Examining the evolutionary history and functional roles of Ll.LtrB, a group II intron from *Lactococcus lactis*

Doctoral Thesis

Felix LaRoche-Johnston

Department of Microbiology and Immunology
McGill University, Montreal
December 2020

A thesis submitted to McGill University in partial fulfillment
of the requirements of the degree of Doctor of Philosophy

Table of Contents

Abstract **7**

Résumé **9**

Acknowledgments **12**

Preface **14**

Contributions of authors **17**

Chapter 1: Literature review and objectives of the thesis **18**

1.1: The history of introns: intragenic elements 18

1.2: The architecture of group II introns: catalytic RNAs with a conserved structure 20

1.2.1: Intron secondary structure 21

1.2.2: Diversity of group II introns 23

1.2.3: Intron tertiary structure 25

1.3: Group II introns as self-splicing elements 26

1.3.1: The branching pathway 26

1.3.2: The hydrolytic pathway 28

1.3.3: The circularization pathway 29

1.3.4: *Trans*-splicing 30

1.4: Group II introns as mobile retroelements 31

1.4.1: Retrohoming 33

1.4.2: Retrotransposition 34

1.5: The evolution of group II introns 36

1.5.1: Origin of group II introns 37

1.5.2: Distribution of group II introns 38

1.5.3: Relationship between group II introns and bacterial hosts 40

1.5.4: Group II intron-derived elements in eukaryotes 41
1.5.5: Group II intron-derived elements in bacteria 43
1.6: Ll.LtrB: a model group II intron from *Lactococcus lactis* 44
1.6.1: Ll.LtrB and group II intron lateral transfer 45
1.6.2: Ll.LtrB as a model system to study group II intron circularization 47
1.7: Objectives of the thesis 48
1.8: References 49
1.9: Figures 60
Chapter 2: Recent horizontal transfer, functional adaptation and dissemination of a bacterial group II intron 65
2.1: Preface 65
2.2: Summary 66
2.3: Introduction 67
2.4: Results 70
2.4.1: A clinical isolate of *Enterococcus faecalis* contains a functional group II intron closely related to Ll.LtrB from *Lactococcus lactis* 70
2.4.2: Ll.LtrB and Ef.PcfG can recognize and invade each other's homing sites 71
2.4.3: The significant difference in mobility efficiency at the *ltrB*-HS between Ll.LtrB and Ef.PcfG is due to sequence variations within the introns 73
2.4.4: The variation in mobility efficiency between Ef.PcfG and Ll.LtrB at the *ltrB*-HS is not due to changes in splicing efficiency 73
2.4.5: Some of the point mutations within Ef.PcfG increase its mobility efficiency to the *ltrB*-HS 74
2.4.6: Full-length variants of Ll.LtrB in *L. lactis* share a conserved pattern of mutation acquisition 75
2.5: Discussion 76

2.6: Conclusions 80
2.7: Experimental procedures 80
2.7.1: Bacterial strains and plasmids 80
2.7.3: Two-plasmid intron mobility and patch hybridization assays 81
2.7.4: RNA extraction, RT-PCR, PCR and poisoned primer extension 81
2.7.5: Dendrogram of group introns present in *L. lactis* 82
2.8: Acknowledgments 82
2.9: References 83
2.10: Figures 87
2.11: Tables 93
Chapter 3: Bacterial group II introns generate genetic diversity by circularization and *trans*-splicing from a population of intron-invaded mRNAs 97
3.1: Preface 97
3.2: Summary 98
3.3: Introduction 99
3.4: Results 101
3.4.1: Some excised Ll.LtrB RNA circles harbor mRNA fragments of various lengths at their splice junction 101
3.4.2: IBS1/2-like sequences are present upstream of both extremities of the mRNA fragments incorporated at the Ll.LtrB circle splice junction 101
3.4.3: Ll.LtrB recognizes both extremities of the mRNA fragments present at intron circle splice junctions through base pairing 102
3.4.4: Models of mRNA fragment incorporation at group II intron circle splice junctions 103
3.4.5: Ll.LtrB lariats reverse splice within *L. lactis* mRNAs downstream of IBS1/2-like sequences 104

3.4.6: Ll.LtrB circularization from intron-interrupted mRNAs generates E1-mRNA and mRNA-mRNA chimeras 105
3.5: Discussion 106
3.6: Experimental procedures 109
3.6.1: Bacterial strains and plasmids 109
3.6.2: RNA extraction, PCR and RT-PCR 109
3.6.3: RNA-seq 110
3.6.4: Sequence consensus 110
3.7: Acknowledgments 110
3.8: References 110
3.9: Figures 113
3.10: Supplementary Figures 123
3.11: Tables 128
Chapter 4: Group II introns generate functional chimeric relaxase enzymes with modified specificities through exon shuffling at both the RNA and DNA level 130
4.1: Preface 130
4.2: Summary 131
4.3: Introduction 132
4.4: Results 135
4.4.1: The Ll.LtrB group II intron from *L. lactis* generates mRNA and protein chimeras between orthologous relaxase genes *in vivo* 135
4.4.2: Chimeric relaxase enzymes are active in *L. lactis* and can confer a gain-of-function phenotype 136
4.4.3: An Ef.PcfG-generated chimeric relaxase enzyme supports conjugation in *L. lactis* 138
4.4.4: The formation of chimeric relaxase transcripts occurs under biologically relevant conditions in *E. faecalis* 139

4.4.5: Phylogenetic analyses unveil group II intron-generated chimeric relaxase genes ... 141
4.5: Discussion ... 143
4.6: Experimental procedures ... 148
4.6.1: Bacterial strains and plasmids ... 148
4.6.2: RNA extraction, RT-PCR and RT-qPCR ... 149
4.6.3: Protein extractions and Western blotting ... 150
4.6.4: Mobility and conjugation assays ... 150
4.6.5: Phylogenetic trees ... 150
4.7: Acknowledgments ... 151
4.8: References ... 152
4.9: Figures ... 155
4.10: Tables ... 163
Chapter 5: Conclusions and perspectives ... 166
5.1: Comparing closely related introns to study intron evolution ... 166
5.1:1: Bacterial group II introns face selective pressure to maintain high mobility efficiency ... 167
5.1.2: Conjugation as a means of intron dissemination ... 168
5.1.3: Linking point mutations to functional differences in homologous introns ... 169
5.2: Elucidating a function for group II introns ... 171
5.2.1: Group II introns generate genetic diversity ... 171
5.2.2: Genetic diversity: similarities between group II introns and the spliceosome ... 173
5.3: Does a novel group II intron function lead to a beneficial relationship with its bacterial host? ... 175
5.3.1: Increasing genetic diversity in *L. lactis* by generating chimeric relaxases with gain-of-function phenotypes ... 175
5.3.2: Chimeric relaxases increase the promiscuity of horizontal transfer ... 178

5.3.3: Expansion of group II introns in *L. lactis*: transient colonization or positive selection? 179

5.4: References 182

5.5: Figures 186

Appendix 189

Abstract

Group II introns are a class of large, versatile ribozymes that interrupt genetic loci in bacteria, archaea and the organelles of certain eukaryotes. Following transcription, these ribozymes autocatalytically excise themselves through self-splicing. The Ll.LtrB group II intron was shown to self-splice through two competing pathways: branching, which generates lasso-like intron lariats, and a lesser-known pathway called circularization that generates circular introns. Lariats produced by the branching pathway can re-insert into unoccupied cognate sites or new ectopic genetic loci, thus behaving as mobile retroelements.

From an evolutionary perspective, these ribozymes are the proposed ancestors of over half of the human genome, notably including the abundant nuclear introns and the RNA core of the eukaryotic splicing machinery: the spliceosome. However, the evolution of bacterial group II introns themselves has proven difficult to study. As ribozymes whose activity depends on their secondary structure, they have very low primary sequence conservation, which complicates phylogenetic studies and our understanding of their evolution in bacteria. Moreover, although group II intron-derived nuclear introns play an essential role in eukaryotes by using splicing to increase overall genetic diversity, bacterial group II introns themselves have no common function that benefits their host. Rather, they have always been considered purely as selfish mobile elements that parasitize bacteria, using splicing only as a means of limiting damage to their host. Here, we used the model group II intron Ll.LtrB from the gram-positive bacterium *Lactococcus lactis* to address outstanding questions regarding their evolution and function.

To study group II intron evolution, we compared Ll.LtrB to Ef.PcfG, a group II intron from the gram-positive bacterium *Enterococcus faecalis* recently discovered by our lab. Since these two introns were nearly identical (99.7%) yet present in different bacterial species, they likely represent a recent horizontal transfer event. We thus hypothesized that analyzing the 8 point mutations distinguishing both introns would yield insight into the evolution of group II introns following their entry into a new cellular environment. We compared the mobility and splicing efficiencies of both introns and found that while there was no significant change in splicing efficiency, the 8 point mutations altered mobility efficiency. Ll.LtrB is significantly more efficient at mobilizing to its own cognate site than Ef.PcfG, while both introns recognize more efficiently the *E. faecalis*

homing site, suggesting that it corresponds to the ancestral site. Finally, a dendrogram representing the distribution of the 8 point mutations within all Ll.LtrB variants in *L. lactis* shows their gradual accumulation from *E. faecalis* to *L. lactis*, indicating that a single instance of horizontal gene transfer likely seeded the subsequent dissemination of Ll.LtrB throughout *L. lactis*.

To study intron function, we sought to understand how a subset of introns circles contained additional nucleotides at their circle splice junctions. Since this phenomenon was previously reported in various group II intron subtypes, we hypothesized that it might represent a conserved and hitherto unknown function of group II introns. We demonstrated that the origin of the additional nucleotides was bacterial and plasmid-encoded mRNAs. After base pairing with specific recognition sites and invading mRNAs encountered within the bacterial cell, Ll.LtrB can *trans*-splice either its cognate exon 1 or a foreign nucleophile to the downstream mRNA, generating a chimeric molecule. The intron circles containing additional nucleotides at their splice junction are generated through alternative circularization to an upstream site, thus acting as stable markers of chimera formation. Ll.LtrB can therefore increase the genetic diversity of *L. lactis* by catalyzing the formation of shuffled mRNA molecules.

To determine the biological relevance of intron-generated chimeric mRNAs, we co-expressed the cognate *ltrB* relaxase gene interrupted by Ll.LtrB with an orthologous gene called *pcfG* and demonstrated that chimeric mRNAs and proteins were generated between *pcfG* and *ltrB*. We showed that the abundance of relaxase chimeras correlated with intron copy number and that chimeric relaxases can exhibit gain-of-function phenotypes where their efficiency surpasses the WT relaxases. Since relaxases are involved in horizontal gene transfer by conjugation, the ability of group II introns to increase genetic diversity by forming chimeric relaxases may thus have played an important role in the rapid dissemination of group II introns throughout *L. lactis*.

Overall, our data experimentally demonstrate that group II introns behave and evolve mostly as mobile elements in bacteria, rather than as splicing elements. Our work furthermore reveals that bacterial group II introns can increase the genetic diversity of their host, an ability that likely emerged over the course of evolution from otherwise selfish behavior. Their capacity to generate novel proteins that functionally benefit their host may thus partly explain how these versatile retroelements have been conserved in bacteria throughout evolution.

Résumé

Les introns de groupe II sont une classe de larges ribozymes qui interrompent certains loci génétiques chez les bactéries, archées et organites de certains eucaryotes. Après avoir été transcrits, ces ribozymes s'excisent de manière autocatalytique par épissage. L'intron de groupe II Ll.LtrB peut s'épisser par deux voies concurrentes : l'embranchement, qui génère des lariats; et la circularisation, un mécanisme moins connu qui génère des introns circulaires. Les lariats produits suite à l'embranchement peuvent se réintroduire dans des loci génétiques identiques au site d'origine mais inoccupés, ou dans de nouveaux loci génétiques ectopiques; se comportant ainsi comme des rétroéléments mobiles.

D'un point de vue évolutif, ces éléments mobiles sont les ancêtres proposés de plus de la moitié du génome humain, notamment les introns nucléaires et le noyau d'ARN de la machinerie d'épissage eucaryote : l'épissosome. Cependant, l'évolution des introns bactériens de groupe II s'est avérée difficile à étudier. Puisque la fonction de ces ribozymes dépend de leur structure secondaire, ils ont une très faible conservation de leur séquence primaire, ce qui complique les études phylogénétiques et notre compréhension de leur évolution chez les bactéries. En outre, bien que les introns nucléaires dérivés des introns de groupe II jouent un rôle essentiel chez les eucaryotes en utilisant l'épissage pour augmenter la diversité génétique, les introns bactériens de groupe II eux-mêmes n'ont pas de fonction commune qui bénéficie à leur hôte. Au contraire, ils ont toujours été considérés comme des éléments mobiles égoïstes qui parasitent les bactéries, utilisant l'épissage uniquement comme moyen de limiter les dommages causés à leur bactérie hôte. Ici, nous avons utilisé l'intron modèle de groupe II Ll.LtrB, provenant de la bactérie gram-positive *Lactococcus lactis*, pour répondre à des questions en suspens concernant leur évolution et leur fonction.

Pour étudier l'évolution de notre intron modèle, nous avons comparé Ll.LtrB à Ef.PcfG, un intron de groupe II récemment découvert par notre laboratoire chez la bactérie gram-positive *Enterococcus faecalis*. Comme ces deux introns étaient presque identiques (99,7%) mais présents dans des espèces bactériennes différentes, ils représentent probablement un événement de transfert horizontal récent. Nous avons donc émis l'hypothèse que l'analyse des 8 mutations ponctuelles entre les deux introns permettrait de mieux comprendre l'évolution des introns de groupe II

lorsqu'ils entrent dans un nouvel environnement cellulaire. Nous avons comparé les efficacités de mobilité et d'épissage des deux introns et avons trouvé que, bien qu'il n'y ait pas de changement significatif dans l'efficacité d'épissage, les 8 mutations modifiaient l'efficacité de mobilité. Ll.LtrB est nettement plus efficace à se mobiliser vers son propre site d'origine que Ef.PcfG, tandis que les deux introns reconnaissent plus efficacement le site d'origine de *E. faecalis*, ce qui suggère qu'il correspond au site ancestral. Enfin, un dendrogramme de la distribution de ces 8 mutations ponctuelles au sein de toutes les variantes de Ll.LtrB chez *L. lactis* montre une accumulation progressive des mutations de *E. faecalis* à *L. lactis*, ce qui suggère qu'un seul cas de transfert horizontal a mené à la dissémination de Ll.LtrB au travers de *L. lactis*.

Pour étudier la fonction des introns, nous avons cherché à comprendre pourquoi certains introns circulaires contenaient des nucléotides supplémentaires d'origine inconnue à leurs jonctions d'épissage. Comme ce phénomène avait déjà été observé dans divers sous-types d'introns de groupe II, nous avions proposé qu'il représentait peut-être une fonction conservée et jusqu'à présent inconnue des introns de groupe II. Nous avons démontré que l'origine de ces nucléotides supplémentaires était des ARNm codés par des chromosomes bactériens et des plasmides. Ll.LtrB peut envahir certains ARNm à des sites de reconnaissance spécifiques, d'où il peut épisser en *trans* soit son exon d'origine, soit un nucléophile étranger, vers l'ARNm en aval; un mécanisme qui catalyse donc la formation de molécules chimériques. Les introns circulaires contenant des nucléotides supplémentaires à leur jonction d'épissage sont ensuite produits par circularisation alternative vers un site de reconnaissance secondaire situé en amont. Ces produits circulaires accumulent ensuite dans la bactérie, agissant ainsi comme des marqueurs stables de la formation de chimères. Ll.LtrB peut donc augmenter la diversité génétique de son hôte bactérien *L. lactis* en générant des molécules d'ARNm chimériques.

Pour déterminer la pertinence biologique des ARNm chimériques générés par l'intron, nous avons exprimé le gène de relaxase *ltrB* interrompu par Ll.LtrB dans la présence d'un gène orthologue appelé *pcfG* et avons trouvé que des ARNm et des protéines chimériques étaient générés entre *ltrB* et *pcfG*. Nous avons démontré que l'abondance des relaxases chimériques est corrélée au nombre de copies d'intron, et que ces enzymes peuvent présenter des phénotypes de gain de fonction, où leur efficacité dépasse celle des relaxases WT. Comme les relaxases sont impliquées dans le transfert horizontal par conjugaison, la capacité des introns de groupe II à

augmenter la diversité génétique en formant des relaxases chimériques pourrait donc avoir joué un rôle important dans la dissémination rapide des introns de groupe II au sein de *L. lactis*.

Somme toute, nos résultats démontrent expérimentalement que les introns de groupe II se comportent et évoluent surtout comme des éléments mobiles dans les bactéries, plutôt que comme des éléments d'épissage. Nos travaux révèlent en outre que les introns bactériens de groupe II peuvent accroître la diversité génétique de leur hôte, une fonction qui est probablement issue au cours de l'évolution d'un comportement par ailleurs égoïste. Leur capacité de générer de nouvelles protéines qui peuvent bénéficier fonctionnellement à leur hôte permet donc d'expliquer en partie comment ce groupe de rétroéléments a été conservé dans les bactéries au cours de l'évolution.

Acknowledgments

I can’t begin to imagine what the last 6 years would have been without the outstanding tutelage and guidance of my supervisor, Dr. Benoit Cousineau. Your boundless energy, forward thinking, limitless insight and positive outlook helped push me out of my comfort zone and kickstart my burgeoning graduate studies into a success. The countless hours we spent talking about science and everyday life, whether over a good gin poured in lab beakers or a nice chat at the end of the day, are a testament to how you were always there for me, and made the time whenever you didn’t have any. Ben, I’m proud to have worked in your lab and alongside you, and I feel privileged that you gave me the opportunity to work on such great projects and that I could share my enthusiasm with you. I can’t thank you enough.

I’d also like to thank the undergraduate students I had the honour of supervising and working with during my PhD, Deeva, Samy, Rafia and Erika. But most of all, I want to thank my amazing colleague Caroline, who had the patience to both train me at my beginnings me and endure me during all this time. Your meticulousness and the perfection with which you did every single experiment always made me jealous, and always gave me an ideal to aspire to. I really enjoyed our time together and I consider myself lucky to have worked with you.

Everyday life in the Duff wouldn’t have been the same without my friends in the department, especially Patrick and Kayla, who found a way to make things even more fun once the Duff was empty and everyone else had moved out. Our end of day (or sometimes midday) drinks, gossiping, venting and exploration of the Duff always made for a good time. I also want to thank Marc-André, my roommate for a time and best friend. Our ruthless Covid-confinement cribbage matches and late-night gaming turned what could have been a bleak year into a great one. Your presence has been an anchoring point of fun and foolishness that helped keep me sane, and I consider myself lucky to have you as a friend.

I’d also like to thank my parents and my siblings, Gabriel and Justine, for their help and support throughout this long journey. Your endless affection and constant enquiring as to how my bacteria were doing always made me smile. You’ve always made me feel that I had your full support, even when I made you suffer through long bouts of vortexing during our phone calls. You were always there when I needed you.

Lastly, I would like to thank Makisha for her limitless love, support, encouragement and patience in listening to me vent about introns that weren't behaving the way I wanted them to. Thank you for helping my artless eye create beautiful posters and presentations. Thank you for sharing my joy and going out of your way to celebrate with me when things went well. Thank you for giving me the warmth, love and McDonalds I needed on dark days when everything seemed to be going badly. Having you by my side all these years has been my greatest joy and pride. You are the light of my life, I love you always.

Preface

This thesis was written in accordance with McGill University's "Guidelines for Thesis Preparation". The candidate has chosen to present in a "Manuscript-based thesis" format following these recommendations:

"As an alternative to the traditional thesis format, the thesis research may be presented as a collection of scholarly papers of which the student is the author or co-author; that is, it can include the text of one or more manuscripts, submitted or to be submitted for publication, and/or published articles reformatted according to the requirements described below. Manuscripts for publication are frequently very concise documents. The thesis is expected to be a more detailed, scholarly work than manuscripts for publication in journals. A manuscript-based thesis will be judged by the examiners as a unified, logically-coherent document in the same way a traditional thesis is judged."

Below is a list of the published manuscripts presented in this thesis, along with their respective contributions to original knowledge. Author contributions to each chapter of the thesis are detailed in the following section.

1. **LaRoche-Johnston F**, Monat C, Cousineau B. 2016. Recent horizontal transfer, functional adaptation and dissemination of a bacterial group II intron. *BMC Evolutionary Biology*, 16(1):223.

In this manuscript (**Chapter 2**), we functionally characterized a new group II intron from *Enterococcus faecalis*, called Ef.PcfG. We demonstrated using mobility and splicing assays that point mutations between this new intron and the model intron Ll.LtrB from *Lactococcus lactis* significantly increase Ef.PcfG's mobility efficiency to the *L. lactis* homing sites. By generating a dendrogram of point mutation accumulation between Ef.PcfG and all Ll.LtrB-variants found throughout *L. lactis*, we uncovered a gradual accumulation of point mutations. The likeliest explanation for our findings was that a single horizontal transfer event introduced an ancestral Ef.PcfG into *L. lactis*, where beneficial mutations enabled its continued dissemination throughout this new host. Overall, our study provided the first functional characterization of a group II intron's

adaptation to a novel cellular environment, and for the first time provided functional support to the theory that bacterial group II introns behave more like retroelements than splicing elements.

2. **LaRoche-Johnston F**, Monat C, Coulombe S, Cousineau B. 2018. Bacterial group II introns generate genetic diversity by circularization and *trans*-splicing from a population of intron-invaded mRNAs. *PLoS Genetics*, 14(11):e1007792.

In this manuscript (**Chapter 3**), we discovered the origin of additional nucleotide stretches found at the splice junction of group II intron circles. We characterized a new intron splicing pathway that accounts for these nucleotides, which combines aspects of branching and circularization. We demonstrated that group II introns invade host mRNAs at specific recognition sites *in vivo*, generating a population of intron-interrupted mRNAs. Introns within these invaded transcripts can *trans*-splice either their cognate E1 or a host mRNA to the downstream interrupted mRNA, respectively generating E1-mRNA and mRNA-mRNA chimeric molecules. Overall, we showed for the first time that bacterial group II introns can increase the genetic diversity of their host, disputing the longstanding claim that group II introns function solely as genetic parasites within bacteria.

3. **LaRoche-Johnston F**, Bosan R, Cousineau B. 2020. Group II introns generate functional chimeric relaxase enzymes with modified specificities through exon shuffling at both the RNA and DNA level. *Molecular Biology and Evolution*.

In this manuscript (**Chapter 4**), we assessed the biological relevance of chimeras generated *in vivo* by group II introns in the model organism *L. lactis*. We proved that chimeric transcripts can be formed by both Ll.LtrB and Ef.PcfG between orthologous relaxase mRNAs, producing chimeric enzymes. We demonstrated quantitatively than when intron copy numbers increase, the amount of these chimeric transcripts also rises. Using conjugation assays, we showed that a specific combination of orthologous relaxase exons yielded higher conjugation efficiencies than either of the WT enzymes, indicating a gain-of-function phenotype achieved by certain compound relaxases. We used phylogenetic tools to unveil the existence of natural chimeric relaxase genes,

where the exons of a single relaxase belong to different evolutionary lineages. Overall, we showed a concrete example of how an increase in genetic diversity caused by group II introns can lead to a beneficial function, refuting the claim that bacterial group II introns always undergo negative/purifying selection and rather suggesting that occasionally, bursts of intron dissemination may be fuelled by positive selection.

Contributions of authors

Chapter 1:

FLJ wrote the literature review.

Chapter 2:

FLJ, CM and BC designed the experiments and analyzed the data. FLJ and CM performed the experiments. FLJ and BC wrote the manuscript, with all three authors editing and approving the final version of the manuscript.

Chapter 3:

FLJ, CM and BC conceptualized and designed the experiments, and analyzed the results. FLJ, CM and SC performed the experiments. FLJ and BC wrote the manuscript, with FLJ, CM and BC editing and reviewing the final version of the manuscript.

Chapter 4:

FLJ and BC conceptualized and designed the experiments, and analyzed the results. FLJ and RB performed the experiments. FLJ and BC wrote, edited and reviewed the final version of the manuscript.

Chapter 5:

FLJ wrote the conclusions and perspectives.

Chapter 1:
Literature review and objectives of the thesis

1.1: The history of introns: intragenic elements

The course of biology and genetics was profoundly altered when the structure of DNA was discovered to be a double-helix (Watson and Crick 1953). Previous work by Oswald Avery had shown through his "transforming principle" that DNA acted as the carrier of hereditary traits for pathogenesis in bacteria (Avery et al. 1944). With DNA established as the repository for genes, the discovery of the former's structure now opened the door to a better understanding of the genetic code and protein synthesis. Francis Crick next laid out two hypotheses which later proved pivotal in bringing about the dawning of modern molecular biology. The first was the so-called Central Dogma, where the flow of information proceeds from DNA to RNA to proteins, such that the main function of DNA is the production of proteins (Crick 1958). The second was the Sequence Hypothesis, where the specific order of bases in a gene — then only defined as the smallest units of genetic information — represented a code which would yield a specific protein sequence (Crick 1958). When work on the genetic landscape of bacteriophages next emerged where genes appeared to be spread throughout DNA like beads on a string (Benzer 1959), all the experimental data seemed to support the prevailing view that a single gene coded for a single polypeptide (Beadle and Tatum 1941; Horowitz 1948).

This view was challenged however by the discovery that some mRNA fragments which had "matured" could no longer hybridize perfectly to the genomic DNA from which they were transcribed (Berget et al. 1977; Chow et al. 1977). Rather than one contiguous string of nucleotides yielding one polypeptide chain, these findings seemed to suggest a genomic architecture of "genes in pieces" (Gilbert 1978). These pieces would thereafter be described using two different terms: the exons, or expressed portions; and introns, denoting their intragenic nature (Gilbert 1978). Due to this specific genetic arrangement being initially found in eukaryotic genes, it was believed that introns were altogether absent from prokaryotes such as bacteria (Mercereau-Puijalon and Kourilsky 1979). Yet later studies would show the versatility of introns that splice from interrupted transcripts, which stem throughout the 3 domains of life.

The first type of introns to be discovered were thus the nuclear introns, which splice with the assistance of a large *trans*-acting machine called the spliceosome. This type of intron, along with some form of the spliceosome, are both found in the nucleus of every single eukaryote sequenced to date (Collins and Penny 2005; Vanacova et al. 2005), yet are completely absent from bacteria or archaea. Sequencing studies of the mitochondrial genome from the *Saccharomyces cerevisiae* yeast (Borst and Grivell 1978) demonstrated very soon thereafter the presence of new classes of introns, which based on the conservation of their secondary structure could be classified as group I and group II introns (Michel et al. 1982; Michel and Dujon 1983). The characteristics of group I introns were studied first, in the context of various unicellular ciliated eukaryotes called *Tetrahymena*, where they interrupt rDNA genes (Wild and Gall 1979). In this setting, group I introns were found to splice out of interrupted genes using a splicing pathway unique to themselves, requiring a free guanosine residue to act as an external nucleophile (Cech et al. 1981). This pathway ligates the flanking upstream and downstream exons, while also releasing linear group I intron molecules that can later undergo circularization (Grabowski et al. 1981). It was later shown that group I introns could self-splice *in vitro*, without the help of any protein co-factors, suggesting that the requirements for catalytic function may be entirely held within the RNA sequence. This discovery led Kruger and colleagues to coin the term “ribozyme”, or RNA enzyme, to denote any RNA molecule that can itself perform catalytic functions (Kruger et al. 1982).

When group II introns were shown to self-splice *in vitro*, the finding proved much more ground-breaking than for group I introns. Firstly, group II introns were shown to excise as lasso-like lariat structures, which branch at an internal adenosine residue very near the 3′ splice site (van der Veen et al. 1986). Moreover, the 2′-5′ covalent bond at the branchpoint differed from the 3′-5′ bond typical of circularized molecules (Peebles et al. 1986). The importance of these findings was immediately recognized due to their similarity with nuclear intron splicing, which also produce lariats branched at an internal adenosine residue through a 2′-5′ linkage (Padgett et al. 1984). Various groups thus proposed that an evolutionary link might exist between group II introns and nuclear introns, where they might share a common ancestor (Cech 1986; Jacquier 1990), while others suggested that observed similarities might simply reflect convergent evolution or biochemical determinism (Weiner 1993).

Since their discovery, research in the field of group II introns has flourished. Initially believed to be confined to the genomes of mitochondria and chloroplasts (Michel et al. 1989), they were quickly thereafter discovered in bacteria (Ferat and Michel 1993), and have since been identified in the genomes of nearly a quarter of all sequenced bacteria (Candales et al. 2012). Work on bacterial model systems has helped accelerate the characterization of group II introns, due to the rapid growth of bacteria and their pliability to modern genetic tools (Matsuura et al. 1997). Early observational evidence of the distribution of group II introns had led many groups to propose that these self-splicing RNAs could also function as mobile elements (Lambowitz 1989). Using model bacterial systems, group II introns were indeed shown to be mobile RNA elements, which use a variety of different mobility pathways (Cousineau et al. 1998; Cousineau et al. 2000), demonstrating their versatility. The following sections will outline the various discoveries made in the fields of group II intron splicing, mobility and evolution, with a special emphasis on the model group II intron Ll.LtrB from the gram positive bacterium *Lactococcus lactis*.

1.2: The architecture of group II introns: catalytic RNAs with a conserved structure

Group II introns are ribozymes that either interrupt coding or non-coding genetic loci. Upon transcription, group II introns can excise themselves from interrupted transcripts through self-splicing while concurrently ligating the flanking upstream and downstream exons (see Section 1.3). The ability of group II introns to self-splice accurately stems from their highly conserved secondary and tertiary structures. The striking similarities between the secondary structures of group II introns and spliceosomal snRNAs are what prompted Philip Sharp to propose that the fragmentation of group II introns into 5 *trans*-acting RNAs may have initiated the evolution of nuclear splicing (Sharp 1991). Since then, research in the field of group II intron structure has expanded immensely, largely driven by an attempt to use group II introns as model systems to better understand the folding and catalysis of the spliceosome (Keating et al. 2010). Solving the complex 3D structures of group II introns using crystallography further led to new insights regarding the complex tertiary interactions governing the folding of these large ribozymes (Toor et al. 2008).

1.2.1: Intron secondary structure

Group II introns have a conserved secondary structure that consists of six domains that radiate from a central wheel (Fig. 1.1) (Jacquier and Michel 1987). Although finer structural differences serve to further separate these self-splicing ribozymes into various subtypes (see Section 1.2.2), the six domains are a conserved facet of group II introns (Fig. 1.1B).

Domain I (DI) is the largest domain and contains a multitude of long-range binding sites. Among these are extensive tertiary interactions with other domains to coordinate proper group II intron folding, enabling DI to act as a molecular scaffold (Zhao et al. 2015). Also essential to intron splicing are the loop regions of DI that mediate recognition of the flanking exons (Jacquier and Michel 1987). For the bacterial group IIA intron Ll.LtrB, these correspond to two loop regions termed Exon Binding Sites 1 and 2 (EBS1/2), which bind through base pairing the Intron Binding Sites 1 and 2 (IBS1/2), a complementary stretch of 11 nucleotides at the 3′ end of the upstream exon 1 (E1) (Fig. 1.1B). The δ region of DI, on the other hand, base pairs with the complementary δ′ region at the start of the downstream exon 2 (E2). These contact points are also important for the accurate recognition of intron homing sites within DNA and RNA substrates during intron mobility (see Section 1.4). DI is the first domain to be transcribed, and its folding is independent of the other domains (Qin and Pyle 1997). Since transcription and translation are coupled in bacteria, it is unclear how the folding of group II introns within interrupted mRNAs is affected by the presence of ribosomes. In the case of certain group I introns, ribosomes actively translating into the 5′ intron sequence prevent proper folding and self-splicing of the intron (Ohman-Heden et al. 1993). However, ribosomes have also been shown as necessary to “iron-out” the intron secondary sequence, allowing for sequential modular folding and assembly (Semrad and Schroeder 1998). Since proper folding of DI is required for the subsequent assembly of other intron domains, DI folding is the rate-limiting step in the formation of an active ribozyme (Su et al. 2005).

Domain II (DII) is not actively involved in catalysis, though it does play important structural roles through tertiary interactions (Costa et al. 1997). Of particular note is the η- η′ interaction with DVI, which is key in mediating the conformational shift that occurs between the two steps of splicing (Fig. 1.1B) (Chanfreau and Jacquier 1996). The joiner region between Domains II and III (J2/3) is one of the most conserved sequences of group II introns, playing an essential role in catalysis (Mikheeva et al. 2000; de Lencastre and Pyle 2008). It contains the γ

nucleotide, which binds to the last nucleotide of the intron (γ′) to ensure proper positioning of the intron's 3′ end within the active site (Fig. 1.1B) (Jacquier and Michel 1990). Moreover, the J2/3 region is inserted within the major groove of DV, directly interacting with the AGC catalytic triad of DV (Fig. 1.1B) to form a triple helix that coordinates the magnesium ions within the catalytic site (Toor et al. 2008). Domain III (DIII) does not play a direct role in catalysis, similarly to DII (Koch et al. 1992). However, it is considered to be a catalytic effector, due to its ability to increase splicing reaction rates (Fedorova et al. 2003; Fedorova and Zingler 2007). DIII binds with high affinity to many parts of the intron (Podar et al. 1995), strengthening the overall fold of the intron through interactions with other domains (Fedorova and Pyle 2005). Among these is the μ-μ′ pairing with DV, which stabilizes the active site structure (Fig. 1.1B). Most importantly, DIII serves as an allosteric effector of catalysis by positioning the J2/3 linker within DV (Toor et al. 2008).

Domain IV (DIV) often contains a large open reading frame (ORF) that can take up as much as 2/3 of the total intron sequence, such as for the bacterial group II intron Ll.LtrB (Fig. 1.1A). DIV is not involved in catalysis, but rather loops out of the catalytic RNA core (Michel et al. 1989). This structural characteristic has led to natural cases of group II intron fragmentation within DIV, which can nevertheless assemble *in trans* (see Section 1.3.4) (Jarrell et al. 1988a; Belhocine et al. 2007a). Moreover, DIV's distance from the RNA active site enables the insertion of additional material within DIV with limited effects on intron splicing or mobility (Cousineau et al. 1998). The ORF codes for an intron-encoded protein (IEP), which invariably contains an N-terminal reverse-transcriptase domain that plays an important role in intron binding and cDNA synthesis during intron mobility (see Section 1.4); followed by a maturase (X) domain whose main role is to coordinate the proper folding of the intron RNA structure (Fig. 1.1A) (Mohr et al. 1993; Matsuura et al. 1997). Some group II intron IEPs, such as LtrA from Ll.LtrB, also contain C-terminal DNA-binding and endonuclease domains that are important in intron mobility (Fig. 1.1A), whereas other mobile group II intron subtypes lack these domains (Toro and Martinez-Abarca 2013). Translation of the Ll.LtrB IEP can occur from the pre-mRNA containing the entire intron sequence, yielding the LtrA protein, yet the majority of LtrA transcripts emanate from an internal promoter within the intron sequence (Zhou et al. 2000). The LtrA enzyme is essential for both splicing and mobility of Ll.LtrB, where the maturase domain binds to multiple contact points throughout the intron RNA sequence to help coordinate accurate folding of the intron (Wank et al. 1999). However, the binding site for which LtrA has highest affinity is a stem-loop in DIVa (Fig.

1.1B) (Watanabe and Lambowitz 2004). This site also overlaps with the LtrA Shine-Dalgarno sequence, thus auto-regulating its own translation through steric hindrance (Singh et al. 2002).

Domain V (DV) contains a very small stem-loop structure, yet its sequence is the most conserved throughout group II introns (Michel and Ferat 1995). This conservation is due to its role as the catalytic core of group II introns. The small stem-loop structure of DV can be defined as having two faces: a binding face, and a chemical face (Fedorova and Zingler 2007). The binding face is important for mediating key tertiary interactions with DI (κ-κ′, λ-λ′, ζ-ζ′) (Boudvillain et al. 2000) and with DIII (μ-μ′), altogether contributing to position DV within the active site (Fig. 1.1B) (Fedorova and Pyle 2005). On the other hand, the chemical face contains the key residues for catalysis. Among the catalytically important motifs of DV is the AC dinucleotide bulge (Fig. 1.1B), whose phosphate backbones bind to magnesium ions essential to catalysis (Schmidt et al. 1996). This occurs in coordination with the catalytic triad of DV, a 3-nucleotide AGC motif that forms a triple helix structure with the J2/3 linker, which also helps coordinate the metal ions (Chanfreau and Jacquier 1994; Toor et al. 2008).

Domain VI (DVI) has the sole function of providing the bulged adenosine for the first step of splicing (Michel and Dujon 1983). The phylogenetic conservation of DVI is accordingly quite low, except for the bulged adenosine itself, which is very conserved (Toor et al. 2001). The use of the branchpoint is highly specific and is supported by several features of the group II intron secondary structure, such as the length of the basal stem of DVI and of the linker between DV and DVI (Chu et al. 2001). DVI also has long-range tertiary interactions with DII (η-η′), which serve to position the bulged adenosine and facilitate the conformational switch in between the two steps of splicing (Fig. 1.1B) (Chanfreau and Jacquier 1996).

1.2.2: Diversity of group II introns

When their secondary structures were first described, group II introns were divided into two subtypes based on their overall structure and their mechanisms of exon recognition during splicing: subclasses IIA and IIB (Jacquier and Michel 1987; Michel et al. 1989). As the number of described group II introns increased so did the resolution of group II intron subtypes, such as a

unique class of primitive group II introns that mobilize downstream of bacterial transcriptional terminators: the IIC introns (Fig. 1.2) (Granlund et al. 2001; Robart et al. 2007).

Subsequent efforts to classify group II introns focused on phylogenetic data. First, the IEPs of retromobile group II introns were compiled and shown to correspond to distinct lineages in chloroplasts and mitochondria, while bacterial ORFs positioned at the base of the tree, thus possibly being ancestral (Zimmerly et al. 2001). Moreover, a pattern of coevolution emerged where the RNA sequences of group II introns consistently grouped into the same clades as their IEPs (see Section 1.5). Together, these observations took the form of the retroelement ancestor hypothesis, in which the ancestor of extant group II introns functioned as a retroelement, and all the different types of group II introns observed today in bacteria and organelles are derived from such retroelements (Toor et al. 2001).

The retroelement hypothesis was further used as a means of classifying group II intron RNAs and IEPs (Simon et al. 2008). Using conserved sequences in the catalytic DV of the RNA sequence and of the RT sequence in IEPs, a comprehensive analysis was done to classify group II introns within distinct clades (Simon et al. 2009). The initial organization of group II introns into classes IIA, IIB and IIC was maintained to describe how group II introns recognize and bind to their flanking exons during splicing. However, additional resolution of IEP clades allowed further classification of group II introns: IIA introns such as Ll.LtrB cluster with mitochondrial-like (ML) ORFs; IIB introns cluster with ORF classes B, D, E, F and chloroplast-like (CL); IIC introns cluster with ORF class C (Fig. 1.2). Although in most cases the intron RNA was grouped in the same clade as their cognate ORFs, there were a few exceptions for IIB introns. Possible reasons for varying evolutionary histories between intron RNAs and their ORFs could be explained by twintrons, where an intron invades the sequence of another intron. In such cases, the invading intron's catalytic RNA may degrade and its ORF could be adopted by the invaded intron RNA (Dai and Zimmerly 2003). However, discordance between intron RNA and ORF sequences could also arise because functional constraints in certain environments lead to convergent evolution (Kelchner 2002). It nevertheless appears clear that in most cases, group II intron RNAs coevolve with their IEPs.

1.2.3: Intron tertiary structure

Insight into group II intron architecture deepened when a crystal structure was obtained for a group IIC intron from the extremophilic bacterium *Oceanobacillus iheyensis* (Toor et al. 2008). Although group IIC introns represent a smaller, more ancient class of group II introns, crystallographic studies revealed key insights about the conserved catalytic core of group II introns. When combined with subsequent crystallographic data of IIB (Robart et al. 2014) and IIA (Qu et al. 2016) group II introns, these studies helped identify conserved facets of group II intron tertiary structures.

Group II introns all have a conserved catalytic core that coordinates two monovalent ions (often potassium ions K1 and K2) and two divalent ions (often magnesium ions M1 and M2) (Marcia and Pyle 2014). Crystal structures show that the phosphate backbones of the catalytic triad in DV serve to create a negatively-charged pocket in the catalytic core, where the two Mg^{2+} ions are recruited and positioned in the sharp kink formed by the catalytic 2-nt bulge (Marcia and Pyle 2014). The two potassium ions, on the other hand, serve to stabilize the correct conformation, since replacement with smaller Li^{+} or Na^{+} ions result in an opened catalytic core where all metal ions are released (Marcia and Pyle 2012). These findings confirmed a longstanding theory that group II intron catalysis is mediated by a two-metal-ion mechanism, where one metal ion activates the nucleophile while the second metal ion stabilizes the leaving group (Steitz and Steitz 1993).

Recently, the crystal structure of the model group IIA intron Ll.LtrB was resolved, where it is spliced as a lariat bound to the LtrA IEP (Qu et al. 2016). The interaction showed that the IEP binds to the DIVa loop through its RT domain, which was previously shown to be highly positively charged (Zhao and Pyle 2016). Once anchored in DIV, the RT domain also interacts with DI to integrate itself into the RNA scaffold. The maturase domain of the IEP next initiates multiple long-range interactions throughout the intron RNA. Most of these interactions occur in DI, notably at the EBS1/2 site of exon 1. In the absence of LtrA, the EBS loop regions of DI become untethered to the upstream exon, indicating an important role of the IEP in strengthening the intron's binding to the 5′ splice site (Qu et al. 2016). Despite the low sequence conservation of maturases throughout group II intron IEPs, they all tend to have high positive charges, thus allowing them to bind negatively charged RNA (Zimmerly et al. 2001; Blocker et al. 2005). Moreover, the ability of the LtrA IEP to increase the reactivity of group II introns appears to occur without direct contact

with either the catalytic core or with DVI. Rather, IEP binding to the intron may involve a more indirect, allosteric process, where LtrA limits the number of spurious conformations the intron RNA structure can adopt and thus increases the likelihood that a catalytic state is reached (Zhao and Pyle 2017).

1.3: Group II introns as self-splicing elements

The culmination of the complex interplay of secondary and tertiary interactions is to allow the intron RNA to catalyze the splicing reaction, where the intron ligates its flanking exons and is itself released. Self-splicing always proceeds through two consecutive transesterification reactions. As mentioned previously, the splicing reactions catalyzed by group II introns and the spliceosome were proposed to have common ancestry due to biochemical similarities in the transesterification reactions they catalyze (Cech 1986). This notion was further supported by the proposal and subsequent crystallographic validation that they both catalyze the splicing reaction through the same process: coordinating two divalent metal ions inside a conserved catalytic core (Steitz and Steitz 1993). This meant that insights gained by studying the mechanistic intricacies of group II intron splicing could lead to novel insights into the inner workings of the spliceosome (Smathers and Robart 2019).

Since the discovery that group II introns self-splice as lariats in a mechanistically identical pathway as nuclear introns (Peebles et al. 1986; van der Veen et al. 1986), new types of splicing pathways have emerged (McNeil et al. 2016). Although many of these were initially described *in vitro* under unphysiologically high salt conditions (Jarrell et al. 1988b), they have since been shown to occur *in vivo* for a number of different group II introns. The following subsections will highlight the main mechanistic attributes of each splicing pathway.

1.3.1: The branching pathway

Branching was the first self-splicing pathway described for group II introns (Peebles et al. 1986), and it has since remained by far the most studied pathway (Pyle 2016). Using the same two biochemical steps as nuclear splicing (Padgett et al. 1984), the 2′ OH of a bulged adenosine residue

in DVI termed “branchpoint” acts as the nucleophile in the first transesterification reaction, targeting the first nucleotide of the group II intron at the 5′ splice site (Fig. 1.3A) (Schmelzer and Schweyen 1986). This reaction generates the lasso-like lariat structure of branched group II introns, where the 5′ phosphate of the intron’s first nucleotide is covalently bound to the 2′ OH of the branchpoint adenosine through a 2′-5′ phosphodiester bond. The first transesterification also releases E1, which nonetheless remains attached to the intron through non-covalent IBS1/2-EBS1/2 base pairing interactions (Fig. 1.3A). Moreover, E1 remains positioned in the intron’s catalytic active site (Marcia and Pyle 2012), where it acts as the nucleophile in the second transesterification reaction. The intron thus toggles between two different active conformations to increase proximity between the reactants during each splicing reaction (Chanfreau and Jacquier 1996). During the second transesterification reaction, the 3′ OH of the last nucleotide of E1 attacks the first nucleotide of E2, thus ligating the two exons together and releasing the group II intron as a branched lariat (Fig. 1.3A).

Once released as lariats, group II introns can also use the reverse branching pathway to re-insert into different genetic loci in RNA or DNA (Fig. 1.3A, double arrows) (Augustin et al. 1990; Morl and Schmelzer 1990). Reversing the branching pathway into new genetic loci provides the basis for group II intron mobility (see Section 1.4). During branching, the same number of phosphate bonds are created and broken, providing the basis for the reversal of the pathway. Since self-splicing through the branching pathway is an energetically neutral process, both “forward” and “reverse” splicing through the branching pathway are equally possible (Robart and Zimmerly 2005). The first step of branching is readily reversible in both forward and reverse splicing reactions (Chin and Pyle 1995), likely having evolved as a means to promote intron mobility, rather than as a proofreading mechanism to prevent mis-splicing (Wang and Silverman 2006). However, the second step of splicing is much faster than the first step, rendering the first step of splicing rate-limiting (Daniels et al. 1996). This implies that the invasion of a new genetic site through reverse-splicing is much less energetically favorable than forward splicing and rather inefficient (Daniels et al. 1996). However, group II introns have evolved other means to increase the efficiency of reverse-splicing, such as by reverse-transcribing the intron in place following the reversal of the second step, forming a kinetic trap (Aizawa et al. 2003).

1.3.2: The hydrolytic pathway

Soon after the discovery of branching, group II introns were rapidly found to also self-splice through different pathways. When the bulged adenosine molecule of DVI, key for the first transesterification reaction in branching, is base-paired to a complementary U residue — or removed altogether — group II introns can still self-splice without the formation of lariat molecules (van der Veen et al. 1987). Under high salt concentrations (high monovalent salt or Mg^{2+}), linear excised group II introns were produced through *in vitro* splicing (Jarrell et al. 1988b) and were later demonstrated to occur *in vivo* as well (Podar et al. 1998). This intron form was proposed to originate in a novel group II intron splicing pathway, in which splicing is initiated at the 5′ splice site by a hydroxyl ion or water molecule rather than the bulged adenosine residue (Fig. 1.3B). The second step is essentially the same as branching, where the 3′ OH of excised E1 attacks the 3′ splice site, releasing ligated exons and a linear intron. Although linear group II introns can reverse the first step of splicing into a target site, they are unable to catalyze complete reverse-splicing the way lariats can (Fig. 1.3B) (Roitzsch and Pyle 2009). Linear introns were nevertheless shown to be mobile in eukaryotes through partial reverse splicing, followed by reverse transcription and non-homologous end-joining (Zhuang et al. 2009).

However, the balance between branching and hydrolysis is not only a function of the branchpoint, as evidenced by its occurrence in conjunction with traditional branching for some group II introns (Daniels et al. 1996). For the O.i.I1 group IIC intron from *Oceanobacillus iheyensis*, the presence of a branchpoint still mainly results in splicing through the hydrolysis pathway, yet increasing the length of the stem at the base of DVI to weaken its interaction with DII leads to an increase in branching (Monachello et al. 2016). This suggests that a short DVI stem leads to its sequestration by DII, where it is prevented from toggling between the two active conformations it needs to adopt for efficient branching (Chanfreau and Jacquier 1996), thus defaulting to hydrolysis.

Various group II intron subclasses were discovered that naturally lack their branchpoint adenosine residues and were found to splice exclusively as linear introns *in vivo* (Vogel and Borner 2002; Li et al. 2011). Hydrolysis has even been proposed to be an ancestral form of self-splicing, from which transesterification could have later evolved. The advantage of branching over hydrolysis would have been the ability to mobilize into new sites due to the complete reversibility

of the pathway (see Section 1.4), eventually displacing hydrolysis as the main method of self-splicing (Podar et al. 1998; Bonen and Vogel 2001).

1.3.3: The circularization pathway

When spliced group II introns were first described by electron microscopy, they appeared and were proposed to be true circular molecules (Arnberg et al. 1980). However, with the discovery that group II introns mainly self-splice using the branching pathway (Peebles et al. 1986; van der Veen et al. 1986), circularization was largely forgotten, until it was functionally demonstrated *in vivo* (Murray et al. 2001). Using branchpoint-mutant group II introns, Murray and colleagues had assumed that the only splicing products they would obtain *in vitro* would be hydrolyzed linear introns, which were well-described at the time (Gaur et al. 1997). Yet they were able to detect circular excised introns, which contained a 2′-5′ linkage at the splice junction similar to intron lariats, yet were resistant to debranching enzymes which turn lariats into linear introns (Ruskin and Green 1985). Since the discovery of circularization *in vivo*, group II introns have been found in natural settings to be lacking a branchpoint and to self-splice as circles, suggesting a conserved splicing pathway (Li-Pook-Than and Bonen 2006).

Using branchpoint mutants of the Ll.LtrB group II intron from *L. lactis*, circularization was later demonstrated to occur through an initial *trans*-splicing of free E1 at the 3′ splice site, yielding ligated exons (Fig. 1.3C) (Monat and Cousineau 2016). The free 3′ end of the intron then attacks the 5′ splice site, generating a circular head-to-tail intron with a 2′-5′ linkage, and releasing additional free E1 (Fig. 1.3C). Since these molecules lack the 3′ OH found on the tails of lariats, they are unable to mobilize by retrohoming (Monat et al. 2015). However, they are likely shielded from degradation by RNases due to their circular nature, since they were shown to accumulate in bacteria over time (Monat et al. 2015). The source of free E1 to initiate circularization is hypothesized to be the Spliced Exon Reopening (SER) reaction (Fig. 1.3), which was first shown to occur *in vitro* (Jarrell et al. 1988b) and later demonstrated *in vivo* (Qu et al. 2018). In this pathway, group II introns target ligated exons for hydrolysis, cleaving them at the intron recognition site and releasing free E1 and E2. This hypothesis was bolstered when short RNAs

corresponding to the last 17 nucleotides of E1 were added *in vitro* to unspliced precursor mRNA, leading to an increased production of circular molecules (Murray et al. 2001).

The transesterification reactions during circularization were shown to occur with some degree of variability, as evidenced by the presence of intron circles *in vivo* containing an additional C residue at their circle splice junction (Molina-Sanchez et al. 2006; Monat et al. 2015), and sometimes harboring longer stretches of additional nucleotides at their circle splice junctions (Li-Pook-Than and Bonen 2006; Monat et al. 2015; Monat and Cousineau 2016). The presence of additional C residues was shown to stem from misrecognition of the 3′ splice site during the first step of splicing, resulting in intron circles containing the first nucleotide of E2 (Monat and Cousineau 2016). However, the origin of the longer stretches of additional nucleotides observed at the junctions of certain intron circles has never been explained (see Chapter 3).

Although the presence of intron circles has now been widely reported in both prokaryotes and eukaryotes, it is unknown whether they all have a similar function which has yet to be discovered, or whether individual and distinct functions have independently emerged throughout evolution (Lasda and Parker 2014). The accumulation of intron circles in *Podospora anserina* has previously been correlated with senescence, but has yet to be validated elsewhere (Osiewacz and Esser 1984; Begel et al. 1999). Various other functions have been attributed to circular RNA such as acting as miRNA sponges *in vivo* and in protein sequestration (Hansen et al. 2013; Du et al. 2017). Yet despite a now detailed understanding of the mechanism of circularization, a clear function for group II intron circles has not been found (see Chapter 3).

1.3.4: *Trans*-splicing

Group II introns can also self-splice through *trans*-splicing, where separate mRNA transcripts harboring fragments of the same group II intron assemble to self-splice using the branching pathway (Fig. 1.3D) (Bonen 1993). During *trans*-splicing, group II intron fragments fold into their respective secondary structures, allowing independently folded fragments to interact with each other through tertiary interactions (Quiroga et al. 2011). Fragmentation sites were found to occur naturally most often in DIV (Bonen 1993; Michel and Ferat 1995), which was demonstrated *in vitro* to support tertiary interactions and allow splicing to occur (Jarrell et al.

1988a). Upon assembly, fragmented group II intron *trans*-splicing next occurs following the same two transesterification reactions as branching (Fig. 1.3D). The result is the release of ligated exons and a branched "Y"-shaped group II intron.

This type of splicing was first suggested upon finding that the chloroplast *psaA* gene from the *Chlamydomonas reinhardtii* alga was split into 3 exons, each widely separated throughout the chloroplast genome (Choquet et al. 1988). Since its discovery, bipartite (2 intron pieces) and tripartite (3 intron pieces) group II introns have been widely documented, all in the organelles of lower eukaryotes and higher plants (Kohchi et al. 1988; Knoop et al. 1997). Despite their notable absence from prokaryotes, the bacterial group II intron Ll.LtrB was fragmented at sites analogous to the natural fragmentation sites of organellar bipartite and tripartite group II introns and was shown to fold and self-splice accurately *in vivo* (Belhocine et al. 2007a). A Tn5 transposon-based fragmentation system further showed that bacterial group II introns can be fragmented in 2 or 3 pieces in a multitude of sites that have never been observed in naturally *trans*-splicing group II introns (Belhocine et al. 2008; Ritlop et al. 2012).

The ability of group II intron fragments to accurately fold, assemble and splice lends experimental credence to the hypothesis that the eukaryotic spliceosome arose through fragmented group II introns, which maintained their function yet began acting *in trans* (Cavalier-Smith 1991; Sharp 1991). However, the ability of fragmented group II introns to *trans*-splice is different from *trans*-splicing carried out by the spliceosome. In the nuclei of eukaryotes, snRNAs can *trans*-splice exons together that originate from separate mRNA transcripts belonging to different genes, thus forming chimeric mRNAs (Lasda and Blumenthal 2011). The nuclear process thus differs from fragmented group II introns, which are bound to *trans*-splice based solely on self-recognition. The ability of group II introns to *trans*-splice cellular mRNA transcripts together has never been demonstrated and would provide an interesting functional link with spliceosomal snRNAs (see Chapter 3).

1.4: Group II introns as mobile retroelements

Though group II introns were initially described as splicing elements, evidence rapidly grew to suggest that these ribozymes were also mobile. Due to observational evidence of their

often patchy distribution (Field et al. 1989; Lambowitz and Belfort 1993), group II introns began to be called infectious, even before a mobility pathway had been elucidated (Lambowitz 1989). Group II introns were shown to mobilize to identical, unoccupied sites in the absence of other mobile group I introns, using a pathway that depended on an intact intron core and associated maturase protein (Skelly et al. 1991).

Studies on the mechanism of group II intron mobility began in eukaryotic organelles, specifically in the mitochondria of yeast. Using such model systems revealed several facets of intron mobility that were later demonstrated to be common throughout all different subtypes, including bacterial group II introns. First, it was shown that the RNA portion of group II introns is responsible for catalyzing insertion into the DNA sense strand of the mobility target site (Zimmerly et al. 1995a). Second, the intron-encoded protein (IEP) of group II introns was demonstrated to function not only as a maturase that assists intron folding to allow accurate splicing, but also as an essential cofactor for mobility through maturase-assisted reverse splicing, as well as its DNA-binding, endonuclease and reverse-transcriptase domains (Fig. 1.1A) (Curcio and Belfort 1996). The endonuclease portion, previously known only as a conserved Zn^{2+} domain, was shown to be responsible for nicking the negative DNA strand, leading to a double-stranded DNA break (Zimmerly et al. 1995a). On the other hand, the RT domain is responsible for generating a cDNA copy of the intron RNA during mobility (Kennell et al. 1993). Third, mobility was shown to occur via a process called target-primed reverse transcription, where the 3′ OH generated by the proteolytic cleavage of the negative strand serves as a primer for reverse-transcription by the RT domain of the IEP (Zimmerly et al. 1995b).

However, once group II introns were discovered in bacteria (Ferat and Michel 1993), they rapidly emerged as model systems that were much more pliable to genetic manipulation. Mobility experiments began in *Lactococcus lactis*, where the resident Ll.LtrB was the first bacterial group II intron shown to both splice and mobilize *in vivo* (Mills et al. 1996; Shearman et al. 1996). The next sections will outline the main breakthroughs that occurred in elucidating the mechanisms that support group II intron mobility in bacteria.

1.4.1: Retrohoming

The first bacterial group II intron mobility pathway to be characterized was retrohoming, where the intron invades a dsDNA target site that is both identical to its original site and is unoccupied (Fig. 1.4A) (Cousineau et al. 1998). The basis for retrohoming is the reversal of the branching pathway by excised group II intron lariats (see Section 1.3.1), where the intron RNA integrates within intronless alleles (Yang et al. 1996). Using the Ll.LtrB group II intron from *L. lactis* in the compatible and more genetically pliable setting of *Escherichia coli* (Matsuura et al. 1997), the mechanistic details of bacterial retrohoming rapidly emerged.

During retrohoming, a bacterial group II intron first self-splices through the branching pathway, releasing an active RNP: a lariat RNA bound to its IEP. The active RNP next binds non-specifically to DNA and scans for a suitable integration site, through facilitated diffusion (Aizawa et al. 2003). Recognition of a cognate intronless allele is achieved through a combination of interactions between the DNA target site and both the RNA and protein component of the RNP (Jacquier and Michel 1987; Guo et al. 1997). Through base pairing interactions, Ll.LtrB recognizes a stretch of nucleotides in E1 directly upstream of the mobility site called Intron Binding Sites 1 and 2 (IBS1/2), which it binds using two loop regions in DI called Exon Binding Sites 1 and 2 (EBS 1/2). The intron also base pairs with the first nucleotide of E2 using the δ-δ′ tertiary interaction. Many more distal interactions take place between the IEP and the DNA target site, spanning nucleotides -20 to +10 (Singh and Lambowitz 2001).

Once the intron is bound to a suitable site, it reverses both transesterification steps used during branching to insert itself into the DNA sense strand, or top strand. The endonuclease domain then nicks the bottom strand of the target DNA slightly downstream of the integration site, generating a double-stranded break. The liberated 3′ OH of the nicked DNA bottom strand is next used by the IEP's RT domain to initiate target-primed reverse-transcription, generating a cDNA copy of the group II intron RNA. The intron now consisting of an RNA/DNA hybrid, host-encoded RNaseH enzymes digest the RNA copy of the group II intron. The process of retrohoming is completed when bacterial DNA pol III synthesizes the second DNA strand of the group II intron, which is finally sealed using DNA ligase (Smith et al. 2005). Interestingly, mobility of Ll.LtrB occurs without coconversion of flanking markers (Cousineau et al. 1998). This is in stark contrast with their counterparts in yeast mitochondria, where mobility is completed when host repair

mechanisms such as homologous recombination use an intron-containing allele to integrate the group II intron into the intronless allele, resulting in part of the upstream exon also appearing in the mobility site (Lazowska et al. 1994). Therefore, the bacterial Ll.LtrB group II intron retrohoming pathway occurs without the aid of host repair pathways such as homologous recombination (Cousineau et al. 1998). The result of retrohoming is thus the stable, RecA-independent RNA-based mobility of a group II intron into a cognate unoccupied genetic locus (Fig. 1.4A).

Although the mechanism described above represents the classical retrohoming pathway, different subtypes of group II introns exist in bacteria that employ mechanistically distinct mobility pathways. Certain group II intron subtypes, including RmInt1 from *Sinorhizobium meliloti*, contain IEPs within separate evolutionary clades that lack an endonuclease domain (Molina-Sanchez et al. 2010). Despite this reduced IEP, such group II introns are nevertheless mobile, also using a RecA-independent pathway (Martinez-Abarca et al. 2000; Martinez-Abarca and Toro 2000). Rather than mobilizing into dsDNA, these introns insert into the ssDNA of replication forks generated during bacterial replication. They can then use the 3′ OH of nascent DNA lagging or leading strands to prime reverse transcription and generate a cDNA copy of themselves (Martinez-Abarca et al. 2004).

Moreover, group IIC introns have evolved a different strategy than mobilizing into ORFs to ensure subsequent transcription and limit damage to the bacterial host. During retrohoming, IIC introns recognize a combination of conserved sequences in DNA that lead to efficient reverse splicing, as well as structural motifs. Most often, mobility occurs directly after transcriptional terminator step-loops (Toor et al. 2006; Robart et al. 2007). However, this structural recognition can also lead to different specificities, such as the group IIC-*attC* intron subclass (Fig. 1.2) that mobilizes into site-specific recombination sequences for integron integrases (Leon and Roy 2009).

1.4.2: Retrotransposition

Bacterial group II introns can also mobilize to non-cognate or ectopic sites, albeit at much lower efficiencies. Although there are some mechanistic differences with retrohoming, the basis for retrotransposition remains the recognition of a potential insertion site through base pairing

interactions with the intron RNP. The advantage of maintaining base pairing with the targeted insertion site as a prerequisite for mobility is that the intron can maintain the IBS1/2-EBS1/2 base pairing required to self-splice following transcription, enabling the continued expression of invaded genes. Retrotransposition in bacteria was first described as an endonuclease-independent, RecA-dependent pathway occurring primarily when a group II intron RNP reverse-splices into an ectopic site of bacterial mRNA rather than DNA (Fig. 1.4B) (Cousineau et al. 2000). After invading the target mRNA, the intron reverse-transcribes the transcript to generate a cDNA copy of itself within the new site. Host-encoded enzymes degrade the intron RNA and perform second-strand synthesis, resulting in a dsDNA intron-interrupted allele (Smith et al. 2005). Mobility is accomplished when the interrupted allele displaces the initial intronless allele in the bacterial chromosome through homologous recombination (Fig. 1.4B).

However, the analysis of retrotransposition events in bacterial genomes demonstrated that many group II introns also reside in intergenic regions, which was inconsistent with a solely mRNA-based mechanism for mobility to ectopic sites (Dai and Zimmerly 2002a). Subsequent pathways that were described relied more closely on retrohoming events that occurred in ectopic DNA sites rather than identical intronless alleles (Ichiyanagi et al. 2002). In this setting, mobility occurs either with or without the endonuclease domain, and can take place in the lagging strand of replication forks or through inaccurate insertion events in double stranded DNA. Insertion events were shown to be biased towards the lagging strand, suggesting a preferential use for ssDNA as a target for reverse splicing and Okazaki fragments as primers for synthesizing cDNA (Fig. 1.4C). Lagging strands were also shown to be favored when group II introns reside in bacteria with short doubling times such as *E. coli*, likely due to the increased frequency of replication forks (Coros et al. 2005). The use of replication forks as targets for retrotransposition may also partly explain the heavy association of group II introns with other mobile elements such as plasmids (Klein and Dunny 2002), since plasmids have a higher number of replication forks per unit of DNA length than bacterial chromosomes (Ichiyanagi et al. 2003).

The ability of group II introns to mobilize into ectopic sites is an important facet of their evolution. It has enabled these mobile retroelements to diversify their niches by introducing them to novel sites that they could adapt to over the course of evolution.

1.5: The evolution of group II introns

Given their wide distribution and the variability of their secondary structure, a concerted effort has been made to classify group II intron subtypes (Fig. 1.2) and identify their phylogenetic relationship both to each other and to other genetic elements (see Section 1.2.2). Recently, a largescale phylogenetic analysis was conducted on group II introns to resolve their evolutionary history and identify their particular clades (Fig. 1.5) (Simon et al. 2009). However, this study highlighted several characteristics of group II introns that render their phylogenetic analysis quite difficult. First, group II intron catalysis relies on the secondary structure of the RNA, resulting in very low primary sequence conservation in the noncoding RNA portion of the intron, except for the very small catalytic DV. This leaves a larger portion in the ORF that can be used as a marker for evolutionary studies. However, residues within the IEP need to be limited to portions of the RT domain that are shared across all group II intron IEPs, and furthermore biases phylogenetic studies to group II introns that contain an IEP, which is not always the case (Simon et al. 2008). Second, group II introns have heterogeneous base compositions that vary according to the inherent mutational biases of the host organisms, especially ones with high GC content (Mooers and Holmes 2000). Third, group II introns are found in such a wide array of organisms that mutational saturation often occurs, especially at the third position of codons, thus randomizing the phylogenetic signal and decreasing the signal-to-noise ratio (Simon et al. 2009). Fourth, group II introns associate frequently with mobile genetic elements that spread laterally throughout populations, lending disproportionately large importance to horizontal rather than vertical transmission (Klein and Dunny 2002).

Overall, the combined RNA and IEP signals that can be reliably used to generate phylogenetic trees are 138 nucleotides and 230 amino acids, respectively (Fig. 1.5). The broad phylogenetic analyses of group II introns support the notion that catalytic RNAs and their IEPs form robust clades and show patterns of coevolution (Fig. 1.5) (Toor et al. 2001). However, while using so few residues as phylogenetic signals has the advantage of allowing sampling across domains, it leaves a considerable lack of resolution at the tips of the trees (Simon et al. 2009). This renders the examination of recent natural instances of intron dispersal (Dai and Zimmerly 2002b; Fernandez-Lopez et al. 2005; Tourasse and Kolsto 2008) the best way to understand the selective

pressures shaping the short term evolution of group II introns, which remains poorly understood (see Chapter 2).

1.5.1: Origin of group II introns

Phylogenetic studies of group II introns have consistently found that the most parsimonious trees have bacterial retroelements at their base (Zimmerly et al. 2001; Simon et al. 2009). Combined with the general pattern of coevolution between group II intron RNAs and IEPs, these findings coalesced into the retroelement ancestor hypothesis, which posits that all extant group II introns descend from an ancestral bacterial group II intron that behaved as a retroelement (Dai and Zimmerly 2002a). By extension, all other group II introns found naturally that are not autonomous retroelements represent various stages of degradation from this ancestral state.

Several hypotheses were put forward to explain the origin of group II introns, though the exact details remain speculative. Many groups have suggested that the RNA-based catalytic abilities of group II introns are a testament to their origin in a primordial RNA World (Gesteland et al. 2006), though conclusive evidence is still lacking (Doolittle 2013). However, an area of scientific consensus is that the ancestral group II intron form consisted of an initial pairing of a self-splicing RNA and an RT-bearing protein in bacteria (Toor et al. 2001). One possibility is that a self-splicing RNA evolved from ncRNA sequences flanking a primitive mobile element, such as a transposon (Curcio and Belfort 1996). This gradual evolution would have been beneficial to the transposon by reducing the negative impacts of transposition, and beneficial to the evolving RNA by promoting its dissemination. An alternative scenario is that self-splicing RNA would have evolved independently as a selfish genetic parasite, either in the RNA World or in bacteria, whose splicing evolved as a means of limiting damage to the host and promoting dissemination. Once in bacteria, one of the many bacterial RTs would associate with a primitive group II intron sequence, initially increasing mobility efficiency and then gradually evolving to also stabilize splicing (Wank et al. 1999). Some of these would have evolved to recognize secondary structures during mobility (group IIC introns), while others would rely more heavily on nucleotide sequences to spread to novel sites (group IIA and IIB introns). Through domain accretion, some group II intron IEPs would have extended to include a C-terminal H-N-H endonuclease (Gorbalenya 1994), which

would have increased mobility efficiency by eliminating the dependency on ssDNA. Some group IIA and IIB introns would have next spread laterally into archaea, where they would remain at low levels (Dai and Zimmerly 2003). The group II introns of mitochondria and chloroplasts, on the other hand, likely derive from group II introns of the ML and CL lineages that were in the α-proteobacterial and cyanobacterial ancestors of the eukaryotic organelles, respectively (Cavalier-Smith 1991).

1.5.2: Distribution of group II introns

As mentioned previously, group II introns are incredibly widespread genetic elements, spanning throughout bacteria, archaea and the organelles of eukaryotes, notably fungi, plants and several protists. Within each of these cellular compartments, they have evolved into very different forms, reaching various levels of specialization and degradation.

In the organelles of eukaryotes, some group II introns maintain an IEP and are mobile elements, such as the aI1 intron from the mitochondria of *Saccharomyces cerevisiae* (Eskes et al. 1997). However, most group II introns in mitochondria and chloroplasts have gradually become specialized as splicing-only elements. Indeed, although they are rarely fragmented, organellar group II introns are often severely degraded. The extent of their degradation is most apparent in a class of introns initially described in *Euglena gracilis* (Christopher and Hallick 1989). Termed group III introns, these genetic elements are extremely small, ranging in size from 95-109 nucleotides, and contain only DI and DVI as identifiable motifs. The notable absence of the catalytic DV from these elements suggests the presence of a common *trans*-acting machinery for group III introns, though none has yet been found. Moreover, most organellar group II introns have either lost their IEP or contain a severely degraded form of it. The loss of their IEP often means that they can no longer mobilize to new sites. They have thus gradually evolved a dependence on host-encoded chaperones/maturases, which nevertheless enable them to properly fold and splice (Brown et al. 2014). Finally, they are frequently associated with housekeeping genes, so mutations that increase their splicing efficiency undergo positive selection.

In bacteria, several facets of group II introns point to a very different type of adaptation to their environment, where they behave mainly as mobile retroelements (Dai and Zimmerly 2002a).

First, they are very rarely associated with housekeeping genes, more frequently interrupting intergenic regions or other mobile genetic elements (Klein and Dunny 2002). Their association with other mobile elements leads to frequent lateral transfer and a patchy distribution, where intron density can vary within a specific bacterial species and where a single bacterium can contain introns of multiple different classes (Candales et al. 2012). Next, group II introns are frequently fragmented, suggesting a more dynamic life cycle than in organelles, involving much higher rates of intron gain and loss: gain through insertions to new genetic loci and loss by fragmentation and degradation (Leclercq and Cordaux 2012).

Despite their widespread distribution, group II introns are all but absent from the organelles of Metazoa, with rare instances representing recent lateral transfer events from bacterial or viral vectors (Valles et al. 2008; Huchon et al. 2015). This likely suggests that the organelles of ancestral higher eukaryotes such as Metazoans shed their group II introns but that their presence is not overtly detrimental. However, one genetic compartment where group II introns have never been reported is the nuclei of eukaryotes, with only a few exceptions that likely represent inert sequences that were transferred together with other portions of mtDNA into the nucleus (Knoop and Brennicke 1994). This exclusion was experimentally demonstrated to reside with various aspects of nuclear transcript maturation that are incompatible with group II intron splicing (Truong et al. 2015). First, group II intron catalysis in bacteria relies on high intracellular levels of Magnesium (1-4mM), which are low in the nuclei of eukaryotes (0.2-1mM) and thus impede group II intron splicing (Mastroianni et al. 2008). This leads to the export of transcripts containing unspliced group II introns into the cytoplasm, where low-level splicing can occur. In cases where cytoplasmic self-splicing occurs, group II introns remain bound to their transcript through base pairing, which sterically hinders ribosome progression and thus inhibits translation. On the other hand, the majority of transcripts that are not transcribed are targeted for nonsense-mediated mRNA decay (Chalamcharla et al. 2010). Overall, these factors appear to have contributed to the complete functional exclusion of group II introns from the nuclei of eukaryotes.

1.5.3: Relationship between group II introns and bacterial hosts

The retroelement behaviour of group II introns in bacteria is a phenomenon that has likely evolved over millions of years, caused by the selective pressures these self-splicing retroelements face in this specific environment. The predominant view is that group II introns undergo net negative selection in bacteria over evolutionary timespans (Leclercq and Cordaux 2012). This contrasts with some other mobile elements such as group I introns in organelles, which are largely considered neutral. Such neutral mobile elements invade target sites until saturation, after which they gradually degrade through random point mutations until they yield immobile, splicing-only elements (Goddard and Burt 1999; Burt and Koufopanou 2004). In bacteria, group II introns are frequently found with abundant homing sites that remain unoccupied. Moreover, they are very rarely found within housekeeping genes, altogether suggesting that negative selective pressure is exercised against bacterial group II introns (Dai and Zimmerly 2002a).

The reasons for negative selection may in part be intron-specific, since some group II introns were shown to reduce the expression levels of the genes they interrupt (Chen et al. 2005). However, it is also likely to be a generalized pattern of selection against selfish mobile elements, as was shown to be the case for IS elements (Touchon and Rocha 2007). In such cases, bacterial genomes undergo dynamic processes of extinction and recolonization, where acquired mobile elements quickly proliferate and are rapidly removed (Wagner 2006). This is likely due to the underlying population genetics of bacterial populations (Lynch 2002). High bacterial population numbers lead to increased purifying selection and genome streamlining, such that even slightly deleterious genetic elements are removed over long periods of time. In contrast, small population sizes such as eukaryotes lead to reduced purifying selection and enhanced genetic drift, so mobile elements that are neutral or even slightly deleterious can be maintained and fixed (Le Rouzic et al. 2007).

Key regulators have been identified that play important roles in mediating the complex relationship between group II introns and their bacterial hosts. Often, these regulators act to keep intron levels low. These include RNaseE, the central component of the bacterial degradosome (Coros et al. 2008). This nuclease may act in concert with enolase, another component of the degradosome, to sense the metabolic state of the bacterium and accordingly prevent group II intron retromobility through degradation. However, other environmental sensor molecules are important

to indirectly enhance intron mobility by relaxing the structure of the bacterial nucleoid, thus increasing the frequency of ectopic retrotransposition events. These global regulators include ppGpp, which signals amino acid starvation, and cAMP that indicates glucose starvation (Coros et al. 2009). An increase in mobility can potentially benefit the bacterial host, by increasing genetic diversity through group II intron insertion events that, when abundant enough, can remodel genomes (Beauregard et al. 2008). In some cases, bacterial group II intron expansion has been extensive, such as in the *Wolbachia* obligate intracellular symbionts (Leclercq et al. 2011). For these organisms, frequent intron mobility has generated large-scale recombination events, which were likely beneficial in accelerating the process of reductive genome evolution. These organisms serve as an interesting example of uncontrolled group II intron proliferation, a phenomenon frequently ascribed to the early evolution of eukaryotes (Koonin 2006).

1.5.4: Group II intron-derived elements in eukaryotes

Despite their notable absence from the nuclei of eukaryotes, group II introns are nevertheless believed to have substantially shaped the origin and evolutionary trajectory of eukaryotes. The earliest eukaryotes are thought to have been formed by the symbiotic relationship that developed between an archaea that engulfed a primitive bacteria, which later became an obligate intracellular endosymbiont called mitochondria, in what is commonly referred to as the endosymbiotic theory of eukaryotic evolution (Sagan 1967). Comparative genomic analyses of extant eukaryotes suggest that intron numbers were very large early in their evolution, while subsequent eukaryotic intron evolution has mostly consisted of intron loss (Koonin 2009). These findings are consistent with an early invasion of primitive eukaryotic genes by group II introns, which would likely have already been contained within the genomes of primitive mitochondria and later chloroplasts (Cavalier-Smith 1991). This heavy invasion would have been fuelled by a reduction in purifying selection due to the initially small population sizes (Koonin 2006). Furthermore, uncontrolled group II intron mobility may have been a driving factor in the evolution of the nuclear membrane in primitive eukaryotes, which would have prevented further gene invasion and also allowed for a separation between transcription and translation, ensuring that all mRNA be intronless at the time of translation (Martin and Koonin 2006).

The most prominent genetic elements in eukaryotes believed to be derived from the historic burst of group II intron mobility are the nuclear introns themselves (Cech 1986). Indeed, these two classes of splicing elements have biochemically identical splicing pathways that yield lariat molecules. Moreover, eukaryotic introns have the same conserved 5′-GU and AY-3′ intron boundary residues as group II introns, and have a remarkable conservation of intron positions, even throughout different kingdoms, suggesting an intron-dense ancestral state (Rogozin et al. 2003). Single-celled eukaryotes likely reached sufficiently high population sizes to enable efficient intron removal through purifying selection, resulting in the intron-poor genomes of extant single-celled eukaryotes (Lynch 2002). However, other organisms such as vertebrates and plants consistently maintained low population sizes, resulting in overall intron conservation and in some cases even intron gain (Charlesworth 2009).

To ensure consistent and accurate splicing of the abundant introns of early eukaryotes, a common *trans*-acting machinery is believed to have evolved through group II intron fragmentation: the spliceosome (Sharp 1991). The snRNAs of the spliceosome are responsible for carrying out catalysis of all nuclear introns, assembling as five short nuclear RNAs (snRNAs U1, U2, U4, U5, U6) with the help of a supporting network as extensive as 170 proteins (Will and Luhrmann 2011). Within the five spliceosomal snRNAs, only U2, U5 and U6 are essential for the catalytic steps of both transesterification reactions, and these have functional equivalents within the domains of group II introns (Valadkhan 2013). Group II intron domains and snRNAs were shown to behave modularly, since functional substitution studies demonstrated that the catalytic DV could substitute U6 during nuclear splicing (Shukla and Padgett 2002), while U5 could substitute a portion of DI and catalyze group II intron splicing (Hetzer et al. 1997). Moreover, *in vivo* splicing experiments using group II introns has shown that fragmentation occurs both naturally and artificially, even though it only ever results in self-splicing due to reassembly (see Section 1.3.4). Finally, one of the most important proteins in the spliceosome, Prp8, bears extensive phylogenetic homology to the RT domains of group II intron IEPs (Dlakic and Mushegian 2011). Prp8 was shown to interact directly with U2, U5 and U6, helping generate the spliceosomal active site: a role analogous to the IEP of self-splicing group II introns (Galej et al. 2013).

A final class of elements believed to descend from group II introns are proteins containing homologous RT domains that have retained their functionality, notably non-LTR retroelements and the telomerase enzyme. Group II intron-encoded RTs contain seven conserved amino acid motifs corresponding to the fingers and palm region of viral retroelements (Fig. 1.1A) (Blocker et al. 2005). However, group II intron RTs also contain a conserved N-terminal motif extension (RT-0) (Fig. 1.1A), which is absent in retroviral RTs and yet is found in non-LTR retroelements such as LINE elements (Xiong and Eickbush 1990). Moreover, the conserved 2a insertion between amino acid motifs 2 and 3 (Fig. 1.1A) is shared between group II intron RTs, LINE elements and telomerase RTs (Lambowitz and Belfort 2015). Finally, both LINE elements and the telomerase RT target DNA and reverse transcribe using a conserved mechanism also shared with group II introns: target-primed reverse transcription (Zimmerly et al. 1995b).

1.5.5: Group II intron-derived elements in bacteria

Group II introns are also posited to share evolutionary relationships with a diverse set of genetic elements in prokaryotes harboring RT motifs (Zimmerly and Wu 2015). These include diversity-generating retroelements (DGRs), which introduce variation within a set of target genes through a process known as mutagenic retrohoming (Wu et al. 2018). The added genetic variation is often beneficial to the host, such as when the BPP-1 bacteriophage uses DGRs to switch the tropism of its tail fibers, ensuring consistent infection of *Bordetella* bacteria despite their highly variable cell surface (Liu et al. 2002). The RTs of DGRs also contain the 2a insertion found between conserved amino acid motifs 2 and 3, which as mentioned above is present in the RTs of group II introns and non-LTR retrotransposons (Fig. 1.1A), yet is absent in retroviral RTs (Malik et al. 1999). This suggests that the RTs of group II introns, non-LTR retrotransposons and DGRs form a distinct subclass with shared ancestry (Doulatov et al. 2004).

The RTs of group II introns have also been proposed to be evolutionarily related to RTs involved in mechanisms of host defense, such as those contained in certain CRISPR-Cas systems (McNeil et al. 2016). These bacterial defense mechanisms are responsible for integrating new spacer elements into the CRISPR array, providing subsequent immunity against infectious elements that encode identical sequences (Nunez et al. 2014). Integration of novel spacer elements

is achieved in part using *Cas1* proteins, which are universally present in all Crispr-Cas systems and have associated RT motifs termed group II-like proteins 1 and 2 (G2L1 and G2L2), either present as stand-alone ORFs or fusion constructs (Simon and Zimmerly 2008). Due to the high degree of similarity between these RT motifs, it is tempting to think that the integration of new spacer elements might occur through a mechanism resembling the target-primed reverse-transcription used during group II intron retromobility, though this remains unclear (Zimmerly et al. 1995b).

1.6: Ll.LtrB: a model group II intron from *Lactococcus lactis*

The discovery of group II introns in bacteria led to large strides in studying various facets of their splicing and mobility (Ferat and Michel 1993). Ll.LtrB rapidly emerged as a model bacterial system when it was demonstrated to be the first group II intron to self-splice *in vivo*. This group IIA intron is 2942 nucleotides and harbors a single, 599 amino acid IEP in DIV termed *ltrA* (Fig. 1.1). Ll.LtrB was first discovered in the pRS01 conjugative plasmid of *Lactococcus lactis*, a gram-positive bacterium, as a genetic element whose integrity was essential for the successful conjugative transfer of pRS01 (Mills et al. 1994). Only later was it understood that disrupting this autonomous group II intron prevented its self-splicing from genes involved in conjugation, causing drastic reductions in the conjugative transfer of the pRS01 plasmid (Mills et al. 1996).

Since its discovery, Ll.LtrB has been shown to self-splice and mobilize in a number of different cellular environments such as *E. coli* (Matsuura et al. 1997), where it has been used to characterize bacterial mobility pathways (see Section 1.4) (Cousineau et al. 1998). Moreover, this retroelement has emerged as the model representative of group IIA introns, and crystallographic studies have detailed exactly how Ll.LtrB uses its secondary and tertiary interactions to fold and self-splice (Qu et al. 2016). Ll.LtrB was also used to functionally address the ability of fragmented group II introns to reassemble and self-splice *in vivo* (see Section 1.3.4), which supported a longstanding evolutionary theory that group II intron fragmentation gave rise to the catalytic snRNAs of the spliceosome (see Section 1.5.4) (Sharp 1991). In the following sections, I will outline how Ll.LtrB has specifically been used as a model system to address evolutionary and functional aspects of group II introns.

1.6.1: Ll.LtrB and group II intron lateral transfer

In its native environment, the Ll.LtrB bacterial group IIA intron interrupts the *ltrB* relaxase gene of several mobile genetic elements in *L. lactis*, including plasmids such as pAH90 (O' Sullivan et al. 2001) and pRS01 (Mills et al. 1996), but also larger elements frequently embedded within the bacterial chromosome such as Sex Factor, an integrative and conjugative element (Shearman et al. 1996). Since its initial discovery, 60 copies of Ll.LtrB and slight Ll.LtrB variants (>95% nucleotide similarity) have been reported throughout strains and sub-species of *L. lactis*, over 50 of which are in putative relaxase genes (Candales et al. 2012).

Relaxases are a group of endonuclease enzymes that nick conjugative elements at their origin of transfer (*oriT*) (Smillie et al. 2010). This is achieved using a conserved tyrosine residue near the N-terminus of relaxase enzymes, which forms a covalent phosphodiester bond with the released 5′ phosphate of *oriT* (Byrd and Matson 1997). Nicking is only achieved when the relaxase enzyme is recruited to *oriT* by the conjugative relaxosome, a protein complex that assembles around *oriT*. In *L. lactis*, these consist of lactococcal transfer genes (ltr), some of which are disposable and thus likely only serve to strengthen the relaxosome complex (*ltrC* and *ltrD*), while others are essential to recruit the relaxase to *oriT* (*ltrF*) (Chen et al. 2007). Once nicking is completed, the relaxosome-*oriT* complex is next trafficked to the bacterial membrane, where the relaxase interacts with the all-alpha domain of a type-4 coupling protein ATPase (T4CP) (Whitaker et al. 2015). The relaxase-*oriT* complex is then actively transferred by the T4CP into the mating pore, a type-4 secretion system (T4SS) that spans the cell membranes of both the donor and recipient cells (Goessweiner-Mohr et al. 2014). Conjugation is deemed successful when the conjugative element is fully transferred into the recipient cell, after which the relaxase reverses the transesterification at *oriT* and second-strand DNA synthesis occurs (Llosa et al. 2002).

The presence of Ll.LtrB within a relaxase gene has several implications for its function and its evolutionary relationship with *L. lactis*. By some measures, the association of *ltrB* with the Ll.LtrB group II intron can be thought of as antagonistic, since Ll.LtrB reduces the amount of *ltrB* transcripts that are translated (Chen et al. 2005) and degrades certain ligated *ltrB* exons by targeting them for hydrolytic SER (Fig. 1.3) (Qu et al. 2018). The enzyme and invading intron can also act

synergistically, since LtrB was shown to randomly nick chromosomal DNA in *L. lactis*, providing targets for Ll.LtrB retrotransposition and thus increasing intron mobility (Novikova et al. 2014). Yet as previously mentioned, the greatest impact of Ll.LtrB is on the process of conjugation itself. Ll.LtrB interrupts a histidine triad in the catalytic core of LtrB, which makes splicing essential for proper function of the enzyme and conjugation to occur. This pairing of self-splicing and conjugation has led to the development of functional assays which measure self-splicing quantitatively, using conjugation efficiency as a sensitive output (Klein et al. 2004).

The link between conjugation and Ll.LtrB was also used to functionally address routes of group II intron dispersal. The natural distribution of group II introns is often patchy, where natural populations of a single bacterial species can contain variable copy numbers of both identical group II introns and group II introns belonging to different subclasses, suggesting transmission by horizontal transfer (Dai and Zimmerly 2002b; Tourasse and Kolsto 2008). Ll.LtrB was experimentally demonstrated to transfer laterally by conjugation to other strains of *L. lactis* (intra-species) and to *Enterococcus faecalis* (inter-species), after which the intron could mobilize to new sites through both retrohoming and retrotransposition (Belhocine et al. 2004). Initially demonstrated for a shuttle vector harboring a segment of the pRS01 plasmid, conjugative transfer of Ll.LtrB was also shown for the chromosomal Sex Factor (Belhocine et al. 2005) and later for the full pRS01 plasmid itself (Belhocine et al. 2007b). Overall, conjugation thus appears to be an important method for group II intron horizontal transfer, also extending to other intron subtypes such as the RmInt1 group II intron (Nisa-Martinez et al. 2007). Moreover, the precise insertion site of Ll.LtrB is proposed to be beneficial for its dissemination. The *ltrB* catalytic triad interrupted by Ll.LtrB is a conserved motif in the IncP family of conjugative relaxases (Pansegrau et al. 1994). Upon arriving in a new bacterium through horizontal transfer, Ll.LtrB was previously shown to recognize this motif in orthologous relaxases and invade them by retrohoming, suggesting that an abundance of conserved mobility sites exist for this group II intron throughout bacteria (Staddon et al. 2004). Although Ll.LtrB has emerged as a powerful system to study group II intron lateral transfer, very little remains known about subsequent selective pressures shaping group II intron evolution in a novel environment (see Chapter 2).

1.6.2: Ll.LtrB as a model system to study group II intron circularization

Ll.LtrB has recently been used as a model system to study an alternative group II intron splicing pathway: circularization (see Section 1.3.3). Although Ll.LtrB-WT was shown to concurrently self-splice as intron lariats and circles *in vivo* (Monat et al. 2015), very little is still known about both the mechanism and purpose of circularization. A model system has emerged for studying group II intron circles in *L. lactis*, since removal of the Ll.LtrB branchpoint adenosine residue (Ll.LtrB-ΔA) results in exclusive self-splicing through the circularization pathway, generating circles with an additional C residue at the splice junction (Monat et al. 2015). Substitution assays demonstrated that the additional C residue originates from imprecise nucleophilic attack of exon 1 at the 3′ splice site during the first transesterification attack (Fig. 1.3C), where the first C of exon 2 remains bound to the intron 3′ end and attacks the 5′ splice site (Monat and Cousineau 2016). Amplification of the Ll.LtrB-ΔA circle splice junction combined with functional assays linking self-splicing to conjugation nevertheless demonstrated that the branchpoint mutant also generates perfect head-to-tail intron circles (Monat and Cousineau 2016). Removing the bulged adenosine residue thus tilts the balance of the first transesterification reaction entirely towards circularization, with a bias to attack the first nucleotide of E2. Overall, Ll.LtrB-ΔA has emerged as a robust system to dissect various aspects of the circularization pathway.

A longstanding question regarding group II intron circles has been the presence of short sequences of additional nucleotides at their circle splice junctions. Initially described by Murray and colleagues for the aI5γ group II intron, these were proposed to be artefacts generated by the reverse-transcriptase encountering a 2′-5′ linkage at the circle splice junction (Murray et al. 2001). The *nad1* intron 2 was later reported to contain longer tracts of 7 nucleotides at its circle splice junction, whose origin was unknown and too short to be reliably identified (Li-Pook-Than and Bonen 2006). Recently, Ll.LtrB was shown to generate circles with even longer stretches of nucleotides at their splice junctions, corresponding to the ribosomal protein L21 gene (31 nucleotides) of the bacterial chromosome and to the chloramphenicol resistance gene (27 nucleotides) of the shuttle vector used to express the group II intron (Monat et al. 2015). Overall, the origin and mechanism through which group II introns incorporate additional nucleotides at their splice junctions remains unknown, perhaps pointing to the presence of a different splicing pathway altogether (see Chapter 3).

1.7: Objectives of the thesis

Group II introns are proposed to have had an enormous impact on the evolution of eukaryotes. However, the rapid evolution of these ribozymes and their low degree of sequence conservation render it difficult to adequately understand the selective forces shaping their current evolution in bacteria (see Section 1.5). Moreover, they have given rise to a plethora of functional genetic elements in bacteria and eukaryotes, most of which are beneficial to their host (see Sections 1.5.4, 1.5.5). Despite the functional versatility of these group II intron-derived elements, group II introns themselves are still considered solely as selfish genetic parasites that provide no functional benefit to their host.

We thus chose to use Ll.LtrB as a model system to address outstanding questions regarding both the evolution and function of group II introns. We began by studying a group II intron recently discovered by our lab within the genome of *Enterococcus faecalis*. This group II intron is nearly identical to Ll.LtrB (99.7% nucleotide identity), yet is present in a different bacterial species, likely representing a natural case of recent horizontal transfer. We thus chose this comparison of two group II introns as a model to study how group II introns adapt to novel cellular environments and to determine which selective pressures affect their subsequent evolution (Chapter 2). Next, we made use of Ll.LtrB-WT, Ll.LtrB-ΔA and several other Ll.LtrB mutants to elucidate the precise mechanism used by our model group II intron to generate circular RNAs with additional nucleotides at their splice junctions (Chapter 3). Finally, we used the native biological context of Ll.LtrB to assess the functional value of the newly described splicing pathway, using conjugation as a functional output to demonstrate how group II introns can be beneficial to their hosts (Chapter 4).

1.8: References

Aizawa Y, Xiang Q, Lambowitz AM, Pyle AM. 2003. The pathway for DNA recognition and RNA integration by a group II intron retrotransposon. *Mol Cell* 11(3):795-805.

Arnberg AC, Van Ommen GJ, Grivell LA, Van Bruggen EF, Borst P. 1980. Some yeast mitochondrial RNAs are circular. *Cell* 19(2):313-9.

Augustin S, Muller MW, Schweyen RJ. 1990. Reverse self-splicing of group II intron RNAs in vitro. *Nature* 343(6256):383-6.

Avery OT, Macleod CM, McCarty M. 1944. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med* 79(2):137-58.

Beadle GW, Tatum EL. 1941. Genetic Control of Biochemical Reactions in Neurospora. *Proc Natl Acad Sci U S A* 27(11):499-506.

Beauregard A, Curcio MJ, Belfort M. 2008. The take and give between retrotransposable elements and their hosts. *Annu Rev Genet* 42:587-617.

Begel O, Boulay J, Albert B, Dufour E, Sainsard-Chanet A. 1999. Mitochondrial group II introns, cytochrome c oxidase, and senescence in Podospora anserina. *Mol Cell Biol* 19(6):4093-100.

Belhocine K, Plante I, Cousineau B. 2004. Conjugation mediates transfer of the Ll.LtrB group II intron between different bacterial species. *Mol Microbiol* 51(5):1459-69.

Belhocine K, Yam KK, Cousineau B. 2005. Conjugative transfer of the Lactococcus lactis chromosomal sex factor promotes dissemination of the Ll.LtrB group II intron. *J Bacteriol* 187(3):930-9.

Belhocine K, Mak AB, Cousineau B. 2007a. Trans-splicing of the Ll.LtrB group II intron in Lactococcus lactis. *Nucleic Acids Res* 35(7):2257-68.

Belhocine K, Mandilaras V, Yeung B, Cousineau B. 2007b. Conjugative transfer of the Lactococcus lactis sex factor and pRS01 plasmid to Enterococcus faecalis. *FEMS Microbiol Lett* 269(2):289-94.

Belhocine K, Mak AB, Cousineau B. 2008. Trans-splicing versatility of the Ll.LtrB group II intron. *RNA* 14(9):1782-90.

Benzer S. 1959. On the Topology of the Genetic Fine Structure. *Proc Natl Acad Sci U S A* 45(11):1607-20.

Berget SM, Moore C, Sharp PA. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A* 74(8):3171-5.

Blocker FJ, Mohr G, Conlan LH, Qi L, Belfort M, Lambowitz AM. 2005. Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. *RNA* 11(1):14-28.

Bonen L. 1993. Trans-splicing of pre-mRNA in plants, animals, and protists. *FASEB J* 7(1):40-6.

Bonen L, Vogel J. 2001. The ins and outs of group II introns. *Trends Genet* 17(6):322-31.

Borst P, Grivell LA. 1978. The mitochondrial genome of yeast. *Cell* 15(3):705-23.

Boudvillain M, de Lencastre A, Pyle AM. 2000. A tertiary interaction that links active-site domains to the 5' splice site of a group II intron. *Nature* 406(6793):315-8.

Brown GG, Colas des Francs-Small C, Ostersetzer-Biran O. 2014. Group II intron splicing factors in plant mitochondria. *Front Plant Sci* 5:35.

Burt A, Koufopanou V. 2004. Homing endonuclease genes: the rise and fall and rise again of a selfish element. *Curr Opin Genet Dev* 14(6):609-15.
Byrd DR, Matson SW. 1997. Nicking by transesterification: the reaction catalysed by a relaxase. *Mol Microbiol* 25(6):1011-22.
Candales MA, Duong A, Hood KS, Li T, Neufeld RA, Sun R, McNeil BA, Wu L, Jarding AM, Zimmerly S. 2012. Database for bacterial group II introns. *Nucleic Acids Res* 40(Database issue):D187-90.
Cavalier-Smith T. 1991. Intron phylogeny: a new hypothesis. *Trends Genet* 7(5):145-8.
Cech TR, Zaug AJ, Grabowski PJ. 1981. In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* 27(3 Pt 2):487-96.
Cech TR. 1986. The generality of self-splicing RNA: relationship to nuclear mRNA splicing. *Cell* 44(2):207-10.
Chalamcharla VR, Curcio MJ, Belfort M. 2010. Nuclear expression of a group II intron is consistent with spliceosomal intron ancestry. *Genes Dev* 24(8):827-36.
Chanfreau G, Jacquier A. 1994. Catalytic site components common to both splicing steps of a group II intron. *Science* 266(5189):1383-7.
Chanfreau G, Jacquier A. 1996. An RNA conformational change between the two chemical steps of group II self-splicing. *EMBO J* 15(13):3466-76.
Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10(3):195-205.
Chen Y, Klein JR, McKay LL, Dunny GM. 2005. Quantitative analysis of group II intron expression and splicing in Lactococcus lactis. *Appl Environ Microbiol* 71(5):2576-86.
Chen Y, Staddon JH, Dunny GM. 2007. Specificity determinants of conjugative DNA processing in the Enterococcus faecalis plasmid pCF10 and the Lactococcus lactis plasmid pRS01. *Mol Microbiol* 63(5):1549-64.
Chin K, Pyle AM. 1995. Branch-point attack in group II introns is a highly reversible transesterification, providing a potential proofreading mechanism for 5'-splice site selection. *RNA* 1(4):391-406.
Choquet Y, Goldschmidt-Clermont M, Girard-Bascou J, Kuck U, Bennoun P, Rochaix JD. 1988. Mutant phenotypes support a trans-splicing mechanism for the expression of the tripartite psaA gene in the C. reinhardtii chloroplast. *Cell* 52(6):903-13.
Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* 12(1):1-8.
Christopher DA, Hallick RB. 1989. Euglena gracilis chloroplast ribosomal protein operon: a new chloroplast gene for ribosomal protein L5 and description of a novel organelle intron category designated group III. *Nucleic Acids Res* 17(19):7591-608.
Chu VT, Adamidi C, Liu Q, Perlman PS, Pyle AM. 2001. Control of branch-site choice by a group II intron. *EMBO J* 20(23):6866-76.
Collins L, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol* 22(4):1053-66.
Coros CJ, Landthaler M, Piazza CL, Beauregard A, Esposito D, Perutka J, Lambowitz AM, Belfort M. 2005. Retrotransposition strategies of the Lactococcus lactis Ll.LtrB group II intron are dictated by host identity and cellular environment. *Mol Microbiol* 56(2):509-24.
Coros CJ, Piazza CL, Chalamcharla VR, Belfort M. 2008. A mutant screen reveals RNase E as a silencer of group II intron retromobility in Escherichia coli. *RNA* 14(12):2634-44.

Coros CJ, Piazza CL, Chalamcharla VR, Smith D, Belfort M. 2009. Global regulators orchestrate group II intron retromobility. *Mol Cell* 34(2):250-6.
Costa M, Deme E, Jacquier A, Michel F. 1997. Multiple tertiary interactions involving domain II of group II self-splicing introns. *J Mol Biol* 267(3):520-36.
Cousineau B, Smith D, Lawrence-Cavanagh S, Mueller JE, Yang J, Mills D, Manias D, Dunny G, Lambowitz AM, Belfort M. 1998. Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell* 94(4):451-62.
Cousineau B, Lawrence S, Smith D, Belfort M. 2000. Retrotransposition of a bacterial group II intron. *Nature* 404(6781):1018-21.
Crick FH. 1958. On protein synthesis. *Symp Soc Exp Biol* 12:138-63.
Curcio MJ, Belfort M. 1996. Retrohoming: cDNA-mediated mobility of group II introns requires a catalytic RNA. *Cell* 84(1):9-12.
Dai L, Zimmerly S. 2002a. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res* 30(5):1091-102.
Dai L, Zimmerly S. 2002b. The dispersal of five group II introns among natural populations of Escherichia coli. *RNA* 8(10):1294-307.
Dai L, Zimmerly S. 2003. ORF-less and reverse-transcriptase-encoding group II introns in archaebacteria, with a pattern of homing into related group II intron ORFs. *RNA* 9(1):14-9.
Daniels DL, Michels WJ, Jr., Pyle AM. 1996. Two competing pathways for self-splicing by group II introns: a quantitative analysis of in vitro reaction rates and products. *J Mol Biol* 256(1):31-49.
de Lencastre A, Pyle AM. 2008. Three essential and conserved regions of the group II intron are proximal to the 5'-splice site. *RNA* 14(1):11-24.
Dlakic M, Mushegian A. 2011. Prp8, the pivotal protein of the spliceosomal catalytic center, evolved from a retroelement-encoded reverse transcriptase. *RNA* 17(5):799-808.
Doolittle WF. 2013. The spliceosomal catalytic core arose in the RNA world... or did it? *Genome Biol* 14(12):141.
Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, Deora R, Simons RW, Zimmerly S, Miller JF. 2004. Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. *Nature* 431(7007):476-81.
Du WW, Yang W, Chen Y, Wu ZK, Foster FS, Yang Z, Li X, Yang BB. 2017. Foxo3 circular RNA promotes cardiac senescence by modulating multiple factors associated with stress and senescence responses. *Eur Heart J* 38(18):1402-1412.
Eskes R, Yang J, Lambowitz AM, Perlman PS. 1997. Mobility of yeast mitochondrial group II introns: engineering a new site specificity and retrohoming via full reverse splicing. *Cell* 88(6):865-74.
Fedorova O, Mitros T, Pyle AM. 2003. Domains 2 and 3 interact to form critical elements of the group II intron active site. *J Mol Biol* 330(2):197-209.
Fedorova O, Pyle AM. 2005. Linking the group II intron catalytic domains: tertiary contacts and structural features of domain 3. *EMBO J* 24(22):3906-16.
Fedorova O, Zingler N. 2007. Group II introns: structure, folding and splicing mechanism. *Biol Chem* 388(7):665-78.
Ferat JL, Michel F. 1993. Group II self-splicing introns in bacteria. *Nature* 364(6435):358-61.

Fernandez-Lopez M, Munoz-Adelantado E, Gillis M, Willems A, Toro N. 2005. Dispersal and evolution of the Sinorhizobium meliloti group II RmInt1 intron in bacteria that interact with plants. *Mol Biol Evol* 22(6):1518-28.

Field DJ, Sommerfield A, Saville BJ, Collins RA. 1989. A group II intron in the Neurospora mitochondrial coI gene: nucleotide sequence and implications for splicing and molecular evolution. *Nucleic Acids Res* 17(22):9087-99.

Galej WP, Oubridge C, Newman AJ, Nagai K. 2013. Crystal structure of Prp8 reveals active site cavity of the spliceosome. *Nature* 493(7434):638-43.

Gaur RK, McLaughlin LW, Green MR. 1997. Functional group substitutions of the branchpoint adenosine in a nuclear pre-mRNA and a group II intron. *RNA* 3(8):861-9.

Gesteland RF, Cech T, Atkins JF. 2006. The RNA world : the nature of modern RNA suggests a prebiotic RNA world. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Gilbert W. 1978. Why genes in pieces? *Nature* 271(5645):501.

Goddard MR, Burt A. 1999. Recurrent invasion and extinction of a selfish gene. *Proc Natl Acad Sci U S A* 96(24):13880-5.

Goessweiner-Mohr N, Arends K, Keller W, Grohmann E. 2014. Conjugation in Gram-Positive Bacteria. *Microbiol Spectr* 2(4):PLAS-0004-2013.

Gorbalenya AE. 1994. Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family. *Protein Sci* 3(7):1117-20.

Grabowski PJ, Zaug AJ, Cech TR. 1981. The intervening sequence of the ribosomal RNA precursor is converted to a circular RNA in isolated nuclei of Tetrahymena. *Cell* 23(2):467-76.

Granlund M, Michel F, Norgren M. 2001. Mutually exclusive distribution of IS1548 and GBSi1, an active group II intron identified in human isolates of group B streptococci. *J Bacteriol* 183(8):2560-9.

Guo H, Zimmerly S, Perlman PS, Lambowitz AM. 1997. Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA. *EMBO J* 16(22):6835-48.

Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J. 2013. Natural RNA circles function as efficient microRNA sponges. *Nature* 495(7441):384-8.

Hetzer M, Wurzer G, Schweyen RJ, Mueller MW. 1997. Trans-activation of group II intron splicing by nuclear U5 snRNA. *Nature* 386(6623):417-20.

Horowitz NH. 1948. The one gene-one enzyme hypothesis. *Genetics* 33(6):612.

Huchon D, Szitenberg A, Shefer S, Ilan M, Feldstein T. 2015. Mitochondrial group I and group II introns in the sponge orders Agelasida and Axinellida. *BMC Evol Biol* 15:278.

Ichiyanagi K, Beauregard A, Lawrence S, Smith D, Cousineau B, Belfort M. 2002. Retrotransposition of the Ll.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. *Mol Microbiol* 46(5):1259-72.

Ichiyanagi K, Beauregard A, Belfort M. 2003. A bacterial group II intron favors retrotransposition into plasmid targets. *Proc Natl Acad Sci U S A* 100(26):15742-7.

Jacquier A, Michel F. 1987. Multiple exon-binding sites in class II self-splicing introns. *Cell* 50(1):17-29.

Jacquier A. 1990. Self-splicing group II and nuclear pre-mRNA introns: how similar are they? *Trends Biochem Sci* 15(9):351-4.

Jacquier A, Michel F. 1990. Base-pairing interactions involving the 5' and 3'-terminal nucleotides of group II self-splicing introns. *J Mol Biol* 213(3):437-47.

Jarrell KA, Dietrich RC, Perlman PS. 1988a. Group II intron domain 5 facilitates a trans-splicing reaction. *Mol Cell Biol* 8(6):2361-6.
Jarrell KA, Peebles CL, Dietrich RC, Romiti SL, Perlman PS. 1988b. Group II intron self-splicing. Alternative reaction conditions yield novel products. *J Biol Chem* 263(7):3432-9.
Keating KS, Toor N, Perlman PS, Pyle AM. 2010. A structural analysis of the group II intron active site and implications for the spliceosome. *RNA* 16(1):1-9.
Kelchner SA. 2002. Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. *Am J Bot* 89(10):1651-69.
Kennell JC, Moran JV, Perlman PS, Butow RA, Lambowitz AM. 1993. Reverse transcriptase activity associated with maturase-encoding group II introns in yeast mitochondria. *Cell* 73(1):133-46.
Klein JR, Dunny GM. 2002. Bacterial group II introns and their association with mobile genetic elements. *Front Biosci* 7:d1843-56.
Klein JR, Chen Y, Manias DA, Zhuo J, Zhou L, Peebles CL, Dunny GM. 2004. A conjugation-based system for genetic analysis of group II intron splicing in Lactococcus lactis. *J Bacteriol* 186(7):1991-8.
Knoop V, Brennicke A. 1994. Promiscuous mitochondrial group II intron sequences in plant nuclear genomes. *J Mol Evol* 39(2):144-50.
Knoop V, Altwasser M, Brennicke A. 1997. A tripartite group II intron in mitochondria of an angiosperm plant. *Mol Gen Genet* 255(3):269-76.
Koch JL, Boulanger SC, Dib-Hajj SD, Hebbar SK, Perlman PS. 1992. Group II introns deleted for multiple substructures retain self-splicing activity. *Mol Cell Biol* 12(5):1950-8.
Kohchi T, Umesono K, Ogura Y, Komine Y, Nakahigashi K, Komano T, Yamada Y, Ozeki H, Ohyama K. 1988. A nicked group II intron and trans-splicing in liverwort, Marchantia polymorpha, chloroplasts. *Nucleic Acids Res* 16(21):10025-36.
Koonin EV. 2006. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct* 1:22.
Koonin EV. 2009. Intron-dominated genomes of early ancestors of eukaryotes. *J Hered* 100(5):618-23.
Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* 31(1):147-57.
Lambowitz AM. 1989. Infectious introns. *Cell* 56(3):323-6.
Lambowitz AM, Belfort M. 1993. Introns as mobile genetic elements. *Annu Rev Biochem* 62:587-622.
Lambowitz AM, Zimmerly S. 2011. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol* 3(8):a003616.
Lambowitz AM, Belfort M. 2015. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol Spectr* 3(1).
Lasda E, Parker R. 2014. Circular RNAs: diversity of form and function. *RNA* 20(12):1829-42.
Lasda EL, Blumenthal T. 2011. Trans-splicing. *Wiley Interdiscip Rev RNA* 2(3):417-34.
Lazowska J, Meunier B, Macadre C. 1994. Homing of a group II intron in yeast mitochondrial DNA is accompanied by unidirectional co-conversion of upstream-located markers. *EMBO J* 13(20):4963-72.
Le Rouzic A, Boutin TS, Capy P. 2007. Long-term evolution of transposable elements. *Proc Natl Acad Sci U S A* 104(49):19375-80.

Leclercq S, Giraud I, Cordaux R. 2011. Remarkable abundance and evolution of mobile group II introns in Wolbachia bacterial endosymbionts. *Mol Biol Evol* 28(1):685-97.
Leclercq S, Cordaux R. 2012. Selection-driven extinction dynamics for group II introns in Enterobacteriales. *PLoS One* 7(12):e52268.
Leon G, Roy PH. 2009. Group IIC intron mobility into attC sites involves a bulged DNA stem-loop motif. *RNA* 15(8):1543-53.
Li-Pook-Than J, Bonen L. 2006. Multiple physical forms of excised group II intron RNAs in wheat mitochondria. *Nucleic Acids Res* 34(9):2782-90.
Li CF, Costa M, Bassi G, Lai YK, Michel F. 2011. Recurrent insertion of 5'-terminal nucleotides and loss of the branchpoint motif in lineages of group II introns inserted in mitochondrial preribosomal RNAs. *RNA* 17(7):1321-35.
Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, Preston A, Maskell DJ, Simons RW, Cotter PA, Parkhill J et al. . 2002. Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. *Science* 295(5562):2091-4.
Llosa M, Gomis-Ruth FX, Coll M, de la Cruz Fd F. 2002. Bacterial conjugation: a two-step mechanism for DNA transport. *Mol Microbiol* 45(1):1-8.
Lynch M. 2002. Intron evolution as a population-genetic process. *Proc Natl Acad Sci U S A* 99(9):6118-23.
Malik HS, Burke WD, Eickbush TH. 1999. The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol* 16(6):793-805.
Marcia M, Pyle AM. 2012. Visualizing group II intron catalysis through the stages of splicing. *Cell* 151(3):497-507.
Marcia M, Pyle AM. 2014. Principles of ion recognition in RNA: insights from the group II intron structures. *RNA* 20(4):516-27.
Martin W, Koonin EV. 2006. Introns and the origin of nucleus-cytosol compartmentalization. *Nature* 440(7080):41-5.
Martinez-Abarca F, Garcia-Rodriguez FM, Toro N. 2000. Homing of a bacterial group II intron with an intron-encoded protein lacking a recognizable endonuclease domain. *Mol Microbiol* 35(6):1405-12.
Martinez-Abarca F, Toro N. 2000. RecA-independent ectopic transposition in vivo of a bacterial group II intron. *Nucleic Acids Res* 28(21):4397-402.
Martinez-Abarca F, Barrientos-Duran A, Fernandez-Lopez M, 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 Res* 32(9):2880-8.
Mastroianni M, Watanabe K, White TB, Zhuang F, Vernon J, Matsuura M, Wallingford J, Lambowitz AM. 2008. Group II intron-based gene targeting reactions in eukaryotes. *PLoS One* 3(9):e3121.
Matsuura M, Saldanha R, Ma H, Wank H, Yang J, Mohr G, Cavanagh S, Dunny GM, Belfort M, Lambowitz AM. 1997. A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes Dev* 11(21):2910-24.
McNeil BA, Semper C, Zimmerly S. 2016. Group II introns: versatile ribozymes and retroelements. *Wiley Interdiscip Rev RNA*.
Mercereau-Puijalon O, Kourilsky P. 1979. Introns in the chicken ovalbumin gene prevent ovalbumin synthesis in E. coli K12. *Nature* 279(5714):647-9.

Michel F, Jacquier A, Dujon B. 1982. Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. *Biochimie* 64(10):867-81.
Michel F, Dujon B. 1983. Conservation of RNA secondary structures in two intron families including mitochondrial-, chloroplast- and nuclear-encoded members. *EMBO J* 2(1):33-8.
Michel F, Umesono K, Ozeki H. 1989. Comparative and functional anatomy of group II catalytic introns--a review. *Gene* 82(1):5-30.
Michel F, Ferat JL. 1995. Structure and activities of group II introns. *Annu Rev Biochem* 64:435-61.
Mikheeva S, Murray HL, Zhou H, Turczyk BM, Jarrell KA. 2000. Deletion of a conserved dinucleotide inhibits the second step of group II intron splicing. *RNA* 6(11):1509-15.
Mills DA, Choi CK, Dunny GM, McKay LL. 1994. Genetic analysis of regions of the Lactococcus lactis subsp. lactis plasmid pRS01 involved in conjugative transfer. *Appl Environ Microbiol* 60(12):4413-20.
Mills DA, McKay LL, Dunny GM. 1996. Splicing of a group II intron involved in the conjugative transfer of pRS01 in lactococci. *J Bacteriol* 178(12):3531-8.
Mohr G, Perlman PS, Lambowitz AM. 1993. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Res* 21(22):4991-7.
Molina-Sanchez MD, Martinez-Abarca F, Toro N. 2006. Excision of the Sinorhizobium meliloti group II intron RmInt1 as circles in vivo. *J Biol Chem* 281(39):28737-44.
Molina-Sanchez MD, Martinez-Abarca F, Toro N. 2010. Structural features in the C-terminal region of the Sinorhizobium meliloti RmInt1 group II intron-encoded protein contribute to its maturase and intron DNA-insertion function. *FEBS J* 277(1):244-54.
Monachello D, Michel F, Costa M. 2016. Activating the branch-forming splicing pathway by reengineering the ribozyme component of a natural group II intron. *RNA* 22(3):443-55.
Monat C, Quiroga C, Laroche-Johnston F, Cousineau B. 2015. The Ll.LtrB intron from Lactococcus lactis excises as circles in vivo: insights into the group II intron circularization pathway. *RNA* 21(7):1286-93.
Monat C, Cousineau B. 2016. Circularization pathway of a bacterial group II intron. *Nucleic Acids Res* 44(4):1845-53.
Mooers AO, Holmes EC. 2000. The evolution of base composition and phylogenetic inference. *Trends Ecol Evol* 15(9):365-369.
Morl M, Schmelzer C. 1990. Integration of group II intron bI1 into a foreign RNA by reversal of the self-splicing reaction in vitro. *Cell* 60(4):629-36.
Murray HL, Mikheeva S, Coljee VW, Turczyk BM, Donahue WF, Bar-Shalom A, Jarrell KA. 2001. Excision of group II introns as circles. *Mol Cell* 8(1):201-11.
Nisa-Martinez R, Jimenez-Zurdo JI, Martinez-Abarca F, Munoz-Adelantado E, Toro N. 2007. Dispersion of the RmInt1 group II intron in the Sinorhizobium meliloti genome upon acquisition by conjugative transfer. *Nucleic Acids Res* 35(1):214-22.
Novikova O, Smith D, Hahn I, Beauregard A, Belfort M. 2014. Interaction between conjugative and retrotransposable elements in horizontal gene transfer. *PLoS Genet* 10(12):e1004853.
Nunez JK, Kranzusch PJ, Noeske J, Wright AV, Davies CW, Doudna JA. 2014. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol* 21(6):528-34.

O' Sullivan D, Ross RP, Twomey DP, Fitzgerald GF, Hill C, Coffey A. 2001. Naturally occurring lactococcal plasmid pAH90 links bacteriophage resistance and mobility functions to a food-grade selectable marker. *Appl Environ Microbiol* 67(2):929-37.

Ohman-Heden M, Ahgren-Stalhandske A, Hahne S, Sjoberg BM. 1993. Translation across the 5'-splice site interferes with autocatalytic splicing. *Mol Microbiol* 7(6):975-82.

Osiewacz HD, Esser K. 1984. The mitochondrial plasmid of Podospora anserina: A mobile intron of a mitochondrial gene. *Curr Genet* 8(4):299-305.

Padgett RA, Konarska MM, Grabowski PJ, Hardy SF, Sharp PA. 1984. Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. *Science* 225(4665):898-903.

Pansegrau W, Schroder W, Lanka E. 1994. Concerted action of three distinct domains in the DNA cleaving-joining reaction catalyzed by relaxase (TraI) of conjugative plasmid RP4. *J Biol Chem* 269(4):2782-9.

Peebles CL, Perlman PS, Mecklenburg KL, Petrillo ML, Tabor JH, Jarrell KA, Cheng HL. 1986. A self-splicing RNA excises an intron lariat. *Cell* 44(2):213-23.

Podar M, Dib-Hajj S, Perlman PS. 1995. A UV-induced, Mg(2+)-dependent crosslink traps an active form of domain 3 of a self-splicing group II intron. *RNA* 1(8):828-40.

Podar M, Chu VT, Pyle AM, Perlman PS. 1998. Group II intron splicing in vivo by first-step hydrolysis. *Nature* 391(6670):915-8.

Pyle AM. 2016. Group II Intron Self-Splicing. *Annu Rev Biophys* 45:183-205.

Qin PZ, Pyle AM. 1997. Stopped-flow fluorescence spectroscopy of a group II intron ribozyme reveals that domain 1 is an independent folding unit with a requirement for specific Mg2+ ions in the tertiary structure. *Biochemistry* 36(16):4718-30.

Qu G, Kaushal PS, Wang J, Shigematsu H, Piazza CL, Agrawal RK, Belfort M, Wang HW. 2016. Structure of a group II intron in complex with its reverse transcriptase. *Nat Struct Mol Biol* 23(6):549-57.

Qu G, Piazza CL, Smith D, Belfort M. 2018. Group II intron inhibits conjugative relaxase expression in bacteria by mRNA targeting. *Elife* 7.

Quiroga C, Kronstad L, Ritlop C, Filion A, Cousineau B. 2011. Contribution of base-pairing interactions between group II intron fragments during trans-splicing in vivo. *RNA* 17(12):2212-21.

Ritlop C, Monat C, Cousineau B. 2012. Isolation and characterization of functional tripartite group II introns using a Tn5-based genetic screen. *PLoS One* 7(8):e41589.

Robart AR, Zimmerly S. 2005. Group II intron retroelements: function and diversity. *Cytogenet Genome Res* 110(1-4):589-97.

Robart AR, Seo W, Zimmerly S. 2007. Insertion of group II intron retroelements after intrinsic transcriptional terminators. *Proc Natl Acad Sci U S A* 104(16):6620-5.

Robart AR, Chan RT, Peters JK, Rajashankar KR, Toor N. 2014. Crystal structure of a eukaryotic group II intron lariat. *Nature* 514(7521):193-7.

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol* 13(17):1512-7.

Roitzsch M, Pyle AM. 2009. The linear form of a group II intron catalyzes efficient autocatalytic reverse splicing, establishing a potential for mobility. *RNA* 15(3):473-82.

Ruskin B, Green MR. 1985. An RNA processing activity that debranches RNA lariats. *Science* 229(4709):135-40.

Sagan L. 1967. On the origin of mitosing cells. *J Theor Biol* 14(3):255-74.
Schmelzer C, Schweyen RJ. 1986. Self-splicing of group II introns in vitro: mapping of the branch point and mutational inhibition of lariat formation. *Cell* 46(4):557-65.
Schmidt U, Podar M, Stahl U, Perlman PS. 1996. Mutations of the two-nucleotide bulge of D5 of a group II intron block splicing in vitro and in vivo: phenotypes and suppressor mutations. *RNA* 2(11):1161-72.
Semrad K, Schroeder R. 1998. A ribosomal function is necessary for efficient splicing of the T4 phage thymidylate synthase intron in vivo. *Genes Dev* 12(9):1327-37.
Sharp PA. 1991. "Five easy pieces". *Science* 254(5032):663.
Shearman C, Godon JJ, Gasson M. 1996. Splicing of a group II intron in a functional transfer gene of Lactococcus lactis. *Mol Microbiol* 21(1):45-53.
Shukla GC, Padgett RA. 2002. A catalytically active group II intron domain 5 can function in the U12-dependent spliceosome. *Mol Cell* 9(5):1145-50.
Simon DM, Clarke NA, McNeil BA, Johnson I, Pantuso D, Dai L, Chai D, Zimmerly S. 2008. Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *RNA* 14(9):1704-13.
Simon DM, Zimmerly S. 2008. A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res* 36(22):7219-29.
Simon DM, Kelchner SA, Zimmerly S. 2009. A broadscale phylogenetic analysis of group II intron RNAs and intron-encoded reverse transcriptases. *Mol Biol Evol* 26(12):2795-808.
Singh NN, Lambowitz AM. 2001. Interaction of a group II intron ribonucleoprotein endonuclease with its DNA target site investigated by DNA footprinting and modification interference. *J Mol Biol* 309(2):361-86.
Singh RN, Saldanha RJ, D'Souza LM, Lambowitz AM. 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. *J Mol Biol* 318(2):287-303.
Skelly PJ, Hardy CM, Clark-Walker GD. 1991. A mobile group II intron of a naturally occurring rearranged mitochondrial genome in Kluyveromyces lactis. *Curr Genet* 20(1-2):115-20.
Smathers CM, Robart AR. 2019. The mechanism of splicing as told by group II introns: Ancestors of the spliceosome. *Biochim Biophys Acta Gene Regul Mech* 1862(11-12):194390.
Smillie C, Garcillan-Barcia MP, Francia MV, Rocha EP, de la Cruz F. 2010. Mobility of plasmids. *Microbiol Mol Biol Rev* 74(3):434-52.
Smith D, Zhong J, Matsuura M, Lambowitz AM, Belfort M. 2005. Recruitment of host functions suggests a repair pathway for late steps in group II intron retrohoming. *Genes Dev* 19(20):2477-87.
Staddon JH, Bryan EM, Manias DA, Dunny GM. 2004. Conserved target for group II intron insertion in relaxase genes of conjugative elements of gram-positive bacteria. *J Bacteriol* 186(8):2393-401.
Steitz TA, Steitz JA. 1993. A general two-metal-ion mechanism for catalytic RNA. *Proc Natl Acad Sci U S A* 90(14):6498-502.
Su LJ, Waldsich C, Pyle AM. 2005. An obligate intermediate along the slow folding pathway of a group II intron ribozyme. *Nucleic Acids Res* 33(21):6674-87.
Toor N, Hausner G, Zimmerly S. 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA* 7(8):1142-52.

Toor N, Robart AR, Christianson J, Zimmerly S. 2006. Self-splicing of a group IIC intron: 5' exon recognition and alternative 5' splicing events implicate the stem-loop motif of a transcriptional terminator. *Nucleic Acids Res* 34(22):6461-71.

Toor N, Keating KS, Taylor SD, Pyle AM. 2008. Crystal structure of a self-spliced group II intron. *Science* 320(5872):77-82.

Toro N, Martinez-Abarca F. 2013. Comprehensive phylogenetic analysis of bacterial group II intron-encoded ORFs lacking the DNA endonuclease domain reveals new varieties. *PLoS One* 8(1):e55102.

Touchon M, Rocha EP. 2007. Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol* 24(4):969-81.

Tourasse NJ, Kolsto AB. 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(14):4529-48.

Truong DM, Hewitt FC, Hanson JH, Cui X, Lambowitz AM. 2015. Retrohoming of a Mobile Group II Intron in Human Cells Suggests How Eukaryotes Limit Group II Intron Proliferation. *PLoS Genet* 11(8):e1005422.

Valadkhan S. 2013. The role of snRNAs in spliceosomal catalysis. *Prog Mol Biol Transl Sci* 120:195-228.

Valles Y, Halanych KM, Boore JL. 2008. Group II introns break new boundaries: presence in a bilaterian's genome. *PLoS One* 3(1):e1488.

van der Veen R, Arnberg AC, van der Horst G, Bonen L, Tabak HF, Grivell LA. 1986. Excised group II introns in yeast mitochondria are lariats and can be formed by self-splicing in vitro. *Cell* 44(2):225-34.

van der Veen R, Kwakman JH, Grivell LA. 1987. Mutations at the lariat acceptor site allow self-splicing of a group II intron without lariat formation. *EMBO J* 6(12):3827-31.

Vanacova S, Yan W, Carlton JM, Johnson PJ. 2005. Spliceosomal introns in the deep-branching eukaryote Trichomonas vaginalis. *Proc Natl Acad Sci U S A* 102(12):4430-5.

Vogel J, Borner T. 2002. Lariat formation and a hydrolytic pathway in plant chloroplast group II intron splicing. *EMBO J* 21(14):3794-803.

Wagner A. 2006. Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. *Mol Biol Evol* 23(4):723-33.

Wang Y, Silverman SK. 2006. Experimental tests of two proofreading mechanisms for 5'-splice site selection. *ACS Chem Biol* 1(5):316-24.

Wank H, SanFilippo J, Singh RN, Matsuura M, Lambowitz AM. 1999. A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. *Mol Cell* 4(2):239-50.

Watanabe K, Lambowitz AM. 2004. High-affinity binding site for a group II intron-encoded reverse transcriptase/maturase within a stem-loop structure in the intron RNA. *RNA* 10(9):1433-43.

Watson JD, Crick FH. 1953. The structure of DNA. *Cold Spring Harb Symp Quant Biol* 18:123-31.

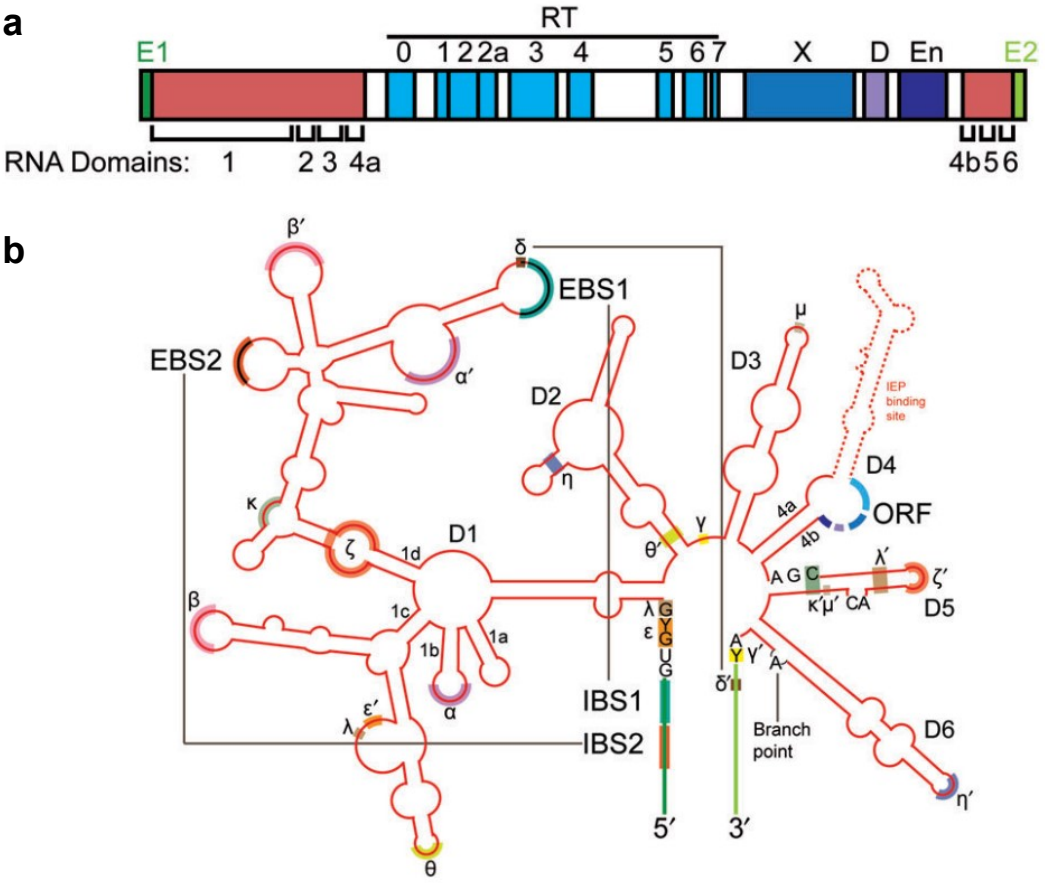
Weiner AM. 1993. mRNA splicing and autocatalytic introns: distant cousins or the products of chemical determinism? *Cell* 72(2):161-4.

Whitaker N, Chen Y, Jakubowski SJ, Sarkar MK, Li F, Christie PJ. 2015. The All-Alpha Domains of Coupling Proteins from the Agrobacterium tumefaciens VirB/VirD4 and Enterococcus

faecalis pCF10-Encoded Type IV Secretion Systems Confer Specificity to Binding of Cognate DNA Substrates. *J Bacteriol* 197(14):2335-49.
Wild MA, Gall JG. 1979. An intervening sequence in the gene coding for 25S ribosomal RNA of Tetrahymena pigmentosa. *Cell* 16(3):565-73.
Will CL, Luhrmann R. 2011. Spliceosome structure and function. *Cold Spring Harb Perspect Biol* 3(7).
Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, Handa S, Khan H, Liu M, Pohlschroder M, Shaw KL et al. . 2018. Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. *Nucleic Acids Res* 46(1):11-24.
Xiong Y, Eickbush TH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* 9(10):3353-62.
Yang J, Zimmerly S, Perlman PS, Lambowitz AM. 1996. Efficient integration of an intron RNA into double-stranded DNA by reverse splicing. *Nature* 381(6580):332-5.
Zhao C, Rajashankar KR, Marcia M, Pyle AM. 2015. Crystal structure of group II intron domain 1 reveals a template for RNA assembly. *Nat Chem Biol* 11(12):967-72.
Zhao C, Pyle AM. 2016. Crystal structures of a group II intron maturase reveal a missing link in spliceosome evolution. *Nat Struct Mol Biol* 23(6):558-65.
Zhao C, Pyle AM. 2017. Structural Insights into the Mechanism of Group II Intron Splicing. *Trends Biochem Sci* 42(6):470-482.
Zhou L, Manias DA, Dunny GM. 2000. Regulation of intron function: efficient splicing in vivo of a bacterial group II intron requires a functional promoter within the intron. *Mol Microbiol* 37(3):639-51.
Zhuang F, Mastroianni M, White TB, Lambowitz AM. 2009. Linear group II intron RNAs can retrohome in eukaryotes and may use nonhomologous end-joining for cDNA ligation. *Proc Natl Acad Sci U S A* 106(43):18189-94.
Zimmerly S, Guo H, Eskes R, Yang J, Perlman PS, Lambowitz AM. 1995a. A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. *Cell* 83(4):529-38.
Zimmerly S, Guo H, Perlman PS, Lambowitz AM. 1995b. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* 82(4):545-54.
Zimmerly S, Hausner G, Wu X. 2001. Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res* 29(5):1238-50.
Zimmerly S, Wu L. 2015. An Unexplored Diversity of Reverse Transcriptases in Bacteria. *Microbiol Spectr* 3(2):MDNA3-0058-2014.

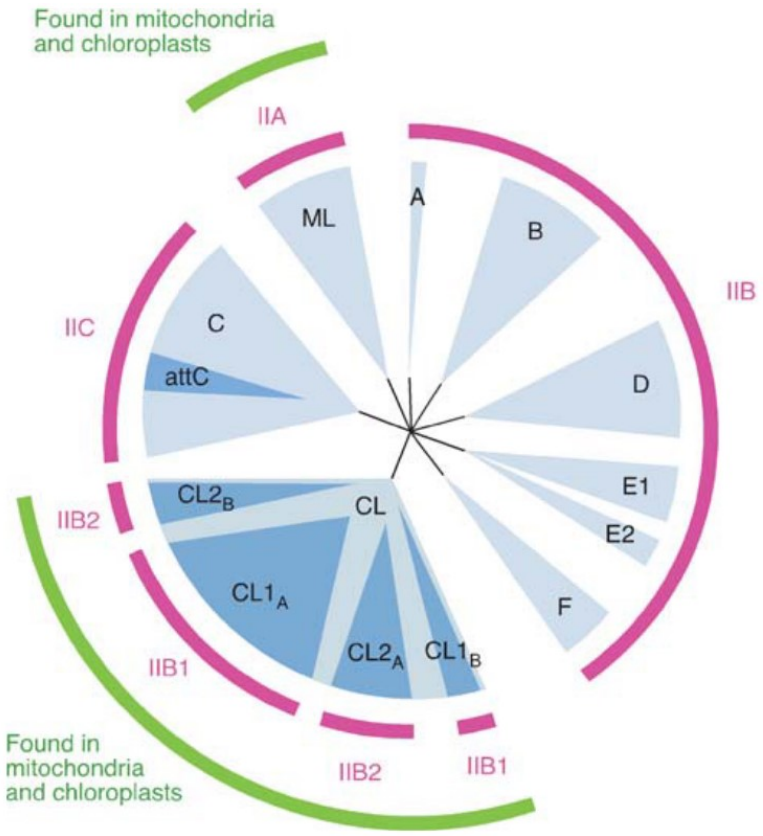
1.9: Figures

Figure 1.1:



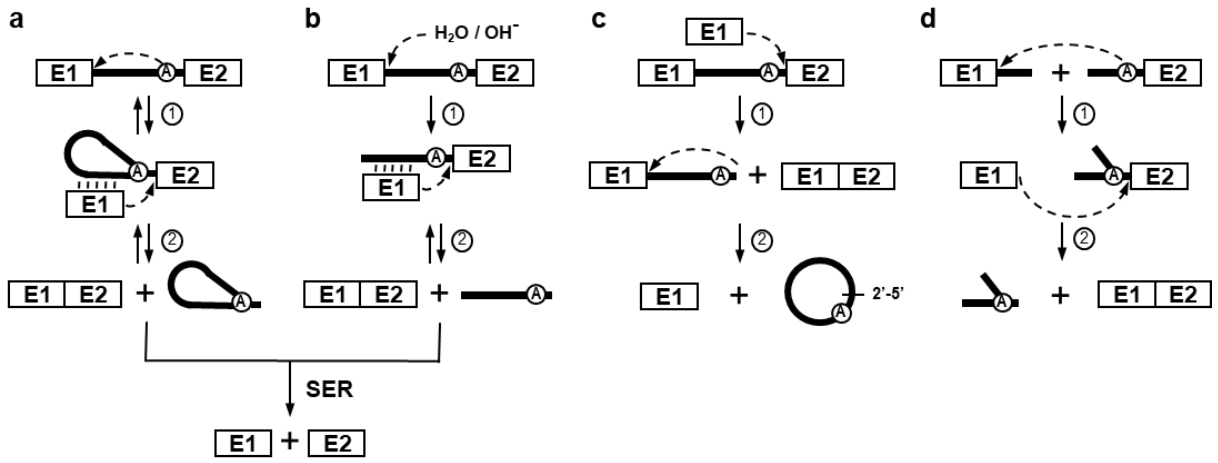
Genomic structure and RNA secondary structure of the Ll.LtrB group II intron (McNeil et al. 2016) (Copyright © 1999-2020 John Wiley & Sons, Inc. All rights reserved). The schematics in both panels correspond to the IIA intron Ll.LtrB of *Lactococcus lactis*. (**a**) Genomic structure. The intron consists of a ribozyme component (red) and protein component (different shades of blue). The RNA component has six structural domains (bracketed below), with domain 4 split into two parts (4a, 4b). The intron-encoded protein consists of RT motifs 0–7 (palm and finger domains of the RT), domain X (thumb domain of the RT and required for maturase activity), a DNA-binding domain D, and an endonuclease domain En, which is lacking from some introns. The intron is nested between two exons, E1 and E2 (green). (**b**) RNA secondary structure. The ribozyme's secondary structure is in red, beginning with the 5′ boundary motif GUGYG and ending with AY. The intron-encoded protein (IEP)'s open reading frame (ORF) is located within the loop of D4 (shades of blue), and the IEP-binding site is indicated by dotted red lines. The 5′ and 3′ exons are in green. Tertiary interactions within the RNA structure are denoted by Greek lettering (e.g., α–α′). For the Ll.LtrB IIA intron, pairings between exons and introns occur through IBS1–EBS1, IBS2–EBS2, and δ–δ′.

Figure 1.2:



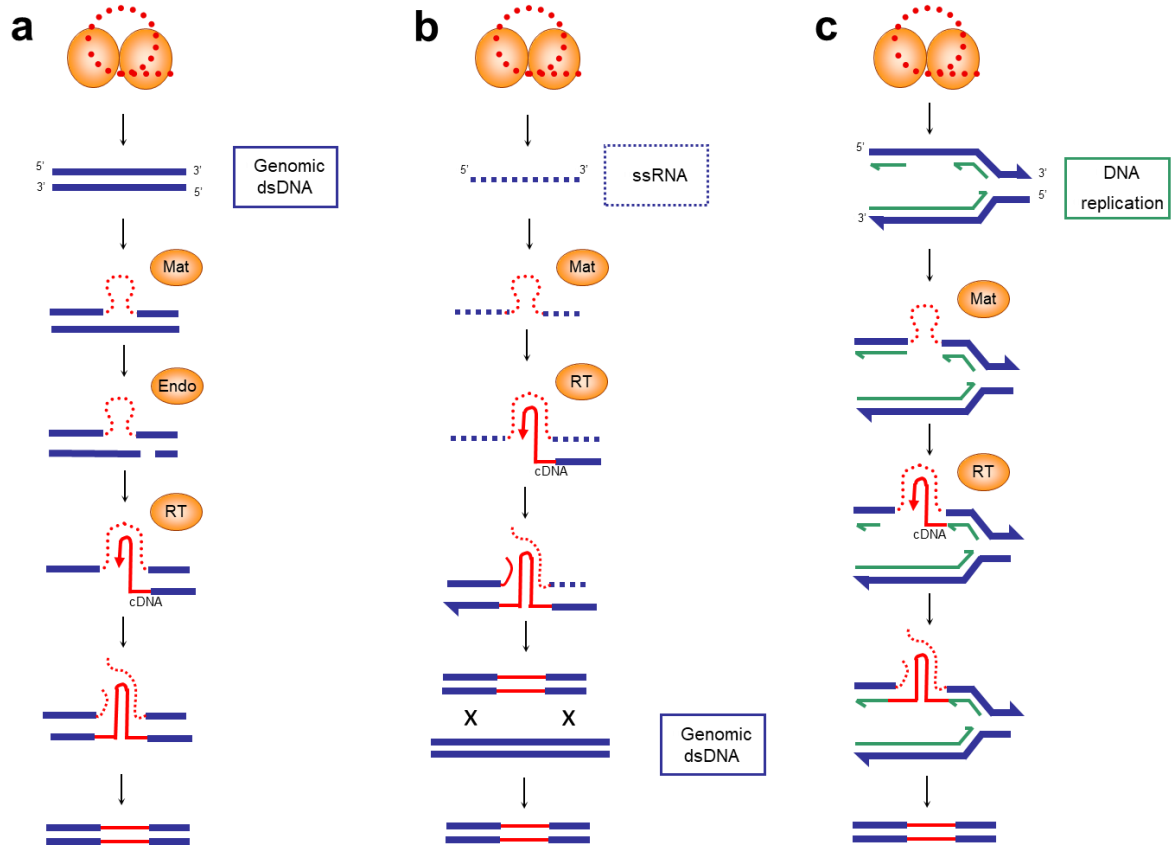
Group II intron lineages (Lambowitz and Zimmerly 2011) (Copyright © 2011 Cold Spring Harbor Laboratory Press; all rights reserved). The major lineages of group II intron IEPs, denoted CL (chloroplast-like), ML (mitochondrial-like), and bacterial classes A-F, are shown as blue sectors. Notable sublineages, including four subdivisions of CL and a subclass of IIC introns that inserts after *attC* sites, are shown as darker blue sectors within the major lineages. RNA structural subgroups that correspond to IEP lineages are shown in magenta. All group II intron lineages and RNA types are found in bacteria. Lineages and RNA types also found in organelles are delineated in green (outer circle). Note that there may be limited exceptions to the overall pattern of coevolution within the CL group, with different sublineages possibly having exchanged IIB RNA structures (Simon et al. 2009).

Figure 1.3:



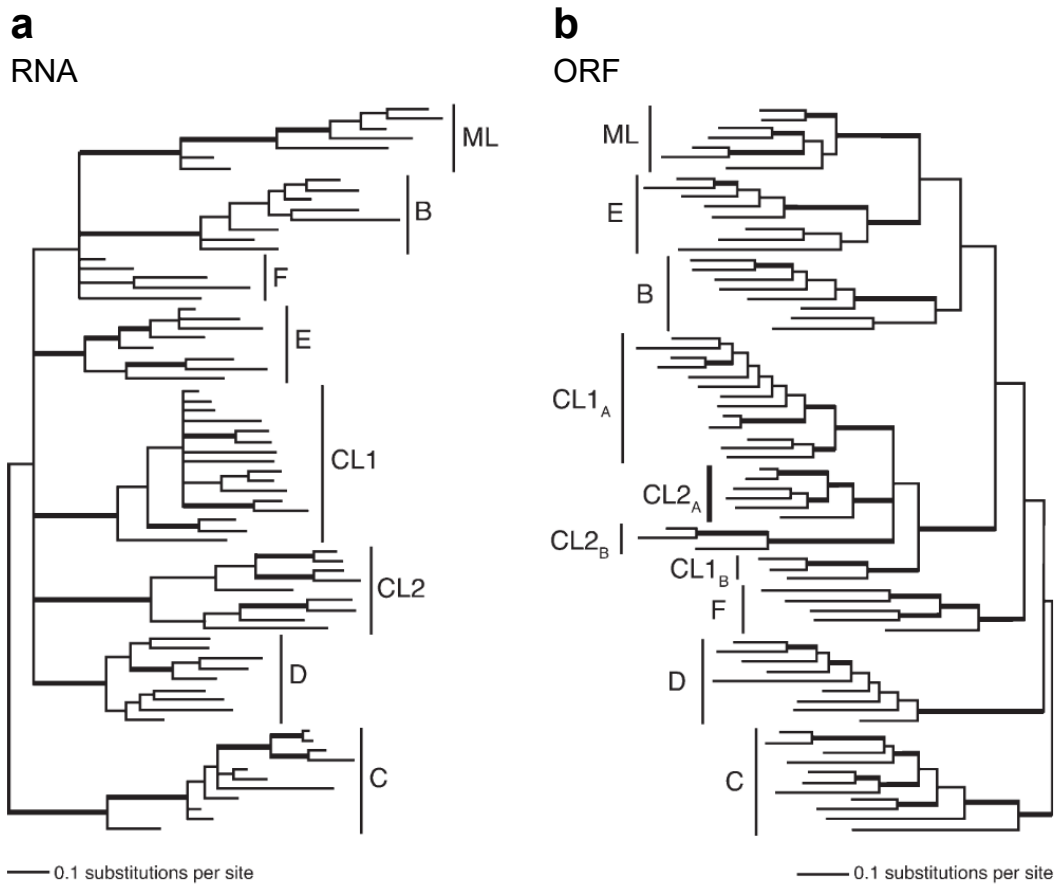
Group II intron splicing pathways. In the group II intron branching pathway (**a**), a bulged adenosine residue called branchpoint uses its 2′ OH to attack the first nucleotide of the intron at the 5′ splice site (step 1), generating a branched lariat molecule. The free E1 remains bound to the group II intron through EBS1/2-IBS1/2 base pairing interactions. The free 3′ OH of E1 next attacks the first nucleotide of E2 at the 3′ splice site (step 2), releasing ligated exons and an intron lariat. In the hydrolysis pathway (**b**), a free water or hydroxyl ion attacks the 5′ splice site instead of the bulged adenosine residue (step 1), releasing E1. The free 3′ OH of E1 next attacks the 3′ splice site (step 2), releasing ligated exons and a linear intron molecule. During circularization (**c**), splicing is initiated by a *trans*-E1 attacking the 3′ splice site (step 1), releasing ligated exons and liberating the intron's 3′ end. The 2′ OH of the intron's last residue next attacks the 5′ splice site (step 2), generating a 2′-5′ head-to-tail intron circle and free E1, which can initiate further instances of circularization. A source of free E1 to initiate circularization may stem from the Spliced Exon Reopening (SER) pathway, where intron lariats and linear introns can hydrolyze ligated exon mRNAs at the ligation point. In group II intron *trans*-splicing (**d**), fragmented intron transcripts (bipartite shown here) assemble using tertiary interactions and use the 2′ OH of the branchpoint adenosine residue to attack the 5′ splice site (step 1), generating a "Y"-shaped intron molecule and releasing E1. The 3′ OH of E1 next attacks the 3′ splice site (step 2), releasing ligated exons and a "Y"-shaped broken lariat.

Figure 1.4:



Mobility pathways of group II introns. In the retrohoming pathway (**a**), a group II intron RNP recognizes an identical or highly homologous homing site within a dsDNA target and initiates complete reverse splicing with the aid of the IEP maturase domain. The endonuclease domain next nicks the bottom strand downstream of the insertion site, providing a primer for target-primed reverse transcription by the RT domain of the IEP, generating a cDNA copy of the intron. Host-encoded RNaseH enzymes next degrade the RNA portion of the intron, and host polymerases and ligases generate an integrated dsDNA copy within the initial dsDNA target site. Group II introns can also mobilize into non-cognate ectopic sites through retrotransposition. Upon recognizing an ectopic homing site within a ssRNA (**b**), the group II intron reverse splices using the IEP maturase domain. Reverse transcription next occurs without the aid of the endonuclease. Host enzymes degrade the initial RNA copy and host polymerases are responsible for second strand synthesis, generating a dsDNA allele of the transcribed gene containing a group II intron. Through homologous recombination, the intron-interrupted allele displaces the intron-free allele, resulting in the genomic insertion of an intron-interrupted gene. Retrotransposition can also occur when the intron recognizes an ectopic mobility site within the ssDNA of a replication fork (**c**). The intron first uses the IEP maturase domain to accurately reverse splice. The RT domain next uses the free 3′ OH of a nascent Okazaki fragment to generate a cDNA copy of the intron, without the help of the endonuclease domain. Host enzymes then degrade the initial RNA copy of the intron and host polymerases and ligases lead to second strand synthesis.

Figure 1.5:



Comparison of intron RNA and ORF phylogenies (Simon et al. 2009) (by permission of Oxford University Press). Consensus trees (50% majority rule) were constructed from two Bayesian runs. Panel (**a**) shows the RNA phylogeny, which is based on 138 nts, and panel (**b**) is the corresponding ORF phylogeny (first and second codon positions; 460 nts). The group IIA introns (ML lineage) including Ll.LtrB from *Lactococcus lactis* show signs of coevolution with the ML lineage ORFs. Group IIC introns, which are much smaller and likely an ancestral form of group II introns, were used as the outgroups. Thick lines indicate nodes with posterior probabilities ≥0.95 and at least one bootstrap value ≥75.

Chapter 2:
Recent horizontal transfer, functional adaptation and dissemination of a bacterial group II intron

2.1: Preface

As mentioned in Chapter 1, group II introns have had an enormous impact on the evolution of eukaryotes (Lambowitz and Belfort 2015). However, their own evolutionary patterns in bacteria are still poorly understood, largely due to the high frequency of horizontal transfer events and the limited conservation of their primary sequence (Simon et al. 2009). Because of these limiting factors, the study of group II intron evolution alongside bacterial hosts has relied heavily on natural cases of recent group II intron dispersal (Dai and Zimmerly 2002a; Fernandez-Lopez et al. 2005; Tourasse and Kolsto 2008). Yet even in these more focused studies, the directionality of lateral transfer and the selective forces shaping the mutations that arise among fixed group II intron variants was never determined.

To address how group II introns evolve in bacteria, we compared the model group II intron Ll.LtrB from *Lactococcus lactis* to a group II intron newly characterized by our lab: Ef.PcfG, from *Enterococcus faecalis*. Since these group II introns are nearly identical (99.7%) yet are present in different bacterial species, we hypothesized that they represented a recent inter-species horizontal transfer event. Our results examined the effects of the 8 point mutations between both introns on their splicing and mobility efficiencies. We obtained experimental evidence that supports the retroelement-like behaviour of group II introns in bacteria, yielding insight on the selective forces that shape their evolution. We furthermore used these findings to propose a directionality for the natural horizontal transfer event that took place between *L. lactis* and *E. faecalis*.

This chapter was adapted from the following manuscript: “Recent horizontal transfer, functional adaptation and dissemination of a bacterial group II intron”. **Félix LaRoche-Johnston**, Caroline Monat and Benoit Cousineau. *BMC Evolutionary Biology*, 2016;16(1):223.

LaRoche-Johnston *et al. BMC Evolutionary Biology* (2016) 16:223
DOI 10.1186/s12862-016-0789-7

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Recent horizontal transfer, functional adaptation and dissemination of a bacterial group II intron

Félix LaRoche-Johnston, Caroline Monat and Benoit Cousineau*

2.2: Summary

Group II introns are catalytically active RNA and mobile retroelements present in certain eukaryotic organelles, bacteria and archaea. These ribozymes self-splice from the pre-mRNA of interrupted genes and reinsert within target DNA sequences by retrohoming and retrotransposition. Evolutionary hypotheses place these retromobile elements at the origin of over half the human genome. Nevertheless, the evolution and dissemination of group II introns was found to be quite difficult to infer.

We characterized the functional and evolutionary relationship between the model group II intron from *Lactococcus lactis*, Ll.LtrB, and Ef.PcfG, a newly discovered intron from a clinical strain of *Enterococcus faecalis*. Ef.PcfG was found to be homologous to Ll.LtrB and to splice and mobilize in its native environment as well as in *L. lactis*. Interestingly, Ef.PcfG was shown to splice at the same level as Ll.LtrB but to be significantly less efficient to invade the Ll.LtrB recognition site. We also demonstrated that specific point mutations between the IEPs of both introns correspond to functional adaptations which developed in *L. lactis* as a response to selective pressure on mobility efficiency independently of splicing. The sequence of all the homologous full-length variants of Ll.LtrB were compared and shown to share a conserved pattern of mutation acquisition.

This work shows that Ll.LtrB and Ef.PcfG are homologous and have a common origin resulting from a recent lateral transfer event followed by further adaptation to the new target site and/or host environment. We hypothesize that Ef.PcfG is the ancestor of Ll.LtrB and was initially acquired by *L. lactis*, most probably by conjugation, via a single event of horizontal transfer. Strong selective pressure on homing site invasion efficiency then led to the emergence of beneficial point mutations in the IEP, enabling the successful establishment and survival of the group II intron in its novel lactococcal environment. The current colonization state of Ll.LtrB in *L. lactis* was probably later achieved through recurring episodes of conjugation-based horizontal transfer as well as independent intron mobility events. Overall, our data provide the first evidence of functional adaptation of a group II intron upon invading a new host, offering strong experimental support to the theory that bacterial group II introns, in sharp contrast to their organellar counterparts, behave mostly as mobile elements.

2.3: Introduction

Group II introns are phylogenetically widespread mobile retroelements present in bacteria, archaea, plant chloroplasts, and mitochondria of fungi and plants (Lambowitz and Zimmerly 2011). However, they are absent, and most likely functionally excluded (Chalamcharla et al. 2010) from the nuclear genomes of eukaryotes that are instead loaded with evolutionarily related intervening sequences called spliceosomal or nuclear introns (Lambowitz and Belfort 2015).

Active group II intron ribonucleoprotein particles (RNPs) are composed of a large and highly structured RNA core associated with two copies of a multifunctional intron-encoded protein (IEP). Following transcription of the intron-interrupted gene, the IEP specifically binds the intervening sequence within the precursor mRNA transcript, assisting the intron to fold into its catalytically active tridimensional conformation. Self-splicing of the intron concurrently leads to the ligation of its flanking exons and the release of active RNPs. Both components of the intron RNPs intimately cooperate in the recognition and invasion of identical or similar sequences using the retrohoming or retrotransposition pathway, respectively (Cousineau et al. 1998; Cousineau et al. 2000; Ichiyanagi et al. 2002; Toro et al. 2007; Lambowitz and Zimmerly 2011).

The architecture and genomic localization of group II introns are quite different depending on whether the host is bacterial or organellar, suggesting that they do not behave the same way in these distinct cellular environments (Dai and Zimmerly 2002b; Nisa-Martinez et al. 2007; Simon et al. 2009; Zimmerly and Semper 2015). Despite having often lost their IEPs, organellar group II introns are mostly splicing-competent and usually interrupt housekeeping genes. These introns are thus more genomically stable and must splice efficiently to ensure adequate expression of the genes they interrupt. In contrast, bacterial group II introns are primarily truncated, inactivated, associated with other mobile genetic elements and located outside housekeeping genes. Taken together, these features suggest that organellar group II introns act almost solely as splicing ribozymes whereas bacterial group II introns behave mostly as mobile genetic elements, cycling through high rates of gain and loss (Wagner 2006). Over evolutionary timescales, bacterial group II introns are believed to be deleterious to their host cells and to survive the streamlining pressure of purifying selection applied on bacterial genomes through repeated instances of extinction and recolonization, previously characterized as the selection-driven extinction model (Leclercq and Cordaux 2012).

On a broad evolutionary perspective, group II introns are thought to have substantially shaped the origin and evolution of contemporary eukaryotic genomes. They are considered as the progenitors of the telomerase enzyme, the very abundant non-LTR retroelements and spliceosomal introns, and the nuclear intron splicing machinery, the spliceosome. Altogether, these presumed group II introns derivatives correspond to more than half of the human genome (Malik et al. 1999; Curcio and Belfort 2007; Chalamcharla et al. 2010; Lambowitz and Zimmerly 2011; Lambowitz and Belfort 2015).

Despite their interesting history, the evolution and dissemination of group II introns was found to be quite difficult to study for several reasons (Tourasse and Kolsto 2008; Simon et al. 2009; Zimmerly and Semper 2015). Indeed, even though the retroelement ancestor hypothesis proposes a general pattern of coevolution between the intron RNA secondary structures and their related IEPs, several caveats remain which hamper the establishment of conclusive group II intron phylogenies (Toor et al. 2001; Dai and Zimmerly 2002b; Chillon et al. 2011). First, both the RNA and protein components have several variant forms with different potential evolutionary histories (Zimmerly and Semper 2015). Second, the use of amino acid sequences from the IEPs as a

phylogenetic marker is quite limited due to the high level of sequence saturation and the potential bias of the amino acid composition depending on the bacterial host, leading to small signal-to-noise ratios and generating uncertainties about the inner nodes of the phylogenetic trees (Simon et al. 2009; Toro and Martinez-Abarca 2013). Third, even if the size of the RNA component is significant (2-3 kb), it is only conserved at the secondary structure and thus evolves rapidly, leaving very limited primary sequence information as potential phylogenetic signal (Fedorova and Zingler 2007). Finally, as retromobile elements that move between genetic locations and that can also be transferred amongst cells within and across species, direct evolutionary links between group II introns as well as with both the genes they interrupt and their host organisms are difficult to infer (Klein and Dunny 2002). Therefore, definitive conclusions about the evolution and dissemination of group II introns can only be drawn by studying homologous introns that diverged relatively recently (Zimmerly and Semper 2015).

The presence of multiple classes of introns in a number of given bacterial species suggests that group II intron horizontal transfer is quite common (Leclercq et al. 2011). Accordingly, the majority of functional bacterial group II introns are found associated with other mobile genetic elements, such as conjugative plasmids, transposons, and IS elements (Klein and Dunny 2002; Zimmerly and Semper 2015). However, only a limited number of natural horizontal transfer events have been conclusively demonstrated, and in every case the precise origin of the transferred intron was impossible to infer (Dai and Zimmerly 2002a; Fernandez-Lopez et al. 2005; Tourasse and Kolsto 2008; Leclercq et al. 2011; Zimmerly and Semper 2015). Previous studies have shown that Ll.LtrB, the model group II intron from the gram-positive bacterium *Lactococcus lactis*, is able to invade conserved loci within orthologous genes in transconjugant bacterial strains by both retrohoming and retrotransposition, following the intra- (*L. lactis* to *L. lactis*) and inter-species (*L. lactis* to *Enterococcus faecalis*) transfer of its host conjugative elements (Belhocine et al. 2004; Staddon et al. 2004; Belhocine et al. 2005; Belhocine et al. 2007b).

Here we describe the functional and evolutionary relationship between Ef.PcfG, a newly discovered group II intron from a clinical strain of *E. faecalis* (SF24397), and the model group II intron from *Lactococcus lactis*, Ll.LtrB. Overall, our data support the hypothesis that Ef.PcfG is ancestral to Ll.LtrB and was acquired by *L. lactis*, most likely by conjugation, through a single horizontal transfer event. Repeated instances of conjugation-based horizontal transfer and

independent intron mobility events led to the current colonization status of *L. lactis*. We also show for the first time the functional adaptation of a group II intron following its acquisition by horizontal transfer, providing strong experimental support to the theory that group II introns behave mostly as mobile elements in bacterial cells.

2.4: Results

2.4.1: A clinical isolate of *Enterococcus faecalis* contains a functional group II intron closely related to Ll.LtrB from *Lactococcus lactis*

We identified by sequence comparison a novel group II intron in the SF24397 strain of *E. faecalis* isolated from the urine sample of a patient in Michigan, USA (McBride et al. 2007). The intron was found on a large contig within a region of high sequence similarity to the *E. faecalis* pTEF2 conjugative plasmid (Paulsen et al. 2003). It interrupts a relaxase gene, *pcfG*, at the exact same conserved position Ll.LtrB interrupts the *ltrB* relaxase gene in *L. lactis* (Pansegrau et al. 1994; Staddon et al. 2004). Because of its origin (*E. faecalis*) and genetic location (*pcfG*) this novel group II intron and its intron-encoded protein were named Ef.PcfG and IepG respectively.

Ef.PcfG is almost identical to Ll.LtrB (99.7%), the model group II intron from the *L. lactis* conjugative plasmid, pRS01 (Mills et al. 1996), exhibiting only eight point mutations out of a total of 2492 nts (Fig. 2.1). The majority of the mutations (7/8) are located in domain IV, within the IEP (Mut #2-Mut #8), while one mutation (Mut #1) is located within the ribozyme portion of the intron in a bulged region of domain III. Five of the mutations in domain IV are missense mutations leading to amino acid changes in either the reverse transcriptase (IEP-RT) or the DNA binding (IEP-DB) domain of the IEP (Fig. 2.1A).

Plasmid isolation from *E. faecalis* SF24397 revealed the presence of two resident plasmids: a large pTEF2-like plasmid harboring the interrupted *pcfG* relaxase gene and the pEF1071 conjugative plasmid (9328 bp), containing an uninterrupted relaxase gene called *mobA* (Balla and Dicks 2005). To study Ef.PcfG in a more convenient genetic environment we generated the SF24397ΔpEF1071 strain by curing the pEF1071 plasmid with novobiocin (Ruiz-Barba et al. 1991).

To determine whether Ef.PcfG is functional in its native environment, we first assessed its ability to splice *in vivo* by looking for both released intron and ligated exons by RT-PCR on *E. faecalis* (SF24397ΔpEF1071) total RNA extracts (Fig. 2.2A) (Monat et al. 2015; Monat and Cousineau 2016). Sequencing of the major amplicons for both ligated exons and released intron showed that they correspond respectively to accurately joined *pcfG* exons (1587 bp) (Fig. 2.2B) and released intron lariats (287 bp) (Fig. 2.2C). Next, we wanted to examine if Ef.PcfG is mobile in *E. faecalis* (SF24397). We took advantage of the endogenous pEF1071 plasmid, which harbors an uninterrupted *mobA* relaxase gene of the same family as *ltrB* and *pcfG* (Balla et al. 2000; Balla and Dicks 2005). Even though the potential Ef.PcfG recognition site within *mobA* is not identical to its original recognition site in *pcfG* (Fig. 2.1B), we were able to detect Ef.PcfG mobility products within *mobA* by PCR (Fig. 2.2D). Using two primer pairs, where one primer is specific for Ef.PcfG and the other specific for a plasmid sequence outside *mobA*, we amplified both the 5′ (E1-Ef.PcfG) (626 bp) and 3′ (Ef.PcfG-E2) (1021 bp) junctions of the intron mobility products (Fig. 2.2D). The sequence of both amplicons confirmed the precise insertion of Ef.PcfG within *mobA* at the expected position (Fig 1B, arrowhead) (Staddon et al. 2004).

These results reveal the presence of a novel group II intron, homologous to Ll.LtrB, in a clinical isolate of *E. faecalis* (SF24397). They also demonstrate that the interrupted *pcfG* relaxase gene is expressed in *E. faecalis* and that following accurate splicing from its relaxase transcript, EF.PcfG can invade, at the exact same conserved position, the *mobA* relaxase gene present within the resident conjugative plasmid pEF1071.

2.4.2: Ll.LtrB and Ef.PcfG can recognize and invade each other's homing sites

Several facts support the hypothesis of a recent event of group II intron horizontal transfer between *E. faecalis* and *L. lactis*: i) the high degree of identity between Ll.LtrB and Ef.PcfG (99.7%) (Fig. 2.1A) compared to their full-length (61%) or proximal flanking exons (72%) (Fig. 2.1B); ii) both introns interrupt related relaxase genes at the exact same conserved position; iii) Ef.PcfG and Ll.LtrB are both splicing and mobile in their native environments; iv) Ll.LtrB is contained within different *L. lactis* conjugative elements previously shown to transfer to *E. faecalis* (Belhocine et al. 2007b) while the *E. faecalis* pTEF2 plasmid, harboring Ef.PcfG, is closely related

to the pheromone-sensitive pCF10 plasmid, that was shown to laterally transfer to *L. lactis* at very high efficiencies (Staddon et al. 2006); v) the *pcfG* gene of pCF10 was previously shown to be a functional target for Ll.LtrB in *E. faecalis* following inter-species conjugation experiments (Staddon et al. 2004). It was thus of interest to compare the functional characteristics of Ef.PcfG and Ll.LtrB to elucidate the evolutionary relationship between these two homologous bacterial group II introns.

Using a two-plasmid intron mobility assay (Fig. 2.3A) (Plante and Cousineau 2006), we first measured and contrasted the efficiency of each intron to recognize and invade the recognition or homing site (HS) of both the *ltrB* and *pcfG* genes. The mobility assay consisted of co-transforming *L. lactis* with an intron donor plasmid, containing either the *ltrB*- (pLE-Pnis-*ltrB*E1-Ll.LtrB-*ltrB*E2) or *pcfG*- (pLE-Pnis-*pcfG*E1-Ef.PcfG-*pcfG*E2) interrupted gene downstream of the nisin-inducible promoter (Pnis), and an intron recipient plasmid harboring the HS of either the *ltrB* (pDL-*ltrB*-HS) or *pcfG* (pDL-*pcfG*-HS) gene (Fig. 2.3A) (Plante and Cousineau 2006). Intron mobility efficiency was subsequently calculated by patch hybridization as the proportion of intron recipient plasmids invaded by the intron (Belhocine et al. 2004).

The mobility efficiency of Ef.PcfG was found to be significantly higher at its own HS, *pcfG*-HS (89.0 %) than at the *ltrB*-HS (42.0 %) (Fig. 2.3B, Wild-type flanking exons). On the other hand, Ll.LtrB invaded the *pcfG*-HS (85.5 %) significantly more efficiently than its own recognition site, the *ltrB*-HS (65.0 %). Our data thus show that while the *ltrB*-HS is invaded significantly more efficiently by Ll.LtrB (65.0 % *vs* 42.0 %), both introns are capable to invade significantly more proficiently, and at similar levels, the *E. faecalis pcfG*-HS (65.0% *vs* 85.5 % and 42.0% *vs* 89.0 %).

Taken together, these results suggest that Ef.PcfG is ancestral to Ll.LtrB, and that following the invasion of the *ltrB* gene, Ef.PcfG adapted and evolved to mobilize significantly more efficiently to its new sequence environment, the *ltrB*-HS, without affecting its proficiency to invade its original recognition site, the *pcfG*-HS.

2.4.3: The significant difference in mobility efficiency at the *ltrB*-HS between Ll.LtrB and Ef.PcfG is due to sequence variations within the introns

Having uncovered a significant difference in the capacity of both introns to invade the *ltrB*-HS, we sought to examine the cause of this difference. Sequence variations exist between Ef.PcfG and Ll.LtrB (Fig. 2.1A) and also among their flanking exons (Fig. 2.1B), both of which may potentially affect intron mobility efficiency (Fedorova and Zingler 2007). We thus exchanged the introns between the interrupted *ltrB* and *pcfG* genes creating two new intron donor plasmids harboring Ll.LtrB flanked by the *pcfG* exons (pLE-Pnis-*pcfG*E1-Ll.LtrB-*pcfG*E2) and Ef.PcfG flanked by the *ltrB* exons (pLE-Pnis-*ltrB*E1-Ef.PcfG-*ltrB*E2) (Fig. 2.3B, Exchanged flanking exons).

Using our two-plasmid intron mobility assay (Fig. 2.3A) we found that the Ll.LtrB intron, despite being flanked by the *pcfG* exons, mobilized to its own HS with similar efficiency (70.5 % *vs* 65.0 %) (Fig. 2.3B). Likewise, the Ef.PcfG intron, interrupting the *ltrB* gene, mobilized to the *ltrB*-HS with comparable efficiency as wild-type Ef.PcfG (44.5 % *vs* 42.0 %). The same trend was observed for the mobility efficiency of both introns to the *pcfG*-HS.

Overall, mobility efficiencies of both introns are not significantly altered regardless of the nature of their flanking exons. Our data thus demonstrate that the difference in mobility efficiency to the *ltrB*-HS between Ll.LtrB and Ef.PcfG is not due to sequence variations amongst the flanking exons, but most likely due to the eight point mutations between the introns.

2.4.4: The variation in mobility efficiency between Ef.PcfG and Ll.LtrB at the *ltrB*-HS is not due to changes in splicing efficiency

Two main factors can influence group II intron mobility efficiency: splicing or RNP release and homing site invasion. Following the observation that the difference in mobility efficiency between Ll.LtrB and Ef.PcfG to the *ltrB*-HS is most likely due to sequence variations within the introns, we wanted to first study if these point mutations affect splicing efficiency.

To evaluate the splicing efficiency of our various intron constructs, we performed poisoned primer extension assays (Fig. 2.4A) (Plante and Cousineau 2006; Monat and Cousineau 2016). This assay compares the ratio of ligated exons to precursor mRNA from total RNA extracts. Since the sequence of the two RNAs are different after the exon 2 junction, the first G residue encountered is at a different distance from the primer, generating differently sized bands for the precursor and the ligated exons (Fig. 2.4A). Our data show that the splicing efficiency of Ll.LtrB and Ef.PcfG are almost identical, varying from 32-35%, regardless of whether they are flanked by their cognate exons or not (Fig. 2.4B).

These results demonstrate that variations in splicing efficiency between Ll.LtrB and Ef.PcfG cannot account for the significant increase in mobility efficiency observed for Ll.LtrB at the *ltrB*-HS and rather suggest that Ll.LtrB is more proficient than Ef.PcfG during the invasion of the *ltrB*-HS.

2.4.5: Some of the point mutations within Ef.PcfG increase its mobility efficiency to the *ltrB*-HS

Having demonstrated that the significant difference in mobility efficiency at the *ltrB*-HS between Ef.PcfG and Ll.LtrB arises from sequence variations within the introns and that these eight mutations (Fig. 2.1A, Mut #1 to Mut #8) do not affect splicing efficiency, we studied the individual effect of these mutations on the mobility efficiency of Ef.PcfG. Eight intron donor plasmids were engineered by site-directed mutagenesis (pLE-Pnis-*pcfG*E1-Ef.PcfG-Mut #1-*pcfG*E2 to pLE-Pnis-*pcfG*E1-Ef.PcfG-Mut #8-*pcfG*E2) and co-transformed independently with pDL-*ltrB*-HS or pDL-*pcfG*-HS in *L. lactis*.

The mobility efficiency of these eight Ef.PcfG mutants was assessed using our two-plasmid mobility assay (Fig. 2.3A). Four mutations in domain IV (Mut #2, #5, #6, #8), within the IepG coding region, increased the mobility efficiency of Ef.PcfG to the *ltrB*-HS (Fig. 2.5, black bars), two of them significantly (Mut #2, #6). In contrast, three mutations did not significantly affect the mobility efficiency of Ef.PcfG to *ltrB*-HS (Mut #1, #3, #4). Mutation #1 is located in a bulged region of domain III not disrupting the predicted secondary structure or any of the previously identified long-range tertiary interactions (Dai et al. 2008). Mutations #3 and #4, in domain IV,

are silent and thus do not change the amino acid sequence of the IepG protein. Finally, mutation #7 lead to a significant decrease of the Ef.PcfG mobility efficiency to *ltrB*-HS. As expected, none of the eight point mutations significantly affected the mobility efficiency of Ef.PcfG to its own recognition site (pDL-*pcfG*-HS) (Fig. 2.5, open bars).

Taken together, these results demonstrate that half of the point mutations between Ef.PcfG and Ll.LtrB improved the mobility efficiency of Ef.PcfG to the *ltrB*-HS, two of them significantly. This supports the hypothesis that Ef.PcfG invaded the *ltrB*-HS and then functionally adapted following evolutionary pressure on the invasion efficiency of its new flanking exons.

2.4.6: Full-length variants of Ll.LtrB in *L. lactis* share a conserved pattern of mutation acquisition

Having demonstrated that some point mutations within Ef.PcfG significantly improve its ability to invade the *ltrB*-HS, we analysed the distribution of the eight point mutations throughout all the homologous full-length group II introns present in *L. lactis* (Fig. 2.6). A total of 24 homologous full-length introns were thus identified, aligned and grouped together in clades based on nucleotide divergence from Ef.PcfG (Fig. 2.6). An additional group II intron, almost identical to Ef.PcfG (1 point mutation) was also identified in *E. faecalis* (533_EFLS). The presence of an additional nucleotide difference between this Ef.PcfG variant and Ll.LtrB (9 point mutations) suggests that the Ef.PcfG variant from SF24397 better represents the ancestral state of the intron that potentially colonized *L. lactis*.

All of the Ll.LtrB variants contain between 5 and 10 point mutations when compared to Ef.PcfG. The eight point mutations between Ef.PcfG and the Ll.LtrB model intron associated with the *L. lactis* pRS01 conjugative plasmid are found to progress throughout the dendrogram with all *L. lactis* introns containing mutations #2, #4, and #6 (Fig. 2.6). Although the distribution of the eight point mutations forms clear clades, introns throughout the dendrogram were found both in *L. lactis* subsp. *lactis* (Fig. 2.6, underlined) and *L. lactis* subsp. *cremoris* (Fig. 2.6, not underlined). These introns were also found interrupting different conserved functional motifs within relaxase genes, notably HLHN-H (Fig. 2.6, thin branches) and HIHN-H (Fig. 2.6, thick branches), with one instance of retrotransposition into an ectopic chromosomal site (SK11).

Overall, these results support the hypothesis that a single horizontal transfer event led to the introduction of an ancestral Ef.PcfG into *L. lactis*, whereby selective pressure caused the intron to adapt and evolve to its new host and/or sequence environment. This newly acquired intron then subsequently disseminated to other strains of *L. lactis*, most likely by conjugation, leading to the intra-species dispersal of group II introns in a new bacterial host.

2.5: Discussion

In this study, we initially identified a functional group II intron, interrupting the *pcfG* relaxase gene of a clinical isolate of *E. faecalis*, which we named Ef.PcfG. This intron is homologous and almost identical to Ll.LtrB, the model group II intron from *L. lactis*, and was shown to splice and mobilize in its native host environment as well as in *L. lactis*. Interestingly, Ef.PcfG was found to splice at the same level as Ll.LtrB in *L. lactis* but to be significantly less efficient to invade the *ltrB*-HS. In contrast, both introns recognize the *pcfG*-HS significantly more proficiently, and at the same level. Finally, we identified, compiled and analyzed the homologous full-length variants of Ll.LtrB present in *L. lactis* and characterized the functional and evolutionary relationship between Ef.PcfG and Ll.LtrB from pRS01.

Overall, our data can be interpreted in different ways regarding both the direction and the order of the lateral transfer of the intron. The intron could have been transferred from *L. lactis* to *E. faecalis*, from *E. faecalis* to *L. lactis* or simply originated from a third partner. The Ll.LtrB variants present in *L. lactis* could also be the result of independent and recent horizontal transfer events rather than a single episode of horizontal transfer. Nevertheless, our results show that Ll.LtrB and Ef.PcfG are homologous and have a common origin resulting from a recent lateral transfer event followed by further adaptation to the new target site and/or host environment.

Even though we cannot rule out the alternative scenarios described above, our favored hypothesis supported by our data is that an ancestral Ef.PcfG colonized *L. lactis* from *E. faecalis*. Ef.PcfG would have most probably been transferred by conjugation relatively recently and initially invaded an orthologous relaxase gene of an *L. lactis* conjugative element. Following this single instance of horizontal transfer, the newly acquired intron would have adapted to its novel host and/or sequence environment in response to selective pressure specifically on target site invasion

independently of splicing. Being associated with a conjugative element, the intron was further disseminated by conjugation between *L. lactis* strains and subspecies, invading the same highly conserved sequence motif present in two different catalytic histidine motifs of relaxase genes associated with various *L. lactis* conjugative elements.

Previous studies on the distribution of group II introns within bacterial populations revealed the presence of homologous group II introns in different bacterial strains and species (Dai and Zimmerly 2002a; Fernandez-Lopez et al. 2005; Tourasse and Kolsto 2008; Leclercq et al. 2011). In these studies, however, the direction of the horizontal transfer, believed to be responsible for the dispersal of these introns, was impossible to infer. The observed genetic differences between these homologous retroelements thus provided no insights either into the nature of potential selective pressures affecting group II introns upon their colonization of a new sequence and/or host environment, or into how they would adapt to these hypothetical selective pressures. In contrast, our experimental data show that Ll.LtrB is significantly more efficient than Ef.PcfG to invade the *ltrB*-HS. This finding supports our hypothesis that the intron was horizontally transferred from *E. faecalis* to *L. lactis* and, combined with the fact that this difference in mobility efficiency is splicing-independent, indicates that the newly acquired intron adapted to its new host and/or sequence environment following selective pressure specifically applied on target site invasion. The adaptation responsible for the increased efficiency of Ll.LtrB to invade the *ltrB*-HS was most likely recent enough that it did not affect its vestigial capacity to invade its ancestral recognition site, the *pcfG*-HS, at a higher efficiency. The analysis of the eight point mutations between Ef.PcfG and Ll.LtrB from pRS01 further supports our hypothesis. Mutations #2, #5, #6, and #8 all increase the Ef.PcfG mobility efficiency towards the *ltrB*-HS, mutations #2 and #6 significantly, while none of these individual mutations affect the mobility efficiency towards the *pcfG*-HS. The location of these amino acid substitutions in the RT (Mut #2, #5, #6) and DNA binding (Mut #8) domains of IepG correlate with a splicing-independent increase in intron mobility.

Analysis of the homologous full-length Ll.LtrB variants in *L. lactis* revealed a pattern of mutation acquisition which, although not extensive enough to construct a conclusive phylogeny, enabled their grouping into clades based on the distribution of the eight point mutations between Ef.PcfG and Ll.LtrB from pRS01. Three mutations were found to be common to all of the analysed

lactococcal introns (Mut #2, #4 and #6). Although it remains possible that these ubiquitous mutations only arose in the *E. faecalis* copy of Ef.PcfG after the horizontal transfer event occurred, two of them (Mut #2 and #6) lead to significant increases in mobility efficiency of Ef.PcfG towards the *ltrB*-HS. The appearance of these mutations could thus represent an adaptive response to a selective pressure on the mobility efficiency of the intron upon entering a new host and/or sequence environment. The third mutation common to all introns in *L. lactis* is a silent mutation (Mut #4), which accordingly showed no significant difference in the mobility efficiency of Ef.PcfG to either the *ltrB*-HS or *pcfG*-HS. It is thus impossible to determine whether this mutation was acquired independently in the *E. faecalis* copy of Ef.PcfG after the horizontal transfer event occurred, or whether it arose in the ancestor of all the Ll.LtrB variants. In either case, the ubiquitous presence of both beneficial mutations within all *L. lactis* introns supports the scenario of a single horizontal transfer event from *E. faecalis* to *L. lactis*, followed by subsequent dissemination and accumulation of additionnal independent mutations. This explanation offers a more parsimonious sequence of events than the alternative view involving multiple independent lateral transfer events and the independent acquisition of identical beneficial mutations by numerous introns.

Group II introns can theoretically be horizontally transferred by either invading a new site in a novel host or by the transfer of a previously interrupted gene. Nevertheless, we have found extensive evidence for independent intron mobility within *L. lactis* rather supporting the initial introduction of Ef.PcfG into the *ltrB*-HS by invasion of that new site rather than by acquisition of the interrupted gene. First, the stark contrast between the homology of Ef.PcfG and Ll.LtrB (99.7%) compared with *pcfG* and *ltrB* (61%) is a good indication that the initial introduction of Ef.PcfG into the *ltrB*-HS was by site-specific invasion. Second, aside the single exception of a retrotransposition event into a chromosomal gene coding for a cell surface protein (strain SK11), the homologous full-length introns were found to interrupt two different catalytic histidine motifs of relaxase genes: HLHN-H and HIHN-H (Ilyina and Koonin 1992). Overall these data suggest that Ll.LtrB was acquired and disseminated through a series of horizontal transfer and independent mobility events.

The pTEF2-like element which harbors Ef.PcfG in SF24397 greatly resembles the pCF10 pheromone-sensitive plasmid, which has been shown to conjugate efficiently to *L. lactis* (Hirt et al. 2005; Staddon et al. 2006). This suggest that the initial introduction of Ef.PcfG into *L. lactis*

was the result of a conjugative transfer. This is consistent with previous studies proposing that conjugation may play an important role for the spread of bacterial group II introns (Belhocine et al. 2004; Belhocine et al. 2005; Belhocine et al. 2007b; Nisa-Martinez et al. 2007). Although pCF10 is not stably maintained in *L. lactis*, the transient introduction of a conjugative plasmid was shown to be sufficient for group II intron expression and the invasion of new loci in recipient cells (Belhocine et al. 2004). On the other hand, lactococcal conjugative plasmids can also transfer to *E. faecalis*, allowing for a potential alternative scenario where an *L. lactis* conjugative plasmid could have been initially introduced into *E. faecalis*, invaded by Ef.PcfG, and then transferred back into *L. lactis* (Chen et al. 2007). Our findings that various intermediates in the acquisition of the eight point mutations are found within various conjugative elements in either *L. lactis* subsp. *lactis* or *L. lactis* subsp. *cremoris*, which have no tendency to cluster in the dendrogram, suggests that upon arrival in *L. lactis*, additional conjugation-based horizontal transfer events lead to further dissemination of the intron.

Bacterial group II introns have been previously characterized as behaving more like retroelements than splicing-only introns (Dai and Zimmerly 2002b). The distribution of group II introns has largely supported this theory by finding markers underlining their behavior as transposable elements in constant movement, such as identical group II introns in different bacterial strains and species, unoccupied HSs in intron-containing bacteria, and numerous intron fragments alongside full-length copies (Dai and Zimmerly 2002a; Fernandez-Lopez et al. 2005; Tourasse and Kolsto 2008; Leclercq et al. 2011). Group II introns have previously been shown to recognize new HSs more efficiently by modifying the exon binding sites 1 and 2 regions of their RNA component (Mohr et al. 2010). However, the mutations increasing the mobility efficiency of Ef.PcfG towards the *ltrB*-HS are located in IepG showing that group II introns are also able to adapt to new sequence and/or host environments by mutating their IEP. Our data thus demonstrate for the first time the functional adaptation of a group II intron following its acquisition by horizontal transfer, providing strong experimental support to the theory that group II introns behave mostly as mobile elements in bacteria.

2.6: Conclusions

This work shows that Ll.LtrB and Ef.PcfG are homologous and have a common origin resulting from a recent lateral transfer event followed by further adaptation to the new target site and/or host environment. We hypothesize that Ef.PcfG is the ancestor of Ll.LtrB and was initially acquired by *L. lactis*, most probably by conjugation, via a unique event of horizontal transfer. Strong selective pressure on homing site invasion efficiency then led to the emergence of beneficial point mutations in the IEP, enabling the successful establishment and survival of the group II intron in its novel lactococcal environment. The current colonization state of Ll.LtrB in *L. lactis* was probably later achieved through recurring episodes of conjugation-based horizontal transfer as well as independent intron mobility events. Overall, our data provide the first evidence of functional adaptation of a group II intron upon invading a new host, offering strong experimental support to the theory that bacterial group II introns, in sharp contrast to their organellar counterparts, behave mostly as mobile elements.

2.7: Experimental procedures

2.7.1: Bacterial strains and plasmids

Lactococcus lactis strain NZ9800Δ*ltrB* (Tet^R) (Ichiyanagi et al. 2002) was grown in M17 media supplemented with 0.5% glucose (GM17) at 30°C without shaking. The *Escherichia coli* strain DH10β was grown in LB at 37°C with shaking. The two strains of *Enterococcus faecalis*, SF24397 (Erm^R/Gen^R) (McBride et al. 2007) and SF24397ΔpEF1071 (Erm^R) (this study), were grown in BHI at 37°C without shaking. To generate the SF24397ΔpEF1071 strain, the resident plasmid pEF1071 was cured from SF24397 by treatment with Novobiocin (15 µg/µl) (Ruiz-Barba et al. 1991). Antibiotics were used at the following concentrations: chloramphenicol (Cam^R), 10 µg/ml; spectinomycin (Spc^R), 300 µg/ml; erythromycin (Erm^R), 300 µg/ml.

Plasmids and primers used in this study are listed in Tables S2.1 and S2.2, respectively. pLE-Pnis-*ltrB*E1-Ll.LtrB-*ltrB*E2 and pLE-Pnis-*pcfG*E1-Ef.PcfG-*pcfG*E2 were constructed by first cloning (BamHI) the nisin-inducible promoter (Pnis) within the shuttle plasmid pLE1 (Cam^R)

(Kuipers et al. 1993; Mills et al. 1997). The *ltrB* and *pcfG* genes interrupted by their respective introns, Ll.LtrB and Ef.PcfG, were then cloned (NotI) downstream of Pnis. The pLE-Pnis-*pcfG*E1-Ll.LtrB-*pcfG*E2 and pLE-Pnis-*ltrB*E1-Ef.PcfG-*ltrB*E2 were generated by swapping a restriction fragment (BsrGI/BsiWI) that contains the eight point mutations between both plasmids. The pLE-Pnis-*pcfG*E1-Ef.PcfG-*pcfG*E2-Mut #1-Mut #8 plasmids were generated independently by site-directed mutagenesis (New England Biolabs® Q5® Site-Directed-Mutagenesis Kit) (primers in Table S2.2). The intron recipient plasmid pDL-*ltrB*-HS contains a 271 bp fragment (HindIII) of the *ltrB* relaxase gene, encompassing the native Ll.LtrB homing site, inserted within the pDL278 plasmid (SmaI) (Mills et al. 1997). Similarly, the pDL-*pcfG*-HS plasmid harbors a 602 bp PCR amplicon (AclI) of the *pcfG* relaxase gene (primers in Table S2.2), cloned into pDL278 (SmaI).

2.7.3: Two-plasmid intron mobility and patch hybridization assays

To assess Ef.PcfG and Ll.LtrB mobility efficiency to both the *ltrB*-HS and *pcfG*-HS, NZ9800Δ*ltrB* cells containing an intron donor and an intron recipient plasmid were induced for intron expression with nisin as previously described (Plante and Cousineau 2006). Plasmid mixes (donor, recipient, mobility product) were extracted and electroporated into *E. coli* strain DH10β, which were then plated on LB/Spc plates to select for pDL-based plasmids (recipient plasmids and mobility products). The percentage of mobility efficiency was then obtained by patching 100 isolated colonies onto a new LB/Spc plate. Patches were lifted on a Hybond-N nylon membrane (Amersham™) and screened with a P^{32}-labelled intron specific probe (Table S2.2) to reveal intron mobility events. Mobility efficiency was then calculated as a percentage of positive hybridization events out of 100 colonies (Plante and Cousineau 2006). Statistical significance was calculated using an unpaired Student's T-test, with $p<0.05$.

2.7.4: RNA extraction, RT-PCR, PCR and poisoned primer extension

Total RNA was isolated from SF24397ΔpEF1071 and NZ9800Δ*ltrB* harboring various plasmid constructs as previously described (Belhocine et al. 2007a). RT-PCR reactions (Belhocine et al. 2007a) and poisoned primer extensions (Plante and Cousineau 2006) were performed on total

RNA preparations of SF24397ΔpEF1071 and NZ9800Δ*ltrB* harboring various intron constructs, respectively (primers in Table S2.2). PCR amplifications of the 5′ and 3′ mobility junctions of Ef.PcfG within *mobA* of pEF1071 (primers in Table S2.2) were performed on a plasmid preparation from *E. faecalis* (SF24397).

2.7.5: Dendrogram of group introns present in *L. lactis*

A BLASTN search was performed using the Ef.PcfG intron as the query throughout all *L. lactis* genome and plasmid sequences available in the NCBI database. The homologous full-length introns (2492 nt) were compiled and aligned using the Clustal Omega alignment software (Sievers et al. 2011). The introns were organized into a Neighbour-joining tree without distance corrections, which was then exported to the interactive tree of life software (iTOL) for visualization (Letunic and Bork 2016).

2.8: Acknowledgments

We thank Bruce Johnston for providing comments on the manuscript. This work was supported by a discovery grant from the Natural Sciences and Engineering Research Council of Canada to B.C. (227826). F.L.J. received a Graduate Excellence Fellowship from McGill University, a CGS-M Fellowship from Natural Sciences and Engineering Research Council of Canada and a Master's Research Scholarship from Fonds de Recherche en Nature et Technologies.

2.9: References

Balla E, Dicks LM, Du Toit M, Van Der Merwe MJ, Holzapfel WH. 2000. Characterization and cloning of the genes encoding enterocin 1071A and enterocin 1071B, two antimicrobial peptides produced by Enterococcus faecalis BFE 1071. *Appl Environ Microbiol* 66(4):1298-304.

Balla E, Dicks LM. 2005. Molecular analysis of the gene cluster involved in the production and secretion of enterocins 1071A and 1071B and of the genes responsible for the replication and transfer of plasmid pEF1071. *Int J Food Microbiol* 99(1):33-45.

Belhocine K, Plante I, Cousineau B. 2004. Conjugation mediates transfer of the Ll.LtrB group II intron between different bacterial species. *Mol Microbiol* 51(5):1459-69.

Belhocine K, Yam KK, Cousineau B. 2005. Conjugative transfer of the Lactococcus lactis chromosomal sex factor promotes dissemination of the Ll.LtrB group II intron. *J Bacteriol* 187(3):930-9.

Belhocine K, Mak AB, Cousineau B. 2007a. Trans-splicing of the Ll.LtrB group II intron in Lactococcus lactis. *Nucleic Acids Res* 35(7):2257-68.

Belhocine K, Mandilaras V, Yeung B, Cousineau B. 2007b. Conjugative transfer of the Lactococcus lactis sex factor and pRS01 plasmid to Enterococcus faecalis. *FEMS Microbiol Lett* 269(2):289-94.

Chalamcharla VR, Curcio MJ, Belfort M. 2010. Nuclear expression of a group II intron is consistent with spliceosomal intron ancestry. *Genes Dev* 24(8):827-36.

Chen Y, Staddon JH, Dunny GM. 2007. Specificity determinants of conjugative DNA processing in the Enterococcus faecalis plasmid pCF10 and the Lactococcus lactis plasmid pRS01. *Mol Microbiol* 63(5):1549-64.

Chillon I, Martinez-Abarca F, Toro N. 2011. Splicing of the Sinorhizobium meliloti RmInt1 group II intron provides evidence of retroelement behavior. *Nucleic Acids Res* 39(3):1095-104.

Cousineau B, Smith D, Lawrence-Cavanagh S, Mueller JE, Yang J, Mills D, Manias D, Dunny G, Lambowitz AM, Belfort M. 1998. Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell* 94(4):451-62.

Cousineau B, Lawrence S, Smith D, Belfort M. 2000. Retrotransposition of a bacterial group II intron. *Nature* 404(6781):1018-21.

Curcio MJ, Belfort M. 2007. The beginning of the end: links between ancient retroelements and modern telomerases. *Proc Natl Acad Sci U S A* 104(22):9107-8.

Dai L, Zimmerly S. 2002a. The dispersal of five group II introns among natural populations of Escherichia coli. *RNA* 8(10):1294-307.

Dai L, Zimmerly S. 2002b. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res* 30(5):1091-102.

Dai L, Chai D, Gu SQ, Gabel J, Noskov SY, Blocker FJ, Lambowitz AM, Zimmerly S. 2008. A three-dimensional model of a group II intron RNA and its interaction with the intron-encoded reverse transcriptase. *Mol Cell* 30(4):472-85.

Erkus O, de Jager VC, Spus M, van Alen-Boerrigter IJ, van Rijswijck IM, Hazelwood L, Janssen PW, van Hijum SA, Kleerebezem M, Smid EJ. 2013. Multifactorial diversity sustains microbial community stability. *Isme j* 7(11):2126-36.

Fallico V, Ross RP, Fitzgerald GF, McAuliffe O. 2012. Novel conjugative plasmids from the natural isolate Lactococcus lactis subspecies cremoris DPC3758: a repository of genes for the potential improvement of dairy starters. *J Dairy Sci* 95(7):3593-608.

Fedorova O, Zingler N. 2007. Group II introns: structure, folding and splicing mechanism. *Biol Chem* 388(7):665-78.

Fernandez-Lopez M, Munoz-Adelantado E, Gillis M, Willems A, Toro N. 2005. Dispersal and evolution of the Sinorhizobium meliloti group II RmInt1 intron in bacteria that interact with plants. *Mol Biol Evol* 22(6):1518-28.

Gasson M, Godon J, Pillidge C, Eaton T, Jury K, Shearman C. 1995. Characterization and exploitation of conjugation in Lactococcus lactis. *International Dairy Journal* 5(8):757-762.

Hirt H, Manias DA, Bryan EM, Klein JR, Marklund JK, Staddon JH, Paustian ML, Kapur V, Dunny GM. 2005. Characterization of the pheromone response of the Enterococcus faecalis conjugative plasmid pCF10: complete sequence and comparative analysis of the transcriptional and phenotypic responses of pCF10-containing cells to pheromone induction. *J Bacteriol* 187(3):1044-54.

Ichiyanagi K, Beauregard A, Lawrence S, Smith D, Cousineau B, Belfort M. 2002. Retrotransposition of the Ll.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. *Mol Microbiol* 46(5):1259-72.

Ilyina TV, Koonin EV. 1992. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Res* 20(13):3279-85.

Klein JR, Dunny GM. 2002. Bacterial group II introns and their association with mobile genetic elements. *Front Biosci* 7:d1843-56.

Kuipers OP, Beerthuyzen MM, Siezen RJ, De Vos WM. 1993. Characterization of the nisin gene cluster nisABTCIPR of Lactococcus lactis. Requirement of expression of the nisA and nisI genes for development of immunity. *Eur J Biochem* 216(1):281-91.

Ladero V, Del Rio B, Linares DM, Fernandez M, Mayo B, Martin MC, Alvarez MA. 2015. Draft Genome Sequence of the Putrescine-Producing Strain Lactococcus lactis subsp. lactis 1AA59. *Genome Announc* 3(3).

Lambie SC, Altermann E, Leahy SC, Kelly WJ. 2014. Draft Genome Sequence of Lactococcus lactis subsp. cremoris HPT, the First Defined-Strain Dairy Starter Culture Bacterium. *Genome Announc* 2(2).

Lambowitz AM, Zimmerly S. 2011. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol* 3(8):a003616.

Lambowitz AM, Belfort M. 2015. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol Spectr* 3(1).

Leclercq S, Giraud I, Cordaux R. 2011. Remarkable abundance and evolution of mobile group II introns in Wolbachia bacterial endosymbionts. *Mol Biol Evol* 28(1):685-97.

Leclercq S, Cordaux R. 2012. Selection-driven extinction dynamics for group II introns in Enterobacteriales. *PLoS One* 7(12):e52268.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res*.

Linares DM, Kok J, Poolman B. 2010. Genome sequences of Lactococcus lactis MG1363 (revised) and NZ9000 and comparative physiological studies. *J Bacteriol* 192(21):5806-12.

Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, Pavlov A, Pavlova N, Karamychev V, Polouchine N et al. . 2006. Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci U S A* 103(42):15611-6.

Malik HS, Burke WD, Eickbush TH. 1999. The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol* 16(6):793-805.

McBride SM, Fischetti VA, Leblanc DJ, Moellering RC, Jr., Gilmore MS. 2007. Genetic diversity among Enterococcus faecalis. *PLoS One* 2(7):e582.

Mills DA, McKay LL, Dunny GM. 1996. Splicing of a group II intron involved in the conjugative transfer of pRS01 in lactococci. *J Bacteriol* 178(12):3531-8.

Mills DA, Manias DA, McKay LL, Dunny GM. 1997. Homing of a group II intron from Lactococcus lactis subsp. lactis ML3. *J Bacteriol* 179(19):6107-11.

Mohr G, Ghanem E, Lambowitz AM. 2010. Mechanisms used for genomic proliferation by thermophilic group II introns. *PLoS Biol* 8(6):e1000391.

Monat C, Quiroga C, Laroche-Johnston F, Cousineau B. 2015. The Ll.LtrB intron from Lactococcus lactis excises as circles in vivo: insights into the group II intron circularization pathway. *RNA* 21(7):1286-93.

Monat C, Cousineau B. 2016. Circularization pathway of a bacterial group II intron. *Nucleic Acids Res* 44(4):1845-53.

Nisa-Martinez R, Jimenez-Zurdo JI, Martinez-Abarca F, Munoz-Adelantado E, Toro N. 2007. Dispersion of the RmInt1 group II intron in the Sinorhizobium meliloti genome upon acquisition by conjugative transfer. *Nucleic Acids Res* 35(1):214-22.

O'Sullivan D, Twomey DP, Coffey A, Hill C, Fitzgerald GF, Ross RP. 2000. Novel type I restriction specificities through domain shuffling of HsdS subunits in Lactococcus lactis. *Mol Microbiol* 36(4):866-75.

Pansegrau W, Schroder W, Lanka E. 1994. Concerted action of three distinct domains in the DNA cleaving-joining reaction catalyzed by relaxase (TraI) of conjugative plasmid RP4. *J Biol Chem* 269(4):2782-9.

Paulsen IT, Banerjei L, Myers GS, Nelson KE, Seshadri R, Read TD, Fouts DE, Eisen JA, Gill SR, Heidelberg JF et al. . 2003. Role of mobile DNA in the evolution of vancomycin-resistant Enterococcus faecalis. *Science* 299(5615):2071-4.

Plante I, Cousineau B. 2006. Restriction for gene insertion within the Lactococcus lactis Ll.LtrB group II intron. *RNA* 12(11):1980-92.

Ruiz-Barba JL, Piard JC, Jimenez-Diaz R. 1991. Plasmid profiles and curing of plasmids in Lactobacillus plantarum strains isolated from green olive fermentations. *J Appl Bacteriol* 71(5):417-21.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J et al. . 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539.

Siezen RJ, Renckens B, van Swam I, Peters S, van Kranenburg R, Kleerebezem M, de Vos WM. 2005. Complete sequences of four plasmids of Lactococcus lactis subsp. cremoris SK11 reveal extensive adaptation to the dairy environment. *Appl Environ Microbiol* 71(12):8371-82.

Simon DM, Kelchner SA, Zimmerly S. 2009. A broadscale phylogenetic analysis of group II intron RNAs and intron-encoded reverse transcriptases. *Mol Biol Evol* 26(12):2795-808.

Staddon JH, Bryan EM, Manias DA, Dunny GM. 2004. Conserved target for group II intron insertion in relaxase genes of conjugative elements of gram-positive bacteria. *J Bacteriol* 186(8):2393-401.

Staddon JH, Bryan EM, Manias DA, Chen Y, Dunny GM. 2006. Genetic characterization of the conjugative DNA processing system of enterococcal plasmid pCF10. *Plasmid* 56(2):102-11.

Toor N, Hausner G, Zimmerly S. 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA* 7(8):1142-52.

Toro N, Jimenez-Zurdo JI, Garcia-Rodriguez FM. 2007. Bacterial group II introns: not just splicing. *FEMS Microbiol Rev* 31(3):342-58.

Toro N, Martinez-Abarca F. 2013. Comprehensive phylogenetic analysis of bacterial group II intron-encoded ORFs lacking the DNA endonuclease domain reveals new varieties. *PLoS One* 8(1):e55102.

Tourasse NJ, Kolsto AB. 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(14):4529-48.

Wagner A. 2006. Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. *Mol Biol Evol* 23(4):723-33.

Wegmann U, O'Connell-Motherway M, Zomer A, Buist G, Shearman C, Canchaya C, Ventura M, Goesmann A, Gasson MJ, Kuipers OP et al. . 2007. Complete genome sequence of the prototype lactic acid bacterium Lactococcus lactis subsp. cremoris MG1363. *J Bacteriol* 189(8):3256-70.

Yang X, Wang Y, Huo G. 2013. Complete Genome Sequence of Lactococcus lactis subsp. lactis KLDS4.0325. *Genome Announc* 1(6).

Yao J, Zhong J, Lambowitz AM. 2005. Gene targeting using randomly inserted group II introns (targetrons) recovered from an Escherichia coli gene disruption library. *Nucleic Acids Res* 33(10):3351-62.

Zimmerly S, Semper C. 2015. Evolution of group II introns. *Mob DNA* 6:7.

2.10: Figures

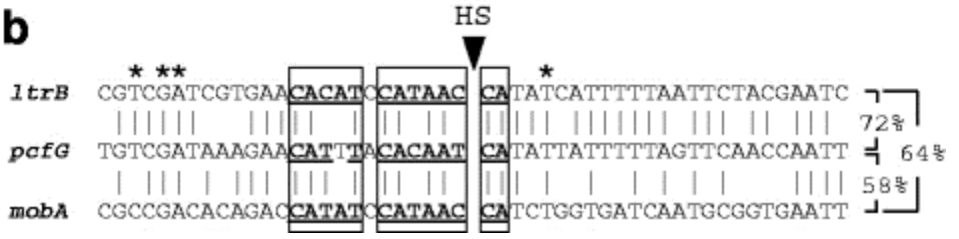
Figure 2.1

a

Mutation (Ef.PcfG / Ll.LtrB)	Location of the mutation (IEP region)	IEP amino acid (IepG / LtrA)
Mut #1 (G_{515} / A_{515})	Domain III	NA
Mut #2 (A_{754} / G_{754})	Domain IV (IEP-RT)	Ser_{61} (P) / Gly_{61} (0)
Mut #3 (G_{996} / A_{996})	Domain IV (IEP-RT)	Thr_{141} (P) / Thr_{141} (P)
Mut #4 (G_{1161} / A_{1161})	Domain IV (IEP-RT)	Ala_{196} (H) / Ala_{196} (H)
Mut #5 (A_{1438} / C_{1438})	Domain IV (IEP-RT)	Thr_{289} (P) / Pro_{289} (0)
Mut #6 (A_{1588} / G_{1588})	Domain IV (IEP-RT)	Lys_{339} (+) / Glu_{339} (-)
Mut #7 (A_{2128} / G_{2128})	Domain IV (IEP-DB)	Lys_{519} (+) / Glu_{519} (-)
Mut #8 (G_{2195} / A_{2195})	Domain IV (IEP-DB)	Arg_{541} (+) / Lys_{541} (+)

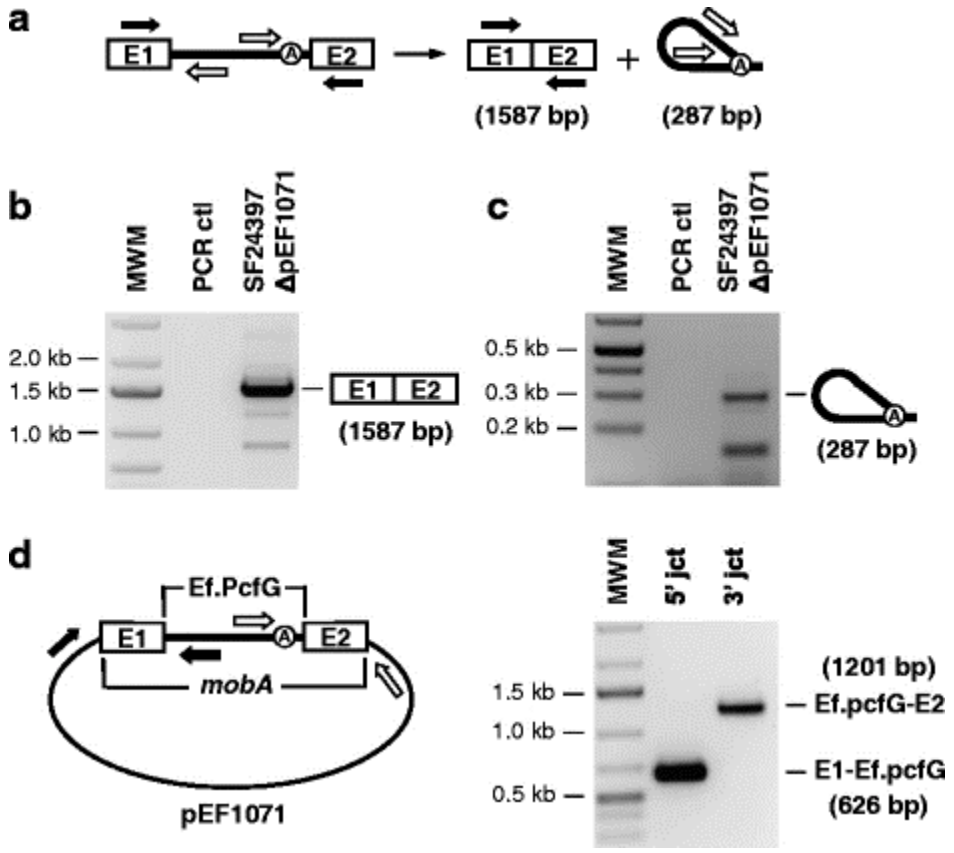
IEP, Intron-Encoded protein; RT, reverse transcriptase domain; DB, DNA binding domain (0), neutral aa; (+), positively charged aa; (-) negatively charged aa; (H) Hydrophobic aa; (P) polar aa; NA, not applicable

b



Comparison between the Ef.PcfG and Ll.LtrB group II introns and various relaxase genes from *L. lactis* and *E. faecalis*. (**a**) Description and location of the eight point mutations (Mut #1-Mut #8) that distinguish Ef.PcfG from Ll.LtrB within pRS01. All mutations except one (Mut #1, domain III) are located in domain IV within the intron-encoded protein (IEP) gene. Five mutations (Mut #2 to Mut #6) are located in the reverse transcriptase domain (IEP-RT) while two (Mut #7, Mut #8) are within the DNA binding domain (IEP-DB) of the IEP. Among the mutations located within the IEP, five are missense (Mut #2, Mut #5 to Mut #8) while the other two are silent (Mut #3, Mut #4). The nucleotide (nt) and amino acid (aa) numberings are in reference to the first nt of the intron (2492 nt) and the first aa of the IEP (599 aa) respectively. (**b**) Sequence alignement of the intron insertion sites in various relaxase genes from *L. lactis* (*ltrB*) and *E. faecalis* (*pcfg* and *mobA*). The intron insertion site or homing site (HS) (black arrowhead) and the percentage of homology between the three sequences are depicted. The IBS1, IBS2 and ∂' sequences are boxed and the nts that are complementary to the EBS1, EBS2 and ∂ sequences are bolded and underlined. Nts of *ltrB*-HS that are known to interact with LtrA are denoted by an asterisk (Yao et al. 2005).

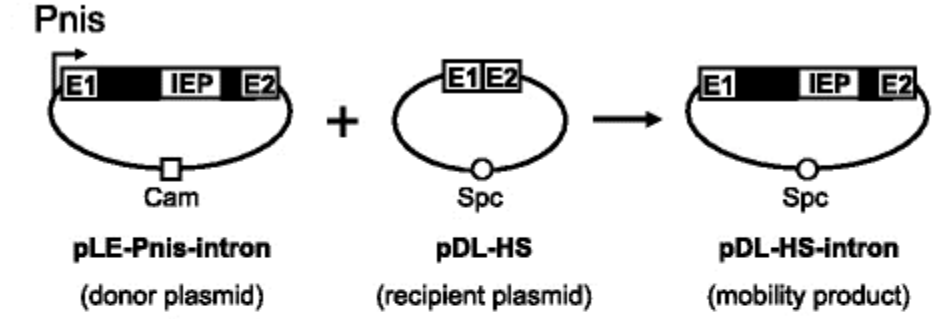
Figure 2.2



Splicing and mobility of the Ef.PcfG intron in its native environment. (**a**) Schematic of the group II intron splicing pathway. Position of the primers (Table S2.2) used to amplify ligated exons (E1/E2) (black arrows, 1587 bp) and the intron splice junction (open arrows, 287 bp) by RT-PCR is depicted. RT-PCR amplifications of ligated exons (**b**) and of intron splice junctions (**c**) were performed on total RNA extracts from *E. faecalis* (SF24397ΔpEF1071). The RT-PCR amplicons corresponding to *pcfG* ligated exons (*B*, 1587 bp) and Ef.PcfG spliced junction (*C*, 287 bp) were excised and directly sequenced. (**d**) Mobility efficiency of Ef.PcfG to the relaxase *mobA* gene (E1/E2) on pEF1071. Position of the primers (Table S2.2) used to amplify the 5′ (black arrows, 626 bp, E1-Ef.PcfG) and 3′ (open arrows, 1201 bp, Ef.PcfG-E2) junctions of Ef.PcfG mobility products in *mobA* by PCR is depicted.

Figure 2.3

a

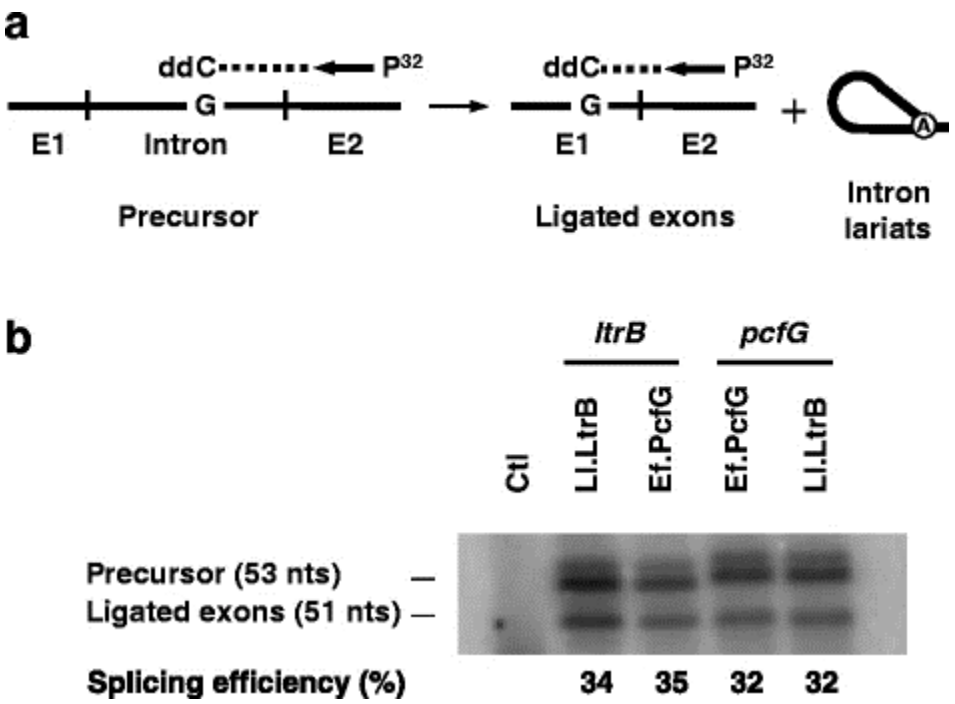


b

	Wild-type flanking exons		Exchanged flanking exons	
	Ll.LtrB (donor)	Ef.PcfG (donor)	Ll.LtrB (donor)	Ef.PcfG (donor)
ltrB-HS (recipient)	65.0% ± 13.9%	42.0% ± 18.0%	70.5% ± 5.4%	44.5% ± 12.3%
pcfG-HS (recipient)	85.5% ± 10.8%	89.0% ± 5.4%	93.8% ± 3.8%	88.2% ± 7.0%

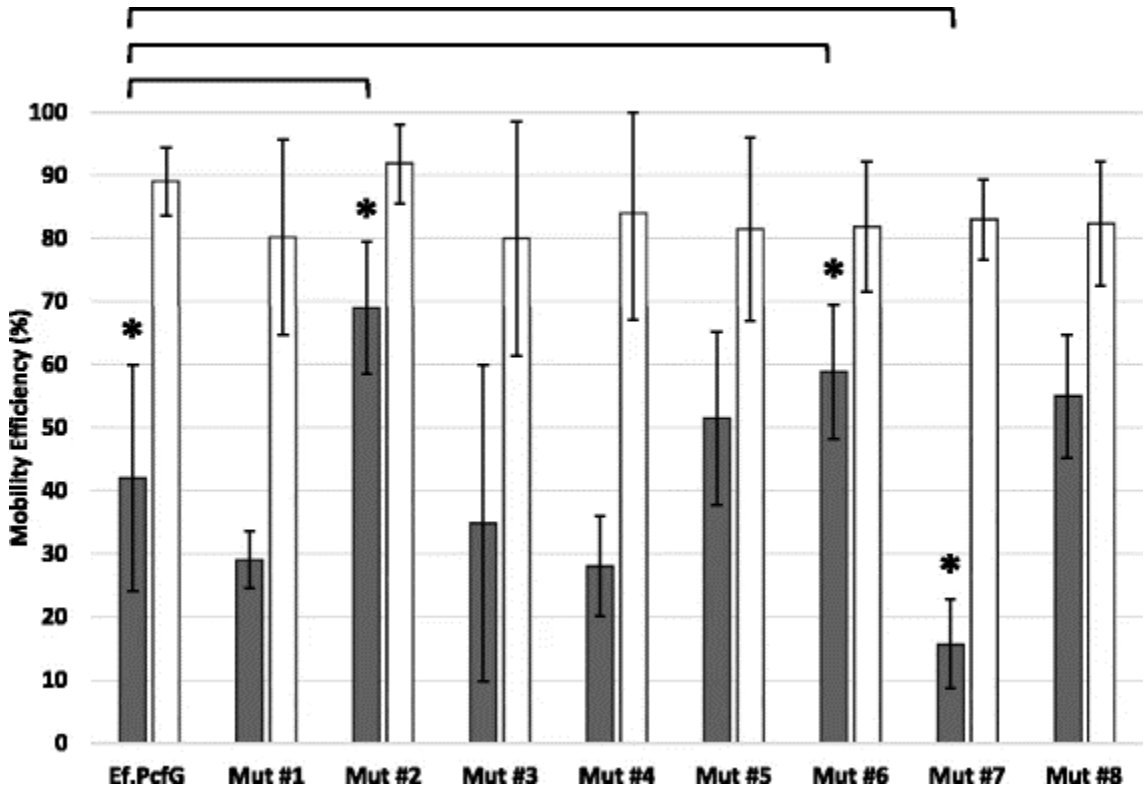
Mobility efficiency of Ef.PcfG and Ll.LtrB in *L. lactis*. (**a**) Schematic of the two-plasmid intron mobility assay. The assay consists of co-transforming both an intron donor and an intron recipient plasmid in *L. lactis* cells (NZ9800Δ*ltrB*) and monitoring for the appearance of intron mobility products. The intron donor plasmid harbors the *pcfG* or *ltrB* genes interrupted by their cognate or exchanged introns under the control of the nisin-inducible promoter (Pnis). The recipient plasmid contains either the *ltrB*-HS or the *pcfG*-HS (E1/E2). Upon nisin induction the intron can move from the donor to the recipient plasmid generating mobility products. Plasmid mixes (donor, recipient, mobility product) from independent mobility assays are recovered and the intron mobility efficiency is calculated as the percentage of recipient plasmids invaded by the intron (mobility product / (recipient + mobility product)). (**b**) Mobility efficiency of Ef.PcfG and Ll.LtrB at their own and each other's homing sites. Two independent series of mobility assays were performed by expressing both introns flanked by either their wild-type (Wild-type flanking exons) or swapped exons (Exchanged flanking exons). Regardless of their flanking exons (Wild-type or exchanged), the mobility efficiency of both introns to the *pcfG*-HS is significantly higher ($p<0.05$) than their efficiency towards the *ltrB*-HS while the mobility efficiency of Ll.LtrB to the *ltrB*-HS is significantly higher ($p<0.05$) than Ef.PcfG. Each mobility efficiency value corresponds to the average of six independent mobility assays.

Figure 2.4



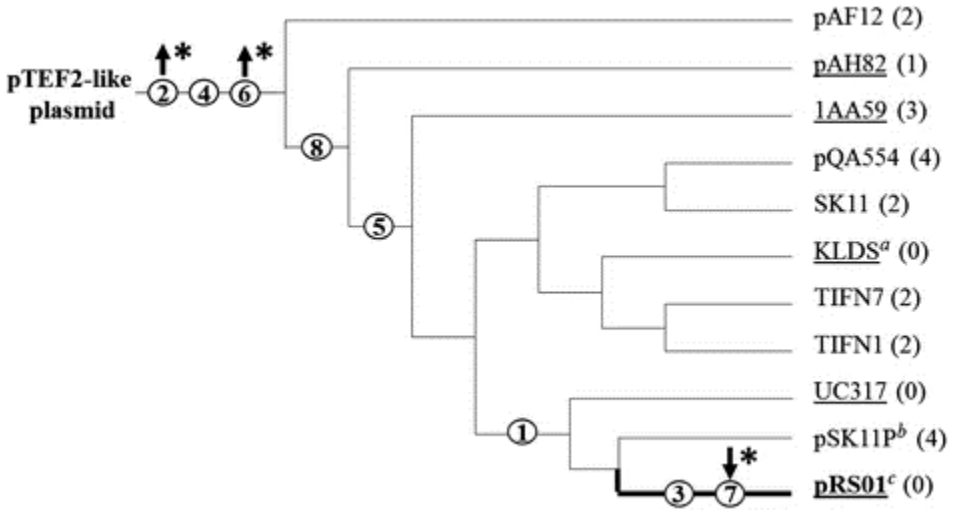
Splicing efficiency of Ef.PcfG and Ll.LtrB in *L. lactis*. (**a**) Group II intron splicing efficiency assessed by poisoned primer extension assay. This assay monitors splicing efficiency by comparing the relative abundance of precursor and ligated exons from total RNA extracts. A P^{32}-labeled primer (Table S2.2) complementary to exon 2 was extended from both the precusor and the ligated exons in the presence of a high concentration of ddCTP. Since the sequence of the two RNAs are different after the exon 2 junction the first G residue encountered is at a different distance from the primer generating differently sized bands for the precursor (53 nts) and the ligated exons (51 nts). (**b**) The splicing efficiency (%) of Ef.PcfG and Ll.LtrB was assessed from their wild-type and exchanged flanking exons and calculated as the relative intensity of the ligated exons and precursor bands (ligated exons / precursor + ligated exons).

Figure 2.5



Graphical representation of the mobility efficiency of Ef.PcfG, flanked by its wild-type exons, to both the *ltrB*-HS (black bars) and the *pcfG*-HS (open bars). The eight point mutations between Ef.PcfG and Ll.LtrB were independently engineered within Ef.PcfG (Mut #1 to Mut #8) and their mobility efficiencies are compared to wild-type Ef.PcfG (*, $p<0.05$). All mutants are significantly more efficient at homing to the *pcfG*-HS than the *ltrB*-HS ($p<0.05$). Each mobility efficiency value corresponds to the average of six independent mobility assays.

Figure 2.6



Dendrogram of mutation accumulation between Ef.PcfG from *E. faecalis* and the various Ll.LtrB introns identified from *L. lactis*. The root of the tree is depicted as the Ef.PcfG intron (pTEF-2-like plasmid), which likely disseminated throughout *L. lactis* following a single horizontal transfer event. The plasmids harboring the Ef.PcfG (pTEF-2-like plasmid) and Ll.LtrB (pRS01) introns discussed in this study are bolded. The eight point mutations (Mut #1 to Mut #8) that distinguish these two introns are shown in circles, with an asterisk adjacent to mutations conferring significant increases (upwards arrow) and decreases (downwards arrow) in mobility efficiency to the *ltrB*-HS (p<0.05). Introns found in the chromosome are designated by the strain's name, whereas introns found within plasmids are represented by the plasmid's name. Strains or plasmids that are underlined belong to *L. lactis* subsp. *lactis*, whereas strains or plasmids that are not underlined belong to *L. lactis* subsp. *cremoris*. Thick branches of the dendrogram represent an insertion event into an HIHN-H relaxase motif, whereas thin branches represent an insertion event into an HLHN-H relaxase motif; the only exception being the intron present in the chromosome of SK11, which is likely a retrotransposition event into an ectopic site (cell surface protein). Numbers between parentheses denote the amount of additional mutations that distinguish a particular intron from Ef.PcfG in relationship to its position in the dendrogram. Three groups are present in the dendrogram which encompass a number of additional identical introns: Group (*a*) contains 6 additional introns (HP, TIFN5, TIFN6, FG2, B40, LMG6897), group (*b*) contains 3 additional intron (p3, SK110, AM2), and group (*c*) contains 4 additional introns (MG1363, pFI430, NZ9000, NCDO763) (Table S2.3).

2.11: Tables

Table S2.1:

Plasmids used in the study

Plasmid name	Size and Antibiotic Resistance [a]	Description
pLE-Pnis-*ltrB*E1-Ll.LtrB-*ltrB*E2	13.3 kb, Cam^R	'Intron Donor'[b], with WT Ll.LtrB intron
pLE-Pnis-*pcfG*E1-Ef.PcfG-*pcfG*E2	13.1 kb, Cam^R	'Intron Donor'[b], with WT Ef.PcfG intron
pLE-Pnis-*ltrB*E1-Ef.PcfG-*ltrB*E2	13.3 kb, Cam^R	'Intron Donor'[b], with Ef.PcfG intron flanked by ltrB exons
pLE-Pnis-*pcfG*E1-Ll.LtrB-*pcfG*E2	13.1 kb, Cam^R	'Intron Donor'[b], with Ll.LtrB intron flanked by pcfG exons
pLE-Pnis-*pcfG*E1-Ef.PcfG-*pcfG*E2-Mut #1	13.1 kb, Cam^R	'Intron Donor'[b], with Ef.PcfG intron containing A→G mutation
pLE-Pnis-*pcfG*E1-Ef.PcfG-*pcfG*E2-Mut #2	13.1 kb, Cam^R	'Intron Donor'[b], with Ef.PcfG intron containing A→G mutation
pLE-Pnis-*pcfG*E1-Ef.PcfG-*pcfG*E2-Mut #3	13.1 kb, Cam^R	'Intron Donor'[b], with Ef.PcfG intron containing A→G mutation
pLE-Pnis-*pcfG*E1-Ef.PcfG-*pcfG*E2-Mut #4	13.1 kb, Cam^R	'Intron Donor'[b], with Ef.PcfG intron containing A→G mutation
pLE-Pnis-*pcfG*E1-Ef.PcfG-*pcfG*E2-Mut #5	13.1 kb, Cam^R	'Intron Donor'[b], with Ef.PcfG intron containing A→C mutation
pLE-Pnis-*pcfG*E1-Ef.PcfG-*pcfG*E2-Mut #6	13.1 kb, Cam^R	'Intron Donor'[b], with Ef.PcfG intron containing A→G mutation
pLE-Pnis-*pcfG*E1-Ef.PcfG-*pcfG*E2-Mut #7	13.1 kb, Cam^R	'Intron Donor'[b], with Ef.PcfG intron containing A→G mutation
pLE-Pnis-*pcfG*E1-Ef.PcfG-*pcfG*E2-Mut #8	13.1 kb, Cam^R	'Intron Donor'[b], with Ef.PcfG intron containing G→A mutation
pDL-*ltrB*-HS	7 kb, Spc^R	'Intron Recipient', with native Ll.LtrB Homing Site (*ltrB* gene)
pDL-*pcfG*-HS	7.3 kb, Spc^R	'Intron Recipient', with native Ef.PcfG Homing Site (*pcfG* gene)

a. Cam^R, Chloramphenicol resistance at (10μg/ml); Spc^R, Spectinomycin resistance at (300μg/ml).

b. Intron-interrupted genes (*ltrB* or *pcfG*) in 'Intron Donor' plasmids are under the control of a Nisin-inducible promoter (Pnis).

Table S2.2:

Primers used in the study

Primer name	Sequence (5′-3′) [a]
Poisoned primer for *ltrB* exon 2	GCCAGTATAAAGATTCGTAGAAT
Poisoned primer for *pcfG* exon 2	ACCTGTTTTTAAATTGGTTGAAC
Released intron (RT)	CGATTGTCTTTAGGTAACTCAT
Released intron (PCR)	CTCTTGTTGTATGCTTTCATTG
	CTTTCCAACCGTGCTCTGTTC
Ligated exons and *pcfG*-HS fragment (RT)	AATGTCGGTTTGCTTCTCTG
Ligated exons and *pcfG*-HS fragment (PCR)	GCTTGCTCATATATTGAGATTGC
	TACGGCTTGTATTTCATGAAGCT
5′ mobility junction of Ef.PcfG in *mobA* of pEF1071 (PCR)	CTGGGACAATCAAGCAACCAAA
	CTTTCCAACCGTGCTCTGTTC
3′ mobility junction of Ef.PcfG in *mobA* of pEF1071 (PCR)	CTCTTGTTGTATGCTTTCATTG
	AACCACCGATAAAATTCGTCCA
Mut #1 Within Ef.PcfG (Nt 515/2492)	TTTACATGGC**A**AAGGGGTACAG
	GGGCGTTATCCTTCTCAG
Mut #2 Within Ef.PcfG (Nt 754/2492)	TACAGCGGAT**G**GCTTTAGTGAAG
	TCATCTAATATTCCTTTTGTGGAAG
Mut #3 Within Ef.PcfG (Nt 996/2492)	GCTGTCACAC**A**GCTTTGAAAAC
	TTCGTTGAGGTCTAAAACC
Mut #4 Within Ef.PcfG (Nt 1161/2492)	TTCTAAAAGC**A**GGTTATCTGGAAAAC
	ATTTATAAATCAATTGGCTCATTTTC
Mut #5 Within Ef.PcfG (Nt 1438/2492)	TAAAAGATTA**C**CCACACTCCCC
	CGTTTTTCTTGATATTCTAAAAGAAC
Mut #6 Within Ef.PcfG (Nt 1588/2492)	CTAAAAATG**G**AATTGAGTGAAGAAAAAAC
	TTGTTATGAATAAAAAGTTTTAATTGTTC
Mut #7 Within Ef.PcfG (Nt 2128/2492)	ATTTACGGAT**G**AGATAAGTCAAGC
	TGATAAGGGGATTTACATTCAC
Mut #8 Within Ef.PcfG (Nt 2195/2492)	TTAAAAGCTA**A**ATGTTGTGAATTATG
	CCTGTTTTCAAGAGTATTCC

a. Pairs of primers which were used for mutagenesis have the substituted nucleotides in bold.

Table S2.3:

Point mutations between Ef.PcfG and highly homologous full-length Ll.LtrB variants in *L. lactis*

Strain of Origin	Intron Location	Interrupted motif	Presence of Mut #1-#8	Additional Mutations			
L. lactis subsp. *cremoris* DPC3758 (O'Sullivan et al. 2000)	Plasmid: pAF12	Relaxase HLHN-H	Mut #2, #4, #6	Nt 1421, A→G (RT) Glu(-)→Gly(o))	Nt 1975, G→A (X) (Ala(o)→Thr(p))		
L. lactis lactis subsp. *lactis* DPC220 (Fallico et al. 2012)	Plasmid: pAH82	Relaxase HLHN-H	Mut #2, #4, #6, #8	Nt 2223, T→C (En) (Silent Mutation)			
L. lactis subsp. *lactis* 1AA59 (Ladero et al. 2015)	Chromosomal (contig)	Relaxase HLHN-H	Mut #2, #4, #5, #6, #8	Nt 59, C→G (DI)	Nt 713, A→G (RT) (Lys(+)→Arg(+))	Nt 2216, C→T (En) (Thr(p)→Ile(o))	
L. lactis subsp. *cremoris* A76 (Bolotin et al. unpublished data)	Plasmid: pQA554	Relaxase HLHN-H	Mut #2, #4, #5, #6, #8	Nt 308, ΔT (DI)	Nt 560, +G (DIVa)	Nt 1230, T→C (RT) (Silent Mutation)	Nt 2109, C→T (D) (Silent Mutation)
. *lactis* subsp. *cremoris* SK11 (Makarova et al. 2006)	Chromosomal	Complement Strand of Cell Surface Protein	Mut #2, #4, #5, #6, #8	Nt 502, C→T (DIII)	Nt 2109, C→T (D) (Silent Mutation)		
L. lactis subsp. *lactis* KLDS (Yang et al. 2013)	Chromosomal	Relaxase HLHN-H	Mut #2, #4, #5, #6, #8	N/A			
L. lactis subsp. *cremoris* TIFN6 (Erkus et al. 2013)	Chromosomal (contig)	Relaxase HLHN-H	Mut #2, #4, #5, #6, #8	N/A			
L. lactis subsp. *cremoris* TIFN5 (Erkus et al. 2013)	Chromosomal (contig)	Relaxase HLHN-H	Mut #2, #4, #5, #6, #8	N/A			
L. lactis subsp. *cremoris* HP (Lambie et al. 2014)	Chromosomal (contig)	Relaxase HLHN-H	Mut #2, #4, #5, #6, 8	N/A			
L. lactis subsp. *cremoris* FG2 (Wels et al. unpublished data)	Chromosomal (contig)	Relaxase HLHN-H	Mut #2, #4, #5, #6, 8	N/A			
L. lactis subsp. *cremoris* B40 (Wels et al. unpublished data)	Chromosomal (contig)	Relaxase HLHN-H	Mut #2, #4, #5, #6, 8	N/A			
L. lactis subsp. *cremoris* LMG6897 (Wels et al. unpublished data)	Chromosomal (contig)	Relaxase HLHN-H	Mut #2, #4, #5, #6, 8	N/A			

L. lactis subsp. *cremoris* TIFN7 (Erkus et al. 2013)	Chromosomal (contig)	Relaxase HLHN-H	Mut #2, #4, #5, #6, #8	Nt 771, ΔA (RT)	Nt 773, ΔA (RT)		
L. lactis subsp. *cremoris* TIFN1 (Erkus et al. 2013)	Chromosomal (contig)	Relaxase HLHN-H	Mut #2, #4, #5, #6, #8	Nt 174, ΔA (DI)	Nt 175, ΔT (DI)		
L. lactis subsp. *lactis* UC317 (Wels et al. unpublished data)	Chromosomal (contig)	Relaxase HLHN-H	Mut #1, #2, #4, #5, #6, #8	N/A			
L. lactis subsp. *cremoris* SK11 (Makarova et al. 2006)	Plasmid: pSK11P	Relaxase HLHN-H	Mut #1, #2, #4, #5, #6, #8	Nt 756, C→T (RT) (Silent Mutation)	Nt 1593, G→T (RT) (Leu(o)→ Phe(o))	Nt 1732, C→T (X) (Leu(o)→ Phe(o))	Nt 1958, G→A (X) (Ser(p)→ Asn(p))
L. lactis subsp. *cremoris* SK11 (Siezen et al. 2005)	Plasmid: Plasmid 3 (p3)	Relaxase HLHN-H	Mut #1, #2, #4, #5, #6, #8	Nt 756, C→T (RT) (Silent Mutation)	Nt 1593, G→T (RT) (Leu(o)→ Phe(o))	Nt 1732, C→T (X) (Leu(o)→ Phe(o))	Nt 1958, G→A (X) (Ser(p)→ Asn(p))
L. lactis subsp. *cremoris* SK110 (Wels et al. unpublished data)	Chromosomal (contig)	Relaxase HLHN-H	Mut #1, #2, #4, #5, #6, #8	Nt 756, C→T (RT) (Silent Mutation)	Nt 1593, G→T (RT) (Leu(o)→ Phe(o))	Nt 1732, C→T (X) (Leu(o)→ Phe(o))	Nt 1958, G→A (X) (Ser(p)→ Asn(p))
L. lactis subsp. *cremoris* AM2 (Wels et al. unpublished data)	Plasmid: Plasmid 3 (p3)	Relaxase HLHN-H	Mut #1, #2, #4, #5, #6, #8	Nt 756, C→T (RT) (Silent Mutation)	Nt 1593, G→T (RT) (Leu(o)→ Phe(o))	Nt 1732, C→T (X) (Leu(o)→ Phe(o))	Nt 1958, G→A (X) (Ser(p)→ Asn(p))
L. lactis subsp. *lactis* ML3 (Mills et al. 1996)	Plasmid: pRS01	Relaxase HIHN-H	Mut #1, #2, #3, #4, #5, #6, #7, #8	N/A			
L. lactis subsp. *cremoris* NZ9000 (Linares et al. 2010)	Chromosomal	Relaxase HIHN-H	Mut #1, #2, #3, #4, #5, #6, #7, #8	N/A			
L. lactis subsp. *cremoris* MG1363 (Gasson et al. 1995)	Plasmid: pFI430	Relaxase HIHN-H	Mut #1, #2, #3, #4, #5, #6, #7, #8	N/A			
L. lactis subsp. *cremoris* MG1363 (Wegmann et al. 2007)	Chromosomal	Relaxase HIHN-H	Mut #1, #2, #3, #4, #5, #6, #7, #8	N/A			
L. lactis subsp. *cremoris* NCDO763 (Wels et al. unpublished data)	Chromosomal	Relaxase HIHN-H	Mut #1, #2, #3, #4, #5, #6, #7, #8	N/A			

Chapter 3:
Bacterial group II introns generate genetic diversity by circularization and *trans*-splicing from a population of intron-invaded mRNAs

3.1: Preface

As discussed in Chapter 1, throughout the course of evolution bacterial group II introns have given rise to genetic elements in eukaryotes such as spliceosomal introns and the snRNAs of the spliceosome, which are beneficial to their host by increasing genetic diversity (Irimia and Roy 2014; Bush et al. 2017). However, bacterial group II introns themselves have always been perceived solely as selfish genetic elements that are detrimental to their bacterial hosts (Dai and Zimmerly 2002; Leclercq and Cordaux 2012). We thus studied the function of circularization for the Ll.LtrB bacterial group II intron, hypothesizing that a secondary splicing pathway conserved throughout many intron subtypes may provide a beneficial function for their hosts.

Here we elucidated the pathway through which group II introns incorporate additional nucleotides in their circle splice junctions (Monat et al. 2015; Monat and Cousineau 2016). Through the isolation of splicing intermediates, we were able to propose a new group II intron splicing pathway that combines aspects of group II intron branching and circularization. Our results suggest that although group II introns are present at low copy numbers within bacteria, they can significantly alter the bacterial host's transcriptome using this new splicing pathway. Finally, we provide experimental evidence to suggest that group II introns may be beneficial to their bacterial hosts by increasing genetic diversity.

This chapter was adapted from the following manuscript: "Bacterial group II introns generate genetic diversity by circularization and *trans*-splicing from a population of intron-invaded mRNAs". **Félix LaRoche-Johnston**, Caroline Monat, Samy Coulombe and Benoit Cousineau. *PLoS Genetics*, 2018;14(11):e1007792.

PLOS | GENETICS

RESEARCH ARTICLE

Bacterial group II introns generate genetic diversity by circularization and *trans*-splicing from a population of intron-invaded mRNAs

Félix LaRoche-Johnston, Caroline Monat, Samy Coulombe, Benoit Cousineau*

Department of Microbiology and Immunology, Microbiome and Disease Tolerance Centre (MDTC), McGill University, Montréal, Québec, Canada

3.2: Summary

Group II introns are ancient retroelements that significantly shaped the origin and evolution of contemporary eukaryotic genomes (Lambowitz and Belfort 2015; Zimmerly and Semper 2015; Novikova and Belfort 2017). These self-splicing ribozymes share a common ancestor with the telomerase enzyme, the spliceosome machinery as well as the highly abundant spliceosomal introns and non-LTR retroelements (Lambowitz and Belfort 2015; Zimmerly and Semper 2015; Novikova and Belfort 2017). More than half of the human genome thus consists of various elements that evolved from ancient group II introns, which altogether significantly contribute to key functions and genetic diversity in eukaryotes (Lambowitz and Belfort 2015; Zimmerly and Semper 2015; Novikova and Belfort 2017). Similarly, group II intron-related elements in bacteria such as abortive phage infection (Abi) retroelements, diversity generating retroelements (DGRs) and some CRISPR-Cas systems have evolved to confer important functions to their hosts (Lambowitz and Belfort 2015; Zimmerly and Semper 2015). In sharp contrast, since bacterial group II introns are scarce, irregularly distributed and frequently spread by lateral transfer (Lambowitz and Belfort 2015; Zimmerly and Semper 2015; LaRoche-Johnston et al. 2016; Novikova and Belfort 2017), they have mainly been considered as selfish retromobile elements with no beneficial function to their host (Dai and Zimmerly 2002). Here we unveil a new group II intron function that generates genetic diversity at the RNA level in bacterial cells. We demonstrate that Ll.LtrB, the model group II intron from *Lactococcus lactis*, recognizes specific sequence motifs within cellular mRNAs by base pairing, and invades them by reverse splicing. Subsequent

splicing of ectopically inserted Ll.LtrB, through circularization, induces a novel *trans*-splicing pathway that generates exon 1-mRNA and mRNA-mRNA intergenic chimeras. Our data also show that recognition of upstream alternative circularization sites on intron-interrupted mRNAs release Ll.LtrB circles harboring mRNA fragments of various lengths at their splice junction. Intergenic *trans*-splicing and alternative circularization both produce novel group II intron splicing products with potential new functions. Overall, this work describes new splicing pathways in bacteria that generate, similarly to the spliceosome in eukaryotes, genetic diversity at the RNA level while providing additional functional and evolutionary links between group II introns, spliceosomal introns and the spliceosome.

3.3: Introduction

Bacterial group II introns are large RNA enzymes that mostly behave as retromobile elements (Dai and Zimmerly 2002). Following their autocatalytic excision from interrupted RNA transcripts, they can reinsert within identical or similar DNA target sequences by retrohoming or retrotransposition, respectively (Cousineau et al. 1998; Cousineau et al. 2000; Ichiyanagi et al. 2002). These retromobile genetic elements are present in archaea, bacteria, and bacterial-derived organelles such as plant and fungal mitochondria, and plant chloroplasts (Lambowitz and Zimmerly 2011). While group II introns are somewhat infrequent in archaea, roughly one quarter of all sequenced bacterial genomes harbor one to a few copies displaying a broad phylogenetic distribution in the bacterial kingdom (Candales et al. 2012). In sharp contrast, no functional group II introns were yet described in the nuclear genome of eukaryotes where they seem to be functionally excluded (Chalamcharla et al. 2010). Although mitochondrial and chloroplastic group II introns mainly interrupt housekeeping genes, bacterial group II introns are generally found in non-coding sequences and associated with other mobile genetic elements (Dai and Zimmerly 2002). Organellar group II introns thus primarily function as classic intervening sequences while bacterial group II introns behave like mobile elements. Bacterial group II introns were also shown to propagate by conjugation within and between species, invading the chromosome or resident plasmids of their new hosts using either the retrohoming or retrotransposition pathways (Belhocine et al. 2004; Belhocine et al. 2005; Nisa-Martinez et al. 2007).

Group II introns require the assistance of RNA binding proteins called maturases to adopt their active three-dimensional conformation and self-splice *in vivo* (Fedorova and Zingler 2007). Specific sequence motifs within group IIA introns mediate the accurate recognition of the 5′ and 3′ splice sites. Exon binding sequence 1 (EBS1) and 2 (EBS2) identify the 5′ splice site by base pairing with complementary intron binding sequence 1 (IBS1) and 2 (IBS2) situated at the 3′ extremity of the upstream exon. The 3′ splice site is recognized by the *∂*-*∂*′ base paring interaction at the 5′ extremity of the downstream exon. Group II introns self-splice from interrupted RNA transcripts through three different splicing pathways (Fig. 3.1) (Fedorova and Zingler 2007). The branching (Fig. 3.1A), hydrolysis (Fig. 3.1B) and circularization (Fig. 3.1C) pathways release the intron as either branched structures called lariats, in linear forms or as closed circles, respectively. Each of these three splicing pathways involve two consecutive transesterification reactions (Fig. 3.1, steps 1 and 2). Branching, however, is the only splicing pathway that is completely reversible where intron lariats can recognize single- and double-stranded nucleic acid substrates (RNA/DNA) through base pairing and reinsert themselves by reverse splicing (Fig. 3.1A, double arrows) (Fedorova and Zingler 2007; Pyle 2016). Since reverse splicing is the initial step of both group II intron mobility pathways, retrohoming and retrotransposition, only released intron lariats are active mobile elements (Pyle 2016).

We recently unveiled and characterized at the molecular level the circularization pathway of Ll.LtrB, the model group II intron, from the gram-positive bacterium *Lactococcus lactis* (Monat et al. 2015; Monat and Cousineau 2016). Our work showed that the intron excises simultaneously through the branching and circularization pathways *in vivo* leading to the accumulation of both intron lariats and circles respectively. While the majority of the excised intron circles were found to have their 5′ and 3′ ends perfectly joined, we identified Ll.LtrB RNA circles harboring additional nucleotides at their splice junction. Here we describe novel group II intron splicing pathways in which the release of intron circles, harboring or not mRNA fragments of various lengths at their splice junctions, occurs concurrently with the generation of intergenic E1-mRNA and mRNA-mRNA chimeras *in vivo*. Overall, this study unveils that, similarly to spliceosomal introns in eukaryotes, bacterial group II introns generate genetic diversity at the RNA level, producing novel splicing products with potential new functions.

3.4: Results

3.4.1: Some excised Ll.LtrB RNA circles harbor mRNA fragments of various lengths at their splice junction

To study the splicing pathway leading to the incorporation of additional nucleotides at the splice junction of group II intron circles (Monat et al. 2015) we performed an RT-PCR reaction across the Ll.LtrB-ΔLtrA+LtrA lariat and circle splice junctions (Fig. 3.2) (Monat et al. 2015; Monat and Cousineau 2016). We cloned and sequenced the amplicons located in the faint smear above the RT-PCR band that corresponds to perfect lariat and circle splice junctions (Fig. 3.2C). They revealed excised intron circles harboring additional nucleotides (nts) between the first and the last nts of the intron (Fig. 3.3A). The stretch of additional nts greatly varied in size (20-576 nts), originated from the *L. lactis* chromosome or the two plasmids used to express the intron (Fig. 3.2A) (Monat et al. 2015; Monat and Cousineau 2016) and mapped to the transcribed strand of annotated genes. Some sequences were identified more than once while others corresponded to different portions of the same gene.

Additional nts within the same size range (26-593 nts) and with identical characteristics (Fig. S3.1) were identified at the circle splice junction of Ll.LtrB-WT (Fig. 3.2C). Taken together, these data show that mRNA fragments are incorporated at the splice junction of Ll.LtrB RNA circles during circularization, regardless if LtrA, the intron-encoded protein, is expressed in *trans* (Fig. 3.3A) or in *cis* (Fig. S3.1).

3.4.2: IBS1/2-like sequences are present upstream of both extremities of the mRNA fragments incorporated at the Ll.LtrB circle splice junction

The flanking sequences on both sides of the mRNA fragments incorporated at the Ll.LtrB-ΔLtrA+LtrA (Fig. 3.3A) and Ll.LtrB-WT (Fig. S3.1) circle splice junctions were retrieved, compiled and analyzed. Directly upstream from the 5′ and 3′ junctions we identified IBS1/2-like sequences partly complementary to the EBS1/2 sequences for both introns (Figs. 3.3A, S3.1). Consensus sequences of 30 nts spanning the 5′ and 3′ junctions of the mRNA fragments confirmed

the presence of IBS1/2-like sequence motifs. The IBS1-like motifs are better defined than the IBS2-like motifs, whereas the upstream IBS1/2-like motifs are stronger for both Ll.LtrB-ΔLtrA+LtrA and Ll.LtrB-WT (Fig. 3.4A-C).

3.4.3: Ll.LtrB recognizes both extremities of the mRNA fragments present at intron circle splice junctions through base pairing

Comparable mRNA fragments of various lengths (43-452 nts) (Fig. 3.3B) were also found at the circle splice junction of Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA (Fig. 3.2C), for which the EBS1 sequence was modified from 5′-GUUGUG-3′ to 5′-CAACAC-3′. Accordingly, the IBS1-like consensus sequence motifs upstream from both mRNA junctions were found to be different from Ll.LtrB-ΔLtrA+LtrA and Ll.LtrB-WT and complementary to the mutated EBS1 sequence (Fig. 3.4E). In addition, similarly to Ll.LtrB-ΔLtrA+LtrA and Ll.LtrB-WT, the IBS1-like sequence motifs are both better defined than the IBS2-like motifs and the upstream IBS1/2-like motif much stronger.

The base pairing potential of Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA is more stringent than Ll.LtrB-ΔLtrA+LtrA and Ll.LtrB-WT because its EBS1 sequence (5′-CAACAC-3′) can perfectly recognize only 1 sequence (5′-GUGUUG-3′). In contrast, both introns harboring the wild-type EBS1 sequence (5′-GUUGUG-3′) can base pair perfectly with 64 different sequence combinations using G=U wobble base pairings. Consequently, the more stringent EBS1 sequence led to a fainter RT-PCR smear (Fig. 3.2C), the identification of fewer mRNA fragments at the intron circle splice junction (Fig. 3.3B), and to much stronger flanking consensus motifs when compared to Ll.LtrB-ΔLtrA+LtrA and Ll.LtrB-WT (Fig. 3.4). These data confirm that both junctions of the incorporated mRNA fragments at intron circle splice junctions are recognized by the EBS1/2 motifs of Ll.LtrB through base pairing interactions during circularization.

Consensus sequences are slightly but consistently stronger when flexibility is allowed at both junctions of the mRNA fragments for all three constructs suggesting that Ll.LtrB does not always process mRNAs precisely downstream from the recognized IBS1/2-like motifs (Figs. S3.2-S3.5). We also identified mRNA fragments, at intron circle splice junctions, that either contained untranslated sequences or spanned two genes including the short intergenic regions of

polycistronic mRNAs (Figs. 3.3, S3.1). This further supports our conclusion that Ll.LtrB can capture *L. lactis* transcripts at intron circle splice junctions during circularization.

3.4.4: Models of mRNA fragment incorporation at group II intron circle splice junctions

Our findings indicate that cellular mRNAs can somehow be incorporated at the Ll.LtrB circle splice junction during the circularization pathway. Two models can explain how mRNA fragments could be incorporated at the splice junction of group II intron circles (Fig. 3.5).

The external nucleophilic attack pathway (Fig. 3.5A) was previously proposed to explain how short stretches of additional nts could be incorporated at the circle splice junction during intron circularization. However, the pathway of integration and the origin of the additional nts were never demonstrated (Li-Pook-Than and Bonen 2006; Monat et al. 2015). Taking into consideration the data presented here, Ll.LtrB would recognize, through base pairing interactions, an IBS1/2-like sequence on an *L. lactis* mRNA and guide its hydrolysis downstream of the recognized sequence (step 1). Next, the 3′-OH of the processed mRNA would induce a transesterification reaction at the exon 1-intron splice junction resulting in its ligation to the 5′ end of the intron and the release of exon 1 (step 2). The 3′-OH of exon 1 would then initiate the next transesterification reaction at the intron-exon 2 splice junction, releasing ligated exons and a linear intron harboring an mRNA fragment at its 5′ end (step 3). The final transesterification reaction would be induced at the intron 5′ end (step 4a) or within the mRNA (step 4b) by the 2′-OH of the last nt of the linear intron, just downstream from IBS1/2-like sequences, resulting in the release of either a head-to-tail circular intron or an intron circle harboring an mRNA fragment at its splice junction respectively.

An alternative pathway (Fig. 3.5B) would rather be initiated by the reverse splicing of an intron lariat within an *L. lactis* mRNA downstream of an IBS1/2-like sequence (step 1). The ectopically inserted group II intron would then excise from the mRNA through circularization (steps 2-4). The 3′-OH of free exon 1 would first attack the phosphodiester bond at the 3′ splice site between the last nt of the intron and the 3′ segment of the mRNA (step 2). This would generate a chimeric mRNA consisting of the *ltrB*-exon 1 (E1) linked to the 3′ segment of the mRNA (E1-

mRNA) and a circularization intermediate where the linear intron is still attached to the 5′ segment of the mRNA. The final transesterification reaction would then be induced at the intron 5′ end (step 3a) or within the mRNA fragment (step 3b) by the 2′-OH of the last nt of the intron, just downstream from IBS1/2-like sequences, resulting in the release of either a head-to-tail circular intron or an intron circle harboring an mRNA fragment at its splice junction respectively.

3.4.5: Ll.LtrB lariats reverse splice within *L. lactis* mRNAs downstream of IBS1/2-like sequences

To investigate the proposed models we looked for unique intermediates of the reverse splicing pathway: the 3′ junction of Ll.LtrB reverse-spliced within mRNAs and chimeric E1-mRNAs (Fig. 3.5B, asterisks). We first detected by RNA-Seq intron-interrupted mRNAs for Ll.LtrB-ΔLtrA+LtrA and Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA but not for the Ll.LtrB-ΔA-ΔLtrA+LtrA control which lacks the essential branch point A residue required for branching and reverse splicing (van der Veen et al. 1987; Monat et al. 2015; Monat and Cousineau 2016) (Fig. 3.6A). The reverse splice sites of Ll.LtrB-ΔLtrA+LtrA (Fig. 3.6B) and Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA (Fig. 3.6C) were shown to be immediately preceded by consensus IBS1/2-like sequence motifs complementary to their respective EBS1/2 sequences. On the other hand, similarly to the junctions between intron circles and mRNA fragments (Figs. 3.4, S3.5), we did not detect a ∂′-like sequence on the 3′ side of the intron insertion sites (Fig. 3.6). This shows that Ll.LtrB can recognize IBS1/2-like sequences on various mRNAs by base pairing with its EBS1/2 sequences and invade them by reverse splicing, generating a population of intron-interrupted mRNAs in *L. lactis*. As expected, the more stringent EBS1 sequence of Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA led to the identification of fewer intron-interrupted mRNAs and a stronger IBS1/2-like consensus sequence upstream of the intron insertion sites compared to Ll.LtrB-ΔLtrA+LtrA.

We next studied in further details the reverse splicing of Ll.LtrB-ΔLtrA+LtrA within the Enolase (*enoA*) and Alanyl-tRNA synthetase (*alaS*) mRNAs. The *enoA* (167 nts) and *alaS* (304 nts) mRNA fragments, previously identified at the Ll.LtrB-ΔLtrA+LtrA circle splice junction, are both flanked by a strong (10/11 nts) and a weak (7/11 and 8/11 nts) IBS1/2-like sequence motif (Fig. 3.3A). We amplified by RT-PCR the 5′ (Fig. 3.7A, F) and 3′ (Fig. 3.7B, E) junctions between

the intron and the two mRNAs. Sequences of the four amplicons confirmed reverse splicing of the intron precisely downstream of the strong IBS1/2-like sequence within the *enoA* (Fig. 3.7C, large open arrowhead) and *alaS* (Fig. 3.7G, large open arrowhead) mRNAs. Importantly, no amplifications were detected for the reverse splicing deficient control, Ll.LtrB-ΔA-ΔLtrA+LtrA. Next, the faint smears above (Fig. 3.7A, E) and below (Fig. 3.7B, F) the main amplicons were cloned and shown to correspond to several independent 5′ and 3′ junctions of the intron inserted downstream of different weak IBS1/2-like sequences (7-9/11 nts) (Fig. 3.7C, G, black and grey arrowheads). The weak IBS1/2-like sequences flanking the mRNA fragments previously identified within intron circles (Fig. 3.3A), were also found invaded by the intron for both *enoA* (Fig. 3.7C, small open arrowhead) and *alaS* (Fig. 3.7G, small open arrowhead). Similarly, the Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA variant was shown to reverse splice at specific strong and weak IBS1/2-like sequences within the *S12/S7* transcript (Fig. 3.8A-C). The identified reverse splice sites also include the strong and weak IBS1/2-like sequences flanking the *S12/S7* mRNA fragment (161 nts) previously identified at the intron circle splice junction (Fig. 3.3B).

Collectively, these results show that IBS1/2-like sequences are widespread within *L. lactis* mRNAs, providing abundant targets for Ll.LtrB reverse splicing. They also support the proposed alternative circularization model by which introns, reverse-spliced at ectopic sites within mRNAs, can circularize alternatively by recognizing upstream IBS1/2-like sequences leading to the capture of mRNA fragments at their splice junction (Fig. 3.5B, step 3b). Accordingly, when additional nts are found at intron circle splice junctions, the upstream IBS1/2-like consensus sequences are consistently stronger (Figs. 3.4, S3.5) suggesting that when the intron reverse splices at a weak IBS1/2-like sequence, it is more likely to release intron circles harboring mRNA fragments by recognizing a stronger upstream alternative IBS1/2-like sequence.

3.4.6: Ll.LtrB circularization from intron-interrupted mRNAs generates E1-mRNA and mRNA-mRNA chimeras

The second distinguishing splicing intermediate between the two proposed models is a chimeric mRNA consisting of *ltrB*-exon 1 (E1) *trans*-spliced to an *L. lactis* mRNA fragment (E1-mRNA) (Fig. 3.5B, asterisk). We specifically screened for E1-*enoA* and E1-*alaS* mRNA chimeras

by RT-PCR. In both cases we detected, exclusively for the reverse splicing-competent intron, E1-mRNA chimeras ligated precisely downstream from the strong IBS1/2-like sequences (Fig. 3.7D, H), the exact sites previously identified at one of the extremities of the mRNA fragments identified at intron circle splice junctions (Fig. 3.3A) and invaded by reverse splicing (Fig. 3.7A, B, E, F). The intron-catalyzed EBS1/2-specific generation of E1-mRNA chimeras was corroborated with the Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA variant again at the previously identified strong IBS1/2-like sequence of the *S12/S7* transcript (Fig. 3.8D). These results show that Ll.LtrB, reverse-spliced at IBS1/2-like sequences of various mRNAs, can recruit free E1 through EBS-IBS base pairing interactions, and catalyze the formation of E1-mRNA chimeras.

Ll.LtrB splicing *via* circularization, from a population of intron-interrupted mRNAs, generates processed mRNA fragments harboring IBS1/2-like sequences at their 3′ end (Fig. 3.5B, step 3a and 3b). We next examined if these splicing products could be recruited by Ll.LtrB, similarly to free E1 through EBS-IBS base pairing, and used to generate intergenic mRNA-mRNA chimeras (Fig. 3.9). We detected by RT-PCR both *alaS-enoA* (Fig. 3.7I) and *enoA-alaS* (Fig. 3.7J) mRNA-mRNA intergenic chimeras joined at specific IBS1/2-like sequences for Ll.LtrB-ΔLtrA+LtrA but not for the Ll.LtrB-ΔA-ΔLtrA+LtrA control. These data show that Ll.LtrB, reverse-spliced within various mRNAs, can recruit through base pairing processed mRNA fragments, harboring IBS1/2-like sequences at their 3′ end, to initiate the circularization splicing pathway (Fig. 3.9, step 2). Ll.LtrB can thus catalyze the shuffling of coding sequences within a population of intron-interrupted mRNAs by a new intergenic *trans*-splicing pathway (Fig. 3.9).

3.5: Discussion

One quarter of currently sequenced bacterial genomes harbor one to a few group II introns (Candales et al. 2012). This paucity, coupled with their irregular distribution and frequent lateral transfer (LaRoche-Johnston et al. 2016), has led to the suggestion that they are selfish retromobile elements with no beneficial function to their host (Dai and Zimmerly 2002). In contrast, many group II intron derivatives provide important functions in both eukaryotes and prokaryotes (Lambowitz and Belfort 2015; Zimmerly and Semper 2015; Novikova and Belfort 2017). For example, the abundant spliceosomal introns, descendants of group II introns, generate significant

genetic diversity and transcriptomic complexity *via* alternative splicing (Bush et al. 2017), intergenic *trans*-splicing (Lei et al. 2016), RNA circle formation (Chen 2016) and by creating new genes through exon shuffling (Franca et al. 2012).

Even though the Ll.LtrB group II intron is present at only one copy in the *L. lactis* genome, the new splicing pathways described here (Figs. 3.5B, 3.9) expand the genetic diversity and complexity of its host transcriptome. This stems from the ability of Ll.LtrB, following its release as RNP particles, to generate a population of intron-interrupted mRNAs through reverse splicing, which we were able to detect by RNA-Seq (Fig. 3.6) and gene-specific RT-PCR (Figs. 3.7, 3.8). Ll.LtrB was recently shown to interact with its cognate ligated exons at the IBS1/2 site *in vivo*, leading to either complete reverse splicing or negative regulation of targeted mRNA through hydrolysis and degradation (Qu et al. 2018). However, when we contrasted the counts per million (CPM) of Ll.LtrB-WT, Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA and Ll.LtrB-ΔA-ΔLtrA+LtrA constructs for *alaS*, the most abundant target for reverse splicing that we identified by RNA-Seq, we obtained differential expression ratios that showed very little change in the abundance of the *alaS* transcript: 0.97 between Ll.LtrB-WT and EBS1/Mut-ΔLtrA+LtrA and 0.96 between Ll.LtrB-WT and Ll.LtrB-ΔA-ΔLtrA+LtrA. This suggests that the IBS1/2-like sites we identified within host mRNAs are not efficient targets for hydrolysis, but rather seem to be used for reverse splicing. Interestingly, several of the reverse-splicing sites found by RNA-Seq were also identified independently at the extremity of mRNA fragments captured at intron circle splice junctions (Figs. 3.3A, S3.1), yet there was only a small overlap of IBS1/2-like motifs between these two sets of data. Moreover, when we analysed the *enoA* and *alaS* genes in greater detail, we found a multitude of additional IBS1/2-like motifs that were used as targets for Ll.LtrB reverse splicing and whose base paring interactions with the intron varied from strong (11/11 nts) to weak (7/11 nts) (Fig. 3.7C, G). Overall, our data thus suggest that the reverse-splicing of group II introns into ectopic sites within host mRNAs is a widespread, dynamic and transient process whose exact scope is hard to determine.

We demonstrated that circularization of Ll.LtrB from interrupted mRNAs, using free E1 or mRNA fragments harboring IBS1/2-like sequences at their 3′ end, generates two types of *trans*-spliced transcripts: E1-mRNA (Fig. 3.5B) and mRNA-mRNA (Fig. 3.9) chimeras respectively. Ll.LtrB was recently found to generate free E1 *in vivo* through hydrolysis of ligated cognate exons

at the IBS1/2 site (Qu et al. 2018). This Spliced Exon Reopening (SER) reaction (Fig. 3.1) could thus produce the initial source of E1 required to initiate Ll.LtrB circularization from both its cognate exons and ectopic insertion sites. In addition, we found that alternative circularization of Ll.LtrB from interrupted mRNAs releases intron circles harboring mRNA fragments at their splice junction (Fig. 3.5B, step 3b). These novel bacterial splicing products, generated by alternative circularization and intergenic *trans*-splicing, may have and/or lead to novel biological functions for their host cell. For instance, chimeric RNAs, intron circles and different circular RNAs that accumulate *in vivo* have been recently associated to a variety of interesting new functions such as RNA sponges, protein sponges and transcriptional regulators in various biological systems (Chen 2016; Cortes-Lopez and Miura 2016; Hsiao et al. 2017). Moreover, the *trans*-spliced E1-mRNA and mRNA-mRNA chimeras could be reclaimed by the host and potentially lead to the creation of new genes. Group II introns may thus serve a beneficial function for their hosts by increasing the complexity and genetic diversity of their transcriptomes (Fig. 3.10) which could explain why they were retained in bacteria.

Our work also unveils two additional functional and evolutionary links between group II introns, spliceosomal introns and the spliceosome. First, the *trans*-splicing of E1 at the 5′ end of various mRNA fragments is analogous to the second step of the spliced leader (SL) *trans*-splicing pathway, which has a patchy evolutionary distribution amongst eukaryotes and whose origin has remained enigmatic (Hastings 2005; Krchnakova et al. 2017). Second, we showed that group II introns, similarly to the spliceosome (Lei et al. 2016), can catalyze the *trans*-splicing of intergenic mRNA-mRNA chimeras in bacteria. Since group II introns are considered as the progenitors of both spliceosomal introns and the snRNAs of the spliceosome (Lambowitz and Belfort 2015; Zimmerly and Semper 2015; Novikova and Belfort 2017), our findings suggest that the spliceosome-dependent formation of SL *trans*-spliced transcripts and intergenic mRNA-mRNA chimeras in eukaryotes both consist of ancient group II intron splicing functions still shared with their contemporary bacterial relatives.

Overall, we described here new group II intron splicing pathways that generate and expand the genetic diversity and complexity of its host transcriptome which represents a new function for these bacterial retroelements. Our work also unveils new functional and evolutionary links with

their nuclear relatives in eukaryotes, and provides a potential explanation of why group II introns were maintained in bacteria.

3.6: Experimental procedures

3.6.1: Bacterial strains and plasmids

Lactococcus lactis strain NZ9800Δ*ltrB* (Tet^R) (Ichiyanagi et al. 2002) was grown in M17 media supplemented with 0.5 % glucose (GM17) at 30°C without shaking. The *Escherichia coli* strain DH10β, used for cloning purposes, was grown in LB at 37°C with shaking. Antibiotics were used at the following concentrations: chloramphenicol (Cam^R), 10 µg/ml; spectinomycin (Spc^R), 300 µg/ml. Previously constructed plasmids (pDL-$P_{23}{}^2$-Ll.LtrB-ΔLtrA (Matsuura et al. 1997), pDL-$P_{23}{}^2$-Ll.LtrB-WT (Matsuura et al. 1997), pLE-$P_{23}{}^2$-LtrA (Belhocine et al. 2007)) were used to study Ll.LtrB splicing. Additional variants were constructed by site-directed mutagenesis (New England Biolabs® Q5® Site-Directed-Mutagenesis Kit): pDL-$P_{23}{}^2$-Ll.LtrB-ΔA-ΔLtrA, pDL-$P_{23}{}^2$-Ll.LtrB-EBS1/Mut-ΔLtrA. The alanyl tRNA synthetase (*alaS*) and enolase (*enoA*) genes were cloned in pLE-$P_{23}{}^2$-LtrA (BssHII), downstream of the second P_{23} promoter, and expressed with the intron in a two-plasmid system. Primers used for mutagenesis and cloning are in Table S3.2.

3.6.2: RNA extraction, PCR and RT-PCR

Total RNA was isolated from NZ9800Δ*ltrB* harboring various plasmid constructs as previously described (Belhocine et al. 2007). RT-PCR reactions (Monat et al. 2015; Monat and Cousineau 2016) were performed on total RNA preparations of NZ9800Δ*ltrB* harboring various intron constructs (primers in Table S3.2).

3.6.3: RNA-seq

RNA-seq was performed on rRNA-depleted total RNA from *L. lactis* (NZ9800Δ*ltrB*) expressing Ll.LtrB-ΔLtrA+LtrA, Ll.LtrB-ΔA-ΔLtrA+LtrA, or Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA using the Illumina HiSeq 2500 paired-end sequencing system (Bentley et al. 2008).

3.6.4: Sequence consensus

Aligned and adjusted consensuses were prepared using the WebLogo software (Crooks et al. 2004). Adjusted consensuses were determined using a code that calculated contiguous nucleotides with the highest capacity of base pairing to EBS1 and EBS2 (total of 11 nucleotides), separated from each other by 0-2 nucleotides, in a region that spanned -14, +4 nts around the junction with the intron.

3.7: Acknowledgments

We thank G. Burger for comments on the manuscript. This work was supported by a discovery grant from the Natural Sciences and Engineering Research Council of Canada to BC (227826). FLJ previously received a Graduate Excellence Fellowship from McGill University, a Canada Graduate Scholarship-Masters from the Natural Sciences and Engineering Research Council of Canada and a Master's Research Scholarship from Fonds de Recherche en Nature et Technologies du Québec; and currently holds an Alexander Graham Bell Canada Graduate Scholarship-Doctoral from the Natural Sciences and Engineering Research Council of Canada.

3.8: References

Belhocine K, Plante I, Cousineau B. 2004. Conjugation mediates transfer of the Ll.LtrB group II intron between different bacterial species. *Mol Microbiol* 51(5):1459-69.

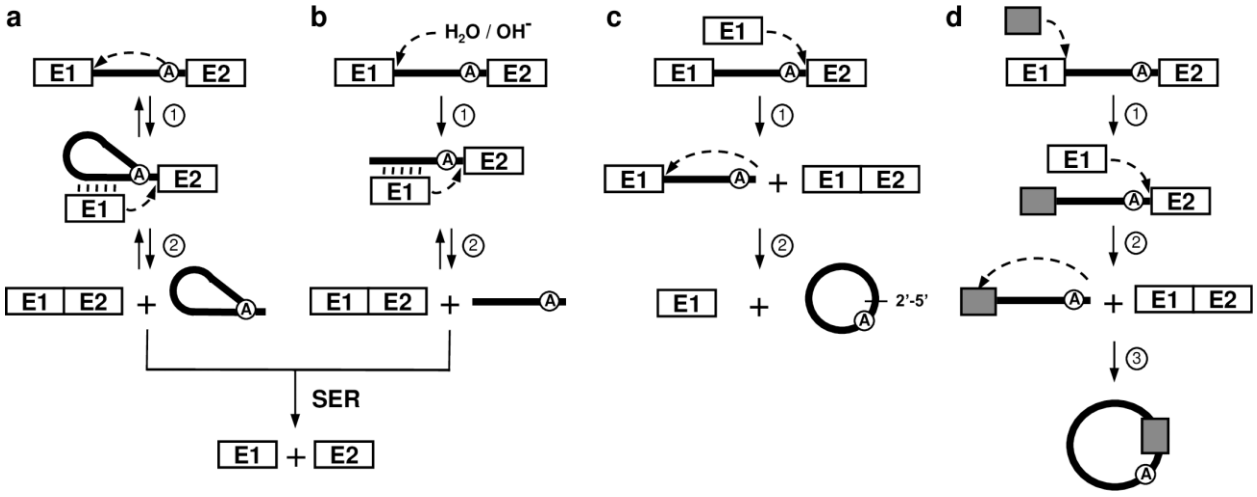
Belhocine K, Yam KK, Cousineau B. 2005. Conjugative transfer of the Lactococcus lactis chromosomal sex factor promotes dissemination of the Ll.LtrB group II intron. *J Bacteriol* 187(3):930-9.

Belhocine K, Mak AB, Cousineau B. 2007. Trans-splicing of the Ll.LtrB group II intron in Lactococcus lactis. *Nucleic Acids Res* 35(7):2257-68.
Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR et al. . 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456(7218):53-9.
Bush SJ, Chen L, Tovar-Corona JM, Urrutia AO. 2017. Alternative splicing and the evolution of phenotypic novelty. *Philos Trans R Soc Lond B Biol Sci* 372(1713).
Candales MA, Duong A, Hood KS, Li T, Neufeld RA, Sun R, McNeil BA, Wu L, Jarding AM, Zimmerly S. 2012. Database for bacterial group II introns. *Nucleic Acids Res* 40(Database issue):D187-90.
Chalamcharla VR, Curcio MJ, Belfort M. 2010. Nuclear expression of a group II intron is consistent with spliceosomal intron ancestry. *Genes Dev* 24(8):827-36.
Chen LL. 2016. The biogenesis and emerging roles of circular RNAs. *Nat Rev Mol Cell Biol* 17(4):205-11.
Cortes-Lopez M, Miura P. 2016. Emerging Functions of Circular RNAs. *Yale J Biol Med* 89(4):527-537.
Cousineau B, Smith D, Lawrence-Cavanagh S, Mueller JE, Yang J, Mills D, Manias D, Dunny G, Lambowitz AM, Belfort M. 1998. Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell* 94(4):451-62.
Cousineau B, Lawrence S, Smith D, Belfort M. 2000. Retrotransposition of a bacterial group II intron. *Nature* 404(6781):1018-21.
Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* 14(6):1188-90.
Dai L, Zimmerly S. 2002. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res* 30(5):1091-102.
Fedorova O, Zingler N. 2007. Group II introns: structure, folding and splicing mechanism. *Biol Chem* 388(7):665-78.
Franca GS, Cancherini DV, de Souza SJ. 2012. Evolutionary history of exon shuffling. *Genetica* 140(4-6):249-57.
Hastings KE. 2005. SL trans-splicing: easy come or easy go? *Trends Genet* 21(4):240-7.
Hsiao KY, Sun HS, Tsai SJ. 2017. Circular RNA - New member of noncoding RNA with novel functions. *Exp Biol Med (Maywood)* 242(11):1136-1141.
Ichiyanagi K, Beauregard A, Lawrence S, Smith D, Cousineau B, Belfort M. 2002. Retrotransposition of the Ll.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. *Mol Microbiol* 46(5):1259-72.
Irimia M, Roy SW. 2014. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb Perspect Biol* 6(6).
Krchnakova Z, Krajcovic J, Vesteg M. 2017. On the Possibility of an Early Evolutionary Origin for the Spliced Leader Trans-Splicing. *J Mol Evol* 85(1-2):37-45.
Lambowitz AM, Zimmerly S. 2011. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol* 3(8):a003616.
Lambowitz AM, Belfort M. 2015. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol Spectr* 3(1).
LaRoche-Johnston F, Monat C, Cousineau B. 2016. Recent horizontal transfer, functional adaptation and dissemination of a bacterial group II intron. *BMC Evol Biol* 16(1):223.

Leclercq S, Cordaux R. 2012. Selection-driven extinction dynamics for group II introns in Enterobacteriales. *PLoS One* 7(12):e52268.
Lei Q, Li C, Zuo Z, Huang C, Cheng H, Zhou R. 2016. Evolutionary Insights into RNA trans-Splicing in Vertebrates. *Genome Biol Evol* 8(3):562-77.
Li-Pook-Than J, Bonen L. 2006. Multiple physical forms of excised group II intron RNAs in wheat mitochondria. *Nucleic Acids Res* 34(9):2782-90.
Matsuura M, Saldanha R, Ma H, Wank H, Yang J, Mohr G, Cavanagh S, Dunny GM, Belfort M, Lambowitz AM. 1997. A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes Dev* 11(21):2910-24.
Monat C, Quiroga C, Laroche-Johnston F, Cousineau B. 2015. The Ll.LtrB intron from Lactococcus lactis excises as circles in vivo: insights into the group II intron circularization pathway. *RNA* 21(7):1286-93.
Monat C, Cousineau B. 2016. Circularization pathway of a bacterial group II intron. *Nucleic Acids Res* 44(4):1845-53.
Nisa-Martinez R, Jimenez-Zurdo JI, Martinez-Abarca F, Munoz-Adelantado E, Toro N. 2007. Dispersion of the RmInt1 group II intron in the Sinorhizobium meliloti genome upon acquisition by conjugative transfer. *Nucleic Acids Res* 35(1):214-22.
Novikova O, Belfort M. 2017. Mobile Group II Introns as Ancestral Eukaryotic Elements. *Trends Genet* 33(11):773-783.
Pyle AM. 2016. Group II Intron Self-Splicing. *Annu Rev Biophys* 45:183-205.
Qu G, Piazza CL, Smith D, Belfort M. 2018. Group II intron inhibits conjugative relaxase expression in bacteria by mRNA targeting. *Elife* 7.
van der Veen R, Kwakman JH, Grivell LA. 1987. Mutations at the lariat acceptor site allow self-splicing of a group II intron without lariat formation. *EMBO J* 6(12):3827-31.
Zimmerly S, Semper C. 2015. Evolution of group II introns. *Mob DNA* 6:7.

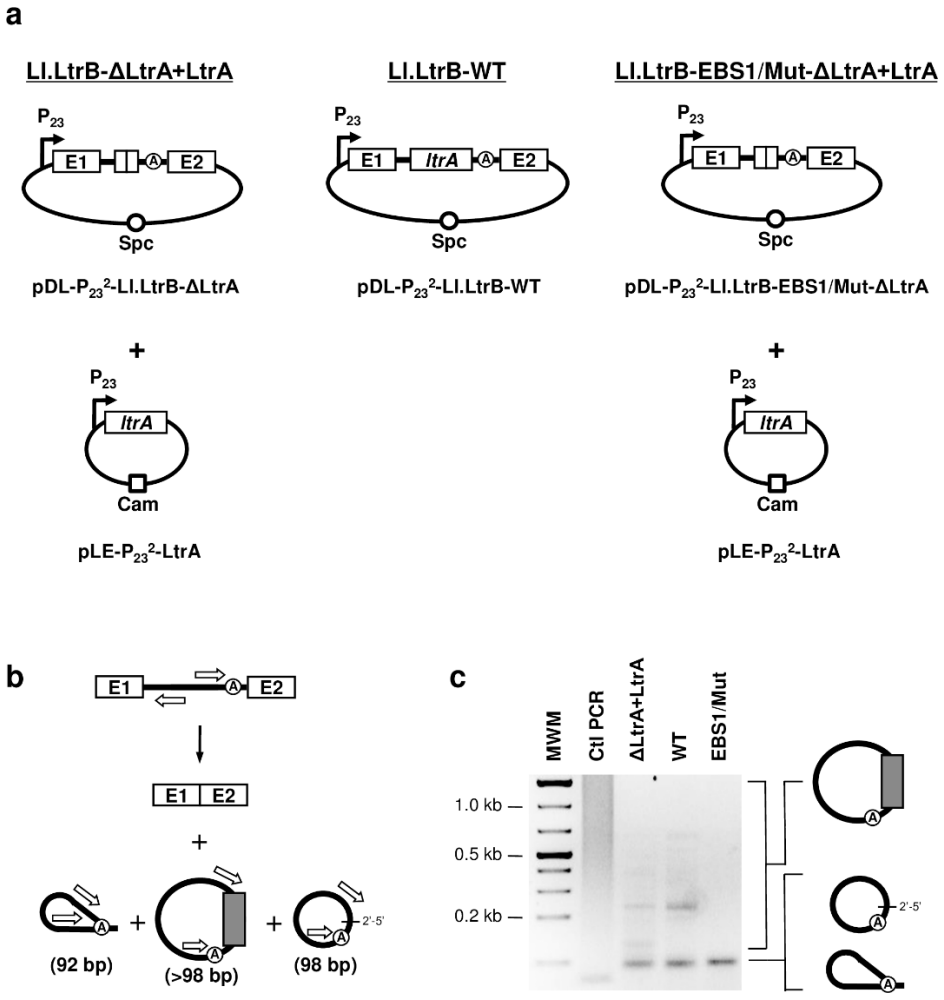
3.9: Figures

Figure 3.1



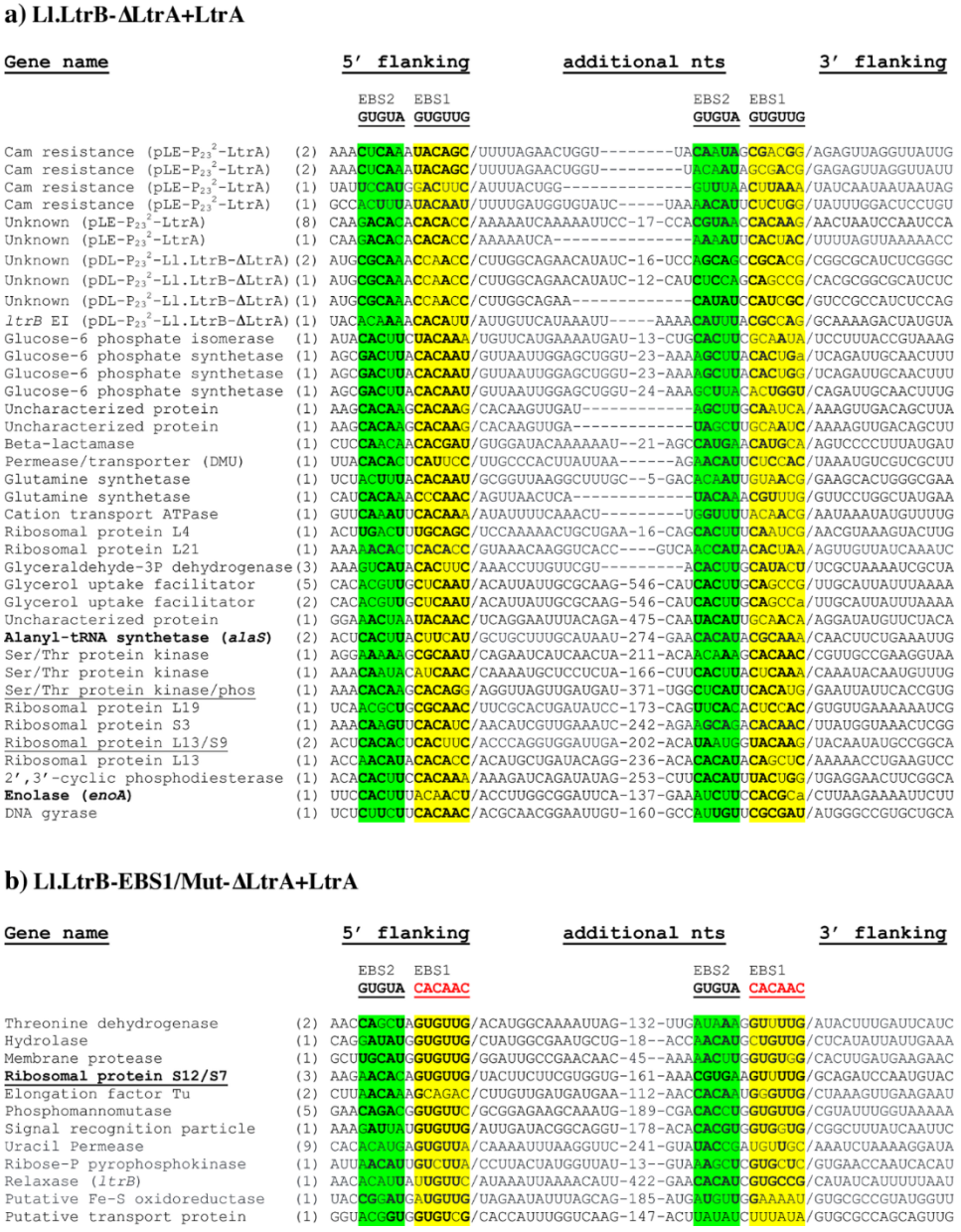
Group II intron splicing pathways. (**a**) Branching pathway. Following transcription of the interrupted gene, the 2′-OH residue of the branch-point nucleotide (A) initiates the first nucleophilic attack at the exon 1-intron junction (step 1). This transesterification reaction connects the 5′ end of the intron to the branch point and releases exon 1 that remains associated to the intron through base pairing interactions (EBS-IBS interactions) (vertical lines). The liberated 3′-OH at the end of exon 1 then initiates a second nucleophilic attack at the intron-exon 2 junction (step 2), ligating the two exons and releasing the intron as a lariat. (**b**) Hydrolytic pathway. A hydroxyl ion or a water molecule initiates the first nucleophilic attack at the exon 1-intron junction (step 1). The second nucleophilic attack at the intron-exon 2 junction is initiated by the liberated 3′-OH at the end of exon 1 (step 2) which ligates the two exons and releases a linear intron. (**c**) Circularization pathway. The first nucleophilic attack takes place at the intron-exon 2 junction and is initiated by the 3′-OH of a free exon 1 (step 1) generating ligated exons and a circularization intermediate where the linear intron is still attached to exon 1. Next, the 2′-OH of the last intron residue is thought to initiate the second nucleophilic reaction at the exon 1-intron junction (step 2) resulting in intron circularization and the release of free exon 1. A potential source of free exon 1 is the spliced exon reopening (SER) reaction where both excised lariats and linear introns can recognize and hydrolyze ligated exons at the splice junction. To explain the presence of additional nts at the splice junction of intron circles, the external nucleophilic attack pathway (**d**) was previously proposed (Li-Pook-Than and Bonen 2006; Monat et al. 2015). The 3′OH residue of a block of external nts (grey box) attacks the exon 1-intron junction, ligating it to the intron 5′ end while concurrently displacing exon 1 (step 1). The 3′OH at the end of exon 1 then attacks the intron-exon 2 junction releasing ligated exons and a linear intron harboring external nts at its 5′ end (step 2). The third transesterification reaction is initiated by the 2′-OH of the last intron residue (step 3). The position of this final nucleophilic attack thus dictates how many additional nts are incorporated at the junction of intron circles.

Figure 3.2



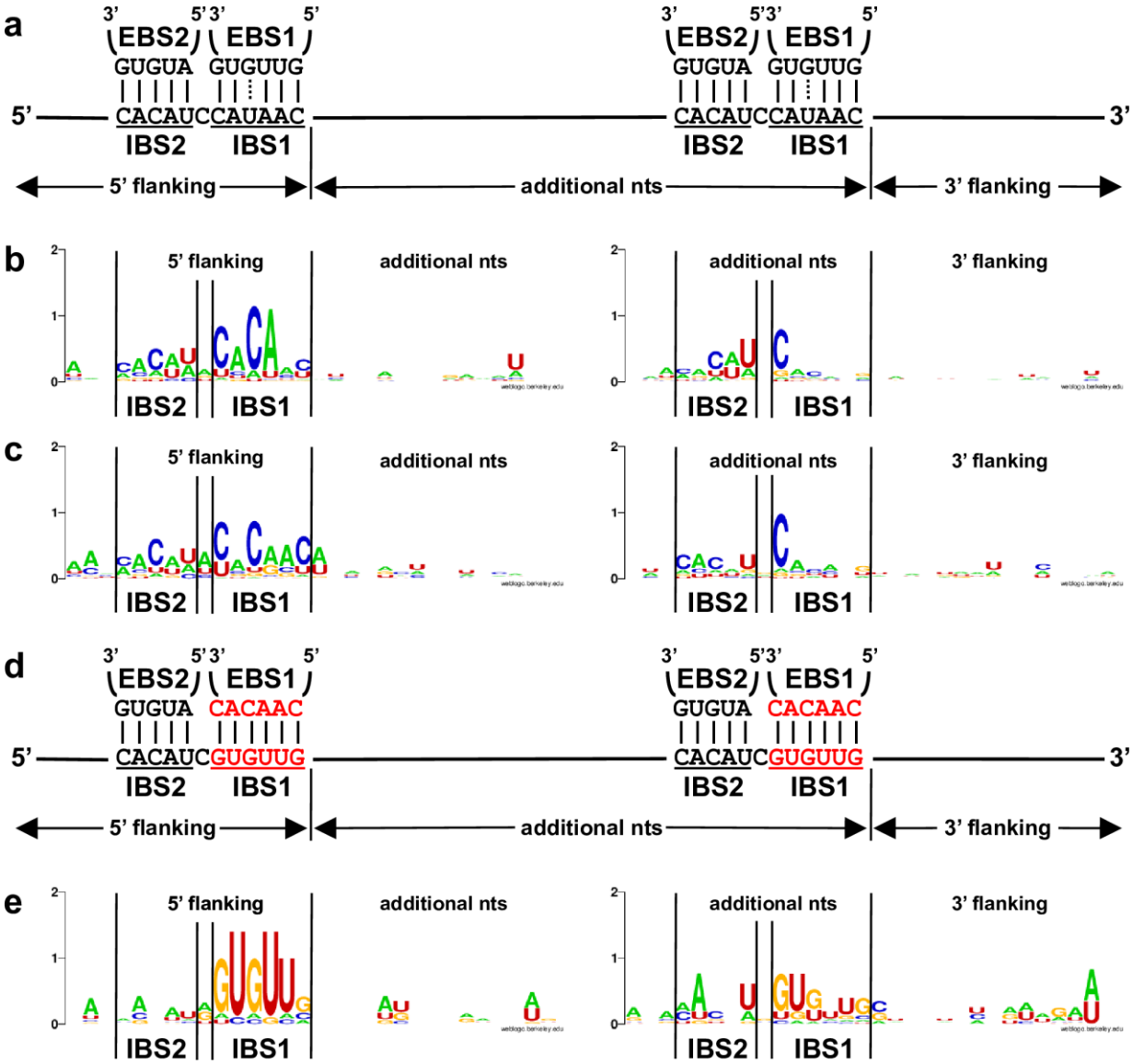
Detection of mRNA fragments at the splice junction of excised intron RNA circles. (**a**) Various Ll.LtrB constructs used in this study where the LtrA protein is provided either in *trans* (Ll.LtrB-ΔLtrA+LtrA, Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA) or in *cis* (Ll.LtrB-WT) (**b**) Schematic of Ll.LtrB self-splicing. Position of the primers (open arrows) (Table S3.2) used to amplify the splice junction of excised introns by RT-PCR is depicted (92 bp (lariat), 98 bp (circle) or >98 bp (circle harboring additional nts)). (**c**) RT-PCR amplifications of intron splice junctions. Amplifications were performed on total RNA extracts from *L. lactis* (NZ9800Δ*ltrB*) harboring different Ll.LtrB constructs expressed under the control of the P_{23} constitutive promoter (ΔLtrA+LtrA: pDL-P_{23}^2-Ll.LtrB-ΔLtrA and pLE-P_{23}^2-LtrA) (WT: pDL-P_{23}^2-Ll.LtrB-WT) (EBS1/Mut: pDL-P_{23}^2-Ll.LtrB-EBS1/Mut-ΔLtrA and pLE-P_{23}^2-LtrA). The EBS1/Mut intron variant was shown to splice accurately and efficiently *in vivo* by RT-PCR amplifications of both released introns and ligated exons. Additional nts incorporated at the junction of intron circles are represented by a gray box.

Figure 3.3



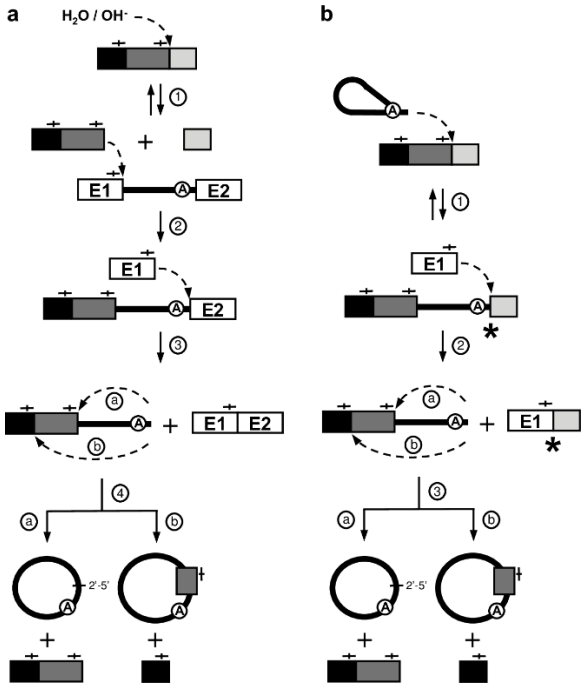
mRNA fragments identified at the splice junction of Ll.LtrB-ΔLtrA+LtrA (a) and Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA (b) circles. Additional nts are shown along with their flanking sequences (5′ flanking) (3′ flanking), their origin (Gene name) and frequency of identification between parentheses. The junctions between the additional nts and their flanking regions (/) as well as the IBS1- (yellow) and IBS2- (green) like sequences are denoted. The bolded nts represent residues from the IBS1- and IBS2-like sequences that can potentially base pair with the intron's EBS1 and EBS2 sequences specified above. Sequences spanning two genes and including a short intergenic region are underlined. The genes in bold (*alaS*, *enoA*, *S12/S7*) were further studied for Ll.LtrB reverse splicing analyses and the detection of E1-mRNA and mRNA-mRNA chimeras (Figs. 3.7, 3.8).

Figure 3.4



Logo representation of the consensus sequences (30 nts) around the 5′ and 3′ extremities of the mRNA fragments identified at intron circle splice junctions. The EBS1-IBS1 and EBS2-IBS2 base pairing interactions for Ll.LtrB-ΔLtrA+LtrA and Ll.LtrB-WT (**a**) as well as Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA (**d**) are depicted. The consensus sequences are shown for Ll.LtrB-ΔLtrA+LtrA (Fig. 3.3A) (**b**), Ll.LtrB-WT (Fig. S3.1) (**c**) and Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA (Fig. 3.3B) (**e**).

Figure 3.5

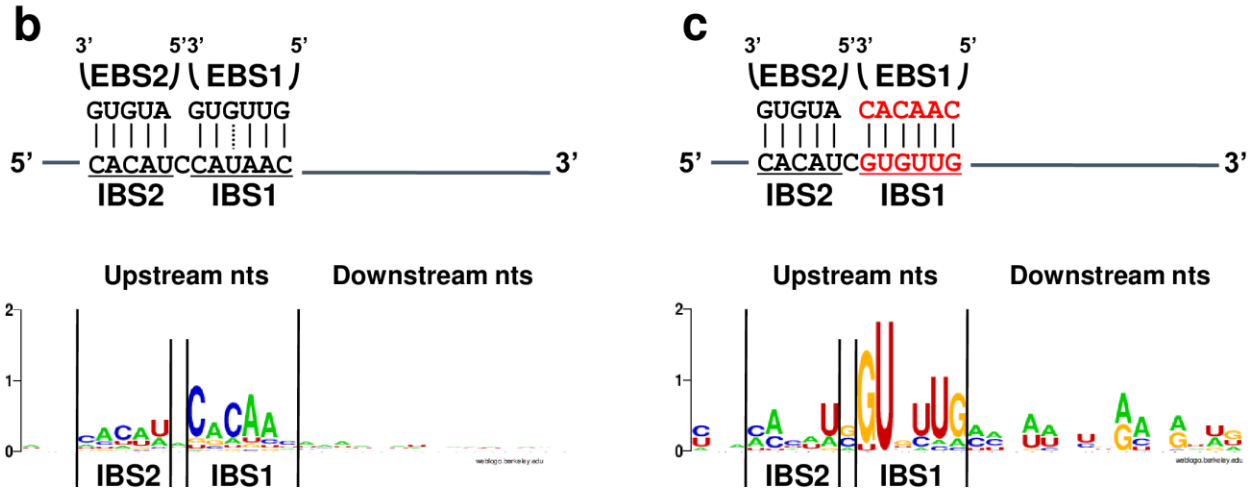


Models for the incorporation of mRNA fragments at the splice junction of intron RNA circles. (**a**) External nucleophilic attack pathway (Li-Pook-Than and Bonen 2006; Monat et al. 2015). The Ll.LtrB group II intron recognizes, through base pairing interactions, an IBS1/2-like sequence (—|—) on an mRNA and guides the first nucleophilic attack induced by an hydroxyl ion or a water molecule downstream of the recognized sequence (step 1). Next, the 3′-OH of the processed mRNA induces a nucleophilic attack at the exon 1-intron splice junction resulting in its ligation to the 5′ end of the intron and the release of exon 1 (step 2). The 3′-OH of exon 1 is then free to initiate the second transesterification reaction at the intron-exon 2 splice junction, releasing ligated exons and a linear intron harboring a fragment of mRNA at its 5′ end (step 3). The final transesterification reaction is induced at the intron 5′ end (a) or within the mRNA (b) by the 2′-OH of the last nt of the linear intron, just downstream from IBS1/2-like sequences (—|—), resulting in the release of either a head-to-tail circular intron (step 4a) or an intron circle harboring an mRNA fragment at its splice junction (step 4b). (**b**) Reverse splicing pathway. This pathway is initiated by the reverse splicing of an intron lariat within a non-cognate mRNA downstream of an IBS1/2-like sequence (—|—) (step 1). The 3′-OH of free exon 1 then attacks the phosphodiester bond at the 3′ splice site between the last nt of the intron and the 3′ segment of the mRNA (step 2). This generates a chimeric mRNA consisting of the *ltrB*-exon 1 (E1) linked to the 3′ segment of the mRNA (E1-mRNA) and a circularization intermediate where the linear intron is still attached to the 5′ segment of the mRNA. The third transesterification reaction is induced at the intron 5′ end (a) or within the mRNA fragment (b) by the 2′-OH of the last residue of the linear intron, just downstream from IBS1/2-like sequences (—|—), resulting in the release of either a head-to-tail circular intron (step 3a) or an intron circle harboring an mRNA fragment at its splice junction (step 3b). The 3′ junction of reverse-spliced introns and the chimeric E1-mRNAs are unique splicing intermediates that distinguish both pathways (asterisks).

Figure 3.6

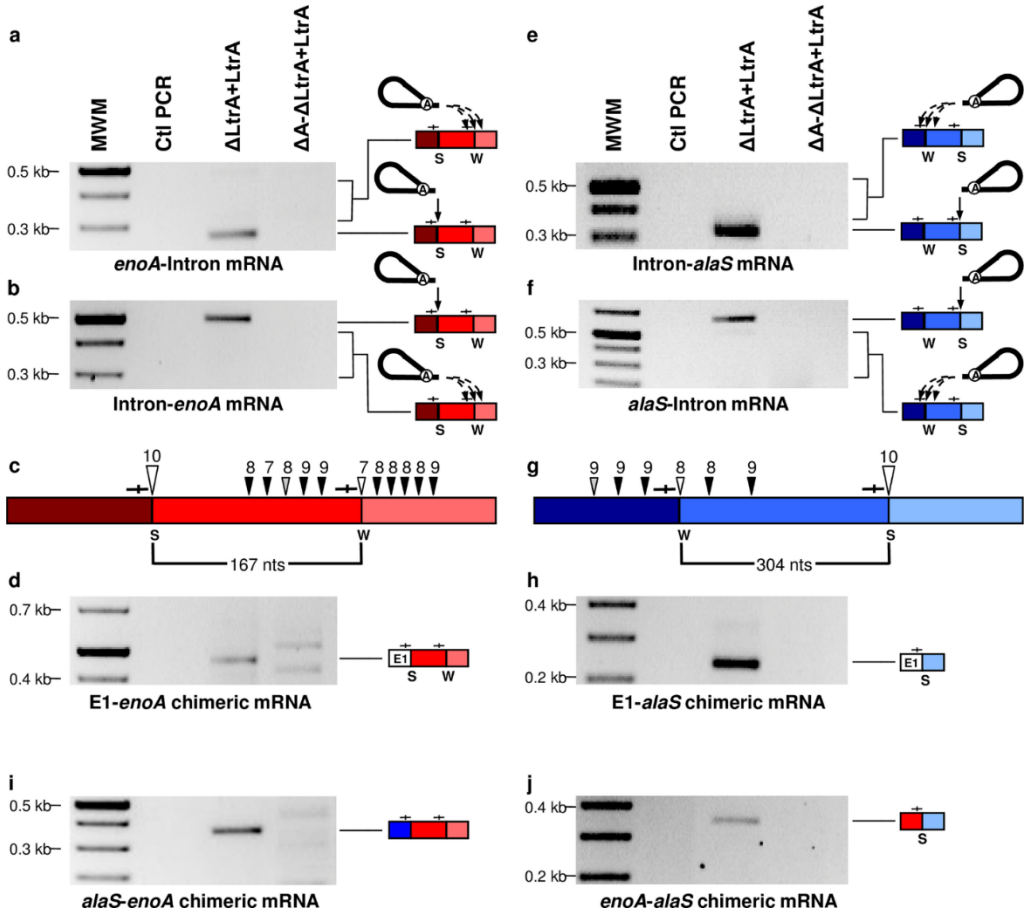
a

Ll.LtrB Variants	WT	EBS1/Mut	ΔA
Independent Reverse Splicing Events	344	19	0
Number of Intron-Interrupted mRNAs	135	13	0



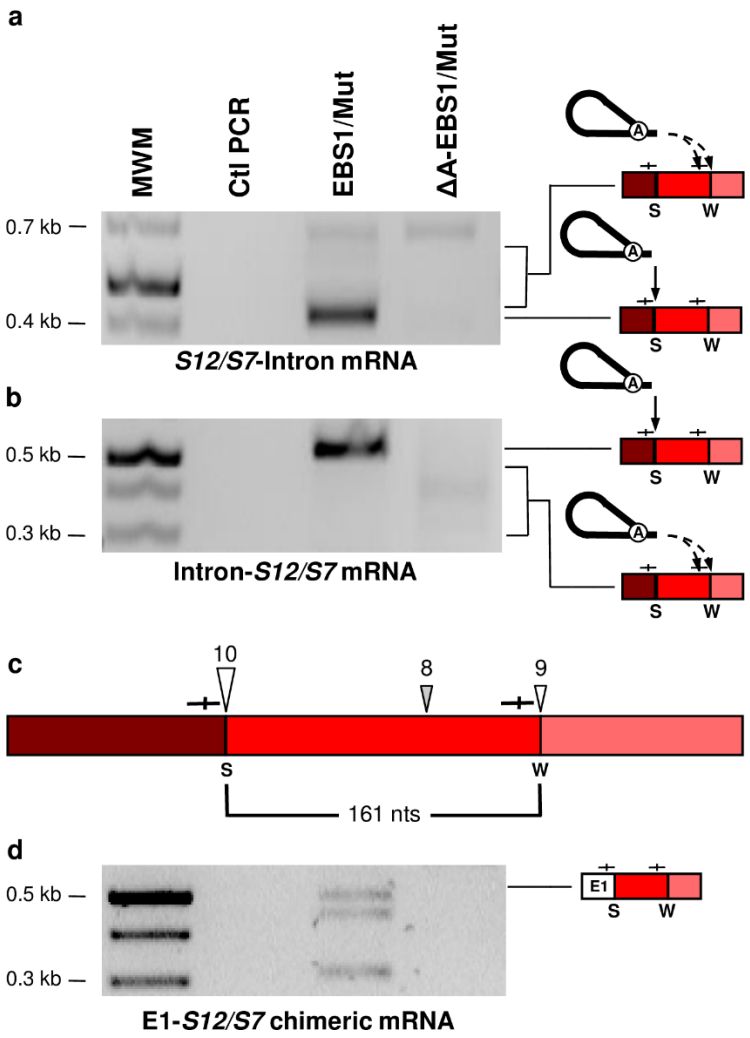
Ll.LtrB reverse splicing within *L. lactis* mRNAs. (**a**) Independent Ll.LtrB reverse splicing events were identified by total RNA-Seq for Ll.LtrB-ΔLtrA+LtrA, Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA and the control Ll.LtrB-ΔA-ΔLtrA+LtrA that exclusively splices through the circularization pathway and cannot reverse splice. The EBS1-IBS1 and EBS2-IBS2 base pairing interactions for Ll.LtrB-ΔLtrA+LtrA (**b**) and Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA (**c**) are depicted. Logo representations of the consensus sequences upstream (15 nts) and downstream (15 nts) from the intron reverse splice sites within the various *L. lactis* mRNAs are also shown.

Figure 3.7



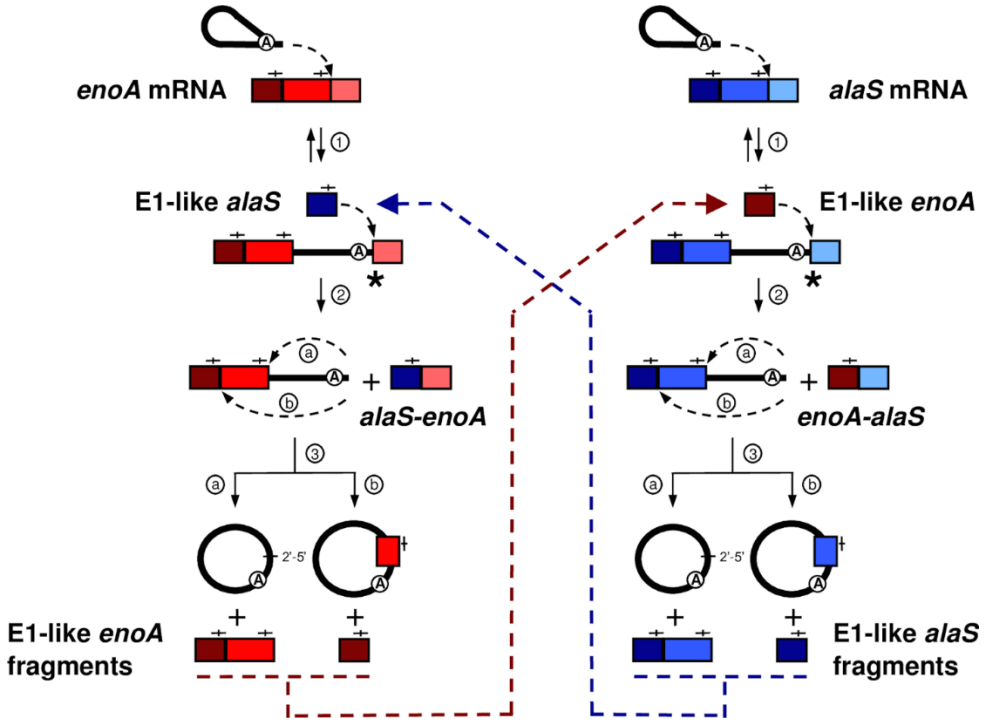
Detection of intermediates unique to the reverse splicing pathway: Ll.LtrB reverse-spliced within *L. lactis* mRNAs, E1-mRNA and mRNA-mRNA chimeras. RT-PCR assays were performed to detect the 5′ and 3′ junctions of Ll.LtrB-ΔLtrA+LtrA and Ll.LtrB-ΔA-ΔLtrA+LtrA reverse splicing events within the *enoA* (red boxes) (**a**, **b**) and *alaS* (blue boxes) (**e**, **f**) mRNAs. Complete and dashed arrows indicate reverse splicing of Ll.LtrB lariats within strong (S) and weak (W) IBS1/2-like sequences (—|—) respectively. The strong (S) (10/11 nts) (large arrowhead) and weak (W) (7-9/11 nts) (small arrowhead) IBS1/2-like sequences invaded by reverse splicing are represented (**c**, **g**). The sites flanking the mRNA fragments (**c**, 167 nts and **g**, 304 nts) initially detected at intron circle splice junctions (Fig. 3.3A) are indicated by open arrowheads. The Ll.LtrB insertion sites were identified in conditions where the *enoA* or the *alaS* genes were overexpressed (small open and black arrowheads) or not (large open and small gray arrowheads) from a P_{23} constitutive promoter. mRNA chimeras between *ltrB*-exon 1 (E1) and *L. lactis* mRNAs (**d**, E1-*enoA*)(**h**, E1-*alaS*) as well as between *L. lactis* mRNAs (**i**, *alaS-enoA*)(**j**, *enoA-alaS*) were also detected by RT-PCR at IBS1/2-like sequences.

Figure 3.8



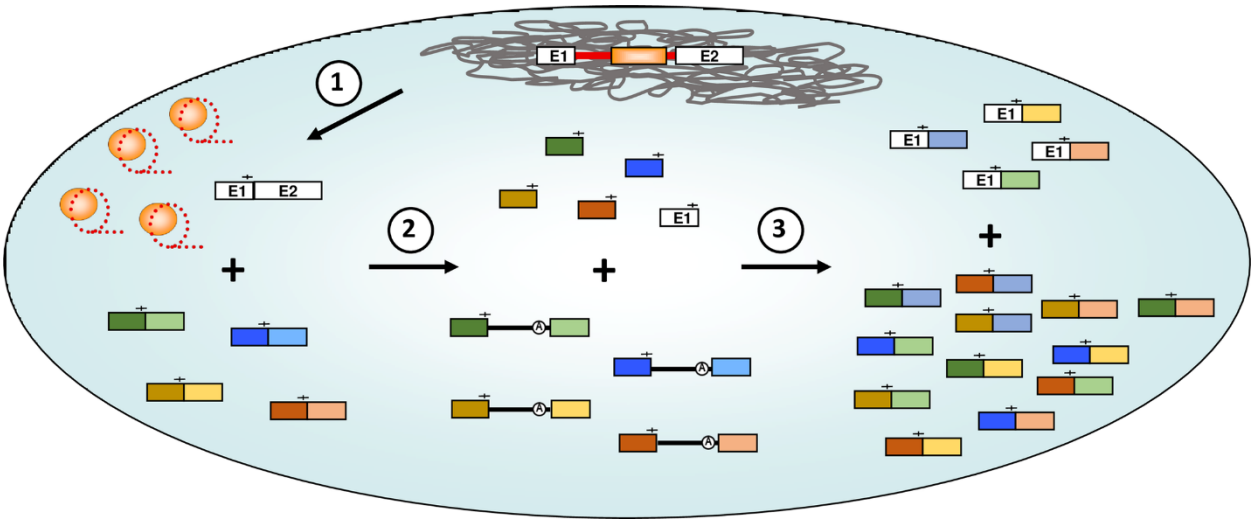
Detection of intermediates unique to the reverse splicing pathway: Ll.LtrB reverse-spliced within an *L. lactis* mRNA and an E1-mRNA chimera. RT-PCR assays were performed to detect the 5′ and 3′ junctions of Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA and Ll.LtrB-ΔA-EBS1/Mut-ΔLtrA+LtrA reverse splicing events within the Ribosomal Protein *S12/S7* mRNA (**a**, **b**). Complete and dashed arrows indicate reverse splicing of Ll.LtrB lariats within strong (S) and weak (W) IBS1/2-like sequences (—|—) respectively. The strong (S) (10/11 nts) (large arrowhead) and weak (W) (8-9/11 nts) (small arrowheads) IBS1/2-like sequences invaded by reverse splicing are represented (**c**). The sites flanking the mRNA fragment (**c**, 161 nts) initially detected at intron circle splice junctions (Fig. 3.3B) are indicated by open arrowheads. mRNA chimeras between *ltrB*-exon 1 (E1) and *L. lactis* mRNAs (**d**, E1-*S12/S7*) were also detected by RT-PCR at IBS1/2-like sequences.

Figure 3.9



The intergenic group II intron *trans*-splicing pathway leading to the generation of *alaS-enoA* and *enoA-alaS* mRNA chimeras. Ll.LtrB first recognizes by base pairing and invades by reverse splicing IBS1/2-like sequences on the *enoA* (red) and *alaS* (blue) mRNAs (step 1). The 3′OH of intron-processed *alaS* mRNA fragments (E1-like *alaS* fragments), harboring IBS1/2-like sequences (—|—) at their 3′ ends (dark blue and blue) (dark blue), can, similarly to free E1, be recruited by the intron to initiate the circularization pathway. These fragments can attack the intron-exon 2 splice junction of the Ll.LtrB-interrupted *enoA* mRNA (red) leading to the generation of *alaS-enoA* mRNA chimeras (step 2). Subsequent excision of the intron by circularization releases *enoA* mRNA fragments (E1-like *enoA* fragments) with IBS1/2-like sequences at their 3′ ends (step 3a, dark red and red) (step 3b, dark red), which can in turn initiate the generation of *enoA-alaS* mRNA chimeras with intron-interrupted *alaS* mRNA (blue).

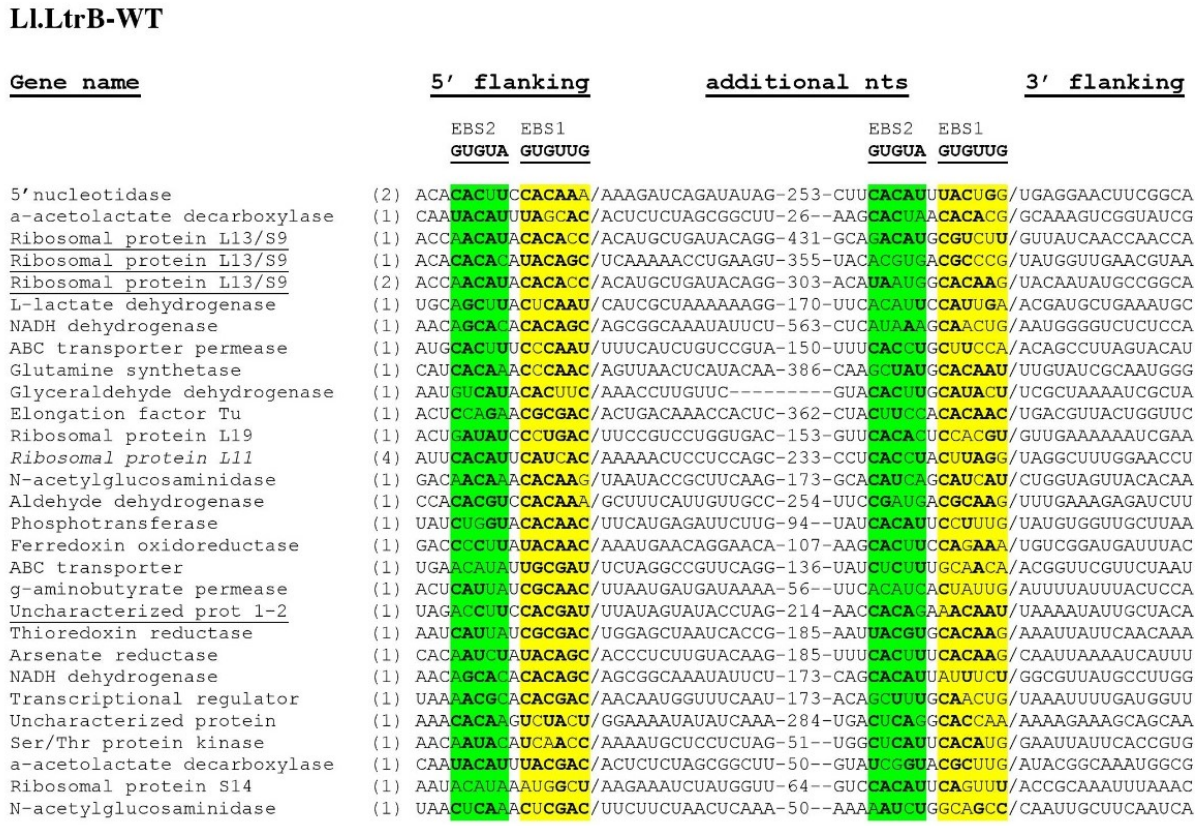
Figure 3.10



Model for group II intron-catalyzed genetic diversity. Upon expression of group II intron-interrupted genes in bacteria, the ribozymes self-splice using the conventional branching pathway, releasing a mix of RNPs (lariats + LtrA) and accurately ligated flanking exons (step 1). Excised RNPs next interact with cellular mRNA transcripts through specific base pairing with IBS1/2-like sequences (—|—). This interaction leads either to complete reverse splicing or hydrolysis at the IBS1/2-like sites, producing a population of intron-invaded mRNA transcripts or hydrolysed mRNA fragments with a free 3′-OH, respectively (step 2). When introns interrupting an ectopic site self-splice using the circularization pathway, they recruit either processed ectopic mRNA fragments or their processed cognate E1, which can both act as external nucleophiles in an intergenic *trans*-splicing reaction (step 3). This produces two distinct populations of chimeric mRNA transcripts: E1-mRNA and mRNA-mRNA products, which together increase the overall diversity of the bacterial host's transcriptome. The presence of a series of group II intron–interrupted mRNAs may potentially lead to a multitude of chimeric mRNA-mRNA combinations.

3.10: Supplementary Figures

Figure S3.1



mRNA fragments identified at the splice junction of Ll.LtrB-WT circles. Additional nts are shown along with their flanking sequences (5′ flanking) (3′ flanking), their origin (Gene name) and frequency of identification between parentheses. The junctions between the additional nts and their flanking regions (/) as well as the IBS1- (yellow) and IBS2- (green) like sequences are denoted. The bolded nts represent residues from the IBS1- and IBS2-like sequences that can potentially base pair with the intron's EBS1 and EBS2 sequences specified above. Sequences spanning two genes and including a short intergenic region are underlined while the mRNA sequence including 3′ untranslated residues is italicized.

Figure S3.2

mRNA fragments identified at the splice junction of Ll.LtrB-ΔLtrA+LtrA circles. Sequences of the additional nts are shown along with their flanking sequences (5′ flanking) (3′ flanking), their origin (Gene name) and frequency of identification between parentheses. The junctions between the additional nts and their flanking regions (/) as well as the IBS1- (yellow) and IBS2- (green) like sequences are denoted. Some IBS1/2-like sequences were adjusted to optimize their potential base pairing with the EBS1/2 sequences of the intron. The number of nts separating the IBS1/2-like sequences was fixed between 0-2 nts, and their maximum distance from the junction with the intron was fixed between -14, +4 nts. The bolded nts represent residues from the IBS1- and IBS2-like sequences that can potentially base pair with the intron's EBS1 and EBS2 sequences specified above. Sequences spanning two genes and including a short intergenic region are underlined. The genes in bold (*alaS* and *enoA*) were further studied for Ll.LtrB reverse splicing analyses and the detection of E1-mRNA and mRNA-mRNA chimeras (Fig. 3.7).

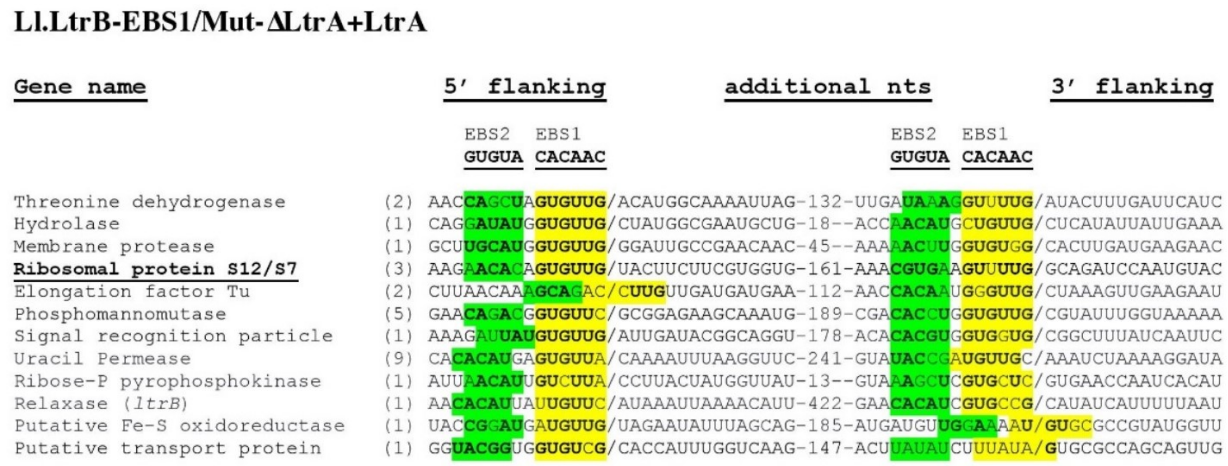
Figure S3.3

Ll.LtrB-WT

Gene name		5′ flanking	EBS2 GUGUA	EBS1 GUGUUG	additional nts	EBS2 GUGUA	EBS1 GUGUUG	3′ flanking
5′nucleotidase	(2)	ACA	CACUU	CCACAA	A/AAAGAUCAGAUAUAG-253-CUU	CACAU	UUACUGG	/UGAGGAACUUCGGCA
a-acetolactate decarboxylase	(1)	CAA	UACAUUUA	GCAC	/ACUCUCUAGCGGCUU-26--AAG	CAC	UAACACA	CG/GCAAAGUCGGUAUCG
Ribosomal protein L13/S9	(1)	ACC	AACAUA	CACACC	/ACAUGCUGAUACAGG-431-GCAGACA	UGCGU	CUU/GUU	AUCAACCAACCA
Ribosomal protein L13/S9	(1)	ACA	CACAGA	UACAGC	/UCAAAAACCUGAAGU-355-UA	CACGU	GACGCCCG	/UAUGGUUGAACGUAA
Ribosomal protein L13/S9	(2)	ACC	AACAUA	CACACC	/ACAUGCUGAUACAGG-303-ACAUA	AUGG	CACAAG	/UACAAUAUGCCGGCA
L-lactate dehydrogenase	(1)	UGC	AGCUUA	CUCAAU	/CAUCGCUAAAAAAGG-170-UU	CACAU	UCCAUUGA	/ACGAUGCUGAAAUGC
NADH dehydrogenase	(1)	AAC	AGCA	CACACAGC	/AGCGGCAAAUAUUCU-563-CU	CAUA	AAGCAAC	UG/AAUGGGGUCUCUCCA
ABC transporter permease	(1)	AUG	CACUUU	CCCAAU	/UUUCAUCUGUCCGUA-150-UUUCA	CCUGC	UUCCA/AC	AGCCUUAGUACAU
Glutamine synthetase	(1)	CAU	CACAA	ACCCAAC	/AGUUAACUCAUACAA-386-CAA	GCUAU	GCACAAU	/UUGUAUCGCAAUGGG
Glyceraldehyde dehydrogenase	(1)	AA	UGU	CA	UACA	CUUC	/AAACCUUGUUC---------	GUA	CACUU	GCAUACU	/UCGCUAAAAUCGCUA
Elongation factor Tu	(1)	ACUC	CAGAA	CGCGAC	/ACUGACAAACCACUC-362-CUA	CUUCC	ACACAAC	/UGACGUUACUGGUUC
Ribosomal protein L19	(1)	ACU	GAUAU	CCCUGAC	/UUCCGUCCUGGUGAC-153-GUUCA	CACUC	CACGU/GUU	GAAAAAAUCGAA
Ribosomal protein L11	(4)	AUU	CACAUU	CAUCAC	/AAAAACUCCUCCAGC-233-CCU	CACC	UACUUAG	G/UAGGCUUUGGAACCU
N-acetylglucosaminidase	(1)	GAC	AACA	AACACAA	G/UAAUACCGCUUCAAG-173-GCA	CAUC	AGCAUCAU	/CUGGUAGUUACACAA
Aldehyde dehydrogenase	(1)	CCA	CACGU	CCACAA	A/GCUUUCAUUGUUGCC-254-UUCC	GAUGA	CGCAAG	/UUUGAAAGAGAUCUU
Phosphotransferase	(1)	UAU	CUGGUA	CACAAC	/UUCAUGAGAUUCUUG-94--UAU	CACAU	UCCUUUG	/UAUGUGGUUGCUUAA
Ferredoxin oxidoreductase	(1)	GAC	CCUUA	UACAAC	/AAAUGAACAGGAACA-107-AAG	CACUU	CCAGAAA	/UGUCGGAUGAUUUAC
ABC transporter	(1)	UGAA	CAUAUU	GCGAU	/UCUAGGCCGUUCAGG-136-UA	UCUCU	UUGCAAC	A/ACGGUUCGUUCUAAU
g-aminobutyrate permease	(1)	ACU	CAU	UAUCGCAAC	/UUAAUGAUGAUAAAA-56--UU	CACAUCAC	UAUUG	/AUUUUAUUUACUCCA
Uncharacterized prot 1-2	(1)	UA	GACCU	UCCACGAU	/UUAUAGUAUACCUAG-214-AAC	CACAG	AAACAAU	/UAAAAUAUUGCUACA
Thioredoxin reductase	(1)	AAU	CAUUA	UCGCGAC	/UGGAGCUAAUCACCG-185-AAU	UACGU	GCACAAG	/AAAUUAUUCAACAAA
Arsenate reductase	(1)	CAC	AAUCU	AUACAGC	/ACCCUCUUGUACAAG-185-UUU	CACUU	UCACAAG	/CAAUUAAAAUCAUUU
NADH dehydrogenase	(1)	AAC	AGCA	CACAGC	/AGCGGCAAAUAUUCU-173-CAG	CACAUUAU	UUCU	/GGCGUUAUGCCUUGG
Transcriptional regulator	(1)	UAA	AACG	CACGAC	/AACAAUGGUUUCAAU-173-AC	AGCU	UUGCAAC	UG/UAAAUUUUGAUGGUU
Uncharacterized protein	(1)	AAA	CACA	AGUCUACU	/GGAAAAUAUAUCAAA-284-UGA	CUCA	GGCACCAA	/AAAAGAAAGCAGCAA
Ser/Thr protein kinase	(1)	AAC	AAUA	CAUCAAC	C/AAAAUGCUCCUCUAG-51--UGG	CUCAU	UCACA	UG/GAAUUAUUCACCGUG
a-acetolactate decarboxylase	(1)	CAA	UACAUU	UACGAC	/ACUCUCUAGCGGCUU-50--GUAUC	GGUA	CGCUG/AU	ACGGCAAAUGGCG
Ribosomal protein S14	(1)	AA	UACAUA	AAUGGC	U/AAGAAAUCUAUGGUU-64--GUCCACA	UUCA	GUUU/ACC	GCAAAUUUAAAC
N-acetylglucosaminidase	(1)	UAA	CUCA	AACUCGAC	/UUCUUCUAACUCAAA-50--AAAA	AUCU	GGCAGC	C/CAAUUGCUUCAAUCA

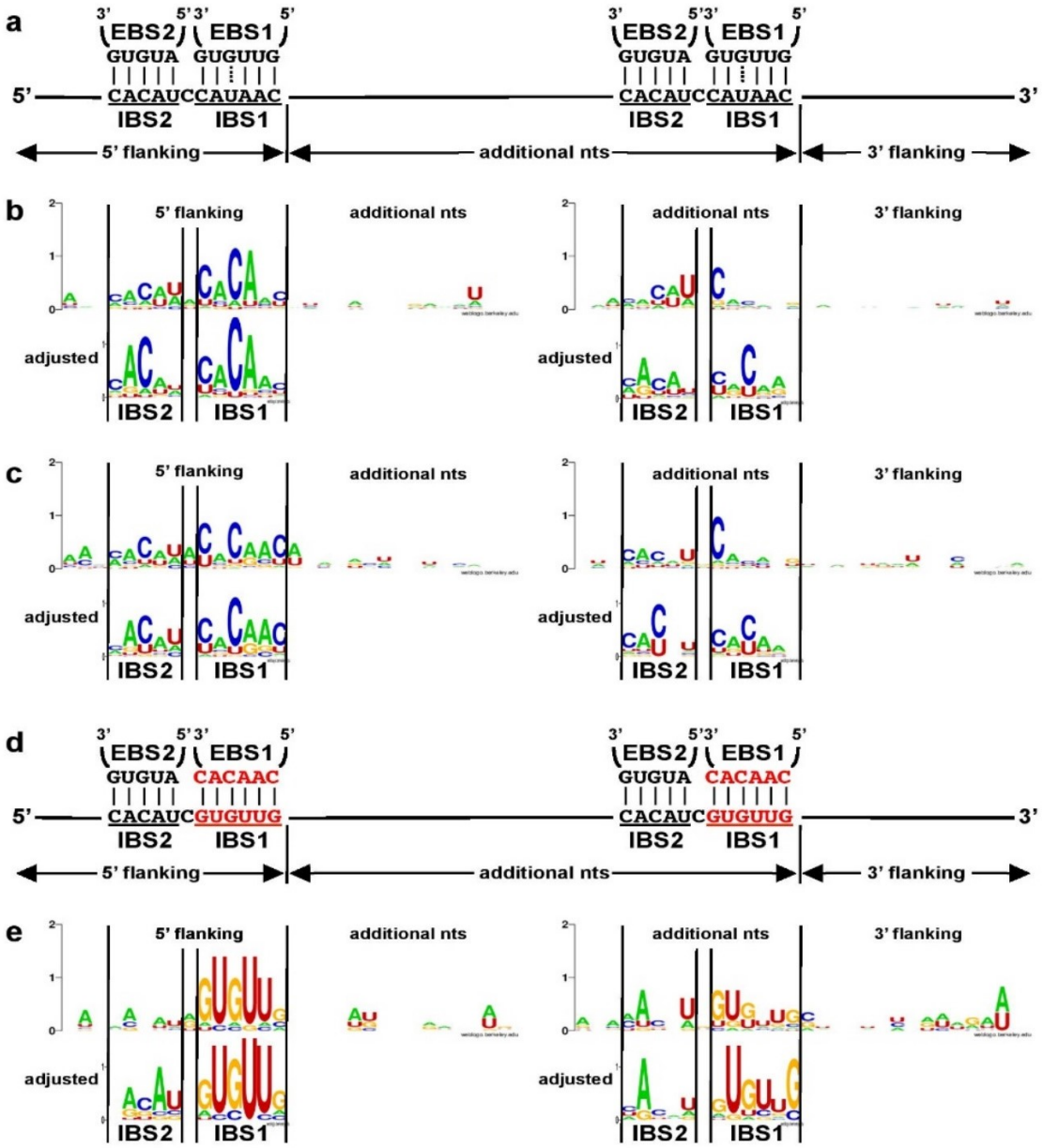
mRNA fragments identified at the splice junction of Ll.LtrB-WT circles. Sequences of the additional nts are shown along with their flanking sequences (5′ flanking) (3′ flanking), their origin (Gene name) and frequency of identification between parentheses. The junctions between the additional nts and their flanking regions (/) as well as the IBS1- (yellow) and IBS2- (green) like sequences are denoted. Some IBS1/2-like sequences were adjusted to optimize their potential base pairing with the EBS1/2 sequences of the intron. The number of nts separating the IBS1/2-like sequences was fixed between 0-2 nts, and their maximum distance from the junction with the intron was fixed between -14, +4 nts. The bolded nts represent residues from the IBS1- and IBS2-like sequences that can potentially base pair with the intron's EBS1 and EBS2 sequences specified above. Sequences spanning two genes and including a short intergenic region are underlined.

Figure S3.4



mRNA fragments identified at the splice junction of Ll.LtrB-EBS1/Mut circles. Sequences of the additional nts are shown along with their flanking sequences (5′ flanking) (3′ flanking), their origin (Gene name) and frequency of identification between parentheses. The junctions between the additional nts and their flanking regions (/) as well as the IBS1- (yellow) and IBS2- (green) like sequences are denoted. Some IBS1/2-like sequences were adjusted to optimize their potential base pairing with the EBS1/2 sequences of the intron. The number of nts separating the IBS1/2-like sequences was fixed between 0-2 nts, and their maximum distance from the junction with the intron was fixed between -14, +4 nts. The bolded nts represent residues from the IBS1- and IBS2-like sequences that can potentially base pair with the intron's EBS1 and EBS2 sequences specified above. Sequences spanning two genes and including a short intergenic region are underlined. The gene in bold (*S12/S7*) was further studied for Ll.LtrB reverse splicing analyses and the detection of E1-mRNA chimeras (Fig. 3.8).

Figure S3.5



Logo representations of the consensus sequences (30 nts) around the 5′ and 3′ extremities of the mRNA fragments identified within circle splice junctions. When WT base pairing interactions for Ll.LtrB-ΔLtrA+LtrA and Ll.LtrB-WT (**a**) as well as Ll.LtrB-EBS1/Mut (**d**) are depicted. The consensus (Figs. 3.3, S3.1) and adjusted consensus (Figs. S3.2-S3.4) sequences are shown for Ll.LtrB-ΔLtrA+LtrA (**b**), Ll.LtrB-WT (**c**) and Ll.LtrB-EBS1/Mut-ΔLtrA+LtrA (**e**).

3.11: Tables

Table S3.1:

Plasmids used in this study

Plasmid name	Size and antibiotic resistance	Description
pDL-P_{23}^{2}-Ll.LtrB-WT	11 kb, Spc^{R} (300µg/µl)	*E. coli-L. lactis* shuttle vector harboring wild-type Ll.LtrB expressed under the control of the P_{23} constitutive promoter
pDL-P_{23}^{2}-Ll.LtrB-ΔLtrA	9.2 kb, Spc^{R} (300µg/µl)	Ll.LtrB lacking its essential intron-encoded protein (LtrA)
pDL-P_{23}^{2}-Ll.LtrB-ΔA-ΔLtrA	9.2 kb, Spc^{R} (300µg/µl)	Ll.LtrB lacking LtrA and the branch point adenosine residue
pDL-P_{23}^{2}- Ll.LtrB-EBS1/Mut-ΔLtrA	9.2 kb, Spc^{R} (300µg/µl)	Ll.LtrB lacking LtrA and mutated in the EBS1 (intron) and IBS1 (exon1) regions to modify base pairing specificity
pDL-P_{23}^{2}- Ll.LtrB-ΔA-EBS1/Mut-ΔLtrA	9.2 kb, Spc^{R} (300µg/µl)	Ll.LtrB lacking LtrA, the branch point adenosine and mutated in the EBS1 (intron) and IBS1 (exon1) regions to modify base pairing specificity
pLE-P_{23}^{2}-LtrA	10.9 kb, Cam^{R} (10µg/µl)	*E. coli-L. lactis* shuttle vector harboring the ltrA gene expressed under the control of the P_{23} constitutive promoter
pLE-P_{23}^{2}-LtrA+AlaS	11.7 kb, Cam^{R} (10µg/µl)	Alanyl-tRNA synthetase gene (*alaS*) from *L. lactis* expressed under the control of the P_{23} constitutive promoter
pLE-P_{23}^{2}-LtrA+EnoA	10.4 kb, Cam^{R} (10µg/µl)	Enolase gene (*enoA*) from *L. lactis* expressed under the control of the P_{23} constitutive promoter

Table S3.2:

Primers used in this study

Primer location	Function	Sequence (5'-3')
Ll.LtrB 5' end (-) strand	RT of splice junction and PCR of 5' junctions of reverse splicing	CGATTGTCTTTAGGTAACTCAT
Ll.LtrB 5' end (+) strand	PCR of circle splice junction	TTAAACTACTTGACTTAACACCC
Ll.LtrB 3' end (-) strand	PCR of circle splice junction	TGTGAACAAGGCGGTACCTC
alaS 5' end (+) strand	Cloning *alaS* gene for overexpression	AAAAGCGCGCGTGGTACCGCGGTATAACTGT
alaS 3' end (-) strand	Cloning *alaS* gene for overexpression	AAAAGCGCGCCTATCCTAATTTTTCAGCAACAGC
enoA 5' end (+) strand	Cloning *enoA* gene for overexpression	AAAAGCGCGCTGTATTAAGAGTGCAGACGCAC
enoA 3' end (-) strand	Cloning *enoA* gene for overexpression	AAAAGCGCGCTTAATGCATTTTTTTAAGGTTGTAGAATGCTTTAA
alaS 3' end (-) strand	RT to screen for *ltrB* E1-*alaS* chimeras	CCAATCCAGCCACTTCGCTC
alaS 3' end (-) strand	PCR to screen for *ltrB* E1-*alaS* chimeras	CTGTCACAGCAATAATCCGGC
ltrB-E1 3' end (+) strand	PCR to screen for *ltrB* E1-*alaS*/*enoA* chimeras	TTGGTCATCACCTCATCCAATC
enoA 3' end (-) strand	RT to screen for *ltrB* E1-*enoA* chimeras, PCR for 3' junction of Ll.LtrB reverse splicing in *enoA*	CATAACGCCATCTTCACCAGC
enoA 3' end (-) strand	PCR to screen for *ltrB* E1-*enoA* chimeras	TAACCAGCTGCTTCGATTGCTT
alaS 3' end (-) strand	RT of 5' and 3' junctions of Ll.LtrB reverse splicing in *alaS*	CAACGATTTGTGAAGCTTGTGG
alaS 5' end (+) strand	PCR of 5' junction of Ll.LtrB reverse splicing in *alaS*	GGTCAAGTGGTGGCAACTGT
alaS 3' end (-) strand	PCR of 3' junction of Ll.LtrB reverse splicing in *alaS*	AACCAATCCAGCCACTTCGCT
Ll.LtrB 3' end (+) strand	PCR of 3' junction of Ll.LtrB reverse splicing in *alaS* and *enoA*	CTCTTGTTGTATGCTTTCATTG
enoA 3' end (-) strand	RT of 5' and 3' junctions of Ll.LtrB reverse splicing in *enoA*	GAATTCTGATGATGCACAGTCG
enoA 5' end (+) strand	PCR of 5' junction of Ll.LtrB reverse splicing in *enoA*	ATGATCGCTCTTGACGGTACT
Ll.LtrB 3' end (-) strand	PCR to remove branch point adenosine residue from Ll.LtrB-ΔLtrA	GGCGGTACCTCCCTCTTCACCATATCAT
Intron 5' end (+) strand	PCR to mutate the EBS1 sequence of Ll.LtrB	GTAAGTTATGCAACACGACTTATCTGTTATCACCACC
Intron 5' end (-) strand	PCR to mutate the EBS1 sequence of Ll.LtrB	CTTTCTTTGTACTAGAGGTTTC
ltrB-E1 3' end (+) strand	PCR to mutate the IBS1 sequence of *ltrB*-E1	TGAACACATCGTGTTGGTGCGCCCAGATAGGGTGTTAAG
ltrB-E1 3' end (-) strand	PCR to mutate the IBS1 sequence of *ltrB*-E1	CGATCGACGTGGGTTGCA

Chapter 4:
Group II introns generate functional chimeric relaxase enzymes with modified specificities through exon shuffling at both the RNA and DNA level

4.1: Preface

As discussed in Chapter 3, group II introns have the potential to combine aspects of branching and circularization to generate chimeric mRNAs and thus expand the complexity of the bacterial host transcriptome. Although we proposed that increasing genetic diversity could be beneficial to bacteria, we failed to specifically demonstrate how this might occur (LaRoche-Johnston et al. 2018). We thus used the native biological context of Ll.LtrB in *Lactococcus lactis* to address the functionality of chimeric mRNA formation.

In this study, we found that chimeric mRNAs generated by Ll.LtrB can be functional and lead to gain-of-function phenotypes. Ll.LtrB can form chimeric mRNAs between the relaxase transcripts of *ltrB* from *L. lactis* and *pcfG* from *E. faecalis* (LaRoche-Johnston et al. 2016). We show that these molecules are formed in natural settings under biologically relevant expression levels, and are consistently produced at the E1-E2 homing site recognized by these introns (Staddon et al. 2004). Finally, we demonstrate the existence of chimeric genes, suggesting the existence of a pathway where genetic diversity catalyzed by group II introns might be fixed in the bacterial chromosome.

This chapter was adapted from the following manuscript: “Group II introns generate functional chimeric relaxase enzymes with modified specificities through exon shuffling at both the RNA and DNA level”. **Félix LaRoche-Johnston**, Rafia Bosan and Benoit Cousineau. 2020. *Molecular Biology and Evolution*, 38(3):1075-1089.

Group II Introns Generate Functional Chimeric Relaxase Enzymes with Modified Specificities through Exon Shuffling at Both the RNA and DNA Level

Félix LaRoche-Johnston,[1] Rafia Bosan,[1] and Benoit Cousineau*,[1]

[1]Department of Microbiology and Immunology, McGill University, Montréal, Québec, Canada

***Corresponding author:** E-mail: benoit.cousineau@mcgill.ca.

Associate editor: Irina Arkhipova

4.2: Summary

Group II introns are large self-splicing RNA enzymes with a broad but somewhat irregular phylogenetic distribution. These ancient retromobile elements are the proposed ancestors of approximately half the human genome, including the abundant spliceosomal introns and non-LTR retrotransposons. In contrast to their eukaryotic derivatives, bacterial group II introns have largely been considered as harmful selfish mobile retroelements that parasitize the genome of their host. As a challenge to this view, we recently uncovered a new intergenic *trans*-splicing pathway that generates an assortment of mRNA chimeras. The ability of group II introns to combine disparate mRNA fragments was proposed to increase the genetic diversity of the bacterial host by shuffling coding sequences.

Here we show that the Ll.LtrB and Ef.PcfG group II introns from *Lactococcus lactis* and *Enterococcus faecalis* respectively can both use the intergenic *trans*-splicing pathway to catalyze the formation of chimeric relaxase mRNAs and functional proteins. We demonstrated that some of these compound relaxase enzymes yield gain-of-function phenotypes, being significantly more efficient than their precursor wild-type enzymes at supporting bacterial conjugation. We also found that relaxase enzymes with shuffled functional domains are produced in biologically relevant settings under natural expression levels. Finally, we uncovered examples of lactococcal chimeric relaxase genes with junctions exactly at the intron insertion site. Overall, our work demonstrates that the genetic diversity generated by group II introns, at the RNA level by intergenic *trans*-splicing and at the DNA level by recombination, can yield new functional enzymes with shuffled exons, which can lead to gain-of-function phenotypes.

4.3: Introduction

Group II introns are large RNA enzymes that are widely found throughout bacteria, some archaea and the organelles of certain eukaryotes (Lambowitz and Zimmerly 2011; McNeil et al. 2016). Following transcription, these versatile ribozymes associate with their intron-encoded protein (IEP) to fold into a catalytically competent three-dimensional RNA structure that concurrently enables ligation of the flanking exons and self-splicing of the intron through different pathways (Fedorova and Zingler 2007; Pyle 2016). The most studied self-splicing pathway is branching, where the intron uses a bulged adenosine residue, located near its 3′ end, as a branchpoint to excise as a lariat. Once released as a lariat, the intron can use the reverse branching pathway to invade target sites in DNA or RNA substrates. Following insertion into DNA substrates, group II introns can function as retromobile elements to form stable DNA copies of themselves by completing either the retrohoming or retrotransposition mobility pathways (Cousineau et al. 1998; Cousineau et al. 2000; Ichiyanagi et al. 2002). Alternatively, bacterial group II introns can initiate self-splicing through less characterized pathways such as circularization, where they catalyze a *trans*-splicing reaction by recruiting an external nucleophile (Monat et al. 2015; Monat and Cousineau 2016).

When seen through an evolutionary lens, group II introns are believed to have played a monumental role in the evolution of eukaryotes. Due to conserved facets of their biochemical and structural properties, group II introns are the proposed evolutionary ancestors of spliceosomal introns and the non-LTR retrotransposons, which together account for over half of the human genome (Lambowitz and Belfort 2015; McNeil et al. 2016). Although spliceosomal introns are largely seen as beneficial to eukaryotes by increasing their genetic diversity and overall complexity through regulated pathways such as alternative splicing and *trans*-splicing (Irimia and Roy 2014; Bush et al. 2017), an opposite view has emerged for their bacterial ancestors. Bacterial group II introns are indeed considered as detrimental, selfish elements subjected to negative selection. They are thought to invade other DNA target sites in order to spread and survive, using splicing solely as a means of preventing damage to their hosts (Dai and Zimmerly 2002; Leclercq and Cordaux 2012). Although ancestral group II introns were most likely selfish and proliferative before the emergence of the first eukaryotes, their current descendants found in various prokaryotic genomes

may have evolved specific benefits to their hosts. In this study, we challenge the current paradigm that modern bacterial group II introns behave exclusively as selfish elements.

We recently characterized a new group II intron intergenic *trans*-splicing pathway that increases the genetic diversity of its bacterial host at the RNA level by combining aspects of both the branching and circularization pathways (LaRoche-Johnston et al. 2018). In this novel splicing pathway, excised group II intron lariats first recognize sequence motifs on bacterial mRNA substrates through base pairing and invade them via reversal of the branching pathway. The intron can target multiple sites on mRNAs since the 11 nt-interaction can occur with some mismatches and also because most of the intron sequence motifs (10/11) are made of Gs and Us that can potentially base pair with Cs or Us and As or Gs, respectively. Once inserted in a non-cognate host mRNA, the intron excises using the circularization pathway, where it *trans*-splices an external RNA to its downstream mRNA fragment, thus forming a chimeric mRNA (LaRoche-Johnston et al. 2018). The stochastic nature of this pathway should most of the time result in mRNA-mRNA chimeras that would code for non-functional proteins. Although we demonstrated the production of a variety of E1-mRNA and mRNA-mRNA chimeras *in vivo*, it remained unclear whether these compound transcripts are functional and biologically relevant. We thus examined the native biological context of the model group II intron Ll.LtrB from the gram-positive bacterium *Lactococcus lactis* to determine whether group II intron-generated chimeric mRNAs are translated and if their corresponding chimeric proteins are functional.

Ll.LtrB interrupts the *ltrB* gene, which codes for a conjugative relaxase enzyme, LtrB. This single-strand endonuclease is part of a DNA processing complex called relaxosome that assembles at the origin of transfer (*oriT*) of conjugative elements (Smillie et al. 2010). A key partner of LtrB in the lactococcal relaxosome is the MobC-family accessory protein LtrF, which binds an inverted repeat directly adjacent to the *oriT*. This interaction is essential for the specific recruitment of LtrB to the *oriT* (Chen et al. 2007). Once the LtrB-LtrF-*oriT* complex is formed, the dsDNA is locally unwound, allowing a direct interaction between LtrB and the *oriT*-ssDNA (Chen et al. 2007). Once bound to ssDNA, LtrB initiates conjugation by nicking *oriT* and remains covalently bound to the liberated 5′-phosphate (Byrd and Matson 1997). Through protein-protein interactions, the relaxosome next binds to a type-4 coupling protein ATPase (T4CP) at the host membrane, which directs the conjugative element through a mating channel formed by a type 4 secretion system

(T4SS) (Chen et al. 2008). By interrupting relaxase genes of various conjugative elements, group II introns use conjugation as a means of survival by spreading to different bacterial strains and species (Belhocine et al. 2004; Belhocine et al. 2005; Belhocine et al. 2007b). Moreover, since the relaxase enzyme is an essential component of the conjugative machinery, the accuracy and efficiency of Ll.LtrB self-splicing from the mRNA essentially controls the conjugation of its host element (Mills et al. 1996; Klein et al. 2004). LtrB was also shown to have off-target DNA nicking activity that stimulates both the frequency and diversity of Ll.LtrB retrotransposition events, revealing yet another link between group II intron dissemination and conjugation (Novikova et al. 2014).

Ll.LtrB interrupts a specific site of *ltrB*, consisting of a highly conserved catalytic histidine triad in the IncP family of relaxases (Pansegrau et al. 1994). The conserved nature of this catalytic motif has been proposed to be advantageous for the dissemination of Ll.LtrB in *L. lactis*, providing an abundance of unoccupied sites to invade in orthologous relaxase genes (Staddon et al. 2004). Indeed, we previously described a recent burst of mobility by Ll.LtrB variants within various *L. lactis* strains and subspecies, nearly all of which specifically invaded the conserved histidine triad (LaRoche-Johnston et al. 2016). We thus hypothesized that the conserved nature of the relaxase recognition site and the frequent exposure of introns to orthologous relaxase genes may enable the consistent production of chimeric relaxase mRNAs containing shuffled exons, which could then be translated into chimeric enzymes with potentially altered functions.

Here we demonstrate that one of the effects of group II introns increasing bacterial genetic diversity is the production of chimeric relaxase mRNAs and active enzymes under biologically relevant conditions. We show that since relaxase exons exert different functions during conjugation, group II intron *trans*-splicing of exons from orthologous relaxase mRNAs can produce chimeric enzymes with gain-of-function phenotypes that enhance the spread of conjugative elements. Finally, we also uncovered examples of chimeric relaxase genes throughout *L. lactis* strains and sub-species with junctions located precisely at the site of group II intron insertion. Overall, our data show for the first time that group II introns can be beneficial to their hosts by producing novel compound transcripts and proteins of functional value to the bacteria, which may have played an important role in the rapid adaptation of *L. lactis* to the dairy environment.

4.4: Results

4.4.1: The Ll.LtrB group II intron from *L. lactis* generates mRNA and protein chimeras between orthologous relaxase genes *in vivo*

We previously demonstrated that the Ll.LtrB group II intron mediates the formation of various E1-mRNA and mRNA-mRNA chimeras in *L. lactis* through a novel intergenic *trans*-splicing pathway (LaRoche-Johnston et al. 2018). To assess whether Ll.LtrB could use this pathway to generate in-frame chimeric relaxase mRNAs that are functional substrates for translation, we built a group II intron *trans*-splicing assay containing the relaxase genes from *L. lactis* and *E. faecalis* (Fig. 4.1). In their native environment, the highly similar and homologous Ll.LtrB and Ef.PcfG group II introns interrupt respectively the *ltrB* and *pcfG* orthologous relaxase genes at the exact same conserved position at the junction between two codons (LaRoche-Johnston et al. 2016) (Fig. 4.1A). As a consequence of the recent lateral transfer of Ef.PcfG from *E. faecalis* to *L. lactis* (LaRoche-Johnston et al. 2016), the flanking exons of the relaxase genes are significantly less conserved (E1:60%, E2: 58%) than the introns they harbor (99.7%) (Fig. 4.1A).

The group II intron *trans*-splicing assay consisted of co-expressing the *ltrB* and *pcfG* relaxase genes in *L. lactis* from the pLE and pDL plasmids respectively (Fig. 4.1B). The *ltrB* gene was under the control of the nisin-inducible promoter (P_{nis}) and interrupted by its cognate Ll.LtrB intron, while the non-interrupted *pcfG* gene was expressed from a constitutive promoter (P_{23}) (Fig. 4.1B). The previously described group II intron intergenic *trans*-splicing pathway (Fig. 4.1C) (LaRoche-Johnston et al. 2018) predicts that free E1 from the interrupted mRNA (*ltrB*E1) should be *trans*-spliced to E2 of the non-interrupted mRNA (*pcfG*E2), precisely at the E1-E2 splice junction (Fig. 4.1C, step 5). We thus performed an RT-PCR assay across the predicted chimeric *ltrB*E1-*pcfG*E2 mRNA splice junction from total RNA extracts of *L. lactis* expressing the *ltrB* and *pcfG* genes (Monat et al. 2015; Monat and Cousineau 2016; LaRoche-Johnston et al. 2018). An amplicon of the expected size was obtained specifically when *ltrB* was interrupted by Ll.LtrB-WT and was absent for the branchpoint mutant, Ll.LtrB-ΔA (Fig. 4.1D). The Ll.LtrB-ΔA negative control cannot support E1 *trans*-splicing since it is unable to splice through the requisite branching (Fig. 4.1C, steps 1-2) and reverse branching pathways (Fig. 4.1C, steps 3-4) (Monat et al. 2015;

Monat and Cousineau 2016). Sequencing of the RT-PCR amplicon confirmed the identity of the *ltrB*E1-*pcfG*E2 chimeric mRNA where E1 of the interrupted *ltrB* mRNA was precisely ligated to E2 of the *pcfG* non-interrupted relaxase transcript.

To analyze whether the chimeric *ltrB*E1-*pcfG*E2 relaxase transcript is translated into a chimeric protein (Fig. 4.1C, step 7), the non-interrupted *pcfG* gene included a 6X His-tag at the C-terminus as well as an in-frame deletion of 366 nts in E1 (Fig. 4.1B). When we performed Western Blots using His-tag primary antibodies on total protein extracts, a strong signal at ~ 52 kDa was detected when *ltrB* was interrupted by either Ll.LtrB-WT or Ll.LtrB-ΔA (Fig. 4.1E). This band corresponds to the contiguous PcfG relaxase protein harboring a deletion in E1 (PcfGE1Δ366-PcfGE2). We also observed an additional band at ~ 67 kDa that corresponds to the size of the chimeric LtrBE1-PcfGE2 relaxase protein exclusively when *ltrB* was interrupted by Ll.LtrB-WT (Fig. 4.1E).

Overall, our data show that the Ll.LtrB group II intron can catalyze the formation of in-frame chimeric relaxase transcripts as well as detectable levels of chimeric relaxase proteins in *L. lactis*.

4.4.2: Chimeric relaxase enzymes are active in *L. lactis* and can confer a gain-of-function phenotype

Having demonstrated that Ll.LtrB can generate chimeric relaxase proteins in *L. lactis*, we next wanted to assess whether these enzymes were active. We thus engineered a conjugation assay to study the activity of chimeric relaxase enzymes using NZ9800Δ*ltrB* and LM0231 as the donor and recipient strain respectively (Fig. 4.2A) (Belhocine et al. 2004). The donor strain contained all the conjugation machinery to support the transfer of conjugative elements harboring an origin of transfer (*oriT*), except for the *ltrB* relaxase gene. Donor cells were co-transformed with two plasmids, one expressing wild-type or chimeric relaxase enzymes while the second plasmid harbored an *L. lactis* or an *E. faecalis oriT* (Fig. 4.2A). This conjugation assay allowed us to study the efficiency with which different relaxase enzymes recognize various *oriT*s and support the transfer of mobilizable plasmids between strains of *L. lactis* (Fig. 4.2B).

We first looked at the ability of four different relaxase enzymes to support the transfer of a plasmid harboring the *oriT* from the *L. lactis* pRS01 plasmid (Mills et al. 1996) (Fig. 4.2B, first row). As expected, the cognate LtrB relaxase from *L. lactis* supported conjugative transfer of the mobilizable plasmid very efficiently at 10^6-fold over background levels. However, the PcfG relaxase from *E. faecalis* was not able to support the transfer of the *L. lactis oriT*-containing plasmid, showing a conjugation efficiency at background levels. However, when we tested both permutations of chimeric relaxases, we found that LtrBE1-PcfGE2 (88-fold) and PcfGE1-LtrBE2 (791-fold) each supported conjugation efficiencies at low levels but nevertheless clearly above background. These data demonstrate that both chimeric relaxase enzymes are translated, fold appropriately and can actively support the transfer of an *L. lactis*-containing *oriT* plasmid, albeit at lower levels than the cognate wild-type LtrB relaxase.

In contrast, none of the four relaxases were able to support the transfer of a plasmid containing an *oriT* from the *E. faecalis* pTEF4 plasmid (LaRoche-Johnston et al. 2016), including its cognate wild-type PcfG relaxase (Fig. 4.2B, second row). These data suggest a functional uncoupling between the *oriT*-PcfG relaxase unit from *E. faecalis* and the lactococcal conjugation machinery (*ltr*-genes) encoded by the *L. lactis* chromosomal sex factor. These two functional units are most likely unable to interact and successfully mediate conjugation of the mobilizable plasmid.

We next assessed the conjugation efficiency of a third mobilizable plasmid that harbored both the *E. faecalis oriT* and *pcfF* accessory gene (Fig. 4.2B, third row). PcfF plays an essential role in *E. faecalis* conjugation, since its deletion results in the complete shutdown of conjugation (Chen et al. 2007). This MobC-family accessory protein is believed to act the same way as the *L. lactis* orthologous LtrF protein, by binding to the palindromic sequence directly adjacent to the *oriT* and recruiting the PcfG relaxase to initiate conjugation (Staddon et al. 2006). This construct revealed a ~ 400-fold increase in conjugation above background level for the wild-type PcfG relaxase which under biological conditions interacts with the *E. faecalis oriT*-PcfF complex (Fig. 4.2B, third row). A smaller ~ 18-fold increase in conjugation was also detected for the wild-type LtrB relaxase suggesting that there is a limited ability of the lactococcal relaxase to recognize and interact with the *E. faecalis oriT*-PcfF complex. However, the PcfGE1-LtrBE2 chimeric relaxase showed the largest significant increase in conjugation over background (~ 9420-fold) as well as a significant increase when compared to both wild-type LtrB (~ 530-fold) and PcfG (~ 24-fold)

enzymes. In sharp contrast, the LtrBE1-PcfGE2 chimeric relaxase was not able to support any level of conjugation.

Taken together, our results demonstrate that chimeric relaxase enzymes are active *in vivo* and can lead to a gain-of-function phenotype when E1 and E2 are associated respectively to their cognate *oriT* and conjugation machinery. These results also suggest that the minimal components required by a chimeric relaxase enzyme to support the conjugative transfer of a mobilizable plasmid are its cognate *oriT* and MobC-family accessory protein.

4.4.3: An Ef.PcfG-generated chimeric relaxase enzyme supports conjugation in *L. lactis*

We next engineered a conjugation assay in *L. lactis* to determine whether chimeric relaxase enzymes produced as a result of group II intron catalysis could also be active and sustain conjugation (Fig. 4.3). Since the highest conjugation efficiency was detected when both the *oriT* and *pcfF* from *E. faecalis* were coupled with the PcfGE1-LtrBE2 chimeric relaxase (Fig. 4.2B, third row), we co-transformed a pLE plasmid containing the *E. faecalis oriT*, the *pcfF* accessory protein and the Ef.PcfG-interrupted *pcfG* relaxase with a pDL-based plasmid expressing the non-interrupted *ltrB* gene (Fig. 4.3A). To reduce background levels of conjugation stemming from both wild-type relaxases, we introduced in-frame deletions of 360 nts in *ltrB*E1 and 936 nts in *pcfG*E2 (Fig. 4.3, red boxes). By inactivating both wild-type relaxases, the only remaining way for the pLE plasmid to be transferred by conjugation from NZ9800Δ*ltrB* to LM0231 is by the generation of an intergenic PcfGE1-LtrBE2 relaxase chimera that can functionally bridge the gap between the *E. faecalis oriT* and the *L. lactis* conjugation machinery.

To validate our system, we first performed an RT-PCR to look for the *pcfG*E1-*ltrB*E2 mRNA chimera. An amplicon was exclusively generated when the *pcfG* gene was interrupted by its Ef.PcfG-WT intron (Fig. 4.3B). When we tested our system functionally, we observed a relatively high conjugation efficiency (3.42×10^{-6}) in the presence of Ef.PcfG-WT, which represented a slight 12-fold decrease from when the chimeric non-interrupted relaxase gene was directly expressed (4.07×10^{-5}) (Fig. 4.2B, third row). Importantly, the conjugation efficiency when the *pcfG* gene was interrupted by Ef.PcfG-WT had a significant 1,600-fold increase over the

branchpoint mutant control, Ef.PcfG-ΔA, where the intron was unable to generate chimeric relaxase mRNA (Fig. 4.3B).

Our data thus show that when group II introns interrupt relaxase genes, they can produce enough chimeric relaxase enzymes in donor cells to mediate significant levels of conjugation. These results also demonstrate that the Ef.PcfG group II intron from *E. faecalis* can, similarly to Ll.LtrB, generate chimeric relaxase mRNAs and active chimeric enzymes in *L. lactis*.

4.4.4: The formation of chimeric relaxase transcripts occurs under biologically relevant conditions in *E. faecalis*

Having shown with our expression vectors that group II introns can catalyze the formation of active chimeric relaxase enzymes in *L. lactis*, we next wanted to see if they could also be produced in biologically relevant conditions under natural expression systems. We chose *E. faecalis* as the bacterial host to co-transform pEF1071 from *E. faecalis* (Balla and Dicks 2005) and pLE12 from *L. lactis* (Mills et al. 1996), because they both harbor relaxase genes under the control of their natural promoters (Fig. 4.4A). pEF1071 harbors the non-interrupted *mobA* relaxase gene that was previously shown to be invaded by Ef.PcfG in *E. faecalis* (LaRoche-Johnston et al. 2016). pLE12 contains the Ll.LtrB-interrupted *ltrB* relaxase gene that stems from the pRS01 *L. lactis* plasmid (Mills et al. 1996). This plasmid was previously shown to transfer laterally from *L. lactis* to the JH2-2 lab strain of *E. faecalis* by conjugation, where it can efficiently replicate and produce active group II intron RNPs (Belhocine et al. 2004). These plasmids were co-transformed in JH2-2 and maintained using their natural origin of replication. Co-transformants contained both plasmids with no apparent additional bands that typically appear when mobility products are generated due to intron mobility into non-interrupted relaxase genes (Fig. 4.4A, bottom panel) (Cousineau et al. 1998; Belhocine et al. 2004). When intron mobility was analyzed by colony hybridization (Belhocine et al. 2004; Plante and Cousineau 2006), it was found to be very limited with on average 5% of *mobA* genes interrupted by Ll.LtrB (Fig. 4.4A). Nevertheless, this showed that the *ltrB* gene is expressed at low levels from its natural promoter in *E. faecalis*, producing relatively small amounts of active Ll.LtrB RNPs.

We next addressed qualitative aspects of the mechanism of chimera formation *in vivo* (Fig. 4.4B). We first found that the natural expression levels of both relaxase genes is sufficient to generate chimeric mRNAs, which is again dependent on the branching pathway since no chimeras are detected when *ltrB* is interrupted by the branchpoint mutant (Fig. 4.4B, top panel, lanes 3-4). Surprisingly, although our model predicted exclusively the production of *ltrB*E1-*mobA*E2 chimeras (Fig. 4.1C), we also detected the presence of counterpart chimeras, *mobA*E1-*ltrB*E2 (Fig. 4.4B, bottom panel, lanes 3-4). A potential explanation is that expression of the Ll.LtrB-interrupted *mobA* gene, resulting from the mobility of Ll.LtrB from pLE12 to pEF1071, leads to the production of free *mobA*E1 (Fig. 4.1C, step 1) and the unexpected *mobA*E1-*ltrB*E2 chimeras (Fig. 4.1C, step 5). We thus next modified our system to simulate a 100% mobility scenario, where both relaxase genes are fully interrupted by Ll.LtrB. We detected stronger RT-PCR amplicons for both mRNA chimeras, suggesting a positive correlation between intron invasion of a target site and mRNA chimera formation (Fig. 4.4B, both panels, lane 5). Finally, to determine whether mRNA chimera formation is a product of group II intron catalysis or some type of RNA and/or DNA recombination event, we modified our assay so that both introns lacked their small catalytic domains (Ll.LtrB-ΔDV) (Zhao and Pyle 2017). Ll.LtrB-ΔDV is completely inactive and unable to perform any type of splicing reaction. However, the small deletion left the bulk of the intron sequences intact (2459/2492 nts) while maintaining perfect sequence homology between both intron copies. Interestingly, neither type of chimeric mRNAs were detected when both relaxase genes are interrupted by Ll.LtrB-ΔDV (Fig. 4.4B, both panels, lane 6).

We next used our assays with one or two wild-type introns to quantitatively address mRNA chimera production by RT-qPCR. We began by assessing the relative amounts of various RNAs being produced for each construct, using the *ltrB* pre-mRNA as the reference. When only one intron is present, interrupting *ltrB* (Fig. 4.4C), we detected about 10 times more *ltrB* pre-mRNA than ligated exons, while the *ltrB*E1-*mobA*E2 and *mobA*E1-*ltrB*E2 chimeras were respectively 55 and 126 times less abundant than ligated exons. When the two relaxase genes were interrupted by Ll.LtrB (Fig. 4.4D), the proportion of mRNA chimeras appeared to increase. While splicing efficiency was similar with about 9 times more *ltrB* pre-mRNA than ligated exons, there were now respectively only about 4 and 9 times fewer *ltrB*E1-*mobA*E2 and *mobA*E1-*ltrB*E2 chimeras than ligated exons. We finally compared the two systems by analyzing the relative normalized expression of each target (Fig. 4.4E). We first found that, as expected, amounts of *ltrB* pre-mRNA

and ligated exons were not significantly different. However, the expression system with two interrupted genes (Fig. 4.4E, red bars) showed significant 19- and 22-fold increases in production of *ltrB*E1-*mobA*E2 and *mobA*E1-*ltrB*E2 chimeras respectively when compared to the expression system with a single intron (Fig. 4.4E, blue bars), again supporting a positive correlation between intron invasion of a target site and mRNA chimera formation.

Taken as a whole, these results show that chimeric transcripts are produced at detectable levels by Ll.LtrB when the two relaxase genes are present on biologically relevant vectors and expressed under the control of their natural promoters in *E. faecalis*. Furthermore, our results demonstrate that chimeric transcript formation is dependent on intron catalysis and increases when more target sites are occupied by group II introns.

4.4.5: Phylogenetic analyses unveil group II intron-generated chimeric relaxase genes

The Ll.LtrB intron from the lactococcal pRS01 plasmid has been a model system to study group II intron splicing, mobility, lateral transfer as well as evolution. However, at least 60 closely related full-length intron variants are present in different species, subspecies and strains of lactococci (Candales et al. 2012). These group II introns have over 95% identity to Ll.LtrB and they mostly interrupt orthologous relaxase genes at the exact same conserved position, suggesting a recent acquisition and dissemination into this group of lactic acid bacteria (LaRoche-Johnston et al. 2016). Since most of these introns interrupt relaxase genes, we wanted to study the phylogenetic relationship between the lactococcal relaxase genes that are interrupted by group II introns.

We first analysed the flanking exon sequences of each intron and found that 53/60 were interrupting genes that could be identified as coding for relaxase enzymes. All intron-containing relaxase genes were interrupted at the exact same position within the conserved catalytic histidine triad common to members of the IncP relaxase family (Pansegrau et al. 1994). To avoid redundancy, we further narrowed our analyses to exclude sequences that were identical on both sides of the intron insertion site (± 25nts), leaving only 16/53 relaxase genes. Phylogenetic trees were then generated using either the full-length genes (Fig. 4.5A), E1 (Fig. 4.5B), or E2 (Fig. 4.5C)

by Maximum Likelihood using PhyML (Guindon et al. 2010), with 1000 bootstraps and the *E. faecalis pcfG* relaxase gene from the pTEF4 plasmid as the outgroup.

We first noticed that even though the overall structure of the trees were very similar, the position of some sequences were changing drastically between trees. To further investigate the evolutionary relationships of all 16 relaxase genes, we performed BLASTn searches using the individual exons of each relaxase as an input sequence against the Nucleotide collection database for *Lactococcus lactis* (taxid: 1358). Our goal was to determine whether the relaxase genes in each cluster were distinct monophyletic groups and belonged to the same evolutionary lineage.

For each of the 16 E1 and E2 queries, the entire gene of the highest nucleotide identity exon match identified by BLASTn was aligned to the whole gene of the input exon, to determine whether nucleotide similarity between exons extended to the remainder of the gene. We found three genes which appeared to be made up of exons from different relaxase families: DmW198 (Fig. 4.5D), pAH82 (Fig. 4.5E) and pSK11P (Fig. 4.5F). All three cases have highest similarities with at least one relaxase gene that does not contain a group II intron (Fig. 4.5D-F, asterisks). Interestingly, nucleotide identity steeply drops off for both exons precisely at the intron insertion site, such that each gene is drastically different from the exons that were not specifically used as input sequences (Fig. 4.5D-F).

To increase the resolution of our initial trees, we generated new phylogenetic trees of E1 (Fig. 4.5G) and E2 (Fig. 4.5H) that included the 5 additional genes found by searching for closest exon matches (Fig. 4.5D-F, asterisks). If relaxase genes had evolved as monophyletic units, we would expect no significant changes between the makeup of phylogenetic trees made for E1 or E2. We found that although the newly added relaxase genes remained in the same distinct clusters of the trees regardless of the exon that was analyzed, the exons of the pAH82 and pSK11P relaxase genes clearly belonged to different clusters. The DmW198 relaxase gene remained in the α cluster for both exons, likely representing a chimera formed between more closely related relaxase genes whose chimeric nature would likely be more obvious if the tree had better resolution.

Overall, these data indicate that certain lactococci contain relaxase genes whose exons have different evolutionary origins, since both exons belong to distinct relaxase phylogenetic lineages. Furthermore, these chimeric relaxase genes were likely generated by group II intron-based exon

shuffling, since the point at which nucleotide homology shifts corresponds precisely to the site of group II intron insertion.

4.5: Discussion

We previously described how reversal of the group II intron branching pathway into ectopic mRNAs produces a population of intron-interrupted cellular transcripts (Fig. 4.1C, steps 3-4). When used as templates for circularization instead of branching, these intron-interrupted mRNAs were shown to generate a population of E1-mRNA and mRNA-mRNA chimeric transcripts (Fig. 4.1C, steps 5-6) (LaRoche-Johnston et al. 2018). We thus proposed that this new group II intron intergenic *trans*-splicing pathway increases genetic diversity at the RNA level by shuffling coding sequences with potential benefits to the host cell. However, it remained unclear whether these chimeric transcripts are recognized by ribosomes and translated into chimeric proteins in sufficient amounts to yield any observable phenotype.

In this study, we took advantage of the native biological context of the Ll.LtrB and Ef.PcfG group II introns from *L. lactis* and *E. faecalis* to address important features of the intergenic *trans*-splicing pathway. Ll.LtrB and Ef.PcfG interrupt respectively the *ltrB* and *pcfG* orthologous relaxase genes at the same position between two codons (Fig. 4.1A), potentially allowing for in-frame exon shuffling between their mRNAs through intergenic *trans*-splicing (Fig. 4.1C) (LaRoche-Johnston et al. 2016).

We first demonstrated that Ll.LtrB can generate *ltrB*E1-*pcfG*E2 chimeric trancripts between the interrupted *ltrB* and non-interrupted *pcfG* relaxase mRNAs (Fig. 4.1D) and that these mRNA-mRNA chimeras are recognized by the translation machinery leading to the production of detectable amounts of LtrBE1-PcfGE2 chimeric proteins (Fig. 4.1E) in *L. lactis*.

Next, we showed that chimeric relaxase enzymes between LtrB and PcfG are active in *L. lactis* and can even confer a gain-of-function phenotype when compared to their precursor wild-type enzymes (Fig. 4.2). Previous *in vitro* work on the specificity determinants of the *L. lactis* and *E. faecalis* conjugative systems had shown that the LtrF accessory protein from *L. lactis* could functionally substitute the *E. faecalis* PcfF protein in recognition of the *E. faecalis oriT* and

recruitment of the PcfG relaxase (Chen et al. 2007). However, we found the conjugation efficiency of the PcfG relaxase in our lactococcal system, where LtrF is expressed from the chromosome, to be at background levels (Fig. 4.2B, second row). When the cognate PcfF accessory protein was provided in *L. lactis*, the transfer efficiencies supported by certain relaxases increased (Fig. 4.2B, third row). Interestingly, even though we detected small increases in conjugation efficiency for the two wild-type relaxase enzymes, the efficiency of the PcfGE1-LtrBE2 relaxase was significantly higher while its counterpart LtrBE1-PcfGE2 was completely inactive. The ability of the PcfGE1-LtrBE2 chimeric relaxase to considerably outperform both wild-type relaxases likely stems from the architecture of these enzymes, which are generally composed of two domains. The N-terminal domain, corresponding to E1, contains 3 distinct motifs which are believed to act in concert to bind and nick ssDNA at the *oriT* and to form a covalent bond between the liberated 5′ phosphate of the ssDNA and a highly conserved tyrosine residue (Byrd and Matson 1997). This was supported by functional assays showing that relaxase enzymes with truncated C-terminal ends were sufficient to recognize and nick their cognate *oriT*, and yet were unable to complete conjugative transfer through the mating pore (van Kregten et al. 2009; Cascales et al. 2013). The larger C-terminal domain, corresponding to E2, is less well characterized, having very little sequence conservation (Pansegrau et al. 1994). However, it plays an essential role during conjugation, and has recently been shown to bind the all-alpha domain (AAD) of type 4 coupling proteins (T4CPs), thus conferring specificity to distinct type IV secretion systems (T4SSs) (Whitaker et al. 2015). Taken together, the separation of functional domains of relaxase enzymes thus support a model where E1 of a chimeric relaxase recognizes, binds and nicks its cognate *oriT*, in conjunction with its MobC-family accessory protein; while the function of E2 is to provide specificity to the larger T4SS through interactions with the T4CPs. This molecular architecture is consistent our conjugation data, which showed that the best suited relaxase to interact with the *E. faecalis oriT* and *L. lactis* T4CP, PcfGE1-LtrBE2, indeed gave the highest conjugation efficiency. Conversely, the worst-suited relaxase to interact with its binding partners is expected to be the complement chimeric relaxase, LtrBE1-PcfGE2, which was accordingly at background levels and the least efficient of all relaxases tested.

We then determined that a chimeric relaxase enzyme, produced through intergenic *trans*-splicing *in vivo*, is abundant enough to yield an observable phenotype. We used an *L. lactis* conjugation assay where the two precursor relaxase genes harbored large deletions in either E1 or

E2, such that conjugation is only detectable when the two functional exons of each relaxase gene are shuffled together (Fig. 4.3A). When the *pcfG* gene was interrupted by Ef.PcfG-WT we observed high conjugation efficiency, which completely disappeared when the intron was mutated to lack the branchpoint adenosine (Fig. 4.3B). The ability of chimeric relaxases, produced in small amounts when compared to contiguous relaxases, to produce a gain-of-function phenotype may be due to the fact that only a limited amount of relaxase enzyme is necessary to successfully mediate conjugation (Chen et al. 2005; Belhocine et al. 2007a). Our data also showed that Ef.PcfG from *E. faecalis* can induce exon shuffling between the *ltrB* and *pcfG* mRNAs in *L. lactis* through intergenic *trans*-splicing similarly to Ll.LtrB.

We subsequently demonstrated that Ll.LtrB can generate mRNA chimeras under biologically-relevant conditions in *E. faecalis*. Despite the fact that the *ltrB* and *mobA* relaxase genes were expressed from their natural promoters, we were able to detect *ltrB*E1-*mobA*E2 chimeras by RT-PCR. In accordance with our previous results, this amplicon was absent when *ltrB* was interrupted by the Ll.LtrB-ΔA branchpoint mutant (LaRoche-Johnston et al. 2018). Unexpectedly, we also detected *mobA*E1-*ltrB*E2 chimeras again exclusively for Ll.LtrB-WT. Our model (Fig. 4.1C) predicted a clear directionality for the intergenic *trans*-splicing pathway, which would favor the sole production of *ltrB*E1-*mobA*E2 chimeras. Expression of the Ll.LtrB-interrupted *mobA* gene from mobility products in pEF1071 seem to contribute to the generation of *mobA*E1-*ltrB*E2 through the production of free *mobA*E1. However, an alternative explanation is that the intron may interact with the non-interrupted *mobA* transcripts in other ways than reverse splicing. For instance, Ll.LtrB lariats may be hydrolyzing the contiguous exons of non-interrupted *mobA* transcripts by the spliced exon reopening (SER) reaction, also leading to the release of free *mobA*E1 (Qu et al. 2018).

We next determined the relative abundance of both *ltrB*E1-*mobA*E2 and *mobA*E1-*ltrB*E2 compared to *ltrB*E1-*ltrB*E2 in contexts where *ltrB* (Fig. 4.4C) or *ltrB* and *mobA* (Fig. 4.4D) are interrupted by Ll.LtrB. Despite the fact that internal comparisons could be slightly biased due to the use of different primer pairs for each target (Yuan et al. 2006), chimeric mRNA formation was much higher than we expected. In a biologically relevant context where only *ltrB* is fully interrupted, in the presence of small amounts of interrupted *mobA* (Ll.LtrB mobility products) (~ 5%) (Fig. 4.4C), both types of chimeras were produced at a proportion of about 1-2% compared

to *ltrB* ligated exons (1.8% for *ltrB*E1-*mobA*E2 and 0.8% for *mobA*E1-*ltrB*E2). When both relaxase genes were fully interrupted by Ll.LtrB (Fig. 4.4D), this ratio increased by approximately 14-fold (25% for *ltrB*E1-*mobA*E2 and 11% for *mobA*E1-*ltrB*E2). Our findings are further supported by the relative normalized expression analysis between the two systems. When both relaxase genes were interrupted by Ll.LtrB, the two types of chimeras increased by about 20-fold (19-fold for *ltrB*E1-*mobA*E2 and 22-fold for *mobA*E1-*ltrB*E2) compared to when only *ltrB* was interrupted (Fig. 4.2E).

The copy number of group II introns is notoriously low within the chromosomes of individual bacteria (Lambowitz and Zimmerly 2011). In addition, group II introns were previously shown to generate recombination events in bacteria when multiple copies were present within the genome (Leclercq et al. 2011). However, using a *trans*-splicing assay where both relaxase genes are interrupted by catalytically inactive introns (Ll.LtrB-ΔDV) we demonstrated that the mRNA chimeras observed are exclusively produced by intron catalysis and not generated through some type of RNA and/or DNA homologous recombination event (Fig. 4.4B).

On the other hand, by making phylogenetic trees outlining the evolutionary relationships of lactococcal relaxase genes interrupted by a group II intron, we found three genes where the two exons belonged to different evolutionary lineages: DmW198, pAH82 and pSK11P. These chimeric relaxase genes may have been generated at the DNA level through homologous recombination since they are interrupted by almost identical group II introns at the exact same position. Indeed, group II introns and other mobile elements in bacteria such as IS elements were previously shown to generate large-scale modifications in bacterial genomes through processes such as recombination (Leclercq et al. 2011). The distinguishing characteristic of chimeric genes generated by group II intron-mediated recombination is their ability to still be expressed as chimeras, due to intron self-splicing at the RNA level, which may limit the potential damage brought on by recombination. If almost identical introns, occupying conserved sites in homologous or orthologous genes, mediate recombination events, the interrupted gene becomes chimeric but may still yield a functional product once the intervening intron splices and ligates its shuffled exons. Alternatively, we cannot completely rule out the possibility that chimeric relaxase genes may have been generated by the reverse transcription of group II intron-generated chimeric mRNAs followed by the fixation of these cDNAs in *L. lactis* genomes and/or plasmids.

The great majority of bacterial conjugative elements harbor at least the basic components to produce their own relaxosomes consisting of an *oriT* and *oriT*-specific relaxase and accessory protein (Smillie et al. 2010). Upon arrival in a new host, a newly transferred non-autonomous mobilizable plasmid is thus in a conjugative cul-de-sac if its relaxase is not recognized by the resident conjugation machinery (Fig. 4.6, scenario 1) or if the relaxase of the resident conjugative element cannot recognize its *oriT* (Fig. 4.6, scenario 4). However, our data suggest that if at least one of the two relaxase genes, encoded either on the acquired or resident conjugative element, is interrupted by a group II intron, then two chimeric relaxase enzymes with shuffled exons can be generated by intergenic *trans*-splicing (Fig. 4.6, scenarios 2 and 3). One of the two chimeric relaxase enzymes, harboring E1 and E2 specific for the *oriT* of the mobilizable plasmid and the T4CP encoded by the resident conjugative element, respectively, could bridge the functional gap between the DNA transfer and replication (Dtr) proteins (relaxases and accessory proteins) and the mating pore formation (Mpf) proteins (T4CPs and T4SSs) (Fig. 4.6, scenario 2) of incompatible conjugative systems. This would allow the transfer of the conjugative element even if its own relaxase is unable to do so. Our work also indicates that if the two relaxase genes are interrupted by homologous group II introns, more chimeric relaxases should be produced, in turn supporting higher levels of conjugation. Of importance, the presence of a group II intron-interrupted relaxase gene within the conjugative element of a bacterial host, like for example the *L. lactis* chromosomal sex factor, should provide a link to its conjugation machinery and stimulate the conjugative transfer of all acquired non-autonomous mobilizable elements.

The ability of the Ll.LtrB group II intron to produce functional chimeric relaxases *in vivo* increases conjugative efficiency, which in turn can affect the relationship of these mobile elements with their bacterial hosts. Group II introns were long thought of as parasitic elements that were solely subjected to negative selection by their bacterial hosts (Leclercq and Cordaux 2012). In the specific case of Ll.LtrB, the interrupted *ltrB* gene was shown to be translated at lower levels than the non-interrupted *ltrB* gene (Chen et al. 2005), and even to be targeted for degradation by the group II intron (Qu et al. 2018). However, we show here that group II introns such as Ll.LtrB and Ef.PcfG can in fact be beneficial to their host conjugative elements by catalyzing the formation of chimeric relaxase enzymes that increase their potential of dispersal by conjugation. This appears especially relevant in the biological context of *L. lactis* where rapid adaptation to the dairy environment largely occurred by shrinking of the bacterial chromosome through reductive

evolution and a concurrent drastic increase in plasmid content, most of which were acquired by conjugation (Cavanagh et al. 2015). For *L. lactis* dairy strains, it is thus likely that the acquisition of new plasmids would have been positively selected for, which may account for the recent dispersal of Ll.LtrB variants that has taken place within different species, sub-species and strains of dairy lactococci (LaRoche-Johnston et al. 2016).

Taken together, our data show that bacterial group II introns can generate active chimeric relaxase enzymes by shuffling coding sequences at both the RNA level by intergenic *trans*-splicing and at the DNA level, most likely by homologous recombination. This is the first demonstration that group II introns can be beneficial to the conjugative elements that harbor them and to their bacterial host cells. Although mobilizing into a new DNA site may be seen purely in terms of intron spreading and survival, the ability to increase genetic diversity by generating a new population of mRNA chimeras as well as chimeric genes that could be beneficial to the host cell may be another factor that positively selects for mobility events and eventually fixes them in a population. The specific benefit of the Ll.LtrB variants in lactococci is illustrated by their positive selection in *L. lactis*, which enabled their recent dissemination in the highly conserved sites of several relaxase genes following the lateral transfer of Ef.PcfG from *E. faecalis* to *L. lactis*.

Although the work presented here defies the paradigm that bacterial group II introns provide no benefits to their hosts, it is nevertheless compatible with the fact that these retroelements behave selfishly in order to spread and survive within bacterial cells. Our work thus provides an interesting case study to describe how beneficial outcomes can arise from selfish behavior throughout the course of evolution.

4.6: Experimental procedures

4.6.1: Bacterial strains and plasmids

Enterococcus faecalis lab strain JH2-2 was grown in BHI media at 37°C without shaking. *Lactococcus lactis* strains NZ9800Δ*ltrB* (Tet^R) (Ichiyanagi et al. 2002) and LM0231 were grown in M17 media supplemented with 0.5% glucose at 30°C without shaking. *Escherichia coli* strains DH10β and TransforMax™ EC100D™ pir^+ were grown in LB media at 37°C with shaking.

Antibiotics were used at the following concentrations: chloramphenicol (Cam^R), 10 µg/ml; spectinomycin (Spc^R), 300 µg/ml; fusidic acid (Fus^R), 25 µg/ml. Plasmids used in this study are listed in Table S4.1. The construction of some plasmids was previously described (pLE-P_{Nis}-*ltrB* (LaRoche-Johnston et al. 2016) and pLE12 (Mills et al. 1996)). The pEF1071 plasmid was isolated from a clinical strain of *E. faecalis* (SF24397) (McBride et al. 2007). Since this clinical strain is difficult to work with, we performed a Tn5 transposition assay to insert a gene conferring resistance to spectinomycin for selection and the R6Kγori *E. coli* origin of replication (Fig. 4.4A), generating the pEF1071::<R6Kγori/Spc^R> plasmid. Other plasmids were constructed by restriction-digestion/ligation reactions (pDL-P_{23}-*pcfG*, pDL-P_{23}-*ltrB*, pLE-*oriT-L. lactis*, pLE-*oriT-E. faecalis*, pLE-*oriT-E. faecalis*-P_{23}-*pcfF-pcfG*-E2Δ936. The pEF1071::<R6Kγori/Spc^R>-Ll.LtrB plasmid was obtained through the invasion of *mobA* by Ll.LtrB *in vivo*. The following plasmids were obtained by using the NEB Site-Directed Mutagenesis kit (pLE12-ΔA, pLE12-ΔDV, pEF1071::<R6Kγori/SpcR>-Ll.LtrB-ΔDV, pDL-P_{23}-*ltrB*-E1Δ360, pLE-P_{23}-*pcfF-pcfG*-Ef.PcfGΔA-E2Δ936, pDL-P_{23}-*pcfG*-E1Δ366-ΔEf.PcfG-His). Primers used for site-directed mutagenesis and cloning are shown in Table S4.2.

4.6.2: RNA extraction, RT-PCR and RT-qPCR

L. lactis cultures were induced with nisin when required and RNA extractions were done on various *L. lactis* and *E. faecalis* strains as previously described (Belhocine et al. 2007a). RT-PCRs for the detection of chimeric mRNAs produced in *L. lactis* and *E. faecalis* was done using branchpoint mutant controls and stringent amplification conditions, as previously described (LaRoche-Johnston et al. 2020). RT-qPCR reactions were done by treating total RNA extracts (10µg/sample) with RNase-free DNase I (New England Biolabs) for 1 hour at 37°C. RNA was then recovered from the reaction (RNeasy Mini Kit, Quiagen). RT reactions were then performed as previously described (Belhocine et al. 2007a), using an annealing temperature of 50°C and 3 RT primers for every target to be analyzed (all primers in Table S4.2). cDNA was then loaded onto a 96-well PCR plate (Progene®), where a qPCR was done using a SYBR-green fluorescent dye (abm) in a qPCR plate reader (Bio-Rad). Results were analyzed using Bio-Rad CFX Manager™. Each data point represents the average of technical duplicates done for biological triplicates. No-

template controls and no RT controls were also added for each target. The *E. faecalis* gene Lactate Dehydrogenase B (*ldhB*) was used as a reference gene across samples for normalization.

4.6.3: Protein extractions and Western blotting

L. lactis cultures were induced with nisin and whole protein extractions were performed as previously described (Hugentobler et al. 2012). Raw protein extracts were run on SDS-PAGE (8%), then transferred on a PVDF membrane for Western blotting as previously described (Hugentobler et al. 2012). 6X-His Tag Monoclonal Antibody (HIS.H8) from Thermo Fisher Scientific (MA1-21315) was used as a primary antibody (1:3000). Goat anti-Mouse IgG (H+L) Cross-Adsorbed Secondary Antibody, conjugated to Horseradish Peroxidase from Thermo Fisher Scientific (G-21040) was used as a secondary antibody (1:5000).

4.6.4: Mobility and conjugation assays

Intron mobility efficiency was calculated by Patch Hybridization as previously described (Plante and Cousineau 2006), using TransforMax^{TM} EC100D^{TM} *pir*$^{+}$ *E. coli* cells. Each data point represents triplicates of independent transformants, along with the standard error. Statistical significance was determined using a Student's unpaired T-test ($p<0.05$). Conjugation assays were done on milk plates between *L. lactis* donor strain NZ9800Δ*ltrB* and *L. lactis* recipient strain LM0231, as previously described (Belhocine et al. 2004). Each conjugative assay was performed in triplicates, and statistical significance was determined using a Student's unpaired T-test ($p<0.05$).

4.6.5: Phylogenetic trees

Input sequences were aligned using Clustal Omega (Sievers et al. 2011). Output files in Philip format were then used to generate maximum likelihood trees in PhyML (Guindon et al. 2010), using nearest neighbor interchange and 1000 bootstraps. The trees were then visualized

using the interactive tree of life (iTOL) software (Letunic and Bork 2011). Matrices and phylogenetic trees were uploaded and made available in the TreeBASE online repository (URL: http://purl.org/phylo/treebase/phylows/study/TB2:S27000).

4.7: Acknowledgments

This work was supported by a discovery grant from the Natural Sciences and Engineering Research Council of Canada to B.C. (227826). F.L.J. received a Graduate Excellence Fellowship from McGill University, a CGS-M Scholarship from Natural Sciences and Engineering Research Council of Canada and a Master's Research Scholarship from Fonds de Recherche en Nature et Technologies; and currently holds an Alexander Graham Bell CGS-D Scholarship from Natural Sciences and Engineering Research Council of Canada.

4.8: References

Balla E, Dicks LM. 2005. Molecular analysis of the gene cluster involved in the production and secretion of enterocins 1071A and 1071B and of the genes responsible for the replication and transfer of plasmid pEF1071. *Int J Food Microbiol* 99(1):33-45.

Belhocine K, Plante I, Cousineau B. 2004. Conjugation mediates transfer of the Ll.LtrB group II intron between different bacterial species. *Mol Microbiol* 51(5):1459-69.

Belhocine K, Yam KK, Cousineau B. 2005. Conjugative transfer of the Lactococcus lactis chromosomal sex factor promotes dissemination of the Ll.LtrB group II intron. *J Bacteriol* 187(3):930-9.

Belhocine K, Mak AB, Cousineau B. 2007a. Trans-splicing of the Ll.LtrB group II intron in Lactococcus lactis. *Nucleic Acids Res* 35(7):2257-68.

Belhocine K, Mandilaras V, Yeung B, Cousineau B. 2007b. Conjugative transfer of the Lactococcus lactis sex factor and pRS01 plasmid to Enterococcus faecalis. *FEMS Microbiol Lett* 269(2):289-94.

Bush SJ, Chen L, Tovar-Corona JM, Urrutia AO. 2017. Alternative splicing and the evolution of phenotypic novelty. *Philos Trans R Soc Lond B Biol Sci* 372(1713).

Byrd DR, Matson SW. 1997. Nicking by transesterification: the reaction catalysed by a relaxase. *Mol Microbiol* 25(6):1011-22.

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

Cascales E, Atmakuri K, Sarkar MK, Christie PJ. 2013. DNA substrate-induced activation of the Agrobacterium VirB/VirD4 type IV secretion system. *J Bacteriol* 195(11):2691-704.

Cavanagh D, Fitzgerald GF, McAuliffe O. 2015. From field to fermentation: the origins of Lactococcus lactis and its domestication to the dairy environment. *Food Microbiol* 47:45-61.

Chen Y, Klein JR, McKay LL, Dunny GM. 2005. Quantitative analysis of group II intron expression and splicing in Lactococcus lactis. *Appl Environ Microbiol* 71(5):2576-86.

Chen Y, Staddon JH, Dunny GM. 2007. Specificity determinants of conjugative DNA processing in the Enterococcus faecalis plasmid pCF10 and the Lactococcus lactis plasmid pRS01. *Mol Microbiol* 63(5):1549-64.

Chen Y, Zhang X, Manias D, Yeo HJ, Dunny GM, Christie PJ. 2008. Enterococcus faecalis PcfC, a spatially localized substrate receptor for type IV secretion of the pCF10 transfer intermediate. *J Bacteriol* 190(10):3632-45.

Cousineau B, Smith D, Lawrence-Cavanagh S, Mueller JE, Yang J, Mills D, Manias D, Dunny G, Lambowitz AM, Belfort M. 1998. Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell* 94(4):451-62.

Cousineau B, Lawrence S, Smith D, Belfort M. 2000. Retrotransposition of a bacterial group II intron. *Nature* 404(6781):1018-21.

Dai L, Zimmerly S. 2002. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res* 30(5):1091-102.

Fedorova O, Zingler N. 2007. Group II introns: structure, folding and splicing mechanism. *Biol Chem* 388(7):665-78.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307-21.

Hugentobler F, Yam KK, Gillard J, Mahbuba R, Olivier M, Cousineau B. 2012. Immunization against Leishmania major infection using LACK- and IL-12-expressing Lactococcus lactis induces delay in footpad swelling. *PLoS One* 7(2):e30945.

Ichiyanagi K, Beauregard A, Lawrence S, Smith D, Cousineau B, Belfort M. 2002. Retrotransposition of the Ll.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. *Mol Microbiol* 46(5):1259-72.

Irimia M, Roy SW. 2014. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb Perspect Biol* 6(6).

Klein JR, Chen Y, Manias DA, Zhuo J, Zhou L, Peebles CL, Dunny GM. 2004. A conjugation-based system for genetic analysis of group II intron splicing in Lactococcus lactis. *J Bacteriol* 186(7):1991-8.

Lambowitz AM, Zimmerly S. 2011. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol* 3(8):a003616.

Lambowitz AM, Belfort M. 2015. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol Spectr* 3(1).

LaRoche-Johnston F, Monat C, Cousineau B. 2016. Recent horizontal transfer, functional adaptation and dissemination of a bacterial group II intron. *BMC Evol Biol* 16(1):223.

LaRoche-Johnston F, Monat C, Coulombe S, Cousineau B. 2018. Bacterial group II introns generate genetic diversity by circularization and trans-splicing from a population of intron-invaded mRNAs. *PLoS Genet* 14(11):e1007792.

LaRoche-Johnston F, Monat C, Cousineau B. 2020. Detection of Group II Intron-Generated Chimeric mRNAs in Bacterial Cells. *Methods Mol Biol* 2079:95-107.

Leclercq S, Giraud I, Cordaux R. 2011. Remarkable abundance and evolution of mobile group II introns in Wolbachia bacterial endosymbionts. *Mol Biol Evol* 28(1):685-97.

Leclercq S, Cordaux R. 2012. Selection-driven extinction dynamics for group II introns in Enterobacteriales. *PLoS One* 7(12):e52268.

Letunic I, Bork P. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* 39(Web Server issue):W475-8.

McBride SM, Fischetti VA, Leblanc DJ, Moellering RC, Jr., Gilmore MS. 2007. Genetic diversity among Enterococcus faecalis. *PLoS One* 2(7):e582.

McNeil BA, Semper C, Zimmerly S. 2016. Group II introns: versatile ribozymes and retroelements. *Wiley Interdiscip Rev RNA*.

Mills DA, McKay LL, Dunny GM. 1996. Splicing of a group II intron involved in the conjugative transfer of pRS01 in lactococci. *J Bacteriol* 178(12):3531-8.

Monat C, Quiroga C, Laroche-Johnston F, Cousineau B. 2015. The Ll.LtrB intron from Lactococcus lactis excises as circles in vivo: insights into the group II intron circularization pathway. *RNA* 21(7):1286-93.

Monat C, Cousineau B. 2016. Circularization pathway of a bacterial group II intron. *Nucleic Acids Res* 44(4):1845-53.

Novikova O, Smith D, Hahn I, Beauregard A, Belfort M. 2014. Interaction between conjugative and retrotransposable elements in horizontal gene transfer. *PLoS Genet* 10(12):e1004853.

Pansegrau W, Schroder W, Lanka E. 1994. Concerted action of three distinct domains in the DNA cleaving-joining reaction catalyzed by relaxase (TraI) of conjugative plasmid RP4. *J Biol Chem* 269(4):2782-9.

Plante I, Cousineau B. 2006. Restriction for gene insertion within the Lactococcus lactis Ll.LtrB group II intron. *RNA* 12(11):1980-92.

Pyle AM. 2016. Group II Intron Self-Splicing. *Annu Rev Biophys* 45:183-205.

Qu G, Piazza CL, Smith D, Belfort M. 2018. Group II intron inhibits conjugative relaxase expression in bacteria by mRNA targeting. *Elife* 7.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J et al. . 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539.

Smillie C, Garcillan-Barcia MP, Francia MV, Rocha EP, de la Cruz F. 2010. Mobility of plasmids. *Microbiol Mol Biol Rev* 74(3):434-52.

Staddon JH, Bryan EM, Manias DA, Dunny GM. 2004. Conserved target for group II intron insertion in relaxase genes of conjugative elements of gram-positive bacteria. *J Bacteriol* 186(8):2393-401.

Staddon JH, Bryan EM, Manias DA, Chen Y, Dunny GM. 2006. Genetic characterization of the conjugative DNA processing system of enterococcal plasmid pCF10. *Plasmid* 56(2):102-11.

van Kregten M, Lindhout BI, Hooykaas PJ, van der Zaal BJ. 2009. Agrobacterium-mediated T-DNA transfer and integration by minimal VirD2 consisting of the relaxase domain and a type IV secretion system translocation signal. *Mol Plant Microbe Interact* 22(11):1356-65.

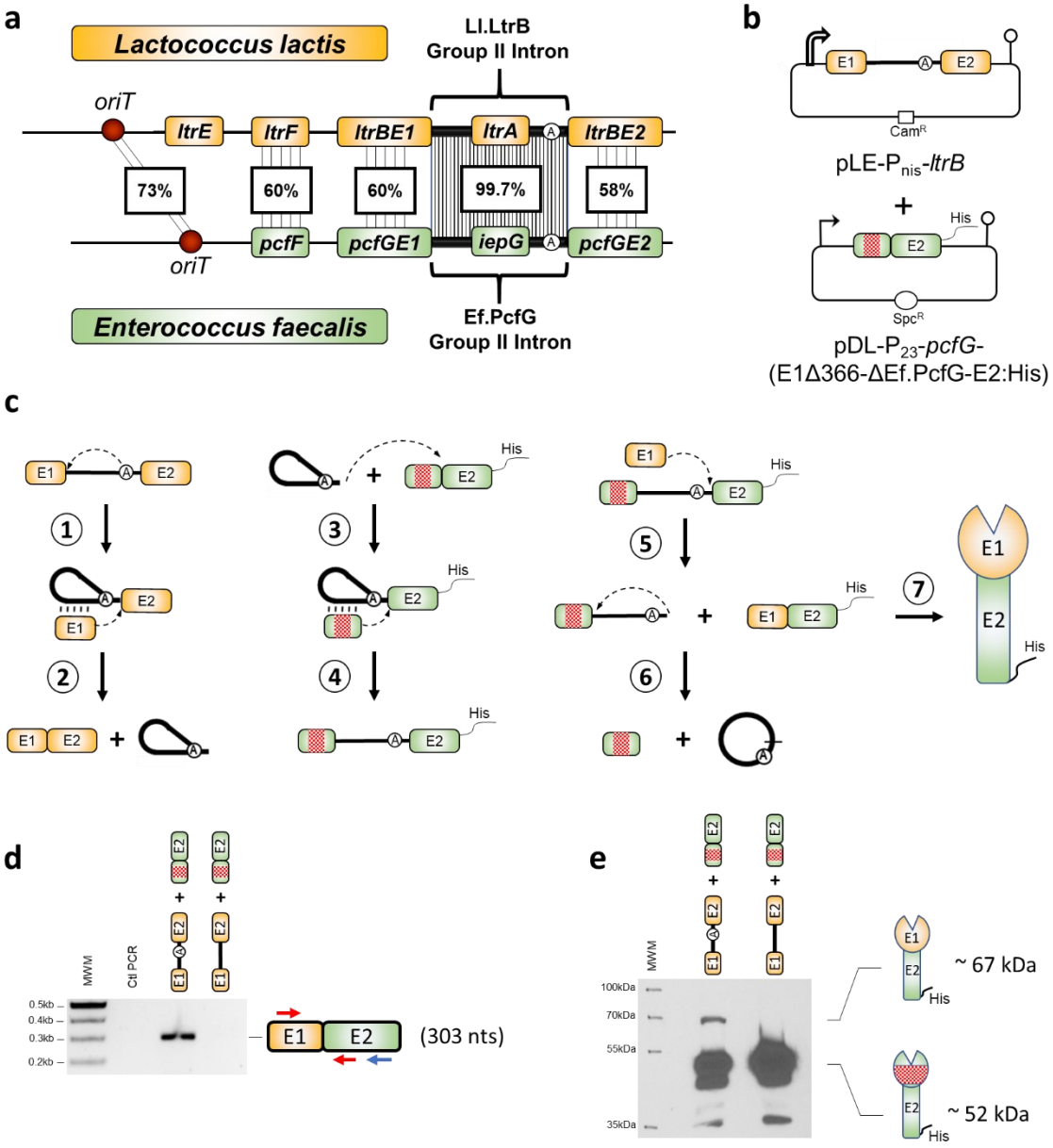
Whitaker N, Chen Y, Jakubowski SJ, Sarkar MK, Li F, Christie PJ. 2015. The All-Alpha Domains of Coupling Proteins from the Agrobacterium tumefaciens VirB/VirD4 and Enterococcus faecalis pCF10-Encoded Type IV Secretion Systems Confer Specificity to Binding of Cognate DNA Substrates. *J Bacteriol* 197(14):2335-49.

Yuan JS, Reed A, Chen F, Stewart CN, Jr. 2006. Statistical analysis of real-time PCR data. *BMC Bioinformatics* 7:85.

Zhao C, Pyle AM. 2017. Structural Insights into the Mechanism of Group II Intron Splicing. *Trends Biochem Sci* 42(6):470-482.

4.9: Figures

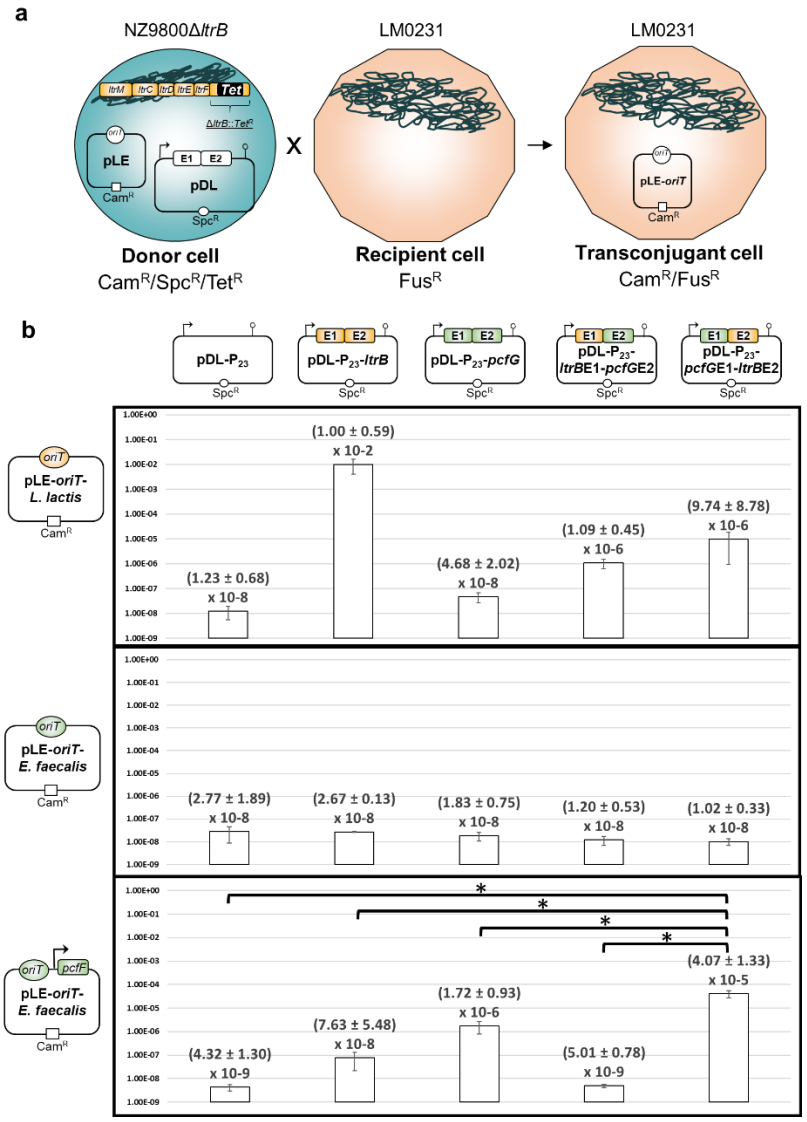
Figure 4.1



The Ll.LtrB group II intron generates chimeric relaxase mRNAs and proteins in *L. lactis*. (a) Comparison of the genetic loci involved in conjugative transfer of the *L. lactis* pRS01 plasmid (orange) harboring the Ll.LtrB group II intron and the *E. faecalis* pTEF4 plasmid (green) harboring the Ef.PcfG group II intron. Both introns interrupt a conserved catalytic motif at the exact same position in the *ltrB* and *pcfG* orthologous relaxase genes. Similarities between orthologous genes are shown as percent nucleotide identity. *oriT* (red circle): origin of conjugative transfer.

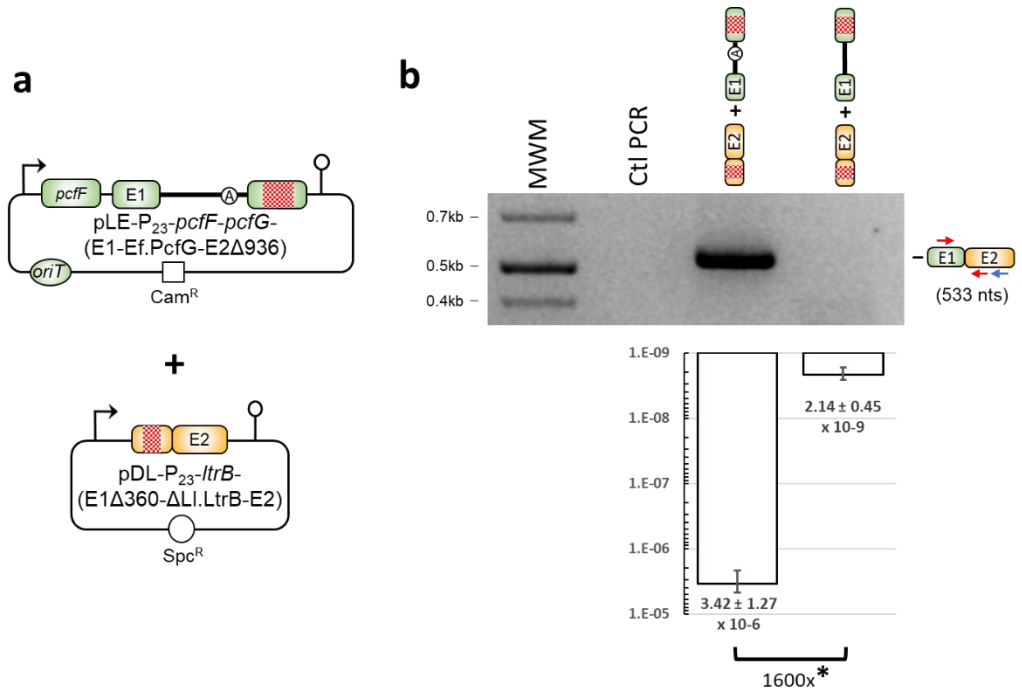
(b) Two-plasmid system (pLE and pDL) used to study the production of chimeric relaxase mRNAs and proteins in *L. lactis*. The *ltrB* relaxase gene (orange) is expressed from a nisin-inducible promoter (open broken arrow) and is interrupted by the Ll.LtrB group II intron while the *pcfG* relaxase gene (green) is expressed from a constitutive P_{23} promoter (broken arrow). The *pcfG* gene also harbors a C-terminal 6X His-tag and a 366 nt in-frame deletion in E1 (red square). **(c)** Group II intron intergenic *trans*-splicing pathway producing chimeric relaxases *in vivo* (LaRoche-Johnston et al. 2018). When Ll.LtrB excises through the branching pathway from the *ltrB* pre-mRNA (orange), the bulged adenosine's 2′ OH attacks the 5′ splice site, forming a branched lariat still attached to exon 2 (E2) while exon 1 (E1) remains associated to the intron solely through base pairing interactions (vertical lines) (Step 1). In the second step of branching, the 3′ OH of the last nt of released E1 attacks the 3′ splice site, ligating both exons and releasing the intron lariat (Step 2). Ll.LtrB intron lariats can base pair with a sequence coding for conserved catalytic residues in the non-interrupted orthologous *pcfG* mRNA (green) and invade it by complete reverse splicing (Steps 3, 4). Introns that interrupt the *pcfG* mRNA can self-splice using the circularization pathway by recruiting free *ltrB*E1 (orange) to attack the 3′ splice site, producing a chimeric relaxase mRNA (*ltrB*E1-*pcfG*E2, orange-green) marked with a 6X His-tag (Step 5). The 2′ OH of the intron's last nucleotide then attacks the 5′ splice site, generating a head-to-tail intron circle and free E1 (Step 6). The chimeric relaxase mRNA (orange-green) can be translated into a His-tagged chimeric relaxase enzyme (orange-green) (Step 7). The two-plasmid expression system shown in panel B was used to screen for the *in vivo* production of chimeric relaxases at the mRNA **(d)** and protein **(e)** levels where the *ltrB* gene was interrupted by either the Ll.LtrB-WT or Ll.LtrB-ΔA intron. Arrows denote the relative position of primers used for RT-PCR (blue for RT, red for PCR) and the expected size for mRNA chimeras containing a perfect *ltrB*E1-*pcfG*E2 junction is shown (303 nts). Expected sizes of translated chimeric (~ 67 kDa) and contiguous (~ 52 kDa) relaxase proteins are also shown.

Figure 4.2



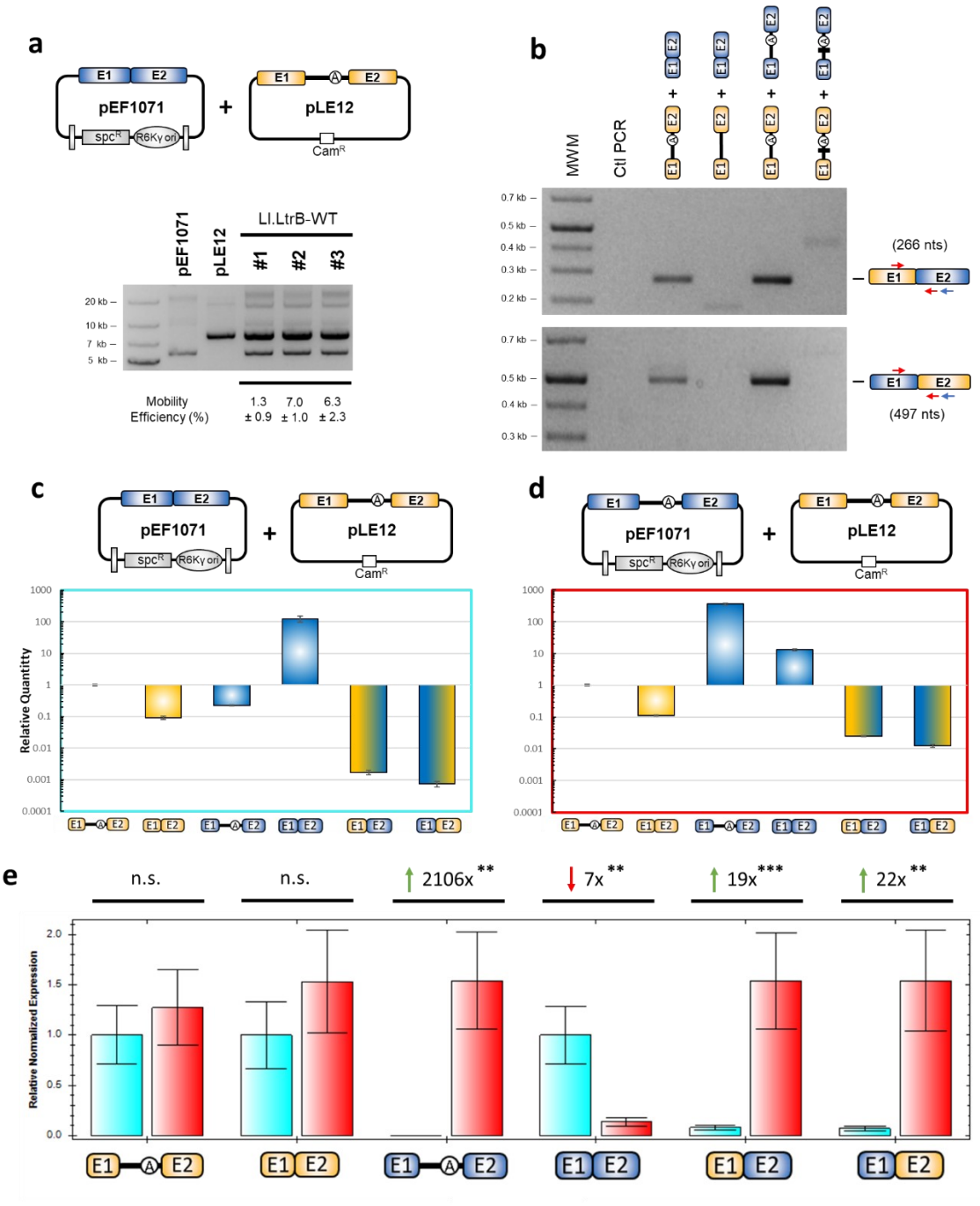
Conjugation efficiencies of plasmids harboring various *oriT*s in the presence of wild-type or chimeric relaxase enzymes. (a) Schematic representation of the conjugative assay. *L. lactis* donor cells (NZ9800Δ*ltrB*) encode the required lactococcal transfer machinery (*Ltr* genes, orange) to fully support conjugation, except for the *ltrB* relaxase gene which was replaced by a gene conferring resistance to tetracycline (Tet^R). Different relaxase genes were supplied in *trans* from a pDL-based plasmid (Spc^R) using a P_{23} constitutive promoter in the presence of a pLE-based plasmid (Cam^R) harboring either the *L. lactis* pRS01 (orange) or the *E. faecalis* pTEF4 (green) *oriT*. *L. lactis* recipient cells (LM0231) are resistant to fusidic acid (Fus^R), and transconjugant cells were isolated by selecting for recipient cells that received the *oriT*-containing plasmid (Cam^R/Fus^R). **(b)** Conjugation efficiencies for wild-type (LtrB, orange-orange and PcfG, green-green) and chimeric (LtrBE1-PcfGE2, orange-green and PcfGE1-LtrBE2, green-orange) relaxases. A pDL-based plasmid without any relaxase was used to determine background levels of conjugation. Conjugation efficiency was calculated by dividing the number of transconjugant cells by the number of donor cells, which is shown together with the standard error. Each data point represents triplicates of independent assays. Bent arrows denote the presence of a P_{23} constitutive promoter. Asterisks denote statistical significance (*: $p < 0.05$).

Figure 4.3



Ef.PcfG-generated chimeric relaxase enzyme supports conjugation in *L. lactis*. (a) Conjugation system used to study the production of chimeric PcfGE1-LtrBE2 relaxase enzyme *in vivo*. The pDL and pLE plasmids are co-transformed in the NZ9800Δ*ltrB* donor strain while the *L. lactis* strain LM0231 is used as the recipient. Red boxes denote in-frame deletions in *pcfG*E2 (936 nts) and *ltrB*E1 (360 nts). Broken arrows represent the P_{23} constitutive promoter. **(b)** The production of chimeric *pcfG*E1-*ltrB*E2 relaxase mRNA was assessed with the *ltrB* gene interrupted by either the Ef.PcfG-WT or the Ef.PcfG-ΔA intron. Arrows denote the relative position of primers used for RT-PCR (blue for RT, red for PCR), and the expected size for the mRNA chimera containing a perfect *pcfG*E1-*ltrB*E2 junction is shown (533 nts). Conjugation efficiencies were measured as the ability of *L. lactis* to transfer the pLE-P_{23}-*pcfF-pcfG*-(E1-Ef.PcfG-E2Δ936) plasmid harboring either the Ef.PcfG-WT or Ef.PcfG-ΔA intron from the donor to the recipient strain. Conjugation efficiency was calculated by dividing the number of transconjugant cells by the number of donor cells and shown with the standard error. Each data point represents triplicates of independent assays. Asterisks denote statistical significance (*: $p < 0.05$).

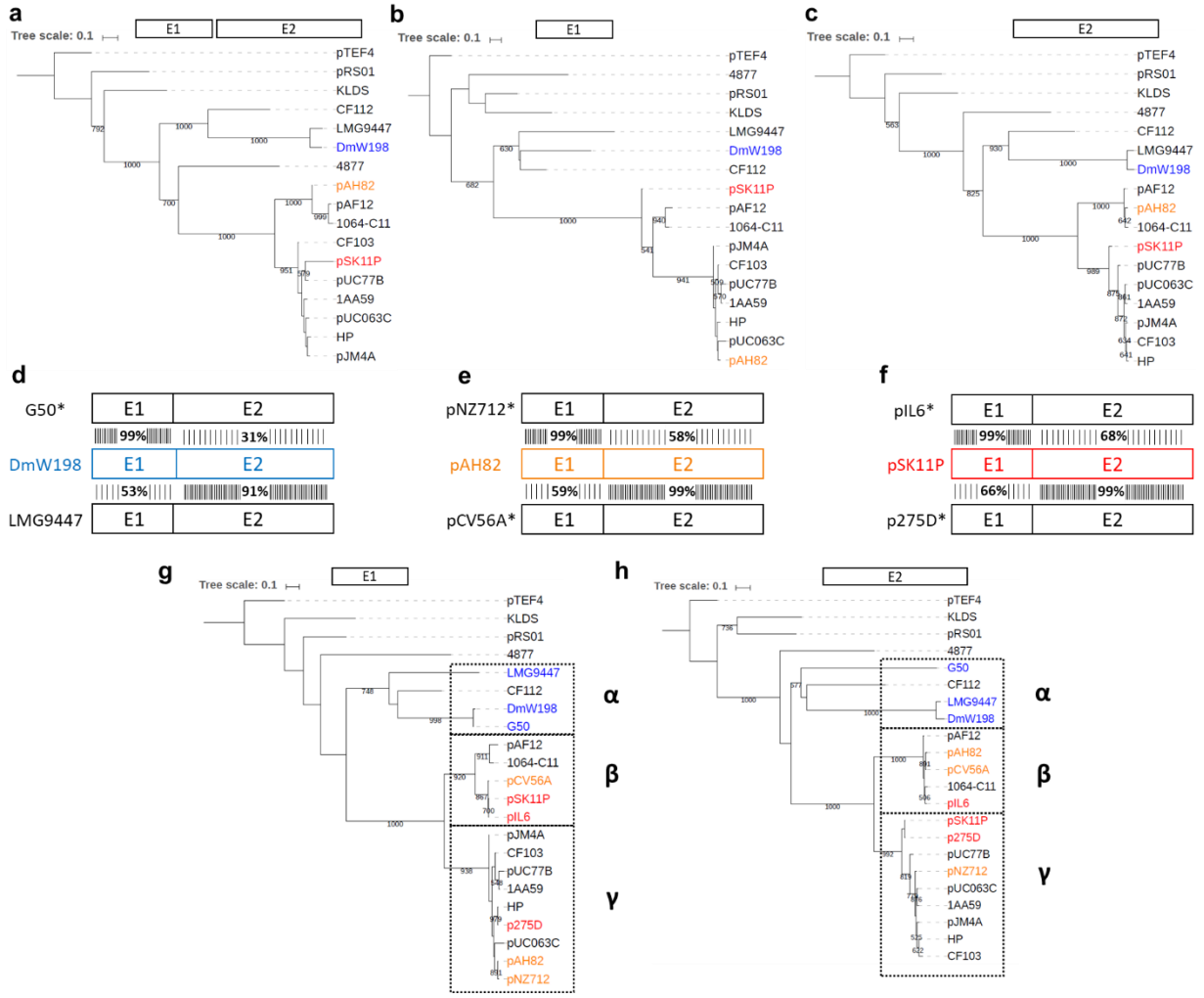
Figure 4.4



Production of chimeric relaxase mRNAs by LI.LtrB in biologically-relevant conditions in *E. faecalis*. (a) The pEF1071 and pLE12 plasmids were co-transformed in the *E. faecalis* lab strain JH2-2. The pEF1071 plasmid expressed the *E. faecalis* non-interrupted relaxase *mobA* gene (blue) while the pLE12 plasmid expressed the *L. lactis* relaxase *ltrB* gene (orange) interrupted by the LI.LtrB group II intron. Both genes were under the control of their

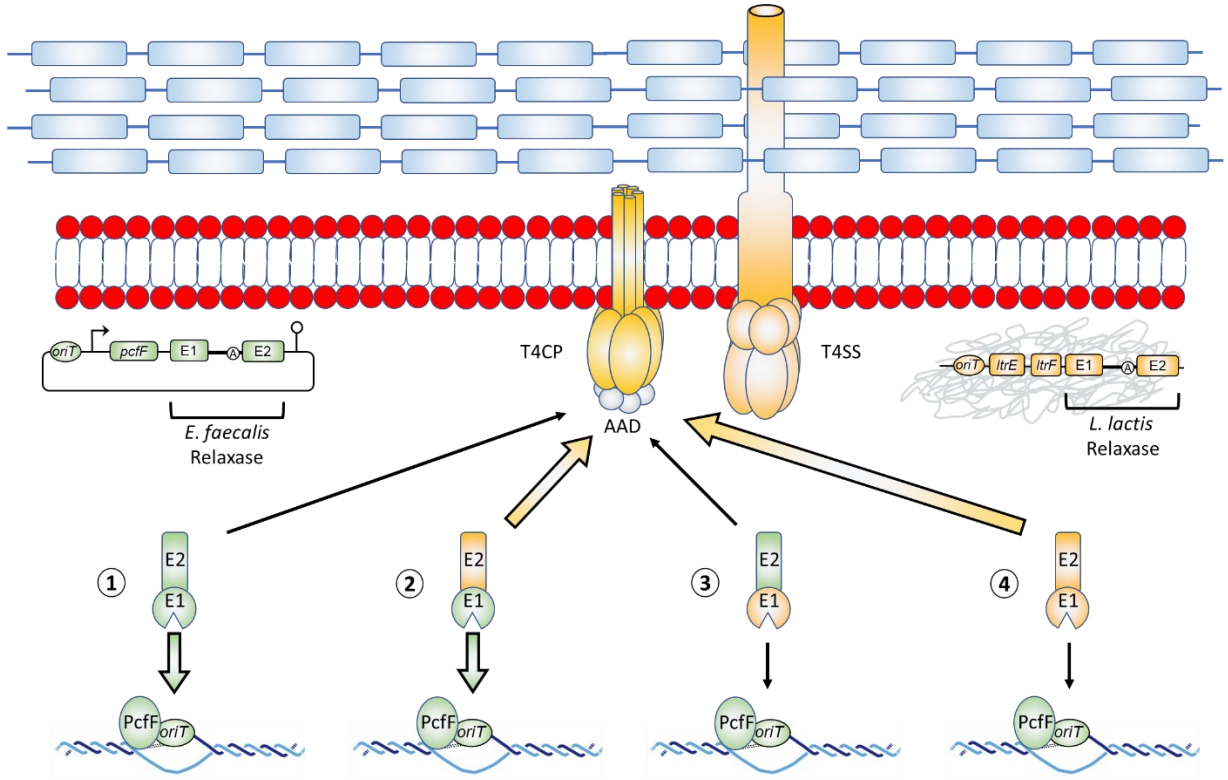
natural promoters. Plasmid preparations from three independent co-transformants were run on an agarose gel and shown to contain both plasmids. Mobility efficiency of Ll.LtrB from pLE12 to pEF1071 was calculated through patch hybridization (Plante and Cousineau 2006) and is shown with the standard error. **(b)** Formation of relaxase mRNA chimeras in different biological settings. RT-PCR was performed to determine the presence of chimeric relaxase mRNAs in JH2-2 strains containing different plasmid combinations. Plasmid combinations from left to right: pLE12 expressing *ltrB* interrupted by the Ll.LtrB-WT intron with pEF1071 expressing non-interrupted *mobA*; pLE12 expressing *ltrB* interrupted by the Ll.LtrB-ΔA intron with pEF1071 expressing non-interrupted *mobA*; pLE12 expressing *ltrB* interrupted by the Ll.LtrB-WT intron with pEF1071 expressing *mobA* also interrupted by the Ll.LtrB-WT intron; pLE12 expressing *ltrB* interrupted by the Ll.LtrB-ΔDV intron with pEF1071 expressing *mobA* also interrupted by the Ll.LtrB-ΔDV intron. Arrows denote the relative position of the RT-PCR primers (blue for RT, red for PCR). The relative amounts of various transcripts were compared to the *ltrB* pre-mRNA precursor by RT-qPCR, in conditions where one **(c)** or both **(d)** relaxase genes were interrupted by Ll.LtrB-WT. **(e)** RT-qPCR of samples producing relaxase chimeras (blue, Ll.LtrB-WT in pLE12; red, Ll.LtrB-WT in both pLE12 and pEF1071). ΔΔCt values were calculated to determine fold change between conditions where either one (blue) or two (red) Ll.LtrB-WT introns were present. Each data point represents technical duplicates done for biological triplicates. Biological replicates were normalized to the *ldhB* housekeeping gene of *E. faecalis*, which was used as a reference gene. Green arrows represent increases of a specific amplicon when two group II introns are present, while the red arrow represents a decrease. Asterisks denote statistical significance for an unpaired Student's T-test (**: $p<0.01$, ***: $p<0.001$).

Figure 4.5



Phylogenetic trees of intron-interrupted relaxase genes from lactococci. The phylogeny of all non-redundant intron-interrupted relaxase genes found in lactococci was assessed by Maximum Likelihood using PhyML with 1000 bootstraps. The phylogenetic trees were produced using either the full length relaxase genes without the group II introns **(a)**, the relaxase sequences preceding the intron (E1) **(b)**, or the relaxase sequences following the intron (E2) **(c).** In each case, the intronless *pcfG* relaxase gene from the *E. faecalis* pTEF4 plasmid was used as the outgroup. Both exons from the DmW198 **(d)**, pAH82 **(e)** and pSK11P **(f)** relaxase genes were individually analysed by searching for the most similar sequences in GenBank using BLASTn. Highest similarity results for E1 are shown above the reference sequence while the highest similarity results for E2 are shown below the reference sequence. Exons of both relaxase genes were then fully compared to the initial query to determine whether homology extended to the entire gene. Relaxase genes with an asterisk denote that they are not interrupted by a group II intron. The relaxase genes marked with an asterisk were included to generate phylogenetic trees with greater resolution for E1 **(g)** and E2 **(h)**. The α, β and γ groups delineate separately evolving lineages of relaxase genes in lactococci. Accession numbers of GenBank sequencing projects containing the relaxase genes used are listed in Table S4.3. Matrices and phylogenetic trees are available in the TreeBASE online repository (URL: http://purl.org/phylo/treebase/phylows/study/TB2:S27000).

Figure 4.6



Some group II intron-generated chimeric relaxase enzymes allow the dissemination of conjugative elements by functionally linking incompatible origins of transfer and conjugation machineries. Schematic representation of a bacterial cell harboring a chromosomally-embedded conjugative element (orange genes). This conjugative element encodes all the genes required for its own lateral transfer by conjugation: a relaxosome, a type 4 coupling protein (T4CP), and a type 4 secretion system (T4SS). When a foreign non-autonomous conjugative element enters the cell, it only contains a minimal set of genes (relaxase, accessory protein) that can form a relaxosome specific for its cognate *oriT* (green genes). This renders its continued lateral transfer by conjugation contingent on successful interactions with its new host conjugation machinery. The non-autonomous conjugative element is in a conjugative cul-de-sac when the C-terminus of its relaxase (E2) cannot be recognized by the host conjugation machinery (scenario 1) and when the N-terminus of the host relaxase (E1) cannot recognize its *oriT* (scenario 4). When a group II intron is found interrupting at least one of the relaxase genes, 2 types of chimeric relaxase enzymes can be generated: green-orange (scenario 2) and orange-green (scenario 3). Conjugative transfer of a non-autonomous conjugative element is most efficient when E1 matches the *oriT* from the mobilizable plasmid (E1, green), and E2 matches the resident conjugative element (E2, orange) (scenario 2). In this scenario, E1 will interact with its MobC-type accessory protein, and *oriT* to initiate nicking while E2 will interact with the host T4CP's all-alpha domain (AAD), which mediates specificity for substrates to be passed along into the T4SS mating pore between the donor and recipient cell during conjugation. The group II intron-generated chimeric relaxase in scenario 2, PcfGE1-LtrBE2, functionally links the *E. faecalis oriT* with the *L. lactis* conjugation machinery leading to efficient conjugative transfer.

4.10: Tables

Table S4.1:

Plasmids used in this study

Plasmid	Antibiotic Resistance	Description and Reference
pLE-Pnis-*ltrB*	Cam^R	ltrB relaxase with Ll.LtrB group II intron under a nisin-inducible promoter (LaRoche-Johnston et al. 2016).
pDL-P_{23}-*pcfG*-E1Δ366-ΔEf.PcfG-His	Spc^R	*pcfG* relaxase lacking its group II inton with a 366 in-frame deletion in exon 1 and a C-terminal 6x His tag.
pLE12	Cam^R	Shuttle vector containing the pRS01 *ltrB* relaxase interrupted by Ll.LtrB (Mills et al. 1996).
pLE12-ΔA	Cam^R	pLE12 with Ll.LtrB lacking its branchpoint adenosine residue.
pLE12-ΔDV	Cam^R	pLE12 with Ll.LtrB lacking its catalytic domain V.
pEF1071::<R6Kγori/Spc^R>	Spc^R	pEF1071 plasmid from *E. faecalis* (Balla and Dicks 2005) containing an R6Kγori-Spc^R transposon.
pEF1071::<R6Kγori/Spc^R>-Ll.LtrB	Spc^R	pEF1071::<R6Kγori/Spc^R> where Ll.LtrB mobilized into the *mobA* Homing Site.
pEF1071::<R6Kγori/Spc^R>-Ll.LtrB-ΔDV	Spc^R	pEF1071::<R6Kγori/SpcR> with the *mobA* relaxase interrupted by Ll.LtrB lacking its catalytic domain V.
pDL-P_{23}	Spc^R	pDL278 shuttle vector (Mills et al. 1997) containing the P_{23} constitutive promoter.
pDL-P_{23}-*ltrB*	Spc^R	pDL278 expressing the *ltrB* relaxase lacking the Ll.LtrB intron from a P_{23} promoter.
pDL-P_{23}-*pcfG*	Spc^R	pDL278 expressing the *pcfG* relaxase lacking the Ef.PcfG intron from a P_{23} promoter.
pDL-P_{23}-*ltrB*E1-*pcfG*E2	Spc^R	Intron-less relaxase containing the sequence upstream of Ll.LtrB (*ltrB*E1) and downstream of Ef.PcfG (*pcfG*E2).
pDL-P_{23}-*pcfG*E1-*ltrB*E2	Spc^R	Intron-less relaxase containing the sequence upstream of Ef.PcfG (*pcfG*E1) and downstream of Ll.LtrB (*ltrB*E2).
pLE-*oriT-L.lactis*	Cam^R	pLE1 shuttle plasmid carrying the *oriT* from pRS01 (Mills et al. 1994).
pLE-*oriT-E.faecalis*	Cam^R	pLE1 shuttle plasmid carrying the *oriT* from pTEF4 (LaRoche-Johnston et al. 2016).
pLE-*oriT-E.faecalis*-P_{23}-*pcfF*	Cam^R	pLE-*oriT-E.faecalis*-PcfF with a P_{23} promoter driving *pcfF* expression.
pLE-*oriT-E.faecalis*-P_{23}-*pcfF-pcfG*-E2Δ936	Cam^R	pLE-*oriT-E.faecalis*-P_{23}-*pcfF* with the Ef.PcfG-interrupted *pcfG* gene, lacking 936 nts in the middle of exon 2.
pDL-P_{23}-*ltrB*-ΔLl.LtrB-E1Δ360	Spc^R	pDL-P_{23}-*ltrB* where the Ll.LtrB intron is missing and *ltrB* has an in-frame deletion of 366 nucleotides in exon 1.
pLE-P_{23}-*pcfF-pcfG*-Ef.PcfGΔA-E2Δ936	Cam^R	pLE-*oriT-E.faecalis*-P_{23}-*pcfF-pcfG*-E2Δ936 with a deletion of the Ef.PcfG branchpoint adenosine.

Table S4.2:

Primers used in this study

Primer	ID	Sequence (5'-3')
pDL-*pcfG* Forward	Bc 1194	AAAGCGGCCGCGTTTAAATGCGCTGGTTCAGAG
pDL-*pcfG* Reverse	Bc 895	AAAGCGGCCGCTTATAGTTTGGGCTTAATGTCGG
pDL-*ltrB* Forward	Bc 270	AAAGCGGCCGCCAGAACGATTTAAAGAAGAATTGAA
pDL-*ltrB* Reverse	Bc 260	AAAGCGGCCGCACTACATCCGTTCATAAACTATAC
pcfG E1-Δ366 Forward	Bc 1440	ACGGGTGGAGAGTATGAATTTG
pcfG E1-Δ366 Reverse	Bc 1441	TGAGCAAGCATTGTTTAATTTATCA
ltrB E1-Δ360 Forward	Bc 1438	ACAGGTGGCGAATATGAATTTGT
ltrB E1-Δ360 Reverse	Bc 1439	GTCTTTTGCCTGGCGTAAATGT
pcfG E2-Δ936 Forward	Bc 899	GTTAGAGAGTAAACTGGAACG
pcfG E2-Δ936 Reverse **and** Chimera PCR *pcfG* E2	Bc 884	TTTGTATGAGTTTGCTTCGGTGA
pcfG 6x His tag Forward	Bc 1498	CACCACCACTAAGCGGCCGCGGATCCT
pcfG 6x His tag Reverse	Bc 1499	ATGATGATGTAGTTTGGGCTTAATGTCGGTTTGC
Chimera RT *pcfG* E2	Bc 885	GACTCTGTTTCGGATTTTG
Chimera PCR *ltrB* E1	Bc 92	TTGGTCATCACCTCATCCAATC
Chimera RT *ltrB* E2 **and** qRT Chimera *ltrB* E2	Bc 301	GAGCCGTTCAATAATAGATTCCA
Chimera PCR *pcfG* E1	Bc 1285	CAACGCCGTTTTAGCACACCA
Chimera PCR *ltrB* E2	Bc 302	CATTTGAGGTTCATCAAGCAGC
Chimera RT *mobA* E2 **and** qRT Chimera *mobA* E2	Bc 1279	GGTCTTTCCAAACCCATTGCC
Chimera PCR *mobA* E2	Bc 1264	ATCTGTAACTGGTGTTCCTGCA
ltrB-ΔLl.LtrB Forward **and** pDL-*ltrB*-ΔE1 Forward	Bc 1421	CATATCATTTTTAATTCTACGAATCTT
ltrB-ΔLl.LtrB Reverse **and** pDL-*ltrB*-ΔE2 Reverse	Bc 1420	GTTATGGATGTGTTCACGATCG
pcfG-ΔEf.PcfG Forward **and** *pcfG* E2-Full Forward	Bc 1305	CATATTATTTTTAGTTCAACCAATTT
pcfG-ΔEf.PcfG Reverse **and** *pcfG* E1-Full Reverse	Bc 1309	ATTGTGTAAATGTTCTTTATCGACAT
pcfG E2-Full Reverse	Bc 1306	TTATAGTTTGGGCTTAATGTCGG
pDL-*ltrB*-ΔE2 Forward	Bc 1303	GGATCCTCTAGAGTCGACCTG
pcfG E1-Full Forward	Bc 1308	GTTTAAATGCGCTGGTTCAGAG
pDL-*ltrB*-ΔE1 Reverse	Bc 1265	TTCAATTCTTCTTTAAATCGTTCTG
E. faecalis oriT Forward	Bc 1228	ACGTCTGCAGATCCATGGCCTTACGAAGAAGCAGCCACTTA
E. faecalis oriT Reverse	Bc 1229	ACGTCTGCAGAATAGTCAACATGGCGAATCTCT
E. faecalis oriT-pcfF Forward	Bc 1373	ACGTCTGCAGGGTCAAAAATGGTAAGTCGAAAC
E. faecalis oriT-pcfF Reverse	Bc 1372	ACGTCTGCAGTTACGATTGTTCCTTTCTCTTTATT
E. faecalis oriT-pcfF-pcfG Forward	Bc 1483	AAAGTCGACGGTCAAAAATGGTAAGTCGAAAC
E. faecalis oriT-pcfF-pcfG Reverse	Bc 1484	AAAGTCGACTTATAGTTTGGGCTTAATGTCGG
P_{23}-*pcfF* Forward	Bc 1521	GATAAAATAGTATTAGAATTGCGAGACTACTTATTATGTAAAAGAAAG
P_{23}-*pcfF* Reverse	Bc 1522	AAAATTTGCTCTTTTTGTCATCACTTAGTTTGACAATTAGCTG
qRT *ldhB E. faecalis* Standard	Bc 1620	CATCTAAGTAAGCTGAGACAGG
qPCR *ldhB E. faecalis* Forward	Bc 1618	ACCATGATTGGTACCAAACCTATT
qPCR *ldhB E. faecalis* Reverse	Bc 1619	TGCGTGCAGTACTCATACCAAT
qPCR *ltrB*E2 Reverse	Bc 93	CTTTAGGAATGACTTTCCAGTC
qPCR *ltrB*E1 Forward	Bc 1595	AAACCATATTAGAATTTACAGGTG
qPCR *mobA*E2 Reverse	Bc 1596	GTCATGGAGGGCAGATACGC
qPCR *mobA*E1 Forward	Bc 1597	TAGAATTGGCCGAGAAAATGGC
qPCR Ll.LtrB Forward	Bc 1598	GGGTACGTACGGTTCCCGA

Table S4.3:

Accession numbers for relaxase genes

Strain/Plasmid	Genbank Accession Number
pTEF4	AJAW01000016
KLDS	CP006766
pRS01	U50902
4877	CALL01000108
LMG9447	LKLS01000214
CF112	OLMC01000017
DmW198	NEQN01000018
G50	CP025500
pAF12	JQ821355
1064-C11	RAGF01000005
pCV56A	CP002366
pSK11P	DQ149245
pIL6	HM021331
pJM4A	CP016729
CF103	OESN01000005
pUC77B	CP016714
1AA59	AZQT01000137
HP	LIYE01000297
p275D	CP016702
pUC063C	CP016717
pAH82	AF243383
pNZ712	KX138409

Chapter 5:
Conclusions and perspectives

The goal of our research project was to address outstanding questions regarding the evolution and function of group II introns. To do so, we used the Ll.LtrB bacterial group II intron from *Lactococcus lactis* as a model system. Nearly 60 Ll.LtrB variants (>95% identical) are present in several strains and subspecies of *L. lactis*, as well as different gram positive bacteria (Candales et al. 2012). This makes our system ideal for studying the various ways that group II introns adapt to novel environments and evolve over time, which remain obscure. Moreover, Ll.LtrB has been a model system to study the versatility of group II intron splicing, yielding novel insight into splicing pathways such as circularization (Monat et al. 2015; Monat and Cousineau 2016). Yet despite a thorough understanding of self-splicing pathways, it remained poorly understood whether any of these mechanisms contributed to functionally benefit the bacterial host or were solely used by the intron to parasitize the host.

5.1: Comparing closely related introns to study intron evolution

To study group II intron evolution, we compared two group II introns that were nearly identical (99.7%) yet located in different bacterial species (*L. lactis* and *E. faecalis*) (see Chapter 2). Their similar nucleotide sequences contrasted with the overall trend of rapid evolution for group II introns. When combined with the much lower conservation of the flanking *ltrB* and *pcfG* genes (Fig. 2.1B), these introns thus likely represented a recent instance of horizontal transfer between bacterial species, providing a unique opportunity to study how laterally transferred introns adapt to their environments.

5.1:1: Bacterial group II introns face selective pressure to maintain high mobility efficiency

In Chapter 2, we examined the evolutionary relationship between the model group II intron Ll.LtrB and a group II intron newly characterized by our lab: Ef.PcfG. We showed that although 8 point mutations distinguish these two introns, both are splicing-competent and mobile in their native environments. When splicing efficiency was compared for both introns in the context of their own and each other's homing sites, no differences were observed. However, when mobility efficiency was measured, both introns were found to be more competent in the *pcfG* site than the *ltrB* site, while Ll.LtrB was more efficient at mobilizing in its own homing site. We later observed that these mobility efficiencies were independent of the flanking exons, suggesting that some of the 8 point mutations were directly involved in the increased mobility of Ll.LtrB to its own homing site. When we tested each of the 8 point mutations independently, we found that two mutations (#2 and #6) significantly increased the mobility efficiency of the Ef.PcfG intron to the *ltrB* homing site. Finally, we generated a dendrogram outlining the distribution of the 8 point mutations throughout Ll.LtrB variants in different *L. lactis* strains and subspecies. Interestingly, we noticed that every single Ll.LtrB variant contained beneficial mutations #2 and #6, while the remaining point mutations gradually accumulated throughout *L. lactis*. The most parsimonious explanation thus appears to be that Ef.PcfG is ancestral to all Ll.LtrB variants sequenced so far, likely having been transferred from *E. faecalis* to *L. lactis* following a single horizontal transfer event.

Taken together, these results point to group II intron mobility efficiency, not splicing efficiency, as being the trait that undergoes strongest selection upon entering a novel bacterial environment. The genomic location of the 8 point mutations supports this view, since nearly all mutations are within the ORF (7/8), and none affect the maturase domain that aids in intron folding, splicing and reverse splicing. The two mutations that increase intron mobility efficiency (#2 and #6) are correspondingly found in the reverse transcriptase domain of the IEP. Overall, this provides experimental evidence supporting the hypothesis — hitherto only bolstered by observational data — that bacterial group II introns behave more like mobile retroelements than splicing elements. Interestingly, the opposite trend was previously observed for group II introns residing in other organisms such as eukaryotic organelles, where they behave mainly as splicing elements and have often lost the ability to mobilize (Dai and Zimmerly 2002a). This difference may in part be due to

the negative selective pressure facing group II introns in bacteria. Since the large population sizes of bacteria lead to high rates of purifying selection and genome streamlining, neutral or even slightly deleterious mutations are eventually removed over large periods of time (Leclercq and Cordaux 2012). Although it remains uncertain precisely how bacteria might purge group II introns over time, one documented mechanism involves the reverse-transcription of mature mRNA where the intron has been excised through self-splicing. This generates an intronless cDNA that can displace the chromosomal allele through homologous recombination, thus removing group II introns from the bacterial chromosome (Jeffares et al. 2006). In bacteria, this specific process might be accomplished by any of the abundant and poorly characterized proteins that harbor a reverse transcriptase (Simon and Zimmerly 2008). The ability of spliced bacterial group II introns to mobilize by retrohoming back into an intronless allele may thus be a mechanism to limit intron displacement and favour intron reinsertion and survival, potentially explaining the increased fitness of group II introns with higher mobility efficiencies.

5.1.2: Conjugation as a means of intron dissemination

Conjugation was previously shown to be a favoured mechanism for the dispersal of Ll.LtrB. Bacterial mating experiments demonstrated the intra-species (*L. lactis* to *L. lactis*) and inter-species (*L. lactis* to *E. faecalis*) transfer of the model group II intron. Although these first involved shuttle vectors that harbored segments of the large native pRS01 plasmid (Belhocine et al. 2004), conjugation was next demonstrated for the large chromosomal Sex Factor that also harbors a copy of Ll.LtrB (Belhocine et al. 2005) and later for the pRS01 plasmid itself (Belhocine et al. 2007). In each case, mobility of the Ll.LtrB intron within both cognate and ectopic homing sites was detected upon arrival into a novel bacterial host, suggesting a pathway for intron dissemination regardless of whether the native plasmid is maintained within the recipient cell. Later experiments showing the spread of different introns by conjugation points to this method of horizontal transfer as a broadly used mechanism for introns to disseminate and seek out new hosts (Nisa-Martinez et al. 2007).

The ecological distribution of the bacterial hosts that harbor Ll.LtrB and Ef.PcfG suggests that many natural horizontal transfer events may have taken place over the course of evolution. *E.*

faecalis is an incredibly abundant member of the adult human microbiota, comprising along with *E. faecium* nearly 1% of total fecal content (Dubin and Pamer 2014). On the other hand, despite not being a consistent member of the human microbiota, the widespread use of *L. lactis* as dairy starters in cheese production and milk fermentation has led to its profuse consumption (Mills et al. 2010). Moreover, genetic marking analyses have demonstrated that *L. lactis* can survive passage through the human gastrointestinal (GI) tract following ingestion (Klijn et al. 1995). Taken together, these characteristics of *L. lactis* and *E. faecalis* suggest that many opportunities for the close physical contact needed during conjugation may have arisen over the course of evolution in the human GI tract, where horizontal gene transfer is a common occurrence (McInnes et al. 2020).

Interestingly, despite the seemingly abundant opportunities of horizontal gene transfer between *E. faecalis* and *L. lactis*, our data suggests that a single instance of conjugation led to the current abundance of group II introns in *L. lactis* (see Chapter 2). This may be due to the relative paucity of group II introns in *E. faecalis*, where widespread sequencing due to their medical relevance has only identified Ef.PcfG itself and a single Ef.PcfG variant (LaRoche-Johnston et al. 2016). However, an interesting alternative may be that multiple transfer events have in fact taken place between *E. faecalis* and *L. lactis*, yet the arriving intron was only maintained within *L. lactis* a single time. A possible explanation is that although the Ef.PcfG intron recognizes the *ltrB* homing site sufficiently well to mobilize through retrohoming, the basal mobility rate (~ 42%) may be too low to prevent subsequent intron removal through purifying selection. However, when an Ef.PcfG variant harboring beneficial mutations #2 (mobility rate of 69%) and #6 (mobility rate of 58%) was introduced into *L. lactis*, its mobility efficiency was high enough to be maintained. The insertion of this ancestral Ll.LtrB variant within the conserved catalytic histidine triad of a relaxase gene next ensured its continued transmission and further dissemination by conjugation within *L. lactis* strains and sub-species.

5.1.3: Linking point mutations to functional differences in homologous introns

Our approach to look at recent events of intron dispersal was not a novel way of studying intron evolution. Previous groups had also analyzed the distribution of group II introns in natural populations of *Escherichia coli* (Dai and Zimmerly 2002b), *Sinorhizobium meliloti* (Fernandez-

Lopez et al. 2005) and *Bacillus cereus* (Tourasse and Kolsto 2008). However, these studies largely provided observational insight into the overall evolution and behavior of group II introns, rather than functional insight. They consistently found that group II introns are frequently fragmented and are often present alongside unoccupied homing sites, validating previous observations that bacterial group II introns behave mostly as retroelements (Dai and Zimmerly 2002a). Moreover, these studies often concluded that recent events of horizontal gene transfer had taken place, based on the distribution of similar group II introns throughout different bacterial strains. Yet in each case, the direction of horizontal transfer and the selective forces shaping newly acquired point mutations was impossible to determine. In Chapter 2, we took the novel approach of testing homologous introns functionally, both in terms of splicing and mobility efficiency. By showing that splicing was unchanged between Ef.PcfG and Ll.LtrB, while mobility efficiency increased significantly in *L. lactis*, we demonstrated that selective pressures affecting bacterial group II introns mainly target mobility efficiency. Our data thus provide functional support to the theory that group II introns behave mostly as retroelements in bacteria, rather than mainly functioning as splicing elements.

While we established that certain point mutations in Ll.LtrB resulted in an increased mobility efficiency in *L. lactis*, we still lack a clear understanding of the underlying mechanistic reasons. Our approach of testing the effects of each point mutation independently revealed that some mutations (3/8) either significantly increased (#2, #6) or decreased (#7) mobility efficiency (see Chapter 2). Interestingly, all three of these mutations are in the *ltrA* IEP (Fig. 2.1). Crystal structures of Ll.LtrB in complex with a copy of the LtrA protein were recently produced, revealing the contact points between Ll.LtrB and its IEP (Qu et al. 2016). It would thus be interesting to assess how these point mutations alter either the interaction between LtrA and Ll.LtrB, or between LtrA and the dsDNA substrate during retrohoming. This approach however still biases functional analyses to understand how each point mutation acts individually, without considering the potentially synergistic nature of the point mutations. Since our study took place (LaRoche-Johnston et al. 2016), new genomic sequencing data has increased the number of characterized Ll.LtrB variants (Fig. 5.1A). This has added resolution and further reinforced our dendrogram of point mutation distribution between Ef.PcfG and Ll.LtrB variants (Fig. 2.6). Given this increased resolution, it would be interesting to test the effects of the 8 point mutations on mobility efficiency in a sequential manner, following the most parsimonious order of mutation acquisition (Fig. 5.1B).

We could thus elucidate whether mutations #2 and #6 increase mobility efficiency independently or synergistically. Likewise, we could assess whether mutation #7, one of the last mutations to appear in Ll.LtrB, maintains its negative impact on mobility efficiency or has a different effect altogether once other beneficial mutations are present.

5.2: Elucidating a function for group II introns

To address group II intron function, we began by studying the circularization pathway. Although it appears to be conserved throughout different group II intron subclasses (Murray et al. 2001; Li-Pook-Than and Bonen 2006; Molina-Sanchez et al. 2006) and to occur concurrently with branching (Monat et al. 2015; Monat and Cousineau 2016), its function has remained enigmatic. We thus wanted to address an outstanding question about group II intron circles: how certain circular molecules were generated harboring additional nucleotides of unknown origin at their splice junctions, rather than being perfect head-to-tail intron circles as the conventional circularization pathway would predict (see Chapter 3).

5.2.1: Group II introns generate genetic diversity

In Chapter 3, we described the precise mechanism underlying the incorporation of additional nucleotides at the circle splice junctions of group II introns. We began by compiling an extensive list of additional nucleotides found at the junction of individual Ll.LtrB-WT circles. We found that nucleotide fragments consistently mapped to the coding strand of genes from the bacterial chromosome or from resident plasmids, indicating that they corresponded to mRNA fragments. Moreover, the specificity of incorporated mRNA fragments changed when Ll.LtrB contained a different EBS1 sequence. Consensus sequences accordingly showed that nucleotide fragment selection depended on base pairing ability with the Ll.LtrB intron.

We next proposed a pathway to explain how group II intron circles might incorporate these additional nucleotides, which depended on a combination of branching and circularization. Our pathway begins by the initial reverse splicing of excised Ll.LtrB lariats into mRNAs, a concept

that was previously demonstrated *in vitro* (Morl and Schmelzer 1990b) but has been poorly described *in vivo*. Our data in *L. lactis* demonstrate the abundance and versatility of this reaction. Gene-specific analyses show that Ll.LtrB invades bacterial mRNAs at multiple sites, where efficiency is correlated with the strength of the EBS-IBS 11 nucleotide base pairs (7/11 nts to 11/11 nts interactions were found). Once fully reverse spliced into an mRNA substrate, the group II intron self-splices through the circularization pathway. A *trans*-acting nucleophile is recruited by the intron, whose 3′ OH attacks the first nucleotide of the downstream mRNA, generating a chimeric mRNA. Through RNA-Seq and gene-specific analyses, we showed that both E1 and various other mRNA fragments containing an IBS1/2-like sequence at their 3′ ends could be recruited as external nucleophiles, thus increasing the diversity of the bacterial transcriptome by generating a population of chimeric E1-mRNA and mRNA-mRNA molecules. The exact scope of this effect on the transcriptome is difficult to assess, but studies from other groups have yielded interesting insights. For example, non-contiguous reads from bacterial transcriptome data are frequently disregarded altogether from RNA-Seq analysis pipelines, due to the paucity of intervening sequences such as group II introns. However, an analysis of discarded RNA-Seq split reads from several bacterial species containing group II introns showed a large number of non-contiguous and circularized reads that could not be explained by known splicing mechanisms (Doose et al. 2013).

Finally, once a chimeric mRNA transcript has been formed, the liberated 3′ end of the intron becomes free to circularize either at the initially recognized site, generating a head-to-tail intron circle, or at an upstream site, generating the initially described intron circles with mRNA fragments at their splice junctions. We thus proposed that intron circles harboring additional nucleotides, which are highly stable and accumulate over time, might serve as indirect indicators of chimera formation, while chimeric mRNA transcripts may themselves be short-lived.

The formation of chimeric transcripts was previously suggested as a group II intron function, based on *in vitro* work (Morl and Schmelzer 1990a). Morl and colleagues found that combining intron lariats with different RNA fragments harboring IBS1/2-like regions yielded recombined RNA molecules. Recently, the Ll.LtrB group II intron was shown by Northern blotting to also generate recombined RNAs *in vivo* (Qu et al. 2018), through a mechanism analogous to the process described by Morl and colleagues. In this pathway, Ll.LtrB initiates the first

transesterification reaction of branching, generating a 2′-5′ branched lariat covalently linked to E2. E1, bound to the intron only through base pairing interactions, is displaced and substituted for another RNA molecule harboring an IBS1/2-like sequence at its 3′ end. If Ll.LtrB completes the branching pathway, the swapped E1 is ligated to the downstream E2, resulting in a recombined RNA molecule. Although we only have limited information regarding the scope of this intron-catalyzed reaction, we cannot rule out that this type of intron-mediated recombination also generates chimeric RNAs *in vivo.* However, the abundance of circular intron RNAs and the diversity of the additional nucleotides at their splice junctions underscores the importance of our newly described pathway, which requires a balance between branching and circularization.

5.2.2: Genetic diversity: similarities between group II introns and the spliceosome

The ability of group II introns to increase genetic diversity resembles the function of another ribozyme: the nuclear splicing machinery called the spliceosome. Although nuclear splicing and group II intron branching occur using a biochemically identical pathway, an important distinction has always been the substrate of the splicing reaction. Group II introns are *cis*-acting ribozymes, which means that their catalytic "self"-splicing is limited to their own sequence. The only known exceptions were bipartite and tripartite *trans*-splicing group II introns, yet even these *trans*-acting molecules assemble and splice based on self-recognition. In contrast, as the generalized splicing machinery in the nuclei of eukaryotes, the spliceosome is responsible for the accurate splicing of all nuclear introns, meaning that it always functions *in trans*. The evolutionary transition from self-splicing to *trans*-splicing is thought to have occurred through the progressive fragmentation of ancestral group II introns into distinct pieces (now snRNAs), which were gradually stabilized with a supporting protein scaffold (Sharp 1991). In bacteria, group II introns function mainly as selfish retroelements, which likely represents the ancestral state (Dai and Zimmerly 2002a). The inability of group II introns to *trans*-splice molecules other than their own sequence was considered a confirmation of their otherwise selfish behavior, in which self-splicing evolved solely as a means of limiting damage to the host. Our observation that group II introns can *trans*-splice together various host mRNAs thus challenges the longstanding view that group II

introns function exclusively as *cis*-acting ribozymes. Moreover, these data present an interesting paradox, where the ability of group II introns to increase bacterial genetic diversity likely originated as a by-product of otherwise selfish behavior. Since group II introns are believed to be the ancestors of nuclear introns, the newly described catalytic function of generating genetic diversity, albeit at low levels, provides a glimpse of the role these bacterial ribozymes would later evolve to fully support in the nuclei of eukaryotes.

The specific mechanism through which group II introns generate genetic diversity has two important parallels in eukaryotes. First, we demonstrated in Chapter 3 that the most efficient way to initiate circularization from an intron-invaded ectopic bacterial mRNA was to use the intron's cognate E1 as the external nucleophile, generating a variety of E1-mRNA chimeras (Fig. 5.2A). This type of reaction resembles the Spliced Leader (SL) *trans*-splicing catalyzed by many forms of lower eukaryotes. During this pathway, a common 5′ exon is *trans*-spliced to many different independently transcribed mRNAs, providing transcripts with a ribosome-binding site that enables their translation (Fig. 5.2B) (Nilsen 1993). The origin of SL *trans*-splicing is still uncertain, mostly due to its patchy phylogenetic distribution (Nilsen 2001). The different evolutionary scenarios that have been proposed for its origin thus involve either multiple instances of lineage-specific gain or loss (Hastings 2005). Given the mechanistic similarities between group II intron chimera formation and SL *trans*-splicing presented herein, and the intron-rich ancestors of the earliest eukaryotes (Koonin 2009), it is tempting to think that certain fragmenting group II introns in early eukaryotic lineages may have maintained a preference for their cognate E1, and accordingly evolved to use a single E1 as a common external nucleophile to be *trans*-spliced to every mRNA harboring an intervening intron during self-splicing.

Second, we demonstrated in Chapter 3 that group II introns increase genetic diversity by forming mRNA-mRNA chimeras, where they *trans*-splice a bacterial mRNA containing an IBS1/2-like sequence at its 3′ end to the downstream sequence of an interrupted mRNA (Fig. 5.2C). This reaction is analogous to the intergenic *trans*-splicing reaction catalyzed in eukaryotes, where exons from different genes are used to create chimeric RNAs (Fig. 5.2D) (Lei et al. 2016). The eukaryotic splicing reaction has remained poorly characterized, since many intergenic chimeras are identified through next-generation sequencing platforms such as RNA-Seq, where the RT-generated libraries can produce artefacts through template switching (Yu et al. 2014). However,

some biologically relevant chimeric mRNAs were shown to *trans*-splice *in vivo* and to be physiologically regulated in healthy tissues, such as the *JAZF1-JJAZ1* transcript (Li et al. 2008). It is thus likely that intergenic *trans*-splicing is an underappreciated facet of eukaryotic gene expression, with the ability to largely increase genetic complexity (Gingeras 2009). Overall, the bacterial group II intron splicing pathways described in Chapter 3 that lead to increased genetic diversity draw new evolutionary parallels with eukaryotic splicing pathways.

5.3: Does a novel group II intron function lead to a beneficial relationship with its bacterial host?

In Chapter 4, we described a specific circumstance where the newly discovered group II intron splicing pathway that increases genetic diversity in bacteria may provide a beneficial function. We began by assessing which potential candidates out of the possible E1-mRNA and mRNA-mRNA chimeras might lead to a novel function that we could detect in bacteria. Since Ll.LtrB has preferential use for its cognate E1 as the external nucleophile, we assumed that an E1-mRNA chimera might be more abundant and yield a more detectable phenotype than an mRNA-mRNA chimera. Moreover, since Ll.LtrB naturally interrupts the *ltrB* relaxase gene at a conserved catalytic motif, we tested whether our model group II intron could produce E1-mRNA chimeras using homologous relaxases, and whether these might exert any novel function.

5.3.1: Increasing genetic diversity in *L. lactis* by generating chimeric relaxases with gain-of-function phenotypes

We began by demonstrating that Ll.LtrB can generate chimeric E1-mRNA relaxases between its cognate *ltrB* relaxase and the orthologous *pcfG* relaxase. The latter enzyme originates from the SF24397 clinical strain of *E. faecalis* (see Chapter 2) and was a good candidate to use for several reasons. First, *pcfG* and *ltrB* are both interrupted by their respective introns in the same frame, so E1-mRNA chimeras produced with either exon combination would yield in-frame mRNAs that could be translated into chimeric proteins. Second, both enzymes belong to the IncP

family of conjugative relaxases, so chimeric products would likely remain functionally active and biochemically stable since both proteins share the same functional motifs (Pansegrau et al. 1994). Third, two factors indicated group II intron recognition of this relaxase: the presence of the native Ef.PcfG intron and our previous data demonstrating that Ll.LtrB can target and invade the *pcfG* homing site precisely within the catalytic motif (see Chapter 2). Finally, the exact *oriT* nucleotide sequence as well as the specificity determinants between the *L. lactis* and *E. faecalis* conjugative systems were previously determined (Chen et al. 2007), facilitating our use of both systems to test the function of chimeric relaxases.

Using the *ltrB* and *pcfG* chimeras, we showed that Ll.LtrB can generate chimeric relaxase mRNAs and that these mRNAs can be translated to yield chimeric proteins. When both an interrupted and uninterrupted relaxase were co-expressed under natural promoters to approximate biologically relevant ratios of pre-mRNA and mature mRNA, we showed that chimeric relaxases are produced at nearly 1-2% of the native ligated exons. Interestingly, this ratio increased 14-fold when both relaxases were interrupted by an identical group II intron. The correlation between increased genetic diversity and higher intron copy number provides a stark contrast to the generally low copy numbers of group II introns in bacterial cells (Lambowitz and Zimmerly 2011). This potentially suggests that an increase in the abundance of introns may undergo positive selection when a beneficial chimera is produced. Moreover, mutational analyses in which the catalytic DV of group II introns was removed led to a complete absence of chimeric relaxases. Group II introns were shown to be five times as recombinogenic as IS elements (Leclercq et al. 2011), yet chimera formation nevertheless appears to be actively driven by catalysis rather than through homologous recombination. When we used conjugation as a functional output to assess the efficiency of various engineered chimeric relaxases, we found a case where one such enzyme, *pcfG*E1-*ltrB*E2, was significantly more efficient at transferring the native *E. faecalis oriT* than both WT relaxases. To assess the biological relevance of the gain-of-function phenotype we observed when expressing a chimeric gene, we tested a conjugation system where the *pcfG*E1-*ltrB*E2 chimeric enzyme could only be generated *in vivo* through *trans*-splicing. We found that the chimera was produced at sufficiently high levels through the novel splicing pathway to maintain conjugative transfer, with only a 12-fold decrease in conjugation efficiency over expression as a contiguous engineered gene (3.42×10^{-6} vs 4.07×10^{-5}).

Finally, a phylogenetic analysis of intron-interrupted relaxases in *L. lactis* revealed the presence of certain natural chimeric genes, where the E1 and E2 of a given relaxase grouped with separate lineages and thus appeared to have distinct evolutionary histories. Since the point at which nucleotide identity suddenly falls corresponds exactly to the intron insertion site, the existence of these chimeric genes is likely caused by the presence of group II introns. Although it is unclear which underlying mechanism is responsible for their formation, a few testable hypotheses exist. The most likely scenario is homologous recombination mediated by two nearly identical group II introns, such as two Ll.LtrB variants whose nucleotide sequence is >95% identical. As previously mentioned, group II introns are very recombinogenic. When uncontrolled proliferation occurs, largescale homologous recombination events can take place that reshape genomes, which may partly explain why purifying selection would favor low intron copy numbers (Leclercq et al. 2011). However, when highly homologous group II introns undergo homologous recombination, they have the unique advantage of frequently maintaining the ability to self-splice, while other mobile elements inactivate the recombined genes. Indeed, if the intron copy remains functional, the recombined flanking exons are ligated and can begin evolving as a functional unit. Another possible scenario would be that group II introns somehow actively produced these chimeric gene, potentially reverse transcribing them into cDNA and integrating them back into the bacterial chromosome.

Overall, the intron-mediated formation of chimeric genes may be a way to fix beneficial chimeras into the bacterial chromosome, where they are expressed as contiguous genes rather than needing to be *trans*-spliced *in vivo* (Fig. 5.3). Indeed, if a chimera generated by a group II intron at the RNA-level (Fig. 5.3 step 1) is beneficial, then an increase of intron copy numbers undergoes positive selection (Fig. 5.3 step 2), since it leads to an increase in the number of chimeric mRNAs (Fig. 5.3 step 3) (see Chapter 4). Once several group II intron copies are present in a bacterial cell, the likelihood of generating a chimeric gene increases (Fig. 5.3 step 4), either through homologous recombination or through an intron-generated cDNA intermediate, which would also undergo positive selection since the expression of a contiguous chimeric gene would bypass the need for *trans*-splicing and would thus further increase the production of the beneficial chimeric mRNA (Fig. 5.3 step 5).

5.3.2: Chimeric relaxases increase the promiscuity of horizontal transfer

As previously mentioned, the gain-of-function phenotype obtained with the *pcfG*E1-*ltrB*E2 chimeric relaxase is an increase in horizontal gene transfer, specifically for the plasmid harboring the *E. faecalis oriT*. This is likely due to the separation of functional motifs for conjugative relaxases of the IncP family, since the catalytic motifs are located at the N-terminus (Kopec et al. 2005) while motifs mediating the protein-protein interactions that confer specificity to the conjugative system are located at the C-terminus (Whitaker et al. 2015). In our specific biological system, the PcfGE1-LtrBE2 protein is thus recruited to its cognate *E. faecalis oriT* by the PcfF accessory protein through interactions at the catalytic N-terminus (Chen et al. 2007), where it nicks *oriT*. Once covalently bound to the released 5′ phosphate (Byrd and Matson 1997), the chimeric relaxase next migrates to the bacterial membrane where it interacts with the lactococcal type 4 coupling protein (T4CP) (Chen et al. 2008). Since the PcfGE1-LtrBE2-T4CP protein-protein interactions also appear cognate due to the C-terminal LtrBE2 component of the chimeric relaxase, conjugation is completed and the ssDNA is transferred through the T4SS mating pore.

In the biological context of *L. lactis* and *E. faecalis*, our data suggest that when conjugative elements harboring a group II intron transfer horizontally from *E. faecalis* to *L. lactis*, they have a greater chance of being consistently transferred throughout *L. lactis* thereafter, rather than becoming "trapped" and unable to transfer laterally. The native mobile element harboring *pcfG* and Ef.PcfG was called pTEF4 (see Chapter 2) due to its high nucleotide identity with the pTEF2 plasmid (Paulsen et al. 2003). Both pTEF2 and pTEF4 are evolutionarily derived from the pCF10 pheromone-sensitive plasmid, whose molecular characteristics have been widely studied (Hirt et al. 2005). These plasmids are termed conjugative elements, since they contain all of the requisite machinery to independently mediate their own horizontal transfer by forming a full relaxosome and an associated mating pore (Smillie et al. 2010). Interestingly, pCF10 can very efficiently transfer a much smaller class of genetic elements called mobilizable plasmids from *E. faecalis* into *L. lactis*, which contain only a small number of proteins required to form a minimal relaxosome: an *oriT*, a relaxase and an accessory protein (Staddon et al. 2006). Once horizontal transfer has occurred and the mobilizable plasmid arrives in *L. lactis*, it becomes entirely dependent on the machinery of conjugative elements within the new host for its continued horizontal transmission (Smillie et al. 2010). When conjugation systems are not identical yet homologous, a limited

amount of transfer can occur (see Chapter 4), either by the partial recognition of the mobilizable plasmid's *oriT* by the *L. lactis* relaxase, or by the partial interaction between the mobilizable plasmid's relaxase and the *L. lactis* T4CP. However, we demonstrated through conjugation assays using various relaxases that when both foreign and resident relaxases are unable to partially mediate conjugation, a chimeric relaxase generated by a group II intron becomes the only way to successfully transfer the newly arrived plasmid.

Collectively, our data thus indicate that the ability of group II introns to generate chimeric relaxases leads to an increase in the horizontal transfer of their conjugative plasmids. We previously demonstrated that the current diversity of Ll.LtrB-variants in *L. lactis* likely originated from a single instance of horizontal transfer from *E. faecalis* (see Chapter 2). The fact that nearly all (53/60) Ll.LtrB-variants in different strains and sub-species of *L. lactis* are still found within putative relaxases (Candales et al. 2012) may thus point to a sustained production of chimeric relaxases throughout evolution, resulting in higher levels of conjugation and likely contributing to the rapid dissemination of Ll.LtrB.

5.3.3: Expansion of group II introns in *L. lactis*: transient colonization or positive selection?

As previously mentioned, group II introns generally have a parasitic relationship with their bacterial host, where they behave mostly as retromobile elements (Dai and Zimmerly 2002a) and need to constantly overcome the purifying selection that would otherwise exclude them from bacteria altogether (Leclercq and Cordaux 2012). As seen in Chapter 2, this leads to the selective advantage of group II introns harboring beneficial mutations that increase mobility efficiency, not splicing efficiency. A potential model to explain bacterial group II intron dynamics is the extinction-recolonization model, where bacteria undergo recurring acquisitions and expansions of group II introns, followed by their rapid decline (Wagner 2006). We thus wanted to study the recent natural burst of Ll.LtrB dissemination throughout *L. lactis* (see Chapter 2), to better understand whether it corresponded to an uncontrolled proliferation in this novel species that simply outpaced purifying selection, or whether it was actually fuelled by positive selection.

Ll.LtrB variants are so similar (all >95% identical) that the likeliest explanation for their wide distribution in *L. lactis* is through horizontal gene transfer (HGT). This method of sharing DNA is very widespread among bacteria, providing an essential supply of novel genetic information, analogous to sexual reproduction. Without HGT, asexual populations that rely solely on vertical transmission rapidly accumulate detrimental mutations that lead to declining populations, a process termed mutational meltdown (Lynch et al. 1993). HGT also has the immediate benefit of providing bacteria with genes that enable rapid adaptation to a novel environment, such as antibiotic resistance in *Enterococci* (Paulsen et al. 2003). In the specific case of *L. lactis*, HGT has had an enormous effect on reshaping its genome, mainly due to the frequent transfer of plasmids by conjugation (Cavanagh et al. 2015). Although *L. lactis* currently inhabits a variety of different ecological niches, the different strains of this lactic acid bacterium can be classified as either environmental (isolated from plants, raw milk and animals) or domesticated (used as dairy starters and in milk production) (Passerini et al. 2010). Several differences exist between environmental and domesticated strains, most notably chromosome size, which is consistently smaller in strains of industrial dairy origin: a phenomenon attributable to reductive evolution (Kelly et al. 2010).

Domesticated dairy lactococci have a much larger plasmid content than non-dairy lactococci, making up nearly 5% of the total genome (Makarova et al. 2006). They often confer traits that lead to a selective advantage for the bacterial host, such as catabolic genes that degrade casein and lactose, enabling their use as sources of amino acids and carbon, respectively, and restriction-modification genes that resist bacteriophage infection (Mills et al. 2006). Interestingly, nearly every Ll.LtrB variant identified in our studies was found within a lactococcal strain isolated from a dairy environment. They are almost always found on mobile elements such as conjugative plasmids, that are correspondingly often associated with traits that benefit the bacterial host, such as Ll.LtrB within the pRS01 plasmid that encodes genes for lactose utilization (Anderson and McKay 1984) and a Ll.LtrB variant within the pSK11P plasmid (see Chapter 4) that encodes genes for cadmium resistance (Siezen et al. 2005). It is thus possible that the burst of intron mobility in *L. lactis* that originated from *E. faecalis* (see Chapter 2) coincided with the population bottleneck of modern industrial dairy *L. lactis* strains (Kelly et al. 2010). The fact that Ll.LtrB preferentially interrupts relaxase genes and is frequently exposed to orthologous relaxases likely resulted in the sustained production of chimeric relaxase enzymes, whose gain-of-function phenotypes would

have allowed for an enhanced dissemination of plasmids containing group II introns. Since these lactococcal plasmids often encode genes conferring selective advantages to the dairy environment, the ensuing promiscuity of horizontal gene transfer would have resulted in an increased dissemination of both beneficial plasmids and group II introns throughout *L. lactis* populations. It is thus likely that the spread of Ll.LtrB variants including Ll.LtrB itself (see Chapter 2) may have been beneficial for *L. lactis*, such that individual introns underwent positive selection due to their ability to increase genetic diversity (see Chapter 3) and enhance horizontal gene transfer (see Chapter 4).

5.4: References

Anderson DG, McKay LL. 1984. Genetic and physical characterization of recombinant plasmids associated with cell aggregation and high-frequency conjugal transfer in Streptococcus lactis ML3. *J Bacteriol* 158(3):954-62.

Belhocine K, Plante I, Cousineau B. 2004. Conjugation mediates transfer of the Ll.LtrB group II intron between different bacterial species. *Mol Microbiol* 51(5):1459-69.

Belhocine K, Yam KK, Cousineau B. 2005. Conjugative transfer of the Lactococcus lactis chromosomal sex factor promotes dissemination of the Ll.LtrB group II intron. *J Bacteriol* 187(3):930-9.

Belhocine K, Mandilaras V, Yeung B, Cousineau B. 2007. Conjugative transfer of the Lactococcus lactis sex factor and pRS01 plasmid to Enterococcus faecalis. *FEMS Microbiol Lett* 269(2):289-94.

Byrd DR, Matson SW. 1997. Nicking by transesterification: the reaction catalysed by a relaxase. *Mol Microbiol* 25(6):1011-22.

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

Cavanagh D, Fitzgerald GF, McAuliffe O. 2015. From field to fermentation: the origins of Lactococcus lactis and its domestication to the dairy environment. *Food Microbiol* 47:45-61.

Chen Y, Staddon JH, Dunny GM. 2007. Specificity determinants of conjugative DNA processing in the Enterococcus faecalis plasmid pCF10 and the Lactococcus lactis plasmid pRS01. *Mol Microbiol* 63(5):1549-64.

Chen Y, Zhang X, Manias D, Yeo HJ, Dunny GM, Christie PJ. 2008. Enterococcus faecalis PcfC, a spatially localized substrate receptor for type IV secretion of the pCF10 transfer intermediate. *J Bacteriol* 190(10):3632-45.

Dai L, Zimmerly S. 2002a. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res* 30(5):1091-102.

Dai L, Zimmerly S. 2002b. The dispersal of five group II introns among natural populations of Escherichia coli. *RNA* 8(10):1294-307.

Doose G, Alexis M, Kirsch R, Findeiss S, Langenberger D, Machne R, Morl M, Hoffmann S, Stadler PF. 2013. Mapping the RNA-Seq trash bin: unusual transcripts in prokaryotic transcriptome sequencing data. *RNA Biol* 10(7):1204-10.

Dubin K, Pamer EG. 2014. Enterococci and Their Interactions with the Intestinal Microbiome. *Microbiol Spectr* 5(6).

Fernandez-Lopez M, Munoz-Adelantado E, Gillis M, Willems A, Toro N. 2005. Dispersal and evolution of the Sinorhizobium meliloti group II RmInt1 intron in bacteria that interact with plants. *Mol Biol Evol* 22(6):1518-28.

Gingeras TR. 2009. Implications of chimaeric non-co-linear transcripts. *Nature* 461(7261):206-11.

Hastings KE. 2005. SL trans-splicing: easy come or easy go? *Trends Genet* 21(4):240-7.

Hirt H, Manias DA, Bryan EM, Klein JR, Marklund JK, Staddon JH, Paustian ML, Kapur V, Dunny GM. 2005. Characterization of the pheromone response of the Enterococcus faecalis conjugative plasmid pCF10: complete sequence and comparative analysis of the

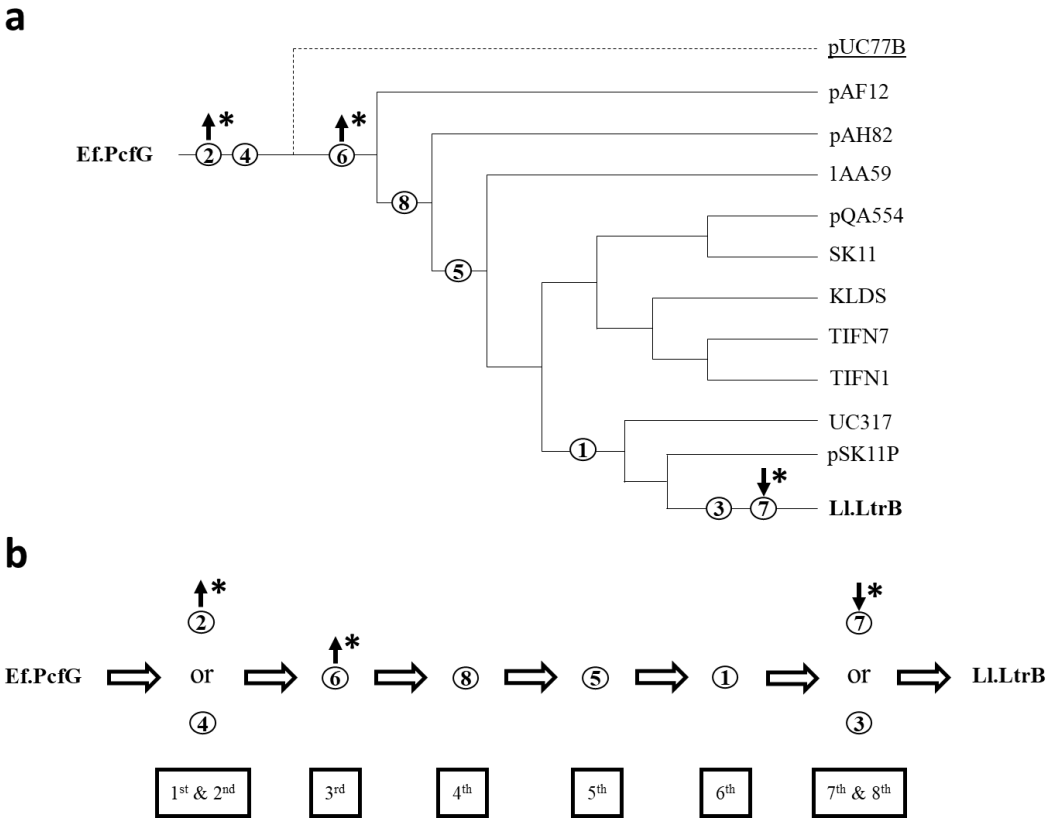
transcriptional and phenotypic responses of pCF10-containing cells to pheromone induction. *J Bacteriol* 187(3):1044-54.
Jeffares DC, Mourier T, Penny D. 2006. The biology of intron gain and loss. *Trends Genet* 22(1):16-22.
Kelly WJ, Ward LJ, Leahy SC. 2010. Chromosomal diversity in Lactococcus lactis and the origin of dairy starter cultures. *Genome Biol Evol* 2:729-44.
Klijn N, Weerkamp AH, de Vos WM. 1995. Genetic marking of Lactococcus lactis shows its survival in the human gastrointestinal tract. *Appl Environ Microbiol* 61(7):2771-4.
Koonin EV. 2009. Intron-dominated genomes of early ancestors of eukaryotes. *J Hered* 100(5):618-23.
Kopec J, Bergmann A, Fritz G, Grohmann E, Keller W. 2005. TraA and its N-terminal relaxase domain of the Gram-positive plasmid pIP501 show specific oriT binding and behave as dimers in solution. *Biochem J* 387(Pt 2):401-9.
Lambowitz AM, Zimmerly S. 2011. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol* 3(8):a003616.
LaRoche-Johnston F, Monat C, Cousineau B. 2016. Recent horizontal transfer, functional adaptation and dissemination of a bacterial group II intron. *BMC Evol Biol* 16(1):223.
Leclercq S, Giraud I, Cordaux R. 2011. Remarkable abundance and evolution of mobile group II introns in Wolbachia bacterial endosymbionts. *Mol Biol Evol* 28(1):685-97.
Leclercq S, Cordaux R. 2012. Selection-driven extinction dynamics for group II introns in Enterobacteriales. *PLoS One* 7(12):e52268.
Lei Q, Li C, Zuo Z, Huang C, Cheng H, Zhou R. 2016. Evolutionary Insights into RNA trans-Splicing in Vertebrates. *Genome Biol Evol* 8(3):562-77.
Li-Pook-Than J, Bonen L. 2006. Multiple physical forms of excised group II intron RNAs in wheat mitochondria. *Nucleic Acids Res* 34(9):2782-90.
Li H, Wang J, Mor G, Sklar J. 2008. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science* 321(5894):1357-61.
Lynch M, Burger R, Butcher D, Gabriel W. 1993. The mutational meltdown in asexual populations. *J Hered* 84(5):339-44.
Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, Pavlov A, Pavlova N, Karamychev V, Polouchine N et al. . 2006. Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci U S A* 103(42):15611-6.
McInnes RS, McCallum GE, Lamberte LE, van Schaik W. 2020. Horizontal transfer of antibiotic resistance genes in the human gut microbiome. *Curr Opin Microbiol* 53:35-43.
Mills S, McAuliffe OE, Coffey A, Fitzgerald GF, Ross RP. 2006. Plasmids of lactococci - genetic accessories or genetic necessities? *FEMS Microbiol Rev* 30(2):243-73.
Mills S, O'Sullivan O, Hill C, Fitzgerald G, Ross RP. 2010. The changing face of dairy starter culture research: From genomics to economics. *International Journal of Dairy Technology* 63(2):149-170.
Molina-Sanchez MD, Martinez-Abarca F, Toro N. 2006. Excision of the Sinorhizobium meliloti group II intron RmInt1 as circles in vivo. *J Biol Chem* 281(39):28737-44.
Monat C, Quiroga C, Laroche-Johnston F, Cousineau B. 2015. The Ll.LtrB intron from Lactococcus lactis excises as circles in vivo: insights into the group II intron circularization pathway. *RNA* 21(7):1286-93.
Monat C, Cousineau B. 2016. Circularization pathway of a bacterial group II intron. *Nucleic Acids Res* 44(4):1845-53.

Morl M, Schmelzer C. 1990a. Group II intron RNA-catalyzed recombination of RNA in vitro. *Nucleic Acids Res* 18(22):6545-51.
Morl M, Schmelzer C. 1990b. Integration of group II intron bI1 into a foreign RNA by reversal of the self-splicing reaction in vitro. *Cell* 60(4):629-36.
Murray HL, Mikheeva S, Coljee VW, Turczyk BM, Donahue WF, Bar-Shalom A, Jarrell KA. 2001. Excision of group II introns as circles. *Mol Cell* 8(1):201-11.
Nilsen TW. 1993. Trans-splicing of nematode premessenger RNA. *Annu Rev Microbiol* 47:413-40.
Nilsen TW. 2001. Evolutionary origin of SL-addition trans-splicing: still an enigma. *Trends Genet* 17(12):678-80.
Nisa-Martinez R, Jimenez-Zurdo JI, Martinez-Abarca F, Munoz-Adelantado E, Toro N. 2007. Dispersion of the RmInt1 group II intron in the Sinorhizobium meliloti genome upon acquisition by conjugative transfer. *Nucleic Acids Res* 35(1):214-22.
Pansegrau W, Schroder W, Lanka E. 1994. Concerted action of three distinct domains in the DNA cleaving-joining reaction catalyzed by relaxase (TraI) of conjugative plasmid RP4. *J Biol Chem* 269(4):2782-9.
Passerini D, Beltramo C, Coddeville M, Quentin Y, Ritzenthaler P, Daveran-Mingot ML, Le Bourgeois P. 2010. Genes but not genomes reveal bacterial domestication of Lactococcus lactis. *PLoS One* 5(12):e15306.
Paulsen IT, Banerjei L, Myers GS, Nelson KE, Seshadri R, Read TD, Fouts DE, Eisen JA, Gill SR, Heidelberg JF et al. . 2003. Role of mobile DNA in the evolution of vancomycin-resistant Enterococcus faecalis. *Science* 299(5615):2071-4.
Qu G, Kaushal PS, Wang J, Shigematsu H, Piazza CL, Agrawal RK, Belfort M, Wang HW. 2016. Structure of a group II intron in complex with its reverse transcriptase. *Nat Struct Mol Biol* 23(6):549-57.
Qu G, Piazza CL, Smith D, Belfort M. 2018. Group II intron inhibits conjugative relaxase expression in bacteria by mRNA targeting. *Elife* 7.
Sharp PA. 1991. "Five easy pieces". *Science* 254(5032):663.
Siezen RJ, Renckens B, van Swam I, Peters S, van Kranenburg R, Kleerebezem M, de Vos WM. 2005. Complete sequences of four plasmids of Lactococcus lactis subsp. cremoris SK11 reveal extensive adaptation to the dairy environment. *Appl Environ Microbiol* 71(12):8371-82.
Simon DM, Zimmerly S. 2008. A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res* 36(22):7219-29.
Smillie C, Garcillan-Barcia MP, Francia MV, Rocha EP, de la Cruz F. 2010. Mobility of plasmids. *Microbiol Mol Biol Rev* 74(3):434-52.
Staddon JH, Bryan EM, Manias DA, Chen Y, Dunny GM. 2006. Genetic characterization of the conjugative DNA processing system of enterococcal plasmid pCF10. *Plasmid* 56(2):102-11.
Tourasse NJ, Kolsto AB. 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(14):4529-48.
Wagner A. 2006. Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. *Mol Biol Evol* 23(4):723-33.
Whitaker N, Chen Y, Jakubowski SJ, Sarkar MK, Li F, Christie PJ. 2015. The All-Alpha Domains of Coupling Proteins from the Agrobacterium tumefaciens VirB/VirD4 and Enterococcus

faecalis pCF10-Encoded Type IV Secretion Systems Confer Specificity to Binding of Cognate DNA Substrates. *J Bacteriol* 197(14):2335-49.
Yu CY, Liu HJ, Hung LY, Kuo HC, Chuang TJ. 2014. Is an observed non-co-linear RNA product spliced in trans, in cis or just in vitro? *Nucleic Acids Res* 42(14):9410-23.

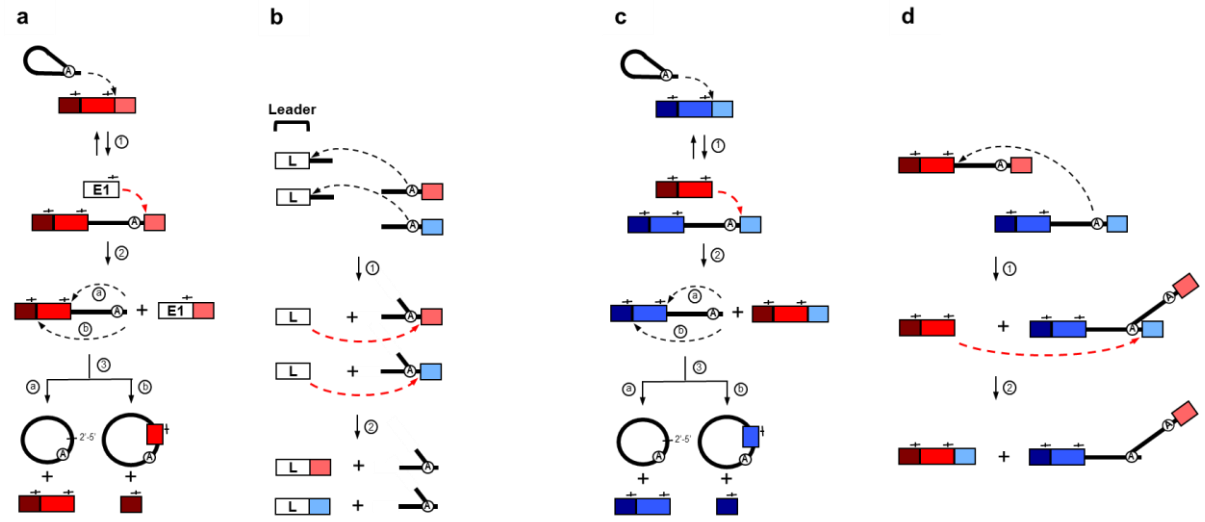
5.5: Figures

Figure 5.1



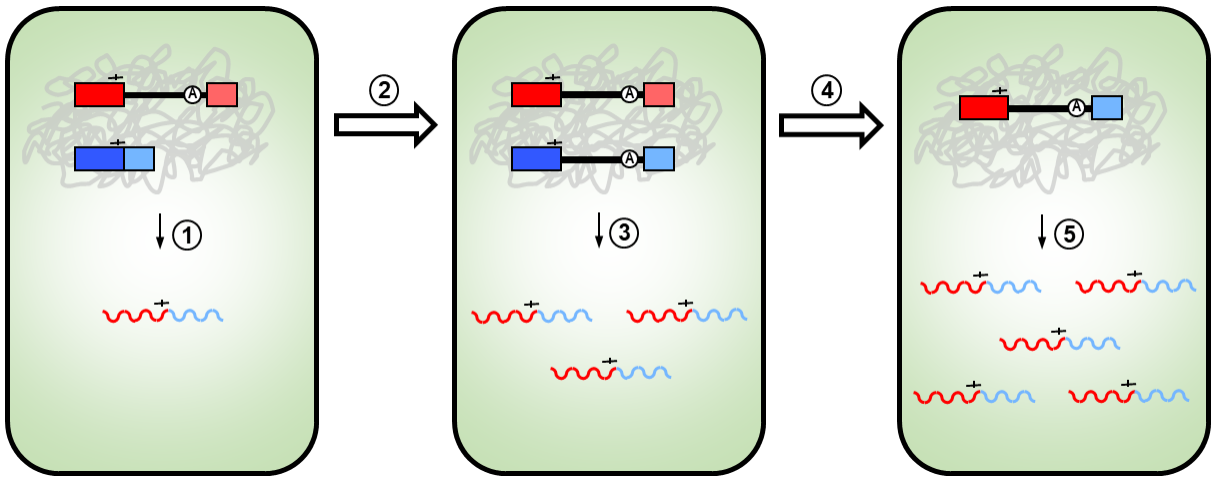
Parsimonious progression of mutations between Ef.PcfG and Ll.LtrB. (**a**) Dendrogram of point mutation accumulation between Ef.PcfG from *E. faecalis* (bold, tree root) and the various Ll.LtrB variants identified from *L. lactis* including the model intron Ll.LtrB (bold). Dashed line indicates a new addition to the previously published dendrogram (LaRoche-Johnston et al. 2016), where pUC77B (underlined) adds increased resolution. (**b**) Likely progression of point mutation acquisition from Ef.PcfG to Ll.LtrB. Empty arrows indicate the acquisition of a new point mutation. Boxes under each mutation (numbered circles) indicate the respective order of appearance of that mutation.

Figure 5.2



Similarities between *trans*-splicing reactions in bacteria and eukaryotes that lead to increased genetic diversity. (**a**) Group II intron *trans*-splicing pathway that generates E1-mRNA intergenic chimeras. Group II introns can increase genetic diversity by first completely reverse splicing into an IBS1/2-like recognition site (—|—) of an mRNA target (step 1). A cognate exon 1 (E1) can be used as an external nucleophile by the group II intron to attack the 3′ splice site, releasing an E1-mRNA chimera (step 2). The intron 3′ end is free to circularize either as a perfect head-to-tail intron circle (step 3a) or to an upstream target site, trapping a stretch of additional molecules at the intron circle splice junction (step 3b). (**b**) Eukaryotic Spliced Leader (SL) *trans*-splicing pathway. During SL *trans*-splicing, most transcripts contain the adenosine branchpoint and 3′ ends of typical introns but lack a conventional 5′ end. Splicing occurs using a conserved Spliced Leader RNA, which contains the upstream Leader "exon" followed by a downstream typical 5′ splice site. The first transesterification occurs when the branchpoint targets the 5′ splice site of SL RNAs, liberating the upstream Leader and forming a "Y"-shaped branched intron (step 1). The Leader RNA next attacks the 3′ splice site of various mRNAs, generating a pool of L-mRNA molecules (step 2). (**c**) Group II intron *trans*-splicing pathway that generates mRNA-mRNA intergenic chimeras. Group II introns can also increase genetic diversity by first completely reverse splicing into an IBS1/2-like recognition site of an mRNA target (step 1). By using another mRNA molecule bearing an IBS1/2-like motif at its 3′ end as an external nucleophile to attack the 3′ splice site, group II introns can generate mRNA-mRNA chimeras (step 2). The intron 3′ end is free to circularize either as a perfect head-to-tail intron circle (step 3a) or to an upstream target site, trapping a stretch of additional molecules at the intron circle splice junction (step 3b). (**d**) Eukaryotic spliceosomal intergenic *trans*-splicing. During intergenic *trans*-splicing, the spliceosome catalyzes a branching reaction where the adenosine branchpoint attacks a *trans* 5′ splice site on a different mRNA transcript rather than its own *cis* 5′ splice site (step 1). This enables the liberated 3′ OH of the foreign mRNA to attack *in trans* the 3′ splice site, generating an mRNA-mRNA chimera (step 2). Mechanistic similarities between the E1-mRNA (**a**) and SL *trans*-splicing pathways (**b**) as well as between the mRNA-mRNA and intergenic *trans*-splicing pathways are shown as red arrows.

Figure 5.3



Chimera formation under various cellular contexts. The left panel denotes a cellular environment in which a single intron copy is present (red gene atop grey chromosome). When an orthologous gene (blue) containing an intron-recognition site (—|—) is also present, low levels of chimeric mRNAs will be generated by the intron through *trans*-splicing (step 1, red-blue). If chimeras produced by the intron are beneficial to the host, intron mobility events into the orthologous gene undergo positive selection (step 2), since these bacterial cells will express higher amounts of beneficial chimeric mRNAs through *trans*-splicing (step 3). Once multiple intron copies are present, the likelihood of generating a chimeric gene increases (step 4), either through homologous recombination or through an intron-generated chimeric cDNA intermediate. These chimeric genes may also undergo positive selection, since their transcription directly produces beneficial chimeras through self-splicing of the intervening intron copy (step 5).

Appendix

Permissions to use copyright material

Chapter 1, Figure 1.1:

Genomic structure and RNA secondary structure of the Ll.LtrB group II intron, from "McNeil BA, Semper C, Zimmerly S. 2016. Group II introns: versatile ribozymes and retroelements. *Wiley Interdiscip Rev RNA*". Permission was granted by John Wiley and Sons and Copyright Clearance Center, provided that an appropriate acknowledgement is given to the author, title of the material/book/journal and the publisher; as well as inclusion of the copyright notice that appears in the Wiley publication.

Chapter 1, Figure 1.2:

Group II intron lineages, from "Lambowitz AM, Zimmerly S. 2011. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol* 3(8):a003616". Permission was granted by Cold Spring Harbor Laboratory Press, provided that a complete reference and copyright are included.

Chapter 1, Figure 1.5:

Comparison of intron RNA and ORF phylogenies, from "Simon DM, Kelchner SA, Zimmerly S. 2009. A broadscale phylogenetic analysis of group II intron RNAs and intron-encoded reverse transcriptases. *Mol Biol Evol* 26(12):2795-808". Permission was granted by Oxford University Press and Copyright Clearance Center, provided that the author and work are properly cited.

Chapter 2:

Recent horizontal transfer, functional adaptation and dissemination of a bacterial group II intron, from "LaRoche-Johnston F, Monat C, Cousineau B. 2016. *BMC Evol Biol* 16(1):223". This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. All co-authors agreed to inclusion of this manuscript to the present thesis.

Chapter 3:

Bacterial group II introns generate genetic diversity by circularization and trans-splicing from a population of intron-invaded mRNAs, from "LaRoche-Johnston F, Monat C, Coulombe S, Cousineau B. 2018. *PLoS Genet* 14(11):e1007792". This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. All co-authors agreed to inclusion of this manuscript to the present thesis.

Chapter 4:

Group II introns generate functional chimeric relaxase enzymes with modified specificities through exon shuffling at both the RNA and DNA level, from "LaRoche-Johnston F, Bosan R, Cousineau B. 2020. *Mol Biol Evol*". This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. All co-authors agreed to inclusion of this manuscript to the present thesis.