Plant Ecology and Evolution 144 (1): 29–43, 2011
doi:10.5091/plecevo.2011.535

PLANT ECOLOGY EVOLUTION

REGULAR PAPER

Deep sequencing of *Ptilidium* (Ptilidiaceae) suggests evolutionary stasis in liverwort plastid genome structure

Laura L. Forrest[1,*], Norman J. Wickett[2], Cymon J. Cox[3] & Bernard Goffinet[1]

[1]University of Connecticut, Ecology and Evolutionary Biology, 75 N. Eagleville Rd., Storrs, CT, 06269-3043, U.S.A.
[2]Pennsylvania State University, Department of Biology, 403 Life Sciences Building, University Park, Pennsylvania, 16802, U.S.A.
[3]Centro de Ciências do Mar (CCMAR), Universidade do Algarve, PT-8005-139 Faro, Portugal
[*]Author for correspondence: laura.forrest@uconn.edu

Background and aims – Organellar genome sampling is patchy for non-vascular groups, with the earliest land plants poorly represented; currently only two liverworts, two mosses and one hornwort have sequenced, annotated plastid genomes. This is in part due to methodological difficulties that have hampered attempts to generate plastid genome data from liverworts. In this paper we present a method that overcomes some of the inherent difficulties by circumventing the need for plastid enrichment, but that also provides other valuable information from nuclear and mitochondrial regions including sequences from loci that may be phylogenetically useful, and potential population-level markers such as single nucleotide polymorphisms and microsatellites.
Methods – A shotgun library developed from total genomic liverwort DNA was subjected to high-throughput pyrosequencing using the Roche 454 platform. Plastid reads were bioinformatically identified, assembled and annotated. To maximize usage of the vast number of reads generated using 454 sequencing technology, combined nuclear, mitochondrial and plastid contigs were also screened for microsatellite markers, and presumed nuclear contigs were scanned for protein domains.
Key Results – This is the first plastid genome to be assembled for a leafy liverwort (i.e. *Ptilidium*) and also the first such genome to be sequenced using next generation technology for any bryophyte. The 119,007 base long plastid genome of *Ptilidium pulcherrimum* contains 88 protein-coding genes, four rRNAs and thirty tRNAs. The Inverted Repeat occurs between *trn*V-GAC and *trn*N-GUU. Functional copies of the two plastid-encoded sulphate import protein-coding genes (*cys*A and *cys*T) are absent, although pseudogenes are present in the same position that the functional genes occupy in *Marchantia*. Microsatellites: 197 novel potential primer pairs for *P. pulcherrimum* were found. Presumed nuclear *Ptilidium* contigs gave multiple hits to Class I transposable elements.
Conclusions – The arrangement of genes is identical to the plastid of the complex thalloid liverwort *Marchantia*, suggesting that structural rearrangements are rare in hepatics. This dataset represents a valuable resource for novel phylogenetic and population level marker design in hepatics.

Key words – chloroplast, liverwort, next generation sequencing, plastome, plastid genome, *Ptilidium*, 454 technology.

INTRODUCTION

Embryophyte plastid genomes are, in general, conserved in structure – from 110 to 130 unique (unduplicated) genes arranged across a circular structure comprising two Inverted Repeats (IR) separating a Large Single-Copy (LSC) region and a Small Single-Copy (SSC) region (Raubeson & Jansen 2005). Changes that are found to occur in genome structure (e.g. changes in gene order, gene duplication and loss) are thought to be relatively rare. Plastid genome rearrangements have proved informative in many groups including embryophytes (Kelch et al. 2004, Mishler & Kelch 2009), vascular plants (Raubeson & Jansen 1992), ferns (Stein et al. 1992), Asteraceae (Jansen & Palmer 1987) and recently in mosses (Goffinet et al. 2005, 2007); rearrangements in mitochondrial gene order have also provided supporting information for liverwort relationships inferred from DNA sequence data (e.g. Wahrmund et al. 2008, Knoop 2010).

One of the first plastid genomes, from the complex thalloid liverwort *Marchantia polymorpha* L., was mapped in

1986 (Ohyama et al. 1986) and sequenced in 1988 (Ohyama et al. 1988). The second liverwort plastid to be sequenced was assembled twenty years later: the simple thalloid liverwort *Aneura mirabilis* (Malmb.) Wickett & Goffinet (Wickett et al. 2008b). This species is non-photosynthetic and the only non-vascular plant known to have adopted a parasitic lifestyle (Schuster 1992, Read et al. 2000, Brundrett 2002, Bidartondo 2005). The gene order in these two liverwort plastid genomes is virtually identical, the one difference being a two-gene inversion (*psb*E and *pet*L) unique to *Aneura mirabilis* among all liverworts surveyed, including other photosynthetic members of the genus *Aneura* Dumort. (Wickett et al. 2008a). Several unique characteristics of the *A. mirabilis* genome appear to correlate with the loss of photosynthesis. Although its annotated plastid genome provides a unique opportunity to independently test hypotheses of genomic evolution following a shift to heterotrophy in a bryophyte, it is not an optimal model for comparing plastid genome structure in photosynthetic organisms.

Another difference between the plastid genomes of *Marchantia* L. and *Aneura* is the absence of functional copies of the plastid-encoded sulphate import protein-coding genes (*cys*A and *cys*T) in the simple thalloid. Gene loss can be a powerful phylogenetic marker, as has been shown in mosses, where transfer of the *rpo*A gene to the nuclear genome denotes a clade comprising nearly 90% of moss species (Sugita et al. 2004, Goffinet et al. 2005). However, a wider taxonomic survey for presence/loss of the *cys*A and *cys*T genes shows that these have most likely been lost multiple times in hepatics (Wickett et al., unpubl. res.).

Currently our ability to survey for plastid gene order rearrangements and losses is limited: few bryophyte plastid genomes are available, and lineage sampling is extremely poor. Apart from the two liverworts, only two mosses – the widely used model organism *Physcomitrella patens* (Hedw.) Bruch & Schimp. (Sugiura et al. 2003) and the desiccation tolerance model *Syntrichia* (*Tortula*) *ruralis* Weber & D.Mohr (Oliver et al. 2010) – and one hornwort, *Anthoceros angustus* Steph. (published under the synonym *A. formosae* Steph.; Kugita et al. 2003), have assembled and annotated plastid genomes. In addition to a 71 kb inversion in the *Physcomitrella* Bruch. & Schimp LSC that has been shown to be a synapomorphy for the Funariales and Encalyptales (Goffinet et al. 2007), only one other structural change, the deletion of *pet*N from *Tortula* Hedw. (Oliver et al. 2010), has been identified in mosses.

The twenty-year time lag between sequencing the first and second bryophyte plastid genome is remarkable, but recently, critical advances in organellar sequencing techniques have been made. Many studies use methods that purify plastids from other cell fractions prior to sequencing (usually after cloning the plastid DNA into vectors) [e.g. Ohyama et al. 1988 (*Marchantia*), Maier et al. 1995 (*Zea* L.), Kugita et al. 2003 (*Anthoceros* L.), Wolf et al. 2005 (*Huperzia* Bernh.), Ravi et al. 2006 (*Morus* L.), Raubeson et al. 2007 (*Nuphar* Smith and *Ranunculus* L.), Tsuji et al. 2007 (*Selaginella* P.Beauv.), Wu et al. 2007 (*Cycas* L.), Oliver et al. 2010 (*Tortula)*]. However, *a priori* separation of the plastid fraction requires a larger amount of starting material than a simple total genomic DNA extraction. In some cases, particularly with small and rare plants, obtaining this living plant material is the rate-limiting step for the entire process. Recently, Wickett et al. (2008b; *Aneura*) cloned fragments from total genomic extracts, creating a fosmid library that was then screened for plastid DNA inserts using PCR probes from well-characterized plastid loci (McNeal et al. 2006). Clones that matched plastid probes were subjected to shotgun Sanger sequencing. Goremykin et al. (2003; *Calycanthus* L.) and Gao et al. (2009; *Alsophila* R.Br.), on the other hand, extracted total genomic DNA then used several targeted long-range PCRs to amplify the whole plastid genome prior to cloning. Today, laboratories are increasingly utilizing the power of next generation pyrosequencing technology. For example, Moore et al. (2006; *Nandina* Thunb. and *Platanus* L.) purified plastids using a sucrose gradient then subjected these to shotgun methods using the Roche 454 platform. Cronn et al. (2008; *Pinus* L. and *Picea* Link.) extracted total DNA then used long-range PCRs to amplify the plastid genome before performing amplicon sequencing using an Illumina Genome Analyzer. Donaher et al. (2009; *Cryptomonas* Ehrenb.) also enriched the plastid component of their extractions, using a caesium-chloride density gradient prior to 454 sequencing.

Liverworts are small plants, and extracting large amounts of high quality DNA is not always feasible. Consequently, plastid isolation methods have proven unsuccessful for a broad sampling of liverworts. We also attempted to construct large insert libraries (40kb fosmid) for four phylogenetically disparate species, but multiple attempts per species produced no clones with plastid inserts. Instead we have chosen to utilize the vast over-sequencing capability of next generation platforms by sequencing fragments from all three plant genomes without any enrichment, and screening the total sequencing products for plastid genome homologues. The method consists of sequencing random fragments from a total DNA isolation. Subsequently, the plastid genome sequences are bioinformatically sorted and assembled. One of the most valuable resources in any investigation is time, and replacing pre-sequencing enrichment, primer design or long-range PCR optimization with bioinformatic pipelines allows more rapid access to sequence data. Additionally, this method provides data that can be mined for other resources, for example microsatellite repeat regions, nuclear genes and mitochondrial genomes. A similar approach is being followed in other labs around the world (e.g. Tangphatsornruang et al. 2010, Wolf et al. 2010), and provides a sound alternative in cases where pre-sequencing plastid enrichment is difficult or impossible.

A well-sampled comparative matrix of organellar genomes is lacking for any non-vascular group. To address this, our objective is to sequence, assemble and annotate entire plastid genomes from all major lineages of liverworts. These genome data should enable inference of phylogenetic relationships among liverworts based both on structural changes and DNA sequence data.

Liverworts are an ancient group: their origin may mark the successful transition of plants to land (Qiu et al. 2006) and the beginning of the diversification of a land flora. They comprise several morphologically distinct lineages. The c. 8,500 extant species (von Konrat et al. 2010) are divided by their vegetative architecture into ‘simple thalloid’, ‘complex thalloid’, and ‘leafy’ forms. Although molecular phylogenies based on nucleotide sequence data (e.g. Forrest & Crandall-

Stotler 2005, Forrest et al. 2006, Heinrichs et al. 2005) have resolved relationships among thalloid forms, the early branching order of major lineages within leafy clades remains controversial (e.g. the position of *Ptilidium* Nees relative to other leafy liverworts). As analyses of sequence data have not yet resolved this question, analysis of plastid genome rearrangement data offers an alternative approach.

To test our technique we have sequenced the plastid genome of the leafy liverwort *Ptilidium*, which is representative of a lineage that diverged from *Marchantia* c. 370 million years ago (mya; Heinrichs et al. 2007). We applied whole genome next generation sequencing, to allow us to 1) assess the efficiency of this technique in recovering complete plastid genomes, 2) identify structural markers of potential phylogenetic utility and 3) explore the large set of sequences obtained for additional data sources such as microsatellites.

MATERIALS AND METHODS

Plant material

Ptilidium, a small genus of only three species (Schuster 1966), occupies a somewhat basal position in the leafy lineages (Forrest et al. 2006). Our targeted species, *P. pulcherrimum* (Weber) Vain., has a holarctic distribution (Schuster 1966). It was collected on a granite boulder in Vermont (N44.99620, W71.69815) by Blanka Shaw (collection no. 6968, 19 May 2008), and sent to the University of Connecticut as living material.

Molecular methods

Total genomic DNA was extracted from c. 2 g fresh gametophyte homogenized with pestle and mortar in liquid nitrogen. Using a modification of the CTAB protocol from McNeal et al. (2006), the powder was incubated for c. 1 hour at 70°C in c. 3 mL extraction buffer (consisting of 1.21 g Tris-HCl, 0.74 g EDTA, 8.18 g sodium chloride, 2 g CTAB and 1 g PEG8000 in 100 mL water) with 30 µL β-mercaptoethanol. Two SEVAG (24:1 chloroform:isoamyl alcohol) extraction steps were performed, then the DNA was precipitated with isopropanol overnight and pelleted by centrifugation. The pellet was washed with 70% ethanol, and suspended in 500 µL TE buffer. A second precipitation step using 1/10 volume 3 M sodium acetate (pH 7.0) and 2.5 volumes freezer-cold 100% ethanol was performed to maximize DNA purity. The resulting pellet was suspended in 165 µL TE. The DNA quality was ascertained using a Nanodrop to obtain Optical Density (OD) 260:280 ratios and DNA was also run on a 0.8% agarose TBE gel stained with SYBRsafe DNA gel stain (Invitrogen, Oregon, U.S.A.), using Invitrogen's High Molecular Weight DNA mass ladder, to check that it was not degraded. Invitrogen's Qubit fluorometer system with a Quant-iT ™ dsDNA BR Assay was used to estimate DNA quantity. Eight µg of DNA were then sent to the IGSP Sequencing Core Facility (119 Biological Sciences Building, Science Drive, Durham, NC 27708) at Duke University for Roche FLX 454 shotgun library preparation using the Titanium system, and the library was run on 3/8 of a PicoTitre Plate (PTP) on a Genome Sequencer FLX Instrument.

Where required for gap filling, Sanger sequences were generated on an ABI 377 capillary system using Applied Biosystems BigDye version 1.3 chemistry (following the manufacturer's protocols except using 10 µL total reaction volume with 0.85 µL BigDye and 2 µL ABI buffer per reaction), using newly designed primers specific to *P. pulcherrimum*.

Plastid assembly, annotation and analyses

The sequences generated were trimmed of adaptor and low quality regions using the sff_extract python application distributed with the MIRA software package. The trimmed sequences were then assembled into contigs using all available reads, with MIRA 2.9.43 (Chevreux et al. 1999) (using commands –job = denovo,genome,accurate,454 -SK:mnr = yes:nrr = 20 -AS:ugpf = no 454_SETTINGS -AS:mrl = 50); BLASTn searches against inferred plastid coding sequences of *Marchantia polymorpha*, *Aneura mirabilis* and *Physcomitrella patens* were used to select the plastid sequences, with an e-value cutoff of e-5. Large contigs were annotated using DOGMA (Wyman et al. 2004) and subsequently joined, using custom perl scripts and CAP3 (Huang & Madan 1999) to iteratively blast the plastid contigs to the contig and read databases, thus elongating the ends of the plastid contigs. Manual editing and joining of the remaining contigs was performed using Sequencher 4.9 (Gene Codes Corporation). Finally, primers were designed using Primer3PLUS (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) in order to verify ambiguous regions using Sanger sequencing. Sanger sequence was generated for over 15,500 bases from fifteen regions, using a total of 39 primers.

The genome was annotated using DOGMA (Wyman et al. 2004; http://dogma.ccbb.utexas.edu/) with 60% cutoff for protein coding genes; 80% cutoff for RNAs; E-value $1e^{-5}$; returning fifteen blast hits; gapped alignment; plant plastid genetic code. Because the fifteen plastid genomes that DOGMA uses for comparison of input genomes only contain one bryophyte, the annotation was tested against coding sequences from all available bryophyte plastids in NCBI GenBank, as follows. *Ptilidium* protein coding nucleotide sequences were extracted from the DOGMA annotation and imported into Sequencher 4.9. Coding regions from the five other bryophytes available on NCBI GenBank on 25 Nov. 2009 (*Aneura* NC_010359, *Anthoceros* NC_004543, *Marchantia* NC_001319, *Physcomitrella* NC_005087, *Syntrichia* Brid. NC_012052) were extracted from the Entrez Genome database, imported to Sequencher, and assembled by locus name with the *Ptilidium* sequences to compare positions of start and stop codons. Probably as a result of high sequence divergence from the other available taxa, the *mat*K gene was not annotated in DOGMA, although a large open reading frame (ORF) was present in the *trn*K-UUU intron. This and other minor corrections to the annotation (e.g. expanding genes to include 3' stop codons, and linking exons) were made manually in DOGMA. The assembled and annotated *Ptilidium* plastid genome is available from GenBank, reference number HM222519. The raw 454 reads have also been deposited in the NCBI sequence read archive (SRA024244.1/Ptilidium_454 Study). A circular gene map of the *Ptilidium*

plastid genome was drawn in OrganellarGenomeDRAW (Lohse et al. 2007) (fig. 1).

The completed *Ptilidium* plastid genome (with IR^B removed, because the presence of a large duplicate region causes problems for the mapping algorithm) was input as a reference genome to GS Mapper (Roche). All 454 reads were then mapped to the plastid genome using default mapping parameters, as an indication of depth of cover. Another assembly of all 235,791 reads was also run using GS de novo Assembler (Roche), with default parameters for a large or complex genome, to estimate the inferred read error for the 454 sequences.

The plastid sequence (with IR^B removed to simplify results; it is already defined as a large repeat) was run through the web-based interface of REPuter (Kurtz & Schleiermacher 1999; http://bibiserv.techfak.uni-bielefeld.de/reputer/submission.html) in order to get a rapid representation of the position of exact short dispersed repeats of eight bases or more throughout the genome. Forward, reverse, complement and reverse complements were assessed separately. The maximum computed repeat number via this interface is 5,000.

Nuclear data mining

Infernal 1.0.2 (Nawrocki et al. 2009) was used to search the raw reads for small subunit ribosomal RNA (‘5 domain) using the Rfam9.1 rf00177 model, in order to check for non-*Ptilidium* contamination, with the default settings. An NCBI BLASTx search (e^{-5}) was used to query the contigs assembled using MIRA against bryophyte mitochondrion and plastid genomes to identify probable nuclear contigs; these nuclear contigs were then scanned with InterProScan (The InterPro Consortium 1999, Quevillon et al. 2005, Mulder et al. 2007) against databases HMMPanther (Hidden Markov Model Protein ANalysis THrough Evolutionary Relationships; classification of genes by functions) and HMMSmart (Hidden Markov Model Simple Modular Architecture Research Tool) to search for protein domains. Singletons were not scanned, although it is possible that some of the longer reads may include identifiable ORFs.

Microsatellite markers

In order to identify microsatellite sequence regions, the total 454 reads, the assembled contigs and the assembled plastid sequence were run separately through msatCommander 0.8.2 (Faircloth 2008), set to automatically design primers using primer3 (Rozen & Skaletsky 2000) as its primer design engine, accepting mononucleotide repeats of ten or more, dinucleotide repeats of six or more, trinucleotide repeats of four or more, tetranucleotide repeats of four or more, pentanucleotide repeats of four or more and hexanucleotide repeats of four or more.

RESULTS

DNA quality

The CTAB-extracted *Ptilidium* DNA had a 260:280 ratio of 2.03, and a 260:230 ratio of 1.81. On an agarose gel it formed a clear high mass band at c. 12,000 bases, with a small amount of smearing that may have been due to some DNA degradation. The final concentration of DNA, estimated by fluorometry, was 51.9 µg/mL.

454 data

DNA sequences were generated from the nuclear, mitochondrial and plastid genomes. 235,791 sequences (mean read length 335 base pairs [bp], median read length 384 bp) were generated, comprising 79,052,599 bp of sequence data. Using MIRA, 17,308 contigs were generated, ranging in length from 50 to 21,948 bp with mean length 494.04 bp, standard deviation 509.25 and median length 404.00 bp. In total, 8.55 megabases of sequence were assembled.

The assembly produced by GS de novo Assembler used 115,213 reads, comprising 35,574,046 bp, i.e. c. 49% of the total reads, and c. 46% of the total base pairs. The assembly had an inferred error rate of 3.16%. GS de novo Assembler aligned the raw sequence reads into 12,843 contigs; these ranged in length from 100 (set as the minimum in GS Assembler) to 41,192 bp, with 478 contigs over 500 bp and 77 contigs over 2,000 bp long. The average large contig size was 1,718 bp (N50 = 1,602 bp; 94% of aligned bases with quality scores of Q40 plus). NCBI BLASTn analyses of select contigs showed that they belonged to all three genomic compartments. As GS de novo Assembler was not available to us during initial data assembly, the contigs generated by it were not utilized in the plastid genome assembly.

Plastid assembly, annotation and analyses

When the sequencing products were screened against known plastid genome sequences to isolate the fraction of 454 sequencing products belonging to the *Ptilidium* plastid genome, approximately 29% of the MIRA assembled contigs matched regions of the plastid genome. These contigs were further assembled into a near-complete circular plastid genome of c.119,000 bases. The sequence of each locus was drawn from the consensus of an average of ten to thirty overlapping sequence reads. Although many ambiguous regions occurred, these were limited to locations rich in polynucleotide repeats. Unless an obvious frame-shift was present that easily restored an expected ORF and did not disagree with the 454 ACE files, in which case the correction was made immediately, primers were designed using the 454 sequence assembly and the regions were sequenced using standard Sanger methods (table 1). A solitary 1,200 base gap in the plastid assembly between IR^A and the SSC, which corresponded to a region with low 454 coverage, was also filled using Sanger sequencing.

Most disagreements in the initial assembly of Sanger sequences with 454 data were due to low quality reads at the starts and ends of the Sanger sequences. After trimming these end regions from the Sanger files, the majority of errors that remained were due to 454 misreads in the numbers of ‘A’ or ‘T’ in mononucleotide runs (the Sanger length of which varied from three to ten nucleotides). G and C mononucleotide repeats are relatively infrequent in the dataset, so the observed bias towards errors in A and T repeats may not be purely due to the known problem for 454 in reading through A and T homopolymers (e.g. Wicker et al. 2006). Only one

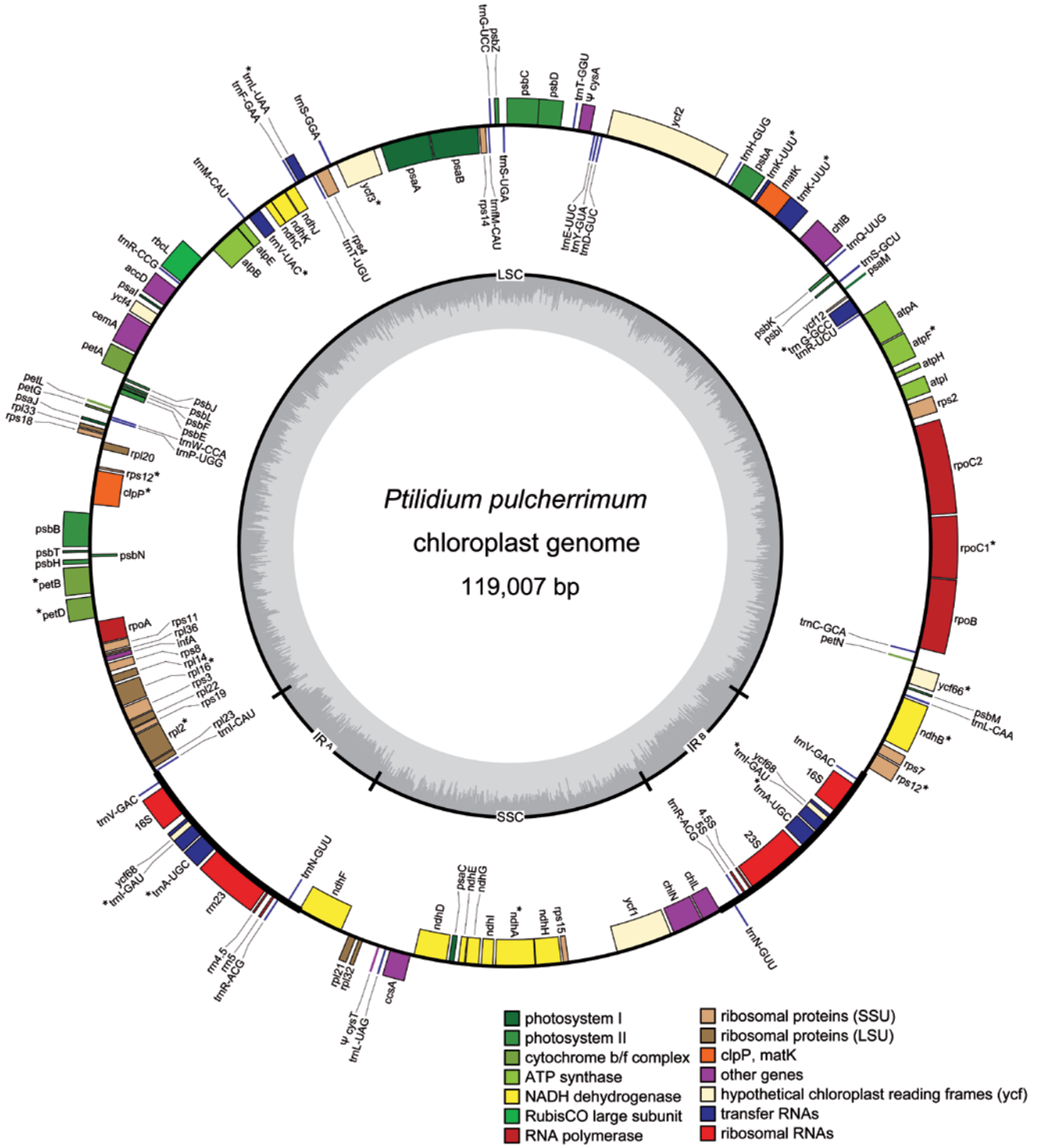

Figure 1 – Plastid genome map of *Ptilidium pulcherrimum*. Genes on the outside of the circle are transcribed in the counterclockwise direction, and genes on the inside of the circle are transcribed in the clockwise direction. Structural components are labeled on the inner circle as LSC and SSC regions, IR^A and IR^B. Inner graph charts % GC composition (dark) across the genome. Asterisks denote split genes. Pseudogenes are notated with a Ψ.

seven-nucleotide G plastid repeat was miscalled, as six nucleotides, in the 454 data.

The completed *Ptilidium* plastid genome is 119,007 bases long, with 109,923 unique base pairs, 9,084 of which make up the IR. The IR occurs between *trn*V-GAC and *trn*N-GUU. The plastome encodes 122 genes, excluding the second IR region, corresponding to 88 protein-coding genes, four rRNAs, and 30 tRNAs (table 2). The status of three poten-

Table 1 – Sanger sequences generated for *Ptilidium* plastid genome.
Additional sequence data required to verify 454 sequence or close gaps in chloroplast assembly.

region	Sanger sequence (bp)	number of primers	reason Sanger sequencing required	problems identified in 454 assembly
*ndh*B	1013	2	stop codons in ORF	3 polyT runs too short; 1 polyA run too long
*cem*A	1409	2	stop codons in ORF	2 polyT runs too short; 4 polyA runs too short
*chl*N	700	2	low coverage	none
*ycf*68	787	2	stop codons in ORF	1 polyA run too short
*rpo*B	718	2	stop codons in ORF	1 polyG run too short; 1 polyT run too short; 1 polyA run too short
*rpo*C2	2360	4	stop codons in ORF	2 polyT runs too long; I polyT run too short; 5 polyA runs too long; 1 polyA run too short
*ycf*2	1618	2	stop codons in ORF	1 polyC run too long; 3 polyA runs too long; 1 polyT run too long; 1 polyA run too short; 4 polyT runs too short
*ndh*F	482	1	stop codons in ORF	1 polyA run too short
*ycf*1	3819	10	stop codons in ORF	21 polyT runs too short; 2 polyA runs too short; 2 polyT runs too long; 2 polyA runs too long
*chl*B (contigs 10-36)	576	2	low coverage	1 extra C added into 454 sequence
*ycf*2 – *trn*Y-GUA (contigs 11-42)	509	2	low coverage	1 small misassembly corrected
*trn*V-GAC – rrn16 (contigs 65-14)	1556	4	low coverage	1 polyT run too long
*ycf*3 – *psa*A (contigs 7–80)	393	2	low coverage	1 polyT run too short; 5 polyA runs too long; 1 A inserted in wrong place
*psa*B (contigs 80–11)	674	2	low coverage	none
IR^A – SSC	1235	6	gap in assembly	none (but most of region had no 454 sequence)
Total	**17,849**	**45**		

tial protein-coding genes is ambiguous – *mat*K in the LSC, *ycf*1 in the SSC and *ycf*68 in the IR all contain interruptions to their ORF. The single stop codon in the short *ycf*68 locus has been verified with Sanger sequencing. Functional copies of two plastid-encoded sulphate import protein-coding genes (*cys*A and *cys*T) present in *Marchantia* are absent from the *Ptilidium* plastome.

Seventeen genes contain introns splitting them in two exons or in the case of *clp*P, *rps*12 and *ycf*3 in three exons. Two of these intron-containing genes are in the SSC region, while the rest are in the LSC region. Sixty-four genes are transcribed in the sense direction (fifty LSC, nine IR^A, four SSC, one IR^B), and 68 are antisense (46 LSC, one IR^A, nine SSC, nine IR^B) (fig. 1).

The GC content of the entire plastome is 33.2% (32.1% with only a single copy of the IR), while that of the IR alone is 47.1% due to elevated GC content in the four ribosomal genes (which together comprise 4,527 bases, 53.7% GC, in each IR segment). The shortest intergenic spacer (between *ndh*A and *ndh*H) is one base pair, while the longest, between *rps*15 and the *ycf*1 ORF, is 2,216 bases long. The reading frames of two genes, *psb*D and *psb*C, overlap – thus they have no intergenic spacer. In total the intergenic spacers cover 25,101 bases, with a GC content of 25.7%. Gene introns vary in length from 69 bases in *trn*I-GAU to 2,099 bases in *trn*K-UUU. In total (but excluding sequence of the *mat*K gene region annotated in the *trn*K-UUU intron) they cover 13,257 bases, with a GC content of 32%. The tRNAs are, in total, 2,708 bases long, with a GC content of 52.8%. Including two annotated pseudogenes (*cys*A and *cys*T), a total of 39,013 bases of the plastome, or nearly 33%, appear to represent non-coding sequence.

Using GS Mapper, a total of 3,133,832 bases (from 9,733 reads) of the total genomic 454 data mapped to the *Ptilidium* plastid genome, meaning that each base was read on average 28.3 times. The plastid-mapping base pairs represent 3.96% of the total bases (leaving 75,918,767 bp that did not contribute to the plastid assembly, and thus presumably represent mitochondrial and nuclear sequence data, as well as any contaminant reads).

Twenty-nine simple sequence repeats (microsatellites) were identified in the *Ptilidium* plastid genome (with the second IR segment excluded) by msatCommander. Sixteen of these were mononucleotide repeats of ten bases or more (all

Table 2 – Gene distribution by plastid region in *Ptilidium*.

	LSC	IR	SSC	Total
ORF	72 (including *mat*K)	1 (*ycf*68)	15 (including *ycf*1)	88
tRNA	24	5	1	30
rRNA	0	4	0	4
Total	96	10	16	122

either A or T), ten were dinucleotide repeats with six to ten copies (all AT), two were trinucleotide repeats with four copies (ATT and AAT), and one was a tetranucleotide repeat with four copies (ATTT).

The *Ptilidium* plastid genome contains over 5,000 of each Short Dispersed Repeat (SDR) type (forward, reverse, complement and reverse-complement); the forward repeats range from 13–40 bases, reverse from 12–37 bases, complement from 12–37 bases, and reverse-complement from 12–82 bases. Most of the longest repeats map to either the same, or very similar, regions of the plastid – e.g. the longest repeat, an 82 base reverse-complement, is a palindrome that occurs between *psa*A and *ycf*3; the longest forward repeat (forty bases) occurs in the spacer between *pet*N and *trn*C-GCA, and is repeated only two bases upstream – it comprises an AT-repeating microsatellite region. However, some multiple repeats map to physically separate regions of the genome; e.g. a 37 base forward repeat that is shared between *psa*B (starting at 43,294) and *psa*A (starting at 45,519); a 25 base reverse repeat that is shared between the *pet*D-*rpo*A spacer (starting at 73,506) and the *ndh*H-*rps*15 spacer (starting at 102,366); a 21 base complement repeat that is shared between *psb*C (starting at 39,868 bases) and the *trn*M-CAU-*atp*E spacer (starting at 53,522 bases).

Nuclear data mining

Using total genomic DNA means that in addition to plastid loci we sequenced templates drawn from the nuclear and mitochondrial genomes. For *Ptilidium*, nearly 76 million bases of the unassembled sequence data do not match plastid loci. These data were screened against known genomes. Only a small proportion of the overlapping sequences matched mitochondrial genomes, although this may still allow for the near-complete assembly of a mitochondrial genome from the data.

Searches using Infernal resulted in 122 hits for the small subunit ribosomal RNA gene. Using BLAST against the GenBank nucleotide (nt) database showed that 95 (78%) of the 122 hits were from *Ptilidium*, while eleven (9%) were from contaminating genomes (bacteria: three probacterium, two planctomycetes, one firmicute, one unknown bacterium; protists: one cercozoan; animals: one tardigrade; fungi: two ascomycetes).

InterProScan identified 40,885 open reading frames in the purportedly nuclear contigs. By default InterProScan (using EMBOSS sixpack) identifies all ORFs regardless of whether they start with methionine. Because we are dealing with incomplete fractions of the genome, our contigs do not necessarily contain the start codon, however, some identified domains may not be expressed genes (e.g. they may be interrupted up- or down-stream).

InterProScan found 2,517 hits to contigs, and 2,648 hits to 614 different domains. Twenty-five genes had ten or more hits (table 3), while 361 had only a single hit (electronic appendix 1).

Microsatellite markers

The pool of 454 sequence data was also screened for potential microsatellite loci for *Ptilidium pulcherrimum*. The set of random 454 sequences obtained for *Ptilidium* contains numerous microsatellite loci (table 4). Screening revealed many mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats. In the raw data 4,260 *Ptilidium* reads contain microsatellite regions, representing 1.81% of the total 454 reads; primer sequences flanking microsatellite regions were successfully generated from 1430 *Ptilidium* reads. In the assembled data, however, 528 contigs contained microsatellite regions, representing 2.05% of the total contigs; primer sequences flanking microsatellite regions were successfully generated for 197 *Ptilidium* loci (electronic appendix 2). Primer pairs can be preferentially selected from these according to the desired characteristics of the microsatellites (e.g. tetra- and pentanucleotide repeats for an amplicon size of 200–300 base pairs).

DISCUSSION

The *Ptilidium* plastidome

Gene order conservation – The architecture of the *Ptilidium* genome is remarkable, given the amount of time since divergence from a common ancestor, in that all genes occur in the exact same order and location as in the *Marchantia* plastome. No gene has changed position or orientation – even remnants of the two pseudogenes *cys*A and *cys*T occur in the same location as their ORFs in *Marchantia*. Few studies have attempted to estimate divergence times within the liverworts using molecular dating; the most comprehensive of these, Heinrichs et al. (2007), estimated the most recent common ancestor for *Ptilidium* and *Marchantia* to have occurred 372.6 mya (+/- 10.2 million years). However, the discovery in 2008 of a simple thalloid liverwort from the mid to late Devonian (c. 392–385 mya) that now represents the earliest hepatic fossil (VanAller Hernick et al. 2008) is likely to push the origin of major liverwort groups even deeper into the past. The contrast between the lack of plastid gene order rearrangements in taxa that probably diverged almost 400 mya, and the relative frequency of gene translocations in some angiosperm lineages is noteworthy. Within the Campanulaceae, for example, a family that diversified less than 40 mya (Roquet et al. 2009) and has the commonplace LSC/SSC/two IR plastid structure, gene order rearrangements are so rampant that researchers struggle to define individual events (Cosner et al. 2004). Likewise, Geraniaceae plastomes are highly variable in size, gene content and order (Guisinger et al. 2010) although the basal split in the family appears to have occurred only 50 mya (Fiz et al. 2008). At the other end of the plant evolutionary spectrum, algal plastids also show great variation in genome arrangement (e.g. Pyke 2009).

Table 3 – **Nuclear genes identified in *Ptilidium* data.**
Genes with at least ten hits from scan of 28,874 unknown 454 contigs (16,580,932 bases, presumably mostly nuclear *Ptilidium pulcherrimum*) against HMMPanther and HHMSmart databases using InterProScan.

protein	number of hits in contigs	percentage of total hits to domains
GAG/POL/ENV polyprotein	767	28.97%
GAG-POL-related retrotransposon	268	10.12%
reverse transcriptases	97	3.66%
HMMSmart: Zinc finger CCHC-type	43	1.62%
ATP binding cassette (ABC) transporter	33	1.25%
sensor histidine kinase-related	31	1.17%
helicase-related	22	0.83%
short-chain dehydrogenases/reductases family member	20	0.76%
modification methylase	18	0.68%
HMMSmart: Chromo domain	15	0.57%
sentrin/sumo-specific protease	14	0.53%
HMMSmart: Zinc finger PHD-type	14	0.53%
HMMSmart: NLI interacting factor	14	0.53%
ATP-binding cassette transporter	14	0.53%
cytochrome P450	13	0.49%
chromobox protein	13	0.49%
ATP-dependent AMP-binding enzyme family member	13	0.49%
NADH dehydrogenase	12	0.45%
chaperonin	12	0.45%
alcohol dehydrogenase related	12	0.45%
zinc finger CCHC domain containing protein	11	0.42%
heat shock protein 70KDA	11	0.42%
sugar transporter	10	0.38%
ATP-dependent CLP protease	10	0.38%

Repeat regions have previously been hypothesized to represent hotspots for non-homologous recombination within the plastid, and thus to lead to genome rearrangements (inversions and transpositions) (e.g. Cosner et al. 2004, Haberle et al. 2008, Lee et al. 2007, Milligan et al. 1989), so it is possible that a genome depauperate in SDRs would have a more stable architecture. Instead, we found that SDRs are present throughout the *Ptilidium* plastid genome. However, the repeats hypothesized to play a part in structural rearrangements within the *Trifolium subterraneum* plastid are over 500 bases long (Milligan et al. 1989), while those identified in the *Ptilidium* plastid have a maximum length of 82 bases. Furthermore, many of the *Ptilidium* repeats occur close together in the same non-coding regions; thus rearrangements between them, involving only spacer or intron sequence, would not be apparent based on gene annotations. Some of the other repeats occur within genes at one or both sites; rearrangements between these are unlikely to be retained as they would disrupt the involved genes. Dispersed repeats with both ends in spacer or intron regions are far less common. However, they do exist, and so the absolute lack of plastid rearrangement observed between *Ptilidium* and *Marchantia* cannot be attributed entirely to the absence of short dispersed repeats in the genome, but must have another explanation.

The conserved plastid gene content and order in liverworts may be influenced by low rates of molecular change in bryophyte plastid genes relative to vascular plants [as shown by Stenøien (2008) for coding regions in mosses]. Whether this observed low rate of nucleotide substitution is due, for example, to intense purifying selection on changes to genes in a haploid-dominant organism, or to more efficient DNA repair mechanisms (as may be expected in plants with little protection from UV radiation and/or that are likely to experience – and recover from – the cellular damage caused by severe desiccation events), or is a function of high organellar genome copy numbers increasing overall organelle population size, is unknown. However, a shortcoming of considering lower rates of molecular change as a factor in plastid gene order stability is that these are nucleotide substitution rates calculated using gene sequences. Calculation of rates of molecular change based on non-coding regions would be hampered because sequence homology in alignments of spacer regions often becomes impossible to assess as phylogenetic distances increase. Furthermore the situation in mitochondrial genomes suggests that gene sequence stability and gene order lability are unrelated: despite extremely reduced levels of molecular change within genes, mitochondrial genomes have frequent genomic rearrangements (Knoop 2004).

Table 4 – **Microsatellite repeats.**
Repeats found in the 454 sequence data set, and the number of microsatellite primer pairs successfully designed, using the program msatCommander.

dataset	reads/contigs	mono-	di-	tri-	tetra-	pent-	hex-	number of primer pairs designed
unassembled	235,791	3669	2185	848	118	52	34	1430
assembled	25,700	62	181	214	53	11	6	197

Another potential stabilizing factor for plastid arrangement is maintenance of polycistronic transcription, where a group of genes are regulated together in an operon. This is a feature of the plastid genome (e.g. Monde et al. 2000, Woodbury et al. 1988) that reduces the number of transcriptional regulation units required. Membership of these co-transcribed gene clusters may be positively selected, for example to include components of a functional pathway (Kallas et al. 1988); evidence for this is also seen in the unicellular green alga *Chlamydomonas* (Cui et al. 2006, Maul et al. 2002) where plastid genome organization is a product of directional selection. Any gene rearrangement causing the transcription direction of some genes to change relative to others in a co-transcribed cluster disrupts the operon. Thus inversion of any parts of an operon results in changes to the transcriptional units. Given that inversion is widely accepted to be the most common cause of plastid rearrangement (e.g. Cui et al. 2006) and that different kinds and degrees of RNA processing occur in different plant lineages (Pyke 2009, Stern et al. 2010), it may be that selection for maintenance of a precise set of polycistronic transcription operons is a stronger stabilizing factor in liverwort plastid organization than in other land plant lineages.

Data generation and quality

Using next generation sequence data has allowed us to generate a relatively fast and inexpensive plastid genome sequence. However, the genome is highly AT biased, and contains many polyA and polyT repeats – sequence motifs that current 454 technology has difficulty in determining accurately (e.g. Moore et al. 2006, Wicker et al. 2006). Errors in polynucleotide repeat size were not always limited to long repeats – Sanger sequencing showed that occasionally repeats as short as three bases were miscalled. In addition, errors in ORFs were apparent during genome annotation, adding considerably to the amount of manual input required. Sequencing (or assembly) errors in non-coding regions of the genome are far less likely to have been identified and corrected, as they do not cause easily identifiable problems with the annotation of the plastid sequence. Reassuringly, however, most miscalls relate to longer polyA or polyT repeats, and so bases (or regions) that may be problematic can be quickly identified.

One confounding factor is that we harvested multiple *Ptilidium* individuals from a single patch of wild-growing plants. Although it is often assumed that monospecific bryophyte clumps are largely clonal, this has not been tested in this species, and thus the data we have generated may contain sequences from a pool of individuals rather than a single clone. This could account for some (but not all) observed variation in mononucleotide repeat lengths between reads, particularly in non-coding regions. It is also likely that many point mutations, found particularly in contigs with deeper coverage, represent real population level genetic variation rather than sequencing errors – suggesting that single nucleotide polymorphism (SNP) genotyping would be rewarding in this liverwort.

Generating a plastid genome sequence for a group of individuals rather than for a single clone is certainly not unique to this study. No other published bryophyte plastid genome [*Aneura* (Wickett et al. 2008b), *Anthoceros* (DNA extraction details in Yoshinaga et al. 1996), *Marchantia* (DNA extraction details in Ohyama et al. 1988), *Physcomitrella* (Sugiura et al. 2003), *Tortula* (Oliver et al. 2010)] or mitochondrial genome [*Marchantia* (Oda et al. 1992), *Nothoceros* (R.M.Schust.) Hasegawa (Li et al. 2009), *Phaeoceros* Prosk. (Xue et al. 2009), *Physcomitrella* (Terasawa et al. 2006), *Pleurozia* (Wang et al. 2009)] uses DNA from plants cultured from a single spore isolate. The issue of sampling from multiple individuals is not often addressed, and indeed methods that involve sequencing large plastid inserts from fosmid clones may mask much intra-individual sequence variation. However, 29 polymorphisms were identified in the *Tortula* plastid genome (Oliver et al. 2010), which was sequenced from plants grown from multiple sporophytes harvested from a single location.

Annotation – A conserved motif that has been annotated as a hypothetical gene in many plastids, *ycf*68, occurs in the IR between *trn*I-GAU and the *rrn*16 locus. In *Ptilidium*, although the sequence seems highly conserved when compared to other bryophytes, it contains a stop codon. The *ycf*68 motif has been reported in several vascular plants, but it is non-functional in many lineages, and Raubeson et al. (2007) surmise that it is probably conserved simply because it occurs in the IR region, which is the slowest evolving region of the plastome, and may not code for a functional gene at all. However, the occurrence of RNA editing in the plastid genome of many plant lineages, including simple thalloid and leafy liverworts (Freyer et al 1997; reviewed in Schmitz-Linneweber & Barkan 2007), provides a complicating factor in assessing the presence/absence of pseudogenes based on missing start codons, premature stop codons, or small insertions or deletions to reading frames, as all these may be edited post-transcription to yield a fully functional gene.

Premature stop codons that were initially identified within two of the largest plastid open reading frames, *ycf*1 and *ycf*2, were shown using Sanger sequencing to be caused by frame shifts due to miscalling of polyN regions. Both regions have been determined to be essential genes in higher plants by plastid transformation experiments (Drescher et al. 2000), although they are not ubiquitous in land plant plastids: they are not present in the plastid genomes of rice and maize (Hiratsuka et al. 1989, Maier et al. 1995), presumably following

transfer to the nucleus (Maier et al. 1995). In *Nicotiana sylvestris* (NC_007500) the *ycf*1 locus is 5,706 bases long and the *ycf*2 locus, 6,843 bases. In *Marchantia*, the *ycf*1 region (ORF464–ORF1068) is 4,662 bases, while *ycf*2 (ORF2136) is annotated at 6,411 bases. Both loci are shorter still in *Ptilidium* – the sequence space between *chl*N and *rps*15, in which *ycf*1 is located, is only 4,913 bases long and the annotated ORF itself is only 2,582 bases. The spacer in which the *ycf*2 locus is found, between *trn*H-GUG and *trn*D-GUC, is 6,016 bases long and contains an 5,240 base *ycf*2 ORF. The erroneously interrupted reading frames observed in the 454 sequence assembly likely represent a combination of the length of the loci (which increases the chances of containing sequencing errors), and the fact that they are rich in A and T mononucleotide repeats, which are particularly problematic for 454 sequencing. Moore et al. (2006) also found a high incidence of 454 sequencing errors in the *ycf*1 gene in two angiosperm plastids, likely related to the number of mononucleotide repeats in the locus.The *mat*K locus, which is situated in the *trn*K-UUU intron, lacks the entire maturase ORF in our annotation (annotated length 1,103; locus interrupted by internal stop codons; verification with Sanger sequencing failed). However, the *mat*K region appears to be essential for plastid function. Previous reports of *mat*K as a pseudogene in the hornwort *Anthoceros* (Kugita et al. 2003) may be premature; despite reporting non-edited nonsense codons, transcripts of *mat*K are detectable in *Anthoceros* (Kugita et al. 2003) (and also in *Phaeoceros* – Barthet & Hilu 2007). Later research shows *mat*K to have an essential function in the splicing of several plastid Group II introns (Barthet & Hilu 2007, Schmitz-Linneweber & Barkan 2007, Zoschke et al. 2010).

Nuclear data mining

The nuclear genome size of *Ptilidium pulcherrimum* was recently estimated, based on flow cytometry readings from four collections from two Austrian provinces, to be in the order of 1.2–1.3 pg (c. 1,158–1,255 Mbp), while one accession of the congener *P. ciliare* had 1.17 pg (c. 1,130 Mbp), and a second had 0.95 pg (c. 917 Mbp) (Temsch et al. 2010). These are at the smaller end of the known range of estimates for liverwort nuclear genome sizes, from 0.21–7.97 pg (c. 203–7,700 Mbp) (Temsch et al. 2010). In contrast, the nuclear genome of *P. pulcherrimum* is around 5.6 times larger than that of the angiosperm *Arabidopsis thaliana*, which has mean haploid genome size 0.215 pg (c. 211 Mbp) (Schmuths et al. 2004), and around 2.3 times larger than moss *Physcomitrella patens*, which has mean haploid genome size 0.53 pg (c. 511 Mbp) (Schween et al. 2003).

Ignoring for now the amount of both plastid and mitochondrial sequence data included in the reads, the 79,052,599 bases of 454 data randomly samples approximately 6.3–7.0% of the nuclear genome of *Ptilidium*. The true figure is, however, a little lower – removing plastid sequences from our total reads leaves just under 78 million bases of nuclear and mitochondrial sequence (reducing the sampling of the nuclear genome to 6.2–6.7%). We have not estimated the amount of mitochondrial sequence in the data, but expect it to be somewhat lower than the amount of plastid data, bringing us to a conservative estimate of a little over 0.05-fold (i.e. 5%) coverage of the *Ptilidium* nuclear genome sequenced for this project.

One proviso in using 454 data for candidate gene identification or other marker development projects is the issue of contamination – a small but significant proportion of identified small subunit ribosomal raw reads in our data were not from plants (c. 9%). Using contigs instead of raw reads can reduce the problem, as the likelihood of contaminating genomes, which should be present in the data at much lower proportions, assembling is far lower than the likelihood of getting assemblies from *Ptilidium* (e.g. only 1/122 small subunit ribosomal reads, or 0.8%, was from tardigrades, a group of animals that frequently inhabit bryophytes; Glime 2007).

Nearly 40% of purportedly nuclear domains identified from the *Ptilidium* data were Class I transposable elements (retroviruses and retrotransposons). Retrotransposons are highly abundant in plant genomes; in several plants they form the main component of nuclear DNA (Kumar & Bennetzen 1999, Marco & Marín 2005). About 50% of the *Physcomitrella* nuclear genome is made up of long terminal repeat retrotransposons (Rensing et al. 2008). Genomes from other plant lineages also contain a large proportion of retrotransposons – e.g. 49–78% in maize (SanMiguel & Bennetzen 1998) and about 68% in wheat (Li et al. 2004).

Microsatellites

Microsatellite development using enrichment or selective hybridization protocols can be time consuming (Squirrell et al. 2003, Zane et al. 2002). Further, enrichment-based protocols can be biased, as they target a subset of microsatellite loci, for example, a particular repeat motif (Castoe et al. 2010). Next generation sequencing data provides a rapid, potentially unbiased and powerful tool for microsatellite primer development: Castoe et al. (2010) sampled about 2% of the haploid genome of a diploid snake using a 454 FLX GS shotgun library (mean read length 215 bases), identifying thousands of microsatellite loci. However, they used unassembled reads to generate their microsatellite primers. We found that using only assembled data leads to a drastic reduction in the number of microsatellites identified. The reasons for this are twofold – firstly, not all the reads are assembled into contigs; c. 65.2% of our reads (i.e. over 153,000 reads) remain as singletons and we did not scan these for microsatellite primers. Secondly, redundancy (multiple hits on the same microsatellite present in different reads) is drastically reduced when data are assembled. Another advantage of assembling reads prior to microsatellite screening is that the regions flanking the microsatellite will often be longer, allowing for better primer design. Also, as mentioned above, reads from contaminant genomes are less likely to form contigs and so non-*Ptilidium* microsatellite loci are less likely to be developed.

On the other hand, assembly of reads prior to microsatellite screening may increase levels of genome bias in the data (Magain et al. 2010), as the genomes that are most likely to assemble are from the multi-copy plastid and mitochondrion, while nuclear regions are far less likely to assemble until the amount of coverage of the nuclear genome is high – at a financial cost that would be impractical in most instances.

Thus nuclear data are more likely to be discarded as singletons. However, if insufficient primers for a particular project are generated from the assembled contig data, or organellar bias is felt to be an issue, singletons can be bioinformatically extracted from the total 454 reads and screened separately.

Future prospects

An initial drawback of 454 sequencing for our research was its requirement for large quantities of high quality DNA (c. 5 µg pure high molecular weight DNA in 100 µL buffer). Such concentrations can only be reached when extracting from relatively large amounts of fresh tissue; for many liverworts this means gathering several hundred stem or thallus apices. This has presented the main hurdle to significant progress, both in obtaining sufficient high quality plant material (months to years of growth are required to generate the quantities of plant tissue required) and in extracting DNAs that meet quality specifications. However, the recent development of a 454 Rapid Library protocol that only requires 1/10 of the amount of input DNA (300–500 ng) has opened the way for data generation for a far greater number of liverworts, and we are now utilizing the protocol in the liverworts *Blasia* L., *Fossombronia* Raddi, *Haplomitrium* Nees, *Moerckia* Gottsche, *Pleurozia* Dumort., *Scapania* (Dumort.) Dumort. and *Treubia* K.I.Goebel and the hornwort *Nothoceros*. Plastid genomes for these taxa are all currently at the assembly or annotation stage. These will provide a valuable resource for plastid marker development within bryophytes, a group wherein even finding broadly amplifying primers for the relatively well-characterized *mat*K locus, as required for successful implementation of the CBOL-approved two-locus land plant DNA barcode, has thus far proved problematic (Hollingsworth et al. 2009).

One of the biological questions prompting the choice of *Ptilidium* for plastid sequencing was its uncertain placement in liverwort phylogeny. Gene order rearrangements could have provided a powerful marker for resolving this. However, multigene analyses using gene sequences generated by liverwort plastid genome sequencing may instead elucidate these early divergences within the leafy liverwort clades.

CONCLUSIONS

Using total genomic 454 data allows rapid and methodologically straightforward assembly of complete plastid genomes; in addition, the sequence data provides a valuable resource for marker development from all three genomic compartments, including microsatellites, SNPs, and novel genes and spacers. Remarkably, annotation of the *Ptilidium* plastid shows that the genome is perfectly collinear with that of *Marchantia*, despite over 370 million years (Heinrichs et al. 2007) having passed since their lineages diverged, and the presence of various short dispersed repeats, hypothesized as a factor in genomic rearrangements, throughout the *Ptilidium* genome. Plastid genes are often grouped into multigene transcriptional units; thus the highly conserved genome structure present in liverworts may reflect a lack of flexibility in which combinations of products are co-transcribed. Lower levels of molecular evolution in bryophyte organellar genomes may also contribute to the conservative nature of the liverwort plastid genome.

SUPPLEMENTARY DATA

Supplementary data are available at *Plant Ecology and Evolution*, Supplementary Data Site (http://www.ingentaconnect.com/content/botbel/plecevo/supp-data), and consist of the following: (1) Protein domains with hits from *Ptilidium* nuclear contigs, generated using InterProScan software with EMBOSS sixpack scanning against HMMPanther and HMMSmart (pdf format); (2) Mononucleotide to pentanucleotide microsatellites in *Ptilidium* and potential primer sequences, generated by an msatCommander 0.8.2 scan of contigs (Excel sheet).

ACKNOWLEDGEMENTS

This research was funded by a grant from the National Science Foundation (EF-0531557 to B.G.). The authors thank Blanka Shaw (Duke) for the provision of fresh plant material. The UConn Bioinformatics Facility provided computing resources for 454 sequence assembly. Goffinet lab members, in particular Juan Carlos Villarreal and Jessica Budke, are thanked for general assistance and encouragement. Comments from Susann Wicke and an anonymous reviewer further improved the manuscript.

REFERENCES

Barthet M.M., Hilu K.W. (2007) Expression of matK: functional and evolutionary implications. American Journal of Botany 94: 1402–1412. DOI: 10.3732%2Fajb.94.8.1402

Bidartondo M.I. (2005) The evolutionary ecology of myco-heterotrophy. New Phytologist 167: 335–352. DOI: 10.1111/j.1469-8137.2005.01429.x

Brundrett M.C. (2002) Coevolution of roots and mycorrhizas of land plants. New Phytologist 154: 275–304. DOI: 10.1046/j.1469-8137.2002.00397.x

Castoe T.A., Poole A.W., Gu W., de Koning A.P.J., Daza J.M., Smith E.N., Pollock D.D. (2010) Rapid identification of thousands of copperhead snake (Agkistrodon contortrix) microsatellite loci from modest amounts of 454 shotgun sequencing. Molecular Ecology Resources 10: 341–347. DOI: 10.1111/j.1755-0998.2009.02750.x

Chevreux B., Wetter, T., Suhai, S. (1999) Genome sequence assembly using trace signals and additional sequence information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99: 45–56. [available at http://www.bioinfo.de/isb/gcb99/talks/chevreux/]

Cosner M.E., Raubeson L.A., Jansen R.K. (2004) Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. BMC Evolutionary Biology 2004, 4: 27. DOI: 10.1186%2F1471-2148-4-27

Cronn R., Liston A., Parks M., Gernandt D.S., Shen R., Mockler, T. (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Research 36: e122. DOI: 10.1093/nar/gkn502

Cui L., Leebens-Mack J., Wang L.-S., Tang J., Rymarquis L., Stern D.B., dePamphilis C.W. (2006) Adaptive evolution of chloroplast genome structure inferred using a parametric boot-

strap approach. BMC Evolutionary Biology 2006, 6:13. DOI: 10.1186/1471-2148-6-13

Donaher N., Tanifuji G., Onodera N.T., Malfatti S.A., Chain P.S.G., Hara Y., Archibald J.M. (2009) The complete plastid genome sequence of the secondarily nonphotosynthetic alga Cryptomonas paramecium: reduction, compaction, and accelerated evolutionary rate. Genome Biology and Evolution 1: 439–448. DOI: 10.1093/gbe/evp047

Drescher A., Ruf S., Calsa T. Jr, Carrer H., Bock. R. (2000) The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. The Plant Journal 22: 97–104. DOI: 10.1046/j.1365-313x.2000.00722.x

Faircloth B.C. (2008) MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. Molecular Ecology Resources 8: 92–94. DOI: 10.1111/j.1471-8286.2007.01884.x

Fiz O., Vargas P., Alarcón M., Aedo C., García J.L., Aldasoro J.J. (2008) Phylogeny and historical biogeography of Geraniacae in relation to climate changes and pollination ecology. Systematic Botany 33: 326–342. DOI: 10.1600/036364408784571482

Forrest L.L., Crandall-Stotler B.J. (2005) Progress towards a robust phylogeny for the liverworts, with particular focus on the simple thalloids. Journal of the Hattori Botanical Laboratory 97: 127–159.

Forrest L.L., Davis E.C., Long D.G., Crandall-Stotler B.J., Hollingsworth M.L., Clark A. (2006) Unravelling the evolutionary history of the liverworts (Marchantiophyta) – multiple taxa, genomes and analyses. The Bryologist 109: 303–334. DOI: 10.1639/0007-2745(2006)109[303:UTEHOT]2.0.CO;2

Freyer R., Kiefer-Meyer M.C., Kössel H. (1997) Occurrence of plastid RNA editing in all major lineages of land plants. Proceedings of the National Academy of Sciences of the USA 94: 6285–6290. DOI: http://dx.doi.org/10.1073/pnas.94.12.6285

Gao L., Yi X., Yang Y.-X., Su Y.-J., Wang T. (2009) Complete chloroplast genome sequence of a tree fern Alsophila spinulosa: insights into evolutionary changes in fern chloroplast genomes. BMC Evolutionary Biology 2009, 9: 130. DOI: 10.1186/1471-2148-9-130

Glime J.M. (2007) Bryophyte Ecology. Volume 1. Physiological Ecology. Ebook sponsored by Michigan Technological University and the International Association of Bryologists. Available from http://www.bryoecol.mtu.edu/ [accessed 7 Sep. 2010]

Goffinet B., Wickett N.J., Shaw A.J., Cox C.J. (2005) Phylogenetic significance of the rpoA loss in the chloroplast genome of mosses. Taxon 54: 353–360. DOI: 10.2307/25065363

Goffinet B., Wickett N.J., Werner O., Ros R.M., Shaw A.J., Cox C.J. (2007) Distribution and phylogenetic significance of the 71 kb inversion in the chloroplast genome in the Funariidae (Bryophyta). Annals of Botany 99: 747–753. DOI: 10.1093/aob/mcm010

Goremykin V., Hirsch-Ernst K.I., Wölfl S., Hellwig F.H. (2003) The chloroplast genome of the "basal" angiosperm Calycanthus fertilis – structure and phylogenetic analyses. Plant Systematics and Evolution 242: 119–135. DOI: 10.1007/s00606-003-0056-4

Guisinger M.M., Kuehl J.V., Boore J.L., Jansen R.K. (2010) Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Molecular Biology and Evolution Advance Access 28: 583–600. DOI: 10.1093/molbev/msq229

Haberle R.C., Fourcade H.M., Boore J.L., Jansen R.K. (2008) Extensive rearrangements in the chloroplast genome of Trachelium caeruleum are associated with repeats and tRNA genes. Journal of Molecular Evolution 66: 350–361. DOI: 10.1007/s00239-008-9086-4

Heinrichs J., Gradstein S.R., Wilson R., Schneider H. (2005) Towards a natural classification of liverworts (Marchantiophyta) based on the chloroplast gene rbcL. Cryptogamie, Bryologie 26: 131–150.

Heinrichs J., Hentschel J., Wilson R., Feldberg K., Schneider H. (2007) Evolution of leafy liverworts (Jungermanniidae, Marchantiophyta): estimating divergence times from chloroplast DNA sequences using penalized likelihood with integrated fossil evidence. Taxon 56: 31–44. [available at http://www.ingentaconnect.com/content/iapt/tax/2007/00000056/00000001/art00004]

Hiratsuka J., Shimada H., Whittier R., Ishibashi T., Sakamoto M., Mori M., Kondo C., Honji Y., Sun C.-R., Meng B.-Y., Li Y.-Q., Kanno A., Nishizawa Y., Hirai A., Shinozaki K., Sugiura M. (1989) The complete sequence of the rice (Oryza sativa) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of cereals. Molecular and General Genetics 217: 185–194. DOI: 10.1007/BF02464880

Hollingsworth P.M., Forrest L.L., Spouge J.L., Hajibabaei M., Ratnasingham S., van der Bank M., Chase M.W., Cowan R.S., Erickson D.L., Fazekas A.J., Graham S.W., James K.E., Kim K.-J., Kress W.J., Schneider H., van AlphenStahl J., Barrett S.C.H., van den Berg C., Bogarin D., Burgess K.S., Cameron K.M., Carine M., Chacón J., Clark A., Clarkson J.J., Conrad F., Devey D.S., Ford C.S., Hedderson T.A.J., Hollingsworth M.L., Husband B.C., Kelly L.J., Kesanakurti P.R., Kim J.S., Kim Y.-D., Lahaye R., Lee H.-L., Long D.G., Madriñán S., Maurin O., Meusnier I., Newmaster S.G., Park C.-W., Percy D.M., Petersen G., Richardson J.E., Salazar G.A., Savolainen V., Seberg O., Wilkinson M.J., Yi D.-K., Little D.P. (2009) A DNA barcode for land plants. Proceedings of the National Academy of Sciences of the USA 106: 12794–12797. DOI: 10.1073/pnas.0905845106

Huang X., Madan A. (1999) CAP3: A DNA sequence assembly program. Genome Research 9: 868–877. DOI: 10.1101/gr.9.9.868

Jansen R.K., Palmer J.D. (1987) A chloroplast DNA inversion marks an ancient split in the sunflower family Asteraceae. Proceedings of the National Academy of Sciences of the USA 84: 5818–5822. DOI: 10.1073/pnas.84.16.5818

Kallas T., Spiller S., Malkin R. (1988) Primary structure of cotranscribed genes encoding the Rieske Fe-S and cytochrome f proteins of the cyanobacterium Nostoc PCC 7906. Proceedings of the National Academy of Sciences of the USA 85: 5794–5798. DOI: 10.1073/pnas.85.16.5794

Kelch D.G., Driskell A., Mishler B.D. (2004) Inferring phylogeny using genomic characters: a case study using land plant plastomes. Monographs in Systematic Botany from the Missouri Botanical Garden 68: 3–12.

Knoop V. (2004) The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Current Genetics 46: 123–139. DOI: 10.1007/s00294-004-0522-8

Knoop V. (2010) Looking for sense in the nonsense: a short review of non-coding organellar DNA elucidating the phylogeny of bryophytes. Tropical Bryology 31: 51–60. [available at http://tropical-bryology.org/online/V31_35]

Kugita M., Kaneko A., Yamamoto Y., Takeya Y., Matsumoto T., Yoshinaga K. (2003) The complete nucleotide sequence of the hornwort (Anthoceros formosae) chloroplast genome: insight into the earliest land plants. Nucleic Acids Research 31: 716–721. DOI: 10.1093/nar/gkg155

Kumar A., Bennetzen J.L. (1999) Plant retrotransposons. Annual Review of Genetics 33: 479–532. DOI: 10.1146/annurev.genet.33.1.479

Kurtz S., Schleiermacher C. (1999) REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15: 426–427. DOI: 10.1093/bioinformatics/15.5.426

Lee H.L., Jansen R.K., Chumley T.W., Kim K.-J. (2007) Gene relocations within chloroplast genomes of Jasminum and Menodora (Oleaceae) are due to multiple, overlapping inversions. Molecular Biology and Evolution 24: 1161–1180. DOI: 10.1093/molbev/msm036

Li L., Wang B., Liu Y., Qiu Y.L. (2009) The complete mitochondrial genome sequence of the hornwort Megaceros aenigmaticus shows a mixed mode of conservative yet dynamic evolution in early land plant mitochondrial genomes. Journal of Molecular Evolution 68: 665–678. DOI: 10.1007/s00239-009-9240-7

Li W., Zhang P., Fellers J.P., Friebe B., Gill B.S. (2004) Sequence composition, organization, and evolution of the core Triticeae genome. The Plant Journal 40: 500–11. DOI: 10.1111/j.1365-313X.2004.02228.x

Lohse M., Drechsel O., Bock R. (2007) OrganellarGenomeDRAW (OGDRAW) – a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Current Genetics 52: 267–274. DOI: 10.1007/s00294-007-0161-y

Magain N., Forrest L.L., Sérusiaux E., Goffinet B. (publ. online 2010) Microsatellite primers in the Peltigera dolichorhiza complex (lichenized ascomycete, Peltigerales). American Journal of Botany. DOI: 10.3732/ajb.1000283

Maier R.M., Neckermann K., Igloi G.L., Kössel H. (1995) Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. Journal of Molecular Biology 251: 614–628. DOI: 10.1006/jmbi.1995.0460

Marco A., Marín I. (2005) Retrovirus-like elements in plants. Recent Research Developments in Plant Science 3: 15–24. [available at http://www.uv.es/genomica/spa/PDF/27.pdf]

Maul J.E., Lilly J.W., Cui L., dePamphilis C.W., Miller W., Harris E.H., Stern D.B. (2002) The Chlamydomonas reinhardtii plastid chromosome: islands of genes in a sea of repeats. The Plant Cell 14: 2659–2679. DOI: 10.1105/tpc.006155

McNeal J.R., Leebens-Mack J.H., Arumuganathan K., Kuehl J.V., Boore J.L., dePamphilis C.W. (2006) Using partial genomic fosmid libraries for sequencing complete organellar genomes. BioTechniques 41: 69–73. DOI: 10.2144/000112202

Milligan B.G., Hampton J.N., Palmer J.D. (1989) Dispersed repeats and structural reorganization in subclover chloroplast DNA. Molecular Biology and Evolution 6: 355–368. [available at http://mbe.oxfordjournals.org/content/6/4/355.full.pdf]

Mishler B.D., Kelch D.G. (2009) Phylogenomics and early land plant evolution. In: Goffinet B., Shaw A.J. (eds) Bryophyte Biology (Ed. 2): 173–197. Cambridge, Cambridge University Press.

Monde R.A., Schuster G., Stern D.B. (2000) Processing and degradation of chloroplast mRNA. Biochimie 82: 573–582. DOI: 10.1016/S0300-9084%2800%2900606-4

Moore M.J., Dhingra A., Soltis P.S., Shaw R., Farmerie W.G., Folta K.M., Soltis D.E. (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biology 2006, 6: 17. DOI: 10.1186/1471-2229-6-17

Mulder N.J., Apweiler R., Attwood T.K., Bairoch A., Bateman A., Binns D., Bork P., Buillard V., Cerutti L., Copley R., Courcelle E., Das U., Daugherty L., Dibley M., Finn R., Fleischmann W., Gough J., Haft D., Hulo N., Hunter S., Kahn D., Kanapin A., Kejariwal A., Labarga A., Langendijk-Genevaux P.S., Lonsdale D., Lopez R., Letunic I., Madera M., Maslen J., McAnulla C., McDowall J., Mistry J., Mitchell A., Nikolskaya A.N., Orchard S., Orengo C., Petryszak R., Selengut J.D., Sigrist C.J.A., Thomas P.D., Valentin F., Wilson D., Wu C.H., Yeats C. (2007) New developments in the InterPro database. Nucleic Acids Research 35: D224–D228. DOI: 10.1093/nar/gkl841

Nawrocki E.P., Kolbe D.L., Eddy S.R. (2009) Infernal 1.0: Inference of RNA alignments. Bioinformatics 25: 1335–1337. DOI: 10.1093/bioinformatics/btp157

Oda K., Yamato K., Ohta E., Nakamura Y., Takemura M., Nozato N., Akashi K., Kanegae T., Ogura Y., Kohchi T., Ohyama K. (1992) Gene organization deduced from the complete sequence of liverwort Marchantia polymorpha mitochondrial DNA – a primitive form of plant mitochondrial genome. Journal of Molecular Biology 223: 1–7. DOI: 10.1016/0022-2836%2892%2990708-R

Ohyama K., Fukuzawa H., Kohchi T., Sano T., Shirai H., Umesono K., Shiki Y., Takeuchi M., Chang Z, Atoa S.i., Inokuchi H., Ozeki H. (1988) Structure and organization of Marchantia polymorpha chloroplast genome. I. Cloning and gene identification. Journal of Molecular Biology 203: 281–298. DOI: 10.1016/0022-2836%2888%2990001-0

Ohyama K., Fukuzawa H., Kohchi T., Shirai H., Sano T., Sano S., Umesono K., Shiki Y., Takeuchi M., Chang Z., Aota S.-i., Inokuchi H., Ozeki H. (1986) Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. Nature 322: 572–574. DOI: 10.1038/322572a0

Oliver M.J., Murdock A.G., Mishler B.D., Kuehl J.V., Boore J.L., Mandoli D.F., Everett K.D.E., Wolf P.G., Duffy A.M., Karol K.G. (2010) Chloroplast genome sequence of the moss Tortula ruralis: gene content, polymorphism, and structural rearrangement relative to other green plant chloroplast genomes. BMC Genomics 11: 143. DOI: 10.1186/1471-2164-11-143

Pyke K. (2009) Plastid Biology. Cambridge, U.K., Cambridge University Press.

Qiu Y.-L., Li L., Wang B., Chen Z., Knoop V., Groth-Malonek M., Dombrovska O., Lee J., Kent L., Rest J., Estabrook G.F., Hendry T.A., Taylor D.W., Testa C.M., Ambros M., Crandall-Stotler B., Duff R.J., Stech M., Frey W., Quandt D., Davis C.C. (2006) The deepest divergences in land plants inferred from phylogenomic evidence. Proceedings of the National Academy of Sciences of the USA 103: 15511–15516. DOI: 10.1073/pnas.0603335103

Quevillon E., Silventoinen V., Pillai S., Harte N., Mulder N., Apweiler R., Lopez R. (2005) InterProScan: protein domains identifier. Nucleic Acids Research 33: W116–W120. DOI: 10.1093/nar/gki442

Raubeson L.A., Jansen R.K. (1992) Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. Science 255: 1697–1699. DOI: 10.1126/science.255.5052.1697

Raubeson L.A., Jansen R.K. (2005) Chloroplast genomes of plants. In: Henry R.J. (ed) Plant diversity and evolution: genotypic and phenotypic variation in higher plants: 45–68. Wallingford, UK, CAB International.

Raubeson L.A., Peery R., Chumley T.W., Dziubek C., Fourcade H.M., Boore J.L., Jansen R.K. (2007) Comparative chloroplast genomics: analyses including new sequences from the angiosperms Nuphar advena and Ranunculus macranthus. BMC Genomics 2007, 8: 174. DOI: 10.1186/1471-2164-8-174

Ravi V., Khurana J.P., Tyagi A.K., Khurana P. (2006) The chloroplast genome of mulberry: complete nucleotide sequence, gene

organization and comparative analysis. Tree Genetics and Genomes 3: 49–59. DOI: 10.1007/s11295-006-0051-3

Read, D.J., Duckett J.D., Francis R., Ligrone R., Russell A. (2000) Symbiotic fungal associations in 'lower' land plants. Philosophical Transactions of the Royal Society of London Ser. B. 355: 815–830. DOI: 10.1098/rstb.2000.0617

Rensing S.A., Lang D., Zimmer A.D., Terry A., Salamov A., Shapiro H., Nishiyama T., Perroud P.-F., Lindquist E.A., Kamisugi Y., Tanahashi T., Sakakibara K., Fujita T., Oishi K., Shin-I T., Kuroki Y., Toyoda A., Suzuki Y., Hashimoto S.-i., Yamaguchi K., Sugano S., Kohara Y., Fjiyama A., Anterola A., Aoki S., Ashton N., Barbazuk W.B., Barker E., Bennetzen J.L., Blankenship R., Cho S.H., Dutcher, S.K. Estelle M., Fawcett J.A., Gundlach H., Hanada K., Heyl A., Hicks K.A., Hughes J., Lohr M., Mayer K., Melkozernov A., Murata T., Nelson D.R., Pils B., Prigge M., Reiss B., Renner T., Rombauts S., Rushton P.J., Sanderfoot A., Schween G., Shiu S.-H., Stueber K., Theodoulou F.L., Tu H., Van de Peer Y., Verrier P.J., Waters E., Wood A., Yang L., Cove D., Cuming A.C., Hasebe M., Lucas S., Mishler B.D., Reski R., Grigoriev I.V., Quatrano, R.S., Boore J.L. (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science 319: 64–69. DOI: 10.1126/science.1150646

Roquet C., Sanmartín I., Garcia-Jacas N., Sáez L., Susanna A., Wikström N., Aldasoro J.J. (2009) Reconstructing the history of Campanulaceae with a Bayesian approach to molecular dating and dispersal-vicariance analyses. Molecular Phylogenetics and Evolution 52: 575–587. DOI: 10.1016/j.ympev.2009.05.014

Rozen S., Skaletsky H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S., Misener S. (eds) Bioinformatics methods and protocols: methods in molecular biology: 365–386. Totowa, NJ, Humana Press.

SanMiguel P., Bennetzen J.L. (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Annals of Botany 82 (Suppl. A): 37–44. DOI: 10.1006/anbo.1998.0746

Schmitz-Linneweber C., Barkan A. (2007) RNA splicing and RNA editing in chloroplasts. In: Bock R. (ed.) Topics in current genetics 19: cell and molecular biology of plastids: 213–248. Berlin, Springer-Verlag.

Schmuths H., Meister A., Horres R., Bachmann K. (2004) Genome size variation among accessions of Arabidopsis thaliana. Annals of Botany 93: 317–321. DOI: 10.1093/aob/mch037

Schuster R.M. (1966) Ptilidium. In: Schuster R.M. (ed.) (1966–1993) Hepaticae and Anthocerotae of North America East of the Hundredth Meridian. Vol. 1: 760–780. New York, Columbia University Press.

Schuster R.M. (1992) The Hepaticae and Anthocerotae of North America, Vol. V. Chicago IL., Field Museum of Natural History.

Schween G., Gorr G., Hohe A., Reski R. (2003) Unique tissue-specific cell cycle in Physcomitrella. Plant Biology 5: 50–58. DOI: 10.1055/s-2003-37984

Squirrell J., Hollingsworth P.M., Woodhead M., Russell J., Lowe A., Gibby M., Powell W. (2003) How much effort is required to isolate nuclear microsatellites from plants? Molecular Ecology 12: 1339–1348. DOI: 10.1046/j.1365-294X.2003.01825.x

Stein D.B., Conant D.S., Ahearn M.E., Jordan E.T., Kirch S.A., Hasebe M., Iwatsuki K., Tan M.K., Thomson J.A. (1992) Structural rearrangements of the chloroplast genome provide and important phylogenetic link in ferns. Proceedings of the National Academy of Sciences of the USA 89: 1856–1860. DOI: 10.1073/pnas.89.5.1856

Stenøien H.K. (2008) Slow molecular evolution in 18S rDNA, rbcL and nad5 genes of mosses compared with higher plants. Journal of Evolutionary Biology 21: 566–571. DOI: 10.1111/j.1420-9101.2007.01479.x

Stern D.B., Goldschmidt-Clermont M., Hanson M.R. (2010) Chloroplast RNA metabolism. Annual Review of Plant Biology 61L: 125–155. DOI: 10.1146/annurev-arplant-042809-112242

Sugita M., Sugiura C., Arkkawa T., Higuchi M. (2004) Molecular evidence of an rpoA gene in the basal moss chloroplast genomes: rpoA is a useful marker for phylogenetic analysis of mosses. Hikobia 14: 171–175.

Sugiura C., Kobayashi Y., Aoki S., Sugita C., Sugita M. (2003) Complete chloroplast DNA sequence of the moss Physcomitrella patens: evidence for the loss and relocation of rpoA from the chloroplast to the nucleus. Nucleic Acids Research 31: 5324–5331. DOI: 10.1093/nar/gkg726

Tangphatsornruang S., Sangsrakru D., Chanprasert J., Uthaipaisanwong P., Yoocha T., Jomchai N., Tragoonrung S. (2010) The chloroplast genome sequence of mungbean (Vinga radiata) determined by high-throughput pyrosequencing: structural organization and phylogenetic relationships. DNA Research 17: 11–22. DOI: 10.1093/dnares/dsp025

Temsch E.M., Greilhuber J., Krisai R. (2010) Genome size in liverworts. Preslia 82: 63–80.

Terasawa K., Odahara M., Kabeya Y., Kikugawa T., Sekine Y., Fujiwara M., Sato N. (2006) The mitochondrial genome of the moss Physcomitrella patens sheds new light on mitochondrial evolution in land plants. Molecular Biology and Evolution 24: 699–709. DOI: 10.1093/molbev/msl198

The InterPro Consortium (Apweiler R., Attwood T.K., Bairoch A., Bateman A., Birney E., Biswas M., Bucher P., Cerutti L., Corpet F., Croning M.D.R., Durbin R., Falquet L., Fleischmann W., Gouzy J., Hermjakob H., Hulo N., Jonassen I., Kahn D., Kanapin A., Karavidopoulou Y., Lopez R., Marx B., Mulder N.J., Oinn T.M., Pagni M., Servant F., Sigrist C.J.A., Zdobnov E.M.) (1999) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Research 29: 37–40. DOI: 10.1093/nar/29.1.37

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, transpositions and many gene losses. Journal of Plant Research 120: 281–290. DOI: 10.1007/s10265-006-0055-y

Van Aller Hernick L., Landing E., Bartowski K.E. (2008) Earth's oldest liverworts – Metzgeriothallus sharonae sp. nov. from the Middle Devonian (Givetian) of Eastern New York. Review of Palaeobotany & Palynology 148: 154–162. DOI: 10.1016/j.revpalbo.2007.09.002

von Konrat M., Söderströmm L., Renner M.A.M., Hagborg A., Briscoe L., Engel J.J. (2010) Early Land Plants Today (ELPT): how many liverwort species are there? Phytotaxa 9: 22–40. [available at http://www.mapress.com/phytotaxa/taxa/Bryophytes.htm]

Wahrmund U., Groth-Malonek M., Knoop V. (2008) Tracing plant mitochondrial DNA evolution: rearrangements of the ancient mitochondrial gene cluster trnA-trnT-nad7 in liverwort phylogeny. Journal of Molecular Evolution 66: 621–629. DOI: 10.1007/s00239-008-9114-4

Wang B., Xue J., Li L., Qiu Y.-L. (2009) The complete mitochondrial genome sequence of the liverwort Pleurozia purpurea reveals extremely conservative mitochondrial genome evolution in liverworts. Current Genetics 55: 601–609. DOI: 10.1007/s00294-009-0273-7

Wicker T., Schlagenhauf E., Graner A., Close T.J., Keller B., Stein N. (2006) 454 sequencing put to the test using the complex genome of barley. BMC Genomics 7: 275. DOI: 10.1186/1471-2164-7-275

Wickett N.J., Fan Y., Lewis P.O., Goffinet B. (2008a) Distribution and evolution of pseudogenes, gene losses and a gene rearrangement in the plastid genome of the non-photosynthetic liverwort, Aneura mirabilis (Metzgeriales, Jungermanniopsida). Journal of Molecular Evolution 67: 111–122. DOI: 10.1007/s00239-008-9133-1

Wickett N.J., Zhang Y., Hansen S.K., Roper J.M., Kuehl J.V., Plock S.A., Wolf P.G., dePamphilis C.W., Boore J.L., Goffinet B. (2008b) Functional gene losses occur with minimal size reduction in the plastid genome of the parasitic liverwort Aneura mirabilis. Molecular Biology and Evolution 25: 393–401. DOI 10.1093/molbev/msm267

Wolf P.G., Karol K.G., Mandoli D.F., Kuehl J., Arumuganathan K., Ellis M.W., Mishler B.D., Kelch D.G., Olmstead R.G., Boore J.L. (2005) The first complete chloroplast genome sequence of a lycophyte, Huperzia lucidula (Lycopodiaceae). Gene 350: 117–128. DOI: 10.1016/j.gene.2005.01.018

Wolf P.G., Der J.P., Duffy A.M., Davidson J.B., Grusz A.L., Pryer K.M. (2010) The evolution of chloroplast genes and genomes in ferns. Plant Molecular Biology. DOI:10.1007/s11103-010-9706-4

Woodbury N.W., Roberts L.L., Palmer J.D., Thompson W.F. (1988) A transcription map of the pea chloroplast genome. Current Genetics 14: 75–89. DOI: 10.1007/BF00405857

Wu C.-S., Wang Y.-N., Liu S.-M., Chaw S.-M. (2007) Chloroplast genome (cpDNA) of Cycas taitungensis and 56 cp protein-coding genes of Gnetum parvifolium: insights into cpDNA evolution and phylogeny of extant seed plants. Molecular Biology and Evolution 24: 1366–1379. DOI: 10.1093/molbev/msm059

Wyman S.K., Jansen R.K., Boore, J.L. (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252–3255. DOI: 10.1093/bioinformatics/bth352

Xue J.-Y., Liu Y., Li L., Wang B. (2009) The complete mitochondrial genome sequence of the hornwort Phaeoceros laevis: retention of many ancient pseudogenes and conservative evolution of mitochondrial genomes in hornworts. Current Genetics 56: 53–61. DOI: 10.1007/s00294-009-0279-1

Yoshinaga K., Iinuma H., Masuzawa T., Uedal K. (1996) Extensive editing of C to U substitution in the rbcL transcripts of hornwort chloroplasts and the origin of RNA editing in green plants. Nucleic Acids Research 24: 1008–1014. DOI: 10.1093/nar/24.6.1008

Zane L., Bargelloni L., Patarnello T. (2002) Strategies for microsatellite isolation: a review. Molecular Ecology 11: 1–16. DOI: 10.1046/j.0962-1083.2001.01418.x

Zoschke R., Nakamura M., Liere K., Sugiura M., Börner T., Schmitz-Linneweber C. (2010) An organellar maturase associates with multiple group II introns. Proceedings of the National Academy of Sciences of the USA 107: 3245–3250. DOI: 10.1073/pnas.0909400107

Manuscript received 28 Sep. 2010; accepted in revised version 21 Dec. 2010.

Communicating Editor: Bart Van de Vijver.