New Phytologist

Phylogenomic structure and speciation in an emerging model: the Sphagnum magellanicum complex (Bryophyta)

Journal:	*New Phytologist*
Manuscript ID	Draft
Manuscript Type:	Full Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Shaw, Jonathan; Duke University, Department of Biology; Duke University, Biology Piatkowski, Bryan; Oak Ridge National Laboratory, Biosciences Division Duffy, Aaron; Duke University, Department of Biology; Duke University Aguero, Blanka; Duke University, Department of Biology Imwattana, Karn; Duke University, Department of Biology Nieto-Lugilde, Marta; Duke University, Department of Biology Healey, Adam; 3HudsonAlpha Institute of Biotechnology, Computational Biology Weston, David; Oak Ridge National Laboratory, Biosciences Patel, Megan; Oak Ridge National Laboratory, Biosciences Schmutz, Jeremy; HudsonAlpha Institute for Biotechnology, Genome Sequencing Center Grimwood, Jane; HudsonAlpha Institute for Biotechnology, Genome Sequencing Center Yavitt, Joseph; Cornell University, Dept. Natural Resources Hassel, Kristian; Norwegian University of Science and Technology, Department of Natural History Stenøien, Hans; Norwegian University of Science and Technology, Systematics and Evolution Group Flatberg, Kjell-Ivar; Norwegian University of Science and Technology, Department of Natural History Bickford, Christopher; Kenyon College, Biology Hicks, Karen; Kenyon College, Department of Biology;
Key Words:	peat mosses, bryophytes, ecological genomics, introgression, peatlands, speciation, Sphagnum
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
Fig1_map.tif Fig2_Cophylogeny.tif Fig4_Splitstree.tif Fig5_RADseqPhylogeny.tif Fig6_STRUCTURE.tif	

Phylogenomic structure and speciation in an emerging model: the *Sphagnum magellanicum* complex (Bryophyta)

A. Jonathan Shaw[1], Bryan Piatkowski[2], Aaron M. Duffy[1], Blanka Aguero[1], Karn Imwattana[1], Marta Nieto-Lugilde[1], Adam Healey[3], David J. Weston[2], Megan N. Patel[2], Jeremy Schmutz[3,4], Jane Grimwood[3], Joseph B. Yavitt[5], Kristian Hassel[6], Hans K. Stenøien[6], Kjell-Ivar Flatberg[6], Christopher P. Bickford[7], Karen A. Hicks[7]

[1]Duke University, Department of Biology, Durham, NC, 27708, USA; [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA; [3]HudsonAlpha Institute of Biotechnology, Huntsville, AL 35806, USA; [4]Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA; [5]Cornell University, Department of Natural Resources, Ithaca, NY, 14853, USA; [6]NTNU University Museum, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway; [7]Kenyon College, Department of Biology, Gambier, OH 43022

Corresponding author: A. Jonathan Shaw

Email: shaw@duke.edu

Total Words: 6475

Introduction Words: 787

Methods Words: 2183

Results Words: 2464

Discussion Words: 1032

Figures: 6 (in-print). Figs. 1,2, 4-6 in color

Supplemental materials: Figs. S1-S17; Tables S1-S8

Summary

- Sphagnum *magellanicum* is one of two *Sphagnum* species for which a reference-quality genome exists to facilitate research in ecological genomics.
- Phylogenetic and comparative genomic analyses were conducted based on resequencing data from 48 samples and RADseq analyses based on 187 samples.
- We report herein that there are four clades/species within the *S. magellanicum* complex in eastern North America, and that the reference genome belongs to *S. divinum*. The species exhibit tens of thousands (RADseq) to millions (resequencing) of fixed nucleotide differences. Two species, however, referred to informally as S. diabolicum and S. magni because they have not been formally described, are differentiated by only hundreds (RADseq) to thousands (resequencing) of differences. Introgression among species in the complex is demonstrated using *D*-statistics and f_4-ratios. One ecologically important functional trait, tissue decomposability, which underlies peat (carbon) accumulation does not differ between segregates in the *S. magellanicum* complex, although previous research showed that many closely related *Sphagnum* species have evolved differences in decomposability/carbon sequestration.
- Phylogenetic resolution and more accurate species delimitation in the *S. magellanicum* complex substantially increase the value of this group for studying the early evolutionary stages of climate adaptation, and ecological evolution more broadly.

Key words: peat mosses, bryophytes, ecological genomics, introgression, peatlands, speciation, *Sphagnum*

INTRODUCTION

Understanding phylogenetic relationships is critical to the development and value of model organisms. In the angiosperms, recognizing close relatives of *Arabidopsis thaliana* has enabled research on genome evolution and speciation across closely related species and genera of Brassicaceae (Beilstein *et al*., 2010; Koch, 2019; Nikolov *et al*., 2019; Brukhin *et al*., 2019). Recent phylogenetic work on the moss family Funariaceae has provided new insights into relationships and evolution of the widely utilized model, *Physcomitrium patens* (formerly *Physcomitrella patens*), opening new lines of research into the evolution of plant life cycles and morphological evolution (McDaniel *et al*., 2009; Beike *et al*., 2014).

The moss genus *Sphagnum* (peatmoss) comprises some 250-450 species and a genus-wide genome sequencing project is underway to assess inter- and intraspecific phylogenetic relationships and genome evolution (Weston *et al*., 2018). An overarching goal of developing *Sphagnum* as a model is to link genome variation with phenotypic traits of ecological importance. Toward that end, a reference quality genome is available for a species thought to be widespread globally, *S. magellanicum*. However, we now know that *S. magellanicum* sensu stricto is restricted to South America and that Northern Hemisphere plants formerly named *S. magellanicum* comprise multiple closely related and morphologically similar taxa (Yousefi *et al*., 2017; Hassel *et al*., 2018). The current research was designed to resolve phylogenetic relationships within the *S. magellanicum* complex to further enable development of this important resource for ecological genomics and speciation research. We sampled plants from around the Northern Hemisphere as well as in Central and South America where they are abundant and important components of high elevation peatlands south to Tierra del Fuego (where they occur at sea level). We investigate phylogenetic divergence at whole-genome and chromosomal scales and show that the genome data generated to represent "*S. magellanicum*" (from Minnesota) belong to the segregate species, *S. divinum*. We also show that the *S. magellanicum* complex comprises at least four species-level clades in North America, with additional clades in South America and Asia. Segregate species in the complex differ in geographic ranges and the climate zones they occupy, and co-occurring species typically occupy different niches that vary relative to local-scale hydrological, light, and nutrient gradients. Because of this variation, and well-supported phylogenetic resolution for the group, the *S. magellanicum* complex takes on even greater value as a model for ecological and evolutionary genomics. We further argue herein that speciation in the group is at an early stage, which

further empowers the group for research on the evolutionary origins of traits that scale up to impact global ecosystems.

Sphagnum-dominated peatlands are estimated to cover about 10% of the Northern Hemisphere boreal zone yet store approximately 25-30% of the total global terrestrial carbon pool (Yu *et al*., 2012). *Sphagnum* grows most abundantly in bogs and fens where they engineer peatland habitats in ways that promote their own persistence and dominance. *Sphagnum* species in general have low decomposition rates that promote carbon accumulation (Rydin & Jeglum, 2013). Peatlands typically have a hummock-hollow (mound-valley) physiognomy created by the peatmosses themselves, as hummock-forming species grow relatively slowly but decompose even more slowly such that hummocks are formed through the accumulation of partially decomposed plant material (*Sphagnum* peat). Hollow-inhabiting species, in contrast, grow quickly but also decompose rapidly (preventing hummock formation) and the plants lie close to or at the water table (Rydin & Jeglum, 2013, Piatkowski *et al*., 2021). A second microenvironmental gradient within peatlands is the poor-rich axis (pH and nutrients), also engineered by *Sphagnum* itself through metabolic processes and cation exchange. Even closely related species of peatmoss vary in traits that scale up to impact ecosystem processes (e.g., Piatkowski *et al*., 2021). In other words, *species matter*. *S*pecies-specific niche differentiation allows 20 or more *Sphagnum* species to occur sympatrically in some peatlands. Because of the well-documented niche differences among species and the relatively simple structure of *Sphagnum*-dominated peatlands, the plants and their ecology have long served as a model for studies of community assembly and structure (e.g., Vitt & Slack, 1984; Rydin & Jeglum, 2013).

Several of the clades resolved in our phylogenetic research do not yet have taxonomic names. We refer to the three well-documented and published species by their established binomials, *S. divinum* Flatberg & K. Hassel, *S. magellanicum* Brid., and *S. medium* Limpr. and refer to unpublished taxa informally without italics; i.e., S. asiaticum, S. diabolicum, S. magni, S. magellanicum-NW. The formal taxonomic name, *S. magellanicum*, is used for plants derived from Tierra del Fuego (Chile, Argentina) as that is where the species was described from. Plants from northern South America and Central America form a distinct clade and are referred to as S. magellanicum-NW. Formal taxonomic establishment of these unnamed clades as species will follow in a subsequent publication.

MATERIALS AND METHODS

Plant Sampling - We obtained whole genome data from 48 accessions representing the *S. magellanicum* complex plus two outgroup taxa in the subgenus *Sphagnum*, *S. affine* Renauld & Cardot and *S. perichaetiale* Hampe. RADseq data were obtained from 185 samples representing the complex plus the same two outgroups. Of the 187 samples included in the RADseq data set, 144 were derived from new extractions and 43 represented *in silico* digests from the resequencing data so those samples could be included in both data sets. Sixteen of the new extractions were from the same collection as an *in silico* digest to test for differences between these methods of generating RADseq data. Samples representing the *S. magellanicum* complex came most abundantly from eastern North America but with additional samples from Central and South America, Europe, Siberia, China, Taiwan, and Japan. Sample localities are shown in Fig. 1. Voucher specimens for each extraction are deposited in the Duke University Herbarium (DUKE). Extractions for RADseq utilize only part of an individual gametophyte; any remaining parts of that plant were placed in a smaller envelope and returned to the specimen herbarium packet. As whole plants were required for the resequencing, each sample is vouchered by a sample from the same clump (handful). A complete list of samples with locality information is provided as Table S1.

DNA extraction, library preparation, and sequencing – Genomic DNA was extracted from each sample using the CTAB protocol as described in Shaw *et al*. (2003). DNA purity was measured with Nanodrop, DNA concentration measured with Qubit HS kit, and DNA size was validated by pulsed field gel electrophoresis. Illumina libraries for the references were prepared as Tight Insert Fragment libraries, 400bp - 2 ug of DNA was sheared to 400 bp using the Covaris LE220 and size selected using the Pippin (Sage Science). The fragments were treated with end-repair, A-tailing, and ligation of Illumina compatible adapters (IDT, Inc) using the KAPA-Illumina library creation kit (KAPA biosystems). The prepared libraries were quantified using KAPA Biosystem's next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The quantified libraries were then prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq Rapid paired-end cluster kit, v2, with the HiSeq2500 sequencer instrument to generate a clustered flowcell for sequencing. Sequencing of the flowcell was performed on the Illumina HiSeq2500 sequencer using HiSeq Rapid SBS sequencing kits, v2, following a 2x150 indexed run recipe.

For RADseq analyses, aliquots of five extractions were included twice in the library as duplicates to allow identification of likely clones in the sample set. RADseq libraries were

prepared following the double digestion restriction site-associated DNA sequencing (ddRADseq) protocol of Parchman *et al*. (2012) with the modifications described in Duffy *et al*. (2020). Each library was cleaned and size-selected for fragments of approximately 350bp using AMPur XP beads (Beckman Coulter), inspected for quality on a BioAnalyzer (Agilent), and sequenced on a single lane of Illumina NextSeq 500 with 150bp single-ended reads at the Genome Sequencing Shared Resource at the Duke Center for Genomic and Computational Biology (https://oit.duke.edu/comp-print/research/).

SNP discovery using genome resequencing data **-** Illumina reads were screened for PhiX and organellar contamination. Reads composed of greater than 95% simple sequence were removed. Libraries were aligned to the *S. divinum* v1.1 reference genome (https://phytozome-next.jgi.doe.gov/info/Smagellanicum_v1_1) using bwa-mem 0.7.12 (Li & Durbin, 2009). Duplicates in these BAM files were marked using Picard v2.2.6.2 (http://broadinstitute.github.io/picard/). Autosomal variants were called for individual samples using HaplotypeCaller and joint genotyping was performed for the cohort using GenotypeGVCFs in GATK v4.2.2.0 (Van de Auwera & O'Connor, 2020). Samples were treated as haploid because this is the dominant phase of the bryophyte life cycle. Biallelic SNPs were separated from indels and invariant sites. The SNPs were then filtered using the following criteria: QD ≥ 2.0, MQ ≥ 40.0, FS ≤ 60.0, SOR ≤ 3.0, MQRankSum ≥ -12.5, and ReadPosRankSum ≥ -8.0. Four datasets were generated that included biallelic sites genotyped in at least 70%, 80%, 90%, and 100% of samples to allow for investigation of the sensitivity of downstream analyses to missing data. For our phylogenetic analyses, variant sites were further filtered to retain those with a minor allele frequency of at least 0.10 and then pruned for linkage disequilibrium (LD) using PLINK v1.90b6.24 (Chang et al. 2015) with a window size of 50 variants, a window shift of 10 variants at the end of each pruning step, and a variance inflation factor threshold of 1.5.

RADseq processing – Raw Illumina reads were quality checked with FastQC v0.11.9 (Andrews, 2010) and RADseq loci were identified with ipyrad v.0.9.50 (Eaton, 2014) using default settings except as noted here. The *in silico* digested reads from 43 genomic resequencing samples were included with the Illumina reads of 149 samples (including duplicates) after the demultiplexing step. Reads were filtered for adapter sequences or low-quality bases, trimmed to a maximum of 92 bases after removing the barcode, and filtered to remove reads shorter than 35bp after trimming. Loci were identified using the reference

assembly method against the *S. divinum* reference genome. Samples were treated as haploid. Three datasets were generated including loci present in at least 70, 80, and 90% of samples to allow for investigation of the sensitivity of downstream analyses to missing data or number of loci.

To generate RADseq-like data from genomic resequencing assemblies, each resequencing assembly was digested *in silico* with EcoRI and MseI using the program "restrict" from the EMBOSS package (Rice *et al*., 2000). Contigs from Illumina libraries were assembled using HipMer (Georganas *et al*., 2015) with a kmer size of 51. Custom scripts were used to filter for digested sequence fragments with an EcoRI cutsite at one end and an MseI cutsite at the other, to mimic the size-selection steps of a RADseq library preparation, to trim the fragments to match the length of Illumina reads, and to write the sequences to a FASTQ formatted file. Each resulting "read" was given a quality score and number of copies sufficient to pass filters during downstream processing.

Phylogenetic reconstruction using nuclear data - A combination of maximum likelihood, distance, and multi-species coalescent methods were used to reconstruct phylogeny. SNPs from genome resequencing data that had been pruned for LD were analyzed using IQ-TREE2 (Minh *et al*., 2020a), SplitsTree (Huson & Bryant, 2006), ASTRAL (Zhang et al. 2018), and SVDquartets (Chifman & Kubatko, 2014). For the maximum likelihood analyses in IQ-TREE2 v2.1.2, we used ModelFinder (Kalyaanamoorthy *et al*., 2017) to perform automatic model selection and kept the most likely tree from 10 independent runs. Model selection incorporated corrections for ascertainment bias (Lewis, 2001). Bipartition support was assessed using the ultrafast bootstrap method with 1E3 pseudoreplicates (Hoang *et al*., 2018) and site concordance factors were estimated from analysis of 1E2 quartets (Minh *et al*., 2020b). To visualize conflicting signals within the dataset, we used SplitsTree v.4.17.1 to estimate a NeighborNet phylogenetic network based on K2P genetic distance with the dataset in which at least 80% of the samples were genotyped at each site. SVDquartets in PAUP* v4.0a build 166 for Unix/Linux was used to identify relationships among species under the multispecies coalescent model and support was evaluated with 200 standard non-parametric bootstrap pseudoreplicates. Additional coalescent analyses were performed using ASTRAL v5.7.8 to reconstruct species trees from the maximum likelihood genealogies of 100-kb non-overlapping genomic windows. To explore how phylogenetic signal was distributed throughout the genome, we also performed the likelihood and coalescent analyses on individual chromosomes using the dataset in which at least 80% of the samples were genotyped at each site. We performed approximately unbiased

(AU) tree topology tests (Shimodaira, 2002) using each dataset in IQ-TREE2 with 1E4 bootstrap replicates generated using the resampling estimated log likelihoods (RELL) method to evaluate statistical support for the various species relationships recovered across analyses. Maximum likelihood and SVDquartets analyses were also performed for the RADseq data.

Phylogenetic reconstruction using plastid data – Plastid reads from 49 resequenced genomes were identified and assembled into contigs using NOVOPlasty (v2.6.7- plastid range 120,000-200,000) (Dierckxsens *et al*., 2017). One resequencing library, IUXC, did not have a plastid assembly. For each genome, contigs were manually aligned to the published *S. palustre* plastid genome (KU726621) and to each other to identify the Inverted Repeat boundaries and generate a single incomplete plastid genome sequence (with missing data represented by strings of Ns) including the Long Single Copy region, one copy of the Inverted Repeat, and the Small Single Copy region. Sequences were aligned with MAFFT v7.490 (Katoh & Standley, 2013) and used to infer phylogeny under maximum likelihood with IQ-TREE2 as described in the preceding section. A cophylogenetic plot showing topological differences between nuclear and plastid phylogenies was generated using the *R* package *phytools* v0.7-90 (Revell, 2012). AU tests were performed for each nuclear resequencing dataset (70%-100% of samples genotyped) to determine the significance of topological differences between nuclear and plastid phylogenies.

Cluster analyses – Using RADseq data, one SNP per locus was used for clustering analyses to analyze SNPs in putative linkage equilibrium. after removing data from the outgroup samples, extraction duplicates, and likely clones. Clones cannot be directly identified from RADseq data due to error, but samples were considered likely clones if they were from the same collection or site, sister to each other in the maximum likelihood analysis, and separated by branch lengths similar to those of the extraction duplicates. Genetic structure among the remaining 141 samples was explored using Bayesian model-based cluster analysis with STRUCTURE v2.3.4 (Pritchard *et al*., 2000). The method of Evanno *et al*. (2005) as implemented in structureHarvester vA.2 (Earl & vonHoldt, 2012) was used to evaluate the most likely number of clusters (*K*), based on ten independent runs using an admixture model with correlated allele frequencies. Each value of *K* from 1 to 10 was evaluated with 1E5 steps of burn-in and 1E6 iterations per run. Matrices of membership coefficients across the multiple runs were used to search for the optimal alignment in CLUMPP version 1.1.2 (Jakobsson & Rosenberg, 2007). A second dataset with groups downsampled to reduce uneven sampling was analyzed to ensure

this was not biasing inferences. For groups with over 15 samples, 15 samples were selected to minimize missing data and maintain geographic coverage.

Isolation by distance (IBD) analysis – Using RADseq data, the relationship between genetic distance and the log of geographic distance (IBD) was explored within each of the named groups in the *S. magellanicum* complex and for the S. diabolicum+S. magni clade. One SNP per locus was selected and within-site likely clones were removed. IBD was tested using Mantel tests with the R package *adegenet* (Jombert, 2011) with 1E3 replicates.

Analysis of introgression - We used the program Dsuite v0.4 r38 (Malinsky *et al*., 2021) to detect signals of admixture among samples in the *S. magellanicum* species complex. This analysis used the genome resequencing dataset with at least 80% of individuals genotyped. We calculated D_{min} (Malinsky *et al*., 2018) for each species trio, which represents the minimum *D* statistic across all possible topologies and tested the null hypothesis that there is no excess allele sharing. We also calculated f_4-ratios that represent the fraction of the genome that is introgressed. Significance of the D_{min} statistic was determined using 1E2 jackknife blocks and the resulting *P*-values were adjusted using the Benjamini-Hochberg procedure. The *f*-branch statistic was calculated with both maximum likelihood and coalescent species tree topologies to determine whether gene flow might have involved internal branches of the phylogeny.

Genetic differentiation and diversity statistics – We estimated relative genetic differentiation between species using Hudson's F_{ST} (Hudson *et al*., 1992; Bhatia *et al*., 2013) for both genome resequencing and RADseq datasets with at least 80% of samples genotyped. Genome resequencing data were filtered for a minor allele frequency of at least 0.10, were not pruned for linkage disequilibrium, and included invariant sites from the GATK pipeline that were absent in our phylogenetic analyses. We also estimated absolute genetic divergence (d_{XY}) between species and nucleotide diversity (π) within species. All three statistics were calculated using pixy v1.2.5.beta1 (Korunes & Samuk, 2021) for 10-kb non-overlapping windows of the genome. We report median values from across these windows and provide a 95% confidence interval from bootstrapping with 1E4 pseudoreplicates using the R package *boot* v1.3-28 (Canty & Ripley, 2021).

We also sought to estimate the number of fixed differences between species in the *S. magellanicum* complex. For genome resequencing data, we used bcftools v1.13 (Danecek *et al*., 2021) and VCFtools v0.1.17 (Danecek *et al*., 2011) to subset and identify loci that were

alternatively fixed in all samples of each species pair with no missing data allowed. For RADseq data, we used a custom script to identify SNPs present in 50% or more of samples from both species of a species pair and alternatively fixed.

Microhabitat occupancy and decomposability – In order to assess whether species in the complex differ in microhabitat as described for *S. divinum* and *S. medium* in Europe (Yousefi *et al*., 2017) we tallied microhabitat data for samples for which collection notes included a statement about whether the collection was made in an "open bog", "bog margin", or [surrounding] "forest". These were a subset of all samples, comprising specimens collected by AJS in New England. Most other samples did not include this information specifically.

As even closely related *Sphagnum* species that differ relative to the hummock-hollow gradient are known to differ in decomposability and peat formation (Bengtsson *et al*., 2016; Piatkowski *et al*., 2021), we measured tissue decomposability in the field at the McLean Bog (Tompkins Co., NY), where a larger genus-wide experiment was undertaken in 2017-2019 using comparable methodologies (Piatkowski *et al*., 2021). The sampling for this experiment included 56 samples representing the four North American species in the *S. magellanicum* complex. Dried litter was placed in 5 x 5 cm fiber 25-micron mesh bags (Product F57, ANKOM Technology), with two technical replicates for most samples. Litter bags were buried in McLean bog just beneath the surface of living plants and left to decompose for approximately 1.85 years. Following harvest of the litter bags, we calculated the exponential decay constant (*K,* yr-1) for each sample using percent mass loss data (Olson, 1963; Turetsky *et al*., 2008). Analysis of variance was performed using *R* v4.1.1 (R Core Team, 2021) to test whether species differed in tissue decomposability.

RESULTS

Genome-wide phylogenetic relationships

Genome resequencing: Analyses were conducted with four genome-wide data sets that varied in amounts of missing data (Table S2). Ancestor-descendent relationships differ among data sets (Figs. S1-3). The genome resequencing data do not include two taxa that are resolved in RADseq analyses as separate clades, S. asiaticum and *S. magellanicum* (from Tierra del Fuego, where the species was described). Phylogenetic relationships among samples in the *S. magellanicum* complex are shown in Fig. 2 (left: nuclear data, right: plastid data). For clarity,

clade support values are not included in that figure but results for nuclear data are provided in Figs. S1-3.

In nearly all phylogenetic analyses, *S. medium* is resolved as reciprocally monophyletic to a clade containing the other taxa among which there exist three plausible histories. (1) Plants from northern South America (S. magellanicum-NW) are sister to a clade containing *S. divinum,* S. diabolicum*,* and S. magni. These relationships appear most frequently in the ML analyses. (2) *S. divinum* is sister to a clade containing S. magellanicum-NW, S. diabolicum, and S. magni. This topology comes from the 70% dataset under ML, most of the coalescent analyses, and is in the credible set of trees for the 90% dataset under ML based on AU tree topological tests. (3) S. magellanicum-NW and *S. divinum* form a clade reciprocally monophyletic to S. magni plus S. diabolicum. This topology comes from the ML analysis of the 70% and 80% RADseq datasets and the SVDquartets coalescent analysis of the 90% resequencing dataset, the 100% resequencing dataset, and several chromosome resequencing datasets. A summary of these analyses is presented in Fig. 3. Despite ambiguities in the backbone, all named groups (e.g., S. magni or *S. divinum*) are resolved as monophyletic with high support (Fig. S1). These inferences are corroborated by the SplitsTree analysis (Fig. 4) which also demonstrates substantial topological conflict.

Analysis of plastid data resolve S. magellanicum-NW as sister to the rest of the complex (Fig. 2; support values shown in Fig. S4), a topology seen in analysis of the 100% nuclear dataset under maximum likelihood. The only taxon that forms a clade in the plastid-only data is S. magellanicum-NW. Most samples of *S. divinum* fall within a single clade that also contains two samples of S. diabolicum. Most samples of *S. medium* form a clade, but one sample is placed in a clade that includes all S. magni and most of S. diabolicum. There appears to be no differentiation between S. diabolicum and S. magni in terms of plastid sequences. The AU tree topology tests provide strong statistical support for the incongruence between plastid and nuclear phylogenies: the plastid tree is rejected for the nuclear genome resequencing datasets and the nuclear trees are rejected for the plastid dataset ($P < 0.001$ in every test).

RADseq analyses: We initially analyzed three data sets varying in levels of missing data (Table S3, Fig. S5). RADseq loci are distributed across all of the 19 chromosomes (Fig. S6). A maximum likelihood reconstruction with all samples is shown in Fig. S7, with a summary of relationships among clades for the 80% sample coverage dataset provided in Fig. 5. RADseq data support monophyly of the same clades as resolved from genome resequencing: *S. divinum*, S. *medium*, S. magni, S. diabolicum, and S. magellanicum-NW. In addition, RADseq

analyses provide evidence of additional clades that were not sampled for resequencing: *S. magellanicum* s. str. (from Tierra del Fuego) and S. asiaticum from China and Taiwan. Reconstructions based on maximum likelihood versus the coalescent methods are completely congruent except for the placement of the *S. divinum* and S. magellanicum-NW clades (Fig. 5). In the coalescent reconstruction these two clades are sister whereas in the likelihood reconstruction they comprise steps in a paraphyletic grade leading to S. diabolicum and S. magni. Both reconstruction methods indicate that S. asiaticum is sister to the rest of the complex, *S. medium* is then sister to the remaining clades, and Fuegan *S. magellanicum* is sister to S. diabolicum and S. magni from eastern North America. These inferences are largely compatible with the results from genome resequencing, notwithstanding the absence of samples representing the S. asiaticum and *S. magellanicum* clades in our genome data set.

Chromosome-level phylogenetic relationships

We assessed whether inferences about phylogenetic relationships varied across the genome by reconstructing relationships at the chromosomal level. Chromosome by chromosome reconstructions from genome resequencing are illustrated in Figs. S8 (likelihood), S9 (SVDquartets), and S10 (ASTRAL). RADseq-based chromosome-level reconstructions are shown in Figs. S11 (likelihood) and S12 (SVDquartets).

Resolved topologies varied with chromosome (1-19), dataset (resequencing, RADseq), and analytical method (coalescent, likelihood). With all seven clades represented, RADseq data provide evidence for nine topologies across chromosomes with varying levels of support (Fig. S13). RADseq analyses of six individual chromosomes resolve S. diabolicum as reciprocally monophyletic relative to S. magni, but all topologies converge on the inference that these two groups are at least partially divergent. All RADseq analyses also converge on a close relationship between S. diabolicum+S. magni and the South American plants. The most likely tree for loci mapped to chromosome 3 is alone in strongly supporting a sister group relationship between the S. diabolicum+S. magni clade and *S. divinum*, rather than with South American samples. *Sphagnum medium* is resolved as sister to the rest of the complex for some chromosomes and S. asiaticum for other chromosomes, but inferences about specific chromosomes are hampered by discordance between results across analytical methods.

Genome resequencing data identify three topologies supported by different chromosomes (Fig. 3). The first two differ in the placement of *S. divinum* versus S. magellanicum-NW as sister to the S. diabolicum+S. magni clade and the third suggests a clade containing both *S. divinum* and S. magellanicum-NW is sister to the S. diabolicum+S. magni

clade. Comparing results from three analytical methods for the resequencing data and two for RADseq, most chromosomes yield discordant results whereas for six, inferences are concordant across analyses (Fig. 3). Genome resequencing lacked several of the clades (species) resolved by the RADseq data, so we compare phylogenetic inferences from the two data sets with a focus on the S. diabolicum+S. magni subclade because these two taxa are especially closely related.

Based on maximum likelihood analysis of genome resequencing data, all 19 chromosomes indicate that S. magni and the more inclusive S. diabolicum+S. magni clade are both monophyletic (Table S4). All but five chromosomes resolve the reciprocal monophylly of S. diabolicum and S. magni; chromosome LG04 suggests monophyly but the branch subtending S. diabolicum is not supported while the other 4 chromosomes indicate that S. diabolicum is paraphyletic, with S. magni nested within it. Based on RADseq data, 18 of the 19 chromosomes support the S. diabolicum+S. magni clade as monophyletic (Table S5). Eleven of the 19 chromosomes resolve S. magni as monophyletic with ultrafast bootstrap support ranging from 95 to 100%. Chromosome LG16 suggests that S. diabolicum is non monophyletic because the Fuegan *S. magellanicum* samples are nested within the clade. Moreover, six chromosomes yield a topology wherein S. diabolicum and S. magni are reciprocally monophyletic (Table S5). Data from only two of those chromosomes (LG08, LG11) provide strong support for that inference (ultrafast bootstrap > 95%). However, eight chromosomes yield a topology where either both groups would be monophyletic or a monophyletic S. magni would be nested within (paraphyletic) S. diabolicum, except for the misplacement of one or two samples from Delaware, New Jersey, or the mountains of North Carolina.

Monophyly of the other five species in the complex is supported by all 19 individual chromosomes and it is clear that S. diabolicum and S. magni are phylogenetically closer to one another than are any other pair of species in the group. Discordant patterns across chromosomes, sister group relationships among the major clades (species), and the relationship between S. diabolicum and S. magni could reflect incomplete lineage sorting or introgression, or some combination of these evolutionary processes.

Population genomics and introgression

Admixture analyses of the RADseq data were accomplished using STRUCTURE, and levels of *K* (number of Hardy-Weinberg genetic groups) from 1 to 10 were explored (Fig. S14). *K*=2 was considered optimal but *K*=5 is also informative, so both results are shown in Fig. 6. At *K*=2, *S. divinum* is distinct from all other species in the complex as it is almost fixed for the

orange genotype group; only two Russian samples show evidence of admixture. South American plants (*S. magellanicum* and S. magellanicum-NW) and S. asiaticum are admixed for that group. At *K*=5, the green genotype group is substantially represented in plants from Chile and Argentina and in S. asiaticum (Fig. 6). The two Russian plants of *S. divinum* have that group and it is minimally represented in plants of S. diabolicum and S. magni. *Sphagnum medium* is clearly distinct at *K*=5, comprised of the blue genotype group that does not occur in any other species, and S. magellanicum-NW is distinct with a yellow group that also occurs at low levels of admixture in S. asiaticum. Sphagnum diabolicum and S. magni are only distinguishable at *K*=5 by the amount of the green genotype present; both are mostly comprised of the grey genotype group. In fact, these two species are not further distinguished at higher levels of resolution for genotype groups (i.e., *K*=5-10; Fig. S14). Reducing the number of samples in each group to have more equally-sized groups does not affect these inferences (Fig. S15).

Nucleotide diversity (π) calculated from genome resequencing data indicates that S. diabolicum and S. magni have the highest levels of genetic diversity in the complex. Much lower levels were detected within S. magellanicum-NW. *Sphagnum medium* and *S. divinum*, the most common species, are characterized by an intermediate level of genetic diversity (Table 1).

Genome-wide F_{ST} and d_{XY} values based on resequencing data corroborate the inference from phylogenetic reconstructions that clades within the *S. magellanicum* complex are strongly differentiated, except for S. diabolicum versus S. magni (Table 2). F_{ST} estimates based on RADseq data provide similar inferences; S. diabolicum and S. magni are weakly differentiated relative to other species pairs (Table 3; Fig. S16), and further corroborate a close genetic relationship between these species and South American plants.

Fixed nucleotide differences among clades/species support inferences from phylogenetic reconstructions and statistics quantifying genetic differentiation (Table 4). The highest number of fixed differences (2,515,839) in the resequencing data is between *S. medium* and S. magellanicum-NW, which are entirely allopatric. Fixed differences among the other species range from 584,419 between *S. divinum* and S. diabolicum to 1,685,899 between *S. divinum* and S. magellanicum-NW. Only 37,565 fixed differences were detected between S. diabolicum and S. magni. Similar patterns are evident from the RADseq data (Table 4; Fig. S16), but here the highest numbers of fixed differences occur between *S. medium* and *S. divinum*, two northern species that sometimes occur sympatrically. In agreement with the resequencing data, plants belonging to *S. medium* and S. magellanicum-NW also exhibited a high number of differences, only slightly lower than between *S. medium* and *S. divinum*. Also in

agreement with results from the resequencing data, S. diabolicum and S. magni exhibit the least fixed differences, 210. RADseq data corroborate the relatively high level of genetic similarity of both S. diabolicum and S. magni to plants from South America, and support a closer relationship of those eastern North American species to plants from southernmost South America (*S. magellanicum*) than to those from northern South and Central America (S. magellanicum-NW). Interestingly, the number of fixed differences between plants in the *S. magellanicum* versus S. magellanicum-NW are as high as between most of the northern Hemisphere pairs, and lower than (Fuegan) *S. magellanicum* to either S. diabolicum or S. magni (Table 4).

As genealogical discordance can reflect retention of standing ancestral polymorphism (i.e., deep coalescence) as well as introgression, we conducted tests to distinguish these processes. Using the resequencing data, we detected significant introgression among multiple pairs of species. We found that D_{min} was significantly elevated in nine of the ten species trios and ranged from 0.012 to 0.055 (Table S6). The corresponding f_4-ratios for these conservative estimates of *D* suggest that between 1% and 11% of the genome is introgressed depending on the species involved. As lower limits for the amount of interspecific gene flow in the complex, these results strongly suggest that a strictly bifurcating tree cannot accurately represent the evolutionary history of *S. magellanicum* species. Some of this gene flow likely represents ancient introgression, perhaps between *S. magellanicum* or *S. divinum* and an ancestor of S. magni and S. diabolicum as *f*-branch statistics would suggest (Fig. S17), but this will be further explored in a subsequent publication.

Microhabitats, geographic structure and functional trait variation

Sphagnum medium grows predominantly in open bog microhabitats, but *S. divinum*, which occurs at some of the same sites, is equally distributed across open bogs, bog margins, and surrounding forests (Table 5). In Norway, *S. divinum* is said to occur primarily at bog margins and in forests. Although *S. medium* is more common in the open parts of the bogs, it does occasionally also occur at the margins and in adjacent forests. These two species have ample ecological opportunity to hybridize, consistent with the observation (above) of substantial introgression between them. Sphagnum diabolicum is less common, but like *S. divinum*, it appears to occur across multiple microhabitats at peatland sites. Limited data are available for S. magni and the boreal microhabitat classification does not correspond well to the warm temperate / subtropical sites where it grows. The habitats of S. magni include poorly drained

areas along roadsides, in pine forests, and, unlike the boreal species, co-occurs with palmettos and other subtropical plants in southeastern U.S. coastal plain areas.

A subjective perusal of RADseq-based relationships among samples (Fig. S7) suggests that while some samples from the same or proximate sites often group together, there is also evidence of close relationships among highly disjunct plants. We see strongly supported clades of *S. divinum*, for example, that group plants from New Hampshire, or West Virginia. A strongly supported clade include most of the samples from western Canada and the U.S. On the other hand, two European samples of *S. divinum* from the Czech Republic are closely related to samples from the northeastern U.S. but are distant from a third Czech sample that is closely related to a Swedish sample.

Similarly, there is some grouping of S. magni samples from North Carolina, but otherwise there appears to be little geographic structure among samples of that species from southeastern and Gulf of Mexico coastal plain samples. One sample of *S. medium*, a relatively uncommon species that appears to be otherwise completely restricted to western Europe and northeastern North America, was collected in Central America (Suriname). We checked this unlikely observation by re-extracting and resequencing, but it appears to be accurate, and we conclude that there was a dispersal from some northern latitude source. It is closely related to a sample from Maine.

We quantified geographic structure within species by estimating the correlation between geographic and genetic distances estimated from RADseq data (IBD; Isolation by Distance). IBD was significant only for *S. medium*, the S. diabolicum+S. magni clade, and (less strongly) for S. diabolicum alone (Table S7).

Tissue decomposability is an important functional trait that underlies carbon sequestration through the accumulation of peat. Our field experiment did not reveal differences in decomposability among the four eastern North American species in the complex (Table S8).

Discussion

This research provides a framework to enable ecological genomics for the genus *Sphagnum* as a model. *Sphagnum* peatmosses have unparalleled ecological importance because they create peatland ecosystems that support a host of other organisms, control hydrology over regional scales, release substantial methane into the atmosphere, accumulate vast amounts of carbon in the form of peat, and provide natural laboratories for investigating niche differentiation within communities. Members of the *S. magellanicum* complex are important players because they are widespread and abundant components of peatland

ecosystems. Newly resolved insights into genomic divergence and phylogenetic relationships within the *S. magellanicum* complex suggest that we have caught these plants in the act of speciation, and have a model system to promote greater understanding of how these ecological functions came to be during the course of evolution.

Species in the *S. magellanicum* complex differ in geographic ranges across climate zones. *Sphagnum divinum*, *S. medium*, and S. diabolicum are boreal species that range from Canada southward in the mountains of eastern North America. Sphagnum magni has a warm temperate to subtropical range in the eastern U.S. and reaches as far south as Lake Okeechobee in south-central Florida. Sphagnum magellanicum-NW and S. asiaticum have tropical-montane ranges, and *S. magellanicum* occupies temperate-subantarctic habitats in Tierra del Fuego. These contrasting climate-correlated distributions will be valuable for identifying genomic features that confer tolerances to heat and cold stresses and can inform predictions about the biotic consequences of climate warming.

Within climate zones, species in the complex differ in niches. Within boreal peatlands, *S. medium* typically forms high hummocks out in the open whereas *S. divinum* more commonly occurs at the margins and in surrounding forests (Yousefi *et al*., 2017). We find that in eastern North America, *S. medium* is also almost completely restricted to open bogs, but *S. divinum* occurs in both the surrounding forested areas and out in the open. When in the open, however, *S. divinum* consistently occurs in lower hummocks and in lawns closer to the water table. Sphagnum diabolicum also occurs both in open bogs, close to the water table, and in surrounding forests. We have been unable to detect an ecological difference between *S. divinum* and S. diabolicum and they frequently occur sympatrically in the northeastern U.S. In contrast, the warm temperate-subtropical species, S. magni, occurs in habitats that have no counterparts in the boreal zone. This species occurs in pine, pine-palmetto, and hardwood forests, and along roadsides where impeded drainage results in standing water for parts of the year.

Ecological differences between S. diabolicum and S. magni, very closely related sister taxa that have not reached reciprocal monophylly across their entire genomes, are especially attractive targets for research on ecological adaptation. The strongly supported inference that S. magni is sister to S. diabolicum, and that this inclusive clade is sister to plants from Tierra del Fuego, raises a series of important issues to pursue. This phylogenetic topology strongly suggests that the warm temperate-subtropical ecology of S. magni is derived from a colder climate ancestry. In addition to heat tolerance per se, their habitats likely differ in nutrient availability, light quality and quantity, microbial associates, and annual photoperiod variation,

among other features. We are currently pursuing analyses to characterize genomic differentiation between S. diabolicum and S. magni, with research on their comparative physiology and microbiomes being initiated.

Much is known about ecological processes that underlie community assembly in *Sphagnum*-dominated peatlands, such as competition between hummock-forming and hollow-inhabiting species (summarized in Rydin & Jeglum, 2013), but few studies have examined *Sphagnum* ecology in an evolutionary/phylogenetic framework. Some ecologically important functional traits (e.g., growth, decomposability) are phylogenetically conserved across *Sphagnum* at the genus level and are correlated with interspecific niche differences (e.g., hummock vs. hollow niches; Bengtsson *et al*., 2016, 2018; Piatkowski & Shaw, 2019; Piatkowski *et al*., 2021). Decomposition rate, a critical trait underlying peat accumulation and therefore carbon sequestration, is conserved across the genus and is correlated with the characteristic positions of species along the hummock-hollow gradient (Piatkowski *et al*., 2021). It is noteworthy that some of the clades/species within the S. diabolicum complex differ in niche relative to height above the water table (hollows to hummocks), but we detected no differences in decomposition rates. This includes S. magni, which occupies very different warm temperate to subtropical habitats where peat accumulation is minimal. In addition to rates of decomposition, growth rates are the other obvious variable underlying how high plants are raised above the water table and how much peat they accumulate. We have in-progress work designed to test if species in the *S. magellanicum* complex differ in growth rates and whether any such differences relate to their niches.

Recent discoveries about phylogenetic relationships and species delimitation have increased accuracy of evolutionary inferences in other bryophyte model organisms including *Physcomitrium patens* (McDaniel *et al*., 2009; Medina *et al*., 2019; Rensing *et al*., 2020), *Marchantia polymorpha* (Linde, 2019), and *Ceratodon purpureus* (Nieto-Lugilde *et al*. 2018a, b). Added value to these systems come from new insights into related species and intraspecific variants with more complex morphologies (*Physcomitrium*), different ecological ranges (*Marchantia*), and genome structure (*Ceratodon*).

These studies of bryophytes that are widely utilized for comparative genomics, including those in the genus *Sphagnum*, show that phylogenetic resolution is critical to maximize their value and to suggest avenues for additional research. Moreover, accurate species delimitation that reflects variation in morphology, ecology, and genomic structure, is critical to evolutionary interpretations and is essential for accurate identification of plant material utilized for any research application. New phylogenetic resolution of clades within the *Sphagnum magellanicum*

complex presented here facilitates accurate downstream comparisons among field-collected samples and raises novel questions about the evolution of ecological variation in the group. Our results suggest that speciation within the *S. magellanicum* complex is in-progress, and appears to be at a particularly early stage for the S. diabolicum-S. magni clade. Other species in the monophyletic *S. magellanicum* complex are more divergent than are S. diabolicum and S. magni, but are nevertheless very closely related in the context of the genus *Sphagnum* more broadly. Clear ecological differences between species makes this group a valuable model for evolutionary/ecological genomics in a group with unparalleled importance to global scale biogeochemistry and climate.

Acknowledgements

This research was supported by NSF grants DEB-1737899 and DEB-1928514 (PI A.J. Shaw). The research was also supported by a grant from the Tom and Bruce Shinn Fund from the North Carolina Native Plant Society. The work (proposal: 10.46936/10.25585/60001030) conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Collection of starting *Sphagnum* was made possible through the SPRUCE project, which is supported by Office of Science; Biological and Environmental Research (BER); US Department of Energy (DOE), Grant/Award Number: DE-AC05–00OR22725. Experimental work and analyses were supported by the DOE BER Early Career Research Program. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the US DOE under contract no DE-AC05–00OR22725. Additional support for diversity collections and analysis by NSF DEB-1737899, 1928514.The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank Min Kim of HudsonAlpha for generating chloroplast contigs for phylogenetic analyses.

Author contributions

AJS conducted the field work, participated in data analysis and interpretation, and wrote the paper. BP, BA, CPB, and KAH, and KI conducted field work, and participated in data analysis, and preparation of the paper. AMD, JG, MN-L, and MNP contributed lab work, participated in data analyses, and paper preparation. AH, DJW, and JS participated in data analysis and paper

preparation. KH, KIF, and HKS contributed to conceptual development and paper preparation. JBY participated in the decomposability experiment and paper preparation.

Data availability

Demultiplexed Illumina reads from RADseq samples, *in silico* digested reads from genomic resequencing samples, and the alignment of chloroplast genome sequences are available in Dryad (https://doi.org/10.5061/dryad.1c59zw3xc). The genome assembly and annotation for the *S. divinum* reference (v1.1) is freely available at Phytozome (https://phytozome-next.jgi.doe.gov/). The whole genome shotgun sequencing project for the *S. divinum* reference genome has been deposited at DDBJ/ENA/GenBank under the accession JAKJHR000000000 and the version described in this paper is version JAKJHR010000000. Sequencing libraries are publicly available within the sequence read archive (SRA) under BioProject PRJNA799298.

References

Andrews S. 2010. FastQC: A quality control tool for high throughput sequence data. v.011.9 [WWW document] URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Beike AK, von Stackelberg M, Schallenberg-Rüdinger M, T Hanke ST, Follo M, Quandt D, McDaniel SF, Reski R, Tan BC, Rensing SA. 2014. Molecular evidence for convergent evolution and allopolyploid speciation within the *Physcomitrium- Physcomitrella* species complex *BMC Evolutionary Biology* **14:**158. DOI: 10.1186/1471-2148-14-158

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proceedings of the National Academy of Science* **107:** 18724-18728.

Bengtsson F, Granath G, Rydin H. 2016 Photosynthesis, growth, and decay traits in *Sphagnum*—a multispecies comparison. *Ecology and Evolution* **6:** 3325–3341.

Bengtsson F, Rydin H, Hájek T. 2018 Biochemical determinants of litter quality in 15 species of *Sphagnum*. *Plant Soil* **425:** 161–176.

Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting F_{ST}: the impact of rare variants. *Genome Research* **23**: 1514-1521

Brukhin V, Jaroslaw V. Osadtchiy JV, Florez-Rueda AM, Smetanin D, Bakin E, Nobre MS, Grossniklaus U. 2019. The *Boechera* genus as a resource for apomixis research. *Frontiers in Plant Science* **10:** 1–19.

Canty A, Ripley B. 2021. Boot: bootstrap R (S-Plus) functions. R package version 1.3-28. [WWW document] URL https://cran.r-project.org/web/packages/boot/index.html.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**: s13742-015-0047-8

Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* **30:** 3317–3324

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarhty SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10: 1–4**. https://doi.org/10.1093/gigascience/giab008

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. 2011. The variant call format and VCFtools. *Bioinformatics* **27:** 2156–2158.

Dierckxsens N, Mardulyn P, Smits G, 2017. NOVOPlasty: de novo assembly of organelle genomes from whole genome data, *Nucleic Acids Research* **45:** e18, https://doi.org/10.1093/nar/gkw955

Duffy A, Aguero B, Stenøien H, Flatberg KI, Ignatov MS, Hassel K, Shaw AJ. 2020. Phylogenetic structure in the *Sphagnum recurvum* complex (Bryophyta: Sphagnaceae) relative to taxonomy and geography. *American Journal of Botany* **107:** 1283–1295.

Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4:** 359–361.

Eaton DAR. 2014. PyRAD: Assembly of *de novo* RADseq loci for phylogenetic analyses. *Bioinformatics* **30:** 1844–1849.

Evanno G, Regnaut S, Gaudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14:** 611–2620.

Georganas E, Buluç A, Chapman JA, Hofmeyr SA, Aluru C, Egan R, Oliker L, Rokhsar DS,

Yelick KA. 2015. HipMer: an extreme-scale de novo genome assembler. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* **14**:1–11. https://doi.org/10.1145/2807591.2807664.

Hassel K, Kyrkjeeide MO, Yousefi N, Prestø T, Stenøien H, Shaw AJ, Flatberg KI. 2018. *Sphagnum divinum* (sp. nov.) and *S. medium* Limpr. and their relationship to *S. magellanicum* Brid. *Journal of Bryology* **3**: 197-222.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* **35:** 518–522.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**: 583-589.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**: 254-267.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23:** 1801–1806.

Jombart T, Ahmed I. 2011. Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27:** 3070–3071.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods* **14:** 587–589.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772-780.

Koch MA. 2019. The plant model system *Arabidopsis* set in an evolutionary, systematic, and spatio-temporal context. *Journal of Experimental Botany* **70:** 55–67.

Korunes KL, Samuk K. 2021. Pixy: unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources* **21**: 1359-1368

Lewis PO. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology* **50**: 913-925.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**:1754-1760.

Linde, A-M. 2019. *Rates and patterns of Bryophyte molecular evolution.* Ph.D Dissertation, Uppsala Universitet, Stockholm, Sweden.

Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, Durbin R. 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology & Evolution* **2**: 1940-1955.

Malinsky M, Matschiner M, Svardal H. 2021. Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources* **21:** 584–595.

McDaniel SF, von Stackelberg M, Richardt S, Quatrano RS, Reski R, Rensing SA. 2009. The speciation history of the *Physcomitrium-Physcomitrella* species. *Evolution* **64:** 217–231.

Medina R, Johnson MG, Liu Y, Wickett NJ, Shaw AJ, Goffinet B. 2019. Phylogenomic delineation of *Physcomitrium* (Bryophyta: Funariaceae) based on targeted sequencing of nuclear exons and their flanking regions rejects the retention of *Physcomitrella, Physcomitridium* and *Aphanorrhegma*. *Journal of Systematics and Evolution* **57**: 404-417.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020a. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* **37:** 1530–1534.

Minh BQ, Hahn MW, Lanfear R. 2020b. New Methods to Calculate Concordance Factors for Phylogenomic Datasets. *Molecular Biology and Evolution* **37:** 2727–2733.

Nieto-Lugilde M, Werner O, McDaniel SF, Ros RM. 2018a. Environmental variation obscures species diversity in southern European populations of the moss genus *Ceratodon*. *Taxon* **67**: 673-692.

Nieto-Lugilde M, Werner O, McDaniel SF, Ros RM. 2018b. Environmental variation obscures species diversity in southern European populations of the moss genus *Ceratodon*. *Taxon* **67:** 673–692.

Nikolov LA, Shushkov P, Nevado B, Gan X, Al-Shehbaz IA, Filatov D, Bailey CD, Tsiantis M. 2019. Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytologist* **222:** 1638–1651.

Olson JS. 1963. Energy storage and the balance of producers and decomposers in ecological systems. *Ecology* **44:** 322-331.

Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA. 2012. Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology* **21:** 2991–3005.

Piatkowski BT, Shaw AJ. 2019. Functional trait evolution in *Sphagnum* peat mosses and its relationship to niche construction. *New Phytologist* **223:** 939–949.

Piatkowski BT, Yavitt JB, Turetsky MR, Shaw AJ. 2021. Natural selection on a carbon cycling trait drives ecosystem engineering by *Sphagnum* (peat moss). *Proceedings of the Royal Society B* **288**: 20210609.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155:** 945–959.

R Core Team. 2021. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. v4.1.1. URL https://www.R-project.org/

Rensing SA, Goffinet B, Meyberg R, Wu S-Z, Bezanillac M. 2020. The moss *Physcomitrium* (*Physcomitrella*) *patens*: a model organism for non-seed plants. *The Plant Cell* **32:** 1361–1376.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**: 217-223.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. *Trends in Genetics* **16:** 276-277.

Rydin H, Jeglum J. 2013. *The Biology of Peatlands.* Ed. 2. Oxford University Press, New York, NY, USA.

Shaw AJ, Cox CJ, Boles SB. 2003. Polarity of peatmoss (*Sphagnum*) evolution: who says bryophytes have no roots? *American Journal of Botany* **90:** 1777–1787.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Systematic Biology* **51:** 492-508.

Turetsky MR, Crow SE, Evans RJ, Vitt DH, Wieder RK. 2008. Trade-offs in resource allocation among moss species control decomposition in boreal peatlands. *Journal of Ecology* **96:** 1297-1305.

Van de Auwera GA, O'Connor BD. 2020. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition)*. O'Reilly Media.

Vitt DH, Slack NG. 1984. Niche diversification of Sphagnum relative to environmental factors in northern Minnesota peatlands. *Canadian Journal of Botany* **62:** 1409-1430.

Weston, D.J., Turetsky, M.R., Johnson, M.G., Granath, G., Lindo, Z., Belyea, L.R., Rice, S.K., Hanson, D.T., Engelhardt, K.A.M., Schmutz, J., Dorrepaal, E., Euskirchen, E.S., Stenøien, H.K., Szövényi, P., Jackson, M., Piatkowski, B.T., Muchero, W., Norby, R.J., Kostka, J.E., Glass, J.B., Rydin, H., Limpens, J., Tuittila, E-S., Ullrich, K.K., Carrell, A., Benscoter, B.W., Chen, J-G., Oke, T.A., Nilsson, M.B., Ranjan, P., Jacobson, D., Lilleskov, E.A., Clymo, R.S., and Shaw, A.J. 2018. The Sphagnome Project: enabling ecological and evolutionary insights through a genus-level sequencing project. *New Phytologist* **217**: 16-25.

Yousefi N, Hassel K, Flatberg KI, Kemppainen P, Trucchi E, Shaw AJ , Kyrkjeeide MO, Szövényi, P, Stenøien HK. 2017. Divergent evolution and niche differentiation within the common peatmoss *Sphagnum magellanicum*. *American Journal of Botany* **104:** 1060–1072.

Yu ZC. 2012. Northern peatland carbon stocks and dynamics: a review. *Biogeosciences* **9:** 4071–4085.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**: 153.

806 **Table 1**: Nucleotide diversity (π) for species in the *S. magellanicum* complex estimated from 10
807 kb non-overlapping genomic windows. Estimates represent medians of average values within
808 windows, and the 95% bootstrap confidence intervals are given in parentheses.

	N	π	π (95% CI)
S. diabolicum	8	0.0103	(0.0102, 0.0104)
S. divinum	24	0.0031	(0.0031, 0.0032)
S. magellanicum-NW	2	0.0019	(0.0018, 0.0019)
S. magni	9	0.0093	(0.0093, 0.0094)
S. medium	5	0.0038	(0.0037, 0.0038)

Table 2. Pairwise comparisons of the fixation index (F_{ST}, above diagonal) and genetic divergence (d_{XY}, below diagonal) between species in the *S. magellanicum* complex based on genome resequencing data. These estimates represent the medians of average values within 10 kb non-overlapping genomic windows and the 95% bootstrap confidence intervals are given in parentheses.

	S. diabolicum	***S. divinum***	**S. magellanicum-NW**	**S. magni**	***S. medium***
S. diabolicum		0.4825 (0.4795, 0.4851)	0.5134 (0.5105, 0.5163)	0.0720 (0.0710, 0.0729)	0.5665 (0.5638, 0.5693)
S. divinum	0.0153 (0.0152, 0.0154)		0.7301 (0.7272, 0.7328)	0.5293 (0.5268, 0.5319)	0.7302 (0.7273, 0.7330)
S. magellanicum-NW	0.0157 (0.0156, 0.0158)	0.0154 (0.0153, 0.0155)		0.5459 (0.5432, 0.5487)	0.7479 (0.7452, 0.7510)
S. magni	0.0115 (0.0114, 0.0115)	0.0158 (0.0157, 0.0159)	0.0158 (0.0157, 0.0159)		0.5944 (0.5290, 0.5969)
S. medium	0.0193 (0.0192, 0.0194)	0.0173 (0.0171, 0.0174)	0.0184 (0.0183, 0.0185)	0.0195 (0.0194, 0.0196)	

Table 3. Pairwise comparisons of the fixation index (F_{ST}, above diagonal) between species in the *S. magellanicum* complex based on RADseq data. These estimates represent medians of average values within 10 kb non-overlapping genomic windows and the 95% bootstrap confidence intervals are given in parentheses.

	S. diabolicum	***S. divinum***	***S. magellanicum***	**S. magellanicum-NW**	**S. magni**	***S. medium***
S. asiaticum	0.6544 (0.6491, 0.6603)	0.7573 (0.7511, 0.7635)	0.9429 (0.9357, 0.9473)	0.8540 (0.8468, 0.8582)	0.6767 (0.6701, 0.6825)	0.7979 (0.7918, 0.8052)
S. diabolicum		0.3967 (0.3881, 0.4043)	0.4054 (0.3950, 0.4164)	0.4786 (0.4702, 0.4865)	0.0876 (0.0853, 0.0898)	0.5411 (0.5353, 0.5495)
S. divinum			0.6353 (0.6241, 0.6457)	0.6689 (0.6610, 0.6772)	0.4453 (0.4359, 0.4537)	0.6558 (0.6501, 0.6643)
S. magellanicum				0.8107 (0.8032, 0.8119)	0.5070 (0.4974, 0.5164)	0.7571 (0.7521, 0.7631)
S. magellanicum-NW					0.4519 (0.4411, 0.4647)	0.7717 (0.7645, 0.7788)
S. magni						0.5654 (0.5577, 0.5725)

820 **Table 4**. Fixed differences between species using data from genome resequencing (above diagonal) and RADseq (below diagonal).
821 Comparisons that could not be made due to lack of data are represented by “na”.
822

	S. asiaticum	**S. diabolicum**	***S. divinum***	***S. magellanicum***	**S. magellanicum-NW**	**S. magni**	***S. medium***
S. asiaticum		na	na	na	na	na	na
S. diabolicum	4,239		584,419	na	1,140,617	37,565	1,406,748
S. divinum	5,295	4,519		na	1,685,899	682,824	1,682,598
S. magellanicum	5,179	1,919	4,783		na	na	na
S. magellanicum-NW	6,075	5,053	8,704	6,664		1,236,519	2,515,839
S. magni	4,396	210	5,361	2,608	5,561		1,509,463
S. medium	6,067	9,732	13,021	7,612	12,129	10,363	

824 **Table 5.** Microhabitat distributions of species in the *S. magellanicum* complex.

825

Species	Open Bog	Bog Margin	Forest
S. diabolicum	2	8	6
S. divinum	14	14	17
S. magni	0	2	2
S. medium	13	2	1

826

827 **Figure Legends**

828

829 **Figure 1**. Map of the collection locations of samples included in the RADseq and genomic
830 resequencing analyses. Colors represent the clades referred to in this study: blue = *S. divinum*,
831 brown = *S. medium*, green = S. diabolicum, violet = S. magni, cyan = *S. magellanicum* and S.
832 magellanicum-NW, red = S. asiaticum.

833

834 **Figure 2**. Phylogenetic relationships among samples based maximum likelihood analyses of
835 resequencing data from the nuclear (left) and plastid (right) genomes suggest cytonuclear
836 discordance. The nuclear tree presented here was reconstructed using the dataset containing
837 loci genotyped in at least 80% of samples. Central lines connect sample position in each
838 phylogeny. Colors represent the clades referred to in this study: blue = *S. divinum*, brown = *S.*
839 *medium*, green = S. diabolicum, violet = S. magni, cyan = S. magellanicum-NW. Scale bars =
840 substitutions/site

841

842 **Figure 3**. Summary of the tree topologies identified from analyses of individual chromosomes
843 (e.g., LG01) and across the genome using nuclear genome resequencing and RADseq data.
844 Data were analyzed using the maximum likelihood (ML) and coalescent (SVD) methods.
845 Asterisks indicate that at least one internal branch is not statistically supported. Parentheses
846 indicate tree topologies that approximately unbiased tests fail to reject despite their suboptimal
847 likelihood. na = not available.

848

849 **Figure 4**. Phylogenetic network based on nuclear genome resequencing loci depicting weighted
850 splits among samples in the *S. magellanicum* complex. Longer edges are splits found more
851 frequently in the dataset and cycles/boxes represent incompatible splits.

852

853 **Figure 5**. Summary of phylogenetic relationships among *Sphagnum magellanicum* complex
854 clades based on RADseq loci. Relationships on the left were estimated using maximum
855 likelihood and branches are labelled with ultrafast bootstrap values. Relationships on the right
856 were estimated using singular value decomposition scores for species quartets and nodes are
857 labelled with bootstrap values.

858

859 **Figure 6**. Results of STRUCTURE analyses of RADseq loci for *Sphagnum magellanicum*
860 complex clades at *K*=2 and *K*=5. Colors represent different genotype groups.

861
862 **Supplemental Online Material**
863
864 **Supplemental Tables:**
865
866 **Table S1**. Voucher table with collection information for samples included in RADseq and
867 genome resequencing analyses.
868
869 **Table S2**. Number of sites used for phylogenetic analyses using genome resequencing data.
870
871 **Table S3**. Summary statistics for the two types of samples used in the RADseq analyses.
872 Illumina RADseq samples were produced by RAD sequencing of plant collections. *In silico*
873 digested genomes were produced from genome resequencing data and "digested" as described
874 in Methods. Datasets including loci with three different minimum sample coverage levels were
875 used to verify that inferences are not affected by the trade-off between number of loci and
876 proportion of missing data.
877
878 **Table S4.** Support values per chromosome for clades containing S. magni and/or S. diabolicum
879 based on maximum likelihood analyses of nuclear resequencing data. Clade support values
880 include ultrafast bootstrap values (left) and site concordance factors (right).
881
882 **Table S5.** Support values per chromosome for clades containing S. magni and/or S. diabolicum
883 based on maximum likelihood analyses of RADseq data. Support values with asterisks indicate
884 clades that do not represent a monophyletic S. magni or S. diabolicum group due to the listed
885 exceptions. Clades labelled as paraphyletic with asterisks indicate clades that would be
886 paraphyletic if not for the indicated exceptions.
887
888 **Table S6.** D_{min} for species trios in the *S. magellanicum* complex estimated from genome
889 resequencing data. For each trio, this statistic represents the minimum possible value for *D*
890 across all possible topologies. Significance was assessed using the block jackknife with 1E
891 blocks and the resulting *P*-values were adjusted using the Benjamini-Hochberg procedure.
892
893 **Table S7**. Isolation by distance (genetic distance versus the base 10 log of geographic
894 distance) for pairwise comparisons between samples of A) *S. divinum*, B) *S. medium*, C) S.

895 diabolicum + S. magni, D) S. diabolicum, E) S. magni, F) S. magellanicum-NW, G) *S.*
896 *magellanicum*, and H) S. asiaticum.
897
898 **S8**. Tissue decomposability (*K,* yr-1) of samples representing North American clades in the *S.*
899 *magellanicum* complex. Analysis of variance failed to recover an effect of lineage on
900 decomposability (F(3,52)=0.274, *P*=0.844).
901
902 **Supplemental Figures:**
903
904 **Fig. S1.** Phylogenetic relationships among samples in the *S. magellanicum* complex estimated
905 using IQ-TREE2 for nuclear genome resequencing data.
906
907 **Fig. S2.** Phylogenetic relationships among clades in the *S. magellanicum* complex estimated
908 using SVDquartets for nuclear genome resequencing data.
909
910 **Fig. S3.** Phylogenetic relationships among clades in the *S. magellanicum* complex estimated
911 using ASTRAL for nuclear genome resequencing data.
912
913 **Fig. S4.** Phylogenetic relationships among samples in the *S. magellanicum* complex estimated
914 using IQ-TREE2 for plastid genome data.
915
916 **Fig. S5.** Phylogenetic relationships among *Sphagnum magellanicum* complex clades or groups
917 based on RADseq loci present in at least 70%, 80%, or 90% of samples.
918
919 **Fig. S6.** Map of RADseq locus positions on chromosomes of the *Sphagnum divinum* genome.
920
921 **Fig. S7.** Phylogenetic relationships among samples based on RADseq loci present in at least
922 70%, 80%, or 90% of samples.
923
924 **Fig. S8.** Phylogenetic relationships among samples in the *S. magellanicum* complex estimated
925 using IQ-TREE2 for nuclear genome resequencing data from individual chromosomes.
926 Branches are labelled with ultrafast bootstrap values (left) and site concordance factors (right).
927 Scale bar units are the number of substitutions per site.
928

Fig. S9. Phylogenetic relationships among clades in the *S. magellanicum* complex estimated using SVDquartets for nuclear genome resequencing data from individual chromosomes.

Fig. S10. Phylogenetic relationships among clades in the *S. magellanicum* complex estimated using ASTRAL for nuclear genome resequencing data from individual chromosomes.

Fig. S11. Phylogenetic relationships among samples based on RADseq loci mapping to each individual chromosome of the *Sphagnum divinum* genome.

Fig. S12. Phylogenetic relationships among samples based on RADseq loci mapping to each individual chromosome of the *Sphagnum divinum* genome.

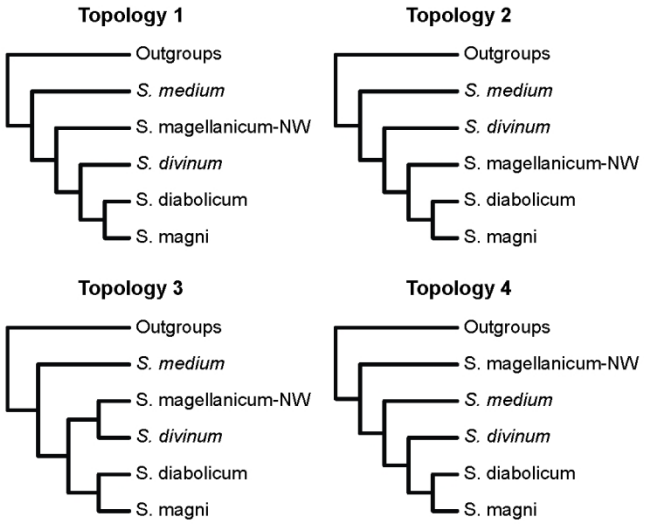
Fig. S13. Summary of the nine different phylogenetic relationships identified based on maximum likelihood and singular value decomposition scores for species quartets using RADseq loci mapping to each chromosome of the *Sphagnum divinum* genome.

Fig. S14. Results of STRUCTURE analyses of RADseq loci for clades in the *Sphagnum magellanicum* complex at all values of *K* from 2 through 10.

Fig. S15. Results of STRUCTURE analyses of RADseq loci with reduced sampling to minimize size differences between clades in the *Sphagnum magellanicum* complex at all values of K from 2 through 10.

Fig. S16. Scatterplot of F_{ST} values (above) and fixed differences (below) for clades in the *S. magellanicum* complex estimated from RADseq and genome resequencing data.

Fig. S17. Plot of *f*-branch statistics calculated using the genome resequencing dataset that includes loci present in at least 80% of samples.



Dataset	Reseq ML.	Reseq. SVD	Reseq. ASTRAL	RAD ML	RAD SVD
LG01	1	2	1	2	2*
LG02	2* (1)	2	2	2	2
LG03	2* (1)	3	2*	1*	1
LG04	2 (1)	2	2	2	2
LG05	1	2	1	1	1
LG06	1	2	1	1*	1*
LG07	2 (1)	2	2	2	2
LG08	2 (1)	2	2	2	2
LG09	1	2	1	1	1*
LG10	1 (2)	3	1	1	1
LG11	1	3	1*	2*	2
LG12	1	3	1	1	1
LG13	2* (1)	2	2	2*	2
LG14	2	2	2	2	2
LG15	1* (2)	2	2	2*	2
LG16	1 (2)	2	1	2	2
LG17	1 (2)	3	1	1	2
LG18	2 (1)	2	2	2	1
LG19	1* (2)	2	2*	1	1*
70% Dataset	2*	2	2	2	3
80% Dataset	1	2	2	2*	3
90% Dataset	1 (2)	3	2	2*	2
100% Dataset	4 (1)	3	1	na	na

Fig. 3

140x251mm (200 x 200 DPI)