American Journal of Botany

American Journal of Botany 98(8): 1263–1275. 2011.

Frequent pseudogenization and loss of the plastid-encoded sulfate-transport gene *cys*A throughout the evolution of liverworts[1]

Norman J. Wickett[2,5], Laura L. Forrest[3], Jessica M. Budke[3], Blanka Shaw[4], and Bernard Goffinet[3]

[2]Department of Biology, Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, Pennsylvania 16802 USA; [3]Department of Ecology and Evolutionary Biology, University of Connecticut, 75 North Eagleville Road, Unit 3043, Storrs, Connecticut 06269-3043 USA; and [4]Department of Biology, Duke University, Durham, North Carolina 27708 USA

- *Premise of the study:* The presence or absence of a functional copy of a plastid gene may reflect relaxed selection, and may be phylogenetically significant, reflecting shared ancestry. In some liverworts, the plastid gene *cys*A is a pseudogene (inferred to be nonfunctional). We surveyed 63 liverworts from all major clades to determine whether the loss of *cys*A is phylogenetically significant, whether intact copies of *cys*A are under selective constraints, and whether rates of nucleotide substitution differ in other plastid genes from taxa with and without a functional copy of *cys*A.
- *Methods:* Primers annealing to flanking and internal regions were used to amplify and sequence *cys*A from 61 liverworts. Two additional *cys*A sequences were downloaded from NCBI. The ancestral states of *cys*A were reconstructed on a phylogenetic hypothesis inferred from seven markers. Rates of nucleotide substitution were estimated for three plastid loci to reflect the intrinsic mutation rate in the plastid genome. The ratio of nonsynonymous to synonymous substitutions was estimated for intact copies of *cys*A to infer selective constraints.
- *Key results:* Throughout liverworts, *cys*A has been lost up to 29 times, yet intact copies of *cys*A are evolving under selective constraints. Gene loss is more frequent in groups with an increased substitution rate in the plastid genome.
- *Conclusions:* The number of inferred losses of *cys*A in liverworts exceeds any other reported plastid gene. Despite frequent losses, *cys*A is evolving under purifying selection in liverworts that retain the gene. It appears that *cys*A is lost more frequently in lineages characterized by a higher rate of nucleotide substitutions in the plastid.

Key words: *cys*A; gene loss; liverwort; plastid; pseudogene.

The plastid genome provides a rich set of phylogenetic markers, including structural rearrangements, coding and noncoding DNA sequences, and differences in gene content. Understanding the evolutionary patterns and processes that govern plastid genome characters is essential for interpreting the phylogenetic hypotheses that are inferred from these markers. The photosynthetic organelle of green plants, the plastid, originated when an early eukaryotic cell engulfed, and subsequently enslaved, a photosynthetic cyanobacterium (Margulis, 1970; Cavalier-Smith, 2002). Through the transfer of much of the cyanobacterial genome to the nucleus of the "host" cell, the endosymbiont lost its autonomy and eventually retained, primarily, genes encoding products of the photosynthetic pathway or involved in the transcription and translation of those genes (Martin et al., 2002). A minority of genes is retained in the plastid genome for nonphotosynthetic functions such as fatty acid biosynthesis (dePamphilis and Palmer, 1990; Schnurr et al., 2002) or cysteine and methionine biosynthesis (Romer et al., 1992; Hell, 1997). The plastid genome that remains in green plants as a ghost of its cyanobacterial origin is typically made up of 120-200 genes that can be mapped as a circular molecule with a large single-copy region and a small single-copy region separated by two large inverted repeats (Martin et al., 1998; Raubeson and Jansen, 2005). In embryophytes, the plastid genome is particularly well-conserved in structure and content, with some dramatic exceptions, e.g., *Epifagus virginiana* (L.) W. Bartram (Orobanchaceae) (Wolfe et al., 1992) and *Pelargonium* ×*hortorum* L.H. Bailey (Geraniaceae) (Chumley et al., 2006).

Conservation in function, content, and structure of the plastid genome has led to the frequent use of plastid-encoded genes in phylogenetic reconstruction, and more recently to the use of whole-genome data to estimate phylogenies (e.g., Jansen et al., 2007; Moore et al., 2007). Structural rearrangements, such as inversions, of the plastid genome have been used to add support to existing hypotheses of phylogenetic relationships (e.g., Goffinet et al., 2007). These changes in gene order and gene content were thought to be relatively rare events (Raubeson and Jansen, 2005) that are therefore less subject to homoplasy and thus have a strong phylogenetic signal (Wolf et al., 2010); however, mutational biases and relaxed selection may lead to frequent rearrangements (e.g., Geraniaceae; Guisinger et al., 2011). The loss of a particular gene from the plastid genome can occur repeatedly in independent lineages. For example, the plastid-encoded RNA polymerase gene, *rpo*A, was shown to be lost twice

[1]Manuscript received 6 January 2011; revision accepted 31 May 2011.

The authors thank P. Lewis (University of Connecticut) for help with the scaling factor calculations, and the bioinformatics facility there for use of the computer cluster. Comments by members of the Goffinet laboratory, two anonymous reviewers and an associate editor who dealt with our submission have improved this paper. This research was supported by National Science Foundation grant Assembling the Liverwort Tree of Life DEB-0531557 to Goffinet, DEB-0531730 to Shaw, and supplemental Research Experiences for Undergraduates (REU) funding.

[5]Author for correspondence (e-mail: njw17@psu.edu)

doi:10.3732/ajb.1100010

throughout the evolutionary history of mosses (Goffinet et al., 2005); a ribosomal protein gene, *rps*16, has been lost independently several times within legumes (Doyle et al., 1995). Repeated loss of any character naturally leads to a discussion of whether ancestral character state reconstructions should be asymmetrically weighted to favor losses over regains of that character. In the case of complex characters that are thought to be more easily lost than gained, the evidence supporting the plausibility of each character state transition (e.g., gains vs. losses) should be evaluated (Doyle et al., 1995; Omland, 1997, 1999; McPherson et al., 2004). Though plastid gene loss is relatively common, and indicative of a long-term evolutionary trend of genome reduction, gene gain is virtually unknown, with the exception of the ancient acquisition of *mat*K in the ancestor of embryophytes and closely related green algae (Turmel et al., 2005). The rapid, irreversible pseudogenization of genes released from selective pressure (Marshall et al., 1994) and the lack of evidence for gene transfer *into* the plastid genome additionally supports the use of an asymmetric weighting scheme that heavily favors gene loss over gain when reconstructing ancestral states. As complete plastid genome sequences are becoming more readily available for greater numbers of taxa at lower taxonomic levels, understanding the dynamics of gene content, its phylogenetic significance, and its evolutionary history is increasingly important.

Fully sequenced plastid genomes are currently available on GenBank from at least 108 angiosperms. However, within bryophytes, only six plastid genomes have been sequenced: the mosses *Physcomitrella patens* (Hedw.) Bruch & Schimp. (Sugiura et al., 2003) and *Syntrichia ruralis* (Hedw.) F. Weber & D. Mohr (Oliver et al., 2010), the hornwort *Anthoceros angustus* Stephani (Kugita et al., 2003), and the liverworts *Marchantia polymorpha* L. (Ohyama et al., 1986), *Aneura mirabilis* (Malmb.) Wickett & Goffinet (Wickett et al., 2008b), and *Ptilidium pulcherrimum* (G. Weber) Hampe (Forrest et al., 2011). Four genes that are present in a green alga (*Chlorella* M. Beijerinck), other bryophytes (*Marchantia* L. and *Anthoceros* L.) and some embryophytes were found absent from the *Physcomitrella patens* plastid genome: *rpo*A, *ccs*A, *cys*A, and *cys*T; *Aneura mirabilis,* a nonphotosynthetic, subterranean liverwort, is characterized by the pseudogenization of several genes encoding proteins of the photosynthetic apparatus (Wickett et al., 2008b). Initially, and despite their absence from the *P. patens* plastid genome, two pseudogenes present in the *A. mirabilis* genome, *cys*A and *cys*T, were thought to be nonfunctional due to the liverwort's unique life history; the plastid genome of this mycoheterotrophic, nonphotosynthetic liverwort includes many other pseudogenes due to relaxed selection on the photosynthetic apparatus. Amplicons of *cys*A and *cys*T from some photosynthetic Aneuraceae also lack an intact open reading frame (ORF) for either, or both loci (Wickett et al., 2008a). However, this survey of *cys*A and *cys*T from photosynthetic liverworts was limited to a clade of simple thalloid liverworts, the Metzgeriales, and did not suggest a broader pattern.

The detection of pseudogenes of *cys*A and *cys*T from photosynthetic liverworts suggests that the selective pressures that maintain these genes are not linked to photosynthesis; furthermore, it appears that liverworts are the sole clade of embryophytes to include members with both genes in their plastid genomes (Cui et al., 2006). Based on sequence similarity, these genes appear to be ABC-type transporter genes involved in the import of sulfate into the plastid (Chen et al., 2003), which would implicate them in the biosynthesis of cysteine and methionine, a plastid function unrelated to photosynthesis. Melis and Chen (2005) suggest, however, that the prokaryotic-type sulfate permease genes were not retained throughout the evolution of embryophytes, which is consistent with the absence of *cys*A and *cys*T from the plastids of mosses and tracheophytes, unlike liverworts and hornworts. In a species of the cyanobacterium *Synechococcus* Nageli, *cys*A, *cys*T, and *cys*W encode, at least in part, structural proteins of the sulfate permease complex and are essential for growth (Laudenbach and Grossman, 1991). Genes with sequence homology to these three transporter genes, in addition to a fourth gene, Sbp, were identified in the nuclear genome of the green alga *Chlamydomonas reinhardtii* P.A. Dangeard (Melis and Chen, 2005). Though plastid encoded *cys*A and *cys*T are found in liverworts, hornworts, and some green algae, *cys*W and Sbp have yet to be identified from either the plastid or nuclear genome of embryophytes.

Here, we present an analysis of *cys*A in the plastid genomes of liverworts to address the molecular evolution of this ancient plant gene. We surveyed 63 species of liverworts to determine whether the loss of *cys*A is a phylogenetically informative character, to reconstruct the ancestral state of the gene in various lineages of liverworts, and to estimate whether the gene is selectively retained in some liverwort lineages.

MATERIALS AND METHODS

Detection of cysA—DNA accessions from 61 liverworts were sampled for the presence of the cysA gene (Appendix 1). Additionally, the cysA sequence was extracted from the plastid genomes of the liverworts *Marchantia polymorpha* (Ohyama et al., 1986) and *Aneura mirabilis* (Wickett et al., 2008b), as well as the hornwort, *Anthoceros formosae* Stephani (Kugita et al., 2003), and the green alga *Chara vulgaris* L. (Turmel et al., 2006). Degenerate PCR primers were initially designed to anneal to flanking tRNA genes (trnT-GGU and trnE-UUC), using the plastid genomes of *A. mirabilis* and *M. polymorpha* as templates. Additional internal primers were designed as cysA sequences were obtained, and were used in combination with the original primers in a nested approach, or on their own. Primer sequences are provided in Appendix 2. PCR amplification of this region was carried out in 25 μl reactions containing 1 unit taq DNA polymerase (5 PRIME, Gaithersburg, Maryland, USA), 1× buffer (5 PRIME taq buffer advanced with Mg^{2+}), 800 μM total dNTPs (200 μM of each dNTPs), and 200 μM of each primer, with 0.5-1 μl total genomic DNA extract. The amplification profile was 94°C for 1 min 30 s, followed by 30 cycles of denaturing (95°C for 20 s), annealing (58°C for 45 s) and extension (68°C for 1 min) with a final 7 min extension step at 68°C. PCR products were visualized under ultraviolet light on a 1% agarose tris-borate-EDTA (TBE) gel using SYBR Safe (Invitrogen, Carlsbad, California, USA) staining, to confirm the presence and size of amplicons. Subsequently, the products were purified using the NucleoSpin Extract II Kit (Machery-Nagel, distributed by Clontech, Mountain View, California, USA) in preparation for cycle sequencing. Amplicons were sequenced using the PCR primers; in longer amplicons internal primers were also used to obtain fully double-stranded sequence reads. Sequencing reactions were performed in 10 μl reactions using the ABI PRISM BigDye Terminators v 1.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, California, USA), using quarter reactions (i.e., 2 μl BigDye Buffer v.3.1 5× and 1 μl BigDye Terminator v.1.1 per reaction). Sequencing products were purified using Sephadex G-50 (Amersham, Piscataway, New Jersey, USA) gel filters, and then separated by capillary electrophoresis using the ABI Prism 3100 Genetic Analyser. Sequences were edited with Sequencher v 4.9 (GeneCodes, Ann Arbor, Michigan, USA), and translated with MacClade 4.07 (Maddison and Maddison, 2000).

Primers internal to *cys*A were developed to facilitate complete double-stranded sequencing of the *cys*A gene when present, particularly in the complex thalloid lineage. Internal primers were also used as a further test for *cys*A presence outside the plastid genome, or in an alternative position within the plastid genome. Amplifications were attempted using all possible combinations of internal primers and multiple concentrations of DNA template per taxon surveyed. A low annealing temperature of 48°C was used to maximize the likelihood of amplifying the *cys*A gene in taxa where it is possible that the gene has been transferred to the nucleus (e.g., Sugiura et al., 2003).

NCBI-GenBank accession numbers for the loci between *trn*T-GGU and *trn*E-UUC, whether it includes an ORF for *cys*A, are given in Appendix 1.

Phylogenetic marker sampling—Seven markers were sampled to reconstruct the phylogenetic relationship among the genera sampled for the presence of *cys*A. A well-resolved phylogeny enables the use of tree-based methods of estimating substitution rates, and is also necessary for ancestral character-state reconstructions.

Sequences for phylogenetic reconstructions were either downloaded from GenBank or generated de novo following methodology given in Davis (2004). Four plastid regions were included: 1) *rbc*L (the ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit); 2) *rps*4 (small ribosomal protein 4); 3) *psb*A (amplicon = partial *trn*K, *trn*K-*psb*A intergenic spacer, the photosystem II 32 kDa protein gene, *psb*A-*trn*H intergenic spacer, and partial *trn*H); and 4) *psb*T (here, we include partial sequences from *psb*T, the entire *psb*N gene, and partial sequences of *psb*H, which encode structural subunits of the photosystem II complex). There were two mitochondrial regions, amplicons from both *nad*1 (NADH dehydrogenase subunit 1) and *nad*5 (NADH dehydrogenase subunit 5, including a group I intron), and lastly an amplicon of the nuclear ribosomal large subunit (alternatively known as LSU or 26S) between primers 0F and 12R. Primer sequences are available in Forrest and Crandall-Stotler (2005), with two exceptions: *psb*T (in Krellwitz et al., 2001), and *nad*1 (in Demesure et al., 1995).

NCBI-GenBank accession numbers for all sequences used to generate the phylogeny are given in Appendix 1.

Phylogeny and Ancestral Character State Reconstruction—The seven loci were aligned using ClustalX. Individual alignments were checked manually, then assembled into a single National Biomedical Research Foundation (NBRF)-format file, which was run through Gblocks (http://molevol.cmima.csic.es/castresana/Gblocks.html; Talavera and Castresana 2007; Castresana 2000) to identify regions of ambiguous alignment using the least stringent settings (minimum number of sequences for a conserved position: 38; minimum number of sequences for a flanking position: 38; maximum number of contiguous nonconserved positions: 8; minimum length of a block: 5; allowed gap positions: with half), as using more stringent settings resulted in most of the data being excluded; the 69 selected blocks were included in the analyses. The resultant matrix includes both coding and noncoding regions and is available from TreeBase, accession number S11390. The model of molecular evolution with the best fit to the data were calculated using jModeltest 0.1.1 (Posada, 2008, using PhyML [Guindon and Gascuel, 2003]) using a fixed BioNeighbour Joining tree estimated using Jukes Cantor distances (BioNJ-JC) for 11 substitution schemes, base frequencies +F, Rate variation +1, four categories (i.e., testing the maximum number of models: 88 models). The Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and hierarchical Likelihood Ratio Testing (hLRT) all selected the General Time Reversible model with invariant positions and a gamma rate distribution (GTR+I+Γ).

Prior to analyzing the combined data from all seven loci, a maximum parsimony (MP) bootstrap analysis was performed for each locus to rule out conflicting phylogenetic signal, indicated by alternative supported topologies. Data from all loci were combined for all subsequent analyses due to the lack of conflict detected among loci. Ten independent replicate GARLI 0.951-GUI (Zwickl, 2006) runs were used to obtain a maximum likelihood (ML) topology, using the GTR+I+Γ model, with parameters estimated from the data. GARLI was also used to obtain ML bootstrap (BS) estimates, with 250 BS replicates. MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) was run with the GTR+I+Γ model of molecular evolution as selected by jModeltest, with 2 parallel runs of 8 chains running for 5 million generations, sampling every 1000th generation. A burnin of 50 000 generations was deleted per run, and the two sets of tree files were checked for compatibility and then combined to maximize effective sample sizes using Tracer v.1.5 (Rambaut and Drummond, 2007). Analyses using the MP criterion were run using PAUP* version 4.0a133 (Swofford, 2003); 1000 random addition replicates were used to find the most parsimonious trees (MPTs) for the matrix. Estimates of parsimony support were generated with 1000 bootstrap pseudoreplicates, with 25 random-addition replicates per bootstrap pseudoreplicate, and saving no more than 25 trees per step.

The *cys*A region was scored according to the presence or absence of an ORF, and entered into Mesquite 2.6 (Maddison and Maddison, 2009) as a linked matrix (available from TreeBase, accession number S11390). Maximum parsimony was used to trace presence of an intact *cys*A ORF, or its absence (referred to as 'loss' below, indicating either pseudogene or loss, here denoted as ψ/-), across the ML phylogeny. Equally weighted and various asymmetrically weighted character transformation models for loss vs. gain were compared in Mesquite 2.6 (Maddison and Maddison, 2009). The distribution of an ORF for *cys*A was mapped under two likelihood models. The Markov single parameter (Mk1) likelihood model was used to reconstruct character-states under a single rate of change. The effect of altering the rate of loss as opposed to gain of an ORF was examined using the Asymmetric 2-parameter (A2P) likelihood model, which allows different rates to be set for character gain and loss: the rates of forward and backward character transformation of "ORF" to "ψ/-" were estimated with the root state frequencies set to the same as the matrix equilibrium. To compensate for the fact that setting the forward and backward rates of character transformation at 1:1 in Mesquite does not generate the same reconstruction as setting them to 2:2 (as in the latter case the rate of evolution is twice as high), the estimated forward rate (12.239) was divided by the estimated reverse rate (7.593) to calculate kappa, *K*, (1.612). A scaling factor used within Mesquite (9.640) was calculated by multiplying the reverse rate by the square root of *K* (or by dividing the forward rate by the square root of *K*). This constant scaling factor was then used to calculate the corresponding forward and reverse rates for a range of *K* values (0.0001, 0.002, 0.005, 0.001, 0.01, 1, 10, and 100). Using a range of *K* values reflects the relative contribution of branch lengths to the reconstruction of character states, and reflects whether character state transitions occur in a more gradualistic or more punctuational manner (Pagel, 1994; Hardy, 2006). A low value of *K* can be interpreted as character state transitions being associated largely with branching events (speciation) (Hinchliff and Roalson, 2009). Reconstructions of character state likelihood were made across this range of *K* values to see the effects on overall tree likelihood and individual node reconstruction of increasing the penalty for secondary regain of the ORF. Given that the restoration of a pseudogene or gene loss in the plastid genome has not been shown, the penalty on ORF regain was increased in iterative reconstructions, until the only change shown was "ORF" to "ψ/-."

Analysis of Substitution Rates—To test whether the pseudogenization and loss of *cys*A commonly occurs in groups with an accelerated rate of substitution in the plastid genome, rates of synonymous substitutions per synonymous site (dS) and nonsynonymous substitutions per nonsynonymous site (dN) were estimated using the beta release of HyPhy version 2.0 (Kosakovsky Pond et al., 2005). Substitution rates were estimated for the combined *rbc*L, *rps*4, and *psb*A loci using the MG94xHKY85_3 × 4 codon model of evolution as implemented in HyPhy. This method requires a hypothesis of relationships among the sequences, represented here by the phylogeny generated for this study. To test whether substitution rates vary among lineages, or among taxa characterized by a particular state of *cys*A, five taxon-partitioning *schemes* were created (as opposed to *partitions*, which instead refers here to each set of taxa resulting from the partitioning itself): 1) complex thalloid and noncomplex thalloid (two taxon partitions); 2) complex thalloid, simple thalloid I, simple thalloid II, leafy (four taxon partitions); 3) complex thalloid, simple thalloid I, simple thalloid II, leafy I, leafy II (five taxon partitions); 4) ORF, pseudogene/loss (two partitions); and 5) ORF-complex thalloid, ORF-noncomplex thalloid, pseudogene/loss (three partitions). Initially, dS and dN were estimated "locally," that is, they were estimated for each terminal and internal branch, and the likelihood of the data was calculated. For each taxon-partitioning scheme defined by monophyletic lineages (i.e., schemes 1, 2, and 3), dS and dN (independently) were constrained to be equal for all terminal and internal branches within each partition. Similarly, dS and dN were constrained to be equal for all terminal branches within each partition in schemes 4 and 5; the constraint on internal branches was determined according to the results of the ancestral character state reconstructions for the loss or retention of *cys*A. With constraints in place, the likelihood of the data was recalculated for each scheme and a Likelihood Ratio Test was carried out in HyPhy for each pair of schemes to determine whether the likelihood is significantly better for more highly partitioned schemes. The value of dN and dS estimated for each partition can then be used to infer directionality of differences in substitution rates (e.g., do complex thalloid liverworts evolve with a lower rate of nonsynonymous substitutions than noncomplex thalloids?).

In addition to estimating substitution rates for three plastid genes universally retained by liverworts, we estimated substitution rates and selective pressure on *cys*A itself. For those taxa characterized by an intact *cys*A (i.e., with an uninterrupted ORF), the *cys*A coding sequences were aligned by eye in MacClade 4.07 (Maddison and Maddison, 2000), using the *Marchantia polymorpha* sequence extract from its plastid genome as a guide. The phylogeny of the intact sequences was inferred using RAxML v 7.0.4 under the GTR+I+Γ model of sequence evolution (Stamatakis et al., 2008), with the intact copy of *cys*A from the green alga *Zygnema* C. Agardh specified as the out-group; the resultant tree was

used as the phylogenetic hypothesis for subsequent estimations of dN and dS and is available from TreeBase (accession number S11390). The ratio of dN to dS (ω) was calculated using HyPhy to infer whether the gene is evolving under purifying selection (ω < 1). The rates of substitution and ω were estimated locally (independently for all terminals and internal branches) using the MG94x-HKY85_3 × 4 codon model of evolution, and re-estimated using several taxon-partitioning schemes: 1) liverworts, *Zygnema*; 2) complex thalloid, noncomplex thalloid, *Zygnema*; 3) complex thalloid, *Cyathodium* Kunze, noncomplex thalloid, *Zygnema*; and 4) complex thalloid, *Cyathodium*, leafy, simple thalloid, *Zygnema*. *Cyathodium* was segregated from other complex thalloids here based on preliminary analyses that suggested its substitution rate was uncharacteristically elevated compared to other complex thalloids. For each taxon-partitioning scheme, ω was constrained to be equal within each partition, and the likelihood of the data was recalculated for each scheme. Likelihood Ratio Tests were performed, as implemented in HyPhy, to determine whether the likelihood of the data was significantly better as the number of partitions increased.

RESULTS

State of cysA—The length of the total *trn*T-GGU to *trn*E-UUC amplicon varied from c. 270 bp (*Odontolejeunea lunulata*) to c. 1640 bp (*Dumortiera hirsuta*); this variation is due both to the presence or absence of an intact *cys*A and to variation in the spacer regions between the gene and its flanking tRNAs. In two genera (*Exormotheca pustulosa* and *Peltolepis quadrata*), a partial sequence was obtained from an amplicon whose migration on the gel was consistent with the migration distance of an amplicon known to be from a *cys*A ORF. As the partial sequences in these cases could be translated and aligned to a full intact gene without interruption of the reading frame, we considered this evidence of an ORF. Furthermore, for *E. pustulosa*, overlapping sequences from two separate accessions could be combined to generate the entire *cys*A ORF. With the addition of sequences from complete plastid genomes, an ORF was detected in 26 of the 63 surveyed liverworts, a pseudogene was detected in 25, and a gene loss was detected in 12. The length of the complete *cys*A sequences varied from 359 to 373 amino acids.

Primer pairs cysAintF and cysAintR, and cysAFnew and cysARnew, successfully amplified the *cys*A region from plants that were already known to have a full-length functional chloroplast *cys*A gene but were not successful in amplifying the region from any plants that lacked a functional chloroplast copy. All accessions sampled from the complex thalloid liverworts, or Marchantiopsida, are characterized by the presence of an intact and full-length *cys*A. The Porellales (leafy I clade) are characterized by the universal absence of an intact *cys*A (Fig. 1).

Phylogeny—The aligned matrix included a total of 5779 characters. Within the liverworts, 2428 (42%) of these characters were variable and 1742 (30.1%) were parsimony informative. This comprised data from: LSU (810 characters, 33.3% variable, 20.6% informative); *nad*1 (945, 27%, 16.2%); *nad*5 (855, 41.2%, 19.5%); *psb*A (1110, 37.7%, 28.3%); *psb*T (334, 61.1%, 47.3%); *rbc*L (1140, 44.4%, 37.5%); and *rps*4 (585, 72.3%, 60.7%).

Maximum parsimony analyses recovered 30 MPTs of length 11815, consistency index 0.3570 (0.3060 when uninformative characters are excluded), retention index 0.5547, and rescaled consistency index 0.1980. ML analyses converged on a topology with –Ln 63871.596.

The phylogenetic relationships among the sampled taxa (Fig. 1) largely agrees with previous analyses (e.g., Heinrichs et al., 2005; Forrest et al., 2006; He-Nygrén et al., 2006) and recovers the major lineages of liverworts (Marchantiopsida or 'complex thalloids', Pelliidae or 'simple thalloid I', Metzgeriidae or 'simple thalloid II', Porellales or 'leafy I', and Jungermanniales or 'leafy II') (Fig. 1). These clades are all well supported (here defined as > 75% bootstrap support and > 0.95 posterior probability) with the exception of the branch subtending all leafy liverworts (not supported by the MP bootstrap), the branch subtending leafy II (low support from the maximum parsimony bootstrap), and the deep branches of leafy I, which receive low to no support from any method. Within the complex thalloid clade, many of the internal relationships are poorly supported. This correlates with short internal branch lengths for most of the nodes (Fig. 2).

Ancestral character-state reconstruction—Using unweighted parsimony, the pattern of observed intact genes, pseudogenes, and gene losses requires 13 steps, and at least eight changes from 'ψ/-' to 'ORF', while three nodes in the leafy II clade are reconstructed unequivocal losses from 'ORF' to 'ψ/-'. With 'ψ/-' to 'ORF' weighted two times higher, there are 18 steps and no changes from 'ψ/-' to 'ORF', although two nodes are equivocal. If the weighting is increased to five times, then there are 18 steps and no regains of 'ORF' from 'ψ/-'.

When likelihood-based methods of ancestral character-state reconstruction are used, under the MK1 model (which uses a single rate of character gain and loss; $K = 1$; rate estimated from the data 7.549; ML under this model = –lnL 35.545), there are at least eight changes from 'ψ/-' to 'ORF', with other nodes equivocal; no unequivocal changes from 'ORF' to 'ψ/-' are seen (reconstruction not shown). Separating the rates of character gain and loss, using the A2P model, results in a more likely reconstruction (–lnL 35.0841, $K = 1.677$, forward rate 9.776, reverse rate 5.830). This reconstruction shows at least ten changes from 'ψ/-' to 'ORF'; two unequivocal changes from 'ORF' to 'ψ/-' are also seen, within the Leafy II clade. Only when K is 0.001 (forward rate 0.239; reverse rate 238.723; –lnL 249.232) are no regains of ORF from a pseudogene state seen (using a 95% likelihood significance value); instead, the character state reconstruction shows 29 separate losses of the *cys*A ORF (Fig. 2). The –lnL value for the reconstruction that lacks 'ψ/-' to 'ORF' transitions is over seven times that for the reconstruction that Mesquite calculates to be most likely (–lnL 249.232, as opposed to –lnL 35.0841).

Substitution rates—We estimated the background rates of synonymous and nonsynonymous substitutions (dS and dN) for a combined data set of three plastid genes (*rbc*L, *rps*4, and *psb*A), and calculated the likelihood of the data when dS and dN were constrained within partitions of five taxon-partioning schemes. Partitioning the accessions into complex thalloids and noncomplex thalloids resulted in a significantly better likelihood than the likelihood of the data under the globally estimated rates (Table 1). In each pair-wise comparison of the five partitioning schemes (every comparison was performed), the scheme with more partitions was significantly more likely than the scheme with fewer partitions. The values of dN and dS for each partition within each scheme allowed us to compare substitution rates among lineages (or partitions). For example, noncomplex thalloids have a higher rate of both dN and dS than complex thalloids (scheme 1 in Table 1). Under all partitioning schemes, the complex thalloids were estimated to have the slowest dS, whereas the simple thalloid II liverworts were estimated to have the fastest dS. Only the leafy II liverworts, under

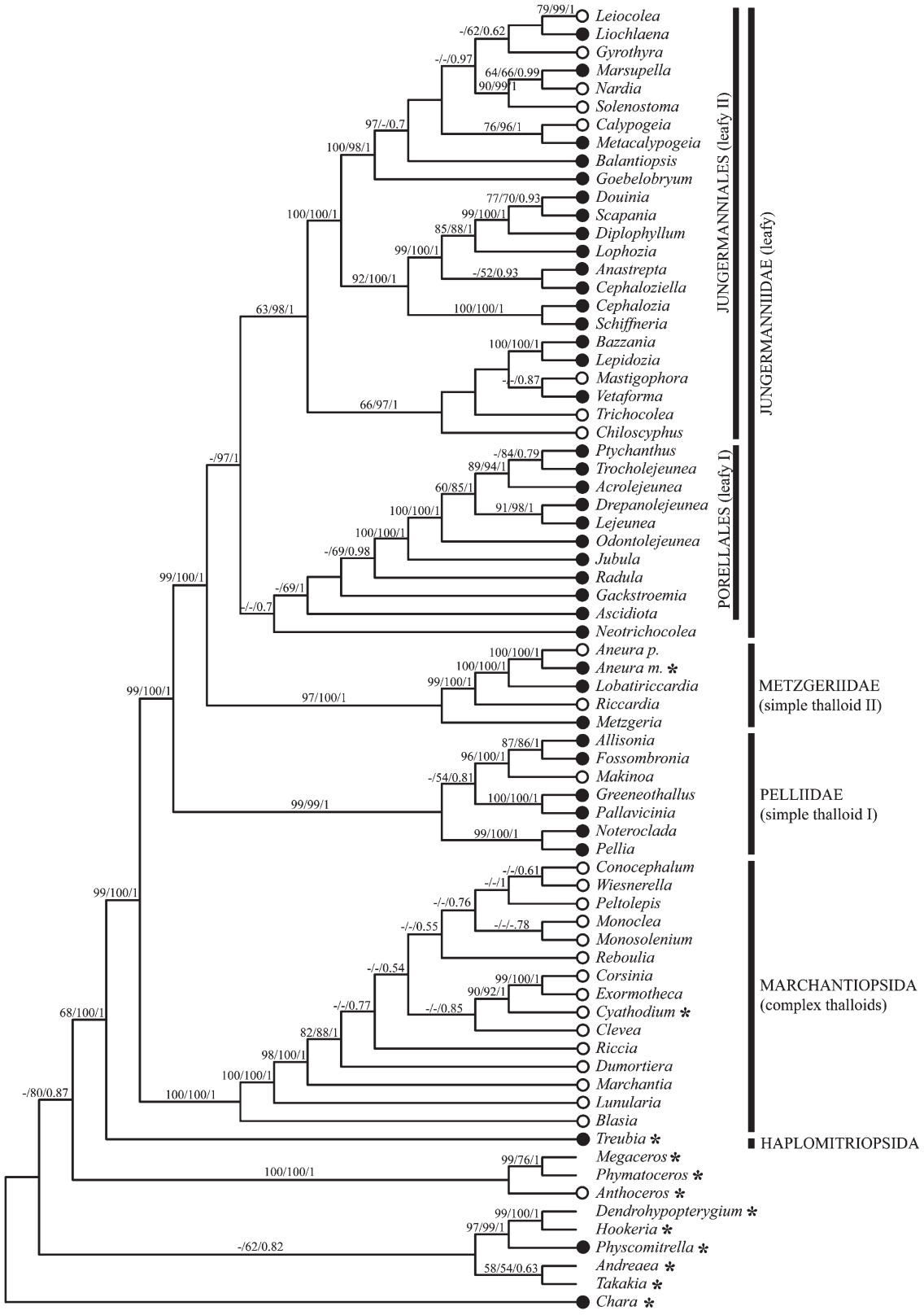

Fig. 1. Hypothesis of relationships of genera included in this study, inferred from sequence data from four chloroplast loci (*psb*A, *psb*T, *rbc*L, and *rps*4), two mitochondrial loci (*nad*1 and *nad*5) and the nuclear ribosomal large subunit. Maximum parsimony and maximum likelihood bootstrap support values and posterior probabilities of clades respectively are shown above the relevant branches. Open circles denote taxa that have an intact ORF, while black circles denote taxa that do not have an ORF, for *cys*A. Asterisks denote taxa that were not included in reconstructions of rates of nonsynonymous and synonymous substitutions per nonsynonymous and synonymous sites (see Table 1).

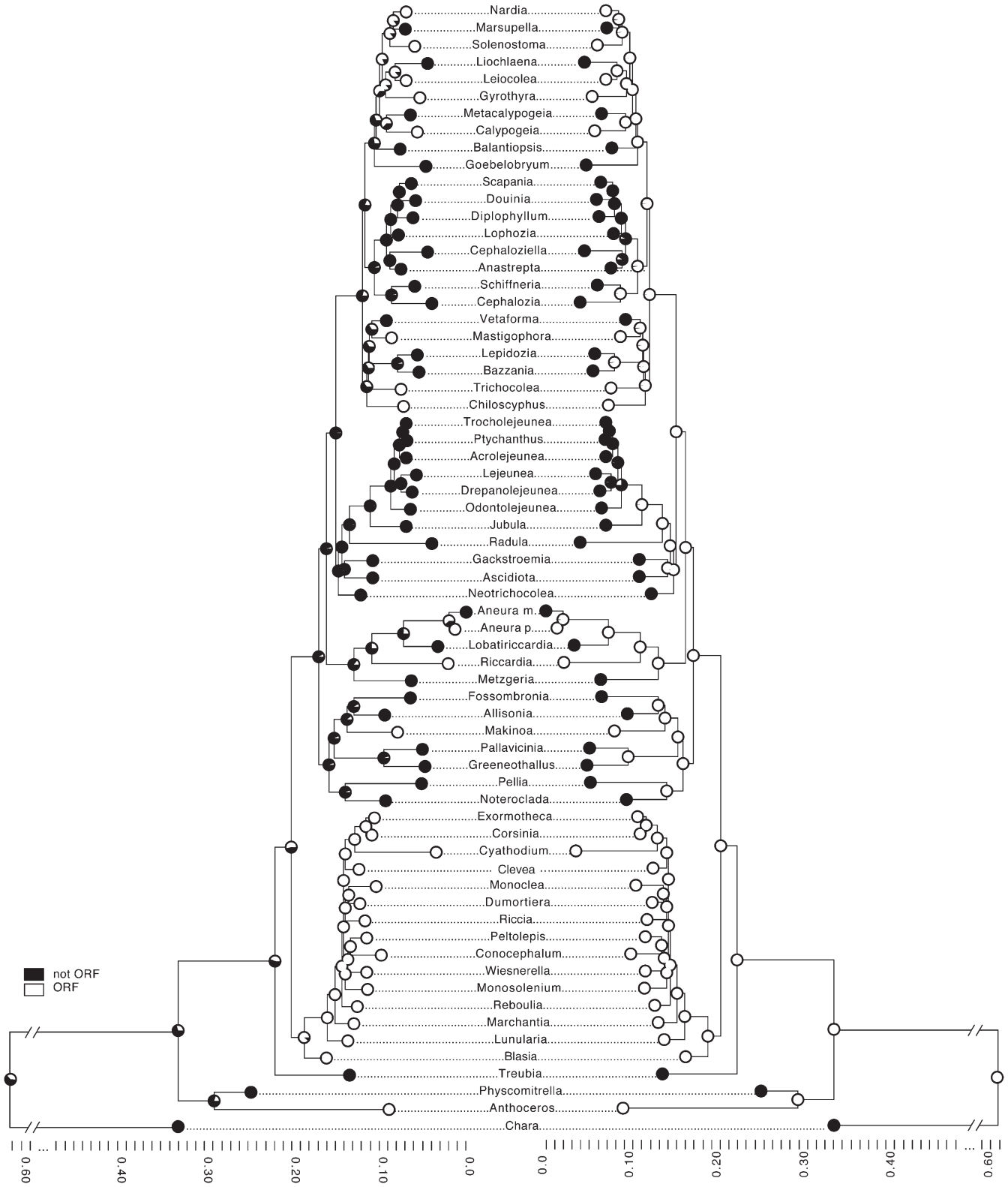

Fig. 2. Character-state likelihood reconstruction for the loss (black) or retention (open) of an intact ORF for *cys*A in the plastid genome under an Asymmetric 2 parameter model. Topology and branch lengths generated by maximum likelihood analysis of sequence data from four chloroplast genes (*psb*A, *psb*T, *rbc*L, and *rps*4), two mitochondrial genes (*nad*1 and *nad*5) and the nuclear ribosomal large subunit. The pie charts on the left-hand phylogram represent the likelihood of character transformation with all parameters estimated for the data, while the pie charts on the right-hand phylogram represent the likelihood of character transformation when kappa (*K*) is scaled so that the regain of an ORF after loss is impossible. Left-hand phylogram: kappa (*K*) = 1.677 (–lnL 35.084). Right-hand phylogram: kappa (*K*) = 0.001 (–lnL 249.232).

TABLE 1. Rates of nonsynonymous (dN) and synonymous (dS) substitutions per nonsynonymous and synonymous site, respectively, for three plastid genes, *psb*A, *rps*4, *rbc*L, and five different taxon-partitioning schemes. Due to paraphyly, simple thalloids are not included as an autonomous partition. Both *Cyathodium tuberosum* and *Aneura mirabilis* were removed from analyses due to known acceerated rates of evolution uncharacteristic for their respective higher order groups. *Treubia* and all out-groups were removed due to their status as singleton partitions (i.e., cannot be informatively placed into a taxon partition with multiple members). Three asterisks indicate a strongly significant ($P < 0.005$) result of the likelihood ratio test.

	Substitution Rates			Likelihood Ratio Test				
Partition	dN	dS	Likelihood	5	4	3	2	1
Local			−30470.3					
global (A)				***	***	***	***	***
liverworts	0.017443	0.25429	−32338.2					
two partitions (scheme 1)				***	***	***	***	
complex thalloid	0.014287	0.11836	−32091.8					
noncomplex thalloid	0.018441	0.29842						
four partitions (scheme 2)				***	***	***		
complex thalloid	0.014289	0.11778	−31711.2					
simple thalloid I	0.028948	0.50805						
simple thalloid II	0.030325	0.58426						
leafy	0.014347	0.22114						
five partitions (scheme 3)				***	***			
complex thalloid	0.014289	0.11778	−31700.0					
simple thalloid I	0.028953	0.50804						
simple thalloid II	0.030331	0.53589						
leafy I	0.018846	0.21153						
leafy II	0.012472	0.22487						
two partitions (scheme 4)				***				
gene loss	0.022983	0.33002	−32216.2					
ORF	0.014260	0.20825						
three partitions (scheme 5)								
gene loss	0.023072	0.31971	−32060.8					
ORF—complex thalloid	0.014298	0.11838						
ORF—other	0.014329	0.27646						

the most highly partitioned scheme (scheme 3 in Table 1), were estimated to have a slower dN than the complex thalloids. The fastest dN was estimated at 0.0303 for the simple thalloid II clade (schemes 2 and 3 in Table 1).

In the single comparison of taxon-partitioning schemes that resulted in the same number of partitions (two partitions, schemes 1 and 4 in Table 1; loss (or pseudogene) vs. ORF and complex thalloid vs. noncomplex thalloid, respectively), the scheme based on relatedness (scheme 1 in Table 1) resulted in a significantly better likelihood than the partition based on the state of *cys*A (scheme 4 in Table 1). However, further partitioning of taxa characterized by an intact *cys*A (ORF) into complex thalloids and noncomplex thalloids resulted in a significantly better likelihood than simply partitioning all taxa into these two groups (scheme 5 vs. scheme 1 in Table 1). In both partitioning schemes for which the accessions were partitioned based on the state of *cys*A (schemes 4 and 5 in Table 1), those taxa characterized by either a pseudogene or a loss were estimated to have a faster dN and a faster dS. Noncomplex thalloids with an intact *cys*A had a negligibly faster dN and a substantially faster dS than the complex thalloids.

We estimated the ratio of dN to dS (dN/dS = ω) for intact *cys*A sequences (ORFs) with sufficient alignable length (26 sequences; Fig. 2). The tree estimated for these *cys*A sequences (Fig. 3) resolved the three major clades of liverworts represented (complex thalloids, simple thalloids, leafy II), however many branches were unsupported within these clades, suggesting that *cys*A evolves too slowly to be phylogenetically informative at this level. For all lineages, *cys*A appears to be evolving under purifying selection ($\omega < 1$; Fig. 3). Constraining ω to be equal within all complex thalloids, and equal within all noncomplex thalloids results in a significantly better likelihood than when ω is globally constrained (Table 2); schemes 1 and 2). Allowing ω to be estimated independently in *Cyathodium* does not result in a significantly better likelihood (schemes 2 and 3 in Table 2). If ω is estimated independently in leafy and simple thalloid liverworts, there is a weakly significant improvement in the likelihood (scheme 4 in Table 2). At the clade level, the simple thalloid clade is characterized by the highest ω, whereas the complex thalloids have the lowest ω (Table 2).

DISCUSSION

Structural rearrangements and the composition of plastid genomes can provide a powerful set of phylogenetic markers, but must sometimes be treated cautiously due to asymmetry in their character-state transitions. Pseudogenization and gene loss, at least in the plastid genome, occur much more frequently than gene gain. In this study we examined the evolution of the seemingly labile *cys*A gene in the plastid genome of liverworts. Of 63 liverworts surveyed, only 26 retain an intact, or inferred intact, *cys*A ORF (Fig. 1). Only the Marchantiopsida (complex thalloids) and Porellales (Leafy I) surveyed here are all characterized by a single state of *cys*A (intact and loss (pseudogene, missing between flanking tRNAs, or nonamplifiable), respectively); however, the monophyly of these clades is well supported in other analyses (e.g., Heinrichs et al., 2005; Forrest et al., 2006; He-Nygrén et al., 2006), suggesting that, at this level, the presence or absence of *cys*A has no additional phylogenetic signal. Although plastid gene content and order has been shown to be phylogenetically informative in other bryophytes (see Goffinet et al., 2007), the presence or absence of *cys*A does not appear to add support to unsupported, or

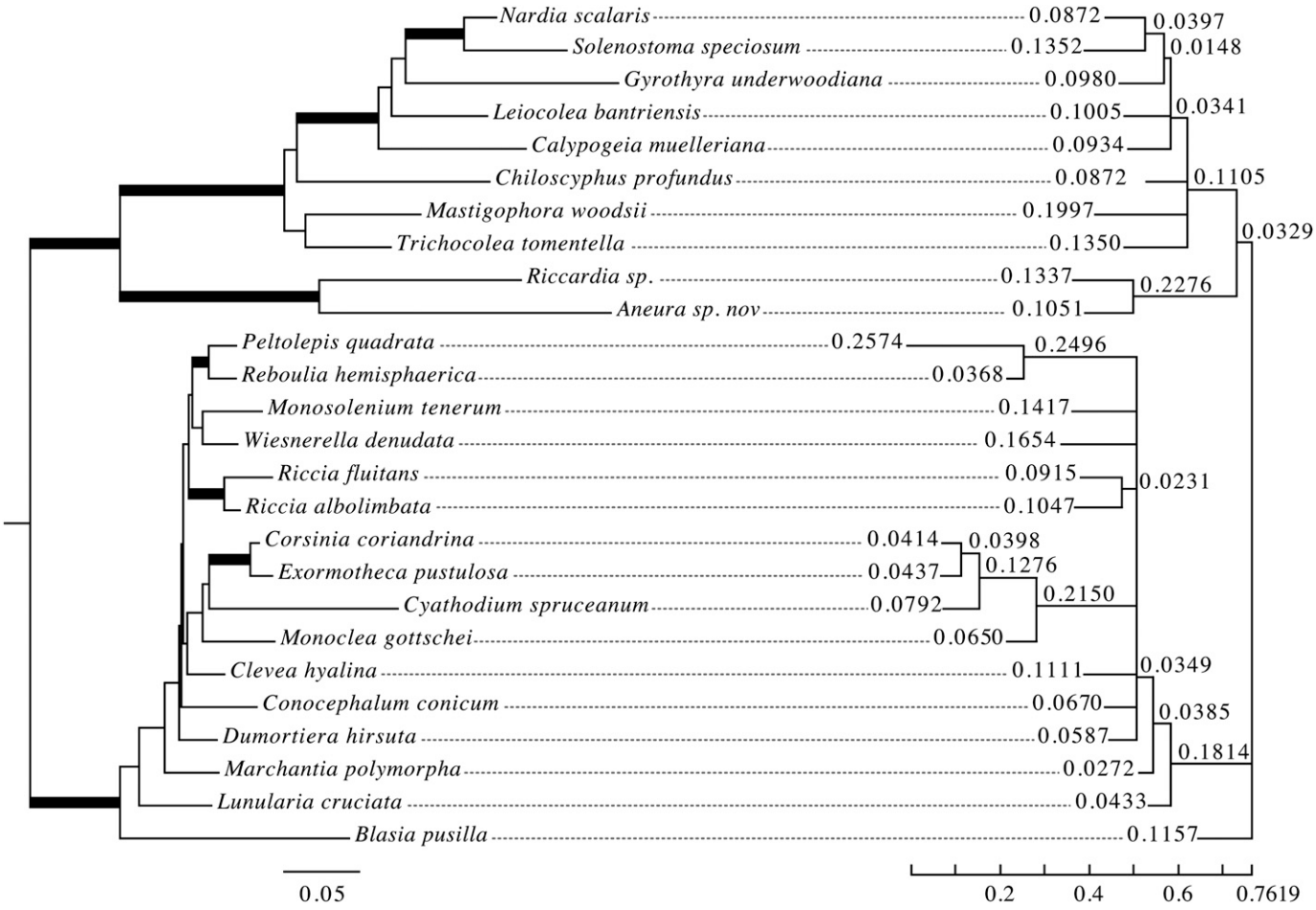

Fig. 3. Left: Maximum likelihood (ML) (1000 bootstrap replicates) phylogenetic reconstruction of intact *cys*A gene sequences (ORFs), with highly supported (> 75% bootstrap support) branches indicated in bold. *Makinoa* Miyake is excluded due to the incomplete nature of its *cys*A sequence. Right: the ML tree scaled by the ratio of the nonsynonymous substitution rate (dN) to the synonymous substitution rate (dS), or omega (ω; values above branches), which was estimated with no constraints applied to a taxon-partitioning scheme (values estimated "locally").

weakly supported, clades for the taxa sampled in this study. The frequency with which this gene is lost from the plastid genome of liverworts suggests that this would be an unreliable character to use when inferring relationships. Broader sampling from lower order taxa, particularly at the level of genus and species, is required to determine whether the presence or absence of *cys*A is an informative character within and among genera.

Reconstructing the ancestral states of *cys*A under MP and ML can require the invocation of both gains and losses. Under both optimality criteria, gains of *cys*A outnumber losses in reconstructions for which gains and losses are weighted equally, or losses are slightly up-weighted. Only when losses are heavily up-weighted do the reconstructions fail to explain the data through gene gains. Under this asymmetric weighting, five more steps are need to explain the data under MP, and a likelihood more than seven-fold greater than the unweighted reconstruction is needed to explain the data. However, the asymmetric reconstructions are the most biologically plausible, given that the reacquisition of a gene lost from the plastid genome in a shared common ancestor has never been reported. In fact, the maturase gene *mat*K, required for successful splicing of plastid introns, appears to be the only gene gain in plastids (Turmel et al., 2005), though this is not a case where a gene previously lost was reacquired.

When the likelihood reconstruction is constrained to allow only losses, 29 independent losses of *cys*A must be invoked (Fig. 2). The loss of a gene from the plastid genome is not necessarily uncommon; however, multiple losses at a rate comparable to *cys*A are rare. One example, in angiosperms, is the loss of a plastid-encoded *inf*A that occurred at least 24 separate times (Millen et al., 2001). Clearly, *cys*A is prone to being lost from the plastid genome, possibly indicating that its function may be either unnecessary or made redundant by alternative transporter genes in those lineages for which *cys*A is retained as an intact ORF. Regardless, *cys*A is retained in the plastid genome of all complex thalloid liverworts surveyed, despite over 400 million years of evolution having passed since they diverged from an ancestor shared with a liverwort lacking *cys*A (e.g., *Treubia* Goebel) (Heinrichs et al., 2007; Wikström et al., 2009).

The dN/dS ratio (ω) is often used to reflect whether selection is acting on a particular gene, where a ratio less than one indicates purifying selection, a ratio equal to one indicates no selection, and a ratio greater than one indicates positive selection. Though it remains much less than one, ω tends to increase for *cys*A clade-wide in the noncomplex thalloid liverworts (Table 2).

TABLE 2. Ratio of nonsynonymous substitutions per synonymous site (dN) to synonymous substitutions per synonymous site (dS) (ω) for intact open reading frames (ORFs) of the plastid gene *cys*A (see Fig. 3). The likelihood of the data given a hypothesis of relationships (phylogeny) was calculated for four taxon-partioning schemes; a likelihood-ratio test was used to determine whether increasing the number of taxon partitions resulted in a significantly better likelihood score. A single asterisk represents a weakly significant result, and "ns" indicates no significant difference.

Partition	Omega	Likelihood	*p*-value	Significance
global (scheme 1)				
liverworts	0.08903	−7709.18	n/a	n/a
two partitions (scheme 2)				
complex thalloid	0.07556			
noncomplex thalloid	0.09933	−7706.65	0.0245	*
three partitions (scheme 3)				
complex thalloid	0.07494			
Cyathodium	0.08082			
noncomplex thalloids	0.09942	−7706.61	0.7846	ns
four partitions (scheme 4)				
complex thalloid	0.07489			
Cyathodium	0.08078			
leafy	0.08995			
simple thalloid	0.12893	−7704.48	0.0386	*

This slight increase in ω for the noncomplex *cys*A sequences is only weakly significant, as indicated by the significantly better likelihood when the aligned *cys*A sequences are partitioned into complex and noncomplex thalloid liverworts. Furthermore, some lineages of complex thalloids actually have a higher value of ω compared to the simple thalloids (Fig. 3). However, based on the results of the partitioned analyses of ω, we might infer that selection may be very slightly relaxed on *cys*A in those noncomplex thalloid liverworts that retain a *cys*A ORF. This is consistent with the fact that *cys*A is lost repeatedly throughout the noncomplex thalloid liverworts, whereas it is universally retained in the complex thalloids. Whether *cys*A is actively functional requires, at a minimum, sequencing of plastid RNA to determine whether it is transcribed.

The frequent loss of *cys*A in noncomplex thalloid liverworts is loosely associated with accelerated rates of both synonymous and nonsynonymous rates of substitution in protein-coding plastid genes. The rate of synonymous substitutions per synonymous site is significantly higher for noncomplex thalloids in three plastid genes measured here (Table 1), and rates of molecular evolution in mitochondrial and plastid genes have previously been observed to be slower in the complex thalloid lineage than in other liverwort lineages (Forrest et al., 2006). If, in lineages characterized by a pseudogene or gene loss, *cys*A is no longer under strong selective constraints, the accumulation of slightly deleterious mutations may occur at a higher rate in noncomplex thalloids, possibly leading to the observed pattern of frequent pseudogenizations and gene losses in this group. Indeed, three protein coding plastid genes are shown here have a higher rate of both dN and dS in those taxa for which there is no *cys*A ORF (Table 1), suggesting that pseudogenization and loss of *cys*A may be correlated with an increase in the rate of substitution of the plastid genome, though this certainly does not imply that this is the sole cause, if at all.

The transfer of plastid DNA to the nucleus is thought to occur at a high rate (Stegemann et al., 2003), and plastid gene losses have been shown to co-occur with the transfer of that gene to the nuclear genome, for example, *inf*A (Millen et al., 2001), *rpl*32 (Ueda et al., 2007), and *rpo*A (Sugiura et al., 2003). Rather than multiple independent transfers of *cys*A to the nuclear genome, we suspect that sulfate import into the chloroplast is maintained either by gene substitution, where a nuclear gene whose function may be homologous to that of *cys*A compensates for the lack of the plastid gene, or by an ancient transfer, and subsequent divergence, of the plastid gene. Degenerate PCR primers, designed to anneal within the coding region of liverwort *cys*A sequences, do not produce an amplicon from total DNA extracted from those liverworts that do not have a plastid copy of the gene. Furthermore, a BLAST search (www.ncbi.nlm.nig.gov/BLAST) (tBLASTn, E-value 1e-5) against the *Physcomitrella patens* genome, a moss that lacks a plastid copy of *cys*A, returned no hits with greater than 45% identity to the *Marchantia polymorpha cys*A sequence. Of the 15 hits in the *P. patens* genome with percent identity greater than 40 and less than 45, only one could be aligned with the *M. polymorpha cys*A sequence for more than 200bp (215bp); these relatively weak hits are likely the result of a high degree of similarity in the conserved ABC transporter domain shared between *cys*A and unrelated genes.

Sulfate reduction, required for the production of cysteine, occurs in green plant plastids, which can provide the large amount of electrons and energy required for this reaction via photosynthesis (Hell, 1997; Chen et al., 2003). It has been hypothesized that *cys*A in *Marchantia polymorpha* encodes the component of a sulfate-transport system that binds ATP to provide the energy required for transport; these ATP-binding proteins belong to a protein superfamily that include ABC-type transporters (Leustek et al., 2000 and references therein). In the cyanobacterium *Synechococcus* sp., *cys*A encodes an ATP-binding protein in the sulfate-transport system that is 50% conserved at the amino acid level to *cys*A in *M. polymorpha* (annotated as *mbp*X) (Laudenbach and Grossman, 1991). The fact that cysteine synthesis occurs in plants lacking a plastid-encoded *cys*A suggests that nuclear-encoded proteins with plastid-transit peptides exist to import sulfate into plastids, though they have yet to be identified in embryophytes (Kopriva et al., 2008; Lindberg and Melis, 2008). However, homologous sequences of *cys*A, *cys*W, *cys*T, and Sbp (encoding structural components of the sulfate permease complex in cyanobacteria) have been identified from the nuclear genome of *Chlamydomonas reinhardtii* (Melis and Chen, 2005; Lindberg and Melis, 2008). The existence of these nuclear-encoded genes, or the existence of an alternative, plastid-localized sulfate-transport system may be uncovered as functional and comparative genomics becomes more accessible and cost-effective. Evidence suggests that plastid-encoded *cys*T follows an evolutionary trajectory similar to *cys*A (Wickett

et al., 2008a); additional evidence from liverwort plastid genomes, expressed sequence tags, and nuclear genomes may allow us to uncover the fate of *cys*T, *cys*W, and Sbp in more detail. With a complete nuclear genome sequence of *M. polymorpha* on the horizon, a comprehensive gene family phylogeny and analysis may be able to resolve these questions with more certainty.

Summary—The plastid-encoded, putative sulfate-transport protein coding gene *cys*A has been lost independently at least 29 times throughout the evolution of liverworts (Fig. 2). Of the 63 liverworts surveyed for the presence of *cys*A, the complex thalloid liverworts (Marchantiopsida) are the only major clade to be characterized by the universal presence of an intact ORF; Porellales (leafy I) is the only clade characterized by the universal absence of an intact *cys*A. Three other major liverworts clades, simple thalloid I & II and leafy II, all comprise taxa both with and without an intact copy of the gene. Based on estimated rates of nucleotide substitution from three other plastid genes, it appears that the complex thalloid liverworts are evolving at a slower rate than other liverworts (Table 1). However, the leafy I clade does not appear to be evolving more rapidly than those clades comprising taxa with both states of *cys*A. That said, those taxa lacking an intact *cys*A appear to be evolving at a faster rate of nucleotide substitution than those noncomplex thalloids that possess an intact *cys*A (Table 1). Regardless of clade membership, all intact copies of *cys*A that were sufficiently sequenced appear to be evolving under purifying selection ($\omega < 1$), though a *slight* relaxation of this selection pressure may occur in the noncomplex thalloid liverworts (Table 2; Fig. 3). Whether *cys*A is expressed and functional in any clade of liverworts, or whether a functional homolog exists in the nuclear genome, will require extensive transcriptome and genome analysis, the source data for which are not yet available.

LITERATURE CITED

CASTRESANA, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17: 540–552.

CAVALIER-SMITH, T. 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *International Journal of Systematic and Evolutionary Microbiology* 52: 297–354.

CHEN, H., K. YOKTHONGWATTANA, A. J. NEWTON, AND A. MELIS. 2003. *SulP*, a nuclear gene encoding a putative chloroplast-targeted sulfate permease in *Chlamydomonas reinhardtii. Planta* 218: 98–106.

CHUMLEY, T. W., J. D. PALMER, J. P. MOWER, H. M. FOURCADE, P. J. CALIE, J. L. BOORE, AND R. K. JANSEN. 2006. The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Molecular Biology and Evolution* 23: 2175–2190.

CUI, L., N. VEERARAGHAVAN, A. RICHTER, K. WALL, R. K. JANSEN, J. LEEBENS-MACK, I. MAKALOWSKA, AND C. W. DEPAMPHILIS. 2006. ChloroplastDB: The chloroplast genome database. *Nucleic Acids Research* 34: D692–D696.

DAVIS, E. C. 2004. A molecular phylogeny of leafy liverworts (Jungermanniidae: Marchantiophyta). *In* B. Goffinet, V. Hollowell, and R. Magill [eds.], Molecular systematics of bryophytes, Monographs in Systematic Botany, vol. 98, 61-86. Missouri Botanical Garden Press, St. Louis, Missouri, USA.

DEMESURE, B., N. SODZI, AND R. J. PETIT. 1995. A set of universal primers for the amplification of polymorphic noncoding regions of mitochondrial and chloroplast DNA in plant. *Molecular Ecology* 4: 129–131.

DEPAMPHILIS, C. W., AND J. D. PALMER. 1990. Loss of photosynthetic and chlororespiratory gene from the plastid genome of a parasitic flowering plant. *Nature* 348: 337–339.

DOYLE, J. J., J. L. DOYLE, AND J. D. PALMER. 1995. Multiple independent losses of 2 genes and one intron from legume chloroplast genomes. *Systematic Botany* 20: 272–294.

FORREST, L. L., AND B. J. CRANDALL-STOTLER. 2005. Progress towards a robust phylogeny of the liverworts, with particular focus on the simple thalloids. *Journal of the Hattori Botanical Laboratory* 97: 127–159.

FORREST, L. L., E. C. DAVIS, D. G. LONG, B. J. CRANDALL-STOTLER, A. CLARK, AND M. L. HOLLINGSWORTH. 2006. Unraveling the evolutionary history of the liverworts (Marchantiophyta): Multiple taxa, genomes and analyses. *Bryologist* 109: 303–334.

FORREST, L. L., N. J. WICKETT, C. J. COX, AND B. GOFFINET. 2011. Deep sequencing of *Ptilidium* (Ptilidiaceae) suggests evolutionary stasis in liverwort plastid genome structure. *Plant Ecology and Evolution* 144: 29–43.

GOFFINET, B., N. J. WICKETT, A. J. SHAW, AND C. J. COX. 2005. Phylogenetic significance of the *rpo*A loss in the chloroplast genome of mosses. *Taxon* 54: 353–360.

GOFFINET, B., N. J. WICKETT, O. WERNER, R. M. ROS, A. J. SHAW, AND C. J. COX. 2007. Distribution and phylogenetic significance of the 71-kb inversion in the plastid genome in Funariidae (Bryophyta). *Annals of Botany* 99: 747–753.

GUINDON, S., AND O. GASCUEL. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52: 696–704.

GUISINGER, M. M., J. V. KUEHL, J. L. BOORE, AND R. K. JANSEN. 2011. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: Rearrangements, repeats, and codon usage. *Molecular Biology and Evolution* 28: 583–600.

HARDY, C. R. 2006. Reconstructing ancestral ecologies: challenges and possible solutions. *Diversity & Distributions* 12: 7–19.

HE-NYGRÉN, X., A. JUSLÉN, I. AHONEN, D. GLENNY, AND S. PIIPPO. 2006. Illuminating the evolutionary history of liverworts (Marchantiophyta) – Towards a natural classification. *Cladistics* 22: 1–31.

HEINRICHS, J., S. R. GRADSTEIN, R. WILSON, AND H. SCHNEIDER. 2005. Towards a natural classification of liverworts (Marchantiophyta) based on the chloroplast gene *rbc*L. *Cryptogamie Bryologie* 26: 131–150.

HEINRICHS, J., J. HENTSCHEL, R. WILSON, K. FELDBERG, AND H. SCHNEIDER. 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.

HELL, R. 1997. Molecular physiology of plant sulfur metabolism. *Planta* 202: 138–148.

HINCHLIFF, C. E., AND E. H. ROALSON. 2009. Stem architecture in *Eleocharis* subgenus *Limnochloa* (Cyperaceae): Evidence of dynamic morphological evolution in a group of pantropical sedges. *American Journal of Botany* 96: 1487–1499.

HUELSENBECK, J. P., AND F. RONQUIST. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford, England)* 17: 754–755.

JANSEN, R. K., C. ZHENGQIU, L. A. RAUBESON, H. DANIELL, C. W. DEPAMPHILIS, J. LEEBENS-MACK, K. F. MÜLLER, ET AL. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences, USA* 104: 19369–19374.

KOPRIVA, S., N. J. PATRON, AND P. KEELING. 2008. Phylogenetic analysis of sulfate assimilation and cysteine biosynthesis in phototrophic organisms. *In* R. Hell, C. Dahl, D. B. Knaff, and T. Lustek [eds.], Sulfur metabolism in phototrophic organisms, 31-58. Springer, Dordrecht, Netherlands.

KOSAKOVSKY POND, S. L., S. D. W. FROST, AND S. V. MUSE. 2005. HyPhy: Hypothesis testing using phylogenies. *Bioinformatics (Oxford, England)* 21: 676–679.

KRELLWITZ, E. C., K. V. KOWALLIK, AND P. S. MANOS. 2001. Molecular and morphological analyses of Bryopsis (Bryopsidales, Chlorophyta) from the western North Atlantic and Caribbean. *Phycologia* 40: 330–339.

KUGITA, M., A. KANEKO, Y. YAMAMOTO, Y. TAKEYA, T. MATSUMOTO, AND K. YOSHINAGA. 2003. The complete nucleotide sequence of the

hornwort (*Anthoceros formosae*) chloroplast genome: Insights into the earliest land plants. *Nucleic Acids Research* 31: 716–721.

LAUDENBACH, D. E., AND A. R. GROSSMAN. 1991. Characterization and mutagenesis of sulfur-regulated genes in a cyanobacterium: Evidence for function in sulfate transport. *Journal of Bacteriology* 173: 2739–2750.

LEUSTEK, T., M. N. MARTIN, J. BICK, AND J. P. DAVIES. 2000. Pathways and regulation of sulfur metabolism revealed through molecular and genetic studies. *Annual Review of Plant Physiology and Plant Molecular Biology* 51: 141–165.

LINDBERG, P., AND A. MELIS. 2008. The chloroplast sulfate transport system in the green alga *Chlamydomonas reinhardtii. Planta* 228: 951–961.

MADDISON, D. R., AND W. P. MADDISON. 2000. MacClade version 4: Analysis of phylogeny and character evolution. Sinauer, Sunderland, Massachusetts, USA.

MADDISON, W. P., AND D. R. MADDISON. 2009. Mesquite: A modular system for evolutionary analysis. Version 2.6. http://mesquiteproject.org [accessed 28 September 2009].

MARGULIS, L. 1970. Origin of eukaryotic cells. Yale University Press, New Haven, Connecticut, USA.

MARSHALL, C. R., E. C. RAFF, AND R. A. RAFF. 1994. Dollo's law and the death and resurrection of genes. *Proceedings of the National Academy of Sciences, USA* 91: 12283–12287.

MARTIN, W., T. RUJAN, E. RICHLY, A. HANSEN, S. CORNELSEN, T. LINS, D. LEISTER, ET AL. 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences, USA* 99: 12246–12251.

MARTIN, W., B. STOEBE, V. GOREMYKIN, S. HANSMANN, M. HASEGAWA, AND K. V. KOWALLIK. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393: 162–165.

MCPHERSON, M. A., M. F. FAY, M. W. CHASE, AND S. W. GRAHAM. 2004. Parallel loss of a slowly evolving intron from two closely related families in Asparagales. *Systematic Botany* 29: 296–307.

MELIS, A., AND H. CHEN. 2005. Chloroplast sulfate transport in green algae—Genes, proteins and effects. *Photosynthesis Research* 86: 299–307.

MILLEN, R. S., R. G. OLMSTEAD, K. L. ADAMS, J. D. PALMER, N. T. LAO, L. HEGGIE, T. A. KAVANAGH, ET AL. 2001. Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. *Plant Cell* 13: 645–658.

MOORE, M. J., C. D. BELL, P. S. SOLTIS, AND D. E. SOLTIS. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences, USA* 104: 19363–19368.

OHYAMA, K., H. FUKUZAWA, T. KOHCHI, H. SHIRAI, T. SANO, S. SANO, K. UMESONO, ET AL. 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature* 322: 572–574.

OLIVER, M. J., A. G. MURDOCK, B. D. MISHLER, J. V. KUEHL, J. L. BOORE, D. F. MANDOLI, K. D. E. EVERETT, ET AL. 2010. Chloroplast genome sequence of the moss *Tortula ruralis*: Gene content, polymorphism, and structural arrangement relative to other green plant chloroplast genomes. *BMC Genomics* 11: 143.

OMLAND, K. E. 1997. Examining two standard assumptions of ancestral reconstructions: Repeated loss of dichromatism in dabbling ducks (Anatini). *Evolution; International Journal of Organic Evolution* 51: 1636–1646.

OMLAND, K. E. 1999. The assumptions and challenges of ancestral state reconstructions. *Systematic Biology* 48: 604–611.

PAGEL, M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. B, Biological Sciences* 255: 37–45.

POSADA, D. 2008. jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution* 25: 1253–1256.

RAMBAUT, A., AND A. J. DRUMMOND. 2007. Tracer v1.4. Website http://beast.bio.ed.ac.uk/Tracer [accessed 10 October 2009].

RAUBESON, L. A., AND R. K. JANSEN. 2005. Chloroplast genomes of plants. *In* R. J. Henry [ed.], Plant diversity and evolution: Genotypic and phenotypic variation in higher plants, 45-68. CAB International, Wallingford, United Kingdom.

ROMER, S., A. D'HARLINGUE, B. CAMARA, R. SCHANTZ, AND M. KUNTZ. 1992. Cysteine synthase from *Capsicum annuum* chromoplasts. *Journal of Biological Chemistry* 267: 17966–17979.

RONQUIST, F., AND J. P. HUELSENBECK. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics (Oxford, England)* 19: 1572–1574.

SCHNURR, J. A., J. M. SCHOCKEY, G. J. DEBOER, AND J. A. BROWSE. 2002. Fatty acid export from the chloroplast: Molecular characterization of a major plastidial acyl-coenzyme A synthetase from *Arabidopsis. Plant Physiology* 129: 1700–1709.

STAMATAKIS, A., P. HOOVER, AND J. ROUGEMONT. 2008. A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology* 57: 758–771.

STEGEMANN, S., S. HARTMANN, S. RUF, AND R. BOCK. 2003. High-frequency gene transfer from the chloroplast genome to the nucleus. *Proceedings of the National Academy of Sciences, USA* 100: 8828–8833.

SUGIURA, C., Y. KOBAYASHI, S. AOKI, S. SUGITA, AND M. SUGITA. 2003. Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: Evidence for the loss and relocation of *rpo*A from the chloroplast to the nucleus. *Nucleic Acids Research* 31: 5324–5331.

SWOFFORD, D. L. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer, Sunderland, Massachusetts, USA.

TALAVERA, G., AND J. CASTRESANA. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* 56: 564–577.

TURMEL, M., C. OTIS, AND C. LEMIEUX. 2005. The complete chloroplast DNA sequences of the charophycean green algae *Staurastrum* and *Zygnema* reveal that the chloroplast genome underwent extensive changes during the evolution of the Zygnematales. *BMC Biology* 3: 22.

TURMEL, M., C. OTIS, AND C. LEMIEUX. 2006. The chloroplast genome sequence of *Chara vulgaris* sheds new light into the closest green algal relatives of land plants. *Molecular Biology and Evolution* 23: 1324–1338.

UEDA, M., M. FUJIMOTO, S. ARIMURA, J. MURATA, N. TSUTSUMI, AND K. KADOWAKI. 2007. Loss of the *rpl32* gene from the chloroplast genome and subsequent acquisition of a preexisting transit peptide within the nuclear gene in *Populus. Gene* 402: 51–56.

WICKETT, N. J., Y. FAN, P. O. LEWIS, AND B. GOFFINET. 2008a. Distribution and evolution of pseudogenes, gene losses, and a gene rearrangement in the plastid genome of the nonphotosynthetic liverwort, *Aneura mirabilis* (Metzgeriales, Jungermanniopsida). *Journal of Molecular Evolution* 67: 111–122.

WICKETT, N. J., Y. ZHANG, S. K. HANSEN, J. M. ROPER, J. V. KUEHL, S. A. PLOCK, P. G. WOLF, ET AL. 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.

WIKSTRÖM, N., X. HE-NYGRÉN, AND A. J. SHAW. 2009. Liverworts (Marchantiophyta). *In* S. B. Hedges and S. Kumar [eds.], The timetree of life, 146-152. Oxford University Press, Oxford, United Kingdom.

WOLF, P. G., J. M. ROPER, AND A. M. DUFFY. 2010. The evolution of chloroplast genome structure in ferns. *Genome* 53: 731–738.

WOLFE, K. H., C. W. MORDEN, AND J. D. PALMER. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proceedings of the National Academy of Sciences, USA* 89: 10648–10652.

ZWICKL, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, University of Texas , Austin, Texas, USA.

APPENDIX 1. List of taxa sampled, and GenBank accession numbers.

Taxon; (genes sampled for reference phylogeny [as GenBank accession numbers]): *nad*5; Nr LSU; *nad*1; *rbc*L; *rps*4; *psb*A; *psb*T; [***Taxon***] (sampled for *cys*A, if different than the species sampled for the phylogenetic analysis); *cys*A state (ORF [open reading frame], pseudogene, or loss); GenBank accession number for *cys*A.

Acrolejeunea aulacophora (Mont.) Stephani;—;—; JF513352; JF513461; JF513479; JF513398;—; loss; JF825929. ***Allisonia cockaynii*** (K.I.Goebel) R.M.Schust.; AY688742; AY688679; JF513357; AY507389; AY507432; AY507472; JF513423; pseudogene; JF825930. ***Anastrepta orcadensis*** (Hook.) Schiffner;—; JF513325; JF513340; JF513450; JF513468; JF513391; JF513407; pseudogene; JF825931. ***Aneura mirabilis*** (Malmb.) Wickett & Goffinet; AY608274; AY608207; JF513376; DQ439683; AY608057; AY607936; JF513441; pseudogene; EU043314. ***Aneura pinguis*** (L.) Dumort.; AY688744; AY877366; JF513358; AY507391;—; AY507474; JF513424; [*Aneura* sp. nov]; ORF; JF825894. ***Ascidiota blepharophylla*** C. Massal.; AY608265; AY608198; JF513383; DQ439678; AY608046; AY607925; JF513447; pseudogene; JF825932.

Balantiopsis tumida Berggr.;—; JF513338; JF513356;—; JF513484; JF513402; JF513422; pseudogene; JF825933. ***Bazzania tricrenata*** (Wahlenb.) Trevis.;—; JF513326; JF513341; JF513451; JF513469; JF513392; JF513408; pseudogene; JF825934. ***Blasia pusilla*** L.; AY688746; AY688683; JF513359; AF536232; AY507436; AY507477;—; ORF; JF825895, JF825896.

Calypogeia muelleriana (Schiffner) K. Mueller;—; AY608203; AY607870;—; AY608052; AY607931; AY608000; ORF; JF825897. ***Cephalozia bicuspidate*** (L.) Dumort.;—;—; JF513385; JF513465; JF513486; JF513405;—; loss; JF825935. ***Cephaloziella hirta*** (Stephani) R.M.Schust.; AY608271; AY608204;—; DQ439682; AY608054; AY607933; JF513433; loss; JF825936. ***Chiloscyphus profundus*** (Nees) J.J.Engel & R.M.Schust.; DQ268932; DQ268889; JF513369; DQ268973; DQ268987; DQ268999; JF513434; ORF; JF825898. ***Clevea hyalina*** (Sommerf.) Lindb.; DQ268911; DQ265773;—; DQ286002; DQ220676; DQ265746;—; ORF; JF825899. ***Conocephalum conicum*** (L.) Dumort.; AY688748; DQ268878; JF513360; AY688778; AY688791; DQ268996; JF513425; ORF; JF825900, JF825901. ***Corsinia coriandrina*** (Spreng.) Lindb.; DQ268917; DQ265777;—; DQ286006; DQ220680; DQ265750;—; ORF; JF825902. ***Cyathodium tuberosum*** Kashyap; DQ268919; DQ265779;—; DQ286008; DQ220682; DQ265752;—; [*Cyathodium spruceanum* Prosk.]; ORF; JF825903.

Diplophyllum obtusifolium (Hook.) Dumort.; AY688749; AY688685; JF513361; AY507397; AY507439; AY507480; JF513426; [*Diplophyllum albicans* (L.) Dumort.]; pseudogene; JF825937. ***Douinia ovata*** (Dicks.) H. Buch;—;—; JF513348; JF513457; JF513475;—; JF513415; loss; JF825938. ***Drepanolejeunea erecta*** (Stephani) Mizut.; JF513387; JF513327; JF513342; JF513452; JF513470; JF513393;—; loss; JF825939. ***Dumortiera hirsuta*** (Sw.) Nees; DQ268920; DQ265780;—; DQ286009; DQ220683; DQ265753;—; ORF; JF825904.

Exormotheca pustulosa Mitt.; DQ268921; DQ265781;—; DQ286010; DQ220684; DQ265754;—; ORF; JF825906, JF825907.

Fossombronia angulosa (Dicks.) Raddi; AY688750; AY688686; JF513362; AY507398; AY507440; AY507481; JF513427; loss; JF825940.

Gackstroemia magellanica (Lam.) Trevis.; JF513389; JF513337; JF513355; JF513463; JF513483;—; JF513421; pseudogene; JF825941. ***Goebelobryum unguiculatum*** (Hook.f. & Taylor) Grolle;—; JF513339; JF513386; JF513466; JF513487; JF513406;—; pseudogene; JF825942. ***Greeneothallus gemmiparus*** Hässel; DQ268923; AY877367; JF513363; AY688780; AY688792; AY688828; JF513428; pseudogene; JF825943. ***Gyrothyra underwoodiana*** M. Howe; DQ268924; DQ268881; JF513364; DQ268970; DQ268985; DQ268997;—; ORF; JF825908.

Jubula hutchinsiae (Hook.) Dumort.; DQ268929; DQ268886; JF513366; AY507408; AY688794; AY507492; JF513430; loss; JF825944.

Liochlaena lanceolata Nees; AY688756;—; JF513367; AY507409; AY507451; AY507493; JF513431; pseudogene; JF825910. ***Leiocolea bantriensis*** (Hook.) Jørg.;—; JF513332; JF513349; JF513458; JF513476;—; JF513416; ORF; JF825909. ***Lejeunea tuberculosa*** Stephani; JF513388; JF513328; JF513344; JF513453; JF513472; JF513394; JF513410; loss; JF825946. ***Lepidozia reptans*** (L.) Dumort.; AY608293; AY608229; JF513380;—; AY608083; AY607962; AY608013; pseudogene; JF825947. ***Lobatiriccardia lobata*** (Schiffner) Furuki; AY688757; AY877373; JF513368; AY507421; AY507462; AY507507; JF513432; pseudogene; JF825948. ***Lophozia ventricosa*** (Dicks.) Dumort.;—;—; JF513350; JF513459; JF513477;—; JF513417; pseudogene; JF825950. ***Lunularia cruciata*** (L.) Dumort. ex Lindb.; DQ268934; DQ268890;—; AY688783; AY688795; AY688830;—; ORF; JF825911.

Makinoa crispata (Stephani) Miyake; AY877386; AY877374; JF513370; AY877390; AY877393; AY877399; JF513435; ORF; JF825912. ***Marchantia inflexa*** Nees & Mont.; DQ268937; DQ268891;—; DQ268974;—; DQ269000;—; ORF; JF825913. ***Marsupella emarginata*** (Ehrh.) Dumort.;—; JF513333; JF513351; JF513460; JF513478; AY607966; JF513418; pseudogene; JF825951. ***Mastigophora woodsii*** (Hook.) Nees;—; JF513329; JF513345; JF513454; JF513473; JF513395; JF513411; ORF; JF825915. ***Metacalypogeia alternifolia*** (Nees) Grolle;—; JF513330;—; JF513455; JF513474; JF513396; JF513412; pseudogene; JF825952. ***Metzgeria furcata*** (L.) Dumort.; DQ268940; DQ268892; JF513371; DQ268975; DQ268988; DQ269001; JF513436; pseudogene; JF825953. ***Monoclea gottschei*** Lindb.; AY877388; AY688691; JF513372; AY507414; AY507455; AY507498; JF513437; ORF; JF825916. ***Monosolenium tenerum*** Griff.; DQ268944; DQ265788;—; DQ286017; DQ220691;—;—; ORF; JF825917.

Nardia scalaris A. Gray;—; JF513335; JF513354; JF513462; JF513481; JF513400; JF513419; ORF; JF825918. ***Neotrichocolea bisseti*** (Mitt.) S. Hatt.; AY608300; AY608237; JF513381; DQ439696; AY608093; AY607972; JF513445; pseudogene; JF825954. ***Noteroclada confluens*** Taylor ex Hook. & Wilson; DQ268946; DQ268895; JF513373; DQ268977; DQ268990; DQ269003; JF513438; pseudogene; JF825955.

Odontolejeunea lunulata (F. Weber) Schiffner; AY608302; AY608239; JF513382; DQ439698; AY608095; AY607974; JF513446; loss; JF825956.

Pallavicinia lyellii (Hook.) Carruth.; AY688761; AY734742; JF513374; AY507416; AY688798; AY507501; JF513439; pseudogene; JF825957. ***Pellia epiphylla*** (L.) Corda; AY688764; AY688693; JF513375; AY688787; AY507457; AY507502; JF513440; pseudogene; JF825958. ***Peltolepis quadrata*** (Saut.) K. Mueller; DQ268947; DQ265790;—; DQ286019; DQ220693;—;—; ORF; JF825919. ***Ptychanthus striatus*** (Lehm. & Lindenb.) Nees;—; JF513334; JF513353; JF513467; JF513480; JF513399; JF513449; loss; JF825959.

Radula* sp**; JF513390;—; JF513384; JF513464; JF513485;—; JF513448; loss; JF825960, JF825961. ***Reboulia hemisphaerica (L.) Raddi; AY877383; DQ268900;—; AY688788; AY688801; AY8773402;—; ORF; JF825920, JF825921. ***Riccardia capillacea*** (Stephani) Meenks & C. De Jong; AY688768; AY877378; JF513377; AY877392; AY688802; AY688833; JF513442; [*Riccardia* sp.]; ORF; JF825922. ***Riccia albolimbata*** S.W. Arnell;—;—;—;—;—;—;—; ORF; JF825923; ***Riccia fluitans*** L.; DQ268956; DQ265795;—; DQ286023; DQ220697; DQ265766;—; ORF; JF825924, JF825925.

Scapania nemorea (L.) Grolle; AY688769; AY877379; AY607893; AY507423; AY507464; AY507509; AY608023; pseudogene; JF825962. ***Solenostoma speciosum*** (Horik.) Hentschel, K. Feldberg, Bombasch, D.G.Long, Vánas & J. Heinrichs;—; JF513331; JF513346; JF513456;—; JF513397; JF513413; [*Scaphophyllum speciosum* (Horik.) Inoue]; ORF; JF825926. ***Schiffneria hyalina*** Stephani; DQ268961; DQ268902; JF513378;—; JF513488; JF513403; JF513443; pseudogene; JF825963.

Treubia lacunosoides T. Pfeiff., W. Frey & M. Stech; AY688775; DQ268905; JF513379; AY507429; AY507469; AY507515; JF513444; pseudogene; JF825964. ***Trichocolea tomentella*** (Ehrh.) Dumort.;—;—; JF513343;—; JF513471;—; JF513409; ORF; JF825927. ***Trocholejeunea infuscata*** (Mitt.) Verd.;—;—; JF513347;—;—;—; JF513414; loss; JF825965.

Vetaforma dusenii (Stephani) Fulford & J. Taylor;—; JF513336;—;—; JF513482; JF513401; JF513420; pseudogene; JF825966.

Wiesnerella denudata (Mitt.) Stephani; DQ268966; DQ265799;—; DQ286027;

DQ220701;—;—; ORF; JF825928.

Summary:In total, 26 taxa retain an ORF for *cys*A, 25 retain a pseudogene, and 12 have no detectable ORF that can be aligned with *cys*A (noted here as a gene loss).

Outgroups:

Andreaea rothii F. Weber & D. Mohr; AY312862; AY312861;—; AY608025; AY312866; AY312863; AY607999; *cys*A not applicable. ***Anthoceros*** sp.; DQ845731; AF226036; DQ845731; AB086179; AB086179; AB086179;—; ORF; AB086179. ***Chara*** sp.; AY267353; DQ026521;—; DQ229107; DQ229107; DQ229107;—; [*Chara vulgaris* L.]; loss; DQ229107. ***Dendrohypopterygium arbuscula*** (Brid.) Kruijer; AY608276; AY608209;—; AY608027; AY608059; AY607938; AY608002; *cys*A not applicable. ***Hookeria lucens*** (Hedw.) Sm.; AY908489; AY330439;—; AY631185; AJ251316; AY312906; AY608009; *cys*A not applicable. ***Megaceros flagellaris*** (Mitt.) Stephani; AY877387; AY877375;—; AY463040;—; AY877400;—; *cys*A not applicable. ***Phymatoceros bulbiculosus*** (Brot.) Stotler, W.T.Doyle & Crand.-Stot.; DQ268948; DQ268898;—; DQ268978;—; DQ269004;—; *cys*A not applicable. ***Physcomitrella patens*** (Hedw.) Bruch & Schimper; AB251495;—; AB251495; AP005672; AP005672; AP005672;—; loss; AP005672. ***Takakia ceratophylla*** (Mitt.) Grolle; DQ268963; DQ268904;—;—; DQ268993; JF513404;—; *cys*A not applicable.

APPENDIX 2. Primers used to amplify and sequence *cys*A.

Primer name	Sequence	Position relative to *cys*A	Direction	Notes
trnDF	GGGGGTAGAGGGACTTGAAC	External (in *trn*D)	Forward	
trnE-cysA_F	CCAGGGGAATTCGAATCCCCGTCGCC	External (in *trn*E)	Forward	
trnE_longF	GCCTAGGACACCTCTCTTT CAAGGAGGCGACGGGGATTCG	External (in *trn*E)	Forward	
trnE_longF2	CGAATCCCCGTCGCCTCCT TGAAAGAGAGGTGTCCTAGGC	External (in *trn*E)	Forward	
trnT-cysA_R	CGCCTTACCATGGCGTTACTCTACCAC	External (in *trn*T)	Reverse	
trnT_longR	GATGACTTACGCCTTACCA TGGCGTTACTCTACCACTGAG	External (in *trn*T)	Reverse	
trnT_GUU-R	GAACCGATGACTTACGCCTTACC	External (in *trn*T)	Reverse	
trnYF	CCGTCCCCATTAACCACTCG	External (in *trn*Y)	Forward	
cysAFNew	GTAGTTTATTACGAATYATTGCAGG	Internal (starts at base 125 *Marchantia cys*A)	Forward	
cysAinternalF	CNAGTTTATTACGAATYATTGCRGG	Internal (starts at base 125 *Marchantia cys*A)	Forward	
cysAFint	CGAATTATTGCRGGTCTTG	Internal (starts at base 136 *Marchantia csy*A)	Forward	
cysAFintB	GCACTTTTTAARCATATGACTG	Internal (starts at base 250 *Marchantia cys*A)	Forward	
cysA_DF	GCRCTTTTYAARCATATGAC	Internal (reverse complement of cysA_CR; starts at base 250 *Marchantia cys*A)	Forward	
cysA_CR	GTCATATGYTTRAAAAGYGC	Internal (reverse complement of DF; starts at base 269 *Marchantia cys*A)	Reverse	
cysAinternalR	CCRTCYAANGCACYRAAAGGTTC	Internal (starts at base 494 *Marchantia cys*A)	Reverse	
cysARNew-Long	GGATANGCYCKTAAAAAAAC	Internal (starts at base 863 *Marchantia cys*A)	Reverse	
cysARint	GAAAAGMTTGATARCCTATYG	Internal (starts at base 1009 *Marchantia cys*A)	Reverse	
cysARNew	CKRAAAGATTGATAACCTATTGG	Internal (starts at base 1010 *Marchantia cys*A)	Reverse	
cysAComFi	GTTTATTACGAATTATTGCAGGTC	Internal (starts at base 128 *Marchantia cys*A)	Forward	Complex thalloid sequencing primer
cysAComRi2	GGAATTTCKATAAGTAATCC	Internal (starts at base 719 *Marchantia cysA*)	Reverse	Complex thalloid sequencing primer
cysAComRi1	GTMATTGAYTCATTTAATTTYGG	Internal (starts at base 740 *Marchantia cys*A)	Reverse	Complex thalloid sequencing primer