Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Ryegrass organelle genomes: phylogenomics and sequence evaluation

Kerstin Diekmann

2010
Submission for Ph.D.

University of Dublin, Trinity College

Declaration

I, the undersigned, hereby declare that I am the sole author of this dissertation and that the work presented in it, unless otherwise referenced, is my own.

I also declare that the work has not been submitted, in whole or in part, to any other university or college for a degree or other qualification.

I authorize the library of Trinity College Dublin to lend or copy this dissertation on request.

Dedicated to my parents and my brother.

Thank you,
for always being there for me!

ACKNOWLEDGEMENTS

I am very grateful to:

My supervisors Dr. Susanne Barth and Dr. Trevor R. Hodkinson for their immense support, guidance and especially motivation throughout the time of this Ph.D. project.

Professor Kenneth Wolfe for his advice and patience with the assembly and annotation of the organelle genomes.

Dr. Dan Milbourne, Dr. Alfonso Blanco, Colum Kennedy and Dr. Paul McDermott for their help with the DNA isolation from perennial ryegrass organelles in general and Dr. Sophie Kiang and Professor Matthew Harmey for their advice on the isolation of mitochondrial DNA in particular.

The Teagasc Oak Park grass breeding group especially Olivia Aylesbury for supplying me with the large amount of seeds necessary to succeed in this project.

The Teagasc Oak Park Administration in particular Cheryl Austin, Eleanor Butler, Connie Conway, Sharon Cosgrove, Sadie Fleming, Philomena Kelly and Eugene Brennan who helped me on many occasions.

The friends that I met in Ireland particularly Ulrike Anhalt, Bicheng Yang, Céline Tomaszewski, Marta Cristilli, Stephen Byrne, Jeanne Mehenni-Ciz and Sinead Phelan. Thank you for your friendship and support that made my life in Ireland in general and the Ph.D. time in particular so much easier and nicer.

My friends back at home in Germany especially Kristin Elvers, Lena and Jörn Uwe Starcke with Julius, Meike Wildung and Luise Lange. Thank you for your friendship in general and for having been so supportive, understanding and patient with me during this time.

Teagasc for funding this Ph.D. project within the Teagasc Walsh Fellow Programme.

Summary

Perennial ryegrass (*Lolium perenne* L.) is the most important forage grass of temperate regions of the world. The main objective in breeding perennial ryegrass cultivars is to increase its biomass. Chloroplasts and mitochondria are two organelles of the plant cell that are actively involved in biomass production. Chloroplasts derive from cyanobacteria and are the location of photosynthesis in plant cells. Mitochondria derive from α-proteobacteria and are involved in cell respiration. Due to their evolutionary history both organelles still contain their own genome which is in general maternally inherited. The interest in chloroplast genome sequences increased in recent years because they offer a useful option for plant genetic engineering. The risk of transgene escape via pollen flow is reduced while the expression of the transgene due to the high number of chloroplast genome copies is increased (in comparison to nuclear genome transformation). Mitochondrial genomes are of special interest because they are involved in cytoplasmic male sterility. Cytoplasmic male sterility is a very important trait in plant breeding programmes because it enables the cost efficient production of hybrid seed. Additionally, both organelle genomes can be used for molecular evolution or phylogenetic studies, as well as for population genetic approaches. Therefore the major aim of this thesis was to sequence the entire chloroplast and mitochondrial genomes of *L. perenne* to provide sequence information for chloroplast genetic engineering approaches, insights into the mitochondrial genome of a male fertile *L. perenne* cultivar and to gather knowledge about sequence variation in both genomes that can be used to design new markers for phylogenetic and population genetic studies.

Chloroplast DNA was extracted with an in this study optimised protocol and sequenced using a shotgun sequencing approach from the *L. perenne* cv. Cashel. After sequence assembly the *L. perenne* chloroplast genome was found to comprise of 135,282 bp, divided into a large (79,972 bp) and a small (12,428 bp) single copy region separated from each other by two copies of an inverted repeat (each 21,441 bp). The chloroplast genome encodes 76 protein-coding, 30 transfer and four ribosomal RNA genes and therefore was, regarding its size and gene content/order, in the range expected for Poaceae chloroplast genomes. Thirty-three protein-coding genes of the *L. perenne* chloroplast genome were searched for RNA editing sites via reverse transcriptase PCR. RNA editing is a posttranscriptional process that can increase the gene conservation, create start and stop codons and restore the functionality of genes. Thirty-one editing sites (all C-U) were detected in 18 *L. perenne* chloroplast protein-coding genes. Chloroplast genomes are generally highly conserved, however, an unexpected and considerable high amount of variation in forms of insertion deletion events and single nucleotide polymorphisms was

found throughout the genome of *L. perenne* cv. Cashel. Nine new Poaceae-universal chloroplast markers were designed from some of those highly variable regions. The markers were tested on a subset of the Teagasc Oak Park grass collection. All markers revealed diversity across the subset although to different extent. Three markers were especially interesting due to their ability to detect high amounts of variation, to distinguish haplotypes via inexpensive agarose gel-electrophoresis, or to distinguish annual from perennial ryegrass.

The mitochondrial genome was sequenced from the *L. perenne* cv. Shandon using a hybrid sequencing approach based on Sanger and GS FLX sequencing. Due to timely restrictions of the project and the very challenging nature of plant mitochondrial genomes, the *L. perenne* mitochondrial genome could not be completely assembled. The draft assembly consists to date of 43 contigs (ca. 560 kb in total), containing all 33 expected protein-coding genes that are commonly found in Poaceae mitochondrial genomes as well as three ribosomal and twenty transfer RNA genes. Three major regions of intracellular gene transfer of approximately 2, 10 and 12 kb, respectively were detected. The fragments had their origin in the large single copy and inverted repeat region of the *L. perenne* chloroplast genome. Analyses for horizontal gene transfer using the mitochondrial genome sequence of the abuscular mycorrhizal fungus *Glomus interradices* revealed a fragment of approximately 300 bp that showed 76% identity to one region in the *L. perenne* genome. Further investigation showed that this fragment is part of a large ribosomal subunit in *Glomus* as well as *Lolium*. Therefore further studies need to be carried out to reveal if this fragment expresses similarity due to horizontal gene transfer or a homologous function in both genomes. Despite the high level of variation across mitochondrial genomes of different Poaceae subfamilies a 9-kb region consisting of five genes and exons (*ccmFN*. *rps1*, *matK*, *nad1* exon 5, *nad5* exon3), respectively, could be detected and might prove suitable for phylogenetic and population genetic studies in the future.

The protein-coding genes of both organelles proved useful in a phylogenetic study using sequence information of all publicly available chloroplast and mitochondrial Poaceae genomes. Analyses revealed that both genomes contain enough sequence information to distinguish between all Poaceae species, the different Poaceae subfamilies and were also able to reveal the relationship between the Poaceae subfamilies. Neither of the whole genome analyses supported the BEP clade hypothesis for the relationships between grass subfamilies. This result showed how promising future analyses based on a combination of sequence information from both organelles could be.

TABLE OF CONTENT

TABLE OF CONTENT I

LIST OF FIGURES VII

LIST OF TABLES X

1. General introduction to 'Ryegrass organelle genomes: phylogenomics and sequence evaluation' 1

1.1 The endosymbiotic theory 1

1.2 The hydrogen hypothesis 4

1.3 Metabolic functions of chloroplasts and mitochondria 6

1.4 Why have organelle genomes been retained in plants? 8

1.5 Mitochondrial and chloroplast mutations in plants 9

1.6 *Lolium perenne* L. 11

1.7 Aims and objectives 12

2. The chloroplast genome of *Lolium perenne* L. 15

2.1 Introduction 15

2.1.1 Discovery of the chloroplast genome 15

2.1.2 General features of chloroplast genomes 17

2.1.3 Poaceae/monocotyledonous chloroplast genomes 19

2.1.4 Aims and objectives 22

2.2 Material and Methods 24

2.2.1 Plant material 24

2.2.2 Preparations of buffers and materials before isolation 24

2.2.3 Isolation of chloroplast DNA 24

2.2.4 Whole genome amplification 29

2.2.5 Sequencing, assembling and annotating the chloroplast genome 30

2.2.6 Variation analysis 32

2.3 Results 33
2.3.1 The chloroplast genome of *Lolium perenne* 33
2.3.2 Indel/SNP analysis 33
2.3.3 Comparison of Poaceae chloroplast genomes 38

2.4 Discussion 42
2.4.1. Sequencing the chloroplast genome of *Lolium perenne* 42
2.4.2 Indel/SNP analysis 44
2.4.3 Comparison of Poaceae chloroplast genomes 46

2.5 Conclusion 48

3. RNA editing sites in *Lolium perenne* **50**

3.1 Introduction 50
3.1.1 RNA editing 50
3.1.2 RNA editing in chloroplast genomes 51
3.1.3 Chloroplast mRNA editing mechanism 53
3.1.4 Evolution of chloroplast mRNA editing sites 55
3.1.5 Aims and objectives 56

3.2 Material and Methods 57
3.2.1 Plant material 57
3.2.2 RNA extraction and cDNA synthesis 57
3.2.3 Primer design and PCR amplification 58
3.2.4 Phylogenetic analyses of editing sites in the *ndhB* gene 59

3.3 Results
3.3.1 mRNA editing sites in the chloroplast genome of *L. perenne* 61
3.3.2 Phylogenetic analyses of editing sites in three genes 62

3.4 Discussion 70
3.4.1 mRNA editing sites in *L. perenne* 70
3.4.2 Phylogenetic analyses of editing sites in three genes 74
3.4.3 Future applications 78

3.5 Conclusion 79

4. Plastid genome diversity within and among *Lolium perenne* cultivars and populations **81**

4.1 Introduction 81

4.1.1 The need for crop genetic diversity 81

4.1.2 Chloroplast microsatellite markers 84

4.1.3 Application of chloroplast microsatellite markers 84

4.1.4 Genetic diversity and *Lolium perenne* 86

4.1.5 Aims and objectives 87

4.2 Material and Methods 88

4.2.1 Characterization of Poaceae chloroplast microsatellites 88

4.2.2 Plant material 89

4.2.3 Primer design 89

4.2.4 Chloroplast microsatellite region amplification 93

4.2.5 Phylogenetic analysis 95

4.2.6 Locus-by-locus AMOVA and genetic distance analysis 96

4.3 Results 97

4.3.1 Characterization of Poaceae chloroplast microsatellites 97

4.3.2 Chloroplast microsatellite analysis 98

4.3.3 Phylogenetic and haplotype analysis 101

4.3.4 Locus-by-locus AMOVA 104

4.3.5 Genetic distance analysis 108

4.4 Discussion 111

4.4.1 Characterization of Poaceae chloroplast microsatellites 111

4.4.2 Chloroplast microsatellite analysis 112

4.4.3 Phylogenetic and haplotype analysis 115

4.4.4 Locus-by-locus AMOVA 118

4.4.5 Genetic distance analysis 120

4.5 Conclusion 121

5. First insights into the mitochondrial genome of *Lolium perenne* L. 123

5.1 Introduction 123

5.1.1 Plant mitochondrial genome feature 124

5.1.2 Horizontal and intracellular gene transfer 127

5.1.3 Cytoplasmic male sterility 129

5.1:4 Aims and Objectives 130

5.2 Material and Methods 131

5.2.1 Plant material 132

5.2.2 Mitochondrial DNA isolation 133

5.2.3 Sequencing the mitochondrial genome of *Lolium perenne* 135

5.2.4 Assembling and annotating the mitochondrial genome of *Lolium perenne* 136

5.2.5 Further analyses 139

5.4 Results 140

5.4.1 Preassemblies of the mitochondrial genome 140

5.4.2 Gene conservation and rearrangement study 142

5.4.3 Further results 154

5.5 Discussion 157

5.5.1 The mitochondrial genome of *Lolium perenne* 157

5.5.2 Horizontal and intracellular gene transfer 160

5.5.3 CMS in *Lolium perenne* 165

5.6 Conclusion 167

6. Applicability of organelle sequence information to phylogenetic studies 168

6.1 Introduction 168

6.1.1 Evolution of phylogenetic studies 168

6.1.2 Phylogenetic studies in Poaceae 171

6.1.3 Aims and objectives 172

6.2 Material and Methods 173

6.2.1 Structural characterization of monocotyledonous chloroplast genomes 173

6.2.2 Phylogenetic analyses 174

6.2.2.1 Chloroplast single gene analysis 176

6.2.2.2 Combined chloroplast and mitochondrion gene analyses 176

6.3 Results 178

6.3.1 Structural characterization of monocotyledonous chloroplast genomes 178

6.3.2 Phylogenetic analyses 181

6.3.2.1 Chloroplast single gene analysis 181

6.3.2.2 Chloroplast multigene analysis 183

6.3.2.3 Mitochondria multigene analysis 183

6.4 Discussion 185

6.4.1 Structural characterization of monocotyledonous chloroplast genomes 185

6.4.2 Single gene analysis 187

6.4.3 Chloroplast multigene analysis 191

6.4.4 Mitochondrion multigene analysis 193

6.5 Conclusion 194

7. General discussion on "Ryegrass organelle genomes: phylogenomics and sequence evaluation" 196

7.1 Introduction 196

7.1.1 The chloroplast genome of *Lolium perenne*. 196

7.1.2 RNA editing sites in *Lolium perenne*. 197

7.1.3 Plastid genome diversity within and among *Lolium perenne* cultivars and populations 198

7.1.4 First insights into the mitochondrial genome of *Lolium perenne*. 199

7.1.5 Applicability of organelle sequence information to phylogenetic studies 200

7.2 Future prospects 201

8. References **206**

9. Appendices **237**

Table A1 237

Published articles 242

LIST OF FIGURES

Figure 1 The complex picture of photosynthesis according to Eberhard et al. 2008.....7

Figure 2 Plant respiratory electron transport chain according to Day 2004. 8

Figure 3 Number of published chloroplast genome sequences since 1986 and amount of publicly available chloroplast genome sequences in November 2009 (www.ncbi.nlm.nih.gov).....18

Figure 4 Lengths of chloroplast genomes from all vascular species publicly available in July 2009 (www.ncbi.nlm.nih.gov). 20

Figure 5 Lengths of monocotyledonous chloroplast genomes (www.ncbi.nlm.nih.gov).....21

Figure 6 Overview over the major steps for the isolation of high purity chloroplast (cp) DNA.....25

Figure 7 Schematic diagram of the GenomiPhi amplification process (Amersham Biosciences Corp 2003).....30

Figure 8 Agarose gel of whole genome amplified *Lolium perenne* chloroplast DNA. .31

Figure 9 Alignement of contiguous *Lolium perenne* chloroplast genome sequences with the chloroplast genome of *Agrostis stolonifera.*.....32

Figure 10 Observed single nucleotide polymorphism in the *trnF-ndhJ* intergenic spacer region of the *Lolium perenne* chloroplast genome. 33

Figure 11 Circular structure of the chloroplast genome of *Lolium perenne*.....34

Figure 12 Single nucleotide polymorphisms (SNPs) in the *psbE* and *atpA* region in the *Lolium perenne* chloroplast genome.....36

Figure 13 Two hairpin loops found in the chloroplast genome of *Lolium perenne*. 38

Figure 14 Amino acid based alignment of the chloroplast *rps18* gene alignment of eleven Poaceae species. 42

Figure 15 Possible chloroplast mRNA editing mechanism as predicted for the *ndhD*-1 editing site in *Arabidopsis* by Okuda et al. (2006). 55

Figure 16 Amplified PCR-products of *Lolium perenne* chloroplast genes ready for sequencing. 60

Figure 17 Alignment of *ndhB*-DNA sequence with cDNA sequences derived from two different primer sets.....63

Figure 18 Sequencing trace files of editing sites in the chloroplast genome of *Lolium perenne*. 67

Figure 19 Bayesian consensus tree derived from the phylogenetic analysis of the *ndhA, ndhB* and *rpoB* gene. 69

Figure 20 Example of amplified products on a 2.5% MetaPhor® Agarose gel (Lonza, Rockland, ME, USA) stained with 1% ethidium bromide (10 mg/ml). 94

Figure 21 Number of microsatellites (>7 nucleotides) in complete Poaceae chloroplast genomes. 98

Figure 22 Nucleotide usage of chloroplast microsatellites (>7 nucleotides) in Poaceae species. 99

Figure 23 Distribution of microsatellites (> 9 nucleotides) in the chloroplast genome of *L. perenne*. 100

Figure 24 Amplification results using marker TeaCpSSR32 on a set of 14 different grass species. 101

Figure 25 Agarose gel picture and alignment of DNA sequences obtained from PCR reactions with TeaCpSSR27 and different individuals from *Lolium perenne* accession PI267059. 102

Figure 26 TeaCpSSR26 PCR products (5 µl/sample) from different individuals of two *Lolium perenne* accessions. 103

Figure 27 Bayesian inference tree using DNA sequences from haplotypes as input file. 106

Figure 28 Number of publicly available mitochondrial genome sequences from moss and higher plant species in October 2009 (www.ncbi.nlm.nih.gov). 124

Figure 29 Lengths of mitochondrial genomes from all vascular plant and moss species publicly available in October 2009 (www.ncbi.nlm.nih.gov). 125

Figure 30 Undigested (a) and restriction enzyme digested (b) *Lolium perenne* mitochondrial DNA. 134

Figure 31 Overview over the assembly process used for gaining a draft assembly of the mitochondrial genome of *Lolium perenne*. 142

Figure 32 Alignment of the complete *Bambusa oldhamii* and the draft *Lolium perenne* mitochondrial genome sequence to the complete and annotated *Triticum aestivum* mitochondrial genome sequence using the Shuffle-Lagan alignment algorithm from the online program Vista (http://genome.lbl.gov/vista/ index.shtml). 146

Figure 33 Comparison of the mitochondrial gene order in species of different Poaceae subfamilies using *T. aestivum* as reference sequence. 148

Figure 34 Comparison of the mitochondrial gene order in different species of the Panicoideae subfamily *using Zea mays ssp. mays* as reference sequence. 150

Figure 35 Comparison of the mitochondrial gene order in the two *Oryza sativa* varieties. 151

Figure 36 Comparison of the mitochondrial gene order of *Triticum aestivum* with the gene clusters found so far in the draft assembly of *Lolium perenne*. 151

Figure 37 Alignment of the *matR – rps1* intergenic spacer region from five different Poaceae species. 152

Figure 38 Alignment of the *Lolium perenne* mitochondrial genome draft assembly with itself to highlight repetitive regions using the NCBI blast tool. 154

Figure 39 Blast result of the horizontal gene transfer analyses showing a sequence alignment of a *Glomus intraradices* mitochondrial DNA fragment with a region of the *rrn26* gene of the *L. perenne* mitochondrial genome. 155

Figure 40 Alignment of the *Lolium perenne* mitochondrial draft assembly with the complete chloroplast genome of *L. perenne*. 156

Figure 41 Alignment of a fragment of the *Lolium perenne* chloroplast (cp) *rpoC2* gene with its homologue in the *L. perenne* mitochondrial (mt) genome. 156

Figure 42 Alignment of the chloroplast derived mitochondrial *trnV*-GAC of *Lolium perenne* with the original chloroplast *trnV*-GAC gene of *L. perenne* showing complete homology between the two sequences. 156

Figure 43 Comparison of the chloroplast genome structure of Poaceae and non-Poaceae species using *Lolium perenne* (Poaceae) and *Dioscorea elephantipes* (non-Poaceae) for the illustration of differences. 179

Figure 44 Illustration of protein-coding gene loss and gene rearrangements in completely sequenced monocotyledonous chloroplast genomes. 180

Figure 45 Number of parsimony informative sites (PIS) in relation to the alignment length for chloroplast and mitochondrial protein coding genes. 181

Figure 46 Results of chloroplast protein-coding single gene analysis for the five genes with the highest bootstrap (beneath lines) and posterior probability (above lines) values. 182

Figure 47 Bayesian inference tree of 76 protein coding chloroplast genes from 19 monocotyledonous species including 14 Poaceae species. 184

Figure 48 Bayesian inference tree of 32 protein coding mitochondrial genes from eleven monocotyledonous and two dicotyledonous species. 184

LIST OF TABLES

Table 1 Gene annotation of the chloroplast genome of *Lolium perenne* 35

Table 2 Indels observed in the chloroplast genome of *Lolium perenne.* 36

Table 3 Single nucleotide polymorphisms in the chloroplast genome of *Lolium perenne* 37

Table 4 Length variation of more than 100 bp for intergenic spacer (IGS) and intron regions in Poaceae chloroplast genomes. 40

Table 5 Variation in length of different chloroplast genes. 41

Table 6 RNA editing sites found in the chloroplast genome of *Lolium perenne* in comparison to the editing sites found in other grasses. 62

Table 7 Names of accessions, their origin and the numbers of individuals used in the *Lolium perenne* diversity study as well as the haplotypes found within them. 91

Table 8 Chloroplast microsatellite regions (>10 nucleotides) found in the *Lolium perenne* chloroplast genome 94

Table 9 Poaceae universal chloroplast microsatellite primers 95

Table 10 Ability of newly designed Poaceae universal primers to detect variation within the chloroplast genome of *Lolium perenne* and other *Lolium* species (*L. hybridum, L. multiflorum, L. persicum, L. remotum, L. subulatum, L. temulentum*) 103

Table 11 Results of the locus-by-locus AMOVA. 107

Table 12 Results from the genetic distance analysis including all accessions and sequences from all markers. 109

Table 13 Results from the genetic distance analysis focussing only on *Lolium perenne* accessions and sequences derived from data of markers TeaCpSSR27 and TeaCpSSR31. 110

Table 14 Overview of the preassemblies of the *Lolium perenne* mitochondrial genome retrieved from the sequencing company. 141

Table 15 Overview over the different contigs, their lengths and the genes they are encoding. 144

Table 16 Comparison of the gene content and gene length (in number of amino acids) of all publicly available Poaceae mitochondrial genome sequences and of one dicotyledonous (*Arabidopsis thaliana*) and one moss (*Marchantia polymorpha*) species. 145

Table 17 Number of pairwise differences in the sequence alignment of the large plant mitochondrial gene cluster '*ccmFN-nad5* exon 3'.. 147

Table 18 List of species and accession numbers for the phylogenetic analyses using chloroplast and mitochondrial genes. 174

Table 19 Seventy-six protein coding chloroplast genes used for analysing the phylogenetic relationship of 19 monocotyledonous species.....177

Table 20 Thirty-two protein coding mitochondrial genes used for analysing the phylogenetic relationship of eleven monocotyledonous and one dicotyledonous species. 178

Table A1 Primer sequences for mRNA editing analysis. 237

Chapter 1

General introduction to
'Ryegrass organelle genomes: phylogenomics and sequence evaluation'.

"The palm behaves so peacefully, so passively, because it is a symbiosis, because it contains a plethora of little workers, green slaves (chromatophores) that work for it and nourish it."
Mereschkowsky (1905)

1.1 The endosymbiotic theory

For a long time it was thought that the complete amount of DNA of a plant cell is located in the nucleus and inherited in a Mendelian way. Evidence that extranuclear factors might also be involved in the inheritance mechanism was first reported by Carl Correns in 1908 based on his experiments with the species *Mirabilis jalapa* (four o'clock flower; Caryophyllales). However, Correns did not believe in an inheritance mechanism other than the Mendelian one until 1937 (Correns 1937; Rheinberger 2000). Even then, he strongly doubted that a cytoplasmic inheritance mechanism could be based on genes and thought it could rather be due to differential regulation of the nuclear genes, with their products translocated into the cytoplasm (Correns 1937). However, with the discovery of extranuclear DNA in the chloroplasts and mitochondria in the mid-1960s, Correns assumption was soon refuted. The amount of DNA located in the chloroplasts and mitochondria was significantly less than the amount of nuclear DNA, nevertheless one question remained: Why do chloroplasts and mitochondria contain DNA at all?

Already in 1883 Schimper and other research groups mentioned that plants possibly originate from a symbiotic combination of a colourless organism with a chlorophyll stained organism (Schimper 1883; Sapp et al. 2002). In 1905 Konstantin Sergejewiz Merezkovskij (Constantin Mereschkowsky) speculated that chloroplasts (chromatophores) were originally free living organisms that were incorporated into

plant cells and lived with them in a symbiotic relationship (Mereschkowsky, 1905). Mereschkowsky based his theory ('endosymbiotic theory') on the following observations: a) chloroplasts never arise *de novo,* only through division of already existing chloroplasts; b) chloroplasts are independent of the nucleus; c) chloroplasts resemble zoochlorellae completely (small green algae that live within the bodies of various freshwater protozoans and invertebrates); d) Cyanophyceae are organisms that can be regarded as free-living chloroplasts; e) Cyanophyceae can invade the cytoplasm of other organisms and live within them. The theory that mitochondria are symbionts, too, was developed in 1918 by Paul Portier based on his experiments and later on discussed by Wallin (1923) based on the observation of similarities between bacteria and mitochondria (Sapp et al. 2002). However, for a long time, the wider scientific community rejected both theories and thus the possibility that chloroplasts and mitochondria are of endosymbiotic origin.

In 1967 Lynn Margulis (married name Sagan) revived the endosymbiotic theory (Sagan 1967). She proposed that an increase of oxygen concentration on earth led to the evolution of photosynthetic prokaryotic cells (oxygen could have increased due to tectonic activity). Eventually, a cell population arose from those photosynthetic cells that used "photoproduced ATP with water as the source of hydrogen atoms in the reduction of CO_2 for the production of cell material" (Sagan 1967). As a consequence from this process more oxygen was set free. From these originally autotrophic microbes, mutants evolved with an aerobic respiration mechanism that occurred in the darkness. Thus the first prokaryotic algae had evolved having both a photosynthetic and respiratory mechanism (Sagan 1967).

The link between atmospheric oxygen concentrations and the evolution of photosynthetic active organisms was also stressed by Knoll et al. (1992). First photosynthetic active organisms originated at around 3,300 million years ago (Knoll et al. 1992; Willis and McElwain 2002). Oxygenic photosynthesis appeared approximately 2,800 million years ago and a drastic increase of oxygen concentration occurred around 2,300 million years ago (Olson 2006). From that time on life became dependent on oxygenic photosynthesis derived products for nutrition. At some point a heterotrophic anaerobic organism engulfed an aerobic prokaryotic microbe into its cytoplasm, the protomitochondrion, which led to the survival of the anaerobic

organism in a highly aerobic environment (Sagan 1967). This endosymbiosis became obligate and resulted in the first aerobic amitotic amoeboid organism. According to Sagan (1967) different photosynthetically active prokaryotes, the protoplastids, were ingested by several heterotrophic protozoans throughout evolution and resulted in the eukaryotic algae and plants as we know them today. The dates of the earliest endosymbiotic events are unknown but could be between 2,400 and 2,800 million years ago and complex eukaryotic organisms appear in the fossil record from approximately 2,100 million years ago; the oldest algal fossils, including *Grypania*, date from this time (Willis and McElwain 2002). Although DNA in both chloroplasts and mitochondria had already been found by the time Sagan published the revived endosymbiotic theory, she still encountered great criticism (e.g. Cavalier-Smith 1975). However, the endosymbiotical origin of organelles particularly chloroplasts is now widely accepted by scientists (Palmer et al. 2004). Palmer et al. (2004) argue that the best definition of a plant is "an organism containing a plastid" and hence under this definition we can deduce that symbiosis led to the evolution of plants.

The further investigation of chloroplast origins revealed cyanobacteria as their ancestors. Primary symbiosis, the original uptake and retention of a cyanobacterium by a eukaryontic organism, was observed for three different lineages - glaucophytes, red algae and green algae (Keeling 2004). Green algae are the ancestors of land plants, thus land plant chloroplasts originate from a primary symbiosis. Plastids derived from primary endosymbiosis characteristically have two membranes that originate from the inner and outer membrane of the ancestral cyanobacterium. The ancestral genome has been drastically reduced by loss or transfer of the majority of genes to the nucleus of the eukaryotic host. However, chloroplasts are functional due to post-translational targeting of nuclear-encoded proteins (Keeling 2004).

Many plastid-bearing eukaryotes of other lineages aquired their plastids via secondary endosymbiosis. Secondary endosymbiosis is the uptake and retention of a primary plastid containing algae into a eukaryotic organism. Plastids that originated from secondary endosymbiosis can be recognized by their additional membranes. Besides the two membranes belonging to the ancestral cyanobacterium, they often show a third membrane which corresponds to the cytoplasmic envelope of the endosymbiotic algae and a fourth membrane which corresponds to the membrane

system of the secondary host (reviewed in Keeling 2004). Secondary symbionts are found in many major groups of eukaryotes including Euglinoids (Excavates), Chlorachniophytes (Rhizaria), and several but not all Chromists and Alveolates (Chromalveolates). See Palmer et al. (2004) for a discussion about the origin and diversification of plastid containing eukaryotes.

1.2 The hydrogen hypothesis

Mitochondria are also believed to have endosymbiotic origins. They very likely derive from the rickettsial subdivision of α-proteobacteria. The most mitochondrial-like eubacterial genome sequenced belongs to the α-proteobacterium *Rickettsia prawazekii* (Gray et al. 1999). As was the case for the ancestral chloroplast genome, it is believed that the ancestral mitochondrial genome either lost or transferred the majority of its genes to the nucleus. The most bacterial-like mitochondrial genome sequenced is that of the protozoon *Reclinomonas americana* (Gray et al. 1999). With 97 genes it is also the most gene rich mitochondrial genome ever sequenced. Species of the related genus *Rickettsia* are intracellular parasites and contain, due to their parasitic lifestyle, a reduced genome compared to their free-living relatives. Comparing the gene content of *Reclinomonas americana* with the gene content of *Rickettsia prawazekii* revealed additional genes that are missing in the bacterial-like *R. americana* mitochondrial genome. However, it is not believed that mitochondria derive directly from *Rickettsia* but rather that Rickettsia and mitochondria have a common ancestor from which they evolved by different genome reducing processes (Gray et al. 1999).

According to the endosymbiotic theory, proto-mitochondria were engulfed by an anaerobic, heterotrophic type of organism via phagocytosis. Whether this organism was already a complete nucleus-containing eukaryote or rather an archaebacterium is not clear. In recent years evidence became available that the nuclear genome contains informational genes of rather archaebacterial origin and operational genes of mainly eubacterial origin (Gray et al. 1999). Therefore it was suggested that the nuclear genome originates from a fusion of an eubacterial with an archaebacterial genome.

Archezoa, a group of eukaryotes lacking mitochondria and capable of phagocytosis, are believed to be the ancestral host of the mitochondrial endosymbiont. However, with the discovery of mitochondrial protein coding genes in Archezoa species it becomes possible that they contained, at some stage, mitochondria and lost them during evolution. Recent research indicates that the common ancestor of eukaryotes already contained an endosymbiont, the protomitochondrion, because even organisms that were believed to be mitochondrion free contain mitochondrial remnant organelles as for example observed for the protozoan pathogen *Giardia intestinalis* (Zimmer 2009; Tovar et al. 2003). Based on these observations Martin and Müller (1998) proposed a new hypothesis explaining the origin of eukaryotes and mitochondria, known as the hydrogen hypothesis.

The hydrogen hypothesis assumes a symbiosis between a free-living H_2 and CO_2 producing eubacterium as symbiont and a methanogenic archaebacterium as host. Both 'met' under anaerobic conditions where CO_2 and H_2 were sufficiently available to ensure the independent viability of the host. However, once the host would be removed from these ideal conditions it became immediately dependent on the symbiont. Therefore, successful hosts would be predicted to have 'sticked' very tightly to their symbionts (Martin & Müller 1998). Hosts with a large cell surface that were able to surround the symbionts and thus increase the contact and the uptake of H_2 and CO_2 into the cytosol, had a selective advantage. But increased contact between host and symbiont also resulted in decreased contact with the environment for the symbiont to feed itself. Hosts that were able to feed their symbionts would have a selective advantage. This advantage could have been gained by, for example, rearranging genes. If the symbiont's genes for importing carbon were transferred to the chromosomes of the host, then the host would theoretically be able to feed the symbiont and itself. Simultaneously the symbiont would become dependent on the host as is the situation in modern eukaryotic cells. Modern eukaryotic cells use organic molecules to create pyruvate so that mitochondria or hydrogenosomes (hydrogen-producing organelles, evolutionary related to mitochondria, found in protozoa, fungi and ciliates; Benchimol 2009) can produce ATP (adenosine triphosphate). The hydrogen hypothesis assumes that the symbiont was originally able to produce ATP aerobically and anaerobically. As more oxygen accumulated in the host's environment the symbiont started generating ATP in an aerobic way which

led the endosymbiont to become a mitochondrion and its host to develop an aerobic nature. Hosts that continued living in an anaerobical environment would have given the symbiont the possibility to loose its aerobic metabolism and to become a hydrogenosome. Hosts that developed a parasitic nature had no need of ATP producing symbionts and consequently lost them (Martin & Müller 1998).

1.3 Metabolic functions of chloroplasts and mitochondria

Chloroplasts have an obvious structure of grana thylakoids (stacked membranes) that contain the characteristic pigment chloropyhll. The thylakoids are surrounded by the colourless stroma. The highly important metabolic process photosynthesis takes place within the chloroplasts. Although only as a by-product, photosynthesis generates oxygen which is essential for life on earth. The photosynthetic process can be divided into a light (in the thylakoids) and dark reaction (in the stroma). During the light reaction chlorophyll molecules that are arranged in light-gathering arrays absorb visible light from a wavelength of 400 to 700 nm. The light harvesting arrays are part of two photosystems which have specialised chlorophyll molecules at their centre (Campbell & Reece 2008). Photosystem I contains chlorophyll P_{700} and photosystem II chlorophyll P_{680}. After initial excitation, the light energy is passed on to the centre of the photosystems. Both photosystems are part of electron transfer chains which result in the production of ATP and reduction of $NADP^+$ to NADPH. Both these products are used in the dark reaction. The dark reaction (Calvin cycle) mainly fixes carbon dioxide in to glucose under reduction of ribulosephosphate and oxidation of NADPH and ATP (Campbell & Reece 2008). The complete photosynthetic process (Figure 1) can be summarized in the formula:

$$6\,CO_2 + 12\,H_2O \xrightarrow{\text{light energy}} C_6H_{12}O_6 + 6\,H_2O + 6\,O_2$$

Chloroplast genomes still code for many of the components that are actively involved in photosynthesis. A well known component is the large subunit of the ribulose-1, 5-biphosphate carboxylase/oxygenase, but there are also many components of photosystems I and II, the cytochrome b/f complex and the ATP synthase complex that are encoded by the chloroplast genome. Only the light-harvesting chlorophyll protein complex is based on completely nuclear encoded proteins.

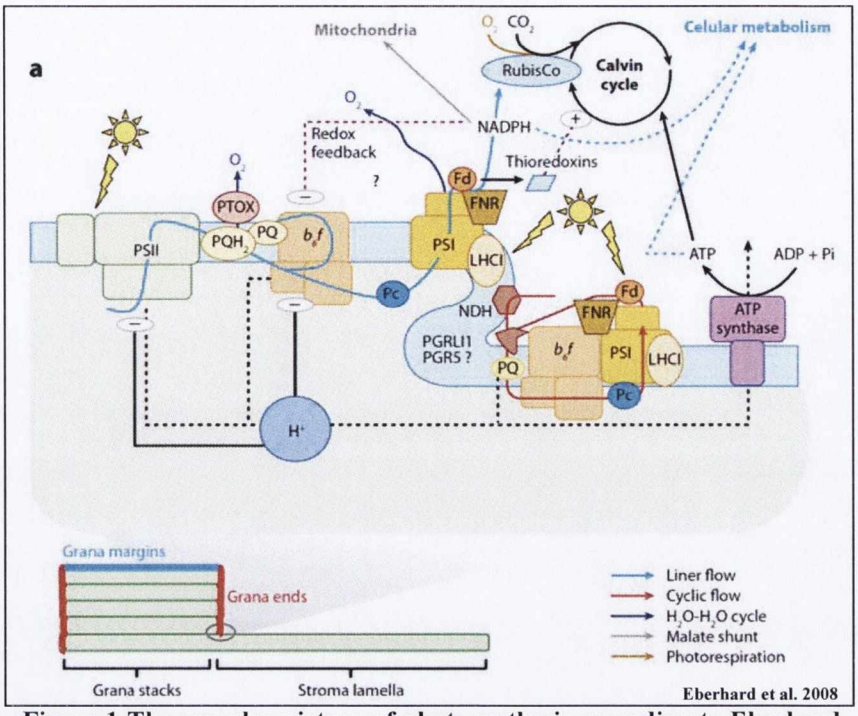

Figure 1 The complex picture of photosynthesis according to Eberhard et al. 2008. Photosystem II is enriched in stacked membranes in the grana regions and Photosystem I is enriched in unstacked membranes in the stroma lamellae regions. Fd: ferredoxin; FNR: Fd-NADP+ reductase; Pc: plastocyanin; PQ: plastoquinone; PTOX: plastoquinone terminal oxidase; NDH: NADP plastoquinone reductase.

Mitochondria are the location of aerobic respiration that provides the cell with the necessary energy to live (Day 2004). The main process behind the synthesis of ATP in the mitochondria is the tricarboxylic acid (TCA) cycle (also citric acid or Krebs cycle) which is dependent on a product from glycolysis. Glycolysis is the metabolic pathway that converts glucose from the photosynthetic process to pyruvate. In the TCA cycle pyruvate is in a first step oxidized to acetyl CoA under removal of carbon dioxide and the addition of coenzyme A. Acetyl CoA together with oxalacetate results in citrate. Citrate is subsequently broken down to isocitrate, α-ketoglutarate, succinyl CoA, succinate, fumarate and oxaloacetate under the reduction of NAD^+ to NADH, FAD^+ to $FADH_2$ and the catalytic splitting of water which both result in the synthesis of ATP. The complete TCA cycle takes place in the mitochondrial matrix except from the oxidation of succinate to fumerate. Succinate dehydrogenase is the only enzyme of the TCA cycle that is located in the mitochondrial membrane and is also part of the respiratory electron transfer chain (Day 2004). The respiratory electron transfer chain oxidizes, in a first step, NADH and $FADH_2$ from the TCA

cycle to NAD^+ (complex I) and FAD^+ (complex II), respectively. The electrons set free by this process are then consequently transferred to complex III and IV where they reduce one molecule of oxygen to water. This process accumulates H^+ ions on the cytosolic side of the inner mitochondrial membrane, which diffuse via complex V back into the matrix. As a result of this process ADP is converted to ATP (Figure 2) (Day 2004).

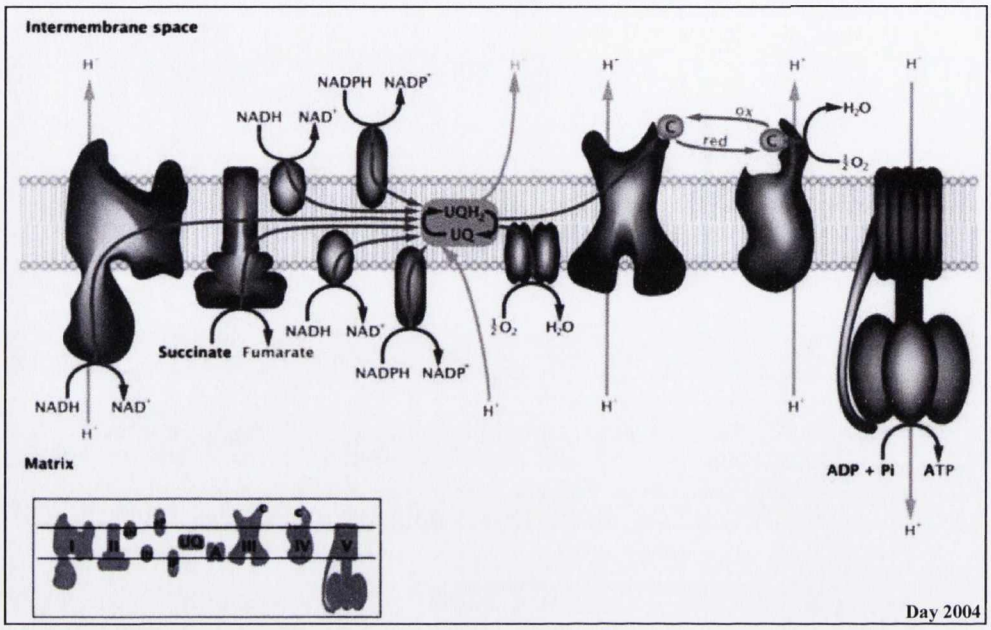

Figure 2 Plant respiratory electron transport chain according to Day 2004. lower panel: I = complex I (NADH-dehydrogenase subunits), II = complex II (succinate dehydrogenase), III = complex III (apocytochrome subunits), IV = complex IV (cytochrome c oxidase subunits) and V = complex V (ATP synthase F1 subunits), UQ = ubiquinone, eN = external NADH dehydrogenase, eP = external NADPH dehydrogenase, iN = external NADH dehydrogenase, iP = external NADPH dehydrogenase, A = alternative oxidase (Aox); c = cytochrome.

In general, all plant mitochondrial genomes encode genes for complex I (NADH-dehydrogenase subunits), complex III (apocytochrome subunits), complex IV (cytochrome c oxidase subunits) and complex V (ATP synthase F1 subunits) (Clifton et al. 2004).

1.4 Why have organelle genomes been retained in plants?

Plant cells have their genetic information distributed over three different locations. The largest amount is located in the nucleus, smaller amounts can be found in the mitochondria and chloroplasts. Although many mitochondrial and chloroplast genes

were already transferred to the nucleus since the incorporation of both organelles in the plant cell, a core of around 110 genes is retained in chloroplasts and of 35 genes in mitochondria. Maintaining three different genomes causes high energetic costs for the plant cell. It also means a high risk of mutations from free-radical by-products of the electron transfer reactions (Woodson & Chory 2008). Therefore it must be advantageous to the plant cell to retain all three genomes. Two hypotheses that could explain the maintenance of the mitochondrial and chloroplast genomes are reviewed in Woodson and Chory (2008). The first hypothesis assumes that the transport of proteins for the photosynthesis and respiration is not efficient enough because the proteins are either too hydrophobic or probably even toxic when accumulated in the cytoplasm. Also an efficient assembly of multi-subunit complexes might require the protein synthesis at the target location. The second hypothesis assumes that in maintaining chloroplast and mitochondrial genomes the plant cell is able to respond to changes in the organelle redox state rapidly and organelle specific.

1.5 Mitochondrial and chloroplast mutations in plants

The most well known form of mutation associated with chloroplasts is probably chlorophyll deficiency. Mutants can be completely white, variegated or viriscent (Hirao et al. 2009). Variegated or viriscent plants are viable and desired in horticultural breeding programmes, but albino plants are not viable due to their inability to photosynthesize and therefore their loss of capability to live autotrophically. Analysing viriscent plants of the coniferous tree species *Cryptomeria japonica* revealed three insertion/deletion events in chloroplast genome coding regions (*matK*, *ycf1*, *ycf2*) that are very likely responsible for the natural occurrence of the defect. This study was facilitated due to the fact that the complete chloroplast genome sequence of a normal *C. japonica* variety was available (Hirao et al. 2009). Ankele et al. (2005) analysed possible chlorophyll deficiency mechanisms in chlorophyll deficient cereals derived from microspore embryogenesis. Microspore embryogenesis is a widely used approach to obtain double haploid plants in plant breeding programmes. However, the adoption of this method for the production of double haploid lines in cereal breeding programmes is not very successful due to a high yield of albino plants. Studies showed that this high amount of albino plants is to

some extent due to rearrangements and deletions in the chloroplast genome. In cereal species the absence of the chloroplast *ycf1* gene causing instability of the chloroplast genome structure and a change in the transcript level of nuclear genes might both lead to the occurrence of albino plants and might be an explanation why this economically valuable approach can not be applied to cereals (Ankele et al. 2005).

Mitochondrial mutations are mainly based on rearrangements or deletions within the mitochondrial genome. The most intensely studied mutations based on rearrangements confer cytoplasmic male sterility (CMS) or restore the fertility of CMS plants. The term 'cytoplasmic male sterility' describes the inability of plants to produce functional pollen. It occurs naturally in wild species and in agricultural crops following interspecific crosses (Kiang et al. 1994, reviewed in Kubo & Newton et al. 2008). Cytoplasmic male sterility is very likely caused by the expression of chimeric proteins that are the result of a mitochondrial genome rearrangement. The fertility of CMS plants is in general restored by nuclear restorers of fertility (*Rf*) genes, which encode mitochondrial-localized PPR proteins, *Rf* genes generally reduce the accumulation of deviant mitochondrial proteins by probably changing post-transcriptionally CMS mRNA transcripts via endonucleolytic cleavage, editing and destabilization (reviewed in Woodson & Chory 2008). However, also cytoplasmic reversion, the restoration of fertility in CMS plants due to rearrangement and deletion mutations in the mitochondrial genome, was observed. This phenomenon is rather rare but was observed to occur spontaneously in CMS lines of maize and common bean. In maize the cytoplasmic reversion seems mainly to be due to rearrangements and smaller deletions while in bean it was found to be associated with the nearly complete loss of a 210 kb fragment that carries the CMS inducing sequence. The loss of this big fragment does not have any further consequences because the essential mitochondrial genes are located on other regions of the genome (reviewed in Newton et al. 2004). Extensively studied mutations based on the deletions of genes result in general in abnormal growth. Because of the mitochondrial gene involvement in essential metabolic functions of the plant, deletions of mitochondrial genes usually result in the death of the plant. Only heteroplasmic plants, plants that contain normal and mutated mitochondria, are able to live with deletion mutations in coding regions. However, these heteroplasmic plants show in general abnormal phenotypes. In *Zea mays*, for example, mutations that are due to mitochondrial deletions resulting in

abnormal growth are termed non-chromosomal stripe (NCS) mutants. Depending on the mitochondrial gene deleted in the NCS mutants, plants can be pale green in some mutated leaf sectors (partially deletion of Complex I subunit gene), have yellow stripes on the leaves (partially deletion of Complex IV gene), exhibit a short deformed and slow growth combined with brown or necrotic leaf sectors (partially deletion of a ribosomal transcript unit). Sectors of kernels can be aborted and tassels reduced (reviewed in Newton et al. 2004). Therefore mitochondrial genome alterations do not only have an impact on the life cycle of a plant in general, they furthermore also influence the most important agricultural aspects such as biomass and seed yield.

1.6 *Lolium perenne* L.

Lolium perenne L. (perennial ryegrass) is the most important perennial agricultural grass (Wilkins 1991) of temperate regions of the world. It is widely used as forage due to its high digestibility combined with a good grazing tolerance and adequate seed production. But it is also used for sport pitches, lawns and other amenity areas. In 2006-2007 more than one third of the world grass seed production was from *L. perenne* (http://www.worldseed.org). Thus *L. perenne* has the highest economic impact as a forage and grassland crop. Species within the genus *Lolium* are believed to have originated in the Mediterranean area, the temperate regions of Europe and Asia as well as the Canary Islands. The first evidence of *L. perenne* dates back to 4000 BC according to subfossil records (reviewed in Lenuweit & Gharadjedaghi 2002). *Lolium perenne* occurs naturally in temperate Europe, Asia and North Africa and was introduced into North and South America, Australia and New Zealand (Hubbard 1992).

Taxonomically, *L. perenne* is a monocot and belongs to the Poaceae family which contains approximately 10,000 species and many of the most important agricultural species of the world, such as *Oryza sativa* (rice), *Saccharum officinarum* (sugar cane), *Triticum aestivum* (wheat) and *Zea mays* (maize). Within Poaceae, *L. perenne* is grouped into the subfamily Pooideae (one of 13 subfamilies; Bouchenak-Khelladi et al. 2008) which also contains *T. aestivum*, *Secale cereale* (rye), *Avena sativa* (oat),

Hordeum vulgare (barley) as well as other grassland species as for example *Agrostis stolonifera* (bentgrass) and *Festuca pratensis*. *Lolium* can be further classified into the tribe Poeae together with *Festuca*.

Lolium perenne is a cross-pollinating, mainly self-incompatible, diploid (2n=14) species. In contrast to most agricultural important annual grass species, the breeding of perennial grasses and specifically *L. perenne* has a rather recent history as it began not even a century ago. The first varieties, which are still in use today, became available around 1930 (Wilkins 1991; Humphreys et al. 2006). Main breeding objectives for *L. perenne* are improvement of the biomass yield, grass quality (e.g. digestability, carbohydrate and protein content), persistency, cold stress tolerance, drought tolerance and disease resistance. Cultivars are generally produced by pair crossings of several distantly related plants with superior performance followed by recurrent selection via the poly- or topcross method. Poly- or topcross plants are then bulk-harvested and form the new 'synthetic' cultivar (Wilkins 1991). Therefore a *L. perenne* cultivar represents, in general, a population with a heterozygous nuclear genomic background.

1.7 Aims and objectives

Lolium perenne is the most important forage grass species in Ireland. The two most important objectives in breeding *L. perenne* are to increase the biomass yield and to improve the grass quality (Wilkins 1991). Both of these objectives are tightly linked in many ways to the processes of photosynthesis and respiration. Many genes for these two processes are still encoded by the organelles in which they mainly take place – the chloroplasts and mitochondria. Few studies on the chloroplast and mitochondrial genome of *L. perenne* alone have been carried out. Thorough research has been conducted on CMS in *L. perenne* (Rouwendal et al. 1992, Kiang 1992, Kiang et al. 1993, Kiang et al. 1994, Kiang & Kavanagh 1996, McDermott et al. 2008) and some research has been carried out on chloroplast (McGrath et al. 2006, McGrath et al. 2007, McGrath 2008) and mitochondrial (Sato et al. 1994) genome variation in *L. perenne*. All studies were generally based on partial coding or non-coding sequences of either organelle genome. Therefore the project that formed the

basis of this thesis "Ryegrass organelle genomes: phylogenomics and sequence evaluation" aimed to sequence the entire chloroplast and mitochondrial genome of *Lolium perenne* to facilitate insights into the organellar evolution of this important forage species and to gain new knowledge that can be applied to improve breeding programmes.

In recent years the interest in chloroplast and plant mitochondrial genomes increased rapidly. Chloroplast and mitochondrial genomes comprise around 150 kb and 450 kb, respectively, and encode both less than 110 different genes. Thus they are much smaller and therefore faster, more cost efficient and easier to sequence and analyse than the nuclear genome. This project started in October 2006 when only very few organelle genomes of grass species had been sequenced completely, among them maize, wheat and rice. The organelle genomes of a turf and forage grass species had not been sequenced, yet, and only little information was available about them supporting the decision to focus in this thesis on sequencing the organelle genomes of *L. perenne*.

Protocols for the isolation of organelle DNA exist for many species, even for some Poaceae species, however, these protocols are in general species specific and/ or require the use of expensive and specific laboratory equipment such as for example an ultracentrifuge. At the beginning of this study the Teagasc Crops Research Centre in Carlow did not own an ultracentrifuge. Therefore the first objective of the present study was to develop a new protocol or modify existing protocols that are not dependent on such specific instruments and enable the isolation of organelle DNA with standard laboratory equipment.

RNA editing is a mechanism that alters genetic information after transcription and leads to the creation of start and stop codons by generally increasing the degree of conservation among species. RNA editing is suspected to be involved in the CMS system of plants and gathering more information about it should lead to a better understanding of both processes. Around 30 editing sites can be expected within chloroplast genomes while around 400 can be found in plant mitochondrial genomes. This study focusses only on chloroplast RNA editing sites due to the timely restriction of the project.

Sequence information from the chloroplast and mitochondrial genomes will be used in phylogenetic analyses of Poaceae species and in the generation of new markers that can be applied to plant breeding. Both organelle genomes are very interesting from a breeding perspective because they are inherited maternally in *L. perenne* (observed exception: the mitochondrial genome was paternally inherited during the creation of interspecific hybrids from *L. perenne* and *Festuca pratensis*; Kiang et al. 1994). Thus markers can be used to trace maternal inheritance, define genetic breeding pools and even assist in the collection of ecotype populations by identifying species. Chloroplast and mitochondrial genome sequence information enables also the genetic modification of plants via organelle engineering. Due to the maternal inheritance of both organelles the risk of spreading the transgenes via outcrossing is reduced. Furthermore a higher expression of the transgene can be observed due to the high amount of organelles and organelle genomes within a single plant cell.

In summary the objectives of the present study were:

(1) To develop efficient, or optimize existing, protocols for the extraction of plastid and mitochondrial DNA from *Lolium perenne*.

(2) To sequence the complete chloroplast genome of *Lolium perenne* and to evaluate its genetic information in comparison to other Poaceae species.

(3) To analyse the chloroplast genome of *L. perenne* for RNA editing sites and their potential evolutionary history.

(4) To analyse the *L. perenne* chloroplast genome for variable regions in comparison to other Poaceae species to facilitate the design of new markers capable of quantifying the diversity within the Teagasc Oak Park grass breeding and Irish genetic resources collection.

(5) To sequence the complete mitochondrial genome of *L. perenne* to gain insights into the mitochondrial genome structure of Poaceae species.

(6) To evaluate the suitability of complete or single chloroplast and mitochondrial protein coding gene sets for their application in phylogenetic studies.

Chapter 2

The chloroplast genome of *Lolium perenne* L.

"In retrospect it is not surprising that the
discovery of DNA in chloroplasts was accepted so readily.
The evidence which appeared in 1962-4 was compelling,
the ground had been prepared by the geneticists,
it was an idea whose time had come."
Kirk (1986)

2.1 Introduction

2.1.1 Discovery of the chloroplast genome

The green alga *Chlamydomonas reinhardtii* is a commonly used model organism. In 1954 Ruth Sager observed a non-Mendelian inheritance of streptomycin-resistance in *C. reinhardtii* (Sager 1954) and this led to the discussion of whether DNA might, not only be found in the nucleus but, also be found in chloroplasts. Ris and Plaut (1962) were the first to find conclusive evidence for the presence of DNA within chloroplasts. Via Feulgen reaction staining they were able to detect small bodies that were Feulgen-stained in a similar intensity as the nucleus in *C. moewusii*. After deoxyribonuclease digestion the Feulgen staining disappeared from both nucleus and chloroplasts. Electron microscopy showed that these small bodies were located within the chloroplast and had the shape of 25 A fibrils. Ris and Plaut (1962) also found similar fibrils in higher plants (e.g. *Helianthus tuberosus, Zea mays*) although in lower concentration. Chun et al. (1963) extracted total DNA from *C. reinhardthii* and subjected it to buoyant density centrifugation in caesium chloride. They observed a major band of density 1.723 g/cm^3 and a minor band of density 1.695 g/cm$^{3.}$ The cellular location of these bands was, however, unknown. Sager and Ishida (1963) isolated DNA from purified chloroplasts of *C. reinhardtii*. By comparing total *C. reinhardtii* DNA extracts with the chloroplast *C. reinhardtii* DNA extracts, Sager and Ishida (1963) observed similar bands as Chun et al.

(1963). However the minor band in the chloroplast DNA extract was much stronger. While it made up 6% of the total amount of DNA in the total DNA extracts, it made up 39% in the chloroplast DNA extracts. This was among the earliest evidence that chloroplasts contain their own DNA. Chun et al. (1963) tried to characterize the chloroplast DNA using caesium chloride buoyant density centrifugation in spinach and beet leaves. They observed three bands of different density: 1.695 g/cm^3 (predominant), 1.705 g/cm^3 (5-15%; GC content ~ 46%) and 1.719 g/cm^3 (5-40%; GC content ~ 60%). Based on earlier observations, they concluded that the predominantly occurring band corresponded to nuclear DNA and that the remaining two bands together were the chloroplast DNA. Also in 1963, Kirk carried out an analysis on the chloroplast DNA of broad beans to determine the base composition using a new chemical approach. He measured GC contents for chloroplast and nuclear DNA of 37.4% and 39.4%, respectively. These values differed very much from the values observed by Chun et al. (1963). Wells and Birnstiel (1967) extracted chloroplasts from spinach and broad beans amongst others and subjected them to a DNase treatment. They yielded, from all species analysed, only a single DNA component with a density of 1.694 - 1.695 g/cm^3. These values were in agreement with the results from Sager and Ishida (1963) for *C. reinhardthii*. The observed densities from Wells and Birnstiel (1967) were therefore in agreement with the GC content of 37.4% measured by Kirk (1963). The differences of these values to the results from Chun et al. (1963) are very likely due to a misinterpretation. While chloroplast DNA has a relatively constant GC content of 37.5±1%, and a buoyant density in caesium chloride of 1.697±0.001 g/cm^3, the GC content and the buoyant density in caesium chloride are very much dependent on the species analysed. Unfortunately in the species used by Chun et al. (1963), spinach and beet, nuclear and chloroplast DNA have the same GC content and buoyant density. Thus it is likely that the observed 1.695 g/cm^3 band is a combination of nuclear and chloroplast DNA, the observed 1.705 g/cm^3 band correspondents to the density of mitochondrial DNA and the remaining band might be due to bacterial contamination because the used leaves were shop-bought (Kirk 1986). Nevertheless, the studies by Chun et al. (1963) and Kirk (1963) were the first studies characterizing the DNA in chloroplasts of higher plants (Kirk 1986). Manning et al. (1971) reported not only the circular structure of the chloroplast DNA in *Euglena gracilis*, but also that several copies of the circular molecule exist within one chloroplast. One of the first physical maps of a chloroplast genome was published by Bedbrook and Bogorad (1976). It was

produced using restriction enzymes for the chloroplast genome of *Zea mays*. This approach also revealed, for the first time, the presence of two copies of a large inverted repeat. Shortly after one of the first physical maps was published, the DNA sequence of the *rrn16* gene in *Zea mays* became available (Schwarz & Kössel 1980). Comparisons to the 16S rRNA of *Escherichia coli* revealed very high homology and thus supported the endosymbiont hypothesis for the origin of plastids. Furthermore it was the first complete analysis of a gene from a higher plant (Schwarz & Kössel 1980). Then in 1986 the first complete chloroplast genome sequences were published. Ohyama et al. (1986) sequenced the chloroplast genome of the liverwort *Marchantia polymorpha,* a lower plant species. The complete chloroplast genome of *M. polymorpha* is 121,024 bp long and contains two copies of the inverted repeat (IR, 10,058 bp), a large single copy region (LSC, 81,095 bp) and a small single copy region (SSC, 19,813 bp). It codes for 128 possible genes including four ribosomal RNAs, 32 transfer RNAs and 55 protein coding genes. Twenty genes contained introns. In this same year, Shinozaki et al. (1986) published the complete chloroplast genome of the angiosperm *Nicotiana tabacum* (tobacco), a higher plant species. The chloroplast genome of *N. tabacum* consisted of 155,844 bp and contained two copies of the inverted repeat (25,339 bp), an LSC (86,684 bp) and an SSC (18,482 bp). *Nicotiana tabacum* chloroplast DNA encodes 146 possible genes including four ribosomal RNAs, 30 transfer RNAs and 88 protein coding genes. Fifteen genes contained introns. Both research groups detected significant homologies to the *Escherichia coli* genome which supported, once more, the endosymbiont hypothesis of plastids.

2.1.2 General features of chloroplast genomes

Since the first chloroplast genomes became publicly available in 1986 and 1989 (Ohyama et al. 1986; Shinozaki et al. 1986; Hiratsuka et al. 1989), the number of published chloroplast genomes has increased rapidly and especially in recent years (Figure 3). More than 100 chloroplast genome sequences are publicly available. Chloroplast genomes from a wide range of species have been sequenced and analysed. Thus sequences from algae, mosses, ferns, gymnosperms, and angiosperms (monocotyledonous and dicotyledonous) are available. With an increase in published sequences more information about chloroplast genomes became available.

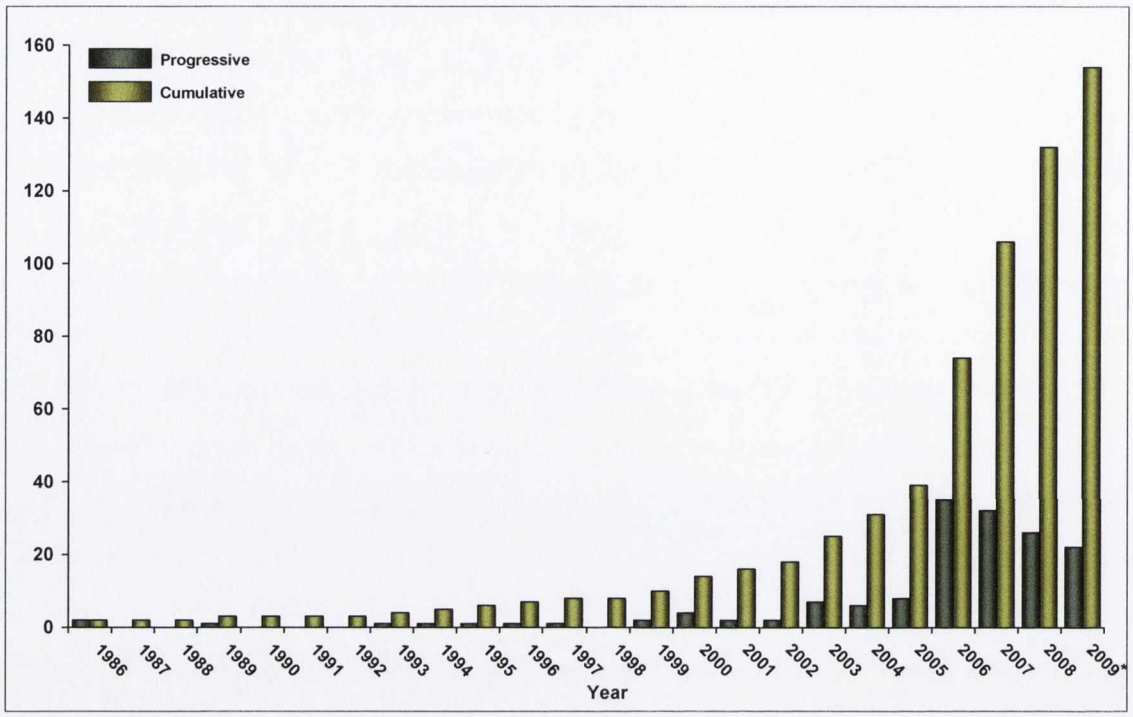

Figure 3 Number of published chloroplast genome sequences since 1986 and amount of publicly available chloroplast genome sequences in November 2009 (www.ncbi.nlm.nih.gov). *preliminary

It is now in generally accepted that chloroplasts derive from cyanobacteria (Mereschkowsky 1905, Margulis 1970, Keeling 2004). Cyanobacteria have a genome size of up to 9 Mb coding for more than 7000 protein coding genes (Timmis et al. 2004). In contrast chloroplast genomes have a considerably smaller size. The non-photosynthetic parasitic vascular plant *Epifagus virginia* has the smallest chloroplast genome detected so far with 70,028 bp (Wolfe et al. 1992) and *Pelargonium* × *hortorum* has the largest chloroplast genome detected so far with 217,942 bp (Palmer et al. 1987, Chumley et al. 2006). However, most of the chloroplast genomes have an average size of approximately 150 kb (Figure 4). The reduction in genome size from the ancestral cyanobacterium to the chloroplast genome is due to gene transfer from the ancestral genome to the genome of its host. Chloroplast genome molecules have mainly a circular structure but can also be found in linear or other forms. The molecules can be mono- or multimeric (Lilly et al. 2001). But in general the genome is represented by a monomeric circular map containing two copies of an IR region which separate an SSC region from an LSC region. This quadripartite structure is typical for most species. However, in a few species the inverted repeat region was found to be absent (*Zygnema circumcarinatum* (Turmel et al. 2005), *Staurastrum punctulatum* (Turmel et al. 2005), *Medicago truncatula* (Genbank Acc.-no.: AC093544), *Cryptomeria japonica* (Hirao et

al. 2008)), or highly reduced (*Pinus koraiensis* (Genbank Acc.-no.: AY228468) and *Pinus thunbergii* (Tsudzuki et al. 1992)).

In most angiosperm species, the chloroplast genome contains approximately 113 different genes (Sugiura 1992) that code for ribosomal and transfer RNA, ribosomal proteins, translational factors, RNA polymerase subunits, Rubisco subunits, Photosystem I and II subunits, Cytochrome b/f complex subunits, ATP synthase subunits and NADH dehyrogenase subunits (Sugiura 1992). The gene content and the gene order are generally highly conserved among angiosperm plant families (Palmer 1987, Ravi et al. 2007).

Chloroplast genomes are usually inherited maternally (Corriveau & Coleman 1988). This property makes them a useful tool for a variety of applications. Chloroplast genome information can be used for defining cytoplasmic breeding pools (McGrath et al. 2007), and for the tracking of parentage in interspecific hybrids such as in *Arabidopsis suecica* (Säll et al. 2003) or *Miscanthus* (Hodkinson et al. 2002). Genetically modifying the chloroplast genome instead of the nuclear genome is also an ideal approach for minimizing the risk of spreading transgenes into wild plants via pollen flow (Daniell et al. 1998). In comparison to nuclear genetic engineering, much higher expression of the transgenic insertion can be obtained because of the high copy number of chloroplast genomes within a single plant cell (e.g. tobacco 10,000 per cell, Daniell et al. 2001). Chloroplast genome sequences are also highly suitable for phylogenetic studies (Wu et al. 2009; Chapter 4).

2.1.3 Poaceae/monocotyledonous chloroplast genomes

In 1989 the third completely sequenced chloroplast genome, *Oryza sativa* 'japonica group' (rice), was published. It was the first complete chloroplast genome sequence of a monocotyledonous species (Hiratsuka et al. 1989). The *O. sativa* chloroplast genome was 134,525 bp long and as previously observed for *M. polymorpha* and *N. tabacum* divided into LSC (80,592 bp), SSC (12,335 bp) and the two copies of the IR (20,799 bp).

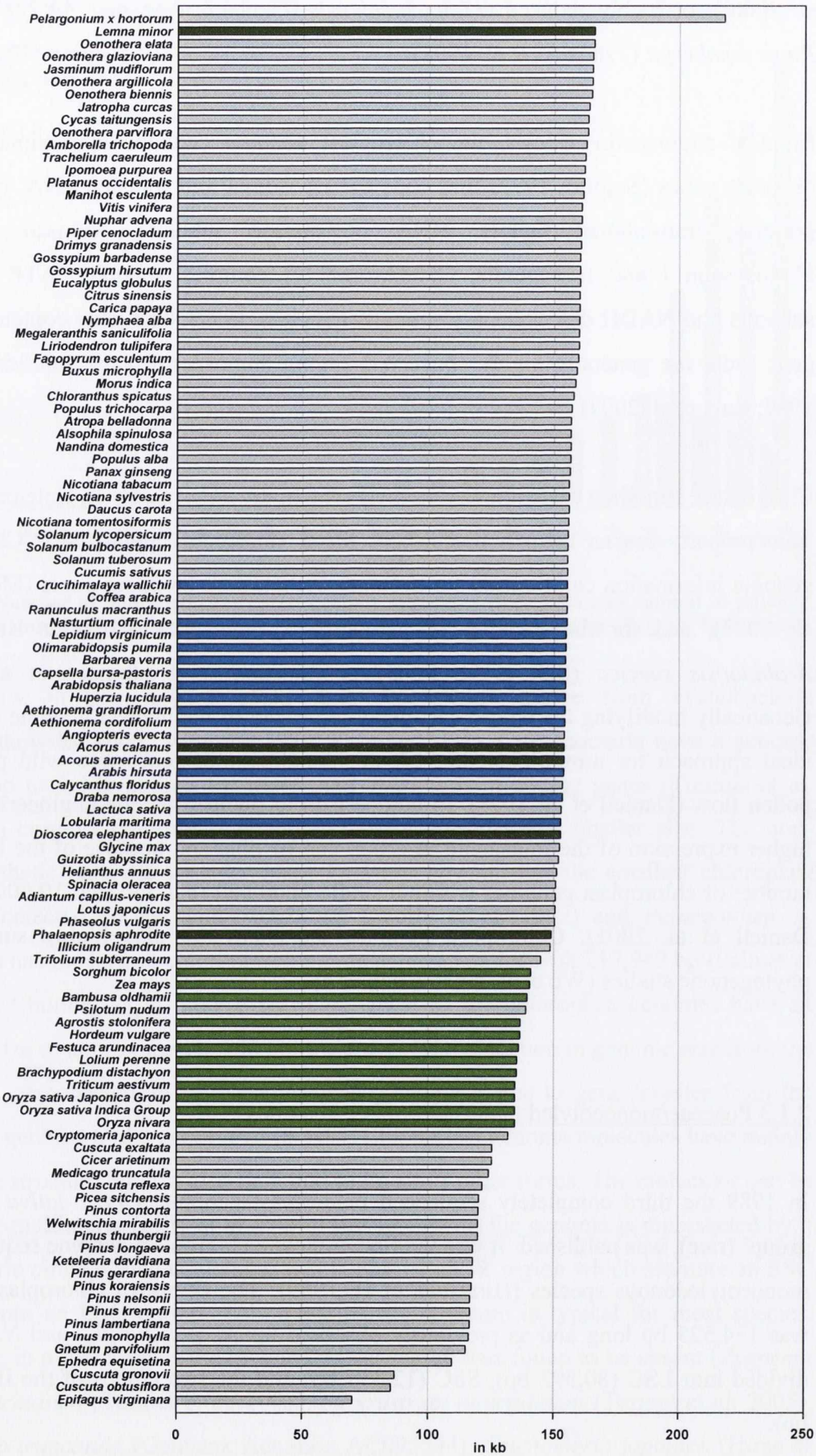

Figure 4 Lengths of chloroplast genomes from all vascular species publicly available in July 2009 (www.ncbi.nlm.nih.gov). bright green: Poaceae species; dark green: non-Poaceae monocotyledonous species, blue: Brassicaceae species (dicotyledonous).

The *O. sativa* chloroplast genome encodes 135 possible genes including four ribosomal RNAs, 30 transfer RNAs and 93 protein coding genes. Sixteen genes contained introns. Despite the similarity with *M. polymorpha* and *N. tabacum*, *O. sativa* showed some major differences. The IR region, although smaller than the one of *N. tabacum*, encompassed three additional genes (*trnH*, *rps19*, *rps15*) more. However, it also showed two major deletions of 4.8 kb and 6 kb within the IR corresponding to two very large reading frames in *N. tabacum* and *M. polymorpha*. Deletions of 1.4 kb were also observed within the LSC in *O. sativa*. Additionally major rearrangements within the LSC were detected for *O. sativa* (Hiratsuka et al. 1989). These rearrangements had also been observed for *Triticum aestivum* and *Zea mays* and thus Hiratsuka et al. (1989) concluded that they are specific for Poaceae.

Since this first published chloroplast genome of a monocotyledonous species, chloroplast genomes of 19 more monocotyledonous species were sequenced. Fourteen of those species belonged to the family Poaceae and among those are the world's most important agricultural crops such as rice (*Oryza sativa*), wheat (*Triticum aestivum*), maize (*Zea mays*) and sugarcane (*Saccharum officinarum*) (http://faostat.fao.org/). Chloroplast genomes of monocotyledonous species vary in size from 134,494 bp (*Oryza nivara*) to 165,955 bp (*Lemna minor*), although the variation within Poaceae is much smaller with *Sorghum bicolor* having the largest genome of 140,754 bp (Figure 5).

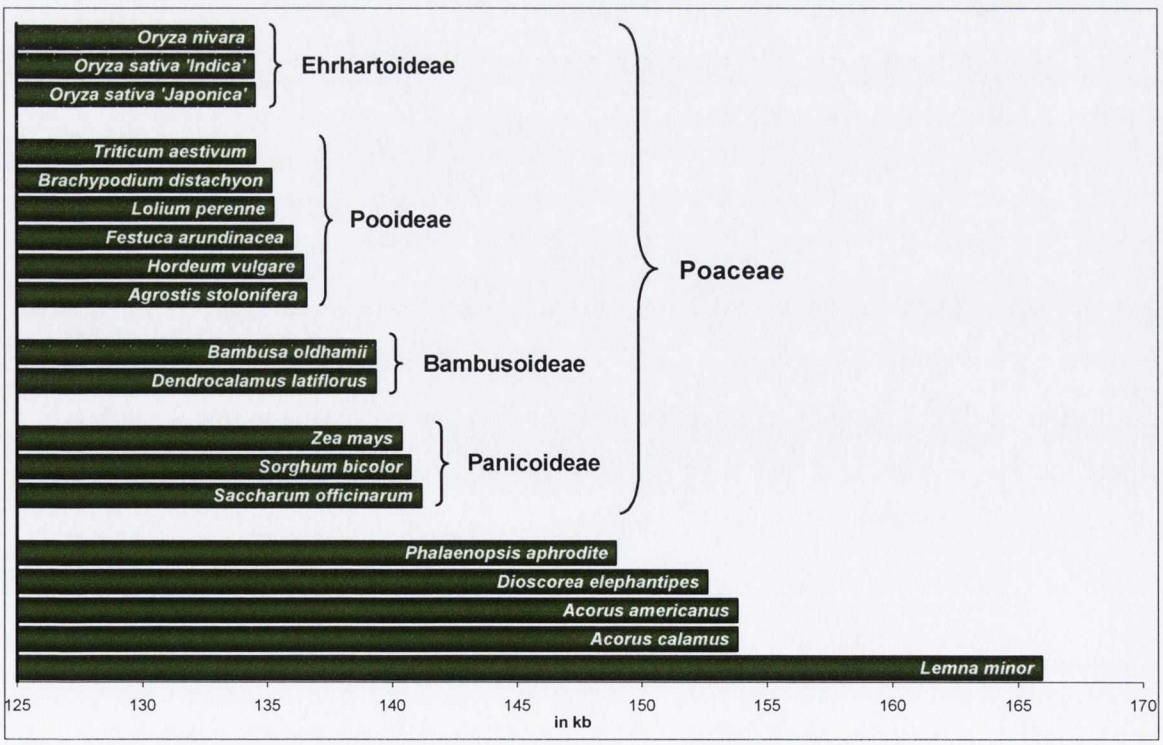

Figure 5 Lengths of monocotyledonous chloroplast genomes (www.ncbi.nlm.nih.gov).

Major differences of Poaceae chloroplast genomes to other chloroplast genomes are the losses of the introns in the genes *clpP* and *rpoC1* and the losses of the genes *accD, ycf1* and *ycf2*. The loss of these three genes are very likely the three major deletions Hiratsuka et al. (1989) observed when first comparing *O. sativa* to *N. tabacum* and *M. polymorpha*. The function of the genes *ycf1* and *ycf2* is still not known but evidence was found that they are essential in dicotyledonous chloroplast genomes (Drescher et al. 2000). Generally it is believed that the loss of genes from the chloroplast genome results into a transfer to the nuclear genome, however, the nuclear homologs to *ycf1* and *ycf2* have not been identified yet (Drescher et al. 2000). *accD* is the gene that produces the prokaryotic form of Acetyl-CoA carboxylase which catalyzes the first step of fatty acid synthesis. Its absence from Poaceae chloroplast genomes results in the susceptibility of grasses to graminicides (Konishi et al. 1996). The loss of these three genes explains most of the length variation found in comparison to dicotyledonous chloroplast genomes (Figure 4).

The Poaceae chloroplast genome itself is otherwise highly conserved regarding genome size, gene content and gene order (Wu et al. 2009). The only known exception to date is the chloroplast genome of *Festuca arundinacea* which surprisingly lacks intact copies of four genes (*psbF*, *rps14*, *rps18, ycf4*; Genbank Acc.-No. FJ466687).

2.1.4 Aims and objectives

In October 2006, when this study started, only six Poaceae chloroplast genomes had been sequenced and published: *Oryza nivara, O. sativa 'indica', O. sativa 'japonica', Triticum aestivum, Saccharum officinarum and Zea mays*. These six taxa were all cereals. However, no forage grass species had been sequenced. Therefore the major objective of this study was to make the first complete chloroplast genome sequence of a Poaceae forage species, *L. perenne*, publicly available. Although genomic libraries of *L. perenne* exist, they were not accessible to us for a screening for chloroplast DNA clones. Also the use of a primer walking strategy using consensus chloroplast DNA primers was not possible because the primers had not been made publicly available, yet. Hence, to facilitate sequencing the first chloroplast genome of Poaceae forage species, pure chloroplast DNA had to be extracted, sequenced, assembled and the genes

annotated. This approach was previously already used for other species as for example *A. thaliana*. Several previously published protocols were tested for their application in *L. perenne* chloroplast DNA isolation (Palmer 1986; Bellaoui 1997; Triboush 1998) however, the chloroplast DNA quantities yielded were too small and it was necessary to modify and optimise existing protocols to extract *L. perenne* chloroplast DNA. Following the annotation of the complete *L. perenne* chloroplast genome, a thorough study of variation within Poaceae chloroplast genomes was planned to be carried out to find highly variable regions. Another objective was to evaluate the molecular diversity within the chloroplast genome of the cross-pollinating species *L. perenne*. Few studies have examined variation of the chloroplast genomes within a population of a species. However, McGrath et al. (2007) discovered more than 500 haplotypes within 1,575 individual plants of *Lolium*, *Festulolium* and *Festuca* populations. Therefore the aim was to assess the chloroplast genome variation within a *Lolium perenne* cultivar. It was anticipated that new highly variable DNA regions would be revealed that are suitable for the design of new markers to assess cytoplasmic breeding pools and to add to population genetic and phylogenetic studies.

2.2 Material and Methods

2.2.1 Plant material

For the isolation of *L. perenne* chloroplast DNA, material was obtained from cultivar Cashel that was sown in soil in the green house. The plants were grown under natural daylight conditions from October to December 2006 in an unheated greenhouse. For each isolation experiment 200 g of leaves of young plants, not older than 6 weeks, were harvested. *Lolium perenne* is an outbreeding species, therefore the harvested leaves represent a mix of genotypes.

2.2.2 Preparations of buffers and materials before isolation

At least one day prior to the chloroplast DNA isolation the sucrose cushions and the sucrose gradients were prepared to achieve a slightly diffuse band in the interphase of the layers. For the sucrose cushions, 16 centrifuge tubes (15 ml; Falcon) were filled with 5 ml of 30 % sucrose solution (in 50 mM Tris-HCl pH 8.0, 25 mM EDTA) and stored on ice until further use. For the gradients, six 50 ml centrifuge tubes (Falcon) were prepared. 18 ml of 52 % sucrose solution (in 50 mM Tris-HCl, pH 8.0, 25 mM EDTA) was transferred into each of them. Subsequently, 7 ml of 30 % sucrose solution was layered carefully on top of the 52% solution. The tubes containing the gradients were stored on ice until further use.

2.2.3 Isolation of chloroplast DNA

Preparation of plant material

All the following steps (Figure 6) were carried out at 4 °C if not otherwise stated.
The harvested young leaves were cut in pieces smaller than 1 cm length. For each 100 g of leaf material, 400 ml of ice-cold buffer 1 (1.25 M NaCl, 50 mM Tris-HCl pH 8.0, 7 mM EDTA, 5 % PVP-40, 0.1 % BSA, 1 mM DTT; Bookjans 1984) were added. The mixture was homogenized using an Ultra-Turrax T 25 basic (IKA -WERKE GmbH & Co. KG, Staufen, Germany) homogenizer. Homogenizations started at 6,500 rpm (step

1 on the homogenizer) and the speed was increased to 13,500 rpm (step 3 on the homogenizer) until the majority of leaf pieces were in solution. The homogenate was filtered once through four layers of miracloth (Merck KGaA, Darmstadt, Germany) with squeezing and once without squeezing. The filtrate was divided into four centrifuge bottles (250 ml), balanced and centrifuged for three minutes at 1,000 rpm (~ 200 *g*) in a Sorvall RC 5B plus (Thermo Fisher Scientific, Waltham, USA) using a fixed angle rotor (SLA 1500, Thermo Fisher Scientific, Waltham, USA). The function of this step was to pellet nuclei and whole cells. This step was determined to be optional since later extractions were done without it and were similarly successful.

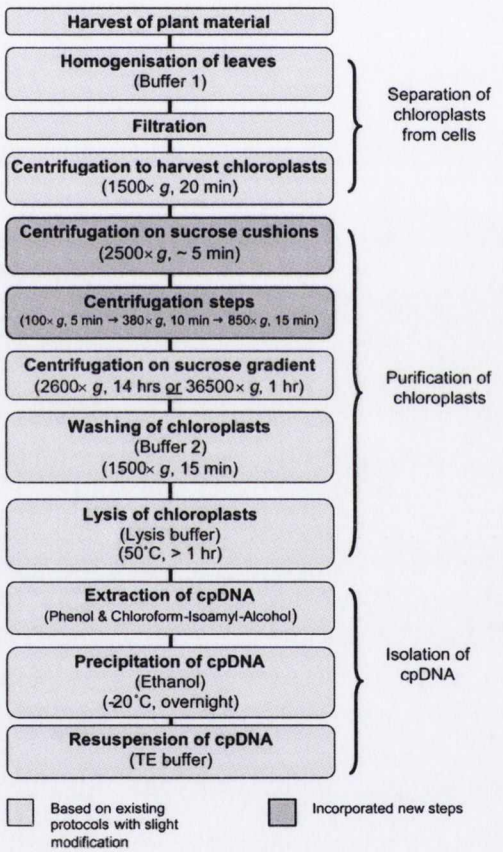

Figure 6 Overview over the major steps for the isolation of high purity chloroplast (cp) DNA. 'Slight modification' refers to the adjustment of some steps due to the use of different laboratory equipment. Highlighted in dark gray are two newly incorporated steps that were not included in the existing protocols.
Diekmann et al. (2008) with modifications.

The supernatant was transferred into four new centrifuge bottles (250 ml), balanced and centrifuged for 20 minutes at 3,000 rpm (~ 1,366 *g*) in the Sorvall RC 5B plus (Thermo Fisher Scientific, Waltham, USA) using the fixed angle rotor (SLA 1500, Thermo Fisher Scientific, Waltham, USA). The supernatant was discarded and each resulting pellet was carefully re-suspended in 100 ml buffer 1 using a soft paintbrush. The

solution was again centrifuged for 20 minutes at 3,000 rpm (~ 1,366 *g*) in the Sorvall RC 5B plus (Thermo Fisher Scientific, Waltham, USA) using the fixed angle rotor (SLA 1500, Thermo Fisher Scientific, Waltham, USA). The supernatant was discarded. For subsequent extractions when the first centrifuge step was omitted, this centrifugation was repeated once more.

Sucrose cushions

Each pellet was dissolved in 25 ml wash buffer (0.35 M Sorbitol, 50 mM Tris-HCl pH 8.0, 25 mM EDTA; Palmer 1986). The dissolved pellets were combined in one centrifuge bottle. Then 6 ml of the solution was layered carefully on top of each of the sucrose cushions. The tubes were balanced and centrifuged for five minutes at 3,420 rpm (~ 2,500 *g*) in a Sorvall Legend RT (Thermo Fisher Scientific, Waltham, USA) in the swinging bucket rotor. The supernatant was discarded and each of the resulting pellets was washed in 5 ml of wash buffer. They were again centrifuged for five minutes at 3,420 rpm (~ 2,500 *g*) in the Sorvall Legend RT (Thermo Fisher Scientific, Waltham, USA) with the swinging bucket rotor. The supernatant was discarded. Eight of the pellets were re-suspended in 5 ml wash buffer each. To reduce the number of tubes, a re-suspended pellet was combined with a tube of a non-re-suspended pellet. The number of tubes was halved. The tubes were centrifuged for three minutes at 3,420 rpm (~ 2,500 *g*) in the Sorvall Legend RT (Thermo Fisher Scientific, Waltham, USA) in the swinging bucket rotor. Extended centrifugation time resulted in very dense hardly solvable pellets. The supernatant was discarded. Each of the resulting pellets was re-suspended in 1.5 ml wash buffer and transferred into a 2 ml microcentrifuge tube. These tubes were centrifuged in a Hettich Mikro 200 R centrifuge (Andreas Hettich GmbH & Co.KG, Tuttlingen, Germany) for five minutes at 1,000 rpm (~ 100 *g*). The supernatant was transferred into new tubes which were then centrifuged in the same centrifuge for ten minutes at 2,000 rpm (~ 380 *g*). The supernatant was then discarded. Each of the resulting pellets was dissolved in 1 ml of wash buffer. The aim of this step was to obtain a purer chloroplast solution without whole cells, cell debris and nuclei. The success of this procedure was checked using a light microscope (400 x magnified). Both resulting pellets contained chloroplasts, but the pellet after the second centrifugation was much purer.

When this protocol was used with bigger amounts of starting leaf material the concentration of chloroplasts was higher and the centrifugation of the sucrose cushions took more time (~ 20 minutes). The resulting pellets were not divided into microcentrifuge tubes but left in the 15 ml centrifuge tubes. The first centrifugation step was then carried out at 670 rpm (~ 100 *g*) for five minutes using the Sorvall Legend RT (Thermo Fisher Scientific, Waltham, USA) with its swing out bucket rotor. The second step was carried out at 1,330 rpm (~ 380 *g*) for ten minutes in the same centrifuge with the same rotor. An additional (third) centrifugation step was necessary at 2,000 rpm for 15 minutes in the same centrifuge. These pellets were dissolved in 2 ml wash buffer. The purest pellets were obtained from the two last centrifugations.

Sucrose gradients

For both amounts of starting material, the volume of two of the tubes (200 g leaf tissue = 2 ml; more leaf tissue = 4 ml) were carefully layered on top of one sucrose gradient. The gradients were centrifuged for 14.5 hours at 3,500 rpm (~ 2,600 *g*) in the Sorvall Legend RT (Thermo Fisher Scientific, Waltham, USA) in the swinging bucket rotor. If an ultracentrifuge with a SW-27 rotor (Beckman Coulter, Inc., Fullerton, California, USA) was available centrifugation was done for 60 minutes at 25,000 rpm (~112,500 *g*). The aim of this step was to purify the chloroplasts more thoroughly and separate them from nuclei and cell debris. The success of the sucrose gradient purification step was observed using light microscopy (400x magnified). At the interphase of the 52% to 30% sucrose layers the chloroplast band was built. The chloroplast band was carefully removed using a wide bore one way plastic pipette. The chloroplast bands of two such gradients were united in one new 50 ml centrifuge tube. The chloroplast solution was then diluted with wash buffer up to 45 ml. It was centrifuged again for 15 minutes at 2,644 rpm (~ 1,500 *g*) in the Sorvall Legend RT (Thermo Fisher Scientific, Waltham, USA) in the swing out bucket rotor. The supernatant was discarded. Each pellet was resuspended in 10 ml buffer 2 (10 mM Tris-HCl pH 8.0; 5 mM EDTA; 100 µg/ml Proteinase K) and centrifuged again for 15 minutes at 2,644 rpm (~ 1,500 *g*). This step was repeated a second time.

Lysis of chloroplasts

The resulting pellets were dissolved in 2 ml of buffer 2 each. To each tube, 1/10 volume of Proteinase K (200 µl) was added. After two minutes at room temperature, 1/5 volume of lysis buffer (20% sodiumsarcosinate, 50 mM Tris-HCl pH 8.0, 25 mM EDTA; Jansen 2005) was added. Then the tubes were incubated at 50 °C for at least 1 hour.

Extraction of chloroplast DNA

The chloroplast DNA was extracted adding an equal volume of saturated phenol (pH 8.0). The solution was well mixed and centrifuged at 3,500 rpm (~ 2,600 *g*) for 20 minutes. The upper clear aqueous phase was transferred into new tubes. An equal volume of chloroform-isoamyl-alcohol was added and the centrifugation was repeated. This step was done twice. Afterwards the aqueous phase was transferred in to new tubes and 2.5 volumes of 96 % ethanol were added. The tubes were stored either for 30 minutes at -80 °C or over night at -20 °C. The content was divided to fit into 2 ml microcentrifuge tubes and centrifuged for one hour at 10,000 rpm (~ 9,500 *g*) in a Hettich Mikro 200 R centrifuge (Andreas Hettich GmbH & Co.KG, Tuttlingen, Germany). The supernatant was placed in a new tube and centrifuged again for one hour at 10,000 rpm in the Hettich Mikro 200 R centrifuge (Andreas Hettich GmbH & Co.KG, Tuttlingen, Germany). The resulting pellets were washed twice in 500 µl 70% ethanol and were spun for 20 minutes at 10,000 rpm.

If a larger amount of leaf material was used, at the start the 96 % ethanol mixture was transferred into centrifuge tubes and centrifuged at 13,000 rpm (~20,000 *g*) in a Sorvall RC 5B plus (Thermo Fisher Scientific, Waltham, USA) using a fixed angle rotor (SS 34, (Thermo Fisher Scientific, Waltham, USA) for 15 minutes. The tubes were placed in a laminar flow hood to allow the remaining alcohol to evaporate. Then the pellets were dissolved in 200 µl TE (1x) and stored at -20 °C until further use.

If an ultracentrifuge was availabe for the chloroplast DNA extraction, it was possible to do caesium chloride gradients (4.5 g of caesium chloride, 60 µl ethidium bromide, 0.3 ml 1.0 M Tris-HCl pH 8.0, 0.5 M 0.3 ml EDTA for each 1.8 ml chloroplast solution) instead of the phenol-chloroform-isoamyl steps. The chloroplast-caesium chloride-solution was transferred into ultracentrifuge tubes (Quick-Seal Polyallomer Tubes,

Beckman, Part.-no. 342412) and centrifuged for 16 hours at 58,000 rpm in a near vertical NVT 90 rotor (Beckman Coulter, Inc., Fullerton, California, USA) in a Beckman Ultracentrifuge Floor Model L5-65 or an Optima L-XP (Beckman Coulter, Inc., Fullerton, California, USA) centrifuge. The chloroplast DNA band was visualised under UV light and found to be positioned in the middle of the tubes. It was extracted with a syringe and needle. The syringe needle was carefully stabbed below the band and then the band was withdrawn from the tube. To remove the ethidium bromide and the caesium chloride from the chloroplast DNA, a protocol from Sambrook and Russel (2001) was followed. An equal volume of water-saturated n-butanol was given to each sample. The tubes were well shaken and centrifuged for three minutes at 1,440 rpm (~ 450 *g*) in the Sorvall Legend RT (Thermo Fisher Scientific, Waltham, USA) with the swinging bucket rotor. The upper pink phase was discarded. These steps were repeated until both phases had lost their pink colour. To remove the caesium chloride three volumes of water were added and the solution was well mixed. Afterwards 8 volumes (1 volume = volume of solution prior to dilution) of 96% ethanol were added and again well mixed. The mixture was stored for at least 15 minutes at 4 °C. The tubes were centrifuged at 13,000 rpm (~20,000 *g*) in a Sorvall RC 5B (Thermo Fisher Scientific, Waltham, USA) plus using a fixed angle rotor (SS 34, Sorvall) for 15 minutes. The pellets were washed once with 70% ethanol and the centrifugation was repeated. The tubes were placed in the laminar flow hood to allow the remaining alcohol to evaporate. Then the pellets were dissolved in 200 µl TE (1x) and stored at -20 °C until further use.

2.2.4 Whole genome amplification

To see if the chloroplast DNA isolation was successful, 3 µl of each of the resulting samples were mixed with 5 µl ddH_2O and 3 µl loading buffer. 50 ml of a 1% Agarose 0.5X TBE gel was prepared containing 3 µl ethidium bromide (10mg/µl). The gel was loaded and run at 60 V for an hour. To enable sequencing, the amount of yielded chloroplast DNA had to be amplified (Figure 7). Thus 1 µl of each successful chloroplast DNA sample was used for a whole genome amplification using a GenomePhi DNA amplification kit (Amersham Biosciences, Prod.-no.: 25-6600-01). For the preparation of the samples, the manufacturer's instructions were followed.

To control the success of the isolation and amplification, 2 µl of each amplified sample were digested with the restriction enzymes EcoRI and HindIII (2 µl amplified sample, 1 µl enzyme, 2 µl buffer, 15 µl ddH_2O) over night. The results of the restriction digest were visualized on a 0.5 x TBE 1% Agarose gel. The gel was run for 2 hours at 120 V (Figure 8).

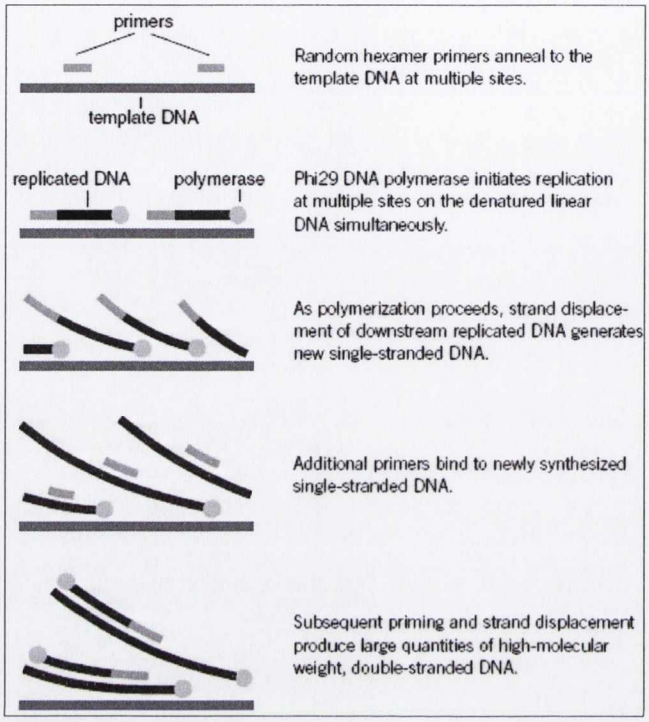

Figure 7 Schematic diagram of the GenomiPhi amplification process (Amersham Biosciences Corp 2003).

2.2.5 Sequencing, assembling and annotating the chloroplast genome

Successfully amplified *L. perenne* chloroplast DNA samples were sequenced using a shotgun sequencing approach. The DNA sample is sheared mechanically and the fragments are cloned into a vector. The DNA fragments are then sequenced using the chain-terminating method and vector specific primers that flank the insert region (Sanger et al. 1977, Sanger et al. 1980). Shotgun sequencing was sourced to a company (GATC Biotech AG, Konstanz, Germany) from which 2,179 read and trace files were obtained. Preassemblies were also carried out with the programs Lasergene (DNAstar, Inc., Madison, Wisconsin) and PHRAP. None of the received preassemblies produced one large complete contiguous sequence (contig) but instead they produced several smaller contigs. Thus, single read files were checked for their chloroplast DNA content. After detection of read files containing nuclear DNA and mitochondrial DNA, all read

files were checked using the blastn program (http://www.ncbi.nlm.nih.gov/blast/). Thus, 1,581 read files consisting of chloroplast DNA were detected. The final assembly was based on the PHRAP based contiguous sequences and done in comparison to the chloroplast genome of bentgrass (*Agrostis stolonifera*, Saski et al. 2007, Figure 9).

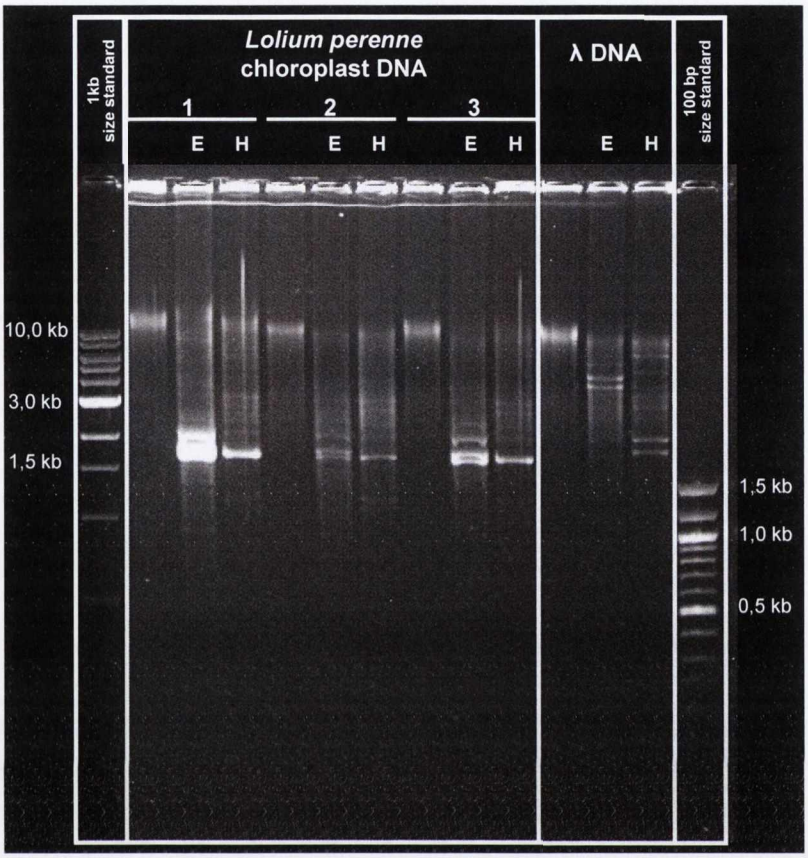

Figure 8 Agarose gel of whole genome amplified *Lolium perenne* chloroplast DNA. 170 ml 1% Agarose gel stained with 1% Ethidium bromide (10mg/µl) run for ~ 2 hours at 130 V. First lane of each sample undigested, lane E= EcoRI (New England Biolabs, Massachusetts, USA) digested, lane H= HindIII (New England Biolabs, Massachusetts, USA) digested. Size standard: 5 µl of mi-1 kb DNA Marker Go (metabion international AG, Martinsried, Germany).

Figure 9 shows that contigs 95, 96, 99, 100 and 101 combined present the complete chloroplast genome of *L. perenne*. Visual inspection of the contig sequences showed that the ends of these contigs were based on tracefiles with low quality, but that all contigs in general had an overlap of several basepairs. Thus this overlap was sought visually by comparing the end of one contig with the start of the next contig. Once found, the low quality sequences were deleted and the contigs combined. Three genome regions were observed with low genome coverage, therefore primers were designed to re-sequence these regions and confirm results (*trnL-trnF*: forward primer (FP): AGTTGTGAGGGTTCAAGTCC / reverse primer (RP): GAACTGGTGACACGAGG ATT; *atpB*: FP: GTTCGTTGCCAACAATCCTA / RP: AGGTAGCTCTAGTCTAT

GGC; *atpB-rbcL*: FP: TGTGGAAGATCTGTGCCTAC / RP: GCTGAGGAGT TACTCGGAAT). Annotation was based on the two online available programs DOGMA (http://dogma.ccbb.utexas.edu/) and tRNA-Scan SE (http://lowelab. ucsc.edu/tRNAscan-SE/) using the default settings. Intron positions were determined after Sugita and Sugiura (1996). The circular chloroplast genome map was produced using the GenomeVx program (Conant & Wolfe 2008).

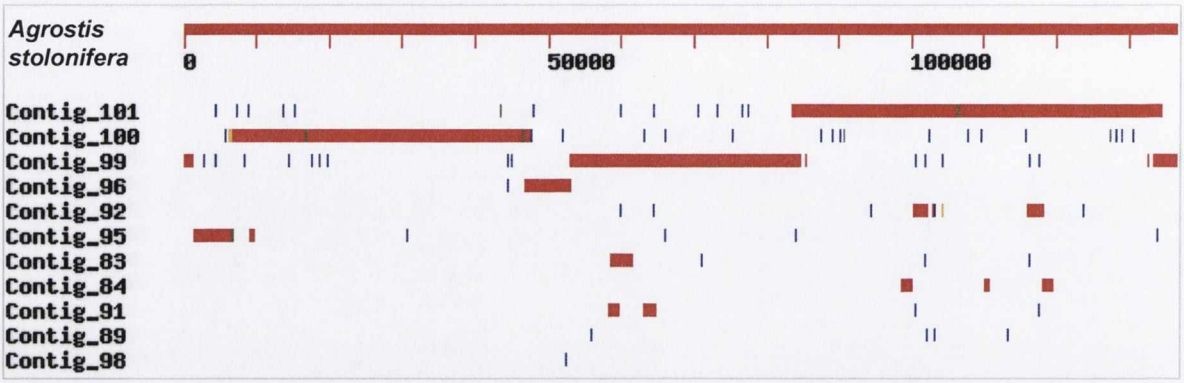

Figure 9 Alignment of contiguous *Lolium perenne* chloroplast genome sequences (contigs) with the chloroplast genome of *Agrostis stolonifera*. Long bar in red at the top of the figure represents the complete *A. stolonifera* chloroplast genome in linear form, numbers below are basepairs. Smaller bars below the *A. stolonifera* reference bar highlight the similarity of the different contigs in comparison to *A. stolonifera* – black: < 40 bp, blue: 40 – 50 bp, green: 50-80, pink: 80-200, red: >200 bp.

2.2.6 Variation analysis

Single nucleotide polymorphisms (SNPs) and insertion/deletion events (indels) within the *L. perenne* chloroplast genome were detected visually by examining the alignment of the read and trace files (Figure 10) in Lasergene (DNAstar, Inc., Madison, Wisconsin). Only SNPs/Indels with very high quality trace files were taken into account. Analysis of changes in the aminoacid sequence for SNPs that were located in coding regions was carried out in MEGA3.1 (Kumar et al. 2004). Differences between the published Poaceae chloroplast genomes were analysed in Excel using gene, intergenic spacer (IGS) and intron lengths which were extracted from the published sequences (*Agrostis stolonifera*: EF115543; *Brachypodium distachyon*: EU325680; *Hordeum vulgare*: EF115541; *Oryza nivara:* AP006728; *Oryza sativa indica:* AY522329; *Oryza sativa japonica:* X15901; *Sorghum bicolor:* EF115542; *Saccharum officinarum:* AP006714; *Triticum aestivum:* AB042240; *Zea mays:* X86563).

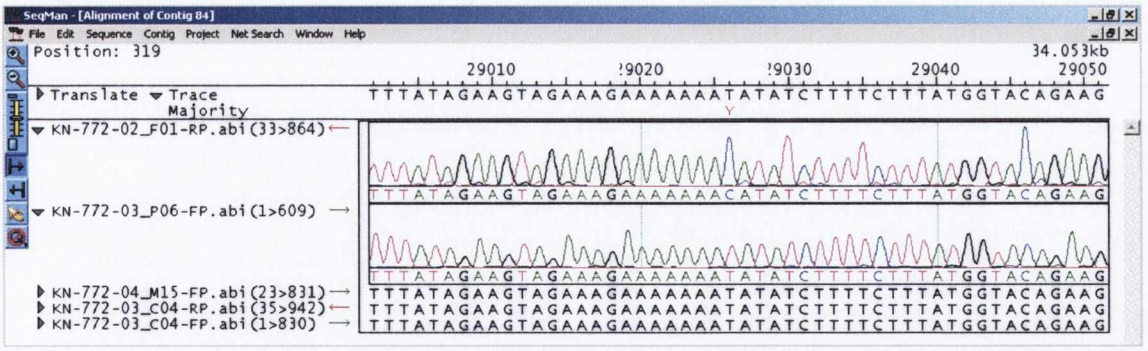

Figure 10 Observed single nucleotide polymorphism in the *trnF-ndhJ* intergenic spacer region of the *Lolium perenne* chloroplast genome. Polymorphism is highlighted with a 'Y' below the consensus sequence at the top of the figure.

2.3 Results

2.3.1 The chloroplast genome of *Lolium perenne*

The chloroplast genome of *Lolium perenne* is 135,282 bp long and has a quadripartite structure of a LSC region (79,972 bp) and a SSC region (12,428 bp) which are separated from each other by two copies of an IR (each 21,441 bp). Eightfold coverage for the *L. perenne* chloroplast genome could be achieved and a GC content of 38% was detected. The genome codes for 128 genes of which 15 contain introns and 18 are duplicated in the inverted repeat region (Figure 11, Table 1). Seventy-six genes are encoding proteins, 30 transfer RNAs and 4 ribosomal RNAs (Table 1). The chloroplast genome sequence of *L. perenne* is deposited at the European Molecular Biology Laboratory (EMBL) under Accession number AM777385.

2.3.2 Indel/SNP analysis

A total of 10 indels (Table 2) and 40 SNPs (Table 3) were found to be polymorphic among the sequencing reads. All indels were located in intergenic regions. Indels occurred in microsatellite regions, resulting in both shortening (one occurrence) and lengthening (nine occurrences) of the sequenced region compared with the length that was observed in the majority of the trace files. Nineteen SNPs were found within IGS regions and introns and 21 within coding regions (Table 3).

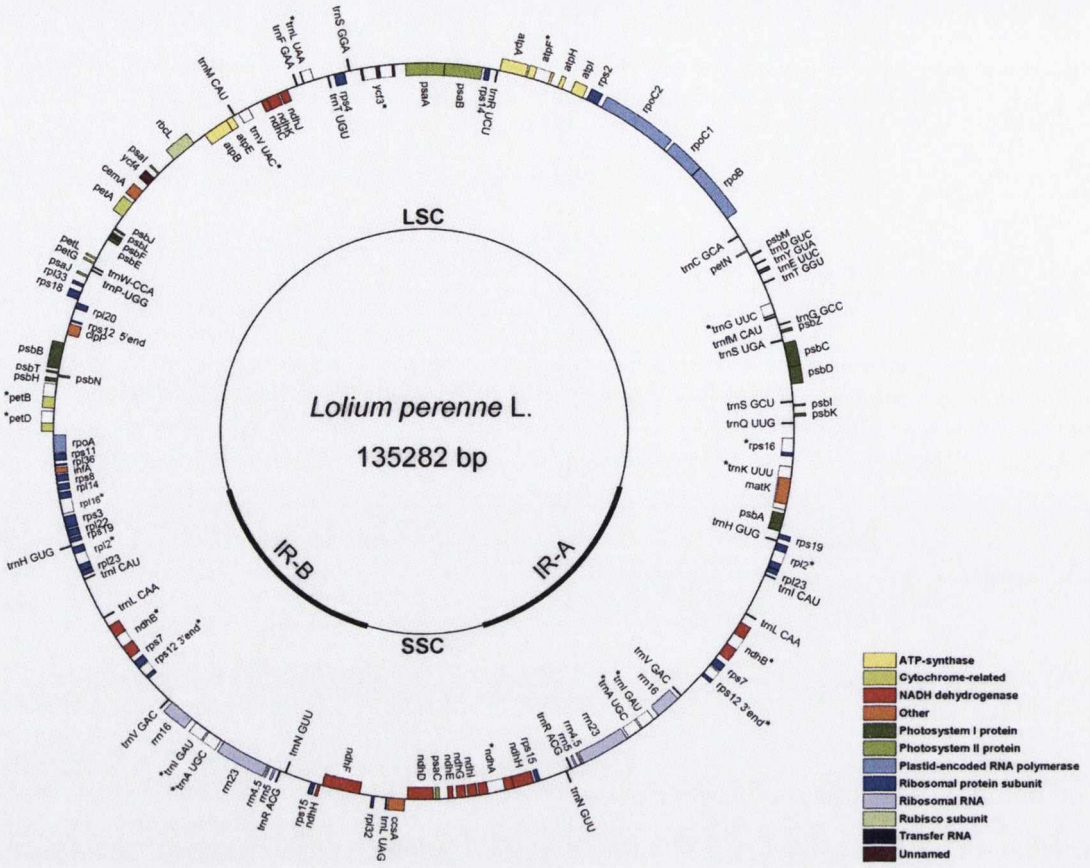

Figure 11 Circular structure of the chloroplast genome of *Lolium perenne*. Genes written on the outside are transcribed clockwise, genes on the inside counter-clockwise, annotated genes are colour coded according to their function, genes containing introns are highlighted with an asterisk; LSC=large single copy region, SSC=small single copy region, IR=inverted repeat.

Most of the SNPs (Figure 12) are due to transition (changes within purines/ pyrimidines) mutations (20 A↔G and 8 C↔T), with 12 transversions (changes between purines and pyrimidines). Closer analysis of the SNPs found at position 100,655 and 100,656 (*trnN-rps15* IGS) revealed that they are caused by a tiny inversion of two nucleotides which are flanked by an IR of 29 bp length forming a stable hairpin secondary structure (Figure 13). The small inversion of TG within the *trnN-rps15* region in the IR is found in 13 out of the 29 trace files covering the region. Another small inversion that was supported by only one sequence read caused SNPs at positions 18, 20, 21 and 23 (*rps19-psbA* IGS). This inversion spans six nucleotides (TTCTAG) that are flanked by an IR of 25 bp length (Figure 13).

Table 1 Gene annotation of the chloroplast genome of *Lolium perenne*.

Start	End	Gene Name	Strand	Exon		Start	End	Gene Name	Strand	Exon
87	1148	*psbA*	-		LSC	43519	43642	*ycf3*	-	Exon
1383	1417	*trnK-UUU*	-	Exon		44239	44325	*trnS-GGA*	+	
3930	3966	*trnK-UUU*	-	Exon		44591	45196	*rps4*	-	
1715	3250	*matK*	-			45516	45588	*trnT-UGU*	-	
4545	4774	*rps16*	-	Exon		77244	77645	*rpl16*	-	Exon
5613	5652	*rps16*	-	Exon		78513	78521	*rpl16*	-	Exon
6467	6538	*trnQ-UUG*	-			78687	79406	*rps3*	-	
6882	7067	*psbK*	+			79479	79922	*rpl22*	-	
7482	7592	*psbI*	+							
7714	7801	*trnS-GCU*	-		Inverted repeat region (IR)	80001	80282	*rps19*	-	
8783	9844	*psbD*	+			80419	80492	*trnH-GUG*	+	
9792	11213	*psbC*	+			80547	80977	*rpl2*	-	Exon
11369	11456	*trnS-UGA*	-			81641	82031	*rpl2*	-	Exon
11812	12000	*psbZ*	+			82050	82331	*rpl23*	-	
12282	12352	*trnG-GCC*	+			82506	82579	*trnI-CAU*	-	
12803	12876	*trnfM-CAU*	-			85031	85111	*trnL-CAA*	-	
12967	13014	*trnG-UCC*	-	Exon		85644	86399	*ndhB*	-	Exon
13694	13716	*trnG-UCC*	-	Exon		87115	87891	*ndhB*	-	Exon
14891	14962	*trnT-GGU*	+			88191	88661	*rps7*	-	
15499	15571	*trnE-UUC*	+			88715	88743	*rps12_3end*	-	Exon
15633	15716	*trnY-GUA*	+			89284	89515	*rps12_3end*	-	Exon
16080	16153	*trnD-GUC*	+			91161	91232	*trnV-GAC*	+	
16627	16731	*psbM*	+			91462	92953	*rrn16*	+	
17010	17099	*petN*	-			93269	93305	*trnI-GAU*	+	Exon
18016	18086	*trnC-GCA*	-			94113	94147	*trnI-GAU*	+	Exon
19271	22501	*rpoB*	+			94213	94250	*trnA-UGC*	+	Exon
22539	24569	*rpoC1*	+			95062	95096	*trnA-UGC*	+	Exon
24773	29173	*rpoC2*	+			95242	98129	*rrn23*	+	
29460	30170	*rps2*	+			98226	98320	*rrn4.5*	+	
30425	31168	*atpI*	+			98548	98668	*rrn5*	+	
31738	31983	*atpH*	+			98897	98970	*trnR-ACG*	+	
32446	32590	*atpF*	+	Exon		99223	99295	*trnN-GUU*	-	
33467	33873	*atpF*	+	Exon		100835	101107	*rps15*	+	
33965	35488	*atpA*	+			101237	101416	*ndhH*	+	
35590	35661	*trnR-UCU*	-							
36041	36352	*rps14*	-		Small single copy region (SSC)	101511	103736	*ndhF*	-	
36513	38717	*psaB*	-			104439	104618	*rpl32*	+	
38743	40995	*psaA*	-			105281	105360	*trnL-UAG*	+	
41651	41809	*ycf3*	-	Exon		105438	106397	*ccsA*	+	
42531	42760	*ycf3*	-	Exon		106540	108048	*ndhD*	-	
46413	46447	*trnL-UAA*	+	Exon		108168	108413	*psaC*	-	
46997	47046	*trnL-UAA*	+	Exon		108922	109227	*ndhE*	-	
47395	47467	*trnF-GAA*	+			75572	75685	*rpl36*	-	
48053	48532	*ndhJ*	-			75787	76110	*infA*	-	
48632	49372	*ndhK*	-			76193	76603	*rps8*	-	
49363	49725	*ndhC*	-			76750	77121	*rpl14*	-	
50569	50603	*trnV-UAC*	-	Exon		109440	109970	*ndhG*	-	
51209	51247	*trnV-UAC*	-	Exon		110086	110628	*ndhI*	-	
51434	51506	*trnM-CAU*	+			110724	111262	*ndhA*	-	Exon
51621	52034	*atpE*	-			112278	112827	*ndhA*	-	Exon
52031	53527	*atpB*	-							
54321	55754	*rbcL*	+		Inverted repeat region (IR)	112829	114010	*ndhH*	-	
56938	57048	*psaI*	+			114140	114412	*rps15*	-	
57357	57914	*ycf4*	+			115952	116023	*trnN-GUU*	+	
58374	59066	*cemA*	+			116277	116350	*trnR-ACG*	-	
59289	60251	*petA*	+			116579	116699	*rrn5*	-	
61047	61169	*psbJ*	-			116927	117021	*rrn4.5*	-	
61333	61449	*psbL*	-			117117	120004	*rrn23*	-	
61472	61591	*psbF*	-			120151	120185	*trnA-UGC*	-	Exon
61602	61853	*psbE*	-			120997	121034	*trnA-UGC*	-	Exon
63134	63229	*petL*	+			121100	121134	*trnI-GAU*	-	Exon
63403	63516	*petG*	+			121942	121978	*trnI-GAU*	-	Exon
63639	63712	*trnW-CCA*	-			122294	123785	*rrn16*	-	
63853	63926	*trnP-UGG*	-			124015	124086	*trnV-GAC*	-	
64264	64392	*psaJ*	+			125732	125963	*rps12_3end*	+	Exon
64825	65025	*rpl33*	+			126504	126532	*rps12_3end*	+	Exon
65324	65794	*rps18*	+			126586	127056	*rps7*	+	
65957	66316	*rpl20*	-			127356	128132	*ndhB*	+	Exon
67014	67127	*rps12_5end*	-			128848	129603	*ndhB*	+	Exon
67269	67919	*clpP*	-			130136	130216	*trnL-CAA*	+	
68457	69983	*psbB*	+			132668	132741	*trnI-CAU*	+	
70152	70268	*psbT*	+			132916	133197	*rpl23*	+	
70317	70448	*psbN*	-			133216	133607	*rpl2*	+	Exon
70552	70773	*psbH*	+			134271	134700	*rpl2*	+	Exon
70903	70908	*petB*	+	Exon		134755	134828	*trnH-GUG*	-	
71655	72296	*petB*	+	Exon		134965	135246	*rps19*	+	
72487	72494	*petD*	+	Exon						
73172	73646	*petD*	+	Exon						
73857	74882	*rpoA*	-							
74947	75378	*rps11*	-							

Large single copy region (LSC)

green = transfer RNAs, blue = ribosomal RNAs

Table 2 Insertion/deletion events (indels) observed in the Lasergene assembly files of the chloroplast genome of *Lolium perenne*. 'Position' refers to the position of the indel within the complete chloroplast genome sequence of *L. perenne*. The hyphen within the major/minor column represents a deletion event, while a nucleotide stands for an insertion event.

position	nucleotide		region	trace files	
	major[a]	minor[a]		absolute[a]	%[a]
8258	-	T	trnS-psbD	2	50.00
18191	-	T	trnC-rpoB	3	50.00
31190	T	-	atpI-atpH	3	27.27
62835	-	A	psbE-petL	3	42.86
62836	-	A	psbE-petL	3	42.86
63107	-	T	psbE-petL	4	50.00
66367	-	A	rpl20-rps12	3	30.00
66368	-	A	rpl20-rps12	3	30.00
80295	-	T	rps19-trnH	2	28.57
93161	-	G	rrn16-trnI	3	16.67

[a](major = most commonly found nucleotide; minor = least commonly found nucleotide; absolute and % columns refer to the amount of trace files containing the under-represented nucleotide)

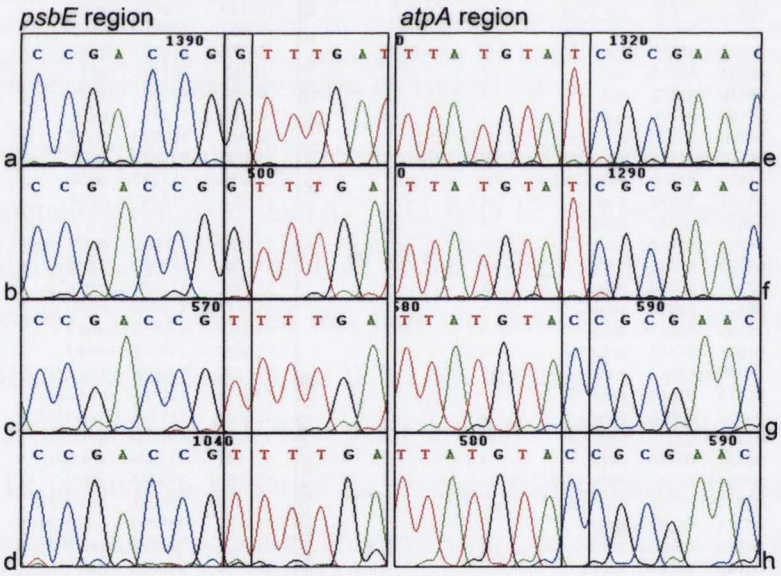

Figure 12 Single nucleotide polymorphisms (SNPs) in the *psbE* and *atpA* region in the *Lolium perenne* chloroplast genome. The SNP in the *psbE* region represents a transversion, the SNP in the *atpA* region a transition event. a, b, c, d = four different trace files of the same region in *psbE* / e, f, g, h = four different trace files of the same region in *atpA*.

Table 3 Single nucleotide polymorphisms in the chloroplast genome of *Lolium perenne*.

position	nucleotide		IUPAC[a]	amino acid change	region	tracefiles	
	[a]major	[a]minor				[a]absolute	[a]%
1618	A	T	W		*trnK* intron	1	20.00
19560	T	C	Y	I → T	*rpoB*	1	25.00
27177	G	A	R	S → N	*rpoC2*	1	25.00
28829	G	A	R	A → T	*rpoC2*	1	14.29
34720	G	A	R	—	*atpA*	4	36.36
37506	T	C	B	—	*psaB*	2	25.00
38978	C	A	M	—	*psaA*	1	7.14
40609	A	G	R	G → V	*psaA*	5	33.33
42894	A	G	R		*ycf3* intron	2	25.00
43270	G	A	R		*ycf3* intron	1	10.00
54360	C	A	M	Q → K	*rbcL*	1	50.00
61647	A	C	M	—	*psbE*	2	28.57
65631	C	T	Y	P → L	*rps18*	1	7.69
69066	G	A	R	A → T	*psbB*	1	16.67
86203	G	A	R	A → V	*ndhB* exon	1	5.56
94732	G	A	R		*trnA* intron	1	4.00
95307	G	A	R		*rrn23*	1	4.17
96920	C	A	M		*rrn23*	2	4.76
96968	C	G	S		*rrn23*	2	5.13
108390	A	G	R	—	*psaC*	6	42.86
109007	G	A	R	A → V	*ndhE*	1	10.00
[a]18	C	T	Y		*rps19-psbA*	1	16.67
[a]20	A	C	M		*rps19-psbA*	1	16.67
[a]21	G	T	K		*rps19-psbA*	1	16.67
[a]23	A	G	R		*rps19-psbA*	1	16.67
45874	A	C	M		*trnT-trnL*	2	50.00
47636	T	C	Y		*trnF-ndhJ*	1	20.00
51379	A	G	R		*trnV-trnM*	1	16.67
62341	T	G	K		*psbE-petL*	3	42.86
62521	A	C	M		*psbE-petL*	4	50.00
63360	G	A	R		*petL-petG*	3	50.00
73849	C	T	Y		*petD-rpoA*	1	8.33
82491	A	G	R		*rpl23-trnI*	1	5.88
83207	G	T	K		*trnI-trnL*	1	7.14
85260	G	A	R		*trnL-ndhB*	1	5.88
[a]100655	C	T	Y		*trnN-rps15*	13	43.33
[a]100656	A	G	R		*trnN-rps15*	13	43.33
103870	A	G	R		*ndhF-rpl32*	2	18.18
105222	C	T	Y		*rpl32-trnL*	1	14.29
108689	G	A	R		*psaC-ndhE*	1	7.69

[a](major = most commonly found nucleotide; minor = least commonly found nucleotide; absolute and % columns refer to the amount of tracefiles containing the under-represented nucleotide; inversions are highlighted in grey. IUPAC = International Union of Pure and Applied Chemistry, code for ambiguous nucleic acids.)

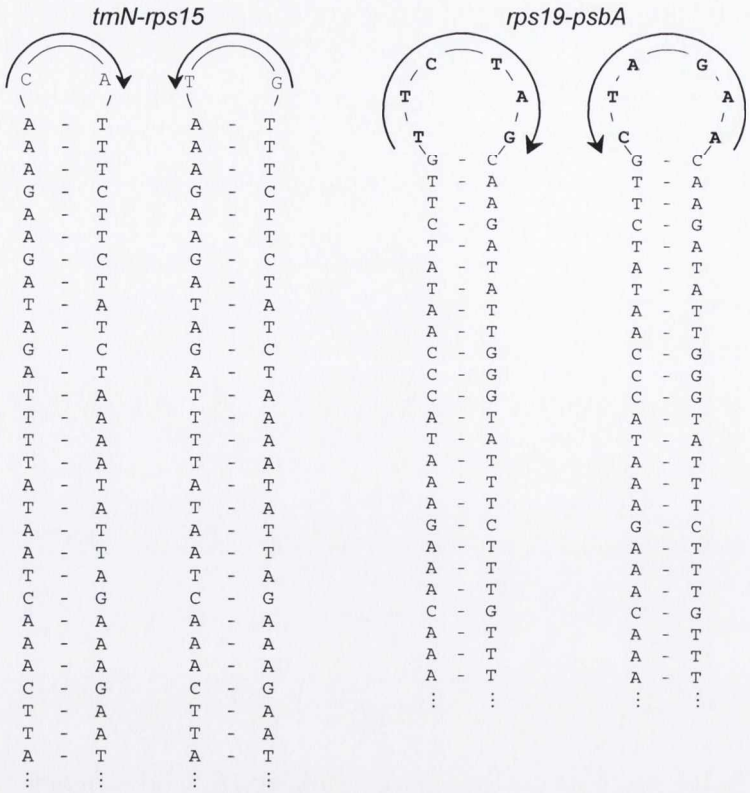

Figure 13 Two hairpin loops found in the chloroplast genome of *Lolium perenne*.

2.3.3 Comparison of Poaceae chloroplast genomes

The average size of publicly available Poaceae chloroplast genomes is 137,091 bp. The subfamily Ehrhartoideae has the smallest genome with an average size of 134,505 bp; subfamily Panicoideae has the largest genome with an average size of 140,876 bp. The subfamily Pooideae, to which *L. perenne* belongs, has an average size of 135,614 bp. Thus the chloroplast genome of *L. perenne* is of average size within Pooideae and of medium size within Poaceae (Figure 5). The gene content and intron content of *L. perenne* chloroplast DNA is the same as that of other grasses (Saski et al. 2007, Calsa Júnior et al. 2004, Maier et al. 1995, Ogihara et al. 2002, Bortiri et al. 2008, Palmer 1991) with 76 protein-coding genes, 30 tRNA genes and four rRNA genes. Eighteen genes are completely duplicated within the IR, as are the 3' exons of the *trans*-spliced gene *rps12* and the 5' part of *ndhH* which overlaps the IR/SSC junctions. As compared to the 'standard gene set' of angiosperm chloroplast genomes, the genes *accD*, *ycf1* and *ycf2* are absent.

Differences in chloroplast genome size of *L. perenne* to other Poaceae species are mainly due to length variations of IGS regions and introns (Table 4). The length of IGS regions and introns varies widely from only a few base pairs up to several hundred. Twenty five IGS regions and four introns were found to vary in length by more than 100 bp (Table 4). The highest variation in size (given in brackets) was found in the *trnI-CAU - trnL-CAA* IGS (2,135 bp), the *trnG-UCC - trnT-GGU* IGS (1,231 bp) and the *rbcL - psaI* IGS (1,221 bp). A comparison between *L. perenne* and the other Poaceae species showed differences in gene length for 26 genes (Table 5). The majority of these genes is in the LSC region. Length variations of more than ten codons were observed in eight genes (codon variation): *matK* (31), *ndhK* (21), *petB* (19), *rpoC2* (68), *rps3* (15), *rps15* (12), *rps16* (27) and *rps18* (14). The variation in gene length for the *rpoC2* gene was more than twice that found in any other gene. *Lolium perenne* and *A. stolonifera* have the shortest *rpoC2* genes (each 4,401 bp). The *rps18* gene in *L. perenne* is up to 14 codons shorter than in the other species. An alignment showed that this was due to the absence of two copies of a repeat and a deletion of 21 bp (Figure 14).The *ndhK* and *rps16* genes are 21 and 27 codons, respectively, longer in *L. perenne* than in *O. nivara*.

Table 4 Length variation of more than 100 bp for intergenic spacer (IGS) and intron regions in Poaceae chloroplast genomes. Bold: shortest length for that IGS/intron; **bold and underlined**: largest length for that IGS/intron. * highlights introns. **difference between smallest and largest value, IR = inverted repeat, SSC = small single copy region

		Ehrhartoideae	Pooideae					Panicoideae			
	genome region	*Oryza nivara*	*Triticum aestivum*	*Brachipodium distachyon*	***Lolium perenne***	*Hordeum vulgare*	*Agrostis stolonifera*	*Zea mays*	*Sorghum bicolor*	*Saccharum officinarum*	variation** in bp
large single copy region	*matK* - *trnK-UUU*	692	599	**444**	680	690	691	**695**	693	688	251
	rps16 - *trnQ-UUG*	1061	1062	784	815	**772**	787	1169	1521	**1530**	758
	trnS-UGA - *psbZ*	347	358	359	356	357	**362**	**261**	344	344	101
	trnG-GCC - *trnfM-CAU*	434	449	438	451	449	438	**494**	**378**	492	116
	trnG-UCC - *trnT-GGU*	1306	1200	**782**	1175	1194	1284	1874	**2013**	1956	1231
	trnT-GGU - *trnE-UUC*	519	**306**	507	**537**	453	470	529	462	535	231
	trnD-GUC - *psbM*	**381**	774	799	474	778	698	1052	**1059**	1052	678
	psbM - *petN*	761	628	797	**279**	717	287	799	808	**811**	532
	petN - *trnC-GCA*	**414**	949	935	917	725	921	955	933	**962**	548
	trnC-GCA - *rpoB*	**1084**	1165	1196	1185	1166	1173	**1273**	1204	1213	189
	atpI - *atpH*	795	572	**387**	570	568	531	818	756	**820**	433
	trnT-UGU - *trnL-UAA*	764	**613**	**829**	825	624	819	813	797	801	216
	**trnL-UAA* - *trnL-UAA*	542	**589**	543	550	569	**424**	459	450	453	165
	trnL-UAA - *trnF-GAA*	**245**	355	357	349	321	341	364	365	**366**	121
	trnF-GAA - *ndhJ*	495	**448**	575	586	587	584	**591**	591	570	143
	ndhC - *trnV-UAC*	**706**	911	816	844	727	926	**941**	924	929	235
	rbcL - *psaI*	**1683**	880	**462**	1184	1604	1561	889	862	945	1221
	ycf4 - *cemA*	420	**475**	426	460	470	455	**330**	373	372	145
	petA - *psbJ*	**1006**	821	835	**796**	821	806	900	900	900	210
	psbE - *petL*	1197	1169	1277	1281	**1162**	**1286**	1237	1265	1214	124
	**petB* - *petB*	**814**	749	809	747	**697**	759	699	702	758	117
	**rpl16* - *rpl16*	1056	1044	1050	**868**	1064	1045	1043	1072	**1080**	212
IR	*trnI-CAU* - *trnL-CAA*	**1498**	1498	2497	2452	2380	2487	3630	3632	**3633**	2135
	rps12_3end - *trnV-GAC*	1724	1726	**1642**	1646	1726	1756	1758	1767	**1767**	125
	**trnI-GAU* - *trnI-GAU*	948	807	806	808	802	**801**	950	948	**952**	151
	rps15 - *ndhF*	343	421	399	404	**422**	413	**107**	**107**	125	315
SSC	*ndhF* - *rpl32*	715	**919**	846	**703**	848	897	839	814	856	216
	rpl32 - *trnL-UAG*	547	690	697	663	**725**	659	531	**522**	523	203
	ndhG - *ndhI*	243	**264**	251	**116**	250	252	184	184	184	148

Table 5 Variation in length of different chloroplast genes. If not otherwise stated numbers shown refer to amount of additional codons; - = no variation to the smallest length observed; IR = inverted repeat, LSC = large single copy region, SSC = small single copy region

				Ehrhartoideae	Pooideae					Panicoideae			
	gene name	length in bp	length in codons	*Oryza nivara*	*Triticum aestivum*	*Brachipodium distachyon*	***Lolium perenne***	*Hordeum vulgare*	*Agrostis stolonifera*	*Zea mays*	*Sorghum bicolor*	*Saccharum officinarum*	codon variation
large single copy region	***atpA***	1515	505	3	-	-	3	-	1	3	3	3	3
	atpF	552	184	-	-	-	-	-	3	-	-	-	3
	infA	324	108	-	6	-	-	6	-	-	-	-	6
	matK	1536	512	-	-	31	-	-	-	2	4	2	31
	ndhK	678	226	-	20	20	21	20	20	2	2	2	21
	petB	648	216	-	-	-	-	19	-	19	19	-	19
	psaB	2205	735	-	-	-	-	-	-	1	-	-	1
	psaJ	129	43	2	-	-	-	-	-	-	-	-	2
	psbK	183	61	1	1	-	1	1	2	1	1	1	2
	psbT	102	34	2	5	2	5	5	5	-	-	-	5
	rbcL	1431	477	1	1	-	1	3	1	-	-	-	3
	rpl16	450	150	-	-	-	-	1	-	-	-	-	1
	rpl22	444	148	2	1	2	-	2	2	1	1	1	2
	rpoA	1014	338	-	2	-	4	2	2	2	2	2	4
	rpoB	3228	1076	-	1	1	1	1	1	-	-	-	1
	rpoC1	2031	677	6	7	6	-	6	6	7	7	7	7
	rpoC2	4401	1467	47	13	37	-	36	-	61	54	68	68
	rps16	189	63	-	23	23	27	23	23	23	23	23	27
	rps18	471	157	7	14	7	-	14	13	14	7	7	14
	rps3	675	225	15	15	15	15	15	15	-	-	-	15
IR	***rps12***	363	121	4	1	-	4	4	4	4	4	4	4
	rps15	237	79	12	12	12	12	12	12	-	-	-	12
SSC	***ccsA***	960	320	2	3	3	-	3	-	2	2	2	3
	ndhD	1503	501	-	-	-	2	-	-	-	-	-	2
	ndhF	2205	735	-	5	7	7	5	5	4	4	4	7
	rpl32	180	60	-	4	-	-	-	2	-	-	-	4

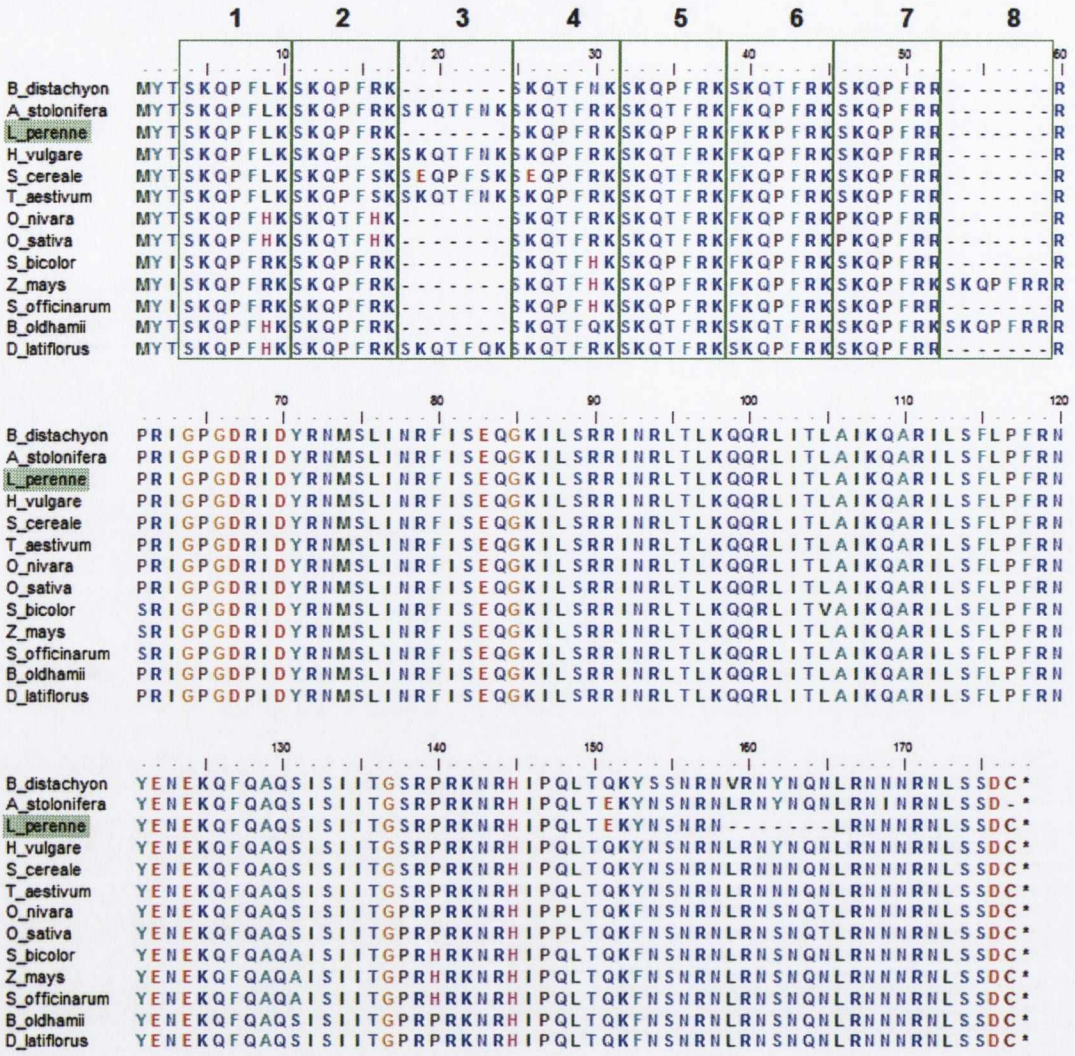

Figure 14 Amino acid based alignment of the chloroplast *rps18* gene alignment of eleven Poaceae species. Highlighted in green boxes is a heptapeptide repeat motif. A = *Agrostis*, B distachyon = *Brachipodium*, B oldhamii = *Bambusa,* H = *Hordeum*, L = *Lolium*, O = *Oryza*, S officinarum = *Saccharum*, S bicolor = *Sorghum*, T = *Triticum*, Z = *Zea,* D = *Dendrocalamus*.

2.4 Discussion

2.4.1 Sequencing the chloroplast genome of *Lolium perenne*

Sequencing the chloroplast genome of *Lolium perenne* was difficult. Chloroplast DNA yields were either of low quantity or of low quality due to high contamination with nuclear DNA. There are several methods which can be used for obtaining complete chloroplast DNA sequences. Chloroplast genomes can be sequenced by using chloroplast DNA clones found as 'contaminations' in genomic libraries, as done for the

Arabidopsis thaliana chloroplast genome (Sato et al. 1999). Sequencing can also be undertaken using the method used for the chloroplast genome of *Nicotiana sylvestris,* the maternal genome donor of *Nicotiana tabacum*: extracted high purity chloroplast DNA was cloned into sequencing vectors (Yukawa et al. 2006). Another approach that recently became available is based on the application of consensus chloroplast DNA sequencing primers in a primer walking strategy as was done for sequencing the chloroplast genome of *Cucumis sativus* (Chung et al. 2007). Another frequently used method involves amplifying the chloroplast genome by rolling circle amplification and then cloning this product into sequencing vectors (Jansen et al. 2005).

For our sequencing, a genomic library of *L. perenne* that could have been screened for chloroplast DNA clones was not accessible. Also consensus chloroplast DNA primers for the primer walking strategy had not been made publicly available. Therefore it was necessary to extract high purity chloroplast DNA that could be cloned into sequencing vectors as was done for example for *A. thaliana*. Several previously published protocols were tested for use in *L. perenne* chloroplast DNA isolation (Palmer 1986; Bellaoui 1997; Triboush 1998). However, the quantities of chloroplast DNA yielded were too small for this approach. The reason for this is not completely understood, but is suspected to be due to the fibrous content of *L. perenne* that complicated the homogenization of the leaves. Most of the tested protocols had been designed for the use in eudicot plant species like pea and sunflower (Bookjans 1984; Triboush 1998) which could have lower leaf fibre contents. Finally a combined approach was used based on the extraction of high purity chloroplast DNA and whole genome amplification to get sufficient amounts of DNA for sequencing. This approach was based on the combination of the protocols from Bookjans (1984), Palmer (1986) and Jansen et al. (2005) and their optimization for the application in *L. perenne*. A high salt step from Bookjans (1984) was used to prevent the co-isolation of nuclear DNA. It was followed up with differential centrifugations and the usage of sugar gradients based on the protocol from Palmer (1986). For the chloroplast lysis the lysis buffer based on the protocol from Jansen et al. (2005) was used. The so yielded high purity DNA was then amplified using a whole genome amplification kit to obtain enough DNA for sequencing. On an agarose gel which was used for visualizing the amplified and restricted chloroplast DNA, several bands fluoresced very strong (Figure 8), suggesting that the whole genome amplification kit is biased and amplifying several regions in the

chloroplast genome preferentially. However, this did not have any negative impact on the sequencing, assembling and annotating of the *L. perenne* chloroplast genome. The protocol for chloroplast DNA extraction has been made available to the public (Diekmann et al. 2008) and was also tested on another grass species (*Miscanthus*).

2.4.2 Indel/SNP analysis

Choosing the shotgun sequencing approach, each region of the *L. perenne* chloroplast genome was sequenced several fold from independent clones which enabled the detection of SNPs and indels. A high number of indels were detected that were based on only one trace file. These indels were not considered for further analysis. All indels were based on mononucleotide repeats and it is very likely that indels found within only one trace file are due to slipped strand mispairing during amplification with the whole genome amplification kit but also during propagation of the bacteria containing the chloroplast DNA fragments for sequencing. Nevertheless the amount of observed SNPs and indels within one cultivar remained unexpectedly high, considering the generally high conservation of the chloroplast genome. *Lolium perenne* is an outcrossing species and cultivars are based on populations (they are not clonal or highly inbred). Thus great variation at the nuclear DNA level but not at the chloroplast DNA level would have been expected. However, a previous study McGrath et al. (2007) revealed also a high cytoplasmic variation within *L. perenne* ecotypes and cultivars using chloroplast microsatellite markers. Thus up to 16 haplotypes within 16 analysed individuals per accession could be detected. The two inversions found in the current study lead, together with the level of observed SNPs, to the conclusion that the chloroplast genome of *L. perenne* 'Cashel' consists of at least two, maybe three haplotypes but potentially scores even more. Due to the lack of recombination all chloroplast DNA loci of an individual are linked with each other. Hence all SNPs and indels observed could occur in only two individuals or more. It is not possible to tell from this analysis how many haplotypes occur in cultivar Cashel because a mixed leaf sample from a large Cashel population was used and it is not possible to assign the results from the sequences to individual plants. However, McGrath et al. (2007) detected five haplotypes in 16 individuals of Cashel using a set of ten primers (McGrath et al. 2006) amplifying eight different regions in the chloroplast genome and sizing the PCR products. From personal

communication with Vincent Connolly, the breeder of *L. perenne* cv. Cashel, it is known that this cultivar is based on eight maternal lines. Therefore it is possible that up to eight different haplotypes could be detected in this cultivar, in future, more thorough, diversity studies.

Minute inversions like the reported inversions of this study have been previously found in studies of other plant species (Kim & Lee 2004, 2005), between genera (Kim & Lee 2004, Kelchner & Wendel 1996) and also within populations of one other species, the conifer *Abies* (Tsumura et al. 2000). Inversions are thought to be due to intramolecular recombination events (Kelchner & Wendel 1996) and are common in chloroplast genomes (e.g. the large inverted repeat in many chloroplast genomes). While it is possible to use information on the structure of the large IR for phylogenetic analysis, minute inversions show some homoplasy and thus are only applicable to a limited extent. Nevertheless are they useful in assessing population structures and the detection of diversity within species.

Chloroplast genomes are known to be highly conserved, thus the extent of variation found within the single *L. perenne* cv. Cashel was unexpected. However this was not the first report about intraspecific variation in the chloroplast genome. Similar observations have been recorded in other species such as in the monocotyledonous species *Agrostis stolonifera*, *Sorghum bicolour* and *Hordeum vulgare* (Saski et al. 2007) and the dicotyledonous species *Solanum lycopersicum, Solanum bulbocastanum* (Daniell et al. 2006), *Lactuca sativa* and *Helianthus annuus* (Timme et al. 2007). However, the current study is the only study known that has quantified SNP variation of the whole chloroplast genome within a cultivar. Most studies of chloroplast DNA variation within a species have assessed populations of individuals with a limited number of markers, from a few selected gene regions, or have sampled wild populations. Although some of the apparent SNPs and indels that were only present in one sequencing read might be due to cloning artefacts, the current results are not surprising in view of the fact that *L. perenne* is an outcrossing species and the cultivar used for sequencing is based on a population of several maternal lines and is thus heterogeneous and heterozygous.

2.4.3 Comparison of Poaceae chloroplast genomes

This study revealed that Poaceae chloroplast genomes vary in size from 134 kb to 141.5 kb. Thus the newly sequenced *L. perenne* chloroplast genome is of average size. Sorting the Poaceae chloroplast genomes according to their size furthermore revealed that genomes of the same subfamily were grouped together (Figure 5). Figure 4 indicates a similar trend for dicotyledonous angiosperms or gymnosperms. For example, all sequenced Brassicaceae genomes are grouped close together (Figure 4) and analysis might reveal that in this family a grouping according to lower taxonomic rank such as subfamily, or tribe, is possible.

Comparison of the Poaceae chloroplast genomes also revealed a high conservation regarding the gene and intron content of this family. Recently the chloroplast genome of *Festuca arundinacea* was published on Genbank (A. B. Cahoon et al, unpublished; accession number FJ466687). Surprisingly, this seems to be the first Poaceae chloroplast genome that does not share this conservation. In addition to the expected absences of *accD*, *ycf1* and *ycf2*, the *Festuca* sequence also lacks intact copies of the genes *psbF*, *rps14*, *rps18* and *ycf4*. Especially *psbF* was so far thought of as being essential due its function in the cytochrome b559 in the photosystem II. This was the first report of the deletion of photosynthetic gene within the chloroplast genome of a photosynthetic active plant. All four genes that are absent from *F. arundinacea* are intact and apparently functional in *L. perenne*. The more surprising is the absence of these genes in *F. arundinacea* which is very closely related to *L. perenne* (Catalan et al. 2004). For sequencing the chloroplast genome a combined strategy of a shotgun sequencing approach and a primer walking approach was used (http://mtsu32.mtsu.edu: 11354 /3250%20Genetics/Sequencing%20Lessons/sequencing%20home.htm). Thus the regions were only sequenced once or twice and sequencing errors due to polymerase slipped-strand mispairing might not have been revealed. Interestingly most of the differences between the *F. arundinacea* and *L. perenne* chloroplast genome are located in microsatellite regions in form of indels. This observation supports the hypothesis of polymerase reading mistakes. However, A. B. Cahoon (personal communication) assured the correctness of their results.

Differences in the chloroplast genome size of *L. perenne* compared to other Poaceae species were mainly due to length variations of IGS regions and introns (Table 4). This finding was consistent with previous observations (Maier et al. 1995, Palmer 1991). The length of IGS regions and introns varies widely from only a few base pairs up to several hundred. Two of the most variable regions were the *trnI-CAU – trnL-CAA* IGS and *rbcL - psaI* IGS. These IGS are sites that contain pseudogenes for *ycf2* and *accD*, respectively, in Poaceae (Maier et al. 1995). Both these pseudogenes and a *ycf1* pseudogene could be also detected in *L. perenne*. Another highly variable site, the *trnG-UCC - trnT-GGU* IGS, is part of a 'divergence hotspot' described by Maier et al. (1995) whose variability is caused by a large number of deletion/insertion events.

Length variations were also observed in coding regions. Especially noteworthy are variations observed in *rps18* and *rpoC2*. In both cases *L. perenne* showed the shortest of all sequences. Sequence variation between monocots and dicots for *rps18* has been described by Weglöhner et al. (1991), based on the occurrence of different numbers of the heptapeptide repeat SKQPFRK near the N-terminus of the protein. This study revealed that length differences among Poaceae *rps18* sequences are mainly based on the same heptapeptide repeat (S/F)K(Q/K)(P/T)F(R/L/H/S/N)(K/R) as described by Weglöhner et al. (1991). This motif is six times present in *rps18* of *L. perenne*, *B. distachyon*, *O. sativa*, *O. nivara*, *S. bicolor* and *S. officinarum* and seven times in *rps18* of *A. stolonifera*, *H. vulgare, T. aestivum* and *Z. mays* (Figure 4). The *L. perenne rps18* gene is the shortest detected so far because it has undergone an additional deletion of seven amino acids near its C-terminus. The deletions do not result in the creation of stop codons. Hence the *L. perenne rps18* gene is expected to be functional. However, the variability of the *rps18* gene and its length reduction in the *L. perenne* genome possibly indicates the ongoing deletion of the *rps18* gene from the chloroplast genome and its transfer to the nuclear or mitochondrial genome. As detailed in chapter 5, intracellular gene transfer is part of the endosymbiotic gene transfer and responsible for the reduction of the chloroplast genome size from several hundred coding sequences in the ancestral chloroplast genome to only around 100 in the nowadays chloroplast genome (Adams & Palmer 2003, Keeling 2004). Figure 14 shows that only repeat copies three and eight are affected from the deletions/insertions and it furthermore shows that deletions/insertions occur parallel in different subfamilies. Thus it might be possible

that the deletion of *rps18* observed by Cahoon et al. (unpublished; accession number FJ466687) for the *F. arundinacea* genome is not completely unlikely.

The largest length variation in Poaceae genes was found in *rpoC2,* of up to 68 codons difference between *L. perenne* and *S. officinarum.* Alignments of the Poaceae *rpoC2* genes showed that this variation is due to several insertion/deletion events. Comparisons of the *rpoC2* gene from dicots and monocots revealed that Poaceae have a unique insertion of approximately 400 bp in the middle of this gene (Igloi et al. 1990; Shimada et al. 1990; Cummings et al. 1994; Chen et al. 1993, 1995). Cummings et al. (1994) demonstrated that this region is highly variable compared to its flanking regions and is rich with tandem repeats. Nearly all the variations found between *L. perenne* and the panicoids are located in this specific insertion region. Analysing cytoplasmic male sterile (CMS) lines of *Sorghum,* Chen et al. (1995) discovered a 165 bp deletion in this insertion region that suggests a possible relation between this deletion in *rpoC2* and the CMS-system. So far this deletion was only observed in *Sorghum* but sequence comparisons revealed that one deletion that results in the shorter *L. perenne rpoC2* gene is located in the same region where the deletion occurs in *Sorghum.* Hence a higher susceptibility to variation in this gene region could be indicated and an investigation of *L. perenne* CMS lines in regard to variation to fertile lines may prove valuable for improving future *Lolium* breeding schemes.

2.5 Conclusion

The chloroplast genome of *L. perenne* is of average size when compared to other species. It has the typical quadripartite structure observed for many species and it lacks the genes *accD*, *ycf1* and *ycf2* which are typical for Poaceae chloroplast genomes. The availability of the *L. perenne* chloroplast genome enables the species-specific design of vectors that target the *L. perenne* chloroplast genome and enable a genetically modification of the plastid genome in the future.

Despite the generally high conservation of chloroplast genomes, sequencing the chloroplast genome of *L. perenne* revealed a surprisingly high amount of variation in form of SNPs and indels, mainly located within the IGS regions, within a single

cultivar. A comparison between the chloroplast genomes of different Poaceae species also revealed that variations in genome length are mainly due to length variation within IGS. Furthermore it was possible to detect genes highly variable in length and possibly in the *rps18* gene in comparison to other Poaceae *rps18* genes an initial horizontal gene transfer event was observed.

Knowledge of the sequence variability of regions within Poaceae chloroplast genomes and particularly the *L. perenne* chloroplast genome can be used to design primers for population genetic and phylogenetic studies and to support breeding schemes via defining cytoplasmic breeding pools. This will be of especially high value for breeding schemes based on interspecific crosses between *Lolium* and *Festuca.*

Chapter 3

RNA editing sites in *Lolium perenne*

"Nature does not always prefer simple solutions over complicated ones. The persistence of costly but seemingly unnecessary processes like RNA editing in plants illustrates quite drastically that evolution is still far from being an open book for us."

Bock (2000)

3.1 Introduction

3.1.1 RNA editing

In 1986 Dutch scientists (Benne et al. 1986) made a rather surprising discovery when they analysed the cytochrome oxidase subunit II (*coxII*) of the mitochondrial genome of trypanosomes (euglenoids). This gene showed a -1 frameshift in three trypanosome species, but appeared nevertheless functional due to its high degree of conservation. They analysed the frameshift region via reverse transcriptase polymerase chain reaction (RT-PCR) using total RNA of each species as a template. This analysis revealed an insertion of four extra nucleotides which lead to the restoration of the frameshift and resulted in a fully functional gene. This discovery was the first observation of RNA editing.

One year later Powell et al. (1987) observed RNA editing for a nuclear encoded gene, the human apolipoprotein B (*apoB*). Here, RNA editing modifies one nucleotide and thereby changes a glutamine codon to a stop codon which results finally in the expression of two different proteins (one of full length and one of partial length) from the same primary transcript (Powell et al. 1987). In 1989, three different research groups reported that RNA editing also happens in plants where it was found to occur in wheat, *Triticum*, (Covello & Gray 1989; Gualberto et al. 1989) and evening primrose, *Oenothera,* (Hiesel et al. 1989) mitochondria. At that time it was still debated whether plant mitochondria use a different genetic code, because of detection of arginine codons

in some species at positions were tryptophan codons were highly conserved in other species. However, due to the observation by these three research groups the answer was clearly that the standard genetic code can be applied to the plant mitochondrial genome because previously observed differences are eliminated by editing. Shortly afterwards, in 1991, the first observation of RNA editing in chloroplast genomes was published based on analyses of the ribosomal protein L2 (*rpl2*) gene of maize, *Zea mays*, (Hoch et al. 1991).

Much more information on RNA editing has become available and evidence for the process has been found in the organelle genomes of a wide range of species (Brennicke et al. 1999). Smith et al. (1997) defined RNA editing as "the co- or post-transcriptional modification of RNA primary sequence from that encoded in the genome through nucleotide deletion, insertion, or base modification". RNA editing was mainly detected in transcripts of protein coding genes, but also in structural transfer RNAs,(tRNA), (Diamond et al. 1990) and ribosomal RNAs (rRNA), (Schuster et al. 1991). Two classes of RNA editing have been reported: 'insertional RNA editing' and 'substitutional RNA editing' (Bock et al. 1996; Brennicke et al. 1999). 'insertional RNA editing' includes all processes that insert or delete nucleotides from the RNA molecule and change the reading frame of an RNA (Mulligan 2004) as observed for the *coxII* gene in trypanosomes (Benne et al. 1986). The edited transcript will be longer or shorter than the primary transcript. In contrast, 'substitutional editing' is based on nucleotide substitution, mainly A to I and C to U changes as found in the nuclear genome of mammals and in the mitochondrial genome of plants, respectively (Brennicke et al. 1999).

3.1.2 RNA editing in chloroplast genomes

Hoch et al. (1991) aligned the maize *rpl2* gene with the *rpl2* gene from other species and recognized that the maize *rpl2* gene lacks a translation start codon. At the position corresponding to the AUG start codon in other plant species, maize had the codon ACG. Sequence analyses of RT-PCR products from cDNA templates of that region revealed that the codon ACG in maize is replaced on mRNA level by the start codon AUG via a

C to U base modification. This modification results in a fully functional gene. Hoch et al. (1991) were the first to show that RNA editing also takes place in chloroplasts.

Many more reports of editing sites in chloroplast genomes followed but the number of editing sites found in chloroplast genomes, approximately 30 in a single species (Tillich et al. 2005), is much smaller than the number found in mitochondrial genes (e.g. 456 in *Arabidopsis thaliana*; Giege & Brennicke 1999; Miyamoto et al. 2002). So far only C-to-U RNA editing sites have been reported with the only exception found in the chloroplast genome of the liverwort (*Anthoceros formosae*) that has been shown to have a large number of U-to-C editing sites (Kugita et al. 2003). Only transcripts of protein coding genes, the messenger RNAs (mRNA), are affected by editing in chloroplast genomes. Although editing can restore translation initiation codons, as reported for the *rpl2* gene in maize, and translation stop codons as reported for the *petL* and *petD* genes in black pine (*Pinus thunbergii*) (Wakasugi et al. 1996), it mainly restores internal codons for conserved amino acids (Giege & Brennicke 1999). mRNA editing affects 0.1 – 0.25% of all codons or approximately 0.02% of all nucleotides in a chloroplast genome (Bock 2000). Editing sites within the chloroplast genome are rather unevenly distributed, many genes are not edited at all, while others like the NADH dehydrogenase subunit 2 (*ndhB*), contain up to nine editing sites (Freyer et al. 1995). Nevertheless editing sites do occur in both of the two major classes of chloroplast genes, the genetic-system and the photosynthesis related genes (Bock 2000). Editing sites are generally located at the second codon position, and there is strong bias to the 5' neighbour nucleotide being a pyrimidine base and the 3' neighbour nucleotide a purine base (Bock 2000). Editing in chloroplast genomes generally results in changes of the physicochemical properties of the amino acids. Amino acid changes from serine to phenylalanine and leucine, respectively, as well as changes from proline to leucine are strongly favoured (Bock 2000). Most editing sites are completely edited, but some genes, like the *rpl2* (Freyer et al. 1993), have been found to be partially edited. The term 'partially edited site' describes the observation that only a subset of transcripts is edited (Mower & Palmer 2006). Partially edited sites were first reported by Gualberto et al. (1989) in wheat mitochondria via evidence from direct cDNA sequencing, and shortly after confirmed by sequencing of *Oenothera* cDNA clones (Schuster et al. 1990). In cDNA sequences partially edited sites can be found in form of SNPs - both nucleotides appear at the partially edited site. Schuster et al. (1990) found that mainly silent editing

sites were affected. However, partially edited sites were also observed in different (root, leaves) or developmentally varying tissue (young leaves, mature leaves) (Bock et al. 1993; Ruf & Kössel 1997), but the functional significance of partially edited sites is not fully understood.

3.1.3 Chloroplast mRNA editing mechanism

Knowledge about the mechanism of editing in chloroplast genomes is rather limited. Comparisons of editing sites have not revealed obvious structures, such as conserved primary sequences or structural motifs, to direct the process (Chaudhuri & Maliga 1996). One of the most recent models predicted by Okuda et al. (2006) (Figure 15) suggests the involvement of a *cis*-acting element of approximately 36 nucleotides, a *trans*-acting factor and an RNA editing enzyme.

Very early on in the attempts to understand the editing mechanism, the involvement of *cis*-elements, regulatory functioning DNA sequences located within the chloroplast genome, was suggested. In 1996 two independent research groups (Bock et al. 1996; Chaudhuri & Maliga 1996) presented results supporting this theory for the first time. They showed that flanking sequences of around 40 nucleotides were essential for the recognition of editing sites (Bock et al. 1996). Furthermore they were able to show that flanking sequences of -12 and -2 nucleotides for editing site IV and V, respectively, of the *ndhB* gene (Bock et al. 1996), and -16 to +5 nucleotides for the editing site of the *psbL* gene (Chaudhuri & Maliga 1996) in tobacco, are essential for editing these sites. Hayes and Hanson (2007) showed that the six nucleotide motif GCCGUU is critical for efficient editing of the tobacco *psbE* transcript. All three research groups confirmed in their results that upstream sequences have a higher impact on editing than downstream sequences. However, as already previously observed only few flanking regions of editing sites share common sequence motifs (Maier et al. 1992), and the GCCGUU motif was only found once within the protein-coding sequences of the tobacco chloroplast genome (Hayes & Hanson 2007). Also Bock et al. (1996) were not able to detect homologous sequences upstream of other chloroplast RNA editing sites and therefore concluded that *cis*-acting DNA sequences are site-specific and are possibly binding sites for site-specific *trans*-acting factors.

Chaudhuri et al. (1995) had already considered the involvement of site specific *trans*-acting factors. The first proof of a *trans*-acting factor, a factor that is not encoded in the chloroplast genome, however was detected by Hirose and Sugiura (2001) who detected a 25 kDa protein which binds to the predicted region of the tobacco *psbL cis*-factor and obviously influences the editing efficiency in the *psbL* transcript. Miyamoto et al. (2002) reported a 56 kDa protein binding specifically to the *cis*-acting element of the tobacco *psbE* gene. The exact *cis*-element motif of the tobacco *psbE* was at that time only roughly determined, but Hayes and Hanson (2007) detected the exact sequence and position of it a few years later. Nevertheless the results of both research groups were in agreement. Also a 70 kDa protein was detected serving obviously as *trans*-acting factor for the *petB* transcript (Miyamoto et al. 2002). The combination of results suggested that each editing site is recognized by a unique RNA-binding protein (Miyamoto et al. 2002). But the number of unknown open reading frames (ORFs) does not correspond to the number needed for translating these unique proteins, therefore Miyamoto et al. (2002) suspected that these proteins are imported from the cytoplasm.

Kotera et al. (2005) discovered a gene (*crr4*) that encodes a pentatricopeptide (PPR) protein. They predicted that this protein was the *trans*-acting factor essential for the recognition of the *ndhD*-I editing site in *Arabidopsis* and a follow-up study by Okuda et al. (2006) confirmed this hypothesis. The PPR protein family is very large and most of its proteins are predicted to target mitochondria and chloroplasts (Small & Peeters 2000). This could be observed for many nuclear encoded CMS related fertility restorer (*Rf*) genes, for example, that code for PPR proteins and target mitochondria (reviewed in Hanson & Bentolila 2004). The PPR protein family is characterized by its PPR motif which consists of 35 amino acids and appears as tandem repeats in proteins (Small & Peeters 2000). Other PPR genes involved in the recognition of editing sites have since been discovered: CRR21 as a *trans*-factor for the editing site *ndhD*-II (Okuda et al. 2007), CRR22 as a *trans*-factor for editing sites in *ndhD*, *ndhB*, *rpoB* and CRR28 as a *trans*-factor for editing sites in *ndhD* and *ndhB* (Okuda et al. 2009), as well as CLB19 as *trans*-factor for editing sites in *rpoA* and *clpP* transcripts (Chateigner-Boutin et al. 2008). It was also discovered that although *cis*-elements seem to be site-specific, *trans*-factors can be involved in several editing sites. According to the model proposed by Okuda et al. (2006) the editing of the nucleotides is carried out by a specific RNA editing enzyme. The *trans*-acting factor and the RNA editing enzyme do not have to be

necessarily two different proteins (Okuda et al. 2006). However, the *trans*-factor proteins detected so far do not possess cytidine deaminase activity (Okuda et al. 2009) and therefore the search for the (possibly) last missing link from our understanding of the RNA editing mechanism in chloroplast genomes goes on.

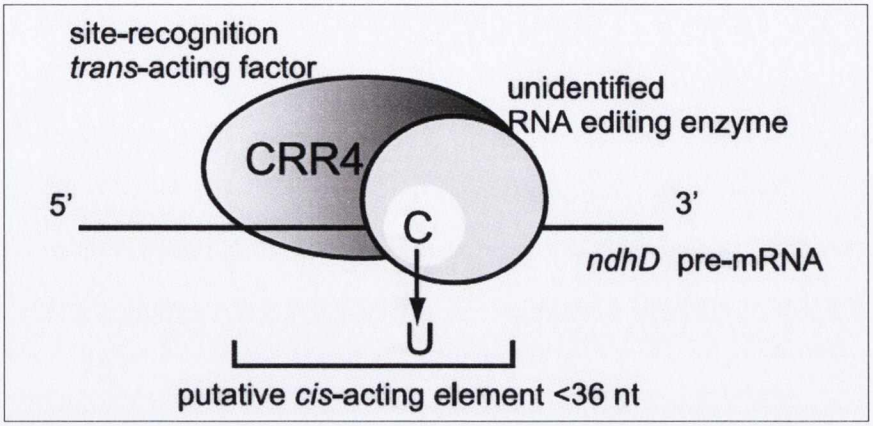

Figure 15 Possible chloroplast mRNA editing mechanism as predicted for the *ndhD*-1 editing site in *Arabidopsis* by Okuda et al. (2006).

3.1.4 Evolution of chloroplast mRNA editing sites

RNA editing in different organelles involves different mechanisms and therefore it would seem to have evolved several times independently during evolution (Bock 1996) either concomitantly or shortly after the evolutionary diversification of land plants (Bock et al. 2000). RNA editing compensates for point mutations in gene sequences and increases the evolutionary variation (Brennicke et al. 1999). Editing in chloroplast genomes is predicted to be of monophyletic origin (Tillich et al. 2006). A subsequent reduction of the number of editing sites was observed from hornworts, over ferns, to seed plants. This reduction could be an indicator that editing is disappearing from seed plants because genome mutations eliminate the need of editing (Tillich et al. 2006). Although Tillich et al. (2006) report that editing sites are predominantly in regions with low point mutation rates, Shields and Wolfe (1997) showed that the evolutionary rate at editing sites is high and predict that selection is working against the editing mechanism.

Editing sites are most frequently observed in non-essential plastid genes where they are not very well conserved among species suggesting some evolutionary flexibility (Fiebig et al. 2004). This observation led to the 'relative neutrality hypothesis' which is based on the fixation of T-to-C point mutations in predominantly non-essential plastid genes.

These point mutations are not fatal and can according to the hypothesis be reversed by establishing mutation compensating editing activities (Fiebig et al. 2004).

Comparison analyses of the *ndhB* editing sites of *Hordeum vulgare* (barley) with two other monocotyledonous and one dicotyledonous angiosperm species showed high conservation among all these species for six out of nine sites (Freyer et al. 1995). However, divergence at two of the remaining three sites was surprisingly higher between *H. vulgare* and the monocotyledonous species than between *H. vulgare* and the dicotyledonous species (Freyer et al. 1995). This is in contradiction to established phylogenetic relationships and shows that individual editing sites alone can not be used as criteria for estimating phylogenetic distances (Freyer et al. 1995). Also the independent gain of the same editing site in different species suggests convergent evolution by independent creation of identical new editing sites and simultaneous acquisition of specific *cis*- and *trans*-factors (Freyer et al. 1997). Nevertheless, other studies based on editotypes (complete set of editing sites in a species, Tillich et al. 2005) instead of single genes revealed that most sites are conserved among taxonomic families (Corneille et al. 2000) and subfamilies (Tillich et al. 2005).

3.1.5 Aims and objectives

The grasses (Poaceae) are the fourth largest plant family and the second largest family within the clade of monocotyledonous species (Tzvelev 1989). Considering the fact that land plants in general are approximately 450 million years old (reviewed in Palmer et al. 2004), the grass family has with an estimated age of 70 million years a rather recent evolutionary history (Bremer 2002). Hence it offers a unique opportunity to study genetic variation at the level of a plant family. Additionally, due to their economically high importance, this family is very well studied which is also reflected in the amount of chloroplast genome sequences published so far. Nineteen monocotyledonous chloroplast genome sequences are available, of which 14 are from Poaceae species. However, information on editing sites within these Poaceae chloroplast genomes was limited until now to only four species, *H. vulgare*, *O. sativa*, *S. officinarum* and *Z. mays*. RNA editing is essential to plants because it compensates for mutations during plant evolution and enables the expression of functional proteins (Shikanai et al. 2006).

Although permanently increasing, the knowledge about the RNA editing mechanism is still limited. But the possible involvement of RNA editing in the CMS-system of plants (Chapter 5) makes it highly interesting from a plant breeding perspective and highlights the importance of gathering information on editing sites in different species as well as to further investigate the mechanism behind it. Therefore the objectives of this chapter were 1) to analyse the chloroplast genome of *L. perenne* for editing sites via a bioinformatic and RT-PCR approach, 2) to assess the efficiency of different approaches for the detection of editing sites such as EST screening, RT-PCR, and combinations of the two, 3) to assess the extent of partial RNA editing in *L. perenne*, 4) to compare the extent and nature of editing sites among monocotyledons and more specifically among the grasses, 5) to evaluate the evolutionary history of editing sites within the grass family and its respective subfamilies and tribes (concentrating on the three genes containing most editing sites in grasses – *ndhA* (four sites in *H. vulgare*), *ndhB* (nine sites in *H. vulgare*) and *rpoB* (four sites in *Z. mays*).

3.2 Material and Methods

3.2.1 Plant material

Small quantities of two week old *L. perenne* cv. Cashel leaves were shock frozen in liquid nitrogen and ground thoroughly with mortar and pestle. 2 ml microcentrifuge tubes were filled half full with the frozen plant powder and stored until further use at -80 °C. *Lolium perenne* cv. Cashel is a population of plants and the obtained plant powder derives from approximately 200 individuals.

3.2.2 RNA extraction and cDNA synthesis

Total RNA was extracted using TRI Reagent® Solution (Ambion Inc., Austin, TX, USA) following the supplier's protocol (http://www.ambion.com/techlib/prot/bp_9738.pdf) with the following modifications: the incubation of the homogenate was extended to 10 min; instead of 100 ml bromochloropropane, 200 ml of ice cold chloroform was used; the steps including the addition of ice cold chloroform, followed by incubation at room temperature and centrifugation at 12,000g were repeated once; in

addition to the 500 ml isopropanol, 0.5 ml Glycogen (Sigma-Aldrich, St Louis, Missouri, USA) was added to enhance the RNA yield; the centrifugation following the addition of isopropanol was extended to 10 min. The RNA was finally dissolved in nuclease free water and treated with DNA-free™ (Ambion Inc., Austin, TX, USA) following the manufacturer's instructions (http://www.ambion.com/techlib/prot/bp_1906.pdf) to remove possible DNA contamination. First strand cDNA was synthesized using SuperScript™ III Reverse Transcriptase (Invitrogen™ Corporation, Carlsbad, CA, USA) following the manufacturer's instructions (http://tools.invitrogen.com /content/sfs/manuals/ superscriptIII_man.pdf). Resulting cDNA was stored at -20 °C.

3.2.3 Primer design and PCR amplification

This analysis focused only on chloroplast genes for which RNA editing has previously been detected in Poaceae species (Calsa Júnior et al. 2004; Corneille et al. 2000; Freyer et al. 1993; Maier et al. 1995; Zeltz et al. 1993) or for which differences from existing expressed sequence tags (ESTs) in Poaceae species have been found (Saski et al. 2007). Thus, thirty-three genes (*atpA, atpB, atpF, clpP, matK, ndhA, ndhB, ndhD, ndhF, ndhG, ndhI, ndhK, petA, petB, psaA, psaB, psaJ, psbC, psbD, psbE, psbJ, psbL, psbZ, rpl2, rpl20, rpoA, rpoB, rpoC1, rpoC2, rps14, rps2, rps8, ycf3*) were incorporated in the analysis, 22 because editing sites had been previously reported and 11 because they showed differences between genome sequences and ESTs.

Primers (Appendix Table A1) for these genes were designed using Primer Express (version 2.0, Applied Biosystems, Foster City, CA, USA) and Primer3 software (http://frodo.wi.mit.edu/). For genes larger than 700 bp, several primer pairs were designed to allow complete coverage of the genes. Primers were designed in the untranslated regions (UTRs) to ensure amplification over the complete lengths of genes. Since the size of the UTRs was not known, the primers were designed in the 30 bp region before and after each gene. The previously synthesized cDNA was used as template for the RT-PCRs. For each gene region, two independent RT-PCR reactions were set up using the following components per 30 µl reaction: 3 µl cDNA, 3 µl 10 x Thermo Buffer (New England Biolabs, Inc., Ipswich, MA, USA), 0.6 µl (10 µM) FP (forward primer), 0.6 µl (10 µM) RP (reverse primer), 0.6 µl dNTPs (metabion

international AG, Martinsried, Germany) (10 mM), 21.9 µl ddH_2O, 0.3 µl (5 units/ µl) Taq-Polymerase (New England Biolabs, Inc.). The PCR program settings were 95 °C 5 min, (95 °C, 1 min; annealing temperature °C, 1 min; 72 °C, 1 min) 35 cycles; 72 °C 10 min. The annealing temperature was adjusted according to the optimal primer requirements (Appendix Table A1). The resulting RT-PCR products were tested on 2.5% MetaPhor® Agarose (Lonza, Rockland, ME, USA) gels for amplification (Figure 16). Amplified RT-PCR products were sequenced twice using for the first independent reaction of a gene region the forward and for the second the reverse primer. Sequencing was outsourced to a company (AGOWA GmbH, Berlin, Germany). The analysis of the editing sites was carried out in MEGA 3.1 (Kumar et al. 2004) by aligning the cDNA sequence results with the corresponding *L. perenne* plastid genome DNA sequences (Diekmann et al. 2009; Chapter 2) and checking visually for SNPs (Figure 17). To assure accuracy all discovered editing sites were confirmed by comparing the genomic DNA tracefiles obtained by sequencing the *L. perenne* chloroplast genome (Chapter 2) with the cDNA tracefiles. Furthermore to avoid overlooking partially edited sites, all tracefiles were completely visually checked for obvious SNPs (Figure 18). Edited sites were counted as partially edited when a peak of the C nucleotide was still visible. This peak had to be clearly distinguishable from the background noise and therefore was only counted when no C nucleotide existed in the directly flanking regions. Background noise from neighbouring C nucleotides could otherwise been accidentally mistaken as a partially edited site.

3.2.4 Phylogenetic analyses of editing sites in the *ndhB* gene

The evolution of editing sites within Poaceae was analysed. The analysis was based on the three genes with the highest number of editing sites, *ndhA*, *ndhB* and *rpoB*. However, the number of reported editing sites is in general very small and also for these three sites restricted. Complete data are only available from the following species: *Hordeum vulgare* (Freyer et al. 1995, Lopez et al. 1997, Zeltz et al. 1993), *Oryza sativa* (Corneille et al. 2000, Inada et al. 2004), *Saccharum officinarum* (Calsa Júnior et al. 2004), *Zea mays* (Mayer et al. 1995) and *Lolium perenne* (this study). Partial data are available for *ndhB* from the non-grass monocotyledonous angiosperm *Acorus calamus* (Freyer et al. 1997). *ndhA*, *ndhB* and *rpoB*-DNA sequences were obtained from whole

genome chloroplast sequences from the following available monocotyledonous species (GenBank accession numbers, http://www.ncbi.nlm.nih.gov/): *Acorus americanus* (EU273602), *Acorus calamus* (AJ879453), *Lemna minor* (DQ400350), *Dioscorea elephantipes* (EF380353), *Zea mays* (X86563), *Saccharum officinarum* (AP006714), *Sorghum bicolour* (EF115542), *Oryza nivara* (AP006728), *Oryza sativa* 'japonica group' (X15901), *Oryza sativa* 'indica group'(AY522329), *Agrostis stolonifera* (EF115543), *Brachypodium distachyon* (EU325680), *Hordeum vulgare* (EF115541), *Triticum aestivum* (AB042240), *Lolium perenne* (this study, AM777385). The orchid *Phalaenopsis aphrodite* was excluded because it lacks the *ndh-gene-complex* and *Festuca arundinacea* was excluded because of afore mentioned (Chapter 2) ambiguities in the DNA sequences of these three genes. Phylogenetic analysis was based on the genomic DNA sequences for all three genes together before editing takes place and performed using PAUP* 4.0 (Swofford 2002) for maximum parsimony with Branch-&-Bound search and MrBayes (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003) for a Bayesian inference approach. For the analysis in PAUP* 4.0 bootstrapping was carried out via Branch-&-Bound searches with 1,000 replicates. Bayesian inference was based on the GTR+G+I-model (**G**eneral **T**ime **R**eversible model + **G**amma-distributed rate variation across sites + a proportion of **I**nvariable sites), using 1,000,000 generations, a sample frequency of 100 and a burnin-value of 1,000. The resulting MrBayes tree-file was used as input for the programs Figtree (Rambaut 2006) and TreeView (Page 1996) and the bootstrap values from PAUP* 4.0 were added manually to the Bayesian phylogenetic tree. Amino acid sequences of *ndhA*, *ndhB* and *rpoB* editing sites before editing had occurred were extracted in MEGA 3.1 (Kumar et al. 2004) and joined manually to the tree graph.

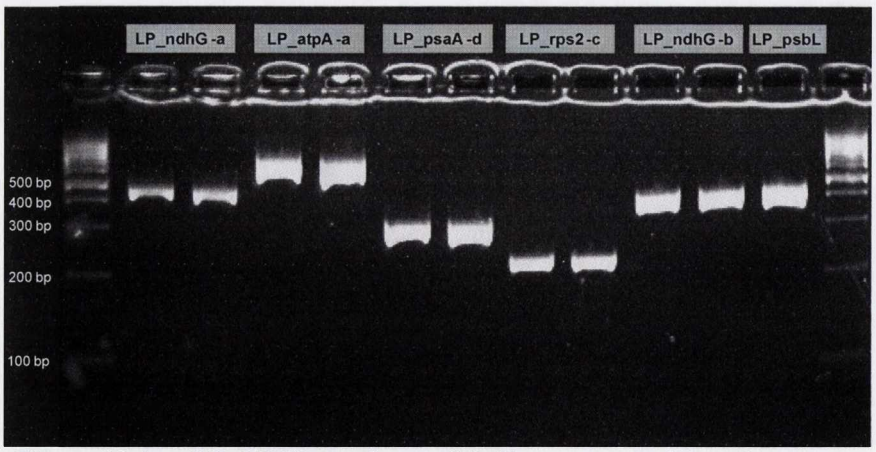

Figure 16 Amplified PCR-products of *Lolium perenne* chloroplast genes ready for sequencing. Size standard = mi-100 bp+ DNA Marker Go (metabion international AG, Martinsried, Germany)

3.3 Results

3.3.1 mRNA editing sites in the chloroplast genome of *L. perenne*

The search of 33 transcripts from the chloroplast genome of *L. perenne* revealed a total of 31 mRNA editing sites from 18 out of the 33 successfully amplified transcripts (Table 6). This number of editing sites equates to 0.15% of codons or 0.02% of nucleotides. 27 editing sites affected the second position of codons and, of those, 26 were flanked by a pyrimidine nucleotide on the 5' side and 24 by a purine on the 3' side. All observed editing sites involved C to U changes. Most frequently, editing resulted in changes of the amino acids serine or proline to leucine. Four editing sites (*ndhA*, site 4; *ndhG,* 5' UTR; *rpoB*, site 4; *rps14*), previously observed in other Poaceae species, were not edited in *L. perenne*. For three of them, the conserved nucleotide U already exists at the DNA level. Site 4 in *rpoB* is not edited, although C is encoded in the genomic DNA. Analysis of editing in the *ndhA* gene was not completed because several primers failed to amplify the region and the obtained sequencing products did not have the full gene length. Thus site 4, which was observed in *O. sativa* (Corneille et al. 2000), *S. officinarum* (Calsa Júnior et al. 2004) and *Z. mays* (Maier et al. 1995), could not be analysed. However, in *L. perenne* this position is a TTC (phenylalanine) codon, which is the same as the codon that is formed by mRNA editing in the three other species. Thus editing is unlikely to happen at this position in *L. perenne*.

Further comparisons to editing sites in other species revealed five new editing sites so far unique to *L. perenne*. Four of these sites are in three genes (*ndhK, psbJ, psbL*) in which editing has not been reported before in Poaceae species. Two out of the five new editing sites result in a synonymous change but three (1 in *ndhK*, 1 in *psbJ,* 1 in *rpoC2*) result in changes of the amino acid to leucine (Table 6).

Partial editing was observed at 14 editing sites (nine genes: *matK, ndhA, ndhD, ndhF, psbL, rpl2, rpl20, rpoB, rpoC2*) (Table 6) (Figure 18). In most of these editing sites, the number of incompletely edited transcripts is small. However, approximately one half and one third of the *matK* and *psbL* transcripts, respectively, are not edited.

3.3.2 Phylogenetic analyses of editing sites in three genes

ndhA and *rpoB* in *L. perenne* each have three editing sites that were previously already observed in other monocotyledonous species (all C-U changes; Table 6), but for both genes the fourth site was absent in *L. perenne*.

Table 6 RNA editing sites found in the chloroplast genome of *Lolium perenne* in comparison to the editing sites found in other grasses. 'Codon position' refers to the amino acid position edited in the analysed gene using the *L. perenne* chloroplast genome as reference. 'Editing sites' specifies the position of the editing site within the *L. perenne* chloroplast genome.

gene	site	codon position	editing sites	edited codon	amino acid change			*Lolium perenne*	*Hordeum vulgare*	*Oryza sativa*	*Saccharum officinarum*	*Zea mays*
atpA	1	383	35112	tCa	S	→	L	+		+	+	+
matK	1	420	1993	Cat	H	→	Y	+[a]	+[3]	+[1]		
ndhA	1	17	111250	tCa	S	→	L	+[a]	+[4]	(-)	+	+
	2	158	112355	tCa	S	→	L	+	+[4]	+	+	+
	3	188	112777	tCa	S	→	L	+[a]	+[4]	+[1]	+	+
	4	357		tCc	S	→	F	(-)	+[4]	+	+	+
ndhB	1	50	87743	tCa	S	→	L	+	+	(-)	(-)	(-)
	2	156	87425	cCa	P	→	L	+	+	+	+	+
	3	196	87306	Cat	H	→	Y	+	+	+	+	+
	4	204	87281	tCa	S	→	L	+	+	+[a]	+	+
	5	235	87188	tCc	S	→	F	+	+	+	(-)	(-)
	6	246	87155	cCa	P	→	L	+	+	+	+	+
	7	277	86347	tCa	S	→	L	+	+	+	+	+
	8	279	86341	tCa	S	→	L	+	+	+	(-)	(-)
	9	494	85696	cCa	P	→	L	+	+	+[a]	+	+
ndhD	1	295 (293)	107165	tCa	S	→	L	+[a]	+[5]	+	+	+
ndhF	1	21	103675	tCa	S	→	L	+[a]		+	+[a]	+
ndhG	1	116	109624	cCa	P	→	L	+		+[1]		(-)
5'UTR		-10						(-)	(-)[6]	+	+	+
ndhK	1	2	49367	gtC	V	→	V	+				
	2	43	49245	cCa	P	→	L	+				
petB	1	204	72259	cCa	P	→	L	+		(-)	+	+
psbJ	1	20	61111	cCt	P	→	L	+				
psbL	1	37	61339	ttC	F	→	F	+[a]		(-)		
rpl2	1	1	82030	aCg	T	→	M	+[a]	+	+[a]	+[a]	+
rpl20	1	103	66009	tCa	S	→	L	+[a]		(-)	+[a]	+
rpoB	1	156	19737	tCa	S	→	L	+[a]	+[2]	+[a]	+	+
	2	182	19815	tCa	S	→	L	+[a]	+[2]	+[a]	+	+
	3	187	19830	tCg	S	→	L	+[a]	+[2]	+[a]$_{uCa}$	+	+
	4	206		cCg	P	→	L	-	-[2]	-	-	+
rpoC2	1	925		tCg	S	→	L	(-)		(-)	+	+
	2	1320	28731	tCa	S	→	L	+[a]				-
rps8	1	61	76422	tCa	S	→	L	+		+	+	+
rps14	1	27		tCa	S	→	L	(-)		+	+[a]	+
ycf3	1	15	43599	tCc	S	→	F	+		(-)	-	+
	2	62	42700	aCg	T	→	M	+		+[a]	+	+[a][7]

(- : no editing although C encoded in DNA; (-): no editing, U encoded in DNA; blank space: editing not yet determined/no information available; [a] : partially edited; grey highlighted: unique for *Lolium perenne*. *Hordeum vulgare* = Freyer et al. 1995, [2] : Zeltz et al.1993, [3] : Vogel et al.1997, [4] : Lopez et al. 1997, [5] : del Campo et al. 1997, [6] : Drescher et al. 2002; *Oryza sativa* = Corneille et al. 2000, [1] : Inada et al. 2004; *Saccharum officinarum* = Calsa Júnior et al. 2004; *Zea mays* = Maier et al. 1995; [7] : Ruf & Kössel et al. 1997)

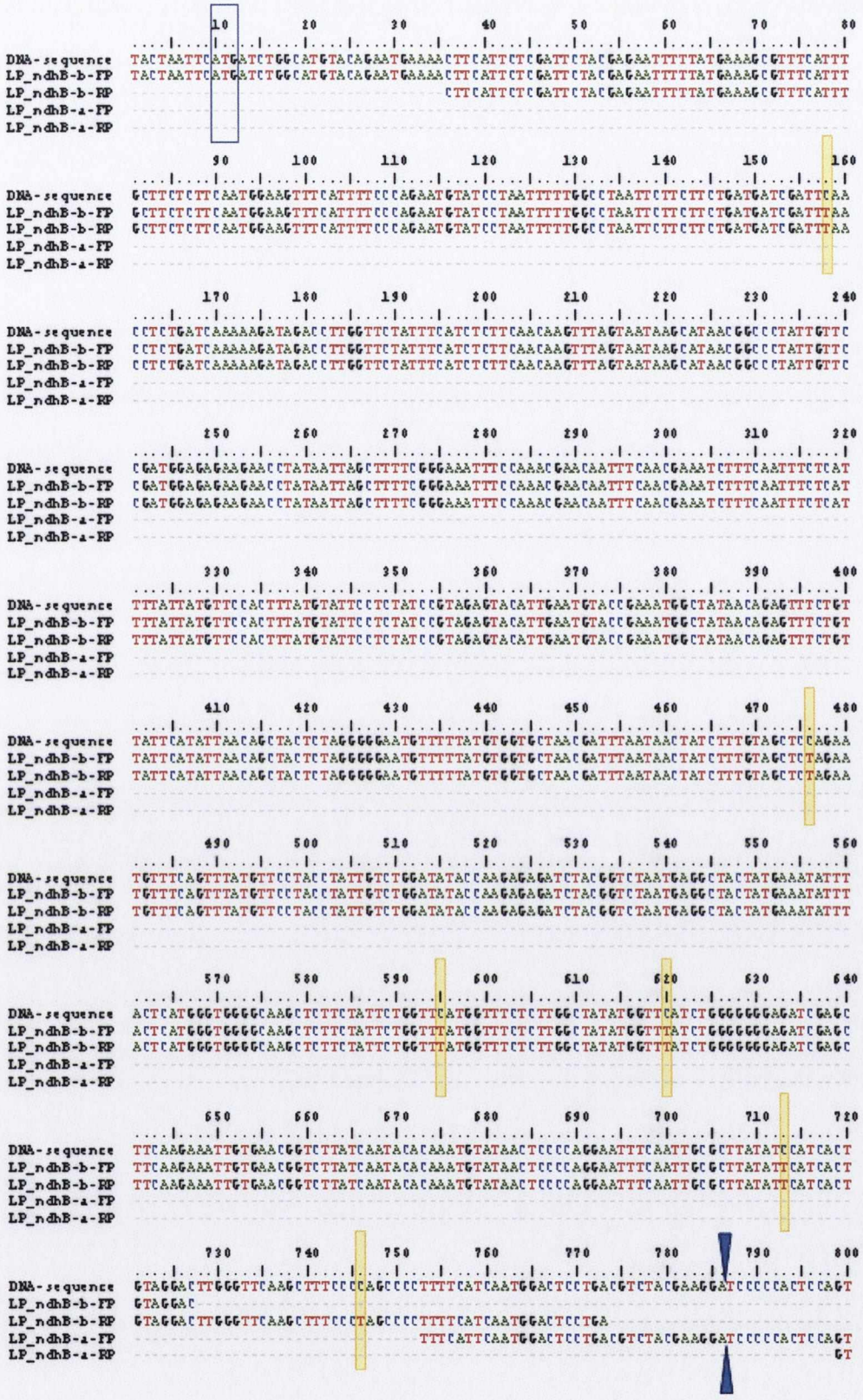

Figure 17 Alignment of *ndhB*-DNA sequence with cDNA sequences derived from two different primer sets. Yellow boxes highlight editing sites, blue boxes highlight start and stop codon of *ndhB*-gene, blue arrows show position of intron (sequence not shown), program used: Bioedit version 7.0.5.3 (Hall 1999).

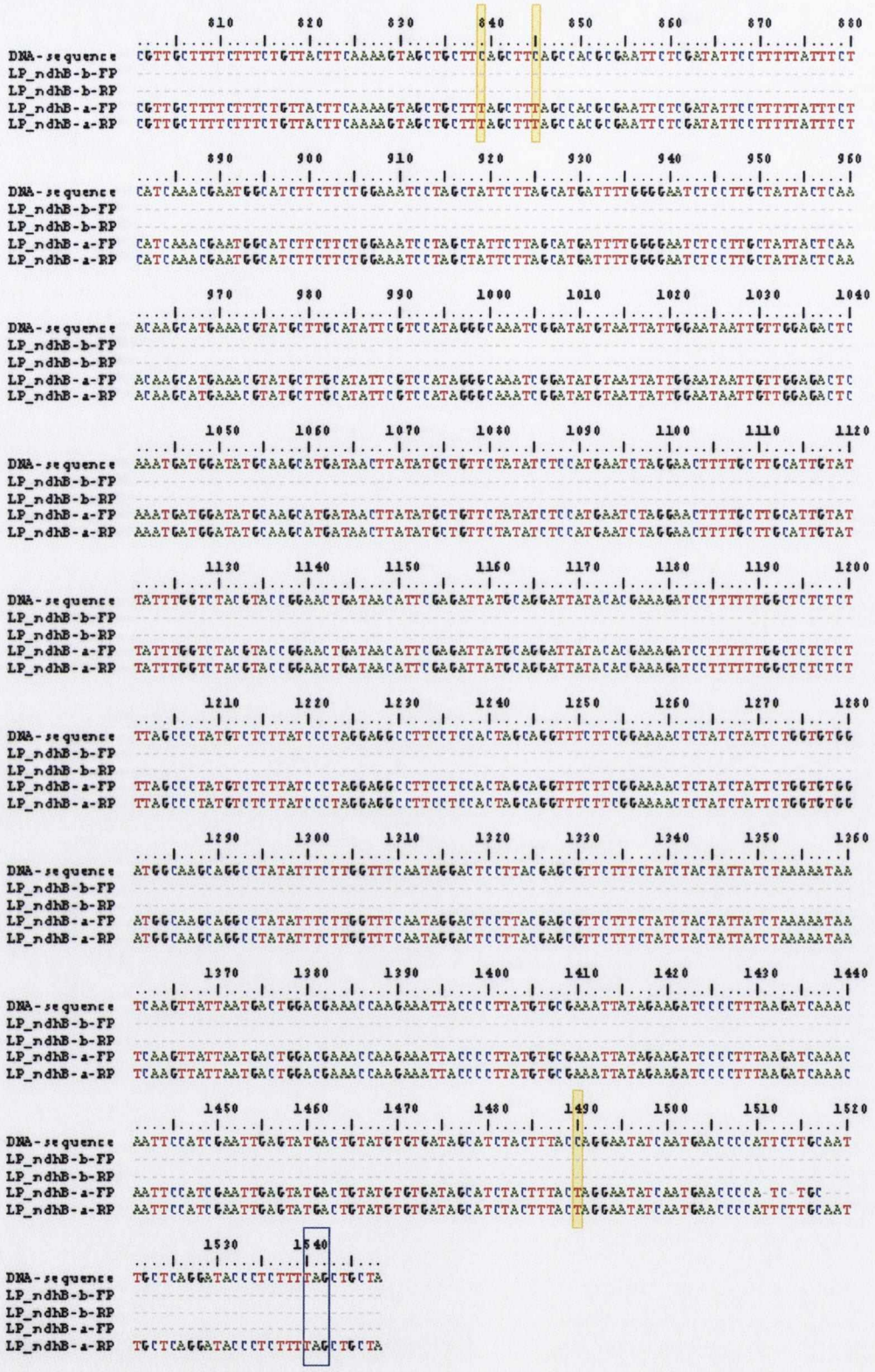

Figure 17 *continued*

Editing site *ndhA*-4 in *L. perenne* already codes for the conserved amino acid and although *rpoB*-4 does not have the conserved amino acid, editing was not observed in this study. *ndhA*-1 (tCa; where the upper case base is the edited base of the codon) exists in all so far analysed species except of *O. sativa*, where the conserved amino acid is already encoded in the genomic DNA (Corneille et al. 2000). *rpoB*-4 was only observed in *Z. mays* (Maier et al. 1995). The *ndhB* gene in *L. perenne* has the same nine editing sites (all C-U), as previously reported for the *H. vulgare ndhB* gene (Freyer et al. 1995). Editing site *ndhB*-1 (tCa) is absent however from *O. sativa, S. officinarum* and *Z. mays* (Table 6). *ndhB*-5 (tCc) and *ndhB*-8 (tCa) are absent from *S. officinarum* and *Z. mays*. All absent sites already have the nucleotide T in the genomic DNA sequence. Figure 19 illustrates the results obtained from the phylogenetic analysis. The analysis resolved the subfamilies of the grasses with high posterior probability (PP) and bootstrap percentage support (BP) (Panicoideae 1.00PP, 100BP; Ehrhartoideae 1,00PP, 100BP; Pooideae 1.00PP, 99BP). Pooideae (1.00PP, 100BP) were resolved as sister to the Ehrhartoideae + Panicoideae clade (0.80PP, 55BP). With *Acorus* as the outgroup (chosen according to Chase 2004), *Dioscorea* and *Lemna* were resolved successively as sister to the clade comprising the grasses (1.00PP, 79BP and 1.00PP, 100BP, respectively). Although the tree distinguishes between the Ehrhartoideae and Panicoideae the support for this relationship is weak. However, the species relationships within these subfamilies are highly supported (Figure 19). The gene tree, based on 15 monocotyledonous species and three protein-coding genes, is also in agreement with a phylogenetic tree based on 76 protein-coding genes of 19 monocotyledonous chloroplast genomes presented in Chapter 6 of this thesis.

Plotting the nine amino acids of the observed *ndhA*, *ndhB and rpoB* editing sites for all 15 species analysed next to the terminal branch of the phylogenetic tree of the three genes revealed that the amino acids are identical within Panicoideae and within Ehrhartoideae. Also within the Pooideae the amino acids are well conserved especially for the genes *ndhB* and *rpoB*. The only differences found in the *ndhB* gene were observed for site *ndhB*-2 in *B. distachyon*. In the *ndhA* gene however the variation is bigger. *ndhA*-1 is lost in *A. stolonifera* and *T. aestivum*, as is the case for *ndhA*-2 in *A. stolonifera* and n*dhA*-4 in *L. perenne*. Although information on editing sites is not available for eight species (*A. americanus*, *A. stolonifera*, *B. distachyon*, *D. elephantipes*, *L. minor*, *O. nivara, S. bicolor*, *T. aestivum*,), conservation among closely

related species (within a subfamily) was observed. Therefore it is assumed that the species have adapted to maintain the prevalent amino acid by modifying via editing the mRNA that does not have the 'correct' codon. For example for editing site *ndhB*-1 (Figure 19), it is known that editing occurs in *H. vulgare* and *L. perenne* (Table 6) but it has never been observed in *O. sativa, S. officinarum* and *Z. mays* which transcribe for the conserved amino acid already in the genomic DNA. Therefore it is assumed here that editing at *ndhB-1* does not occur in Panicoideae and Ehrhartoideae because all analysed species already possess the conserved codon. But it is assumed that it happens in the Pooideae as well as the outgroup species because they do not code for the conserved amino acid. Based on this assumption, editing sites that exist in the outgroup were lost in Panicoideae and Ehrhartoideae, but gained in Pooideae. Similar parallel mutations (e.g. *ndhB*-5: edited in *D. elephantipes* (outgroup) and Ehrhartoideae species and Pooideae species) can be found.

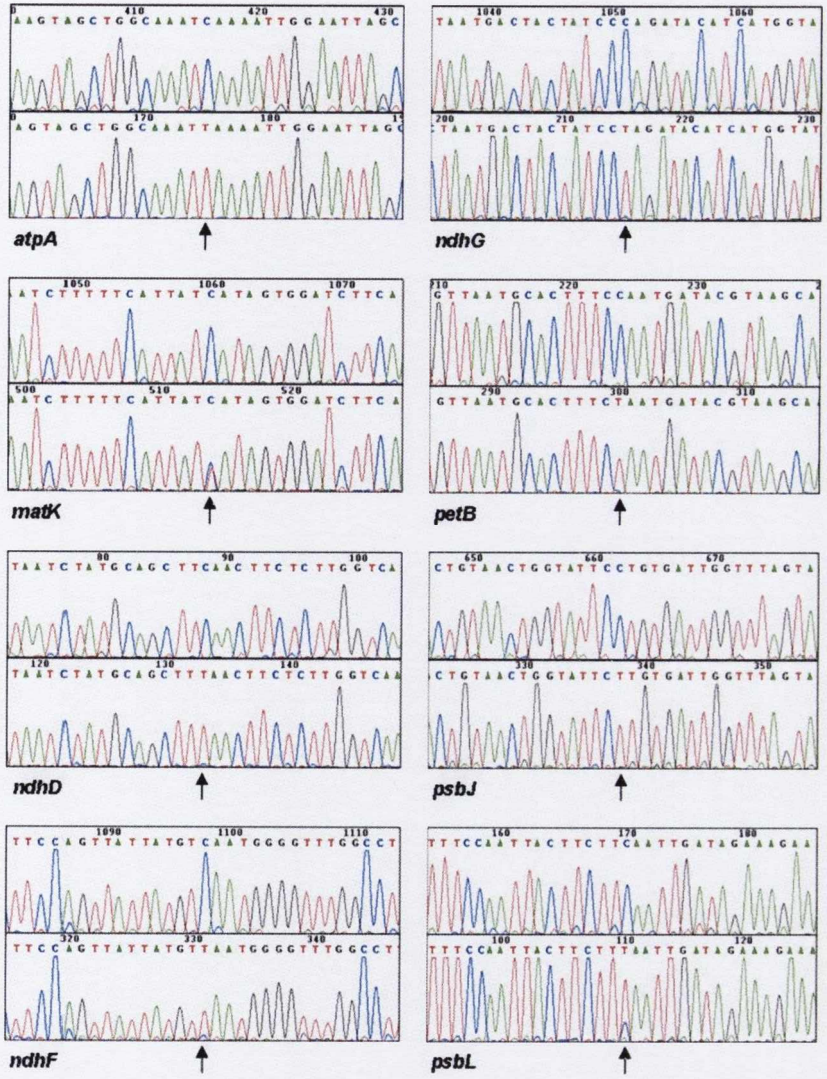

Figure 18 Sequencing trace files of editing sites in the chloroplast genome of *Lolium perenne*. Upper files show DNA sequence, lower files show cDNA sequence, editing sites are indicated by arrows.

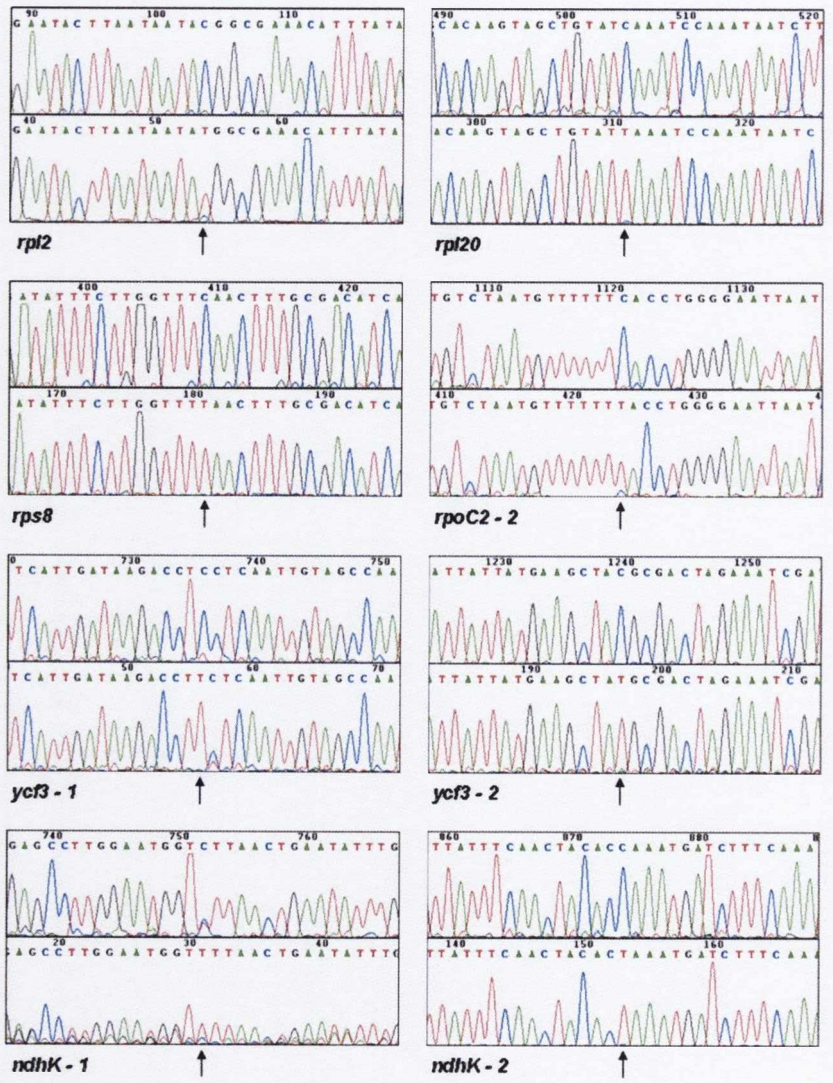

Figure 18 *continued*

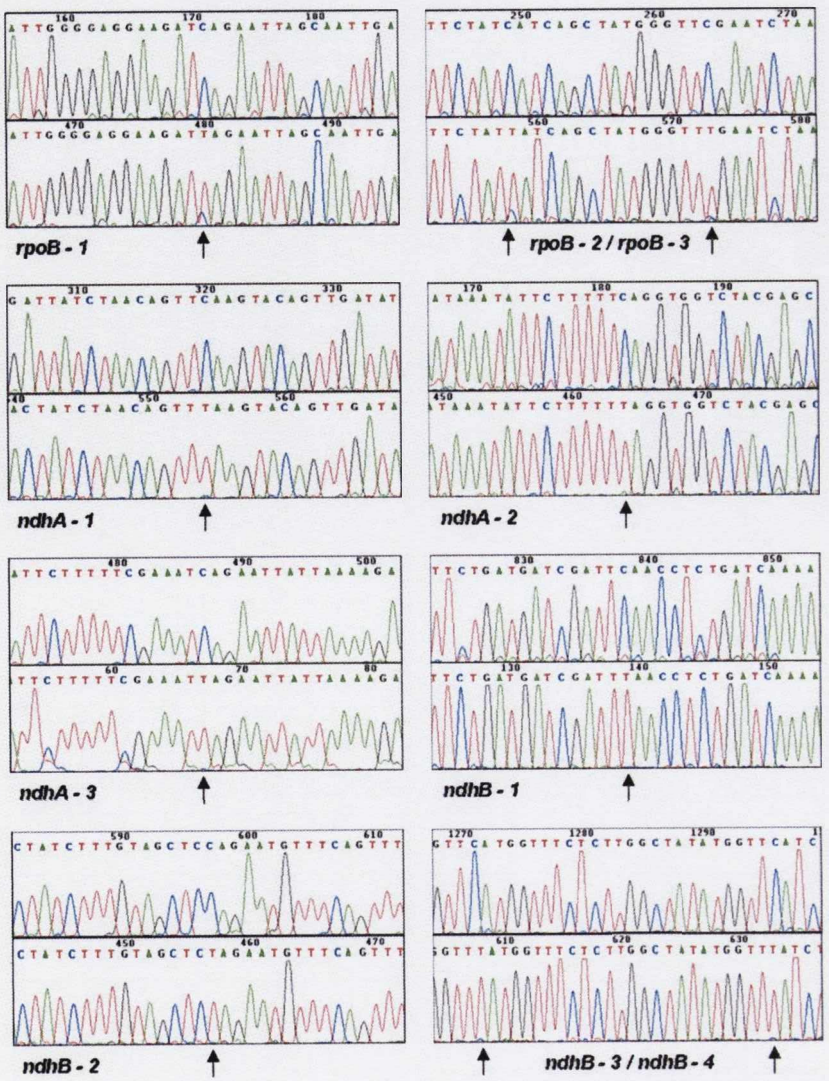

Figure 18 *continued*

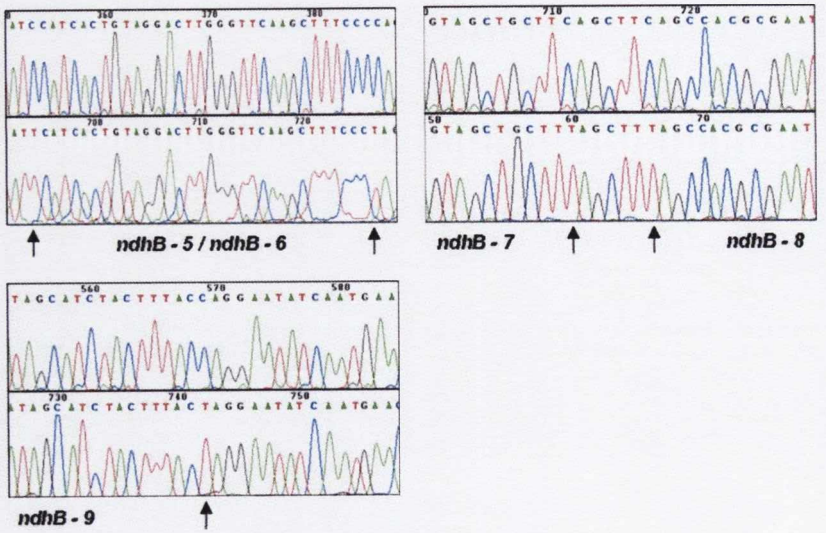

Figure 18 *continued*

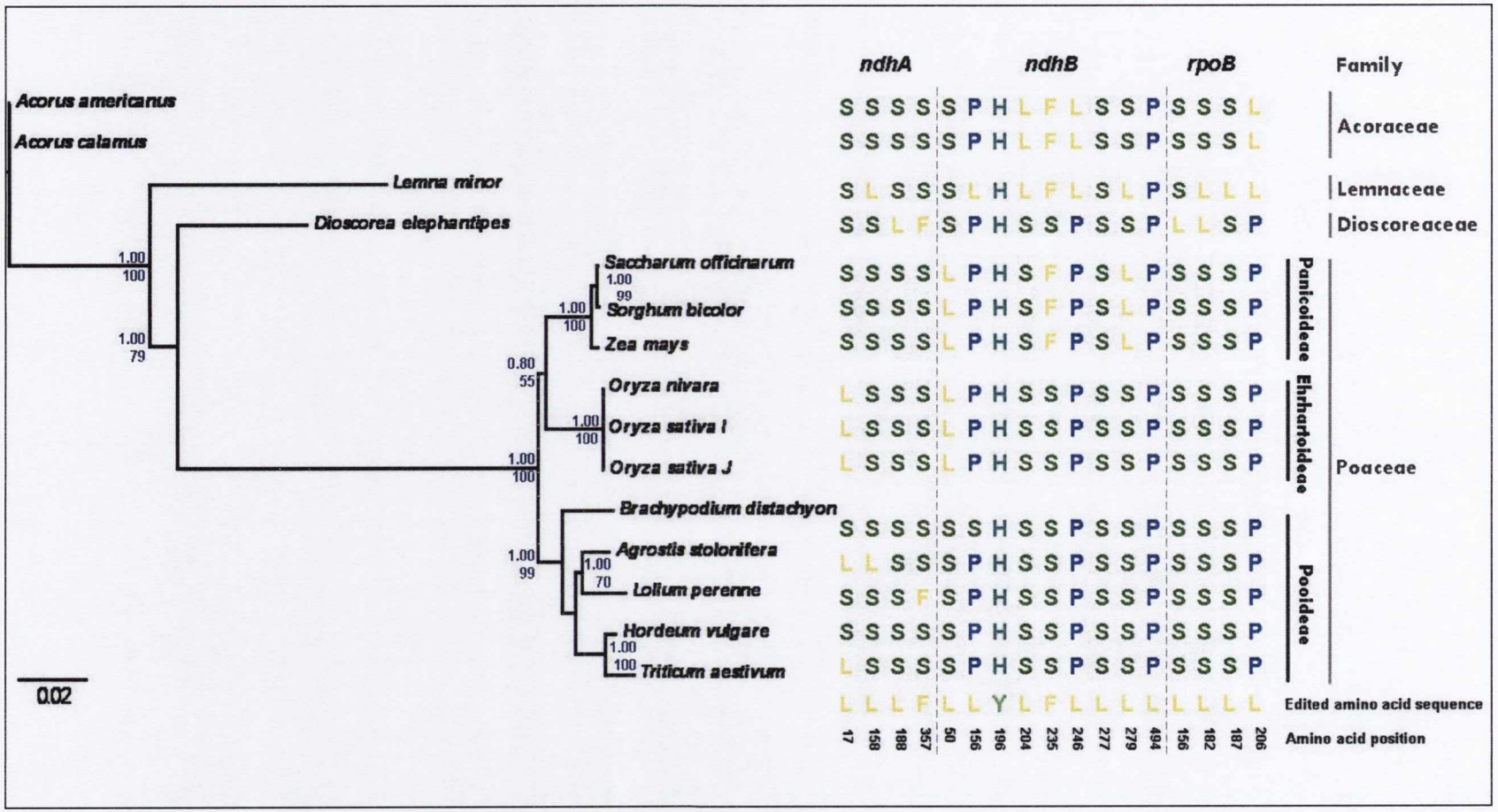

Figure 19 Bayesian consensus tree derived from the phylogenetic analysis of the *ndhA*, *ndhB* and *rpoB* gene. Known editing sites (amino acids shown) are aligned to the tree. Above branches are posterior probability values, below bootstrap percentage values. S = serine, L = leucine, P = phenylalanine, H = histidine, Y = tyrosine. Tree was obtained by using FigTree (Rambaut 2006), aminoacids were extracted from MEGA 3.1 (Kumar et al. 2004). The color of the amino acid letters was automatically obtained from MEGA 3.1 during the extraction and refers to the biochemical properties of the different amino acids.

3.4 Discussion

3.4.1 mRNA editing sites in *L. Perenne*

In this study, 33 out of 76 possible protein-coding chloroplast genes were analysed based on previous publications for other grass chloroplast genomes (Calsa Júnior et al. 2004; Corneille et al. 2000; Freyer et al. 1993; Maier et al. 1995; Zeltz et al. 1993). 31 mRNA editing sites in 18 genes of the chloroplast genome of *L. perenne* were detected. This number is in the expected range of editing sites found so far in other genomes (Tillich et al. 2005) and also within the expected 'editing sites : chloroplast genome ratio' (Bock 2000). Therefore it is likely that this study has reported the majority of the *L. perenne* chloroplast RNA editing sites. However, the possibility of finding additional sites in the unanalysed 43 genes, as well as in non-coding regions, still exists. For example, this study revealed five new editing sites (*ndhK*-1, *ndhK*-2, *psbJ*-1, *psbL*-1, *rpoC2*-2) that have never been observed in other Poaceae species and which are so far unique to *L. perenne*. Four of these (all except *rpoC2*) new sites were found in genes for which editing has not been observed before in Poaceae. These genes were included in this study only due to detected variations in ESTs (Saski et al. 2007). However, only site *ndhK*-1 predicted from ESTs for *S. bicolor* (Saski et al. 2007) was shown to be a real editing site in *L. perenne*. New editing sites were found also in *rpoC2*, *psbJ* and *psbL* but the exact position of editing sites within these genes was not predicted correctly from ESTs. Predicting mRNA editing sites based solely on published EST sequences as done for *A. stolonifera*, *S. bicolour* and *H. vulgare* therefore is not sufficiently accurate. Timme et al. (2007) also showed that editing sites can be easily overlooked, or SNPs can be falsely interpreted as editing sites when using ESTs only. Only six of the genes analysed via EST comparisons by Saski et al. (2007) have had editing sites experimentally confirmed in other species. The variations found in EST sequences by Saski et al. (2007) were generally not based on C-U changes and thus are highly unlikely to be editing sites. Most of the variations found by comparing ESTs to cpDNA sequences were very likely based on the use of different varieties, or on poor quality sequencing data. In this study SNPs (Chapter 2) and editing sites were analysed in the same variety of *L. perenne* to ensure that newly detected sites, with either complete or partial editing, were identified correctly as editing sites and not accidentally mistaken as SNPs, or vice versa.

Nevertheless, based on the results from Saski et al. (2007) more genes were included into this study and hence new edited genes detected. Editing sites can generally be predicted by comparing protein-coding DNA sequences with the corresponding DNA sequence of the liverwort *Marchantia polymorpha*. Editing does not occur in *M. polymorpha* (Ohyama 1996). Thus the degree of conservation between *M. polymorpha* and the species to be analysed can increase at candidate editing sites if mRNA editing takes place. Candidate editing sites need then to be tested experimentally. This approach is efficient at identifying functional mRNA editing sites, but it does not detect silent sites, as for example the two found in this study (*ndhK*-1 and *psbL*-1) or the one found in the *atpA* gene of tobacco (Hirose et al. 1996). Editing sites in non-coding regions will also not be detected. Thus, so far only a few editing sites in non-coding regions have been reported as such as the *psbL/J* intergenic spacer region of *Gingko* (Kudla & Bock 1999) or in the 5' UTR of *ndhG* in various Poaceae species (Drescher et al. 2002). Three explanations for editing sites within non-coding regions are possible: a) although in a non-coding region editing of this site is of functional relevance; b) this editing site is an evolutionary relic but is maintained due to lack of selective pressure and c) this site is edited due to coincidental sequence similarity of its recognition site with the recognition site of another functionally important editing site (Kudla & Bock 1999).

Recently a web-based program called CURE (Du et al. 2009) was designed to detect mRNA editing sites within chloroplast genome sequences. Using the *L. perenne* chloroplast genome sequence as an input-file, this program detected 34 editing sites (using the default settings; *ndhA*-1 – *ndhA*-4, *ndhD*-1, *ndhF*-1, *ndhB*-1 – *ndhB*-9, *rpl2*, *rps8*-1, *rpoA*-1, *rpl20*-1, *rpl33*-1, *psbF*-1, *ycf3*-2, *matK*-1, *petN-trnC*, *rpoB*-1 – *rpoB*-4, *rpoC2*-1, *rpoC2*-2, *atpA*-1, *rbcL*-1, *rbcL*-2, *petB*-1). A comparison of the experimentally found sites with the predicted sites from CURE showed that only 25 of the 31 sites were found by both approaches, while the experimental approach detected six and CURE nine more sites. CURE did not detect the four editing sites in the genes *ndhK*, *psbJ* and *psbL*, for which editing has not been observed before in grasses. Comparing the nucleotides of the five new editing sites found in *L. perenne* with the corresponding nucleotides in *M. polymorpha* reveals that editing restores conserved nucleotides/codons for editing site *psbJ*-1, *psbL*-1, *rpoC2*-2 and *ndhK*-2 and thus supports the results obtained in this study. However, editing of *ndhK*-1 increases codon diversity instead of reducing it, because the codons found at DNA level in *M.*

polymorpha and *L. perenne* were identical. This is not an unusual result because similar sites have already been observed in other species (Inada et al. 2004). The occurrence of silent editing sites can be, for example, explained by the involvement of flanking regions showing accidentally a similar or the same DNA sequence as the one for *cis*-elements of non-silent sites. These accidentally similar flanking regions might also lead to the observation of editing sites in which modification results in a decrease of conservation.

Comparing the additional editing sites found by CURE with the *M. polymorpha* sequence reveals, however, that most of the additional CURE sites are not very likely to be edited because they have already conserved nucleotides. On this basis, only two of the additional nine sites found by CURE could be possible editing sites; one at position 17,309 and one at position 61,511 of the chloroplast genome. Nucleotide 17,309 is in the intergenic spacer region of *petN* and *trnC* where editing is unlikely to have a functional impact. Nucleotide 61,511 is located in the *psbF* gene but this editing site would be silent because it is at the third position of the codon. The authors of CURE report that this program has a sensitivity of 70% when using whole chloroplast genome sequences (Du et al. 2009). Considering that from 34 predicted sites, only seven seem unlikely to be edited, this is in agreement with the author's proposal of 70% sensitivity. However, CURE was not able to detect six editing sites which were experimentally confirmed and which were also in agreement with the *M. polymorpha* sequence or results from previous studies. The program CURE is based on previously reported editing sites from a range of angiosperm species (Du et al. 2009). Considering that RNA editing has been systematically analysed for only eleven species (Guzowska-Nowowieskja et al. 2009) and not always completely (Corneille et al. 2004), this foundation might not be good enough to detect all editing sites. Therefore predicting editing sites based on the *M. polymorpha* sequence or CURE or a combination of both does not guarantee that all sides will be found and the predicted sites will be correct. It is very likely that editing sites in rarely edited genes or in non-coding regions will be overseen. Thus the experimental analysis is important to confirm sites and to detect those which can not be predicted.

There was evidence that nearly half of all editing sites detected in *L. perenne* were partially edited. Only six of these sites from four genes (*rpoB*, *rpl2*, *rpl20* and *ndhF*)

had previously been observed to be partially edited (Corneille et al. 2000; Calsa Júnior et al. 2004). Differences in editing efficiency were observed for silent editing sites (Kempken et al. 1991), different tissues (Ruf & Kössel 1997) and tissues of different developmental stages (Miyata & Sugita 2004). This study used young leaves (two to three weeks old) for which the editing efficiency is supposed to be between 80% and 100% (Peeters & Hanson 2002). The observed editing efficiency was indeed in most sites higher than 80%. Only the transcripts of *psbL* and *matK* showed greater amounts of partial editing with around 30% and 50%, respectively (Figure 18) Evaluating the efficiency of partial editing is not an easy task due to background noise in some trace files. Mower and Palmer (2006) counted editing sites as partially edited when peaks of both the C and T nucleotide were present and clearly above the background noise. However they were also aware of their possible underestimation of partially edited sites. To evaluate partial editing in our *Lolium* study, we looked for the clear presence of both peaks and the absence of a C nucleotide in the direct flanking regions whose background could cause a misinterpretation. Thus the amount of partially edited sites is more likely to be estimated correctly. However, the possibility of overlooking partially edited sites can also not be completely excluded especially in those cases where the amount of non-edited transcripts is lower than 10%.

Although most of the observed partially edited sites might be due to the use of young leaf material, partial editing was also observed at one of the two silent editing sites. The *psbL*-1 editing site was one of the editing sites with the lowest editing efficiency. Although the trace files for the other silent editing site *ndhK*-1 are of low quality, there also seems to be the indication that this site is partially edited. Thus, the observations of partially editing at silent editing sites (Kempken et al. 1991) are supported by this study. The emergence of silent editing sites is thought to have happened accidentally, when the flanking region of a C nucleotide showed a similar or the same DNA sequence as the one for a *cis*-element of a non-silent site (Chateigner-Boutin & Hanson 2003). If a silent and a non-silent editing site share the same *cis*- and *trans*-factors there could be a rather low selective pressure against the silent site due to the fact that both transcripts of the silent site are completely functional and the editing factors are necessary for modifying the non-silent site. Under these conditions the presence of edited and non-edited transcripts for silent editing sites is not surprising.

The significance of non-silent partial editing is not completely understood, but several hypotheses exist. In case of *matK* it could be possible that different forms of that protein exist but so far no supporting evidence has been found (Guzowska-Nowowiejska et al. 2009). Editing might also act as regulation factor at the transcript level (Guzowska-Nowowiejska et al. 2009) as suggested for the *rpoA* gene. By partial editing, this gene could possibly regulate the abundance of active RNA polymerase (Hirose et al. 1999). Another explanation for the occurrence of partially edited sites is that these editing sites are in the process of being eliminated over time (Guzowska-Nowowiejska et al. 2009). A consequence of partial editing is that some proteins have probably incorporated the 'wrong' amino acid; therefore selection pressure for C to T transitions on genomic DNA level could increase and eliminate editing sites (Guzowska-Nowowiejska et al. 2009). It must also be taken into account that editing in general may have selective disadvantages for the organisms that use it because it should be more efficient for them not to edit. This hypothesis would further support the theory that partial editing in non-silent editing sites is, at least to some extent, caused by ongoing selection against the editing mechanism itself. Another explanation for the observation of partially edited sites could be based on the usage of a plant population of several individuals for this analysis. If selection is working against the editing mechanism, there could be individuals within a population that have lost these editing sites and show the T nucleotide already at genomic DNA level. In that case, the observed 'partially edited' sites would possibly not only show a variation in editing efficiency but also reflect a mixture of transcripts for which editing is and is not required. However, if some individuals would have lost their editing sites and others still have them, then this would have been detected in the single nucleotide polymorphism study carried out in Chapter 2, but none of the detected editing sites was found to be polymorphic at genomic DNA level. Therefore the partially edited sites reflect only a variation in editing efficiency and future studies need to be carried out to fully understand the meaning and importance of partial editing.

3.4.2 Phylogenetic analyses of editing sites in three genes

The *ndhA, ndhB* and *rpoB* gene were chosen for an approach to study the evolution of editing sites within monocotyledonous species, especially grasses. These genes, particularly *ndhB* are suited for this kind of analysis due to their abundance of editing

sites (Freyer et al. 1995). Unfortunately editing sites for most species included in this study needed to be predicted based on the observed editing sites in *H. vulgare, L. perenne, O. sativa, S. officinarum* and *Z. mays*, because experimental results are not available. Additional information was only available for *A. calamus* for which it was experimentally confirmed that editing occurs at sites *ndhB*-2, *ndhB*-3, *ndhB*-7 and *ndhB*-8 (Freyer et al. 1997). Freyer et al. (1997) included the monocot *Narcissus pseudonarcissus* in their study of *ndhB*, which contained the editing sites *ndhB*-2 to *ndhB*-7 and an additional site in codon 181. This study showed that predicting editing sites based on the available but limited information from other species is possible and leads to an under- rather than an over-estimation of editing sites in monocotyledonous species. It is very likely that editing takes place at most of the predicted editing sites, however, this study (*rpoB*-4) and previous studies also showed that such predicted editing sites are not always edited (Freyer et al. 1997). Point mutations that affect regulatory *cis*- and *trans*-elements can lead to the non-functionality of an editing site (Freyer et al. 1997). Tillich et al. (2006) analysed editing sites in five land species (*Arabidopsis thaliana, Atropa belladonna, Nicotiana tabacum, Pinus thunbergii* and *Zea mays*) and found in total 155 sites of which more than half were unique to the individual species. However, as a mixture of monocotyledonous (*Z. mays*) and dicotyledonous (*A. thaliana*, *A. belladonna*, *N. tabacum*) angiosperms as well as one gymnosperm (*P. thunbergii*), these species were in general only distantly related.

The analysis as carried out in this study has focussed on a taxonomic family (Poaceae) instead of a group with much higher taxonomic rank (e.g. seed plants) and therefore reveals a higher degree of editing site conservation. For example, at the subfamily level most editing sites were conserved especially for Panicoideae and the Ehrhartoideae. Editing sites in Pooideae were in general also conserved, but one difference was detected for *ndhB* (*Brachypodium distachyon*: amino acid serine at position 156 instead of proline) and several for the *ndhA (Agrostis stolonifera*: amino acid leucine at positions 17 and 158 instead of serine; *Triticum aestivum*: amino acid leucine at position 17 instead of serine; *Lolium perenne*: amino acid phenylalanine at position 357 instead of serine). Serine is the most edited amino acid and the possibility exists that editing site *ndhB*-2 in *B. distachyon* is also edited to leucine. The differences found in the *ndhA* gene are very likely due to the loss of editing sites because the conserved nucleotide is already present at the genomic DNA level and editing is not required.

The phylogenetic tree based on the three chloroplast genes *ndhA*, *ndhB* and *rpoB* is fully resolved and all sister group relationships strongly supported with only one exception. Pooideae are in this analyses sister to a Panicoideae + Ehrhartoideae group, which is in contrast to some earlier studies (e.g. GPWG 2001), but this relationship yields only low bootstrap and posterior probability support. The tree is in general agreement with trees from Chase (2004) and Bouchenak-Khelladi et al. (2008) regarding the relationship of the orders, families and tribes and is completely consistent with the phylogenetic tree obtained in Chapter 6 when using 76 protein-coding genes in a Bayesian inference analyses.

Nevertheless this tree enables the construction of hypotheses about the loss and gain of editing sites based on parsimony principles – a scenario that explains the complete evolutionary process where the smallest amount of changes is most likely to be true (Nei 1996). At editing site *ndhB*-1, for example, serine is edited to leucine in Pooideae but not in Panicoideae or Ehrhartoideae. It is however, edited from serine to leucine in all the outlying taxa (*Dioscorea*, *Lemna* and *Acorus*). If serine is assumed as the ancestral state, it is more parsimonious to assume that there has been a gain of leucine in the ancestor of Panicoideae + Ehrhartoideae (loss of the editing site; one change) as it is to infer that there was a gain of leucine in the ancestor to all grasses and then a loss to serine in only Pooideae (two changes). Focussing on editing site *ndhB*-5 serine is found in all grasses except in the Panicoideae and found only in one outgroup taxon, *Dioscorea elephantipes*. Assuming that phenylalanine was the ancestral state then serine would have evolved independently in *D. elephantipes (gain of editing site; 1 change)* and the two branches leading to Pooideae (1 change) and Ehrhartoideae (1 change), respectively (3 changes in total). But it can also be assumed that serine was ancestral and phenylalanine evolved independently in all outgroup genera and Panicoideae, and got lost again in *D. elephantipes* (3 steps in total). Similar results can be obtained when considering a tree that is completely consistent with the results from GPWG (2001) thus Panicoideae are sister to Ehrhartoideae + Pooideae. In case of site *ndhB-1* again both scenarios are equally parsimonious: Considering serine as ancestral state, it could have been lost in the branch leading to the Poaceae and gained in the branch leading to the Pooideae and if leucine was present in the ancestor serine could have been gained independently in the branch leading to the non-Poaceae species and in the branch leading to the Pooideae species. For site *ndhB*-5 the most parsimonious assumption

would be that phenylalanine is the ancestral state and serine evolved twice, once in *D. elephantipes* and once in the branch leading to Ehrhartoideae and Pooideae. If serine would have been present in the ancestor phenylalanine needs to evolve independently in the outgroup species and the Panicoideae and would have been lost again in *D. elephantipes*. These examples based on the *ndhB* gene demonstrate equally the conflict between the gain and loss of editing sites (dependent also on the tree topology). Considering the possibility that the requirement of RNA editing might have a selective disadvantage in general and additionally that the mechanism itself is based on different factors that have to evolve simultaneously, it is probably more likely that editing sites were lost during evolution rather than gained. However, taking a closer look at the Pooideae editing sites in the *ndhA* gene shows that this question can not be answered so easily. *ndhA*-1 does not exist in *A. stolonifera* and *T. aestivum*. While it could be assumed that the presence of this editing site is the ancestral state because it exists in all outgroup species, and the loss of the site happened independently three times – once on the Ehrhartoideae branch, once in the Aveneae tribe (including *Agrostis*) and once in the Triticeae tribe (including *Triticum*), this hypotheses seems to become unstable when taking into account that *H. vulgare* of Triticeae does not have this editing site.

The conflict between loss and gain of editing sites is not surprising considering the fact, that the editing site itself consists of only one nucleotide and also that *cis*-elements, which are important to detect the editing site, are only based on a few nucleotides (Hayes & Hanson 2007). Thus, point mutations can easily result in the loss of editing sites, either due to a substitution in the editing site itself or a mutation in the *cis*-element which can result in the loss of the transfactor binding site. Also the transfactor itself might undergo mutations and not bind anymore to the *cis*-element (Zeltz et al. 1993; Drescher et al. 2002). The gain of editing sites is proposed to evolve based on so called "haphazard co-editing", which is based on the similarity of upstream sequences where the *trans*-factor can bind to (Chateigner-Boutin & Hanson 2003). This suggested model is supported by results from Tillich et al. (2005) for an editing site in the *ndhB* gene of some dicotyledonous species. The *cis*-sequence for that specific site is nearly identical to the *cis*-sequence for an editing site in *matK*. While restoring a conserved amino acid in *matK*, editing deletes a conserved amino acid in *ndhB*. This suggests that the particular *ndhB* editing site which evolved recently is rather a secondary editing target. However, editing sites gained in this way might be conserved during evolution when

they provide a selective advantage. Tillich et al. (2006) consider the chloroplast genome of a single plant cell a highly polyploid population, which is capable of repairing minor mutations. The fact that *trans*-factors seem to be encoded in the nuclear genome means that this repair mechanism can be highly efficient and act simultaneously in different chloroplast genomes. Therefore gaining new editing sites is slightly more complex than loosing them, but once established these sites have a high efficiency.

3.4.3 Future applications

The PPR protein family is not only involved in RNA editing but also in CMS. Wang et al. (2006) examined CMS of rice with Boro II cytoplasms. They found two *Rf* genes coding for PPR proteins which restored the fertility. Furthermore they detected that one of these two genes, *Rf1a*, was involved in RNA editing of the *atp6* gene. Although *Rf1a* was not an essential factor for RNA editing it had an editing promoting impact. Wang et al. (2006) hypothesized that the original function of *Rf1a* was to enhance editing and a new gained function of the gene was the restoration of fertility. However, they did not observe a direct connection between the edited or unedited *atp6* gene and CMS. Nevertheless, this observation was of high importance due to the significance of the CMS system to plant breeding and plant cultivation. As explained more thoroughly in Chapter 4 and Chapter 5, the discovery of CMS resulted in a great cost reduction for the hybrid seed production. But the utilization of CMS for hybrid seed production is not easily applicable to all species. CMS was reported for more than 140 angiosperm species (Laser & Lersten 1972), but is utilized only in four species (maize, oilseed rape, rice, beet) (Ruiz & Daniell 2005) because sterility inducing and maintaining genes as well as fertility restoring genes need to be found and stably maintained in a species (www.freepatentsonline.com/6951970.html). The Indian life science company Avesthagen developed a method to easily obtain CMS-lines from various species by genetically modifying the plant cell. Although the information on this method is rather limited, it seems to be based on the usage of edited and unedited forms of a mitochondrial inner membrane protein encoding gene (www.avesthagen.com/docs/ white_paper/ hybrid.pdf). More specified is a similar method invented by G. Brown (www.freepatentsonline.com/6951970.html) for use in oilseed rape and other crops to enhance naturally occuring CMS. Therefore a gene construct consisting of a

mitochondrial transit peptide together with an unedited form of *atp6* and *orf224* from the *Brassica napus* mitochondrial genome was introduced into the nuclear genome of a plant cell. Fertility restoration was achieved by transfer of a normal edited mitochondrial gene co-transcribed with an unusual CMS-associated mitochondrial gene (www.freepatentsonline.com/6951970.html). Both inventions apply information on RNA editing sites to enhance or introduce CMS in plants and thus facilitate future hybrid seed production. Furthermore both inventions highlight a possible natural connection between editing and CMS. Hence, gathering knowledge about editing sites and the editing mechanism is as import as analysing the mechansims behind CMS because it will lead to a better understanding of both systems. Thus it can lead to a more efficient use of CMS in various species that might not need to be based on genetical modification.

3.5 Conclusion

Thirty-one *L. perenne* chloroplast RNA editing sites were detected in 18 out of 33 analysed genes. Editing sites were in general conserved among Poaceae; however the phylogenetic analysis of *ndhA*, *ndhB* and *rpoB* showed the fast evolution of editing sites and the conflict of loosing and gaining sites. RNA editing is an important tool for predicting the functionality of genes and identifying potentially functional new ORFs (Schuster et al. 1991). Studying RNA editing in chloroplast genomes however requires the detection of sites in various species to enable understanding the mechanism and evolution behind it. Although a lot of research has been carried out in recent years to improve the availability of information on editing in chloroplast genomes, knowledge is still very restricted. The number of published chloroplast genome sequences (to date more than 130 sequences) has increased rapidly in recent years, but information on editing sites within these genomes is very limited and available only for eleven species (reviewed in Zeng et al. 2007). Information on editing sites was only available for four of the 17 published Poaceae chloroplast genome species so far. Analysing the *L. perenne* chloroplast genome for editing sites provided knowledge about editing in a fifth Poaceae species. Information is now available for one tribe of the subfamiliy Panicoideae (Andropogoneae: *Zea mays* and *Saccharum officinarum;* but two of its subtribes Tripsaceinae and Saccharinae respectively), one tribe of the subfamily

Ehrhartoideae (Oryzeae: *Oryza sativa*) and two tribes of the subfamily Pooideae (Triticeae: *Hordeum vulgare*; Poeae: *Lolium perenne*). Recently a new approach was developed for the detection and quantification of RNA editing sites based on the high resolution melting analysis of amplicons which is a rapid, high-throughput method (Chateigner-boutin & Small 2007). This approach is six times cheaper than previous methods and much faster. The application of this new approach to the chloroplast genome of many different species will hopefully result in an increased amount of knowledge of editing sites for various species in the near future and help us further understand the editing mechanism and its evolution.

Chapter 4

Plastid genome diversity within and among *Lolium perenne* cultivars and populations

"Over the past century, the development and successful application of plant breeding methodologies has produced the high-yielding crop varieties on which modern agriculture is based. Yet, ironically, it is the plant-breeding process itself that threatens the genetic base on which breeding depends."
Tanksley & McCouch (1997)

4.1 Introduction

4.1.1 The need for crop genetic diversity

In 1970 the U.S. corn belt (~ Western Indiana, Illinois, Iowa, Missouri, Eastern Nebraska, Eastern Kansas) experienced tropical weather with high temperatures and persistent rain. These were ideal growth, reproduction and spreading conditions for the fungus *Helminthosporium maydis* which causes the disease 'Southern corn leaf blight' in maize plants. As a result of this disease, the U.S. corn crop market suffered dramatic yield losses of up to 70% (~710 million bushels) in 1970. Field losses of 50% and more were common. However, approximately 15% of the fields in heavily diseased areas remained nearly disease free (Adams et al. 1971).

One year earlier it had been detected that corn plants containing the Texas male sterile cytoplasm were very susceptible to the fungus, but plants with normal cytoplasm were resistant. The susceptibility of the Texas male sterile plants was due to a host specific toxin which was produced by a new race of *H. maydis* (Adams et al. 1971, Martinson 2007). Male sterile lines are desirable in maize breeding programmes which are mainly based on the production of hybrid seed (and dependent on heterosis). For the production of hybrid seed, maize plants need to be de-tasselled to avoid self-pollination. This is, however, a very time and labour consuming process which can be omitted by the use of

male sterile lines. Texas male sterility was discovered in one maize plant in 1944 (Martinson 2007). Eighty-five percent of the U.S. corn plants grown in 1970 could be maternally traced back to that one plant, despite the fact that around 30 other male sterility inducing cytoplasms were known (Tatum 1971). The corn blight serves as a powerful example of the dangers of over-reliance on a narrow gene pool for cultivar development and use.

Unfortunately, this is not the only example of a species, with such economically high importance as maize, whose cultivars were developed from a very narrow genetic base. As reviewed in Tanksley & McCouch (1997) the majority of hard red winter wheat varieties in the United States derive from only two lines that were imported from Poland and Russia.

Narrowing of the genetic base of crops has been recorded in a wide range of crops using a range of molecular and morphological markers. For example, Provan et al. (1999) analysed 101 *Hordeum vulgare* ssp. *vulgare* cultivars and 51 accessions of *Hordeum vulgare* ssp. *spontaneum*, the wild progenitor of *Hordeum vulgare* ssp. *vulgare*. They found a total of eleven different plastid DNA haplotypes in the 51 accessions of the wild progenitor but only one haplotype within the 101 cultivars. Grau Nersting et al. (2006) analysed six different groups of Nordic oat (*Avena sativa*) using morphological and molecular approaches. One group contained 17 accessions of different oat landraces, the remaining groups represented cultivars from 20 years of breeding beginning in 1898 until today. They could clearly observe the breeding progress due to the better performance of more recent developed cultivars and the adaption of cultivars to the Norwegian climate conditions. However, they also observed a decrease in diversity when comparing the landraces with the different cultivar groups. While 46 different nuclear alleles were found within the landraces, the modern cultivars had lost up to 24 of them (Grau Nersting et al. 2006).

The domestication of crops from wild species started around 10,000 years ago (Tanksley & McCouch 1997; Gepts 2002). It is not exactly known in which form it happened, but it is likely that a selection for specific traits (known as the 'domestication syndrome traits'; Koinange et al. 1996) was started and thus the genetic base for future crops began to narrow. This bottleneck got even narrower with the introduction of

efficient modern plant breeding systems (Tanksley & McCouch 1997). Epidemics such as the Southern corn leaf blight disease in 1970 (Adams et al. 1971, Martinson 2007) or the potato late blight disease in 1845 in Ireland, which caused severe famine might have been avoidable if cultivars with a more diverse genomic background had been used. These examples illustrate the importance of monitoring the diversity in plant breeding material and broadening the genetic base of breeding schemes by introducing new wild germplasm (Tanksley & McCouch 1997).

Monitoring diversity in plant breeding material can be easily done using molecular techniques such as Restriction Fragment Length Polymorphism (RFLP), Random Amplified Polymorphic DNA (RAPD), Amplified Fragment Length Polymorphism (AFLP), Simple Sequence Repeat (SSR, microsatellite) and Single Nucleotide Polymorphism (SNP) markers. These approaches were first applied to the nuclear genome but have also been used for analysing chloroplast genomes. Studies involving the mitochondrial genome have also been undertaken but the mitochondrial genome has generally been less suitable due to its very low mutation rate and high level of intramolecular recombination (Provan et al. 2001).

Chloroplast genomes are generally known as being highly conserved. Thus it was very surprising when Powell et al. (1995a) reported a high amount of SSRs within the chloroplast genome that showed variation between individuals of the same cultivar. Until that report, SSRs had mainly been observed in the nuclear genome where they occurred very frequently in form of di-, tri- and tetra-nucleotide repeats with up to 1,000 copies (reviewed in Cuadrado et al. 2008). Nuclear microsatellites (ncSSR) proved to be very variable and thus were used to characterize species. In contrast to ncSSRs, chloroplast microsatellites (cpSSRs) are based mainly on mononucleotide repeats of not more than 20 bp in length. CpSSRs can be found throughout the chloroplast genome but occur mainly in the large single copy region. Only very few cpSSRs can be found in the inverted repeat regions (Powell et al. 1995b). The amount of cpSSRs found in different species varies widely, as well as their position in the different species. However, cpSSRs found within the inverted repeat seem to be conserved across species (Powell et al. 1995b). Despite the conserved nature of chloroplast genomes, many microsatellite regions analysed show a high degree of variation, are genetically well defined, and are easily assayed by PCR. Hence the use of chloroplast microsatellites can result in a

higher resolution of diversity than the use of chloroplast RFLP and RAPD markers (Echt et al. 1998).

4.1.2 Chloroplast microsatellite markers

Chloroplast microsatellite regions are highly susceptible to mutation due to slipped strand mispairing during replication (reviewed in Kelchner 2000) which results in small insertion/deletion events (Powell et al. 1995a). In general it has been observed that the likelihood of mutation increases with an elongation of the cpSSR. However, imperfect cpSSRs, those interrupted by another nucleotide, show in general a lower mutation rate (Jakobsson et al. 2007) which is likely due to a stabilizing effect of the interrupting nucleotide (Rolfsmeier & Lahue 2000). Nevertheless, chloroplast microsatellite regions seem to be suitable as markers (Powell et al. 1995a). With the design of universal primers (primers amplifying across a very wide range of species) (Taberlet et al. 1991, Dumolin-Lapegue et al. 1997), the possibility arose to analyse microsatellite regions in a wide range of species. Interestingly available primers mainly amplify markers in intergenic spacer regions of transfer RNAs (Provan et al. 2001). This is mainly due to the fact that the degree of conservation is high in the transfer RNA genes (for primer development) and the length of the intergenic spacers sufficient to accumulate enough polymorphism for use as molecular markers at infraspecific levels (Taberlet et al. 1991). Universal primers will in general not resolve all the detectable variation in a species because they were designed in highly conserved regions, hence the region they amplify will have a higher degree of conservation (Provan et al. 2001). Thus, several authors have developed species or taxon specific primers for chloroplast microsatellite markers (Flannery et al. 2006) which are able to amplify regions of greater variation within those taxa.

4.1.3 Application of chloroplast microsatellite markers

To date only two nuclear plant genomes (*Oryza sativa, Arabidopsis thaliana*) are completely sequenced, annotated and publicly available (http://www.ncbi.nlm. nih.gov).

Although the sequencing and annotation of six further nuclear plant genomes (*Glycine max, Medicago trunculata, Populus trichocarpa, Sorgum bicolor, Vitis vinifera, Zea mays*) is close to completion and sequence data are partially already available, this number is nevertheless much lower than the 100 sequenced and publicly available chloroplast genomes. This fact clearly illustrates one advantage of chloroplast microsatellite markers as supporting tools for plant breeding schemes. Chloroplast microsatellite regions can be easily identified by using computer programmes to scan DNA sequence files or even identified by screening chloroplast DNA sequences by eye. Primers can be designed based on the sequences of closely related species if sequence information from the target species is missing, due to the high conservation of chloroplast DNA sequences (Olmstead & Palmer 1994). Chloroplast genomes are known to be non-recombinant, thus different microsatellite loci are linked with each other and individual haplotypes can be easily detected by using a set of different chloroplast microsatellite markers (Bryan et al. 1999a). The application of cpSSRs for phylogenetic analysis is often most effective in combination with nuclear markers (e.g. internal transcribed spacer (ITS) sequences) because the nuclear and plastid genomes can often have different phylogenetic histories (Wendel et al. 1995) and due to the risk of size homoplasy (alleles are identical by state but not by descent (Jarne & Lagoda 1996))(Hale et al. 2004) that can lead to wrong conclusions. Furthermore microsatellite rich regions complicate the application of otherwise suitable non-coding regions for barcoding purposes (CBOL Plant Working Group 2009). However, there are many other significant applications for chloroplast microsatellites. cpSSRs enable the monitoring of pollen flow in gymnosperms where the chloroplast genome is paternally inherited (Powell et al. 1995a), the seed mediated gene flow in angiosperms where the chloroplast genome is generally maternally inherited (Corriveau & Coleman 1988) and thus facilitate the monitoring of gene flow in general. This is of great importance, for example, for assessing the risk of transferring transgenes into wild populations or creating wild genetically modified populations (Ryan et al. 2006). Chloroplast markers can be used for detecting the parentage in interspecific hybrids (Atienza et al. 2007), somatic hybrids (Bastia et al. 2001, Bryan et al. 1999b) and intraspecific hybrids (Hodkinson et al. 2002a, Akkak et al. 2007). Powell et al. (1995b) reported that intraspecific chloroplast variation is not random with regard to geographical localization, thus chloroplast microsatellites are very useful tools for detecting the phylogeographic population structure of species (Balfourier et al. 2000, McGrath et al.

2007). Furthermore chloroplast microsatellites are well suited to detect population genetic bottlenecks in natural populations and to assess the cytoplasmic diversity and genetic variation that exists in plant breeding material (Provan et al. 2001, Grau Nersting et al. 2006).

4.1.4 Genetic diversity and *Lolium perenne*

Maintaining the genetic diversity in *L. perenne* and related species is as important as maintaining the diversity in other species such as maize or potatoes. However, generally the situation is a little bit different from that in maize and potatoes due to the use of different breeding systems. Maize is an outcrossing self-compatible monoecious species with male and female flowers separated from each other. Therefore new cultivars are created by use of the hybrid breeding system. This system reduces the genetic diversity because it is based on inbred lines. Inbred lines can be easily generated due to the self-compatibility of maize. Controlled crosses of maize plants can be easily carried out by means of mechanical or chemical emasculation and the use of CMS lines. Potato breeding is based on clone breeding. After an initial crossing the seedlings are maintained and propagated vegetatively. Hence a potato cultivar consists of one fixed genotype. Thus the risk of reducing genetic diversity when breeding maize and potato is extremly high.

The situation when breeding *L. perenne* and closely related species such as *L. multiflorum* or even *Festuca pratensis* is slightly different. *Lolium perenne* and its relatives are like maize monoecious but male and female flowers are combined making a mechanical emasculation complicated. Furthermore are all three species highly self-incompatible and thus the creation of inbred lines is very difficult. All three species can theoretically be propagated vegetatively but the one objective in breeding *L. perenne* and its relatives is the seed production. Therefore *L. perenne* is mainly bred via population breeding and creation of synthetic cultivars (Wilkins et al. 1991, Guthridge et al. 2001). A synthetic cultivar is based on several selected paternal plants with desired traits which are crossed via open pollination with each other, afterwards propagated via seed and then form the new cultivar (Wilkins et al. 1991, Guthridge et al. 2001). Thus each plant of an *L. perenne* cultivar can be a different genotype. Hence

the risk of reducing the diversity within *L. perenne* is not as severe as for example for maize or potato. Nevertheless, it is important to monitor the diversity within *L. perenne* and to introduce new genetic resources into breeding populations to avoid the reduction of genetic diversity. Introducing new genetic resources into the breeding population can be done by use of existing cultivars from other countries and continents, by use of gene bank material (gene banks comprise collections of various species from many different climatic regions from all over the world) and by use of own collection material from the same country. *Lolium perenne* can naturally be crossed with some other *Lolium* species or even *Festuca* species, which can increase the genetic diversity. This happened for example when introducing CMS to *L. perenne*. Although the use of CMS generally bares the risk of diversity reduction its initial introduction into *L. perenne* by a cross with *F. pratensis* (www.teagasc.ie/research/reports/ crops/3495/eopr-3495.asp) resulted in an increase in genetic diversity.

4.1.5 Aims and objectives

Breeding can reduce genetic diversity as was shown for example for *Avena sativa* (oat) (Grau Nersting et al. 2006). This reduction of genetic diversity has contributed to recurrent plant disease epidemics with severe economical and fatal losses, because new varieties are created from a very narrow genetic base. Therefore a major goal of modern plant breeding schemes needs to be the monitoring and increase of the genetic diversity within the plant breeding material. Monitoring genetic diversity can be easily done by using cpSSR markers. Primers for universal chloroplast microsatellite markers exist (Taberlet et al. 1991, Dumolin-Lapegue et al. 1997) but also more family specific primers became available in recent years (Flannery et al. 2006, Bryan et al. 1999b). Five Poaceae specific primers had been designed by Provan et al. (2004) for the *trnK* intron, the *psbK-psbI* intergenic spacer, *rpoC2-rps2* intergenic spacer, the *atpI-atpH* intergenic spacer and the *atpB-rbcL* intergenic spacer. Twelve additional primer sets for the Poaceae had been published by McGrath et al. (2006). Five of these additional primers represented an optimization of previously existing sets while seven sets were completely new. These new primer sets amplified chloroplast microsatellite markers in the following regions and genes: *23S-5S* internal transcribed spacer, *psbA*, *trnV*, *rpl2* – *trnH* intergenic spacer, *rpoA*, *rpoC2* and *ndhF*.

One objective of this thesis was to sequence and annotate the chloroplast genome of *Lolium perenne*. During this work, a high amount of variation was revealed within the chloroplast genome of a single cultivar of *L. perenne* (Diekmann et al. 2009, Chapter 2). Chloroplast genomes are known as being highly conserved in comparison to nuclear and mitochondrial genomes; thus this observation of high variation was unexpected particularly in a cultivar. Variation was found in form of indels in cpSSR regions, SNPs and inversions. Therefore this chapter, generally aimed to test if regions with observed variation in the chloroplast genome of *L. perenne* can be of potential value for providing new diverse chloroplast microsatellite markers that are able to detect genetic diversity within *L. perenne* cultivars and populations. More specifically the objectives are 1) to evaluate the amount of mononucleotide repeats within Poaceae and detect the regions where they most frequently accumulate by using the *L. perenne* chloroplast genome as an example, 2) to combine the information on variable regions from Chapter 2 with the information of the position of long mononucleotide repeats to design new Poaceae universal cpSSR primers, 3) to test these new primers on a small set of different *L. perenne* ecotypes (Irish and European) and cultivars, as well as different *Lolium* species, *Festuca* species and other grass species for their potential value as new chloroplast microsatellite markers, and to quantify the amount of diversity they are able to detect in our genetic resource collection of *Lolium*.

4.2 Material and Methods

4.2.1 Characterization of Poaceae chloroplast microsatellites

Thirteen Poaceae chloroplast genomes (*Agrostis stolonifera* (EF115543), *Bambusa oldhamii* (FJ970915)[1], *Brachypodium distachyon* (EU325680), *Dendrocalamus latiflorus* (FJ970916)[1], *Festuca arundinacea* (FJ466687), *Hordeum vulgare* (EF115541), *Lolium perenne* (this study, AM777385), *Oryza nivara* (AP006728), *Oryza sativa* 'japonica group' (X15901), *Oryza sativa* 'indica group'(AY522329), *Saccharum officinarum* (AP006714), *Sorghum bicolor* (EF115542), *Triticum aestivum* (AB042240), *Zea mays* (X86563); (Genbank accession numbers,

[1] *kindly released to us by Choun-Sea Lin prior to its publication*

http://www.ncbi.nlm.nih.gov/)) were searched for chloroplast microsatellites using the microsatellite finder tool find_microsat_Win32 (Salamin unpublished). The search focussed on mononucleotide repeats of seven and more nucleotides. The microsatellites had to be interrupted from each other by at least one base. To determine the microsatellite rich regions in the chloroplast genome the results for *L. perenne* were investigated more thoroughly by comparing the microsatellite positions given by find_microsat_Win32 (Salamin unpublished) with the *L. perenne* chloroplast genome annotation.

4.2.2 Plant material

Old permanent grasslands often contain a great reservoir of genetic diversity in comparison to highly managed, fertilized and reseeded grasslands. To preserve this diversity, *ex situ*, an extensive programme of sampling of old pasture ecosystems was carried out from 1979 to 1983 by Teagasc. Either at least 50 vegetative tillers were collected in spring or at least 80 seed heads collected in the autumn from different grassland species including *L. perenne*. Thus 534 sites for *L. perenne* were sampled. These 534 ecotypes were then propagated under isolation. In 1994 the European *Lolium* core collection programme was started and 163 different accessions from different gene banks were included in this programme to assess the genetic diversity within this species (Connolly 2000). Sixteen Irish *L. perenne* ecotypes, 15 European *L. perenne* ecotypes, nine *L. perenne* cultivars as well as six different *Lolium* species and 14 other grass species were used for this study and they derived mainly from these two collections (Table 7). The number of individuals analysed per species varied from one to 16. Total DNA from these individuals was extracted during an earlier project by McGrath (2008) and is described thoroughly in their publications (McGrath et al. 2006, 2007).

4.2.3 Primer design

For amplifying chloroplast microsatellite regions across different Poaceae species, primers were designed based on the following criteria: a) only mononucleotide repeats

of more than 10 bp length were considered due to the higher likelihood in observing length variations, b) only regions that have not been a focus of a chloroplast microsatellite study in grasses before were considered, c) preferably those microsatellite regions with an observed high variation (Chapter 2) were taken into account, d) primers should bind also to non-*Lolium perenne* chloroplast DNA to enable studies across different grass species (universality in the grass family). Based on this chloroplast microsatellite study, 14 regions consisting of 17 microsatellites of more than 10 bp in length (including incomplete repeats) were found (Table 8). Two of these regions had been subjects of earlier studies (Provan et al. 2004, McGrath et al. 2006) and thus were

Table 7 Names of accessions, their origin and the numbers of individuals used in the *Lolium perenne* diversity study as well as the haplotypes found within them using Arlequin 3.11. Unique = unique haplotype among all accessions analysed.

species	accession number	individuals	country of origin	county	Seed Source	haplotypes						
						total	unique	name				
***Lolium perenne* - Irish ecotypes**												
Lolium perenne	IRL-OP-02007	8	Ireland	Cork	Teagasc Oak Park	2		1	2			
Lolium perenne	IRL-OP-02018	2	Ireland	Wicklow	Teagasc Oak Park	-	-	-	-	-	-	-
Lolium perenne	IRL-OP-02059	8	Ireland	Clare	Teagasc Oak Park	3		1	2	3		
Lolium perenne	IRL-OP-02078	12	Ireland	Galway	Teagasc Oak Park	5		1	2	3	4	50
Lolium perenne	IRL-OP-02128	15	Ireland	Kerry	Teagasc Oak Park	4	2	1	2	7	8	
Lolium perenne	IRL-OP-02173	15	Ireland	Waterford	Teagasc Oak Park	3		1	2	9		
Lolium perenne	IRL-OP-02192	7	Ireland	Cork	Teagasc Oak Park	1		1				
Lolium perenne	IRL-OP-02230	2	Ireland	Galway	Teagasc Oak Park	2		1	2			
Lolium perenne	IRL-OP-02267	7	Ireland	Tipperary	Teagasc Oak Park	3	1	1	10	11		
Lolium perenne	IRL-OP-02269	8	Ireland	Tipperary	Teagasc Oak Park	3	1	1	2	12		
Lolium perenne	IRL-OP-02274	2	Ireland	Limerick	Teagasc Oak Park	1		2				
Lolium perenne	IRL-OP-02312	9	Ireland	Cork	Teagasc Oak Park	3		1	2	13		
Lolium perenne	IRL-OP-02337	3	Ireland	Carlow	Teagasc Oak Park	3		1	2	14		
Lolium perenne	IRL-OP-02419	8	Ireland	Roscommon	Teagasc Oak Park	3		1	2	3		
Lolium perenne	IRL-OP-02442	6	Ireland	Mayo	Teagasc Oak Park	4		1	2	3	4	
Lolium perenne	IRL-OP-02491	8	Ireland	Wexford	Teagasc Oak Park	3		1	2	14		
***Lolium perenne* - European ecotypes**												
Lolium perenne	21806	6	Ukraine		N/A	3		1	4	10		
Lolium perenne	16-7-62-2Nordic	8	Norway		Teagasc Oak Park	3		1	2	13		
Lolium perenne	3013 Romania	2	Romania		Teagasc Oak Park	1		2				
Lolium perenne	3199 Romania	5	Romania		Teagasc Oak Park	2		1	2			
Lolium perenne	ABY-Ba 11478	7	Greece		IGER	2		1	4			
Lolium perenne	ABY-Ba 12896	6	Denmark		IGER	2		1	2			
Lolium perenne	IV-51-161 Hungary	9	Hungary		Teagasc Oak Park	2		1	2			
Lolium perenne	NGB 14250	4	Sweden		Nordic Gene Bank	1		1				
Lolium perenne	PI 267059	8	Poland		GRIN	3	1	3	9	15		
Lolium perenne	PI 321397	7	Czech Republic		GRIN	2		1	2			
Lolium perenne	PI 547390	1	Iran		GRIN	-	-	-	-	-	-	-
Lolium perenne	PI 598445	8	Netherlands		GRIN	5		1	2	4	13	60
Lolium perenne	W6 11325	8	Turkey		GRIN	4		1	2	4	14	
Lolium perenne	W6 9286	8	France		GRIN	1		2				
Lolium perenne	W6 9339	8	Wales		GRIN	4		1	2	4	10	

Table 7 *continued*

species	accession number	individuals	country of origin	county	Seed Source	haplotypes						
						total	unique	name				
Lolium perenne – cultivars												
Lolium perenne	cv. Aurora	8	N/A		IBERS	2		1	2			
Lolium perenne	cv. Barlenna	5	N/A		Barenbrug Holland BV	3		1	2	10		
Lolium perenne	cv. Cashel	5	N/A		Teagasc Oak Park	3		1	2	3		
Lolium perenne	cv. Greengold	8	N/A		Teagasc Oak Park	2		1	2			
Lolium perenne	cv. Magician	2	N/A		Teagasc Oak Park	1		1				
Lolium perenne	cv. Manhattan	8	N/A		Alan Stewart NZ	3		1	3	14		
Lolium perenne	cv. Navan	2	N/A		DARDNI	1		2				
Lolium perenne	cv. Odenwaelder	2	N/A		IPK Gatersleben	1		1				
Lolium perenne	cv. Portstewart	6	N/A		DARDNI	2	1	2	17			
Lolium sp.												
Lolium hybridum	GR 11849/49	5	N/A		IPK Gatersleben	3	1	3	25	60		
Lolium multiflorum	cv. Multimo	8	N/A		N/A	3		2	21	22		
Lolium multiflorum	cv. Nivack	7	N/A		N/A	3		2	21	22		
Lolium persicum	PI 229764	6	Iran		GRIN	3	1	5	27	50		
Lolium remotum	GR 11839/99a	3	Germany		IPK Gatersleben	1	1	2				
Lolium subulatum	PI 197310	7	Argentina		GRIN	4		5	9	60	50	
Lolium temulentum	ABY-Ba 8917	3	Iran		IGER	1		9				
Lolium temulentum	ABY-Ba 13643	7	Morocco		IGER	2	2	1	19			
other species												
Agrostis stolonifera	PI 439027	8	Uzbekhistan		GRIN	-	-	-	-	-	-	-
Avena sativa	cv. Evita	1	N/A		Lochow-Petkus	-	-	-	-	-	-	-
Bromus erectus	PI 619490	4	Hungary		GRIN	-	-	-	-	-	-	-
Cynosurus cristatus	PI 509441	2	Romania		GRIN	-	-	-	-	-	-	-
Festuca	cv. Dovey	1	N/A		Barenbrug Holland BV	-	-	-	-	-	-	-
Festuca ovina	PI 634304	6	China		GRIN	5	4	2	33	34	35	36
Festuca pratensis	cv. Wendelmold	1	N/A		N/A	-	-	-	-	-	-	-
Festuca pratensis	cv. Northland	8	N/A		PGG-Wrightson	1	1	3				
Festuca rubra	IRL-OP-02174	5	Ireland		Teagasc Oak Park	4	4	3	38	39	40	
Festuca vivipara	PI 251118	8	Yugoslavia		GRIN	4	3	2	29	30	31	
Hordeum vulgare	cv. Regina	1	N/A		Cebeco Zaden BV	-	-	-	-	-	-	-
Poa pratensis	PI 539060	4	Siberia		GRIN	-	-	-	-	-	-	-
Secale cereale	cv. Protector	1	N/A		Cebeco Zaden BV	-	-	-	-	-	-	-
Triticum aestivum	cv. Robicum	1	N/A		CPB Twyford UK	-	-	-	-	-	-	-
xTriticosecale	cv. Lupus	1	N/A		Nordsaat Saatzucht GmbH	-	-	-	-	-	-	-

of no interest for this study. Primer sets were finally designed for nine regions amplifying twelve microsatellites (Table 9) using the Primer3 software (http://frodo.wi.mit.edu/). Six primer sets amplified repeats in non-coding regions and three sets amplified repeats within genes. However, only one primer set amplified exclusively a gene region. The amplicon lengths varied from 195 bp to 658 bp depending on the primer set used. The amplicons were, thus, considerably longer than the microsatellite repeat of interest. This provided potential for recording other polymorphisms outside the SSR repeat and would allow them to be more broadly applied (for example for phylogenetic studies).

4.2.4 Chloroplast microsatellite region amplification

The DNA templates were divided up into four 96-well master plates. Each plate included a negative control. For each primer set and 96 samples a PCR mastermix for 100 reactions was set up consisting of the following components: 600 µl 5x Phusion™ HF Buffer (New England Biolabs, Inc., Ipswich, MA, USA), 60 µl FP (10 µM), 60 µl RP (10 µM), 60 µl dNTPs (metabion international AG, Martinsried, Germany) (10 mM), 1896 µl ddH_2O, 24 µl Phusion™ Hot Start High-Fidelity DNA Polymerase (New England Biolabs, Inc.). To each well, 3 µl DNA template and 27 µl of PCR master mix were added. The PCR program settings were 98 °C 5 min, (98 °C 1 min, 60 °C 1 min, 72 °C 1 min) 35 cycles, 72 °C 10 min. 3 µl of each PCR product were checked for amplification using 2.5% MetaPhor® Agarose (Lonza, Rockland, ME, USA) gels (Figure 20). Amplified PCR products were sequenced once using forward primers. Sequencing was outsourced to a company (AGOWA GmbH, Berlin, Germany or GATC Biotech AG, Konstanz, Germany).

Originally 10 x Thermo Buffer (New England Biolabs, Inc., Ipswich, MA, USA) and *Taq*-Polymerase (New England Biolabs, Inc.) were used for amplification but the sequencing results of primer sets TeaCpSSR34 and TeaCpSSR35 were not sufficiently good enough because the sequencing reactions stopped within the microsatellite motif. By recommendation of the sequencing company (AGOWA GmbH, Berlin, Germany) Phusion™ Hot Start High-Fidelity DNA Polymerase (New England Biolabs, Inc.) with proof reading ability was used. However, sequences of PCR products achieved with the

Phusion polymerase were also unsuccessful when sent to the same sequencing company (AGOWA GmbH, Berlin, Germany) but were good when sequenced by another company (GATC Biotech AG, Konstanz, Germany). This implies that problems observed were rather due to technical problems at the sequencing company than to issues with the polymerase.

Table 8 Chloroplast microsatellite regions (>10 nucleotides) found in the *Lolium perenne* chloroplast genome.

chloroplast microsatellite	position		region	analysed in
	start	end		
$(A)_{16}$	78484	78499	*rpl16* intron	-
$(C)_{15}$	47279	47293	*trnL-trnF*	McGrath et al. 2006
$(C)_{11}$	14120	14130	*trnG-trnT*	-
$(T)^{11}$	80296	80306	*rps19-trnH*	this study
$(A)_7g(A)_7$	33357	33371	*atpF* intron	this study
$(A)_{10}$	3593	3602	*matK-trnK*	Provan et al. 2004
$(T)_{11}$	62836	62846	*psbE-petL*	this study
$(T)_{10}$	63032	63041	*psbE-petL*	this study
$(A)_{10}$	63108	63117	*psbE-petL*	this study
$(T)_{10}$	76076	76085	*infA*	this study
$(T)_{10}$	76094	76103	*infA*	this study
$(A)_7c(A)_8$	36489	36504	*rps14-psaB*	this study
$(T)_{12}$	49337	49348	*ndhK*	this study
$(T)_{10}$	58739	58748	*cemA*	this study
$(T)_{11}$	76655	76665	*rps8-rpl14*	this study
$(C)_{10}$	93162	93171	*rrn16-trnI*	this study
$(A)_8c(A)_{13}$	43483	43504	*ycf3* intron	-

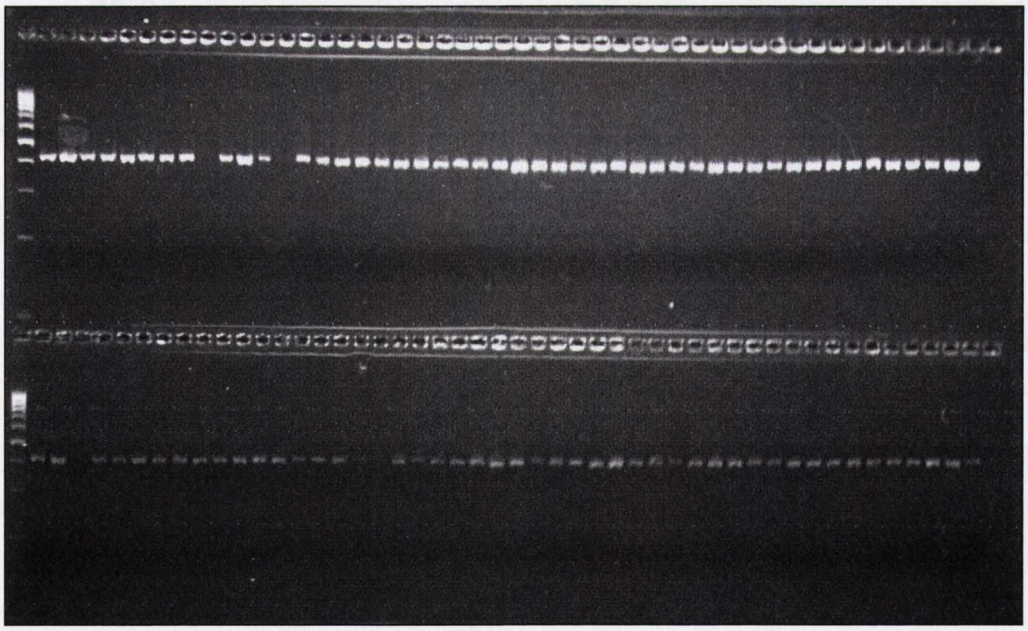

Figure 20 Example of amplified products on a 2.5% MetaPhor® Agarose gel (Lonza, Rockland, ME, USA) stained with 1% ethidium bromide (10 mg/ml). Size standard: 3 µl of mi-100 bp+ DNA Marker Go (metabion international AG, Martinsried, Germany), run for ~ 2 hours at 130 V. Primer set used: TeaCpSSR28, templates: different *Lolium perenne* ecotypes.

Table 9 Poaceae universal chloroplast microsatellite (cpSSR) primers. CpSSR position is given in relation to the chloroplast genome of *L. perenne*.

primer	Sequence FP = forward primer RP = reverse primer	amplicon length	genome region (Figure 23)	cpSSR	cpSSR position	amplified genome region (Figure 23)
TeaCpSSR27	FP: AATGCCGAATCGACGACCTA RP: CAATGGTCCCTCTACGCAAT	390	*atpF* intron	$(A)_7g(A)_7$	33357	*atpF* intron
TeaCpSSR28	FP: TGCAATTTTTCTCGCATTTTC RP: TTTCCATTGTGCAAGCAAGA	419	*rps14-psaB*	$(A)_7c(A)_8$	36489	*rps14-psaB*
TeaCpSSR29	FP: GGTACCAATCCATAACGATC RP: GCGCTAGTTTTTGTTGTTTT	415	*ndhK*	$(T)_{12}$	49337	*ndhK-ndhC*
TeaCpSSR30	FP: GGATTAAGAATTGGTGGAATACC RP: AAATGACTACAAGCGAAGGAGA	570	*cemA*	$(T)_{10}$	58739	*cemA*
TeaCpSSR31	FP: GGTCGTGGAATGCTTTTCTT RP: TCCACGAATCTCAATGACCA	658	*psbE-petL*	$(T)_{11}$	62836	*psbE-petL*
				$(T)_{10}$	63032	
				$(A)_{10}$	63108	
TeaCpSSR32	FP: ACGTCCCTTGCTTGAATCAT RP: TCGAGGGATAATGACAGATCG	386	*infA*	$(T)_{10}$	76076	*infA- rps8*
				$(T)_{10}$	76094	
TeaCpSSR33	FP: TGTCGCAAAGTTGAAACCAA RP: AATCCACTGCCTTGATCCAC	482	*rps8-rpl14*	$(T)_{11}$	76655	*rps8-rpl14*
TeaCpSSR34	FP: CCCAATTTGCGACCTACCAT RP: AATCCACTGCCTTGATCCAC	382	*rps19-trnH*	$(T)_{11}$	80296	*rps19-trnH*
TeaCpSSR35	FP: GGAGGCCTGCTACGCC RP: TGGAAGTCTTCTTTCGTTTAGGGT	195	*rrn16- trnI*	$(C)_{10}$	93162	*rrn16- trnI*

4.2.5 Phylogenetic analysis

Sequences from each of the different loci were first aligned in MEGA 3.1 (Kumar et al. 2004) and checked manually for indels, SNPs and length variation in microsatellite regions. Observed variations were always confirmed by comparisons with the original sequence trace files. Indels and gaps were coded for by using other nucleotides than the ones already existing in the individuals that did not show indels or gaps. This step was necessary especially for the indels. Leaving them unchanged each additional nucleotide would otherwise stand for an evolutionary event of gain or loss, while it is more likely that the complete indel was lost/gained at one time. Individuals for whom amplification or sequencing failed were included in the analysis coding each absent nucleotide with a question mark. Sequences from the different primer sets were joined in MEGA 3.1 (Kumar et al. 2004) to one file. The following adjustments had to be made: Sequences derived from DNA template IRL-OP-02018 were completely omitted from the analysis; data from single individuals from ABYBa-12896 (one), IRL-OP-02078 (one) and IRL-OP-02269 (three) had also to be removed from the analysis. These adjustments were required because of poor amplification or failed sequencing results. The phylogenetic analysis was performed in PAUP* 4.0 (Swofford 2002) using UPGMA (Unweighted

Pair Group Method with Arithmetic mean/ using the *p*-distance method) and the individual sequences from the alignment. The *p* distance is the simple proportion of nucleotide differences in an alignment. Many methods that estimate also the total number of substitutions between any pair of sequences exist, however they mostly have a very large variance. Thus simple measures such as the *p* distance produce very often the correct tree topology when the substitution rate is constant and the number of analysed nucleotides small (Nei 1996). Here, the combination of UPGMA and *p* distance was simply used to provide a first overview over the sequencing data. Due to the large input file, the phylogenetic tree from this approach was however, too big, for further studies. Therefore the different haplotypes of the data set were distinguished using the software Arlequin 3.11 (Excoffier et al. 2005). Arlequin 3.11 counts loci with missing data as extra haplotypes thus a manual search based on the Arlequin 3.11 results had also to be carried out to prevent overestimating the number of haplotypes. That search was conducted in MEGA 3.1 (Kumar et al. 2004). The resulting haplotype sequences were used as input file for a Bayesian inference approach using MrBayes (settings: **nst** = 6 and **rates** = invgamma (= GTR+G+I-model (**G**eneral **T**ime **R**eversible model + **G**amma-distributed rate); **ngen** (generations) = 1,000,000; **samplefreq** (samplefrequency) = 100; **burnin** = 1000) and a Maximum parsimony approach combined with bootstrapping using PAUP* 4.0 (Swofford 2002) (settings: **Search** = Heuristic; **NRep** (replicates) = 1000; **/RearrLimit** = (rearrangements limited to) 1,000,000). The tree from this approach was modified in FigTree v1.2.1 2006-2009 (http://tree.bio.ed.ac.uk/) for figure production.

4.2.6 Locus-by-locus AMOVA and genetic distance analysis

The locus-by-locus AMOVA was based on the complete DNA sequences to include polymorphic information outside the microsatellite regions. As before IRL-OP-02018 was completely deleted and some individuals from ABYBa-12896 (one), IRL-OP-02078 (one) and IRL-OP-02269 (three) were removed from the analysis because of incomplete data. Data of populations which were represented by only one individual were also removed from the analysis. Thus a total of 54 populations were included and they were divided into six groups: 1) *Lolium perenne* – Irish ecotypes, 2) *Lolium perenne* – European ecotypes, 3) *Lolium perenne* – cultivars, 4) *Lolium* species, 5)

Festuca species and 6) other grass species. The analysis was performed for all groups and then stepwise reduced (Table 11). The input-files had the following header: **NbSamples=54** (depending on the amount of samples (= populations) used), **DataType=DNA**, **GenotypicData=0** (= haploid data), **GameticPhase=1** (=known), **LocusSeparator=NONE** (= each nucleotide will be counted as locus), **RecessiveData=0** (= co-dominant data), **MissingData='?'**. For the analysis the value for the allowed missing data was set to 0.26 so that all data were included. Pairwise differences were computed with 10,000 permutations at three different significance levels ($P<0.5$, $P<0.1$ and $P<0.01$). The same analysis was additionally carried out using only groups 1), 2) and 3) and sequence data from primer sets TeaCpSSR31 and TeaCpSSR27 because these sets had very interesting results in the chloroplast microsatellite study. In both analyses carried out negative values were found. These values were replaced by the value 0 according to Nei (1978).

4.3 Results

4.3.1 Characterization of Poaceae chloroplast microsatellites

The chloroplast genomes of thirteen different grass species were searched for chloroplast microsatellite regions. Although this search was limited to mononucleotide repeats with at least seven nucleotides, the number found was high with 241 (*Dendrocalamus latiflorus*) to 282 (*Brachypodium distachyon*) microsatellites per genome. Interestingly the number of microsatellites found in the Pooideae and Panicoideae subfamilies was clearly higher than the number found in the Ehrhartoideae and Bambusoideae (Figure 21). In all genomes analysed the number of microsatellites found decreases with an increase of the number of repeats. Thus the highest amount of microsatellites can be found for repeats of seven nucleotides and the lowest for repeats of more than eleven nucleotides (Figure 21).

Chloroplast microsatellites in the Pooideae are mainly based on the nucleotide A while for the other subfamilies they are relatively equally based on the nucleotides A and T. Although the amount of published chloroplast genome sequences for the Bambusoideae is limited to date (to only two), it seems that in this subfamily the amount of

microsatellites that are based on the nucleotide C is higher than in the other species (Figure 22).

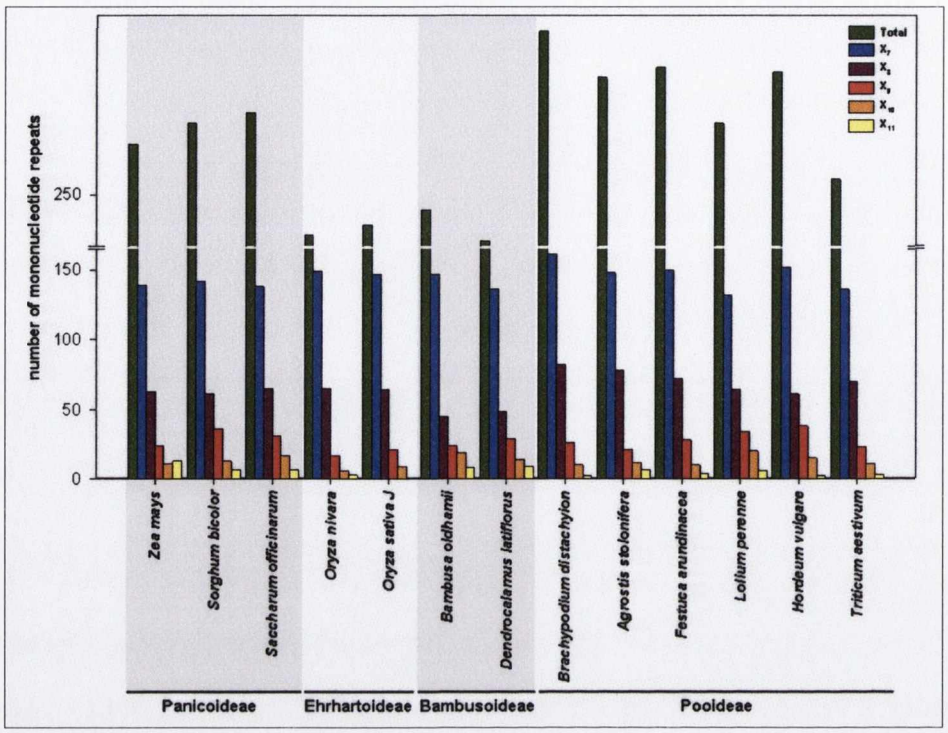

Figure 21 Number of microsatellites (>7 nucleotides) in complete Poaceae chloroplast genomes. X =the repeat length (7 to 11).

In the chloroplast genome of *L. perenne* 78 (33.05%) of the microsatellites were located within genes and 158 (66.94%) within intergenic spacers and introns. The microsatellites were, in general, distributed all over the genome; however, repeats with nine and more nucleotides were clearly less abundant in the inverted repeat region (Figure 23). Furthermore it can be seen that repeats based on ten and more nucleotides cluster mainly in four regions of the large single copy region: *matK – rpoB*, *psaA – trnV* UAC, *infA – trnH* GUG and in the intergenic spacer region between *psbE* and *petL*.

4.3.2 Chloroplast microsatellite analysis

Nine new primer sets for chloroplast microsatellite region markers were designed. They were demonstrated to be able to detect polymorphism, although to different extents. All markers were able to distinguish between *Lolium perenne* and non-*Lolium* grass species (Figure 24). Table 10 gives an overview of the marker performance within the *Lolium*

genus (among species). TeaCpSSR29 was the only marker not able to detect variation among *Lolium* species. TeaCpSSR32 detected single nucleotide polymorphisms within one individual of *L. persicum*. TeaCpSSR30 detected single nucleotide polymorphisms for a range of individuals in different *Lolium* accessions. TeaCpSSR33 detected SNPs but additionally found variation in the cpSSR regions. TeaCpSSR35 detected cpSSR variations. TeaCpSSR34 found a great degree of variation in microsatellite regions.

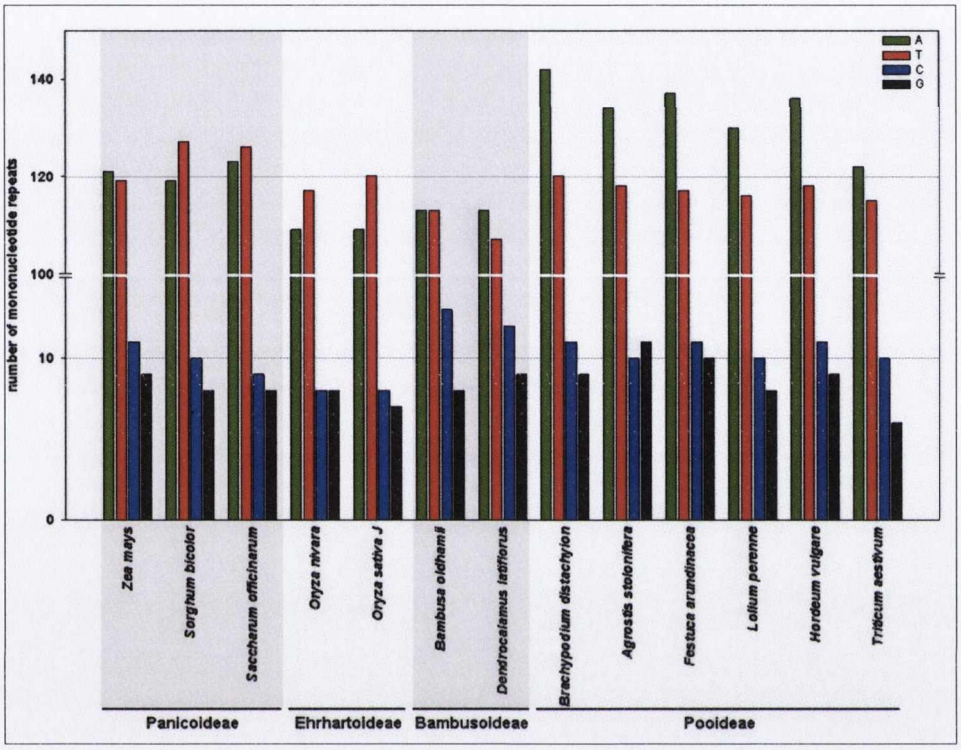

Figure 22 Nucleotide usage of chloroplast microsatellites (>7 nucleotides) in Poaceae species.

However, the three most informative markers were TeaCpSSR27, TeaCpSSR28 and TeaCpSSR31. TeaCpSSR28 was able to distinguish between all tested *Lolium* species and *L. multiflorum*. The difference between *L. multiflorum* and the other *Lolium* species was based within the A_8 mononucleotide repeat of that marker. This repeat showed an elongation of one A nucleotide for all *L. multiflorum*. TeaCpSSR31 combines a high amount of cpSSR length variation with a considerably high amount of SNPs. SNPs detected by this marker happened to be in the same individuals, SNPs had been observed before by TeaCpSSR30 and TeaCpSSR33. Testing the products of marker TeaCpSSR27 for amplification via relatively low/moderate resolution (in comparison to automated sequencer sizing) gel electrophoresis revealed variation within some *L. perenne* accessions (Figure 25). Sequencing revealed that this is due to an insertion/deletion event which results in the lack of a copy of a repeat of 44 nucleotides.

Interestingly all the haplotypes for which SNPs had been observed in the PCR products of TeaCpSSR30, TeaCpSSR31 and TeaCpSSR33 showed additionally this deletion in the sequencing results of TeaCpSSR27. The gel electrophoresis approach was mainly used as a control for amplification before the products were sent off for sequencing and thus was not appropriate to achieve a good separation between the two different haplotypes. Therefore the PCR for primer TeaCpSSR27 was repeated with more individuals of the accessions that showed this polymorphism (Figure 26) and confirmed the observations from the test gels.

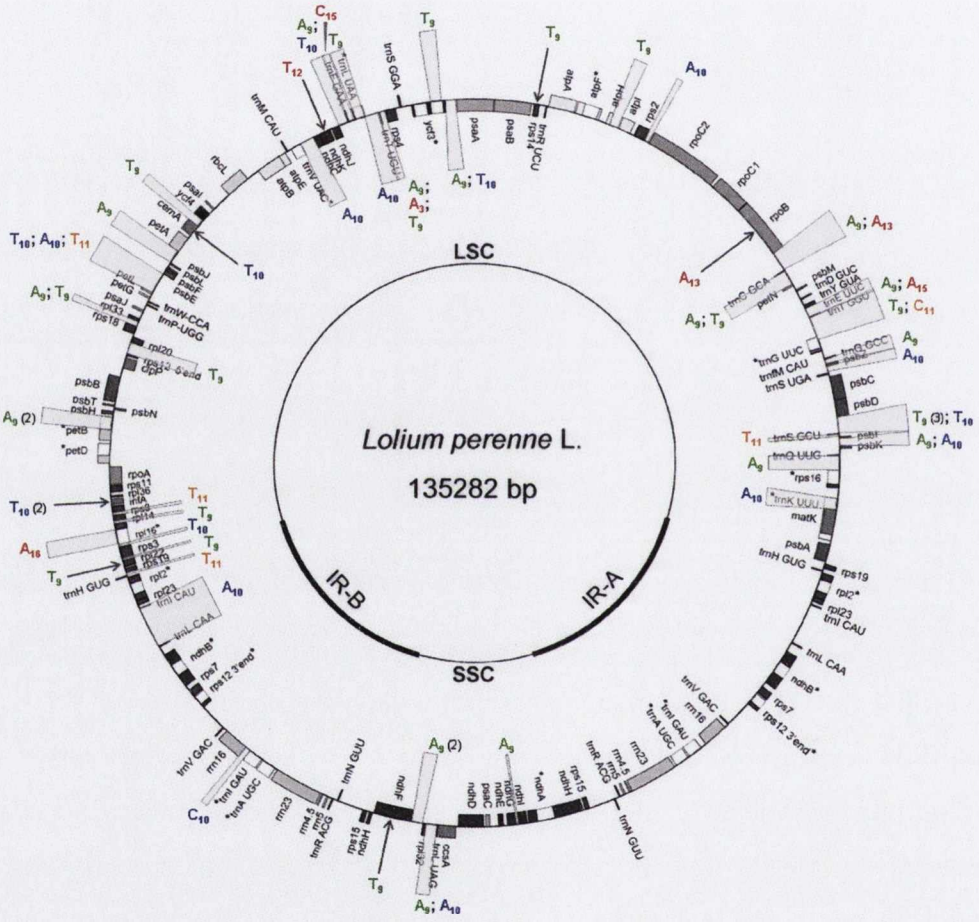

Figure 23 Distribution of microsatellites (> 9 nucleotides) in the chloroplast genome of *L. perenne*. Capital letters represent the nucleotide the repeat is based on, subscript numbers refer to the length of the mononucleotide repeat, numbers in brackets stand for the quantity of repeats within that specific region of the chloroplast genome. Colours were used based on the microsatellite length: green = 9 bp, blue = 10 bp, orange = 11 bp, red = >12 bp. Because both inverted repeats (IR) are identical to each other microsatellites for IR-A were not highlighted.

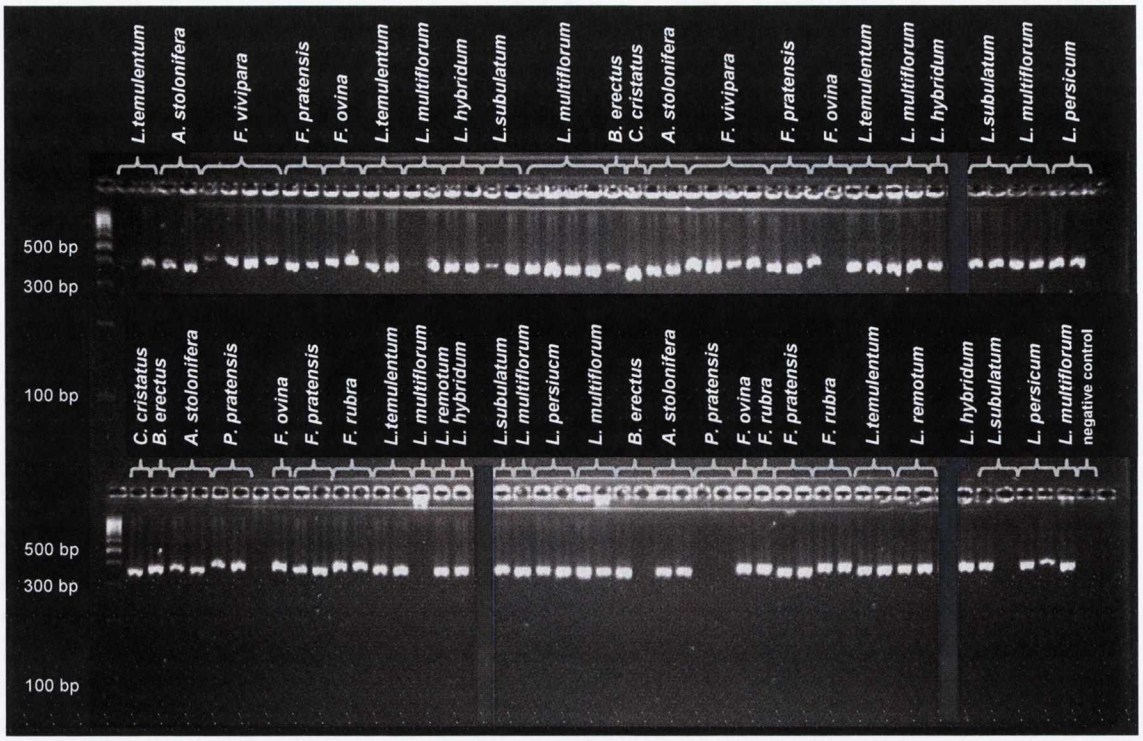

Figure 24 Amplification results using marker TeaCpSSR32 on a set of 14 different grass species. Sequencing revealed a length variation of 22 nucleotides – *Cynosurus cristatus* having the shortest product and the *Festuca* species except of *F. pratensis* the longest. *A.* = *Agrostis*, *B.* = *Bromus*, *C.* = *Cynosurus*, *F.* = *Festuca*, *L.* = *Lolium*, *P.* = *Poa*. / Products on a 2.5% MetaPhor® Agarose gel (Lonza, Rockland, ME, USA) stained with 1% ethidium bromide (10mg/ml). Size standard: 3 µl of mi-100 bp+ DNA Marker Go (metabion international AG, Martinsried, Germany), run for ~ 2 hours at 130 V.

4.3.3 Phylogenetic and haplotype analysis

Using UPGMA (results not shown) the 352 individual DNA sequences of this study were analysed to get an overview of the sequence results. This analysis revealed mainly two large *L. perenne* groups and several minor groups. It also separated the different *Lolium* species in general from the large *L. perenne* group, it resolved the *Festuca* species as one group with the exception of *F. pratensis* Northland, which was grouped together with the different *Lolium* species, and it recognized other grass species and the cereals as the most outlying taxa.

Arlequin 3.11 (Excoffier et al. 2005) detected 39 different haplotypes of which 33 different haplotypes were finally taken into account (Table 7, haplotype names do not refer to amount of haplotypes observed in total). Up to five haplotypes could be found within a single accession and only a few haplotypes were found that were unique to some of the accessions (private haplotypes). Haplotypes 1 and 2 were the most frequent,

of which at least one was found in each *L. perenne* accession. Most of the non-*Lolium perenne* species showed high variation with up to four haplotypes per accession although the accession sample size was rather small.

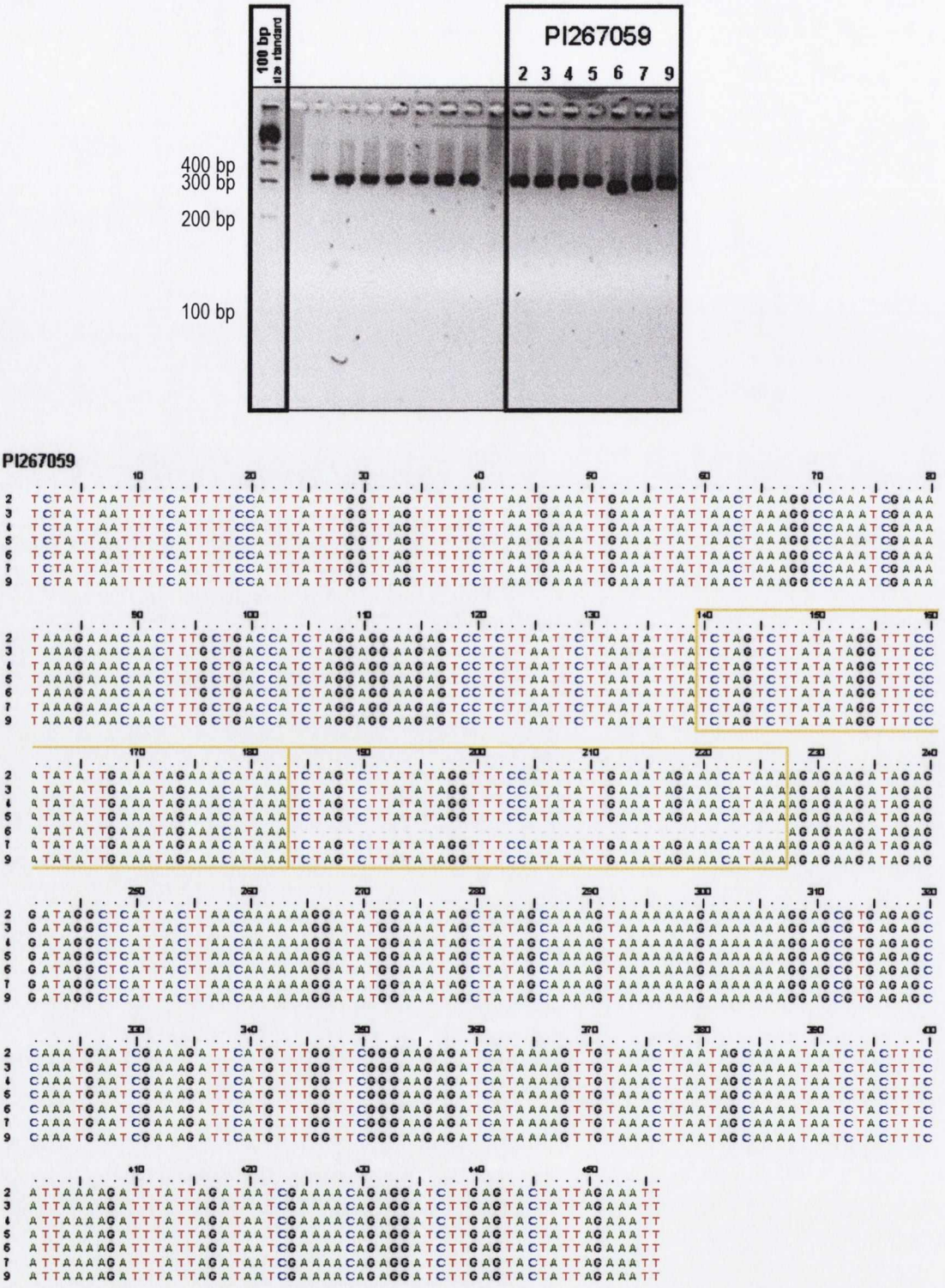

Figure 25 Agarose gel picture and alignment of DNA sequences obtained from PCR reactions with TeaCpSSR27 and different individuals (no. in front of alignment) from *Lolium perenne* accession PI267059. 2.5% MetaPhor® Agarose gel (Lonza, Rockland, ME, USA) with 1% ethidium bromide (10 mg/ml). Size standard: 3 µl of mi-100 bp+ DNA Marker Go (metabion international AG, Martinsried, Germany), run for ~ 2 hours at 130 V.

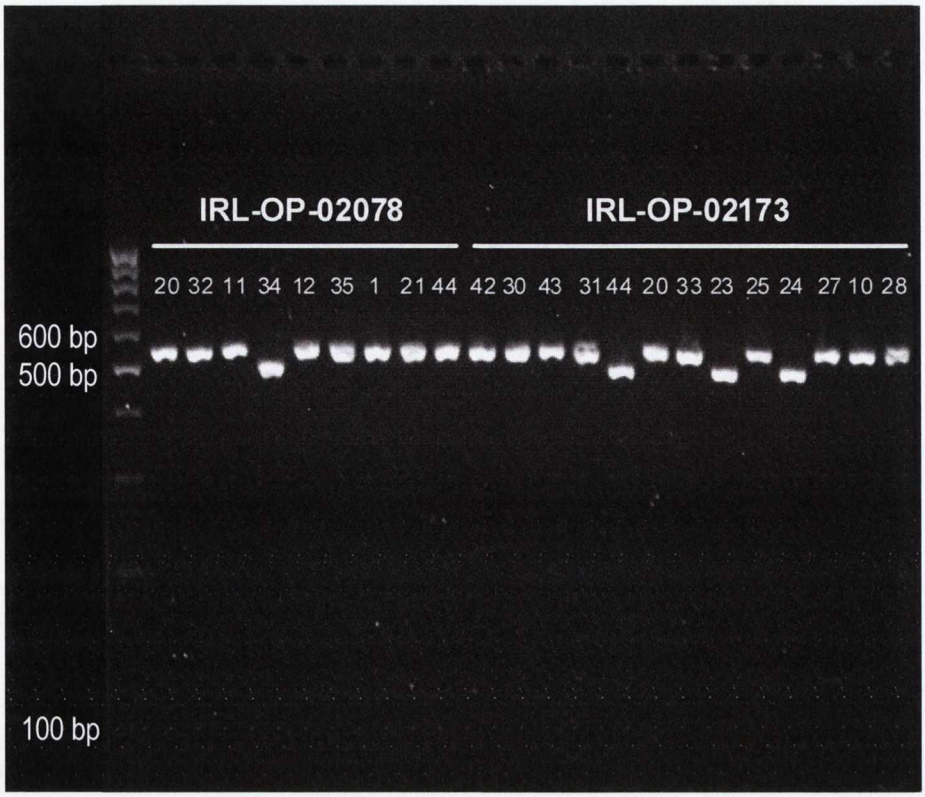

Figure 26 TeaCpSSR27 PCR products (5 µl/sample) from different individuals of two *Lolium perenne* accessions. Shorter amplicon products are due to the lack of one copy of a 44 base pair long repeat. 100 ml of a 2.0% MetaPhor® Agarose gel (Lonza, Rockland, ME, USA) stained with 1% ethidium bromide (10mg/ml). Size standard: 5 µl of mi-100 bp+ DNA Marker Go (metabion international AG, Martinsried, Germany), run for ~ 3 hours at 80 V.

Table 10 Ability of newly designed Poaceae universal primers to detect variation within the chloroplast genome of *Lolium perenne* and other *Lolium* species (*L. hybridum, L. multiflorum, L. persicum, L. remotum, L. subulatum, L. temulentum*).

Primer	*Lolium* spp.	*Lolium perenne*	form of variation		
			cpSSR	SNP	Indel
TeaCpSSR27	✓	✓	–	–	✓
TeaCpSSR28	✓	–	✓	–	–
TeaCpSSR29	–	–	–	–	–
TeaCpSSR30	✓	✓	–	✓	–
TeaCpSSR31	✓	✓	✓	✓	–
TeaCpSSR32	✓	–	–	✓	–
TeaCpSSR33	✓	✓	✓	✓	–
TeaCpSSR34	✓	✓	✓	–	–
TeaCpSSR35	✓	✓	✓	–	–

The haplotype DNA sequences as well as one sequence of *Avena sativa*, *Triticum aestivum*, *Secale cereale* and *Hordeum vulgare* were used in an input file for a Bayesian inference and parsimony analyses. The results were very similar to the ones previously

obtained with UPGMA when all individual sequences were considered. In both analyses *A. stolonifera* and *Avena sativa* from tribe Avenae were revealed as outgroup. High posterior probability and bootstrap values were obtained for *Bromus erectus* as sister to the Triticeae species, *S. cereale* being sister to *T. aestivum*, and *P. pratensis* being sister to all *Festuca* and *Lolium* species as well as to *Cynosurus cristatus*. High posterior probability values were only obtained for *F. ovina* as sister to a group of *F. rubra* and most *F. vivipara* individuals (but note *F. vivipara* was not monophyletic) and for all *Festucas* (except of *F. pratensis)* being sister to the *Lolium* species (Figure 27). The tree combining results of both approaches illustrates again that two haplotypes occur frequently in all the analysed accessions. It also shows that the ecotypes have a greater variation than the cultivars. *Festuca pratensis* Northland is again grouped within the *Lolium multiflorum* accessions as observed before in the UPGMA approach. Interestingly, the haplotypes of the ecotype accessions PI267059, IRL-OP-02173, IRL-OP-02078 and IRL-OP-02269 for which variation was observed already by gel electrophoresis grouped mainly with the other *Lolium* species. An analysis to verify that these plants were true *L. perenne* individuals was not carried out. Therefore a contamination of those populations with other *Lolium* species can not be excluded.

4.3.4 Locus-by-locus AMOVA

Each accession had some missing data for some of the analysed regions due to a lack of amplification of the PCR product or sequencing problems. In order to test the assigned genetic structure across all nine regions a locus-by-locus AMOVA had to be done as recommended by Excoffier et al. (2005). The number of polymorphic loci decreased from 158 if all six groups are included to 16 when considering only the European *L. perenne* ecotypes and *L. perenne* cultivars (Table 11). The highest variation could be found among groups when performing the AMOVA including all six groups. While there was still around 35% variation when using four groups, performing an AMOVA based on only *L. perenne* accessions showed no variation among the groups, but very high variation within the different populations analysed. Thus the fixation index F_{ST} decreased from 0.95 (all six groups) to 0.27 (only *L. perenne* accessions). All the variation found was highly significant at P = 99.9%. The AMOVA between the different *L. perenne* accessions revealed high partitioning of variation within

populations of the groups 1) and 2) (76.32%) rather than among them (25.14%) as well as with populations of groups 1) and 3) (83.32%). The results of the separate AMOVA analyses on the different *L. perenne* groups reveals that the largest percentage of variation within population relative to among populations, can be found within the Irish ecotypes (89.89% within vs. 10.11% among) while the cultivars and European ecotypes only showed ca. 50% within populations relative to ca. 50% among populations.

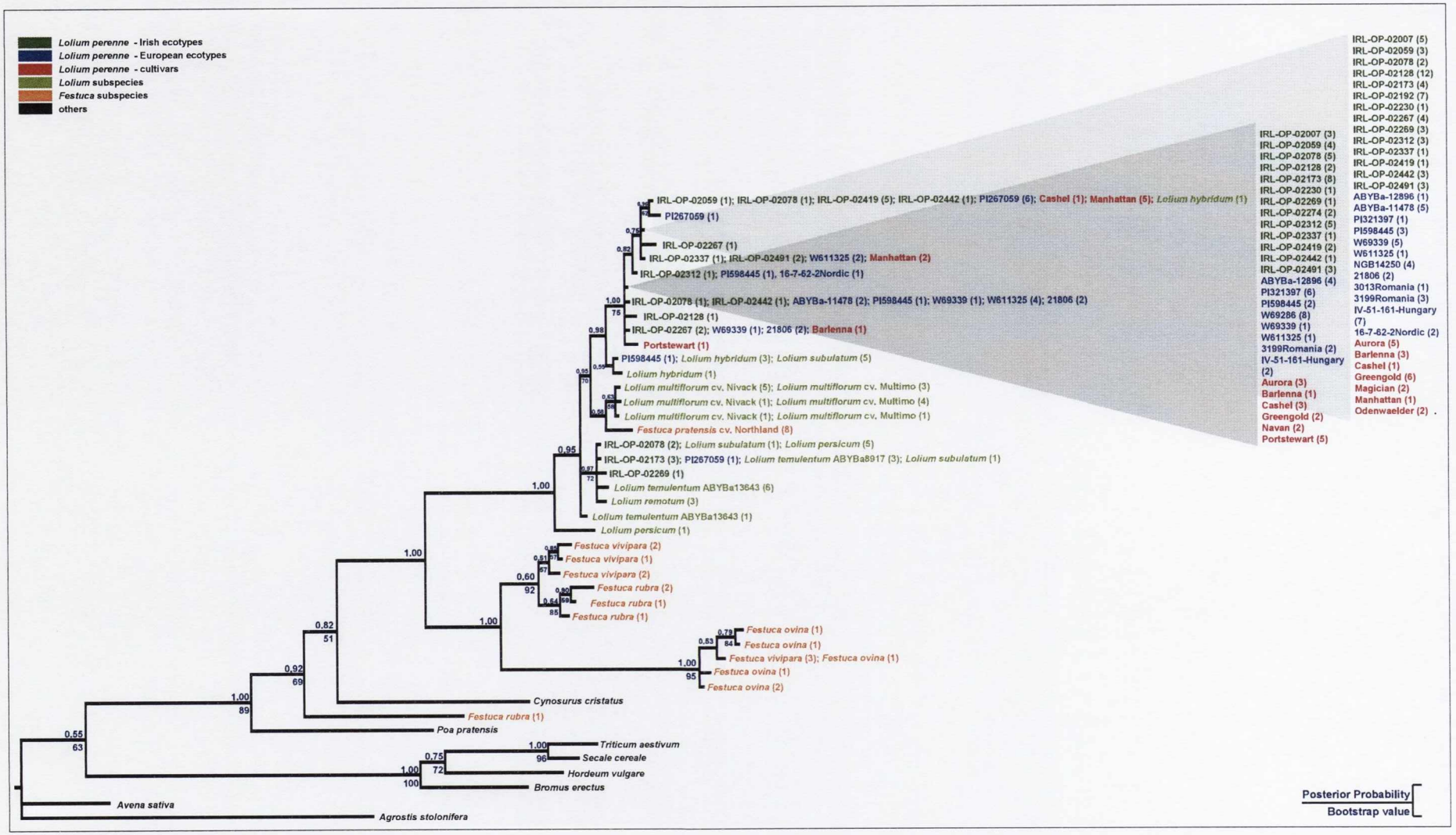

Figure 27 Bayesian inference tree using DNA sequences from haplotypes as input file. Haplotype names are replaced by the names of the accessions they were found in. Posterior probability values were obtained in MrBayes and Bootstrap values in PAUP* 4.0.

Table 11 Results of the locus-by-locus AMOVA. *** = highly significant, ns = not significant, 1 = *Lolium perenne* (Irish ecotypes), 2 = *Lolium perenne* (European ecotypes), 3 = *Lolium perenne* cultivars, 4 = *Lolium* species, 5 = *Festuca* species, 6 = other grasses

group composition		populations	groups	% variation			FST	FSC	FCT	P-value			polymorphic loci
				among groups	among populations / within groups	within populations				Vc & FST	Vb & FSC	Va & FCT	
				Va	Vb	Vc							
1-6	***Lolium* spp. + *Festuca* spp. + other grasses**	**54**	**6**	61.22	34.25	4.53	0.95	0.88	0.61	***	***	***	158
1-5	***Lolium* spp. + *Festuca* spp.**	**50**	**5**	60.00	30.42	9.57	0.90	0.76	0.60	***	***	***	77
1-4	***Lolium* spp.**	**46**	**4**	37.37	28.68	33.95	0.66	0.46	0.37	***	***	***	27
1-3	***Lolium perenne***	**38**	**3**	-2.13	28.65	73.48	0.27	0.28	-0.02	***	***	ns	21
1+2	***Lolium perenne* (Irish + European ecotypes)**	**29**	**2**	-1.46	25.14	76.32	0.24	0.25	-0.01	***	***	ns	19
1+3	***Lolium perenne* (Irish ecotypes + cultivars)**	**24**	**2**	-1.16	17.34	83.82	0.16	0.17	-0.01	***	***	ns	18
2+3	***Lolium perenne* (European ecotypes + cultivars)**	**23**	**2**	-5.93	54.52	51.41	0.49	0.51	-0.06	***	***	ns	16
3	***Lolium perenne* (cultivars)**	**9**	**1**		49.86	50.14	0.50			***			6
2	***Lolium perenne* (European ecotypes)**	**14**	**1**		49.59	50.41	0.50			***			15
1	***Lolium perenne* (Irish ecotypes)**	**15**	**1**		10.11	89.89	0.10			***			16

Not surprisingly the amount of polymorphic loci found by Arlequin decreases from 158 (all groups) to six (only *L. perenne* cultivars).

4.3.5 Genetic distance analysis

The results from the genetic distance (Nei and Li 1979) analyses are given in Table 12 and Table 13. The sample size of the populations was in general very small (less than ten individuals per population) therefore the values displayed in these tables are the corrected values obtained from Arlequin 3.11 (Excoffier et al. 2005). Thus the genetic distance found between the samples is less likely to be overestimated. In general the analysis of genetic distance confirmed the earlier results obtained with the other statistical approaches. The highest distance values (on average) were obtained between *L. perenne* and *Agrostis stolonifera* (82.68), *Bromus erectus* (69.35), *Cynosurus cristatus* (43.79) and *Poa pratensis* (47.37), respectively. The distance between *L. perenne* and *Festuca* species was in general lower. The largest distance was observed between *L. perenne* and *Festuca ovina* (52.65) and the lowest was obtained for *L. perenne* and *Festuca pratensis* cv. Northland (8.87). The distance between *L. perenne* and the other *Lolium* species decreased to 2.00 (*L. hybridum*) to 9.13 (*L. remotum*). Distance values obtained for *Lolium* species, *Festuca* species and the other grasses were generally significant to highly significant. Genetic distance values between the different *L. perenne* accessions varied from 0.00 to 3.32 (PI267059/IRL-OP-02274) and reached the highest values within the populations (4.55; IRL-OP-02173). However, most observed distances were minor and rarely significant. Significant values were only observed for few pairs of accessions and involved only the accessions: IRL-OP-02078, IRL-OP-02192, PI267059, 21806 and cv. Manhattan. The extent of significant distances between the different European accessions seemed to be higher than between the Irish accessions. The genetic distance analysis for *L. perenne* accessions using only primer sets TeaCpSSR27 and TeaCpSSR31, confirmed generally the results obtained when using all sequences and samples. The distances within the different populations decreased to nearly half its size. In addition to the already mentioned four accessions, above that had a high amount of significant distances, two more accessions showed an abundance of significant values: W69286 and W611325.

Table 12 Results from the genetic distance analysis including all accessions and sequences from all markers. not significant, significant (P<0.05), significant (P<0.01), significant (P<0.001), grey highlighted are genetic distances within the accessions.

		Lolium perenne - Irish ecotypes													*Lolium perenne* - European ecotypes												*Lolium perenne* - cultivars								*Lolium* subspecies							*Festuca* subspecies				other Poaceae species			
		IRL-OP-02067	IRL-OP-02069	IRL-OP-02078	IRL-OP-02128	IRL-OP-02173	IRL-OP-02192	IRL-OP-02230	IRL-OP-02267	IRL-OP-02269	IRL-OP-02274	IRL-OP-02312	IRL-OP-02337	IRL-OP-02419	IRL-OP-02442	IRL-OP-02491	ABYBa-12896	ABYBa-11478	PI321397	PI267059	PI598445	W69286	W69339	W611325	NGB14250	21806	3013Romania	3199Romania	IV-51-161-Hungary	16-7-62-2Nordic	Aurora	Barlenna	Cashel	Greengold	Magician	Manhattan	Navan	Odenwaelder	Portstewart	*L. temulentum* ABYBa13643	*L. temulentum* ABYBa8917	*L. multiflorum* Nivack	*L. remotum*	*L. hybridum*	*L. subulatum*	*L. persicum*	*L. multiflorum* Multimo	*F. vivipara*	*F. pratensis* Northland	*F. ovina*	*F. rubra*	*Agrostis stolonifera*	*Bromus erectus*	*Cynosurus cristatus*	*Poa pratensis*
Lolium perenne - Irish ecotypes	IRL-OP-02067	1.07																																																
	IRL-OP-02069	0.00	1.64																																															
	IRL-OP-02078	0.87	0.53	2.91																																													
	IRL-OP-02128	0.29	0.22	0.71	4.28																																												
	IRL-OP-02173	0.34	0.13	0.00	0.64	4.55																																											
	IRL-OP-02192	0.16	0.32	1.29	0.52	0.74	0.00																																										
	IRL-OP-02230	0.00	0.00	0.32	0.00	0.00	0.00	2.00																																									
	IRL-OP-02267	0.00	0.00	0.55	0.11	0.28	0.00	0.00	2.29																																								
	IRL-OP-02269	0.36	0.20	0.00	0.00	0.00	0.41	0.00	0.06	3.70																																							
	IRL-OP-02274	0.71	0.43	1.00	1.36	0.46	1.86	0.00	0.86	0.95	0.00																																						
	IRL-OP-02312	0.00	0.00	0.69	0.40	0.09	0.55	0.00	0.02	0.29	0.29	0.97																																					
	IRL-OP-02337	0.00	0.00	0.58	0.20	0.15	0.00	0.00	0.00	0.02	0.67	0.00	2.00																																				
	IRL-OP-02419	0.61	0.38	0.62	0.54	0.78	0.61	0.26	0.27	0.57	1.70	0.65	0.03	1.86																																			
	IRL-OP-02442	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.73	0.00	0.00	0.00	1.53																																		
	IRL-OP-02491	0.00	0.00	0.64	0.29	0.21	0.16	0.00	0.00	0.15	0.75	0.00	0.00	0.34	0.00	1.50																																		
Lolium perenne - European ecotypes	ABYBa-12896	0.00	0.00	0.22	0.34	0.00	0.49	0.00	0.00	0.00	0.00	0.00	0.00	0.77	0.40	0.00	0.40																																		
	ABYBa-11478	0.00	0.00	0.00	0.00	0.00	0.24	0.00	0.00	0.00	1.00	0.16	0.00	0.32	0.11	0.00	0.66	0.00																																
	PI321397	0.00	0.00	0.34	0.51	0.00	0.73	0.00	0.10	0.12	0.00	0.00	0.00	0.79	0.52	0.00	0.00	0.86	0.00																																
	PI267059	1.94	1.55	0.79	1.57	1.62	1.48	1.51	1.25	0.77	3.32	1.97	1.11	0.13	0.72	1.42	2.10	0.77	2.07	3.11																															
	PI598445	0.00	0.00	0.21	0.00	0.00	0.45	0.00	0.05	0.15	0.32	0.00	0.00	0.80	0.18	0.00	0.00	0.21	0.00	1.82	1.11																														
	W69286	0.09	0.00	0.43	0.52	0.00	0.86	0.00	0.29	0.35	0.00	0.00	0.00	1.07	0.73	0.13	0.00	0.93	0.00	2.61	0.07	0.00																													
	W69339	0.00	0.00	0.24	0.00	0.00	0.16	0.00	0.00	0.00	0.50	0.00	0.00	0.31	0.00	0.00	0.13	0.16	0.25	1.31	0.04	0.38	1.00																											
	W611325	0.45	0.31	0.69	0.67	0.41	0.54	0.10	0.00	0.26	1.54	0.56	0.00	0.11	0.08	0.08	0.96	0.66	1.07	0.59	0.78	1.29	0.27	0.43																										
	NGB14250	0.12	0.30	1.23	0.02	0.74	0.00	0.00	0.07	0.55	1.75	0.49	0.00	0.76	0.07	0.16	0.60	0.21	0.86	1.73	0.61	1.00	0.22	0.66	0.00																									
	21806	0.76	0.35	0.50	1.04	0.02	1.48	0.17	0.38	0.32	0.67	0.48	0.11	0.45	0.46	0.38	0.60	1.19	0.57	1.01	0.61	0.67	0.46	0.45	1.58	0.00																								
	3013Romania	0.00	0.00	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00	1.51	0.00	0.00	0.00	0.10	0.00	0.17	2.00																							
	3199Romania	0.00	0.00	0.56	0.00	0.03	0.14	0.00	0.00	0.19	0.60	0.00	0.00	0.57	0.00	0.00	0.00	0.00	0.00	1.62	0.00	0.00	0.00	0.41	0.10	0.67	0.00	1.20																						
	IV-51-161-Hungary	0.00	0.01	0.90	0.00	0.40	0.05	0.00	0.00	0.34	1.08	0.13	0.00	0.63	0.00	0.00	0.11	0.02	0.31	1.74	0.13	0.42	0.00	0.49	0.03	1.05	0.00	0.00	0.72																				
	16-7-62-2Nordic	0.03	0.00	0.62	0.31	0.02	0.71	0.00	0.09	0.28	0.18	0.00	0.00	0.73	0.00	0.04	0.00	0.25	0.00	2.02	0.00	0.00	0.00	0.65	0.65	0.43	0.00	0.00	0.23	0.89																			
Lolium perenne - cultivars	Aurora	0.00	0.00	0.88	0.23	0.34	0.16	0.00	0.00	0.31	0.71	0.00	0.00	0.61	0.00	0.00	0.00	0.02	0.00	1.88	0.00	0.09	0.00	0.45	0.12	0.78	0.00	0.00	0.00	0.03	1.07																		
	Barlenna	0.08	0.00	0.19	0.65	0.00	0.86	0.00	0.00	0.00	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.46	0.00	0.47	0.00	0.00	0.00	0.11	0.65	0.02	0.00	0.00	0.26	0.00	0.06	0.50																	
	Cashel	0.14	0.00	0.11	0.00	0.00	0.86	0.00	0.00	0.00	0.55	0.09	0.00	0.00	0.00	0.00	0.11	0.09	0.24	0.47	0.07	0.35	0.00	0.19	0.60	0.28	0.00	0.07	0.28	0.10	0.14	0.00	0.90																
	Greengold	0.00	0.00	0.78	0.44	0.15	0.11	0.00	0.00	0.02	0.89	0.05	0.00	0.31	0.00	0.00	0.00	0.11	0.07	1.29	0.00	0.14	0.00	0.30	0.08	0.73	0.00	0.00	0.00	0.13	0.00	0.14	0.19	0.71														
	Magician	0.00	0.00	0.41	0.36	0.00	0.00	0.00	0.00	0.00	1.00	0.07	0.00	0.00	0.00	0.00	0.20	0.29	0.43	0.51	0.01	0.60	0.00	0.22	0.00	0.83	0.00	0.00	0.00	0.18	0.00	0.25	0.15	0.00	0.00												
	Manhattan	1.40	1.16	1.42	1.56	1.82	1.11	1.13	0.79	1.10	2.95	1.57	0.57	0.02	0.49	0.96	1.82	1.14	1.86	0.03	1.72	2.20	1.02	0.37	1.32	1.07	1.13	1.37	1.32	1.70	1.40	0.65	0.65	1.00	0.57	0.36											
	Navan	0.71	0.43	0.77	0.58	0.22	1.86	0.00	0.93	0.95	0.00	0.29	0.67	1.70	0.73	0.75	0.00	0.86	0.00	3.07	0.32	0.00	0.50	1.54	1.75	0.67	0.00	0.60	1.08	0.18	0.71	0.35	0.55	0.89	1.00	2.95	0.00										
	Odenwaelder	0.21	0.43	1.64	0.68	1.12	0.00	0.00	0.14	0.75	2.00	0.63	0.00	0.82	0.07	0.25	0.60	0.29	0.86	1.95	0.57	1.00	0.25	0.66	0.00	1.67	0.00	0.20	0.08	0.80	0.21	0.75	0.75	0.14	0.00	1.32	2.00	0.00									
	Portstewart	0.68	0.33	0.87	0.85	0.28	1.78	0.00	0.80	0.85	0.07	0.30	0.57	1.39	0.49	0.67	0.03	0.68	0.07	2.78	0.16	0.07	0.38	1.35	1.57	0.57	0.00	0.37	0.85	0.16	0.59	0.38	0.42	0.88	0.98	2.62	0.00	1.90	0.53								
Lolium subspecies	*L. temulentum* ABYBa13643	8.93	8.28	3.18	7.43	6.60	8.37	8.31	8.19	3.89	9.45	8.33	8.36	6.86	6.86	8.36	7.42	5.04	6.80	4.80	7.38	7.38	7.35	6.85	8.74	5.78	8.31	8.06	8.82	8.15	8.93	5.20	6.57	7.67	5.81	6.92	8.62	8.60	8.30	0.24								
	L. temulentum ABYBa8917	8.30	8.80	3.48	7.74	6.98	8.76	8.50	8.52	3.82	9.67	8.68	8.67	8.28	7.40	8.67	7.67	5.33	7.10	6.65	7.65	7.67	7.63	7.12	9.08	5.94	8.50	8.27	8.94	8.43	9.21	5.48	6.28	8.02	6.17	8.36	8.67	10.00	8.62	0.88	0.00							
	L. multiflorum Nivack	6.15	6.98	3.94	5.69	5.10	5.15	6.33	5.46	2.40	7.48	6.32	5.86	4.98	5.11	5.40	5.96	2.64	6.50	4.60	4.98	6.55	4.96	4.94	5.30	5.96	6.33	5.28	5.53	6.32	6.56	5.06	4.05	5.73	4.62	5.67	6.33	6.33	6.90	6.56	6.67	0.76						
	L. remotum	10.55	10.22	4.48	9.09	7.06	9.81	9.83	10.33	4.75	10.67	9.88	10.67	10.57	9.23	10.25	8.67	6.86	8.33	8.57	8.53	8.83	9.13	9.74	10.17	8.50	9.83	9.33	10.05	8.60	10.58	8.96	7.55	9.23	7.17	11.32	9.50	11.33	9.73	2.60	2.00	8.48	0.00					
	L. hybridum	2.46	2.03	0.24	1.76	1.02	2.69	1.95	2.52	0.70	1.75	1.93	2.35	2.60	1.98	2.10	1.95	1.38	1.72	3.10	1.85	1.96	2.03	2.66	3.15	1.75	1.95	2.16	2.52	1.76	2.56	0.74	0.98	1.97	1.15	3.62	1.65	3.55	1.72	5.15	5.48	3.65	6.35	2.90				
	L. subulatum	2.84	2.27	0.51	2.17	0.77	3.01	2.21	2.42	0.34	2.21	2.46	2.45	2.30	1.69	2.23	1.99	1.62	2.05	2.07	1.88	2.18	1.93	2.29	3.25	1.81	2.21	2.39	2.74	2.27	2.86	1.39	1.48	2.48	1.64	3.16	2.07	3.79	2.47	2.54	2.55	3.26	3.60	0.00	1.86			
	L. persicum	6.03	5.74	2.36	4.41	3.21	5.55	5.50	5.71	1.52	6.17	5.90	6.00	6.59	4.04	5.62	4.13	2.36	4.19	4.43	4.22	4.29	4.44	5.04	5.67	4.76	5.33	5.23	5.58	5.74	5.94	4.15	4.42	5.41	3.83	6.30	5.67	6.67	6.15	0.81	0.39	4.40	0.50	3.18	1.98	2.00		
	L. multiflorum Multimo	6.10	5.97	4.00	5.61	5.16	5.05	6.32	5.61	2.25	7.32	6.29	6.07	5.24	5.30	5.45	5.92	2.43	6.43	5.14	4.77	6.45	5.01	5.44	5.20	6.45	6.32	5.12	5.40	6.27	6.54	5.30	4.07	5.75	4.63	6.25	6.07	6.32	6.83	6.93	7.07	0.12	6.32	3.52	3.39	4.30	1.11	
Festuca subspecies	*F. vivipara*	41.94	42.39	33.55	37.78	40.06	36.00	44.54	44.79	22.59	30.41	43.40	46.45	41.23	37.66	35.46	40.19	25.52	41.86	37.87	36.19	38.66	39.69	42.84	42.79	39.91	38.54	40.76	40.91	41.65	40.05	31.09	33.99	39.58	31.91	43.84	35.29	47.91	38.21	31.92	28.49	42.87	38.87	36.81	34.59	36.10	44.14	1.18			
	F. pratensis Northland	10.09	9.43	6.34	8.85	8.68	9.55	10.13	9.34	5.03	10.13	9.57	9.46	8.45	8.69	9.00	9.93	5.27	9.98	5.45	8.17	10.13	8.56	8.47	9.59	8.46	10.13	8.93	8.29	9.02	10.46	7.98	5.83	9.64	7.63	9.51	8.63	11.13	8.19	10.70	10.75	6.30	13.08	6.20	4.89	8.21	6.20	44.68	0.00		
	F. ovina	57.24	58.54	39.15	48.51	57.96	52.99	57.97	58.54	37.02	55.97	55.55	59.41	57.95	54.31	55.57	56.10	36.87	53.28	56.36	51.09	56.34	55.88	56.06	56.43	47.66	53.80	54.90	56.25	53.27	54.99	38.48	40.58	51.67	41.63	56.60	54.80	61.47	50.34	54.78	55.74	61.44	59.63	53.92	37.96	41.72	61.95	25.61	64.30	2.40	
	F. rubra	31.91	33.30	26.43	27.51	31.49	28.96	34.30	34.21	13.69	25.20	34.06	35.30	32.15	27.13	27.45	29.98	17.99	32.27	28.22	25.80	29.93	30.30	32.89	32.85	30.83	28.00	31.14	31.74	32.20	29.14	24.93	26.01	31.74	25.70	33.85	28.60	37.70	29.03	27.12	23.90	34.12	30.67	29.77	27.96	29.87	34.85	3.69	35.83	5.27	13.40
other Poaceae species	*Agrostis stolonifera*	90.57	91.15	61.75	78.49	89.82	80.67	93.50	93.46	59.33	78.50	87.93	94.83	89.56	85.03	83.66	88.60	60.27	86.18	85.54	81.98	88.63	88.38	87.30	91.47	76.06	86.00	89.10	89.25	84.27	87.53	57.63	85.06	90.21	62.13	87.87	83.06	98.38	78.94	78.04	79.58	93.03	90.54	83.45	60.41	69.81	95.52	71.01	93.56	76.47	44.55	0.00						
	Bromus erectus	74.78	75.05	64.59	65.55	74.36	69.50	76.38	76.57	42.60	63.13	75.93	78.25	74.01	66.69	67.06	70.40	46.57	75.29	69.26	61.04	71.84	71.34	78.04	72.56	72.38	67.75	70.70	70.64	70.08	69.62	63.90	60.05	74.02	62.00	77.42	65.75	83.00	69.11	64.77	59.33	77.73	68.92	64.75	61.07	66.21	80.32	85.94	77.31	48.43	54.95	54.09	0.00					
	Cynosurus cristatus	44.96	47.93	39.50	42.02	46.26	42.64	50.50	48.71	21.15	39.00	47.29	49.50	46.26	44.32	40.63	47.60	30.79	48.93	42.61	39.13	45.75	46.69	49.41	46.75	46.50	42.50	46.10	45.25	44.93	42.71	38.75	36.35	45.96	41.50	48.45	43.25	52.50	41.23	42.38	36.17	50.90	44.83	40.65	36.79	41.92	52.95	43.66	53.00	30.80	32.00	52.06	63.50	0.00				
	Poa pratensis	53.95	51.20	44.08	43.15	50.37	47.58	54.21	53.16	31.63	43.58	52.65	54.00	50.03	42.94	46.74	43.93	29.30	48.01	47.44	40.87	46.33	45.93	48.40	49.33	49.00	51.46	50.28	47.75	47.70	52.11	45.08	40.98	51.36	37.33	53.12	45.08	57.33	51.61	40.54	38.25	52.06	44.00	42.93	40.15	41.96	54.90	42.06	54.08	29.22	31.08	44.80	53.02	41.46	0.33	

Table 13 Results from the genetic distance analysis focussing only on *Lolium perenne* accessions and sequences derived from data of markers TeaCpSSR27 and TeaCpSSR31. not significant, significant (P<0.05), significant (P<0.01), significant (P<0.001), grey highlighted are genetic distances within the accessions.

		Lolium perenne - Irish ecotypes													*Lolium perenne* - European ecotypes													*Lolium perenne* - cultivars					
		IRL-OP-02007	IRL-OP-02059	IRL-OP-02078	IRL-OP-02128	IRL-OP-02173	IRL-OP-02192	IRL-OP-02267	IRL-OP-02269	IRL-OP-02312	IRL-OP-02337	IRL-OP-02419	IRL-OP-02442	IRL-OP-02491	ABYBa-12896	ABYBa-11478	PI321397	PI267059	PI598445	W69286	W69339	W611325	NGB14250	21806	3199Romania	IV-51-161-Hungary	16-7-62-2Nordic	Aurora	Barlenna	Cashel	Greengold	Manhattan	Portstewart
Lolium perenne - Irish ecotypes	IRL-OP-02007	0.54																															
	IRL-OP-02059	0.00	1.07																														
	IRL-OP-02078	0.73	0.45	1.71																													
	IRL-OP-02128	0.00	0.01	0.78	0.31																												
	IRL-OP-02173	0.27	0.11	0.00	0.34	2.30																											
	IRL-OP-02192	0.11	0.21	1.02	0.04	0.65	0.00																										
	IRL-OP-02267	0.00	0.00	0.56	0.00	0.24	0.02	1.81																									
	IRL-OP-02269	0.28	0.11	0.00	0.27	0.00	0.45	0.05	2.50																								
	IRL-OP-02312	0.00	0.00	0.64	0.11	0.15	0.36	0.10	0.28	0.54																							
	IRL-OP-02337	0.00	0.00	0.48	0.00	0.09	0.00	0.00	0.02	0.00	1.33																						
	IRL-OP-02419	0.67	0.40	0.74	0.59	0.83	0.75	0.33	0.53	0.79	0.17	1.50																					
	IRL-OP-02442	0.05	0.00	0.48	0.00	0.31	0.07	0.00	0.08	0.22	0.00	0.11	1.53																				
	IRL-OP-02491	0.00	0.00	0.49	0.00	0.14	0.14	0.00	0.07	0.03	0.00	0.39	0.00	0.96																			
Lolium perenne - European ecotypes	ABYBa-12896	0.11	0.01	0.59	0.27	0.11	0.60	0.21	0.31	0.00	0.07	0.95	0.40	0.14	0.40																		
	ABYBa-11478	0.04	0.09	0.81	0.01	0.40	0.00	0.00	0.25	0.25	0.00	0.50	0.00	0.04	0.47	0.00																	
	PI321397	0.27	0.11	0.57	0.47	0.07	0.86	0.34	0.32	0.05	0.19	0.96	0.52	0.27	0.00	0.71	0.00																
	PI267059	1.78	1.36	1.04	1.59	1.62	1.71	1.20	1.02	1.97	1.02	0.10	0.77	1.25	2.14	1.33	2.05	1.71															
	PI598445	0.03	0.00	0.55	0.13	0.17	0.39	0.11	0.24	0.00	0.00	0.83	0.25	0.07	0.00	0.25	0.05	1.99	0.96														
	W69286	0.36	0.21	0.77	0.58	0.25	1.00	0.52	0.55	0.11	0.33	1.25	0.73	0.39	0.00	0.83	0.00	2.52	0.14	0.00													
	W69339	0.00	0.00	0.58	0.00	0.19	0.14	0.00	0.14	0.03	0.00	0.39	0.00	0.00	0.14	0.04	0.27	1.30	0.07	0.39	0.96												
	W611325	0.63	0.47	0.94	0.57	0.68	0.66	0.15	0.44	0.78	0.00	0.22	0.08	0.26	0.96	0.51	1.07	0.58	0.79	1.29	0.26	0.43											
	NGB14250	0.11	0.21	1.14	0.04	0.71	0.00	0.10	0.55	0.36	0.00	0.75	0.07	0.14	0.60	0.00	0.86	1.77	0.39	1.00	0.14	0.66	0.00										
	21806	0.82	0.46	0.56	0.95	0.23	1.33	0.38	0.28	0.65	0.22	0.42	0.46	0.43	0.60	1.11	0.57	0.77	0.60	0.67	0.43	0.37	1.33	0.00									
	3199Romania	0.00	0.00	0.69	0.00	0.23	0.10	0.00	0.25	0.00	0.00	0.65	0.03	0.00	0.06	0.03	0.21	1.77	0.00	0.30	0.00	0.61	0.10	0.77	0.60								
	IV-51-161-Hungary	0.00	0.02	0.81	0.00	0.39	0.03	0.00	0.31	0.11	0.00	0.67	0.02	0.00	0.27	0.00	0.47	1.71	0.14	0.58	0.00	0.61	0.03	0.99	0.00	0.39							
	16-7-62-2Nordic	0.04	0.00	0.45	0.17	0.00	0.45	0.04	0.12	0.00	0.00	0.70	0.20	0.04	0.00	0.34	0.04	1.71	0.00	0.07	0.04	0.76	0.45	0.57	0.00	0.17	0.36						
Lolium perenne - cultivars	Aurora	0.00	0.00	0.73	0.00	0.27	0.11	0.00	0.28	0.00	0.00	0.67	0.05	0.00	0.11	0.04	0.27	1.78	0.03	0.36	0.00	0.63	0.11	0.82	0.00	0.00	0.04	0.54					
	Barlenna	0.32	0.05	0.18	0.48	0.00	0.83	0.02	0.00	0.15	0.00	0.33	0.18	0.10	0.10	0.67	0.12	0.89	0.10	0.17	0.10	0.41	0.83	0.06	0.27	0.49	0.11	0.32	0.33				
	Cashel	0.08	0.00	0.30	0.09	0.06	0.25	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.25	0.13	0.36	0.39	0.11	0.50	0.00	0.16	0.25	0.33	0.05	0.11	0.13	0.08	0.08	1.00			
	Greengold	0.00	0.00	0.38	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.36	0.00	0.00	0.11	0.07	0.29	1.14	0.00	0.36	0.00	0.46	0.11	0.69	0.00	0.00	0.05	0.00	0.27	0.00	0.29		
	Manhattan	1.38	1.07	1.31	1.24	1.47	1.32	0.81	1.02	1.60	0.65	0.01	0.49	0.95	1.82	1.07	1.86	0.00	1.60	2.20	0.95	0.37	1.32	0.82	1.37	1.32	1.50	1.38	0.95	0.35	1.02	0.36	
	Portstewart	0.25	0.09	0.54	0.45	0.04	0.83	0.31	0.28	0.04	0.17	0.92	0.48	0.25	0.00	0.69	0.00	1.98	0.04	0.00	0.25	1.04	0.83	0.56	0.20	0.45	0.03	0.25	0.11	0.33	0.27	1.80	0.00

4.4 Discussion

4.4.1 Characterization of Poaceae chloroplast microsatellites

Thirteen Poaceae chloroplast genomes were searched for mononucleotide repeat regions of more than seven base pairs in length. All genomes analysed contained more than 200 cpSSR repeats. This is in agreement with the result obtained by Jakobssen et al. (2007) when searching the *Arabidopsis thaliana* chloroplast genome for microsatellites. However, the total amount of microsatellites found within the chloroplast genome can vary greatly between different species (Powell et al. 1995b) as was also confirmed by this study. Furthermore the obtained results provide evidence that the number of microsatellites observed in different subfamilies is not random. Pooideae and Panicoideae have a significant (P = 99.99%) higher amount of chloroplast microsatellites than Ehrhartoideae and Bambusoideae. When comparing the size of complete chloroplast genome sequences of Poaceae species it was already observed that the species group in size according to their subfamilies (Chapter 2).

Many chloroplast microsatellite analyses focus on long cpSSRs (above 10 bp) as was also done in this study, because they are thought to show higher variation (Echt et al. 1998). This, however, reduces the amount of usable cpSSRs dramatically, because less than 13% of the microsatellites have lengths of ten base pairs and more (Figure 21). Primers designed in this study amplified also shorter microsatellites, but revealed no or only minor variation at intraspecific level (TeaCpSSR31, TeaCpSSR35). Jakobsson et al. (2007) obtained very similar results for *Arabidopsis*. They did not detect any variation for microsatellites shorter than seven base pairs, only moderate variation for microsatellites up to twelve base pairs long but complete variation in microsatellites of more than twelve base pairs in length (Jakobsson et al. 2007). Thus microsatellite studies focussing mainly on repeats with at least ten base pairs are more likely to detect variation.

Many research groups found a strong bias of the nucleotides A and T for mononucleotide repeats (Flannery et al. 2006, Powell et al. 1995b, Rajendrakumar et al. 2006). However, this analysis revealed that, although the nucleotides A and T are clearly favoured across all species, also the nucleotide usage seems to be subfamily

specific, with Pooideae clearly favouring nucleotide A and Ehrhartoideae clearly favouring nucleotide T (Figure 22). This observation is to our knowledge the first of its kind. However, more Poaceae chloroplast genomes, especially for the subfamilies Ehrhartoideae and Bambusoideae need to be sequenced to confirm this result. The two sequenced bamboos are closely related to each other and some taxonomists even combine them in the same genus, also the two rice taxa analysed here belong in only one genus of Ehrhartoideae (*Oryza*). However, despite this uncertainty, there are clearly differences between species.

The majority of microsatellites were located in intergenic spacer regions (67%). This observation is in agreement with results from Powell et al. (1995b) and Rajendrakumar et al. (2006). Approximately 70% of the microsatellites observed in *Arabidopsis* were located in non-coding regions (Jakobsson et al. 2007), additionally confirming the observation made in *L. perenne*. Most of the cpSSRs were furthermore located in the large single copy region as already observed by Powell et al. (1995b). Interestingly Powell et al. (1995b) observed similar regions for the accumulation of long repeats for *Oryza sativa* as this study did, which illustrates again how conserved the chloroplast genome is. It is very likely that long microsatellites evolve due to slipped-strand mispairing (Kelchner 2000). The accumulation of microsatellites in the LSC is thus not surprising because this region of the chloroplast genome is known to be more variable than the IR region. In the IR a mutation correction mechanism exists that keeps both IRs identical, thus the nucleotide substitution is much lower (Wolfe et al. 1987). Hence mutations are less likely to accumulate.

4.4.2 Chloroplast microsatellite analysis

The newly designed nine cpSSR primer sets were able to detect polymorphism across all species tested, but only six sets (TeaCpSSR27, TeaCpSSR30, TeaCpSSR31, TeaCpSSR33, TeaCpSSR34, TeaCpSSR35) were able to detect variation at an intraspecific level. This is not surprising considering the fact that one objective of this analysis was to design primers that preferably amplify across a wide range of Poaceae species and thus some amount of conservation was required (Dumolin-Lapegue et al. 1997, Provan et al. 2001, Taberlet et al. 1991). However, most primers produced

amplicon lengths of 380 bp and more and thus the non-detection of variation might be due rather to the location of the microsatellite than to the location of the primers. Very poor intraspecific variation was found for marker TeaCpSSR29 which is located within the *ndhK* gene, TeaCpSSR32 which is situated in the *infA* gene and TeaCpSSR28 which is found in the IGS of *rps14-psaB*. No variation at microsatellites in coding regions was also observed for *A. thaliana* (Jakobsson et al. 2007). This low variation in microsatellites in coding regions is expected because variation in form of one to two nucleotide indels would lead to frameshifts and thus to non-functionality of the respective gene (Metzgar et al. 2000). Mutations would be more readily removed by natural selection. Marker TeaCpSSR28 is based on an interrupted microsatellite. As previously shown by Jakobsson et al. (2007) interrupted microsatellites found in *A. thaliana* were also not variable confirming the results obtained with TeaCpSSR28. The interruption of a microsatellite seems to have a stabilizing effect (Rolfsmeier & Lahue 2000). One of the main aims of this study was to design new primers for investigating intraspecific variation in the *L. perenne* chloroplast genome. Although primers TeaCpSSR29, TeaCpSSR32 and TeaCpSSR28 are not able to detect intraspecific variation, they are nevertheless able to detect interspecific variation and thus can still be useful in future studies at that level. Primer TeaCpSSR28 looks especially promising due to its ability to detect *L. multiflorum* among other *Lolium* species. This ability would be important for seed testing agencies for testing the purity of seed lots especially because the polymorphism is based on a mononucleotide repeat and thus can be easily adopted to high throughput genotyping applications. *Lolium perenne* is widely used as turf grass in the United States (Floyd & Barker 2002). Contamination of seed lots with *L. multiflorum* which has a brighter foliage colour, broader leaves and less tillering is unwanted. The detection of *L. multiflorum* within *L. perenne* seed lots is mainly done by seedling root fluorescence (SRF) tests. Roots of *L. multiflorum* seedlings grown on white filter paper fluoresce under ultraviolet light, while the majority of *L. perenne* seedlings do not fluoresce. However, the SRF trait was introgressed into *L. perenne* and breeders document now the variety fluorescence level (VFL). But studies showed that SRF and VFL are dependent on the environment the seed was grown in and thus should be an additional but not exclusive test for determining the contamination of *L. perenne* seed lots with *L. multiflorum* (Floyd & Barker 2002). Warnke et al. (2002) showed that the locus of the enzyme superoxide dismutase (*Sod-1*) can be used for distinguishing *L. perenne* from *L. multiflorum*. In further studies this locus was mapped to chromosome 7

and the trait SRF to chromosome 1 (Warnke et al. 2004). Warnke et al. (2004) showed that the SRF locus (*Pgi-2*) and the 8-h flowering characteristics on chromosome 1 can distinguish between *L. multiflorum* and *L. perenne* and the design of species specific markers from those regions is possible. However, although nuclear SSR markers are able to distinguish between *L. perenne* and *L. multiflorum* the finding of TeaCpSSR28 is, due to its easy applicability, of extremely high value.

The most informative marker was TeaCpSSR31 because it revealed nearly as much haplotype information as that detectable by using a combination of the other primers. This marker is located in the *psbE-petL* intergenic region. This region became a focus of this study because a high amount of variation was observed in this region while sequencing the chloroplast genome of *L. perenne*. Shaw et al. (2007) also detected this region when searching for highly variable non-coding regions within the chloroplast genome of different angiosperm species. High variation in non-coding regions can be due to the loss of genes (Diekmann et al. 2009, Chapter 2), but no gene loss in the *psbE-petL* intergenic spacer region had been reported so far. However, sequencing the chloroplast genome of the parasitic liverwort *Aneura mirabilis* also revealed that this region seems to be a 'divergence hotspot' as previously described by Maier et al. (1995) for the *trnG-UCC - trnT-GGU* intergenic spacer region, because it is the only region that is inverted in the parasitic *A. mirabilis* compared to the non-parasitc liverwort *Marchantia polymorpha* (Wickett et al. 2008). The *psbE-petL* region had never been the target of a chloroplast microsatellite variation analysis in Poaceae before and is thus of great value.

Marker TeaCpSSR27 is also a new primer set of considerable value. This marker is not variable at the microsatellite region but showed variation at a 44 bp long repeat that showed one and two copies respectively in four different accessions (IRL-OP-02078 (2x one copy/9x two copies), IRL-OP-02173 (3/12), IRL-OP-02269 (1/4), PI267059 (1/7)) samples. This variation was already detectable via agarose gel electrophoresis and thus this primer set is of special value because the sequencing step for the detection of variation can be omitted.

Sequence information obtained for the chloroplast genome of different species and their individuals (when polymorphisms were observed) using the new markers was submitted

to Genbank (Acc.-No.: TeaCpSSR27: HM172869 – HM 172889; TeaCpSSR 27: HM173009 - HM173027; TeaCpSSR29: HM172936 - HM172954; TeaCpSSR30: HM172890 – 172912; TeaCpSSR31: HM172955 - HM172983; TeaCpSSR32: HM172913 - HM172935; TeaCpSSR33: HM172984 - HM173008; TeaCpSSR34: HM173028 - HM173057, TeaCpSSR35: HM173058 - HM173079) and will be publicly available in January 2011.

4.4.3 Phylogenetic and haplotype analysis

Thirty-three different haplotypes were detected in this study and analysed using Bayesian inference and Parsimony approaches. These trees were rooted on *Agrostis stolonifera* and *Avena sativa* as the outgroup to all other species analysed following GPWG (2001). *Bromus erectus,* the remaining cereals, *Poa pratensis* and *Cynosurus cristatus* were resolved as outlying the majority of *Festuca* and *Lolium* accessions.

This tree showed considerably high resolution and consistency with taxonomy taking into account that most of the variation between the different haplotypes is based on microsatellites and thus size homoplasy is possible (Hale et al. 2004). However, nearly all the individuals of the different *L. perenne* accessions were grouped together. The other *Lolium* species and the *Festuca* species with exception to *Festuca pratensis* formed each a separate group. The *Festuca* group was furthermore divided into two smaller groups thus separating clearly *F. vivipara* and *F. rubra* from *F. ovina* (except for one *F. vivipara* haplotype in the *F. ovina* group). There was also evidence that the inbreeding (*L. persicum, L. remotum, L. subulatum, L. temulentum*; (Kubik et al. 1999, Balfourier et al. 2000) and outbreeding *Lolium* species could be distinguished from each other in this analysis. One exception however was the grouping of *L. subulatum*. The majority of the inbreeding *L. subulatum* individuals were grouped close to the outbreeding *Lolium* species. Unfortunately six out of the seven *L. subulatum* individuals showed haplotypes which were not clearly distinguishable from each other due to a high amount of missing data. This might have had an influence on the results.

The positions obtained for *L. hybridum* accessions are of note. *Lolium hybridum* is an interspecific highly fertile hybrid between *L. perenne* and *L. multiflorum* that occurs

naturally very frequently. In this study, *L. hybridum* individuals are grouped either within or as sister to *L. perenne,* while the *L. multiflorum* individuals are grouped as sister to *L. perenne*. This reveals *L. perenne* as maternal genome donor of *L. hybridum* and *L. multiflorum* as paternal genome donor in all these analysed accessions. It has not occurred the other way around with *L. multiflorum* as the maternal parent.

Lolium perenne individuals that lacked one copy of the repeat of the TeaCpSSR27 marker had a haplotype very similar to the inbreeding *Lolium* species (Figure 27). Although interspecific hybrids between the different outbreeding *Lolium* species are possible as mentioned above, interspecific hybrids between in- and outbreeding *Lolium* species are normally very rare due to hybridization barriers. For obtaining viable plants after reciprocal crosses between *L. temulentum* and *L. perenne*, embryo rescue techniques have to be carried out (Yamada 2001). The the results obtained with TeaCpSSR27 are more surprising because they clearly indicate that some natural hybridization occurred between the two groups containing species with different propagation systems. Assuming that this explanation is not a possible option, the only explanation left is that during ecotype sampling plants had been misidentified. Propagation of the Irish ecotype plants had been done by bulk harvesting (Connolly 2000) but is assumed to have been done in a similar way with the European ecotype plants which seeds derived from different gene banks. Thus seed from only one wrong identified plant could lead to a greater contamination. Also mix-ups during the DNA extraction step or the PCR set-ups can not be completely excluded.

The grouping of *Festuca pratensis* within the *Lolium* accessions is also of note. It is positioned relatively far away from the other *Festuca* species. *Festuca* species can be grouped into two major categories – fine-leaved and broad-leaved (Torrecilla & Catalán 2002). In phylogenetic analyses the genus *Lolium* is found within the broad-leaved *Festuca* group (Torrecilla & Catalán 2002, Catalán et al. 2004). *Festuca pratensis* belongs to the broad-leaved category while *F. ovina, F. rubra* and *F. vivipara* belong to the fine-leaved category. This is in accordance to the extensive *Lolium* diversity study of McGrath et al. (2007) where, on a phylogenetic tree, *F. pratensis* grouped closer to the *Lolium* species and where the fine-leaved *Festuca* species were clearly resolved as an outlying group. Thus it is not surprising that *F. pratensis* as only representative of the broad-leaved fescues grouped closely to *Lolium* in this study. That it is grouped within

Lolium might either indicate that this cultivar was derived from an interspecifc hybrid between *Lolium* and *Festuca* species, or that it should be classified as *Lolium* and not *Festuca*. The former hypothesis is most likely and with chloroplast DNA a species can have the plastid type of one parental group and a predominantly nuclear DNA component from another group. No nuclear DNA markers were used in this study to verify this hypothesis but the results confirm that a 'chloroplast capture' event has happened in this taxon. However, the accession used in this study is an ecotype collected in the north of New Zealand above Auckland. *Festuca pratensis* is a very rare species in New Zealand and this accession derived from one of only two known populations. It most likely derived from a *F. pratensis* seed lot imported from the United Kingdom into New Zealand in the early 1900s which was sown in that specific region. Sequence alignments carried out in the company PGG Wrightson revealed a good alignment of accession Northland with other *F. pratensis* accessions (personal communication with Alan Stewart). Nevertheless, it would be still possible that it derived from a backcross with *L. multiflorum*. However, this specific line of *F. pratensis* was used for endophyte content analyses. The endophyte is maternally inherited (as the chloroplast genome) and was identified as derived from *F. pratensis* in this specific accession. Thus hybridization between *L. multiflorum* and *F. pratensis* seems to be unlikely. In 1993 Stephen J. Darbyshire suggested to rename parts of the *Festuca* genus. While the fine-leaved fescues are considered to be true fescues and represent a separate branch in phylogenetic analyses, the broad-leaved fescues such as *F. pratensis* or *F. arundinacea* generally group within the genus *Lolium* as was also observed in this study. Thus Darbyshire (1993) proposed to rename *F. pratensis* into *Lolium pratense*. This study also shows that *F. pratensis* groups with species of genus *Lolium* and hence confirms previous observations. Therefore the grouping of *F. pratensis* within the genus *Lolium* can more likely be explained with a misclassification of this species rather than hybridization events between species of two different genera.

Focussing only on the *L. perenne* group revealed two major haplotypes and several minor haplotypes. While all ecotypes and cultivars are based on at least one of the major haplotypes, only three of the cultivars had some of the minor haplotypes. Although the number of analysed individuals per cultivar is very small (with not more than eight individuals per cultivar), it still represents an obvious decrease in diversity within the *L. perenne* ecotypes.

4.4.4 Locus-by-locus AMOVA

For the locus-by-locus AMOVA the different accessions were divided into six different groups (Table 11). When all six groups were included in the study the highest variation was found among groups, which is not surprising, taken into account that some of the groups represented different species. When only *Lolium* and *Festuca* species were analysed the results still showed that the highest variation was found among groups; however the population fixation index F_{ST} and the variation within the different groups decreases while the variation found within the different populations increases. Focussing only on the *Lolium* species reveals that approximately 34% of the variation found can be attributed to variation within the populations. And when, finally, focusing only on *L. perenne* accessions, one can see a very low fixation index of 0.27 and within-population variation accounting for 73.48% of the total variation.

Analysing the three different *L. perenne* groups separately from each other reveals a very low fixation index of 0.10 for the Irish ecotypes while the within population variation is around 90%. This is in very strong contrast to the F_{ST} for the cultivars and European ecotypes which is 0.50. However, it is in agreement with other studies using nuclear markers on *L. perenne* and other outbreeding grass species. Ubi et al. (2003), for example, examined the genetic diversity of rhodesgrass (*Chloris gayana*) using AFLP markers. Analysing 15 individuals of 13 different cultivars they obtained within population variation values of around 82% (Ubi et al. 2003). Another study based on *Dactylis glomerata* and using RFLP markers analysed the genetic diversity of drought responsive genes in six ecotype populations (Trejo-Calzada & O'Connell 2005). The within population variation in this study was on average 81%. Fjellheim and Rognli (2005) analysed the genetic diversity of thirteen *Festuca pratensis* cultivars that derived from Denmark, Iceland, Norway and Sweden via AFLP markers. Cultivars were selected to represent old and new varieties from each country. The within cultivar variation was again around 80%. Furthermore did their study show that the level of genetic diversity was in new cultivars as large as in older cultivars and a decrease of genetic diversity caused by breeding *F. pratensis* was not observed. Also in *L. perenne* genetic diversity studies have been carried out using nuclear markers. Guthridge et al. (2001), for example, analysed genetic diversity using two different sets of plant material. The first set consisted of four cultivars with different breeding background in

regards to genetic base and numbers of parents and here a within population variation of 89% was observed. The second set was based on three cultivars which derived from a low number of parents and the within population variation was 79%. Surprisingly, the amount of variation found within the cultivars and European ecotypes used in this study was rather small compared to the results achieved by other research groups. This could have been caused by the low number of individuals per accession used. Most of the aforementioned studies used at least 15 individuals per cultivar. The amount of individuals per accession used in this study was on average eight for the Irish ecotypes, six for the European and five for the cultivars. Thus the results could be strongly biased due to the sampling. Another explanation for the different results could also be the marker system used in this study. The chloroplast genome is generally rather conserved and although the markers chosen for this study were located in chloroplast genome regions with a considerable high degree of variation, the variation that can be observed in the nuclear genome is much higher. This assumption is also confirmed by results from McGrath (2007, 2008) who used also chloroplast microsatellite markers for assessing the genetic diversity of *L. perenne* ecotypes and cultivars. While the within population variation of Irish ecotypes was 82%, the variation among populations was 18%. The variation within populations of European ecotypes accounted for only 61% compared to 39% of variation among populations (McGrath 2007, 2008). Although the values differ slightly from the results obtained in this study, which might be due to different sample size (and different marker system), the trend is the same. This result is not surprising taking into account that Ireland is an island and thus the possibility of pollen and seed mediated gene flow is limited to a much smaller area than on the European continent. The distances between populations is in general less and hence differentiation of populations is lower. Genetic diversity analyses are generally influenced by factors such as the breeding background of the accessions analysed, sampling strategies used for collecting ecotypes, the marker system applied to the samples and the number of accessions and individuals per accession used. However, the factor breeding background should not have any major influence on the comparison of data from this study and data from the other studies mentioned before because all studies comprised ecotypes and synthetic cultivars independent on the species analysed. Only six polymorphic loci were found within the *L. perenne* cultivars, while there were 15 within the European ecotypes and more than 16 within the Irish ecotypes. This could already indicate a decrease in the size of the genetic pool that is considered in breeding

L. perenne cultivars. However, this result could also be due to a bias through sampling. Only a small amount of individuals per accession was used and the focus on this study was on Irish ecotypes including them preferentially in this study. Nevertheless, does this result still highlight the diversity of the Irish ecotype collection and the value of it for future breeding programmes.

4.4.5 Genetic distance analysis

The genetic distance analysis was firstly carried out with all groups and sequence data and then with only the *L. perenne* accession groups and sequence data from TeaCpSSR27 and TeaCpSSR31. The genetic distances between the *Lolium* species and species from the outgroup were very large. This result was expected because of the distant relationship of the species. However, also within the *Lolium* populations some of the distance values were considerably high. Comparing the results from the two different analyses reveals that in general the observed distances were reduced when the data set was reduced. This is very likely due to a reduction of missing data which might have some influence on the analysis. Marker TeaCpSSR31 alone detects an equivalent amount of variation on its own as that detected with a combination of the different markers and therefore it can be expected that nearly all the possible variation can be detected using only this marker.

However, the genetic distances are possibly overestimated due to the small sample size of less than eight individuals for some accessions. IRL-OP-02337, for example, was represented in this study by three individuals that revealed in total three different haplotypes. IRL-OP-02491 was represented by eight individuals among whom the same three different haplotypes had been observed as for IRL-OP-02337. However the genetic distances are 1.33 for IRL-OP-02337 and 0.96 for IRL-OP-02491. For calculating the genetic distances Arlequin 3.11 (Excoffier et al. 2005) considers the number of the distinct haplotypes in the two populations, the frequency with that one haplotype occurs in a population and the number of nucleotide differences between the different haplotypes. While the number of haplotypes and the number of nucleotide differences is independent from the population size, is the frequency with which a haplotype occurs depending on the population size. For IRL-OP-02337 the frequency of

each haplotype was 0.33 because only three individuals were analysed. However, it is very unlikely that the frequency in the whole population for all three haplotypes is 0.33 and thus the genetic distance is very likely overestimated.

Nevertheless the genetic distance analyses carried out in this study provide an idea about the diversity within the data set. Distances within European ecotypes and cultivars were less than within Irish ecotypes, confirming earlier results. Some of the European accessions were significantly distant to the Irish ecotypes (e.g. PI267059, W69286 and W61125) and could prove to be of interest for introduction into the Irish plant breeding material to increase the genetic diversity.

4.5 Conclusion

This study revealed that chloroplast simple sequence repeats can be characterized according to the species they are found in. For example, Pooideae species had a higher number of cpSSRs that were mainly based on the nucleotide A than Ehrhartoideae (*Oryza*) species which cpSSRs were mainly based on the nucleotide T. Most large microsatellites were found in the LSC region of the genome. However, the main objectives of this study were to find new Poaceae chloroplast DNA markers using information about the location of variable regions previously obtained when sequencing the *L. perenne* chloroplast genome. Thus, nine new primer sets were designed which all successfully amplified polymorphic regions either in or outside (as indels or SNPs) the target SSRs. The most informative markers were TeaCpSSR31 (*psbE-petL*), because it alone detected nearly all the variation otherwise only found with a combination of primers, and TeaCpSSR27 which enabled the distinction of haplotypes solely via agarose gel electrophoresis without the need for sequencing or more accurate sizing methods. However, TeaCpSSR28 which is not able to detect variation within *L. perenne* was found to be able to particularly informative because it could detect *L. multiflorum* within other *Lolium* species. This marker although based on a microsatellite region might prove specifically useful in barcoding approaches when commonly used broadly applied universal markers (such ast *matK* and *rbcL*) are not able to detect variation within species of the same genus.

The new markers were tested on a small set of different *Lolium* and *Festuca* accessions and were able to detect the diversity within these accessions. Although the breeding history of *L. perenne* is rather young (approximately 90 years, Wilkins 1991) this study shows that existing cultivars derive from a narrow genetic pool. 15 to 16 polymorphic loci were found within the ecotype accessions, but only six were found within the cultivars. An increase of diversity or at least a monitoring of the diversity within the plant breeding material should therefore be considered. Plant breeders normally try to avoid the introduction of new diverse material because wild germplasm rarely shows a superior performance to the established breeding material. However, thorough studies in tomato revealed that wild tomatoes with non-desirable phenotypes (eg. small green fruits) can nevertheless contain alleles that can improve desirable traits in existing cultivars. Bernacchi et al. (1998) screened wild tomatoes for quantitative trait loci (QTL) and introduced these regions via backcrossing into adapted plant breeding material. Thus breeding lines were created that had a higher performance of up to 48% compared to existing cultivars. QTL maps for different traits in *L. perenne* exist (e.g. for biomass yield (Anhalt et al. 2009), vernalization response (Jensen et al. 2005), herbage quality (Cogan et al. 2005) and a controlled increase in diversity by using a QTL assisted approach might be already possible. Especially markers TeaCpSSR27 and TeaCpSSR31 will prove useful for monitoring this process from a cytoplasmic level and for monitoring this increased diversity in the future.

Chapter 5

First insights into the mitochondrial genome of *Lolium perenne* L.

"For the isolation of systemin, the first plant peptide messenger detected, 28 kg of tomato leaves was required and 30000 tomato plants had to be processed to produce 1 µg of the peptide."

Pearce et al. (1991);

Cited in Drescher et al. (2002)

5.1 Introduction

Only shortly after Ris and Plaut (1962) had found evidence for the presence of DNA within chloroplasts, Nass and Nass (1963) reported the observation of fibres within mitochondria of chick embryos via electron microscopy and drew the conclusion that these fibres contain DNA. Schatz et al. (1964) delivered the biochemical evidence that the mitochondria of *Saccharomyces cerevisiae* (baker yeast) contain a significant quantity of extranuclear DNA. Shortly afterwards, Kislev et al. (1965), using *Beta vulgaris* var. cicla (Swiss chard), obtained the first biochemical evidence that plant mitochondria also contain DNA. The first mitochondrial DNA sequences became publicly available in 1979 such as a transfer RNA gene of the fungus *Neurospora crassa* (Heckman et al. 1979) and a region surrounding the origin of mitochondrial DNA replication in humans (Crews et al. 1979). However, it was not until 1981 that the first complete mitochondrial genome, the one of humans, became available (Anderson et al. 1981) and was determined to be 16,569 bp in length. Since the publication of the human genome many more animal mitochondrial genomes have been sequenced and by October 2009 1,778 animal mitochondrial genomes had been published. The genome sizes varied from 11 to 43 kb, with the majority of genomes being between 16 to 17 kb long as observed for the human mitochondrial genome (http://www.ncbi.nlm.nih.gov/genomes/ OrganelleResource.cgi?taxid=33208). Most of the genomes contained 13 protein coding genes, 22 transfer and two ribosomal RNA

genes. The order of these genes is generally conserved within major groups (Boore 1999).

5.1.1 Plant mitochondrial genome features

First plant mitochondrial DNA sequences became available during the mid 1980s such as the 18S ribosomal RNA gene of maize (Chao et al. 1984). The first complete plant mitochondrial genome sequenced was from the liverwort *Marchantia polymorpha* (Oda et al. 1992) - roughly ten years after the first animal mitochondrial genome. It measured 186,609 bp and thus was more than eleven times bigger than most animal mitochondrial genomes. In 1997, the first complete mitochondrial genome of a higher land plant species, *Arabidopsis thaliana,* became available (Unseld et al. 1997) and was 366,924 bp in length. Plant mitochondrial genomes vary widely in their size. The *Brassica napus* mitochondrial genome with 222 kb is the smallest detected so far (Handa 2003). The largest mitochondrial genome found, but as yet unsequenced, belongs to *Cucumis melo* (muskmelon) and consists of 2,400 kb (Ward et al. 1981). It is thus more than 100 times larger than average animal mitochondrial genomes. In November 2009 24 mitochondrial genomes of different land plant species had been sequenced (Figure 28) and they varied in size from ~42 kb to ~770 kb (Figure 29).

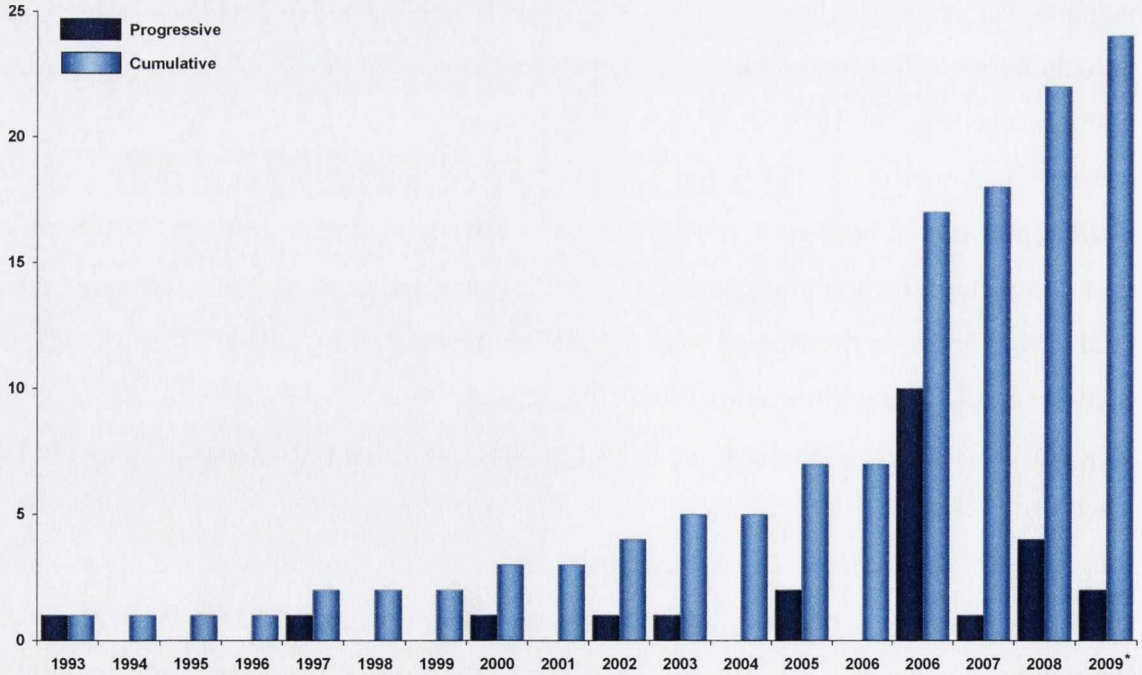

Figure 28 Number of publicly available mitochondrial genome sequences from moss and higher plant species in October 2009 (www.ncbi.nlm.nih.gov). *preliminary data

Mitochondrial genomes are generally illustrated in the form of circular maps which are named mastercircles or masterchromosomes. The genome of plant mitochondria is very rich in repetitive structures and in all plants analysed to date large repeats of direct or inverted orientation were observed. These repeats are thought to be involved in intramolecular recombination. Recombination between inverted repeats results in sequence isomerization (flip – flop) as for example for the inverted repeat in chloroplast genomes. Recombination between direct repeats leads to loop-outs that can increase the amount of subgenomic circles (Backert et al. 1997). However, electron-microscopic (EM) examinations showed that the mitochondrial genome structure is much more complicated as implied by the model of master- and subcircles. Besides the regular observation of circular molecules also linear molecules were found. However, circular molecules found are rarely of the size expected for the master- and subgenomic circles. Instead most of them are significantly smaller than the complete genome size or the size of the predicted subgenomic circles. Furthermore, linear molecules of all sizes can be found, even molecules that are much longer than the detected genome size. Besides the general mixture of linear and circular molecules also much more complex structures such as interlocked circles, branched linear molecules or molecules with a rosette-like structure have been observed. While the majority of mitochondrial DNA will be double-stranded also single stranded molecules of various forms were already found (reviewed in Backert et al. 1997).

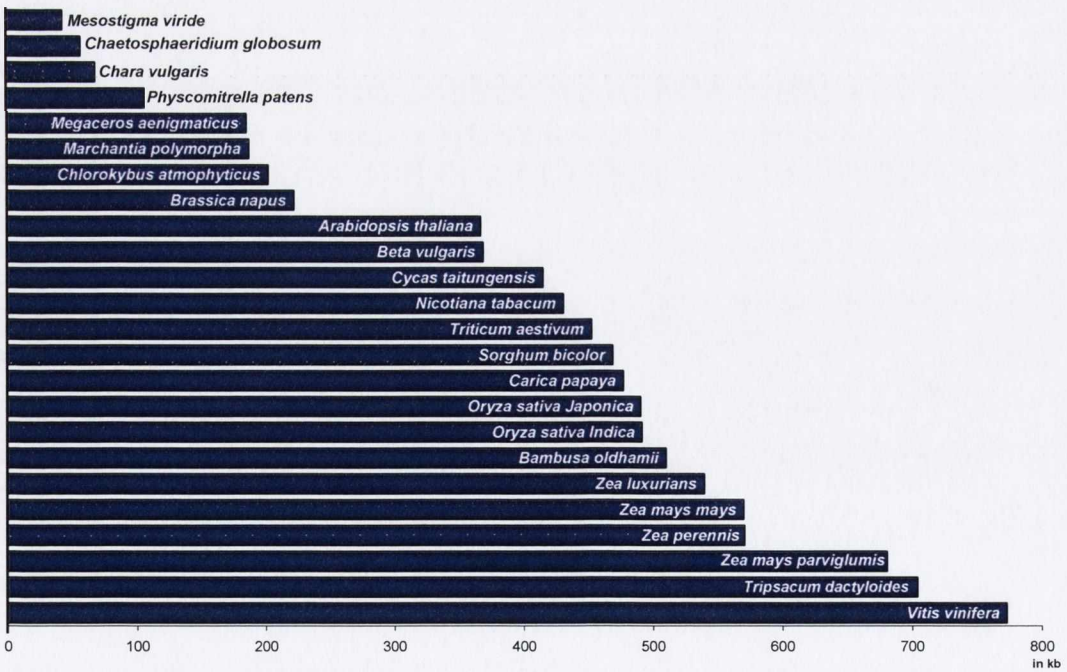

Figure 29 Lengths of mitochondrial genomes from all vascular plant and moss species publicly available in October 2009 (www.ncbi.nlm.nih.gov).

Despite their great variation in shape, plant mitochondrial genomes are rather conserved regarding the genes they encode. In most angiosperm species, 50 to 60 genes of three different types, protein-coding, tRNA and rRNA, can be found. In general several genes are always present including the three ribosomal RNA genes, genes of the *nad*-complex (complex I), the *cob*-complex (complex III), the *cox*-complex (complex IV), the *ccm*-complex, *atp*-complex (complex V) as well as the maturase and translocator gene (Adams & Palmer 2003, Kubo & Mikami 2007). The highest gene content variation can normally be observed for the ribosomal protein-coding genes (Adams & Palmer 2003). Mitochondrial genes of angiosperms follow generally the universal genetic code, however, a few exceptions were observed regarding translation initiation codons (Kubo & Mikami 2007). The gene sequences of mitochondrial genomes are highly conserved which generally disables their use for example in phylogenetic studies (but see Chapter 6). The GC-content of mitochondrial genomes is rather constant between the different species sequenced so far, varying from 43 to 46% in vascular plants.

Despite the rather conserved gene content, gene order is highly variable. So far no mitochondrial genome has been sequenced that has the exact same gene order as another one. Even the closely related varieties *Oryza sativa* indica-group and *Oryza sativa* japonica-group differ in their gene order. A comprehensive study on the *Triticum aestivum* mitochondrial genome showed that only six shared gene clusters could be found when comparing the gene order of *Triticum aestivum* with *Oryza sativa* and *Zea mays*, respectively. One gene cluster comprised five genes (*ccmFN*, *rps1*, *matR*, *nad1e*, *nad5c*) and five gene clusters contained each two genes (*rps13-nad1bc*; *rrn18-rrn5*; *rps3-rpl16*; *nad9-nad2cde, nad3-rps12*) (Ogihara et al. 2005). For eight plant mitochondrial genes, group II introns were detected which show *cis*- and *trans* splicing (*cis*: *nad1*, *nad2*, *nad4*, *nad5*, *nad7*, *ccmFC*, *rps3*, *cox2*; *trans*: *nad1*, *nad2*, *nad5*). The intergenic regions of plant mitochondrial genomes comprise 70 to 90% of the total genome and are responsible for most of the observed genome size variation (Kubo & Mikami 2007). Due to the non-conserved gene order, these intergenic regions are highly variable. Around 55 to 80% of newly sequenced mitochondrial genome sequences do not show any homology to sequences published in Genbank (Sugiyama et al. 2005).

In all sequenced plant mitochondrial genomes a complete set of tRNA genes that is able to read all the codons has never been found. Generally 10 to 12 tRNA genes are native

to the mitochondrial genome; around 9 tRNA genes are very likely of chloroplast origin and approximately 10 further tRNA genes need to be imported from the cytosol (Kubo & Mikami 2007).

In plant mitochondrial genomes RNA editing is an essential process for gene expression and generally more than 400 editing sites can be found (Kubo & Mikami 2007). Hence, although mitochondrial genomes encode fewer genes than chloroplast genomes, they have 10 times more editing sites.

5.1.2 Horizontal and intracellular gene transfer

The term 'horizontal gene transfer (HGT)' (=lateral gene transfer) describes the movement of genetic information across normal mating barriers, between more or less distantly related organisms. It has to be distinguished from the vertical gene transfer which occurs when genes are transferred from the parents to the offspring (Keeling & Palmer 2008).

Using mitochondrial genome sequences for phylogenetic analyses can reveal contradictionary results to analyses based on sequences of different origins especially when using genes that appear in duplicate or even triplicate in the mitochondrial genome. *Plantago* species, for example, contain two *atp1* genes. One is an intact copy of the gene being inherited by normal vertical gene transfer, while the other *atp1* copy contains some mutation frameshifts and is not functional. Phylogenetic analyses of the pseudo-*atp1* gene revealed that one *Plantago* group of closely related species is grouped with species of the genus *Bartsia* and another group of closely related *Plantago* species is grouped with species of the genus *Cuscuta*. Both genera, contain parasitic species and it is very likely that HGT had taken place via direct contact of the parasitic with the host plant (Mower & Palmer 2006). *Amborella* contains three copies of the *nad5* gene. Phylogenetic analyses show that one copy is native to *Amborella* but two are of foreign origins, grouping with angiosperms and mosses, respectively. It is not known how the transfer of the foreign *nad5* genes took place, however, studies showed that in most cases a direct contact between donor and recipient plant is involved (Keeling & Palmer 2008).

Besides the large, complex mitochondrial genome plant mitochondria can also contain extrachromosomal autonomously replicating linear and circular DNA molecules (plasmids) of high abundance. The circular molecules comprise generally around 1 to 2 kb but no open reading frames of significant length. The linear molecules range in size from 2 to 13 kb, with terminal inverted repeats and long expressed open reading frames. It is believed that linear mitochondrial plasmids have a viral origin due to structure similarities between plasmids and DNA viruses. However, only very few plant species contain linear plasmids and while for example *Brassica rapa* and *B. napus* both contain an 11.6-kb linear plasmid other Brassicacea-species do not have this plasmid. Therefore it is assumed that these linear plasmids were also acquired via HGT probably from a fungal phytopathogen to a plant species (reviewed in Handa 2008).

Stern and Lonsdale (1982) reported a chloroplast derived 12 kb DNA fragment in the mitochondrial genome of maize. This was the first observation that DNA can, not only be transferred horizontally or vertically, but also intracellularly between different cell compartments (Timmis et al. 2004). Shortly after this discovery, the presence of chloroplast and mitochondrial DNA fragments in the nuclear genome was reported and the term 'promiscuous DNA' introduced for the DNA fragments that originated in other genomes. Intracellular gene transfer (IGT) is an ongoing process and experiments in tobacco showed that it can happen with considerable frequency (Sheppard et al. 2008). IGT is part of the endosymbiotic gene transfer and happens mainly from the chloroplasts and mitochondria to the nucleus. Nowadays plant mitochondrial genomes contain around 50 to 60 genes, but α-Proteobacteria, the mitochondrial ancestors, contained up to three times more genes. Mitochondria (and also chloroplasts) need several hundred proteins for proper functioning but their genome encodes only few. The majority of the genes have been transferred from the mitochondria or chloroplasts to the nucleus from where the essential proteins are now transferred back to the mitochondria and chloroplasts, respectively (Adams & Palmer 2003, Keeling 2004). Protein coding genes that were frequently lost in the mitochondrial genome of angiosperm species during evolution belong to the succinate dehydrogenase group (complex II) and to the group of ribosomal proteins (Adams & Palmer 2003). But not only single genes were transferred to the nucleus. The complete mitochondrial genome of *Arabidopsis thaliana* was found in a region close to the centromere of chromosome 2 (Stupar et al. 2001). And a nearly complete copy of the choroplast genome was found on rice chromosome

10 (The Rice Chromosome 10 Sequencing Consortium 2003). Although IGT occurs more frequently from the two endosymbiotic genomes to the host genome, occasionally it can be observed from the nucleus to mitochondrion or from the chloroplast to the mitochondrion. Plant mitochondria contain between 1.1–6.3% chloroplast derived and 0.1-13.4% nuclear derived promiscuous DNA (Kubo & Mikami 2007). While the nuclear derived DNA is mainly of retrotransposonical nature, chloroplast DNA comprises protein-coding as well as tRNA genes. The chloroplast derived protein-coding genes are generally, due to frame shift mutations, disfunctional; however the tRNA genes can build a clover like structure and are therefore very likely functional considering the fact that the mitochondrial genome itself does not code for all essential tRNA genes.

5.1.3 Cytoplasmic male sterility

The term 'cytoplasmic male sterility (CMS)' describes an inherited condition under which a plant is not able to produce functional pollen. CMS was discovered at the beginning of the 19th century. One of the first observations of CMS happened in a flax breeding program. Crosses of plants of two different flax strains obtained male sterile F2 plants when crossed in one direction but completely fertile plants when crossed in the other implying that extranuclear inheritance had taken place (Bateson & Gardner 1921). By 1972, CMS had been reported for more than 140 angiosperm species (Laser & Lersten 1972). It is maternally inherited and associated with unusual chimeric open reading frames in the mitochondrial genome (Schnable & Wise 1998). These chimeric open reading frames code for cytotoxic proteins that inhibit the development of functional pollen (Wang et al. 2006). The occurrence of CMS in natural plant populations is a disadvantage; however, it is a huge advantage for conventional plant breeding programs where it is used widely for the production of hybrid seeds (Eckardt 2006). Fertility can be restored by nuclear genes that suppress or alter the expression of CMS-linked genes (Schnable & Wiese 1998). Therefore crosses in hybrid programs can be controlled using a CMS-line as mother and a male line that contains restorer genes. Although the resulting F1 hybrid seed will carry the CMS cytoplasm the resulting plants will be fertile due to nuclear restoration genes. The application of CMS in plant

breeding schemes reduced breeding costs dramatically because mechanical emasculation can be omitted.

CMS is also used in breeding schemes for *L. perenne*, however it is much more complicated to utilize than for example in maize. *Lolium perenne* has a self-incompatibility (SI) mechanism making it very difficult to obtain inbred lines for hybrid seed production. Due to the self-incompatibility mechanism the breeding of *L. perenne* is based on population breeding schemes specifically on the creation of so called synthetic cultivars that consist of a mixture of many genotypes. Nevertheless, the SI mechanism of *L. perenne* occasionally breaks down so that it is possible to generate inbred lines (Thorogood & Hayward 1992). But in general inbred lines can not be generated efficiently enough to base the complete breeding system of *L. perenne* on it.

5.1.4 Aims and Objectives

The majority of published plant mitochondrial genomes are from monocotyledonous species, specifically only from Poaceae, However, no mitochondrial genomes of a pure forage grass species have been sequenced so far. The aims of this study were therefore to sequence the entire mitochondrial genome of *Lolium perenne* to better understand mitochondrial genome evolution in grasses and to allow application of the new information to plant breeding. Although *L. perenne* is a forage grass and the interest lies especially in breeding varieties with high biomass yield rather than seed yield, the application of CMS lines in the breeding process is also of high interest. The induced heterosis effect by the usage of the CMS system results not only in higher seed production but also in more vigorous growth and thus in higher biomass yield. Several studies on CMS in *L. perenne* have been carried out (Rouwendal et al. 1992, Kiang et al. 1993, Kiang & Kavanagh 1996, McDermott et al. 2008). Rouwendal et al. (1992) suspected that the mitochondrial genes *atp6* and *coxI* were involved in CMS in *L. perenne*. Kiang et al. (1993), Kiang and Kavanagh (1996) as well as McDermott et al. (2008) investigated the involvement of a mitochondrial linear plasmid in *L. perenne* CMS. However, none of those studies focussed on the complete mitochondrial genome. Thus the availability of the complete mitochondrial genome would be a huge advantage.

Therefore, the detailed aims and objectives of this chapter were to isolate pure mitochondrial DNA of *L. perenne* suitable for sequencing, to assemble and annotate the sequenced mitochondrial genome and to analyse it in comparison to other published mitochondrial genomes of Poaceae species. The mitochondrial sequence information should be also used to investigate intracellular and horizontal gene transfer. Furthermore, the availability of the complete mitochondrial genome sequence of *L. perenne* will enable the investigation of the linear plasmid sequence detected by McDermott et al. (2008) in a CMS line, in the male fertile *L. perenne* cv. Shandon.

5.2 Material and Methods

Several protocols for the isolation of plant mitochondrial DNA were tested (Honda et al. 1966, Pring & Levings 1978, Mulligan et al. 1988, adaptation of Lang & Burger 2007 and Diekmann et al. 2008 for plant mitochondrial DNA), but the isolation of high quality mitochondrial DNA proved to be difficult and even more complicated than the isolation of chloroplast DNA (Chapter 2). One major difficulty was to monitor the progress of the extraction, as done for the chloroplasts, due to the colourless nature of the mitochondria. Several dyes for staining mitochondria were tested; however, these dyes (MitoTracker® Red CMXRos, MitoTracker® Orange CMTMRos, MitoTracker® Deep Red FM; Invitrogen™, Carlsbad, California, USA) were mainly designed for animal mitochondria and also seemed to stain the cell fibres, making it hard to distinguish the cell debris from the mitochondria. Therefore the monitoring of the isolation progress via light microscopy was finally omitted. Chloroplast DNA of *Angiopteris evecta* was retrieved by using a flow sorting approach (Roper et al. 2007) that sorted the three organelles (nucleus, chloroplast and mitochondrion) from each other and resulted in a pure chloroplast fraction. This approach was adapted for the isolation of mitochondria from *L. perenne*. It is generally based on staining the three DNA containing organelles with different fluorescent colours. Chloroplasts are by nature stained green, nuclei were stained using 4'-6-Diamidino-2-phenylindole (DAPI, Sigma-Aldrich, Saint Louis, Missouri, USA) which fluoresces blue and mitochondria were stained red with MitoTracker® Orange CMTMRos (Invitrogen™, Carlsbad, California, USA). Although a sample was retrieved from this approach, and sent off for

test sequencing, the results were not sufficiently good to allow the sequencing of the entire mitochondrial genome.

A second difficulty, apart from the isolation of mitochondrial DNA, was to evaluate the quality and purity of isolated DNA. It was expected that pure mitochondrial DNA would form a clear restriction digest banding pattern as observed for various other plant species (e.g. Kiang et al. 1994). Hence, DNA samples that showed such a banding pattern were sent away to a sequencing company for test-sequencing. The sequencing company sequenced 48 clones and blasted the resulting sequences against their internal databank of published sequences. Three samples had been sent away of which the first showed five mitochondrial hits, the second eight and the third which was obtained via flow sorting, none at all. Next to the few mitochondrial hits that had been observed in the two other samples some chloroplast DNA hits were observed as well as nuclear DNA hits. However around half of the sequences from the test-sequencing did not show any similarity to any sequence ever published in Genbank. Due to horizontal gene transfer it is possible to obtain sequences that are similar to chloroplast DNA or nuclei DNA and researchers that sequenced other plant mitochondrial genomes also frequently observed a very high amount of 'no hits' (up to 80%). However the low amount of mitochondrial hits implied that the DNA samples might not be pure mitochondrial DNA and be contaminated with other organellar DNA and thus were not used for further sequencing.

Finally, the *L. perenne* mitochondrial DNA was isolated following the method described by Kiang (1992) and Kiang et al. (1993). This method (see below) included a DNase I step which was supposed to digest the DNA from broken chloroplast and nuclei while the mitochondria were still intact.

5.2.1 Plant material

During the time this project was carried out a similar project was carried out in the Smurfit Institute of Genetics (Trinity College Dublin) dealing with sequencing mitochondrial DNA of *Lolium perenne* cv. Cashel. Therefore, to not interfere with the ongoing work of the Smurfit Institute it was decided to isolate mitochondria from a

CMS *L. perenne* line. Thus it would have enabled the direct comparison of a male fertile *L. perenne* line (cv. Cashel) with a male sterile line (a CMS line). Unfortunately it was not possible to yield enough leaf material from any CMS line tested due to germination issues and therefore the male fertile *L. perenne* cv. Shandon was chosen. Approximately 250 g seeds of *L. perenne* cv. Shandon were sterilized in 10% Sodium hypochlorite for one hour and washed with autoclaved tap water three times. They were then thinly spread on moistened filter paper in deep plastic trays. The trays were covered in black foil and kept in a climate chamber at 22 °C. They were watered using autoclaved tap water. The grown seedlings were harvested after approximately 14 days when they had a length of circa 4 cm.

5.2.2 Mitochondrial DNA isolation

All steps were carried out at 4 °C if not otherwise stated. 200 g of leaves of the 7-day old etiolated seedlings were homogenised in 1.5 l extraction buffer (0.44 M Sucrose, 50 mM Tris-HCl pH 8.0, 3 mM EDTA pH 8.0, 0.5% bovine serum albumin, 1 mM DTT) using mortar and pestle. The homogenate was filtered through four layers of Miracloth and centrifuged for 30 min (5,000 *g*, SLA3000) to remove cell debris, nuclei and chloroplasts. The supernatant was centrifuged for 35 min (16,000 *g*, SLA3000) to pellet the mitochondria. The organelles were resuspended in 20 ml DNase buffer (0.44 M sucrose, 50 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$) using a soft paintbrush. Suspension was divided up into two SS34 tubes and 10 mg of DNase I was added to each tube. The digestion was carried out on ice for one hour. For terminating the digestion, EDTA was added to a final concentration of 25 mM by underlaying the organelle suspension with 20 ml washbuffer (0.6 M sucrose, 25 mM EDTA pH 8.0, 50 mM Tris-HCl pH 8.0) per tube. The tubes were centrifuged at 16,000 *g* for 35 min to pellet the mitochondria again. The pellets were once more resuspended using 10 ml of DNase buffer (without DNase I !) and underlayed with 20 ml washbuffer per tube. Both tubes were again centrifuged at 16,000 *g* for 35 min. The mitochondria were then lysed resuspending each pellet in 2 ml NTE buffer (10mM NaCl, 10mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0) and adding 1% SDS. The solution was incubated in a waterbath at 37 °C for 45 min. Purification was done by five phenol-chloroform extractions (for faster phase building the tubes were centrifuged for 10 min at 12,000 rpm in the Hettich microcentrifuge) and the DNA was precipitated with two volumes of 99% ethanol at –

20 °C overnight. DNA was pelleted by centrifuging (45 min, 13,000 rpm, Hettich), washed twice in 70 % ethanol, dried and resuspended in 300 µl 1x TE.

To see if the mitochondrial DNA isolation was successful 20 µl of the sample was mixed with 3 µl loading buffer and run on a 0.8% agarose 0.5X TBE gel containing 0.6 µl ethidium bromide (10mg/ml) at 60 V for an hour (Figure 30a). For RNA free samples, 2 µl/100 µl sample of RNase were added and digestion was done for 30 min at room temperature. To enable sequencing the sample had to be whole genome amplified as previously described for the chloroplast DNA (Chapter 2) using the GenomePhi DNA Amplification Kit (Amersham Biosciences, Prod. no.: 25-6600-01) following the manufacturer's instructions. For a further control of the isolation and amplification success 2 µl of each amplified sample were digested with the restriction enzyme HindIII (2 µl amplified sample, 1 µl enzyme, 2 µl buffer, 15 µl ddH_2O) over night. The results of the restriction digest were visualized on a 0.5 x TBE 1% agarose gel that was run for 2 hours at 120 V (Figure 30b).

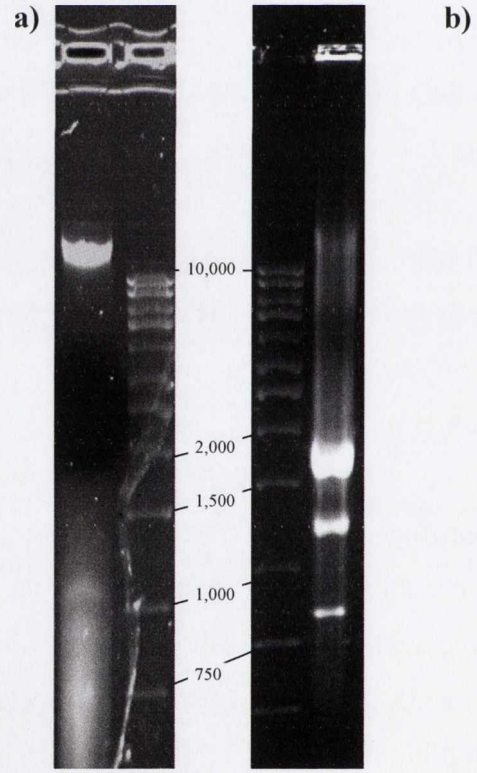

Figure 30 Undigested (a) and restriction enzyme digested (b) *Lolium perenne* mitochondrial DNA. Panel a): 20 µl DNA sample (prior to amplification) were mixed with 3 µl loading buffer and run on a 0.8% agarose 0.5X TBE gel containing 0.6 µl ethidium bromide (10mg/ml) at 60 V for an hour. Panel b): 20 µl of the whole genome amplified sample were digested with HindIII (2 µl enzyme, 3 µl buffer, 5 µl ddH_2O) and run on a 300 ml 0.8% agarose 0.5X TBE gel containing 3 µl ethidium bromide (10mg/ml) at ~180 V for 3.5 hours. Size standard: 5 µl 1 kb DNA ladder (Promega, Wisconsin, USA).

5.2.3 Sequencing the mitochondrial genome of *Lolium perenne*

The mitochondrial genome of *L. perenne* was expected to be around 450 to 500 kb large. Because of the general complicated nature of mitochondrial genomes regarding gene order and repetetive structures of mitochondrial genomes in general (as well as the improvement of sequencing technologies) a hybrid sequencing approach was chosen. This approach consisted of two different sequencing methods, the conventional Sanger sequencing technique (Chapter 2, Sanger et al. 1977, Sanger et al. 1980) and the more recently developed 454 sequencing technique.

Sanger sequencing was the most frequently method used in the last 30 years. However, Sanger sequencing is rather expensive and very time consuming and in recent years the so called "next generation" sequencing techniques have been developed. The first of these next generation sequencing techniques was introduced by 454 Life Science, a company member of Roche, in 2005. 454 sequencing is much cheaper and faster than Sanger. It can process a broad range of DNA samples such as PCR products, genomic DNA or cDNA. Long DNA sequences as for example genomes are mechanically sheared into 300 to 800 bp fragments. Adapters are added to each resulting fragment, which are needed during purification, amplification and sequencing. Single stranded DNA (sstDNA) fragments with adaptors build the sample library which is used for sequencing. This library is mixed with an excess of sepharose beads that carry oligonucleotides which are complimentary to the adaptors. Most of the beads carry afterwards a unique sstDNA fragment. The library is then emulsified with amplification reagents in a mixture of water and oil. Each of these beads is then within its own microreactor where the clonal amplification of sstDNA fragments occurs. Afterwards the beads are added to an incubation mix that contains DNA polymerase and layered with enzyme beads that contain sulfurylase and luciferase on a Picotiterplate. This plate is 7 x 7.5 cm large and contains 1.6 million wells with a diameter of 44 µm per well. The beads are approximately 30 µm large, therefore only one bead fits per well. The layer of enzyme beads ensures that the beads stay in the wells. The PicoTiterPlate is then placed into the Genome Sequencer (GS) FLX^{TM} machine. The fluidics sub-system flows sequencing reagents across the wells of the plate. The nucleotides are flowed sequentially in a specific order across the plate during sequencing. When a nucleotide is complementary to the template strand the polymerase extends the existing strand by

adding the nucleotide. This addition results in a reaction which generates a light signal that is detected and recorded by a specific camera. During the flow of the nucleotides thousands of beads are sequenced simultaniously. The average read length is around 500 bp. The accuracy of 454 sequencing is at least as good as the Sanger sequencing or even better (reviewed in Droege & Hill 2008).

The successfully amplified DNA sample was finally sequenced via 4 384-well plates Sanger sequencing (both directions) using an insert size of 2.5 kb, and 200,000 reads of GS FLX sequencing. Expecting an average read length of the Sanger sequences of 800 bp and of the GS FLX data of 500 bp, the Sanger approach would cover a genome with an expected size of 500 kb roughly 5 times and the GS FLX approach would cover it 200x if no contamination exists.

5.2.4 Assembling and annotating the mitochondrial genome of *Lolium perenne*

The hybrid sequencing approach had been carried out by GATC Biotech AG (Konstanz, Germany) which also undertook the first assembly of the data from both sequencing approaches. These pre-assemblies were carried out with the programs Phrap (http://www.phrap.org/) (only Sanger sequences), Lasergene (DNAstar, Inc., Madison, Wisconsin, USA) and Newbler (454 Life Sciences, Branford, Connecticut, USA) (Figure 31).

Contigs that were based on more than 100 read files were extracted from the Lasergene hybrid assembly based on the assumption that contigs that are made up of less than 100 read files are more likely to contain fragments from contaminating DNA. The extracted contigs were each visually checked for correct assembling and then re-assembled in Lasergene using the proassembler option which is specifically designed for large amounts of data. Except from the settings for the 'maximum mismatch end bases' which was increased to 50 and the 'minimum match percentage' which was increased to 90 the default settings were kept. The increase of the 'maximum mismatch end bases' takes into account that the ends of the contigs may not be accurate due to a reduction in the quality of read files. The 'minimum match percentage' was increased to account for contigs that contain different copies of a repeat and should prevent the combination of

these contigs with each other. Resulting contigs were checked visually in Lasergene for assembly quality. In case of discrepancies, all the readfiles from the relevant preassembled contigs (which were forming the new contig with the discrepancies) were extracted and then completely newly assembled using the same settings. New contigs that contained non-coding sequence information were blasted against each other using the ALIGN function from NCBI to exclude the possibility that two contigs contain exactly the same sequence but could not be assembled because of low quality contig ends that were larger than 50 bp. Thus further contigs could be joined (Figure 31). However, contigs that showed similarities but exceeded the 'maximum mismatch end bases' by more than 500 bp were not joined.

To annotate the resulting contigs, protein-coding genes that all sequenced Poaceae mitochondrial genomes have in common were extracted from mainly the mitochondrial genomes of *Triticum aestivum* and *Oryza sativa* 'japonica'. Using the amino acid sequence of these protein-coding mitochondrial genes in a tblastn (http://blast.ncbi.nlm.nih. gov/Blast.cgi) approach both hybrid assemblies in an earlier approach and then later on the new assembled contigs were screened to find out which contigs contain which genes. Contigs from the new assembly that were larger than 10 kb or contained protein coding genes were searched via tRNA Scan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/) and blastn (http://blast.ncbi.nlm.nih. gov/Blast.cgi) for tRNAs and rRNAs. The gene annotation approach enabled a further joining of contigs when, for example, a non-splicing gene had one exon located on the end of one contig and the other exon on the beginning of another contig (Figure 31).

While annotating the contigs, genes were found that did not have conserved starts or ends or contained frameshifts in comparison to genes from other species, which would result in non-functional genes. The frameshifts were mainly based on indels in microsatellite regions. To confirm the sequence of these genes nine primer pairs (mtLp-atp4-total-FP: TAGTTTCATTGGCCCTCGTC, mtLp-atp4-total-RP: AGGGCATGA CATCCCAAAT/ mtLp-cox1-end-FP: GAAAGTCCTTGGGCTGTTGA, mtLp-cox1-end-RP: TTGAGCGCGTATTCACTCTG/ mtLp-nad6-end-FP: GACGAAATGCC TTGGATTCT, mtLp-nad6-end-RP: TCGAACAGGTAACGCAGTTG/ mtLp-rps1-total-FP: AGTGGAAACGCCTCACAATC, mtLp-rps1-total-RP: TGCGTCTCCCT AGTCGAAGT/ mtLp-rps2-a-FP: AACTTCACCCGCCTTTCTTT, mtLp-rps2-a-RP:

GGGAAGAGGTTGAGGAGGAG/ mtLp-rps2-b-FP: GAAGATCCATTTTGGGC TGA, mtLp-rps2-b-RP: TCTCCCACGTA GGCACATAA/ mtLp-rps2-c-FP:AAACCTGGAATGGACGTGAG, mtLp-rps2-c-RP: TCCCACTTCAACCAGA ATCC/ mtLp-rps3-ex2a-FP: ACGTCCACCTACGAAACTCG, mtLp-rps3-ex2a-RP: GCACCGAAGAAAGGAATGAG/ mtLp-rps3-ex2b-FP: GAAAGCTCTACCGG CGATAA, mtLp-rps3-ex2b-RP: TGCTTCAATGGCTCGATATG) were designed and used on the original non-whole genome amplified DNA sample for PCR amplification (per PCR reaction: 1 µl mitochondrial DNA sample , 1 µl 5x Phusion™ HF Buffer (New England Biolabs, Inc., Ipswich, MA, USA), 0.3 µl forward primer (FP)(10 mM), 0.3 µl reverse primer (RP) (10 mM), 0.3 µl dNTPs (metabion international AG, Martinsried, Germany) (10 mM), 7 µl ddH_2O, 0.1 µl Phusion™ Hot Start High-Fidelity DNA Polymerase (New England Biolabs, Inc.). The PCR program settings were 98 °C 5 min, (98 °C 1 min, 60 °C 1 min, 72 °C 1 min) 35 cycles, 72 °C 10 min.). Each region was amplified twice and sequenced from both ends using once the forward and once the reverse primer (outsourced to the sequencing company).

To enable further assembly of the contigs, a rearrangement study of the gene order within the mitochondrial genomes of so far sequenced Poaceae species was carried out. The rearrangement study was based on a dot-plot approach similar to that published earlier by Ogihara et al. (2005). The order of the protein-coding genes and the three ribosomal RNA genes was manually extracted from all Poaceae mitochondrial genomes published on Genbank (http://www.ncbi.nlm.nih.gov/Genbank/). Gene names were coded with a unique number and exons with a unique letter. The analysis was then carried out in Excel. In a first attempt the mitochondrial genome of *T. aestivum* was used as reference sequence because it is the most closely related sequenced mitochondrial genome to *L. perenne*. The gene order of *T. aestivum* was compared to the gene order of *Bambusa oldhamii*, *Oryza sativa*, *Sorghum bicolor*, *Tripsacum dactyloides* and *Zea mays ssp. mays*. In a second approach *Z. mays ssp. mays* was used as the reference sequence and compared to the other *Zea* species and *T. dactyloides* to analyse the conversation of gene order within one Poaceae subfamily (Panicoideae), one Poaceae genus (*Zea*) and one Poaceae species (*Zea mays*). A third analysis was carried out comparing the two *Oryza sativa* varieties (Vaughan et al. 2008) with each other. Finally also the gene clusters that were found within *L. perenne* were compared to the gene order in *T. aestivum*.

Because a complete assembly of the mitochondrial genome of *L. perenne* is not yet possible the online based software VISTA (http://genome.lbl.gov/vista/index.shtml) was used to visualize the draft assembly. VISTA is a graphical program that takes completely assembled and annotated genomes as well as draft assemblies and aligns them to each other. The huge advantage of VISTA is that it has a function that enables the comparison of rearranged genomes. Therefore this study enables another view on the conservation of mitochondrial genomes. For the analysis, the annotated *T. aestivum* mitochondrial genome was used as reference sequence and aligned with the *Bambusa oldhamii* mitochondrial genome and the *L. perenne* draft assembly. Because VISTA is able to compare rearranged genomes, the order of the *L. perenne* mitochondrial contigs was irrelevant and the contigs were added to each other to create a single sequence that was then used as input for VISTA. To gain a first insight into the number of repeats present in the mitochondrial genome as well as to show that repeats were not discarded by focussing only on the reduced set of contigs, the draft assembly was analysed for repeats larger than 50 bp with the Reputer program (http://bibiserv.techfak.uni-bielefeld.de/reputer/) and for visualization it was aligned to itself using the "Blast 2 sequences" function from NCBI (http://blast. ncbi.nlm.nih.gov/Blast.cgi).

5.2.5 Further analyses

Horizontal gene transfer was investigated by using the "Blast 2 sequences" function and aligning the contigs of the draft *L. perenne* mitochondrial genome to the complete mitochondrial genome of the mycorrhizal fungus *Glomus intraradices* (FJ648425; Lee & Young 2009).

To enable investigation of intracellular (inter organelle) gene transfer in *L. perenne*, the draft *L. perenne* mitochondrial genome was used as reference sequence and aligned against the complete *L. perenne* chloroplast genome using VISTA and "Blast 2 sequences". Chloroplast genome regions that were copied to the mitochondrial genome were then detected visually and specified by comparison to the chloroplast genome annotation.

During all the BlastN analyses, sequences were detected with similarities to a linear plasmid sequence of *L. perenne* which is supposed to be related to CMS in *L. perenne* (AM998372, McDermott et al. 2008). This sequence was extracted and blasted against the draft assembly of this study. *L. perenne* cv. Shandon which was used for extracting the mitochondrial DNA used in this study is a normal fertile variety. Therefore it was hoped to gain insights into the nature of this plasmid in a fertile population.

The rearrangement study revealed one particularly highly conserved region in nearly all Poaceae mitochondrial genomes. This region was extracted from GenBank sequences of some species and aligned with the corresponding region in *L. perenne* using MEGA 3.1 to gain insights into the conservation of intergenic spacer regions within this gene cluster and to estimate its suitability for diversity studies in *L. perenne* cultivars and *Lolium* species in general.

5.4 Results

5.4.1 Preassemblies of the mitochondrial genome

The hybrid sequencing approach resulted in 3,542 Sanger sequences and 383,368 GS FLX sequences. The pre-assemblies carried out by the sequencing company varied widely regarding contig lengths and read files used per contig (Table 14). The Sanger approach only covered the mitochondrial genome 4 times and was therefore expected to generate fewer contigs, with shorter lengths and less read files per contig when assembled on its own. The Newbler hybrid pre-assembly resulted in around 4,000 contigs from 100 bp to 25,000 bp while the Lasergene based pre-assembly produced around 8,000 contigs varying in size from 37 bp to 50,000 bp. The Lasergene assembly contained 4,615 'normal' contigs and 4,114 additionally contigs that were based on putative repetitive regions according to the sequencing company. None of the received pre-assemblies produced one large complete contig.

Table 14 Overview of the preassemblies of the *Lolium perenne* mitochondrial genome retrieved from the sequencing company.

	assembly programme	number of contigs	contig lengths in bp	number of read files/contig	combined contig lengths in bp
Sanger	Phrap	227	89 - 8220	1 - 677	431405
Sanger + GS FLX	Newbler	4185	100 - 24912	1 - 3910	2041920
Sanger + GS FLX	Lasergene	8729	37 - 49829	2 - 43322	6204640

Initial analyses of both hybrid pre-assemblies for coding regions within the contigs revealed that less than 40 contigs contained the expected amount of mitochondrial genes (Figure 31). Neither pre-assemblies contained a complete set of genes, however the Lasergene based pre-assembly contained more genes than the Newbler based pre-assembly and generally seemed to be better and more accurate. Therefore it was decided to base further analyses on the Lasergene pre-assembly. Adding the lengths of all Lasergene derived contigs that contained protein-coding sequences together, revealed that they only accounted for approximately half of the expected *L. perenne* mitochondrial genome size. The contig lengths varied from around 1,000 bp to up to 35,000 bp. Thus, major parts of the mitochondrial genome were still hidden in the huge amount of pre-assembled Lasergene contigs.

Sorting the Lasergene contigs according to the amount of read files they are based on, revealed that only 140 out of more than 8,000 contigs were based on more than 100 read files. These 140 contigs were used as input for a new assembly approach in Lasergene. The re-assembly could reduce the number of possible mitochondrial DNA containing contigs to 99 by joining some contigs. Thirty one of the new contigs had a length of more than 10 kb and a total length of 490,381 bp (Figure 31). This was in the expected size range for the mitochondrial genome of *L. perenne* and it was believed that these 31 contigs contain all the protein-coding genes. However, analyses showed that approximately half of the protein-coding gene information was located in the contigs that were smaller than 10 kb. Further studies showed that some of these contigs could still be joined with the bigger contigs when assuming that up to 500 bp at the end of the contigs could be inaccurate.

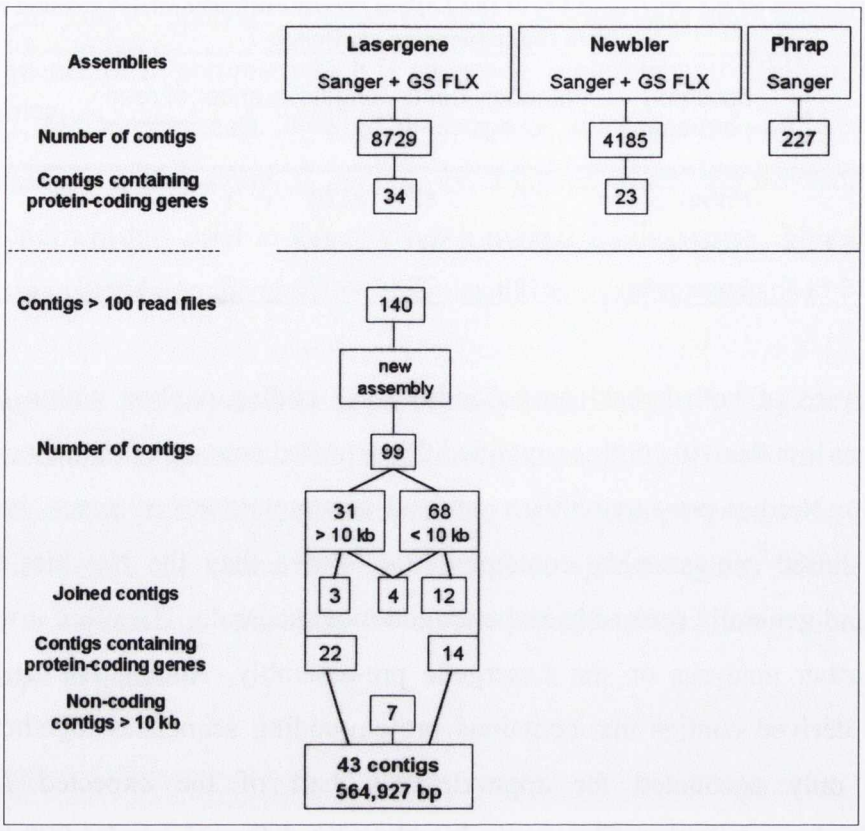

Figure 31 Overview over the assembly process used for gaining a draft assembly of the mitochondrial genome of *Lolium perenne*.

Taking all contigs that contain protein-coding sequences or are larger than 10 kb into account, a final number of 43 contigs with a combined length of 564,927 bp was achieved (Figure 31, Table 15). Within these 43 contigs we detected 33 protein-coding genes, three ribosomal RNA and 20 transfer RNA genes (14 unique tRNAs) (Table 15, Figure 32). Two contigs were found that encode *nad3*. It was not possible to combine these two contigs because they differed by up to 2,000 bp from each other. Thus it is possible that two copies of the *nad3* gene exist in the *L. perenne* mitochondrial genome. The order and orientation of the protein-coding genes could not be established in this study.

5.4.2 Gene conservation and rearrangement study

The mitochondrial genomes of all so far sequenced Poaceae species, as well as one dicotyledonous species (*Arabidopsis thaliana*) and one liverwort species (*Marchantia polymorpha*) were compared to assess conservation of mitochondrial gene content

across species. Thirty-three protein coding genes were common in all Poaceae species with only a few exceptions (*atp4*: absent from *Oryza sativa*, *rpl16*: absent from *Oryza sativa japonica*, *rpl5*: present only in *Oryza sativa*, *Bambusa oldhamii*, *Lolium perenne*, *Triticum aestivum*). Three genes (*cox3*, *nad4* and *nad4L*) were highly conserved regarding their length in all species, six (*ccmB*, *cox2*, *nad1*, *nad7*, *nad9*, *rps12*) in all Poaceae species and *A. thaliana* and eight in most of the Poaceae species (Table 16). All these 17 genes were also highly conserved regarding their length in *L. perenne*.

Analysing the order of these 33 protein-coding genes and the three ribosomal RNA genes revealed that mitochondrial genomes are highly rearranged even within subfamilies, genera and species (Figure 33 – Figure 36). Analysing the gene order of the subfamily Panicoideae shows that the less related species are, the less conserved is their gene order in comparison to each other. While the *Zea* subspecies showed ten clusters of up to twelve linked genes the comparison of maize with *Tripsacum dactyloides* showed only eight clusters with a maximum of four genes (Figure 34). Even two cultivars of the same species show a considerable number of rearrangements as illustrated for the two *Oryza* varieties (Figure 35).

Table 15 Overview over the different contigs, their lengths and the genes they are encoding. Contigs in grey do not contain any protein-coding information.

Contig	length	genes	start	end	strand
9	49881	ccmC	19996	19274	-
		nad4L	46422	46724	-
		trn-Phe	46827	46899	-
25	26893	rrn18	1822	3798	+
		rrn5	3912	4033	-
		nad1-d	15345	15402	-
		nad6	13407	14300	-
		ccmFC-a	22671	23428	+
		ccmFC-b	24428	24994	+
15	24773	atp8	5957	6424	+
		trn-His	18599	18672	+
12	23185	trn-Pro	831	905	+
24	23070	nad1-b	7470	7663	+
		nad1-c	9078	9159	+
		rps13	9697	10047	-
20+79+96_1	20632	nad5-e	310	459	-
		cob	3026	4219	-
		rpl5	5476	6050	-
		trn-Gln	11033	11104	-
		cox2-a	17913	18302	+
		cox2-b	19587	19979	+
6	19927	nad4-a	8389	8849	+
		nad4-b	9881	10395	+
		nad4-c	13018	13440	+
		nad4-d	16237	16325	+
17+68	15775	nad5-c	694	714	-
		cox1	12413	14017	+
18	16982	cox3	9578	10375	-
		nad9	10692	11264	-
10	14885	ccmB	12741	13361	-
30	14728				
19	14476				
32	13978	trn-Ser	8644	8730	-
26	13475	atp1	9342	10872	+
31	13068	trn-Tyr	8197	8279	+
		nad2-c	8917	9077	-
		nad2-d	11489	12061	+
48+67	13000	nad5-c	1624	1644	+
		nad1-e	2007	2264	-
		matR	2781	4817	-
		rps1	8270	8794	-
		ccmFN	9040	10806	-
16	12594				
74	12246				

Contig	length	genes	start	end	strand
34	12088				
37	11978				
35	11709	trn-Met	9644	9717	-
(33+45)2	17136	trn-Glu	10945	11016	+
42	11481	rrn26	1090	4552	+
36	11398	trn-Pro	1542	1615	+
		trn-Trp	1759	1832	+
23	11289	rps4	6451	7485	-
13	11090	trn-Ser	4943	5029	+
		rps7	5795	6241	+
		trn_Lys	9013	9085	+
		trn-Lys	9976	10048	-
3+(33+45)1	16017	trn-Asp	3739	3812	-
		atp4	5888	6496 ?	-
89	10099	trn-Val	7997	8068	+
		trn-Leu	1846	1926	-
39+44	19372				
60+77+78	9111	trn-Cys	310	380	+
		nad7-e	2156	2418	-
		nad7-d	4124	4367	-
		nad7-c	5367	5832	-
		nad7-b	7144	7212	-
		nad7-a	8016	8158	-
21+93	6982	nad5-b	2417	3632	-
		nad5-a	4503	4733	-
86+11+87	11300	atp9	9161	9385	-
51	6787	rps2	4325 ?	5477	-
1+91	7535	rps3-a	2868	2941	+
		rps3-b	3877	5499	+
		rpl16	5345	5902	+
55+56	5692	mttB	2965	3780	-
54	5355	atp6	1623	2786	-
62	5149	trn-Ser	442	530	-
		nad3	887	1243	+
		rps12	1289	1666	+
70	5131	nad2-a	1319	1710	-
22	4481	nad2-b	2351	2503	+
65	4136	nad1-a	2496	2882	-
		trn-Pro	4062	4136	-
85	2620	nad3	272	629	-
		trn-Ser	1144	1231	-
20+79+96_2	1782	nad5-d	731	1125	+
94	1641	nad2-e	1155	1342	-
total length	**564927**				

Table 16 Comparison of the gene content and gene length (in number of amino acids) of all publicly available Poaceae mitochondrial genome sequences and of one dicotyledonous (*Arabidopsis thaliana*) and one moss (*Marchantia polymorpha*) species. Highlighted in gray are genes with a conserved length, - indicates loss of gene, empty space indicates not annotated in the respective genome.

Marchantia	*Arabidopsis*	gene	*Bambusa*	*Lolium*	*Oryza sativa*	*Oryza sativa*	*Sorghum*	*Tripsacum*	*Triticum*	*Zea luxurians*	*Zea mays*	*Zea mays ssp.*	*Zea*
513	507	***atp1***	509	509	509	509	513	508	509	508	508	508	508
183		***atp4***	196	202	-	-	208	212	192	221	219	221	221
252	385	***atp6***	446	387	337	334	382	409	386	409	409	409	409
172		***atp8***	155	155	155	155	155	153	156	153	153	153	153
74	85	***atp9***	80	74	75	74	86	84	80	74	74	74	74
277	206	***ccmB***	206	206	206	206	206	206	206	206	206	206	206
228	203	***ccmC***	240	240	240	240	240	240	240	240	240	240	240
509	256	***ccmFC***	438	438	435	435	437	433	438	433	433	433	434
169	382	***ccmFN***	607	588	594	594	608	605	589	602	620	620	602
404	393	***cob***	399	397	397	397	388	388	398	388	388	388	388
522	527	***cox1***	524	535	524	524	530	710	524	528	528	528	528
251	260	***cox2***	260	260	260	260	260	260	260	260	260	260	260
265	265	***cox3***	265	265	265	265	265	265	265	265	265	265	265
	672	***matR***	678	678	678	678	678	675	678	670	673	673	673
		mttB	271	271	269	269	271	271	271	271	271	271	271
328	325	***nad1***	325	325	325	325	325	325	325	325	325	325	325
489	499	***nad2***	488	488	488	488	488	488	488	488	488	488	488
118		***nad3***	118	118	155	155	118	118	118	118	118	118	118
495	495	***nad4***	495	495	495	495	495	495	495	495	495	495	495
100	100	***nad4L***	100	100	100	100	100	100	100	100	100	100	100
669	669	***nad5***	670	670	670	670	670	670	670	670	670	670	670
199	205	***nad6***	205	296	205	205	271	333	247	375	220	220	209
	394	***nad7***	394	394	394	394	394	394	394	392	394	394	394
212	190	***nad9***	190	190	190	190	190	190	287	190	190	190	190
135	179	***rpl16***	185	185	185	-	185	185	185	185	185	185	185
188	185	***rpl5***	188	189	188	188	-	-	189	-	-	-	-
270		***rps1***	174	174	172	172	183	177	174	172	172	172	172
237		***rps2***	421	387	483	483	447	559	360	530	531	531	533
430	556	***rps3***	558	566	538	538	559	559	537	556	559	559	559
196	362	***rps4***	344	344	352	352	383	356	357	358	358	358	358
230	148	***rps7***	148	148	148	148	148	148	148	148	148	148	148
126	125	***rps12***	125	125	125	125	125	125	125	125	125	125	125
120		***rps13***	116	116	116	116	116	116	116	116	116	116	116

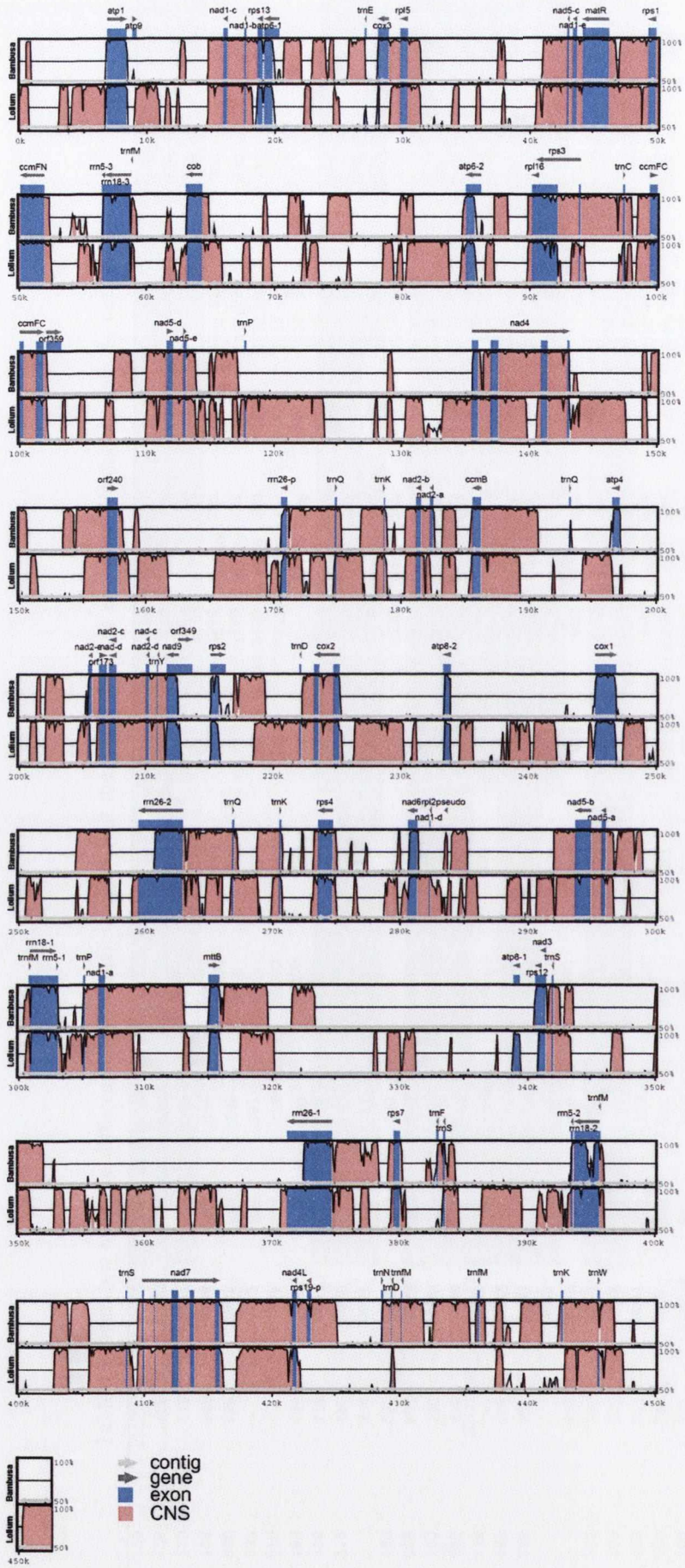

Figure 32 Alignment of the complete *Bambusa oldhamii* and the draft *Lolium perenne* mitochondrial genome sequence to the complete and annotated *Triticum aestivum* mitochondrial genome sequence using the Shuffle-Lagan alignment algorithm from the online program Vista (http://genome.lbl.gov/vista/index.shtml). The reference sequence *T. aestivum* is shown on top of the alignment indicated as one continuous sequence by the annotated genes and at the bottom of the alignment indicated by the length in kilo base pairs (k). Dark grey arrows within the reference sequence indicate genes and their orientation. Blue boxes/peaks indicate exons within the reference sequence but also within the aligned sequences.

Light grey arrows indicate the different contigs as a result of the Shuffle-Lagan algorithm. They do not correspond to the contigs of the *L. perenne* mitochondrial genome draft assembly.

The alignment shows only similarity of both *B. oldhamii* and *L. perenne* to *T. aestivum*, not to each other. The higher the percentage of similarity between one of the species and the reference sequence, the higher the peaks.

Red: conserved non-coding sequence (CNS)
Blue: conserved coding sequence (exon)

Nevertheless, comparing the gene order of different Poaceae genomes with *T. aestivum* revealed six groups of at least two genes and exons, respectively, that seem to be linked in all genomes (Figure 33). Always found linked together are the genes *rrn5* and *rrn18*; *rps12* and *nad3*; *nad2* exon number 3 and *nad2* exon numbers 4, 5; *rpl16* and *rps3*; *rps1* and *ccmFN* as well as *nad1* exon number 5 and *matR*.

In the draft assembly of the *L. perenne* mitochondrial genome, nine contigs contained more than one protein-coding gene or exons of different genes. Comparing the gene clusters found in *L. perenne* with the gene order of *T. aestivum* (Figure 36) revealed six clusters that the two species have in common (*rrn5* and *rrn18* / *rps12* and *nad3* / *rpl16* and *rps3* / *ccmFN, rps1, matR, nad1* exon 5 and *nad5* exon 3 / *nad6* and *nad1* exon 4 / *nad1* exon 2,3 and *rps13*). Only the cluster of *nad6* and *nad1* exon number 4 was specific for the Pooideae species.

Figure 37 shows the alignment of the sequence of the large gene cluster 'ccmFN - nad5 exon 3'. It showed that not only the genes and their order but also the intergenic spacer regions are highly conserved. The highest sequence variation was observed between *Zea mays ssp. mays* and *Triticum aestivum* with 149 out of 9,381 bp and 84 out of 4,796 bp (only intergenic spacer regions), respectively (Table 17). The highest conservation was observed for the short intergenic spacer between *ccmFN* and *rps1* and the lowest conservation for the intergenic spacer region between *matR* and *rps1*.

Table 17 Number of pairwise differences in the sequence alignment of the large plant mitochondrial gene cluster 'ccmFN-nad5 exon 3'. Total = the complete alignment is taken into account, IGS = only intergenic spacer regions were considered, *Lolium* = *Lolium perenne*, *Triticum* = *Triticum aestivum*, *Oryza* = *Oryza sativa*, *Sorghum* = *Sorghum bicolor*, *Zea* = *Zea mays ssp. mays*. Grey = only *rps1 – matR* intergenic spacer.

total					IGS			
Lolium	*Triticum*	*Oryza*	*Sorghum*		*Lolium*	*Triticum*	*Oryza*	*Sorghum*
85				*Triticum*	57 (46)			
126	134			*Oryza*	72 (59)	73 (57)		
123	127	126		*Sorghum*	71 (58)	69 (51)	70 (50)	
142	149	139	34	*Zea*	83 (63)	84 (59)	78 (51)	21 (12)

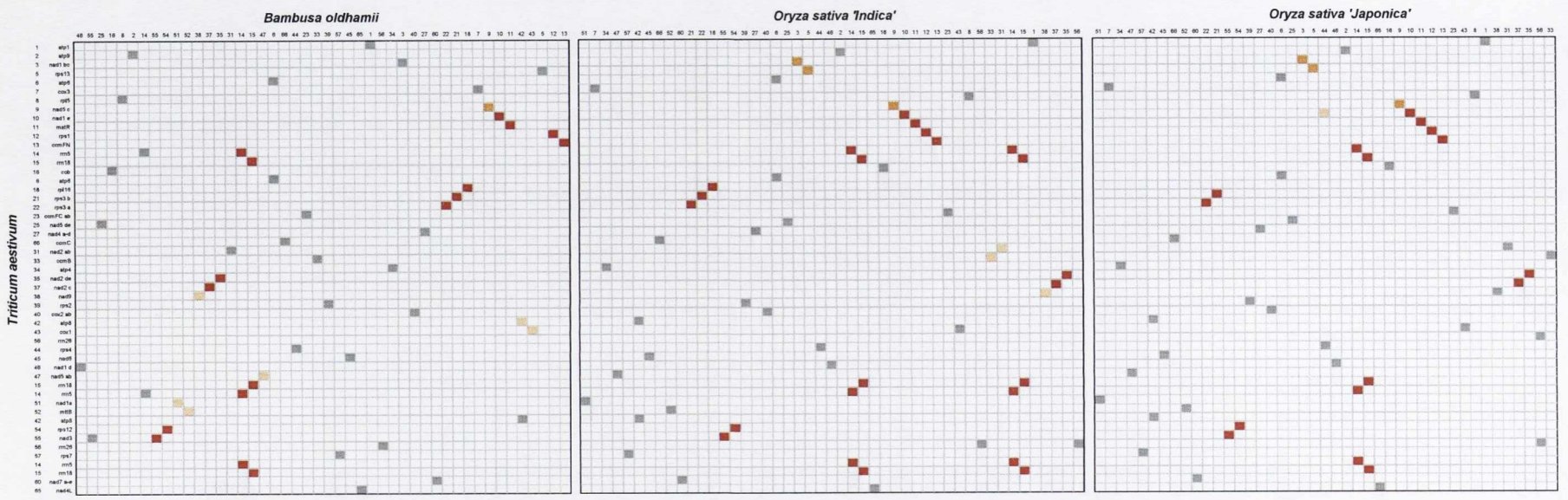

Figure 33 Comparison of the mitochondrial gene order in species of different Poaceae subfamilies using T. aestivum as reference sequence. The more intense the colour the more conserved is this gene order in the different species (grey highlighted boxes = no conservation, red highlighted boxes = complete conservation).

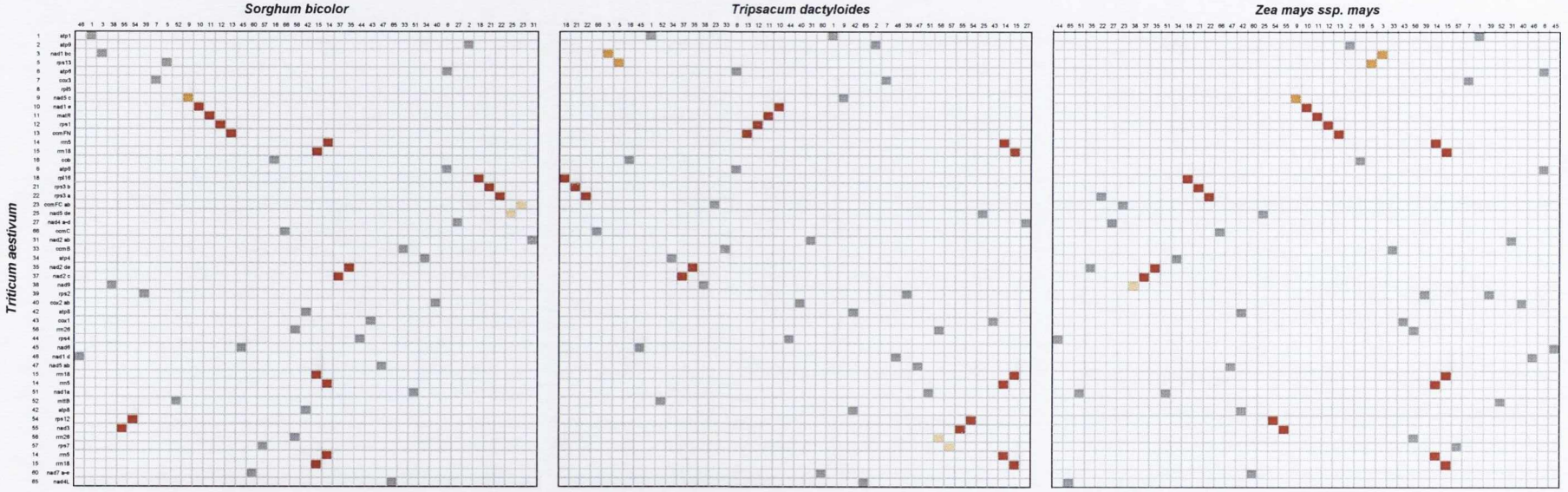

Figure 33 *continued*

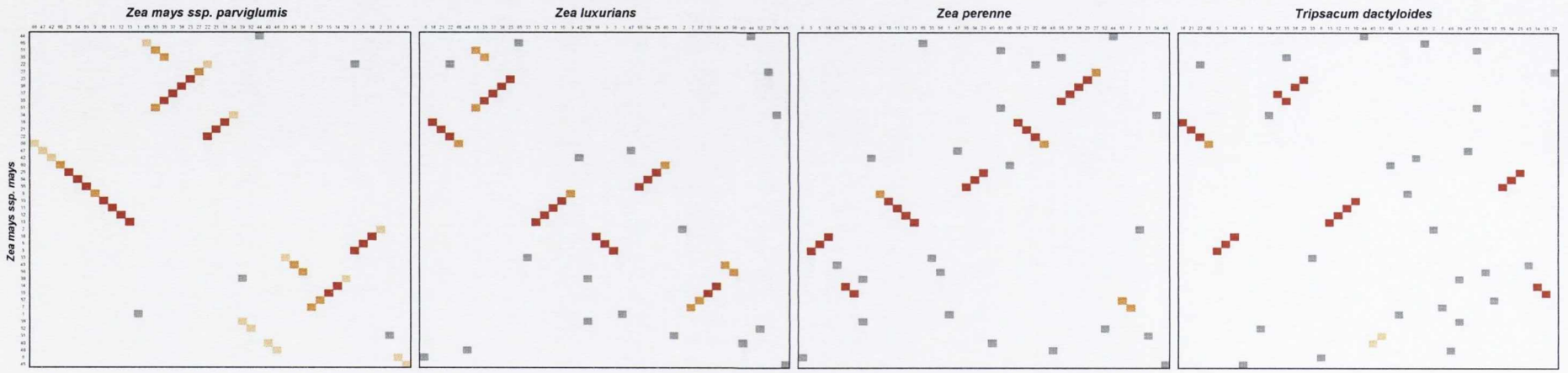

Figure 34 Comparison of the mitochondrial gene order in different species of the Panicoideae subfamily *using Zea mays ssp. mays* as reference sequence. The more intense the colour the more conserved is this gene order in the different species (grey highlighted boxes = no conservation, red highlighted boxes = complete conservation). Genes are coded with numbers based on the *Triticum aestivum* gene codes used in Figure 5.

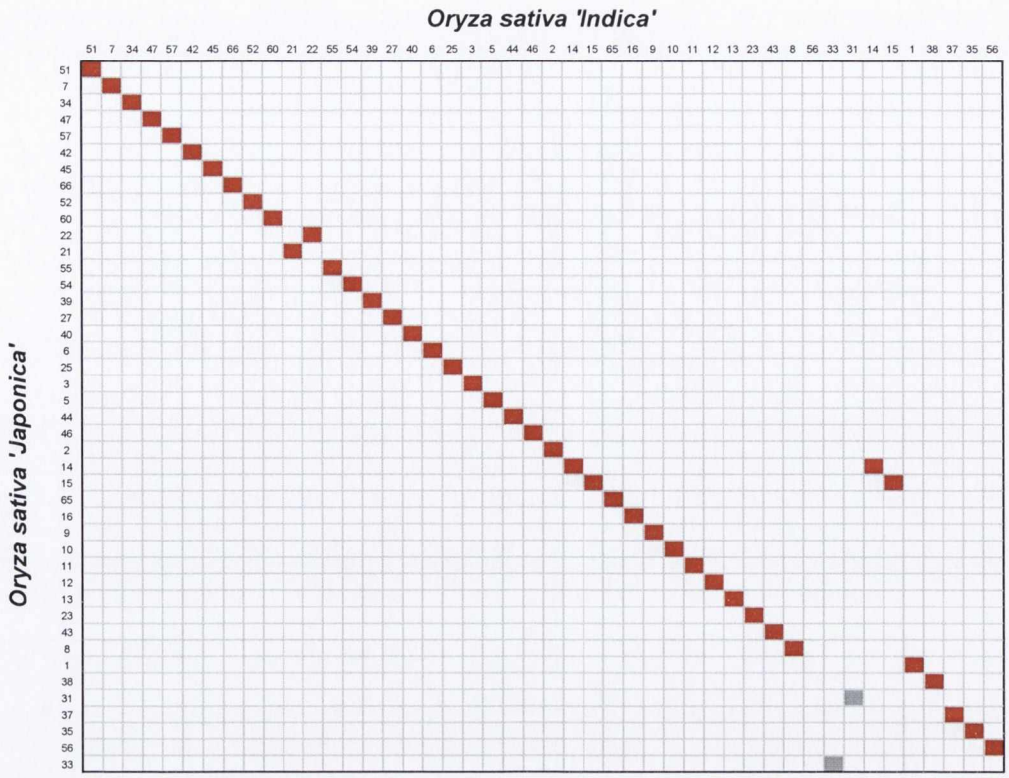

Figure 35 Comparison of the mitochondrial gene order in the two *Oryza sativa* varieties. Genes are coded with numbers based on the *Triticum aestivum* gene codes used in Figure 5; grey highlighted boxes = no conservation, red highlighted boxes = complete conservation.

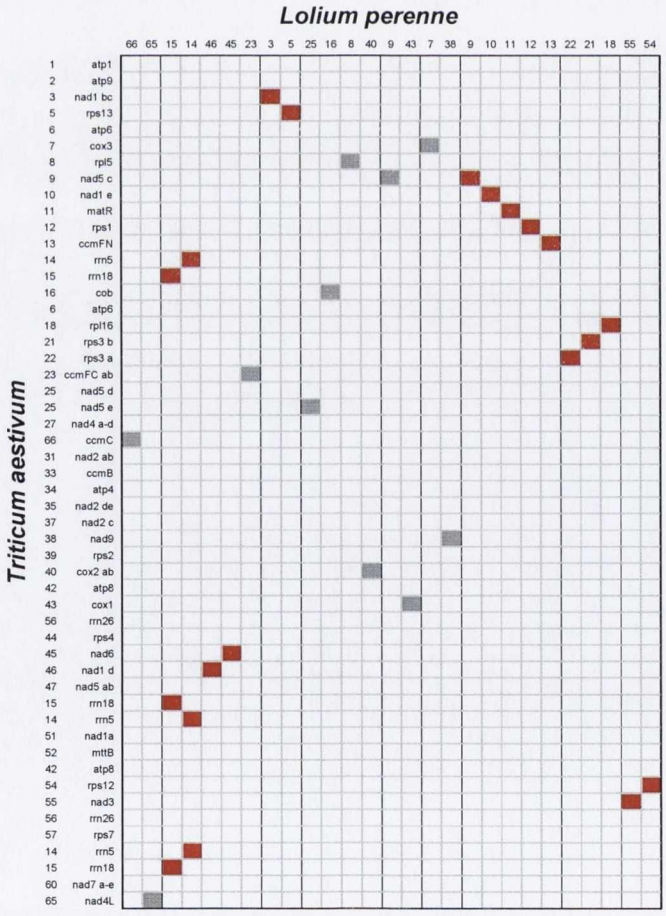

Figure 36 Comparison of the mitochondrial gene order of *Triticum aestivum* with the gene clusters found so far in the draft assembly of *Lolium perenne*. Grey highlighted boxes = no conservation, red highlighted boxes = complete conservation.

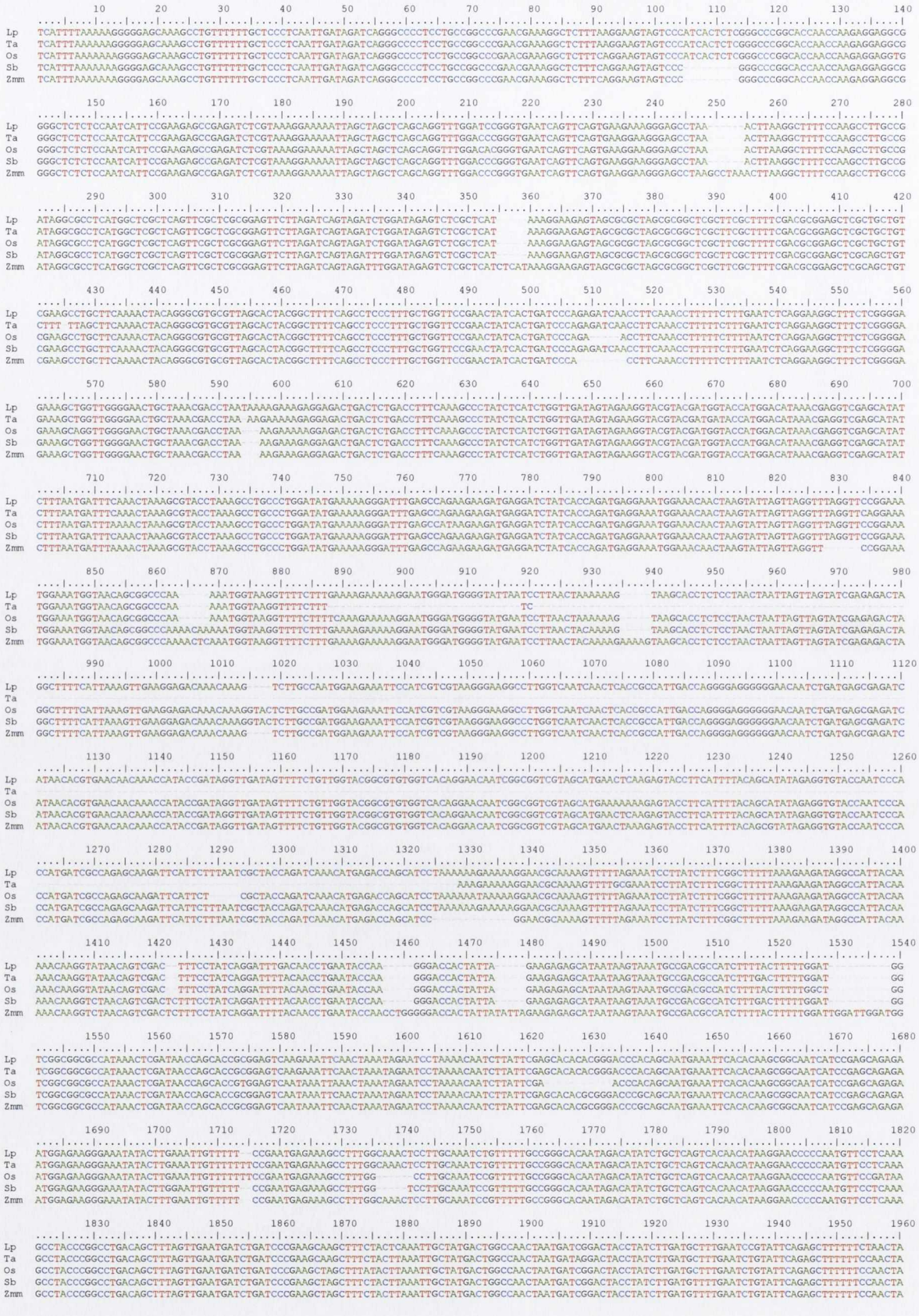

Figure 37 Alignment of the *matR – rps1* intergenic spacer region from five different Poaceae species.
Lp = *Lolium perenne*, Ta = *Triticum aestivum*, Os = *Oryza sativa*,
Sb = *Sorghum bicolor*, Zmm = *Zea mays ssp. mays*.

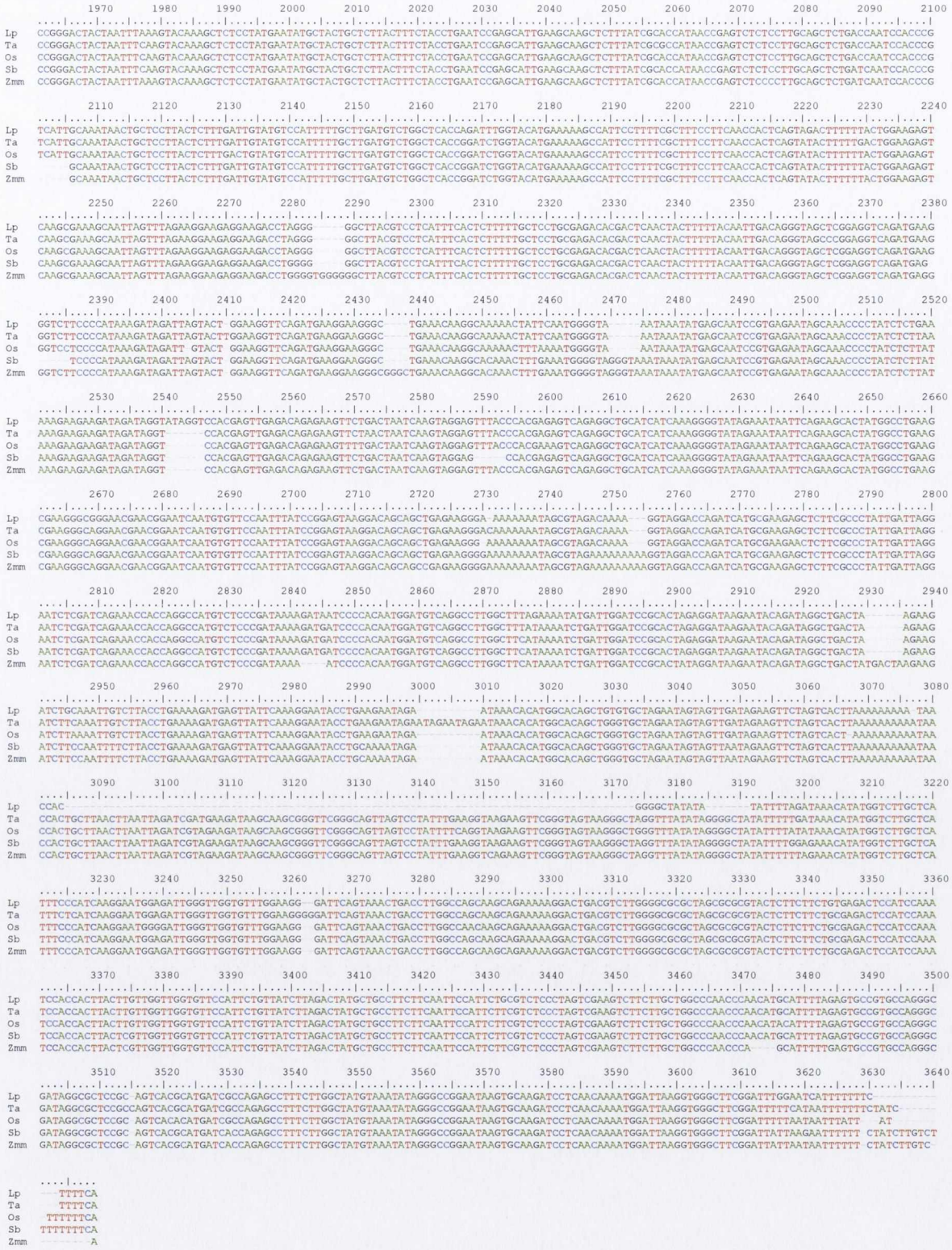

Figure 37 *continued*

Also the visualization of the *L. perenne* mitochondrial genome in comparison to *T. aestivum* (Figure 32) showed a considerable high amount of conserved regions but also long stretches in the sequence with no sequence similarities at all. The *Bambusa oldhamii* sequence revealed a similar pattern when compared to *T. aestivum* and confirms the *L. perenne* results (Figure 32). Surprisingly this alignment did not reveal

the presence of the genes *atp9* and *cox3*, although both genes were present in the input sequence and also the annotation in the *T. aestivum* reference sequence was correct.

The Reputer repeat analysis revealed 370 repeats, direct and inverted ones, with lengths from 50 bp to 1151 bp. Some repeats were present in more than only two copies. However, as shown in Figure 38 the repeats were distributed throughout all contigs.

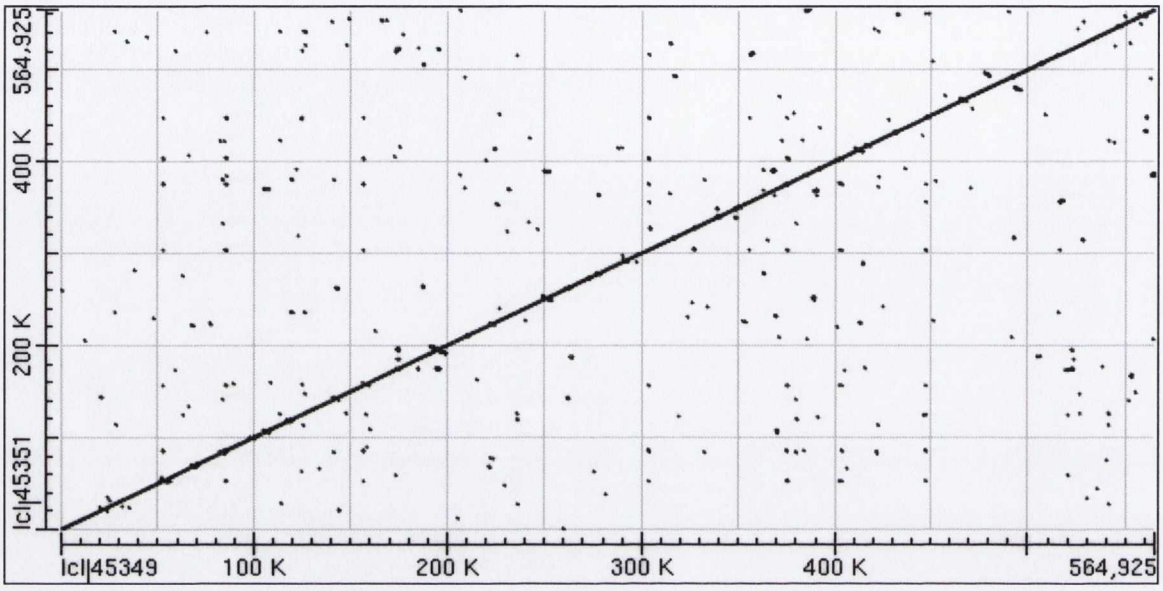

Figure 38 Alignment of the *Lolium perenne* mitochondrial genome draft assembly with itself to highlight repetitive regions using the NCBI blast tool. X and Y axis both represent the mitochondrial draft genome, the diagonal line represents also the draft assembly aligned with itself, surrounding dots and lines represent all repetitive regions within the assembly.

5.4.3 Further results

The horizontal gene transfer analyses revealed a region comprising 322 bp that showed a sequence similarity of 76% between the *L. perenne* and sequences from the *G. intraradices* mitochondrial genome (Figure 39). This region belonged to the large ribosomal subunit in *Glomus* as well as *Lolium*. Further analyses revealed similar hits in other Poaceae as well as dicotyledonous mitochondrial genome species.

An alignment of the mitochondrial draft assembly with the *L. perenne* chloroplast genome revealed two large regions of chloroplast DNA that were obviously transferred to the mitochondrial genome (Figure 40). One region (IGT 1) derives from the large single copy region of the chloroplast genome, is 12,344 bp long and encodes in the chloroplast genome the genes *rpoC1*, *rpoC2*, *rps2*, *atpI*, *atpH*, *atpF*. The other region

(IGT 2), 10,127 bp long, derives from the inverted repeat region which codes for the chloroplast genes *rrn16*, *trnV-GAC*, *rps12* 3'-end, *rps7*, *ndhB* and *trnL-CAA*. Also a third region from the chloroplast inverted repeat already earlier described by Watanabe et al. (1994) was found in the draft assembly (IGT 3; 2070 bp). All three IGT events were found on different contigs (IGT1 = contig 89, IGT2 = contig 74, IGT3 = contig 15) and did not, with exception to contig 15, contain any mitochondrial protein-coding information. However, all contigs that did not contain protein-coding genes were blasted on NCBI and showed, in all cases, high homology to plant mitochondrial DNA sequences. Mitochondrial protein-coding genes of chloroplast origin are known to have lost their function due to frameshifts in the coding sequence. This was also observed for protein-coding genes of chloroplast origin in *L. perenne* as illustrated for the *rpoC2* gene in Figure 41. Chloroplast derived transfer ribosomal genes are generally completely conserved in their sequence and believed to be functional. Figure 42 shows that this was also the case in the *L. perenne* mitochondrial genome using the *trnV*-GAC gene as example.

```
>lcl|56355 gi|222839796:4017-4321 Glomus intraradices mitochondrion, complete
genome
Length=305

 Score =  214 bits (236),  Expect = 1e-58
 Identities = 245/322 (76%), Gaps = 18/322 (5%)
 Strand=Plus/Minus

Query  1999  GAATTTCGCTACCTTAGGACAGTTAGAGTTACTGCCGCCGTTTACCGGGGCTTCCATTCA  2058
              ||||||||||||||||||||| || |  ||||| |||||| || ||  | ||||||  ||
Sbjct  305   GAATTTCGCTACCTTAGGACCGTCAAGGTTACAGCCGCCATTGACTAGAGCTTCCTTTGC  246

Query  2059  AAGCTTATCACACTTCTCCTTCCGACCTTCCAGCACCGGGCAGGTGTCAGACTCTATACA  2118
               | | ||  ||   | |  ||  || | || |||| |||||||| ||||||||||||||
Sbjct  245   CATCCTAAAACGGCT-TAGTT--GATCATCTAGCATCGGGCAGGTATCAGACTCTATACA  189

Query  2119  TCGT-GTTACCACTTAGCAGAGTCCTGTGTTTTTAATAAACAGTCGCTACCCCCTGGTAT  2177
              ||   | ||    |||||||||||| ||||||| | ||||||||||  ||||||| |
Sbjct  188   TCCACGCTAGTG-TTAGCAGAGTCCGGTGTTTTAAGTAAACAGTCGCGCCCCCCTGTTTA  130

Query  2178  GTGCCGCTTTCCTAATCAAAAGATAGGAGAGCACCCCTTCTCCCGAAGTTACGGGGTCAT  2237
              |||||||    || || |        | | ||||||||||||||||||||||||||||||
Sbjct  129   GTGCCGCCA--CTTATTA--------GTG-GCACCCCTTCTCCCGAAGTTACGGGGTCAT  81

Query  2238  TTTGCCGAGTTCCTTCGACATGGTTCTCTCGACGCGCCCTAGTATACTCTACTTGTTCAC  2297
              |||||||||||||||| | | | || |||| |  |||| |||||| |||||||||   ||
Sbjct  80    TTTGCCGAGTTCCTTCAAGACGATTATCTCTA--CGCCTTAGTATGCTCTACTTGCAGAC  23

Query  2298  CTGTGTCGGTTTGGGGTACGGT  2319
              |||||||||||| |||||||||
Sbjct  22    CTGTGTCGGTTTAGGGTACGGT  1
```

Figure 39 Blast result of the horizontal gene transfer analyses showing a sequence alignment of a *Glomus intraradices* mitochondrial DNA fragment with a region of the *rrn26* gene of the *L. perenne* mitochondrial genome. Query = *Lolium perenne* sequence, Sbjct = *Glomus intraradices*.

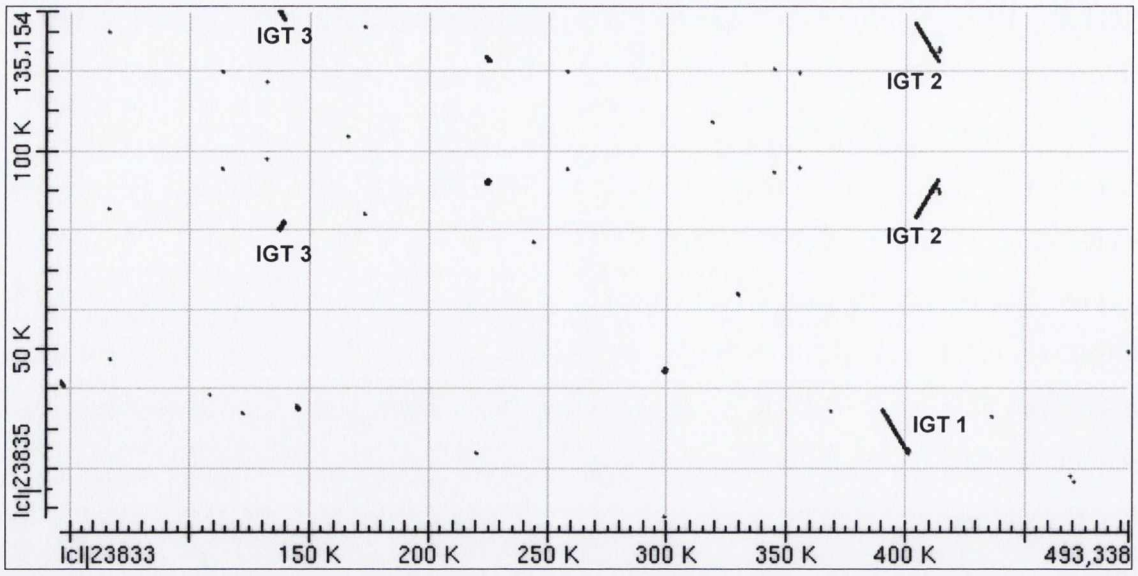

Figure 40 Alignment of the *Lolium perenne* mitochondrial draft assembly with the complete chloroplast genome of *L. perenne*. X-axis = chloroplast genome, Y-axis = mitochondrial draft assembly, IGT = intracellular gene transfer events.

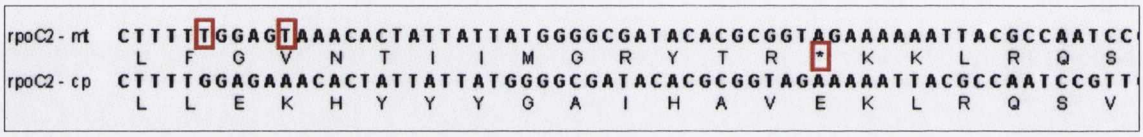

Figure 41 Alignment of a fragment of the *Lolium perenne* chloroplast (cp) *rpoC2* gene with its homologue in the *L. perenne* mitochondrial (mt) genome. The chloroplast *rpoC2* gene was copied to the *L. perenne* mitochondrial genome. The mitochondrial *rpoC2* gene is a pseudogene due to several frameshifts caused by insertion/deletion events (red box, T) that result in stopcodons (red box *.)

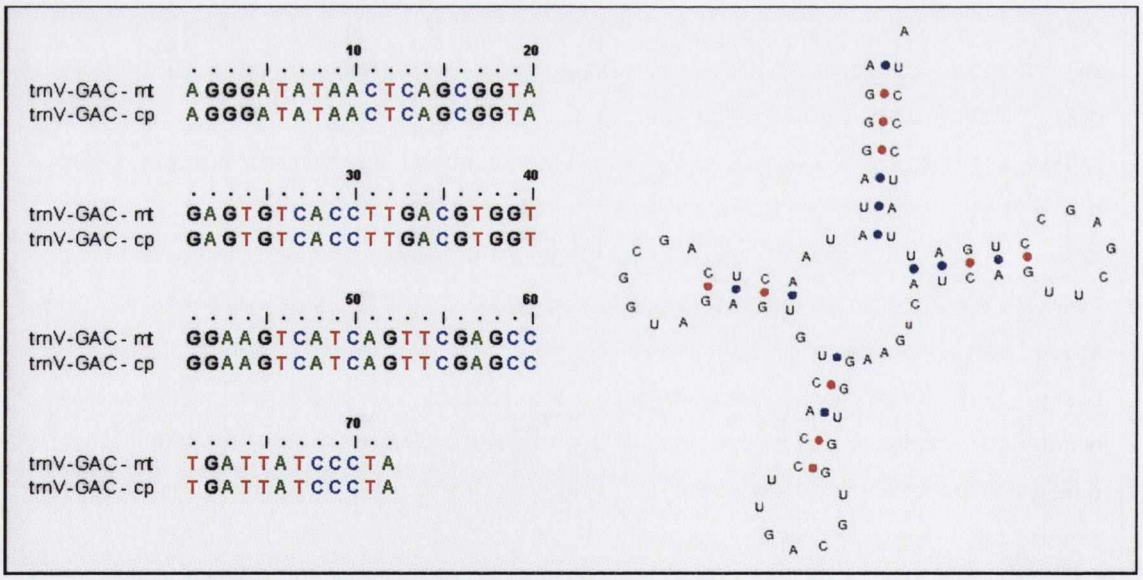

Figure 42 Alignment of the chloroplast derived mitochondrial *trnV*-GAC of *Lolium perenne* with the original chloroplast *trnV*-GAC gene of *L. perenne* showing complete homology between the two sequences. Both genes are able to form a cloverleaf-like structure indicating their functionality (right part of figure).

5.5 Discussion

5.5.1 The mitochondrial genome of *Lolium perenne*

Plant mitochondrial genomes are known for their very complex nature consisting of only few coding regions, many repetitive regions of various sizes, a highly rearranged structure and sequences derived from horizontal and intracellular gene transfer. The *L. perenne* mitochondrial genome is unfortunately (from a genome assembly perspective) not an exception and its isolation and assembly were very challenging. Therefore it was not possible to complete the assembly and annotation of the *L. perenne* mitochondrial genome within this project due to time restrictions. However, it was possible to narrow the huge data set of more than 13,000 contigs from three different assemblies down to a data set of 43 contigs from only one assembly that very likely contains the complete sequence of the mitochondrial genome of *L. perenne*. Within this data set we found all the protein-coding genes, three ribosomal RNA genes and 20 transfer RNA genes that had previously been recorded from other angiosperm mitochondrial genomes. Before starting this project, the estimated size of the *L. perenne* mitochondrial genome was around 500 kb based on restriction fragment analyses (Kiang et al. 1992) and previously sequenced Poaceae mitochondrial genomes. Combining the length of contigs from this study that are definitely mitochondrial DNA reveals a genome size of 565 kb. However, it is expected that this size will be reduced in later studies, when further possibly overlapping contigs can be joined together. In this study contigs that were overlapping by more than 500 bp (more than 500 bp would need to be deleted, before contigs could be joined) were ignored because they were possibly based on inaccurate assemblies or repeated regions. However, these contigs might be useful in further studies to enable the completion of the genome assembly.

This is not the first study encountering difficulties in assembling a plant mitochondrial genome (Clifton et al. 2004). Compared to assembling the very first plant mitochondrial genome (*Arabidopsis thaliana*, Unseld et al. 1999) there were only minor advantages of assembling the mitochondrial genome of *L. perenne.* These advantages included 1) knowledge about the expected number of protein-coding sequences, 2) some conserved regions in the Poaceae family had already been detected (Ogihara et al. 2005) and 3) some chloroplast genome regions that can be copied to the mitochondrial genome via

intracellular gene transfer were known. However, only very few studies based on the mitochondrial genome of *L. perenne* or one of its very close relatives have been carried out to date. Most analyses were based on CMS in *L. perenne* and dealt mainly with a linear mitochondrial plasmid of *Lolium* (Kiang & Kavanagh 1996 and McDermott et al. 2008). Only one mitochondrial gene of *L. perenne*, *atp9*, had been published prior to this study (McDermott et al. 2008) and the order of genes within the mitochondrial genome of *L. perenne* was unknown. Also, the closest relative of *L. perenne,* for which a published mitochondrial genome is available, is *T. aestivum*. The rearrangement study carried out in this chapter showed that only sequences from very closely related species are helpful for the assembly of the mitochondrial genome. They need to be of at least the same genus or even better the same species to contain valuable information about the gene order. The more distantly related species are the less well conserved in their genetic information (Ogihara et al. 2005). Hence, *Triticum aestivum* is from a different tribe (Triticeae) within the Pooideae and not very suitable for gaining information about the gene order in *L. perenne* (tribe Poeae).

Despite the difficulties encountered during the assembly of the genome it was still possible to detect all genes that mitochondrial genomes of Poaceae species have in common and thus to enable mitochondria based phylogenetic studies in the grass family (Chapter 6). Plant mitochondrial genes are known to be highly conserved and this was also detectable when monitoring their length. Approximately half of the mitochondrial genes had exactly the same length in all Poaceae species and the remaining ones showed a high degree of length conservation. Adding the genes from *L. perenne* to the list of genes from published Poaceae species confirmed this observation. However, plant mitochondrial genes do not only show a high degree of conservation regarding their length. Their sequences are also generally highly conserved. This was observed not only when aligning the genes of the different species with each other but also when doing whole genome comparison using the VISTA software. The highest regions of conservation were mainly found in coding regions as already previously observed by Ogihara et al. (2005). Exceptions were only found for four genes in this study: *atp4*, *atp6*, *mttB* and *nad6*. While the former three had non-conserved starts, the latter one had a non-coserved ending. Initially it was thought that *L. perenne* sequences that diverge widely from the *T. aestivum* gene sequence are based on wrongly assembled read files. However, sequence alignments showed that these generally did not have conserved

starts or ends. In the case of *atp6* it was observed that this gene can be present in two different molecular forms in mitochondrial genomes and that the 'pre-sequences' from both forms can be highly different from each other in Poaceae species (Bonen & Bird 1988, Gibala et al. 2004). Studies on radish showed that these pre-sequences are cleaved post-translationally and it is believed that this is also the case in other species (Gibala et al. 2004). Thus all mature *atp6* proteins start with the conserved S/T-P-L amino acid sequence. This sequence was also found in *L. perenne*. No literature could be found about non-conserved regions in *atp4*, *nad6* and *mttB*. Considering that these genes are not completely conserved across all Poaceae species, however, supports the sequences revealed for *L. perenne* by this study. Future studies could show that similar mechanisms work in *atp4*, *nad6* and *mttB,* cleaving the non-conserved regions off the genes. Especially for the *nad6* gene which does not have a conserved ending, mitochondrial mRNA editing might be involved, which was not analysed in the present study.

Mitochondrial genomes are normally illustrated as a circular map, the mastercircle. However their real structure is not really known and expected to be much more variable. Mitochondrial genomes also consist of many repeats that, for example, form 15.2% of *T. aestivum* sequence (Ogihara et al. 2005). Repeats in *T. aestivum* ranged in size from 104 to 9,882 bp (Ogihara et al. 2005). The *L. perenne* mitochondrial genome also consists of many repeats but the largest repeat (found to date) was only 1,151 bp long. It is possible that the approach used for assembling the mitochondrial genome accidentally eliminated larger repeats, or that this genome simply does not contain larger repeats. Mitochondrial genome repeats are of high importance because they are thought to be actively involved in the rearrangement of genome structure. It is believed that recombination between the copies of the repeats occurs that results in smaller subgenomic circles (Ogihara et al. 2005). This recombination is very likely also responsible for the creation of highly divergent intergenic spacers and for the spatial rearrangement of the genes.

To gain further insights into mitochondrial genome rearrangements, the published Poaceae sequences were analysed for their gene order. Only six groups of at least two genes or two exons of different genes were found across nearly all species. The largest gene cluster was found in a region containing *rps1, ccmFN, matR, nad5* exon number 3

and *nad1* exon number 5 (not complete in *T. distachyon* and *B. oldhamii*). Sequence analysis showed that not only the gene order is conserved in that region but also the sequence of intergenic spacer regions is highly conserved (Table 17). However, there might still be enough variation to use these regions for population genetic or phylogenetic studies as shown for the intergenic spacer of *rps1* and *matR*. Generally the intergenic spacer regions of mitochondrial genomes are not suitable for those studies. This is due to the abundance of rearrangements which results in highly divergent intergenic spacers between genomes of different species that in BLASTN analyses are not even recognized as mitochondrial DNA due to their uniqueness (Unseld et al. 1997). For example, Ogihara et al. (2005) determined that 57% of the mitochondrial genome of *T. aestivum* is based on unique sequences. Thus primers can only be designed in coding regions which flank the intergenic spacer regions, but those will be in many cases too far away from each other for easy amplification and would need the application of expensive long-range PCRs combined with complex sequencing approaches. Although the *rps1-matR* intergenic spacer is around 3,500 bp long it is conserved enough to design primers within it and still contains the most variable regions within the cluster comprising around one third of the complete variability found within the intergenic regions of the cluster. Thus it could prove suitable for application in population or even phylogenetic studies.

5.5.2 Horizontal and intracellular gene transfer

The examination of horizontal gene transfer in the *L. perenne* mitochondrial genome was very limited due to the lack of mitochondrial genome sequences from symbiotic and pathogenic organisms. The only publicly available mitochondrial genome sequence belonged to *Glomus intraradices*, a widespread arbuscular mycorrhizal fungus (Lee & Young 2009). Arbuscular mycorrhizal fungi live in symbiosis with plant roots (intracellularly) and can increase their nutrient uptake, especially phosphorus, while the fungus receives carbohydrates from the plant (Allen 1991). The sequence analysis revealed a 322 bp fragment that showed 76% identity between the mitochondrial genome of *L. perenne* and *G. intraradices*. However, further analyses also showed that this fragment is not unique to *Lolium*, but could be found in other Poaceae and even

dicotyledonous species. It was also found in the chloroplast genome where it aligned to the large ribosomal subunit *rrn23*. At this point it can not be excluded that this fragment originates from *Glomus* and was transferred into the ancestor of land plants, or that it was transferred from a land plant species to *Glomus*. However, considering the fact that this sequence can be found in many plant species and in different organelle genomes it might be more likely to assume that the observed similarity between the two fragments is rather due to the function of the gene it was found in rather than due horizontal gene transfer. However, further and more thorough analyses are necessary before the possibility of HGT can be completely excluded. Even more promising than the analyses of HGT between *L. perenne* and *Glomus* would be a HGT analyses including *Neotyphodium* and *L. perenne*.

Most wild European *L. perenne* populations and around a quarter of the European commercial *L. perenne* seed contain by nature the endophytic fungus *Neotyphodium lolii*. *Lolium perenne* provides the fungus mainly with the necessary nutrition for its survival while *L. perenne* in return benefits from this symbiosis by an increase in the growth rate, an increase in the tolerance of abiotic stress and an increase in the resistance against herbivores (reviewed in Lenuweit & Gharadjedaghi 2002). The resistance against herbivores is based on the production of secondary metabolistic chemicals by the fungus such as alkaloids or neurotoxins. The herbivore resistance of infected plants results on the one side in a higher resistance against pathogenic insects, but also leads to two cattle and sheep diseases, known as ryegrass staggers and ryegrass fever. While insect resistance is welcomed by *L. perenne* breeders and farmers, the cause of the diseases in the animal stock is unwanted. However, the economic advantages of infected *L. perenne* seed, because of insect resistance, generally outweighs the economic disadvantages caused by ryegrass staggers or fever; therefore infected seeds are, for example, widely used in New Zealand. Irish accessions generally contain the fungus but the concentration of the alkaloid lolitrem B which causes ryegrass staggers is more than five times lower than in New Zealand *L. perenne* lines (do Valle Ribeiro et al. 1995). Nevertheless a lot of research is carried out to find an optimal way of using this endophyte efficiently for fighting pathogenic insects by causing as little harm as possible to cattle and sheep. Another endophyte, *Epichloe festucae*, has similar effects on *L. perenne* plants and animals as *N. lolii*. Furthermore *E.*

festucae and *N. lolii* are very closely related to each other and it is suspected that *N. lolii* derives from *E. festucae* (reviewed in Lenuweit & Gharadjedaghi 2002)

Nuclear and mitochondrial genome sequences of both fungal species were determined in recent years and provide the perfect opportunity to analyse horizontal gene transfer from the endophyte to the host. However, the sequences are unfortunately not yet publicly available (personal communication with German C. Spangenberg), and therefore this analysis can not be carried out. The draft assembly of the mitochondrial genome of *L. perenne* consists of two copies of the *nad3* gene located on two different contigs that can not be joined together. In case that these two copies of the *nad3* gene are still present in the final assembly of the *L. perenne* mitochondrial genome, this would be an optimal starting point to further investigate HGT in the *L. perenne* mitochondrial genome. As reported by Mower and Palmer (2006) for the *atp1* gene in *Plantago* species and by Keeling and Palmer (2008) for the *nad5* gene in *Amborella*, the presence of several copies of the same gene could indicate that HGT had occurred during evolution. Horizontal gene transfer was observed in *Plantago* species where a pseudo-*atp1* copy was found which probably derived from HGT from a species of the parasitic plant genus *Cuscuta* (Mower and Palmer 2006). Therefore HGT can be also expected to occur between an endophytic fungus such as *N. lolii* and its host *L. perenne*.

While the investigation of HGT in the *L. perenne* mitochondrial genome is currently only to a very limited extent possible, investigations regarding intracellular gene transfer were possible due to the availability of the complete chloroplast genome sequence of *L. perenne* (of this thesis). This study revealed mainly three regions that were transferred from the chloroplast genome to the mitochondrial genome in *L. perenne* during evolution (promiscuous DNA). The first observation of promiscuous DNA was made by Stern and Lonsdale (1982) in maize who reported a chloroplast homologous DNA fragment in the maize mitochondrial genome of 12 kb that comprised the chloroplast genes *rrn16*, *tRNA-Ile* and *tRNA-Val*. A similar fragment (IGT 2) was also found in this study although it was 2 kb shorter and did not contain *tRNA-Ile*. In *T. aestivum* that fragment was even shorter and contained only the genes *ndhB*, *rps7* and *rps12* 3'end (Ogihara et al. 2005). A 12 kb fragment (IGT1) from the large single copy region of the chloroplast genome was also present in the draft assembly of the *L. perenne* mitochondrial genome. This fragment had not been observed previously in any

published mitochondrial genome, although some of the chloroplast genes that were located on this fragment were found seperately in some of the genomes (Clifton et al. 2004, Ogihara et al. 2005). The third fragment found (IGT3) was nearly identical to the fragment described by Watanabe et al. (1994) for Italian ryegrass (*Lolium multiflorum*) and also exists in a similar form in maize (Panicoideae), rice (Ehrhartoideae) and sorghum (Panicoideae) (Watanabe et al. 1994). Another chloroplast fragment frequently found in Poaceae mitochondrial DNA contains the *rbcL* gene. However, this fragment was so far not found in the present draft assembly of *L. perenne*, which is also in agreement with Watanabe et al. (1994) who did not find this fragment in *L. multiflorum*. The fact that many Poaceae genera (of subfamilies Ehrhartoideae, Panicoideae and Pooideae) contain fragment IGT3 is very likely due to the fact that their common ancestor aquired this fragment and it was lost during evolution in *T. aestivum* (Pooideae) (Watanabe et al. 1994). The *rbcL*-fragment has so far not found in Pooideae species which could indicate that it was lost in this subfamily after divergence from the other subfamilies.

Protein-coding genes aquired by mitochondrial genomes via IGT are in general not functional due to frame shifts in the coding regions. Comparing the mitochondrial chloroplast fragments with the respective region on the chloroplast genome, revealed a considerably high amount of SNPs and Indels (data not shown). Indels occurred mainly in microsatellite regions and resulted in frame shifts. However, the sequence assembly showed that GS FLX sequencing does not have strength in determining the accurate sequence in microsatellite rich regions. Therefore, it cannot be said if the observed indels are based on sequencing errors or are indeed existing indels that cause frame shifts and make protein-coding regions disfunctional. Considering that the chloroplast derived protein-coding genes are not needed for the proper function of mitochondria it could be that slipped strand mispairing that causes these possible indels is not the target of a repair mechanism of the mitochondrial genome. Transferred ribosomal genes from those regions seem in general to be functional, indicated by their capability of forming a cloverleaf like structure. This observation is also in agreement with earlier studies and the fact that mitochondrial genomes do not encode all the necessary transfer ribosomal genes themselves (Unseld et al. 1997).

Although the complete sequence of the *L. perenne* mitochondrial genome is not yet established, the amount of chloroplast derived sequences can be estimated already. The three major IGT sequences observed in this study account for in total 24,541 bp. Considering that the *L. perenne* mitochondrial genome is very likely to have a length between 490,381 bp to 564,927 bp the amount of chloroplast derived sequences can be estimated to be in a range of four to five per cent, depending on the final sequence length and the amount of smaller chloroplast DNA derived fragments. The observed amount of chloroplast derived DNA fragments varies in the so far sequenced mitochondrial genome from 1.1% (*Arabidopsis thaliana*) to 8.8% (*Vitis vinifera*) (Unseld et al. 1997, Sugiyama et al. 2005, Goremykin et al. 2008). The amount of chloroplast derived sequences within Poaceae genomes ranges from 3.0% in *T. aestivum* to 6.2% in *Oryza sativa*. Therefore the estimated range of four to five per cent for *L. perenne* is reasonable.

Intracellular gene transfer also occurs from the nucleus to the mitochondrial genome mainly in form of transposable and retrotransposable elements and can account for 4.4% (*Arabidopsis thaliana* (Unseld et al. 1997); 1.4% *Oryza sativa* (Ogihara et al. 2005)) of the mitochondrial genome. During the assembly process some similarity hits were found to a repetitive probably retrotransposable sequence found in *L. perenne* (Acc. no. AF063225). However, searches of the final 43 contigs for this sequence and other transposable and retrotransposable elements known for *L. perenne* and *L. multiflorum* did not reveal any of those elements in these contigs anymore. This can be due to the fact that these sequences which occur in high copy number in the nuclear genome were contaminating DNA sequences and were extinguished by the final assembly approach chosen. However it is also possible that they are among the sequences which did not contain any protein-coding genes and were smaller than 10kb. As shown for the protein-coding sequences, these sequences although smaller in length can not be underestimated regarding their importance for a complete assembly of the mitochondrial genome. Although they do not contain any of the major coding information they might contain nuclear elements which are generally less than 500 bp large and future studies might reveal them.

5.5.3 CMS in *Lolium perenne*

The *L. perenne* mitochondrial genome used in this study derived from the fertile cultivar Shandon. However, in 1984 Connolly and Wright-Turner induced cytoplasmic male sterility in *L. perenne* lines from the Teagasc Oak Park grass collection via intergeneric hybridization of *Festuca pratensis* with *L. perenne*. Cytoplasmic male sterility is of special value to breeders for the easy and inexpensive production of hybrid seed. The CMS mechanism is generally believed to be based on the mitochondrial genome but also the chloroplast genome might be involved in this complex system (Chen et al. 1995). Often the presence of chimeric mitochondrial open reading frames or chimeric transcripts could be linked to CMS.

The *atp6* gene, for example, is believed to be associated with CMS in *L. perenne* (Rouwendale et al. 1992) and other species. In this study non-conserved pre- and post-sequences for *atp6* and three other genes (*atp4*, *nad6, mttB*) were observed. Yamamoto et al. (2005) showed using sugar beet mitochondrial genomes that the *atp6* pre-sequence is translated, associated with the mitochondrial membrane and part of a unique protein complex in the mitochondrial genome sugar beet CMS-lines. The *atp6* region is known to be regularly involved in recombination events (Yamamoto et al. 2005). Also the genes *atp4*, *nad6* and *mttB* are not located within one of the few conserved regions within the Poaceae mitochondrial genome. Thus the non-conserved pre- and post-sequences of all four genes evolved possibly by genome rearrangements. Those rearrangements can lead to functional chimeric genes that can be involved in CMS (Schnable & Wise 1998). Thus these four regions might be of special interest if the *L. perenne* CMS-system should be investigated in the future.

In *L. perenne*, both organelle genomes are in general maternally inherited. However, surprisingly all the CMS lines derived from the intergeneric hybridization in which *F. pratensis* was used as maternal parent showed the paternal *L. perenne* mitochondrial genome (Kiang et al. 1993). Analysing the mitochondrial genome of these CMS lines and comparing it with the mitochondrial genome of fertile lines revealed a major difference -the absence of a 5.6 kb HindIII fragment from the CMS lines. Further analyses of that region revealed the replacement of a 1.8 kb fragment with a 9.6 kb fragment in the CMS line (Kiang & Kavanagh 1996). Later on analyses revealed that

this 9.6 kb fragment is a linear plasmid that was incorporated into the mitochondrial genome of the CMS line in close proximity to the gene *atp9* (McDermott et al. 2008). It is suspected that the integration of this linear plasmid caused the cytoplasmic male sterility in the *L. perenne* lines. Therefore it is expected that this plasmid although very likely present in the mitochondrion of *L. perenne* cv. Shandon is not integrated in its mitochondrial genome. Sequence analysis revealed the gene *atp9* on contig 86+11+87 of this study; blasting this contig against the plasmid sequence from the CMS line (Acc.-No. AM998372) shows that nearly the complete plasmid sequence can be found on that contig. Only 800 bp are missing in contig 86+11+87 that are present in the plasmid sequence. Thus part of the *rpo*-like sequence that is commonly found in linear mitochondrial plasmids is missing. However, searching the Lasergene pre-assembly showed that the coding sequence for the complete *rpo*-like open reading frame exists. Hence it is possible that either the contig 86+11+87 is not completely or correctly assembled or parts of the *rpo*-like sequence are located elsewhere in the mitochondrial genome of *L. perenne*.

Furthermore the presence of the plasmid like sequence within the mitochondrial genome sequence of the fertile *L. perenne* cv. Shandon does not mean that this sequence is integrated *in vivo* into its genome. The mitochondrial DNA that was used in the present study was isolated from complete mitochondria thus also DNA of possibly non-integrated linear plasmids was isolated. Given that the plasmid is able to be integrated into the mitochondrial genome of the CMS line indicates that there are specific sites in the DNA sequence that enable this integration. It is likely that these sites exist also in fertile lines (if not with the exact sequence but maybe with a high similarity). Therefore when assembling the mitochondrial genome the assembly programs might automatically integrate the linear plasmid sequence into the mitochondrial DNA sequence. The fact that the plasmid sequence was found on only a single contig and that it was not possible to join this contig with further contigs, could indicate that in this plasmid is not integrated into the mitochondrial genome. However, final conclusions can not be drawn until the complete mitochondrial genome of *L. perenne* is assembled.

5.6 Conclusion

The mitochondrial genome of *L. perenne* was sequenced, and partially assembled and annotated. The complex nature of plant mitochondrial genomes prohibited the complete assembly and annotation due to time restrictions of the project. Nevertheless, the availability of sequence information for the mitochondrial genome is of great value as *Lolium* is one of the most important forage grasses of the northern hemisphere. Only 24 plant mitochondrial genome sequences are publicly available to date and although ten of them are from grass species (with a strong bias to Panicoideae species) no forage grass species had until now been sequenced. Furthermore, the number of published articles on sequenced genomes is very restricted and thus background information to the published mitochondrial genomes only available for *O. sativa* (Nutso et al. 2002), *T. aestivum* (Ogihara et al. 2005) and *Z. mays* (Clifton et al. 2004). Although the final order of mitochondrial genes within the *L. perenne* mitochondrial genome could not be established, conserved regions across many Poaceae species that might be suitable for future population and phylogenetic studies were detected. During the assembly of the *L. perenne* mitochondrial genome, microsatellite regions could be detected but were not further analysed in this study. These regions might be of interest in future studies as additional tools for population and phylogenetic studies. Also insights into the extent of intracellular gene transfer from the chloroplast genome to the mitochondrial genome were gained. Once mitochondrial genome sequences of *L. perenne* endophytes become available it will be interesting to see to what extent horizontal gene transfer can be observed in *L. perenne*. It will be also interesting to analyse the complete mitochondrial genome for further open reading frames that might be involved in the CMS-mechanism with special emphasis on the role of the linear plasmid sequence described by McDermott et al. (2008) in the mitochondrial genome of a fertile *L. perenne* cultivar.

Chapter 6

Applicability of organelle sequence information to phylogenetic studies

"The success of phylogenetic analysis depends upon the discovery of character sets that provide accurate information regarding phylogenetic relationships among natural lineages. "
Davis et al. (1998)

6.1 Introduction

6.1.1 Evolution of phylogenetic studies

Phylogenetic studies were, until the beginning of the 1970s, mainly based on morphological and biochemical data. In 1971 type II (specifically cutting) restriction enzymes were discovered (Danna & Nathans 1971) and restriction site mapping from DNA fragment patterns became another phylogenetic tool. Shortly after, Sanger et al. (1977) and Maxam and Gilbert (1977) presented reliable sequencing techniques which were quickly utilized for phylogenetic studies. Phylogenetic studies based on restriction site mapping and sequencing techniques were, from then on, called molecular phylogenetic analyses.

Generally, DNA from the genomes of all three plant organelles (nucleus, chloroplasts, mitochondria) is suitable for these studies. However, the chloroplast genome has several advantages due to its a) relatively small size, b) conserved gene content and order, c) slow evolution, d) low substitution rate, e) uniparental inheritance and f) lack of recombination (Palmer et al. 1988, Wolfe et al. 1987). The mapping of chloroplast restriction sites was one of the first tools used for plant molecular phylogenetic studies to assess the relationship of different lineages based on rearrangements of the chloroplast genome. Palmer and Zamir (1982) demonstrated that the mapping of restriction sites is a valuable approach for investigating relationships of *Lycopersicon* (tomato) species. In an additional step, genes can be hybridized to restriction fragments

and thus facilitate the detection of rearrangements between the chloroplast genome of different species as demonstrated by Knox et al. (1993) for species of the genus *Lobelia*. The first molecular phylogenetic study based on sequencing was carried out on a single chloroplast gene, *rbcL* (the large subunit of ribulose biphosphae carboxylase), for which sufficient sequencing data had been available at that time (Palmer et al. 1988). Chase et al. (1993) published a landmark paper on seed plant phylogenetics using *rbcL* and over 500 species. This was followed by several other general seed plant studies (Soltis et al. 1999; Qiu et al. 1999; Savolainen & Chase 2003) using plastid, nuclear and mitochondrial genes.

With an increase in published chloroplast genomes since 2003 the possibility arose for more complex analyses involving whole genomes instead of single genes (or combinations of a few genes). One of the first comprehensive studies was based on 61 chloroplast genes that were common in 13 plant species (Goremykin et al. 2003), followed shortly after by several other studies based on the same genes but 20 (Chang et al. 2006) and 24 plant species (Leebens-Mack et al. 2005) all trying to discover phylogenetic pattern among the phylogenetically most outlying angiosperm species. Since 2006, approximately 20 – 30 chloroplast genomes have been published per year, thus the number of genomes within different clades and families has increased and enabled more informative phylogenetic analysis of complete chloroplast genomes (Bortiri et al. 2008).

Hebert et al. (2003) proposed the establishment of a standard system for facilitating and enhancing the identification of species. They suggested basing this system on DNA sequences that serve as species identifying 'barcodes' similar to the barcodes used in supermarkets for identifying products. Tautz et al. (2003) also advocated the barcoding approach but it has also been the subject of much controversy (Moritz & Cicero 2004). Consequently the term 'DNA barcoding' was introduced. 'DNA barcoding' describes an approach of large-scale screening of one or few reference genes to identify unknown individuals and to discover new species (Hebert et al. 2003). The mitochondrial *coi*-gene was recommended as a suitable DNA barcode for animals (Hebert et al. 2003); however plant mitochondrial gene sequences are highly conserved and are not generally suitable for DNA barcoding. Therefore plant taxonomists focussed on chloroplast genes and intergenic spacers as barcodes for plants. Genes and intergenic spacers promoted as

suitable candidates for plant DNA barcode include: *trnH-psbA* intergenic spacer (Kress et al. 2005), *rbcL* (Kress & Erickson 2007), *matK*, *rpoC1*, *rpoB* (Chase et al. 2007), *psbK-psbI* and *atpF-atpH* intergenic spacers (Kim et al. (reviewed in Pennisi 2007)). In 2009, the CBOL (Consortium for the Barcode of Life) Plant Working Group proposed a combination of the chloroplast genes *rbcL* and *matK* as barcode for plants. This gene combination was able to distinguish 72% of a test set of 550 species and could sort the remaining ones into congeneric groups with 100% success (CBOL Plant Working Group 2009).

Although chloroplast DNA is a very powerful source of variation for molecular systematics its use can sometimes be limited to higher taxonomic levels due to its slow evolutionary rate (Wolfe et al. 1987). It sometimes does not provide enough sequence divergence for studies on lower taxonomic levels. At these levels nuclear DNA proved to be of great value, especially the ribosomal gene cluster 18S-5.8S-26S and its internal transcribed spacers (ITS) (Baldwin et al. 1995, Bult et al. 1995; Hodkinson et al. 2002). The 18S-26S ITS region is highly repeated in the nuclear genome, the genes evolve rapidly and in concert and the region is generally small (<700 bp) with conserved sequences flanking both ITS and facilitating the use of universal primers for amplification across species (reviewed in Baldwin et al. 1995). But also the use of the ITS region is restricted. It is not suitable for analysing high taxonomic levels, because it is too diverged to align the sequences, but it is sometimes also too conserved for analysing very low taxonomic levels e.g. closely related species or intraspecific relationships because it is too conserved (Baldwin et al. 1995; Hodkinson et al. 2002). Analysing those low taxonomic levels can be facilitated by using rapid evolving introns of low-copy nuclear genes such as *adh* (alcohol dehydrogenase) and *pgiC* (cytosolic phosphoglucose isomerise) genes (reviewed in Sang 2002).

In contrast to chloroplast and nuclear DNA, the third organelle DNA type, mitochondrial DNA, has received less attention for plant molecular phylogenetic analyses. While mitochondrial DNA is widely used in animal molecular phylogenetic analyses (Hebert et al. 2004), plant mitochondrial DNA is rarely used for phylogenetic analyses due to its complex structure compared to animal mitochondria. Animal mitochondria, like chloroplasts, are rather conserved and of small genome size (10 – 45 kb). Plant mitochondrial genomes are two to five times larger than chloroplast genomes

(200 – 800 kb), contain around two and half times less genes and consist of large intergenic spacers. The gene order of plant mitochondrial DNA is less conserved than plastid DNA due to frequent rearrangements and so far no genomes have been sequenced with the same gene order (Palmer et al. 1992, 2000). The intergenic spacers vary greatly regarding length and sequence even across closely related species. Hence universal primer design to amplify and phylogenetically analyse non-coding plant mitochondrial regions has generally not been successful. While the non-coding regions are highly variable the coding regions are highly conserved and have much lower nucleotide substitution rates than the nuclear and also the chloroplast genomes (Wolfe et al. 1987). Therefore molecular phylogenetic analyses on low taxonomic levels are not promising (Palmer et al. 1992). However analyses on high taxonomic levels are possible and useful for several applications such as to identify the free-living relatives of parasitic plants (Barkman et al. 2007). Plant mitochondrial genome sequences are also of high value in combination with nuclear and chloroplast sequences when exploring the question about the relationships of major seed plant groups and the earliest angiosperms (Qiu et al. 1999; Savolainen and Chase, 2003). Only a few mitochondrial genes have been used so far in plant phylogenetics. Many genes of the mitochondrial genome are too short (less than 500 bp), trans-spliced and thus not easy to amplify or they are simply too conserved. Mitochondrial genes that have been used in phylogenetic studies are *atp1* (Qiu et al. 1999), *matR* (Qiu et al. 1999), *nad1* exons b and c (Bakker et al. 2000) and *rrn19* (Duff & Nickrent 1999).

6.1.2 Phylogenetic studies in Poaceae

The grass family Poaceae consists of thirteen subfamilies (Anomochlooideae, Aristidoideae, Arundinoideae, Bambusoideae, Centothecoideae, Chloridoideae, Danthonioideae, Ehrhartoideae, Micrairoideae, Panicoideae, Pharoideae, Pooideae, Puelioideae; GPWG 2001, reviewed in Bouchenak-Khelladi et al. 2008) and around 900 genera containing approximately 10,000 species (Tzvelev 1989). Poaceae include some of the world's most important crops such as cereals, turf and forage species, sugarcane, *Miscanthus*, reeds and bamboos. Many phylogenetic studies have been carried out to shed light on the evolution of this family. Molecular phylogenetic studies in the grass family included the analysis of chloroplast restriction sites, of chloroplast genome

regions and of nuclear genome regions. The first large-scale chloroplast genome restriction analyses was carried out in 1993 with 31 species (Davis & Soreng 1993) and extended in 1998 to include 72 species (Soreng & Davis 1998). For single chloroplast gene analyses several different genes were used such as *matK* (Hilu et al. 1999), *ndhF* (Clark et al. 1995), *rbcL* (Barker 1997), *rpoC2* (Cummings et al. 1994) and *rps4* (Nadot et al. 1995), but also non-coding regions of the chloroplast genome e.g. the intron and intergenic spacer of the *trnL-trnF* region (Bouchenak-Khelladi et al. 2008) or the *rpl16* intron (Zhang 2000) proved to be suitable for assessing the relationship of the different Poaceae subfamilies. A comprehensive study using nuclear genome sequences was carried out by Hsiao et al. (1999) based on the ITS region (reviewed in GPWG 2001). Only one year later the first detailed nuclear gene study carried out in Poaceae was published and based on the gene phytochrome B (*phyB*) (Mathews et al. 2000).

In 1996 the Grass Phylogeny Working Group (GPWG) was founded. The objectives of the working group were to combine a series of existing data sets to produce a comprehensive phylogeny of the grass family, to help focus the further development of existing data sets and to re-evaluate the sub-familial classification (GPWG 2001). Information obtained from morphological data was also included. The major results of the GPWG were: a) Poaceae are monophyletic; b) Anomochlooideae, Pharoideae and Puelioideae, respectively are the oldest lineages to diverge from the rest of the grasses; c) all remaining subfamilies form one clade; d) each of the remaining subfamilies is monophyletic; e) the PACCAD clade (Panicoideae, Arundinoideae, Chloridiodeae, Centothecoideae, Aristidoideae and Danthonioideae (since 2007 PACCMAD clade – M = Micrairoideae as reviewed in Bouchenak-Khelladi et al. 2008) is monophyletic. The work provided a general, common and accepted basis for future phylogenetic studies to enable better resolution of the complete Poaceae tree.

6.1.3 Aims and objectives

In September 2009, 19 chloroplast genomes of monocotyledonous species were publicly available. Two species belonged to the order Acorales, one to the order Alismatales, one to Asparagales and one to Dioscoreales; but fourteen genomes belonged to species of the family Poaceae. These fourteen species belonged to four different subfamilies and

represented seven different tribes (Bambusoideae – Bambuseae; Ehrhartoideae – Oryzeae; Panicoideae – Andropogoneae; and Pooideae – Aveneae, Brachypodieae, Poeae, Triticeae). Also in September 2009 complete mitochondrial genomes of eleven Poaceae species (including the partially assembled but likely completely sequenced *Lolium perenne* mitochondrial genome of this thesis) belonging to three different subfamilies and five different tribes (Bambusoideae – Bambuseae; Ehrhartoideae – Oryzeae; Panicoideae – Andropogoneae; and Pooideae – Poeae, Triticeae) were available. The availability of these chloroplast and mitochondrial genomes within one plant family presented the possibility to analyse their applicability to grass phylogenetic studies (no other family is so well represented in terms of complete genomes available in GenBank). While Goremykin et al. (2003) showed that chloroplast protein coding genes are generally able to infer the phylogenetic relationship of angiosperm species, Bortiri et al. (2008) also showed that complete chloroplast genomes can be used for grass phylogenetic studies. Neither of both research groups focussed only on all the protein-coding genes of the Poaceae family. A comprehensive study based on all protein coding mitochondrial genomes has not been published to date.

Thus the aims and objectives of this study were generally to analyse the applicability of protein-coding chloroplast and mitochondrial genes of Poaceae species to phylogenetic studies. In detail the objectives were a) to characterize structural differences in the chloroplast genome of Poaceae and non-Poaceae species, b) to evaluate the application of single chloroplast protein coding genes for studies within the Poaceae subfamilies Bambusoideae, Ehrhartoideae, Panicoideae and Pooideae and to gain insight into their evolution, c) to evaluate the application of all chloroplast protein coding genes in phylogenetic studies within the Poaceae and d) to evaluate the application of all mitochondrial protein coding genes in phylogenetic studies in the Poaceae.

6.2 Material and methods

6.2.1 Structural characterization of monocotyledonous chloroplast genomes

The gene order of protein coding genes from Poaceae chloroplast genomes and non-Poaceae chloroplast genomes (listed in Table 18) were compared. Differences were

illustrated by manually highlighting divergent regions on the circular maps of *L. perenne* and *Dioscorea elephantipes* as representatives of the Poaceae and non-Poaceae species, respectively.

6.2.2 Phylogenetic analyses

The phylogenetic analyses were based on 76 chloroplast protein-coding genes (Table 19) from 19 different monocotyledonous species including 14 grass species (Table 18), and on 32 mitochondrial protein-coding genes (Table 20) from 13 different species (Table 18), including eleven grass species. All non-Poaceae species in the chloroplast gene analyses served as outgroups, while in the mitochondrial gene analyses two dicotyledonous species were used, due to a lack of information (sequenced genomes) on non-Poaceae monocotyledonous species. For analysing the different data sets, Maximum Parsimony (combined with parsimony bootstrapping) and Bayesian inference approaches were chosen.

Table 18 List of species and accession numbers for the phylogenetic analyses using chloroplast and mitochondrial genes. M=monocotyledonous, D=Dicotyledonous, **P=Poaceae,** indicated in colours are the different Poaceae subfamilies: **B**ambusoideae, **E**hrhartoideae, **Pa**nicoideae, **Po**oideae

			chloroplast accession number	mitochondrion accession number
Acorus americanus	M		EU273602	
Acorus calamus	M		AJ879453	
Agrostis stolonifera	P	Po	EF115543	
Arabidopsis thaliana	D			Y08501
Bambusa oldhamii	P	B	FJ970915	EU365401
Brachypodium distachyon	P	Po	EU325680	
Dendrocalamus latiflorus	P	B	FJ970916	
Dioscorea elephantipes	M		EF380353	
Festuca arundinacea	P	Po	FJ466687	
Hordeum vulgare	P	Po	EF115541	
Lemna minor	M		DQ400350	
Lolium perenne	P	Po	AM777385	not available yet
Nicotiana tabacum	D			BA000042
Oryza nivara	P	E	AP006728	
Oryza sativa 'indica-group'	P	E	AY522329	DQ167399
Oryza sativa 'japonica-group'	P	E	X15901	BA000029
Phalaenopsis aphrodite	M		AY916449	
Saccharum officinarum	P	Pa	AP006714	
Sorghum bicolor	P	Pa	EF115542	DQ984518
Tripsacum dactyloides	P	Pa		DQ984517
Triticum aestivum	P	Po	AB042240	AP008982
Zea luxurians	P	Pa		DQ645537
Zea mays	P	Pa	X86563	
Zea mays mays	P	Pa		AY506529
Zea mays parviglumis	P	Pa		DQ645539
Zea perennis	P	Pa		DQ645538

Maximum Parsimony methods search for the tree with the minimum amount of evolutionary changes that can explain the sequence evolution and in doing so maximise the congruence between characters (reviewed in Takahashi & Nei 2000). Generally more than one most-parsimonous tree will be found and for estimating the support of the clades, bootstrapping is carried out. The term "bootstrapping" is a statistical method which is based on resampling the data randomly to create pseudoreplicate datasets that are then each analysed and summarized. While the original Maximum Parsimony analysis will consider every character column in the data set once, bootstrapping chooses randomly columns and might duplicate some of the characters and loose others in the sampling process. This approach is based on the assumption that each character evolves independently and contains information about the underlying phylogeny (Felsenstein 1985). Generally the bootstrapping process is repeated several times and a majority-rule consensus tree is built from the results of each run. On the majority-rule consensus tree it will be recorded in per cent how often each of the bootstrap subsets appeared, thus giving an estimate of the reliability of the branching order (Felsenstein 1985).

Bayesian inference analysis is based on the Bayes' theorem. For the application of this formula an assumption must be made (prior probability) and at the end of the calculation it will be said how likely or not likely this assumption was (posterior probability) (Huelsenbeck et al. 2002). For calculating the posterior probability of a specific tree all possible trees have to be taken into account. However, this is due to the computational effort generally not possible, therefore the posterior probabilities for the trees must be estimated. This is done by using a Marcov Chain Monte Carlo (MCMC) model. In a first step to start the chain a tree is chosen. Then in a second step based on few criteria (Huelsenbeck et al. 2002) another new tree is proposed. The third step accepted the new proposed tree with probability R (= the proposed state becomes the next state of the chain). In a fourth step a uniform random number will be created on the interval, if this number is smaller than R the new state is accepted. The fifth step is to go back to step two and start over again. The steps two to five are repeated many times. The fraction of time that the chain finds a specific tree is a valid estimate of the posterior probability of the tree (Huelsenbeck et al. 2002). One advantage of this method of phylogenetic inference is that is generates most likely trees and assesses their robustness (in terms of posterior probabilities) in the same analysis. Parsimony and

Maximum Likelihood approaches require separate analyses to generate the trees and assess their robustness (via bootstrapping or similar approaches such as jacknifing; see Salamin et al. 2003).

6.2.2.1 Chloroplast single gene analysis

The genes of the different species were individually aligned in Mega3.1 (Kumar et al. 2004) and then transferred into a Nexus-file format. The individual chloroplast genes were first all analysed using Branch and Bound in PAUP* 4.0 (Swofford 2002) (settings: **Search** = BandB; **NRep** (replicates) = 1000). Genes that revealed one of the subfamilies as sister to the three other subfamilies were further analysed in MrBayes (Huelsenbeck & Ronquist 2001, Ronquist & Huelsenbeck 2003) (settings: **nst** = 6, **rates** = invgamma (= GTR+G+I-model (**G**eneral **T**ime **R**eversible model + **G**amma-distributed rate); **ngen** = 1,000,000, **samplefreq** = 100; **burnin** = 500). For generating trees the program FigTree v1.2.1 2006-2009 (http://tree.bio.ed.ac.uk/) was used with MrBayes (Huelsenbeck & Ronquist 2001, Ronquist & Huelsenbeck 2003) tree files as input file.

6.2.2.2 Combined chloroplast and mitochondrion gene analyses

The Nexus files from the 76 chloroplast genes from the chloroplast single gene analysis were combined in an interleaved format. The 32 mitochondrial genes of the different species were individually aligned in Mega3.1 (Kumar et al. 2004) and also transferred into a Nexus-file also with an interleaved format. The analyses were carried out in PAUP* 4.0 usinig Branch and Bound (Swofford 2002) (settings: **Search** = BandB (Branch & Bound); **NRep** (replicates) = 1000) and MrBayes (Huelsenbeck & Ronquist 2001, Ronquist & Huelsenbeck 2003) (settings: **nst** = 6, **rates** = invgamma (= GTR+G+I-model (**G**eneral **T**ime **R**eversible model + **G**amma-distributed rate); **ngen** = 2,000,000, **samplefreq** = 100; **burnin** = 1,000/ 500 (chloroplast genes/mitochondria genes). The MrBayes (Huelsenbeck & Ronquist 2001, Ronquist & Huelsenbeck 2003) tree files were used as input files for the program FigTree v1.2.1 2006-2009 (http://tree.bio.ed.ac.uk/) for displaying the trees.

Table 19 Seventy-six protein coding chloroplast genes used for analysing the phylogenetic relationship of 19 monocotyledonous species.
PIS = parsimony informative sites; BP = bootstrap value; PP = Posterior propability; Panicoideae, Ehrhartoideae, or Pooideae are outgroup to rest, respectively.

gene	Synonym	Product Name	Length	PIS	BP	PP
atpA		ATP synthase CF1 alpha subunit	1524	271	78	0.79
atpB		ATP synthase CF1 beta subunit	1500	263	60	0.53
atpE		ATP synthase CF1 epsilon subunit	414	104		
atpF		ATP synthase CF0 B subunit	580	110		
atpH		ATP synthase CF0 C subunit	246	27		
atpI		ATP synthase CF0 A subunit	744	120		
ccsA	*ycf5*	cytochrome c biogenesis protein	996	267	51	
cemA	*ycf10*	envelope membrane protein	693	184		
clpP		ATP-dependent Clp protease proteolytic subunit	651	163		
infA		translation initiation factor 1	342	79		
matK		maturase K	1614	551	93	0.95
ndhA		NADH dehydrogenase subunit 1	1116	236		
ndhB		NADH dehydrogenase subunit 2	1548	61		
ndhC		NADH dehydrogenase subunit 3	366	59	51	
ndhD		NADH dehydrogenase subunit 4	1509	304	53	
ndhE		NADH dehydrogenase subunit 4L	306	51		
ndhF		NADH dehydrogenase subunit 5	2274	616		
ndhG		NADH dehydrogenase subunit 6	534	117		
ndhH		NADH dehydrogenase subunit 7	1194	223		
ndhI		NADH dehydrogenase subunit I	546	115		
ndhJ		NADH dehydrogenase subunit J	480	80		
ndhK		NADH dehydrogenase subunit K	726	132		
petA		cytochrome f	966	166		
petB		cytochrome b6	705	91		
petD		cytochrome b6/f complex subunit IV	570	78	81	0.99
petG		cytochrome b6/f complex subunit V	114	13		
petL	*ycf7*	cytochrome b6/f subunit VI	96	17		
petN	*ycf6*	cytochrome b6/f complex subunit VIII	90	9		
psaA		photosystem I P700 chlorophyll a apoprotein A1	2256	297		
psaB		photosystem I P700 chlorophyll a apoprotein A2	2220	284	87	0.99
psaC		photosystem I subunit VII	246	43		
psaI		photosystem I subunit VIII	111	17	51	0.64
psaJ		photosystem I subunit IX	136	22		
psbA		photosystem II protein D1	1062	117	54	0.99
psbB		photosystem II 47 kDa protein	1536	206		
psbC		photosystem II 44 kDa protein	1389	195	61	0.58
psbD		photosystem II protein D2	1062	139	65	
psbE		photosystem II protein V	252	24		
psbF		photosystem II protein VI	120	15		

Locus	Synonym	Product Name	Length	PSI	BP	PP
psbH		photosystem II protein H	222	50	78	0.92
psbI		photosystem II protein I	111	17		
psbJ		photosystem II protein J	132	17		
psbK		photosystem II protein K	195	42		
psbL		photosystem II protein L	180	8		
psbM		photosystem II protein M	105	15		
psbN		photosystem II protein N	132	11		
psbT		photosystem II protein T	117	16		
psbZ	*ycf9*	photosystem II protein Z	189	29	74	0.54
rbcL		ribulose-1,5-bisphosphate carboxylase/ oxygenase large subunit	1473	261		
rpl2		ribosomal protein L2	825	64	64	0.81
rpl14		ribosomal protein L14	372	68	69	
rpl16		ribosomal protein L16	453	86	77	0.70
rpl20		ribosomal protein L20	369	92	74	0.51
rpl22		ribosomal protein L22	321	105	53	
rpl23		ribosomal protein L23	282	40		
rpl32		ribosomal protein L32	210	55		
rpl33		ribosomal protein L33	207	53		
rpl36		ribosomal protein L36	114	24		
rpoA		RNA polymerase alpha subunit	1041	232	88	0.60
rpoB		RNA polymerase beta subunit	3261	610		
rpoC1		RNA polymerase beta' subunit	2076	438	86	0.64
rpoC2		RNA polymerase beta'' chain	4788	1186	63	
rps2		ribosomal protein S2	711	143	82	
rps3		ribosomal protein S3	726	187	54	
rps4		ribosomal protein S4	609	120	58	0.78
rps7		ribosomal protein S7	471	38		
rps8		ribosomal protein S8	411	85		
rps11		ribosomal protein S11	441	96		
rps12		ribosomal protein S12	375	34	63	0.53
rps14		ribosomal protein S14	312	65		
rps15		ribosomal protein S15	288	59	51	0.60
rps16		ribosomal protein S16	285	54		
rps18		ribosomal protein S18	537	90		
rps19		ribosomal protein S19	285	64	64	0.57
ycf3		photosystem I assembly protein Ycf3	519	73		
ycf4		photosystem I assembly protein Ycf4	576	103		

Table 20 Thirty-two protein coding mitochondrial genes used for analysing the phylogenetic relationship of eleven monocotyledonous and one dicotyledonous species. PIS = parsimony informative sites.

gene	Synonym	Product name	length	PIS
atp1		ATPase subunit 1	1542	111
atp4	*orf25*	ATPase subunit 4	675	94
atp6		ATPase subunit 6	804	91
atp8	*orfB*	ATPase subunit 8	480	56
atp9		ATPase subunit 9	261	23
ccmB		cytochrome c biogenesis B	621	30
ccmC		cytochrome c biogenesis C	771	28
ccmFC		cytochrome c biogenesis FC	1338	91
ccmFN		cytochrome c biogenesis FN	1893	101
cob		apocytochrome b	1197	45
cox1		cytochrome c oxidase subunit 1	1605	86
cox2		cytochrome c oxidase subunit 2	789	59
cox3		cytochrome c oxidase subunit 3	798	22
matR		maturase related protein	2043	139
mttB	*orfX*	transport membrane protein	825	59
nad1		NADH dehydrogenase subunit 1	978	29
nad2		NADH dehydrogenase subunit 2	1467	42
nad3		NADH dehydrogenase subunit 3	357	15
nad4		NADH dehydrogenase subunit 4	1488	64
nad4L		NADH dehydrogenase subunit 4L	303	17
nad5		NADH dehydrogenase subunit 5	2013	67
nad6		NADH dehydrogenase subunit 6	1149	130
nad7		NADH dehydrogenase subunit 7	1185	23
nad9		NADH dehydrogenase subunit 9	573	25
rpl16		ribosomal protein L16	540	24
rps1		ribosomal protein S1	555	47
rps2		ribosomal protein S2	615	96
rps3		ribosomal protein S3	1736	106
rps4		ribosomal protein S4	1245	111
rps7		ribosomal protein S7	447	2
rps12		ribosomal protein S12	378	26
rps13		ribosomal protein S13	351	5

6.3 Results

6.3.1 Structural characterization of monocotyledonous chloroplast genomes

A comparison of the chloroplast genomes of different monocotyledonous species revealed that the protein-coding gene content among them is highly conserved apart from a few exceptions. All species belonging to the Poaceae have lost three genes: *accD*, *ycf1* and *ycf2*. Genes have also been lost in the non-Poaceae monocotyledonous species: *accD* in *Acorus calamus* and *A. americanus*, *rps16* in *Dioscorea elephantipes*, *infA* in *Lemna minor* and the complete complex of *ndh*-genes in *Phalaenopsis aphrodite* (Figure 43). The chloroplast gene order among different monocotyledonous species is highly conserved.

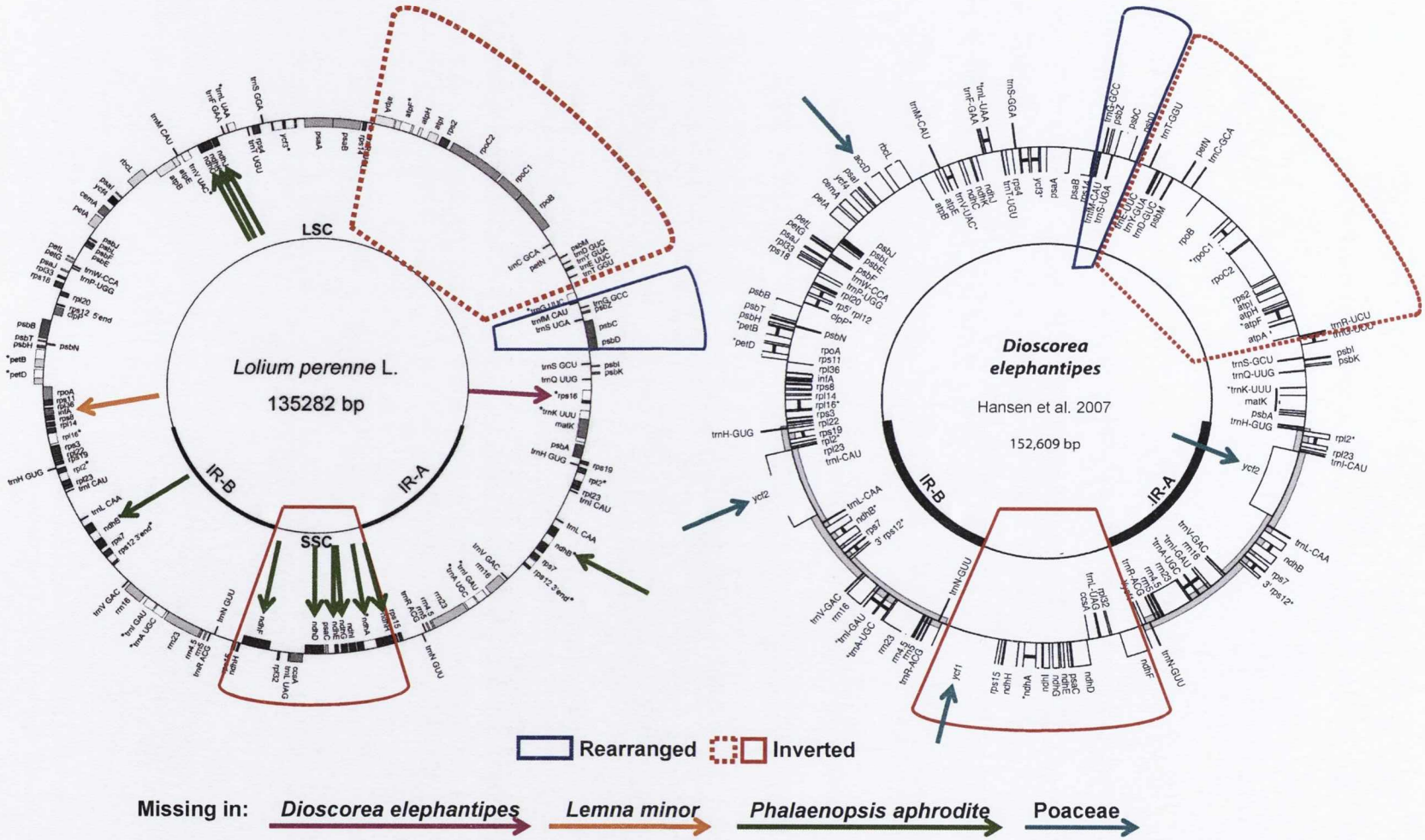

Figure 43 Comparison of the chloroplast genome structure of Poaceae and non-Poaceae species using *Lolium perenne* (Poaceae) and *Dioscorea elephantipes* (non-Poaceae) for the illustration of differences. Arrows point to the name of the absent gene.

Only three major rearranged regions were found: 1.) the single copy region of *Dioscorea elephantipes* had an inverted gene order in comparison to all other species; 2.) a block of 16 genes was inverted in the Poaceae family compared to the non-Poaceae species and 3.) a block of six genes had been rearranged in the Poaceae to a different position in the genome (Figure 43). Figure 44 illustrates gene losses and rearrangements on a cladogram obtained from the analysis of all protein-coding chloroplast genes (s. 6.3.2.2 Chloroplast multigene analysis). Considering the most parsimonious solution it is more likely that genes were in general lost in the different lineages instead of gained. Only the gene *accD* could have been possibly lost in all lineages and then independently gained in *L. minor*, *D. elephantipes* and *P. aphrodite* due to its absence in *Acorus*. Nevertheless, the absence of *accD* in some lineages is considered as loss, because it seems unlikely that a complete gene can be gained, taking into account that there is in general no intracellular gene transfer to the chloroplast genome.

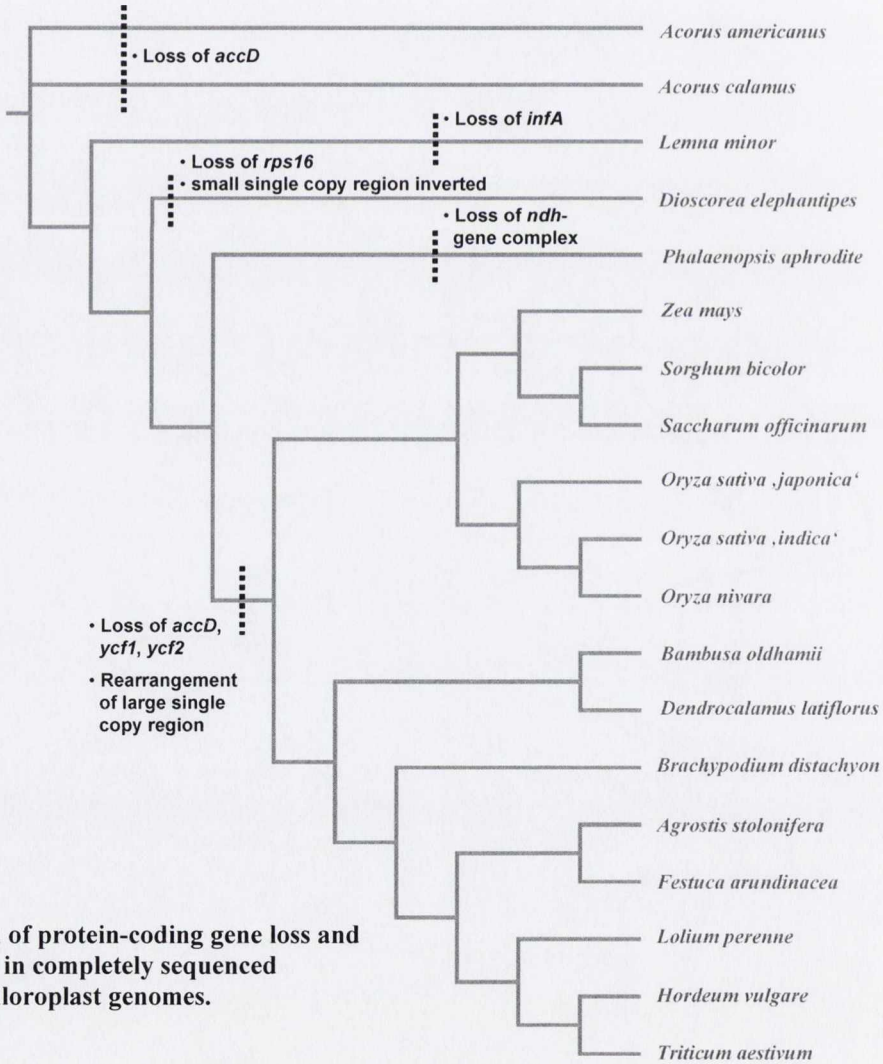

Figure 44 Illustration of protein-coding gene loss and gene rearrangements in completely sequenced monocotyledonous chloroplast genomes.

6.3.2 Phylogenetic analyses

6.3.2.1 Chloroplast single gene analysis

For the chloroplast single gene analysis the non-Poaceae species *Acorus americanus*, *A. calamus*, *Dioscorea elephantipes*, *Lemna minor* and *Phalaeonopsis aphrodite* were used as outgroup. The ingroup was based on four different subfamilies: Bambusoideae (two species), Ehrhartoideae (three species), Panicoideae (three species) and Pooideae (six species). In general all genes contained enough information to group the species into their respective taxonomic subfamily. The length of the gene alignments used for the chloroplast single gene analysis varied from 90 nucleotides with nine parsimony-informative sites (*petN*) to 4788 nucleotides with 1186 parsimony informative sites (*rpoC2*). In general it was found that the longer the gene sequences were (and hence alignments) the more informative sites they contained (Figure 45). But the number of genes that contained information about the relationship of the different subfamilies to each other was rather limited. Using PAUP* 4.0 (Swofford 2002) only 28 genes out of 76 were able to distinguish the subfamilies and for only eleven genes the bootstrap support was above 70%. Analysing these 28 genes in MrBayes (Huelsenbeck & Ronquist 2001, Ronquist & Huelsenbeck 2003) left only five genes (*matK*, *petD*, *psaB*, *psbA, psbH*) with posterior probability values higher than 0.90.

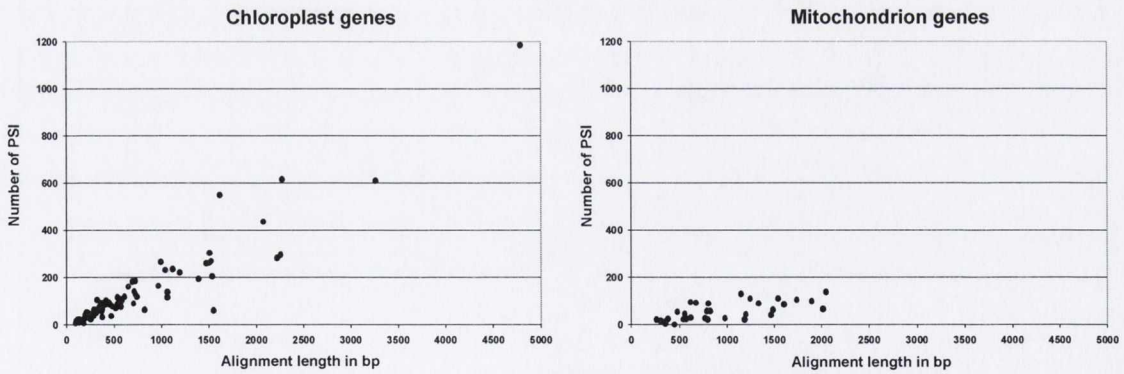

Figure 45 Number of parsimony informative sites (PIS) in relation to the alignment length for chloroplast and mitochondrial protein coding genes.

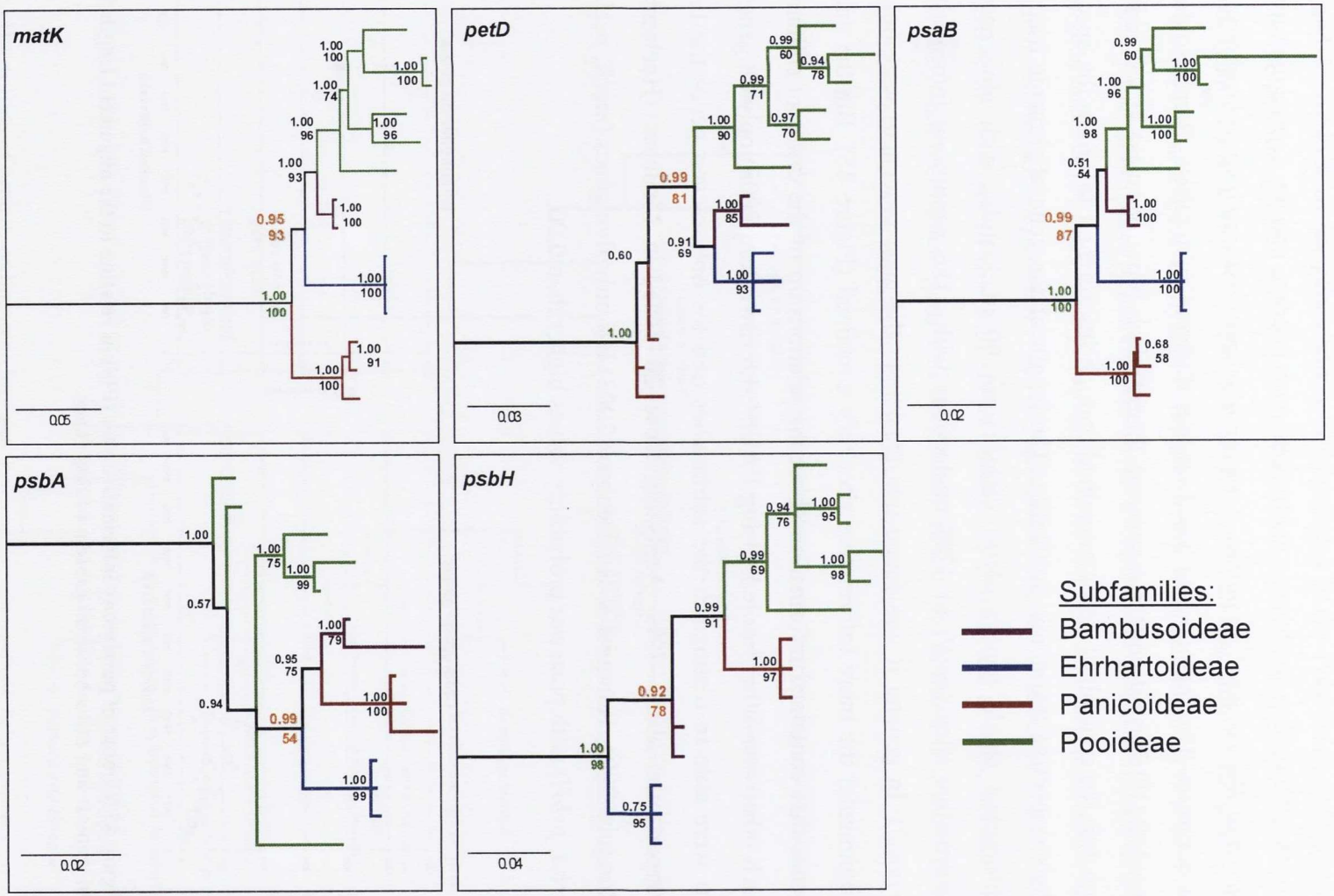

Figure 46 Results of chloroplast protein-coding single gene analysis for the five genes with the highest bootstrap (beneath lines) and posterior probability (above lines) values.

Three of these genes, *matK*, *petD* and *psaB*, supported the hypothesis that the subfamily Panicoideae are the most outlying family included in the analysis with Pooideae, Bambusoideae and Erhartoideae grouping together (Hypothesis **A**), *psbA* supported the hypothesis that Panicoideae, Bambusoideae and Ehrhartoideae can be grouped (Hypothesis **B**) and *psbH* resolved Ehrhartoideae as outlying to the rest (Hypothesis **C**) (Figure 46).

6.3.2.2 Chloroplast multigene analysis

The analysis of the combined matrix of 76 protein coding chloroplast genes of 19 different monocotyledonous species using Parsimony and Bayesian inference approaches resulted in highly congruent and resolved trees with very well supported branches (Figure 47). It clustered the subfamilies Panicoideae, Ehrhartoideae, Bambusoideae and Pooideae and also distinguished the relationships within the Pooideae of the tribes Triticeae, Aveneae and Poeae –with *Brachypodium* outlying all other Pooideae taxa. The Bayesian inference grouped Panicoideae and Ehrhartoideae together as sister to Bambusoideae and Pooideae while the analysis with parsimony in PAUP* 4.0 (Swofford 2002) revealed the Panicoideae as sister to the other Poaceae subfamilies (data not shown). Both results had very high bootstrap and posterior propability values.

6.3.2.3 Mitochondria multigene analysis

The mitochondria multigene analysis was based on 32 protein coding genes (Table 47). The alignment of the genes varied from 261 (*atp9*) to 2043 bp (*matR*) with two to 139 parsimony informative sites. Figure 45 illustrates the relationship between parsimony informative sides and the sequence length of chloroplast protein coding sides compared to the relation between parsimony informative sites and the sequence length of mitochondrial protein coding sites. It shows that the chloroplast genome has generally more genes, a higher percentage of long genes and that the chloroplast protein coding genes have in average more parsimony informative sites than the mitochondrion genes. This is also mirrored in a single gene analysis carried out on the mitochondrial protein-

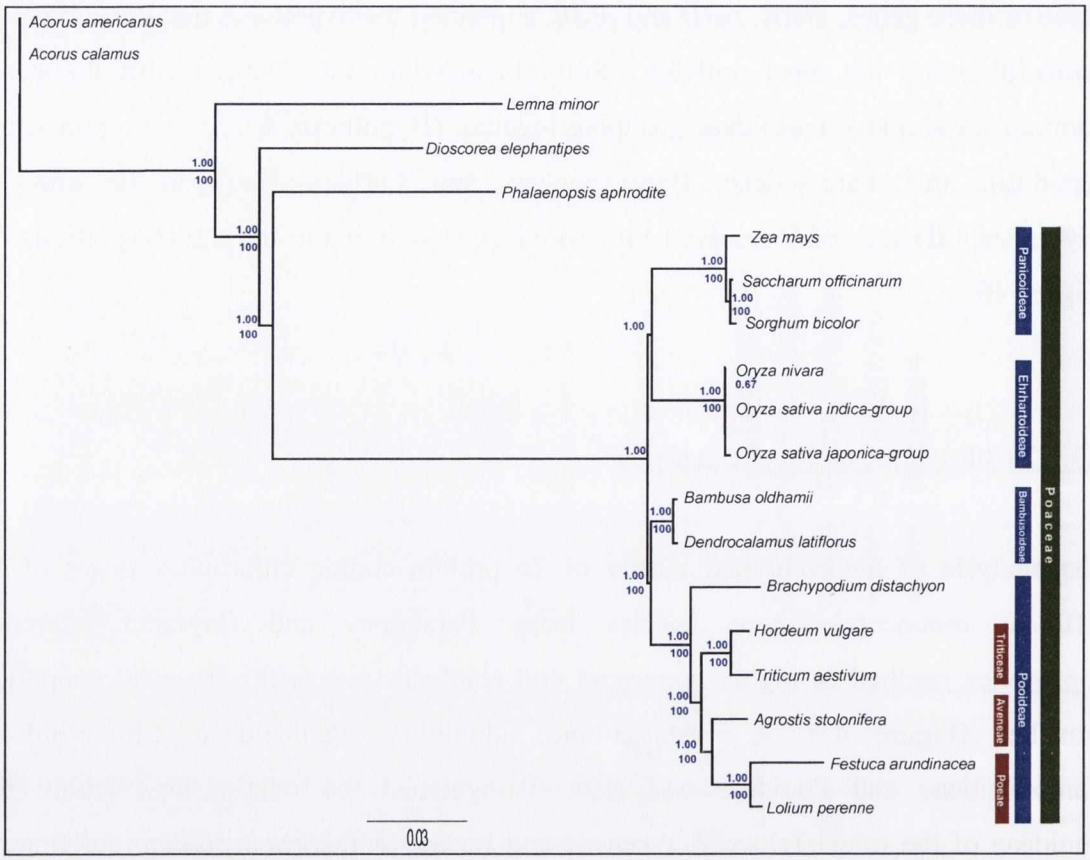

Figure 47 Bayesian inference tree of 76 protein coding chloroplast genes from 19 monocotyledonous species including 14 Poaceae species. Above line = posterior probability values, below line = bootstrap values

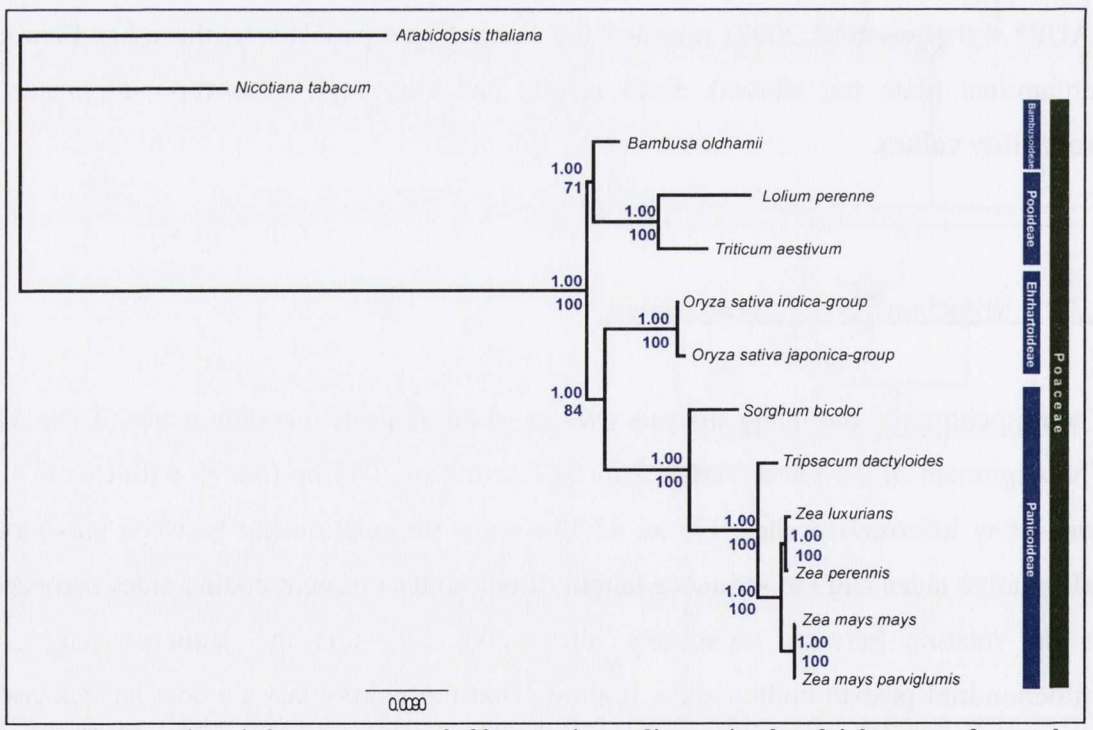

Figure 48 Bayesian inference tree of 32 protein coding mitochondrial genes from eleven monocotyledonous and two dicotyledonous species. Above line = posterior probability values, below line = bootstrap values

coding genes (data not shown). Although the dicotyledonous outgroup of *Nicotiana tabacum* and *Arabidopsis thaliana* in general was resolved and the Poaceae species were also grouped into the correct tribes and most of the time also into the correct subfamilies, so was the bootstrap and posterior propability support in general much weaker as observed for most chloroplast protein coding genes (data not shown). A conflict in the relationship of the different subfamilies was also observed in the mitochondrial gene data (data not shown), however for more thorough studies the amount of available sequences was not enough. Nevertheless, the multigene analysis on the mitochondrial genome resolved all branches completely and with very high bootstrap support and posterior propability values (Figure 48). Both approaches grouped Bambusoideae as sister to the Pooideae species and the Ehrhartoideae species as sister to the Panicoideae species. However, the bootstrap support for these groupings was much lower (71 and 84 %, respectively) as for the rest of the tree. The mitochondrial tree has resolved the same pattern of interrelationships between subfamilies as the combined chloroplast tree (with an Ehrhartoideae +Panicoideae group sister to a Pooideae + Bambusoideae group).

6.4 Discussion

6.4.1 Structural characterization of monocotyledonous chloroplast genomes

Chloroplast genomes are generally highly conserved regarding gene content and gene order (Palmer et al. 1988). However, few exceptions to this rule can be found and in the monocotyledonous family Poaceae the number of exceptions is extraordinarily high, compared to all other chloroplast genomes analysed so far. Comparative studies between the dicotyledonous species *Nicotiana tabacum* (Solanaceae) and the monocotyledonous species *Oryza sativa* (Poaceae) revealed six major rearrangements in *O. sativa* compared to *N. tabacum*: a) three inversions within the large single copy region, b) loss of the intron of *rpoC1*, c) an insertion in *rpoC2*, d) deletion of the two large open reading frames *ycf1* and *ycf2* in the inverted repeat, e) loss of the functionality of *accD* and f) translocation and loss of the *rpl23* gene (reviewed in Katayama & Ogihara 1996). Analysing 58 monocotyledonous species from twenty-two different families and eleven different orders, Katayama and Ogihara (1996) revealed

that the combination of these rearrangements is unique to Poaceae. The present study focussed only on the protein-coding gene content and gene order to highlight major differences between the outgroup (non-Poaceae) and ingroup (Poaceae) species used in this study. However, using this approach it was possible to detect some of the Poaceae specific rearrangements. The 28 kb inversion (inversion I, *psbM* - *atpA*) and the 6 kb inversion (inversion II, *psbC* – *psbZ*) (Howe et al. 1988) are absent from all outgroup species and thus confirm earlier results by Katayama and Ogihara (1996) which found these two inversions only in closely related subfamilies such as Joinvillaceae and Restionaceae. Interestingly, from the four analysed Restionaceae species only one showed inversion II (Katayama & Ogihara 1996). Other structural differences between the outgroup and ingroup species was the loss of *ycf1* and *ycf2* which are both present in the five outgroup species. However, genes were also lost in the outgroup species. The largest gene loss was observed in *Phalaenopsis aphrodite*. This species lost the complete *ndh*-gene complex which was presumably transferred to the nucleus (Chang et al. 2006). *Acorus calamus* and *A. americanus* lost the *accD* gene (Goremykin et al. 2005), *Dioscorea elephantipes* the *rps16* gene (Hansen et al. 2007) and *Lemna minor* the *infA* gene (Mardanov et al. 2008). These gene losses seem to be, in general, family specific because except from the *accD* gene all these genes are present in Poaceae, implying that they were lost after diverging from a common ancestor. The loss of these genes is not unique to monocotyledonous chloroplast genomes but was for the *ndh*-gene complex also observed in *Epifagus virginiana* (Wolfe et al. 1992) and *Pinus thunbergii* (Wakasugi et al. 1994), for *infA* in *Arabidopsis thaliana*, *Lotus japonicus* and *Medicago truncatula* (Millen et al. 2001), and for *rps16* in *Pinus thunbergii* (Wakasugi et al. 1994).

In recent years a rapid increase in published chloroplast genome sequences has occurred providing the opportunity for phylogenetic studies based on complete chloroplast genome sequences (Bortiri et al. 2008). However, the presence of rearrangments within the monocotyledonous clade rather limits the application of complete monocotyledonous chloroplast genome sequences for phylogenetic analyses. Alignments based on complete genome sequences from species belonging to different families will not be possible unless rearranged regions are excluded from the analyses or alignment programs available that rearrange the sequences based on a reference sequence. Such alignment programs exist and MultiPipMaker (Schwartz et al. 2000) is

probably the first which became available. Although other programs for example VISTA (Mayor et al. 2000) were designed which included even phylogenetic analyses based on the created alignments, MultiPipMaker seems to be the only tool to date that allows the extraction of the created alignments for further studies in various phylogenetic programs. This is especially important for phylogenetic analyses on rearranged genomes, because VISTA does not support these type of analyses. However, the quality and also accuracy of alignments from rearranged genomes created by these programs is not known and thus the use of extracted chloroplast protein-coding genes from different species for phylogenetic analyses might be more precise.

6.4.2 Single gene analysis

Single gene analyses were carried out based on all the 76 protein-coding genes that exist in Poaceae chloroplast genomes. The amount of phylogenetically informative characters varied widely across genes. As little as nine parsimony informative sites were found in the shortest gene alignment (*petN*) to as many as 1186 parsimony informative sites in the longest gene alignment (*rpoC2*). Although short genes like *petN* are too conserved to resolve the relationship between the Poaceae species they generally recognized differences relative to the outgroup species suggesting that they might be suitable for phylogenetic studies at higher taxonomic levels. However, while long genes were more likely to contain phylogenetic information this information was not always precise enough to group not only closely related species into their subfamilies but also to reveal the relationships of the different subfamilies.

The relationship of the different subfamilies was of specific interest due to a continous conflict based on the sequencing regions chosen for analyses (reviewed in GPWG 2001). Using sequences of the chloroplast gene *rbcL* or the nuclear ITS region, restriction site mapping or morphological characters supported a grouping of the PACCMAD clade with the Pooideae while sequences of the chloroplast genes *ndhF* and *rpoC2* as well as the nuclear gene *phyB* supported the existence of a PACCMAD and BEP clade (GPWG 2001, Sánchez-Ken et al. 2007, Bouchenak-Khelladi et al. 2008).

The single chloroplast protein-coding gene analyses revealed that the majority of chloroplast genes are able to distinguish the different subfamilies but are not able to reveal the relationship between them (Figure 46). Only 28 out of 76 genes contained enough phylogenetic information for robust tree reconstruction (highly resolved and well supported clades). Most of the single gene analyses (16 out of 28) were consistent with the PACCMAD and BEP clade hypothesis. However, bootstrap and posterior propability support were for most of the genes weak. Only five genes (*matK*, *petD*, *psaB*, *psbA*, *psbH*) showed very high posterior propability supports (>0.90). The majority of these genes (*matK*, *petD*, *psaB*) supported hypothesis **A**, that the PACCMAD clade, represented in this study by the subfamily Panicoideae, is sister to the BEP clade. This result was very well supported by bootstrap and posterior propability values above 80% and 0.90. *psbA* supported hypothesis **B** (Pooideae are sister to Panicoideae, Ehrhartoideae, Bambusoideae). However, the bootstrap value for this hypothesis was quite low despite the high posterior propability value. The *psbH* gene supported hypothesis **C** (Ehrhartoideae are sister to Pooideae, Panicoideae, Bambusoideae). No gene revealed a relationship that grouped Pooideae with Panicoideae. Few genes (*rpoC2*, *rps8*; data not shown) supported a grouping of Panicoideae with Ehrhartoideae either in the bootstrapping or MrBayes (Huelsenbeck & Ronquist 2001, Ronquist & Huelsenbeck 2003) analysis. However the support for this relationship was always weak. Thus the single chloroplast protein-coding gene analysis revealed that only five out of 76 genes are suitable in phylogenetic analyses to investigate the relationship of the Poaceae subfamilies but that even between these considerable incongruence existed.

However, the majority of the single gene analysis was consistent with the hypothesis that the PACCMAD clade is sister to the BEP clade confirming earlier studies (Clark et al. 1995; GPWG, 2001). Surprisingly none of the genes that were used in the earlier studies revealed this grouping in this study. The *ndhF* gene, a chloroplast gene that is frequently used in phylogenetic studies was able to distinguish between the four subfamilies although with only moderate support, however it did not reveal the relationships of the four subfamilies to each other. Furthermore none of the genes of the *ndh*-complex were able to reveal this relationship. This could be due to the small number (19) of taxa used in this study. Clark et al. (1995) used 45 species in their analysis which was based on the *ndhF* gene and revealed the PACC clade as sister to

the BEP clade (Micrairoideae, Aristidoideae and Danthoniodeae (MAD) were not included). However, their sampling was strongly biased toward species of the Bambusoideae, because of their aim to investigate them more thoroughly. This study includes only two Bambusoideae species belonging to different but closely related genera (Sungkaew et al. 2009). Also the Ehrhartoideae are rather underrepresented in this study, with two species. The *rpoC2* gene revealed different groupings in this study depending on the program used. While PAUP* 4.0 (Swofford 2002) supported hypothesis **C**, MrBayes (Huelsenbeck & Ronquist 2001, Ronquist & Huelsenbeck 2003) revealed Ehrhartoideae and Panicoideae as sister to Pooideae and Bambusoideae. But in both studies the support for either relationship was rather weak. Again more thorough sampling could improve the resolution. The *rbcL* gene, one of the most frequently used genes in plant phylogenetic studies (Chase et al. 1993), grouped Pooideae with the PACCMAD clade in studies of GPWG (2001). However, in an earlier study by Doebley et al. (1990) the Ehrhartoideae were grouped with the Panicoideae as found for the *rpoC2* gene in this study when using Bayesian inference. The *rps8* gene in the bootstrap analyses of this study revealed also a very similar topology. One possible explanation for these results could be again the small number of taxa used in this study and used in the study by Doebley et al. (1990) compared to the large number of taxa sampled by the GPWG (2001). Nevertheless Doebley et al. (1990) also mention that the phylogenetic information within *rbcL* was not enough in their study to statistically clearly resolve the branching order for the subfamilies Ehrhartoideae, Panicoideae and Pooideae, suggesting that they diverged from each other in a very short time frame. However, these results are still surprising and of relevance because *rbcL* is one of the two gene regions recommended for the proposed plant DNA barcode (CBOL Plant Working Group 2009) and a higher resolution would have been expected.

The GPWG (2001) found contradictory results when analysing two chloroplast genes which are supposed to be linked and thus inherited in the same manner (GPWG 2001). But their analyses showed that the *ndhF* gene supported a different branching order between the subfamilies than the *rbcL* gene. Taking into the account the results of this study it was clearly shown that some genes might have or infer a different evolutionary history than others. This study found evidence that the *psbA* gene and *psbH* gene evolved in a different manner. While the *psbA* result might be arguable because the Pooideae family was not at all or only with minor support grouped together, the *pbsH*

result received high bootstrap and posterior propability values for a relationship that sees Ehrhartoideae as sister to Bambusoideae, Panicoideae and Pooideae. A previous study conducted before the sequence of the two bamboo species became available, showed even higher support for *psbH* in regard to the branching order of the three subfamilies. That study had been conducted in the same way as the present one, however, with the exception of *psbH,* the highest bootstrap value for hypothesis **B** was obtained for *infA* and for hypothesis **A** by *ndhE*. Interestingly after including the Bambusoideae into the study neither *infA* nor *ndhE* contained enough information to group these subfamilies at all. The three genes which show in the present study the highest bootstrap and posterior propability values for hypothesis **A**, showed already in that earlier analyses high values (boostrap values without Bambusoideae for *matK*: 86, *petD*: 81, *psaB*:81) but increased them in the present one even. *psbA* is supporting hypothesis **B** in the present study. However, in the earlier study this gene was not even informative enough for detecing the branching order between these three species at all. Therefore the genes *matK*, *petD* and *psaB* seem to be very suitable for future phylogenetic studies including more than the present taxa. The gene *psbH* is an interesting candidate for further investigating the evolutionary background of chloroplast genes. It seems to be worthwhile to phylogenetically investigate this gene in more taxa, to see if the different branching order will remain. The gene *psbA* showed to be highly depending on the number of taxa included in a study and might not be very useful in future phylogenetic analyses. Three other genes *rps4*, *rps12* and *rps19*, were observed in both studies supporting hypothesis **B** although with very weak support. Nadot et al. (1994, 1995) conducted comprehensive phylogenetic studies on monocotyledonous species using *rps4* and obtained the same branching order as this study suggesting a different evolutionary background for this gene, too.

The gene with the highest resolution in this single gene analysis is *matK*. It supports the hypothesis **A** with the highest combination of bootstrap and posterior propability values. This result is not very surprising because *matK* is one of the two genes proposed as plant DNA barcode (CBOL Plant Working Group 2009). Therefore the results obtained in the present study are confirmed by the results of the CBOL Plant Working Group and show that *matK* is definitely a good choice as barcoding gene for monocotyledonous species.

The fact that different chloroplast genes might either have different evolutionary histories or vary in their ability to detect phylogenetic patterns can have several reasons. Most phylogeny programmes consider the possibility of a variation in the substitution rate across nucleotide or amino acid data; however, variation can also occur in the evolutionary rate of a single site through time which is rarely taken into account (Ané et al. 2005). One way to explain the different phylogenetic patterns observed in the single gene analyses of this study, might be with 'the covarion (for concomitantly variable codon) hypothesis of molecular evolution' that proposes a change of the selective pressure on nucleic and amino acid sites throughout time and hence, a change in the evolutionary rate of that specific site (reviewed Ané et al. 2005). Another possibility to explain the different results obtained could be the low taxon sampling in this study. Basing phylogenetic analyses on one gene and only 19 taxa is possibly not sufficient enough as discussed in many scientific articles (reviewed in Soltis et al. 2004). Poor taxon sampling, as for example taxa with large amounts and biased patterns of nucleotide change can lead to wrong conclusions. Grasses are known for their high chloroplast nucleotide substitution rate. A high nucleotide substitution rate can, due to homoplasy, result in incorrect trees because of long-branch-attraction (Sanderson et al. 2000, Soltis et al. 2004).

6.4.3 Chloroplast multigene analysis

Acorus americanus, *Acorus calamus*, *Lemna minor*, *Dioscorea elephantipes* and *Phalaenopsis aphrodite* were used as outgroup taxa and in both parsimony and Bayesian analyses the four Poaceae subfamilies Bambusoideae, Ehrhartoideae, Pooideae and Panicoideae were resolved. Within the Pooideae the three tribes Triticeae, Poeae and Aveneae were also consistently resolved. *Lolium perenne* was grouped with *Festuca arundinacea* as was to be expected based on earlier studies (Torrecilla & Catalán 2002). The *F. arundinacea* branch was rather long compared to the *L. perenne* branch, which is very likely due to sequencing mistakes that exist in the *F. arundinacea* sequence. During the alignment of the individual genes some polymorphism were observed that occurred only in *F. arundinacea* and resulted in frameshifts which would cause non-functionality of the respective genes.

The parsimony and Bayesian analyses revealed two different branching orders in regard to the interrelationships of the four subfamilies. While the bootstrap analyses supported Panicoideae as sister to the BEP clade (P, BEP), the Bayesian inference analysis grouped the Panicoideae + Ehrhartoideae as sister to a Bambusoideae +Pooideae group (PE, BP). Most surprising was that both topologies were very well supported. The bootstrapping topology supporting Panicoideae as sister to the BEP clade was in agreement with several earlier multigene analysis such as: GPWG (2001) using a set of chloroplast restriction sites, chloroplast and nuclear genes and morphology data; Bouchenak-Khelladi et al. (2008) based on the chloroplast genes *matK* and *rbcL* as well as on the chloroplast intergenic spacer *trnL-trnF*; Leebens-Mack et al. (2005) and Saski et al. (2007) based on 61 chloroplast genes; Bortiri et al. (2008) based on complete chloroplast genomes. The results obtained with the Bayesian inference approach are very similar to results previously obtained in the single gene analysis using *rpoC2* and *rps8* although the support for this grouping was weak. Also Hilu et al. (1999) obtained this relationship when analysing the *matK* gene of 62 Poaceae species. Reasons for the difference between the two topologies could be that the number of taxa used in this study was relatively small - some subfamilies were only represented by two species (Soltis et al. 2004); that missing data influenced the output of the different programs in different ways and that the four subfamilies evolved from each other in such a narrow time frame that chloroplast gene sequences might not contain enough information to resolve the relationship between them. Bouchenak-Khelladi et al. (2008) observed in their comprehensive analyses that an addition of taxa with missing data decreased the support of several branches although the topology in general stayed the same. Missing data in this study were only based on a lack of a gene in either of the species. This study was designed to evaluate the application of chloroplast multigene analyses in the Poaceae, therefore the only missing data included in this study were located in the outgroup species. In total thirteen genes were absent from one or the other outgroup species, but it is unlikely that their absence should be the cause of the different topologies within the grasses.

It is also noteworthy that a study which was carried out before the two Bambusoideae species became available gave for both approaches the same result - Panicoideae as sister to Pooideae and Ehrhartoideae - which was in complete agreement with the studies carried out by Leebens-Mack et al. (2005) and Saski et al. (2007). It is not

known how the addition of another subfamily to the 24 (Leebens-Mack et al. 2005) and 38 taxa (Saski et al. 2007), respectively, would change the inferred relationships of the Poaceae subfamilies.

The present study differs from previous chloroplast multigene studies such as the ones by Leebens-Mack et al. (2005) and Saski et al. (2007) in the amount of taxa and the amount of genes used. While Leebens-Mack et al. (2005) and Saski et al. (2007) based their studies on 24 and 38 taxa, respectively, including gymnosperm and angiosperm species and using 61 chloroplast genes, this study focussed only on monocotyledonous species, specifically Poaceae and thus was able to include 76 chloroplast genes. Among the fifteen extra chloroplast genes were all *ndh*-complex genes. Although the chloroplast single gene analysis showed that their individual use in phylogenetic studies might be questionable they still contain valuable information especially on high taxonomic levels. Also the *ndhF* gene was previously already used for phylogenetic analyses among monocotyledonous species and might be of specific value.

6.4.4 Mitochondrion multigene analysis

Mitochondria genes are much more conserved (Wolfe et al. 1987) than chloroplast genes. The amount of completed mitochondrial genome sequences from the Poaceae family is to date very limited. However, eleven Poaceae species are available and therefore it was worthwhile to evaluate their applicability to phylogenetic studies as well as to compare the results of the two multigene analyses (chloroplast and mitochondria). Because no mitochondrial genome of a non-Poaceae monocotyledonous species is available to date, *Arabidopsis thaliana* and *Nicotiana tabacum* (both eudicot species) were chosen as outgroup due to their status as model species. In preparation to the multigene analysis all mitochondrial gene alignments were analysed seperately using parsimony bootstrapping and although in general many genes were able to group the species according to their subfamily, the support for the groupings was generally much lower than results from analyses with single chloroplast genes. No single gene was able to distinguish the different subfamilies and their branching order with high supportive values. This result is consistent with the hypothesis that mitochondrial genes are highly conserved and that their substitution rate is much lower than that of

chloroplast protein coding genes. This difference is also illustrated in Figure 3 which compares the number of parsimony informative sites and gene lengths found in the two organellar genomes. While in average every seventh nucleotide of chloroplast genes is a parsimony informative sites, in the mitochondrial genes only every 28th nucleotide is informative. However, when analysing the mitochondrial genes combined both bootstrapping and Bayesian inference analysis revealed all subfamilies very well supported. Both methods grouped the Ehrhartoideae with the Panicoideae species, and Bambusoideae with Pooideae. The best sampled subfamily in this study is the Panicoideae. Here, *Sorghum bicolor* and *Tripsacum dactyloides* are grouped with the genus *Zea* which is represented with three species and two subspecies. *Tripsacum* is sister to *Zea* in the chloroplast analysis as expected from morphological taxonomic knowledge (Clayton and Renvoize, 1986). The results of the mitochondrial multigene analysis are therefore also generally in agreement with GPWG (2001).

6.5 Conclusion

This study did not reveal any new information for the phylogeny of the grasses, it rather confirmed results obtained from other research groups such as APGIII and GPWG. This study in fact showed that it is possible to use Poaceae chloroplast and mitochondrial sequence information for phylogenetic studies despite their rather limited amount of evolutionary information (in comparison to the nuclear genome). There are several ways to apply this organellar sequence information and one is to use single genes. However, the chloroplast single gene analyses carried out in this study showed an incongruity in the evolutionary information that the different genes contain which could be due the application of not optimized phylogenetic methods, erroneous sequences, long-branching attractions and biased data (Sanderson et al. 2000, Duvall et al. 2008). Therefore, it can not be recommended to base phylogenetic studies on only one or few genes. The best way to conduct phylogenetic analyses might be to use complete genome sequences. However, the amount of sequenced chloroplast and mitochondrial genomes is still limited and the computer capacity needed for this approach is huge. Our study revealed that analyses based on either all chloroplast or all mitochondrial protein-coding genes are able to obtain the same phylogenetic results as more comprehensive studies that are based on a combination of morphological, biochemical and sequence data. The

chloroplast multigene analyses in the Poaceae family using all 76 protein-coding genes was generally able to resolve the relationships of the 19 taxa used in this study with high support. Also, the mitochondrial multigene analyses based on 32 protein coding genes and 13 taxa proved suitable for distinguishing between the different subfamilies and even genera. Chloroplasts and mitochondria are generally both maternally inherited and it is thus expected that they show a similar evolutionary history. Thus future phylogenetic studies can possibly be based on a combined multigene approach using protein-coding sequences from both, chloroplasts and mitochondriaand might reveal an even better picture of the evolution of grasses and other families.

Chapter 7

General discussion on "Ryegrass organelle genomes: phylogenomics and sequence evaluation"

7.1 Introduction

The main objectives of this work were to sequence the entire chloroplast and mitochondrial genomes of the important forage grass species *Lolium perenne* and to extract information from both organelle genomes that is applicable in molecular evolution, phylogenetic and population genetic studies as well as in plant breeding programmes.

7.1.1 The chloroplast genome of *Lolium perenne*

A protocol was developed for the optimal extraction of chloroplast DNA from *Lolium* plants (Diekmann et al. 2008). It was found that the best yields were obtained when a combined approach of differential centrifugation and sucrose gradients/cushions for chloroplast purification was used. This yielded sufficient and pure DNA for chloroplast genome sequencing. The chloroplast genome of *L. perenne* cv. Cashel consists of 135,282 bp of DNA that are divided into a large (79,972 bp) and a small (12,428 bp) single copy region which are separated from each other by two copies of an inverted repeat (each 21,441 bp). The *L. perenne* chloroplast genome encodes 76 protein-coding, 30 transfer and four ribosomal RNA genes (Diekmann et al. 2009; Chapter 2). The genome was, regarding its size and gene content/order, in the range expected for Poaceae chloroplast genomes (e.g. Saski et al. 2007, Bortiri et al. 2009). Variation in the size of the different Poaceae chloroplast genomes were found to be mainly attributed to length variations of intron and intergenic spacer regions. An unexpected and considerably high amount of variation in the form of indels and SNPs was found throughout the genome of a single *L. perenne* cultivar (Diekmann et al. 2008, 2009). *Lolium perenne* is an outbreeding species and cultivars generally represent heterogeneous plant populations that are based on several maternal lines

(Wilkins 1991). The detection of highly variable regions within this cultivar enabled the targeted design of chloroplast DNA markers for phylogenetic and population/genetic resource studies (Chapter 4). The availability of the complete chloroplast DNA sequence of *L. perenne* enables the design of further chloroplast markers applicable in phylogenetic studies but also in plant breeding programmes for tracing and monitoring maternal inheritance. Furthermore does it provide the possibility to genetically engineer the chloroplast genome of *L. perenne* which would have a lower risk of spreading the transgene into wild populations than nuclear genetic engineering due to its maternal inheritance.

7.1.2 RNA editing sites in *Lolium perenne*

RNA editing is "the co- or post-transcriptional modification of RNA primary sequence from that encoded in the genome through nucleotide deletion, insertion, or base modification" (Smith et al. 1997). This process is very common in chloroplast genomes and past studies have shown that around 30 to 40 sites can be detected within a single species (Tillich et al. 2005). The complete *L. perenne* genome was not analysed for editing sites, instead this study focussed mainly on genes for which editing sites had been previously reported (Calsa Júnior et al. 2004; Corneille et al. 2000; Freyer et al. 1993; Maier et al. 1995; Zeltz et al. 1993). By taking this approach, 31 sites were found in 18 genes and it is believed that these are likely to include most of the existing editing sites in *L. perenne* (Diekmann et al. 2009). The examination of the *L. perenne* chloroplast genome for editing sites is essential because RNA editing can create translation initiation and termination codons. Therefore it is an important tool for predicting the functionality of genes and identifying potentially functional new open reading frames (Schuster et al. 1991). The knowledge of RNA editing sites within the chloroplast genome of *L. perenne* could be of specific interest to plant breeding because of the possibility that RNA editing is involved in CMS. Already now it is possible to induce and enhance CMS as well as to restore fertility in other species by introducing edited and non-edited gene constructs into the nucleus.

7.1.3 Plastid genome diversity within and among *Lolium perenne* cultivars and populations

Several chloroplast markers that amplify in *L. perenne* have been designed in recent years (Provan et al. 2001, McGrath et al. 2007). However, because these markers were optimised to amplify across a broad range of species, they may not amplify the most variable chloroplast regions in *L. perenne* and closely related species. Sequencing the *L. perenne* chloroplast genome enabled the design of nine new markers from highly variable regions of the *L. perenne* chloroplast genome, which additionally also amplify across a wide range of Poaceae species. Although all markers are suitable for population and phylogenetic analyses three markers performed particularly well on a small set of different *Festuca* and *Lolium* accessions from the Oak Park grass breeding collection. Marker TeaCpSSR31 was able to detect nearly all the variation within this set that was otherwise only found when combining the rest of the new markers. TeaCpSSR27 enabled the differentiation of haplotypes solely via amplicon size differentiation by agarose gel electrophoresis. Although this marker presents only a small part of the complete variation that could be found within the samples it provides the opportunity for a simple and inexpensive screening for different haplotypes within genetic resource collections of *Lolium* such as the Oak Park Grass collection. The new chloroplast markers were applied to a small sample set of cultivars and ecotypes. The study revealed a slightly reduced diversity for the cultivars in comparison to the Irish ecotypes. However, this could be due to the small sample set used. Nevertheless was the study able to show that these new markers are able to monitor genetic diversity and trace maternal inheritance. These markers can also define cytoplasmic breeding pools. This study showed that they are Poaceae universal primers thus amplify also in other agricultural important grasses such as bentgrass, wheat and barley. Of special interest for breeding companies and seed testing agencies could be TeaCpSSR28. TeaCpSSR28 is able to distinguish reliably between *L. multiflorum* and *L. perenne* which is especially important when analying seed lots of *L. perenne* for contamination with *L. multiflorum*. Although nuclear markers might show a greater diversity than chloroplast markers due to the conservation of the chloroplast genome, they need to be linked to specific agricultural important traits to use them in plant breeding programmes (Fjellheim & Rognli 2005). Some of them might be able to distinguish cultivars, however, their

application to an outbreeding synthetic population might be limited. Chloroplast microsatellite markers will in general not be able to distinguish cultivars, but they are able to detect species reliably and could be of great value in the evaluation of ecotype collections to assure the correct species identification.

7.1.4 First insights into the mitochondrial genome of *Lolium perenne*

One objective of the present study was to sequence, assemble and annotate the entire mitochondrial genome of *L. perenne*. Unfortunately, this could only be partially achieved due to difficulties encountered during the extraction of pure mitochondrial DNA and because of the complex nature of plant mitochondrial genomes in general (Clifton et al. 2004). However, despite all these issues a draft assembly based on 43 contigs was obtained. These 43 contigs very likely contain all the protein-coding and ribosomal RNA genes that can be found within the *L. perenne* mitochondrial genome (Ogihara et al. 2005). Mitochondrial genomes are generally not used in phylo- or population genetic studies because they are on the one side too conserved (the coding regions) and on the other side too diverse (intergenic spacer regions). Comparing the gene order of Poaceae mitochondrial genomes revealed one conserved region that is present in most species and comprises 5 genes and exons (*ccmFN*. *rps1*, *matK*, *nad1* exon 5, *nad5* exon3), respectively. The intergenic spacer regions of that gene cluster are conserved enough, among Poaceae species, to design primers within them, but also might exhibit enough variation to enable phylogenetic studies in the future. Furthermore it could even enable the design of mitochondrial markers that can be easily applied and are able to detect cytoplasmic variability among *L. perenne* cultivars as previously reported by Sato et al. (1995). The *L. perenne* mitochondrial genome was too rearranged to be assembled by aligning the contigs with the most closely related available genome sequence, that from *Triticum*. *Triticum* is in the same subfamily (Pooideae) but belongs to a different tribe to *Lolium* (Triticeae). The comparative studies in this thesis (e. g. between and within *Oryza* species and *Zea* species) indicate that this sort of approach to assemble the genome may only be fully successful by aligning *Lolium* with another *Lolium* species or perhaps a sister species but certainly not with species more divergent than that. The significance of this part

of the project for plant breeding lies in the involvement of the mitochondrial genome in CMS. *Lolium perenne* is self-incompatible and hence breeding of *L. perenne* does not focus on the generation of inbred lines and CMS. Nevertheless do most breeding companies have a few *L. perenne* CMS lines. Analysing the complete mitochondrial genome of *L. perenne* could lead to a better understanding of CMS in *L. perenne* and thus lead to a more efficient use of it in future breeding programmes.

7.1.5 Applicability of organelle sequence information to phylogenetic studies

Chloroplast genome sequences are frequently used in phylogenetic studies, solely or in combination with nuclear DNA while mitochondrial genome sequences are far less used and often only in combination with either chloroplast or nuclear DNA (Qiu et al. 1999; Savolainen and Chase, 2003). The complete set of chloroplast and mitochondrial protein coding sequences that exist in *L. perenne* were used in two separate phylogenetic studies using all available complete Poaceae organelle genomes. The results were in total agreement with the phylogenetic patterns found by the GPWG (2001) and APG III (2009). The degree of variation found within the mitochondrial protein-coding sequences was great enough to distinguish all species from each other. Therefore it was shown that mitochondrial DNA sequences should possibly be considered more often in future phylogenetic studies, because they contain valuable information not only on higher taxonomic levels. The sequences produced in this thesis will be highly useful for the design of primers to amplify suitable regions for future studies. Furthermore, from a plant breeding perspective, could this study enable the detection of horizontal gene transfer from, for example, the mitochondrial genome of fungi to the mitochondrial genome of *Lolium* once the genome assembly is completed. *Neotyphodium lolii* is an endophytic fungus which lives often in symbiosis with *L. perenne*. Infected *L. perenne* plants have a greater vigour and greater tolerance to biotic and abiotic stress. Horizontally genes transferred from this or other fungi group differently in phylogenetic studies and their detection might enable the design of new markers that could enable a better understanding of the interaction between these two different organisms.

7.2 Future prospects

The availability of the complete *L. perenne* chloroplast genome enables the genetic modification of *L. perenne* via chloroplast genetic engineering. The public acceptance of genetically modified plants is low due to a number of perceived risks including health risks and the risk of transgene escape via pollen flow (Daniell et al. 2005). Escaped transgenes can propagate independently in wild populations of the same or closely related species. For example, an escape of herbicide resistance inducing genes would make weed control via simple and inexpensive herbicide application difficult. The first observation of a transgene escape via pollen and seed flow into natural populations was reported by Reichman et al. (2006) for *Agrostis stolonifera* (creeping bentgrass) and highlights the risks of genetically engineering plant genomes. *Agrostis stolonifera* is a turf grass used widely for example on golf courses. It is also one of the very few turf grass species that were genetically modified making it resistant to glyphosate the major compound of the all-round herbicide RoundUp. Herbicide resistant *A. stolonifera* was thought to be desirable to facilitate the maintenance of golf courses more cost efficiently. In 2003, the United States Department of Agriculture firstly approved a 162 ha field trial with genetically modified *A. stolonifera* plants permitted to flower. Genetically modified *A. stolonifera* plants were set free in nine fields distributed rather randomly within an 4,453 ha bentgrass control area (Reichman et al. 2006). During 2004-2005 thorough field sampling took place outside the control area collecting 20,400 individual *A. stolonifera* plants. Analyses showed that nine out of the 20,400 plants had the transgene. They were found at a distance to the genetically modified *A. stolonifera* fields of up to 3.8 km. This was the first description of transgene escape via pollen and seed flow into natural populations (Reichman et al. 2006) highlighting the risks of genetically engineering plant genomes. Escaped transgenes can lead to a reduction of the natural genetic diversity among plant populations (Wolfenbarger et al. 2000).

Chloroplast genomes are maternally inherited in most angiosperm species (Corriveau & Coleman 1988). Thus engineering the chloroplast genome does not normally result in the risk of the escape of transgenes via pollen flow. This approach has become more important in recent years. Chloroplast genomes are generally highly conserved and it is theoretically possible to genetically modify a chloroplast genome without

knowing the exact sequence of the insert site. However, studies showed that the design of species specific vectors is of huge advantage for yielding successfully modified plants (Daniell et al. 2005). Modifying the *L. perenne* chloroplast genome genetically was not an objective of the present study, but was an objective of the complete project in which this study was included (and was the focus of the PhD project of Rob van den Bekerom). Therefore attempts on modifying the *L. perenne* chloroplast genome are ongoing and species specific vectors have been designed based on the chloroplast genome sequence established in this study. The extent to which genetically modified *L. perenne* plants will feature in agricultural systems can not currently be accurately predicted. As with chloroplast genomes, mitochondrial genomes are also maternally inherited in *L. perenne*. However, crossing *Festuca pratensis* with *L. perenne*, a commonly used approach in grass breeding programs, resulted in paternal inheritance of the mitochondrial genome (Kiang et al. 1994). *Lolium perenne* is able to form intergeneric/interspecific hybrids in nature (Hubbard 1984). If the chloroplast genome would also be inherited paternally via those hybridizations, the risk of spreading transgenes would increase again despite the fact that the genes are located within the chloroplast genome. *Lolium perenne* pollen flow studies carried out in Teagasc Oak Park showed that pollen flow can occur at least up to a distance of ca. 200 m (Mullins et al. 2009). Should genetically modified *L. perenne* plants ever be tested in field trials it will be nearly impossible to keep areas of more than 200 m surrounding the field trials completely free from *L. perenne* and other closely related species to avoid the escape of transgenes. Furthermore, the escape of transgenes via seed flow has to be considered which makes it even more unlikely that genetically modified plants would be used in agricultural systems in the near future. In *Lolium rigidum*, naturally occurring glyphosate resistance was detected (reviewed in Wolfenbarger et al. 2000). *Lolium rigidum* can be crossed with *L. perenne* and thus the resistance gene can bred naturally into *L. perenne* plants if the interest in glyphosate resistant *L. perenne* plants arises. Nevertheless, with the progressing climate change and the need to breed new cultivars adapted to their new environments, a system that would enable the safe genetic modification of *L. perenne* could still be advantageous.

Plant mitochondrial genomes are very challenging targets for genetic engineering. Mitochondria are in general much smaller than chloroplasts and modification of the

genome using a gene gun is not very promising. Furthermore, mitochondrial genomes are not conserved (except from the coding regions) and therefore sequences for the insert sites would have to be determined for every species nearly independent on how closely related the species are. The sequencing of the *L. perenne* mitochondrial genome in this study was done partly because of the involvement of mitochondrial genomes in CMS. CMS is a mechanism used by plant breeders for the rather inexpensive and controlled production of hybrid seed which shows the positive effects of heterosis (e.g. increased seed and biomass production). McDermott et al. (2008) showed that CMS in *L. perenne* lines derived from intergeneric hybrids is based on the insertion of a linear mitochondrial plasmid into the mitochondrial genome. The same sequence to the plasmid was located within the draft assembly of the *L. perenne* mitochondrial genome of this study, which was rather surprising because a male fertile cultivar was used. However, it is not yet possible to say if this plasmid is *in vivo* inserted into the *L. perenne* mitochondrial genome or if it was only accidentally assembled into the mitochondrial genome. Only the complete assembly of the *L. perenne* mitochondrial genome from the fertile *L. perenne* cv. Shandon can shed light into this question.

Besides the application of chloroplast genome sequences to genetic engineering approaches and the use of the mitochondrial genome for investigative studies of CMS in *L. perenne*, both genomes proved to contain valuable information for comparative and phylogenetic studies. *Lolium perenne* chloroplast genome sequences are already in use for phylogenetic and population genetic studies (e.g. McGrath et al. 2007, McGrath 2008) and additional primers amplifying highly variable regions of Poaceae chloroplast genomes applicable in both approaches were designed in this study. However, the availability of the complete *L. perenne* chloroplast genome sequence combined with the knowledge of highly variable regions within this sequence enables the further search for additional valuable regions that can be used for studying Poaceae phylogeny as well as the diversity within grass collections. Mitochondrial genomes are rarely used for those approaches due to difficulties in finding conserved regions for primer design that surround variable regions suitable for phylogenetic and population genetic analyses. However, restriction fragment analyses on the mitochondrial genome of different *L. perenne* cultivars revealed a considerable high amount of cytoplasmic variation (Sato et al. 1995, Shimamoto et al.1997). During the

L. perenne mitochondrial genome assembly attention was drawn to several regions within the genome that comprised long microsatellites or exhibited single nucleotide polymorphisms within *L. perenne* cv. Shandon itself. These observations combined with the results from earlier analyses by Sato et al. (1995) and Shimamoto et al. (1997) indicate the possible potential of mitochondrial genome sequences for population genetic analyses or even for DNA barcoding applications. Furthermore, this study showed also that protein-coding sequences are variable enough to be used for phylogenetic analyses at high taxonomic levels (Chapter 6).

Therefore future work should definitely aim to finish the mitochondrial genome assembly of this *L. perenne* cultivar. One of the biggest issues which prevented the completion of the assembly within the present study was a lack of knowledge regarding the orientation and order of the 43 mitochondrial DNA contigs obtained. For sequencing the *L. perenne* mitochondrial genome, a hybrid approach was used which included the creation of a shotgun library with insert sizes of 2.5 kb. The clones of that library were sequenced from both ends. Therefore in a first attempt for finishing the assembly the draft assembly should be searched for read files of one clone that are located on the ends of two different contigs. Finding these files and contigs would enable the orientation and combination of the contigs. The gaps between the contigs can then be sequenced based on primers designed on both contigs. However, this approach is very much limited to gaps between contigs of not more than 900 bp due to the choice of the insert size. Larger gaps might need to be closed using a primer walking approach, in which primers are developed for the sequence of one contig, which will be extended via sequencing until another contig is met. The quick success of this method, however, is very much dependent on the size of the gaps between the different contigs. If the gaps are too long, the amount of necessary PCR and sequencing reactions increases the costs immensly. Furthermore this approach would require a huge amount of template DNA to guarantee its success. Using total DNA is not an option because it might be that the primer walking needs to start in an intracellular transferred chloroplast region. Hence primer design from such a region would very likely automatically result in the amplification of chloroplast rather than mitochondrial DNA. If a primer walking approach is unable to finish the assembly, the design of a fosmid library from the mitochondrial DNA sample could be another option. Fosmids can contain inserts of up to 40 kb although

the insert size for the purpose of finishing the mitochondrial genome assembly should be more around 10 kb. The ends of all fosmids could then be sequenced and the sequences compared to the contigs in a similar way as done for the Sanger read files in the first approach. That way it should be also possible to orientate the contigs to each other even if the gaps between them are bigger. In case that a fosmid clone contains several smaller contigs it might also be an option to design probes from the different contigs that can be hybridized to the different clones. That way it will be possible to firstly find out which contigs are located on which fosmid and then to sequence specifically from one contig to another on the fosmid clone.

Once the *L. perenne* mitochondrial genome is completely assembled it can be searched for editing sites. Plant mitochondrial genomes generally contain around 400 editing sites. For screening the *L. perenne* mitochondrial genome, a fast, efficient and inexpensive is needed such as the high resolution melting analysis as described by Chateigner-Boutin and Small (2007). Determining the mitochondrial editing sites within *L. perenne* might not only solve some of the observed differences in coding regions between the different species, it is also a very valuable approach for detecting functional open reading frames to discover new genes that are possibly involved in CMS or have other metabolic functions. Once completed, the mitochondrial genome should also be search for regions that can be used in population genetic studies and might provide information for phylogenetic studies.

To date, the amount of published chloroplast genome sequences is considerably high, while the amount of published mitochondrial genome sequences is rather low. With the addition of the *L. perenne* chloroplast genome sequence, and in the near future the complete *L. perenne* mitochondrial genome sequence, valuable information becomes available for investigating molecular evolution of organelle genomes. The presented organelle genomes and their features are already of enormous value despite the fact that the number of published Poaceae organelle genomes is still limited. Therefore, once more genomes (especially those of Poaceae species) become available; the genome sequences produced in this thesis will become even more valuable for future comparative studies.

8. References

Adams KL, Palmer JD. (2003) Evolution of mitochondrial gene content: Gene loss and transfer to the nucleus. *Molecular Phylogenetics and Evolution*. 29:380-395.

Adams MW, Ellingboe AH, Rossman EC. (1971) Biological uniformity and disease epidemics. *BioScience*. 21(21):1067-1070.

Akkak A, Boccacci P, Botta R. (2007) 'Cardinal' grape parentage: a case of a breeding mistake. *Genome*. 50(3):325-8.

Allen MF. (1991) *The ecology of mycorrhizae*. Cambridge University Press. Cambridge.

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG. (1981) Sequence and organization of the human mitochondrial genome. *Nature*. 290(5806):457-65.

Ané C, Burleigh JG, McMahon MM, Sanderson MJ. (2005) Covarion structure in plastid genome evolution: a new statistical test. *Molecular Biology and Evolution*. 22(4):914-24.

Anhalt UCM, Heslop-Harrison (Pat) JS, Piepho HP, Byrne S, Barth S. (2009) Quantitative trait loci mapping for biomass yield traits in a *Lolium* inbred line derived F2 population. *Euphytica*. 170(1-2):99-107.

Ankele E, Heberle-Bors E, Pfosser MF, Hofinger BJ. (2005) Searching for mechanisms leading to albino plant formation in cereals. *Acta Physiologiae Plantarum*. 27:651-664.

Atienza SG, Martín AC, Ramírez MC, Martín A, Ballesteros J. (2007) Effects of *Hordeum chilense* cytoplasm on agronomic traits in common wheat. *Plant Breeding*. 126(1):5-8.

Backert S, Nielsen BL, Börner T. (1997) The mystery of the rings: structure and replication of mitochondrial genomes from higher plants. *Trends in Plant Science. Reviews*. 2(12):477-483.

Bakker FT, Culham A, Pankhurst CE, Gibby M. (2000) Mitochondrial and chloroplast DNA-based phylogeny of *Pelargonium* (Geraniaceae). *American Journal of Botany*. 87(4):727-734.

Baldwin BG, Sanderson MJ, Porter JM, Wojciechowski MF, Campbell CS, Donoghue MJ. (1995) The ITS region of nuclear ribosomal DNA: A valuable source of

evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden*. 82(2):247-277.

Balfourier F, Imbert C, Charmet G. (2000) Evidence for phylogeographic structure in *Lolium* species related to the spread of agriculture in Europe. A cpDNA study. *Theoretical and Applied Genetics*. 101(1-2):131-138.

Barker NP. (1997) The relationship of *Amphipogon*, *Elytrophorus* and *Cyperochloa* (Poaceae) as suggested by *rbcL* sequence data. *Telopea*. 7(3):205-213.

Barkman TJ, McNeal JR, Lim S-H, Coat G, Croom HB, Young ND, dePamphilis CW. (2007) Mitochondrial DNA suggests at least 11 origins of parasitism in angiosperms and reveals genomic chimerism in parasitic plants. *BMC Evolutionary Biology*. 7:248.

Bastia T, Scotti N, Cardi T. (2001) Organelle DNA analysis of *Solanum* and *Brassica* somatic hybrids by PCR with 'universal primers'. *Theoretical and Applied Genetics*. 102(8):1265-1272.

Bateson W, Gairdner AE. (1921) Male sterility in flax, subject to two types of segregation. *Journal of Genetics*. 11(3):269-275.

Bedbrook JR, Bogorad L. (1976) Endonuclease recognition sites mapped on *Zea mays* chloroplast DNA. *Proceedings of the National Academy of Science USA*. 73(12):4309-4313.

Bellaoui M, Pelletier G, Budar F. (1997) The steady-state level of mRNA from the Ogura cytoplasmic male sterility locus in *Brassica* cybrids is determined post-transcriptionally by its 3' region. *The EMBO Journal*. 16(16):5057-68.

Benchimol M. (2009) Hydrogenosomes under microscopy. *Tissue and Cell*. 41:151-168.

Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, Tromp MC. (1986) Major transcript of the frameshifted *coxII* gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell*. 46(6):819-26.

Bernacchi D, Beck-Bunn T, Eshed Y, Lopez J, Petiard V, Uhlig J, Zamir D, Tanksley S. (1998) Advanced backcross QTL analysis in tomato. I. Identification of QTLs for traits of agronomic importance from *Lycopersicon hirsutum*. *Theoretical and Applied Genetics*. 97:381-397.

Bock R. (2000) Sense from nonsense: how the genetic information of chloroplasts is altered by RNA editing. *Biochimie*. 82(6-7):549-57.

Bock R, Hagemann R, Kössel H, Kudla J. (1993) Tissue- and stage-specific modulation of RNA editing of the *psbF* and *psbL* transcript from spinach plastids - a new regulatory mechanism? *Molecular and General Genetics*. 240(2):238-44.

Bock R, Hermann M, Kössel H. (1996) In vivo dissection of *cis*-acting determinants for plastid RNA editing. *The EMBO Journal*. 15(18):5052-9.

Bonen L, Bird S. (1988) Sequence analysis of the wheat mitochondrial *atp6* gene reveals a fused upstream reading frame and markedly divergent N termini among plant ATP6 proteins. *Gene*. 73(1):47-56.

Bookjans G, Stummann BM, Henningsen KW. (1984) Preparation of chloroplast DNA from pea plastids isolated in a medium of high ionic-strength. *Analytical Biochemistry*. 141:244–247.

Boore JL. (1999) Animal mitochondrial genomes. *Nucleic Acids Research*. 27(8):1767-1780.

Bortiri E, Coleman-Derr D, Lazo GR, Anderson OD, Gu YQ. (2008) The complete chloroplast genome sequence of *Brachypodium distachyon*: sequence comparison and phylogenetic analysis of eight grass plastomes. *BMC Research Notes*. 1:61.

Bouchenak-Khelladi Y, Salamin N, Savolainen V, Forest F, Bank M, Chase MW, Hodkinson TR. (2008) Large multi-gene phylogenetic trees of the grasses (Poaceae): progress towards complete tribal and generic level sampling. *Molecular Phylogenetics and Evolution*. 47(2):488-505.

Bremer K. (2002) Gondwanan evolution of the grass alliance of families (Poales). *Evolution*. 56(7):1374-87.

Brennicke A, Marchfelder A, Binder S. (1999) RNA editing. *FEMS Microbiology Reviews*. 23(3):297-316.

Bryan GJ, De Jong W, Provan J, McNicoll J, Davidson J, Ramsay G, Waugh R. (1999a) *Potato genomics: a general strategy for the molecular genetic characterisation of* Solanum *germplasm*. http://www.scri.ac.uk/publications/1998/99annual report. Accessed: 18.12.2009.

Bryan GJ, McNicoll J, Ramsay G, Meyer RC, De Jong WS. (1999b) Polymorphic simple sequence repeat markers in chloroplast genomes of Solanaceous plants. *Theoretical and Applied Genetics*. 99(5):859-867.

Bult CJ, Sweere JA, Zimmer EA. (1995) Cryptic sequence simplicity, nucleotide composition bias, and molecular coevolution in the large subunit of ribosomal

DNA in plants: Implications for phylogenetic analyses. *Annals of the Missouri Botanical Garden.* 82(2):235-246.

Calsa Júnior T, Carraro DM, Benatti MR, Barbosa AC, Kitajima JP, Carrer H. (2004) Structural features and transcript-editing analysis of sugarcane (*Saccharum officinarum* L.) chloroplast genome. *Current Genetics*. 46(6):366-73.

Campbell NA, Reece JB. (2005) *Biology*. 8th edn. Addison Wesley. Munich.

Catalán P, Torrecilla P, López Rodríguez JA, Olmstead RG.(2004) Phylogeny of the festucoid grasses of subtribe Loliinae and allies (Poeae, Pooideae) inferred from ITS and *trnL-F* sequences. *Molecular Phylogenetics and Evolution*. 31(2):517-41.

Cavalier-Smith T. (1975) The origin of nuclei and of eukaryotic cells. *Nature*. 256:463-468.

CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Science USA*. 106(31):12794-12797.

Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, Cheng CH, Lin CY, Liu SM, Chang CC, Chaw SM. (2006) The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. *Molecular Biology and Evolution*. 23(2):279-91.

Chao S, Sederoff R, Levings III CS. (1984) Nucleotide sequence and evolution of the 18S ribosomal RNA gene in maize mitochondria. *Nucleic Acids Research*. 12(16):6629-6644.

Chase MW (2004) Monocot relationships: an overview. *American Journal of Botany*. 91:1645-1655.

Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, Madriñán S, Petersen G, Seberg O, Jørgsensen T, Cameron KM, Carine M, Pedersen N, Hedderson TAJ, Conrad F, Salazar GA, Richardson JE, Hollingsworth ML, Barraclough TG, Kelly L, Wilkinson M. (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon*. 56(2):295-299.

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu Y-L, Kron KA, Rettig JH, Conti E, Palmer JD, Manhart JR, Sytsma KJ, Michaels HJ, Kress WJ, Karol KG, Clark WD, Hedren M, Gaut BS, Jansen RK, Kim K-J, Wimpee CF, Smith JF, Furnier GR, Strauss SH, Xiang Q-Y, Plunkett GM, Soltis PS, Swensen SM, Williams SE, Gadek PA,

Quinn CJ, Eguiarte LE, Golenberg E, Learn Jr. GH, Graham SW, Barrett SCH, Dayanandan S, Albert VA. (1993) Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden*. 80(3):528-548+550-580.

Chateigner-Boutin AL, Hanson MR. (2003) Developmental co-variation of RNA editing extent of plastid editing sites exhibiting similar *cis*-elements. *Nucleic Acids Research*. 31(10):2586-94.

Chateigner-Boutin AL, Ramos-Vega M, Guevara-García A, Andrés C, de la Luz Gutiérrez-Nava M, Cantero A, Delannoy E, Jiménez LF, Lurin C, Small I, León P. (2008) CLB19, a pentatricopeptide repeat protein required for editing of *rpoA* and *clpP* chloroplast transcripts. *The Plant Journal*. 56(4):590-602.

Chateigner-Boutin AL, Small I. (2007) A rapid high-throughput method for the detection and quantification of RNA editing based on high-resolution melting of amplicons. *Nucleic Acids Research*. 35(17):e114.

Chaudhuri S, Carrer H, Maliga P. (1995) Site-specific factor involved in the editing of the *psbL* mRNA in tobacco plastids. *The EMBO Journal*. 14(12):2951-7.

Chaudhuri S, Maliga P. (1996) Sequences directing C to U editing of the plastid *psbL* mRNA are located within a 22 nucleotide segment spanning the editing site. *The EMBO Journal*. 15(21):5958-64.

Chen Z, Muthukrishnan S, Liang GH, Schertz KF, Hart GE. (1993) A chloroplast DNA deletion located in RNA polymerase gene *rpoC2* in CMS lines of S*orghum*. *Molecular and General Genetics*. 236:251–259.

Chen Z, Schertz KF, Mullet JE, DuBell A, Hart GE. (1995) Characterization and expression of *rpoC2* in CMS and fertile lines of sorghum. *Plant Molecular Biology*. 28:799–809.

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK. (2006) The complete chloroplast genome sequence of *Pelargonium* x *hortorum*: Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Molecular Biology and Evolution*. 23(11):2175-2190.

Chun EHL, Vaughan jr. MH, Rich A. (1963) The isolation and characterization of DNA associated with chloroplast preparations. *Journal of Molecular Biology*. 7:130-141.

Chung SM, Gordon VS, Staub JE. (2007) Sequencing cucumber (*Cucumis sativus* L.) chloroplast genomes identifies differences between chilling—tolerant and—susceptible cucumber lines. *Genome*. 50:215–225.

Clark LG, Zhang W, Wendel JF. (1995) A phylogeny of the grass family (Poaceae) based on *ndhF* sequence data. *Systematic Botany*. 20(4):436-460.

Clayton WD, Renvoize SA. (1986) *Genera Graminum: Grasses of the World*. Kew Publishing. London.

Clifton SW, Minx P, Fauron CM-R, Gibson M, Allen JO, Sun H, Thompson M, Barbazuk WB, Kanuganti S, Tayloe C, Meyer L, Wilson RK, Neton KJ. (2004) Sequence and comparative analysis of the maize NB mitochondrial genomes. *Plant Physiology*. 136:3486-3508.

Cogan NO, Smith KF, Yamada T, Francki MG, Vecchies AC, Jones ES, Spangenberg GC, Forster JW. (2005) QTL analysis and comparative genomics of herbage quality traits in perennial ryegrass (*Lolium perenne* L.). *Theoretical and Applied Genetics*. 110(2):364-80.

Conant GC, Wolfe KH. (2008) GenomeVx: Simple web-based creation of editable circular chromosome maps. *Bioinformatics Applications Note*. 24(6):861-862.

Connolly V. (2000). Collection and conservation of forage species from old pasture ecosystems. *Irish Journal of Agricultural and Food Research*. 39: 476.

Connolly V, Wright-Turner R. (1984) Induction of cytoplasmic male-sterility into ryegrass (*Lolium perenne*). *Theoretical and Applied Genetics*. 68(5):449-453.

Corneille S, Lutz K, Maliga P. (2000) Conservation of RNA editing between rice and maize plastids: are most editing events dispensable? *Molecular and General Genetics*. 264(4):419-24.

Correns C. (1937) *Nicht mendelnde Vererbung*. Handbuch der Vererbungswissenschaft Band II H (Hrsg. Baur E, Hartmann M). Gebrüder Borntraeger Verlag. Berlin.

Corriveau JL, Coleman AW. (1988) Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species. *American Journal of Botany*. 75(10):1443-1458.

Covello PS, Gray MW. (1989) RNA editing in plant mitochondria. *Nature*. 341(6243):662-6.

Crews S, Ojala D, Posakony J, Nishiguchi J, Attardi G. (1979) Nucleotide sequence of a region of human mitochondrial DNA containing the precisely identified origin of replication. *Nature*. 277:192-198.

Cuadrado A, Cardoso M, Jouve N. (2008) Physical organisation of simple sequence repeats (SSRs) in Triticeae: structural, functional and evolutionary implications. *Cytogenetic and Genome Research*. 120(3-4):210-9.

Cummings MP, Mertens King L, Kellogg EA. (1994) Slipped-strand mispairing in a plastid gene: *rpoC2* in grasses (Poaceae). *Molecular Biology and Evolution*. 11(1):1-8.

Daniell H, Datta R, Varma S, Gray S, Lee S-B. (1998) Containment of herbicide resistance through genetic engineering of the chloroplast genome. *Nature Biotechnology*. 16:345-348.

Daniell H, Lee SB, Grevich J, Saski C, Quesada-Vargas T, Guda C, Tomkins J, Jansen RK. (2006) Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. *Theoretical and Applied Genetics*. 112:1503–1518.

Daniell H, Lee SB, Panchal T, Wiebe PO. (2001) Expression of the native cholera toxin B subunit gene and assembly as functional oligomers in transgenic tobacco chloroplasts. *Journal of molecular biology*. 311(5):1001-1009.

Daniell H, Kumar S, Dufourmantel N. (2005) Breakthrough in chloroplast genetic engineering of agronomically important crops. *Trends in Biotechnology*. 23(5):238-245.

Danna K, Nathans D. (1971) Specific cleavage of simian virus 40 DNA by restriction endonuclease of hemophilus influenzae. *Proceedings of the National Academy of Science USA*. 68(12):2913-2917.

Darbyshire SJ. (1993) Realignment of *Festuca* subgenus *schedonorus* with the genus *Lolium* (Poaceae). *Novon*. 3(3):239-243.

Davis JID, Simmons MP, Stevenson DW, Wendel JF. (1998) Data decisiveness, data quality, and incongruence in phylogenetic analysis: An example from the monocotyledons using mitochondrial *atpA* sequences. *Systematic Biology*. 47(2):282-310.

Davis JID, Soreng RJ. (1993) Phylogenetic structure in the grass family (Poaceae) as inferred from chloroplast DNA restriction site variation. *American Journal of Botany*. 80(12):1444-1454.

Day DA. (2004) Mitochondrial structure and function in plants. In: Day DA, Millar AH, Whelan J. (eds). *Plant Mitochondria: From genome to function*. Kluwer Academics. Dordrecht.

del Campo EM, Sabater B, Martin M. (1997) Plastid *ndhD* gene of barley (*Hordeum vulgare* L.), sequence and transcript editing (Accession No. Y12258). *Plant Physiology*. 114:748.

Diamond AM, Montero-Puerner Y, Lee BJ, Hatfield D. (1990) Selenocysteine inserting tRNAs are likely generated by tRNA editing. *Nucleic Acids Research*. 18(22):6727.

Diekmann K, Hodkinson TR, Fricke E, Barth S. (2008) An optimized chloroplast DNA extraction protocol for grasses (Poaceae) proves suitable for whole plastid genome sequencing and SNP detection. *PLoS ONE*. 3(7): e2813. doi:10.1371/journal.pone.0002813.

Diekmann K, Hodkinson TR, Wolfe KH, van den Bekerom R, Dix PJ, Barth S. (2009) Complete chloroplast genome sequence of a major allogamous forage species, perennial ryegrass (*Lolium perenne* L.). *DNA Research*. 16(3):165-76.

do Vallo Ribeiro MAM, Gurney KA, Bush LP. (1996) Endophyte *Acremonium lolii* in ecotypes of perennial ryegrass (*Lolium perenne* L.) collected in old Irish pastures. *Irish Journal of Agricultural and Food Research*. 35(2):151-157.

Doebley J, Durbin M, Golenberg EM, Clegg MT, Ma DP. (1990) Evolutionary analysis of the large subunit of carboxylase (*rbcL*) nucleotide sequence among the grasses (Gramineae). *Evolution*. 44(4):1097-1108.

Drescher A, Hupfer H, Nickel C, Albertazzi F, Hohmann U, Herrmann RG, Maier RM. (2002) C-to-U conversion in the intercistronic *ndhI/ ndhG* RNA of plastids from monocot plants: conventional editing in an unconventional small reading frame? *Molecular Genetics and Genomics*. 267(2):262-9.

Droege M, Hill B. (2008) The Genome Sequencer FLX^{TM} System - Longer reads, more applications, straight forward bioinformatics and more complete data sets. *Journal of Biotechnology*. 136:3-10.

Du P, Jia L, Li Y. (2009) CURE-Chloroplast: a chloroplast C-to-U RNA editing predictor for seed plants. *BMC Bioinformatics*. 10:135.

Duff RJ, Nickrent DL. (1999) Phylogenetic relationships of land plants using mitochondrial small-subunit rDNA sequences. *American Journal of Botany*. 86(3):372-386.

Dumolin-Lapegue S, Pemonge MH, Petit RJ. (1997) An enlarged set of consensus primers for the study of organelle DNA in plants. *Molecular Ecology*. 6(4):393-7.

Duvall MR, Robinson JW, Mattson JG, Moore A. (2008) Phylogenetic analyses of two mitochondrial metabolic genes sampled in parallel from angiosperms find fundamental interlocus incongruence. *American Journal of Botany*. 95(7):871-884.

Eberhard S, Finazzi G, Wollman FA. (2008) The dynamics of photosynthesis. *Annual review of genetics*. 42:463-515.

Echt CS, Deverno LL, Anzidei M, Vendramin GG. (1998) Chloroplast microsatellites reveal population genetic diversity in red pine, *Pinus resinosa* Ait. *Molecular Ecology*. 7(3):307-316.

Eckardt NA. (2006) Cytoplasmic male sterility and fertility restoration. *The Plant Cell*. 18:515-517.

Excoffier L, Laval G, Schneider S. (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*. 1:47-50.

Felsenstein J. (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*. 39(4):783-791.

Fiebig A, Stegemann S, Bock R. (2004) Rapid evolution of RNA editing sites in a small non-essential plastid gene. *Nucleic Acids Research*. 32(12):3615-22.

Fjellheim S, Rognli OA. (2005) Molecular diversity of local Norwegian meadow fescue (*Festuca pratensis* Huds.) populations and Nordic cultivars-consequences for management and utilisation. *Theoretical and Applied Genetics*. 111(4):640-650.

Flannery ML, Mitchell FJ, Coyne S, Kavanagh TA, Burke JI, Salamin N, Dowding P, Hodkinson TR. (2006) Plastid genome characterisation in *Brassica* and Brassicaceae using a new set of nine SSRs. *Theoretical and Applied Genetics*. 113(7):1221-31.

Floyd DJ, Barker RE. (2002) Change of ryegrass seedling root fluorescence expression during three generations of seed increase. *Crop Science*. 42:905-911.

Freyer R, Hoch B, Neckermann K, Maier RM, Kössel H. (1993) RNA editing in maize chloroplasts is a processing step independent of splicing and cleavage to monocistronic mRNAs. *The Plant Journal*. 4(4):621-9.

Freyer R, Kiefer-Meyer MC, Kössel H. (1997) Occurrence of plastid RNA editing in all major lineages of land plants. *Proceedings of the National Academy of Science USA*. 94(12):6285-90.

Freyer R, López C, Maier RM, Martín M, Sabater B, Kössel H. (1995) Editing of the chloroplast *ndhB* encoded transcript shows divergence between closely related members of the grass family (Poaceae). *Plant Molecular Biology*. 29(4):679-84.

Gepts P. (2002) Ten thousand years of crop evolution. In: Chrispeels M, Sadava D. (eds). *Plants, Genes, and Crop Biotechnology*. 2nd edn. Bartlett and Jones. Sudbury. Massachusetts.

Gibala M, Szczesny B, Kieleczawa J, Janska H. (2004) The pea mitochondrial *atp6*: RNA editing and similarity of presequences in the Vicieae tribe. *Current genetics*. 46:235-239.

Giegé P, Brennicke A. (1999) RNA editing in Arabidopsis mitochondria effects 441 C to U changes in ORFs. *Proceedings of the National Academy of Science USA*. 96(26):15324-9.

Goremykin VV, Hirsch-Ernst KI, Wölfl S, Hellwig FH. (2003) Analysis of the *Amborella trichopoda* chloroplast genome sequence suggest that *Amborella* is not a basal angiosperm. *Molecular Biology and Evolution*. 20(9):1499-1505.

Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH. (2005) Analysis of *Acorus calamus* chloroplast genome and ist phylogenetic implications. *Molecular Biology and Evolution*. 22(9):1813-1822.

Goremykin VV, Salamini F, Velasco R, Viola R. (2008) Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. *Molecular Biology and Evolution*. 26(1):99-110.

Grass Phylogeny Working Group. (2001) Phylogeny and subfamilial classification of the grasses (Poaceae). *Annals of the Missouri Botanical Garden*. 88: 373–457.

Grau Nersting L, Bode Andersen S, von Bothmer R, Gullord M, Bagger Jørgensen R. (2006) Morphological and molecular diversity of Nordic oat through one hundred years of breeding. *Euphytica*. 150(3):327-337(11).

Gray MW, Burger G, Lang BF. (1999) Mitochondrial evolution. Science. 283(5407):1476-81.

Gualberto JM, Lamattina L, Bonnard G, Weil JH, Grienenberger JM. (1989) RNA editing in wheat mitochondria results in the conservation of protein sequences. *Nature*. 341(6243):660-2.

Guthridge KM, Dupal MP, Kölliker R, Jones ES, Smith KF, Forster JW. (2001) AFLP analysis of genetic diversity within and between populations of perennial ryegrass (*Lolium perenne* L.). *Euphytica*. 122(1):191-201.

Guzowska-Nowowiejska M, Fiedorowicz E, Plader W. (2009) Cucumber, melon, pumpkin, and squash: are rules of editing in flowering plants chloroplast genes so well known indeed? *Gene*. 434(1-2):1-8.

Hale ML, Borland AM, Gustafsson MH, Wolff K. (2004) Causes of size homoplasy among chloroplast microsatellites in closely related *Clusia* species. *Journal of Molecular Evolution*. 58(2):182-90.

Hall, TA. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*. 41:95-98.

Handa H. (2003) The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus* L.): comparative analysis of the mitochondrial genomes of rapeseed and *Arabidopsis thaliana*. *Nucleic Acids Research*. 31(20):5907-5916.

Handa H. (2008) Linear plasmids in plant mitochondria: Peaceful coexistences or malicious invasions? *Mitochondrion*. 8:15-25.

Hansen DR, Dastidar SG, Cai Z, Penaflor C, Kuehl JV, Boore JL, Jansen RK. (2007) Phylogenetic and evolutionary implications of complete chloroplast genome sequences of four early-diverging angiosperms: *Buxus* (Buxaceae), *Chloranthus* (Chloranthaceae), *Dioscorea* (Dioscoreaceae), and *Illicium* (Schisandraceae). *Molecular Phylogenetics and Evolution*. 45(2):547-63.

Hanson MR, Bentolila S. (2004) Interactions of mitochondrial and nuclear genes that affect male gametophyte development. *The Plant Cell*. 16:S154-S169.

Hayes ML, Hanson MR. (2007) Identification of a sequence motif critical for editing of a tobacco chloroplast transcript. *RNA*. 13(2):281-8.

Hebert PD, Cywinska A, Ball SL, deWaard JR. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*. 270(1512): 313–321.

Hebert PD, Stoeckle MY, Zemlak TS, Francis CM. (2004) Identification of Birds through DNA Barcodes. *PLoS Biology*. 2(10): e312.

Heckman JE, Alzner-Deweerd B, Rajbhandary UL. (1979) Interesting and unusual features in the sequence of *Neurospora crassa* mitochondrial tyrosine transfer RNA. *Proceedings of the National Academy of Science USA*. 76(2):717-721.

Hiesel R, Wissinger B, Schuster W, Brennicke A.(1989) RNA editing in plant mitochondria. *Science*. 246(4937):1632-4.

Hilu KW, Alice LA, Liang H. (1999) Phylogeny of Poaceae inferred from *matK* sequences. *Annals of the Missouri Botanical Garden*. 86(4):835-851.

Hirao T, Watanabe A, Kurita M, Kondo T, Takata K. (2008) Complete nucleotide sequence of the *Cryptomeria japonica* D. Don. chloroplast genome and comparative chloroplast genomics: Diversified genomic structure of coniferous species. *BMC Plant Biology*. 8:70.

Hirao T, Watanabe A, Kurita M, Kondo T, Takata K. (2009) A frameshift mutation of the chloroplast *matK* coding region is associated with chlorophyll defficiency in the *Cryptomeria japonica* virescent mutant Wogon-Sugi. *Current Genetics*. 55:311-321.

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun CR, Meng BY, Li Y-Q, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M. (1989) The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. *Molecular and General Genetics*. 217(2-3):185-94.

Hirose T, Fan H, Suzuki JY, Wakasugi T, Tsudzuki T, Kössel H, Sugiura M. (1996) Occurrence of silent RNA editing in chloroplasts: its species specificity and the influence of environmental and developmental conditions. *Plant Molecular Biology*. 30(3):667-72.

Hirose T, Kusumegi T, Tsudzuki T, Sugiura M. (1999) RNA editing sites in tobacco chloroplast transcripts: editing as a possible regulator of chloroplast RNA polymerase activity. *Molecular and General Genetics*. 262(3):462-7.

Hirose T, Sugiura M. (2001) Involvement of a site-specific trans-acting factor and a common RNA-binding protein in the editing of chloroplast mRNAs: development of a chloroplast in vitro RNA editing system. *The EMBO Journal*. 20(5):1144-52.

Hoch B, Maier RM, Appel K, Igloi GL, Kössel H. (1991) Editing of a chloroplast mRNA by creation of an initiation codon. *Nature*. 353(6340):178-80.

Hodkinson TR, Chase MW, Lledó MD, Salamin N, Renvoize SA. (2002a) Phylogenetics of *Miscanthus*, *Saccharum* and related genera (Saccharinae, Andropogoneae, Poaceae) based on DNA sequences from ITS nuclear ribosomal DNA and plastid *trnL* intron and *trnL-F* intergenic spacers. *Journal of Plant Research*. 115(5):381-92.

Hodkinson TR, Chase MW, Takahashi C, Leitch IJ, Bennett MD, Renvoize SA. (2002b) The use of DNA sequencing (ITS and *trnL-F*), AFLP, and fluorescent in situ hybridization to study allopolyploid *Miscanthus* (Poaceae). *American Journal of Botany*. 89(2):279-286.

Honda SI, Hongladarom T, Laties GG. (1966) A new isolation medium for plant organelles. *Journal of Experimental Botany*. 17(3):460-472.

Howe CJ, Barker RF, Bowman CM, Dyer TA. (1988) Common features of three inversions in wheat chloroplast DNA. *Current Genetics*. 13(4):343-349.

Hsiao C, Jacobs SWL, Chatterton NJ, Asay KH. (1999) A molecular phylogeny of the grass family (Poaceae) based on the sequences of nuclear ribosomal DNA (ITS). *Australian Systematic Botany*. 11(6):667 - 688.

Hubbard C. (1984) *Grasses - A guide to their structure, identification, uses and distribution in the British Isles*. 3rd edn. Penguin Books. London.

Huelsenbeck JP, Larget B, Miller RE, Ronquist F. (2002) Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology*. 51(5):673-688.

Huelsenbeck JP, Ronquist F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*. 17(8):754-5.

Humphreys MW, Yadav RS, Cairns AJ, Turner LB, Humphreys J, Skøt L. (2006) A changing climate for grassland research. *New Phytologist*. 169(1):9-26.

Igloi GL, Meinke A, Döry I, Kössel H. (1990) Nucleotide sequence of the maize chloroplast *rpo* B/C1/C2 operon: comparison between the derived protein primary structures from various organisms with respect to functional domains. *Molecular and General Genetics*. 221:379–394.

Inada M, Sasaki T, Yukawa M, Tsudzuki T, Sugiura M. (2004) A systematic search for RNA editing sites in pea chloroplasts: an editing event causes diversification from the evolutionarily conserved amino acid sequence. *Plant Cell Physiology*. 45(11):1615-22.

Jakobsson M, Säll T, Lind-Halldén C, Halldén C. (2007) Evolution of chloroplast mononucleotide microsatellites in *Arabidopsis thaliana*. *Theoretical and Applied Genetics*. 114(2):223-35.

Jansen RK, Raubeson LA, Boore JL, dePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ, Fourcade HM, Kuehl JV, McNeal JR, Leebens-Mack J, Cui L. (2005) Methods for obtaining and

analyzing whole chloroplast genome sequences. *Methods in Enzymology.* 395:348-84.

Jarne P, LAGODA PJL. (1996) Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution.* 11:424-430

Jensen LB, Andersen JR, Frei U, Xing Y, Taylor C, Holm PB, Lübberstedt T. (2005) QTL mapping of vernalization response in perennial ryegrass (*Lolium perenne* L.) reveals co-location with an orthologue of wheat VRN1. *Theoretical and Applied Genetics.* 110(3):527-36.

Katayama H, Ogihara Y. (1996) Phylogenetic affinities of the grasses to other monocots as revealed by molecular analysis of chloroplast DNA. *Current Genetics.* 29:572-581.

Keeling PJ, Palmer JD. (2008) Horizontal gene transfer in eukaryotic evolution. *Nature Reviews. Genetics.* 9:605-618.

Keeling PJ. (2004) Diversity and evolutionary history of plastids and their hosts. *American Journal of Botany.* 91(10)1481-1493.

Kelchner SA, Wendel JF. (1996) Hairpins create minute inversions in non-coding regions of chloroplast DNA. *Current Genetics.* 30(3):259-62.

Kelchner SA. (2000) The evolution of non-coding chloroplast DNA and its application in plant systematics. *Annals of the Missouri Botanical Garden.* 87(4):482-498.

Kempken F, Mullen JA, Pring DR, Tang HV. (1991) RNA editing of *Sorghum* mitochondrial *atp6* transcripts changes 15 amino acids and generates a carboxy-terminus identical to yeast. *Current Genetics.* 20(5):417-22.

Kiang A-S, Connolly V, McConnell DJ, Kavanagh TA. (1993) Cytoplasmic male sterility (CMS) in *Lolium perenne* L.: 1. Development of a diagnostic probe for the male-sterile cytoplasm. *Theoretical and Applied Genetics.* 86(6):781-787.

Kiang A-S, Connolly V, McConnell DJ, Kavanagh TA. (1994) Paternal inheritance of mitochondria and chloroplasts in *Festuca pratensis-Lolium perenne* intergeneric hybrids. *Theoretical and Applied Genetics.* 87(6):681-688.

Kiang A-S. (1992) *Molecular analysis of cytoplasmic male sterility in ryegrass hybrids.* Ph.D. thesis. Trinity College Dublin. Ireland.

Kiang A-S, Kavanagh TA. (1996) Cytoplasmic male sterility (CMS) in *Lolium perenne* L. 2. The mitochondrial genome of a CMS line is rearranged and contains a chimaeric *atp9* gene. *Theoretical and Applied Genetics.* 92(3-4):308-315.

Kim KJ, Lee HL. (2004) Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Research*. 11(4):247-61.

Kim KJ, Lee HL. (2005) Widespread occurence of small inversions in the chloroplast genomes of land plants. *Molecular and Cells*. 19(1):104-113.

Kirk JT0. (1963). The deoxyribonucleic acid of broad bean chloroplasts. *Biochimica et Biophysica Acta*. 76:417-424.

Kirk JTO. (1986) The discovery of chloroplast DNA. *BioEssays*. 4(1):36-38.

Kislev N, Swift H, Bogorad L. (1965) Nucleic acids of chloroplasts and mitochondria in Swiss chard. *The Journal of Cell Biology*. 25:327-344.

Knoll AH. (1992) The early evolution of eukaryotes: A geological perspective. *Science*. 256(5057):622-627.

Knox EB, Downie SR, Palmer JD. (1993) Chloroplast genome rearrangements and the evolution of giant lobelias from herbaceous ancestors. *Molecular Biology and Evolution*. 10(2): 414-430.

Konainge EMK, Singh SP, Gepts P. (1996) Genetic control of the domestican syndrome in common bean. *Crop Science*. 36:1037-1045.

Konishi T, Shinohara K, Yamada K, Sasaki Y. (1996) Acetyl-CoA carboxylase in higher plants: most plants other than gramineae have both the prokaryotic and the eukaryotic forms of this enzyme. *Plant and Cell Physiology*. 37(2):117-22.

Kotera E, Tasaka M, Shikanai T. (2005) A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. *Nature*. 433(7023):326-30.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH. (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Science USA*. 102(23):8369-74.

Kress WJ, Erickson DL. (2007) A two-locus global DNA barcode for land plants: The coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PLoS ONE*. 2(6): e508. doi:10.1371/journal.pone.0000508.

Kubik C, Meyer WA, Gaut BS. (1999) Assessing the abundance and polymorphism of simple sequence repeats in perennial ryegrass. *Crop Science*. 39:1136-1141.

Kubo T, Newton KJ. (2008) Angiosperm mitochondrial genomes and mutations. *Mitochondrion*. 8:5-14.

Kubo T, Mikami T. (2007) Organization and variation of angiosperm mitochondrial genome. *Physiology Plantarum*. 129:6-13.

Kudla J, Bock R. (1999) RNA editing in an untranslated region of the Ginkgo chloroplast genome. *Gene*. 234(1):81-6.

Kugita M, Kaneko A, Yamamoto Y, Takeya Y, Matsumoto T, Yoshinaga K. (2003) The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the earliest land plants. *Nucleic Acids Research*. 31(2):716-21.

Kumar S, Tamura K, Nei M. (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics*. 5:150-163.

Lang BF, Burger G. (2007) Purification of mitochondrial and plastid DNA. *Nature Protocols*. 2(3):652-660.

Laser KD, Lersten NR. (1972) Anatomy and cytology of microsporogenesis in cytoplasmic male sterile angiosperms. *Botanical Review*. 38(3):425-454.

Lee J, Young JPW. (2009) The mitochondrial genome sequence of the arbuscular mycorrhizal fungus *Glomus intraradices* isolate 494 and implications for the phylogenetic placement of *Glomus*. *New Phytologist*. 183:200-211.

Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, dePamphilis CW. (2005) Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. *Molecular Biology and Evolution*. 22(10):1948-1963.

Lenuweit U, Gharadjedaghi B. (2002) Biologische Basisdaten zu *Lolium perenne*, *Lolium multiflorum*, *Festuca pratensis* and *Trifolium repens*. (Hrsg. Umweltbundesamt, Berlin, Germany). http://www.umweltbundesamt.de/uba-info-medien/dateien/2103.htm. Accessed: 17.12.2009.

Lilly JW, Havey MJ, Jackson SA, Jiang J. (2001) Cytogenomic analyses reveal the structural plasticity of the chloroplast genome in higher plants. *The Plant Cell*. 13:245-254.

Lopez C, Freyer R, Guera A, Maier RM, Martin M, Sabater B, Kössel H. (1997) Sequence of *ndhA* gene of barley (*Hordeum vulgare* L.) plastid (accession Nos. Y13729 & Y13730). Transcript editing in Graminean organs. *Plant Physiology*. 115:313.

Maier RM, Neckermann K, Hoch B, Akhmedov NB, Kössel H. (1992) Identification of editing positions in the *ndhB* transcript from maize chloroplasts reveals

sequence similarities between editing sites of chloroplasts and plant mitochondria. *Nucleic Acids Research*. 20(23):6189-94.

Maier RM, Neckermann K, Igloi GL, Kössel H. (1995) Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *Journal of Molecular Biology*. 251:614–628.

Manning JE, Wolstenholme DR, Ryan RS, Hunter JA, Richards OC. (1971) Circular chloroplast DNA from *Euglena gracilis*. *Proceedings of the National Academy of Science USA*. 68(6):1169-1173.

Mardanov AV, Ravin NV, Kuznetsov BB, Samigullin TH, Antonov AS, Kolganova TV, Skyabin KG. (2008) Complete sequence of the duckweed (*Lemna minor*) chloroplast genome: structural organization and phylogenetic relationships to other angiosperms. *Journal of Molecular Evolution*. 66(6):555-64.

Margulis L (1970) *Origin of eukaryotic cells*. Yale University Press. New Haven.

Martin W, Müller M. (1998) The hydrogen hypothesis for the first eukaryote. *Nature*. 392(6671):37-41.

Martinson C. (2007) *Looking back at nearly 50 years at Iowa State*. In: Essays on the college of agriculture's history. http://www.ag.iastate.edu/coa150/martinson.php. Accessed: 18.12.2009.

Mathews S, Tsai RC, Kellogg EA. (2000) Phylogenetic structure in the grass family (Poaceae): Evidence from the nuclear gene phytochrome B. *American Journal of Botany*. 87(1):96-107.

Maxam AM, Gilbert W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Science USA*. 74(2):560-564.

Mayor C, Brudno M, Schwartz JR, Pollakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I. (2000) VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics Applications Note*. 16(11):1046-1047.

McDermott P, Connolly V, Kavanagh TA. (2008) The mitochondrial genome of a cytoplasmic male sterile line of perennial ryegrass (*Lolium perenne* L.) contains an integrated linear plasmid-like element. *Theoretical and Applied Genetics*. 117(3):459-70.

McGrath S, Hodkinson TR, Barth S. (2007) Extremely high cytoplasmic diversity in natural and breeding populations of *Lolium* (Poaceae). *Heredity*. 99(5):531-44.

McGrath S, Hodkinson TR, Salamin N, Barth S. (2006) Development and testing of novel chloroplast microsatellite markers for *Lolium perenne* and other grasses (Poaceae) from de novo sequencing and *in silico* sequences. *Molecular Ecology Notes*. 6(2):449-452.

McGrath S. (2008) *The characterisation of genetic diversity of a collection of perennial ryegrass (*Lolium perenne *L.)*. Ph.D. thesis. Trinity College Dublin. Ireland.

Mereschkowsky C. (1905) Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. *Biologisches Centralblatt*. 25:593-604.

Metzgar D, Bytof J, Wills C. (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Research*. 10(1):72-80.

Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW, Calie PJ, Jermiin LS, Wolfe KH. (2001) Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. *The Plant Cell*. 13:645-658.

Miyamoto T, Obokata J, Sugiura M. (2002) Recognition of RNA editing sites is directed by unique proteins in chloroplasts: biochemical identification of *cis*-acting elements and trans-acting factors involved in RNA editing in tobacco and pea chloroplasts. *Molecular and Cellular Biology*. 22(19):6726-34.

Miyata Y, Sugita M. (2004) Tissue- and stage-specific RNA editing of *rps14* transcripts in moss (*Physcomitrella patens*) chloroplasts. *Journal of Plant Physiology*. 161(1):113-5.

Moritz C, Cicero C. (2004) DNA barcoding: promise and pitfalls. *PLoS Biology*. 2(10):e354.

Mower JP, Palmer JD. (2006) Patterns of partial RNA editing in mitochondrial genes of *Beta vulgaris*. *Molecular Genetics and Genomics*. 276(3):285-93.

Mower JP, Stefanović S, Young GJ, Palmer JD. (2006) Plant genetics: Gene transfer from parasitic to host plants. *Nature*. 432:165-166.

Mulligan RM. (2004) RNA editing in plant organelles. In: Daniell H, Chase C, editors. *Molecular biology and biotechnology of plant organelles*. Dordrecht: Springer. pp. 239–260.

Mulligan RM, Dement E, Andre C, Walbot V. (1988) Isolation of RNA and DNA from rice seedling mitochondria. *Rice Genetics Newsletter*. 5:58.

Mullins E, Ryan E, Meade C. (2009) *Potential for gene-flow from cultivated Irish grasses and cereals.* End of Project Report 5181. http://www.teagasc.ie/research/_reports/crops/5181/eopr5181.asp. Accessed: 19.12.2009.

Nadot S, Bajon R, Lejeune B. (1994) The chloroplast gene rps4 as a tool for the study of Poaceae phylogeny. *Plant Systematics and Evolution.* 191:27-38.

Nadot S, Bittar G, Carter L, Lacroix R, Lejeune B. (1995) A phylogenetic analysis of monocotyledons based on the chloroplast gene *rps4*, using parsimony and a new numerical phenetics method. *Molecular Phylogenetics and Evolution.* 4(3): 257-282.

Nass MMK, Nass S. (1963) Intramitochondrial fibers with DNA characteristics. I. Fixation and electron staining reactions. *The Journal of Cell Biology.* 19:593-611.

Nei M. (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics.* 89(3):583-590.

Nei M. (1996) Phylogenetic analysis in molecular evolutionary genetics. *Annual Review of Genetics.* 30:371-403.

Nei M, Li W-H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Science USA.* 76(10):5269-5273.

Oda K, Yamato K, Ohta E, Nakamura Y, Takemura M, Nozato N, Akashi K, Kanegae T, Ogura Y, Kohchi T, Ohyama K. (1992) Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA. A primitive form of plant mitochondrial genome. *Journal of Molecular Biology.* 223(1):1-7.

Ogihara Y, Isono K, Kojima T, Endo A, Hanaoka M, Shiina T, Terachi T, Utsugi S, Murata M, Mori N, Takumi S, Ikeo K, Gojobori T, Murai R, Murai K, Matsuoka Y, Ohnishi Y, Tajiri H, Tsunewaki K. (2002) Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. *Molecular genetic and genomics.* 266:740-746.

Ogihara Y, Yamazaki Y, Murai K, Kanno A, Terachi T, Shiina T, Miyashita N, Nasuda S, Nakamura C, Mori N, Takumi S, Murata M, Futo S, Tsunewaki K. (2005) Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. *Nucleic Acids Research.* 33(19):6235-6250.

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, Aota S, Inokuchi H, Ozeki H. (1986) Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature*. 322:572 - 574.

Ohyama K. (1996) Chloroplast and mitochondrial genomes from a liverwort, *Marchantia polymorpha* - gene organization and molecular evolution. *Bioscience, Biotechnology, and Biochemistry*. 60(1):16-24.

Okuda K, Chateigner-Boutin AL, Nakamura T, Delannoy E, Sugita M, Myouga F, Motohashi R, Shinozaki K, Small I, Shikanai T. (2009) Pentatricopeptide repeat proteins with the DYW motif have distinct molecular functions in RNA editing and RNA cleavage in *Arabidopsis* chloroplasts. *Plant Cell*. 21(1):146-56.

Okuda K, Myouga F, Motohashi R, Shinozaki K, Shikanai T. (2007) Conserved domain structure of pentatricopeptide repeat proteins involved in chloroplast RNA editing. *Proceedings of the National Academy of Science USA*. 104(19):8178-83.

Okuda K, Nakamura T, Sugita M, Shimizu T, Shikanai T. (2006) A pentatricopeptide repeat protein is a site recognition factor in chloroplast RNA editing. *The Journal of Biological Chemistry*. 281(49):37661-7.

Olmstead RG, Palmer JD. (1994) Chloroplast DNA systematics: A review of methods and data analysis. *American Journal of Botany*. 81(9):1205-1224.

Olson JM (2006) Photosynthesis in the Archean Era. *Photosynthesis Research*. 88:109-117.

Page, RDM. (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences*. 12(4):357-8.

Palmer JD. (1986) Isolation and structural-analysis of chloroplast DNA. *Methods in Enzymology*. 118:167-186.

Palmer JD. (1987) Chloroplast DNA evolution and biosystematic uses of chloroplast DNA variation. *The American Naturalist*. 130:S6.

Palmer JD. (1991) *Plastid chromosomes: structure and evolution*. In: Vasil IK, Bogorad L. (eds.) *Cell Culture and Somatic Cell Genetics of Plants*. vol 7A. The Molecular Biology of Plastids. Academic Press. San Diego, pp. 5–53.

Palmer JD. (1992) *Mitochondrial DNA in plant systematics*. In: Soltis PS, Soltis DE, Doyle JJ. (eds). *Molecular systematics of plants*. Chapman and Hall. New York. USA.

Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu Y-L, Song K. (2000) Dynamic evolution of plant mitochondrial genomes: Mobile genes and introns and highly variable mutation rates. *Proceedings of the National Academy of Science USA*. 97(13):6960-6966.

Palmer JD, Jansen RK, Michaels HJ, Chase MW, Manhart JR. (1988) Chloroplast DNA variation and plant phylogeny. *Annals of the Missouri Botanical Garden*. 75(4):1180-1206.

Palmer JD, Soltis DE, Chase MW. (2004) The plant tree of life: An overview and some points of view. *American Journal of Botany*. 91(10):1437-1445.

Palmer JD, Zamir D. (1982) Chloroplast DNA evolution and phylogenetic relationships in *Lycopersicon*. *Proceedings of the National Academy of Science USA*. 79:5006-5010.

Peeters NM, Hanson MR. (2002) Transcript abundance supercedes editing efficiency as a factor in developmental variation of chloroplast gene expression. RNA. 8(4):497-511.

Pennisi E. (2007) Wanted: A barcode for plants. *Science*. 318(5848):190-191.

Pearce G, Strydom D, Johnson S, Ryan CA. (1991) A polypeptide from tomato leaves induces wound-inducible proteinase inhibitor proteins. *Science*. 253(5022):895-898.

Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J. (1987) A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell*. 50(6):831-40.

Powell W, Morgante M, Andre C, McNicol JW, Machray GC, Doyle JJ, Tingey SV, Rafalski JA. (1995b) Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome. *Current Biology*. 5(9):1023-9.

Powell W, Morgante M, McDevitt R, Vendramin GG, Rafalski JA. (1995a) Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines. *Proceedings of the National Academy of Science USA*. 92(17):7759-63.

Pring DR, Levings III CS. (1978) Heterogeneity of maize cytoplasmic genomes among male-sterile cytoplasms. *Genetics*. 89:121-136.

Provan J, Biss PM, McMeel D, Mathews S. (2004) Universal primers for the amplification of chloroplast microsatellites in grasses (Poaceae). *Molecular Ecology Notes*. 4:262-264.

Provan J, Powell W, Hollingsworth PM. (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends in Ecology and Evolution*. 16(3):142-147.

Provan J, Russell JR, Booth A, Powell W. (1999) Polymorphic chloroplast simple sequence repeat primers for systematic and population studies in the genus *Hordeum*. *Molecular Ecology*. 8(3):505-11.

Qiu Y-L, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW. (1999) The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature*. 402:404-407.

Rajendrakumar P, Biswal AK, Balachandran SM, Srinivasarao K, Sundaram RM. (2006) Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions. *Bioinformatics*. 23(1):1-4.

Rambaut A (2007) *FigTree, a graphical viewer of phylogenetic trees*. http://tree.bio.ed.ac.uk/software/figtree/. Accessed: 18.12.2009.

Ravi V, Khurana JP, Tyagi AK, Khurana P. (2007) An update on chloroplast genomes. *Plant Systematics and Evolution*. 271(1-2):101-122.

Reichman JR, Watrud LS, Lee EH, Burdick CA, Bollman MA, Storm MJ, King GA, Mallory-Smith C. (2006) Establishment of transgenic herbicide-resistant creeping bentgrass (*Agrostis stolonifera* L.) in nonagronomic habitats. *Molecular Ecology*. 15:4243-4255.

Rheinberger HJ. (2000) Mendelian inheritance in Germany between 1900 and 1910. The case of Carl Correns (1864–1933). *Comptes rendus de l'Académie des sciences. Série III. Sciences de la vie*. 323:1089-1096.

Ris H, Plaut W. (1962) Ultrastructure of DNA-containing areas in the chloroplast of *Chlamydomonas*. *The Journal of Cell Biology*. 13:383-391.

Rolfsmeier ML, Dixon MJ, Lahue RS. (2000) Mismatch repair blocks expansions of interrupted trinucleotide repeats in yeast. *Molecular Cell*. 6(6):1501-7.

Ronquist F, Huelsenbeck JP. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 19(12):1572-4.

Roper JM, Hansen SK, Wolf PG, Karol KG, Mandoli DF, Everett KDE. (2007) The complete plastid genome sequence of *Angiopteris evecta* (G. Forst.) Hoffm. (Marattiaceae). *American Fern Journal*. 97(2):95-106.

Rouwendal GJA, Creemers-Molenaar J, Krens FA. (1992) Molecular aspects of cytoplasmic male sterility in perennial ryegrass (*Lolium perenne* L.): mtDNA and RNA differences between plants with male-sterile and fertile cytoplasm and restriction mapping of their *atp6* and *coxl* homologous regions. *Theoretical and Applied Genetics*. 83:330-336.

Ruf S, Kössel H.(1997) Tissue-specific and differential editing of the two *ycf3* editing sites in maize plastids. *Current Genetics*. 32(1):19-23.

Ruiz ON, Daniell H. (2005) Engineering cytoplasmic male sterility via the chloroplast genome by expression of β-ketothiolase. *Plant physiology.* 138(3):1232-1246.

Ryan E, Mullins E, Burke J, Downes M, Meade C. (2006) Tracing field hybridization in ryegrass species using microsatellite and morphological markers. *Environmental Biosafety Research*. 5(2):111-7.

Sagan L. (1967) On the origin of mitosing cells. *Journal of Theoretical Biology*. 14(3):255-74.

Sager R. (1954) Mendelian and non-mendelian inheritance of streptomycin resistance in *Chlamydomonas reinhardi*. *Proceedings of the National Academy of Science USA*. 40(5):356-63.

Sager R, Ishida MR. (1963) Chloroplast DNA in *Chlamydomonas*. *Proceedings of the National Academy of Science USA*. 50:725-730.

Salamin N, Chase MW, Hodkinson TR, Savolainen V. (2003) Assessing internal support with large phylogenetic DNA matrices. *Molecular Phylogenetics and Evolution*. 27(3):528-39.

Säll T, Jakobsson M, Lind-Hallden C, Hallden C. (2003) Chloroplast DNA indicates a single origin of the allotetraploid *Arabidopsis suecica*. *Journal of Evolutionary Biology*. 16:1019-1029.

Sambrook J, Russel DW. (2001) *Molecular cloning*. A labaratory manual. 3rd edn. Cold Spring Harbor. New York.

Sanderson MJ, Wojciechowski MF, Hu JM, Khan TS, Brady SG. (2000) Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. *Molecular Biology and Evolution*. 17(5):782-97.

Sang T. (2002) Utility of low-copy nuclear gene sequences in plant phylogenetics. Critical reviews in biochemistry and molecular biology. 37(3):121-47.

Sanger F, Coulson AR, Barrell BG, Smith AJ, Roe BA. (1980) Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *Journal of Molecular Biology*. 143(2):161-78.

Sanger F, Nicklein S, Coulson AR. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Science USA*. 74(12):5463-5467.

Sapp J, Carrapiço F, Zolotonosov M. (2002) Symbiogenesis: the hidden face of Constantin Merezhkowsky. *History and philosophy of the life sciences*. 24(3-4):413-40.

Saski C, Lee SB, Fjellheim S, Guda C, Jansen RK, Luo H, Tomkins J, Rognli OA, Daniell H, Clarke JL. (2007) Complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes. *Theoretical and Applied Genetics*. 115(4):571-90.

Sato M, Yamashita M, Kato S, Mikami T, Shimamoto Y. (1995) Mitochondrial DNA analysis reveals cytoplasmic variation within a single cultivar of perennial ryegrass (*Lolium perenne* L.). *Euphytica*. 83(3):205-208.

Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S. (1999) Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Research*. 6:283–90.

Sánchez-Ken JG, Clark LG, Kellogg EA, Kay EE. (2007) Reinstatement and emendation of subfamily Micrairoideae (Poaceae). *Systematic Botany*. 32(1):71-80.

Savolainen V, Chase MW. (2003) A decade of progress in plant molecular phylogenetics. *Trends in Genetics*. 19(12):717-724.

Schatz G, Haslbrunner E, Tuppy H. (1964) Deoxyribonucleic acid associated with yeast mitochondria. *Biochemical and Biophysical Research Communications*. 15(2):127-132.

Schimper AFW (1883) Über die Entwickelung der Chlorophyllkörner und Farbkörper. *Botanische Zeitung*. 41:105-113, 121-131, 137-146, 153-162.

Schnable PS, Wise RP. (1998) The molecular basis of cytoplasmic male sterility and fertility restoration. *Trends in Plant Science*. 3(5):175-180.

Schuster W, Ternes R, Knoop V, Hiesel R, Wissinger B, Brennicke A. (1991) Distribution of RNA editing sites in *Oenothera* mitochondrial mRNAs and rRNAs. *Current Genetics*. 20(5):397-404.

Schuster W, Wissinger B, Unseld M, Brennicke A. (1990) Transcripts of the NADH-dehydrogenase subunit 3 gene are differentially edited in *Oenothera* mitochondria. *The EMBO Journal*. 9(1):263-9.

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W. (2000) PipMaker - a web server for aligning two genomic DNA sequences. *Genome research*. 10(4):577-86.

Schwarz ZS, Kössel H. (1980) The primary structure of 16S rDNA from *Zea mays* chloroplast is homologous to *E. coli* 16S rRNA. *Nature*. 283(5749):739-742.

Shaw J, Lickey EB, Schilling EE, Small RL. (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III. *American Journal of Botany*. 94(3):275-288.

Sheppard AE, Ayliffe MA, Blatch L, Day A, Delaney SK, Khairul-Fahmy N, Li Y, Madesis P, Pryor AJ, Timmis JN. (2008) Transfer of plastid DNA to the nucleus is elevated during male gametogenesis in tobacco. *Plant Physiology*. 148:328-336.

Shields DC, Wolfe KH. (1997) Accelerated evolution of sites undergoing mRNA editing in plant mitochondria and chloroplasts. *Molecular Biology and Evolution*. 14(3):344-9.

Shikanai T. (2006) RNA editing in plant organelles: machinery, physiological function and evolution. *Cellular and molecular life sciences.* 63(6):698-708.

Shimada H, Fukuta M, Ishikawa M, Sugiura M. (1990) Rice chloroplast RNA polymerase genes: the absence of an intron in *rpoC1* and the presence of an extra sequence in *rpoC*. *Molecular and General Genetics*. 221:395–402.

Shimamoto Y, Tominaga Y, Sato M, Mikami T. (1997) Mitochondrial genome polymorphism in *Lolium perenne*. Proceedings XVIII International Grassland Congress.Winnepeg. Manitoba

Small ID, Peeters N. (2000) The PPR motif - a TPR-related motif prevalent in plant organellar proteins. *Trends in Biochemical Sciences*. 25(2):46-7.

Smith HC, Gott JM, Hanson MR. (1997) A guide to RNA editing. RNA. 3(10):1105-23.

Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu Y-L, Chase MW, Farris JS, Stefanovic´ S, Rice DW, Palmer JD, Soltis PS. (2004) Genome-scale data, angiosperm relationships, and 'ending incongruence': A cautionary tale in phylogenetics. *Trends in Plant Science*. 9(10):477-483.

Soltis PS, Soltis DE, Chase MW. (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature*. 402:402-404.

Soreng RJ, Davis JI. (1998) Phylogenetics and character evolution in the grass family (Poaceae): Simultaneous analysis of morphological and chloroplast DNA restriction site character sets. *Botanical Review*. 64(1):1-85.

Stern DB, Lonsdale DM. (1982) Mitochondrial and chloroplast genomes of maize have a 12-kilobase DNA sequence in common. *Nature*. 299:698-702.

Stupar RM, Lilly JW, Town CD, Cheng Z, Kaul S, Buell CR, Jiang J. (2001) Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: Implication of potential sequencing errors caused by large-unit repeats. *Proceedings of the National Academy of Science USA*. 98(9):5099-5103.

Sugita M, Sugiura M. (1996) Regulation of gene expression in chloroplasts of higher plants. *Plant Molecular Biology*. 32:315–326.

Sugiura M. (1992) The chloroplast genome. *Plant Molecular Biology*. 19:149–168.

Sugiyama Y, Watase Y, Nagase M, Makita N, Yagura S, Hirai A, Sugiura M. (2005) The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. *Molecular Genetics and Genomics*. 272:603-615.

Sungkaew S, Stapleton CMA, Salamin N, Hodkinson TR. (2009) Non-monophyly of the woody bamboos (Bambueae; Poaceae): A multi-gene region phylogenetic analysis of Bambusoideae s.s. *Journal of Plant Research*. 122:95-108.

Swofford DL. (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Taberlet P, Gielly L, Pautou G, Bouvet J. (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology*. 17(5):1105-9.

Takahashi K, Nei M. (2000) Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum

likelihood when a large number of sequences are used. *Molecular Biology and Evolution*. 17(8):1251-1258.

Tanaka M, Wakasugi T, Sugita M, Shinozaki K, Sugiura M. (1986) Genes for the eight ribosomal proteins are clustered on the chloroplast genome of tobacco (*Nicotiana tabacum*): similarity to the S10 and spc operons of *Escherichia coli*. *Proceedings of the National Academy of Science USA*. 83(16):6030-4.

Tanksley SD, McCouch SR. (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. *Science*. 277(5329):1063-6.

Tatum LA. (1971) The southern corn leaf blight epidemic. *Science*. 171(3976):1113-1116.

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP. (2003) A plea for DNA taxonomy. *Trends in Ecology and Evolution*. 18(2):70-74.

The Angiosperm Phylogeny Group. (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society*. 161:105-121.

The Rice Chromosome 10 Sequencing Consortium (2003) In-depth view of structure, activity, and evolution of rice chromosome 10. Science. 300:1566-1569.

Thorogood D, Hayward MD. (1992) Self-compatibility in *Lolium temulentum* L: its genetic control and transfer into *L. perenne* L. and *L. multiflorum* Lam. *Heredity*. 68:71-78.

Tillich M, Funk HT, Schmitz-Linneweber C, Poltnigg P, Sabater B, Martin M, Maier RM. (2005) Editing of plastid RNA in *Arabidopsis thaliana* ecotypes. *The Plant Journal*. 43(5):708-15.

Tillich M, Lehwark P, Morton BR, Maier UG. (2006) The evolution of chloroplast RNA editing. *Molecular Biology and Evolution*. 23(10):1912-21.

Timme RE, Kuehl JV, Boore JL, Jansen RK. (2007) A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. *American Journal of Botany*. 94(3): 302–312.

Timmis JN, Ayliffe MA, Huang CY, Martin W. (2004) Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nature Reviews. Genetics*. 5:123-135.

Torrecilla P, Catalán P. (2002) Phylogeny of broad-leaved and fine-leaved *Festuca* lineages (Poaceae) based on nuclear ITS sequences. *Systematic Botany*. 27(2):241-251.

Tovar J, León-Avila G, Sánchez LB, Sutak R, Tachezy J, van der Giezen M, Hernández M, Müller M, Lucocq JM. (2003) Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. *Nature*. 426:172-176.

Trejo-Calzada R, O'Connell MA. (2005) Genetic diversity of drought-responsive genes in populations of the desert forage *Dactylis glomerata*. *Plant Science*. 168:1327-1335.

Triboush SO, Danilenko NG, Davydenko OG. (1998) A method for isolation of chloroplast DNA and mitochondrial DNA from sunflower. *Plant Molecular Biology Reporter* 16:183-189.

Tsudzuki J, Nakashima K, Tsudzuki T, Hiratsuka J, Shibata M, Wakasugi T, Sugiura M. (1992) Chloroplast DNA of black pine retains a residual inverted repeat lacking rRNA genes: nucleotide sequences of *trnQ*, *trnK*, *psbA*, *trnI* and *trnH* and the absence of *rps16*. *Molecular and General Genetics*. 232(2):206-14.

Tsumura Y, Suyama Y, Yoshimura K. (2000) Chloroplast DNA inversion polymorphism in populations of *Abies* and *Tsuga*. *Molecular Biology and Evolution*. 17(9):1302-1312.

Turmel M, Otis C, Lemieux C. (2005) The complete chloroplast DNA sequences of the charophycean green algae *Staurastrum* and *Zygnema* reveal that the chloroplast genome underwent extensive changes during the evolution of the Zygnematales. *BMC Biology*. 3:22.

Tzvelev NN. (1989) The system of grasses (Poaceae) and their evolution. *Botanical Review*. 55(3):141-204.

Ubi BE, Kölliker R, Fujimori M, Komatsu T. (2003) Genetic diversity in diploid cultivars of rhodesgrass determined on the basis of amplified fragment length polymorphism markers. *Crop Science*. 43:1516-1522.

Unseld M, Marienfeld JR, Brandt P, Brennicke A. (1997) The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. *Nature Genetics*. 15(1):57-61.

Vaughan DA, Lu B-R, Tomooka N. (2008) The evolving story of rice evolution. *Plant Science*. 174:394-408.

Vogel J, Hübschmann T, Börner T, Hess WR. (1997) Splicing and intron-internal RNA editing of *trnK-matK* transcripts in barley plastids: support for *matK* as an essential splice factor. *Journal of Molecular Biology*. 270(2):179-87.

Wakasugi T, Hirose T, Horihata M, Tsudzuki T, Kössel H, Sugiura M. (1996) Creation of a novel protein-coding region at the RNA level in black pine chloroplasts: the pattern of RNA editing in the gymnosperm chloroplast is different from that in angiosperms. *Proceedings of the National Academy of Science USA*. 93(16):8766-70.

Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M. (1994) Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. *Proceedings of the National Academy of Science USA*. 91:9794-9798.

Wallin IE. (1923) The mitochondria problem. *The American Naturalist*. 57(650):255-261.

Wang Z, Zou Y, Li X, Zhang Q, Chen L, Wu H, Su D, Chen Y, Guo J, Luo D, Long Y, Zhong Y, Liu Y-G. (2006) Cytoplasmic male sterility of rice with Boro II cytoplasm is caused by a cytotoxic peptide and is restored by two related PPR motif genes via distinct modes of mRNA silencing. *The Plant Cell*. 18:676-687.

Ward BL, Anderson RS, Bendich AJ. (1981) The mitochondrial genome is large and variable in a family of plants (Cucurbitaceae). *Cell*. 25:793-803.

Warnke SE, Barker RE, Brilman LA, Young III WC, Cook RL. (2002) Inheritance of superoxide dismutase (*sod-1*) in a perennial × annual ryegrass cross and its allelic distribution among cultivars. *Theoretical and Applied Genetics*. 105:1146-1150.

Watanabe N, Nakazono M, Kanno A, Tsutsumi N, Hirai A. (1994) Evolutionary variations in DNA sequences transferred from chloroplast genomes to mitochondrial genomes in the Gramineae. *Current Genetics*. 26(5-6):512-518.

Weglöhner W, Subramanian AR. (1991) A heptapeptide repeat contributes to the unusual length of chloroplast ribosomal protein S18. Nucleotide sequence and map position of the *rpl33-rps18* gene cluster in maize. *FEBS Letters*. 279(2):193-197.

Wells R, Birnstiel ML. (1967) A rapidly renaturing deoxyribonucleic acid component associated with chloroplast preparations. *Proceedings of the biochemical society. The Biochemical Journal*. 105:53-54.

Wendel JF, Schnabel A, Seelanan T. (1995) An unusual ribosomal DNA sequence from *Gossypium gossypioides* reveals ancient, cryptic, intergenomic introgression. *Molecular Phylogenetics and Evolution*. 4(3):298-313.

Wickett NJ, Zhang Y, Hansen SK, Roper JM, Kuehl JV, Plock SA, Wolf PG, DePamphilis CW, Boore JL, Goffinet B. (2008) Functional gene losses occur with minimal size reduction in the plastid genome of the parasitic liverwort *Aneura mirabilis*. *Molecular and Biology Evolution*. 25(2):393-401.

Wilkins PW. (1991) Breeding perennial ryegrass for agriculture. *Euphytica* 52:201-214.

Willis KJ, McElwain C. (2002) *The evolution of plants*. Oxford University Press.

Wolfe KH, Li W-H, Sharp PM. (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Science USA*. 84:9054-9058.

Wolfe KH, Morden CW, Palmer JD. (1992) Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proceedings of the National Academy of Science USA*. 89(22):10648-52.

Wolfenbarger LL, Phifer PR. (2000) The ecological risks and benefits of genetically engineered plants. *Science*. 290:2088-2093.

Wu F-H, Kan D-P, Lee S-B, Daniell H, Lee Y-W, Lin C-C, Lin N-S, Lin C-S. (2009) Complete nucleotide sequence of *Dendrocalamus latiflorus* and *Bambusa oldhamii* chloroplast genomes. *Tree Physiology*. 29(6):847-56.

Yamada T. (2001) Introduction of a self-compatible gene of *Lolium temulentum* L. to perennial ryegrass (*Lolium perenne* L.) for the purpose of the production of inbred lines of perennial ryegrass. *Euphytica*. 122(2):213-217.

Yamamoto MP, Kubo T, Mikami T. (2005) The 5¢-leader sequence of sugar beet mitochondrial *atp6* encodes a novel polypeptide that is characteristic of Owen cytoplasmic male sterility. *Molecular Genetics and Genomics*. 273:342-349.

Yukawa M, Tsudzuki T, Sugiura M. (2006) The chloroplast genome of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*: complete sequencing confirms that the *Nicotiana sylvestris* progenitor is the maternal genome donor of *Nicotiana tabacum*. *Molecular Genetics and Genomics*. 275:367–373.

Zeltz P, Hess WR, Neckermann K, Börner T, Kössel H. (1993) Editing of the chloroplast *rpoB* transcript is independent of chloroplast translation and shows different patterns in barley and maize. *The EMBO Journal*. 12(11):4291-6.

Zeng WH, Liao SC, Chang CC. (2007) Identification of RNA editing sites in chloroplast transcripts of *Phalaenopsis aphrodite* and comparative analysis with those of other seed plants. *Plant and Cell Physiology*. 48(2):362-8.

Zhang W. (2000) Phylogeny of the grass family (Poaceae) from *rpl16* intron sequence data. *Molecular Phylogenetics and Evolution*. 15(1):135-146.

Zimmer C. (2009) On the origin of eukaryotes. *Science*. 325:666-668.

9. Appendix

Table A1 Primer sequences for mRNA editing analysis. First sequence = forward primer, second sequence = reverse primer, T_a = annealing temperature in °C

gene	name	genome region	sequence	T_a
atpA	LP_atpA-a	33936 - 34546	GCTTCCGAAAAAAGAGTAAAGA	60
			ACATATTACATTTTGTCCTTTTTGA	
	LP_atpA-b	34474 - 35074	GGACAGACAGACTGGCAAAAC	60
			CGCGGATCCTACTCTGGAA	
	LP_atpA-c	34914 - 35514	TACCAATAGTTGAGACTCAAT	55
			AACAAGAAAGAGTAGACGTG	
atpB	LP_atpB-a	52011 - 52834	CAGTACATAAAGATTTAATT	55
			GTAGCTCTAGTCTATGGC	
	LP_atpB-b	52637 - 53127	ACCCACTGCAGAAGGCATTC	60
			GATCTGCACCCGCCTTTATC	
	LP_atpB-c	53035 - 53565	CACGTCGATAAGGAGC	50
			AAATTTCGACGAATTCT	
atpF	LP_atpF	32425 - 33903	AAGACTATAAGAGGAGAGCATA	53
			AAGTAGAAGCTCAAAGCCTA	
clpP	LP_clpP	67246 - 67946	TAACTTTACAAGAAATCTGGAC	48
			CTTAAATTTAATAATATTAATCTAAAT	
matK	LP_matK-a	1681 - 1882	CAACCCTTTTCTGTTTATTAGTTCTG	52
			CAACGATTAGGTTCGGCATT	
	LP_matK-b	1715 - 2503	TTAATTAAGAGGCTTCACCA	53
			CACTTTTCTAAGAAGATGGAA	
	LP_matK-c	2397 - 3249	CCTTGATATCGAACATAATGC	60
			TGGAAAAATTCGAAGGGTATT	
ndhA	LP_ndhA-a	110695 - 111247	ATATTCTTACTGTTATTGTATTATTTTCTTTATAGT	53
			GTACAGTTGATATAGTTGAAGCACAGTCTA	
	LP_ndhA-b	111150 - 112845	AGAAGAAATTAGAAAAACCAGAAA	50
			GGGAAGTTGATCGTTGAAA	
ndhB	LP_ndhB-a	85624 - 87150	GAAAGAAATAGACCTAGCAG	48
			CCTTTTCATCAATGGACT	
	LP_ndhB-b	87115 - 87911	TCCTTCGTAGACGTCAG	48
			TTGGATGCAGTTACTAATTC	
ndhD	LP_ndhD-a	106513 - 107326	CATTAATAGACCGGACTGGT	65
			GGAAGCGCATTATAGTACATG	
	LP_ndhD-b	107096 - 107495	ATGAAACCCATGTGAGATA	55
			TGGGCTTATATGGTTATGGT	

gene	name	genome region			sequence	T_a
	LP_ndhD-c	107313	-	108075	TATAATGCGCTTCCCCATGG	65
					AACGTATCTTGTCTTTACCACTTAATCAT	
ndhF	LP_ndhF-a	101485	-	102015	GACAAGATACTCGTAAATAGCTTAAGT	52
					AACCCTATTATCAAAGTGGTTAACTC	
	LP_ndhF-b	101992	-	102779	GTTAACCACTTTGATAATAGGGTTAA	50
					GCCTATTCTACAATGTCTCAATT	
	LP_ndhF-c	102724	-	103330	GAACCTATACCTAGAGCTAACATCA	65
					TGGGATTGGTTACTAGCTCC	
	LP_ndhF-d	103020	-	103440	CGCAACTGCACCAAGGAA	60
					AATTACTACTGTAGGAATCCTGGTTCTT	
	LP_ndhF-e	103311	-	103761	GGAGCTAGTAACCAATCCCA	55
					AGTAAAAATTGCAATTTCTTTTC	
ndhG	LP_ndhG-a	109402	-	109866	TTTGCAATTCTTATAATTTTTACTCTTT	60
					TGCCTTTTCGTTGGGATTAG	
	LP_ndhG-b	109579	-	110019	TGGTTTGATCTTGTCGTCCA	60
					AGTCAGTTCATGAAAAATTTTATACT	
ndhI	LP_ndhI	110045	-	110658	AAAGGGATTAGGATATTTTTATAAC	50
					ACATATCTTTTTCATAGATATAGAATGA	
ndhK	LP_ndhK-a	48608	-	49057	CTACACAAAAAAGCCCCATC	55
					ATGGCTCCCTCCTTAGTGAG	
	LP_ndhK-b	48964	-	49393	TGAACATTCCCCCTGTAATA	55
					ATGGCGAAAAGGAGCCTT	
petA	LP_petA-a	59259	-	59767	TTGTAGAAATTCAGGGATCC	53
					TCACTATATTTCTTACCGGGAA	
	LP_petA-b	59669	-	60279	GTTGAAAGAAAAGATAGGAAATC	53
					CCAACTTGATGGTAAGAAATC	
petB	LP_petB-a	70886	-	72031	TTCGGTATCTCTGGAATATG	55
					GTCAAAACAGCCAAAACC	
	LP_petB-b	71927	-	72318	GGTTCTAATGATGATCCTG	55
					AATTCTTTATGATATGCCTT	
psaA	LP_psaA-a	38724	-	39509	TTTCAAATCCTCCTAGCCACTATCC	53
					CTCCTGGTGCAACAACAAGTACC	
	LP_psaA-b	39430	-	40180	TAACAAAGCCACTTTACCGCCT	48
					TCTGACTTTTCGCGGAGGAC	
	LP_psaA-c	40121	-	41025	GCAATATCAGTCAGCCATAGACCAC	48
					GTATATCTTGTCCGGAAAGAGGAGG	
	LP_psaA-d	40657	-	41016	CTGAGCACTGGGTCCAATGT	60
					GTCCGGAAAGAGGAGGACTT	

gene	name	genome region			sequence	T_a
psaB	LP_psaB-a	36483	-	37252	CAGTACAAAAAAACAAAAAAAATAAC	53
					GTAGAAACCTATGGTTGCCAG	
	LP_psaB-b	37224	-	37934	CAACCATCCTGGCAACC	53
					CATCCACAAACACAAAGTTTG	
	LP_psaB-c	37833	-	38323	GAAGTTAGTTCGATACATGTGACCA	48
					ATGACGATCTTTATACTGGAGCTC	
	LP_psaB-d	38234	-	38742	TCCATTTGGGTTGTAGATGT	53
					TGGCTAGGAGGATTTGAAAG	
psaJ	LP_psaJ	64236	-	64432	GTATTTCAACAAGGAAGGAGGAT	53
					GGGCATGTCACAAAGAATTC	
psbC	LP_psbC-a	9772	-	10501	GGCAGCTCAGGATCAGCCT	48
					TGTCCACCAATTATATCTTCTAAATCATC	
	LP_psbC-b	10298	-	10701	TCTAGGTTCTTTTCTTTTAGTACTTAAGGCTC	55
					GTCCATAAAACTCACTCGGATAAGC	
	LP_psbC-c	10468	-	11243	GTGTGGATGATTTAGAAGATA	52
					AACATTAGAACAGGTATAAATAAG	
	LP_psbC-d	10971	-	11268	GCAAGAACGACGTTCAGCAG	55
					ACCGAGCCAGAACAGAAAAA	
psbD	LP_psbD-a	8753	-	9341	TGGTTGAAGTAATTGAATAGGAG	53
					GGAAGAAAAGGATGAATCG	
	LP_psbD-b	9290	-	9872	TTTGCGCCGAGTTTTGG	48
					GTCACGACCAGCTAAAACGA	
psbE	LP_psbE	61575	-	61860	GTTCGATCTATGGTCATTG	55
					GCTCAGCATGTCTGGA	
psbJ	LP_psbJ	61031	-	61196	CTGGTCCCTCCGATT	53
					GATAAAATGTGGAGGAAAGT	
psbL	LP_psbL	61035	-	61481	TCCCTCCGATTACTATAGAGATGAA	55
					CCAACGATAAACAAAATTCCAAC	
psbZ	LP_psbZ	11784	-	12029	TTTGATGGTAGTTTGGAGG	52
					TGTATCGGGCTACTAAATACT	
rpl20	LP_rpl20	65910	-	66347	GGGAATTCGGTTTTTATTAT	50
					ATTAGTTATTCATTAAGGTTAATTTG	
rpl2	LP_rpl2-a	80517	-	81700	AATGCAATTTCCATATTCCT	50
					GAGATACTATTGTTTCTGGTACAAA	
	LP_rpl2-b	80901	-	82081	TGCTGCTCTAGCTAATTGC	53
					CTATTCCACTTCTAGATAGAGAAAA	

gene	name	genome region			sequence	T_a
rpoA	LP_rpoA-a	73845	-	74317	ATTGCAATGGATTTAAAACGAGA	55
					GTGGGAATGGAAATGAAAAACAC	
	LP_rpoA-b	74313	-	74914	CCCACAAGAAAAAATACTATAATT	50
					ATTCAATGATCAAATAATAATACTAG	
rpoB	LP_rpoB-a	19232	-	19552	TGGAATTTTTCTTTGATAAACA	55
					CTTTTGCACATCGAAACC	
	LP_rpoB-b	19261	-	20081	TAAGATTAAGATGCTCCGG	60
					CAAGATTAAGCCTCCGA	
	LP_rpoB-c	19989	-	20792	CCTTATGTGAGGAATTACAAAA	60
					CTGGAACAACCTGCTCTTC	
	LP_rpoB-d	20642	-	21453	CCGTTTTATGAAATATCTGCTGAGA	60
					CGGAGTAAATGTGCTTCTAGATG	
	LP_rpoB-e	21380	-	22182	ACGGATACGACAAGCCAA	52
					GGCCCAGTCGAACGC	
	LP_rpoB-f	22020	-	22550	CAAAAAATCCATGGGTA	50
					TTGGTCAATCATAAAATAGAAA	
rpoC1	LP_rpoC1-a	22523	-	23291	TTCTTATTTCTATTTTATGATTGACCAA	48
					CGCCAATTGCATGCGTC	
	LP_rpoC1-b	23173	-	23916	TAGAGAATTCCTTGGTCGAATG	52
					TATGCAGGGTAGGTGCTCTATTT	
	LP_rpoC1-c	23860	-	24587	TTCAAGAAGTTATGCGAGG	55
					CGAATCTAACTTCCTGGTTT	
rpoC2	LP_rpoC2-a	24750	-	25010	AGCCGAAAAGGGGGTACTTA	56
					AACTCTGTTGTTCCGCGTCT	
	LP_rpoC2-b	24775	-	25612	GGCGGAACGGGCCA	52
					TGCTCGAAAGGCAGTTATGAATC	
	LP_rpoC2-c	25489	-	26309	TGTCCAAACACTAATTGGTCGTGT	48
					TTTCCCTATCTTTTTTACCATAAGTATTCAT	
	LP_rpoC2-d	26227	-	27124	TAGATGCAGATCCAGTATAGCGTCTT	52
					TGTGAGTTGTGTATCCACTCCAATAA	
	LP_rpoC2-e	26968	-	27779	TTTGATTGAGCATCGAGGAACA	65
					GTTGGCTAAGTGATTCAATCTTCG	
	LP_rpoC2-f	27715	-	28583	TAGGTTAGATCACACCAATCTAAATTCC	65
					CTATGGTTAGCTCAGCTCCAATCA	
	LP_rpoC2-g	28307	-	29187	GCAGCTAAACCTTATTTGGCCAC	48
					TCCTTTATGTATCCTTAACTCGTATAAATGG	
	LP_rpoC2-h	28998	-	29232	ATCGTTCCCCACAAGACAAG	60
					CCGTTTAGTGTTCTAAGTTCAAAAGTC	

gene	name	genome region	sequence	T_a
rps14	LP_rps14	36006 - 36361	AGAAATTCATACAAATACGAAAGGAGATT	60
			CGGCACATTATGGCAAAAAAA	
rps2	LP_rps2-a	29434 - 29785	AAAAATCAAAAGGAAATGTGGAAAA	52
			GACCAATTAGTTAACATACCACTGAACC	
	LP_rps2-b	29751 - 30169	AAAAAGTGGTTCAGTGGTAT	55
			CAATGATTTCTTATATAGAGAGAA	
	LP_rps2-c	29942 - 30199	GCCAGACATTGTGATCGTCCT	60
			GTTGCCCAAGAATTCACTATTCTTC	
rps8	LP_rps8	76167 - 76636	GAATTCGGGTACTATAGCCACAGC	65
			GAAAAATTTGGAGGAACCTAGAATTAGA	
ycf3	LP_ycf3	41623 - 43653	TTATGAAAAAGAAGGAGCGTGGTC	65
			GGTTGGGAATTATGCCTAGATCC	