doi:10.1111/evo.12547

Evolution of niche preference in *Sphagnum* peat mosses

Matthew G. Johnson,[1,2,3] Gustaf Granath,[4,5,6] Teemu Tahvanainen,[7] Remy Pouliot,[8] Hans K. Stenøien,[9] Line Rochefort,[8] Håkan Rydin,[4] and A. Jonathan Shaw[1]

[1]*Department of Biology, Duke University, Durham, North Carolina 27708*

[2]*Current Address: Chicago Botanic Garden, 1000 Lake Cook Road Glencoe, Illinois 60022*

[3]*E-mail: mjohnson@chicagobotanic.org*

[4]*Department of Plant Ecology and Evolution, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36, Uppsala, Sweden*

[5]*School of Geography and Earth Sciences, McMaster University, Hamilton, Ontario, Canada*

[6]*Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, SE-750 07, Uppsala, Sweden*

[7]*Department of Biology, University of Eastern Finland, P.O. Box 111, 80101, Joensuu, Finland*

[8]*Department of Plant Sciences and Northern Research Center (CEN), Laval University Quebec, Canada*

[9]*Department of Natural History, Norwegian University of Science and Technology University Museum, Trondheim, Norway*

Received March 26, 2014
Accepted September 23, 2014

Peat mosses (*Sphagnum*) are ecosystem engineers—species in boreal peatlands simultaneously create and inhabit narrow habitat preferences along two microhabitat gradients: an ionic gradient and a hydrological hummock–hollow gradient. In this article, we demonstrate the connections between microhabitat preference and phylogeny in *Sphagnum*. Using a dataset of 39 species of *Sphagnum*, with an 18-locus DNA alignment and an ecological dataset encompassing three large published studies, we tested for phylogenetic signal and within-genus changes in evolutionary rate of eight niche descriptors and two multivariate niche gradients. We find little to no evidence for phylogenetic signal in most component descriptors of the ionic gradient, but interspecific variation along the hummock–hollow gradient shows considerable phylogenetic signal. We find support for a change in the rate of niche evolution within the genus—the hummock-forming subgenus *Acutifolia* has evolved along the multivariate hummock–hollow gradient faster than the hollow-inhabiting subgenus *Cuspidata*. Because peat mosses themselves create some of the ecological gradients constituting their own habitats, the classic microtopography of *Sphagnum*-dominated peatlands is maintained by evolutionary constraints and the biological properties of related *Sphagnum* species. The patterns of phylogenetic signal observed here will instruct future study on the role of functional traits in peatland growth and reconstruction.

KEY WORDS: **Bryophyte, comparative methods, peatland ecology, phylogenetic signal.**

Boreal peatlands are not just dominated by *Sphagnum* peat mosses—they are engineered by them (van Breemen 1995). Habitat variation within a peatland ecosystem can be substantial, and is generally characterized along two gradients (Rydin and Jeglum 2013)—an electrochemical gradient (defined by pH and other cations) and a hydrological gradient (variation in the availability of ground water due to microtopography). Some *Sphagnum* species both create and inhabit the raised microtopographic features (hummocks) because of their growth forms (Laing et al. 2014), water transport abilities (Granath et al. 2010), and low decay rates (Belyea 1996). The plants produce large amounts of organic acids, contributing to a lower pH, and yet maintain an effective uptake of solutes through cation exchange in extremely nutrient poor environments (Hemond 1980). By creating an environment that is wet, acidic, and anoxic (Clymo 1963), *Sphagnum* decomposes slowly and thereby triggers peat accumulation.

EVOLUTION
INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

Within these gradients, *Sphagnum* species are known to differentiate into narrow microhabitat preferences: in one survey in New York State, *Sphagnum contortum* was found only in areas with pH above 6.0, whereas *S. majus* was found only below pH 5.0 (Andrus 1986). Similar differentiation has been observed in other peatlands along the hummock–hollow and electrochemical gradients (Vitt and Slack 1984; Gignac 1992; Rochefort et al. 2012; Rydin and Jeglum 2013). Experimental transplants have revealed that while hummock-preferring species can survive more aquatic environments, a hollow-preferring species cannot survive the more stressful hummock environment (Rydin et al. 2006). Within hummock environments, some hummock species depend on the presence of other specific species for optimal establishment and growth (Chirino et al. 2006). The development and maintenance of boreal peatland ecosystems thus depends on the facilitation and competition of many species within the same genus.

What makes the microhabitat differentiation in *Sphagnum* more remarkable is the relatively young age of most *Sphagnum* species. The class Sphagnopsida is one of the earliest diverging groups of mosses, splitting from the rest of Bryophyta about 380 million years ago (mya; Newton et al. 2009). However, nearly all extant *Sphagnum* species originate from a radiation just about 14 mya (Shaw et al. 2010b), coinciding with the end of the mid-Miocene climatic optimum and the appearance of peatland ecosystems in the northern boreal zone. Of the 250–300 extant species of *Sphagnum* resulting from this radiation, approximately 40 of these species have circumboreal distributions and can be commonly found in peatlands throughout the high latitudes of the Northern Hemisphere. In a relatively small amount of geologic time, these 40 species have shaped peatland ecosystems through their extended phenotypes and microhabitat preferences.

Given the recent radiation of species, their narrow observed preferences and perhaps narrow physiological tolerances, it is reasonable to expect that microhabitat preferences in *Sphagnum* exhibit "phylogenetic signal"—closely related species are expected to be more similar than randomly selected species on a phylogeny (Blomberg and Garland 2002). However, despite many years of observing ecology of *Sphagnum* (reviewed in Clymo and Hayward 1982; Rydin and Jeglum 2013), the presence of phylogenetic signal has not been tested.

When considering the evolution of ecological niche descriptors, it is useful to distinguish between β-niche—climatic tolerances or macrohabitat affinity—and α-niche, within-community microhabitat affinity (Ackerly et al. 2006). Many studies model ecological niches using climatic BIOCLIM data from public databases, for example (Boucher et al. 2012), and focus on β-niches because data on α-niches are unavailable or impractical to collect. In cases where the α-niche is considered, phylogenetic signal can suggest whether habitat preferences underlie community assembly (Cavender Bares et al. 2004) and whether phylogenetic signal has been overwhelmed and erased for evolutionarily labile traits (Eterovick et al. 2010). Labile traits such as behavior (Blomberg et al. 2003) and ecological niche (Losos 2008) may not show phylogenetic signal. For ecological traits, phylogenetic signal must be demonstrated before inferences about, for example, community assembly or niche conservatism can be made.

Subgeneric classification in *Sphagnum* already gives some clue about phylogenetic signal of microhabitats in the genus. Two monophyletic subgenera, *Cuspidata* and *Subsecunda*, are generally characterized by species living at or near the water table (hollow), whereas members of subgenera *Acutifolia* and *Sphagnum* (also monophyletic) are more likely to form hummocks high above the water table. It was recently shown that although *Sphagnum* has a large cation exchange capacity, it does not exceed the capacity of other peatland mosses (such as brown mosses, Soudzilovskaia et al. 2010). This suggests that peatland acidification along the fen–bog gradient is due to peat accumulation, not to the actions of live *Sphagnum* plants. Therefore, phylogenetic signal may be more easily detected in hummock/hollow microhabitat descriptors, compared to the pH/ionic gradient.

The evolution of continuous traits on a phylogeny is commonly modeled using Brownian motion (BM), which predicts that trait variance increases along the phylogeny from root to tip (Felsenstein 1985). The BM pattern, however, may be masked by several factors, each of which is addressed by additional models. If the rate of trait evolution is not constant along the phylogeny, or the trait has accumulated more variance than is predicted by BM, the model may be a poor fit for the phylogeny and trait. Pagel (1999) developed models to detect phylogenetic signal under these conditions: a *lambda* model allowing for greater trait variance, and a *delta* model predicting that trait variance has accumulated faster at the root of the phylogeny compared to the tips.

The presence of one or more optimal trait values for *Sphagnum* species would constrain the trait evolution to values close to these optima. For instance, there may be an "ideal" pH preference for *Sphagnum* species, and therefore evolution of this niche descriptor would be constrained among *Sphagnum* species due to forces such as stabilizing selection (sensu Hansen 1997). Finally, if the rate of evolution in microhabitat preference is unconstrained or extremely fast, then phylogenetic signal for that trait may become undetectable.

Demonstration of phylogenetic signal for microhabitat preference in *Sphagnum* would further suggest that the underlying functional traits (such as growth rate, decomposition rate, water retention, or cation exchange ability, see, e.g., Rice et al. 2008 and Turetsky et al. 2008) would also show similar patterns. Presence of phylogenetic signal would provide information on how

contrasting peatland habitats (fens and bogs) and microhabitats (hummock and hollows) have developed over evolutionary time. This would guide the focus of future studies on functional traits and the Neogene development of peatland ecosystems.

In this study, we test whether *Sphagnum* microhabitat descriptors show phylogenetic signal using a variety of comparative models to test the tempo, direction, and heterogeneity of microhabitat niche evolution in the genus. To do this, we use ecological niche data for 39 *Sphagnum* species from three large published studies in northern Europe and North America, construct a phylogenetic tree using sequences from 18 genes, and analyze the comparative dataset containing eight univariate niche descriptors and two principal components representing the environmental gradients. Using methods designed to account for phylogenetic uncertainty and within-species measurement error, we test whether any of the niche descriptors (1) has phylogenetic signal; (2) whether this signal corresponds to or deviates from BM; and (3) if changes in evolutionary rates can be detected within *Sphagnum*.

Materials and Methods

NICHE DIFFERENTIATION

Peatland ecologists have noted the specificity of *Sphagnum* species along electrochemical and hydrological gradients for more than 40 years (Clymo 1973; Vitt and Slack 1984; Andrus 1986), and ideally, we would have used the microhabitat data from all available studies. However, we chose to focus on three recent major surveys of *Sphagnum* microhabitat specificity to ensure consistent measurements, the largest selection of species, and the most modern *Sphagnum* taxonomy. The three selected large surveys each recorded data from eight niche descriptors: Height above water table (HWT), percent vascular plant cover (as an indicator of shade), pH, electrical conductivity (EC), and several ionic concentrations (Ca, K, Mg, and Na). Each study represents a number of sites, and within each site, data were recorded for a number of plots along transects. Plot sizes varied among studies, with 25 × 25 cm square plots in Estonia, 50 × 50 cm square plots in Finland, and 70 cm diameter circular plots in Canada. In each plot, the eight niche descriptors were recorded, as well as the presence and relative abundance of each species in the plot. Each plot, therefore, may represent a datapoint for one or more species.

The first survey covered 498 sites in eastern (2647 datapoints) and western (944 datapoints) Canada (Gignac et al. 2004). The second study also included two areas of Canada: 23 sites in Quebec and New Brunswick (1369 datapoints) and one area in Estonia (Europe) where 11 sites were surveyed (389 datapoints, Pouliot 2011). The third study included 36 sites (714 datapoints across 29 mire complexes) in eastern Finland located in the mid-boreal zone (Tahvanainen 2004). Two of the mire complexes were sampled intensively in a separate substudy of 270 plots (258 datapoints; Tahvanainen et al. 2002). Taken together, these data represent 6533 observations of *Sphagnum* microhabitat associations, by far the most comprehensive dataset of its kind.

Fusion of the three major studies allows us to be confident that if a species was not observed in any plots, it is not a major contributor to boreal peatland diversity in Canada or northern Europe. A total of 40 species were recorded, but we excluded *S. auriculatum* because of low sample size ($N = 1$), yielding 39 species in the final dataset. Data were summarized across the three studies by weighting the means and SDs of each species by percent cover of the sampled plots, that is, giving more weight to plots where the species covered a larger area. Because most species occur in all regions covered by the three studies, we estimated the overall mean and variation in niche descriptors, across all sites and plots. This estimate will therefore not account for different ecotypes or large-scale (continental) differences in environmental conditions, but is instead a generalized estimate of the realized niche for each *Sphagnum* species.

The niche descriptor (ecological trait) for each species was transformed so that its mean was zero and its SD across the genus was 1. In addition to univariate descriptors, we also investigated the evolution of microhabitat niche in a multivariate sense, using a principal components transformation on all eight niche descriptors. The principal component analysis (PCA) scores from the first two ordinates (Fig. 1) were included in the analyses below. We also repeated the analyses using nonmetric multidimensional scaling (NMDS), but representing multivariate niche by this alternative ordination did not alter our major conclusions (results not shown).

DNA EXTRACTION, AMPLIFICATION, AND SEQUENCING

For each of the 39 *Sphagnum* species, we sampled representative DNA sequences from GenBank and from a database maintained by AJS; most sequences have been submitted previously, previously unpublished samples are identified as such in Table S1. We also selected one sample each of *Flatbergium serecium* and *Eosphagnum inretortum*, representing early diverging members of class Sphagnopsida, to serve as outgroups (Shaw et al. 2010a). Previous studies (Shaw et al. 2003b, 2010a) used 24 species to demonstrate that *Sphagnum* has four major monophyletic subgenera: *Sphagnum*, *Subsecunda*, *Cuspidata*, and *Acutifolia*. Our sampling of 39 species covers all four subgenera (Fig. 1) with more species in the latter two subgenera.

We followed protocols described in (Shaw et al. 2003b) to sample sequences from the following genes: photosystem II (PSII) reaction center protein D1 (*psbA*), PSII reaction center protein T (*psbT-H*), ribulose-bisphosphate carboxylase gene

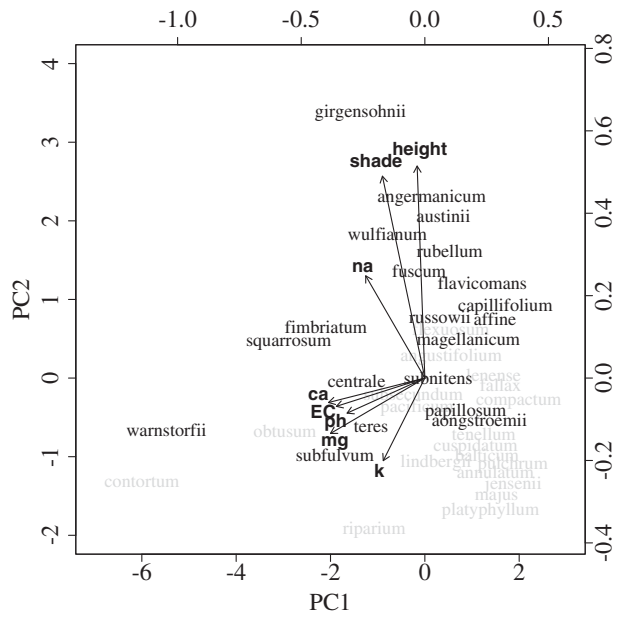

Figure 1. Biplot of principal components analysis (PCA) for eight microhabitat preferences in 39 species of *Sphagnum*. Each species is plotted in Euclidian space for the first two principal components, which cumulatively represent 61.4% of the total variance. Loadings upon each axis are indicated by arrows and lines—PC1 (43.9% of total variance) is a "pH–ionic gradient," whereas PC2 (17.5% of total variance) is predominantly a "hummock–hollow" gradient. Species in black are in subgenera characterized by hummock habitats (*Sphagnum* and *Acutifolia*), whereas species in gray are in subgenera characterized by hollow habitats (*Subsecunda* and *Cuspidata*). Left and bottom axes represent PC scores, right and top axes represent niche trait loadings upon the principal components.

(*rbcL*), plastid ribosomal gene (*rpl16*), RNA polymerase subunit beta (*rpoC1*), ribosomal small protein 4 (*rps4*), tRNA(Gly) (UCC) (*trnG*), and the *trnL* (UAA) 59 exon-*trnF* (GAA) region (*trnL*) from the plastid genome; introns within NADH protein-coding subunits 5 and 7 (*nad5*, *nad7*, respectively) from the mitochondrial genome; the nuclear ribosomal internal transcribed spacer (ITS) region, two introns in the nuclear LEAFY/FLO gene (*LL* and *LS*), three anonymous nuclear regions (*rapdA*, *rapdB*, *rapdF*), and two sequenced nuclear microsatellite loci (*ATGc89* and *A15*) from the nuclear genome. Primer sequences for amplifying and sequencing for most loci were provided in Shaw et al. (2003b). For *rpoC1*, we used primers described in the Royal Botanic Gardens, Kew, web page: DNA Barcoding, phase 2 protocols (http://www.kew.org/barcoding/protocols.html). For the two microsatellite-containing loci, we used primer sequences: A15—F: 5′TGTGGAGACCCAAGTGAATG3′ R: 5′GGTGATGCTCAAAGGGCTTA3′; ATGc89—F: 5′CGTCGAACGGATTCAAAAAT3′ R: 5′AGGGGAAGAGACCATCAGGT3′. We used the Duke University Sequencing facility for Sanger sequencing of all samples. For GenBank accession numbers, see Table S1.

PHYLOGENETIC RECONSTRUCTION

Although phylogenetic relationships within the genus are not the primary focus of this study, it is worth noting that our taxon sampling (39 species) and genomic sampling (seven nuclear, eight chloroplast, and two mitochondrial genes) are the largest species-level phylogenetic analysis of *Sphagnum* to date.

Individual genes were aligned using MUSCLE (Edgar 2004) and adjusted manually using PhyDE (Muller et al. 2010). When concatenated, the dataset contained 14,918 characters, of which 636 were parsimony informative (Table S1). To obtain ultrametric trees required for phylogenetic comparative methods, we reconstructed the *Sphagnum* phylogeny via Bayesian inference on a concatenated 18-gene alignment, using BEAST (Drummond et al. 2012). For each gene, we chose a substitution model using the Bayesian information criterion from jModelTest (Guindon and Gascuel 2003; Posada 2008; Table S1). Branch lengths were inferred using uncorrelated relaxed clock model and a lognormal branch length prior, one model for each gene separately. We confirmed convergence to the same joint posterior distribution by replicating the BEAST analysis ($N = 2$), and visualizing the likelihood and parameter estimates from the two runs using Tracer version 1.75 (Rambaut and Drummond 2014). In each analysis, the chain ran for 200 million generations, sampling every 10,000 steps following a 20 million generation burnin. We summarized the 18,000 trees from the posterior distribution using a maximum credibility tree calculated by TreeAnnotator (Drummond et al. 2012), with node heights set to the median branch lengths. To marginalize phylogenetic uncertainty (topology and branch lengths) in the comparative methods, we randomly selected 1000 trees from the posterior density for most analyses.

EVOLUTION OF NICHES: MODEL CHOICE

Testing models of comparative evolution has recently become much easier because all of the models can be implemented and connected using the phylogenetic package ape (Paradis et al. 2004) in the statistical programming environment R (R Core Development Team, www.R-project.org). On each ecological niche descriptor, we evaluated the fit of three main models of evolution (Table 1). (1) White noise (*WN*)—the trait values are independent of phylogenetic distance; this represents our baseline model. Under this model, all internodes on the phylogeny are set to zero length, creating a star phylogeny—all trait evolution occurs at the tips, and phylogeny and trait variance are therefore completely unrelated. By using WN as a baseline, we assert that alternative models (below) must demonstrate better fit to the data than a model where the phylogeny does not contribute to trait evolution. Any model with a sample-size corrected Akaike information

Table 1. Detailed information about the eight models of trait evolution tested.

Model	Abbreviation	Description	Parameters	Equivalent to
White noise	WN	Trait values independent of phylogenetic distance	Covariance	
Brownian motion	BM	Trait variance increases with phylogenetic distance	β—Rate of evolution	WN if $\beta = \infty$
Brownie 2-rate	BM2	Separate rates of evolution in *Acutifolia* versus *Cuspidata*	β—Rate of evolution (one for each group)	
Ornstein–Uhlenbeck	OU	Random walk with central tendency (stabilizing selection)	β—Rate of evolution; α—strength of selection; θ—trait optimum	BM if $\alpha = 0$; WN If $\alpha = \infty$
Ornstein–Uhlenbeck	OU2	OU model with different optima for *Acutifolia* versus *Cuspidata*	β—rate of evolution; α—strength of selection; θ—trait optimum (one for each group)	
Lambda	lambda	Internal branch lengths multiplied; deviation from pure BM	β—rate of evolution, λ—multiplier	BM if $\lambda = 1$; WN if $\lambda = 0$
Delta	delta	Internal branch lengths raised to a power; if $\delta > 1$: evolution concentrated in tree tips	β—rate of evolution; δ—multiplier	BM if $\delta = 1$

For each model, the parameters estimated by maximum likelihood and the nesting of each model are also indicated.

criterion (AICc) score exceeding the score for WN is not a plausible alternative.

(2) BM—the trait increases in variance through evolutionary time at a constant rate (beta). Although this is the standard phylogenetic comparative model, signal may be masked by several other patterns of trait evolution, which are addressed with the remaining models. (3) Ornstein–Uhlenbeck (OU) model (Martins and Hansen 1997)—although the evolution of the trait contains phylogenetic signal, evolution is constrained by a strength parameter (alpha), causing the trait to trend toward an optimum value (theta). Two of the other models are nested within the OU model: BM (alpha = 0) and WN (alpha = infinite).

If either of the alternative models (OU or BM) is accepted, we further evaluate the fit of these models through two evolutionary parameters: The *Lambda* parameter (Pagel 1994)—the trait has phylogenetic signal, but deviates from a pure BM process. Specifically, the phylogenetic covariance is multiplied by a scalar, which is inferred via maximum likelihood. The WN model (lambda = 0) and BM model (lambda = 1) are nested within the lambda model, in which lambda is inferred as a free parameter. Values between 0 and 1 correspond to an "imperfect" BM model, where only some proportion (lambda) of the trait variance can be explained by phylogeny. The *Delta* parameter (Pagel 1997)—all node depths are raised to the power delta—values less than 1 provide evidence that much/most trait evolution occurred deep (early) in the phylogeny, whereas values greater than 1 indicate trait evolution concentrated in the tips. The BM (delta = 1) and WN (delta = infinite) models are nested within the WN model. For both the lambda and delta models, we can infer whether it is a better fit than the BM model (via a likelihood ratio test) and whether the maximum likelihood values inferred on 1000 trees significantly deviates from WN (lambda = 0) or BM (lambda and delta = 1) using one-tailed tests.

We fit the WN, BM, and lambda models using the R package phytools (Revell 2011), the delta model with geiger (Harmon et al. 2008), and the OU model was fitted using the pmc_fit method of the package pmc (Boettiger et al. 2012).

Many sources of error exist in the estimation of mean trait values for species, and phylogenetic comparative methods are improved when they account for measurement error (Ives et al. 2007). For each niche descriptor, we used methods in phytools for the BM, lambda, and delta models to incorporate measurement

error (SE, incorporating both SD and sample size); incorporation of measurement error is not implemented in pmc_fit, so it is absent from the OU models.

RATE CHANGES WITHIN *SPHAGNUM*

All of the methods above assume constant conditions on the entire *Sphagnum* phylogeny. To incorporate the possibility of different rates of niche evolution within the tree, which would mask the pattern when considering the entire genus, we used two different methods. In our first approach, we pruned the phylogeny to contain only members of subgenera *Acutifolia* and *Cuspidata*, which represented the two largest subgenera sampled. Every branch on the phylogeny was classified as an *Acutifolia* or a *Cuspidata* lineage. We tested whether a model allowing different rates of niche evolution in the two lineages (*BM2*) was supported over a single-rate model (*BM1*, the "Brownie" model; O'Meara et al. 2006), using the brownie.lite method in phytools. We also tested whether an OU model with different trait optima for the *Acutifolia* and *Cuspidata* lineages (*OU2*) was supported over a single-optimum *OU1* model, using pmc.

To visualize phylogenetic signal and rate change within the reduced dataset, we created traitgrams for the two principal components. A traitgram is constructed by reconstructing the ancestral traits for every node on the chronogram. The *x*-axis in a traitgram corresponds to time, whereas the *y*-axis corresponds to the reconstructed trait values. Our trait reconstructions and traitgrams were plotted using the "phenogram" function in the phytools package.

We also used a Bayesian MCMC approach on the full *Sphagnum* phylogeny to identify nodes where rate changes have occurred (Revell et al. 2012). This method samples evolutionary rates and locations of exceptional rate shifts in proportion to their posterior probability, which does not require any a priori hypothesis about the location of rate shifts. We ran the MCMC implemented in phytools under the default priors for evolutionary rate and proposal frequency, for 10,000 generations, sampling every 100 generations (the first 20 samples were discarded as burnin).

TAXON SAMPLING AND PHYLOGENETIC UNCERTAINTY

We conducted a sensitivity analysis to test whether individual species affected the fit of evolutionary models—for example, due to the wide variance in sampling frequency among *Sphagnum* species. Using the maximum credibility tree from BEAST, we analyzed each model for each descriptor 39 times, deleting one *Sphagnum* species each time. We compared the support for each model on the reduced trees to the WN model to assess the sensitivity for each descriptor.

Most phylogenetic comparative methods also unrealistically assume that the tree (topology and branch lengths) is known without error. To incorporate phylogenetic uncertainty into the model fitting procedure, we tested each model on 1000 trees randomly sampled from the BEAST posterior distribution. We recorded, for each descriptor and tree, the AICc scores for each model. The distribution of the AICc scores for each model and descriptor is an indication of model fit, averaged over phylogenetic uncertainty (Boucher et al. 2012). For the Bayesian MCMC approach, we used only the maximum credibility tree from BEAST.

For descriptors found to have significant phylogenetic signal, we used a phylogenetic generalized least squares (PGLS) model (Freckleton et al. 2002) to evaluate their correlated evolution, using the R package caper (Orme et al. 2011). For this analysis, the residuals were modeled using the best-supported model in the full analysis.

Results

NICHE DIFFERENTIATION

There is much variation in within-species sample sizes in the ecological dataset, from three (*S. wulfianum*) to 1055 (*S. fuscum*), reflecting the relative abundances of species in the study sites (Table S2). Among-species SD was lowest for pH and highest for shade. Microhabitats are grouped into two principal components (Fig. 1): PC1 representing an ionic gradient (excluding Na), and PC2 representing the "hummock–hollow gradient" (sodium along with HWT and percent shade). The first two PC axes accounted for 47.3% and 17.7% of the total variance, respectively. Variations along the sodium gradient may reflect the proximity to the sea, which was not tracked in the present study.

The covariation of shade and HWT mainly reflects the abundance of dwarf shrubs on hummocks and the relative scarcity of vascular plants in hollow habitats. The differentiation among subgenera confirms the picture that *Acutifolia* are largely hummock species (higher on PC2), and *Cuspidata* largely hollow inhabitants (lower on PC2), but there are some species deviating from this general pattern (Fig. 1). For example, *S. subfulvum* (subgenus *Acutifolia*) has a low PC2 score, whereas *S. flexuosum* (subgenus *Cuspidata*) is high on that scale. Notably, the subgenus *Sphagnum* is quite variable in HWT. On the ionic gradient, there is less agreement with subgeneric classification.

PHYLOGENETIC RECONSTRUCTION

Each gene in the DNA sequence matrix had varying amounts of missing data, ranging from two sequences missing (ITS) to 29 (nad5 and nad7), whereas sampling for each species ranged from two genes to the full 18 (Table S1). The maximum credibility tree from the Bayesian inference, using BEAST, is presented in Figure S1. The amount of missing data in the alignment does not appear to deflate support for the maximum credibility tree. All major subgenera are resolved at 99% posterior probability or

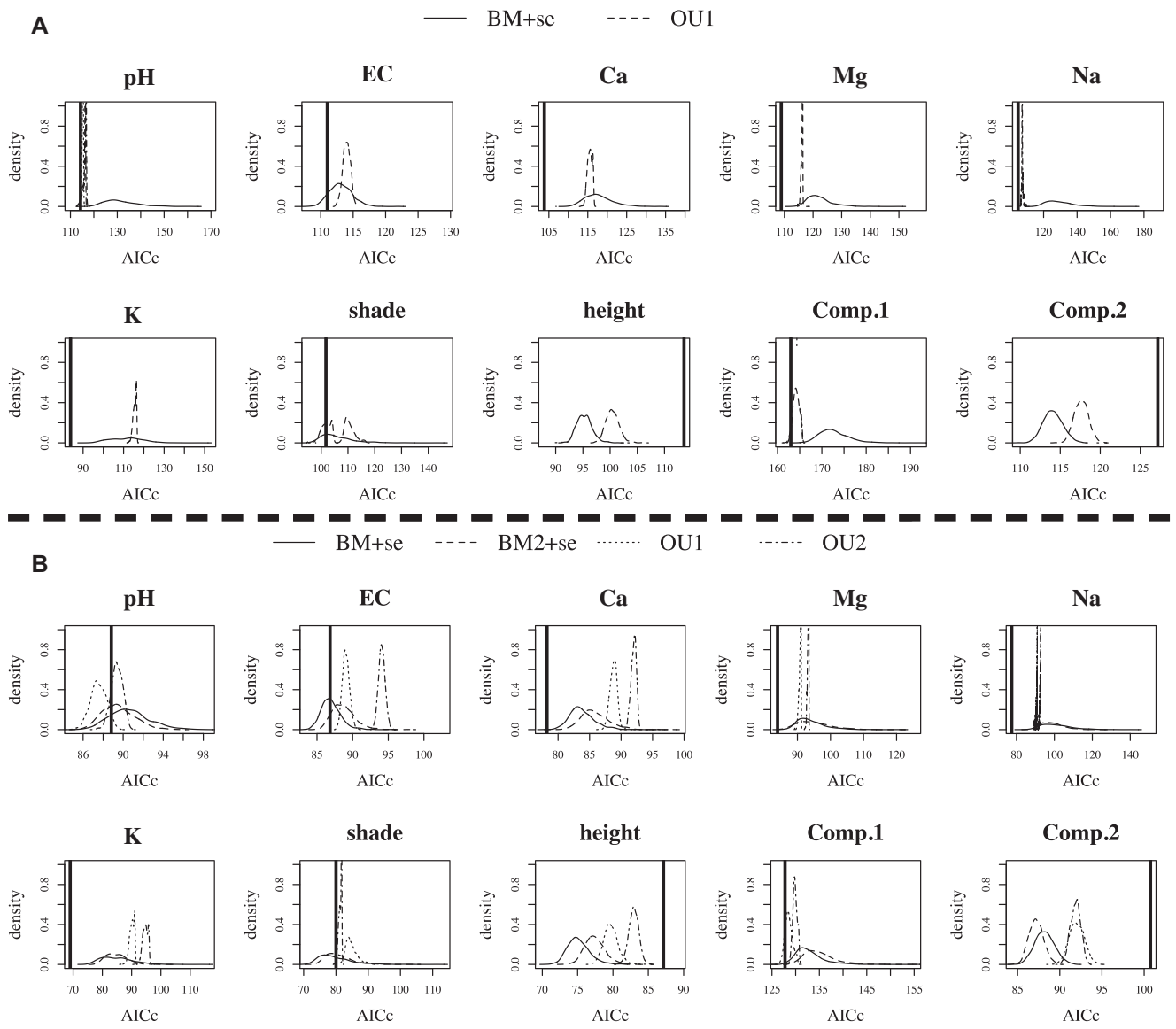

Figure 2. **Model choice using AICc distributions for alternative models of continuous trait evolution, on six niche descriptors—pH, electrical conductivity (EC), concentrations of potassium (K), sodium (Na), magnesium (Mg), and calcium (Ca), percent shade cover, and height above water table (HWT) and the first two principal components across 1000 trees. (A) The full dataset (all of *Sphagnum*). For each niche descriptor, the distribution of AICc scores is shown for Brownian motion (BM) and Ornstein–Uhlenbeck stabilizing selection model (OU). (B) The reduced dataset (subgenera *Acutifolia* and *Cuspidata* only), used to detect changes in niche preference evolution within the genus. For each niche descriptor, the AICc curves for BM1 (one rate) versus BM2 (separate rates) and OU1 (one optimum) versus OU2 (separate optima) are plotted. In each panel, the thick black line indicates the AICc score for white noise (no phylogenetic signal). Lower AICc scores are better; models with AICc distributions falling mostly or entirely to the left of the WN line are preferred.**

greater, while relationships among subgenera are less supported. This is consistent with previous reconstructions of *Sphagnum* phylogeny when both chloroplast and nuclear genomes are used (Shaw et al. 2010a). Notably, among-subgenera median branch lengths are very short; therefore, comparative methods that consider only phylogenetic distance (and not topology) should be relatively unaffected by topological uncertainty.

FULL DATASET: MODEL CHOICE

For five of the six ionic niche dimensions (pH, Ca, Mg, Na, and K), the model that best fits the data across all trees was *WN*, based on the AICc criterion, indicating a lack of phylogenetic signal for these niche descriptors (Table S3). These niche descriptors contribute primarily to the pH–ionic first principal component (except Na, Fig. 1), the evolution of which also is best fit by the white noise

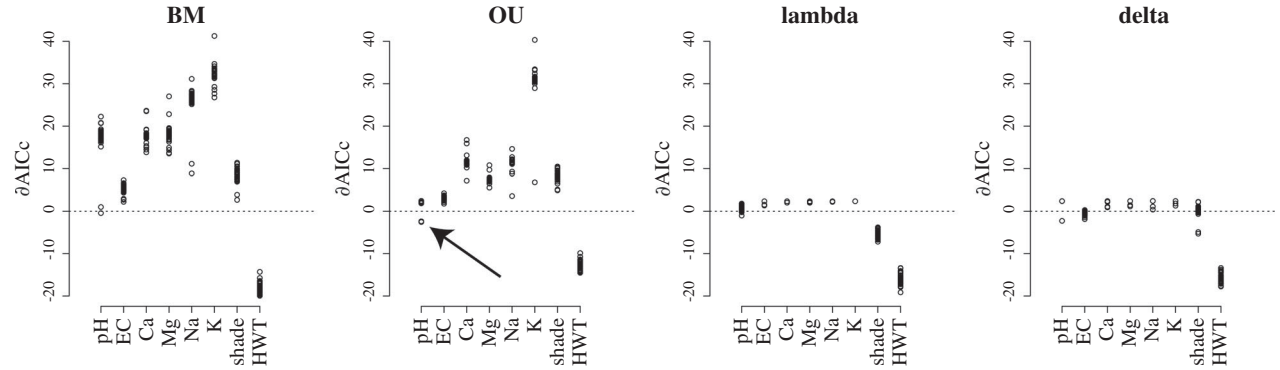

Figure 3. **Sensitivity analysis for model selection in the full *Sphagnum* dataset. Each panel shows one of the four models of evolution at each of the eight niche descriptors (see Table 1 for a key to models). Each point represents the support for the model when individual species were removed from the maximum credibility tree. The *y*-axis is the ∂AICc score compared to WN (no phylogenetic signal). If the points for a model cross the line, it means that deletion of specific species from the analysis changes the interpretation of that model. For example, the arrow indicates two points, representing *S. magellanicum* and *S. centrale*. When either of these species is deleted, AICc supports the OU model (single optimum preference in *Sphagnum*) for pH. In all other combinations, the OU model is rejected for pH (points above the line).**

model (Table S3 and Fig. 2A). There was little variability in the fit of the lambda model across all trees for a few niche descriptors, such as Ca and K (Fig. 2A). In these cases, the values of lambda inferred are very close to zero, providing additional evidence for lack of phylogenetic signal in these descriptors. On 79.7% of the trees, AICc supports *delta* model over *WN* for EC. Values of delta ranging from 2.33 to 18.51 suggest microhabitat evolution is extremely concentrated at the tips—as the value of delta increases to infinity, the delta model collapses to the *WN* model.

For pH, inferred lambdas range from 0.17 to 0.40, but *lambda* never exceeded *WN* in AICc on any of the 1000 trees. Additionally, a likelihood ratio test between *lambda* and *WN* on each tree fails to achieve significance at the $P < 0.05$ level on any tree (results not shown).

In contrast, models of phylogenetic signal are unambiguously a better fit than white noise for two traits—percent cover (shade) and HWT (Fig. 2A and Table S3). The *lambda* model best fits the data for shade, with values of lambda ranging from 0.50 to 0.71. Besides *lambda*, none of the other models were a better fit than *WN* for shade. Among the univariate traits, HWT shows the highest support for phylogenetic signal. The best model was *BM* with a single rate across *Sphagnum*, although all models tested have better AICc scores than *WN*. The distributions of AICc scores for shade, HWT, and PC2 (the hummock–hollow gradient) all indicate phylogenetic signal is strongly supported on all 1000 trees (Fig. 2A).

Sensitivity analyses indicate that the data are generally robust to influence from individual species. In nearly all cases, the AICc score difference between a model and *WN* changes very little, and we almost never observe a model losing support after deletion of individual species (Fig. 3). There are two exceptions: deletion of either *S. magellanicum* or *S. centrale* results in support for the *OU* model for pH, each of which showed a ∂AICc > 7, compared to *WN* (Fig. 2A).

Without phylogenetic correction, the species means for shade and HWT are significantly positively correlated ($t = 2.55$, $r = 0.36$, $P = 0.015$). Using the maximum credibility tree, a test for correlated evolution using lambda as a free parameter was not significant ($t = 1.92$, $r = 0.07$, $P = 0.062$). Because the correlation weakens when accounting for phylogeny, the small but significant correlation observed between shade and HWT may be derived from phylogenetic signal.

RATE CHANGE WITHIN *SPHAGNUM*

The reduced dataset used to investigate rate changes contains only species from subgenera *Acutifolia* (17 species) and *Cuspidata* (13 species). These subgenera contain the largest species sampling, represent one largely hummock (*Acutifolia*) and one largely hollow (*Cuspidata*) clade, and do not share a recent common ancestor within the genus (Fig. S1). For the eight niche descriptors and PC1, neither the *OU2* model nor the *BM2* models were supported (long-dashed line in Fig. 2B). On PC2, however, 91% of the trees supported the *BM2* model over the *BM1* model in the reduced dataset with an average ∂AICc of 1.01 (both models were always better than *WN*, Fig. 2B). The *BM2* model for PC2 inferred a mean evolutionary rate of 500 (range 220–1200) for subgenus *Acutifolia* and a mean evolutionary rate of 190 (range 81–500) for subgenus *Cuspidata*. A paired Student's *t*-test of AICc scores for BM1 versus BM2 on all 1000 trees indicates high support for separate rates of PC2 evolution between the subgenera (mean rate difference: 320, $P < 0.0001$). Traitgrams,

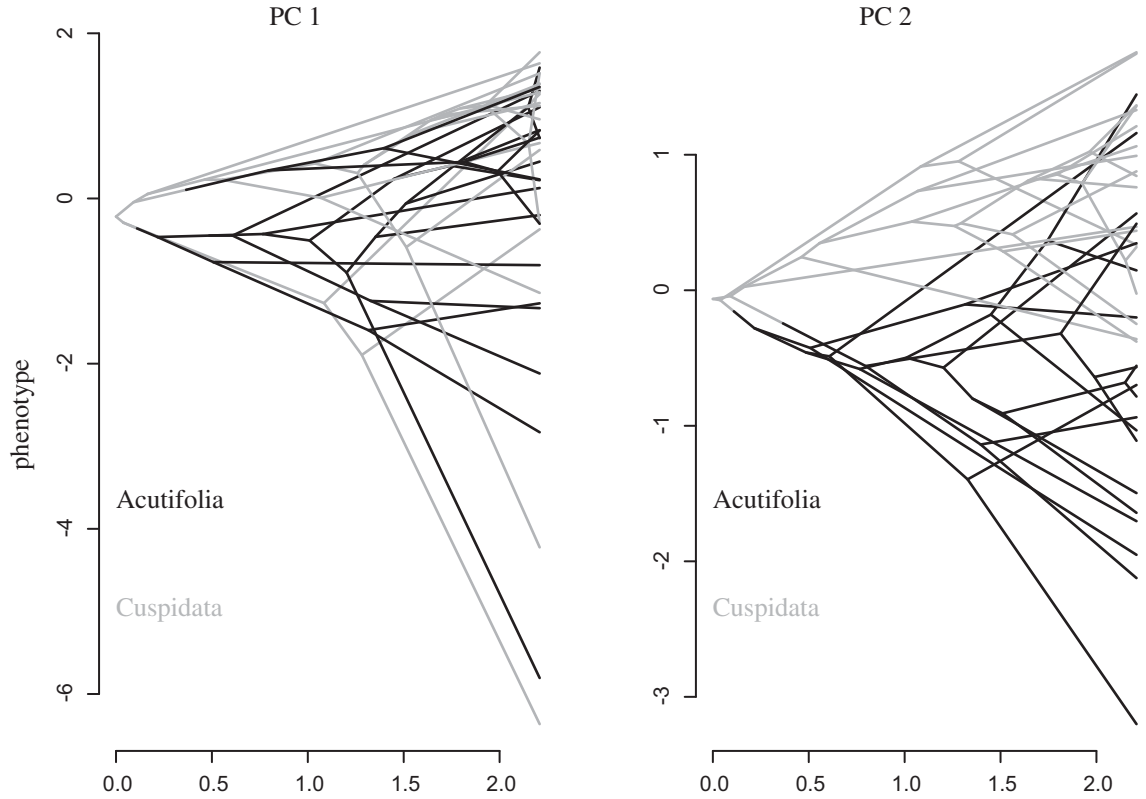

Figure 4. Traitgrams illustrating the phylogenetic signal and rate change within the genus, for two principal components. Using the reduced dataset, ancestral states for the first two principal components were estimated using the maximum credibility chronogram from BEAST. For each tree, the position on the *x*-axis represents time, whereas the position on the *y*-axis represents reconstructed trait values. Dark branches correspond to subgenus *Acutifolia*, whereas lighter branches are subgenus *Cuspidata*. The left panel shows the fast evolution of microhabitat preference in the electrochemical gradient (PC1); the right panel illustrates phylogenetic signal in the hummock–hollow gradient (PC2) along with a difference in evolutionary rate between the two subgenera.

reconstructed for the two principal components (Fig. 4), illustrate the evidence for phylogenetic signal and rate change in PC2 (right) but not PC1 (left).

Although there was no support for an *OU2* model for pH, the *OU1* model was supported in the reduced dataset—953 of the 1000 trees had better AICc scores for the *OU1* model than for *WN* (Table S3). The *OU* model was not supported in the full dataset; but as noted, the *OU* model was supported when either *S. magellanicum* or *S. centrale* were deleted in the sensitivity analysis (arrow in Fig. 3). Both species are in subgenus *Sphagnum*, and were therefore not included in analysis of the reduced dataset.

The Bayesian MCMC approach to identifying exceptional evolutionary rate changes within a phylogeny produces posterior probabilities for each node on the tree for each niche descriptor and microhabitat gradient. Only four descriptors had nodes with a mean posterior probability exceeding 10%. For pH, a rate change was supported within subgenus *Sphagnum*, either on the branch leading to *S. centrale* and *S. magellanicum* (29%) or on an immediately ancestral branch including *S. papillosum* (43%; Fig. 5A). Further evidence of an increase in evolutionary rate comes from the difference in pH preference between the closely related *S. centrale* (mean pH 5.75) and *S. magellanicum* (mean pH 4.14). This is a very large difference compared to other pairs of closely related species in the phylogeny (Fig. 5B).

Although the primary motivation for using the Revell method was to investigate the support for *OU* in pH preference in the sensitivity analysis, rate changes were moderately supported in a few other cases: on the terminal branches leading to *S. contortum*for EC (33%; Fig. S2) and the clade containing *S. fallax* and *S. pacificum* for Na (46%; Fig. S2). Finally, there is support for a rate change in K, either on a terminal branch leading to *S. riparium* (37%) or on the immediately ancestral branch that includes *S. lindbergii* (41%; Fig. S2). No rate change was found for PC2, either in the full dataset or in the reduced dataset (Fig. S2).

Discussion

Individual *Sphagnum* species inhabit narrowly defined microhabitat niches that are an extended phenotype of physical and chemical properties of the genus (Clymo and Hayward 1982).

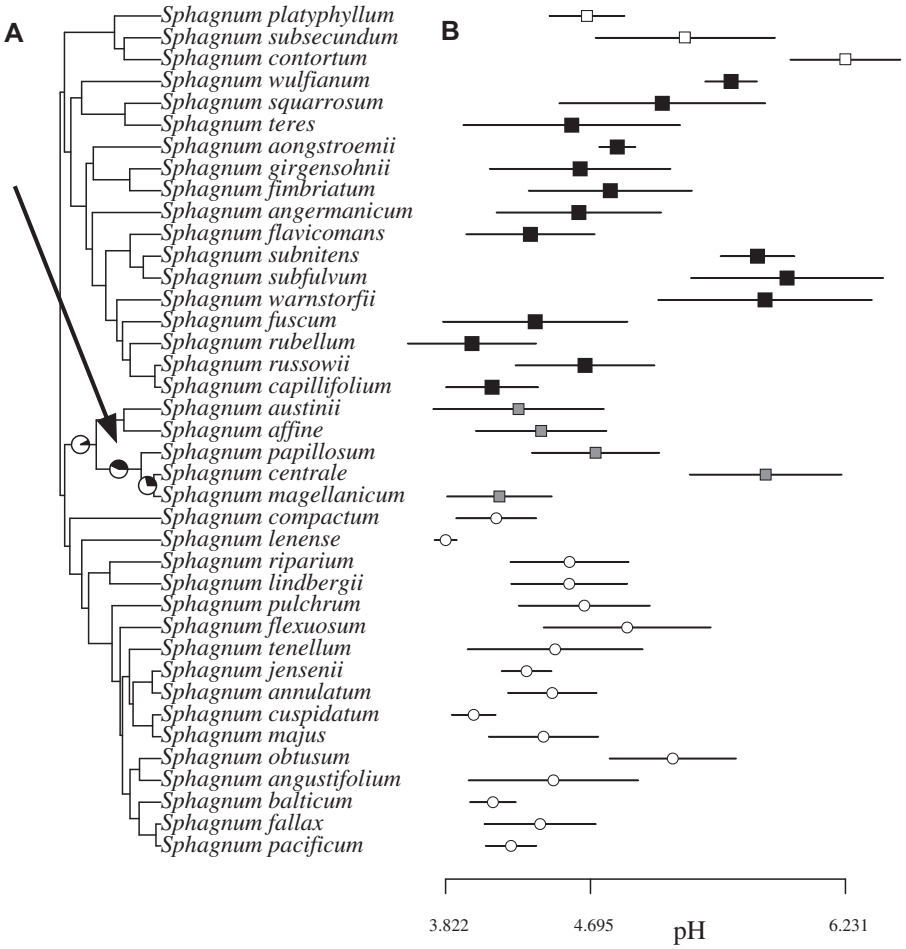

Figure 5. Evidence for exceptional rate change in evolution of pH preference in *Sphagnum*. (A) Evidence of extreme pH preference shift via Bayesian MCMC (Revell et al. 2012)—pie charts indicate nodes receiving at least 10% posterior probability for a rate change. Black portions of each pie chart represent the support for a rate change at that node. The arrow indicates a 77% posterior probability for a rate change in subgenus *Sphagnum*. (B) Phylogenetic diversity of pH preference breadth in *Sphagnum*, by mean and SDs. The symbols represent species in subgenus *Subsecunda* (open squares), *Acutifolia* (black squares), *Sphagnum* (gray squares), or *Cuspidata* (open circles). Additional figures for the other niche descriptors can be found in the Supporting Information.

Therefore, demonstration of phylogenetic signal in microhabitat preference (strongest for HWT) in *Sphagnum* suggests that constrained evolution of microhabitat preferences shapes peatlands with assemblages of related species within similar microhabitats. By contrast, the abiotic electrochemical gradient (pH and ions) may not be constrained, and thus preferences evolve too quickly for phylogenetic signal to be detected. Our tests for phylogenetic signal in *Sphagnum* also show the importance of incorporating several models of trait evolution, as signal may be masked by changes in the rate of trait evolution.

HWT, SHADE, AND MULTIVARIATE NICHE GRADIENTS

Our results clearly show the presence of phylogenetic signal in relation to the hummock/hollow gradient. Species in the major subgenera of *Sphagnum* are generally differentiated along this gradient. We find evidence for rate change in a multivariate niche gradient (encompassing shade and HWT) that suggests a higher rate of niche evolution in subgenus *Acutifolia*, which contains mostly hummock species, than in subgenus *Cuspidata*, which contains mostly hollow species. The strength of the phylogenetic signal indicates that across trees in the dataset, microhabitat preference for height is maintained within, as well as among subgenera. There is also phylogenetic signal in the shade cover of *Sphagnum* species (lambda model, Fig. 2A), and the shade and HWT values are correlated. However, when phylogenetic relatedness is removed with the PGLS model, the strength and significance of the correlation is highly reduced. The bulk of the relationship between HWT and shade is phylogenetically related, reflecting an ecological correlation between HWT and shading—ligneous vascular plants are dependent on oxygen for root

respiration and mycorrhiza, and grow almost exclusively in hummocks, where they provide shade.

We find additional support for a change in the rate of evolution of the multivariate niche gradient encompassing shade and HWT (PC2 axis, Fig. 1). Subgenus *Acutifolia* appears to be evolving faster along the shade–HWT gradient than is subgenus *Cuspidata*. This is apparent in the reconstructed traitgrams (Fig. 4), which show subgenus *Acutifolia* (black) spreading through the trait space much more rapidly than subgenus *Cuspidata* (gray). However, we did not find evidence for separate "optimum" values (*OU2*) in the two subgenera (Fig. 1). Instead, it appears that HWT preference may be more evolutionarily constrained in *Cuspidata*. The range of heights corresponding to "hollow" habitats (0–10 cm) is narrower than the range corresponding to "hummock" habitats (10–30 cm and above). Further, there is growing evidence for a physiological trade-off between hummock and hollow species in growth strategies. Hollow species tend to concentrate growth in the capitulum, maximizing photosynthesis while remaining sparsely packed at the water table (Rice et al. 2008). Conversely, plants with small capitula grow higher above the water table and yet maintain water availability by growing in densely packed hummocks, and thus avoid water stress. The driver behind this trade-off is related to the water flux (capillary rise, water retention) and the need to minimize surface roughness with increasing HWT to decrease water loss (Price and Whittington 2010).

Our results suggest that the classic microtopography of *Sphagnum*-dominated peatlands is caused by an extended phenotype of related species. Shoots of hollow species have high growth rate but decompose faster than hummock species (Turetsky et al. 2008). Because microhabitat preference on the hummock–hollow gradient contains phylogenetic signal, studies of *Sphagnum* functional traits related to this gradient (e.g., leaf and stem morphology, carbon allocation, decomposition rate) should also account for phylogenetic signal. It is likely that the trade-offs mentioned here largely contributed to the observed phylogenetic signal and possibly there is an evolutionary driver behind the microtopographic patterns in peatlands. Consequently, studies of community assembly in *Sphagnum*-dominated peatlands, and studies of functional traits may need to account for the phylogenetic relatedness of peat moss species, as similar habitats along the hummock–hollow will tend to be inhabited by related species.

IONIC GRADIENTS

In contrast, we find that evidence for phylogenetic signal in "ionic" preferences is mostly absent (all cations) or is concentrated in the tips of the phylogeny (EC). Despite the small niche breadth observed in many studies of *Sphagnum*, and that these microhabitat preferences make up much of the major axis of among-species niche variation, the lack of signal is consistent with the observation that the four species with highest PC1 scores ("ionic" niche descriptors excluding Na) represent different subgenera (Fig. 1).

A notable exception is pH, for which a complex pattern possibly including stabilizing selection and a rate change is suggested. Several pieces of evidence, when taken together, suggest that the evolution of the pH niche does in fact contain phylogenetic signal in *Sphagnum*. Although the full dataset failed to support any evolutionary model better than *WN*, the sensitivity analysis (Fig. 3) shows that deletion of either *S. magellanicum* or *S. centrale* provides support for an *OU* model in microhabitat pH evolution. When these species and other members of subgenus *Sphagnum* (and subgenus *Subsecunda*) are removed in the reduced dataset, there is strong support for an *OU* model with a single optimum for the whole genus (Fig. 2B). Moreover, the Bayesian analysis of exceptional rate changes (Revell method) showed strong support for a change in pH niche evolution within subgenus *Sphagnum* (Fig. 5A). These data therefore indicate that pH niche evolution in *Sphagnum* has two phases: (1) An *OU* model, where pH niche evolution deviates from a pure BM process by trending toward a genus-wide optimum of 5.5. Typically, support for an *OU* model is interpreted as evidence of stabilizing selection (Hansen 1997), but can also be interpreted as a bounded BM process. (2) An exceptional rate change occurred within subgenus *Sphagnum*, which masks the signal of the *OU* model when considering the entire genus.

Additional descriptors show evidence of exceptional rate change using the Bayesian MCMC method (Revell et al. 2012), and many of the branches identified are located near the tips of the tree (e.g., *S. contortum* for EC). If the purported rate changes were masking phylogenetic signal in these descriptors, as we suggest for pH, the sensitivity analysis should show model support when these tips are removed. However, none of the other sensitivity analyses indicate support for any model for any of the descriptors where rate changes are proposed by the Bayesian MCMC method. This suggests it is less likely for a rate change to obscure phylogenetic signal in these descriptors, compared to pH. The lack of support for an exceptional rate change in the evolution of the preference along the shade–HWT gradient seems to conflict with our other results, which show evidence for separate rates of PC2 evolution between subgenus *Acutifolia* and subgenus *Cuspidata*. However, the Bayesian MCMC approach was taken with the full dataset, where the rate change signal may be masked by the presence of the other two subgenera.

Several studies besides ours have found very limited intraspecific variation of ionic niche occupancy in *Sphagnum* (Vitt and Slack 1984; Andrus 1986; Gignac 1992). It therefore seems unlikely that the lack of phylogenetic signal is explained by new species preferring ionic microhabitats at random. Rather, microhabitat preference is more evolutionarily labile for these traits, and perhaps phenotypic plasticity or among-species interactions are

more important than phylogeny for the ionic microhabitat preferences (Eterovick et al. 2010). Several bog species have been shown to tolerate more minerotrophic waters from rich fens (Granath et al. 2010), suggesting that these species may have broader tolerances on the ionic gradient than suggested by their observed occurrences. Both of these factors could increase the rate of ionic habitat preference evolution beyond the ability of the comparative methods to detect phylogenetic signal. This would explain why models where trait evolution is concentrated on terminal branches (*delta* model with high value of delta) or completely eliminated in internal branches (*WN* model) are more highly supported for ionic preferences.

It is worth noting here that *Sphagnum*, as a bryophyte, has a haploid dominant life stage. Although allopolyploidy is common in *Sphagnum* (Karlin et al. 2010; Ricca and Shaw 2010), peatlands are primarily engineered by haploid plants. Any mutations that allow for broader physiological tolerances would be immediately exposed to natural selection. This may account for some of the increased rate of microhabitat preference evolution along the electrochemical gradient.

SPECIES INTERACTIONS AND UNCERTAINTIES

Because *Sphagnum* itself is largely responsible for its external microhabitat, and the fact that many *Sphagnum* species establish in patches of other *Sphagnum* species, additional studies are required to investigate the importance of interspecific interactions in definition of narrow microhabitat niches within peatlands. Observations and experiments involving damaged peatlands show that hummocks form several years after reestablishment of *Sphagnum* in a peatland (Pouliot et al. 2012), and that vigorous growth of some species (*S. magellanicum*) depends on the presence of other species (such as *S. fuscum*; Chirino et al. 2006). Therefore, it is clear that interspecies interactions play some role in the formation and maintenance of species diversity in peatlands. A more detailed study could test the role of species interactions serving as a filter in *Sphagnum* community assembly at the hummock/hollow, mineralogical, and peatland scales, by sampling the species diversity at hierarchal scales within one or more peatlands.

In general, our findings are robust to uncertainty introduced by within-species measurement error and phylogenetic uncertainty. Accounting for the former improved the model fits for a few niche descriptors, but did not alter any conclusions. This is not to suggest that within-species variability is unimportant. In their current forms, the methods employed here assume that error estimation of a species mean decreases with sample size, and does not explicitly model the niche breadth of each species. Topological phylogenetic uncertainty was low in our case, but the observations of overlapping AICc distributions, for example, in PC2 in the reduced dataset, indicates the necessity of including phylogenetic error in comparative methods to account for branch length uncertainty.

Conclusions

We have demonstrated the presence of phylogenetic signal in *Sphagnum* for microhabitat preference along the hummock–hollow gradient. Preference for narrow ranges on the ionic gradient appears to be uncorrelated with phylogeny, and further study may confirm whether phenotypic plasticity or infraspecific competition plays roles in eliminating phylogenetic signal. One exception is pH, for which we demonstrate a constraint on pH preference around a genus-wide optimum, although this signal is masked by an exceptional rate change in subgenus *Sphagnum.* The evolution of preferences on the hummock–hollow gradient, however, has a large component explained by phylogeny. The rate of evolution is heterogeneous; lineages classified as preferring hollow environments have lower rates of evolution and are constrained to prefer different multivariate microhabitat optima than hummock lineages.

Because our data represent the realized niches, we are in fact interpreting the combined evolution of physiological tolerances and biotic interactions. Niche preferences demonstrating phylogenetic signal may be more likely to have underlying functional traits related to *Sphagnum* peatland engineering, and may be more likely to be involved in peatland community assembly. The obvious next stage would be to gather data on the basic physiological and morphological traits behind the niches to trace their evolution. The importance of this study and its implications for functional trait evolution in *Sphagnum* are amplified by the recent acceptance of a proposal (A. J. Shaw and D. J. Weston, Principal Investigators) to the Joint Genome Institute (U.S. Department of Energy) to sequence a *Sphagnum* genome, with complementary analyses of gene expression using transcriptomics. This is in recognition of the global importance of *Sphagnum* for carbon sequestration, opening the possibility to link niche and functional trait evolution with global biogeochemistry and climate change.

ACKNOWLEDGMENTS

We thank D. Vitt, N. Slack, M. Poulin, and D. Gignac for providing their raw data, J. Meireles, B. Shaw, and L. Pokorny for comments on earlier drafts, and the r-sig-phylo discussion group for technical support. We also thank two anonymous reviewers for their insightful comments. The sequencing for this study was funded in part by National Science Foundation (NSF) grant DEB-0918998 to AJS and B. Shaw.

DATA ACCESSIBILITY

All DNA sequences have been deposited in GenBank; see Table S1 for accession information. Summarized ecological data, DNA alignments, and phylogenetic trees can be found on Dryad and R scripts used to analyze the data can be found at github.com/mehmattski.

DATA ARCHIVING

The doi for our data is: 10.5061/dryad.0p36h.

LITERATURE CITED

Ackerly, D. D., D. W. Schwilk, and C. O. Webb. 2006. Niche evolution and adaptive radiation: testing the order of trait divergence. Ecology 87:50–61.

Andrus, R. 1986. Some aspects of *Sphagnum* ecology. Can. J. Bot. 64:416–426.

Belyea, L. R. 1996. Separating the effects of litter quality and microenvironment on decomposition rates in a patterned peatland Oikos. 77:529–539.

Blomberg, S. P., and T. Garland. 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. J. Evol. Biol. 15:899–910.

Blomberg, S. P., T. Garland, and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745.

Boettiger, C., G. Coop, and P. Ralph. 2012. Is your phylogeny informative? Measuring the power of comparative methods. Evolution 66:2240–2251.

Boucher, F. C., W. Thuiller, C. Roquet, R. Douzet, S. Aubert, N. Alvarez, and S. Lavergne. 2012. Reconstructing the origins of high-alpine niches and cushion life form in the genus *Androsace S.L.* (Primulaceae). Evolution 66:1255–1268.

Cavender Bares, J., D. D. Ackerly, D. A. Baum, and F. A. Bazzaz. 2004. Phylogenetic overdispersion in Floridian oak communities. Am. Nat. 163:823–843.

Chirino, C., S. Campeau, and L. Rochefort. 2006. *Sphagnum* establishment on bare peat: the importance of climatic variability and *Sphagnum* species richness. Appl. Veg. Sci. 9:285–294.

Clymo, R. S. 1963. Ion exchange in *Sphagnum* and its relation to bog ecology. Ann. Bot. 27:309–324.

———. 1973. The growth of *Sphagnum*: some effects of environment. J. Ecol. 849–869.

Clymo, R. S., and P. M. Hayward. 1982. The ecology of *Sphagnum*. Pp. 229–289 *in* A. J. E. Smith, ed. Bryophyte ecology. Chapman and Hall, Lond.

Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Eterovick, P. C., C. R. Rievers, K. Kopp, M. Wachlevski, B. P. Franco, C. J. Dias, I. M. Barata, A. D. M. Ferreira, and L. G. Afonso. 2010. Lack of phylogenetic signal in the variation in anuran microhabitat use in southeastern Brazil. Evol. Ecol. 24:1–24.

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

Freckleton, R. P., P. H. Harvey, and M. Pagel. 2002. Phylogenetic analysis and comparative data: a test and review of evidence. Am. Nat. 160:712–726.

Gignac, L. D. 1992. Niche structure, resource partitioning, and species interactions of mire bryophytes relative to climatic and ecological gradients in Western Canada. The Bryologist 95:406–418.

Gignac, L. D., R. Gauthier, L. Rochefort, and J. Bubier. 2004. Distribution and habitat niches of 37 peatland Cyperaceae species across a broad geographic range in Canada. Can. J. Bot. 82:1292–1313.

Granath, G., J. Strengbom, and H. Rydin. 2010. Rapid ecosystem shifts in peatlands: linking plant physiology and succession. Ecology 91:3047–3056.

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

Hansen, T. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Harmon, L. J., J. T. Weir, C. D. Brock, R. E. Glor, and W. Challenger. 2008. GEIGER: investigating evolutionary radiations. Bioinformatics 24:129–131

Hemond, H. F. 1980. Biogeochemistry of Thoreau's Bog, Concord, Massachusetts. Ecol. Monogr. 50:507–526.

Ives, A. R., P. E. Midford, and T. Garland Jr. 2007. Within-species variation and measurement error in phylogenetic comparative methods. Syst. Biol. 56:252–270.

Karlin, E. F., G. Gardner, K. Lukshis, and S. B. Boles. 2010. Allopolyploidy in *Sphagnum mendocinum* and *S. papillosum* (Sphagnaceae). Bryologist 113:114–119.

Laing, C. G., G. Granath, L. R. Belyea, K. E. Allton, and H. Rydin. 2014. Tradeoffs and scaling of functional traits in *Sphagnum* as drivers of carbon cycling in peatlands. Oikos 123:817–824.

Losos, J. B. 2008. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. Ecol. Lett. 11:995–1003.

Martins, E., and T. Hansen. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am. Nat. 149:646–667.

Muller, K., J. Muller, and D. Quandt. 2010. PhyDE—Phylogenetic Data Editor. Available at http://www.phyde.de.

Newton, A. E., N. Wikström, and A. J. Shaw. 2009. Mosses (Bryophyta). Pp 138–145 *in* S. B. Hedges and S. Kumar, eds. The timetree of life. Oxford Univ. Press, Oxford, U.K.

O'Meara, B. C., C. Ané, M. J. Sanderson, and P. C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933.

Orme, C. D. L., R. P. Freckleton, G. H. Thomas, T. Petzoldt, S. Fritz, and N. Isaac. 2011. caper: comparative analyses of phylogenetics and evolution in R. Available at http://CRAN.R-project.org/package-caper.

Pagel, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc. R. Soc. Lond. B 255:37–45.

———. 1997. Inferring evolutionary processes from phylogenies. Zool. Scripta 26:331–348.

———. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Posada, D. 2008. jModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25:1253–1256.

Pouliot, R. 2011. Initiation du patron de butted et de dépressions dans les tourbiéres ombrotrophes boréales. Ph.D. dissertation, l'Université Laval, Quebec, Canada.

Pouliot, R., L. Rochefort, and E. Karofeld. 2012. Initiation of microtopography in revegetated cutover peatlands. Appl. Veg. Sci. 14:158–171.

Price, J. S., and P. N. Whittington. 2010. Water flow in *Sphagnum* hummocks: mesocosm measurements and modelling. J. Hydrol. 381:333–340.

Rambaut, A., and A. J. Drummond. 2014. Tracer v1.6. Available at http://beast.bio.ed.ac.uk/Tracer.

Revell, L. J. 2011. phytools: an R package for phylogenetic comparative biology (and other things). Meth. Ecol. Evol. 3:217–223.

Revell, L. J., D. L. Mahler, P. R. Peres-Neto, and B. D. Redelings. 2012. A new phylogenetic method for identifying exceptional phenotypic diversification. Evolution 66:135–146.

Ricca, M., and A. J. Shaw. 2010. Allopolyploidy and homoploid hybridization in the *Sphagnum subsecundum* complex (Sphagnaceae: Bryophyta). Biol. J. Linn. Soc. 99:135–151.

Rice, S. K., L. Aclander, and D. T. Hanson. 2008. Do bryophyte shoot systems function like vascular plant leaves or canopies? Functional trait relationships in *Sphagnum* mosses (Sphagnaceae). Am. J. Bot. 95:1366–1374.

Rochefort, L., M. Strack, M. Poulin, J. S. Price, M. Graf, A. Derochers, C. Lavoie, and L. Lapointe. 2012. Northern peatlands. Pp. 119–134 *in* D. P. Batzer and A. Baldwin, eds. Wetland habitats of North America: ecology and conservation concerns. Univ. California Press, Berkeley, CA.

Rydin, H., and J. K. Jeglum. 2013. The biology of peatlands. 2nd ed. Oxford Univ. Press, Oxford, U.K.

Rydin, H., U. Gunnarsson, and S. Sundberg. 2006. The role of *Sphagnum* in peatland development and persistence. Pp. 47–65 *in* K. Wieder and D. H. Vitt, eds. Boreal peatland ecosystems. Springer, Berlin, Germany.

Shaw, A. J., C. J. Cox, and S. B. Boles. 2003a. Global patterns in peatmoss biodiversity. Mol. Ecol. 12:2553–2570.

———. 2003b. Polarity of peatmoss (*Sphagnum*) evolution: who says bryophytes have no roots? Am. J. Bot. 90:1777–1787.

Shaw, A. J., C. J. Cox, W. R. Buck, N. Devos, A. M. Buchanan, L. Cave, R. Seppelt, B. Shaw, J. Larraín, R. Andrus, et al. 2010a. Newly resolved relationships in an early land plant lineage: Bryophyta class Sphagnopsida (peat mosses). Am. J. Bot. 97:1511–1531.

Shaw, A. J., N. Devos, C. J. Cox, S. B. Boles, B. Shaw, A. M. Buchanan, L. Cave, and R. Seppelt. 2010b. Peatmoss (*Sphagnum*) diversification associated with Miocene Northern Hemisphere climatic cooling? Mol. Phylogen. Evol. 55:1139–1145.

Soudzilovskaia, N. A., J. H. C. Cornelissen, H. J. During, R. S. P. van Logtestijn, S. I. Lang, and R. Aerts. 2010. Similar cation exchange capacities among bryophyte species refute a presumed mechanism of peatland acidification. Ecology 91:2716–2726.

Tahvanainen, T. 2004. Water chemistry of mires in relation to the poor-rich vegetation gradient and contrasting geochemical zones of the north-eastern Fennoscandian Shield. Folia Geobot. 39:353–369.

Tahvanainen, T., T. Sallantaus, R. Heikkila, and T. Tolonen. 2002. Spatial variation of mire surface water chemistry and vegetation in northeastern Finland. Ann. Bot. Fenn. 39:235–251.

Turetsky, M. R., S. E. Crow, R. J. Evans, D. H. Vitt, and R. K. Wieder. 2008. Trade-offs in resource allocation among moss species control decomposition in boreal peatlands. J. Ecol. 96:1297–1305.

van Breemen, N. 1995. How *Sphagnum* bogs down other plants. Trends Ecol. Evol. 10:270–275.

Vitt, D., and N. G. Slack. 1984. Niche diversification of *Sphagnum* relative to environmental factors in northern Minnesota peatlands. Can. J. Bot. 62:1409–1430.

Associate Editor: D. Polly

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Table S1. GenBank accession numbers for each species at each gene.
Table S2. Species mean and SD for eight niche descriptors, summarized over five ecological sampling studies.
Table S3. Model selection for trait evolution using AICc in eight niche descriptors and two microhabitat gradients.
Figure S1. Maximum credibility tree from BEAST analysis, created using TreeAnnotator.
Figure S2. Bayesian inference of rate change in niche preference for eight niche descriptors and two multivariate niche gradients.
Figures S3. Distributions of niche preferences in eight niche characters, aligned with the maximum credibility tree.

Heredity (2016), 1–8

www.nature.com/hdy

ORIGINAL ARTICLE

The effects of quantitative fecundity in the haploid stage on reproductive success and diploid fitness in the aquatic peat moss *Sphagnum macrophyllum*

MG Johnson[1] and AJ Shaw

A major question in evolutionary biology is how mating patterns affect the fitness of offspring. However, in animals and seed plants it is virtually impossible to investigate the effects of specific gamete genotypes. In bryophytes, haploid gametophytes grow via clonal propagation and produce millions of genetically identical gametes throughout a population. The main goal of this research was to test whether gamete identity has an effect on the fitness of their diploid offspring in a population of the aquatic peat moss *Sphagnum macrophyllum*. We observed a heavily male-biased sex ratio in gametophyte plants (ramets) and in multilocus microsatellite genotypes (genets). There was a steeper relationship between mating success (number of different haploid mates) and fecundity (number of diploid offspring) for male genets compared with female genets. At the sporophyte level, we observed a weak effect of inbreeding on offspring fitness, but no effect of brood size (number of sporophytes per maternal ramet). Instead, the identities of the haploid male and haploid female parents were significant contributors to variance in fitness of sporophyte offspring in the population. Our results suggest that intrasexual gametophyte/gamete competition may play a role in determining mating success in this population.
Heredity advance online publication, 24 February 2016; doi:10.1038/hdy.2016.13

INTRODUCTION

A major focus in population biology is the estimation of mating success among individuals in a population and the consequences of parental mating patterns on offspring fitness (Fisher, 1930; Gustafsson, 1986; Merilä and Sheldon, 1999). In animals and most plants, 'parents' are defined as the diploid individuals that mate to produce diploid offspring. This conceptual formulation of parentage is limited because (except in severely inbred diploid individuals) the gametes they produce are genetically variable and will therefore have variable affects on offspring fitness. Although the variation can be studied in aggregate (as the result of recombination in the diploid parent), the expected fitness impact of individual gametes cannot be studied.

In agriculture, the genetic contribution of (diploid) parents to the fitness of their offspring is often deduced using controlled breeding experiments (Griffing, 1956). The 'general combining ability' of a parent is assessed through crosses with many mates (Falconer and Mackay, 1996). The genetic contribution of (diploid) parents to offspring is also studied in aggregate in sexual selection studies, where 'cryptic female choice' refers to apparent nonrandom mating success among haploid sperm in the female reproductive tract (Gasparini *et al.*, 2010; see, for example, Firman and Simmons, 2010; Boschetto *et al.*, 2011). In seed plants, pollen competition experiments frequently indicate that seed set (average) fitness increases with the number of pollen donors (Mulcahy and Mulcahy, 1975; Winsor *et al.*, 2000; Zhang *et al.*, 2010). Both examples suggest that intrasex competition at the haploid stage is important for determining offspring fitness (Snow and Spira, 1991; Marshall and Diggle, 2001), but the success of individual sperm or egg genotypes cannot be directly assessed.

From the perspective of a gamete, fecundity can be defined as the number of fertilization events the gamete participates in. For animals and seed plants, millions of genetically distinct gametes are produced by the diploid parent, and each gamete either participates in fertilization or not. As gametes in these organisms do not generally clonally replicate, 'parental' reproductive success is applied to diploid individuals rather than their haploid gametes (which are, more specifically, the actual parents). Gamete fecundity may be viewed as a binary trait in these organisms (that is, successful or not).

Because of their homosporous, haploid-dominant life cycle, bryophytes present a unique opportunity to study the effects of the haploid stage on mating success and offspring fitness. Individual haploid gametophytes produce thousands of genetically identical gametes via mitosis, and reproductive success of gamete genotypes is therefore quantitative rather than binary. In addition, clonal growth of the haploid stage before gamete production can result in genetically identical zygotes produced simultaneously throughout the population. These life-cycle features permit estimation of reproductive success for individual gamete genotypes, as well as the fitness consequences of specific haploid mating combinations.

In this study, we present an investigation of *haploid* parental contributions to fertilization success in a natural population of the peat moss *Sphagnum macrophyllum*, and the consequences of various gamete matings for zygote fitness. We sought to test the hypothesis

Department of Biology, Duke University, Durham, NC, USA
Correspondence: Dr MG Johnson, Department of Plant Sciences, Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, IL 60022, USA.
E-mail: mjohnson@chicagobotanic.org
[1]Current address: Department of Plant Sciences, Chicago Botanic Garden, Glencoe, IL, USA.
Received 16 August 2015; revised 23 November 2015; accepted 29 December 2015

that the identity of parental gamete genotypes affects the fitness of diploid offspring. Toward that end, we (1) assess variance in haploid fecundity (reproductive success) among individuals, (2) correlate haploid fecundity with the number of different mates for each individual and (3) determine the relative contribution of haploid parental genotype to diploid offspring fitness, compared with other factors including inbreeding depression and maternal resource limitation.

Our study population and the bryophyte life cycle allows us to study the 'Bateman gradient' in haploids for the first time. We calculated the mating success (number of genetically different mates) and fecundity (number of sporophyte offspring) for each maternal and inferred paternal genotype. In its original formulation, sexual selection was inferred from extreme secondary sexual characteristics, and defined by competition between individuals to mate and produce offspring (Darwin, 1859). The quantitative description of sexual selection later became synonymous with the relationship between fecundity (the number of offspring) and mating success (the number of different mates) (Bateman, 1948). When the strength of this relationship differs between the sexes, the sex with the reduced variance in mating success benefits from the ability to choose among many potential mates (Arnold, 1994). Our results demonstrate the utility of bryophytes to answer questions about reproductive biology that are intractable or impossible in seed plants or animals.

MATERIALS AND METHODS

The bryophyte life cycle

As with all land plants, bryophytes have an alternation of generations between a diploid spore-producing stage (sporophyte) and a haploid gamete-producing stage (gametophyte) (Figure 1). Both stages are multicellular and gametes are produced via mitosis from haploid gametophytes. Peat mosses (*Sphagnum*) are quasi-broadcast spawners; sperm are released into the water, whereas eggs are retained on female plants. The diploid sporophyte remains attached to the maternal gametophyte and is essentially a sphere containing, at maturity, meiotically produced spores. Although immature sporophytes are green, they lose their chlorophyll as they mature and are dependent on the maternal gametophyte for nourishment (Duckett *et al.*, 2009). Before spore release the sporophyte is raised by a pseudopodium of maternal gametophyte tissue. Meiotically produced spores germinate and grow into large, persistent haploid gametophytes that can multiply via clonal propagation. Each independent gametophyte (ramet) may share an identical genotype with hundreds of other disconnected ramets—collectively, the haploid genet. Depending on the clonal structure of the population, genets may be represented by few to many independent ramets, and these ramets may be dispersed throughout the population.

In an idealized bryophyte population, let there be one male and one female genet. The (haploid) male genet undergoes vegetative propagation and each of the many ramets produce sperm by mitosis, giving rise to genetically identical sperm. The same is true of the (haploid) female clone that produces many genetically identical eggs. Fertilization between gametes produced by the one male clone and one female clone would produce a population of diploid sporophytes that are genetically identical because they result from fusion of the same male and female gamete genotypes. An important implication of this life cycle is that genetically equivalent fertilization events may be replicated in the population. In a genetically variable (multiclonal) population, the fecundity (fertilization success) of each male (and female) gametophyte genet need not be binary. Instead, haploid bryophyte fecundity is a quantitative trait. For example, one haploid male gametophyte may participate in one to many fertilization events, or none. This is in contrast to seed plants or animals, in which a gamete (or gametophyte) can participate in a maximum of one fertilization event—haploid fecundity is binary.

The bryophyte life cycle also allows for unequivocal resolution of haploid parentage for diploid offspring. Because the sporophyte remains physically attached to the maternal gametophyte throughout its life (Figure 1), maternity is obvious and paternity can be inferred by subtracting the maternal haploid genotype from the diploid sporophyte genotype (Szövényi *et al.*, 2009).

Study site and species

The study site is in the Hell Hole Bay Wilderness Area in the Francis Marion National Forest, Berkeley County, South Carolina (33.218° N 79.712° W). The sampling site is a break in the otherwise dense swamp pocosin: an open canopy of *Taxodium* trees standing in up to a meter of water. The understory contains several trees and shrubs including *Nyssa biflora*, *Lyonia lucida* and *Vaccinium formosum*. The aquatic herbaceous plants, *Dulichium arundinaceum*, *Carex striata* and *Nymphaea odorata*, occur scattered in the open water with several other peat moss species, including *Sphagnum cuspidatum*, *S. magellanicum* and *S. portoricense*. However, the dominant plant, floating in the water in nearly continuous mats covering an area of roughly one acre (4000 m^2), is the aquatic peat moss *S. macrophyllum*.

S. macrophyllum is in the subgenus *Subsecunda* (Shaw *et al.*, 2008a), and occurs in North America along the coastal plains of the Atlantic Ocean and Gulf of Mexico from Newfoundland to Texas (McQueen and Andrus, 2009). It usually grows in acidic waters such as roadside ditches, the margins of small ponds and in wet areas of pine savannahs, especially where there is underlying sandy soil (Anderson *et al.*, 2009). Based on extensive field work throughout eastern North America, Hell Hole Swamp has the largest biomass of *S. macrophyllum* we have observed at a single site. It is also the only site where *S. macrophyllum* undergoes consistent and abundant sexual reproduction; the

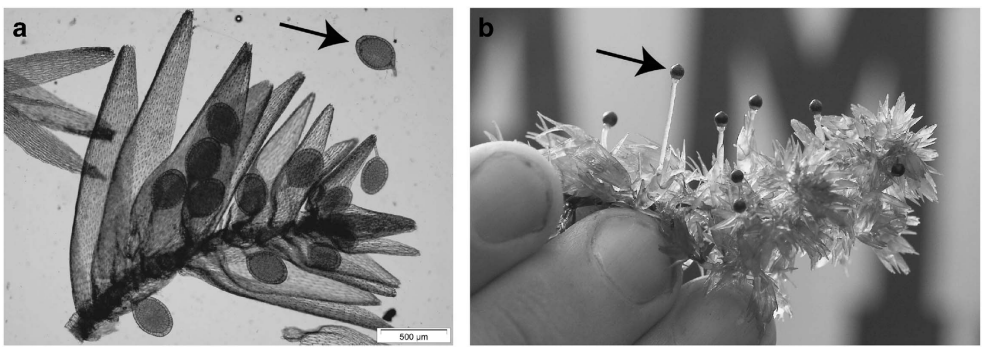

Figure 1 Reproductive structures in the peat moss *S. macrophyllum*, a species that has separate male and female haploid gametophytes. (**a**) In *S. macrophyllum*, male plants produce sperm in antheridia (arrow), of which at least 12 can be seen. The gametes are produced by mitosis from the perennial male plants; therefore, all sperm produced in all antheridia shown are genetically identical to each other, and to the male gametophyte that produced them. Sperm are released and swim through water to female plants to effect fertilization. Photo by MG Johnson. (**b**) A haploid female gametophyte (large leafy plant) can be identified by the presence of diploid sporophytes (arrow; small black spheres). This female ramet has a sporophyte brood of at least eight. Each sporophyte in this brood thus shares exactly half of its diploid genotype; genetic differences within the brood could only result from having different fathers. The translucent tissue that exerts each sporophyte is haploid, produced by the maternal gametophyte. Photo by B Shaw (Duke University, Durham, NC, USA).

plants produced sporophytes every year since we began sampling the site in 2005 (MG Johnson *et al.*, unpublished). Sporophytes are otherwise uncommon in this species.

Reproductive phenology survey

To describe the sex ratio and reproductive phenology of the population, we sampled gametophytes throughout a breeding season from December 2009 to June 2010. We established eight 'neighborhoods' throughout the population of floating mats of *S. macrophyllum*, with each neighborhood marked by a PVC pipe so that sampling would occur in the same area each month (for a schematic of the population, see Supplementary Figure 1). At each visit we sampled 12 gametophytes from within 1 m of the pipe, for a total of 96 plants sampled each month. We then assessed the sexual expression of gametophytes under a dissecting microscope.

S. macrophyllum, like all members of the subgenus *Subsecunda*, has unisexual gametophytes that produce either sperm (male) or eggs (female) but not both. There is little morphological sex differentiation, but male gametophytes are identified by antheridia (the male gametangia that produce sperm). Our observations indicate that antheridia are produced between December and March; plants are not identifiable as male outside this temporal range.

Female gametophytes are more challenging to identify in the field and require microscopic examination; archegonia (the female gametangia that produce eggs) are less conspicuous than are antheridia and often difficult to locate. Archegonia usually turn a reddish color after fertilization and this can help. The most reliable way to identify a gametophyte as female is by the presence of sporophytes because they remain attached to their maternal gametophyte. Females are identifiable beginning in late December, and young, macroscopic sporophytes are obvious by February. Sporophytes mature from April through June, and spores are released by July.

We characterized genetic diversity among male gametophytes as well as those gametophytes with no evidence of sexual expression ('nonexpressing'). However, male and female plants become sexually mature at different times, and because gametangia (or attached sporophytes in the case of females) are required to confirm gametophyte sex, we sampled plants throughout the breeding season from December through June. We marked 8 locations (neighborhoods) and randomly sampled 12 gametophytes (ramets) within 1 m^2 at each neighborhood each month.

We determined the sex of each gametophyte by microscopic dissection, classifying each as 'male' (presence of antheridia), 'female' (presence of archegonia or attached sporophytes) or 'nonexpressing' (neither gametangia nor sporophytes observed). In order to determine whether female plants had already mated at the time of collection, all female gametophytes were transferred to individual tubes containing sterile water and observed for subsequent maturation of sporophytes. All gametophytes identified as females in the 2009–2010 breeding season ($N = 126$, Table 1) eventually produced sporophytes in the lab, indicating that each had already mated at the time of collection.

Table 1 Phenology of sexual expression and sporophyte production in *S. macrophyllum* at Hell Hole Swamp

Month	*Male*	*Female*	*Nonexpressing*	*Sporophytes*
December 2009	46	7	43	0
January 2010	40	16	40	66
February 2010	44	16	36	76
March 2010	40	31	25	182
April 2010	28	18	50	82
May 2010	0	24	72	151
June 2010	0	14	82	45
Total	198	126	348	602

A total of 96 plants were sampled at the same 8 locations (12 per location) within the population throughout one breeding season. Male plants were identified by the presence of antheridia (sperm-producing structures) during microscopic dissection. Female plants were identified through the presence of archegonia (egg-producing structures) or sporophytes. Plants with no sexual structures of either type were recorded as nonexpressing.

Sporophyte sampling and fitness proxy

We sampled mature sporophytes over 2 years: April–May 2009 and March–June 2010. In 2009, sporophytes were sampled throughout the population; in 2010, we sampled mature sporophytes that were attached to the female gametophytes we sampled during the monthly phenology survey. *Sphagnum* sporophytes are roughly spherical capsules containing, at maturity, haploid spores (products of meiosis). The diameter of the capsule is a reliable predictor of spore number across multiple species in *Sphagnum* (Sundberg and Rydin, 1998). We measured the diameters of all sporophytes in the study when they were at full maturity and we use this measure as a proxy for sporophyte fitness (Szövényi *et al.*, 2009). This phenotypic measurement allows for a simple estimate of potential reproductive output for the sporophyte without damaging the tissue necessary for genotyping the sporophyte.

Genotyping

Using material collected from the phenology survey, we genotyped every gametophyte identified as female (75 samples), a representative sample of male (58 samples), and nonexpressing (65 samples) gametophytes from each neighborhood each month. Sporophytes were sampled throughout the population in April and May 2009, in addition to all mature sporophytes sampled during the phenology survey in March through June 2010 (321 total sporophytes).

For DNA extractions of gametophyte tissue, a portion of the gametophyte's capitulum (the dense cluster of branches at the apex of *Sphagnum* plants) was removed. For sporophytes, the entire capsule was used. DNA extraction followed the cetyltrimethylammonium bromide protocol (Shaw *et al.*, 2003). Briefly, tissue was flash frozen in tubes using liquid nitrogen and ground into a powder using steel beads. Cellulose and proteins were removed using cetyltrimethylammonium bromide and chloroform mixed with β-mercaptoethanol in a 24:1 ratio. DNA pellets were precipitated using cold isopropanol and eluted in 50 μl of TE buffer. In preparation for PCR amplification, we diluted the DNA of gametophytes 7:1, and diluted sporophytes 2:1.

All plants were genotyped at 15 microsatellite loci, labeled as in Shaw *et al.* (2008b), 1, 4, 7, 9, 10, 14, 17, 18, 19, 20, 29, 65, 68, 75 and 93, using previously described protocols.

Paternity and genet assignment

Because 'parentage' is defined at the haploid stage and because sporophytes remain attached to their maternal gametophyte, inferring the multilocus paternal haploid genotype is straightforward. Paternal alleles can be identified by simply subtracting the maternal haploid genotype from an attached sporophyte's diploid genotype. In the case of homozygous loci, the paternal and maternal genotypes were assumed to be identical. The inferred haploid paternal multilocus genotype for every sporophyte was determined using the custom Python script. The script determined the inferred paternal genotype by comparing the observed sporophyte genotype with its maternal haploid genotype. All sporophyte alleles not present in the maternal genotype were assumed to have originated in the paternal genotype. All of the sporophytes analyzed had an allele matching its maternal gametophyte, indicating that mutation rates at the microsatellite loci are minimal at this scale.

To characterize clonal structure of the population, a second Python script was used to assign every sample to unique multilocus genotypes. All haploid multilocus genotypes were included: female, male and nonexpressing gametophytes, as well as inferred paternal genotypes. The script first sorted samples with no missing data into separate multilocus genotypes. Samples with missing data were then reconciled with the initial multilocus genotypes and could have three outcomes: (1) the sample is unambiguously assigned to just one genotype, (2) the sample could be assigned to more than one genotype and (3) the sample represents a new genotype not seen previously. Samples were included in further analysis only if they unambiguously matched to exactly one unique multilocus genotype.

Genetic analyses

Population genetic statistics, including PhiPT (a multilocus equivalent of F_{ST}), analysis of molecular variance and principal coordinate analysis, were calculated using GenAlEx version 6.5 (Peakall and Smouse, 2012). Missing data were not interpolated in any analysis using GenAlEx.

For estimating diversity and inbreeding statistics, we needed to take into account that sporophytes attached to the same maternal gametophyte are not random draws from the population (that is, they share the same haploid mother). To partially address this bias, we generated 1000 'subsets' of each population using a modified version of the bootstrap method of Szövényi *et al.* (2009). Each subset contained a random draw of one sporophyte per (haploid) maternal brood. For instance, in a population with three maternal gametophytes (A, B and C), each bearing four sporophytes (numbered 1–4), a subset population would contain one sporophyte randomly drawn from each mother, such as (A1, B2, C3). We then repeated this procedure by randomly drawing another sporophyte (with replacement) from each mother and recalculating each statistic on this new subset, such as (A3, B2, C4). This was done a total of 1000 times, resulting in a range of values for each population statistic. F_{IS} was calculated from the observed and expected heterozygosity of sporophytes within this subset, and this was repeated 1000 times to estimate a range of inbreeding coefficients for the whole population.

We calculated the 'Bateman gradient' for the haploid genotypes with a linear regression of fecundity versus mating success in each sex. A slope significantly different from zero indicates that more mates leads to more offspring, and can indicate the overall mating pattern for the population: polygamous, polyandrous or polygynous. We also compared the slopes using an analysis of covariance. If the slope of the relationship between fecundity and mating success significantly differs between the sexes, it suggests the presence of sexual selection.

We explored three main explanations for variance in fitness in the population of sporophytes. First, we calculated the relationship between heterozygosity (percentage of heterozygous loci) and fitness at the level of the individual sporophyte and at the level of the brood. Each statistic was calculated using the pseudo-population approach described above to avoid bias because of non-independence of sporophytes attached to the same maternal gametophyte. Second, we investigated whether the number of sporophytes in a brood affected the average size of sporophytes in the brood. Third, we conducted an analysis of variance to determine whether maternal or paternal genet identity was a significant predictor of sporophyte fitness, using Type III sums of squares.

All statistical analyses were conducted in R (version 3.1, R Core Development Team, 2012). R and Python scripts used to generate these analyses are available at github.com/mossmatters/moss_popgen_scripts, and data tables containing the multilocus genotypes of gametophytes and sporophytes are available in the Dryad data repository (http://dx.doi.org/10.5061/dryad.qj3hn).

RESULTS

Hell Hole Swamp: phenology and genetic diversity

The phenology survey in *S. macrophyllum* in the 2010 breeding season revealed a heavily male-biased sex ratio (Table 1) at the ramet level. Male gametophytes could be identified (by the presence of antheridia) in 5 of the 7 months, and males always outnumbered females. In most months, a large percentage of gametophyte samples could not be determined as either male or female and were recorded as nonexpressing. Though this could reflect the greater difficulty of identifying female plants, the nonexpressing samples we genotyped most often matched male genotypes (see below). Overall, there were more male-expressing samples early in the breeding season, and female-expressing samples appeared later. However, the seven females identified in December 2009 all produced sporophytes when separately stored in sterile water, indicating that they were fertilized before collection. This observation, combined with field observations of exerted sporophytes in March, indicates that *S. macrophyllum* sporophytes require 3–4 months to fully mature.

Of the 197 gametophytes (ramets) genotyped, 149 were unambiguously assigned to one of 65 unique multilocus haploid genotypes (genets). In all, 17 genets included at least one demonstrably male ramet (identified by presence of antheridia) and 22 genets included at least 1 female ramet (identified by the presence of sporophytes). Of the nonexpressing ramets, 47 were unambiguously assigned to one genet—21 nonexpressing ramets matched a male genet whereas 8 ramets matched a female genet. The remainder of the nonexpressing ramets sorted into genets containing only nonexpressing individuals.

Among the female ramets, half (37 of 75) belong to a single genet that occurred in 6 of the 8 neighborhoods. The next most common genet contained just four ramets. One genet comprising 25 of 78 ramets dominated the male plants.

Male and female genets were genetically divergent (Figure 2). Female genets were significantly differentiated from male genets (PhiPT = 0.103, $P < 0.001$) and from nonexpressing genets (PhiPT = 0.039, $P < 0.05$). Male genets were not differentiated from nonexpressing genets (PhiPT = 0.00). A significant proportion of genetic differentiation among all ramets at the site was among neighborhoods (PhiPT = 0.088, d.f. = 7, $P < 0.01$).

Fecundity and mating success

A total of 321 mature sporophytes were sampled and represent 10 unique maternal genets and 91 different inferred paternal genets. One single female, mentioned above, mated with 60 different paternal genets to produce 68.5% (220 of 321) of the sporophytes (Figure 3).

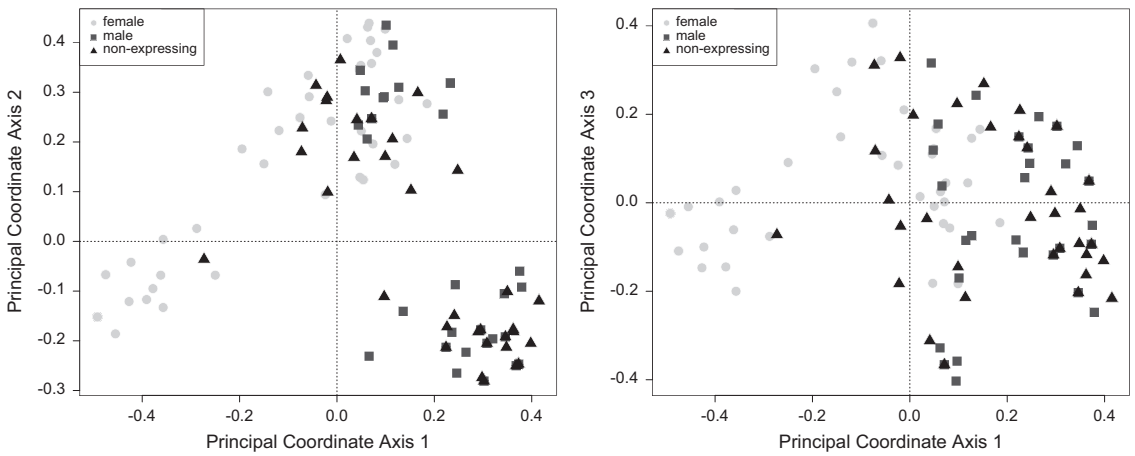

Figure 2 Principal coordinate analysis of genetic diversity in gametophytes. All gametophytes were classified as male (presence of antheridia) or female (presence of sporophytes). All other samples are classified as nonexpressing. When a nonexpressing ramet (triangles) has the same multilocus genotype as a male (squares) or female (circles) ramet, the points overlap. Percent of total variance: axis 1: 39.0%; axis 2: 23.5%; axis 3: 13.5%.

Each of the other maternal genets also mated with multiple paternal genotypes.

The skew in fecundity was also pronounced for the paternal genotypes (Figure 3). Three of the most commonly inferred genotypes together accounted for 36% of the sporophytes. Comparatively few of the inferred paternal genotypes matched male gametophytes we sampled: only 7 of the 91 inferred paternal genotypes match any gametophytes actually sampled in the population. A total of 19 inferred paternal genets were identified in at least 3 sporophytes; 6 of these fathers mated with more than one maternal genet.

The correlation between fecundity (number of offspring) and mating success (number of different genetic mates) was significant for both males ($r^2 = 0.644$, $P < 2 \times 10^{-16}$) and females ($r^2 = 0.983$, $P < 2.2 \times 10^{-8}$) (Figure 3). The slope of the male gradient was significantly steeper than the female gradient (male slope = 11.25, female slope = 3.70; analysis of covariance d.f. = 98, F = 63.14, $P < 3.3 \times 10^{-8}$). Both the relationship between mating success and fecundity for females, and the difference in slopes, remained significant after excluding the most common female (results not shown).

With the exception of the one very large female clone, there is no relationship between the number of ramets in a female genet and the number of mates. One female genet was identified from just one ramet but had sporophytes with eight different inferred fathers, whereas another genet identified from four ramets mated with only two different fathers. In addition, the relationship between the number of mates and the number of offspring is significant even when excluding the largest female clone. This suggests that size is not the only factor controlling mating success or fecundity.

Sporophyte fitness

The mean sporophyte size (our proxy for diploid offspring fitness) in the population was 1508 μm (±108 μm s.d.). We investigated several sources of variance in sporophyte size: inbreeding, limitation of resources from the maternal gametophyte and the genotype of the maternal and paternal gametophyte parents.

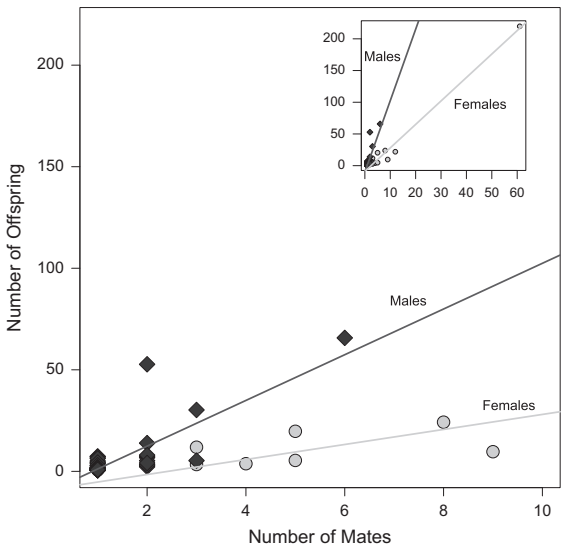

Figure 3 Bateman gradient for *S. macrophyllum* at Hell Hole Bay. The effect of mating success (number of different mates) on fecundity (number of sporophyte offspring) is shown for male (diamonds) and female (circles) gametophytes. The inset shows the full data set, including the very large female genet (top right), and the main plot is zoomed in to show the differentiation between male and female fecundity. Trend lines for each sex include all data points in both plots.

Using 1000 pseudoreplicate populations, the average F_{IS} was −0.11 (±0.04 s.d.). In the three neighborhoods where sporophytes were most common, F_{IS} was even more negative: neighborhood 3 ($F_{IS} = -0.28$, ±0.06 s.d.), neighborhood 4 ($F_{IS} = -0.34$, ±0.05 s.d.) neighborhood 6 ($F_{IS} = -0.14$, ±0.02 s.d.). The average multilocus heterozygosity was 54%, and individual sporophytes ranged from 20 to 100% heterozygous loci.

There was a small but significant relationship between heterozygosity and sporophyte fitness (Figure 4a). The number of sporophytes attached to a female gametophyte ramet ('brood size') could limit the amount of maternal resources available to each sporophyte, potentially reducing the average fitness of the brood. However, there was no effect of brood size on the average brood fitness (Figure 4b, range 2–18, $n = 76$, $r^2 = 0$).

Instead, the genetic identity of the maternal (d.f. = 5, F = 4.24, $P < 0.01$) and paternal (d.f. = 22, F = 3.58, $P < 0.001$) (haploid) parents explained much of the variance in sporophyte size (analysis of variance, Type III SSS, singletons removed). We detected no maternal × paternal interaction effect (d.f. = 8, F = 1.13, $P > 0.1$), and paternal identity (9.2%) accounted for greater variance than maternal identity (3.6%). Because of low sample size, the effects of specific maternal or paternal genotypes could not be assessed.

To further investigate the effect of paternal identity on sporophyte fitness, we considered only the sporophytes attached to the most common female genet and classified the sporophytes based on paternal identity. In this analysis, all sporophytes in a category have the same haploid mother and the same haploid father, and therefore represent repeated fertilization events across the population. Despite considerable variation within each paternal group, there was a highly significant effect of haploid paternal identity on sporophyte fitness (Figure 5, analysis of variance, Type III SSS $F_{16} = 5.25$, $P < 10^{-7}$).

DISCUSSION

Phenology and genetic structure

In contrast to the female-biased sex ratio found in most natural populations of bryophytes (reviewed in Bisang and Hedenäs, 2005), sexual expression in *S. macrophyllum* at Hell Hole Swamp is male biased (3:1 male/female ratio among sampled gametophytes). In other bryophytes, the low level of male sex expression has been attributed to a 'shy male hypothesis'; genetically male gametophytes do not always produce antheridia, obscuring an even sex ratio (Stark *et al.*, 2010). If the sex ratio in our study population were truly even, then there must instead be the opposite effect— the presence of 'shy females' that are not producing archegonia or are not successfully fertilized. However, male and female gametophytes form distinct clusters on the principal coordinate analysis (Figure 2), and plants that we classified as nonexpressing are genetically distinct from the female, but not the male cluster. The high diversity of inferred paternal genotypes compared with the observed maternal genotypes also suggests sex bias at the genet level, although there may be a large unsampled pool of nonexpressing female genotypes.

Given the high amount of genetic diversity in the population, and the lack of a sex-specific genetic marker in *Sphagnum*, an even more comprehensive sampling of gametophytes would be required to determine whether the population is male biased at the genet level as well. However, it is worth mentioning that our sampled nonexpressing gametophytes were more likely to match the microsatellite profiles of male plants (21 nonexpressing ramets) than match genets including known females (8 nonexpressing ramets). Sex ratio bias

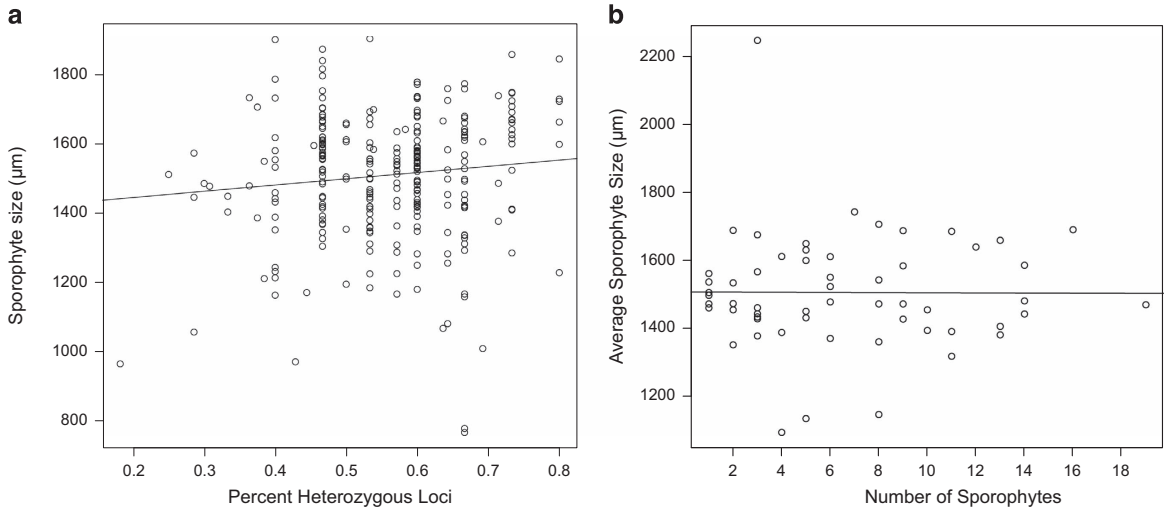

Figure 4 Effects of inbreeding depression and resource limitation on sporophyte fitness. (**a**) Relationship between percent heterozygosity and sporophyte size ($r^2 = 0.018$, $P < 0.05$). (**b**) Effect of 'brood size' (number of sporophytes supported by a maternal gametophyte) on average brood fitness ($r^2 = 0.00$).

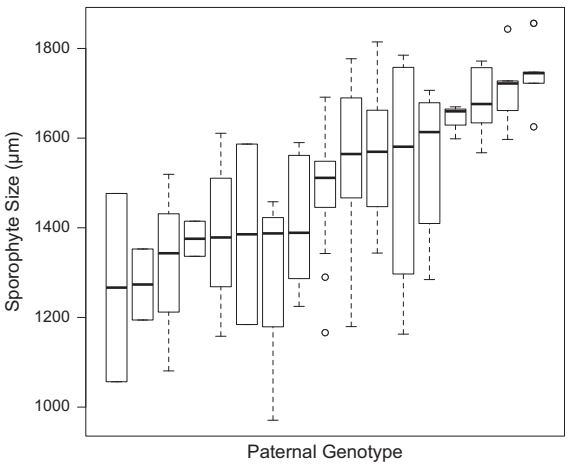

Figure 5 Distribution of sporophyte fitness by genetic father. All sporophytes shown have the same (haploid) genetic mother. Each box represents identical diploid sporophytes formed throughout the population; all variance within a box is environmental. All sporophytes within one column have the same haploid mother and the same haploid father, generating repeated fertilization events across the population. Despite the variance found within each box, there is a highly significant effect of paternal identity on sporophyte fitness for this subset (analysis of variance (ANOVA), $F = 5.25$, d.f. = 16, $P < 10^{-7}$), indicating the effect of haploid paternal identity on sporophyte fitness must have a large genetic component.

among ramets in bryophytes may be a result of a bias in sex expression ('shy males'), a bias in growth rate between males and females or a result of biased spore death in sporophytes (Norrell *et al.*, 2014). Distinguishing among these possibilities will require more detailed study at the level of sporelings and the development of sex-specific genetic markers for *Sphagnum*, as has been done with other mosses (Korpelainen *et al.*, 2008; Hedenaes *et al.*, 2010). Peat mosses are perennial plants that can reproduce asexually and hence the realized sex ratio bias in the population is likely to be stable over many years and breeding seasons. This is consistent with trends in seed plants, where sex ratio bias is more likely to be found in perennial plants (such as tree species) (Field *et al.*, 2013).

A pronounced feature of sporophytes we sampled is an excess of heterozygosity, resulting in estimates of F_{IS} that are significantly below zero. Using similar methods, Szövényi *et al.* (2009) also found a slightly negative F_{IS} for *Sphagnum lescurii*. Other *Sphagnum* populations of species with unisexual gametophytes generally exhibit inbreeding coefficients below zero (Johnson and Shaw, 2015). Using isozymes, Eppely *et al.* (2007) found low (though positive) inbreeding coefficients in several other bryophytes with unisexual gametophytes. In sexually reproducing organisms with separate male and female sexes, small population sizes can lead to an excess of heterozygosity because of the effects of binomial sampling with small sample sizes (Rasmussen, 1979). Although considerable microsatellite variation exists among genets of *S. macrophyllum*, some may be a result of somatic mutations rather than sexual recombination. If this is true, then the true effective population size is very small. Because there is strong differentiation by sex in this population (Figure 2), fertilization events are likely to be from very different genotypes, resulting in highly heterozygous sporophytes. A broader survey of mating patterns in *Sphagnum* populations revealed that a highly negative F_{IS} is especially common in populations of other aquatic species with unisexual gametophytes (Johnson and Shaw, 2015).

Sporophyte size does increase with heterozygosity in the study population (that is, evidence of inbreeding depression), but the effect is weak (Figure 4a). However, a large majority of sporophytes in this study had >50% heterozygous loci. The excess of observed heterozygosity may result from homozygous sporophytes that did not reach maturity. In another aquatic species of subgenus *Subsecunda* there was a stronger relationship between heterozygosity and sporophyte fitness (Szövényi *et al.*, 2009), but results from other unisexual species of *Sphagnum* suggest that inbreeding depression is not universal in *Sphagnum* (Johnson and Shaw, 2015).

Variance in fecundity and mating success

We found a high variance in the number of different mates, and the number of offspring they yield, for gametophytes of both sexes. Multiple paternity in sporophyte 'broods' raised by the same gametophyte mother was high, as has been previously described in *S. lescurii* (Szövényi *et al.*, 2009). Unlike the previous study (where male gametophytes were monogamous or mated with few females), most paternal genotypes in *S. macrophyllum* mated with multiple female genets. Our study demonstrates abundant promiscuity in both sexes.

The high variance in fecundity and mating success in both sexes provided an opportunity to test for the Bateman gradient among haploid plants in this population. The strength of the association in males compared with the association in females implies the underlying mating system for the population (Arnold and Duvall, 1994). In *S. macrophyllum*, there was a significant association between fecundity and mating success in both sexes (Figure 3). The greater slope of the gradient in male gametophytes suggests a polygamous mating system. In many other organisms, a significant difference in slopes between the sexes on the Bateman gradient has been interpreted as a signal of sexual selection (Levitan, 1998; Becher and Magurran, 2004; Tatarenkov *et al.*, 2008; Anthes *et al.*, 2010).

It is not clear what role, if any, sexual selection could play in a population of aquatic mosses. Sexual selection is expected to be most effective in mating systems where one sex has an upper fecundity limit that would put strong pressure on the other sex to compete for fertilization success (Arnold and Duvall, 1994). For *S. macrophyllum*, the Bateman gradient is linear for both sexes, although steeper for males (Figure 3). The trend in females is primarily because of one very large female genet that accounted for 50% of female ramets and 68% of sporophytes produced. Therefore, an alternative explanation for the relationship between fecundity and mating success is natural selection for clone size, rather than more direct competition for fertilization success within sexes.

There could also be sexual selection for clonal growth patterns. Given equivalent biomass, a clone that creates a large mate in one location may not contact as many potential mates as a clone that spreads outward and/or fragments. Clonal growth patterns are known to be extremely variable in *S. cribrosum*, the sister species to *S. macrophyllum* (Johnson *et al.*, 2012). Given the higher slope of the Bateman gradient in males (Figure 3), there may be greater competition for exposure to multiple females as a mechanism to increase mating success and thus fecundity. In this study it was not possible to determine the effect of clone spatial breadth (vs clone size), because only a few of the male gametophytes sampled matched the inferred paternal genotypes.

Although purely speculative, the possibility that sexual selection could operate on gamete recognition in bryophytes is certainly feasible, given that sperm must swim to the eggs. Experimental tests could determine whether, as in seed plants (Snow and Spira, 1991) and sea urchins (Levitan, 2004), the number of available sperm donors affects fertilization success. In bryophytes, there may be competition between genetically identical sperm produced by the same genetic individual, and among male genotypes, an additional level of sperm competition not possible in other organisms. Experimental manipulation of sperm availability could test whether there is sperm competition, whether female genotypes choose among sperm, whether there is variance in chosen sperm genotypes and whether selection results in increased sporophyte fitness.

Haploid parental effects on sporophyte fitness

The strongest determinant of sporophyte fitness in the study population is the identity of the maternal and paternal haploid genotypes. This finding would be intractable or impossible to dissect in seed plants or animals, because it relies on the potential for quantitative fecundity in the haploid generation. The clonal propagation of gametophytes and the production of gametes via mitosis permits multiple, repeated fertilization events. In the most extreme case, one haploid female genet was fertilized by the same haploid male genet 55 times in 6 of the 8 neighborhoods. All of these diploid sporophytes are genetically identical; any size variance among these sporophytes must be environmental. Despite this, both maternal and paternal identity explained a significant proportion of sporophyte fitness variance.

This effect is in some ways analogous to general and specific combining ability in breeding experiments (Falconer and Mackay, 1996). A haploid male parent would have high general combining ability if, when crossed with multiple females, its offspring are on average more fit than the rest of the population. In quantitative genetics, this effect is important because it can be used to estimate the additive genetic variance in the population. Specific combining ability is the performance of one particular cross and represents nonadditive genetic variance. We did observe a higher relative fitness of certain males in the population when they crossed with the most common female genet (Figure 5). However, because this is a natural population and not a designed cross, we did not observe all possible combinations of males and females. This means there was not enough power to detect whether specific males or females produced larger offspring.

There is the potential for intergenome conflict among haploid parental genotypes in diploid sporophytes. Because all sporophyte offspring in *Sphagnum* are related to the maternal gametophyte equally, there could be selective pressure for maternal gametophytes to spread resources among all attached sporophytes. In contrast, there may be competition among the paternal genomes in a brood, with each attempting to maximize additional resources (Haig and Wilczek, 2006). The maternal gametophyte is responsible for supplying nutrients to young sporophytes (Duckett *et al.*, 2009) and raises the capsule on a stalk of gametophyte tissue (Figure 1). Although nurturing sporophytes clearly presents resource demands on maternal gametophytes, there does not appear to be maternal resource limitation in this population. Though the large number of *S. macrophyllum* sporophytes (up to 18) found attached to some maternal ramets is extreme for *Sphagnum* (MG Johnson and AJ Shaw, unpublished), brood size does not seem to affect average brood fitness. It is possible that an effect may arise in other proxies for fitness not measured here, such as spore viability. Future experiments should follow the fitness consequences observed in the sporophyte to the next generation of gametophytes to determine the heritability of gametophyte fecundity.

CONCLUSIONS

We recorded, for the first time, how individual haploid genotypes (that is, gametes) vary in fertilization success and in fitness of their diploid offspring in a natural population. The identity of both maternal and paternal genotypes had significant affects on sporophyte size. We also found a relationship between the number of mates and the number of offspring, characteristic of a Bateman gradient suggesting the potential for sexual selection. Although it is unclear what traits may be selected, the extremely biased male sex ratio and the connection between gametophyte parental identity and sporophyte fitness both provide suggestive evidence that intrasexual competition for fertilization success occurs. To distinguish between natural and sexual selection, further experiments are necessary to determine the major life-history factors that determine fecundity, mating success and fitness. We have demonstrated the utility of the bryophyte system to provide unique windows into the often hidden haploid component of mating patterns in natural populations.

DATA ARCHIVING

R and Python scripts used to generate our analyses are available at github.com/mossmatters/moss_popgen_scripts, and data tables containing the multilocus genotypes of gametophytes and sporophytes are available in the Dryad data repository (http://dx.doi.org/10.5061/dryad.qj3hn). Specimen vouchers from the reproductive phenology

study, including the gametophytes selected for microsatellite genotyping, were deposited in the DUKE herbarium.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank R Johnson, M Ricca, L Clark, A Grusz and B Shaw for assistance in collecting the samples, S Boles and N Devos for help with microsatellite genotyping and C Fitzpatrick, I Liu and three anonymous reviewers for comments on earlier drafts of this manuscript. This research was supported by the Sigma-Xi Grant G200810150253 to MG Johnson and NSF Grant No. DEB-0918998 to AJ Shaw. The NSF grant was jointly awarded to AJ Shaw and B Shaw, but B Shaw was not involved in this study. B Shaw should still be listed as co-PI of this grant, which is documented here: http://www.nsf.gov/awardsearch/showAward?AWD_ID=0918998.

Anderson L, Shaw B, Shaw AJ (2009). *Peatmosses of the Southeastern United States*. In Buck WR (ed). New York Botanical Garden: NY, USA.

Anthes N, David P, Auld JR, Hoffer JNA, Jarne P, Koene JM *et al*. (2010). Bateman gradients in hermaphrodites: an extended approach to quantify sexual selection. *Am Nat* **176**: 249–263.

Arnold SJ (1994). Is there a unifying concept of sexual selection that applies to both plants and animals? *Am Nat* **144**: S1–S12.

Arnold SJ, Duvall D (1994). Animal mating systems: a synthesis based on selection theory. *Am Nat* **143**: 317–348.

Bateman AJ (1948). Intra-sexual selection in Drosophila. *Heredity* **2**: 349–368.

Becher SA, Magurran AE (2004). Multiple mating and reproductive skew in Trinidadian guppies. *Proc Biol Sci* **271**: 1009–1014.

Bisang I, Hedenäs L (2005). Sex ratio patterns in dioicous bryophytes re-visited. *J Bryol* **27**: 207–219.

Boschetto C, Gasparini C, Pilastro A (2011). Sperm number and velocity affect sperm competition success in the guppy (Poecilia reticulata). *Behav Ecol Sociobiol* **65**: 813–821.

Darwin C (1859). *The Origin of Species*. John Murray: London.

Duckett JG, Pressel S, P'ng KMY, Renzaglia KS (2009). Exploding a myth: the capsule dehiscence mechanism and the function of pseudostomata in Sphagnum. *New Phytol* **183**: 1053–1063.

Eppley SM, Taylor PJ, Jesson LK (2007). Self-fertilization in mosses: a comparison of heterozygote deficiency between species with combined versus separate sexes. *Heredity* **98**: 38–44.

Falconer DS, Mackay T (1996). *Introduction to Quantitative Genetics*. 4th edn Pearson Prentice Hall: Harlow, UK.

Field DL, Pickup M, Barrett SCH (2013). Comparative analyses of sex-ratio variation in dioecious flowering plants. *Evolution* **67**: 661–672.

Firman RC, Simmons LW (2010). Experimental evolution of sperm quality via postcopulatory sexual selection in house mice. *Evolution* **64**: 1245–1256.

Fisher RA (1930). *The Genetical Theory of Natural Selection*. Clarendon Press: Oxford.

Gasparini C, Simmons LW, Beveridge M, Evans JP (2010). Sperm swimming velocity predicts competitive fertilization success in the green swordtail Xiphophorus helleri. *PLoS One* **5**: e12146.

Griffing B (1956). Concept of general and specific combining ability in relation to diallel crossing systems. *Aust J Biol Sci* **9**: 463–493.

Gustafsson L (1986). Lifetime reproductive success and heritability: empirical support for Fisher's fundamental theorem. *Am Nat* **128**: 761–764.

Haig D, Wilczek A (2006). Sexual conflict and the alternation of haploid and diploid generations. *Philos T Roy Soc B* **361**: 335–343.

Hedenaes L, Bisang I, Korpelainen H, Cronholm B (2010). The true sex ratio in European Pseudocalliergon trifarium (Bryophyta: Amblystegiaceae) revealed by a novel molecular approach. *Biol J Linn Soc* **100**: 132–140.

Johnson MG, Shaw AJ (2015). Genetic diversity, sexual condition, and microhabitat preference determine mating patterns in Sphagnum (Sphagnaceae) peat-mosses. *Biol J Linn Soc* **115**: 96–113.

Johnson MG, Shaw B, Zhou P, Shaw AJ (2012). Genetic analysis of the peatmoss Sphagnum cribrosum (Sphagnaceae) indicates independent origins of an extreme infraspecific morphology shift. *Biol J Linn Soc* **106**: 137–153.

Korpelainen H, Bisang I, Hedenäs L, Kolehmainen J (2008). The first sex-specific molecular marker discovered in the moss Pseudocalliergon trifarium. *J Hered* **99**: 581–587.

Levitan D (1998). Does Bateman's principle apply to broadcast-spawning organisms? Egg traits influence in situ fertilization rates among congeneric sea urchins. *Evolution* **52**: 1043–1056.

Levitan D (2004). Density-dependent sexual selection in external fertilizers: variances in male and female fertilization success along the continuum from sperm limitation to sexual conflict in the sea urchin Strongylocentrotus franciscanus. *Am Nat* **164**: 298–309.

Marshall DL, Diggle PK (2001). Mechanisms of differential pollen donor performance in wild radish, Raphanus sativus (Brassicaceae). *Am J Bot* **88**: 242–257.

McQueen CB, Andrus RE (2009). *Sphagnaceae*. In: Flora of North America Editorial Committee (ed), Flora of North America North of Mexico. vol. 27. Oxford: New York, NY, USA.

Merilä J, Sheldon BC (1999). Genetic architecture of fitness and nonfitness traits: empirical patterns and development of ideas. *Heredity* **83**: 103–109.

Mulcahy D, Mulcahy G (1975). The influence of gametophytic competition on sporophytic quality in Dianthus chinensis. *Theor Appl Genet* **46**: 277–280.

Norrell TE, Jones KS, Payton AC, McDaniel SF (2014). Meiotic sex ratio variation in natural populations of Ceratodon purpureus (Ditrichaceae). *Am J Bot* **101**: 1572–1576.

Peakall R, Smouse PE (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research–an update. *Bioinformatics* **28**: 2537–2539.

Rasmussen DI (1979). Sibling clusters and genotypic frequencies. *Am Nat* **113**: 948–951.

Shaw AJ, Boles SB, Shaw B (2008a). A phylogenetic delimitation of the 'Sphagnum subsecundum complex' (Sphagnaceae, Bryophyta). *Am J Bot* **95**: 731–744.

Shaw AJ, Cao T, Wang L, Flatberg KI, Flatberg B (2008b). Genetic variation in three Chinese peat mosses (Sphagnum) based on microsatellite markers, with primer information and analysis of ascertainment bias. *Bryologist* **111**: 271–281.

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

Snow A, Spira T (1991). Pollen vigour and the potential for sexual selection in plants. *Nature* **352**: 796–797.

Stark LR, McLetchie DN, Eppley SM (2010). Sex ratios and the shy male hypothesis in the moss Bryum argenteum (Bryaceae). *Bryologist* **113**: 788–797.

Sundberg S, Rydin H (1998). Spore number in Sphagnum and its dependence on spore and capsule size. *J Byol* **20**: 1–16.

Szövényi P, Ricca M, Shaw AJ (2009). Multiple paternity and sporophytic inbreeding depression in a dioicous moss species. *Heredity* **103**: 394–403.

Tatarenkov A, Healey CIM, Grether GF, Avise JC (2008). Pronounced reproductive skew in a natural population of green swordtails, Xiphophorus helleri. *Mol Ecol* **17**: 4522–4534.

Winsor J, Peretz S, Stephenson A (2000). Pollen competition in a natural population of Cucurbita foetidissima (Cucurbitaceae). *Am J Bot* **87**: 527–532.

Zhang C, Tateishi N, Tanabe K (2010). Pollen density on the stigma affects endogenous gibberellin metabolism, seed and fruit set, and fruit quality in Pyrus pyrifolia. *J Exp Bot* **61**: 4291–4302.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)

HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-Throughput Sequencing Reads Using Target Enrichment

Author(s): Matthew G. Johnson, Elliot M. Gardner, Yang Liu, Rafael Medina, Bernard Goffinet, A. Jonathan Shaw, Nyree J. C. Zerega, and Norman J. Wickett
Source: Applications in Plant Sciences, 4(7)
Published By: Botanical Society of America
DOI: http://dx.doi.org/10.3732/apps.1600016
URL: http://www.bioone.org/doi/full/10.3732/apps.1600016

Applications in Plant Sciences

Applications in Plant Sciences 2016 4(7): 1600016

Software Note

HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment[1]

Matthew G. Johnson[2,6], Elliot M. Gardner[2,3], Yang Liu[4], Rafael Medina[4], Bernard Goffinet[4], A. Jonathan Shaw[5], Nyree J. C. Zerega[2,3], and Norman J. Wickett[2,3]

[2]Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, Illinois 60022 USA; [3]Plant Biology and Conservation, Northwestern University, 2205 Tech Drive, Evanston, Illinois 60208 USA; [4]Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, Storrs, Connecticut 06269 USA; and [5]Department of Biology, Duke University, Box 90338, Durham, North Carolina 27708 USA

- *Premise of the study:* Using sequence data generated via target enrichment for phylogenetics requires reassembly of high-throughput sequence reads into loci, presenting a number of bioinformatics challenges. We developed HybPiper as a user-friendly platform for assembly of gene regions, extraction of exon and intron sequences, and identification of paralogous gene copies. We test HybPiper using baits designed to target 333 phylogenetic markers and 125 genes of functional significance in *Artocarpus* (Moraceae).
- *Methods and Results:* HybPiper implements parallel execution of sequence assembly in three phases: read mapping, contig assembly, and target sequence extraction. The pipeline was able to recover nearly complete gene sequences for all genes in 22 species of *Artocarpus*. HybPiper also recovered more than 500 bp of nontargeted intron sequence in over half of the phylogenetic markers and identified paralogous gene copies in *Artocarpus*.
- *Conclusions:* HybPiper was designed for Linux and Mac OS X and is freely available at https://github.com/mossmatters/HybPiper.

Key words: bioinformatics; Hyb-Seq; phylogenomics; sequence assembly.

Targeted sequence capture, or target enrichment, has emerged as an efficient, cost-effective method for generating phylogenomic data sets for nonmodel organisms (Cronn et al., 2012). The procedure works by reducing genomic DNA complexity through the use of short (80 to 120 nucleotide) bait sequences that hybridize with template sequences. By selectively retaining only genomic fragments bound to baits, high-throughput sequencing libraries are enriched for target sequences. Many samples may be multiplexed and sequenced together, and target enrichment has the potential to generate DNA sequence for hundreds of loci and dozens of samples simultaneously. The methods for generating enriched libraries have been extensively described elsewhere (e.g., Gnirke et al., 2009; Mamanova et al., 2010).

[1]Manuscript received 10 February 2016; revision accepted 1 June 2016.

We would like to thank A. DeVault at MycroArray for assistance optimizing the target enrichment protocol, and the Field Museum for use of its DNA sequencers. The authors thank B. Faircloth and two anonymous reviewers for helpful comments on an earlier version of the manuscript. This research was funded by National Science Foundation grants to A.J.S. (DEB-1239980), B.G. (DEB-1240045 and DEB-1146295), N.J.W. (DEB-1239992), and N.J.C.Z. (DEB-0919119), and by a grant from the Northwestern University Institute for Sustainability and Energy (N.J.C.Z.). Data generated for this study can be found at www.artocarpusresearch.org, www.datadryad.org (http://dx.doi.org/10.5061/dryad.3293r), and the NCBI Sequence Read Archive (SRA; BioProject PRJNA301299).

[6]Author for correspondence: mjohnson@chicagobotanic.org

doi:10.3732/apps.1600016

Several recent papers have demonstrated the efficacy of target enrichment to resolve relationships in a variety of organisms (Mandel et al., 2014; Mariac et al., 2014; Bragg et al., 2015). In one strategy, ultra-conserved elements can be used to anchor baits in slow-evolving portions of the genome, and analysis is focused on more variable flanking regions (Faircloth et al., 2012; Lemmon et al., 2012). Another approach is to focus on exon sequences, because reference sequences across phylogenetic scales can be efficiently generated using transcriptome sequencing (Bi et al., 2012; Hugall et al., 2016). An extension of this approach is Hyb-Seq (Weitemier et al., 2014), which combines exon capture with genome skimming of a "splash zone"—intronic and intergenic regions that flank target exons, potentially of use for shallower phylogenetic applications.

In previously published studies using Hyb-Seq data, three main bioinformatics issues have arisen: (1) how to efficiently sort high-throughput sequencing reads into separate loci (e.g., Stull et al., 2013), (2) how to assemble sequences at each locus that can be aligned for phylogenetic inference (e.g., Stephens et al., 2015), and (3) how to extend sequence recovery beyond the coding sequence into the more variable intron regions (e.g., Folk et al., 2015). Data need to be handled in an efficient, streamlined manner because many Hyb-Seq projects involve dozens or hundreds of samples.

We developed HybPiper to efficiently turn sequencing reads generated by the Hyb-Seq method into organized gene files ready for phylogenetic analysis. HybPiper is a suite of Python scripts that wrap and organize bioinformatics tools for target sequence extraction from high-throughput sequencing reads. The primary

output of the pipeline is a nucleotide and translated amino acid sequence for every gene that can be assembled from the sequencing reads. HybPiper also includes several postprocessing scripts for retrieving sequences from multiple samples run through the pipeline, visualization of summary statistics such as recovery efficiency and coverage depth, and extraction of flanking intron sequences. We designed the pipeline to be easy-to-use in a modular design that allows the user to rerun portions of the pipeline to adapt parameter settings (i.e., *E*-value thresholds, assembly coverage cutoffs, or percent identity filters) for individual samples as needed. Although other bioinformatics pipelines are available to process target enrichment data, such as PHYLUCE (Faircloth, 2015) and alignreads.py (Straub et al., 2011), HybPiper is designed specifically for the Hyb-Seq approach: targeting exons and flanking intron regions.

METHODS AND RESULTS

***Input data*—**Here, we demonstrate the utility of HybPiper using 22 species of *Artocarpus* J. R. Forst. & G. Forst. (Moraceae) and six outgroups. Sequencing libraries were hybridized to a bait set comprising 458 target nuclear coding regions. We describe the development of bait sequences from a draft genome sequence in a companion paper (Gardner et al., 2016). Briefly, 333 loci intended for phylogenetic analysis were selected by identifying long exons homologous between the *Artocarpus* draft genome and the published genome of *Morus notabilis* C. K. Schneid. (Moraceae) (He et al., 2013). For the "phylogenetic genes," the bait set included 120-bp baits designed from both *Artocarpus* and *Morus* sequences. A set of 125 additional genes were targeted for their functional significance: 98 MADS-box genes and 27 genes that have been implicated in floral volatiles. For the genes of functional significance, baits were designed from the *A. camansi* Blanco draft genome alone (Gardner et al., 2016). A set of 20,000 baits (biotinylated RNA oligonucleotides, the smallest MYbaits kit) with 3× tiling was manufactured by MYcroarray (Ann Arbor, Michigan, USA).

Sequencing libraries for 22 *Artocarpus* species and two of the outgroups (Table 1) were prepared with the Illumina TruSeq Nano HT DNA Library Preparation Kit (Illumina, San Diego, California, USA) following the manufacturer's protocol, with a target mean insert size of 550 bp. Libraries were hybridized to the bait set in four pools of six libraries each at 65°C for approximately 18 h, following the manufacturer's protocol. The enriched libraries were reamplified with 14 PCR cycles. The four pools of enriched libraries were sequenced together in a single flow cell of Illumina MiSeq (600 cycle, version 3 chemistry). This run produced 9,503,831 pairs of 300-bp reads. Four additional outgroup libraries generated in a separate hybridization were sequenced as part of a separate run and generated an additional 3,716,390 pairs of reads. Demultiplexed and adapter-trimmed reads (cleaned automatically by Illumina BaseSpace) were quality trimmed using Trimmomatic (Bolger et al., 2014), with a quality cutoff of 20 in a 4-bp sliding window, discarding any reads trimmed to under 30 bp. Only pairs with both mates surviving were used for HybPiper. An average of 391,505 reads per sample survived the trimming process across all 28 samples. The reads have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (BioProject PRJNA301299).

The inputs of HybPiper are the read file (or files, for paired-end data) and a curated "target file." HybPiper is built to operate at the locus level; if target sequences were designed from multiple exons within the same gene, these may be concatenated (with no gaps or intervening sequence) to generate a single coding sequence for each gene. This allows HybPiper to detect intron sequences during coding sequence extraction. In the case of the *Artocarpus*/*Morus* bait set, two orthologous sequences are retrieved for most loci. The presence of multiple sequences for each locus is specified in the sequence IDs within the target file; for example, "*Artocarpus*-g001" and "*Morus*-g001" indicate to HybPiper that both sequences represent locus g001. Phase 1 of HybPiper, in part, determines which sequence is the more appropriate reference sequence for each gene and sample separately. This flexibility allows the use of the same target file for samples that span a wide range of phylogenetic distances. Additional orthologous sequences for each gene may be added to the target file as desired by the user, which may increase the efficiency of sorting reads and generate new orthologous loci for phylogenetics.

***Phase 1: Sorting sequencing reads by target gene*—**Target enrichment is typically conducted on multiple samples that have been pooled during bait hybridization and sequencing. HybPiper maps reads against the target genes for each sample separately. This is a different procedure than several other target enrichment analysis pipelines (Straub et al., 2011; Bi et al., 2012; Faircloth, 2015), which typically begin with de novo assembly for each sample, and then attempt to match contigs to target loci. In HybPiper, reads are first sorted based on whether they map to a target locus. We explored two methods for aligning reads to the targets: (1) BLASTX (Camacho et al., 2009), which uses peptide sequences as a reference, and (2) BWA (Li and Durbin, 2009), which uses nucleotide sequences. In principle, the BLASTX approach should be more forgiving to substitutions between the target sequence and sample reads, because alignments are conducted at the peptide level and may detect similarity between more distant sequences than BWA. The BWA approach may result in fewer overall reads mapping to a distantly related target sequence, but is several times faster than the BLASTX method.

HybPiper sorts reads into separate directories for each gene using Biopython (Cock et al., 2009) to efficiently parse the FASTA format. In our tests of the BLASTX method, we set an *E*-value threshold of 1×10^{-5} to accept alignments, but the user can change this. For the BWA method, all alignable reads are sorted into each gene directory using a Python wrapper around SAMtools (Li et al., 2009). We calculate the enrichment efficiency as the percentage of trimmed, filtered reads that were sorted into a gene directory.

For the *Artocarpus* reads, an average of 71.9% of reads were on target (range 64.4–79.9%), based on the BLASTX method. Enrichment efficiency was lower for some of the outgroup samples, which ranged from just 5.0% for *Antiaropsis* K. Schum. to 71.6% for *Ficus* L. To address whether the presence of duplicate reads affects our estimate of enrichment efficiency, we removed paired duplicate reads using SuperDeduper (http://dstreett.github.io/Super-Deduper/). Most samples had between 6% and 18% duplicate read pairs, and a similar percentage of the duplicate read pairs mapped to the target loci (Appendix S1). One outlier was *Ficus*, which had 34% duplicate reads, 42% of which mapped to targets. After adjusting for duplicate reads, our estimates of enrichment efficiency were reduced by about 4% on average (Table 1). Removing duplicate reads did not affect the extraction of exon sequences in HybPiper for this data set.

The phylogenetic distance to *Artocarpus* did not seem related to percent enrichment. However, the two outgroup samples that were pooled in a hybridization with *Artocarpus* in the first sequencing run had much lower enrichment efficiency than ingroup samples (Table 1). This suggests that multiplexing at the hybridization stage should be nonrandom, and only libraries of taxa that are relatively equidistant from the taxa used to design the bait sequences should be pooled. This strategy has been previously recommended in other studies (McGee et al., 2016).

***Phase 2: Sequence assembly*—**Some previous methods for target enrichment assembly have used mapping-based approaches to reassemble target loci (Straub et al., 2011; Hugall et al., 2016), which may be inefficient when there is high sequence divergence between the sample reads and the target reference. HybPiper instead conducts a de novo assembly for each gene separately; reads are assembled into contiguous sequences (contigs) using SPAdes (Bankevich et al., 2012). Multiple contigs may be assembled per gene, due to incomplete sequencing of intron sequences, paralogous gene sequences, or alleles. Additional contigs may be assembled from weakly aligning reads with low identity to the target sequences. These contigs are sorted by sequencing depth and are aligned to the reference protein sequence in the next phase. HybPiper decreases the amount of time needed for assembly and alignment stages by using GNU Parallel, a tool for executing commands simultaneously, using the multiple threads available on modern processors (Tange, 2011).

***Phase 3: Alignment of exons*—**In 333 of 458 genes in our test data set, baits were designed from homologous sequences in the *A. camansi* draft genome and the previously published *M. notabilis* genome. For each gene, HybPiper decides whether the *Artocarpus* or *Morus* target sequence should serve as the reference by tallying all alignment scores from reads aligned to the gene during Phase 1. In the BWA version of the alignment, the "mapping score" is tallied for all reads mapping to the target gene; for the BLASTX method, the bit score is used.

Target enrichment is generally carried out using genomic DNA; however, bait sequences are often designed from only the coding regions of a target, such as assembled transcripts. To extract the coding sequence portion of the assembled contigs that likely contain partial intron sequence, HybPiper uses Exonerate (Slater and Birney, 2005). For each target, assembled contigs are aligned to target peptide sequences using the "protein2genome" alignment model. If the

Applications in Plant Sciences 2016 4(7): 1600016
doi:10.3732/apps.1600016

TABLE 1. Sample information, sequencing run and hybridization pool, summary of sequencing, and target enrichment results for the *Artocarpus*/*Morus* bait set.

Sample ID	Species	Run	Pool	Paired reads	Paired surviving QC	Percent reads on target	Genes recovered	Subgenus/Tribe[a]
NZ866	*Artocarpus odoratissimus* Blanco	1	4	237,638	212,318	70.1	456	*Artocarpus*
NZ728	*Artocarpus rigidus* Blume	1	1	657,701	592,305	75.2	458	*Artocarpus*
NZ739	*Artocarpus lanceifolius* Roxb.	1	2	410,273	343,182	73.8	456	*Artocarpus*
NZ606	*Artocarpus anisophyllus* Miq.	1	3	507,744	456,512	65.3	457	*Artocarpus*
NZ814	*Artocarpus brevipedunculatus* (F. M. Jarrett) C. C. Berg	1	1	757,804	697,873	76.7	458	*Artocarpus*
NZ612	*Artocarpus kemando* Miq.	1	3	590,801	502,324	68.1	458	*Artocarpus*
EG92	*Artocarpus tamaran* Becc.	1	3	422,739	383,077	68.4	458	*Artocarpus*
EG87	*Artocarpus elasticus* Reinw. ex Blume	1	4	508,620	456,063	72.6	458	*Artocarpus*
NZ771	*Artocarpus sericicarpus* F. M. Jarrett	1	4	437,596	368,357	71.7	458	*Artocarpus*
NZ946	*Artocarpus teysmannii* Miq.	1	3	409,715	379,410	64.7	457	*Artocarpus*
MW_lowii-2	*Artocarpus lowii* King	1	4	417,260	350,643	72.4	458	*Artocarpus*
NZ780	*Artocarpus excelsus* F. M. Jarrett	1	3	328,567	291,565	64.7	458	*Artocarpus*
NZ918	*Artocarpus integer* Merr.	1	3	296,053	273,231	64.4	457	*Cauliflori*
EG98	*Artocarpus heterophyllus* Lam.	1	2	634,153	523,372	75.5	458	*Cauliflori*
NZ694	*Artocarpus peltatus* Merr.	1	4	444,734	369,717	72.4	458	*Pseudojaca*
NZ687	*Artocarpus primackii* Kochummen	1	2	353,376	316,194	75.7	458	*Pseudojaca*
NZ420	*Artocarpus lacucha* Roxb. ex Buch.-Ham.	1	2	425,368	386,539	77.9	457	*Pseudojaca*
NZ911	*Artocarpus nitidus* Trécul	1	4	403,279	340,166	72.3	457	*Pseudojaca*
NZ402	*Artocarpus thailandicus* C. C. Berg	1	1	208,369	188,696	77.5	458	*Pseudojaca*
NZ929	*Artocarpus fretessii* Tiejsm. & Binn. ex Hassk.	1	2	385,183	345,495	76.0	458	*Pseudojaca*
GW1701	*Artocarpus sepicanus* Diels	1	2	146,460	129,755	74.4	458	[*Artocarpus*]
NZ609	*Artocarpus limpato* Miq.	1	1	520,398	478,292	72.8	457	*Prainea*
NZ281	*Antiaropsis decipiens* K. Schum.	2	1	1,122,018	866,706	5.0	380	Castilleae
EG139	*Maclura pomifera* (Raf.) C. K. Schneid.	2	4	441,498	236,834	56.7	417	Maclureae
EG78	*Streblus glaber* Corner	2	1	484,680	297,401	23.9	392	Moreae
EG30	*Ficus macrophylla* Desf. ex Pers.	2	4	1,522,294	1,047,142	71.6	423	Ficeae
NZ311	*Dorstenia hildebrandtii* Engl.	1	1	91,831	83,447	16.3	294	Dorstenieae
NZ874	*Parartocarpus venenosus* Becc.	1	1	54,069	45,534	31.4	378	[unnamed tribe-level clade]

[a]Brackets indicate subgenera or tribes with uncertain taxonomic designation.

BWA method was used for alignment, the peptide sequence is generated by direct translation using Biopython. Sample sequences homologous to the reference coding sequence are extracted in FASTA format with a customized header specifying alignment start and end locations and percent identity between the sample and reference, using the "roll your own" feature in Exonerate.

Within HybPiper, Exonerate (Slater and Birney, 2005) is used to extract likely coding sequences (introns removed) aligned to the reference protein sequence. These alignments must be nonoverlapping and exceed a percent identity threshold (default: 60%) between the contig and the protein sequence. Alignments are sorted by position (relative to the reference sequence), and the longest contig that does not overlap with other contigs is retained. However, if the overlap between the ends of two contig alignments is less than 20 bp, both contigs are retained. This is to reduce errors in alignment at the ends of exons. Any contigs with slightly overlapping ranges are concatenated into a "supercontig" and a second Exonerate analysis is conducted to detect the true intron-exon junctions. At this stage, the coding sequence that aligns to the reference amino acid sequence and statistics about the contigs retained are saved into the gene sequence directory.

***Identification of paralogous sequences, alleles, or contaminants*—**In many target enrichment analysis pipelines, correct orthology of enriched sequences is inferred using BLAST searches to the target proteome (e.g., Bi et al., 2012; Bragg et al., 2015), but this method will be inefficient when genomic resources in the target taxa are incomplete. In HybPiper we provide a streamlined method for identification of potential paralogs that can be further analyzed using gene phylogenies. Typically, if HybPiper identifies a single contig that subsumes the range of other contigs, it is retained and the smaller contigs are discarded. However, sequences assembled using SPAdes occasionally result in multiple, long contigs, each representing the entire target sequence. During the extraction of exon sequences, the HybPiper script exonerate_hits.py identifies contigs that span more than 85% of the length of the reference sequence. HybPiper will generate a warning that indicates multiple long-length matches to the reference sequence have been found. HybPiper chooses among multiple full-length contigs by first using a sequencing coverage depth cutoff—if one contig has a coverage depth 10 times (by default) greater than the next best full-length contig, it is chosen. If the sequencing depth is similar among all full-length contigs, the percent identity with the reference sequence is used as the criterion. Genes for which multiple long-length sequences exist should be examined further to detect whether they represent paralogous genes, alleles, or contaminants. We discuss identification of paralogs in more detail below (see "Separating paralogous gene copies in *Artocarpus*").

***Extraction of flanking intron sequences*—**Following the identification of exon sequences in the assembled contigs, HybPiper attempts to identify intron regions flanking the exons using the script "intronerate.py." This is done by rerunning Exonerate on the supercontigs used in Phase 3 and retaining the entire gene sequence, rather than just the exon sequence. Even if the entire intron was not recovered, Exonerate can detect the presence of splice junctions based on the alignment of the supercontig to the reference protein sequence. HybPiper generates an annotation of the supercontig in genomic feature format (GFF). The annotations are sorted and filtered in the same manner as the exon sequences during Phase 3 (alignment of exons) of the main HybPiper script, to remove overlapping annotations. Following the intron annotation, HybPiper also produces two additional sequence files at each locus: (1) the supercontig and (2) only the intron sequences (exons removed from supercontig).

***Postprocessing: Visualization of recovery efficiency*—**After running HybPiper on multiple samples, we have provided a series of helper scripts to collect and visualize summary statistics across samples. Running these scripts from a directory containing all of the output of each sample takes advantage of the standardized directory structure generated by HybPiper. The script get_gene_lengths.py will summarize the length of the sequences recovered and will return a warning if a sequence is more than 50% longer than the corresponding target sequence. This file can be used as the input for an R script included with HybPiper (gene_recovery_heatmap.R) to visualize the recovery efficiency using a heat map (Fig. 1).

In the heat map, each row represents a sample, and each column represents a target. Each cell is shaded based on the length of the sequence recovered by HybPiper for that gene, as a percentage of the length of the reference sequence. The heat map can be used as a first glance at the efficiency of target recovery and help identify difficult-to-recover loci (columns with lighter shading) and low-enrichment samples (rows with lighter shading).

For the *Artocarpus* data set, we were able to recover the vast majority of loci for in-group samples (Table 2). Using 10 processors on a Linux computer,

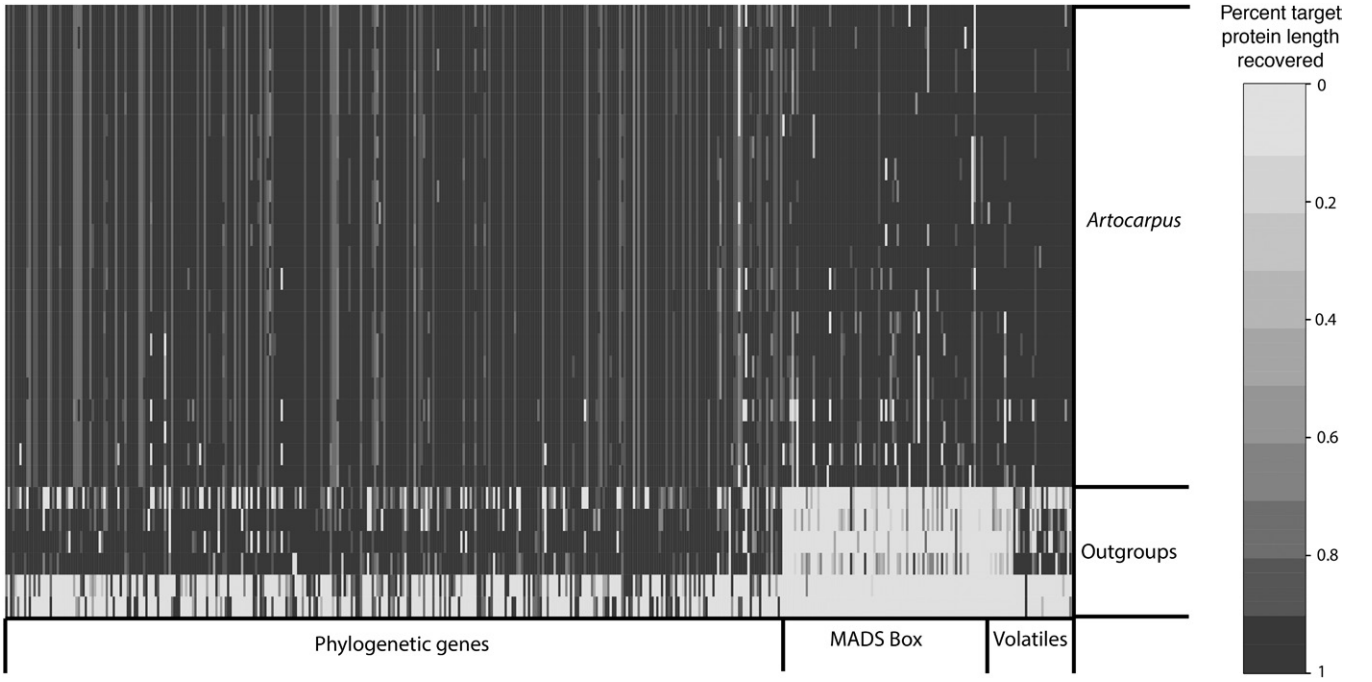

Fig. 1. Heat map showing recovery efficiency for 458 genes enriched in *Artocarpus* and other Moraceae and recovered by HybPiper using the BWA method. Each column is a gene, and each row is one sample. The shade of gray in the cell is determined by the length of sequence recovered by the pipeline, divided by the length of the reference gene (maximum of 1.0). Three types of genes were enriched: phylogenes (left), MADS-box genes (center), and volatiles (right) for 22 *Artocarpus* samples (top) and six outgroup species (bottom). Full data for this chart can be found in Appendices S2 and S3.

TABLE 2. Recovery efficiency of HybPiper for 22 *Artocarpus* and six other Moraceae, using two methods for assigning reads to loci—BLASTX (mapping to protein sequences) and BWA (mapping to nucleotide sequences).

		BLASTX method			BWA method		
Taxon	*N*	Phylogenetic loci	MADS box	Volatiles	Phylogenetic loci	MADS box	Volatiles
Total genes in array		333	98	27	333	98	27
Artocarpus subg. *Artocarpus*	12	329.6	96.4	26.4	331	94	26.5
Artocarpus subg. *Cauliflori*	2	328.5	96.5	26.5	330.5	93	26.5
Artocarpus subg. *Pseudojaca*	6	326.7	95.3	26.2	329.8	89.3	26.2
Artocarpus (other)	2	326	95	26	327	87.5	25.5
Antiaropsis	1	257	41	13	259	8	10
Maclura	1	307	53	21	299	10	19
Streblus	1	318	30	17	315	3	17
Ficus	1	315	56	25	311	17	20
Dorstenia	1	135	15	2	118	0	1
Parartocarpus	1	127	21	4	120	0	3

Note: *N* = number of individuals sampled.

HybPiper completed in about 9 h using the BWA method and 24 h using BLASTX for all 28 samples. For all samples, including outgroups, HybPiper was able to recover the 333 "phylogenetic genes" with more efficiency than the MADS-box or "volatile" genes. For instance, HybPiper recovered 93% (311 of 333) of the phylogenetic loci for *Ficus*, but just 30% (37/125) of the MADS-box and volatile loci using the BWA method (Table 2). The most likely explanation for this is that the phylogenetic loci had bait sequences derived from two different sources (*Artocarpus* baits and *Morus* baits). The remaining 125 loci had baits designed only from *Artocarpus* draft genome sequence. Substitutions between the *Artocarpus* sequences and our outgroup samples may have reduced the hybridization efficiency for the MADS-box and volatile genes. The dissimilarity between target sequence and sample sequences was compounded by using the BWA method, which allows for less dissimilarity than the BLASTX method for aligning reads (Table 2). Additional information about the recovery of loci using the BLASTX and BWA methods can be found in the supplemental material (Appendices S2 and S3).

Alternatively, the increased hybridization efficiency for the phylogenetic genes may be the result of twice as many bait sequences for each target. The pooling strategy, during hybridization and sequencing, may have also had an effect. In our first sequencing run, the two outgroup species had about one third of the average number of reads, and about one tenth the average number of reads on target, compared to the *Artocarpus* species. Therefore, pooling outgroup samples together in separate hybridizations may be advisable in future Hyb-Seq analyses.

***Recovery of intron sequences in* Artocarpus—**The utility of phylogenomic approaches is maximized when the resolution at individual loci is sufficient for the phylogenetic depth of the analysis (Salichos and Rokas, 2013). This is especially true for "species tree" methods that require gene tree reconstructions as input (Mirarab et al., 2014). For adaptive radiations and species complexes, ultraconserved elements may not be variable enough to resolve internodes with only a few parsimony informative characters per locus (Smith et al., 2013; e.g., Giarla and Esselstyn, 2015; Manthey et al., 2016). Newer sequencing technologies, such as the 2 × 300 (paired-end) MiSeq chemistry from Illumina, produce reads that are longer than the typical bait length, resulting in sequence fragments containing pieces of exon (i.e., on-target) that may also extend hundreds of base pairs into intron or intergenic regions. This is an attractive solution because the same bait sets designed for deeper phylogenetic questions, where exon variability may be sufficient, could also be used at shallower scales.

We explored the capture efficiency of intron regions in the *Artocarpus* bait set by aligning reads of *Artocarpus* samples to the reference genome scaffolds using BWA. Two patterns emerged: for short introns (<500 bp), little difference was detected in the depth of coverage between exons and introns (Fig. 2). For longer introns, depth steadily decreased but was typically still above 10× up to 500 bp away from the end of the exon (Fig. 2), even with duplicate reads removed (Appendix S4). This suggests that long-read technologies such as MiSeq 2 × 300 paired-end sequencing are well-suited for recovering intronic regions using Hyb-Seq.

Recent studies have explored the feasibility of using introns extracted from Hyb-Seq (Folk et al., 2015; Brandley et al., 2015), but have not presented an automated method for extracting intron sequence from capture data. HybPiper can generate "supercontigs" containing all assembled contigs (exon and intron sequence) for a gene. This sequence is annotated in genome feature format (GFF), which can be used to extract intronic and/or intergenic regions into separate files for alignment and analysis. We observe the greatest reliability of extracting intronic regions by generating multiple sequence alignments of the supercontigs from multiple samples. The exon regions serve as an anchor for the alignment, and extraneous sequence that appears in only one or a few sequences can be trimmed by downstream analysis, such as trimAl (Capella-Gutierrez et al., 2009).

For the 333 loci developed as phylogenetic markers, extracting intron data vs. exon data alone using HybPiper increased the average length of the loci from 1135 bp (range 201–3171 bp) to 1784 bp (range 528–4267 bp) (Appendix S5). When all samples were aligned using MAFFT and trimmed using trimAl, the total alignment length increased to 594,149 bp and added 138,982 characters to the concatenated alignment. The number of parsimony informative characters within *Artocarpus* increased from 35,935 using exons only to 138,932 using supercontigs. Intron sequence recovery efficiency was variable across the loci; for 54 loci the full alignment length was within 100 bp of the exon-only alignment. In contrast, 172 loci had a final alignment length 500 bp or longer than that of exons alone.

***Separating paralogous gene copies in* Artocarpus—**The genus *Artocarpus* has undergone at least one whole genome duplication since its divergence from the rest of the Moraceae (Gardner et al., 2016). As a result, many genes that appear single copy in the *Morus* reference genome are multicopy in *Artocarpus*. HybPiper identified paralogous gene copies in *Artocarpus* for 123 of our 333 phylogenetic loci (Appendix S6). For these genes, multiple full-length contigs were assembled with similar sequencing coverage. To investigate the paralagous copies further, we extracted the paralogous exon sequence from each contig using two scripts included in HybPiper: paralog_investigator.py identifies whether multiple contigs that are at least 85% of the target reference length are present and flags these as possible paralogs. It then extracts exon sequences from putative paralogs using exonerate_hits.py. A second script, paralog_retriever.py, collects the inferred paralogous sequences across many samples for one gene. If no paralogs are identified for a sample, the coding sequence extracted during Phase 3 of the main script is included.

We generated gene family phylogenies for several genes where multiple copies were identified in *Artocarpus*. Nucleotide sequences were aligned using MAFFT (Katoh and Standley, 2013) (--localpair --maxiterate 1000) and phylogenies were reconstructed using RAxML using a GTRGAMMA substitution model and 200 "fast-bootstrap" pseudoreplicates. In each case, two separate clades of *Artocarpus* samples are observed (Appendix S7), indicating that the multiple full-length contigs likely result from the paleopolyploidy event, and that they can be distinguished by HybPiper. For further phylogenetic analysis, the paralog with the highest percent identity to the *Artocarpus camansi* reference genome sequence was selected for each sample, because this paralog represents the closest homology to the sequence from which the baits were designed.

***Phylogeny reconstruction*—**After running HybPiper on multiple samples, multisequence FASTA files can be generated for each gene using a script included with the pipeline (retrieve_sequences.py). After aligning each gene separately with MAFFT, we reconstructed phylogenies with two nucleotide supermatrix data sets, (1) the full matrix and (2) the exons alone, in RAxML

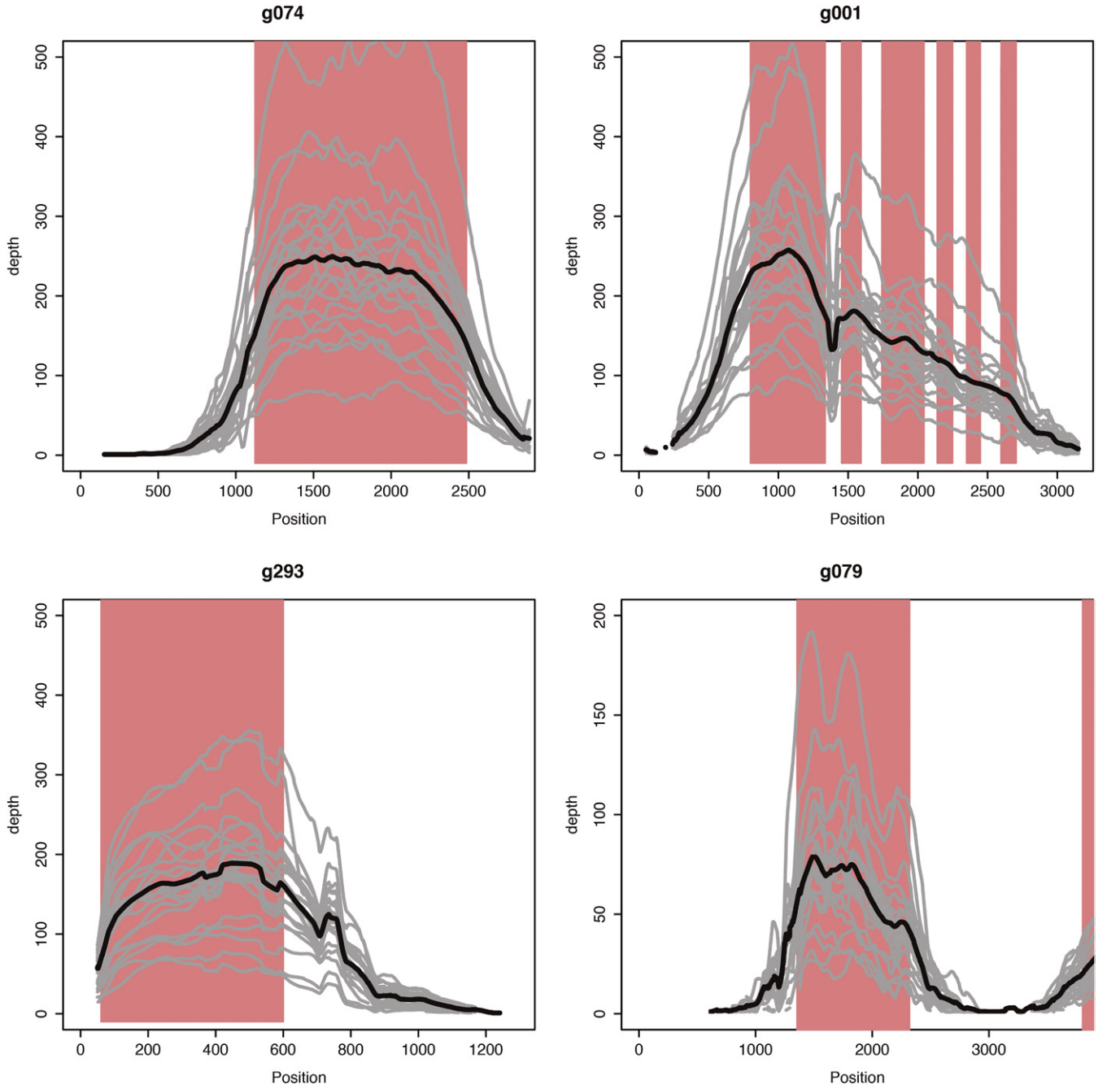

Fig. 2. Depth-of-coverage plots for four exemplar loci based on reads aligned to the *Artocarpus camansi* draft genome. Each gray line represents a rolling average depth across 50 bp for one of 22 *Artocarpus* species. The dark line represents the average depth of coverage. Red bars indicate the location of exon boundaries predicted in the *Artocarpus camansi* draft genome.

(Stamatakis, 2014) using a separate GTRGAMMA partition for each locus, along with 200 "fast-bootstrap" pseudoreplicates.

The phylogenies are largely similar in topology and level of support (Appendix S8), particularly for most subgenera circumscribed by Zerega et al. (2010). Rearrangements or changes in bootstrap support are occasional and minor. Only one rearrangement was characterized by high support in both positions (*A. limpato* Miq.). Although the phylogenies are for the most part well resolved, they should be treated with caution due to the low taxon sampling (only 22 out of ca. 70 species). We recognize here that using a supermatrix approach alone is insufficient to fully understand the influence of conflicting gene-tree signal due to processes such as incomplete lineage sorting. We present, however, this phylogeny simply as an example of one possible analysis that can be performed using the output of HybPiper. The output files generated by HybPiper are suitable for use with whichever phylogeny reconstruction method is favored by the user.

CONCLUSIONS

HybPiper can be used to efficiently assemble gene regions from enriched sequencing libraries designed using the Hyb-Seq method, extract exon and intron sequences, and assemble sequence data that are ready to use in phylogenetic analysis. The

pipeline is flexible and modular, and can be adapted to analyses at deep phylogenetic depths (by using the BLASTX method) or within genera (by incorporating intron sequence). The pipeline has been tested on Linux and Mac OS X, and is freely available under a GPLv3 license at: https://github.com/mossmatters/HybPiper.

LITERATURE CITED

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, et al. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19: 455–477.

Bi, K., D. Vanderpool, S. Singhal, T. Linderoth, C. Moritz, and J. M. Good. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics* 13: 403.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* 30: 2114–2120.

Bragg, J. G., S. Potter, K. Bi, and C. Moritz. 2015. Exon capture phylogenomics: Efficacy across scales of divergence. *Molecular Ecology Resources* doi:10.1111/1755-0998.12449.

Brandley, M. C., J. G. Bragg, S. Singhal, D. G. Chapple, C. K. Jennings, A. R. Lemmon, E. M. Lemmon, et al. 2015. Evaluating the performance of anchored hybrid enrichment at the tips of the tree of life: A phylogenetic analysis of Australian *Eugongylus* group scincid lizards. *BMC Evolutionary Biology* 15: 62.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10: 421.

Capella-Gutierrez, S., J. M. Silla-Martinez, and T. Gabaldon. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)* 25: 1972–1973.

Cock, P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, et al. 2009. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)* 25: 1422–1423.

Cronn, R., B. J. Knaus, A. Liston, P. J. Maughan, M. Parks, J. V. Syring, and J. Udall. 2012. Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany* 99: 291–311.

Faircloth, B. C. 2015. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics (Oxford, England)* 32: 786–788.

Faircloth, B. C., J. E. McCormack, N. G. Crawford, M. G. Harvey, R. T. Brumfield, and T. C. Glenn. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* 61: 717–726.

Folk, R. A., J. R. Mandel, and J. V. Freudenstein. 2015. A protocol for targeted enrichment of intron-containing sequence markers for recent radiations: A phylogenomic example from *Heuchera* (Saxifragaceae). *Applications in Plant Sciences* 3(8): 1500039.

Gardner, E. M., M. G. Johnson, D. Ragone, N. J. Wickett, and N. J. C. Zerega. 2016. Low-coverage, whole-genome sequencing of *Artocarpus camansi* (Moraceae) for phylogenetic marker development and gene discovery. *Applications in Plant Sciences* 4(7): 1600017.

Giarla, T. C., and J. A. Esselstyn. 2015. The challenges of resolving a rapid, recent radiation: Empirical and simulated phylogenomics of Philippine shrews. *Systematic Biology* 64: 727–740.

Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology* 27: 182–189.

He, N., C. Zhang, X. Qi, S. Zhao, Y. Tao, G. Yang, T.-H. Lee, et al. 2013. Draft genome sequence of the mulberry tree *Morus notabilis. Nature Communications* 4: 2445.

Hugall, A. F., T. D. O'Hara, S. Hunjan, R. Nilsen, and A. Moussalli. 2016. An exon-capture system for the entire Class Ophiuroidea. *Molecular Biology and Evolution* 33: 281–294.

Katoh, K., and D. M. Standley. 2013. MAFFT: Multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Lemmon, A. R., S. A. Emme, and E. M. Lemmon. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* 61: 727–744.

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

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 25: 2078–2079.

Mamanova, L., A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, et al. 2010. Target-enrichment strategies for next-generation sequencing. *Nature Methods* 7: 111–118.

Mandel, J. R., R. B. Dikow, V. A. Funk, R. R. Masalia, S. E. Staton, A. Kozik, L. Rieseberg, and J. M. Burke. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. *Applications in Plant Sciences* 2(2): 1300085.

Manthey, J. D., L. C. Campillo, K. J. Burns, and R. G. Moyle. 2016. Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: A test in cardinalid tanagers (Aves, Genus: *Piranga*). *Systematic Biology* doi:10.1093/sysbio/syw005.

Mariac, C., N. Scarcelli, J. Pouzadou, A. Barnaud, C. Billot, A. Faye, A. Kougbeadjo, et al. 2014. Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. *Molecular Ecology Resources* 14: 1103–1113.

McGee, M. D., B. C. Faircloth, S. R. Borstein, J. Zheng, C. D. Hulsey, P. C. Wainwright, and M. E. Alfaro. 2016. Replicated divergence in cichlid radiations mirrors a major vertebrate innovation. *Proceedings of the Royal Society B, Biological Sciences* 283: 20151413.

Mirarab, S., R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014. ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics (Oxford, England)* 30: i541–i548.

Salichos, L., and A. Rokas. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497: 327–331.

Slater, G., and E. Birney. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6: 31.

Smith, B. T., M. G. Harvey, B. C. Faircloth, T. C. Glenn, and R. T. Brumfield. 2013. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Systematic Biology* 63: 83–95.

Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)* 30: 1312–1313.

Stephens, J. D., W. L. Rogers, C. M. Mason, L. A. Donovan, and R. L. Malmberg. 2015. Species tree estimation of diploid *Helianthus* (Asteraceae) using target enrichment. *American Journal of Botany* 102: 910–920.

Straub, S. C., M. Fishbein, T. Livshultz, Z. Foster, M. Parks, K. Weitemier, R. C. Cronn, and A. Liston. 2011. Building a model: Developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* 12: 211.

Stull, G. W., M. J. Moore, V. S. Mandala, N. A. Douglas, H.-R. Kates, X. Qi, S. F. Brockington, et al. 2013. A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Applications in Plant Sciences* 1(2): 1200497.

Tange, O. 2011. GNU Parallel: The Command-Line Power Tool. *USENIX Magazine* 36: 42–47.

Weitemier, K., S. Straub, R. C. Cronn, and M. Fishbein. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2(9): 1400042.

Zerega, N. J. C., M. N. Nur Supardi, and T. J. Motley. 2010. Phylogeny and recircumscription of Artocarpeae (Moraceae) with a focus on *Artocarpus. Systematic Botany* 35: 766–782.

Molecular Phylogenetics and Evolution 98 (2016) 29–40

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

A phylotranscriptomic analysis of gene family expansion and evolution in the largest order of pleurocarpous mosses (Hypnales, Bryophyta) ☆

Matthew G. Johnson [a,*], Claire Malley [b], Bernard Goffinet [c], A. Jonathan Shaw [d], Norman J. Wickett [a,b,*]

[a] Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, IL 60022, United States
[b] Program in Biological Sciences, Northwestern University, 2205 Tech Drive, O.T. Hogan Hall, Room 2-144, Evanston, IL 60208, United States
[c] Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Rd., Storrs, CT 06269, United States
[d] Department of Biology, Duke University, Box 90338, Durham, NC 27708, United States

ARTICLE INFO

Article history:
Received 8 October 2015
Revised 7 January 2016
Accepted 11 January 2016
Available online 23 January 2016

Keywords:
Bryophyte
Phylogenomics
Rapid radiation
Orthology

ABSTRACT

The pleurocarpous mosses (i.e., Hypnanae) are a species-rich group of land plants comprising about 6,000 species that share the development of female sex organs on short lateral branches, a derived trait within mosses. Many of the families within Hypnales, the largest order of pleurocarpous mosses, trace their origin to a rapid radiation less than 100 million years ago, just after the rise of the angiosperms. As a result, the phylogenetic resolution among families of Hypnales, necessary to test evolutionary hypotheses, has proven difficult using one or few loci. We present the first phylogenetic inference from high-throughput sequence data (transcriptome sequences) for pleurocarpous mosses. To test hypotheses of gene family evolution, we built a species tree of 21 pleurocarpous and six acrocarpous mosses using over one million sites from 659 orthologous genes. We used the species tree to investigate the genomic consequences of the shift to pleurocarpy and to identify whether patterns common to other plant radiations (gene family expansion, whole genome duplication, or changes in the molecular signatures of selection) could be observed. We found that roughly six percent of all gene families have expanded in the pleurocarpous mosses, relative to acrocarpous mosses. These gene families are enriched for several gene ontology (GO) terms, including interaction with other organisms. The increase in copy number coincident with the radiation of Hypnales suggests that a process such as whole genome duplication or a burst of small-scale duplications occurred during the diversification. In over 500 gene families we found evidence of a reduction in purifying selection. These gene families are enriched for several terms in the GO hierarchy related to "tRNA metabolic process." Our results reveal candidate genes and pathways that may be associated with the transition to pleurocarpy, illustrating the utility of phylotranscriptomics for the study of molecular evolution in non-model species.

1. Introduction

The Bryophyta (mosses) are one of three phyla of non-vascular land plants, and comprise more than 13,000 species (Magill, 2014). Although it is one of the oldest groups of land plants, with fossils dating to at least the lower Permian (Smoot and Taylor, 1986), a significant amount of genus-level diversity has been generated in bursts that are coincident with the diversification of extant ferns and angiosperms in the Mesozoic (Laenen et al., 2014). Approximately 42% of moss species diversity (Crosby et al., 1999) belong to the pleurocarpous mosses or Hypnanae (Buck et al., 2005) a monophyletic "crown group" of mosses typically defined by the development of female gametangia on short, lateral branches lacking differentiated vegetative leaves (La Farge-England, 1996). This is in contrast to the ancestral growth form of mosses, acrocarpy and cladocarpy, where gametangia (and thus sporophytes) are usually terminal on upright shoots, or branches bearing well developed leaves. Recent fossil discoveries have pushed the origin of the pleurocarpous growth form into the late Permian (de Souza et al., 2012), but the major diversification of extant pleurocarpous moss lineages occurred much later; in fossil-calibrated phylogenies, the most speciose pleurocarpous moss families (i.e., Amblystegiaceae, Hypnaceae, or Brachytheciaceae) are estimated to have diversified within the last 100 million years (Laenen et al., 2014; Newton et al., 2007; Shaw et al., 2003).

☆ This paper was edited by the Associate Editor Stefanie M. Ickert-Bond.

* Corresponding authors at: Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, IL 60022, United States (N.J. Wickett).
E-mail addresses: mjohnson@chicagobotanic.org (M.G. Johnson), nwickett@chicagobotanic.org (N.J. Wickett).

http://dx.doi.org/10.1016/j.ympev.2016.01.008
1055-7903/

The superorder Hypnanae comprises four orders: Hypnodendrales, Ptychomniales, Hookeriales, and Hypnales, with the latter including most of the diversity, namely ±4000 species in about 430 genera and 42 families (Goffinet et al., 2009; Huttunen et al., 2013). Just eight families hold the majority of genera, 5% of the genera hold the majority of species and nearly 200 genera are monospecific. Suprageneric taxa within Hypnanae were traditionally circumscribed using morphological traits and habitat type. Given the uneven distribution of species among genera and families and the rapid tempo of diversification in Hypnales, it is not surprising that most of these morphologically-defined taxa are not monophyletic (Cox et al., 2010). Lineage-through-time plots revealed that Hypnales, unlike its sister group (Hookeriales), underwent a rapid, explosive diversification rather than a gradual diversification early in its history (Shaw et al., 2003). Many of the families within Hypnales likely diversified after the radiation of angiosperm forests (Laenen et al., 2014; Newton et al., 2007), which is hypothesized to have been a major driver of diversification in other epiphytic and understory-dwelling plants, such as leptosporangiate ferns and liverworts (Feldberg et al., 2014; Schneider et al., 2004). The diversification of leptosporangiate ferns parallels that of major families in Hypnales in both timing (late Cretaceous) and extant diversity in similar habitats (e.g., low-light, epiphytic; (Li et al., 2014)), which raises the possibility that a signature of the radiation can be found in the genomes of pleurocarpous mosses.

Recent genomic evidence has revealed that whole genome duplication (WGD) events are associated with many gene family expansions in land plants (Barker et al., 2008; Blanc and Wolfe, 2004a; Jiao and Paterson, 2014; Jiao et al., 2011; Li et al., 2015; Maere et al., 2005). Although most of the gene copies generated by WGD events are lost due to fractionation and subsequent "red iploidization" or nonfunctionalization (Jiao et al., 2011), in *Arabidopsis* some of the duplications led to neo- and/or sub-functionalization (Moore and Purugganan, 2005), resulting in the evolution of parallel gene networks (Blanc and Wolfe, 2004a) with copy-specific gene regulation (Spangler et al., 2012). Paleopolyploidy events in plants may provide opportunities for positive selection (or, at the very least reduced purifying selection) on retained duplicated copies of genes. Gene duplications in some gene families have been associated with key innovations in angiosperms; for example, the expansion of the CYCLOIDEA gene family is linked to the development of zygomorphic flowers. Members of the CYCLOIDEA gene family that have been implicated in symmetry contain conserved domains with sites under positive selection following duplication (e.g. Chapman et al., 2008).

The identification of genomic signatures, such as WGD, associated with radiations in non-model organisms has benefited from the emergence of sequencing techniques that reduce genomic complexity, such as transcriptome sequencing (Cannon et al., 2015; Yang et al., 2015). These methods allow researchers to generate large, comparative data sets without having to invest considerable resources in sequencing entire genomes, which can vary enormously in size. By using only the coding portion of the genome in a phylotranscriptomic approach, it is possible to provide an evolutionary context to inferred paleopolyploidy, gene family expansion, and shifts in selective regimes of duplicated genes, without having to sequence the extensive non-coding portion of a genome. To date, all phylogenetic analyses involving pleurocarpous mosses have sampled few discrete nuclear coding loci and focused on dense taxon sampling (Cox et al., 2010; Huttunen et al., 2012; Shaw et al., 2003). In contrast, transcriptome sequencing generates enough data to leverage methods that have the potential to identify genomic signatures, such as evidence of WGD, associated with the radiation of Hypnales, as has been observed in other land plant radiations.

Here, we identify homologous and orthologous genes from sequenced transcriptomes using a phylogeny-based approach (Yang and Smith, 2014). We construct a species tree from over 650 nuclear, protein-coding loci for 21 pleurocarpous mosses, five acrocarpous mosses, and the well-annotated proteome of the model moss *Physcomitrella patens*. Our sampling includes *Aulacomnium palustre*, an exemplar of the Aulacomniales resolved by Bell et al. (2007) as the sister-group to the Hypnanae. We use the species tree to infer the history of gene duplication events, gene family expansions, signatures of whole genome duplications, and shifts in rates of selection. Functional analysis of gene families that underwent expansion during the diversification of Hypnales will inform future studies on the genetic basis of the radiation of this most speciose lineage of pleurocarpous mosses.

2. Materials and methods

2.1. RNA extraction and sequencing

We generated 25 transcriptomes for this study from 21 pleurocarpous and four acrocarpous moss species (Table 1). Two of our samples are from the same species, *Aulacomnium palustre*, which was sequenced twice due to its critical phylogenetic position. All samples were wild-collected and then placed in clear plastic containers within a growth chamber for at least one week prior to RNA extraction. Tissue was sampled from young, green shoots and flash-frozen in liquid nitrogen, then ground to a powder with a mortar and pestle. Total RNA was extracted using the Spectrum Plant Total RNA Kit (Sigma–Aldrich, St. Louis, MO, USA) with no modification to the standard protocol. RNA was quantified using a Qubit Fluorometer (Life Technologies, Grand Island, NY, USA). RNA quality was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies Inc, Santa Clara, CA, USA) at the Northwestern University Center for Genetic Medicine (Chicago, IL, USA). RNA-Seq libraries for Illumina sequencing were prepared at BGI (Shenzhen, China) using the TruSeq RNA Sample Preparation Kit v2 (Illumina Inc., San Diego CA, USA). Three of the libraries (NW_1, NW_2, and NW_3, see Table 1) were multiplexed and sequenced on one lane of Illumina HiSeq2000 (2×100 Paired End) in February, 2014. A second batch of ten samples (accession numbers NW_45 through NW_62, see Table 1) was multiplexed and sequenced across two lanes in July, 2014, and the remaining fourteen samples were multiplexed and sequenced across two lanes in November, 2014. All sequencing was carried out at BGI (Shenzhen, China). All unedited sequence reads were deposited in the NCBI Sequence Reads Archive (BioProject PRJNA296787).

2.2. Transcriptome assembly and filtering

Raw reads were demultiplexed and adapters were removed by BGI (Shenzhen, China) prior to delivering the data. We further trimmed the sequences with Trimmomatic (Bolger et al., 2014) using the following parameters: LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN: 36. We assembled each transcriptome from the filtered reads with the Trinity pipeline (Grabherr et al., 2011; Haas et al., 2013) using default parameters. We applied a hierarchal filtering approach in order to reduce the complexity of downstream analyses, and to reduce the possibility of contamination of our datasets by transcripts from associated organisms. First, the transcripts were translated using the Transdecoder (version 20140704, http://transdecoder.github.io) tool included with Trinity. Transdecoder chooses the most valid open reading frames from each transcript using a likelihood approach that incorporates domain similarity matches to the Pfam database (Finn et al., 2014) using the hmmscan function in HMMER (Johnson et al.,

Table 1
Transcriptome assembly statistics and herbarium voucher information. Buck–NY, Goffinet, Quandt–CONN, Shaw–DUKE.

ID	Species	Voucher collection	Paired reads (Millions)	Trinity transcripts (thousands)	Percent of transcripts passing filter (BLAST and Transdecoder) (%)	Masked transcripts kept	Homologous gene family trees	BUSCOs
NW-1	*Climacium americanum*	Goffinet 11684	58.86	129.4	33.7	20,018	12,295	390
NW-2	*Thuidium delicatulum*	Goffinet 11686	67.17	137.6	46.1	20,289	12,501	391
NW-3	*Hypnum cupressiforme*	Goffinet 11687	72.38	235.6	31.7	22,035	12,735	392
NW-45	*Pleurozium schreberi*	Goffinet 11700	23.11	108.1	37.3	18,098	11,965	388
NW-51	*Aulacomnium palustre*	Goffinet 11701	38.23	135.4	40.9	19,894	12,313	390
NW-53	*Bryoandersonia illecebra*	Buck 63016	34.75	111.8	41.8	20,891	12,702	392
NW-55	*Rhytidiadelphus subpinnatus.*	Goffinet 11708	34.76	82.4	48.0	18,490	12,068	391
NW-56	*Kindbergia praelonga*	Goffinet 11707	31.73	132.8	34.5	21,146	12,528	387
NW-57	*Callicladium haldanianum*	Goffinet 11688	25.38	169.4	33.6	19,332	11,991	381
NW-59	*Hylocomium brevirostre*	Goffinet 11699	30.56	211.8	31.1	25,976	13,205	386
NW-60	*Hylocomium splendens*	Goffinet 11696	26.69	147.6	35.3	20,902	12,375	391
NW-61	*Calliergon cordifolium*	Goffinet 11693	29.07	300.3	27.0	21,494	12,417	389
NW-62	*Aulacomnium palustre*	Goffinet 11702	23.79	147.0	39.3	20,489	12,335	390
NW-65	*Anomodon rostratus*	Goffinet 11711	12.88	115.3	35.8	19,720	12,253	390
NW-66	*Thelia asprella*	Goffinet 11712	22.01	135.6	37.6	19,994	12,421	391
NW-69	*Pilotrichella flexillis*	Quandt DR158/ WP281	22.91	128.0	37.5	20,181	12,535	391
NW-71	*Antitrichia curtipendula*	Shaw 17555	21.87	101.5	44.2	20,736	12,481	385
NW-72	*Meteoridium remotifolium*	Quandt DR152/ WP281	22.78	102.5	42.0	18,649	12,066	387
NW-74	*Rhytidiopsis robusta*	Shaw 17554	19.18	113.1	37.8	20,312	12,453	386
NW-76	*Forrstroemia trichomitria*	Shaw 17557	21.27	113.8	40.9	19,161	12,365	389
NW-77	*Rhodobryum ontariense*	Goffinet 11803	13.34	53.3	60.6	11,296	9470	382
NW-79	*Dicranum scoparium*	Goffinet 11775	20.08	120.7	37.1	13,577	9843	383
NW-84	*Platyhypnidium riparioides*	Goffinet 11802	19.28	100.4	39.1	18,969	12,011	385
NW-85	*Plagiothecium laetum*	Goffinet 11851	17.19	144.7	29.5	20,620	12,560	382
NW-86	*Leucobryum glaucum*	Goffinet 11773	20.64	171.5	30.3	12,566	9450	386

2010). All transcripts with a valid translation were searched against a custom BLAST database containing protein sequences from 22 land plant nuclear genomes, including the moss *Physcomitrella patens*, downloaded from Phytozome (phytozome.jgi.doe.gov). We accepted protein matches in blastp (version 2.2.29) with an e-value below 10^{-10}. For the next stages of analysis, we included all transcripts that had a significant hit to the proteome database, as well as all Trinity-annotated isoforms of that same transcript.

2.3. Clustering transcripts into homologous gene families

In order to cluster transcripts from all species, remove redundant isoforms, and construct a species tree from low-copy genes, we employed the phylogenetic clustering method described by Yang and Smith (2014). In this pipeline (hereafter, the Yang/Smith Pipeline), all transcripts that passed the above filtering procedures were first grouped using an all-vs-all BLAST search of nucleotide sequences from every species. Significant hits (e-value < 10^{-5}) were clustered, using the software MCL (Enright et al., 2002), into gene families using a hit-fraction of 0.3 and an inflation parameter of 2.0. All clusters containing sequences from at least four species (of 26 total, including the *Physcomitrella* proteome) were aligned using MAFFT (Katoh and Standley, 2013) and nucleotide gene trees were reconstructed from peptide sequences with RAxML (Stamatakis, 2014) under the CAT model.

At this stage in the pipeline, the gene clusters may include isoforms as inferred by Trinity. To account for the possibility that some of these may actually be paralogs, rather than alternative splice forms, the Yang/Smith Pipeline extracts monophyletic or paraphyletic groupings of transcripts from a single taxon. These clades are reduced to contain only the sequence with the longest

unambiguous alignment. We also trimmed the gene family trees to remove terminal branches whose length exceeded an absolute (0.3 substitutions/site) or relative (more than 10× longer than its sister branch) cutoff. These terminal branches represent transcripts with potentially spurious homology to the other transcripts in the gene family tree, and were removed from further analysis. This approach retains multiple isoforms from the same Trinity component if they are not part of the same clade on the gene tree; however, recent lineage-specific paralogs will be lost. Since the goal of this project is to identify genomic changes prior to, or coincident with the diversification of Hypnales, rather than species-specific changes, the masking of lineage-specific duplications does not impact our conclusions. We refer to the subset of transcripts that remain following this phylogenetic transcript clustering as the **masked** dataset.

When using transcriptomes for gene discovery, rather than quantifying relative expression, it is important to assess whether or not it is likely that sequences for all possible transcripts were recovered. One method to approximate whether or not we sequenced all possible transcripts that were present in the tissues we sequenced is to determine how well we have recovered a core set of genes for our taxonomic group. Although no such set is known for mosses (or even land plants), a set of core genes is defined for all eukaryotes. The Basic Universal Single Copy Orthologs (BUSCOs) are curated from all metazoan and fungal genomes, and maintained as a set of profile Hidden Markov Models (HMMs), and an inferred ancestral amino acid sequence for each orthogroup is provided (Simão et al., 2015). To determine whether our masked dataset contained BUSCOs, we first searched the translated amino acids from our masked dataset against the ancestral sequences using BLASTP. Sequences with hits were then searched against the 429 BUSCO profile HMMs using HMMER. In order to accept a match, the hmmsearch score had to exceed a minimum score threshold defined for each BUSCO.

2.4. Species tree reconstruction

Many methods of species tree reconstruction rely on the identification of orthologous sequences, that is, sequences that arose by speciation rather than duplication. The Yang/Smith Pipeline identifies sets of orthologous genes (orthogroups) by decomposing the unrooted homologous gene family trees into subtrees where a monophyletic outgroup (here, acrocarpous mosses) is sister to a monophyletic ingroup (pleurocarpous mosses). The extracted subtrees represent inferred orthologous gene families, i.e. gene families for which the most recent common ancestor (the ancestral node) underwent a speciation event and not a duplication event. More than one of these orthologous subtrees may appear within a homologous gene tree if, for example, a gene duplication occurred prior to the divergence of the ingroup and outgroup.

We further filtered the orthologous gene trees by requiring that each species be represented by exactly one transcript, and refer to this subset as the **one-to-one orthologs.** The final matrix for species tree reconstruction represented 659 orthogroups where all 26 species are represented, with a total alignment length of 361,745 amino acid residues. Transdecoder produces an amino acid file and a coding domain sequence (CDS; nucleotides) for each putative protein. We aligned the proteins from each orthogroup with MAFFT, and back-translated the sequences using the corresponding nucleotide sequences using TrimAl (Capella-Gutierrez et al., 2009). We concatenated all of the coding regions into a supermatrix using phyutility (Smith and Dunn, 2008). This matrix comprised 659 genes and 1,062,897 nucleotides, with all 26 species represented for each gene. We reconstructed the species tree in RAxML using the GTRGAMMA model using two partitions per gene (one partition for the first and second codon positions, and another partition for the third). We evaluated nodal support using 200 bootstrap replicates.

We also reconstructed the Maximum Quartet Support Species Tree (MQSST) using ASTRAL (Mirarab et al., 2014). Individual gene trees were reconstructed using RAxML with the GTRGAMMA model, including a single maximum likelihood tree as well as 200 "fast bootstrap" trees. We evaluated support on the ASTRAL trees with a "gene-wise jackknife method." We generated 200 pseudoreplicates of the dataset by sampling 10% of the maximum likelihood gene trees without replacement and calculated a MQSST tree using ASTRAL on each subset.

We also repeated both the supermatrix and MQSST reconstruction methods using the corresponding amino acid alignment. The PROTGAMMA models were used for the supermatrix and individual gene tree reconstructions in RAxML.

2.5. Gene family expansion in Hypnales

We generated a table of gene family occupancy by counting the number of transcripts present for each species in each of the 27,299 homolog gene family trees generated by the Yang/Smith Pipeline. Unlike the one-to-one ortholog set of gene trees used for phylogenetic reconstruction, each gene homolog gene family can contain many transcripts from each species. To track the phylogenetic history of these gene families and identify expansions, we used the program Count (Csurös, 2010) to reconstruct ancestral states. Count uses Wagner parsimony (with a 20% penalty for gains) to identify nodes where: (a) a gene family has more than one member and (b) the gene family has exactly one member at the immediately ancestral node.

To identify gene family expansions associated with the diversification of Hypnales (the largest order of pleurocarpous mosses), we were interested in gene family expansions reconstructed at the following nodes (Fig. 1): (A) the common ancestor of all Hypnales minus *Plagiothecium* (a genus revealed to be sister to the rest of Hypnales), (B) the common ancestor of all Hypnales (including *Plagiothecium*), (C) the common ancestor of Hypnales and *Aulacomnium*, and (D) the common ancestor of the Bryidae (includes Hypnales, *Aulacomnium*, and *Rhodobryum*). In order to assign GO annotations to transcripts of non-model organisms, we used the annotation pipeline Trinotate (trinotate.github.io). Briefly, the pipeline searches transcripts (and their protein translations) against curated functional annotation databases, including Pfam and SwissProt. We assigned GO annotations to each homologous gene family cluster by recording all GO annotations from Trinotate made to each transcript in the cluster.

We performed a gene ontology enrichment analysis using the orthogroups with inferred expansions on one of the four nodes of interest, using the Python package goatools (version 0.5.4, github.com/tanghaibao/goatools). All GO categories annotated for all 15,459 homologous gene family clusters was used as the baseline, and we controlled for multiple testing using the False Discovery Rate method (Benjamini and Hochberg, 1995).

2.6. Evidence of paleopolyploidy

Whole genome duplication (WGD) events can be detected from transcriptome data by finding pairs of paralogous sequences within the transcriptome (Barker et al., 2009; Blanc and Wolfe, 2004b; Yang et al., 2015). The synonymous substitution rate (Ks) is calculated from each pair of paralogs. In principle, the distribution of Ks values should approximate an exponential distribution, reflecting the age distribution of gene duplication events—many pairs of genes with low Ks values, and fewer pairs with larger Ks values. This distribution could arise from many, ongoing small-scale duplications occurring throughout the history of the lineage,

followed by the nonfunctionalization of one duplicate. However, a WGD event would result in a very large number of paralogous pairs all having the same age. If a histogram of Ks values among pairs of parologs in a transcriptome has multiple peaks at intermediate values of Ks, it may be evidence of a paleopolyploidy event.

We could not analyze the transcriptome data for the presence of recent paralogs using the masked dataset, which we used for phylogenetic analysis and ortholog detection. Masking removes recent paralogs, which would bias the estimation of Ks from paralog pairs. We also could not use the raw output from Trinity, which produces transcripts with names such as "c1250_g1_i1." The first field refers to a "component" of the DeBruijn graph, the second field to a "gene" identifier, and the third field to a putative "isoforms" for the transcript. However, these isoforms may not correspond to real alternative splice variants, but may also contain parologous gene sequences. Therefore, rather than keeping only the longest isoform, or the isoform with the highest coverage for each gene component, we used CD-Hit-EST (Fu et al., 2012) to cluster the transcripts with a high percent identity threshold (−c option, 98%) and high alignment overlap threshold (−aS option, 90%) to reduce the isoforms to a set of non-redundant transcripts for each species.

To detect ancient parology, we began with the protein sequences that matched the non-redundant transcript set from CD-Hit-EST cluster for each species separately. We clustered these protein sequences for each species again with CD-Hit, but with much lower thresholds for percent identity (−c 0.4) and alignment overlap (−aS 0.75) to maximize cluster inclusiveness. For each cluster that was not a singleton, we constructed pairwise amino acid alignments among all proteins in the cluster using MAFFT. The corresponding nucleotide transcript sequences were forced into the amino acid alignments using pal2nal (Suyama et al., 2006), and all gap regions and internal stop codons were removed.

For each pair of paralogous nucleotide sequences, we calculated the synonymous substitution rate (Ks) with KaKs_Calculator (Zhang et al., 2006), using the "GY" method (Goldman and Yang, 1994), also known as F3x4. We investigated the presence of multiple normal distributions of Ks values using mixture models, implemented in the R package mclust (Fraley et al., 2012). We evaluated mixture models with between one and ten components, and the best fit model was chosen using the Bayesian Information Criterion (BIC).

2.7. Detecting changes in selection

We investigated the effect of the Hypnales radiation on signatures of molecular selection using the codeml package implemented in PAML (version 4.7; (Yang, 2007). For this analysis we used a subset of orthogroups where (1) all six acrocarpous mosses were present and (2) at least ten pleurocarpous mosses were present. We aligned protein sequences with MAFFT and back-translated using the corresponding CDS sequences using trimAL (version 1.4.rev15 Capella-Gutierrez et al., 2009), which also removed codons in the sequence matrix if they were present in fewer than five sequences. We calculated a tree for each orthogroup using FastTree (Price et al., 2010). The common ancestor of pleurocarpous moss sequences in each orthogroup was determined using the Python package ETE2 (Huerta-Cepas et al., 2010), which also assisted in running codeml. All branches descending from the common ancestor of pleurocarp sequences were marked as "foreground" for branch-model analysis. We estimated the ratio of non-synonymous to synonymous nucleotide substitution rate (dN/dS or omega) under two models: in the "M0" model, all branches are assumed to have the same omega, but in the "bfree" model, separate omegas are estimated for the "foreground" and "background" branches. Because the M0 model is nested within the bfree model, the significance of the bfree model can be determined with a Likelihood Ratio Test (LRT) with one degree of freedom. We tested the significance of the LRT against a chi-squared distribution, and accounted for multiple tests by accepting p-values less than 0.0001.

For genes with evidence of different rates of evolution in pleurocarps and acrocarps, we conducted a GO Enrichment analysis using the same procedure as above. We summarized the GO Enrichment results using ReviGO (Supek et al., 2011), which produces a visualization of the semantic similarity among GO categories.

A conceptual diagram illustrating our entire analysis can be found in Supplemental Fig. 1.

3. Results and discussion

3.1. Transcriptome assembly and orthology detection

Across our 25 assembled transcriptomes, we recovered between 53,000 and 301,000 transcripts (Table 1). After applying our filtering steps, assuring that the transcript contained a valid protein using Transdecoder and that the protein had BLAST hits against known land plant proteomes, we retained between 32,349 and 81,018 protein coding sequences per species. The Yang/Smith Pipeline then clustered these transcripts from all 25 transcriptomes and the *Physcomitrella patens* proteome into 27,299 homologous gene families that contained transcripts from at least four different species. We then produced the "masked" dataset by removing sequences if transcripts from the same species form monophyletic or paraphyletic groups on the multispecies gene family trees. On average, 19,393 translated proteins were retained (Table 1), and each transcriptome had at least one sequence in an average of 12,054 homologous gene families. Following the filtering and clustering steps, we retained between 12,566 and 25,976 transcripts within 27,299 homolog family trees (Table 1).

To approximate whether we sampled the pool of all possible transcripts as deeply as possible, we searched for the 429 BUSCOs (universal orthologous genes) defined for all eukaryotes (Simão et al., 2015). We were able to detect between 382 and 391 BUSCOs in our 25 transcriptomes. For comparison, we could find 391 BUSCOs in the *Physcomitrella* proteome, suggesting that this is the maximum number for mosses, and that the remaining 38 orthogroups are not universal to all eukaryotes if plants are included. The authors of the BUSCOs pipeline have begun development of a more plant-specific set of universal orthologs (buscos.ezlab.org). We therefore accept 391 as the maximum number of eukaryote BUSCOs that can be found in mosses.

Because we used different multiplexing schemes, we also tested whether the number of reads correlated with our assessment metrics, but the number of reads per sample did not correlate with the number of Trinity transcripts ($F_{23} = 3.2$, $r^2 = 0.08$, $p = 0.09$) or the number of proteins retained in the masked dataset ($F_{23} = 3.1$, $r^2 = 0.08$, $p = 0.09$). The correlation between the number of reads and the number of BUSCOs recovered was significant ($F_{23} = 8.0$, $r^2 = 0.23$, $p = 0.01$). On average, we recovered three additional BUSCOs from the eight samples with more than 30 million reads, compared to the seventeen samples with fewer than 30 million reads. Overall, this suggests that additional sequencing of each species would not change the inferences made here.

We also assessed the completeness of our transcriptomes by comparison to the well-curated proteome of *Physcomitrella patens*. Although 21,312 proteins from *Physcomitrella* clustered with transcripts from at least four of our transcriptomes, only 7778 gene families in the masked data set contain a protein from the model moss proteome. This is likely due to the masking step of the Yang/Smith pipeline, which removes sequences if they form a

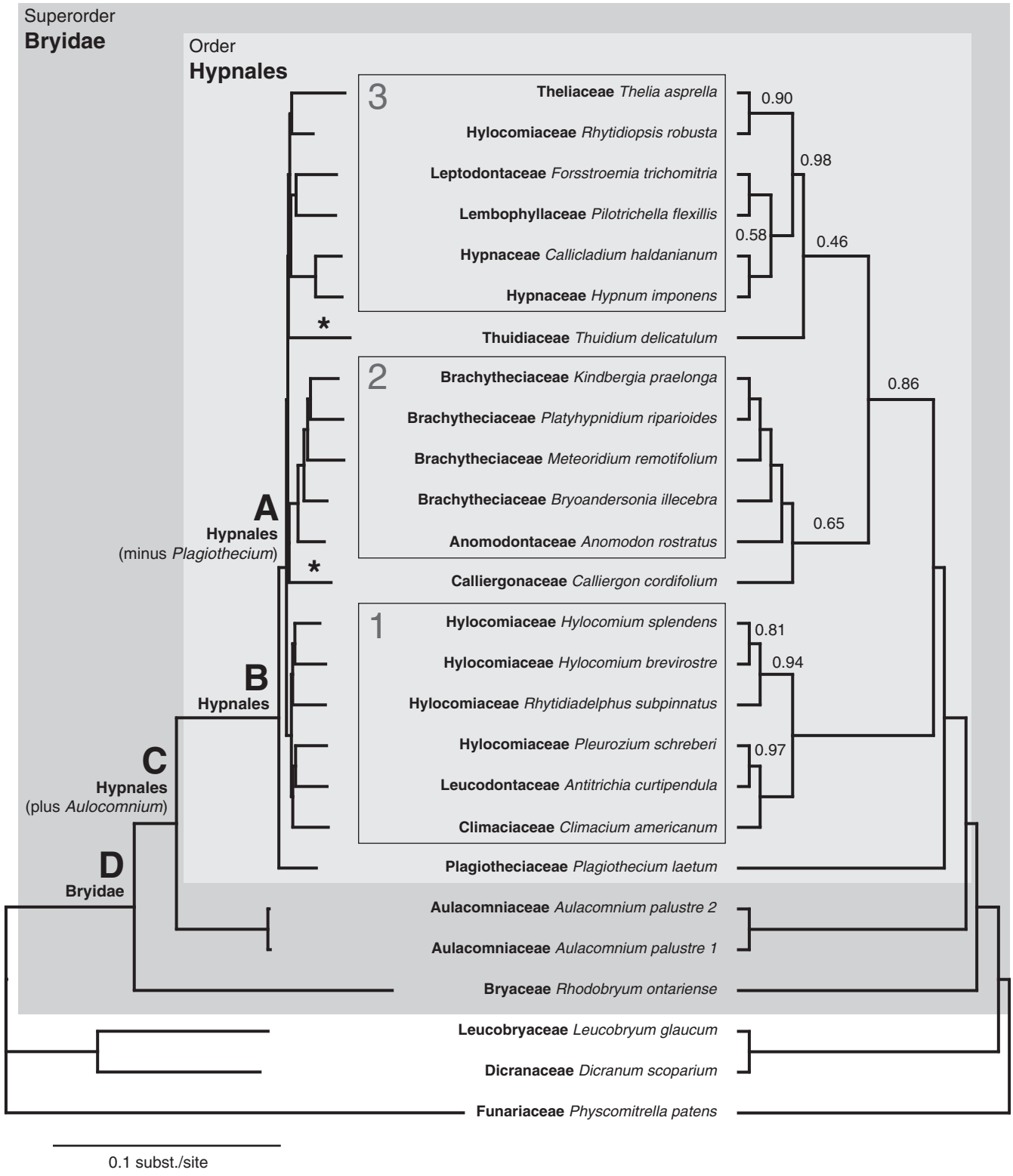

Fig. 1. Species trees constructed from 659 orthogroups in 21 pleurocarpous and five acrocarpous mosses, and the number of homologous gene family expansions at key nodes on the tree. Left: Maximum-likelihood nucleotide tree from RAxML. All nodes are supported at 100% bootstrap support (200 pseudoreplicates) except for the nodes indicated by stars. The letters indicate four key nodes at which gene family expansions were calculated: A: Hypnales minus *Plagiothecium*, B: Hypnales, C: Hypnales mosses plus *Aulacomnium*, D: Superorder Bryidae. Right: Maximum Quartet Support Species Tree from ASTRAL, reconstructed from nucleotide gene trees estimated in RAxML. Nodal support is 100% except where indicated by genewise-jackknife.

monophyletic group that includes only sequences from the same taxon. As a result of the whole genome triplication event that occurred during the diversification of Funariaceae (Rensing et al., 2007), many *Physcomitrella* paralogs clustered together in our gene family trees at a 3–1 ratio relative to other mosses. On the initial gene family trees, the *Physcomitrella* paralogs would share a common ancestor within *Physcomitrella*, and the masking step would retain only one of these paralogs per gene family.

There are also a large percentage of homolog gene families (over 19,000) that do not contain a representative from *Physcomitrella*, but do contain transcripts from at least four different transcriptomes. The transcriptome sequences could only progress to this point in our pipeline if they had a significant (*e*-value < 10^{-5}) BLAST hit to a land plant proteome. Therefore, some of these gene families may be ancestral land plant families that have since been lost in the lineage leading to *Physcomitrella*. Alternatively, because Yang/Smith Pipeline is intended primarily as a way of identifying low-copy orthologs for phylogenetic inference, gene families may be circumscribed narrowly. Higher-order clustering may reveal homology between gene families. Additional genome sequences from other bryophytes would assist in this effort.

3.2. Species tree reconstruction

We constructed species trees from 659 orthogroups that had a one-to-one orthology between acrocarpous and pleurocarpous mosses, as well as a representative sequence from all 26 taxa. Both the Maximum Likelihood (ML) Tree and the (Maximum Quartet Support Species Tree (MQSST) have strong backbone support, reflecting relationships among acrocarpous mosses, and their relationships to Hypnales (Fig. 1). When rooted using *Physcomitrella patens*, *Aulacomnium palustre* (both accessions) is sister to the pleurocarpous mosses (Hypnales), as expected (Bell et al., 2007; Cox et al., 2010). Sister to this clade is *Rhodobryum*; together with *Aulacomnium* and Hypnales these species represent the subclass Bryidae (Goffinet et al., 2009; Stech and Frey, 2008). Previous phylogenetic evidence supports Bryidae as sister to Dicranidae (Chang and Graham, 2011; Cox et al., 2010), which in our dataset is represented by *Leucobryum* and *Dicranum*.

Within the pleurocarpous mosses, *Plagiothecium* is resolved as sister to the rest of Hypnales with maximum support in both tree reconstruction approaches (Fig. 1), consistent with previous efforts using one or a few genes (Cox et al., 2010; Huttunen et al., 2012; Merget and Wolf, 2010). Two other multi-species relationships are maximally supported using all methods. Clade 1 contains four of our five samples from the Hylocomiaceae plus *Climacium* (Climaciaceae) and *Antitrichia* (Leucodontaceae), and is resolved as sister to the remaining Hypnales (minus *Plagiothecium*) with maximal support in both the RAXML and ASTRAL trees. Clade 2 contains four Brachytheciaceae, which compose a monophyletic sister lineage to *Anomodon* (Anomodontaceae), and relationships within this clade are fully resolved in both trees. Clade 3 contains the remaining species of Hylocomiaceae (Rhytidiopsis) along with two species of Hypnaceae, Forsstroemia (Leptodontaceae), Pilotrichella (Lembophyllaceae), and Thelia (Theliaceae). Rhytidiopsis has been accommodated in the Hylocomiaceae (Buck and Vitt, 1986) but affinities to Thelia had first been proposed by Chiang and Schaal (2000), and then by Huttunen et al. (2012). These three well-resolved clades are consistent with those recovered from inferences from discrete loci from all genomic compartments (Huttunen et al., 2012).

Two branches on the maximum likelihood tree did not receive 100% support. Inspection of bootstrap trees reveals that full phylogenetic resolution within Hypnales is impeded by the positions of two species: *Thuidium delicatulum* and *Calliergon cordifolium* (Supplemental Fig. 2). Lineage movement analysis of the bootstrap replicates revealed the two species are as likely to be sister-species (25%) as they are in their maximum likelihood arrangement (26%). Likewise, the placement of *Thuidium* sister to Clade 3 (46%) and *Calliergon* as sister to Clade 2 (65%) have the lowest gene-wise jackknife values of any relationship on the ASTRAL tree. We expect that denser taxon sampling of transcriptomes within Hypnales and the less species-rich orders of pleurocarpous mosses would allow us to more confidently reconstruct the affinities of *Thuidium* and *Calliergon*. However, while we are unable to completely resolve relationships in these clades, the focus of this study is to reconstruct the history of gene family evolution at higher phylogenetic levels.

The backbone of the phylogeny was reconstructed with equal confidence using amino acid characters (Supplemental Fig. 3). The resolution of clades within Hypnales was less certain; Clades 1 and 2 were fully supported using both RAxML and ASTRAL, but the support for Clade 3 was similarly reduced by the placement of *Thuidium* and *Calliergon*. The relationships among the three major clades within Hypnales were unresolved with both methods using the amino acid matrix.

Our results suggest that intra-genomic phylogenetic conflict complicates the resolution of family-level relationships within Hypnales. Earlier studies, which focused effort on taxon sampling, with few genes, had similar difficulty resolving the same relationships (Cox et al., 2010; Huttunen et al., 2012). Although we have employed a phylotranscriptomic approach, it is likely that the relationships within Hypnales may not be resolved without a broader taxonomic sampling. Specifically, several major families of Hypnales were not sampled in our phylogeny, and the addition of transcriptomes from the other orders of pleurocarpous mosses (e.g. Hookeriales) would root the Hypnales phylogeny more accurately. Though not suitable for taxonomic revision, our data do present a large step forward in the genomic sampling effort, increasing the number of genes sequenced in mosses by two orders of magnitude over previous studies. We anticipate that the phylogenetic framework presented here, particularly with respect to the identification of orthologous gene families, will provide a foundation for more taxon-dense phylogenetic studies in the future.

3.3. Gene family expansion analysis

Despite some instances of low phylogenetic resolution within Hypnales, the strong support among backbone clades of the mosses enables us to infer patterns of gene family evolution. Because we are using transcriptomes, gene family membership may be reduced due to incomplete expression of the proteome. However, we can treat the number of distinct transcripts per species (an approximation of gene copy number, not relative expression) as a discrete trait that evolves along the species tree. By reconstructing the "gene family occupancy" for each gene family at each node on the species tree, we minimize the noise associated with gene loss and under-expression in terminal taxa. Specifically, we identified expansions in homologous gene families at four robustly supported (100% in all methods), nested nodes, marked in Fig. 1: (A) Hypnales minus *Plagiothecium*, (B) Hypnales, (C) Hypnales plus *Aulacomnium*, and (D) Bryidae.

We reconstructed gene family occupancy using a Wagner parsimony approach in the software Count (Csurös, 2010). Expansions were identified by two criteria at each of the four target nodes: the gene family had to be reconstructed as (1) containing multiple paralogs at the target node and (2) exactly one paralog at the immediately ancestral node. Gene family contractions were identified by inverse criteria. Our analysis revealed that homologous gene family expansions at the four target nodes were extremely enhanced compared to homologous gene family contractions (Fig. 2A, Supplemental Table 2). The highest discrepancy between expansions and contractions occurred at the Hypnales + *Aulacomnium* (C) node, with 799 expansions and only 11 contractions.

Across all four target nodes, 1712 homologous gene families exhibited low ancestral occupancy and high occupancy in Hypnales. Sixty-one GO categories (Fig. 2B, Supplemental Table 1) were enriched among the homologous gene families that had expanded in Hypnales. The enriched categories included "post-embryonic morphogenesis" (GO:0009986) and "developmental process

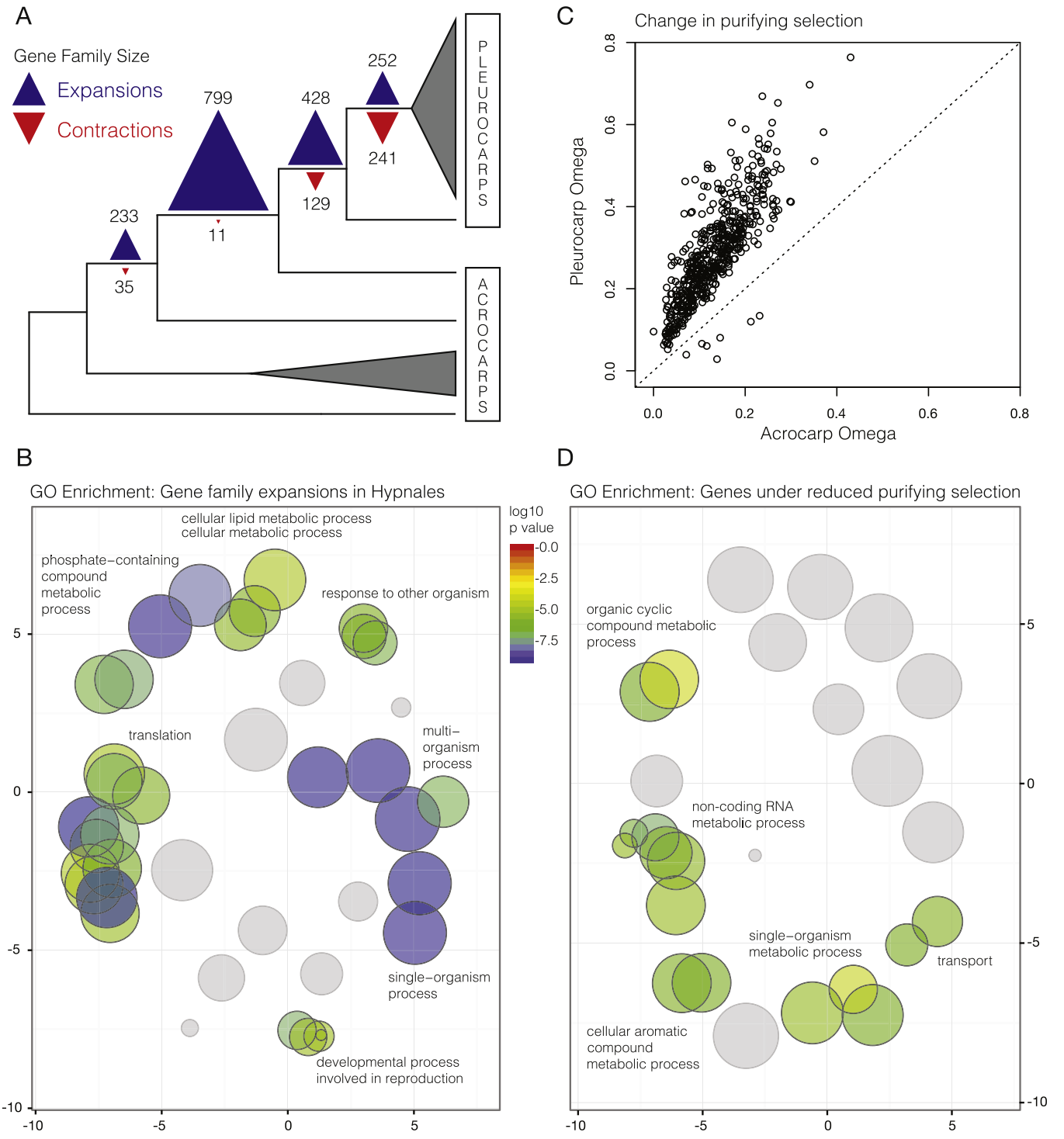

Fig. 2. Gene family expansion in pleurocarpous mosses is associated with enriched gene ontology (GO) categories and reduced purifying selection. (A) Expansions and contractions in homologous gene families reconstructed by Wagner parsimony at four nested nodes that represent common ancestors of the pleurocarpous mosses. (B) Multidimensional scaling of semantic similarity among GO categories for gene families that have expanded in pleurocarpous mosses. Two GO terms (circles) have similar semantic similarity if they are siblings in the GO heirarchy or are related by inheritance. Overlapping circles of terms share similar characteristics and are labeled in the figure with one representative GO term. (C) Comparison of omega values (ratio of non-synonymous to synonymous substitution rate) in 526 orthogroups where a two-omega model was preferred to a single omega for the tree. The remaining 2220 gene family trees for which the two-omega model was not preferred are not shown. Each point represents a gene tree, where the horizontal axis is the omega inferred from the "background" branches (acrocarpous mosses), and the vertical axis is the omega inferred from the "foreground" branches (pleurocarpous mosses). The dotted line represents equivalent omegas in the two sets of branches; only seven genes have a lower inferred omega in pleurocarpous mosses, compared to acrocarpous mosses. An omega over 1.0 represents evidence of positive selection, but an increased omega in the foreground branches may represent reduced purifying selection. (D) Semantic similarity plot of GO terms that are enriched in the set of gene families shown to have reduced purifying selection. Overlapping sets of GO categories are represented by one GO term. For a full list of all enriched categories, see the supplemental information.

involved in reproduction" (GO:0003006). Of particular interest were several categories involved in interactions among organisms, such as "response to external biotic stimulus" (GO:0043207) and "defense response to other organism" (GO:0098542).

These functional annotations are intriguing because the shared derived growth form of pleurocarpous mosses (a shift in reproductive structure from terminal to lateral, and a generally more prostrate and branching growth form) may be associated with traits that later facilitated the rapid radiation of Hypnales. Because most of the major families in Hypnales diversified concurrent with the diversification of other plants (Laenen et al., 2014), the opportunity for new biotic interactions may have arisen. For example, the diversification of angiosperms may have presented mosses with a novel substrate, driving neofunctionalization of genes responsible for external stimuli and defense responses. It is clear from our results that many gene families have expanded coincident with the radiation of the pleurocarpous mosses. However, it is unclear when the gene families expanded during the evolution of pleurocarpous mosses, because our sampling was limited to Hypnales. Future studies could identify whether many of the gene families instead expanded in the shared ancestor of Hypnales and Hookeriales (the other major order of pleurocarpous mosses), or perhaps prior to the divergence of earlier pleurocarpous lineages, namely the Hypnodendrales and the Ptychomniales (Bell et al., 2007). Detailed studies involving gene knockouts are also needed to confirm that these gene families have significant impact on life history traits in pleurocarpous mosses.

3.4. Evaluation of paleopolyploidy in pleurocarpous mosses

The increase in gene family occupancy that we observe can be explain either by (1) a whole genome duplication (WGD) event, or (2) many small-scale duplications (SSDs). If the large increase in gene family occupancy in the pleurocarpous mosses indicates a whole genome duplication, it does not appear to have been accompanied by an increase in the base chromosome number. Most of the species in our transcriptome dataset have a base chromosome count of 10, 11, or 12 (Fritsch, 1991), below a threshold that has been used previously for inferring polyploidy in mosses (Crawford et al., 2009).

We observe a consistent pattern of gene family occupancy ratios (2:1 or 3:1 pleurocarp:acrocarp) in about 6% of the gene family trees in the masked dataset. Many gene family duplications localize to specific branches on our species tree (Figs. 1 and 2A), particularly the common ancestor of Hypnales and *Aulacomnium*, suggesting that the gene family expansions share a common age, indicative of WGD. However, Hypnales shares a more recent common ancestor with other groups of pleurocarpous mosses (such as Hookeriales) that were not sampled as part of this study. The apparent increase in gene family occupancy at the common ancestor of Hypnales and *Aulacomnium* may be the result of SSDs that occurred along the long branch that separates these groups (Fig. 1), but could not be reconstructed to more specific nodes due to our taxon sampling. Data from additional groups of mosses, including the other orders of pleurocarpous and protopleurocarpous mosses, would be necessary to pinpoint the age of gene duplications and better distinguish WGD from SSD using the gene family occupancy reconstruction method.

We were also unable to detect a clear signal of WGD in our 25 transcriptomes using a Ks-based method. Although our pipeline reliably recovered the WGD previously described from *Physcomitrella* (Rensing et al., 2007), none of our transcriptomes showed an obvious intermediate "peak" of Ks values between 0.5 and 2.0 (Fig. 3, Supplemental Fig. 1). Using the mclust method to fit Gaussian distributions to the Ks values, the best fit (evaluated by BIC score) was typically between 6 and 9 components (Supplemental Table 3), which would suggest evidence of several WGD events. However, in most cases the BIC values for several values of "g" (the number of components) were very similar. When only the model with the best BIC value was considered for each transcriptome, the means of distributions for each species did not show consistent overlap (Supplemental Fig. 4). We therefore consider any signal of WGD with this method to be weak. If an ancient WGD event occurred, the intermediate peak of Ks values may be obscured due to the age of the duplication event, as seen in the *Amborella* proteome (Amborella Genome Project et al., 2013). Additionally, we may not be able to observe consistent peaks across Hypnales if the substitution rates are too variable (Barker et al., 2009).

3.5. Molecular signatures of selection

When a gene is duplicated, one or both copies may take on a new function (neofunctionalization), and these innovations result from positive, or reduced purifying selection acting on one or both copies (Blanc and Wolfe, 2004a; Freeling, 2009). We investigated the signature of selection in all orthologous gene trees found by the Yang/Smith pipeline that contained sequences from all acrocarpous mosses and at least ten pleurocarpous mosses (2746 orthogroups). We reconstructed gene trees for each orthogroup from back-translated nucleotide sequences and estimated branch-wise models of molecular evolution using CodeML. In the "M0" model, a single ratio of synonymous to nonsynonymous substitution rates (dN/dS, or omega) was inferred for the entire tree. In the "two-omega" model, separate omegas are estimated for the "pleurocarpous" and "acrocarpous" branches, and we determined whether this was a significantly better fit to the data using a Likelihood Ratio Test ($p < 0.0001$ to correct for multiple tests). Of the 2746 orthogroups tested, the two-omega model was preferred in 526 orthogroups, and for 519 of these, omega was greater in the pleurocarpous moss lineages (Fig. 2C). None of the inferred omega values were greater than one (which would suggest positive selection), but the large number of genes with elevated omegas in pleurocarpous lineages relative to acrocarpous lineages supports a hypothesis of reduced purifying selection.

We performed a GO enrichment analysis of the functionally annotated *Physcomitrella* proteins in the 519 orthogroups with reduced purifying selection, and revealed 19 enriched GO categories (Fig. 2D, Supplemental Table 4). Many of these GO categories belong to a single "family" of GO categories (Supplemental Fig. 5), the most specific of which is "metabolic process" (GO: 006399). To determine whether this was related to codon usage bias, we analyzed all transcriptomes and the *Physcomitrella* coding domain sequences using four metrics of codon usage calculated by the program codonW (http://codonw.sourceforge.net/). However, none of the metrics showed significant differences in codon usage between acrocarps and pleurocarps (Supplemental Table 5).

For seven orthogroups for which the two-omega model was preferred, omega was greater in the background (acrocarp) branches (Table 2). The *Physcomitrella* proteins in these orthogroups have been studied in controlled differential expression experiment, and several pairs of these seven genes are known to have correlated co-expression (see phytozome.jgi.gov). *Physcomitrella* genes Phpat.011G004300.1 (Membrane-associated hemotopoietic protein) and Phpat.011G069500.1 (Ankyrin repeat and protein kinase domain-containing protein) are known to be significantly co-expressed (Pearson's coefficient 0.82). In contrast, Phpat.001G030000.1 (Putative RNA Polymerase II regulator) and Phpat.024G039100.1 (histone H-3) are known to be significantly inversely expressed (Pearson's coefficient −0.33). If the co-expression of these genes is maintained in Hypnales, it would suggest that entire gene networks have shifted regimes of molecular

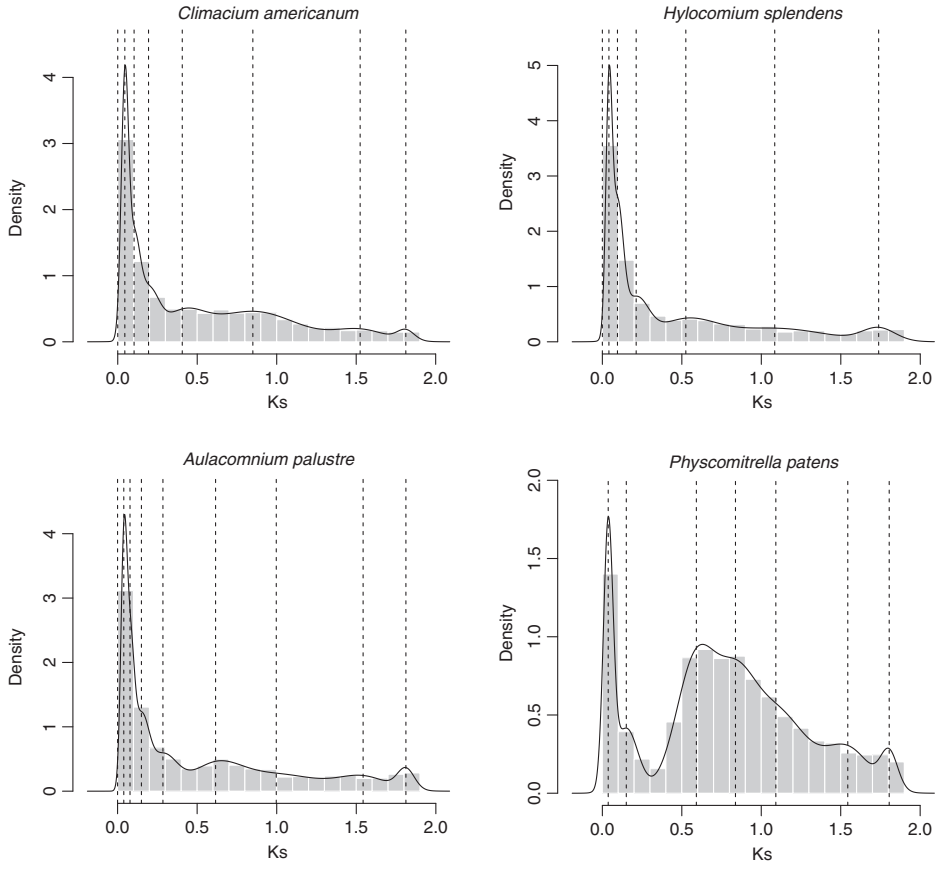

Fig. 3. Distribution of synonymous substitution rates (Ks) among pairs of paralogous genes within selected species. The dotted vertical lines represent the mean values of component distributions inferred by mclust under the model with the highest BIC score. The solid curved line is the inferred density distribution for the mixture model by mclust. For plots of all species, see Supplemental Fig. 2. *Climacium americanum* and *Hylocomium splendens* are pleurocarpous mosses, *Aulacomnium palustre* is a "protopleurocarpous" moss generally considered to be acrocarpous, and *Physcomitrella patens* is acrocarpous.

Table 2
Description of gene families with increased purifying selection in Hypnales, relative to the acrocarpous mosses.

Cluster ID	Acrocarp omega	Pleurocarp omega	*Physcomitrella* Gene ID	*Physcomitrella* gene annotation
cluster7933	0.145	0.051	Phpat.001G030000.1	Putative RNA Polymerase II regulator
cluster4410	0.212	0.12	Phpat.011G069500.1	Ankyrin repeat and protein kinase domain-containing protein
cluster5502	0.071	0.039	Phpat.011G004300.1	Membrane-associated hemotopoietic protein
cluster6993	0.106	0.066	Phpat.024G039100.1	Histone-H3
cluster10685	0.139	0.028	Phpat.007G045900.1	U2 small nuclear ribonucleoprotein B
cluster8238	0.116	0.061	Phpat.001G146700.1	ATP Binding/DNA Binding/Helicase
cluster3968	0.232	0.134	Phpat.012G062300.1	Hypothetical protein F4|10.140

selection coincident with the radiation of Hypnales. A controlled-condition differential expression study is needed to determine if the genes that have shifted regimes of selection have retained correlated expression in Hypnales pleurocarpous mosses.

It is possible that selection may act differently upon gene copies that result from small-scale gene duplications (SSDs), compared to genes that result from WGD. For example, a WGD generates gene copies in roughly equal proportions throughout enzymatic pathways, while SSDs may cause dosage imbalances and result in poor pathway flux (Lynch and Conery, 2000). As such, an alternative explanation to WGD is that the excess of gene family occupancy in Bryidae (and in Hypnales) is the result of several SSDs. In an investigation of the fate of gene copies resulting from small-scale duplications in land plants, Carretero-Paulet and Fares (2012) found that gene copies resulting from SSDs had reduced purifying selection in three angiosperm species, unlike gene copies resulting from WGD events. However, they did not observe this effect in *Physcomitrella patens*. Because we see a similar pattern (reduced purifying selection) in gene families that have expanded in Hypnales, small-scale duplications may be more plausible than a WGD. Additional taxon sampling and genome-wide analyses of synteny, particularly of other lineages within Bryidae and among other orders of the pleurocarpous mosses, are required to further distinguish between whole genome and small-scale duplications as sources of expanded gene families in pleurocarpous mosses.

4. Conclusions

This study is the first to investigate the genomic signatures associated with a rapid radiation in the largest order of pleurocarpous mosses, the lineage that accounts for the largest proportion of extant moss diversity. We describe here a set of 659 orthologous gene families and demonstrate their utility for phylogenetic

reconstruction in pleurocarpous mosses. These genes and analyses will likely form the foundation for future analyses of pleurocarp diversity, and our phylogenetic hypothesis provides a starting point to ask whether genomic features common to other rapid radiations in land plants occurred in pleurocarps. Our results suggest that both gene family expansion and a relaxation of purifying selection on many genes are significant features of the radiation of Hypnales and provide a set of candidate genes that, with further refinement, may be used in functional studies of pleurocarp development. The utility of transcriptome data for the phylogenetic analysis of molecular evolution depends on careful curation of datasets, including removal of contaminants, detection of homologous and orthologous sequences, and data analysis that allows the presence of missing data. Future phylotranscriptomic work, in this group and others, will rely on the continued development of bioinformatics pipelines to handle the challenges of working with transcriptome data. However, we have shown here that the use of transcriptomes to discover fundamental evolutionary processes that underlay the radiation of pleurocarpous mosses yields significant results and shows great promise for testing evolutionary hypotheses more broadly in mosses.

Acknowledgments

We would like to thank the Genomics Core Facility at the Northwestern University Center for Genetic Medicine Center and BGI for quality assurance and sequencing. We also thank D. Quandt (University of Bonn, Germany) for supplying two of our moss specimens. We thank B. Shaw for permission to use the photos in the graphical abstract. The masked transcriptome dataset, orthogroup assignment, individual gene alignments, and GO categories can be found in the Dryad depository: http://dx.doi.org/10.5061/dryad.475g7. This research was funded by National Science Foundation grants to AJS (DEB-1239980), BG (DEB-1240045), and NJW (DEB-1239992).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2016.01.008.

References

Amborella Genome Project, Albert, V.A., Barbazuk, W.B., dePamphilis, C.W., Der, J.P., Leebens-Mack, J., Ma, H., Palmer, J.D., Rounsley, S., Sankoff, D., Schuster, S.C., Soltis, D.E., Soltis, P.S., Wessler, S.R., Wing, R.A., Ammiraju, J.S.S., Chamala, S., Chanderbali, A.S., Determann, R., Ralph, P., Talag, J., Tomsho, L., Walts, B., Wanke, S., Chang, T.H., Lan, T., Arikit, S., Axtell, M.J., Ayyampalayam, S., Burnette, J.M., De Paoli, E., Estill, J.C., Farrell, N.P., Harkess, A., Jiao, Y., Liu, K., Mei, W., Meyers, B.C., Shahid, S., Wafula, E., Zhai, J., Zhang, X., Carretero-Paulet, L., Lyons, E., Tang, H., Zheng, C., Altman, N.S., Chen, F., Chen, J.Q., Chiang, V., Fogliani, B., Guo, C., Harholt, J., Job, C., Job, D., Kim, S., Kong, H., Li, G., Li, L., Liu, J., Park, J., Qi, X., Rajjou, L., Burtet-Sarramegna, V., Sederoff, R., Sun, Y.H., Ulvskov, P., Villegente, M., Xue, J.Y., Yeh, T.F., Yu, X., Acosta, J.J., Bruenn, R.A., de Kochko, A., Herrera-Estrella, L.R., Ibarra-Laclette, E., Kirst, M., Pissis, S.P., Poncet, V., 2013. The amborella genome and the evolution of flowering plants. Science 342, 1241089. http://dx.doi.org/10.1126/science.1241089.

Barker, M.S., Kane, N.C., Matvienko, M., Kozik, A., Michelmore, R.W., Knapp, S.J., Rieseberg, L.H., 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol. Biol. Evol. 25, 2445–2455. http://dx.doi.org/10.1093/molbev/msn187.

Barker, M.S., Vogel, H., Schranz, M.E., 2009. Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales. Genome Biol. Evol. 1, 391–399. http://dx.doi.org/10.1093/gbe/evp040.

Bell, N.E., Quandt, D., O'Brien, T.J., Newton, A.E., 2007. Taxonomy and phylogeny in the earliest diverging pleurocarps: square holes and bifurcating pegs. The Bryologist. 110, 533–560. http://dx.doi.org/10.1639/0007-2745(2007)110[533:TAPITE]2.0.CO;2.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. Ser. B (Methodological) 57, 289–300. http://dx.doi.org/10.2307/2346101?ref=no-x-route:68fa25040cda789006fe1143e3a461e1.

Blanc, G., Wolfe, K.H., 2004a. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16, 1679–1691. http://dx.doi.org/10.1105/tpc.021410.

Blanc, G., Wolfe, K.H., 2004b. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16, 1667–1678. http://dx.doi.org/10.1105/tpc.021345.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. http://dx.doi.org/10.1093/bioinformatics/btu170.

Buck, W.R., Cox, C.J., Shaw, A.J., Goffinet, B., 2005. Ordinal relationships of pleurocarpous mosses, with special emphasis on the Hookeriales. System. Biodiv. 2, 121–145. http://dx.doi.org/10.1017/S1477200004001410.

Buck, W.R., Vitt, D.H., 1986. Suggestions for a new familial classification of pleurocarpous mosses. Taxon 35, 21. http://dx.doi.org/10.2307/1221034.

Cannon, S.B., McKain, M.R., Harkess, A., Nelson, M.N., Dash, S., Deyholos, M.K., Peng, Y., Joyce, B., Stewart, C.N., Rolf, M., Kutchan, T., Tan, X., Chen, C., Zhang, Y., Carpenter, E., Wong, G.K.-S., Doyle, J.J., Leebens-Mack, J., 2015. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. Mol. Biol. Evol. 32, 193–210. http://dx.doi.org/10.1093/molbev/msu296.

Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973. http://dx.doi.org/10.1093/bioinformatics/btp348.

Carretero-Paulet, L., Fares, M.A., 2012. Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. Mol. Biol. Evol. 29, 3541–3551. http://dx.doi.org/10.1093/molbev/mss162.

Chang, Y., Graham, S.W., 2011. Inferring the higher-order phylogeny of mosses (Bryophyta) and relatives using a large, multigene plastid data set. Am. J. Bot. 98, 839–849. http://dx.doi.org/10.3732/ajb.0900384.

Chapman, M.A., Leebens-Mack, J.H., Burke, J.M., 2008. Positive selection and expression divergence following gene duplication in the sunflower CYCLOIDEA gene family. Mol. Biol. Evol. 25, 1260–1273. http://dx.doi.org/10.1093/molbev/msn001.

Chiang, T.-Y., Schaal, B.A., 2000. The internal transcribed spacer 2 region of the nuclear ribosomal DNA and the phylogeny of the moss family Hylocomiaceae. Plant System. Evol. 224, 127–137.

Cox, C.J., Goffinet, B., Wickett, N.J., Boles, S.B., Shaw, A.J., 2010. Moss diversity: a molecular phylogenetic analysis of genera. Phytotaxa 9, 175–195.

Crawford, M., Jesson, L.K., Garnock-Jones, P., 2009. Correlated evolution of sexual system and life history traits in mosses. Evolution 63, 1129–1142. http://dx.doi.org/10.1111/j.1558-5646.2009.00615.x.

Crosby, M.R., Magill, R.E., Allen, B., He, S., 1999. A checklist of the Mosses [WWW Document]. <http://www.mobot.org/MOBOT/tropicos/most/checklist.shtml> (accessed 8.28.15).

Csurös, M., 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26, 1910–1912. http://dx.doi.org/10.1093/bioinformatics/btq315.

de Souza, I.C.C., Recardi Branco, F.S., Vargas, Y.L., 2012. Permian bryophytes of Western Gondwanaland from the Paraná Basin in Brazil. Palaeontology 55, 229–241. http://dx.doi.org/10.1111/j.1475-4983.2011.01111.x.

Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucl. Acids Res. 30, 1575–1584. http://dx.doi.org/10.1093/nar/30.7.1575.

Feldberg, K., Schneider, H., Stadler, T., Schäfer-Verwimp, A., Schmidt, A.R., Heinrichs, J., 2014. Epiphytic leafy liverworts diversified in angiosperm-dominated forests. Sci. Rep. 4. http://dx.doi.org/10.1038/srep05974.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., Punta, M., 2014. Pfam: the protein families database. Nucl. Acids Res. 42, D222–D230. http://dx.doi.org/10.1093/nar/gkt1223.

Fraley, C., Raftery, A.E., Murphy, T.B., Scrucca, L., 2012. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.

Freeling, M., 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu. Rev. Plant Biol. 60, 433–453. http://dx.doi.org/10.1146/annurev.arplant.043008.092122.

Fritsch, R., 1991. Index to Bryophyte Chromosome Counts, Bryophytorum Biblioteca, Bryophytorum Bibliotheca. Science Publishers, Stuttgart, DE.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152. http://dx.doi.org/10.1093/bioinformatics/bts565.

Goffinet, B., Buck, W.R., Shaw, A.J., 2009. Morphology, anatomy, and classification of the Bryophyta. In: Goffinet, B., Shaw, A.J. (Eds.), Bryophyte Biology. Cambridge, pp. 55–138.

Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11, 725–736.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman,

N., Regev, A., . Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. http://dx.doi.org/10.1038/nbt.1883.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., Macmanes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protocols 8, 1494–1512. http://dx.doi.org/10.1038/nprot.2013.084.

Huerta-Cepas, J., Dopazo, J., Gabaldón, T., 2010. ETE: a python environment for tree exploration. BMC Bioinform. http://dx.doi.org/10.1186/1471-2105-11-24.

Huttunen, S., Bell, N., Bobrova, V.K., Buchbender, V., Buck, W.R., Cox, C.J., Goffinet, B., Hedenäs, L., Ho, B.-C., Ignatov, M.S., Krug, M., Kuznetsova, O., Milyutina, I.A., Newton, A., Olsson, S., Pokorny, L., Shaw, J.A., Stech, M., Troitsky, A., Vanderpoorten, A., Quandt, D., 2012. Disentangling knots of rapid evolution: origin and diversification of the moss order Hypnales. J. Bryol. 34, 187–211. http://dx.doi.org/10.1179/1743282012Y.0000000013.

Huttunen, S., Ignatov, M.S., Quandt, D., Hedenäs, L., 2013. Phylogenetic position and delimitation of the moss family Plagiotheciaceae in the order Hypnales. Bot. J. Lin. Soc. 171, 330–353. http://dx.doi.org/10.1111/j.1095-8339.2012.01322.x.

Jiao, Y., Paterson, A.H., 2014. Polyploidy-associated genome modifications during land plant evolution. Philos. Trans. Roy. Soc. B 369, 20130355. http://dx.doi.org/10.1098/rstb.2013.0355.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., Soltis, D.E., Clifton, S.W., Schlarbaum, S. E., Schuster, S.C., Ma, H., Leebens-Mack, J., dePamphilis, C.W., 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473, 97–100. http://dx.doi.org/10.1038/nature09916.

Johnson, L.S., Eddy, S.R., Portugaly, E., 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinform. 11, 431. http://dx.doi.org/10.1186/1471-2105-11-431.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780. http://dx.doi.org/10.1093/molbev/mst010.

La Farge-England, C., 1996. Growth form, branching pattern, and perichaetial position in mosses: cladocarpy and pleurocarpy redefined. The Bryologist 99, 170. http://dx.doi.org/10.2307/3244546.

Laenen, B., Shaw, B., Schneider, H., Goffinet, B., Paradis, E., Désamoré, A., Heinrichs, J., Villarreal, J.C., Gradstein, S.R., McDaniel, S.F., Long, D.G., Forrest, L.L., Hollingsworth, M.L., Crandall-Stotler, B., Davis, E.C., Engel, J., von Konrat, M., Cooper, E.D., Patiño, J., Cox, C.J., Vanderpoorten, A., Shaw, A.J., 2014. Extant diversity of bryophytes emerged from successive post-Mesozoic diversification bursts. Nat. Commun. 5, 5134. http://dx.doi.org/10.1038/ncomms6134.

Li, F.-W., Villarreal, J.C., Kelly, S., Rothfels, C.J., Melkonian, M., Frangedakis, E., Ruhsam, M., Sigel, E.M., Der, J.P., Pittermann, J., Burge, D.O., Pokorny, L., Larsson, A., Chen, T., Weststrand, S., Thomas, P., Carpenter, E., Zhang, Y., Tian, Z., Chen, L., Yan, Z., Zhu, Y., Sun, X., Wang, J., Stevenson, D.W., Crandall-Stotler, B.J., Shaw, A. J., Deyholos, M.K., Soltis, D.E., Graham, S.W., Windham, M.D., Langdale, J.A., Wong, G.K.-S., Mathews, S., Pryer, K.M., 2014. Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. Proc. Natl. Acad. Sci. USA 111, 6672–6677. http://dx.doi.org/10.1073/pnas.1319929111.

Li, Z., Baniaga, A.E., Sessa, E.B., Scascitelli, M., Graham, S.W., Rieseberg, L.H., Barker, M.S., 2015. Early genome duplications in conifers and other seed plants. Sci. Adv. 1, e1501084. http://dx.doi.org/10.1126/sciadv.1501084.

Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequences of duplicate genes. Science 290, 1151–1155.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., Van de Peer, Y., 2005. Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. USA 102, 5454–5459. http://dx.doi.org/10.1073/pnas.0501102102.

Magill, R.E., 2014. Moss diversity: new look at old numbers. Phytotaxa.

Merget, B., Wolf, M., 2010. A molecular phylogeny of Hypnales (Bryophyta) inferred from ITS2 sequence-structure data. BMC Res. Notes 3, 320. http://dx.doi.org/10.1186/1756-0500-3-320.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30, i541–i548. http://dx.doi.org/10.1093/bioinformatics/btu462.

Moore, R.C., Purugganan, M.D., 2005. The evolutionary dynamics of plant duplicate genes. Curr. Opin. Plant Biol. 8, 122–128. http://dx.doi.org/10.1016/j.pbi.2004.12.001.

Newton, A., Wikstrom, N., Bell, N., Lowe Forrest, L., Ignatov, M., 2007. Dating the diversification of the pleurocarpous mosses. In: Newton, A.E., Tangney, R.S. (Eds.), Pleurocarpous Mosses: Systematics and Evolution, Systematics and Evolution. CRC Press, pp. 337–366. http://dx.doi.org/10.1201/9781420005592.ch17.

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS ONE 5, e9490. http://dx.doi.org/10.1371/journal.pone.0009490.

Rensing, S.A., Ick, J., Fawcett, J.A., Lang, D., Zimmer, A., Van de Peer, Y., Reski, R., 2007. An ancient genome duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens. BMC Evol. Biol. 7, 130. http://dx.doi.org/10.1186/1471-2148-7-130.

Schneider, H., Schuettpelz, E., Pryer, K.M., Cranfill, R., Magallón, S., Lupia, R., 2004. Ferns diversified in the shadow of angiosperms. Nature 428, 553–557. http://dx.doi.org/10.1038/nature02361.

Shaw, A.J., Cox, C.J., Goffinet, B., Buck, W.R., Boles, S.B., 2003. Phylogenetic evidence of a rapid radiation of pleurocarpous mosses (Bryophyta). Evolution 57, 2226–2241. http://dx.doi.org/10.2307/3448774.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. http://dx.doi.org/10.1093/bioinformatics/btv351.

Smith, S.A., Dunn, C.W., 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24, 715–716. http://dx.doi.org/10.1093/bioinformatics/btm619.

Smoot, E.L., Taylor, T.N., 1986. Structurally preserved fossil plants from Antarctica: II. A permian moss from the transantarctic mountains. Am. J. Bot. 73, 1683. http://dx.doi.org/10.2307/2444234.

Spangler, J.B., Subramaniam, S., Freeling, M., Feltus, F.A., 2012. Evidence of function for conserved noncoding sequences in *Arabidopsis thaliana*. New Phytol. 193, 241–252. http://dx.doi.org/10.1111/j.1469-8137.2011.03916.x.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. http://dx.doi.org/10.1093/bioinformatics/btu033.

Stech, M., Frey, W., 2008. A morpho-molecular classification of the mosses (Bryophyta). Nova Hedw. 86, 1–21. http://dx.doi.org/10.1127/0029-5035/2008/0086-0001.

Supek, F., Bošnjak, M., Škunca, N., Šmuc, T., 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE 6, e21800. http://dx.doi.org/10.1371/journal.pone.0021800.

Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucl. Acids Res. 34, W609–W612. http://dx.doi.org/10.1093/nar/gkl315.

Yang, Y., Moore, M.J., Brockington, S.F., Soltis, D.E., Wong, G.K.-S., Carpenter, E.J., Zhang, Y., Chen, L., Yan, Z., Xie, Y., Sage, R.F., Covshoff, S., Hibberd, J.M., Nelson, M.N., Smith, S.A., 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. Mol. Biol. Evol. http://dx.doi.org/10.1093/molbev/msv081.

Yang, Y., Smith, S.A., 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. Mol. Biol. Evol. 31, 3081–3092. http://dx.doi.org/10.1093/molbev/msu245.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591. http://dx.doi.org/10.1093/molbev/msm088.

Zhang, Z., Li, J., Zhao, X.-Q., Wang, J., Wong, G.K.-S., Yu, J., 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genom. Proteom. Bioinform. 4, 259–263. http://dx.doi.org/10.1016/S1672-0229(07)60007-2.

The LINNEAN SOCIETY of London BIOLOGICAL Journal of the Linnean Society

Biological Journal of the Linnean Society, 2015, **115**, 96–113. With 4 figures

Genetic diversity, sexual condition, and microhabitat preference determine mating patterns in *Sphagnum* (Sphagnaceae) peat-mosses

MATTHEW G. JOHNSON* and A. JONATHAN SHAW

Biology Department, Duke University, 130 Science Drive, Box 90338, Durham, NC 27708, USA

Received 26 September 2014; revised 19 December 2014; accepted for publication 20 December 2014

In bryophytes, the possibility of intragametophytic selfing creates complex mating patterns that are not possible in seed plants, although relatively little is known about patterns of inbreeding in natural populations. In the peat-moss genus *Sphagnum*, taxa are generally bisexual (gametophytes produce both sperm and egg) or unisexual (gametes produced by separate male and female plants). We sampled populations of 14 species, aiming to assess inbreeding variation and inbreeding depression in sporophytes, and to evaluate correlations between sexual expression, mating systems, and microhabitat preferences. We sampled maternal gametophytes and their attached sporophytes at 12–19 microsatellite loci. Bisexual species exhibited higher levels of inbreeding than unisexual species but did generally engage in some outcrossing. Inbreeding depression did not appear to be common in either unisexual or bisexual species. Genetic diversity was higher in populations of unisexual species compared to populations of bisexual species. We found a significant association between species microhabitat preference and population genetic diversity: species preferring hummocks (high above water table) had populations with lower diversity than species inhabiting hollows (at the water table). We also found a significant interaction between sexual condition, microhabitat preference, and inbreeding coefficients, suggesting a vital role for species ecology in determining mating patterns in *Sphagnum* populations. *Biological Journal of the Linnean Society*, 2015, **115**, 96–113.

ADDITIONAL KEYWORDS: bryophytes – haploid – inbreeding depression – mating systems.

INTRODUCTION

A central issue in plant reproductive biology is intra- and interspecific variation in inbreeding, and the avoidance of inbreeding depression. In angiosperms, self-incompatibility systems and dioecy both reduce the probability of mating between gametes produced by the same sporophyte (i.e. intergametophytic selfing). The most common explanation for this avoidance of self-fertilization is the accumulation of slightly deleterious recessive alleles, especially in outcrossing populations (Nei, Maruyama & Chakraborty, 1975; Charlesworth & Charlesworth, 1999). Inbreeding 'exposes' these alleles in homozygotes that carry reduced fitness (inbreeding depression). However, in populations where inbreeding is common, deleterious alleles may be purged from the population, and subsequent inbreeding has a reduced fitness cost (Lande & Schemske, 1985; Goodwillie, Kalisz & Eckert, 2005). General surveys confirm a wide variety of outcrossing rates in angiosperms, from obligate outcrossing to obligate selfing and every rate inbetween (Barrett, 2003).

Homosporous spore-producing plants, such as ferns and bryophytes, are capable of an additional type of selfing not possible in seed plants. Intragametophytic selfing (i.e. the union of genetically identical sperm and egg) produces a completely homozygous offspring in a single round of mating. In ferns, the rate of intragametophytic selfing is generally low (Soltis & Soltis, 1992) and, in some species, is prevented by the production of chemicals (antheridiogens) that block production of sperm in bisexual gametophytes (i.e. those producing both antheridia and archegonia, Chiou & Farrar, 1997). Therefore, there may also be

*Corresponding author. E-mail: mgj4@duke.edu

selective pressure to avoid intragametophytic selfing, even though exposure of deleterious alleles in homozygous sporophytes would likely result in these alleles being quickly purged from the population (Hedrick, 1987).

This is especially true for bryophytes because of their haploid-dominant life cycle. Genes co-expressed in both the haploid and diploid stages are likely to experience less genetic load than genes expressed in just the diploid stage (Shaw & Beer, 1997; Joseph & Kirkpatrick, 2004). However, compared to seed plants, relatively few studies have tested theoretical predictions about the cost of inbreeding in natural populations of bryophytes. A few have made indirect inferences about mating patterns from genetic diversity (Stoneburner, Wyatt & Odrzykoski, 1991; Shaw, 2009), although only one (Eppley, Taylor & Jesson, 2007) compared levels of inbreeding in moss species with uni- versus bisexual gametophytes in natural populations. They documented significantly higher inbreeding coefficients in bisexual species.

There is also a paucity of data regarding inbreeding depression in bryophytes. Taylor, Eppley & Jesson (2007) found sporophytic inbreeding depression in a species with unisexual gametophytes (i.e. producing archegonia or antheridia but not both) but not in a species with bisexual gametophytes. Jesson *et al*. (2012) did not observe reduced sporophyte fitness in self-fertilized sporophytes of *Atrichum undulatum* (Hedw.) P. Beauv., a species that produces both unisexual and bisexual gametophytes. Szövényi, Ricca & Shaw (2009) found a significant association between sporophyte size [correlated with fitness (spore number)] and heterozygosity in *Sphagnum lescurii* Sull., a unisexual species.

Even if inbreeding depression is a significant selective pressure in bryophytes, their mating patterns may also be influenced by abiotic environmental factors. Similar to all seed-free plants, bryophyte sperm must swim to the egg to effect fertilization. Sperm dispersal distance is generally quite limited (Wyatt, 1977; Bisang, Ehrlén & Hedenäs, 2004) and strongly dependent upon water availability (Shortlidge, Rosenstiel & Eppley, 2012). Species that prefer dryer habitats are more likely to experience self-fertilization as a result of limited sperm dispersal. Therefore, fertilization success may depend on species ecology, especially for unisexual species, where intragametophytic selfing is not possible.

Sphagnum L. (peatmosses) presents an excellent case study for investigating the influences of sexual condition and ecology on mating behaviour in a group of closely-related species because they vary in gametophyte sexuality (uni- versus bisexual) and microhabitat. An early diverging lineage of the mosses (Bryophyta), *Sphagnum* includes approximately 250–400 species worldwide, organized in six monophyletic groups that have been classified as subgenera (Shaw *et al*., 2010). *Sphagnum* is best known for its prominence in Northern Hemisphere peatlands that have huge impacts on biogeochemistry and global climate (Wieder & Vitt, 2006; Tuba, Slack & Stark, 2011). Twenty or more species can co-occur in wetland communities and many species are ecologically differentiated in relation to microenvironmental variation within those peatlands (Rydin & Jeglum, 2013).

As with all bryophytes, the sporophyte generation is short lived and remains attached to the maternal gametophyte throughout its life. It is therefore simple to determine the maternal haploid component of the diploid genotype, and to infer the paternal haploid genotype by subtraction (Szövényi *et al*., 2009). Previous studies have demonstrated a high correlation across multiple species between the diameter of a *Sphagnum* sporophyte capsule (sporangium) and the number of spores that it contains (Sundberg & Rydin, 1998), presenting an easily measured proxy for fitness in *Sphagnum* sporophytes. Based on studies of seed plants, we can hypothesize that bisexual species are more inbred than unisexual species, that unisexual but not bisexual species might exhibit inbreeding depression, and (in bryophytes) that water availability impacts outcrossing rates.

Of the 91 species of *Sphagnum* that occur in North America, 14 are known to have bisexual gametophytes, 58 have unisexual gametophytes, and the rest have unknown sexual conditions (McQueen & Andrus, 2009). Studies of mating patterns in *Sphagnum* are facilitated by a set of microsatellite markers that amplify in almost every species in the genus without evidence of ascertainment bias (Shaw *et al*., 2008a; Shaw, Terracciano & Shaw, 2009; Karlin *et al*., 2010, 2011). Microhabitat preference (in the sense of very narrow, species-specific, realized niches) likely plays a role in determining mating patterns in *Sphagnum*. Species are generally differentiated with respect to their ranges along two ecological gradients: a mineotrophic gradient (pH and other micronutrients) and a hydrological gradient (height above the water table) (Vitt & Slack, 1984; Andrus, 1986). The hydrological gradient is of particular interest because *Sphagnum* species that generate hummocks high above the water table are likely to experience limited water availability for the dispersal of sperm.

In the present study, we investigated 18 populations across 14 species of *Sphagnum* aiming to address three questions. (1) What is the diversity of mating patterns in *Sphagnum*? We tested this using standard statistics to measure genetic diversity and inbreeding. (2) Is inbreeding depression common in populations of unisexual species and absent in populations of bisexual species, as predicted by population

genetics theory? (3) Are there connections between mating patterns and sexual condition or microhabitat preference?

MATERIAL AND METHODS

Bryophyte life cycle

Most moss species (approximately 60%) have separate sexes, comprising unisexual gametophytes that produce either sperm or eggs but not both (Wyatt & Anderson, 1984). Many species (approximately 30%) have bisexual gametophytes that can produce both archegonia (containing eggs) and antheridia (containing sperm). Sexual system (uni- versus bisexual gametophytes) is generally a species trait, although a minority of species can be polymorphic for gametophyte sexuality: gametophytes may be unisexual or bisexual. Haploid plants germinate from spores and form leafy gametophytes at maturity; many species also accomplish various forms of asexual reproduction via branching, fragmentation or the production of specialized vegetative propagules. At sexual maturity, males (Fig. 1A) produce sperm via mitosis in antheridia and females (Fig. 1B) produce eggs in archegonia, also via mitosis. Thus, all sperm produced by a single male gametophyte are genetically identical to each other and to the male gametophyte that produced them, with a comparable pattern for females. In mosses with bisexual gametophytes, all eggs and sperm are genetically identical to one another, as well as to the parental gametophyte.

In all species, water is required for the sperm to reach an egg-containing archegonium and effect fertilization. The resulting sporophyte begins development within the archegonium and remains attached to the maternal gametophyte for its entire life cycle. At maturity, the sporophyte forms haploid spores via meiosis within a single sporangium. Individual gametophytes often bear multiple sporophytes (Fig. 1B).

Sexual conditions in *Sphagnum*

When considering mating patterns, the possibility of intragametophytic selfing is arguably the most important distinction for homosporous plants. For the present study, intragametophytic selfing is defined as fertilization between genetically identical sperm and egg, producing offspring that are completely homozygous at every locus. Thus, a large vegetative clone may produce many ramets that have identical genotypes, although fertilization within genets is always intragametophytic. The critical distinction among species is the classification of sexual condition, and depends on whether a gametophyte is capable of producing both kinds of gametes ('bisexual') or only one type of gamete ('unisexual'). Although a few bryophyte species have variable sexual systems (Jesson *et al*., 2012), in *Sphagnum*, gametophytic sex expression is generally a fixed species trait (Szövényi *et al*., 2009; Ricca *et al*., 2011). Therefore, in the present study, we describe a species that produces male gametes and female gametes from the same plant as a 'bisexual species'. A species that produces gametes in separate male and female gametophytes is a 'unisexual species'.

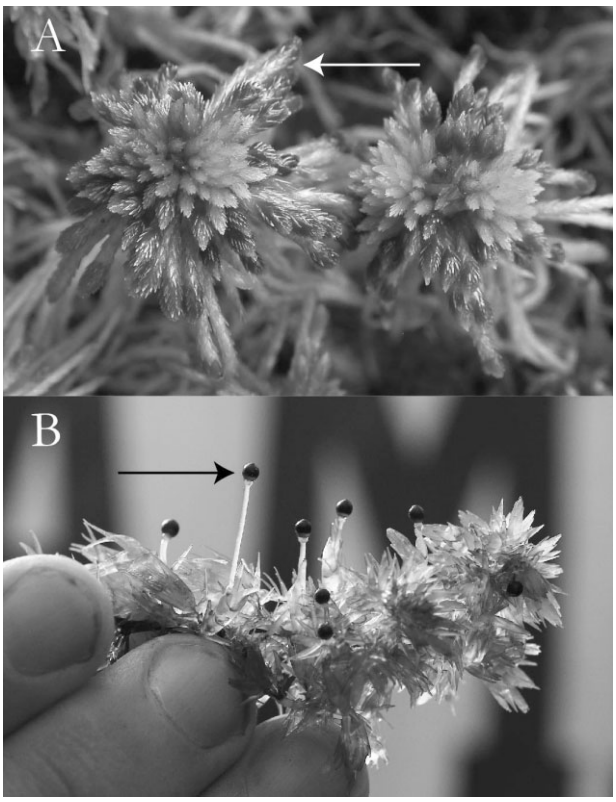

Figure 1. Unisexual gametophytes in *Sphagnum*. A, male gametophyte of *Sphagnum capillifolium* produces sperm in antheridia on specialized branches (white arrow). B, female gametophyte of *Sphagnum macrophyllum*, bearing a 'brood' of sporophytes (at least eight sporophytes are visible). The sporophyte is a single spheroid capsule (sporangium, black arrow). Photos courtesy of Blanka Shaw.

We classified each species in the present study as 'unisexual' or 'bisexual' based on personal observation of sexual conditions in the field, including thousands of *Sphagnum* specimens personally collected on five continents spanning more than a decade. Our classifications generally agree with two major monographs of *Sphagnum* (Crum, 1984; McQueen & Andrus, 2009); one notable exception is discussed below.

Species and population sampling

In general, our sampling of sporophyte-producing populations of *Sphagnum* was opportunistic. Plants

with sporophytes were collected when encountered during field work, and so we include multiple population samples for some species but only one population for others. We chose to maximize numbers of species rather than infraspecific populations, although additional data be will be valuable in the future to assess mating pattern variation among populations within species.

When a population with sporophytes was identified, our sampling strategy depended on the nature of the population. In populations where the sporophytes were spread across a large area, we sampled gametophytes and sporophytes along transects. In other cases, when the populations consisted of one or a few patches only, we sampled from within each patch. Because we had no information a priori about the genetic structure of these populations, our sampling strategy aimed at collecting a minimum number of maternal ramets. Populations are primarily from sites in Maine, Alaska, British Columbia, and the south-eastern USA. Collection and DNA voucher information are provided in the Appendix. All vouchered specimens were deposited in the Duke University Herbarium, and are identified as vouchers in the herbarium database (http://plantdb.biology.duke.edu:8080/BryoFullSearch/advanced.jsp).

We sampled populations from species in each of the major subgenera of *Sphagnum*, including both bisexual species and unisexual species. Although allopolyploidy is a common feature in *Sphagnum*, all fourteen species sampled have haploid gametophytes (no fixed heterozygosity in gametophytes). We sampled no more than two populations of each species; most species are represented by a single population. Because of this, our results are not intended to be extrapolated to represent the 'mating system' of particular species. Many previous studies have demonstrated that mating systems can vary within species (Goodwillie *et al*., 2005). Instead, our results are intended to show, across species, the relationship between sexual condition and mating patterns.

Several decades of peatland ecology research show that individual *Sphagnum* species have narrow microhabitat ranges and, in particular, are differentiated along a hydrological gradient (Vitt & Slack, 1984; Andrus, 1986; Gignac, 1992). This differentiation can be generally described as species preferring either a 'hollow' habitat (i.e. living right at the water table or even aquatically) or a 'hummock' habitat (i.e. forming high mounds or cushions well above the water table) (Rydin and Jeglum, 2013).

We assigned each species in the present study to a microhabitat group based on data collected from peatland ecological surveys that measured species preferences on the hummock–hollow gradient (Johnson *et al*., 2015b). From these data, we designated five 'hummock' species (mean height above water table > 10.0 cm): *Sphagnum angustifolium* (Warnst.) C.E.O. Jensen, *Sphagnum austinii* Sull., *Sphagnum fallax* H. Klinggr., *Sphagnum magellanicum* Brid., *Sphagnum squarrosum* Crome and four 'hollow' species: *Sphagnum compactum* Lam. & DC., *Sphagnum cuspidatum* Ehrh. ex Hoffm., *Sphagnum pulchrum* (Lindb.) Warnst., and *Sphagnum tenellum* (Brid.) Brid. The remaining species included in the present study did not have data in the ecological surveys; however, *Sphagnum macrophyllum* Bernh. ex Brid. and *Sphagnum cribrosum* Lindb. are both aquatic species that grow in open water (Anderson *et al*., 2009), and are classified here as 'hollow', whereas *Sphagnum molle* Sull. and *Sphagnum strictum* Sull. both form cushions, and are classified here as 'hummock' species.

Genotyping

For DNA extractions of gametophyte tissue, we sampled a portion of the gametophyte's capitulum (the dense cluster of branches at the apex of *Sphagnum* plants). In addition, all sporophytes attached to the female gametophyte were sampled. Extractions followed a CTAB protocol (Shaw, Cox & Boles, 2003). In preparation for polymerase chain reaction amplification, we diluted the DNA of gametophytes 7 : 1, and diluted sporophytes 2 : 1.

Plants were genotyped at nineteen microsatellite loci using previously described protocols (Shaw *et al*., 2008a). Primers were multiplexed in five sets. Set 1: p17, p22, p65, p78; Set 2: p1, p7, p12, p68; Set 3: p4, p10, p30; Set 4: p18, p19, p29, p93; and Set 5: p9, p14, p20, p56. Locus designations follow Shaw *et al*. (2008a). These microsatellite loci have been shown to be variable in a multitude of *Sphagnum* species from every subgenus.

In the present study, not all loci were successful in each species but, using this standard set of loci, we ensured that at least 10–12 loci consistently amplified in each species. Loci were also discarded if multiple individuals had more than one allele in the haploid stage or more than two in the diploid stage (a total of three loci across all populations were discarded). Missing data for within-species matrices ranged from 2.7% to 15.9%. Microsatellite allele tables were submitted to the Dryad data repository.

To assess paternity, we first inferred the haploid paternal genotype from every sporophyte by subtracting the haploid genotype of the maternal gametophyte. A custom PYTHON script that accounted for missing data was used to assign multilocus paternal microsatellite genotypes to genets. The script began by sorting samples with no missing data into multilocus genotypes. If a sample could be unambigu-

ously assigned to just one multilocus genotype despite missing data, it was retained; otherwise, it was discarded.

Mating patterns: statistics

To characterize mating patterns in *Sphagnum*, we focused on two main properties of the sporophyte generation in each species: genetic diversity and inbreeding. Sporophytes attached to the same maternal gametophyte necessarily share half of their multilocus genotype, and are not therefore independent draws from mating events in the population. Particularly fecund maternal gametophytes (those with many sporophytes) will introduce a bias when estimating allele frequencies.

To partially address this bias, we generated 1000 'subsets' of each population. Each subset contained a random draw of one sporophyte per maternal brood. For example, imagine a population with three maternal gametophytes (A, B, and C), each of which bears four sporophytes (numbered 1–4). A subset population would contain one sporophyte randomly drawn from each mother, such as: [A1, B2, C3]. We calculated diversity and inbreeding statistics on this subset population. We then repeated this procedure by randomly drawing another sporophyte (with replacement) from each mother and recalculating each statistic on this new subset, such as: [A3, B2, C4]. This was carried out a total of 1000 times, resulting in a range of values for each population statistic.

We calculated three measures of sporophyte genetic diversity at each locus. The included the effective number of alleles (also referred to as Simpson's Index: $1/\Sigma(p_i^2)$, where p_i is the frequency of the i_{th} allele at the locus), Shannon's Diversity Index [$1/\Sigma(p - \ln(p)$], and the inbreeding coefficient [$F_{IS} = 1 - (H_O/H_E)$, where H_O is the observed heterozygosity and H_E is the expected heterozygosity in the subsampled population].

We used sporophyte diameter as a proxy for sporophyte fitness. Sundberg & Rydin (1998) demonstrated a correlation between the diameter of a *Sphagnum* sporophyte and the number of spores it contains. Sundberg and Rydin found that this correlation holds for a phylogenetically diverse set of eight *Sphagnum* species, including three sampled in the present study. This phenotypic measurement allows for a simple estimate of potential reproductive output for the sporophyte without damaging the tissue necessary for genotyping the sporophyte. Other studies have also used the height of the sporophyte as a proxy for fitness in bryophytes (Eppley *et al*., 2007; Taylor *et al*., 2007; Jesson *et al*., 2012) because sporangial height may correlate with dispersal ability. However, each of those studies focused on mosses in the class Bryopsida, which have sporangia exerted on seta derived from sporophytic tissue. By contrast, *Sphagnum* sporangia are raised on a pseudopodium derived from haploid maternal gametophyte tissue, and thus height would be an inappropriate proxy for fitness of the diploid phase in *Sphagnum*. Sporophyte diameter was also used by Szövényi *et al*. (2009) to assess the presence of inbreeding depression in *S. lescurii*.

For each maternal haploid gametophyte sampled, we removed all attached sporophytes and photographed the sporophytes using a calibrated digital microscope camera (Olympus BX41; Olympus Imaging America, Inc.) under a dissecting microscope at × 10 power. We measured the diameter of each sporophyte from the image using MICROSUITE, version 5 (Olympus Imaging America, Inc.). Measurements were accurate to 0.1 μm.

We estimated the extent of sporophytic inbreeding depression using the linear regression of heterozygosity (percentage of loci with two alleles) on fitness (sporophyte size) using two methods. Our main approach was to test the significance of the correlation including all sporophytes in a population. However, to consider the possibility that sporophytes may not be independent draws from the population, we also used a pseudoreplication approach as above and determined the significance of the linear relationship in 1000 pseudoreplicates. The pseudoreplicate approach will have less power to detect significant relationships because they will be based on the sample size of mothers in the population (rather than of sporophytes).

RESULTS

Microsatellite diversity

Out of our panel of 19 microsatellite loci, we genotyped at least 12 loci in each population (mean loci scored: 13.5; range 12–15) (Table 1). The loci with successful amplification differed among species and, in one case (*S. strictum*), successful amplification of two loci (17 and 22) were specific to populations within species (Table 1). By using a standard panel of loci, we ensured amplification of sufficient loci necessary to assess genetic variability and mating patterns within populations. The allelic diversity of loci varied among species and subgenera of *Sphagnum*: for example, locus 1 was diverse in the subgenera *Cuspdiata* and *Acutifolia* but was generally invariable in the other subgenera. Across loci, allelic diversity was highest in subgenus *Subsecunda* (4.5–6.0 alleles/locus) (Table 1), the group that includes the species for which the microsatellite loci were originally designed *S. lescurii* (Shaw *et al*., 2008a). Allelic diversity was also high in *Cuspidata* (2.2–5.9 alleles/locus) and *Acutifolia* (1.8–3.3 alleles/locus), although

Table 1. Allelic diversity of sporophytes at 19 microsatellite markers across *Sphagnum* populations

Species	Population	*Subgenus*	1	4	7	9	10	12	14	17	18	19	20	22	29	30	56	65	68	78	93	Average unique
Sphagnum angustifolium	ME-S	*Acut*	3	NA	4	4	5	1	7	4	1	NA	4	3	NA	3	2	3	2	NA	NA	3.3
*Sphagnum molle**	SC-36	*Acut*	1	1	1	NA	3	1	3	2	3	NA	2	1	1	2	3	NA	1	NA	NA	1.8
*Sphagnum tenellum**	ME-W	*Cusp*	5	NA	6	2	4	NA	2	2	1	7	2	NA	NA	NA	2	1	1	2	NA	2.8
Sphagnum warnstorfii	AK-M	*Acut*	2	NA	1	5	4	2	6	NA	4	4	2	3	1	2	NA	NA	NA	NA	4	3.1
Sphagnum cuspidatum	GA-28	*Cusp*	6	3	6	12	12	1	15	7	2	NA	3	NA	NA	4	8	NA	3	1	NA	5.9
Sphagnum cupidatum	NC-JL	*Cusp*	2	3	5	5	5	1	6	4	1	NA	2	NA	NA	2	5	NA	1	1	NA	3.1
Sphagnum fallax	ME-S	*Cusp*	3	NA	5	NA	5	2	7	3	1	NA	4	2	3	2	5	3	1	NA	NA	3.3
Sphagnum fallax	ME-W	*Cusp*	5	4	11	NA	8	2	8	4	1	NA	4	3	5	4	NA	5	1	NA	NA	4.6
Sphagnum pulchrum	ME-C	*Cusp*	3	2	1	2	5	1	7	3	1	NA	2	1	NA	1	NA	1	NA	1	NA	2.2
Sphagnum compactum	AK-PC	*Rig*	1	NA	NA	1	1	NA	4	NA	1	1	6	NA	NA	1	5	1	1	1	1	1.9
Sphagnum compactum	AK-PH	*Rig*	1	NA	NA	1	1	NA	3	NA	1	1	3	NA	NA	1	3	1	1	1	1	1.5
*Sphagnum strictum**	SC-36	*Rig*	1	NA	1	1	1	1	2	NA	2	2	1	1	NA	2	NA	NA	1	1	NA	1.3
*Sphagnum strictum**	NC-LL	*Rig*	1	NA	1	1	1	1	1	1	1	1	1	NA	NA	1	NA	NA	1	1	NA	1.0
Sphagnum cribrosum	GA-32	*Subs*	NA	5	4	6	10	NA	8	4	3	NA	4	NA	2	NA	NA	2	3	3	NA	4.5
Sphagnum macrophyllum	SC-39	*Subs*	1	4	3	13	12	NA	11	5	9	6	5	NA	3	NA	NA	2	3	6	7	6.0
Sphagnum macrophyllum	SC-36	*Subs*	NA	4	5	8	12	NA	9	2	4	NA	6	NA	2	NA	NA	2	1	5	NA	5.0
Sphagnum austinii	VI	*Spha*	1	NA	2	2	3	2	1	2	2	1	1	1	1	3	NA	NA	NA	NA	1	1.6
Sphagnum magellanicum	NC-JL	*Spha*	2	1	2	2	2	1	2	1	2	NA	2	2	2	2	NA	2	2	NA	NA	1.8
*Sphagnum squarrosum**	AK-W	*Squa*	1	NA	1	2	2	NA	5	1	4	2	2	NA	NA	2	5	NA	3	1	3	2.4
		Average unique	2.3	3.0	3.5	4.2	5.1	1.3	5.6	3.0	2.3	2.8	2.9	1.9	2.2	2.1	4.2	2.1	1.6	2.0	2.8	

NA, locus did not amplify in the population. The mean number of unique alleles across populations (columns) or within populations (rows) is also indicated. Asterisks indicate bisexual species. Subgenus abbreviations: Acut, *Acutifolia*; Cusp, *Cuspidata*; Rig, *Rigida*; Subs, *Subsecunda*; Spha, *Sphagnum*; Squa, *Squarrosa*. Information on each population is ptovided in the Appendix.

it was much lower in populations of species in subgenera *Squarrosa* (2.4 alleles/locus), *Sphagnum* (1.6–1.8 alleles/locus), and *Rigida* (1.0–1.9 alleles/locus). Overall, there was a significant subgenus effect on within-population allelic diversity (analysis of variance: $F_5 = 7.1$, $P < 0.01$). However, as shown below, the subgenus effect is confounded by the effects of sexual condition and species ecology.

Bisexual populations

All seven populations of five bisexual *Sphagnum* species showed high levels of intragametophytic selfing and low genetic diversity (I) (Table 2). Inbreeding was lowest [$F_{IS} = 0.35$; 95% confidence interval (CI) = 0.24–0.47] in the AK-PH population of *S. compactum*, in which 46% of sporophytes were completely homozygous despite moderate genetic diversity in the population ($I = 0.63$, 95% CI 0.55–0.69). Intragametophytic selfing was highest in one population of *S. strictum* from North Carolina; 100% of sporophytes were homozygous at every locus and there was zero genetic diversity ($I = 0$) in this population. The other population of *S. strictum* we sampled, from South Carolina, had a low, but nonzero genetic diversity ($I = 0.19$, 95% CI = 0.12–0.30), and also had several sporophytes that were heterozygous at exactly one locus (locus p19). The two populations, which were approximately 80 miles apart, had fixed differences at several loci (results not shown). We also observed high intragametophyte selfing rates in the population of *S. molle*, despite a comparatively larger amount of genetic diversity in that species. Although the one population of *S. squarrosum* that we sampled had the highest genetic diversity among populations of bisexual species ($I = 0.81$, 95% CI = 0.75–0.88), it also had a very high inbreeding coefficient ($F_{IS} = 0.71$, 95% CI = 0.65–0.88). In most bisexual populations, homozygous sporophytes exhibit a large range in fitness that is equivalent to the ranges observed for heterozygous sporophytes (Fig. 2).

The linear regression of heterozygosity on sporophyte fitness was significant for only two populations, both of *S. compactum* (Fig. 1). In one population (AK-PH), 23 of 39 sporophytes (59%) were completely homozygous. These sporophytes were, on average, larger than heterozygous sporophytes in the population ($b = -25.1$, d.f. = 21, $r^2 = 0.12$, uncorrected $P < 0.05$), suggesting outbreeding depression. The pseudoreplication approach indicated significant linear relationships between sporophyte size and percent heterozygosity in 231 of 1000 pseudoreplicates of the AK-PH population (see Supporting information, Table S1).

By contrast, the other population of *S. compactum* (AK-PC) showed a strong positive association between heterozygosity and fitness (inbreeding depression, Fig. 2); homozygous sporophytes are significantly

Table 2. Genetic diversity and mating patterns in bisexual *Sphagnum* species

Species	*Sphagnum compactum*	*Sphagnum compactum*	*Sphagnum molle*	*Sphagnum squarrosum*	*Sphagnum strictum*	*Sphagnum strictum*	*Sphagnum tenellum*
Subgenus	Rig	Rig	Acut	Squ	Rig	Rig	Cusp
Population	AK-PH	AK-PC	SC-36	AK-W	SC-36	NC-LL	ME-W
Ecology	hol	hol	hum	hum	hum	hum	hol
Female gametophytes	16	10	10	23	8	6	15
Average sporophytes per brood	2.4	2.9	3.2	2.5	4.5	4.5	2.5
Sporophytes	39	29	32	58	36	48	37
Maternal genotypes	5	6	4	11	2	1	10
Inferred paternal genotypes	8	17	7	15	4	1	23
Ambiguous paternity	0	10	2	7	1	0	16
Homozygous sporophytes	23	14	28	42	32	48	23
Effective alleles (N_E)	1.54	1.9	1.52	2.22	1.23	1	1.78
Shannon's Diversity (I)	0.51	0.59	0.34	0.81	0.19	0	0.63
Inbreeding coefficient (F_{IS})	0.35	0.42	0.86	0.71	0.6	NA	0.41
Heterozygosity versus size	OUT	IN	NO	NO	NO	NA	NO

NA, there are no heterozygous sporophytes to calculate the value.
Significant relationships between sporophyte heterozygosity and size indicate sporophytic inbreeding depression (IN, positive correlation) or outbreeding depression (OUT, negative correlation). Only sporophytes that could be unambiguously assigned an inferred paternal genotype were included in the analysis; sporophytes that had to be discarded due to missing data are indicated for each population. Key to subgenera: Acut, *Acutifolia*; Cusp, *Cuspidata*; Sph, *Sphagnum*; Rig, *Rigida*. Key to species microhabitat preference: Hum = hummock, Hol = hollow. More information on the populations is provided in the Appendix.

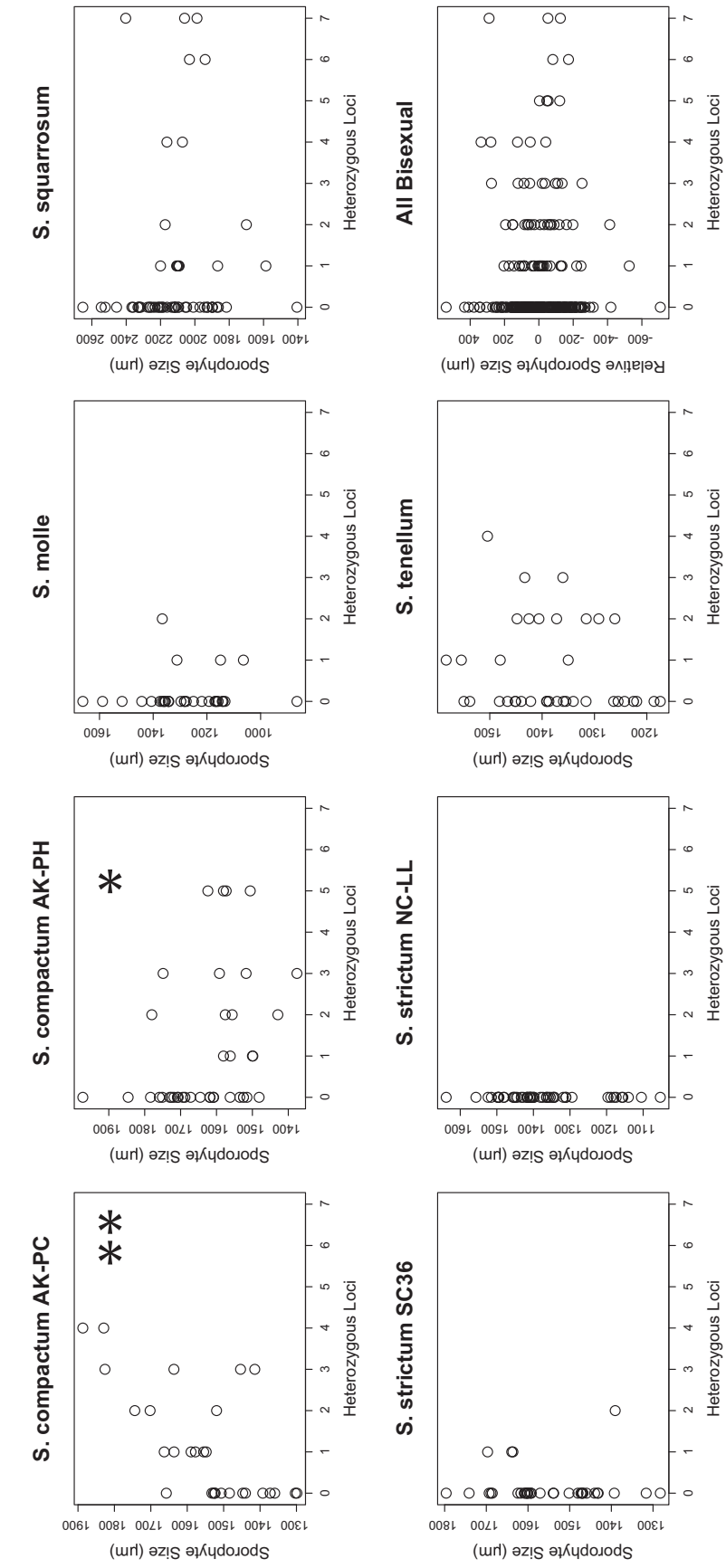

Figure 2. Relationship between heterozygosity and sporophyte size in seven populations of bisexual *Sphagnum* species. A significant positive correlation indicates inbreeding depression, whereas a significant negative correlation indicates outbreeding depression. Bottom right: sporophyte sizes have been standardized within populations (mean of zero) to allow comparison of all bisexual populations. Asterisks indicate significance of the correlations (*$P < 0.05$; **$P < 0.001$).

smaller than heterozygous sporophytes ($b = 74.1$, d.f. = 26, $r^2 = 0.43$, uncorrected $P < 0.001$). The pseudoreplication approach also identified a large percentage of pseudoreplicates with significant relationships (480; see Supporting information, Table S1). In this population, the 15 heterozygous sporophytes were sired by 15 unique inferred paternal genotypes. Six unique maternal genotypes were present, and five of these mothers raised sporophytes that were self-fertilized.

Unisexual populations

We observed a greater variance in genetic diversity and mating patterns among populations of *Sphagnum* species with separate sexes (Table 3). Five of the eleven populations have genetic diversity estimates (N_E and I) greater than any of the bisexual populations. One population, of *S. austinii*, showed no genetic variation at any locus among maternal gametophytes. Additional alleles appeared at two loci in the sporophytes but 42 of 48 sporophytes were completely homozygous. Excepting this population, genetic diversity was lowest in our population of *S. pulchrum* ($I = 0.49$, 95% CI = 0.43–0.56) and highest in the GA28 population of *S. cuspidatum* ($I = 1.29$, 95% CI = 1.23–1.36).

Populations varied from a mixed mating in the GA28 *S. cuspidatum* population ($F_{IS} = 0.56$, 95% CI = 0.50–0.63), to the population of *S. magellanicum* ($F_{IS} = -0.98$, 95% CI = −0.98 to −1.00) in which almost all sporophytes were heterozygous at all variable loci (Fig. 3). Seven of the eleven unisexual populations have significantly negative inbreeding coefficients. Excepting our population of *S. austinii*, the only unisexual population with any completely homozygous sporophytes was the GA28 population of *S. cuspidatum*; three sporophytes, all attached to the same maternal ramet, had zero heterozygosity.

There was a strong positive, significant relationship between genetic diversity (I) and inbreeding coefficient in unisexual populations (d.f. = 10, $r^2 = 0.44$, $P < 0.05$) (Fig. 4). The relationship was even stronger when considering the effective number of alleles (d.f. = 10, $r^2 = 0.53$, $P < 0.01$).

Inbreeding depression appeared to be absent in all but two unisexual populations, both of *S. macrophyllum* – the SC36 population has a stronger association between heterozygosity and fitness ($b = 30.5$, d.f. = 44, $r^2 = 0.15$, uncorrected $P < 0.01$) than the SC39 population ($b = 13.5$, d.f. = 273, $r^2 = 0.02$, uncorrected $P < 0.05$). Using the pseudoreplication approach, neither population showed strong assotiations between heterozygosity and size (see Supporting information, Table S1). A highly significant *negative* relationship (outbreeding depression) between heterozygosity and size was found in the population of *S. cribrosum* ($b = -64.7$, d.f. = 44, $r^2 = 0.31$, uncorrected $P < 0.0001$). The pseudoreplication approach revealed 533 out of 1000 pseudoreplicates had significant linear relationships (see Supporting information, Table S1).

In two populations, genetic diversity among the inferred paternal genotypes was low or absent. For the population of *S. magellanicum*, only two unique paternal genotypes were inferred; one of these was found in only two sporophytes, and differed from the other genotype at only one locus. In the population of *S. austinii*, almost all sporophytes were genetically identical and homozygous, and only three different parental genotypes could be inferred.

Effect of ecology and sexual condition on mating patterns

In addition to assignment based on sexual condition (bisexual versus unisexual), species were assigned to ecological groups based on microhabitat preference (hummock versus hollow). We investigated the relationship between each of the genetic diversity and mating pattern statistics, and found four associations related to ecology and sexual condition. First, as mentioned above, there was a significant correlation between genetic diversity and inbreeding coefficient, but only for unisexual populations (Fig. 4).

Genetic diversity (Shannon's Diversity) was significantly associated with sexual condition (analysis of variance: type III SS $F_1 = 9.83$, $P < 0.01$) (Fig. 4) and species ecology ($F_1 = 5.89$, $P < 0.05$), although there was no interaction effect ($F_1 = 0.11$, $P > 0.5$). A similar pattern was found for effective number of alleles per locus (results not shown). For inbreeding coefficient, there was a significant interaction effect between ecology and sexual condition ($F_1 = 6.5$, $P < 0.05$) (Fig. 4). F_{IS} tends to be higher in bisexual hummock populations than bisexual hollow populations, although the reverse is true for unisexual populations.

DISCUSSION

Inbreeding depression

Only three studies on bryophytes have previously investigated the relationship between inbreeding and sporophyte fitness. Taylor *et al*. (2007) found that heterozygosity was associated with reduced sporophyte height but not reduced spore output in a unisexual moss. In a population of *S. lescurii*, an aquatic unisexual species, Szövényi *et al*. (2009) found a significant increase in sporophyte size with increased heterozygosity (with a stronger linear relationship

Table 3. Genetic diversity and mating patterns in unisexual *Sphagnum* species

Species	*Sphagnum angustifolium*	*Sphagnum austinii*	*Sphagnum cribrosum*	*Sphagnum cuspidatum*	*Sphagnum cuspidatum*	*Sphagnum fallax*	*Sphagnum fallax*	*Sphagnum macrophyllum*	*Sphagnum macrophyllum*	*Sphagnum magellanicum*	*Sphagnum pulchrum*	*Sphagnum warnstorfii*
Subgenus	Cusp	Sph	Subs	Susp	Cusp	Cusp	Cusp	Subs	Subs	Sph	Cusp	Acut
Population	ME-S	VI	GA-32	GA-28	NC-JL	ME-S	ME-W	SC-36	SC-39	NC-JL	ME-C	AK-M
Ecology	hum	hum	hol	hol	hol	hum	hum	hol	hol	hum	hol	hum
Female gametophytes	7	24	9	18	8	7	10	12	41	6	8	18
Sporophytes	36	48	46	68	39	31	36	46	275	32	35	43
Average sporophytes per brood	5.1	2	5.1	3.7	4.9	4.4	3.6	6.7	5.3	5.3	4.4	2.4
Maternal genotypes	2	1	8	17	2	3	1	10	11	1	7	8
Inferred paternal genotypes	14	3	25	43	12	8	27	91	33	2	13	22
Ambiguous paternity	6	0	2	0	0	2	6	0	46	0	3	11
Homozygous sporophytes	0	42	0	3	0	0	0	0	0	0	0	0
Effective alleles (N_E)	1.78	1.51	2.73	3.78	2.13	1.96	1.90	3.18	2.65	2.01	1.77	2.35
Shannon's diversity (I)	0.58	0.48	1.08	1.29	0.79	0.63	0.74	1.17	1.04	0.67	0.49	0.85
Inbreeding coefficient (F_{IS})	−0.46	0.12	0.09	0.56	−0.14	−0.23	−0.32	0.12	−0.13	−0.98	−0.09	−0.06
Heterozygosity versus size	NO	NO	**OUT**	NO	NO	NO	NO	**IN**	**IN**	NO	NO	NO

Significant relationships between sporophyte heterozygosity and size indicate sporophytic inbreeding depression (IN, positive correlation), outbreeding depression (OUT, negative correlation) or no significant correlation. Only sporophytes that could be unambiguously assigned an inferred paternal genotype were included in the analysis; sporophytes that had to be discarded as a result of missing data are indicated for each population Key to subgenera: Acut, *Acutifolia*; Cusp, *Cuspidata*; Sph, *Sphagnum*; Subs, *Subsecunda*. Key to species microhabitat preference: Hum, hummock; Hol, hollow. More information on the populations is provided in the Appendix.

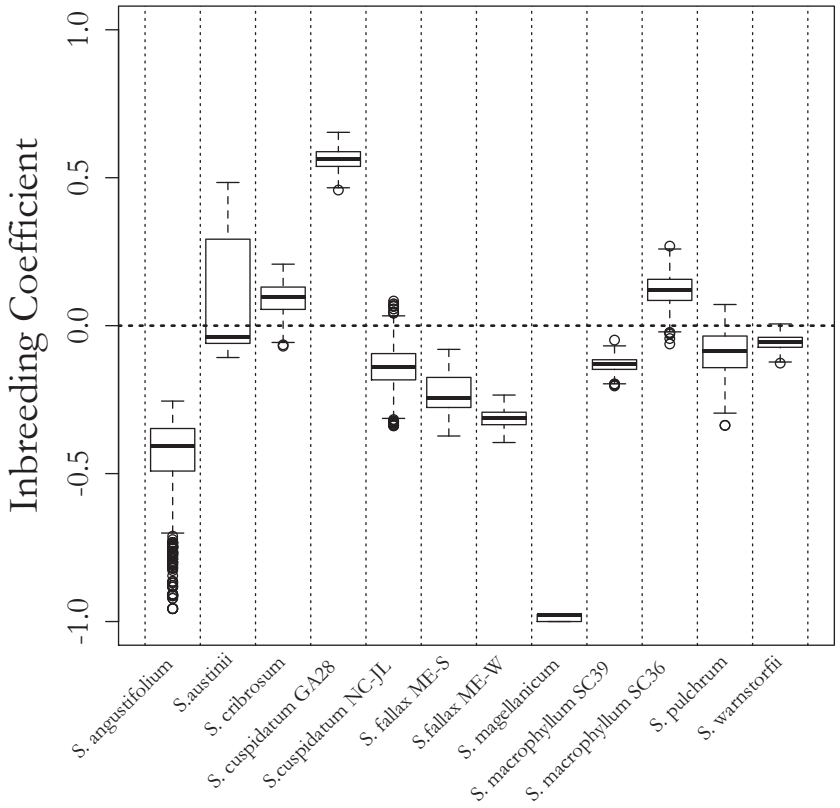

Figure 3. Variation in inbreeding coefficients (F_{IS}) among unisexual *Sphagnum* populations. Inbreeding coefficients were calculated by randomly sampling one sporophyte from each maternal gametophyte. The boxplot shows means and standard variations for 1000 subsamples from each population.

than is found in any population in our study, $b = 35.7$). Our data suggest that these studies cannot be extrapolated to a general conclusion that unisexual moss populations show inbreeding depression. Only two of our unisexual populations, both in the aquatic species *S. macrophyllum*, showed a significant relationship between sporophyte size and heterozygosity. In both, the inbreeding depression effect was smaller than found in *S. lescurii*. In addition, if we correct our *P*-values for multiple comparisons (Bonferonni correction), the significance of the relationship between heterozygosity and size is not maintained for any population. We would therefore caution against interpreting our findings as evidence for significant inbreeding depression in any population. More intense sampling may be necessary to detect weak inbreeding depression in unisexual *Sphagnum* species. Sporophytic inbreeding depression may be masked in some populations as a result of reproductive compensation or early sporophyte abortion; these possibilities require further study.

Most theoretical predictions suggest a small role for inbreeding depression in species capable of intragametophytic selfing (Hedrick, 1987; Stenøien & Såstad, 2001), and this is supported by our findings. If our results are typical for bisexual *Sphagnum* species, it is difficult to imagine a scenario in which inbreeding depression can be maintained in a population where almost 50% of the sporophytes are completely homozygous in each round of mating. For the lone bisexual population with evidence of inbreeding depression, it is possible that the effect we observed resulted from the specific combination of parental genotypes, rather than from heterozygosity per se, although this would require additional study with a larger sample size.

Several studies have suggested that inbreeding depression may not be a universal feature in outcrossing populations when there is substantial overlap in gene expression between the haploid gametophyte and diploid sporophyte generations (Shaw & Beer, 1997; Joseph & Kirkpatrick, 2004). Indeed, approximately 70% of genes expressed in the diploid stage of the bisexual moss *Funaria hygrometrica* are also expressed in the haploid stage, compared to just 3% in *Arabidopsis* (Szövényi *et al*., 2011). If the same is true for *Sphagnum*, then deleterious alleles could be purged from the population during the

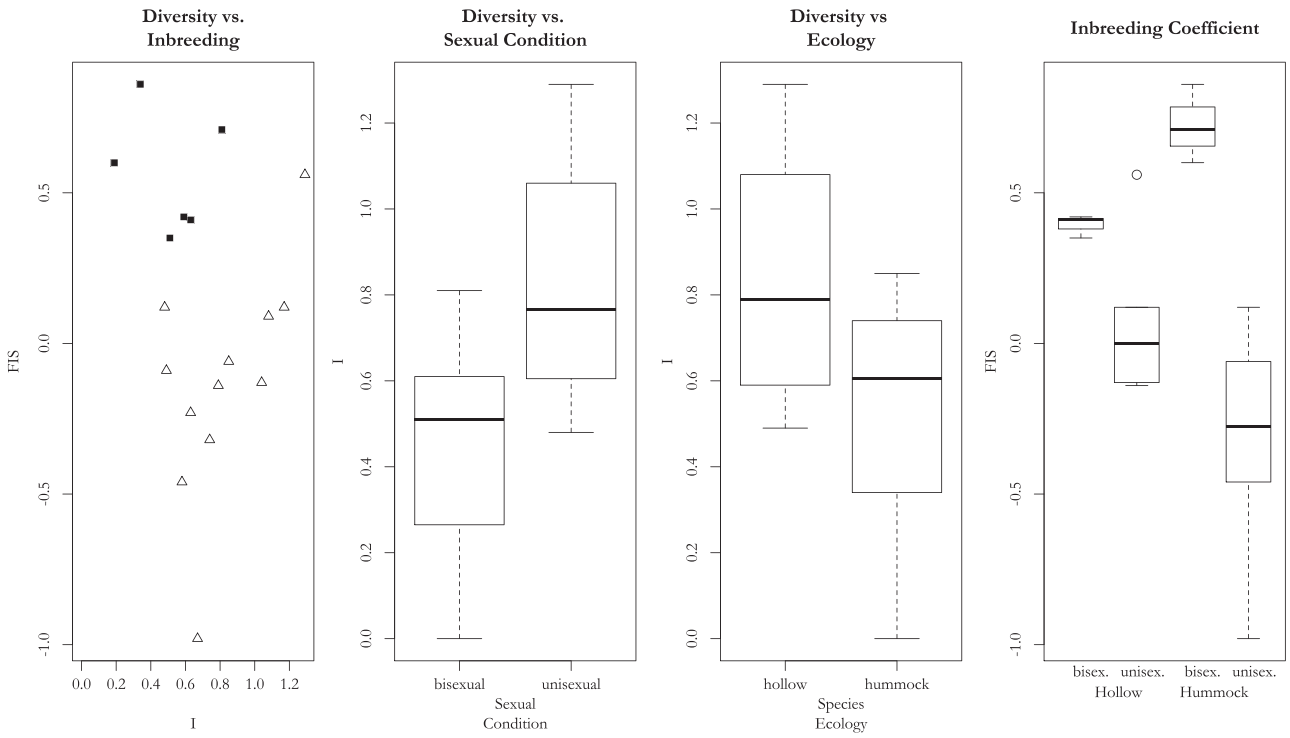

Figure 4. Effect of sexual condition and microhabitat preferences on mating pattern and genetic diversity. Far left: relationship between genetic diversity (*I*, Shannon's Allelic Diversity) and inbreeding coefficient (F_{IS}) for bisexual (circles) and unisexual (triangles) populations. Left centre: relationship between sexual condition (uni- versus bisexual) and genetic diversity. Right centre: relationship between species ecology (hummock versus hollow) and genetic diversity. Far right: interaction effect between sexual condition and ecology (hummock – hollow) versus F_{IS}.

perennial haploid stage, where there is no 'shielding' by dominance, and many unisexual mosses might therefore show no inbreeding depression.

Outbreeding depression (a negative relationship between heterozygosity and fitness) was observed in the AK-PH population of bisexual *S. compactum*, and in the unisexual species, *S. cribrosum*. Outbreeding depression is feasible in populations of predominantly selfing plants, because many generations of local adaptation may result in strong linkage disequilibrium, which would be broken up by hybridization between inbred lines (Fischer & Matthies, 1997; Edmands, 2006). If the intragametophytic selfing rate observed in our *S. compactum* population is typical, this would explain why heterozygous sporophytes are less fit, although a multi-generation study would be necessary to confirm this.

More typically, outbreeding depression is seen between distant populations, such as between populations of the moss *Ceratadon purpureus* from New York and Ecuador (McDaniel, Willis & Shaw, 2008). Genetic incompatibilities developed in isolation have been suggested as a type of postzygotic isolation important in speciation (Dobzhansky, 1950; Lynch, 1991). By contrast, we observed outbreeding depression in crosses that naturally occurred within a single population. Interestingly, the *S. cribrosum* population investigated in the present study is known to have the highest genetic diversity of any population in the species (Johnson *et al.*, 2012). In that previous study, we suggested that the GA32 population has been a source of genetic diversity for other populations of *S. cribrosum* in the eastern USA. However, if the population were instead a sink, receiving migrants from all over the distribution of *S. cribrosum*, it would make outbreeding depression more likely. Similarly, outbreeding depression was seen in natural immigrants to an inbred population of song sparrows (Marr, Keller & Arcese, 2002) and among genetically diverse individuals in a population of *Ipomopsis aggregate* (Pursh) V.E. Grant (Waser, Price & Shaw, 2000).

Overall, we find a lack of association between heterozygosity and fitness in many natural populations of *Sphagnum*. This is unexpected, given the high incidence of unisexual species in *Sphagnum* and other mosses. This suggests that, in contrast to flowering plants, selection for inbreeding avoidance mechanisms might be weak. Bryophytes have perennial haploid gametophytes that are exposed to

environmental selection pressures for a much longer proportion of the life cycle than the haploid gametophytes of flowering plants. This, coupled with relatively high co-expression of genes in the haploid and diploid moss generations (Szövényi *et al*. 2011), may generate weaker selection for inbreeding avoidance. Additional work on a diversity of moss species is needed to determine whether there is a consistent difference in mosses relative to flowering plants in the occurrence of inbreeding depression. Other metrics, such as germination rate and survival of spores from inbred sporophytes, would be useful to further characterize the effect of inbreeding on bryophyte fitness.

Sexual condition, mating patterns, and sex determination

We find sexual condition to be the major factor affecting mating patterns in *Sphagnum*. Most of the unisexual populations had inbreeding coefficients near or below zero, whereas the bisexual populations had significantly higher inbreeding coefficients. In unisexual populations, there was a strong correlation between genetic diversity and inbreeding, as predicted by theory (Charlesworth, Morgan & Charlesworth, 1993).This correlation was, however, absent in bisexual populations.

F_{IS} values indicate a high level of intragametophytic selfing in populations of bisexual *Sphagnum* species, consistent with values found in a limited survey of several other moss species (Eppley *et al*., 2007). Mixed mating (F_{IS} near 0.50) is a common feature in bisexual *Sphagnum* populations. This is in contrast to the theoretical predictions for angiosperms, in which outcrossing (F_{IS} near zero) and inbreeding (F_{IS} near 1) are the theoretical preferred stable states (Lande & Schemske, 1985). However, Lande & Schemske (1985) also postulate that reproductive compensation may lead to stable mixed mating when the threat of inbreeding depression is low. In these cases, inbreeding is advantageous for local populations but outcrossing is preferable for colonization of new areas with potentially different selective pressures (Lande & Schemske, 1985). More recent surveys of natural populations of angiosperms have revealed a wide variety of mating regimes, frequently variable within species (Barrett, 2003).

For species with separate sexes, a low effective population size may generate excesses of heterozygosity because of binomial sampling error (Rasmussen, 1979). A female *Sphagnum* plant in a species with unisexual gametophytes must, by necessity, mate with a different genetic individual. If the overall genetic diversity of the population is low, it is possible that many matings occur between two very different genotypes, generating highly heterozygous offspring. In spore-producing bryophytes characterized by highly effective dispersal, individual populations of gametophytes may be highly diverse, representing genotypes from across the continental range of the species (Ramaiya *et al*., 2010; Johnson *et al*., 2012), and clonal propagation rather than spore-mediated reproduction within populations may be most important for population maintenance (Sundberg, Hansson & Rydin, 2006). In this scenario, the binomial sampling issue may be accentuated, and the inbreeding coefficient (F_{IS}) may be strongly negative.

Traditionally, F_{IS} is often interpreted as the 'probability of identity by descent', although this definition complicates interpretation of negative values of F_{IS}. Instead, it is more meaningful to think in terms of Wright's original definition of the inbreeding coefficient as the 'correlation between uniting gametes' (Wright, 1922, 1965). A high correlation results in inbred offspring, whereas a low correlation results in highly heterozygous offspring. This may be the more appropriate interpretation for bryophytes, especially given low levels of genetic diversity, coupled with extensive clonal reproduction within populations.

Of the unisexual species studied here, two populations contained homozygous sporophytes: the GA28 population of *S. cuspidatum*, and (the only population we sampled of) *S. austinii*. These outliers could be explained by either incorrect assignments of their sexuality, or very low genetic diversity, limiting the ability to distinguish maternal and paternal genotypes in the sporophytes. For *S. austinii*, the latter is more likely to be correct because there was only one allele detected at each locus and all sporophytes were 100% homozygous. A broader survey of *S. austinii* in North America has revealed very low diversity compared to other members of subgenus *Sphagnum* (M. Kyrkjeeide, pers. comm.), and a survey of *S. austinii* in Europe with enzyme markers also revealed low genetic diversity (Thingsgaard, 2002). There are no known bisexual species within subgenus *Sphagnum*, making misdiagnosis of sexual condition unlikely, and so we conclude that the higher F_{IS} value for *S. austinii* likely reflects very low levels of genetic diversity rather than misdiagnosis of its sexual condition.

Many of these same microsatellites were used in other studies of species across all subgenera of *Sphagnum* (Ricca *et al*., 2008; Shaw *et al*., 2008b, 2009; Szövényi *et al*., 2009; Karlin *et al*., 2010; Stenøien *et al*., 2011; Johnson *et al*., 2012). In one study, 15 of these loci were used in a global study of the bisexual species *Sphagnum subnitens* Russ. & Warnst. (Karlin *et al*., 2011). Karlin *et al*. (2011) found high genetic diversity in European populations but a single multilocus haploid genotype in western North America.

Their conclusion was that the microsatellites were sufficiently variable to distinguish genetic variability in that species, and that the single genotype was indeed a very large clone. Taken together, these studies suggest no evidence of ascertainment bias, and that the microsatellite loci contain sufficient polymorphism to assess genetic diversity with populations.

By contrast, the GA28 population of *S. cuspidatum* exhibited a high degree of genetic variability. Despite this, the population inbreeding coefficient was more similar to bisexual populations than to other unisexual populations. In addition, three sporophytes, all attached to the same maternal ramet, were completely homozygous. One explanation is that these sporophytes are heterozygous but undifferentiated at these loci; given the high number of alleles at each locus in the population, this appears to be unlikely. An alternative explanation is that *S. cuspidatum* is not always unisexual. There has been some disagreement about the sexual condition of this species. McQueen & Andrus (2009) state that *S. cuspidatum* is dioicous (unisexual gametophytes), whereas Crum (1984) described the species as 'monoicous, and apparently also dioicous'.

Although sexual condition is generally fixed within species, a few species have been recorded as having both unisexual and bisexual gametophytes ('polyoicous', c.f. Cronberg, 1991). Although none of the species identified as polyoicous were included in the present study, it is safe to say that the determination of sexual condition in *Sphagnum* is poorly understood, and it may be polymorphic within some species or populations. Gene expression analysis, along with careful tracking of individual ramets within populations, would be very beneficial for understanding genetic differences between male and female gametophytes, and to assess whether a single gametophyte can vary temporally in sexual expression. Our results suggest that *S. cuspidatum* would be a very good choice for this type of study.

Microhabitat preference affects mating patterns

Sphagnum species are known to prefer narrow ranges of microhabitats within peatlands (Vitt & Slack, 1984; Andrus, 1986), including the tendency of each species to prefer either 'hollow' conditions (aquatic microhabitats at the water table) or 'hummock' conditions (cushions forming high above the water table). Water availability (necessary for fertilization in bryophytes) is reduced in hummock habitats, and we predicted that this would reduce fertilization probabilities, resulting in lower genetic diversity and higher inbreeding rates. We confirmed this prediction, finding that populations of species preferring hummock habitats have significantly reduced genetic diversity compared to populations of species preferring hollow habitats.

We found the inbreeding coefficient to be significantly correlated with species microhabitat distribution along the hummock–hollow gradient (Fig. 4). Rather than direct effects, there is an interaction with sexual condition. Inbreeding coefficients were higher in bisexual hummock-preferring populations than bisexual hollow populations. The limited water availability atop hummocks means that sperm, which must swim to effect fertilization, may have fewer opportunities for intergametophytic mating, especially if genetic diversity within hummocks is low. Our data support this because the same inferred paternal genotypes were rarely found in sporophytes attached to mothers from different parts of the population in hummock-preferring species.

We found the opposite pattern in unisexual populations: hummock-preferring populations had lower inbreeding coefficients than hollow populations (Fig. 4). Intriguingly, this pattern could be explained by the same phenomenon of limited water availability. As discussed earlier, an overabundance of one type of gamete in the mating pool could increase observed heterozygosity above expected, generating negative values of F_{IS}. By definition, a unisexual gametophyte must mate with a different genet to produce sporophytes; if these opportunities are limited to local genotypes, it could explain the extremely negative values found in unisexual hummock populations.

Bisexual species tend to be colonizers (Sundberg *et al.*, 2006); a disproportionate percentage of hummock preferences among bisexual species (such as *S. molle* and *S. strictum* in the present study) may thus explain the interaction between habitat preference, sexual condition, and mating patterns. Microhabitat preference along the hummock-hollow gradient is phylogenetically conserved within *Sphagnum*. (Johnson *et al.*, 2015a) Therefore, our findings about the connection between species ecology and population mating patterns suggest that related species are expected to retain mating system characteristics along with microhabitat preferences at macroevolutionary time scales.

Conclusions

We find that sexual condition, genetic diversity, and microhabitat preference all correlate with mating patterns in *Sphagnum* populations. Sexual condition is the most prevalent; the effect of microhabitat is detectable but subtler and more complex. Multiple paternity appears to be very common in *Sphagnum* but paternity skew is most pronounced in unisexual

populations living high above the water table. The possibility of labile sexual conditions within some species, such as *S. cuspidatum*, requires future studies investigating the genetic components of sexual determination in *Sphagnum*. We have greatly expanded the knowledge of inbreeding depression in natural populations of mosses, and have revealed that it is not a universal phenomenon in either unisexual or bisexual populations.

ACKNOWLEDGEMENTS

The authors thank S. Boles for technical laboratory support, as well as three anonymous reviewers for their comments and suggestions. This research was supported by NSF grant number DEB-0918998 to AJS and Blanka Shaw.

REFERENCES

Anderson L, Shaw B, Shaw AJ, Buck WR. 2009. *Peatmosses of the Southeastern United States*. Bronx, NY: New York Botanical Garden.

Andrus RE. 1986. Some aspects of *Sphagnum* ecology. *Canadian Journal of Botany* **64:** 416–426.

Barrett SCH. 2003. Mating strategies in flowering plants: the outcrossing-selfing paradigm and beyond. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences* **358:** 991–1004.

Bisang I, Ehrlén J, Hedenäs L. 2004. Mate limited reproductive success in two dioicous mosses. *Oikos* **104:** 291–298.

Charlesworth B, Charlesworth D. 1999. The genetic basis of inbreeding depression. *Genetical Research* **74:** 329–340.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134:** 1289–1303.

Chiou W, Farrar D. 1997. Antheridiogen production and response in Polypodiaceae species. *American Journal of Botany* **84:** 633–633.

Cronberg N. 1991. Reproductive biology of *Sphagnum*. *Lindbergia* **17:** 69–82.

Crum H. 1984. *Sphagnopsida, Sphagnaceae*. Bronx. NY: New York Botanical Garden.

Dobzhansky T. 1950. Genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of *Drosophila pseudoobscura*. *Genetics* **35:** 288.

Edmands S. 2006. Between a rock and a hard place: evaluating the relative risks of inbreeding and outbreeding for conservation and management. *Molecular Ecology* **16:** 463–475.

Eppley SM, Taylor PJ, Jesson LK. 2007. Self-fertilization in mosses: a comparison of heterozygote deficiency between species with combined versus separate sexes. *Heredity* **98:** 38–44.

Fischer M, Matthies D. 1997. Mating structure and inbreeding and outbreeding depression in the rare plant *Gentianella germanica* (Gentianaceae). *American Journal of Botany* **84:** 1685–1685.

Gignac LD. 1992. Niche structure, resource partitioning, and species interactions of mire bryophytes relative to climatic and ecological gradients in Western Canada. *The Bryologist* **95:** 406.

Goodwillie C, Kalisz S, Eckert CG. 2005. The evolutionary enigma of mixed mating systems in plants: occurrence, theoretical explanations, and empirical evidence. *Annual Review of Ecology, Evolution, and Systematics* **36:** 47–79.

Hedrick PW. 1987. Genetic load and the mating system in homosporous ferns. *Evolution* **41:** 1282.

Jesson LK, Perley DS, Cavanagh AP, Cameron JAC, Kubien DS. 2012. Mating and fitness consequences of sexual system in the moss *Atrichum undulatum* s.l. (Polytrichaceae). *International Journal of Plant Sciences* **173:** 16–25.

Johnson MG, Granath G, Tahvanainen T, Pouliot R, Stenøien HK, Rochefort L, Rydin H, Shaw AJ. 2015a. Evolution of niche preference in *Spahgnum*peat mosses. *Evolution* **69:** 90–103.

Johnson MG, Shaw AJ. 2015b. Data from: genetic diversity, sexual condition, and microhabitat preference determine mating patterns in *Sphagnum* (Sphagnaceae) peat-mosses. *Dryad Digital Repository*. doi:10.5061/dryad.73b9h.

Johnson MG, Shaw B, Zhou P, Shaw AJ. 2012. Genetic analysis of the peatmoss *Sphagnum cribrosum* (Sphagnaceae) indicates independent origins of an extreme infra-specific morphology shift. *Biological Journal of the Linnean Society* **106:** 137–153.

Joseph S, Kirkpatrick M. 2004. Haploid selection in animals. *Trends in Ecology & Evolution* **19:** 592–597.

Karlin EF, Andrus RE, Boles SB, Shaw AJ. 2011. One haploid parent contributes 100% of the gene pool for a widespread species in northwest North America. *Molecular Ecology* **20:** 753–767.

Karlin EF, Giusti MM, Lake RA, Boles SB, Shaw AJ. 2010. Microsatellite analysis of *Sphagnum centrale*, *S. henryense*, and *S. palustre* (Sphagnaceae). *The Bryologist* **113:** 90–98.

Lande R, Schemske DW. 1985. The evolution of self-fertilization and inbreeding depression in plants. I. Genetic models. *Evolution* **39:** 24.

Lynch M. 1991. The genetic interpretation of inbreeding depression and outbreeding depression. *Evolution* **45:** 622.

Marr AB, Keller LF, Arcese P. 2002. Heterosis and outbreeding depression in descendants of natural immigrants to an inbred population of song sparrows (*Melospiza melodia*). *Evolution* **56:** 131–142.

McDaniel SF, Willis JH, Shaw AJ. 2008. The genetic basis of developmental abnormalities in interpopulation hybrids of the moss *Ceratodon purpureus*. *Genetics* **179:** 1425–1435.

McQueen CB, Andrus RE. 2009. *Sphagnaceae*. New York, NY: Flora of North America North of Mexico.

Nei M, Maruyama T, Chakraborty R. 1975. The bottleneck effect and genetic variability in populations. *Evolution* **29:** 1.

Ramaiya M, Johnson MG, Shaw B, Heinrichs J, Hentschel J, von Konrat M, Davison PG, Shaw AJ.

2010. Morphologically cryptic biological species within the liverwort *Frullania asagrayana*. *American Journal of Botany* **97:** 1707–1718.

Rasmussen DI. 1979. Sibling clusters and genotype frequencies. *The American Naturalist* **113:** 948–951.

Ricca M, Beecher FW, Boles SB, Temsch E, Greilhuber J, Karlin EF, Shaw AJ. 2008. Cytotype variation and allopolyploidy in North American species of the *Sphagnum subsecundum* complex (Sphagnaceae). *American Journal of Botany* **95:** 1606–1620.

Ricca M, SzöVéNyi P, Temsch EM, Johnson MG, Shaw AJ. 2011. Interploidal hybridization and mating patterns in the *Sphagnum subsecundum* complex. *Molecular Ecology* **20:** 3202–3218.

Rydin H, Jeglum J. 2013. *The biology of peatlands*. Oxford: Oxford University Press.

Shaw AJ. 2009. Broyphyte species and speciation. In: Shaw AJ, Goffinet B, eds. *Bryophyte biology*. Cambridge: Cambridge University Press, 445–485.

Shaw AJ, Beer S. 1997. Gametophyte-sporophyte variation and covarition in mosses. *Advances in Bryology* **6:** 35–63.

Shaw AJ, Cao T, Wang L, Flatberg KI, Flatberg B, Shaw B, Zhou P, Boles S, Terracciano S. 2008a. Genetic variation in three Chinese peat mosses (*Sphagnum*) based on microsatellite markers, with primer information and analysis of ascertainment bias. *The Bryologist* **111:** 271–281.

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.

Shaw AJ, Cox CJ, Buck WR, Devos N, Buchanan AM, Cave L, Seppelt R, Shaw B, Larrain J, Andrus R, Grielhuber J, Temsch EM. 2010. Newly resolved relationships in an early land plant lineage: Bryophyta class Sphagnopsida (peat mosses). *American Journal of Botany* **97:** 1511–1531.

Shaw AJ, Pokorny L, Shaw B, Ricca M, Boles S, Szövényi P. 2008b. Genetic structure and genealogy in the *Sphagnum subsecundum* complex (Sphagnaceae: Bryophyta). *Molecular Phylogenetics and Evolution* **49:** 304–317.

Shaw B, Terracciano S, Shaw AJ. 2009. A genetic analysis of two recently described peat moss species, *Sphagnum atlanticum* and *S. bergianum* (Sphagnaceae). *Systematic Botany* **34:** 6–12.

Shortlidge EE, Rosenstiel TN, Eppley SM. 2012. Tolerance to environmental desiccation in moss sperm. *New Phytologist* **194:** 741–750.

Soltis DE, Soltis PS. 1992. The distribution of selfing rates in homosporous ferns. *American Journal of Botany* **79:** 97.

Stenøien HK, Såstad SM. 2001. Genetic variability in bryophytes: does mating system really matter? *Journal of Bryology* **23:** 313–318.

Stenøien HK, Shaw AJ, Stengrundet K, Flatberg KI. 2011. The narrow endemic Norwegian peat moss *Sphagnum troendelagicum* originated before the last glacial maximum. *Heredity* **106:** 370–382.

Stoneburner A, Wyatt R, Odrzykoski I. 1991. Applications of enzyme electophoresis to bryophyte phylogenetics and population biology. *Advances in Bryology* **4:** 1–27.

Sundberg S, Hansson J, Rydin H. 2006. Colonization of *Sphagnum* on land uplift islands in the Baltic Sea: time, area, distance and life history. *Journal of Biogeography* **33:** 1479–1491.

Sundberg S, Rydin H. 1998. Spore number in *Sphagnum* and its dependence on spore and capsule size. *Journal of Bryology* **20:** 1–16.

Szövényi P, Rensing SA, Lang D, Wray GA, Shaw AJ. 2011. Generation-biased gene expression in a bryophyte model system. *Molecular Biology and Evolution* **28:** 803–812.

Szövényi P, Ricca M, Shaw AJ. 2009. Multiple paternity and sporophytic inbreeding depression in a dioicous moss species. *Heredity* **103:** 394–403.

Taylor PJ, Eppley SM, Jesson LK. 2007. Sporophytic inbreeding depression in mosses occurs in a species with separate sexes but not in a species with combined sexes. *American Journal of Botany* **94:** 1853–1859.

Thingsgaard K. 2002. Taxon delimitation and genetic similarities of the *Sphagnum imbricatum* complex, as revealed by enzyme electrophoresis. *Journal of Bryology* **24:** 3–15.

Tuba Z, Slack NG, Stark L. 2011. *Bryophyte ecology and climate change*. Cambridge: Cambridge University Press.

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

Waser NM, Price MV, Shaw RG. 2000. Outbreeding depression varies among cohorts of *Ipomopsis aggregata* planted in nature. *Evolution* **54:** 485–491.

Wieder RK, Vitt DH. 2006. *Boreal peatland ecosystems*. Berlin: Springer-Verlag.

Wright S. 1922. Coefficients of inbreeding and relationship. *American Naturalist* **56:** 330–338.

Wright S. 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* **19:** 395.

Wyatt R. 1977. Spatial pattern and gamete dispersal distances in *Atrichum angustatum*, a dioicous moss. *The Bryologist* **80:** 284.

Wyatt R, Anderson L. 1984. Breeding systems in bryophytes. In: Dyer AF, Duckett JG, eds. *The experimental biology of bryophytes*. London: Academic Press, 40–64.

SHARED DATA

Data deposited in the Dryad digital repository (Johnson & Shaw, 2015b).

APPENDIX

We describe the twenty populations sampled for the survey of mating patterns in *Sphagnum*.

Sphagnum angustifolium Population: ME-S. Washington County, Maine, USA. Locality: near Steuben, E side of East Side Rd, 0.8 miles N of US1, 44.5221°N, 67.9500°W, elevation 35 m. Description: poor fen,

approximately 100 m from open water. Fruiting plants forming several hummocks in an area of approximately 100 m^2. Duke Herbarium collections: Matt Johnson 118–120. 15 June 2009.

Sphagnum austinii. Population: VI. Vancouver Island, Canada. Locality: Bamfield Area, W coast of island, on Bamfield Rd, 2.35 km S of Nuthatch Rd. 48.8155°N, 125.1275°W. Description: Poor-medium fen on NE side of road, around small shallow lake dominated by *Juncus*, *Carex*, and *Myrica*. Plants spread along two large hummocks approximately 300 m apart. Duke Herbarium collection: Jonathan Shaw 16578, 16584, 16585, 16590, and 16591.

Sphagnum compactum. Population: AK-PC. Matanuska-Susitna Borough, Alaska, USA. Locality: Petersville Rd, 17.2 miles W of jct Parks Hwy (AK2), N 62.3674° W 150.7121°, elevation 375 m. Description: Plants collected along an approximately 25-m transect parallel to the road in a floating fen. Duke Herbarium Collections: Matt Johnson 143 and 144, Jonathan Shaw 16913 and 16917. 9 August 2010.

Sphagnum compactum. Population: AK-PH. Denali Borough, Alaska, USA. Locality: Parks Hwy, approximately 30 miles S of Cantwell at 180 Mi Lake. 63.0822°N, 149.5252°W, elevation 560 m. Description: Fen on S side of Rd, with rich areas (*Tomenthypnum*, *Paludella*, etc.), and poorer depressions (with *Sphagnum lindbergii*, *Sphagnum balticum* and *Sphagnum* cf. *orientale*) plus higher hummocks (with *Sphagnum fuscum*, *Sphagnum lenense, Sphagnum capillifolium*). Plants collected in an approximately 100-m transect in low depressions between huummocks. Duke Herbarium collections: Matt Johnson 146–149. 9 August 2010.

Sphagnum cribrosum. Population: GA32. Long County, Georgia, USA. Locality: US-84, 1.3 miles NE of US-25 in Ludowici. 31.7232°N, 81.7270°W. Description: Wet trenches running parallel and perpendicular to W side of road. Growing intermixed in an area of approximately 25 m^2 area with *S. macrophyllum*. 8 May 2009.

Sphagnum cuspidatum. Population GA28. Ware County, Georgia, USA. Locality: W side of GA 177, 0.5 miles S of US 1, entrance to Okeefenokee Swamp. 31.1225°N, 82.2723°W. Description: Pondcypress depression dominated by *Taxodium ascendens, Ilex myrtifolia, Nyssa biflora* and *Carex striata*, on S side of road, powerline right-of-way, plants collected along an approximately 200-m transect, floating in the water. Duke Herbarium collections: Matt Johnson 101-113. 6 February 2009.

Sphagnum cuspidatum. Population NC-JL. Bladen County, North Carolina, USA. Locality: Jones Lake State Park, along Cedar Loop Trail on E shore. 34.6884°N, 78.5960°W, elevation 21 m. Description: Wet shaded depression in bay forest, plants collected from one large 10 m^2 patch. Duke Herbarium collections: Blanka Shaw 9746. 16 May 2009.

Sphagnum fallax. Population: ME-S. Washington County, Maine, USA. Locality: near Steuben, E side of East Side Rd, 0.8 miles N of US1, 44.5221°N, 67.9500°W, elevation 35 m. Description: poor fen, approximately 50 m from open water. Fruiting plants forming several hummocks in an approximately 100 m^2 area. Deeper into the fen than the S. angustifolium population, the hummocks were nearly dry. Duke Herbarium collections: Matt Johnson 121–123. 15 June 2009.

Sphagnum fallax. Population: ME-W. Washington County, Maine, USA. Locality: Great Wass Island Preserve, NE of Black Duck Cove Rd, 0.4 miles N of Preserve parking lot. 44.5221°N, 67.9500°W, elevation 14 m. Description: fruiting plants forming one large hummock at the edge of a moderate fen dominated by *S. tenellum*. Duke Herbarium collections: Matt Johnson 124–125. 16 June 2009.

Sphagnum macrophyllum Population SC-36. Jasper County, South Carolina, USA. Locality: 6.2 miles E of US 601 on SC 462, near Coosawatchie, SC. 32.6193°N, 81.0653°W. Description: Disturbed pond cypress dome S of highway, plants collected on an approximately 100 m transect arcing around the wettest areas beneath *Hypericum crux-andreae*, *Ilex myrtifolia*, and *Nyssa biflora*. Duke Herbarium collections: Matt Johnson 95–100. 5 March 2009.

Sphagnum macrophyllum. Population SC-39. Berkeley County, South Carolina, USA. Locality: Hell Hole Bay Wilderness Area in Francis Marion National Forest, FR 161 (Hell Hole Rd), 0.5 miles NE of FR 138. 33.218°N, 79.712°W. Description: Open canopy of *Taxodium distichum* surrounding approximately 2000 m^2 of open water approximately 1 m deep. Understory of *Lyonia lucida*, *Nyssa biflora*, and *Vaccinium formosum*. In the water, plants form almost continuous mats around *Nymphaea odorata*, *Dulichium arundinaceum*, and *Carex striata*. Collected as part of intensive survey of phenology and mating patterns (Chapter 2), April–May 2009 and December 2009–June 2010.

Sphagnum magellanicum. Population: NC-JL. Bladen County, North Carolina, USA. Locality: Jones Lake State Park, along Cedar Loop Trail on E shore. 34.6884°N, 78.5960°W, elevation 21 m. Description: Wet, partically shaded depression in old growth bay forest, with poison ivy. Plants formed one single hummock approximately 0.5 m^2. Duke Herbarium collection: Blanka Shaw 9745. 16 May 2009.

Sphagnum molle. Population: SC36. Jasper County, South Carolina, USA. Locality: 6.2 miles E of US 601 on SC 462, near Coosawatchie, SC. 32.6193°N, 81.0653°W. Description: Disturbed pond cypress dome S of highway. Plants collected in several isolated

hummocks in approximately 25 m^2 area, hidden beneath *Hypericum* and *Lyonia*. Duke Herbarium collection: Matt Johnson 115. 5 Mar 2009.

Sphagnum pulchrum. Population ME-C. Hancock County, Maine, USA. Locality: Gouldsboro Twp, E from Prospect Harbor at Corea Heath Bog (NWR). 44.4056°N, 67.9812°W, elevation 17 m. Description: Poor fen along edges and ombotrophic bog with raised mud flats. Plants collected near trailway at edge of fen forming many fruiting patches along an approximately 50 m^2 transect. Duke Herbarium collections: Jonathan Shaw 16085 and 16087. 11 June 2009.

Sphagnum squarrosum Population AK-W. Fairbanks North-Star Borough, Alaska, USA. Locality: Milepost 13 on Elliot Hwy (AK-2) just south of Willow Creek. 65.1004°N, 147.7464°W, elevation 180 m. Description: Moderately rich fen with hummocks/hollows dominated by *Sphagnum teres, S. obtusum* and *Vaccinium uliginosum* and *Salix* shrubs. Plants collected from several hummocks along an approximately 50 m^2 transect beneath blueberry bushes. Duke Herbarium collections: Jonathan Shaw 16849 and 16851, Matt Johnson 131–134. 4 August 2010.

Sphagnum strictum Population: SC-36. Jasper County, South Carolina, USA. Locality: 6.2 miles E of US 601 on SC 462, near Coosawatchie, SC. 32.6193°N, 81.0653°W. Description: Disturbed pond cypress dome S of highway. Plants collected in several isolated hummocks in an approximately 25 m^2 area, hidden beneath *Hypericum* and *Lyonia*. Duke Herbarium collection: Matt Johnson 116–117. 5 Mar 2009.

Sphagnum strictum. Population NC-LL. Wake County, North Carolina, USA. Locality: Lizard Lich granitic rock outcrops, approximately 4 miles WNW of Zebulon near intersection of Marshburn Rd. and Riley Hill Rd. 35.8447°N, 78.3731°W, elevation 80 m. Description: granitic flat outcrop, in partial shade in pine-juniper woodland. Plants collected from one cushion with a conspicuously large number of sporophytes. Duke Herbarium collection: Blanka Shaw 7202. 16 November 2008.

Sphagnum tenellum Population: ME-W. Washington County, Maine, USA. Locality: Great Wass Island Preserve, NE of Black Duck Cove Rd, 0.4 miles N of Preserve parking lot. 44.5221°N, 67.9500°W, elevation 14 m. Description: poor fen near trail, many fruiting plants in an approximately 25 m^2 hollow. Duke Herbarium collections: Jonathan Shaw 16037. 9 June 2009.

Sphagnum warnstorfii Population: AK-M. Valdez-Cordova Census Area, Alaska, USA. Locality: McCarthy Rd, 9.1 miles W of McCarthy foot bridge. 61.3861°N, 143.1821°W. Description: Extreme rich fen with hummock-hollow structure, scattered *Salix*, *Betula, Picea* and *Calex*. Plants collected from several hummocks. Duke Herbarium collections: Jonathan Shaw 16714, 16715, 16717. 24 July 2010.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Table S1. Assessment of relationship between percent heterozygosity and sporophyte size in *Sphagnum* using a pseudoreplication approach. One sporophyte per maternal shoot was chosen at random. The significance of the linear correlation and the r^2 value were recorded for each simulation. The number of pseudoreplicates with $P < 0.05$ and the mean r^2 across all pseudoreplicates for a population are shown.