An evolutionary and biochemical characterization of a self-splicing group II intron and its encoded LAGLIDADG homing endonuclease in *Leptographium truncatum*

by

Sahra-Taylor Mullineux

A thesis submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfilment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Department of Microbiology

University of Manitoba

Winnipeg, Manitoba

Canada

ABSTRACT

Evolutionary relationships amongst strains of the fungal genus *Leptographium* and related taxa were inferred using the internal transcribed spacer (ITS) region of the nuclear ribosomal DNA repeat. To generate robust sequence alignments for phylogenetic analysis the relationship between DNA sequence variability and RNA structural conservation of ITS segments was examined. The results demonstrate that structural conservation of helical regions is facilitated by compensatory base changes, compensating insertions/deletions, and, possibly, RNA strand slippage. A high mol % G+C bias for ITS1 and ITS2 and structural constraints at the RNA level appear to limit the types of changes observed.

Fifty strains of *Leptographium* were screened for the presence of introns within mitochondrial genes. Superimposing intron survey data onto the ITS-derived phylogenetic tree reveals that introns are absent from the small ribosomal RNA (*rns*) gene of all strains of *L. procerum* yet are found in all strains of *L. lundbergii*. Amongst members of *L. wingfieldii*, *L. terebrantis*, and *L. truncatum* intron distribution is stochastic and is not correlated to the evolutionary relationships amongst strains.

A group II intron/LAGLIDADG homing endonuclease gene (HEG) composite element from the mt *rns* gene of *L. truncatum* strain CBS929.85 was characterized. Intron-catalyzed splicing was tested using ORF-less and ORF-containing precursor transcripts, and both versions of the intron readily self-splice under moderate temperature and ionic conditions (37 °C and 6 mM $MgCl_2$). Cleavage activity of the intron-encoded protein (I-LtrII) was tested using an N-terminal His_6-tagged and near native protein. The homing endonuclease cleaves double-stranded DNA 2 nucleotides

upstream of the intron insertion site within the exon, generating 4 nucleotide 3' OH overhangs. Intron splicing is not enhanced by the addition of I-LtrII and RNA-binding assays indicate that the His_6-tagged protein does not bind to the intron. Phylogenetic relationships amongst the *rns* gene, intron, and amino acid sequences were inferred. An evolutionary model of the composite element is proposed in which the HEG invaded a group II intron and mobilized it. The mobile genetic element may be transmitted vertically amongst *L. lundbergii* strains and horizontally through lateral gene transfer amongst strains of *L. wingfieldii*, *L. terebrantis*, and *L. truncatum*.

ACKNOWLEDGEMENTS

Financial support was provided by: grants from the Natural Sciences and Engineering Research Council (NSERC) and the University of Manitoba's University Research Grants Program to Dr. Georg Hausner; graduate scholarships awarded to S.T. Mullineux (NSERC Postgraduate Scholarship M award, 2004-2006; University of Manitoba Graduate Fellowship, 2006-2007; Manitoba Health Research Council, 2007-2008); and the Faculties of Science and Graduate Studies (University of Manitoba).

I would like to express my deepest gratitude to Dr. Georg Hausner for giving me the opportunity to study in his lab. I thank you for your outstanding patience, support, encouragement, and guidance; I could not have asked for a better mentor. I wish to thank the members of my committee for their guidance and encouragement: Drs. Peter Loewen, Deborah Court, and Michele Piercey-Normore, as well as my external examiner Dr. Linda Bonen (University of Ottawa). I gratefully acknowledge the invaluable support, patience, and expertise given to me by my excellent collaborators at the Centre de Génétique Moléculaire, Centre National de la Recherche Scientifique (Gif-sur-Yvette, France): Drs. François Michel, Maria Costa, and Gurminder Bassi. I gratefully acknowledge the support from my lab colleagues: Drs. James Reid, Jyothi Sethuraman, and Mahmood Iranpour, Shelly Rudski, and Mohamed Hafez.

Many thanks are due to my professors and my previous lab colleagues: Drs. Ivan Oresnick, Kathleen Londry, Richard Sparling, Betty Worobec, Linda Cameron, Vladimir Yurkov, and Christopher Rathgeber, and Julius Csotonyi – I learned so much from all of you and this helped me to gain confidence in myself as a scientist. In

dition, I wish to express my gratitude to the faculty, staff, and students in the Department of Microbiology, especially Mrs. Sharon Berg, Vikash Jha, Rahul Singh, and Robin Imperial.

To my beloved family and friends: Mum, Elinor, Elisabeth, and Charla; Carmen, Audrey, and Grant; Nicoleta, Darsh, Yang and Tatsuya; Mickey and Shauna: thank you for being there to give me encouragement, for sharing your wisdom, making me laugh, and helping me to look on the bright side of things. Your support and encouragement have gone a long way to make this research possible!

TABLE OF CONTENTS

	Page
Abstract	II
Acknowledgements	IV
Table of contents	VI
List of tables	XIII
List of figures	XV
List of abbreviations	XIX
General introduction	XXI
Chapter 1. Literature review	1
1.1. *Leptographium*: fungal associates of insects and agents of blue-stain	1
1.1.1. General features of members of the fungal genus *Leptographium*	1
1.1.2. Taxonomy of *Leptographium*	6
1.1.3. Phylogenetic studies of *Leptographium* species	7
1.1.4. Association with insect vectors	11
1.1.5. Plant pathology and blue-stain production of members of *Leptographium*	11
1.2. Internal transcribed spacers in the nuclear ribosomal RNA gene cluster	13
1.2.1. Internal transcribed spacers and ribosome biogenesis	13
1.2.2. Sequence and structural features of ITS1 and ITS2	18
1.2.3. Applications of ITS sequences in phylogenetic and barcoding studies	22
1.3. Fungal mitochondrial genomes	24

1.3.1. Conformation and genetic organization of fungal mitochondrial DNA 24
1.3.2. Characterization of codon usage and promoter sequences 26
1.3.3. Optional genetic elements: introns and intron-encoded genes 27
1.3.4. Mitochondrial plasmids 27
1.3.5. Plasmid-like elements 28
1.4. Mobile genetic elements: introns and homing endonucleases 29
1.4.1 Group I and group II introns: the discovery and initial characterization of self-splicing introns 29
1.4.2. Group III and archaeal introns 31
1.4.3. Genes encoding homing endonucleases 31
1.5. Group I introns 32
1.5.1. Intron Distribution 32
1.5.2. RNA secondary and tertiary structure 33
1.5.3. Intron splicing 37
1.6. Group II introns 37
1.6.1. Distribution of group II introns 37
1.6.2. Secondary structure 41
1.6.3. Tertiary structure 43
1.6.4. Mechanisms of intron splicing 46
1.6.5. Intron-encoded proteins 48
1.6.6. Unusual group II introns encoding homing endonuclease genes 50
1.6.7. Host factors involved in intron splicing 51
1.6.8. Intron mobility: retrohoming and retrotransposition 52

1.7. Homing Endonucleases 57

1.7.1. Families of homing endonucleases 57

1.7.2. Nomenclature 57

1.7.3. Structural studies of LAGLIDADG homing endonucleases 58

1.7.4. Role as DNA endonucleases 59

1.7.5. Role as RNA maturases 62

1.8. Applications of group II introns and LAGLIDADG homing endonucleases 63

1.8.1. Genetic manipulation using group II introns and homing endonucleases 63

1.8.2. The development of group II intron gene targeting vectors ("targetrons") 63

1.8.3. Application of homing endonucleases as gene targeting elements 64

1.8.4. Engineering homing endonucleases 65

1.9. Research objectives 66

1.9.1. Identifying constraints that influence the evolution of insertion segments in nuclear ribosomal DNA and assessing the impact on phylogenetic studies 66

1.9.2. Screening for the presence of optional genetic elements in the mitochondrial DNA of *Leptographium* species 67

1.9.3. Biochemical and evolutionary characterization of group II introns and their putative LAGLIDADG homing endonucleases genes 68

Chapter 2. General materials and methods 69

2.1. Strains and growth conditions 69

2.2. DNA extraction and purification 69

2.3. PCR amplification and product purification 70

2.3.1. ITS1-5.8S-ITS2 nuclear rDNA sequences 70

2.3.2. Mitochondrial DNA 71

2.4. Cloning of PCR products, purification of plasmid DNA, and DNA sequencing 81

2.5. Analyses of sequence data 82

2.5.1. ITS1-5.8S-ITS2 nuclear rDNA region in *Grosmannia*, *Leptographium*, and related taxa 82

2.5.2. ITS1-5.8S-ITS2 nuclear rDNA region in *Ophistoma*, *Pesotum*, and related taxa 84

2.5.3. ITS1-5.8S-ITS2 nuclear rDNA region in *Graphium* species 84

2.5.4. The mitochondrial *rns* gene: exon, intron, and ORF sequences 86

2.6. Sequence and structural analyses using logos 90

2.6.1. ITS1-5.8S-ITS2 nuclear rDNA region 90

2.6.2. The mitochondrial *rns* gene and putative LAGLIDADG homing endonucleases 91

2.7. Design of precursor constructs, synthesis and purification of intron-containing transcripts, and self-splicing assays 91

2.8. Expression and purification of I-LtrII 95

2.8.1. N-terminal His_6-tagged I-LtrII 95

2.8.2. Near-native I-LtrII 97

2.9. Maturase and filter-binding assays 98

2.10. Endonuclease assays 99

2.11. Mapping of the cleavage site 101

Chapter 3. The molecular evolution of internal transcribed spacer sequences in nuclear ribosomal RNA genes 103

3.1. Introduction and research objectives 103

3.2. Methods overview 106

3.3. Results 107

3.3.1. Characteristics of the ITS region in *Leptographium* and related taxa 107

3.3.2. Characterizing the DNA sequence of ITS1 and ITS2 with DNA sequence logos 113

3.3.3. Models of RNA secondary structure of ITS1 and ITS2 122

3.3.4. Evolution of the ITS region 122

3.3.5. Possible constraints involved in conservation of RNA structure 133

3.4. Discussion 142

3.4.1. Mechanisms involved in the conservation of the RNA secondary structures of ITS segments 142

3.4.2. GC balance between ITS1 and ITS2 143

3.4.3. The use of the ITS region as a phylogenetic marker 145

Chapter 4. Bioprospecting novel insertions within ribosomal- and protein-coding genes in the mitochondrial DNA of *Leptographium* species 147

4.1. Introduction and research objectives 147

4.2. Methods overview 148

4.3. Results 151

4.3.1. The distribution of insertions within the mitochondrial *rns* gene 151

4.3.2. The presence/absence of insertions within the U7 and U11 regions of the mitochondrial *rnl* gene 157

4.3.3. The presence/absence of insertions within the mitochondrial *cob* gene 160

4.4. Discussion 163

4.4.1. Distribution of insertion elements within mitochondrial genes in *Leptographium* species 163

4.4.2. Evaluation of the use of mitochondrial insertions as a DNA marker for species identification 165

Chapter 5. A biochemical and evolutionary study of a group II intron ribozyme and its encoded LAGLIDADG homing endonuclease 167

5.1. Introduction and research objectives 167

5.2. Methods overview 170

5.3. Results 176

5.3.1. The mitochondrial *rns* gene contains a group IIB1 intron with an LHEG 176

5.3.2. The Lt.SSU/1 intron self-splices *in vitro* in the absence of protein factors 184

5.3.3. The I-LtrII protein does not bind RNA or enhance the efficiency of intron splicing 188

5.3.4. I-LtrII cleaves the mitochondrial *rns* gene upstream of the intron insertion site 194

5.3.5. Phylogenetic analyses of the mitochondrial *rns* gene, intron, and LHEases sequences 202

5.3.6. Sequence characterization of I-LtrII using logos 211

5.4. Discussion 214

5.4.1. Group II introns with LHEGs 214

5.4.2. The Lt.SSU/1 intron self-splices *in vitro* 215

5.4.3. Intron-encoded I-LtrII does not enhance intron splicing 216

5.4.4. I-LtrII targets the mt *rns* gene and is capable of promoting intron mobility 217

5.4.5. Speculations on the origins and evolutionary dynamics of the group II ribozyme/LHEG composite element 217

Chapter 6. Conclusions 221

6.1. Major findings of the thesis 221

6.1.1. Evolutionary dynamics of the ITS1-5.8S-ITS2 nuclear rDNA region 221

6.1.2. Distribution of insertions within mitochondrial DNA in members of the fungal genus *Leptographium* 223

6.1.3. Group II introns and their encoded LAGLIDADG homing endonucleases within the mitochondrial *rns* gene 224

6.1.4. Proposing a model describing the life cycle of group II intron/ LAGLIDADG homing endonuclease genes amongst *Leptographium* species 225

6.2. Future prospectives 231

Chapter 7. Appendix 235

References 286

LIST OF TABLES

Page

Table 1.1. Species of *Leptographium* for which only an anamorphic (asexual) state has been observed. 4

Table 1.2. Species of *Leptographium* for which a teleomorphic (sexual) state has been described. 5

Table 2.1. Primer sequences used for amplification reactions in each of the studies. 74

Table 2.2. Description of amplification conditions for the mt *rns* gene. 75

Table 2.3. Description of amplification conditions for the U11 region of the mt *rnl* gene. 77

Table 2.4. Description of amplification conditions for the mt *cob* gene. 79

Table 2.5. Description of bioinformatics programs used in sequence analyses. 89

Table 3.1. Strains of *Grosmannia*, *Leptographium*, and related taxa used in the ITS study, along with GenBank accession numbers. 108

Table 3.2. Nucleotide composition and GC content of nuclear ITS1 and ITS2 sequences in strains of *Grosmannia*, *Leptographium*, and related taxa. 110

Table 4.1. Strains of *Grosmannia* and *Leptographium* species used in the PCR survey. 149

Table 4.2. PCR survey of insertions within the mitochondrial *rns*, *rnl*, and *cob* genes within strains of *Grosmannia* and *Leptographium*. 153

Table 5.1. Strains of ascomycetous fungi and of *Leptographium* species used

in the analysis of the mt *rns* gene, along with GenBank accession numbers. 172

Table 5.2. Sequences used in the mitochondrial group II intron phylogeny. 174

Table 5.3. Sequences used in the mitochondrial LHEase phylogeny. 175

Table 5.4. Sequence characteristics of group II introns and their putative LHEGs within the mt *rns* gene. 178

Supplementary Table 7.1. Nested primers that were designed specifically for sequencing mt genes. 236

Supplementary Table 7.2. Strains of *Ophiostoma*, *Pesotum*, and related taxa used in this study, along with GenBank accession numbers. 239

Supplementary Table 7.3. Strains of *Graphium* and related taxa used in this study, along with GenBank accession numbers. 241

LIST OF FIGURES

Page

Figure 1.1. Phylogenetic analysis of nuclear ITS1-5.8S-ITS2 rDNA sequences in strains of *Grosmannia*, *Leptographium*, and related fungal taxa. 9

Figure 1.2. The pathway for processing of the 18S, 5.8S, and 25S nuclear ribosomal RNA transcripts in *S. cerevisiae*. 16

Figure 1.3. Pan-eukaryotic four-fingered hand model of ITS2. 19

Figure 1.4. Structure and catalysis of group I introns. 35

Figure 1.5. General structure and catalysis of group II introns. 39

Figure 1.6. Subdivision of group II introns based on differences in secondary structure and tertiary interactions and the arrangement of group II-encoded RT ORFs. 43

Figure 1.7. Group II intron mobility by target DNA-primed reverse transcription. 55

Figure 1.8. Homing of introns and homing endonuclease genes is promoted by the DNA cleavage activity of the homing endonuclease. 60

Figure 2.1. Schematic of the gene regions that were amplified and the primers that were used in the amplification. 72

Figure 3.1. DNA sequence and RNA structural features of the ITS1 segment. 116

Figure 3.2. DNA sequence and RNA structural features of the ITS2 segment. 119

Figure 3.3. Phylogenetic analysis of nuclear ITS1-5.8SrDNA-ITS2 DNA sequences in strains of *Grosmannia*, *Leptographium*, and related fungal taxa. 127

Figure 3.4. Phylogenetic analysis of the ITS1 rDNA segment in strains of *Grosmannia*, *Leptographium*, and related fungal taxa. 129

Figure 3.5. Phylogenetic analysis of the ITS2 rDNA segment in strains of *Grosmannia*, *Leptographium*, and related fungal taxa. 131

Figure 3.6. Hypothesized shifts from one ITS RNA structure to another as a result of compensatory base changes, insertions and deletions, and potential RNA strand slippage. 135

Figure 4.1. The distribution of insertions within the mt *rns* gene of strains of *Grosmannia* and *Leptographium*. 155

Figure 4.2. The distribution of insertions within the U11 region of the mt *rnl* gene of strains of *Grosmannia* and *Leptographium*. 158

Figure 4.3. The distribution of insertions within the mt *cob* gene of strains of *Grosmannia* and *Leptographium*. 161

Figure 5.1. Insertion site of group II introns and model of the RNA secondary structure of group IIB1 introns in *Leptographium*. 180

Figure 5.2. Purification of a LAGLIDADG-type homing endonuclease encoded by a group II intron. 182

Figure 5.3. Gel electrophoresis of *in vitro* splicing reactions by Lt.SSU/1 intron under various temperature and ionic conditions. 186

Figure 5.4. Assaying for maturase activity of the N-terminal His_6-tagged I-LtrII. 189

Figure 5.5. Assaying for enhancement of intron splicing by near-native I-trII. 192

Figure 5.6. Cleavage activity of near-native I-LtrII protein. 195

Figure 5.7. Mapping of the cleavage site of the I-LtrII LHEase within the mt

rns gene. 199

Figure 5.8. Phylogenetic analysis of the mt *rns* gene sequence in strains of *Grosmannia*, *Leptographium*, and related fungal taxa. 203

Figure 5.9. Phylogenetic analysis of the mt *rns* group II intron sequence in strains of *Grosmannia*, *Leptographium*, and related fungal taxa. 206

Figure 5.10. Phylogenetic analysis of LHEase amino acid sequences in strains of *Grosmannia*, *Leptographium*, and related fungal taxa. 209

Figure 5.11. Sequence logo of the 13 LAGLIDADG homing endonucleases encoded by group II introns within the mt *rns* gene. 212

Figure 6.1. A model describing the spread of group II introns/LAGLIDADG homing endonuclease genes in the mt *rns* gene of *Leptographium* species. 229

Figure 6.2. Outline of the proposed experiment used to demonstrate I-LtrII-promoted homing of the Lt.SSU/1 intron/LHEG composite element. 233

Supplementary Figure 7.1. Alignment of the DNA sequence of the nuclear ITS1-5.8S-ITS2 rDNA region of strains of *Grosmannia*, *Leptographium*, and related taxa. 242

Supplementary Figure 7.2. Alignment of the DNA sequence of the nuclear ITS1-5.8S-ITS2 rDNA region of strains of *Ophiostoma*, *Pesotum*, and related taxa. 250

Supplementary Figure 7.3. Phylogenetic analysis of nuclear ITS1-5.8S rDNA-ITS2 sequences of *Ophiostoma*, *Pesotum*, and related taxa. 256

Supplementary Figure 7.4. Alignment of the DNA sequence of the nuclear ITS1-5.8S-ITS2 rDNA region of strains of *Graphium* and related taxa. 258

Supplementary Figure 7.5. Phylogenetic analysis of nuclear ITS1-5.8S rDNA-ITS2 sequences of *Graphium* and related taxa. 262

Supplementary Figure 7.6. Alignment of the DNA sequence of the mt *rns* gene. 264

Supplementary Figure 7.7. Alignment of the DNA sequence of the mt *rns* group II intron encoding a putative LHEG. 276

Supplementary Figure 7.8. Alignment of the LHEase amino acid sequence. 279

Supplementary Figure 7.9. (A) Optimized DNA sequence of I-LtrII. (B) Amino acid sequence of I-LtrII, based on the optimized DNA sequence. 282

Supplementary Figure 7.10. Amino acid sequence alignment of putative LHEases in members of *Leptographium*. 284

LIST OF ABBREVIATIONS

atp: ATP synthase gene

BSRD: black-stain root disease

CBC: compensatory base change

cDNA: complementary DNA

CL buffer: cell lysis buffer

CS: cleavage site

cob: cytochrome b gene

cox: cytochrome oxidase gene

cp: chloroplast

DSBR: double strand break repair

DTT: dithiothreitol

EBS: exon binding site

ENase: endonuclease

ETS: external transcribed spacer

(L)HEase: (LAGLIDADG) homing endonuclease

(L)HEG: (LAGLIDADG) homing endonuclease gene

IBS: intron binding site

IEP: intron-encoded protein

IGS: internal guide sequence

indel: insertion/deletion

IPTG: isopropyl β-D-1-thiogalactopyranoside

IS: insertion site

ITS: internal transcribed spacer

LB: Luria Bertani

LSU: large subunit

mRNA: messenger RNA

mt: mitochondrial

nad: NADH dehydrogenase gene

NJ: Neighbor Joining

nt: nucleotide

ORF: open reading frame

PCR: polymerase chain reaction

pLME: plasmid-like element

PRL: progenote rRNA linker

RFLP: restriction fragment length polymorphisms

rnl: large subunit ribosomal RNA gene

RNP: ribonucleoprotein

rns: small subunit ribosomal RNA gene

rRNA: ribosomal RNA

RT: reverse transcriptase

SCID: severe combined immunodeficiency

SDS: sodium dodecyl sulphate

SSU: small subunit

TPRT: target DNA-primed reverse transcription

tRNA: transfer RNA

GENERAL INTRODUCTION

Fungal mitochondrial (mt) DNA is a rich source of unusual elements, such as plasmids, mt plasmid-like elements (pLMEs), self-splicing introns, and homing endonuclease genes (HEGs, Griffiths, 1995; Bullerwell et al., 2003; Hausner, 2003; Monteiro-Vitorello et al., 2009). Variation in the size of mt genomes across fungal taxa is due to differences in the presence or absence of these DNA elements and in the lengths of intergenic spacers (reviewed in Bullerwell et al., 2003; Hausner, 2003).

Circular pLMEs are derived from regions of mt DNA and may accumulate within mitochondria. These elements have been identified in several species of filamentous fungi, including *Neurospora crassa*, *Cryphonectria parasitica*, and *Ophiostoma novo-ulmi*. The latter two organisms are economically significant plant pathogens, causing, respectively, chestnut blight and Dutch elm disease; the accumulation of pLMEs has been shown to be associated with, although not the cause of, a hypovirulent phenotype (Hausner et al., 2006a). A recent study of the mt small subunit ribosomal (*rns*) gene in *C. parasitica* identified an autonomously replicating pLME in hypovirulent and respiratory mutants *mit1* and *mit2*. Sequence analysis showed that the pLME was derived from a region of the *rns* gene that included exon 1 and the majority of the first intron (Monteiro-Vitorello et al., 2009). Group I and group II introns and their associated open reading frames (ORFs) not only contribute towards the size variation of fungal mt DNA but they can also be associated with the formation of plasmids and possibly mt DNA instabilities (Pöggeler and Kempken, 2004; Hausner et al., 2006a). There is a potential link between intron/HEG sequences, mt DNA instabilities, and hypovirulence in fungi (Monteiro-Vitorello et al., 2009).

Mobile genetic elements, including group I and group II introns and HEGs, are wide-spread in bacterial and bacteriophage genomes, in nuclear ribosomal genes (group I introns exclusively), and in ribosomal- and protein-coding genes in the organellar DNA of yeasts and filamentous fungi, algae, plants, and some metazoan animals (Saldanha et al., 1993; Michel and Ferat, 1995; Belfort et al., 2002; Lambowitz and Zimmerly, 2004; Vallès et al., 2008). Group I and group II introns differ from spliceosomal introns by their ability to catalyze their excision from primary transcripts *in vitro* in the absence of protein factors. Group I and group II introns are distinguished from each other based on differences in structure and self-splicing mechanisms and are evolutionarily unrelated (reviewed in Saldanha et al., 1993; Michel and Ferat, 1995; Lambowitz and Zimmerly, 2004).

Some group II introns are mobile "retroelements" (mobile via an RNA intermediate) and code for a multifunctional protein with maturase, DNA endonuclease (ENase), and reverse transcriptase (RT) activities (reviewed in Lambowitz and Belfort, 1993; Michel and Ferat, 1995; Lambowitz et al., 1999; Lambowitz and Zimmerly, 2004). The maturase domain facilitates efficient intron splicing by stabilizing the catalytically active structure and the ENase and RT domains promote intron mobility. During expression of the host gene the intronic ORF is translated and a ribonucleoprotein (RNP) particle is formed between the intron lariat and intron-encoded protein (IEP). In a process referred to as "retrohoming" the intron and its ORF colonize typically intron-less cognate alleles through a site-specific DNA integration mechanism. Retrohoming involves both the intron RNA and the IEP and is completed by host DNA repair mechanisms (Curcio and Belfort, 1996; Cousineau et al., 1998; Lambowitz and Zimmerly, 2004; Smith et al., 2005).

Comparative sequence analyses identified group II introns with ORFs encoding putative LAGLIDADG homing endonucleases (LHEases) in the mt small and large subunit (SSU and LSU, respectively) ribosomal (r) RNA genes of fungi belonging to the Ascomycota and Basidiomycota (Michel and Ferat, 1995; Toor and Zimmerly, 2002; Monteiro-Vitorello et al., 2009). However, neither the intron nor the ORF were functionally characterized in these studies. Genes coding for LHEases are typically associated with group I and archaeal introns or with inteins, or they may be present as free-standing ORFs (Dujon, 1980; Dalgaard et al., 1993; Dürrenberger and Rochaix, 1993; Belfort and Roberts, 1997; Jurica and Stoddard, 1999; Chevalier et al., 2005; Gibb and Hausner, 2005; Stoddard, 2006; Sethuraman et al., 2008, 2009a; Bae et al., 2009; Singh et al., 2009). Therefore, the association of a group II intron and LHEG is atypical, and the evolutionary and biological aspects of their relationship are unknown.

LHEase proteins are named for the LAGLIDADG amino acid α-helical motif that comprises each half of the enzyme's active site. LHEases recognize and bind to long (greater than 20 bp) DNA target sites and exhibit flexibility in sequence recognition (Aagaard et al., 1997; Jurica et al., 1998; Lucas et al., 2001; Moure et al., 2003; Eastberg et al., 2007). LHEases promote homing by generating a double-stranded staggered cut in DNA, and the break is repaired by the host's double-stranded break repair processes using the intron/LHEG-containing allele as a template (reviewed in Belfort and Roberts, 1997; Belfort et al., 2002; Stoddard, 2006; Edgell, 2009). Some LHEases have been shown to function as maturases and promote the splicing of their host group I intron and occasionally related introns (Lazowska et al., 1989; Ho et al., 1997; Ho and Waring, 1999; Bassi et al., 2002; Bassi and Weeks, 2003; Belfort, 2003; Longo et al., 2005).

In this thesis, insertions, introns in particular, in the mt genome were examined in fifty strains belonging to the genus *Leptographium*, filamentous asexual fungi associated with members of the kingom Ascomycota. Some species of *Leptographium* are of particular interest to the forestry industry in Canada; members of this group of fungi, which are vectored by bark beetles, are agents of blue-stain and some species are notable plant pathogens (Harrington, 1993; Jacobs and Wingfield, 2001; Hausner et al., 2005).

This research program comprised three major studies. The first study involved an examination of the molecular evolution of nuclear internal transcribed spacer (ITS) 1 and ITS2. The DNA sequences of the ITS segments evolve rapidly, making this region useful for resolving evolutionary relationships between closely-related species and strains. Although ITS sequences are widely-used in phylogenetic and taxonomic studies, little is known about their mode of evolution. Despite hypervariability at the DNA level, models of RNA secondary structure indicate that structural features may be conserved across higher-order taxa (Mai and Coleman, 1997; Schultz et al., 2005; Hausner and Wang, 2005; Beiggi and Piercey-Normore, 2007; Coleman, 2007; Mullineux and Hausner, 2009), suggesting functional constraints may be involved in the evolution of the DNA sequence. The analysis of ITS segments, including the use of conserved features of RNA secondary structure to guide the DNA sequence alignment of ambiguous regions, has become an important tool in phylogenetic analyses (Coleman, 2003, 2007; Goertzen et al., 2003; Krüger and Gargas, 2008; Schultz et al., 2006; Young and Coleman, 2004; Aguilar and Sánchez, 2007). The major objective of the study was to document the types of nucleotide (nt) changes occurring in the ITS segments and describe the effects on RNA secondary structure.

In the second study, ribosomal- and protein-coding genes in the mt DNA of fifty strains of *Leptographium* were screened for the presence or absence of introns using the polymerase chain reaction (PCR). The objectives of the survey were: (i) to document the distribution pattern and size variation of insertions and (ii) to determine if their presence or absence correlated to the evolutionary relationships of the host strain. The potential evolutionary affiliations between these elements and their host gene/organism and the origins of these elements (via horizontal/vertical transmission, random gain/loss) were examined within the phylogenetic framework established by the ITS study.

The final component was a biochemical and evolutionary study of one group II intron and its putative LAGLIDADG-type ORF. Self-splicing of the intron was tested *in vitro* in the absence of protein factors and DNA cleavage activity of the LHEase was examined. This specific component represents groundbreaking, fundamental research, for few group II introns have been shown to self-splice *in vitro* in the absence of protein factors under moderate temperature and ionic conditions. Group II introns with putative LHEGs were briefly discussed by Ferat and Michel (1995), yet they were not reported on again until the studies by Toor and Zimmerly (2002) and Monteiro-Vitorello et al. (2009), and the activities of the intron and the putative LHEase were not examined at that time. To the best of my knowledge, this research constitutes the first detailed evolutionary and functional analysis of group II introns that encode proteins of the LAGLIDADG family rather than the RT family and will shed light on the nature of the evolutionary and biological relationships between group II introns and LHEGs.

CHAPTER 1. LITERATURE REVIEW

1.1. *Leptographium*: fungal associates of insects and agents of blue-stain

1.1.1. General features of members of the fungal genus *Leptographium*

Members of the genus *Leptographium* Lagerb. & Melin are filamentous, obligately aerobic, and asexual (anamorphic) fungi. The genus was established by Lagerberg et al. (1927) to describe penicillately-branched isolates from blue-stained sapwood in Sweden. Genus characteristics include: tolerance to the antibiotic cycloheximide; synthesis of dark mononematous conidiophores; a hyaline (colourless) conidiogenous apparatus with multiple cylindrical branches; and hyaline aseptate conidia produced by sympodial or annelidic proliferation. Conidia accumulate in a sticky aggregate at the apex of conidiophores (Lagerberg et al., 1927; Wingfield, 1993; Jacobs and Wingfield, 2001; Jacobs et al., 2001). At the time of its original description the genus comprised the sole species *Leptographium lundbergii* Lagerb. & Melin, which represents the type species of the genus (Lagerberg et al., 1927; Jacobs et al., 2005; Zipfel et al., 2006).

There are more than sixty species of *Leptographium* currently listed in the literature (Wingfield, 1993; Jacobs et al., 2001; Zipfel et al., 2006; Index Fungorum, 2010). Approximately forty species represent true anamorphs, as no teleomorphic, or sexual, state has been observed (Table 1.1). Both the teleomorphic and anamorphic states have been described in more than twenty species, and the *Leptographium* epithet is applied to the anamorphic form only (Table 1.2). Because of the absence of a sexual state; that is, the absence of meiotic spores, members of *Leptographium* representing true anamoraphs are called imperfect fungi. However, phylogenetic

analyses of *Leptographium* spp. indicate that members of the genus are allied to ophiostomatoid fungi (Benny and Kimbrough, 1980), and teleomorphic fungi that have anamorphic *Leptographium* states have been placed in the genus *Grosmannia* Goid. (Jacobs et al., 2001; Zipfel et al., 2006). The genus *Grosmannia* belongs to the order *Ophiostomatales* (Ascomycota, Hibbett et al., 2007), and traditionally species of *Grosmannia* were accommodated in the genus *Ophiostoma* (Upadhyay, 1981), which includes such notable plant pathogens as *O. novo-ulmi*, the fungus responsible for Dutch elm disease (Gibb and Hausner, 2005; Sethuraman et al., 2008).

The distinction of taxa based on anamorphic and teleomorphic states can be rather complicated. For example, strains of *Leptographium wageneri* (Kendr.) Wingf., which are formally described as varieties, are unusual for the reason that several varieties are exclusively anamorphic and one variety has a teleomorphic state. *Leptographium wageneri* var. *ponderosae* Harr. & Cobb has a teleomorph in the genus *Grosmannia* (*Grosmannia wageneri* Goheen & Cobb), whereas *L. wageneri* var. *wageneri* Kendr. and *L. wageneri* var. *pseudotsugae* Harr. & Cobb are anamorphs (Zipfel et al., 2006). These varieties are also distinguished by isozyme profiles, fine morphological features, and pathogenic characteristics.

Members of *Leptographium* have a broad geographical distribution; strains have been isolated from New Zealand, Asia, Africa, Scandanavia, the United Kingdom, Europe, the United States, and Canada (Harrington, 1988; Hausner et al., 2000; Jacobs et al., 2001; Hausner et al., 2005). Fungal samples have been recovered from the tissue of softwoods, including several pine species (*Pinus sylvestris*, *P. radiata*, *P. taeda*), and from the pine shoot beetle *Tomicus piniperda* (Hausner et al., 2005). Despite such diverse global origins many of these organisms use common

insect vectors, such as bark beetles, for dispersal (Harrington, 1988; Klepzig and Six, 2004; Hausner et al., 2005; Lu et al., 2009).

Table 1.1. Species of *Leptographium* for which only an anamorphic (asexual) state has been observed.

Specific epithet		
Leptographium abieticolens	*Leptographium euphyes*	*Leptographium pityophilum*
Leptographium abietinum	*Leptographium fruticetum*	*Leptographium procerum*
Leptographium albopini	*Leptographium galleciae*	*Leptographium profanum*
Leptographium alethinum	*Leptographium guttulatum*	*Leptographium pruni*
Leptographium antibioticum	*Leptographium hughesii*	*Leptographium pyrinum*
Leptographium bhutanense	*Leptographium hymenaeae*	*Leptographium reconditum*
Leptographium bistatum	*Leptographium longiclavatum*	*Leptographium sibiricum*
Leptographium brachiatum	*Leptographium lundbergii*	*Leptographium sinoprocerum*
Leptographium calophylli	*Leptographium neomexicanum*	*Leptographium terebrantis*
Leptographium costaricense	*Leptographium peucophilum*	*Leptographium truncatum*
Leptographium douglasii	*Leptographium pineti*	*Leptographium wingfieldii*
Leptographium elegans	*Leptographium pini-densiflorae*	*Leptographium yunnanense*
Leptographium eucaluptophilum	*Leptographium piriforme*	

Table 1.2. Species of *Leptographium* for which a teleomorphic (sexual) state has been described.

Specific epithet of anamorph	Specific epithet of teleomorph
Leptographium aenigmaticum	*Grosmannia aenigmatica*
Leptographium americanum	*Grosmannia americana*
Leptographium aureum	*Grosmannia aurea*
Leptographium brevicolle	*Ophiostoma brevicolle*
Leptographium clavigerum	*Grosmannia clavigera*
Leptographium crassivaginatum	*Grosmannia crassivaginata*
Leptographium dryocoetidis	*Grosmannia dryocoetidis*
Leptographium francke-grosmanniae	*Grosmannia francke-grosmanniae*
Leptographium grandifoliae	*Grosmannia grandifoliae*
Leptographium huntii	*Grosmannia huntii*
Leptographium koreanum	*Ophiostoma koreanum*
Leptographium laricis	*Grosmannia laricis*
Leptographium leptographioides	*Grosmannia leptographioides*
Leptographium microsporum	*Ophiostoma microsporum*
Leptographium penicillatum	*Grosmannia penicillata*
Leptographium piceaperdum	*Grosmannia piceaperda*
Leptographium robustum	*Grosmannia robusta*
Leptographium serpens	*Grosmannia serpens*
Leptographium trinacriforme	*Ophiostoma trinacriforme*
Leptographium wageneri	*Grosmannia wageneri*
Leptographium valdivianum	*Ophiostoma valdiviana*
Leptographium yunnanense	*Grosmannia yunnanense*

1.1.2. Taxonomy of *Leptographium*

Historically fungal taxonomy relied heavily on morphological characters to identify strains and define taxa. For *Leptographium*-like strains taxonomically-relevant characters have traditionally included: size and shape of conidia; number and arrangement of metulae; size and branching pattern of the stipe; hyphal growth patterns; and the colour and form of colonies (Zambino and Harrington, 1992; Wingfield, 1993). The caveat to a morphology-based taxonomic system is that evolutionarily-unrelated organisms may have acquired similar morphological characters due to convergent evolution, meaning that the shared character is the result of an adaptation to similar niches rather than having been derived from a common ancestor. Potential examples of convergent evolution amongst *Leptographium* and *Leptographium*-like isolates include conidial development and morphology, since the shape, size, and sticky quality of the conidia are well-adapted for insect dispersal. Thus, organisms that rely on, or whose survival is enhanced by, insect vectors would be under strong selective pressure to evolve or maintain "*Leptographium*"-like conidiophores and conidia (Wingfield, 1993; Hausner et al., 2000). There has been considerable confusion surrounding the taxonomy of *Leptographium* isolates, especially *L. lundbergii*, due to the lack of herbarium samples from the original specimen of *L. lundbergii* that was described by Lagerberg and Melin (Lagerberg et al., 1927). A new strain, CBS352.29, which was collected in 1929, was subsequently designated as a neotype of *L. lundbergii* Lagerb. & Melin sensu Jacobs & Wingfield (Jacobs and Wingfield, 2001).

1.1.3. Phylogenetic studies of *Leptographium* species

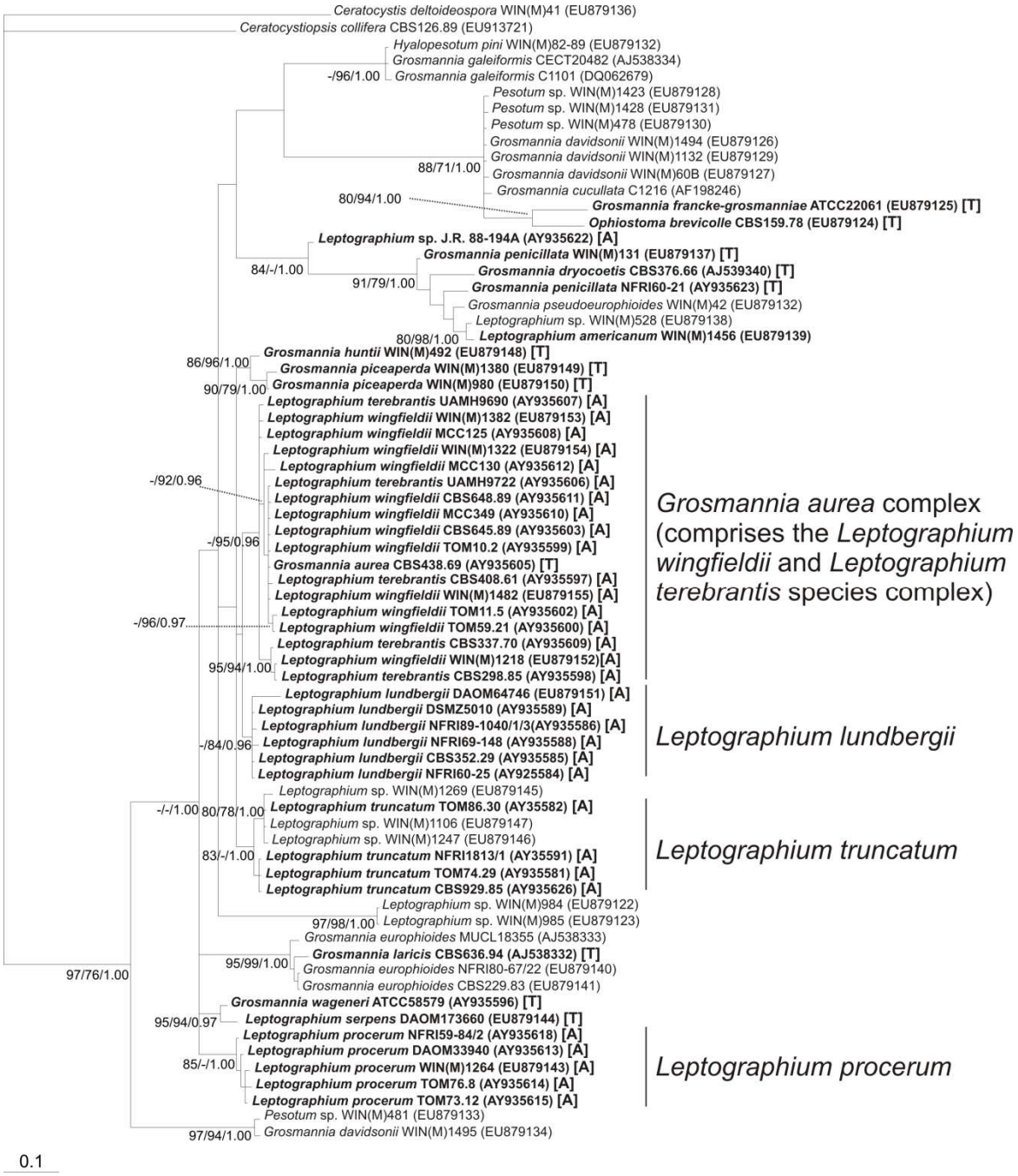
A number of studies have sought to clarify the appropriate classification of both genuine anamorphic strains of *Leptographium* and isolates with teleomorphic representatives in the genus *Grosmannia*. Using a combination of isozyme and DNA sequence analyses, these studies identified morphological characters that are useful for the taxonomy of *Leptographium*-like isolates and clarified the evolutionary relationships amongst *Leptographium* spp. (Zambino and Harrington, 1992; Hausner et al., 2000, 2005; Jacobs et al., 2001, 2005; Zipfel et al., 2006). Some of the most significant results were obtained using sequence analyses of nuclear small subunit (SSU) and large subunit (LSU) ribosomal (r) RNA genes and intergenic spacers, including internal transcribed spacers (ITS), although most phylogenetic analyses focussed on the relatively conserved ITS2 segment and partial LSU sequences (Hausner et al., 2000; Jacobs et al., 2001; Jacobs et al., 2005).

A study by Hausner et al. (2000) examined the evolutionary relationships of *Leptographium* spp. representing true anamorphs and those that have teleomorphic representatives in the genus *Ophiostoma* (and now also *Grosmannia*, Zipfel et al., 2006; Hibbett et al., 2007). Phylogenetic inference using partial SSU and LSU rDNA sequences indicated that the anamorphic species have a scattered distribution and are not monophyletic, prompting the authors to suggest that anamorphic isolates may have evolved independently from several different ophiostomatoid holomorphs (teleomorphs) who lost their sexual state (Hausner et al., 2000). A later study by the same group (Hausner et al., 2005) examined the evolutionary relationships of strains of *Leptographium lundbergii* Lagerb. & Melin sensu Jacobs & Wingfield, *L. procerum* (Kendr.) Wingf., *L. terebrantis* Barras & Perry, *L. truncatum* (Wingf. & Marasas) Wingf., and *L. wingfieldii* Morelet using phylogenetic analyses of partial

sequences of a group I intron in the U11 region of the mitochondrial (mt) LSU rRNA (*rnl*) gene and sequence data from the intergenic spacer and ITS (consisting of the 3' end of the SSU-ITS1-5.8S-ITS2-5'end of LSU) regions of nuclear rDNA and analysis of restriction fragment length polymorphisms (RFLP). This was the first study to use the combined data from ribosomal and intron sequences, as well as RFLP analysis, to examine the taxonomy of *Leptographium* isolates. One of the major findings of this study was that evolutionary relationships inferred from sequence analyses did not correspond well to key morphological characters, such as features of the conidiophores, which had been used previously to delineate species. Their results indicated that these *Leptographium* spp. form four major clades represented by species of *L. procerum*, *L. lundbergi*, *L. truncatum*, and the so-called *O. aureum* (now *G. aurea*, Zipfel et al., 2006) species complex, which includes strains of *L. wingfieldii* and *L. terebrantis*. These results were confirmed in a subsequent phylogenetic study that also included teleomorphic representatives (Mullineux and Hausner, 2009), as shown in Figure 1.1.

.

Figure 1.1. Phylogenetic analysis of nuclear ITS1-5.8S-ITS2 rDNA sequences in strains of *Grosmannia*, *Leptographium*, and related fungal taxa. Species of *Leptographium* that represent true anamorphs (that is, no sexual state has been observed) are in bold and marked by [A]. Species of *Grosmannia* for which a *Leptographium* anamorph has been observed are in bold and marked by a [T]. Branch lengths were determined using the Bayesian consensus outfile. Values at the nodes were determined using algorithms implemented by DNA PARS/Tree Puzzle/Mr Bayes programs. "-" indicates the node is absent or the posterior probability or bootstrap value is not well supported (a posterior probability value of less than 0.95 for Bayesian analysis and bootstrap values less than 70 % for parsimony and maximum likelihood analyses). The figure was generated using information from Tables 1.1 and 1.2 and from Figure 3.1 of this thesis.


Ceratocystis deltoideospora WIN(M)41 (EU879136)
Ceratocystiopsis collifera CBS126.89 (EU913721)
Hyalopesotum pini WIN(M)82-89 (EU879132)
Grosmannia galeiformis CECT20482 (AJ538334)
-/96/1.00
Grosmannia galeiformis C1101 (DQ062679)
Pesotum sp. WIN(M)1423 (EU879128)
Pesotum sp. WIN(M)1428 (EU879131)
Pesotum sp. WIN(M)478 (EU879130)
Grosmannia davidsonii WIN(M)1494 (EU879126)
Grosmannia davidsonii WIN(M)1132 (EU879129)
88/71/1.00
Grosmannia davidsonii WIN(M)60B (EU879127)
Grosmannia cucullata C1216 (AF198246)
80/94/1.00
Grosmannia francke-grosmanniae ATCC22061 (EU879125) [T]
Ophiostoma brevicolle CBS159.78 (EU879124) [T]
Leptographium sp. J.R. 88-194A (AY935622) [A]
Grosmannia penicillata WIN(M)131 (EU879137) [T]
84/-/1.00
Grosmannia dryocoetis CBS376.66 (AJ539340) [T]
91/79/1.00
Grosmannia penicillata NFRI60-21 (AY935623) [T]
Grosmannia pseudoeurophioides WIN(M)42 (EU879132)
Leptographium sp. WIN(M)528 (EU879138)
80/98/1.00
Leptographium americanum WIN(M)1456 (EU879139)
Grosmannia huntii WIN(M)492 (EU879148) [T]
86/96/1.00
Grosmannia piceaperda WIN(M)1380 (EU879149) [T]
90/79/1.00
Grosmannia piceaperda WIN(M)980 (EU879150) [T]
Leptographium terebrantis UAMH9690 (AY935607) [A]
Leptographium wingfieldii WIN(M)1382 (EU879153) [A]
Leptographium wingfieldii MCC125 (AY935608) [A]
Leptographium wingfieldii WIN(M)1322 (EU879154) [A]
Leptographium wingfieldii MCC130 (AY935612) [A]
-/92/0.96
Leptographium terebrantis UAMH9722 (AY935606) [A]
Leptographium wingfieldii CBS648.89 (AY935611) [A]
Leptographium wingfieldii MCC349 (AY935610) [A]
Leptographium wingfieldii CBS645.89 (AY935603) [A]
-/95/0.96
Leptographium wingfieldii TOM10.2 (AY935599) [A]
Grosmannia aurea CBS438.69 (AY935605) [T]
Leptographium terebrantis CBS408.61 (AY935597) [A]
Leptographium wingfieldii WIN(M)1482 (EU879155) [A]
Leptographium wingfieldii TOM11.5 (AY935602) [A]
-/96/0.97
Leptographium wingfieldii TOM59.21 (AY935600) [A]
Leptographium terebrantis CBS337.70 (AY935609) [A]
Leptographium wingfieldii WIN(M)1218 (EU879152)[A]
95/94/1.00
Leptographium terebrantis CBS298.85 (AY935598) [A]
Grosmannia aurea complex (comprises the Leptographium wingfieldii and Leptographium terebrantis species complex)
Leptographium lundbergii DAOM64746 (EU879151) [A]
Leptographium lundbergii DSMZ5010 (AY935589) [A]
Leptographium lundbergii NFRI89-1040/1/3(AY935586) [A]
-/84/0.96
Leptographium lundbergii NFRI69-148 (AY935588) [A]
Leptographium lundbergii CBS352.29 (AY935585) [A]
Leptographium lundbergii NFRI60-25 (AY925584) [A]
Leptographium lundbergii
Leptographium sp. WIN(M)1269 (EU879145)
-/-/1.00
80/78/1.00
Leptographium truncatum TOM86.30 (AY35582) [A]
Leptographium sp. WIN(M)1106 (EU879147)
Leptographium sp. WIN(M)1247 (EU879146)
83/-/1.00
Leptographium truncatum NFRI1813/1 (AY35591) [A]
Leptographium truncatum TOM74.29 (AY935581) [A]
Leptographium truncatum CBS929.85 (AY935626) [A]
Leptographium truncatum
Leptographium sp. WIN(M)984 (EU879122)
97/98/1.00
Leptographium sp. WIN(M)985 (EU879123)
Grosmannia europhioides MUCL18355 (AJ538333)
95/99/1.00
Grosmannia laricis CBS636.94 (AJ538332) [T]
Grosmannia europhioides NFRI80-67/22 (EU879140)
Grosmannia europhioides CBS229.83 (EU879141)
97/76/1.00
Grosmannia wageneri ATCC58579 (AY935596) [T]
95/94/0.97
Leptographium serpens DAOM173660 (EU879144) [T]
Leptographium procerum NFRI59-84/2 (AY935618) [A]
Leptographium procerum DAOM33940 (AY935613) [A]
85/-/1.00
Leptographium procerum WIN(M)1264 (EU879143) [A]
Leptographium procerum TOM76.8 (AY935614) [A]
Leptographium procerum TOM73.12 (AY935615) [A]
Leptographium procerum
Pesotum sp. WIN(M)481 (EU879133)
97/94/1.00
Grosmannia davidsonii WIN(M)1495 (EU879134)
0.1

1.1.4. Association with insect vectors

Strains of *Leptographium* have been isolated from several insect species, including the pine shoot beetle *T. piniperda* (Coleoptera: Scolytidae), *Dendroctonus terebrans*, and *D.valens* (Harrington, 1988; Jacobs et al., 2000; Jacobs and Wingfield, 2001; Hausner et al., 2005; Lu et al., 2009). The conidia of *Leptographium* accumulate at the apex of the conidiophores in a sticky mass, which allows the conidia to adhere to the insect, thus facilitating their dispersal. In turn, the beetles are also well-adapted as fungal vectors and some have evolved such specialized structures, referred to as mycangia, as pits (shallow or deep, with and without glands), punctures, and setal brushes used for carrying fungal spores/conidia (Klepzig and Six, 2004). The major targets of beetle invasion/infection are the cambium tissue of soft woods, such as species of pine, although there are some examples of infections on hard wood species (Harrington, 1988; Hausner et al., 2005; Jacobs et al., 2004; Klepzig and Six, 2004; Lu et al., 2009). The beetle commonly penetrates the tree through lesions in the bark and burrows through the wood, feeding on the tissue and forming beetle galleries (tunnels). Fungal conidia have been recovered from the galleries and it is likely that fungi serve as an important source of nutrients for the beetle and its brood (Klepzig and Six, 2004).

1.1.5. Plant pathology and blue-stain production of members of *Leptographium*

Some species of *Leptographium* are known plant pathogens, while others have been isolated from damaged and dying trees and are likely only weakly pathogenic or saprotrophic (Jacobs and Wingfield, 2001). *Leptographium wageneri* is a well-known pathogen and is the cause of black-stain root disease (BSRD) on conifers in the northwestern United States, although European varieties are also sensitive to the

disease. There appears to be host specificity between the tree and the variety of *L. wagneri*. *Leptographium wageneri* var. *wageneri* targets *Pinus monophylla* and *P. edulis* species of pines, while the douglas fir *Pseudotsuga menziesii* is susceptible to infection by *L. wageneri* var. *pseudotsugae*. *Leptographium wageneri* var. *ponderosae* attacks *P. ponderosa*, *P. contorta*, and *P. jeffreyi*. The three varieties of *L. wageneri* can be distinguished by mophology, by the presence of a teleomorphic state in *L. wageneri* var. *ponderosa* (*G. wageneri*), isozyme profiles, and random amplified polymorphic markers (Jacobs and Wingfield, 2001). The key symptoms of BSRD include: needle chlorosis, due to the production of phytotoxins, which inhibit photosynthesis and transpiration; needle cast, caused by black fruiting bodies on needles; resinous lesions on lower stems; and stunted growth. One of the most characteristic symptoms is the production of a streaky staining pattern by *L. wageneri*; the streaky pattern is the result of fungal hyphae present in the tracheids, rather than the parenchyma (Jacob and Wingfield, 2001).

Leptographium procerum is associated with the disease white pine root decline, which affects *P. strobus* (eastern white pine), *P. resinosa* (red pine), and *P. sylvestris* (Scots Pine). The major vector of the fungus is the weevil (Coleoptera: Curculionidae). Fungal infection can occur concomitantly with insect infestation and/or after elimination of the insect. Symptoms of the disease include: browning and wilting of needles, resinous black-streaked wood at the bases of stems, and basal cankers.

Some species of *Leptographium*, including *L. lundbergii*, *L. terebrantis*, and *L. wingfieldii*, are known agents of blue-strain and sapwood that has been inoculated with the fungi may develop a bluish discoloration. These fungi have been observed to colonize the trees after the xylem has been damaged (Jacobs and Wingfield, 2001).

While the blue stain itself does not cause death of the tree, it can interfere with transpiration. In general, trees that are most susceptible to damage are those under nutritional or environmental stress. Infection of healthy trees by the beetles stimulates production of resin, which "pitches out" the beetle (Hausner et al., 2005). Production of a blue stain on trees and stored lumber by *Leptographium* spp. diminishes the value of the wood, which, coupled to the pathogenic effects of some species, can lead to economic losses to the forestry and lumber industries, and some countries may refuse to import stained lumber.

While pathogenicity has been established in several species, including *L. wageneri* and *L. procerum*, some species, notably *L. wingfieldii*, still require further investigation to confirm this. Further, it is possible that *L. wingfieldii* is a species complex comprising several distinct species (in terms of the phylogenetic species concept). Thus, the identification and characterization of novel genetic markers in the DNA, including mt DNA, of these organisms may be useful to support the identification of strains as probable pathogens (based on high degrees of similarity with genuine pathogens at specific genetic loci) and also to recover the extent of genetic diversity of these strains and clarify their appropriate taxonomic positions.

1.2. Internal transcribed spacers in the nuclear ribosomal RNA gene cluster

1.2.1. Internal transcribed spacers and ribosome biogenesis

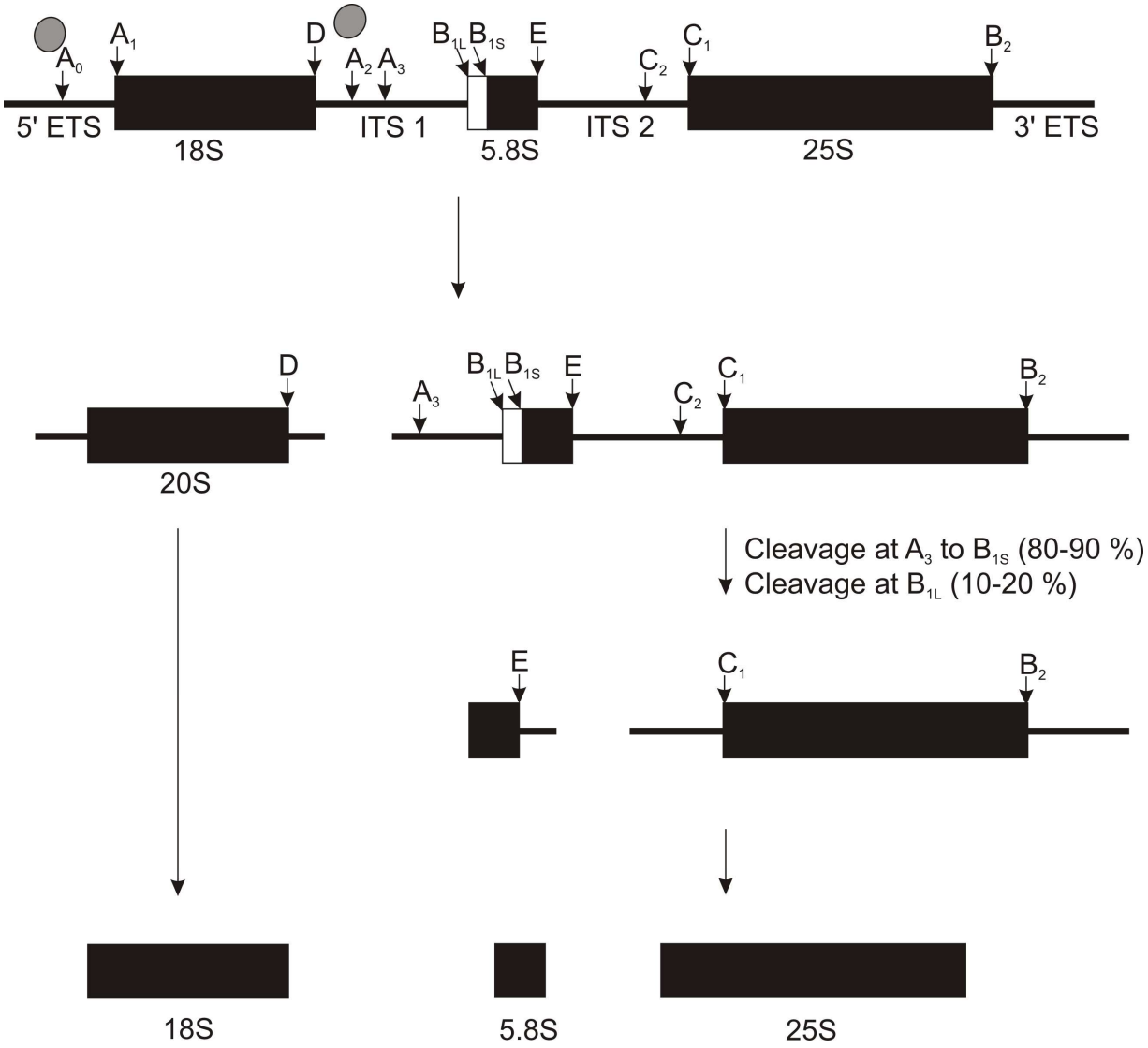
In the majority of eukaryotes the nuclear ribosomal (r) RNA genes are arranged in tandem repeats comprising the 5' external transcribed spacer (ETS) segment, 16-18S rRNA gene, ITS1, 5.8S rRNA gene, ITS2, 25-28S rRNA gene, and 3' ETS. This repeat is transcribed in the nucleolus by RNA polymerase I into a 35-

45S rRNA precursor (Good et al., 1997; Lalev and Nazar, 1999; Lafontaine, 2004; Wery et al., 2009). Several exceptions to the tandem arrangement of rRNA genes have been found: the SSU and LSU genes in species of the protozoan *Plasmodium* are not arranged in tandem repeats (Clark, 1987), and in the microsporidia, which are amitochondriate and obligately intracellular parasites, the ITS2 segment is absent and the 5.8S gene is instead fused to the LSU gene (Keeling and Fast, 2002). In general, the gene encoding the 5S rRNA, also a component of the mature ribosome, is translated by RNA polymerase III rather than RNA polymerase I (reviewed in Lafontaine, 2004).

This precursor is processed primarily in the nucleolus and the components are then exported to the cytoplasm where the remaining processing and assembly steps occur, yielding the mature 40S and 60S ribosomal subunits. Biochemical evidence suggests that the ITS segments are indispensable in rRNA processing because they promote cleavage domains to be arranged in close proximity and they contain binding sites for nucleolar proteins involved in ribosome biogenesis (van der Sande et al., 1992; Lafontaine and Tollervey, 1995; van Nues et al., 1995; Lalev et al., 2000; Nazar 2003, 2004). The ITS segments contain processing sites (A_2 and A_3 in ITS1 and C_1 and C_2 in ITS2) that are cleaved during rRNA processing and data have also shown that there is an interdependence of the ETS and ITS regions during maturation of the nucleolar rRNA precursor particle (Good et al., 1997; Lalev and Nazar 1998, 1999, 2001; Lalev et al., 2000; Abeyrathne et al., 2002). A large number of protein factors are involved in rRNA processing and ribosome assembly: endo- and exonucleases; putative ATP-dependent RNA helicases; chaperones; and small nucleolar RNP particles (Lalev et al., 2000; Fromont-Racine et al., 2003; Nazar, 2003).

The pathway of rRNA processing in *Saccharomyces cerevisiae* is summarized in Figure 1.2. During ribosome biosynthesis in yeast spliceosomal proteins bind to cleavage sites in the 5' and 3' ETS regions (A_0 and A_1, and B_2, respectively), and in the ITS1 (A_2 and A_3) and ITS2 (C_1 and C_2) segments (Lafontaine and Tollervey, 1995; Fromont-Racine et al., 2003; Lafontaine, 2004; Wery et al., 2009). Cleavage within the 5' ETS and ITS1 (A_0 to A_2) segments generate a 20S product, which is ultimately dimethylated and cleaved at the 3' terminus (site D), generating the mature 18S transcript. The 5' end of the 5.8S gene is trimmed and the 5' and 3' regions of the ITS2 segment are cleaved (van Nues et al., 1995; Fromont-Racine et al., 2003; Nazar, 2003; Wery et al., 2009). Investigation of ribosome biogenesis in *S. cerevisiae* has shown that synthesis and proper processing of rRNA genes, including base modification and cleavage of spacers, and assembly of mature ribosomes are subject to surveillance and quality control mechanisms (Lalev et al., 2000; Wery et al., 2009; Lafontaine, 2010). For example, components of the SSU that fail to assemble correctly are polyadenylated by TRAMP5 and targeted for degradation by the exosome, an exoribonuclease with 3' to 5' activity (Wery et al., 2009).

Figure 1.2. The pathway for processing of the 18S, 5.8S, and 25S nuclear ribosomal RNA transcripts in *S. cerevisiae*. Processing sites in the ETS and ITS region are indicated. Various exo- and endo-ribonucleases are shown by circular characters. Only the pathway showing cleavage at the A_3 to B_{1S} sites, the major processing pathway, is shown. The figure was generated using information from Lafontaine and Tollervey (1995), Fromont-Racine et al. (2003), and Lafontaine (2004).

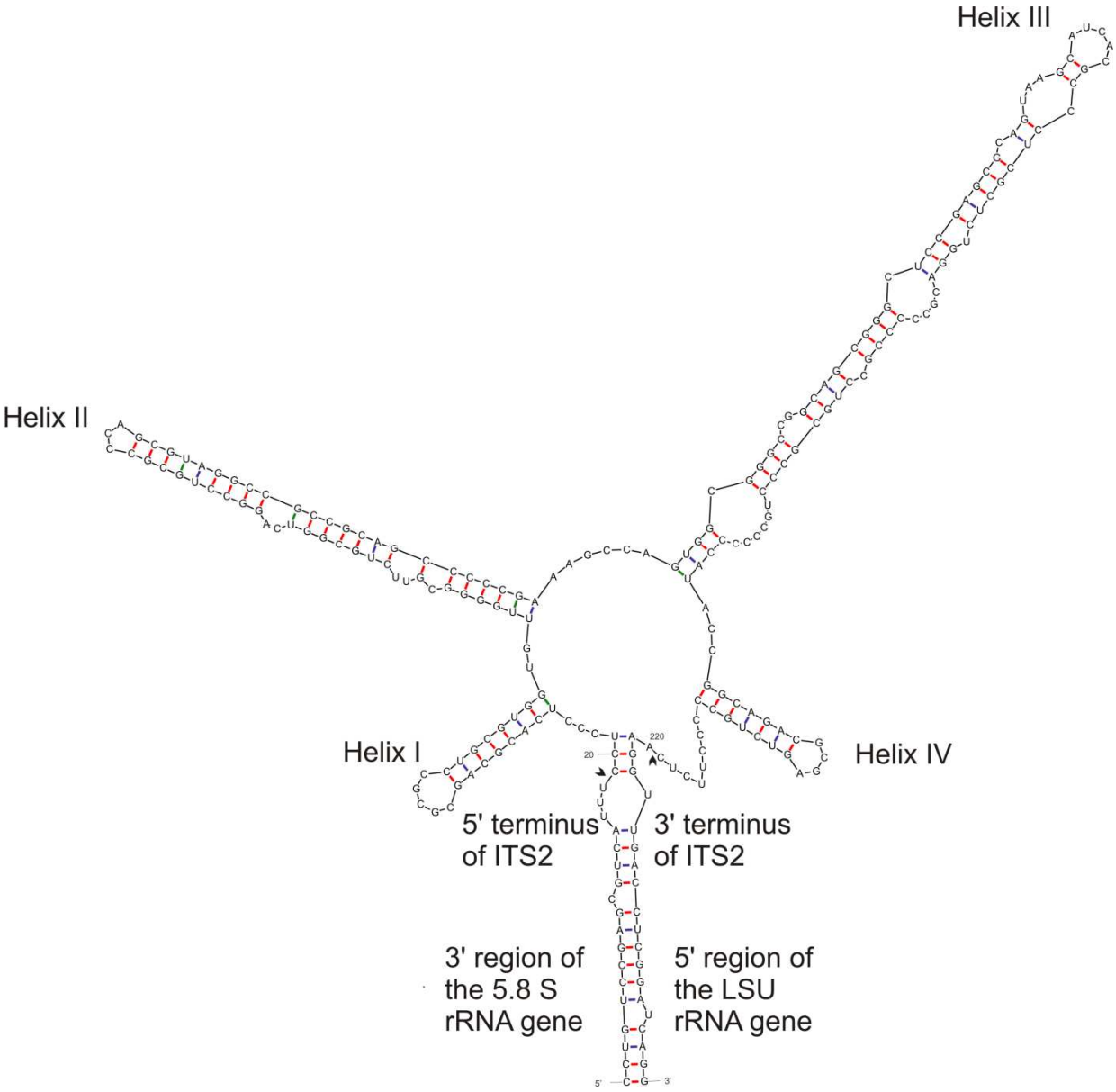

A0
A1
D
A2
A3
B1L
B1S
E
C2
C1
B2
5' ETS
18S
ITS 1
5.8S
ITS 2
25S
3' ETS
D
20S
A3
B1L
B1S
E
C2
C1
B2
Cleavage at A3 to B1S (80-90 %)
Cleavage at B1L (10-20 %)
E
C1
B2
18S
5.8S
25S

1.2.2. Sequence and structural features of ITS1 and ITS2

Models of the RNA secondary structure of the ITS segments were developed using data from biochemical and genetic experiments and homology structural modelling, which uses information obtained from comparative sequence analysis to construct structural models (Yeh and Lee, 1990; Yeh et al., 1990; van Nues et al., 1995; Lalev and Nazar, 1998; Lalev et al., 2000). Using a combination of chemical and enzymatic probes, Yeh et al. (1990) proposed a model for the RNA secondary structure of ITS1 that was composed of several stable helical structures. Three major stem-loop regions were found to be conserved amongst several species of *Saccharomyces* and in the more distantly-related filamentous fungus *N. crassa* despite differences in length and GC content. Using comparative sequence analysis, van Nues et al. (1995) identified evolutionarily conserved regions in the ITS1 segments from *S. cerevisiae*, *Kluyveromyces lactis*, *Torulaspora delbrueckii*, and *Hansenula wingei* and showed that ITS1 segments from *S. cerevisiae* retained their function when conserved regions were replaced by homologous counterparts from the related yeast taxa.

Analysis of ITS2 using biochemical and mutational approaches uncovered two major structural conformations: the hairpin and ring structures (Côté and Peculis, 2001; Côté et al., 2002). The latter conformation of ITS2 consists of a central ring structure surrounded by three to four helices, with helix III as the longest stem region. Numerous studies of the ITS2 region from various eukaryotic taxa, such as green algae, yeasts and filamentous fungi, have produced similar ring models of RNA secondary structure despite great variability in sequence and size; this structure, shown in Figure 1.3, is now described by the pan-eukaryotic four-fingered hand model (Mai and Coleman, 1997; Schultz et al., 2005, 2006; Coleman, 2007).

Figure 1.3. Pan-eukaryotic four-fingered hand model of ITS2. The model of the RNA secondary structure of the filamentous fungus, *Leptographium truncatum*, is shown. The 5' and 3' termini are marked by the arrow heads.


Helix III
Helix II
Helix I
Helix IV
5' terminus of ITS2
3' terminus of ITS2
3' region of the 5.8 S rRNA gene
5' region of the LSU rRNA gene
20
220
5'
3'

Data from both biochemical studies and comparative sequence analysis of the ITS regions indicated that ITS RNA transcripts form extended helical structures, which are essential to their biological function (Mai and Coleman, 1997; Lalev and Nazar, 1999; Lalev et al., 2000; Côté and Peculis, 2001; Schultz et al., 2005; Wolf et al., 2005; Keller et al., 2009). The ITS1 region was proposed to function as a sort of "biological spring", bringing into close proximity the appropriate termini of the 18S and 5.8S genes and of the ITS segments during processing of the pre-RNA precursor (Nazar et al., 1987; Lalev and Nazar, 1998; Lalev et al., 2000). Helical regions within the ITS segments likely function as binding sites for soluble proteins and co-factors. Purification of ITS1 RNA-protein complexes showed that the segment interacts with at least twenty proteins (Lalev et al., 2000). This functional constraint may provide sufficient selective pressure to conserve base pairing interactions. In addition, complementarity between the 3' end of the 5.8S gene and 5' end of the 25S gene brings into close proximity the termini of ITS2.

Evolutionarily-conserved structural elements in ITS1 have been shown to be essential for rRNA processing. GC-rich segments may have a role in forming and/or stabilizing the extensive base pairing of helical regions in the RNA secondary structure, such as the central extended hairpin of ITS1 and the three to four helices radiating from the "palm" structural model for ITS2 (Mai and Coleman, 1997; Schultz et al., 2005; Wolf et al., 2005) and the base pairing interactions between the 5' terminus of ITS2 and the 3' terminus of the LSU (Figure 1.3; Lalev and Nazar, 1999). These observations suggest that the preservation of RNA secondary and/or tertiary structure and, ultimately, biological function, is a major constraint acting on the evolution of ITS sequences.

1.2.3. Applications of ITS sequences in phylogenetic and barcoding studies

Amongst the eukaryotes ITS regions vary greatly in size and sequence (reviewed in Hausner and Wang, 2005). Because of sequence variability, the ITS region, especially the ITS2 segment, has become widely used as a marker in phylogenetic and taxonomic studies, and phylogenetic analysis of the ITS region has been enhanced by the use of RNA secondary structural features to guide the alignment of ambiguous regions (Coleman 2003, 2007, 2009; Goertzen et al., 2003; Young and Coleman, 2004; Won and Renner, 2005; Schultz et al., 2006; Aguilar and Sánchez, 2007; Müller et al., 2007; Krüger and Gargas, 2008; Ryberg et al., 2008; Seibel et al., 2008; Wolf et al., 2008; Schultz and Wolf, 2009). The ITS2 Database III contains models of ITS2 segments that have been developed using homology-based structural analysis. The webserver allows users to upload the ITS2 sequence and obtain a predicted secondary structure based on homology with sequences in the database (Schultz et al., 2006; Selig et al., 2008; Koetschan et al., 2010). Thus, the automation of structural prediction should facilitate phylogenetic studies that incorporate structure-guided sequence alignments in the analysis.

Although other genes, such as *cox1* (*COI*; Seifert et al., 2007) or *nad6* (*ND6*; Santamaria et al., 2009) may be used in conjunction with rRNA-coding genes in the future for DNA barcoding studies in fungi, the utility of the multicopy rDNA sequences, with regards to primer design and PCR amplification, will ensure the continued use of ITS sequences for phylogenetic studies. This in particular applies to situations, including studies using herbarium specimens or environmental samples, in which only limited amounts of DNA can be recovered.

In addition to their use as a marker for phylogenetic studies, ITS sequences are utilized in the identification of fungal species, such as in culture-independent, high-

throughput automated approaches for the characterization of fungal communities from various habitats (Ranjard et al., 2001; Horton, 2002; Buchan et al., 2002; Kennedy and Clipson, 2003; Druzhinina et al., 2005; Mitchell and Zuccaro, 2006; Nilsson et al., 2009). In the case of pathogenic organisms sequence analysis of ITS regions may be useful in the identification of native or exotic plant pathogens (Hausner et al., 2005; Zhang et al., 2008).

ITS2 was also found to be a superior biomarker for the identification of medicinal plants (Chen et al., 2010). Several regions were examined as potential DNA barcodes, including: *psbA-trnH*, *matK*, *rbcL*, and *rpoC1* genes of the plastid genome; the *ycf5* gene; and the ITS region. The authors examined 8557 species and related samples belonging to higher and lower plants, and it was found that the ITS2 segment was readily amplified (amplicons were obtained for 93.8 % of taxa) and exhibited the highest inter-specific divergence and intra-specific variation. The authors also evaluated the accuracy of ITS2 sequences in correctly identifying samples to the genus and species levels using both the BLAST1 and nearest genetic distance method. In both methods the query sequence was compared to sequences in a database; in the BLAST1 approach, sample identity was determined by the best hit (with an E-value below a designated cut-off value) that was retrieved, whereas in the second method, identity was determined based on which sample in a database had the smallest genetic distance from the query (Chen et al., 2010). Both methods gave comparable results: 99.8 % and 92.7 % of the samples were correctly identified to the genus and species levels, respectively, with BLAST1 and 99.7 % and 90.3 % of the samples were identified to the genus and species levels, respectively, with nearest genetic distance method (Chen et al., 2010). This recent study highlights the continued wide-spread use of the ITS2 for as a marker for rapid species identification.

In summary, RNA secondary structure models of ITS1 and ITS2 indicate that the segments fold into extensive hairpin structures. While RNA secondary structural features are conserved across taxa, the sequence and size of the ITS segments are hypervariable, making this region useful in phylogenetic and barcoding studies. By resolving phylogenetic relationships between closely-related taxa, analysis of the ITS region may help elucidate the evolutionary dynamics of mobile genetic elements, such as group I and group II introns and HEGs. Superimposing the presence/absence of genetic insertions onto the phylogeny of host organisms may reveal greater detail on the evolutionary dynamics and origins; that is, the inheritance of mobile genetic elements by vertical descent or horizontal transmission.

1.3. Fungal mitochondrial genomes

1.3.1. Conformation and genetic organization of fungal mitochondrial DNA

Current models describing the origin of eukaryotic organisms propose that mitochondria arose as a result of either an endosymbiotic event between an α-proteobacterium and a pre-eukaryotic organism (reviewed in Gray et al., 2001; Bullerwell et al., 2003; Hausner, 2003) or, as more recently proposed, of a merger between an α-proteobacterium symbiont and archaebacterial host (Martin and Koonin, 2006). Sequence analyses suggest that the original prokaryotic symbiont was likely related to *Rickettsia* spp., as about 10 % of the mt genes are related to those found in *R. prowasekii* (reviewed in Bullerwell et al., 2003; Hausner, 2003). In general, the mt genome comprises genes involved in protein synthesis and oxidative phosphorylation. This core gene complement includes the SSU and LSU rRNA genes, subunits of the cytochrome oxidase (*cox1-3*), NADH dehydrogenase (*nad1-4*, *4L-6*), components of

the ATP synthase (*atp*), and transfer (t) RNA genes. It is worth noting that *nad* genes are absent from the mt genomes of the yeasts *S. cerevisiae* and *S. castelli* (reviewed in Bullerwell et al., 2003). Interestingly, these conserved, essential genes are often the targets of invasive mobile elements, such as group I and group II introns and HEGs.

In general, mt DNA is AT-rich and often either circular or circular-mapping, if the genome is arranged in linear head-to-tail concatemers. There are, however, examples of unit length molecules, as in the chytridiomycete *Hyaloraphidium curvatum*. In *Spizellomyces punctatus*, another chytridiomycete, the mt genome is found in three segments of 1.1-, 1.4-, and 58.8-kb in size (Grivell, 1995; Paquin et al., 1997; Bullerwell et al., 2003). There is great variation in the sizes of mt genomes, from 19.4 kb in *Schizosaccharomyces pombe* to nearly 86 kb in *S. cerevisiae* (Bullerwell et al., 2003) and up to 175 kb in *Agaricus bisporus* (Hausner, 2003). Variation in the size of mt genomes across fungal taxa is due to several factors, including: the presence/ absence of *nad* genes; the number of tRNA genes; the presence/absence of insertion elements, including introns and HEGs; and differences in the length of intergenic spacers (Clark-Walker, 1992; Bullerwell et al., 2003; Hausner, 2003). Some insertion elements have been shown to be mobile and invade new sites in the genome through a process involving DNA cleavage and duplication of the insertion element into the damaged site. It is tantalizing to suppose that this process of mobility may be an on-going, dynamic factor that could continually influence the size and arrangement of the genome.

Mitochondrial DNA is present in multiple copies within each mitochondrion. The hyphae of filamentous ascomycetous fungi are segregated by septae; however, the cytoplasmic contents, including mitochondria, in each compartment may be exchanged via pores in the septa, and if mitochondria fuse, there is the potential for

mixing of mt DNA from different mitochondria. Also, hyphae from different fungi can occasionally make contact and temporarily fuse (hyphal anastomosis), allowing for an exchange of cytoplasm and, possibly, mitochondria. Both mechanisms are means by which mitochondria can fuse and the genomes can mix, allowing for possible recombination (reviewed in Hausner, 2003; Okamoto and Shaw, 2005).

1.3.2. Characterization of codon usage and promoter sequences

Genes encoded by fungal mt DNA are not always translated using the universal genetic code, and, in fact, there are some notable examples of deviations from this code. For instance, the stop codon UGA encodes a tryptophan in many fungal mt DNA; AUA encodes a methionine, rather than isoleucine; and CUN codes for threonine and not leucine (Paquin et al., 1997; Bullerwell et al., 2003; Schäfer, 2003). Identification of mt promoter sequences is challenging because of the lack of consensus sequence conservation amongst more distantly-related taxa. A putative promoter sequence 5'-ATATAAGTA-3' was identified upstream of known transcription initiation sites in *S. cerevisiae* and in strains of its close relatives *Kluyveromyces lactis* and *Torulopsis glabrata*; however, this consensus sequence is absent in more distant relatives, including *S. pombe*, *S. japonicus*, and *S. octosporus* (Grivell, 1995; Bullerwell et al., 2003; Schäfer, 2003). Putative promoter sequences for the *rns*, *rnl*, and cytochrome b (*cob*) genes have been identified in *N. crassa* (Kennell and Lambowitz, 1989); the consensus sequence was identified as 5'-TAGAR R(T/G)R(T/G)A - 3'.

1.3.3. Optional genetic elements: introns and intron-encoded genes

Insertion sequences corresponding to self-splicing introns and homing endonuclease genes (HEGs) were identified within the *rnl*, *rns*, *cox*, and *cob* genes of several species of ascomycetous fungi (Michel et al., 1982). Comparative sequence analyses identified two major families of self-splicing introns present in mt DNA, group I and group II introns, and each family is distinguished by conserved sequence and RNA structural features. Moreover, a number of these introns contain large ORFs. Group I introns are typically associated with ORFS that are related to maturases or DNA ENases, while group II IEPs contain motifs that are related to retroviral RTs.

1.3.4. Mitochondrial plasmids

Mitochondrial plasmids are defined as autonomously replicating elements that do not share sequence homology with the parental mt genome (Dujon and Belcour, 1989; Griffiths, 1995; Hausner, 2003). Mitochondrial plasmids may be transmitted vertically from parent to progeny or horizontally; the latter mechanism may occur during sexual crosses or during hyphal anastomosis, when fungal hyphae from two strains temporarily fuse (Hausner, 2003). Plasmids are generally linear, although some examples of circular plasmids have been found. In *N. crassa* two linear plasmids (referred to as *kalilo* and *maranhar*) and four circular plasmids (*mauriceville*, *varkud*, *fiji*, and *laBelle*) have been described. In general, mt plasmids are classified into three types based on plasmid conformation and mode of replication (Monteiro-Vitorello et al., 2000; Cahan and Kennell, 2005). Type 1 plasmids are retroplasmids that are either circular (type 1A) or linear (type 1B), encode an RT, and replicate via an RNA intermediate. The *mauriceville* and *varkud* plasmids of *N. crassa* are well-studied examples of type 1 retroplasmids. Type 2 plasmids are circular and encode DNA

polymerases; the *fiji* and *laBelle* plasmids are examples of type 2 plasmids. Type 3 plasmids are linear, code for DNA and RNA polymerases, and have terminal inverted repeats, which may be important in protein-primed replication (Hausner, 2003; Cahan and Kennell, 2005).

1.3.5. Plasmid-like elements

Circular mt plasmid-like elements (pLMEs) are derived from regions of mt DNA and, once formed, tend to accumulate and co-exist with wild-type mt DNA. They have been identified in several species of filamentous fungi, including *N. crassa*, *C. parasitica*, and *O. novo-ulmi*. The latter two organisms are economically significant plant pathogens, causing chestnut blight and Dutch elm disease, respectively, and the accumulation of pLMEs has been shown to be associated with, although not the cause of, a hypovirulent phenotype (Hausner et al., 2006a; Monteiro-Vitorello et al., 2009). Previous work in *N. crassa* showed that some types of pLMEs can be transmitted from pLME-carrier strains to wild-type strains via heterokaryosis or to asexual (vegetative) progeny through mitosis but may be eliminated during meiosis; that is, sexual crosses (Hausner et al., 2006a). However, results from the characterization of a second pLME, pstp-BI (2.2 kb), derived from a region near the *cox2* gene in *N. crassa* indicated that this element can survive meiosis. In this case, the element is transmitted maternally to sexually-derived progeny, although the mode of replication is unknown (Hausner et al., 2006a). Characterization of the mt *rns* gene in *C. parasitica* revealed the presence of an autonomously replicating pLME in respiratory mutants *mit1* and *mit2* that was derived from a region of the *rns* gene that included exon 1 and the majority of the first intron, an unusual group II intron containing a putative LAGLIDADG-type HEG (LHEG; Monteiro-Vitorello et al.,

2009). In *Podospora anserina*, analysis of a plasmid-like sequence associated with senescence revealed it to share sequence homology with an RT-encoding group II intron within the *cox1* gene (Osiewacz and Esser, 1984; Cummings et al., 1985).

1.4. Mobile genetic elements: introns and homing endonucleases

1.4.1 Group I and group II introns: the discovery and initial characterization of self-splicing introns

Group I and group II introns are intervening sequences that catalyze their own removal from precursor messenger (m) RNA transcripts. Since they are absent from the mature transcript, their deleterious effects on the host genome are minimized. Because they are autocatalytic, group I and group II introns are commonly referred to as ribozymes, or RNA-derived enzymes (Cech and Bass, 1986; Cech, 1987; Dujon, 1989; Jaeger, 1997; Lambowitz et al., 1999; Doudna and Cech, 2002; Khan and Lal, 2003; Lehmann and Schmidt, 2003; Lambowitz and Zimmerly, 2004; Scott, 2007). At the RNA level these introns fold into compact tertiary structures, bringing into close proximity the flanking exon sequences.

Intron-catalyzed splicing proceeds by two transesterification reactions; however, fundamental differences in their primary sequence, RNA secondary and tertiary structures, mechanism of splicing, and distribution indicate that these elements have different origins and have evolved independently of each other (reviewed in Lambowitz and Belfort, 1993; Saldanha et al., 1993; Michel and Ferat, 1995; Lambowitz et al., 1999; Lambowitz and Zimmerly, 2004). Features of RNA structure and of splicing mechanisms of group I and group II introns will be discussed in sections 1.5 and 1.6, respectively. The ability to self-splice *in vitro* in the absence

of protein co-factors is a major distinction between the group I and group II introns and the nuclear spliceosomal introns; the latter is a component of RNP particles and requires an additional 300+ protein co-factors for intron splicing (reviewed in Lambowitz and Belfort, 1993; Saldanha et al., 1993; Michel and Ferat, 1995; Lambowitz et al., 1999; Patel and Bellini, 2008).

On the basis of results obtained from comparative sequence and RNA secondary structural analyses two distinct intron families were first proposed, group I and group II (Michel et al., 1982; Michel and Dujon, 1983). Sequence analysis of the mt genome of *S. cerevisiae* revealed the presence of introns, six of which contained putative ORFs, in the *rnl*, cytochrome oxidase subunit 1 (*oxi3*), and apocytochrome b (*cob*) genes. Sequence comparison of the *rnl* intron in *S. cerevisiae* and in the related fungus *Kluyveromyces thermotolerans* indicated that the two introns were closely-related. Additional introns were subsequently identified and sequenced; conserved core secondary structures were identified using data obtained from comparative sequence analysis and modelling of RNA secondary structures, and members of the group I intron family were further divided into subgroups on the basis of fine differences in the core structure. Moreover, it was observed that several introns appeared to be evolutionarily-conserved, with similar introns present in the same loci in distantly-related organisms (Michel et al., 1982; Lang, 1984; Lang et al., 1985). The introns were apparently optional, non-essential elements, for there was no phenotypic difference between strains with and without these elements. Analysis of the putative ORFs led to some intriguing suggestions for the functional role of IEPs. It was found for several intronic ORFs in the *cob* gene that trans-recessive mutations prevented intron splicing, suggesting that the ORF encoded a protein that functioned

as a “spligase”, now referred to as a maturase, to facilitate intron splicing (Michel et al., 1982).

1.4.2. Group III and archaeal introns

Group III introns are apparently degenerated group II introns: while group III introns contain domain VI, domains I to IV are often absent (reviewed in Lambowitz and Belfort, 1993). Group III introns are abundant in *Euglena* chloroplast (cp) genomes and are typically 93 to 120 nt in size, which is significantly smaller than group I and group II introns (Hallick et al., 1993; Copertino and Hallick, 1993). Characterization of the fourth intron in the *psbC* gene of *E. gracilis* revealed it to be a composed of a series of three nested introns that self-splice sequentially, starting from the innermost intron, and are excised as lariats (Copertino et al., 1994). At least fifteen nested group III “twintrons” are present in euglenoid cp DNA.

Archaeal introns are also diminutive in size, ranging from 110 to 600 nt (if an LHEG is present). Unlike group I and group II introns, archaeal introns lack a conserved internal structure and are not true ribozymes. Instead, they form bulge-helix-bulge motifs at the exon-intron junctions that are recognized by endoribonucleases (Lambowitz and Belfort, 1993). The introns are inserted in various helical and loop regions in archaeal tRNA and rRNA genes.

1.4.3. Genes encoding homing endonucleases

Genes encoding homing endonucleases (HEases) have been identified in phage and archaeal genomes and in the nuclear and organellar genomes of fungi and plants where they tend to function as “selfish”, parasitic elements. They self-propagate and persist in the genome by coding for a protein product that cleaves

double-stranded DNA and by doing so stimulate DNA repair mechanisms and gene conversion events that result in the duplication of the HEG into a new site in the genome. This process, referred to as homing, will be discussed in greater detail in section 1.7. Putative HEGs have been identified within non-spliceosomal introns (group I and group II introns), within other ORFs (in which case the corresponding protein is referred to as an intein, or self-splicing protein) or as free-standing ORFs (Belfort and Roberts, 1997; Chevalier and Stoddard, 2001; Burt and Koufopanou, 2004; Stoddard, 2006; Edgell, 2009).

Homing endonucleases recognize long (14 to 40 bp) DNA target sites and make double-stranded cuts. Homing endonucleases generally exhibit high sequence specificity, yet they they are able to tolerate some polymorphisms in their target site. This give-and-take is essential to the "survival" of the HEG. On the one hand, high sequence specificity helps to ensure that the HEase acts as a rare-cutting enzyme and toxicity to the cell, due to random or excessive cleavage, is minimized. On the other hand, flexibility in sequence recognition is key to the continued propagation of the element within and across genomes; the HEase may recognize and cleave, albeit at a lower efficiency, new target sites.

1.5. Group I introns

1.5.1. Intron Distribution

Group I introns are abundant in the nuclear and organellar DNA of eukaryotes, and they are the major class of introns in fungal mt DNA. They have been found as well in the genomes of eubacteria and bacteriophage. Group I introns are present within the nuclear SSU and LSU ribosomal genes of fungi and algae, protein coding

genes, LSU ribosomal and tRNA genes in cp DNA, bacterial tRNA genes, in phage genes encoding proteins involved in DNA synthesis, and they have recently been identified in the mitochondria of sea anemones and soft corals (Rochaix et al., 1985; Shub et al., 1988; Lambowitz and Belfort, 1993; Saldanha et al., 1993; Carbone et al., 1995; Lambowitz et al., 1999; van Oppen, 2000; Belfort et al., 2002; Hausner, 2003; Gibb and Hausner, 2005).

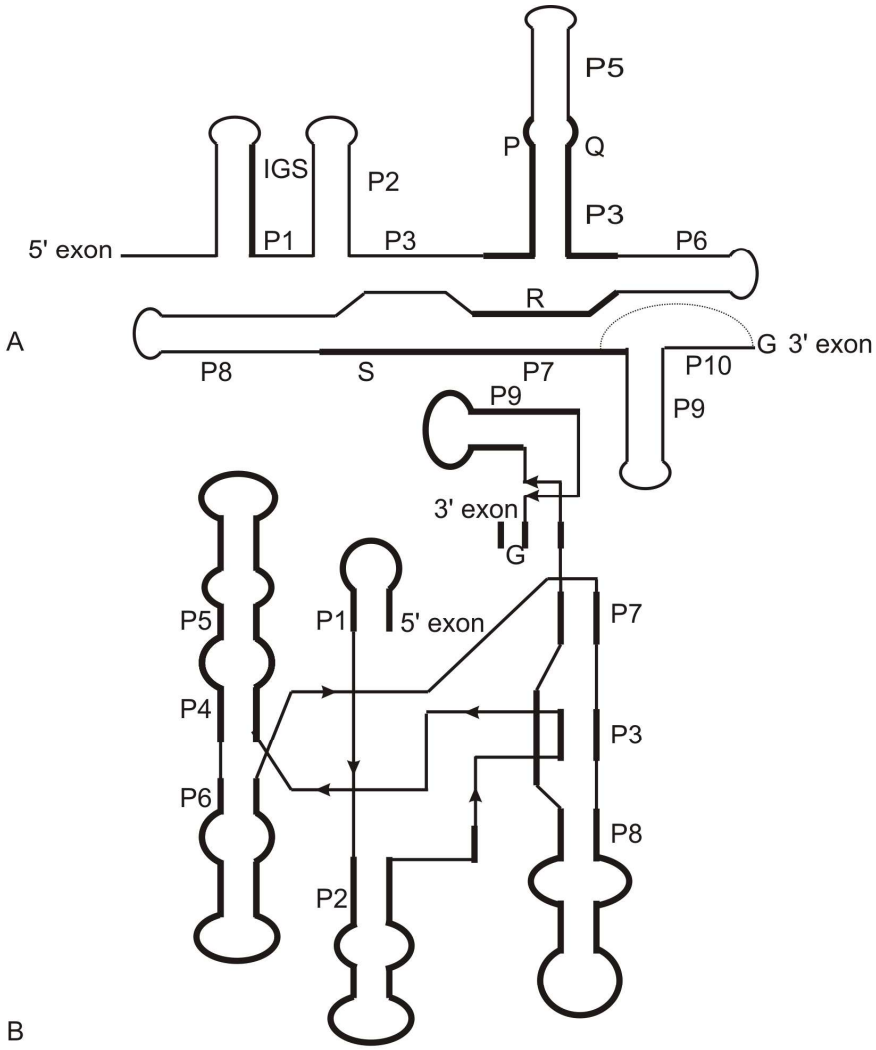
Intron distribution is often random and introns may either be present or absent amongst closely-related strains or species (Lambowitz and Belfort, 1993). There are examples of multiple introns inserted within the same gene (Michel et al., 1982; Michel and Dujon, 1983) and of related introns in different genes. The IEP of the mt *cob*-I4 intron in *S. cerevisiae* promotes the splicing of both its own intron and that of the *cox1*-I4α intron, suggesting that the introns share homologous sequence and/or RNA structural motifs (Lambowitz and Belfort, 1993). The stochastic distribution of introns and the presence of related introns in different genetic loci are a reflection of intron mobility within homologous alleles in different organisms and intronic transposition to ectopic sites within the genome (Lambowitz and Belfort, 1993, Saldanha et al., 1993).

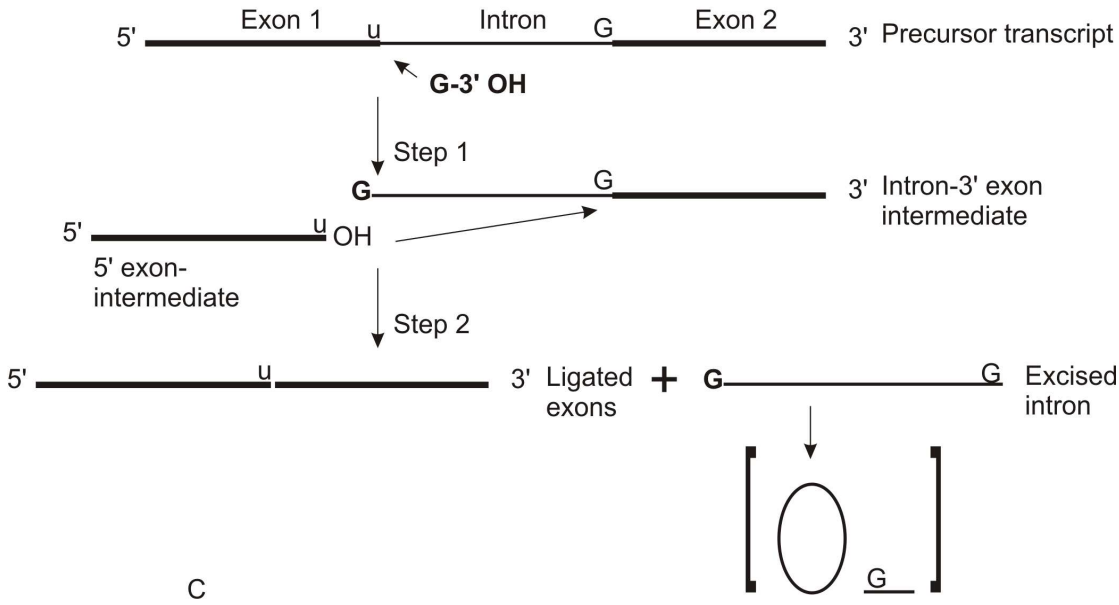
1.5.2. RNA secondary and tertiary structure

Group I introns range in size from 200 nt to approximately 3000 nt, if an ORF is present. Large ORFs have been identified in approximately one-third of group I introns and may be inserted in any of the protruding loops on either the sense or antisense strand (Michel et al., 1982; Schäfer, 2003). The ORF sequence may also extend into the intron sequence, itself, becoming part of the intron core structure, in which case the ORF is fused in-frame to the upstream exon and is translated as a

fusion protein (Shub et al., 1988; Gibb and Hausner, 2005; Sethuraman et al., 2008). Unrelated introns are variable in sequence but comparative sequence analysis and RNA secondary structure modelling revealed conserved sequence elements P, Q, R, and S and a core secondary structure that is dominated by ten stem regions (P1 through P10), as shown in Figure1.4A and B (reviewed in Cech, 1988; Michel and Westhof, 1990; Lambowitz and Belfort, 1993; Saldanha et al., 1993). Variations in sequence and secondary structure led to the subdivision of group I introns into five major classes (IA to IE). The intron's internal guide sequence (IGS) base pairs with flanking exon sequences, forming the P1 and P10 hairpin domains. The P1 base pairing interactions determine the 5' splice site, while the 3' splice site is determined by both the P10 interaction and interactions between P9.0 and the catalytic core, formed by helces P3 to P8 (Burke et al., 1990; Michel et al., 1990). Three dimensional models show that the helices P3-P4-P6 and P9.0-P7-P3-P8 are co-axially stacked and form a cleft, which is the active site of the ribozyme (Kim and Cech, 1987; Michel and Westhof, 1990).

Figure 1.4. Structure and catalysis of group I introns. (A) Secondary structure of group I introns. Conserved sequence elements P, Q, R, S and the IGS and hairpin regions P1 to P10 in group I introns are indicated. (B) Model of the RNA secondary structure based on crystallographic data. (C) Splicing of group I introns consists of two transesterification reactions, initiated by an external guanosine co-factor. The intron may cyclize in a third transesterification step, releasing the 5' G as part of a short oligonucleotide. Panels A and B were generated from information obtained from Saldanha et al. (1993) and Vicens and Cech (2006).


P5
P
Q
IGS
P2
P3
P1
P3
P6
5' exon
R
A
P8
S
P7
P10
G 3' exon
P9
P9
3' exon
G
P5
P1
5' exon
P7
P4
P3
P6
P8
P2
B


5'
Exon 1
u
Intron
G
Exon 2
3'
Precursor transcript
G-3' OH
Step 1
G
G
3'
Intron-3' exon
intermediate
5'
u
OH
5' exon-
intermediate
Step 2
5'
u
3'
Ligated
exons
+
G
G
Excised
intron
G
C

1.5.3. Intron splicing

Group I introns self-splice via a two-step transesterification reaction (Figure 1.4C) and splicing is dependent upon Mg^{2+}, which is essential for stabilizing RNA secondary and tertiary structure, as well as the transition state structure, and for activating the 3' OH group of the nucleophile (reviewed in Saldanha et al., 1993; Jaeger, 1997). The reaction also requires exogenous GTP, which acts as a co-factor; its binding site is located within the catalytic cleft. Amino acids, such as arginine, or antibiotics with guanido groups, streptomycin, for an example, act as competitive inhibitors of splicing by binding the guanosine binding pocket (reviewed in Saldanha et al., 1993). The guanosine initiates a nucleophilic attack at the 5' splice site, cleaving at the 5' splice junction and binding to the 5' terminus of the intron. The 5' exon and intron-3' exon intermediates interact via the IGS sequences. In the last step the hydroxyl group of the 5'exon attacks the 3' splice site, resulting in the ligation of exons and release of the intron-5'G molecule. Both reactions are reversible and do not require additional energy. The intron may cyclize in a third transesterification step, releasing the 5' G as part of a short oligonucleotide (reviewed in Cech and Bass, 1986; Lambowitz and Belfort, 1993; Saldanha et al., 1993).

1.6. Group II introns

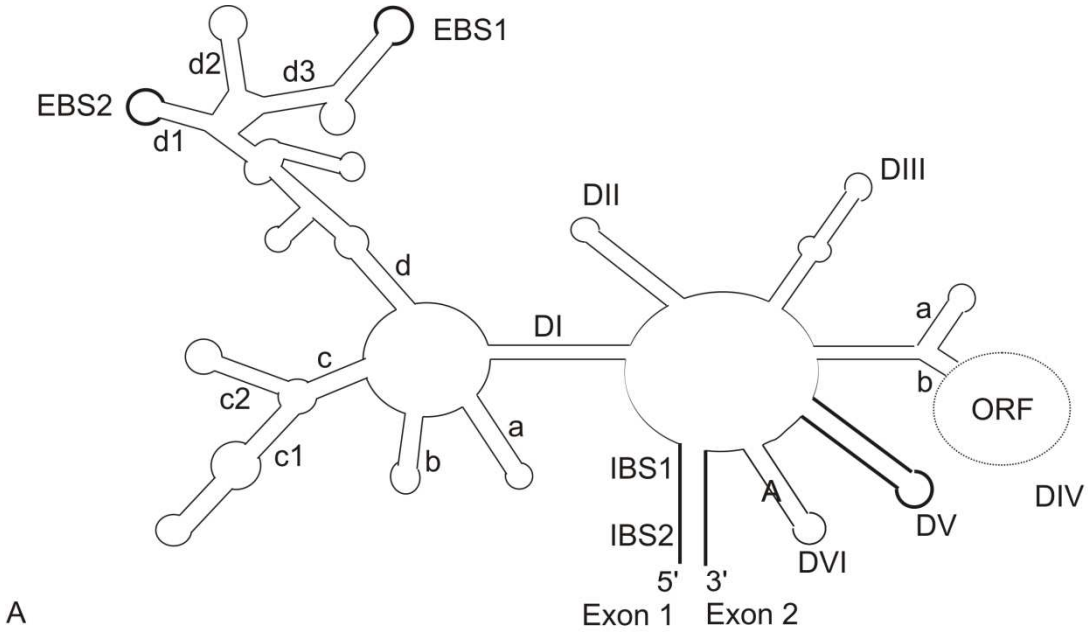
1.6.1. Distribution of group II introns

Group II introns are the major class of introns within cp and mt DNA of higher plants. They are wide-spread in the organellar genomes of fungi, algae, and plants, and although they are common in the genomes of eubacteria (including the proposed ancestors of the mitochondria and chloroplasts), group II introns are rare in the

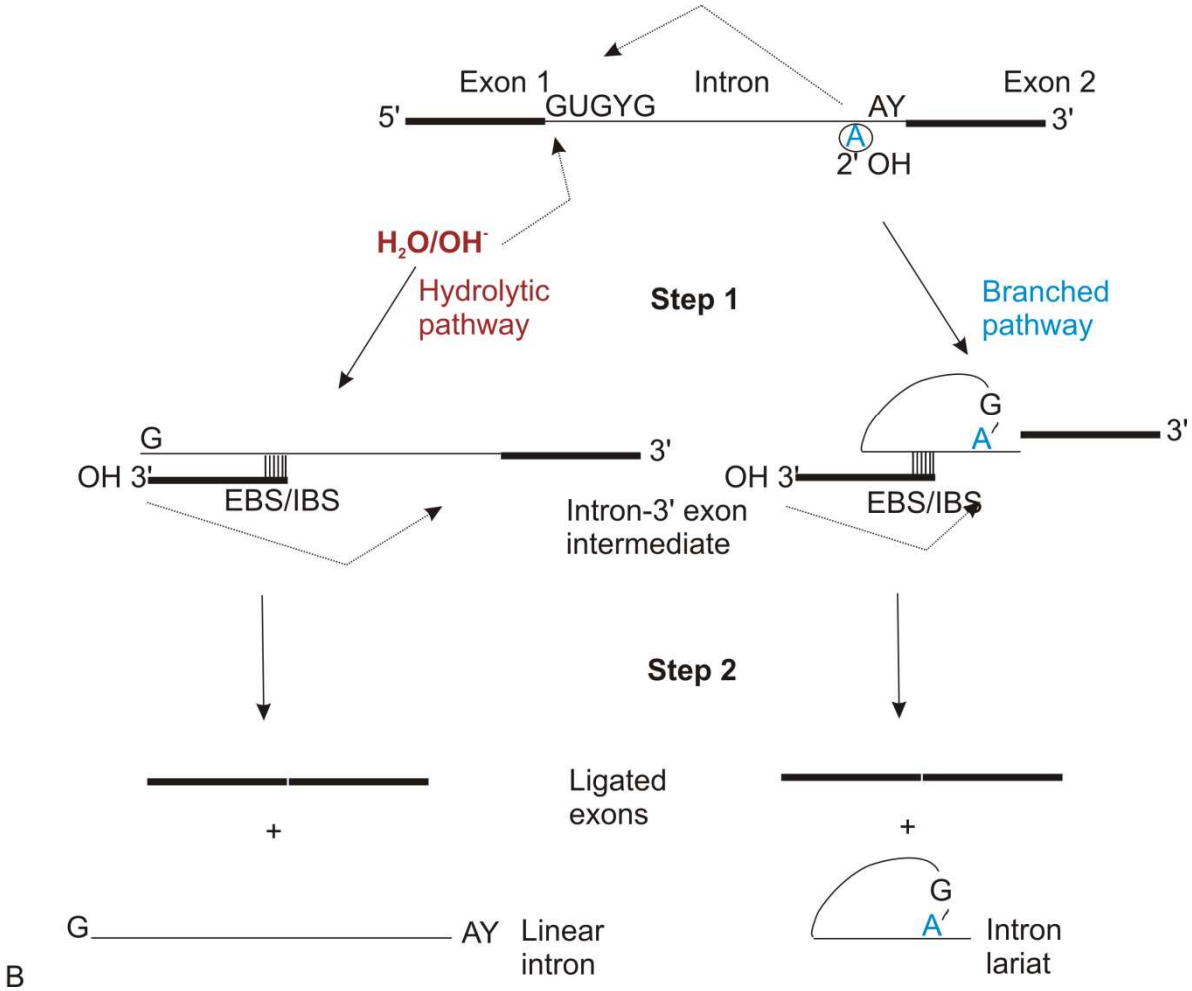
archaea (Ferat and Michel, 1993; Michel and Ferat, 1995; Lambowitz and Zimmerly, 2004; Simon et al., 2008; Kamikawa et al., 2009). A group II intron encoding a putative RT was identified within the mt *cox1* gene DNA of a bilaterian worm, a metazoan animal (Vallès et al., 2008). Group IIB1 introns were recently identified in the mt *rnl* gene in the wine yeast *Candida zemplinina* (Pramateftaki et al., 2008); this observation was unusual because group IIA, not group IIB, introns are usually predominant in fungal mt DNA (Pramateftaki et al., 2008). Moreover, these introns were found to contain RT ORFs of the eubacterial lineage; this finding is interesting because RT-type ORFs associated with group IIB1 introns are typically of the cp lineage (Toor et al., 2001).

Group II introns have not been found in nuclear genomes; however, structural and biochemical similarities between group II introns and nuclear spliceosomal introns suggest a possible evolutionary relationship between group II introns and nuclear spliceosomal introns (Cech, 1986; Jacquier and Rosbash, 1986; Michel et al., 1989; Copertino and Hallick, 1993; Lambowitz and Belfort, 1993; Saldanha et al., 1993; Michel and Ferat, 1995; Costa et al., 1998; Martin and Koonin, 2006; Valadkan, 2007; Michel et al., 2009; Keating et al., 2010). Both classes contain a bulged adenosine residue, whose 2' OH group initiates a nucleophilic attack at the 5' splice site, forming a 2'-5' phosphodiester bond, and the intron is released as a lariat molecule (Figure 1.5).

Figure 1.5. General structure and catalysis of group II introns. (A) Secondary structure of group II introns. Helical regions, DI to DVI, are indicated. Intron sequences EBS1 and EBS2 determine the 5' splice site by base pairing interactions with IBS1 and IBS2 in the 5' exon. Domain V, indicated in bold is more conserved in sequence than the other domains. Sequences in the intron and exon are in capital and small letters, respectively. Exons are shown as dark lines. (B) Splicing of group II introns follows the branching and hydrolytic pathways. In the branching pathway, the attack at the 5' splice site is initiated by a bulged adenosine residue (circled A) in domain VI and the intron is released as a branched, or lariat, molecule with a characteristic 2'-5' phosphodiester bond. In the hydrolytic pathway, the 5'splice site is cleaved by a water molecule or hydroxyl ion and the intron is excised as a linear molecule. Panel A was generated using information from Saldanha et al. (1993), Michel and Ferat (1995), Costa et al. (2000), and Lambowitz and Zimmerly (2004).


EBS1
d2
d3
EBS2
d1
d
DII
DIII
DI
a
c
b
ORF
c2
a
c1
b
IBS1
A
DV
DIV
IBS2
DVI
5'
3'
Exon 1
Exon 2

A


Exon 1
Intron
Exon 2
5'
GUGYG
AY
3'
A
2' OH
H2O/OH-
Hydrolytic pathway
Step 1
Branched pathway
G
3'
OH 3'
EBS/IBS
Intron-3' exon intermediate
G
A
3'
OH 3'
EBS/IBS
Step 2
Ligated exons
+
+
G
AY
Linear intron
G
A
Intron lariat

B

1.6.2. Secondary structure

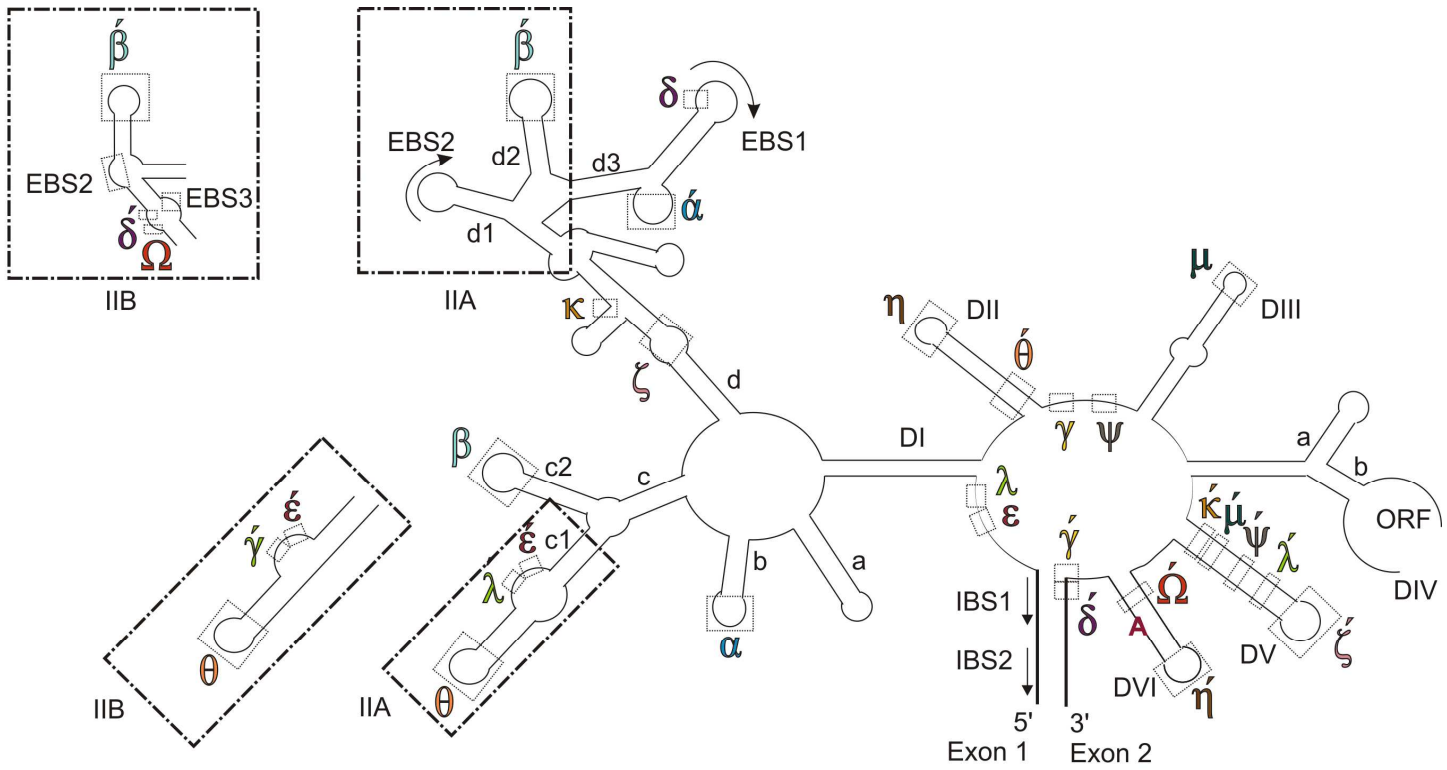
In contrast to group I introns, members of group II are considerably larger in size. They can be in excess of 500 nt in size (Simon et al., 2009) and fungal mt introns range in size from approximately 900 to 2500 nt, if an ORF is present (reviewed in Belfort and Lambowitz, 1993). As shown in Figure 1.5A, the core RNA secondary structure of group II introns consists of a ring structure surrounded by six major stem-loop domains (DI to DVI). Based on fine structural differences and the presence/absence of tertiary interactions, group II introns have been subdivided into three major classes: IIA and IIB and bacterial classes IIC-IIE, and minor differences within IIA and IIB have led to further division of these subclasses into IIA1, IIA2, IIB1, and IIB2 (Michel et al., 1989; Toor et al., 2001; Zimmerly et al., 2001; Lehmann and Schmidt, 2003; Lambowitz and Zimmerly, 2004).

Using a combination of comparative sequence analysis, base substitution studies, and chemical probing with RNA-modifying chemicals such as dimethylsulfate, diethylpyrocarbonate, and kethoxal, a number of regions involved in canonical (Watson-Crick) and non-canonical base pairing interactions, as well as tertiary interactions, have been identified and mapped onto the secondary structure (Michel et al., 1989, 2009; Michel and Ferat, 1995; Costa and Michel, 1995; Costa et al., 1997a, 1997b, 1998, 2000; Qin and Pyle, 1998; Lehmann and Schmidt, 2003; Lambowitz and Zimmerly, 2004; Fedorova and Zingler, 2007). As shown in Figure 1.6A, DI is the largest, comprising several smaller stem-loop subdomains (labelled 1a, 1b, 1c1, etc.). Domain I has an essential role in exon sequence recognition during splicing and intron mobility, and it also acts as a scaffold by forming multiple tertiary interactions with the other domains during intron folding. Exon binding sites (EBS) 1 and 2, located in DI, are involved in 5' splice site recognition and form canonical base

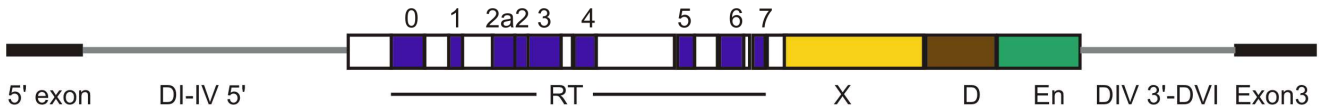
pairing interactions, typically 5 to 7 nt, with sequences in the 5' exon (Jacquier and Michel, 1987; Michel et al., 1989). The EBS2 site, present in group IIA introns, also interacts with the 5' exon; this site is absent in a number of group IIB introns (Costa et al., 2000). In members of the group IIA introns, the 3' splice site is defined by the Watson-Crick base pair formed between the first nucleotide of the 3' exon [intron binding site (IBS) 3] and the nucleotide immediately 5' to the EBS1 segment (δ, Michel et al., 1989; Costa et al., 2000). However, in group IIB introns, the δ position interacts with δ' and the 3' splice site is defined instead by the EBS3-IBS3 interaction (Costa et al., 2000).

Neither DII nor DIII are required for catalysis but they may promote efficient splicing by stabilizing tertiary structure. Interactions between DII and DVI (θ-θ' and η-η') are involved in a conformational switch during intron splicing and DIII acts as a "catalytic effector" by enhancing the efficiency of splicing. In canonical group II introns, the ORF is located in a looped-out segment in DIV (or, more rarely, in DII); this region is dispensable for self-splicing and likely contains protein-binding sites. Domain V is the only region phylogenetically conserved in sequence; together DI and DV are the essential components of the group II ribozyme and form the minimal catalytic core (Lambowitz and Zimmerly, 2004; Fedorova and Zingler, 2007). Several key tertiary interactions between DI and DV have been identified. These include the docking sites for the GNRA tetraloop of DV with a receptor in DId (the ζ-ζ' interaction), the κ-κ' interaction, and the λ-λ' interaction. Domain VI contains a bulged adenosine residue, which is involved in branch-point splicing.

Figure 1.6. Subdivision of group II introns based on differences in secondary structure and tertiary interactions and the arrangement of group II-encoded RT ORFs. (A) Secondary structure of group II introns. Structural differences that differentiate two of the subclasses of group II introns (IIA, IIB) are highlighted in boxes. Regions involved in tertiary interactions are marked by Greek symbols and the partner in the interaction is marked by '. Exon sequences are in black boxes. The 5' splice site is determined by base pairing interactions between the exon (IBS1 and IBS2) and DI of the intron (EBS1 and EBS2) and the 3' splice site is determined by the δ-δ' (IIA) or EBS3-IBS3 (IIB) interactions. (B) Arrangement of group II intron-encoded RTs. RT, reverse transcriptase; X, maturase; D, DNA binding; En, ENease domains. Shown is the arrangement of the RT encoded by the *Lactococcus lactis* intron Ll.LtrB. There are variations amongst RTs in the size of spacer regions (white) and the presence/absence of D and En domains. Exon and intron regions are indicated by black and gray lines, respectively. The figure was generated using information from Michel et al. (1989), Michel and Ferat (1995), and Lambowitz and Zimmerly (2004).


EBS2
EBS3
IIB
EBS2
d2
d3
EBS1
d1
IIA
DII
DIII
DI
d
c
c2
c1
b
a
IBS1
IBS2
5'
Exon 1
3'
Exon 2
DV
DVI
ORF
DIV
IIB
IIA

A


0 1 2a2 3 4 5 6 7
5' exon
DI-IV 5'
RT
X
D
En
DIV 3'-DVI
Exon3

B

1.6.3. Tertiary structure

Tertiary interactions were identified using data obtained from comparative sequence analysis, chemical probing, circular permutation analysis, and cross-linking studies. Combining data on known and newly identified interactions, Costa et al. (2000) identified two new interactions in group IIB introns: δ-δ' and EBS3-IBS3. The former interaction is involved in stabilizing the interactions between the 5' exon and DI, while the latter interaction is involved in 3' splice site recognition. In the same study, the authors also demonstrated a direct interaction between DI and DV that brings into close proximity the catalytic component, DV, of the ribozyme and the flanking exon sequences (Costa et al., 2000). Dai et al. (2008) developed a three-dimensional model of a bacterial group II intron. In their model, DV docks into DI, which folds over both sides of DV. The DIII catalytic effector contacts both halves of DI. The EBSs are on the surface of the ribozyme; however, in group IIB introns exons are positioned internally.

Recently the first crystal structural of a spliced bacterial group IIC intron from *Oceanobacillus iheyensis* was published (Toor et al., 2008, reviewed in Pyle, 2010). Bacterial group II introns are considerably smaller than introns belonging to either the mt- or cp-classes. The construct used in the crystallization study was further truncated and lacks such key components as the branch-point and surrounding RNA secondary structure domain (Toor et al., 2008). A high-affinity binding site for the IEP of the *Lactococcus lactis* intron, Ll.LtrB, has been well characterized, but this constitutes only a small part of the intron RNA-protein interface and the location of the rest of the site remains controversial (reviewed in Michel et al., 2009). Crystallization of RNA is challenging, particularly with larger molecules. Because of the relatively uniformly-charged phosphate backbone and reduced chemical variation, it is more likely that

molecules will crystallize out of register (Westhof and Michel, 1998). Thus, the crystallization of the *O. iheyensis* intron represents a major breakthrough in visualizing the structural and catalytic organization of group II ribozymes and also provides clues as to the molecular organization of the spliceosome.

1.6.4. Mechanisms of intron splicing

Despite sequence variability, group II introns adopt evolutionarily conserved RNA secondary and tertiary structures essential for catalyzing autosplicing in the absence of protein factors (reviewed in Michel and Ferat, 1995; Lehmann and Schmidt, 2003; Lambowitz and Zimmerly, 2004; Robart and Zimmerly, 2005; Fedorova and Zingler, 2007). A number of group II introns in yeast, algae, and bacteria have been shown to catalyze their own removal from primary transcripts (Schmelzer and Schweyen, 1986; Peebles et al., 1986; van der Veen et al., 1986; Schmidt et al., 1990; Costa et al., 1997a; Toor et al., 2006). Several organellar group II introns have been shown to self-splice *in vitro*, including Sc.cob/1 in the *cob* gene and Sc.cox1/1 and Sc.cox1/5γ in the gene encoding cytochrome oxidase subunit 1 in *S. cerevisiae* (reviewed in Michel and Ferat, 1995). For these introns, however, optimal ribozyme activity was observed under elevated temperature and ionic conditions: for example, Sc.cox1/5γ self-splices at 45 °C in the presence of 0.5 M $(NH_4)_2SO_4$ and 0.1 M Mg^{2+} (Jarrell et al., 1988). In contrast, intron 2 in the LSU rRNA gene of the brown alga *Pylaiella littoralis*, Pl.LSU/2, was shown to have an unusually low requirement for Mg^{2+}, as low as 0.1 mM (Costa et al., 1997a).

As shown in Figure 1.5B, self-splicing via the branching mechanism proceeds by two transesterification reactions and leads to the excision of the intron as a branched, or lariat, molecule with a characteristic 2'-5' phosphodiester bond (Daniels

et al., 1996; van der Veen et al., 1986; van der Veen et al., 1987; Vogel and Börner, 2002). In the first step, the 2' OH of the branch-point adenosine residue initiates a nucleophilic attack at the phosphate group at the 5' splice site, forming a lariat intron intermediate with a 2'-5' phosphodiester linkage. The 5' exon interacts with the intron-lariat-3'exon intermediate via IBS1 and IBS2. In the second step, the 3' OH of the 5' exon carries out a nucleophilic attack at the 3' splice site, ligating the exons and releasing the intron as a lariat molecule. The two transesterification reactions are reversible and the rate constants for the forward and reverse reactions in the first step are similar. The first step, or branching reaction, is the rate-limiting step. Both steps are carried out in a single active site; however, base modification interference studies have demonstrated that there is a conformational change during the second step involving a GNRA tetraloop in DVI and its receptor in DII that releases the lariat-3'exon intermediate and aligns the 5' exon for the second attack (Chanfreau and Jacquier, 1996).

Under some circumstances, such as deletion of the branch-point sequence/structure or non-physiological salt concentrations, the process may be initiated by hydrolysis at the 5' splice site, with release of the intron as a linear molecule (Michel et al., 1989, Fedorova and Zingler, 2007). The type of monovalent ion may also influence the splicing pathway. For instance, NH_4^+ promotes splicing via the branched pathway, whereas K^+ leads to splicing predominately via the hydrolytic pathway. In the latter pathway, the initiating nucleophile is water or a hydroxyl ion. The second step is the same in both pathways. However, in terms of reverse splicing *in vivo* the branched intron structure is essential for retrohoming, since the second transesterification step in the reverse splicing reaction generates at the 5' splice site a

3’-5’ phosphodiester bond at the expense of the lariat 2’-5’ bond (Roitzsch and Pyle, 2009).

1.6.5. Intron-encoded proteins

RT ORFs are found in nearly all bacterial group II introns but are often absent or degenerated in organellar group II introns (reviewed in Lambowitz and Zimmerly, 2004; Simon et al., 2009). Intron-encoded proteins (IEPs) are multifunctional proteins with RT, maturase (X), DNA binding (D), and ENase (En) domains (Figure 1.6B), although there are examples of RTs that lack an En domain (Lambowitz and Belfort, 1993; Michel and Ferat, 1995; Matsuura et al., 1997; Lambowitz et al., 1999; Lambowitz and Zimmerly, 2004). The RT domain is situated at the N-terminus of the protein and contains the conserved sequence block RT-0, which is similar to RTs found in non-long terminal repeat retrotransposons, and sequence blocks RT-1 to -7, which corresponds to the palm and finger region of retroviral and other RTs (Zimmerly et al., 2001; Lambowitz and Zimmerly, 2004). Downstream of the RT segment is the X domain, and both the RT and X domains bind to the intron RNA during splicing. Both the D and En domains are located at the C-terminus; the D domain is predicted to contain an α-helix and the En domain has an H-N-H motif, analogous to that found in the H-N-H family of HEases (reviewed in Chevalier and Stoddard, 2001). These latter two regions are absent in some RTs (Lambowitz and Zimmerly, 2004). Maturase activity facilitates efficient intron splicing *in vivo* by stabilizing the catalytically active structure. The D and En domains promote intron mobility by binding and unwinding the DNA and nicking the bottom strand, and the RT region synthesizes a cDNA copy of the intron during retrohoming and retrotransposition (reviewed in Lambowitz and Zimmerly, 2004). Deletion of both D

and En domains blocks reverse splicing of the intron into double-stranded DNA, although the intron can still reverse splice into single-stranded DNA at lower levels (Lambowitz and Zimmerly, 2004). Presumably the RT and/or X domains are also involved in DNA binding (Lambowitz and Zimmerly, 2004).

Notably, a survey of group II introns and their RT ORFs in bacteria and organelles indicated that the two elements appear to co-evolve (Toor et al., 2001). Phylogenetic analysis of the RT and X domains revealed that the RT ORFs fall into several subgroups (A to F, mt-like, cp1-like, and cp2-like), and each ORF class is associated with a particular intron structural class (Toor et al., 2001). Analysis of the co-evolution of group II intron structures and RTs led to the proposal of the retroelement ancestor hypothesis, a model describing the continuous evolution of group II introns (Toor et al., 2001). The authors proposed that all group II introns originated from a single group II intron encoding an RT in bacteria and structural differences represent different stages of degeneration (Toor et al., 2001; Simon et al., 2009). According to the model, intron RNA structures have co-evolved with the RT IEP, likely because of selective pressures on the intron and IEP to maintain interactions involved in formation of the RNP particle, which is essential for both efficient intron splicing and promoting intron mobility. Moreover, mobility originated in bacteria, and ORF-less variants of cp- and mt-like introns are the result of ORF loss. However, a recent large-scale phylogenetic analysis of group II intron and intron RT sequences in bacteria revealed a single phylogeny for the intron sequences, which were aligned using conserved structural features as a guide, and two phylogenies for the ORF, one for the amino acid sequence and another for the nucleotide sequence (Simon et al., 2009). Several major issues that could potentially affect phylogenetic inferrence were identified: mutational saturation; base composition heterogeneity,

especially in the third codon position; low support for short nodes; and taxon sampling sensitivity, particularly in the ORF analysis based on amino acid sequence (Simon et al., 2009). These observations prompted the authors to suggest that the co-evolution of these elements, the foundation of the retroelement ancestor hypothesis, may be imperfect.

1.6.6. Unusual group II introns encoding homing endonuclease genes

Unusual group II introns with ORFs encoding putative LHEases were identified within the mt SSU and LSU rRNA genes of fungi belonging to the Ascomycota and Basidiomycota (Michel and Ferat, 1995; Toor and Zimmerly, 2002; Monteiro-Vitorello et al., 2009), although neither the introns nor the ORFs were characterized functionally in these studies. Genes coding for LHEases are commonly associated with group I and archaeal introns or with inteins (sequences that self-splice post-translationally) or they exist as free-standing ORFs (Dujon, 1980; Dürrenberger and Rochaix, 1993; Belfort and Roberts, 1997; Jurica and Stoddard, 1999; Chevalier et al., 2005; Gibb and Hausner, 2005; Stoddard, 2006; Sethuraman et al., 2008, 2009a; Bae et al., 2009; Singh et al., 2009). Homing endonucleases typically promote the propagation of their ORF sequence and may also promote intron mobility.

Some LHEases have been shown to function as maturases and promote the splicing of their host intron and sometimes related introns (Lazowska et al., 1989; Ho et al., 1997; Ho and Waring, 1999; Bassi et al., 2002; Bassi and Weeks, 2003; Belfort, 2003; Longo et al., 2005). On the other hand, the relationship between a group II intron and its RT is more specific and the intron RNA and IEP form an RNP particle, which is essential for splicing and mobility *in vivo*. Models of RNA secondary structure indicate that the introns are members of the IIB1 subclass. Introns related to

intron 3 (designated C.p.SSUi3 or mS952) inserted within the mt *rns* gene of *C. parasitica* contain the ORF in DIII (Toor and Zimmerly, 2002; Mullineux et al., 2010), a region known to enhance the efficiency of splicing. Conversely, in both intron 1 of the mt *rns* gene of *C. parasitica* (C.p.SSUi1, or mS785) and intron 5 of the mt *rnl* gene of *Agrocybe aegerita* the ORF is inserted in DIV (Toor and Zimmerly, 2002), which is the insertion site (IS) of most RT ORFs in canonical group II introns.

1.6.7. Host factors involved in intron splicing

Intron-catalyzed splicing *in vitro* requires unphysiological temperatures and salt concentrations, so efficient splicing *in vivo* is expected to depend upon proteins that are encoded by the intron and/or by the host, which have been co-opted to function as maturases or chaperones (Cui et al., 2004; Huang et al., 2005; Mohr et al., 2006; Solem et al., 2006; Halls et al., 2007; Zingler et al., 2008). Splicing of ORF-less introns certainly requires host-encoded factors, although one cannot discount the possibility that other IEPs may moonlight as splicing factors for ORF-less introns. The nuclear-encoded MSS116 protein in *S. cerevisiae* contains a DEAD-box motif found in RNA helicases and has been shown to promote efficient splicing of both group I and group II introns (Niemer et al., 1995; Huang et al., 2005; Solem et al., 2006; Halls et al., 2007). These proteins likely recognize consensus sequences or core structures and either stabilize catalytically-active structures or prevent the intron from adopting kinetically stable but catalytically inactive structures under cellular conditions, in which the temperature and ionic concentrations are usually below the optimum for self-splicing in the absence of protein co-factors (Weeks, 1997; Ho and Waring, 1999; Zingler et al., 2008).

An interesting variation on the theme of group II introns co-opting host-encoded splicing factors is found within plants. The *trnK*I1 intron is a member of group IIA1 that is inserted within the cp *trnK* gene, encoding $tRNA^{LYS}$, and the intron encodes the *matK* ORF (Hausner et al., 2006b). The *matK* ORF is a degenerate RT-type ORF with a conserved X domain and is a widely-used phylogenetic marker in plants, as it is ubiquitous, on the one hand, and rapidly evolving, on the other (Hausner et al., 2006b and references therein). Sequence and structural analyses of the intron in organisms ranging from algae to angiosperms revealed that the intron is degenerated. In the non-photosynthetic plant *Epifagus virginiana*, the *trnK* gene, along with its intron, is absent but the *matK* gene has been retained as a free-standing ORF, leading to the suggestion that MatK may function as a general splicing factor (Wolfe et al., 1992; Hausner et al., 2006b).

1.6.8. Intron mobility: retrohoming and retrotransposition

During expression of the host gene, the intronic ORF is translated and an RNP particle is formed between the intron lariat and IEP. In a process referred to as "retrohoming" the intron and its ORF colonize typically intron-less cognate alleles through a site-specific DNA integration mechanism. Retrohoming involves both the intron RNA and the IEP and is completed by host DNA repair mechanisms (Curcio and Belfort, 1996; Cousineau et al., 1998; Smith et al., 2005). In stark contrast to group I intron mobility, mutations in the group II sequence that interfered with intron splicing also inhibited mobility, suggesting that the intron RNA was involved as a mobility intermediate (reviewed in Lambowitz and Zimmerly, 2004). This was confirmed by Cousineau et al. (1998). In this study, the authors inserted a self-splicing group I intron and kan^{R} marker in DIV and showed that the group I intron was absent

after homing, demonstrating that the mobility intermediate is RNA. At lower frequencies, the intron may target ectopic sites and this form of mobility is referred to as retrotransposition.

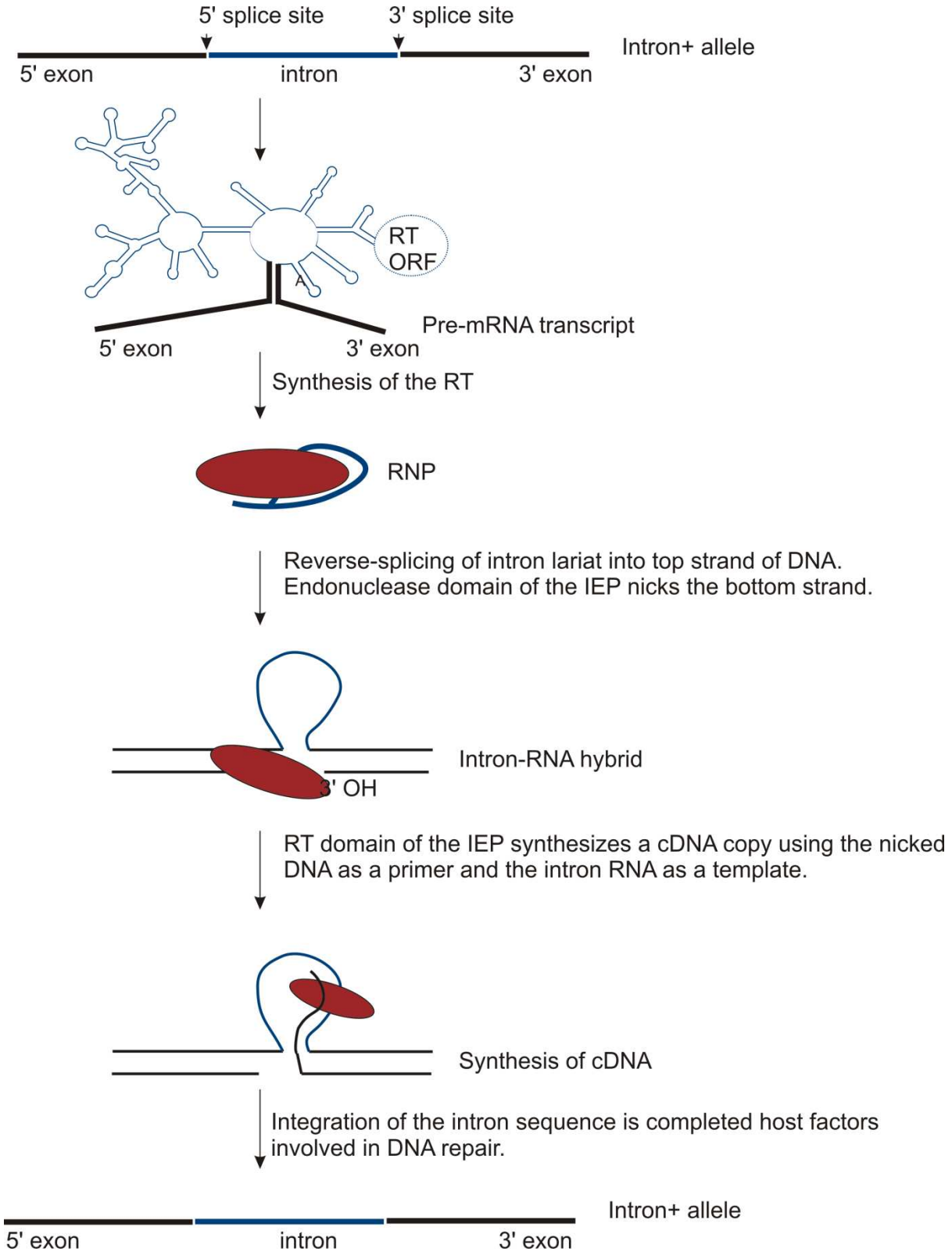
Multiple pathways for intron mobility have been described (reviewed in Lambowitz and Zimmerly, 2004). One of the major mechanisms of intron mobility is target DNA-primed reverse transcription (TPRT). The RNA component of the RNP recognizes the DNA target site by base pairing interactions between the EBSs and complementary bases in the target site; the intron cleaves the top (sense strand) and reverse splices, forming an RNA-DNA hybrid (Figure 1.7). The ENase domain of the IEP cleaves the antisense strand, and the RT domain uses the free 3' OH of the DNA strand as a primer for cDNA synthesis (Zimmerly et al., 1995a, 1995b). In ENase-independent mobility, the intron may reverse splice into single-stranded DNA near replication forks and use either a nascent leading or lagging strand or a random nick on the opposite strand to prime cDNA synthesis (Zhong and Lambowitz, 2003; Lambowitz and Zimmerly, 2004; Martínez-Abarca et al., 2004). Intron mobility may also occur via homologous recombination at double-stranded gaps in the DNA and may or may not be accompanied by coconversion of one or both flanking exon sequences.

A large fraction of organellar group II introns are presumed to have lost RT-catalyzed retrohoming as either they are ORF-less or the polymerase domain of the RT ORF is degenerated. Interestingly, relief from the selective pressures associated with protein-supported retrohoming has not led to the widespread adoption of splicing by the hydrolytic pathway: in the cp genome of angiosperms, the only RT-devoid system investigated in detail (Vogel and Börner, 2002), evidence for release of the intron as a lariat structure was obtained for all but one of the seventeen group II

introns. Interestingly, this pathway is exactly the same as the one used by the spliceosome to excise intron sequences from eukaryotic nuclear pre-mRNA transcripts.

Recently, global regulator molecules in *E. coli*, cAMP and ppGpp, which are involved in the stringent response to nutrient deprivation, were shown to play a role in regulating retromobility (Coros et al., 2009). Disruption of the *cyaA* and *spoT* genes, which respectively encode adenylate cyclase and ppGpp synthetase II, involved in the synthesis of cAMP and ppGpp, inhibited retromobility. Conversely, retromobility in wild-type *E. coli* was stimulated under low glucose or amino acid levels or by supplementation of cAMP to mutant cells. In addition, invasion of sites in plasmid DNA was favoured over chromosomal targets. The authors suggested that the greater efficiency of plasmid invasion may be due to changes in the nucleoid structure that are unfavourable for retrohoming events, including repositioning of the RNA polymerase throughout the nucleoid. The availability of replication forks was also proposed to be a likely factor (Coros et al., 2009). The efficiency of retrotransposition was greater when the replication fork stalled, allowing the IEP to use a nascent leading or lagging strand to prime cDNA synthesis at replication forks.

Figure 1.7. Group II intron mobility by target DNA-primed reverse transcription. The IEP binds to the intron RNA in DIV and, via its maturase activity, promotes intron splicing. The intron component of the RNP recognizes and binds to the DNA target site using base pairing interactions between the intron (EBS sites) and exons (IBS sites). The intron nicks the sense strand and reverse splices. The antisense strand is cleaved by the En domain of the IEP, generating a primer for reverse transcription. The RT domain synthesizes a cDNA copy of the intron, and host-encoded factors involved in DNA repair likely complete the integration of the intron sequence into the site.


5' splice site
3' splice site
Intron+ allele
5' exon
intron
3' exon
RT
ORF
A
Pre-mRNA transcript
5' exon
3' exon
Synthesis of the RT
RNP
Reverse-splicing of intron lariat into top strand of DNA.
Endonuclease domain of the IEP nicks the bottom strand.
Intron-RNA hybrid
3' OH
RT domain of the IEP synthesizes a cDNA copy using the nicked
DNA as a primer and the intron RNA as a template.
Synthesis of cDNA
Integration of the intron sequence is completed host factors
involved in DNA repair.
Intron+ allele
5' exon
intron
3' exon

1.7. Homing Endonucleases

1.7.1. Families of homing endonucleases

There are currently five families of HEases recognized, and each family is defined by conserved amino acid motifs: LAGLIDADG; H-N-H; His-Cys box; GIY...YIG; and PD-D/E-XK (Aggarwal, 1997; Belfort and Roberts, 1997; Chevalier and Stoddard, 2001; Burt and Koufopanou, 2004; Chevalier et al., 2005; Marcaida et al., 2010). GIY...YIG HEases are found within organellar group I introns and in bacteriophage genomes, where they are present as free-standing ORFs and within introns. Members of the H-N-H and His-Cys box share structural similarities, including a conserved ββαMe motif, which have led to the suggestion that the two families should be amalgamated into one superfamily. The H-N-H motif is also present within the ENase domain of group II IEPs of the RT family. The largest and best characterized is the LAGLIDADG family. Proteins in this class have either one or two of the dodecapeptide (referred to as P1 and P2) motifs. LAGLIDADG-type HEases with two motifs are believed to have arisen from a gene duplication event of an ancestral gene with a single dodecapeptide motif within an LSU gene (Haugen and Bhattacharya, 2004).

1.7.2. Nomenclature

The proposed nomenclature of HEases resembles that of restriction enzymes (Roberts et al., 2003); the designation contains the first letter of the genus name and first two letters of the species name. Preceding this is an additional designation, indicating whether the HEase ORF is intron-encoded (I), intein-encoded or protein-inserted (PI), or free-standing (F). Lastly, the HEase is assigned a roman numeral

based on the order in which it has been characterized and published. Thus, "I-SceII" would represent the second intron-encoded HEase characterized from *S. cerevisiae.*

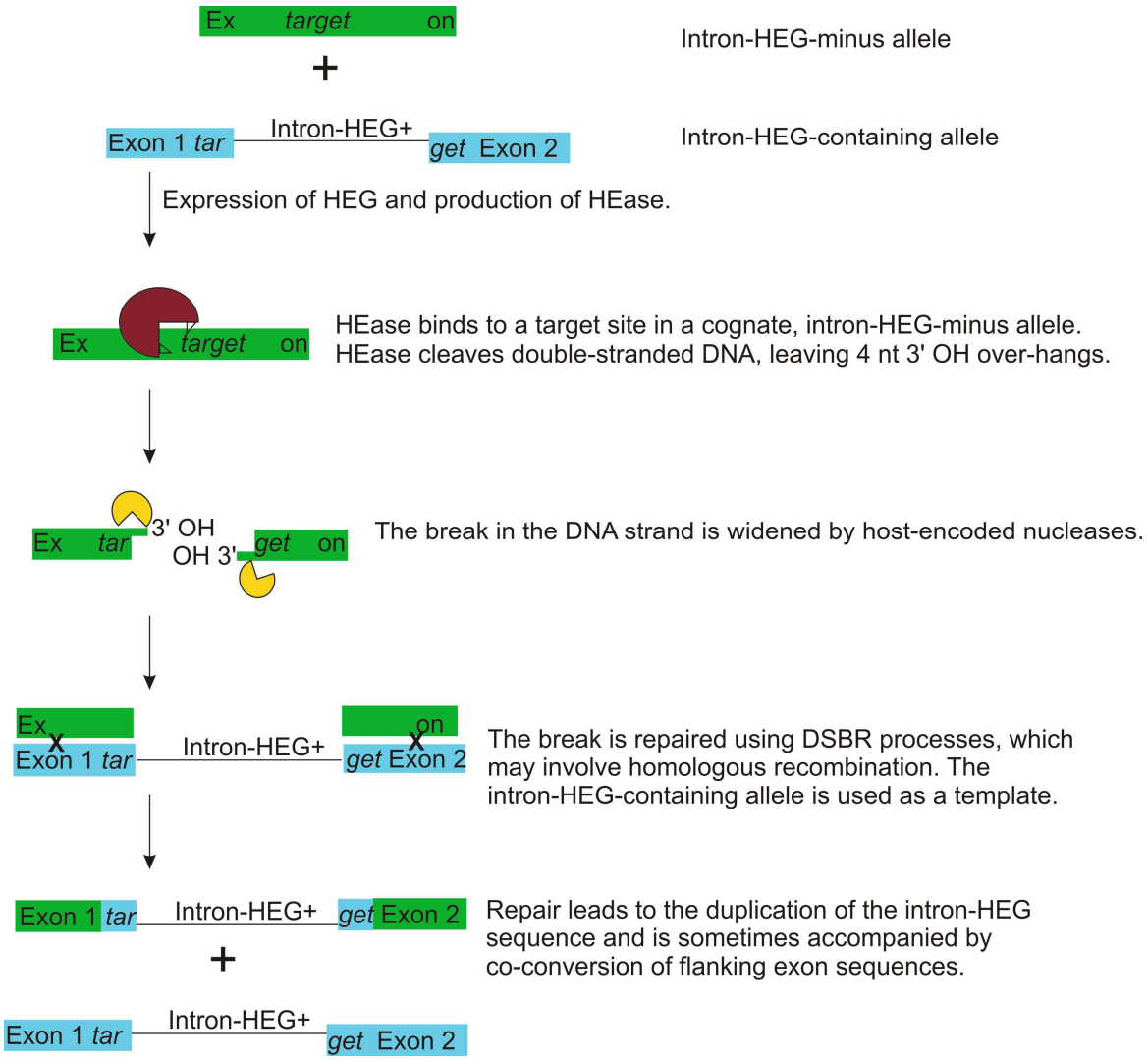
1.7.3. Structural studies of LAGLIDADG homing endonucleases

Crystal structures are available for several LHEases bound to their DNA substrates, including PI-SceI, I-CreI, I-AniI, I-SceI, and I-MsoI (Heath et al., 1997; Jurica et al., 1998; Moure et al., 2002, 2003; Longo et al., 2005). Despite amino acid sequence variability amongst the LHEases, structural studies indicate that members of this family share a common conserved αββαββα topology (Aagaard, 1997; Chevalier and Stoddard, 2001; Chevalier et al., 2005). The LAGLIDADG motif is situated near the base of an α-helix; the active site consists of two LAGLIDADG motifs, with the α-helices packed closely together, and each motif is involved in cleaving the DNA sugar-phosphate backbone. The four β-sheets are antiparallel, forming a saddle across the major groove of DNA. The β strands are responsible for recognizing and binding the DNA target site, which occurs via phased, subsaturating interactions between the amino acid residues in the strands and the bases and backbone groups in the DNA target site. LHEases with single P1 motifs, including I-CreI, I-CeuI, and I-MsoI, form homodimers and recognize palindromic target sites, whereas double-motif LHEases, such as I-AniI, I-DmoI, and I-SceI, are monomers and recognize pseudopalindromic target sites (Aagaard et al., 1997; Jurica et al., 1998; Lucas et al., 2001; Moure et al., 2003; Eastberg et al., 2007).

1.7.4. Role as DNA endonucleases

LAGLIDADG-type HEases catalyze cleavage of double-stranded DNA using a canonical two-metal ion phosphoryl hydrolysis pathway (reviewed in Jurica and Stoddard, 1999; Chevalier et al., 2005; Stoddard, 2006; Marcaida et al., 2010). Each LAGLIDADG motif contributes one catalytically essential residue; this conserved residue (the penultimate D or E residue) coordinates bound metal ions, such as Mg^{2+} or Mn^{2+}. Mutation of key residues can abolish cleavage activity entirely or result in single-strand cleavage, thus generating a so-called "nickase" (McConnell Smith et al., 2009). The protein cleaves across the minor groove of DNA, generating 4 nt 3' OH overhangs. The initial double-stranded break is often expanded by exonucleases, and the gap is repaired by homologous recombination or by nonhomologous end joining. A homologous donour sequence is used as the template for DNA repair in the former pathway and in the latter pathway the gap is repaired by rejoining the DNA ends; some loss of sequence information may result from exonuclease activity (McConnell Smith et al., 2009). Homing may also be accompanied by coconversion of flanking exon sequences (Figure 1.8).

Figure 1.8. Homing of introns and homing endonuclease genes is promoted by the DNA cleavage activity of the homing endonuclease. Double-stranded breaks in DNA stimulate host DNA repair mechanisms, typically via the double-strand break repair (DSBR) pathway (reviewed in Lambowitz and Belfort, 1993; Belfort et al., 2002). Repair of the break in the intron-HEG-minus allele may use the intron-HEG-plus allele as a template, which leads to the duplication of the intron-HEG sequence into the cognate allele. The recognition site of the HEase is disrupted as a result. Homologous recombination may be accompanied by co-conversion of flanking exon sequences.


Ex target on
+
Exon 1 tar
Intron-HEG+
get Exon 2
Intron-HEG-minus allele
Intron-HEG-containing allele
Expression of HEG and production of HEase.
Ex target on
HEase binds to a target site in a cognate, intron-HEG-minus allele.
HEase cleaves double-stranded DNA, leaving 4 nt 3' OH over-hangs.
Ex tar
3' OH
OH 3'
get on
The break in the DNA strand is widened by host-encoded nucleases.
Ex
Exon 1 tar
Intron-HEG+
on
get Exon 2
The break is repaired using DSBR processes, which may involve homologous recombination. The intron-HEG-containing allele is used as a template.
Exon 1 tar
Intron-HEG+
get Exon 2
+
Exon 1 tar
Intron-HEG+
get Exon 2
Repair leads to the duplication of the intron-HEG sequence and is sometimes accompanied by co-conversion of flanking exon sequences.

1.7.5. Role as RNA maturases

Some LHEases have been shown to function as maturases and promote the splicing of their host group I intron and occasionally related introns (Lazowska et al., 1989; Ho et al., 1997; Ho and Waring, 1999; Bassi et al., 2002; Bassi and Weeks, 2003; Belfort, 2003; Longo et al., 2005). Maturase activity was initially demonstrated genetically. Mutations in mt group I IEPs within *cob*-I2, -I3, and -I4 and group II IEPs within *cox1*-I1 and -I2 that prevent splicing of the intron could be repressed if the wild-type protein was provided *in trans* (reviewed in Lambowitz and Belfort, 1993). Ho and co-workers (1997) developed an assay for maturase activity *in vitro* with the AnCOB maturase, encoded by a group I intron ORF within the apocytochrome b gene of *A. nidulans*. Interestingly, the AnCOB maturase is bifunctional; it also moonlights as an active ENase, cleaving 1 nt upstream of the intron IS (Ho et al., 1997).

In a group I intron-encoded LAGLIDADG-type maturase protein, the bI3 maturase, a mutation in the active site disrupted ENase activity; this protein likely evolved from a DNA-binding to an RNA-binding protein (Bassi et al., 2002; Bassi and Weeks, 2003; Longo et al., 2005) and facilitated intron splicing at the expense of mobility. The authors proposed a model for the evolution of the bI3 maturase in which they suggest that initially the LHEase was an active DNA ENase and invaded the target site and subsequent mutations resulted in the evolution of RNA binding properties. Genes encoding LHEases spread rapidly in the genome and possibly the population (by horizontal transmission) and may also insert into ectopic sites at lower frequency because of flexibility in DNA target site recognition. Once all possible homing sites have been occupied, there is little selective pressure on the ENase activity and the ORF sequence can accumulate mutations that may fortuitously result in maturase activity. The LHEase is then effectively co-opted to promote efficient

splicing of its own or related intron (Longo et al., 2005). The life cycle of HEases will be discussed in greater detail in section 5.3.

1.8. Applications of group II introns and LAGLIDADG homing endonucleases

1.8.1. Genetic manipulation using group II introns and homing endonucleases

Both group II introns and HEases have applications in biotechnology, including the use of catalytic RNA molecules for developing new strategies for repairing mutant genes, site-specific genome modifications, cleaving viral transcripts, down-regulating oncogenes or cleaving mRNAs from defective genes (Guo et al., 2000; Karberg et al., 2001; Khan and Lal, 2003; Frazier et al., 2003; Perutka et al., 2004; Jones et al., 2005; Yao et al., 2005; Yao and Lambowitz, 2007; Shao et al., 2007; Nazari and Joshi, 2008; Mastroianni et al., 2008; Zhuang et al., 2009). "*In vitro*" evolution of ribozymes and "directed evolution" of HEases for genomic engineering and other biotechnological applications frequently use naturally-occurring ribozymes or HEases as starting material (Chen and Zhao, 2005; Thyme et al., 2009; Marcaida et al., 2010). Monomeric LHEases are potential candidates for engineering novel ENases; their long (greater than 20 bp) target sites make them useful as rare-cutting enzymes for whole-genome studies. Also, their recognition of non-palindromic target sites renders them amenable to genetic engineering for novel target site recognition (Thyme et al., 2009).

1.8.2. The development of group II intron gene targeting vectors ("targetrons")

Two features of group II introns make them potential candidates as gene targeting vectors, or "targetrons". Firstly, it is theoretically possible to direct the

intron to insert site-specifically into a desired target simply by changing the intronic sequences of the EBS1, EBS2, and EBS3/δ sites to base pair with sequences in the DNA target site. Secondly, the intron can be used as a shuttle by inserting a sequence or gene of interest into DIV. The application of group II introns as targetrons has been tested for both gene disruption and gene repair strategies (Guo et al., 2000; Karberg et al., 2001; Frazier et al., 2003; Perutka et al., 2004; Jones et al., 2005; Yao et al., 2005; Yao and Lambowitz, 2007; Shao et al., 2007; Mastroianni et al., 2008; Zhuang et al., 2009). To generate an unconditional gene disruption, the intron can be directed to insert into the antisense strand from which it cannot splice. Conditional gene disruptions, on the other hand, can be generated by targeting an ORF-less group II intron to the sense strand, where intron splicing, and ultimately gene expression, is linked to the inducible expression of the IEP *in trans* (reviewed in Lambowitz and Zimmerly, 2004). In gene repair strategies the gene or sequence of interest is cloned into DIV of the intron, and the IEP, involved in RNP formation and reverse splicing, is expressed *in trans*.

1.8.3. Application of homing endonucleases as gene targeting elements

Rare-cutting enzymes have great potential as tools for genomic engineering, including gene repair and viral inactivation (Windbichler et al., 2007; Grizot et al., 2009; McConnell Smith et al., 2009; Traver et al., 2009; Marcaida et al., 2010; Gao et al., 2010). One example involves the homodimer I-*Cre*I, which was engineered to target the human *RAG1* gene; mutations in this gene are responsible for severe combined immunodeficiency (SCID). Because two LAGLIDADG motifs are present in the active enzyme, it was possible to target specific loci using chimeric HEases or engineered heterodimers; recombination efficiencies with the engineered I-C*re*I

reached 6 % in transfected human cells (Grizot et al., 2009). The one caveat to this approach is the potential for the two homodimers to form, in addition to the heterodimer, thus, interfering with DNA cleavage. In the context of developing effective and reliable human gene therapy approaches, this may be catastrophic because of the risk of toxicity or undesirable cleavage (Grizot et al., 2009). One possible solution proposed by the authors was the design of obligate heterodimers (Grizot et al., 2009). In this approach the C-terminus of the first LAGLIDADG motif was connected to the N-terminus of the second motif by a linker region, whose propensity to form secondary structures, including α-helices, could be predicted and the length and sequence composition could be manipulated. The dimerization surface could also be optimized through amino acid substitution to prevent homodimers from forming (Grizot et al., 2009).

1.8.4. Engineering homing endonucleases

The recent study by Thyme et al. (2009) highlighted the importance of determining substrate-enzyme binding energies when engineering HEases using computational-based directed evolutionary approaches. The effects of single-base mutations in the target site on the binding affinity of the monomeric LHEase I-*Ani*I were found to be asymmetric. Using fluorescence-binding assays, the authors demonstrated that substrate binding was reduced by mutations on the "left-hand" side of the target site (positions -10 to -3) but only slightly reduced by mutations in the "right half" (-2 to +10). While target site mutations decreased the binding affinity overall, mutations occurring in the left-hand side of the site (which affect enzyme-substrate binding energy) led specifically to a decrease in enzyme activity by increasing both the disassociation constant (K_D) and the pseudo-Michaelis-Menten

constant under single-turnover conditions (K_M*). In contrast, base changes in the right-hand side of the target, which affect the stability of the transition state, led to a reduction in the apparent turnover number (k_{cat}*). These observations were confirmed using surface display and tethered cleavage assays. In this approach, a DNA duplex containing the mutant site was tethered to the enzyme, which was displayed on the surface of a yeast cell. For substrates containing left-side mutations, the increased local concentration of the substrate "suppressed" the decreased ground-state binding affinity and cleavage activity was restored. However, the cleavage efficiency of substrates with right-side mutations was not rescued. These results indicated that despite the apparent two-fold symmetry of the enzyme-substrate complex, substrate binding and transition-state stabilization were, in fact, separate from each other.

1.9. Research objectives

1.9.1. Identifying constraints that influence the evolution of insertion segments in nuclear ribosomal DNA and assessing the impact on phylogenetic studies

To better understand the evolutionary dynamics of insertion elements, a robust phylogenetic analysis of the host organisms was required. For this purpose, phylogenetic relationships amongst strains of *Leptographium* and related taxa were determined using the ITS region in nuclear rDNA. At the DNA level, ITS segments are rapidly evolving, while RNA structural features remain conserved. Thus, structural features may be used to guide the alignment of ambiguous regions, which may result in a more robust phylogenetic analysis. However, the mechanisms involved in structural conservation, such as compensatory base changes (CBCs),

insertions/deletions (indels), and possible RNA strand slippage, are not well-understood.

The goals of the study were: (i) to identify mechanisms involved in conserving RNA secondary structure in spite of the rapid evolution of the DNA sequence and (ii) to assess both the level of branch support and the effects on tree topology obtained from the phylogenetic analyses of the entire ITS1-5.8S-ITS2 region and of each segment individually. To address the first issue comparative sequence analysis was used to develop models of RNA secondary structure and to identify the potential effects of CBCs, hemi-CBCs, and indels on secondary structure. This may assist in generating improved structure-guided DNA sequence alignments of rapidly-evolving regions, including ITS. To examine the second issue, phylogenetic analyses were conducted on the entire ITS-5.8S rDNA region and each ITS segment individually using multiple approaches.

1.9.2. Screening for the presence of optional genetic elements in the mitochondrial DNA of *Leptographium* species

The major objectives of the PCR screen were: (i) to identify, based on size and phylogenetic distribution, potentially novel insertion elements in the mt DNA of members of *Leptographium*; (ii) to understand the mode of transmission (that is, acquisition by horizontal or vertical transmission or by random gain); and (iii) to assess whether or not there was a relationship between the distribution of insertions and the phylogenetic relationship of strains and if the presence/absence of insertions could be used as a molecular marker for the rapid identification of potential plant pathogens. To address the second and third issues, data on the presence/absence of

insertions were superimposed onto the phylogenetic tree of the host organisms; this tree was inferred using the nuclear ITS region, as described in 1.9.1.

1.9.3. Biochemical and evolutionary characterization of group II introns and their putative LAGLIDADG homing endonucleases genes

This study is subdivided into three components. The first part involves: (i) the identification of group II introns/LHEG composite elements within mt *rns* gene fifty strains belonging to *L. lundbergii*, *L. truncatum*, *L. procerum*, *L. terebrantis*, *L. wingfieldii*, *Grosmannia aurea*, and *G. penicillata* and (ii) the characterization of these elements in terms of genomic location (host gene), size, and conserved amino acid sequence motifs. Following the identification of group II intron/LHEG composite elements, the next component focussed on investigating the catalytic activities of the intron and LHEase separately and together as a composite element. The specific questions that were examined were: (i) Can ORF-less and ORF-containing intron constructs self-splice *in vitro* in the absence of protein co-factors? (ii) Does the LHEase have maturase activity and enhance the efficiency of splicing *in vitro*? and (iii) Does the putative ORF encode an active LHEase, and if so what is the cleavage site? Lastly, phylogenetic relationships amongst the host gene (mt *rns*), group II introns, and LHEase sequences were examined for the purpose of determining the evolutionary affiliations between the elements and their host gene/organism.

CHAPTER 2. GENERAL MATERIALS AND METHODS

2.1. Strains and growth conditions

Fungal cultures were grown on Malt Extract Agar, containing (g/L): malt extract (20); yeast extract (1); and bacteriological agar (20), at 25 °C for up to 5 days. Segments of the fungal mat were then transferred to 50 ml of PYG medium, containing (g/L): peptone (1); yeast extract (1); and glucose (3), and the cultures were incubated without shaking at 25 °C for up to 5 days.

2.2. DNA extraction and purification

The extraction and purification of fungal genomic DNA was carried out as previously described (Kim et al., 1990; Hausner et al., 1992). Briefly, fungal mycelia were harvested by filtering still-grown cultures through Whatman Grade no. 1 filter paper, with a particle retention size of 11 µm (GE Healthcare, Baie d'Urfe, Canada). Mycelia were disrupted by vortexing in 3 ml of extraction buffer [1 M Tris-Cl (pH 8.0), 0.5 M $Na_2EDTA{\cdot}2H_2O$ (pH 8.0), 5 M NaCl, 10 % (w/v) cetyltrimethylammonium bromide] and 4 ml of glass beads, and the lysate was then incubated at 55 °C for a minumum of 2 hours, though typically overnight, in 3 ml of extraction buffer and 660 µl of 20 % (w/v) sodium dodecyl sulphate (SDS). Contaminants were removed by chloroform extraction, followed by centrifugation at 3000 rpm for 30 minutes at room temperature. DNA was precipitated with 2.5 volumes of ice-cold 95 % ethanol overnight at -20 °C and then washed with 1 ml of 70 % ethanol. The DNA was pelleted by centrifugation at 3000 rpm for 30 minutes at room temperature. The dried pellet was resuspended in 300 µl of 1X TE buffer [10 mM Tris-Cl (pH 7.6) and 1 mM $Na_2EDTA{\cdot}2H_2O$ (pH 8.0)].

2.3. PCR amplification and product purification

2.3.1. ITS1-5.8S-ITS2 nuclear rDNA sequences

The ITS1-5.8S-ITS2 region was amplified by PCR using the 1.5X PCR Enhancer System (Life Technologies, Burlington, Canada), containing: 5 µl of 10X PCRx amplification buffer; 4 µl of 2.5 mM dNTP (200 µM, final concentration, of each dNTP); 1.5 µl of $MgSO_4$ (1.5 mM, final concentration); 0.5 µl (20 pmoles) of each primer; 7.5 µl of 10X PCRx enhancer solution; 0.25 µl Taq polymerase (approximately 1.25 units), 1 µl of genomic DNA (approximately 50-100 ng); and 29.75 µl of H_2O. The oligonucleotide primers (Alpha DNA, Montréal, Canada) used for amplification, SSZ, LS4, SS3, and LS2 (Hausner et al., 1993a; Hausner and Wang, 2005; Mullineux and Hausner, 2009), are listed in Table 2.1. The region amplified, along with the relative positions of the primers, is shown in Figure 2.2A. Amplification was carried out using the following standard conditions: initial denaturation (93 ºC, 2 min); and 30 cycles of denaturation (93 ºC, 1 min), primer annealing (50 ºC, 1 min), and extension (72 ºC, 1 min). When necessary, the temperature and length of the annealing and/or extension steps, template concentration, and the number of cycles were optimized to obtain single, sharp bands, which were visualized on an ethidium bromide-stained agarose gel. PCR products were purified using the Wizard® SV Gel and PCR Clean-Up System (Promega Corp., Madison, USA). Briefly, the solution containing the PCR product was mixed with an equal volume of binding solution (Promega Corp.), loaded onto a mini-column, incubated for 1 minute, and then centrifuged at 14 000 rpm for 1 minute. The sample was washed with 700 µl of membrane wash solution (Promega Corp.) and then centrifuged as above. The sample was washed a second time with 500 µl of

membrane wash solution and centrifuged as above for 5 minutes. The purified PCR product was eluted in 35-50 µl of nuclease-free H_2O (Promega Corp.) and stored at -20 °C.

2.3.2. Mitochondrial DNA

The following regions of mt DNA were PCR-amplified: the SSU rRNA (*rns*) gene; the U7 and U11 regions of the LSU (*rnl*) gene; and the cytochrome b (*cob*) gene. Amplification reactions were typically carried out in a solution containing: 5 µl of 10x *Taq* DNA polymerase buffer [10 mM Tris-Cl (pH 8.8), 1.5 mM $MgCl_2$, 50 mM KCl, 0.001 % (w/v) gelatin, final concentrations] (Stratagene, Agilent Technologies, La Jolla, USA); 0.75 µl of $MgCl_2$ (0.75 mM, final concentration); 4 µl of 2.5 mM dNTP (200 µM, final concentration, of each dNTP); 0.5 µl (20 pmoles) of each primer; 0.25 µl Taq polymerase (approximately 1.25 units), 1 µl of genomic DNA (approximately 50-100 ng); and 38 µl of H_2O. The concentration of $MgCl_2$ was optimized as required. The oligonucleotide primers used for amplification are described in Table 2.1 and the primer pairs IP1 and IP2 have been previously described (Bell et al., 1996). The regions amplified, showing the relative positions of the primers, are illustrated in Figure 2.2B-D. In general, the amplification conditions were: initial denaturation (93 ºC, 3 min); and 25 cycles of denaturation (93 ºC, 1 min), primer annealing (52.9 ºC, 1 min 30 s), and extension (70 ºC, 4 min). Amplification conditions were optimized to produce single sharp bands, which were visualized using an ethidium bromide-stained agarose gel, and the specific parameters used for amplification are described in Table 2.2 (the mt *rns* gene), Table 2.3 (the U11 region of the mt *rnl* gene), and Table 2.4 (the mt *cob* gene). PCR products were purified using the Wizard® SV Gel and PCR Clean-Up System (Promega Corp.).

Figure 2.1. Schematic of the gene regions that were amplified and the primers that were used in the amplification. (A) ITS1-5.8S-ITS2 nuclear rDNA region; (B) mt *rns*; (C) mt *rnl*; and (D) mt *cob* genes. The gene regions are not to scale relative to each other.

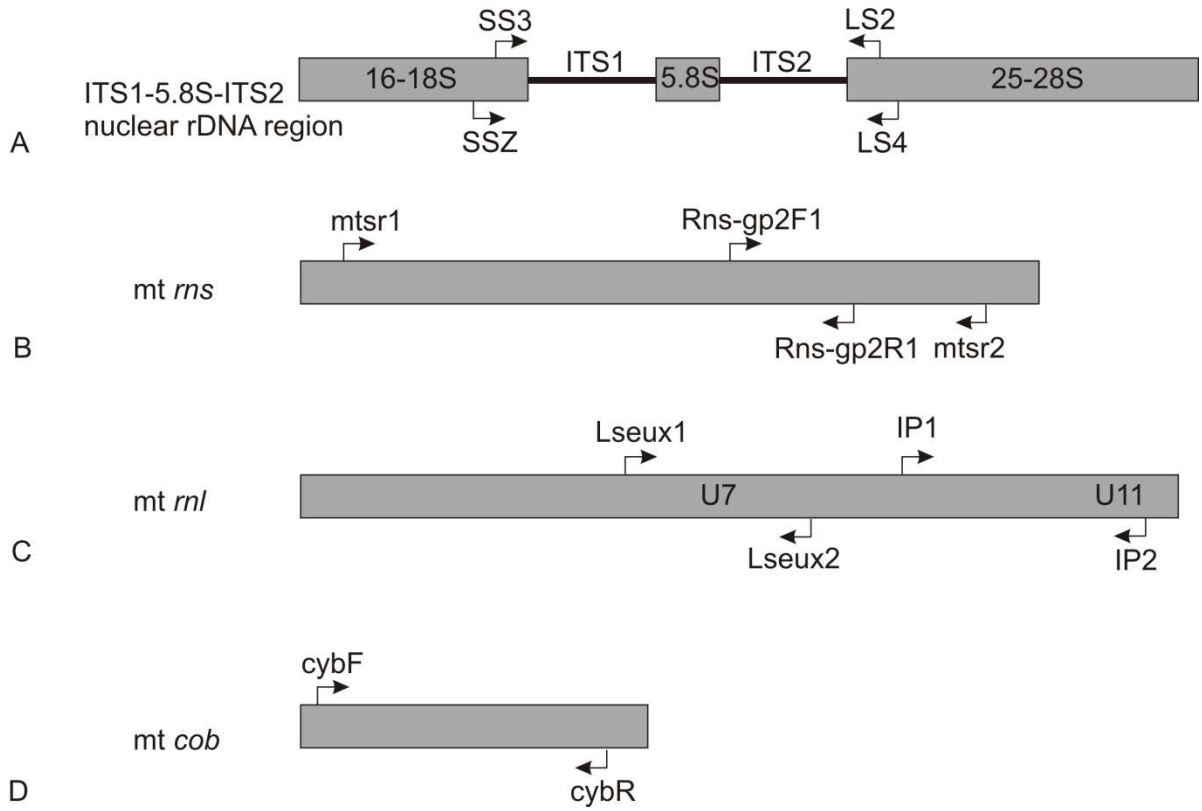

A
ITS1-5.8S-ITS2
nuclear rDNA region
SS3
SSZ
16-18S
ITS1
5.8S
ITS2
LS2
LS4
25-28S
B
mt rns
mtsr1
Rns-gp2R1
Rns-gp2F1
mtsr2
C
mt rnl
Lseux1
Lseux2
U7
IP1
IP2
U11
D
mt cob
cybF
cybR

Table 2.1. Primer sequences used for amplification reactions in each of the studies.

Gene region/Insertion	Primer name	Primer sequence (5'-3')	DNA strand orientation
nuclear ITS1-5.8S-ITS2 rDNA	SSZ	ATAACAGGTCTGTGATG	sense
	SS3	GTCGTAACAAGGTCTCCG	sense
	LS2	GATATGCTTAAGTCAGCG	antisense
	LS4	TTTGTGCGCTATCGGTCTC	antisense
mt *rns*	mtsr1	AGTGGTGTACAGGTGAG	sense
	mtsr2	CGAGTGGTTAGTACCAATCC	antisense
mt *rns* 3' terminal region	Rns-gp2F1	GAGTAACGTGGCAACACGGAAACTG	sense
	Rns-gp2R1	CATTAACTGGAAACAGCCGTGCAAC	antisense
mt *rns* cleavage site	mtsrF-2	AAATGATGAATGTCATAGG	sense
	mtsrR-2	TAGTTTCACAACATTAACTG	antisense
mt *rnl* U7 region	Lseux2F	GACCGCATTTAACGGCCAAGG	sense
	Lseux1R	GCTAGTAGAGAATACGAAGGC	antisense
mt *rnl* U11 region	IP1	GGAAAAGCTACGCTAGGG	sense
	IP2	GTTGCGCAAATTAGCC	antisense
mt *cob*	cybF[1]	TATTTACAYATAGGAGAGG	sense
	cybR[2]	AGGWATAGATCTTAATATAGC	antisense
LtrΔORF.3 intron construct	Ltr1	CCCAAACCCTGCAGAAACACCTCAAGAGTAACGTG	sense
	Ltr2	AAAGCTTACGCTTAAATTACTTCAGAATCAAGTCGTAGC	antisense
	Ltr3	ATTTAAGCGTAAGCTTTAATATATT	sense
	Ant7	AACAGCTATGACCATGATTACG	sense
	24-mer	CGCCAGGGTTTTCCCAGTCACGAC	antisense
I-LtrII gene (N-terminal His_6-tagged)	1435HEGF	CACCATGATTAACCTGAAGAATAAC	sense
	1435HEGR	TATTAAATACGGTTGGTGTTC	antisense

[1] Y = C, T.

[2] W = A, T.

Table 2.2. Description of amplification conditions for the mt *rns* gene.

Strain	$MgCl_2$ (mM final)	Initial denaturation	Denaturation	Annealing	Extension	Number of cycles
Grosmannia aurea CBS438.69	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1.5 m	70 °C, 4 m, final 70 °C, 10 m	30
Grosmannia penicillata DAOM63691	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	30 °C, 5 m	30
Grosmannia penicillata NFRI60-21	2.25-4.0	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 5 m	30
Leptographium sp. J.R.88-194A	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
Leptographium lundbergii CBS352.29	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii DAOM60397	1.5	93 °C, 3 m	93 °C, 1 m	51 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii DAOM63692	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1.5 m	72 °C, 4 m, final 70 °C, 10 m	25
Leptographium lundbergii DSMZ5010	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii NFRI60-25	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii NFRI69-148	2.25	93 °C, 3 m	93 °C, 1 m	55 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii NFRI89-1040/1/3	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
Leptographium lundbergii NFRI1502/1	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium procerum DAOM33940	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium procerum J.R.88-409A	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium procerum NFRI59-84/2	2.25-4.0	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 5 m	30
Leptographium procerum TOM55.35	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium procerum TOM62.30	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium procerum TOM73.12	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium procerum TOM76.36	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium procerum TOM76.8	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium procerum TOM86.19	2.25	93 °C, 3 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	30
Leptographium procerum UAMH9724	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium terebrantis CBS298.85	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1.5 m	72 °C, 4 m, final 70 °C, 10 m	25
Leptographium terebrantis CBS337.70	2.25-4.0	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 5 m	30
Leptographium terebrantis CBS408.61	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
Leptographium terebrantis UAMH9690	2.25	93 °C, 3 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	30
Leptographium terebrantis UAMH9722	2.25	93 °C, 3 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	30
Leptographium truncatum CBS647.89	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium truncatum CBS929.85	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium truncatum DAOM60396	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30

Leptographium truncatum Forintek C34	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
Leptographium truncatum J.R.88-324	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium truncatum J.R.88-449	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium truncatum NFRI59-7/3	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium truncatum NFRI1813/1	2.25	93 °C, 3 m	93 °C, 1 m	55 °C, 1 m	70 °C, 3 m	25
Leptographium truncatum TOM74.29	2.25-4.0	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 5 m	30
Leptographium truncatum TOM86.30	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium wingfieldii CBS345.90	2.25-4.0	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 5 m	30
Leptographium wingfieldii CBS645.89	2.25	93 °C, 3 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	30
Leptographium wingfieldii CBS648.89	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
Leptographium wingfieldii MCC125	2.25-4.0	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 5 m	30
	2.25	93 °C, 3 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	30
Leptographium wingfieldii MCC130	2.25-4.0	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 5 m	30
Leptographium wingfieldii MCC349	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
Leptographium wingfieldii NFRI88-369/11	2.25-4.0	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 5 m	30
Leptographium wingfieldii TOM1.3	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
Leptographium wingfieldii TOM5.1	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
Leptographium wingfieldii TOM9.4	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	30
Leptographium wingfieldii TOM10.2	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
Leptographium wingfieldii TOM11.5	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
	2.25	93 °C, 3 m	93 °C, 1 m	54 °C, 1 m	70 °C, 4.5 m	30
Leptographium wingfieldii TOM59.21	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
	2.25 -4.0	93 °C, 3 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	30

Table 2.3. Description of amplification conditions for the U11 region of the mt *rnl* gene.

Strain	$MgCl_2$ (mM final)	Initial denaturation	Denaturation	Annealing	Extension	Number of cycles
Grosmannia aurea CBS438.69	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Grosmannia penicillata DAOM63691	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Grosmannia penicillata NFRI60-21	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographiumlundbergii CBS352.29	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium lundbergii DAOM60397	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium lundbergii DAOM63692	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
Leptographium lundbergii DSMZ5010	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
Leptographium lundbergii NFRI60-25	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii NFRI69-148	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii NFRI89-1040/1/3	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium lundbergii NFRI1502/1	2.25	93 °C, 3 m	93 °C, 1 m	55 °C, 1 m	70 °C, 3 m	25
Leptographium procerum DAOM33940	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium procerum J.R.88-409A	2.25	94 °C, 2 m	93 °C, 1 m	51 °C, 1 m	70 °C, 5 m	30
Leptographium procerum NFRI59-84/2	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium procerum TOM55.35	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium procerum TOM62.30	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium procerum TOM73.12	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium procerum TOM76.36	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium procerum TOM76.8	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium procerum TOM86.19	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium procerum UAMH9724	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium terebrantis CBS298.85	2.25	94 °C, 2 m	93 °C, 1 m	51 °C, 1 m	70 °C, 5 m	30
Leptographium terebrantis CBS337.70	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium terebrantis CBS408.61	2.25	94 °C, 2 m	93 °C, 1 m	51 °C, 1 m	70 °C, 5 m	30
Leptographium terebrantis UAMH9690	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium terebrantis UAMH9722	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium truncatum CBS647.89	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium truncatum CBS929.85	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium truncatum DAOM60396	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30

Leptographium truncatum Forintek C34	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
Leptographium truncatum J.R.88-324	2.25	93 °C, 3 m	93 °C, 1 m	52.9 °C, 1 m	70 °C, 4 m	30
Leptographium truncatum J.R.88-449	2.25	93 °C, 3 m	93 °C, 1 m	55 °C, 1 m	70 °C, 3 m	25
Leptographium truncatum NFRI59-7/3	2.25	94 °C, 2 m	93 °C, 1 m	51 °C, 1 m	70 °C, 5 m	30
Leptographium truncatum NFRI1813/1	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium truncatum TOM74.29	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium truncatum TOM86.30	2.25	94 °C, 2 m	93 °C, 1 m	51 °C, 1 m	70 °C, 5 m	30
Leptographium wingfieldii CBS345.90	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium wingfieldii CBS645.89	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium wingfieldii CBS648.89	2.25	94 °C, 2 m	93 °C, 1 m	51 °C, 1 m	70 °C, 5 m	30
Leptographium wingfieldii MCC125	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium wingfieldii MCC130	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium wingfieldii MCC349	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium wingfieldii NFRI88-369/11	2.25	94 °C, 2 m	93 °C, 1 m	51 °C, 1 m	70 °C, 5 m	30
Leptographium wingfieldii TOM1.3	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium wingfieldii TOM5.1	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium wingfieldii TOM9.4	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium wingfieldii TOM10.2	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium wingfieldii TOM11.5	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30
Leptographium wingfieldii TOM59.21	2.25	93 °C, 3 m	93 °C, 1 m	52 °C, 1 m	70 °C, 3 m	30

Table 2.4. Description of amplification conditions for the mt *cob* gene.

Strain	$MgCl_2$ (mM final)	Initial denaturation	Denaturation	Annealing	Extension	Number of cycles
Grosmannia aurea CBS438.69	2.25	94 °C, 2 m	95 °C, 2 m	52.9 °C, 1.5 m	70 °C, 4.5 m, final 70 °C, 10 m	25
Grosmannia penicillata DAOM63691	2.25	95 °C, 2 m	95 °C, 1 m	52 °C, 1 m	70 °C, 5 m	35
Grosmannia penicillata NFRI60-21	2.25	95 °C, 2 m	95 °C, 1 m	52 °C, 1 m	70 °C, 5 m	35
Leptographium sp. J.R.88-194A	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii CBS352.29	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii DAOM60397	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii DAOM63692	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii DSMZ5010	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii NFRI60-25	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii NFRI69-148	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii NFRI89-1040/1/3	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium lundbergii NFRI1502/1	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium procerum DAOM33940	2.25	95 °C, 3 m	95 °C, 2 m	52.9 °C, 1.5 m	70 °C, 4 m, final 70 °C, 10 m	25
Leptographium procerum J.R.88-409A	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium procerum NFRI59-84/2	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium procerum TOM55.35	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium procerum TOM62.30	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium procerum TOM73.12	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium procerum TOM76.36	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium procerum TOM76.8	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium procerum TOM86.19	2.25	95 °C, 3 m	95 °C, 2 m	53 °C, 1.5 m	70 °C, 5 m, final 70 °C, 10 m	35
Leptographium procerum UAMH9724	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium terebrantis CBS337.70	2.25	95 °C, 3 m	95 °C, 2 m	53 °C, 1.5 m	70 °C, 5 m, final 70 °C, 10 m	35
Leptographium terebrantis UAMH9690	2.25	95 °C, 2 m	95 °C, 1 m	51-57 °C, 1.5 m	70 °C 5 m, final 70 °C, 10 m	35
Leptographium terebrantis UAMH9722	2.25	95 °C, 2 m	95 °C, 1 m	51-57 °C, 1.5 m	70 °C 5 m, final 70 °C, 10 m	35
Leptographium truncatum CBS647.89	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium truncatum CBS929.85	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium truncatum DAOM60396	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium truncatum Forintek C34	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25

Leptographium truncatum J.R.88-324	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium truncatum J.R.88-449	2.25	93 °C, 3 m	93 °C, 1 m	50-55 °C, 1 m	70 °C, 3 m	25
Leptographium truncatum NFRI59-7/3	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium truncatum NFRI1813/1	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium truncatum TOM74.29	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium truncatum TOM86.30	2.25	95 °C, 2 m	95 °C, 1 m	52 °C, 1 m	70 °C, 4 m	30
Leptographium wingfieldii CBS345.90	2.25	95 °C, 2 m	95 °C, 1 m	52 °C, 1.5 m	70 °C, 5 m, final 70 °C, 5 m	35
Leptographium wingfieldii CBS645.89	2.25	95 °C, 2 m	95 °C, 1 m	52 °C, 1.5 m	70 °C, 5 m, final 70 °C, 5 m	35
Leptographium wingfieldii CBS648.89	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium wingfieldii MCC125	2.25	95 °C, 2 m	95 °C, 1 m	52 °C, 1.5 m	70 °C, 5 m, final 70 °C, 5 m	35
Leptographium wingfieldii MCC130	2.25	95 °C, 2 m	95 °C, 1 m	52 °C, 1.5 m	70 °C, 5 m, final 70 °C, 5 m	35
Leptographium wingfieldii MCC349	2.25	95 °C, 2 m	95 °C, 1 m	52 °C, 1 m	70 °C, 5 m	35
Leptographium wingfieldii TOM1.3	2.25	95 °C, 2 m	95 °C, 1 m	52 °C, 1.5 m	70 °C, 5 m, final 70 °C, 5 m	35
Leptographium wingfieldii TOM5.1	2.25	95 °C, 3 m	95 °C, 2 m	53 °C, 1.5 m	70 °C, 5 m, final 70 °C, 10 m	35
Leptographium wingfieldii TOM9.4	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium wingfieldii TOM10.2	1.5	94 °C, 2 m	93 °C, 1 m	54 °C, 1 m	70 °C, 3 m	25
Leptographium wingfieldii TOM11.5	2.25	95 °C, 3 m	95 °C, 2 m	53 °C, 1.5 m	70 °C, 5 m, final 70 °C, 10 m	35
Leptographium wingfieldii TOM59.21	2.25	95 °C, 2 m	95 °C, 1 m	52 °C, 1.5 m	70 °C, 5 m, final 70 °C, 5 m	35

2.4. Cloning of PCR products, purification of plasmid DNA, and DNA sequencing

Five µl (approximately 200 ng) of the purified amplicon were mixed with 1 µl (10 ng) of the pCR®4-TOPO® vector, and the entire reaction mix was transformed into 50 µl of One Shot® MAX Efficiency® DH5α-T1^R Chemically Competent *E. coli* (transformation efficiency approximately 1 x 10^9 cfu/µg DNA) according to the recommended procedure for the TOPO-TA Cloning® Kit for Sequencing (Life Technologies). The culture was inoculated onto Luria Bertani (LB) agar (in a petri dish that was 100 mm by 15 mm in size) to which 4.5 mg of ampicilin (75 µl of a 60 mg/ml stock solution) was spread over the surface of the agar just prior to inoculation. Putative transformants were identified using blue-white selection and were screened by PCR. The composition of the PCR reaction mix and the amplification conditions were identical to those used to originally amplify the gene region, except that the number of cycles was set to 30. Plasmid DNA was harvested and purified using the Wizard® *Plus* Minipreps DNA purification system according to the manufacturer's protocol (Promega Corp.) and eluted in a final volume of 35 µl of nuclease-free H_2O (Promega Corp.) at a final concentration of approximately 150 to 350 ng/µl for use in DNA sequencing.

The cycle-sequencing of PCR products and plasmid DNA was carried according to the Big Dye (version 3.1) protocols supplied by the manufacturer (Applied Biosystems, Foster City, USA). Nested primers that were designed specifically for sequencing mt genes are described in Supplementary Table 7.1. For sequencing reactions carried out at the University of Manitoba, the reaction mix contained: 0.75 µl (100 to 250 ng) of plasmid DNA, which had been denatured at 99 °C for 10 minutes and then snap-cooled on ice; 0.75 µl (3 pmoles) of primer; 1.5 µl of

2.5X Big Dye terminator (version 3.1) ready reaction premix; 1.5 µl of 5X Big Dye sequencing buffer; and 10.5 µl of H_2O. The resulting sequencing products were purified using the ethanol/EDTA precipitation method as described by the manufacturer (for sequencing reactions purified at the University of Manitoba) and resolved on an automated fluorescent DNA sequence analysis system using an Applied Biosystems 3730XL 96 capillary sequencer (Applied Biosystems) for sequencing carried out at the University Core DNA Services (University of Calgary, Calgary, Canada) or using an Applied Biosystems 3130 4 capillary sequencer (Applied Biosystems) for sequencing carried out at the University of Manitoba (Winnipeg, Canada). Chromatograms were visualized using the program BioEdit version 7.0.9.0 (Hall, 1999).

2.5. Analysis of sequence data

2.5.1. ITS1-5.8S-ITS2 nuclear rDNA region in *Grosmannia*, *Leptographium*, and related taxa

Sequence data were obtained from GenBank and from strains housed at the WIN(M) herbarium (University of Manitoba). Identical sequences were identified using DAMBE (Xia, 2000) and discarded for this study, giving data sets of 70 (entire ITS-5.8S region), 55 (ITS1 segment), and 46 (ITS2 segment) unique sequences.

ITS sequences were aligned manually using GeneDoc V2.7.000 (Nicholas et al., 1997), and the alignment is provided in the Appendix as Supplementary Figure 7.1. Programs contained within PAUP* version 4.0b10 (Swofford, 2002), PHYLIP Version 3.68 (Felsenstein, 2008), MrBayes v3.1.2 (Ronquist and Huelsenbeck, 2003), Tree-Puzzle version 5.2 (Schmidt et al., 2002), and the Willi Hennig Society edition

of TNT (Goloboff et al., 2008) were utilized for phylogenetic analyses. *Ceratocystis deltoideospora* strain WIN(M)41 was selected as the outgroup for all phylogenetic analyses, and *Ceratocystiopsis collifera* strain CBS126.89 was included as a second outgroup (Zipfel et al., 2006). Phylogenetic estimates using parsimony were evaluated with the bootstrap procedure (SEQBOOT: 1000 replicates, jumble 3 times) and CONSENSE in PHYLIP. For Bayesian analysis, Modeltest 3.7 (Posada and Crandall, 1998) was used to select models of evolution. Based on the Akaike Information Criterion (AIC), the TVM+I+G model was selected for analysis of the ITS-5.8S region and ITS1, while ITS2 was analyzed with the TVM+G model. In all cases, the parameters were estimated by MrBayes. The analyses were all run for 15 million generations and the sampling frequency was set to 1000. To generate 50 % majority rule consensus trees with posterior probability values, 50 % of the trees were discarded. Analysis using maximum likelihood, as implemented by Tree-Puzzle (maximum likelihood phylogenetic analysis using quartets and parallel computing), employed the following settings for the quartet puzzling algorithms: 25 000 puzzling steps; transition/transversion parameter estimated from the data sets; and HKY evolutionary model (Hasegawa et al., 1985). The topology of the tree obtained from the TNT program confirmed the topology of the trees obtained using PHYLIP and the analysis was not included in this study. The phylogenetic trees presented were drawn with the Tree View program version 1.6.6 (Page, 1996), using the Bayesian consensus outfile and annotations were added to the figure using Corel Draw version 14.0.0.701 (Corel, Ottawa, Canada).

2.5.2. ITS1-5.8S-ITS2 nuclear rDNA region in *Ophiostoma*, *Pesotum*, and related taxa

The strains used in this study are listed in Supplementary Table 7.2, and sequences were obtained from strains maintained in the WIN(M) collection and from GenBank. ITS sequences were aligned with ClustalX (Thompson et al., 1997) and alignments were manually edited using GeneDoc (Nicholas et al., 1997). The alignment is shown in Supplementary Figure 7.2. Programs contained within PHYLIP (Felsenstein, 2002) and Tree-Puzzle (Schmidt et al., 2002) were applied to resolve phylogenetic relationships amongst the tested sequences. The aligned ITS data set, comprising 56 sequences, was analyzed with DNAPARS (maximum parsimony) and DNADIST (F84 setting). From the latter, the distance matrix generated for each set was utilized in the NEIGHBOR program (NJ setting) for inferring a phylogenetic tree. Phylogenetic estimates were evaluated using the bootstrap procedure (SEQBOOT: NJ, 1000 replicates; parsimony, 500 replicates; and CONSENSE) in PHYLIP. Analysis with the Tree-Puzzle program used settings for the quartet puzzling algorithms as follows: 25 000 puzzling steps; transition/transversion parameter estimated from the data sets; and HKY evolutionary model (Hasegawa et al., 1985). The phylogenetic tree was drawn with the Tree View program (Page, 1996) using the PHYLIP/Tree Puzzle outfiles and annotations were added to the figure using Corel Draw. The phylogenetic tree is shown in Supplementary Figure 7.3.

2.5.3. ITS1-5.8S-ITS2 nuclear rDNA region in *Graphium* species

The strains used in this study are listed in Supplementary Table 7.3, and sequences were obtained from strains maintained in the WIN(M) collection and from GenBank. Identical sequences were identified using DAMBE (Xia, 2000) and

discarded for this study, giving a data set composed of 23 unique sequences. ITS sequences were aligned manually using GeneDoc V2.7.000 (Nicholas et al., 1997), and the sequence alignment is provided in Supplementary Figure 7.4. Programs contained within PHYLIP Version 3.68 (Felsenstein, 2008), MrBayes v3.1.2 (Ronquist and Huelsenbeck, 2003), and Tree-Puzzle version 5.2 (Schmidt et al., 2002) were utilized for inferring phylogenetic relationships. *Ceratocystis coerulescens* strain CL13-12 was selected as the outgroup for all phylogenetic analyses. For Neighbor Joining (NJ) and maximum parsimony, as implemented by PHYLIP, phylogenetic estimates were evaluated using the bootstrap procedure (SEQBOOT: NJ, 1000 replicates; parsimony, 1000 replicates and jumble 3 times) and CONSENSE. For Bayesian analysis lset nst was set to six and the rate was set to gamma. The analysis was run for 10 million generations and the sampling frequency was set to 1000. To generate 50 % majority rule consensus trees with posterior probability values, 50 % of the trees were discarded. Analysis with the Tree-Puzzle program (maximum likelihood phylogenetic analysis using quartets and parallel computing) used the following settings for the quartet puzzling algorithms: 1000 puzzling steps; transition/transversion parameter estimated from the data sets (1.71, with a standard error of 0.22); and HKY evolutionary model (Hasegawa et al., 1985). The phylogenetic trees presented were drawn with the Tree View program version 1.6.6 (Page, 1996), using the Bayesian consensus outfile and annotations were added to the figure using Corel Draw version 14.0.0.70. The phylogenetic tree is shown in Supplementary Figure 7.5.

2.5.4. The mitochondrial *rns* gene: exon, intron, and ORF sequences

Sequence data were obtained from strains housed at the WIN(M) herbarium (University of Manitoba) and additional sequences were obtained from GenBank (National Center for Biotechnology Information) using those sequence data as queries in blastn searches, employing the database corresponding to "Others (nr etc.)" for the mt *rns* and group II intron data sets and in blastp searches for the amino acid data set. A description of the internet-based bioinformatics programs used in this work is provided in Table 2.5. Identical sequences were identified using DAMBE (Xia, 2000) and discarded for this study, leaving data sets of 65 sequences (mt *rns* exon), 18 (mt *rns* group II introns), and 39 (LHEases).

DNA sequences corresponding to the mt *rns* exon (from which intronic sequences were removed) were first aligned using ClustalX 2.0.10 (Larkin et al., 2007) and the alignment was refined manually using GeneDoc V2.7.000 (Nicholas et al., 1997). Regions of the mt *rns* sequence in which the alignment was ambiguous were removed; the alignment used for phylogenetic analyses is provided in Supplementary Figure 7.6. Programs contained within PHYLIP Version 3.68 (Felsenstein, 2008), MrBayes v3.1.2 (Ronquist and Huelsenbeck, 2003), and Tree-Puzzle version 5.2 (Schmidt et al., 2002) were utilized for phylogenetic analyses. The mt *rns* gene sequence of *Kluyveromyces thermotolerans* was selected as the outgroup, and the data set was analyzed with DNAPARS (maximum parsimony) and DNADIST (F84 setting). From the latter, the distance matrix generated for each set was utilized in the NEIGHBOR program (NJ setting) for inferring a phylogenetic tree. Phylogenetic estimates were evaluated using the bootstrap procedure (SEQBOOT: NJ, 1000 replicates; parsimony, 1000 replicates and jumble 1 time) and CONSENSE in PHYLIP. Analysis with the Tree-Puzzle program used the following settings for the

quartet puzzling algorithms: 25 000 puzzling steps; transition/transversion parameter estimated from the data sets; and HKY evolutionary model (Hasegawa et al., 1985). For Bayesian analysis, the data set comprised 65 taxa and 1242 characters, lset nst was set to six, and the rate was set to gamma. The analyses were run for 10 million generations and the sampling frequency was set to 1000. To generate 50 % majority rule consensus trees with posterior probability values, 50 % of the trees were discarded. The phylogenetic trees presented were drawn with the Tree View program version 1.6.6 (Page, 1996), using the Bayesian consensus outfile and annotations were added to the figure using Corel Draw version 14.0.0.701.

DNA sequences corresponding to the group II intron (from which sequences from the putative ORF start codon to the stop codon were removed) were first aligned using ClustalX 2.0.10 (Larkin et al., 2007); the alignment was refined manually in GeneDoc V2.7.000 (Nicholas et al., 1997) using conserved helices and loops in RNA secondary structure models (Toor and Zimmerly, 2002; Mullineux et al., 2010) as a guide. Regions in DIII to DIV in which the alignment was ambiguous were removed; the alignment is provided in Supplementary Figure 7.7. For phylogenetic analyses of the ORF-less group II intron, intron 1 from *C. parasitica* (C.p.SSUi1) was selected as the outgroup. Phylogenetic estimates were evaluated using the bootstrap procedure (SEQBOOT: NJ, 1000 replicates; parsimony, 1000 replicates and jumble 3 times) and CONSENSE in PHYLIP. Analysis with the Tree-Puzzle program used the following settings for the quartet puzzling algorithms: 10 000 puzzling steps; transition/transversion parameter estimated from the data sets; and HKY evolutionary model (Hasegawa et al., 1985). For Bayesian analysis, the data set comprised 18 taxa and 739 characters and was analyzed as for the mt *rns* exon data set. The phylogenetic trees presented were drawn with the Tree View program version 1.6.6 (Page, 1996),

using the Bayesian consensus outfile and annotations were added to the figure using Corel Draw version 14.0.0.701.

The amino acid sequences of putative LHEGs were automatically aligned with PRALINE (Heringa, 1999, 2000, 2002; Simossis and Heringa, 2003, 2005) using the default parameters: exchange weights matrix, BLOSUM62; open gap penalty, 12; extension, 1; progressive alignment strategy, PSI-BLAST pre-profile processing (homology-extended alignment); iterations, 3; e-value cut-off, 0.01; DSSP-defined secondary structure search; and secondary structure prediction, PSIPRED. In all cases the alignments were refined manually using GeneDoc V2.7.000 (Nicholas et al., 1997). Ambiguous regions were removed and the alignment of the LHEase amino acid sequences is provided in Supplementary Figure 7.8 in the Appendix. For analysis of the amino acid sequence of the LHEases, the LHEase from the fifth intron of the *cox1* gene from *Podospora anserina* (cox1i5) was selected as the outgroup. For maximum parsimony, phylogenetic estimates were evaluated as for the mt *rns* exon data set. Analysis with the Tree-Puzzle program used the following settings for the quartet puzzling algorithms: 10 000 puzzling steps; uniform rate of heterogeneity; and the Mueller-Vingron Model (Müller and Vingron, 2000). For Bayesian analysis the data set comprised 39 taxa and 356 characters. The parameters were estimated by MrBayes: the amino acid model was Poisson and a gamma rate was used. The analyses were run for 5 million generations and the sampling frequency was set to 1000. To generate 50 % majority rule consensus trees with posterior probability values, 50 % of the trees were discarded. The phylogenetic trees presented were drawn with the Tree View program version 1.6.6 (Page, 1996), using the Bayesian consensus outfile and annotations were added to the figure using Corel Draw version 14.0.0.701 (Corel Corportation, Ottawa, Canada).

Table 2.5. Description of bioinformatics programs used in sequence analyses.

Program	URL	Features
blastn (Basic Local Alignment Search Tool, NCBI server)	http://blast.ncbi.nlm.nih.gov/	Search nt databases using a DNA sequence query. Database used is "Others (nr etc.)".
blastp (Basic Local Alignment Search Tool, NCBI server)	http://blast.ncbi.nlm.nih.gov/	Search protein databases using an amino acid sequence query. Database used is the "non-redundant protein sequences (nr)".
ORF Finder (NCBI server)	http://www.ncbi.nlm.nih.gov/projects/gorf/	Identify putative ORF sequences and conserved amino acid domains. The genetic code used is "4 Mold, protozoan, and coelenterate mitochondria and the mycoplasma code".
Mfold	http://mfold.bioinfo.rpi.edu/cgi-bin/rna-form1.cgi	The webserver predicts folding of nucleic acids. The user can input structural constraints that either force or prohibit base pairing.
DNA weblogos	http://weblogo.berkeley.edu/logo.cgi	Create sequence logos and rapidly identify regions of conserved sequence and the relative frequency of nucleotides or amino acid residue at each position
RNA structure logos	http://www.cbs.dtu.dk/~gorodkin/appl/slogo.html	A structural logo is created based on the Vienna notation (in which brackets indicate base pairing and dots represent unpaired nucleotides) for a representative sequence and the nt sequence of the remaining samples are aligned with respect to the reference. Logos provide "mutual information" (Gorodkin et al., 1997); that is, sites at which nucleotide substitutions occur that maintain base pairing.

2.6. Sequence and structural analyses using logos

2.6.1. ITS1-5.8S-ITS2 nuclear rDNA region

The Weblogo program (Crooks et al., 2004) was used to generate sequence logos (Schneider and Stephens, 1990) for assessing both sequence conservation within ITS1 and ITS2 and the relative frequency of the nucleotides at each position. To generate the sequence logos, the sequences for the phylogenetic outgoups were removed from the alignment.

Models of RNA secondary structure of ITS1 and ITS2 were developed using Mfold (Matthews et al., 1999; Zuker, 2003) and the number of constraints (either forcing or prohibiting base pairing interactions) was kept to a minimum. For ITS1, motifs common to the majority of strains, such as the loops/helices near the 5' and 3' termini and features within the major helix, were identified. Structural models for the remaining strains were determined based on comparative sequence analysis using these features as a guide. Proposed models for ITS2 were compared to those available in the ITS2 Database II (Selig et al., 2008) and were found to conform to those in the database. Structures were modeled using conserved sequence/structural features in helices I to IV as a template.

To generate RNA structure logos (Gorodkin et al., 1997) for each ITS segment, the two phylogenetic outgroups were removed from the sequence alignment. *Pesotum* sp. strain WIN(M)481, a strain that is distantly-related to the remaining strains in the study, was used as the reference in this alignment. RNA structural models of each ITS segment for this strain were converted to Vienna notation (in which brackets indicate base pairing and dots represent unpaired nucleotides) in Mfold (Matthews et al., 1999; Zuker, 2003; Meyer and Miklós, 2007) and the text file

containing the Vienna notation of *Pesotum* sp. strain WIN(M)481 and the DNA sequence of the remaining strains was uploaded to the RNA structure logo site (http://www.cbs.dtu.dk/~gorodkin/appl/slogo.html). Logos were generated using the default parameters: type 2 logo, *a priori* nucleotide distribution probability of 0.25 for each of the 4 ribonucleotides and a base pair weight probability of 1.0 for AU, CG, and GU base pairs. The logos were examined for "mutual information" (Gorodkin et al., 1997); that is, sites where there is base pairing in the absence of nucleotide conservation.

2.6.2. The mitochondrial *rns* gene and putative LAGLIDADG homing endonucleases

The Weblogo program (Crooks et al., 2004) was used to generate sequence logos (Schneider and Stephens, 1990) for assessing both sequence conservation in the mt *rns* exon sequences flanking the intron insertion site (IS) and the relative frequency of the nucleotides at each position. The logo comprised 147 nt. The program was also used to generate an amino acid sequence logo for the 13 LHEases encoded by mS952-type group II introns (that is, introns related to intron 3 of the mt *rns* gene of *C. parasitica*) in the mt *rns* gene.

2.7. Design of precursor constructs, synthesis and purification of intron-containing transcripts, and self-splicing assays

In collaboration with the laboratory of Dr. François Michel (Centre de Génétique Moléculaire, Centre National de la Recherche Scientifique, Gif-sur-Yvette, France), the LtrΔORF.3 precursor was generated in three PCR reactions from the mtsr1435-3 clone (that is, the clone of the intron/LHEG-plus allele of the mt *rns* gene

in *L. truncatum* strain CBS929.85) by truncating the 5' exon at position 726 (based on the numbering of the intron-plus allele of clone mtsr1435-3) and deleting the ORF sequence. The region between the truncated 5' exon and upstream of the ORF in DIII was amplified using primer pairs Ltr1 and Ltr2, while the segment between the 3' exon and sequences downstream of the ORF in DIII was amplified using primer pairs Ltr3 and mtsr2. The reaction mix contained: 10 µl of 5X Phusion® HF buffer; 0.4 µl of 2.5 mM dNTPs (20 µM, final concentration, of each dNTP); 5 µl (50 pmoles) of each primer; 1 µl (195 ng) of DNA template; 0.5 µl (1 unit) of Phusion® High-Fidelity DNA polymerase (New England Biolabs Inc., Saint-Quentin-en-Yvelines, France); and 28.1 µl of H_2O. The amplification conditions were: initial denaturation (92 °C, 3 min); followed by 15 cycles of denaturation (92 °C, 15 s), annealing (47 °C, 45 s), extension (72 °C, 3 min); ramp of 8.0; and 25 cycles of denaturation (92 °C, 15 s), annealing (56 °C, 45 s), extension (72 °C, 3 min); followed by a final extension (72 °C, 5 min). The two segments were joined using self-priming PCR with the Ltr1 and mtsr2 primers. The amplification conditions were: initial denaturation (92 °C, 3 min); followed by 10 cycles of denaturation (92 °C, 15 s), annealing (42 °C, 45 s), extension (72 °C, 3 min); ramp of 8.0; and 5 cycles of denaturation (92 °C, 15 s); annealing (56 °C, 45 s), extension (72 °C, 3 min); followed by a final extension (72 °C, 5 min). The amplicon was purified with an illustra Sephadex™ G-50 DNA grade column (GE Healthcare Europe GmbH, Munich, Germany), visualized on an ethidium bromide-stained agarose gel, and gel extracted using the QIAquick Gel extraction kit (Qiagen, Courtaboeuf, France). The purified amplicon, 1342 bp in size, was digested with PstI and BmgBI and then ligated with the two fragments of plasmid pTZ19U, which had been triple digested with Ecl136III, AlwNI, and PstI. The ligated product was

amplified using the Ant7 primer and 24-mer primer. The final construct consisted of 98 nt of the 5' exon, 819 nt of the ORF-less intron, and 262 nt of the 3' exon.

For the transcription of intron-containing precursors, plasmid DNA was linearized completely in a reaction mix containing: 20 µl of 10X NEBuffer 3 [100 mM NaCl, 50 mM Tris-Cl, 10 mM $MgCl_2$, 1 mM dithiothreitol (DTT), final concentration, New England Biolabs]; 2 µl of 100X NEB bovine serum albumin (New England Biolabs); 25 µl (8 µg) of plasmid DNA; 5 µl (100 units) of EcoRI (New England Biolabs); and 148 µl of H_2O, and the digest was incubated at 37 °C for 3.5 h. Linearized DNA was purified by extraction with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1). The mixture was centrifuged at maximum speed for 5 minutes at room temperature. The aqueous phase was transferred to a new microfuge tube and the extraction steps were repeated. One volume of pure chloroform was added to the aqueous phase and the sample was centrifuged as above; this step was repeated twice. DNA was precipitated with 1/10 volume of 3 M sodium acetate and 3 volumes of ice-cold absolute ethanol and incubated at -20 °C for at least 30 minutes. DNA was pelleted by centrifugation at the maximum speed for 20 minutes at 4 °C and to the pellet 400 µl of ice-cold 75 % ethanol was added; the sample was centrifuged as in the previous step and the pellet was resuspended in 15 µl of 1X TE buffer.

The LtrΔORF.3 precursor was transcribed in a reaction mix containing: 3 µl of 10X transcription buffer [40 mM Tris-Cl (pH 7.5), 12 mM $MgCl_2$, 10 mM DTT, 2 mM spermidine, final concentrations]; 6 µl of rNTPs (25 mM rATP, rGTP, rCTP, and 12.5 mM UTP, final concentrations); 0.6 µl of DTT (10 mM, final concentration); 0.6 µl of $MgCl_2$ (20 mM, final concentration); 8 µl (80 µCi) of α-^{32}P-UTP; 12 µl (116 nM) of PCR template; and 0.6 µl of T7 RNA polymerase. The transcription mix was

incubated for 2.5 h at 37 °C. Three µl (3 units) of RQ1 RNase-free DNase (Promega Corp.) was then added to the reaction and incubated for 30 minutes at 37 °C, and the reaction was stopped with 3 µl (417 mM, final concentration) of 500 mM $Na_2EDTA{\cdot}2H_2O$ (pH 8.0). 19U/Ltr.2, derived from LtrΔORF.3 comprised the full-length, ORF-containing, intron sequence (1840 nt), 98 nt of the 5' exon (exon 1), and 262 nt of the 3' exon (exon 2). Transcription of 0.5 µg/µl of plasmid DNA (3.75 µg), digested with EcoRI, was carried out in a volume of 50 µl under the same conditions as for LtrΔORF.3, except that the DNase digestion step was omitted.

Transcripts were purified as previously described (Michel et al., 1992). Briefly, the transcript was precipitated with the addition of 1/3 the reaction volume of 6 M ammonium acetate and 3 volumes of absolute ethanol. The sample was incubated at -70 °C for 30 minutes and then centrifuged at 10 000 rpm at 10 °C. The sample was loaded onto a preparative gel (4 % acrylamide 19:1 and 7-8 M urea). The labelled transcript was gel excised and frozen at -70 °C overnight and extracted with n-butanol (neat). Self-splicing assays were carried out as described by Costa et al. (1997a). Essentially, 2 µl of transcript was added to an equal volume of splicing buffer [0.1 or 1 M NH_4Cl or KCl; 10 or 50 mM $MgCl_2$; 40 mM Tris-Cl (pH 7.5); 0.02 % (w/v) SDS, final concentrations] and the reaction mix was overlaid with mineral oil to prevent evaporation. The reaction was stopped by the addition of 4.5 µl of formamide blue solution [deionized formamide, 80 mM $Na_2EDTA{\cdot}2H_2O$ (pH 8), 0.025 % (w/v) bromophenol blue] and placed on dry ice. Samples were incubated at -20 °C.

2.8. Expression and purification of I-LtrII

2.8.1. N-terminal His_6-tagged I-LtrII

Because of differences in codon use and codon bias in fungal mt protein-coding sequences, the ORF sequence was optimized for expression in *E. coli* (Supplementary Figure 7.9) and synthesized commercially (Bio S & T, Montréal, Canada). The optimized ORF sequence (mtsr1435HEG), 915 nt, was cloned into a pBluescript II SK+ vector by the supplier (Bio S & T). The mtsr1435HEG construct was PCR-amplified using primers 1435HEGF and 1435HEGR in a reaction mix containing: 5 µl of 10X High Fidelity PCR buffer [60 mM Tris-SO_4 (pH 8.9), 18 mM ammonium sulphate, final concentrations, Life Technologies]; 2 µl of $MgSO_4$ (2 mM, final concentration); 1 µl of dNTP (200 µM of each dNTP, final concentration); 1 µl (10 pmoles) of each primer; 1 µl of plasmid DNA, diluted either 10^{-2} or 10^{-3} (approximately 0.25 to 3.5 ng); 0.2 µl (1 unit) of Platinum® *Taq* High Fidelity polymerase (Life Technologies); and 38.8 µl of H_2O. Amplification was carried out using a temperature gradient thermocycler (Tgradient thermocycler, Biometra GmbH, Goettingen, Germany) under the following conditions: initial denaturation (94 ºC, 2 min); followed by 25 cycles of denaturation (94 ºC, 30 s), primer annealing (57.5 ± 2.5 ºC, 30 sec), extension (68 ºC, 1 min 30 s); and a final extension (70 °C, 7 min). The mtsr1435HEG amplicon was cloned into the pET200/D-TOPO® vector (15 to 20 ng) supplied in the Champion™ pET Directional TOPO® expression kit (Life Technologies), following the manufacturer's recommended protocol. The orientation and appropriate reading frame were confirmed by sequencing.

Protein expression and purification were carried out in collaboration with the laboratory of Dr. François Michel (Centre de Génétique Moléculaire). For expression,

the pET200/Dmtsr1435HEG construct was transformed into *E. coli* strain RosettaBlue™ (DE3) pLysS (Novagen®, Merch KGaA, Darmstadt, Germany) and 2 ml of an overnight culture were inoculated into 400 ml of 2x YT medium, containing (g/L): tryptone (16); yeast extract (10); and NaCl (5), supplemented with 50 µg/ml of kanamycin. Protein expression was induced at an OD_{600} of 0.8 to 1.0 with 100 µM of isopropyl β-D-1-thiogalactopyranoside (IPTG) at 25 °C for 3 to 4 h.

The protein was harvested by resuspending the cell pellet in 15 ml of cell lysis (CL) buffer [40 mM HEPES (pH 7.5), 800 mM NaCl, 10 % (w/v) glycerol, 6mM β-mercaptoethanol, and one-half tablet of the complete EDTA-free protease inhibitor cocktail (Roche Diagnostics)]. The N-terminal His_6-tagged protein was purified using Ni-NTA resin (Qiagen). Cells were lysed using the cell press method. The cell lysate was added to 2 ml of Ni-NTA resin (Qiagen) and 20 ml of CL buffer was added to the pelleted resin-bound protein. The following series of washing steps were carried out: wash 1, 40 ml of CL buffer with 2 mM of imidazole; wash 2 (3 times), 40 ml of CL buffer with 9 mM of imidazole; and wash 3, 20 ml of CL buffer with 15 mM of imidazole. The protein was eluted in buffers containing 40 mM HEPES, 800 mM NaCl, 20 % (w/v) glycerol, and either 120 mM or 200 mM imidazole (pH 8). The sample was collected in 3 fractions (0.75 to 1 ml) with 120 mM imidazole and then in a final fraction with 200 mM imidazole (1 ml). Excess imidazole was removed by dialysis in buffer A [40 mM HEPES (pH 7.5), 200 mM NaCl, and 3 mM β-mercaptoethanol] using a slide-a-lyzer dialysis cassette with a 7 kDa molecular weight cut-off (Millipore, Billerica, USA) according to the manufacturer's suggested protocol.

A second purification step was carried out using a HiTrap™ heparin HP column (GE Healthcare Europe). The sample was washed with one column volume of

wash buffer over a range of 0.2 to 1.5 M NaCl. In each wash fraction the NaCl concentration was increased by 0.1 M; NaCl was adjusted by mixing the appropriate volumes of buffers A and B [40 mM HEPES (pH 7.5), 1.5 M NaCl, and 3 mM β-mercaptoethanol]. The fractions recovered with 600 and 700 mM NaCl were combined and the sample was then concentrated using a Centricon Ultracel YM-30 centrifugation device (Millipore), following the manufacturer's guidelines. The sample was diluted to a final volume of 9 ml in a storage buffer [40 mM HEPES (pH 7.5), 400 mM NaCl, 0.5 mM DTT, and 10 % (w/v) glycerol], and centrifuged at 4000 xg at 4 °C until the sample was concentrated in a final volume of 500 µl.

2.8.2. Near-native I-LtrII

The mtsr1435HEG construct was subcloned into the pTYB4 vector (New England Biolabs) for purification using the Impact (Intein Mediated Purification with an Affinity Chitin-binding Tag) system (New England Biolabs, Pickering, Canada). For expression, pTYB4-L.trORF was transformed into 50 µl of BL21 Star™ (DE3) One Shot® Chemically competent *E. coli* (Life Technologies, transformation efficiency greater than 1 x 10^8 cfu/µg plasmid DNA). 400 ml of LB supplemented with 100 µg/ml of ampicillin were inoculated with 1.5 ml of the overnight culture, expression was induced at an OD_{600} of 0.8 to 1.0 with 250 µM of IPTG, and the culture was incubated at 30 °C for 3 h.

The protein was harvested and purified according to the manufacturer's recommended protocol (New England Biolabs). Pellets were resuspended and lysed in buffer CBD1 [30 mM HEPES (pH 7.5), 1 M NaCl, 0.4 mM $Na_2EDTA{\cdot}2H_2O$ (pH 8.0)] to which one-half tablet of the complete EDTA-free protease inhibitor cocktail (Roche Diagnostics, Laval, Canada) had been added. The sample was washed with

100 ml of CBD2 [30 mM HEPES (pH 7.5) and 700 mM NaCl], and cleaved in CBD2 buffer that was supplemented with 2 % (w/v) glycerol and 35 mM DTT; the cleavage reaction was carried out overnight (16 to 20 h) at 4 °C. The protein was eluted in 3 ml of CBD2 supplemented in 4 % glycerol and was de-salted in an exchange buffer [40 mM HEPES (pH 7.5), 300 mM NaCl, 0.5 mM DTT, and 10 % (w/v) glycerol] using an Amicon Ultra 3K NMWL centriprep (Millipore). The I-LtrII protein expressed and purified with this system is near-native, containing an additional glycine residue at its C-terminus. The protein sample was quantified by measuring the absorbances at 260 nm and 280 nm and the concentration was determined using the formula: protein concentration (mg/ml) = $1.55*Abs_{280} – 0.76*Abs_{260}$.

2.9. Maturase and filter-binding assays

In collaboration with the laboratory of Dr. François Michel (Centre de Génétique Moléculaire), maturase activity of N-terminal His_6-tagged I-LtrII was assayed using 20 nM of internally labelled LtrΔORF.3 RNA incubated at 37 °C with increasing concentrations (up to 80 nM) of protein in splicing buffer C1 [40 mM Na_2HEPES (pH 7.6), 150 mM KCl, 6 mM $MgCl_2$] or C2 [40 mM Na_2HEPES (pH 7.6), 150 mM KCl, 12 mM $MgCl_2$] over a period of 60 minutes. Maturase activity of near-native I-LtrII was also tested with the full-length precursor, 19U/Ltr.2. 100 nM of internally-labelled RNA was incubated with 19.5 nM of I-LtrII at 37 °C in splicing buffer [40 mM Tris-Cl (pH 7.6 at 37 °C), 100 mM NH_4Cl, 6 mM $MgCl_2$] over a period of 60 minutes. A molecular weight standard was generated using the products of the splicing reaction catalyzed by Pl.LSU/2 (Costa et al., 1997a) at 45 °C in splicing buffer [40 mM Tris-Cl (pH 7.6 at 37 °C), 1 M KCl, 50 mM $MgCl_2$]. Filter-binding assays were carried out with 5 nM of internally-labelled LtrΔORF.3 RNA

incubated with increasing concentrations of N-terminal His_6-tagged I-LtrII in buffers C1 and C2 at 37 °C based on the procedure of Bassi and co-workers (2002).

2.10. Endonuclease assays

The substrates used for the cleavage assays were plasmid DNA harvested and purified using the Wizard® *Plus* Minipreps DNA purification system (Promega Corp.). The samples were quantitated by measuring the absorbance spectrum over 220 to 300 nm and using the absorbance maximum in the following formula: concentration (ng/µl) = Abs_{max} *50 µg/ml. The potential cleavage substrates were proposed to be either the exon sequences flanking the intron insertion site or the intron sequences in DIII flanking the ORF. Construct pCR4mtrnsEx was used as the intron-minus allele (generated from clone mtsr1435-16, a naturally-occurring intron-minus allele of the mt *rns* gene of *L. truncatum* strain CBS929.85). The second potential substrate comprised 50 bp of the intron sequence upstream and downstream of the ORF: 5'-TTTTATATTATAAAGGATTTTTCAAGACATATAAATTAAATC AGTTAAATGCGTAAGCTTTAATATATTATATTTAAGAATAAATAATATGTT TATTAGT-3'. This sequence was synthesized commercially (Bio S & T) and cloned by the supplier into a pBluescript II SK+ vector to create the construct pSKmtrnsORFis. The control substrate consisted of the intron-plus allele (clone mtsr1435-3), cloned into the pCR®4-TOPO® vector as described above, generating the pCR4mtrnsExIn construct, where both potential cleavage substrates are interrupted by the intron and ORF sequences. To assess random cleavage of the vector itself, an unrelated gene, the *cob* gene from *L. lundbergii* strain DAOM60397, which was cloned into the pCR®4-TOPO® vector, was also included.

The cleavage reaction mix contained: 2 µl of 10X REact® Buffer 3 [50 mM Tris-Cl (pH 8.0), 10 mM $MgCl_2$, 100 mM NaCl (Life Technologies)] supplemented with 1 mM DTT; 1 µl (315 ng) of plasmid DNA; 3 µl of near-native I-LtrII (4 µM, final concentration); and 14 µl of nuclease-free H_2O (Promega Corp.). Cleavage reactions were incubated at 37 °C for the indicated length of time and were stopped by the addition of 3 µl (25 mM, final concentration) of $Na_2EDTA{\cdot}2H_2O$ (pH 8.0) and 1 µl of proteinase K (1 mg/ml), with subsequent incubation for 20 minutes at 37 °C. For the zero time sample, I-LtrII was excluded and the sample was incubated for 60 minutes. An equal volume of stop solution [200 mM $Na_2EDTA{\cdot}2H_2O$ (pH 8.0) and 0.4X TBE in Hi-Di™ formamide (Applied Biosystems)], was added to all samples but those that were stopped after 30 minutes and the samples were then stored at -20 °C. Samples were separated through an 0.8 % (w/v) agarose gel.

Cleavage activity was also tested against linearized plasmid DNA. To generate the substrates used for the cleavage assays, plasmid DNA, harvested and purified using the Wizard® *Plus* Minipreps DNA purification system (Promega), was linearized with either NcoI (constructs cloned into the pCR®4-TOPO® vector) or NotI (construct cloned into the pBluescript II SK+ vector). DNA was linearized in a 50 µl reaction volume, containing: 5 µl of 10X NEBuffer 3 (New England Biolabs); 10 µl of plasmid DNA (1.95 to 5.9 µg); 2 µl (20 units) of NcoI (New England Biolabs); and 33 µl of H_2O. The pBluescript II SK+ vector was linearized as above, except that 30 units of NotI (Life Technologies) was used in REact® Buffer 3 (Life Technologies). DNA was purified using the Wizard® SV Gel and PCR Clean-Up System (Promega Corp.) and quantified by measuring the absorbance spectrum over 220 to 330 nm and then using the absorbance maximum in the following formula: concentration (ng/µl) = Abs_{max} *50 µg/ml. The cleavage reaction mix contained: 2 µl of Invitrogen's 10X

REact® Buffer 3 supplemented with 1 mM DTT, 6 µl (210 ng) of linearized plasmid DNA; 3 µl of near-native I-LtrII (4 µM, final concentration); and 14 µl of nuclease-free H_2O (Promega). Cleavage reactions were incubated at 37 °C for the indicated length of time and were stopped by the addition of 3 µl (25 mM, final concentration) of $Na_2EDTA{\cdot}2H_2O$ (pH 8.0) and 1 µl of proteinase K (1 mg/ml), with subsequent incubation for 20 minutes at 37 °C. For the zero time sample, I-LtrII was excluded and the sample was incubated for 45 minutes and the samples were then stored at -20 °C. Samples were separated through an 0.8 % (w/v) agarose gel.

2.11. Mapping of the cleavage site

In collaboration with the laboratory of Dr. François Michel (Centre de Génétique Moléculaire) the cleavage site was mapped by amplifying a 248-bp region of the pCR4mtrnsEX construct, labelling either the top or bottom strand. The amplification mixture contained: 5 µl of 10X ThermoPol reaction buffer [20 mM Tris-Cl; 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$, 0.1 % Triton X-100 (pH8.8), final concentrations (New England Biolabs)]; 1 µl of dNTPs (200 µM, final concentration, of each dNTP); 5 µl (5 pmoles) of γ-^{32}P-labelled and 1.67 µl (5 pmoles) of unlabelled primers; 1 µl (10 ng) of plasmid DNA; 0.5 µl (2.5 units) of *Taq* DNA polymerase (New England Biolabs); and 35.83 µl of H_2O. The amplification was carried out under the following conditions: initial denaturation (95 °C, 2 min); followed by 30 cycles of denaturation (93 °C, 45 s), annealing (50 °C, 30 s), extension (70 °C, 30 s); and a final extension (70 °C, 2 min). The amplicon was purified using the GenElute™ PCR clean-up kit (Sigma-Aldrich Chimie, Lyon, France). Sequencing ladders of the top and bottom strands of the cleavage substrate were generated using the *fmol*® DNA sequencing system (Promega Corp.). 5' end-labelling of the primers

with γ-^{32}P-ATP and generation of the sequencing ladders was carried out according to the manufacturer's suggested protocol, with the following modification: prior to 5' end-labelling the primer with T4 polynucleotide kinase, the oligonucleotide primer was denatured in water at 100 °C for 2 minutes and then snap-cooled on ice. The cleaved product was generated in a reaction mix containing: 2 µl of 10X NEBuffer 3 (New England Biolabs); 6 µl (approximately 288 ng) of PCR substrate; 1 µl of I-LtrII (4 µM, final concentration); and 11 µl of H_2O, and was incubated at 37 °C for 60 minutes. The products of the sequencing reaction were resolved alongside cleaved and uncleaved products on a 5 % (w/v) polyacrylamide/8 M urea gel. The salt concentration was normalized for all samples in order to reduce aberrations in migration due to salt differences.

CHAPTER 3. THE MOLECULAR EVOLUTION OF INTERNAL TRANSCRIBED SPACER SEQUENCES IN NUCLEAR RIBOSOMAL RNA GENES

Work described in this chapter was published in the journal Fungal Genetics and Biology in 2009: Mullineux T, Hausner G. 2009. Evolution of rDNA ITS1 and ITS2 sequences and RNA secondary structures within members of the fungal genera *Grosmannia* and *Leptographium*. *Fungal Genet. Biol.* 46:855-867.

3.1. Introduction and research objectives

In the majority of eukaryotes the primary product of RNA polymerase I is a 35-45S rRNA precursor, comprising the 5' external transcribed spacer (ETS), 16-18S ribosomal (r) RNA, internal transcribed spacer (ITS) 1, 5.8S rRNA, ITS2, 23-28S rRNA, and 3' ETS segments (Good et al., 1997; Lafontaine, 2004). The ITS segments are thought to have independent origins, with ITS1 being derived from an intergenic spacer (Clark, 1987) and ITS2 from an expansion segment of the large subunit (LSU) rRNA gene (Nazar, 1980; Hershkovitz et al., 1999). It is possible that these spacers evolved from progenote rRNA linker (PRL) regions, whose biological role in early cells was to interact with and stabilize rRNA in the early, ribosome-like molecules (Clark, 1987). As the assembly and stability of these "quasi-ribosomes" improved and the fidelity of transcription increased, the selective pressure on PRLs likely decreased and the segments drifted in size and sequence. Biochemical evidence indicates that these segments contain sites recognized by protein and RNA components involved in ribosome biogenesis, suggesting that the spacers were derived from PRLs that were co-opted to interact with ribosomal proteins (Clark, 1987). In plants the spacer regions

appear to be convergent in length and substitution patterns (reviewed in Won and Renner, 2005). Such observations support the suggestion that the two spacers co-evolve, which is possibly due to the functional constraints that maintain the reciprocal biochemical interdependence of the ITS segments during rRNA processing.

It was previously shown that there is a positive correlation between the lengths of the ITS1 and ITS2 sequences amongst members of the Ascomycota, which is also suggestive of co-evolution (Hausner and Wang, 2005). The study also showed that for this broad group of fungal genera the RNA secondary structure of the ITS segments was conserved, despite variability in the DNA sequence. Due to its wide coverage of fungal taxa and, consequently, poor sequence conservation of the ITS region, the authors were unable to identify the full extent of compensatory substitutions and slippage events that may maintain the RNA secondary structure of the ITS region despite the high degree of variability at the DNA level. To examine this relationship, the present study focused on a narrow selection of phylogenetically-related fungal species and strains that share a recent common ancestor. This reduces ambiguity in the sequence alignment and permits detailed comparative analysis for assessing the mutational dynamics of the ITS region, such as potential biases regarding GC content, expansion and contraction of the segments, compensatory base changes [CBCs (G-C to A-U, and vice versa)], hemi-CBCs (G-C to G-U, and vice versa), and possible strand slippage events (Hancock and Dover, 1990).

The strains in this study belong to the anamorphic (asexual) fungal genus *Leptographium* Lagerb. & Melin and to species that have a *Leptographium* conidial state but due to the presence of a meiotic state (teleomorph) are accommodated in the genus *Grosmannia* Goid. (Jacobs et al., 2001; Zipfel et al., 2006). The genus *Grosmannia* belongs to the order *Ophiostomatales* (Ascomycota) (Hibbett et al.,

2007). Typically members of the genus *Leptographium* are phylogenetically-allied to members of *Grosmania* and like many other ophiostomatoid fungi they are associated with bark beetles that can vector these fungi to new plant hosts/sites. In many areas *Leptographium* species are becoming increasingly problematic as new species are introduced due to the arrival or introduction of bark beetle species that serve as potential vectors for the dispersal of fungal spores (Hausner et al., 2005). These fungi are of particular interest to the forestry industry, as they are blue-stain fungi that can reduce the value of stored lumber and some species are serious plant pathogens (Harrington, 1993).

In addition to their use as a marker for phylogenetic studies, ITS sequences are utilized in the identification of fungal species, such as in culture-independent, high-throughput automated approaches for the characterization of fungal communities from various habitats (Ranjard et al., 2001; Horton, 2002; Buchan et al., 2002; Kennedy and Clipson, 2003; Druzhinina et al., 2005; Mitchell and Zuccaro, 2006; Nilsson et al., 2009). In the case of pathogenic organisms sequence analysis of ITS regions may be useful in the identification of native or exotic plant pathogens (Hausner et al., 2005; Zhang et al., 2008).

The goals of the current study were: (i) to identify both conserved regions in the DNA sequences and RNA secondary structures of the ITS segments and the types of nucleotide changes occurring at the DNA level and (ii) to assess the level of phylogenetic resolution and the effects on tree topology obtained from the analysis of entire ITS1-5.8S-ITS2 region and of each segment individually. To address the first issue comparative sequence analysis was used to develop models of RNA secondary structure and to identify the potential effects of CBCs, hemi-CBCs, and insertions/deletions (indels) on secondary structure. This may assist in generating

improved structure-guided DNA sequence alignments of rapidly evolving regions, such as ITS. This approach makes use of conserved features of secondary structures to facilitate the alignment of ambiguous regions and as such may be useful for analyses based on ITS sequences, especially given the wide-spread application of this region as a marker in phylogenetic and fungal barcoding studies. To examine the second issue, phylogenetic relationships were inferred based on sequence analysis of the entire ITS-5.8S rDNA region and each ITS segment individually.

3.2 Methods overview

Sequence data were obtained from GenBank and from strains housed at the WIN(M) herbarium (University of Manitoba). As described in Chapter 2, identical sequences in the entire ITS region and in each ITS segment were identified using DAMBE (Xia, 2000) and discarded for this study. The DNA sequence alignment was analyzed using multiple phylogenetic approaches, including: parsimony; maximum likelihood with quartet puzzling; and Bayesian analyses. A DNA sequence logo of each ITS segment was also constructed, allowing one to easily identify conserved and variable regions of the ITS segments. Models of RNA secondary structure were developed using Mfold. RNA structure logos, developed using both the RNA secondary structure models and the DNA sequence alignment, were used to identify mutually informative sites.

3.3. Results

3.3.1. Characteristics of the ITS region in *Leptographium* and related taxa

The data set was composed of 70 sequences for the ITS region (Table 3.1). Amongst these 70 sequences, 55 comprised unique ITS1 sequences and only 46 consisted of unique ITS2 sequences. The lengths and GC contents of ITS1, ITS2, and the 5.8S gene for the strains used in this study are described in Table 3.2. The size of the ITS1 and ITS2 segments ranged, respectively, from 157 to 232 nt (mean, 202 nt) and 177 to 233 nt (mean, 197 nt). The ITS regions were GC-rich: mol % G+C values ranged from 54.4 to 69.6 (mean, 65.3) in the ITS1 segment and 63.5 to 77.9 (mean, 72.3) in ITS2. The GC content of the 5.8S gene was considerably lower, ranging from 48.4 to 50.6 % (mean, 49.9 %).

Table 3.1. Strains of *Grosmannia*, *Leptographium*, and related taxa used in the ITS study, along with GenBank accession numbers.

Strain[1]	Accession Number	Strain	Accession Number	Strain	Accession Number
Ceratocystiopsis collifera CBS126.89	EU913721	*Leptographium* sp. J.R.88-194A	AY935622	*Leptographium terebrantis* UAMH9722	AY935605
Ceratocystis deltoideospora WIN(M)41	EU879121	*Leptographium* sp. WIN(M)528	EU879138	*Leptographium truncatum* CBS929.85	AY935626
Grosmannia aurea CBS438.69	AY935605	*Leptographium* sp. WIN(M)984	EU879122	*Leptographium truncatum* NFRI1813/1	AY935591
Grosmannia cucullata C1216	AF198246	*Leptographium* sp. WIN(M)985	EU879123	*Leptographium truncatum* TOM74.29	AY935581
Grosmannia davidsonii WIN(M)60B	EU879127	*Leptographium* sp. WIN(M)1106	EU879147	*Leptographium truncatum* TOM86.30	AY935582
Grosmannia davidsonii WIN(M)1132	EU879129	*Leptographium* sp. WIN(M)1247	EU879146	*Leptographium serpens* WIN(M)1214	EU879144
Grosmannia davidsonii WIN(M)1494	EU879126	*Leptographium* sp. WIN(M)1269	EU879145	*Leptographium wingfieldii* CBS645.89	AY935603
Grosmannia davidsonii WIN(M)1495	EU879134	*Leptographium americanum* WIN(M)1456	EU879139	*Leptographium wingfieldii* CBS648.89	AY935611
Grosmannia dryocoetis CBS376.66	AJ538340	*Leptographium lundbergii* CBS352.29	AY935585	*Leptographium wingfieldii* MCC125	AY935608
Grosmannia europhioides CBS229.83	EU879141	*Leptographium lundbergii* DAOM64746	EU879151	*Leptographium wingfieldii* MCC130	AY935612
Grosmannia europhioides MUCL18355	AJ538333	*Leptographium lundbergii* DSMZ5010	AY935589	*Leptographium wingfieldii* MCC349	AY935610
Grosmannia europhioides NFRI80-67/22	EU879140	*Leptographium lundbergii* NFRI60-25	AY925584	*Leptographium wingfieldii* TOM10.2	AY935599
Grosmannia francke-grosmanniae ATCC22061	EU879125	*Leptographium lundbergii* NFRI69-148	AY935588	*Leptographium wingfieldii* TOM11.5	AY935602
Grosmannia galeiformis C1101	DQ062679	*Leptographium lundbergii* NFRI89-1040/1/3	AY935586	*Leptographium wingfieldii* TOM59.21	AY935600
Grosmannia galeiformis CECT20482	AJ538334	*Leptographium procerum* DAOM33940	AY935613	*Leptographium wingfieldii* WIN(M)1218	EU879152
Grosmannia huntii WIN(M)492	EU879148	*Leptographium procerum* NFRI59-84/2	AY935618	*Leptographium wingfieldii* WIN(M)1322	EU879154
Grosmannia laricis CBS636.94	AJ538332	*Leptographium procerum* TOM73.12	AY935615	*Leptographium wingfieldii* WIN(M)1382	EU879153
Grosmannia penicillata NFRI60-21	AY935623	*Leptographium procerum* TOM76.8	AY935614	*Leptographium wingfieldii* WIN(M)1482	EU879155
Grosmannia penicillata WIN(M)131	EU879137	*Leptographium procerum* WIN(M)1264	EU879143	*Ophiostoma brevicolle* CBS150.78	EU879124
Grosmannia piceaperda WIN(M)980	EU978150	*Leptographium terebrantis* CBS298.85	AY935598	*Pesotum* sp. WIN(M)478	EU879130
Grosmannia piceaperda WIN(M)1380	EU879149	*Leptographium terebrantis* CBS337.70	AY935609	*Pesotum* sp. WIN(M)481	EU879133
Grosmannia pseudoeurophioides WIN(M)42	EU879136	*Leptographium terebrantis* CBS408.61	AY935597	*Pesotum* sp. WIN(M)1423	EU879128
Grosmannia wageneri ATCC58579	AY935596	*Leptographium terebrantis* UAMH9690	AY935607	*Pesotum* sp. WIN(M)1428	EU879131
Hyalopesotum pini WIN(M)82-89	EU879132				

[1]J.R., J. Reid; CBS, Centraal Bureau voor Schimmelcultures, Utrecht, The Netherlands; WIN(M), University of Manitoba, Microbiology/Botany (J.R.'s personal collection); UAMH, University of Alberta Microfungus Collection & Herbarium, Devonian Botanic Garden, Edmonton, AB, Canada, T6G 2E1; ATCC, American Type Culture Collection, Rockville, MD, USA; NFRI, Norwegian Forest Research Institute, AS, Norway; DAOM, Cereal and Oilseeds Research, Agriculture & Agri-Food Canada, Ottawa, Ont., Canada; MCC, culture collection of H. Masuya; TOM, Isolation designation, Canadian Forest Service, Great Lakes Forestry Centre, 1219 Queen St., Sault Ste. Marie, ON, P6A 5M7.

Table 3.2. Nucleotide composition and GC content of nuclear ITS1 and ITS2 sequences in strains of *Grosmannia*, *Leptographium*, and related taxa.

Strain	ITS 1						ITS 2						5.8 S rDNA	
	Length (nt)	**% GC**	**A**	**T**	**G**	**C**	**Length (nt)**	**% GC**	**A**	**T**	**G**	**C**	**Length (nt)**	**% GC**
Outgroup[1]														
Ceratocystis deltoideospora WIN(M)41	202	68.3	29	35	57	81	190	77.9	22	20	66	82	159	49.1
Ceratocystiopsis collifera CBS126.89	160	54.4	40	33	32	55	181	71.3	23	29	54	75	159	49.1
Clade I[2]														
Hyalopesotum pini WIN(M)82-89	158	69.6	26	22	45	65	185	76.8	22	21	61	81	159	50.3
Grosmannia galeiformis C1101	158	69.6	26	22	45	65	185	76.8	22	21	62	80	159	50.3
Grosmannia galeiformis CECT20482	157	69.4	26	22	45	64	185	76.8	22	21	61	81	159	50.3
Minimum-maximum (mean)	157-158 (158)	69.4-69.6 (69.5)												
Clade II														
Grosmannia cucullata C1216	174	63.8	30	33	50	61	194	67.0	27	37	60	70	159	49.7
Grosmannia davidsonii WIN(M)60B	175	62.9	31	34	50	60	194	66.0	27	39	60	68	159	49.7
Grosmannia davidsonii WIN(M)1132	175	62.9	31	34	50	60	194	66.5	27	38	60	68	159	49.7
Grosmannia davidsonii WIN(M)1494	174	63.2	31	33	50	60	194	66.5	27	38	60	68	159	49.7
Grosmannia francke-grosmanniae ATCC22061	173	63.6	32	31	47	63	192	66.7	26	38	66	62	159	49.7
Ophiostoma brevicolle CBS150.78	171	59.1	33	37	47	54	192	63.5	30	40	58	64	159	50.3
Pesotum sp. WIN(M)478	174	62.6	31	34	50	59	194	66.5	27	38	60	69	159	49.7
Pesotum sp. WIN(M)1423	174	62.6	31	34	50	59	194	67.0	27	37	60	70	159	49.7
Pesotum sp. WIN(M)1428	174	63.2	30	34	51	59	194	66.5	27	38	60	70	159	49.7
Minimum-maximum (mean)	171-175 (174)	59.1-63.8 (62.7)					192-194 (194)	63.5-67.0 (66.2)						49.7-50.3 (49.8)
Clade III														
Grosmannia dryocoetis CBS376.66	179	65.4	30	32	48	69	185	70.3	28	27	60	70	159	48.4
Grosmannia pseudoeurophioides WIN(M)42	177	65.5	29	32	50	66	188	69.1	30	28	59	71	159	48.4
Leptographium sp. J.R.88-194A	177	63.8	31	33	49	64	201	75.1	23	27	64	87	159	50.3
Leptographium sp. WIN(M)528	177	66.1	29	31	50	67	187	65.8	34	30	68	55	159	48.4
Leptographium americanum WIN(M)1456	177	66.7	29	30	50	68	188	66.5	33	30	56	69	159	48.4
Grosmannia penicillata NFRI60-21	178	63.5	31	33	48	65	187	68.4	31	28	58	70	159	48.4
Grosmannia penicillata WIN(M)131	177	63.8	31	33	49	64	187	67.4	31	30	57	69	159	48.4
Minimum-maximum (mean)	177-179 (177)	63.5-66.7 (65.0)					185-201 (189)	65.8-75.1 (68.9)						48.4-50.3 (48.7)

Clade IV														
Grosmannia huntii WIN(M)492	216	65.7	30	44	63	79	199	74.4	23	28	63	85	159	50.3
Grosmannia piceaperda WIN(M)980	214	66.4	30	42	60	82	203	75.9	22	27	64	90	159	50.3
Grosmannia piceaperda WIN(M)1380	232	66.8	34	43	69	86	233	75.1	25	33	77	98	159	50.3
Minimum-maximum (mean)	214-232 (221)	65.7-66.8 (66.3)					199-233 (212)	74.4-75.9 (75.1)						
Clade Va														
Grosmannia aurea CBS438.68	221	65.2	32	45	65	79	200	74.0	24	28	64	84	159	50.3
Leptographium terebrantis CBS298.85	227	65.2	34	45	66	82	200	74.0	24	28	64	84	159	50.3
Leptographium terebrantis CBS337.70	226	65.5	33	45	66	82	200	74.0	24	28	64	84	159	50.3
Leptographium terebrantis CBS408.61	220	64.5	32	46	64	78	200	74.0	24	28	64	84	159	50.3
Leptographium terebrantis UAMH9690	222	65.3	33	44	64	81	200	74.5	24	27	64	85	159	50.3
Leptographium terebrantis UAMH9722	221	65.2	32	45	65	79	200	73.5	25	28	64	83	159	50.3
Leptographium wingfieldii CBS645.89	220	65.0	32	45	64	79	200	74.0	24	28	64	84	159	50.3
Leptographium wingfieldii CBS648.89	220	65.0	32	45	64	79	201	74.1	24	28	64	85	160	50.6
Leptographium wingfieldii MCC125	224	66.1	32	44	67	81	200	74.5	24	27	64	85	159	50.3
Leptographium wingfieldii MCC130	223	65.5	33	44	64	82	200	74.0	24	28	64	84	159	48.4
Leptographium wingfieldii MCC349	220	65.0	32	45	64	79	200	74.0	24	28	64	84	156	49.4
Leptographium wingfieldii TOM10.2	217	64.5	32	45	61	79	200	74.0	24	28	64	84	159	50.3
Leptographium wingfieldii TOM11.5	217	64.5	32	45	61	79	200	74.0	24	28	63	85	159	50.3
Leptographium wingfieldii TOM59.21	217	64.5	32	45	61	79	200	73.5	25	28	63	84	159	50.3
Leptographium wingfieldii WIN(M)1218	225	65.3	33	45	65	82	200	74.0	24	28	64	84	159	50.3
Leptographium wingfieldii WIN(M)1322	223	65.9	32	44	66	81	200	73.5	24	29	64	83	159	50.3
Leptographium wingfieldii WIN(M)1382	222	65.3	33	44	64	81	200	74.5	24	27	64	85	159	50.3
Leptographium wingfieldii WIN(M)1482	222	65.3	32	45	66	79	200	74.0	24	28	64	84	159	50.3
Minimum-maximum (mean)	217-227 (221)	64.5-66.1 (65.2)					200-201 (200)	73.5-74.5 (74.0)					156-160 (159)	48.4-50.6 (50.2)
Clade Vb														
Leptographium lundbergii CBS352.29	218	67.9	33	37	63	85	199	74.9	23	27	63	86	159	50.3
Leptographium lundbergii DAOM64746	223	64.1	36	44	63	80	200	74.0	23	29	63	85	159	50.3
Leptographium lundbergii DSMZ5010	218	67.9	33	37	63	85	198	74.7	23	27	62	86	159	50.3
Leptographium lundbergii NFRI60-25	217	68.2	33	36	63	85	199	74.9	23	27	63	86	159	50.3
Leptographium lundbergii NFRI69-148	218	67.9	33	37	63	85	200	74.0	24	28	62	86	159	49.7
Leptographium lundbergii NFRI89-1040/1/3	218	67.9	33	37	63	85	199	74.4	24	27	63	85	159	50.3
Minimum-maximum (mean)	217-223 (219)	64.1-68.2 (67.3)					198-200 (199)	74.0-74.9 (74.5)						49.7-50.3 (50.2)
Clade Vc														
Leptographium sp. WIN(M)1106	219	68.5	31	38	63	87	200	74.5	23	28	63	86	159	50.3

Leptographium sp. WIN(M)1247	215	67.4	31	39	63	82	200	74.5	23	28	63	86	159	50.3
Leptographium sp. WIN(M)1269	219	68.5	30	39	65	85	200	74.5	23	28	63	86	159	50.3
Leptographium truncatum CBS929.85	214	68.7	30	37	63	84	200	75.5	23	26	64	87	159	50.3
Leptographium truncatum NFRI1813/1	215	68.8	30	37	63	85	200	75.5	23	26	64	87	159	50.3
Leptographium truncatum TOM74.29	215	68.8	30	37	63	85	200	75.5	23	26	64	87	159	50.3
Leptographium truncatum TOM86.30	215	67.9	31	38	63	83	200	74.5	23	28	63	86	159	50.3
Minimum-maximum (mean)	214-219 (216)	67.4-68.8 (68.4)						74.5-75.5 (74.9)						
Clade VI														
Leptographium sp. WIN(M)984	196	64.8	35	34	56	71	209	68.4	31	35	63	80	159	49.7
Leptographium sp. WIN(M)985	196	64.8	35	34	56	71	210	68.1	31	36	63	80	159	49.7
Clade VII														
Grosmannia europhioides CBS229.83	224	61.2	40	47	61	76	200	69.5	23	38	62	77	159	50.3
Grosmannia europhioides MUCL18355	222	59.9	41	48	58	75	200	70.5	23	36	62	79	159	50.3
Grosmannia europhioides NFRI80-67/22	224	61.2	40	47	61	76	199	69.8	23	37	62	77	159	50.3
Grosmannia laricis CBS636.94	223	60.5	40	48	60	75	200	70.0	23	37	62	78	159	50.3
Minimum-maximum (mean)	222-224 (223)	59.9-61.2 (60.7)					199-200 (200)	69.5-70.5 (69.9)						
Clade VIII														
Grosmannia wageneri ATCC58579	202	67.3	31	35	56	80	198	73.7	25	27	64	82	159	50.3
Leptographium serpens DAOM173660	206	68.4	28	37	59	82	198	73.2	26	27	64	81	159	49.7
Clade IX														
Leptographium procerum DAOM33940	174	64.4	28	34	44	68	203	71.9	26	31	63	83	159	50.3
Leptographium procerum NFRI59-84/2	184	63.6	30	37	49	68	203	71.9	26	31	63	83	159	50.3
Leptographium procerum TOM73.12	207	67.1	30	38	60	79	203	71.9	26	31	63	83	159	50.3
Leptographium procerum TOM76.8	207	67.1	30	38	60	79	203	71.9	26	31	63	83	159	50.3
Leptographium procerum WIN(M)1264	207	67.1	30	38	61	78	203	71.9	26	31	63	83	159	50.3
Minimum-maximum (mean)	174-207 (196)	63.6-67.1 (65.9)												
Clade X														
Grosmannia davidsonii WIN(M)1495	171	66.1	27	31	49	64	177	75.7	25	18	59	75	159	49.1
Pesotum sp. WIN(M)481	171	66.1	27	31	49	64	177	75.7	25	18	60	74	159	49.1

[1] The strains used as the phylogenetic outgroups as shown in Figure 3.3.

[2] Clades are described in Figure 3.3.

3.3.2. Characterizing the DNA sequence of ITS1 and ITS2 with DNA sequence logos

DNA sequence logos (Figures 3.1A and 3.2A) were constructed for both ITS segments in order to easily visualize conserved regions and local sequence biases, such as expansion or contraction segments and mono- and di-nucleotide repeats. The level of sequence conservation at each position within the sequence is indicated by the overall height of the stack in the logo, while the relative frequency of the individual nucleotides is indicated by the relative heights of each symbol within the stack. Within ITS1, there were two GC-rich regions in the hairpin that were potential hotspots for compensatory changes, including nucleotide substitutions and indels; the latter were responsible for the expansion or contraction of the length of the hairpin (Figure 3.1B). These base pairing interactions may be formed between nucleotides that are separated by a great distance in the primary sequence: the segment near the 5' terminus, which was dominated by GA repeats (positions 58 to 68, as numbered in Figure 3.1A), formed an extended hairpin with a downstream C-/CT-rich sequence (positions 199 to 207). Near the terminus of the helix (Figure 3.1B), the 5' CT stretch (positions 100 to 117) formed a hairpin with a series of Gs on the 3' side (positions 123 to 142) of the stem.

Comparing the sequence logo with the original sequence alignment revealed sites of potential replication slippage, which are defined in this study as sites at which the sequence for one or two strains contains a stretch of the same base or a repeat of a short nucleotide pattern that is not found in the remaining strains. Putative replication slippage was identified in *G. piceaperda* strain WIN(M)1380: within a GC-rich stretch there was an insert of CGG (positions 188 to 190), followed by a stretch of Gs (see Figure 3.1A, positions 209 to 212). There were helical-loop regions near the

termini of ITS1 in which the nucleotides were highly conserved. These regions are part of the central helix, which previous biochemical studies have shown is involved in forming a complex with soluble protein factors involved in rRNA maturation/ ribosome biogenesis (Lalev and Nazar, 1999).

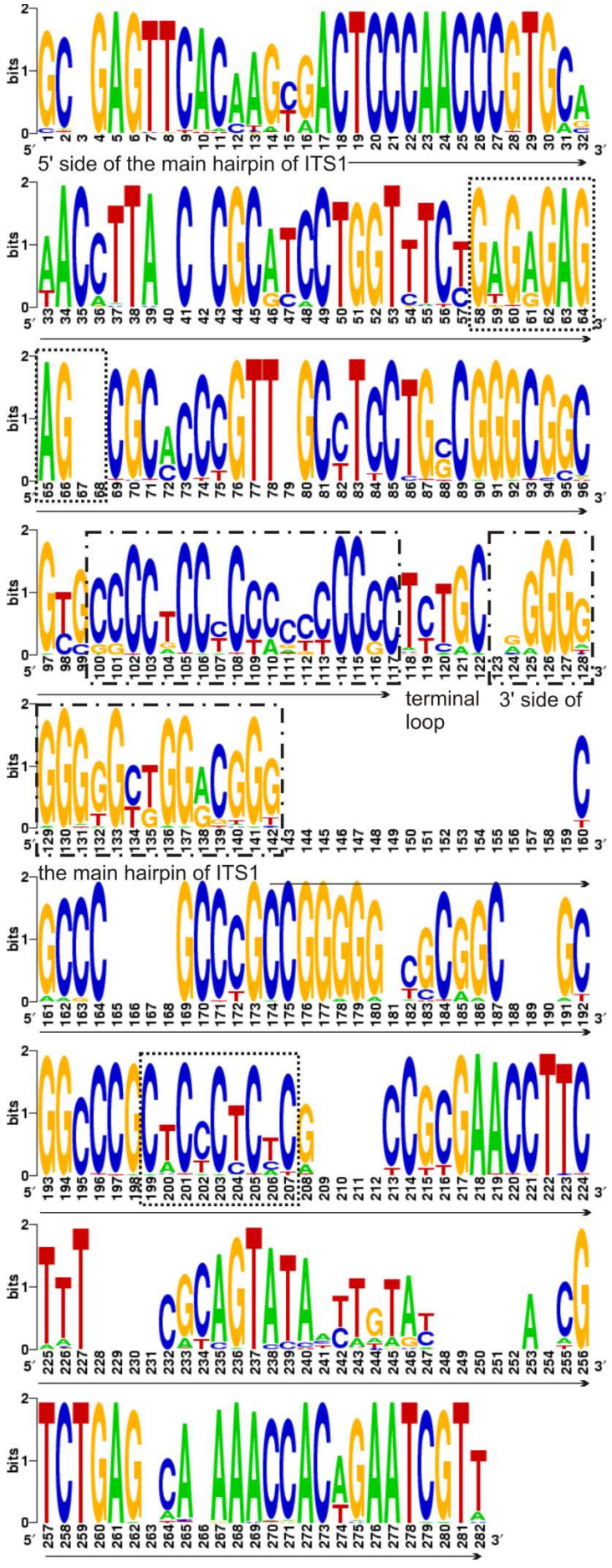
Within the ITS1 logo shown in Figure 3.1A, a number of empty positions are apparent; these represent sites at which there was no sequence conservation, often resulting from an insertion in one or several sequences. Positions (based on the numbering of the logo) 3, 40, 42, 79, 181, and 263 corresponded to an insertion of an A, G, A, C, G, and T at the respective position. Positions 67 to 68 corresponded to an insertion of AG, whereas at position 123 there were insertions of 2 As and 2 Gs and 1 C in five different taxa. Insertions in two taxa corresponding to A and C, G and A, G and C, and G and C were found at positions 248, 252, 254, and 266, respectively. Positions 165 to 158, 228 to 231, and 249 to 251 corresponded to insertions of AGGC, AACA, and GAA, respectively. A lengthy insertion was present between positions 143 to 159 in the outgroup taxon *Ceratocystis deltoideospora* WIN(M)41.

In ITS2, the 5' terminus was nearly invariant (Figure 3.2A): the 3' terminus of the 5.8S gene and the first three nucleotides of ITS2 formed a hairpin with the 5' terminus of the LSU (Figure 3.2B). An earlier investigation of the core secondary structure of ITS2 in green algae and flowering plants found the sequence and length of helix I to be variable (Mai and Coleman, 1997). In contrast, in strains of *Leptographium* and *Grosmannia* the terminal loop was variable, while the stem region exhibited little variation in length or sequence. Further, in *G. piceaperda* strain WIN(M)1380 there was evidence suggestive of replication slippage in helix I (Figure 3.2A, B; positions 29 to 31). Helix II was less conserved, particularly in the terminal loop. There were numerous inserts as well as sites of putative replication slippage in

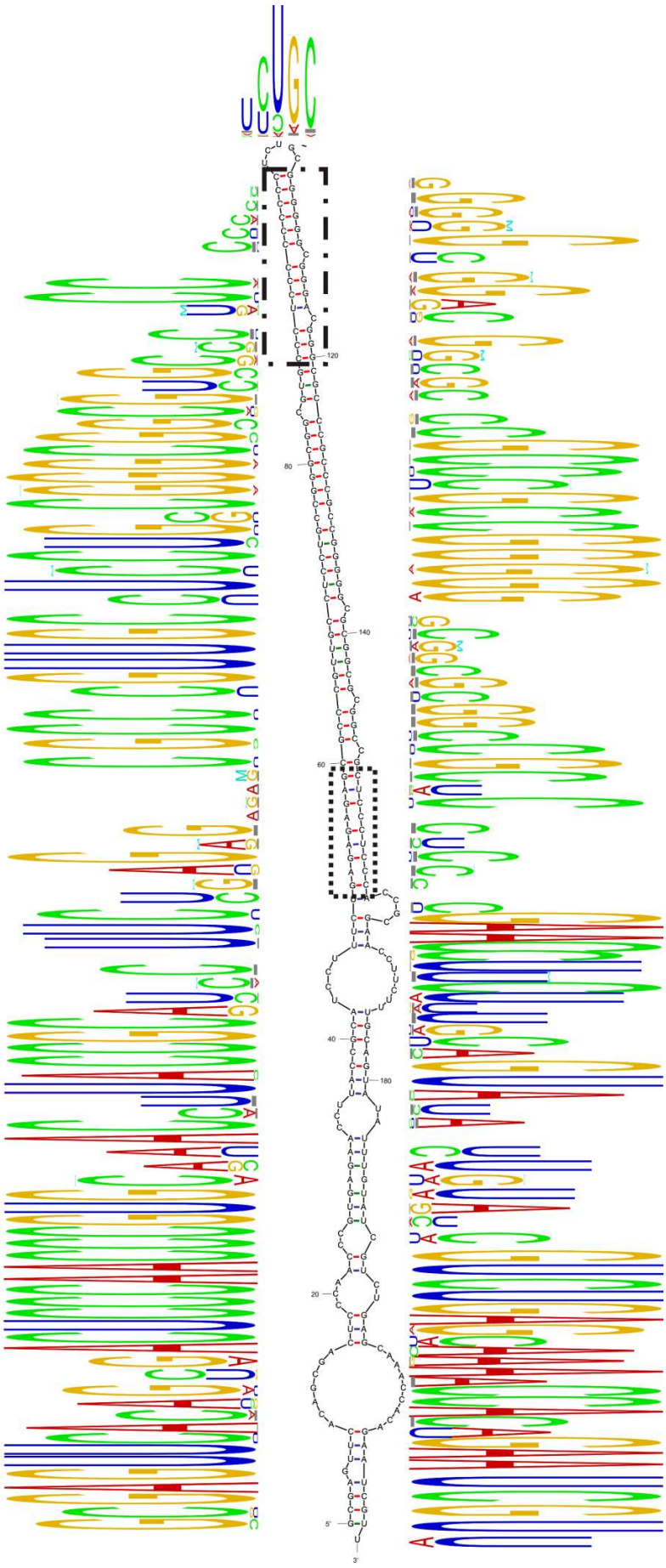
helix II; the stretch of Gs (positions 103 to 110) observed in *G. piceaperda* strain WIN(M)1380 represents one such example. The 3' terminus at the base of helix III was hypervariable and was one of the main sites of potential RNA strand slippage. Helix IV was variable in length and sequence; this variation is potentially influenced by RNA strand slippage events occurring in helix III and in the single-stranded "palm" region between helices III and IV.

Within the ITS2 logo shown in Figure 3.2A, a number of empty positions are also visible, representing sites at which there is an insertion in one to three sequence(s). At positions (based on the numbering of positions in the logo) 7, 51, 211, 223, and 271, insertions of C, T, A, T, and A, respectively, were present in only one of the sequences. Five sequences contained an insertion of a T at position 43. Position 43 was the site of an insertion of an A in one sequence and a G in another. Likewise, one sequence contained an insertion of a C at position 138, while another sequence contained a T at that position. Three sequences had insertions at positions 85 (G, C, or T) and 178 (A, C, or G). Di-nucleotide insertions were identified at positions 58 to 59 (GC), 164 to 165 (CC), and 246 to 247 (CG). Tri-nucleotide insertions were present at positions 145 to 147 (ATC) and 180 to 182 (TTA).

Figure 3.1. DNA sequence and RNA structural features of the ITS1 segment. (A) DNA weblogo of the ITS1 segment in strains of *Grosmannia*, *Leptographium*, and related taxa. Sequence stretches representing potential tandem repeats or expansion/contraction segments are enclosed in boxes (...., for the region near the base of the stem;_._., for the region near the terminal loop). As described in the text, empty positions represent sites at which there is an insertion in one or several sequences. (B) Model of the RNA secondary structure of the ITS1 segment of *L. truncatum* strain TOM86.30. The RNA structure logo is superimposed onto the model. The proposed model suggests that nucleotides found in the series of loops and short hairpins near the 5' and 3' termini are highly conserved. Regions of the main stem representing potential tandem repeats or expansion/contraction segments are enclosed in boxes as described in panel A.

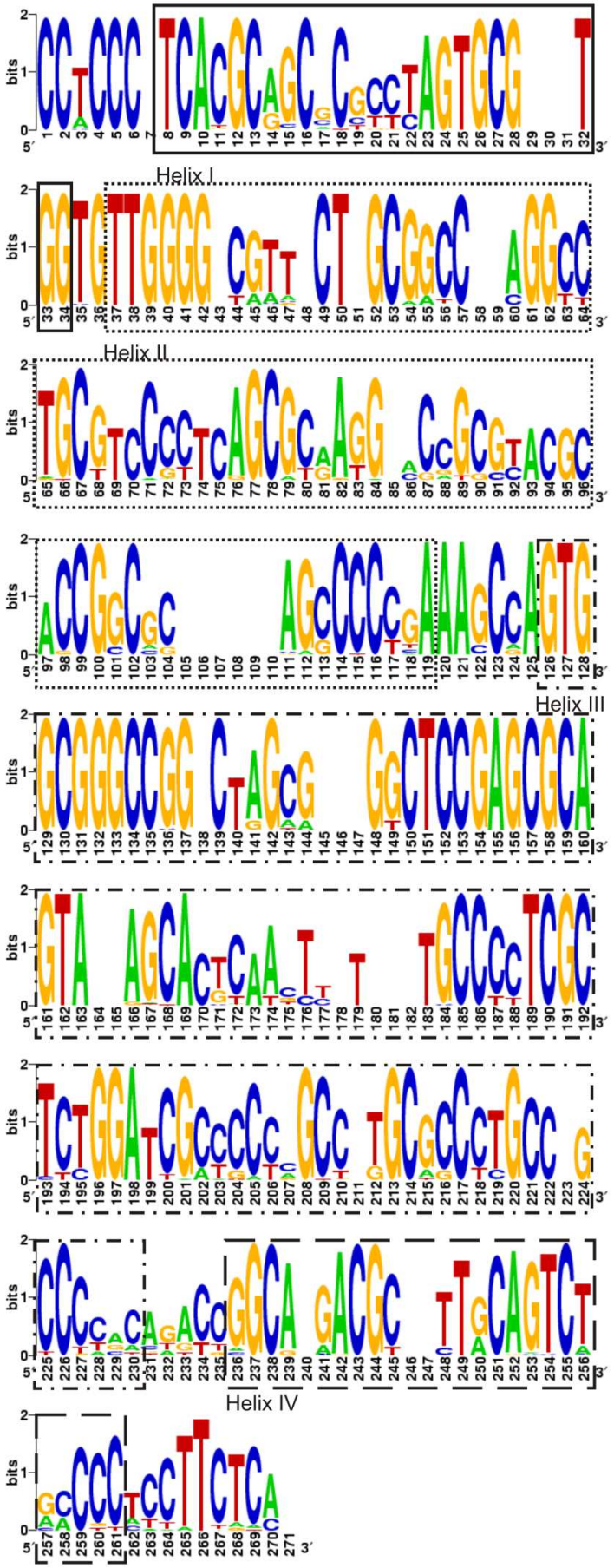

bits
5′
3′
5' side of the main hairpin of ITS1
terminal loop
3' side of
the main hairpin of ITS1

A

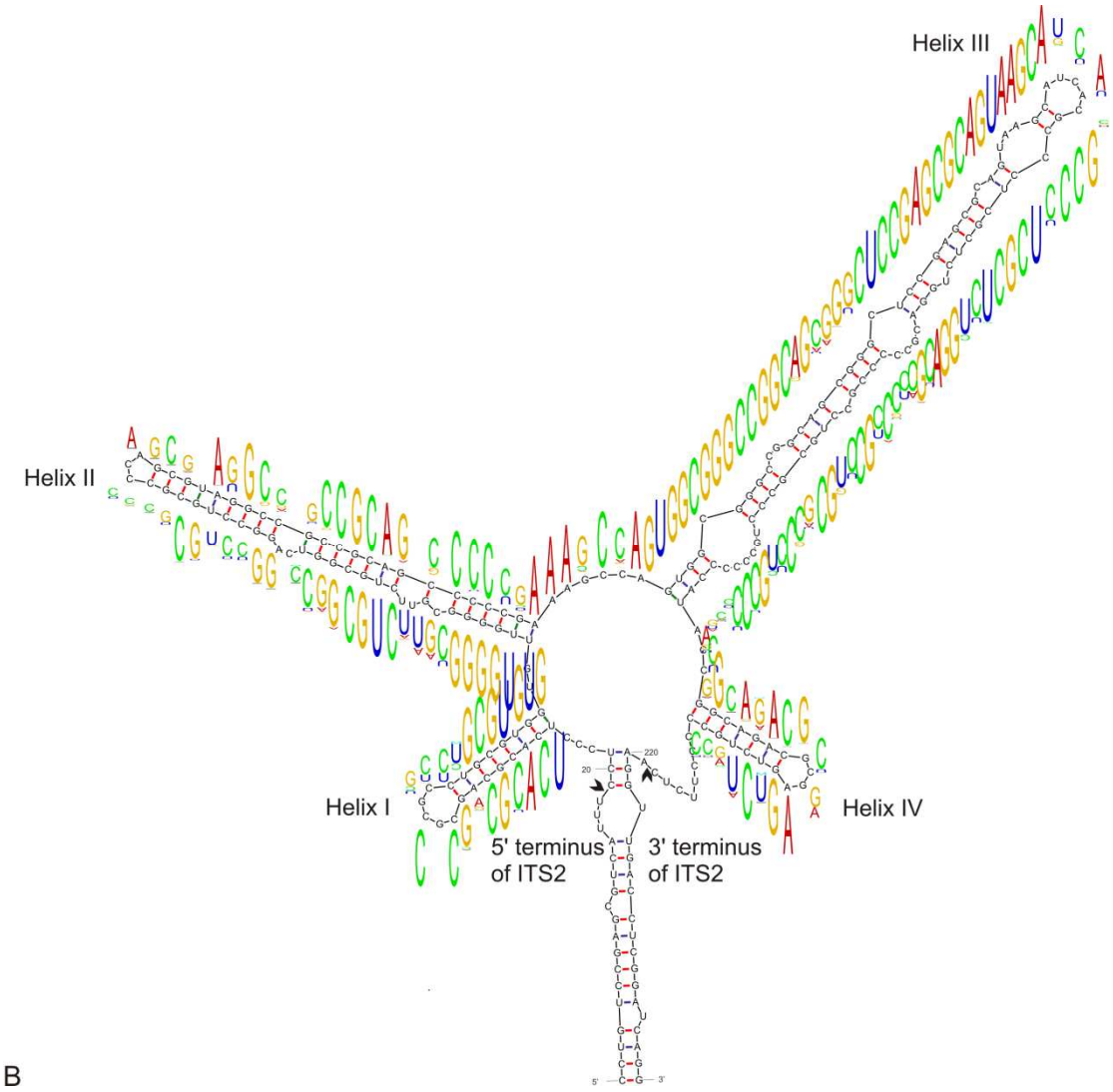


B

Figure 3.2. DNA sequence and RNA structural features of the ITS2 segment. (A) DNA weblogo of the ITS2 region in strains of *Grosmannia*, *Leptographium*, and related taxa. Sequences forming helices I to IV are enclosed in boxes. As described in the text, empty positions represent sites at which there is an insertion in one or several sequences. (B) Model of the RNA secondary structure of the ITS2 segment of *L. truncatum* strain TOM86.30. The RNA structure logo is superimposed onto the model. The model conforms to the pan-eukaryotic four-fingered hand model proposed by Coleman (2007). Nucleotides at the 5' terminus are conserved and base pair with the 5' terminus of the LSU. The 5' and 3' termini of the ITS2 segment are indicated with the black arrow heads.



A


Helix III
Helix II
Helix I
Helix IV
5' terminus
of ITS2
3' terminus
of ITS2
20
220
5'
3'

B

3.3.3. Models of RNA secondary structure of ITS1 and ITS2

Models of RNA secondary structure were generated for ITS1 and ITS2 (Figures 3.1B, 3.2B). For the ITS1 region, an extended hairpin structure that lacked any side helices was identified, as well as several conserved loops and shorter helices (Figure 3.1B). This "core" structure was subsequently used as a template to guide the folding of ITS1 sequences for more distantly-related taxa using comparative sequence analysis, which involves the identification of CBCs to support proposed base pairing interactions. One exception to this model occurred in the RNA structure for *L. procerum* strain NFRI59-84/2, which contained a small lateral helix on the 5' side of the hairpin as a result of a large deletion on the 3' side of the helix that disrupted base pairing with opposing nucleotides on the 5' side. Alternative structures containing small side helices, as described by Nazar et al. (1987), were obtained, indicating that in these GC-rich segments alternative interactions between bases on opposing sides of the hairpin are possible. Flexibility in base pairing is especially relevant in RNA due to GU "wobble" interactions. RNA structural models of ITS2 (Figure 3.2B) conformed to the previously proposed ITS2 "core" secondary structural model: a model of four helices radiating from a central loop, with helix III as the longest hairpin (Coleman, 2007; Keller et al., 2009). This model has been examined using functional genetic assays (Côté et al., 2002).

3.3.4. Evolution of the ITS region

Phylogenetic analyses were carried out on the ITS1-5.8S-ITS2 region and on each segment individually for the purpose of: (i) examining the evolution of DNA sequence and RNA secondary structure amongst closely-related sequences and (ii)

assessing branch support and tree topology obtained from the entire region (Figure 3.3) versus the ITS1 (Figure 3.4) or ITS2 (Figure 3.5) segments.

Ten major clades with strong Bayesian support (0.96 to 1.00 posterior probability values) and moderate to strong (71 to 99 %) support from parsimony or maximum likelihood analysis were identified in the Bayesian tree obtained from analysis of the ITS-5.8S region (Figure 3.3). Clade I is composed of strains of *Hyalopesotum pini* and *Grosmannia galeiformis*. Clades II and X include undescribed strains of *Pesotum* spp. and strains of *Grosmannia davidsonii*; taxonomic novelties with respect to *G. davidsonii* will be addressed in a future study. Clade III includes strains of *Grosmannia penicillata*, *Leptographium americanum*, which has a teleomorphic state in the genus *Grosmannia* (*G. americana*), and *Leptographium* sp. strain J.R. 88-194A [= WIN(M)1376, Hausner et al. (2005)]. Strains of *G. piceaperda* form clade IV. Clade V is subdivided into three subclades, comprising strains of *L. wingfieldii-L. terebrantis-G. aurea* species complex (clade Va), *L. lundbergii* (clade Vb), and *L. truncatum* (clade Vc). Undescribed strains of *Leptographium* spp. are grouped separately in clade VI. Strians of *G. europhioides* and *G. laricis* are grouped together in clade VII. *Grosmannia wageneri*, which has an anamorphic, *Leptographium*, state, is related to *L. serpens*, and together they form clade VIII. Lastly, strains of *L. procerum* group in clade IX. As suggested by the topology of the tree and its branch lengths ITS variability is more pronounced between the clades, rather than within the species examined in this study.

Phylogenetic analysis of the ITS1 segment alone resulted in a tree with the same ten major clades obtained from the analysis of the entire ITS-5.8S region (compare Figures 3.3 and 3.4); however, many of the clades were without support from Bayesian analysis and the overall topology of the tree, especially the relative

positions of the clades, differs. When ITS2 is used as the phylogenetic marker, species distinction and branch support are significantly reduced and several clades are reduced to polytomies (Figure 3.5). This is especially evident with members of clade IX: all five strains of *L. procerum* share identical ITS2 segments and are differentiated from each other based on the sequence of ITS1. Thus, the ability to differentiate amongst these strains of *L. procerum* is lost when ITS2 is used as the sole molecular marker. While sequence comparison of ITS1 recovered thirteen of the eighteen unique strains within the *L. wingfieldii-L. terebrantis-G. aurea* species complex, only eight strains could be identified based on sequence identity of ITS2. In addition, *G. aurea* strain CBS438.89, a strain with a known teleomorphic (sexual) state was indistinguishable from the anamorphic (asexual) *Leptographium* strains based on its ITS2 sequence. Also, clades Va (comprising members of the *G. aurea-L. wingfieldii-L. terebrantis*), III, and IV are split and rearranged in the ITS2-derived tree.

An extreme example of distinct strains sharing identical ITS segments was represented by the *Leptographium* sp. strain J.R. 88-194A [= WIN(M)1376, Hausner et al. (2005)]: sequence identity of ITS1 between *Leptographium* sp. strain J.R. 88-194A and *G. penicillata* strain WIN(M)131 was 100 %, while the sequence identity of ITS2 between *Leptographium* sp. strain J.R. 88-194A and *L. truncatum* strains NFRI1813/1, TOM74.29 [=WIN(M)1246; AY935181; see Figure 3.3], and CBS929.85 was 99 %. Unlike the examples described in the previous section, phylogenetic analysis of the ITS-5.8S region indicated that these strains do not share a recent common ancestor (Figure 3.3). Consequently, the phylogenetic position of *Leptographium* sp. strain J.R. 88-194A (identified by the black arrow in Figures 3.3 to 3.5) appeared to be unstable, as it depended on the particular segment within the ITS region that was subjected to phylogenetic analysis. When ITS1 alone, or the entire

ITS region was analyzed, *Leptographium* sp. strain J.R. 88-194A and *G. penicillata* strain WIN(M)131 were grouped in the same clade (clade III in Figures 3.3 and 3.4). However, analysis based on just the ITS2 segment placed *Leptographium* sp. strain J.R. 88-194A in clade Vc, with strains of *L. truncatum*. A similar phenomenon has been recently observed between fungal isolates of *Ceratobasidium oryzae-sativae* and *Thanatephorus cucumeris* (Xie et al., 2008). In the study the authors suggest the possibility of a rare intergeneric hybridization event that may have allowed for a mitotic genetic recombination event that generated biological chimeric ITS forms, but the mechanism behind this phenomenon was not described. One could assume it involved rare hyphal fusion between two different fungal species that allowed for the transfer of nuclei. This could have eventually led to the fusion of nuclei between two different species and mitotic recombination, a situation analogous to a parasexual cycle (Tinline and MacNeil 1969; Croll et al. 2009).

Amongst the fungi treated in this study, ITS1 sequences were noted to be less conserved than ITS2 sequences. This has been observed in other fungal taxa (Goertzen et al., 2003; Schultz et al., 2005; Wolf et al., 2005; Piercey-Normore et al., 2006), but it has been noted recently that there are many examples within the fungi in which the ITS2 segment is more variable (Nilsson et al., 2008). Overall, within the data set a greater variability was observed in the size of the ITS1 segment compared to that of ITS2. The ITS1 segment was also more variable in sequence than ITS2; of the 70 strains in the data set, there were 55 unique ITS1 sequences and only 46 unique ITS2 sequences. Furthermore, amongst closely-related strains of ophiostomatoid fungi the ITS region was less variable, but between more distantly-related taxa, there was increased sequence ambiguity in the variable regions of the ITS segments. The ITS1 segment exhibited the greatest variability in length and GC content in the *L.*

wingfieldii-L. terebrantis-G. aurea species complex compared to the other subclades. Two strains of *L. terebrantis* CBS298.85 and *L. terebrantis* CBS337.70 were remarkable for having numerous insertions in the ITS1 region, yet their ITS2 sequence was identical to each other and to *L. wingfieldii* strain WIN(M)1218 (Figures 3.4 and 3.5). Diversity in the ITS region amongst strains belonging to clade IV was the result of numerous insertions and potential replication slippage events observed in *G. piceaperda* strain WIN(M)1380 (Figure 3.1A). Members of clade IX exhibited the greatest variation in the ITS1 segment. *Leptographium procerum* strains NFRI59-84/2 and DAOM33940 had unusually short ITS1 sequences of 184 and 174 nt, respectively, as a result of large deletions in GC-rich regions on the 3' side of the hairpin.

In contrast to the variability observed in the size of ITS1 segment, there is little variation in the size of the ITS2 region, even in those strains with reduced ITS1 segments (Table 3.2, Figure 3.3). In *Pesotum* sp. strain WIN(M)481 and *G. davidsonii* strain WIN(M)1495 the ITS segments were similar in size: ITS1 and ITS2 were, respectively, 171 nt and 177 nt in length. In strain *G. piceaperda* strain WIN(M)1380 both segments were greater in length (232 nt and 233 nt for ITS1 and ITS2, respectively) than the average size for the other two members of clade IV (average was 221 nt and 212 nt for ITS1 and ITS2, respectively). The increase in size may have been due to possible replication slippage (Figures 3.1 and 3.2).

Figure 3.3. Phylogenetic analysis of nuclear ITS1-5.8SrDNA-ITS2 DNA sequences in strains of *Grosmannia*, *Leptographium*, and related fungal taxa. Branch lengths were determined using the Bayesian consensus outfile. For Bayesian analysis, the TVM+I+G model (based on AIC) was selected. Values at the nodes were determined using algorithms implemented by DNA PARS/Tree Puzzle/Mr Bayes programs. "-" indicates the node is absent or the posterior probability or bootstrap value is not well supported (a posterior probability value of less than 0.95 for Bayesian analysis and bootstrap values less than 70 % for parsimony and maximum likelihood analyses). The parsimony analysis combined with bootstrap analysis as performed by the TNT program (Tree analysis using New Technology; http://www.zmuc.dk/public/ Phylogeny/TNT/) generated a tree topology essentially identical to that obtained from DNAPARS analysis. The ten major clades are indicated in roman numerals. Based on morphological criteria and molecular data *Ophiostoma brevicolle* should be included within the genus *Grosmannia* Goid. (Hausner et al., 1993b; Zipfel et al., 2006). *Leptographium* sp. strain J.R. 88-194A is marked by the black arrow; its position varies depending on which segment of the ITS region is used as the phylogenetic marker.

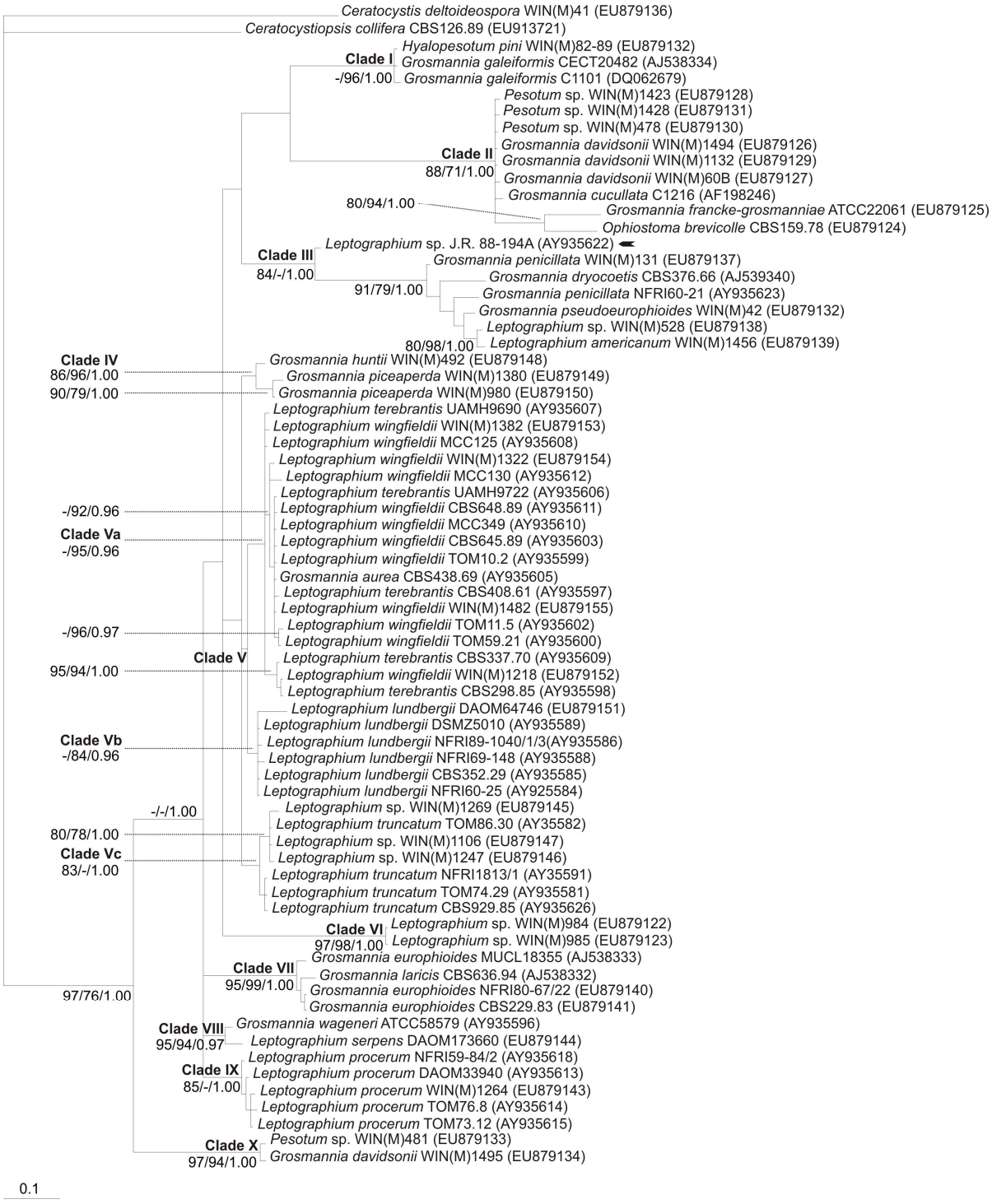
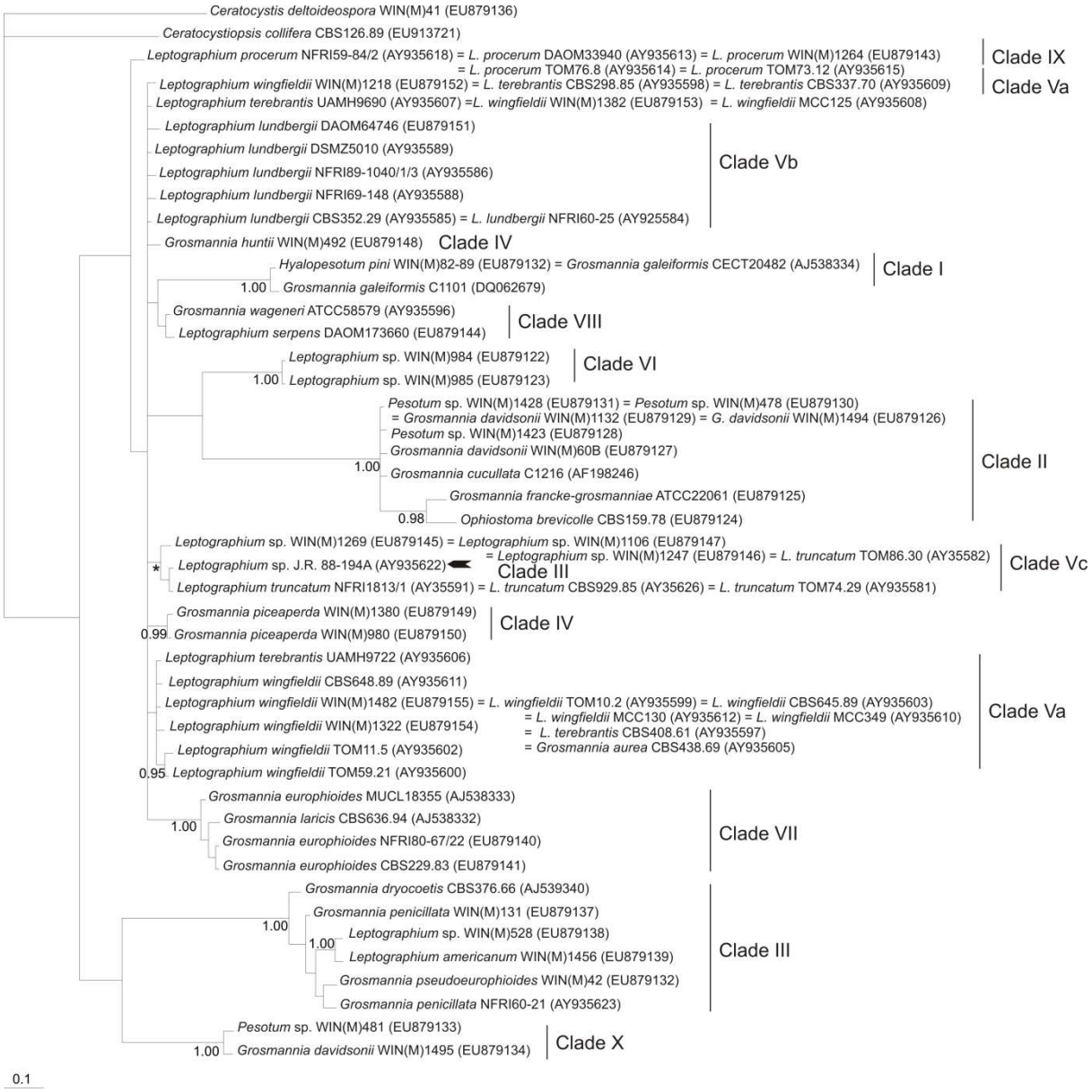

Ceratocystis deltoideospora WIN(M)41 (EU879136)
Ceratocystiopsis collifera CBS126.89 (EU913721)
Clade I
-/96/1.00
Hyalopesotum pini WIN(M)82-89 (EU879132)
Grosmannia galeiformis CECT20482 (AJ538334)
Grosmannia galeiformis C1101 (DQ062679)
Clade II
88/71/1.00
Pesotum sp. WIN(M)1423 (EU879128)
Pesotum sp. WIN(M)1428 (EU879131)
Pesotum sp. WIN(M)478 (EU879130)
Grosmannia davidsonii WIN(M)1494 (EU879126)
Grosmannia davidsonii WIN(M)1132 (EU879129)
Grosmannia davidsonii WIN(M)60B (EU879127)
Grosmannia cucullata C1216 (AF198246)
80/94/1.00
Grosmannia francke-grosmanniae ATCC22061 (EU879125)
Ophiostoma brevicolle CBS159.78 (EU879124)
Clade III
84/-/1.00
Leptographium sp. J.R. 88-194A (AY935622)
Grosmannia penicillata WIN(M)131 (EU879137)
Grosmannia dryocoetis CBS376.66 (AJ539340)
91/79/1.00
Grosmannia penicillata NFRI60-21 (AY935623)
Grosmannia pseudoeurophioides WIN(M)42 (EU879132)
Leptographium sp. WIN(M)528 (EU879138)
80/98/1.00
Leptographium americanum WIN(M)1456 (EU879139)
Clade IV
86/96/1.00
90/79/1.00
Grosmannia huntii WIN(M)492 (EU879148)
Grosmannia piceaperda WIN(M)1380 (EU879149)
Grosmannia piceaperda WIN(M)980 (EU879150)
Leptographium terebrantis UAMH9690 (AY935607)
Leptographium wingfieldii WIN(M)1382 (EU879153)
Leptographium wingfieldii MCC125 (AY935608)
Leptographium wingfieldii WIN(M)1322 (EU879154)
Leptographium wingfieldii MCC130 (AY935612)
Leptographium terebrantis UAMH9722 (AY935606)
-/92/0.96
Leptographium wingfieldii CBS648.89 (AY935611)
Clade Va
-/95/0.96
Leptographium wingfieldii MCC349 (AY935610)
Leptographium wingfieldii CBS645.89 (AY935603)
Leptographium wingfieldii TOM10.2 (AY935599)
Grosmannia aurea CBS438.69 (AY935605)
Leptographium terebrantis CBS408.61 (AY935597)
Leptographium wingfieldii WIN(M)1482 (EU879155)
-/96/0.97
Leptographium wingfieldii TOM11.5 (AY935602)
Leptographium wingfieldii TOM59.21 (AY935600)
Clade V
95/94/1.00
Leptographium terebrantis CBS337.70 (AY935609)
Leptographium wingfieldii WIN(M)1218 (EU879152)
Leptographium terebrantis CBS298.85 (AY935598)
Leptographium lundbergii DAOM64746 (EU879151)
Leptographium lundbergii DSMZ5010 (AY935589)
Clade Vb
-/84/0.96
Leptographium lundbergii NFRI89-1040/1/3(AY935586)
Leptographium lundbergii NFRI69-148 (AY935588)
Leptographium lundbergii CBS352.29 (AY935585)
Leptographium lundbergii NFRI60-25 (AY925584)
-/-/1.00
Leptographium sp. WIN(M)1269 (EU879145)
80/78/1.00
Leptographium truncatum TOM86.30 (AY35582)
Clade Vc
83/-/1.00
Leptographium sp. WIN(M)1106 (EU879147)
Leptographium sp. WIN(M)1247 (EU879146)
Leptographium truncatum NFRI1813/1 (AY35591)
Leptographium truncatum TOM74.29 (AY935581)
Leptographium truncatum CBS929.85 (AY935626)
Clade VI
97/98/1.00
Leptographium sp. WIN(M)984 (EU879122)
Leptographium sp. WIN(M)985 (EU879123)
Grosmannia europhioides MUCL18355 (AJ538333)
Clade VII
95/99/1.00
Grosmannia laricis CBS636.94 (AJ538332)
Grosmannia europhioides NFRI80-67/22 (EU879140)
97/76/1.00
Grosmannia europhioides CBS229.83 (EU879141)
Clade VIII
95/94/0.97
Grosmannia wageneri ATCC58579 (AY935596)
Leptographium serpens DAOM173660 (EU879144)
Leptographium procerum NFRI59-84/2 (AY935618)
Clade IX
85/-/1.00
Leptographium procerum DAOM33940 (AY935613)
Leptographium procerum WIN(M)1264 (EU879143)
Leptographium procerum TOM76.8 (AY935614)
Leptographium procerum TOM73.12 (AY935615)
Clade X
97/94/1.00
Pesotum sp. WIN(M)481 (EU879133)
Grosmannia davidsonii WIN(M)1495 (EU879134)
0.1

Figure 3.4. Phylogenetic analysis of the ITS1 rDNA segment in strains of *Grosmannia*, *Leptographium*, and related fungal taxa. Branch lengths were determined using the Bayesian consensus outfile. For Bayesian analysis, the TVM+G+I model (based on the AIC) was selected and nodes with posterior probabilities of 0.95 and greater are marked. *Leptographium* sp. strain J.R. 88-194A is indicated by the black arrow. The labelling of clades on the right side of the figure is based on the identification of clades obtained from analysis of the entire ITS-5.8S region (see Figure 3.3).


Ceratocystis deltoideospora WIN(M)41 (EU879136)
Ceratocystiopsis collifera CBS126.89 (EU913721)
Leptographium sp. WIN(M)984 (EU879122) =Leptographium sp. WIN(M)985 (EU879123)
Pesotum sp. WIN(M)481 (EU879133) = Grosmannia davidsonii WIN(M)1495 (EU879134)
Hyalopesotum pini WIN(M)82-89 (EU879132) = Grosmannia galeiformis C1101 (DQ062679)
Grosmannia galeiformis CECT20482 (AJ538334)
Grosmannia penicillata WIN(M)131 (EU879137) = Leptographium sp. J.R. 88-194A (AY935622)
Leptographium sp. WIN(M)528 (EU879138)
Leptographium americanum WIN(M)1456 (EU879139)
Grosmannia pseudoeurophioides WIN(M)42 (EU879132)
Grosmannia penicillata NFRI60-21 (AY935623)
Grosmannia dryocoetis CBS376.66 (AJ539340)
Leptographium procerum NFRI59-84/2 (AY935618)
Leptographium procerum DAOM33940 (AY935613)
Leptographium procerum WIN(M)1264 (EU879143)
Leptographium procerum TOM76.8 (AY935614)
Leptographium procerum TOM73.12 (AY935615)
Grosmannia wageneri ATCC58579 (AY935596)
Leptographium serpens DAOM173660 (EU879144)
Grosmannia davidsonii WIN(M)1494 (EU879126)
Pesotum sp. WIN(M)1423 (EU879128) = WIN(M)478 (EU879130)
Pesotum sp. WIN(M)1428 (EU879131)
Grosmannia davidsonii WIN(M)1132 (EU879129) = G. davidsonii WIN(M)60B (EU879127)
Grosmannia cucullata C1216 (AF198246)
Grosmannia francke-grosmanniae ATCC22061 (EU879125)
Ophiostoma brevicolle CBS159.78 (EU879124)
Grosmannia huntii WIN(M)492 (EU879148)
Grosmannia piceaperda WIN(M)1380 (EU879149)
Grosmannia piceaperda WIN(M)980 (EU879150)
Leptographium terebrantis UAMH9722 (AY935606) = Grosmannia aurea CBS438.69 (AY935605)
Leptographium wingfieldii MCC349 (AY935610) = L. wingfieldii CBS645.89 (AY935603) = L. wingfieldii CBS648.89 (AY935611)
Leptographium wingfieldii TOM10.2 (AY935599) = L. wingfieldii TOM11.5 (AY935602) =L. wingfieldii TOM59.21 (AY935600)
Leptographium terebrantis CBS408.61 (AY935597)
Leptographium wingfieldii WIN(M)1482 (EU879155)
Leptographium wingfieldii WIN(M)1322 (EU879154)
Leptographium wingfieldii MCC130 (AY935612)
Leptographium terebrantis UAMH9690 (AY935607)
Leptographium wingfieldii WIN(M)1382 (EU879153)
Leptographium wingfieldii MCC125 (AY935608)
Leptographium terebrantis CBS337.70 (AY935609)
Leptographium wingfieldii WIN(M)1218 (EU879152)
Leptographium terebrantis CBS298.85 (AY935598)
Grosmannia europhioides NFRI80-67/22 (EU879140) = G. europhioides CBS229.83 (EU879141)
Grosmannia laricis CBS636.94 (AJ538332)
Grosmannia europhioides MUCL18355 (AJ538333)
Leptographium lundbergii DAOM64746 (EU879151)
Leptographium lundbergii NFRI69-148 (AY935588) = L. lundbergii NFRI89-1040/1/3 (AY935586) = L. lundbergii DSMZ5010 (AY935589)
= L. lundbergii CBS352.29 (AY935585)
Leptographium lundbergii NFRI60-25 (AY925584)
Leptographium sp. WIN(M)1269 (EU879145)
Leptographium sp. WIN(M)1247 (EU879146)
Leptographium sp. WIN(M)1106 (EU879147)
Leptographium truncatum TOM86.30 (AY35582)
Leptographium truncatum NFRI1813/1 (AY35591)
Leptographium truncatum CBS929.85 (AY935626)
Leptographium truncatum TOM74.29 (AY935581)
1.00
1.00
1.00
1.00
0.99
1.00
0.97
1.00
1.00
0.96
1.00
0.99
Clade VI
Clade X
Clade I
Clade III
Clade IX
Clade VIII
Clade II
Clade IV
Clade Va
Clade VII
Clade Vb
Clade Vc
0.1

Figure 3.5. Phylogenetic analysis of the ITS2 rDNA segment in strains of *Grosmannia*, *Leptographium*, and related fungal taxa. Branch lengths were determined using the Bayesian consensus outfile. For Bayesian analysis, the TVM+G model (based on the AIC) was selected and nodes with posterior probabilities of 0.95 and greater are indicated. The node marked by the asterisk, which groups *Leptographium* sp. strain J.R. 88-194A (indicated by the black arrow) with related strains based on ITS2, has a posterior probability value of 0.90. The labelling of clades on the right side of the figure is based on the identification of clades obtained from analysis of the entire ITS-5.8S region (see Figure 3.3).


Ceratocystis deltoideospora WIN(M)41 (EU879136)
Ceratocystiopsis collifera CBS126.89 (EU913721)
Leptographium procerum NFRI59-84/2 (AY935618) = L. procerum DAOM33940 (AY935613) = L. procerum WIN(M)1264 (EU879143)
= L. procerum TOM76.8 (AY935614) = L. procerum TOM73.12 (AY935615)
Clade IX
Leptographium wingfieldii WIN(M)1218 (EU879152) = L. terebrantis CBS298.85 (AY935598) = L. terebrantis CBS337.70 (AY935609)
Clade Va
Leptographium terebrantis UAMH9690 (AY935607) =L. wingfieldii WIN(M)1382 (EU879153) = L. wingfieldii MCC125 (AY935608)
Leptographium lundbergii DAOM64746 (EU879151)
Leptographium lundbergii DSMZ5010 (AY935589)
Leptographium lundbergii NFRI89-1040/1/3 (AY935586)
Leptographium lundbergii NFRI69-148 (AY935588)
Leptographium lundbergii CBS352.29 (AY935585) = L. lundbergii NFRI60-25 (AY925584)
Clade Vb
Grosmannia huntii WIN(M)492 (EU879148)
Clade IV
Hyalopesotum pini WIN(M)82-89 (EU879132) = Grosmannia galeiformis CECT20482 (AJ538334)
Clade I
1.00
Grosmannia galeiformis C1101 (DQ062679)
Grosmannia wageneri ATCC58579 (AY935596)
Leptographium serpens DAOM173660 (EU879144)
Clade VIII
Leptographium sp. WIN(M)984 (EU879122)
Clade VI
1.00
Leptographium sp. WIN(M)985 (EU879123)
Pesotum sp. WIN(M)1428 (EU879131) = Pesotum sp. WIN(M)478 (EU879130)
= Grosmannia davidsonii WIN(M)1132 (EU879129) = G. davidsonii WIN(M)1494 (EU879126)
Pesotum sp. WIN(M)1423 (EU879128)
Grosmannia davidsonii WIN(M)60B (EU879127)
1.00
Grosmannia cucullata C1216 (AF198246)
Clade II
Grosmannia francke-grosmanniae ATCC22061 (EU879125)
0.98
Ophiostoma brevicolle CBS159.78 (EU879124)
Leptographium sp. WIN(M)1269 (EU879145) = Leptographium sp. WIN(M)1106 (EU879147)
= Leptographium sp. WIN(M)1247 (EU879146) = L. truncatum TOM86.30 (AY35582)
Clade Vc
*
Leptographium sp. J.R. 88-194A (AY935622)
Clade III
Leptographium truncatum NFRI1813/1 (AY35591) = L. truncatum CBS929.85 (AY35626) = L. truncatum TOM74.29 (AY935581)
Grosmannia piceaperda WIN(M)1380 (EU879149)
Clade IV
0.99
Grosmannia piceaperda WIN(M)980 (EU879150)
Leptographium terebrantis UAMH9722 (AY935606)
Leptographium wingfieldii CBS648.89 (AY935611)
Leptographium wingfieldii WIN(M)1482 (EU879155) = L. wingfieldii TOM10.2 (AY935599) = L. wingfieldii CBS645.89 (AY935603)
= L. wingfieldii MCC130 (AY935612) = L. wingfieldii MCC349 (AY935610)
Leptographium wingfieldii WIN(M)1322 (EU879154)
= L. terebrantis CBS408.61 (AY935597)
Leptographium wingfieldii TOM11.5 (AY935602)
= Grosmannia aurea CBS438.69 (AY935605)
Clade Va
0.95
Leptographium wingfieldii TOM59.21 (AY935600)
Grosmannia europhioides MUCL18355 (AJ538333)
Grosmannia laricis CBS636.94 (AJ538332)
1.00
Grosmannia europhioides NFRI80-67/22 (EU879140)
Grosmannia europhioides CBS229.83 (EU879141)
Clade VII
Grosmannia dryocoetis CBS376.66 (AJ539340)
Grosmannia penicillata WIN(M)131 (EU879137)
1.00
Leptographium sp. WIN(M)528 (EU879138)
1.00
Leptographium americanum WIN(M)1456 (EU879139)
Grosmannia pseudoeurophioides WIN(M)42 (EU879132)
Grosmannia penicillata NFRI60-21 (AY935623)
Clade III
Pesotum sp. WIN(M)481 (EU879133)
1.00
Grosmannia davidsonii WIN(M)1495 (EU879134)
Clade X
0.1

3.3.5. Possible constraints involved in conservation of RNA structure

For detailed structural analysis of the major helix of ITS1 and helices I to IV of ITS2, ten strains were selected in order to highlight unusual sequence/structural features. One of these strains, *Pesotum* sp. strain WIN(M)481 which of the ingroup sequences was the most distantly-related to the remaining sequences in the data set (Figure 3.3), was selected as the reference strain from which structural comparisons were made.

Based on the RNA modeling the following observations were made: (i) variability in the length of the major helix of ITS1 could be the result of indels and possible RNA strand slippage events (Figures 3.1 and 3.6A); (ii) CBCs and hemi-CBCs maintained helical structure in ITS segments and also supported the proposed models for the major hairpin stem in ITS1; and (iii) indels were accommodated by corresponding indels on the opposite site of the helix and/or by potential slippage of the RNA strand to reform the hairpin (Figure 3.6). The terminal hairpin loop of ITS1 was a hot spot for slippage events: its size contracted and expanded, and nucleotides from the 3' side of the helix may have potentially slid over the top loop and positioned themselves on the 5' side to base pair with neighbouring nucleotides (Figure 3.6A). Moreover, strand slippage appeared to be more common on the 3' side of the major helix. Near the base of the helix, deletions were typically compensatory on both sides of the hairpin, whereas at the top indels, as well as substitutions in some sequences, may also have been accommodated by potential RNA strand slippage.

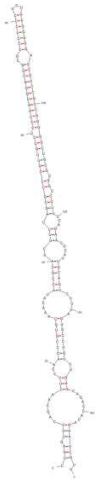
A noteworthy exception to the observation of compensatory/slippage events balancing indels was found in the unusually short ITS1 sequence of *L. procerum* strain NFRI59-84/2. As described above, there was a deletion of 22 nt on the 3' side of the major helix near its midpoint, while the sequence on the 5' side was apparently

intact. These unpaired nucleotides could potentially interact with each other to form a short lateral helix (Figure 3.6A).

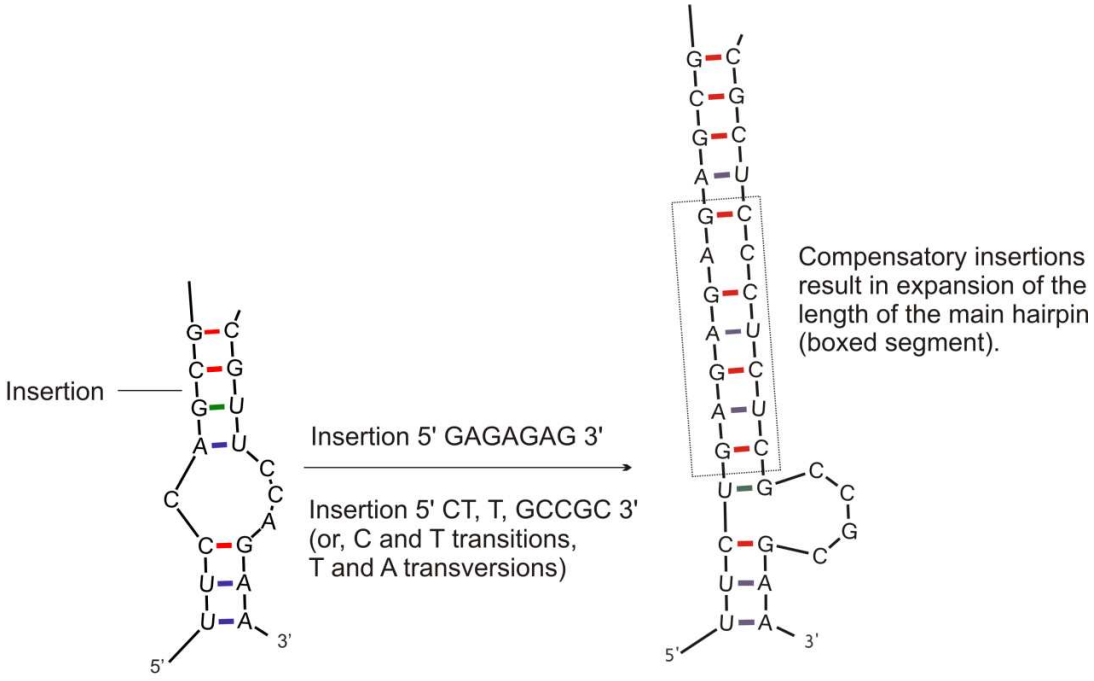
The extended hairpin in ITS1 contained a plethora of so-called mutually informative sites; that is, sites which are involved in base pairing in the absence of nucleotide conservation. As shown in Figure 3.1B, mutually informative sites were found at the base of the major helix in the GA/CT stretch, near the midpoint where there are polyGC tracts on both sides of the helix, and in the G-rich region on the 3' side of the hairpin, near the tip. The proximal region of the helix also contained numerous sites with CBCs and hemi-CBCs.

In ITS2, mutually informative sites were identified in all four helices (Figure 3.2B). The final two nucleotide positions (positions 118 to 119) of helix II were dynamic sites: G to T transversion of the penultimate nucleotide in the stem was potentially accompanied by slippage of the RNA strand in the central loop into helix II to base pair with the first G (position 39). In helix III, the nucleotides on the 5' side of the stem were more conserved than those on the opposing side of the stem. The 3' terminal was the most variable region of the stem and one of the main sites of possible strand slippage events that led to the expansion or contraction of both helix IV and the unpaired nucleotides in the loop between helices III and IV.

Figure 3.6. Hypothesized shifts from one ITS RNA structure to another as a result of compensatory base changes, insertions and deletions, and potential RNA strand slippage. Hypothesized shifts from one structure to another, due to indels, CBCs, and possible RNA strand slippage are indicated by boxes and arrows. (A) Proposed ITS1 rRNA secondary structure model for *Pesotum* sp. WIN(M) 481. Variability in the length of the major helix of ITS1 is caused by indels and putative RNA strand slippage. Helical regions may be maintained by compensating indels, CBCs, hemi-CBCs, and possible slippage of the RNA strand. The 5' and 3' sides of the hairpin at the proximal and distal sections are potential hotspots for expansion/contraction of sequence motifs. (B) Proposed ITS2 rRNA secondary structure model for *Pesotum* sp. WIN(M) 481. Substitution, insertion, and deletion events in helices I to IV of ITS2 may be accommodated by compensatory and slippage events that maintain RNA secondary structure. The 3' terminus at the base of helix III is hypervariable and is one of the main sites of potential RNA strand slippage. Helix IV is variable in length and sequence and variation is influenced by RNA strand slippage events occurring in helix III and in the single-stranded junction between helices III and IV.

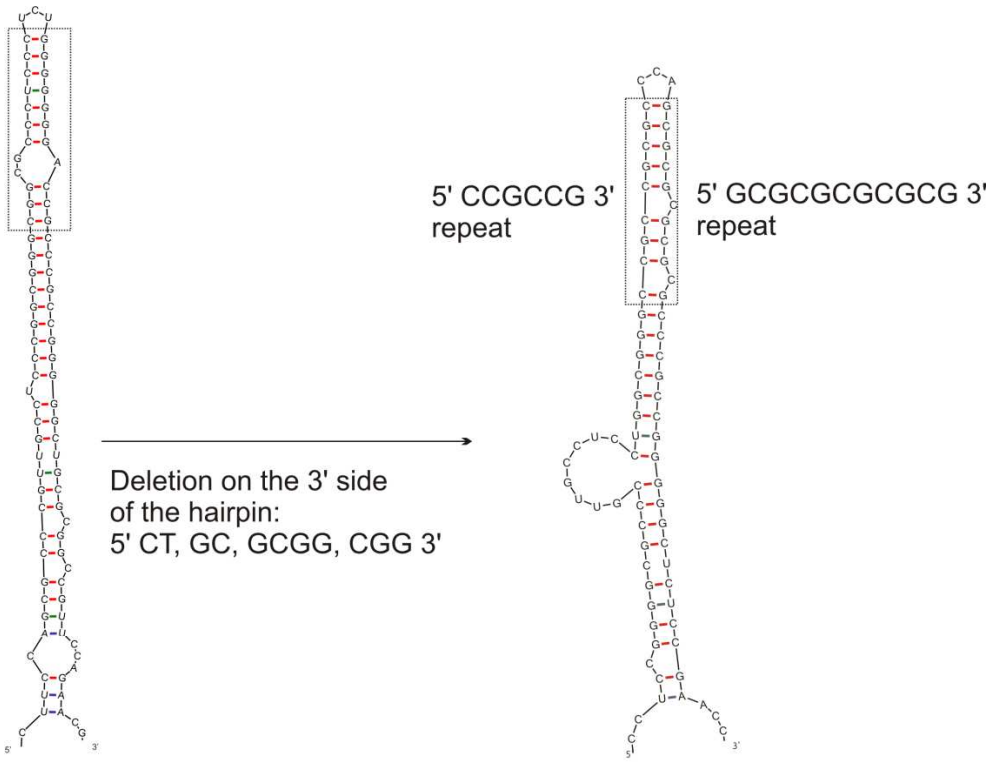


Model of RNA secondary structure of the ITS1 segment in *Pesotum* sp. WIN(M)481.



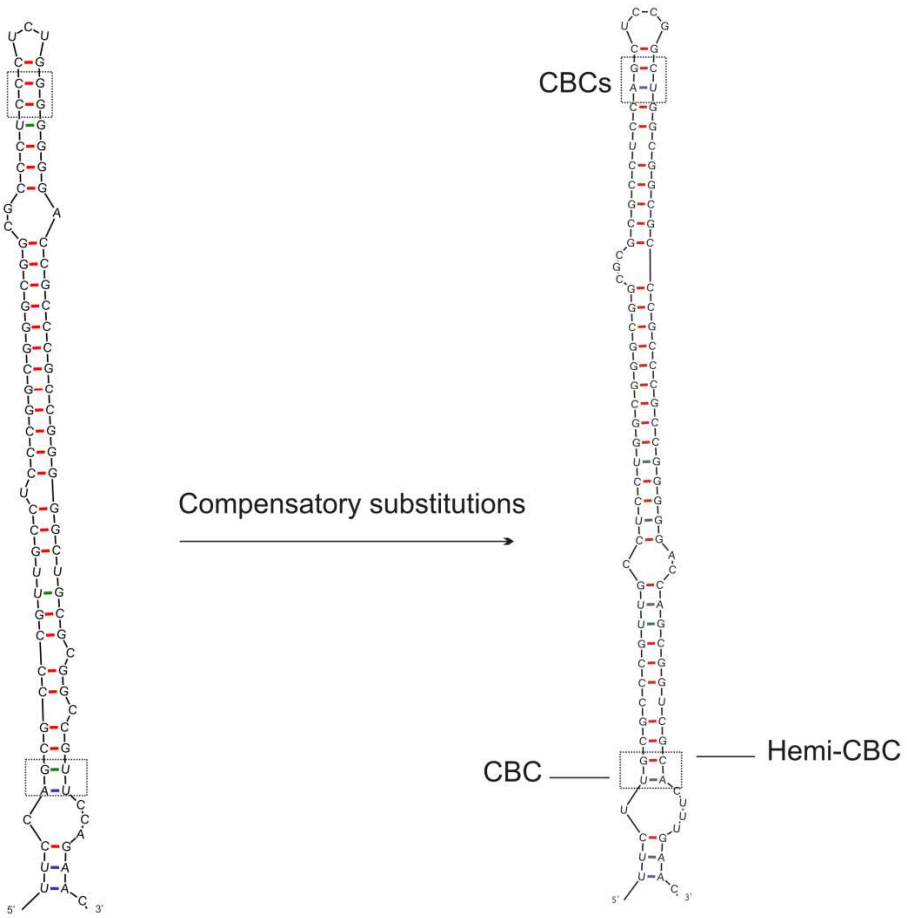
Pesotum sp. WIN(M)481 *L.wingfieldii* TOM10.2

A



Pesotum sp. WIN(M)481

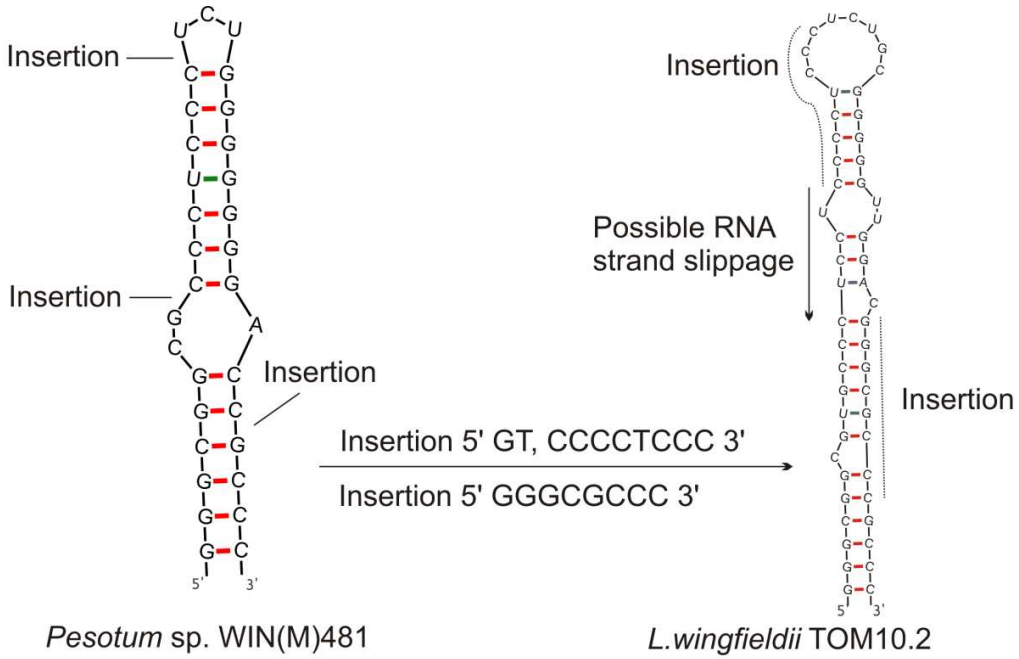
H. pini WIN(M)82-89



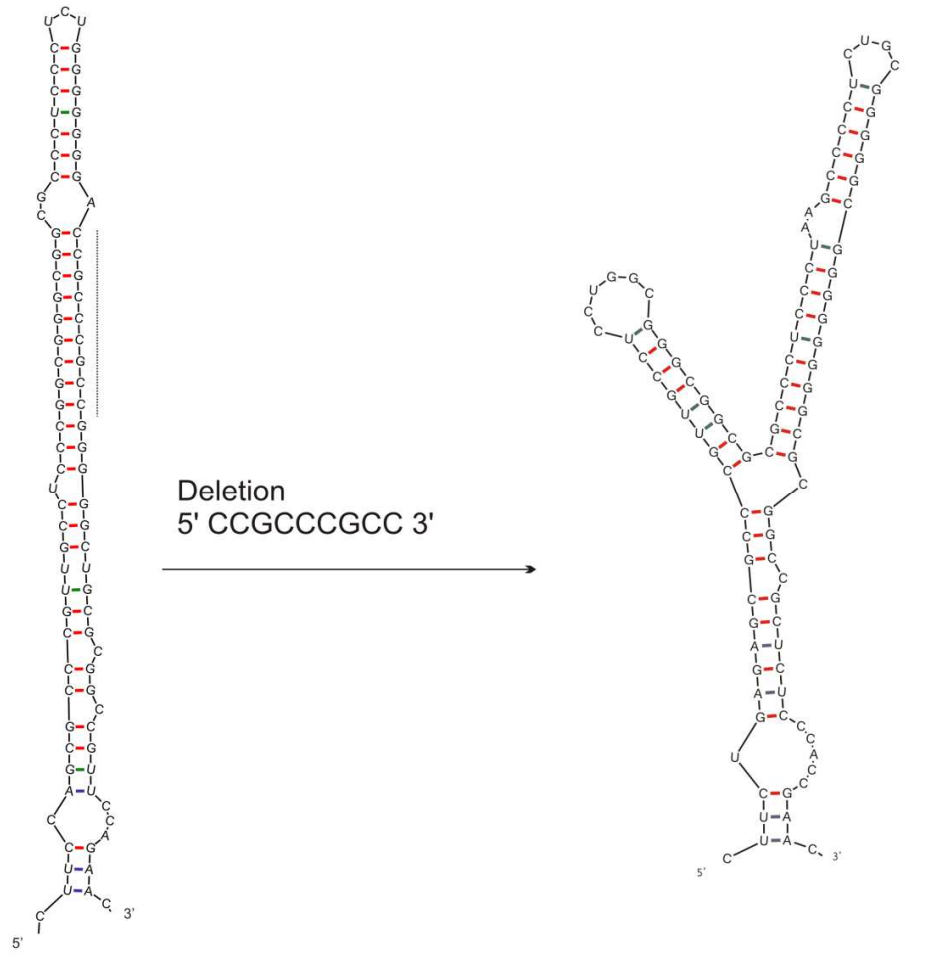
Pesotum sp. WIN(M)481

G. penicillata WIN(M)131

A (continued)

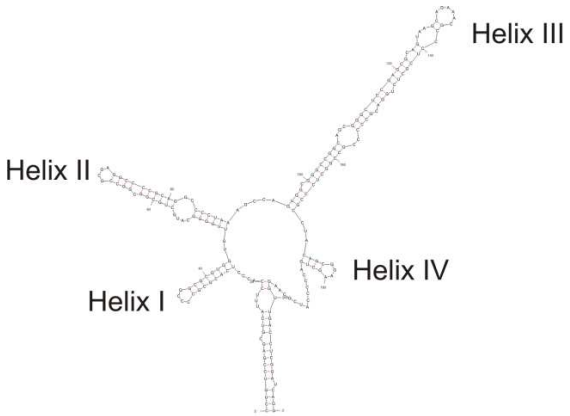


Pesotum sp. WIN(M)481 *L.wingfieldii* TOM10.2

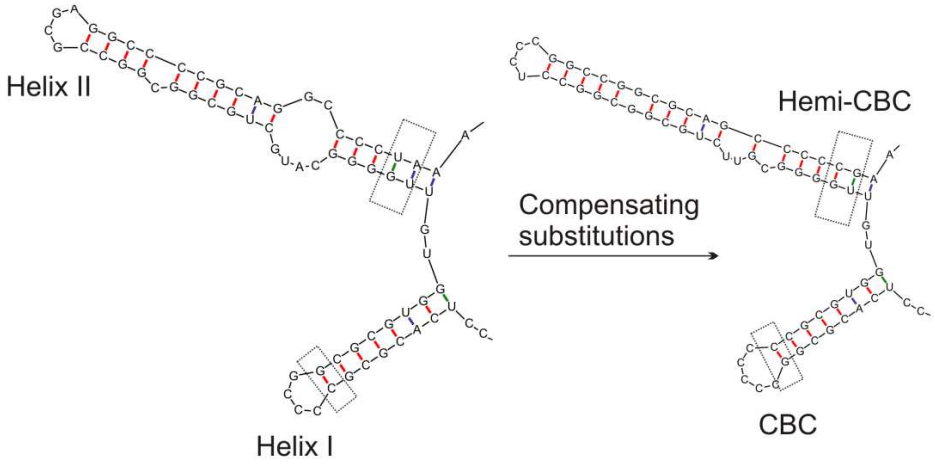


Pesotum sp. WIN(M)481 *L. procerum* NFRI59-84/2

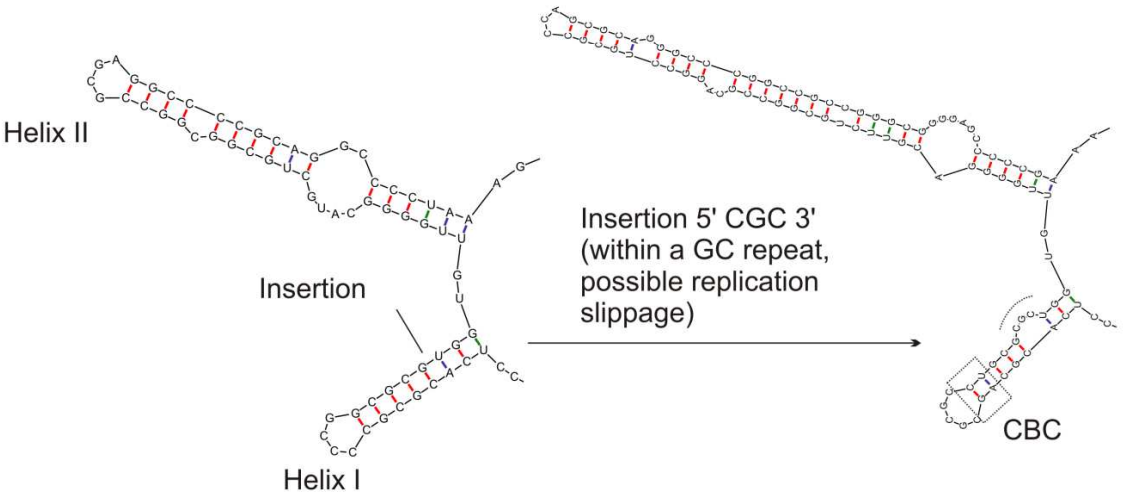
A (continued)



Model of RNA secondary structure of the ITS2 segment in *Pesotum* sp. WIN(M)481.

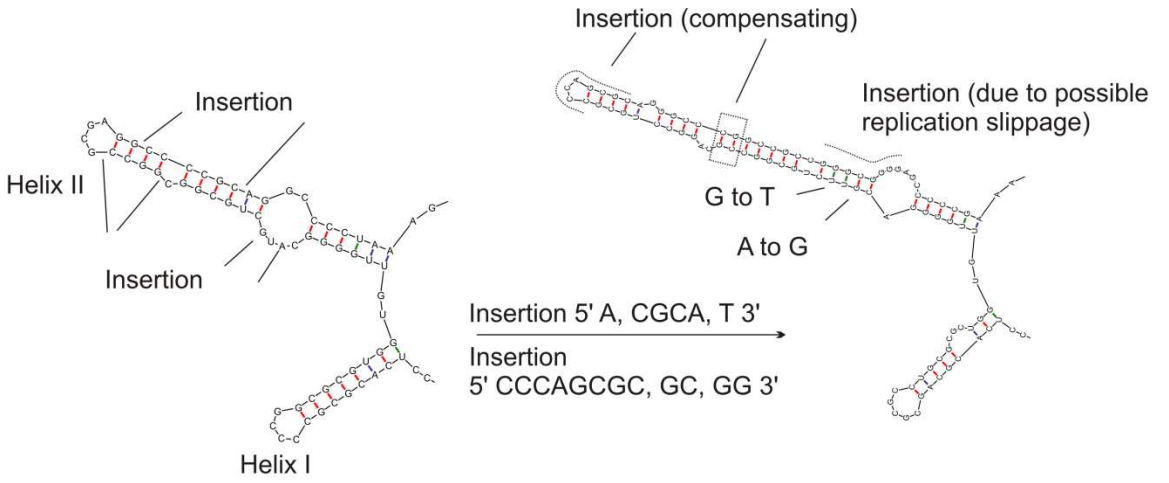


Pesotum sp. WIN(M)481 *H. pini* WIN(M)82-89



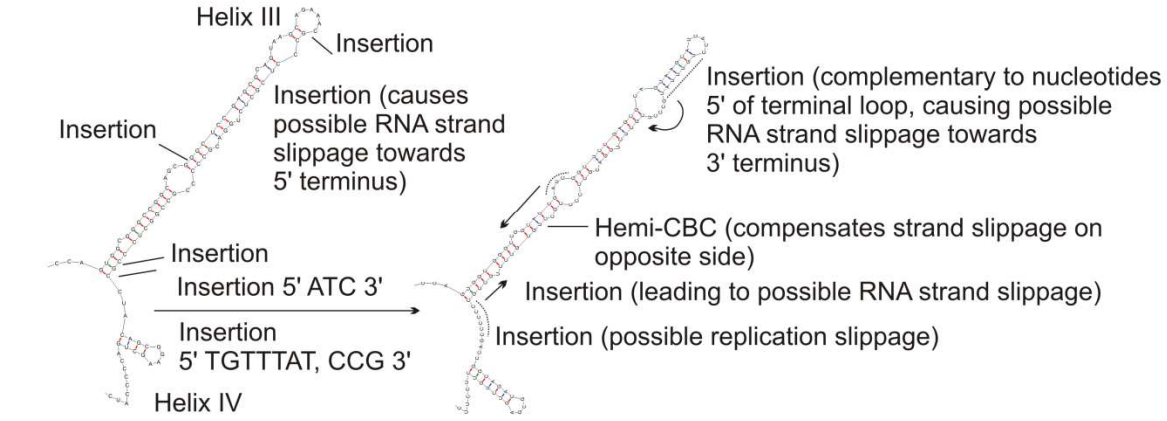
Pesotum sp. WIN(M)481 *G. piceaperda* WIN(M)1380

B



Pesotum sp. WIN(M)481

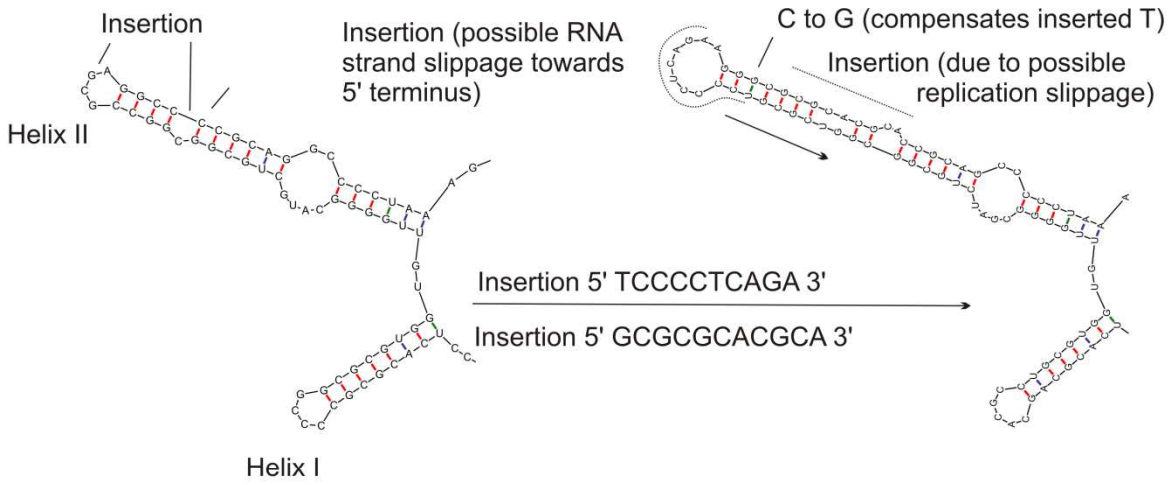
G. piceaperda WIN(M)1380



Pesotum sp. WIN(M)481

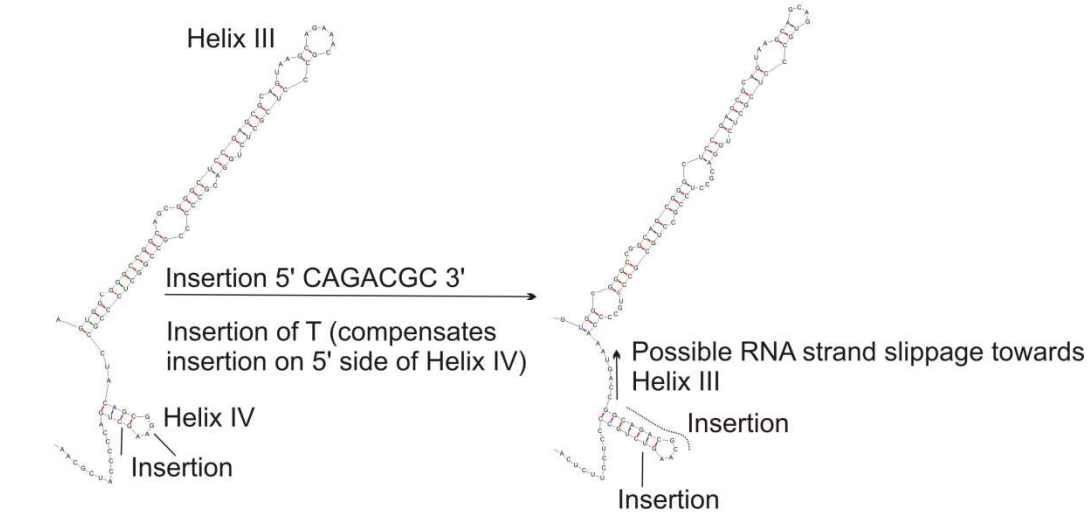
G. piceaperda WIN(M)1380

B (continued)



Pesotum sp. WIN(M)481

Leptographium sp.WIN(M)984



Pesotum sp. WIN(M)481

L. procerum NFRI59-84/2

B (continued)

3.4. Discussion

3.4.1. Mechanisms involved in the conservation of the RNA secondary structures of ITS segments

One goal of this study was to identify the potential effects on RNA secondary structure of CBCs, hemi-CBCs, and indels. For the purpose of identifying potential RNA structural constraints that may influence the evolution of ITS sequences, the following questions were posed: (i) What are the effects on the RNA secondary structure of nucleotide substitutions and indels at the DNA level? and (ii) Do sequence changes disrupt base pairing (and to what extent), or are there compensatory changes to maintain the overall structure (compensating indels, CBCs, hemi-CBCs, or possible RNA strand slippage that restore hairpin structures)? Analysis of variations in the DNA sequence and RNA secondary structure may be useful for elucidating the evolutionary pressures on ITS segments.

Mono- and di-nucleotide repeats are viewed as potential markers for replication slippage (Levinson and Gutman, 1987), particularly if the duplication or deletion events occurred 5' to the existing repeat. Analysis of the sequence alignment for each ITS sequence (excluding the phylogenetic outgroups) with the Weblogo program (Crooks et al., 2004) revealed as conserved motifs several long G and C mono-nucleotide repeats, suggesting that replication slippage may be an important mechanism for generating ITS sequence diversity. It has been shown that in variable regions of the small subunit (SSU) gene replication slippage occurs frequently (Hancock and Dover, 1990; Hancock and Vogler, 2000). It was noted that within the SSU rDNA, indels generated by slippage frequently result in compensatory events (compensatory slippage or point mutations) that allow RNA secondary structures to

be maintained. It appears that in this study similar consequences for slippage-derived sequences within ITS1 and ITS2 were uncovered. This raises an important issue, as it has been suggested that slippage-derived sequences tend to be self-complementary and, thus, at the RNA level stem-loop structures could be self-organizing as a consequence of random replication slippage events and point mutations (Hancock and Vogler, 2000). There is the potential risk that similar ITS RNA structures are observed in a wide range of organisms due to convergent evolution; that is, the structures are the result of the underlying mutational mechanisms implied by the patterns of nucleotide changes observed in this study and not necessarily due to functional constraints.

3.4.2. GC balance between ITS1 and ITS2

Previously, Hausner and Wang (2005) suggested that molecular co-evolution exists between ITS1 and ITS2 with respect to the length of these sequence elements, although no strong functional basis for this observation was proposed. Intraspecific ITS variability within the fungi was recently assessed by Nilsson et al. (2008) and they noted that variation in the ITS1 and ITS2 regions were highly correlated, suggesting the two regions do not evolve independently. Torres et al. (1990) observed that amongst many eukaryotes a GC balance appears to exist between ITS1 and ITS2 such that the mol % G+C content of ITS1 is nearly identical to that of ITS2, irrespective of the GC content of the corresponding 5.8 S gene. This phenomenon has been observed in many plants (Jobst et al., 1998) and it was also observed in the fungal ITS sequences in this study that there is a GC balance and GC bias in the rDNA ITS regions compared to the 5.8 S gene sequence. The ITS2 segments in these fungal strains were unusually GC-rich compared to green algal strains belonging to

the Volvocales, whose mol % G+C ranges from 35.7 for *Pleodorina japonica* to 58.6 for *Chlamydomonas callosa* (Mai and Coleman, 1997). The size and GC contents of both ITS regions in these strains are comparable to those described by Nazar and co-workers (1987) for the thermophilic fungus *Thermomyces lanuginosus.* In this species, the length of ITS1 is 208 nt and its G+C is 65.4 %, and the size of ITS2 is 170 nt, with a G+C of 68.2 %. The authors attributed the unusually high GC content as an adaptation to the elevated temperature of the environment. Within the ITS region of the powdery mildews (oomycetes) a GC balance was observed between the two ITS regions and a positive correlation was noted between GC content and ITS length; however, no explanations could be provided (Takamatsu et al., 1998). A GC balance between ITS1 and ITS2 (71.84 % and 73.25 %, respectively) was also documented in species of cacti belonging to the genus *Mammillaria*, family *Cactaceae* (Harpke and Peterson, 2006). The fungal strains examined in this particular work originate from the Northern Hemisphere and temperate zones and are mesophilic, growing poorly at temperatures above 25 ºC (Upadhyay, 1981). Thus, it is likely that the high GC content is not an adaptation to environmental temperature in the fungal strains examined here, and there may be an alternative or additional evolutionary pressure.

It is not apparent what types of mechanisms would promote the molecular co-evolution that has been noted within the fungal ITS1 and ITS2 segments in terms of length and GC balance; however, structural constraints might set the limits on how ITS sequences evolve. Data from both biochemical studies and comparative sequence analyses of the ITS regions indicate that ITS RNA transcripts form extended helical structures, which are essential to their biological function (Lalev and Nazar, 1999; Côté and Peculis, 2001; Mai and Coleman, 1997; Schultz et al., 2005; Wolf et al.,

2005; Keller et al., 2009). The ITS1 region appears to act as a "biological spring" to bring into close proximity the appropriate termini of 18S, 5.8S, and ITS segments during processing of the pre-rRNA precursor (Nazar et al., 1987; Lalev and Nazar, 1998), and this functional constraint may provide sufficient evolutionary pressure to conserve base pairing. GC-rich segments may have a role in forming and/or stabilizing the extensive base pairing of helical regions in the RNA secondary structure such as the central extended hairpin of ITS1 and the three to four helices radiating from the "palm" structural model for ITS2 (Mai and Coleman, 1997; Schultz et al., 2005; Wolf et al., 2005) and the base pairing between the 5' terminus of ITS2 and the 3' terminus of the LSU (Lalev and Nazar, 1999). These observations suggest that the preservation of RNA secondary and/or tertiary structure, and, ultimately, biological function, is a major constraint acting on the evolution of ITS sequences.

Future efforts in elucidating the mechanisms that allow for evolutionary changes and constraints that limit change and possible interdependence of ITS1 and ITS2 sequences amongst the fungi would benefit from comparative analysis of ITS region sequences from other fungal groups such as the well-characterized members of the genus *Saccharomyces*.

3.4.3. The use of the ITS region as a phylogenetic marker

A second objective of this study was to assess both the level of phylogenetic resolution and effects on tree topology obtained from the analysis of the entire region, including the highly-conserved 5.8S region that separates ITS1 and ITS2, and of each segment individually. The latter goal stems from the wide-spread application of the ITS region as a marker in phylogenetic and taxonomic studies. It was observed that the number of strains distinguished using ITS1 sequences was greater than when ITS2

sequences were used (compare Figures 3.4 and 3.5), and analysis of just the latter region (that is, excluding ITS1 and 5.8S sequences) resulted in phylogenetic trees with numerous unresolved polytomies. Within the *Hypocreales*, specifically the *Gibberella fujikuroi* complex, the ITS regions, when analyzed individually, performed poorly, and again the ITS2 segment was noted to be less informative when compared to the ITS1 region (Lieckfeldt and Seifert, 2000).

The results of the phylogenetic analyses indicate that polytomies in the phylogenetic tree are reduced when the phylogenetic analysis incorporates the entire ITS-5.8S region rather than either segment on its own (compare Figures 3.3 to 3.5) and the distinction between strains and species of *Leptographium*/*Grosmannia* is better represented by sequence analyses that incorporate the entire ITS-5.8S region. Moreover, the ITS phylogenetic analysis does indeed confirm that *Leptographium* species represent mitotic derivatives of *Grosmannia* species (Zipfel et al., 2006).

CHAPTER 4. BIOPROSPECTING NOVEL INSERTIONS WITHIN RIBOSOMAL- AND PROTEIN-CODING GENES IN THE MITOCHONDRIAL DNA OF *LEPTOGRAPHIUM* SPECIES

4.1. Introduction and research objectives

Insertion sequences, such as self-splicing introns and homing endonuclease genes (HEGs), were identified within genes encoding the small and large ribosomal (r) RNA genes (*rns* and *rnl*, respectively), subunits of the cytochrome oxidase (*cox*), and cytochrome b (*cob*) of several species of ascomycetous fungi (Michel et al., 1982). Group I introns are especially abundant in fungal mitochrondrial (mt) DNA (Lang, 1984; Lang et al., 1985; Lambowitz and Belfort, 1993; Gibb and Hausner, 2005). The mt genome of the ascomycete *Podospora anserina* contains thirty-six introns, sixteen of which are found in the *cox1* gene alone. In fact, the large size of this gene, 25 kb, is clearly due to its high intron content (Cummings et al., 1989). In addition to coding for putative homing endonucleases (HEases) some group I introns within the U11 region of the mt *rnl* gene have been shown to encode an essential host ribosomal protein, rps3 (Burke and RajBhandary, 1982; Gibb and Hausner, 2005; Sethuraman et al., 2008, 2009a, 2009b).

In this study, ribosomal- and protein-coding genes in the mt DNA of fifty strains of *Leptographium* were screened for the presence or absence of insertions using a PCR-based approach. The main objectives of the PCR screen were: (i) to identify potentially novel insertion elements in the mt DNA of members of *Leptographium* and (ii) using phylogenetic data of the host organisms, examine if these elements are transmitted vertically or horizontally. Analysis of the variations in the presence/absence of the intron and/or intronic open reading frame (ORF) may

elucidate the evolutionary dynamics of optional genetic elements and provide clues regarding the spread of insertion elements amongst closely-related fungi. The potential evolutionary affiliations between these elements and their host gene/organism and the origins of these elements (vertical/horizontal transmission, random gain/loss events) were examined within the phylogenetic framework established by the ITS study in Chapter 3.

A third goal of the survey was to assess whether or not the presence/absence of insertions could be used as a molecular marker for the rapid identification of *Leptographium* species, particularly potential plant pathogens and members of the *G. aurea-L. wingfieldii-L. terebrantis* species complex.

4.2. Methods overview

The strains of *Leptographium* and of an allied teleomorphic genus, *Grosmannia*, which were submitted to PCR screening, are listed in Table 4.1. The primers used to amplify mt genes encoding the small and large subunit RNAs (*rns* and *rnl*, respectively) and cytochrome b (*cob*) are shown in Figure 2.1. Data (presence or absence of introns and sizes of amplicons) obtained from the PCR screening were superimposed onto the phylogenetic tree of the host organism that was generated using the nuclear ITS-5.8S rDNA region, as described in Chapter 3. Selected representatives of intron-plus and intron-minus alleles were cloned into the TOPO-TA Cloning® Kit for Sequencing (Life Technologies) and plasmid DNA was sequenced using the Big Dye (version 3.1) protocols supplied by the manufacturer (Applied Biosystems, Foster City, USA).

Table 4.1. Strains of *Grosmannia* and *Leptographium* species used in the PCR survey.

Strain[1]	WIN(M)[5]	Substrate	Country of origin
Grosmannia aurea CBS438.69[2]	809	*Pinus contorta*	Canada
Grosmannia penicillata DAOM63691	544	*Picea abies*	Sweden
Grosmannia penicillata NFRI60-21	27	*P. abies*	Sweden
Leptographium sp. J.R.88-194A	1376	*Pinus austriaca* Hoess	New Zealand
Leptographium lundbergii CBS352.29[3]	1115	N/A[6]	Sweden
Leptographium lundbergii DAOM60397	1129	wood	Sweden
Leptographium lundbergii DAOM63692	1197	*Pinus* sp.	Sweden
Leptographium lundbergii DSMZ5010	1194	*Picea* sp.	Germany
Leptographium lundbergii NFRI60-25	68	*Pinus sylvestris*	Sweden
Leptographium lundbergii NFRI69-148	967	*P. sylvestris*	Norway
Leptographium lundbergii NFRI89-1040/1/3	1131	*P. sylvestris*	Sweden
Leptographium lundbergii NFRI1502/1	69	*P. sylvestris*	Norway
Leptographium procerum DAOM33940	1199	*Pinus strobus*	Canada
Leptographium procerum J.R.88-409A	1375	*Pinus* sp.	New Zealand
Leptographium procerum NFRI59-84/2	33	Timber	Norway
Leptographium procerum TOM55.35	1211	*P. sylvestris*	Canada
Leptographium procerum TOM62.30	1210	*P. sylvestris*	Canada
Leptographium procerum TOM73.12	1244	*P. sylvestris*	Canada
Leptographium procerum TOM76.36	1254	*Tomicus piniperda*	Canada
Leptographium procerum TOM76.8	1250	*T. piniperda*	Canada
Leptographium procerum TOM86.19	1272	*P. sylvestris*	Canada
Leptographium procerum UAMH9724	796	*Pinus nigra*	New Zealand
Leptographium terebrantis CBS298.85	1184	N/A	USA
Leptographium terebrantis CBS337.70[2]	1183	Pupal chamber	USA
Leptographium terebrantis CBS408.61	1185	Gymnosperm, wood	Germany
Leptographium terebrantis UAMH9690	662	*Pinus contorta*	Canada
Leptographium terebrantis UAMH9722	468	*P. contorta*	Canada
Leptographium truncatum CBS647.89	1434	*P. sylvestris*	France
Leptographium truncatum CBS929.85[2]	1435	*Pinus taeda*	South Africa
Leptographium truncatum DAOM60396	1128	N/A	Sweden
Leptographium truncatum Forintek C34	254	N/A	Sweden
Leptographium truncatum J.R.88-324	1029	*Pinus* sp.	New Zealand
Leptographium truncatum J.R.88-449	1028	*Pinus radiata*	New Zealand
Leptographium truncatum NFRI59-7/3	660	*P. sylvestris*	Norway
Leptographium truncatum NFRI1813/1	174	N/A	Norway
Leptographium truncatum TOM74.29	1246	*P. sylvestris*	Canada
Leptographium truncatum TOM86.30	1274	*P. sylvestris*	Canada
Leptographium wingfieldii CBS345.90	1182	*T. piniperda*	United Kingdom
Leptographium wingfieldii CBS645.89[2]	1118	*T. piniperda*	France
Leptographium wingfieldii CBS648.89[4]	1181	*Pinus brutia*	Greece
Leptographium wingfieldii MCC125	1123	*Pinus densiflora*	Japan
Leptographium wingfieldii MCC130	1124	*P. densiflora*	Japan
Leptographium wingfieldii MCC349	1126	*T. piniperda*	Japan
Leptographium wingfieldii NFRI88-369/11	1130	*P. sylvestris*	Sweden
Leptographium wingfieldii TOM1.3	1100	*P. sylvestris*	Canada
Leptographium wingfieldii TOM5.1	1192	*P. sylvestris* (trap log)	Canada
Leptographium wingfieldii TOM9.4	1120	*P. sylvestris*	Canada
Leptographium wingfieldii TOM10.2	1103	*P. sylvestris* (trap log)	Canada
Leptographium wingfieldii TOM11.5	1121	*P. sylvestris*	Canada
Leptographium wingfieldii TOM59.21	1209	*P. sylvestris* (trap log)	Canada

[1]J.R., J. Reid; CBS, Centraal Bureau voor Schimmelcultures, Utrecht, The Netherlands; UAMH, University of Alberta Microfungus Collection & Herbarium, Devonian Botanic Garden, Edmonton, AB, Canada, T6G 2E1; ATCC, American Type Culture Collection, Rockville, MD, USA; NFRI, Norwegian Forest Research Institute, AS, Norway; DAOM, Cereal and Oilseeds Research, Agriculture & Agri-Food Canada, Ottawa, Ont., Canada; MCC, culture collection of H. Masuya; TOM, Isolation designation, Canadian Forest Service, Great Lakes Forestry Centre, 1219 Queen St., Sault Ste. Marie, ON, P6A 5M7.

[2] Ex-holotype.

[3] Neotype.

[4] Ex-paratype.

[5] WIN(M), University of Manitoba, Microbiology/Botany (J.R.'s personal collection).

[6] N/A, not available.

4.3. Results

4.3.1. The distribution of insertions within the mitochondrial *rns* gene

The mt *rns* gene region was amplified using primer pair mtsr1 and mtsr2 and the 3' region of the gene was amplified using primer pair Rns-Gp2F1 and Rns-Gp2R1 (Figure 2.1). The sizes of amplicons obtained from the screening are listed in Table 4.2. Insert-minus alleles of the mt *rns* gene produced an amplicon of 1.2 kb in size, while insert-plus alleles produced amplicons that ranged in size from 2.6 to 5 kb, corresponding to inserts of 1.4 to 3.8 kb.

Amplicons of 1.2 kb were observed in all strains of *L. procerum*, indicating that insertions/introns are absent within the mt *rns* gene in members of this taxon. Conversely, amplicons of 3 to 3.5 kb, corresponding to insertions, likely introns with ORFs, of 1.8 to 2.3 kb, were obtained for all strains of *L. lundbergii.* Heteroplasmy; that is, the presence of both intron-minus and intron-plus alleles of the mt *rns* gene, was observed in *L. lundbergii* strains NFRI69-148, NFRI89-1040/1/3, NFRI1502/1, and CBS352.29. The latter two strains share identical sequences in the ITS region (indicated by the "=" sign in Figure 4.1). They are also identical at the ITS sequence level to strains DAOM60397 and DAOM63692, which yielded amplicons of 3 kb and 3.5 kb, respectively.

Amongst members of the *G. aurea-L. wingfieldii-L. terebrantis* species complex, amplicons were observed that ranged in size from 1.2 to 5 kb, and inserts, when present, ranged in size from 1.8 to 3.8 kb. Within strains of *L. wingfieldii*, insertions were absent in all strains isolated from Europe and Asia and most strains isolated from Ontario. In fact, *L. wingfieldii* strains TOM1.3 and TOM9.4, both collected in Ontario, are the sole isolates containing inserts (1.8 kb in size).

Leptographium wingfieldii strains CBS948.89 and TOM9.4, which are identical at the ITS sequence level, were differentiated based on the presence of an insert in the latter strain only. Two strains of *L. terebrantis*, strains CBS298.85 and CBS337.70, isolated from the United States, contained inserts of 1.8 kb, and heteroplasmy (insert-plus and insert-minus alleles) was observed in strain CBS298.85. Inserts were absent in *L. terebrantis* strains CBS408.61, UAMH690, and UAMH9722. Amongst strains of *L. truncatum* all European isolates contained an insertion of 1.8 kb, with the exception of strain CBS647.89, for which a 4-kb amplicon was detected, corresponding to an insert of 2.8 kb. Both isolates from New Zealand, however, lacked an insertion. Amongst the strains isolated from Ontario, strain TOM86.30 contained an insertion of 2.8 kb, while no insert was found in strain TOM74.29. As described in Chapter 5, sequencing of insert-plus and insert-minus alleles indicated that the insert corresponds to a group II intron, containing a putative ORF encoding a putative LAGLIDADG-type HEase (LHEase), rather than the RT-type ORFs typically associated with ORF-containing group II introns. No ORF-less introns, however, were found; that is, the intron and ORF sequence were found together as a composite element in all insert-plus alleles. The biochemistry and evolutionary dynamics of group II introns and LHEG composite elements are described in Chapter 5.

Table 4.2. PCR survey of insertions within the mitochondrial *rns*, *rnl*, and *cob* genes within strains of *Grosmannia* and *Leptographium*.

Strain	PCR product (kb)/Insert present[1]				
	rns	3' region of *rns*	U7 region of *rnl*	U11 region of *rnl*	*cob*
Grosmannia aurea CBS438.69	5/+	nd[2]	0.35/-	1.65/+	1.65/+
Grosmannia penicillata DAOM63691	4/+	2/+	0.35/-	< 3/+	3/+
Grosmannia penicillata NFRI60-21	4/+	3/+	nd	< 3/+	3/+
Leptographium sp. J.R.88-194A	1.2/-	0.2/-	0.35/-	nd	1.65/+
Leptographium lundbergii CBS352.29	3, 1.2/+, -[3]	2/+	0.35/-	< 3, 1.65/+	0.5/-
Leptographium lundbergii DAOM60397	3/+	2/+	0.35/-	1.65/+	3/+
Leptographium lundbergii DAOM63692	3.5/+	2/+	0.35/-	1.65/+	0.5/-
Leptographium lundbergii DSMZ5010	3/+	2/+	0.35/-	1.65/+	0.85, 0.5/-
Leptographium lundbergii NFRI60-25	3/+	2/+	0.35/-	1.65/+	0.5/-
Leptographium lundbergii NFRI69-148	3, 1.2/+, -	2/+	0.35/-	1.65/+	0.5/-
Leptographium lundbergii NFRI89-1040/1/3	3, 1.2/+, -	nd	0.35/-	1.65/+	0.5/-
Leptographium lundbergii NFRI1502/1	3, 1.2/+, -	2/+	0.35/-	1.65/+	0.5/-
Leptographium procerum DAOM33940	1.2/-	0.2/-	0.35/-	1.65/+	1.65/+
Leptographium procerum J.R.88-409A	1.2/-	0.2/-	0.35/-	1.65/+	< 2/+
Leptographium procerum NFRI59-84/2	1.2/-	0.2/-	0.35/-	1.65/+	1.65, 0.5/+, -
Leptographium procerum TOM55.35	1.2/-	0.2/-	0.35/-	1.65/+	1.65/+
Leptographium procerum TOM62.30	1.2/-	0.2/-	0.35/-	< 1.65/+	1.65/+
Leptographium procerum TOM73.12	1.2/-	0.2/-	0.35/-	1.65/+	1.65, 0.5/+, -
Leptographium procerum TOM76.36	1.2/-	0.2/-	0.35/-	1.65/+	1.65/+
Leptographium procerum TOM76.8	1.2/-	0.2/-	0.35/-	1.65/+	1.65/+
Leptographium procerum TOM86.19	1.2/-	0.2/-	nd	1.65/+	nd
Leptographium procerum UAMH9724	1.2/-	0.2/-	0.35/-	1.65/+	< 2/+
Leptographium terebrantis CBS298.85	3/+	nd	0.35/-	1.65/+	nd
Leptographium terebrantis CBS337.70	3, 1.2/+, -	2/+	0.35/-	1.65/+	0.5/-
Leptographium terebrantis CBS408.61	1.2/-	nd	0.35/-	1.65/+	nd
Leptographium terebrantis UAMH9690	1.2/-	0.2/-	0.35/-	1.65/+	5/+
Leptographium terebrantis UAMH9722	1.2/-	0.2/-	0.35/-	1.65/+	5, 1.65, 0.5/+, -
Leptographium truncatum CBS647.89	3/+	2/+	0.35/-	< 3/+	1.65/+
Leptographium truncatum CBS929.85	3, 1.2/+, -	nd	0.35/-	1.65/+	< 3/+
Leptographium truncatum DAOM60396	3/+	2/+	0.35/-	< 3/+	1.65/+
Leptographium truncatum Forintek C34	3/+	2/+	0.35/-	< 3/+	1.65/+

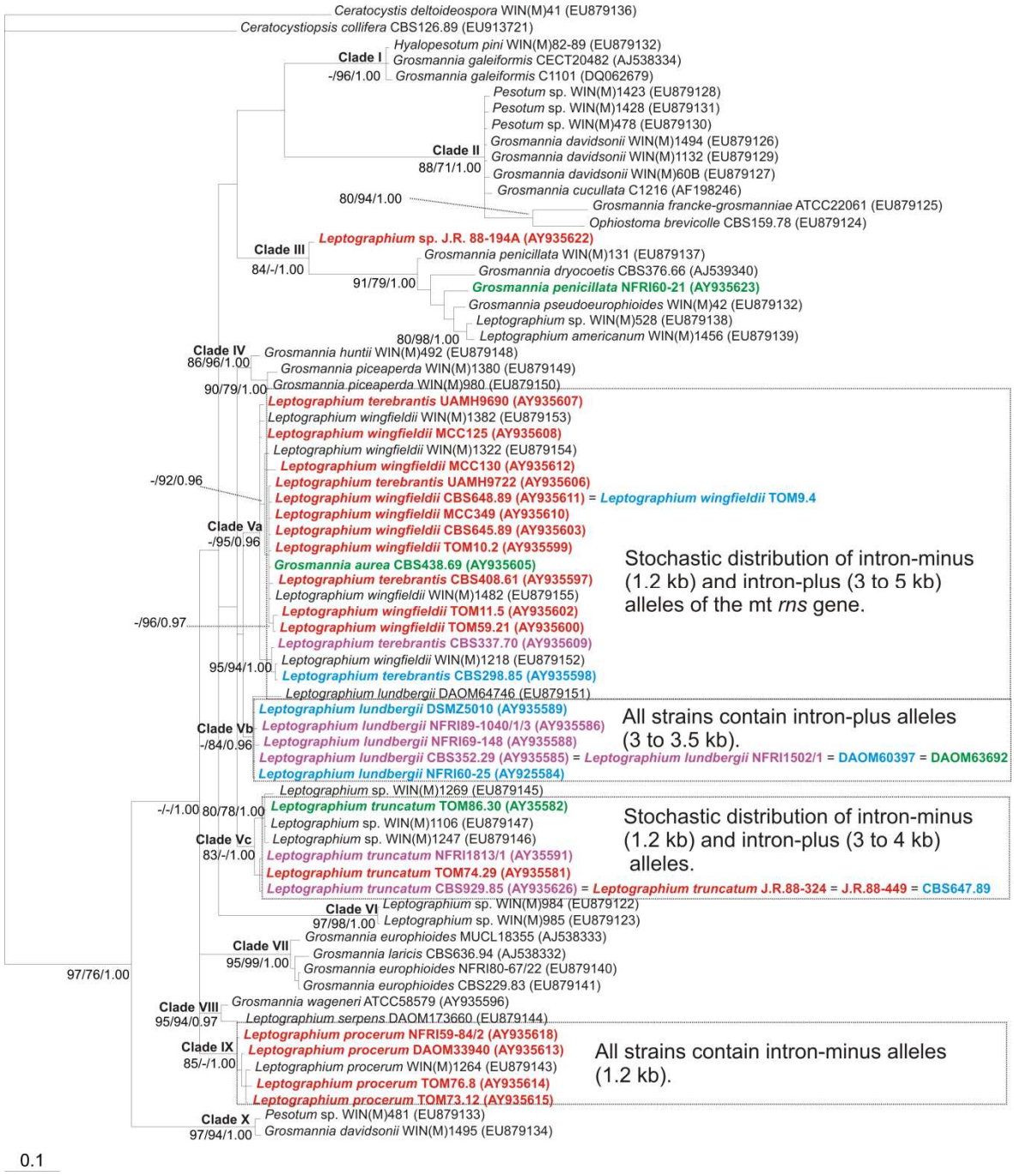
Leptographium truncatum J.R.88-324	1.2/-	0.2/-	0.35/-	< 3/+	1.65/+
Leptographium truncatum J.R.88-449	1.2/-	0.2/-	0.35/-	< 3/+	1.65/+
Leptographium truncatum NFRI59-7/3	3/+	nd	0.35/-	1.65/+	1.65/+
Leptographium truncatum NFRI1813/1	3, 1.2/+, -	nd	0.35/-	< 3/+	1.65/+
Leptographium truncatum TOM74.29	1.2/-	0.2/-	0.35/-	< 3/+	1.65/+
Leptographium truncatum TOM86.30	4/+	2/+	0.35/-	< 3/+	1.65/+
Leptographium wingfieldii CBS345.90	1.2/-	0.2/-	0.35/-	1.65/+	0.5/-
Leptographium wingfieldii CBS645.89	1.2/-	0.2/-	0.35/-	1.65/+	0.5/-
Leptographium wingfieldii CBS648.89	1.2/-	nd	0.35/-	nd	5/+
Leptographium wingfieldii MCC125	1.2/-	0.2/-	0.35/-	1.65/+	0.5/-
Leptographium wingfieldii MCC130	1.2/-	0.2/-	0.35/-	1.65/+	0.5/-
Leptographium wingfieldii MCC349	1.2/-	0.2/-	0.35/-	1.65/+	5/+
Leptographium wingfieldii NFRI88-369/11	1.2/-	0.2/-	0.35/-	1.65/+	nd
Leptographium wingfieldii TOM1.3	3/+	2/+	0.35/-	1.65/+	1.65, 0.5/+, -
Leptographium wingfieldii TOM5.1	1.2/-	nd	0.35/-	1.65/+	5/-
Leptographium wingfieldii TOM9.4	3/+	nd	0.35/-	1.65/+	4/+
Leptographium wingfieldii TOM10.2	1.2/-	0.2/-	0.35/-	1.65/+	5/+
Leptographium wingfieldii TOM11.5	1.2/-	0.2/-	0.35/-	1.65/+	5, 1.65/+
Leptographium wingfieldii TOM59.21	1.2/-	0.2/-	0.35/-	1.65/+	1.65/+

[1] +, insert present; -, insert absent.

[2] nd, not determined.

[3] Both + and – indicate that the strain is heteroplasmic and contains both intron-plus and intron-minus alleles.

Figure 4.1. The distribution of insertions within the mt *rns* gene of strains of *Grosmannia* and *Leptographium*. Intron-minus alleles all yield an amplicon of 1.2 kb in size, while intron-plus alleles produced amplicons ranging in size from 3 to 5 kb. Some strains were heteroplasmic and contained an intron-minus allele (1.2 kb) and intron-plus allele (3 kb). Strains containing insert-minus alleles are indicated in red. Strains with insert-plus alleles of 3 kb are indicated in cyan, while those with amplicons greater than 3 kb (3.5 to 5 kb) are shown in green. Strains that are heteroplasmic (contain insert-minus and insert-plus alleles) are highlighted in purple. Strains that were not included in the screen or for which no amplicon was obtained are shown black; sequences for these strains were obtained from GenBank and from cultures maintained in the WIN(M) collection at the University of Manitoba. The phylogenetic tree is that which is shown in Figure 3.3, modified by including all strains in the PCR screen for which ITS-5.8S sequence data were available. The "=" sign is used to indicate those strains that are identical at the ITS-5.8S sequence level.

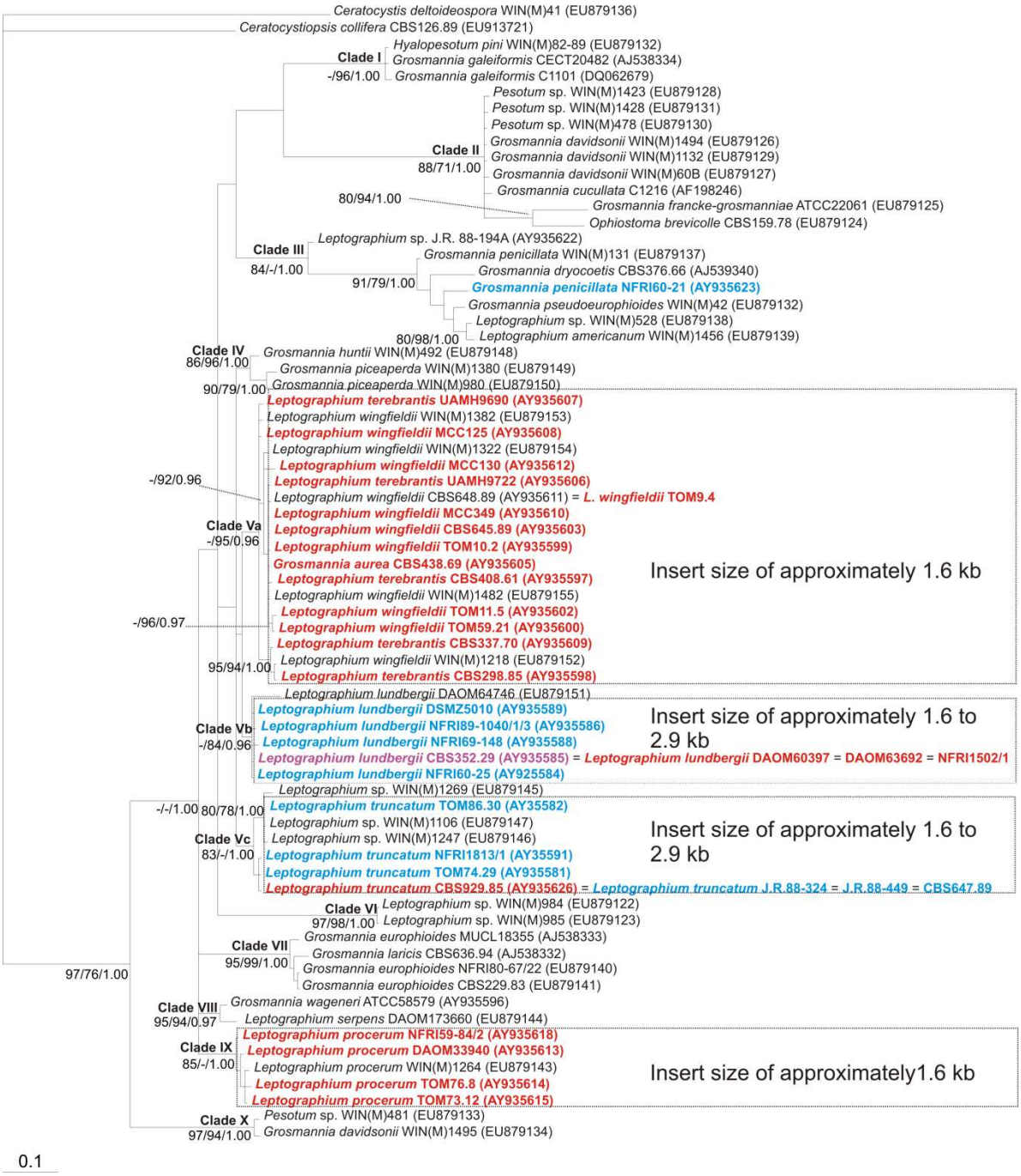

Ceratocystis deltoideospora WIN(M)41 (EU879136)
Ceratocystiopsis collifera CBS126.89 (EU913721)
Clade I
-/96/1.00
Hyalopesotum pini WIN(M)82-89 (EU879132)
Grosmannia galeiformis CECT20482 (AJ538334)
Grosmannia galeiformis C1101 (DQ062679)
Clade II
88/71/1.00
Pesotum sp. WIN(M)1423 (EU879128)
Pesotum sp. WIN(M)1428 (EU879131)
Pesotum sp. WIN(M)478 (EU879130)
Grosmannia davidsonii WIN(M)1494 (EU879126)
Grosmannia davidsonii WIN(M)1132 (EU879129)
Grosmannia davidsonii WIN(M)60B (EU879127)
Grosmannia cucullata C1216 (AF198246)
80/94/1.00
Grosmannia francke-grosmanniae ATCC22061 (EU879125)
Ophiostoma brevicolle CBS159.78 (EU879124)
Clade III
84/-/1.00
Leptographium sp. J.R. 88-194A (AY935622)
Grosmannia penicillata WIN(M)131 (EU879137)
91/79/1.00
Grosmannia dryocoetis CBS376.66 (AJ539340)
Grosmannia penicillata NFRI60-21 (AY935623)
Grosmannia pseudoeurophioides WIN(M)42 (EU879132)
Leptographium sp. WIN(M)528 (EU879138)
80/98/1.00
Leptographium americanum WIN(M)1456 (EU879139)
Clade IV
86/96/1.00
Grosmannia huntii WIN(M)492 (EU879148)
Grosmannia piceaperda WIN(M)1380 (EU879149)
90/79/1.00
Grosmannia piceaperda WIN(M)980 (EU879150)
Leptographium terebrantis UAMH9690 (AY935607)
Leptographium wingfieldii WIN(M)1382 (EU879153)
Leptographium wingfieldii MCC125 (AY935608)
Leptographium wingfieldii WIN(M)1322 (EU879154)
Leptographium wingfieldii MCC130 (AY935612)
-/92/0.96
Leptographium terebrantis UAMH9722 (AY935606)
Leptographium wingfieldii CBS648.89 (AY935611) = Leptographium wingfieldii TOM9.4
Leptographium wingfieldii MCC349 (AY935610)
Clade Va
-/95/0.96
Leptographium wingfieldii CBS645.89 (AY935603)
Leptographium wingfieldii TOM10.2 (AY935599)
Grosmannia aurea CBS438.69 (AY935605)
Leptographium terebrantis CBS408.61 (AY935597)
Leptographium wingfieldii WIN(M)1482 (EU879155)
Leptographium wingfieldii TOM11.5 (AY935602)
-/96/0.97
Leptographium wingfieldii TOM59.21 (AY935600)
Leptographium terebrantis CBS337.70 (AY935609)
95/94/1.00
Leptographium wingfieldii WIN(M)1218 (EU879152)
Leptographium terebrantis CBS298.85 (AY935598)
Leptographium lundbergii DAOM64746 (EU879151)
Stochastic distribution of intron-minus (1.2 kb) and intron-plus (3 to 5 kb) alleles of the mt rns gene.
Leptographium lundbergii DSMZ5010 (AY935589)
Clade Vb
-/84/0.96
Leptographium lundbergii NFRI89-1040/1/3 (AY935586)
Leptographium lundbergii NFRI69-148 (AY935588)
Leptographium lundbergii CBS352.29 (AY935585) = Leptographium lundbergii NFRI1502/1 = DAOM60397 = DAOM63692
Leptographium lundbergii NFRI60-25 (AY925584)
All strains contain intron-plus alleles (3 to 3.5 kb).
Leptographium sp. WIN(M)1269 (EU879145)
-/-/1.00
80/78/1.00
Leptographium truncatum TOM86.30 (AY35582)
Leptographium sp. WIN(M)1106 (EU879147)
Clade Vc
83/-/1.00
Leptographium sp. WIN(M)1247 (EU879146)
Leptographium truncatum NFRI1813/1 (AY35591)
Leptographium truncatum TOM74.29 (AY935581)
Leptographium truncatum CBS929.85 (AY935626) = Leptographium truncatum J.R.88-324 = J.R.88-449 = CBS647.89
Stochastic distribution of intron-minus (1.2 kb) and intron-plus (3 to 4 kb) alleles.
Clade VI
97/98/1.00
Leptographium sp. WIN(M)984 (EU879122)
Leptographium sp. WIN(M)985 (EU879123)
Clade VII
95/99/1.00
Grosmannia europhioides MUCL18355 (AJ538333)
Grosmannia laricis CBS636.94 (AJ538332)
Grosmannia europhioides NFRI80-67/22 (EU879140)
97/76/1.00
Grosmannia europhioides CBS229.83 (EU879141)
Clade VIII
95/94/0.97
Grosmannia wageneri ATCC58579 (AY935596)
Leptographium serpens DAOM173660 (EU879144)
Leptographium procerum NFRI59-84/2 (AY935618)
Clade IX
85/-/1.00
Leptographium procerum DAOM33940 (AY935613)
Leptographium procerum WIN(M)1264 (EU879143)
Leptographium procerum TOM76.8 (AY935614)
Leptographium procerum TOM73.12 (AY935615)
All strains contain intron-minus alleles (1.2 kb).
Clade X
97/94/1.00
Pesotum sp. WIN(M)481 (EU879133)
Grosmannia davidsonii WIN(M)1495 (EU879134)
0.1

4.3.2. The presence/absence of insertions within the U7 and U11 regions of the mitochondrial *rnl* gene

The U7 region of the mt *rnl* gene region was amplified using primer pair Lseux2F and Lseux1R (Figure 2.1). The sizes of amplicons obtained from the screening are listed in Table 4.2. Amplification of the U7 region of the mt *rnl* gene yielded products of 0.35 kb in every tested strain, indicating the absence of inserts in this region. Amplification of the U11 region was carried out using primer pairs IP1 and IP2 (Bell et al., 1996), shown in Figure 2.1. Amplicons ranged in size from 1.65 kb to less than 3 kb in every tested strain (Table 4.2), and the smaller, 1.65 kb, PCR product was detected in the majority of strains.

Data on intron distribution were superimposed onto the phylogenetic tree for the host strains (Figure 4.2). The 1.65 kb product was observed in all strains of *L. procerum*, members of the *G. aurea-L. wingfieldii-L. terebrantis* species complex (regardless of the geographic location from which the strains were isolated), and in all strains of *L. lundbergii*, with the exception of strain CBS352.29. In this particular strain two amplicons of 1.65 kb and less than 3 kb were observed. In eight out of ten strains of *L. truncatum* inserts of approximately 2.4 kb were observed, and in strains CBS929.85 and NFRI59-7/3 inserts of 1 kb were detected instead. It is worth noting that strain CBS929.85 is identical at the ITS sequence level to strains CBS647.89, J.R.88-324, and J.R.88-449.

Figure 4.2. The distribution of insertions within the U11 region of the mt *rnl* gene of strains of *Grosmannia* and *Leptographium*. Strains for which amplicons of 1.65 kb were observed are indicated in red, while those that with amplicons between 1.65 and 3 kb are indicated in cyan. Two amplicons were recovered from *Leptographium lundbergii* strain CBS352.29, indicated in purple: one amplicon was 1.65 kb and the other was less than 3 kb. Strains that were not included in the screen or for which no amplicon was obtained are shown black; sequences for these strains were obtained from GenBank and from cultures maintained in the WIN(M) collection at the University of Manitoba. The phylogenetic tree is that which is shown in Figure 3.3, modified by including all strains in the PCR screen for which ITS-5.8S sequence data were available. The "=" sign is used to indicate those strains that are identical at the ITS-5.8S sequence level.

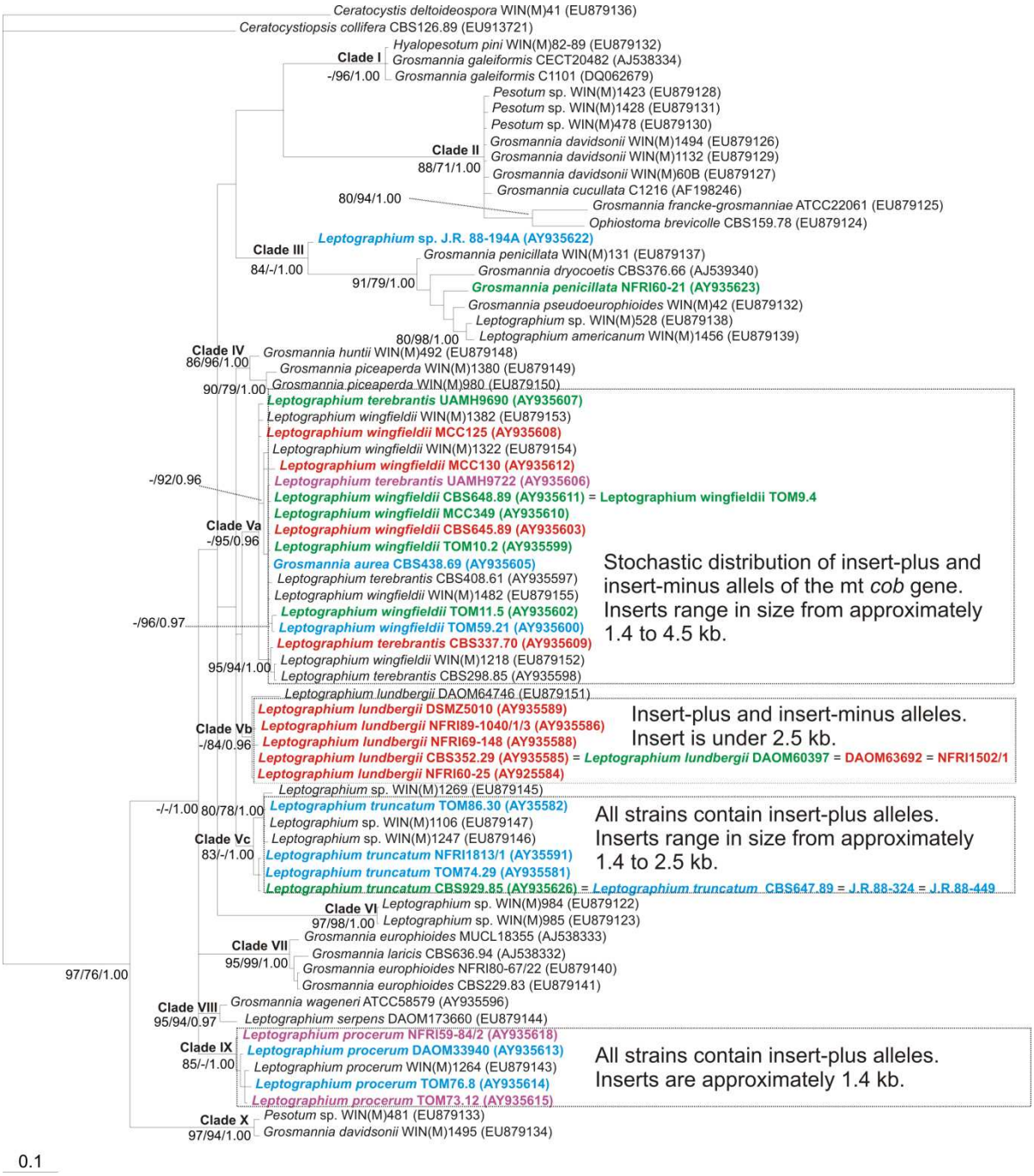

Ceratocystis deltoideospora WIN(M)41 (EU879136)
Ceratocystiopsis collifera CBS126.89 (EU913721)
Clade I
-/96/1.00
Hyalopesotum pini WIN(M)82-89 (EU879132)
Grosmannia galeiformis CECT20482 (AJ538334)
Grosmannia galeiformis C1101 (DQ062679)
Clade II
88/71/1.00
Pesotum sp. WIN(M)1423 (EU879128)
Pesotum sp. WIN(M)1428 (EU879131)
Pesotum sp. WIN(M)478 (EU879130)
Grosmannia davidsonii WIN(M)1494 (EU879126)
Grosmannia davidsonii WIN(M)1132 (EU879129)
Grosmannia davidsonii WIN(M)60B (EU879127)
Grosmannia cucullata C1216 (AF198246)
80/94/1.00
Grosmannia francke-grosmanniae ATCC22061 (EU879125)
Ophiostoma brevicolle CBS159.78 (EU879124)
Clade III
84/-/1.00
Leptographium sp. J.R. 88-194A (AY935622)
Grosmannia penicillata WIN(M)131 (EU879137)
91/79/1.00
Grosmannia dryocoetis CBS376.66 (AJ539340)
Grosmannia penicillata NFRI60-21 (AY935623)
Grosmannia pseudoeurophioides WIN(M)42 (EU879132)
Leptographium sp. WIN(M)528 (EU879138)
80/98/1.00
Leptographium americanum WIN(M)1456 (EU879139)
Clade IV
86/96/1.00
Grosmannia huntii WIN(M)492 (EU879148)
Grosmannia piceaperda WIN(M)1380 (EU879149)
90/79/1.00
Grosmannia piceaperda WIN(M)980 (EU879150)
Leptographium terebrantis UAMH9690 (AY935607)
Leptographium wingfieldii WIN(M)1382 (EU879153)
Leptographium wingfieldii MCC125 (AY935608)
Leptographium wingfieldii WIN(M)1322 (EU879154)
Leptographium wingfieldii MCC130 (AY935612)
-/92/0.96
Leptographium terebrantis UAMH9722 (AY935606)
Leptographium wingfieldii CBS648.89 (AY935611) = L. wingfieldii TOM9.4
Leptographium wingfieldii MCC349 (AY935610)
Clade Va
-/95/0.96
Leptographium wingfieldii CBS645.89 (AY935603)
Leptographium wingfieldii TOM10.2 (AY935599)
Grosmannia aurea CBS438.69 (AY935605)
Leptographium terebrantis CBS408.61 (AY935597)
Leptographium wingfieldii WIN(M)1482 (EU879155)
-/96/0.97
Leptographium wingfieldii TOM11.5 (AY935602)
Leptographium wingfieldii TOM59.21 (AY935600)
Leptographium terebrantis CBS337.70 (AY935609)
95/94/1.00
Leptographium wingfieldii WIN(M)1218 (EU879152)
Leptographium terebrantis CBS298.85 (AY935598)
Insert size of approximately 1.6 kb
Leptographium lundbergii DAOM64746 (EU879151)
Leptographium lundbergii DSMZ5010 (AY935589)
Clade Vb
-/84/0.96
Leptographium lundbergii NFRI89-1040/1/3 (AY935586)
Leptographium lundbergii NFRI69-148 (AY935588)
Leptographium lundbergii CBS352.29 (AY935585) = Leptographium lundbergii DAOM60397 = DAOM63692 = NFRI1502/1
Leptographium lundbergii NFRI60-25 (AY925584)
Insert size of approximately 1.6 to 2.9 kb
Leptographium sp. WIN(M)1269 (EU879145)
-/-/1.00
80/78/1.00
Leptographium truncatum TOM86.30 (AY35582)
Leptographium sp. WIN(M)1106 (EU879147)
Clade Vc
83/-/1.00
Leptographium sp. WIN(M)1247 (EU879146)
Leptographium truncatum NFRI1813/1 (AY35591)
Leptographium truncatum TOM74.29 (AY935581)
Leptographium truncatum CBS929.85 (AY935626) = Leptographium truncatum J.R.88-324 = J.R.88-449 = CBS647.89
Insert size of approximately 1.6 to 2.9 kb
Clade VI
97/98/1.00
Leptographium sp. WIN(M)984 (EU879122)
Leptographium sp. WIN(M)985 (EU879123)
Clade VII
95/99/1.00
Grosmannia europhioides MUCL18355 (AJ538333)
Grosmannia laricis CBS636.94 (AJ538332)
Grosmannia europhioides NFRI80-67/22 (EU879140)
Grosmannia europhioides CBS229.83 (EU879141)
97/76/1.00
Clade VIII
95/94/0.97
Grosmannia wageneri ATCC58579 (AY935596)
Leptographium serpens DAOM173660 (EU879144)
Leptographium procerum NFRI59-84/2 (AY935618)
Clade IX
85/-/1.00
Leptographium procerum DAOM33940 (AY935613)
Leptographium procerum WIN(M)1264 (EU879143)
Leptographium procerum TOM76.8 (AY935614)
Leptographium procerum TOM73.12 (AY935615)
Insert size of approximately1.6 kb
Clade X
97/94/1.00
Pesotum sp. WIN(M)481 (EU879133)
Grosmannia davidsonii WIN(M)1495 (EU879134)
0.1

4.3.3. The presence/absence of insertions within the mitochondrial *cob* gene

The mt *cob* gene was amplified using primer pair cybF and cybR (Figure 2.1) and the sizes of the amplicons are listed in Table 4.2. Amplification of the mt *cob* gene yielded PCR products ranging in size from 0.5 kb (corresponding to the insert-minus allele) to 5 kb (insert-plus allele).

In all strains of *L. procerum* amplicons of 1.65 kb were detected, and heteroplasmy in the mt *cob* gene was observed in strains NFRI59-84/2 and TOM73.12 (Figure 4.3). In all strains of *L. truncatum* but one amplicons of 1.65 kb were detected. The exception was strain CBS929.85, for which a PCR product of approximately 3 kb was detected; this strain is identical at the ITS sequence level to strains CBS647.89, J.R.88-324, and J.R.88-449. Amongst strains of *L. lundbergii* insertions were absent in all strains except strain DAOM60397, which contained an insert of 2.5 kb; this strain shares identical ITS sequences with strains CBS352.29, DAOM63692, and NFRI1502/1. Members of the *G. aurea-L. wingfieldii-L. terebrantis* species complex exhibited a stochastic distribution of inserts. Amongst strains of *L. wingfieldii* and *L. terebrantis* amplicons ranged in size from 0.5 to 5 kb. In *L. truncatum* strain UAMH9722 3 amplicons of 0.5, 1.65, and 5 kb were detected, suggesting a variable arrangement of inserts within the mt *cob* gene.

Figure 4.3. The distribution of insertions within the mt *cob* gene of strains of *Grosmannia* and *Leptographium*. Strains containing insert-minus alleles (0.5 kb) are indicated in red. Strains for which inserts of less than 2 kb were observed are indicated in cyan, while those for which inserts between 2 and 5 kb in size were detected are indicated in green. Strains that are heteroplasmic, containing insert-minus (0.5 kb amplicons) and insert-plus (amplicons of 5 and 1.65 kb) alleles are highlighted in purple. Strains that were not included in the screen or for which no amplicon was obtained are shown black; sequences for these strains were obtained from GenBank and from cultures maintained in the WIN(M) collection at the University of Manitoba. The phylogenetic tree is that which is shown in Figure 3.3, modified by including all strains in the PCR screen for which ITS-5.8S sequence data were available. The "=" sign is used to indicate those strains that are identical at the ITS-5.8S sequence level.


Ceratocystis deltoideospora WIN(M)41 (EU879136)
Ceratocystiopsis collifera CBS126.89 (EU913721)
Clade I
-/96/1.00
Hyalopesotum pini WIN(M)82-89 (EU879132)
Grosmannia galeiformis CECT20482 (AJ538334)
Grosmannia galeiformis C1101 (DQ062679)
Clade II
88/71/1.00
Pesotum sp. WIN(M)1423 (EU879128)
Pesotum sp. WIN(M)1428 (EU879131)
Pesotum sp. WIN(M)478 (EU879130)
Grosmannia davidsonii WIN(M)1494 (EU879126)
Grosmannia davidsonii WIN(M)1132 (EU879129)
Grosmannia davidsonii WIN(M)60B (EU879127)
Grosmannia cucullata C1216 (AF198246)
80/94/1.00
Grosmannia francke-grosmanniae ATCC22061 (EU879125)
Ophiostoma brevicolle CBS159.78 (EU879124)
Clade III
84/-/1.00
Leptographium sp. J.R. 88-194A (AY935622)
Grosmannia penicillata WIN(M)131 (EU879137)
91/79/1.00
Grosmannia dryocoetis CBS376.66 (AJ539340)
Grosmannia penicillata NFRI60-21 (AY935623)
Grosmannia pseudoeurophioides WIN(M)42 (EU879132)
Leptographium sp. WIN(M)528 (EU879138)
80/98/1.00
Leptographium americanum WIN(M)1456 (EU879139)
Clade IV
86/96/1.00
Grosmannia huntii WIN(M)492 (EU879148)
Grosmannia piceaperda WIN(M)1380 (EU879149)
90/79/1.00
Grosmannia piceaperda WIN(M)980 (EU879150)
-/92/0.96
Clade Va
-/95/0.96
Leptographium terebrantis UAMH9690 (AY935607)
Leptographium wingfieldii WIN(M)1382 (EU879153)
Leptographium wingfieldii MCC125 (AY935608)
Leptographium wingfieldii WIN(M)1322 (EU879154)
Leptographium wingfieldii MCC130 (AY935612)
Leptographium terebrantis UAMH9722 (AY935606)
Leptographium wingfieldii CBS648.89 (AY935611) = Leptographium wingfieldii TOM9.4
Leptographium wingfieldii MCC349 (AY935610)
Leptographium wingfieldii CBS645.89 (AY935603)
Leptographium wingfieldii TOM10.2 (AY935599)
Grosmannia aurea CBS438.69 (AY935605)
Leptographium terebrantis CBS408.61 (AY935597)
Leptographium wingfieldii WIN(M)1482 (EU879155)
-/96/0.97
Leptographium wingfieldii TOM11.5 (AY935602)
Leptographium wingfieldii TOM59.21 (AY935600)
Leptographium terebrantis CBS337.70 (AY935609)
95/94/1.00
Leptographium wingfieldii WIN(M)1218 (EU879152)
Leptographium terebrantis CBS298.85 (AY935598)
Stochastic distribution of insert-plus and insert-minus alleles of the mt cob gene. Inserts range in size from approximately 1.4 to 4.5 kb.
Leptographium lundbergii DAOM64746 (EU879151)
Clade Vb
-/84/0.96
Leptographium lundbergii DSMZ5010 (AY935589)
Leptographium lundbergii NFRI89-1040/1/3 (AY935586)
Leptographium lundbergii NFRI69-148 (AY935588)
Leptographium lundbergii CBS352.29 (AY935585) = Leptographium lundbergii DAOM60397 = DAOM63692 = NFRI1502/1
Leptographium lundbergii NFRI60-25 (AY925584)
Insert-plus and insert-minus alleles. Insert is under 2.5 kb.
Leptographium sp. WIN(M)1269 (EU879145)
-/-/1.00
80/78/1.00
Leptographium truncatum TOM86.30 (AY35582)
Leptographium sp. WIN(M)1106 (EU879147)
Clade Vc
83/-/1.00
Leptographium sp. WIN(M)1247 (EU879146)
Leptographium truncatum NFRI1813/1 (AY35591)
Leptographium truncatum TOM74.29 (AY935581)
Leptographium truncatum CBS929.85 (AY935626) = Leptographium truncatum CBS647.89 = J.R.88-324 = J.R.88-449
All strains contain insert-plus alleles. Inserts range in size from approximately 1.4 to 2.5 kb.
Clade VI
97/98/1.00
Leptographium sp. WIN(M)984 (EU879122)
Leptographium sp. WIN(M)985 (EU879123)
Clade VII
95/99/1.00
Grosmannia europhioides MUCL18355 (AJ538333)
Grosmannia laricis CBS636.94 (AJ538332)
Grosmannia europhioides NFRI80-67/22 (EU879140)
Grosmannia europhioides CBS229.83 (EU879141)
97/76/1.00
Clade VIII
95/94/0.97
Grosmannia wageneri ATCC58579 (AY935596)
Leptographium serpens DAOM173660 (EU879144)
Clade IX
85/-/1.00
Leptographium procerum NFRI59-84/2 (AY935618)
Leptographium procerum DAOM33940 (AY935613)
Leptographium procerum WIN(M)1264 (EU879143)
Leptographium procerum TOM76.8 (AY935614)
Leptographium procerum TOM73.12 (AY935615)
All strains contain insert-plus alleles. Inserts are approximately 1.4 kb.
Clade X
97/94/1.00
Pesotum sp. WIN(M)481 (EU879133)
Grosmannia davidsonii WIN(M)1495 (EU879134)
0.1

4.4. Discussion

4.4.1. Distribution of insertion elements within mitochondrial genes in *Leptographium* species

Amongst strains of *L. truncatum* all European isolates contained an insertion of 1.8 to 2.8 kb in the mt *rns* gene, yet both strains from New Zealand lacked an insertion. Amongst the strains isolated from Ontario, strain TOM86.30 contained an insertion of 2.8 kb, while no insert was found in TOM74.29. Thus, while the geographical isolation of New Zealand from Europe and North America may have prevented genetic insertions in the mt *rns* gene, geography does not appear to be a factor amongst North American strains. It is possible that in the case of isolates from New Zealand, no donours were available for initiating the spread of this intron within the mt *rns* gene.

The observation that the intron/LHEG composite element was present in the mt *rns* gene of all strains of *L. lundbergii* suggests that the element was transmitted vertically from parent to progeny. The stochastic distribution amongst strains of *L. truncatum* and amongst members of the *G. aurea-L. wingfieldii-L. terebrantis* species complex suggests that the element may be transferred laterally between strains and species, although one cannot ignore the possibility that the random presence and absence is due to rapid and random gain or loss of the element. Intron loss can occur during homologous recombination between intron-plus alleles and a cDNA copy of the spliced mRNA (reviewed in Belfort and Lambowitz, 1993). Introns may be present or absent amongst closely-related strains, such as members of *L. truncatum*. Heteroplasmy, or intron-plus and intron-minus states within an organism, may also occur, and this could be a by-product of the multi-copy state of mt DNA within each

mitochondrion. In addition, heteroplasmy could result from horizontal transmission events. Lateral gene transfer of optional genetic elements, such as introns and intronic ORFs, is facilitated by the pipeline-like state of fungdue to the presence of pore structures in hyphal septae, which allow for the movement of cytoplasm, including nuclei and organelles, from different hyphal compartments. If hyphae from different strains contact and fuse, which occurs during hyphal anastomosis, mitochondria can also fuse, allowing for a mixing of genomes and possible recombination between intron-plus and intron-minus alleles (reviewed in Hausner, 2003; Okamoto and Shaw, 2005).

Insertions within the U11 region of the mt *rnl* gene were found in all tested strains and the size variability, ranging from aproximately 1.6 to 2.9 kb, suggests that this region is a hot spot for invasion by different types of introns and/or putative HEGs. Interestingly, a recent study examining the arrangement of group I introns and intron ORFs in the U11 region in members of *Ophiostoma* identified a mosaic, gene-within-gene organization involving LHEGs and *rps3* genes (Sethuraman et al., 2009a, 2009b). The authors identified putative LHEGs inserted into N- and C-terminal regions of the *rps3* gene, encoded within the mL2449 group I intron and the insertion of the LHE ORF displaced a segment of the resident *rps3* sequence, while the LHEG became fused in-frame to the remaining *rps3* gene. This event compensated for the displacement of the host ORF sequence, thus, potentially regenerating a full-length rps3 ORF.

In all members of *L. procerum* and of the *G. aurea-L. wingfieldii-L. terebrantis* speceis complex, the 1.6-kb insert was detected, suggesting that this element may have become fixed in those populations and is inherited vertically. In contrast, insertions up to approximately 2.9 kb were observed in strains of *L.*

truncatum and *L. lundbergii*, and both types of inserts were detected in *L. lundbergii* strain CBS352.29. While transmission of optional genetic elements may be promoted by the homing activity of HEases, the group I intron inserted within the U11 region may have become fixed in the population by coding for an essential host protein.

The mt *cob* gene appears to have a mosaic arrangement and the observation of both heteroplasmy and stochastic distribution of insertions within taxa is strongly suggestive of horizontal transmission, likely involving several different introns and/or HEGs. Preliminary sequence analysis indicates that this region has been invaded multiple times by group I introns and putative HEGs, such as GIY...YIG- and LAGLIDADG-type HEGs.

4.4.2. Evaluation of the use of mitochondrial insertions as a DNA marker for species identification

The third objective of the screen was to assess whether or not the presence/absence of insertions could be used as a molecular marker for the rapid identification of potential plant pathogens. The mixed distribution of insertions within the mt *rns*, *rnl*, and *cob* genes suggests that there is intra- and inter-species transfer of these elements, either due to horizontal transmission or rapid and random gain/loss. Since intron distribution is not well-correlated to the evolutionary relationships amongst the taxa, optional genetic elements may not be used as a reliable molecular marker.

However, strains that are identical at the ITS sequence level may be distinguished in a limited manner by differences in amplicon size for these gene regions. For example, *L. lundbergii* strains CBS352.29 or NFRI1502/1 are both heteroplasmic for the mt *rns* gene and can be distinguished from strains DAOM60397

and DAOM63692 by both the number and size of amplicons. Likewise, *L. lundbergii* strains CBS648.85 and TOM9.4 are differentiated from each other based on the presence of the insert in the latter strain only. There is a major caveat, however, to using these elements to distinguish between strains: because they may be potentially horizontally transferred, it is possible that environmental samples could show different insert distribution than strains housed in culture collections or that the distribution profile of laboratory strains could change over time with subculturing due to intron loss or mobility events initiated by HEase activity. Thus, it is clear that optional introns cannot be used as reliable, stable molecular markers for the identification of *Leptographium* species.

CHAPTER 5. A BIOCHEMICAL AND EVOLUTIONARY STUDY OF A GROUP II INTRON RIBOZYME AND ITS ENCODED LAGLIDADG HOMING ENDONUCLEASE

Work described in this chapter was submitted for publication to the journal RNA and is in press: Mullineux ST, Costa M, Bassi GS, Michel F, Hausner G. 2010. A group II intron encodes a functional LAGLIDADG homing endonuclease and self-splices under moderate temperature and ionic conditions. *RNA* (In press, RNA/2010/021840).

5.1. Introduction and research objectives

Many group II introns are mobile retroelements that are capable of site-specific insertion into homologous intron-less alleles with the assistance of an intron-encoded protein (IEP, Moran et al., 1995). These mobile group II introns typically contain an open reading frame (ORF) encoding a multifunctional protein with maturase, DNA endonuclease (ENase), and reverse transcriptase (RT) activities in addition to their ribozyme component, which catalyzes splicing (reviewed in Lambowitz and Belfort, 1993; Saldanha et al., 1993; Michel and Ferat, 1995; Lambowitz et al., 1999; Lambowitz and Zimmerly, 2004). Ribozyme-catalyzed splicing via the branching pathway leads to the excision of the intron as a lariat, or branched, molecule with a characteristic 2'-5' phosphodiester bond. More rarely, self-splicing can occur via the hydrolytic pathway, resulting in the excision of the intron as a linear molecule (Daniels et al., 1996; Vogel and Börner, 2002). A number of group II introns in yeast, algae, and bacteria have been shown in fact to catalyze their own removal from primary transcripts (Peebles et al., 1986; Schmelzer and Schweyen,

1986; van der Veen et al., 1986; Schmidt et al., 1990; Ferat and Michel, 1993; Costa et al., 1997a; Robart & Zimmerly, 2005).

The maturase domain of the IEP facilitates efficient intron splicing *in vivo* by stabilizing the catalytically active structure and the ENase and RT domains promote intron mobility. A survey of group II introns and their RT ORFs in bacteria and organelles indicated that the two elements appear to co-evolve (Toor et al., 2001), although the results of a recent large-scale phylogenetic analysis of group II intron and RT sequences suggested that the co-evolution of these elements may be imperfect (Simon et al., 2009). During expression of the host gene the intronic ORF is translated and a ribonucleoprotein particle is formed between the intron lariat and the IEP. In a process referred to as retrohoming, the intron and its ORF colonize typically intron-less cognate alleles through a site-specific DNA integration mechanism known as target DNA-primed reverse transcription (Zimmerly et al., 1995a, 1995b). Retrohoming involves both the intron RNA and the IEP and is completed by host DNA repair mechanisms (Schäfer et al., 2003; Lambowitz and Zimmerly, 2004).

In contrast to their bacterial counterparts, group II introns in organellar genomes are generally inserted within conserved genes encoding rRNA, tRNA, and components of the electron transport chain (reviewed in Saldanha et al., 1993; Michel and Ferat, 1995; Lambowitz and Zimmerly, 2004; Robart and Zimmerly, 2005). The use of insertion sites (IS) that coincide with highly conserved sequence stretches is important for the persistence of these elements in the genome because their elimination would likely require precise deletion so as to not disrupt the host gene.

Also peculiar to organellar genomes was the identification by sequence analysis of several group II introns with ORFs encoding putative LAGLIDADG homing endonucleases (LHEases) in the mitochondrial (mt) small and large subunit

(SSU and LSU, respectively) rRNA genes of fungi belonging to the Ascomycota and Basidiomycota (Michel and Ferat, 1995; Toor and Zimmerly, 2002; Monteiro-Vitorello et al., 2009). However, neither the introns nor the ORFs were characterized functionally in these studies. Genes coding for LHEases are traditionally associated with group I and archaeal introns or with inteins, or they may be present as free-standing ORFs (Dujon, 1980; Dalgaard et al., 1993; Dürrenberger and Rochaix, 1993; Belfort and Roberts, 1997; Jurica and Stoddard, 1999; Chevalier et al., 2005; Gibb and Hausner, 2005; Stoddard, 2006; Sethuraman et al., 2008, 2009a; Bae et al., 2009; Singh et al., 2009). LHEase proteins are named for the LAGLIDADG amino acid α-helical motif that comprises each half of the enzyme's active site. LHEases recognize and bind to long (greater than 20 bp) DNA target sites and exhibit flexibility in sequence recognition (Aagaard et al., 1997; Jurica et al., 1998; Lucas et al., 2001; Moure et al., 2003; Eastberg et al., 2007). LHEases promote homing by generating a double-stranded cut with 4 nucleotide (nt) 3' OH overhangs in DNA; the break is repaired by the host's double-stranded break repair processes using the intron/LHEase gene (LHEG)-containing allele as a template (reviewed in Belfort and Roberts, 1997; Belfort et al., 2002; Stoddard, 2006; Edgell, 2009). In meiotic crosses this can lead to super-Mendelian inheritance of the LHEG/intron sequence. Some LHEases have been shown to function as maturases and promote the splicing of their host group I intron and occasionally related introns (Lazowska et al., 1989; Ho et al., 1997; Ho and Waring, 1999; Bassi et al., 2002; Bassi and Weeks, 2003; Belfort, 2003; Longo et al., 2005).

During the study of the arrangement of the mt SSU (*rns*) gene within *Leptographium* species described in Chapter 4, several closely-related composite group II introns with putative LHEGs were uncovered. Comparative sequence

analysis and RNA modelling indicate that the LHEase ORF is situated in domain (D) III, a ribozyme component that acts as a catalytic effector in intron splicing (Lehmann and Schmidt, 2003; Fedorova and Zingler, 2007; Pyle, 2010). To the best of my knowledge this study constitutes the first biochemical analysis of a group II intron that encodes an HEase of the LAGLIDADG family rather than the RT family. The specific goals of this study were: (i) to determine if ORF-less and ORF-containing versions of the intron are self-splicing *in vitro*; (ii) to determine if the LHEase enhances intron-catalyzed splicing *in vitro*; and (iii) to assess if the ORF encodes an active LHEase and to map its cleavage site.

5.2. Methods overview

As described in Chapter 4, the mt *rns* gene of forty-seven strains belonging to the fungal genus *Leptographium* was screened for the presence of group II introns introns using PCR. Amplicons corresponding to intron-minus (1.2 kb) and intron-plus (3 kb) alleles from representative strains of *Leptographium* were cloned into the TOPO-TA Cloning® Kit for Sequencing (Life Technologies) and plasmid DNA was sequenced using the Big Dye (version 3.1) protocols supplied by the manufacturer (Applied Biosystems, Foster City, USA).

The group II intron/LHEG composite element from *L. truncatum* strain CBS929.85 was selected for functional characterization. Self-splicing of the intron *in vitro* was examined using both ORF-minus and ORF-plus versions of the precursor transcript. The LHEG sequence was optimized for expression in *E. coli* and two versions of the protein were expressed; one with an N-terminal His_6 tag and the other, near-native, with a C-terminal glycine residue. RNA-binding activity of the N-terminal His_6-tagged protein was examined using RNA filter-binding assays, and

enhancement of intron splicing with the addition of protein was tested using both versions of the intron-encoded protein. DNA cleavage activity of the LHEase was also tested and the cleavage site was mapped.

Evolutionary relationships amongst the host (*rns*) gene, group II intron, and LHEase amino acid sequences were examined using neighbour joining, parsimony, maximum likelihood (as implemented by Tree Puzzle), and Bayesian analyses. Sequences were obtained from GenBank and from cultures maintained at the University of Manitoba WIN(M) culture collection. The sequences used in the mt *rns* gene phylogeny are listed in listed in Table 5.1 and sequences used in the intron and LHEase amino acid phylogenies are provided in Tables 5.2 and 5.3, respectively. Sequence logos were generated for exon sequences flanking the cleavage site and for the group II intron-encoded LHEases.

Table 5.1. Strains of ascomycetous fungi and of *Leptographium* species used in the analysis of the mt *rns* gene, along with GenBank accession numbers.

Strain[1]	Accession Number	Strain	Accession Number	Strain	Accession Number
Aspergillus niger	DQ207726	*Hypocrea lutea*	AB027362	*Magnaporthe grisea*	AF056626
Aspergillus tubingensis	DQ217399	*Hypomyces chrysospermus*	AB027363	*Metarhizium anisopliae*	AB027361
Beauveria bassiana	U91338	*Kluyveromyces thermotolerans* CBS6340	AJ634268	*Neurospora crassa*	Z34001
Beauveria brongniartii	AB027359	*Lecanicillium muscarium*	AF487277	*Ophiocordyceps konnoana*	AB031194
Ceratocystis ossiformis WIN(M)52	GU989595	*Leptographium lundbergii* DAOM60397	GU983686	*Ophiocordyceps sobolifera*	AB027350
Cordyceps sp. 97003	AB027352	*Leptographium lundbergii* NFRI69-148	GU983684	*Ophiostoma coronatum* WIN(M)43	GU989601
Cordyceps sp. 97009	AB027356	*Leptographium lundbergii* NFRI89-1040/1/3[2]	GU983685	*Ophiostoma grande* CBS250.78	GU989598
Cordyceps coccidiicola	AB031197	*Leptographium lundbergii* NFRI89-1040/1/3[3]	GU983687	*Ophiostoma longirostellatum* CBS134.51	GU989594
Cordyceps inegoensis	AB027344	*Leptographium lundbergii* NFRI1502/1[2]	GU983688	*Ophiostoma microsporum* WIN(M)513	GU989596
Cordyceps japonica	AB027342	*Leptographium lundbergii* NFRI1502/1[3]	GU983689	*Ophiostoma pseudonigrum* WIN(M)47	GU989602
Cordyceps militaris	AB027357	*Leptographium terebrantis* CBS337.70[2]	GU983690	*Paecilomyces tenuipes*	AB027358
Cordyceps ophioglossoides	AB027343	*Leptographium terebrantis* CBS337.70[3]	GU983691	*Penicillium chrysogenum*	Z23072
Cordyceps paradoxa	AB027345	*Leptographium truncatum* CBS929.85[2]	GU949593	*Penicillium expansum*	AY157626
Cordyceps prolifica	AB027346	*Leptographium truncatum* CBS929.85[3]	GU949595	*Penicillium glabrum*	AF245271
Cordyceps ramosopulvinata	AB027348	*Leptographium truncatum* Forintek C34	GU983692	*Penicillium marnefei*	AY347307
Cryphonectria parasitica	AF029891	*Leptographium truncatum* J.R.88-449	GU983693	*Penicillium thomii*	AY157627
Emericella nidulans	J01390	*Leptographium truncatum* NFRI59-7/3	GU983694	*Podospora anserina*	X14734
Epidermophyton floccosum	AY916130	*Leptographium truncatum* NFRI1813/1[2]	GU949594	*Saccharomyces* sp. IFO1802	AF114914
Grosmannia crassivaginata WIN(M)1589	GU989600	*Leptographium truncatum* NFRI1813/1[3]	GU949596	*Saccharomyces cerevisiae*	AJ011856
Grosmannia penicillata NFRI60-21	GU989599	*Leptographium truncatum* TOM74.29	GU949597	*Saccharomyces paradoxus* NRRLY-17217	AF114922
Grosmannia pseudoeurophioides WIN(M)42	GU989597	*Leptographium wingfieldii* TOM1.3	GU983695	*Verticillium dahliae*	DQ351941
Hypocrea jecorina	AF447590	*Leptographium wingfieldii* TOM9.4	GU983696		

[1]J.R., J. Reid; CBS, Centraal Bureau voor Schimmelcultures, Utrecht, The Netherlands; WIN(M), University of Manitoba, Microbiology/Botany (J.R.'s personal collection); NFRI, Norwegian Forest Research Institute, AS, Norway; DAOM, Cereal and Oilseeds Research, Agriculture & Agri-Food Canada, Ottawa, Ont., Canada; TOM, Isolation designation, Canadian Forest Service, Great Lakes Forestry Centre, 1219 Queen St., Sault Ste. Marie, ON, P6A 5M7.

[2]The strain is heteroplasmic for the mt *rns* gene. This sequence represents the intron-plus allele.

[3] The strain is heteroplasmic for the mt *rns* gene. This sequence represents the intron-minus allele.

Table 5.2. Sequences used in the mitochondrial group II intron phylogeny.

Strain	Intron designation[1]	Accession number
Agrocybe aegerita	A.a.rnli5	AF087656
Cordyceps sp. 97003	C.97003.rnsi1	AB027352
Cordyceps sp. 97009	C.97009.rnsi1	AB027356
Cordyceps ramosopulvinata	C.r.rnsi1	AB027348
Cryphonectria parasitica	C.p.rnsi1	AF029891
Cryphonectria parasitica	C.p.rnsi3	AF029891
Leptographium lundbergii DAOM60397	L.lu.rnsi1	GU983686
Leptographium lundbergii NFRI89-1040/1/3	L.lu.rnsi1	GU983685
Leptographium lundbergii NFRI1502/1	L.lu.rnsi1	GU983688
Leptographium terebrantis CBS337.70	L.te.rnsi1	GU983690
Leptographium truncatum CBS929.85	L.SSU/1	GU949593
Leptographium truncatum Forintek C34	L.tr.rnsi1	GU983692
Leptographium truncatum NFRI59-7/3	L.tr.rnsi1	GU983694
Leptographium truncatum NFRI1813/1	L.tr.rnsi1	GU949594
Leptographium wingfieldii TOM1.3	L.wi.rnsi1	GU983695
Leptographium wingfieldii TOM9.4	L.wirnsi1	GU983696
Ophiocordyceps konnoana	C.k.rnsi1	AB031194
Ophiocordyceps sobolifera	O.s.rnsi1	AB027350

[1] Nomenclature of the intron follows that used by Monteiro-Vitorello et al. (2009) and Mullineux et al. (2010).

Table 5.3. Sequences used in the mitochondrial LHEase phylogeny.

Strain	Locus[1]	Intron name (ORF designation)[2]	Accession number
Amoebidium parasiticum	*rnl*	A.p.rnli2	AF538042
Agrocybe aegerita	*rns*	A.a.rnsi1	AAB50391
Cordyceps sp. 97009	*rns*	C.97009.rnsi1	AB027356
Cordyceps ramosopulvinata	*rns*	C.r.rnsi1	AB027348
Cryphonectria parasitica	*rnl*	C.p.rpmi	AAC24230
Cryphonectria parasitica	*rns*	C.p.rnsi1	AF029891
Cryphonectria parasitica	*rns*	C.p.rnsi2	AF029891
Cryphonectria parasitica	*rns*	C.p.rnsi3	AF029891
Cryphonectria parasitica	*rns*	C.p.rnsi4	AF029891
Hypocrea jecorina	free-standing	HyjefMp02	NP_570143
Hypocrea jecorina	free-standing	HyjefMp12	NP_570153
Neurospora crassa	ATPase subunit 6	N.c.ATP6i2	T50468
Neurospora crassa	ND4L	N.c.ND4Li1	CAA28761
Neurospora crassa	ND5	N.c.ND5i1	CAA28764
Leptographium lundbergii DAOM60397	*rns*	I-LluI-P	GU983686
Leptographium lundbergii NFRI89-1040/1/3	*rns*	I-LluII-P	GU983685
Leptographium lundbergii NFRI1502/1	*rns*	I-LluIII-P	GU983688
Leptographium terebrantis CBS337.70	*rns*	I-LteI-P	GU983690
Leptographium truncatum CBS929.85	*rns*	I-LtrII	GU949593
Leptographium truncatum Forintek C34	*rns*	I-LtrIII-P	GU983692
Leptographium truncatum NFRI1813/1	*rns*	I-LtrIV-P	GU949594
Leptographium wingfieldii TOM1.3	*rns*	I-LwiI-P	GU983695
Ophiocordyceps konnoana	*rns*	C.k.rnsi1	AB031194
Ophiocordyceps sobolifera	*rns*	O.s.rnsi2	AB027350
Ophiocordyceps sobolifera	*rns*	O.s.rnsi2-3	AB027350
Podospora anserina	*cob*	P.a.cobi1	NP_074921
Podospora anserina	*cox1*	P.a.cox1i5	NP_074928
Podospora anserina	*nad1*	P.a.ND1i4	NP_074960
Podospora anserina	*nad3*	P.a.ND3i1	NP_074914
Podospora anserina	*nad4*	P.a.ND4i1	P15564
Podospora anserina	*nad4L*	P.a.ND4Li1	NP_074942
Podospora anserina	*nad4L*	P.a.ND4Li2	CAA38797
Podospora anserina	*nad5*	P.a.ND5i2	NP_074945
Podospora anserina	*rnl*	P.a.rnli1	NP_074910
Ophiostoma ips	*rnl*	O.i.rnli1	ABI15908
Ophiostoma minus	*rnl*	O.m.rnli1	ABI15909
Ophiostoma novo-ulmi subsp. *americana*	*rnl*	O.n.u.rnli1	ABI15906
Ophiostoma novo-ulmi subsp. *americana*	*rnl*	I-Onu1	AAY59060
Sclerotinia sclerotiorum	*rns*	S.s.rnsi1	AAC48982

[1] *rnl*, gene encoding the large subunit rRNA; *rns*, gene encoding the small subunit rRNA; *nad*, genes encoding subunits of the mitochondrial NADH dehydrogenase protein complex; free-standing, based on GenBank accession no. AAL74167; *cox1*, gene encoding cytochrome oxidase subunit1; *cob*, gene encoding apocytochrome b.

[2] Nomenclature of the intron follows that used by Monteiro-Vitorello et al. (2009) and Mullineux et al. (2010).

5.3. Results

5.3.1. The mitochondrial *rns* gene contains a group IIB1 intron with an LHEG

Amplification of the mt *rns* gene in members of *Leptographium* and *Grosmannia* yielded an amplicon of either 1.2 kb (corresponding to the intron-minus allele) or 3 to 5 kb (representing the intron-plus allele), and nineteen strains were found to contain insertions between 1.8 and 3.8 kb in size. Sequence analyses of intron-minus and intron-plus alleles of the mt *rns* gene in strains of *Leptographium* revealed the presence of a single intron belonging to the group II intron family inserted at position 952 (based on the numbering of positions in the 16S rRNA gene of *E. coli* strain J01695). The intron IS is conserved in members of *Leptographium* and corresponds to intron 3 (I3), or the mS952 intron, of the mt *rns* gene of *Cryphonectria parasitica* (Figure 5.1A). The intron encodes a putative ORF of the LAGLIDADG family of HEases. Characteristics of the mS952 intron and intronic ORF are described in Table 5.4.

The Lt.SSU/1 intron in *L. truncatum* strain CBS929.85 is 1840 nt in length and encodes an ORF of 915 nt that is inserted after intron position 685. The PCR screening of transformants that were obtained from cloning the amplicon of the mt *rns* gene revealed the presence of both intron-minus (1.2 kb) and intron-plus (3 kb) alleles in *L. truncatum* strain CBS929.85, and both versions of the *rns* gene were sequenced. Heteroplasmy in the mt *rns* gene was detected as well in other strains of *L. truncatum* (Figure 5.1A). Consequently, RT-PCR-based approaches or Northern analysis could not be used to determine the intron IS or demonstrate intron splicing *in vivo*, for it would not be possible to unambiguously distinguish between ligated exons and

intron-minus alleles. Therefore, the intron IS was delimited by comparison of intron-plus and intron-minus alleles.

A model of the RNA secondary structure showed that the intron belongs to the IIB1 subclass (Michel et al., 1989) and the ORF is inserted in DIII (Figure 5.1B). Exon binding site 1 (EBS1), which defines the 5' splice site by pairing with intron binding site 1 (IBS1), spans a segment of 7 nt. The EBS3-IBS3 base pairing interaction, involved in 3' splice site recognition, tertiary interactions (α-α', γ-γ', δ-δ', ε-ε', ζ-ζ', κ-κ', θ-θ') that are involved in ribozyme folding, and the bulged adenosine residue in DVI (involved in the branch reaction) were identified using previously published structural models (Costa et al., 1997a, 1998, 2000; Pyle et al., 2007).

Analysis of the ORF sequence using ORF Finder (National Center for Biotechnology Information) and the ProtParam (Gasteiger et al., 2005) programs identified a putative gene encoding an LHEase of 304 amino acids that contains two LAGIDADG motifs: ICGLVDAEG and LAGFIEGEA. Analysis using the Blastp program (National Center for Biotechnology Information) showed that this particular LHEG is related to that encoded within the mS952 group II intron present in the mt *rns* gene of *C. parasitica* (Toor and Zimmerly, 2002; Monteiro-Vitorello et al., 2009). The LHEase that was expressed in *E. coli* was 35 kDa, as determined by SDS-PAGE analysis (Figure 5.2). Based on the nomenclature of HEases proposed by Roberts et al. (2003), the LHEase has been designated I-LtrII, as it represents the second intron-encoded HEase described from *L. truncatum* (Sethuraman et al., 2009a).

Table 5.4. Sequence characteristics of group II introns and their putative LHEGs within the mt *rns* gene.

Strain	Intron sequence			ORF sequence				
	Insertion site[1]	Length (nt)	% GC	ORF start[2]	Length (nt/aa)[4]	% GC	LAGLIDADG motifs	Evidence for degeneration
Cordyceps sp. 97003	952	1094-1095	29.3	719[3]	990/156, 47	28.2	IIGLV...IAEG[5] LVGFMEGES	Yes
Cordyceps sp. 97009	952	871	29.2	591	933/309	24.7	VSGLVDAEG, LAGFMEGES	No
Cordyceps ramosopulvinata	952	822	28.7	592	939/312	24.7	VSGLVDAEG, LEGFMEGES	No
Cryphonectria parasitica Ep155	952	1006	34.0	752	1062/353	32.2	ITGLVDAEG, LAGFTEGES	No
Ophiocordyceps konnoana	952	796	29.4	585	1032/343	24.9	VSGLVDAEG, LTGFMEGES	No
Ophiocordyceps sobolifera	952	977	27.2	604	977/127, 164	23.7	ISGLVDAEG, LVGFMEGES	Yes
Leptographium lundbergii DAOM60397	952	925	29.1	685	948/127, 157	25.2	ICGLVDAEG, LAGFIEGEA	Yes
Leptographium lundbergii NFRI1502/1	952	925	29.1	685	948/127, 157	25.2	ICGLVDAEG, LAGFIEGEA	Yes
Leptographium lundbergii NFRI89-1040/1/3	952	925	29.1	685	948/127, 157	25.2	ICGLVDAEG, LAGFIEGEA	Yes
Leptographium truncatum CBS929.85	952	925	29.0	685	915/ 304	23.9	ICGLVDAEG, LAGFIEGEA	No
Leptographium truncatum Forintek C34	952	925	29.0	685	915/ 304	24.0	ICGLVDAEG, LAGFIEGEA	No
Leptographium truncatum NFRI59-7/3	952	925	29.0	685	915/304	24.0	ICGLVDAEG, LAGFIEGEA	No
Leptographium truncatum NFRI1813/1	952	925	29.1	685	915/148, 139	23.9	ICGLVDAEG, LAGFIEGEA	Yes
Leptographium terebrantis CBS337.70	952	961	30.5	721	915/304	24.0	ICGLVDAEG, LAGFIEGEA	No
Leptographium wingfieldii TOM1.3	952	961	30.4	721	915/304	23.9	ICGLVDAEG, LAGFIEGEA	No

L. wingfieldii TOM9.4	952	961	30.4	721	915/304	23.9	ICGLVDAEG, LAGFIEGEA	No

[1] The numbering of the positions is based on the numbering of the 16S rRNA gene of *E. coli* J01695.

[2] The numbering of the positions is based on the numbering of the intron sequence.

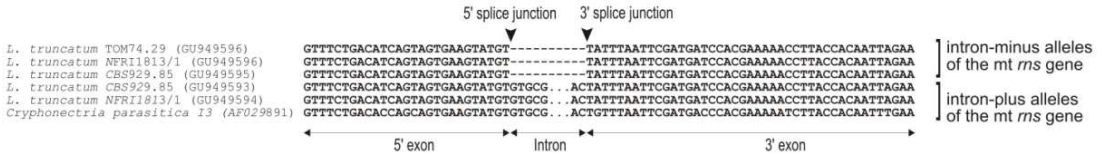
[3] The ORF sequence boundaries (including start and stop codons) were identified by sequence comparison with the remaining strains.

[4] Where there is apparent evidence for degeneration of the ORF and the LAGLIDADG motifs are split by non-coding sequence the length of the amino acid sequence for each of the two "putative" proteins are indicated.

[5] The P1 motif is split by: VFFFYKKK.

Figure 5.1. Insertion site of group II introns and model of the RNA secondary structure of group IIB1 introns in *Leptographium*. (A) Sequence alignment showing the IS of group II introns encoding putative LHEases in the mt *rns* gene of *Leptographium* spp. The intron IS was delimited by comparison of intron-minus and intron-plus alleles of the mt *rns* gene. The corresponding sequence of the mt *rns* gene of *C. parasitica*, which contains a related intron (intron 3), is included for comparison. Both the 5' exon and 3' exon sequences that flank the intron IS and the intron sequences at the 5'- and 3'-termini, which follow the group II consensus GUGYG and AY, respectively, are shown. The rest of the intervening sequence is represented by dots. The 5' and 3' splice junctions are designated by black arrows. Heteroplasmy (intron-plus and intron-minus alleles) has been observed in *L. truncatum* strains CBS929.85 and NFRI1813/1. (B) Secondary structure model of the group IIB1 intron, Lt.SSU/1, of *L. truncatum* showing the insertion of the ORF in a looped-out region in DIII. The 5' and 3' splice sites are marked by black arrowheads and the bulged adenosine residue in DVI is indicated by the asterisk. EBS and IBS nucleotides are shown, as well as nucleotides/segments involved in tertiary interactions (Greek letters).

A


5' splice junction
3' splice junction
L. truncatum TOM74.29 (GU949596) GTTTCTGACATCAGTAGTGAAGTATGT----------TATTTAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA
L. truncatum NFRI1813/1 (GU949596) GTTTCTGACATCAGTAGTGAAGTATGT----------TATTTAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA
L. truncatum CBS929.85 (GU949595) GTTTCTGACATCAGTAGTGAAGTATGT----------TATTTAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA
L. truncatum CBS929.85 (GU949593) GTTTCTGACATCAGTAGTGAAGTATGTGTGCG...ACTATTTAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA
L. truncatum NFRI1813/1 (GU949594) GTTTCTGACATCAGTAGTGAAGTATGTGTGCG...ACTATTTAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA
Cryphonectria parasitica I3 (AF029891) GTTTCTGACACCAGCAGTGAAGTATGTGTGCG...ACTGTTTAATTCGATGACCCACGAAAAATCTTACCACAATTTGAA
intron-minus alleles of the mt rns gene
intron-plus alleles of the mt rns gene
5' exon
Intron
3' exon

B

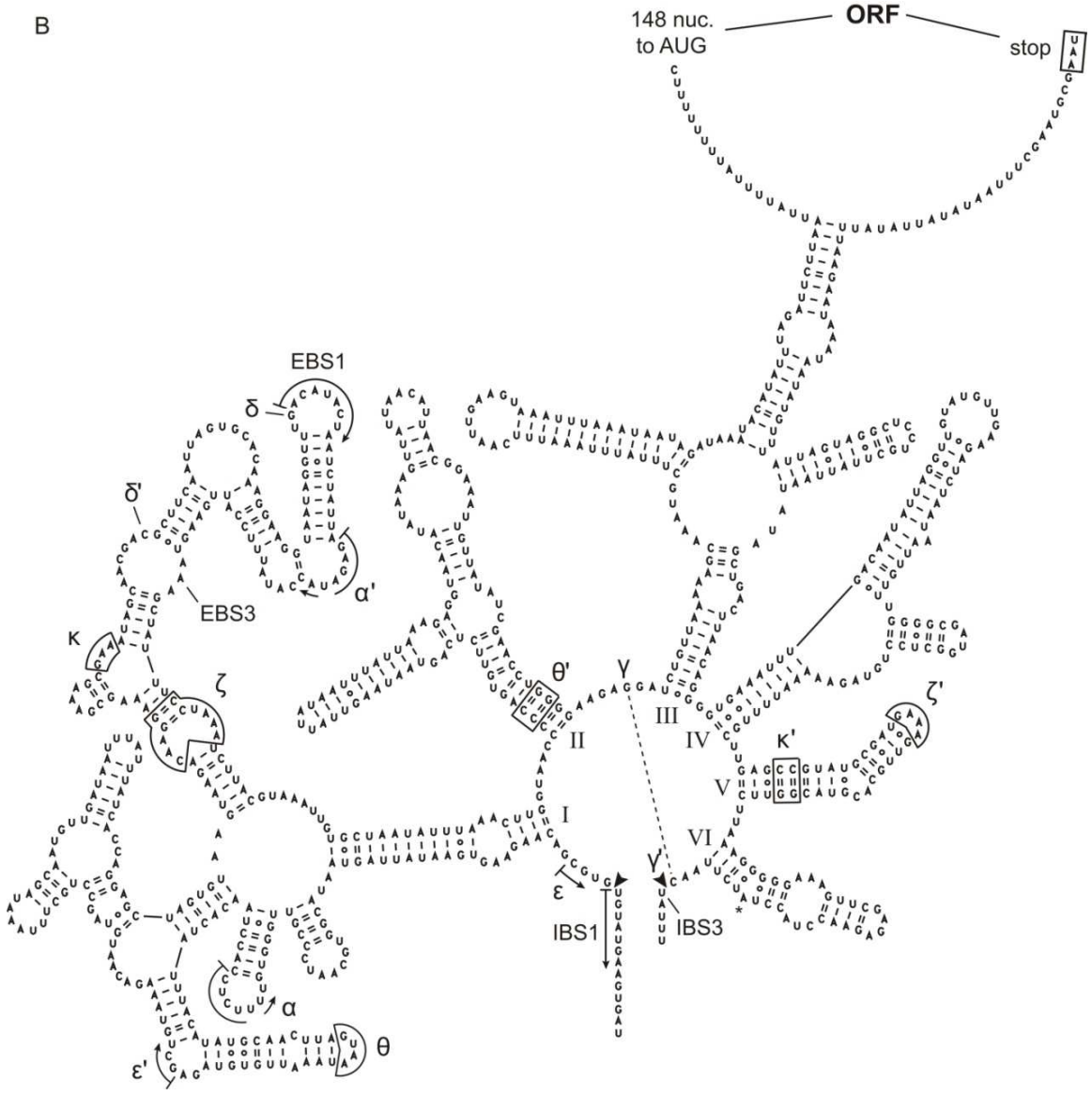
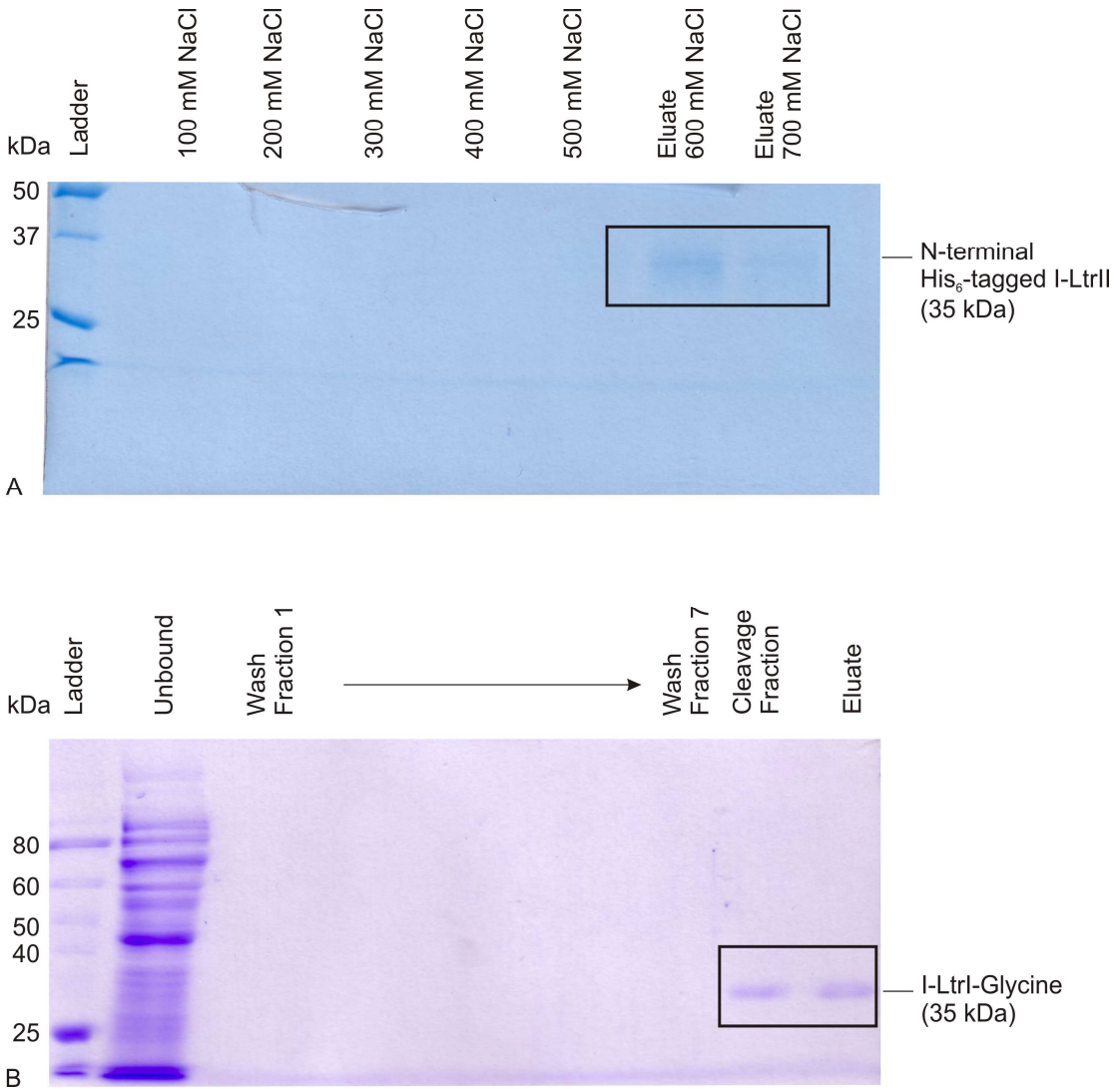

148 nuc. to AUG
ORF
stop
EBS1
δ
δ'
α'
EBS3
κ
ζ
θ'
γ
III
IV
κ'
ζ'
V
II
I
VI
γ'
ε
IBS1
IBS3
α
θ
ε'

Figure 5.2. Purification of a LAGLIDADG-type homing endonuclease encoded by a group II intron. (A) SDS-PAGE gel of purified N-terminal His_6-tagged I-LtrII purified HiTrap™ heparin HP column. (B) SDS-PAGE gel of I-LtrII purified using the Intein Mediated Purification with an Affinity Chitin-binding Tag system (New England Biolabs). The sample collected from the cleavage fraction contains protein that was eluted just after DTT was rapidly flushed through the column. DTT was used to cleave the target protein from the intein/chitin binding domain tag. The fractions containing purified protein are enclosed in the box, and the eluted protein has an estimated molecular weight of 35 kDa.


kDa
Ladder
100 mM NaCl
200 mM NaCl
300 mM NaCl
400 mM NaCl
500 mM NaCl
Eluate 600 mM NaCl
Eluate 700 mM NaCl
50
37
25
A
N-terminal His6-tagged I-LtrII (35 kDa)
kDa
Ladder
Unbound
Wash Fraction 1
Wash Fraction 7
Cleavage Fraction
Eluate
80
60
50
40
25
B
I-LtrI-Glycine (35 kDa)

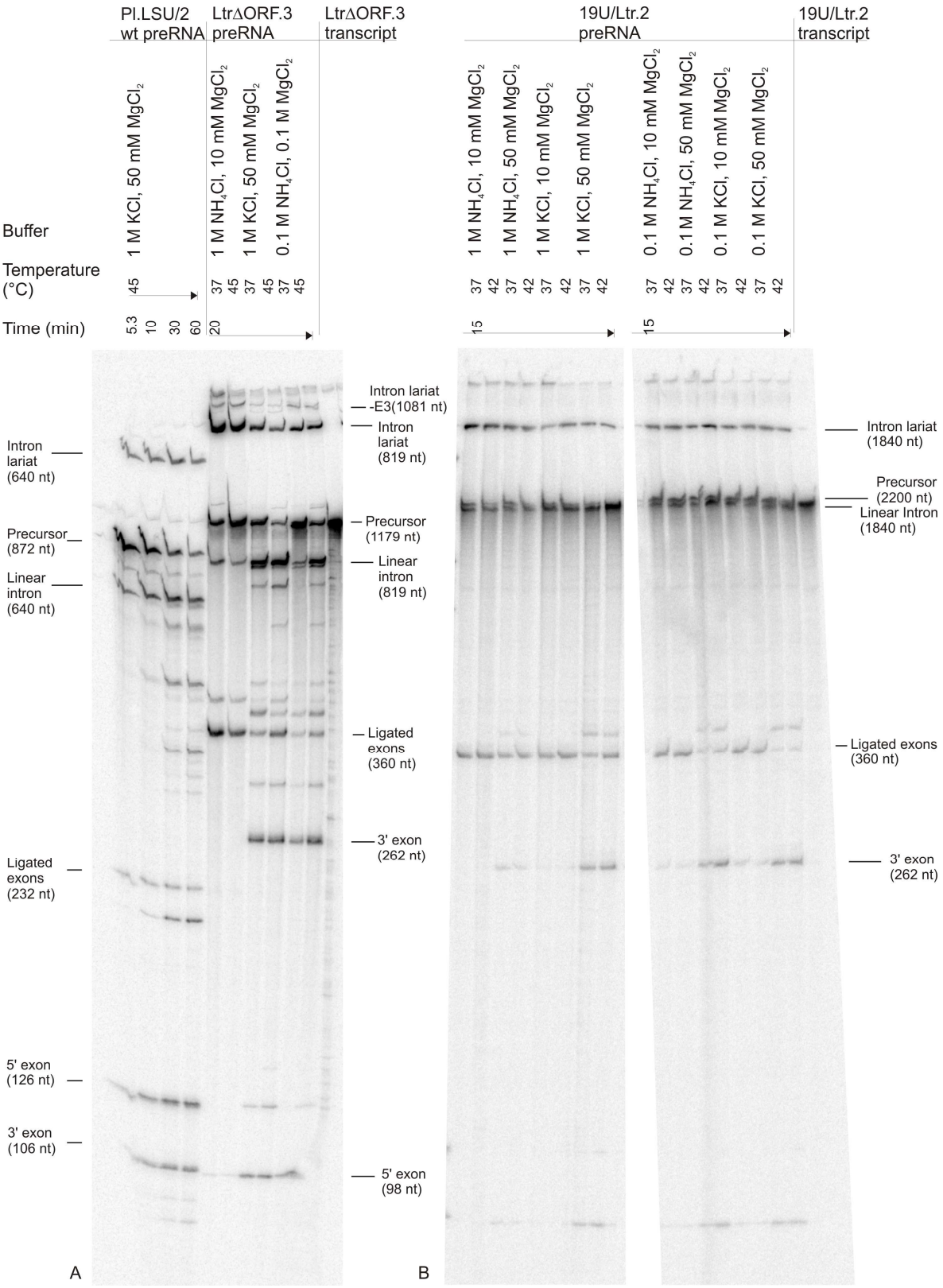
5.3.2. The Lt.SSU/1 intron self-splices *in vitro* in the absence of protein factors

The ability of Lt.SSU/1 to self-splice in the absence of protein factors was examined for both ORF-less and ORF-containing versions of the intron RNA. The *in vitro* transcribed precursor contained 98 nt of the 5' exon and 262 nt of the 3' exon. The ORF-less intron sequence (LtrΔORF.3) was 819 nt in length, while the full-length, ORF-containing, intron (19U/Ltr.2) was 1840 nt. Self-splicing of the *in vitro*-generated RNA transcripts was tested under low (0.1 M) or high (1 M) concentrations of either NH_4Cl or KCl and 10 to 100 mM $MgCl_2$ (Figure 5.3). The LtrΔORF.3 readily self-spliced during the 20-minute reaction period at 37 °C and 45 °C, and both intron lariat and linear molecules were observed. Self-splicing in the presence of KCl or at an elevated temperature (45 °C) resulted in the enhancement of splicing via hydrolysis at the expense of the branching mechanism, although both pathways operated simultaneously under the conditions tested (Figure 5.3A). Bands corresponding to additional products were visible in potassium-containing reactions, as well as in some reactions carried out at elevated magnesium concentrations. The former conditions are known to promote hydrolysis at sites that can pair with the EBS1 sequence: it is worth noting that sequences AAUAUGU and UGUAUGU, which resemble the IBS1 sequence, are present at intron positions 615 and 688, respectively, of the ORF-less construct.

Self-splicing of the full-length (ORF-plus) intron sequence was also examined to determine if the presence of the ORF sequence in DIII affected ribozyme activity; the DIII region is involved in promoting the catalytic efficiency of the active ribozyme (Lehmann and Schmidt, 2003; Fedorova and Zingler, 2007; Pyle, 2010). Self-splicing of the ORF-containing precursor RNA was examined within a 15-minute reaction period, and the production of lariat and linear intron molecules was observed

at 37 °C and 42 °C, 0.1 to 1 M of monovalent cation, and 10 to 50 mM $MgCl_2$. As was observed with the ORF-less version of the intron RNA, splicing occurs via the branching and hydrolytic pathways (Figure 5.3B).

Figure 5.3. Gel electrophoresis of *in vitro* splicing reactions by Lt.SSU/1 intron under various temperature and ionic conditions. Internally-labelled precursor transcripts were incubated for the length of time indicated prior to loading onto a 4 % polyacrylamide/8 M urea gel. (A) The LtrΔORF.3 version of the intron, from which the ORF sequence is absent, comprises 819 nt. Products from a self-splicing reaction of the Pl.LSU/2 intron from *P. littoralis* (Costa et al., 1997) were used to generate a molecular weight calibration. (B) The 19U/Ltr.2 version of the intron (1840 nt) contains the full-length ORF sequence. "LtrΔORF.3" and "19U/-Ltr.2 transcript" designate unincubated precursor transcript.

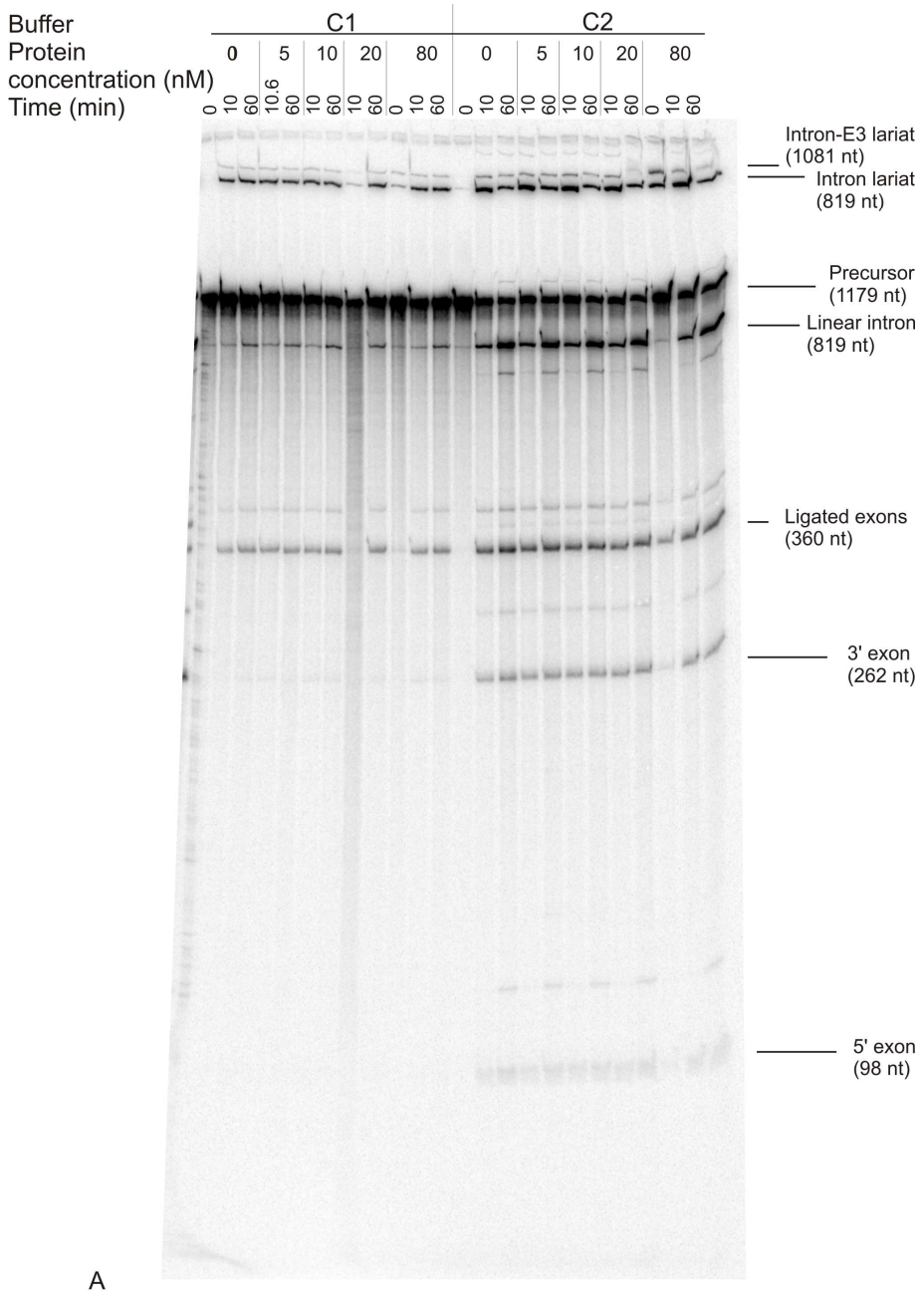

Pl.LSU/2 wt preRNA
LtrΔORF.3 preRNA
LtrΔORF.3 transcript
19U/Ltr.2 preRNA
19U/Ltr.2 transcript
Buffer
Temperature (°C)
Time (min)
1 M KCl, 50 mM MgCl2
1 M NH4Cl, 10 mM MgCl2
1 M KCl, 50 mM MgCl2
0.1 M NH4Cl, 0.1 M MgCl2
0.1 M NH4Cl, 10 mM MgCl2
0.1 M NH4Cl, 50 mM MgCl2
0.1 M KCl, 10 mM MgCl2
0.1 M KCl, 50 mM MgCl2
Intron lariat (640 nt)
Precursor (872 nt)
Linear intron (640 nt)
Ligated exons (232 nt)
5' exon (126 nt)
3' exon (106 nt)
Intron lariat -E3(1081 nt)
Intron lariat (819 nt)
Precursor (1179 nt)
Linear intron (819 nt)
Ligated exons (360 nt)
3' exon (262 nt)
5' exon (98 nt)
Intron lariat (1840 nt)
Precursor (2200 nt)
Linear Intron (1840 nt)
Ligated exons (360 nt)
3' exon (262 nt)
A
B

5.3.3. The I-LtrII protein does not bind RNA or enhance the efficiency of intron splicing

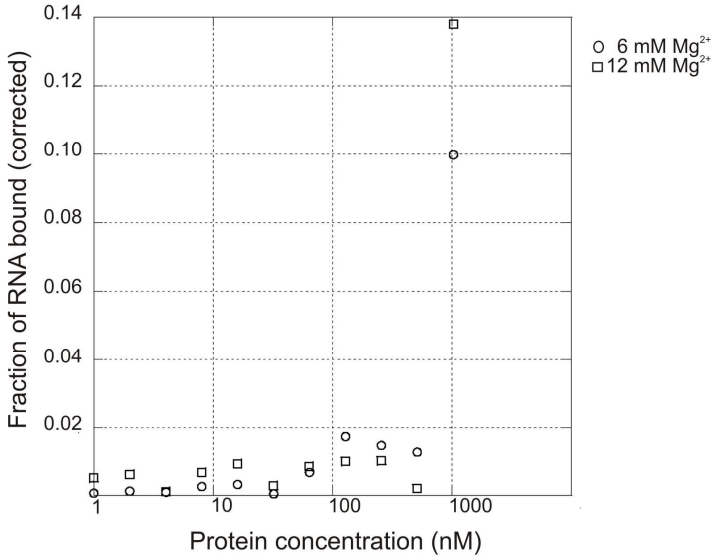
Potential maturase activity of the I-LtrII protein was examined for both the N-terminal His_6-tagged form of the purified protein and a near-native version with an additional glycine residue at the C-terminus. N-terminal His_6-tagged I-LtrII (at a concentration of 0 to 80 nM) was added to 20 nM of internally-labelled LtrΔORF.3 precursor RNA in a splicing buffer that contained 150 mM KCl and either 6 or 12 mM $MgCl_2$. The reaction was incubated at 37 °C and terminated after 10 and 60 minutes. Under the conditions tested, the LtrΔORF.3 intron was capable of self-splicing; however, no increase in the production of linear and lariat intron RNAs under increasing concentrations of I-LtrII was observed, suggesting that it did not enhance the efficiency of intron splicing (Figure 5.4A).

The binding affinity of the I-LtrII protein for the intron RNA precursor was examined using RNA filter-binding assays. Addition of N-terminal His_6-tagged I-LtrII (at a concentration of 0 to 512 nM) to 5 nM of internally-labelled LtrΔORF.3 RNA precursor in the presence of 6 or 12 mM $MgCl_2$ showed negligible RNA binding (Figure 5.4B). The apparent binding of 10 % (in the presence of 6 mM $MgCl_2$) and 14 % (in the presence of 12 mM $MgCl_2$) of LtrΔORF.3 RNA precursor by I-LtrII (at a concentration of 1024 nM) is a reflection of aggregation. The low affinity of the I-LtrII protein for intron RNA indicates that it likely does not have RNA-binding activity.

Figure 5.4. Assaying for maturase activity of the N-terminal His_6-tagged I-LtrII. (A) Gel electrophoresis of *in vitro* splicing reactions of the LtrΔORF.3 precursor transcript in the presence of N-terminal His_6-tagged I-LtrII. Internally-labelled precursor transcripts were incubated with increasing concentrations of I-LtrII at 37 °C for the length of time indicated prior to loading onto a 4 % polyacrylamide/8 M urea gel. The composition of buffers C1 and C2 are described in Materials and Methods. (B) RNA-binding assays of N-terminal His_6-tagged I-LtrII. Values were corrected for binding of the eluate to an RNA-binding Hybond N+ filter, and also for the residual binding observed in the absence of protein.


Buffer
C1
C2
Protein concentration (nM)
0
5
10
20
80
Time (min)
0 10 60 10.6 60 10 60 10 60 0 10 60
Intron-E3 lariat (1081 nt)
Intron lariat (819 nt)
Precursor (1179 nt)
Linear intron (819 nt)
Ligated exons (360 nt)
3' exon (262 nt)
5' exon (98 nt)

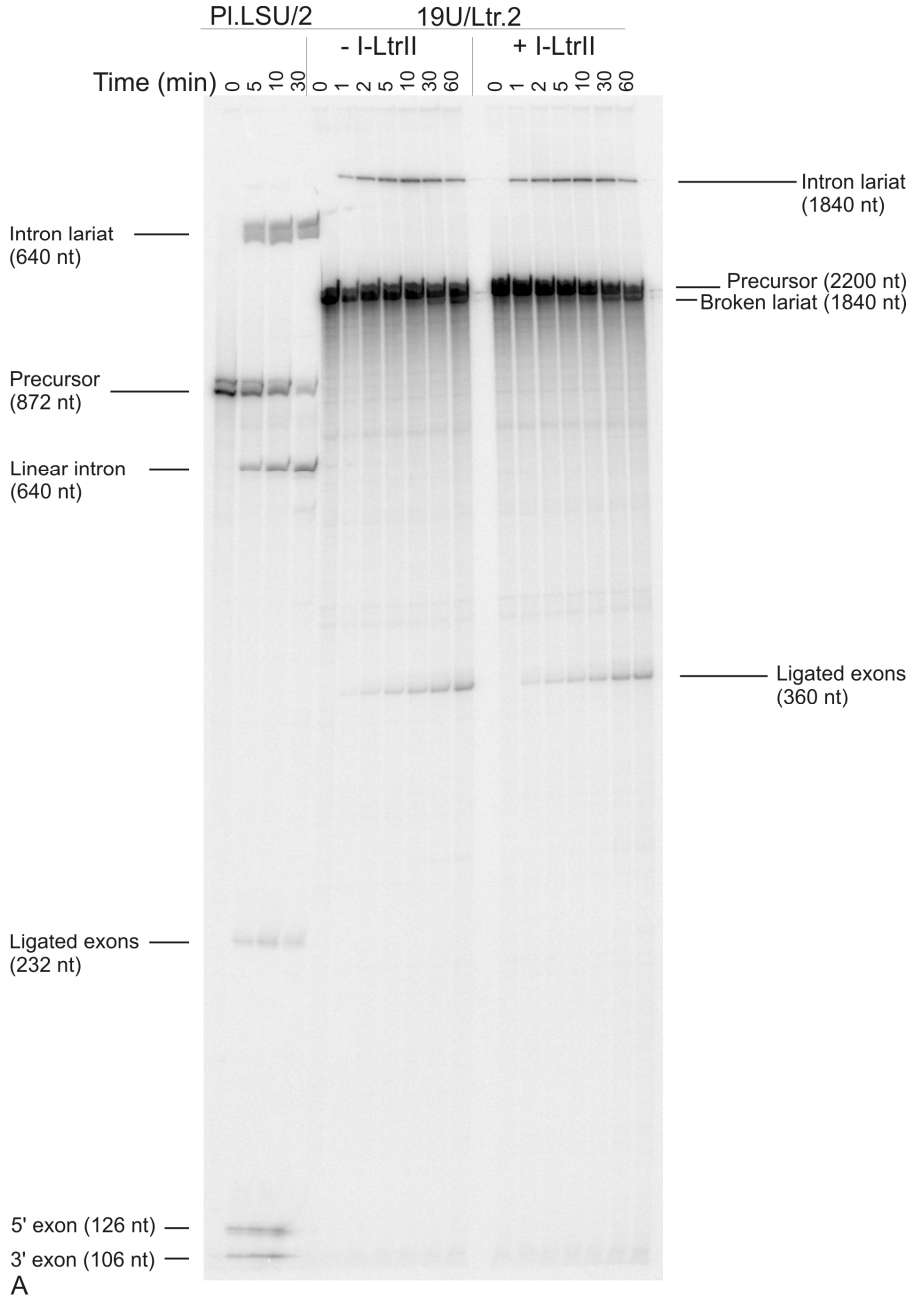
A


Fraction of RNA bound (corrected)
0.14
0.12
0.10
0.08
0.06
0.04
0.02
1
10
100
1000
Protein concentration (nM)
6 mM Mg2+
12 mM Mg2+

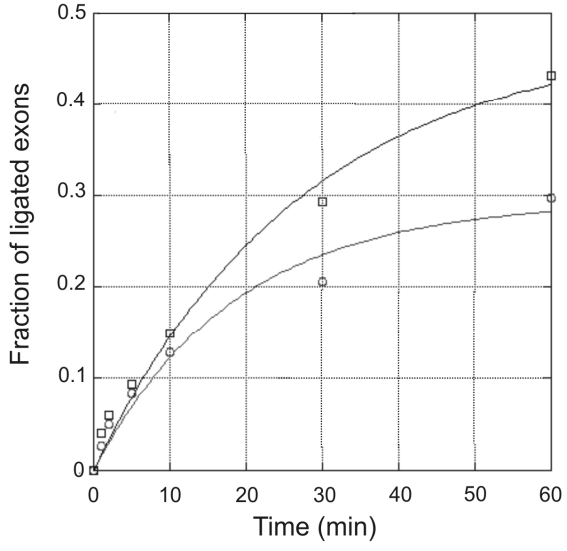
B

Since removal of the ORF sequence could have eliminated intron sequences involved in RNA-binding by the LHEase, self-splicing of the ORF-containing RNA precursor (19U/Ltr.2) in the presence of near-native I-LtrII was examined over the course of 60 minutes. Self-splicing of 100 nM of internally-labelled precursor transcript in the presence of 6 mM $MgCl_2$ was compared for samples to which near-native I-LtrII was added (at a concentration of 19.5 nM) or excluded. The results show that addition of near-native I-LtrII does not significantly enhance the formation of ligated exons and release of the full-length (ORF-plus) intron RNA (Figure 5.5A). The progress of the reaction was quantitated by determining the ratio of the molar concentration of ligated exons to the sum of the precursor and ligated exons (Figure 5.5B). While near-native I-LtrII appears to have a minor effect on the extent and rate of the formation of ligated exons by the ORF-containing intron precursor, the difference is far too small to be considered significant. Finally, maturase activity was also assayed at a magnesium concentration (3 mM) that does not support *in vitro* self-splicing but failed to restore splicing with the addition of the I-LtrII protein (data not shown).

Figure 5.5. Assaying for enhancement of intron splicing by near-native I-LtrII. (A) Gel electrophoresis of *in vitro* splicing reactions with the 19U/Ltr.2 precursor transcript in the presence of near-native I-LtrII. Internally-labelled precursor transcripts were incubated with near native I-LtrII at 37 °C for the length of time indicated prior to loading onto a 4 % polyacrylamide/8 M urea gel. Products from the self-splicing reaction of the Pl.LSU/2 intron from *P. littoralis* were used to generate a molecular weight calibration. (B) Quantitation of the progress of the splicing reaction based on the molar ratio of [ligated exons] over ([precursor] + [ligated exons]) (the calculation did not take into account the relatively greater vulnerability of the lengthy precursor molecules to degradation over extended reaction times). Curves were fitted to the equation m1 * (1 - exp(-m2 * t)) using the KaleidaGraph software. Circles (m1 = 0.294 ± 0.025; m2 = 0.054 ± 0.012; Pearson's R = 0.986) and squares (m1 = 0.477 ± 0.043; m2 = 0.036 ± 0.007; Pearson's R = 0.993) correspond to reactions in the absence and presence, respectively, of the I-LtrII protein.


PI.LSU/2
19U/Ltr.2
- I-LtrII
+ I-LtrII
Time (min)
0 5 10 30
0 1 2 5 10 30 60
0 1 2 5 10 30 60
Intron lariat (640 nt)
Precursor (872 nt)
Linear intron (640 nt)
Ligated exons (232 nt)
5' exon (126 nt)
3' exon (106 nt)
Intron lariat (1840 nt)
Precursor (2200 nt)
Broken lariat (1840 nt)
Ligated exons (360 nt)

A

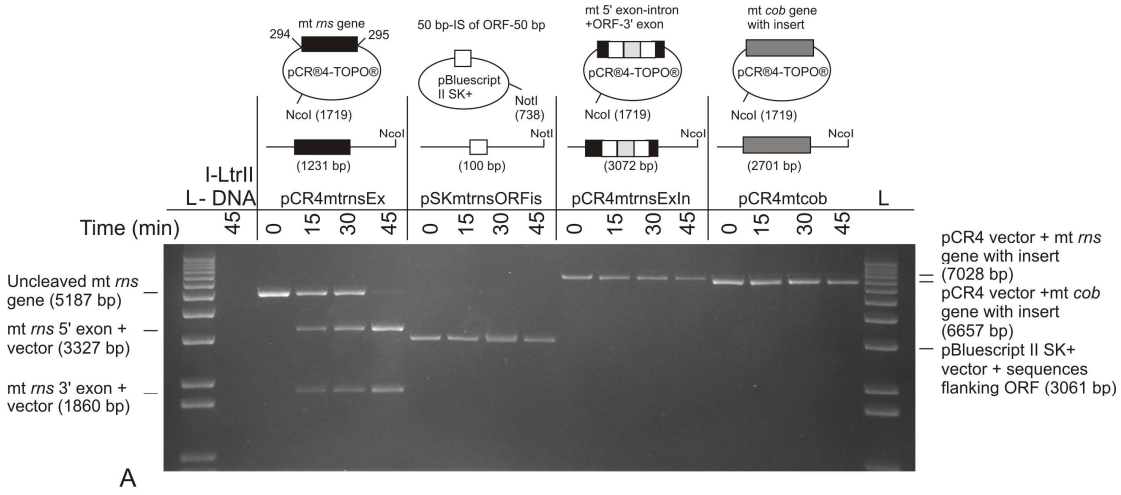

Fraction of ligated exons
0
0.1
0.2
0.3
0.4
0.5
0
10
20
30
40
50
60
Time (min)

B

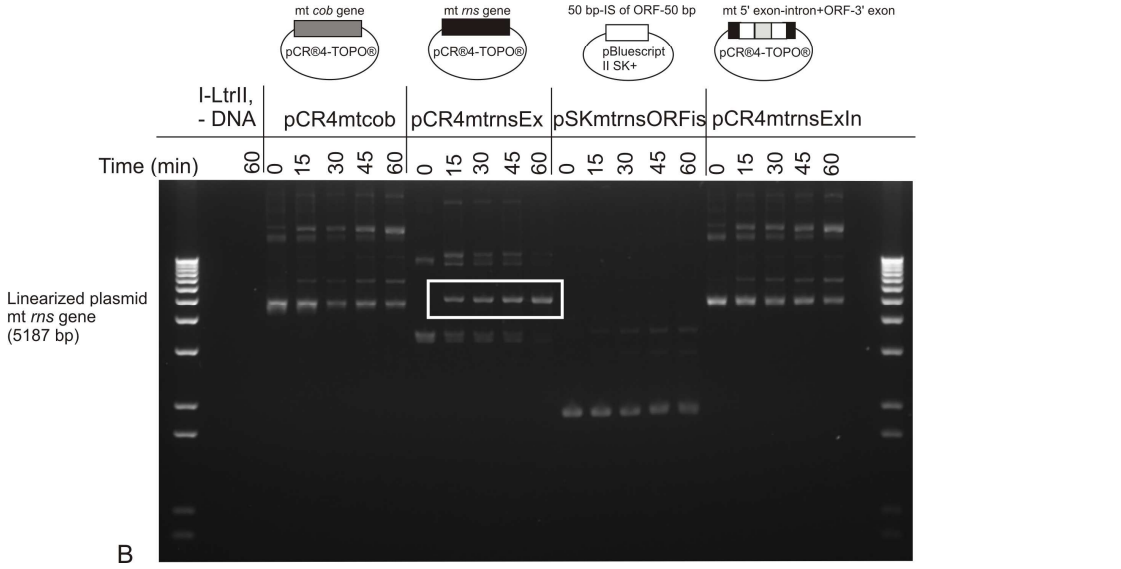
5.3.4. I-LtrII cleaves the mitochondrial *rns* gene upstream of the intron insertion site

Cleavage activity of I-LtrII was characterized using two potential substrates: pCR4mtrnsEx, consisting of the intron-minus allele of the mt *rns* gene (1.2 kb), and pSKmtrnsORFis, which is a synthetic 100-bp fragment composed of intronic sequences that flank the ORF (50 bp were included on each side of the ORF). The cleavage assays were carried out using non-labelled plasmid DNA that was first linearized with either NcoI (pCR4mtrnsEx) or NotI (pSKmtrnsORFis). Negative controls were NcoI-linearized plasmid DNA containing the intron-plus allele of the mt *rns* gene (pCR4mtrnsExIn) or the mt cytochrome b (*cob*) gene (construct pCR4mtcob); the latter sequence is an unrelated gene cloned into the pCR®4-TOPO® vector. The substrates and controls were incubated with near-native I-LtrII at 37 °C for 45 minutes, with time points taken every 15 minutes. The DNA samples were separated by agarose gel electrophoresis and the results show that I-LtrII efficiently cleaves only the linear plasmid containing the exon sequences; that is, I-LtrII cleaves the mt *rns* gene (Figure 5.6).

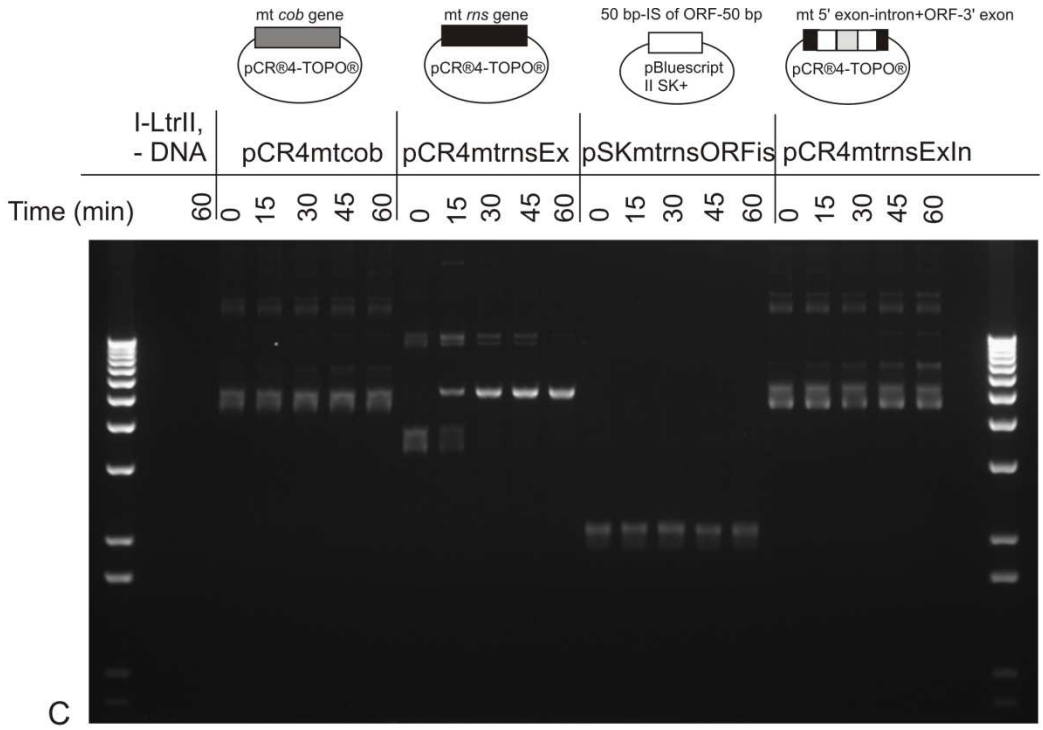
Figure 5.6. Cleavage activity of near-native I-LtrII protein. (A) The ability of I-LtrII to cleave linearized plasmid DNA containing exon sequences (pCR4mtrnsEX) and the intron sequences flanking the LHEase ORF (pSKmtrnsORFis) was tested in REact® Buffer 3. The cleavage assays were carried out using non-labelled plasmid DNA that was first linearized with either NcoI (pCR4mtrnsEx) or NotI (pSKmtrnsORFis). Negative controls were NcoI-linearized plasmid DNA containing the intron-plus allele of the mt *rns* gene (pCR4mtrnsExIn), in which both potential homing sites are disrupted, or the mt cytochrome b (*cob*) gene (construct pCR4mtcob), an unrelated gene, that was cloned into the pCR®4-TOPO® vector. The molecular weight marker, L, used was the 1Kb Plus DNA ladder (Life Technologies). The ability of I-LtrII to cleave plasmid DNA containing exon sequences (pCR4mtrnsEX) and the intron sequences flanking the LHEase ORF (pSKmtrnsORFis) was tested in REact® Buffer 3. The bands corresponding to the pCR4mtrnsEX substrate after it was linearized by I-LtrII are enclosed in the box. In construct pCR4mtrnsExIn both potential homing sites in the mt *rns* gene and intron are disrupted by the intron and ORF, respectively. An unrelated gene, the mt *cob* gene, that was cloned into the pCR®4-TOPO® vector was included as a control. (C) Cleavage activity of I-LtrII against uncut plasmid DNA was also tested in REact® Buffer 1, containing: 50 mM Tris-Cl (pH 8.0) and 10 mM $MgCl_2$, supplemented with 1 mM DTT, final concentation. (D) Cleavage activity of I-LtrII against uncut plasmid DNA was also tested in REact® Buffer 4, containing: 20 mM Tris-Cl (pH 7.4); 5 mM $MgCl_2$; and 50 mM KCl, supplemented with 1 mM DTT, final concentation.

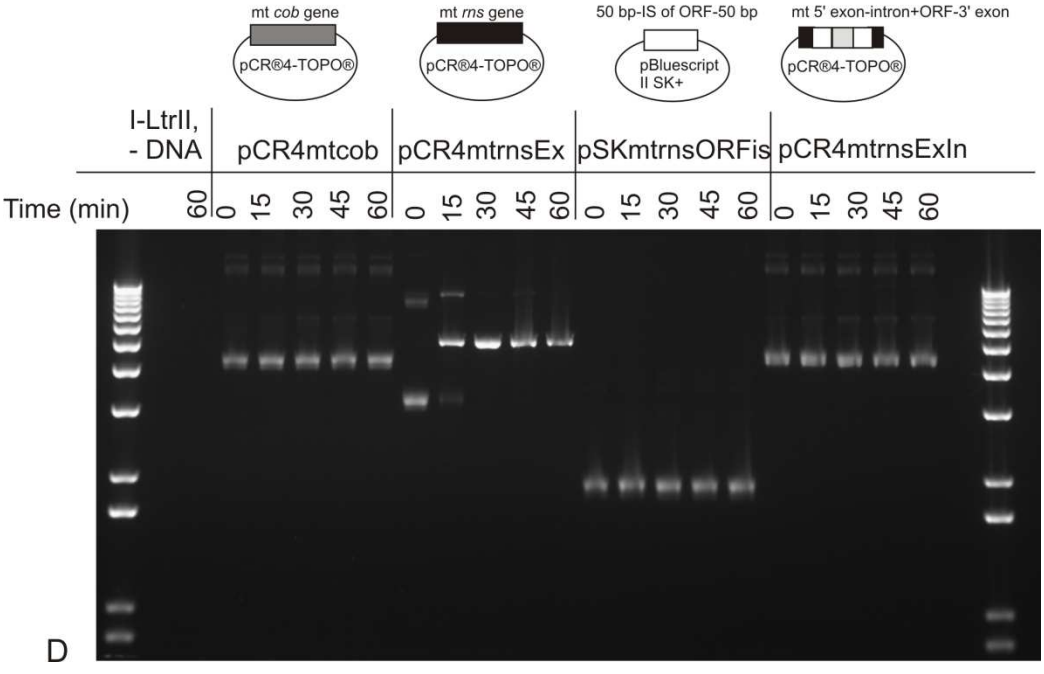

mt rns gene
294
295
pCR®4-TOPO®
NcoI (1719)
NcoI
(1231 bp)
50 bp-IS of ORF-50 bp
pBluescript II SK+
NotI (738)
NotI
(100 bp)
mt 5' exon-intron +ORF-3' exon
pCR®4-TOPO®
NcoI (1719)
NcoI
(3072 bp)
mt cob gene with insert
pCR®4-TOPO®
NcoI (1719)
NcoI
(2701 bp)
I-LtrII
L- DNA
pCR4mtrnsEx
pSKmtrnsORFis
pCR4mtrnsExIn
pCR4mtcob
L
Time (min)
45
0 15 30 45
0 15 30 45
0 15 30 45
0 15 30 45
Uncleaved mt rns gene (5187 bp)
mt rns 5' exon + vector (3327 bp)
mt rns 3' exon + vector (1860 bp)
pCR4 vector + mt rns gene with insert (7028 bp)
pCR4 vector +mt cob gene with insert (6657 bp)
pBluescript II SK+ vector + sequences flanking ORF (3061 bp)

A


mt cob gene
pCR®4-TOPO®
mt rns gene
pCR®4-TOPO®
50 bp-IS of ORF-50 bp
pBluescript II SK+
mt 5' exon-intron+ORF-3' exon
pCR®4-TOPO®
I-LtrII,
- DNA
pCR4mtcob
pCR4mtrnsEx
pSKmtrnsORFis
pCR4mtrnsExIn
Time (min)
60
0 15 30 45 60
0 15 30 45 60
0 15 30 45 60
0 15 30 45 60
Linearized plasmid mt rns gene (5187 bp)

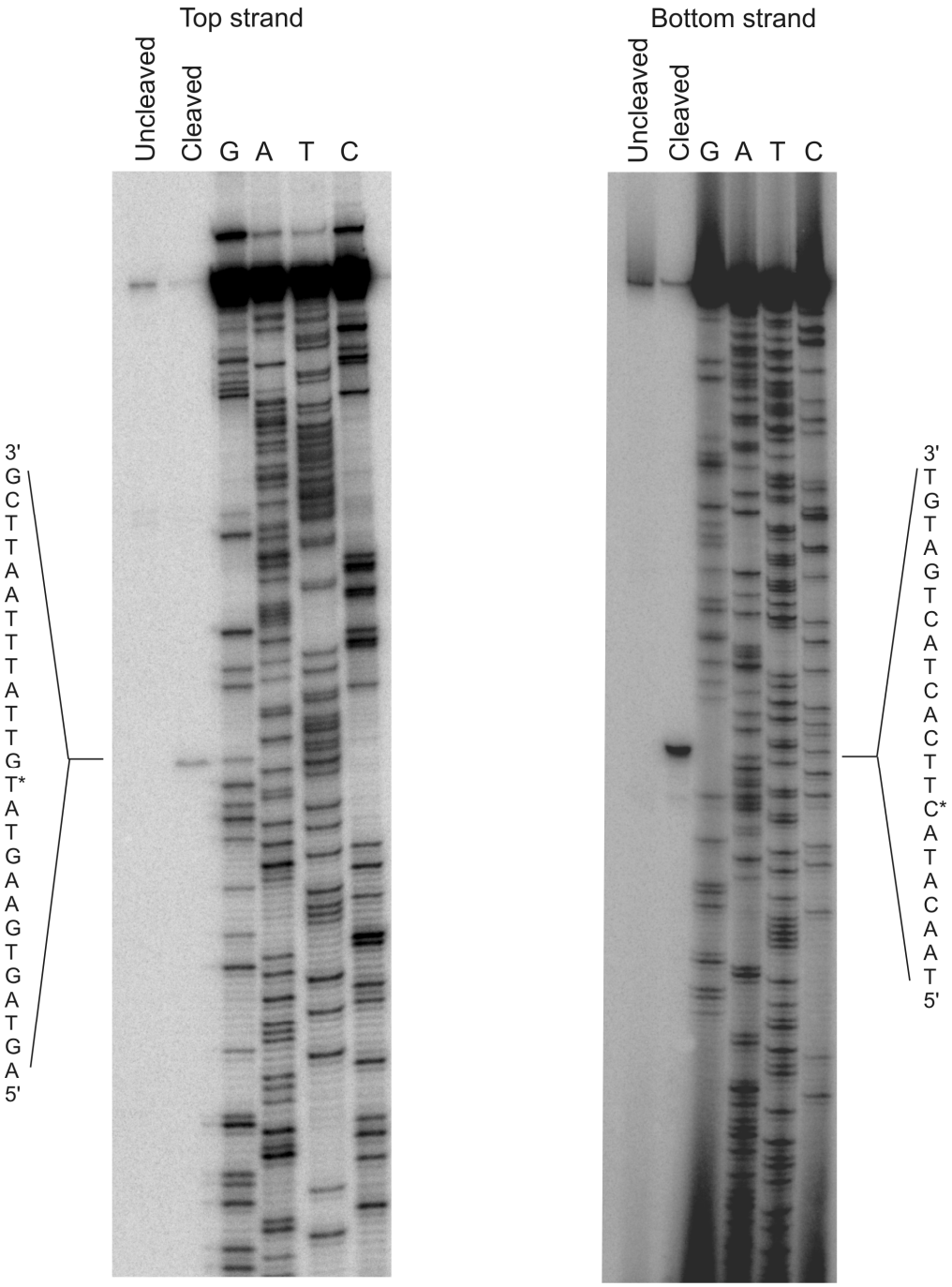
B

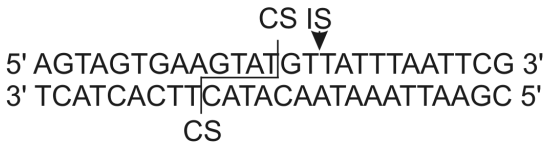

mt cob gene
mt rns gene
50 bp-IS of ORF-50 bp
mt 5' exon-intron+ORF-3' exon
pCR®4-TOPO®
pCR®4-TOPO®
pBluescript II SK+
pCR®4-TOPO®
I-LtrII, - DNA
pCR4mtcob
pCR4mtrnsEx
pSKmtrnsORFis
pCR4mtrnsExIn
Time (min)
60 0 15 30 45 60 0 15 30 45 60 0 15 30 45 60 0 15 30 45 60
C


mt cob gene
mt rns gene
50 bp-IS of ORF-50 bp
mt 5' exon-intron+ORF-3' exon
pCR®4-TOPO®
pCR®4-TOPO®
pBluescript II SK+
pCR®4-TOPO®
I-LtrII, - DNA
pCR4mtcob
pCR4mtrnsEx
pSKmtrnsORFis
pCR4mtrnsExIn
Time (min)
60 0 15 30 45 60 0 15 30 45 60 0 15 30 45 60 0 15 30 45 60
D

To map the cleavage site, a 248-bp PCR product of the substrate was amplified with the same primers used to generate the sequencing ladder. Cleavage products were 5' end-labelled on either the top (sense) or bottom (antisense) strand and these products were denatured and separated alongside the sequencing ladder. The top strand comigrated with the T at position 950 (based on the numbering of positions in the *E. coli* 16S rRNA gene) in the 5' exon (Figure 5.7A), while the bottom strand comigrated with the C at position 946 (Figure 5.7B), which correspond to 2 nt and 6 nt, respectively, upstream of the intron IS (Figure 5.7C). I-LtrII binds and cleaves double-stranded DNA within the mt *rns* gene, generating the characteristic 4 nt 3' OH overhangs expected for LAGLIDADG-type HEases. A DNA sequence logo was generated for the exon sequences flanking the cleavage and intron insertion sites. The alignnment used to synthesize the logo comprised all sequences in the mt *rns* dataset (Table 5.1). DNA sequence logos may be used to examine sequence conservation; the level of sequence conservation at each position within the sequence is indicated by the overall height of the stack in the logo, while the relative frequency of the individual nucleotides is indicated by the relative heights of each symbol within the stack. The region downstream of the cleavage and intron insertion sites (that is, the 3' exon, or exon 2) is more conserved than the corresponding upstream region. Out of 13 positions downstream of the CS, 12 contain a bit score of 2.

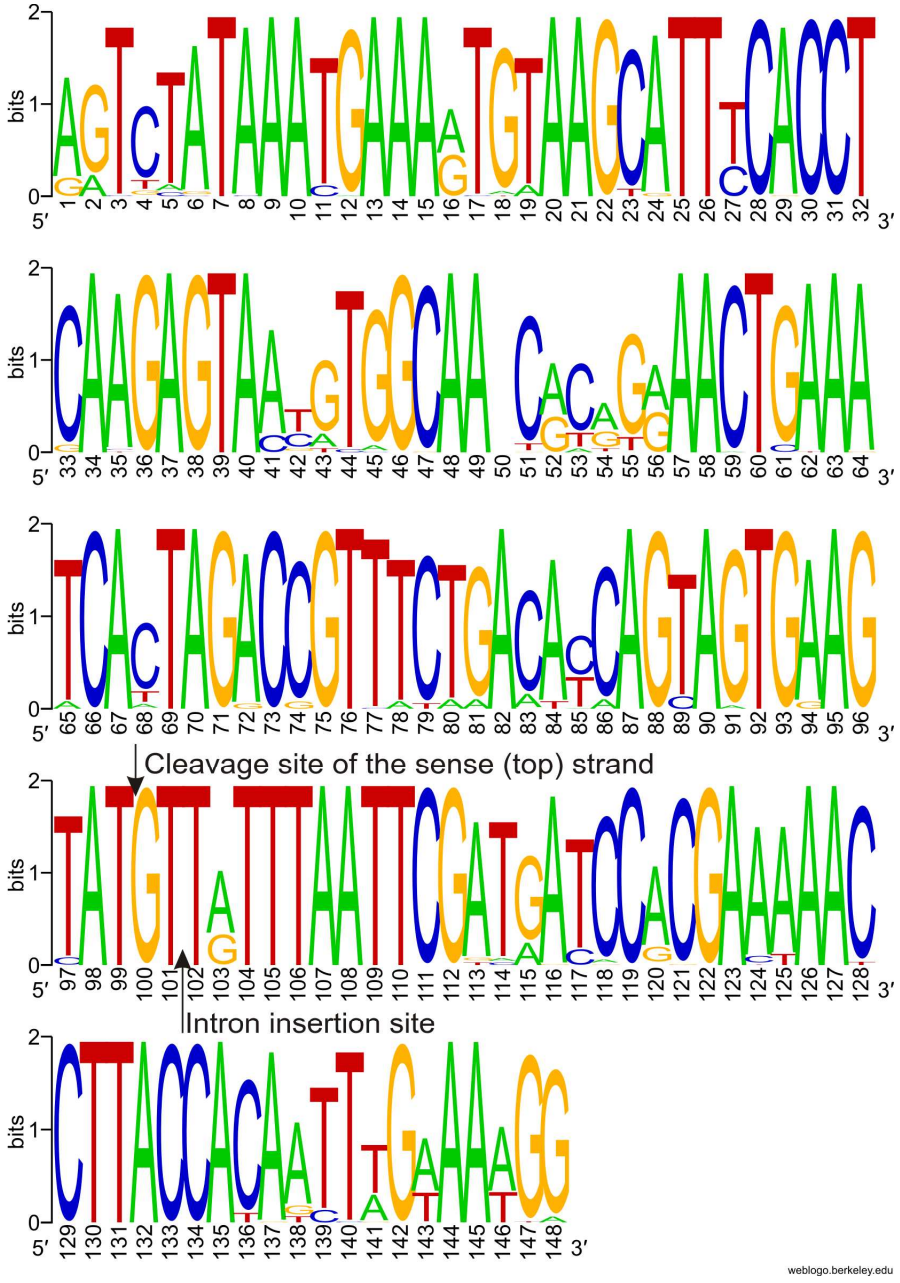
Figure 5.7. Mapping of the cleavage site of the I-LtrII LHEase within the mt *rns* gene. Shown is a representative sequencing ladder generated for the top (A) and bottom (B) DNA strands that flank the intron IS. The uncleaved product represents the 248-bp PCR amplicon that served as the template for the sequencing reaction. The cleaved product was generated by incubating this amplicon with near-native I-LtrII. The asterisk marks the nucleotide that is immediately 5' to the cut site. (C) Schematic of the region of the mt *rns* gene that flanks the intron IS and includes the cleavage site (CS) of I-LtrII. (D) DNA weblogo of the 65 sequences in the data set for the mt *rns* gene. The nucleotides shown flank the CS and intron IS.


Top strand
Uncleaved
Cleaved
G A T C
3' G C T T A A T T T A T T G T* A T G A A G T G A T G A 5'
Bottom strand
Uncleaved
Cleaved
G A T C
3' T G T A G T C A T C A C T T C* A T A C A A T 5'
A
B


CS IS
5' AGTAGTGAAGTATGTTATTTAATTCG 3'
3' TCATCACTTCATACAATAAATTAAGC 5'
CS

C

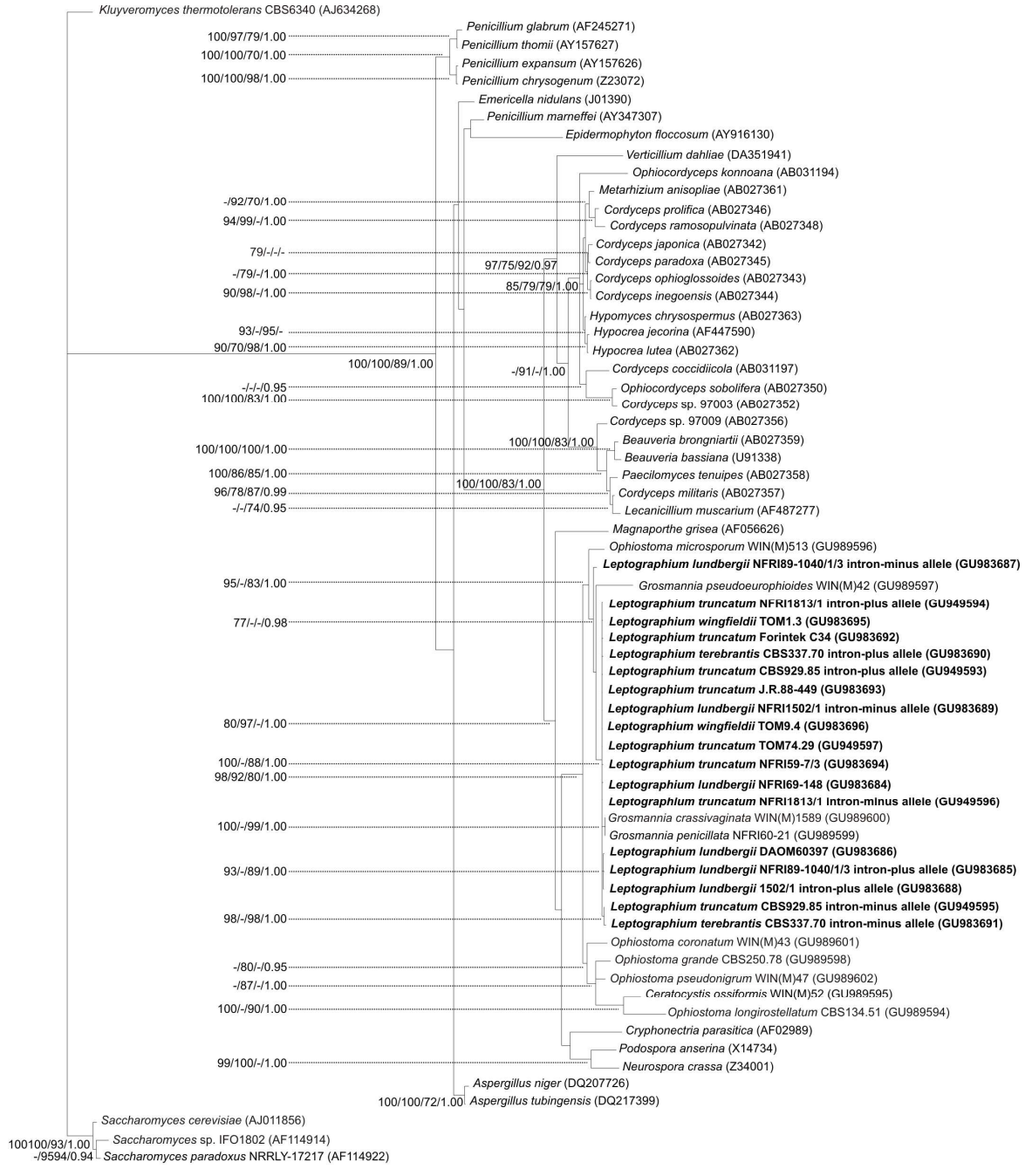
D


bits
5′
3′
Cleavage site of the sense (top) strand
Intron insertion site
weblogo.berkeley.edu

5.3.5. Phylogenetic analyses of the mitochondrial *rns* gene, intron, and LHEases sequences

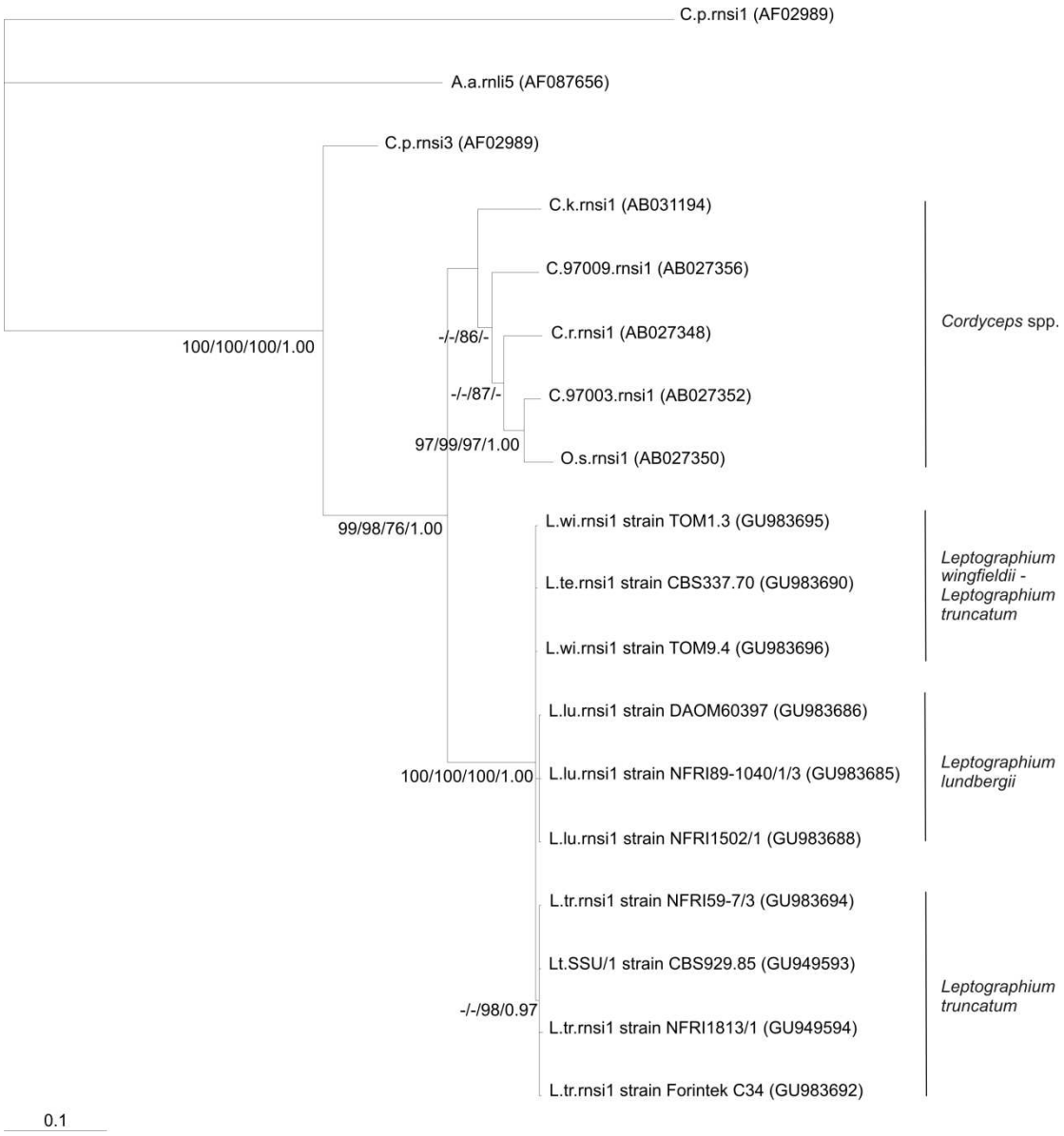
The mt *rns* sequences of intron-minus and intron-plus alleles of *Leptographium* strains were aligned with sequences from representatives of the *Sordariomycetes*, which include *C. parasitica*, *N. crassa*, and *Cordyceps* spp. and members of the *Saccharomycetales*, which include *Saccharomyces* spp. and *Kluyveromyces thermotolerans*; the latter species was used as the outgroup in phylogenetic analysis (Monteiro-Vitorello et al., 2009). The mt *rns* gene of *Leptographium* spp. groups with sequences obtained from members of teleomorphic (sexual) genera *Grosmannia*, *Ophiostoma*, and *Ceratocystis*, forming a clade with numerous polytomies (Figure 5.8). Within this complex, intron-minus alleles from *L. truncatum* strain CBS929.85 and *L. terebrantis* strain CBS337.70 are clustered within a single subclade, which received a strong posterior probability value (1.00) and moderate (88 %) to strong (100 %) support from maximum likelihood and NJ analyses, respectively. Intron-plus alleles from *L. lundbergii* strains DAOM60397, NFRI89-1040/1/3, and NFRI1502/1also form a subclade with a strong posterior probability value (1.00) and strong support (98 %) from maximum likelihood and NJ analyses. Intron-minus and intron-plus alleles from the remaining strains of *L. terebrantis*, *L. truncatum*, and *L. wingfieldii* grouped with members of *Grosmannia*, *Ophiostoma*, and *Ceratocystis* (*Ceratocystis ossiformis* should be transferred to the genus *Ophiostoma*). The clade most closely-related to this one is composed of *C. parasitica*, *P. anserina*, and *N. crassa*. The *rns* gene sequences from species of *Cordyceps*, of which some members contain an mS952-like group II intron/LHEG composite element are more distantly-related.

Figure 5.8. Phylogenetic analysis of the mt *rns* gene sequence in strains of *Grosmannia*, *Leptographium*, and related fungal taxa. Branch lengths were determined using the Bayesian consensus outfile. Values at the nodes were determined using algorithms in NJ/DNA PARS/Tree Puzzle/Mr Bayes programs. "-" indicates the node is absent or the posterior probability or bootstrap value is not well supported (a posterior probability value of less than 0.90 for Bayesian analysis and bootstrap values less than 70 % for parsimony and maximum likelihood analyses). *Ceratocystis ossiformis* should be transferred to the genus *Ophiostoma*.


Kluyveromyces thermotolerans CBS6340 (AJ634268)
100/97/79/1.00
Penicillium glabrum (AF245271)
100/100/70/1.00
Penicillium thomii (AY157627)
Penicillium expansum (AY157626)
100/100/98/1.00
Penicillium chrysogenum (Z23072)
Emericella nidulans (J01390)
Penicillium marneffei (AY347307)
Epidermophyton floccosum (AY916130)
Verticillium dahliae (DA351941)
Ophiocordyceps konnoana (AB031194)
-/92/70/1.00
Metarhizium anisopliae (AB027361)
94/99/-/1.00
Cordyceps prolifica (AB027346)
Cordyceps ramosopulvinata (AB027348)
79/-/-/-
Cordyceps japonica (AB027342)
97/75/92/0.97
Cordyceps paradoxa (AB027345)
-/79/-/1.00
Cordyceps ophioglossoides (AB027343)
85/79/79/1.00
90/98/-/1.00
Cordyceps inegoensis (AB027344)
Hypomyces chrysospermus (AB027363)
93/-/95/-
Hypocrea jecorina (AF447590)
90/70/98/1.00
Hypocrea lutea (AB027362)
100/100/89/1.00
-/91/-/1.00
Cordyceps coccidiicola (AB031197)
-/-/-/0.95
Ophiocordyceps sobolifera (AB027350)
100/100/83/1.00
Cordyceps sp. 97003 (AB027352)
Cordyceps sp. 97009 (AB027356)
100/100/83/1.00
100/100/100/1.00
Beauveria brongniartii (AB027359)
Beauveria bassiana (U91338)
100/86/85/1.00
Paecilomyces tenuipes (AB027358)
100/100/83/1.00
96/78/87/0.99
Cordyceps militaris (AB027357)
-/-/74/0.95
Lecanicillium muscarium (AF487277)
Magnaporthe grisea (AF056626)
Ophiostoma microsporum WIN(M)513 (GU989596)
Leptographium lundbergii NFRI89-1040/1/3 intron-minus allele (GU983687)
95/-/83/1.00
Grosmannia pseudoeurophioides WIN(M)42 (GU989597)
Leptographium truncatum NFRI1813/1 intron-plus allele (GU949594)
77/-/-/0.98
Leptographium wingfieldii TOM1.3 (GU983695)
Leptographium truncatum Forintek C34 (GU983692)
Leptographium terebrantis CBS337.70 intron-plus allele (GU983690)
Leptographium truncatum CBS929.85 intron-plus allele (GU949593)
Leptographium truncatum J.R.88-449 (GU983693)
Leptographium lundbergii NFRI1502/1 intron-minus allele (GU983689)
80/97/-/1.00
Leptographium wingfieldii TOM9.4 (GU983696)
Leptographium truncatum TOM74.29 (GU949597)
100/-/88/1.00
Leptographium truncatum NFRI59-7/3 (GU983694)
98/92/80/1.00
Leptographium lundbergii NFRI69-148 (GU983684)
Leptographium truncatum NFRI1813/1 Intron-minus allele (GU949596)
100/-/99/1.00
Grosmannia crassivaginata WIN(M)1589 (GU989600)
Grosmannia penicillata NFRI60-21 (GU989599)
Leptographium lundbergii DAOM60397 (GU983686)
93/-/89/1.00
Leptographium lundbergii NFRI89-1040/1/3 intron-plus allele (GU983685)
Leptographium lundbergii 1502/1 intron-plus allele (GU983688)
Leptographium truncatum CBS929.85 intron-minus allele (GU949595)
98/-/98/1.00
Leptographium terebrantis CBS337.70 intron-minus allele (GU983691)
Ophiostoma coronatum WIN(M)43 (GU989601)
-/80/-/0.95
Ophiostoma grande CBS250.78 (GU989598)
-/87/-/1.00
Ophiostoma pseudonigrum WIN(M)47 (GU989602)
Ceratocystis ossiformis WIN(M)52 (GU989595)
100/-/90/1.00
Ophiostoma longirostellatum CBS134.51 (GU989594)
Cryphonectria parasitica (AF02989)
Podospora anserina (X14734)
99/100/-/1.00
Neurospora crassa (Z34001)
Aspergillus niger (DQ207726)
100/100/72/1.00
Aspergillus tubingensis (DQ217399)
Saccharomyces cerevisiae (AJ011856)
100100/93/1.00
Saccharomyces sp. IFO1802 (AF114914)
-/9594/0.94
Saccharomyces paradoxus NRRLY-17217 (AF114922)
0.1

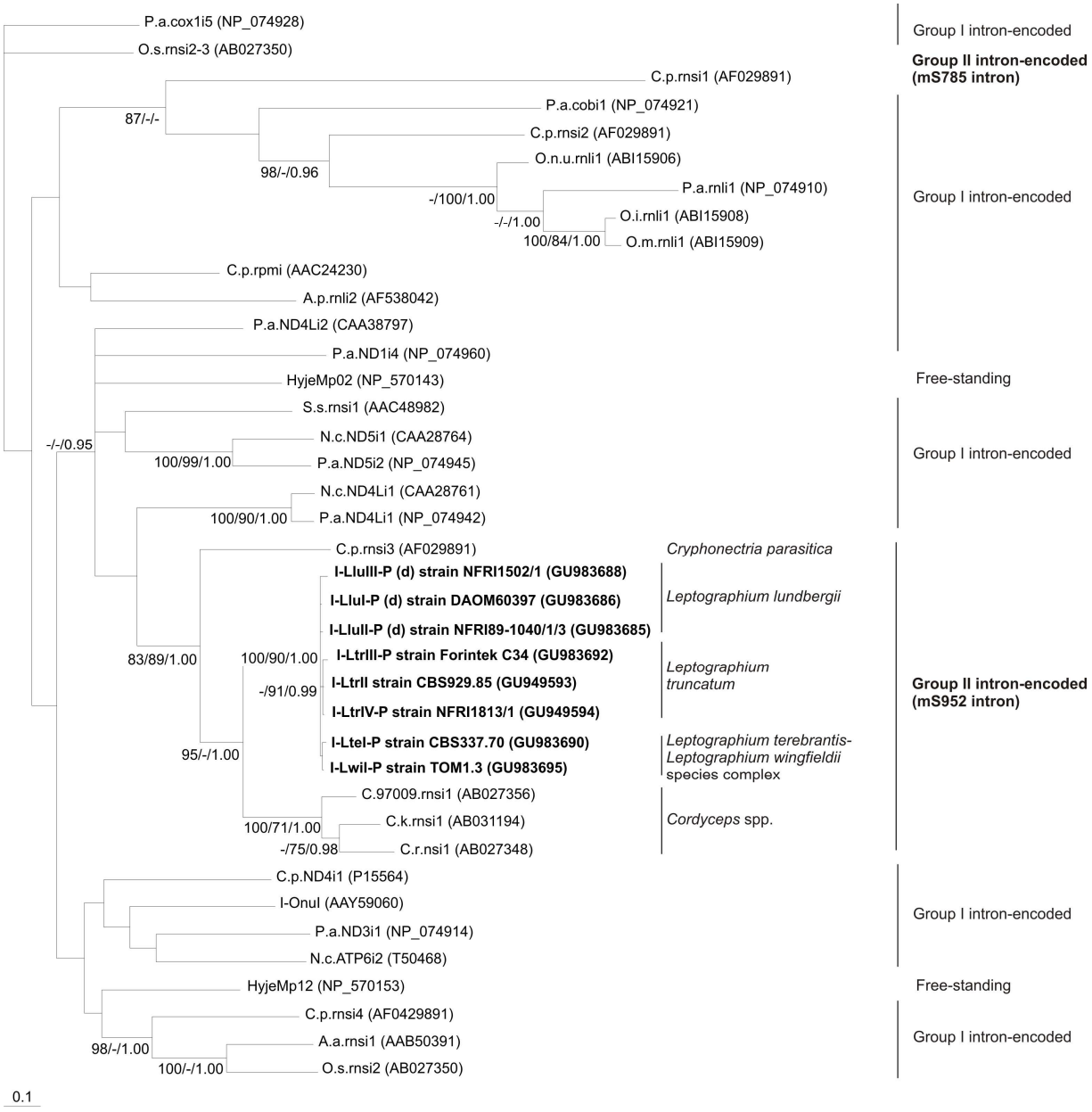
For the phylogenetic analysis of the intron, sequences from the putative start and stop codons were removed; the putative ORF sequences are more rapidly evolving than the flanking introns and create ambiguity in the alignment. Phylogenetic analysis of the mt *rns* group II intron sequence (Figure 5.9) indicates that the mt *rns* intron (I) 1 in members of *Leptographium* spp. forms a single clade, receiving strong (100 %) bootstrap and posterior probability (1.00) supports, and the topology of the tree shows that the arrangement of the intron sequences resembles that of the host organism (see Figure 3.3); that is, three smaller clades are formed comprising introns found in the *L. wingfieldii*-*L. terebrantis* species complex, *L. lundbergii*, and *L. truncatum*. However, only the clade composed of the *L. truncatum* intron sequences received support from maximum likelihood (98 %) and Bayesian (0.97) analyses. The *Leptographium* intron sequences are most closely-related to the introns found within *Cordyceps* spp. rather than intron 3 (or, the mS952 intron) of *C. parasitica*, in contrast to the topology observed in the phylogenetic tree of host gene (Figure 5.8).

Figure 5.9. Phylogenetic analysis of the mt *rns* group II intron sequence in strains of *Grosmannia*, *Leptographium*, and related fungal taxa. Sequences from the putative start codon to the putative stop codon were removed from the sequence alignment. Branch lengths were determined using the Bayesian consensus outfile. Values at the nodes were determined using algorithms in NJ/DNA PARS/Tree Puzzle/Mr Bayes programs. "-" indicates the node is absent or the posterior probability or bootstrap value is not well supported (a posterior probability value of less than 0.95 for Bayesian analysis and bootstrap values less than 70 % for parsimony and maximum likelihood analyses). Nomenclature of the intron follows that used by Monteiro-Vitorello et al. (2009) and Mullineux et al. (2010).


C.p.rnsi1 (AF02989)
A.a.rnli5 (AF087656)
C.p.rnsi3 (AF02989)
C.k.rnsi1 (AB031194)
C.97009.rnsi1 (AB027356)
-/-/86/-
C.r.rnsi1 (AB027348)
100/100/100/1.00
-/-/87/-
C.97003.rnsi1 (AB027352)
97/99/97/1.00
O.s.rnsi1 (AB027350)
Cordyceps spp.
99/98/76/1.00
L.wi.rnsi1 strain TOM1.3 (GU983695)
L.te.rnsi1 strain CBS337.70 (GU983690)
L.wi.rnsi1 strain TOM9.4 (GU983696)
Leptographium wingfieldii - Leptographium truncatum
L.lu.rnsi1 strain DAOM60397 (GU983686)
100/100/100/1.00
L.lu.rnsi1 strain NFRI89-1040/1/3 (GU983685)
L.lu.rnsi1 strain NFRI1502/1 (GU983688)
Leptographium lundbergii
L.tr.rnsi1 strain NFRI59-7/3 (GU983694)
Lt.SSU/1 strain CBS929.85 (GU949593)
-/-/98/0.97
L.tr.rnsi1 strain NFRI1813/1 (GU949594)
L.tr.rnsi1 strain Forintek C34 (GU983692)
Leptographium truncatum
0.1

Phylogenetic analysis of the LHEase amino acid sequence indicates that mS952-like group II-encoded LHEases within the mt *rns* gene of *Leptographium* and *Cordyceps* spp. form a single clade; the LHEase encoded within I3 (mS952) from the mt *rns* gene of *C. parasitica* is more distantly-related. The topology of the clade encompassing species of *Leptographium* corresponds to that observed with the phylogenetic trees of the intron (Figure 5.8) and the ITS (Figure 3.3) sequences. In *L. truncatum* strain NFRI1813/1 and *L. lundbergii* strains DAOM60397, NFRI89-1040/1/3, and NFRI1502/1 insertions after the first LAGLIDADG motif appear to have split the ORF in between the two motifs. The term "d" refer to the edited sequence in which the sequences for N- and C-terminal fragments were joined and amino acid residues that were not in alignment with closely-related taxa were replaced with gaps (-). The original alignment, showing both fragmented and ligated putative LHEases is shown in Supplementary Figure 7.10. A second group II intron/LHEG composite element was identified at position 785 (mS785, or intron 1) of the mt *rns* gene of *C. parasitica*. The phylogeny of the LHEase amino acid sequence indicates that this LHEase is unrelated to that associated with the mS952 intron; instead it is distantly related to LHEases associated with SSU and LSU introns in *Ophiostoma* and intron 2 of *C. parasitica*, as well as intron 1 of the intron 1 of the *cob* gene of *P. anserina*.

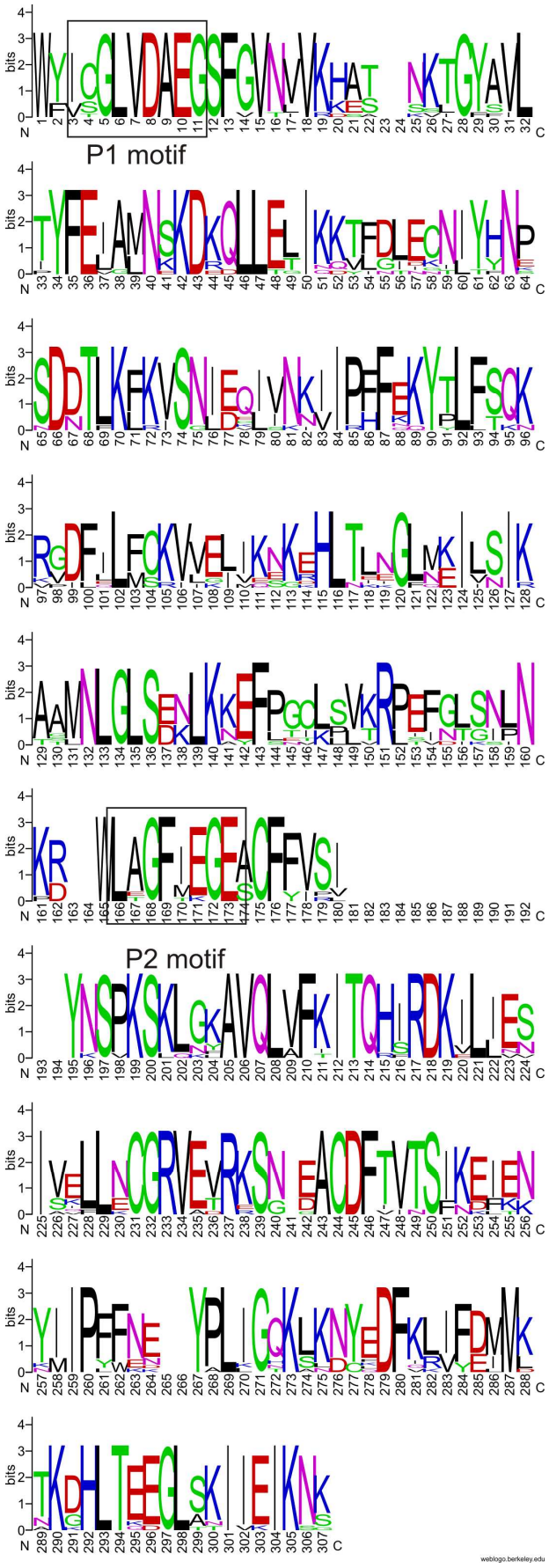
Figure 5.10. Phylogenetic analysis of LHEase amino acid sequences in strains of *Grosmannia*, *Leptographium*, and related fungal taxa. Branch lengths were determined using the Bayesian consensus outfile. Values at the nodes were determined using algorithms in DNA PARS/Tree Puzzle/Mr Bayes programs. "-" indicates the node is absent or the posterior probability or bootstrap value is not well supported (a posterior probability value of less than 0.95 for Bayesian analysis and bootstrap values less than 70 % for parsimony and maximum likelihood analyses). LHEase sequences encoded within *Leptographium* spp. are indicated in bold. The "(d)" indicates that the full-length amino acid sequence was obtained by ligating the N- and C-terminal fragments of the LHEase. Nomenclature of the intron follows that used by Monteiro-Vitorello et al. (2009), Sethuraman et al. (2009a) and Mullineux et al. (2010).


P.a.cox1i5 (NP_074928)
O.s.rnsi2-3 (AB027350)
C.p.rnsi1 (AF029891)
P.a.cobi1 (NP_074921)
C.p.rnsi2 (AF029891)
O.n.u.rnli1 (ABI15906)
P.a.rnli1 (NP_074910)
O.i.rnli1 (ABI15908)
O.m.rnli1 (ABI15909)
C.p.rpmi (AAC24230)
A.p.rnli2 (AF538042)
P.a.ND4Li2 (CAA38797)
P.a.ND1i4 (NP_074960)
HyjeMp02 (NP_570143)
S.s.rnsi1 (AAC48982)
N.c.ND5i1 (CAA28764)
P.a.ND5i2 (NP_074945)
N.c.ND4Li1 (CAA28761)
P.a.ND4Li1 (NP_074942)
C.p.rnsi3 (AF029891)
I-LlulII-P (d) strain NFRI1502/1 (GU983688)
I-LlulI-P (d) strain DAOM60397 (GU983686)
I-LlulI-P (d) strain NFRI89-1040/1/3 (GU983685)
I-LtrIII-P strain Forintek C34 (GU983692)
I-LtrII strain CBS929.85 (GU949593)
I-LtrIV-P strain NFRI1813/1 (GU949594)
I-LteI-P strain CBS337.70 (GU983690)
I-LwiI-P strain TOM1.3 (GU983695)
C.97009.rnsi1 (AB027356)
C.k.rnsi1 (AB031194)
C.r.nsi1 (AB027348)
C.p.ND4i1 (P15564)
I-OnuI (AAY59060)
P.a.ND3i1 (NP_074914)
N.c.ATP6i2 (T50468)
HyjeMp12 (NP_570153)
C.p.rnsi4 (AF0429891)
A.a.rnsi1 (AAB50391)
O.s.rnsi2 (AB027350)
87/-/-
98/-/0.96
-/100/1.00
-/-/1.00
100/84/1.00
-/-/0.95
100/99/1.00
100/90/1.00
83/89/1.00
100/90/1.00
-/91/0.99
95/-/1.00
100/71/1.00
-/75/0.98
98/-/1.00
100/-/1.00
0.1
Group I intron-encoded
Group II intron-encoded (mS785 intron)
Group I intron-encoded
Free-standing
Group I intron-encoded
Cryphonectria parasitica
Leptographium lundbergii
Leptographium truncatum
Leptographium terebrantis-Leptographium wingfieldii species complex
Cordyceps spp.
Group II intron-encoded (mS952 intron)
Group I intron-encoded
Free-standing
Group I intron-encoded

5.3.6. Sequence characterization of I-LtrII using logos

A sequence logo was constructed for the 13 LHEase amino acid sequences encoded by group II introns in the mt *rns* gene (Figure 5.11). The LAGLIDADG motifs, P1 and P2, are enclosed in boxes. The P1 and P2 motifs, which form the catalytic core are more conserved than the intervening region, which are involved in recognizing and binding to the DNA target site.

Figure 5.11. Sequence logo of the 13 LAGLIDADG homing endonucleases encoded by group II introns within the mt *rns* gene. The two LAGLIDADG motifs are enclosed in boxes.


P1 motif
P2 motif
bits
weblogo.berkeley.edu

5.4. Discussion

5.4.1. Group II introns with LHEGs

The discovery of novel group II introns that encode ORFs usually associated with group I introns (Michel and Ferat, 1995; Toor and Zimmerly, 2002) was intriguing; the presence of LHEGs within group II introns could represent the evolution of a new type of composite mobile element or, less interestingly, might simply reflect the existence of yet another genomic niche within which the promiscuous LHEGs may persist as phenotypically silent insertions.

Sequence analysis of group II introns/LHEGs in *Leptographium* spp. revealed two forms of these composite elements, based on sequence characteristics of the intron and on the position of the putative start codon of the intronic ORF (Table 5.4). The intron identified in strains of *L. lundbergii* and *L. truncatum* is 925 nt and has a % G+C of 29.0-29.1, and the putative start codon of the ORF occurs after intron position 685. The intron found in strains of *L. wingfieldii* and *L. terebrantis*, members of the *G. aurea-L. wingfieldii-L. terebrantis* species complex is 961 nt and has a slightly higher % G+C (30.4-30.5) and the putative start codon follows intron position 721.

In this study, the LAGLIDADG protein encoded by a group II intron was purified and submitted to biochemical assays in order to determine: (i) if it is required to assist the group II ribozyme in splicing; that is, it acts as a maturase, as has been seen for several LHEases associated with group I introns (Ho & Waring, 1999; Bassi et al., 2002; Longo et al., 2005) and (ii) if it has DNA cleavage activity, and if that is the case, whether its cleavage site is situated at the proximity of the intron IS, in keeping with the possibility that the intron is actually mobilized by the LHEG.

5.4.2. The Lt.SSU/1 intron self-splices *in vitro*

One of the primary goals of this study was to functionally characterize the group II intron in the mt *rns* gene of *L. truncatum* strain CBS929.85. Ribozyme activity was examined using two versions of the intron, one in which the ORF sequence had been deleted and the other in which the ORF sequence was present. Several organellar group II introns have been shown to self-splice *in vitro*, including Sc.cob/1 in the *cob* gene and Sc.cox1/1 and Sc.cox1/5γ in the gene encoding cytochrome oxidase subunit 1 in *Saccharomyces cerevisiae* (reviewed in Michel and Ferat, 1995). For these introns, however, optimal ribozyme activity was observed under elevated temperature and ionic conditions. Sc.cox1/5γ, for example, self-splices at 45 °C in the presence of 0.5 M $(NH_4)_2SO_4$ and 0.1 M Mg^{2+} (Jarrell et al., 1988). In contrast, intron 2 in the LSU rRNA gene of the brown alga *Pylaiella littoralis*, Pl.LSU/2, was shown to have an unusually low requirement for Mg^{2+}, as it retains activity at concentrations as low as 0.1 mM (Costa et al., 1997). Both versions of Lt.SSU/1 readily self-splice under moderate temperature (37 °C) and ionic (as low as 6 mM Mg^{2+}) conditions and the presence of the ORF sequence in DIII did not appear to interfere with the proper folding of the group II ribozyme. In most group II introns, the ORF is located in a looped-out region in DIV; this region is dispensable for self-splicing. In contrast, the LHEase ORF has invaded a different helical region, DIII, and genetic and biochemical studies previously indicated that this region enhances splicing efficiency (reviewed in Lehmann and Schmidt, 2003; Fedorova and Zingler, 2007; Pyle, 2010).

5.4.3. Intron-encoded I-LtrII does not enhance intron splicing

Proteins encoded by group I and group II introns have been shown to assist in the self-splicing reactions of the ribozymes so that the resulting mature RNAs can proceed with their cellular functions either as components of the ribosome or as mRNAs. This "maturase" activity is probably a derived feature that has contributed to the evolutionary success of IEPs by making them essential to the host organism (unless the latter manages to eliminate the entire intron). It may also be interpreted as an advanced stage of symbiosis in which interdependence of the ribozyme and IEP has developed beyond cooperation in mobility.

For members of the group I intron class it has been suggested that once the intron is invaded by an HEG, the encoded protein gains the ability to promote homing of the intron/HEG composite element and that eventually, the LHEase may acquire maturase activity to ensure the long term survival of the composite intron (Belfort, 2003; Gogarten & Hilario, 2006). In this study evidence for maturase activity of the group II intron-encoded LAGLIDADG protein was not detected. N-terminal His_6-tagged I-LtrII failed to bind ORF-less intron RNA precursor transcripts and near-native I-LtrII neither enhanced splicing of full-length intron RNA precursor transcripts nor rescued it under a prohibitively low (3 mM) Mg^{2+} concentration. This could suggest that the association between the intron and protein results from a relatively recent event or there was little opportunity for selection of maturase activity since, as demonstrated in this study, the group II intron appears to self-splice efficiently in the absence of protein factors.

5.4.4. I-LtrII targets the mt *rns* gene and is capable of promoting intron mobility

Determining the cleavage site of I-LtrII was important as it was supposed to provide a clue as to the possible genetic target(s) of the LHEase; namely, the mt *rns* gene or the intron sequences that flank the ORF. If the target had been the latter, then it would have suggested that while the LHEG had invaded the group II sequence, it probably did not benefit the host intron. In this case, the LHEG likely would act as a truly selfish element that fortuitously found a neutral site to invade and promotes its own mobility. However, not only is the *rns* gene a target, but I-LtrII cleaves specifically the intron-less version of the gene in the immediate vicinity (2 nt and 6 nt upstream) of the Lt.SSU/1 intron IS, as observed for other LHEases (Jurica and Stoddard, 1999; Sethuraman et al., 2009a). This leaves little doubt that the protein is capable of promoting the mobility of the group II intron. The LHEG and group II ribozyme may be regarded as having a mutually beneficial relationship in which the intron provides the LHEG with a non-deleterious IS and the LHEG, by virtue of its ENase activity, ensures the mobility of the intron and LHEG. Therefore, this novel arrangement of a LHEG within a group II intron must represent the evolution of a new category of mobile composite element.

5.4.5. Speculations on the origins and evolutionary dynamics of the group II ribozyme/LHEG composite element

Genes encoding LHEases are often associated with group I introns and inteins (self-splicing proteins), and their broad distribution in unrelated elements suggests that ORF sequences evolve independently of their host intron or intein and that splicing and mobility functions originated independently (Dalgaard et al., 1993; Loizos et al., 1994). The generation of mobile introns is described by the ENase gene

invasion hypothesis (Belfort, 2003). HEGs invade DNA sequences encoding self-splicing introns; since the intron is removed from the mature transcript, it represents a phenotypically silent site for HEG invasion. If the HEase also has maturase activity, it further benefits the intron by ensuring that the sequence is efficiently spliced out of the RNA transcript, thus reducing the negative impact of the intron (and, consequently, HEG) sequence on the host fitness. Following the HEG invasion event, the target, or homing, site recognized by the HEase is disrupted and the intron becomes resistant to self-cleavage (Loizos et al., 1994). For non-coding intron sequences, there may be little pressure for sequence conservation and the sequences that flank the HEG may begin to accumulate mutations (Haugen et al., 2005). However, if the DNA cleavage activity of the HEase happens to evolve so as to target intron-less versions of the host gene, the HEG then benefits the intron by mobilizing it and allowing it to spread in the genome or be transferred horizontally with the HEG (Loizos et al., 1994; Zeng et al., 2009).

Once the HEG has invaded and mobilized the intron, it begins the HEG life cycle (Goddard and Burt, 1999; Burt and Koufopanou, 2004); this simplified model describes the "birth" and "death" and re-birth of the HEase. The cycle begins with the HEG invasion of the DNA sequence of a self-splicing intron, mobilizing the intron, and the HEG then spreads in the population through the homing activity of the HEase. Once all possible homing sites have been invaded, the HEG is no longer under purifying selection pressure and begins to accumulate mutations and degenerate. The HEG and intron are ultimately lost from the genome, which regenerates a possible homing site, allowing the cycle to begin again.

Some HEGs also code for maturases, which are RNA-binding proteins that promote intron splicing by stabilizing the ribozyme's tertiary structure. Following the

characterization of the bI3 maturase, Longo and co-workers (2005) proposed an evolutionary model for LHEase maturases in which the LHEase degenerates and loses DNA cleavage activity and evolves maturase activity. Consequently, the LHEase becomes essential for RNA splicing, and ultimately host fitness, and in this way the LHEG can escape from the homing life cycle and become fixed in the genome.

A recent study by Yahara et al. (2009) demonstrated that so-called degenerated HEGs that code for maturases are relatively more advantageous over HEGs that encode ENases, although both still impose a cost on the host fitness compared to genomes that lack HEGs. The authors developed simulated models for the behaviour of HEGs in meiotic populations by incorporating the relative costs on the host fitness of a gene encoding a functional HEase, a degenerated HEG (possibly one encoding a maturase), or the absence of a HEG. They described a long-term life cycle based on autonomous periodic oscillations in the population and proposed that in the absence of horizontal transmission HEGs may persist in a population over several thousand meiotic generations. The observation that some HEGs appear to have become fixed in the genome or in the population despite the absence of horizontal transfer by coding for maturases has led to the modification of the model for the HEG life cycle (Gogarten and Hilario, 2006; Yahara et al., 2009) and to the development of the asynchronous homing cycle model (Gogarten and Hilario, 2006). In this model HEGs persist when there is a low frequency of homing such that the HEG does not become fixed in the population, although there is still purifying selection on the HEG. Moreover, various subpopulations may be at different stages of the homing cycle.

The demonstration that the Lt.SSU/1 intron efficiently self-splices in the absence of protein factors under moderate temperature and ionic conditions indicates

that both the ribozyme and protein components of this system are active and can function independently. The intron, by virtue of its efficient ribozyme activity, has provided the LHEG with a phenotypically neutral site (a "safe haven") for invasion. In turn, the LHEG benefits the intron by encoding an active LHEase that may promote the mobility of the intron.

This study describes a self-splicing group II intron that has been invaded by a LHEG, and together they form a composite element that has the potential to target the mt *rns* gene, thus behaving like group I/HEG composite elements. Group II introns with LHEGs are so far confined to a single ribozyme structural subgroup (IIB1) and just three sites in mt rRNA genes of fungi (Toor and Zimmerly, 2002; Monteiro-Vitorello et al., 2009; Mullineux et al., 2010). This work demonstrates how novel mobile elements can evolve by shuffling components of unrelated, pre-existing mobile genetic units.

CHAPTER 6. CONCLUSIONS

6.1. Major findings of the thesis

6.1.1. Evolutionary dynamics of the ITS1-5.8S-ITS2 nuclear rDNA region

Amongst the eukaryotes, internal transcribed spacer (ITS) regions vary greatly in size and sequence (reviewed in Hausner and Wang, 2005). The analysis of ITS sequences, including the use of RNA secondary structural features to guide the alignment of ambiguous regions, has become an important tool in phylogenetic analyses (Coleman, 2003, 2007, 2009; Goertzen et al., 2003; Young and Coleman, 2004; Won and Renner, 2005; Schultz et al., 2006; Aguilar and Sánchez, 2007; Müller et al., 2007; Krüger and Gargas, 2008; Ryberg et al., 2008; Seibel et al., 2008;Wolf et al., 2008; Schultz and Wolf, 2009). One caveat to the wide-spread application of ITS segments in phylogenetic and taxonomic studies is the potential influence on the rate and types of nucleotide changes arising from the possible co-evolution of the two spacer regions and compensatory slippage (interdependence of sites within the spacer regions). This is problematic, as phylogenetic algorithms assume that every nucleotide position evolves independently. Previous studies have shown that the ITS RNA structures might be more conserved than the actual nucleotide sequences (Joseph et al., 1999; Gottschling and Plottner, 2004; Schultz et al., 2005; Hausner and Wang, 2005), suggesting a functional constraint, connected with the RNA fold, is influencing the evolution of these sequences. While convergence and other factors do not strictly affect the use of ITS sequences as a bar code, the utility of the ITS region for developing probes to rapidly identify organisms (El Karkouri et al., 2007; Landis and Gargas, 2007) may be affected by the observation noted in Chapter 3 that the ITS2

segment is more conserved than ITS1 in this group of ophiostomatoid fungi. Taxonomically-distinct organisms may share identical ITS2 segments, which could lead to erroneous identification of unknown strains.

The observation that replication slippage may promote compensatory events that maintain thermodynamically favorable RNA folds (either due to convergent evolution or functional constraints) may pose difficulties regarding the use of secondary structure to ameliorate ITS sequence alignments. Positional homology of a nucleotide within the primary sequence may not correspond to the placement of this base within an RNA fold. Slippage of the RNA strand to compensate for an event at the primary sequence level (due to insertions or deletions, as an example) could move a base into a different structural location from its original position within the fold. Alignments guided by structural criteria may essentially align non-homologous positions; that is, positions that happen to occupy the same "site" within a fold but do not share a common ancestral sequence. It might be more beneficial to analyze alignments for evidence of positions that are prone to or the result of replication slippage events. Compensatory mutations may then be more readily recognized and down-weighted in phylogenetic analyses. Although other genes, such as *cox1* (*COI*, Seifert et al., 2009) or *nad6* (*ND6*, Santamaria et al., 2009) along with rRNA-coding genes may be used in the future for DNA barcoding studies, the utility of the multicopy rDNA sequences, with regards to PCR primer design and amplification, will ensure the continued use of ITS sequences for phylogenetic studies. This in particular applies to situations in which only limited DNA amounts can be recovered, such as studies using herbarium specimens or environmental samples. Gaining a better understanding of the mode of fungal ITS sequence evolution, the nature of the GC bias, and the constraints on the ITS sequences may assist in developing improved

evolutionary models for phylogenetic analysis using ITS sequences. Finally, this work highlights the mechanisms that maintain the pan-eukaryotic ITS2 structure (Coleman, 2007) and the hairpin type structures commonly observed the fungal ITS1 region.

6.1.2. Distribution of insertions within mitochondrial DNA in members of the fungal genus *Leptographium*

The primary objective of the PCR screen was to identify, based on size and phylogenetic distribution, potentially novel insertion elements within the mt *rns*, *rnl*, and *cob* genes in members of *Leptographium*. The PCR-based survey and subsequent sequencing of amplicons identified a novel type of composite genetic element in the mt *rns* gene: group II introns encoding LAGLIDADG-type homing endonucleases (LHEases) rather than reverse transcriptases. Superimposing intron distribution onto the host phylogenetic tree suggests that group II introns and their LHEGs appear to be transmitted vertically amongst strains of *L. lundbergii*; however, its mixed distribution amongst members of *L. truncatum* and the *G. aurea-L. wingfieldii-L. terebrantis* species complex suggests that there is intra- and inter-species transfer of these elements, although rapid and random gain and loss may also contribute to the stochastic distribution of insertions in the mt *rns* gene.

Insertions within the U11 region of the mt *rnl* gene were found in all tested strains and the size variability, ranging from aproximately 1.6 to 2.9 kb, suggests that this region is a hot spot for invasion by different types of introns and/or putative HEGs. In all members of *L. procerum* and of the *G. aurea-L. wingfieldii-L. terebrantis* species complex, a 1.6-kb insert was detected in the U11 region of the *rnl* gene, suggesting that this element may have become fixed in those populations and is inherited vertically.

The mt *cob* gene appears to have a mosaic arrangement and the observation of both heteroplasmy and stochastic distribution of insertions within taxa is strongly suggestive of horizontal transmission, likely involving several different introns and/or HEGs.

One of the purposes of the PCR survey and phylogenetic study of the host organisms was to assess whether or not the distribution of inserts could be used as a molecular marker for the rapid identification of potential plant pathogens; this would permit rapid species identification based on the size of the PCR amplicon and would bypass the necessity of first amplifying and then sequencing the nuclear ITS region, and possibly other gene regions. The mixed distribution of insertions within the mt *rns*, *rnl*, and *cob* genes suggests that there is intra- and inter-species transfer of these elements, either due to horizontal transmission or rapid and random gain/loss, thus, eliminating the possibility of using optional genetic elements molecular markers. These observations clearly indicate that the distribution of optional genetic elements is not sufficiently stable for use as a marker in species/strain identification of members of *Leptographium.*

6.1.3. Group II introns and their encoded LAGLIDADG homing endonucleases within the mitochondrial *rns* gene

In this thesis a self-splicing group II intron that has been invaded by a LHEG is described, and together they form a composite element that has the potential to target the mt *rns* gene, thus behaving like group I/HEG composite elements. Therefore, we may be witnessing the evolution of a potentially new combination of mobile element, even though based on currently available data, the LAGLIDADG-type ORFs encoded by group II introns have only a limited distribution, being

restricted to just three sites in mt rRNA genes of fungi and the IIB1 intron structural class (Toor and Zimmerly, 2002; Monteiro-Vitorello et al., 2009; Mullineux et al., 2010). The intron-encoded protein, I-LtrII, was not found to bind intron RNA or ehance intron splicing, and the observation that the intron, Lt.SSU/1, self-splices efficiently in the absence of protein factors suggests that there may not be strong selection pressure on the LHEG to evolve maturase activity. The demonstration that I-LtrII cleaves the mt *rns* gene in the vicinity of the intron insertion site (2 nt and 6 nt upstream) strongly suggests that I-LtrII promotes the mobility of the intron and LHEG sequences. This work demonstrates how novel mobile elements can evolve by shuffling components of unrelated, pre-existing mobile genetic units.

6.1.4. Proposing a model describing the life cycle of group II intron/ LAGLIDADG homing endonuclease genes amongst *Leptographium* species

To understand the transmission of group II intron/LHEG composite elements amongst *Leptographium* spp., a reliable phylogenetic tree of the host organisms was essential. Evolutionary relationships of *Leptographium* spp. were inferred using the nuclear ITS region and to generate a robust sequence alignment conserved regions in the DNA sequence and RNA secondary structure, as well as regions that are potentially prone to strand slippage, were used to guide the alignment of ambiguous regions (Chapter 3).

To correlate intron distribution with the evolutionary relationships of the host organism, data obtained from the intron survey were superimposed onto this phylogenetic tree. Based on the intron survey presented in Chapter 4 and subsequent sequence analysis of selected amplicons, it is clear that group II intron/LHEG composite elements are present in all tested strains of *L. lundbergii*; this is strongly

indicative of transmission of the element by vertical descent. The distribution of the element is random amongst strains of *L. truncatum* and members of the *L. wingfieldii-L. terebrantis* species complex, which suggests that the element is spreading by horizontal transmission, although the stochastic distribution could be also explained by random gain or loss of the element. Horizontal transmission throughout a population and heteroplasmy (intron-plus and intron-minus alleles within the same organism) can result from transient hyphal fusion events (anastomosis) that allow for the transfer of mitochondria. Mitochondria, in turn, can also fuse, allowing for genetic recombination events to occur. All introns that were sequenced have been invaded by LHEGs; that is, no ORF-less introns were found, suggesting that these composite elements could be mobile as a unit.

The biochemical characterization of the *L. truncatum* group II intron, Lt.SSU/1, and its encoded protein, I-LtrII, demonstrated that both components are catalytically active and the results obtained in Chapter 5 strongly support the suggestion that I-LtrII may promote mobility of the composite element. Combining data from all three studies (Chapters 3 to 5) has led to the development of a model describing the life cycle of group II intron-encoded LHEases in *Leptographium* spp. (Figure 6.1).

Taking into consideration the cycles proposed by Goddard and Burt (1999) and Burt and Koufopanou (2004), the LHEG may have occupied all homing sites due to the efficient cleavage activity of the LHEase. It is tantalizing to speculate that during the degeneration phase described by the Goddard and Burt model, the LHEG may have accumulated mutations that lead to the recognition of novel target sites, rather than degeneration and loss of the ORF. This would likely re-initiate the spreading phase of the life cycle, allowing the LHEG to escape the degeneration

phase and, at the same time, promote mobility of the intron sequence. The following observations support the suggestion that the LHEG encoded by the group II intron is in a second phase of spreading: (i) the I-LtrII ENase is active and efficiently cleaves the exon sequences and (ii) the host organism is heteroplasmic; that is, there are intron-minus alleles present alongside intron-plus alleles, suggesting that there are still potential intact homing sites available.

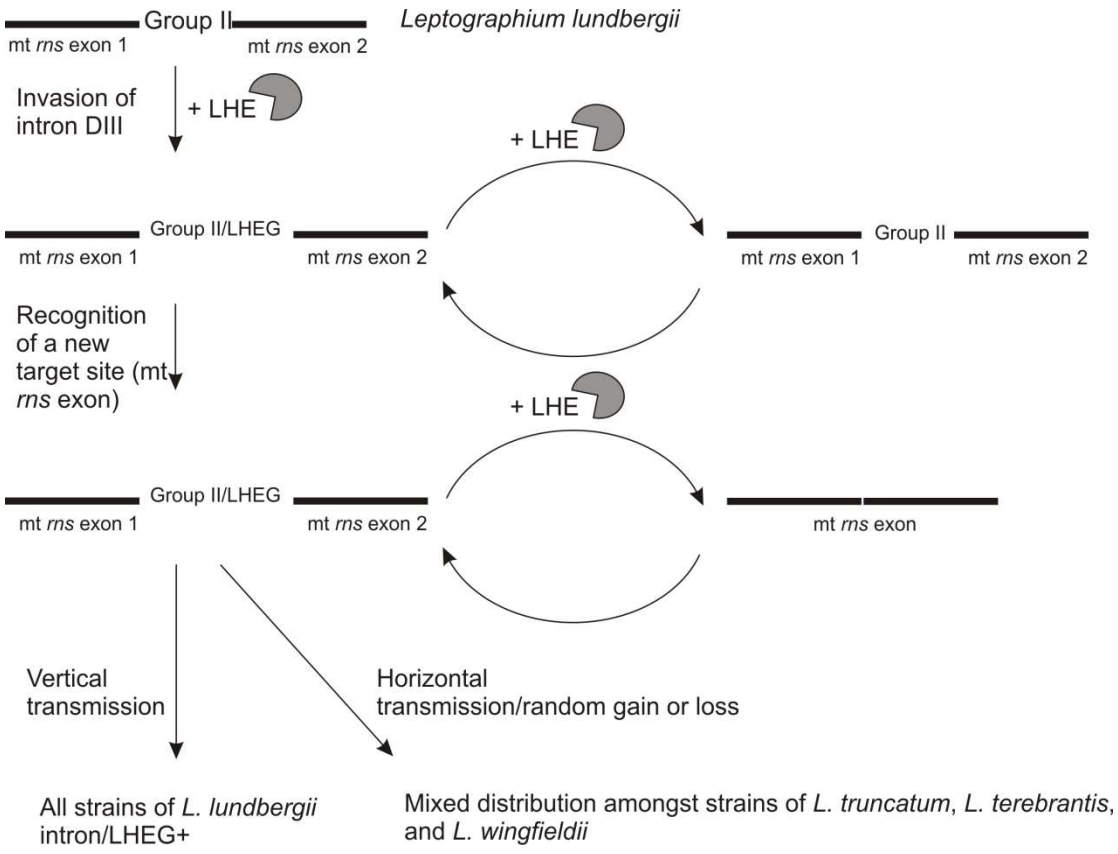
In terms of LHEGs encoded within group II introns in the mt *rns* gene of members of *Leptographium* spp., it is possible that the LHEG may have originally targeted intron sequences in DIII. One could speculate that after the initial spreading phase of the LHEG into DIII of other group II introns, it began to accumulate mutations, and rather than causing degeneration of the ORF, these mutations fortuitously led to the recognition of a novel target site, namely the mt *rns* exon sequences near intron insertion site. Given that the intron is present in all strains of *L. lundbergii*, it is possible that the intron and/or composite element was first acquired by a strain of *L. lundbergii*. The original target site of the LHEG was probably intron DIII, and the LHEG spread into DIII of all available group II introns, which could explain the observation that no ORF-less group II introns were found. Another possibility is that the LHEG invaded DIII in one group II intron and then the ORF sequence changed over time and evolved new specificities such as binding/cleavage of the exon sequence. This composite element could have successfully colonized other mt *rns* sequences.

Overall the results of this study are consistent with the "endonuclease-gene invasion" hypothesis (Belfort and Roberts, 1997; Belfort, 2003) which argues that HEGs in the course of their evolution have invaded self-splicing elements numerous times. The observation that different LAGLIDADG ORFs exist within group II

introns provides more evidence on the invasive nature of HEGs. The *Leptographium rns* group II introns and their ORFs are phylogenetically allied to similar group II introns inserted at the same *rns* position in species of *Cryphonectria* and *Cordycep*. The origin of this intron may have been the invasion of an ORF-less group II intron by a LHEG. The phylogenetic trees for these species based on rDNA data would suggest that this intron (mS952) has a spotty distribution, suggestive of lineages gaining the intron/LHEG combination via horizontal gene transfer and that in some lineages, such as *L. lundbergii*, this composite intron was vertically transmitted. Moreover, on some occasions the LHEase ORF accumulated frame shift mutations or the intron was lost.

An alternate explanation is that in unrelated lineages the *rns* group II intron was invaded by a member of the same LAGLIDADG family, although this explanation however requires numerous evolutionary events/steps and, as such, is a less parsimonious model to the one described above. Due to the rapid evolution of intron sequences and the lack of biochemical/functional data on the mS952 introns and LHEGs encoded within *Cryphonectria* and *Cordyceps* spp. one cannot speculate if models such as collaborative homing (Zeng at al., 2009) could be applied to the evolution of the composite element. Also, one cannot discount the possibility that the group II intron/LHEG composite element could have originated in a different genetic location before it was mobilized into the *rns* gene.

Figure 6.1. A model describing the spread of group II introns and LHEG composite elements in the mt *rns* gene of *Leptographium* species.


Group II
mt rns exon 1
mt rns exon 2
Leptographium lundbergii
Invasion of intron DIII
+ LHE
+ LHE
Group II/LHEG
mt rns exon 1
mt rns exon 2
Group II
mt rns exon 1
mt rns exon 2
Recognition of a new target site (mt rns exon)
+ LHE
Group II/LHEG
mt rns exon 1
mt rns exon 2
mt rns exon
Vertical transmission
Horizontal transmission/random gain or loss
All strains of L. lundbergii intron/LHEG+
Mixed distribution amongst strains of L. truncatum, L. terebrantis, and L. wingfieldii

6.2. Future prospectives

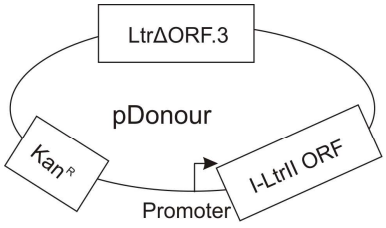
The demonstration of self-splicing activity in ORF-less and ORF-containing versions of the Lt.SSU/1 group II intron and of DNA cleavage activity of intron-encoded I-LtrII LHEase has laid the groundwork for further biochemical studies of this novel composite element. Further experimental work, carried out in an *E. coli* background, would be invaluable for determining if the composite element is truly mobile. An outline of the experimental approaches that could be used to demonstrate homing is shown in Figure 6.2. Briefly, one plasmid, referred to as pDonour, would contain an ORF-less group II intron (LtrΔORF.3) and the LHEG, consisting of the sequence optimized for expression in a recombination-competent strain of *E.coli*, upstream of an IPTG-inducible promoter. A second plasmid, pSubstrate, would be co-transformed; substrate sequence would consist of 50 bp on either site of the I-LtrII cleavage site. Following expression of I-LtrII, plasmid DNA would be harvested, purified, and linearized. A successful homing event would be evidenced by a linearized plasmid with a molecular weight larger than linearized pSubstrate but smaller than that of linearized pDonour.

Another priority is to solve the crystal structure of I-LtrII in complex with the DNA substrate. Replacement of Mg^{2+} with Ca^{2+} inhibits cleavage and should allow for crystallization of the DNA substrate-I-LtrII complex. A crystal structure is essential for identifying which amino acid residues and nucleotides are contact during DNA binding; this would facilitate future efforts in engineering I-LtrII to target specific DNA target sites. Moreover, since LHEases exhibit flexibility in target site recognition (reviewed in Chevalier et al., 2005), it would be of interest to test cleavage activity of I-LtrII against substrates with sequence differences. These

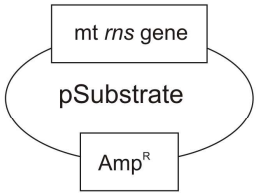
substrates could be synthesized randomly *in vitro* or could comprise sequences from *rns* genes obtained from more distantly-related organisms.

Lastly, in several strains of *L. truncatum* and *L. lundbergii*, the LHEase ORF is apparently "split" in two by a frame-shift mutation, yet the LAGLIDADG motifs appear to be intact. It would be interesting to test whether these fragments could form an active LHEase. In these experiments the N-terminal and C-terminal fragments would be synthesized separately and cloned individually. The two fragments would be expressed and purified either separately to asses if they form an active homodimer or together to determine if they form a heterodimer.

Figure 6.2. Outline of the proposed experiment used to demonstrate I-LtrII-promoted homing of the Lt.SSU/1 intron/LHEG composite element. Construct pDonour contains: the ORF-less group II intron (LtrΔORF.3), 1179 nt in length (262 nt of the 5' exon, 819 nt of the ORF-less intron, and 98 nt of the 3' exon); I-LtrII sequence, optimized for expression in a recombination-competent strain of *E. coli*, upstream of an IPTG-inducible promoter; and a kanamycin resistance gene. Construct pSubstrate contains 50 bp on either side of the cleavage site and an ampicillin resistance gene. Homing would be evidenced by a linearized plasmid with a molecular weight larger than linearized pSubstrate but smaller than that of linearized pDonour.



Schematic of constructs



pDonour construct
ORF-less intron (819 nt) + 5' exon (98 nt) + 3' exon (262 nt).
I-LtrII ORF (optimized for expression in *E. coli*) downstream of an IPTG-inducible promoter.
Kan^R marker.

pSubstrate construct
Target site within mt *rns* gene (50 bp on either side of the cleavage site.
Amp^R marker.

Transform into recombination-competent *E. coli.*
Inoculate into LB broth supplemented with kanamycin, ampicillin, and IPTG.
Isolate plasmid DNA, amplify linearize donour and substrate sequences/amplify pDonour and pSubstrate sequences, and compare sizes of expected products.

CHAPTER 7. APPENDIX

Supplementary Table 7.1. Nested primers that were designed specifically for sequencing mt genes.

Gene region/insertion	Primer name	Primer sequence (5'-3')	DNA strand orientation
mt *rns*	1100mtsr1	AGTGTTACTCATTTTTAATAGG	sense
	1183mtsr1	AACGATGAAATTTGCCTG	
	1118mtsr1	ACCTTTTATGAAAAACTG	
	174mtsr1	ATGTCTAGACTAATTCCGAG	
	468mtsr1	AAGTATACCTAACCTATGAC	
	33mtsr1(F1)	ATACCGATAAAATAGTAGGGCCAAAG	
	mtsr11000F2	ATTTCTATCTAGAGTTTAATATAAGAAG	
	mtsr69(F5)	AATGGCCTATCTTATTAAGTAAATAGTAAG	
	mtsr69(F6)	AACCCCCAGTGTTCTCAGTAATAAG	
	mtsr1435-7(F1)	TATTTAATTCGATGATCCACGAAAAACC	
	mtsr1183-20F8	AAAAAGAGTTTCCAGGATGCC	
	mtsr1183-20F9	ATTTAAACTGATACATTTGTGG	
	mtsr1183-20F14	AAGTATCTATAATTCACCTAAATC	
	mtsr1183-20F15	ACCGGAAATTGTTATATCGAAC	
	mtsr1183-20F16	AAACTGATACATTTGTCC	
	mtsr1435-3F1	AACACGAATAGGATTAGATACCC	
	mtsr1435-3F2	AGGCGAAAGCAACCTTTTATG	
	mtsr69-4F10	AACGACTGAACTGGCAACTTG	
	mtsr69-4F11	ACGAGTGTTGCACGGCTGTTTC	
	mtsr1183-20F17	ATAAGACCGGTTATGCAG	
	mtsr1183-20F18	AAATTGTTATATCGAACCTGG	
	mtsr1183-20F19	ACTGGCAACTTGGAGGAATGG	
	mtsr1274F1	ATGGTCTAATGAAAATTC	
	mtsr1274F2	AAAAGAGTTTCCAGGATGCC	
	mtsr1435-7(F1)	TATTTAATTCGATGATCCACGAAAAACC	
	1100mtsr2	AACCTAATATTGTCAAATTTC	Antisense
	1183mtsr2	AATTCTTAAGTTTTTGTCCG	
	1118mtsr2	AATAACATACTTCACTAC	
	174mtsr2	ATTAATTTAAAATCCACG	
	468mtsr2	ATCATCGAATTAAATAAG	
	69mtsr2	AGATCTTCAACATACAACC	

mt *rns*	1115mtsr2	ATTAGATCTTCAACATACAAC	Antisense
	1118mtsr2n	ATTGTGGTAAGGTTTTTCGTGG	
	1118(3)mtsr2	AATTCTAATTGTGGTAAGGTTTTTCGTGG	
	mtsr2-1183(2)	ACTCTTCCACAATTTAATAGTTCTACG	
	33mtsr2-R1	ACTAAGAGAATTTTGGGCTAGTTAAG	
	mtsr1100R2	ACCCCTGTTAAGTCAGCTATATTAATAAG	
	mstr1115R2	AATAGATGTAACAGTAAAATCACAAGC	
	mtsr1183R4	AAGTTTTTGTCCTATTAAAGGATATTC	
	mtsr69-4R1	AGTCAGCTATATTAATAAGCAGGAGCCTAC	
	69mtsr(r3)	ATTAAAGCTTACGCTTAAATTCTATTTG	
	1183mtsr(r4)	AGCTTACGCTTAAATTCTATTTGTATTC	
	1183mtsr(R5)	ATTAAAATCTTGTCTCTTATGTGTTGAG	
	69mtsr(R4)	ATTTACTCCAAAACTTCCTTCCGCGTC	
	27mtsr(R2)	ACCATAAATACCTTGTTTATCTGATACAG	
	69mtsr(R5)	ATTACTAATATTCACTTCTTGTCGCACAC	
	69mtsr(R6)	ATTTCAAAATAAGTTAAAACTGCATAACCGG	
	mtsr1435-7(R1)	ACCGCGACTGCTGGCACGTAATTAGTCAAG	
	mtsr1183-20(R1)	ATGTAACAGTAAAATCACAAGCTTCATTTG	
	mtsr1183-20R2	ATTTAATAGTTCTACGATAGATTC	
	mtsr1183-20R3	ATGTAACAGTAAAATCACAAGCTTC	
	mtsr1183-20R4	AGTAAAATCACAAGCTTCATTTG	
	mtsr1183-20R5	AATTTAATAGTTCTACG	
	mtsr1183-20R6	ATGTAACAGTAAAATCACAAGC	
	mtsr1183-20R7	AATTTCTTTAATAGATGTAACAG	
	mtsr1183-20R8	AAAAGGCATCCTGGAAACTC	
	mtsr1183-20R9	ACTAAACCACAAATGTATCAG	
	mtsr1183-20F10	ACAATTTAATAGTTCTACG	
	mtsr1183-20F11	ATTTTCTAACTTCTACTCTTCC	
	mtsr1183-20F12	ATGTAACAGTAAAATCACAAGC	
	mtsr1183-20F13	ATTTCTTTAATAGATGTAACAG	
	mtsr1183-20R14	ACGTAAGATTTAGGAAATAGC	
	mtsr1183-20R15	ATTATAATAACTTATTACTGAGAACAC	
	mtsr1183-20R16	ACCTTCCTTTGTGCACTAATTG	
	mtsr1435-3R1	AAGTCGTAGCGCTTCAAAAG	
	mtsr1435-3R2	ATTATTTAAATTTACTTCAATTG	
	mtsr69-4R10	ATTTGTTCTATATTTGATAC	

mt *rns*	mtsr1183-20R17	AAAACTTCCTTCCGCGTC	Antisense
	mtsr1183-20R18	AGGTTCGATATAACAATTTC	
	mtsr1274R1	GGTAGTAGTTCTGTAAAT	
	mtsr1274R2	CGTAATAATACCTTTCTTT	
	mtsr1435-7(R1)	ACCGCGACTGCTGGCACGTAATTAGTCAAG	
I-LtrII gene (N-terminal His_6-tagged)	1435HegFi1	AACGCGGTGACTTTATCCTG	Sense
	1435HegFi2	AACGTCCAGAATTTGGTCTGAG	
	1435HegFi3	ACGTTGGCTGGCAGGCTTCATTG	
	1435HegFi4	AAGAACACCTGACTCTGAACG	
	1435HegRi1	AGTTCAACCACTTTGCAGAACAGG	Antisense
	1435HegRi2	ATCAGGCCGTTCAGAGTCAGGTG	
	1435HegRi3	GACCAAATTCTGGACGTTTAACGC	
	1435HegRi4	AGGTTCATCGCGGCTTTGATGGAC	
mt *cob*	cyb1435F1	AACTGAAACTAGAGTAAATC	Sense
	cyb68F1	AATAATGCAACTTTAAATAGATTCTTTG	
	cyb1435F2	AATTCTAATTGTGGTAAGGTTTTTCGTGG	
	cyb1435R1	ATAATAAAATTTCAGCATC	Antisense
	cyb1435R2	ATTCTCTATTGATCTAACGCTTGTTAC	

Supplementary Table 7.2. Strains of *Ophiostoma*, *Pesotum*, and related taxa used in this study, along with GenBank accession numbers.

Strain[1]	Accession Number	Strain	Accession Number
Ceratocystis deltoideospora WIN(M)41	EU879121	*Ophiostoma himal-ulmi* C1306	AF198234
Ceratocystis introcitrina WIN(M)40	NA[2]	*Ophiostoma megalobrunneum* WIN(M)580	NA
Ceratocystis ips C327	AF198244	*Ophiostoma narcissi* IMI349579	AF484451
Ceratocystis ips C994	DQ062980	*Ophiostoma nigrocarpum* ATCC22391	AF484474
Ceratocystis minor AU58.4	AF234834	*Ophiostoma nigrocarpum* C1142	DQ062978
Ceratocystis pilifera 2/97	AF221071	*Ophiostoma novae-zelandiae* WIN(M)147	NA
Ceratocystis pilifera CBS129.32	AF221070	*Ophiostoma novae-zelandiae* WIN(M)163b	NA
Ceratocystis pluriannulata ATCC8714	NA	*Ophiostoma novo-ulmi* subsp. *americana* C510	AF198236
Ceratocystis pluriannulata C1033	DQ062971	*Ophiostoma novo-ulmi* subsp. *novo-ulmi* CBS298.87	AF198235
Ceratocystis pluriannulata MUCL18373	AY934517	*Ophiostoma perfectum* CBS636.66	DQ062970
Ceratocystis pluriannulata WIN(M)869	NA	*Ophiostoma piceae* AU160-9	AF081132
Ceratocystis stenoceras CBS798.73	AF484475	*Ophiostoma piceae* CBS799.73	AF198231
Ceratocystis tetropii C01-002	AY194489	*Ophiostoma piceae* CMW7647	AF493248
Graphium fragrans C990	DQ062976	*Ophiostoma piceae* CNB13	AJ538341
Ophiostoma sp. C1097	DQ062974	*Ophiostoma piceae* H2009	AF081130
Ophiostoma sp. C1626	DQ062975	*Ophiostoma piceae* H2154	AF081131
Ophiostoma sp. WIN(M)89	NA	*Ophiostoma sejunctum* Ophi1A	AY934519
Ophiostoma sp. WIN(M)465	NA	*Ophiostoma setosum* AU160-38	AF128929
Ophiostoma sp. WIN(M)1136	NA	*Ophiostoma ssiori* MAFF410971	AB096207
Ophiostoma sp. WIN(M)1426	NA	*Ophiostoma subalpinum* MAFF410924	AB096211
Ophiostoma canum NFRI60-165	NA	*Ophiostoma tenellum* CBS189.86	AY934523
Ophiostoma canum MUCL18759	AJ538342	*Ophiostoma ulmi* CBS102.63	AF198232
Ophiostoma canum NCC60-16	AF198229	*Pesotum* sp. WIN(M)1449	NA
Ophiostoma catonianum CBS263.35	AF198243	*Pesotum cupulatum* C1194	AF198230
Ophiostoma coronatum IMI176533	DQ062969	*Sporothrix schenkii* IFM5890	AB089139
Ophiostoma dentifundum CMW13017	AY495435		

[1]CBS, Centraal Bureau voor Schimmelcultures, Utrecht, The Netherlands; WIN(M), University of Manitoba, Microbiology/Botany (J.R.'s personal collection); ATCC, American Type Culture Collection, Rockville, MD, USA; NFRI, Norwegian Forest Research Institute, AS, Norway; MUCL, Mycothèque de l'Universite catholique de Louvain Place Croix du Sud 3, B-1348 Louvain-la-Neuve.

[2] NA, not available (strain has not been formally described and the accession number has yet to be submitted to GenBank).

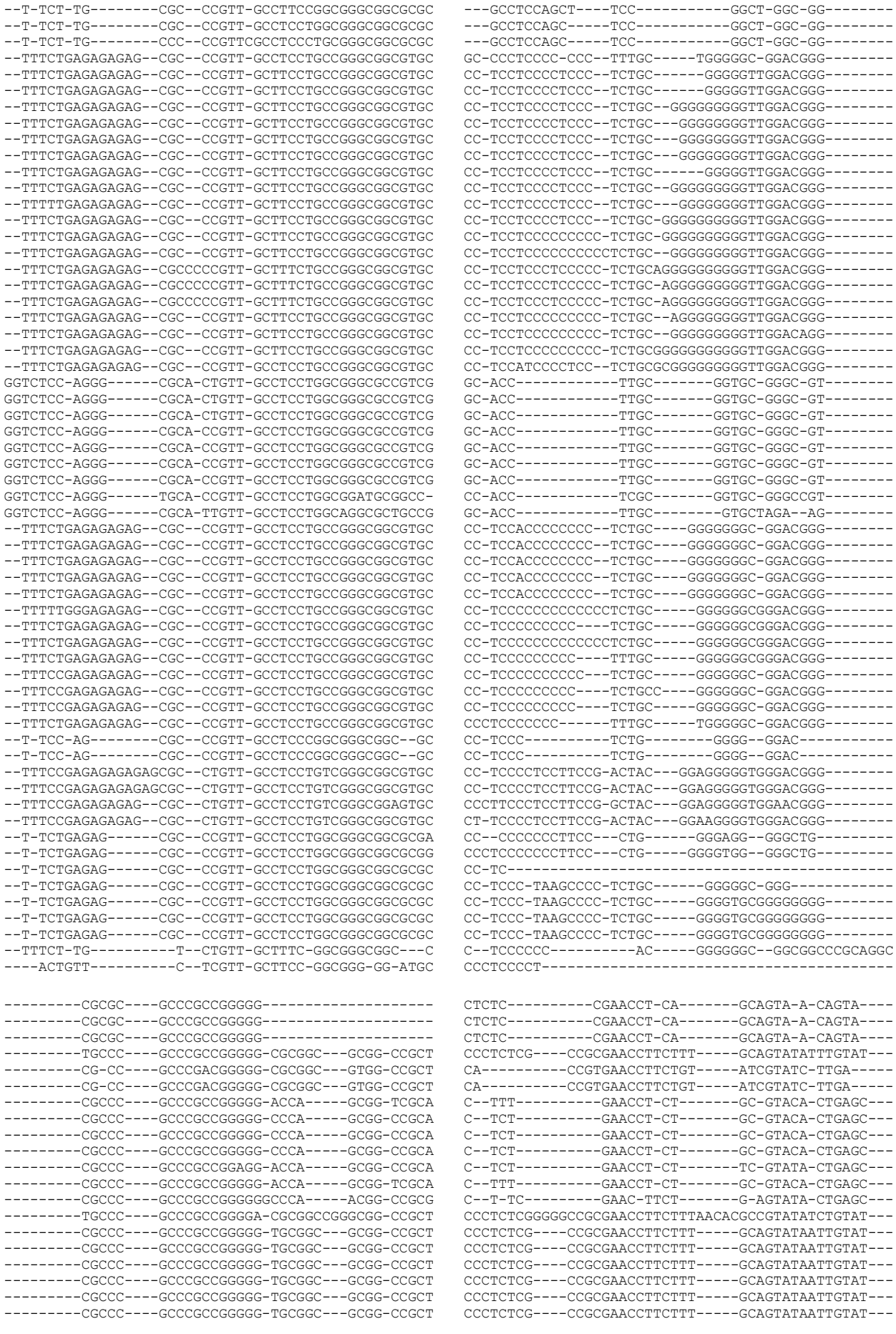
Supplementary Table 7.3. Strains of *Graphium* and related taxa used in this study, along with GenBank accession numbers.

Strain[1]	Accession Number	Strain	Accession Number
Ceratocystis coerulescens CL13-12	AY214000	*Graphium fimbriasporum* CBS870.95 [as WIN(M)1436]	HM244372
Ceratocystis fimbriata WIN(M)931	DQ318204	*Graphium fimbriasporum* CMW3352	AY148178
Ceratocystis polonica WIN(M)450	DQ318199	*Graphium fimbriasporum* CMW3353	AY148179
Graphium sp. 3YT3P.1-G1	DQ268586	*Graphium fimbriasporum* CMW5605	AY148177
Graphium sp. 3YT3P.2-G1	DQ268587	*Graphium fimbriasporum* CMW5606	AY148180
Graphium sp. CCF 3566	AM267264	*Graphium fimbriasporum* CMW5610	AY148176
Graphium sp. CCF 3570	AM267265	*Graphium laricis* CMW5598	AY148181
Graphium sp. DAOM234026 [as WIN(M)1544]	HM244376	*Graphium laricis* CMW5601	AY148183
Graphium sp. DAOM234027 [as WIN(M)1545]	HM244377	*Graphium laricis* CMW5603	AY148182
Graphium sp. DAOM234028 [as WIN(M)1546]	HM244378	*Graphium laricis* CMW5604	AY148165
Graphium sp. JCM 7440	AB038424	*Graphium penicillioides* JCM 9301	AB038428
Graphium sp. JCM 9299	AB038426	*Graphium penicillioides* JCM 10496	AB038430
Graphium sp. JCM 9331	AB038429	*Graphium penicillioides* JCM 10497	AB038431
Graphium sp. WIN(M)718	HM244373	*Graphium penicillioides* JCM 10498	AB038432
Graphium sp. WIN(M)1389	HM244374	*Graphium penicillioides* JCM 10499	AB038433
Graphium sp. WIN(M)1490	HM244375	*Graphium penicillioides* WIN(M)1189	HM244371
Graphium basitruncatum JCM 8083	AB038425	*Graphium pseudormiticum* CMW503	AY148186
Graphium basitruncatum JCM 9300	AB038427	*Graphium pseudormiticum* CMW5611	AY148185

[1]CBS, Centraal Bureau voor Schimmelcultures, Utrecht, The Netherlands; CCF, Culture Collection of Fungi, Department of Botany, Charles University, Prague; CMW, culture collection of the Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, South Africa; DAOM, Cereal and Oilseeds Research, Agriculture & Agri-Food Canada, Ottawa, Ont., Canada JCM, Japan Collection of Microorganism, Riken BioResource Center, Japan; WIN(M), University of Manitoba, Microbiology/Botany (J.R.'s personal collection).

Supplementary Figure 7.1. Alignment of the DNA sequence of the nuclear ITS1-5.8S-ITS2 rDNA region of strains of *Grosmannia*, *Leptographium*, and related taxa.

```
  H. pini WIN(M)82-89                          GC-GAGTTCAC-AGCGACTCCCAACCCGTGCATACC-TA-C-CGCGCCC-
  G. galeiformis CECT20482                     G--GAGTTCAC-AGCGACTCCCAACCCGTGCATACC-TA-C-CGCGCCC-
  G. galeiformis C1101                         GC-GAGTTCAC-AGCGACTCCCAACCCGTGCATACC-TA-C-CGCGCCC-
  G. huntii WIN(M)492                          GC-GAGTTCAC-TGCGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  Leptographium sp. WIN(M)984                  GC-GAGTTCAC-AGCAACTCCCAACCCGTGCAAACC-TA-C-CGCATCC-
  Leptographium sp. WIN(M)985                  GC-GAGTTCAC-AGCAACTCCCAACCCGTGCAAACC-TA-C-CGCATCC-
  Leptographium sp. J.R.88-194A                GC-GAGTTTA--AACAACTCCCAACCCGTGCATACA-TA-C-CGCAT---
  Leptographium sp. WIN(M)528                  GC-GAGTTCA--AGCAACTCCCAACCCGTGCATACA-TA-C-CGCAT---
  L. americanum WIN(M)1456                     GC-GAGTTCA--AGCAACTCCCAACCCGTGCATACA-TA-C-CGCAT---
  G. pseudoeurophioides WIN(M)42               GC-GAGTTCA--AGCAACTCCCAACCCGTGCATACA-TA-C-CGCAT---
  G. penicillata NFRI60-21                     GC-GAGTTCA--AGCAACTCCCAACCCGTGCATACA-TA-C-CGCAT---
  G. penicillata WIN(M)131                     GC-GAGTTTA--AACAACTCCCAACCCGTGCATACA-TA-C-CGCAT---
  G. dryocoetis CBS376.66                      CGAGAGTT-AC-ATCAACTCCCAACCCGTGCATACA-TC-C-CGCA----
  G. piceaperda WIN(M)1380                     GC-GAGTTCAC-TGCGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. wingfieldii TOM11.5                       GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. wingfieldii TOM59.21                      GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. terebrantis UAMH9722                      GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. wingfieldii CBS648.89                     GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. wingfieldii MCC349                        GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. wingfieldii CBS645.89                     GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. wingfieldii TOM10.2                       GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  G. aurea CBS438.69                           GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. terebrantis CBS408.61                     GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. wingfieldii WIN(M)1482                    GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. wingfieldii WIN(M)1322                    GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. wingfieldii MCC130                        GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. terebrantis CBS337.70                     GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. wingfieldii WIN(M)1218                    GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. terebrantis CBS298.85                     GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTAGCACGCATCC-
  L. terebrantis UAMH9690                      GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. wingfieldii WIN(M)1382                    GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. wingfieldii MCC125                        GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. lundbergii DAOM64746                      GC-GAGTTCAC-AGCGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  Pesotum sp. WIN(M)1423                       GC-GAGTTCACAAGTGACTCCCAACCCGTGCCAAC-TTA-C-CGCGCCCT
  Pesotum sp. WIN(M)1428                       GC-GAGTTCACAAGTGACTCCCAACCCGTGCCAAC-TTA-C-CGCGCCCT
  Pesotum sp. WIN(M)478                        GC-GAGTTCACAAGTGACTCCCAACCCGTGCCAAC-TTA-C-CGCGCCCT
  G. davidsonii WIN(M)1494                     GC-GAGTTCACAAGTGACTCCCAACCCGTGCCAAC-TTA-C-CGCGCCCT
  G. davidsonii WIN(M)1132                     GC-GAGTTCACAAGTGACTCCCAACCCGTGCCAAC-TTA-C-CGCGCCCT
  G. davidsonii WIN(M)60B                      GC-GAGTTCACAAGTGACTCCCAACCCGTGCCAAC-TTA-C-CGCGCCCT
  G. cucullata C1216                           GC-GAGTTCACAAGTGACTCCCAACCCGTGCCAAC-TTA-C-CGCGCCCT
  G. francke-grosman. ATCC22061                GC-GAGTTCAA-CATAACTCCCAACCCGTGCCAAC-TTA-C-CGCGCCCT
  O. brevicolle CBS150.78                      GC-GAGTTCAA-AGTGACTCCCAACCCGTGCCAAC-TTA-C-CGCGCCCT
  L. lundbergii DSMZ5010                       GC-GAGTTCAC-AGCGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. lundbergii NFRI89-1040/1/3                GC-GAGTTCAC-AGCGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. lundbergii NFRI69-148                     GC-GAGTTCAC-AGCGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. lundbergii CBS352.29                      GC-GAGTTCAC-AGCGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. lundbergii NFRI60-25                      GC-GAGTTCAC-AGCGACTCCCAACCCGTGCAAACCTTA-C-CGCA-CC-
  Leptographium sp. WIN(M)1269                 GC-GAGTTCAC-AGGGACTCCCAACCCGTGAGAACCTTA-C-CGCATCC-
  L. truncatum TOM86.30                        GC-GAGTTCAC-AGCGACTCCCAACCCGTGAGAACCTTA-C-CGCATCC-
  Leptographium sp. WIN(M)1106                 GC-GAGTTCAC-AGCGACTCCCAACCCGTGAGAACCTTA-C-CGCATCC-
  Leptographium sp. WIN(M)1247                 GC-GAGTTCAC-AGCGACTCCCAACCCGTGAGAACCTTA-C-CGCATCC-
  L. truncatum NFRI1813/1                      GC-GAGTTCAC-AGCGACTCCCAACCCGTGAGAACCTTA-C-CGCATCC-
  L. truncatum TOM74.29                        GC-GAGTTCAC-AGCGACTCCCAACCCGTGAGAACCTTA-C-CGCATCC-
  L. truncatum CBS929.85                       GC-GAGTTCAC-AGCGACTCCCAACCCGTGAGAACCTTA-C-CGCATCC-
  G. piceaperda WIN(M)980                      GC-GAGTTCAC-TGCGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  Pesotum sp. WIN(M)481                        CT-GAGTTC-C-AGCAACTCCCAACCCGTGCAAACA-TA-C-CGCAT-C-
  G. davidsonii WIN(M)1495                     CT-GAGTTC-C-AGCAACTCCCAACCCGTGCAAACA-TA-C-CGCAT-C-
  G. europhioides NFRI80-67/22                 GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATAC-
  G. europhioides CBS229.83                    GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATAC-
  G. laricis CBS636.94                         GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATAC-
  G. europhioides MUCL18355                    GC-GAGTTCAC-AGTGACTCCCAACCCGTGCAAACCTTA-C-CGCATAC-
  G. wageneri ATCC58579                        GC-GAGTTCAC-AGCGACTCCCAACCCGTGCAAACCTTA-C-CGCATCC-
  L. serpens WIN(M)1214                        GC-GAGTTCAC-AGCGACTCCCAACCCGTGCAAACCTTA-C-CGCGTCC-
  L. procerum DAOM33940                        GC-GAGTTCAC-AGCGACTCCCAACCCGTGCATACCTTA-C-CGCAT-C-
  L. procerum NFRI59-84/2                      GC-GAGTTCAC-AGCGACTCCCAACCCGTGCATACCTTA-C-CGCAT-C-
  L. procerum WIN(M)1264                       GC-GAGTTCAC-AGCGACTCCCAACCCGTGCATACCTTA-C-CGCAT-C-
  L. procerum TOM76.8                          GC-GAGTTCAC-GACGACTCCCAACCCGTGCATACCTTA-C-CGCAT-C-
  L. procerum TOM73.12                         GC-GAGTTCAC-AGCGACTCCCAACCCGTGCATACCTTA-C-CGCAT-C-
  C.deltoideospora WIN(M)41                    GC-GAGCGCGTCAGCGACTCCCAACCCGTGCGAACGCTA-C-C-------
  Cops. collifera CBS126.89                    CA-GAGTTTTC-TA-AACTCCCAACCCTTGCGAACCATA-C-C-CAT---

----TCC-GGGG------CGC--CCGTT-GCCTCCTGGCGGGCCGC-CGC   ---GCCC------------AGC----------GC-G--------------
----TCC-GGGG------CGC--CCGTT-GCCTCCTGGCGGGCCGC-CGC   ---GCCC------------AGC----------GC-G--------------
----TCC-GGGG------CGC--CCGTT-GCCTCCTGGCGGGCCGC-CGC   ---GCCC------------AGC----------GC-G--------------
--TTTCTGAGAGAGAG--CGC--CCGTT-GCCTCCTGCCGGGCGGCGTGC   CCCTCCCCCCC------TTTGC-----TGGGGGC-GGATGGGGG------
--TGCCCGAG--------CGC--CCGTT-GCCTCCTGGCGGGCGGCGCG-   ---TCAGCCCCC-----TGTAAAA---CGGGAGCT-GAT-GC--------
--TGCCCGAG--------CGC--CCGTT-GCCTCCTGGCGGGCGGCGCG-   ---TCAGCCCCC-----TGTAAAA---CGGGAGCT-GAT-GC--------
--T-TCT-TG--------CGC--CCGTT-GCCTCCTGGCGGGCGGCGCGC   ---GCCTCCAGC-----TCC-----------GGCT-GGC-GG--------
--T-TCT-TG--------CGC--CCGTT-GCCTCCTGGCGGGCGGCGCGC   ---GCCTCCAGC-----TTC-----------GGCT-GGC-GG--------
--T-TCT-TG--------CGC--CCGTT-GCCTCCTGGCGGGCGGCGCGC   ---GCCTCCAGC-----CTC-----------GGCT-GGC-GG--------
--T-TCT-TG--------CGC--CCGTT-GCCTCCTGGCGGGCGGCGCGC   ---GCCTCTAGC-----TTC-----------GGCT-GGC-GG--------
```



```
---------CGCCC----GCCCGCCGGGGG-TGCGGC---GCGG-CCGCT   CCCTCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-TGCGGC---GCGG-CCGCT   CCCTCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-TGCGGC---GCGG-CCGCT   CCCTCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-TGCGGC---GCGG-CCGCT   CCCTCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-TGCGGC---GCGG-CCGCT   CCCTCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-TGCGGC---GCGG-CCGCT   CCCTCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCTCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCTCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCTCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-TGCGGC---GCGG-CCGCT   CCCTCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-TGCGGC---GCGG-CCGCT   CCCTCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-TGCGGC---GCGG-CCGCT   CCCTCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------TGCCC----GCCCGCCGGGGA-CGTGAC---GCGG-CTGCT   CCCCCTCG----CCGCGAACCTTCTTT-----ACAGTATAATTGTAT---
------------------GCCTGCCGGGGG----------------CGCA   C--------------CGAACCTTCTTT-----GCAGTA---TTATAC---
------------------GCCTGCCGGGGG----------------CGCA   C--------------CGAACCTTCTTT-----GCAGTA---TTGTAC---
------------------GCCTGCCGGGGG----------------CGCA   C--------------CGAACCTTCTTT-----GCAGTA---TTATAC---
------------------GCCTGCCGGGGG----------------CGCA   C--------------CGAACCTTCTTT-----GCAGTA---TTATAC---
------------------GCCTGCCGGGGG----------------CGCA   C--------------CGAACCTTCTTTT----GCAGTA---TTATAC---
------------------GCCTGCCGGGGG----------------CGCA   C--------------CGAACCTTCTTTT----GCAGTA---TTATAC---
------------------GCCTGCCGGGGG----------------CGCA   C--------------CGAACCTTCTTT-----GCCGTA---TTATAC---
------------------GCCTGCCGGGGG----------------CGCA   C--------------CGAACCTTCTTT-----GCAGTA---CAGTAC---
------------------GCTTGCCGGGGG----------------CGCA   C--------------TGAACCTTCTTT-----GTAGTA---TAGTAC---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCCCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCCCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCCCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCCCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCCCTCG----CCGCGAACCTTCTTT-----GCAGTATAATTGTAT---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCTCCCA----CCGCGAACCTTCTTT-----GCAGTATATTTGTAT---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCTCCCA----CCGCGAACCTTCTTT-----GCAGTATATTTGTAT---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCTCCCA----CCGCGAACCTTCTTT-----GCAGTATATTTGTAT---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCTCCCA----CCGCGAACCTTCTTT-----GCAGTATATTTGTAT---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCTCCCG----CCGCGAACCTTCTTT-----GCAGTATATTTGTAT---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCTCCCG----CCGCGAACCTTCTTT-----GCAGTATATTTGTAT---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CCCTCCCG----CCGCGAACCTTCTTT-----GCAGTATATTTGTAT---
---------TGCCC----GCCCGCCGGGGA-CGCGGC---GCGG-CCGCT   CCCTCTCG----CCGCGAACCTTCTTT-----GCCGTATATCTGTAT---
-------------C----GCCCGCCGGGGG-CT--GC---GCGG-CCG--   TTC-CA----------GAACG--CATT-----GCAGTCTT-TAGTGCC--
-------------C----GCCCGCCGGGGG-CT--GC---GCGG-CCG--   TTC-CA----------GAACG--CATT-----GCAGTCTT-TAGTGCC--
---------CACCC----GCCCGCCGGGGG-CGCGGC---GTGGCCCGCT   CCCTCTCG----TCTTGAACCTTCTTT-----ACAGTATACCTATAA---
---------CACCC----GCCCGCCGGGGG-CGCGGC---GTGGCCCGCT   CCCTCTCG----TCTTGAACCTTCTTT-----ACAGTATACCTATAA---
---------CACCC----GCCCGCCGGGGG-CGCGGC---GTGGCCCGCT   CCCTCTCG----TCTTGAACCTTCTTT-----ACAGTATACCTATAA---
---------CACCC----GCCCGCCGGGGG-CGCGGC---ATGGCCCGCT   CCCTCTCG----TCTTGAACCTTCTTT-----ACAGTATACCTATAA---
---------CGACC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CTCTCCC--------CGAACCTTCTTT-----ACAGT-TACCATGAC---
---------CGACC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCC   CTCTCCC--------CGAACCTTCTTT-----GCAGTATACCATTAC---
------------CC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CTCCCAC--------CGAACCTTCTAT------TAGTATACCTTTAC---
-------------------------GGGGG------C---GCGG-CCGCT   CTCCCAC--------CGAACCTTCTAT------TAGTATACCTTTAC---
---------CGCCC----GCCCGCCGGGGG-GGCGGC---GCGG-CCGCT   CTCCCAC--------CGAACCTTCTAT------TAGTATACCTTTAC---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CTCCCAC--------CGAACCTTCTAT------TAGTATACCTTTAC---
---------CGCCC----GCCCGCCGGGGG-CGCGGC---GCGG-CCGCT   CTCCCAC--------CGAACCTTCTAT------TAGTATACCTTTAC---
CTGGTGCCACGCCCAGGCGCCCGCCGGAG----CGAC--------CCAAC   CAAACCTT-------TGATTCTCTTCT---CCGC-GTCTCGCGCGCA---
-----------CCC----GC--GCCGGGGG---CGGC---G-GG-CC-CT   -----AT---------GAACCTT-T-------AC-ATCTCAACCACAAGA

----CGTCTGAG-CA-AAACCACTGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACTGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACTGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CAGAAACCACTGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCAAACTGCGATAAGTAA
----CGTCTGAG-CAGAAACCACTGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCAAACTGCGATAAGTAA
----AGTCTGAG-CA-AAACCA-TGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----AGTCTGAG-CA-AAACCA-TGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----AGTCTGAG-CA-AAACCA-TGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----AGTCTGAG-CA-AAACCA-TGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----AGTCTGAG-AA-AAACCA-TGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----AGTCTGAG-CA-AAACCA-TGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----AGTCTGAG-CA-AAACCA-TGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
-GAGCGTCTGAGTCA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAAACAACGG   ATCTCTTGGTTCTG-CATCGATGAAGAACGCAGCGAACTGCGATAAGTA-
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAA-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCAATGAAAAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG   ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
```

```
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-AA-AAACCACAGAATCGTAAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-AA-AAACCACAGAATCGTAAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-AA-AAACCACAGAATCGTAAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-AA-AAACCACAGAATCGTAAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-AA-AAACCACAGAATCGTAAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-AA-AAACCACAGAATCGTAAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-AA-AAACCACAGAATCGTAAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-AC-AAACCACAGAATCGTAAAAACTTTTAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-AC--AACCACAGAATCGTAAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAAAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AA-CCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AA-CCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AA-CCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AA-CCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AA-CCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AA-CCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AA-CCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----TGTCTGAG-CA-GA-CCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----TGTCTGAG-CA-GA-CCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----TGTCTGAG-TA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----TGTCTGAG-TA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----TGTCTGAG-TA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----TGTCTGAG-TA-AAACCACAGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACTGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACTGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACTGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACTGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACTGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACTGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CA-AAACCACTGAATCGTTAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATAAGTAA
----CGTCTGAG-CACAACTGACACAATCGTCAAAACTTTCAA-CAACGG    ATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAACTGCGATACGTAA
AAACCGTCTGAG-CT-AAAAAC-AAATAATCAAAACTTTCAA-CAACGG    ATCTCTTGGCTCTGGCATCGATGAAGAACGCAGCGAAATGCGATACGTAA

TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGTCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCACTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGTCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCACTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGTGAATTGCAGAATTCAGTGAACCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCACCC-
TGTGAATTGCAGAATTCAGTGAACCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCACCC-
TGTGAATTGCAGAATTCAGTGAACCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCACCC-
TGTGAATTGCAGAATTCAGTGAACCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCACCC-
TGTGAATTGCAGAATTCAGTGAACCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCACCC-
TGTGAATTGCAGAATTCAGTGAACCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCACCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGGGCGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-C--GCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAAAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGTATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGTATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGTATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGTATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGTATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGTATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGTATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
```

```
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGTGAATTGCAGAATTCAGCGAACCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCACCC-
TGTGAATTGCAGAATTCAGCGAACCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCACCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGTATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGCGAATTGCAGAATTCAGCGAGCCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGCATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCTCCC-
TGTGAATTGCAGAATTCAGTGAACCATCGAATCTTTGAACGCACATTGCG    CCTGCCAGCACTCTGG-CAGGCACGCCTGTCCGAGCGTCATTTCCACCCC
TGTGAATTGCAGAATTCAGCGAACCATCGAATCTTTGAACGCACATTGCG    CCCGCCAGTATTCTGG-CGGGCATGCCTGTCCGAGCGTCATTTCCCCCC-

TCACGCGGCCCCCCC---GCG---TGGTGTTGGGG-CGTT-CT-GCGGC-    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-GGCG----GCTCCGAGC
TCACGCGGCCCCCCC---GCG---TGGTGTTGGGG-CGTT-CT-GCGGC-    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-GGCG----GCTCCGAGC
TCACGCGGCCCCCCC---GCG---TGGTGTTGGGG-CGTT-CT-GCGGC-    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-GGCG----GCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCACGCCT---GCG---TGGTGTTGGGG-CGAT-CT-GCGGC-    ---AGCCCCCTAAAGCCAGTGGCGGGCCGG-C-GGCG---GGCTCCGAGC
TCACGCAGCACGCCT---GCG---TGGTGTTGGGG-CGAT-CT-GCGGC-    ---AGCCCCCTAAAGCCAGTGGCGGGCCGG-C-GGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGTC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGC-GCCCCTTTAGTGCG---TGGTGTTGGGG-CATA-CT-GCAGC-    ---AGGCCCTAAAACCCAGTGGCGGGCCGG-C-AGAA---GGCTCCGAGC
TCACGC-GCCCCTTCAGTGCG---TGGTGTTGGGG-CATA-CT-GCGGC-    ---AGGCCCTAAAACCCAGTGGCGGGCCGG-C-AGAA---GGCTCCGAGC
TCACGC-GCCCCTTCAGTGCG---TGGTGTTGGGG-CATA-CT-GCGGC-    ---AGGCCCTGAAACCCAGTGGCGGGCCGG-C-AGAA---GGCTCCGAGC
TCACGC-GCCCCTTTAGTGCG---TGGTGTTGGGG-CATA-CT-GCGGC-    ---AGGCCCTGAAACCCAGTGGCGGGCCGG-C-AGAA---GGCTCCGAGC
TCACGC-GCCCCTTTAGTGCG---TGGTGTTGGGG-CATA-CT-GCGGC-    ---AGGCCCTGAAACCCAGTGGCGGGCCGG-C-AGAA---GGCTCCGAGC
TCACGC-GCCCCTTTAGTGCG---TGGTGTTGGGG-CGT--CT-GCGGC-    ---AGGCCCCGAAACCCAGTGGCGGGCCGG-C-AGAA---GGCTCCGAGC
TCACGCAGCGCGCCT---GCGCGCTGGTGTTGGGGACGTT-CT-GCGGCC    GGGAGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCGATCGGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-TGTT-CT-GCGGCC    ---CACCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-TGTT-CT-GCGGCC    ---AACCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACG-AGCGCGCCT---GCG---TGGTGTTGGGG-TGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-TGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-TGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-TGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-TGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-TGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-TGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-TGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-TGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-TGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCGGC-CTCTC---GCG---TGGTGTTGGGG-CGAT-CT-GCGACC    ---AGCCCCTCAAAGCAAGTGGCGGGCCGG-C-AGCG---GTCTCCGAGC
TCACGCGGC-CTTTC---GCG---TGGTGTTGGGG-CGAT-CT-GCGACC    ---AGCCCCTCAAAGCAAGTGGCGGGCCGG-C-AGCG---GTCTCCGAGC
TCACGCGGC-CTTTC---GCG---TGGTGTTGGGG-CGAT-CT-GCGACC    ---AGCCCCTCAAAGCAAGTGGCGGGCCGG-C-AGCG---GTCTCCGAGC
TCACGCGGC-CTTTC---GCG---TGGTGTTGGGG-CGAT-CT-GCGACC    ---AGCCCCTCAAAGCAAGTGGCGGGCCGG-C-AGCG---GTCTCCGAGC
TCACGCGGC-CTTTC---GCG---TGGTGTTGGGG-CGAT-CT-GCGACC    ---AGCCCCTCAAAGCAAGTGGCGGGCCGG-C-AGCG---GTCTCCGAGC
TCACGCGGC-CTTTC---GCG---TGGTGTTGGGG-CGAT-CT-GCGACC    ---AGCCTCTCAAAGCAAGTGGCGGGCCGG-C-AGCG---GTCTCCGAGC
TCACGCGGC-CTTTC---GCG---TGGTGTTGGGG-CGAT-CT-GCGACC    ---AGCCCCTCAAAGCAAGTGGCGGGCCGG-C-AGCG---GTCTCCGAGC
TCATGCGGC-CTTTC---GCG---TGGTGTTGGGG-CGAT-CT-GCGGCC    ---AGCCCCCGAAAGCGAGTGGCGGGCCGG-C-AGTG---GTCTCCGAGC
TCATGCGGC-CTTTC---GCG---TGGTGTTGGGG-CGAT-CT-GCAACC    ---AGCCCCTGAAAGCGAGTGGCGGGCCGG-C-AGTG---GTCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGAC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAACGCGCCT---GCG---TGGTGTTGGGGGCGTT-CTTGCG-CC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGTC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGTC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGTC    ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
```

```
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGTC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGTC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGTC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGTC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCGCCCCGGC----GCG---TGGTGTTGGGG-CATG-CT-GCGGC-   ---AGGCCCCTAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCGCCCCGGC----GCG---TGGTGTTGGGG-CATG-CT-GCGGC-   ---AGGCCCCTAAAACCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCACGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCACGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCACGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCACGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTT-CT-GCGGCC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTTTCT-GCGGCC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTTTCT-GCGGCC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTTTCT-GCGGCC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTTTCT-GCGGCC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCACGCAGCGCGCCT---GCG---TGGTGTTGGGG-CGTTTCT-GCGGCC   ---AGCCCCCGAAAGCCAGTGGCGGGCCGG-C-AGCG---GGCTCCGAGC
TCAGGC-GCTTGCC----------TGGCGTTGGGG-C-CC-CT-----CC   ---AGGCCCCGAAAGCGAGTGGCGGGCCGGCCTGGCG---GGCTCCGAGC
TCACGCGCCCCGCGC---------TGGTGTTGGGG-C-TC-CT-----CC   ---GGGCCCCTAAAGCGAGTGGCGGGCCCGTCTGGCG----GCTCCGAGC

GCAGTA--AGCA-CC-AGT-------GCCC-TCGCTCCGGA-CGACCACC   GCC-TGCGCCCTG----CCC--AGAAC--GCG-GACGC----G-AGTCTC
GCAGTA--AGCA-CC-AGT-------GCCC-TCGCTCCGGA-CGACCACC   GCC-TGCGCCCTG----CCC--AGAAC--GCG-GACGC----G-AGTCTC
GCAGTA--AGCA-CC-AGT-------GCCC-TCGCTCCGGA-CGACCACC   GCC-TGCGCCCTG----CCC--AGAAC--GCG-GACGC----G-AGTCTC
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCC-CAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-GC-ATT-------GCCC-TCGCTCTGGATTGCCCCCA   GCC-TGCGCCCTG----CCCT--AGAC-AGCA-GACGC----A-AGTCTA
GCAGTA--AGCA-GC-ATT-------GCCC-TCGCTCTGGATTGCCCCCA   GCC-TGCGCCCTG----CCCT--AGAC-AGCA-GACGC----A-AGTCTA
GCAGTA--AGCA-TC-AC--------GCCCTTCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCCCCAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TT-AC--------GCCC-TCGCTCCGGA-CGCCCCT-   GCT-GGCAGCTCG----CCCTGC-GGC-GG-A-AACGC--TTGCAG-CAA
GCAGTA--AGCA-TT-AC--------GCCC-TCGCTCCGGA-CGCCCCT-   GCT-GGCAGCTCG----CCCTGC-GGC-GG-A-AACGC--TTGCAG-CAA
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCCGGA-CGCCCCT-   GCT-GGCAGCTCG----CCCTGC-GGC-GG-A-AACGC--TTGCAG-CAA
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCCGGA-CGCCCCT-   GCT-GGCAGCTCG----CCCTGC-GGC-GG-A-AACGC--TTGCAG-CAA
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCT-   GCT-GGCAGCTCG----CCCTGC-GGC-GG-A-AACGC--TTGCAG-CAA
GCAGTA--AGCA-TC-A---------GCCC-TCGCTCTGGA-CGCCCCT-   GCT-GGCAGCTCG----CCCTGC-GGC-GG-A-AACGC--TTGCAG-CAA
GCAGTA--AGCA-TC-ACCTGTTTATGCCC-TCGCTCTGGA-CGCCCCCC   GCC-TGCGCCCTGCC-GCCCCCCCGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----TCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----TCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----TCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCTCCC-   GCC-TGCGCCCTG----CCCCACAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTTTGGA-CGCCCCC-   GCC-TGCGCCCTG----TCCCCCATACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TT-TTT-------GCCTCTCGCTCTGGA-CGATGCT-   GCC-TGCGCCCTG----CCCC-C-AAT--GCA-GACGC----A-AGTCTA
GCAGTA--AGCA-TT-TTT-------GCCTCTCGCTCTGGA-CGATGCT-   GCC-TGCGCCCTG----CCCC-C-AAT--GCA-GACGC----A-AGTCTA
GCAGTA--AGCA-TT-TTT-------GCCTCTCGCTCTGGA-CGATGCT-   GCC-TGCGCCCTG----CCCC-C-AAT--GCA-GACGC----A-AGTCTA
GCAGTA--AGCA-TT-TTT-------GCCTCTCGCTCTGGA-CGATGCT-   GCC-TGCGCCCTG----CCCC-C-AAT--GCA-GACGC----A-AGTCTA
GCAGTA--AGCA-TT-TTT-------GCCTCTCGCTCTGGA-CGATGCT-   GCC-TGCGCCCTG----CCCC-C-AAT--GCA-GACGC----A-AGTCTA
GCAGTA--AGCA-TT-TTT-------GCCTCTCGCTCTGGA-CGATGCT-   GCC-TGCGCCCTG----CCCC-C-AAT--GCA-GACGC----A-AGTCTA
GCAGTA--AGCA-TT-TTT-------GCCTCTCGCCCTGGA-CGATGCT-   GCC-TGCGCCCTG----CCCC-C-AAT--GCA-GACGC----A-AGTCTA
GCAGTA--AGCA-TC-TTT-------GCCTCTCGCTCTGGA-CGAAGCT-   GCC-TGCGCCCTG----CCT-----AT--GCA-GACGC----A-AGTCTA
GCAGTA--AGCA-TC-TTT-------GCCTCTCGCTCTGGA-TAACGCT-   GCC-TGCGCCCTG----CCT----TAT--GCA-GACGC----A-AGTCTA
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCC-CATACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCC-CATACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCC-CATACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCC-CATACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCC-CATACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCCCCATACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCCCCATACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCCCCATACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCCCCATACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCCCCAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCCCCAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCCCCAGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-TC-AC--------GCCC-TCGCTCTGGA-CGCCCCCC   GCC-TGCGCCCTGCC--CCCCCCCGACCGGCA-GACGC----G-AGTCTG
GCAGTA--AGCA-GA-AAC-------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-GGCTCCC-G----CCTA--CAGC-GG------------A-AG-CTG
GCAGTA--AGCA-GC-AAC-------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-GGCTCCC-G----CCTA--CAGC-GG------------A-AG-CTG
GCAGTA--AGCA-GC-AGT-------GCCC-TCGCTTTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCT-TATTTTGGCA-GACGT----A-AGTCTG
GCAGTA--AGCA-GC-AGT-------GCCC-TCGCTTTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCT-TATTTTGGCA-GACGT----A-AGTCTG
GCAGTA--AGCA-GC-AGT-------GCCC-TCGCTTTGGA-CGCCCCC-   GCC-TGCGCCTTG----CCCC-CATTTTGGCA-GACGT----A-AGTCTG
```

```
GCAGTA--AGCA-GC-AGT-------GCCC-TCGCTTTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCC-TATTCTGGCA-GACGT----A-AGTCTG
GCAGTA--AGCA-GC-AGT-------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCCAAAGAC-GGCA-GACGC----G-AGTCTC
GCAGTA--AGCA-GC-AGT-------GCCC-TCGCTCTGGA-CGCCCCC-   GCC-TGCGCCCTG----CCCCAAAGAC-GGCA-GACGC----G-AGTCTC
GCAGTA--AGCA-GC-AGT-------GCCC-TCGCTCTGGA-CGCCTCC-   GCC-TGCGCCCTGC---CCCAAATGACCGGCA-GACGC----A-AGTCTG
GCAGTA--AGCA-GC-AGT-------GCCC-TCGCTCTGGA-CGCCTCC-   GCC-TGCGCCCTGC---CCCAAATGACCGGCA-GACGC----A-AGTCTG
GCAGTA--AGCA-GC-AGT-------GCCC-TCGCTCTGGA-CGCCTCC-   GCC-TGCGCCCTGC---CCCAAATGACCGGCA-GACGC----A-AGTCTG
GCAGTA--AGCA-GC-AGT-------GCCC-TCGCTCTGGA-CGCCTCC-   GCC-TGCGCCCTGC---CCCAAATGACCGGCA-GACGC----A-AGTCTG
GCAGTA--AGCA-GC-AGT-------GCCC-TCGCTCTGGA-CGCCTCC-   GCC-TGCGCCCTGC---CCCAAATGACCGGCA-GACGC----A-AGTCTG
GCAGTA--GTTACGCAAGTCAT---TGC----CGCTCCGGA-CGCCCCCG   GCGAGGCGCCCAG----CCGAGG-GACGCGCG-CGCGCCGC-GCGCCCCG
GCAGTACCGA-ACGCAAGTTC----TCCCC-TCGCCC-GGA-CGCCCC--   GTC-GGCGCCTTGCCAGCCCAAAT--CGCGCATGACGT----GCAA-CT-

--GCACCT-CAC--
--GCACCT-CAC--
--GCACCT-CAC--
--CCTCCTTCTC--
CCCTTT-TTTTAC-
CCCTTTTTTTTAC-
--CCCCCTTCTC--
--CCAATTT-T---
--CCAATTTTT---
--CCATTTTCT---
--CCATCTTA----
--CCAT-TTCT---
--CCATATTC----
--TCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTT--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--------TCTCA-
--------TCTCA-
--------TCTCA-
--------TCTCA-
--------TCTCA-
--------TCTCA-
--------TCTCA-
--------TCTCA-
--------TCTCA-
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCCCCTTCTC--
--CCCCCTTCTC--
--CCCCCTTCTC--
--CCCCCTTCTC--
--CCCCCTTCTC--
--CCCCCTTCTC--
--CCCCCTTCTC--
--TCTCCTTCTC--
A-CCCCCATCGC--
A-CCCCCATCGC--
--CCTCCT-CTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
--CCTCCTTCTC--
CCCCTCCTTCTC--
CCCCTCCTTCTC--
CCCCTCCTTCTC--
CCCCTCCTTCTC--
CCCCTCCTTCTC--
CCCCCCCTTCTCAA
---CTTCTT-AC--
```

Supplementary Figure 7.2. Alignment of the DNA sequence of the nuclear ITS1-5.8S-ITS2 rDNA region of strains of *Ophiostoma*, *Pesotum*, and related taxa.

```
Ophiostoma sp. WIN(M)1426                CATTAGC---GAG-------TTCACAAGTGACTCCC-AACCCGTGCCAA-CTTACCGCGC
Pesotum sp. WIN(M)1449                   CATTACAAAAGAG--------T---GAAAA-CTCC--AACCCGTGTGAACCATA-CCGAC
G. fragrans C990                         CATTACAAAAGAG--------T---GAAAA-CTCC--AACCCGTGTGAACCATA-CCGAC
C. introcitrina WIN(M)40                 CATTACA---GAGTCTTTTCTTC-GGAAAAACTCCCTAACCCACGCGAACTTTA------
C. deltoideospora WIN(M)41               CATTAGC---GAG----CGCGTC-AGCGA--CTCCC-AACCCGTGCGAACGCTA------
O. dentifundum CMW13017                  CATTACA---GAG-----TTTTC--ACAA--CTCCC-AACCCTTGCGAACCGTA------
Ophiostoma sp. WIN(M)89                  CATTAGC---GAG-----TTTTC----A---CTCCC-AACCCTTGCGAACCGTA------
C. ips C327                              CATTAGC---GAG-----TTTTC----AA--CTCCC-AACCCTTGCAAACCGTA------
C. ips C994                              CATTAGC---GAG-----TTTTC----AA--CTCCC-AACCCTTGCAAACCGTA------
Ophiostoma sp. WIN(M)1136                CATTAGC---GAG-----TTTTC----A---CTCCC-AACCCTTGCCAACCGTA------
S. schenkii IFM5890                      CATTACA---GAG-----TTTTC--ACAA--CTCCC-AACCCTTGCGAACCGTA------
O. himal-ulmi C1306                      CATTACA---GAG----TTTTT-----AA--CTCCC-AACCCTTGCGAACCGTA------
O. ulmi CBS102.63                        CATTACA---GAG----TTTTTC----AA--CTCCC-AACCCTTGCGAACCGTA------
O. piceae CMW7647                        CATTACA---GAG----TTTTT-----AA--CTCCC-AACCCTTGCGAACCGTA------
O. novae-zelandiae WIN(M)147             CATTACA---GAG----TTTTT-----AA--CTCCC-AACCCTTGCGAACCGTA------
O. piceae AU160-9                        CATTACA---GAG----TTTTT-----AA--CTCCC-AACCCTTGCGAACCGTA------
O. catonianum CBS263.35                  CATTACA---GAG----TTTTT-----AA--CTCCC-AACCCTTGCGAACCGTA------
O. novo-ulmi CBS298.87                   CATTACA---GAG---TTTTTTC----AA--CTCCC-AACCCTTGCGAACCGTA------
P. cupulatum C1194                       CATTAGC---GAG-----TTTTC----AA--CTCCC-AACCCTTGCGAACCGTA------
Ophiostoma sp. WIN(M)465                 CATTAGC---GAG-----TTTTC----AA--CTCCC-AACCCTTGCGAACCGTA------
O. canum NFRI60-165                      CATTAGC---GAG-----TTTTC----AA--CTCCC-AACCCTTGCGAACCGTA------
O. piceae CNB13                          CATTAGC---GAG-----TTTTC----AA--CTCCC-AACCCTTGCGAACCGTA------
O. subalpinum MAFF410924                 CATTAGC--GGGG-----TTTT-----ATA-CTCCC--ACCCTTGCGAACCGTA------
O. setosum AU160-38                      CATTAGC---GAG-----TTTTC----AA--CTCCC-AACCCTTGCGAACCGTA------
C. tetropii C01-002                      CATTAGC---GAG-----TTTTC----AA--CTCCC-AACCCTTGCGAACCGTA------
O. piceae H2009                          CATTAGC---GAG-----TTTTC----AA--CTCCC-AACCCTTGCGAACCGTA------
O. ssiori MAFF410971                     CATTAGC---GAG-----TTT-C-GACA---CTCCC--ACCCTTGCGAACCGTA------
O. piceae CBS799.73                      CATTAGC---GAG-----TTTTC----AA--CTCCC-AACCCTTGCGAACCGTA------
C. minor AU58.4                          CATTAGC---GAG-----TTTTC----AA--CTCCC-AACCCTTGCGAACCGTA------
O. coronatum IMI176533                   CATTACC---GAG-------TTCCAGCAA--CTCCC-AACCCGTGCGAACCGTA------
O. tenellum CBS189.86                    CATTACC---GAG-------TTCCAGCAA--CTCCC-AACCCGTGCGAACCGTA------
O. megalobrunneum WIN(M)580              CATTACC-AAGAG-------TTCCAGCAA--CTCCC-AACCCTTGCGAACCGTA------
O. sejunctum Ophi1A                      CATTACC---GAG-------TTCCAGCAA--CTCC--AACCCTTGCGAACCGTA------
O. narcissi IMI349579                    CATTACA---GAG-----TTTTCT---AAA-CTCCC-AACCCTTGCGAACCATA------
C. stenoceras CBS798.73                  CATTACA---GAG-----TTTTCT---AAA-CTCCC-AACCCTTGCGAACCATA------
C. pilifera 2/97                         CATTAGC---GAG-----TCTTC----AA--CTCCC-AACCCTTGCCAACCGTA-----C
C. pilifera CBS129.32                    CATTAGC---GAG-----TTTTC----AA--CTCCC-AACCCTTGCCAACCGTA-----C
O. nigrocarpum C1142                     CATTACA---GAG-----TTTTCC---AAA-CTCCC-AACCCTTGCGAACCATA------
O. nigrocarpum ATCC22391                 CATTACA---GAG-----TTTTCC---AAA-CTCCC-AACCCTTGCGAACCATA------
C. pluriannulata WIN(M)869               CATTATC---GAG-----TTATCC---AA--CTCCC-AACCCTTGCGAATCGTA------
Ophiostoma sp. C1097                     CATTATC---GAG-----TTATCC---AA--CTCCC-AACCCTTGCGAATCGTA------
O. perfectum CBS636.66                   CATTATC---GAG-----TTGTCC---AA--CTCCC-AACCCTTGCGAACCGTA------
```

```
CTTACCGCGC CCTGGT-CTC CAGGGCGCAC T-GTTGCCTC CTGGCGGGCG   ---------- ---------- ---------- ---------- ----CCGTCG
CATA-CCGAC GTTTGTGCT- ------GTCC T-GTTG-CTT TCGGCCGGCG   GCTC-TTTGG A--GCC-CGC AAGG--T--- ---------- ----CCTT-G
CATA-CCGAC GTTTGTGCT- ------GTCC T-GTTG-CTT TCGGCCGGCG   GCTCTTTTGG A--GCC-CGC AAGG--T--- ---------- ----CCTT-G
TTTA------ --CCA--CT- ------GTCC T-GTTG-CTT TTGGCAGGCG   G--------- ---------- ---------- ---------- ---CCCCCTG
GCTA------ --CCTTTCTT ------GT-C T-GTTG-CTT TCGGCGGGCG   G--------- ---------- ---------- -------CCC TCCCCCCACG
CGTA------ -CCCATTCT- ------GTTC TCGTTG-CTT CCGGCGCGGG   GGGACGGGGG GG-------- ---------- ---------- ----CCTCCG
CGTA------ CCCCGTTCC- ------GTTC TCGTTG-CCT CTGGCGGGAG   GGGGCGGGGG GC-------- --CGC----- ---------- ----CCCTCG
CGTA------ CCCCGTTCC- ------GTTC TCGTTG-CTT CTGGCGGGAG   GCGGGAGGGG GGA------- --CGCGTCGC T--------- ----CCTTCG
CGTA------ CCCCGTTCC- ------GTTC TCGTTG-CTT CTGGCGGGAG   GCGGGAGGGG GGA------- --CGCGTCGC T--------- ----CCTTCG
CGTA------ CCCCGTCCC- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGGGGGG-- ---------- --CGCGT--- ---------- ----CCTTCG
CGTA------ CCCAAT-CTC ------GTTC TCGTTG-CTT CTGGCGGGGG   GGAGCGGGGG GG-------- ---------- ---------- ----CGCCCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGAGGGG-- ---------- --CGCGT--- ---------- ----CCTTCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGAGGGG-- ---------- --CGCGT--- ---------- -----CTTCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGAGGGG-- ---------- --CGCGT--- ---------- ----CCTTCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGAGGGG-- ---------- --CGCGT--- ---------- ----CCTTCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGAGGGG-- ---------- --CGCGT--- ---------- ----CCTTCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGAGGGG-- ---------- --CGCGT--- ---------- ----CCTTCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGAGGGG-- ---------- --CGCGT--- ---------- ----CCTTCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGGGGGG-- ---------- --CGCGT--- ---------- ----CCTTCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGGGGGG-- ---------- --CGCGT--- ---------- ----CCTTCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGGGGGGG- ---------- --CGCGT--- ---------- ----CCTTCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGGGGGG-- ---------- --CGCGT--- ---------- ----CCTTCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTGGCTT C-GGCGGGAG   GGGGGGGG-- ---------- --CGCGT--- ---------- ----CCTTGG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGGGGGG-- ---------- --CGCGT--- ---------- ----CCTCCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGAGGGG-- ---------- --CGCGT--- ---------- ----CCTTCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGGGGGG-- ---------- --CGCGT--- ---------- ----CCTCCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGAAGGG-- ---------- --CGCTT--- ---------- ----CCTTCA
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGGGG---- ---------- --CGCGT--- ---------- ----CCTTCT
CGTA------ CCCCGTCCC- ------GTTC TCGTTG-CTT CTGGCGGGAG   GGGAGGGG-- ---------- --CGCGT--- ---------- ----CCTTCG
CGTA------ -CCCCTTCT- ------GTTC TCGTTG-CCT CTGGCGGGCG   GCGGGCGCGG GACGGGGGGC GGCGCGT--- ---------- ----ACCCCG
CGTA------ -CCCCTTCT- ------GTTC TCGTTG-CCT CTGGCGGGCG   GCGGGCGCGG GACGGGGGGC GGCGCGT--- ---------- ----ACCCCG
CGTA------ CCCCCTCCT- ------GTTC TCGTTG-CCT CCGGCGGCGG   GGG------- ---CGGGGGC GGCGCCC--- ---------- ---------A
CGTA------ CCCAGT-CT- ------GTTC TCGTTG-CTT CTGGCGGCGG   GGG------- ---------- ---------- ---------- --------CG
CATA------ -CCCATACT- ------GTTC TCGTTG-CTT CCGGCGAGGG   ACGGGGCGGG GT-------- ---GC-T--- ---------- ----CTCTCG
CATA------ -CCCATACT- ------GTTC TCGTTG-CTT CCGGCGGGGG   ACGGGGCGGG GT-------- ---GC-T--- ---------- ----CTCTCG
CGTA-----C CCCCGTCCC- ------GTTC TCGTTG-CCT CTGGCGGG--   -CGGGGCTGG GT-------C GTCGCGTCCT CTCTCTCTCC CCCCCCTTCT
CGTA-----C CCCCGTCCC- ------GTTC TCGTTG-CCT CTGGCGGG--   -CGGGGCTGG GT-------C GTTGCGTCCT CTCTC----- ---CCCCTCT
CATA------ -CCCATACT- ------GTTC TCGTTG-CTT CCGGCGGGGG   AC------GG GT------GG GGGGC-TG-- ---------- ----CCCTCG
CATA------ -CCCATACT- ------GTTC TCGTTG-CTT CCGGCGGGGG   AC------GG GT------GG GGGGC-TG-- ---------- ----CCCTCG
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GC-GGACTGG GTGGCG-GGC AAGGC-TGG- ---------- -----CTTCG
```

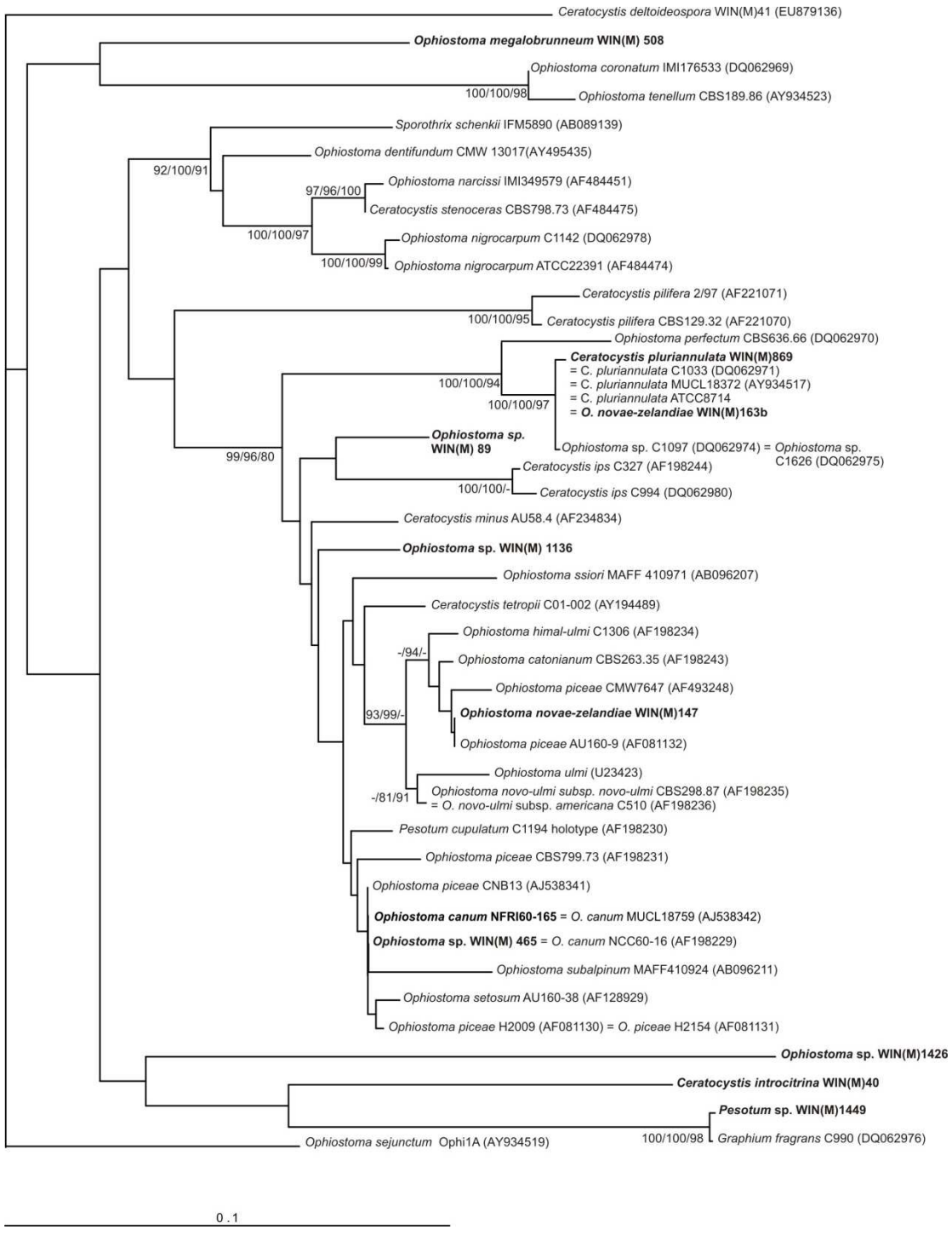
```
CGTA------ CCCCGTTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GC-GGACTGG GTGGCG-GGC AAGGC-TGG- ---------- -----CTTCG
CGTA------ CCCAATTCT- ------GTTC TCGTTG-CTT CTGGCGGGAG   GC-GGACTGG GCGGCG-GGC ATGGC-TGG- ---------- -----CTTCG

G---CAC--- ---------- ---------- ---------- ----------   ---------- CTTGCGGTG- CGGGC----- GTGCCTGCCG GGGGC--GC-
G---CGC--- ---------- ---------- ---------- ----------   ---------- --TTCGGCG- CGAG----GG GCACCGGCCG GGGGT-GGC-
G---CGC--- ---------- ---------- ---------- ----------   ---------- --TTCGGCG- CGAG----GG GCACCGGCCG GGGGT-GGC-
GGAGGGA--- ---------- ------GCCC G-CT------ ----------   -----AGG-C CC--CGGTGT CGAGCCCGGG GTGCCTGCCA G-AG---GC-
GGGGG----- ---------- CGGCG-GCCC G-C------- ----------   -----AGG-C C---TGGTGC CACGCCC-AG GCGCCCGCCG G-AG-CGACC
GG--C----- ---------- --------TC TCCCCGTCC- ----------   ---------- ---------- ---------- -CGCCCGCCG GGGG-CGGC-
GGG-CGCGCC CCCCT----- ------GCCT C-C--CCCC- ----------   ------GGTC CCCTCACG-- -------GG GCGCCCGCCA GCGGAC-GC-
GGGGCGCGC- ---------- -GCCCCGCCT CCCCCCCCT- ----------   -----AGGTC CCCTCG---- -------GG GCGCCCGCCA GCGGCCAGC-
GGGGCGCGC- ---------- -GCCCCGCCT CCCCCCCCT- ----------   -----AGGTC CCCTCG---- -------GG GCGCCCGCCA GCGGCCAGC-
GGG-CGC--- ---------- ------GCCC CCTCCTCCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCC-GC-
AC-------- ---------- ------ACGG CCCCCTCC-- ----------   -------GCC CC-------- ---------- -CGCCCGCCA GGGG-CGGC-
GGG-CGC--- ---------- -----CGCCT CTCTCTCCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCC-GC-
GGG-CGC--- ---------- -----TGCCT CTCTCTCCC- ----------   -----AGGTC C-TTC----- -------GG GCGCCCGCCA -CGGCC-GC-
GGG-TG---- ---------- -----TGCCT CTCTCTCCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCC-GC-
GGG-CG---- ---------- -----TGCCT CTCTCTCCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCC-GC-
GGG-CG---- ---------- -----TGCCT CTCTCTCCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCC-GC-
GGG-CGC--- ---------- -----TGCCT CTCTCTCCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCC-GC-
GGG-CGC--- ---------- -----TGCCT CTCTCTCCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCC-GC-
GGG-CGC--- ---------- ------GCTC CTCTCTCCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCC-GC-
GGG-CGC--- ---------- ------GCCT CTCTCTCCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCC-GC-
GGG-CGC--- ---------- ------GCCT CTCTCTCCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCC-GC-
GGG-CGC--- ---------- ------GCCT CTCTCTCCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCC-GC-
GGG-CGC--- ---------- ------GCCT CTCTCTCCC- ----------   -----AGGTC CCTTGG---- -------GG GCGCCGGCCA ACGGCC-GC-
GGG-CGC--- ---------- ------GCCT CTCTCTCCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCC-GC-
GGGGGG---- ---------- ------GCCT CTCTCTCCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCT-GC-
GGG-CGC--- ---------- ------GCCT CTCTCTCCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCT-GC-
GGAGAGAG-- ---------- ------A-CT CTCTCCCCC- ----------   -----AGGTC CCTCCG---- -------GG GCGCCCGCCA GCGGCC-GC-
GGG-CGC--- ---------- ------GCCT CTCT--CCC- ----------   -----AGGTC CCTTCG---- -------GG GCGCCCGCCA GCGGCC-GC-
GGG-CGC--- ---------- ------GCCT CTCCCCCCC- ----------   -----AGGTC CCCTTG---- -------GG GCGCCCGCCA GCGGCC-GC-
AAACCCGGC- --------GC C--ATCGCCA GGGGGGAGGG G-CGGCGCC-   ------TGTC CCCCGGCCCA AGCCCGC--- -CGCCCGCCA GCGG-CGGC-
AAACCCGGC- --------GC C--ATCGCCA GGGGGGAGGG G-CGGCGCC-   ------TGTC CCCCGGCCCA AGCCCGC--- -CACCCGCCA GCGG-CGGC-
GCT------- --------GC C--CTCACCT C--------- ----------   ---------- ---------- ---------- ---CCCGCCG GCGG-CGGC-
GGCGGCCCCC ---------- -----AGCGG GC-------- ----------   ------GGCC CCCCC----- ---------- ---CCCGCCA GGGG-CGGC-
AG--C----- ---------- ------GCCC CTCCCCTCC- ----------   ---------- ---------- ---------- -CGC--GCCG GGGG-CGGC-
AG--C----- ---------- ------GCCC CTCCCCTCC- ----------   ---------- ---------- ---------- -CGC--GCCG GGGG-CGGC-
GGGGGGGGGG GAGGCCTGGC C--GC-GCCG CCCCTGCCCC CCCGGCGCCC   CCTCCTAG-- ----CGGGGG GTG-----GT GCGCCCGCCG GCGGAC-GC-
GGGGG----- --GGCCTGGC C--GC-GCCG CCCCTGCCCC CCCGGCGCCC   CCTCCTAG-- ----CGGGGG GTG-----GT GCGCCCGCCG GCGGAC-GC-
GGCT------ ---------- ------GCCC CTCCCCTCC- ----------   ---------- ---------- ---------- -CGC--GCCG GGGG-CGGC-
GGCT------ ---------- ------GCCC CTCCCCTCC- ----------   ---------- ---------- ---------- -CGC--GCCG GGGG-CGGC-
GCCGGCCCT- --------GC TCTGCTGCCC GGGCCTCCCC --CGGTCCTC   CCTC-TGG-C CC-TCGGGCC AGGGACG-AG GCGCCCGCCA GCGGTCAGC-
GCCGGCCCT- --------GC TCTGCTGCCC GGGCCTCCCC --CGGTCCTC   CCTC-TGG-C C-TTCGGGCT AGGGACG-AG GCGCCCGCCA GCGGTCAGC-
GCCGGTCTT- --------GC CCTGCTGCCC GGGCCCCCCC T-CGGTCCTC   CCTC-TGG-C C-TTCGGGCT GGAGACG-AG GCGCCCGCCA GCGGTCGGC-

-A-CC----- ---------G AACCTTCTT- --TG------ CAGTA-TTAT   A--CCGTCTG AG--AAA--- ---AACCA-- CAGAATCGTA AAAACTTTCA
-AGCTCGGCA GTTTGTTTTA AACGCTTGC- -CTGAT---A TTGTGGCT--   ---C-GTCTG AGCAAAAAGA AAAGACAAAA --TAG---TC AAAACTTTCA
-AGCTCGGCA GTTTGTTTTA AACGCTTGC- -CTGAT---A TTGTGGCT--   ---C-GTCTG AGCAAAAAGA AAAGACAAAA --TAG---TC AAAACTTTCA
-AGCT----- -----TTTTA AACTCTTTTG -CT------- --GTT-CCAG   G--CTGTCTG AGTAATT--- --TAA-AAAA --TAAA--TC AAAACTTTCA
CAACC-AAAC C-----TTTG A-TTCTCTTC TCCGC----- --GTCTCGCG   CGCACGTCTG AGC--AC--- ---AACTGA- CACAATCGTC AAAACTTTCA
GGGCCCTAC- ---------G AACCTTTATA TCTCAA---C CACTG---GA   AA-CTGTCTG AGCAAAC--- --AAACAAAA --TAA---TC AAAACTTTCA
GAGCC--GCC C---------G AACTCTTTTA AACAA----C C-GTAACCGT   C--CCGTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCC C---------G AAC-TTATC- AATCAA---- CAGCAACCG-   TT-CCGTCTG AGC-AAC--- --AAACAAAA ATTAG----C AAAACTTTCA
GAGCC--GCC C---------G AAC-TTATC- AATCAA---- CAGCAACCG-   TT-CCGTCTG AGC-AAC--- --AAACAAAA ATTAG----C AAAACTTTCA
GAGCC--GCC C---------G AAC--TCTT- -CTAAA---- CAGTAACTGA   A--C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GGGCCCTAC- ---------G AACCTTTGTA TCTCAA---C CACTA---GA   AAACCGTCTG AG-GAAA--- --AAACAAAA --TAA---TC AAAACTTTCA
GAGCC--GCC T---------G AACCTTTTC- -ATAAA---C CAGTAAC-GA   AA-C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCC T---------G AACCTTTT-- -CTAAA---C CAGTAAC-GA   AA-C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCC T---------G AACTTTTT-- -ATAAA---C CAGTGTACGA   A--CCGTCTG AG-AAAC--- --AAACAAAG A-CAG---CC AAAACTTTCA
GAGCC--GCC T---------G AACTTTTT-- -ATAAA---C CAGTAAC-GA   AA-C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
-AGCC--GCC T---------G AACTTTTT-- -ATAAA---C CAGTAA--GA   AA-C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCC T---------G AACTTTTT-- -ATAAA---C CAGTAAC-GA   AA-C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCC C---------G AACCTTTT-- -CCAAA---C CAGTAAC-GA   AA-C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCC C---------G AACCTTTT-- -ATAAA---C T-GTAACTGA   AA-C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCC T---------G AACTTTTTTT ---AAA---C C-GTAAC-GA   AA-C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCC T---------G AACTTTTTTT ---AAA---C C-GTAAC-GA   AA-C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCC T---------G AACTTTTTT- ---AAA---C C-GTAAC-GA   AA-C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCT T---------G AACTTTTTTT ---AAA---C C-GTAAC-GA   AA-C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCC T---------G GACTTTTTTT ---AAA---C C-GTAAC-GA   AA-C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCC T---------G AACTTTTTTT ---AAA---C C-GTAAC-GA   A--CCGTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCC T---------G AACTTTTTTT ---AAA---C C-GTAAC-GA   AA-C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCC C---------G AACTTTTTT- --CAAA---C C-GTAACCGA   A--C-GTCTG AG-AAAC--- --AAACAAAA -CCAG---CC AAAACTTTCA
GAGCC--GCC T---------G AACTTTATT- --CAAA---C C-GTAACTAA   AA-C-GTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GAGCC--GCC T---------G AACC-T-TT- AAAAAA---C C-GTTACCAA   --ACCGTCTG AG-AAAC--- --AAACAAAA A-CAG---CC AAAACTTTCA
GGGCCT--CC T---------G AACGCTTTTT T-CAAAC--C CAATA-C-AG   TGTCCTTCTG AG-CC-C--- -AAAACAAA- -CCAA---TC AAAACTTTCA
GGGCCT--CT T---------G AACGCTTTTT T-CAAAC--C CAATA-C-AG   TGTCCTTCTG AG-CC-T--- -AAAACAAA- -CCAA---TC AAAACTTTCA
GAGCCCTGCA C---------G AACCTCTTTT TCT------C TAGCA-C---   ---C-GTCTG AG-GCAC--- --ACACAAA- --TGAA--TC AAAACTTTCA
GAGCCC-ACC ---------G AACC-TCTT- -GAGAACAGC CCCT--GCAG   -CGCCTTCTG AG-AAAC--A AAAAACG--- ---AA----C AAAACTTTCA
GGGCCCTAT- ---------G AACCTTTATA TCTCAA---C CACTA---GA   AAACCGTCTG AG-AAAC--- --AAACAAAA --TAA---TC AAAACTTTCA
GGGCCCTAT- ---------G AACCTTTATA TCTCAA---C CACTA---GA   AA-CCGTCTG AG-AAAC--- --AAACAAAA --TAA---TC AAAACTTTCA
GAGCC--GTC CC--------G AAC--TCTT- -CTAAA---C C-GTCCACTG   TAAC-GTTTG AG-AAAA--- --AAACAAAA A-TAAA---C AAAACTTTCA
GAGCC--GTC CC--------G AAC--TCTT- -CTAAA---C C-GTCCACTG   TAAC-GTCTG AG-AAAA--- --AAACAAAA A-TAAA---C AAAACTTTCA
GGGCCCTAT- ---------G AACCTTTACA TCTCAA---C CAC-AA--GA   AAACCGTCTG AGTCTAA--- --AAACAAA- --TAA---TC AAAACTTTCA
GGGCCCTAT- ---------G AACCTTTACA TCTCAA---C CAC-AA--GA   AAACCGTCTG AGTTTAA--- --AAACAAA- --TAA---TC AAAACTTTCA
GAGCC--ACT CT--------G AAC--TCTT- -ATAAA---C C-GTT-TTAG   TTACTGTCTG AG-AAAC--- -AAAACAAAA A-TAG---CC AAAACTTTCA
GAGCC--ACT CT--------G AAC--TCTT- -ATAAA---C C-GTT-TTAG   TTACTGTCTG AG-AAAC--- -AAAACAAAA A-TAG---CC AAAACTTTCA
```

```
GAGCC--ACC CT-------G AAC--TCTT- AA-CAA---C CTGTTTT-AG   TTACTGTCTG AG-AAAC--- -AAAACAAAA -TTAG---CC AAAACTTTCA

ACAACGGATC TCTTGGTTCT GG-CATCGAT GAAGAACGCA GCGAACTGCG   ATAAGTAATG CGAATTGCAG AATTCAGCGA GCCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGTTCT GG-CATCGAT GAAGAACGCA GCGAAACGCG   ATAAGTAATG TGAATTGCAG AATTTAGTGA ATCATCAAAT CTTTGAACGC
ACAACGGATC TCTTGGTTCT GG-CATCGAT GAAGAACGCA GCGAAACGCG   ATAAGTAATG TGAATTGCAG AATTTAGTGA ATCATCAAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATAAGTAATG TGAATTGCAG AATTCAGTGA ATCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGTTCT GG-CATCGAT GAAGAACGCA GCGAACTGCG   ATACGTAATG TGAATTGCAG AATTCAGTGA ACCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG TGAATTGCAG AATTCAGCGA ACCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG TGAATTGCAG AATTCAGCGA ACCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCAAAATGCG   ATACGTAATG CGAATTGCAG AATTCAG-GA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GGCCATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACAGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAGATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG A-TTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG TGAATTGCAG AATTCCGCGA ACCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG TGAATTGCAG AATTCCGCGA ACCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG TGAATTGCAG AATTCAGCGA ACCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG TGAATTGCAG AACTCAGCGA ACCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG TGAATTGCAG AATTCAGCGA ACCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG TGAATTGCAG AATTCAGCGA ACCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG TGAATTGCAG AATTCAGCGA ACCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG TGAATTGCAG AATTCAGCGA ACCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC
ACAACGGATC TCTTGGCTCT GG-CATCGAT GAAGAACGCA GCGAAATGCG   ATACGTAATG CGAATTGCAG AATTCAGCGA GTCATCGAAT CTTTGAACGC

ACATTGCGCC CGCC-AGTAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCTCCC-T CA-CGCGGCC --TCT----- --------CG CG--T-GGTG
ACATTGCGCC TGCC-AGTAT TCTGGCAGGC -ATGCCTGTC CGAGCGTCAT   TTCCACCC-T CAG-GCG-CC ----T---AG --------TG CGCCT-GCTG
ACATTGCGCC TGCC-AGTAT TCTGGCAGGC -ATGCCTGTC CGAGCGTCAT   TTCCACCC-T CAG-GCG-CC ----T---AG --------TG CGCCT-GCTG
ACATTGCGCC TGCC-AGTAT TCTGGCAGGC -ATGCCTGTC CGAGCGTCAT   TTCAACCC-T CGA-GCGGCC CGTCT----- ---------- -GC-TCGGTG
ACATTGCGCC TGCC-AGCAC TCTGGCAGGC -ACGCCTGTC CGAGCGTCAT   TTCCACCCCT CAG-GCG-C- ---TT----- ---------- -GCCT-GGCG
ACATTGCGCC CGCC-AGTAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CA-CGCG-CC CCGTT---G- --------CG CGC-T-GGTG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCGCG-CC CCTTT--CGG GGGGAAG--G CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCGCG-C- --TCT----- --------CG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCGCG-C- --TCT----- --------CG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCGCG-CC C-TC----GG G-------TG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CA-CGCG-CC CCGTT---G- --------CG CGC-T-GGTG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCATG-CC CC-TT---AG GG------TG CGC-T-GGCG
ACATTGGGCC CGCC-AGTAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAG-ATG-CC CTTTTT--GG GAGG------ -GC-T-GGTG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCATA-CC C-TTT---GG G-------TG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCATA-CC C-TTT---GG G-------TG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCATA-CC C-TTT---GG G-------TG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCATA-CC C-TTT---GG G-------TG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCATG-CC C-TTTT--GG G-------TG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCACG-CC C-TTT---GG G-------TG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCACG-CC C-TTT---GG G-------TG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCACG-CC C-TTT---GG G-------TG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCACG-CC C-TTT---GG G-------TG CGC-T-GGCG
ACATTGCGCC CGCCCAGCA- TCTGGCGGGC AATGCCTGTC CGAGCGCCAT   TTCCCCCC-T CAGCACG-CC C-TTT---GG --------TG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCACG-CC C-TTT---GG G-------TG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCATA-CC C-TCT---GG G-------TG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCACG-CC C-TTT---GG G-------TG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGG-C -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCGCG-CC C-TCT---GG GA-------G CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCACG-CC CTTCTTGCAA GGGA-----G CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCGCG-CC C-TCT---GG GA-------G CGC-T-GGCG
ACATTGCGCC TGCC-AGCAC TCTGGCAGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCGCG-CC C-TCCAGCGG --------C- CGC-T-GGTG
ACATTGCGCC TGCC-AGCAC TCTGGCAGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCGCG-CC C-TCCAGCGG --------C- CGC-T-GGTG
ACATTGCGCC TGCC-AGCAT TCTGGCAGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCCCT CAGCGCG-CC CCC-----GG --------C- CGC-T-GGTG
ACATTGCGCC TGCC-AGCAT TCTGGCAGGC -ATGCCTGTC CGAGCGTCAT   TTCACCCC-T CAGCGCG-CC CC----GCGG --------CG CGC-T-GGTG
ACATTGCGCC CGCC-AGTAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CA-CGCG-CC CCGTT---G- --------CG CGC-T-GGTG
ACATTGCGCC CGCC-AGTAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CA-CGCG-CC CCGTT---G- --------CG CGC-T-GGTG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCGCG-CC CCTCTC-CAG GGGACTCCCG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CAGCGCG-CC CCTCTC-CAG GGGACTCCCG CGC-T-GGCG
ACATTGCGCC CGCC-AGTAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CA-CGCG-CC CCGTT---G- --------CG CGC-T-GGTG
ACATTGCGCC CGCC-AGTAT TCTGGCGGGC -ATGCCTGTC CGAGCGTCAT   TTCCCCCC-T CA-CGCG-CC CCGTT---G- --------CG CGC-T-GGTG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTT CGAGCGTCAT   TTCCCCCC-T CAGCACG-CC -TTC----GG --------CG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTT CGAGCGTCAT   TTCCCCCC-T CAGCACG-CC -TTC----GG --------CG CGC-T-GGCG
ACATTGCGCC CGCC-AGCAT TCTGGCGGGC -ATGCCTGTT CGAGCGTCAT   TTCCCCCC-T CAGCACG-CC --TTT---AG --------CG CGC-T-GGCG
```

```
TTGGGGCGAT CTGCGACCGG -TCGCTGCAG C--GATGCGG CG-TCCGGCG   CAGCCCCTC- AAA-GCAAGT GGCGGGCC-G GC---AG--C GGTCTCCGAG
TTGGGGCATT CTGCGGCCT- ---A------ ---------- ---GCCGC--   AGG--CCTCG AAA-GCAAGT GGCGGGCT-G GC-T-AG-TC GG-CTCCGAG
TTGGGGCATT CTGCGGCCT- ---A------ ---------- ---GCCGC--   AGG--CCTCG AAA-GCAAGT GGCGGGCT-G GC-T-AG-TC GG-CTCCGAG
TTGGGGCG-- CTACGGCGCC CGTTAAACGT TTAGTTTTAA CGGGCGGCG-   TAGG-CCCTG AAAACC-ATT GGCGG-CCCT GC-TGGGATT G--TTCCGCG
TTGGGGCC-C CTCCGCCACC -G-------- ---------- ---GCGG---   AGG-CCCC-G AAA-GCGAGT GGCGGGCC-G GCCT-GG-CG GG-CTCCGAG
TTGGGGC-CC CTCCGCC-T- ---G------ ---------- ---GCGG---   AGGGCCCC-G AAA-GCGAGT GGCGGGCCCG GC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC CGCG------ ---------- ---GCGGC--   AGGGCCCCC- AAATCC-AGT GGCGGGCCCG GC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC -G-G------ ---------- ---GCGGC--   AGGGCCCTC- AAAACC-AGT GGCGGGCCCA GC-T-GG--T GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC -G-G------ ---------- ---GCGGC--   GGG-CCCTC- AAAACC-AGT GGCGGGCCCA GC-T-GG--T GG-CTCCGAG
TTGGGGCT-C CTCCGCCCCC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAATAC-AGT GGCGGGCCCG TC-T-GG-CT GG-CTCCGAG
TTGGGGCGCC CTCCGCC-T- ---G------ ---------- ---GCGG---   GGGGCCCCCG AAA-GCGAGT GGCGGGCCCT G--T-GGA-T GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAAACC-AGT GGCGGGCCCG TC-T-GG-TT GG-CTCCGAG
TT-GGGCT-C CT--GCCCTC TGCGT----- ---------- ---GCG----   AGGGCCCTC- AAAACC-AGT -GCGGGCCCG CC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAAACC-AGT GGCGGGCCCG TC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAAACC-AGT GGCGGGCCCG TC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAAACC-AGT GGCGGGCCCG TC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC T--G------ ---------- ---GCGGC--   AGGGCCCTC- AAAACC-AGT GGCGGGCCCG TC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAAACC-AGT GGCGGGCCCG TC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAATCC-AGT GGCGGGCCCG TC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAATCC-AGT GGCGGGCCCG TC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAATCC-AGT GGCGGGCCCG TC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAATCC-AGT GGCGGGCCCG TC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCC-TC TGTG------ ---------- ---GCGGC--   AGGG-CCTC- AAATCC-AGT GGCGGGCC-G TC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAATCC-AGT GGCGGGCCCG TC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAATTC-AGT GGCGGGCCCG CC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAATCC-AGT GGCGGGCCCG TC-T-GG-TT GG-CTCCGAG
TTGGGGCA-C CTCCGCCCTC CGCG------ ---------- ---GCGGC--   AGGGCCCTC- AAATCC-AGT GGCGGGCCCG CC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAATCC-AGT GGCGGGCCCG GC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAATCC-AGT GGCGGGCCCG TC-T-GG-CT GG-CTCCGAG
TTGGGGCA-C C-CCGGC--- -GGG------ ---------- ---GCCG---   GGGGCCCT-G AAATGC-AGT GGCGGGCCCG GC---GGACC GG-CCCTGAG
TTGGGGCA-C C-CCGCCGCC CGCCCCGCCT AC-GCGCGGG CGGGCCG---   GGGGCCCT-G AAATGC-AGT GGCGGGCCCG GC---GGACC GG-CCCTGAG
TTGGGGCG-C CTCCGCC-T- ---G------ ---------- ---GCGG---   AGGGCCCT-G AAA-GCGAGT GGCGGGCCTG GCC-----CT GG-CACCGAG
CTGGGGCG-C CTCTGCC-T- ---G------ ---------- ---GCAG---   AGGGCCCC-G AAA-GCGAGT GGCGGGCCTG GC-T-GG-CT GG-CTCCGAG
TTGGGGCT-C CTCTGCC-T- ---G------ ---------- ---GCAG---   AGGGCCCCT- AAA-GCGAGT GGCGGGCCCG GC-T-GG--T GG-CTCCGAG
TTGGGGCT-C CTCTGCC-T- ---G------ ---------- ---GCAG---   AGGGCCCCC- AAA-GCGAGT GGCGGGCCCG GC-T-GG--T GG-CTCCGAG
TTGGGGCT-C CTCCGTCCCC CGCG------ ---------- ---GCGGC--   AGGGCCCTC- AAATCC-GGT GGCGGGCCCG TC-T-GG-CT GG-CTCCGAG
TTGGGGCT-C CTCCGTCCCC CGCG------ ---------- ---GCGGC--   AGGGCCCTC- AAATCC-AGT GGCGGGCCCG TC-T-GG-CT GG-CTCCGAG
TTGGGGCT-C CTCCGCC-T- ---------- ---------- ---GGGG---   AGGGCCCCT- AAA-GCGAGT GGCGGGCCCG TC-T-GG--C GG-CTCCGAG
TTGGGGCT-C CTCCGCC-T- ---G------ ---------- ---GCGG---   AGGGCCCCT- AAA-GCGAGT GGCGGGCCCG TC-T-GG--C GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAATCC-AGT GGCGGGCCCG AC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCGCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAATCC-AGT GGCGGGCCCG AC-T-GG-TT GG-CTCCGAG
TTGGGGCT-C CTCCTCCCTC TGTG------ ---------- ---GCGGC--   AGGGCCCTC- AAATCC-AGT GGCGGGCCCG AC-T-GG-TT GG-CTCCGAG

CGCAGT---A AGCA----TT TTTGCCTCTC GC-TCTGGAC -GATGCT-GC   C-TGCGCCCT GCC-CC-CAA ----TGCA-G AC-------- -----GCAAG
CGCAGTA--- A-------TT TTTCT---TC GC-TTTGGAC --GCCCTGGC   C-TGCGCCTT GCC-GTAAAA GCC------- ---------- ----------
CGCAGTA--- A-------TT TTTCT---TC GC-TTTGGAC --GCCCTGGC   C-TGCGCCCT GCC-GTAAAA GCC------- ---------- ----------
CGCAGTA--G A------TGT -CTCTCT-TC GC-ACCGGGC --GCCCC-GC   GGAA-GCC-A GCC-GT--AA AACGGCCC-- GC-------T TTTTAG----
CGCAGTAGTT -ACGC-AAGT CAT-TG-C-C GC-TCCGGAC -GCCCCCGGC   GAGGCGCCCA GCC-GA-GGG A-CGCGCG-C GC-G--CC-- GC---GCGCC
CGCAGTACCG AACGC-AAGT TCTCCC-CTC GC-TC-GGAC --GCCCC-GC   C-GGCGCCCT GCC-GT-CAA AACGCGCATG AC-GT----- -----GCAA-
CGCAGTACCG AACGC-AAGT TCTCTCCCTC GC-TCTG--C -AGCCCCGGC   C-GGCGCCCA GCC-GT-CAA GCCGCGCA-G GCCGG-CCC- ------CAGC
CGCAGTACCG AACGCCTTGT TCTC-CTCTC GC-TCTG--C -AGCTCCGGC   C-GGCGCTCT GCC-GT-CAA GCCGCGCG-G GC-GG-C--- ------CAGC
CGCAGTACCG AACGCTTTGT TCTC-CTCTC GC-TCTG--C -AGCTCCGGC   C-GGCGCTCT GCC-GT-CAA GCCGCGCG-G GT-GG-C--- ------CAGC
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGC   C-GGCGCCCA GCC-GT-CAA GCCGCGCA-G GC-GA-CCCT -CT---CGGG
CGCAGTACCG AACGC-ATGT TCTCCC-CTC GC-TCCGGAC --GCCCC-CC   A-GGCGCCCT GCC-GT-GAA AACGCGCATG AC-GC----- -----GCAG-
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGT   C-GGTGCCCA GCC-GT-CAA GCCGCGCA-G G--AGGCTCT GTTTTACAGG
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGT   C-GGTGCCCA GCC-GT-CAA -CGGCGCT-- -C-GG-CCCT GCTT-GCAGG
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGT   C-GGTGCCCA GCC-GT-CAA ACCGCGCA-G G--AAGCTCT GCTT-GCAGA
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGT   C-GGTGCCCA GCC-GT-CAA ACCGCGCA-G G--AGGCTCT GCTT-GCAGA
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGT   C-GGTGCCCA GCC-GT-CAA ACCGCGCA-G G--AGGCTCT GCTT-GCAGA
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGT   C-GGTGCCCA GCC-GT-CAA ACCGCGCA-G G--AGGCTCT GTTT-ACAGG
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGT   C-GGTGCCCA GCC-GT-CAA GCCGCGCA-G GC-GG-CCCT GCTT-GCAGG
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGC   C-GGCGCCCA GCC-GT-CAA GCCGCGCA-G GC-GA-CCCT -CT---CAGG
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGC   C-GGCGCCCA GCC-GT-CAA GCCGCGCA-G GC-GA-CCCT -CT---CGAG
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGC   C-GGCGCCCA GCC-GT-CAA GCCGCGCA-G GC-GA-CCCT -CT---CGAG
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C AAGCCCCGGC   C-GGCGCCCA GCC-GT-CAA GCCGCGCA-G GC-GA-CCCT -CT---CGAG
CGCAGTACCG AA-GC-AACT TCTCTCTCTC GC-TCTG--C -AGCCCCGGC   C-GGCGCCCA ACC-GT-CAA GCCGCGCG-R GC-GTA-CCT -CT---CGAG
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGC   C-GGCGCCCA GCC-GT-CAA GCCGCGCA-G GC-GA-CCCT -CT---CGAG
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCAGT   C-GGTGCCCA GCC-GT-CAA GCCGCGCA-G GC-GA-CTCT -TT---CGAG
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGC   C-GGCGCCCA GCC-GT-CAA GCCGCGCA-G GC-GA-CCCT -CT---CGAG
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGG   C-GGCGCCCA GCC-GT-CAA GCCGCGCA-G GC-GA-CCCT -CC---CAGG
CGCAGTACCG AACGC-AAGT TCTCTCTCTC GC-TCTG--C -AGCCCCGGC   C-GGCGCCCT GCC-GT-CAA GCCGCGCA-G GC-GAACCCT -CT---CGGG
CGCAGTACCG AACGC-AAGT TCTCTTTCTC GCTTC-G--T -AGCCCCGGC   C-GGTGCCCA GCC-GT-CAA GCCGCGCA-G GC-GA-CCCT TTT---CGGG
CGCAGCACCG AACGCACCGT TCTCTC-CTC GC-TC-AG-C -GGCTCCCCC   CCGGCGCCCA GCC-GT-CAA AA--CACG-C AC----CCC- GC---GCGG-
CGCAGCACCG AACGCACCGT TCTCTC-CTC GC-TC-AG-C -GGCTCCCCC   CCGGCGCCCA GCC-GT-CAA AA--CACG-C AC----CCC- GC---GCGG-
CGCAGCACTG GACACACCGT -CTTCACCTC GC--CCGCGT -TGCCCTGGC   C-GGCGCCCA GCC-GT-CAA GCAAGGCG-G TC-GGGC--- G----GCAGG
CGCAGTACTG GACGCGACGT TCTCT--CTC GC-TCCG--C -GGCCGCCGC   C-GGCGCCCT GCC-GT-CAA AAGACGCG-C GC-GG-CCCC ------CGGG
CGCAGTACCG AACGC-AAGT TCTCCATCTC AC-TC-GGAC -AGCCCC-GT   C-GGCGCCTT GCCAGTTCAA ATCGCGCATG AC-GT----- -----GCAA-
CGCAGTACCG AACGC-AAGT TCTCCATCTC AC-TC-GGAC --GCCCC-GT   C-GGCGCCTT GCCAGTTCAA ATCGCGCATG AC-GT----- -----GCAA-
CGCAGTACTG GGCGA-AAGC TCTCTCCCTC GC--CCG--T -GGCCCCGGA   C-GGCGCCCA GCC-GT-CAA GACGCGCAGG A--GAAGGAA GA---GG-AG
CGCAGTACTG AGCGA-AAGC TCTCTCCCTC GC--CCG--T -GGCCCCGGA   C-GGCGCCCA GCC-GT-CAA GACGCGCAGG A--GAGGGAG GA---GGAAG
CGCAGTACCG AACGC-AAGT TCTCCC-CTC GC--CCGGAC --GCCCC-GT   C-GGCGCCTT GCCAGCCCAA ATCGCGCATG AC-GT----- -----GCAA-
CGCAGTACCG AACGC-AAGT TCTCCC-CTC GC--CCGGAC --GCCCC-GT   C-GGCGCCTT GCCAGCCCAA ATCGCGCATG AC-GT----- -----GCAA-
CGCAGTACCG AACGC-AAGT TCTCT-TCTC GC-TCTG--C -AGCCCCGGC   C-GGTGCCCA GCC-GT-CAA GCC-TGCG-A GT-GG-CC-T -C---ACAG-
CGCAGTACCG AACGC-AAGT TCTCT-TCTC GC-TCTG--C -AGCCCCGGC   C-GGTGCCCA GCC-GT-CAA GCC-TGCG-A GC-GG-CC-T -C---ACAG-
CGCAGTACCG AACGC-AAGT TCTCT-TCTC GC-TCTGA-- -AGCCCCGGC   C-GGTGCCCA GCC-GT-TAA GCC-TGCG-A G--TGGCC-T CC----CAG-
```

```
TCTA------ ---------- ---------T CTC------- -AAAGG--TG    A-
---------- ---------- ---------- TTAA------ -TTTTTTT-T    A-
---------- ---------- ---------- TTAA------ -TTTTTT--T    A-
---------- ---------- -------CG- ---GC-A-C- ---TG--ACT    A-
---------- ---------- ------CCGC CCCCC---C- -TTCT----C    A-
---------- ---------- ---------- CTC------- -TTTTT--AC    A-
GT-TCG-CGC TGCGG----- ------CCGC CTCGC-ATC- -TTT----AC    A-
GTCTCTTCGC TG-------- ------CCGC CCCGC-A-C- -TTTGT-ACC    A-
GTCTTTTCGC TG-------- ------CCGC CCCGC-A-C- -TTTGT-ACC    A-
GG-------- ---------- ------CCGC CTCGC-A-C- -TTCT-----    A-
---------- ---------- ---------- CTC------- -TTTTT--AC    A-
A--------- ---------- ------CCGC CTCGC-AT-- -TTT---CAC    A-
A--------- ---------- ------CCGC CTCGC-A-C- TTTTT--CAC    A-
A--------- ---------- ------CCGC CTCGC-A-C- -TTTT---AC    A-
A--------- ---------- ------CCGC CTCGC-AT-- -TTTT---AC    A-
A--------- ---------- ------CCGC CTCGC-AT-- -TTTT---AC    A-
A--------- ---------- ------CCGC CTCGC-AT-- -TTTT---AC    A-
A--------- ---------- ------CCGC CTCGC-A-C- -TTTT--CAC    A-
G--------- ---------- ------CCGC CTCGC-A-C- -TTTT---AC    A-
GG-------- ---------- ------CCGC CTCGC-AT-- -TTTT---AC    A-
GG-------- ---------- ------CCGC CTCGC-AT-- -TTTT---AC    A-
GG-------- ---------- ------CCGC CTCGC-AT-- -TTTT---AC    A-
G--------- ---------- ------CCGC CTCGC-AT-- -TTTGT--CG    A-
GG-------- ---------- ------CCGC CTCGC-AT-- -TTTT---AC    A-
GG-------- ---------- ------CCGC CTCGC-A-C- TTTTT----C    A-
GG-------- ---------- ------CCGC CTCGC-AT-- -TTTT---AC    A-
G--------- ---------- ------CCGC CTCGC-A-CT TTTTTAT-GC    A-
GG-------- ---------- ------CCGC CTCGC-A-C- TTTTT---AC    A-
---------- ---------- ------CCGC CTCGC-A-CT TTTTT---AC    A-
---------- ---------- ---------- -T-GC-ACCC CCCC---ACC    A-
---------- ---------- ---------- -T-GC-ACCC CCCC---ACC    A-
GGACCTCTCC TG-------- ------TCGT CCCGTCCCCG AACCACACAC    A-
---------- ---------- ------CCGC C--GC-ACGC ACGACGCTGC    A-
---------- ---------- ---------- CTC------- -TTTTT--AC    A-
---------- ---------- ---------- CTC------- -TTTTT--AC    A-
G---CCCCCG GGCCCCCTCC CCCCCCCCTC CTCGC-ACCC CA-----AAC    AC
G---CCCCCG GGCCCCCT-- CCCCCCCCTC CTCGC-ACCC CA-----AAC    AC
---------- ---------- ---------- CTC------- -TTTTT--AC    A-
---------- ---------- ---------- CTC------- -TTTTT--AC    A-
---------- ---------- ------CCGC C-CGCTATC- -TTTT---AT    A-
---------- ---------- ------CCGC C-CGCTATC- -TTTT---AT    A-
---------- ---------- ------CCGC TT-GCTATCT CTTTT---AC    A-
```

Supplementary Figure 7.3. Phylogenetic analysis of nuclear ITS1-5.8S rDNA-ITS2 sequences of *Ophiostoma*, *Pesotum*, and related taxa. Branch lengths were determined using NJ. Values at the nodes were obtained with NJ/DNA PARS/Tree Puzzle programs. "-" indicates level of support not significant/node absent.


Ceratocystis deltoideospora WIN(M)41 (EU879136)
Ophiostoma megalobrunneum WIN(M) 508
Ophiostoma coronatum IMI176533 (DQ062969)
100/100/98
Ophiostoma tenellum CBS189.86 (AY934523)
Sporothrix schenkii IFM5890 (AB089139)
Ophiostoma dentifundum CMW 13017(AY495435)
92/100/91
Ophiostoma narcissi IMI349579 (AF484451)
97/96/100
Ceratocystis stenoceras CBS798.73 (AF484475)
100/100/97
Ophiostoma nigrocarpum C1142 (DQ062978)
100/100/99
Ophiostoma nigrocarpum ATCC22391 (AF484474)
Ceratocystis pilifera 2/97 (AF221071)
100/100/95
Ceratocystis pilifera CBS129.32 (AF221070)
Ophiostoma perfectum CBS636.66 (DQ062970)
Ceratocystis pluriannulata WIN(M)869
= C. pluriannulata C1033 (DQ062971)
= C. pluriannulata MUCL18372 (AY934517)
= C. pluriannulata ATCC8714
= O. novae-zelandiae WIN(M)163b
100/100/94
100/100/97
Ophiostoma sp.
WIN(M) 89
Ophiostoma sp. C1097 (DQ062974) = Ophiostoma sp.
C1626 (DQ062975)
99/96/80
Ceratocystis ips C327 (AF198244)
100/100/-
Ceratocystis ips C994 (DQ062980)
Ceratocystis minus AU58.4 (AF234834)
Ophiostoma sp. WIN(M) 1136
Ophiostoma ssiori MAFF 410971 (AB096207)
Ceratocystis tetropii C01-002 (AY194489)
Ophiostoma himal-ulmi C1306 (AF198234)
-/94/-
Ophiostoma catonianum CBS263.35 (AF198243)
Ophiostoma piceae CMW7647 (AF493248)
93/99/-
Ophiostoma novae-zelandiae WIN(M)147
Ophiostoma piceae AU160-9 (AF081132)
Ophiostoma ulmi (U23423)
-/81/91
Ophiostoma novo-ulmi subsp. novo-ulmi CBS298.87 (AF198235)
= O. novo-ulmi subsp. americana C510 (AF198236)
Pesotum cupulatum C1194 holotype (AF198230)
Ophiostoma piceae CBS799.73 (AF198231)
Ophiostoma piceae CNB13 (AJ538341)
Ophiostoma canum NFRI60-165 = O. canum MUCL18759 (AJ538342)
Ophiostoma sp. WIN(M) 465 = O. canum NCC60-16 (AF198229)
Ophiostoma subalpinum MAFF410924 (AB096211)
Ophiostoma setosum AU160-38 (AF128929)
Ophiostoma piceae H2009 (AF081130) = O. piceae H2154 (AF081131)
Ophiostoma sp. WIN(M)1426
Ceratocystis introcitrina WIN(M)40
Pesotum sp. WIN(M)1449
100/100/98
Graphium fragrans C990 (DQ062976)
Ophiostoma sejunctum Ophi1A (AY934519)
0.1

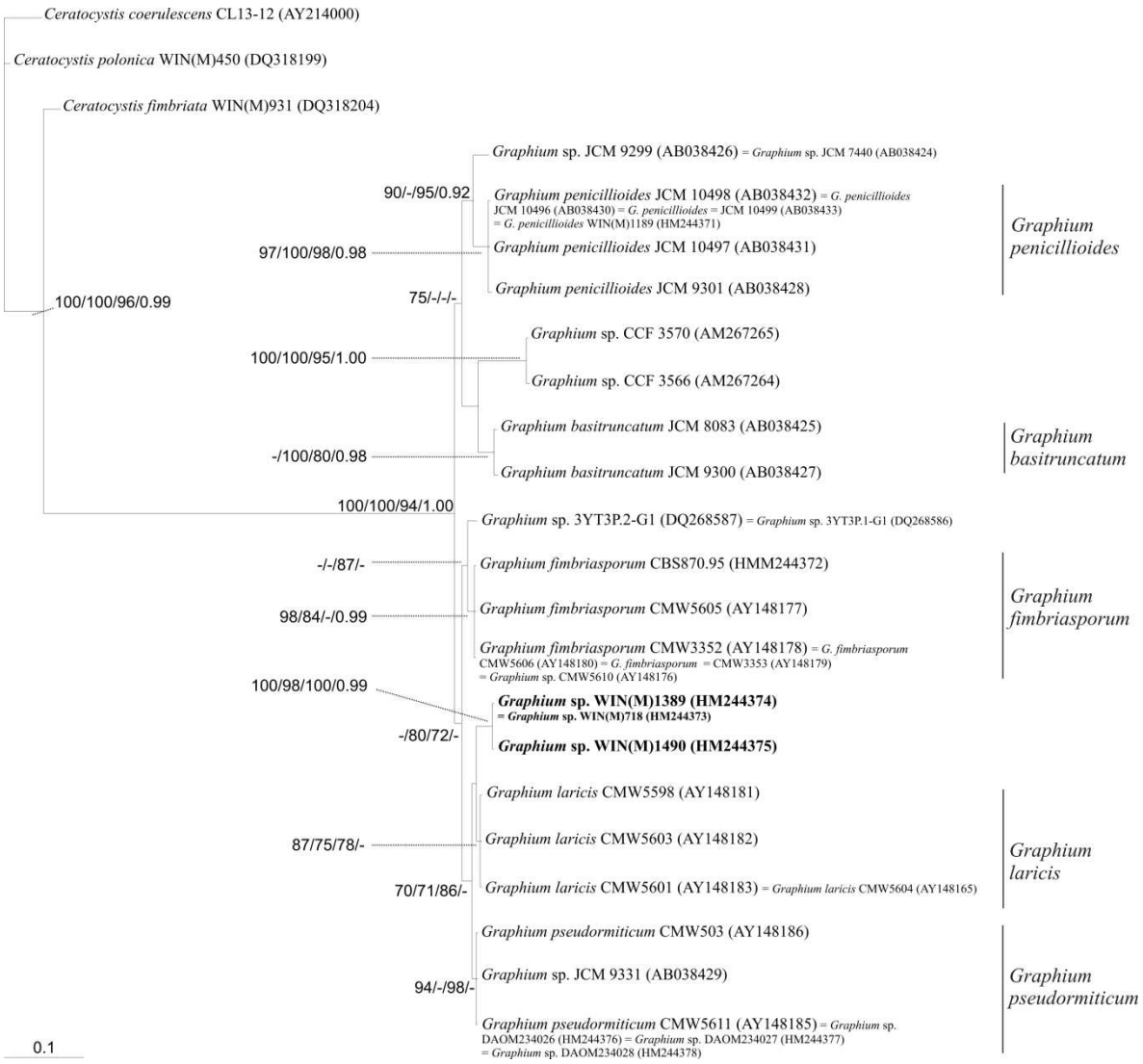
Supplementary Figure 7.4. Alignment of the DNA sequence of the nuclear ITS1-5.8S-ITS2 rDNA region of strains of *Graphium* and related taxa.

```
C. coerulescens CL13-12                   CTGAGTTTTTA-ACTC-TTAAACC-ATATGTG-AACAT-ACCTTT-TTAG
C. polonica WIN(M)450                     CTGAGTTTTTA-ACTC-TTAAACC-ATATGTG-AACAT-ACCTTTTTTAG
C. fimbriata WIN(M)931                    TCGAGTTTTTA-ACTC-TTAAACC-ATATGTG-AACAT-ACCTTTTCTAG
G. penicillioides JCM 10498               CCGAGTGTTTC-ACT--CCAAACCCAC-TGTG-AACCTTACCACT-GTCG
G. penicillioides JCM 10497               CCGAGTGTTTC-ACT--CCAAACCCAC-TGTG-AACCTTACCACT-GTCG
G. penicillioides JCM 9301                CCGAGTGTTTC-ACT--CCAAACCCAC-TGTG-AACCTTACCACT-GTCG
Graphium sp. JCM 9299                     CCGAGTGTTTACGCTCACCAAACCCAC-TGTGTAACCTTACCACT-GTCG
G. fimbriasporum CBS870.95                CCGAGTTTT-ACACTCACCAAACCCAC-TGTG-AACCTTACCAAC-GTCG
G. fimbriasporum CMW5605                  CCGAGTTTTCA-ACTCACCAAACCCAC-TGTG-AACCTTACCAAC-GTCG
G. fimbriasporum CMW3352                  CCGAGTTTTCA-ACTCACCAAACCCAC-TGTG-AACCTTACCAAC-GTCG
Graphium sp. WIN(M)1389                   CCGAGTTTCCACACTCACCAAACCCAC-TGTG-AACCTTACCACC-GTCG
Graphium sp. WIN(M)1490                   CCGAGTTTCCACACTCACCAAACCCAC-TGTG-AACCTTACCACC-GTCG
G. laricis CMW5598                        CCGAGTTTTCACACTCACCAAACCCAC-TGTG-AACCTTACCACT-GTCG
G. laricis CMW5603                        CCGAGTTTTCACACTCACCAAACCCAC-TGTG-AACCTTACCACT-GTCG
G. laricis CMW5601                        CCGAGTTTTCACACTCACCAAACCCAC-TGTG-AACCTTACCACT-GTCG
G. pseudormiticum CMW503                  CCGAGTTTTCA-ACTCACCAAACCCAC-TGTG-AACCTTACCACT-GTCG
Graphium sp. JCM 9331                     CCGAGTTTTCACACTCACCAAACCCAC-TGTG-AACCTTACCACT-GTCG
Graphium sp. 3YT3P.2-G1                   CCGAGTTTTCA-ACTCACCAAACCCAC-TGTG-AACCTTACCACC-GTCG
Graphium sp. CCF 3570                     CCGAGTGTTTC-ACTC-TC-AACCCACCTGTG-AACCTTACC-CC-GTCG
Graphium sp. CCF 3566                     --GAGTGTTTC-ACTC-TC-AACCCACCTGTG-AACCTTACC-CC-GTCG
Graphium basitruncatum JCM 8083           CCGAGTGTTTCTACTC-T-AAACCCAC-TGTG-AACCTTACCACT-GTCG
Graphium basitruncatum JCM 9300           CCGAGTGTTTCTACTC-T-AAACCCAC-TGTG-AACCTTACCACT-GTCG
Graphium pseudormiticum                   CCGAGTTTTCACACTCACCAAACCCAC-TGTG-AACCTTACCACT-GTCG

CTGCTTTGGCAGG-CT-TGGG----TTT---------------GAAA---   -----AAAA--CAAGTC-TG------CCGGCAGTATATAAAAAACACTCT
CTGCTTTGGCAGGTCT-TGG-------T---------------AA-----   -----AA----CAAGTCCTG------CCGGTAGTATTT-ACAAAAACTCT
CTGCTTTGGCAGG-CC---------TTT----------------------   ----CGGGG------TC-TG------CCGGTAGCATTT---AAAAACTCT
TTGCTTCGGCGGG-CG--AAA-CCCCCC---CCC---------GG-----   -----GGGGC-CAG-CC-CG------CCGGCGGCACCC-----AAACTCT
TTGCTTCGGCGGG-CG--AAA-CCCCCC---CC----------GG-----   -----GGGGC-CAG-CC-CG------CCGGCGGCACCC-----AAACTCT
TTGCTTCGGCGGG-CG--AAA-CCCCCC---CC----------G------   -----GGGGC-CAG-CC-CG------CCGGCGGCACCC-----AAACTCT
TTGCTTCGGCGGG-CG-CAGA-CCCCCC---CC----------GGAAGGG   --GA-GGGGCGCGGGCC-CG------CCGACGGCACCC-----AAACTCT
TTGCTTCGGCGGG-CG-CAGA----CCCTC-CCCTC-------GGGG---   -----GGGGCGCGGGCC-CG------CCGGCAGCACCC-----AAACTCT
TTGCTTCGGCGGG-CG-CAGA----CCCTC-CCCTC-------GGGG---   -----GGGGCGCGGGCC-CG------CCGGCAGCACCC-----AAACTCT
TTGCTTCGGCGGG-CG-CAGA----CCCTC-CCCTC-------GGGGG--   -----GGGGCGCGGGCC-CG------CCGGCAGCACCC-----AAACTCT
TTGCTTCGGCGGG-CG-CAGA-CCCCCCT--CCCTCCCC----GGGGGG-   -CGG-GGGGCGCGGGCC-CG------CCGGCAGCACCC----AAAACTCT
TTGCTTCGGCGGG-CG-CAGA-CCCCCCT--CCCTCC----CCGGGGGG-   -CGG-GGGGCGCGGGCC-CG------CCGGCAGCACCC----AAAACTCT
TTGCTTCGGCGGG-CG-CAGA-CCCCCCT--CCCTCC---TCCGGGGGGG   GCG-CGGGGCGCGGGTC-CG------CCGGCAGCACCC----AAAACTCT
TTGCTTCGGCGGG-CG-CAGA-CCCCCCT--CCCTCC---TCCGGGGGGG   -CG-CGGGGCGCGGGTC-CG------CCGGCAGCACCC----AAAACTCT
TTGCTTCGGCGGG-CG-CAGA--CCCCCT--CCCTCC---TCCGGGGGGG   -CG-CGGGGCGCGGGTC-CG------CCGGCAGCACCC----AAAACTCT
TTGCTTCGGCGGG-CG-CAGA---CCCCTC-TCCTCCC--TCCGGGGGGG   -CG-CGGGGCGCGGGCC-CG------CCGGCAGCACCC-----AAACTCT
TTGCTTCGGCGGG-CG-CAGA---CCCCTC-TCCTCCC--TCCGGGGGGG   GCG-CGGGGCGCGGGCC-CG------CCGGCAGCACCCC----AAACTCT
TTGCTTCGGCGGG-CG-CAGA---CCCCTGACCCTCCCCCTC-GGGGGGG   GCGACGGGGCGCGGGCC-CG------CCGGCGGCACCC-----AAACTCT
TTGCTTCGGTGCG-CGCCAGGGACCCCCCG-CCCTCCCC----GGG----   ----CGGGG--CGGGCCCCGGCCGTGCCGGCAGCACCC-----ACATTCT
TTGCTTCGGTGCG-CGCCAGGGACCCCCCG-CCCTCCCC----GGG----   ----CGGGG--CGGGCCCCGGCCGTGCCGGCAGCACCC-----ACATTCT
TTGCTTCGGCGGG-CG------------------TCC------G------   -----------CG--CC-CG------CCGGCAGCACCC-----AAACTCT
TTGCTTCGGCGGG-CG------------------TCC------G------   -----------CG--CC-CG------CCGGCAGCACCC-----AAACTCT
TTGCTTCGGCGGG-CG-CAGA---CCCCTC-TCCTCCC--TCCGGGGGGG   -CG-CGGGGCGCGGGCC-CG------CCGGCAGCACCC----AAAACTCT

TTAATATTTC-TAGAGAATTTATTTTC-ATTGCTGAGTGGC-ATTA-ACA   TAA-TAAGTTAAAACTTTCAACAACGGATCTCTTGGCTCTAGCATCGATG
TTAATATTTC-TAGAGAATTTATTTTC-ATTGCTGAGTGGC-ATTA-ACA   TAA-TAAGTTAAAACTTTCAACAACGGATCTCTTGGCTCTAGCATCGATG
TT-ATATTTC-TATAGAATT-AT--TC-ATTGCTGAGTGGC-ATTA-AC-   TAAATAAGTTAAAACTTTCAACAACGGATCTCTTGGCTCTAGCATCGATG
T--ATAT--CTTA--C-CTA-GC-GTCTCTT-CTGAGT-AC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
T--ATAT--CTTA--C-CTA-GC-GTCTCTT-CTGAGT-AC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
T--ATAT--CTTA--C-CTA-GC-GTCTCTT-CTGAGT-AC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
TCTGTAT--CTTG--CTCTA-GC-GTCTCTT-CTGAGTAAC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
TCTGTAT--C-TA--CCCTA-GC-GTCTCTT-CTGAGT-AC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
TCTGTAT--C-TA--CCCTA-GC-GTCTCTT-CTGAGT-AC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
TCTGTAT--C-TA--CCCTA-GC-GTCTCTT-CTGAGT-AC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
TCTGTAT--A-TA--CCCTA-GC-GTCTCTT-CTGAGT-AC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
TCTGTAT--A-TA--CCCTA-GC-GTCTCTT-CTGAGT-AC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
TCTGTAT--A-TA--CCCTA-GC-GTCTCTT-CTGAGT-AC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
TCTGTAT--A-TA--CCCTA-GC-GTCTCTT-CTGAGT-AC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
TCTGTAT--A-TA--CCCTA-GC-GTCTCTT-CTGAGT-AC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
TCTGTAT--A-TA--CCCTA-GC-GTCTCTT-CTGAGT-AC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
TCTGTAT--A-TA--CCCTA-GC-GTCTCTT-CTGAGT-AC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
TCTGTAT--C-TA--CCCTA-GC-GTCTCTT-CTGAGT-ACAAAAAGAC-   -AAACAA-TCAAA-CTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
T----ATT-T-TA--ACCTA-GA-GTCTCTT-CTGAGT-ACACAAAGAT-   TAAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
T----ATT-T-TA--ACCTA-GA-GTCTCTT-CTGAGT-ACACAAAGAT-   TAAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
T--G-ATTTT-TG-----TA--T-GCCTCT--CTGAGT-AC-TAAAGAC-   AAAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
T--G-ATTTT-TG-----TA--T-GCCTCT--CTGAGT-AC-TAAAGAC-   AAAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG
TCTGTAT--A-TA--CCCTA-GC-GTCTCTT-CTGAGT-AC-AAAAGAC-   -AAACAA-TCAAAACTTTCAACAACGGATCTCTTGGCTGTGGCATCGATG

AAGAACGCAGCGAAATGCGATACGTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAATCTTTGAGCGCACATTGCGCCTGGCAGTATTCTGCCAGGCAT
AAGAACGCAGCGAAATGCGATACGTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAATCTTTGAACGCACATTGCGCCTGGCAGTATTCTGCCAGGCAT
AAGAACGCAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAATCTTTGAACGCACATTGCGCCTGGCAGTATTCTGCCAGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
```

```
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT
AAGAACGCAGCGAAATGCGATAACTAATGTGAATTGCAGAATTCAGTGAA   TCATCGAGTCTTTGAACGCACATTGCGCCCGCTGGTATTCCGGCGGGCAT

GCCTGTCCGAGCGTCATTTCACCACTCAAGCTC---TGCTTGGTGTTGGA   GG-ACCCGCGT-CTTT-TT--TGTTTG-CGGGCCGCC-G-AAATGCATCG
GCCTGTCCGAGCGTCATTTCACCACTCAAGCTC---TGCTTGGTGTTGGA   GG-ACCCGCAT-C-TT-----G------CGGGCCGCC-G-AAATGCATCG
GCCTGTCCGAGCGTCATTTCACCACTCAAGCTC---TGCTTGGTGTTGGA   GG-ACCCGCGT----T-----A--CCG-CGGGCCGCC-G-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CA-CCCCGCGAACCCCCCCGGGGGTCGGCGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CA-CCCCGCGAACCCCCCCGGGGGTCGGCGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CA-CCCCGCGAACCCCCCCGGGGGTCGGCGGG-CGCCCA-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTC-CTCAAGCCCCC-CGCTTGGTGTTGGG   CA--CCCGCCG--CCT--C--GGG-C--CGGG-CGCCCCCAAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CG-CCCCGCGG-CCCT--G--GCGTCG-CGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CG-CCCCGCGG-CCCT--G--GCGTCG-CGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CG-CCCCGCGG-CCCT--G--GCGTCG-CGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CG-ACCCGCGG-CCCC--A-CGG-CCG-CGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CG-ACCCGCGG-CCCC--A-CGG-CCG-CGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CG-CCCCGCGG-CCCC--C--GG-TCG-CGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CG-CCCCGCGG-CCCC--C--GG-TCG-CGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CG-CCCCGCGG-CCCC--C--GG-TCG-CGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CG-CCCCGCGG-CCCT--C--GG-TCG-CGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CG-CCCCGCGG-CCCT--C--GG-TCG-CGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CG-ACCCGCGG-CCCC--C--GCGCCGGCGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CG-CCCCGCGGGTCC------GACCCG-CGGG-CGCCTC-GAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CG-CCCCGCGGGTCC------GACCCG-CGGG-CGCCTC-GAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CGTCCCCGCGGGCCTT--CACGG-CCGGCGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CGTCCCCGCGGGCCTT--CAAGG-CCGGCGGG-CGCCCC-AAATGCATCG
GCCTGTCCGAGCGTCATTTCGTCCCTCAAGCCCCCGCGCTTGGTGTTGGG   CG-CCCCGCGG-CCCT--C--GG-TCG-CGGG-CGCCCC-AAATGCATCG

GCTGTT-ATATTTGCAGCTTCCCTGTGTAGTAATATTTATTTT-ACGCTT   TGAAACTCTTTT--ACAACAT-GCCG-TAAAACCCTAGATTAAATTTTAA
GCTGTT-GAATTTGCAGCTTCCCTGTGTAGTAATATTTATTTT-ACGCTT   TGAAACTCTTTT--ACAACATCGCCG-TAAAACCC--CCCTCAATTATAT
GCTGTT-GTATTTGCAGCTTCCCTGTGTAGTAATGCTTAGCTT-ACACTT   TGAAGCTCTTAT--ATGACAT-GCCGGT-AAAC-----CCTCAAT--T-T
GCGGTCCCGCCTGGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGCGCGCGGCCG-GCCGCT-AAAC----CCCCCAATCGTAC
GCGGTCCCGCCTGGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGCGCGCGGCCG-GCCGCT-AAAC----CCCCCAATCGTAC
GCGGTCCCGC-TGGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGCGCGCGGCCG-GCCGCT-AAAC----CCCCCAATCGTAC
GCGGTCCCGCCTGGCGCC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGC--GCGGCCG-GCCGCT-AAAC-C--CCCCCAATCGTAC
GCGGTCCCGCCTGGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGC--GCGGCCA-GCCGCT-AAAC-C--CCCCCAATCGTAC
GCGGTCCCGCCTGGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGC--GCGGCCA-GCCGCT-AAAC-C--CCCCCAATCGTAC
GCGGTCCCGCCTGGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGC--GCGGCCA-GCCGCT-AAACCC--CCCCCAATCGTAC
GCGGCTCCGCCTGGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGC--GCGGCCT-GCCGCT-AAACAA--CCCCCAATCGTAC
GCGGCTCCGCCTGGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGC--GCGGCCT-GCCGCT-AAACAA--CCCCCAATCGTAC
GCGGTCCCGCCTGGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGC--GCGGCCT-GCCGCT-AAACAA---CCCCAATCGTAC
GCGGTCCCGCCTGGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGC--GCGGCCT-GCCGCT-AAACAA---CCCCAATCGTAC
GCGGTCCCGCCTGGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGC--GCGGCCT-GCCGCT-AAACAA---CCCCAATCGTAC
GCGGTCCCGCCTGGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGC--GCGGCCT-GCCGCT-AAACAA--CCCCCAATCGTAC
GCGGTCCCGC-TGGCG-C-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCC-GGGC--GCGGCCT-GCCGCT-AAACAA--CCCCCAATCGTAC
GCGGTCCCGCCTGGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGC--GCGGCCA-GCCGCT-AAACCC--CCCCCAATCGTAC
GCGGTCCCGCCCAGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGCCCCGGGCGCGCGGCCA-GCCGCC-AAAC----CCCCCCATCGTAC
GCGGTCCCGCCCAGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGCCCCAGGCGCGCGGCCA-GCCGCC-AAAC----CCCCCCATCGTAC
GCGGTCCCGC-TGGCGGC-TCCCTGCGTAGTAGAATCTCTTCTCGCA--T   CGGGTCCCGGGCGCGCGGCCA-GCCGCT-AAAC----CCCCCAATCGTAC
GCGGTCCCGC-TGGCGGC-TCCCTGCGTAGTAGAATCTCTTCTCGCA--T   CGGGTCCCGGGCGCGCGGCCA-GCCGCT-AAAC----CCCCCAATCGTAC
GCGGTCCCGCCTGGCGGC-TCCCTGCGTAGTAGAACCTCTTCTCGCA--T   CGGGTCCCGGGC--GCGGCCT-GCCGCT-AAACAA--CCCCCAATCGTAC

TTTTTTTAAAGTTTATTTTTTTTTTCAAAAGG
-----------------------TTGAA-AGG
-----------------------TTGAA-AGG
-------------------------CAA-CGG
-------------------------CAA-CGG
-------------------------CAA-CGG
-------------------------CAA-CGG
-------------------------CAA-CGG
-------------------------CAA-CGG
-------------------------CAA-CGG
-------------------------CAA-CGG
-------------------------CAA-CGG
-------------------------CAA-CGG
-------------------------CAA-CGG
-------------------------CAA-CGG
-------------------------CAA-CGG
-------------------------CAA-CGG
-------------------------CAA-CGG
-------------------------CAA-CGG
```

```
--------------------------CAA-CG-
--------------------------CAA-CGG
--------------------------CAA-CGG
--------------------------CAA-CGG
```

Supplementary Figure 7.5. Phylogenetic analysis of nuclear ITS1-5.8SrDNA-ITS2 DNA sequences in strains of *Graphium* and related fungal taxa. Branch lengths were determined using the Bayesian consensus outfile. Values at the nodes were determined using algorithms implemented by NJ/DNA PARS/Tree Puzzle/Mr Bayes programs. "-" indicates the node is absent or the posterior probability or bootstrap value is not well supported (a posterior probability value of less than 0.90 for Bayesian analysis and bootstrap values less than 70 % for neighbor joining, parsimony and maximum likelihood analyses). The major clades of *Graphium* species are indicated at the right. Three novel isolates, strains WIN(M)718, WIN(M)1389, and WIN(M)1490, are indicated in bold font.

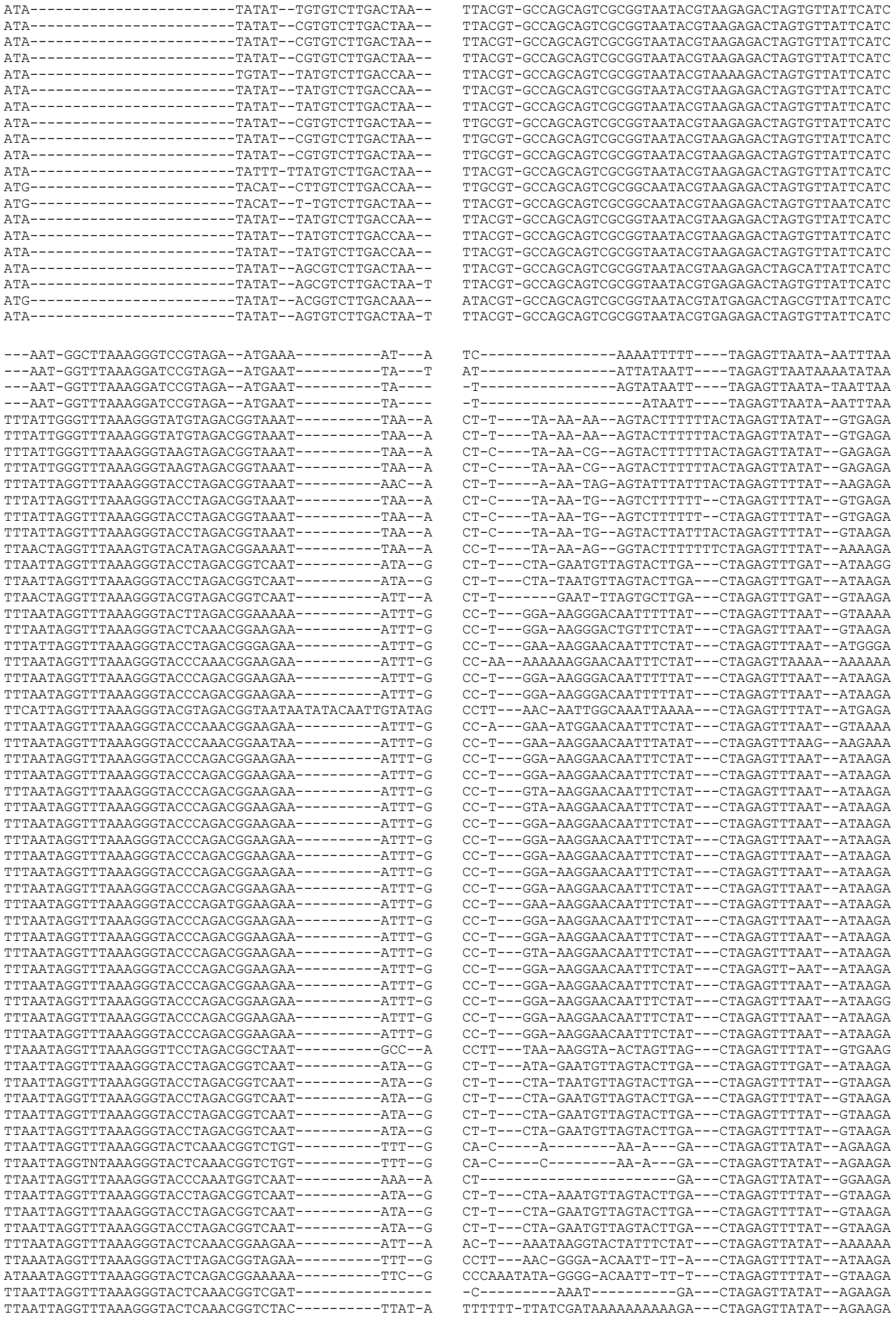

Ceratocystis coerulescens CL13-12 (AY214000)
Ceratocystis polonica WIN(M)450 (DQ318199)
Ceratocystis fimbriata WIN(M)931 (DQ318204)
100/100/96/0.99
Graphium sp. JCM 9299 (AB038426) = *Graphium* sp. JCM 7440 (AB038424)
90/-/95/0.92
Graphium penicillioides JCM 10498 (AB038432) = *G. penicillioides* JCM 10496 (AB038430) = *G. penicillioides* = JCM 10499 (AB038433) = *G. penicillioides* WIN(M)1189 (HM244371)
97/100/98/0.98
Graphium penicillioides JCM 10497 (AB038431)
Graphium penicillioides JCM 9301 (AB038428)
Graphium penicillioides
75/-/-/-
Graphium sp. CCF 3570 (AM267265)
100/100/95/1.00
Graphium sp. CCF 3566 (AM267264)
Graphium basitruncatum JCM 8083 (AB038425)
-/100/80/0.98
Graphium basitruncatum JCM 9300 (AB038427)
Graphium basitruncatum
100/100/94/1.00
Graphium sp. 3YT3P.2-G1 (DQ268587) = *Graphium* sp. 3YT3P.1-G1 (DQ268586)
-/-/87/-
Graphium fimbriasporum CBS870.95 (HMM244372)
98/84/-/0.99
Graphium fimbriasporum CMW5605 (AY148177)
Graphium fimbriasporum CMW3352 (AY148178) = *G. fimbriasporum* CMW5606 (AY148180) = *G. fimbriasporum* = CMW3353 (AY148179) = *Graphium* sp. CMW5610 (AY148176)
Graphium fimbriasporum
100/98/100/0.99
***Graphium* sp. WIN(M)1389 (HM244374)** = ***Graphium* sp. WIN(M)718 (HM244373)**
-/80/72/-
***Graphium* sp. WIN(M)1490 (HM244375)**
Graphium laricis CMW5598 (AY148181)
87/75/78/-
Graphium laricis CMW5603 (AY148182)
70/71/86/-
Graphium laricis CMW5601 (AY148183) = *Graphium laricis* CMW5604 (AY148165)
Graphium laricis
Graphium pseudormiticum CMW503 (AY148186)
94/-/98/-
Graphium sp. JCM 9331 (AB038429)
Graphium pseudormiticum CMW5611 (AY148185) = *Graphium* sp. DAOM234026 (HM244376) = *Graphium* sp. DAOM234027 (HM244377) = *Graphium* sp. DAOM234028 (HM244378)
Graphium pseudormiticum
0.1

Supplementary Figure 7.6 Alignment of the DNA sequence of the mt *rns* gene.

```
K. thermotolerans CBS6340                                TG-GTTT-AGGTAGTAGGTAGAGTAA-GA-GTTTA-CCTAGCC-AACGAT
S. cerevisiae                                            TG-GTTT-AGGTAGTAGGTTTATTAA-GA-GTTAAACCTAGCC-AACGAT
Saccharomyces sp. IFO 1802                               TG-GTTT-AGGTAGTAGGTTTTAAATAAAAGTTAA-CCTAGCC-AACGAT
S. paradoxus  NRRL Y-17217                               TG-GTTT-AGGTAGTAGGTTTATTAA-GA-GTTAAACCTAGCC-AACGAT
P. glabrum                                               TC-G-GTGGAGTAGTGGGAGAGGTAATGA-ATTTT-TCTAGC--AGTAA-
P. thomii                                                TC-G-GT-GAGTAGTAGGAAAGGTAATGA-CTTTT--CTAGC--AGTAA-
P. expansum                                              TC-G-GT-GAGTAGTAGGAAAGGTAATGA-CTTTT--CTAGC--AATAAA
P. chrysogenum                                           TC-G-GT-GAGTAGTAGG-AAGGTAATGA-CTTTT--CTAGC--AATAAA
P. marnefei                                              TC-G-GTGAAGTAGTAGAAGAGGTAATGT-CTTTT-TCTAGCT-AACA--
A. niger                                                 TC-G-GA-GAGTAGTAGGAAAGGTAATGA-CTTTT--CTAGCT-AATA-T
A. tubingensis                                           TC-G-GA-GAGTAGTAGGAAAGGTAATGA-CTTTT--CTAGCT-AATA-T
E. nidulans                                              AC-G-GT-GAGTAGTAGTAAAGGTAATGA-CTTTA--CTAGC--TAAA-T
E. floccosum                                             TC-G-GT-TTGTAATAGTTAAA-TAATAA-AT-AA--TTAAC--AAAA--
C. prolifica                                             TC-GAGATAGGTAGTAGTTAAGGTGATGG-CTTAA--CTAGCC-AAAACT
M. anisopliae                                            TC-GGGATAGGTAGTAGTTAAGGTGATGG-CTTAA--CTAGCCTAAAACT
O. konnoana                                              TC-GAGAAAGGTAGTAGTTAAGGTAATGA-CTTAA--CTAGCCTAGAACT
O. microsporum WIN(M)513                                 TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCTACGATT
O. coronatum WIN(M)43                                    TA-GAGATAGGTAGTAGTTAAGGTAATGG-CTTAA--CTAGCCCATGATT
G. pseudoeurophioides WIN(M)42                           TA-GAGATAGGTAGTTGTTAAGGN-ATGG-CTTAA--CTAGCCCAAAA--
C. ossiformis WIN(M)52                                   AA-GGGATAGGTAGTTGTTAAGGTAATGG-TTTAA--CTAGCCGAATATT
G. crassivaginata WIN(M)1589                             TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
G. penicillata NFRI60-21                                 TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
M. grisea                                                TC-GGGTTAGGCAGTAGTTAAGGTAATGG-CTTAA--CTAGCTTACAATA
O. grande CBS250.78                                      CA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCAATTATT
O. pseudonigrum WIN(M)47                                 TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCAATTATT
L. truncatum NFRI1813/1 intron+                          TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. wingfieldii TOM1.3                                    TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. lundbergii DAOM60397                                  TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. lundbergii NFRI89-1040/1/3 intron+                    TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. truncatum Forintek C34                                TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. terebrantis CBS337.70 intron+                         TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. truncatum CBS929.85 intron+                           TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. truncatum CBS929.85 intron-                           TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. truncatum J.R.88-449                                  TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. lundbergii NFRI89-1040/1/3 intron-                    TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCACGATT
L. lundbergii NFRI1502/1 intron-                         TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. wingfieldii TOM9.4                                    TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. lundbergii NFRI1502/1 intron+                         TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. terebrantis CBS337.70 intron-                         TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAGATT
L. truncatum TOM74.29                                    TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. truncatum NFRI59-7/3                                  TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. lundbergii NFRI69-148                                 TA-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
L. truncatum NFRI1813/1 intron-                          -A-GAGATAGGTAGTTGTTAAGGTAATGG-CTTAA--CTAGCCCAAAATT
C. parasitica                                            TC-GGGATAGGTAGTAGTTAAGGTAATGG-CTTAA--CTAGCCAATAATT
C. ramosopulvinata                                       TC-GAGATAGGTAGTAGTTAAGGTGATGG-CTTAA--CTAGCC-AAAACT
C. japonica                                              TC-GAGATAGGTAGTAGTTAAGGTGATGG-CTTAA--CTAGCCTGAAACT
C. ophioglossoides                                       TC-GAGATAGGTAGTAGTTAAGGTGATGG-CTTAA--CTAGCCTAAAACT
C. inegoensis                                            TC-GAGATAGGTAGTAGTTAAGGTGATGG-CTTAA--CTAGCCTAAAACT
C. paradoxa                                              TC-GAGATAGGTAGTAGTTAAGGTGATGG-CTTAA--CTAGCCTGAAACT
B. bassiana                                              TC-GGGATAGGTAGTAGTGAAGGTTATGT-CTTCA--CTAGCCTC-AACT
B. brongniartii                                          TC-GGGATAGGTAGTAGTGAAGGTTATGT-CTTCA--CTAGCCTT-AACT
Cordyceps sp. 97009                                      TC-GAGGTAGGTAGTAGTGAAGGTTATGT-CTTGA--CTAGCCTT-AATT
H. jecorina                                              TC-GGGATAGGTAGTAGTTAAGGTGATGG-CTTAA--CTAGCCTAAAACT
H. lutea                                                 TC-GGGATAGGTAGTAGTTAAGGTGATGG-CTTAA--CTAGCCTAAAACT
H. chrysospermus                                         TC-GGGATAGGTAGTAGTTAAGGTGATGG-CTTAA--CTAGCCTAAAACT
G. penicillata NFRI60-21                                 AA-GAAATAGGTAGTTGTTAAGGTAATTA-CTTAA--CAAGCCAATTATT
P. anserina                                              TCGGGGAAAGGTAGTTGTTAAGGTAATTG-CTTAA--CAAGCCTAGTATT
N. crassa                                                TCAGAGATAGGTAGTTGTTAAGGTCATGG-CTTAA--CAAGCCAATAATT
C. militaris                                             TC-GGGATAGGTAGTAGTGAAGGTTATGT-CTTCA--CTAGCCTT-AACT
P. tenuipes                                              AC-GAGATAGGTAGTAGTGAAGGTTATGT-CTTCA--CTAGCCTT-AACT
L. muscarium                                             TC-GGAATAGGTAGTAGTGAAGGTTATGT-CTTCA--CTAGCCTT-AACT
C. coccidiicola                                          TC-GGGATAGGTAGTAGTTAAGGTGATGA-CTTAA--CTAGCCGACAACT
O. sobolifera                                            TC-GGGATAGGTAGTAGTTAAGGTGATGA-CTTAA--CTAGCCTTAAACT
V. dahliae                                               TC-GAAATAGGTAGTAGTTAAGGTGATGA-CTTAA--CTAGCT-TAGCAT
Cordyceps sp. 97003                                      TC-GGAATAGGTAGTAGTCAAGGTGATGG-CTTAA--CTAGCCTTAAACT

C--CATAATCGATAATTAA---AGTTAGAACGATCACGTTGACTCTGAAA   CATAGTCAATATCTAAATG-ATACAGCAGTGAGGAATATTGGGCAATGAT
C--CATAATCGATAATGAA---AGTTAGAACGATCACGTTGACTCTGAAA   TATAGTCAATATCTATAAG-ATACAGCAGTGAGGAATATTGGACAATGAT
C--CATAATCGATAATGAA---AGTTAGAACGATCACGTTGACTCTGAAA   TATAGTCAATATCTATAAG-ATACAGCAGTGAGGAATATTGGACAATGAT
C--CATAATCGTTAATGAA---AGTTAGAACGATCACGTTGACTCTGAAA   TATAGTCAATATCTATAAG-ATACAGCAGTGAGGAATATTGGACAATGAT
C--CGTAGTCGTAACTGGAA--AGTT-GATCGACCACATTGGGTCTGAAA   AAAACCCAATGCTTTTAT--GTACAGCAGTGAGGGATATTGGTCAATGGC
C--CGTAGTCGTAACTGGAA--AGTT-GATCGACCACATTGGGTCTGAAA   AAAACCCAATGCTTTTAT--GTACAGCAGTGAGGAATATTGGTCAATGGC
C--CGTAGTCGTAACTGGAA--AGTT-GATCGACCACATTGGGTCTGAAA   AAAACCCAATGCTTTTAT--GTACAGCAGTGAGGAATATTGGTCAATGGC
C--CGTAGTCGTAACTGGAA--AGTT-GATCGACCACATTGGGTCTGAAA   AAAACCCAATGCTTTTAT--GTACAGCAGTGAGGAATATTGGTCAATGGC
CT-CGTAGTCGTGACTGAGA--GGTC-GATCGACCACATTGGGTCTGAAA   AAACCCCAATGCGTATTA--GTACAGCAGTGAGGGATATTGGTCAATGGT
C--CGTAGTCGTGACTGAGA--GGTC-GATCGACCACATTGGGTCTGAGA   AAACCCCAATGCGTATTA--GTACAGCAGTGAGGAATATTGGTCAATGGC
C--CGTAGTCGTGACTGAGA--GGTC-GATCGACCACATTGGGTCTGAGA   AAACCCCAATGCGTATTA--GTACAGCAGTGAGGAATATTGGTCAATGGC
C--CGTAGTCGTGACTGAGA--GGTC-GATCGACCACATTGGGTCTGAAA   AAACCCCAATGCGTTTTA--GTACAGCAGTGAGGAATATTGGTCAATGGC
C--CGTAGTCGTAACTGAGA--AGTT-GATCGACCACATTAGGACTGAGA   AACTCCTAATGCGCTTTAA-GTACAGCAGTGAGGGATATTGGTCAATGAC
CT-CGTAGTCGAATCTGAAA--GGTT-GATCGACCACATTGGGTCTGAAA   AAATCCCAATGCAAGTTA--GTACAGCAGTGAGGAATCTTGGTCAATGGC
CT-CGTAGTCGAATCTGAAA--GGTT-GATCGACCACATTGGGTCTGAAA   AAACCCCAATGCAAGTTA--GTACAGCAGTGAGGAATCTTGGTCAATGGC
```

```
CT-CGTAGTCGAAGCTGAAA--GGTT-GATTGTCCACATTGGGTCTGAAA   AAAGCCCCATACTATGTAAAGTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGATGAAA   AAATCCCAAAACAATCGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAA--GGTT-GATCGACCACATTGGGGATGAAA   AAATCCCAAGGCAAGCGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTTAAT--GGTT-GATCGACCACATTGGGAATGAAA   AAATCCCAATACATGCAA--GTACGGCTTTGGGCAATATTGGTCATTGGT
CT-CTTAGTCGAAACTGAAAG-GGTT-GATCGACCACATTGGGGATGAGA   AAAACCCAAGGCAAGCGA--GTACAGCAGTGGGGAATTTTAGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CA-CGTAGTCGAAGCTGAGA--GGTT-GATCGACCACATTGGGGATGAAA   AAATCCCAAGGCAAGTAG--GTACAGCAGTGAGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGATGAGA   AAATCCCAAGGCAAGCGA--GTACAGCAGTGGGGAATTTTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGATGAGA   AAATCCCAAGGCAAGCGA--GTACAGCAGTGGGGAATTTTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAGTACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGT
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGT
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGATGAAA   AAATCCCAAGACAAGCGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGT
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CTTAGTCGAAACTGAAAT-GGTT-GATCGACCACATTGGGGCTGAAA   AAAACCCAATACAAGTGA--GTACAGCAGTGGGGAATATTGGTCAATGGC
CT-CGGAGTCGAAGCTGAAAT-GGTT-GATCGACCACATTGGGAATGAAA   AAATCCCAAGGCATAATT--GTACAGCAGTGAGGAATTTTGGTCAATGCG
CT-CGTAGTCGAATCTGAAA--GGTT-GATCGACCACATTGGGTCTGAAA   AAACCCCAATGCAAGTTA--GTACAGCAGTGAGGAATCTTGGTCAATGGC
CT-CGTAGTCGAATCTGAAA--GGTT-GATCGACCACATTGGGTCTGAGA   AAACCCCAATGCAAGTTA--GTACAGCAGTGAGGAATATTGGTCAATGGC
CT-CGTAGTCGAATCTGAAA--GGTT-GATCGACCACATTGGGTCTGAAA   AAATCCCAATGCAAGTTA--GTACAGCAGTGAGGAATCTTGGTCAATGGC
CT-CGTAGTCGAATCTGAAA--GGTT-GATCGACCACATTGGGTCTGAAA   AAACCCCAATGCAAGTTA--GTACAGCAGTGAGGAATCTTGGTCAATGGC
CT-CGTAGTCGAATCTGAAA--GGTT-GATCGACCACATTGGGTCTGAAA   AAACCCCAATGCAAGTTA--GTACAGCAGTGAGGAATCTTGGTCAATGGC
CC-CGTAGTCGAAATTGAAA--GGTT-GATCGACCACATTGGGTCTGAAA   AAAGCCCAATGCAAGTTA--GTACAGCAGTGAGGAATATTGGTCAATGGC
CT-CGTAGTCGAAGCTGAAA--GGTT-GATCGACCACATTGGGTCTGAAA   AAAGCCCAATGCAAGTTA--GTACAGCAGTGAG-AATATTGGTCAATGGC
CT-CGTAGTCGAAACTGAAA--GGTT-AATCGACCACATTGGGTCTGAAA   AAATCCCAATGCAAGTTA--GTACAGCAGTGAGGAATATTGGTCAATGGC
CT-CGTAGTCGAAGCTGAAA--GGTT-GATCGACCACATTGGGTCTGAAA   AAACCCCAATGCAAGTTA--GTACAGCAGTGAGGAATTTTGGTCAATGGC
CT-CGTAGTCGAAGCTGAAA--GGTT-GATCGACCACATTGGGTCTGAAA   AAACCCCAATGCAAGTTA--GTACAGCAGTGAGGAATTTTGGTCAATGGC
CT-CGTAGTCGAAGCTGAAA--GGTT-GATCGACCACATTGGGTCTGAAA   AAACCCCAATGCAAGTTA--GTACAGCAGTGAGGAATTTTGGTCAATGGC
TT-CTTAGTCGTAACTGAAA--GGTT-GATCGATCACATTGGGAATGAGA   AAAGCCCAAGACAAATGA--GTACAGCAGTGGGGAATATTAGTCAATGGC
CT-CTTAGTCGAAGCTGAGAAAGGTT-GATCGACCACATTGGGGATGAAA   AAATCCCAAGGCAAACGAA-GTACAGCAGTGAGGAATATTGGTCAATGGC
CT-CTTAGTCGAAGCTGAAAA-GGCT-GATCGACCACATTGGGAATGAAA   AAATCCCAAGGCAAATAG--GTACAGCAGTGAGGAATCTTGGTCAATGGG
CT-CGTAGTCGAAGCTGAAA--GGTT-AATCGACCACATTGGGTCTGAAA   AAAGCCCAATGCAAGTTA--GTACAGCAGTGAGGAATATTGGTCAATGGC
CT-CGTAGTCGAAGCTGAAA--GGTT-AATCGACCACATTGGGTCTGAAA   AAAGCCCAATGCAAGTTA--GTACAGCAGTGAGGAATATTGGTCAATGGC
CT-CGTAGTCGAAGCTGA----AGTT-AATCGACCACATTGGGTCTGAAA   AAAGCCCAATGCAAGTTA--GTACAGCAGTGAGGA-TATTGGTCAAT--C
CT-CGTAGTCGAAGCTGAAA--GGTT-GATCGACCACATTGGGTCTGAAA   AAACCCCAATGCAAATTA--GTACAGCAGTGAGGAATTTTGGTCAATGGC
TT-CGTAGTCGAATCTGAGA--GGTT-GATCGACCACATTGGGTCTGAAA   AAACCCCAATGCAATTTA--GTACAGCAGTGAGGAATCTTGGTCAATGGT
TTTCGTAACCGTAACTGAAA--AGTT-GATCGGTCACATTGGGTCTGAAA   AAAGCCTAATGCAATATAACGTACAGCAGTGGGGAATCTTGGTCAAAT-C
TT-CGTAGTCGAATCTGAGA--GGTT-GATCGACCACATTGGGTCTGAAA   AAACCCCAATGCAATTTA--GTACAGCAGTGAGGAATCTTGGTCAATGGC

CGAAAGATTGACCCAGTTACTTATTAGGATGA-TATAGATATTAGCTA--   --------CC---ATAATAAATA-T-TATGGATAAA---ATT---AAT--
CGAAAGATTGATCCAGTTACTTATTAGGATGA-TATAGATATTAATTA--   --------CC---ATATATATTTTTATATGGATATATATATTAATAAT-A
CGAAAGATTGATCCAGTTACTTATTAGGATGA-TATAGATATTAATTA--   --------CT---ATATAAAAAA-TATATAGATATAAAAAAA-GTAAT-A
CGAAAGATTGATCCAGTTACTTATTAGGATGA-TATAGATATTAATTA--   --------CC---ATATAATAAT-AATATGGATATATAATAAT-T-AT-A
CGCAAGGCTGAACCAGTAACTTGGAAGAATGT--AACGA--TTAAT----   --------TC---GCATAAAATTCTAAATAGATTGTATATAATGACAT-A
CGCAACG-TGAACCAGTAACTTGGAAGAATGT--AACGA--TTAAT----   --------TC---GCATAAAATTCTAAATAGATTGTATATAATGACAT-A
CGAAACG-TGAACCAGTAACTTGGAAGAATGA--GAAGA--TTAAC----   --------TC---GCATAAAATTCTAAATAGATTGTATATAATGACAT-A
CGAAACG-TGAACCAGTAACTTGGAAGAATGT--CTAGA--TTAAC----   --------TC---GCATAAAATTCTAAATAGATTGTATATAATGACAT-A
CTAACGACTGAACCAGTAACTTGGAAGAATGT----AGA--TTAAT----   --------TC---GTATAAAATTCTAAATAGGATAATGATAATGACAATT
CGAAAGGCTGAACCAGTAACTTGGAAGAATGT--AAAGA--TTAAC----   --------TC---GTATAAAATTCTAAATAGGATAATGATAATGACAATT
CGAAAGGCTGAACCAGTAACTTGGAAGAATGT--AAAGA--TTAAC----   --------TC---GTATAAAATTCTAAATAGGATAATGATAATGACAATT
CGAAAGGC-GAACCAGTAACTTGGAAGAATGA--AAGGA--TTATA----   --------TC---GTATAAAATTCTAAATAGAATAATGATAATGACAATT
CTAACGGTTGAACCAGCAACTTGGAAGAATGC---AAAA--TTAA-----   CGTGTAA-AC---GTATAAAATTCTAAATAGATTAAAGATAATGACAATT
CTAACGGCTGAACTGGCAACTTGGAGAAGTGG-TAAGGA--TTGTA----   TTAAAA--TT---GAATAAAGCTTTGTTTATATA-TTGATAATGACAG-T
CTAACGGCTGAACTGGCAACTTGGAGAAGTGG-CAAGGA--TTGTA----   TTAAAA--TT---GAATAAAGCTTTGTTTATATA-TCGATAATGACGG-T
CTAACGGCTGAACTGGCAACTTGAAAAAGTGG-CAAGGA--TTCGA----   TTAAAAT-TT---GAATGAAGCTTTGTTTATATATAATATAATGACAA-T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATAAATC---GAATAAAATTCTAAATATATTTATTATAATGAC---T
CTAACGGCTGAACTGGCAATTTGGAGGAATGG-CCTAGG--TTAAA----   TCGAAGT-TC---GAATAAAATTCTGAATATATATGTTATAATGAC---T
CTAACGACTGAACTGGCTTCTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTC-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAATTTA   TCGCNTGGTA---AAATAAAATTCTAAATATATAACTTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGA--GAGGA--TTAAA----   TCGAACT-TC---GAATAAAATTCTGTATGTATATGGTATAATGACAA-T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGAATA-TC---GAATAAAATTCTAAATATATAGTTTATAATGAC---A
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGAATA-TC---GAATAAAATTCTAAATATATAATTTATAATGAC---A
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGACTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGACTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
```

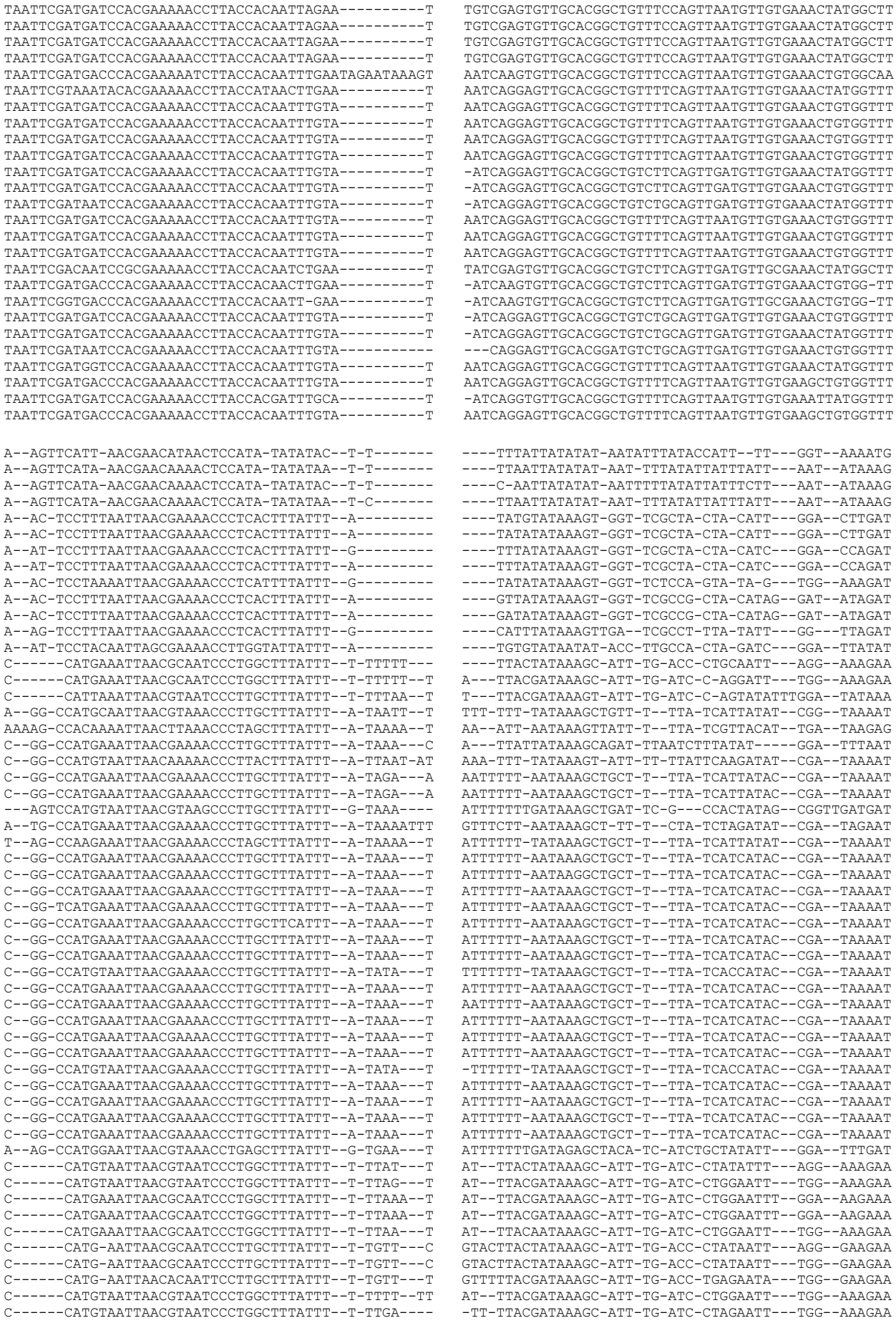
```
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTATATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAGATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGACTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGAGGAATGG-CCTAGG--TTAAA----   TCGATTA-TC---GAATAAAATTCTAAATATATAAATTATAATGAC---T
CTAACGGCTGAACTGGCAACTTGGATGAACGGACCGAGG--TGCAGGAGA   TCTTTT--TTAG-GAGTAAAGTTCTAAATATATGATTGATAATGACAA-T
CTAACGGCTGAACTGGCAACTTGGAGAAGTGG-TAAGGA--TTGTA----   TTAAAA--TT---GAATAAAACTTTGTTTACATA-TTGATAATGACAA-T
CTAACGGCTGAACTGGCAACTTGGAGAAGTGG-CAAGGA--TTGTA----   TTAAAA--TT---GAATAAAGCTTTGTTTATATA-TTGATAATGACAG-T
CTAACGGCTGAACTGGCAACTTGGAAAAGTGG-CAAGGA--TTGTA----   TTAAAA--TT---GAATAAAGCTTTGTTTATATA-TTGATAATGACAG-T
CTAACGGCTGAACTGGCAACTTGGAGAAGTGG-CAAGGA--TTGTA----   TTAAAA--TT---GAATAAAGCTTTGTTTATATA-TTGATAATGACAG-T
CTAACGGCTGAACTGGCAACTTGGAGAAGTGG-TAAGGA--TTGTA----   TTAAAA--TT---GAATAAAGCTTTGTTTATATA-TTGATAATGACAG-T
CTAACGGCTGAACTGGCAACTTGGAGAAGTGT-TAAGGA--TTAAA----   TTAGTA--TT---GAATGAAACTTTGTTTATATA-TCGATAATGACGA-T
CTAACGGCTGAACTGGCAACTTGGAGAAGTGT-TAAGGA--TTAAA----   TTAGTA--TT---GAATGAAACTTTGTTTATATA-TCGATAATGACGA-T
CTAACGGCTGAACTGGCAACTTGGAGAAGTGA-TAAGGA--TTAAA----   TTAGTA--TT---GAATAAAACTTTGTCTATATA-TTGATAATGACGA-T
CTAACGGCTGAACTGGCAACTTGGAGAAGTGG-CAAGGA--TTGTA----   TTAAAA--TT---GAATAAAGCTTTGTTTATATA-TTGATAATGACAG-T
CTAACGGCTGAACTGGCAACTTGGAGAAGTGG-CAAGGA--TTGTA----   TTAAAA--TT---GAATAAAGCTTTGTTTATATA-TTGATAATGACAG-T
CTAACGGCTGAACTGGCAACTTGGAGAAGTGG-CAAGGA--TTGTA----   TTAAAA--TT---GAATAAAGCTTTGTTTATATA-TTGATAATGACAG-T
CTAACGGCTGAACCGGCAACTTGGAGGAATGG-CCTAGG--TTAAATTTT   TCTTTT--AA---AAATAAAATTCTGATTGTATATAATATAATGAC---T
CTAACGGCTGAACTGGTAACTTGGAAGAATGG-C---AG--TTAA-----   TCGAT---TTATCGAATAAAATTCTGTATACATA-CCTATAATGACGG-T
CCCACGCCTGAACTGGTAACTTGGAGGAATGA-GGGGGAC-TCCAC---G   TCGAT---TTATCGAGTAAAATTCTGTATACATATTGGATAATGACAA-T
CTAACGGCTGAACTGGCAACTTGGAGAAGTGT-TAAGGA--TTAAA----   TTAGTA--TT---GAATGAAACTTTGTTTATATA-TCGATAATGACGA-T
CTAACGGCTGAACTGGCAACTTGGAGAAGTGT-TAAGGA--TTAAA----   TTAGTA--TT---GAATGAAACTTTGTTTATATA-CCGATAATGACGA-T
GTAACGGCTGAACTGGCAACTTGGAGAAGTGT-TAAGGA--TTAAA----   TTAGTA--TT---GAATGAAACTTTGTTTATATA-TCGATAATGACGA-T
CTAACGGCTGAACTGGCAACTTGGAAAAGCGG-CAAGGA--TTTTA----   TTGTAA--AA---GAATGAAGCTTTGTTTATATA-TTGATAATGACAA-T
CTAAAGACTGAACTGGCAACTTGGAGAAGGGG-TAAGGA--TTGAA----   TTAGGTA-AT---AAATGAAACTTTGTTTATATA-TAGATAATGACAA-T
CTAACGGCCGAACTGGCAACTTGAAAGAATGG-GAAGAA--TTGA-----   TTAA----TC---GAATAAAATTTTAAATACATT-GATATAATGACGA-T
CTAATGGCTGAACTGGCAACTTGGAGAAGTGG-TAAGGA--TTGAA----   TTAAGAA-AG---AAATGAAACTTTGTTTATATA-TAGATAATGACAG-T

AT-TTAATTT-----------------TATTTTAATAGTCCTAACCAA-A   ATTTGT-GCCAGCAGCTGCGGTAAGACAATGGGGGCTAGCGTTAATCAT-
ATATTAATTTTATTATTATTAATAATATATTTTAATAGTCCTGACTAA-T   ATTTGT-GCCAGCAGTCGCGGTAACACAAAGAGGGCGAGCGTTAATCAT-
AT-TTA-TT--ATTATC-TTCTTTTTATATTTTAATAGTCCTGACTAA-T   ATTTGT-GCCAGCAGTCGCGGTAACACAAAGAGGGCAAGCGTTAATCAT-
ATATTAATTT-ATTATAATAAATTATATATTTTAATAGTCCTGACTAA-T   ATTTGT-GCCAGCAGTCGCGGTAACACAAAGAGGGCGAGCGTTAATCAT-
ATC------------------------TATTT-ATAAGTCTTGACCAA-A   CTACGTTGCCAGCAGTCGCGGTAATACGTAGAAGGCTAGTGTTAGTTATC
ATC------------------------TATTT-ATAAGTCTTGACCAA-A   CTACGT-GCCAGCAGTCGCG-TAATACGTAGAAGGCTAGTGTTAGTTATC
ATC------------------------TATTT-ATAAGTCTTGACCAA-A   CTACGT-GCCAGCAGTCGCG-TAATACGTAGGAGGCTAGTGTTAGTTATC
ATC------------------------TATTT-ATAAGTCTTGACCAA-A   CTACGT-GCCAGCAGTCGCG-TAATACGTAGGAGGCTAGTGTTAGTTATC
TCC------------------------TATTT-ATAAGTCTTGACCAA-A   TTACGT-GCCAGCAGTCGCGGTAATACGTAGAAGACTAGTGTTAGTCATC
TCC------------------------TATTT-ATAAGTCTTGACCAA-A   TTACGT-GCCAGCAGTCGCGGTAATACGTAGAAGACTAGTGTTAGTCATC
TCC------------------------TATTT-ATAAGTCTTGACCAA-A   TTACGT-GCCAGCAGTCGCGGTAATACGTAGAAGACTAGTGTTAGTCATC
TTC------------------------TATTT-ATAAGTCTTGACCAA-A   CTACGT-GCCAGCAGTCGCGGTAATACGTAGAAGACTAGTGTTAATCATC
ATC------------------------TATTT-ATTAGTCTCGACCAATA   CTACGT-GCCAGCAGTCGCGGTAAGACGTAAGAGACAAGTGTAAGTCATC
ATA------------------------TGTAT--CGTGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTATTCATC
ATA------------------------TATAT--CGTGTCTTGGCTAA--   TTACGT-GCCAGCAGCCGCGGTAATACGTAAGAGACTAGTGTTATTCATC
ATA------------------------TATAT--TTAGTCTTGACTAA--   TTGCGT-GCCAGCAGTCGCGGTAATACGTAAAAGATAAGTTTTAATCATG
ATA------------------------TATAT-TTTTGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTATTCATC
GTA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTATTCATC
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTAATCATT
ATG------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTATTCATC
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATCT--TATGTCTTGACAAA--   TTTCGT-GCCAGCAGTCGCGGTAATACGAAAGAGACTAGTGTTATTCATC
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTATTCATC
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTATTCATC
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TAT-T--TATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCAT-
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT-TTATGTCTTGACTAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTACTCATT
ATA------------------------TATAT--TATGTCCTGACCAA--   TTACGT-GCCAGCAGTCGCGGTAATACGTATGGACTAGTGTTATTCATC
ATG------------------------TATAT--TATGTCTTGACTAA--   TACGT-GCCAGCAGTCGCGGTAATACGTAAGAGACTAGTGTTATTCATC
```



```
TTAATTAGGTTTAAAGGGTACTCAAACGGTCCTG----------TTC--G   CA-T-----A--------AA-A--AGA---CTAGAGTTATAT--AGAAGA
TTAATTGGGTTTAAAGGGTACCTAGACGGTCAAT----------ATA--G   CT-T---ATA-GAATGTTAATACTTGA---CTAGAGTTTTAT--ATGAGA
TTAATTGGGTTTAAAGGGTACCTAGACAGTCAAT----------ATA--A   CT-T---CTA-GAATGCTAATACTTGA---CTAGAGTTTTAA--GTAAAA
TTAAATAGGTTTAAAGGGTACTCAAACGGTCAGA----------TTAT-C   CT-T----TA-ATA-GGTAATATTTGT---CTAGAGTTTTAT--GTAAGA
TTAATTGGGTTTAAAGGGTACCTAGACAGTCAAT----------ATA--A   CT-T---CTA-GAATGCTAATACTTGA---CTAGAGTTTTAA--GTAAGA

TTAAAGAATT-ATAATAGT-AAGGATGAAATGATAATAATCAT-TA---T   AAGACT-GGTATGTGTGAAGG-TAT-TAATTAAATATTAACTGAC-ATTG
TTAAAGAATT-ATAATAGT-AAAGATGAAATAATAATAATAAT-TA---T   AAGACT-AATATATGTGAAAA-TAT-TAATTAAATATTAACTGAC-ATTG
TTAAAGAATT-ATAATAGT-AAAGATGAAATAATAATAATAAT-TA---T   AAGACT-AATATATGTGAAAA-TAT-TAATTAATAATTAACTGAC-ATTG
TTAAAGAATT-ATAATAGT-AAAGATGAAATAATAATAATAAT-TA---T   AAGACT-AATATATGTGAAAA-TAT-TAATTAAATATTAACTGAC-ATTG
AGGAAGAATT-CTTGGAGT-AGAGATAGAATT-TGTTGATACC-AA---G   AGGACT-GTCAACGGCGAAGG-CGTCNTTCTATGTAATAACTGAC-GTTG
AGGAAGAATT-CTTGGAGT-AGAGATAGAATT-TGTTGATACC-AA---G   AGGACT-GTCAACGGCGAAGG-CGTCCTTCTATGTAATAACTGAC-GTTG
AGGAAGAATT-TCTGGAGT-AGAGATAAAATT-TGATGATACC-AG---G   AGGACT-GTCAACGGCGAAGG-CGTCCTTCTATGTAATAACTGAC-GTTG
AGGAAGAATT-TCTGGAGT-AGAGATAAAATT-TGATGATACC-AG---G   AGGACT-GTCAACGGCGAAGG-CGTCCTTCTATGTAATAACTGAC-GTTG
AGGAAGAACT-TCTGGAGT-AGTGATAAAATA-TGTTGATACC-AG---A   GGGACT-GGTAACGGCGAAGG-CATCCTTCTATGTAAAAACTGAC-GTTG
AGGAAGAACT-TCTGGAGT-AGGGATAAAATA-CGCAAATACC-AG---A   AGGACT-GATAACGGCGAAGG-CGTCCTTCTATGTAAAAACTGAC-GTTG
AGGAAGAACT-TCTGGAGT-AGGGATAAAATA-CGCAAATACC-AG---A   AGGACT-GATAACGGCGAAGG-CGTCCTTCTATGTAAAAACTGAC-GTTG
AGGAAGAAGC-TCTGGAGT-AGTGATATAATA-CGTATATACC-AG---A   GAGACT-GGTAACGGCGAAGG-CATCCTTCTATGTAAAAACTGAC-GTTG
AGAATTGAATTTTTGAAGA-AAAGATAGAATT-TGCAAATAAC-AA---A   AAGACA-GGTAGCAGTTAATA-CAACCTTCTTTATAAAAACTGAC-GTTG
GGGTAGTACT-TGAGGAGG-AGAGATGAAATT-CTATGATACC-AA---A   GGGACTCGGTAAAGGCGAAGG-CAGCCCTCTATGTAGAAACTGAC-GTTG
GGGCAGTACT-TGAGGAGG-AGAGATGAAATT-CTATTATACC-AA---A   GGGACTCGGTAAAGGCGAAGG-CAGCCCTCTATGTAAAAACTGAC-GTTG
GAGGGGTACT-CGTGGTGT-AGAGATTAAATT-CTATGATACC-AA---G   GGGACACGGTAAAGGCGAAGG-CAGCTCTTTATGTAAAAACTGAC-GTTG
AGGTCGTACC-CGCGGAGT-AGAGATAAAATT-CTTTGATACCTAATAAG   GGGACT-GGTAAAGGCGAAGG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTAGTACC-TGTGGTGT-AGAGATAAAATT-CGTTGATACCTGTTAAA   GGGACT-GGTAATGGCGTAGG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTAGAACC-TAAGGAGT-AGAGATTGAATT-TGTTTATACCTGT---A   GGGACT-GATAAGGGCGAAGG-CAACCTTTTATGTAATAACTGAC-GTTG
AGGTGAAATC-TTTACAGG-AGGATTAAAAAA-CAAAGATTCT-TT---A   GGGACG-GGTAA-GGCGAAGG-CAACCTTTTAGATAATAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGATCGTACT--ATTGGTGGAGAGATGAAATT-CTGTGATACC-TC---T   GGGACT-GGTTAAGGCGAAGG-CAATTTTCTATGTAAAAACTGAC-GTTC
AGGTAGTACC-TATGGAGT-AGAGATAAAATT-CGCTAATCCC-AATATA   GAGACT-GGTAA-GGCGAAGG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTAGTACC-TAAGGAGT-AGAGATAGAATT-CGTTTATACC-AAT--A   GGGACA-GGTAAAGGCGAAGG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAGT-AGAGATGAAATT-CTTTGATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTCG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGCGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACT-TGTGGAAT-AGAGATGAAATA-CTTTTATACCTAATA-A   GGGACT-GGTAAAGGCGAAAG-CAACCTTTTATGTAAAAACTGAC-GTTG
AGGTCGTACTTTTGAGTGG-AGAGTTTAAATT-CTTTGATACCTGG---A   -GGACG-GATAAAGGCGAAGG-CAACCTCTTATGTAAAAACTGAC-GTTG
GGGCAGTACT-TGAGGAGG-AGAGATGAAATT-TTATGATACC-AA---A   GGGACTCGGTAAAGGCGAAGG-CAGCTCTTTATGTAAAAACTGAC-GTTG
GGGCAGTACT-TGAGGAGG-AGAGATGAAATT-CTATGATCCC-AA---A   GGGACTCGGTAAAGGCGAAGG-CAGCCCTCTATGTATAAACTGAC-GTTG
GGGTAGTACT-TGAGGAGG-AGAGATGAAATT-CTATGATACC-AA---A   GGGACTCGGTAAAGGCGAAGG-CAACCCTCTATGTAAAAACTGAC-GTTG
GGGTAGTACT-TGAGGAGG-AGAGATGAAATT-CTATGATACC-AA---A   GGGACTCGGTAAAGGCGAAGG-CAACCCTTTATGTAAAAACTGAC-GTTG
GGGCAGTACT-TGAGGAGG-AGAGATGAAATT-CTATGATACC-AA---A   GGGACTCGGTAAAGGCGAAGG-CAGCTCTCTATGTAAAAACTGAC-GTTG
AGGTAGTACC-TTAAGTGT-AGAGATAAAATT-TAGTTATACT-GG---A   GGGACTTAGAAAAGGCGTAGG-CAACCTTCTAAGTAATAACTGAC-GTTG
AGGTAGTACC-TTAAGTGT-AGAGATAAAATT-TTGTTATACA-GG---A   GGGACTTAGAAAAGGCGTAGN-CAACCTTCTAAGTAATAACTGAC-GTTG
AGGCAGTACC-TTAAGTGT-AGAGATTAAATG-CTGTCATACT-AG---A   GGGACTCTGAAAAGGCGTAGG-CAGCCTTCTAAGTAATAACTGAC-GTTG
GGGCAGTACT-TGAGGAGG-AGAGATGAAATT-CTATGATACC-AA---A   GGGACTCGGTAAAGGCGAAGG-CAGCCTTCTATGTAAAAACTGAC-GTTA
GGGCAGTACT-TGAGGAGG-AGAGATGAAATT-CTATGATACC-AA---A   GGGACTCGGTAAAGGCGAAGG-CAGCCCTCTATGTAAAAACTGAC-GTTA
GGGCAGTACT-TGAGGAGG-AGAGATGAAATT-CTATGATACC-AA---A   GGGACTCGGTAAAGGCGAAGG-CAGCCCTCTATGTAAAAACTGAC-GTTG
AGGTCGAACA-TTTGGAGT-AGGTATGAAAAC-CTTAAATACC-TAT--C   TTGACG-GGTAA-GGCGAAGG-CAACCTTTTATAAAATAACTGAC-GTTG
AGATCTTACTCTAGAGTGT-AGAGTTGTAATT-CATTAATACC-TA---G   GGGACT-GGTAAAGGCGAAGG-CAATCTTTTATATAAAAACTGAC-GTTG
AGGTCGTACTCTAGAGTGG-AGAGATAAAATT-CTGTGATACC-TA---G   GGGACG-GGTAAAGGCGAAGG-CAATCTTTTATGTAAAAACTGAC-GTCG
AGGTAGTACC-TTAAGTGT-AGAGATTATATT-TGGTTATACT-GG---A   GGGACTTAGAAAAGGCGTAGG-CAACCTTCTAAGTAATAACTGAC-GTTG
AGGTAGTACC-TTAAGTGT-AGAGATTATATT-TGGTTATACT-GT---A   GGGACTTAGAAAAGGCGTAGG-CAACCTTCTAAGTAATAACTGAC-GTTG
AGGTAGTACC-TTAAGTGT-AGAGATTATATT-TGGTCATACT-CG---A   GGGACTTAGAAAAGGCGTAGGTCAACCTTCTAAGCAATAACTGACTGTTG
GGGGAGTACT-TGCGGAGG-AGAGATGAAATT-CTATCATACC-AA---A   GGGACTCGGTAAAGGCGAAGG-CACCCCTTTATGTAAAAACTGAC-GTTG
GGGAAGTACT-TGAGGAGTAAGAGATGAAATATCTGTGATACC-AA---A   GGGACTCGGTAAAGGCGAAGG-CATCCCTTTATCTAAAAACTAAC-GTTG
AGACAGTACT-TAGAGTGG-AGAGATTATAT-TCTGTGATACC-AA---A   GGGACTGGC-AACGGCGTAGG-CAGTCTTTTATGTAAAAACTGAC-GTTG
GGGAAGTACT-TAAGGAGTAAGAGATGAAATATCTGTGATACC-AA---A   GGGACTCCGTAAAGGCGAAGG-CATTCCTTTATCTAAAAACTAAC-GTTG

AGGGACGAAAGC-TAGAGTAGCAATACGGATTCGATACCCGAGTAGTTCT   AGCTGTAAACTATGAACACCATTTATTTC-A-A--------C-AAAATAA
AGGGATTAAAAC-TAGAGTAGCGAAACGGATTCGATACCCGTGTAGTTCT   AGTAGTAAACTATGAATACAATTATTTATAATATATAT-TATATAAAT-A
AGGGATTAAGAC-TAGAGTAGCGAAACGGATTCGATACCCGTGTAGTTCT   AGTAGTAAACTATGAATACAATTATTTAT-A-A------TATATAAAT-A
AGGGATTAAAAC-TAGAGTAGCGAAACGGATTCGATACCCGTGTAGTTCT   AGTAGTAAACTATGAATACAATTATTTATAATAAATATATATATAAAT-A
AGAGACGAAGGC-TTGGGTAGCAAACAGGATTAGATACCCTAATAGTCCA   AGCAGACAATGATGAATGTCATACATTAG-ATA-A------TTT-AATGT
AGAGACGAAGGC-TTGGGTAGCAAACAGGATTAGATACCCTAATAGTCCA   AGCAGACAATGATGAATGTCATACATTAG-ATA-T------TTT-AATGT
AGAGACGAAGGC-TTGGGTAGCAAACAGGATTAGATACCCTAATAGTCCA   AGCAGACAATGATGAATGTCATACATTAG-ATA-T---AAATTT-AATGT
AGAGACGAAGGC-TTGGGTAGCAAACAGGATTAGATACCCTAATAGTCCA   AGCAGACAATGATGAATGTCATACATTAG-ATA-T---AATTTT-AATGT
AGGGACGAAGGC-TTGGGTAGCAATAAGGATTAGATACCCTAGTAGTCCA   AGCAGAAAATGATGAATGTCATAGACTAG-ATA-A---AAATTT-AGTCT
```

```
AGGGACGAAGGC-TCGGGTAGCGATAAGGATTAGATACCCTAGTAGTCCA   AGCAGACAATGATGAATGTCATAGACTAG-ATTTA---ATATTT-AGTCT
AGGGACGAAGGC-TCGGGTAGCGATAAGGATTAGATACCCTAGTAGTCCA   AGCAGACAATGATGAATGTCATAGACTAG-ATTTA---ATATTT-AGTCT
AGGGACGAAGGC-TTGGGTAGCAAGAAGGATTAGATACCCTAGTAGTCCA   AGCAGACAATGATGAATGTCATAGACTAG-AAA-A----AGTCT-AGTCT
AGGTACGAAGGC-TTGGGTAACGAGCAGGATTAGATACCCTAGTAGTCCA   AGCAGAAAATTCTGAATGTCATATGCTAA-AAA-A---AAATTT-AGTTT
AGGGACGAAGGC-ACAGAGCACAAACAGGATTAGATACCCAAGTAGTCTT   TGCAGTAAATGATGAATGCCATAGGTTAG-ATT-A---AGCTTT-AGTCT
AAGGACGAAGGC-ACAGAGCACAAACAGGATTAGATACCCAAGTAGTCTT   TGCAGTAAATGATGAATGTCATAGGTTAG-ATT-A---AGCTTT-AGTCT
AGGGACGAAGGC-TTAGAGCACAAAGAGGATTAGGTACCCTAGTAGTCTT   TGCAGTAAATGCTGAACGCTTTAGGTTAG-ATT-T---AGATTT-GGTCT
AAGGACGAAGGC-TTAGAGCACAAATAGGATTAGATACCCTAGTAGTCTT   AGCAGATAATGATGAATGCCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TAAGGGCACGAACAGGATTAGACACCCTAGTAGTCTT   AGCAGACAATACTAAATGCTATGGATTAG-ATT-A----TATTT-AGTCT
AAGTACGAAGGC-TTAGGTAACGAATAGGATTAGATACCCTAGTAGTCTA   AGCAGATAACGATGAATGCCATAAGTTAG-GAT-A----TACTT-AACTT
AAGGACGAAGGC-TAAGGTAGCGAAAAGGATTAGAGACCCTTGTAGTCTT   AGCAGTAAACTATGAATGCTACTTATTAA--TA-A----TATTT-AATTC
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AGGGACGAAGGC-TCAGGACACCAAAAGGATTAGGTACCCAAGCAATCTT   TGCAGAGAATTATGAATGTCATAGGTTAG-ATTTT----TATTT-AGTCT
AAGGACGAAGGC-TAAGGACATCAACAGGATTAGATACCCTAGTAGTCTT   AGCAGTAAATTATGAATGCCATAAATTAG-ATT-A----TATTT-AATTT
AAGGACGAAGGC-TAAGGGCGCGAACAGGATTAGATACCCTAGTAGTCTT   AGCAGTAAATTATGAATGCCATAAGTTGG-ATT-A----TATTT-AATTT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGGACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAGCACGAACAGGATTAGATACCCTAGTAGTCTT   AGCAGATAATGATGAATGCCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACATGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCC
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTGGTAGTCTT   AGCAGAAAATGGTGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AAGGACGAAGGC-TTAGAACACGAATAGGATTAGATACCCTAGTAGTCTT   AGCAGAAAATGATGAATGTCATAGGTTAG-GTT-A----TACTT-AGTCT
AGGAACGAAGGCATTGCGTCACGAAAAGGATTAGGGACCCTAGTAGTCTT   TGCAGAAAATGATGAGTGCCATAGATTAG-GTT-AA--TTCT-T-AGTCT
AAGGACGAAGGC-ACAGAACACAAACAGGATTAGATACCCAAGTAGTCTT   TGCAGTAAATGATGAATGCCATAAGTTAG-ATT-A---AGCTTT-AGTTT
AAGGACGAAGGC-ACAGAGAACAAACAGGATTAGATACCCAAGTAGTCTT   TGCAGTAAATGATGAATGCCATAGGTCAG-ATT-A---AGATTT-GGTCT
AAGGACGAAGGC-ACAGAGAACAAACAGGATTAGATACCCAAGTAGTCTT   TGCAGTAAATGATGAATGCCATAGGTCAG-ATT-A---AGATTT-GGTCT
AAGGACGAAGGC-ACAGAGAACAAACAGGATTAGATACCCAAGTAGTCTT   TGCAGTAAATGATGAATGCCATAGGTCAG-ATT-A---AGATTT-GGTCT
AAGGACGAAGGC-ACAGAGAACAAACAGGATTAGATACCCAAGTAGTCTT   TGCAGTAAATGATGAATGCCATAGGTCAG-ATT-A---AGATTT-GGTCT
AAGGACGAAGGC-ATAGGTCACGAACAGGACTAGTAACCCAAGTAGTCTT   TGCAGTCAATGATGAATGCCATAGGTTAG-ATA-A---AAATTT-AGTCT
AAGGACGAAGGC-ATAGGTCACGAACAGGACTAGTAACCCAAGTAGTCTT   TGCAGTCAATGATGAATGCCATAGGTTAG-ATA-A---AAATTTAAGTCT
AAGGACGAAAGC-ATAGATCACGAACAGGATTAGATACCCAAGTAGTCTT   TGCAGTCAATGATGAATGCCATAGGTTAG-A-A-A---AAATTT-AGTCT
AAGGACGAAGGC-ACAGAGCACAAACAGGATTAGATACCCAAGTAGTCTT   TGCAGTAAATGATGAATGCCATAGGTTAG-ATT-A---AGATTT-GGTCT
AAGGACGAAGGC-ACAGAGCACAAACAGGATTAGATACCCAAGTAGTCTT   TGCAGTAAATGATGAATGCCATAGGTTAG-ATT-A---AGATTT-GGTCT
AAGGACGAAGGC-ACAGAGAACAAACAGGATTAGATACCCAAGTAGTCTT   TGCAGTAAATGATGAATGCCATAGGTTAG-ATT-A---AGATTT-GGTCT
AAGGACGAAGGC-TAAGGAAGCGAAAAGGATTAGATACCCTTGTAGTCTT   AGCAGAAAATTATGAATGCTACTTATTAA-AAT------TATTT-AATTC
AAGGACGAAGGC-TTAGGTCACAAAAAGGATTAGATACCCAAGTAGTCTT   AGCAGTCAATTATGAATGCCATAGGTTAG-ATATAGTTAAAT-T-AGTCT
AAGGACGAAGGC-AAAGGGAACAAAAAGGATTAGATACCCCAGTAGTCTT   TGCAGACAATTATGAATGCCATAGGTTAG-ATT-TT--TAATTT-AGTCT
AAGGACGAAGGC-ATAGGTCACGAACAGGATTAGATACCCAAGTAGTCTT   TGCAGTCAATGATGAATGCCATAGGTTAG-ATA-A---AAATTT-AGTCT
AAGGACGAAGGC-ATAGGTCACGAACAGGATTAGATACCCTAGTAGTCTT   TGCAGTCAATGATGAATGCCATAGGTTAG-ATA-A----AATTT-AGTCT
AAGGACGAAGGCTATAGGTCACGAACAGGATTAGATACCCAAGTAGTCTT   TGCAGTCAATGATGAATGCCATAGGTTAG-ATA-A---AAATTT-AGTCT
AGGGACGAAGGC-ACAGAGAACAAACAGGATTAGATACCCTAGTAGTCTT   TGCAGTAAATGATGAATGCCATAGATTAG-ATAAA---CAATTT-GGTCT
AAGGACGAAGGC-TTAGATAACAAATAGGATTAGAGACCCTAGTAGTCTA   AGCAGTAAACGATGAATGCTATAGGTTAG-ATT-A---ACATTT-GGTCT
AAGGACGAAGGC-ATAGATCGCAAACAGGATTTGACACCCCAGTAGTCTT   TGCAGACAATGATGAATGCCGTAGATTAG-ATG-A---GAATTT-AGTCT
AAGGACGAAGGC-TTAGATAACAAATAGGATTAGAGACCCTAGTAGTCTA   AGCAGTAAACGATGAATGCTATAGGTTAG-ATT-A---ATATTT-GGTCT

ATAAATGAAAATGAAAGTGTTCCGCCTGAAGAGTACGTTCGCAA-GAATG   AAACTCAAGACAATAGACGGTTACAGACTTAAGCAGTGGAGCATGTTATT
ATAAATGAAAATGAAAGTATTCCACCTGAAGAGTACGTTAGCAA-TAATG   AAACTCAAAACAATAGACGGTTACAGACTTAAGCAGTGGAGCATGTTATT
ATAAATGAAAATGAAAGTATTCCACCTGAAGAGTACGTTAGCAA-TAATG   AAACTCAAAACAATAGACGGTTACAGACTTAAGCAGTGGAGCATGTTATT
ATAAATGAAAATGAAAGTATTCCACCTGAAGAGTACGTTAGCAA-TAATG   AAACTCAAAACAATAGACGGTTACAGACTTAAGCAGTGGAGCATGTTATT
ATAAACGAAAGTGTAAGCATTCCACCTCAAGAGTACTATGGCAA-CATAT   AAACTGAAATCATTAGACCGTTTCTGAAACCAGTAGTGAAGTATGTTATT
ATAAACGAAAGTGTAAGCATTCCACCTCAAGAGTACTATGGCAA-CATAT   AAACTGAAATCATTAGACCGTTTCTGAAACCAGTAGTGAAGTATGTTATT
ATAAACGAAAGTGTAAGCATTCCACCTCAAGAGTACTATGGCAA-CATAT   AAACTGAAATCATTAGACCGTTTCTAAAACCAGTAGTGAAGTATGTTATT
ATAAACGAAAGTGTAAGCATTCCACCTCAAGAGTACTATGGCAA-CATAT   AAACTGAAATCATTAGACCGTTTCTAAAACCAGTAGTGAAGTATGTTATT
ATAAATGAAAGTGTAAGCATTCCACCTCAAGAGTACTATGGCAA-CATAT   AAACTGAAATCATTAGACCGTTTCTGAAACCAGTAGTGAAGTATGTTATT
ATAAATGAAAGTGTAAGCATTCCACCTCAAGAGTAATATGGCAA-CATAT   AAACTGAAATCATTAGACCGTTTCTGAAACCAGTAGTGAAGTATGTTATT
ATAAATGAAAGTGTAAGCATTCCACCTCAAGAGTAATATGGCAA-CATAT   AAACTGAAATCATTAGACCGTTTCTGAAACCAGTAGTGAAGTATGTTATT
ATAAATGAAAGTGTAAGCATTCCACCTCAAGAGTAATATGGCAA-CATAT   AAACTGAAATCATTAGACCGTTTCTGAGACCAGTAGTGAAGTATGTTATT
ATAAATAAAAGTGTAAGCATTCCACCTCAAGAGTACAATGGCAA-CATTA   AAACTGAAATCATTAGGCCGTCTTTGAAACCAGTAGTGAAGTATGTTATT
ATAAATGAAAGTGTAAGCATTCCACCTCAAGAGTAATGTGGCAA-CGCAG   GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAGTGTAAGCATTCCACCTCAAGAGTAATGTGGCAA-CGCAG   GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAAT-TAAGCGTTTCACCTCAAGAGTAATGTGGCAA-CGCTG   AAACTGAAATCACTAGACCGTTTCTAACACCAGTAGTGAAGTATGTTGTT
ATAAACGAAAGTGTAAGCATTTCACCTCAAGAGTAACATGGCAA-CATGG   GAACTGAAATCACTAGGCCGTTTCTGACACCAGTAGTGAAGTATGTTATT
ATAAACGAAAGTGAAAGCATTTCACCTCAAGAGTAATGTGGCAA-CACAG   GAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
CTAAATGAAAATATAAGCATTCCACCTCAAGAGTAACGTGGCAA-CACGG   AAACTGTAATCACTAGACCGTTTCTGACAACAGTAGTGAAGTATGTTATT
GTAAATGAAAATGTAAGCATTTCACCTCATGAGTAATGTGGCAA-CGCAG   GAACTGAAATCACTAGACCGTTTTTGACACCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG   AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG   AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAGTGTAAGCATTCCACCTCAAGAGTAAGGTGGGAAACATCG   GAAATGAAATCAGTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-CACAG   GAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
```

```
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGGCCGTTTCTGACATCAGTAATGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAACGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    GAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAACGTGGCAA-CACGG    AAACTGAAATCACTAGACCGTTTCTGACATCAGTAGTGAAGTATGTTATT
ATAAATGAAAGTGTAAGCATTCCACCTCAAGAGTAATGCGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGCAGTGAAGTATGTTGTT
ATAAATGAAAGTGTAAGCATTCCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAATGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAAAGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
GTAAATGAAAATTTAAGCATTTCACCTCACGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTTAGACACCAGTAGTGAAGTATGTTATT
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGCAGTGAAGTATGTTATT
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAAGGCGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGCAATGAAGTATGTTATT
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAGTGTAAGCATTTCACCTCAAGAGTAATGTGGCAA-TGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTGTT
ATAAATGAAAGTGTAAGCATTCCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACAAAAGTAGTGAAGTATGTTGTT
ATAAATGAAAATGTAAGCATTCCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTAACACCAGTAGTGAAGTATGTTGTT
ATCAATGAAAATAAAAGCATTTCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTGACACCAGTAGTGAAGTATGTTATT
ATAAATGAAAATGTAAGCATTCCACCTCAAGAGTAATGTGGCAA-CGCAG    GAACTGAAATCACTAGACCGTTTCTAACACCAGTAGTGAAGTATGTTGTT

TAATTCGATAATCCCCGACTAACCTTACCATATTTTGAA----------T    ATTCAGGCGTTACATCGTTGTCTTCAGTTCGTGCCGCAAGGTCTTAGATT
TAATTCGATAATCCACGACTAACCTTACCATATTTTGAA----------T    ATTCAGGCGTTACATTGTTGTCTTTAGTTCGTGCTGCAAAGTTTTAGATT
TAATTCGATAATCCACGACTAACCTTACCATATTTTGAA----------T    ATTCAGGCGTTACATTGTTGTCTTTAGTTCGTGCTGCAAAGTTTTAGATT
TAATTCGATAATCCACGACTAACCTTACCATATTTTGAA----------T    ATTCAGGCGTTACATTGTTGTCTTTAGTTCGTGCTGCAAAGTTTTAGATT
TAATTCGATGATCCGCGAAAAACCTTACCACAGCTTGTA----------T    AGCCAAGCGCTGCATGGCTGTCTTTAGTTAATGTCGTGANATC-TGG-TT
TAATTCGATGATCCGCGAAAAACCTTACCACAGCTTGTA----------T    AGCCAAGCGCTGCATGGCTGTCTTTAGTTAATGTCGTGAGATC-TGG-TT
TAATTCGATGATCCGCGAAAAACCTTACCACAGCTTGTA----------T    AGCCAAGCGCTGCATGGCTGTCTTTAGTTAATGTCGTGAGATT-TGG-TT
TAATTCGATGATCCGCGAAAAACCTTACCACAGCTTGTA----------T    AGCCAAG-GCTGCATGGCTGTCTTTAGTTAATGTCGTGAGATT-TGG-TT
TAATTCGATAATCCGCGAATAACCTTACCACAATTTGAA----------T    A-TCAAGCGCTGCACGGCTGTCTTTAGTTAATGTCGTGAGATC-TGG-TT
TAATTCGATAACCCGCGAAAAACCTTACCACAGTTTGAA----------T    AACCAAGCGCTGCACGGCTGTCTTTAGTTAATGTCGTGAGATC-TGG-TT
TAATTCGATAACCCGCGAAAAACCTTACCACAGTTTGAA----------T    AACCAAGCGCTGCACGGCTGTCTTTAGTTAATGTCGTGAGATC-TGG-TT
TAATTCGATGATCCGCGAAAAACCTTACCACAGTTTGAA----------T    -ATCAAGCGCTGCACGGCTGTTTTTAGTTAATGTCGTGAGATC-TGG-TT
TAATTCGATAACCCGCGATTAATCTTACCACAATTTGAA----------T    AATCATGCGCTGCACGGCTGTCTTTAGTTAATGTCGTGAGATT-AGG-TT
TAATTCGTAGATACACGAAAAACCTTACCACAACTTGAA----------T    AATCAGGAGTTGCACGGCTGTTTTCAGTTAATGTTGTGAAACTATGGTTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTTGAA----------T    -ATCAGGAGTTGCACGGCTGTTTTCAGTTAATGTTGTGAAACTGTGGTTT
TAATTCGACAATGCACGAAAAACCTTACCACATTTTGTA----------T    AATCGGGAGTTGCACGGCTGTTTTCAGTTAATGTTGTGAAATTATGGTTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGTA----------T    TGTCGAGTGTTGCACGGCTGTTTTCAGTTAATGTTGTGAAACTGTGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTTGAA----------T    TATCGAGTGTTGCACGGCTGTTTGCAGTTAATGTTGTGAAACTGTGGCTT
TAATTCGATGATCCTCGAAAAACCTTACCACAATTGGAA----------T    TTTCGAGTGTTGCACGGCTGTTTTCGGTTAATGTTGTGAAACTGTGGCTA
TAATTCGATGATCCGCGAAAAACCTTACCACAATTTGAA----------T    TATCTAGTGTTGCACGGCTGTCTTCAGTTAATGTTGTGAAACTGTGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TATCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TATCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
TAATTCGGTGATCCAGGAAAAACCTTACCACAATTTGTA-----------    ---CAGGTGTTGCACGGCTGTTTTCAGTTAATGTTGTGAAATTATGGTTA
TAATTCGATAATCCACGAAAAACCTTACCACAATTAGAA----------T    TACCAAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTGTGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TATCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTGTGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TGTCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TGTCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TGTCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TGTCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TGTCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TGTCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TGTCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TGTCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TGTCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
TAATTCGATGACCCACGAAAAACCTTACCACAATTGGAA----------T    TGTCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTGTGGTTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TGTCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TGTCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TGTCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
TAATTCGATGATCCACGAAAAACCTTACCACAATTAGAA----------T    TGTCGAGTGTTGCACGGCTGTTTCCAGTTAATGTTGTGAAACTATGGCTT
```



```
C------CATGAAATTAACGCAATCCCTGGCTTTATTT--TTTTGA----   ATT-TTACGATAAAGC-ATT-TG-ATC-CTGGAATT---TGG--AAAGAA
A--AG-CCATGAAATTGACCAAAACCTTTACTTTATTT--A-TTAA---T   ATA-TTTTTATAAAGT-ATT-TC-ATC-TTGACAAT---AGA--TGTTAC
C--GGTCCATGGAATTAACGAAAACCCTTGCTTTATTT--G-TAAA----   -TT-CTAAAATAAAGCAGTT-C--ACC-TGTATATGG--GTT--ATGATA
C--GTTCCATGGAATTAACGTAAACCCTTGCTTTATTT--G-TAA-----   ----ATATTATAAAGCAGTT-C--ACC-TTTATATAG--GAA--ATGATA
C------CATG-AATTAACACAATCCCTTGCTTTATTTTTG-TAAGG-CT   AGTGTTACTATAAAGC-ATT-TG-ACT-CAAGAATA---TGG--GAAGAA
C------CATG-AATTAACACAATCCCTTGCTTTATTTTTG-TATTA--T   AAGCTTACTATAAAGC-ATT-TA-ACT-CAAGAATA---TGG--GAAGAA
C------CATG-AATTAACTCAATCCCTTGCTTTATTTTTG-TAAA----   ----TTACTATAAAGC-ATT-TG-ACC-CAAGAATA---TGG--GAAGAA
C------CATGAAATTAACGAAATCCTTGGCTTTTTTTTTA-CCCTT-GT   GTG-CTACAATAAAGC-ATT-AGATCC-CCAATATG---GGG--ATAGAA
C------CATGAAATTAACGTAATCCCTGGCTTTATTT-TA-TAA----T   TTT-TTTCGATAGAGCAGATTTT-TCCTTGATATGG---GGG--AAAAAA
T-----TCCATGAATTAACAGAATCCTTTGCTTTATTT--A-TTAT---T   ----CT-AAATACAGC-ACT-CT-TT-TTAA-AGAA---TGA--AAAGAT
C------CATGAAATTAACGAAATCCCTGGCTTTATTT-TA-TAAT---T   TT--TTTCGATAGAGCAGATTTT-TTCTTGATATGG---GGA--AAAAAA

AA-AGGAATTAAGACAAATCA-TGATGATCCTT-ATAATATGGGTAATAG   ACGTGCTA-TAATAAAATGATAACAAAGTTATATAA-------TTAATT-
AA-AGGAATTAAGACAAATCA-TAATGATCCTT-ATAATATGGGTAATAG   ACGTGCTA-TAATAAAATGATAATAAAATTATATAAAATATATTTAATT-
AA-AGGAATTAAGACAAATCA-TAATGATCCTT-ATAATATGGGTAATAG   ACGTGCTA-TAATAAAATAAGAATAAAATTATAAAAAATATA--TAAAT-
AA-AGGAATTAAGACAAATCA-TAATGATCCTT-ATAATATGGGTAATAG   ACGTGCTA-TAATAAAATGATAATAAAATTATATAAAATATATAAAAAA-
AATAGGGATTAAGACAAGTCA-TTATGG-CCTAAATGCTGTGGGCTATAG   ACGTGCCA-CATACGCCTT--TACAAAGGGATGCTATATTGTGAAATGG-
AATAGGGATTAAGACAAGTCA-TTATGG-CCTTAATGCTGTGGGCTATAG   ACGTGCCA-CATACGCCTT--TACAAAGGGATGCTATATTGTGAAATGG-
AAAAGGGATTAAGACAAGTCA-TTATGG-CCTTAATGCTGTGGGCTATAG   ACGTGCCA-CATACGCCTT--TACAACGGGATGCGATATTGTGAAATGG-
AATAGGGATTAAGACAAGTCA-TTATGG-CCTTAATGCTGTGGGCTATAG   ACGTGCCA-CATACGCCTT--TACAACGGGATGCGATATTGTGAAATGG-
AAAAGGGATTAAGACAAGTCA-TCATGG-CCTAAATATTGTGGGCTATAG   ACGTGCCA-CATATGCCTC--TACAAAGGGATGCTATATTGCGAAATGG-
AATAGGGATTAAGACAAGTCA-TCATGG-CCTTAATACTGTGGGCTATAG   ACGTGCCA-CATACGCCTT--AACAAAGGGATGCGATATTGTGAAATGG-
AATAGGGATTAAGACAAGTCA-TCATGG-CCTTAATACTGTGGGCTATAG   ACGTGCCA-CATACGCCTT--AACAAAGGGATGCGATATTGTGAAATGG-
AATAGGGATTAAGACAAGTCA-TCATGG-CCTTAATACTGTGGGCTATAG   ACGTGCCA-CATATGCCTT--TACAAAGGGATGCGATTTTGTGAAATTG-
ATATAGGATCAAGACAAGTCA-TCATGG-CCAAAATATTGTGGGCTATAG   ACGTGCCA-CATTTATCAT--TACAAAAGGATGTTATATTGTGAAAATA-
AAAGGGGACAAAGACAAGTCA-TCATGG-CCTTGATGTTGTGGGCTATAG   ACGTGCCA-CATATACCTA--GACAAAGAGATGCAAAAATGTGAATTTG-
AAAGGGGACAAAGACAAGTCA-TCATGG-CCTTGATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--CACAAAGAGATGCAAAAATGTGAATTTA-
AA-AGGGACAAAGACAAGTCC-TCATGA-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATTTGCCTA--TACAAAGAGATGCAAAAATGTGAATTTA-
AATAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGTGAGTTTA-
AATAGGGCCTAAGACAAGTCA-TCATGG-CCTTTATGTTGTGGGCTATAG   ACGTGCCA-CATATACCTA--GACAAAGAGATGCAAGAATGCGAATTTA-
AATAGGGCCAAAGACAAGTCA-TCATGA-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGTGAATTTG-
AACAGGGACAAAGACAAGTCA-TCATGG-CCTTTATGTTGTGGGCTATAG   ACGTGCCA-CATTTACCTA--GACAAAGAGATGCAAAAATGAAAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AATAGGGACTAAGACAAGTCA-TCATGG-CTTTTACATTGTGGGCTATAG   ACGTGCCA-CATACTCCTA-TAACAAAGAGATGCAAAAATGCGAATTTA-
AACGGGGCCTAAGACAAGTCA-TCATGG-CCTTTGTATTGTGGGCTATAG   ACGTGCTG-CATATGCCTA--GACAAAGAGATGCAAAAATGAAAATTTA-
AATAGGGCCTAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATACCTA--TACAAAGAGATGCAAAAATGAGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCT-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AATAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGAGGCGAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCG---------------
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCCAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCG---------------
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AGTAGGGCCAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--GACAAAGAGATGCGAAAATGCGAATTTA-
AACAGGGACTAAGACAAGTCA-TCATGG-CC-TAATATTGTGGGCTAAAG   ACGTGCCA-CATATTCCTT--AACAAAGAGATGCGAAATAGCAATATTG-
AAAGGGGACAAAGACAAGTCA-TCATGG-CCTTGATGTTATGGGCTATAG   ACGTGCCA-CATATACCTA--AACAAAGAGATGCAAAAATGTGAATTTG-
AAAGGGGACAAAGACAAGTCA-TCATGG-CCTTAATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--AACAAAGAGATGCAAAAATGTGAATTTA-
AA-GGGGACAAAGACAAGTCA-TCATGG-CCTTGATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--AACAAAGAGATGCAAAAATGTGAATTTA-
AA-GGGGACAAAGACAAGTCA-TCATGG-CCTTGATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--AACAAAGAGATGCAAAAATGTGAATTTA-
AAAGGGGACAAAGACAAGTCA-TCATGG-CCTTGATATTGTGGGCTATAG   ACGTGCCA-CATATTCCTA--AACAAAGAGATGCAAAAATGTGAATTTA-
AAAAGGGAGTAAGACAAGTCA-TCATGG-CCTTAATATTGTGGGCTATAG   ACGTGCCA-CATAAGCCTA--TACAAAGAGATGCGAAAATGTGAATTCT-
AAAAGGGAGTAAGACAAGTCA-TCATGG-CCTTAATACTGTGGGCTATAG   ACGTGCCA-CATAAGCCTA--TACAAAGAGATGCGAAAATGTGAATTCT-
AAAAGGGAGTAAGACAAGTCA-TCATGG-CCTTAATATTGTGGGCTATAG   ACGTGCCA-CATAAGCCTA--AACAAAGAGATGCAAAAATGTGAATTCT-
AAAGGGGACAAAGACAAGTCA-TCATGG-CCTTGATATTGTGGGCTATAG   ACGTGCCA-CATATGCCTG--GACAAAGAGATGCAAAAATGTGAATTTA-
AAAGGGGACAAAGACAAGTCA-TCATGG-CCTTAATATTGTGGGCTATAG   ACGTGCCA-CATATGCCTG--GACAAAGAGATGCAAAAATGTGAATTTA-
AAAGGGGACAAAGACAAGTCA-TCATGG-CCTTTATATTGTGGGCTATAG   ACGTGCCA-CATATGCCTA--GACAAAGAGATGCGAAAATGTGAATTTA-
GTTAGGGACCAAGATAAGTCA-TCATGA-CCTCTATATTGTGGGCTATAG   ACGTGCCA-CATTTTCCTA--GACAAAAAGATGCAAAAATGCAAATTTA-
AA-AGGGATCAAGACAAGTCC-TCATGG-CCTTGATGTTGTGGGCTATAG   ACGTGCCA-CATGTACCTA--AACAAAGAGATGCAAAAATGTGAATTTT-
AA-AGGGATC-AGACAAGTCA-TCATGG-CCTAAATATTGTGGGCTATAG   ACGCGCCA-CATTTTCCTA--AACAAAGAGATGCAAAAATGTGAATTTT-
AAAAGGGAGTAAGACAAGTCA-TCATGG-CCTTAATATTGTGGGCTATAG   ACGTGCCA-CATAAGCCTA--TACAAAGAGATGCGAAAATGTGAATTCT-
AAAAGGGAGTAAGACAAGTCA-TCATGG-CCTTAATATTGTGGGCTATAG   ACGTGCCA-CATAAGCCTA--TACAAAGAGATGCAAAAATGTGAATTCT-
AAAAGGGAGTAAGACAAGTCAATCATGG-CCTTAATATTGTGGGCTATAG   ACGTGCCAACATAAGCCTA--TACAAAGAGATGCAAAAATGTGAATTGTT
AAAAGGGACAAAGACAAGTCA-TCATGG-CCTTAATATTGTGGGCTATAG   ACGTGCCA-CATCTTCCTA--GACAAAGAGATGCAAAAATGTGAATTTA-
AAAGGGGACAAAGACAAGTCA-TCATGG-CCCTAATATTGTGGGCGATAG   ACGTGCCA-CATATGCCTA--GACAAAGAGATGCAAAAATGTGAATTGA-
AAAAAGGAATAAGACAAGTCA-TCATGG-CCCTTATATTGTGGGCTATAG   ACGTGCCA-CATGTTCCTT--AACAAAGAGAAGCAAAAATGTGAATTTA-
AAAGGGGACAAAGACAAGTCA-TCATGG-CCCTAATATTGTGGGCGATAG   ACGTGCCA-CATATGCCTA--CACAAAGAGATGCAAAAATGTGAATTGA-

ATAATATAAATTAATAACATAAAACATTTTAATCTTTAATATATTTATAT   ATTATATA-TAGTATGG-ATTATAATCTGAAATTCG-ATTATATGAAAAA
--ATATTTAATTAATAATATAAAACATTTTAATTTTTAATATATTTTTTT   ATTATATATTAATATGA-ATTATAATCTGAAATTCG-ATTATATGAAAAA
ATATATTTAATTAATAATCTAAAATATTTTAACTTTTAATATATTTTTAT   ATTATATATTAATATGA-ATTATAATCTGAAATTCG-ATTATATGAAAAA
```

```
ATATATTTAATTAATAATATAAAACATTTTAACTTTTAATATATTTTTAT   ATTATATATTAATATGA-ATTATAATCTGAAATTCG-ATTATATGAAAAA
AGCTAATCCCCCAAA-AAGGAAATAT------------------------   ------------TATGG-ATTGTAGTCTGTAACTCC-ACTACATGAAAAA
AGCTAATCCCCAAAA-AAGGAAATAT------------------------   ------------TATGG-ATTGTAGTCTGTAACTCG-ACTACATGAAAAA
AGCTAATCCCCAAAA-AAGGAAATAA------------------------   ------------TATGG-ATTGTAGTCTGTAACTCG-ACTACATGAAAAA
AGCTAATCCCCAAAA-AAGGAAATAA------------------------   ------------TATGG-ATTGTAGTCTGTAACTCG-ACTACATGAAAAA
AGCTAATCCCCAAAA-TTGGATATAA------------------------   ------------TATGG-ATTGTTGTCTGTAATTCG-ACAGCATGAATAA
AGCTAATCCCCAAAA-TAGGATATAA------------------------   ------------TATGG-ATTGTAGTCTGTAACTCG-ACTACATGAATAA
AGCTAATCCCCAAAA-AAGGATATAA------------------------   ------------TATGG-ATTGTAGTCTGTAACTCG-ACTACATGAATAA
AGCTAATCCCCCAAAAAAGGATATAA------------------------   ------------TATGG-ATTGTAGTCTGTAACTCG-ACTACATGAATAA
AGCTAATCCTTAAAT-TTGA--TTAA------------------------   ------------TATGG-ATTGTAGTCTGTAACTCG-ACTACATGAATAA
AGCTAATCTCTAAAG-TAGGATATAA------------------------   --------ATTTAAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCTAAAAATAGGATATAA------------------------   --------ATTTTAAGG-ATTATAGTCTGAAACTCG-ACTATATGAATAA
AGCTAATCTAAAAAAATAGGATATAA------------------------   --------ATTCTAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATATGG-ATTGTAGTCTGAAATTCG-ACTACATGAATAA
AGCTAATCTCAAAAA-TAGGATAAAA------------------------   -------AAGGATAAGG-ATTGTAGTC-----------------------
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATATGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCCAAAA-TAGGATAAAA------------------------   --------AGGATATGG-ATTGTAGTCTGAAACTCG-CCTATATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCATAAAAAAGGATAAAA------------------------   --------AAAA------AAAAAAATCTAGA--------------------
AGCTAATCTCAAAAA-TAGGATAAAA------------------------   -------AAGGATATTG-ATTGTAGTCTGTAACTCG-ACTATATGAATAA
AGCTAATCTCGAAAA-TAGGATAAAA------------------------   --------AAGATAAGG-ATTGTAGTCTGTAACTCG-ACTATATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTTG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATA-GG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
--------------------------------------------------   --------------------------------------------------
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
--------------------------------------------------   --------------------------------------------------
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------TAGGATAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAAGAGGATAAAA------------------------   --------AATATATGG-ATTGTAGCCTGAAACTCG-GCTGCATGAATAA
AGCTAATCTCTAAAA-TAGGTTATAA------------------------   --------ATTTAAAGG-ATTGTAGTCTGAAACTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATATAA------------------------   --------GTTTTAAGG-ATTGTAGTCTGAAACTCG-ACTATATGAATAA
AGCTAATCTCAAAAAATAGGATATAA------------------------   --------GTTTTAAGG-ATTGTAGTCTGAAACTCG-ACTATATGAATAA
AGCTAATCTCAAAAAATAGGATATAA------------------------   --------GTTTTAAGG-ATTGTAGTCTGAAACTCG-ACTATATGAATAA
AGCTAATCTCAAAAAATAGGGTATAA------------------------   --------GTTTTAAGG-ATTGTAGTCTGAAACTCG-ACTATATGAATAA
AGCAAATCTCCAAAA-TAGGAAATAA------------------------   ------------AAAGG-ATTGTAGTCTGAAATTCG-ACTATATGAATAA
AGCAAATCTC-AAAA-TAGGAAATAA------------------------   ------------AAAGGAATTCTAGTCTGTTTAACG-ACTATATGAATAA
AGCAAATCTCCAAAA-TAGGAAATAA------------------------   ------------AAAGG-ATTGTAGTCTGAAATTCG-ACTATATGAATAC
AGCTAATCTCAAAAAACAGGATATAA------------------------   --------ATTTTAAGG-ATTGTAGTCTGAAATTCG-ACTATATGAATAA
AGCTAATCTCAAAAAACAGGATATAA------------------------   --------ATTTTAAGG-ATTGTAGTCTGAAATTCG-ACTATATGAATAA
AGCTAATCTCAAAAAATAGGATATAA------------------------   --------ATTTTAAGG-ATTGTAGTCTGAAATTCG-ACTATATGAATAA
AGCTAATCTAAAAAA-TAGGATAAAA------------------------   -------AAGGATATGG-ATTATAGTCTGAAATTCG-ACTATATGAATAA
AGCTAATCTCCCAAAATAGGATAAAA------------------------   -------TTAGATATGG-ATTGTAGTCTGAAACTCG-GCTATATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   -------ATATACAAGG-ATTGTAGTCTGAAATTCGGACTGCATGAATAA
AGCAAATCTCCAAAA-TAGGAAATAA------------------------   ------------AAAGG-ATTGTAGTCTGAAATTCG-ACTATATGAATAA
AGCAAATCTCCAAAA-TAGGAAATAA------------------------   ------------AAAGG-ATTGTAGTCTGAAATTCG-ACTATATGAATAA
AGCAAATCTCCAAAA-TAGGAAATAA------------------------   ------------AAAGG-ATTGTAGTCTGAAATTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATAAAA------------------------   --------ATTTAAAGG-ATTGTAGTCTGAAATTCG-ACTACATGAATAA
AGCTAATCTCAAAAAATAGGATATAA------------------------   ----------T--AAGG-ATTGTAGTCTGAAATTCG-ACTACATGAATAA
AGCTAATCTCAAAAAAAAGGATATAA------------------------   ------ATTATACAAGG-ATTGTAGTCTGAAACTCG-ACTATATGAATAA
AGCTAATCTCAAAAAATAGGATATAA------------------------   ------------GAAGG-ATTGTAGTCTGAAATTCG-ACTACATGAATAA

AGAATTGCTAGTAATACGTAAATTAGTAT-GTTACGGTGAAT--ACTCTA   AC-TGTTTCGCACTAATCACTCATCACGCGTTGAAACATGAT
AGAATTGCTAGTAATACGTAAATTAGTAT-GTTACGGTGAAT--ATTCTA   AC-TGTTTCGCACTAATCACTCATCACGCGTTGAAACATATT
AGAATTGCTAGTAATACGTAAATTAGTAT-GTTACGGTGAAT--ATTCTA   AC-TGTTTCGCACTAATCACTCATCACGCGTTGAAACATCTT
AGAATTGCTAGTAATACGTAAATTAGTAT-GTTACGGTGAAT--ATTCTA   AC-TGTTTCGCACTAATCACTCATCACGCGTTGAAACATATT
GGAATTACTAGTAAT-CGTGAATCACCATCGTCACGGTGAAT--TTTTTC   TC-AGTTAGGTACTAACTACTCGTCAGGCGCTGAAAGAAAAA
GGAATTACTAGTAAT-CGTGAATCACCATCGTCACGGTGAAG--TTTTTC   TTCAGTTAGGTACTAACTACTCGTCAGGCGCTGAAAGAAGAA
GGAATTACTAGTAAT-CGTGAATCACCAACGTCACGGTGAAT--TTTATC   TT-AGTTAGGTACTAACCACTCGTCAGGCGCTGAAAAAAGAA
GGAATTACTAGTAAT-CGTGAATCACCAACGTCACGGTGAAT--TTTATC   TT-AGTTAGGTACTAACCACTCGTCAGGCGCTGAAAAAAGAA
GGAATTACTAGTAAT-CGTGAATCACCATCGTCACGGTGAAT--CTAATC   TC-AGCTAGGTACTAACCACTCGTCAGGCGCTGAAAGGAGTA
GGAATTACTAGTAAT-CGTGTATCACCAACGGCCCGGTGGAT--TTATTC   TC-AGCTCGGGCCTAGCCACTCGTCACGCGCTGCAAGAAGTA
GGAATTACTAGTAAT-CGTGTATCACCAACGGCACGGTGAAT--TTATTC   TC-AGCTAGGTACTAACCCCTCGTCAGGCGCTGAAAGAAGTA
GGAATTACTAGTAAT-CGTGAATCACCATCGTCACGGTGAAT--TAAAAA   TC-AGCTGGGTACTAACCACTCGTCAGGCGCTGAAAGAAGTA
GGAATTACTAGTAAT-CGTTAATCATCA-CGTAACGGTGAAT--CGTTTC   TC-AATTAGGTACTAACCACTCGTCAGGCGCTGAAAGAAGAA
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--T-AACC   TT-GGATTGGTACTAACCACTCGTCACATGCTGAAAAGAGCA
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--T-TACC   TT-GGATTGGTACTAACCACTCGTCACATGCTGAAAAGAGCA
GTAATTACTAGTAAT-CGTGAATTAGCAT-GTCACGGTGAAT--A-ATTC   TT-GGATTGGTACTAACAACTCGTCGCATGCTGAAAAGAGAA
GTAATTACTAGTAAT-C---------------------------------   ------------------------------------------
--------------------------------------------------   ------------------------------------------
```

```
GTAATTACTAGTAAT-C---------------------------------    ------------------------------------------
GTAATTACTAG---------------------------------------    ------------------------------------------
GTAAT---------------------------------------------    ------------------------------------------
GTAATTACTAGTAAT-CGTGAATC--------------------------    ------------------------------------------
--------------------------------------------------    ------------------------------------------
GTAATTACTAGTAAT-C---------------------------------    ------------------------------------------
GTAATTACTAGTAA------------------------------------    ------------------------------------------
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--AATATC    TC-GGATTGGTACTAACCACTCG-------------------
GTAATTACTAGTAAT-CG--------------------------------    ------------------------------------------
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--AATATC    TC-GGATTGGTACTAACCACTCG-------------------
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--AATATC    TC-GGATTGGTACTAACCACTCG-------------------
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--A-----    ------------------------------------------
GTAATTACTAGTAAT-CG--------------------------------    ------------------------------------------
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--AATATC    TC-GGATTGGTACTAACCACTCG-------------------
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--AATATC    TC-GGATTGGTACTAACCACTCG-------------------
GTAATTACTAGTAAT-CGTGAATCACCAT-GTC-----------------    ------------------------------------------
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--ATTACC    TC-GGATTGGTACTAACCACTCG-------------------
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--AATATC    TC-GGATTGGTACTAACCACTCG-------------------
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--AATATC    TC-GGA------------------------------------
--------------------------------------------------    ------------------------------------------
GTAA----------------------------------------------    ------------------------------------------
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--AATATC    TC-GGATT----------------------------------
--------------------------------------------------    ------------------------------------------
GTAATTACTAGTAA------------------------------------    ------------------------------------------
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--AATATC    TC-GGATTGGTACTAACCACTCG-------------------
GAAATTGCTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--AAAACC    TT-GGATTGGTACTAACCACTCGTCACATGCTGAAAGGAGTA
GTAATTACTAGTAAT-CGTGAATCACCAC-GTCACGGTGAAT--T-AACC    TT-GGATTGGTACTAACCACTCGTCACATGCTGAAAAGAGCA
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--T-AACC    TT-GGATTGGTACTAACCACTCGTCACATGCTGAAAAGAGCA
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--T-AACC    TT-GGATTGGTACTAACCACTCGTCACATGCTGAAAAGAGCA
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--T-AACC    TT-GGATTGGTACTAACCACTCGTCACATGCTGAAAAGAGCA
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--T-AACC    TT-GGATTGGTACTAACCACTCGTCACATGCTGAAAAGAGCA
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--A-TTTA    TC-GGATTGGTACTAACCACTCGTCGCATGCTGAAAGGAGTT
GTAATTACTAGTAAT-CGTGAATAAGGAT-GTCACGGAGAAT--A-TTTA    TG-GGATTGGTACTAACCACTCGTCG-ATGCTGAAAGGAGTT
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--A-TATC    TC-GGATTGGTACTAATCACTCGTCGCATGCTGAAAGGAGTT
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--T-AACC    TT-GGATTGGTACTAACCACTCGTCACATGCTGAAAAGAGCA
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--T-AACC    TT-GGATTGGTACTAACCACTCGTCACATGCTGAAAAGAGCA
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--T-AACC    TT-GGATTGGTACTAACCACTCGTCACATGCTGAAAAGAGCA
GTAATTACTAGTAAT-CGTGAATCACCATCGTCACGGTGAATTTTCTATC    TC---ATTGGTACTAACTACTCGTCGCATGCTGAAAAGTGTA
GAAATTGCTAGTAAT-CGTGGATCACCAT-GACACGGTGAAT--ATAACT    TC-GGATTGGTACTAACCACTCGTCACATGCTGAAAGGAGTA
GAAATTGCTAGTAAT-CGTGAATCACCAT-GACACGGTGAAT--ATTCCC    TC-GGATTGGTACTAACCACTCGTCACATGCTGAAAGGAGTG
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--G-TTTC    TC-GGATTGGTACTAACCACTCGTCGCATGCTGAAAGGAGTT
GTAATTACTAGTAAT-CGTGAATCAGTAT-GTCACGGTGAAT--A-TATC    TC-GGATTGGTACTAACCACTCGTCGCATGCTGAAAGGAGTT
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--G-TTTC    TC-GGATTGGTACTAACCACTCGTCGCATGCTGAAAGGAGTT
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--A-AATC    TC-GGATTGGTACTAACCACTCGTCACATGCTGAAAAGAGCT
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--T-AACC    TT-GGATTGGTACTAACCACTCGTCACATGCTGAAAGGAGCA
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--T--TTC    TTCGGATTGGTACTAATTACTCGTCGCATGCTGAAAGGAGCG
GTAATTACTAGTAAT-CGTGAATCACCAT-GTCACGGTGAAT--T-AACC    TT-GGATTGGTACTAACCACTCGTCACATGCTGAAAGGAGCA
```

Supplementary Figure 7.7. Alignment of the DNA sequence of the mt *rns* group II intron encoding a putative LHEG.

A. aegerita (A.a.rnli5)	TTGCGACTAGCTAAGTTTATATTAATAATGT-GGTGAAA---------------------
C. parasitica (C.p.rnsi1)	GTGCGATCTGAAGTCTTTG-ATTAGAACTGTTGACAAAATTATTATAAATAGTAAACATT
C. parasitica (C.p.rnsi3)	GTGCGACAAGAAGCGGAT--GTTAGTAATGT-GGTGAGA---------------------
O. konnoana (C.k.rnsi1)	GTGCGACAAGAAGCAGAT--ATTAGTAATAC-GGTGTGA---------------------
C. ramosopulvinata (C.r.rnsi1)	GTGCGACAAGAAGTGGAT--ATTAGTAATAC-GGTGTGA---------------------
Cordyceps sp. 97009 (C.97009.rnsi1)	GTGCGACAAGAAGCGGGT--ATTAGTAATAC-GGTGTGA---------------------
Cordyceps sp. 97003 (C.97003.rnsi1)	GTGCGACAAGAAGCGGGT--ATTAGTAATAC-GGTGTGA---------------------
O. sobolifera (O.s.rnsi1)	GTGCGACAAGAAGCAGAT--ATTAGTAATAC-GGTGTGA---------------------
L. wingfieldii TOm1.3 (L.w.rnsi1)	GTGCGACAAGAAGTGAAT--ATTAGTAATAC-GGTGCAA---------------------
L. terebrantis CBS337.70 (L.te.rnsi1)	GTGCGACAAGAAGTGAAT--ATTAGTAATAC-GGTGCAA---------------------
L. wingfieldii TOM9.4 (L.w.rnsi1)	GTGCGACAAGAAGTGAAT--ATTAGTAATAC-GGTGCAA---------------------
L. lundbergiii DAOM60397 (L.lu.rnsi1)	GTGCGACAAGAAGTGAAT--ATTAGTAATAC-GGTGCAA---------------------
L. lundbergii NFRI89-1040/1/3 (L.lu.rnsi1)	GTGCGACAAGAAGTGAAT--ATTAGTAATAC-GGTGCAA---------------------
L. truncatum NFRI59-7/3 (L.tr.rnsi1)	GTGCGACAAGAAGTGAAT--ATTAGTAATAC-GGTGCAA---------------------
L. lundbergii NFRI1502/1 (L.lu.rnsi1)	GTGCGACAAGAAGTGAAT--ATTAGTAATAC-GGTGCAA---------------------
L. truncatum CBS929.85 (Lt.SSU/1)	GTGCGACAAGAAGTGAAT--ATTAGTAATAC-GGTGCAA---------------------
L. truncatum NFRI1813/1 (L.tr.rnsi1)	GTGCGACAAGAAGTGAAT--ATTAGTAATAC-GGTGCAA---------------------
L. truncatum Forintek C34 (L.tr.rnsi1)	GTGCGACAAGAAGTGAAT--ATTAGTAATAC-GGTGCAA---------------------

--	-----------------CCCCACCAAATAATCCAT-T-ATT-TGA-ACT--CTAAACA-G
AAGGGTGGATAGCCATCGTTTGCTGCGTACTAAGTATGTATGCACCCTGTAGATGATTTT	GTTTTATAAAATAAAGATTCAACAATGTATTTCGTTTTACATTAG-TCTATCTAGATATG
--	-----------------TTCCACCAGGTATTTTCTCT-ATC-TGACACTACCTACACATG
--	-----------------TTCCGTTTGGTATTTTCTCT-ACC-TAACACTATTTACACATG
--	-----------------TTCCGTTTGGTATTTTCTCT-ACC-TAACACTATTTACACATG
--	-----------------TTCCGTTTGGTATTTTCTCT-ACC-TAACACTATTTACACATG
--	-----------------TTCCGTTTGGTATTTTCTCT-ACC-TAACACTATTTACACATG
--	-----------------TTCCGTTTGGTATTTTCTCT-ACC-TAACACTATTTACATATG
--	-----------------TCCCGTTGGGTGTTTTCTCC-ACC-TAACACTATTTACATATG
--	-----------------TCCCGTTGGGTGTTTTCTCC-ACC-TAACACTATTTACATATG
--	-----------------TCCCGTTGGGTGTTTTCTCC-ACC-TAACACTATTTACATATG
--	-----------------TCCCGTTGGGTGTTTTCTCC-ACC-TAACACTATTTACATATG
--	-----------------TCCCGTTGGGTGTTTTCTCC-ACC-TAACACTATTTACATATG
--	-----------------TCCCGTTGGGTGTTTTCTCC-ACC-TAACACTATTTACATATG
--	-----------------TCCCGTTGGGTGTTTTCTCC-ACC-TAACACTATTTACATATG
--	-----------------TCCCGTTGGGTGTTTTCTCC-ACC-TAACACTATTTACATATG
--	-----------------TCCCGTTGGGTGTTTTCTCC-ACC-TAACACTATTTACATATG
--	-----------------TCCCGTTGGGTGTTTTCTCC-ACC-TAACACTATTTACATATG

CATGCGTT-AGAAG-TAATTTTAAAGCATGAAGC-CTGTAAA----------------TT	AAATTA--------------------------------------AG--AGT-T-AGAT-A
CATA--ACGGGTAACTGTTTT----GTATAAAGCTCTA-GAGAAAAGTAT----------	------------------------------------TCT-AAAAAGAATATTGGCTTTCT
CAAA--GC-AGTAA-TGGTTT----GTGTGAAGC--TGTAGGGACAATGTAGCCTGCCTT	AAATGGTGATGCAGATA--------------AAAAATCTTCAGGAGCTAGTGTAAGTA-A
TAAA--TT-AGTAA-TGATTT----ATGTGAAGC--TGTAATGACAATGTAGCCTCTTTT	AAATGGGAATGTTAATATAGAT--ATTCTTTTTATATTATCAGGAGCTAGTGTAAGTA-A
CAAA--TT-AGTAA-TAATTT----ATGTGAAGC--TGTAAAGACAATGTAGCCTCCCTT	AAATGGGAATGTTAATAT------ATTT-GTATATATTACCAGGAGCTAGTGTAAGTA-A
CAAA--TC-AGTAA-TGATTT----GTGTGAAGC--TGTAAAGACAATGTAGCCTCTCTT	AAATGAGAATGTTAATAT------AAA----TTATATTACCAGGAGCTAGTGTAAATA-A
CAAA--TT-AGCAA-TGATTT----GTGTAAAGC--TGTAAAGACAATGTAGCCTCTCTT	AAATGGGAATGTTAATGTAGATACTTTTT--CTATATTACCTGGAGCTAGTGTAAGTA-A
CAAA--TT-AGTAA-TGATTT----GTGTAAAGC--TGTAAAGACAATGTAGCCTCTCTT	AAATGGGAATGTTAATGTAGATACGTTT---CTATATTACCTGGAGCTAGTGTAAGTA-A
CAAC--TT-AGTAA-TAAATT----GTGTAGAGC--TGTAAAGACAATGTAGCCTGCTTT	AAATAGCAATGTTGATA-------ATTT---ATTTATCACCAGGAGCTAGTGTAAGTA-A
CAAC--TT-AGTAA-TAAATT----GTGTAGAGC--TGTAAAGACAATGTAGCCTGCTTT	AAATAGCAATGTTGATA-------ATTT---ATTTATCACCAGGAGCTAGTGTAAGTA-A
CAAC--TT-AGTAA-TAAATT----GTGTAGAGC--TGTAAAGACAATGTAGCCTGCTTT	AAATAGCAATGTTGATA-------ATTT---ATTTATCACCAGGAGCTAGTGTAAGTA-A
CAAC--TT-AGTAA-TAAATT----GTGTAGAGC--TGTAAAGACAATGTAGCCTGCTTT	AAATAGCAATGTTGATA-------ACTT---ATTTATCACCAGGAGCTAGTGTAAGTA-A
CAAC--TT-AGTAA-TAAATT----GTGTAGAGC--TGTAAAGACAATGTAGCCTGCTTT	AAATAGCAATGTTGATA-------ACTT---ATTTATCACCAGGAGCTAGTGTAAGTA-A
CAAC--TT-AGTAA-TAAATT----GTGTAGAGC--TGTAAAGACAATGTAGCCTGCTTT	AAATAGCAATGTTGATA-------ATTT---ATTTATCACCAGGAGCTAGTGTAAGTA-A
CAAC--TT-AGTAA-TAAATT----GTGTAGAGC--TGTAAAGACAATGTAGCCTGCTTT	AAATAGCAATGTTGATA-------ACTT---ATTTATCACCAGGAGCTAGTGTAAGTA-A
CAAC--TT-AGTAA-TAAATT----GTGTAGAGC--TGTAAAGACAATGTAGCCTGCTTT	AAATAGCAATGTTGATA-------ATTT---ATTTATCACCAGGAGCTAGTGTAAGTA-A
CAAC--TT-AGTAA-TAAATT----GTGTAGAGC--TGTAAAGACAATGTAGCCTGCTTT	AAATAGCAATGTTGATA-------ATTT---ATTTATCACCAGGAGCTAGTGTAAGTA-A
CAAC--TT-AGTAA-TAAATT----GTGTAGAGC--TGTAAAGACAATGTAGCCTGCTTT	AAATAGCAATGTTGATA-------ATTT---ATTTATCACCAGGAGCTAGTGTAAGTA-A

CTTATGGATAATGTAATGTGAATTCACGGTT-GAAGTGTAAA---TATTAAGAGTA--GG	AAA-AT-------------T---TCCCGCATTTT--CC----------TTAAT-GGGT-C
TCTATGGTTAGAGAAA-TCGAATCAATGTAATTAAAATTTTAC-TTATTCTTTA-A--GG	ATACATTATATTATATATAGTGAACT-GGCATT-TAACGTGTCCTTGGTTAAATGAAAAC
GACATGGAAAGCGAAA-GTTAAAGTAGCAAT-GACGCTTCG-TAGTACTAGTCATATTGG	AAG-GT------------AATAGATTCTACATACTAATCT-------ATTAGA-GAAA-C
GACATGGAAAGCGAAA-GCAAAATTAGCAAC-GACGCTTCGATATTTGTGTGTTAA--GG	AAG-GT------------AATAGGTTTTACATACTAATCT-------ATTAGA-GAAA-C
GATATGGAAAGCGAAA-GCGAAATTAGCAAT-GACGCTTTAATACTAGTGCACAAA--GG	AAG-GT------------AATAGATTTGACATGCTAATCT-------ATTAGA-GAAA-C
GATATGGAAAGCGAAA-GCGAAATTAGCAAT-GACGCTTCAATAATACAACACAGA-AGG	AAG-GT------------AATAGATTCGACATACTAATCT-------ATTAGA-GAAA-C
GATATGGAAAGCGAAA-GCGAAATTAGCTAC-GACGCTTTAATACTAGTGCACAAA--GG	AAG-GT------------GATAGGTTCGACATACTAATCT-------ATTAGA-GAAA-C
GATATGGAAAGCGAAA-GCGAAATTAGCTAC-GACGCTTTAAT--------ACAAA--GG	AAG-GT------------AATAGATTCGACATACTAAGCT-------ATTAGA-GAAA-C
GACATGGAAAGCGAAA-GCGAAA-TAGCAAC-GACGCTTCAAT--TAGTGCACAAA--GG	AAG-GT------------AATAGGTTTGACATACTAATCT-------ATTAGA-GATA-C
GACATGGAAAGCGAAA-GCGAAA-TAGCAAC-GACGCTTCAAT--TAGTGCACAAA--GG	AAG-GT------------AATAGGTTTGACATACTAATCT-------ATTAGA-GATA-C
GACATGGAAAGCGAAA-GCGAAA-TAGCAAC-GACGCTTCAAT--TAGTGCACAAA--GG	AAG-GT------------AATAGGTTTGACATACTAATCT-------ATTAGA-GATA-C
GACATGGAAAGCGAAA-GCGAAA-TAGCAAC-GACGCTTCAAT--TAGTGCACAAA--GG	AAG-GT------------AATAGGTTTGACATACTAATCT-------ATTAGA-GATA-C
GACATGGAAAGCGAAA-GCGAAA-TAGCAAC-GACGCTTCAAT--TAGTGCACAAA--GG	AAG-GT------------AATAGGTTTGACATACTAATCT-------ATTAGA-GATA-C
GACAAGGAAAGCGAAA-GCGAAA-TAGCAAC-GACGCTTCAAT--TAGTGCACAAA--GG	AAG-GT------------AATAGGTTTGACATACTAATCT-------ATTAGA-GATA-C
GACATGGAAAGCGAAA-GCGAAA-TAGCAAC-GACGCTTCAAT--TAGTGCACAAA--GG	AAG-GT------------AATAGGTTTGACATACTAATCT-------ATTAGA-GATA-C
GACAAGGAAAGCGAAA-GCGAAA-TAGCAAC-GACGCTTCAAT--TAGTGCACAAA--GG	AAG-GT------------AATAGGTTTGACATACTAATCT-------ATTAGA-GATA-C
GACAAGGAAAGCGAAA-GCGAAA-TAGCAAC-GACGCTTCAAT--TAGTGCACAAA--GG	AAG-GT------------AATAGGTTTGACATACTAATCT-------ATTAGA-GATA-C
GACAAGGAAAGCGAAA-GCGAAA-TAGCAAC-GACGCTTCAAT--TAGTGCACAAA--GG	AAG-GT------------AATAGGTTTGACATACTAATCT-------ATTAGA-GATA-C

```
---CGCAT----GTTAATATTACATATAGT-GGAATCCTAAGAGTATCATTAAA-GTGTT   AATATAAAAACTAG-GTAACCCCAATTACATCCAATTTT-TA-T-----------ATTG
GAACACGTAGTTAGTAAAA-----TAAAGTTGGAATC-TAAGG--A--A-GAGTTGTTCT   AATC-AAGGAAAAGAGTAACACCTATAATTTGCAG-GATATA-A-----GA-----AGTC
---CACATTTCCAGCGAAG-----TAAAGCTACTTTCCTAAATCTTATGTAATTTGTGCT   AATATTTGAACTTG-GTAACTCCT-TTATTTTCAGTAATGAGGGCGCAACGTCTTCATTA
---CAGATTTCCGACGAAG-----TAAAGCT-ATTTCCTAAATCTTATGTAAATTGTGCT   GATATCTGAACTTG-GTAACCCCT-TTATTTTCAGTAATATATTAG-ATA-T-AT-ATTA
---CATATTTCCGTTAAAG-----TAAAGCTAATTTCCTAAGTCTTACGTAAATTGTGCT   AATATCTGAACTTG-GTAACCCCT-TTATTTTCAGTAATATA-T-T-ATA-T----ATTA
---CATATTTCCGTTGAAG-----TAAAGCTAATTTCCTAAGTCTTATGTAAATTGTGCT   AATATCTGAACTTG-GTAACCCCT-TTATCTTCAGTAATGTATT-T-ATA-T----ATTA
---CATATTTCCGTTGAAG-----TAAAGCTAATTTCCTAAGTCTTACATAAATTGTGCT   AATATCTGAACTTG-GTAACCCCT-TTATCTTCAGTAATATACTAT-ATA-T----ATTA
---CATATTTCCGTTAAAG-----TAAAGCTGATTTCCTAAGTCTTACGTAAATTGTACT   AATATCTGAACTTG-GTAACCCCT-TTATCTTCAGTAATGTACTAG-ATA-T----ATTA
---CATATTTCCATTGAAG-----TAAAGCT-ATTTCCTAAATCTTACGTAAATTGTGCT   AATATTTAAACTTG-GTAACCCCCAGTGTTCTCAGTAATAAGTTATTATAAT-TT-ATTA
---CATATTTCCATTGAAG-----TAAAGCT-ATTTCCTAAATCTTACGTAAATTGTGCT   AATATTTAAACTTG-GTAACCCCCAGTGTTCTCAGTAATAAGTTATTATAAT-TT-ATTA
---CATATTTCCATTGAAG-----TAAAGCT-ATTTCCTAAATCTTACGTAAATTGTGCT   AATATTTAAACTTG-GTAACCCCCAGTGTTCTCAGTAATAAGTTATTATAAT-TT-ATTA
---CATATTTCCATTGAAG-----TAAAGCT-ATTTCCTAAATCTTACGTAAATTGTGCT   AATATTTAAACTTG-GTAACCCCCAGTGTTCTCAGTAATAAGTTATTATAAT-TT-ATTA
---CATATTTCCATTGAAG-----TAAAGCT-ATTTCCTAAATCTTACGTAAATTGTGCT   AATATTTAAACTTG-GTAACCCCCAGTGTTCTCAGTAATAAGTTATTATAAT-TT-ATTA
---CATATTTCCATTGAAG-----TAAAGCT-ATTTCCTAAATCTTACGTAAATTGTGCT   AATATTTAAACTTG-GTAACCCCCAGTGTTCTCAGTAATAAGTTATTATAAT-TT-ATTA
---CATATTTCCATTGAAG-----TAAAGCT-ATTTCCTAAATCTTACGTAAATTGTGCT   AATATTTAAACTTG-GTAACCCCCAGTGTTCTCAGTAATAAGTTATTATAAT-TT-ATTA
---CATATTTCCATTGAAG-----TAAAGCT-ATTTCCTAAATCTTACGTAAATTGTGCT   AATATTTAAACTTG-GTAACCCCCAGTGTTCTCAGTAATAAGTTATTATAAT-TT-ATTA
---CATATTTCCATTGAAG-----TAAAGCT-ATTTCCTAAATCTTACGTAAATTGTGCT   AATATTTAAACTTG-GTAACCCCCAGTGTTCTCAGTAATAAGTTATTATAAT-TT-ATTA
---CATATTTCCATTGAAG-----TAAAGCT-ATTTCCTAAATCTTACGTAAATTGTGCT   AATATTTAAACTTG-GTAACCCCCAGTGTTCTCAGTAATAAGTTATTATAAT-TT-ATTA

--GAGGTTTCATTTGTGAAG----------GT-GAA------ATAACTGTAAGTGGGTAA   AGGA--AATTCAAAAAAGCAAAAGC-------AGGGATTAAACTAATATTGAAACTT-AT
TTGCTGGGCTTACTGTAAA-----------GT-AAG------CTATCGATTAGAGGTGAA   GGGACAAAGT--AAAAAGCGAATGC-------AAGTATAAAATAGTTAGACGCAGTTAAT
AAGAGGTAACATATAAGGTT-AT----AT-GCTGAAGTTGTTATAGCGAATAAAGGAGAG   AGGAGAAGTT--AAAAAGCAAATGC-------GGGGCGAGGGCTCCTGTAGAAAATTTAT
AAGAGGTAACATATAAGGTT-TTAT--TTAACCGAAGTTGTTATACCCATTAAAGGGAAG   AGGATTTGTT--AAAAAGCAAATGC-------GGGGCGATGGCTCCTGTAGAAAAATTTT
AAGAGGTAACATATAAGACT-TTA---GTAGTTGAAATTGTTATATCCGATAAAGGGGAG   AGGATTTGTT--AAAAAGCAAATGC-------GGAGCGATGGCTTCTGTAGAAAGATTTT
AAGAGGTAACATATAAAGCT-TAA---AAGGCAGAAATTGTTATATCAGATAAAGGGAAG   AGGATTTGTT--AAAAAGCAAATGC-------GGGGCGATGGCTCCTGTAGAAAAATTTT
AAGAGGTAATATATAAGACT-TTC---TTAGTTGAAATTGTTATATCCGATAAAGGGGAG   AGGATTTGTT--AAAAAGCAAATGC-------GGGGCGATGGCTCCTGTAGAAAATTTTT
AAGAGGTAATATATAAGACT-TTA---TAAGTTGAAATTGTTATATCCGATAAAGGGAAG   AGGATTTGTT--AAAAAGCAAATGC-------GGGGCGAAGGCTCCTGTAGAAAATTTTT
AAGAGGTAACATATAAAGTTATTAACATTAACGGAAATTGTTATATCGAACCTGGGGAAG   AGGATCTGTT--AAAAAGCAAATGC-------GGGGCGATGGCTCCTGTAGAAAAATTTT
AAGAGGTAACATATAAAGTTATTAACATTAACGGAAATTGTTATATCGAACCTGGGGAAG   AGGATCTGTT--AAAAAGCAAATGC-------GGGGCGATGGCTCCTGTAGAAAAATTTT
AAGAGGTAACATATAAAGTTATTAACATTAACGGAAATTGTTATATCGAACCTGGGGAAG   AGGATCTGTT--AAAAAGCAAATGC-------GGGGCGATGGCTCCTGTAGAAAAATTTT
AAGAGGTAACATATAAAGTTATTAACATTAACGGAAATTGTTATATCGAACCTGGGGAAG   AGGATCTGTT--AAAAAGCAAATGC-------GGGGCGATGGCTCCTGTAGAAAAATTTT
AAGAGGTAACATATAAAGTTATTAACATTAACGGAAATTGTTATATCGAACCTGGGGAAG   AGGATCTGTT--AAAAAGCAAATGC-------GGGGCGATGGCTCCTGTAGAAAAATTTT
AAGAGGTAACATATAAAGTTATTAACATTAACGGAAATTGTTATATCGAACCTGGGGAAG   AGGATCTGTT--AAAAAGCAAATGC-------GGGGCGATGGCTCCTGTAGAAAAATTTT
AAGAGGTAACATATAAAGTTATTAACATTAACGGAAATTGTTATATCGAACCTGGGGAAG   AGGATCTGTT--AAAAAGCAAATGC-------GGGGCGATGGCTCCTGTAGAAAAATTTT
AAGAGGTAACATATAAAGTTATTAACATTAACGGAAATTGTTATATCGAACCTGGGGAAG   AGGATCTGTT--AAAAAGCAAATGC-------GGGGCGATGGCTCCTGTAGAAAAATTTT
AAGAGGTAACATATAAAGTTATTGACATTAACGGAAATTGTTATATCGAACCTGGGGAAG   AGGATCTGTT--AAAAAGCAAATGC-------GGGGCGATGGCTCCTGTAGAAAAATTTT
AAGAGGTAACATATAAAGTTATTAACATTAACGGAAATTGTTATATCGAACCTGGGGAAG   AGGATCTGTT--AAAAAGCAAATGC-------GGGGCGATGGCTCCTGTAGAAAAATTTT

GCTTGAGCTGTATGCGATGAAAGTCGCACGTACAGTTCTTAGAGGATGTTAATATAGTAA   TATTTAACGTAGCCTCAAC
GTTTGAGCCGTATGCCATGAAAGTGGCACGTACGGTTCTGAGGGGGGGAAAGTCTGAGAG   GACCT-ACCTA-TCCCAAC
GTTTGAGCCGTATGCGATGAAAGTCGCACGTACGGTTCTTAGAGGGGGAAAGTTCGAGAG   GACCT-ACCTA-TCTCAAC
GCTTGAGCCGTATGCGATGAAAGTTGCACGTACGGTTCTTTAAGGGGGAAAGTCCGAGAG   GACCT-ACCTA-TCTTGAC
GCTTGACCCGTATGCGATGAAAGTTGCACGTACGGTTCTTTAAGGGGGAAAATCTGAGAG   GATCT-ACCTA-TCTTGAC
GCTTGAGCCGTATGCGATGAAAGTTGCACGTACGGTTCTTTAAGGGGGAAAGTTTGAGAG   GACCT-ACCTA-TCTTAAC
GCTTGAGCCGTATGCGATGAAAGTTGCACGTACGGTTCTTTAAGGGGGAAAGTTCGAGAG   GACCT-ACCTA-TCTTGAC
ACTTGAGCCGTATGCGATGAAAGTTGCACGTACGGTTCTTTAAGGGGGAAAGTCCGAGAG   GGCCT-ACCTA-TCTTGAC
GCTTGAGCCGTATGCGATGAAAGTTGCACGTACGGTTCTTAAAGGGGGAAAGTTCGAGAG   AACCT-ACCTA-TCTTAAC
GCTTGAGCCGTATGCGATGAAAGTTGCACGTACGGTTCTTAAAGGGGGAAAGTTCGAGAG   AACCT-ACCTA-TCTTAAC
GCTTGAGCCGTATGCGATGAAAGTTGCACGTACGGTTCTTAAAGGGGGAAAGTTCGAGAG   AACCT-ACCTA-TCTTAAC
GCTTGAGCCGTATGCGATGAAAGTTGCACGTACGGTTCTTAAAGGGGGAAAGTTCGAGAG   AACCT-ACCTA-TCTTAAC
GCTTGAGCCGTATGCGATGAAAGTTGCACGTACGGTTCTTAAAGGGGGAAAGTTCGAGAG   AACCT-ACCTA-TCTTAAC
GCTTGAGCCGTATGCGATGAAAGTTGCACGTACGGTTCTTAAAGGGGGAAAGTTCGAGAG   AACCT-ACCTA-TCTTAAC
GCTTGAGCCGTATGCGATGAAAGTTGCACGTACGGTTCTTAAAGGGGGAAAGTTCGAGAG   AACCT-ACCTA-TCTTAAC
GCTTGAGCCGTATGCGATGAAAGTTGCACGTACGGTTCTTAAAGGGGGAAAGTTCGAGAG   AACCT-ACCTA-TCTTAAC
GCTTGAGCCGTATGCGATGAAAGTTGCACGTACGGTTCTTAAAGGGGGAAAGTTCGAGAG   AACCT-ACCTA-TCTTAAC
GCTTGAGCCGTATGCGATGAAAGTTGCACGTACGGTTCTTAAAGGGGGAAAGTTCGAGAG   AACCT-ACCTA-TCTTAAC
```

Supplementary Figure 7.8 Alignment of the LHEase amino acid sequence.

```
P.a.COX1i5              WIVVGFSDAESSFMIRIRKNSK--YKTGWT-VVAVFSIAVDKKDLFLLESIKTFFGGLGS
C.p.rnsi2               WFV-GFSDGESSFQIQIKYSDQ--DKTKIRGVNFSFTISLHIDDIAVLKFIKDKLE-IGN
O.n.u.rnli1 (U7)        WFV-GFTDAEGCFHIKKKTMSN--GN-T-R-YYFEFVIKLHIDNVALLKYIQEFLG-IGR
O.i.rnli1 (U7)          WF-SGFVDAEGCFHIKSKTMSN--GN-K-R-YYFEFVIKLHIDNVLLVKFIHETLG-IGR
O.m.rnli1 (U7)          WF-SGFVDAEGCFHIKSKVMSN--GN-K-R-YYFEFVIKLHIDNVLLVKFIHETLG-IGR
P.a.rnli1               WV-AGFIDAEGCFHIKRKIMSN---N-KMR-YYFEFVIKLHIDDEAVLRTICHILG-VGR
N.c.ND5i1               WVWSGLIDGEGSFNIIVDRNKS--RKLGWR-AQLKFQLSLHTKDLNLLYLLQQYLGGIGS
P.a.ND5i2               GVWSGLIDGEGSFSIILVRNST--RKLGWR-IEPKFQLGLHKKDYDVLLQLQLFLGGAGG
P.a.ND4Li1              WFITGFTDGEGCFWIGVRRDPR--NKLGWC-VQAFFQIALHKKDEALLKAIQKYFGGIGA
C.p.ND4i1               WWITGFTDAEGSFGLYIYKNSK--FKTGWH-AYLVFSISLHEKDRDILSQIQNFFG-FGG
N.c.ND4Li1              SFITGFSDAEGSFVVTILKNPR--YKIGWN-VQARFQIKLNEKDRALLLLIQNYFDNIGY
P.a.ND4Li1              SFISGFSDAEGSFVVTILNNPR--YKIGWN-VQARFQIKLHEKDRALLLLIQNYFYNVGY
I-LtrIII-P (rns)        WYICGLVDAEGSFGVNVVKHAT--NKTGYA-VLTYFELAMNSKDKQLLELIKKTFDLECN
I-LtrII (rns)           WYICGLVDAEGSFGVNVVKHAT--NKTGYA-VLTYFELAMNSKDKQLLELIKKTFDLECN
I-LteI-P (rns)          WYICGLVDAEGSFGVNVVKHAA--NKTGYA-VLTYFEIAMNSKDKQLLELIKKTIDLECN
I-LwiI-P (rns)          WYICGLVDAEGSFGVNVVKHAT--NKTGYA-VLTYFEIAMNSKDKQLLELIKKTFDLECN
I-LluIII-PD (rns)       WYICGLVDAEGSFGVNVVKHAT--NKTGYA-VLTYFEIAMNSKDKQLLELIKKTFDLECN
I-LluI-PD (rns)         WYICGLVDAEGSFGVNVVKHAT--NKTGYA-VLTYFEIAMNSKDKQLLELIKKTFDLECN
I-LluII-PD (rns)        WYICGLVDAEGSFGVNVVKHAT--NKTGYA-VLTYFEIAMNSKDRQLLELIKKTFDLECN
I-LtrIV-PD (rns)        WYICGLVDAEGSFGVNVVKHAT--NKTGYA-VLTYFELAMNSKDKQLLELIKKTFDLECN
C.k.rnsi1               WFVSGLVDAEGSFGVNVVRKESLRSKIGYS-ILIYFEVALNKKDKQLLETIKQVLNIENT
C.97009.rnsi1           WFVSGLVDAEGSFGVNLVKKES---KLGYA-VLTYFEIGLNKKDEQLLETINKVLGTNKI
C.r.rnsi1               WYVSGLVDAEGSFGVTLVKKES--SVTGYS-VLIYFEIALNKKDKQLLETIKQVLGIDKN
P.a.cobi1               WF-SGFTDAEGNFLISIDRN--------Y--VKLRFKISLHIDDLKVLQVIQSNL-NIGR
O.s.rnsi2-3             WFITGFCDAEGSFSIYLRKKSG--GST-YS--EARFGISLHKKDLDTLKRIQVYFKGKGS
P.a.ND1i4               EWITGFTDGEGSFGIKIYKAKG--HKLGWR-LDPFFQIKLNARDLDILYRIKEYFGGVGT
P.a.ND3i1N              SYISGFVDGEGSFMLTIIKDNK--YKLGWR-VVCRFVISLHKKDLSLLNKIKEFFD-VGN
C.p.rnsi1               WF-TGFCDAEGNFQVYP-KKRV--LKSGE--V-SKYNVGLGFHLSLHSRDAELLHVIHDK
I-OnuI (U11 rnl)        WILTGFADAEGSFLLRI-RNNN-KSSVGYS-TELGFQITLHNKDKSILENIQSTWKV-GV
C.p.rpmi                WVVTGLVDAEGSFNITVVKNNN--SKLGWV-TKLRLEMSMLKKDRFTLEQLINYFGG-GG
HyjeMp12                NFLTGFTDAEGSFVLSITKSDN--VKSGWV-IKPRFSISLHKKDKFVLEAIKNYLGV-GE
HyjeMp02                RPQGGFSDGEATFTFNIIKSPS--NPVKYE-LQLFYIITLHKKDLPLLALIQKYFDNKGQ
N.c.ATP6i2              DYITGFVDGEGCFSLSLFKDDR--RLNGWQ-VKPIFSISLHKKDISLLEAIQRTFKV-GK
A.a.rnsi1               WFITGFTDAEGSFMVSITKNEN--SKLKWG-IYPSFAIHIHNKDISLLNQIQKTLGV-GN
O.s.rnsi2               WFITGFSDAESSFIISIYKDEN--NKLKWR-VSAYFSIHVHIKDLPLLELIQKTLGV-GI
C.p.rnsi4               WFLTGFSDAECSFSILIQANSK--YSTGWR-IKPVFAIGLHKKDLELLKRIQSYLGV-GK
A.p.rnli2               FFVTGITDAEGCFSVGISKHKQ--LRTGWC-VKPVFSITLHERDEALLKQLLSFFGK-GG
C.p.rnsi3               WFITGLVDAEGSFVVNIVKDSS--KSLGFF-VLPYFEIALNEKDRDLLTGIQDYLGI-GN
S.s.rnsi1               WFLTGFIDGEGCFRISLTKV-T--RAIGWR-VQLFFQINLHKKDIALLEDIRDYFGV-GM

I-KKHGKG-TFSYRIESSEQIMKFII-PFFDKYPLITEKLGDYLLFKKVVEMLNNKEHLT   ETGLYKIV-SLKAS-INKGLSEELQAAF-PQCIPVFR---PTVY----NKII---PDP-N
ITVKKTRA-ACVFSVTNQEGL-NKLI-SIFDKYNLNTTKYLDYLDFRKAFLLYQDKDINF   NKENLNILINQIVDLKS---RMNSERTF--FDMSVCN---TSFSTNSLHSG----INKN-
IAIRTNSN-TCSFEVGSEKDL--RILIDIFDNIELMGNKYLDFLSFKKAFFLYFDRSSLV   TESLIESI-EAIRSNHN---IKRTDYTI-PVDFKCVI---TDY-----------------
IAIRSNSN-SCTFEVGSEKDL--RILIDLLDKCPLIGAKVLDFLSFKKAFFFYFDRPGLV   TDEIIDAI-EKIRNNHN---TRRTDFKL-PLDFKSSI---TDY-----------------
IAIRSNSN-SCTFEVGSEKDL--RILIDLLDKYSLIGSKVLDFLSFKKAFFLYFDRPGLV   TNEIIDAI-EKIRNNHN---TRRTDFKL-PLDFKCII---TDY-----------------
VVLRPKYN-TCSFEVGSEKEL--RILIDLLDEFQLMGDKYLDYINFREAFFLYFDRSDSV   TESLKTKI-EEIRANHN---TKRASTTL-PVNYKRVI---TDY-----------------
IHLARNRD-IVNYSIDSIEDL-NKLII-HLENYPLLTQKAADFFLFKQAVKLVNNKAHLT   VEGLNQIV-NIKAS-MNLGLSDTLKSEL-AGYTPVER---PVINCDNVFLD----PY---
IYSARNGE-FVNYIISSIKDL-NKLVV-HLEHYPLLTQKSVDFFFFKQIIDLINNKAHFT   VEGLNKIV-NLKAS-MNLGLSDKLKLEF-PGYQVMER---PIFNCIEVAANIS--PS---
V-KIRG-D-KCTFIVSSLSQI-TKVIIPHFDSYPLITNKLADYFLWKGIIEIMKAKKHLT   LEGLNEIV-RMKAQ-LNLGLSDELKTAF-AETFSTSAIIKRILNKESIPHGM--------
I-HSHGSN-SLKYTVKSLNEL-Q-IIIDHFDKYLLITSKLNDYKLFKLAYNLFINKENLS   IEGIEKLV-AIKSS-MNLGLKSQLKLAF-PQISKHKDIYRDCFAYSNI-KKI---PDPY-
ISKINDRS-TVEFRVSDITSL-NNIIIPHFEKYQLITNKYGDLVIFKQIVSLMLENKHTT   LEGLKEIL-EHRAS-LNWGLSKTLKESF-PSIIPVKRVKIENNILSNL-SSLPLLPRGGN
ISTLKGGS-TVEFRVSDITSL-NNIIIPHFEKYTLITNKYKDFIIFKQIVSLMSENKHTT   LEGLKEIL-EYRAS-LNWGLSKNLKESF-PSIVPVKRVEIEDNILSNLPSN---------
IYHNPSDD-TLKFKVSNIEQIVNKII-PFFEKYTLFSQKRGDFILFCKVVELIKNKEHLT   LNGLMKII-SIKAA-MNLGLSENLKKEF-PGCLSVKR---PEFGLSNLNKR---------
IYHNPSDD-TLKFKVSNIEQIVNKII-PFFEKYTLFSQKRGDFILFCKVVELIKNKEHLT   LNGLMKIL-SIKAA-MNLGLSENLKKEF-PGCLSVKR---PEFGLSNLNKR---------
IYHNPSDD-TLKFKVSNIEQIVNKII-PFFEKYTLFSQKRGDFILFCKVVELIKNKEHLT   LNGLMKIL-SIKAA-MNLGLSENLKKEF-PGCLLVKR---PEFGLSNLNKR---------
IYHNPSDD-TLKFKVSNIEQIVNKII-PFFEKYTLFSQKRGDFILFCKVVELIKNKEHLT   LNGLMKIL-SIKAA-MNLGLSENLKKEF-PGCLLVKR---PEFGLSNLNKR---------
IYHNPSDD-TLKFKVSNIEQIVNKII-PFFEKYTLFS-----------------------   ------------------------KKEF-PGCLSVKR---PEFGLSNLNKR---------
IYHNPSDD-TLKFKVSNIEQIVNKII-PFFEKYTLFS-----------------------   ------------------------KKEF-PGCLSVKR---PEFGLSNLNKR---------
IYHNPSDD-TLKFKVSNIEQIVNKII-PFFEKYTLFS-----------------------   ------------------------KKEF-PGCLSVKR---PEFGLSNLNKR---------
IYHNPSDD-TLKFKVSNIEQIVNKII-PFFEKYTLFSQKRGDFILFCKVVELIKNKEHLT   LNGLMKIL-SIKAA-MNLGLSEN----------------------SNLNKR---------
LYYNESDN-TLKLKVSNLDKLINNVI-PHFKKYPLFTQKKVDFLLMCKVLKIVKEKRHLT   IIGLNEIL-SIKTT-MNLGLSDKLKNEF-LNIKPLTR---LIVNTGIPNKD---------
LYYNKSDD-TLKLKISNIDELINNVI-PHFKKYPLFTQKRVDFELMSKVLEIIQSKRHLT   EEGLNEIV-SIKAG-MNLGLSDKLKNEF-STIKPLTR---LTINTGIPNKD---------
LYYNENDN-TLKLRISNLDVLIKNVI-PHFNKYPLFTQKRVDFLLMCKIIELVEEKKHLT   KEGLNEIL-NIKAA-MNLGLSDKLKNEF-LTVKSLTR---LSIDTGIPNKD---------
I--T-VEEVRSRCSFIVEDISGINTICSIFNNYPLHTSKKLDFQDFYEALLIRTKEKIIS   PTNLEKIL-CLKNG-MNS---KR-EI-F--TY-DITK-----------SQII---IDP-N
I-VKHGED-SIQYAITSIEQL-STLVIPHFDNYPLISKKHADYLLFRKAVLLIRNKKHLT   IEGLQEIV-SIKAS-MNKGLSYELKEVY-PNTLIVPR---PLVP-NSII------PDPE-
I-SFEK-N-TAKYIIRKLSDI-VDILIPHFEAYPLLTKKFADFELFRQIILIIKNESSLS   EQGFIKIL-NLRYY-LNKGISEELKELY-PNLVPVMR---PEVPERKIQPE---------
V-FLMTKD-SAQYRVESLKGL--DLIINHFDKYPLITKKQADYKLFKMAHNLIKNKSHLT   KEGLLELV-AIKAV-INNGLNNDLSIAF-PGINTILR---PDTSLPQ-ILN----PF---
LNNVGVYY-EFKNRNEARIAVNDRAGL-KVIVD-VFGIFPLVTIHQYTRYCLLKEYLIND   IKEFKTLE-KFNAF-KDKCL-DNIDLSS-KVLSGLE---SQINSGL--LDS---------
I-ANSGDN-AVSLKVTRFEDL-KVII-DHFEKYPLITQKLGDYMLFKQAFCVMENKEHLK   INGIKELV-RIKAK-LNWGLTDELKKAF-PEIISKER--SLINKNI-PNFK---------
IRVL-DDN-NLRFYIESLKDL-TTVI-NHFDKYPLFTKKQEDYLLFKQVFELFKNKNHLT   MEGLRQIV-AIKALLRNKGLSNSLLEAF-PGITPATL---PITDTITFGVDKSIL-SP--
IYTQ-GTD-SIQYRVFSIKDL-QLVI-DHFDKYPLISQKFGDYSLFKQAYLLLINKEHLT   PEGLLKII-AIRAS-INNGLSEALKEAF-PDVIPVMK---PVPRGSRNNISIYN-PQ---
ELQV-TDD-NVILTISSIKDL-EAVI-KHFDEYPLITQKLVDYLLFKQVYTIICNKEHLT   LEGFQKIV-AIKSV-LNKDLSDKLKYEF-KDIIPVTR---PLVKLPTTVL-----PS--Y
IYKH-GID-SIQYRVSSLKNL-Q-IITDHFDSYPLITQKRVDYLLFKQAIALIKNKEHLS   LEGLLKLV-GIKAT-L-RSSWPNLKKVF-PTVKAAVR-PSVIYITSDVKVKSLN------
VRKN-SKT-TALFRVDNLKEL-QVII-NHFDKYPLVSFKASDYIIFKKCYNLIMQKQHLT   QKGFEEIL-SLKYK-LNKGLPDNLKKTF-PHIIPIER---PEY--IFINI-----PNPY-
VRKN-NKN-TVLFRVSNMQEL-QIVI-DHFKRYPLISAKYSDFLLFEQCYNLIKQKEHLT   QKGMEKIL-ALKSN-LNKGLSNELKDAFAANIVPVSR---LEY--KFTGI-----PSPF-
IHIH-GKD-SIQFRIDSPKEL-EVII-NHFENYPLVTAKWADYTLFKKALDVILLKEHLS   QKGLLKLV-GIKAS-LNLGLNGSLKEAF-PNWEELQI---DRPSYVFKGI-----PDPN-
IFRH-GPT-TLQVRFESKNAL-WAVV-KHFDKYPLISQKQADYLLWKKAVELFSDKEHLT   LAGLFEVV-ALRAS-INWGLPPRLRKAF-PLIASKIR---PEVKKP--VV-----PNPE-
ITFNSSDN-TFKLKVSGLDGLSNIVI-RHFQQYYLLSKNVPILQLFSRVVGIMKSGDHLN   LKGFQEIV-NIRAS-LNLGLSDKLKAYF-PETLPVVR---PEFN-I-KSI-----PDPW-
IHKS-GTN-LVQYRIQTFDEL-SILI-NHLNDYPLVSQKKWDFELFKQAHELVKMNEHLN   KEGILKIV-SLKAS-LNLGLSEALKLAF-PNVKNATK-----LTSSTVSI-----PDPH-
```

```
WLAGFVSGEGCFKSIL--------------KKSSSVKVGFQSILIFQVTQHARDEKLME-    --------------SLI---SYFKCG-YI-EKDPR-G--PWLSYIVSNFTDIYTKIIP-F
WLLGFIEAEGSFFISI-------------TDIEPSFSIELSNAQMF-LLEKIKDFLISDL    G--------FD-GYSLFQL-KSSSFSVISV-NKQKVK--PSAILLIKNIRVLNNYLVP-Y
KLLGLFEGDGSFVIQK-------------QGLSPKFEIELTKTQRP-LLESIQNYLINEL    ----------KLGDS-----ELLVKQIRIIELPAK-G---NSKPTVRLTVTGVNLLHN-Y
KLLGLFEGDGSFVIQT-------------QSLTPKFEIELTEVQKL-LLNEIQKYLINYF    L---------GLGLSL----EQLEKAIKIQDLKAK-G---NSKATVRLTITGIDFLNN-Y
KLLGLFEGDGSFIIQT-------------QSLTPKFEIELTEVQKL-LLNEIKEFLINYF    L---------ELGLSL----EQLEKAIKIQHLKAK-G---NSKATVRLTVTGIDFFYN-Y
KLLGFIEGEGSFIVQT-------------RGLTPKFELELTSAQKP-LLIDICEYLMLRI    Q---------LIADATQPLNLPSKGGIKIRDRKAK-G---NSKPSVRLEITGIDFLHN-Y
WISGFVSAEGNFDVRM--------------PST-NSKLGYRVQLRFRISTLRVNIRLME-    --------------KIV---EYFGSG-KI-YKYGG-K--SAVSLTIVDFTDITNILVP-F
WISGFVSAEGNFDVRT--------------PKT-NSKTGYRVQLRFRITQHVRDIILME-    --------------KLV---QYLGGG-KV-YKY-T-R--SGVHLSIVDFSLITNRIIP-I
WMAGFTSGEGSFFVNI--------------FKSSHHKVGYQIRLEFEIAQHIRDELLML-    --------------SFT---EFFGCG-II-SRY-S-S--DMVKFRCTKFSDISTIIVP-F
WVSGFTSGEGNFMIDI--------------IKSKSSKLGITAGLRFTITQHFRDKQLMI-    --------------SLI---DFFNCG-HL--NIRN-N--NCFNLTIRKFTDLDTKIIP-F
WVAGFSSGEANFFITM---------------------SGTKVWLRFSIAQDSRDILLLK-    --------------SLV---KFFNCG-YI-AQYKN-R--KVCEFIVTKINDIIIYIIP-F
WVAGFSAGESNFFITI---------------------SGNKVWLRFSIAQDSRDILLLK-    --------------SLV---EFFNCG-YI-AQYKN-R--KVCEFIVTKINDIIIYIIP-F
WLAGFIEGEACFFVSI--------------YNSPKSKLGKAVQLVFKITQRIRDKILIE-    --------------SIV---ELLNCG-RV-EVRKS-N--EACDFTVTSIKEIENYIIP-L
WLAGFIEGEACFFVSI--------------YNSPKSKLGKAVQLVFKITQHIRDKILIE-    --------------SIV---ELLNCG-RV-EVRKS-N--EACDFTVTSIKEIENYIIP-F
WLAGFIEGEACFFVSI--------------YNSPKSKLGKAVQLVFKITQHIRDKILIE-    --------------SIV---ELLNCG-RV-EVRKS-N--EACDFTVTSIKEIENYIIP-F
WLAGFIEGEACFFVSI--------------YNSPKSKLGKAVQLVFKITQHIRDKILIE-    --------------SIV---ELLNCG-RV-EVRKS-N--EACDFTVTSIKEIENYIIP-F
WLAGFIKGEACFFVSI--------------YNSPKSKLGKAVQLVFKITQHIRDKILIV-    --------------SIV---VLLNCG-RV-EVRES-N--EACDFTVTSIKEIENYIIP-F
WLAGFIEGEACFFVSI--------------YNSPKSKLGKAVQLVFKITQHIRDKILIE-    --------------SIV---ELLNCG-RV-EVRKS-N--EACDFTVTSIKEIENYIIP-F
WLAGFIEGEACFFVSI--------------YNSPKSKLGKAVQLVFKITQHIRDKILIE-    --------------SIV---ELLNCG-RV-EVRKS-N--EACDFTVTSIKEIENYIIP-F
WLAGFIEGEACFFVSI--------------YNSPKSKLGKAVQLVFKITQHIRDKILIE-    --------------SIV---ELLNCG-RV-EVRKS-N--EACDFTVTSIKEIENYIIP-F
WLTGFMEGESCFYISV--------------YNSPKSLLNYAVQLVFKITQHSRDKVLLE-    --------------NIS---KLLECG-RV-ETRKS-G--DACDFTITSFNEFKKYMIP-F
WLAGFMEGESCFYVSI--------------YNSPKSKLNYAVQLVFKITQHFRDKVLLE-    --------------NIS---ILLECG-RV-ETRKS-G--DACDFTVTSFNKFKKNMIP-Y
WLEGFMEGESCFFVRV--------------YNSPKSKLKLAVQLAFIITQHSRDKILLE-    --------------NIS---KLLKCG-RV-ETRKS-G--DACDFVVTSFKDFTKYMIP-Y
WFIGFMEGEGTFGIKT--------------GSSLYLQVAQKNTSQECLNGITNYLIALS-    --------------NNT---KYFHTQ-SN-KILPLNV--TSAINAKTSVVSLVVSNIDAL
WVAGFTSGEGCFMIKI--------------SKSPASKLGFGIQLIFQLTQNNRDEALMK-    --------------CIL---TYFGCG-TL-VKD---G--TKIVFFVRKFSNIKDIIIP-F
WLVGFIDGEGSFNIITV--------EKKSSDAPSILSTSYKVWLHFQITQHSRDTTLME-    --------------RIV---TFFDCG-SV-KKRNT----DAVDFKLNKFELLENIIIP-F
WLSGFVDAEGCFSVVVF--------------KSKTSKLGEAVKLSFILTQSNRDEYLIK-    --------------SLI---EYLGCG-NT-SLDPR----GTIDFKVTNFSSIKDIIVP-F
WIVGLINGEGCFYLNK------------------GRCNFFIEHTDLQALEIIKKRLSFS-    -------------PKIV---ARSPRARDV-GKDIK---PTYM-LIVSSKKDIES-LI-EL
WLAGFTSGEGCFFVNL--------------IKS-KSKLGVQVQLVFSITQHIKDKNLMN-    --------------SLI---TYLGCG-YI-KEKNK-SEFSWLDFVVTKFSDINDKIIP-V
WLAGFTSGDGNFYISI--------------GEGPK-H--VQVQLVFTLTQHIRDQALMN-    --------------SLI---SYLGCG-NI-KHNEK-N--SWLQFIVTKFSDIDEKIIP-I
WLAGFTSGEGSFGVKV----------RK-AKEN--SN-AF-IELIFQINQHVRDKQLMA-    --------------CIA---EYLECG-KI-Y-KHS-L--NAIVYRVSKTSNLTEKVIP-F
WLAGFTSAKGHFGVKL-------------SKDS-TTITGYQVVLKFTLTQQLRDKLLLE-    --------------NFI---SFYNCG-TF-WKSES-K--DWGVFEVTKFSDVESIIIP-F
WIRGFIEGEGCFQVIT--------------QNS-XXPKGRNVWLRFSLTQHIKDEELLK-    --------------DIA---IYLNMG-RY-YKSPTRN--EG-QYLITIFSDINNKLIP-F
WISGFASGDSTFSVSI--------------ENS-NNKLGKRVRLIFGTCLHIRDKELLIG    MANYFYNSFENLNKSVLNKQNKLSDSIEHKYKYNY-DSPTTSLLQIKKYSDIRDIIIP-F
WISGFATGDSSFSVSI--------------EKS-TSKVGKRVRLIYGTCLHIRDKDLLRG    MAKYF-NSLP----SGCI--DVVNK--EISIHCN--EINNTSLLQIKNNSDIENKIIP-F
WISGFASGDSSFNVKI--------------SNSPTSLLNKRVQLRFGIGLNIREKALIQY    LVAYF---------------DL-S---DN-LKNIYFD-LNSARFEVVKFSDITDKIIP-F
WMAGFISGEGCFIVEIL-------------ENK-EMRLGYSVKIKFNIPQHKRDIILLQ-    ---------------VLN--EYLGGG-GN-FHQAKNR--DVMEFRFQDFSIIDEQIVP-F
WLAGFTEGESCFFIPLRRAVPPLKGAQARVYKSVKSKQGFAVQLAFTITQHSRDKELLN-    ---------------VIA--QFLECG-RV-AD-RS-N-GGACDFIVNSIKDLKNKIIP-F
WFSGFTSAEGCFMVGI--------------AKSKESTTGYQVYLSFIVTQHVRDELLLK-    --------------CLI---DYFSCG-RL--ARK--R--DVYEYQVSKFSDVEKFID--F

FL----Q--YNIIGSKNLD---FNDWCKIATLMQDKHHLT--TEGLNK-IISIKGG
LST---E--V-FKSKKGQD---FEDWKLICEAVYKGSHKI--DE-IRELILRLTYS
FNTKLKN--LKFFTIKGKD---FESFSLICKVLFNKAHLK--DENIKNLIIKLIN-
FNVYLKK--LQFFSDKSKD---FEIFCTICESLFNKAHIN--NKEVKNSLLNLAK-
FNVYLKE--LQFFSDKSKD---FEIFCTICESLFNKAHIN--NKEVKNSLLNLAK-
FNVFLSK--LEFYSRKKED---FKSFCFICNKLFSKAHID--NEEVKILLLLRLSE-
FN----K--YPIIGIKLYD---YLDWCKIHSLMINRSHLT--VEGINS-IS-LLLG
FK----E--NPLVGVKSKD---FEDWCKIHELMVNRSHLT--IKGMNF-IQEIKLG
FK----E--YLPEGNKIKD---FQDFCKVCELMENNAHLT--EEGLNQ-IREIKAR
FS----K--YPIIGDKLLN---FQNFIEASKLINNKDHLT--IEGLKK-INEIKNG
FD----Q--YKIEGSKYND---YVKFKEAAILIKNKEHLT--EKGLNK-IIELKNS
FE----K--YKIEGSKYKD---YVNFKEAAIPIKNKEHLT--EKGLNK-IIELKNS
FN----E--YPLIGQKLKN---YEDFKLIFDMMKTKDHLT--EEGLSK-IIEIKNK
FN----E--YPLIGQKLKN---YEDFKLIFDMMKTKDHLT--EEGLSK-IIEIKNK
FN----E--YPLIGQKLKN---YEDFKLIFDMMKTKDHLT--EEGLSK-IIEIKNK
FN----E--YPLIGQKLKN---YEDFKLIFDMMKTKDHLT--EEGLSK-IIEIKNK
FN----E--YPLIGQKLKN---YEDFKLIFDMMKTKDHLT--EEGLSK-IIEIKNK
FN----E--YPLIGQKLKN---YEDFKLIFDMMKTKDHLT--EEGLSK-IIEIKNK
FN----E--YPLIGQKLKN---YEDFKLIFDMMKTKDHLT--EEGLSK-IIEIKNK
FN----E--YPLIGQKLKN---YEDFKLIFDMMKTKDHLT--EEGLSK-IIEIKNK
FE----NYRYALKGQKAND---CKDFQRVFEIMLNKGHLT--EEGLAK-IIEIKNS
WE----N--YPLIGHKSKD---YFDFKRVYEIMLAKGHLT--KEGLAN-IIEIKNS
WK----D--YPLIGNKSKD---CLDFIRVYEIMLNKGHLT--EEGLTK-IVEIKNS
YYY---I--LPLLDESKFYSRKFEDFK-LWR-MALILKIR----GYYY-TLEGK--
FN----D--HKIVGVKLQD---FLDWYKAAEIVKAKGHLT--PLGLDE-LKKIKVG
FQ----K--YPLQSAKVFD---FLSFIEAANVIKSKKTRQWTAEEFAK-IQNIQ--
FI----K--YPLKGNKNLD---FTDFCEVVRLMENKSHLT--KEGLDQ-IKKIRNR
LD-SKRV--VPLAGHKLMQ---YNEWKHTWF-----------------------NK
FQ----E--NTLIGVKLED---FEDWCKVAKLIEEKKHLT--ESGLDE-IKKIKLN
FK----E--NKILGEKFKD---FQDWCRAAELMKNKAHLT--PSGLEE-IRKLKAG
FI----K--YPILGIKALD---FKDFCSLAELISNKAHYT--EEGLNK-IISIKAN
FQ----K--YKVVGVKAKD---FASFCFICKMMKEGRHKT--EQGLLE-IQAIKNN
LK----E--YPLLGVKQED---FLDFVKIAKLIESKTHLT--DEGLDT-IKLIQSN
FN----K--YPILGVKHLD---FYDFKLISNLIDNKEHLT--QDRF-K--------
FN----E--YPILGVKKLD---YEDFKKVAELVKNKEHLN--VEGLNK-IIKIAEG
FD----K--YSIQGKKSLD---YINFKEVADIIKSKNHLT--SEGFQE-ILDIKAS
IR----S--FPIQGMKIHD---FNDWCQAIDIVRRKGHLT--PEGLAE-IQKLEEG
FN----K--YPLLGTKSLN---YQDFVKVFELMDTKKHLT--QDGLAT-IIKIKSG
FD----K--YPILGEKSKD---YLDFRTVSEIMRSKDHLT--EVGVAK-VRIIKQG
```

Supplementary Figure 7.9. (A) Optimized DNA sequence of I-LtrII. (B) Amino acid sequence of I-LtrII, based on the optimized DNA sequence.

5’-

ATGATTAACCTGAAGAATAACATCGAATATCTGAATTGGTACATTTGTGGCCTGGTGGATG
CTGAGGGTTCCTTCGGTGTTAACGTAGTAAAACACGCTACTAACAAAACCGGTTACGCGG
TTCTGACCTATTTTGAACTGGCTATGAACAGCAAAGACAAACAGCTGCTGGAGCTGATCA
AGAAGACTTTCGATCTGGAGTGCAACATCTACCACAACCCGAGCGATGATACCCTGAAAT
TCAAGGTGTCTAACATTGAACAGATCGTGAACAAAATCATTCCATTCTTCGAGAAATATAC
CCTGTTTTCTCAAAAACGCGGTGACTTTATCCTGTTCTGCAAAGTGGTTGAACTGATCAAA
AACAAAGAACACCTGACTCTGAACGGCCTGATGAAAATTCTGTCCATCAAAGCCGCGATG
AACCTGGGTCTGTCCGAGAACCTGAAAAAGGAATTCCCGGGCTGTCTGAGCGTTAAACGT
CCAGAATTTGGTCTGAGCAATCTGAACAAACGTTGGCTGGCAGGCTTCATTGAAGGTGAA
GCCTGCTTTTTCGTTTCTATCTACAACTCCCCGAAGTCCAAACTGGGCAAGGCAGTACAGC
TGGTGTTTAAAATCACCCAGCATATCCGCGACAAAATCCTGATCGAATCCATTGTAGAACT
GCTGAACTGTGGTCGTGTTGAGGTTCGTAAATCTAACGAAGCGTGCGACTTCACCGTAACC
TCTATCAAAGAAATCGAAAACTACATCATCCCGTTCTTCAATGAATATCCGCTGATCGGCC
AGAAACTGAAAAACTACGAAGACTTCAAACTGATCTTCGACATGATGAAAACCAAAGATC
ACCTGACTGAAGAAGGCCTGTCTAAGATTATCGAAATTAAAAACAAAATGAACACCAACC
GTATTTAA

– 3’

A

MINLKNNIEYLNWYICGLVDAEGSFGVNVVKHATNKTGYAVLTYFELAMNSKDKQLLELIKK
TFDLECNIYHNPSDDTLKFKVSNIEQIVNKIIPFFEKYTLFSQKRGDFILFCKVVELIKNKEHLTL
NGLMKILSIKAAMNLGLSENLKKEFPGCLSVKRPEFGLSNLNKRWLAGFIEGEACFFVSIYNSP
KSKLGKAVQLVFKITQHIRDKILIESIVELLNCGRVEVRKSNEACDFTVTSIKEIENYIIPFFNEYP
LIGQKLKNYEDFKLIFDMMKTKDHLTEEGLSKIIEIKNKMNTNRI

B

Supplementary Figure 7.10. Amino acid sequence alignment of putative LHEases in members of *Leptographium*. In *L. truncatum* strain NFRI1813/1 and *L. lundbergii* strains DAOM60397, NFRI89-1040/1/3, and NFRI1502/1 insertions after the first LAGLIDADG motif appear to have split the ORF, in between the two motifs. The terms "edited" refer to the edited sequence in which the sequences for N- and C-terminal fragments were joined and amino acid residues that were not in alignment with closely-related taxa were replaced with gaps (-).

```
L. truncatum CBS929.85 (Lt.SSU/1)                          MINLKNNIEY LNWYICGLVD AEGSFGVNVV KHATNKTGYA VLTYFELAMN
L. terebrantis CBS337.70 (L.te.rnsi1)                      MINLKNNIEY LNWYICGLVD AEGSFGVNVV KHAANKTGYA VLTYFEIAMN
L. wingfieldii TOm1.3 (L.w.rnsi1)                          MINLKNNIEY LNWYICGLVD AEGSFGVNVV KHATNKTGYA VLTYFEIAMN
L. truncatum Forintek C34 (L.tr.rnsi1)                     MINLKNNIEY LNWYICGLVD AEGSFGVNVV KHATNKTGYA VLTYFELAMN
L. truncatum NFRI1813/1 (L.tr.rnsi1) Edited                MINLKNNIEY LNWYICGLVD AEGSFGVNVV KHATNKTGYA VLTYFELAMN
L. truncatum NFRI1813/1 (L.tr.rnsi1) N-terminus            MINLKNNIEY LNWYICGLVD AEGSFGVNVV KHATNKTGYA VLTYFELAMN
L. truncatum NFRI1813/1 (L.tr.rnsi1) C-terminus            ---------- ---------- ---------- ---------- ----------
L. lundbergii NFRI1502/1 (L.lu.rnsi1) Edited               MINLKNNIEY LNWYICGLVD AEGSFGVNVV KHATNKTGYA VLTYFEIAMN
L. lundbergii NFRI1502/1 (L.lu.rnsi1) N-terminus           MINLKNNIEY LNWYICGLVD AEGSFGVNVV KHATNKTGYA VLTYFEIAMN
L. lundbergii NFRI1502/1 (L.lu.rnsi1) C-terminus           ---------- ---------- ---------- ---------- ----------
L. lundbergiii DAOM60397 (L.lu.rnsi1) Edited               MINLKNNIEY LNWYICGLVD AEGSFGVNVV KHATNKTGYA VLTYFEIAMN
L. lundbergiii DAOM60397 (L.lu.rnsi1) N-terminus           MINLKNNIEY LNWYICGLVD AEGSFGVNVV KHATNKTGYA VLTYFEIAMN
L. lundbergiii DAOM60397 (L.lu.rnsi1) C-terminus           ---------- ---------- ---------- ---------- ----------
L. lundbergii NFRI89-1040/1/3 (L.lu.rnsi1) Edited          MINLKNNIEY LNWYICGLVD AEGSFGVNVV KHATNKTGYA VLTYFEIAMN
L. lundbergii NFRI89-1040/1/3 (L.lu.rnsi1) N-terminus      MINLKNNIEY LNWYICGLVD AEGSFGVNVV KHATNKTGYA VLTYFEIAMN
L. lundbergii NFRI89-1040/1/3 (L.lu.rnsi1) C-terminus      ---------- ---------- ---------- ---------- ----------

SKDKQLLELI KKTFDLECNI YHNPSDDTLK FKVSNIEQIV NKIIPFFEKY   TLFSQKRGDF ILFCKVVELI KNKEHLTLNG LMKILSIKAA MNLGLSENLK
SKDKQLLELI KKTIDLECNI YHNPSDDTLK FKVSNIEQIV NKIIPFFEKY   TLFSQKRGDF ILFCKVVELI KNKEHLTLNG LMKILSIKAA MNLGLSENLK
SKDKQLLELI KKTFDLECNI YHNPSDDTLK FKVSNIEQIV NKIIPFFEKY   TLFSQKRGDF ILFCKVVELI KNKEHLTLNG LMKILSIKAA MNLGLSENLK
SKDKQLLELI KKTFDLECNI YHNPSDDTLK FKVSNIEQIV NKIIPFFEKY   TLFSQKRGDF ILFCKVVELI KNKEHLTLNG LMKIISIKAA MNLGLSENLK
SKDKQLLELI KKTFDLECNI YHNPSDDTLK FKVSNIEQIV NKIIPFFEKY   TLFSQKRGDF ILFCKVVELI KNKEHLTLNG LMKILSIKAA MNLGLSEN--
SKDKQLLELI KKTFDLECNI YHNPSDDTLK FKVSNIEQIV NKIIPFFEKY   TLFSQKRGDF ILFCKVVELI KNKEHLTLNG LMKILSIKAA MNLGLSEN--
---------- ---------- ---------- ---------- ----------   ---------- ---------- ---------- ---------- ----------
SKDKQLLELI KKTFDLECNI YHNPSDDTLK FKVSNIEQIV NKIIPFFEKY   TLFS------ ---------- ---------- ---------- ---------K
SKDKQLLELI KKTFDLECNI YHNPSDDTLK FKVSNIEQIV NKIIPFFEKY   TLFSVFPLVF DPKGRPKAKK RGFYIIL--- ---------- ----------
---------- ---------- ---------- ---------- ----------   ---------- ---------- ---------- ---------- -------IRK
SKDKQLLELI KKTFDLECNI YHNPSDDTLK FKVSNIEQIV NKIIPFFEKY   TLFS------ ---------- ---------- ---------- ---------K
SKDKQLLELI KKTFDLECNI YHNPSDDTLK FKVSNIEQIV NKIIPFFEKY   TLFS------ ---------- ---------- ---------- ----------
---------- ---------- ---------- ---------- ----------   ---------- ---------- ---------- ---------- ---------K
SKDRQLLELI KKTFDLECNI YHNPSDDTLK FKVSNIEQIV NKIIPFFEKY   TLFS------ ---------- ---------- ---------- ---------K
SKDRQLLELI KKTFDLECNI YHNPSDDTLK FKVSNIEQIV NKIIPFFEKY   TLFS------ ---------- ---------- ---------- ----------
---------- ---------- ---------- ---------- ----------   ---------- ---------- ---------- ---------- ---------K

KEFPGCLSVK RPEFGLSNLN KRWLAGFIEG EACFFVSIYN SPKSKLGKAV   QLVFKITQHI RDKILIESIV ELLNCGRVEV RKSNEACDFT VTSIKEIENY
KEFPGCLLVK RPEFGLSNLN KRWLAGFIEG EACFFVSIYN SPKSKLGKAV   QLVFKITQHI RDKILIESIV ELLNCGRVEV RKSNEACDFT VTSIKEIENY
KEFPGCLLVK RPEFGLSNLN KRWLAGFIEG EACFFVSIYN SPKSKLGKAV   QLVFKITQHI RDKILIESIV ELLNCGRVEV RKSNEACDFT VTSIKEIENY
KEFPGCLSVK RPEFGLSNLN KRWLAGFIEG EACFFVSIYN SPKSKLGKAV   QLVFKITQRI RDKILIESIV ELLNCGRVEV RKSNEACDFT VTSIKEIENY
---------- ------SNLN KRWLAGFIEG EACFFVSIYN SPKSKLGKAV   QLVFKITQHI RDKILIESIV ELLNCGRVEV RKSNEACDFT VTSIKEIENY
---------- ---------- ---------- ---------- ----------   ---------- ---------- ---------- ---------- ----------
---------- -----MSNLN KRWLAGFIEG EACFFVSIYN SPKSKLGKAV   QLVFKITQHI RDKILIESIV ELLNCGRVEV RKSNEACDFT VTSIKEIENY
KEFPGCLSVK RPEFGLSNLN KRWLAGFIKG EACFFVSIYN SPKSKLGKAV   QLVFKITQHI RDKILIVSIV VLLNCGRVEV RESNEACDFT VTSIKEIENY
---------- ---------- ---------- ---------- ----------   ---------- ---------- ---------- ---------- ----------
KEFPGCLSVK RPEFGLSNLN KRWLAGFIKG EACFFVSIYN SPKSKLGKAV   QLVFKITQHI RDKILIVSIV VLLNCGRVEV RESNEACDFT VTSIKEIENY
KEFPGCLSVK RPEFGLSNLN KRWLAGFIEG EACFFVSIYN SPKSKLGKAV   QLVFKITQHI RDKILIESIV ELLNCGRVEV RKSNEACDFT VTSIKEIENY
---------- ---------- ---------- ---------- ----------   ---------- ---------- ---------- ---------- ----------
KEFPGCLSVK RPEFGLSNLN KRWLAGFIEG EACFFVSIYN SPKSKLGKAV   QLVFKITQHI RDKILIESIV ELLNCGRVEV RKSNEACDFT VTSIKEIENY
KEFPGCLSVK RPEFGLSNLN KRWLAGFIEG EACFFVSIYN SPKSKLGKAV   QLVFKITQHI RDKILIESIV ELLNCGRVEV RKSNEACDFT VTSIKEIENY
---------- ---------- ---------- ---------- ----------   ---------- ---------- ---------- ---------- ----------
KEFPGCLSVK RPEFGLSNLN KRWLAGFIEG EACFFVSIYN SPKSKLGKAV   QLVFKITQHI RDKILIESIV ELLNCGRVEV RKSNEACDFT VTSIKEIENY

IIPFFNEYPL IGQKLKNYED FKLIFDMMKT KDHLTEEGLS KIIEIKNKMN   TNRI-
IIPFFNEYPL IGQKLKNYED FKLIFDMMKT KDHLTEEGLS KIIEIKNKMN   TNRI-
IIPFFNEYPL IGQKLKNYED FKLIFDMMKT KDHLTEEGLS KIIEIKNKMN   TNRI-
IIPLFNEYPL IGQKLKNYED FKLIFDMMKT KDHLTEEGLS KIIEIKNKMN   TNRI-
IIPFFNEYPL IGQKLKNYED FKLIFDMMKT KDHLTEEGLS KIIEIKNKMN   TNRI-
---------- ---------- ---------- ---------- ----------   ----
IIPFFNEYPL IGQKLKNYED FKLIFDMMKT KDHLTEEGLS KIIEIKNKMN   TNRI-
IIPFFNEYPL IGQKLKNYED FKLIFDMMKT KDHLTEEGLS KIIEIKNKMN   TNRI-
---------- ---------- ---------- ---------- ----------   -----
IIPFFNEYPL IGQKLKNYED FKLIFDMMKT KDHLTEEGLS KIIEIKNKMN   TNRI-
IIPFFNEYPL IGQKLKNYED FKLIFDMMKT KDHLTEEGLS KIIEIKNKMN   TNRI-
---------- ---------- ---------- ---------- ----------   -----
IIPFFNEYPL IGQKLKNYED FKLIFDMMKT KDHLTEEGLS KIIEIKNKMN   TNRI-
IIPFFNEYPL IGQKLKNYED FKLIFDMMKT KDHLTEEGLS KIIEIKNKMN   TNKI-
---------- ---------- ---------- ---------- ----------   -----
IIPFFNEYPL IGQKLKNYED FKLIFDMMKT KDHLTEEGLS KIIEIKNKMN   TNKI-
```

REFERENCES

Aagaard C, Awayez MJ, Garret RA. 1997. Profile of the DNA recognition site of the archaeal homing endonuclease I-DmoI. *Nucleic Acids Res.* 25:1523-1530

Abeyrathne PD, Lavlev AI, Nazar RN. 2002. A RAC protein-binding site in the internal transcribed spacer 2 of Pre-rRNA transcripts from *Schizosaccharomyes pombe*. *J. Biol. Chem*. 277:21291–21299

Aguilar C, Sánchez JA. 2007. Phylogenetic hypotheses of gorgoniid octocorals according to ITS2 and their predicted RNA secondary structures. *Mol. Phylogenet. Evol*. 43:774-786

Bae H, Kim KP, Song JM, Kim JH, Yang JS, Kwon ST. 2009. Characterization of intein homing endonuclease encoded in the DNA polymerase gene of *Thermococcus marinus*. *FEMS Microbiol. Lett.* 297:180-188

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

Bassi GS, Weeks KM. 2003. Kinetic and thermodynamic framework for assembly of the six-component bI3 group I ribonucleoprotein catalyst. *Biochem.* 42:9980-9988

Beiggi S, Piercey-Normore MD. 2007. Evolution of ITS ribosomal RNA secondary structures in fungal and algal symbionts of selected species of *Cladonia* sect. *Cladonia* (Cladoniaceae, Ascomycotina). *J. Mol. Evol.* 64:528-542

Belfort M, Roberts RJ. 1997. Homing endonucleases: keeping the house in order. *Nucleic Acids Res.* 25:3379-3388

Belfort M, Derbyshire V, Parker MM, Cousineau B, Lambowitz AM. 2002. Mobile introns: pathways and proteins. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, eds. *Mobile DNA II*. Washington, DC: ASM Press. pp 761-783

Belfort M. 2003. Two for the price of one: a bifunctional intron-encoded DNA endonuclease-RNA maturase. *Genes Dev.* 17:2860-2863

Bell JA, Monteiro-Vitorello CB, Hausner G, Fulbright DW, Bertrand H. 1996. Physical and genetic map of the mitochondrial genome of *Cryphonectria parasitica* Ep155. *Curr. Genet.* 30:34-43

Benny GL, Kimbrough JW. 1980. A synopsis of the orders and families of Plectomycetes with keys to genera. *Mycotaxon* 12:1-91

Buchan A, Newell SY, Moreta JI, Moran MA. 2002. Analysis of internal transcribed spacer (ITS) regions of rRNA genes in fungal communities in a southeastern U.S. salt marsh. *Microb. Ecol.* 43:329-340

Bullerwell CE, Leigh J, Seif E, Longcore JE, Lang BF. 2003. Evolution of the fungi and their mitochondrial genomes. *Appl. Mycol. Biotechnol*. 3:133-159

Burke JM, RajBhandary UL. 1982. Intron within the large rRNA gene of N. crassa mitochondria: a long open reading frame and a consensus sequence possibly important in splicing. *Cell* 31:509-520

Burke JM, Esherick JS, Burfeind WR, King JL. 1990. A 3' splice site-binding sequence in the catalytic core of a group I intron. *Nature* 344:80-82

Burt A, Koufopanou V. 2004. Homing endonuclease genes: the rise and fall and rise again of a selfish element. *Curr. Opin. Genet. Dev.* 14:609-615

Cahan P, Kennell JC. 2005. Identification and distribution of sequences having similarity to mitochondrial plasmids in mitochondrial genomes of filamentous fungi. *Mol. Genet. Genomics* 273:462-473

Carbone I, Anderson JB, Kohn LM. 1995. A group-I intron in the mitochondrial small subunit ribosomal RNA gene of *Sclerotinia sclerotiorum*. *Curr. Genet.* 27:166-176

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

Cech TR, Bass BL. 1986. Biological catalysis by RNA. *Annu. Rev. Biochem.* 55:599-629

Cech TR. 1987. The chemistry of self-splicing RNA and RNA enzymes. *Science* 236:1532-1539

Cech TR. 1988. Conserved sequences and structures of group I introns: building an active site for RNA catalysis – a review. *Gene* 73:259-271

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

Chen Z, Zhao H. 2005. A highly sensitive selection method for directed evolution of homing endonucleases. *Nucleic Acids Res*. 33:e154

Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, Luo K, Li Y, Li X, Jia X, Lin Y, Leon C. 2010. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One* 5:e8613

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

Chevalier B, Monnat RJ Jr, Stoddard BL. 2005. The LAGLIDADG homing endonuclease family. In: Belfort M, Derbyshire V, Stoddard BL, Wood DL, eds. *Homing endonucleases and inteins*. New York, NY: Springer. pp 33-47

Clark CG. 1987. On the evolution of ribosomal RNA. *J. Mol. Evol*. 25:343–350

Clark-Walker GD. 1992. Evolution of mitochondrial genomes in fungi. *Int. Rev. Cytol.* 141:89-127

Coleman AW. 2003. ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends Genet*. 19:370-375

Coleman AW. 2007. Pan-eukaryote ITS2 homologies revealed by RNA secondary structure. *Nucleic Acids Res*. 35:3322-3329

Coleman AW. 2009. Is there a molecular key to the level of "biological species" in eukaryotes? A DNA guide. *Mol. Phylogenet. Evol.* 50:197-203

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

Copertino DW, Hall ET, Van Hook FW, Jenkins KP, Hallick RB. 1994. A group III twintron encoding a maturase-like gene excises through lariat intermediates. *Nucleic Acids Res.* 22:1029-1036

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

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

Costa M, Fontaine JM, Loiseaux-de Goër S, Michel F. 1997a. A group II self-splicing intron from the brown alga *Pylaiella littoralis* is active at unusually low

magnesium concentrations and forms populations of molecules with a uniform conformation. *J. Mol. Biol.* 274:353-364

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

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

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

Côté CA, Peculis BA. 2001. Role of the ITS2-proximal stem and evidence for indirect recognition of processing sites in pre-rRNA processing in yeast. *Nucleic Acids Res*. 29:2106-2116

Côté CA, Greer CL, Peculis BA. 2002. Dynamic conformational model for the role of ITS2 in pre-rRNA processing in yeast. *RNA* 8:786-797

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

Croll D, Giovannetti M, Koch AM, Sbrana C, Ehinger M, Lammers PJ, Sanders IR. 2009. Nonself vegetative fusion and genetic exchange in the arbuscular mycorrhizal fungus *Glomus intraradices*. *New Phytol*. 181:924-937

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. Weblogo: A sequence logo generator. *Genome Res.* 14:1188-1190

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

Cummings DJ, MacNeil IA, Domenico J, Matsuura ET. 1985. Excision-amplification of mitochondrial DNA during senescence in *Podospora anserina*. DNA sequence analysis of three unique 'plasmids'. *J. Mol. Biol.* 185:659-680

Cummings DJ, Michel F, McNally KL. 1989. DNA sequence analysis of the 24.5 kilobase pair cytochrome oxidase subunit I mitochondrial gene from *Podospora anserina*: a gene with sixteen introns. *Curr. Genet.* 16:381-406

Curcio MJ, Belfort M. 1996. Retrohoming: cDNA-mediated mobility of group II introns requires a catalytic RNA. *Cell* 84:9-12

Dai L, Zimmerly S. 2002a. The dispersal of five group II introns among natural populations of *Escherichia coli*. *RNA* 8:1294-1307

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

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

Dalgaard JZ, Garrett RA, Belfort M. 1993. A site-specific endonuclease encoded by a typical archaeal intron. *Proc. Natl. Acad. Sci. USA* 90:5414-5417

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.* 256:31-49

Doudna JA, Cech TR. 2002. The chemical repertoire of natural ribozymes. *Nature* 418:222-228

Druzhinina IS, Kopchinskiy AG, Komoń M, Bissett J, Szakacs G, Kubicek CP. 2005. An oligonucleotide barcode for species identification in *Trichoderma* and *Hypocrea*. *Fungal Genet. Biol.* 42:813-828

Dujon B. 1980. Sequence of the intron and flanking exons of the mitochondrial 21S rRNA gene of yeast strains having different alleles at the omega and rib-1 loci. *Cell* 20:185-197

Dujon B. 1989. Group I introns as mobile genetic elements: facts and mechanistic speculations – a review. *Gene* 82:91-114

Dujon B, Belcour L. 1989. Mitochondrial instabilities and rearrangements in yeasts and fungi. In: Berg DA, Howe MM, eds. *Mobile DNA*. Washington, DC: AMS Press. pp 861-878

Dürrenberger F, Rochaix JD. 1993. Characterization of the cleavage site and the recognition sequence of the I-*Cre*I DNA endonuclease encoded by the chloroplast ribosomal intron of *Chlamydomonas reinhardtii*. *Mol. Gen. Genet.* 236:409-414

Eastberg JH, McConnell Smith A, Zhao L, Ashworth J, Shen BW, Stoddard BL. 2007. Thermodynamics of DNA target site recognition by homing endonucleases. *Nucleic Acids Res.* 35:7209-7221

Edgell DR. 2009. Selfish DNA: homing endonucleases find a home. *Curr. Biol.* 19:R115-117

El Karkouri K, Murat C, Zampieri E, Bonfante P. 2007. Identification of internal transcribed spacer sequence motifs in truffles: a first step toward their DNA bar coding. *Appl. Environ. Microbiol*. 73:5320-5330

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

Felsenstein J. 2008. PHYLIP: Phylogeny Inference Package Version 3.68. Distributed by the author. Department of Genome Sciences and Department of Biology, University of Washington, Seattle, Wash. Available from *http://evolution.genetics.washington.edu/phylip/getme.html*

Felsenstein, J. 2002. *http://evolution.genetics.washington.edu/phylip/getme.html*

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

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

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

Fromont-Racine M, Senger B, Saveanu C, Fasiolo F. 2003. Ribosome assembly in eukaryotes. *Gene* 313:17-42

Gao K, Smith J, Yang M, Jones S, Djukanovic V, Nicholson MC, West A, Bidney D, Falco SC, Jantz D, Lyznik LA. 2010. Heritable targeted mutagenesis in maize using a designed endonuclease. *Plant J.* 61:176-87

Gibb EA, Hausner G. 2005. Optional mitochondrial introns and evidence for a homing-endonuclease gene in the mtDNA *rnl* gene in *Ophiostoma ulmi s.lat. Mycol. Res.* 109:1112-1126

Gimble FS. 2007. Engineering home endonucleases to modify complex genomes. *Gene Ther. Regul.* 3:33-50

Goddard MR, Burt A. 1999. Recurrent invasion and extinction of a selfish gene. *Proc. Natl. Acad. Sci. USA* 96:13880-13885

Goertzen LR, Cannone JJ, Gutell RR, Jansen RK. 2003. ITS secondary structure derived from comparative analysis: implications for sequence alignment and phylogeny of the Asteraceae. *Mol. Phylogenet. Evol.* 29:216-234

Goloboff PA, Farris JS, Nixon KC. 2008. TNT, a free program for phylogenetic analysis. *Cladistics* 24:774-786

Good L, Intine RV, Nazar RN. 1997. Interdependence in the processing of ribosomal RNAs in *Schizosaccharomyces pombe*. *J. Mol. Biol.* 273:782-788

Gorodkin J, Heyer LJ, Brunak S, Stormo GD. 1997. Displaying the information contents of structural RNA alignments: the structure logos. *Comput. Appl. Biosci.* 13:583-586

Gottschling M, Plötner J. 2004. Secondary structure models of the nuclear internal transcribed spacer regions and 5.8S rRNAin Calciodinelloideae (Peridiniaceae) and other dinoflagellates. *Nucleic Acids Res.* 32:307-315

Gray MW, Burger G, Lang BF. 2001. The origin and early evolution of mitochondria. *Genome Biol.* 2:1018.1-1018.5

Griffiths AJ. 1995. Natural plasmids of filamentous fungi. *Microbiol. Rev.* 59:673-685

Grivell LA. 1995. Nucleo-mitochondrial interactions in mitochondrial gene expression. *Crit. Rev. Biochem. Mol. Biol.*30:121-164

Grizot S, Smith J, Daboussi F, Prieto J, Redondo P, Merino N, Villate M, Thomas S, Lemaire L, Montoya G, Blanco FJ, Pâques F, Duchateau P. 2009. Efficient targeting of a SCID gene by an engineered single-chain homing endonuclease. *Nucleic Acids Res.*16:5405-5419

Guo H, Karberg M, Long M, Jones JP 3rd, Sullenger B, Lambowitz AM. 2000. Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science* 289:452-457

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser*. 41:95-98

Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A, Stutz E. 1993. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Res.* 21:3537-3544

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

Hancock JM, Dover G A. 1990. 'Compensatory slippage' in the evolution of ribosomal RNA genes. *Nucleic Acids Res.* 18:5949-5954

Harpke D, Peterson A. 2006. Non-concerted ITS evolution in *Mammillaria* (Cactaceae). *Mol. Phylogenet. Evol*. 41:579-593

Harrington TC. 1988. *Leptographium* species, their distributions, hosts and insect vectors. In: Harrington TC, Cobb FW Jr, eds. Leptographium *root diseases on conifers.* St. Paul, Minnesota: APS Press

Harrington TC. 1993. Biology and taxonomy of fungi associated with bark beetles. In: Schowalter TD, Filip GM, eds. *Beetle-Pathogen Interactions in Conifer Forests*. London: Academic Press. pp 37-58

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22:160-174

Haugen P, Bhattacharya D. 2004. The spread of LAGLIDADG homing endonuclease genes in rDNA. *Nucleic Acids Res.* 32:2049-2057

Haugen P, Wikmark OG, Vader A, Coucheron DA, Sjøttem, Johansen SD. 2005. The recent transfer of a homing endonuclease gene. *Nucleic Acids Res.* 33:2734-2741

Hausner G, Reid J, Klassen GR. 1992. Do galeate-ascospore members of the *Cephaloascaceae, Endomycetaceae* and *Ophiostomataceae* share a common phylogeny? *Mycologia* 84:870-881

Hausner G, Klassen GR, Reid J. 1993a. Unusually compact ribosomal RNA gene cluster in *Sphaeronaemella fimicola*. *Curr. Genet.* 23:357-359

Hausner G, Reid J, Klassen GR. 1993b. On the phylogeny of *Ophiostoma*, *Ceratocystis* s.s., *Microascus*, and relationships within *Ophiostoma* based on partial ribosomal DNA sequences. *Can. J. Bot.* 71:1249-1265

Hausner G, Reid J, Klassen GR. 2000. On the phylogeny of members of *Ceratocystis* s.s. and *Ophiostoma* that possess different anamorphic states, with emphasis on the anamorph genus *Leptographium*, based on partial ribosomal DNA sequences. *Can. J. Bot*. 78:903-916

Hausner G. 2003. Fungal mitochondrial genomes, plasmids and introns. *Appl. Mycol. Biotechnol.* 3:101-131

Hausner G, Wang X. 2005. Unusual compact rDNA gene arrangements within some members of the Ascomycota: evidence for molecular co-evolution between ITS1 and ITS2. *Genome* 48:648-660

Hausner G, Iranpour M, Kim JJ, Breuil C, Davis CN, Gibb EA, Reid J, Loewen PC, Hopkin AA. 2005. Fungi vectored by the introduced bark beetle *Tomicus piniperda* in Ontario, Canada, and comments on the taxonomy of *Leptographium lundbergii*, *Leptographium terebrantis*, *Leptographium truncatum*, and *Leptographium wingfieldii*. *Can. J. Bot.* 83:1222-1237

Hausner G, Nummy KA, Bertrand H. 2006a. Asexual transmission and meiotic extinction of small plasmid-like derivatives of the mitochondrial DNA in *Neurospora crassa*. *Fung. Genet. Biol.* 43:90-101

Hausner G, Olson R, Simon D, Johnson I, Sanders ER, Karol KG, McCourt RM, Zimmerly, S. 2006b. Origin and evolution of the chloroplast *trnK* (*matK*) intron: a model for evolution of group II intron RNA structures. *Mol. Biol. Evol.* 23:380-391

Heath PJ, Stephens KM, Monnat RJ Jr, Stoddard BL. 1997. The structure of I-*Cre*I, a group I intron-encoded homing endonuclease. *Nat. Struct. biol.* 4:468-476

Heringa J. 1999. Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Comput. Chem.* 23:341-364

Heringa J. 2000. Computational methods for protein secondary structure prediction using multiple sequence alignments. *Curr. Protein Pept. Sci.* 1:273-301

Heringa J. 2002. Local weighting schemes for protein multiple sequence alignment. *Comput. Chem.* 26:459-477

Hershkovitz MA, Zimmer EA, Hans WJ. 1999. Ribosomal DNA sequences and angiosperm systematics. In: Hollingsworth PM, Bateman RM, Gornall RJ,

eds. *Molecular systematics and plant evolution*. London: Taylor and Francis. pp 268-326

Hibbett DS, Binder M, Bischoff JF, Blackwell M, Cannon PF, Eriksson OE, Huhndorf S, James T, Kirk PM, Lücking R, Thorsten Lumbsch H, Lutzoni F, Matheny PB, McLaughlin DJ, Powell MJ, Redhead S, Schoch CL, Spatafora JW, Stalpers JA, Vilgalys R, Aime MC, Aptroot A, Bauer R, Begerow D, Benny GL, Castlebury LA, Crous PW, Dai YC, Gams W, Geiser DM, Griffith GW, Gueidan C, Hawksworth DL, Hestmark G, Hosaka K, Humber RA, Hyde KD, Ironside JE, Kõljalg U, Kurtzman CP, Larsson KH, Lichtwardt R, Longcore J, Miadlikowska J, Miller A, Moncalvo JM, Mozley-Standridge S, Oberwinkler F, Parmasto E, Reeb V, Rogers JD, Roux C, Ryvarden L, Sampaio JP, Schüssler A, Sugiyama J, Thorn RG, Tibell L, Untereiner WA, Walker C, Wang Z, Weir A, Weiss M, White MM, Winka K, Yao YJ, Zhang N. 2007. A higher-level phylogenetic classification of the fungi. *Mycol. Res*. 111:509-547

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

Ho Y, Waring RB. 1999. The maturase encoded by a group I intron from *Aspergillus nidulans* stabilizes RNA tertiary structure and promotes rapid splicing. *J. Mol. Biol.* 292:987-1001

Horton TR. 2002. Molecular approaches to ectomycorrhizal diversity studies: variations in ITS at a local scale. *Plant Soil* 244:29-39

Huang HR, Rowe CE, Mohr S, Jiang Y, Lambowitz AM, Perlman PS. 2005. The splicing of yeast mitochondrial group I and group II introns requires a DEAD-

box protein with RNA chaperone function. *Proc. Natl. Acad. Sci. USA* 102:163-168

Index Fungorum. www.indexfungorum.org, CABI Bioscience, CBS, Landcare Research. Accessed December 27, 2009; January 5, 2010

Jacobs K, Wingfield MJ. 2001. *Leptographium* species: tree pathogens, insect associates and agents of blue-stain. St. Paul, Minnesota: APS Press

Jacobs K, Wingfield MJ, Wingfield BD. 2001. Phylogenetic relationships in *Leptographium* based on morphological and molecular characters. *Can. J. Bot.* 79:719-732

Jacobs K, Bergdahl DR, Wingfield MJ, Halik S, Seifert KA, Bright DE, Wingfield BD. 2004. *Leptographium wingfieldii* introduced into North America and found associated with exotic *Tomicus piniperda* and native bark beetles. *Mycol. Res*. 108:411-418

Jacobs K, Solheim H, Wingfield BD, Wingfield MJ. 2005. Taxonomic re-evalutation of *Leptographium lundbergii* based on DNA sequence comparisons and morphology. *Mycol. Res.*109:1149-1161

Jacquier A and Rosbash M. 1986. Efficient trans-splicing of a yeast mitochondrial RNA group II intron implicates a strong 5' exon-intron interaction. *Science* 234:1099-1104

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

Jaeger L. 1997. The new world of ribozymes. *Curr. Opin. Struct. Biol.* 7:324-335

Jarrell KA, Peebles CL, Dietrich RC, Romiti SL, Perlman PS. 1988. Group II intron self-splicing. Alternative reaction conditions yield novel products. *J. Biol. Chem.* 263:3432-3439

Jobst J, King K, Hemleben V. 1998. Molecular evolution of the internal transcribed spacers (ITS1 and ITS2) and phylogenetic relationships among species of the family Cucurbitaceae. *Mol. Phylogenet. Evol*. 9:204-219

Jones JP 3rd, Kierlin NM, Coon RG, Perutka J, Lambowitz AM, Sullenger BA. 2005. Retargeting mobile group II introns to repair mutant genes. *Mol. Ther*. 11:687-694

Joseph N, Krauskopf E, Vera MI, Michot B. 1999. Ribosomal internal transcribed spacer 2 (ITS2) exhibits a common core of secondary structure in vertebrates and yeast. *Nucleic Acids Res*. 27:4533-4540

Jurica MS, Monnat RJ Jr, Stoddard BL. 1998. DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-*Cre*I. *Mol. Cell* 2:469-476

Jurica MS, Stoddard BL. 1999. Homing endonucleases: structure, function and evolution. *Cell Mol. Life Sci.* 55:1304-1326

Kamikawa R, Masuda I, Demura M, Oyama K, Yoshimatsu S, Kawachi M, Sako Y. 2009. Mitochondrial group II introns in the Raphidophycean flagellate *Chattonella* spp. suggest a diatom-to-*Chattonella* lateral group II intron transfer. *Protist* 160:364-375

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

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

Keeling PJ, Fast NM. 2002. Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annu. Rev. Microbiol.* 56 :93-116

Keller A, Schleicher T, Schultz J, Müller T, Dandekar T, Wolf, M. 2009. 5.8S-28S rRNA interaction and HMM-based ITS2 annotation. *Gene* 430:50-57

Kennedy N, Clipson N. 2003. Fingerprinting the fungal community. *Mycologist* 17:158-164

Kennell JC, Lambowitz AM. 1989. Development of an in vitro transcription system for *Neurospora crassa* mitochondrial DNA and identification of transcription initiation sites. *Mol. Cell Biol.* 9:3603-3613

Khan AU, Lal SK. 2003. Ribozymes: a modern tool in medicine. *J. Biomed. Sci.* 10:457-467

Kim SH, Cech TR. 1987. Three-dimensional model of the active site of the self-splicing rRNA precursor of *Tetrahymena*. *Proc. Natl. Acad. Sci. USA* 84:8788-8792

Kim WK, Mauthe W, Hausner G, Klassen GR. 1990. Isolation of high molecular weight DNA and double-stranded RNAs from fungi. *Can. J. Bot*. 68:1898-1902

Klepzig KD, Six DL. 2004. Bark beetle-fungal symbiosis: context dependency in complex associations. *Symbiosis* 37:189-205

Koetschan C, Förster F, Keller A, Schleicher T, Ruderisch B, Schwarz R, Müller T, Wolf M, Schultzz J. 2010. The ITS2 Database III – sequences and structures for phylogeny. *Nucleic Acids Res.* D275-D279

Krüger D, Gargas A. 2008. Secondary structure of ITS2 rRNA provides taxonomic characters for systematic studies – a case in Lycoperdaceae (Basidiomycota). *Mycol. Res.* 112:316-330

Lafontaine D, Tollervey D. 1995. *Trans*-acting factors in yeast pre-rRNA and pre-snoRNA processing. *Biochem. Cell Biol.* 73:803-812

Lafontaine DL. 2004. Eukaryotic ribosome synthesis. In: Nierhaus K, ed. *Protein synthesis and ribosome structure*. Wiley-InterScience. pp 107-143

Lafontaine DL. 2010. A 'garbage can' for ribosomes: how eurkaryotes degrade their ribosomes. *Trends Biochem. Sci*. 35:267-277

Lagerberg T, Lundberg G, Melin E. 1927. Biological and practical researches into blueing in pine and spruce. *Sv. Skogsvardsf. Tidskr.* 25:145-272

Lalev AI, Nazar RN. 1998. Conserved core structure in the internal transcribed spacer 1 of the *Schizosaccharomyces pombe* precursor ribosomal RNA. *J. Mol. Biol*. 284:1341-1351

Lalev AI, Nazar RN. 1999. Structural equivalence in the transcribed spacers of pre-rRNA transcripts in *Schizosaccharomyces pombe*. *Nucleic Acids Res*. 27:3071-3078

Lalev AI, Abeyrathne PD, Nazar RN. 2000. Ribosomal RNA maturation in *Schizosaccharomyces pombe* is dependent on a large ribonucleoprotein complex of the internal transcribed spacer 1. *J. Mol. Biol.* 302:65-77

Lalev AI, Nazar RN. 2001. A chaperone for ribosome maturation. *J. Biol. Chem.* 276:16655-16659

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

Lambowitz AM, Caprara MG, Zimmerly S, Perlman PS. 1999. Group I and group II ribozymes as RNPs: clues to the past and guides to the future. In: Gesteland RF, Cech TR, Atkins JF, eds. *The RNA World*. New York, NY: Cold Spring Harbor Laboratory Press. pp 451-485

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

Landis FC, Gargas A. 2007. Using ITS2 secondary structure to create species-specific oligonucleotide probes for fungi. *Mycologia* 99:681-692

Lang BF. 1984. The mitochondrial genome of the fission yeast *Schizosaccharomyces pombe*: highly homologous introns are inserted at the same position of the otherwise less conserved *cox1* genes in *Schizosaccharomyces pombe* and *Aspergillus nidulans*. *EMBO J.* 3:2129-2136

Lang BF, Ahne F, Bonen L. 1985. The mitochondrial genome of the fission yeast *Schizosaccharomyces pombe*. The cytochrome *b* gene has an intron closely related to the first two introns in the *Saccharomyces cerevisiae cox1* gene. *J. Mol. Biol.* 184:353-366

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947-2948

Lazowska J, Claisse M, Gargouri A, Kotylak Z, Spyridakis A, Slonimski PP. 1989. Protein encoded by the third intron of *cytochrome b* gene in *Saccharomyces cerevisiae* is an mRNA maturase. Analysis of mitochondrial mutants, RNA transcripts, proteins and evolutionary relationships. *J. Mol. Biol.* 205:275-289

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

Levinson G, Gutman GA. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4:203-221

Lieckfeldt E, Seifert KA. 2000. An evaluation of the use of ITS sequences in the taxonomy of the Hypocreales. *Studies in Mycol*. 45:35-44

Loizos N, Tillier ER, Belfort M. 1994. Evolution of mobile group I introns: recognition of intron sequences by an intron-encoded endonuclease. *Proc. Natl. Acad. Sci USA* 91:11983-11987

Longo A, Leonard CW, Bassi GS, Berndt D, Krahn JM, Hall TM, Weeks KM. 2005. Evolution from DNA to RNA recognition by the bI3 LAGLIDADG maturase. *Nat. Struc. Mol. Biol.* 12:779-787

Lu Q, Decock C, Zhang XY, Maraite H. 2009. Ophiostomatoid fungi (Ascomycota) associated with *Pinus tabuliformis* infested by *Dendroctonus valens* (Coleoptera) in northern China and an assessment of their pathogenicity on mature trees. *Antonie Leeuwenhoek* 96:275-293

Lucas P, Otis C, Mercier JP, Turmel M, Lemieux C. 2001. Rapid evolution of the DNA-binding site in LAGLIDADG homing endonucleases. *Nucleic Acids Res.* 29:960-969

Mai JC, Coleman AW. 1997. The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. *J. Mol. Evol*. 44:258-271

Marcaida MJ, Muñoz IG, Blanco FJ, Prieto J, Montoya G. 2010. Homing endonucleases: from basics to therapeutic applications. *Cell. Mol. Life Sci.* 67:727-748

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

Martínez-Abarca F, Zekri S, Toro N. 1998. Characterization and splicing *in vivo* of a *Sinorhizobium meliloti* group II intron associated with particular insertion sequences of the IS630-Tc1/IS3 retroposon superfamily. *Mol. Microbiol.* 28:1295-1306

Martínez-Abarca F, Barrientos-Durán A, Fernández-López 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.* 32:2880-2888

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

Mathews DH, Sabina J, Zucker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288:911-940

Matsuura M, Saldanha R, Ma H, Wank H, Yang J, Mohr G, Cavanagh S, Dunny GM, Belfort M, Lambowitz AM. 1997. A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes Dev.* 11:2910-2924

McConnell Smith A, Takeuchi R, Pellenz S, Davis L, Maizels N, Monnat RL Jr, Stoddard BL. 2009. Generation of a nicking enzyme that stimulates site-specific gene conversion from the I-AniI LAGLIDADG homing endonuclease. *Proc. Natl. Acad. Sci. USA* 106:5099-5104

Meyer IM, Miklós I. 2007. SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput. Biol.* 3:e149

Michel F, Jacquier A, Dujon B. 1982. Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. *Biochimie* 64:867-881

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

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

Michel F, Westhof E. 1990. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* 216:585-610

Michel F, Netter P, Xu MQ, Shub DA. 1990. Mechanism of 3' splice site selection by the catalytic core of the sunY intron of bacteriophage T4: the role of a novel base-pairing interaction in group I introns. *Gene Dev.* 4:777-788

Michel F, Jaeger L, Westhof E, Kuras R, Tihy F, Xu MQ, Shub DA. 1992. Activation of the catalytic core of a group I intron by a remote 3' splice junction. *Genes Dev.* 6:1373-1385

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

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

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

Mitchell JI, Zuccaro A. 2006. Sequences, the environment and fungi. *Mycologist* 20:62-74

Mohr S, Matsuura M, Perlman PS, Lambowitz AM. 2006. A DEAD-box protein alone promotes group II intron splicing and reverse splicing by acting as an RNA chaperone. *Proc. Natl. Acad. Sci.USA* 103:3569-3574

Monteiro-Vitorello CB, Baidyaroy D, Bell JA, Hausner G, Fullbright DW, Bertrand H. 2000. A circular mitochondrial plasmid incites cytoplasmically-transmissible hypovirulence in some strains of *Cryphonectria parasitica*. *Curr. Genet.* 37:242-256

Monteiro-Vitorello CB, Hausner G, Searles DB, Gibb EA, Fulbright DW, Bertrand H. 2009. The *Cryphonectria parasitica* mitochondrial *rns* gene: plasmid-like elements, introns and homing endonucleases. *Fungal Genet. Biol.* 11:837-848

Moure CM, Gimble FS, Quiocho FA. 2002. Crystal structure of the intein homing endonuclease PI-*Sce*I bound to its recognition sequence. *Nat. Struct. Biol.* 9:764-770

Moure CM, Gimble FS, Quiocho FA. 2003. The crystal structure of the gene targeting homing endonuclease I-*Sce*I reveals the origins of its target site specificity. *J. Mol. Biol.* 334:685-695

Müller T, Vingron M. 2000. Modelling amino acid replacement. *J. Comput. Biol.* 7:761-776

Müller T, Philippi N, DandekarT, Schultz J, Wolf M. 2007. Distinguishing Species. *RNA* 13:1469-1472

Mullineux T, Hausner G. 2009. Evolution of rDNA ITS1 and ITS2 sequences and RNA secondary structures within members of the fungal genera *Grosmannia* and *Leptographium*. *Fungal Genet. Biol.* 46:855-867

Mullineux ST, Costa M, Bassi GS, Michel F, Hausner G. 2010. A group II intron encodes a functional LAGLIDADG homing endonuclease and self-splices under moderate temperature and ionic conditions. *RNA* (In press, RNA/2010/021840)

Nazar RN. 1980. A 5.8 S rRNA-like sequence in prokaryotic 23 S rRNA. *FEBS Lett.* 119:212-214

Nazar RN, Wong WM, Abrahamson JL. 1987. Nucleotide sequence of the 18-25 S ribosomal RNA intergenic region from a thermophile, *Thermomyces lanuginosus*. *J. Biol. Chem.* 262:7523-7527

Nazar RN. 2003. Ribosome biogenesis in yeast: rRNA processing and quality control. *Appl. Mycol. Biotechnol.* 3:161–183

Nazar RN. 2004. Ribosomal RNA processing and ribosome biogenesis in eukaryotes *IUBMB Life* 56:457–465

Nazari R, Joshi S. 2008. Exploring the potential of group II introns to inactivate human immunodeficiency virus type 1. *J. Gen. Virol.* 89:2605-2610

Nicholas KB, Nicholas HB Jr, Deerfield DW II. 1997. GeneDoc: analysis and visualization of genetic variation. *EMBNEW.NEWS* 4:14

Niemer I, Schmelzer C, Börner GV. 1995. Overexpression of DEAD box protein pMSS116 promotes ATP-dependent splicing of a yeast group II intron *in vitro.* *Nucleic Acids Res.* 23:2966-2972

Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, Larsson KH. 2008. Intraspecific ITS variability in the kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evol. Bioinfo.* 4:193-201

Nilsson RH, Bok G, Ryberg M, Kristiansson E, Hallenberg N. 2009. A software pipeline for processing and identification of fungal ITS sequences. *Source Code Biol. Med*. 4:1

Okamoto K, Shaw JM. 2005. Mitochondrial morphology and dynamics in yeast and multicellular eukaryotes. *Annu. Rev. Genet.* 39:503-536

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

Page RD. 1996. TREEVIEW: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* 12:357-358

Paquin B, Laforest MJ, Forget L, Roewer I, Wang Z, Longcore J, Lang BF. 1997. The fungal mitochondrial genome project: evolution of fungal mitochondrial genomes and their gene expression. *Curr. Genet.* 31:380-395

Patel SB, Bellini M. 2008. The assembly of a spliceosomal small nuclear ribonucleoprotein particle. *Nucleic Acids Res.* 36:6482-6493

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

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

Piercey-Normore MD, Coxson D, Goward T, Goffinet B. 2006. Phylogenetic position of a Pacific Northwest North American endemic cyanolichen, *Nephroma occultum* (Ascomycota, *Peltigerales*). *The Lichenologist* 38:441-456

Pöggeler S, Kempken F. 2004. Mobile genetic elements in mycelial fungi. In: Kück U ed. *The Mycota II*. Berlin-Heidelberg: Springer-Verlag. pp 165-197

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817-818

Pramateftaki PV, Kouvelis VN, Lanaridis P, Typas MA. 2008. Complete mitochondrial genome sequence of the wine yeast *Candida zemplinina*: intraspecies distribution of a novel group-IIB1 intron with eubacterial affiliations. *FEMS Yeast Res* 8: 311-327

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

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

Ranjard L, Poly F, Lata JC, Mougel C, Thioulouse J, Nazaret S. 2001. Characterization of bacterial and fungal soil communities by automated ribosomal intergenic spacer analysis fingerprints: biological and methodological variability. *Appl. Environ. Microbiol.* 67:4479–4487

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

Robart AR, Seo W, Zimmerly S. 2007. Insertion of group II intron retroelements after intrinsic transcriptional terminators. *Proc. Natl. Acad. Sci. USA* 104:6620-6625

Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM, Degtyarev SKh, Dryden DT, Dybvig K, Firman K, Gromova ES, Gumport RI, Halford SE, Hattman S, Heitman J, Hornby DP, Janulaitis A, Jeltsch A, Josephsen J, Kiss A, Klaenhammer TR, Kobayashi I, Kong H, Krüger DH, Lacks S, Marinus MG, Miyahara M, Morgan RD, Murray NE, Nagaraja V, Piekarowicz A, Pingoud A, Raleigh E, Rao DN, Reich N, Repin VE, Selker EU, Shaw PC, Stein DC, Stoddard BL, Szybalski W, Trautner TA, Van Etten JL, Vitor JM, Wilson GG, Xu SY. 2003. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* 31:1805-1812

Rochaix JD, Rahire M, Michel F. 1985. The chloroplast ribosomal intron of *Chlamydomonas reinhardii* codes for a polypeptide related to mitochondrial maturases. *Nucleic Acids Res.* 13:975-984

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

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572-1574

Ryberg M, Nilsson RH, Kristiansson E, Töpel M, Jacobsson S, Larsson E. 2008. Mining metadata from unidentified ITS sequences in GenBank: A case study in *Inocybe* (Basidiomycota). *BMC Evol. Biol.* 8:50

Saldanha R, Mohr G, Belfort M, Lambowitz AM. 1993. Group I and group II introns. *FASEB J.* 7:15-24

Santamaria M, Vicario S, Pappadà G, Scioscia G, Scazzocchio C, Saccone C. 2009. Towards barcode markers in Fungi: an intron map of Ascomycota mitochondria. *BMC Bioinformatics* 10 (Suppl 6):S15

Schäfer B. 2003. Genetic conservation versus variability in mitochondria: the architecture of the mitochondrial genome in the petite-negative yeast *Schizosaccharomyces pombe*. *Curr. Genet.* 43:311-326

Schäfer B, Gan L, Perlman PS. 2003. Reverse transcriptase and reverse splicing activities encoded by the mobile group II intron cobI1 of fission yeast mitochondrial DNA. *J. Mol. Biol.* 329:191-206

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

Schmidt U, Riederer B, Mörl M, Schmelzer C, Stahl U. 1990. Self-splicing of the mobile group II intron of the filamentous fungus *Podospora anserina* (CO1 I1) *in vitro*. *EMBO J. 9*:2289-2298

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502-504

Schneider TD, Stephens RM. 1990. Sequence logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.* 18:6097-6100

Schultz J, Maisel S, Gerlach D, Müller T, Wolf M. 2005. A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *RNA* 11:361-364

Schultz J, Müller T, Achtziger M, Seibel PN, Dandekar T, Wolf M. 2006. The internal transcribed spacer 2 database-a web server for (not only) low level phylogenetic analyses. *Nucleic Acids Res.* 34:W704-707

Schultz J, Wolf M. 2009. ITS2 sequence-structure analysis in phylogenetics: a how-to manual for molecular systematics. *Mol. Phylogenet. Evol.* 52:520-523

Scott WG. 2007. Ribozymes. *Curr. Opin. Struc. Biol.* 17:280-286

Seibel PN, Müller T, Dandekar T, Wolf M. 2008. Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE. *BMC Res. Notes* 1:91

Seifert KA, Samson RA, Dewaard JR, Houbraken J, Lévesque CA, Moncalvo JM, Louis-Seize G, Hebert PD. 2007. Prospects for fungus identification using *CO1* DNA barcodes, with *Penicillium* as a test case. *Proc. Natl. Acad. Sci. USA* 104:3901-3906

Selig C, Wolf M, Müller T, Dandekar T, Schultz J. 2008. The ITS2 Database II: homology modelling RNA structure for molecular systematics. *Nucleic Acids Res.* 36:D377-380

Sethuraman J, Okoli CV, Majer A, Corkery TLC, Hausner G. 2008. The sporadic occurence of a group I intron-like element in the mtDNA *rnl* gene of *Ophiostoma novo-ulmi* subsp. *americana*. *Mycol. Res.* 112:564-582

Sethuraman J, Majer A, Friedrich NC, Edgell DR, Hausner G. 2009a. Genes within genes: multiple LADLIDADG homing endonucleases target the ribosomal protein S3 gene encoded within an *rnl* group I intron of *Ophiostoma* and related taxa. *Mol. Biol. Evol.* 26:2299-2315

Sethuraman J, Majer A, Iranpour M, Hausner G. 2009b. Molecular evolution of the mtDNA encoded *rps3* gene among filamentous ascomycetes fungi with an emphasis on the Ophiostomatoid fungi. *J. Mol. Evol.* 69:372-385

Shao L, Hu S, Yang Y, Gu Y, Chen J, Yang Y, Jiang W, Yang S. 2007. Targeted gene disruption by use of a group II intron (targetron) vector in *Clostridium acetobutylicum. Cell Res*. 17:963-965

Shub DA, Gott JM, Xu MQ, Lang BF, Michel F, Tomaschewski J, Pedersen-Lane J, Belfort M. 1988. Structural conservation among three homologous introns of bacteriophage T4 and the group I introns of eukaryotes. *Proc. Natl. Acad. Sci. USA* 85:1151-1155

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* 14:1404-1713

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

Simossis VA, Heringa J. 2003. The PRALINE online server: optimizing progressive multiple alignment on the web. *Comput. Biol. Chem.* 27:511-519

Simossis VA, Heringa J. 2005. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res.* 33:W289-W294

Singh P, Tripathi P, Silva GA, Pingoud A, Muniyappa K. 2009. Characterization of *Mycobacterium leprae* RecA intein, a LAGLIDADG homing endonuclease, reveals a unique mode of DNA binding, helical distortion, and cleavage compared with a canonical LAGLIDADG homing endonuclease. *J. Biol. Chem.* 284:25912-25928

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

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

Stoddard BL. 2006. Homing endonuclease structure and function. *Q. Revs. Biophys.* 38:49-95

Swofford DL. 2002. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Massachusetts: Sinauer Associates

Thompson JD, Gibson TJ, Plewniak G, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25:4876-4882

Thyme SB, Jarjour J, Takeuchi R, Havranek JJ, Ashworth J, Scharenberg AM, Stoddard BL, Baker D. 2009. Exploitation of binding energy for catalysis and design. *Nature* 461:1300-1304

Tinline RD, MacNeill BH. 1969. Parasexuality in plant pathogenic fungi. *Annu. Rev. Phytopathol.* 7:147-168

Toor N, Hausner G, Zimmerly S. 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA* 7:1141-1152

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

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

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

Torres RA, Ganal M, Hemleben V. 1990. GC balance in the internal transcribed spacers ITS 1 and ITS 2 of nuclear ribosomal RNA genes. *J. Mol. Evol.* 30:170-181

Traver BE, Anderson MAE, Adelman ZN. 2009. Homing endonucleases catalyze double-stranded DNA breaks and somatic transgene excision in *Aedes aegypti*. *Insect Mol. Biol.* 18:623-633

Upadhyay HP. 1981. Monograph of *Ceratocystis* and *Ceratocystiopsis*. University of Georgia Press, GA.

Valadkhan S. 2007. The spliceosome: a ribozyme at heart? *Biol. Chem.* 388:693-697

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

van der Sande CA, Kwa M, van Nues RW, van Heerikhuizen H, Raué HA, Planta RJ. 1992. Functional analysis of internal transcribed spacer 2 of *Saccharomyces cerevisiae* ribosomal DNA. *J. Mol. Biol.* 223:899-910

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

van der Veen R, Arnberg AC, Grivell LA. 1987. Self-splicing of a group II intron in yeast mitochondria: dependence on 5' exon sequences. *EMBO J.* 6:1079-1084

van Nues RW, Venema J, Rientjes JMJ, Dirks-Mulder A, Raué HA. 1995. Processing of eukaryotic pre-rRNA: the role of the transcribed spacers. *Biochem. Cell Biol.* 73:789-801

van Oppen MJ, Catmull J, McDonald BJ, Hislop NR, Hagerman PJ, Miller D. 2000. The mitochondrial genome of *Acropora tenuis* (Cnidaria; Scleractinia) contains a large group I intron and a candidate control region. *J. Mol. Evol.* 55:1-13

Vicens Q and Cech TR. 2006. Atomic level architecture of group I introns revealed. *Trends Biochem. Sci*. 31: 41-51

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

Weeks KM. 1997. Protein-facilitated RNA folding. *Curr. Opin. Struct. Biol.* 7:336-342

Wery M, Ruidant S, Schillewaert S, Leporé N, Lafontaine DL. 2009. The nuclear poly(A) polymerase and exosome cofactor Trf5 is recruited cotranscriptionally to nucleolar surveillance. *RNA* 15:406-419

Westhof E, Michel F. 1998. Ribozyme architectural diversity made visible. *Science* 282:251-252

Windbichler N, Papathanos PA, Catteruccia F, Ranson H, Burt A, Crisanti A. 2007. Homing endonuclease mediated gene targeting in Anopheles gambiae cells and embryos. *Nucleic Acids Res.* 35:5922-5933

Wingfield MJ. 1993. *Leptographium* species as anamorphs of *Ophiostoma*: progress in establishing acceptable generic and species concepts. In: Wingfield MJ, Seifert KA, Webber JF, eds. Ceratocystis and Ophiostoma: *taxonomy, ecology and pathogenecity.* St. Paul, Minnesota: APS Press

Wolf M, Achtziger M, Schultz J, Dandekar T, Müller T. 2005. Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures. *RNA* 11:1616-1623

Wolf M, Ruderisch B, Dandekar T, Schultz J, Müller T. 2008. ProfDistS: (profile-) distance based phylogeny on sequence-structure alignments. *Bioinformatics* 24:2401-2402

Wolfe KH, Morden CW, Ems SC, Palmer JD. 1992. Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes. *J. Mol. Evol.* 35:304–317

Won H, Renner SS. 2005. The internal transcribed spacer of nuclear ribosomal DNA in the gymnosperm *Gnetum*. *Mol. Phylogenet. Evol*. 36:581-597

Xia X. 2000. Data Analysis in Molecular Biology and Evolution. Kluwer Academic Publishers

Xie J, Fu Y, Jiang D, Li G, Huang J, Li B, Hsiang T, Peng Y. 2008. Intergeneric transfer of ribosomal genes between two fungi. *BMC Evol. Biol*. 8:87

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

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

Yeh LC, Lee JC. 1990. Structural analysis of the internal transcribed spacer 2 of the precursor ribosomal RNA from *Saccharomyces cerevisiae*. *J. Mol. Biol.* 211:699-712

Yeh, LC, Thweatt R, Lee JC. 1990. Internal transcribed spacer 1 of the yeast precursor ribosomal RNA. Higher order structure and common structural motifs. *Biochemistry* 29:5911-5918

Young I, Coleman AW. 2004. The advantages of the ITS2 region of the nuclear rDNA cistron for analysis of phylogenetic relationships of insects: a *Drosophila* example. *Mol. Phylogenet. Evol.* 30:236-242

Zambino PJ, Harrington TC. 1992. Correspondence of isozyme characterization with morphology in the asexual genus *Leptographium* and taxonomic implication. *Mycologia* 84:12-25

Zeng Q, Bonocora RB, Shub DA. 2009. A free-standing homing endonuclease targets an intron insertion site in the *psbA* gene of cyanophages. *Curr. Biol.* 19:218-222

Zhang N, McCarthy ML, Smart CD. 2008. A macroarray system for the detection of fungal and oomycete pathogens of solanaceous crops. *Plant Dis.* 92:953-960

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

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

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

Zimmerly S, Guo H, Eskes R, Yang J, Perlman PS, Lambowitz AM. 1995b. A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. *Cell* 83:529-538

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

Zingler N, Solem A, Pyle AM. 2008. Protein-facilitated ribozyme folding and catalysis. *Nucleic Acids Symp. Ser.* 52:67-68

Zipfel RD, de Beer ZW, Jacobs K, Wingfield BD, Wingfield MJ. 2006. Multi-gene phylogenies define *Ceratocystiopsis* and *Grosmannia* distinct from *Ophiostoma*. *Stud. Mycol*. 55:75-97

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