

Phylogenetic analyses of nuclear, mitochondrial and plastid multi-gene datasets support the placement of *Mesostigma* in the Streptophyta

Naiara Rodríguez-Ezpeleta*¹, Hervé Philippe*^{1,2}, Henner Brinkmann*, Burkhard Becker[†], and Michael Melkonian^{†,2}

Research Article

*Département de Biochimie, Centre Robert Cedergren, Canadian Institute of Advanced Research, Université de Montréal;

[†]Botanisches Institut, Lehrstuhl I, Universität zu Köln

¹ These authors contributed equally to this work

² Corresponding authors: herve.philippe@umontreal.ca; michael.melkonian@uni-koeln.de

Key words: *Mesostigma*, *Chlorokybus*, phylogenomics, taxon sampling, systematic error, heterotachy.

Running head: Support for the placement of *Mesostigma* in the Streptophyta

Abstract

All extant green plants belong to one of two major lineages, commonly known as the Chlorophyta (most of the green algae) and the Streptophyta (land plants and their closest green algal relatives). The scaly green flagellate *Mesostigma viride* has an important place in the debate on the origin of green plants. However, there have been conflicting results from molecular systematics as to whether *Mesostigma* diverges before the Chlorophyta/Streptophyta split or is an early diverging flagellate member of the Streptophyta. Previous studies employed either a limited taxon sampling (plastid and mitochondrial genomes) or a small number of phylogenetically informative sites (single nuclear genes). Here, we use large datasets from the nuclear (125 proteins; 29,319 positions), mitochondrial (33 proteins; 6,622 positions), and plastid (50 proteins; 10,137 positions) genomes with an expanded taxon sampling (21, 13, and 28 species, respectively) to reevaluate the phylogenetic position of *Mesostigma*. Our study supports the placement of *Mesostigma* in the Streptophyta (as an early diverging lineage) and provides evidence that systematic biases have played a role in generating some of the previous conflicting results. Importantly, we demonstrate that using an increased taxon sampling as well as more realistic models of evolution allows increasing congruence among the nuclear, mitochondrial and plastid dataset.

Introduction

Since its discovery in 1894 (Lauterborn 1894), the ubiquitous, but non-abundant, freshwater green flagellate *Mesostigma* has had an inconspicuous research history until Manton and Ettl (1965) described the ultrastructure of the spectacular scale covering of the cell and its two flagella, which allied the organism with the then newly recognized green algal class

Prasinophyceae Christensen 1962. Ultrastructural studies in the early 1970s renewed interest in the Prasinophyceae, when morphologically similar scales were found on flagellate reproductive cells of algae such as the stonewort *Chara corallina* (Moestrup 1970), that had been implicated in the evolution of embryophyte land plants based on ultrastructural features of mitosis/cytokinesis and the flagellar apparatus (Pickett-Heaps and Marchant 1972). The search for a flagellate member of the land plant lineage among the prasinophyte flagellates remained unsuccessful until a cruciate flagellar root system with two multilayered structures (MLS) was eventually described in *Mesostigma viride* (Rogers et al. 1981; Melkonian 1983, 1989), which linked this organism to both major lineages of green plants, the Chlorophyta (with cruciate flagellar root systems) and the Streptophyta (with unilateral flagellar root systems and a multilayered structure), the latter including the embryophyte land plants (Bremer et al. 1987). Since many structural features of *Mesostigma*, however, resembled those of other prasinophyte flagellates such as *Pyramimonas*, *Mesostigma* was often regarded as closely related to such prasinophytes (Moestrup and Thronsdén 1988, Melkonian 1990, Moestrup 1991), and it was not until molecular phylogenetic analyses of nuclear-encoded genes (SSU rDNA and actin) revealed the affiliation of *Mesostigma* with the Streptophyta (Melkonian et al. 1995; Bhattacharya et al. 1998, Marin and Melkonian 1999) that this view changed (but see Moestrup 2002, for a traditional taxonomic treatment of *Mesostigma*). The initial molecular phylogenetic analyses renewed interest in this organism and sparked a flurry of further studies aiming to determine the “true” position of *Mesostigma* in the tree of life. Unfortunately, depending on the dataset used, conflicting conclusions were reached (for reviews, see Lewis and McCourt 2004; McCourt et al. 2004): whereas phylogenetic analyses based on chloroplast and mitochondrial genes and genomes led to the conclusion that *Mesostigma* is the earliest green plant divergence branching before the split into the Chlorophyta and Streptophyta (Lemieux et al. 2000; Turmel et al 2002a,

2002b), other studies using a single chloroplast gene (*rbcL*; Delwiche et al. 2002) or a combination of four genes from the chloroplast, mitochondrial and nuclear genomes (Karol et al. 2001) concluded that *Mesostigma* represents the earliest divergence within the Streptophyta. In the meantime, molecular data accumulated that revealed a number of embryophyte traits in *Mesostigma*, thus refuting its close affiliation with other prasinophyte flagellates (Simon et al. 2006; Nedelcu et al. 2006; Petersen et al. 2006). The availability of EST data from *Mesostigma* (Simon et al. 2006) afforded the possibility to perform a large-scale multigene phylogenetic analysis to reevaluate its phylogenetic position. Here, we report multigene phylogenetic analyses of large nuclear, mitochondrial, and chloroplast datasets. The phylogenetic analyses based on the nuclear, mitochondrial and plastid datasets are consistent with the conclusion that *Mesostigma* is a member of the Streptophyta. We discuss possible reasons for the incongruence between previous studies.

Material and Methods

Construction of the nuclear, plastid and mitochondrial datasets

The nuclear dataset is based on an available alignment (TREEBASE, Acc. No. SN2312) to which EST and trace sequences downloaded from NCBI (<http://www.ncbi.nlm.nih.gov/>) from *Mesostigma viride*, *Selaginella moellendorffii*, *Volvox carteri*, *Scenedesmus obliquus*, *Prototheca wickerhamii*, *Helicosporidium sp.*, *Scherffelia dubia*, *Ostreococcus tauri*, *Galdieria sulphuraria* and *Chondrus crispus* were added as described (Philippe et al. 2004). Among the Viridiplantae, the 15 species with the largest number of genes sequenced, that represent the major groups were selected. Four red algae and two glaucophytes were used to root the tree. Unambiguously aligned sequence blocks were extracted with Gblocks (Castresana 2000) and manually verified.

Potential paralogs were identified and removed as described (Philippe et al. 2004). When including all orthologous proteins that are available from at least 11 out of the 21 used species, the dataset contains 125 proteins, totaling 29,319 amino acid positions. On average, 37% of the amino acids are missing.

The mitochondrial dataset consists of 33 proteins, a total of 6,622 amino acid positions, from 13 species, including the jakobid *Reclinomonas* and four red algae that were used as an outgroup. The plastid dataset consists of 50 plastid encoded proteins, a total of 10,137 amino acid positions, from 28 species, including a glaucophyte and eight red plastid-containing eukaryotes used as an outgroup. 13% and 5% of the amino acid positions are missing in the mitochondrial and plastid datasets respectively. Both datasets were aligned with CLUSTAL W (Thompson et al. 1994), refined manually using MUST (Philippe 1993), and filtered from ambiguously aligned positions with Gblocks (Castresana 2000).

Phylogenetic analyses

The concatenated datasets of nuclear, mitochondrial and plastid sequences were analyzed by maximum parsimony (MP), and maximum likelihood (ML). MP analyses were performed using PAUP* 4.0 b10 (Swofford 2002) with the Tree Bisection and Reconnection search and 10 random addition of species. ML analyses were performed using PhyML 2.4 (Guindon and Gascuel 2003) and TREEFINDER (Jobb et al. 2004) with an evolutionary model consisting of the WAG matrix of amino acid substitution (Whelan and Goldman 2001), estimated amino acid frequencies and a gamma distribution split into 4 categories to model the rate heterogeneity among sites (WAG+F+ Γ). The reliability of each internal branch was evaluated based on 1,000 (MP) or 100 (ML) bootstrap replicates.

Because the probability of getting trapped on a local maximum using heuristic searches is high when large datasets are used (Salter 2001), we also performed exhaustive ML analyses on a set of topologies obtained by constraining several groups. In general, groups supported by 100% bootstrap values (BV) in previous analyses were constrained. In some cases, because constraining all the groups supported at 100% BV resulted in a too small set of topologies, we let some of them free (see information about constrained groups in the Figures). In all cases, the three alternative positions for *Mesostigma* were exhaustively tested. The constraints allowed us to reduce the number of topologies to be analyzed to 81, 405 and 675 for the nuclear, plastid and mitochondrial datasets respectively. For each dataset, the likelihood of each topology was calculated using TREE-PUZZLE (Schmidt et al. 2002) with the WAG+F+ Γ model on the concatenated dataset. Site-wise likelihood values were calculated by PAML (Yang 1997) and used to perform RELL bootstrap analyses (Kishino et al. 1990; Baptiste et al. 2002) with 10,000 replicates to assess the statistical support for each unconstrained branch.

Amino acid coding to reduce compositional bias

To reduce the possible impact of compositional bias, we recoded the amino acids into four functional groups. To do that, we used the same coding as Hrady et al. (2004) but combined the aromatic (FYW) and the hydrophobic (MVIL) amino acids in a single category and coded the rare cysteine as missing data. This allowed the use of a general-time-reversible (GTR) matrix with four character states implemented in most programs. The parameters of the GTR matrix were estimated by PAUP using a neighbour-joining tree. The three sets of topologies (from nuclear, plastid and mitochondrial datasets) were exhaustively analyzed by TREE-PUZZLE with

a GTR+F+ Γ model. RELL bootstrap analyses (10,000 replicates) were performed as described above.

Removal of the fast evolving sites to reduce the impact of some systematic errors

To reduce the impact of systematic errors, we eliminated the fast evolving sites (Philippe et al. 2005a). To do that, we calculated the site-wise rates by PAML on an alignment that does not contain *Mesostigma* using the ML topology from which *Mesostigma* was pruned. This strategy was used because the phylogenetic position of *Mesostigma* may influence the estimation of the site-wise evolutionary rates potentially biasing the site removal approach (Rodriguez-Ezpeleta et al. *in press*). The sites were then sorted according to their evolutionary rates and progressive removals of the fastest sites (1,000 each time) were performed. RELL bootstrap analyses (1,000 replicates) were performed after each removal and plotted against the alignment size.

Separate analysis to reduce the impact of heterotachy

To reduce systematic errors due to rate heterogeneity among sites through time (heterotachy), we performed separate analyses on previously defined partitions of the dataset. To do that, we proceeded as described above for the exhaustive analysis, but allowed branch lengths, alpha parameter of the gamma distribution and stationary amino acid frequencies to be estimated independently for each partition. In order to evaluate the performance of the separate versus the concatenate model, we used the second order Akaike Information Criterion (Akaike, 1973), AIC_c (Hurvich and Tsai, 1989): $AIC_c = 2\log L + 2K + 2K(K+1)/n-K-1$, K being the

number of free parameters of the model and n the number of positions in the dataset. The number of free parameters was calculated as 1 alpha parameter + $2s - 3$ branch lengths (s being the number of species) and 19 amino acid frequencies.

Results

Phylogenetic analysis based on nuclear genes

We assembled a dataset of 125 evolutionarily conserved orthologous proteins (29,319 amino acid positions) from 15 different taxa of Viridiplantae and four red algae and two glaucophytes as outgroups to gain more insight into the phylogenetic position of *Mesostigma viride*. The maximum likelihood phylogeny inferred from the concatenation of the 125 proteins clearly placed *Mesostigma* in a sister-group position to the six other streptophyte taxa included in the analysis (Fig. 1). This position of *Mesostigma* is supported by 100% bootstrap values in all analyses (Fig. 1). Almost all nodes in the tree are well-resolved (except for the two long branch thermophilic red algae), and the overall tree topology agrees well with phylogenies derived from SSU rDNA sequence comparisons (e.g. Marin and Melkonian 1999). The amino acid coding into four functional groups (to reduce the impact of compositional bias), the removal of fast evolving sites or the separate analysis of two partitions (ribosomal and non-ribosomal proteins) did not change the results (data not shown).

Phylogenetic analysis based on mitochondrial genes

The dataset contained 33 mitochondrion-encoded orthologous proteins (6,622 amino acid positions) from 8 different taxa of Viridiplantae and four red algae and the jakobid flagellate *Reclinomonas* as outgroups. The maximum likelihood phylogeny inferred from the

concatenation of the 33 proteins again placed *Mesostigma* in a sister-group position to the five other streptophyte taxa included in this analysis (Fig. 2). This position is, however, only weakly supported by bootstrap values (no support in maximum parsimony, 65-85% BV support in the maximum likelihood (TREEFINDER and PhyML) and exhaustive analyses) (Fig. 2). Again we note that the tree topology is not only largely congruent with that of the nuclear-encoded dataset (considering the different taxon sampling), but overall is also not in conflict with several previous single-gene phylogenies. Our phylogeny, however, is in conflict with the phylogenetic analysis based on mitochondrial proteins presented by Turmel et al. (2002a), who showed that *Mesostigma* branched at the base of the Viridiplantae. To understand this discrepancy, we did the following additional experiments (summarized in the Supplementary Material, Figs S1, S2): (1) we reduced our dataset to the same proteins used by Turmel et al. (2002a), i.e. 4842 amino acids; (2) we reduced the taxon sampling to the same species used by these authors (8 species); (3) we reduced both the number of proteins and the taxon sampling; (4) finally we analyzed all datasets with (Fig. S1) and without (Fig. S2) taking rate heterogeneity among sites into account. The results of Turmel et al. (2002a) could only be reproduced using their taxon sampling and without using a gamma distribution to model the rate heterogeneity among lineages. Applied to our original dataset, the amino acid coding into four functional groups or the separate analysis of two partitions (ribosomal and non-ribosomal proteins) does not change the results; however, the removal of fast evolving sites improves the bootstrap support value for the placement of *Mesostigma* as sistergroup of streptophytes (97 % when the 1,500 fastest evolving sites are eliminated; data not shown).

Phylogenetic analysis based on plastid genes

The dataset consists of 50 plastid-encoded orthologous proteins (10,137 amino acid positions) from 19 taxa of Viridiplantae and eight eukaryote taxa with red-algal type plastids and the glaucophyte *Cyanophora* as outgroups. The maximum likelihood tree inferred from the concatenation of the 50 proteins places *Mesostigma* together with the streptophyte genus *Chlorokybus* (Fig. 3A) with strong support (100% BV in all analyses). *Mesostigma* + *Chlorokybus* (MC) emerge at the base of the Viridiplantae with low bootstrap support in the ML analysis (64-68% BV), and no support in the MP analysis. If *Chlorokybus* was excluded from the analysis, the support for the basal position of *Mesostigma* increased significantly in the ML analyses (85-93% BV; summarized in the Supplementary Material, Fig. S3). If *Mesostigma*, however, was excluded from the analysis (Fig. 3B), *Chlorokybus* grouped with the other Streptophyta in a basal position with bootstrap values of 75-92%. In these three cases, the amino acid coding into four functional groups or the removal of fast evolving sites did not change the results (not shown).

However, these two approaches do not overcome tree reconstruction artefacts due to heterotachy, i.e, rate variation across sites through time (Lopez et al., 2002), which are most likely present in plastid datasets (Lockhart et al., 2006). In an attempt to detect heterotachy, we divided the dataset into three functional classes: translation (ribosomal proteins), RNA polymerase (A, B and B' subunits of the RNA polymerase) and photosynthesis (the remaining proteins, all directly implicated in photosynthesis except two –acetylCoA carboxylase and cytochrome biogenesis protein). As shown in Fig. 4, the differences on the branch lengths inferred from the three datasets are extreme. In particular, streptophytes and chlorophytes evolve about three times faster with respect to the remaining species (including *Mesostigma* and *Chlorokybus*) in the RNA polymerase dataset and not in the other two. As expected, important differences in the BV for the position of MC are observed when analyzing these tree datasets

independently (Table 1). The translation dataset (2,199 amino acid positions) supports the placement of MC in the Streptopyta with 96% BV, whereas the support for the same relationship is non-existent with the RNA polymerase dataset (1,498 amino acid positions). Albeit being the largest (6,449 amino acid positions), the photosynthesis dataset does not discriminate among the three alternatives.

A way to reduce artifacts due to heterotachous behaviors is to use a separate model (Kolaczkowski and Thornton 2004; Yang 1996). However, a fully separate analysis considering independent parameters for each one of the 50 proteins only slightly decreases the support for the probably incorrect basal position of MC. In fact, there are too many free parameters (3,650) in this model, and the AIC (Table 1) indicates that it is the one that worst fits the data. In contrast, when only three partitions (translation, RNA polymerase and photosynthesis) are considered, only 146 additional parameters with respect to the concatenate model are needed, and the fit is better in this case (Table 1). Interestingly, support for the sistergroup of MC and streptophytes increases from 32 to 57% with the partially separate model, indicating that heterotachy accounts at least in part for the misplacement of MC in the plastid dataset.

Discussion

The incongruence between two single-gene phylogenies can be the result of (1) paralogy (generated by gene duplication), lateral gene transfer or lineage sorting, (2) stochastic error, derived from the use of too few phylogenetically informative sites, and (3) systematic error, arising from the presence of non-phylogenetic signal in the data that is not accounted for in the tree-reconstruction models employed (Philips et al. 2004). Non-phylogenetic signals mainly derive from variable evolutionary rates across lineages leading to the well-known long branch attraction (LBA) artifact, heterogeneous nucleotide/amino acid compositions leading to artificial

attraction of taxa with the same bias, and heterogeneity of the evolutionary rate of a given position through time, i.e. heterotachy (Philippe et al. 2005a).

The extent of taxon sampling may also account for incongruence between different phylogenies. Empirical evidence (e.g. Brinkmann et al. 2005; Philippe et al. 2005b) argues for a rich taxon sampling in phylogenetic analyses, because this enables better detection of multiple substitutions and thus recognition of long branch attraction (LBA) artifacts (Felsenstein 1978; Hendy and Penny 1989). Sometimes, however, the number of extant lineages is extremely sparse and it may never be possible to attain a rich or balanced taxon sampling.

Previous single gene/few-gene phylogenetic analyses trying to assess the phylogenetic position of *Mesostigma* may have suffered from some or all of the above problems and yielded conflicting results. Given these restrictions, the conclusions from these analyses are often contradictory and somewhat limited. Phylogenies based on the nuclear-encoded SSU rDNA (Melkonian et al. 1995; Marin and Melkonian 1999) and actin-coding genes (Bhattacharya et al. 1998) suggested the divergence of *Mesostigma* at the base of the Streptophyta, although the basal divergences within the Streptophyta remained unresolved. Phylogenies derived from plastid-encoded genes are incongruent. The concatenation of the two plastid-encoded rRNA genes reported phylogenies that consistently placed *Mesostigma* at the base of the Chlorophyta and Streptophyta (Turmel et al. 2002b), whereas phylogenies based on the *rbcL* gene or on four genes from the three genomes (*atpB* and *rbcL* from the plastid genome; *nad5* from the mitochondrial genome and SSU rDNA from the nuclear genome) placed *Mesostigma* at the base of the Streptophyta (Delwiche et al. 2002; Karol et al. 2001).

Phylogenomics, the genome-scale approach to phylogenetic inference, is thought to overcome the limitations of single gene phylogenies by combining many genes, and ultimately complete genomes (Philippe et al. 2005a). The use of large datasets theoretically overcomes

incongruence because such datasets reduce the impact of stochastic error when more and more genes are considered. Several empirical studies have confirmed these predictions (e.g. Madsen et al. 2001; Baptiste et al. 2002; Qiu et al. 1999); however, conflicting results have also emerged (such as the question of the monophyly of the Ecdysozoa, Lophotrochozoa and Protostomia; see Philippe et al. 2005b). Systematic error, caused by the presence of non-phylogenetic signals in the data, is not expected to disappear with the addition of data, because, unlike stochastic error, it does not average out over sites. If non-phylogenetic signal is strong enough, it will cause the tree reconstruction method to be inconsistent and lead to an incorrect, but statistically strongly supported tree (Felsenstein 1978; Phillips et al. 2004).

Two multigene analyses have been carried out to date that specifically address the phylogenetic position of *Mesostigma* in the Viridiplantae, and both refer to organelle phylogenomics. Lemieux et al. (2000) sequenced the entire chloroplast DNA of *Mesostigma viride* (118,360 bp) and analyzed a subset (53) of the 135 proteins encoded on the plastome (10,629 amino acid positions) with a taxon sampling that contained three Embryophyta and three Chlorophyta (and *Cyanophora paradoxa* as the outgroup). The tree topology in which *Mesostigma* diverged before the Streptophyta and Chlorophyta was strongly favored over alternative topologies which placed *Mesostigma* either at the base of the Streptophyta or the Chlorophyta. It must be noted, however, that their taxon sampling was very limited and lacked e.g. other streptophyte algae, and only one outgroup taxon was used. Furthermore, the probabilistic methods used (maximum likelihood analysis under the JTT-F model) for tree reconstruction assumed a uniform rate of substitution. Additional studies from this group increased taxon sampling by adding complete plastid genomes of both streptophyte algae and Chlorophyta (Turmel et al. 2002c; Pombert et al. 2005, 2006; Turmel et al. 2005, 2006) but did not address the phylogenetic position of *Mesostigma*. Other multigene analyses using chloroplast

proteins gave mixed results concerning the placement of *Mesostigma* (Martin et al. 2002, 2005).

In a second approach, Turmel et al. (2002a) sequenced the mitochondrial genome of *Mesostigma viride* (42,424 bp) and analyzed a subset (19) of the 65 proteins encoded on the chondriome (4,139 amino acid positions) with a taxon sampling that contained two embryophytes and two Chlorophyta (plus three red algae as outgroups). The tree topology in which *Mesostigma* diverged before the Streptophyta and Chlorophyta was supported by BV of 100% in PROTML, distance and maximum parsimony analyses assuming a uniform substitution rate across sites. However, when rate variation across sites (eight gamma categories) was taken into consideration, in maximum likelihood analyses (JTT model) the BV dropped significantly and support for this topology was low (63 or 67%, excluding or including invariant sites respectively).

How do the results obtained in the present study compare with those previously published? We have assembled a large nuclear dataset (125 proteins, 29,319 positions) with a reasonable taxon sampling (15 Viridiplantae, among them 6 streptophytes and 8 chlorophytes + *Mesostigma*). Phylogenetic analyses involving different methods of tree reconstruction and addressing likely causes of systematic error such as compositional bias and fast evolving sites, leads us to conclude that *Mesostigma* is an early branching member of the Streptophyta. Since only one other streptophyte alga (*Closterium*) was included in the analysis, we cannot yet address the relationship between *Mesostigma* and other streptophytes such as *Chlorokybus*, *Klebsormidium* or *Chaetosphaeridium*, which must await the generation of EST data from these organisms. Similarly, we note that early branching chlorophytes such as *Pyramimonas* (Nakayama et al. 1998) are still lacking. In general, the results obtained from multigene analyses of the nuclear dataset corroborate earlier analyses of nuclear-encoded single genes (SSU rDNA, actin), that placed *Mesostigma* in the Streptophyta.

For the mitochondrial dataset, our results are in accordance with the nuclear dataset but are in conflict with the mitochondrial protein phylogeny of Turmel et al. (2002a). Additional analyses adjusting the dataset (from 6,622 to 4,842 amino acid positions) and taxon sampling (from 13 to 8 taxa) to those used by Turmel et al. (2002a) revealed the likely reasons for the discrepancy: poor taxon sampling in the ingroup (lack of other streptophyte algae) as well as in the outgroup (only the long-branch red algae were chosen) appeared to be responsible, the number of positions used was less important as long as the rate heterogeneity among lineages was modeled (see Results). However, when no gamma distribution was used, bootstrap support for the placement of *Mesostigma* with the streptophytes was abolished. This is in accordance with the data of Turmel et al. (2002a) who showed that the bootstrap support for the position of *Mesostigma* at the base of the Streptophyta and Chlorophyta was lowered when the rate heterogeneity among lineages was modeled. We conclude that the phylogeny of mitochondrial proteins places *Mesostigma* in the Streptophyta when taxon sampling is improved and the rate heterogeneity among lineages is modeled.

The phylogeny derived from the plastid dataset reveals that the inclusion of *Chlorokybus*, which strongly groups with *Mesostigma*, decreases the bootstrap support for the basal position of *Mesostigma* in the Viridiplantae (Fig 3), and that different subsets of the complete dataset provide conflicting results (Fig 4, Table 1). When the translational proteins are used, significant support for the placement of *Mesostigma* and *Chlorokybus* (MC) in the Streptophyta is obtained (this is also true for the datasets with *Mesostigma* or *Chlorokybus* only), whereas the RNA polymerase dataset significantly rejects this relationship (Table 1, S1, 2). This can be explained by the disproportionately fast evolutionary rate of the streptophyte RNA polymerase, which attracts this group to either Chlorophyta or to the outgroup (Figure 4C). Albeit its large size (6,449 positions) and no apparent evolutionary rate heterogeneities (Figure 4B), the

photosynthesis dataset does not support or reject any of the three alternatives (Tables 1, S1,2). This lack of resolution is likely due to the slow evolutionary rate of these proteins.

The plastid dataset illustrates tree reconstruction artifacts due to heterotachy (Kolaczowski and Thornton, 2004). There exist highly heterotachous branch lengths for the different functional classes that cannot be acknowledged when a single set of branch lengths is used to analyze the concatenation of the 50 plastid proteins. As a result, an erroneous topology (MC at the base of green plants) is recovered. Although a separate model a priori defined by three function-based gene partitions may correct for heterotachy, the improvement with respect to concatenation is only marginal (from 32% to 57% BV for the sister-group of MC and streptophytes). This suggests that even if the branch length differences are extreme between the three partitions, among-gene heterotachy is not the only cause of systematic error in this dataset (Philippe et al. 2005c). For instance, changes in the proportion of variable sites across the phylogeny within genes, which is the case for the RNA polymerase subunits (Lockhart et al. 2006), are not corrected by the type of separate model we used.

The separate model of the 50 proteins has a lower fit to the data because it implies much more free parameters, but only slightly improving the likelihood –note that previous studies (Baptiste et al. 2002; Rodriguez-Ezpeleta et al. 2005; Philippe et al. 2004) suggested, using an inadequate AIC approximation (see Posada and Buckley 2004), that the separate model was better. In any case, the difficulty of correctly locating *Mesostigma* with plastid proteins constitutes an interesting case study for testing new, more realistic, models of sequence evolution.

Apparently, *Mesostigma* attracts *Chlorokybus* to a position in the tree which does not conform to its typical streptophyte traits such as a subapical flagellar insertion and unilateral flagellar root system in the zoospores, which are morphological synapomorphies of the

Streptophyta (Rogers et al. 1980; Lewis and McCourt 2004). Given that the evolutionary rate of *Mesostigma* is higher than that of *Chlorokybus*, this suggests that the position of *Mesostigma* in phylogenetic trees based on plastid datasets is affected by systematic errors. According to Jeffroy et al. (2006), phylogenies based on multiple genes can be biased by systematic errors and should be carefully scrutinized for possible tree reconstruction errors. Supporting their conclusions, here, we have shown the importance of the use of (1) probabilistic methods that aim to capture real substitution patterns (Steel 2005) –in the mitochondrial dataset, and (2) an increased taxon sampling to corroborate the results –in the plastid and mitochondrial datasets.

In conclusion, the inclusion of *Mesostigma* in the Streptophyta, likely as an early branching lineage, is weakly supported by the mitochondrial and plastid datasets, but significantly by the nuclear dataset. This corroborates exciting recent findings about land plant-specific molecular/biochemical traits in *Mesostigma* such as the *GapA/B* gene duplication (Simon et al. 2006; Petersen et al. 2006), plant-type peroxisomal glycolate oxidase (Stabenau and Winkler 2005; Simon et al. 2006), the bud-induced (*BIP*) multi-gene family (Nedelcu et al. 2006), and F-box family proteins (Simon et al. 2006), suggesting that many typical embryophyte traits may have evolved at the level of the unicellular ancestor of the streptophytes before the transition to land, presumably when such a flagellate adapted to a freshwater/brackish habitat (Simon et al. 2006). This underpins the pivotal role that *Mesostigma* is likely to play in the coming years as a model to unravel the intricacies of the early steps in the evolution of streptophytes.

Supplementary Material

Figures S1-S3, Tables S1-S2 and the alignments used are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgements

HP acknowledges Génome Québec, the Canadian Research Chair and the Université de Montréal for financial support, and the Réseau Québécois de Calcul de Haute Performance for computational resources. NRE has been supported by ‘Programa de Formación de Investigadores del Departamento de Educación, Universidades e Investigación’ (Government of Basque Country). Part of this work was supported by grants from the Deutsche Forschungsgemeinschaft (Be 1779/7-1 and Be 1779/7-2).

Literature Cited

- Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In Proceedings 2nd International Symposium on Information Theory, Csaki, ed. (Budapest: Akademia Kiado), pp. 267-281.
- Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Durufle L, Gaasterland T, Lopez P, Müller M, Philippe H. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc Natl Acad Sci USA 99:1414-1419.
- Bhattacharya D, Weber K, An SS, Berning-Koch W. 1998. Actin phylogeny identifies *Mesostigma viride* as a flagellate ancestor of the land plants. J Mol Evol 47:544-550.
- Bremer K, Humphries CJ, Mishler BD, Churchill SP. 1987. On cladistic relationships in green plants. Taxon 36:339-349.

Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol* 54:743-757.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17:540-552.

Delwiche CF, Karol KG, Cimino MT, Sytsma KJ. 2002. Phylogeny of the genus *Coleochaete* (Coleochaetales, Charophyta) and related taxa inferred by analysis of the chloroplast gene *rbcL*. *J Phycol* 38:394-403.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* 27:401-410.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52: 696-704.

Hendy, MD, Penny, D. (1989). A framework for the quantitative study of evolutionary trees. *Syst Zool* 38: 297-309.

Hrdy I, Hirt RP, Dolezal P, Bardonova L, Foster PG, Tachezy J, Embley TM. 2004. *Trichomonas* hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* 432:618-622.

Hurvich, CM, Tsai C-L. 1989. Regression and time series model selection in small samples. *Biometrika* 76:297-307.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *Trends Genet* 22:225-231.

Jobb G, von Haeseler A, Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol* 4:18.

Karol KG, McCourt RM, Cimino MT, Delwiche CF. 2001. The closest living relatives of land plants. *Science* 294:2351-2353.

Kishino H, Miyata T, Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny, and the origin of chloroplasts. *J Mol Evol* 31:151-160.

Kolaczowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431:980-984.

Lauterborn R. 1894. Über die Winterfauna einiger Gewässer der Oberrheinebene. *Biol Zbl* 14:390-398.

Lemieux C, Otis C, Turmel M. 2000. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. *Nature* 403:649-652.

Lewis LA, McCourt RM. 2004. Green algae and the origin of land plants. *Am J Bot* 91:1535-1556.

Lockhart P, Novis P, Milligan BG, Riden M, Rambaut A, Larkum T. 2006. Heterotachy and Tree Building: A Case Study with Plastids and Eubacteria. *Mol Biol Evol.* 23:40-45.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. *Mol Biol Evol.* 19:1-7.

Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, Adkins R, Amrine HM, Stanhope MJ, de Jong WW, Springer MS. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* 409:610-614.

Manton I, Ettl H. 1965. Observations on the fine structure of *Mesostigma viride* Lauterborn. *J Linn Soc (Bot)* 59:175-184.

Marin B, Melkonian M. 1999. Mesostigmatophyceae, a new class of streptophyte green algae revealed by SSU rRNA sequence comparisons. *Protist* 150:399-417.

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA 99:12246-12251.

Martin W, Deusch O, Stawski N, Grünheit N, Goremykin V. 2005. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci 10:203-209.

McCourt RM, Delwiche CF, Karol KG. 2004. Charophyte algae and land plant origins. Trends Ecol Evol 19:661-666.

Melkonian M. 1983. *Mesostigma*, a key organism in the evolution of two major classes of green algae and related to the ancestry of land plants. Br Phycol J 18:206.

Melkonian M. 1989. Flagellar apparatus ultrastructure in *Mesostigma viride* (Prasinophyceae). Plant Syst Evol 164:93-122.

Melkonian M. 1990. Phylum Chlorophyta: class Prasinophyceae. In: Margulis L, Corliss JO, Melkonian M, Chapman DJ, editors. Handbook of Protoctista. Boston: Jones & Bartlett Publ. p 600-607.

Melkonian M, Marin B, Surek B. 1995. Phylogeny and evolution of the algae. In: Arai K, Kato M, Doi Y, editors. Biodiversity and Evolution. Tokyo: The National Science Museum Foundation. p 153-176.

Moestrup Ø. 1970. The fine structure of mature spermatozoids of *Chara corralina*, with special reference to microtubules and scales. Planta 93:295-308.

Moestrup Ø. 1991. Further studies of presumably primitive green algae, including the description of Pedinophyceae class. Nov. and *Resultor* gen. nov. J Phycol 27:119-133.

- Moestrup Ø. 2002. Phylum Prasinophyta. In: John DM, Whitton BA, Brook AJ, editors. The Freshwater Algal Flora of the British Isles. Cambridge: University Press. p. 281-286.
- Moestrup Ø, Thronsen J. 1988. Light and electron microscopical studies on *Pseudoscourfieldia marina*, a primitive scaly green flagellate (Prasinophyceae) with posterior flagella. Can J Bot 66:1415-1434.
- Nakayama T, Marin B, Kranz HD, Surek B, Huss VAR, Inouye I, Melkonian M. 1998. The basal position of scaly green flagellates among the green algae (Chlorophyta) is revealed by analyses of nuclear-encoded SSU rRNA sequences. Protist 149:367-380.
- Nedelcu AM, Borza T, Lee RW. 2006. A land plant-specific multigene family in the unicellular *Mesostigma* argues for its close relationship to Streptophyta. Mol Biol Evol 23:1011-1015.
- Petersen J, Teich R, Becker B, Cerff R, Brinkmann H. 2006. The GapA/B gene duplication marks the origin of Streptophyta (charophytes and land plants). Mol Biol Evol 23:1109-1118.
- Philippe H. 1993. MUST, a computer package of Management Utilities for Sequences and Trees. Nucleic Acids Res 21:5264-5272.
- Philippe H, Snell EA, Baptiste E, Lopez P, Holland PW, Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 21:1740-1752.
- Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005a. Phylogenomics. Annu Rev Ecol Evol Syst 36:541-562.
- Philippe H, Lartillot N, Brinkmann H. 2005b. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol 22:1246-1253.
- Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005c. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5:50.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol* 21:1455-1458.

Pickett-Heaps JD, Marchant HJ. 1972. The phylogeny of the green algae: A new proposal. *Cytobios* 6:255-264.

Pombert J-F, Otis C, Lemieux C, Turmel M. 2005. The chloroplast genome sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of chlorophyte lineages. *Mol Biol Evol* 22:1903-1918.

Pombert J-F, Lemieux C, Turmel M. 2006. The complete chloroplast DNA sequence of the green alga *Oltmannsiellopsis viridis* reveals a distinctive quadripartite architecture in the chloroplast genome of early diverging ulvophytes. *BMC Biology* 4:3.

Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol* 53:793-808.

Qiu YL, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* 402:404-7.

Rodriguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Loffelhardt W, Bohnert H, Philippe H, Lang BF. 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr Biol* 15:1325-1330.

Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol*. *In press*.

Rogers CE, Mattox KR, Stewart KD. 1980. The zoospore of *Chlorokybus atmophyticus*, a charophyte with sarcinoid growth habit. *Am J Bot* 67:774-783.

Rogers CE, Domozych DS, Stewart KD, Mattox KR. 1981. The flagellar apparatus of *Mesostigma viride* (Prasinophyceae): multilayered structures in a scaly green flagellate. *Plant Syst Evol* 138:247-258.

Salter LA. 2001. Complexity of the likelihood surface for a large DNA dataset. *Syst Biol* 50:970-978.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502-504.

Simon A, Glöckner G, Felder M, Melkonian M, Becker B. 2006. EST analysis of the scaly green flagellate *Mesostigma viride* (Streptophyta): Implications for the evolution of green plants (Viridiplantae). *BMC Plant Biol* 6:2.

Stabenau H, Winkler U. 2005. Glycolate metabolism in green algae. *Physiol Plant* 123:235-245.

Steel M. 2005. Should phylogenetic models be trying to 'fit an elephant'? *Trends Genet* 21:307-309.

Swofford DL. 2002. PAUP* Phylogenetic Analysis using Parsimony (*and other Methods). Version 4. Sunderland: Sinauer Associates.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673-4680.

Turmel M, Otis C, Lemieux C. 2002a. The complete mitochondrial DNA sequence of *Mesostigma viride* identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. *Mol Biol Evol* 19:24-38.

Turmel M, Ehara M, Otis C, Lemieux C. 2002b. Phylogenetic relationships among streptophytes as inferred from chloroplast small and large subunit rRNA gene sequences. *J Phycol* 38:364-375.

Turmel M, Otis C, Lemieux C. 2002c. The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: Insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. *Proc Natl Acad Sci USA* 99:11275-11280.

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

Turmel M, Otis C, Lemieux C. 2006. The chloroplast genome sequence of *Chara vulgaris* sheds new light into the closest green algal relatives of land plants. *Mol Biol Evol* 23:1324-1338.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18:691-699.

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *J Mol Evol* 42:587-596.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555-556.

Table1: Bootstrap support values obtained with different datasets for the three alternative positions of *Mesostigma* and *Chlorokybus*

	Number of positions	With Streptophyta	Basal	With Chlorophyta	AIC _c ¹
Complete CON	10,137	32	62	6	380,298.65
Complete SEP50	10,137	33	50	17	380,878.36
Complete SEP3	10,137	57	33	10	376,995.03
Translation CON	2,199	97	3	0	
Photosynthesis CON	6,449	22	41	37	
RNApol CON	1,489	0	59	41	

Bootstrap support values are based on 10,000 RELL replicates on the exhaustive analysis (see Materials and Methods). CON, same branch lengths, stationary amino acid frequencies and alpha parameter were used for the whole dataset; SEP, each protein (SEP50) or each partition (translation, RNA polymerase and photosynthesis; SEP3) was allowed to have its own branch lengths, stationary amino acid frequencies and alpha parameter. The solution with the highest BV in each case is highlighted in grey.

¹ AIC_c = -2LogL + 2K + 2K(K+1)/n-K-1, where K is the number of free parameters and n, the number of positions.

Figure Legends

Figure 1: Phylogenetic analysis of 125 nuclear encoded genes

Optimal maximum likelihood tree obtained by the analysis of 125 concatenated nuclear-encoded proteins (29,319 amino acid positions). Numbers represent (in order) support values obtained for 1,000 bootstrap replicates in maximum parsimony analyses, 100 in unconstrained maximum likelihood (TreeFinder and PhyML) and 10,000 in the constrained maximum likelihood analyses. No numbers indicate that the branch was supported by 100% bootstrap value with all methods. Black dots indicate the groups constrained in the exhaustive maximum likelihood analysis. The scale bar denotes the estimated number of amino acid substitutions per site.

Figure 2: Phylogenetic analyses of 33 mitochondrial encoded genes

Optimal maximum likelihood tree obtained by the analysis of 33 concatenated mitochondrial-encoded proteins (6,622 amino acid positions). See Fig. 1 for details.

Figure 3: Phylogenetic analyses of 50 plastid encoded genes

Optimal maximum likelihood tree obtained by the analysis of 50 concatenated plastid-encoded proteins (10,137 amino acid positions) with the complete dataset (A) or when *Mesostigma* was excluded (B). See Fig. 1 for details

Figure 4: Phylogenetic trees obtained from three subsets of the plastid dataset

Optimal maximum likelihood trees obtained by the analysis of the translation (A), photosynthesis (B) and RNA polymerase datasets (C). Numbers represent support values for the position of *Mesostigma* and *Chlorokybus* obtained for 100 bootstrap replicates in unconstrained (TreeFinder) (above) and 10,000 in the constrained maximum likelihood analyses (below). The scale bar denotes the estimated number of amino acid substitutions per site.

Figure 1

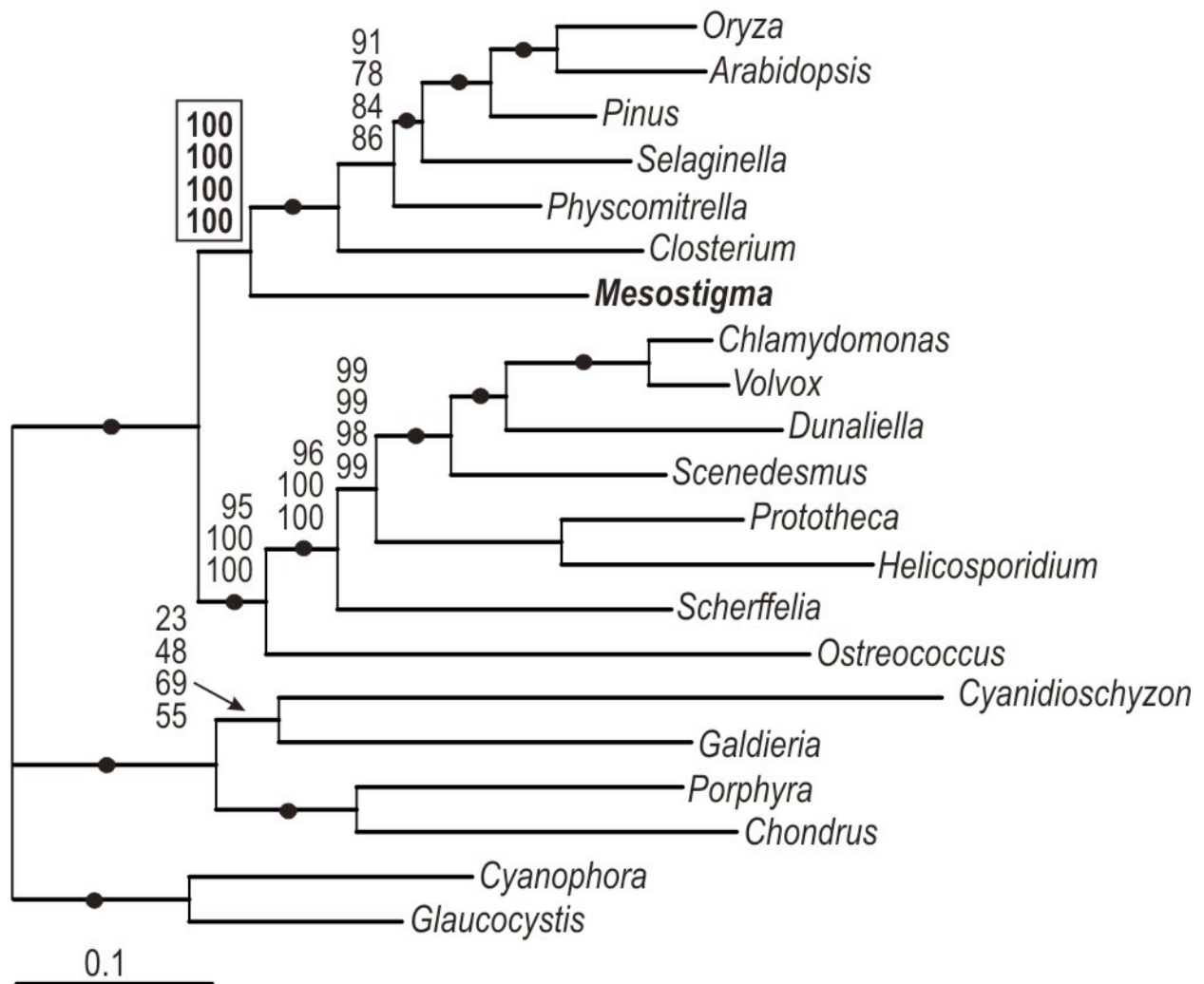


Figure 2

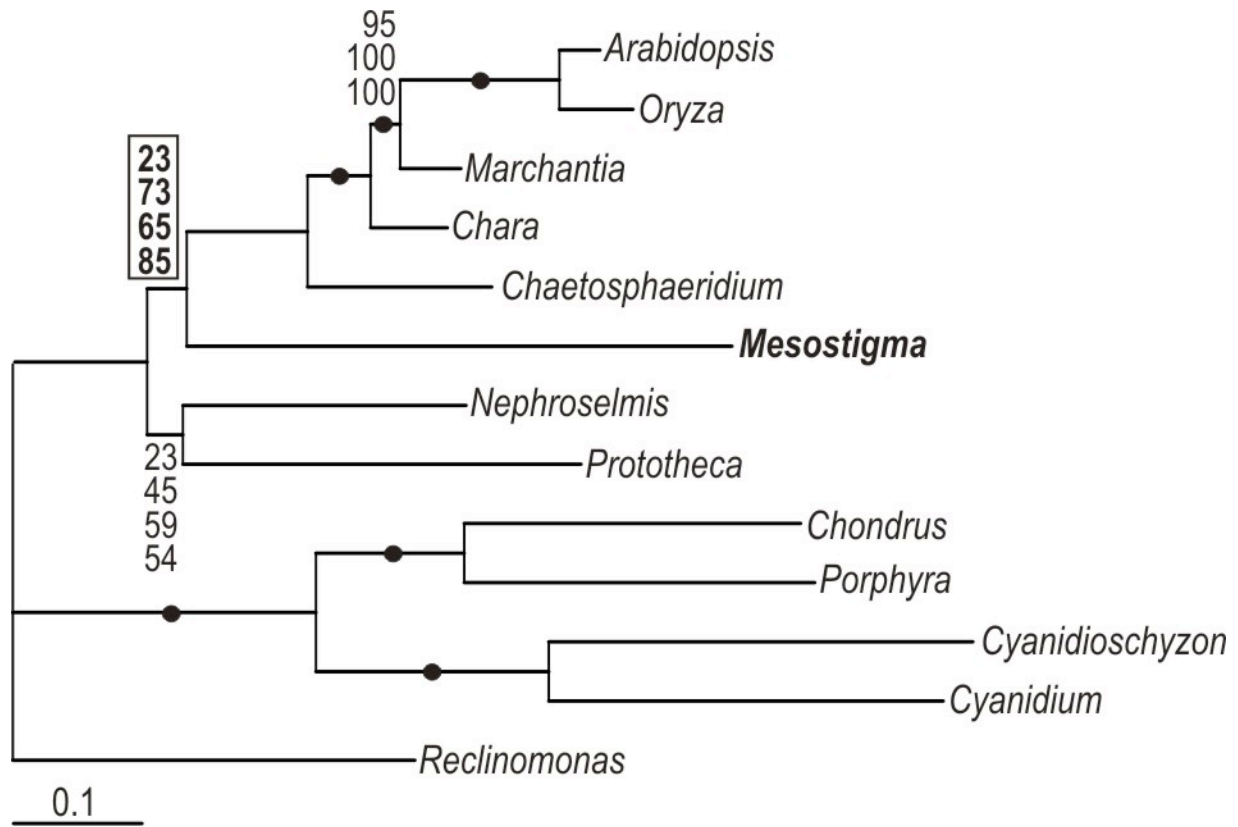


Figure 3

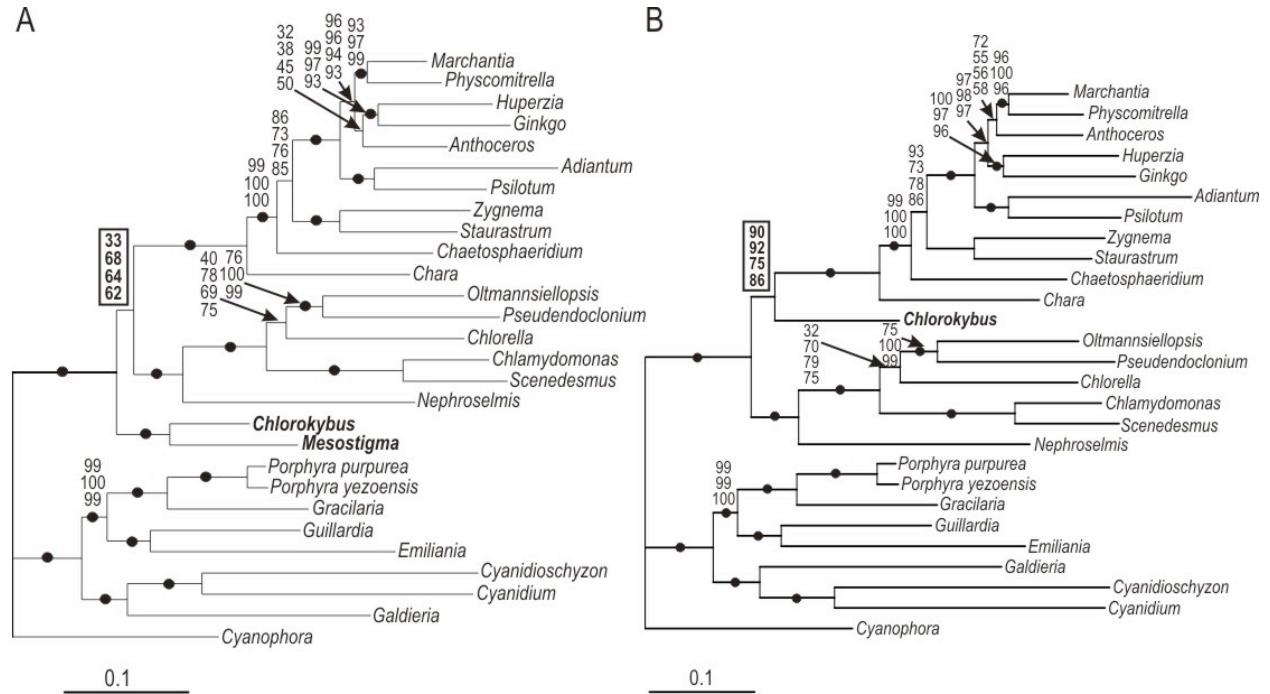


Figure 4

