Rooting morphologically divergent taxa – slow-evolving sequence data might help

Jorge Flores[1*], Alexander C. Bippus[2], Alexandru Tomescu[3], Neil Bell[4] & Jaakko Hyvönen[1,5*]

1 Finnish Museum of Natural History (Botany), PO Box 7, FI-00014 Univ. Helsinki, Finland

2 Dept. Botany & Plant Pathology, Oregon St. Univ., Corvallis, OR, 97331, USA

3 Dept. Biol. Sci., Humboldt St. Univ., Arcata, CA 95521, USA

4 Royal Botanic Garden Edinburgh, Scotland

5 Organismal & Evol. Biol. & Viikki Plant Sci. Centre, PO Box 65, FI-00014 Univ. Helsinki, Finland

* Contact authors: jorge.flores@helsinki.fi, jaakko.hyvonen@helsinki.fi

Abstract

When fossils are sparse and the lineages studied are very divergent morphologically, analyses based exclusively on morphology may lead to conflicting and unexpected hypotheses. Through integration of data from conservative genes/gene regions the terminals including these data can anchor or constrain the search, thereby practically circumscribing the search space of the combined analyses. In this study, we revisit the phylogeny of a highly divergent group of mosses, class Polytrichopsida. We supplemented the morphological matrix by adding sequence data of the nuclear gene 18S, chloroplast genes *rbc*L and *rps*4, plus the mitochondrial gene *nad*5. For the phylogenetic analyses we used parsimony as the optimality criterion. Analyses that included all the terminals resulted in one most parsimonious tree with a clade comprised of *Alophosia azorica* and the fossil *Meantoinea alophosioides* representing the basal-most lineage. Analyses with different outgroup sampling produced the same topology for most ingroup relationships. An analysis excluding morphological characters and the four terminals for which only morphological characters were scored (the two fossil and two extant terminals) resulted in one optimal tree with identical topology to the one obtained when including all terminals. These results are largely congruent with those obtained in the recent analyses based exclusively on sequence level data of a larger number of terminals. Our results indicate that large size and complexity of the gametophyte have evolved independently in several lineages. Notably, the nodes of the backbone of the most parsimonious tree have very low support values, thus these inferred relationships could change if new additional information conflicts with the current data. Future studies should be aimed at incorporating all terminals into phylogenetic analyses, which is not an unrealistic goal for a group with less than 200 species. Also, additional fossils, some of which await detailed examination and description, need to be included. Whether these will affect the overall pattern of phylogeny presented here remains to be seen. In a group that is obviously very ancient,

we cannot assume, *a priori*, that currently known fossil taxa, which go back in time less than 140 Ma, represent the oldest lineages of the group.

Our understanding of organismal diversity and the most robust hypotheses about its history suggest that evolution by default produces divergence. It is also clear that extant organisms represent only a small fraction of the total biodiversity that has lived on this planet throughout the more than 3.5 Ga history of life (e.g. Marshall, 2017), thus, extinction was a key factor shaping extant diversity. Such rampant extinction leads to situations where defining homologies can be difficult. In turn, this leads to difficulties in phylogenetic analyses. Fossils are one the most widely recognized sources of information that can provide more insight in these situations. Numerous classic cases highlighting the importance of the fossils are provided, for example, by Forey & al. (1992), Smith (1994), Crane & al. (2004), Hilton & Bateman (2006), and Rothwell & Nixon (2006). On the other hand, when fossils are sparse and the lineages studied are very divergent morphologically, hypotheses based exclusively on morphology may lead to conflicting and unexpected hypotheses (e.g. Bippus & al., 2018). Sequence data of slowly evolving regions of the genome could offer a remedy in such cases. It is widely recognized that rates of evolution do vary between characters and, thus, when morphology is exceptionally divergent between lineages, at least some gene regions may still be highly similar. We know, for example, that the nuclear gene coding for 18S RNA is highly conserved, enabling comparisons of organisms even across major lineages (e.g. Hedderson & al., 1996, Hovmöller & al., 2002). Likewise, the chloroplast gene coding for the large subunit of rubisco (*rbc*L) is broadly conserved and was among the first genes widely used in phylogenetic studies of plants (e.g. Chase & al., 1993). The use of such highly conservative (i.e. slow-evolving) genes enables alignment of sequences and unequivocal assumptions about homology at the level of individual nucleotide positions. By using data from conservative genes/gene regions the terminals including these data can anchor or constrain the search, thereby practically circumscribing the search space of the combined analyses that integrate molecular and morphological data. As a result of this, more character state changes can be optimized on topologies; conversely, character state changes that were optimal based only on morphology might lose optimality upon combination of different data sources.

In this study, we revisit the phylogeny of a highly divergent group of mosses, class Polytrichopsida. We used a total evidence approach, following the example of Flores & al. (2017) wherein we supplemented the morphological matrix by adding sequence data into the analyses. Our analysis includes two fossils (*Eopolytrichum antiquum* Konopka & al. and *Meantoinea alophosioides* Bippus & al.) and over 40 extant species representing all genera of the Polytrichopsida, except for the recently described *Delongia* N. E. Bell & al. Because the class is morphologically isolated (i.e., highly divergent) from all other extant moss lineages, the root of the Polytrichopsida has been elusive, with different ingroup topologies recovered with different outgroup sampling regimes when only morphological characters are used in the phylogenetic analyses (Bippus & al., 2018). We now supplemented the morphological matrix with sequences of the

nuclear gene 18S, chloroplast genes *rbc*L and *rps*4, plus the mitochondrial gene *nad*5. All these genes have been used in previous phylogenetic analyses of the group (e.g. Bell & Hyvönen, 2010, Hyvönen & al., 1998) and thus sequences for almost all the extant terminals included by Bippus & al. (2018) were readily available. Most of the sequences used were those used by Bell & Hyvönen (2010), with the exception of *rbc*L gene. Obtaining the whole sequence for this region was problematic and only about half of the gene sequence was available in Bell & Hyvönen (2010). Accordingly, we downloaded *rbc*L sequences from the GenBank as listed in the Appendix 1. Sequences of the genes listed above were also downloaded for the outgroup terminals. In cases where more than one sequence had been uploaded to the repository we used the longest sequence available. In order to avoid random attraction, or repulsion, of the terminals by sequences of unequal length that are due to sequencing artefacts, we were very conservative in our choice of the regions to be used. We excluded regions adjacent to the 5´and 3´ ends of the gene sequences where we had only few representatives. For *nad*5 we used only 492 nt from the 3´ end that also allowed alignment with the sequences obtained from the outgroup terminals. In order to maximize positional homology for the sequences downloaded we performed alignment using ClustalX (Larkin & al., 2007) under default settings. For the phylogenetic analyses we used parsimony as the optimality criterion. Discussions of strengths and weaknesses of different optimality criteria in phylogenetic analysis abound in the recent literature, and are beyond scope of this study. Nevertheless, we refer the reader to Flores & al. (2017), and particularly to Goloboff & al. (2018) for such discussions relative to analyses using morphological characters.

Initially, we removed parsimony uninformative characters from gene regions using the "mop uninformative characters" function of Winclada (Nixon, 2002). This produced a matrix of 519 molecular characters. Of these, 83 nt were from the 18S, 238 from the *rbc*L, 158 from the *rps*4, and 40 nt from the 3´ end of the mitochondrial *nad*5 gene. This dataset was combined with the matrix of Bippus & al. (2018), which includes 100 morphological characters, with 11 of which were coded as continuous and additive (Goloboff & al., 2006), while all the other (discrete) characters were treated as non-additive (unordered). The small number of terminals (45) enabled analysis using traditional search algorithms of the program TNT (Goloboff & Catalano 2016), with gaps treated as missing data. We performed analyses with different outgroup sampling regimes, similar to those of Bippus & al. (2018), i.e. with only *Alophosia azorica* Card. of Polytrichaceae to root the tree; or various samplings of the outgroup terminals, i.e. with using *Oedipodium griffithianum* Schwaegr., *Sphagnum palustre* L. and *Tetraphis pellucida* Hedw, as outgroup terminals, or with the latter three supplemented by *Andreaea rupestris* Hedw., *Buxbaumia aphylla* Hedw., *Diphyscium foliosum* D. Mohr and *Funaria hygrometrica* Hedw. as outgroup terminals. All analyses were initiated with the random seed set to "0", i.e. the CPU time is used as the random seed to randomize the

order of the terminals. 1000 replicates of RAS (Wagner trees) were used in each search with 10 trees saved per replicate and TBR as a swapping algorithm. Tree searches were also performed with the same settings as used by Bippus & al. (2018). Support values were calculated using jackknife with the default values, i.e. with the results output as frequency differences (GC; Goloboff & al., 2003)

All analyses that included all the terminals included resulted in one optimal parsimonious tree with a length of 2089.52. In the other analyses we obtained exactly the same topology for most of the ingroup relationships. The only differences were present in the tree obtained using only three outgroup terminals (*Sphagnum palustre, Tetraphis pellucida* and *Oedipodium griffithianum*). In this tree, as compared to other trees, the fossil *Meantoinea alophosioides* was resolved as sister to the clade formed by two species of *Lyellia* R. Br., and not as sister to *Alophosia azorica*. Additionally, *Polytrichadelphus magellanicus* (Hedw.) Mitt. was resolved as sister to the clade formed by the two species of *Dawsonia* R. Br., while in the trees using other sampling sampling of the outgroup terminals *Polytrichadelphus* (Müll. Hal.) Mitt. and *Dawsonia* spp. formed a paraphyletic group basal to the large clade including most of the other genera (Fig. 1). An analysis excluding morphological characters and the four terminals for which only morphological characters were scored (the two fossil terminals plus *Pogonatum philippinense* (Broth.) Touw and *P. volvatum* (Müll. Hal.) Paris) resulted in one optimal tree with identical topology to the one obtained for all the 45 terminals.

It can be argued that in this analysis, morphology is swamped by sequence data because morphological characters are only ca. 16% of the total dataset. However, as discussed for example by Gatesy & al. (1999), the outcome of the total evidence analyses cannot be predicted from the sheer number, or proportions, of different types of characters and is, instead, determined by the amount of character congruence present in the dataset. Furthermore, as succinctly stated by Wheeler & al. (1993), the most logical explanation for the shared "signal" of diverse sources of information, whether morphological or sequence data, is their shared history. This perspective implies that combined analyses of molecular and morphological data provide the most robust hypotheses of relationships within a group (e.g. Flores & al., 2019). The results of our analyses show that the positions of the fossils and of the terminals that lack, at the moment, sequence level data are identical between the trees obtained with different outgroup sampling regimes. The only exception is the different placement of *Meantoinea alophosioides* and *Polytrichadelphus magellanicus* (Hedw.) Mitt. in a tree with only three outgroup terminals. However, it is important to note that this specific outgroup selection, which leaves out representatives of the largest clade of the peristomate mosses, cannot be defended. Therefore, we consider that the topology of Fig. 1 to represent our best estimate of the phylogeny of the major lineages of Polytrichopsida. This topology is largely congruent with the results obtained in analyses based exclusively on sequence level data of larger number of

terminals (Bell & Hyvönen, 2010). When compared to the results of analyses based exclusively on morphological characters (Bippus & al., 2018), it is evident that the most obvious conflict is the absence of the large gametophyte clade found in all analyses of the latter study. When the monophyly of this large gametophyte clade is enforced in our analyses, we obtain a tree that is 19 steps longer is obtained. Thus, our current analyses indicate that large size and complexity of the gametophyte has evolved independently in several lineages. It should be noted that while large size and overall leaf structure are shared by these groups, there are also significant differences between them. For example, the internal structure of the stem differs in most species of *Dawsonia* (hydroids mixed with sclerenchyma) from that observed in other large gametophytes of Polytrichaceae (Smith 1971). *Dawsonia* species also have a unique type of the peristome, different from that of other peristomate Polytrichaceae. On the other hand, it should be noted that the nodes of the backbone of our optimal tree obtained have very low support values, and thus, these relationships can be easily overturned if the new additional information conflicts with the current data.

Ideally, the potential of total evidence analyses that use highly conservative genes to help in situations with conflicting hypotheses derived from morphology should be "tested" in other groups of organisms with a more completely known fossil record. Our knowledge of the bryophyte fossil record has improved in the last decades (Tomescu & al., 2018), but extinct bryophytes are still scarce and poorly known as compared to other embryophytes and, thus, few are available for inclusion in this kind of studies. In comparison, seed plants are represented today only by a fraction of their former diversity, yet recent detailed studies like the one on the conifer family Araucariaceae by Escapa & Catalano (2013) demonstrate how this kind of approach can successfully yield robust phylogenetic hypotheses.

Future studies of Polytrichopsida should be aimed at incorporating all terminals into phylogenetic analyses. This is not an unrealistic goal with a group of this size, with approximately less than 200 species (Bell & Hyvönen, 2010). Ideally future analyses should also include multiple terminals of the species that have wide geographic ranges. At the same time the aim should also be to use also the data from more variable gene regions. Difficulties in aligning these regions, particularly in the case of outgroup terminals, have been the main reason for exclusion of outgroups from previous analyses of relationships within the group (e.g. Bell & Hyvönen, 2010) that rooted trees with *Alophosia azorica* on the basis of prior studies using conserved regions and wider sampling (e.g. Bell & Hyvönen, 2008). Ideally, future analyses can be performed using direct optimization (Wheeler, 1996) in order to avoid unwarranted assumptions about homology between nucleotide positions that occur when alignment and phylogenetic analyses are separated. Of equal importance, additional fossils, some of which await detailed examination and

description, need to be included in future analyses (e.g., Bippus & al., 2018). Whether these fossils will affect the overall pattern of phylogeny presented here remains to be seen. As we have seen with *Eopolytrichum*, in a group that is obviously very ancient, we cannot assume, *a priori*, that the currently known fossils, which go back in time less than 140 Ma, represent the oldest lineages of the group.

Acknowledgements

We thank Kevin Nixon and Willi Hennig Society for making the programs Winclada and TNT freely available.

References

Bell, N. E. & Hyvönen, J. 2008. Rooting the Polytrichopsida: the phylogenetic position of *Atrichopsis* and the independent origin of the polytrichopsid peristome. In: Mohamed, H. & al. (eds.) Bryology in the new millennium. Kuala Lumpur: University of Malaya, pp. 227-239.

Bell, N. E. & Hyvönen, J. 2010. Phylogeny of the moss class Polytrichopsida (Bryophyta): generic level structure and incongruent gene trees. *Molecular Phylogenetics and Evolution* 55: 381–398.

Bippus, A., Escapa, I. E. & Tomescu, A. M. F. 2018. Wanted dead or alive (probably dead): stem group Polytrichaceae. *American Journal Botany* 105: 1-21.

Chase, M. & al. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbc*L. *Annals of the Missouri Botanical Garden* 80: 528-580.

Crane, P. R., Herendeen, P. & Friis, E. M. 2004. Fossils and plant phylogeny. *American Journal Botany* 91: 1683-1699.

Escapa, I. H. & Catalano, S. A. 2013. Phylogenetic analysis of Araucariaceae: integrating molecules, morphology, and fossils. *International Journal of Plant Sciences* 174: 1153–1170.

Flores, J. R., Catalano, S. A., Muñoz, J. & Suárez, G. M. 2017. Combined phylogenetic analysis of the subclass Marchantiidae (Marchantiophyta): towards a robustly diagnosed classification. *Cladistics* 34: 517-541.

Flores, J. R., Suarez, G., & Hyvönen, J. 2019. Reassessing the role of morphology in bryophyte phylogenetics: Combined data improves phylogenetic inference despite character conflict. *Molecular Phylogenetics and Evolution* 143: 106662.

Forey, P. L., Humphries, C. J., Kitching, I., Scotland, R. W., Siebert, D. J. & Williams, D. M. 1992. Cladistics: a practical course in systematics. Oxford Univ. Press.

Gatesy, J., O´Grady, P. & Baker, R. H. 1999. Corroboration among data sets in simultaneousanalysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics* 15: 271-313.

Goloboff, P. A. & Catalano, S. A. 2016. TNT version 1.5, including a full implementation of phylogenetic morphometrics. *Cladistics* 32: 221–238.

Goloboff, P., Farris, J. S., Källersjö, M., Oxelman, B., Ramirez, M. J. & Szumik, C. A. 2003. Improvements to resampling measures of group support. *Cladistics* 19: 324-332.

Goloboff, P. A., Mattoni, C. I. & Quinteros, A. S. 2006. Continuous characters analyzed as such. *Cladistics* 22: 589-601.

Goloboff, P. A., Torres, A. & Arias, J. S. 2018. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics* 34: 407-437.

Hedderson, T. A., Chapman, R. L. & Rootes, W. L. 1996. Phylogenetic relationships of bryophytes inferred from nuclear-encoded rRNA gene sequences. *Plant Systematics & Evolution* 200: 213-224.

Hilton, J. & Bateman, R. M. 2006. Pteridosperms are the backbone of seed-plant phylogeny. *Journal of the Torrey Botanical Society* 133: 119-168.

Hovmöller, R., Pape, T. & Källersjö, M. 2002. The Palaeoptera problem: basal pterygote phylogeny inferred from 18S and 28S rDNA sequences. *Cladistics* 18: 313-323.

Hyvönen, J., Hedderson, T. A., Smith Merrill, G.L., Gibbings, J.G. & Koskinen, S., 1998. On phylogeny of the Polytrichales. *Bryologist* 101: 489–504.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947-2948.

Marshall, C. R. 2017. Five palaeobiological laws needed to understand the evolution of the living biota. *Nature Ecology & Evolution* 1: 0165.

Nixon, K.C., 2002. Winclada, Version 1.00.08. Software published by the author, Ithaca, NY. Available online @ www.diversityoflife.org/winclada/

Rothwell, G. R. & Nixon, K. C. 2006. How does the inclusion of fossil data change our conclusions about the phylogenetic history of euphyllophytes? *International Journal of Plant Sciences* 167: 737–749.

Smith, A. B. 1994. Systematics and the fossil record: documenting evolutionary patterns. Oxford: Blackwell.

Smith, G. L. 1971. Conspectus of the genera of Polytrichaceae. *Memoirs of the New York Botanical Garden* 21: 1-83.

Tomescu, A. M. F., Bomfleur, B., Bippus, A. C., Savoretti, A. 2018. Why are bryophytes so rare in the fossil record? A spotlight on taphonomy and fossil preservation. In: Transformative Paleobotany, Academic Press, 375-416 pp.

Wheeler, W. C. 1996. Optimization alignment: The end of multiple sequence alignment in phylogenetics. *Cladistics* 12: 1–9.

Wheeler, W. C., Cartwright, P. & Hayashi, C. Y. 1993. Arthropod phylogeny: a combined approach. *Cladistics* 9: 1-39.

Appendix

GenBank accessions nos. of the sequences used in the analyses supplementing those used in Bell & Hvyönen (2010). Outgroup terminals listed first, followed by the ingroup and the nos. of their *rbc*L used in the analysis.

	nc 18S	cp *rbc*L	*rps*4	mt *nad*5	
Andreaea rupestris	U18490	AB469555	AF478248	AJ001227	
Buxbaumia aphylla	Y17603	GQ368610	AF231897	KC662862	
Diphyscium foliosum	Y17765	AF478220	AF223034	AY312874	
Funaria hygrometrica	X74114	AF226818	AJ250120	JF501564	
Oedipodium griffithianum	AF228668	AF478202	AF478255	AY312880	
Sphagnum palustre	AF126290	AF231887	MF362454	KC662871	
Tetraphis pellucida	Y17604	U87091	AY908021	KC662872	
Timmia sibirica	AF023678	AJ275166	AF023775	--	-

Alophosia azorica AY312924, *Atrichopsis compressa* EU927307, *Atrichum angustatum* DQ645986, *A. crispum* KP881775, *A. undulatum* AB917062, *Bartramiopsis lescurii* AF208409, *Dawsonia papuana* AF208410, *D. superba* AY118237, *Dendroligotrichum dendroides* AF208411, *Hebantia rigida* AY118240, *Itatiella ulei* AF208412, *Lyellia aspera* AF208413, *L. crispa* JX241626, *Notoligotrichum australe* AF208414, *N. crispulum* GU569425, *Oligotrichum hercynicum* AY118243, *O. parallelum* AF208415, *Pogonatum camusii* KU852692, *P. contortum* AY118247, *P. nudiusculum* ////////, *P. pensilvanicum* AY118253, *P. piliferum* ////////, *P. sinense* DQ120779, *P. urnigerum* AY118256, *Polytrichadelphus magellanicus* AY118257, *Polytrichastrum alpinum* GU569464, *P. lyallii* AY118241, *Polytrichum formosum* AY118259, *P. longisetum* AY118260, *P. commune* U87087, *P. piliferum* AY118263, *Psilopilum laevigatum* AF208416, *Steereobryon subulirostrum* AY118265

Figure 1. Tree obtained from combined analysis of the morphological (11 of these continuous) and sequence level (nc 18S 83, cp *rbc*L 238, *rps*4 158, mt *nad*5 40 nt) data, length 2089.52 steps. Jackknife support values >50 marked below branches.