequence. Self-splicing tested at 0 and 5 minute time points. Location of RNA markers are shown to the left of the gels. Intron precursor is denoted as P, intron-3′ exon intermediate product as I, and linear intron as L.

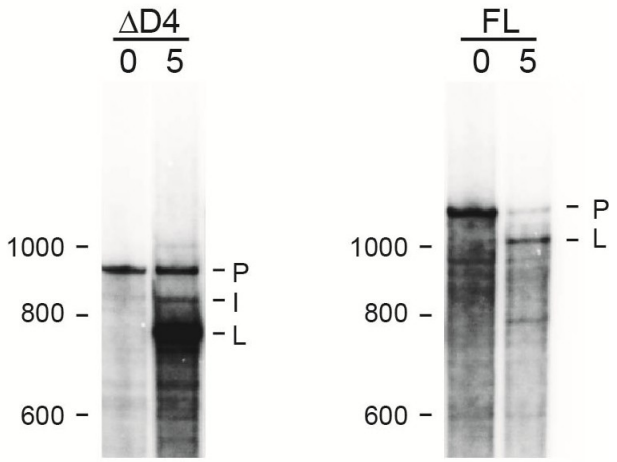


Figure 30. Transcription buffer optimization.

Buffer A is standard transcription buffer which contains a final concentration of 20 mM $MgCl_2$ and also contains spermidine. Buffer B, C and D are buffers that lack spermidine and contain 20 mM, 10 mM and 2 mM $MgCl_2$ respectively. Buffers E, F and G all contain spermidine but have 5 mM, 2 mM and 1 mM $MgCl_2$ respectively. Transcripts shown here are labeled with ^{35}S-UTP.

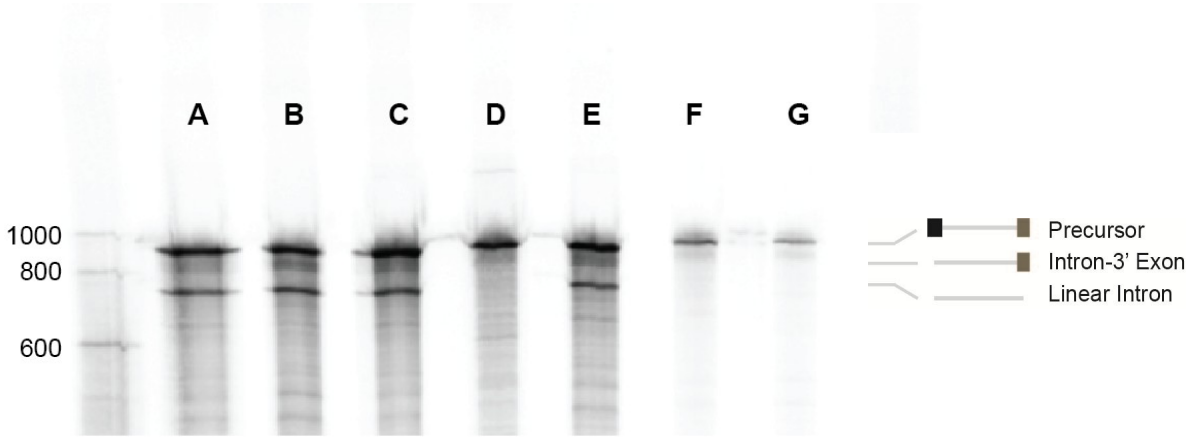


Figure 31. Self-splicing time course of WT *C.te*.I1

Reactions are started by the addition of 2X splicing buffer (80 mM Tris-HCl, pH 7.5, 1 M NH_4Cl, 200 mM $MgCl_2$) and stopped following the desired time interval with EDTA. Exposure is increased in the adjacent panels of ligated exons, 5′ and 3′ exons. The light grey box surrounding the intron precursor, intron 3′-exon intermediate and linear intron products shows the area shown in some subsequent gel figures.

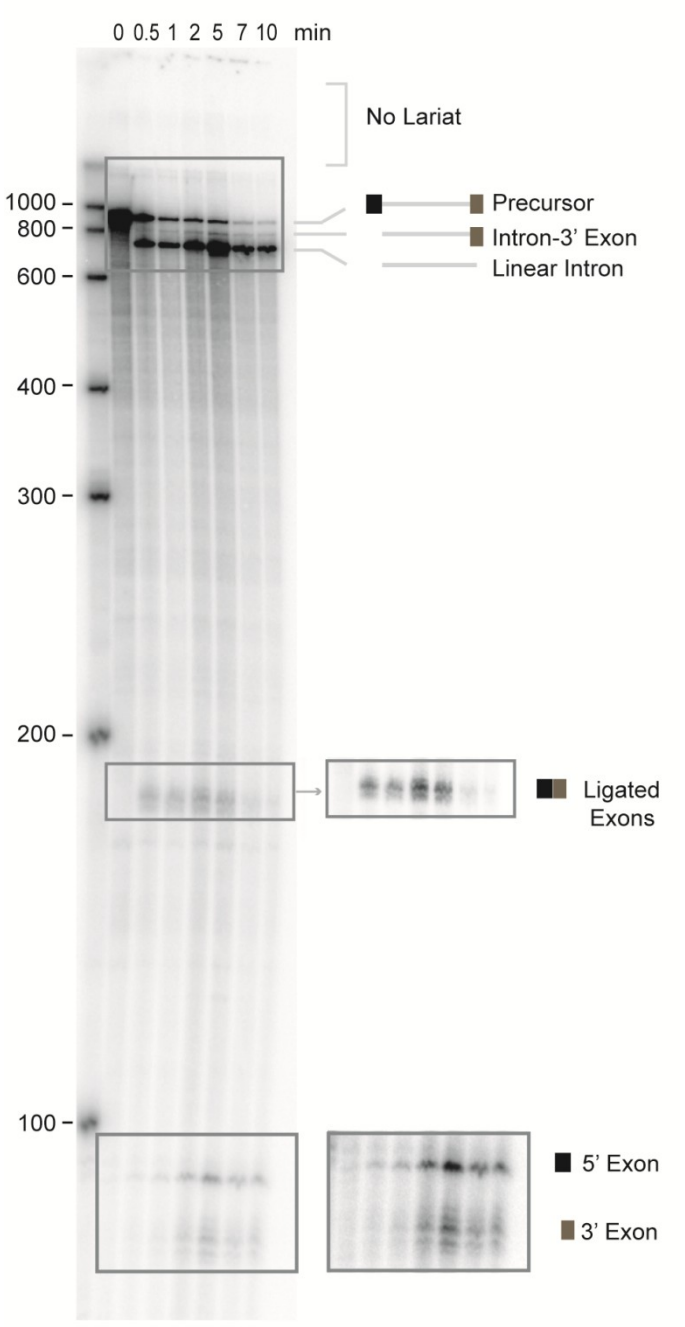
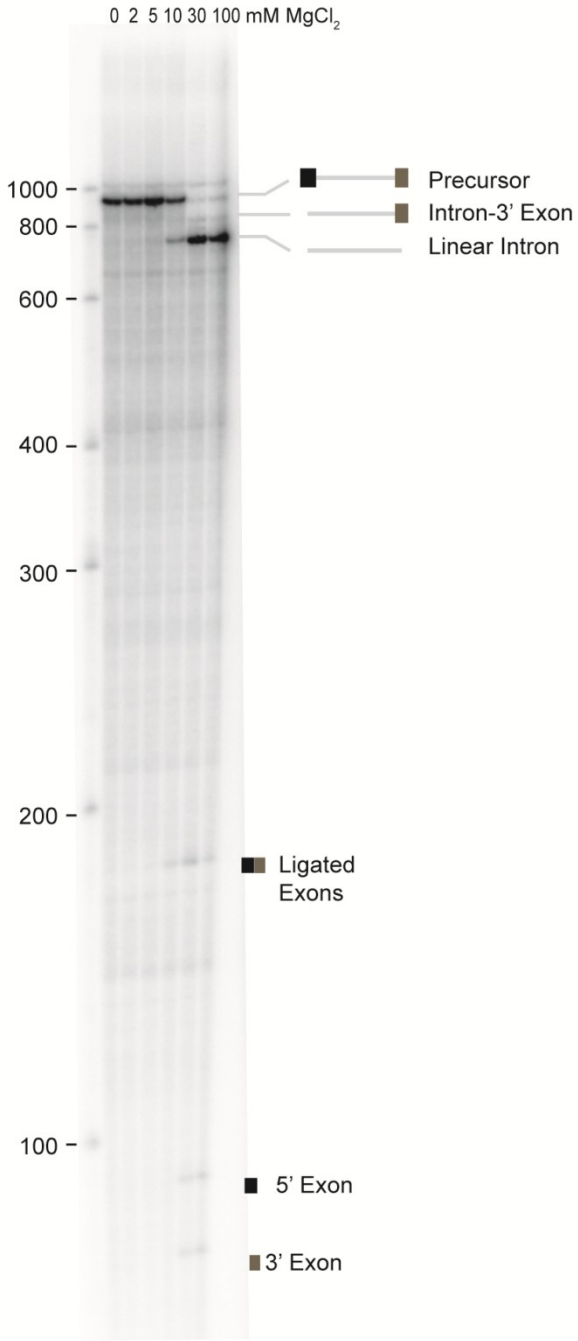


Figure 32. *In vitro* self-splicing in differing $MgCl_2$ concentrations.

Wild-type *C.te*.I1 construct spliced in differing concentrations of $MgCl_2$, but maintaining 0.5 M NH_4Cl and 40 mM Tris-HCl, pH 7.5. Splicing reactions were performed for 10 minutes at 53°C.



intron had reached completion of the splicing reaction with ~95% of the intron precursor having reacted to produce linear intron and ligated exons. Very little intermediate product of intron-3′ exon is observed for the WT intron under these conditions and splicing is completely through the hydrolysis pathway with no evidence of lariat intron being produced. After the 5 minute time point a substantial amount of spliced exon reopening (SER), hydrolysis of ligated exons, is observed.

In addition to analyzing the self-splicing of the *C.te.*I1 WT construct through radiolabeled polyacrylamide gel electrophoresis, RT-PCR was conducted to detect ligated exon products. Amplified cDNAs of ligated exons were gel extracted and sequenced. Sequencing revealed that the splice site used *in vitro* was also shifted 8 nucleotides upstream of the canonical 5′ intron boundary (5′GUGYG). As *in vitro* self-splicing assays are conducted in a protein-free environment, it can be concluded that selection of this shifted 5′ splice site is a property intrinsic to the ribozyme itself.

The *C.te*.I1 intron is capable of self-splicing under low $MgCl_2$ concentrations, as evidenced by the self-splicing observed during the transcription reaction. During 5 minute reactions the intron was shown to self-splice in conditions as low as 10 mM $MgCl_2$, but splicing was not observed in buffers containing 2 mM or 5 mM Mg^{2+} (Figure 32). Given longer incubation periods and the presence of spermidine, such as observed during the transcription buffer optimization, it is evident that this intron can splice in $MgCl_2$ concentrations as low as 5 mM (Figure 30).

The hydrolysis pathway is used by *C.te*.I1 in all monovalent salt conditions tested, indicating that this property is intrinsic to *C.te*.I1 and not an artifact of the conditions used (Figure 33). This is consistent with the recently identified subset of mitochondrial

IIB1 introns that possess similar features (Li et al. 2011b), and is consistent with the loss of the branchpoint adenosine motif in D6. The use of KCl as the monovalent salt (lane B) results in a doublet band of linear intron, which suggests that off-target splicing occurs under these conditions. Sodium chloride seems to allow for a very slight increase in the accumulation of intermediate intron-3′ exon product (lanes F, G, H) but is otherwise comparable to NH_4Cl. Consistent with these findings NH_4Cl is used as the monovalent salt in the splicing buffer for all subsequent assays as it appears to promote complete and accurate splicing; however *C.te.*I1 splices very well in all of the monovalent salt conditions tested, again indicating the robustness of the intron's splicing reaction.

Self-splicing reactions of the WT-*C.te.*I1 construct were also carried out at a range of temperatures from 38-63°C. While the construct was found to splice optimally at 53°C, substantial levels of splicing can be observed across a variety of temperatures Interesting, a small portion of the population of RNA molecules are seen to splice even at the lowest temperature tested, 38°C, which is very near physiological temperatures.

Standard self-splicing protocols heat the intron precursor to 95°C to denature the RNA structure and then proceed through a gradual cooling of the RNA ("folding") allowing the majority of the RNA molecules to find their active native conformations rather than become kinetically trapped misfolded structures [For reviews of RNA folding kinetics see (Treiber and Williamson 2001; Chen 2008)]. Approximately 50% of *C.te.*I1 RNA molecules fold well adopting an active conformation capable of self-splicing without the gradual cooling used to fold group II introns (Figure 34), while essentially all of the RNA molecules adopt an active conformation after folding. These results combined with the evidence that splicing occurs during the transcription reaction for

Figure 33. Effect of monovalent salts on self-splicing of the WT *C.te.*I1 construct**.**

Buffer A is the control showing transcript spliced in buffer containing 0 mM $MgCl_2$. Buffer B contains 1M KCl. Buffers C, D and E contain 1M, 0.5 M and 0.25 M NH_4Cl respectively. Buffers F, G and H contain 1 M, 0.5 M, and 0.25 M NaCl respectively. All reactions are for 5 minutes at 53°C. All other components of the buffer remain constant.

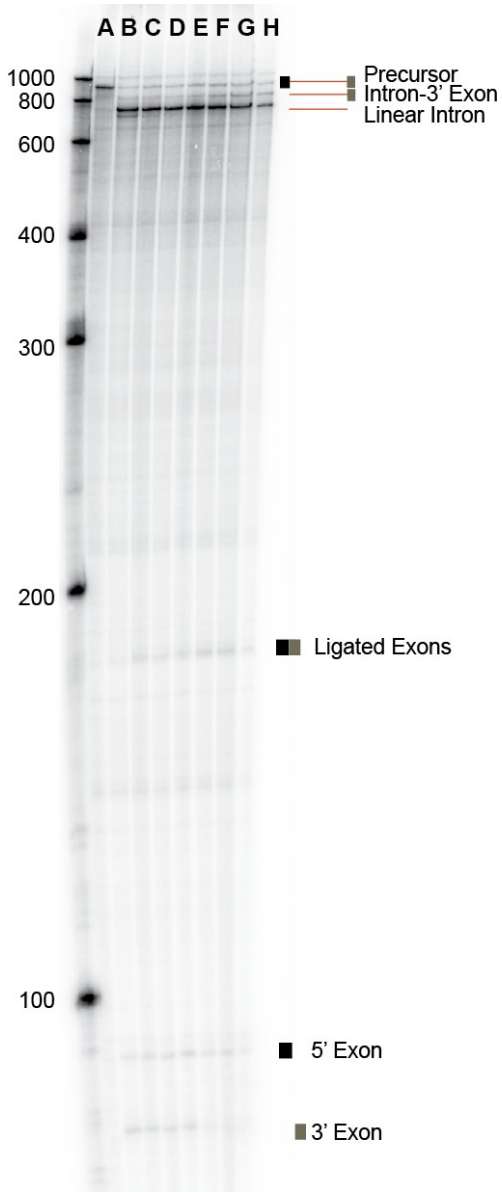
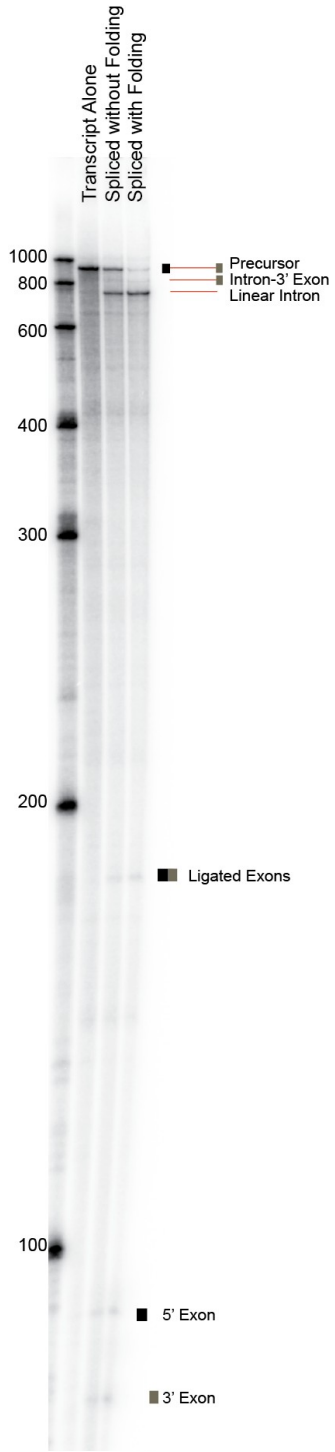


Figure 34. Effects of the RNA folding step on WT *C.te*.I1 self-splicing.

Intron precursor was either heated to 53°C (without folding lane) or heated to 90°C for 1 min, 75°C for 5 min, followed by an incremental cooling to 53°C over 15 min (with folding). Self-splicing reactions were initiated through the addition of 2X splicing buffer (80 mM Tris-HCl, pH 7.5, 1 M NH_4Cl, 200 mM $MgCl_2$) and stopped after 10 minutes by the addition of EDTA.



*C.te.*I1 suggests that a substantial portion of the population of RNA molecules is capable of finding active conformations without optimized conditions. As kinetic traps are normally avoided *in vivo* through functions of the IEP and other host proteins (Pyle et al. 2007), the fact that *C.te.*I1 readily adopts an active conformation *in vitro* in a variety of conditions such as low magnesium concentrations (5-10 mM), near physiological temperatures and without the aid of a "folding" program and gradual cooling may indicate overall adaptations of the intron to allow robust splicing *in vivo* without a self-encoded maturase protein. While the intron may still utilize host factors, such as RNA binding proteins and various chaperone proteins for *in vivo* splicing, *C.te*.I1 may be adapted to splice at more physiologically relevant conditions such that it does not require the aid of an IEP to splice *in vivo*.

3.3.3 Secondary Structure Verification

As the predicted secondary structure of *C.te*.I1 possessed a number of unique structural features that would be located at or near the active site of the intron (Dai et al. 2008; Toor et al. 2008), structural mutagenesis was performed to verify the proposed structure of D1. Alternate foldings of the distal end of the Id stem are possible and are shown alongside the predicted structure in Figure 35. Site-directed mutagenesis of the I(ii) stem (Stem 1 in Figure 35) was carried out to verify the helical pairing that essentially bisects the intron structure Disrupting either side of the stem eliminated splicing by disrupting the overall structure of the intron while a compensatory mutant between the two sides of the stem was found to restore function, thus verifying the pairing in this helix (Figure 36). The α/α′ pairing was also verified (Figure 36). Self-splicing assays for these constructs were carried out under 30 mM $MgCl_2$, as using the

Figure 35. Alternate RNA D1 secondary structures for *C.te*.I1

The Domain 1 secondary structure of *C.te*.I1 is shown along with alternate secondary structures that are possible at the distal end of D1. Locations of the various stems mutated for structural verification experiments are shown on each of the potential structures (highlight with light grey boxes and numbered). The location of the IBS1-EBS1 interaction and α-α′ are shown. The possible EBS2 and IBS2 sequences are also indicated. The verified 5′ splice site is indicated by a black star, while a black diamond indicates the standard 5′ GUGYG intron boundary.

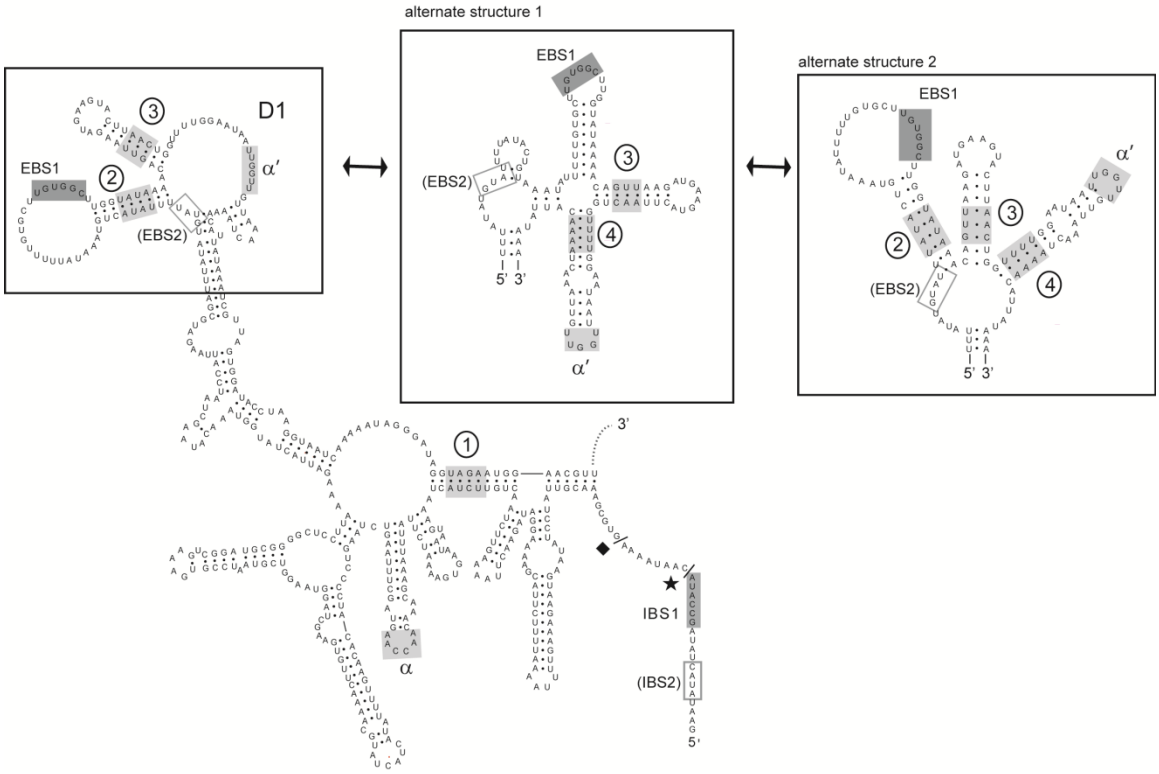
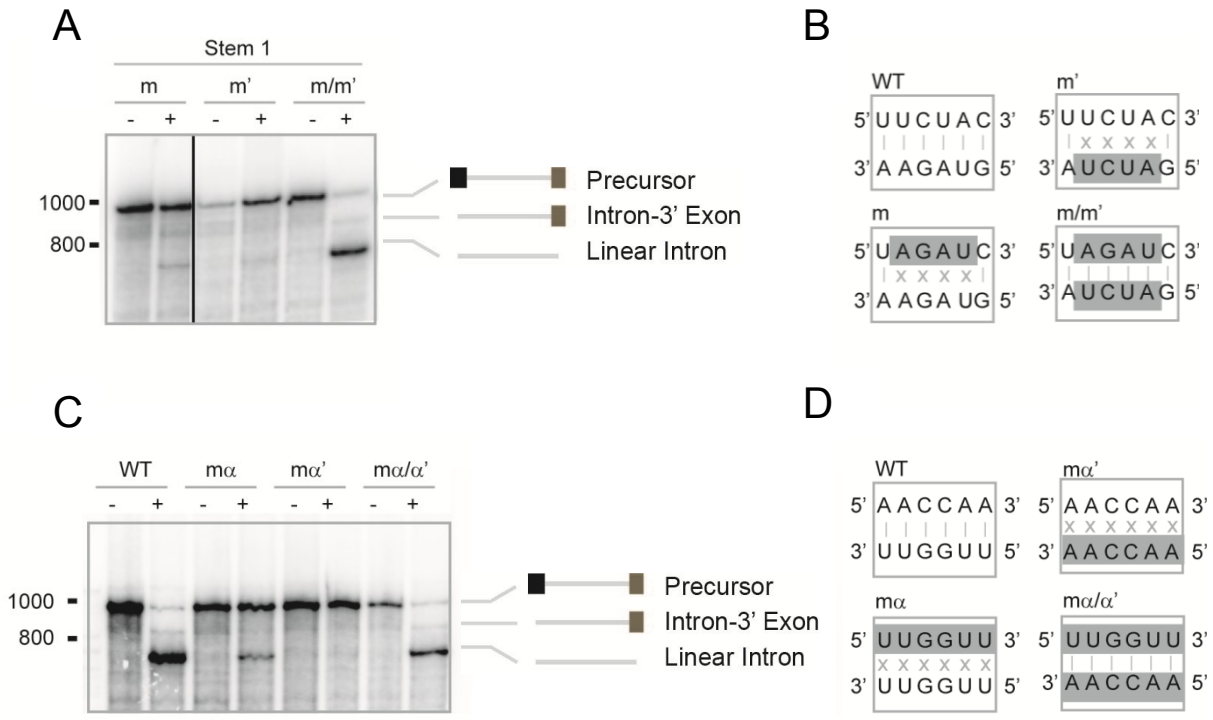


Figure 36. Verification of the I(ii) stem and α-α′ interaction.

(A) Stem I (ii) verification (shown as Stem 1 in Figure 35). Mutation of either the 5′ (m) or 3′ (m′) side of the stem shows complete elimination of self-splicing in 100 mM $MgCl_2$. The compensatory mutation (m/m′) combining both the 5′ and 3′ mutations restores splicing to WT levels. (B) Schematics of Stem I(ii) mutations. (C) Confirmation of the α-α′ pairing. Self-splicing conditions of 30 mM $MgCl_2$ (rather than 100 mM) were used to reveal the disruption and rescue. (D) Schematics of the mutations shown in panel (C).



100 mM $MgCl_2$ used in the other self-splicing reactions compensated for the mutation of α. Disruption of the α sequence had a notable disruptive effect compared to WT, while the α′ disruption completely eliminated splicing. Importantly, when the two mutations were combined, the compensatory mutation (mα/α′) restored splicing to wild-type levels, supporting the sequence specific pairing interaction between these two sequences.

In intron classes D, E and F, the α′ sequence is at the distal end of a stem-loop, and such a structure is observed in an alternate folding of *C.te.*I1 (Stem 4 in Figure 35). Although this feature is not seen in the α-lineage of Class B introns, either the typical α-lineage structure or the stem loop structure typical of D, E and F introns can be formed in introns belonging to the β-lineage. As such, the potential helical pairing of the stem was tested in *C.te.*I1; however, none of the mutations affected splicing at discernible levels (Figure 37). As disruption of the hypothetical pairing does not disrupt splicing, it shows that these sequences are not involved in interactions that are critical for the function of the ribozyme and these sequences are likely not paired. This supports the proposed secondary structure and is consistent with other structures of class B introns.

The helical pairing of the β-lineage specific stem (Stem 3 in Figure 35) was also tested. Similar to stem 4, mutations of either side of the helix do not disrupt splicing (Figure 37). However, as no alternate structures are predicted for this region and the feature is conserved within the β-lineage of Class B introns, I have chosen to maintain the drawing of this feature in the secondary structure but note that the structure has no known function.

The EBS1 stem pairing was tested by the simultaneous mutation of four base pairs (Stem 2; Figure 35 and Figure 38). In this case, mutation of the 5′ strand had a

Figure 37. Verification of stems 3 and 4.

(A) Mutations of the 5′ and 3′ sides of the additional stem present following the EBS1 stem loop (Stem 3, Figure 35) show no effect on splicing. Mutation of the alpha prime stem sequences (Stem 4, Figure 35) again show no effect on splicing, suggesting either that a stem is not formed by this region or that the stem is non-essential at the elevated magnesium concentrations tested. (B) Sequences of the specific mutations used during disruption of Stem 3 and Stem 4.

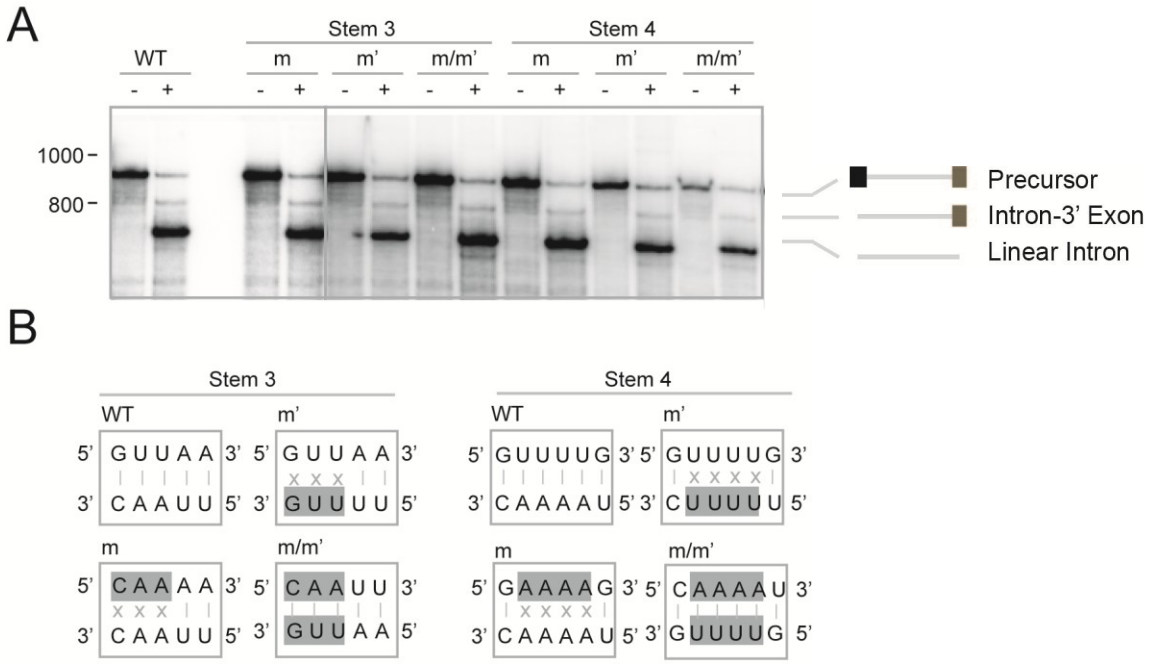
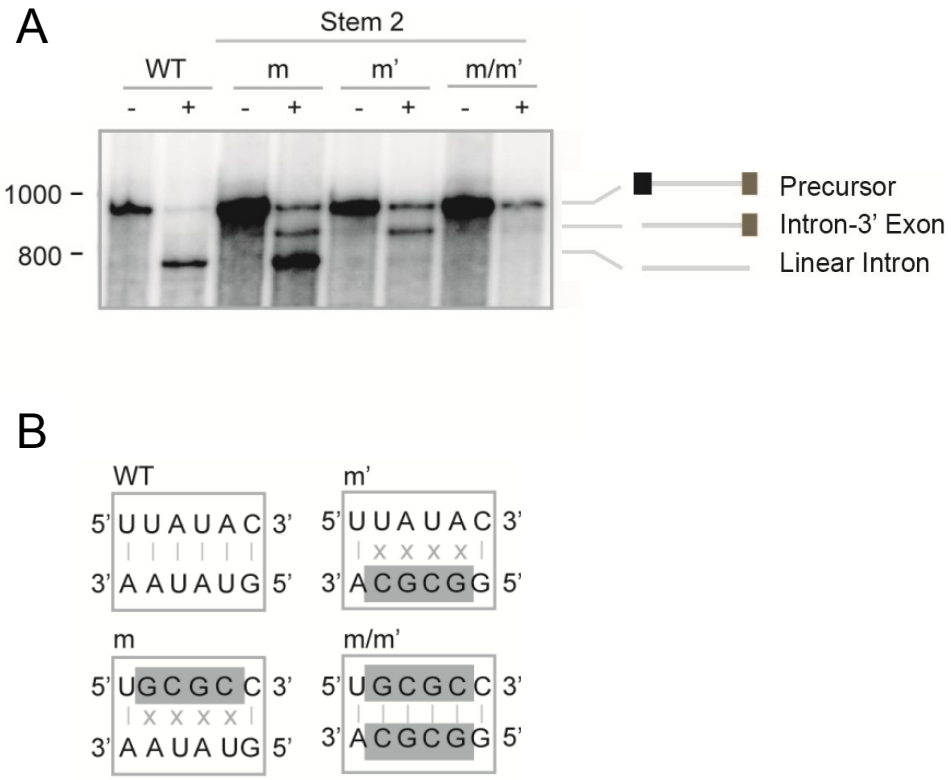


Figure 38. Effects of mutations in the EBS1 stem.

(A) Mutation of the 5′ side (m) of the EBS1 stem loop (Stem 2, Figure 35) results in a slight disruption of splicing. Mutation of the 3′ side (m′) of the stem further disrupts splicing. The combination of 5′ and 3′ mutations eliminates splicing rather than restoring it. (B) Schematics of the mutated helix corresponding to the mutants spliced in panel (A).

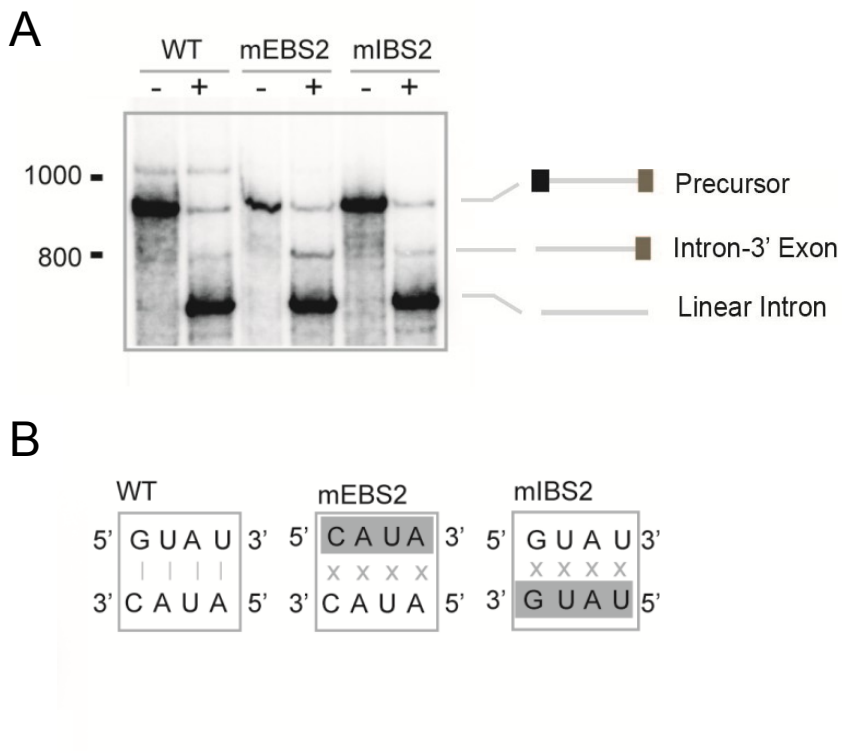


modest effect on splicing, while mutation of the 3′ strand produced a severe defect that is blocked for the second step of splicing. Rather than rescuing the reaction, the compensatory mutations abolished all reactivity of the ribozyme, thus failing to support the proposed pairing. In further experiments, I showed that the large loop containing EBS1 cannot be involved in important pairings as its mutation and deletion have little effect (below). This leaves few options for the EBS1 stem other than the predicted stem (denoted Stem 2 in Figure 35). It is possible that these mutations (Stem 2) disrupt critical interactions within the ribozyme, in addition to the helical pairing being tested. In some other introns this stem contains the ω′ motif that docks the EBS1 motif to helix Id near the ζ-ζ' interaction (Toor et al. 2008). If a similar interaction exists for *C.te.*I1, then a more subtle mutation may be necessary to test the hypothesized pairing.

Finally, the possibility of an EBS2 motif and structure was tested (Figure 39). A typical EBS2 motif was not initially evident in the secondary structure; however, an alternate pairing allowed a modest 4 bp EBS2-IBS2 rather than the usual 6 bp interaction (Figure 35). As such, mutations were made both within the potential EBS2 and potential IBS2 sequences. Neither mutation had a noticeable effect on splicing. Thus, there is no evidence for the EBS2 structure or EBS2-IBS2 pairing. The structure presented in Figure 25 (primary structure in Figure 35) is considered to be the most likely secondary structure of *C.te.*I1, despite the ambiguities that still remain after these experiments. Comparison with the set of ten mitochondrial IIB1 introns possessing 5′ terminal extensions (Li et al. 2011b) lend support to this structure and will be discussed in the discussion section.

Figure 39. Mutagenesis of potential EBS2 and IBS2 sequences.

(A) Mutations of potential EBS2 and IBS2 sequences (locations shown in Figure 35) have no effect on the ribozyme mediated self-splicing reaction, thus suggesting that EBS2-IBS2 pairing is not necessary for 5′ splice site selection. This suggests that the structure lacking the EBS2 motif is a correct structure. (B) Schematics of the mutations used in the self-splicing shown in panel A.



3.3.4 C.te.I1 *Recognizes the 5′ Splice Site Through a Non-Canonical EBS1-IBS1*

For most IIA and IIB introns, 5′ exons are positioned by two pairing interactions, exon binding sites 1 and 2 (EBS1 and 2), located in domain 1, and intron binding sites 1 and 2 (IBS1 and 2), located immediately upstream of the splice site (Jacquier and Michel 1987). For the *C.te*.I1 intron, a potential EBS1 sequence is present in a loop at the distal end of domain 1; however, an EBS2 motif is not evident in the secondary structure model of *C.te.*I1, and is not supported by the mutagenesis data. The putative EBS1 sequence lies in a 24 nt loop that is dramatically larger than the typical EBS1 loop size of 8-9 nt that is standard for β-lineage of Class B introns (Figure 27).

To test whether the candidate EBS1 and IBS1 sequences pair with each other and are responsible for 5′ splice site selection, the sequences were mutated alone and in combination (Figure 40). Consistent with the candidate sequences being involved in the EBS1-IBS1 pairing, mutation of either sequence blocked self-splicing while the compensatory mutations restored splicing, thus demonstrating a bona fide pairing interaction between the two sequences. This pairing is located adjacent to the shifted 5′ splice site that was observed both *in vivo* and *in vitro*.

The EBS1 and IBS1 mutations also shows that when either the IBS1 or EBS1 sequence is mutated, slowly migrating bands appear that could correspond to either intron lariat or circles. These bands are not formed for the compensatory mutation, suggesting that they are due to a side reaction caused by mis-pairings resultant from the mutated EBS1 and IBS1 sequences. RT-PCR of total reaction products for the mIBS1 construct revealed a junction in which the 3′ end of the intron is linked to a cryptic splice site 38 nt upstream of the usual *C.te*.I1 5′ splice site, suggesting the slow migrating product

observed for these mutations are intron circles rather than lariats. Although production of ligated exons is not efficient for these mutants, RT-PCR of ligated exons for the mIBS1 intron also show the use of a cryptic splice site 2 nt downstream of the standard *C.te*.I1 5′ splice site, due to a weak pairing created between the EBS1 sequence and the mutated 3′ exon sequence at this site. This illustrates the importance of complementarity between these two sequences in specifying the location for splicing.

In addition to the verified EBS1-IBS1 pairing at the observed splice site, another potential 6 nt pairing exists between sequence within the enlarged 24 nt EBS1-containing loop (sequence B′) and the sequence immediately upstream of the canonical 5′ GUGYG intron boundary sequence (sequence B). To investigate whether this potential pairing was also functional, site-directed mutagenesis was again performed; however, mutagenesis of sequences B and B′ had no noticeable effect on the self-splicing reaction (Figure 40). Hence neither of these two sequences have critical sequence specific roles. As such, one may conclude that the single 6 bp EBS1-IBS1 is the only pairing that is necessary for selection of the shifted 5′ splice site.

Surprisingly when the 8 nt sequence between the standard 5′ GUGYG intron boundary and the actual 5′ splice site for *C.te*.I1 was deleted (Δ 8 nt IBS1, Figure 41), splicing was virtually eliminated. This suggests that although the sequence of the region is not important as the mB construct has little effect on splicing that a spacer is necessary in this region in order to correctly position the EBS1-IBS1 pairing and allow for efficient splicing. As this deletion construct positions the EBS1-IBS1 pairing at the canonical 5′ GUGYG intron boundary, it also indicates that the intron has adapted not only to utilize a

Figure 40. Mutagenesis of 5′ splice site recognition elements.

(A) Self-splicing assays for mutations in the IBS1-EBS1 interaction and the hypothesized pairing B-B′. Lanes "+" and "-" denote incubation in self-splicing buffer or in self-splicing buffer lacking magnesium. For each of the EBS1-IBS1 mutations, slow-migrating product bands are formed that could potentially be intron lariat or circle (asterisks). (B) Mutations and compensatory mutations for the IBS1-EBS1 and B-B' constructs tested in Panel A. The EBS1 stem-loop is shown on top and the 5′ exon-intron junction is below. The star and diamond denote the actual and predicted intron boundaries, respectively.

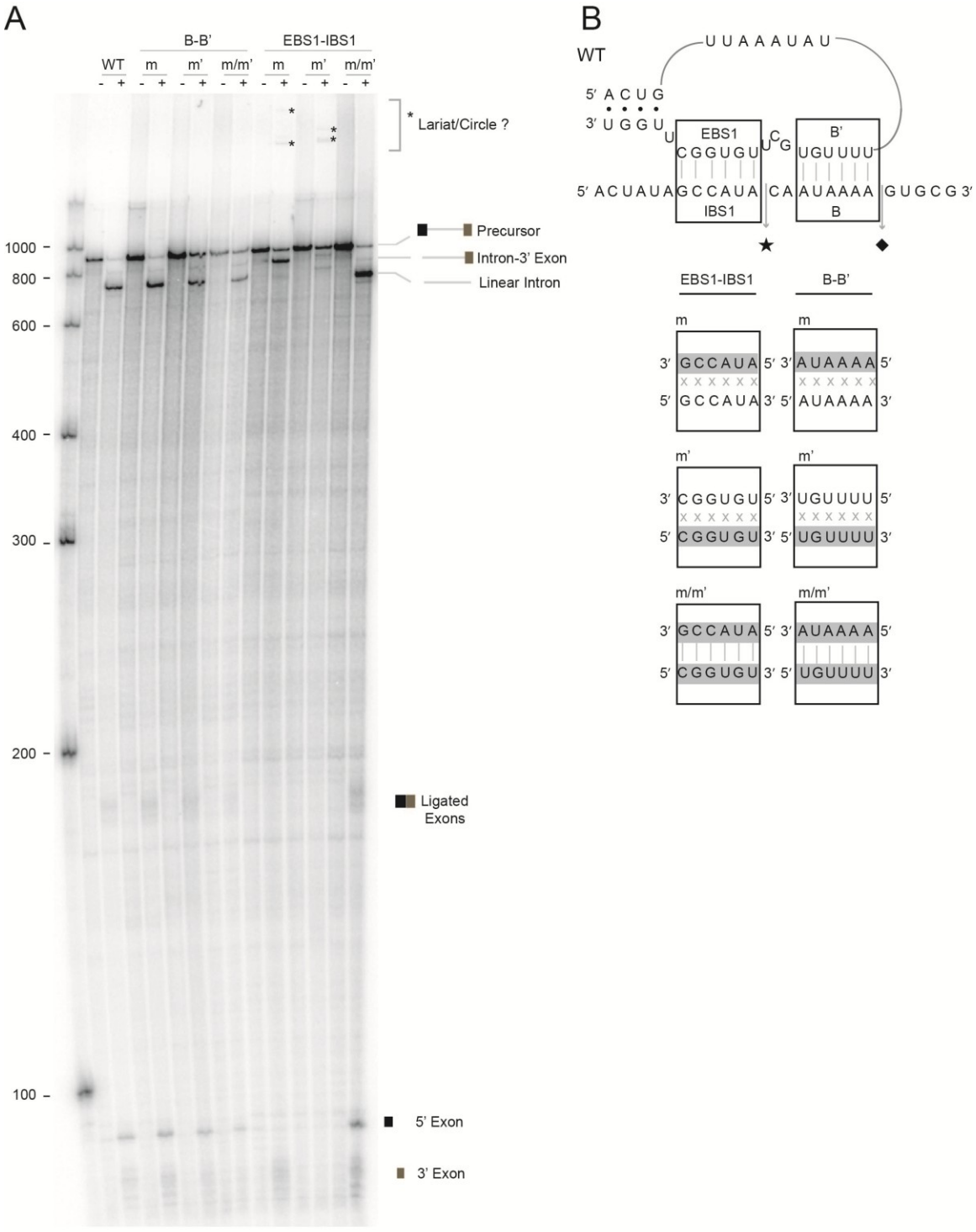
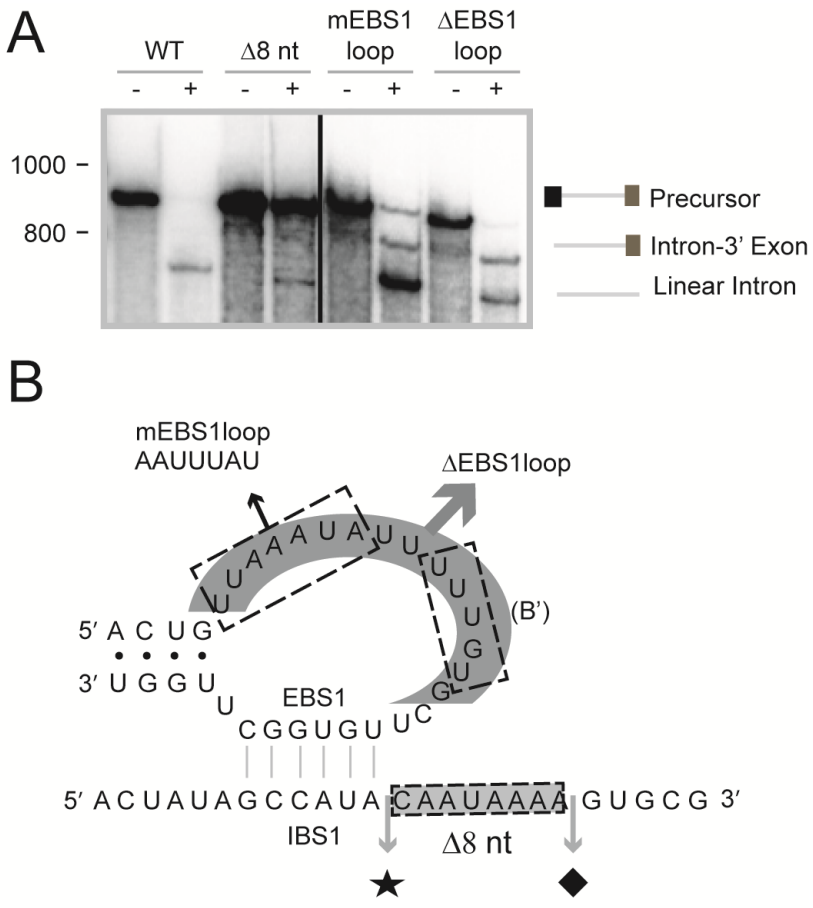

A
B-B'
EBS1-IBS1
WT
m
m'
m/m'
m
m'
m/m'
- + - + - + - + - + - +
* Lariat/Circle ?
Precursor
Intron-3' Exon
Linear Intron
1000
800
600
400
300
200
100
Ligated Exons
5' Exon
3' Exon
B
WT
U U A A A U A U
5' A C U G
3' U G G U
U
EBS1
C G G U G U
U C G
B'
U G U U U U
5' A C U A U A G C C A U A
IBS1
C A A U A A A A
B
G U G C G 3'
★
◆
EBS1-IBS1
B-B'
m
m
3' G C C A U A 5' 3' A U A A A A 5'
5' G C C A U A 3' 5' A U A A A A 3'
m'
m'
3' C G G U G U 5' 3' U G U U U U 5'
5' C G G U G U 3' 5' U G U U U U 3'
m/m'
m/m'
3' G C C A U A 5' 3' A U A A A A 5'
5' C G G U G U 3' 5' U G U U U U 3'

Figure 41. Mutagenesis of other elements potentially involved in 5′ exon recognition. (A) Self-splicing assays of the mutations ΔEBS1loop, mEBS1loop, and Δ8nt IBS1. (B) Schematic of mutations. The sequence mutated in mEBS1loop is shown by the dotted box and arrow, and the sequences deleted in the ΔEBS1loop and Δ8nt IBS1 mutations are shown by grey shading. In addition, the positions mutated for B′ are shown.



shifted 5′ splice site but it is unable to splice efficiently adjacent to the canonical 5′ intron boundary.

It seemed likely that additional sequences present in the EBS1loop, would be responsible for the altered properties of 5′ splice site recognition. As such constructs were made that deleted the EBS1loop down to a standard β-lineage loop (ΔEBS1loop) and that mutated the additional 7 nt of sequence other than B′ and EBS1 (mEBS1loop). Interestingly, the ΔEBS1loop deletion only modestly affected splicing, resulting in slightly elevated levels of intron-3′ exon intermediate (Figure 41), indicating an inhibition of the second step of the splicing reaction. The mEBS1loop construct showed a negligible influence on the splicing reaction (Figure 41). Together, these experiments suggest that the enlarged loop does not have a crucial role in the ribozyme activity since it can be mutated and deleted with little effect on splicing; however, accumulation of the intron-3′ exon intermediate product indicated that the region may play a role in the second reaction step of splicing. These experiments also fail to support the alternate secondary structures presented in Figure 35 as the alternate structures predict critical stem structures formed by the mutated and deleted sequences.

3.3.5 C.te.*I1 uses IIA-like Mechanism of 3′ Splice Site Selection*

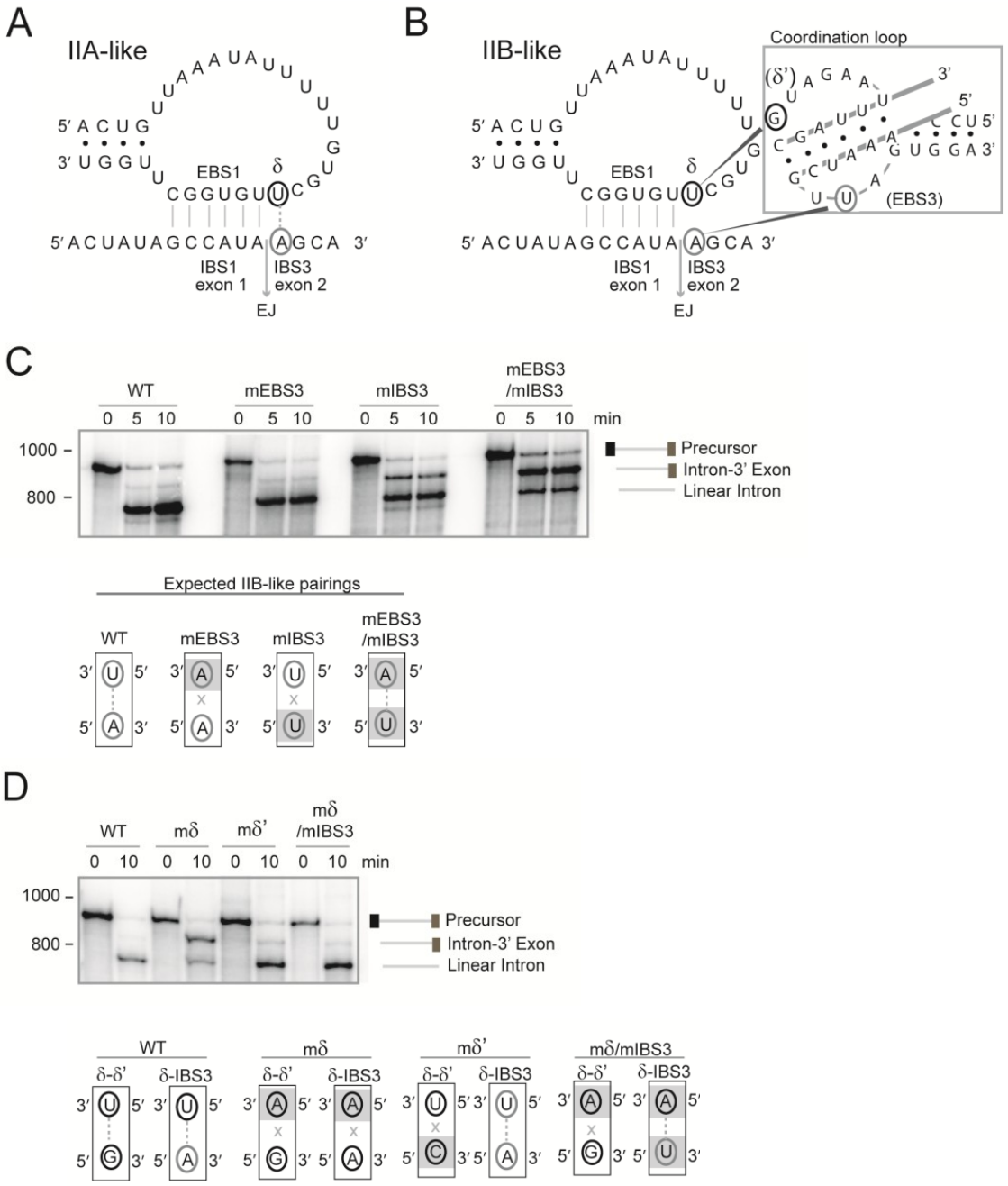
IIB and IIC introns, including bacterial B introns, recognize their 3′ exons through a conserved set of interactions involving the EBS1 stem-loop motif and the coordination loop (Figure 42). The IBS3-EBS3 interaction is a single base pair interaction between EBS3 in the coordination loop and the first nucleotide of the 3′ exon. The coordination loop itself is anchored by the δ-δ′ base pairing between the nucleotide adjacent to EBS1 (δ) and a nucleotide on the 5′ side of the coordination loop (δ′). In contrast, IIA introns

lack the coordination loop structure and recognize the 3′ exon through a direct pairing between δ and the first nucleotide of the 3′ exon (called δ′ rather than IBS3 by standard IIA conventions; however it is labelled in Figure 42 as IBS3 to be consistent with the otherwise IIB conventions). A secondary structure motif resembling the coordination loop can be found for the *C.te.*I1 intron, and since the intron is assumed to be derived from a Class B intron, the initial assumption was that the 3′ exon should be recognized through the IIB intron mechanism.

To test the predicted interaction, mutations were made between the potential IBS3 and EBS3 interactions. Mutation of the IBS3 residue in the 3′ exon indeed led to an accumulation of intron-3′ exon intermediate, as would be expected for a deficiency in the second step of splicing (Figure 42). However, mutation of the putative EBS3 residue had no effect on splicing, arguing against its involvement. Surprisingly, a compensatory mutation expected to restore the IBS3/EBS3 pairing showed an even greater splicing defect than for the IBS3 mutation (Figure 42). This strongly argues against the formation of a typical EBS3-IBS3 pairing. An additional element in the IIB mechanism of 3′ splice site recognition is a pairing between δ and δ′, which is necessary for correct positioning of the EBS3 nucleotide (mentioned above). Mutation of δ had a substantial effect on the accumulation of intron-3′ exon intermediate; however, mutating δ′ had no effect on the splicing reaction compared to the wild-type sequence and did not rescue the δ mutation (Figure 42). Together, the data strongly argue against the typical IIB mode of 3′ exon recognition (δ-δ′ and IBS3-EBS3) even though apparent motifs exist in the secondary structure model.

Figure 42. Mutagenesis and self-splicing assays of 3′ exon recognition elements.

(A) Summary of the mechanisms for 3′ exon recognition for IIA introns, applied to the *C.te.*I1 intron sequence. For IIA introns, the nucleotide directly upstream of EBS1 base pairs with the first position of the 3′ exon, and is called the δ-δ′ pairing, although we call it the δ -IBS3 pairing here to avoid confusion. (B) For IIB introns, the nucleotide directly upstream of EBS1 base pairs with a position in the "coordination loop" of domain I (δ-δ′), while a nearby residue pairs with the first nucleotide of the 3′ exon (EBS3-IBS3). (C) Self-splicing assays of EBS3-IBS3 mutants, showing that mutation of EBS3 does not affect splicing, while mutation of IBS3 partially disrupts the second step of splicing (greater amount of intron-3′ exon intermediate). Mutant sequences shown below the self-splicing gel indicate the predicted pairings and disruptions based on the presumed IIB mechanism of 3′ exon recognition. Mutated nucleotides are indicated by grey shading. (D) Self-splicing assays of δ-δ′ mutants, showing that mutation of δ′ has no impact on self-splicing, while mutation of δ disrupts the second step of splicing. The compensatory mutations of δ and IBS3 rescue the second step deficiency, supporting a IIA-like mechanism of 3′ exon recognition. Shown below the splicing gel are the expected pairings or disruptions for each of the mutants for both IIA (δ-IBS3) and IIB (δ-δ′) –like mechanisms. Mutated nucleotides are indicated by grey shading.


A
IIA-like
5′ A C U G
3′ U G G U
EBS1
δ
C G G U G U U C G
5′ A C U A U A G C C A U A A G C A 3′
IBS1
exon 1
IBS3
exon 2
EJ
B
IIB-like
Coordination loop
(δ′) A G A A
3′
5′
C U 5′
G A A G U G G A 3′
(EBS3)
C
WT
mEBS3
mIBS3
mEBS3
/mIBS3
0 5 10
min
1000 –
800 –
Precursor
Intron-3′ Exon
Linear Intron
Expected IIB-like pairings
D
WT
mδ
mδ′
mδ
/mIBS3
0 10
min
1000 –
800 –
Precursor
Intron-3′ Exon
Linear Intron
δ-δ′
δ-IBS3
mδ/mIBS3

In the absence of a IIB-like mechanism, the possibility of a IIA-like mechanism was considered. In this scenario, the nucleotide adjacent to EBS1 (δ) should pair with the first nucleotide of the 3′ exon (δ′ by IIA convention but labelled IBS3 in Figure 42 for clarity). Indeed, the combined mutations of δ and IBS3 restored the splicing defects seen for the individual δ and IBS3 mutations, showing that the two nucleotides form a base pair, and indicating that *C.te.*I1 uses a IIA-like mechanism to recognize its 3′ exon. This is another divergence of the *C.te.*I1 ribozyme from its ancestral form, and could be an adaptation to its role in alternative splicing.

Interestingly, examination of the exon sequences of *C.te.*I1 reveals that all four 3′ exons recognized *in vivo* begin with the sequence AGCA (Figure 43). The GCA bases are in fact complementary to the nucleotides adjacent to δ, raising the possibility that the δ-IBS3 pairing may extend up to four base pairs. To test this possibility, the positions adjacent to δ and IBS3 were mutated. Mutation of either strand led to a modest but reproducible increase of intron-3′ exon intermediate, suggesting a perturbation in the second step of splicing as would be expected for a mutation in a functional component near the 3′ splice site (Figure 43). The compensatory mutation of the trinucleotide base pairing restored splicing to WT levels. Thus, these mutations suggest that the δ-IBS3 interaction in *C.te.*I1 is longer than the one base pair typical of IIA introns and may extend up to four base pairs.

Individual mutations of the nucleotides proposed to be involved in this extended IIA-like mechanism show only slight effects when mutated aside from the G at the second position of the 3′ exon (Figure 44) that shows a marked effect similar (perhaps stronger) to that of mutating IBS3 or δ. Mutation of the corresponding C in the EBS1loop

Figure 43. Three nucleotides adjacent to δ appear to form an extended δ-IBS3 interaction.

(A) Diagram of the potential extended IIA-like 3′ exon recognition and the mutations tested in panel A. In each of the mutations the δ-IBS3 pairing is maintained. (B) Self-splicing assays of the *C.te*.I1 intron mutated for the three base pairs adjacent to δ-IBS3. (C) Sequence of the exon junctions observed for each of the four splice forms (SF) *in vivo*. The first four base pairs are conserved in all four of the alternative 3′ exons.

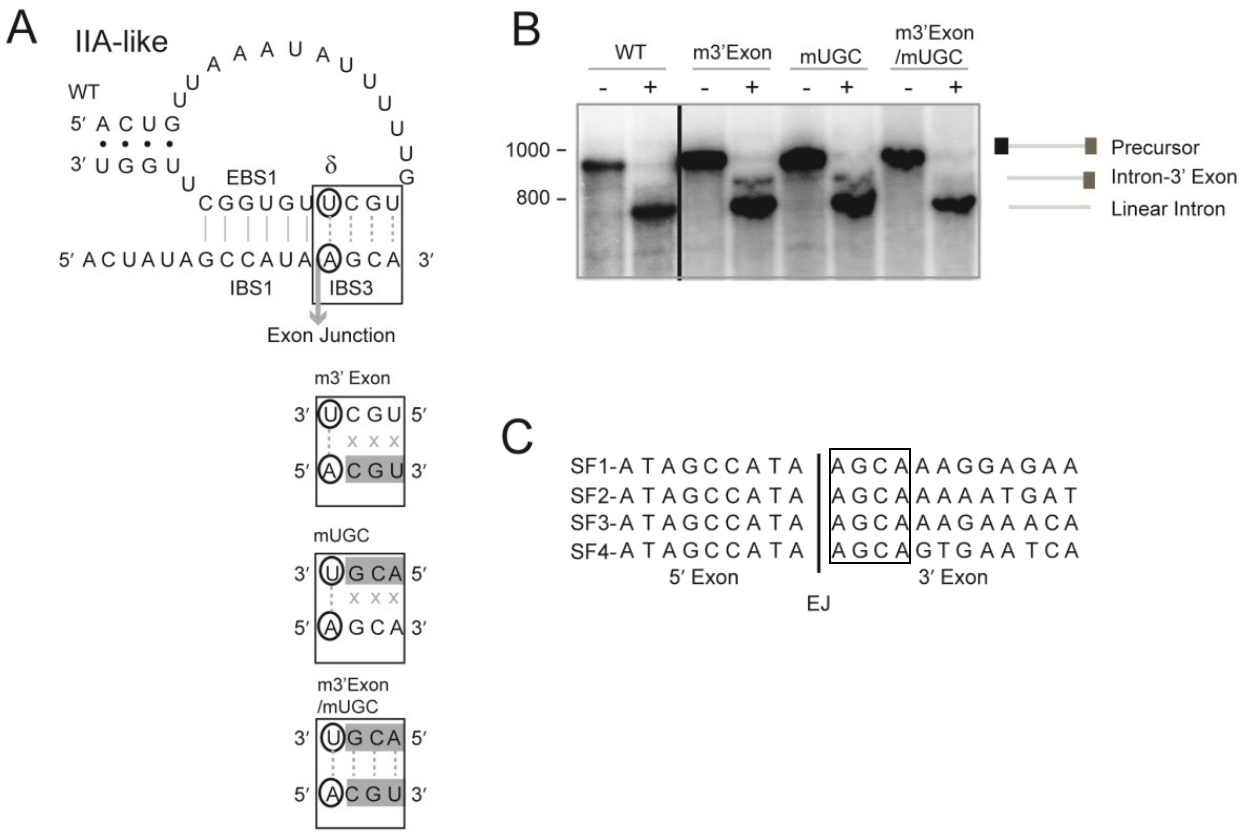
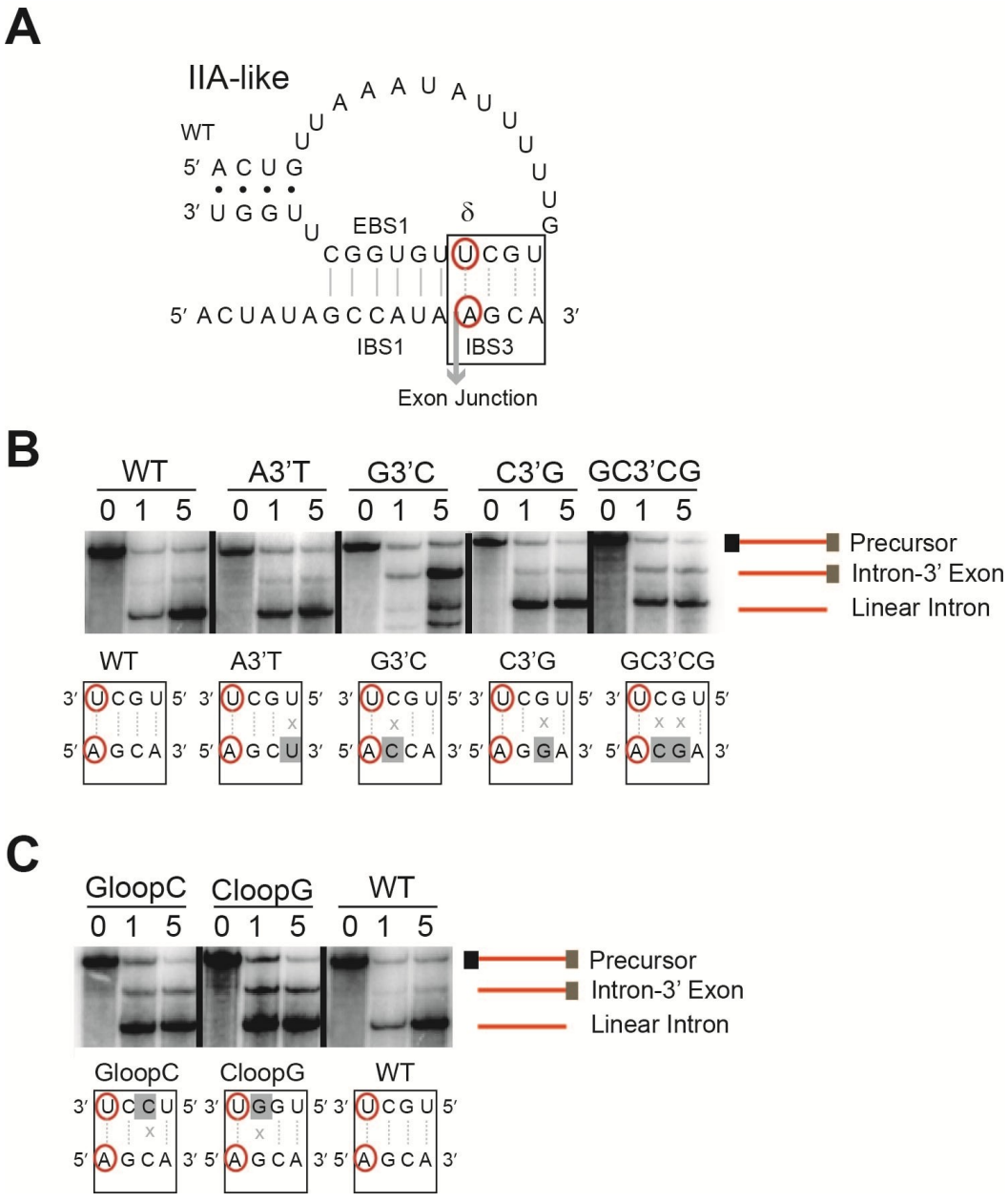


Figure 44. Contributions of single nucleotides in an extended IIA-like mechanism of 3′ splice site recognition.

(A) Shows the proposed extended IIA-like mechanism of 3′ exon recognition (B) Self-splicing assays of single and double nucleotide mutations in the 3′ exon. Schematics of mutations and predicted effects are shown below the splicing panel. (C) Self-splicing assays of single nucleotide mutations in the EBS1 loop sequence. Schematics of mutations and predicted effects are shown below the splicing panel. Mutants are named such that nucleotides indicated on the left of the mutant name indicate the position that was mutated, while the final nucleotide name indicates what it was mutated to (ie. A3′T indicates that the 4^{th} position of the 3′ exon A is mutated to a T, while GloopC indicated that the G in the UGC sequence in the loop was mutated to a C).


A
IIA-like
WT
5′ A C U G
3′ U G G U
EBS1
δ
C G G U G U U C G U
5′ A C U A U A G C C A U A A G C A 3′
IBS1
IBS3
Exon Junction
B
WT
A3'T
G3'C
C3'G
GC3'CG
0 1 5
Precursor
Intron-3' Exon
Linear Intron
C
GloopC
CloopG
WT
0 1 5
Precursor
Intron-3' Exon
Linear Intron

also showed a strong accumulation of intron-3′ exon intermediate. No compensatory mutations were made between these nucleotides, however and as such no direct evidence exists suggesting that their pairing is the reason for the marked effects on the splicing reaction. Mutation of the C in the third position of the 3′ exon and the corresponding G in the EBS1loop also appear to have an effect while mutating the A at the fourth position of the 3′ exon had no observable effect on the splicing reaction in the conditions tested.

3.4 Discussion

In this section, I will discuss two main points related to the structural features and catalytic properties displayed by *C.te*.I1. First, I will analyze how *C.te*.I1 differs from standard Class B introns and how those novel features may be adaptations to the intron's unique role in alternative splicing. Secondly, I will compare *C.te*.I1 to a subgroup of mitochondrial rRNA introns that have been found to possess 5′ terminal insertions and share many other structural features with *C.te*.I1. Together these comparisons will explain why *C.te*.I1 exhibits many of the properties it does and how the novel structure features function to adapt it to its role in alternative splicing.

3.4.1 *Comparison of Standard Class B Introns to* C.te.*I1*

Typical bacterial class B introns possess a conserved secondary structure that can be subdivided into two lineages, α and β based on shared structural features (Stabell et al. 2009). Both lineages possess a bulged EBS2 sequence followed by additional stem loop structures. Following the EBS2 motif, an 8-10 bp stem possessing an internal bulge is capped by an 8-11 nt loop that contains the EBS1 and δ sequences. The β lineage introns possess an 8-9 nt loop, while the α-lineage possesses an 11 nt loop. The β-lineage then

possesses an additional stem loop sequence that is inserted immediately downstream of the EBS1 stem.

Class B introns fall under the broad umbrella of the IIB RNA structural type and as such are expected recognize their 5′ splice sites through both EBS1-IBS1 and EBS2-IBS2 pairings. Recognition of the 3′ splice site during the second step of splicing is then expected to be coordinated by the δ-δ′ pairing which positions EBS3 to pair with the first nucleotide of the 3′ exon, IBS3.

The alternatively splicing intron, *C.te*.I1, however, appears to lack an EBS2 sequence and differs from the expected Class B consensus in the region of EBS1. The stem lacks the internal bulge sequence and possesses an enlarged loop sequence of 24 nucleotides rather than the typical 8-9 nt loop. D6 also differs from consensus as it contains an atypical bulged A region and lacks the GNRA tetraloop that normally caps the stem. *C.te*.I1 also lacks the ORF sequence normally present within D4. Despite these structural anomalies, most long range tertiary interactions appear to be maintained and *C.te*.I1 maintains robust splicing *in vitro*, albeit through the use of a hydrolysis pathway rather than through the typical branching pathway.

Some of the novel structural properties of *C.te*.I1 contribute to the intron's ability to alternatively splice. Specifically the intron utilizes a novel EBS1-IBS1 pairing that specifies the shifted 5′ splice site that eliminates the upstream ORF stop codon and allows each of the alternate downstream ORFs to be fused in frame. Despite the use of the shifted splice site a potential pairing (B-B′) is present at the canonical 5′ GUGYG intron boundary sequence. Mutations that disrupt the EBS1-IBS1 sequence, however, do not result in splicing at the canonical splice site and the Δ8 nt IBS1 mutation that positions

the IBS1 sequence adjacent to the canonical boundary splices at extremely low efficiency. Therefore not only is the intron adapted to utilize a novel EBS1-IBS1 pairing that shifts the 5′ splice site and eliminates the upstream stop codon, it is unable to splice at the canonical boundary. It seems then that the intron has adapted to exclusively utilize the shifted 5′ splice site through the novel EBS1-IBS1 pairing and therefore is only capable of splicing or alternative splicing resulting in correct exon ligation. The use of the novel splice site and the inability to utilize the canonical EBS1-IBS1 argue against these features being mere coincidence and suggest that the novel features represent beneficial changes to the intron. Additionally, the intron has lost the EBS2 sequence and motif structure. As the IBS2 sequence should lie adjacent to the IBS1 sequence in the 5′ exon, loss of this pairing interaction may have been critical to the adaptation of the intron to utilize an alternate EBS1-IBS1 by relaxing the need for increased sequence complementarity and the positional requirement for the IBS1 to lie between the canonical GUGYG and the IBS2 sequence.

In addition to the adaptations of the intron at the 5′ splice site (shifted EBS1-IBS1 pairing and loss of the EBS2 motif), the intron also diverges from standard class B introns in the recognition of the 3′ splice site. Mutagenesis indicated that instead of pairings between EBS3 in the "coordination loop" and IBS3, the first nucleotide of the 3′ exon, the δ nucleotide adjacent to EBS1 pairs with the first nucleotide of the 3′ exon. Reasons why the intron would have evolved to utilize a IIA-like mechanism rather than the IIB mechanism typically used for Class B introns are not completely clear. However, the mutagenesis data suggest that the pairing extends beyond a single nucleotide pairing and might extend up to 4 nt in length. The use of the extended IIA-like mechanism of

recognition at the 3′ splice site may be beneficial *in vivo* during the alternative splicing reactions. Each of the alternate 3′ exon sequences begin with the same 4 nt AGCA sequence and an extended pairing may, therefore, allow for better recognition.

Although experimentally unconfirmed, bioinformatic predictions of promoter and terminator sequences (discussed in Chapter 2) suggest that alternative splicing occurs *in vivo* through a *trans*-splicing mechanism with each of the downstream D5/6 and 3′ exon ORFs encoded on separate transcripts. In the case of *trans*-mediated splicing, the extended IIA-like recognition of the 3′ exon may function to provide better recognition of the alternate 3′ exon transcripts *in vivo*.

It is possible that *C.te*.I1 uses its unique ribozyme structure to splice without the IEP *in vivo*, as no other group II intron related RT-like proteins are encoded within the *C. tetani* genome. The *in vitro* self-splicing studies lend plausibility to the idea as *C.te*.I1 displays robust and rapid splicing. In addition to splicing at near physiological $MgCl_2$ concentrations which has only been shown for a few group II introns [*O.i*.I1, the crystallized IIC intron from *O. iheyenis* and the Pl.LSU/2 intron from *Pylaiella littoralis* (Costa et al. 1997b; Toor et al. 2008)], the intron has been shown to be capable of splicing at physiological temperatures. These combined unique properties of this intron for self-splicing raise the possibility that *C.te*.I1 may be capable of self-splicing *in vivo* and may not need to rely on host-encoded proteins to aid in the splicing reaction. However, it still remains possible that a yet unidentified host splicing-factor may play a role in the alternative splicing of this unique intron *in vivo*.

3.4.2 Comparison of C.te.*I1 to Mitochondrial IIB1 Introns with 5′ Extensions*

The intron, *C.te*.I1 resembles a recently described class of IIB1 mitochondrial group II introns that possess 5′ terminal insertions, which range in size from 1 to 33 nts. *C.te.*I1 possesses an eight nt insertion falling well within this range. These introns also possess several structural features like those observed for *C.te*.I1, including a shortened D6 stem, an atypical branch A motif, and the lack of an EBS2 motif (Li et al. 2011b). The introns are all located with rRNA genes and one of the introns, the *Pycnoporellus* SSU788 intron, was tested for self-splicing. Like *C.te.*I1, the intron was found to splice exclusively through the hydrolysis pathway (Li et al. 2011b). As both introns lack the bulged A in D6, splicing through a hydrolysis pathway is what would be expected.

These introns also lack the RT based IEP typical of group II introns and 4 of the 10 intron subset encode homing endonucleases of the LAGLIDADG family instead. The LAGLIDADG ORFs are expected to provide intron mobility through the DNA-based homing endonuclease pathway that has been studied for group I introns (Lambowitz and Belfort 1993; Chevalier and Stoddard 2001). Because mobility does not proceed through an RNA intermediate, intron splicing is not required for mobility. The intron sequence and splicing abilities then become inconsequential to the mobility mechanism and the lack of selective pressure suggests that retromobility-associated ribozyme characteristics may be lost.

While the branched lariat structure is typically viewed as necessary for retromobility of group II introns (Zimmerly et al. 1995; Belfort et al. 2002; Lambowitz and Zimmerly 2004), it would seem no longer necessary in introns that have lost retromobility. As noted previously (Li et al. 2011b), following the loss of retromobility

the stringency of selective pressure to maintain branching is likely relaxed and degeneration, such as the insertion of 5′ terminal nucleotides and loss of the D6 bulged A (Van der Veen et al. 1987; Jacquier and Jacquesson-Breuleux 1991) can occur.

The original observations about the set of mitochondrial rRNA introns predicted that the multiple structural anomalies were not coincidental, but causally related (Li et al. 2011b). The loss of the EBS2 motif was correlated with the loss of the branching pathway, and it was suggested that the loss of the EBS2-IBS2 pairing precedes and might even be necessary for the loss of branching (Li et al. 2011b). This hypothesis was not unprecedented as the only structural class of group II introns known to lack the EBS2-IBS2 pairing, IIC introns, are also known to initiate splicing through hydrolysis (Granlund et al. 2001; Toor et al. 2006). The occurrence of the same full set of modifications in *C.te*.I1 supports the prediction that the irregularities are causally linked, as it provides a second, independent example of the co-occurrence of all the structural anomalies.

3.4.3 Summary

*C.te.*I1 utilizes novel mechanisms of both 5′ and 3′ splice site recognition that presumably represent adaptations of the intron to its unique role in alternative splicing. The loss of IEP, retromobility properties and subsequent loss of branching may be essential to the unique properties displayed by this intron. The adaptation of the *C.te*.I1 intron structural to its role in alternative splicing illustrates the plasticity of group II introns and underscores their versatility to be utilized in affecting gene expression.

Chapter Four: **CONCLUSIONS**

A unique ORF-less group II intron, *C.te*.I1, has been discovered in the genome of the human pathogen, *C. tetani*. It possesses both an unusual intron structure and an unusual genomic arrangement. The genomic arrangement utilizes downstream copies of D5/6 to function as alternate 3′ splice sites for the intron. Alternative splicing is detected at the RNA level in the *C. tetani* strain ATCC10779 resulting in the production of four unique mRNA splice forms as well as unspliced transcript.

C.te.I1 marks the first alternative spliced group II intron that truly resembles eukaryotic alternative splicing in that it produces multiple protein coding sequences from a single gene region. Splicing of *C.te*.I1 illustrates a novel mechanism for the regulation of gene expression not previously observed in bacteria. Additionally, it represents a novel mechanism to introduce variety into bacterial surface layers.

The intron, *C.te*.I1 also represents the first bacterial group II intron that is known to splice without an IEP encoded within the intron or elsewhere in the host genome. The loss of the IEP suggests the intron has been domesticated by the host and is, in itself, an adaptation to its newly acquired role in alternative splicing. As the region is actively transcribed, ample opportunity would be present for intron mobility if an IEP was present. Loss of the IEP shows that the intron has been harnessed to perform the specific alternative splicing function in the host and ensures the overall fitness of the host as an active mobile intron could be detrimental.

The *C.te*.I1 ribozyme itself is unique and possesses a number of structural variations which serve as additional adaptations it to its role in alternative splicing. Sequencing of the alternatively spliced RNAs revealed that they utilize an altered 5′

splice site shifted 8 nt upstream of the canonical 5′ intron-exon boundary both *in vitro* and *in vivo*. Use of the shifted splice site is critical to alternative splicing of the intron as it allows for the in-frame ligation of the 3′ exon ORFs to the upstream 5′ exon ORF. It was found that use of this shifted splice site is specified by a novel EBS1-IBS1 pairing. The EBS1 sequence involved in this pairing is located in one of the most striking features of the ribozyme, an enlarged EBS1 loop that contains sequences that can pair at both the canonical 5′ intron boundary and at the observed intron boundary. Despite both possible pairings, the intron relies exclusively on the use of the EBS1-IBS1 pairing at the observed splice site, and pairing of the other potential sequence (B/B′) at the canonical boundary was not necessary for splicing *in vitro*. However, the additional sequence upstream of the canonical 5′ intron boundary is necessary to correctly position the splice site. The intron also lacks the EBS2-IBS2 pairing.

In addition to using a shifted EBS1-IBS1 pairing to specify the 5′ splice site, *C.te.*I1 has evolved a IIA-like mechanism of 3′ splice site recognition. Despite a structure that resembles the coordination loop being present in the secondary structure of *C.te*.I1, it is not used in the recognition of the 3′ splice site, instead a pairing between δ and the first nucleotide of the 3′ exon specifies the splice site. Each of the 3′ exon sequences recognized by *C.te.*I1 begin with the same AGCA sequences which can form base pairs with sequence adjacent to δ within the EBS1 loop. Mutation of each of these sequences at least partially inhibit the second step of splicing suggesting that the IIA-like mechanism of splice site recognition utilized by the intron is extended up to 4 base pairs.

All of the above listed structural features represent adaptations of *C.te.*I1 to its role in alternative splicing. The adaptations of the intron highlight the malleable nature of RNA and how it can be evolutionarily molded to perform novel functions within the cell. This work also highlights that novel mechanisms of splice site recognition can be employed by group II introns. Novel mechanisms of splice site selection may mean that some of the predicted splice sites of bacterial group II introns may be incorrect. As many bacterial group II intron sequences are located near ORF sequences but not inserted directly into the ORF sequence, use of non-canonical non-adjacent splice sites could mean that more bacterial group II introns are regulating the expression of nearby ORF sequences in some manner than would be predicted.

As it seems highly likely that *C.te*.I1 is recognizing the downstream copies of D5/6 through *trans*-splicing, this provides increased opportunities for bacterial group II introns to be involved in gene expression. Many of the bacterial group II introns identified are currently viewed as degenerated or fragmented remnants of group II introns; however, these sequences may function through *trans*-splicing even if a functional IEP is not present within the cell. As ORF-less bacterial group II introns have the capacity to be overlooked and combined with the possible functionality of introns currently viewed as degenerate, alternative splicing in bacteria may be more prevalent than the current single example identifies.

The discovery of *C.te.*I1 and its novel properties represent only one of the potential ways that group II introns can be domesticated by bacterial cells. It adds to the growing list of functions known to be carried out by RNAs and opens the door to a world of potentially unknown regulation by group II introns. Additionally, it adds support to the

hypothesis that group II introns are the evolutionary ancestors of spliceosomal introns and snRNAs and shows that alternative splicing could have originated early in eukaryotic evolutionary history, before the genesis of the spliceosome.

REFERENCES

Adamidi C, Fedorova O, Pyle AM. 2003. A group II intron inserted into a bacterial heat-shock operon shows autocatalytic activity and unusual thermostability. *Biochemistry* **42**: 3409-3418.

Alt FW, Bothwell A, Knapp M, Siden E, Mather E, Koshland M, Baltimore D. 1980. Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. *Cell* **20**: 293 - 301.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.

Ast G. 2004. How did alternative splicing evolve? *Nat Rev Genet* **5**: 773-782.

Augustin S, Müller MW, Schweyen RJ. 1990. Reverse self-splicing of group II intron RNAs *in vitro*. *Nature* **343**: 383-386.

Bahl H, Scholz H, Bayan N, Chami M, Leblon G, Gulik-Krzywicki T, Shechter E, Fouet A, Mesnage S, Tosi-Couture E et al. 1997. Molecular biology of S-layers. *FEMS Microbiol Rev* **20**: 47-98.

Baldermann C, Lupas A, Lubieniecki J, Engelhardt H. 1998. The regulated outer membrane protein Omp21 from *Comamonas acidovorans* is identified as a member of a new family of eight-stranded β-sheet proteins by its sequence and properties. *J Bacteriol* **180**: 3741-3749.

Barkan A. 2004. Intron splicing in plant organelles. in *Molecular biology and biotechnology of plant organelles* (eds. H Daniell, C Chase), pp. 281-308. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Bartel DP, Szostak JW. 1993. Isolation of new ribozymes from a large pool of random sequences. *Science* **261**: 1411-1418.

Bassi GS, de Oliveira DM, White MF, Weeks KM. 2002. Recruitment of intron-encoded and co-opted proteins in splicing of the bI3 group I intron RNA. *Proc Natl Acad Sci USA* **99**: 128-133.

Belfort M, Derbyshire V, Parker MM, Cousineau B, Lambowitz AM. 2002. Mobile introns: pathways and proteins. in *Mobile DNA II* (ed. RC N.L. Craig, M. Gellert, A.M. Lambowitz), pp. 761-783. ASM Press, Washington, DC.

Belhocine K, Mak AB, Cousineau B. 2007. *Trans*-splicing of the Ll. LtrB group II intron in *Lactococcus lactis*. *Nucleic Acids Res* **35**: 2257-2268.

Belhocine K, Mak AB, Cousineau B. 2008. *Trans*-splicing versatility of the Ll. LtrB group II intron. *RNA* **14**: 1782-1790.

Blaser M, Smith P, Repine J, Joiner K. 1988. Pathogenesis of *Campylobacter fetus* infections. Failure of encapsulated *Campylobacter fetus* to bind C3b explains serum and phagocytosis resistance. *J Clin Investig* **81**: 1434-1444.

Bonen L. 1993. *Trans*-splicing of pre-mRNA in plants, animals, and protists. *FASEB J* **7**: 40-46.

Bonen L. 2008. *Cis*- and *trans*-splicing of group II introns in plant mitochondria. *Mitochondrion* **8**: 26-34.

Bonen L, Vogel J. 2001. The ins and outs of group II introns. *TRENDS Genet* **17**: 322-331.

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.

Boudvillain M, Pyle AM. 1998. Defining functional groups, core structural features and inter-domain tertiary contacts essential for group II intron self-splicing: a NAIM analysis. *EMBO J* **17**: 7091-7104.

Boutz PL, Stoilov P, Li Q, Lin CH, Chawla G, Ostrow K, Shiue L, Ares M, Black DL. 2007. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev* **21**: 1636-1652.

Breitbart RE, Andreadis A, Nadal-Ginard B. 1987. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annu Rev Biochem* **56**: 467-495.

Brüggemann H, Bäumer S, Fricke WF, Wiezer A, Liesegang H, Decker I, Herzberg C, Martínez-Arias R, Merkl R, Henne A. 2003. The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. *Proc Natl Acad Sci USA* **100**: 1316-1321.

Brüggemann H, Gottschalk G. 2004. Insights in metabolism and toxin production from the complete genome sequence of *Clostridium tetani*. *Anaerobe* **10**: 53-68.

Brunger AT, Rummel A. 2009. Receptor and substrate interactions of clostridial neurotoxins. *Toxicon* **54**: 550-560.

Buzayan JM, Hampel A, Bruening G. 1986. Nucleotide sequence and newly formed phosphodiester bond of spontaneously ligated satellite tobacco ringspot virus RNA. *Nucleic Acids Res* **14**: 9729-9743.

Byrappa S, Gavin DK, Gupta KC. 1995. A highly efficient procedure for site-specific mutagenesis of full-length plasmids using Vent DNA polymerase. *Genome Res* **5**: 404-407.

Candales MA, Duong A, Hood KS, Li T, Neufeld RAE, Sun R, McNeil BA, Wu L, Jarding AM, Zimmerly S. 2012. Database for bacterial group II introns. *Nucleic Acids Res* **40**: D187-D190.

Carthew RW, Sontheimer EJ. 2009. Origins and mechanisms of miRNAs and siRNAs. *Cell* **136**: 642-655.

Cech TR, Zaug AJ, Grabowski PJ. 1981. *In vitro* splicing of the ribosomal RNA precursor of tetrahymena: Involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* **27**: 487-496.

Centrón D, Roy PH. 2002. Presence of a group II intron in a multiresistant *Serratia marcescens* strain that harbors three integrons and a novel gene fusion. *Antimicrob Agents Chemother* **46**: 1402-1409.

Chan RT, Robart AR, Rajashankar KR, Pyle AM, Toor N. 2012. Crystal structure of a group II intron in the pre-catalytic state. *Nat Struct Mol Biol* **19**: 555-557.

Chanfreau G, Jacquier A. 1994. Catalytic site components common to both splicing steps of a group II intron. *Science* **266**: 1383-1387.

Chanfreau G, Jacquier A. 1996. An RNA conformational change between the two chemical steps of group II self-splicing. *EMBO J* **15**: 3466-3476.

Chauvaux S, Matuschek M, Beguin P. 1999. Distinct affinity of binding sites for S-layer homologous domains in *Clostridium thermocellum* and *Bacillus anthracis* cell envelopes. *J Bacteriol* **181**: 2455-2458.

Chen PJ, Kalpana G, Goldberg J, Mason W, Werner B, Gerin J, Taylor J. 1986. Structure and replication of the genome of the hepatitis delta virus. *Proc Natl Acad Sci USA* **83**: 8774-8778.

Chen S-J. 2008. RNA folding: conformational statistics, folding kinetics, and ion electrostatics. *Annu Rev Biophys* **37**: 197-214.

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

Chevalier BS, Stoddard BL. 2001. Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic Acids Res* **29**: 3757-3774.

Cho WCS. 2007. Proteomics: technologies and challenges. *Genomics, Proteomics & Bioinformatics* **5**: 77-85.

Chu VT, Adamidi C, Liu Q, Perlman PS, Pyle AM. 2001. Control of branch-site choice by a group II intron. *EMBO J* **20**: 6866-6876.

Claus H, Akça E, Debaerdemaeker T, Evrard C, Declercq J-P, Harris JR, Schlott B, König H. 2005. Molecular organization of selected prokaryotic S-layer proteins. *Can J Microbiol* **51**: 731-743.

Collins L, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol* **22**: 1053-1066.

Copertino DW, Hallick RB. 1993. Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns. *Trends Biochem Sci* **18**: 467-471.

Coppins RL, Hall KB, Groisman EA. 2007. The intricate world of riboswitches. *Curr Opin Microbiol* **10**: 176-181.

Costa M, Christian EL, Michel F. 1998. Differential chemical probing of a group II self-splicing intron identifies bases involved in tertiary interactions and supports an alternative secondary structure model of domain V. *RNA* **4**: 1055-1068.

Costa M, Dème E, Jacquier A, Michel F. 1997a. Multiple tertiary interactions involving domain II of group II self-splicing introns. *J Mol Biol* **267**: 520-536.

Costa M, Fontaine J-M, Goër SL-d, Michel F. 1997b. 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, Westhof E. 2000. A three-dimensional perspective on exon binding by a group II self-splicing intron. *EMBO J* **19**: 5007-5018.

Cousineau B, Lawrence S, Smith D, Belfort M. 2000. Retrotransposition of a bacterial group II intron. *Nature* **404**: 1018-1021.

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.

Dai L, Chai D, Gu SQ, Gabel J, Noskov SY, Blocker FJH, 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.

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

Dai L, Zimmerly S. 2003. ORF-less and reverse-transcriptase-encoding group II introns in archaebacteria, with a pattern of homing into related group II intron ORFs. *RNA* **9**: 14-19.

Daniels DL, Michels Jr WJ, Pyle AM. 1996. Two competing pathways for self-splicing by group II introns: A quantitative analysis of *in vitro* reaction rates and products. *J Mol Biol* **256**: 31-49.

Das R, Kwok LW, Millett IS, Bai Y, Mills TT, Jacob J, Maskel GS, Seifert S, Mochrie SG, Thiyagarajan P. 2003. The fastest global events in RNA folding: electrostatic relaxation and tertiary collapse of the *Tetrahymena* ribozyme. *J Mol BIol* **332**: 311-319.

De Bock K, Cauwenberghs S, Carmeliet P. 2011. RNA structure and the mechanisms of alternative splicing. *Curr Opin Genetics Dev* **21**: 73-79.

De Lencastre A, Hamill S, Pyle AM. 2005. A single active-site region for a group II intron. *Nat Struct Mol Biol* **12**: 626-627.

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 USA* **103**: 8751-8756.

Dème E, Nolte A, Jacquier A. 1999. Unexpected metal ion requirements specific for catalysis of the branching reaction in a group II intron. *Biochemistry* **38**: 3157-3167.

Dlakić M, Mushegian A. 2011. Prp8, the pivotal protein of the spliceosomal catalytic center, evolved from a retroelement-encoded reverse transcriptase. *RNA* **17**: 799-808.

Dobos KM, Khoo K-H, Swiderek KM, Brennan PJ, Belisle JT. 1996. Definition of the full extent of glycosylation of the 45-kilodalton glycoprotein of *Mycobacterium tuberculosis*. *J Bacteriol* **178**: 2498-2506.

Dorn R, Reuter G, Loewendorf A. 2001. Transgene analysis proves mRNA trans-splicing at the complex *mod (mdg4)* locus in *Drosophila*. *Proc Natl Acad Sci USA* **98**: 9724-9729.

Dworkin J, Blaser MJ. 1997. Molecular mechanisms of *Campylobacter fetus* surface layer protein expression. *Mol Microbiol* **26**: 433-440.

Early P, Rogers J, Davis M, Calame K, Bond M, Wall R, Hood L. 1980. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* **20**: 313-319.

Ebbole DJ, Jin Y, Thon M, Pan H, Bhattarai E, Thomas T, Dean R. 2004. Gene discovery and gene expression in the rice blast fungus, *Magnaporthe grisea*: analysis of expressed sequence tags. *Mol Plant Microbe Interact* **17**: 1337-1347.

Edge A. 2003. Deglycosylation of glycoproteins with trifluoromethanesulphonic acid: elucidation of molecular structure and function. *Biochem J* **376**: 339-350.

Edge AS, Faltynek CR, Hof L, Reichert Jr LE, Weber P. 1981. Deglycosylation of glycoproteins by trifluoromethanesulfonic acid. *Anal Biochem* **118**: 131-137.

Egelseer EM, Danhorn T, Pleschberger M, Hotzy C, Sleytr UB, Sára M. 2001. Characterization of an S-layer glycoprotein produced in the course of S-layer variation of *Bacillus stearothermophilus* ATCC 12980 and sequencing and cloning of the *sbsD* gene encoding the protein moiety. *Arch microbiol* **177**: 70-80.

Egelseer EM, Leitner K, Jarosch M, Hotzy C, Zayni S, Sleytr UB, Sára M. 1998. The S-layer proteins of two *Bacillus stearothermophilus* wild-type strains are bound via their N-terminal region to a secondary cell wall polymer of identical chemical composition. *J Bacteriol* **180**: 1488-1495.

Ellington AD, Szostak JW. 1990. *In vitro* selection of RNA molecules that bind specific ligands. *Nature* **346**: 818-822.

Emerson JE, Reynolds CB, Fagan RP, Shaw HA, Goulding D, Fairweather NF. 2009. A novel genetic switch controls phase variable expression of CwpV, a *Clostridium difficile* cell wall protein. *Mol Microbiol* **74**: 541-556.

Engelhardt H. 2007. Are S-layers exoskeletons? The basic function of protein surface layers revisited. *J Struct Biol* **160**: 115-124.

Erickson P, Herzberg M. 1993. Evidence for the covalent linkage of carbohydrate polymers to a glycoprotein from *Streptococcus sanguis*. *J Biol Chem* **268**: 23780-23783.

Escalante R, Moreno N, Sastre L. 2003. *Dictyostelium discoideum* developmentally regulated genes whose expression is dependent on MADS box transcription factor SrfA. *Eukaryotic Cell* **2**: 1327-1335.

Fang X-W, Thiyagarajan P, Sosnick T, Pan T. 2002. The rate-limiting step in the folding of a large ribozyme without kinetic traps. *Proc Natl Acad Sci USA* **99**: 8518-8523.

Fedorova O, Mitros T, Pyle AM. 2003. Domains 2 and 3 interact to form critical elements of the group II intron active site. *J Mol Biol* **330**: 197-209.

Fedorova O, Pyle AM. 2005. Linking the group II intron catalytic domains: tertiary contacts and structural features of domain 3. *EMBO J* **24**: 3906-3916.

Fedorova O, Zingler N. 2007. Group II introns: structure, folding and splicing mechanism. *Biol Chem* **388**: 665-678.

Ferat J-L, Michel F. 1993. Group II self-splicing introns in bacteria. *Nature* **364**: 358-361.

Ferat JL, Le Gouar M, Michel F. 2003. A group II intron has invaded the genus *Azotobacter* and is inserted within the termination codon of the essential groEL gene. *Mol Microbiol* **49**: 1407-1423.

Folk J. 1980. Transglutaminases. *Annu Rev Biochem* **49**: 517-531.

Fontaine JM, Goux D, Kloareg B, Loiseaux-de Goër S. 1997. The reverse-transcriptase-like proteins encoded by group II introns in the mitochondrial genome of the brown alga *Pylaiella littoralis* belong to two different lineages which apparently coevolved with the group II ribozyme lineages. *J Mol Evol* **44**: 33-42.

Forster AC, Symons RH. 1987. Self-cleavage of plus and minus RNAs of a virusoid and a structural model for the active sites. *Cell* **49**: 211-220.

Fox-Walsh KL, Dou Y, Lam BJ, Hung S, Baldi PF, Hertel KJ. 2005. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc Natl Acad Sci USA* **102**: 16176.

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

Gautheret D, Lambert A. 2001. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J Mol Biol* **313**: 1003-1011.

Gilbert W. 1986. Origin of life: The RNA world. *Nature* **319**: 618.

Goldschmidt-Clermont M, Choquet Y, Girard-Bascou J, Michel F, Schirmer-Rahire M, Rochaix JD. 1991. A small chloroplast RNA may be required for *trans*-splicing in *Chlamydomonas reinhardtii*. *Cell* **65**: 135-143.

Gordon PM, Fong R, Piccirilli JA. 2007. A second divalent metal ion in the group II intron reaction center. *Chem Biol* **14**: 607-612.

Gordon PM, Piccirilli JA. 2001. Metal ion coordination by the AGC triad in domain 5 contributes to group II intron catalysis. *Nat Struct Mol Biol* **8**: 893-898.

Gottesman S. 2004. The small RNA regulators of *Escherichia coli*: roles and mechanisms. *Annu Rev Microbiol* **58**: 303-328.

Granlund M, Michel F, Norgren M. 2001. Mutually exclusive distribution of IS1548 and GBSi1, an active group II intron identified in human isolates of group B streptococci. *J Bacteriol* **183**: 2560-2569.

Graveley BR. 2001. Alternative splicing: increasing diversity in the proteomic world. *TRENDS Genet* **17**: 100-107.

Griffin Jr EA, Qin Z, Michels Jr WJ, Pyle AM. 1995. Group II intron ribozymes that cleave DNA and RNA linkages with similar efficiency, and lack contacts with substrate 2'-hydroxyl groups. *Chem Biol* **2**: 761-770.

Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**: 849-857.

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 RS, Gao B. 2009. Phylogenomic analyses of clostridia and identification of novel protein signatures that are specific to the genus *Clostridium sensu stricto* (cluster I). *Int J Syst Evol Microbiol* **59**: 285-294.

Hamill S, Pyle AM. 2006. The receptor for branch-site docking within a group II intron active site. *Mol Cell* **23**: 831-840.

Harris-Kerr CL, Zhang M, Peebles CL. 1993. The phylogenetically predicted base-pairing interaction between alpha and alpha' is required for group II splicing *in vitro*. *Proc Natl Acad Sci USA* **90**: 10658.

Ho Y, Kim S-J, Waring RB. 1997. A protein encoded by a group I intron in *Aspergillus nidulans* directly assists RNA splicing and is a DNA endonuclease. *Proc Natl Acad Sci USA* **94**: 8994-8999.

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.

Ilk N, Kosma P, Puchberger M, Egelseer EM, Mayer HF, Sleytr UB, Sára M. 1999. Structural and functional analyses of the secondary cell wall polymer of *Bacillus sphaericus* CCM 2177 that serves as an S-layer-specific anchor. *J Bacteriol* **181**: 7643-7646.

Inoue H, Nojima H, Okayama H. 1990. High efficiency transformation of *Escherichia coli* with plasmids. *Gene* **96**: 23-28.

Irimia M, Rukov J, Penny D, Roy S. 2007. Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. *BMC Evol Biol* **7**: 188.

Isaacs FJ, Dwyer DJ, Collins JJ. 2006. RNA synthetic biology. *Nat Biotechnol* **24**: 545-554.

Jacquier A, Jacquesson-Breuleux N. 1991. Splice site selection and role of the lariat in a group II intron. *J Mol Biol* **219**: 415-428.

Jacquier A, Michel F. 1987. Multiple exon-binding sites in class II self-splicing introns. *Cell* **50**: 17-29.

Jacquier A, Michel F. 1990. Base-pairing interactions involving the 5' and 3'-terminal nucleotides of group II self-splicing introns. *J Mol Biol* **213**: 437-447.

Jarrell K, Peebles C, Dietrich R, Romiti S, Perlman P. 1988. Group II intron self-splicing. Alternative reaction conditions yield novel products. *J Biol Chem* **263**: 3432-3439.

Jarrell KA. 1993. Inverse splicing of a group II intron. *Proc Natl Acad Sci USA* **90**: 8624-8627.

Jenison RD, Gill SC, Pardi A, Polisky B. 1994. High-resolution molecular discrimination by RNA. *Science* **263**: 1425-1429.

Jenkins K, Hong L, Hallick R. 1995. Alternative splicing of the *Euglena gracilis* chloroplast roaA transcript. *RNA* **1**: 624-633.

Johnston NC, Aygun-Sunar S, Guan Z, Ribeiro AA, Daldal F, Raetz CRH, Goldfine H. 2010. A phosphoethanolamine-modified glycosyl diradylglycerol in the polar lipids of *Clostridium tetani*. *J Lipid Res* **51**: 1953-1961.

Johnston WK, Unrau PJ, Lawrence MS, Glasner ME, Bartel DP. 2001. RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science* **292**: 1319-1325.

Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* **11**: 345-355.

Khalil AS, Collins JJ. 2010. Synthetic biology: applications come of age. *Nat Rev Genet* **11**: 367-379.

Kirby JM, Ahern H, Roberts AK, Kumar V, Freeman Z, Acharya KR, Shone CC. 2009. Cwp84, a surface-associated cysteine protease, plays a role in the maturation of the surface layer of *Clostridium difficile*. *J Biol Chem* **284**: 34666-34673.

Klein JR, Dunny GM. 2002. Bacterial group II introns and their association with mobile genetic elements. *Front Biosci* **7**: d1843-d1856.

Koch J, Boulanger S, Dib-Hajj S, Hebbar S, Perlman P. 1992. Group II introns deleted for multiple substructures retain self-splicing activity. *Mol Cell Biol* **12**: 1950-1958.

Kohchi T, Umesono K, Ogura Y, Komine Y, Nakahigashi K, Komano T, Yamada Y, Ozeki H, Ohyama K. 1988. A nicked group II intron and *trans*-splicing in liverwort, *Marchantia polymorpha,* chloroplasts. *Nucleic Acids Res* **16**: 10025-10036.

Koonin EV. 2009. Intron-dominated genomes of early ancestors of eukaryotes. *J Hered* **100**: 618-623.

Kornblihtt AR. 2007. Coupling transcription and alternative splicing. *Adv Exp Med Biol* **623**: 175-189.

Kotiranta A, Haapasalo M, Kari K, Kerosuo E, Olsen I, Sorsa T, Meurman JH, Lounatmaa K. 1998. Surface structure, hydrophobicity, phagocytosis, and

adherence to matrix proteins of *Bacillus cereus* cells with and without the crystalline surface protein layer. *Infect Immun* **66**: 4895-4902.

Koval S, Hynes S. 1991. Effect of paracrystalline protein surface layers on predation by *Bdellovibrio bacteriovorus*. *J Bacteriol* **173**: 2244-2249.

Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. 1982. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell* **31**: 147-157.

Kuen B, Sleytr UB, Lubitz W. 1994. Sequence analysis of the *sbsA* gene encoding the 130-kDa surface-layer protein of *Bacillus stearothermophilus* strain PV72. *Gene* **145**: 115-120.

Kuo M, Sharmeen L, Dinter-Gottlieb G, Taylor J. 1988. Characterization of self-cleaving RNA sequences on the genome and antigenome of human hepatitis delta virus. *J Virol* **62**: 4439-4444.

Laemmli UK. 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**: 680-685.

Lambowitz AM, Belfort M. 1993. Introns as mobile genetic elements. *Annu Rev Biochem* **62**: 587-622.

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.

Lee ER, Baker JL, Weinberg Z, Sudarsan N, Breaker RR. 2010. An allosteric self-splicing ribozyme triggered by a bacterial second messenger. *Science* **329**: 845-848.

Li-Pook-Than J, Bonen L. 2006. Multiple physical forms of excised group II intron RNAs in wheat mitochondria. *Nucleic Acids Res* **34**: 2782-2790.

Li C-F, Costa M, Michel F. 2011a. Linking the branchpoint helix to a newly found receptor allows lariat formation by a group II intron. *EMBO J* **30**: 3040-3051.

Li CF, Costa M, Bassi G, Lai YK, Michel F. 2011b. Recurrent insertion of 5'-terminal nucleotides and loss of the branchpoint motif in lineages of group II introns inserted in mitochondrial preribosomal RNAs. *RNA* **17**: 1321-1335.

Liang Joe C, Bloom Ryan J, Smolke Christina D. 2011. Engineering biological systems with synthetic RNA molecules. *Mol Cell* **43**: 915-926.

Lin S, Fu XD. 2007. SR proteins and related factors in alternative splicing. *Adv Exp Med Biol* **623**: 107-122.

Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M et al. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402**: 761-768.

Lopez AJ. 1998. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu Rev Genet* **32**: 279-305.

Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. 2001. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res* **29**: 4724-4735.

Makeyev EV, Zhang J, Carrasco MA, Maniatis T. 2007. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol Cell* **27**: 435-448.

Malek O, Knoop V. 1998. *Trans*-splicing group II introns in plant mitochondria: the complete set of *cis*-arranged homologs in ferns, fern allies, and a hornwort. *RNA* **4**: 1599-1609.

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.

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

Martínez-Abarca F, Toro N. 2000. RecA-independent ectopic transposition *in vivo* of a bacterial group II intron. *Nucleic Acids Res* **28**: 4397-4402.

Martinez-Contreras R, Cloutier P, Shkreta L, Fisette JF, Revil T, Chabot B. 2007. hnRNP proteins and splicing control. *Adv Exp Med Biol* **623**: 123-147.

Matlin AJ, Clark F, Smith CW. 2005. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* **6**: 386-398.

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.

McNeil BA, Simon DM, Zimmerly S. 2013. Alternative splicing of a group II intron in a surface layer protein gene in *Clostridium tetani*. *Nucleic Acids Res*: gkt1053.

Mekalanos JJ. 1992. Environmental signals controlling expression of virulence determinants in bacteria. *J Bacteriol* **174**: 1-7.

Meng Q, Wang Y, Liu X-Q. 2005. An intron-encoded protein assists RNA splicing of multiple similar introns of different bacterial genes. *J Biol Chem* **280**: 35085-35088.

Mengaud J, Ohayon H, Gounon P, Mège R-M, Cossart P. 1996. E-cadherin is the receptor for internalin, a surface protein required for entry of *L. monocytogenes* into epithelial cells. *Cell* **84**: 923-932.

Mesnage S, Fontaine T, Mignot T, Delepierre M, Mock M, Fouet A. 2000. Bacterial SLH domain proteins are non-covalently anchored to the cell surface via a conserved mechanism involving wall polysaccharide pyruvylation. *The EMBO journal* **19**: 4473-4484.

Michel F, Costa M, Doucet AJ, Ferat JL. 2007. Specialized lineages of bacterial group II introns. *Biochimie* **89**: 542-553.

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 J. 1995. Structure and activities of group II introns. *Annu Rev Biochem* **64**: 435-461.

Michel F, Jacquier A. 1987. Long-range intron-exon and intron-intron pairings involved in self-splicing of class II catalytic introns. *Cold Spring Harbor Symp Quant Biol* **52**: 201-212.

Michel F, Umesono K, Ozeki H. 1989. Comparative and functional anatomy of group II catalytic introns--a review. *Gene* **82**: 5-30.

Michels WJJ, Pyle AM. 1995. Conversion of a group II intron into a new multiple-turnover ribozyme that selectively cleaves oligonucleotides: elucidation of reaction mechanism and structure/function relationships. *Biochemistry* **34**: 2965-2977.

Mohr G, Perlman PS, Lambowitz AM. 1993. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Res* **21**: 4991-4997.

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 USA* **103**: 3569-3574.

Molina-Sánchez MD, Martinez-Abarca F, Toro N. 2006. Excision of the *Sinorhizobium meliloti* group II intron RmInt1 as circles *in vivo*. *J Biol Chem* **281**: 28737-28744.

Montecucco C, Schiavo G. 1995. Structure and function of tetanus and botulinum neurotoxins. *Quart Rev Biophys* **28**: 423-472.

Moran JV, Mecklenburg KL, Sass P, Belcher SM, Mahnke D, Lewin A, Perlman P. 1994. Splicing defective mutants of the *COXI* gene of yeast mitochondrial DNA: initial definition of the maturase domain of the group II intron aI2. *Nucleic Acids Res* **22**: 2057-2064.

Mörl M, Schmelzer C. 1990. Integration of group II intron bl1 into a foreign RNA by reversal of the self-splicing reaction *in vitro*. *Cell* **60**: 629-636.

Mueller JH, Miller PA. 1945. Production of tetanal toxin. *J Immunol* **56**: 143-147.

Mullineux S-T, Costa M, Bassi GS, Michel F, Hausner G. 2010. A group II intron encodes a functional LAGLIDADG homing endonuclease and self-splices under moderate temperature and ionic conditions. *RNA* **16**: 1818-1831.

Munn C, Ishiguro E, Kay W, Trust T. 1982. Role of surface components in serum resistance of virulent *Aeromonas salmonicida*. *Infect Immun* **36**: 1069-1075.

Murray HL, Mikheeva S, Coljee VW, Turczyk BM, Donahue WF, Bar-Shalom A, Jarrell KA. 2001. Excision of group II introns as circles. *Mol Cell* **8**: 201-211.

Neves G, Zucker J, Daly M, Chess A. 2004. Stochastic yet biased expression of multiple *Dscam* splice variants by individual cells. *Nat Genet* **36**: 240-246.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457-463.

Noller H. 1993. Peptidyl transferase: protein, ribonucleoprotein, or RNA? *J Bacteriol* **175**: 5297-5300.

Okazaki K, Niwa O. 2000. mRNAs encoding zinc finger protein isoforms are expressed by alternative splicing of an in-frame intron in fission yeast. *DNA Res* **7**: 27-30.

Packer NH, Pawlak A, Kett WC, Gooley AA, Redmond JW, Williams KL. 1997. Proteome analysis of glycoforms: A review of strategies for the microcharacterisation of glycoproteins separated by two-dimensional polyacrylamide gel electrophoresis. *Electrophoresis* **18**: 452-460.

Padgett RA, Podar M, Boulanger SC, Perlman PS. 1994. The stereochemical course of group II intron self-splicing. *Science* **266**: 1685-1688.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**: 1413-1415.

Park JW, Parisky K, Celotto AM, Reenan RA, Graveley BR. 2004. Identification of alternative splicing regulators by RNA interference in *Drosophila*. *Proc Natl Acad Sci USA* **101**: 15974-15979.

Pavkov-Keller T, Howorka S, Keller W. 2011. Chapter 3 - The structure of bacterial S-layer proteins. in *Progress in molecular biology and translational science* (ed. H Stefan), pp. 73-130. Academic Press.

Peebles C, Benatan E, Jarrell K, Perlman P. 1987. Group II intron self-splicing: development of alternative reaction conditions and identification of a predicted intermediate. *Cold Spring Harb Symp Quant Biol* **52**: 223-232.

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

Pellizzari R, Rossetto O, Schiavo G, Montecucco C. 1999. Tetanus and botulinum neurotoxins: mechanism of action and therapeutic uses. *Philos Trans R Soc Lond B Biol Sci* **354**: 259-268.

Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C. 2007. Transcript specificity in yeast pre-mRNA splicing revealed by mutations in core spliceosomal components. *PLoS Biol* **5**: e90.

Podar M, Perlman PS, Padgett RA. 1995. Stereochemical selectivity of group II intron splicing, reverse splicing, and hydrolysis reactions. *Mol Cell Biol* **15**: 4466-4478.

Powner MW, Gerland B, Sutherland JD. 2009. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **459**: 239-242.

Pyle AM. 2010. The tertiary structure of group II introns: implications for biological function and evolution. *Crit Rev Biochem Mol* **45**: 215-232.

Pyle AM, Fedorova O, Waldsich C. 2007. Folding of group II introns: a model system for large, multidomain RNAs? *Trends Biochem Sci* **32**: 138-145.

Qazi O, Brailsford A, Wright A, Faraar J, Campbell J, Fairweather N. 2007. Identification and characterization of the surface-layer protein of *Clostridium tetani*. *FEMS Microbiol Lett* **274**: 126-131.

Qin PZ, Pyle AM. 1998. The architectural organization and mechanistic function of group II intron structural elements. *Curr Opin Struct Biol* **8**: 301-308.

Quiroga C, Roy PH, Centrón D. 2008. The *S.ma*.I2 class C group II intron inserts at integron *attC* sites. *Microbiology* **154**: 1341-1353.

Rest JS, Mindell DP. 2003. Retroids in Archaea: Phylogeny and Lateral Origins. *Mol Biol Evol* **20**: 1134-1142.

Rich A. 1962. On the problems of evolution and biochemical information transfer. in *Horizons in biochemistry* (eds. M Kasha, B Pullman), pp. 103-126. Academic Press: New York, NY, USA.

Robart AR, Montgomery NK, Smith KL, Zimmerly S. 2004. Principles of 3' splice site selection and alternative splicing for an unusual group II intron from *Bacillus anthracis*. *RNA* **10**: 854-862.

Robart AR, Seo W, Zimmerly S. 2007. Insertion of group II intron retroelements after intrinsic transcriptional terminators. *Proc Natl Acad Sci USA* **104**: 6620.

Roegener J, Lutter P, Reinhardt R, Blüggel M, Meyer HE, Anselmetti D. 2003. Ultrasensitive detection of unstained proteins in acrylamide gels by native UV fluorescence. *Anal Chem* **75**: 157-159.

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

Roitzsch M, Pyle AM. 2009. The linear form of a group II intron catalyzes efficient autocatalytic reverse splicing, establishing a potential for mobility. *RNA* **15**: 473-482.

Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet* **7**: 211-221.

Russell R, Millett IS, Tate MW, Kwok LW, Nakatani B, Gruner SM, Mochrie SG, Pande V, Doniach S, Herschlag D. 2002. Rapid compaction during RNA folding. *Proc Natl Acad Sci USA* **99**: 4266-4271.

Salis HM, Mirsky EA, Voigt CA. 2009. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* **27**: 946-950.

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.

Sánchez R, L. 2008. Sex-determining mechanisms in insects. *Int J Dev Biol* **52**: 837-856.

Sára M, Dekitsch C, Mayer HF, Egelseer EM, Sleytr UB. 1998. Influence of the secondary cell wall polymer on the reassembly, recrystallization, and stability properties of the S-layer protein from *Bacillus stearothermophilus* PV72/p2. *J Bacteriol* **180**: 4146-4153.

Saville BJ, Collins RA. 1990. A site-specific self-cleavage reaction performed by a novel RNA in *Neurospora* mitochondria. *Cell* **61**: 685-696.

Schäffer C, Messner P. 2001. Glycobiology of surface layer proteins. *Biochimie* **83**: 591-599.

Schiavo GG, Benfenati F, Poulain B, Rossetto O, de Laureto PP, DasGupta BR, Montecucco C. 1992. Tetanus and botulinum-B neurotoxins block neurotransmitter release by proteolytic cleavage of synaptobrevin. *Nature* **359**: 832-835.

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.

Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL. 2000. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**: 671-684.

Scholz HC, Riedmann E, Witte A, Lubitz W, Kuen B. 2001. S-layer variation in *Bacillus stearothermophilus* PV72 is based on DNA rearrangements between the chromosome and the naturally occurring megaplasmids. *J Bacteriol* **183**: 1672-1679.

Schwartz S, Silva J, Burstein D, Pupko T, Eyras E, Ast G. 2008. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res* **18**: 88-103.

Serganov A, Nudler E. 2013. A Decade of Riboswitches. *Cell* **152**: 17-24.

Sharp PA. 1991. Five easy pieces. *Science* **254**: 663-663.

Shin C, Manley JL. 2004. Cell signalling and the control of pre-mRNA splicing. *Nat Rev Mol Cell Biol* **5**: 727-738.

Shukla GC, Padgett RA. 2002. A catalytically active group II intron domain 5 can function in the U12-dependent spliceosome. *Mol Cell* **9**: 1145-1150.

Sigel RKO, 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 RKO, Vaidya A, Pyle AM. 2000. Metal ion binding sites in a group II intron core. *Nat Struct Mol Biol* **7**: 1111-1116.

Simon DM, Clarke NAC, McNeil BA, Johnson I, Pantuso D, Dai L, Chai D, Zimmerly S. 2008. Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *RNA* **14**: 1704-1713.

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.

Sinha J, Reyes SJ, Gallivan JP. 2010. Reprogramming bacteria to seek and destroy an herbicide. *Nat Chem Biol* **6**: 464-470.

Sinniger F, Chevaldonné P, Pawlowski J. 2007. Mitochondrial genome of *Savalia savaglia (Cnidaria, Hexacorallia*) and early metazoan phylogeny. *J Mol Evol* **64**: 196-203.

Sleytr UB, Beveridge TJ. 1999. Bacterial S-layers. *TRENDS Microbiol* **7**: 253-260.

Sleytr UB, Messner P, Pum D, Sara M. 1996. Crystalline bacterial cell surface proteins. *Mol Microbiol*: 911-916.

Smit E, Oling F, Demel R, Martinez B, Pouwels PH. 2001. The S-layer Protein of *Lactobacillus acidophilus* ATCC 4356: Identification and Characterisation of Domains Responsible for S-protein Assembly and Cell Wall Binding. *J Mol Biol* **305**: 245-257.

Solovyev V, Salamov A. 2011. Automatic Annotation of Microbial Genomes and Metagenomic Sequences. in *Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies* (ed. RW Li), pp. 61-78 Nova Science Publishers.

Sontheimer EJ, Steitz JA. 1993. The U5 and U6 small nuclear RNAs as the active site components of the s-pliceosome. *Science* **262**: 1989-1996.

Stabell FB, Tourasse NJ, Kolstø AB. 2009. A conserved 3' extension in unusual group II introns is important for efficient second-step splicing. *Nucleic Acids Res* **37**: 3202-3214.

Stabell FB, Tourasse NJ, Ravnum S, Kolstø AB. 2007. Group II intron in *Bacillus cereus* has an unusual 3' extension and splices 56 nucleotides downstream of the predicted site. *Nucleic Acids Res* **35**: 1612-1623.

Steinberg TH, Top KPO, Berggren KN, Kemper C, Jones L, Diwu Z, Haugland RP, Patton WF. 2001. Rapid and simple single nanogram detection of glycoproteins in polyacrylamide gels and on electroblots. *Proteomics* **1**: 841-855.

Steitz TA, Moore PB. 2003. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem Sci* **28**: 411-418.

Steitz TA, Steitz JA. 1993. A general two-metal-ion mechanism for catalytic RNA. *Proc Natl Acad Sci USA* **90**: 6498-6502.

Stimson E, Virji M, Makepeace K, Dell A, Morris HR, Payne G, Saunders JR, Jennings MP, Barker S, Panico M. 1995. Meningococcal pilin: a glycoprotein substituted with digalactosyl 2, 4-diacetamido 2, 4, 6-trideoxyhexose. *Mol Microbiol* **17**: 1201-1214.

Storz G, Vogel J, Wassarman Karen M. 2011. Regulation by Small RNAs in Bacteria: Expanding Frontiers. *Mol Cell* **43**: 880-891.

Su LJ, Qin PZ, Michels WJ, Pyle AM. 2001. Guiding ribozyme cleavage through motif recognition: the mechanism of cleavage site selection by a group II intron ribozyme. *J Mol Biol* **306**: 655-668.

Su LJ, Waldsich C, Pyle AM. 2005. An obligate intermediate along the slow folding pathway of a group II intron ribozyme. *Nucleic Acids Res* **33**: 6674-6687.

Takeoka A, Takumi K, Koga T, Kawata T. 1991. Purification and characterization of S layer proteins from *Clostridium difficile* GAI 0714. *J Gen Microbiol* **137**: 261-267.

Toledo-Arana A, Repoila F, Cossart P. 2007. Small noncoding RNAs controlling pathogenesis. *Curr Opin Microbiol* **10**: 182-188.

Toor N, Hausner G, Zimmerly S. 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA* **7**: 1142-1152.

Toor N, Keating KS, Fedorova O, Rajashankar K, Wang J, Pyle AM. 2010. Tertiary architecture of the *Oceanobacillus iheyensis* group II intron. *RNA* **16**: 57-69.

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

Toor N, Robart AR, Christianson J, Zimmerly S. 2006. Self-splicing of a group IIC intron: 5' exon recognition and alternative 5' splicing events implicate the stem-loop motif of a transcriptional terminator. *Nucleic Acids Res* **34**: 6461-6471.

Toor N, Zimmerly S. 2002. Identification of a family of group II introns encoding LAGLIDADG ORFs typical of group I introns. *RNA* **8**: 1373-1377.

Toro N, Jiménez-Zurdo JI, García-Rodríguez FM. 2007. Bacterial group II introns: not just splicing. *FEMS Microbiol Rev* **31**: 342-358.

Toro N, Molina-Sánchez MD, Fernández-López M. 2002. Identification and characterization of bacterial class E group II introns. *Gene* **299**: 245-250.

Treiber DK, Williamson JR. 2001. Beyond kinetic traps in RNA folding. *Curr Opin Struct Biol* **11**: 309-314.

Tucker BJ, Breaker RR. 2005. Riboswitches as versatile gene control elements. *Curr Opin Struct Biol* **15**: 342-348.

Valadkhan S, Manley JL. 2001. Splicing-related catalysis by protein-free snRNAs. *Nature* **413**: 701-707.

Valadkhan S, Mohammadi A, Jaladat Y, Geisler S. 2009. Protein-free small nuclear RNAs catalyze a two-step splicing reaction. *Proc Natl Acad Sci USA* **106**: 11901-11906.

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

Van der Auwera GA, Andrup L, Mahillon J. 2005. Conjugative plasmid pAW63 brings new insights into the genesis of the *Bacillus anthracis* virulence plasmid pXO2 and of the *Bacillus thuringiensis* plasmid pBT9727. *BMC genomics* **6**: 103.

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.

Van der Veen R, Kwakman J, Grivell L. 1987. Mutations at the lariat acceptor site allow self-splicing of a group II intron without lariat formation. *EMBO J* **6**: 3827-3831.

Vilardell J, Chartrand P, Singer RH, Warner JR. 2000. The odyssey of a regulated transcript. *RNA* **6**: 1773-1780.

Vogel J, Börner T. 2002. Lariat formation and a hydrolytic pathway in plant chloroplast group II intron splicing. *EMBO J* **21**: 3794-3803.

Waldsich C, Pyle AM. 2006. A folding control element for tertiary collapse of a group II intron ribozyme. *Nat Struct Mol Biol* **14**: 37-44.

Waldsich C, Pyle AM. 2008. A kinetic intermediate that regulates proper folding of a group II intron RNA. *J Mol Biol* **375**: 572-580.

Waligora A-J, Hennequin C, Mullany P, Bourlioux P, Collignon A, Karjalainen T. 2001. Characterization of a cell surface protein of *Clostridium difficile* with adhesive properties. *Infect Immun* **69**: 2144-2153.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470-476.

Wank H, SanFilippo J, Singh RN, Matsuura M, Lambowitz AM. 1999. A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. *Mol Cell* **4**: 239-250.

Wei-Mei C, Carl M, Dasch GA. 1992. Mapping of monoclonal antibody binding sites on CNBr fragments of the S-layer protein antigens of *Rickettsia typhi* and *Rickettsia prowazekii*. *Mol Immunol* **29**: 95-105.

Weigand JE, Sanchez M, Gunnesch E-B, Zeiher S, Schroeder R, Suess B. 2008. Screening for engineered neomycin riboswitches that control translation initiation. *RNA* **14**: 89-97.

Wieland M, Benz A, Klauser B, Hartig JS. 2009. Artificial ribozyme switches containing natural riboswitch aptamer domains. *Angew Chem Int Ed* **121**: 2753-2756.

Winkler WC, Nahvi A, Roth A, Collins JA, Breaker RR. 2004. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* **428**: 281-286.

Wochner A, Attwater J, Coulson A, Holliger P. 2011. Ribozyme-catalyzed transcription of an active ribozyme. *Science* **332**: 209-212.

Woodson SA. 2005. Metal ions and RNA folding: a highly charged topic with a dynamic future. *Curr Opin Chem Biol* **9**: 104-109.

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.

Xie J, Black DL. 2001. A CaMK IV responsive RNA element mediates depolarization-induced alternative splicing of ion channels. *Nature* **410**: 936-939.

Xiong Y, Eickbush TH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* **9**: 3353-3362.

Yatzkan E, Yarden O. 1999. The B regulatory subunit of protein phosphatase 2A is required for completion of macroconidiation and other developmental processes in *Neurospora crassa*. *Mol Microbiol* **31**: 197-209.

Yu Y, Maroney PA, Denker JA, Zhang XHF, Dybkov O, Lührmann R, Jankowsky E, Chasin LA, Nilsen TW. 2008. Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. *Cell* **135**: 1224-1236.

Zaher HS, Unrau PJ. 2007. Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. *RNA* **13**: 1017-1026.

Zhao G, Ali E, Sakka M, Kimura T, Sakka K. 2006. Binding of S-layer homology modules from *Clostridium thermocellum* SdbA to peptidoglycans. *Appl Microbiol Biotechnol* **70**: 464-469.

Zhao Z, Aliwarga Y, Willcox MD. 2007. Intrinsic protein fluorescence interferes with detection of tear glycoproteins in SDS-polyacrylamide gels using extrinsic fluorescent dyes. *J Biomol Tech* **18**: 331-335.

Zhong J, Lambowitz AM. 2003. Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. *EMBO J* **22**: 4555-4565.

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

Zhuo D, Madden R, Elela SA, Chabot B. 2007. Modern origin of numerous alternatively spliced human introns from tandem arrays. *Proc Natl Acad Sci USA* **104**: 882-886.

Zimmerly S, Guo H, Perlman PS, Lambowltz AM. 1995. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* **82**: 545-554.

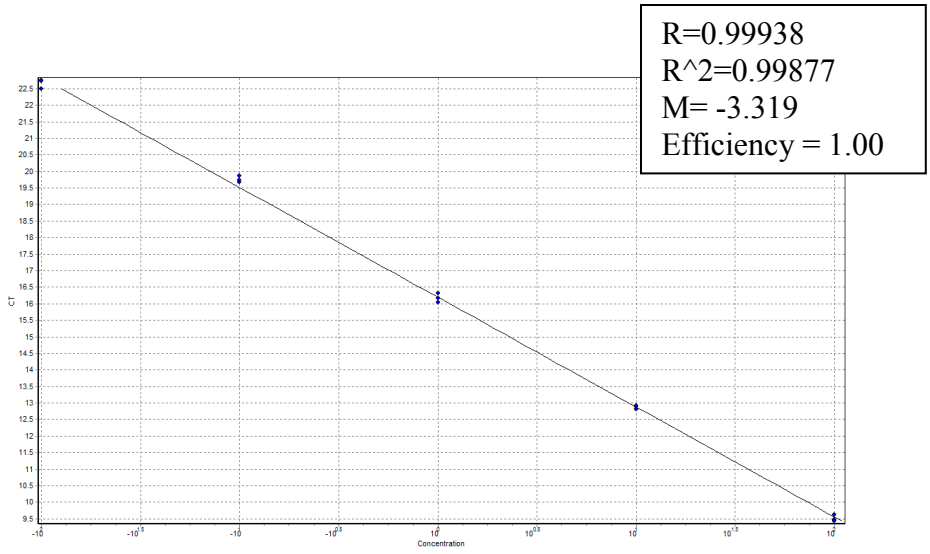
Zimmerly S, Hausner G, Wu X. 2001. Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res* **29**: 1238-1250.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406-3415.

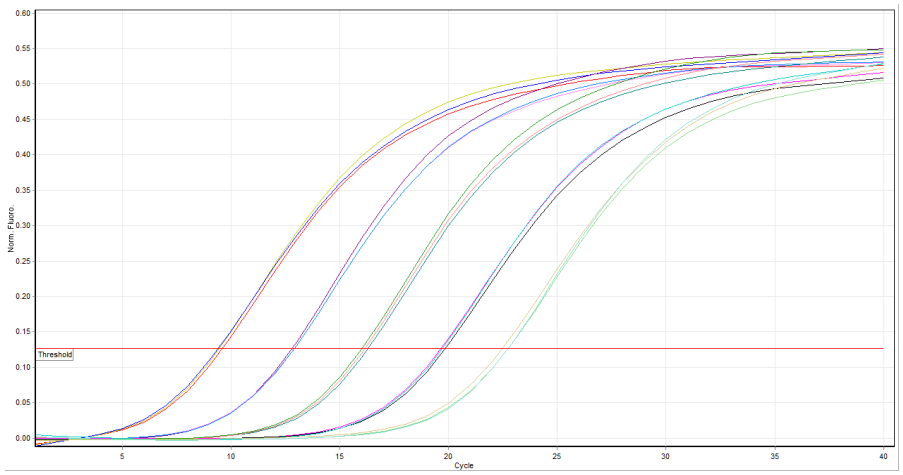
APPENDIX A

Standard curve (A), efficiency of amplification (B) and melt curves (C) for SF1

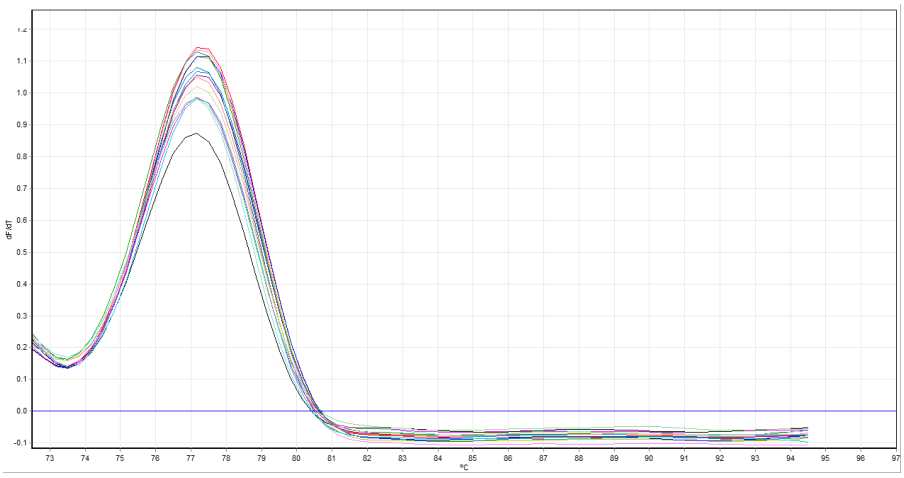
A



B

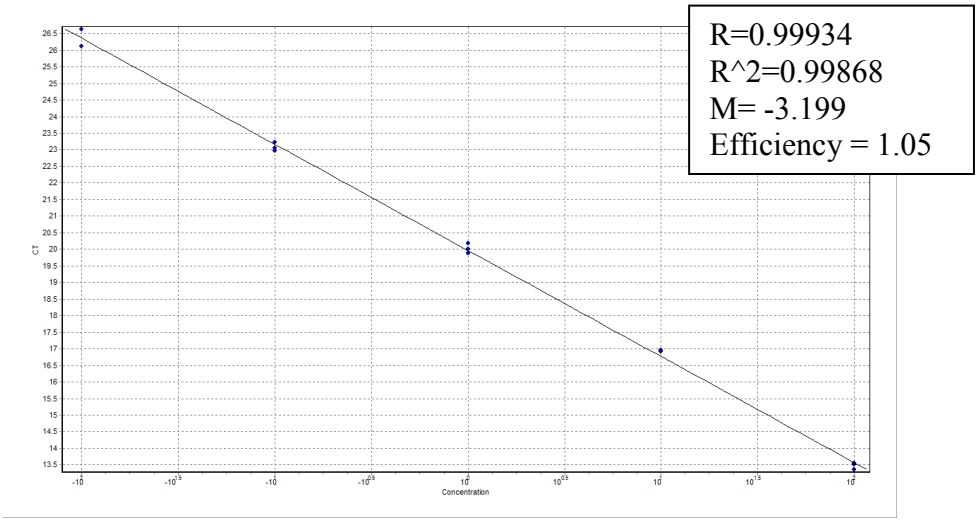


C

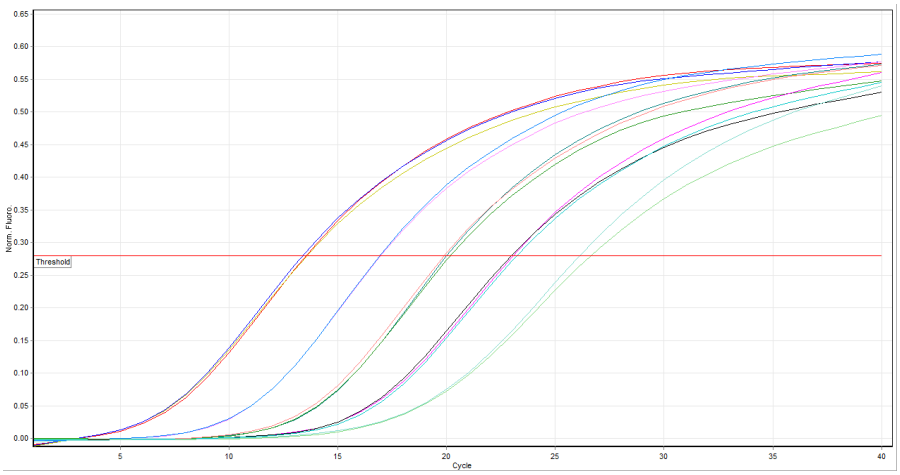


Standard curve (A), efficiency of amplification (B) and melt curves (C) for SF2

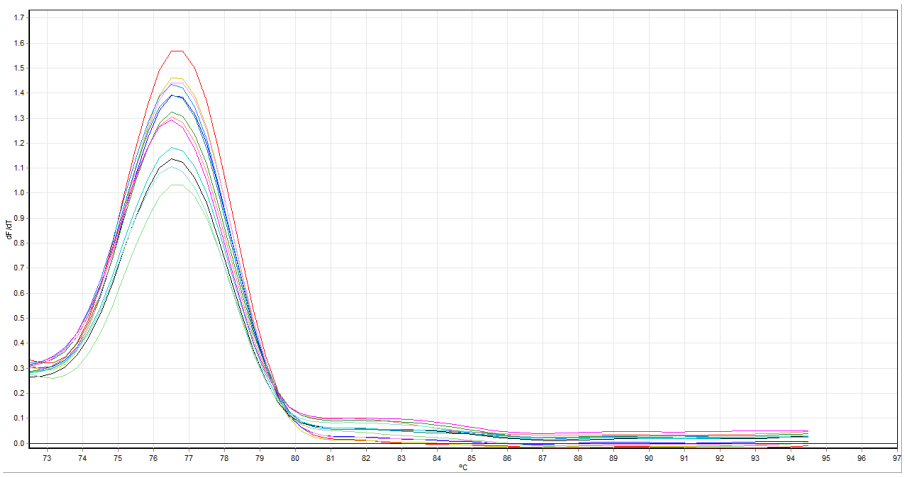
A



B

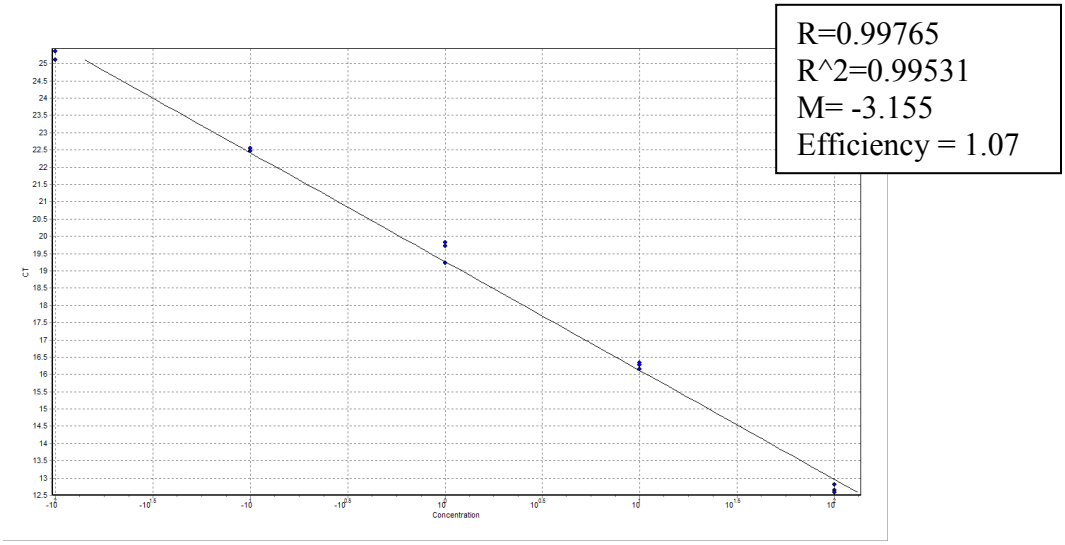


C

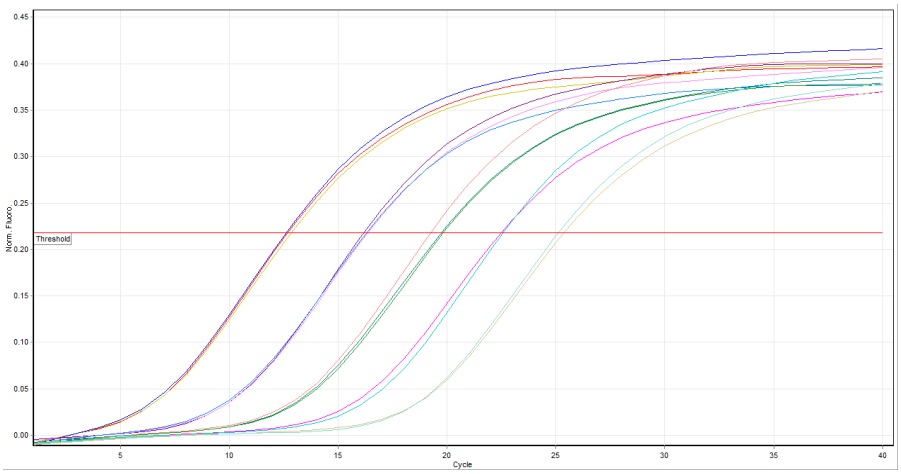


Standard curve (A), efficiency of amplification (B) and melt curves (C) for SF3

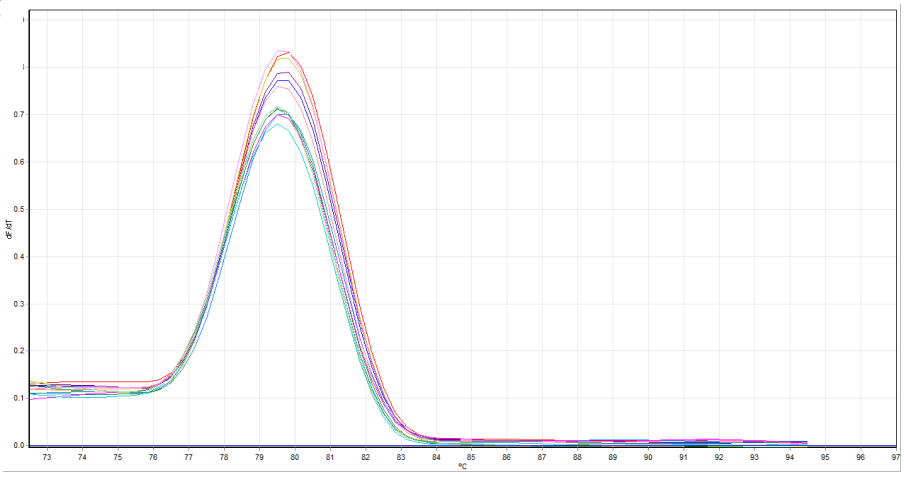
A



B

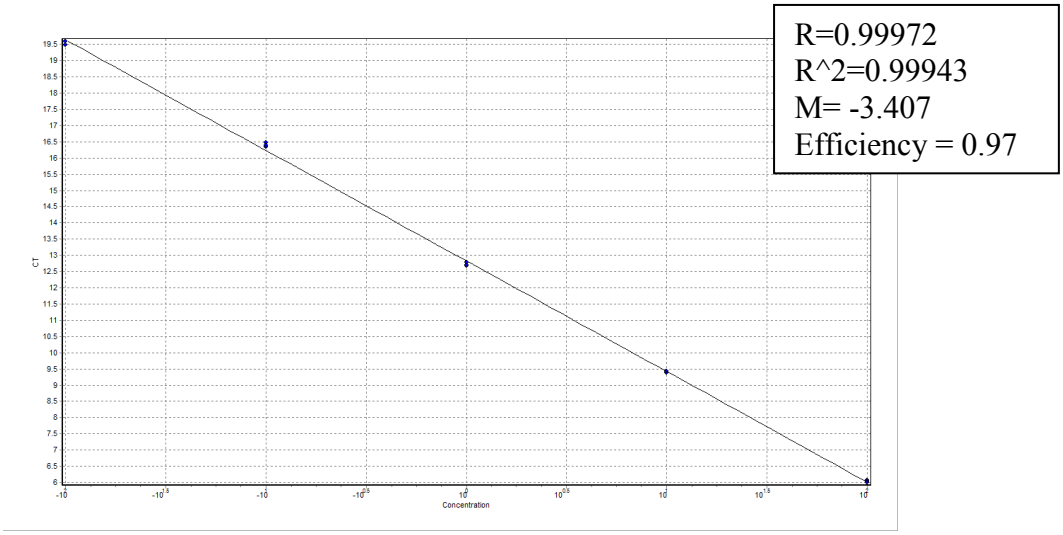


C

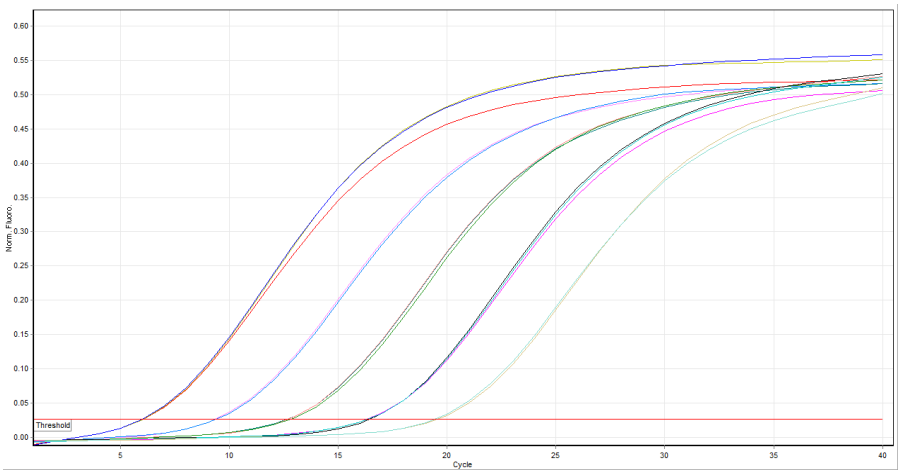


Standard curve (A), efficiency of amplification (B) and melt curves (C) for SF4

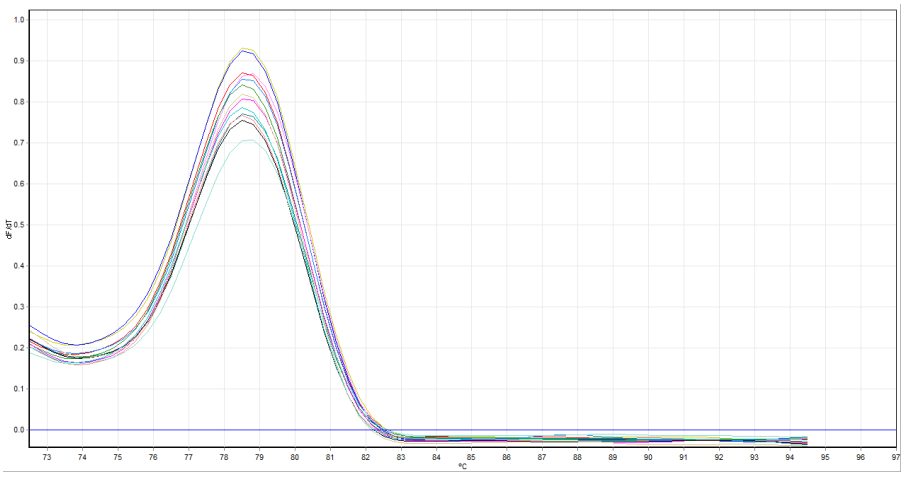
A



B

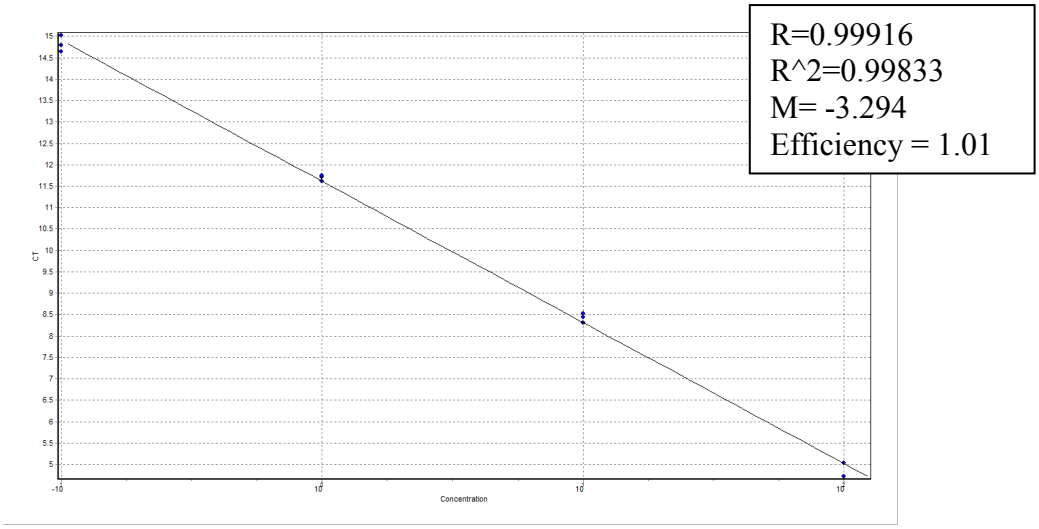


C

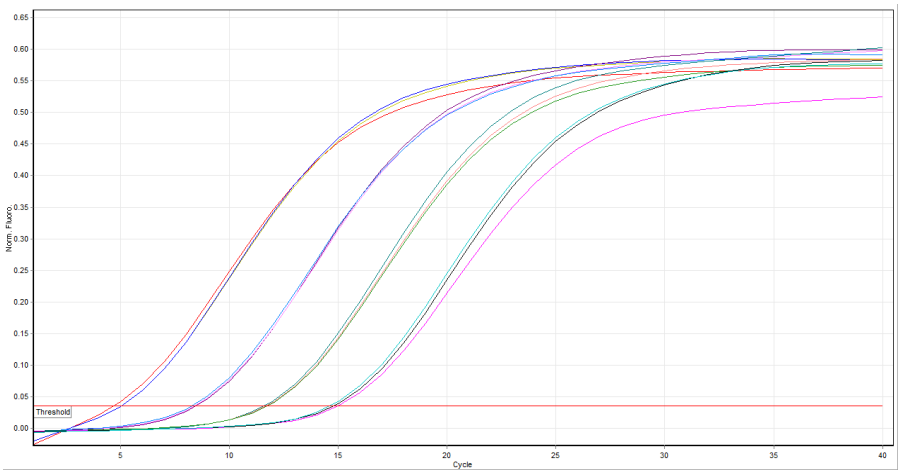


Standard curve (A), efficiency of amplification (B) and melt curves (C) for unspliced

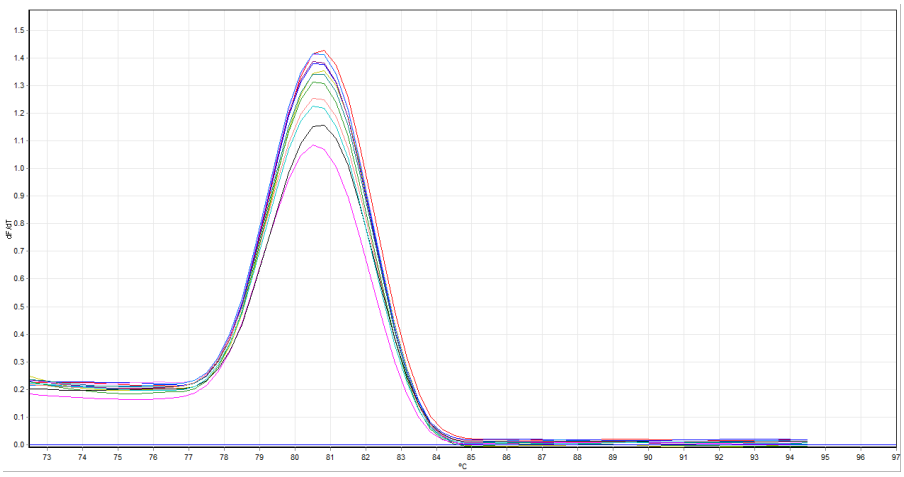
A



B

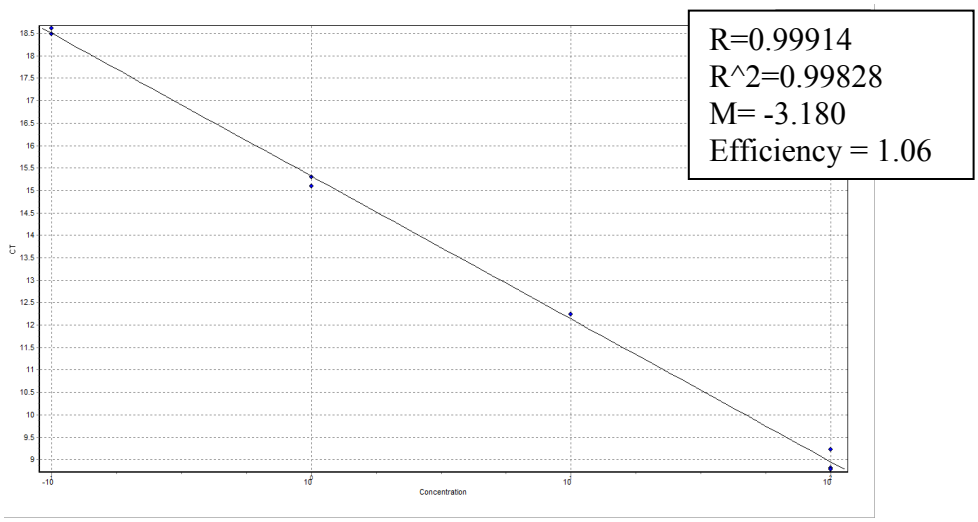


C

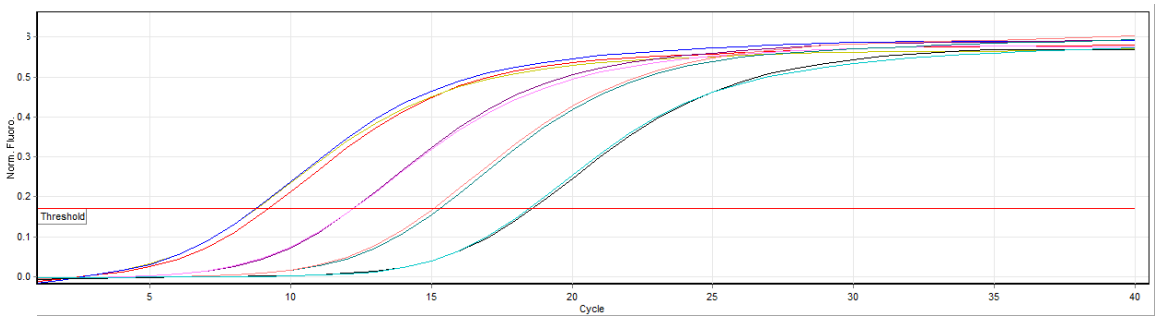


Standard curve (A), efficiency of amplification (B) and melt curves (C) for 5′ Exon

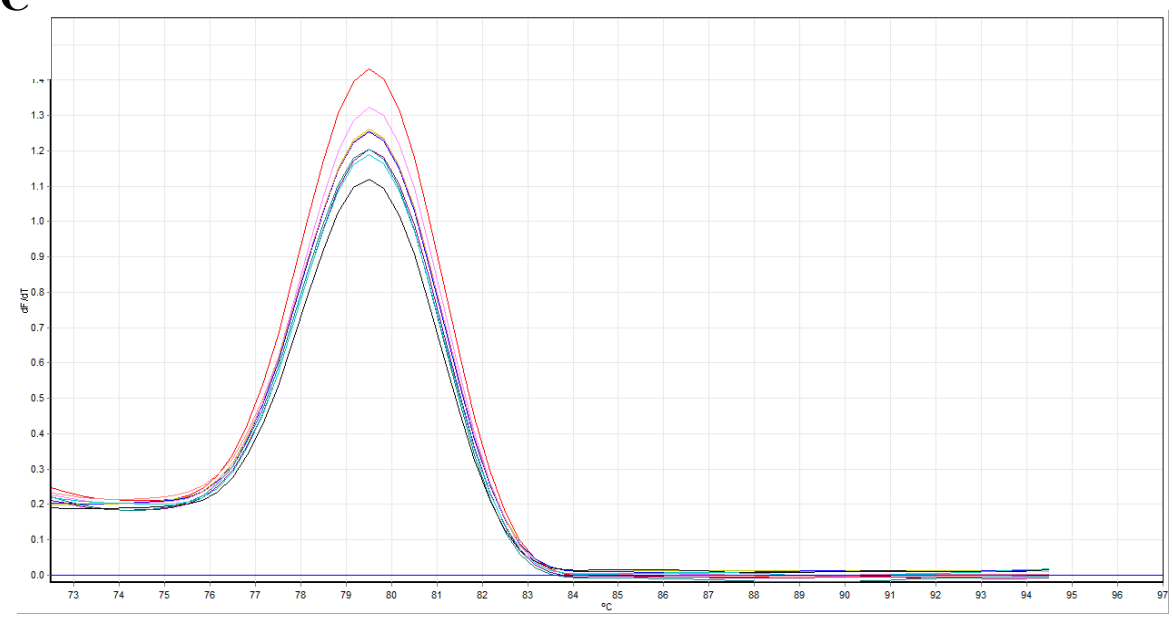
A



B



C



APPENDIX B - Copy Right Permissions

OXFORD UNIVERSITY PRESS LICENSE TERMS AND CONDITIONS

Dec 03, 2013

This is a License Agreement between Bonnie A McNeil ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3273921498719
License date	Nov 21, 2013
Licensed content publisher	Oxford University Press
Licensed content publication	Nucleic Acids Research
Licensed content title	Alternative splicing of a group II intron in a surface layer protein gene in Clostridium tetani:
Licensed content author	Bonnie A. McNeil, Dawn M. Simon, Steven Zimmerly
Licensed content date	11/08/2013
Type of Use	Thesis/Dissertation
Institution name	None
Title of your work	Alternative splicing of an ORF-less group II intron in Clostridium tetani
Publisher of your work	n/a
Expected publication date	Nov 2013
Permissions cost	0.00 USD
Value added tax	0.00 USD

TotalTotal	**0.00 USD**
TotalTotal	**0.00 USD**

Terms and Conditions

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details

provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

12. Other Terms and Conditions:

v1.4

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK501165274.
Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support:customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

OXFORD UNIVERSITY PRESS LICENSE TERMS AND CONDITIONS

Dec 03, 2013

This is a License Agreement between Bonnie A McNeil ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3273921304173
License date	Nov 21, 2013
Licensed content publisher	Oxford University Press
Licensed content publication	Nucleic Acids Research
Licensed content title	Alternative splicing of a group II intron in a surface layer protein gene in Clostridium tetani:
Licensed content author	Bonnie A. McNeil, Dawn M. Simon, Steven Zimmerly
Licensed content date	11/08/2013
Type of Use	Thesis/Dissertation
Institution name	None
Title of your work	Alternative splicing of an ORF-less group II intron in Clostridium tetani
Publisher of your work	n/a
Expected publication date	Nov 2013
Permissions cost	0.00 USD
Value added tax	0.00 USD
TotalTotal	**0.00 USD**

TotalTotal **0.00 USD**

Terms and Conditions

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

12. Other Terms and Conditions:

v1.4

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK501165272.
Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support:customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.