Group II introns in the *Bacillus cereus* group with unusual splicing properties

by

Fredrik Bernhard Stabell

Thesis for the degree of Doctor Philosophiae

Department of Pharmaceutical Bioscience

Faculty of Mathematics and Natural Sciences

University of Oslo

2009

© Fredrik Bernhard Stabell, 2009

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 913

ISSN 1501-7710

All rights reserved. No part of this publication may be
reproduced or transmitted, in any form or by any means, without permission.

Cover: Inger Sandved Anfinsen.
Printed in Norway: AiT e-dit AS, Oslo, 2009.

Produced in co-operation with Unipub AS.
The thesis is produced by Unipub AS merely in connection with the
thesis defence. Kindly direct all inquiries regarding the thesis to the copyright
holder or the unit which grants the doctorate.

Unipub AS is owned by
The University Foundation for Student Life (SiO)

Contents

Acknowledgements 3
List of papers 4
Introduction of mobile genetic elements 8
The prokaryotic mobilome 12
  Intercellular mobility 12
    Plasmids 13
    Bacteriophages 13
  Intracellular mobility 14
    Transposons 15
    Homing endonucleases 16
Introduction of introns 18
  Group II Intron 19
  Splicing reactions 20
    Branchpoint splicing (Forward and reverse) 20
    Hydrolytic splicing 21
    Other reactions: circularization and exon reopening 22
  Structure / Folding / Compaction / Catalysis 23
    Overview of the group II intron RNA domains and their major features 23
    Defining splicing boundaries 25
    The active catalytic site and stabilizing interactions 28
    Protein-assisted splicing *in vivo* 30
  Mobility: Retrohoming and retrotransposition 32
  Evolution: Origin and spread 34
The *Bacillus cereus* group of bacteria 37
    Bacillus cereus group “mobilome” 39
Aims of study 47
Summary of papers 48
Discussion 51
  General features of group II introns in the *Bacillus cereus* group 51
    Group II introns are widespread in the *B. cereus* group of bacteria 51
    Different structural classes of group II introns in the *B. cereus* group 52
    Localization and impact on host 54

The κ region ($ID^{(ii)}$) 56
Group II introns with an unusual 3' extension.. 60
The 3' extension may represent a new functional domain 60
Does the 3' extension fit into the active ribozyme, and if yes, how? 61
The origin of the extension 69
Concluding remarks 73
References.. 75
APPENDIX: PAPERS 1-5..89

Acknowledgements

This work was carried out from August 2003 to August 2009 at the Biotechnology Center and at the School of Pharmacy, Department of Pharmaceutical Biosciences, both at the University of Oslo. The work was funded by The Research Council of Norway (CAMST and FUGE II) and School of Pharmacy. I gratefully acknowledge this support.

I would like to thank Professor Anne-Brit Kolstø for taking me in and giving me the opportunity to stay in her research group. I am also thankful for her excellent supervision, but most of all for her enthusiasm and support under these years. It has been a great learning experience and I am extremely grateful for the freedom and implied trust that Anne-Brit has given me in my research.

My deepest thanks go also to my second supervisor and colleague Nicolas J. Tourasse. Working together with you has been very inspiring. You have always been extremely enthusiastic, ready to discuss and speculate.

I would also like to thank Ole Andreas Økstad for some important years working together on the *bcr*1 project, being both fruitful and enjoyable. Especially great thanks to my colleagues at the Bacillus cereus group, for all your help and encouragement, but most of all for creating such a stimulating and pleasant working environment.

Mom and Dad, I can never thank you enough. You have always wanted the best for me, supported, and encouraged me.

At last I want to thank Kaja, for all your support and patience during these years. Thanks for bearing with me when frustrated, and making me see things clear. It has been crucial for finishing this thesis.

Oslo, September 2009

Fredrik Stabell

List of papers

Paper 1

"Unusual group II introns in bacteria of the Bacillus cereus group."

J Bacteriol. 2005 Aug;187(15):5437-51.

Tourasse NJ; Stabell, FB; Reiter L; Kolstø AB.

Paper 2

"Group II intron in Bacillus cereus has an unusual 3' extension and splices 56 nucleotides downstream of the predicted site."

Nucleic Acids Res. 2007 Feb; 35(5):1612-23.

Stabell FB, Tourasse NJ, Ravnum S, Kolsto AB.

Paper 3

"A conserved 3' extension in unusual group II introns is important for the second splicing step."

Nucleic Acids Res., 2009 June; 37(10): 3202-3214

Stabell FB, Tourasse NJ, Kolstø AB.

Paper 4

"Group II introns carrying an unusual 3' extension show different splicing properties."

Stabell FB, Tourasse NJ, Kolstø AB.

Submitted

Paper 5

"Dispersal of a group II intron carrying an unusual 3' extension in bacteria of the *Bacillus cereus* group"

Tourasse NJ[§]*, Stabell FB[§], Kolstø AB.

Submitted

Other relevant papers not included:

“The bcr1 DNA repeat element is specific to the Bacillus cereus group and exhibits mobile element characteristics.”

J Bacteriol. 2004 Nov;186(22):7714-25.

Økstad OA; Tourasse NJ; Stabell, FB.; Sundfær CK.; Egge-Jacobsen WM; Risøen PA; Read TD.; Kolstø AB.

”Exploring the evolution of the Bacillus cereus group repeat element bcr1 by comparative genome analysis of closely related strains.

Microbiology. 2007 Nov;153(Pt 11):3894-908

Klevan A, Tourasse NJ, Stabell FB, Kolsto AB, Økstad OA.

“Orf2 from the Bacillus cereus linear plasmid pBClin15 encodes a double-stranded DNA binding protein.”

Letters in Applied Microbiology. 2009 Jan;48(1):51-7.

Stabell FB; Egge-Jacobsen WM; Risøen, P A; Kolstø,AB; Økstad, OA.

Summary

Mobile genetic elements have had, and still have an impact on the evolution of the genomes providing means for adaptation and structural organization. These elements are one of the major driving forces for the general evolution of all life forms. For the organisms and their genomes these elements are essential for development and adaptation to different environments.

The *Bacillus cereus* group of bacteria includes the related species *B. cereus* (*sensu stricto*), *B. thuringiensis*, *B. weihenstephanensis*, *B. mycoides*, *B. pseudomycoides*, and *B. anthracis*. These bacteria are very closely related at the genomic level, both in terms of gene content and synteny. Nevertheless, they show different phenotypic characteristics and pathogenic properties, and are altogether found worldwide in diverse habitats. Several of the major phenotypic characteristics of the members of the *B. cereus* group are determined by the different plasmids they have acquired. Besides these vehicles of genetic information, mobile genetic elements like transposons and group II introns do also induce some of the genetic and phenotypic variation and could therefore influence the dynamic behavior of the *B. cereus* group of bacteria.

Introns, or intervening sequences (IVS), are elements interrupting the sequence of genes. These are present in precursor mRNA and are removed by a process called splicing. The group II introns are a type of mobile retrotransposons that can also perform self-splicing. Group II introns are classified according to features of their RNA structure and the sequence of their intron-encoded reverse-transcriptase protein. The typical structure is made of six RNA domains (I-VI), which are involved in a network of tertiary interactions that fold the ribozyme into its catalytically active structure. Self-splicing proceeds in two steps via branching or hydrolysis pathways, releasing a lariat or linear intron, respectively *In vivo* the active intron need its own intron-encoded protein for splicing as well as for mobility.

The work presented in this Thesis starts with the classification and functional characterization of a total of eight group II introns present in the genomes of two strains of *B. cereus*, ATCC 14579 and ATCC 10987. The splice boundaries were as expected except for the *B.c.*I4 intron of *B. cereus* ATCC 10987, which spliced 56 nucleotides

downstream of the predicted 3' splice site. This extraordinary intron was then investigated in more detail. We showed that the extra 56-bp 3' segment is an integral part of the intron RNA molecule downstream of domain VI, while splicing through branching still occurred at the expected site. *B.c.*I4 represented therefore a unique arrangement never seen before, and our studies imply that the intron must have adapted to splice with the 3' extension. RNA secondary structure predictions suggest that the 56-bp segment folds into two stable stem-loop structures.

We later identified four new group II introns, *B.th.*I5, *B.th.*I6(a and b), and *B.th.*I7 from *B. thuringiensis* BGSC 4D1 that harbor a 3' extension similar to that of *B.c.*I4. This showed that the presence of a 3' extension was more common that previously thought and that *B.c.*I4 was not an isolated case. Surprisingly, these introns do not form a single evolutionary lineage even though the structure and sequence of the extensions are highly conserved. Furthermore, our *in vitro* splicing studies demonstrated that the larger of the two stems in the 3' extension is important for an efficient second-step splicing with the extension. Though the initial studies showed that the whole extension of *B.c.*I4 was not essential for splicing, later studies suggested that it has an effect on the balance between splicing via hydrolyis and splicing via branching. Most remarkably, analysis of *B.th.*I6 revealed that this intron does not appear to be able to perform an efficient second splicing step when the extension is removed as opposed to *B.c.*I4. This difference may come from evolutionary divergence that is accompanied by differences in specific (sub)domains of the secondary structure.

We have further reported five divergent copies of the *B.th.*I6 group II intron in five *B. cereus* and *B. thuringiensis* strains. By using sequence comparisons and phylogenetic analysis of the host gene of these introns from 43 different *B. cereus* group strains, we could infer several separate events of mobility, thus strongly indicating that the *B.th.*I6 intron is mobile with the 3' extension.

Altogether, the results presented here indicate that the 3' extension can be regarded as a functional domain VII that does contribute to the splicing properties, when present as an integral part of the intron. In addition to illustrating the adaptability and flexibility of group II introns, the study of these unusual introns has shed light on the structural and functional evolution of group II ribozymes in general.

Introduction of mobile genetic elements

"*Mobile DNA has been described as the genome's "dark matter": a significant part of its mass, difficult to understand, and often ignored. Transposable elements may also be seen as "dark energy," a dynamic force that not only accelerates expansion but also helps set the warp and weft of genomes, for better and for worse*" (Goodier and Kazazian 2008).

Barbra McClintock first discovered mobile genetic elements in the 1940's. She described a transposon in the chromosome of maize as a certain "mutable loci" responsible for phenotypic traits. Studies in the 1950's showed that plasmids and viruses could insert or excise pieces of DNA in the bacterial chromosome (Craig 2002). When one later realized the ubiquity of mobile-genetic elements, one first suggested that these elements and other noncoding DNA were largely a byproduct ("junk" DNA) of "selfish" elements proliferating within host genomes until opposed by natural selection (Doolittle and Sapienza 1980; Orgel and Crick 1980). Many have doubted this view, assuming that these elements, which account for about 45% of the human genome as opposed to only 1% that are dedicated to protein-coding sequences, would have been removed through natural selection and evolution if they were only useless and harmful. More and more evidence has pointed out that the genomes have coevolved with the mobile elements, restraining them from spreading all over and at the same time gaining from their presence. Mobile elements have and continue to impact the evolution of the genomes by providing means for adaptation and structural organization, and are also involved in regulating gene expression (Goodier and Kazazian 2008). The mobile DNA has certainly played an important role in the structure and evolution of genes and genomes from bacteria to humans.

The prokaryotic genomes are much more streamlined than the more complex eukaryotic genomes due to a strong selection for metabolic efficiency in large populations that are under strong selection (Lynch and Conery 2003; Lynch 2006). This "burden of bureaucracy" is certainly one of the important factors that have influenced

prokaryotic genome size and complexity, whereas the emergence of large eukaryotic genomes is proposed to have occurred passively in response to long-term reductions in population size that accompanied increases in organism size. According to this hypothesis, the genomic restructuration that has lead to the eukaryotic complexity has been first mediated by events in a non-adaptive process, which has then provided a substrate for further evolution of multi cellular species via a selective process (Lynch and Conery 2003). Mobile genetic elements in prokaryotes, which will be the focus in this introduction, have been important in the evolution that has lead to the complex eukaryotic genomes, and are also essential in the prokaryotic organisms and their genomic development and adaptation to different environments.

Impact of mobile elements on prokaryotic genome and evolution

Prokaryotic genomes can differ greatly in size, ranging from about 160 to 13 000 kb, where nearly 90% of the sequences (in most bacterial and archaeal genomes) specify proteins. Usually mobile genetic elements in prokaryotes do not account for a substantial part of the genomes as they do for higher eukaryotes. Even so, they are clearly one of the driving forces, together with genome degradation and streamlining, in the evolution and adaptation of prokaryotic species, and have played key roles in the development of bacterial pathogens (Ochman 2005; Pallen and Wren 2007; Koonin and Wolf 2008) Figure 1 illustrates the different factors involved in genome evolution (Koonin and Wolf 2008).

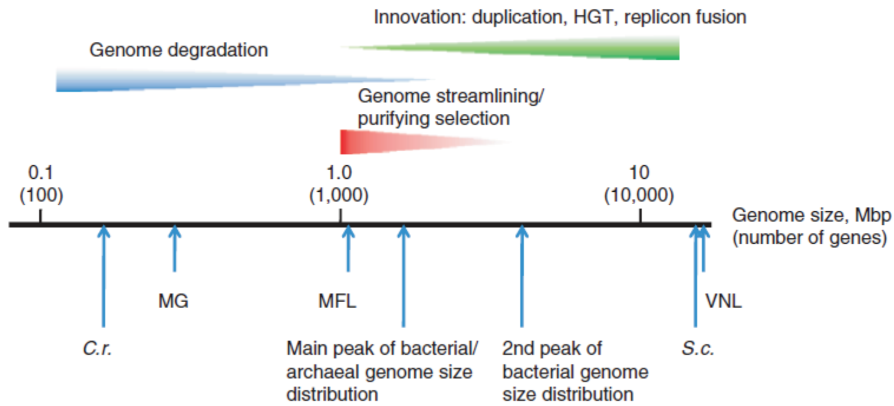


Figure 1. The figure is taken from (Koonin and Wolf 2008), and shows the principal forces of evolution in prokaryotes and their effects on archaeal and bacterial genomes. The colored triangles denote the effects of Genome degradation, Genome streamlining, and Innovation. They are positioned over the ranges of genome size for which the corresponding effects are thought to be most pronounced.

Prokaryotic organisms use multiple strategies involving acquisition, loss, and/or development of genes to adapt to new environments and optimize functionality under the strong selection pressure. Bacteria have a low mutation rate of 10^{-6} to 10^{-9} per generation and do not have, unlike most eukaryotic organisms, a mode of sexual reproduction which can exchange alleles within a population (Lynch 2006). Mobile elements, including viruses (bacteriophages), plasmids and transposable elements, are therefore the most dominant and critical force for acquisition and development of genes in the prokaryotic world (Koonin and Wolf 2008). These elements, referred to as the mobilome, constitute an enormous source of genetic material that can be exchanged between different bacteria, and are in constant exchange with the more stable chromosomes. The mobilome also serve as vehicles for horizontal transfer for other genetic material that regularly becomes passenger under mobility (See 'Intra- and Intercellular mobility' section below).

The content of mobile elements in the genomes varies with the ecology and lifestyle of the organism. This might reflect the importance of these elements in different environments and under certain conditions. Facultative and obligate intracellular bacteria that replicate within eukaryotic host cells usually have many and very few mobile

elements, respectively, while free-living bacteria generally have numbers in-between (Ochman and Davalos 2006). The reduced number of mobile elements in obligate intracellular bacteria can be explained by the accelerated deletion rates and the deleterious effects these elements have under the increased selection pressure and genome streamlining process affecting the former organisms (Lynch and Conery 2003; Moran and Plague 2004). In the initial stages when bacteria are becoming host-dependent, facultative intracellular, there appears to be a rapid burst in frequency of transposable elements together with many genomic rearrangements, while ancient, obligate intracellular organisms appear to have lost these elements. The proliferation of mobile elements in newly host-restricted bacteria is explained by the reduced effective population size that lowers the efficiency by which purifying selection maintains genes and this allows the "selfish" mobile elements to multiply with resulting increase in genomic rearrangement and pseudogene formation (Lynch and Conery 2003; Moran and Plague 2004). This transition stage, with higher frequency in mobile elements, may have a beneficial effect with an increased possibility to adapt to specific hosts and niches. Increased horizontal gene transfer might give a greater opportunity to acquire genetic material of selective advantage (Koonin and Wolf 2008). The spread of mobile elements within the genome can give more genome plasticity, because they can mediate gene deletion, genome rearrangement or have a direct impact on neighboring gene expression (Moran and Plague 2004). It may also be that these events are setting the stage for the more deleterious effects leading to genome streamlining and reduction, as in the "typical" ancient host-dependent bacteria (Parkhill et al. 2003; Moran and Plague 2004). It has further been argued that, more generally, an expansion in mobile elements in any organism could give a transitory selective advantage to the host (Siguier et al. 2006a; Wagner 2006). The expansion can give a beneficial increase in lateral gene transfers and genomic rearrangements, but in the longer term this could be detrimental to the host.

The prokaryotic mobilome

Mobility events of genetic material can be classified according to whether the mobility occurs within genomes (intracellular mobility) or between bacterial cells (intercellular mobility)- however, these two phenomena are usually closely connected. In the sections below, the main groups of mobile elements and mobility processes are presented briefly in the sections below.

Intercellular mobility

It has long been known that horizontal gene transfer occurs between different bacteria, but it was only with the rapid increase of genome sequences and the genome comparisons that the massive extent of genetic transfer has been recognized (Frost et al. 2005; Koonin and Wolf 2008). There are three classical mechanisms that mediate lateral exchange of genetic material between bacteria: natural transformation, transduction and conjugation (Frost et al. 2005).

Natural transformation is the ability to take up naked DNA from the environment, integrate and functionally express the foreign DNA (Chen and Dubnau 2004). This free DNA may come from other dead, lyzed cells. The transformation usually require that the cells come in a physiological state named *competence*, which is a specific response to environmental conditions regulated by a specific set of genes (Thomas and Nielsen 2005). The naked extracellular DNA is actively transported as single-stranded DNA (ssDNA) into the cytosolic compartment by a translocation machinery.

Horizontal genetic transfer by conjugation differs from transformation as it involves direct cell-to-cell contact for the transfer of DNA. The two cells establish contact through specialized transfer pores, by which the DNA is translocated from donor to recipient cell (Frost et al. 2005). The exchanged DNA usually consists of independently replicating genetic elements such as conjugative plasmids or transposons, which encode specific sets

of genes mediating the transfer process. Other plasmids that lack the conjugation system can be mobilized by conjugative elements.

Transduction is the genetic transfer mediated by bacteriophages (see also 'Bacteriophages' section below) (Lengeler et al. 1999). This transfer is an accident of the replication of the phages (general transduction). At a low frequency, DNA from the host is brought with the viral DNA and injected into the next infected cell (specialized transduction). Through recombination or phage integration, the foreign DNA becomes part of the host cell's genetic material.

Plasmids

Plasmids are extra chromosomal DNA elements that are common in prokaryotes. Plasmids can vary in size from 2 kb to more than 1 Mbp, i.e., larger than the smallest known chromosomes. This genetic material forms a stable, self-replicating entity (Phillips and Funnell 2004a). Plasmids encode the genes essential for their own replication, but also genes not required for essential cellular function. The latter ones might give a selective advantage in certain environments, encoding for example antibiotic resistance, secondary metabolic capabilities or virulence factors (Phillips and Funnell 2004b). These genetic entities can usually be horizontally transferred by conjugation, but also by transformation. They display greater genomic plasticity than chromosomes, and are therefore more adaptable and can act as reservoirs for horizontal genetic transfer.

Bacteriophages

Bacteriophages are viruses that infect bacteria, and are dependent on their host for replication. Phages are known to contribute to fitness and pathogenesis of the bacteria as they can be gene transfer particles to shuttle pathogenicity islets or random samples of chromosomal DNA (Brussow et al. 2004). The phages carry first of all their own genetic material that can either be double stranded DNA (dsDNA), singelstranded DNA (ssDNA), dsRNA or ssRNA. This material is usually enclosed by a protein capsid. The

phages recognize and attach to specific receptors features in the membranes of the bacterial cell wall before it injects the genetic material into the bacteria. This specificity for certain receptors do in turn determine what bacteria strains the phages can infect, though it has recently been shown that the infection can happen more interspecies than previously thought (Chen and Novick 2009).

The phages can integrated their genome with host bacterial DNA or become established as plasmids, and then be reproduced with the host cell which is allowed to continue to survive and replicate. These, called prophages, may provide benefits to the host bacterium while they are dormant by adding new functions to the bacterial genome in a phenomenon called lysogenic conversion (Brussow et al. 2004).

Intracellular mobility

There are a wide variety of mobile elements that can mediate their own, and other genetic elements, transfer to new genomic locations. Transposons are a class of elements that is a defined segment of DNA with the ability to move, or copy itself, into a second location without requirement for DNA homology (Curcio and Derbyshire 2003). Transposable elements are the focus here, but other mobile elements as homing endonucleases and repeated sequences are presented.

Intercellular mobility events may both activate and inactivate genes depending on the location of their target (upstream or within a gene, respectively). These elements can promote inversions and deletions of chromosomal DNA, as a result of an intramolecular transposition event or by providing dispersed regions of homology that can be recognized by the DNA recombination machinery of the host. Under mobility the DNA flanking certain elements can also be mobilized, as with transduction, and so provide yet another means of rearranging host genes. On the other side you have the group II introns, a retrotransposons, which by its ribozyme activity can splice the host gene exons and therefore minimize the effect of the transposition.

Transposons

Transposable elements can be divided into two major groups with the DNA transposons, class II elements, and the retroelements, class I elements by which mobility involve reverse transcription of an RNA intermediate. Both the two classes can be further divided into additional classes based on their variation of mechanisms (Curcio and Derbyshire 2003; Goodier and Kazazian 2008). Their might also be some confusion about class I and II transposons, as DNA transposons has also been divided into class I and II based on that the latter represented transposons that carried additional genes not needed for mobility (Craig 2002). Anyway, the major division here is by DNA transposon and retrotransposon (Curcio and Derbyshire 2003 ; Beauregard et al. 2008).

DNA transposons

All the transposase enzymes possess a nuclease activity that allows them to cleave DNA in order to excise transposon DNA, which is subsequently inserted into a new location. Some of these transposons cut out the defined DNA sequence, whereas others only make a copy of the original sequences. Similarly, some of these elements 'paste', whereas others 'copy', themselves into the target. To cleave the DNA substrate different types of nucleophiles are used depending on the system. The phosphorus atom of a backbone phosphate group can be attacked by water that is activated by enzyme-bound metal ions, a hydroxyl group at the 5' or 3' end of a DNA strand, or a hydroxyl-group bearing amino acid in the active site of the transposase itself (Curcio and Derbyshire 2003; Grindley et al. 2006). The different mechanisms are used to classify the transposase into four different protein families that mediate transposition (Curcio and Derbyshire 2003). These families are DDE transposases, which are the most abundant class, Y- and S-transposases, and Y2 enzymes (Craig 2002; Curcio and Derbyshire 2003).

Insertion sequences (IS) are the smallest and most frequent transposable elements in prokaryotes (Craig 2002; Curcio and Derbyshire 2003). IS elements are widespread in eukaryotes, prokaryotes, phages, and plasmids. These genetic elements are

usually flanked by inverted terminal DNA repeats (IRs) of between 10 and 40 bp, which is recognized by the transposase. There are several exceptions - for example the widespread class of IS200/605 that are defined by secondary DNA structures close to the cleavage site (Kersulyte et al. 2002; Ronning et al. 2005; Ton-Hoang et al. 2005).

Retrotransposons

Retrotransposons generate a copy of their DNA (cDNA) by reverse transcription of their RNA. The insertion into new genomic locations can occur through different mechanisms, where they can be divided into two main classes (Beauregard et al. 2008). These two classes are named TP and EP retrotransposons, also known as non-LTR and LTR retrotransposons. TP retrotransposon, that stands for **t**arget-**p**rimed retrotransposon are defined by a mechanism where the cDNA copy is transcribed after the RNA element has inserted into the new DNA loci. EP retrotransposon on the other side, **e**xtrachromosomally **p**rimed retrotransposon transcribed its cDNA copy before inserting into the new site.
EP retrotransposons encodes a recombinase or integrase in addition to reverse transcriptase, which through a recombination event inserts the reverse transcribed dsDNA. Examples of these longs terminal repeats elements are Ty elements of *Saccharomyceerevisiae,* but also retroviruses that also use a DNA based integration method (Curcio and Derbyshire 2003; Beauregard et al. 2008). The TP retrotransposons comprises of many different elements, such as the mammalian long interspersed nuclear elements (LINEs), short-interspersed nuclear elements (SINEs) and the group II introns.

Homing endonucleases

Homing endonucleases are encoded by open reading frames that are usually embedded within group I, group II, archael introns and inteins (intervening sequences that splice out at the protein level), but can also be standalone copies (Stoddard 2005). The endonuclease recognizes and generates a double-strand break at homologous intron-/homing endonuclease-less DNA sites, which is then repaired following the double-strand

break repair (DSBR) or the synthesis dependent strand annealing (SDSA) pathway (Mueller et al. 1996; Edgell et al. 2000; Craig 2002). Both these pathways are dependent on homologous recombination between exon sequences and lead to copy of the intron as well as conversion of part of the flanking exons from the intron donor DNA into the recipient DNA. Homing endonucleases are generally very site-specific, 10-40 bp target sequence, allowing insertion into cognate target sites as well as in additional ectopic sites that broaden the range homing endonuclease mobility (Stoddard 2005).

Repeated sequences

Noncoding repeated sequences are present in the genomes of various bacteria. These sequence elements can be non-autonomous miniature inverted repeat transposable elements (MITE) (Redder et al. 2001; Feschotte and Wessler 2002; Siguier et al. 2006b). The MITEs are sequences are relatively short and flanked by conserved terminal repeats. These sequences are believed to be derived from, and to be trans-mobilized by related transposons outside the repeated element with similar ends (Siguier et al. 2006b). Repeated elements as REP and ERIC are other classes of highly repeated elements with similar structures to MITE (De Gregorio et al. 2005; Tobes and Ramos 2005).

The repeated elements have been associated with a diverse set of possible functions, where some are promoter activity, transcription termination, regulation of mRNA stability, and DNA uptake or recombination signals (Siguier et al. 2006b; Delihas 2008). These sequences participate in maintenance and evolution of chromosome structure and function as suggested above other mobile elements in general by mediating genome plasticity. Suggestions have been made that repetitive DNA elements can act as a source of mutation that convey adaptive benefits likely to happen at a higher frequency in the genome (Schmidt and Anderson 2006).

Introduction of introns

Introns are also called intervening sequence (IVS) and constitute the DNA regions in a gene that are not translated into proteins. These non-coding sections are present in precursor mRNA (pre-mRNA) and are removed by a process called splicing. The main classes of introns are the group I introns, group II introns and the spliceosomal introns. They are all characterized by that they splice by two transesterification reactions, but differ in RNA structure and what they use as catalyst for the reaction. Group I intron differentiate from the two other in RNA structure and sequence, but also by that it use primarily an external nucleophile (free guanine nucleoside) to initiate the first splicing step (Woodson 2005; Stahley and Strobel 2006). This intron was the first RNA molecule assigned with catalytic properties and Thomas Cech was awarded the Nobel Prize in chemistry for the discovery in 1989(Kruger et al. 1982). Both the group I and II introns are true RNA enzymes, ribozymes that can catalyze their own splicing reaction. These are ribozymes that usually dependent on divalent ions to fold into their active structure (Lehmann and Schmidt 2003). The group II introns usually have an unpaired adenosine in their conserved secondary RNA structure which acts as the nuclophile that initiates the first splicing step (See later). Spliceosomal intron is a large ribonucleoprotein complex consisting of several small nuclear RNAs and proteins factors. These introns, that usually reside in the eukaryotic protein coding genes, are due to similarities in important structures features and splicing mechanisms believed to derive from the group II intron (Robart and Zimmerly 2005; Seetharaman et al. 2006). However all the similarities, it is still debated whether the huge spliceosomal complex is a true ribozyme or not (Collins and Guthrie 2000; Valadkhan 2007; Valadkhan et al. 2007; Abelson 2008; Michel et al. 2009). There are other introns that are non-selfsplicing, tRNA and archeal introns, which rely on host factors for splicing (Calvin and Li 2008).

Group II Intron

The group II introns are a type of mobile TP-retrotransposons that can also perform splicing. Group II introns were discovered more than 25 years ago (Michel and Dujon 1983). They are present in mitochondria and chloroplast of plants, fungi and lower eukaryotes, where they are relatively abundant. They are also found in ~25% of the sequenced bacterial genomes and in a few archaea (Dai and Zimmerly 2002a; 2003; Toro 2003). Excision from RNA precursor molecule occurs primarily through a branching or hydrolytic splicing reaction, and the intron can subsequently insert into a new DNA site through the reverse reaction, reverse-splicing (Lehmann and Schmidt 2003; Lambowitz and Zimmerly 2004; Pyle and Lambowitz 2006; Fedorova and Zingler 2007; Toro et al. 2007). The intron RNA is defined by a highly conserved secondary structure that typically consists of six domains (I to VI) connected by a network of tertiary interactions. Domain IV usually contains an open reading frame coding for a multifunctional intron-encoded protein (IEP), which is required for both splicing and mobility events *in vivo*. Only a few introns have been functionally characterized *in vitro* or *in vivo*, but at least one for each of three major structural classes is represented. Further, most of the sections below build on the *in vitro* studies conducted without the IEP on an even narrower group of introns.

Splicing reactions

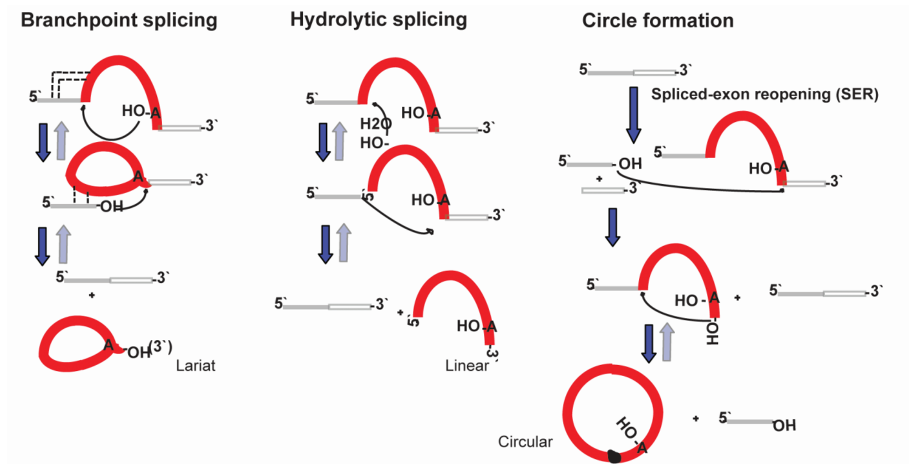


Figure 2. Schematic drawings of the reaction pathways of group II introns. The intron is shown as a thick solid red line and the 5' and 3' exons as a grey and white box, respectively. The adenosine residue and the water molecule acting as nucleophiles are drawn with their 2' hydroxyl group that initiates the first step of branchpoint splicing and hydrolytic splicing, respectively. The minor reaction pathway leading to intron circle formation is also drawn. This reaction is believed to result from the spliced-exon reopening reaction. Nucleophilic attacks are indicated by black arrows. Dark blue and grey arrows indicate the forward and reverse direction of a given reaction step, respectively) See text for more details

Branchpoint splicing (Forward and reverse)

Ribozyme activity was the first property assigned to group II introns, i.e., splicing is catalyzed by the intron RNA (van der Veen et al. 1986; Peebles et al. 1987). A major pathway by which group II intron excise themselves from the exons is the branchpoint pathway (branching). Branch point splicing occurs by two transesterification reactions as a two-step process (Figure 2) (Lehmann and Schmidt 2003). This starts with a nucleophilic attack at the 5' splice site by the 2'OH group of a specific unpaired adenosine in domain VI (the branchsite)(See Figure 2). The 5' splice site is put by several interactions in close proximity for the 2'OH group to attack and break the phosphodiester bond of the 5' junction in a S_N2 displacement mechanism (Padgett et al. 1994; Podar et al. 1995). This releases the 3'OH group in the 5'exon and the intron forms a lariat intermediate with a 2'–5' linkage between the branch site adenosine and the first intron nuclotide, i.e. a branched RNA circle with a 3'-tail still covalently attached to the 3' exon

(van der Veen et al. 1986; Lehmann and Schmidt 2003). After the first cleavage reaction the 5'exon is still tightly linked to the intron via base-pairing interactions (Jacquier and Michel 1987; Jacquier and Jacquesson-Breuleux 1991), and its 3'hydroxyl group is positioned for attacking the 3' splice site in the second splicing reaction. This leads to the release of a free intron lariat and ligated exons. The second step proceeds via a S_N2 displacement mechanism like the frist step, but a phosphate substitution at the two splice sites has revealed inverted stereoisomeric preferences (Padgett et al. 1994; Podar et al. 1995). Group I introns differs in this contrast markedly as the two steps appear to be simple reversal of one another (McSwiggen and Cech 1989; Pyle and Lambowitz 2006). Group II introns are dependent on divalent metal ions for folding and catalysis and have a two-metal ion coordination for the leaving groups at the catalytic center, like group I introns and protein enzymes that catalyze phosphoester transfer (Piccirilli 2008; Toor et al. 2008a).

Both these two splicing reactions are reversible (Figure 2)(Pyle and Lambowitz 2006). The first step is the rate limiting for most self-splicing group II introns, and the rate constant is equal in the forward and reverse direction (Chin and Pyle 1995). The intermediates are usually not detected as the second forward reaction is much faster than the reverse and thus drives the forward reaction to completion. However, under suitable reaction conditions, reverse splicing can be rather efficient (Muller et al. 1991; Aizawa et al. 2003). This property is not limited to RNA substrates as reverse splicing also can introduce the introns into DNA molecules and thus provides the basis for intron mobility (Lambowitz and Zimmerly 2004).

Hydrolytic splicing

In addition to the lariat splicing pathway, where the nucleophile is internal, the group II intron can excise itself via a hydrolytic pathway where the nucleophile attacking the 5'exon-intron junction in the first splicing step is water or a hydroxyl ion (Lehmann and Schmidt 2003; Pyle and Lambowitz 2006). This releases the 5' exon and a linear intron attached to the 3' exon (Figure 2). In contrast to branching, hydrolysis is

irreversible. The second step is identical to that of the branching pathway and the end products are ligated exons and a linear intron. *In vitro* the balance between the branching and hydrolysis reactions is strongly influenced by the choice of monovalent cation used (Daniels et al. 1996). This balance may also depend on the sub class of group II intron and *in vitro* some introns have been shown to only splice only through the hydrolytic pathway(Granlund et al. 2001). Further, *in vivo* the hydrolytic reaction is an active pathway for introns lacking the branchpoint nucleotide (Podar et al. 1998; Vogel and Börner 2002). It has recently been shown that a linear intron can reverse the second splicing in a efficient way, and may suggest an alternative pathway for mobility (Roitzsch and Pyle 2009).

Other reactions: circularization and exon reopening

Both products the free lariat or linear intron, released during the branching and hydrolytic splicing pathway, respectively, have been shown *in vitro* to reopen spliced exons (Jarrell et al. 1988; Daniels et al. 1996). This alternative reaction, shown for some group II introns, is the hydrolysis of the 5’ -3’ exon junction after recognition by excised intron molecule (Lehmann and Schmidt 2003; Fedorova and Zingler 2007). This is actually true ribozyme activity, leaving the intron unchanged, occurs surprisingly with same stereo chemistry as second reaction although it cleavages after the same position as the first (Podar et al. 1995; Lehmann and Schmidt 2003; Michel et al. 2009). This spliced exon reopening (SER) reaction has also been implicated for generation of intron circles. A fully circular intron form, first discovered as a by-product in *in vitro* splicing, has been shown *in vivo* for bacteria and in plant mitochondria (Murray et al. 2001; Li-Pook-Than and Bonen 2006; Molina-Sanchez et al. 2006). In the circularization pathway, a free 5’exon, from the SER reaction is suggested to attack the 3’ splice site of an unspliced precursor mRNA, leaving a 5’exon still covalently linked to the intron. The free 3’ end of the intron then attacks the 5’ splice site, releasing the 5‘exon and a circular intron with a 2’-5’linkage at the circle junction (Murray et al. 2001).

Structure / Folding / Compaction / Catalysis

Overview of the group II intron RNA domains and their major features

Group II introns are the largest catalytic RNAs known, with a size of about about 400-1000 nt (excluding the IEP open reading frame (ORF)), and the size is one of the most diffucult challenges when analyzing the structure/function relationships of group II introns compared with other ribozymes. Strikingly, even though these RNA elements lack sequence conservation, they show conservation in secondary structure features and organization. Group II introns typically form a secondary structure made up of six domains, that radiate out from a central core bringing the 5' and 3' splice sites in close proximity (Figure 3)(Lehmann and Schmidt 2003; Pyle and Lambowitz 2006; Fedorova and Zingler 2007). These domains of group II introns have specific roles in folding, conformational rearrangements, and/or catalysis. Domain I serves as a scaffold for the assembly of other domains into a catalytically active structure, and is essential for exon recognition (see section below), which explain this domain's importance for splicing and mobility (Pyle and Lambowitz 2006). Domain V is the main catalytic center (heart) of these ribozymes, and together with domain I are the only elements that are absolutely essential for minimal catalytic activity of the intron (Koch et al. 1992). Domain VI is necessary for the branching pathway, as it contains the branch point adenosine (Lehmann and Schmidt 2003). Domain II and III have been shown to enhance catalytic efficiency, but are not essential for the intron (Qin and Pyle 1998; Fedorova et al. 2003; Fedorova and Pyle 2005). The intron domain IV is the most varying region in secondary RNA structure, and can contain the multifunctional intron encoded protein (IEP). This structure does not directly contribute to catalysis under the splicing, but when present it influences both splicing and mobility (Fedorova and Zingler 2007). Lastly, the linker regions between the different domains have been shown to be important in several aspects of intron folding and catalysis (De Lencastre and Pyle 2008). Even though all group II introns fold into a similar overall secondary structure, they can divide into three major subclasses, IIA, IIB and IIC by IEP sequence analysis and correlating specific secondary structural features (Toor et al. 2001; Toro 2003; Simon et al. 2008).

Group II introns fold into a compact catalytically active tertiary structure by means of an extended network of long-range tertiary interactions that are distributed throughout the RNA secondary structure, forming a somewhat similar catalytic active center (Pyle and Lambowitz 2006; Fedorova and Zingler 2007; Waldsich and Pyle 2007; Dai et al. 2008).

Until recently most work/studies related to interactions catalytic activity, folding has been done on IIA and B class introns and therefore most interactions classified has been based on them (Lehmann and Schmidt 2003; Noah and Lambowitz 2003; Pyle and Lambowitz 2006; Dai et al. 2008). The recently determined 3D X-ray crystal structures of the group IIC intron of *Oceanobacillus iheyensis* have confirmed and extended the conclusions of numerous biochemical and genetic studies and have given new insights into the compaction and tertiary arrangement of group II ribozymes, which will be reviewed the next two sections (Toor et al. 2008a; Toor et al. 2008b). However, some aspects of the 3D structure are still unresolved (Michel et al. 2009).

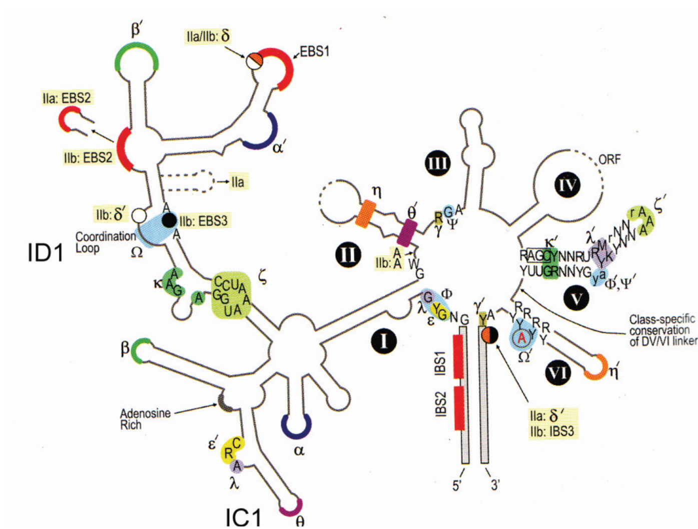


Figure 3. The figure taken from (Pyle and Lambowitz 2006) and slightly) modified, is an illustrative model of the secondary and tertiary structure of group II introns from the IIA and IIB structural classes. The six main RNA domains are designated by Roman numerals (I-VI). Domain IV typically encodes the multifunctional IEP open reading frame (ORF). Tertiary interaction motifs are colored and denoted with their respective Greek letters. The motifs specific for IIA or IIB introns are indicated. The location of subdomains IC1 and ID1 in domain I, which are specifically discussed in the text below, has been included.

Defining splicing boundaries

Group II introns RNAs recognize their targets sites, either being RNA or DNA in for the forward or reverse splicing reaction, via specific basepairing (bp) interactions with the exon sequences (Lehmann and Schmidt 2003; Lambowitz and Zimmerly 2004; Pyle and Lambowitz 2006). For group IIA and IIB introns the 5’ exon is defined through two 5-6 bp interactions with domain I, where the exon binding sequences (EBS) 1 and 2 in the intron pair to their corresponding intron binding sequences (IBS) 1 and 2 spanning the last 12-15 bp of the 5’ exon (See Figure 3). The resulting two recognition duplexes are what mediates the high-interaction specificity and cleavage-site fidelity, giving the proper conformation of the 5’splice site for transesterification or hydrolytic cleavage (Jacquier

and Michel 1987). Class IIC introns differ from IIA and IIB introns with respect to 5'exon definition, as they preferentially insert down-stream of a transcriptional terminator stem–loop structure which substitutes in part for the missing IBS2–EBS2 interaction (Toor et al. 2006; Robart et al. 2007). The 3'exon is defined by two, single-base-pair interactions, which also vary between the RNA structural classes. These interactions are: (1) γ-γ', involving a nucleotide located in the J2/3 linker region between domains II and III (γ) and the last intron base (γ'); and (2) δ-δ' for IIA introns or EBS3-IBS3 for IIB and IIC introns, where δ/EBS3 are in different locations in domain I and δ'/IBS3 is the first base of the 3' exon (Jacquier and Michel 1990; Costa et al. 2000). Class IIB and IIC introns also have δ-δ' interactions, but with a different δ' nucleotide positioned in the coordination loop of domain I and involved in a different aspect of exon recognition (see below). The γ-γ' and EBS-IBS3 interactions seem to play a minor role in splice site recognition for IIB introns, as disruption of these interactions affect mainly the efficiency of the second splicing step but not the fidelity of 3' splice site selection (Costa et al. 2000; Lehmann and Schmidt 2003). In contrast, IIA introns appear to be somewhat more sensitive to substitutions in the γ-γ' nucleotides, which can lead to the use of cryptic 3' splice sites (Lehmann and Schmidt 2003). Furthermore, domain VI is thought to guide the 3' intron-exon junction into the catalytic active site in passive way, ensuring an efficiency second splicing step with high fidelity (Jacquier and Jacquesson-Breuleux 1991; Lehmann and Schmidt 2003)

Biochemical cross-linking experiments have shown that both exons, the branch site, domain V and other elements critical to splicing are proximal and aligned in the correct orientation in a single catalytic center before the first splicing reaction (de Lencastre et al. 2005). These and other studies showed that (the formation of) this active center is in part facilitated by an internal asymmetric loop in subdomain Id in class IIB and IIC introns, which is referred to as the coordination loop (See Figure 3)(Costa et al. 2000; Noah and Lambowitz 2003; de Lencastre et al. 2005; Hamill and Pyle 2006). It plays a critical role in catalysis as it functions as a receptor for the branchsite and surrounding nucleotides in domain VI, and coordinates the docking of all components essential for splicing (de Lencastre et al. 2005; Hamill and Pyle 2006). The coordination loop contains EBS3 involved in the EBS3-IBS3 interaction, and the δ' nucleotide

involved in the δ-δ’ basepairing with the δ base located 5’ of EBS1. Crystallization of the *O. iheyensis* class IIC intron with ligated exon substrate has revealed that the exon junction is presented as a continuous strand over the important active sites in domain V (Toor et al. 2008b). The study also confirmed that the EBS1 and EBS3 motifs are linked together in a common exon binding interface by the δ-δ’ interaction in the coordination loop. Biochemical cross-linking studies suggest that binding of the branch site and the 3’exon bind the coordination loop independently and are energetically uncoupled (Hamill and Pyle 2006). Even though class IIA introns do not harbor the coordination loop, they are also suggested to form an overall topology with a similar catalytic core (Noah and Lambowitz 2003; Dai et al. 2008). These introns form the continuous binding interface for 5’ and 3’ exon recognition with EBS1 and the δ nucleotide that respectively binds IBS1 and the first 3’ exon nucleotide δ’(Jacquier and Jacquesson-Breuleux 1991). The main difference is presumed to be that for the class IIB introns the exons are largely internalized as opposed to the model for IIA introns, where they are mostly bound to the surface of the ribozyme beside the splice junction (Pyle and Lambowitz 2006; Dai et al. 2008).

The active catalytic site and stabilizing interactions

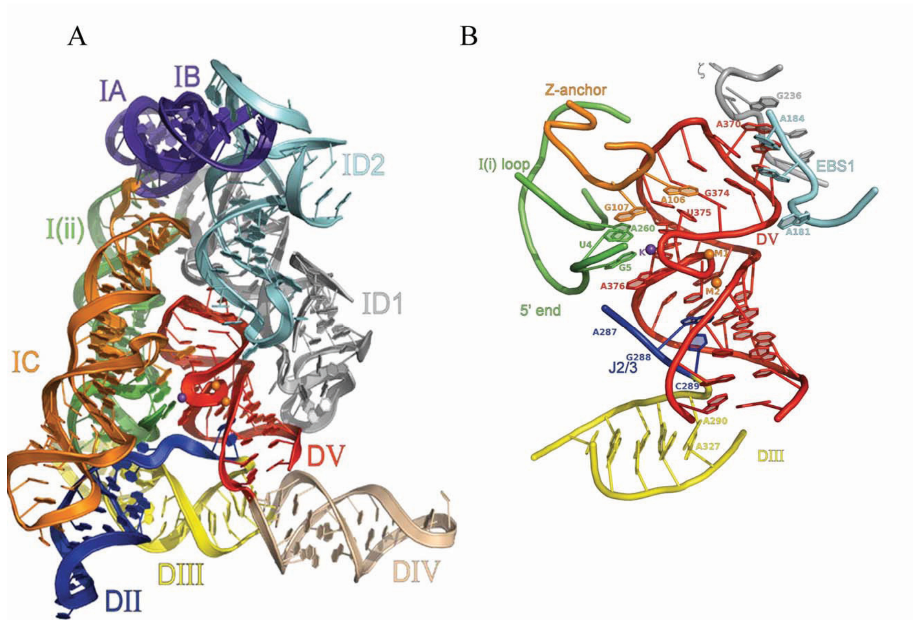


Figure 4. The pictures are taken from (Toor et al. 2008a) and show A) the overall structure in a ribbon representation of the group IIC intron of *Oceanobacillus iheyensis* based on the X-ray crystal, and B) domain V and residues in its close proximity with the interactions between domain V, domain I and the linker region J2/3. The various domains or subdomains of the group II intron are indicated by roman letters and drawn in distinctive colors.

Domain V (DV) is the catalytic heart of the intron. Almost every (most) nucleotide in DV has a major role in the intron's function and this domain is the most phylogenetically conserved primary sequence part of the entire intron (Lehmann and Schmidt 2003; Fedorova and Zingler 2007). At the 5' base of DV there is the catalytic triad AGC, or CGC, and together with a two-nucleotide bulge at the 3' side forms a negatively charged pocket by its backbone moieties that binds the two coordinating metal ions, consistent with a two-metal ion mechanism for catalysis (Sigel et al. 2000; Zhang and Doudna 2002; Lehmann and Schmidt 2003; Sigel et al. 2004; Pyle and Lambowitz 2006; Toor et al. 2008a; Toor et al. 2008b). The catalytic center is assembled via an extensive network of tertiary interactions between the multiple intron domains, which leads to a structure with a highly internalized core. As mentioned above, the exons are presented into the pre-

formed intron structure, where DI forms a scaffold for all the domains, with DV in the middle and the other domain structures stacking upon each other (Dai et al. 2008; Toor et al. 2008a; Toor et al. 2008b; Michel et al. 2009). The large DI is held together through the α-α' Watson-Crick basepairing interaction, and the additional β-β' pairing in some intron subclasses (Toor et al. 2001; Simon et al. 2008), and takes part in the pre-organization of the intron structure to the active form (Figure 3 and 4). The conserved α-α' pairing was shown to be functionally important (required) for self splicing *in vitro* (Harris-Kerr et al. 1993). DI folds independently of the other domains, which is the rate limiting step for the total folding of the intron (Fedorova and Zingler 2007). It has been demonstrated that the a small substructure/ region in domain, ID1, is the most crucial for compaction and folding (See figure 3 and 4) (Waldsich and Pyle 2007). This region, designated the folding control element, harbors the docking sites for domains V (the κ and ζ elements) and VI (the coordination loop), therefore suggesting that proper folding here ensures the specificity and accuracy of group II intron ribozyme catalysis (Fedorova and Zingler 2007; Waldsich and Pyle 2007). The interactions between ID1 and DV are named ζ-ζ' and κ-κ', where the former is a tetraloop-receptor interaction and the latter is similar but of less defined geometry (Figure 3) (Costa and Michel 1995; Boudvillain and Pyle 1998; Keating et al. 2008). Other motifs that are essential in (the formation of) the active site are the ε-ε' and λ-λ' interactions that place the 5' splice site near the catalytic core in DV (Jacquier and Michel 1990; Boudvillain et al. 2000; de Lencastre et al. 2005; De Lencastre and Pyle 2008). The X-ray crystal structure of the *O. iheyensis* IIC intron has revealed that the ε-ε' and λ-λ' are components of a functional substructure in subdomain IC1(See Figure 3), called the z-anchor, that makes multiple contacts between domains I and V and nucleotides at the 5' end of the intron, thereby mediating the structural integrity of the core (Toor et al. 2008a). The linker between domain II and III (J2/3), which is one of the most conserved sequences among group II introns, has long been known to be important for efficient splicing activity and placed it close to the core of the ribozyme (Fedorova et al. 2003; de Lencastre et al. 2005; Pyle and Lambowitz 2006; De Lencastre and Pyle 2008). The x-ray structure shows that J2/3 and the bulge bases in DV form a triple helix with the DV catalytic triad, bringing together the catalytic essential residues of the intron (Toor et al. 2008a). DIII is internalized in the core and

participates in interactions with J2/3, ε-ε' and the bulge of DV, stabilizing the active site (Fedorova et al. 2003; Fedorova and Pyle 2005; 2008). Several studies has suggested that all reaction components are aligned in close proximity in a single active site prior to splicing and that the configuration of the core is maintained throughout the whole splicing process (de Lencastre et al. 2005; Hamill and Pyle 2006). This could suggest that there is no large conformational changes occurring between the two splicing steps- however the degree of such changes is r somewhat debatable (Chanfreau and Jacquier 1996; Costa et al. 1997; Michel et al. 2009).

The recent crystal structures has shown that most of the group II intron structure with all its interactions are essential for inducing the active catalytic (relevant) form/structure of DV that forms a metal binding platform and the active site (Toor et al. 2008a). The crosslinking studies and three-dimensional model built by Dai *et al*. for a IIA intron show the same, suggesting that the catalytic side of domain V is orientated inward (Dai et al. 2008).

Protein-assisted splicing *in vivo*

Most studies of the mechanism of group II intron folding and catalysis *in vitro* are conducted in relatively extreme reaction settings, with high salt and magnesium concentrations (>100 mM) and elevated temperature (>40°C), compared to the physiological conditions in the cell (Jarrell et al. 1988; Lehmann and Schmidt 2003). This is necessary to ensure high enough splicing reactivity. Under near-physiological conditions, intron folding is very slow and the structure is unstable (Fedorova et al. 2007; Fedorova and Zingler 2007). Therefore, most or all group II introns probably require protein factors to stabilize active structure or resolve misfolded (non-native) intermediates and allow efficient splicing *in vivo* (Lehmann and Schmidt 2003; Pyle and Lambowitz 2006; Fedorova et al. 2007). The best-characterized protein factors that participate in splicing are the proteins encoded by the mobile group II introns themselves (IEPs), usually located in domain IV. These proteins usually have four conserved domains RT (reverse transcriptase), X (maturase/splicing), D (DNA binding) and En (endonuclease), where the latter two are required only for the mobility event (see next

section) (Lambowitz and Zimmerly 2004). The RT and X domains participate in RNA binding and maturase activity of the IEP (Lehmann and Schmidt 2003; Cui et al. 2004; Lambowitz and Zimmerly 2004). Studies conducted with the *Lactococcus lactis* intron Ll.LtrB intron showed that the IEP binds specifically to the intron and exerts maturase activity by stabilizing the catalytically active RNA structure, enabling *in vitro* splicing at near-physiological conditions (Matsuura et al. 2001; Pyle and Lambowitz 2006). The binding is very strong and specific in a region including the Shine-Dalgarno sequence and start codon of the IEP ORF in domain IV, thereby providing a mechanism for autoregulating the translation of the ORF (Wank et al. 1999; Singh et al. 2002). The Ll.LtrB protein also makes contacts, though in a weaker fashion, to catalytically important regions in domains I, II and VI (Matsuura et al. 2001; Pyle and Lambowitz 2006; Dai et al. 2008). The protein binds the intron RNA as a dimer, and it is suggested that it holds domain I together, and induces and stabilizes tertiary interactions between the domains, and therefore promotes the formation of the intron's catalytically active structure (Rambo and Doudna 2004; Pyle and Lambowitz 2006; Dai et al. 2008). The IEP of the RmInt1 intron of *Sinorhizobium meliloti* has also been suggested to promote the formation of the correct EBS1/IBS1 exon-intron interaction (Molina-Sanchez et al. 2006). This study also implicates that this protein controls the balance between lariat and circle splicing pathway.

While bacterial group II introns typically encode an IEP, organellar introns often lack the IEP and require host-encoded proteins for splicing (Dai and Zimmerly 2002a; Robart and Zimmerly 2005). A number of nuclear-encoded splicing factors have been shown to be involved in group II intron splicing *in vivo*, either alone or in conjunction with other proteins, in chloroplasts of plants and mitochondria of fungi. These proteins are very diverse and generally represent proteins with additional cellular functions that have been recruited for splicing by group II introns during evolution (reviewed in (Pyle and Lambowitz 2006; Fedorova and Zingler 2007))

Mobility: Retrohoming and retrotransposition

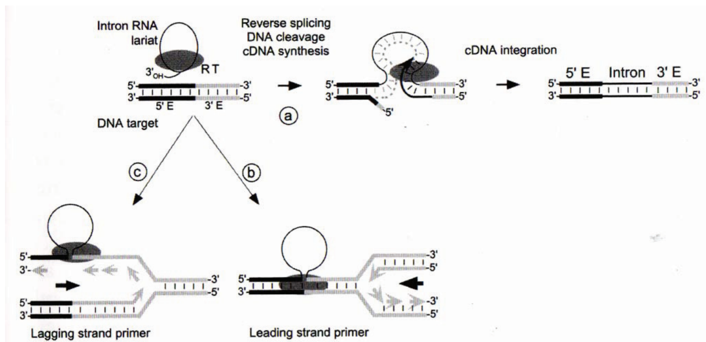


Figure 5. The figure is taken from (Pyle and Lambowitz 2006), illustrating different mobility pathways of group II introns. (a) Endonuclease (En)-dependent retrohoming pathway. While associated with the IEP, forming the RNP, the intron lariat reverse splices into one strand of a duplex DNA. The En domain of the IEP cleaves the opposite strand and uses this 3' end as a primer for reverse transcription. (b, c) Alternative En-independent pathways where the nascent strand in the DNA replication fork is used as primer for reverse transcription. In pathway (b) the intron reverse splices into double-stranded DNA before passage of the replication fork, while in pathway (c) the intron inserts at the replication fork, which is transiently single-stranded. These pathways can be used for retrohoming by introns whose IEP lacks the En activity, or for retrotransposition into ectopic sites. In (b) and (c) the black arrow indicates the direction of replication. See text for more details.

After the splicing reaction has taken place, group II introns can invade genomic DNA sites. The ribonucleoprotein (RNP) complex formed by the IEP and the lariat intron RNA during splicing medites the mobility event of the group II intron (Lambowitz and Zimmerly 2004; Pyle and Lambowitz 2006; Toro et al. 2007). The mobility starts by recognition and binding of the DNA target site by the IEP. Group II introns usually insert into cognate sequences, which extend to 30-35bp (covering positions -25 to +10 relative to the insertion site), an event called retrohoming. A different substrate specificity is seen for IIC introns which insert downstream of intrinsic transcriptional terminator stem-loop structures (Robart et al. 2007). The IEPs of ~60% the bacterial group II introns lack the En domain, and the retromobility events of bacterial group II introns can be divided into endonuclease dependent or endonuclease-independent pathways (Lambowitz and Zimmerly 2004; Robart and Zimmerly 2005; Pyle and Lambowitz 2006). The best-

characterized pathways are those of yeast mitochondrial ai1 and ai2 introns, the *L. lactis* Ll.LtrB and *S. meliloti* RmInt1 introns. Studies with ai1, ai2, and Ll.LtrB (which belong to class IIA and whose IEPs have an En domain) showed that the IEP first binds upstream of the insertion site and promotes local unwinding of the DNA (Singh and Lambowitz 2001; Aizawa et al. 2003). This enables the intron RNA to base pair with rest of the target site and then reverse splice into the top strand (See Figure 5a)(Zimmerly et al. 1995a; Yang et al. 1996). This reaction is mechanistically the reverse of the splicing reaction, and requires the analogous (same) intron/ exon base-pairings with the target sequence (EBS1/IBS1, EBS2/IBS2, and δ-δ') (Mohr et al. 2000; Singh and Lambowitz 2001). Then, the En domain of the IEP cleaves the bottom strand 9-10 bases inside the 3'exon. This generates a free 3'-OH group that the RT domain uses as a primer for reverse transcription and cDNA synthesis of the inserted intron RNA (Zimmerly et al. 1995b). Overall, this process is called **t**arget-**p**rimed **r**everse **t**ranscription (TPRT), which show many similarities to reactions performed by non-LTR retroelements (see earlier chapter on Retrotransposon) (Malik et al. 1999; Robart and Zimmerly 2005; Beauregard et al. 2008).

Many IEPs in bacteria lack the En domain, and these use the other mobility pathway that requires a primer provided by the DNA replication fork (Ichiyanagi et al. 2003; Zhong and Lambowitz 2003). RmInt1, a IIB intron whose IEP lacks the En domain uses two endonuclease-independent retrohoming pathways. A major pathway occurs by reverse-splicing of the intron RNA into ssDNA at the replication fork and the nascent lagging strand is used as a primer for reverse-transcription(See Figure 5b) (Martinez-Abarca et al. 2004; Pyle and Lambowitz 2006). A minor, replication-independent process, involves retrohoming in the opposite orientation using the nascent leading strand as primers(See Figure 5c)(Martinez-Abarca et al. 2004). For both the endonuclease and replication dependent/ -independent mobility pathway, the intron recruits several host factors to complete the integration into the new genomic location (Read et al. 2002; Beauregard et al. 2008).

While retrohoming is the predominant mobility pathway, at a much lower frequency group II introns are also able to invade noncognate (ectopic) sites through retrotransposition. The retrotransposition mobility events of Ll.LtrB in L.latctis follow

the same mechanism as the main retrohoming pathway described above for RmInt1 with insert into ssDNA (Ichiyanagi et al. 2002; Ichiyanagi et al. 2003). The target sequences usually have good match for IBS1, but not for IBS2 or the sequence recognized by the IEP. Different host organism may also influence which mobility pathways each intron use, as the Ll.LtrB intron in *E.coli* retrotransposition by inserting into dsDNA with varying priming mechanism (Coros et al. 2005). The retrotransposition, with its relaxed sequence requirements is evolutionary important because this has allowed the spread of group II introns to new and different genomic locations. The reverse splicing under the mobility event ensures that the intron can splice out of mRNA transcript and thereby minimizing the damage on the host.

Evolution: Origin and spread

Among the three major families of group II introns, elements of the IIA and IIB classes are almost twice the size (~800 nt, excluding the IEP ORF) of those from the IIC class (~450nt), which is presumed to be the most ancient (Toor et al. 2001; Toro 2003; Pyle and Lambowitz 2006). Phylogenetic analysis of the IEP of the growing number of intron sequences has shown that the group II introns can be further subdivided into nine distinct groups; (IEP classes Mitochondrial-like, Chloroplast-like 1, Chloroplast-like 2, Bacterial B, C, D, E1/E2, and F correspond to RNA structural classes A1, B1, B2, B4 (or B2-like), C, B3, B5, and novel, respectively. (Figure 6, page 54))(Michel et al. 1989; Toor et al. 2001; Zimmerly et al. 2001; Toro 2003; Simon et al. 2008). The different subclasses can be found in a mix of host organisms, although some subclasses seem to be somewhat restricted to particular bacterial phylogenetic groups. Single species can harbor introns of several subclasses clearly showing that the introns are mobile elements or that there has been a lot of horizontal transfers of introns (Dai and Zimmerly 2002b; Robart and Zimmerly 2005; Simon et al. 2008). Interestingly, comparison of the RNA secondary structures between the different subclasses indicates that the catalytic RNA has specific features that are unique to each group and strongly suggests that RNA structure has co-evolved with the sequences of the IEP (Toor et al. 2001; Simon et al. 2008). The co-evolution is suggested to be due to the strong biochemical interactions that exist between

the IEP and the catalytic RNA. This relies on the fact that both the protein and ribozyme RNA are required for the splicing reaction and the mobility event of group II introns *in vivo* (see two previous chapters)(Toor et al. 2001; Lambowitz and Zimmerly 2004). During evolution group II introns have developed different modes of target site recognition, especially IIC intron group (see chapter Defining splice boundaries above), but this differentiation is also reflected in differences in catalytic reactivity. IIA, IIB and IIC introns have been shown to exhibit quite different behaviors in autocatalytic splicing under various reaction conditions *in vitro* (Granlund et al. 2001; Lehmann and Schmidt 2003; Toor et al. 2006; Toor et al. 2008a).

The close relationship between the IEP and intron RNA structure and the presence of the IEP in all intron subclasses have also led to the hypothesis that the group II intron ancestor was essentially a retroelement (Toor et al. 2001; Dai and Zimmerly 2002b). The presence of the IEP in a similar location in domain IV suggests that the IEP was acquired once by insertion into an already catalytic ribozyme, or, alternatively, the self-splicing ability might have been developed later by a retroelement in order to prevent host damage. The "retroelement ancestor hypothesis" predicts that the various structural lineages of group II introns arose by coevolution with the IEP from an ancestor intron in bacteria, which had an RNA structure with a mix of features and a compact reverse transcriptase ORF (Toor et al. 2001; Robart and Zimmerly 2005). Bacterial introns are usually not found in important housekeeping genes, but rather they have inserted in intergenic regions or other mobile elements (Dai and Zimmerly 2002a; Ichiyanagi et al. 2003; Robart and Zimmerly 2005). These properties also support their (selfish) retroelement character, as they insert in genomic locations that minimize their impact on the host and/or will favorer their spread. Two of the group II intron subclasses are predicted to have migrated to the organelles of eukaryotes (the mitochondrial and chloroplast-like lineages), which was followed by loss of the IEP and degeneration in several RNA features, especially in plants where almost all group II introns are ORF-less (Toor et al. 2001). Organellar group II introns are inserted in many highly conserved genes essential for respiration and photosynthesis and therefore must retain efficient splicing properties. As opposed to the bacterial introns, organellar elements behave more

like splicing-only elements and rely on host-encoded splicing factors (Toor et al. 2001; Lehmann and Schmidt 2003; Robart and Zimmerly 2005).

There are several similarities in RNA structure and splicing mechanism between group II introns and the nuclear spliceosomal introns (Valadkhan 2007). An evolutionary hypothesis is that group II introns invaded the eukaryotic nucleus and then have been successively fragmented, while retaining the fundamental catalytic mechanism of self-splicing with the evolution from *cis*-acting elements to *trans*-acting small RNAs that became dependent on host protein factors for the splicing reaction (Sharp 1991). This theory is based on, and supported by the fact that fragmented bacterial and organellar group II introns can perform efficient splicing *in trans* (Knoop et al. 1997; Belhocine et al. 2008)

In addition, there is also a relationship between group II introns and the non-LTR retroelements found in higher eukaryotes (Malik et al. 1999; Robart and Zimmerly 2005). The RT segments of the two types of elements are phylogenetically and structurally related and mechanistically both are mobile through a similar TPRT mechanism (see Mobility section above) (Lambowitz and Zimmerly 2004; Beauregard et al. 2008). Together this has put up the scenario that mobile group II introns may be the ancestors of spliceosomal introns and non-LTR retroelements, and therefore may have played a substantial role in the evolution of the eukaryotic genome as predecessors of the spliceosome and retrotransposons (Pyle and Lambowitz 2006).

The *Bacillus cereus* group of bacteria

The *Bacillus cereus* group is a group of endosporeforming bacteria belonging to the Firmicutes phylum, i.e. low G+C% Gram-positive bacteria, and is a subgroup of the genus *Bacillus* (Porwal et al. 2009). Bacteria in this genus are rod-shaped in appearance and are aerobic or facultatively anaerobic.

The *Bacillus cereus* group (*B. cereus sensu lato*) includes the related species *B. cereus* (*sensu stricto*), *B. thuringiensis*, *B. weihenstephanensis*, *B. mycoides*, *B. pseudomycoides*, and *B. anthracis*, which show different phenotypical characteristics and pathologenic properties. Altogether, these Bacilli are found worldwide in diverse habitats including soil, the gut of soil-dwelling invertebrates, the plant rhizosphere, and food processing units (Jensen et al. 2003).

Bacillus cereus sensu stricto has its primary ecological niche in soil environments, and can colonize invertebrate guts as a symbiont. It is an opportunistic human pathogen involved in food-poisoning incidents and contaminations in hospitals. Food poisoning is characterized by either diarrhea and abdominal distress or nausea and vomiting, where the latter case is due to the production of an emetic toxin encoded on a large plasmid (Hoton et al. 2005; Ehling-Schulz et al. 2006). *B. cereus* is also known to cause several other types of infections (Drobniewski 1993). *Bacillus thuringiensis* has for a long time only been regarded as an insect pathogen, but can also be an opportunistic human pathogen like *B. cereus* (Rasko et al. 2005). These two species are distinguishable only by the fact that *B.thuringiensis* produces during sporulation insecticidal crystal proteins (the *cry* and *cyt* proteins), which are in most cases encoded on large plasmids (Aronson 2002). Because of this property, *B. thuringiensis* is widely used for biological control of insects. Its natural environment is thought to be the insect intestinal system, but it is found ubiquitously in soil (Rasko et al. 2005). *Bacillus anthracis* is the causative agent of the animal and human disease anthrax, and full virulence is again dependent on two large plasmids named pXO1 and pXO2 (Okinaka et al. 1999a; Okinaka et al. 1999b;

Mock and Fouet 2001). These plasmids encode a toxin complex and a protective poly-γ-D-glutamic capsule, respectively. *B. anthracis* shares the same ecological niches as *B. cereus* and *B. thuringiensis* and can be found in the soil in most of the parts of the world, but it is not clear whether it can grow vegetatively outside the host (Mock and Fouet 2001).

The last three members of the *B. cereus* group, *B. weihenstephanensis, B. mycoides* and *B. pseudomycoides,* are less studied, but have been argued to have the potential to be pathogenic because they possess enterotoxins common to the *B. cereus* group (Stenfors et al. 2002; Hendriksen et al. 2006; Stenfors Arnesen et al. 2008). *B. weihenstephanensis* is psychrotolerant, being able to grow at 4-7°C, and is a common contaminant in dairies. Specific signatures for this species are found in the sequences of the 16S rDNA and *cspA* cold shock genes (Lechner et al. 1998). *Bacillus mycoides* can also be cold-tolerant and is characterized by a rhizoidal colony shape (Nakamura and Jackson 1995). *B. pseudomycoides* is indistinguishable from *B. mycoides* by physiological and morphological characteristics, but is phylogenetically divergent from the other *B. cereus* group species and exhibits a distinct fatty acid composition (Nakamura 1998).

The species of the *B. cereus* group are despite so different characteristics very closely related at the genomic level, both in terms of gene content and synteny (Read et al. 2003; Rasko et al. 2004; Rasko et al. 2007). Their genomes are usually 5.2–5.4 Mb in size including a single circular chromosome and most strains carry one or several extra-chromosomal plasmids. In fact the *B. cereus* group is one of the most sequenced bacterial groups, with currently 36 genomes available in Genbank (including 17 *B. anthracis* strains), and more than 50 others are underway (Genomes Online database, http://www.genomesonline.org/). Genome sequence comparison of several *B. cereus* group members has revealed genetic elements on the chromosomes that appear to be unique to this group of bacteria and is not found in any other organism (Okstad et al. 2004b; Tourasse et al. 2006). Sequencing of the 16S rDNA shows over 99% similarity for the members of the group and numerous studies have shown that, with the exception of *B. anthracis* which is highly homogeneous (monomorphic), strains of the other species

are often phylogenetically intermixed and thus the species are hard to distinguish on basis of their genetic background (Ash et al. 1991; Carlson et al. 1994; Helgason et al. 2004; Priest et al. 2004; Sorokin et al. 2006). Although the species share a common genetic background, the *Bacillus cereus* group is a dynamic population. This is reflected by certain lineages of isolates that have evolved to form clonal complexes when adapting to a particular niche or host (Tourasse et al. 2006). *B. anthracis* is the clearest example, but clonal complexes of *B. thuringiensis* as well as *B. cereus* isolates of clinical origin have also been identified.

As mentioned above, several of the major phenotypic characteristics of the members of the *B. cereus* group are determined by the different plasmids they have acquired. Due to this, it has been suggested that the various members could be considered as one species based on their overall genetic similarities at the chromosomal level and represent different ecotypes, pathotypes, or plasmidovars of *B. cereus* (Helgason et al. 2000b; Rasko et al. 2005). Species definition within the *B. cereus* group has been an intense debate but in regard to the importance of these bacteria in both economic and health issues the current species nomenclature is likely to be maintained. Plasmids are vehicles for exchange of genetic information and, as in other bacteria, have played a significant adaptive role in pathogenicity and ecology of the isolates (Rasko et al. 2005; Rasko et al. 2007). In addition to factors contributing to pathogenesis or niche adaptation, plasmids also carry mobile genetic elements such as transposons and group II introns which again could induce genetic and phenotypic variation and influence the dynamic behavior of the *Bacillus cereus* group of bacteria (Tourasse et al. 2006; Tourasse and Kolstø 2008).

Bacillus cereus group “mobilome”

B. cereus group organisms have numerous mobile and/or repeated genetic components indicative of a dynamic genome. These elements include plasmids, phages, transposons, mobile introns, and diverse repeat sequences.

Plasmids

Isolates from the *B. cereus* group harbor a large range of plasmids and certain strains have more than 10 different plasmids. The diversity of these plasmids is great with sizes ranging from ~2 kb to almost 600 kb (Carlson et al. 1994; Helgason et al. 2000a; Rasko et al. 2005; Rasko et al. 2007). While some of these plasmids do encode genes responsible for critical differences in virulence and host specificity, many are cryptic.

Much attention has been put on the two plasmids found in *B. anthracis*, pXO1 and pXO2. The anthrax toxin genes encoded on pXO1 and the capsule genes on pXO2 (and their associated regulators) are located in pathogenicity islands (PAI) (Okinaka et al. 1999b; Van der Auwera and Mahillon 2005). Several *B. cereus* and *B. thuringiensis* strains have been identified to harbor plasmids sharing extensive sequence similarity and synteny with pXO1 or pXO2 (Pannucci et al. 2002; Rasko et al. 2005; Van der Auwera and Mahillon 2005; Ehling-Schulz et al. 2006; Rasko et al. 2007; Hu et al. 2009). In three cases of human pneumonia resembling anthrax the isolated strains were *B. cereus* that possessed a plasmid that show ed as much as 99.6% similarities to pXO1 and including the anthrax toxin genes, however these strains do not carry pXO2 (Hoffmaster et al. 2004; Hoffmaster et al. 2006). Other variants of the pXO1 and pXO2 plasmids have been identified in several *B. cereus* and *B. thuringiensis* strains, some of them being of clinical origin (Rasko et al. 2004; Rasko et al. 2005; Van der Auwera et al. 2005; Ehling-Schulz et al. 2006; Rasko et al. 2007). These plasmids lack the respective PAI region that is either replaced by unique DNA or missing. The plasmids are believed to harbor other genes that define the ecological or pathogenic potential of the isolate in a specific niche (Ehling-Schulz et al. 2005; Rasko et al. 2007). The pXO1 and pXO1-like plasmids share a common backbone of genes including similarity in the replication origin and genes, and

likewise for pXO2 and pXO2-like elements. Altogether, this indicates that pXO1-like and pXO2-like plasmids are two distinct groups of related plasmids that are in the genetic pool shared by the *Bacillus cereus* group of bacteria (Han et al. 2006; Rasko et al. 2007; Hu et al. 2009). The presence of different variants of these plasmids illustrate how these plasmids, like in other bacteria, are responsible for much of the genetic reservoir enabling adaption to a specific niche or to multiple niches.

How these related groups of plasmids are horizontally transferred among the *B. cereus* group of bacteria is not understood for all of them. The pXO2-like plasmids contain large conserved conjugation modules and direct evidence for transfer by conjugation in *B. cereus* and *B. thuringiensis* has been observed in both laboratory cultures and foodstuffs (Wilcks et al. 1998; Van der Auwera et al. 2005; Van der Auwera et al. 2007; Van der Auwera et al. 2008) The pXO2 plasmid in *B. anthracis* harbors however critical mutations and, like pXO1, is not directly self-transmissible, but both plasmids can be mobilized by other conjugative plasmids such as pXO14 from *B. thuringiensis* (Reddy et al. 1987). pXO1 and pXO1-like plasmids encode genes that that show weak similarity to conjugation genes, but these have not been experimentally characterized (Grynberg and Godzik 2004; Rasko et al. 2005). Transformation by natural competence has been little addressed in this group of bacteria (Heierson et al. 1987). Sequence comparison and analysis has though shown that many of the *B. cereus* group members have gene sets homologous to those involved in competence in *Bacillus subtilis* (Read et al. 2003) (N.J. Tourasse, unpublished data).

Bacteriophages

Bacteriophages, like plasmids, are important source of genetic drift in bacteria by facilitating horizontal gene transfer and promote genomic rearrangements which may contribute to the emergence of bacterial pathogens (Brussow et al. 2004). Some phages have been shown to infect a various members of the Bacillus cereus group (Koretskaia et al. 1989; Koehler 2002; Schuch et al. 2002).

Several integrated prophages have been found in the various genomes that have been sequenced in the *B. cereus* group of bacteria (Rasko et al. 2005). Four prophages are

present in the genome of *B. anthracis* and are conserved in ~300 isolates sequenced (Sozhamannan et al. 2006). All four prophages can excise from the chromosome at low frequencies, but are apparently defective in phage production. *B. cereus* E33L shares, at least partly, one of these *B. anthracis* prophages, and quite remarkably, is predicted to contain up to 27 chromosomally-encoded prophages plus several more on plasmids (Rasko et al. 2005). The *B. cereus* type strain, ATCC 14579, and *B. cereus* G9241 both contain non-integrated phages, designated pBClin15 and pBClin29, respectively (Ivanova et al. 2003; Hoffmaster et al. 2004). The pBClin15 show close relationship with the other *B. thuringiensis* phages (Stromsten et al. 2003; Verheust et al. 2003; Verheust et al. 2005) The pBClin29 appears more cryptic as a phage, as it both encodes phage-like protein and a pXO2 like plasmid replicon (Hoffmaster et al. 2004).

DNA transposons

As mentioned earlier, transposable elements are contributors to the genome dynamics of all life forms. In the *B. cereus* group of bacteria several types of transposable elements have been discovered (Mahillon et al. 1994; Leonard et al. 1997). One of these is the insertion sequence (IS) element IS231 for which several copies have been found in all of *B. cereus* group members (Mahillon et al. 1985). The IS231 transposase has been shown to trans-mobilize related transposase-less cassettes (miniature insertion cassettes, MICs), where these cassettes can contain genes coding for putative antibiotic resistance or regulatory factors (De Palmenaer et al. 2004). Furthermore, a putative cointegrative transposon, named Tn*XO1*, is present on the pXO1 plasmid of *B.* anthracis (Van der Auwera and Mahillon 2005). This element, that bears a transposase and a site-specific recombinase, appears also to contain a germination operon. Remarkably, this putative transposon is located inside the PAI region containing several of the *B. anthracis* virulence factors. This island is bordered by inverted IS*1627* elements that are presumably capable of mobilizing the entire stretch of DNA between them, which is suggested by the fact that in some isolates of *B. anthracis* the PAI is inverted (Okinaka et al. 1999b; Read et al. 2002). Other examples of transposons in the *B. cereus* group of bacteria are the Tn*916*-like transposon that carries a tetracycline resistance gene or the

mercury resistance transposon, Tn*MERII*, found among many *Bacillus* strains in the environment (Bogdanova et al. 1998; Agerso et al. 2002). Numerous insecticidal *cry* genes in *B. thuringiensis* plasmids are associated with IS elements and transposons, and this is thought to facilitate the spread, exchange, and creation of toxin diversity among *B. thuringiensis* isolates (Mahillon et al. 1994; Berry et al. 2002).

Interspersed DNA repeat sequences *Bcr*1-18

Sequence analysis of the intergenic regions of the growing number of genomes from the *B. cereus* group has revealed that there are as many as 18 groups of repeated sequence elements, named *bcr1* to *bcr18* (Tourasse et al. 2006). These are between 100 and 400 bp in length and are found either at conserved locations or in unique loci in the different strains. About half of the repeats appear to be unique to these organisms. The first repeat discovered and the most studied is the ~155 bp *bcr*1 element that is found on the chromosome of *B. cereus* group organisms (Okstad et al. 1999a; Okstad et al. 1999b). *bcr*1 exhibits several characteristics of a mobile element, as it is found in variable copy numbers and often in different genomic locations between strains (Okstad et al. 2004a; Klevan et al. 2007). Furthermore, it is flanked by a pentameric TTTAT direct repeat motif, which is duplicated at the insertion site, a feature of certain transposable elements. Hybridization experiments demonstrated that *bcr*1 is present on transcripts of the size of a single repeat, but also on larger mRNA transcripts, and it has been suggested that the repeat could have some stabilizing effect on the mRNA or act as a small regulatory RNA (Okstad et al. 2004a; Klevan et al. 2007).

Group I intron

The group I intron inserted in the *rec*A gene of *B. anthracis*, encoding a recombinase essential for DNA recombination and repair, was the first group I intron discovered that is inserted in a standalone bacterial chromosomal protein-coding gene and is not contained within a prophage (Ko et al. 2002). This intron was later identified in several *B. cereus* group bacteria (Tourasse et al. 2005; Tourasse et al. 2006; Tourasse and

Kolstø 2008). Furthermore, group I introns are located in other essential genes that do not have a phage associated copy. Many bacteria of the *B. cereus* group have introns inserted in the genes coding for the α and β subunits of ribonucleotide reductase (*nrd*E and *nrd*F), an enzyme critical for DNA synthesis (Tourasse et al. 2005; Tourasse et al. 2006; Nord and Sjöberg 2008). In fact, the *nrd*E gene contains four different group I introns in four insertion sites that are occupied in different subsets of isolates (Nord et al. 2007; Tourasse and Kolstø 2008). The *recA* and *nrdEF* introns (which belong to various structural classes) are predicted to have come from phage introns, as they share sequence homologies with introns present in similar host genes in various bacteriophages. In addition, several other group I introns are also found inserted in different prophage genes in *B. cereus* group organisms (Tourasse and Kolstø 2008). Altogether, this indicates that, like in other bacteria, phages are important vectors for disseminating group I introns in the *B. cereus* group.

IStrons

A 1.9-kb composite element, consisting of a group I intron and an IS element, similar to that found in *Clostridium difficile* and termed IStron has been identified in most of the genomes of the *B. cereus* group (Braun et al. 2000; Hasselmayer et al. 2004; Tourasse et al. 2006). Copies of this element, named Bc*ISt1*, are 94% identical and often found in unique chromosomal loci and inserted after a consensus pentanucleotide AAGGG. Such a distribution, with insertion in unique rather than conserved loci suggests that these elements are mobile and follow a pattern more similar to that of IS element mobility than of group I introns (Tourasse et al. 2006). Recent work demonstrates that the IS element does excise alone from the IStron sequence (Stabell, Tourasse et al. in preparation). Furthermore, this study also suggests that the IS element's ORF can activate larger parts of the IStron. Though not proven directly in these assays, the results indicate that the whole IStron element may be mobilized at a low frequency and under certain conditions.

Group II introns

The *B. cereus* group of bacteria harbors a wide range of group II introns, from different structural classes and evolutionary origins. A recent survey has shown that among 77 group II introns identified in 29 *B. cereus* group genomes, 68 belong to the IIB class, eight were from the IIA class and there was a single one from the IIC class (Tourasse and Kolstø 2008). The introns, as in other bacteria, appear to use plasmids as vectors for horizontal spread between isolates. This can be seen from the fact that group II introns are common on large plasmids (>15 kb), including pXO1-like and pXO2-like plasmids, and identical intron copies are found on both chromosome and plasmids in the same and in different isolates(Tourasse et al. 2006; Tourasse and Kolstø 2008). In contrast to other bacteria, however, the group II introns in the *B. cereus* group are not particularly associated with IS elements or transposons (Tourasse and Kolstø 2008). One of the most widespread introns in the *B. cereus* group is *B.c.*I1, which is always found in the 3' untranslated region of various genes in the genome of several isolates (Tourasse et al. 2006; Tourasse and Kolstø 2008). This intron appears to have disseminated by site-specific retrohoming as all insertion loci are well conserved in positions -22 to +11 around the homing site (Tourasse and Kolstø 2008). The IEP of *B.c.*I1 lacks the En domain and most likely uses nascent strands at the replication fork as primers for cDNA synthesis (see 'Group II intron' chapter above).

Besides the work reported in this thesis, only two studies have conducted functional analysis of group II introns in bacteria from the *B. cereus* group (Robart et al. 2004; Van der Auwera and Mahillon 2008). Firstly, Robart and co-workers investigated the two introns found in the pXO1 plasmid of *B. anthracis*, both *in vivo* and *in vitro*. Both introns, *B.a.*I1 and *B.a.*I2, spliced *in vivo*, but only *B.a.*I2 showed efficient splicing *in vitro* (Robart et al. 2004). Further characterization of *B.a.*I2 revealed that this intron uses a 3' splice site shifted one nucleotide downstream of the expected position, which disrupts its host gene. The study also revealed a minor splicing event occurring four

nucleotides downstream of this site both *in vivo* and *in vitro*, which restores the expression of the host gene. This flexibility in 3' splice site selection was shown to be partly the result of structural irregularities in the base of the domain V stem and was suggested to be an adaptation to allow (regulate) expression of the host gene (Robart et al. 2004).

Secondly, two introns on the pXO2-like pAW63 plasmid of *B. thuringiensis kurstaki* HD73, located in genes believed to be involved in conjugation, were analyzed for splicing *in vivo* (Van der Auwera et al. 2005; Van der Auwera and Mahillon 2008). Surprisingly, these experiments revealed that the intron named *B.th.*I2, which does lack the IEP, was able to splice *in vivo*, while the *B.th.*I1 intron, which has the IEP, did not. It was suggested that the *B.th.*I2 splicing could be trans-activated by the IEP of *B.th.*I1, which was shown to be transcribed due to the sequence similarities between the extremities of the two introns. However, this and the above results need to be confirmed and it is also possible that an IEP-containing intron more closely related to *B.th.*I2 is present elsewhere in the genome.

Aims of study

The main aim of this theises was from the beging to identify and functional characterize group II introns in the Bacillus cereus group of bacteria. As a part of this characterization we wanted to classify structurally and phylogenetic those group II introns we identified in this bacteria group. These results were then to be compared with other studies of group II introns in bacteria.

Under these studies an unprecedented group II introns with a 56-nucleotide long 3' extension was discovered in B.cereus ATCC 10987. Much of the further aim became therefore to study different aspects of this special variant of group II introns named *B.c.*I4.

First, we wanted to show that that the extension was part of the intron that splice out of the mRNA and see what effects it had one the introns catalytic RNA splicing.
We further wanted to identify if any specific part or residue(s) of the extension was important for splicing, and is so, could there be interaction partners in other domains of the intron.

Our other goal was so to search for introns that harboured a similar 3' extension. If so, we wanted to classify these structurally and phylogenetic, as well as functional characterize them. By comparing sequence and functional analysis with *B.c.*I4 and see if this could reveal the nature and origin of the special 3' extension. In this context we also wanted to screen for more copies in similar positions in different bacteria, as this together with bacteria phylogeny could sheed more light on the origin of these unusual group II introns.

Summary of papers

Paper 1

In this study we define and functionally characterize the group II introns present in the genomes of two strains of *Bacillus cereus* ATCC 14579 and ATCC 10987. Four of the seven group introns identified in strain ATCC10987 were inserted on the plasmid, whereas in strain ATCC 14579 one intron is harboured on its chromosome. Sequence comparison and secondary structural predictions of the intron RNA revealed that they either fold into B1 class or B2-like structural class. Both B.c.I1 and B.c.I2 can be found in almost identical copies in different genomic locations and implicates therefore that these introns are mobile.

For functional analysis by RT-PCR on total RNA from both strains verified that all but *B.c.*I1a in *B.cereus* ATCC14579 were functional for splicing *in vivo* under the conditions tested. Sequencing of splice products revealed splice boundaries as expected besides for *B.c.*I4, which spliced 56 nucleotides downstream of the predicted 3' splice site. Another striking observation is the *B.c.*I2a that is located on the antisense strand of a predicted cell surface protein. PCR screen of 92 strains revealed that those that harbored the host gene appeared to correlate to strains of clinical origin.

Paper 2

Here in this study we examine the splicing of the extraordinary group II intron B.c.I4 intron in more detail using *in vivo* and *in vitro* experiments. We show that the extra 56-bp 30 segment is an integral part of the intron RNA molecule downstream of domain VI, while branching is still maintained at the expected site. This represents therefore a unique arrangement and shows that 3' splice site selection can be more flexible than ever seen before.

The results in vitro demonstrate that the wild-type intron splices somewhat more efficient than the construct deleted of the extension. Thus the intron must have adapted to

splice with the extension, maybe through some conformational changes as the B.c.I4 appear to be as a usual intron. Secondary RNA structure predictions suggest that the 56 bp segment folds into two stable stem-loop structures.

Paper 3

Here, we report of four new group II introns, *B.th*.I5, *B.th*.I6(a and b), and *B.th*.I7 from *Bacillus thuringiensis* 4D1 that harbor a 3' extension similar to that of *B.c*.I4. These splice *in vivo* at analogous positions after the extensions that form the two conserved stem loop secondary structures. Sequence analysis show that they are only 47-61% identical to each other , and surprisingly they do not form a single evolutionary lineage in the Bacterial B class intron they all belong to. *In vitro* mutagenesis showed that the larger of the two stems is important for an efficient second step splicing with the extension, and this unusual mode 3' splice site positioning may suggest that these introns could form a new functional class.

Paper 4

Group II introns that harbor a conserved 3'extension have been identified in the Bacillus cereus group of bacteria. This element forms two stemloop structures where one of them, S2, has been shown to be important for positioning 3' exon for efficient second splicing step with the extension. Here in, we show that the whole extension may also have an effect on the balance of first step of splicing by hydrolyis and transesterification. Most remarkable though, *B.th*.I6 analyzed here appears to not be able to have an efficient second step without the extension as opposed to *B.c*.I4, which do splice equally efficient without the extension. These two belong to different evolutionary branches and may have undergone changes that lead to the different splicing properties.

Paper 5

Here we report the presence of additional atypical group II introns in the *B. cereus* group of bacteria. Several are identical or closely related copies of the previously identified *B. thuringiensis* intron *B.th.*I6. The study gives a detailed sequence and phylogenetic analysis of the *B.th.*I6 intron copies and the plasmid borne host gene. We conduct a detailed sequence and phylogenetic analysis of these intron copies and their host genes and strains, lending evidence that the unusual group II introns are mobile with their 3' extension.

Discussion

General features of group II introns in the *Bacillus cereus* group

Group II introns are widespread in the *B. cereus* group of bacteria

The publications of the genome sequences of the *B. cereus* strains ATCC 14579 and ATCC 10987 (Ivanova et al. 2003; Rasko et al. 2004) were instrumental in the discovery of the group II introns in these strains reported in paper **1**. Genome comparison and detailed sequence analysis including the identification of the IEP and boundaries of introns and host genes revealed 1 and 7 group II introns in *B. cereus* ATCC 14579 and ATCC 10987, respectively. Together with RNA secondary structure prediction, this enabled intron characterization and structural classification. The experimental work in paper **1** confirmed the predicted intron boundaries and demonstrated that these introns spliced *in vivo*. Besides the identification and functional studies of the two group II introns in the *B. anthracis* pXO1 plasmid (Robart et al. 2004) no other work had been done on group II introns from the *B. cereus* group of bacteria prior to paper **1**. In fact, the analysis done on the 7 *B. cereus* ATCC 10987 introns doubled the number of bacterial group II introns that had been functionally tested since the discovery of these elements in bacteria 15 years ago (Ferat and Michel 1993), and only a handful of introns have been tested since then. Later, in the whole genome or plasmid sequences of additional *B. cereus* group strains a varying number of introns were identified (Tourasse and Kolstø 2008; Van der Auwera and Mahillon 2008), ranging from zero to as much as 13 introns per strain. The genome analyses as well as screening by PCR for the host genes of specific introns (*B.c.*I2a, *B.c.*I4 and *B.th.*I6a) in papers **1**, **2**, and **5** revealed that group II introns are relatively widespread in the *B. cereus* group, and that virtually identical copies of introns are found in multiple strains of diverse origins and geographical locations. Although the observed distribution of introns can be interpreted as a result of both vertical inheritance and horizontal transfers, in particular via plasmid vectors, the presence of identical introns in spatially and temporally unrelated strains is puzzling. Possible explanations may be that the different strains have acquired introns from similar

donors, or that the donors and/or recipient strains have been disseminated by human and animal transport.

With respect to the different number of group II introns found in *B. cereus* group strains, it should be pointed out that strains carrying the highest numbers of group II introns, *B. cereus* ATCC 10987 and the emetic *B. cereus* strains AH187 (F4810/72) and H3081.97, appear to have generally higher numbers of mobile elements such as group I introns, transposons, IS elements, IStrons and/or other putative mobile repeated elements compared to other known genomes of the *B. cereus* group of bacteria (Okstad et al. 2004b; Rasko et al. 2004; Tourasse et al. 2006; Kolsto et al. 2009). The relative large number of such elements could imply that these strains are facultative intracellular strains, or are in a transition stage toward becoming host-restricted, and therefore have been subject to an increase in mobile element frequency (Lynch and Conery 2003; Moran and Plague 2004; Ochman and Davalos 2006). Little is known about the lifestyles of these three strains besides the fact that *B. cereus* ATCC 10987 was found in spoiled cheese almost 80 years ago (Rasko et al. 2004) and that *B. cereus* AH187(F4810/72) and H3081.97 produce an emetic food-poisoning toxin and have been isolated from vomit and food, respectively. Other explanations could be that each of these strains are found in unique environments that have an influence on mobility of genetic elements or it could be differences in content of bacterial-cell factors available to be recruited for the mobility events (Coros et al. 2005; Smith et al. 2005; Coros et al. 2009).

Different structural classes of group II introns in the *B. cereus* group

The features of the RNA secondary structure of the eight group II introns identified in the two *B. cereus* strains analyzed in paper **1** revealed that these introns belonged to the same two structural classes as the introns identified earlier in *B. anthracis* (Robart et al. 2004), i.e., the B2-like and B1 structural classes within the main group IIB intron class (Toro 2003; Toro et al. 2007; Simon et al. 2008). These structural classes correlate with the IEP phylogenetic classes “Bacterial B” and “Chloroplast-like class 1” (CL1), respectively. The four additional introns discovered in papers **3** and **5** also belonged to the B2-like

class, which appears to be the dominant class of group II introns found in the *B. cereus* group of bacteria (Tourasse and Kolstø 2008; Van der Auwera and Mahillon 2008). Striking is the fact that all but ten of 174 copies of the 80 different known introns that belong to the ”Bacterial B” group are found in Firmicutes, essentially in Bacilli and Clostridia (Paper **3,** (Simon et al. 2008) and N.J. Tourasse, unpublished results). Unless this is the result of sampling bias, as the *B. cereus* group and Clostridia are among the most sequenced Gram-positive bacteria, one could think that this class of group II introns has adapted to the molecular and/or physiological conditions in Firmicutes. There may also be unique properties in their RNA structure or IEP that give them an advantage in splicing and/or easier propagation in these genomes (Coros et al. 2005; Coros et al. 2008; Simon et al. 2008; Coros et al. 2009). It could also be that there are less lateral transfers between Firmicutes and distantly related bacteria.

In papers **3** and **5** we showed that the Bacterial B IEP class can be divided into two subgroups (α and β) and that these coincide with subtle changes in RNA secondary structure. In paper **5** point out that many of the unique RNA features of the β subgroup are shared with introns belonging to the Bacterial A phylogenetic IEP class. The RNA secondary structure of introns from the β subgroup and the Bacterial A class is hybrid, showing features common to group IIA and IIB introns (Toor et al. 2001; Dai and Zimmerly 2002b). In contrast, the structure of the α subgroup is IIB. In addition, introns of the mitochondrial-like (ML) IEP class, which are IIA, emerge in-between the Bacterial A and B groups (see Figure 6 below and (Simon et al. 2008)). One parsimonious way to interpret this complex phylogenetic and structural pattern would be that all these classes evolved from a common ancestral intron group that had a IIA/IIB hybrid structure. In such a scenario, introns on the ML line became true IIA introns, while others evolved to form the Bacterial B class. While elements from the β subgroup retained most of ancestral hybrid features, those evolving into the α subgroup became true IIB introns.

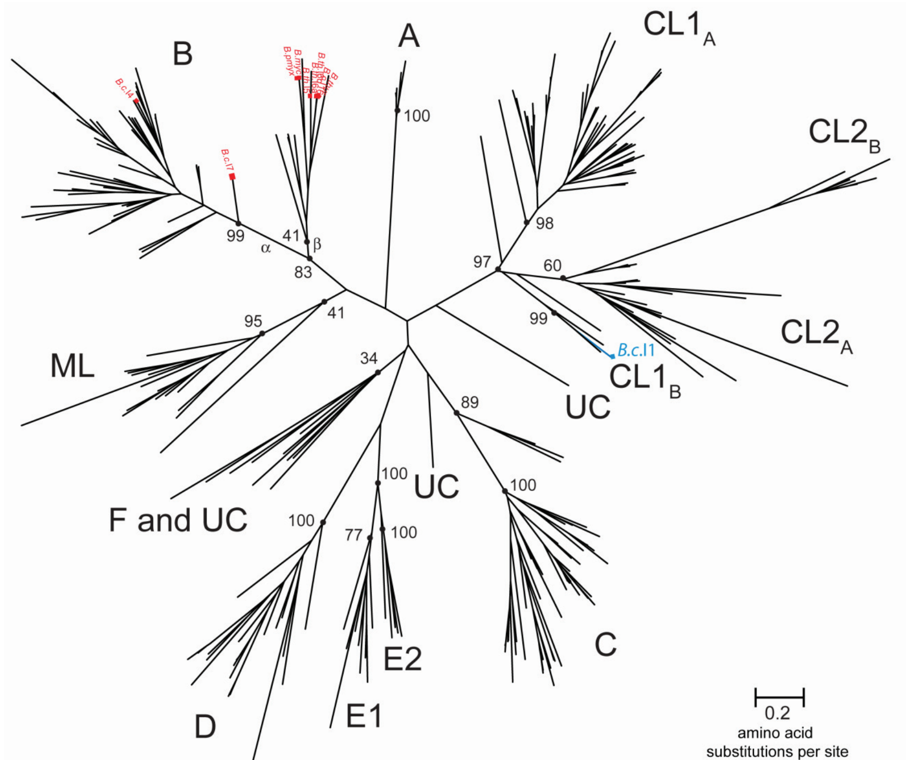


Figure 6. Unrooted phylogenetic tree of 277 bacterial group II intron-encoded proteins (N.J. Tourasse, unpublished). As in paper 3, Figure 3A, and in (Simon et al. 2008), the tree was reconstructed using the maximum likelihood method (RAxML program) and was based on the RT domains of the proteins. The major intron classes are named as in (Simon et al. 2008): A-F, ML (mitochondrial-like), CL (chloroplast-like), and UC (unclassified). The unusual *B. cereus* and *B. thuringiensis* group II introns carrying a 3' extension are indicated by red names and red squares. These are all found in the bacterial B class, that is shown in more detail in Figure 10. In light blue is indicated the *B. cereus* intron *B.c.*I1 that is found in the $CL1_B$ class. Numbers next to branch nodes indicate bootstrap support percentage values (out of 1000 replicates).

Localization and impact on host

Introns in non-coding regions

Although introns from the "Bacterial B" class are the most common in the *B. cereus* group, the B1 class intron *B.c.*I1 appears to be among the most widespread introns

in this group of bacteria (Tourasse et al. 2006; Tourasse and Kolstø 2008). In paper **1** three copies of this intron were found on three different genomic entities (replicons) in the *B. cereus* strains ATCC 10987 and ATCC 14579. All these elements are inserted into genomic locations predicted to be non-coding intergenic regions. One to three copies of the *B.c.*I1 intron were subsequently found in four additional strains and are all located in the 3' untranslated region (3' UTR) of predicted mRNA transcripts of unrelated genes (Tourasse and Kolstø 2008). Whether this distribution is specific only to this particular intron, given some structural RNA or IEP features that have developed to specifically recognize 3' UTRs in the genome, is not known. Among the introns listed in the Mobile group II intron database (http://www.fp.ucalgary.ca/group2introns/), not all B1 class introns target 3' UTRs, and it is therefore not a general feature for this structural class (Dai et al. 2003). The B1 class intron A*vi.groEL* of *Azotobacter vinelandii* on the other hand belongs to a peculiar clade of bacterial group II introns with a 5′-subterminal insertion, whose members are preferentially associated with signals for translation termination and initiation (Adamidi et al. 2003; Ferat et al. 2003; Michel et al. 2007). What makes the target so specifically associated with genetic signals is unknown, but it is suggested to affect the protein expression (Michel et al. 2007). Similarly, most introns of the IIC class specifically recognize transcriptional signals and the mobility events are guided and associated with rho-independent terminators (Dai and Zimmerly 2002a; Toro 2003; Toor et al. 2006; Robart et al. 2007). In this context of development of specialized lineages, it is interesting to see from the phylogenetic tree in paper **3** and Figure 6 that the *B.c.*I1 intron is part of a single lineage branching out at the bottom of the CL1 class, within the $CL1_B$ subclass (Michel et al. 2007; Simon et al. 2008). This could suggest that *B.c.*I1 has developed features reflecting adaptation to 3' UTR targets. The preference of the *B.c.*I1 group II intron for insertion in 3' UTRs might also not be due to specific RNA or IEP features; mobility events occurring in coding loci may be selected against, as for example a low splicing efficiency could have a negative effect on host gene or protein expression. The strategy to insert in non-coding regions is quite common for bacterial group II introns, and fits with the picture of group II introns in bacteria being true selfish, mobile retrolements rather than "splicing-only" entities (Dai and Zimmerly 2002a; Lambowitz and Zimmerly 2004; Robart and Zimmerly 2005). Therefore, *B.c.*I1

represents yet another variant of the strategy used by group II introns to spread in non-coding regions.

Introns in coding regions

Previously group II introns were believed to be relatively rare in bacteria, and if present they would usually be found associated with other mobile elements, in intergenic regions, or in non-conserved genes (Dai and Zimmerly 2002a). However, an increasing amount of group II introns described are inserted in single-copy, important bacterial genes (Michel et al. 2007). Of the 7 *B. cereus* group intragenic introns investigated in papers **1**, **3** and **5** all but a single one (*B.c.*I2a) are located in an ORF predicted to be involved in aspects of DNA replication, recombination, repair, or the metabolism of nucleotides. Indeed, a clear majority of the intragenic group II introns identified in 29 *B. cereus* group genomes recently surveyed are located in these types of genes ((Tourasse and Kolstø 2008) and N.J. Tourasse, personal communication). A similar host gene preference has been observed for other bacterial group II introns in other organisms (Liu et al. 2003; Chee and Takami 2005; Michel et al. 2007). This host gene bias is reminiscent of the distribution of group II introns in organellar genomes and is also seen for group I introns and inteins (Dai and Zimmerly 2002a; Robart and Zimmerly 2005; Tourasse and Kolstø 2008). One suggestion for this bias is that the intron will be expressed together with the proteins required for repairing DNA breaks under mobility events and therefore increase the probability of successful retromobility (Edgell et al. 2000; Michel et al. 2007; Coros et al. 2009). Another hypothesis suggests that the presence of introns in functionally related genes provides an opportunity for coordinated regulation of expression at the level of splicing (Michel et al. 2007).

The κ region (ID$^{(ii)}$)

The κ–κ’ tertiary interaction between domains I and V is an important part of the network of interactions that form the structural unit that supports splicing (Toor et al.

2009). In subdomain ID$^{(ii)}$ of many introns there is a three-way junction where one finds the κ tetraloop, which interacts with the κ' receptor basepairs in domain V, as well as another possible tetraloop adjacent to κ (shown in green in Figure 7A) (Boudvillain and Pyle 1998). Another study suggested that this loop lies close to the bulged branchpoint adenosine in domain VI in the intron 3D structure because these regions formed crosslinks (de Lencastre et al. 2005; Hamill and Pyle 2006). However, the two recent crystal structures of the *O. iheyensis* group IIC intron indicate that the κ tetraloop and the additional tetraloop are on the side of the domain V helix opposite to that from which domain VI with the bulged A resides when attacking the 5' junction (Toor et al. 2008a; Toor et al. 2008b; Michel et al. 2009).

In paper **3** we searched for tetraloops that could be candidates for interacting with the internal loop of stem S2 in the 3' extension, as this loop shows characteristics of the 11-nt tetraloop receptor motif (Costa and Michel 1995). In *B.c.*I4 a putative GAAA tetraloop next to the κ loop can be predicted if the guanosine in position 311 forms a G:C pair with the cytosine in position 334 (Figure 7A B.c.I4v2). This additional tetraloop would have a form somewhat similar to κ and would be followed by a short helix with a terminal loop (ID$^{(ii)}$a). This predicted structure is conserved in many B2-like introns, including *B.th.*I7, while in others it may take a different form, if present, as exemplified by *B.th.*I5 and *B.th* I6 (that do not have the short stem following the two GAAA loops) (Figure 7A).

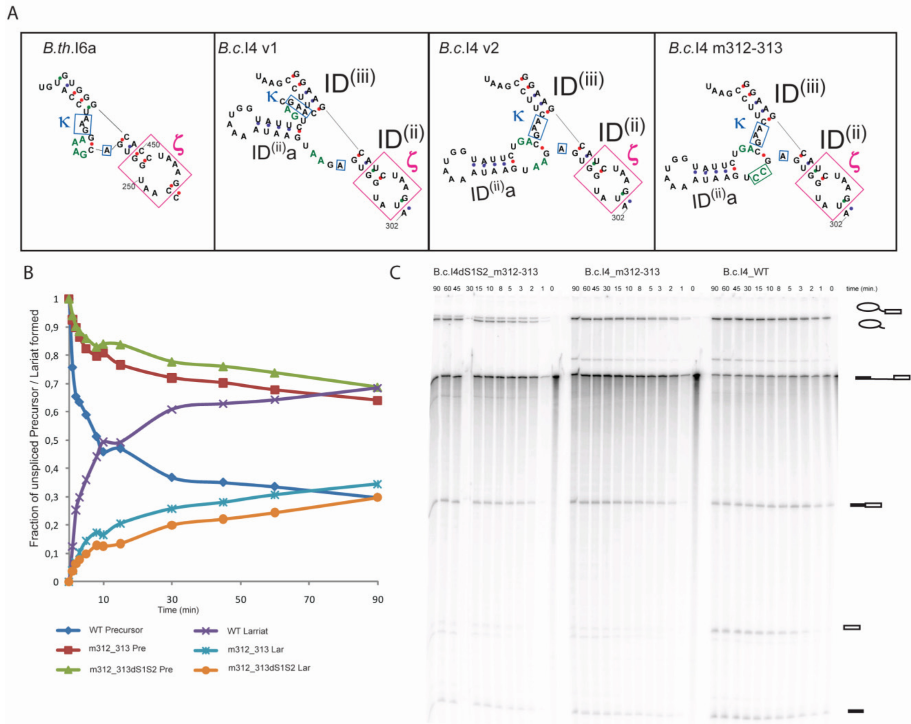


Figure 7. Secondary structure predictions and effects of mutations of nucleotides in possible tetraloop structure close to the κ tetraloop in subdomain ID(ii). (A) Predicted secondary structure of subdomain ID(ii) for *B.th.*I6a and *B.c.*I4 wildtype (WT) and mutant intron constructs. Two possible secondary structures are shown for *B.c.*I4. Structure v1 refers to that shown in paper 2. Nucleotides of the κ tetraloop, which interacts with the κ' receptor basepairs in domain V, are boxed in blue. Nucleotides participating in a possible additional candidate tetraloop are colored in green, and mutated residues are boxed in green. (B and C) Time-course analysis of *in vitro* self-splicing of *B.c.*I4 WT and mutants *B.c.*I4m312-313 and *B.c.*I4dS1S2m312-313. Splicing was performed in 40 mM MOPS, pH=7.5, 100 mM $MgCl_2$ and 500 mM $(NH_4)_2SO_4$ at 47°C. In (B) the relative fractions of unspliced precursor RNA and released lariat intron were computed from the intensities of the radioactive bands using a phosphorimager. In (C) the 7 M urea 4% polyacrylamide gel with samples used for quantification. Schematic drawings of the different reaction products are shown on the right.

Mutations conducted on the analogous GAAA loop in the yeast mitochondrial ai5γ group II intron revealed no effect on the branching reaction studied by a *trans*-splicing assay (Boudvillain and Pyle 1998). In contrast, when the last two adenosines of the potential GAAA loop next to κ were substituted into cytosines in *B.c.*I4, in positions 312-313(mutants *B.c.*I4m312-313 and *B.c.*I4dS1S2m312-313), there was a clear negative effect on the efficiency of splicing (see Figure 7B). This effect appeared to be mainly on

the first step of splicing with a ~50 % reduction, and the results were similar with or without the presence of the 3’ extension. The reduced splicing efficiency observed for the mutations in this potential GAAA loop can suggest that this loop is functionally important, although it may also result from an indirect effect due to that the mutations may affect the folding of the critical κ loop region.

In any case, the most striking observation may be that the relative fraction of released 3’ exon is clearly reduced for the tetraloop mutants compared to the wild type *B.c.*I4 intron, as judged from the gel picture (Figure 7C). Free 3’ exon is only known to come from the SER reaction, where the released lariat catalyzes the hydrolysis of the spliced exons (see ‘Other reactions’ section above; (Jarrell et al. 1988; Podar et al. 1995; Lehmann and Schmidt 2003; Michel et al. 2009). Interestingly, the SER reaction has mainly been observed for the structural class IIB introns and can rarely be observed for IIA class introns (Lehmann and Schmidt 2003). This reaction actually represents a true ribozyme activity of a group II intron since the intron RNA remains unaltered when hydrolyzing the ligated exons (Michel et al. 2009). Comparing the consensus secondary structures for the different phylogenetic groups in the Mobile group II intron database reveals that the additional GNRA tetraloop next to the κ loop is conserved in many IIB class introns, but not all, and none of the IIA class introns (Dai et al. 2003). The close proximity of this possible GNRA loop to the κ loop and possible domain V and VI could suggest some involvement in the ribozyme catalytic activity. The wild speculation would be that it is for example involved in some switch that can activate the ribozyme structure for the SER reaction. Anyway, possible side effects of the mutations on other residues will have to be ruled out before any conclusion can be drawn from these interesting, though preliminary results.

Group II introns with an unusual 3' extension

The 3' extension may represent a new functional domain

As mentioned in the Introduction group II introns are divided into classes based on phylogenetic analysis of their IEP and predicted RNA secondary structure (Michel et al. 1989; Toor et al. 2001; Simon et al. 2008). The numbers of classes have been growing with the increasing number of introns identified, revealing the great diversity of these ribozymes. The major structural classes IIA, IIB and IIC, and their subgroups have in many cases revealed different splicing properties *in vitro* that could be linked to their structural differences (Granlund et al. 2001; Lehmann and Schmidt 2003; Toor et al. 2006; Michel et al. 2007). These structural differences are often due to addition or deletion of particular stem-loop substructures within the different RNA domains (Toor et al. 2001; Toor et al. 2006; Michel et al. 2007).

While the secondary structure of typical group II introns is made up of six domains, the major contribution of the work presented in this Thesis was the discovery and characterization of unusual introns carrying an extension of 53-56 bases at the 3' end. The unusual 3' extension of the *B.c.*I4 intron in *B. cereus* ATCC 10987, identified in paper **1** and further functionally characterized in paper **2**, first represented an extraordinary example of the group II introns' ability to acquire structural elements. This extension is predicted to form two stem-loop structures, S1 and S2. The discovery in papers **3** and **5** of additional introns with the same conserved 3' stem-loop structures would have suggested that these introns were part of a specialized lineage. It was therefore a big surprise when phylogenetic analysis demonstrated that these unusual introns did not represent a single lineage of group II introns, but are distributed among the α and β subgroups in the "Bacterial B" class of group IIB introns. However, these introns exhibit a unique structural arrangement with the addition of an extra element at the 3' end. Other structural classes do often only have variants of the six structural domains and therefore, as suggested in a recent review (Fedorova and Zingler 2007), the

two conserved stem-loop structures constituting the extension could be referred to as a domain VII. Furthermore, the functional evidence presented in papers **3** and **4** shows that the extension can influence different aspects of splicing, both in the first and second steps, and implicate that the extension could represent a new domain, a structural tool that can confer different properties to different group II intron ribozymes, as exemplified by the *B.c.*I4 and *B.th.*I6a introns.

Does the 3' extension fit into the active ribozyme, and if yes, how?

The different results obtained when deleting the 3' extension from unusual introns in papers **2** and **4** imply that introns must have adapted differently to function with the extension. In the case of *B.c.*I4 the extension influences the balance between hydrolysis and branching in the first splicing step, whereas in *B.th.*I6a the extension is required for an efficient second splicing step. On the other hand, the results from paper **3** point to that the S2 stem within this conserved 3' intron segment may have a similar role and function for all the unusual introns, as it is important for efficient second-step splicing in several unusual introns. This would imply that there is a common type of interaction to position the 3' splice site in an efficient manner when the extension is present. Whether this involves similar sites or whether it is specific for each individual intron is an open question.

Ensuring an efficient second splicing step with the extension

The mutational studies in paper **3** did not reveal any alternative 3' splice site when mutating various sites in the extension, thus implying that the extension does not have any impact on the fidelity of 3' splice site selection. In paper **2,** mutations in the 3' splice site were conducted, disrupting the γ –γ' and EBS3-IBS3 interactions. These studies showed that a 3' splice site located 2 nt downstream of the wildtype site could be used, but still the wildtype site was used by 55% of the intron molecules. These results fit with

those observed for other (typical) class IIB introns, for which these two interactions have been shown to play a minor role in recognition of the 3’ splice site, supporting correct recognition only when the 3’ splice site is in position in the catalytic center (Jacquier and Michel 1990; Schmidt et al. 1993; Costa et al. 2000; Lehmann and Schmidt 2003). For the structural class IIB the domain VI is believed to play a dominant role in correct positioning of the 3’ splice site, while the γ –γ’ and EBS3-IBS3 interactions ensure the efficiency of the second splicing step (Lehmann and Schmidt 2003). An assumption one could draw from this and the above results would be that the extension does not interfere with the action of domain VI in positioning the 3’ splice site and the S2 stem ensures the efficiency of the second splicing step together with the γ –γ’ and EBS3-IBS3 interactions. To explain why *B.th.*I6a is dependent on the extension for splicing, one could suggest that there may have been some structural changes that made the γ –γ’ and EBS3-IBS3 interactions not work as they should, and the intron thus totally relies on the extension for an efficient exon ligation. Such changes could for example be reflected in the RNA features that distinguish introns of the α and β subgroups in the Bacterial B class (papers **3** and **5**).

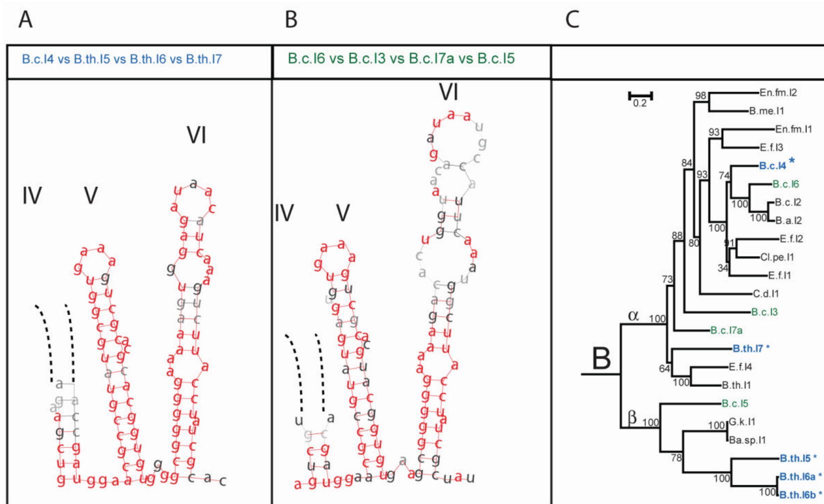


Figure 8. The figure illustrates that domain VI appears to be more conserved in sequence among the group II introns that harbor the 3’ extension than between these and closely related introns that do not carry an extension. The *B. cereus* and *B. thuringiensis* group II introns with (A) and without (B) the extension used to compare the conservation of domains V and VI are respectively colored in blue

and green in the phylogenetic tree of Bacterial B class introns (C). In (A) and (B) red-colored nucleotides are conserved among all four introns listed on the top, and lighter shades of grey indicate less conservation. The comparisons were made using RNAForester (Höchsmann et al. 2003; Höchsmann et al. 2004). The tree in panel C is redrawn from Figure 3B in paper 3.

Domains V and VI are the most conserved domains with respect to sequence and structure in group II introns, due to their critical activity in splicing. Nevertheless, it is interesting to see that these domains are more conserved among the introns carrying an extension, which may be relatively distant in the phylogenetic tree of the Bacterial B class than between these introns and their closest neighbors without such an extension (See Figure 8). The joint conservation of domains V and VI and the 3' extension among the unusual introns could possibly reflect a functional relationship. For example, could there be an interaction between the S2 stem and domain VI that ensures continued correct and efficient positioning of the 3' splice site? If yes, this would support the importance of domains V and/or VI for the extension, but it could also merely reflect the general importance of these domains in catalysis.

Splicing effects due to the extension resemble those related to the interaction between domains II and VI.

We suggest in paper **4** that the extension could be involved in an interaction similar to the η−η' interaction involving domains II and VI. This is based on t the observation that deletion of the whole extension from *B.c.*I4 and *B.th.*I6a are similar to those reported when modifying and/or deleting nucleotides in domain II or VI, more specifically the residues involved in the η−η' interaction (Chanfreau and Jacquier 1996; Costa et al. 1997). Disrupting the η−η' interaction in a class IIB intron or the equivalent interaction in a IIA intron leads to either a dramatic slowdown of the second splicing step or a shift in balance between hydrolysis or transesterification occurring in the first step (Chanfreau and Jacquier 1996; Costa et al. 1997). These effects are analogous to those resulting from the deletion of the extension in *B.th.*I6a and *B.c.*I4, respectively. The η−η' interaction is conserved in most IIB introns and usually takes the form of a GNRA terminal loop in domain VI interacting with two conserved C:G pairs in domain II (Lehmann and Schmidt 2003). In IIA class introns this is less preserved and suggests

replacement by other variants of such interaction partners or by an exchange of location of the loop and receptor. Remarkably, it has also been shown that replacing a domain VI from a IIA intron with one from a IIB gave an increase in first step splicing by transesterification (Schmidt et al. 1993).

It has also been demonstrated that when rendering the η–η' interaction stronger than the wildtype, the first splicing step is shifted towards hydrolysis only instead of primarily a transesterification reaction (Costa and Michel 1995; Costa et al. 1997). For *B.c.*I4 there is more hydrolysis when the extension is removed and could therefore indicate that the η–η' interaction is increased in strength. This could suggest that any interaction involving the extension works in the opposite orientation of the η–η' interaction. For the *B.th.*I6a intron it may be that the interaction of the extension completely compensates for a non-existing η–η' interaction. On the other hand, the functional difference involving the η–η' interaction between the IIA and IIB structural classes may imply that interaction to similar places/residues may lead to a somewhat different outcome (Costa et al. 1997; Michel et al. 2009).

Furthermore, the conservation of domain VI among the introns with an extension may be interesting as the theory above would suggest that the extension interacts with domain VI and/or II. The different splicing effects observed for the two introns may again rely on the secondary structural differences that are predicted to divide the introns of the Bacterial B class into two different subgroups, with *B.th.*I6a residing in the β subgroup and *B.c.*I4 in the α subgroup. Comparing the members of the β subgroup from paper **3** as well as additional introns recently identified (Figure 9C; N.J. Tourasse, unpublished results) revealed an interesting structural pattern. For some introns, the terminal loop of domain VI contains a GUAAUG loop matching the generic GN_n/RA tetraloop motif (Costa and Michel 1995; Abramovitz and Pyle 1997; Costa and Michel 1997), and those also have a branched stem in domain II (see schematic figures in Figure 9 A and B). In contrast, the introns without that particular terminal tetraloop have a single stem in domain II. These include *B.th.*I6a/b/c, *B.c.*I16 *B.th.*I5, and the recently discovered introns from *B.mycoides* and *pseudomycoides* that carry an extension (colored red), as well as two without an extension (colored pink). These introns have two internal loops within domain VI, both being adenosine–rich, and the topmost one has a guanosine on the 5'

side, except in *B.th.*I6/*B.c.*I16 (which has only an adenosine). One may speculate that this internal loop could be the tetraloop that interacts with domain II (e.g., with the two conserved G:C pairs), thus maintaining a η–η' interaction in these introns. This would imply that those introns colored red in figure 9C, except *B.th.*I6/*B.c.*I16, should be independent of the extension for an efficient second splicing step. This could be supported by the pink introns which has a similar structure in domain II and VI, but do not harbor an extension. It would be of great interest to investigate the splicing properties *in vitro* of *B.th.*I5 deleted of the extension, as this could narrow down and shed light on these possible candidates. A further test in that regard would be to see if "re-installing" such a η–η' interaction in *B.th.*I6a could render this intron independent of the extension.

A

IIB
IIA
III
I
IC2
IV
ORF
IC1
IB
IA
Ib
V
Ia
VI

B

III
II
I
IV
IC2
ORF
V
IC1
S1
IB
IA
Ib
Ia
VI
S2

C

α
100
B.c.I5
G.k.I1_
Geobacill
Ba.sp.I1
Geobacill
Clostridi
Subdoligr
B.my.I 1
B.pmy.I1
Geobacill
B.th.I5
B.th.I6a
B.th.I6c
B.th.I6b
B.c.I16a
B.c.I16b
β

Figure 9. Panels A, B and C illustrate a possible co-variation between domains II and VI in group II introns in the β subgroup of the Bacterial B class. Introns, where non harbor the extension, are indicated by blue-colored names in the phylogenetic tree in (C) and their characteristic features and motifs in domains II and VI are boxed in blue in the representative schematic secondary structure shown in (A), which is based on *B.c.*I5. Introns with the extension are colored red in (C) and their domain II and VI features are boxed in red in the representative secondary structure in (B), which is based on *B.th.*I5. In (B) the displayed nucleotides in domain VI are from *B.th.*I5, while the adenosine in parentheses is unique to *B.th.*I6/*B.c.*I16. The pink names refer to group II introns without an extension, but that show domains II and VI similar to the introns with the extension (colored red). Asterix in phylogenetic tree in (C) suggests a division point for these possible structural subgroups. The pound sign is ment to point to mutation of the guanosine to adenosine in domain VI, as illustrated in (B), that might the cause for the dependency of *B.th.*I6a and closely related introns. The phylogenetic tree shown in (C) is a subset of the tree in Figure 10 (see legend of the latter figure for details).

Both *B.c.*I4 and *B.th.*I7 of the α subgroup do harbor a conserved terminal loop with the characteristics of a tetraloop in domain VI, which is conserved between them and could play a similar role in a η–η' interaction. However, no obvious receptor site in domain II can be found and it is difficult to conclude from only two introns with an extension in the α subgroup. *B.th.*I7 and *B.th.*I5 intron constructs deleted of the extension need to be tested if one wants to draw some more generalized conclusion about the properties of the two subgroups within the bacterial B class.

The conserved residues in subdomain IC1

It is interesting in the light of the discussion above to remember that the conserved nucleotides in subdomain IC1 investigated in paper **3** showed a possible link between the extension and the first splicing step in *B.c.*I4. Mutation of these nucleotides gave a somewhat reduced first step of splicing, but no shifting of the balance between hydrolysis and branching, and the effects disappeared when removing the extension. In light of the new data in paper **4** showing that the extension of *B.c.*I4 also influences the first splicing step, these effects may appear more meaningful. For many group II introns there is an established θ–θ' interaction between subdomain IC1 and domain II, again in the form of a terminal tetraloop and a receptor site, respectively (Costa et al. 1997). This interaction is not obvious in all group II introns, but is believed to take place in some

other form (Lehmann and Schmidt 2003). Mutations of the θ–θ' interaction sites have indicated a primary role in stabilization of the correct folding of the catalytic core (Costa et al. 1997). Structurally, this interaction has been shown to position the ε' site, also residing in the IC1 stem, inside the core (Toor et al. 2008a). The 3' extension could in the scenario above replace or compensate for domain II in the η–η' and θ–θ' interactions, and therefore share a common interface with domains VI and IC1, bringing the ε–ε' interaction and the 5' intron-exon junction close to the bulged branchpoint A. In that way the extension could influence the balance between hydrolysis and transesterification in addition to the first step kinetics, along with efficient guiding of the 3' splice site. Removing the extension however improves the first splicing step when the two conserved basepairs in subdomain IC1 are mutated, therefore in *B.c.*I4 these basepairs may have a role in stabilizing the core only when the 3' extra segment is in place. It could be for example, as suggested in paper **3**, that the two basepairs in IC1 have a role in guiding or positioning the extension properly in the ribozyme and the effects seen on first-step splicing are only secondary results.

A conformational change in the group II intron core?

The hypothesis about the extension replacing the η–η' interaction and/or domain II is based on the fact that the splicing effects are similar when the sites are deleted. There is however some "controversy" on how these results should be interpreted, or at least how significant they are. Chanfreau and Jacquier (1996) suggested that the η–η' interaction initiates a conformational switch between the two splicing steps, allowing the 3' splice site to enter the catalytic core and the second step to take place. In other studies this has been deemed unlikely or at least that there can be no major rearrangements in the core (Costa et al. 1997; de Lencastre et al. 2005; Fedorova and Zingler 2007). UV photocrosslinking experiments showed that all the important residues in the (model) class IIB intron ai5γ that are involved in each of the splicing reactions are situated in close proximity before the onset of the first step (de Lencastre et al. 2005). A recent study revealed on the other hand several splicing-dependent crosslinks close to the η–η' interaction of a (model) class IIA intron Ll.LtrA (Dai et al. 2008). The recent crystal

structure of a group IIC intron with both exons indicates that both splicing steps are tightly coordinated within the active site with a continuous exon binding interface (Toor et al. 2008b). Supporting the notion that there are no major changes in the catalytic center is a study of the ai5γ intron showing that the bulged branchsite A crosslinked to the 5' end of the intron and still retained some splicing activity, suggesting that there would be no drastic change before the second splicing step (Hamill and Pyle 2006). The study does not rule out that there could be some additional catalytic strategies required for chemical reactivity, and a recent review points to aspects that could support a change in conformation in the core (Michel et al. 2009). The argumentation is that the triple helix constituted by domain V and the J2/3 linker region revealed by the crystal 3D structure should be put in place after the first step because the highly conserved GA dinucleotide in J2/3 is known to affect only the second step of splicing (Mikheeva et al. 2000; Michel et al. 2009). The various studies mentioned above were done on different intron classes and this could reflect why there is currently no clear uniform data on whether the catalytic core of group II introns undergoes some conformational changes or not. If there are such conformational changes under splicing of the group II intron, our scenario above implies that the 3' extension, instead of domain II, initiates these changes.

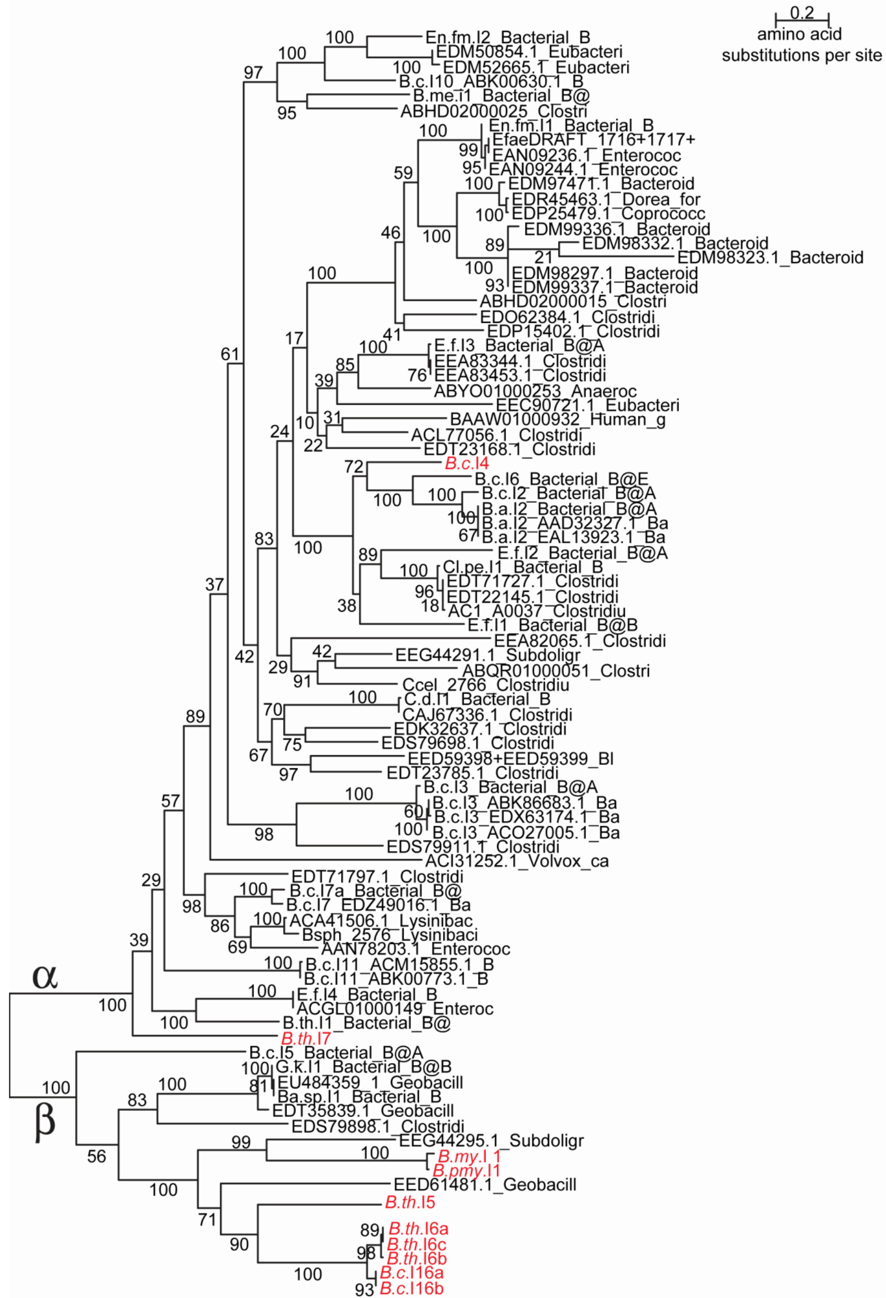
The origin of the extension

With the initial discovery of *B.c.*I4 in papers **1** and **2** the new 3' extension was thought to be only a single instance that demonstrated an extreme example of the flexibility and adaptability of group II introns. The discovery of the additional introns in papers **3** and **5** changed this view, suggesting that the 3' extension may be a more general feature among group II intron ribozymes. One of the most remarkable findings revealed by the phylogenetic analysis of the IEP in paper **3** is the fact that these special group II introns do not form a single lineage and, as discussed above, suggests that these group II introns do not represent a clonal evolutionary group or branch that developed due to the acquisition of the extension. Underlining the fact that these unusual introns are not all closely related are the different splicing effects observed when removing the extension from *B.th.*I6 and *B.c.*I4 in papers **2** and **4.** These observations can suggest that one intron

(*B.th.*I6) has undergone an evolution rendering it dependent on the extension for splicing whereas the other (*B.c.*I4) has not. This makes it difficult to understand the origin and evolution of the extension as there appears to be different reasons for maintaining the extension.

The various unusual group II introns may have adapted differently to splice with the extension when they each obtained it or they may have diverged from each other after one common acquisition event. One could suggest at least two scenarios; either the extension was acquired independently for each of the different introns, or the element has been part of these introns' evolution for a long time and the introns evolved into different lines as a part of the general evolution of group II introns. Both suggestions are reasonable and it could be a combination of the two. How the first acquisition happened in either case is a puzzling question, but it is more likely to be a single event that did not affect the splicing capabilities in any drastic way, as with *B.c.*I4, rather than a result of several events "building" the extra structural 3' element.

The joint conservation in sequence and structure of domains V, VI and the extension between the unusual introns could suggest that these elements are evolutionarily more closely related than other domains in these introns (see Figure 10). One explanation may be that they are the result of some sort of recombination of DV-DVII in events involving group II introns from the α and β subgroups of the Bacterial B class at the RNA or DNA level (Draper et al. 2008) (A theory that could, if domain IV is involved, explain why the extension is only present in introns from only one evolutionary class). However, as mentioned earlier, the similarities are likely to be due to that the conserved domains are important in general and/or that these are involved in interaction(s) with the extension and have therefore been conserved throughout the evolution for the unusual introns. Another type of independent acquisition event might be that the extension mediates its own mobility, and thus can target group II introns that are not very close evolutionarily. This would probably suggest some sort of new mobility mechanism.



0.2
amino acid
substitutions per site
En.fm.I2_Bacterial_B
EDM50854.1_Eubacteri
EDM52665.1_Eubacteri
B.c.I10_ABK00630.1_B
B.me.I1_Bacterial_B@
ABHD02000025_Clostri
En.fm.I1_Bacterial_B
EfaeDRAFT_1716+1717+
EAN09236.1_Enterococ
EAN09244.1_Enterococ
EDM97471.1_Bacteroid
EDR45463.1_Dorea_for
EDP25479.1_Coprococc
EDM99336.1_Bacteroid
EDM98332.1_Bacteroid
EDM98323.1_Bacteroid
EDM98297.1_Bacteroid
EDM99337.1_Bacteroid
ABHD02000015_Clostri
EDO62384.1_Clostridi
EDP15402.1_Clostridi
E.f.I3_Bacterial_B@A
EEA83344.1_Clostridi
EEA83453.1_Clostridi
ABYO01000253_Anaeroc
EEC90721.1_Eubacteri
BAAW01000932_Human_g
ACL77056.1_Clostridi
EDT23168.1_Clostridi
B.c.I4
B.c.I6_Bacterial_B@E
B.c.I2_Bacterial_B@A
B.a.I2_Bacterial_B@A
B.a.I2_AAD32327.1_Ba
B.a.I2_EAL13923.1_Ba
E.f.I2_Bacterial_B@A
Cl.pe.I1_Bacterial_B
EDT71727.1_Clostridi
EDT22145.1_Clostridi
AC1_A0037_Clostridiu
E.f.I1_Bacterial_B@B
EEA82065.1_Clostridi
EEG44291.1_Subdoligr
ABQR01000051_Clostri
Ccel_2766_Clostridiu
C.d.I1_Bacterial_B
CAJ67336.1_Clostridi
EDK32637.1_Clostridi
EDS79698.1_Clostridi
EED59398+EED59399_Bl
EDT23785.1_Clostridi
B.c.I3_Bacterial_B@A
B.c.I3_ABK86683.1_Ba
B.c.I3_EDX63174.1_Ba
B.c.I3_ACO27005.1_Ba
EDS79911.1_Clostridi
ACI31252.1_Volvox_ca
EDT71797.1_Clostridi
B.c.I7a_Bacterial_B@
B.c.I7_EDZ49016.1_Ba
ACA41506.1_Lysinibac
Bsph_2576_Lysinibaci
AAN78203.1_Enterococ
B.c.I11_ACM15855.1_B
B.c.I11_ABK00773.1_B
E.f.I4_Bacterial_B
ACGL01000149_Enteroc
B.th.I1_Bacterial_B@
B.th.I7
α
β
B.c.I5_Bacterial_B@A
G.k.I1_Bacterial_B@B
EU484359_1_Geobacill
Ba.sp.I1_Bacterial_B
EDT35839.1_Geobacill
EDS79898.1_Clostridi
EEG44295.1_Subdoligr
B.my.I 1
B.pmy.I1
EED61481.1_Geobacill
B.th.I5
B.th.I6a
B.th.I6c
B.th.I6b
B.c.I16a
B.c.I16b

Figure 10. Detailed rooted phylogenetic tree of 85 Bacterial B class group II introns, built the same way as in paper 3, Figure 3B, based on amino acid sequences covering the full length of the intron encoded ORF. Group II introns that harbor a 3' extension have their name colored in red. Numbers next to branch nodes indicate bootstrap support percentage values (out of 1000 replicates). The phylogenetic tree shows several new Bacterial B class introns, with and without the 3' extension, compared to that presented in paper 3 (N.J. Tourasse, unpublished data).

The other scenario would be that the extension has been part of the group II intron evolution from the early ancestors. The extension could originate as far back as the RNA world, and merely followed the general evolution of bacterial group II introns (Toor et al. 2001). If the extension has been part of the general intron evolution for a long time one would expect to find related extensions in other classes of group II introns (see Figure 6). However, these might deviate in structure and/or sequence, and therefore might not have been detected by sequence and structural searches. Alternatively, the extension could have entered at a later stage during intron evolution, e.g., in the ancestor of the Bacterial B class. In any case, the first acquisition event may have been neutral for ribozyme activity, being neither beneficial nor disadvantageous. The difference in distribution may suggest that in the α subgroup the extension may have been lost more frequently and only been kept in those two instances where it may have given some selective advantage. The close relatedness of the introns with the extension in the β subgroup could imply that they have become more or completely dependent on the extension for activity (See Figure 10). Either did all unusual introns in this subgroup become dependent on the extension because of the overall structural changes, or did the dependency occur specifically in *B.th.*I6. This could have happened, as suggested above, with the structural evolution of domains II and VI. The difference in distribution may suggest that the introns with the extension in the β subgroup have become more or completely dependent on the extension for activity. For the introns in the α subgroup, this theory suggests that the extension may have been lost more frequently and only been kept in those two instances where it may have given some selective advantage. To verify such a theory, *B.th.*I5 and *B.th.*I7 must be tested to determine if they are dependent on the extension or not. However, the distribution pattern may merely arise due to sampling bias, as there are many more group II introns known from the α subgroup than from the β subgroup.

It is striking to see that the unusual introns are only found in the Bacterial B class so far, but it is maybe more striking that these introns are actually only found in the *B. cereus* group of bacteria. Whether that means that the extension can only work in this group II intron class and bacterial species is not known. This peculiar distribution could suggest that the first example of these unusual introns may have arisen after group II introns entered these organisms, although this could also be sampling bias as discussed in papers **3** and **5**.

Concluding remarks

The work described herein shows how a new structural element can be acquired to fulfill properties/functions that are solved by other elements in other (typical) group II introns. However, this study does not present new catalytic reactions or properties that can be assigned to the novel extension. This work has also given some insights to formulate and test hypotheses regarding how the 3' segment may fit into the large ribozyme structure and its subsequent effects on splicing. The open question of origin of the extension is also of great interest.

Future perspectives

If the extension does actively interact with the ribozyme structure, (which has to be presumed from the results of this work), what type of interaction does that involve and to what intronic element(s) or sites? As discussed above, there are several candidate domains and areas that could be interacting with the extension. In paper **3** we point out that the asymmetric internal loop in stem S2 shows strong resemblance to the 11-nt tetraloop receptor motif, which is a common building block of catalytic RNAs (Costa and Michel 1995), and this conserved region would be an obvious site of interaction creating a strong, compact interaction with a GAAA loop. Quite intriguing is that the S1 stem with its terminal loop shows a strong resemblance to a generic GN_n/RA loop, either GAAAU or GUAAA, and similarly this is also a strong candidate for interaction with other

intronic elements. Although most of the mutations in the S1 and S2 regions conducted on *B.c.*I4 in ammonium sulfate buffer did not give any clear effect on splicing, more visible effects may be detected under other splicing conditions or in other introns. Therefore, it would be of great interest to test some of the previously studied extension mutants in the potassium chloride buffer and/or conduct similar mutations in *B.th.*I6a, which might give a clearer picture of any important sites inside the extension due to the dependency of this intron. Besides the terminal loop of S1 and the internal loop of S2, the basal stem of S2 may be a candidate for further functional studies *in vitro* as it contains several conserved basepairs.

Identifying any important areas or residues might also give some indications of what type of residues or sequence motifs one should look for in other parts of the intron. The lack of conservation between the unusual introns outside domains V, VI, and the extension may of course make it hard to find one specific site, but on the other hand analysis of the phylogenetic and structural differences between the introns might give clues on possibly different strategiesfor adaptation to the extension. In this regard, as mentioned above, it will be important to determine if the two other introns *B.th.*I5 and *B.th.*I7 are dependent, or not, on the extension for function. Such comparative analyses might provide valuable information, but further functional biochemical studies will have to be conducted to give hard evidence for any interaction with the ribozyme structure. These biochemical studies could include nucleotide modification by DMS or DEPC, incorporation of nucleotide analogues (such as NAIM or NAIS technique) or photoreactive residues for photocrosslinking experiments that can be used to determine or identify intronic residues that are in direct contact or close proximity of the S1 and/or S2 stem. This may provide insights into how the 3’ extension is acommodated in the overall tertiary structure of the group II ribozyme, which is a fascinating question.

References

Abelson, J. (2008) Is the spliceosome a ribonucleoprotein enzyme? *Nature structural & molecular biology* **15**, 1235-1237.

Abramovitz, D.L. and Pyle, A.M. (1997) Remarkable morphological variability of a common RNA folding motif: the GNRA tetraloop-receptor interaction. *Journal of molecular biology* **266**, 493-506.

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

Agerso, Y., Jensen, L.B., Givskov, M. and Roberts, M.C. (2002) The identification of a tetracycline resistance gene tet(M), on a Tn*916*-like transposon, in the *Bacillus cereus* group. *FEMS microbiology letters* **214**, 251-256.

Aizawa, Y., Xiang, Q., Lambowitz, A.M. and Pyle, A.M. (2003) The pathway for DNA recognition and RNA integration by a group II intron retrotransposon. *Molecular cell* **11**, 795-805.

Aronson, A. (2002) Sporulation and delta-endotoxin synthesis by *Bacillus thuringiensis*. *Cell Mol Life Sci* **59**, 417-425.

Ash, C., Farrow, J.A., Dorsch, M., Stackebrandt, E. and Collins, M.D. (1991) Comparative analysis of *Bacillus anthracis*, *Bacillus cereus*, and related species on the basis of reverse transcriptase sequencing of 16S rRNA. *Int J Syst Bacteriol* **41**, 343-346.

Beauregard, A., Curcio, M.J. and Belfort, M. (2008) The take and give between retrotransposable elements and their hosts. *Annual review of genetics* **42**, 587-617.

Belhocine, K., Mak, A.B. and Cousineau, B. (2008) Trans-splicing versatility of the Ll.LtrB group II intron. *RNA (New York, NY* **14**, 1782-1790.

Berry, C., O'Neil, S., Ben-Dov, E., Jones, A.F., Murphy, L., Quail, M.A., Holden, M.T., Harris, D., Zaritsky, A. and Parkhill, J. (2002) Complete sequence and organization of pBtoxis, the toxin-coding plasmid of *Bacillus thuringiensis* subsp. *israelensis*. *Applied and environmental microbiology* **68**, 5082-5095.

Bogdanova, E.S., Bass, I.A., Minakhin, L.S., Petrova, M.A., Mindlin, S.Z., Volodin, A.A., Kalyaeva, E.S., Tiedje, J.M., Hobman, J.L., Brown, N.L. and Nikiforov, V.G. (1998) Horizontal spread of mer operons among gram-positive bacteria in natural environments. *Microbiology (Reading, England)* **144 (Pt 3)**, 609-620.

Boudvillain, M., de Lencastre, A. and Pyle, A.M. (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. and Pyle, A.M. (1998) Defining functional groups, core structural features and inter-domain tertiary contacts essential for group II intron self-splicing: a NAIM analysis. *The EMBO journal* **17**, 7091-7104.

Braun, V., Mehlig, M., Moos, M., Rupnik, M., Kalt, B., Mahony, D.E. and von Eichel-Streiber, C. (2000) A chimeric ribozyme in *Clostridium difficile* combines features of group I introns and insertion elements. *Molecular microbiology* **36**, 1447-1459.

Brussow, H., Canchaya, C. and Hardt, W.D. (2004) Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev* **68**, 560-602.

Calvin, K. and Li, H. (2008) RNA-splicing endonuclease structure and function. *Cell Mol Life Sci* **65**, 1176-1185.

Carlson, C.R., Caugant, D.A. and Kolsto, A.B. (1994) Genotypic Diversity among Bacillus cereus and Bacillus thuringiensis Strains. *Applied and environmental microbiology* **60**, 1719-1725.

Chanfreau, G. and Jacquier, A. (1996) An RNA conformational change between the two chemical steps of group II self-splicing. *The EMBO journal* **15**, 3466-3476.

Chee, G.J. and Takami, H. (2005) Housekeeping *recA* gene interrupted by group II intron in the thermophilic *Geobacillus kaustophilus*. *Gene* **363**, 211-220.

Chen, I. and Dubnau, D. (2004) DNA uptake during bacterial transformation. *Nat Rev Microbiol* **2**, 241-249.

Chen, J. and Novick, R.P. (2009) Phage-mediated intergeneric transfer of toxin genes. *Science (New York, NY* **323**, 139-141.

Chin, K. and Pyle, A.M. (1995) Branch-point attack in group II introns is a highly reversible transesterification, providing a potential proofreading mechanism for 5'-splice site selection. *RNA (New York, NY* **1**, 391-406.

Collins, C.A. and Guthrie, C. (2000) The question remains: is the spliceosome a ribozyme? *Nature structural biology* **7**, 850-854.

Coros, C.J., Landthaler, M., Piazza, C.L., Beauregard, A., Esposito, D., Perutka, J., Lambowitz, A.M. and Belfort, M. (2005) Retrotransposition strategies of the *Lactococcus lactis* Ll.LtrB group II intron are dictated by host identity and cellular environment. *Molecular microbiology* **56**, 509-524.

Coros, C.J., Piazza, C.L., Chalamcharla, V.R. and Belfort, M. (2008) A mutant screen reveals RNase E as a silencer of group II intron retromobility in *Escherichia coli*. *RNA (New York, NY*.

Coros, C.J., Piazza, C.L., Chalamcharla, V.R., Smith, D. and Belfort, M. (2009) Global regulators orchestrate group II intron retromobility. *Molecular cell* **34**, 250-256.

Costa, M., Deme, E., Jacquier, A. and Michel, F. (1997) Multiple tertiary interactions involving domain II of group II self-splicing introns. *Journal of molecular biology* **267**, 520-536.

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

Costa, M. and Michel, F. (1997) Rules for RNA recognition of GNRA tetraloops deduced by *in vitro* selection: comparison with *in vivo* evolution. *The EMBO journal* **16**, 3289-3302.

Costa, M., Michel, F. and Westhof, E. (2000) A three-dimensional perspective on exon binding by a group II self-splicing intron. *The EMBO journal* **19**, 5007-5018.

Craig, N.L., Craigie, Robert, Gellert, Martin, Lambowitz, Alan M. (2002) *Mobile DNA II*. Washington, D.C.: ASM Press.

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

Curcio, M.J. and Derbyshire, K.M. (2003) The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Biol* **4**, 865-877.

Dai, L., Chai, D., Gu, S.Q., Gabel, J., Noskov, S.Y., Blocker, F.J., Lambowitz, A.M. and Zimmerly, S. (2008) A Three-Dimensional Model of a Group II Intron RNA and Its Interaction with the Intron-Encoded Reverse Transcriptase. *Molecular cell* **30**, 1-14.

Dai, L., Toor, N., Olson, R., Keeping, A. and Zimmerly, S. (2003) Database for mobile group II introns. *Nucleic acids research* **31**, 424-426.

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

Dai, L. and Zimmerly, S. (2002b) The dispersal of five group II introns among natural populations of *Escherichia coli*. *RNA (New York, NY* **8**, 1294-1307.

Dai, L. and 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 (New York, NY* **9**, 14-19.

Daniels, D.L., Michels, W.J., Jr. and Pyle, A.M. (1996) Two competing pathways for self-splicing by group II introns: a quantitative analysis of *in vitro* reaction rates and products. *Journal of molecular biology* **256**, 31-49.

De Gregorio, E., Silvestro, G., Petrillo, M., Carlomagno, M.S. and Di Nocera, P.P. (2005) Enterobacterial repetitive intergenic consensus sequence repeats in yersiniae: genomic organization and functional properties. *Journal of bacteriology* **187**, 7945-7954.

de Lencastre, A., Hamill, S. and Pyle, A.M. (2005) A single active-site region for a group II intron. *Nature structural & molecular biology* **12**, 626-627.

De Lencastre, A. and Pyle, A.M. (2008) Three essential and conserved regions of the group II intron are proximal to the 5'-splice site. *RNA (New York, NY* **14**, 11-24.

De Palmenaer, D., Vermeiren, C. and Mahillon, J. (2004) IS231-MIC231 elements from *Bacillus cereus sensu lato* are modular. *Molecular microbiology* **53**, 457-467.

Delihas, N. (2008) Small mobile sequences in bacteria display diverse structure/function motifs. *Molecular microbiology* **67**, 475-481.

Doolittle, W.F. and Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601-603.

Draper, W.E., Hayden, E.J. and Lehman, N. (2008) Mechanisms of covalent self-assembly of the Azoarcus ribozyme from four fragment oligonucleotides. *Nucleic acids research* **36**, 520-531.

Drobniewski, F.A. (1993) *Bacillus cereus* and related species. *Clin Microbiol Rev* **6**, 324-338.

Edgell, D.R., Belfort, M. and Shub, D.A. (2000) Barriers to intron promiscuity in bacteria. *Journal of bacteriology* **182**, 5281-5289.

Ehling-Schulz, M., Fricker, M., Grallert, H., Rieck, P., Wagner, M. and Scherer, S. (2006) Cereulide synthetase gene cluster from emetic *Bacillus cereus*: structure and location on a mega virulence plasmid related to *Bacillus anthracis* toxin plasmid pXO1. *BMC microbiology* **6**, 20.

Ehling-Schulz, M., Svensson, B., Guinebretiere, M.H., Lindback, T., Andersson, M., Schulz, A., Fricker, M., Christiansson, A., Granum, P.E., Martlbauer, E., Nguyen-The, C., Salkinoja-Salonen, M. and Scherer, S. (2005) Emetic toxin formation of *Bacillus cereus* is restricted to a single evolutionary lineage of closely related strains. *Microbiology (Reading, England)* **151**, 183-197.

Fedorova, O., Mitros, T. and Pyle, A.M. (2003) Domains 2 and 3 interact to form critical elements of the group II intron active site. *Journal of molecular biology* **330**, 197-209.

Fedorova, O. and Pyle, A.M. (2005) Linking the group II intron catalytic domains: tertiary contacts and structural features of domain 3. *The EMBO journal* **24**, 3906-3916.

Fedorova, O. and Pyle, A.M. (2008) A conserved element that stabilizes the group II intron active site. *RNA (New York, NY* **14**, 1048-1056.

Fedorova, O., Waldsich, C. and Pyle, A.M. (2007) Group II intron folding under near-physiological conditions: collapsing to the near-native state. *Journal of molecular biology* **366**, 1099-1114.

Fedorova, O. and Zingler, N. (2007) Group II introns: structure, folding and splicing mechanism. *Biological chemistry* **388**, 665-678.

Ferat, J.L., Le Gouar, M. and Michel, F. (2003) A group II intron has invaded the genus *Azotobacter* and is inserted within the termination codon of the essential *groEL* gene. *Molecular microbiology* **49**, 1407-1423.

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

Feschotte, C. and Wessler, S.R. (2002) Mariner-like transposases are widespread and diverse in flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 280-285.

Frost, L.S., Leplae, R., Summers, A.O. and Toussaint, A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* **3**, 722-732.

Goodier, J.L. and Kazazian, H.H., Jr. (2008) Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* **135**, 23-35.

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

Grindley, N.D., Whiteson, K.L. and Rice, P.A. (2006) Mechanisms of Site-Specific Recombination. *Annu Rev Biochem*.

Grynberg, M. and Godzik, A. (2004) NERD: a DNA processing-related domain present in the anthrax virulence plasmid, pXO1. *Trends in biochemical sciences* **29**, 106-110.

Hamill, S. and Pyle, A.M. (2006) The receptor for branch-site docking within a group II intron active site. *Molecular cell* **23**, 831-840.

Han, C.S., Xie, G., Challacombe, J.F., Altherr, M.R., Bhotika, S.S., Brown, N., Bruce, D., Campbell, C.S., Campbell, M.L., Chen, J., Chertkov, O., Cleland, C., Dimitrijevic, M., Doggett, N.A., Fawcett, J.J., Glavina, T., Goodwin, L.A., Green, L.D., Hill, K.K., Hitchcock, P., Jackson, P.J., Keim, P., Kewalramani, A.R., Longmire, J., Lucas, S., Malfatti, S., McMurry, K., Meincke, L.J., Misra, M., Moseman, B.L., Mundt, M., Munk, A.C., Okinaka, R.T., Parson-Quintana, B., Reilly, L.P., Richardson, P., Robinson, D.L., Rubin, E., Saunders, E., Tapia, R., Tesmer, J.G., Thayer, N., Thompson, L.S., Tice, H., Ticknor, L.O., Wills, P.L., Brettin, T.S. and Gilna, P. (2006) Pathogenomic sequence analysis of *Bacillus cereus* and *Bacillus thuringiensis* isolates closely related to *Bacillus anthracis*. *Journal of bacteriology* **188**, 3382-3390.

Harris-Kerr, C.L., Zhang, M. and Peebles, C.L. (1993) The phylogenetically predicted base-pairing interaction between alpha and alpha' is required for group II splicing in vitro. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 10658-10662.

Hasselmayer, O., Nitsche, C., Braun, V. and von Eichel-Streiber, C. (2004) The IStron Cd*ISt*1 of *Clostridium difficile*: molecular symbiosis of a group I intron and an insertion element. *Anaerobe* **10**, 85-92.

Heierson, A., Landen, R., Lovgren, A., Dalhammar, G. and Boman, H.G. (1987) Transformation of vegetative cells of Bacillus thuringiensis by plasmid DNA. *Journal of bacteriology* **169**, 1147-1152.

Helgason, E., Caugant, D.A., Olsen, I. and Kolstø, A.B. (2000a) Genetic structure of population of *Bacillus cereus* and *B. thuringiensis* isolates associated with periodontitis and other human infections. *J Clin Microbiol* **38**, 1615-1622.

Helgason, E., Tourasse, N.J., Meisal, R., Caugant, D.A. and Kolstø, A.B. (2004) Multilocus sequence typing scheme for bacteria of the *Bacillus cereus* group. *Applied and environmental microbiology* **70**, 191-201.

Helgason, E., Økstad, O.A., Caugant, D.A., Johansen, H.A., Fouet, A., Mock, M., Hegna, I. and Kolsto, A.B. (2000b) *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*--one species on the basis of genetic evidence. *Applied and environmental microbiology* **66**, 2627-2630.

Hendriksen, N.B., Hansen, B.M. and Johansen, J.E. (2006) Occurrence and pathogenic potential of *Bacillus cereus* group bacteria in a sandy loam. *Antonie Van Leeuwenhoek* **89**, 239-249.

Hoffmaster, A.R., Hill, K.K., Gee, J.E., Marston, C.K., De, B.K., Popovic, T., Sue, D., Wilkins, P.P., Avashia, S.B., Drumgoole, R., Helma, C.H., Ticknor, L.O., Okinaka, R.T. and Jackson, P.J. (2006) Characterization of *Bacillus cereus* isolates associated with fatal pneumonias: strains are closely related to *Bacillus anthracis* and harbor *B. anthracis* virulence genes. *J Clin Microbiol* **44**, 3352-3360.

Hoffmaster, A.R., Ravel, J., Rasko, D.A., Chapman, G.D., Chute, M.D., Marston, C.K., De, B.K., Sacchi, C.T., Fitzgerald, C., Mayer, L.W., Maiden, M.C., Priest, F.G., Barker, M., Jiang, L., Cer, R.Z., Rilstone, J., Peterson, S.N., Weyant, R.S., Galloway, D.R., Read, T.D., Popovic, T. and Fraser, C.M. (2004) Identification of anthrax toxin genes in a *Bacillus cereus* associated with an illness resembling inhalation anthrax. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 8449-8454.

Hoton, F.M., Andrup, L., Swiecicka, I. and Mahillon, J. (2005) The cereulide genetic determinants of emetic Bacillus cereus are plasmid-borne. *Microbiology (Reading, England)* **151**, 2121-2124.

Hu, X., Van der Auwera, G., Timmery, S., Zhu, L. and Mahillon, J. (2009) Distribution, diversity and potential mobility of extra-chromosomal elements related to the Bacillus anthracis pXO1 and pXO2 virulence plasmids. *Applied and environmental microbiology*.

Höchsmann, M., Töller, T., Giegerich, R. and Kurtz, S. (2003) Local similarity in RNA secondary structures. *Proc IEEE Comput Soc Bioinform Conf* **2**, 159-168.

Höchsmann, M., Voss, B. and Giegerich, R. (2004) Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE/ACM Trans Comput Biol Bioinform* **1**, 53-62.

Ichiyanagi, K., Beauregard, A. and Belfort, M. (2003) A bacterial group II intron favors retrotransposition into plasmid targets. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 15742-15747.

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

Ivanova, N., Sorokin, A., Anderson, I., Galleron, N., Candelon, B., Kapatral, V., Bhattacharyya, A., Reznik, G., Mikhailova, N., Lapidus, A., Chu, L., Mazur, M., Goltsman, E., Larsen, N., D'Souza, M., Walunas, T., Grechkin, Y., Pusch, G., Haselkorn, R., Fonstein, M., Ehrlich, S.D., Overbeek, R. and Kyrpides, N. (2003) Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. *Nature* **423**, 87-91.

Jacquier, A. and Jacquesson-Breuleux, N. (1991) Splice site selection and role of the lariat in a group II intron. *Journal of molecular biology* **219**, 415-428.

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

Jacquier, A. and Michel, F. (1990) Base-pairing interactions involving the 5' and 3'-terminal nucleotides of group II self-splicing introns. *Journal of molecular biology* **213**, 437-447.

Jarrell, K.A., Peebles, C.L., Dietrich, R.C., Romiti, S.L. and Perlman, P.S. (1988) Group II intron self-splicing. Alternative reaction conditions yield novel products. *The Journal of biological chemistry* **263**, 3432-3439.

Jensen, G.B., Hansen, B.M., Eilenberg, J. and Mahillon, J. (2003) The hidden lifestyles of *Bacillus cereus* and relatives. *Environ Microbiol* **5**, 631-640.

Keating, K.S., Toor, N. and Pyle, A.M. (2008) The GANC tetraloop: a novel motif in the group IIC intron structure. *Journal of molecular biology* **383**, 475-481.

Kersulyte, D., Velapatino, B., Dailide, G., Mukhopadhyay, A.K., Ito, Y., Cahuayme, L., Parkinson, A.J., Gilman, R.H. and Berg, D.E. (2002) Transposable element IS*Hp608* of *Helicobacter pylori*: nonrandom geographic distribution, functional organization, and insertion specificity. *Journal of bacteriology* **184**, 992-1002.

Klevan, A., Tourasse, N.J., Stabell, F.B., Kolstø, A.B. and Økstad, O.A. (2007) Exploring the evolution of the *Bacillus cereus* group repeat element *bcr1* by comparative genome analysis of closely related strains. *Microbiology (Reading, England)* **153**, 3894-3908.

Knoop, V., Altwasser, M. and Brennicke, A. (1997) A tripartite group II intron in mitochondria of an angiosperm plant. *Mol Gen Genet* **255**, 269-276.

Ko, M., Choi, H. and Park, C. (2002) Group I self-splicing intron in the recA gene of *Bacillus anthracis*. *Journal of bacteriology* **184**, 3917-3922.

Koch, J.L., Boulanger, S.C., Dib-Hajj, S.D., Hebbar, S.K. and Perlman, P.S. (1992) Group II introns deleted for multiple substructures retain self-splicing activity. *Molecular and cellular biology* **12**, 1950-1958.

Koehler, T.M. (2002) Bacillus anthracis genetics and virulence gene regulation. *Current topics in microbiology and immunology* **271**, 143-164.

Kolsto, A.B., Tourasse, N.J. and Okstad, O.A. (2009) What Sets Bacillus anthracis Apart from Other Bacillus Species? *Annu Rev Microbiol*.

Koonin, E.V. and Wolf, Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic acids research* **36**, 6688-6719.

Koretskaia, N.G., Svetoch, O., Kachaturian, S.V. and Dobrina, A.P. (1989) [Transduction ability of mutants of phage CP51, virulent for bacteria of the Bacillus cereus group]. *Genetika* **25**, 1013-1020.

Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E. and Cech, T.R. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31**, 147-157.

Lambowitz, A.M. and Zimmerly, S. (2004) Mobile group II introns. *Annual review of genetics* **38**, 1-35.

Lechner, S., Mayr, R., Francis, K.P., Pruss, B.M., Kaplan, T., Wiessner-Gunkel, E., Stewart, G.S. and Scherer, S. (1998) *Bacillus weihenstephanensis* sp. nov. is a new psychrotolerant species of the *Bacillus cereus* group. *Int J Syst Bacteriol* **48 Pt 4**, 1373-1382.

Lehmann, K. and Schmidt, U. (2003) Group II introns: structure and catalytic versatility of large natural ribozymes. *Crit Rev Biochem Mol Biol* **38**, 249-303.

Lengeler, J.W., Drews, G. and Schlegel, H.G. (1999) *Biology of the prokaryotes*. Stuttgart ; New York
Malden, MA: Thieme ;
Distributed in the USA by Blackwell Science.

Leonard, C., Chen, Y. and Mahillon, J. (1997) Diversity and differential distribution of IS231, IS232 and IS240 among Bacillus cereus, Bacillus thuringiensis and Bacillus mycoides. *Microbiology (Reading, England)* **143 (Pt 8)**, 2537-2547.

Li-Pook-Than, J. and Bonen, L. (2006) Multiple physical forms of excised group II intron RNAs in wheat mitochondria. *Nucleic acids research* **34**, 2782-2790.

Liu, X.Q., Yang, J. and Meng, Q. (2003) Four inteins and three group II introns encoded in a bacterial ribonucleotide reductase gene. *The Journal of biological chemistry* **278**, 46826-46831.

Lynch, M. (2006) Streamlining and simplification of microbial genome architecture. *Annu Rev Microbiol* **60**, 327-349.

Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. *Science (New York, NY* **302**, 1401-1404.

Mahillon, J., Rezsohazy, R., Hallet, B. and Delcour, J. (1994) IS231 and other Bacillus thuringiensis transposable elements: a review. *Genetica* **93**, 13-26.

Mahillon, J., Seurinck, J., van Rompuy, L., Delcour, J. and Zabeau, M. (1985) Nucleotide sequence and structural organization of an insertion sequence element (IS231) from Bacillus thuringiensis strain berliner 1715. *The EMBO journal* **4**, 3895-3899.

Malik, H.S., Burke, W.D. and Eickbush, T.H. (1999) The age and evolution of non-LTR retrotransposable elements. *Molecular biology and evolution* **16**, 793-805.

Martinez-Abarca, F., Barrientos-Duran, A., Fernandez-Lopez, M. and Toro, N. (2004) The RmInt1 group II intron has two different retrohoming pathways for mobility using predominantly the nascent lagging strand at DNA replication forks for priming. *Nucleic acids research* **32**, 2880-2888.

Matsuura, M., Noah, J.W. and Lambowitz, A.M. (2001) Mechanism of maturase-promoted group II intron splicing. *The EMBO journal* **20**, 7259-7270.

McSwiggen, J.A. and Cech, T.R. (1989) Stereochemistry of RNA cleavage by the Tetrahymena ribozyme and evidence that the chemical step is not rate-limiting. *Science (New York, NY* **244**, 679-683.

Michel, F., Costa, M., Doucet, A.J. and Ferat, J.L. (2007) Specialized lineages of bacterial group II introns. *Biochimie* **89**, 542-553.

Michel, F., Costa, M. and Westhof, E. (2009) The ribozyme core of group II introns: a structure in want of partners. *Trends in biochemical sciences*.

Michel, F. and Dujon, B. (1983) Conservation of RNA secondary structures in two intron families including mitochondrial-, chloroplast- and nuclear-encoded members. *The EMBO journal* **2**, 33-38.

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

Mikheeva, S., Murray, H.L., Zhou, H., Turczyk, B.M. and Jarrell, K.A. (2000) Deletion of a conserved dinucleotide inhibits the second step of group II intron splicing. *RNA (New York, NY* **6**, 1509-1515.

Mock, M. and Fouet, A. (2001) Anthrax. *Annu Rev Microbiol* **55**, 647-671.

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

Molina-Sanchez, M.D., Martinez-Abarca, F. and Toro, N. (2006) Excision of the *Sinorhizobium meliloti* group II intron RmInt1 as circles *in vivo*. *The Journal of biological chemistry* **281**, 28737-28744.

Moran, N.A. and Plague, G.R. (2004) Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev* **14**, 627-633.

Mueller, J.E., Clyman, J., Huang, Y.J., Parker, M.M. and Belfort, M. (1996) Intron mobility in phage T4 occurs in the context of recombination-dependent DNA replication by way of multiple pathways. *Genes & development* **10**, 351-364.

Muller, M.W., Stocker, P., Hetzer, M. and Schweyen, R.J. (1991) Fate of the junction phosphate in alternating forward and reverse self-splicing reactions of group II intron RNA. *Journal of molecular biology* **222**, 145-154.

Murray, H.L., Mikheeva, S., Coljee, V.W., Turczyk, B.M., Donahue, W.F., Bar-Shalom, A. and Jarrell, K.A. (2001) Excision of group II introns as circles. *Molecular cell* **8**, 201-211.

Nakamura, L.K. (1998) *Bacillus pseudomycoides* sp. nov. *Int J Syst Bacteriol* **48 Pt 3**, 1031-1035.

Nakamura, L.K. and Jackson, M.A. (1995) Clarification of the taxonomy of *Bacillus mycoides*. *Int J Syst Bacteriol* **45**, 46-49.

Noah, J.W. and Lambowitz, A.M. (2003) Effects of maturase binding and Mg2+ concentration on group II intron RNA folding investigated by UV cross-linking. *Biochemistry* **42**, 12466-12480.

Nord, D. and Sjöberg, B.M. (2008) Unconventional GIY-YIG homing endonuclease encoded in group I introns in closely related strains of the *Bacillus cereus* group. *Nucleic acids research* **36**, 300-310.

Nord, D., Torrents, E. and Sjöberg, B.M. (2007) A functional homing endonuclease in the *Bacillus anthracis nrdE* group I intron. *Journal of bacteriology* **189**, 5293-5301.

Ochman, H. (2005) Genomes on the shrink. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 11959-11960.

Ochman, H. and Davalos, L.M. (2006) The nature and dynamics of bacterial genomes. *Science (New York, NY* **311**, 1730-1733.

Okinaka, R., Cloud, K., Hampton, O., Hoffmaster, A., Hill, K., Keim, P., Koehler, T., Lamke, G., Kumano, S., Manter, D., Martinez, Y., Ricke, D., Svensson, R. and Jackson, P. (1999a) Sequence, assembly and analysis of pXO1 and pXO2. *J Appl Microbiol* **87**, 261-262.

Okinaka, R.T., Cloud, K., Hampton, O., Hoffmaster, A.R., Hill, K.K., Keim, P., Koehler, T.M., Lamke, G., Kumano, S., Mahillon, J., Manter, D., Martinez, Y., Ricke, D., Svensson, R. and Jackson, P.J. (1999b) Sequence and organization of pXO1, the large *Bacillus anthracis* plasmid harboring the anthrax toxin genes. *Journal of bacteriology* **181**, 6509-6515.

Okstad, O.A., Gominet, M., Purnelle, B., Rose, M., Lereclus, D. and Kolsto, A.B. (1999a) Sequence analysis of three Bacillus cereus loci carrying PIcR-regulated genes encoding degradative enzymes and enterotoxin. *Microbiology (Reading, England)* **145 (Pt 11)**, 3129-3138.

Okstad, O.A., Hegna, I., Lindback, T., Rishovd, A.L. and Kolsto, A.B. (1999b) Genome organization is not conserved between Bacillus cereus and Bacillus subtilis. *Microbiology (Reading, England)* **145 (Pt 3)**, 621-631.

Okstad, O.A., Tourasse, N.J., Stabell, F.B., Sundfaer, C.K., Egge-Jacobsen, W., Risoen, P.A., Read, T.D. and Kolsto, A.B. (2004a) The bcr1 DNA repeat element is specific to the Bacillus cereus group and exhibits mobile element characteristics. *Journal of bacteriology* **186**, 7714-7725.

Okstad, O.A., Tourasse, N.J., Stabell, F.B., Sundfaer, C.K., Egge-Jacobsen, W., Risøen, P.A., Read, T.D. and Kolstø, A.B. (2004b) The *bcr1* DNA repeat element is specific to the *Bacillus cereus* group and exhibits mobile element characteristics. *Journal of bacteriology* **186**, 7714-7725.

Orgel, L.E. and Crick, F.H. (1980) Selfish DNA: the ultimate parasite. *Nature* **284**, 604-607.

Padgett, R.A., Podar, M., Boulanger, S.C. and Perlman, P.S. (1994) The stereochemical course of group II intron self-splicing. *Science (New York, NY* **266**, 1685-1688.

Pallen, M.J. and Wren, B.W. (2007) Bacterial pathogenomics. *Nature* **449**, 835-842.

Pannucci, J., Okinaka, R.T., Sabin, R. and Kuske, C.R. (2002) *Bacillus anthracis* pXO1 plasmid sequence conservation among closely related bacterial species. *Journal of bacteriology* **184**, 134-141.

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L.D., Thomson, N., Harris, D.E., Holden, M.T., Churcher, C.M., Bentley, S.D., Mungall, K.L., Cerdeno-Tarraga, A.M., Temple, L., James, K., Harris, B., Quail, M.A., Achtman, M., Atkin, R., Baker, S., Basham, D., Bason, N., Cherevach, I., Chillingworth, T., Collins, M., Cronin, A., Davis, P., Doggett, J., Feltwell, T., Goble, A., Hamlin, N., Hauser, H., Holroyd, S., Jagels, K., Leather, S., Moule, S., Norberczak, H., O'Neil, S., Ormond, D., Price, C., Rabbinowitsch, E., Rutter, S., Sanders, M., Saunders, D., Seeger, K., Sharp, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Unwin, L., Whitehead, S., Barrell, B.G. and Maskell, D.J. (2003) Comparative analysis of the

genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. *Nat Genet* **35**, 32-40.

Peebles, C.L., Benatan, E.J., Jarrell, K.A. and Perlman, P.S. (1987) Group II intron self-splicing: development of alternative reaction conditions and identification of a predicted intermediate. *Cold Spring Harbor symposia on quantitative biology* **52**, 223-232.

Phillips, G. and Funnell, B.E. (2004a) *Plasmid biology*. Washington, D.C.: ASM Press.

Phillips, G. and Funnell, B.E. (2004b) *Plasmid biology*. Washington, D.C.: ASM Press.

Piccirilli, J.A. (2008) Biochemistry. Toward understanding self-splicing. *Science (New York, NY* **320**, 56-57.

Podar, M., Chu, V.T., Pyle, A.M. and Perlman, P.S. (1998) Group II intron splicing *in vivo* by first-step hydrolysis. *Nature* **391**, 915-918.

Podar, M., Perlman, P.S. and Padgett, R.A. (1995) Stereochemical selectivity of group II intron splicing, reverse splicing, and hydrolysis reactions. *Molecular and cellular biology* **15**, 4466-4478.

Porwal, S., Lal, S., Cheema, S. and Kalia, V.C. (2009) Phylogeny in aid of the present and novel microbial lineages: diversity in Bacillus. *PLoS ONE* **4**, e4438.

Priest, F.G., Barker, M., Baillie, L.W., Holmes, E.C. and Maiden, M.C. (2004) Population structure and evolution of the *Bacillus cereus* group. *Journal of bacteriology* **186**, 7959-7970.

Pyle, A.M. and Lambowitz, A.M. (2006) Group II introns: ribozymes that splice RNA and invade DNA. In *The RNA world, third edition* eds. Gesteland, R.F., Cech, T.R. and Atkins, J.F. pp.469-505. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Qin, P.Z. and Pyle, A.M. (1998) The architectural organization and mechanistic function of group II intron structural elements. *Current opinion in structural biology* **8**, 301-308.

Rambo, R.P. and Doudna, J.A. (2004) Assembly of an active group II intron-maturase complex by protein dimerization. *Biochemistry* **43**, 6486-6497.

Rasko, D.A., Altherr, M.R., Han, C.S. and Ravel, J. (2005) Genomics of the *Bacillus cereus* group of organisms. *FEMS Microbiol Rev* **29**, 303-329.

Rasko, D.A., Ravel, J., Økstad, O.A., Helgason, E., Cer, R.Z., Jiang, L., Shores, K.A., Fouts, D.E., Tourasse, N.J., Angiuoli, S.V., Kolonay, J., Nelson, W.C., Kolstø, A.B., Fraser, C.M. and Read, T.D. (2004) The genome sequence of *Bacillus cereus* ATCC 10987 reveals metabolic adaptations and a large plasmid related to *Bacillus anthracis* pXO1. *Nucleic acids research* **32**, 977-988.

Rasko, D.A., Rosovitz, M.J., Økstad, O.A., Fouts, D.E., Jiang, L., Cer, R.Z., Kolstø, A.B., Gill, S.R. and Ravel, J. (2007) Complete sequence analysis of novel plasmids from emetic and periodontal *Bacillus cereus* isolates reveals a common evolutionary history among the *B. cereus*-group plasmids, including *Bacillus anthracis* pXO1. *Journal of bacteriology* **189**, 52-64.

Read, T.D., Peterson, S.N., Tourasse, N.J., Baillie, L.W., Paulsen, I.T., Nelson, K.E., Tettelin, H., Fouts, D.E., Eisen, J.A., Gill, S.R., Holtzapple, E.K., Økstad, O.A., Helgason, E., Rilstone, J., Wu, M., Kolonay, J.F., Beanan, M.J., Dodson, R.J., Brinkac, L.M., Gwinn, M., DeBoy, R.T., Madpu, R., Daugherty, S.C., Durkin, A.S., Haft, D.H., Nelson, W.C., Peterson, J.D., Pop, M., Khouri, H.M., Radune, D., Benton, J.L., Mahamoud, Y., Jiang, L., Hance, I.R., Weidman, J.F., Berry, K.J., Plaut, R.D., Wolf, A.M., Watkins, K.L., Nierman, W.C., Hazen, A., Cline, R., Redmond, C., Thwaite, J.E., White, O., Salzberg, S.L., Thomason,

B., Friedlander, A.M., Koehler, T.M., Hanna, P.C., Kolstø, A.B. and Fraser, C.M. (2003) The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* **423**, 81-86.

Read, T.D., Salzberg, S.L., Pop, M., Shumway, M., Umayam, L., Jiang, L., Holtzapple, E., Busch, J.D., Smith, K.L., Schupp, J.M., Solomon, D., Keim, P. and Fraser, C.M. (2002) Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science (New York, NY* **296**, 2028-2033.

Redder, P., She, Q. and Garrett, R.A. (2001) Non-autonomous mobile elements in the crenarchaeon Sulfolobus solfataricus. *Journal of molecular biology* **306**, 1-6.

Reddy, A., Battisti, L. and Thorne, C.B. (1987) Identification of self-transmissible plasmids in four Bacillus thuringiensis subspecies. *Journal of bacteriology* **169**, 5263-5270.

Robart, A.R., Montgomery, N.K., Smith, K.L. and Zimmerly, S. (2004) Principles of 3' splice site selection and alternative splicing for an unusual group II intron from *Bacillus anthracis*. *RNA (New York, NY* **10**, 854-862.

Robart, A.R., Seo, W. and Zimmerly, S. (2007) Insertion of group II intron retroelements after intrinsic transcriptional terminators. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 6620-6625.

Robart, A.R. and Zimmerly, S. (2005) Group II intron retroelements: function and diversity. *Cytogenet Genome Res* **110**, 589-597.

Roitzsch, M. and Pyle, A.M. (2009) The linear form of a group II intron catalyzes efficient autocatalytic reverse splicing, establishing a potential for mobility. *RNA (New York, NY*.

Ronning, D.R., Guynet, C., Ton-Hoang, B., Perez, Z.N., Ghirlando, R., Chandler, M. and Dyda, F. (2005) Active site sharing and subterminal hairpin recognition in a new class of DNA transposases. *Molecular cell* **20**, 143-154.

Schmidt, A.L. and Anderson, L.M. (2006) Repetitive DNA elements as mediators of genomic change in response to environmental cues. *Biol Rev Camb Philos Soc* **81**, 531-543.

Schmidt, U., Sagebarth, R., Schmelzer, C. and Stahl, U. (1993) Self-splicing of a Podospora anserina group IIA intron in vitro. Effects of 3'-terminal intron alterations on cleavage at the 5' and 3' splice site. *Journal of molecular biology* **231**, 559-568.

Schuch, R., Nelson, D. and Fischetti, V.A. (2002) A bacteriolytic agent that detects and kills Bacillus anthracis. *Nature* **418**, 884-889.

Seetharaman, M., Eldho, N.V., Padgett, R.A. and Dayie, K.T. (2006) Structure of a self-splicing group II intron catalytic effector domain 5: parallels with spliceosomal U6 RNA. *RNA (New York, NY* **12**, 235-247.

Sharp, P.A. (1991) "Five easy pieces". *Science (New York, NY* **254**, 663.

Sigel, R.K., Sashital, D.G., Abramovitz, D.L., Palmer, A.G., Butcher, S.E. and Pyle, A.M. (2004) Solution structure of domain 5 of a group II intron ribozyme reveals a new RNA motif. *Nature structural & molecular biology* **11**, 187-192.

Sigel, R.K., Vaidya, A. and Pyle, A.M. (2000) Metal ion binding sites in a group II intron core. *Nature structural biology* **7**, 1111-1116.

Siguier, P., Filee, J. and Chandler, M. (2006a) Insertion sequences in prokaryotic genomes. *Curr Opin Microbiol* **9**, 526-531.

Siguier, P., Filée, J. and Chandler, M. (2006b) Insertion sequences in prokaryotic genomes. *Curr Opin Microbiol* **9**, 526-531.

Simon, D.M., Clarke, N.A., McNeil, B.A., Johnson, I., Pantuso, D., Dai, L., Chai, D. and Zimmerly, S. (2008) Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *RNA (New York, NY* **14**, 1704-1713.

Singh, N.N. and Lambowitz, A.M. (2001) Interaction of a group II intron ribonucleoprotein endonuclease with its DNA target site investigated by DNA footprinting and modification interference. *Journal of molecular biology* **309**, 361-386.

Singh, R.N., Saldanha, R.J., D'Souza, L.M. and Lambowitz, A.M. (2002) Binding of a group II intron-encoded reverse transcriptase/maturase to its high affinity intron RNA binding site involves sequence-specific recognition and autoregulates translation. *Journal of molecular biology* **318**, 287-303.

Smith, D., Zhong, J., Matsuura, M., Lambowitz, A.M. and Belfort, M. (2005) Recruitment of host functions suggests a repair pathway for late steps in group II intron retrohoming. *Genes & development* **19**, 2477-2487.

Sorokin, A., Candelon, B., Guilloux, K., Galleron, N., Wackerow-Kouzova, N., Ehrlich, S.D., Bourguet, D. and Sanchis, V. (2006) Multiple-locus sequence typing analysis of *Bacillus cereus* and *Bacillus thuringiensis* reveals separate clustering and a distinct population structure of psychrotrophic strains. *Applied and environmental microbiology* **72**, 1569-1578.

Sozhamannan, S., Chute, M.D., McAfee, F.D., Fouts, D.E., Akmal, A., Galloway, D.R., Mateczun, A., Baillie, L.W. and Read, T.D. (2006) The Bacillus anthracis chromosome contains four conserved, excision-proficient, putative prophages. *BMC microbiology* **6**, 34.

Stahley, M.R. and Strobel, S.A. (2006) RNA splicing: group I intron crystal structures reveal the basis of splice site selection and metal ion catalysis. *Current opinion in structural biology* **16**, 319-326.

Stenfors Arnesen, L.P., Fagerlund, A. and Granum, P.E. (2008) From soil to gut: Bacillus cereus and its food poisoning toxins. *FEMS Microbiol Rev* **32**, 579-606.

Stenfors, L.P., Mayr, R., Scherer, S. and Granum, P.E. (2002) Pathogenic potential of fifty *Bacillus weihenstephanensis* strains. *FEMS microbiology letters* **215**, 47-51.

Stoddard, B.L. (2005) Homing endonuclease structure and function. *Q Rev Biophys* **38**, 49-95.

Stromsten, N.J., Benson, S.D., Burnett, R.M., Bamford, D.H. and Bamford, J.K. (2003) The Bacillus thuringiensis linear double-stranded DNA phage Bam35, which is highly similar to the Bacillus cereus linear plasmid pBClin15, has a prophage state. *Journal of bacteriology* **185**, 6985-6989.

Thomas, C.M. and Nielsen, K.M. (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* **3**, 711-721.

Tobes, R. and Ramos, J.L. (2005) REP code: defining bacterial identity in extragenic space. *Environ Microbiol* **7**, 225-228.

Ton-Hoang, B., Guynet, C., Ronning, D.R., Cointin-Marty, B., Dyda, F. and Chandler, M. (2005) Transposition of ISHp608, member of an unusual family of bacterial insertion sequences. *The EMBO journal* **24**, 3325-3338.

Toor, N., Hausner, G. and Zimmerly, S. (2001) Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA (New York, NY* **7**, 1142-1152.

Toor, N., Keating, K.S. and Pyle, A.M. (2009) Structural insights into RNA splicing. *Current opinion in structural biology*.

Toor, N., Keating, K.S., Taylor, S.D. and Pyle, A.M. (2008a) Crystal structure of a self-spliced group II intron. *Science (New York, NY* **320**, 77-82.

Toor, N., Rajashankar, K., Keating, K.S. and Pyle, A.M. (2008b) Structural basis for exon recognition by a group II intron. *Nature structural & molecular biology* **15**, 1221-1222.

Toor, N., Robart, A.R., Christianson, J. and 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 research* **34**, 6461-6471.

Toro, N. (2003) Bacteria and Archaea Group II introns: additional mobile genetic elements in the environment. *Environ Microbiol* **5**, 143-151.

Toro, N., Jimenez-Zurdo, J.I. and Garcia-Rodriguez, F.M. (2007) Bacterial group II introns: not just splicing. *FEMS Microbiol Rev* **31**, 342-358.

Tourasse, N.J., Helgason, E., Økstad, O.A., Hegna, I.K. and Kolstø, A.B. (2006) The *Bacillus cereus* group: novel aspects of population structure and genome dynamics. *J Appl Microbiol* **101**, 579-593.

Tourasse, N.J. and Kolstø, A.B. (2008) Survey of group I and group II introns in 29 sequenced genomes of the *Bacillus cereus* group: insights into their spread and evolution. *Nucleic acids research* **36**, 4529-4548.

Tourasse, N.J., Stabell, F.B., Reiter, L. and Kolstø, A.B. (2005) Unusual group II introns in bacteria of the *Bacillus cereus* group. *Journal of bacteriology* **187**, 5437-5451.

Valadkhan, S. (2007) The spliceosome: a ribozyme at heart? *Biological chemistry* **388**, 693-697.

Valadkhan, S., Mohammadi, A., Wachtel, C. and Manley, J.L. (2007) Protein-free spliceosomal snRNAs catalyze a reaction that resembles the first step of splicing. *RNA (New York, NY* **13**, 2300-2311.

Van der Auwera, G. and Mahillon, J. (2005) Tn*XO1*, a germination-associated class II transposon from *Bacillus anthracis*. *Plasmid* **53**, 251-257.

Van der Auwera, G.A., Andrup, L. and 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 Auwera, G.A. and Mahillon, J. (2008) Transcriptional analysis of the conjugative plasmid pAW63 from *Bacillus thuringiensis*. *Plasmid* **60**, 190-199.

Van der Auwera, G.A., Timmery, S., Hoton, F. and Mahillon, J. (2007) Plasmid exchanges among members of the *Bacillus cereus* group in foodstuffs. *Int J Food Microbiol* **113**, 164-172.

Van der Auwera, G.A., Timmery, S. and Mahillon, J. (2008) Self-transfer and mobilisation capabilities of the pXO2-like plasmid pBT9727 from Bacillus thuringiensis subsp. konkukian 97-27. *Plasmid* **59**, 134-138.

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

Verheust, C., Fornelos, N. and Mahillon, J. (2005) GIL16, a new gram-positive tectiviral phage related to the Bacillus thuringiensis GIL01 and the Bacillus cereus pBClin15 elements. *Journal of bacteriology* **187**, 1966-1973.

Verheust, C., Jensen, G. and Mahillon, J. (2003) pGIL01, a linear tectiviral plasmid prophage originating from Bacillus thuringiensis serovar israelensis. *Microbiology (Reading, England)* **149**, 2083-2092.

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

Wagner, A. (2006) Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. *Molecular biology and evolution* **23**, 723-733.

Waldsich, C. and Pyle, A.M. (2007) A folding control element for tertiary collapse of a group II intron ribozyme. *Nature structural & molecular biology* **14**, 37-44.

Wank, H., SanFilippo, J., Singh, R.N., Matsuura, M. and Lambowitz, A.M. (1999) A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. *Molecular cell* **4**, 239-250.

Wilcks, A., Jayaswal, N., Lereclus, D. and Andrup, L. (1998) Characterization of plasmid pAW63, a second self-transmissible plasmid in *Bacillus thuringiensis subsp. kurstaki* HD73. *Microbiology (Reading, England)* **144 (Pt 5)**, 1263-1270.

Woodson, S.A. (2005) Structure and assembly of group I introns. *Current opinion in structural biology* **15**, 324-330.

Yang, J., Zimmerly, S., Perlman, P.S. and Lambowitz, A.M. (1996) Efficient integration of an intron RNA into double-stranded DNA by reverse splicing. *Nature* **381**, 332-335.

Zhang, L. and Doudna, J.A. (2002) Structural insights into group II intron catalysis and branch-site selection. *Science (New York, NY* **295**, 2084-2088.

Zhong, J. and Lambowitz, A.M. (2003) Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. *The EMBO journal* **22**, 4555-4565.

Zimmerly, S., Guo, H., Eskes, R., Yang, J., Perlman, P.S. and Lambowitz, A.M. (1995a) A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. *Cell* **83**, 529-538.

Zimmerly, S., Guo, H., Perlman, P.S. and Lambowitz, A.M. (1995b) Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* **82**, 545-554.

Zimmerly, S., Hausner, G. and Wu, X. (2001) Phylogenetic relationships among group II intron ORFs. *Nucleic acids research* **29**, 1238-1250.

I

II

1612–1623 *Nucleic Acids Research, 2007, Vol. 35, No. 5* *Published online 14 February 2007*
doi:10.1093/nar/gkm031

Group II intron in *Bacillus cereus* has an unusual 3′ extension and splices 56 nucleotides downstream of the predicted site

Fredrik B. Stabell, Nicolas J. Tourasse, Solveig Ravnum and Anne-Brit Kolstø*

Department of Pharmaceutical Biosciences, University of Oslo, Oslo, Norway

Received July 21, 2006; Revised December 20, 2006; Accepted January 5, 2007

ABSTRACT

All group II introns known to date fold into six functional domains. However, we recently identified an intron in *Bacillus cereus* ATCC 10987, *B.c.*I4, that splices 56 nt downstream of the expected 3′ splice site *in vivo* (Tourasse *et al*. 2005, *J. Bacteriol*., 187, 5437–5451). In this study, we confirmed by ribonuclease protection assay that the 56-bp segment is part of the intron RNA molecule, and computational prediction suggests that it might form a stable stem-loop structure downstream of domain VI. The splicing of *B.c.*I4 was further investigated both *in vivo* and *in vitro*. Lariat formation proceeded primarily by branching at the ordinary bulged adenosine in domain VI without affecting the fidelity of splicing. In addition, the splicing efficiency of the wild-type intron was better than that of a mutant construct deleted of the 56-bp 3′ extension. These results indicate that the intron has apparently adapted to the extra segment, possibly through conformational adjustments. The extraordinary group II intron *B.c.*I4 harboring an unprecedented extra 3′ segment constitutes a dramatic example of the flexibility and adaptability of group II introns.

INTRODUCTION

Group II introns are a class of genetic retroelements that are capable of self-splicing and mobility. They are able to excise out of RNA transcripts and to ligate their flanking exons (self-splicing), and excised introns may subsequently insert (reverse splice) into identical intron-less DNA sites (process called homing) or into novel genomic locations (retrotransposition) [see (1–4) for reviews]. Group II introns are found in bacteria, archaea and the organelles of fungi, plants and lower eukaryotes, and they either interrupt genes or are inserted in intergenic regions (5–7). These retroelements typically consist of a catalytic RNA (ribozyme) containing an internal open reading frame (ORF) encoding a multifunctional reverse-transcriptase protein, although ORF-less introns do exist, especially in eukaryotic organelles (5). While some introns have been shown to self-splice *in vitro* in the absence of protein, demonstrating that the splicing reactions are intrinsically catalyzed by the intron RNA, the intronic protein is required for both the splicing and insertion events *in vivo* (1–4).

To date, known group II intron ribozymes normally fold into conserved secondary structures consisting of six domains that are linked by tertiary interactions, where domain V contains a nucleotide triad that is the presumed catalytic center (3,8,9). However, many degenerate introns in eukaryotic organelles lack RNA substructures (4,10). Differences in structure are used to divide the introns into subclasses (9). Typically, the splicing occurs via a two-step transesterification mechanism that requires magnesium (Mg^{2+}) ions as cofactors (2–4). In the first transesterification step, the 2′ hydroxyl (2′-OH) group of a bulged adenosine (A) located in domain VI makes a nucleophilic attack on the 5′ exon–intron junction phosphodiester bond, leading to cleavage of the 5′ exon and formation of a lariat (circle with tail) with 2′–5′ linkage. Due to similarities in this splicing mechanism that involves the formation of a branched lariat RNA molecule and shared structural features group II introns are thought to be the ancestors of the nuclear spliceosomal introns of eukaryotes (11,12). In the second transesterification step, the free 3′-OH of the cleaved 5′ exon attacks the 3′ splice junction, leading to exon ligation and release of the intron lariat. However, an alternative hydrolytic pathway, in which the first splicing reaction is initiated by nucleophilic attack by a water molecule leading to release of a linear intron after transesterification in the second step of splicing has been shown to occur both *in vitro* and *in vivo* (13–15). Group II introns have also been reported

*To whom correspondence should be addressed. Tel: +47 22 85 69 23; Fax: +47 22 84 49 44; Email: a.b.kolsto@farmasi.uio.no

© 2007 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

to be excised as full or partial circles and several pathways have been suggested (13,16).

Splice-site selection is determined by base-pairing interactions between the intron and its flanking exons that help position the splice junctions in the active site of the intron (Figure 1). At the 5′ splice site two 5–6 nucleotide motifs, IBS1 and IBS2 (intron binding sites), located just upstream of the intron insertion site base-pair with the complementary EBS1 and EBS2 (exon binding sites) in intron domain I. Two single base pair interactions guide 3′ splice-site selection, but they differ between major intron classes. For introns of class IIB, the first nucleotide of the 3′ exon (IBS3) pairs with a nucleotide within domain I [EBS3; (2–4,17)]. The second interaction involves the last base of the intron (γ′) and a base between domain II and III (γ) (Figure 1). The location of the 3′ splice-site relative to the base of domain VI is also important. For group IIB introns, the 3′ splice site occurs 3 or 4 nt past domain VI (10,18,19). It should be noted that the correct positioning of the 3′ splice site is actually highly dependent on domain VI, which helps stabilize the intron structure and the catalytic center (3).

We have recently reported a group IIB intron (B2-like class), *B.c.*I4, on the pBc10987 plasmid of *Bacillus cereus*

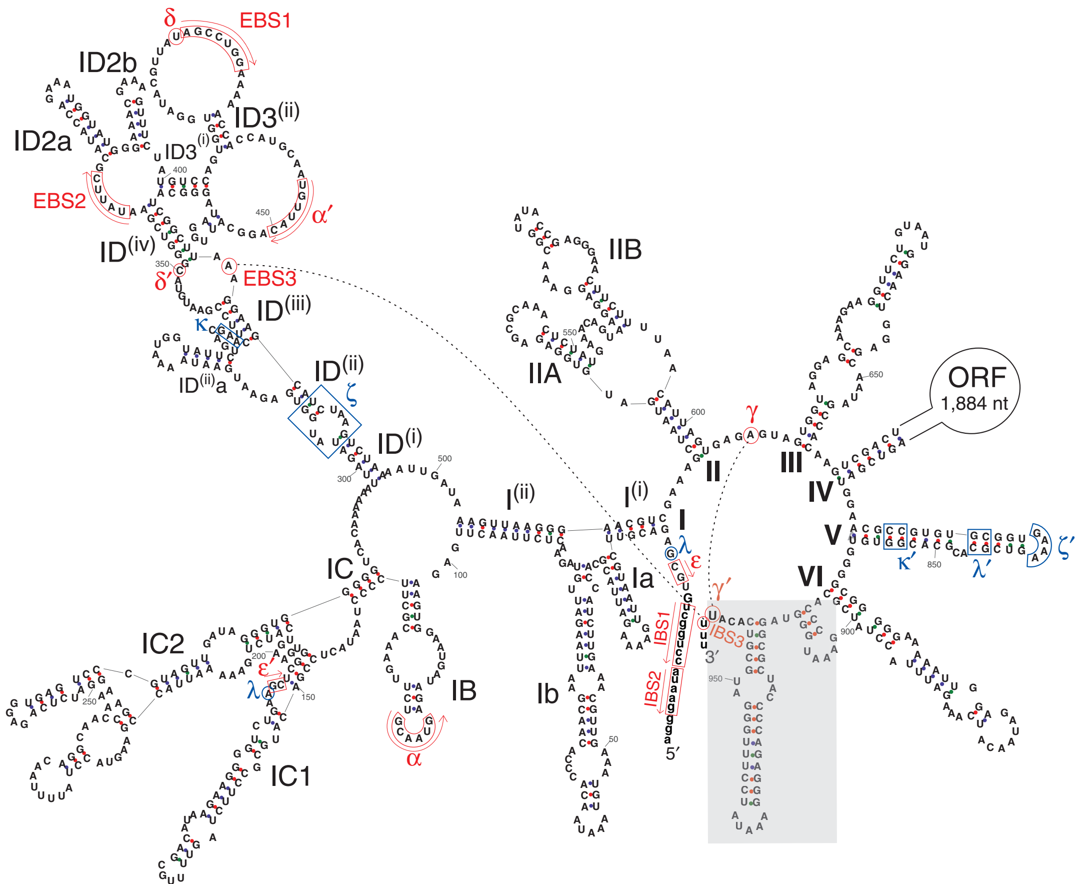


Figure 1. Predicted secondary structure of the *B.c.*I4 group II intron from *B. cereus* ATCC 10987. Exon nucleotides are in lowercase. Roman numerals (I to VI) indicate the six typical functional RNA domains and their subdomains are designated by combinations of letters, numbers and superscripts according to the nomenclature of (10). The extra 56-nt 3′ segment harbored by *B.c.*I4 is boxed in grey. Sites corresponding to consensus positions involved in tertiary interactions (8,9) are indicated by pairs of Greek letters or EBS/IBS (exon binding sites and intron binding sites). Sites of tertiary base-pairing interactions are boxed or circled in red with arrows indicating the orientation of complementarity. Sites implicated in other tertiary contacts are boxed or circled in blue. The IBS3–EBS3 and γ–γ′ base-pair interactions involved in 3′ splice-site selection are indicated by black dotted lines. The δ′ nucleotide was set at the expected location (C:350) according to (17), however it is not complementary to the δ site, while nucleotide A:349 is. ORF, intron-encoded multifunctional open reading frame. Numbering of residues does not include the ORF. A–U, G–C and G–U base pairs are linked by blue, red and green dots, respectively.

ATCC 10987 that splices *in vivo* 56 nt downstream of domain VI and the 3′ splice site that would be expected from secondary structure predictions (19). That is, an extra 56-bp segment downstream of the intron was not present in the ligated exons, as indicated by RT-PCR and sequencing. *B.c.*I4 is inserted within a gene encoding a hypothetical protein with a DNA primase domain (BCEA0033 + BCEA0036), and, interestingly, the unusual splicing puts the ligated exons in-frame and would produce a protein of exactly the same size as the intron-less counterpart in other *B. cereus* group strains. Furthermore, the observed 3′ splice site would also give the correct γ–γ' and IBS3–EBS3 interactions, while non-canonical pairings would form three or four bases downstream of domain VI. In the present study, we have examined the splicing of this extraordinary *B.c.*I4 intron in more detail using *in vivo* and *in vitro* experiments. We show that the extra 56-bp 3′ segment is an integral part of the intron RNA molecule downstream of domain VI, which represents a unique arrangement, and that 3′ splice-site selection can be more flexible than ever seen before, while branching is still maintained at the expected site.

MATERIALS AND METHODS

Secondary structure predictions

The secondary structure of the *B.c.*I4 intron RNA (ORF removed) was computationally predicted by constrained folding using the MFOLD 3.1 package (20) following the consensus structure of group IIB introns from B2-like class (9,19). That is, conserved and identifiable sequence motifs corresponding to the consensus structure were forced during the folding computation.

DNA and RNA isolation

Bacillus cereus ATCC 10987 was grown on Luria Bertani (LB) agar plates at pH 7 and 30°C. An overnight culture (16 h) was inoculated for 3.5 h in 10 ml LB, and then cells were lysed with 10 mg/ml lysozyme. DNA isolation was performed using the Genomic DNA Midi kit (Qiagen) as described by the supplier. Total RNA isolation was conducted as in (19).

PCR and RT-PCR

PCR was performed with Dynazyme II using both the forward and reverse primers at a concentration of 0.4 µM, each deoxynucleoside triphosphate at a concentration of 0.8 mM, and 1 U of Dynazyme (Finnzymes). PCR was primarily run with one denaturation step at 94°C for 5 min, followed by 30–38 cycles of 1 min denaturation at 94°C, 50 s annealing at 58°C and 50 s extension at 72°C, followed by a final extension step of 7 min at 72°C. For RT-PCR, the synthesis of cDNA was initiated with the reverse primers using Superscript III (Invitrogen) and 5 µg of total RNA or 0.1 µg *in vitro* spliced RNA as a template according to the supplier's protocol. A negative control was conducted without addition of reverse transcriptase. A portion of the cDNA and negative control was then amplified by PCR. A complete listing of all the primers used in this study is given in Table 1.

PCR performed with 2.5 U of *Pfu* Turbo DNA polymerase (Stratagene), with same primers and dNTP concentrations as earlier, was run with one denaturation step at 95°C for 2 min, followed by 30 cycles of 30 s denaturation at 95°C, 30 s annealing at 58°C and 1 min/kb extension at 72°C, followed by a final extension step of 10 min at 72°C.

Table 1. List of primers used in the study

Name	Sequence	Location
I4A_left	TCGGATTTTTGCCGTTAGAG	5′ exon
I4A_right	ACCCCTTCTTATGTCGCAAA	intron (5′ end)
I4A_right_nested	CAATCTTAATTCGTTGTGGGTGT	intron (5′ end)
I4B_left	CCTGATGTGATCGGAGGTCT	intron (3′ end)
I4B_left_lariat	GACATTAAACAGTCGATGGAACG	intron (3′ end)
I4B_left_BamHI[a]	**CCCGGATCC**CCTGATGTGATCGGAGGTCT	intron (3′ end)
I4B_right	TGGTTTCGGAATGGAATCAT	3′ exon
I4B_right_T7[a]	***TAATACGACTCACTATAGGGAGA*****GCGGGATCC**TGGTTTCGGAATGGAATCAT	3′ exon
5p_left_BamHI[a]	**CGGGATCCCG**AAATTGTGTTGGAAGATGATAATACG	5′ exon
3p_right_ClaI[a]	**CCATCGATGG**GGTCCAGATTTCACATGAACAG	3′ exon
5p_right	ATTCCATGAGCGATTGAGGT	intron
3p_left	AACCTTTAGATTGAGGAAACACAAA	intron
d3ps_sense[b]	CCTTTGGGATGCGTCACA**AA**TTATGAAATGAAGAAAGGACAAC	3′ splice junction
d3ps_antisense[b]	GTTGTCCTTTCTTCATTTCATAA**TT**TGTGACGCATCCCAAAGG	3′ splice junction
d56_sense[b]	CATCAAAGATTTACCTATCGCA[ ]ATTTTATGAAATGAAGAAAGGAC	3′ splice junction
d56_antisense[b]	GTCCTTTCTTCATTTCATAAAAT[ ]TGCGATAGGTAAATCTTTGATG	3′ splice junction
Restore_sense[b]	CATCAAAGATTTACCTATCGCAA**TT**GAAATGGGTAGGCGCTAC	intron (3′ end)
Restore_antisense[b]	GTAGCGCCTACCCATTTC**AA**TTGCGATAGGTAAATCTTTGATG	intron (3′ end)

[a]Nucleotides in boldface have been added to the primers in order to include restriction sites or promoters. The restriction sites for BamHI or ClaI are underlined. The T7 promoter sequence is in italics.

[b]The sense and antisense primers are complementary to each other and were used to generate the d3ps, d56, and 'Restore' mutant constructs using Quikchange II (Stratagene). The mutated nucleotides are in bold. For the d56 mutant, the deletion point is indicated by brackets.

Radioactive RT-PCR assay

The radioactive RT-PCR assay was mainly conducted as described by Robart *et al.* (21). Primer I4B_right was 5′-end-labeled with $\gamma^{[32\text{-}P]}$ATP (3000 Ci/mmol, 10 mCi/ml) using T4 kinase (New England Biolabs). Labeled primer was purified with Nucleotide purification kit (Qiagen). Reverse transcription (cDNA synthesis) and PCR were conducted as described earlier, except that PCR was conducted with ~5000 cpm of labeled primer I4B_right together with unlabeled counterpart (1 pmol) and primer I4A_left (4 pmol) (Table 1). Amplification products were phenol–CIA (25:24:1 phenol:chloroform:isoamyl alcohol) extracted, ethanol-precipitated with 0.3 M NaOAc, pH 5.2 and digested with SnaBI (New England Biolabs) to ensure homogeneous ends, and these samples were again precipitated and eluted in formamide loading buffer (Ambion). Samples were separated on 7 M urea 6% polyacrylamide gel after heating at 90°C for 2 min and cooling on ice.

Cloning and site-directed mutagenesis

RT-PCR products, either taken directly or gel-purified from 1X TAE gel (QIAquick gel extraction Kit, Qiagen), were cloned into TA cloning vector (Invitrogen) and subsequently sequenced.

Plasmid constructs for *in vitro* self-splicing experiments were made by cloning a PCR product covering the entire *B.c.*I4 intron and parts of the flanking exons, amplified with primers 5p_left_BamHI and 3p_right_ClaI (Table 1), into the BamHI site of pBluescript II KS+ (Stratagene), in orientation for T7 transcription. The intron-containing inserts were then amplified by inverse PCR with outward primers, 5p_right and 3p_left, using *Pfu* Turbo in order to remove the ORF encoded in domain IV of *B.c.*I4, and then ligated with T4 ligase (New England Biolabs), giving a wild-type construct containing 96 bp of the 5′ exon, 884 bp of intron and 151 bp of the 3′ exon.

Site-directed mutagenesis to generate point mutation and deletion constructs was performed with Quikchange II (Stratagene) according to the manufacturer's instructions using two complementary oligonucleotides (of ~40 bases) containing the desired mutation(s) (see Table 1 for details). All constructs were verified by sequencing.

In vitro transcription

One microgram of plasmid construct was linearized by XhoI for transcription reactions with 30 U T7 RNA polymerase (Ambion) according to the manufacturer's instructions. For radiolabeled transcripts, transcription was performed using 20 μCi$^{[\alpha-32P]}$ UTP (800 Ci/mmol, Amersham Pharmacia Biosciences), 0.1 mM UTP and 0.5 mM other NTPs. After DNAse treatment, transcripts were gel-purified and resuspended in 10 mM MOPS, pH 7.5. Unlabeled transcripts were synthesized and purified as for labeled transcripts, but with concentrations of 0.5 mM for each NTP and transcripts were gel-purified after visualization with Fluor-coated TLC plates (Ambion).

Ribonuclease protection assay

RNase T1/A protection assay (RPA) was performed using the Ambion RPA III kit following the manufacturer's protocol. The RNA probe was synthesized from a PCR product made using primers I4B_right_T7, containing a T7 promoter and a 9-nt non-homologous sequence, and I4B_left_BamHI (Table 1), creating a 422-nt long product spanning over the 3′ splice junction of intron *B.c.*I4. The probe was synthesized by *in vitro* transcription as described earlier, using 0.2 μg of PCR product as template. RPA reactions were performed using 10–25 μg total RNA and ~60 000 cpm of gel-purified probe as described by the manufacturer. After hybridization and digestion, probe was separated on a denaturing 6% polyacrylamide/7 M urea gel. For imaging, gels were exposed and analyzed using a Molecular Dynamics Storm 860 Phosphorimager.

In vitro self-splicing of ribozyme

In vitro generated transcripts were denatured and refolded using a GenAmp 2700 PCR machine (Applied Biosystems), by incubating the transcripts in 20 mM MOPS, pH 7.6 at 90°C for 1 min, 75°C for 5 min and then cooling to the splicing temperature over 5 min. Intron transcripts were spliced with 50 000 cpm RNA or ~0.1 μg unlabeled transcripts in 40 mM MOPS, pH 7.6, 100 mM $MgCl_2$ and 500 mM $(NH_4)_2SO_4$, NH_4Cl or KCl at the temperatures indicated in the text. Reactions were initiated by adding pre-warmed splicing buffer to the transcript RNA giving a total reaction volume of 40 μl. At each time point 2 μl were taken, quenched with loading buffer (Ambion) and by placing samples on dry ice. Samples were then heated to 95°C and cooled on ice before being separated on a 7 M urea 4% polyacrylamide gel. Gels were then vacuum dried, exposed and analyzed using a Molecular Dynamics Storm 860 Phosphorimager.

For subsequent RT-PCR and sequencing of these splicing products either unlabeled spliced transcripts, purified with Nucleotide purification kit (Qiagen), or labelled spliced transcript species, excised from gels, were used as templates.

For kinetic analysis the intensities of the radioactive bands were quantified using the ImageQuant 5.0 software and corrected by the number of uridines. The relative fraction of unspliced precursor RNA was computed from the intensities of the radioactive bands. Data were fitted to a biphasic exponential kinetic model [Equation (6) in (22)] by the non-linear least squares method using the GNU gretl 1.5.1 software (http://gretl.sourceforge.net/).

Reverse transcriptase primer extension

The I4B_right primer was 5′-end-labeled with 40 μCi of γ-[^{32}P]ATP and 15 U of T4 polynucleotide kinase. Lariat, lariat with 3′ exon and precursor RNA were incubated with 4 μl of 5× first-strand buffer (Invitrogen), 0.5 μl of RNasOUT (40 U/μl), 2 pmol labeled primer at 85°C for 10 min and then transferred to 55°C for 15 min. The reaction mixture was supplemented with 1 μl of 0.1 M DTT, 40 U of SuperScript III (Invitrogen), 1 μl of 10 mM dNTPs. For primer extension reactions with precursor

RNA 1 μl of 4 mM ddCTP was also added. The reaction mixtures with total volume of 20 μl were incubated at 55°C for 25 min, and stopped by heating to 70°C for 15 min. Reaction products were ethanol-precipitated in 0.3 M NaOAc, resuspended in Gel loading buffer II (Ambion), heated to 95°C for 2 min and resolved in 7.5% polyacrylamide 7 M urea gels and visualized using a Storm 860 scanner (Molecular Dynamics).

RESULTS

The extra 56-bp segment downstream of *B.c.*I4 is part of the intron RNA molecule that splices *in vivo*

A ribonuclease protection assay was performed to investigate whether the 56-bp sequence immediately downstream of the *B.c.*I4 intron was part of the excised intron *in vivo* as it was known to be absent from the spliced exons (19). RNase digestion using a riboprobe covering the intron-3′ exon junction, including the 56-bp sequence, hybridized to total RNA from *B. cereus* ATCC 10987 gave two bands: (1) a band migrating at ∼400 nt, which matches the size expected for the unspliced (precursor) RNA (398 nt) and (2) a band at 319 nt that corresponds to the size of the spliced intron with the extra 3′ 56-bp segment (Figure 2A). The assay gave no indication of an excised intron without the extra 3′ segment. To detect if there could be a small amount of ligated exons that include the extra 56 nt a sensitive radioactive RT-PCR assay with total RNA was conducted where spliced exons were amplified with 5′-end-labeled primer I4B_right and unlabeled primer I4A_left (see Methods section). A single band at ∼147 bp was obtained, corresponding to the size of the splice junction sequenced previously with the same primers (19), but no band containing the extra element was detected (Figure 2B). We therefore conclude that the 56-bp sequence 3′ of *B.c.*I4 is part of the spliced intron RNA *in vivo*.

The extra 3′ segment of *B.c.*I4 is predicted to fold into a stable stem-loop structure

Computational secondary structure predictions suggest that the 56-bp element at the 3′ end of *B.c.*I4 would be able to form a long stable stem-loop structure as shown in Figure 3, although the folding of the first part of the element is unsure. If folding is conducted by constraining the structure of domain VI to be as it would be in a typical group IIB intron, then the first 12 nt of the 56-nt segment would fold into a small stem–loop element (Figure 3A), whereas they would partly extend the base of the domain VI stem if no constraint is applied on domain VI (Figure 3B). The overall secondary structure of *B.c.*I4 (excluding the extra segment) displays all the characteristics typical of group IIB introns belonging to the B2-like class (9), in particular two stem–loop structures inserted between subdomains I$^{(i)}$ and I$^{(ii)}$, absence of subdomain IA, a 3-nt linker between domains V and VI and a stretch of 4 bp below the bulged A (A:899; Figure 1 and ref. (19)). Furthermore, all sites involved in tertiary interactions, with the exception of γ′ and IBS3, can be predicted at the expected locations and the motifs match the consensus elements of group IIB2-like introns (Figure 1). Therefore, the structure shown in Figure 3B, which would imply important changes like the disruption of the V–VI linker and an extension of domain VI, might be less likely to form than the structure shown in Figure 3A. Interestingly, no canonical γ–γ′ and IBS3–EBS3 base-pairings could form at the expected

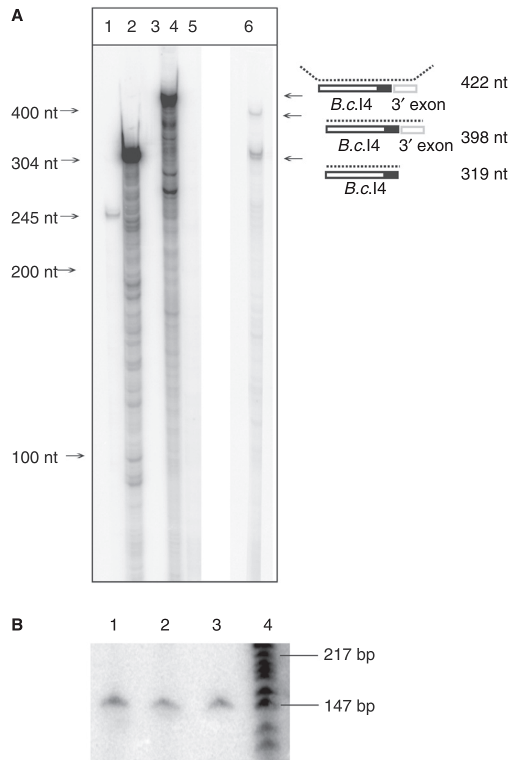
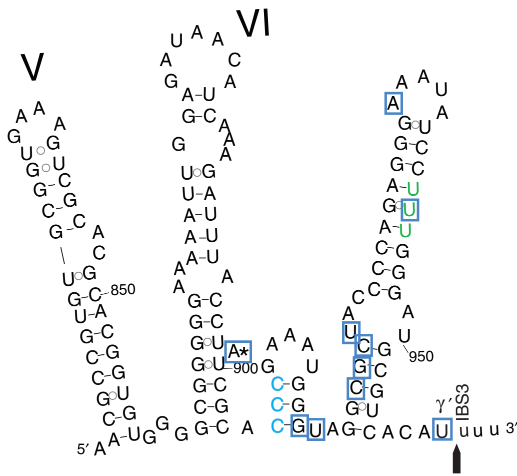
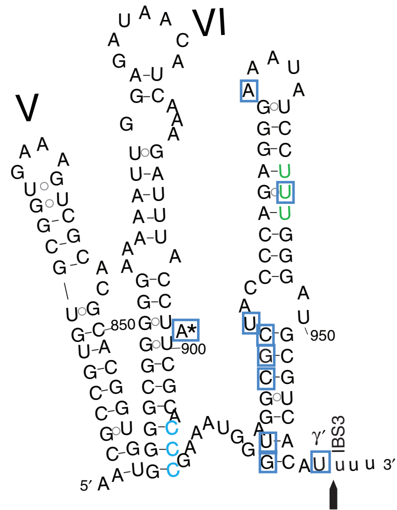
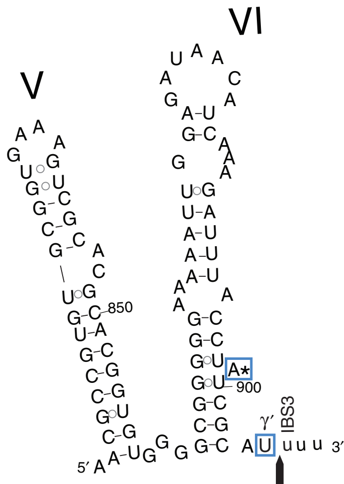


Figure 2. RNase T1/A protection assay (**A**) and radioactive RT-PCR (**B**) showing that the extra 56-nt element 3′ of the *B.c.*I4 intron is part of the intron RNA and not part of the exons. In **A**, lanes 1, 2 and 3 show positive controls based on mouse RNA, and lanes 4, 5 and 6 show the results based on *B. cereus* RNA. Lane 1: digested antisense mouse β-actin RNA probe hybridized with mouse liver RNA; lane 2: same probe as in lane 1, undigested; lane 3: same probe as in lane 1, digested, without mouse liver RNA; lane 4: undigested *B.c.*I4-3′exon junction probe hybridized to *B. cereus* ATCC 10987 total RNA; lane 5: same probe as in lane 4, digested, without RNA sample; lane 6: same probe as in lane 4, digested, with RNA sample. A schematic of the experiment illustrating the location of the probe and the expected products is shown on the right. The black area represents the extra 56-nt element. In **B**, lanes 1, 2 and 3: RT-PCR conducted with exon-specific primers I4B_right (radiolabeled) and I4A_left (Table 1) using as template total RNA sample isolated from *B. cereus* ATCC 10987 at 3, 4 and 6 h of growth, respectively. Lane 4: $\gamma^{[32\text{-}P]}$ATP 5′-end-labeled pBR322 DNA digested with MspI (New England Biolabs), as marker.







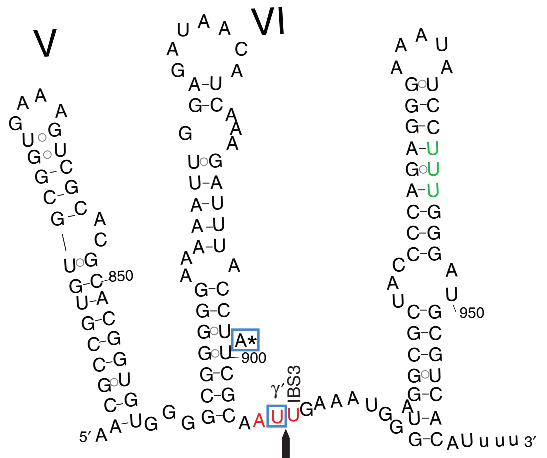


Figure 3. Predicted secondary structure of the 3′ end of the *B.c.*I4 group II intron from *B. cereus* ATCC 10987. Domains V and VI and the extra 56-nt 3′ segment are shown. A and B; two alternative structures for the wild-type intron. **C**; d56 mutant construct, in which the 56-nt element was removed. **D**; 'restore' construct, in which the 3′ splice site was restored at the expected location typical for group II introns, three bases downstream of domain VI. The nucleotides at the expected 3′ splice site in the wild-type intron, CCC, are colored in blue. These bases were mutated to AUU (colored in red) in the 'restore' mutant. The observed 3′ splice junction is indicated by an arrow. Exon residues are in lowercase. The three uridine residues that might create potential 3′ splice sites within the 56-nt 3′ segment, which were not observed to be used, are colored in green. Nucleotides boxed in blue were found to be linked to the intron's 5′ end in inverse RT-PCR experiments. The ordinary branchpoint bulged adenosine (A:899) in domain VI is marked by an asterisk.

splice site three or four bases downstream of domain VI, and both structure predictions suggest that the potential γ′ and IBS3 nucleotides (C:906 and C:907) at this site would already be base-paired and unable to interact with any other nucleotides. The stem–loop structure(s) in the extra segment might enable the 3′ end of the intron to get in closer contact with the rest of the intron core and bring the observed 3′ splice site to a

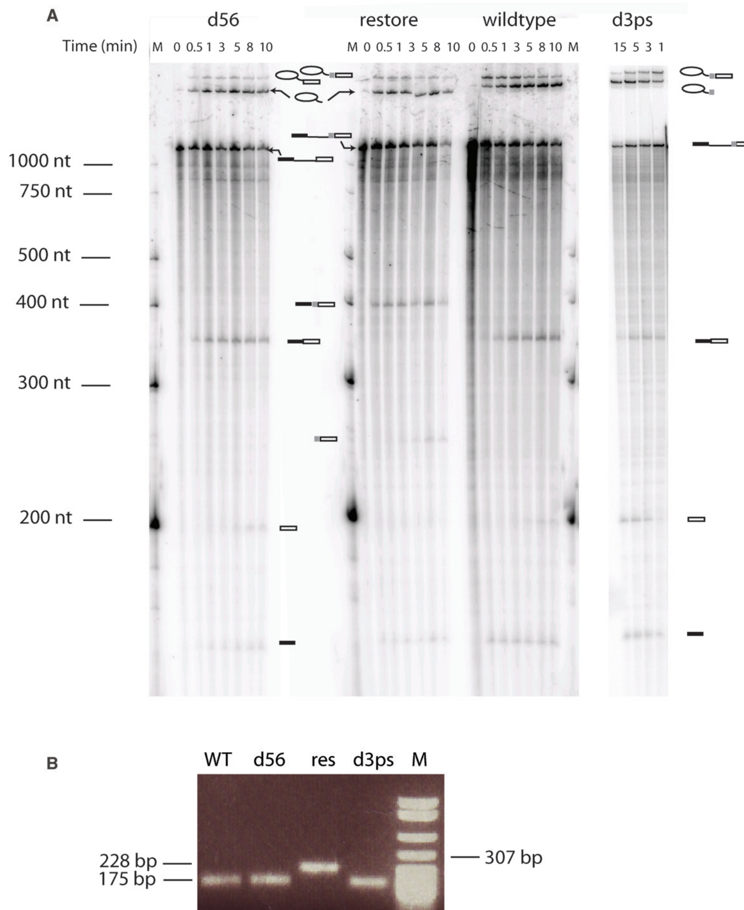


Figure 4. *In vitro* self-splicing (**A**) of *B.c.*I4 wild-type and mutant constructs and subsequent RT-PCR (**B**). In **A**, lane M shows the marker, $\gamma^{[32\text{-}P]}$ATP 5′-end-labeled RNA Century-Plus Marker (Ambion). Splicing was performed in 40 mM MOPS (pH 7.5), 500 mM $(NH_4)_2SO_4$, and 100 mM $MgCl_2$ at 45°C. Samples were separated on a 7 M urea 4% polyacrylamide gel. Schematic drawings are shown next to the bands corresponding to the different splicing products. The light grey box represents the extra 56-nt element. In **B**, RT-PCR with I4B_right and 5p_left_BamH1 primers (Table 1) using *in vitro* splicing products as templates, confirming the size of the ligated exons. Lane M, pBR322 DNA digested with MspI (New England Biolabs), as marker. Samples were separated on a 1% agarose gel.

location close to the site expected for typical group IIB introns.

In vitro splicing analysis of the *B.c.*I4 intron

To study the role or impact of the extra 3′ segment on the splicing of the *B.c.*I4 intron, *in vitro* self-splicing experiments were conducted using wild-type and mutant constructs in the absence of intron-encoded protein. As can be seen in Figure 4A, splicing of the *B.c.*I4 wild-type (WT) intron done with 40 mM MOPS (pH 7.5), 500 mM $(NH_4)_2SO_4$ and 100 mM $MgCl_2$ at 45°C followed by separation of the splicing products on a polyacrylamide gel revealed that the *B.c.*I4 intron could splice out as a true ribozyme. The size of the ligated exon band observed on the gel (Figure 4A) and RT-PCR with primers I4B_right and 5p_left_BamHI (Table 1) and subsequent sequencing confirmed that the 56 nt 3′ of the intron were not part of the ligated exons, verifying that *in vitro* splicing of *B.c.*I4 was the same as *in vivo* (Figure 4B). When the extra 3′ segment was deleted from the intron (d56 construct), while maintaining the last three nucleotides before the 3′ splice site (Figure 3C), the intron could still splice. As expected the spliced exons were of the same sizes as for the wild type, while the size of the lariat was smaller than the wild-type lariat (Figure 4A). The smaller size of the d56 lariat compared to wild type also confirms that the extra element was part of the lariat form of the wild-type intron that had

spliced out. RT-PCR and sequencing confirmed that the ligated exons had the same sequence as in the wild-type case and therefore the intron had spliced at the correct site (Figure 4B). These results indicate that extra 3′ element is not essential for splicing *in vitro*.

We then investigated whether an intron construct carrying the extra 3′ sequence could splice at the location expected for typical group IIB introns by restoring the 3′ splice site three bases downstream of domain VI, i.e. upstream of the extra 3′ segment. In the 'Restore' construct, the nucleotides ACCC immediately after domain VI were changed to AATT, thus providing a potential 3′ splice site that could interact with the γ' and EBS3 nucleotides of *B.c.*I4, without changing the wild-type 3′ splice-site sequence (Figure 3D). Remarkably, the intron was able to splice at the restored site, as judged by the spliced exons being ~50 nt longer in size than the wild-type exons, but no product indicative of splicing at the wild-type site 56 bp downstream of domain VI could be observed (Figure 4A). RT-PCR and sequencing of splicing products using primers located in the exons (I4B_right and 5p_left_BamHI; Table 1) confirmed that ligated exons corresponded to splicing at the restored site (Figure 4B). Interestingly, secondary structure predictions using the restored mutant only produced one possible folding in the 3′ end of *B.c.*I4, whether or not constraints are applied to domain VI. In that structure (Figure 3D), domain VI would adopt the same conformation as in typical group IIB introns, which might explain why the restored site is preferred over the wild-type site. In addition, the possible long linker that may form between the restored site and the stem of the extra 56-bp element may bring the wild-type site too far away from the central core of the intron, thus preventing it from splicing there.

Comparing the splicing of the wild-type *B.c.*I4 intron with different salts such as $(NH_4)_2SO_4$, NH_4Cl and KCl at the same concentration (500 mM) by a 90-min quantitative time-course analysis showed that the splicing efficiency was best with $(NH_4)_2SO_4$. Splicing with this latter salt at various temperatures revealed that the rate of splicing, as judged by the fraction of unreacted precursor RNA, was highest between 47 and 50°C (Figure 5A). While previous results obtained with the d56 mutant construct showed that the *B.c.*I4 intron could splice without the extra 56-bp segment (Figures 3C and 4), a time-course comparison under the presumed optimal conditions revealed that, overall, the d56 mutant intron was not more efficient in splicing than the wild-type *B.c.*I4 (Figure 5B).

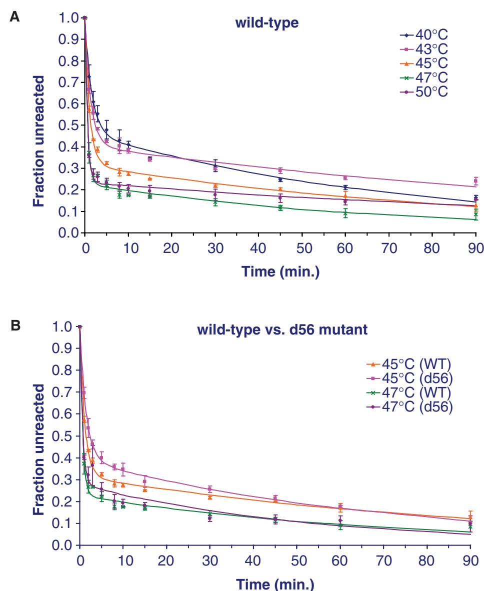


Figure 5. Time-course analysis of *in vitro* self-splicing of *B.c.*I4 wild-type (WT) and d56 mutant constructs. Splicing was performed in buffers containing 100 mM $MgCl_2$ and 500 mM $(NH_4)_2SO_4$. The relative fractions of unspliced precursor RNA were computed from the intensities of the radioactive bands using a phosphorimager. Reactions were repeated 3× for each construct, and are expressed as averages with standard deviations. Data were fitted to a biphasic exponential kinetic model (Equation (6) in (22); rate constant estimates are provided in supplementary Table 1).

3′ splice-site selection

The use of an alternative 3′ splice site by a bacterial group II intron has been shown to occur both *in vivo* and *in vitro* in *B. anthracis*, a close relative of *B. cereus* (21). Interestingly, the 3′ splice junction of *B.c.*I4, which is delimited by two uridines (U) corresponding to the γ' and IBS3 residues, is immediately followed by two more uridines (Figure 3A and B). These latter nucleotides could thus make two possible alternative splice sites where the γ–γ' and EBS3–IBS3 pairings with the γ and EBS3 adenosines would be maintained. However, sequencing of eight clones of RT-PCR products from *in vivo* ligated exons did not reveal any use of either two possible alternative splice sites by the *B.c.*I4 intron, and no alternatively spliced exons could be observed in radioactive RT-PCR assays performed on splicing products from the wild-type RNA both *in vivo* and *in vitro* (Figures 2B and 6). Nevertheless, this absence of detection could be due to tiny amounts of alternative splicing. To investigate further whether use of alternative splice sites could be possible, the two first U residues at the intron-3′ exon junction of the wild-type construct were mutated to A nucleotides. Splicing of this mutant, d3ps (Figure 4A), followed by RT-PCR (Figure 4B), cloning and sequencing of five clones of ligated exon products revealed that, in one case, the *B.c.*I4 intron could splice in between the two uridines at the alternative 3′ site located at position +2 downstream of the wild-type splice site. Very surprising was the fact that the other four clones showed sequences corresponding to splicing at the mutated 3′ splice site, i.e. in between the two adenosines. In this case, no canonical IBS3–EBS3 pairing would form as no U is present in the EBS3 internal loop, while a uridine located in the linker between domain II and III

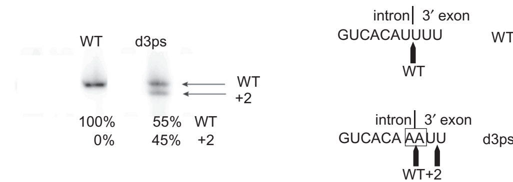


Figure 6. Radioactive RT-PCR assay for detecting the use of alternative 3′ splice sites. The assay was conducted with exon-specific primers I4B_right (radiolabeled) and 5p_left_BamHI (Table 1) with wild-type (WT) and d3ps mutant constructs spliced *in vitro* in 40 mM MOPS (pH 7.5), 500 mM $(NH_4)_2SO_4$ and 100 mM $MgCl_2$ at 45°C. While no alternative splicing was detected for the WT *B.c.*I4 intron, the d3ps mutant could use a 3′ splice site at position +2 downstream of the mutated wild-type site. Quantification of the bands using a phosphorimager is shown, expressed as percentage of total radioactivity. The intron-3′ exon splice junction is shown on the right, with mutated nucleotides boxed.

would be the only nucleotide complementary to γ′ adenosine, the last nucleotide of the d3ps intron (Figure 1). This suggests that either splicing occurs without γ–γ′ and/or IBS3–EBS3 pairings or that non-canonical interactions form. Quantification of the 3′ splice site usage by the d3ps intron using radioactive RT-PCR indicated that the mutated and alternative sites were used in 55 and 45% of the splicing, respectively (Figure 6). It should also be noted that there is a potential 3′ splice site inside the extra element, as three uridines can be found (U:943–U:945; green bases in Figure 3), where no splicing was observed.

Lariat and circle formation

In typical group II introns domain VI contains a bulged A, which is the branchpoint of the lariat, and has been shown to be important for guiding 3′ splice-site selection usually 7–8 nt downstream of the branchpoint (3). Since intron *B.c.*I4 splices 56-nt downstream of domain VI, it raises the question of whether the extra 3′ element could have an impact on the branchpoint selection.

*B.c.*I4 harbors a bulged A at the expected location in domain VI (A:899). To investigate lariat formation *in vivo* inverse nested RT-PCR experiments were performed using outward-directed primers I4A_right/I4A_right_nested and I4B_left (Table 1), located in the 5′ and 3′ end of the intron, respectively, and total RNA isolated from *B. cereus* ATCC 10987 (Figure 7A). Gel electrophoresis of products revealed three different bands, of which only the smallest one was clearly visible each time the assay was reproduced (Figure 7B). This band matched the size of a lariat branching at the expected bulged A:899 in domain VI and this was confirmed by cloning and sequencing. In the sequence, the bulged A residue was converted to T, which is a characteristic error due to the reverse-transcriptase which incorporates an A at the cDNA level when it passes through a branched adenosine (15,23). This result shows that *B.c.*I4 can branch at the typical site within domain VI despite the presence of the extra 3′ segment. The sequence of the largest inverse RT-PCR product (Figure 7B, black arrow) corresponded to

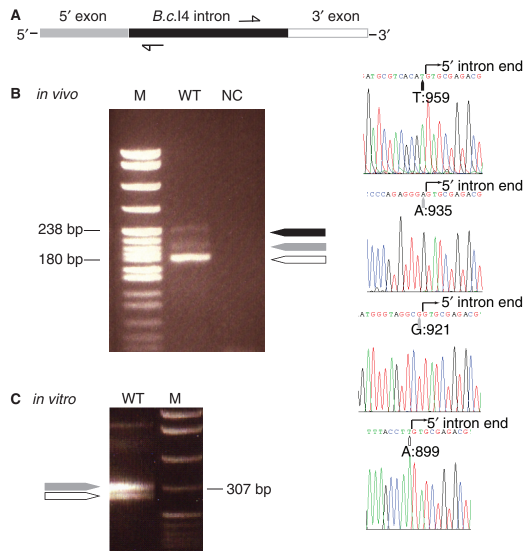


Figure 7. Inverse RT-PCR analysis for detection of *B.c.*I4 lariat and other circular molecules. **A**; schematic drawing illustrating the inverse PCR strategy and the location of primers. **B**; *in vivo* analysis; gel electrophoresis of nested RT-PCR products obtained from total RNA isolated from *B. cereus* ATCC 10987 cultures. **C**; *in vitro* analysis; gel electrophoresis of RT-PCR products obtained from *in vitro* self-splicing of *B.c.*I4 wild-type (WT). Sequencing chromatograms of selected inverse RT-PCR products are shown to the right. The black, grey and white arrows indicate full-length circles, molecules linked within the extra 56-nt segment, and lariats branched in the bulged A:899, respectively. In **B** and **C**, lane M shows the size marker, pBR322 DNA digested with MspI (New England Biolabs); in **B**, lane NC shows a negative control done without reverse-transcriptase. Primers I4A_right/I4A_right_nested and I4B_left were used in **B**; I4A_right and I4B_left_lariat were used in **C** (Table 1). Samples were separated on a 3.5% NuSieve GTG agarose gel (Cambrex).

molecules in which the first and last intron nucleotides were linked, potentially representing full-length intron circles, while in the middle-size product (Figure 7B, grey arrow) the first intron nucleotide was linked to an internal guanosine residue (G:921; Figure 3A and B) in the extra 3′ element, i.e. the last 38 bases of the intron were missing. In the radioactive RT-PCR experiment shown in Figure 2B, no product corresponding to ligated exons containing the last 38 bases of the intron could be observed, indicating that the G:921 residue was not used as an alternative 3′ splice site. The middle-size product could then represent an alternative lariat form or a partial circle. Therefore, from the earlier results it appears that *B.c.*I4 can generate a variety of heterogeneous products *in vivo*. The band corresponding to lariat branching in the expected bulged A:899 was by far the strongest (Figure 7B, white arrow), indicating that this pathway was the major one *in vivo* under the growth conditions used.

When inverse RT-PCR was conducted on the products of *in vitro* splicing with the WT construct using primers I4A_right and I4B_left_lariat (Table 1), similar results

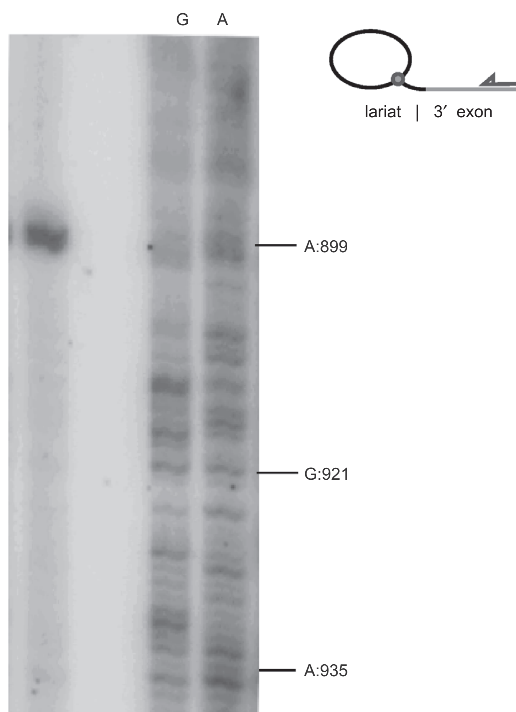


Figure 8. Reverse transcriptase primer extension for detection of *B.c.*I4 lariat RNA. Primer extension was conducted on the 'lariat + 3′ exon' fraction obtained from *in vitro* splicing of the wild-type *B.c.*I4 intron (top band in Figure 4a) using primer I4B_right located in the 3′ exon. The same reaction was performed on precursor RNA using ddCTP and ddTTP to produce the G and A ladder, respectively. A schematic of the experiment is drawn to the right.

were obtained (Figure 7C). The three mutant introns (d3ps, d56 and restore) also gave rise to clones corresponding to full circles (data not shown). To investigate whether the molecules linked within the 56-bp 3′ segment could be lariats, inverse RT-PCR was run separately on the fractions corresponding to 'lariat' and 'lariat + 3′ exon' for the WT construct (top two bands in Figure 4A). Although most of the clones (8 of 13 in 'lariat' and 12 of 17 in 'lariat + 3′ exon') were lariats branched in A:899, molecules in which the 5′ intron nucleotide was linked to G:921 were found in both fractions, and clones linked to a variety of other sites present in either fraction (Figure 3A and B and 7). Furthermore, a primer extension analysis was conducted on the 'lariat + 3′ exon' fraction (top band in Figure 4A) using primer I4B_right located in the 3′ exon. The strongest stop to reverse-transcription was by far at position A:899 confirming that this position is the main branch site (Figure 8). There were no clear other bands that could be indications of stops representing minor ectopic branching events within the extra 56-bp segment of *B.c.*I4, although this could also be due to a very low frequency of these events, in agreement with the observed frequencies in clones of inverse RT-PCR run on the same fraction.

DISCUSSION

In this article, we have investigated the splicing of the atypical group IIB (B2-like class) intron *B.c.*I4 of *B. cereus* ATCC 10987 which splices, both *in vivo* and *in vitro*, 56 nt downstream of the 3′ site that would be predicted from the classical 6-domain secondary structure of group II introns (19). Here, we present data from three different types of experiments, i.e. ribonuclease protection assay, radioactive RT-PCR and *in vitro* self-splicing, clearly demonstrating that the extra 3′ 56-nt sequence is a part of the intron that splices out and is not part of the ligated exons (Figures 2 and 4). This is further supported by the fact that this 56-nt sequence stretch is not present in the sequence of the orthologous intron-less gene in *B. anthracis* (ORF pXO1-70), *B. cereus* ATCC 43881 and *B. thuringiensis* ATCC 33679 (19,24). This extraordinary arrangement raises the question of how the intron ribozyme can accommodate the extra element and be functional. In this respect, it is remarkable that the predicted secondary structure of *B.c.*I4 (apart from the extra segment) conforms perfectly to the consensus structure of group IIB2-like introns (Figure 1). All motifs involved in tertiary interactions are also conserved and no obvious feature deviating from the B2-like consensus is apparent. This might suggest that *B.c.*I4 has been able to accommodate the extra 3′ segment without major reorganization of the core structure. Use of the downstream 3′ splice site may be facilitated by the formation of a long stable stem–loop structure in the extra 56-nt element that could bring the observed 3′ site closer to the (catalytic) core of the folded intron (Figure 3). The predicted folding of this unusual sequence is supported by the fact that no splicing was observed to occur in the U:943–U:945 stretch, which could make a potential 3′ splice site with IBS3–EBS3 and γ–γ' pairings, as would be expected if these nucleotides are part of a double-stranded stem (green bases in Figure 3A and B). Furthermore, the nucleotides at the expected 3′ splice site are probably paired with other nucleotides either from the 56-nt element or the linker between domains V and VI (blue bases in Figure 3A and B), therefore preventing splicing at the location typical of group II introns. Even though a more precise determination of the secondary structure would require experimental probing, it is likely that the structure downstream of domain VI is important for the intron to be able to splice after the 56-nt long 3′ element. It would also be interesting to determine whether this element is involved in tertiary interactions with the other intron domains and/or the intron-encoded protein.

When the downstream 3′ segment is removed, as in the d56 construct, the intron can still splice *in vitro*, indicating that the extra sequence is not essential for *B.c.*I4 to splice (Figure 4A). On the other hand, the splicing efficiency of the d56 mutant *in vitro* in $(NH_4)_2SO_4$ buffer is not better

than that of the wild-type overall (as judged from fractions of unspliced precursor during *in vitro* time-course analysis in Figure 5), implying that the intron has adapted to function with the extra 3′ segment. Another sign of adaptation is given by the fact that the expected branchpoint (A:899) is still the most predominant both *in vivo* and *in vitro*, and that *B.c.*I4 has a high fidelity in branch-site selection (Figures 7 and 8). It has been extensively demonstrated that branch-site selection occurs with a very high fidelity in typical group II introns (25). Thus, this observed fidelity of *B.c.*I4 is quite remarkable compared to other group II introns in which the bulged A in domain VI is normally located 7–8 nt upstream of the 3′ splice site (3), as opposed to 60 bases.

Mutation of the 3′ splice site (UU→AA) revealed that the *B.c.*I4 intron could use an alternative site 2 nt downstream *in vitro* (Figure 6), indicating that *B.c.*I4 retains some flexibility allowing the selection of an alternative 3′ site. This suggests that the 3′ splice site could come in close proximity to the catatylic center of *B.c.*I4. On the other hand, the d3ps mutant also spliced at the mutated site, where canonical IBS3–EBS3 and/or γ–γ' interactions are not expected to occur, implying that there may be structural constraints limiting somewhat the flexibility of the *B.c.*I4 intron. It could also be that the IBS3–EBS3 and/or γ–γ' interactions are not critical for 3′ splice-site selection by *B.c.*I4, as these interactions support correct 3′ splice-site choice only when the 3′ splice site is located near the active center of the intron (3). Alternative splicing of a bacterial group II intron has been reported in *B. anthracis*, a close relative of *B. cereus* (21). However, no evidence for such an event could be obtained for the wild-type *B.c.*I4 intron both *in vivo* and *in vitro* under the conditions tested (Figures 2B and 6).

Although the *B.c.*I4 intron follows a two-step transesterification splicing reaction using the bulged A in domain VI, results from inverse RT-PCR could indicate that minor products may be formed, as suggested by the detection of molecules in which the 5′ end of the intron was linked to a nucleotide inside the extra 56-nt 3′ element or linked to the 3′ end of the intron. These products were obtained both *in vivo* and *in vitro*, thus indicating that their formation is not due to the action of the intron-encoded protein or host factors (e.g. RNA ligase). Molecules in which the 5′ and 3′ intron ends are joined are suggestive of fully circularized intron RNAs. Alternatively, they could be the result of reverse splicing into precursor RNA (26,27) or template switch by the reverse-transcriptase during cDNA synthesis (28,29) leading to head-to-tail intron tandems. Fully circular group II intron RNAs have been reported for a number of elements from organelles of fungi and plants (13,15,16) and recently in the bacterium *Sinorhizobium meliloti* (30,31), either *in vitro* or *in vivo*, and some of these introns exhibit obvious unusual features in their 3′ ends. If circle formation does occur, it seems to be a minor pathway for the wild-type *B.c.*I4 intron under the *in vivo* and *in vitro* conditions tested in this study. With respect to the molecules that are linked within the extra 56-nt 3′ segment, whether they represent ectopic lariats or small circles is unclear. Circular products lacking 3′ terminal nucleotides of the intron have been observed *in vivo* in plant mitochondria (13) and chloroplasts (15). In the case of *B.c.*I4, this type of molecules does not seem to be a result of splicing or hydrolysis at these ligation points, as the corresponding exonic products were not observed (Figures 2B, 4A and 6). Another alternative is that these molecules are lariats branched within the extra 3′ segment. As shown by Chu *et al.* (25), intron mutants exhibiting ectopic branching still give correct joined exon products. Recently, ectopic branching was suggested to explain different splicing products obtained when mutating the bulged A in the *Lactococcus lactis* LtrB intron (32). Similar events in *B.c.*I4 could be due to structural interference by the 56-nt 3′ element, e.g. when domain VI docks into domain I (33), in a minority of cases. However primer extension assay did not reveal evidence for ectopic branchpoints and we cannot exclude the possibility that these products may be the result of RT-PCR artifacts or structural effects.

The intron characterized in this study, *B.c.*I4, carrying a 3′ extra segment gives a dramatic new example of the flexibility and adaptability of group II introns. *B.c.*I4 has been able to accommodate an additional 56 nt and still remains functional for splicing with high fidelity of branch-site and splice-site selection. This is probably due to some favorable conformational adjustments. The versatility of group II introns can also be seen in the splicing ability of elements lacking the branching A or exhibiting various unusual features in domains V and/or VI, which are still able to splice via alternative pathways like hydrolytic splicing or formation of various circular forms (13,15). This adaptability may have contributed to the survival and maintenance of group II introns in genomes where they have to cope with varying physiological conditions affecting splicing.

Finally, a puzzling question relates to the origin of this extra 56-bp element. How did *B.c.*I4 acquire it and from where? Since the 3′ element is absent from the intron-less gene orthologous to *B.c.*I4's host gene in other *B. cereus* group strains, this element probably originated from another genomic context. A PCR screening of a set of 25 *B. cereus* group strains from our collection using primers specific to *B.c.*I4 and the 3′ exon (BCEA0036) covering the intron-3′ exon junction indicated that similar *B.c.*I4 copies including the 56-nt segment are present in homologous host genes in three other strains from the group (results not shown). Therefore, *B.c.*I4 and its 3′ extra segment are not unique to *B. cereus* ATCC 10987. However, the latter segment does not show similarity to any other sequence in the public databases (both at the nucleotide and amino-acid levels), thus providing no clues about its origin. It would also be interesting to determine whether *B.c.*I4 is capable of mobility and whether the extra 56-nt segment has any impact on the ability to reverse splice into DNA sites. Indeed, the *B.c.*I4-encoded protein contains the endonuclease motif involved in mobility (19).

SUPPLEMENTARY DATA

Supplementary Data is available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by grants from the Norwegian Research Council through the Strategic University Programme (SUP) and the Functional Genomics (FUGE) and Consortium for Advanced Microbial Sciences and Technologies (CAMST) platform. Funding to pay the Open Access publication charge was provided by CAMST.

Conflict of interest statement. None declared.

REFERENCES

1. Bonen,L. and Vogel,J. (2001) The ins and outs of group II introns. *Trends Genet*., **17**, 322–331.
2. Lambowitz,A.M. and Zimmerly,S. (2004) Mobile group II introns. *Annu. Rev. Genet*., **38**, 1–35.
3. Lehmann,K. and Schmidt,U. (2003) Group II introns: structure and catalytic versatility of large natural ribozymes. *Crit. Rev. Biochem. Mol. Biol*., **38**, 249–303.
4. Robart,A.R. and Zimmerly,S. (2005) Group II intron retroelements: function and diversity. *Cytogenet. Genome Res*., **110**, 589–597.
5. Dai,L. and Zimmerly,S. (2002) Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res*., **30**, 1091–1102.
6. Dai,L. and 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.
7. Toro,N. (2003) Bacteria and archaea group II introns: additional mobile genetic elements in the environment. *Environ. Microbiol*., **5**, 143–151.
8. Qin,P.Z. and Pyle,A.M. (1998) The architectural organization and mechanistic function of group II intron structural elements. *Curr. Opin. Struct. Biol*., **8**, 301–308.
9. Toor,N., Hausner,G. and Zimmerly,S. (2001) Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA*, **7**, 1142–1152.
10. Michel,F., Umesono,K. and Ozeki,H. (1989) Comparative and functional anatomy of group II catalytic introns–a review. *Gene*, **82**, 5–30.
11. Seetharaman,M., Eldho,N.V., Padgett,R.A. and Dayie,K.T. (2006) Structure of a self-splicing group II intron catalytic effector domain 5: parallels with spliceosomal U6 RNA. *RNA*, **12**, 235–247.
12. Valadkhan,S. (2005) snRNAs as the catalysts of pre-mRNA splicing. *Curr. Opin. Chem. Biol*., **9**, 603–608.
13. Li-Pook-Than,J. and Bonen,L. (2006) Multiple physical forms of excised group II intron RNAs in wheat mitochondria. *Nucleic Acids Res*., **34**, 2782–2790.
14. Podar,M., Chu,V.T., Pyle,A.M. and Perlman,P.S. (1998a) Group II intron splicing *in vivo* by first-step hydrolysis. *Nature*, **391**, 915–918.
15. Vogel,J. and Börner,T. (2002) Lariat formation and a hydrolytic pathway in plant chloroplast group II intron splicing. *EMBO J*., **21**, 3794–3803.
16. Murray,H.L., Mikheeva,S., Coljee,V.W., Turczyk,B.M., Donahue,W.F., Bar-Shalom,A. and Jarrell,K.A. (2001) Excision of group II introns as circles. *Mol. Cell*., **8**, 201–211.
17. Costa,M., Michel,F. and Westhof,E. (2000) A three-dimensional perspective on exon binding by a group II self-splicing intron. *EMBO J*., **19**, 5007–5018.
18. Dai,L., Toor,N., Olson,R., Keeping,A. and Zimmerly,S. (2003) Database for mobile group II introns. *Nucleic Acids Res*., **31**, 424–426.
19. Tourasse,N.J., Stabell,F.B., Reiter,L. and Kolstø,A.B. (2005) Unusual group II introns in bacteria of the *Bacillus cereus* group. *J. Bacteriol*., **187**, 5437–5451.
20. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*., **31**, 3406–3415.
21. Robart,A.R., Montgomery,N.K., Smith,K.L. and 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.
22. Daniels,D.L., Michels,W.J.Jr and Pyle,A.M. (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.
23. Vogel,J., Hess,W.R. and Börner,T. (1997) Precise branch point mapping and quantification of splicing intermediates. *Nucleic Acids Res*., **25**, 2030–2031.
24. Pannucci,J., Okinaka,R.T., Sabin,R. and Kuske,C.R. (2002) *Bacillus anthracis* pXO1 plasmid sequence conservation among closely related bacterial species. *J. Bacteriol*., **184**, 134–141.
25. Chu,V.T., Adamidi,C., Liu,Q., Perlman,P.S. and Pyle,A.M. (2001) Control of branch-site choice by a group II intron. *EMBO J*., **20**, 6866–6876.
26. Yang,J., Mohr,G., Perlman,P.S. and Lambowitz,A.M. (1998) Group II intron mobility in yeast mitochondria: target DNA-primed reverse transcription activity of aI1 and reverse splicing into DNA transposition sites *in vitro*. *J. Mol. Biol*., **282**, 505–523.
27. Zimmerly,S., Guo,H., Eskes,R., Yang,J., Perlman,P.S. and Lambowitz,A.M. (1995) A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. *Cell*, **83**, 529–538.
28. Sellem,C.H., Begel,O. and Sainsard-Chanet,A. (2000) Recombinant mitochondrial DNA molecules suggest a template switching ability for group-II-intron reverse transcriptase. *Curr. Genet*., **37**, 24–28.
29. Tuschl,T., Sharp,P.A. and Bartel,D.P. (1998) Selection *in vitro* of novel ribozymes from a partially randomized U2 and U6 snRNA library. *EMBO J*., **17**, 2637–2650.
30. Costa,M., Michel,F., Molina-Sanchez,M.D., Martinez-Abarca,F. and Toro,N. (2006) An alternative intron-exon pairing scheme implied by unexpected *in vitro* activities of group II intron RmInt1 from *Sinorhizobium meliloti*. *Biochimie*, **88**, 711–717.
31. Molina-Sanchez,M.D., Martinez-Abarca,F. and Toro,N. (2006) Excision of the *Sinorhizobium meliloti* group II intron RmInt1 as circles *in vivo*. *J. Biol. Chem*., **281**, 28737–28744.
32. Slagter-Jäger,J.G., Allen,G.S., Smith,D., Hahn,I.A., Frank,J. and Belfort,M. (2006) Visualization of a group II intron in the 23 s rRNA of a stable ribosome. *Proc. Natl. Acad. Sci. U.S.A*., **103**, 9838–9843.
33. Hamill,S. and Pyle,A.M. (2006) The receptor for branch-site docking within a group II intron active site. *Mol. Cell*., **23**, 831–840.

III

3202–3214 *Nucleic Acids Research, 2009, Vol. 37, No. 10* *Published online 21 March 2009*
doi:10.1093/nar/gkp186

A conserved 3′ extension in unusual group II introns is important for efficient second-step splicing

Fredrik B. Stabell, Nicolas J. Tourasse and Anne-Brit Kolstø*

Laboratory for Microbial Dynamics (LaMDa), Department of Pharmaceutical Biosciences, University of Oslo, Oslo, Norway

Received November 19, 2008; Revised March 5, 2009; Accepted March 6, 2009

ABSTRACT

The *B.c.*I4 group II intron from *Bacillus cereus* ATCC 10987 harbors an unusual 3′ extension. Here, we report the discovery of four additional group II introns with a similar 3′ extension in *Bacillus thuringiensis kurstaki* 4D1 that splice at analogous positions 53/56 nt downstream of domain VI *in vivo*. Phylogenetic analyses revealed that the introns are only 47–61% identical to each other. Strikingly, they do not form a single evolutionary lineage even though they belong to the same Bacterial B class. The extension of these introns is predicted to form a conserved two-stem–loop structure. Mutational analysis *in vitro* showed that the smaller stem S1 is not critical for self-splicing, whereas the larger stem S2 is important for efficient exon ligation and lariat release in presence of the extension. This study clearly demonstrates that previously reported *B.c.*I4 is not a single example of a specialized intron, but forms a new functional class with an unusual mode that ensures proper positioning of the 3′ splice site.

INTRODUCTION

Group II introns are self-splicing ribozymes that are able to excise themselves from precursor mRNA transcripts. They are also retroelements which encode a multifunctional reverse-transcriptase (RT) open reading frame (ORF) and through reverse-splicing they are able to invade new DNA locations (1–5). They are found in the genomes of bacteria, archaea and eukaryotic organelles. Phylogenetically, group II introns can be divided into several major subfamilies based on RNA secondary structure features and ORF sequences [Figure 3A; (1,5–7)]. The secondary structure of group II intron RNA consists of six domains that are linked by a network of tertiary in-
te-
ractions (2,8–10). In particular, domain I forms the scaffold for intron assembly and domain V is essential for catalysis. The other structural elements are important for compaction, stabilization and/or catalysis. Group II intron splicing proceeds through two transesterification reactions. The first reaction is mediated via nucleophilic attack on the 5′ intron–exon junction either by the 2′ hydroxyl group of the bulged adenosine in domain VI or by water. Subsequently the flanking exons are ligated and a branched intron lariat or a linear intron form is respectively released (1–3).

We previously showed that a group II intron *B.c.*I4, from *Bacillus cereus* ATCC 10987, has unusual properties by splicing 56 nt downstream of the predicted 3′ splice site (11,12). *In vivo* and *in vitro* analyses revealed that this intron harbors a 3′ extension that is a part of the RNA molecule that splices out. In addition, these studies showed that *B.c.*I4 has adapted to splice with the extra element, as the splicing efficiency *in vitro* is slightly better than that of a construct deleted of the 3′ extension, and this extra substructure has been referred to as a domain VII (13). In this study we report four new group II introns, *B.th.*I5, *B.th.*I6(a and b) and *B.th.*I7 from *Bacillus thuringiensis* 4D1 that harbor a 3′ extension similar to that of *B.c.*I4. *Bacillus cereus* and *B. thuringiensis* are genetically very closely related and are members of the *B. cereus* group of bacteria (14,15). The extensions of all these introns form two conserved stem–loop secondary structures and *in vitro* mutagenesis showed that the larger of the two stems is needed for an efficient second-step splicing with the extension.

MATERIAL AND METHODS

Bioinformatic searches

The unusual group II introns *B.th.*I5, *B.th.*I6(a and b) and *B.th.*I7 in *B. thuringiensis kurstaki* 4D1 were identified in preliminary genome sequence data (Økstad,O.A. and Nederbragt,L., University of Oslo, Norway, unpublished

*To whom correspondence should be addressed. Tel: +47 22 85 69 23; Fax: +47 22 84 49 44; Email: a.b.kolsto@farmasi.uio.no

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

data) or in fragments from private sequence data (Papazisi,L. and Peterson,S.N., J. Craig Venter Institute, USA, unpublished data), using BLASTN (16) search with the *B.c.*I4 group II intron and its 56-nt 3′ extension as query. The extensions of *B.th.*I5, *B.th.*I6(a and b) and *B.th.*I7 were also used to search the Genbank and EMBL nucleotide sequence databases. BLASTN was run using two sets of parameters, one using lowered gap penalties (opening cost $G = 2$ and extension cost $E = 1$), and the other using increased reward for nucleotide match (match reward, $r = 2$). Other parameters set to nondefault values were: word size of 7 ($W = 7$), *E*-value of 1 ($e = 1$), and no filtering of low-complexity regions ($F = F$).

For structural searches, domains V and VI of the bacterial group II intron sequences available at the Group II intron database [http://www.fp.ucalgary.ca/group2introns/; (17)] were used as queries to search the Genbank and EMBL databases using BLASTN. The hits were extracted along with 150 bases of downstream flanking sequence. This set of sequences was then scanned using the RNAMotif program (18) with a descriptor representing the 3′ extensions of the unusual introns.

Secondary structure predictions

The secondary structures of the *B.th.*I5, *B.th.*I6 and *B.th.*I7 intron RNAs (ORF removed) were computationally predicted by constrained folding using the MFOLD 3.1 package (19) following the consensus structures of group IIB (B class) introns (6,11). That is, conserved and identifiable sequence motifs corresponding to the consensus structures were forced during the folding computation.

Phylogenetic analysis

Bacterial group II intron sequences were taken from the Group II intron database (17). Amino acid sequences of the ORFs of these introns and the new *B. thuringiensis kurstaki* 4D1 introns were aligned using CLUSTALW (20) followed by manual correction. An unrooted phylogenetic tree of a total of 221 introns was reconstructed using the maximum likelihood method as in ref. 21, based on all the RT domains, by means of the program RAxML 7.0.4 with the amino acid substitution model RtREV + Γ + F (22). After removing ambiguously aligned regions, the alignment contained 221 amino acid sites. Statistical support for the groupings in the tree was assessed using 1000 bootstrap replicates (23). The same procedure was employed to build a tree of B class introns only, except that in this case the full-length intron-encoded ORFs could be aligned (438 amino acid positions).

DNA and RNA isolation

Bacillus thuringiensis kurstaki 4D1 was grown on Luria Bertani (LB) agar plates at pH 7 and 30°C. An overnight culture (16 h) was inoculated for 3.5 h in 10 ml LB, and then cells were lyzed with 10 mg/ml lysozyme. DNA isolation was performed using the Genomic DNA Midi kit (Qiagen) as described by the supplier. Total RNA isolation was conducted as in (12).

PCR and RT-PCR

PCR and RT-PCR were performed as described in (11), with the exception that the annealing temperature was set to 59°C for PCR. A listing of all the primers used in this study is given in Supplementary Table 1.

Cloning and site-directed mutagenesis

RT-PCR products, either taken directly or gel purified from 1× TAE gel (QIAquick gel extraction Kit, Qiagen), were cloned into TA cloning vector (Invitrogen) and subsequently sequenced.

*B.th.*I5 and *B.th.*I6a were cloned into pBluescript KS+ or TA-topo vector pCRII respectively, using primers B.th.I5_right/left and B.th.I6a_right/left, based on sequence fragments of *B. thuringiensis kurstaki* 4D1 and orthologous genes in *Bacillus anthracis* or *B. cereus* ATCC 10987. The intron-containing inserts were then amplified by PCR with outward primers, B.th.I5dORF_right/left and B.th.I6adORF_right/left, in order to remove the ORF encoded in domain IV.

Site-directed mutagenesis to generate point mutation and deletion constructs was performed with Quikchange II (Stratagene) according to the manufacturer's instructions using two complementary oligonucleotides (of ~40 bases) containing the desired mutation(s) with *B.c.*I4, *B.th.*I5 or *B.th.*I6a ΔORF as a template (11). Primers are listed in Supplementary Table 1. All constructs were verified by sequencing.

In vitro transcription

One microgram of plasmid construct was linearized by XhoI for transcription reactions with 30 U T7 or Sp6 RNA polymerase (Ambion) according to the manufacturer's instructions. Transcription and gel-purification of radiolabeled and unlabeled RNA were conducted as previously described (11).

In vitro self-splicing of ribozyme

In vitro generated transcripts were denatured and refolded using a GenAmp 2700 PCR machine (Applied Biosystems), by incubating the transcripts in 10 mM MOPS, pH 7.5 at 90°C for 1 min, 75°C for 5 min, and then slow cooling to the splicing temperature at 47°C. Intron transcripts were spliced with 70 000 c.p.m. RNA or ~0.1 µg unlabeled transcripts in 40 mM MOPS, pH 7.5, 100 mM $MgCl_2$ and 500 mM $(NH_4)_2SO_4$ at 47°C. Reactions were initiated by adding prewarmed splicing buffer to the transcript RNA giving a total reaction volume of 40 µl. At each time point, 2 µl were taken out, quenched with loading buffer (Ambion) and storing samples on dry ice. Samples were then heated to 95°C and cooled on ice, before being separated on a 7 M or 8.5 M Urea 4% polyacrylamide gel. Gels were then vacuum dried, exposed and analyzed using a Molecular Dynamics Storm 860 Phosphorimager.

For subsequent RT-PCR and sequencing of these splicing products, either unlabeled spliced transcripts, purified with Nucleotide purification kit (Qiagen), or labeled

spliced transcript species, excised from gels were used as templates.

For kinetic analysis, the intensities of the radioactive bands were quantified using the ImageQuant 5.0 software and corrected for the number of uridines. The relative fractions of unspliced precursor and free lariat RNA were computed from the intensities of the radioactive bands of all intron-containing products. Data were fitted to a biphasic exponential kinetic model [Equations (6) and (8) in ref. 24] by the nonlinear least squares method using the GNU gretl 1.6.5 software (http://gretl.sourceforge.net/).

Sequence availability

The nucleotide sequences of the *B.th.*I5, *B.th.*I6a, *B.th.*I6b and *B.th.*I7 group II introns have been deposited in the EMBL database under accession numbers FM992108, FM992109, FM992110 and FM992111, respectively.

RESULTS

New group II introns with a 3′ extension in *B. thuringiensis kurstaki* 4D1

Through a sequence similarity search of private sequence collections using BLAST (16) and the 3′ extension of the unusual *B.c.*I4 intron of *B. cereus* ATCC 10987 (11,12) as query, four sequence fragments exhibiting similarity to the *B.c.*I4 extension were identified in *B. thuringiensis kurstaki* BGSC 4D1 (also known as AH248 or KB004). Further cloning, sequencing and computational secondary structure predictions revealed that all of these sequences contained a full group II intron with the six typical domains and a 3′ extension (Figure 1). No additional group II introns with this unusual extension were identified in a similar search of public sequence databases, as well as in a structural search with RNAMotif (18) (see 'Material and Methods' section). Three of the

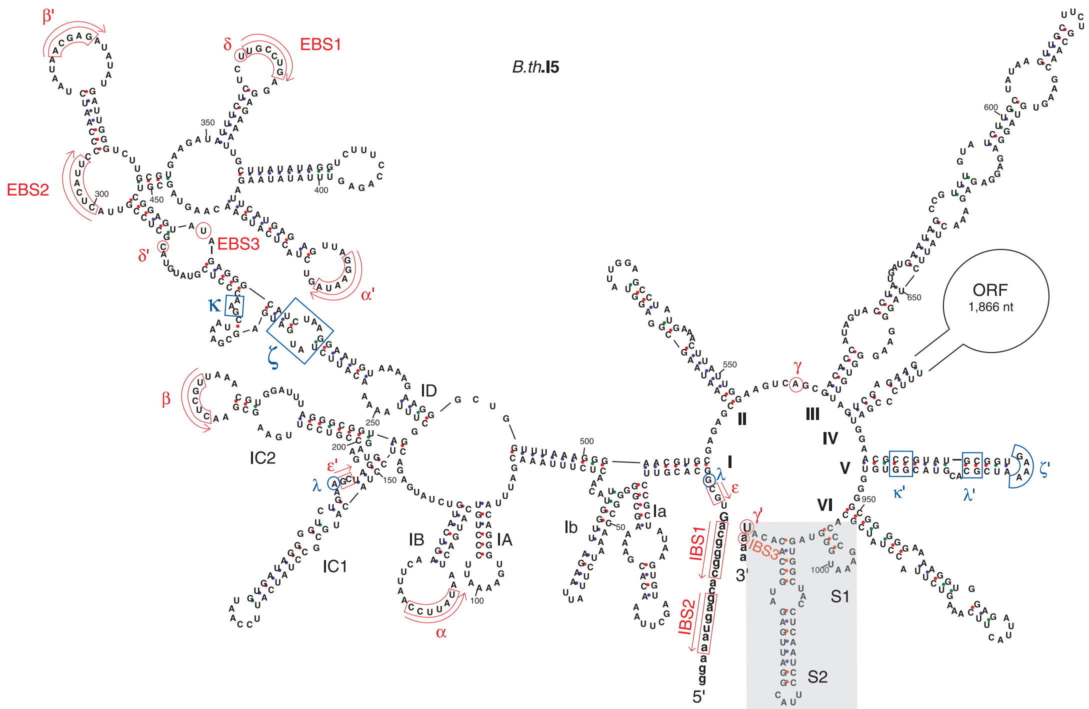


Figure 1. Predicted secondary structures of the *B.th.*I5, *B.th.*I6 and *B.th.*I7 group II introns from *B. thuringiensis kurstaki* 4D1. Exon nucleotides are in lowercase. Roman numerals (I to VI) indicate the six typical functional RNA domains, and subdomains within domain I are designated following the nomenclature of (40). The extra 53/54-nt 3′ segment harbored by the three introns is boxed in gray. Sites corresponding to consensus positions involved in tertiary interactions (6,41) are indicated by pairs of Greek letters or EBS/IBS (exon/intron-binding sites). Sites of tertiary base-pairing interactions are boxed or circled in red with arrows indicating the orientation of complementarity. Sites implicated in other tertiary contacts are boxed or circled in blue. For *B.th.*I5 and *B.th.*I7, the δ′ nucleotide was set at the expected location (C:290 and C:332, respectively) according to (26), however it is not complementary to the δ site, while the adenosine 5′ of δ′ is. Two copies of the *B.th.*I6 intron (*B.th.*I6a and *B.th.*I6b) were found in separate genomic locations and the nucleotide differences in *B.th.*I6b compared to *B.th.*I6a are shown by green boxes and letters. ORF, intron-encoded multifunctional open reading frame. Numbering of residues does not include the ORF. The lengths of the *B.th.*I5, *B.th.*I6a and *B.th.*I7 RNAs (excluding the ORF) are 1044, 936 and 905 nt, respectively (*B.th.*I6b has one extra adenosine compared to *B.th.*I6a). A-U, G-C and G-U base pairs are linked by blue, red and green dots, respectively.

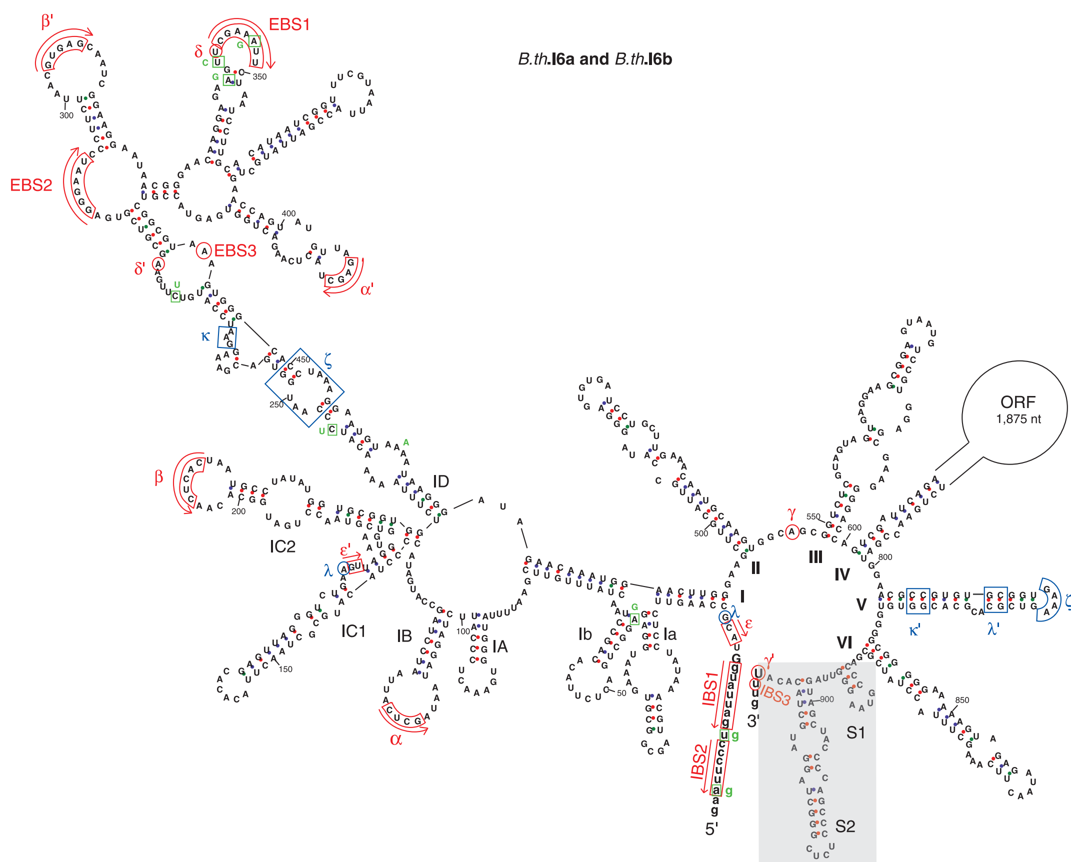


Figure 1. Continued.

B. thuringiensis kurstaki introns were located in genes homologous to genes from the pXO1 plasmid of *B. anthracis* (Table 1). These introns were named (according to the nomenclature used in ref. 25) *B.th.*I5, *B.th.*I6a and b. The latter two are inserted in different genomic contigs and homologs of a predicted conjugation gene (~80% nucleotide sequence identity). *B.th.*I6a and b are 98.4% identical overall and inserted in the same sites thereby representing two copies of the same intron *B.th.*I6. The remaining intron *B.th.*I7 was inserted in a homolog of a hypothetical gene from the pBc10987 plasmid of *B. cereus* ATCC 10987 (Table 1). RT-PCR conducted on total RNA from *B. thuringiensis kurstaki* 4D1 with host gene-specific primers showed that the *B. thuringiensis* introns spliced and thus were functional *in vivo* (Figure 2 and Supplementary Table 1). Sequencing of the RT-PCR products confirmed that the 3′ splice sites of *B.th.*I5, *B.th.*I6 and *B.th.*I7 were located respectively 53, 54 and 54 nucleotides downstream of domain VI, as opposed to the usual three or four bases.

In addition, *in vitro* splicing of *B.th.*I5 and *B.th.*I6a confirmed the splice boundaries observed *in vivo* for these introns (see below). Furthermore, as in the *B.c.*I4 intron, potential EBS3-IBS3 and γ–γ′ base-pair interaction sites, which are important for 3′ splice site selection (2,26,27) could be identified at the observed 3′ splice site (Figure 1). *B.th.*I5, *B.th.*I6 and *B.th.*I7 therefore represent new examples of bacterial group II introns carrying a 3′ extension. Overall, the four different *B. cereus* and *B. thuringiensis* introns are only 47–61% identical at the nucleotide level (31–52% amino acid sequence identity between the ORFs). Phylogenetic analysis of the ORFs of 221 bacterial group II introns available at the group II intron database (17) revealed that these four unusual introns belong to the B class (according to the nomenclature mentioned in ref. 6; Figure 3). However, they do not group in a single lineage but are located in two subgroups herein named α and β. *B.c.*I4 and *B.th.*I7 belong to subgroup α, while *B.th.*I5 and *B.th.*I6 are in subgroup β. The division of the four introns harboring an extra 3′

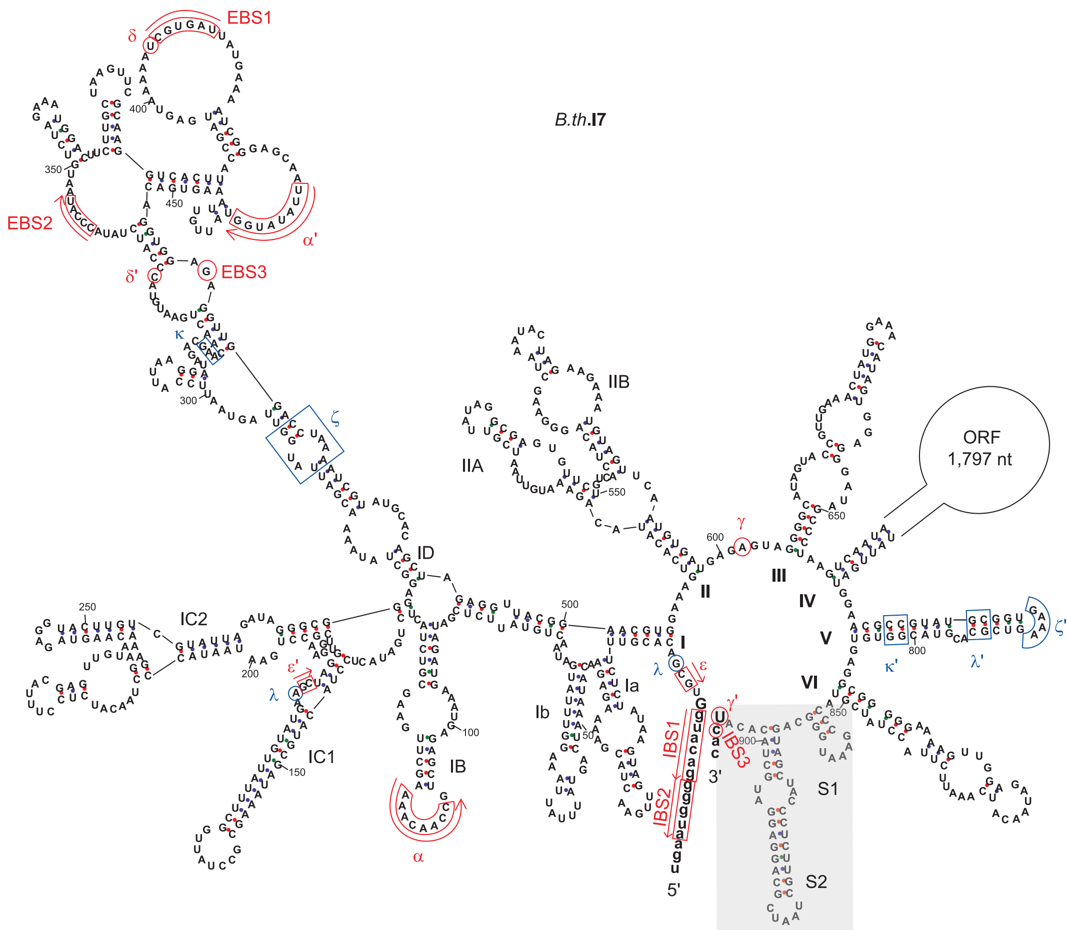


Figure 1. Continued.

segment based on ORF sequence relatedness is supported by the fact that the introns share structural features common to each subgroup (Supplementary Figure 1). These features are all located in domain I of the RNA secondary structure. Despite this divergence, sequence and secondary structure comparisons done both manually and using RNAForester (28,29) revealed that all four introns share several conserved regions (see Supplementary Figures 2 and 3). Besides domains V and VI, which are highly conserved, several nucleotides within the 3′ extra segment are identical (marked in red in Figure 4A). The 3′ extension of the *B. thuringiensis* introns could fold into two stem–loop structures (S1 and S2) similar to those in *B.c.*I4 of *B. cereus*, where the most conserved sites in structure and sequence are within the small stem S1 and the asymmetric internal loop of the longer stem S2. The sequence and folding conservation, together with the occurrence of compensatory mutations in S2, strongly suggest that the 3′ extension forms a stable structure downstream of domain VI for these four unusual introns, and thus might indicate the importance of maintaining this structure for intron function. Remarkably, the S2 internal loop and its surrounding base pairs show a striking resemblance to, and matches the consensus of, the 11-nt tetraloop receptor motif 5′ [CCUAAG…UA UGG] 3′ (30). This is a common RNA motif that participates in the tertiary folding of several catalytic RNAs by interacting with tetraloops of the generic GN_n/RA family (30–32). In addition to the 3′ end, there is a high sequence and structure conservation in the stem of subdomain IC1 in *B.c.*I4, *B.th.*I5 and *B.th.*I6, while *B.th.*I7 shows a lower sequence conservation (see Figures 1 and 4F). Intriguingly, the conserved area is contiguous to the bulged region containing the ε′- and λ sites (z-anchor)

Table 1. Unusual group II introns with 3′ extensions in *B. cereus* and *B. thuringiensis*

Species and strains	Intron name	Intron ORF domains[a]	Intron ORF length (bp/aa)	Intron length (bp)	Intron ORF phylogenetic class[b]	Host gene	Closest intron relative (% aa identity)[c]
B. cereus ATCC 10987	*B.c.*I4	RT-X-En	1884/627	2843	B	BCEA0036 + BCEA0033; hypothetical protein (DNA primase domain)	*Clostridium perfringens* CPE F4969 (BAE79013, 61%)
B. thuringiensis kurstaki 4D1	*B.th.*I5	RT-X-En	1866/621	2910	B	Homolog of pXO1-08 from *B. anthracis*, hypothetical protein with two helicase domains	*Bacillus sp.* EA1 (ABN04186, 40%)
	*B.th.*I6a[d]	RT-X-En	1875/624	2811	B	Homolog of pXO1-42 from *B. anthracis*, conjugation protein of the traG/traD family	*Geobacillus sp.* WCH70 (EDT35839, 41%)
	*B.th.*I6b[d]	RT-X-En	1875/624	2812	B	Homolog of pXO1-42 from *B. anthracis*, conjugation protein of the traG/traD family	*Geobacillus sp.* WCH70 (EDT35839, 41%)
	*B.th.*I7	RT-X-En	1797/598	2702	B	Homolog of BCEA0037 gene from *B. cereus* ATCC 10987, encoding a hypothetical protein.	*B. thuringiensis kurstaki* HD73 (AAZ06578, 47%)

[a]RT, reverse transcriptase domain; X, maturase (splicing) domain; En, endonuclease domain.
[b]According to the classification of Toor *et al.* (6). See also Figure 3.
[c]Genbank accession numbers and amino acid sequence identity between intron ORFs are given in parentheses (top hit of BLAST search of the Genbank database).
[d]*B.th.*I6a and b are overall 98.4% identical at the nucleotide sequence level and represent two copies of the same intron, *B.th.*I6, inserted in different genomic locations.

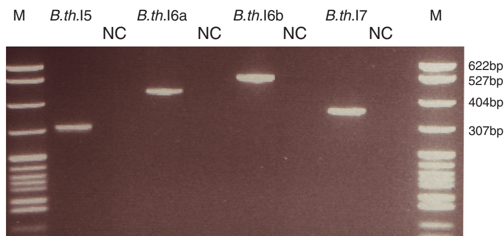


Figure 2. *In vivo* splicing of unusual group II introns in *B. thuringiensis kurstaki* 4D1. RT-PCR was conducted on total RNA with primers located in the flanking exons. The gel picture shows the RT-PCR products of the spliced exons (names of the products related to each intron are given on top) and the corresponding negative controls run without reverse transcriptase (lanes marked with NC). Lane M, pBR322 DNA digested with MspI (New England Biolabs), as marker. Samples were separated on a 2.8% NuSieve GTG agarose gel (Cambrex).

that form interactions with the 5′ end of the intron and/or domain V (9,33).

Mutational analysis of the *B.c.*I4 intron's 3′ extension

Since the extra 3′ element is conserved in structure and partially in sequence between the four *B. cereus* and *B. thuringiensis* introns described here, we conducted an *in vitro* mutational analysis of this element in *B.c.*I4 in order to investigate whether it contributes to the splicing activity of the intron. *In vitro* splicing was conducted under the same conditions as in (11), i.e. at 47°C in 0.5 M ammonium sulfate ($(NH_4)_2SO_4$), 40 mM MOPS, pH = 7.5 and 100 mM $MgCl_2$ (see 'Material and Methods' section). First, two deletion mutants were made by removing separately each of the two stem–loop structures of the 3′ extension (S1 and S2) from the *B.c.*I4 wild-type (WT) ΔORF construct (Figure 4B and C). For the mutant dS2 in which the longer S2 stem was deleted, there was a drastic reduction of the amount of free lariat formed after the second step of splicing, together with a clear increase of the first step intermediate 'lariat with 3′ exon', compared to WT intron (Figure 5, two top bands). A time-course kinetic analysis showed that the fraction of unreacted dS2 precursor RNA was ~20% higher than that of WT, while the relative fraction of lariat released by dS2 was decreased by ~60% on average (Figure 6A and B). Altogether, these results indicate it is mostly the second splicing reaction that is severely slowed down when the S2 stem–loop structure of the 3′ extension is removed from the intron. To support this argument no clear band corresponding to the ligated exons could be observed for this deletion mutant, even though RT-PCR analysis revealed that it did occur, suggesting that the efficiency of exon ligation was decreased. To determine whether the observed phenomenon applies for other introns carrying the extensions, corresponding deletions of S2 were performed on the *B. thuringiensis* *B.th.*I5 and *B.th.*I6a intron constructs. For both introns, a drastic reduction in the efficiency of the second splicing step was also observed compared to the WT, thus pointing to a general importance of S2 for the splicing of the unusual introns

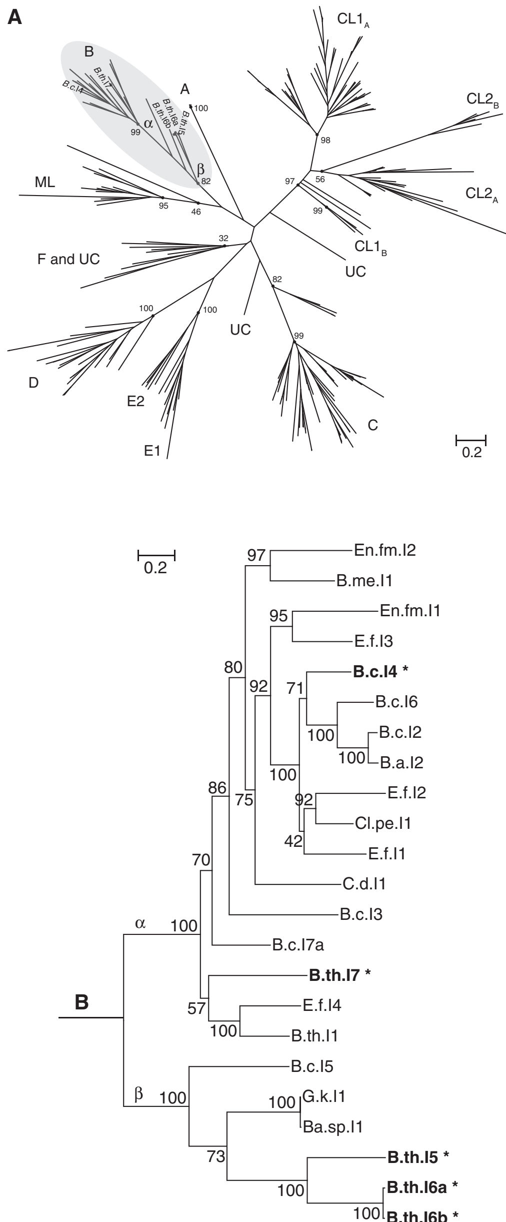


Figure 3. (**A**) Unrooted phylogenetic tree of 221 bacterial group II intron-encoded proteins. The tree was reconstructed using the maximum-likelihood method (RAxML program) and was based on the RT domains of the proteins. The major intron classes are named as in ref. 21: A-F, ML (mitochondrial-like), CL (chloroplast-like) and UC (unclassified). B-class introns are shaded, with the unusual *B. cereus* and *B. thuringiensis* group II introns carrying the 3′ extension indicated by name and a black square. Unlike in the tree as in ref. 21, introns from the $CL2_A$ and $CL2_B$ subclasses are grouped together in the present tree. (**B**) Detailed rooted phylogenetic tree of the B class group II introns, built the same way as in (A), but based on amino acid sequences covering the full length of the intron encoded ORF. The unusual introns are shown in bold and indicated with asterisks. Information, sequences and secondary structure models of all other introns can be found in the Group II intron database [http://www.fp.ucalgary.ca/group2introns/; (17)]. Species names are abbreviated as follows: *Ba.sp*, *Bacillus sp.*; *B.a*, *Bacillus anthracis*; *B.c*, *Bacillus cereus*, *B.me*, *Bacillus megaterium*, *B.th*, *Bacillus thuringiensis*, *C.d*, *Clostridium difficile*, *Cl.pe*, *Clostridium perfringens*, *E.f*, *Enterococcus fæcalis*, *En.fm*, *Enterococcus fæcium*; and *G.k*, *Geobacillus kaustophilus*. In (A) and (B) 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. Proposed subgroupings within the B class are labeled α and β.

(Figures 5 and 6A and B, and Supplementary Figure 4). In sharp contrast to the dS2 mutant, the *B.c.*I4 construct in which the smaller S1 stem–loop structure was deleted, dS1, showed a splicing efficiency equal or better than that of the WT construct with respect to both the amount of precursor processed and lariat formed (Figure 6A and B). Furthermore, mutating the sequence of the terminal loop of S1 (mutant mS1, Figure 4C) revealed no negative effect on either of the two splicing steps. Therefore, the smaller S1 stem–loop part of the 3′ extension does not appear to be critical for splicing under the conditions tested in this study.

To further investigate whether the deficiency in the second step of splicing observed for the dS2 constructs could be due to specific sites within the S2 stem, a number of modifications were made within that stem in *B.c.*I4 (Figure 4B) in order to test for structural or sequence-dependent effects. Mutations of the conserved internal asymmetric loop within S2 were first considered. These included closing the internal loop by re-establishing base pairings (mIntclosed mutant), shuffling the sequence of either one or both sides of the loop while preserving its asymmetry (mutants mIntL, mIntR and mIntL/R), and substituting the conserved C922:G951 bp 5′ to this loop to a A:U pair (mutant m922_951). Unlike the dS2 construct, none of these modifications had a major effect on the second step of splicing and the rate of lariat formation (Figure 6D). They led to slightly higher fractions of unreacted precursor suggesting a less efficient first step of splicing (Figure 6C). In contrast, mispairing C922:G951 by substituting C:922 with A (mutant m922), which would create a larger internal loop, had a drastic effect with respect to the second splicing step comparable to that of the dS2 construct (~30% decrease in free lariat fraction; Figures 5 and 6C). We then addressed the upper part of S2 possessing a 8–9-bp stem with a terminal (top) loop that appears to be conserved in structure by compensatory mutations in the four introns (Figure 4A). Changing the sequence of the terminal loop of S2 from AAAUA to C ACGA (mS2_TL construct; Figure 4B) or shortening the upper stem by 5 bp with or without modifying the internal and terminal loop (dS2_SL and dS2_SS mutants) had little effect on the amount of spliced intron overall, compared to the WT construct (Figure 6C and D), although the rate of the first splicing reaction was

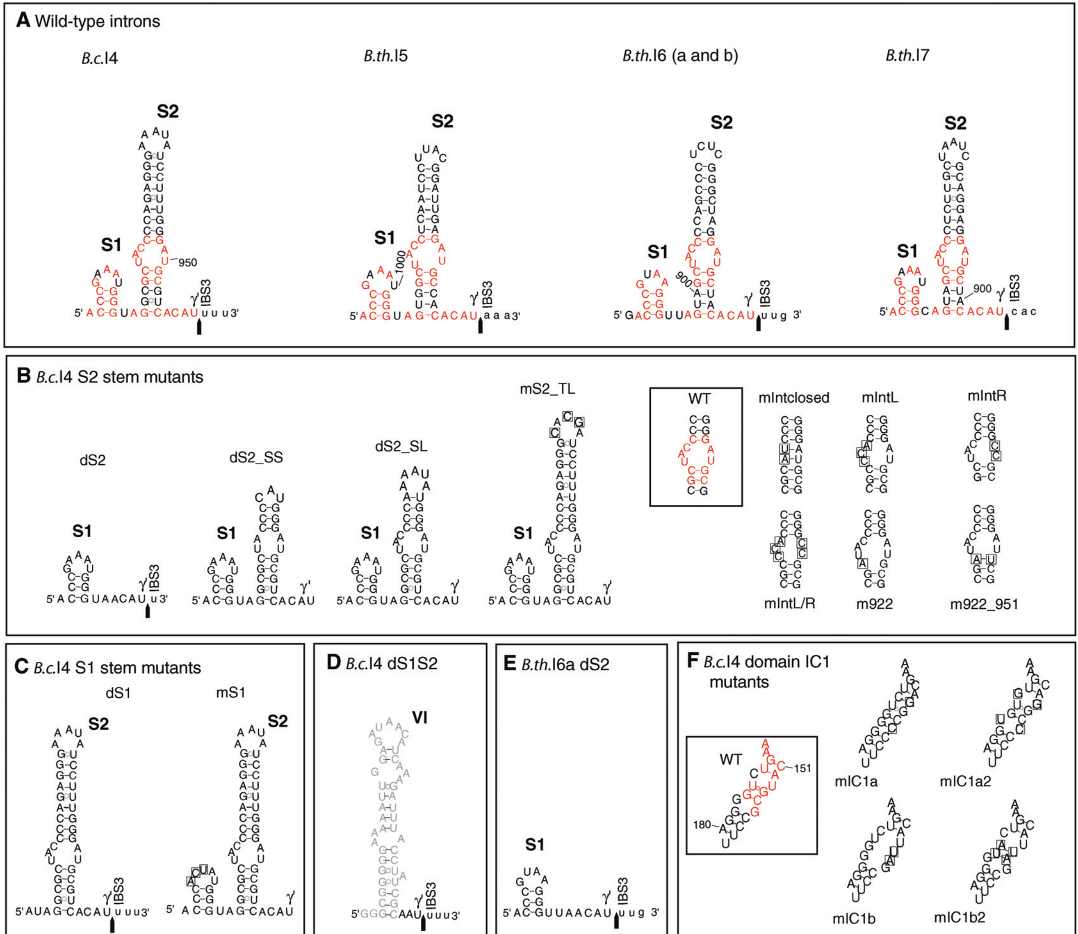


Figure 4. Mutational analysis of the group II introns *B.c.*I4 from *B. cereus* ATCC 10987 and *B.th.*I6a from *B. thuringiensis kurstaki* 4D1. (**A**) Predicted secondary structure of the extra 53/56-nt 3′ segment of *B.c.*I4 and the newly discovered group II introns *B.th.*I5, *B.th.*I6 (a and b) and *B.th.*I7, forming stems S1 and S2. Nucleotides within the 3′ extension that are identical between all four introns are colored in red. The observed 3′ splice junction is indicated by an arrow. Exon residues are in lowercase. (**B**) and (**C**) *B.c.*I4 constructs mutated in the S2 or S1 stem of the 56-nt 3′ extension, respectively. Substituted nucleotides are boxed. (**D**) *B.c.*I4 construct deleted of the whole 3′ extension (previously named d56; see ref. 11). Domain VI is drawn in gray. (**E**) *B.th.*I6a construct deleted of the S2 stem. (**F**) *B.c.*I4 constructs mutated in the IC1 stem of domain I. Nucleotides that are identical between all four introns are colored in red, and substituted nucleotides are boxed. WT, wild-type *B.c.*I4.

reduced for dS2_SL and dS2_SS (Supplementary Table 2). This could indicate that the upper section of the S2 stem is less important for splicing under the conditions tested here.

Mutational analysis of the subdomain IC1 of *B.c.*I4

The area adjacent to the ε′ and λ sites in subdomain IC1 is highly conserved in sequence and structure between the *B.c.*I4, *B.th.*I5 and *B.th.*I6 unusual group II introns, and therefore it was mutated in *B.c.*I4 in order to assess its possible role in intron function and whether there could be a relationship with the extra 3′ segment. There was no clear effect on the efficiency of either splicing steps when base pairing or changing the conserved mispairs U153:C186 and G156:G183 in the IC1 stem (mutants mIC1a or mIC1a2; Figures 4F, 6E and F). These observations, combined with the fact that these nucleotide positions are not conserved in *B.th.*I7 imply that they are less important for splicing. However, when mutating both of the two pairs G154:U185 and C155:G184, which are shared between all four introns

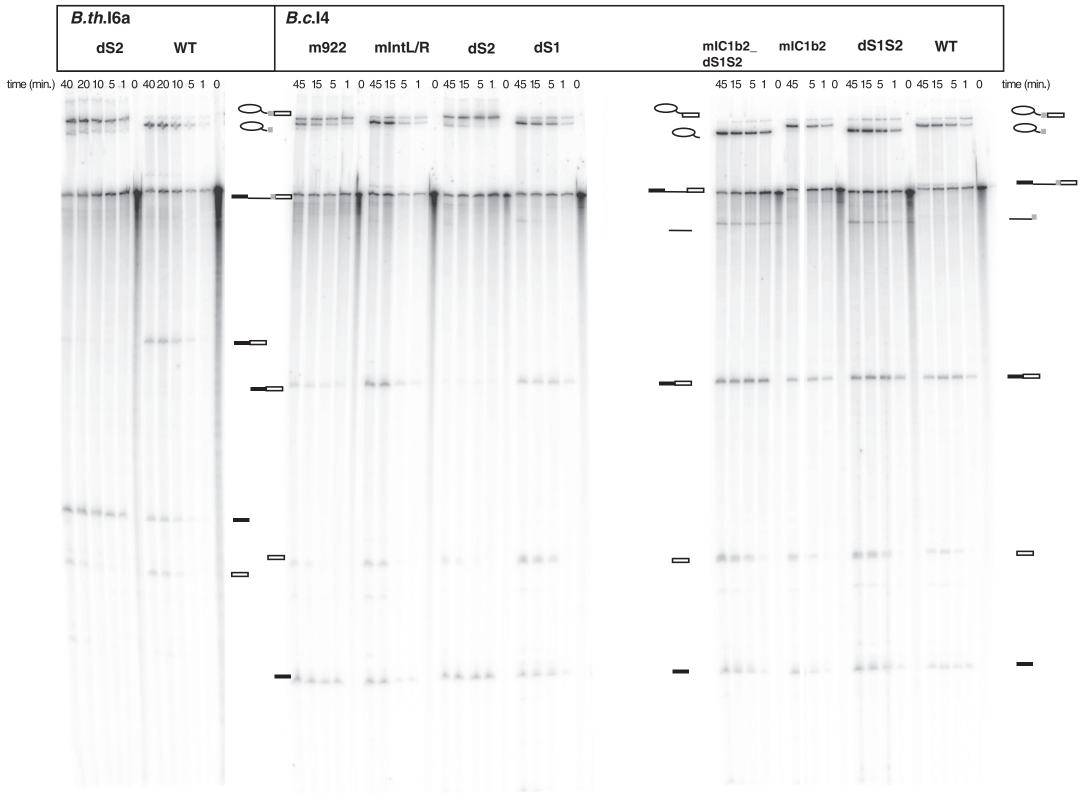


Figure 5. *In vitro* self-splicing of *B.c.*I4 and *B.th.*I6a wild-type (WT) and mutant constructs. Splicing was performed in 40 mM MOPS (pH 7.5), 500 mM $(NH_4)_2SO_4$ and 100 mM $MgCl_2$ at 47°C. Samples were separated on a 7 M urea 4% polyacrylamide gel. Schematic drawings are shown next to the bands corresponding to the different splicing products. The light gray box represents the extra 54/56-nt element. Lariat-containing products, precursor and ligated exons were identified by gel excision, and subsequent RT-PCR and sequencing. The linear intron form and free exons were determined by size. The mutants are described in Figure 4.

into U:A pairs (mutants mIC1b and mIC1b2; Figure 4F), a visible reduction in the first splicing reaction was observed as indicated by the fraction of unreacted precursor RNA, which was ~25% higher than for the WT intron (Figure 6E). The fraction of released lariat was also decreased and this could be a consequence of the slower first splicing step. These results are similar to the results of the mutations within the internal loop of S2, although they are somewhat stronger (compare Figure 6C and E). Interestingly, the reduction in the first splicing reaction was abolished when the extra 56-nt 3′ element was deleted from the mIC1b2 mutant (mIC1b2_dS1S2).

Finally, it should be noted that sequencing of the spliced exons generated by all the mutant introns used in this study confirmed that they all used the same 5′ and 3′ splice sites as the WT construct with the 56-nt extension, indicating that none of the mutations created affected the fidelity of the splicing process.

DISCUSSION

In this study we have discovered four additional group II introns from *B. thuringiensis kurstaki* 4D1 with a 53/54-nt 3′ extension. This extension is similar in structure and partly in sequence to that of the unusual *B.c.*I4 intron identified earlier in *B. cereus* ATCC 10987 (11,12). These four additional introns, named *B.th.*I5, *B.th.*I6a and b and *B.th.*I7, all splice *in vivo* in *B. thuringiensis* (Figure 2), and imply that the unusual 3′ extension may be common to more group II introns than previously thought. Strikingly, all the introns carrying a 3′ extension belong to the same phylogenetic class in terms of ORF sequence (B class); however they do not form a single evolutionary branch within that class (Figure 3). This subgrouping correlates with subtle differences in the RNA secondary structure and is in agreement with the demonstrated coevolution of group II intron structure and ORF (6) (Supplementary Figure 1). Thus, it is not

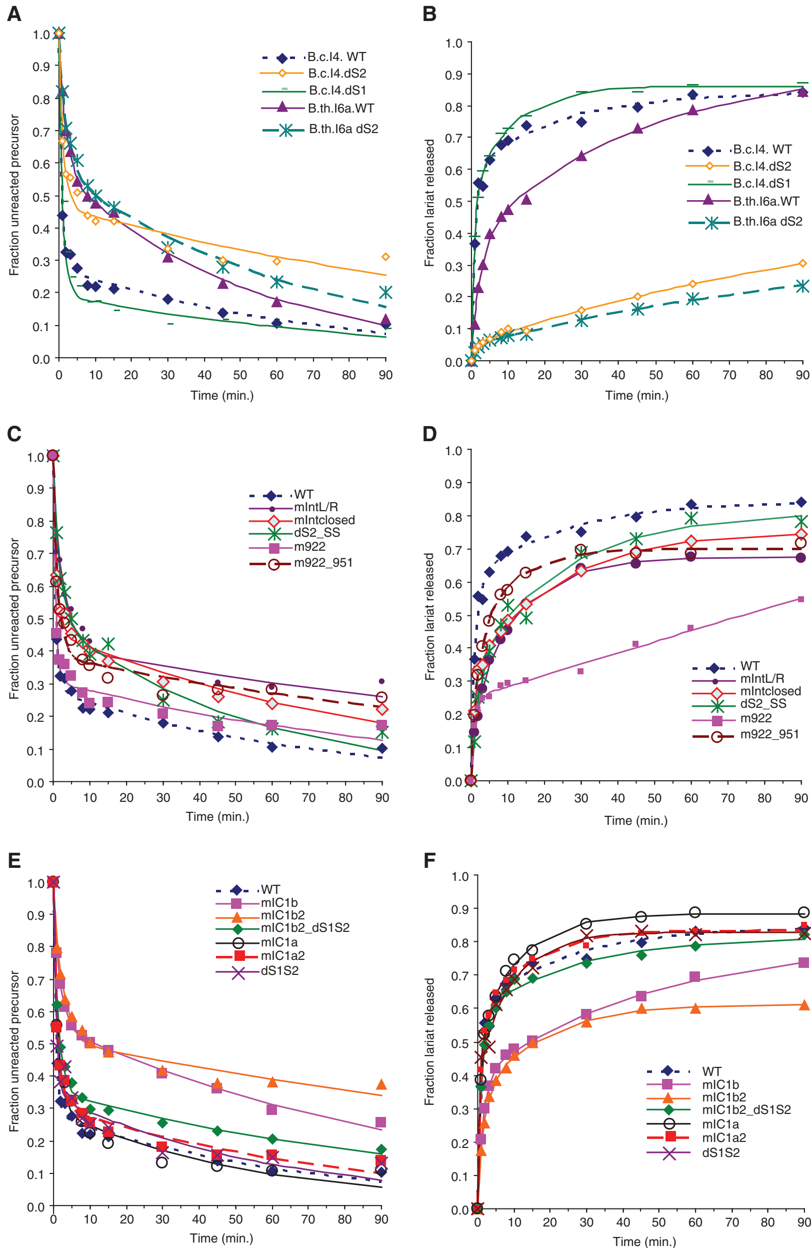


Figure 6. Time-course analysis of *in vitro* self-splicing of *B.c.*I4 and *B.th.*I6a wild-type (WT) and mutant constructs carrying changes in the 3′ extension or in subdomain IC1 (see Figure 4). dS1S2 is a *B.c.*I4 construct lacking the entire 3′ extension (previously named d56; see ref. 11). Splicing was performed in 40 mM MOPS, pH = 7.5, 100 mM $MgCl_2$ and 500 mM $(NH_4)_2SO_4$ at 47°C. The relative fractions of unspliced precursor RNA (**A**, **C** and **E**) and released lariat intron (**B**, **D** and **F**) were computed from the intensities of the radioactive bands using a phosphorimager. Reactions were repeated three times for each construct, and are expressed as averages. Data were fitted to a biphasic exponential kinetic model [Equations (6) and (8) in ref. 24, and rate constants are given in Supplementary Table 2].

clear whether the extension was acquired from a common ancestor. The extra segment might represent a kind of mobile element targeting group II introns. Alternatively, the extension may have been acquired independently by each of these introns, or one of the introns ancestral to the B class obtained the extra segment and spread in different environments and bacterial hosts, and still evolved like the other group II introns in the B class. Indeed, several introns in this class appear to be more flexible in their 3′ splice-site selection, a property that could have enabled some of them to acquire and adapt to this extra segment (21,34). Interestingly, the fact that *B. thuringiensis kurstaki* 4D1 contains representatives of the two different subgroups of the B class (*B.th.*I7 and *B.th.*I5/I6 in α and β subgroup, respectively) suggests that these introns have been acquired separately by the bacterium via different mobility events. Furthermore, the nucleotide sequence identity between the *B.th.*I6a and b introns is significantly higher than the similarity between their homologous host genes (98.4% as opposed to ~80%). Unless this is the result of a duplication event and a strong selective pressure to maintain intron sequence and structure, this observation implies a recent mobility event of *B.th.*I6 with the extra segment within *B. thuringiensis kurstaki* 4D1. It could therefore support the hypothesis of an ancestral origin and spread of a group II intron carrying this 3′ extension. All B class introns known to date are in Gram-positive host bacteria and the unusual introns are only found in the *B. cereus* group of bacteria so far. However, the unusual elements are not typical of the group II introns found in this bacterial group as they only represent a small number of them; 5 out of 81 (35). No similar segment could be identified through sequence and secondary structure search of public databases. It cannot be excluded that additional variants of the extension were not detected by the search procedure used in this study. Therefore, it would be necessary and of great interest to find more examples to determine if this is just sampling bias (as half of the B class introns known are from the *B. cereus* group of bacteria; Figure 3B) or if this is due to a specific property of this type of introns or hosts.

The 53/56-nt 3′ segment is predicted to fold into two stem–loop structures (S1 and S2) and the observation of compensatory substitutions between the *B. cereus* and *B. thuringiensis* introns strongly suggests that the 3′ element forms a stable structure that must be maintained for intron structure and/or activity (Figure 4A). Indeed, deletion of the S2 stem from the *B. cereus B.c.*I4 intron led to an accumulation of the 'lariat with 3′ exon' intermediate and very little exon ligation, i.e. a slower second step of splicing under the conditions tested here (ammonium sulfate splicing buffer; Figures 5 and 6A and B). Similar results were obtained for *B. thuringiensis B.th.*I5 and *B.th.*I6a constructs indicating that this is a general effect. However, the complete deletion of the extension did not affect splicing (11). Therefore, while the 3′ extension is not essential for splicing, the S2 stem is important for maintaining an efficient second step of splicing in presence of the extension.

The conserved asymmetric internal loop of S2, which resembles a tetraloop receptor could be a candidate for facilitating the second splicing step through interaction with other parts of the intron. However, mutational analysis of these nucleotides showed no visible impact on the second step of splicing, but rather a moderate reduction of the first step. Thus, these residues do not appear to be directly responsible for the effect observed when deleting S2. Interestingly, a more pronounced slowdown of the first step also occurred when mutating the two conserved base pairs in subdomain IC1 (mutant mIC1b2), an effect that was abolished when removing the whole 3′ extension (mutant mIC1b2_dS1S2; Figure 6E and F). The faster first step for the latter mutant could be explained by more splicing through the hydrolytic pathway, as constructs without the extension (mIC1b2_dS1S2 and dS1S2) appear to release more linear form of the intron in addition to the lariat (Figure 5). Another possibility is that the two conserved base pairs in IC1 may be involved in accommodating the extension properly into the intron structure via some interactions. The negative effect on the first splicing step observed when mutating the residues within IC1 or in the S2 internal loop may be a consequence of disrupted interaction(s) and/or subtle changes in RNA structure or conformation which may interfere with elements involved in the first step of splicing, e.g. the coordination loop in domain I with the branch point in domain VI, and the *z*-anchor in subdomain IC1 with the 5′ end of the intron (9,33,36–39). The S2 deletion also had an effect on the first step, but the clearly reduced exon ligation and lariat release strongly suggest that this stem is mainly required for efficient 3′ splice site recognition with the extension. This could be due to specific interaction sites other than those investigated in this study, and/or to structural constraints of stem S2. The strong effect of mispairing basepair C922:G951, which is predicted to form a larger internal loop within S2, could point to the latter interpretation. Furthermore, in a *trans*-splicing assay conducted with a *B.c.*I4 construct containing the 5′ exon and domains I to VI and one covering the 3′ extension and the 3′ exon, the ligated exon product was detected by RT-PCR, which may suggest that the 3′ extra segment has interacted with the rest of the intron in a way permitting correct splicing (data not shown). Additional thorough mutagenesis and biochemical experiments will be needed in order to reveal and characterize any interaction partners.

The four new group II introns discovered in this study, which carry a 3′ extension, show that the previously reported *B.c.*I4 is not a single example of a specialized intron, but forms a new functional class with an unusual mode of ensuring proper positioning of the 3′ splice site. All these introns have a conserved two-stem–loop structure at the 3′ end and splice at analogous positions 53/56-nt downstream of domain VI. Mutagenesis showed that the larger stem S2 is important for self-splicing, while the smaller stem S1 is not, and suggests that the S2 stem helps bring the 3′ splice site close to the ribozyme's active site. These findings add support to the proposal of the extension as domain VII (13). A surprising

finding was that the five introns do not form a monophyletic group within class B introns. Therefore, the origin of this extension and why the introns maintain it (as it is not essential for splicing) are open questions. Further work is needed to elucidate how the introns have adapted to the extra segment, which would shed light on the structural and functional evolution of these ribozymes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Leka Papazisi and Scott N. Peterson, J. Craig Venter Institute, USA, and Ole Andreas Økstad and Lex Nederbragt, University of Oslo, Norway, for providing unpublished preliminary genomic sequence data. We are grateful to Murali S. Srinivasan for proofreading the text. We thank anonymous referees for their constructive comments.

FUNDING

The Norwegian Functional Genomics (FUGE II) and the Consortium for Advanced Microbial Sciences and Technologies (CAMST) platform of the Research Council of Norway. Funding for open access charge: Norwegian Functional Genomics (FUGE II) platform of the Research Council of Norway and the University of Oslo.

Conflict of interest statement. None declared.

REFERENCES

1. Lambowitz,A.M. and Zimmerly,S. (2004) Mobile group II introns. *Annu. Rev. Genet*., **38**, 1–35.
2. Lehmann,K. and Schmidt,U. (2003) Group II introns: structure and catalytic versatility of large natural ribozymes. *Crit. Rev. Biochem. Mol. Biol*., **38**, 249–303.
3. Pyle,A.M. and Lambowitz,A.M. (2006) In Gesteland,R.F., Cech,T.R. and Atkins,J.F. (eds), *The RNA World*. 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., Vol. Cold Spring Harbor monograph series; 43, pp. 469–505.
4. Robart,A.R. and Zimmerly,S. (2005) Group II intron retroelements: function and diversity. *Cytogenet. Genome Res*., **110**, 589–597.
5. Toro,N., Jimenez-Zurdo,J.I. and Garcia-Rodriguez,F.M. (2007) Bacterial group II introns: not just splicing. *FEMS Microbiol. Rev*., **31**, 342–358.
6. Toor,N., Hausner,G. and Zimmerly,S. (2001) Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA*, **7**, 1142–1152.
7. Zimmerly,S., Hausner,G. and Wu,X. (2001) Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res*., **29**, 1238–1250.
8. Dai,L., Chai,D., Gu,S.Q., Gabel,J., Noskov,S.Y., Blocker,F.J., Lambowitz,A.M. and 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**, 1–14.
9. Toor,N., Keating,K.S., Taylor,S.D. and Pyle,A.M. (2008) Crystal structure of a self-spliced group II intron. *Science*, **320**, 77–82.
10. Waldsich,C. and Pyle,A.M. (2007) A folding control element for tertiary collapse of a group II intron ribozyme. *Nat. Struct. Mol. Biol*., **14**, 37–44.
11. Stabell,F.B., Tourasse,N.J., Ravnum,S. and Kolstø,A.B. (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.
12. Tourasse,N.J., Stabell,F.B., Reiter,L. and Kolstø,A.B. (2005) Unusual group II introns in bacteria of the *Bacillus cereus* group. *J. Bacteriol*., **187**, 5437–5451.
13. Fedorova,O. and Zingler,N. (2007) Group II introns: structure, folding and splicing mechanism. *Biol. Chem*., **388**, 665–678.
14. Rasko,D.A., Altherr,M.R., Han,C.S. and Ravel,J. (2005) Genomics of the *Bacillus cereus* group of organisms. *FEMS Microbiol. Rev*., **29**, 303–329.
15. Tourasse,N.J., Helgason,E., Økstad,O.A., Hegna,I.K. and Kolstø,A.B. (2006) The *Bacillus cereus* group: novel aspects of population structure and genome dynamics. *J. Appl. Microbiol*., **101**, 579–593.
16. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
17. Dai,L., Toor,N., Olson,R., Keeping,A. and Zimmerly,S. (2003) Database for mobile group II introns. *Nucleic Acids Res*., **31**, 424–426.
18. Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res*., **29**, 4724–4735.
19. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*., **31**, 3406–3415.
20. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
21. Simon,D.M., Clarke,N.A., McNeil,B.A., Johnson,I., Pantuso,D., Dai,L., Chai,D. and Zimmerly,S. (2008) Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *RNA*, **14**, 1704–1713.
22. Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.
23. Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.
24. Daniels,D.L., Michels,W.J. Jr. and Pyle,A.M. (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.
25. Dai,L. and Zimmerly,S. (2002) Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res*., **30**, 1091–1102.
26. Costa,M., Michel,F. and Westhof,E. (2000) A three-dimensional perspective on exon binding by a group II self-splicing intron. *EMBO J*., **19**, 5007–5018.
27. Jacquier,A. and 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.
28. Höchsmann,M., Töller,T., Giegerich,R. and Kurtz,S. (2003) Local similarity in RNA secondary structures. *Proc. IEEE Comput. Soc. Bioinform. Conf*., **2**, 159–168.
29. Höchsmann,M., Voss,B. and Giegerich,R. (2004) Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE ACM Trans. Comput. Biol. Bioinform. IEEE ACM*, **1**, 53–62.
30. Costa,M. and Michel,F. (1995) Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J*., **14**, 1276–1285.
31. Abramovitz,D.L. and Pyle,A.M. (1997) Remarkable morphological variability of a common RNA folding motif: the GNRA tetraloop-receptor interaction. *J. Mol. Biol*., **266**, 493–506.
32. Hendrix,D.K., Brenner,S.E. and Holbrook,S.R. (2005) RNA structural motifs: building blocks of a modular biomolecule. *Q. Rev. Biophys*., **38**, 221–243.
33. De Lencastre,A. and Pyle,A.M. (2008) Three essential and conserved regions of the group II intron are proximal to the 5′-splice site. *RNA*, **14**, 11–24.

34. Robart,A.R., Montgomery,N.K., Smith,K.L. and 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.
35. Tourasse,N.J. and Kolstø,A.B. (2008) Survey of group I and group II introns in 29 sequenced genomes of the *Bacillus cereus* group: insights into their spread and evolution. *Nucleic Acids Res.*, **36**, 4529–4548.
36. Hamill,S. and Pyle,A.M. (2006) The receptor for branch-site docking within a group II intron active site. *Mol. Cell*, **23**, 831–840.
37. Toor,N., Rajashankar,K., Keating,K.S. and Pyle,A.M. (2008) Structural basis for exon recognition by a group II intron. *Nat. Struct. Mol. Biol.*, **15**, 1221–1222.
38. Boudvillain,M., de Lencastre,A. and Pyle,A.M. (2000) A tertiary interaction that links active-site domains to the 5′ splice site of a group II intron. *Nature*, **406**, 315–318.
39. De Lencastre,A., Hamill,S. and Pyle,A.M. (2005) A single active-site region for a group II intron. *Nat. Struct. Mol. Biol.*, **12**, 626–627.
40. Michel,F., Umesono,K. and Ozeki,H. (1989) Comparative and functional anatomy of group II catalytic introns–a review. *Gene*, **82**, 5–30.
41. Qin,P.Z. and Pyle,A.M. (1998) The architectural organization and mechanistic function of group II intron structural elements. *Curr. Opin. Struct. Biol.*, **8**, 301–308.

IV

Group II introns carrying an unusual 3' extension show different splicing properties.

Fredrik B. Stabell, Nicolas J. Tourasse*, and Anne-Brit Kolstø

Laboratory for Microbial Dynamics (LaMDa), Department of Pharmaceutical Biosciences, University of Oslo, Oslo, Norway.

Running title: Splicing pathways of introns with a 3' extension

Keywords: group II introns, unusual extension, splicing, branching, hydrolysis

*To whom correspondence should be addressed. Mailing address: Department of Pharmaceutical Biosciences, University of Oslo, P.O. Box 1068 Blindern, 0316 Oslo, Norway. Phone: (+47) 22 84 49 41. Fax: (+47) 22 84 49 44. e-mail: nicolat@biotek.uio.no.

Abstract

Five group II introns harboring a conserved extension of 53/56 nucleotides at the 3' end have been identified previously in the *Bacillus cereus* group of bacteria. The extra segment forms two stem-loop structures where the largest, S2, has been shown to be important for positioning the 3' exon for efficient second-step splicing with the extension. Here, we show, by *in vitro* mutational and kinetic analysis, that for the *B.c.*I4 intron of *B. cereus* the entire extension also affects the balance between hydrolysis and transesterification in the first step of splicing. Most remarkably, the *B.th.*I6 intron of *B. thuringiensis* is unable to perform an efficient second step when the extension is removed as opposed to *B.c.*I4 which splices equally efficiently whether or not the extension is present. These two introns belong to different evolutionary branches and may have undergone sequence and/or structural changes that led to the different splicing properties.

1 **Dispersal of a group II intron carrying an unusual 3' extension in**

2 **bacteria of the *Bacillus cereus* group**

3

4

5 Nicolas J. Tourasse[§]*, Fredrik B. Stabell[§], and Anne-Brit Kolstø

6

7 Laboratory for Microbial Dynamics (LaMDa), Department of Pharmaceutical

8 Biosciences, University of Oslo, Oslo, Norway.

9

10

11 Running title: Dispersal of introns with a 3' extension

12

13 Keywords: group II intron, unusual extension, pXO1-42, plasmid, mobility, horizontal

14 gene transfer

15

16 [§]These authors contributed equally to the work.

17

18 *To whom correspondence should be addressed. Mailing address: Department of

19 Pharmaceutical Biosciences, University of Oslo, P.O. Box 1068 Blindern, 0316 Oslo,

20 Norway. Phone: (+47) 22 84 49 41. Fax: (+47) 22 84 49 44. e-mail:

21 nicolat@biotek.uio.no.

1 **Abstract**

2 Five unusual group II introns carrying a unique 3’ extension of 53/56 bases have
3 been identified previously in bacteria of the *Bacillus cereus* group. In the present study,
4 we report the identification of additional copies of these atypical elements. Remarkably,
5 the predicted secondary structure of most of these introns is hybrid, exhibiting features of
6 group IIA and IIB introns. Five of the new introns are highly similar to *B.th.*I6, originally
7 discovered in *B. thuringiensis kurstaki* BGSC 4D1, and inserted in the same location in
8 homologues of the pXO1-42 gene from the pXO1 plasmid of *B. anthracis*. From
9 combined sequence comparisons and phylogenetic analyses of introns, host gene, plasmid
10 and chromosome of 43 *B. cereus* group strains, several possible separate events of
11 mobility involving plasmids and *B.th.*I6-like introns could be identified. Altogether this
12 study lends further evidence that the unusual group II introns are mobile with their 3’
13 extension.

14