Plant Mol Biol (2009) 71:627–639
DOI 10.1007/s11103-009-9545-3

Unparalleled GC content in the plastid DNA of *Selaginella*

David Roy Smith

Received: 3 July 2009 / Accepted: 21 August 2009 / Published online: 23 September 2009

Abstract One of the more conspicuous features of plastid DNA (ptDNA) is its low guanine and cytosine (GC) content. As of February 2009, all completely-sequenced plastid genomes have a GC content below 43% except for the ptDNA of the lycophyte *Selaginella uncinata*, which is 55% GC. The forces driving the *S. uncinata* ptDNA towards G and C are undetermined, and it is unknown if other *Selaginella* species have GC-biased plastid genomes. This study presents the complete ptDNA sequence of *Selaginella moellendorffii* and compares it with the previously reported *S. uncinata* plastid genome. Partial ptDNA sequences from 103 different *Selaginella* species are also described as well as a significant proportion of the *S. moellendorffii* mitochondrial genome. Moreover, *S. moellendorffii* express sequence tags are data-mined to estimate levels of plastid and mitochondrial RNA editing. Overall, these data are used to show that: (1) there is a genus-wide GC bias in *Selaginella* ptDNA, which is most pronounced in South American articulate species; (2) within the Lycopsida class (and among plants in general), GC-biased ptDNA is restricted to the *Selaginella* genus; (3) the cause of this GC bias is arguably a combination of reduced AT-mutation pressure relative to other plastid genomes and a large number of C-to-U RNA editing sites; and (4) the mitochondrial DNA (mtDNA) of *S. moellendorffii* is also GC biased (even more so than the ptDNA) and is arguably the most GC-rich organelle genome observed to date—the high GC content of the mtDNA also appears to be influenced by RNA editing. Ultimately, these findings provide convincing support for the earlier proposed theory that the GC content of land-plant organelle DNA is positively correlated and directly connected to levels of organelle RNA editing.

Keywords Chloroplast · Lycophyte · Nucleotide composition · GC content · RNA editing

Accession numbers The GenBank accession numbers of the *S. moellendorffii* organelle genome sequences described in this study are FJ755183 (ptDNA) and GQ246802-GQ246808 (mtDNA).

Electronic supplementary material The online version of this article (doi:10.1007/s11103-009-9545-3) contains supplementary material, which is available to authorized users.

D. R. Smith (✉)
Department of Biology, Dalhousie University, Halifax, NS, Canada
e-mail: smithdr@dal.ca

Introduction

A prominent feature of plastid DNA (ptDNA) is its low guanine and cytosine (GC) content. Indeed, all of the 150 completely-sequenced plastid genomes available at the National Center for Biotechnology Information (NCBI) as of February 2009 have a GC content between 19.5 and 42.1% (average = 36.2%; SD = 4.6%), with the exception of the *Selaginella uncinata* ptDNA, which is 54.8% GC—a complete compilation is shown in Table S1 (see Supplementary Materials). The evolutionary forces shaping ptDNA nucleotide landscape are unknown; however, several hypotheses have been proposed. For instance, some argue that a neutral process such as AT-mutation pressure or AT-biased gene conversion caused the low GC content of ptDNA (Howe et al. 2003; Kusumi and Tachida 2005; Khakhlova and Bock 2006). Others invoke selection for translational efficiency to explain the lack of G and C observed in plastid genomes (Morton 1993, 1998). There is also the possibility that plastids originate from an AT-rich bacteria, but it is generally thought that ptDNA has become

GC poor since endosymbiosis (Howe et al. 2003). Interestingly, convergent evolution towards a reduced GC content is seen in mitochondrial DNA (mtDNA), nucleomorph DNA, and in the genomes of symbionts, parasites, and pathogenic bacteria (Dybvig and Voelker 1996; Ogata et al. 2001; Lane et al. 2007; Smith and Lee 2008a); if the forces biasing these genomes against G and C are similar to those acting on ptDNA, then understanding nucleotide composition from a plastidial framework could have wide-reaching implications.

The plastid genome of the lycophyte *S. uncinata* is exceptional in that it has a GC content above 50% (Tsuji et al. 2007). In addition to being GC biased, the *S. uncinata* ptDNA has several other distinguishing characteristics: (1) it encodes only 12 distinct tRNAs, which is currently one of the most reduced tRNA-coding repertoires of any completely-sequenced plastid genome (land-plant plastid genomes typically contain more than 30 tRNA-coding genes); (2) it contains a unique ptDNA gene order, unlike *Huperzia lucidula* (the only other lycophyte with a completely sequenced ptDNA), which has a ptDNA gene arrangement similar to bryophytes (Wolf et al. 2005); and (3) it experiences extremely high levels of RNA editing, potentially higher than any other ptDNA sequence examined to date. It is predicted that RNA editing in *S. uncinata* restores the 79 non-standard start and stop codons found in the ptDNA protein-coding regions to their canonical state (Tsuji et al. 2007).

Although the ptDNA of *S. uncinata* has been described in detail, the plastid genomes from other *Selaginella* species remain unexplored and it is unknown if GC-biased ptDNA is a trait common to all members of the *Selaginella* genus or if it is restricted to only *S. uncinata*. One of the only reported cases of a GC-rich mitochondrial genome is that of the green algae *Polytomella capuana*. Interestingly, the mtDNA from other *Polytomella* species are AT rich (Mallet and Lee 2006; Smith and Lee 2008a); thus, it would be fascinating to see if the same trend is apparent for *Selaginella* ptDNA. It would also be intriguing to explore the mitochondrial and nuclear genomes from *Selaginella* taxa—if they too are GC rich, then their sequences may help pinpoint the processes that are biasing *Selaginella* nucleotide composition.

The *Selaginella* genus belongs to the class Lycopsida, the members of which are called lycophytes. Fossil records and phylogenetic analyses indicate that lycophytes are an ancient, monophyletic group of vascular plants [~400-million years old (Kenrick and Crane 1997)] comprised of three known families: the Isoetaceae (quillworts), the Lycopodiaceae (club mosses), and the Selaginellaceae. *Selaginella*, the only recognized genus within the Selaginellaceae, is an eclectic genus, containing around 700 species, which are spread throughout the world and cover an impressive range of habitats, including desert, tropical-rain-forest, alpine, and arctic habitats (Mabberley 1997). Significant efforts have been directed at resolving the evolutionary relationships among *Selaginella* species (Korall et al. 1999). Phylogenetic analyses using the plastid-encoded *rbcL* gene and nuclear-encoded rRNA genes suggest that the ancestries of *Selaginella* species are complex with many subgroups existing (Korall and Kenrick 2002, 2004); these analyses have also shown that *rbcL* substitution rates among *Selaginella* species are high relative to those observed within other land-plant families (Korall and Kenrick 2002); though, to a less dramatic extent when compared to other spore-producing vascular plants (Pryer et al. 2004).

One *Selaginella* species that has recently gained widespread attention is *Selaginella moellendorffii*; this is because in 2007 its nuclear genome was completely sequenced by the United States Department of Energy Joint Genome Institute (DOE JGI). *S. moellendorffii* was chosen as a candidate for sequencing because its nuclear genome is especially small (~110 mega bases [Mb]) relative to the nuclear DNA (nucDNA) of other land plants, and also because lycophytes represents an important evolutionary link between vascular plants and the nonvascular mosses, liverworts, and hornworts (Banks 2009). Now that the *S. moellendorffii* nuclear genome is sequenced, *Selaginella* is emerging as a model genus for comparative plant genomics.

The study presented here takes advantage of publicly-available DNA-and RNA-sequence data to investigate the evolution of nucleotide landscape in the plastid and mitochondrial genomes of *S. moellendorffii* and other *Selaginella* species. It is concluded that there is a genus-wide GC bias in *Selaginella* ptDNA and potentially one in the mtDNA as well, and that both of these nucleotide biases are affiliated with high levels of RNA editing. The overall implications of this nucleotide bias are discussed.

Methods

Assembly and verification of the *S. moellendorffii* organelle-genome sequences

The complete plastid-genome sequence of *S. moellendorffii* was generated by collecting and assembling ptDNA trace files produced by the DOE JGI *S. moellendorffii* nuclear-genome sequencing project (http://genome.jgi-psf.org/Selmo1/Selmo1.home.html). Trace files were data-mined from the NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml) using the *S. uncinata* ptDNA sequence as a BLAST (blastn 2.2.21+) query—similar approaches for assembling organelle genomes have been

used in previous studies (Smith and Lee 2008b; Smith and Lee 2009). The BLAST parameters were as follows: an expectation value (*E*-value) of 10; a word size of 11; match and mismatch scores of 2 and -3, respectively; and gap-cost values of 5 (existence) and 2 (extension). Trace files showing >90% sequence identity to the *S. uncinata* ptDNA in BLAST alignments were downloaded and assembled using CodonCode Aligner Version 2.0.6 (CodonCode Corporation, Dedham, MA, USA), which employs the Phred, Cross-match, and Phrap algorithms for base calling, sequence comparison, and sequence assembly, respectively. Assemblies were performed with a minimum-percent-identity score of 98, a minimum-overlap length of 500 nucleotides (nt), a match score of 1, a mismatch penalty of -2, a gap penalty of -2, and an additional first-gap penalty of -3. Assembly of the *S. moellendorffii* ptDNA trace files ultimately gave a complete plastid-genome sequence with >50-fold coverage.

To verify that no nuclear-genome-located ptDNA-like sequences (NUPTs) were collected, the entire *S. moellendorffii* nucDNA sequence was scanned for regions that show similarity to ptDNA. This was performed by blasting (blastn version 2.2.21+) the *S. moellendorffii* ptDNA sequence against the *S. moellendorffii* nuclear-genome sequence (v1.0) using the same parameters that are listed above. Only the first 150 scaffolds of the nuclear-genome assembly were analyzed: approximately 93.5% of the diploid nuclear genome is contained in these 150 scaffolds and their cumulative length is 198.93 Mb. PtDNA sequences that mapped to the nucDNA with >80% sequence identity and at least 30 nt of aligned length (in BLAST alignments) were counted as hits.

The same general approach as that described for the ptDNA was used to collect, assemble, and validate the 56 kilobases (kb) of *S. moellendorffii* mtDNA-sequence data presented in this study. The *Physcomitrella patens* and *Marchantia polymorpha* mitochondrial genomes (GenBank accession numbers NC_007945 and NC_001660, respectively) were used as BLAST queries to data-mine *S. moellendorffii* mtDNA trace files from the NCBI Trace Archive.

Scanning the *S. moellendorffii* nuclear genome for plastid-targeted sequences

The *S. moellendorffii* nuclear genome was scanned for plastid-targeted sequences by constructing a custom BLAST databank of the first 150 nucDNA scaffolds and then blasting this databank with ptDNA queries using an *E*-value of 5, a word size of 7, a match score of 2, a mismatch penalty of -3, a gap open score of 5, and a extend value of 2. All of the queries came from the *H. lucidula* ptDNA—specifically, the pool of genes that are located in the *H. lucidula* ptDNA but absent from the *S. moellendorffii* plastid genome. The TargetP server was employed for the prediction of plastid transit peptide sequences (Emanuelsson et al. 2007).

S. moellendorffii express sequence tags

Express-sequence-tag (EST) data for *S. moellendorffii* were obtained from the DOE JGI *S. moellendorffii* Genome Portal (v1.0) (http://genome.jgi-psf.org/Selmo1/Selmo1.home.html) on 1 January 2009. Plastid and mitochondrial RNA-derived ESTs were collected by blasting (employing the same BLAST parameters that were used for finding plastid-targeted sequences) this EST databank using *S. moellendorffii* ptDNA and mtDNA sequences as queries. All hits were subsequently checked against the *S. moellendorffii* nucDNA sequence to insure that they were not derived from nuclear-genome-located ptDNA-like or mtDNA-like sequences (NUPTs or NUMTs). The *S. moellendorffii* ESTs that map to the plastid and mitochondrial genomes are shown in Table S2 (see Supplementary Materials).

RbcL sequence data

The *Selaginella rbcL* sequences employed in this study come from either Korall and Kenrick (2002, 2004) or are unpublished data deposited in GenBank. A list of the *Selaginella* species from which *rbcL* sequences were data-mined (including GenBank accession numbers) is described in Table S3 (see Supplementary Materials)—note, the GC content of these sequences has neither been presented nor discussed elsewhere.

The other non-*Selaginella rbcL* sequences described in this study were collected by downloading from the NCBI nucleotide-sequence repository all of the entries that have an *rbcL* annotation and any of the following taxonomic identifications: Charophyceae, Marchantiophyta, Bryophyta, Lycopodiophyta, Moniliformopses, Coniferophyta, Cycadophyta, Ginkgophyta, Gnetophyta, and Magnoliophyta. Partial *rbcL* sequences were accepted as long as they were >900 nt in length.

XLSTAT-Pro, an add-in software package for Microsoft Excel, was employed for all statistical analyses of the *rbcL* dataset, including Tukey's Honestly Significant Difference (HSD) test.

Nucleotide-composition analyses

Nucleotide-composition analyses, including the GC content of first-, second-, and third-position codons sites, were determined with DAMBE (Xia and Xie 2001). The GC content of fourfold-degenerate sites (i.e., synonymous

sites) was calculated with INCA (Supek and Vlahovicek 2004) by measuring the proportion of G or C at third-position codon sites that can tolerate any of the four nucleotides without altering the amino acid specified.

Results and discussion

General features of the *S. moellendorffii* plastid genome

The entire ptDNA sequence of *S. moellendorffii* was produced by data-mining and assembling publicly-available sequences generated by the DOE JGI *S. moellendorffii* nuclear-genome sequencing project—see "Methods" for a detailed description of how this was performed. To ensure that no nuclear-genome-located ptDNA-like sequences (NUPTs) were collected, the complete *S. moellendorffii* nuclear genome was analyzed for regions that show similarity to ptDNA. The results of this analysis, described in Table 1, demonstrate that there are very few ptDNA-like sequences embedded in the nuclear genome: ~21.5 kb distributed over 307 sites in the diploid nucDNA sequence (this is at the lower end of what is observed for other land plants [Richly and Leister 2004]). These findings are a strong indication that the sequences used to assemble the *S. moellendorffii* plastid genome are derived from ptDNA and are not nuclear-encoded ptDNA-like sequences.

The *S. moellendorffii* plastid genome is 143.8 kb in length and assembles as a circular molecule (Fig. 1). Fifty-four percent (78 kb) of the genome codes for proteins and structural RNAs; the remaining 45.8% (65.8 kb) represents noncoding DNA, which can be subdivided into intergenic regions (57.8 kb) and introns (8 kb). A pair of inverted repeats, each with a length of 12.1 kb, divide the genome into a large- (83.7 kb) and a small-single-copy region (35.9 kb), referred to as the LSC and SSC regions. These statistics are similar to those of the *S. uncinata* ptDNA, with the exception that the *S. uncinata* plastid genome is 390 nt longer and its LSC and SSC regions have lengths of 77.7 and 40.9 kb, respectively—these size discrepancies are primarily due to the fact that the *S. uncinata* ptDNA harbours four pseudogenes and four gene duplicates that are absent from the *S. moellendorffii* ptDNA, and also because three genes in the SSC region of the *S. uncinata* ptDNA are found in the LSC region of the *S. moellendorffii* ptDNA (see Fig. 1 for details). The only other complete ptDNA sequence from a lycophyte, that of the club moss *H. lucidula* (Wolf et al. 2005), is ~10 kb longer than its *Selaginella* counterparts (because of a larger gene repertoire) and has a significantly larger LSC region (104.1 kb) and a much smaller SSC region (19.5 kb).

Annotation of the *S. moellendorffii* ptDNA sequence revealed 99 genes, 7 of which are duplicates found in the inverted repeats (Fig. 1); when ignoring these duplicates,

Table 1 Number of nucleotides in the *S. moellendorffii* diploid nuclear genome that share similarity with the plastid genome (by ptDNA subcategory)

	# Of similarity regions[a]	Mean length of similarity region (nt)	Longest similarity length (nt)	Cumulative length of similarity regions (nt)	Fraction of nuclear genome
PtDNA subcategories[b]					
Protein-coding genes[c]	100	72	229	7,207	3.62×10^{-5}
Structural-RNA genes[d]	142	55	91	7,746	3.89×10^{-5}
Introns[e]	19	91	301	1,726	0.86×10^{-5}
Intergenic spacers[f]	93	52	138	4,823	2.42×10^{-5}
Complete plastid genome[g]	307	67	530	21,502	10.79×10^{-5}

Nuclear DNA analyses are based on the *S. moellendorffii* draft nuclear-genome sequence (v1.0). Only the first 150 scaffolds of the nuclear-genome assembly were analyzed; ~93.5% of the *S. moellendorffii* nucDNA is contained in these 150 scaffolds and their cumulative length is 198.93 mega bases (note: this length represents two haplotypes)

[a] The number of distinct regions in the *S. moellendorffii* nucDNA that show >80% sequence identity and at least 30 nt of aligned length to the plastid DNA

[b] Refers to the region of the plastid genome to which the nucDNA maps

[c] Includes all of the identified protein-coding genes

[d] Includes all of the identified tRNA- and rRNA-coding genes

[e] Includes all of the identified introns

[f] Includes all of the identified intergenic regions, including pseudogenes

[g] Some of the "ptDNA similarity regions" in the *S. moellendorffii* nucDNA consist of coding and noncoding ptDNA-like sequences; thus, the overall sum of similarity regions (307) is less than the sum of the similarity regions based on ptDNA subcategory (354)

Fig. 1 Genetic map of the *Selaginella moellendorffii* plastid genome. Transfer RNA-coding regions are designated by the single-letter abbreviation of the amino acid they specify (their anticodon sequence is shown in *brackets*). Introns are *black* (all introns are of group-II affiliation). Pseudogenes are labeled with a ψ. *Gray blocks* highlight regions of the genome for which cDNA sequences were data-mind. Differences in the ptDNA gene order between *S. moellendorffii* and *S. uncinata* are shown with *blocks*, *arrows*, and *symbols* (\$ and #). The genes and pseudogenes in the *S. uncinata* plastid genome that are absent from the *S. moellendorffii* ptDNA are signified as follows: $\otimes^1$ = *rpl23* and the 5′ end of *rpl2*; $\otimes^2$ = *psaM* pseudogene; $\otimes^3$ = *chlL* pseudogene; $\otimes^4$ = *rps12* pseudogene; $\otimes^5$ = *rpl21* pseudogene; $\otimes^6$ = duplicate copy of *psbK* and *trnQ*

there are 75 protein-, 4 rRNA-, and 13 tRNA-coding genes (including tRNA *fMet*), which is among the most reduced ptDNA gene contents from any photosynthetic land plant examined to date. Pseudogenes of *accD*, *rpl33*, and *infa* were identified; the presumed functional copies of these loci were discovered in the nucDNA (see "Methods" for details). Eleven group-II introns, all within protein-coding genes, were also discerned from the ptDNA sequence (Fig. 1). The ptDNA gene complement of *S. moellendorffii*, including introns and pseudogenes, mirrors that of *S. uncinata*, with some exceptions: (1) the *S. uncinata* ptDNA contains duplicate copies of *psbK*, *trnQ*, *rpl23*, and the 5′-end of *rpl2*, whereas in the *S. moellendorffii* plastid genome these genes are present only once; (2) the *S. uncinata* ptDNA harbours pseudogenes for *chlL*, *psaM*, *rps12*, and *rpl21* (the latter three loci exist in the ptDNA only as pseudogenes), whereas the *S. moellendorffii* plastid genome contains only a functional *chlL* and has neither functional nor pseudogene copies of *psaM, rps12 or rpl21*. A scan of the *S. moellendorffii* nucDNA did not expose functional copies of these loci. They most likely exist in the nucDNA but were not uncovered because of their small size and relatively nonconserved sequence; and (3) the *S. moellendorffii* ptDNA encodes *trnL*, a gene that is absent from the *S. uncinata* ptDNA. Compared to the *H. lucidula* ptDNA, the *S. moellendorffii* plastid genome has 16 fewer tRNA-coding genes and 10 fewer protein-coding genes.

The relatively reduced ptDNA gene repertoires of *S. moellendorffii* and *S. uncinata* are reflections of the surprisingly small number of tRNAs encoded in these genomes (13 and 12, respectively, not including duplicates). Their nearest rivals are the plastid genomes of the alveolates *Babesia bovis* and *Theileria parva*, which each encode 24 tRNAs, and the ptDNA of the parasitic angiosperm *Epifagus virginiana*, which encodes 23 tRNAs. It is unknown how *S. moellendorffii* and *S. uncinata* compensate for the tRNA-coding genes that appear to be absent from their plastid genomes. One hypothesis is that they are encoded in the nuclear genome and imported to the plastid from the cytosol—a similar process is known to occur for plant mitochondria (Glover et al. 2001). A scan of the *S. moellendorffii* nucDNA for the missing plastidial tRNAs

(using plastid-encoded tRNAs from a close relative as search queries) revealed only one putative plastid-bound tRNA: *trnP*-CGG. An alternative hypothesis is that the missing tRNAs are imported to the plastid from the mitochondria—a process also proposed for *E. virginiana* (Modern et al. 1991). However, analysis of a 56 kb portion of the *S. moellendorffii* mitochondrial genome uncovered no tRNA-coding genes, suggesting that the mtDNA of *S. moellendorffii*, like those from other land plants, has a reduced tRNA-coding suite. A final hypothesis is that novel tRNAs are generated from those encoded in the ptDNA through RNA editing, a topic discussed in more detail below.

The ptDNA gene order for *S. moellendorffii* is similar to that of *S. uncinata*, with one significant difference: the *S. moellendorffii* plastid genome lacks a 20-kb inversion (from *trnC* to *psbI*) found in the *S. uncinata* ptDNA. This inversion, which is also absent from the *H. lucidula* ptDNA and available plastid-genome sequences from bryophytes, is commonly found in the ptDNA of higher ferns and seed plants (Palmer and Stein 1986; Raubeson and Jansen 1992). In addition, *rpl23* and the 5′-end of *rpl2*, which are a part of the inverted repeat in the *S. uncinata* ptDNA, are in the LSC region of the *S. moellendorffii* plastid genome; and the position of one protein-coding- and four tRNA-coding-genes in the *S. moellendorffii* ptDNA (*petN, trnD*, *trnE*, *trnF,* and *trnY*) differ from that in the *S. uncinata* ptDNA. In a general sense, the *S. moellendorffii* ptDNA gene order is intermediary to that of *S. uncinata* and *H. lucidula*, and shares more similarities with bryophyte ptDNA than with those of other vascular plants. The discrepancies in gene order and gene content between the *S. moellendorffii* and *S. uncinata* plastid genomes are outlined with blocks, arrows, and symbols on Fig. 1.

Nucleotide landscape of the *S. moellendorffii* plastid genome

The overall GC content of the *S. moellendorffii* ptDNA is 51%, which is less than that of *S. uncinata* (54.8%) but still the second most GC-rich plastid genome observed to date. A schematic compilation comparing the ptDNA GC content of *S. moellendorffii* to that of completely-sequenced plastid genomes is shown in Fig. 2. From this plot it is apparent that the nucleotide composition of ptDNA forms a continuum from approximately 20–40% GC, to the exclusion of the *S. moellendorffii* and *S. uncinata* plastid genomes, which are positioned outside of this continuum, well above all other available ptDNA sequences in terms of GC content. Note that the lycophyte *H. lucidula* has a more typical ptDNA GC content of 36.2% (Fig. 2).

Among the different portions of the *S. moellendorffii* plastid genome, RNA-coding regions have the highest GC content (57.8%), followed by protein-coding regions (55.5%), introns (50.3%), and intergenic spacers (49.9%). The inverted repeats are more GC-rich (55.7%) than the SSC region (50.5%) and the LSC region (49.9%). The allocation of G versus C (GC skew) on the main sense strand (the strand depicted in Fig. 1) is negligible with a value of only 0.0003. These trends parallel that of the *S. uncinata* ptDNA. It is noteworthy that the RNA-coding regions from other plastid genomes also tend to be GC-biased, having an average GC content of 52.9% (SD = 4.9%) among

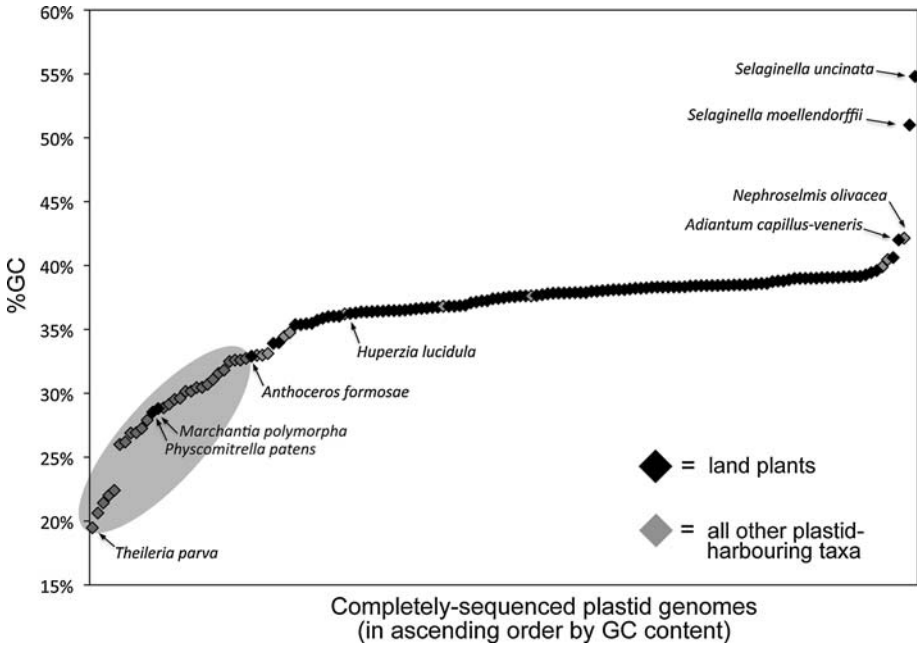

Fig. 2 The GC content of completely-sequenced plastid genomes. Genomes are organized in ascending order by GC content. The data points corresponding to species of interest are labeled. The *shaded oval* highlights some of the species for which plastid RNA-editing is believed to be either absent or restricted to less than five edited sites. The ptDNA GC-content data from which this graph was plotted are listed in Table S1 (see Supplementary Materials)

completely-sequenced ptDNAs; however, the intergenic-spacer, intronic and protein-coding regions from plastid genomes are generally skewed towards A and T.

Within the protein-coding ptDNA of *S. moellendorffii*, the average GC content of first-position codon sites (55.9%) exceeds that of second- (50.8%) and third-positions (44.8%). A comparison of these data with those from *S. uncinata* and other available plastid-genome sequences is presented in Figure 3 (the raw data from which this figure was derived are shown in Table S1 [see Supplementary Materials]). It is evident from this figure that the overall GC content of the *S. moellendorffii* (and *S. uncinata*) protein-coding regions is the result of a relatively inflated GC content at all three codon-site positions; although, among codon sites, third-position synonymous sites in the *S. moellendorffii* and *S. uncinata* ptDNAs (46.5 and 51% GC, respectively) depart most significantly in GC content from those of other available plastid genome sequences, which on average are 25.3% GC (SD 7.7%). GC-rich codons (those that code for the amino acids alanine, glycine, proline, and in some cases arginine) represent 30% of the codons found in the *S. moellendorffii* plastid genome. The proportion of GC-rich codons in the *S. uncinata* ptDNA is even greater at 34%, whereas the GC-rich-codon composition from other completely-sequenced ptDNAs is on average only 17%.

The ptDNA GC content of other *Selaginella* species

The observation of GC-rich ptDNA in *S. moellendorffii* and *S. uncinata* raises questions regarding the phylogenetic distribution of GC content within the Selaginellaceae, such as: is GC-biased ptDNA a trait common to many (or all) members of the *Selaginella* genus? And if yes, is *Selaginella* truly an outlier in terms of ptDNA nucleotide composition, or are there other plant lineages with similarly high GC contents? To address these questions, *rbcL* ptDNA sequences from a series of diverse plant taxa, including over 100 *Selaginella* species (representing most of the species diversity within the genus), were data-mined from NCBI and assessed for their GC-content. The *rbcL* gene was chosen as a ruler for assessing the overall plastid-genome nucleotide composition because it is one of the only ptDNA genes whose sequence is readily available for many plant species (due to the fact that it is often used for phylogenetic analyses) and because its GC content scales reasonably well with the overall plastid-genome GC content: for complete ptDNA sequences, the Pearson correlation coefficient between the *rbcL* GC content and the whole-genome GC content is 0.82 ($r^2 = 0.76$). Altogether, *rbcL* sequences were collected for 167 charophytes, 911 liverworts, 811 mosses, 62 hornworts, 103 *Selaginella* species, 87 "non-*Selaginella*" lycophytes, 2,848 monilophytes, 855 gymnosperms, and 2,100 angiosperms. Summary statistics of the *rbcL* GC contents for these different plant lineages are shown in Fig. 4.

The mean *rbcL* GC content for *Selaginella* species (52%; SD = 1.7%) is significantly higher than that from other plant lineages (Fig. 4), including other lycophytes, which have an average *rbcL* GC composition of only 42.7% (SD = 0.9%). The monilophytes and gymnosperms are the closest to *Selaginella* with respect to *rbcL* GC content with values of 46.2% (SD = 2.3%) and 44.3% (SD = 1.3%), respectively. The charophytes and liverworts have the lowest observed mean *rbcL* GC contents at 39.6%, with standard deviations of 3.2% (charophytes) and 2.0% (liverworts). Overall, these

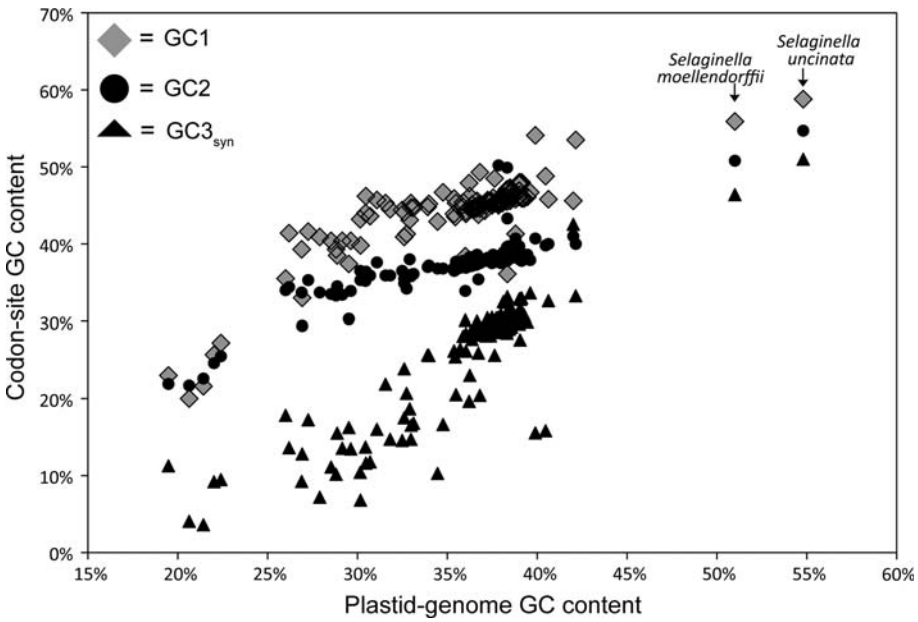

Fig. 3 Scaling of plastid-genome GC content with GC content at different codon-site positions. GC1, GC2, and $GC3_{syn}$ represent the GC content at first-position, second-position, and third-position-synonymous codon sites, respectively. The ptDNA GC-content data from which this graph was plotted are listed in Table S1 (see Supplementary materials)

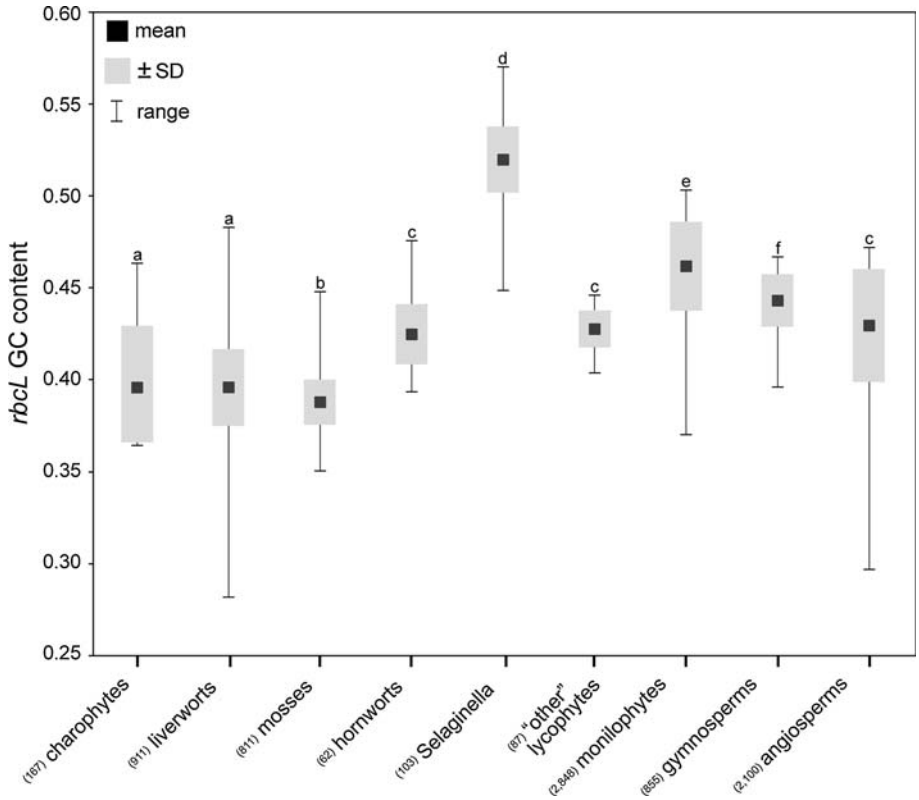

Fig. 4 The *rbcL* GC content for major plant lineages. The number of species sampled from each lineage are shown in *brackets*. Plots with the *same letter* are not significantly different from one another (under the Tukey–Kramer method)

findings suggest that *Selaginella* ptDNA has become GC biased since the Selaginellaceae diverged from their common ancestor with quillworts and club mosses, which is believed to have occurred at least 400 million years ago (Kenrick and Crane 1997; Banks 2009).

All 103 *Selaginella* species that were analyzed have an *rbcL* GC content above 50%, except for *Selaginella sinensis*, which has an *rbcL* GC content of 44.8% (Figure S1 and Table S3 [see Supplementary Materials]). The most extreme *rbcL* GC content is observed for *Selaginella fragilis* (57.0%), which is greater than that of *S. moellendorffii* (50.6%) and *S. uncinata* (53.2%), and suggests that the ptDNA of *S. fragilis* may have a higher overall GC content than *S. uncinata* (i.e., >55%). The seven highest *rbcL* GC contents (ranging from 55–57% GC) come from *Selaginella* species that belong to the South American articulate subclade (Figure S1 [see Supplementary Materials]). Support for this subclade come from parsimony and Bayesian analyses using *rbcL* (Korall and Kenrick 2002, 2004) and from the observation that the *Selaginella* taxa that form this subclade possess a unique morphological marker: the rhizosphore develops from the upper surface of the stem and loops over the branch to grow downwards whereas in other *Selaginella* species it develops on the lower surface of the stem (Korall and Kenrick 2002). It should be mentioned that when maximum-likelihood analyses were performed on the *rbcL* dataset used by Korall and Kenrick (2002, 2004) the South American articulate subclade is still observed (data not shown). The phylogenetic affiliation of *S. sinensis*, the only *Selaginella* species shown to have an *rbcL* GC content below 50%, remains problematic. Parsimony analyses using *rbcL* place it (with low bootstrap support) as a sister to a clade containing all other species in the genus (Korall and Kenrick 2004). This could be an indication that the occurrence of GC-rich ptDNA in *Selaginella* taxa evolved after the split between the lineage that gave rise to *S. sinensis* and that leading to the other *Selaginella* species investigated in this study. That being said, parsimony inferred phylogenies are particularly sensitive to nucleotide composition biases (Eyre-Walker 1998), meaning they can cause distantly related organisms with similar GC contents to look more closely related than they actually are. Bayesian analyses with the same *rbcL* dataset (Korall and Kenrick 2004) place the GC-poor *S. sinensis* in a well-supported subclade with GC-rich *Selaginella* species—this position of *S. sinensis* is also supported by parsimony and Bayesian inferred phylogenies of 26S rDNA sequence data from *Selaginella* species (Korall and Kenrick 2004).

Evolution of nucleotide composition in *Selaginella* ptDNA

Why is *Selaginella* ptDNA GC-biased? Or rather, why is *Selaginella* ptDNA not enriched in A and T like other available plastid-genome sequences? For virtually all completely-sequenced plastid genomes the AT content is highest at what are assumed to more neutrally evolving

positions, such as fourfold-degenerate sites and noncoding regions (collectively defined as silent sites), and it is lowest at the more functionally constrained sites (first- and second-position codons sites and RNA-coding regions). Thus, it is generally believed that a neutral process, such as AT-mutation pressure or AT-biased gene conversion, is driving the nucleotide composition of most plastid genomes towards A and T (Howe et al. 2003; Kusumi and Tachida 2005; Khakhlova and Bock 2006). Could the observed GC content of *Selaginella* ptDNA be caused by the absence of either AT-mutation pressure or AT-biased gene conversion, or both? In the plastid genomes of *S. moellendorffii* and *S. uncinata* the GC content of silent sites is on average 48.9 and 52.5%, respectively. Similar values are also seen for the *rbcL* data from the different *Selaginella* species where the average GC content of fourfold degenerate sites is 50.5% (Table S3 [see Supplementary Materials]). Taken as a whole, these findings on the silent-site nucleotide composition of *Selaginella* ptDNA could (because %GC ≈ %AT) be a reflection of an unbiased mutation/gene-conversion process. There is also the possibility that two opposing neutral forces, such as AT-mutation pressure coupled with a GC-biased gene conversion mechanism (or GC-mutation pressure coupled with an AT-conversion bias) are balancing the silent-site nucleotide composition of *Selaginella* ptDNA resulting in a GC content of ~50%. The fact that *rbcL* synonymous substitution rates among *Selaginella* species are exceptionally high relative to those observed within most other land-plant families (Korall and Kenrick 2002) may be an indication of an elevated mutation rate in *Selaginella* ptDNA; if true, this may imply a scenario where biased mutation pressure (AT or GC) is offset by a biased gene conversion mechanism.

There is also the possibility that natural selection is influencing the nucleotide composition of *Selaginella* ptDNA, and this may explain why, in addition to an elevated silent-site GC content, the more functionally constrained positions in *Selaginella* ptDNA, such as first- and second-position codon sites, are also skewed towards G and C relative to other plastid genomes (Fig. 3). GC-richness could be interpreted as an adaptation for thermostability or UV-light tolerance. The thermostability hypothesis seems unlikely when considering that *Selaginella* species from northern climates, like Alaska, Canada, and Siberia have equally high *rbcL* GC contents as those from tropical and desert habitats. That being said, both hot- and cold-climate *Selaginella* species tend to grow in environments with reasonably high levels of UV radiation (Table S3 [see Supplementary Materials]). Another adaptive hypothesis could be that there is selection for translational efficiency. Approximately 1/3 of the codons in the *S. moellendorffii* and *S. uncinata* plastid genomes are GC-rich, which may correlate with the specific pool of tRNA anticodons that are available for plastid-gene translation. This topic is difficult to address because both the *S. moellendorffii* and *S. uncinata* plastid genomes encode a limited number of tRNAs. Nevertheless, of the 13 and 12 unique tRNAs that are respectively encoded in the *S. moellendorffii* and *S. uncinata* plastid genomes, all except one (*trnR*-ACG) are cognate to AT-rich codons. There is reason to believe, however, that many of the GC-rich codons in both the *S. moellendorffii* and *S. uncinata* plastid genomes are changed into AT-rich codons through RNA editing. If true, RNA editing may be influencing the GC content of *Selaginella* ptDNA.

RNA editing in *Selaginella* ptDNA and its impact on nucleotide composition

A few observations indicate that RNA editing is an important, widespread, and frequently occurring phenomenon in the plastids of *Selaginella* species. For instance, of the 75 protein-coding genes encoded in the *S. moellendorffii* plastid genome, 41 contain non-canonical start codons (ACG instead of ATG) and 23 have non-canonical stop codons (CGA, CAA, or CAG instead of TGA, TAA or TAG); the incidence of irregular start/stop codons in the *S. uncinata* ptDNA is even more prevalent with 50 and 29 non-canonical start and stop codons, respectively (Tsuji et al. 2007). If these codons were left unedited in mature transcripts, it would imply that 66% of the protein-coding genes in the *S. moellendorffii* ptDNA and 83% of those in the *S. uncinata* ptDNA are non-functional. However, preliminary investigations of plastid-complementary-DNA (cDNA) sequence data from *S. moellendorffii* (this study) and *S. uncinata* (Tsuji et al. 2007) indicate that RNA editing restores these irregular start/stop codons to their canonical states and induces C→U conversions in other plastid-RNA regions as well.

Analyses of 8,291 nt of cDNA sequence data from the *S. moellendorffii* plastid genome (4,501 nt from coding regions, 901 nt from intronic regions, and 2,889 nt from intergenic regions) reveals 104 edited sites, all of them corresponding to C→U changes (Table 2). Fifty-eight of these edited sites map to protein-coding ptDNA, 16 to intronic portions of the genome, and 30 to intergenic regions (Table 2). Of the cDNA data that covers nonstandard stop/start codons, all were restored to their canonical states. For *S. uncinata*, cDNA studies of the *rbcL* and *atpB* genes uncovered 54 and 111 C→U edited sites, respectively (Tsuji et al. 2007); this is one of the more massive examples of plastid RNA editing observed to date. These data, albeit providing only a small window into the degree of plastid RNA editing in *S. moellendorffii* and *S. uncinata*, indicate that RNA editing is a critical and prevalent process in these two taxa, and operates at a greater level than that

Table 2 RNA-editing sites in the *Selaginella moellendorffii* plastid genome

	Length (nt)[a]	# Of RNA editing sites[b]
Protein-coding (by gene)		
ndhA	411	25
ndhH	128	6
psbH	234	21
psbN	132	1
psbT	102	5
Intronic (by gene)		
ndhA-intron	750	16
Intergenic (by region)		
ndhA/ndhH	127	4
psbC/psbZ	451	1
psbN/psbH	86	3
psbH/petB	800	22

RNA-editing data could only be collected for regions described above; these data were derived from EST sequences produced by the DOE JGI *S. moellendorffii* nuclear-genome sequencing project

[a] Refers to the length of the EST sequence covering the given region

[b] All observed RNA editing sites involve C→U changes

currently observed in other land plants. Studies suggest that in the plastid mRNAs of seed plants there are approximately 15–44 C→U editing sites [see Tillich et al. (2006) for a review]. A significantly larger number of plastid RNA editing sites are observed for the fern *Adiantum capillus-veneris*, which has 315 C→U and 35 U→C editing sties (Wolfe et al. 2004), and the hornwort *Anthoceros formosae*, which has 509 C→U and 433 U→C editing sites (Kugita et al. 2003). Organelle RNA editing in land plants is believed to be of monophyletic origin (Tillich et al. 2006). Although, among different species of land plant, the levels of RNA editing and the sites that get edited appear to be highly lineage specific (Jobson and Qiu 2008), leaving open the possibility that RNA editing has arisen multiple times in land-plant evolution.

Considering that plastid RNA editing mostly involves C→U changes (with the exception of *A. formosae*), and that first- and second-position codon sites are generally the most edited positions in plastid genomes (Tillich et al. 2006; Jobson and Qiu 2008), then the elevated GC content of first- and second-position codon sites in the *S. moellendorffii* and *S. uncinata* ptDNAs may be a reflection of a large number of RNA editing sites in these genomes—an idea also suggested for *S. uncinata* by Tsuji et al. (2007). If true, this would imply a positive correlation between the ptDNA GC content and the number of C→U RNA editing sites. A cursory scan of complete land-plant ptDNA sequences reveals that those with a large number of C→U RNA editing sites, such as *A. capillus-veneris*, are more GC-rich than those with only a few RNA editing sites (Fig. 1). Indeed, next to *S. moellendorffii* and *S. uncinata*, *A. capillus-veneris* has the most GC-biased land-plant plastid genome observed to date (42% GC), whereas *Marchantia polymorpha*, which appears to lack plastid RNA editing (Freyer et al. 1997), and *Physcomitrella patens*, which is believed to have less than five plastid RNA editing sites (Miyata and Sugita 2004; Rüdinger et al. 2009), are the two most AT-rich land-plant plastid genomes sampled thus far (Fig. 1). The correlation between genomic GC content and levels of RNA editing has been highlighted in other studies (Malek et al. 1996; Jobson and Qiu 2008). Given these observations, one could suggest that RNA editing is acting as a genomic buffer against GC-biased mutation/conversion pressure by neutralizing T→C mutations, specifically those at functionally important first- and second-position codon sites. That being said, the evolution of RNA editing is a complicated topic and many sophisticated (and well articulated) models for its origins exist (Covello and Gray 1993; Lynch et al. 2006; Tillich et al. 2006; Jobson and Qiu 2008). I favor the model of Covello and Gray (1993), which posits that both the origin of RNA editing activity and the fixation of mutations at editable sites evolved primarily through random genetic drift but that the maintenance of RNA editing activity at specific sites is the result of natural selection. A salient point for any debate on RNA editing is that the genomic and nucleotide-composition contexts under which RNA editing evolved are unknown. The RNA editing machinery may have originated in a species with a moderately AT-rich organelle genome, but one whose descendants were exposed to increasing GC pressure [see Reviewers' comments in Jobson and Qiu (2008)]. If RNA editing is linked to the high GC content of *Selaginella* ptDNA, studies on the different *Selaginella* species, especially those at either extremes of the nucleotide-composition spectrum, may give insight into the link between RNA editing and GC content.

The GC content in other genetic compartments of *Selaginella* species

The observation that *Selaginella* species have relatively GC-rich plastid genomes raises questions regarding their mtDNA—is it also GC-rich? An attempt at answering this question was made by collecting 56 kb of *S. moellendorffii* mtDNA sequence data. These data were generated by assembling mtDNA trace files produced by the DOE JGI *S. moellendorffii* nuclear-genome sequencing project (see "Methods" for details). All of the mtDNA sequences were subsequently blasted against the *S. moellendorffii* nucDNA assembly to ensure that no nuclear-genome-located mtDNA-like sequences (NUMTs) were collected. The results of these analyses suggest that there are a relatively small number of NUMTs in the nucDNA, <20 kb distributed over ~50 sites (based on the 56 kb of mtDNA that

Table 3 General feature of the mtDNA sequences collected from *S. moellendorffii*

	Length (nt)	%GC	GC1	GC2	GC3	# Of RNA editing sites[c]
All sites	55,780	67.8	–	–	–	–
rRNA-coding (*26S*)	4,150	65.1	–	–	–	–
Protein-coding (overall)	7,742	62.0	64.2	60.2	61.5	–
Protein-coding (by gene)						
atp1	1,515	64.2	69.3	64.4	58.8	78
atp9[a]	74	62.2	33.3	70.8	79.2	–
cox1[a]	864	63.9	61.5	66.7	63.5	74
nad2	1,055	63.2	68.1	60.4	61.3	–
nad4[a]	879	61.5	58.0	65.2	61.5	–
nad5	1,704	60.1	59.5	54.2	66.5	–
nad7	1,096	59.8	68.5	54.0	56.7	–
nad9	555	62.2	67.5	60.0	58.9	–
Intronic (overall)	27,490	69.1	–	–	–	–
Intronic (by gene)						
atp9	3,547	69.7	–	–	–	–
cox1	6,825	69.4	–	–	–	–
nad2	4,798	68.6	–	–	–	–
nad4	2,013	66.9	–	–	–	–
nad5	5,580	69.5	–	–	–	–
nad7	4,727	69.0	–	–	–	–
Intergenic (overall)	16,398	68.9	–	–	–	–
Intergenic (by region)[b]						
?/*atp1*	930	67.1	–	–	–	–
atp1/*nad5*	461	65.7	–	–	–	–
nad5/?	2,916	70.6	–	–	–	–
?/*atp9*	356	71.3	–	–	–	–
?/*nad2*	769	66.8	–	–	–	–
nad2/?	652	67.8	–	–	–	–
?/*nad4*	283	67.8	–	–	–	–
nad4/?	175	58.3	–	–	–	–
?/*nad7*	540	65.7	–	–	–	–
nad7/?	317	69.1	–	–	–	–
nad9/*26S*	8,999	69.3	–	–	–	–

GC1, GC2, and GC3 are the GC contents at first-, second-, and third-position codon sites, respectively

[a] Partial sequence

[b] A question mark (?) is used when the adjacent gene is undetermined

[c] RNA-editing data could only be collected for *cox1* and *atp9*; these data were derived from EST sequences produced by the DOE JGI *S. moellendorffii* nuclear-genome sequencing project

were collected), and the NUMTs that are present are highly degenerate. Several attempts were made to complete the *S. moellendorffii* mitochondrial-genome sequence; however, the mtDNA of *S. moellendorffii*, like that from other land-plant taxa, contains an abundance of repeats, which have spread throughout most of the intergenic and intronic regions. This feature of the *S. moellendorffii* mitochondrial genome means that the mtDNA trace files corresponding to intergenic and intronic regions collapse on top of one another upon assembly, resulting in a network of spurious repetitive motifs. Nonetheless, enough mtDNA sequence data were characterized to confidently describe the nucleotide landscape of this genome. In total, eight protein-coding genes, 1 rRNA-coding gene, and 11 intergenic spacers were collected; the lengths and nucleotide compositions of these regions are summarized in Table 3. The overall GC content of the 56 kb of mtDNA sequence data is 67.8%, ranging from 62% for protein-coding genes,

65.1% for rRNA-coding genes, 69.1% for intronic regions, and 68.9% for intergenic spacers (Table 3). These nucleotide composition statistics suggests that the *S. moellendorffii* mtDNA is the most GC-rich mitochondrial genome observed to date, exceeding that of the green alga *Polytomella capuana*, which is 57% GC (Smith and Lee 2008a). Could similar processes be driving the nucleotide composition of the *S. moellendorffii* mitochondrial and plastid genomes towards G and C? Analyses of cDNA sequences for *atp1* and *nad5* revealed 78 and 74 C→U RNA editing sites, respectively (Table 3). This implies that the *S. moellendorffii* mitochondrial genome is, like its plastid counterpart, experiencing exceptionally high levels of RNA editing; moreover, this provides support for the notion that RNA editing is connected to the high GC content of the *S. moellendorffii* organelle DNA. Interestingly, it is believed that in land plants the same machineries are responsible for editing both the mitochondrial and plastid derived transcripts (Freyer et al. 1997; Steinhauser et al. 1999).

On a final note, the GC content of the *S. moellendorffii* nuclear genome is ~45%—based on analyses of the first 150 scaffolds of the diploid genome assembly (~93.5% of the complete nuclear-genome sequence), which is unremarkable in comparison to the nuclear genomes from other land plants. The discordance between the nucleotide composition of the *S. moellendorffii* nucDNA and organelle DNA could be a reflection of different mutation/conversion biases in these genomes. One crucial point, however, is that land plant nucDNA, unlike its organelle counterparts, is believed to experience very little (if any) RNA editing, thus, reinforcing the notion that RNA editing is connected to the high GC content of the *S. moellendorffii* organelle genomes.

Conclusions

There is a genus-wide GC bias in *Selaginella* ptDNA, which is most pronounced in South American articulate species. GC-rich ptDNA appears to be something unique to *Selaginella* species and is absent from lycophytes outside of the Selaginellaceae. It is argued that the cause of this GC bias is a combination of reduced AT-mutation pressure relative to other plastid genomes and a large number of C→U RNA editing sites. Partial-genome analysis of the *S. moellendorffii* mtDNA indicates that it is also GC biased (even more so than the ptDNA) and is arguably the most GC-rich organelle genome observed to date—the high GC content of the mtDNA also appears to be influenced by RNA editing. These findings provide convincing support for the earlier proposed theory that the GC content of land-plant organelle DNA is positively correlated (and directly connected) to the levels of organelle RNA editing.

Acknowledgments Many thanks to Robert W. Lee for insightful comments and to Jo Ann Banks for giving me permission to use the *S. moellendorffii* nuclear genome project trace file data. This work was supported by a grant to R.W.L. from the Natural Sciences and Engineering Research Council (NSERC) of Canada. D.R.S. is an Izaak Walton Killam Memorial Scholar and holds a Canada Graduate Scholarship from NSERC.

References

Banks JA (2009) Selaginella and 400 million years of separation. Annu Rev Plant Biol 60:223–238

Covello PS, Gray MW (1993) On the evolution of RNA editing. Trends Genet 9:265–268

Dybvig K, Voelker LL (1996) Molecular biology of mycoplasmas. Ann Rev Microbiol 50:25–57

Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP, and related tools. Nat Protoc 2:953–971

Eyre-Walker A (1998) Problems with parsimony in sequences of biased base composition. J Mol Evol 47:686–690

Freyer R, Kiefer-Meyer MC, Kössel H (1997) Occurrence of plastid RNA editing in all major lineages of land plants. Proc Natl Acad Sci 94:6285–6290

Glover KE, Spencer DF, Gray MW (2001) Identification and structural characterization of nucleus-encoded transfer RNAs imported into wheat mitochondria. J Biol Chem 276:639–648

Howe CJ, Barbrook AC, Koumandou VL, Nisbet RER, Symington HA, Wightman TF (2003) Evolution of the chloroplast genome. Philos Trans R Soc Lond B Biol Sci 358:99–107

Jobson RW, Qiu YL (2008) Did RNA editing in plant organellar genomes originate under natural selection or through genetic drift? Biol Direct 3:43

Kenrick P, Crane PR (1997) The origin and early evolution of plants on land. Nature 389:33–39

Khakhlova O, Bock R (2006) Elimination of deleterious mutations in plastid genomes by gene conversion. Plant J 46:85–94

Korall P, Kenrick P (2002) Phylogenetic relationships in Selaginellaceae based on *rbcL* sequences. Am J Bot 89:506–517

Korall P, Kenrick P (2004) The phylogenetic history of Selaginellaceae based on DNA sequences from the plastid and nucleus: extreme substitution rates and rate heterogeneity. Mol Phylogenet Evol 31:852–864

Korall P, Kenrick P, Therrien JP (1999) Phylogeny of Selaginellaceae: evaluation of generic/subgeneric relationships based on *rbcL* gene sequences. Int J Plant Sci 160:585–594

Kugita M, Yamamoto Y, Fujikawa T, Matsumoto T, Yoshinaga K (2003) RNA editing in hornwort chloroplasts makes more than half the genes functional. Nucleic Acids Res 31:2417–2423

Kusumi J, Tachida H (2005) Compositional properties of green-plant plastid genomes. J Mol Evol 60:417–425

Lane CE, van den Heuvel K, Kozera C, Curtis BA, Parsons B, Bowman S, Archibald JM (2007) Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. Proc Natl Acad Sci USA 104:19908–19913

Lynch M, Koskella B, Schaack S (2006) Mutation pressure and the evolution of organelle genomic architecture. Science 311:1727–1730

Mabberley DJ (1997) The plant-book: a portable dictionary of the vascular plants. Cambridge University Press, Cambridge

Malek O, Lättig K, Hiesel R, Brennicke A, Knoop V (1996) RNA editing in bryophytes and a molecular phylogeny of land plants. EMBO J 15:1403–1411

Mallet M, Lee RW (2006) Identification of three distinct *Polytmella* lineages based on mitochondrial DNA features. J Eukaryot Microbiol 53:79–84

Miyata Y, Sugita M (2004) Tissue- and stage-specific RNA editing of *rps14* transcripts in moss (*Physcomitrella patens*) chloroplasts. J Plant Physiol 161:113–115

Modern CW, Wofe K, dePamphilis CW, Palmer JD (1991) Plastid translation and transcription genes in a non-photosynthetic plant: intact, missing and pseudo genes. EMBO J 10:3281–3288

Morton BR (1993) Chloroplast DNA codon use: evidence for selection at the psb A locus based on tRNA availability. J Mol Evol 37:273–280

Morton BR (1998) Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. J Mol Evol 46:449–459

Ogata H, Audic S, Renesto-Audiffren P et al (2001) Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science 293:2093–2098

Palmer JD, Stein DB (1986) Conservation of chloroplast genome structure among vascular plants. Curr Genet 10:823–833

Pryer KM, Schuettpelz E, Wolf PG, Schneider H, Smith AR, Cranfill R (2004) Phylogeny and evolution of ferns (monilophytes) with a focus on the early leptosporangiate divergences. Am J Bot 91:1582–1598

Raubeson LA, Jansen RK (1992) Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. Science 255:1697–1699

Richly E, Leister D (2004) NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. Mol Biol Evol 21:1972–1980

Rüdinger M, Funk HT, Rensing SA, Maier UG, Knoop V (2009) RNA editing: only eleven sites are present in the *Physcomitrella patens* mitochondrial transcriptome and a universal nomenclature proposal. Mol Genet Genomics 281:473–481

Smith DR, Lee RW (2008a) Mitochondrial genome of the colorless green alga *Polytomella capuana*: a linear molecule with an unprecedented GC content. Mol Biol Evol 25:487–496

Smith DR, Lee RW (2008b) Nucleotide diversity in the mitochondrial and nuclear compartments of *Chlamydomonas reinhardtii*: investigating the origins of genome architecture. BMC Evol Biol 8:156

Smith DR, Lee RW (2009) The mitochondrial and plastid genomes of *Volvox carteri*: bloated molecules rich in repetitive DNA. BMC Genomics 10:132

Steinhauser S, Beckert S, Capesius I, Malek O, Knoop V (1999) Plant mitochondrial RNA editing. J Mol Evol 48:303–312

Supek F, Vlahovicek K (2004) INCA: synonymous codon usage analysis and clustering by means of self-organizing map. Bioinformatics 20:2329–2330

Tillich M, Lehwark P, Morton BR, Maier UG (2006) The evolution of chloroplast RNA editing. Mol Biol Evol 23:1912–1921

Tsuji S, Ueda K, Nishiyama T, Hasebe M, Yoshikawa S, Konagaya A, Nishiuchi T, Yamaguchi K (2007) The chloroplast genome from a lycophyte (microphyllophyte), *Selaginella uncinata*, has a unique inversion, transposition and many gene losses. J Plant Res 120:281–290

Wolf PG, Karol KG, Mandoli DF, Kuehl J, Arumuganathan K, Ellis MW, Mishler BD, Kelch DG, Olmstead RG, Boore JL (2005) The first complete chloroplast genome sequence of a lycophyte *Huperzia lucidula* (Lycopodiaceae). Gene 350:117–128

Wolfe PG, Rowe CA, Hasebe M (2004) High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern *Adiantum capillus-veneris*. Gene 339:87–89

Xia X, Xie Z (2001) DAMBE: data analysis in molecular biology and evolution. J Hered 92:371–373