DNA Research **16**, 165–176, (2009) doi:10.1093/dnares/dsp008

Complete Chloroplast Genome Sequence of a Major Allogamous Forage Species, Perennial Ryegrass (*Lolium perenne* L.)

Kerstin Diekmann[1,2,*], Trevor R. Hodkinson[2], Kenneth H. Wolfe[3], Rob van den Bekerom[1,4], Philip J. Dix[4], and Susanne Barth[1]

Teagasc Crops Research Centre, Oak Park, Carlow, Ireland[1]; School of Natural Sciences, Trinity College Dublin, Dublin 2, Ireland[2]; Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland[3] and Department of Biology, National University of Ireland Maynooth, Maynooth, Ireland[4]

(Received 10 February 2009; accepted 12 April 2009; published online 4 May 2009)

Abstract

***Lolium perenne* L. (perennial ryegrass) is globally one of the most important forage and grassland crops. We sequenced the chloroplast (cp) genome of *Lolium perenne* cultivar Cashel. The *L. perenne* cp genome is 135 282 bp with a typical quadripartite structure. It contains genes for 76 unique proteins, 30 tRNAs and four rRNAs. As in other grasses, the genes *accD*, *ycf1* and *ycf2* are absent. The genome is of average size within its subfamily Pooideae and of medium size within the Poaceae. Genome size differences are mainly due to length variations in non-coding regions. However, considerable length differences of 1–27 codons in comparison of *L. perenne* to other Poaceae and 1–68 codons among all Poaceae were also detected. Within the cp genome of this outcrossing cultivar, 10 insertion/deletion polymorphisms and 40 single nucleotide polymorphisms were detected. Two of the polymorphisms involve tiny inversions within hairpin structures. By comparing the genome sequence with RT–PCR products of transcripts for 33 genes, 31 mRNA editing sites were identified, five of them unique to *Lolium*. The cp genome sequence of *L. perenne* is available under Accession number AM777385 at the European Molecular Biology Laboratory, National Center for Biotechnology Information and DNA DataBank of Japan.**

Key words: chloroplast genome; *Lolium perenne*; Poaceae; chloroplast DNA variation; RNA editing

1. Introduction

Chloroplasts (cps), plant cell organelles derived from independent living cyanobacteria,[1–3] contain their own small genome averaging 150 kb in flowering plants. The cp genome molecules can be circular or linear, mono- or multimeric,[4] but the genome can be represented by a monomeric circular map containing two copies of an inverted repeat (IR) region (~23 kb) which separate a small single copy (SSC) region (~18 kb) from a large single copy (LSC) region (~84 kb). In most angiosperm species, the cp genome contains ~113 different genes[5] that primarily encode for proteins and RNAs for the photosynthetic system and that are generally highly conserved in terms of content and order among plant families.[6] Cp genomes are usually inherited maternally,[7] and this property is useful for several applications such as for defining cytoplasmic breeding pools in plant breeding, and tracking parentage in interspecific hybrids (e.g. *Arabidopsis suecica*[8]). Cp genetic engineering is also an ideal approach for minimizing the risk of spreading transgenes into wild plants via pollen.[9] In comparison with nuclear genetic engineering, much higher expression of the transgenic insertion can also be obtained because of

Edited by Satoshi Tabata

* To whom correspondence should be addressed. Tel. +353 59-9170-243. Fax. +353 59-9142-423. E-mail: kerstin.diekmann@teagasc.ie

the high copy number of cp genomes within a single plant cell. Cp genome sequences are also highly suitable for phylogenetic studies.[10]

To date (February 2009), entire cp genome sequences of 117 streptophytic species are publicly available (http://www.ncbi.nlm.nih.gov/genomes/ORGANELLES /plastids_tax.html). Only 18 of these genome sequences belong to the monocot group of angiosperms, and of these 13 are from the grass family Poaceae. Poaceae include the most important agricultural plant species from a socio-economic perspective as they contain the cereals and forage species.[11] *Lolium perenne* (perennial ryegrass) is globally one of the most important grassland species especially for the northern hemisphere (http://www.worldseed.org). In 2006–2007, more than one-third of world grass seed production was from *L. perenne*. Thus *L. perenne* has the highest economic impact as a forage and grassland crop. It is a cross-pollinating species and cultivar populations consist of a heterozygous nuclear genome background.

Several methods exist for obtaining complete cpDNA sequences. The *Arabidopsis thaliana* cp genome, for example, was sequenced using cpDNA clones found as 'contaminations' in genomic libraries.[12] The cp genome of *Nicotiana sylvestris*, the maternal genome donor of *Nicotiana tabacum*, was obtained by sequencing extracted high-purity cpDNA that was cloned into sequencing vectors.[13] A commonly used method involves amplifying the cp genome by rolling circle amplification and then cloning this product into sequencing vectors.[14] Recently, consensus cpDNA sequencing primers have become available for sequencing cp genomes using a primer walking strategy.[15] For sequencing the cp genome of *L. perenne*, we extracted high-purity cpDNA which we amplified with a whole genome amplification kit and used a shotgun sequencing approach. Thus each region of the genome was sequenced several fold from independent clones, which allowed us to detect SNPs and indels.

Few studies have examined variation of the cp genome within a population of a species. However, McGrath et al.[16] discovered more than 500 haplotypes within 1575 individual plants of *Lolium*, *Festulolium* and *Festuca* populations. We hope to add to this information by assessing cp genome variation within a *Lolium* cultivar by detecting SNPs and indels. This assessment should reveal highly variable regions in the *Lolium* cp genome, from which markers can be designed for assessing cytoplasmic breeding pools and to add to population genetic and phylogenetic studies.

In this study, we also analysed RNA editing sites in *L. perenne* cp transcripts. RNA editing is a repair mechanism that alters the genetic information of land plant organelles at the transcript level. It is a post-transcriptional modification (mostly C to U conversion) of the nucleotide sequence of pre-mRNAs by inserting, deleting or substituting nucleotides in order to yield functional RNA species.[17,18] Editing in cps was first discovered by Hoch et al.[19] for the cp *rpl2* gene in maize, where it creates a start codon and hence restores the functionality of the *rpl2* gene. Knowledge about RNA editing sites is essential for describing the functional capability of cp genes, characterizing different species and obtaining a better understanding of how these sites have evolved.

2. Materials and methods

2.1. Sequencing, assembling and annotating the cp genome

cpDNA was isolated from the *L. perenne* cultivar Cashel following a protocol from Diekmann et al.[20] Approximately 400 g of 3-week-old leaf material derived from ca. 200 g of a heterozygous Cashel seed population was used. Sequencing of the cpDNA was sourced to a commercial company (GATC Biotech/Germany). A shotgun sequencing approach was used resulting in 2179 trace files. A pre-assembly was carried out with the program PHRAP (http://www.phrap.org/index.html). The final assembly was based on the contiguous sequences obtained from PHRAP and done in comparison with the cpDNA sequence of *Agrostis stolonifera* (bentgrass).[21] For three genome regions with low coverage, primers were designed to re-sequence these regions (*trnL-trnF*: forward primer (FP): AGTTGTGAGGGTTC AAGTCC and reverse primer (RP): GAACTGGTGACAC GAGGATT; *atpB*: FP: GTTCGTTGCCAACAATCCTA and RP: AGGTAGCTCTAGTCTATGGC; *atpB-rbcL*: FP: TGTGG AAGATCTGTGCCTAC and RP: GCTGAGGAGTTACTCG GAAT).

The annotation of the cp genome was based on 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 following Sugita and Sugiura.[22] The circular cp genome map was drawn using the GenomeVx program.[23] Differences between the available cp genomes were analysed based on gene, intergenic spacer (IGS) and intron lengths which were extracted from the published cp genome sequences (*A. 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).

2.2. SNP and indel analysis

Because the cpDNA had been extracted from a population of plants belonging to the cultivar Cashel, several SNPs and indels could be detected. A thorough SNP and indel analysis was carried out by manually checking the alignment of the read and trace files from which the genome assembly was undertaken using the programme Lasergene (DNAstar, Inc., Madison, Wisconsin). Only SNPs and indels supported by trace files with low background and clear, distinguishable peaks were recorded. Indels were only taken into account if they were supported by at least two trace files and not located in coding regions where they would cause a frame shift. This way the possibility of cloning and sequencing artefacts was considered.

2.3. RNA editing analysis

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 analysed for RNA editing sites. Of these, 22 were chosen for study because they had been previously reported to be edited in other monocot plants,[24–28] and 11 were included because of observed differences from existing expressed sequence tags (EST) in Poaceae,[21] but no information was previously available for *Lolium*. Primers (Supplementary Table S1) 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 $>$ 700 bp, several primer pairs were designed to cover the complete gene region. Primers were designed in the untranslated regions (UTR) to ensure complete coverage of genes. Since the length of the UTR of genes was not known, the primers were designed in the 30 bp region before and after each gene.

cDNA was used as template for the RT–PCRs. 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 µl bromochloropropane, 200 µl of ice cold chloroform was used; the steps including the addition of ice cold chloroform, followed by incubation at room temperature and centrifugation at 12 000*g* were repeated once; in addition to the 500 µl isopropanol, 0.5 µl 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 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.

For each gene region, two independent RT–PCR reactions were set up using the following components per 30 µl PCR reaction: 3 µl cDNA, 3 µl 10 x Thermo Buffer (New England Biolabs, Inc., Ipswich, MA, USA), 0.6 µl FP, 0.6 µl RP, 0.6 µl dNTPs (metabion international AG, Martinsried, Germany) (10 mM), 21.9 µl ddH_2O, 0.3 µl Taq-Polymerase (New England Biolabs, Inc.). The PCR programme settings were 95°C 5 min, (95°C 1 min, 55°C 1 min, 72°C 1 min) 35 cycles, 72°C 10 min. The annealing temperature was adjusted according to the optimal primer requirements. The resulting RT–PCR products were sequenced twice using both forward and reverse primers. The analysis of the editing sites was carried out in MEGA 3.1[29] by aligning the cDNA sequence results with the corresponding DNA sequences and checking visually for SNPs.

3. Results and discussion

Using the shotgun sequencing approach an average eightfold genome coverage was achieved. The cp genome of *L. perenne* has a total length of 135 282 bp with a quadripartite structure typical of angiosperms. The LSC region consists of 79 972 bp, the SSC of 12 428 bp and the IRs of 21 441 bp each (Fig. 1). The genome has a GC content of 38% and codes for 128 genes of which 18 are duplicated in the IR region. The genome contains 264 simple sequence repeats (SSRs) with mononucleotide repeats of 7–16 bp in length. The cp genome sequence of *L. perenne* is deposited at the European Molecular Biology Laboratory under Accession number AM777385.

3.1. Comparison to other species

The average size of publicly available Poaceae cp 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 *L. perenne* is of average size within Pooideae and of medium size within Poaceae (Fig. 2).

The gene content and intron content of *L. perenne* cpDNA are the same as that of other

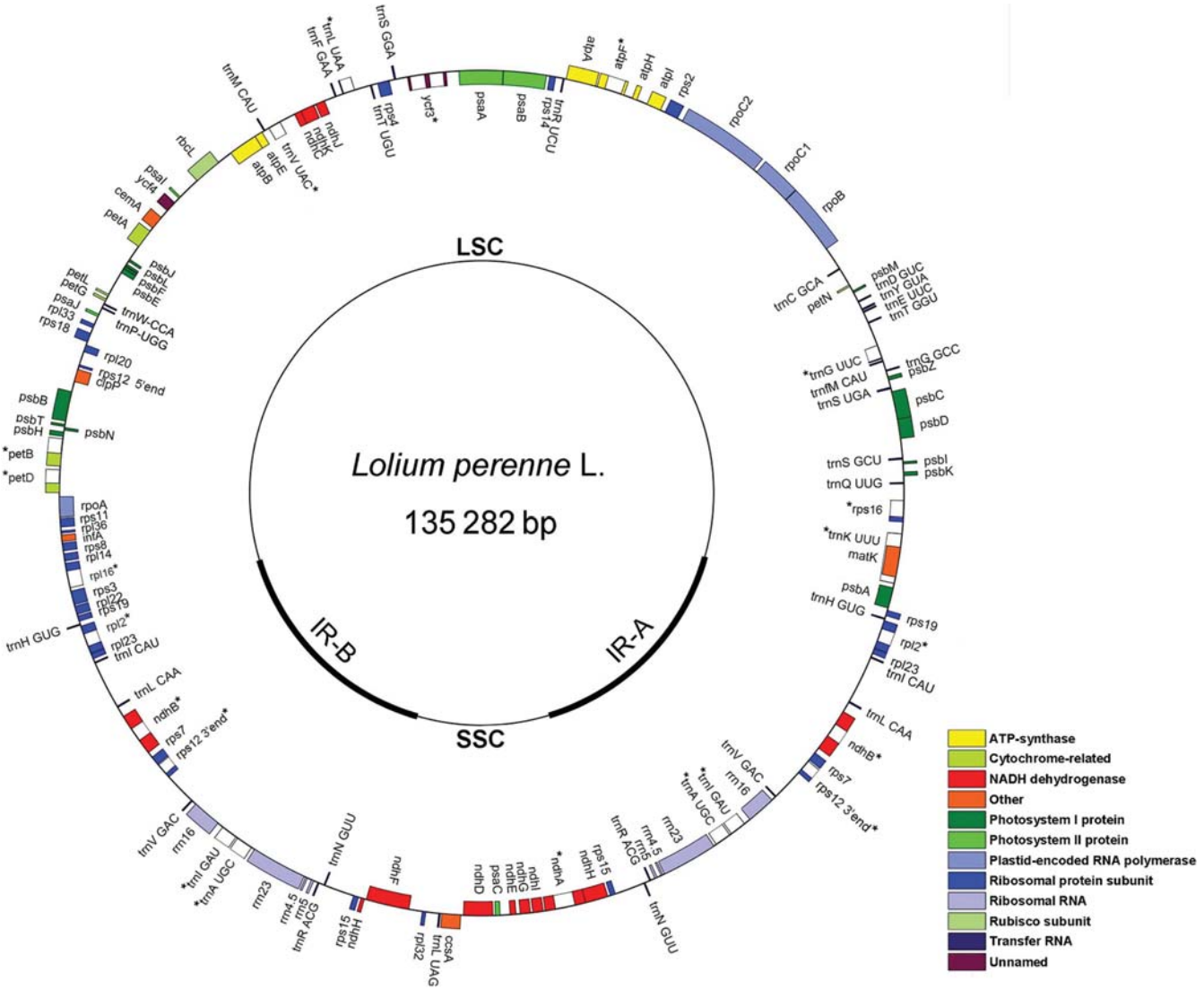

Figure 1. Circular structure of the chloroplast genome of *Lolium perenne*. Genes written on the outside are transcribed clockwise, genes on the inside counter-clockwise, annotated genes are colour coded according to their function, genes containing introns are highlighted with an asterisk; LSC, large single copy region; SSC, small single copy region; IR, inverted repeat.

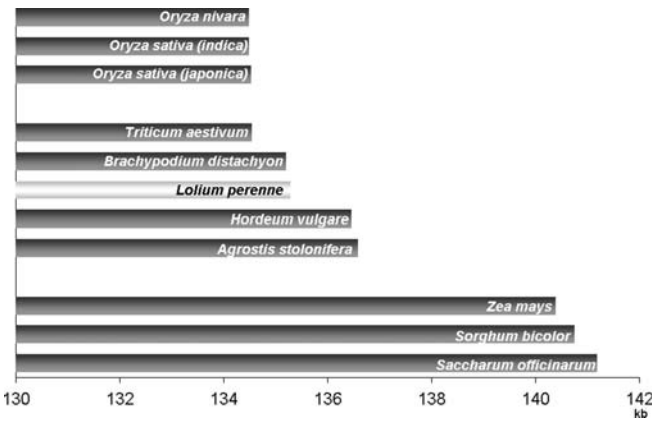

Figure 2. Chloroplast genome sizes of 11 different Poaceae species grouped by taxonomic sub-families.

grasses,[21,24,27,30,31,32] 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. When compared with the standard set of genes in angiosperm cp genomes, the genes *accD*, *ycf1* and *ycf2* are absent. After our analysis was completed, the cp genome sequence of the very closely related species *Festuca arundinacea* became available in GenBank (Cahoon et al., unpublished data; accession number FJ466687). Rather surprisingly, 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*. All four of these

Table 1. Length variation of > 100 bp for intergenic spacer and intron regions in Poaceae chloroplast genomes

Genome region	Ehrhartoideae	Pooideae					Panicoideae			Variation (bp)[a]
	Oryza nivara	*Triticum aestivum*	*Brachipodium distachyon*	*Lolium perenne*	*Hordeum vulgare*	*Agrostis stolonifera*	*Zea mays*	*Sorghum bicolor*	*Saccharum officinarum*	
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
[b]*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
[b]*petB–petB*	**814**	749	809	747	**697**	759	699	702	758	117
[b]*rpl16–rpl16*	1056	1044	1050	**868**	1064	1045	1043	1072	**1080**	212
Inverted repeat										
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
[b]*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
Small single copy region										
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

Bold numbers show the shortest length for that intergenic spacer/intron. Bold and underlined numbers show the largest length for that intergenic spacer/intron.

[a]Difference between smallest and largest values.

[b]Highlights introns.

Table 2. Variation in length of different chloroplast genes

Gene name	Length in		Ehrhartoideae	Pooideae					Panicoideae			Codon variation
	bp	codons	*Oryza nivara*	*Triticum aestivum*	*Brachipodium distachyon*	*Lolium perenne*	*Hordeum vulgare*	*Agrostis stolonifera*	*Zea mays*	*Sorghum bicolor*	*Saccharum officinarum*	
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
Inverted repeat												
rps12	363	121	4	1	—	4	4	4	4	4	4	4
rps15	237	79	12	12	12	12	12	12	—	—	—	12
Small single copy region												
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

If not otherwise stated numbers shown refer to amount of additional codons; —, no variation to the smallest length observed.

genes are intact and apparently functional in *L. perenne*.

Differences in the cp genome size of *L. perenne* compared with other Poaceae species are mainly due to length variations of IGS regions and introns (Table 1) and this finding was consistent with previous observations.[27,32] 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 1). The highest variation in size (given in brackets) was found in the *trnI-CAU–trnL-CAA* IGS (2135 bp), the *trnG-UCC–trnT-GGU* IGS (1231 bp) and the *rbcL–psaI* IGS (1221 bp). The *trnI-CAU–trnL-CAA* IGS and *rbcL–psaI* IGS are sites that contain pseudogenes for *ycf2* and *accD*, respectively, in Poaceae.[27] Both these pseudogenes and a *ycf1* pseudogene were detected in *L. perenne*. The *trnG-UCC - trnT-GGU* IGS is part of a 'divergence hotspot' described by Maier et al.[27] whose variability is caused by a large number of deletion/insertion events.

A comparison between *L. perenne* and the other Poaceae species showed differences in gene length for 26 genes (Table 2). 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. *L. 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. The *ndhK* and *rps16* genes are 21 and 27 codons, respectively, longer in *L. perenne* than in *O. nivara*.

The length variations observed in *rps18* and *rpoC2* are noteworthy. 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.,[33] based on the occurrence of different numbers of the heptapeptide repeat SKQPFRK near the N terminus of the protein. Our 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.[33] (Supplementary Fig. S1). This motif is present six times 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*. 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 and we expect the *L. perenne rps18* gene to be functional.

The largest length variation in Poaceae genes was found in *rpoC2*, of up to 68 codons difference between *L. perenne* and *S. officinarum*, and is due to several insertion/deletion events (data not shown). Comparisons of the *rpoC2* gene from dicots and monocots revealed that Poaceae have a unique insertion of ~400 bp in the middle of this gene.[34–36] Cummings et al.[36] demonstrated that this region is highly variable compared with 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.[37] discovered a 165 bp deletion in this insertion region that suggests a possible relation between this deletion in *rpoC2* and the CMS-system.[38] So far this deletion was only observed in *Sorghum* but sequence comparisons (data not shown) 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.

3.2. Indel/SNP analysis

A total of 10 indels (Table 3) and 40 SNPs (Table 4) were found to be polymorphic among our sequencing reads. All indels are 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. Knowledge gained about the sequence variability of these regions can be used to design primers around those microsatellites for population genetic and phylogenetic studies and can be also used 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*.

Nineteen SNPs were found within IGS regions and introns and 21 within coding regions (Table 4). Most of the SNPs are due to transition mutations (20 A↔G and 8 C↔T), with 12 transversions. Closer analysis of the SNPs found at position 100 655 and 100 656 (*trnN-rps15* IGS) revealed that these SNPs 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 (Fig. 3). The small inversion of TG within the *trnN-rps15* region in the IR is found in 13 of the 29 trace files covering the region. We also noticed another small inversion that was supported by only one sequence read and caused SNPs at position 18, 20, 21 and 23 (*rps19-psbA* IGS). This inversion spans six nucleotides (TTCTAG) that are flanked by an IR of 25 bp length (Fig. 3).

Small inversions like the ones revealed by our study have been found between species,[39,40] between genera[39,41] and also within populations of one other species, the conifer *Abies*.[42] The two inversions found in the current study lead, together with the

Table 3. Indels observed in the cp genome of *Lolium perenne*

Position	Nucleotide		Region	Trace files	
	Major	Minor		Absolute	%
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

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.

Table 4. Single nucleotide polymorphisms in the chloroplast genome of *Lolium perenne*

Position	Nucleotide		IUPAC	Amino acid change	Region	Trace files	
	Major	Minor				Absolute	%
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
10390	A	G	R	—	*psaC*	6	42.86
109007	G	A	R	A→V	*ndhE*	1	10.00
18[a]	C	T	Y		*rps19-psbA*	1	16.67
20[a]	A	C	M		*rps19-psbA*	1	16.67
21[a]	G	T	K		*rps19-psbA*	1	16.67
23[a]	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
100655[a]	C	T	Y		*trnN-rps15*	13	43.33
100656[a]	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

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.
[a]Inversions.

level of observed SNPs, to the conclusion that the cp genome of *L. perenne* cv Cashel consists of at least two haplotypes but potentially scores more. McGrath et al.[16] detected five haplotypes in 16 individuals of Cashel using a set of ten primers[43] amplifying eight different regions in the cp genome and sizing

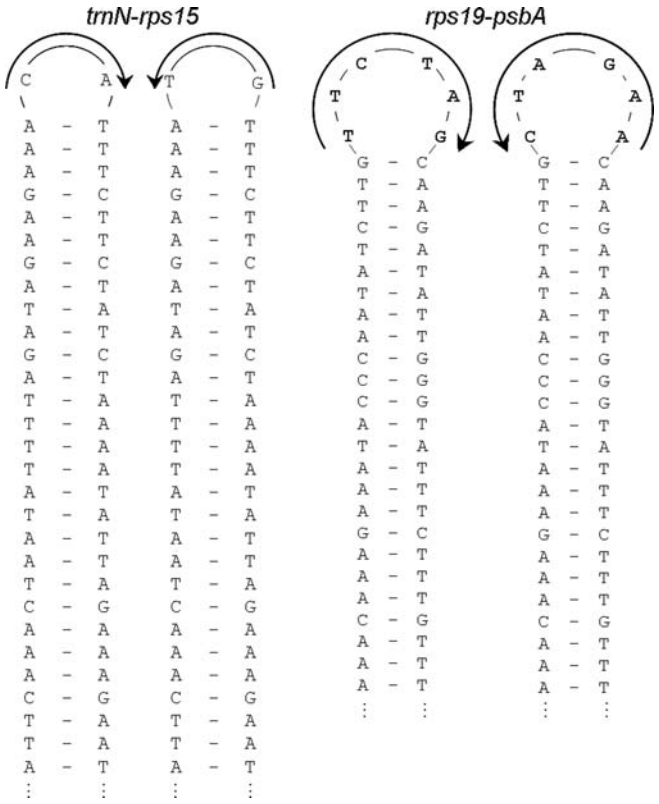

Figure 3. Two hairpin loops found in the chloroplast genome of *Lolium perenne*.

the PCR products. Eight maternal lines were included in breeding Cashel (Vincent Connolly, personal communication). Thus further analyses based on DNA sequences could reveal up to eight haplotypes, differing by SNPs and indels that include the ones found in this study.

Although cp genomes are known to be highly conserved, similar observations of intraspecific cp DNA variation have been recorded in other species.[21,44,45] However, this is the only study we know of that has quantified SNP variation of the whole cp genome within a cultivar. Most studies of cp 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. Tsumura et al.[42] studied natural populations of *Abies* and also detected many minor variations like indels and inversions within species. Although some of the apparent SNPs 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 we used for sequencing is based on a population of several maternal lines and is thus heterogeneous and heterozygous. However, to discover this extent of SNP variation within a single cultivar of *Lolium* was surprising.

3.3. RNA editing sites

In total, 31 RNA editing sites were detected in 18 genes (Table 5). All editing sites are C to U changes. Most frequently, editing results in changes of the amino acid from serine or proline to leucine. Four editing sites (*ndhA*, site 4; *ndhG*, 5′ UTR; *rpoB*, site 4; *rps14*), which were previously observed in other Poaceae species, were not edited in *L. perenne*. For three of them, the conserved nucleotide U exists already at the DNA level. Site 4 in *rpoB* is not edited, although C is encoded in the DNA.

Analysis of editing in the *ndhA* gene was not completed because several primers failed to amplify and the obtained sequencing products did not have the full gene length. Thus site 4, which was observed in *O. sativa*, *S. officinarum* and *Z. mays*, 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*.

The editing analysis revealed five new editing sites that are so far unique to *L. perenne*. Four of these sites are in three genes (*ndhK, psbJ, psbL*) in which editing has never been reported before in Poaceae species. Two of the five new editing sites are synonymous but three result in changes of the amino acid to leucine.

Partial editing was observed at eight editing sites (six genes). In most of these editing sites, the amount of incompletely edited transcripts is small. However, approximately one-half and one-third of the *matK* and *psbL* transcripts, respectively, are not edited.

This study of RNA editing sites in the cp genome of *L. perenne* demonstrates that predicting editing sites based solely on published EST sequences is not sufficient. Timme et al.[45] also showed that editing sites can be easily overlooked, or SNPs can be falsely interpreted as editing sites, using that approach. For example only six of the genes analysed via EST comparisons by Saski et al.[21] have had editing sites experimentally confirmed in other species. The SNPs found in EST sequences by Saski et al.[21] were in general not based on C–U changes and thus are highly unlikely to be editing sites. Most of the SNPs found by comparing ESTs to cpDNA sequences will be based on the use of different varieties, or on poor quality sequencing data. Our approach of analysing SNPs and editing sites in the same variety of *L. perenne* ensured that newly detected sites with either complete or partial editing were evaluated

Table 5. RNA editing sites found in the chloroplast genome of *Lolium perenne* in comparison with the editing sites found in other monocots

Gene	Site	Codon position	Editing sites	Edited codon	Amino acid change	*Lolium perenne*	*Hordeum vulgare*[26]	*Oryza sativa*[25]	*Saccharum officinarum*[24]	*Zea mays*[27]
atpA	1	383	35112	tCa	S→L	+		+	+	+
matK	1	420	1993	Cat	H→Y	+[a]	+[46]	+[50]		
ndhA	1	17	111 250	tCa	S→L	+	+[47]	(—)	+	+
	2	158	112 355	tCa	S→L	+	+[47]	+	+	+
	3	188	112 777	tCa	S→L	+	+[47]	+[50]	+	+
	4	357		tCc	S→F	(—)	+[47]	+	+	+
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)	107 165	tCa	S→L	+[a]	+[48]	+	+	+
ndhF	1	21	103 675	tCa	S→L	+		+	+[a]	+
ndhG	1	116	109 624	cCa	P→L	+		+[50]		(—)
5′UTR		−10				(—)	(—)[49]	+	+	+
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]	+[28]	+[a]	+	+
	2	182	19815	tCa	S→L	+[a]	+[28]	+[a]	+	+
	3	187	19830	tCg	S→L	+[a]	+[28]	+[a]uCa	+	+
	4	206		cCg	P→L	—	—[28]	—	—	+
rpoC2	1	925		tCg	S→L	(—)		(—)	+	+
	2	*1320*	*28731*	*tCa*	*S→L*	+				—
rps8	1	61	76422	tCa	S→L	+		+	+	+
rps14	1	27		tCa	S→L	(—)		+	+[a]	+
ycf3	1	15	43599	tCc	S→F	+		(—)	—	+
	2	62	42 700	aCg	T→M	+		+[a]	+	+[51a]

—, editing although C encoded in DNA; (—), no editing, U encoded in DNA; blank space, editing not yet determined/no information available; italic text: unique for *Lolium perenne*.
[a]Partially edited.

correctly as editing sites and not accidentally mistaken as SNPs or vice versa.

Acknowledgements: We thank Dr Jiri Ködding and Dr Gavin Conant for all their support while sequencing, assembling and annotating the chloroplast genome. K.H.W. is supported by Science Foundation Ireland. K.D. and R.vdB. were financed under the Teagasc Walsh Fellowship Scheme.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

The project funding was obtained from the Teagasc 'Vision' programme.

References

1. Gray, M. W. 2004, The evolutionary origins of plant organelles, In: Daniell and Chase, H. (eds.), *Molecular Biology and Biotechnology of Plant Organelles*, Springer: Dordrecht, pp. 87–108.

2. Keeling, P. J. 2004, Diversity and evolutionary history of plastids and their hosts, *Am. J. Bot.*, **91**, 1481–1493.
3. Margulis, L. 1970, Origin of eukaryotic cells, Yale University Press: New Haven.
4. Lilly, J. W., Havey, M. J., Jackson, S. A. and Jiang, J. 2001, Cytogenomic analyses reveal the structural plasticity of the chloroplast genome in higher plants, *Plant Cell*, **13**, 245–254.
5. Sugiura, M. 1992, The chloroplast genome, *Plant Mol. Biol.*, **19**, 149–168.
6. Palmer, J. D. 1987, Chloroplast DNA evolution and biosystematic uses of chloroplast DNA variation, *Am. Nat.*, **130**, S6.
7. Corriveau, J. L. and Coleman, A. W. 1988, Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species, *Am. J. Bot.*, **75**, 1443–1458.
8. Säll, T., Jakobsson, M., Lind-Halldén, C. and Halldén, C. 2003, Chloroplast DNA indicates a single origin of the allotetraploid *Arabidopsis suecica*, *J. Evol. Biol.*, **16**, 1019–1029.
9. Daniell, H., Datta, R., Varma, S., Gray, S. and Lee, S. B. 1998, Containment of herbicide resistance through genetic engineering of the chloroplast genome, *Nat. Biotechnol.*, **16**, 345–348.
10. Wu, F. -H., Kan, D. -P., Lee, S. -B., et al. 2009, Complete nucleotide sequence of *Dendrocalamus latiflorus* and *Bambusa oldhamii* chloroplast genomes, *Tree Physiol.*, doi:10.1093/treephys/tpp015.
11. Hodkinson, T. R., Waldren, S., Parnell, J. A. N., Kelleher, C. T., Salamin, K. and Salamin, N. 2007, DNA banking for plant breeding, biotechnology and biodiversity evaluation, *J. Plant Res.*, **120**, 17–29.
12. Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E. and Tabata, S. 1999, Complete structure of the chloroplast genome of *Arabidopsis thaliana*, *DNA Res.*, **6**, 283–90.
13. Yukawa, M., Tsudzuki, T. and 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*, *Mol. Genet. Genomics*, **275**, 367–373.
14. Jansen, R. K., Raubeson, L. A., Boore, J. L., et al. 2005, Methods for obtaining and analyzing whole chloroplast genome sequences, *Methods Enzymol.*, **395**, 348–384.
15. Chung, S. M., Gordon, V. S. and Staub, J. E. 2007, Sequencing cucumber (*Cucumis sativus* L.) chloroplast genomes identifies differences between chilling—tolerant and—susceptible cucumber lines, *Genome*, **50**, 215–225.
16. McGrath, S., Hodkinson, T. R. and Barth, S. 2007, Extremely high cytoplasmic diversity in natural and breeding populations of *Lolium* (Poaceae), *Heredity*, **99**, 531–544.
17. Bock, R. 2000, Sense from nonsense: how the genetic information of chloroplasts is altered by RNA editing, *Biochimie*, **82**, 549–557.
18. Tsudzuki, T., Wakasugi, T. and Sugiura, M. 2001, Comparative analysis of RNA editing sites in higher plant chloroplasts, *J. Mol. Evol.*, **53**, 327–332.
19. Hoch, B., Maier, R. M., Appel, K., Igloi, G. L. and Kössel, H. 1991, Editing of a chloroplast mRNA by creation of an initiation codon, *Nature*, **353**, 178–180.
20. Diekmann, K., Hodkinson, T. R., Fricke, E. and 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**, e2813.
21. Saski, C., Lee, S. B., Fjellheim, S., et al. 2007, Complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes, *Theor. Appl. Genet.*, **115**, 591.
22. Sugita, M. and Sugiura, M. 1996, Regulation of gene expression in chloroplasts of higher plants, *Plant Mol. Biol*, **32**, 315–326.
23. Conant, G. C. and Wolfe, K. H. 2008, GenomeVx: simple web-based creation of editable circular chromosome maps, *Bioinformatics*, **24**, 861–862.
24. Calsa Júnior, T., Carraro, D. M., Benatti, M. R., Barbosa, A. C., Kitajima, J. P. and Carrer, H. 2004, Structural features and transcript-editing analysis of sugarcane (*Saccharum officinarum* L.) chloroplast genome, *Curr. Genet*, **46**, 366–373.
25. Corneille, S., Lutz, K. and Maliga, P. 2000, Conservation of RNA editing between rice and maize plastids: are most editing events dispensable, *Mol. Gen. Genet.*, **264**, 419–424.
26. Freyer, R., Hoch, B., Neckermann, K., Maier, R. M. and Kössel, H. 1993, RNA editing in maize chloroplasts is a processing step independent of splicing and cleavage to monocistronic mRNAs, *Plant J.*, **4**, 621–629.
27. Maier, R. M., Neckermann, K., Igloi, G. L. and Kössel, H. 1995, Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing, *J. Mol. Biol.*, **251**, 614–628.
28. Zeltz, P., Hess, W. R., Neckermann, K., Börner, T. and Kössel, H. 1993, Editing of the chloroplast *rpoB* transcript is independent of chloroplast translation and shows different patterns in barley and maize, *EMBO J.*, **12**, 4291–4296.
29. Kumar, S., Tamura, K. and Nei, M. 2004, MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment, *Brief. Bioinform.*, **5**, 150–163.
30. Ogihara, Y., Isono, K., Kojima, T., et al. 2002, Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA, *Mol. Genet. Genomics*, **266**, 740–746.
31. Bortiri, E., Coleman-Derr, D., Lazo, G. R., Anderson, O. D. and Gu, Y. Q. 2008, The complete chloroplast genome sequence of *Brachypodium distachyon*: sequence comparison and phylogenetic analysis of eight grass plastomes, *BMC Res. Notes*, **31**, 61.
32. Palmer, J. D. 1991, Plastid chromosomes: structure and evolution, In: Vasil, I. K. and 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.

33. Weglöhner, W. and Subramanian, A. R. 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 Lett.*, **279**, 193–197.
34. Igloi, G. L., Meinke, A., Döry, I. and 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, *Mol. Gen. Genet.*, **221**, 379–394.
35. Shimada, H., Fukuta, M., Ishikawa, M. and Sugiura, M. 1990, Rice chloroplast RNA polymerase genes: the absence of an intron in *rpoC1* and the presence of an extra sequence in *rpoC*, *Mol. Gen. Genet.*, **221**, 395–402.
36. Cummings, M. P., King, L. M. and Kellogg, E. A. 1994, Slipped-strand mispairing in a plastid gene: *rpoC2* in grasses (Poaceae), *Mol. Biol. Evol.*, **11**, 1–8.
37. Chen, Z., Muthukrishnan, S., Liang, G. H., Schertz, K. F. and Hart, G. E. 1993, A chloroplast DNA deletion located in RNA polymerase gene *rpoC2* in CMS lines of sorghum, *Mol. Gen. Genet.*, **236**, 251–259.
38. Chen, Z., Schertz, K. F., Mullet, J. E., DuBell, A. and Hart, G. E. 1995, Characterization and expression of *rpoC2* in CMS and fertile lines of sorghum, *Plant Mol. Biol.*, **28**, 799–809.
39. Kim, K. J. and Lee, H. L. 2004, Complete chloroplast genome sequences from *Korean Ginseng* (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants, *DNA Res.*, **11**, 247–261.
40. Kim, K. J. and Lee, H. L. 2005, Widespread occurrence of small inversions in the chloroplast genomes of land plants, *Mol. Cells*, **19**, 104–113.
41. Kelchner, S. A. and Wendel, J. F. 1996, Hairpins create minute inversions in non- coding regions of chloroplast DNA, *Curr. Genet*, **30**, 259–262.
42. Tsumura, Y., Suyama, Y. and Yoshimura, K. 2000, Chloroplast DNA inversion polymorphism in populations of *Abies* and *Tsuga*, *Mol. Biol. Evol.*, **17**, 1302–1312.
43. McGrath, S., Hodkinson, T. R., Salamin, N. and 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, *Mol. Ecol. Notes*, **6**, 449–452(4).
44. Daniell, H., Lee, S. B., Grevich, J., et al. 2006, Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes, *Theor. Appl. Genet.*, **112**, 1503–1518.
45. Timme, R. E., Kuehl, J. V., Boore, J. L. and Jansen, R. K. 2007, A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats, *Am. J. Bot.*, **94**, 302–312.
46. Vogel, J., Hübschmann, T., Börner, T. and Hess, W. R. 1997, Splicing and intron- internal RNA editing of *trnK–matK* transcripts in barley plastids: support for *matK* as an essential splice factor, *J. Mol. Biol.*, **270**, 179–187.
47. Lopez, C., Freyer, R., Guera, A., et al. 1997, Sequence of *ndhA* gene of barley (*Hordeum vulgare* L.) plastid (accession nos. Y13729 & Y13730). Transcript editing in Graminean organs, *Plant Physiol.*, **115**, 313.
48. del Campo, E. M., Sabater, B. and Martin, M. 1997, Plastid *ndhD* gene of barley (*Hordeum vulgare* L.), sequence and transcript editing (Accession no. Y12258), *Plant Physiol.*, **114**, 748.
49. Drescher, A., Hupfer, H., Nickel, C., et al. 2002, C-to-U conversion in the intercistronic *ndhI/ ndhG* RNA of plastids from monocot plants: conventional editing in an unconventional small reading frame, *Mol. Genet. Genomics*, **267**, 262–269.
50. Inada, M., Sasaki, T., Yukawa, M., Tsudzuki, T. and 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 Physiol.*, **45**, 1615–1622.
51. Ruf, S. and Kössel, H. 1997, Tissue-specific and differential editing of the two *ycf3* editing sites in maize plastids, *Curr. Genet.*, **32**, 19–23.