American Fern Journal 109(3):267–282 (2019)
Published on 16 September 2019

Using RAD Data to Confirm Parentage of Polyploids in a Reticulate Complex of Ferns

Sylvia P. Kinosian
Department of Biology, Utah State University, 5305 Old Main Hill, Logan, UT 84322, USA.
E-mail: sylvia.kinosian@gmail.com

Weston L. Testo
Department of Biology, University of Florida, Box 118525, Gainesville, FL 32611, USA.
E-mail: westontesto@gmail.com

Sally M. Chambers
Marie Selby Botanical Gardens, 811 South Palm Avenue, Sarasota, FL 34236, USA.
E-mail: schambers@selby.org

Emily B. Sessa*
Department of Biology, University of Florida, Box 118525, Gainesville, FL 32611, USA.
E-mail: emilysessa@ufl.edu

Abstract.—Reticulate evolution, in which phylogenetic relationships are not strictly bifurcating (tree-like), is a common feature of fern evolution. Ferns are prone to hybridization and whole genome duplication, two processes that can make untangling phylogenetic relationships among species challenging. Next-generation sequencing technologies have greatly increased the amount of data available for analyzing various aspects of evolutionary history, and here we test the ability of one next-generation sequencing approach to identify the progenitors of allopolyploids. We produced and analyzed double-digest restriction-site-associated DNA (ddRAD) sequences from six species of North American *Dryopteris*, including two allopolyploids and their respective diploid parents. The relationships of these species have been confidently established in previous studies, and our goal was to determine the extent to which RAD data are capable of identifying these known relationships. Analyses of the genetic structure in our samples reliably separated the diploids from one other, but in general each polyploid sample resembled one or the other of its progenitors, or had genetic variation unassignable to either parent. None of the polyploid samples had unambiguous genetic contributions from both known parents, as we had expected. These results may have been influenced by small overall sample size, different numbers of samples from the two diploid parents in each pair, and the large divergence times between the diploids. These are all potentially important issues to consider when designing similar studies, and our results therefore have useful implications for researchers interested in using a RAD approach to study polyploid complexes.

Key Words.—allopolyploid, next-generation sequencing, phylogenetics, reticulate evolution

Fern enthusiasts, amateur and professional alike, are captivated by these plants for a multitude of reasons. For researchers interested in evolutionary processes, ferns are a particularly fascinating lineage to study because of their propensity for polyploidy and hybridization. These two processes, which are particularly common in ferns compared to other land plants (Wood *et al.*, 2009), have the potential to influence many aspects of evolution, particularly

* corresponding author

those related to genome size, structure, and complexity, as well as phylogenetic relationships (Soltis, Visger, and Soltis, 2014). Hybridization occurs when members of two distinct evolutionary lineages interbreed and produce offspring. Polyploidy, or whole genome duplication, is a complete doubling of the genome that results in offspring with at least twice the number of chromosomes and genetic content of their progenitors. These processes can occur independently or in synchrony, producing organisms known as allopolyploids that are the product of both hybridization and genome doubling.

The non-bifurcating phylogenies that result from these reticulate evolutionary processes require extra effort to decipher. The traditional workhorse of plant phylogenetics, the chloroplast genome, is maternally inherited in most plants, including in ferns (Gastony and Yatskievych, 1992; Vogel *et al.*, 1998), and can therefore identify only one parent of a putative hybrid or allopolyploid. Identifying the second, paternal parent, requires information from biparentally inherited nuclear markers. For the last two decades, obtaining data from these markers has relied on painstaking and time-consuming laboratory procedures to isolate each homoeologous copy (using an *Escherichia coli* vector that replicates via cloning), so that each can be sequenced independently. Genes such as *gapCp* (Schuettpelz *et al.*, 2008), *pgiC* (Ishikawa *et al.*, 2002), and many others (Rothfels, Li, *et al.*, 2015) have been analyzed in this way and produced the first DNA-sequencing based confirmations of parentage in many fern polyploid complexes (e.g., in *Dryopteris* Adans. (Sessa, Zimmer, and Givnish, 2012b), *Polystichum* Roth (Jorgensen and Barrington, 2017), various Pteridaceae genera (Beck *et al.*, 2010; Grusz, Windham, and Pryer, 2009), and many others). However, the rise of next-generation sequencing (NGS) technologies has inspired a natural desire to use these powerful and data-rich approaches as an alternative to labor-intensive gene-by-gene cloning and sequencing for analyzing polyploid complexes. Despite this, few studies have used NGS approaches to resolve reticulate evolutionary histories in ferns (but see Rothfels, Pryer, and Li, 2017).

A principal limitation to the use of NGS data in studies of polyploids has been the challenge of correctly assembling homoeologous copies, especially when sequence data are generated on platforms with relatively short read lengths. Overcoming this and related bioinformatic challenges will be a critical step before widespread use of NGS approaches in the study of polyploid complexes can become feasible. An additional problem for ferns is their large genomes, which have made them less tractable than other plant lineages as study groups for NGS methods (for example, ferns were the last major lineage of land plants to have a reference nuclear genome sequenced, due in part to their massive genomes (F.-W. Li *et al.*, 2018; Sessa *et al.*, 2014).

"Reduced representation" sequencing methods seek to minimize the complexity of genome assembly by sequencing only a subset of the complete genome (Andrews *et al.*, 2016; Rowe, Renaut, and Guggisberg, 2011). While the cost of next-generation sequencing has decreased steadily, assembling the millions of short (typically 100-150 base pairs) sequencing reads produced by

these approaches remains an immense challenge. For many study systems, sequencing and assembling an entire genome remains out of reach, due either to financial limitations, assembly issues, or a combination of the two (e.g., difficult-to-assemble genomes benefit from long-read sequencing, which is more expensive than short-read sequencing). Whole genome sequencing may also be unnecessary for addressing questions of interest, such as identifying the parents of a polyploid species, a query for which sequence data from only one or a handful of markers is typically sufficient. In groups like ferns, where whole-genome sequencing is still impractical for the average researcher, reduced representation strategies are ideal for capitalizing on next-generation sequencing approaches while using resources efficiently.

In the present study, we evaluated the utility of one reduced representation, next-generation sequencing approach – restriction-site-associated DNA sequencing (known as RAD or RADseq) – for identifying the progenitors of polyploid ferns. RAD and the related genotyping-by-sequencing (GBS) are both approaches that utilize the restriction enzyme cut sites that occur naturally across the genome (Andrews *et al.*, 2016). Whole genomic DNA is first digested with restriction enzymes that cut the DNA only at specific sequences that are unique to each enzyme; these cut sites typically occur thousands of times throughout the genome, at different frequencies for different enzymes. The resulting DNA fragments will span a range of sizes, and those in the ideal range for next-generation sequencing can be selected at a later step in library preparation. By sequencing only fragments that are adjacent to restriction enzyme cut sites, RAD targets a non-random portion of the genome and therefore increases the likelihood of sequencing homologous regions across samples. This is especially important for organisms with large genomes, like ferns, where approaches such as genome skimming or "shotgun" sequencing (which theoretically sequence a truly random subset of the genome) are less likely to capture homologous sequences from different samples and species. Because SNPs are identified from individual (unassembled) reads in the RAD data analysis pipeline, this method does not suffer from the issues associated with assembling homoeologous copies from short-read data. However, while RAD and other reduced-representation approaches attempt to deal with the issue of large genome sizes, the analytical complications introduced by polyploidy still haunt these approaches, since the programs available for analyzing the datasets typically operate under the assumption that included taxa are diploid, a fundamental assumption (with numerous implications for expected copy numbers and locus behavior, among other things) that is not readily altered to accommodate data from known polyploids. For example, two of the most commonly used pipelines for processing GBS and RADseq data, Stacks (Catchen *et al.*, 2013) and TASSEL (Bradbury *et al.*, 2007), assume genotypes are diploid and do not permit allele frequencies to deviate from those expected in diploids, typically treating as noise information that may be in fact be the signal of polyploid genotypes.

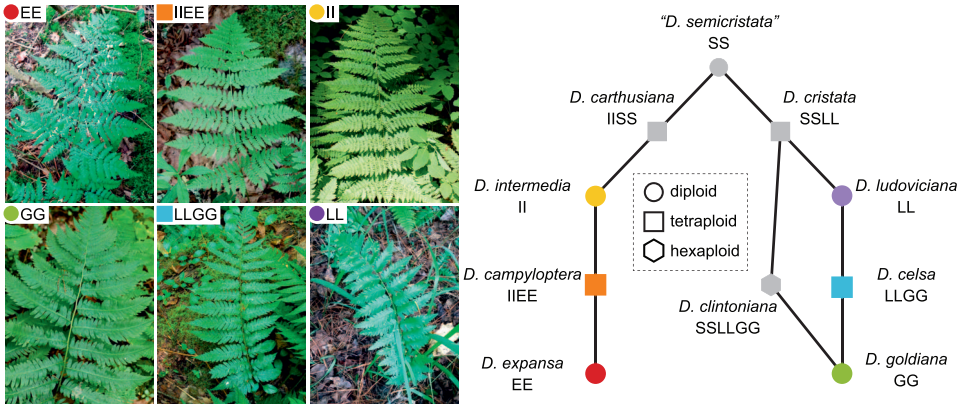

FIG. 1. Relationships of diploid and polyploid species in the North American *Dryopteris* reticulate complex (based on (Sessa, Zimmer, and Givnish, 2012a, 2012b, 2012c). Letters below species names refer to their genomic designations. Pictures of all species are included at left (all photos by EB Sessa).

We applied a double-digest RAD approach (ddRAD; so called because two restriction enzymes are used) to a small dataset for a polyploid complex in which relationships are known with confidence: North American *Dryopteris*. These ferns have been the focus of previous study by our research group (Sessa, Zimmer, and Givnish, 2012a, 2012b, 2012c; Sessa *et al.*, 2015; Sessa and Givnish, 2014; Testo, Watkins, and Barrington, 2015), and the North American complex as a whole includes one extinct and four extant diploids, four allotetraploids, and one allohexaploid (Fig. 1). The parents of the allopolyploids were first hypothesized based on morphology and cytological observations (Walker, 1955, 1959, 1961, 1962, 1969), and later confirmed using sequences of *gapCp* and *pgiC* (Sessa, Zimmer, and Givnish, 2012b). Here we focused on two of the allotetraploids, *D. campyloptera* (Kunze) Clarkson and *D. celsa* (W. Palmer) Knowlt., Palmer & Pollard, and their respective diploid parents: *D. intermedia* (Willd.) A. Gray and *D. expansa* (C. Presl) Fraser-Jenk. & Jermy for *D. campyloptera*, and *D. ludoviciana* (Kunze) Small and *D. goldiana* (Hook. ex Goldie) A. Gray for *D. celsa* (Fig. 1).

Our goal was to determine whether analysis of ddRAD sequences from a small sampling of these six species could recover evidence of their known polyploid-progenitor relationships by correctly grouping sequences from the two allotetraploids with their respective parent taxa, ideally with each polyploid showing evidence of equal genetic contributions from the two progenitors. The six species we selected are ideal for this study because all diploid progenitors are extant and can be sampled, and the diploids are sufficiently diverged from one another (Sessa, Zimmer, and Givnish, 2012a) that sequences from the polyploids can theoretically be assigned unambiguously to each of the four diploids.

TABLE 1. Sample ID and voucher information for samples included in this study. Institution acronyms follow Index Herbariorum (Thiers, 2018).

	Genus	Ploidy	Sample ID	U.S. state collected	Voucher
1	*Dryopteris campyloptera*	4x	E648	VA	EBS 9722006 (FLAS)
2	*Dryopteris campyloptera*	4x	E650	VA	EBS 9722005 (FLAS)
3	*Dryopteris campyloptera*	4x	E655	VA	EBS 9722004 (FLAS)
4	*Dryopteris campyloptera*	4x	E649	VA	EBS 9722016 (FLAS)
5	*Dryopteris celsa*	4x	E657	GA	EBS 27 (WIS)
6	*Dryopteris celsa*	4x	E658	SC	EBS 49 (WIS)
7	*Dryopteris celsa*	4x	E661	LA	Price 94-2 (NY)
8	*Dryopteris celsa*	4x	E664	MO	3479307 (MO)
9	*Dryopteris expansa*	2x	E734	AK	5710532 (MO)
10	*Dryopteris goldiana*	2x	E700	NY	EBS 65A (WIS)
11	*Dryopteris goldiana*	2x	E701	NY	EBS 72 (WIS)
12	*Dryopteris goldiana*	2x	E715	NY	EBS 12 (WIS)
13	*Dryopteris intermedia*	2x	E713	WV	EBS 9722021 (FLAS)
14	*Dryopteris intermedia*	2x	E714	VA	EBS 9722010 (FLAS)
15	*Dryopteris intermedia*	2x	E736	MO	5198765 (MO)
16	*Dryopteris intermedia*	2x	E739	SC	EBS 48 (WIS)

MATERIALS AND METHODS

Taxon sampling and DNA extraction.—We included sixteen samples representing six *Dryopteris* species, with all but two species represented by multiple accessions (Table 1). We extracted total genomic DNA using a DNeasy Plant Mini Kit (Qiagen, Valencia, California, USA) following the manufacturer's protocol.

ddRAD library preparation and sequencing.—We followed the ddRAD library construction protocol established by (Peterson *et al*., 2012), with a few modifications. Because of the large genome sizes of the focal taxa (average in *Dryopteris* diploids is 1C = 7.63 pg; Bainard *et al*., 2011), we replicated each sample three times during library preparation. To obtain equal numbers of reads for all individuals, we standardized DNA quantity prior to library preparation.

We used two enzymes, *Mse*I and *Eco*RI – a frequent cutter and an infrequent cutter, respectively (referring to the distribution of the enzymes' cut sites across the genome) – to digest 6 μL of genomic DNA from each sample. We then ligated enzyme-specific, double-stranded adaptors (8–14 base pairs in length) to the digested DNA fragments, with the *Eco*RI adapter containing a unique barcode specific to each sample and replicate. We ensured successful ligation of the adaptors to the digested DNA by inspecting PCR products via gel electrophoresis visualization. We then pooled the restriction ligation product from each of the successful libraries and cleaned this pooled product using a QIAquick PCR Purification Kit (Qiagen, Valencia, California, USA). The pooled product was then run on an Elf Pippin Bioanalyzer (Sage Science, Massachusetts, USA) to select genomic fragments ranging from 350 to 700 bp

(size selection service provided by the University of Florida Interdisciplinary Center for Biotechnology Research; UF ICBR). We checked the success of the fragment size selection via gel electrophoresis and analysis on a TapeStation 2200 Automated Electrophoresis (Agilent, California, USA) system.

We performed a final round of PCR to anneal the Illumina sequencing primers to the digested DNA fragments, working with 1 μL at a time of digested, size-selected, pooled DNA. To reduce the opportunity for PCR errors we performed eighteen separate reactions and then combined the resulting PCR products for sequencing. We visualized a subset of the pooled product via gel electrophoresis to check for amplification success, and cleaned the remaining product using a QIAquick PCR Purification Kit. Pooled, clean PCR product was submitted to the UF ICBR where it was cleaned further using Ampure beads to remove unincorporated adaptors before being sequenced on an Illumina NextSeq500 platform, generating 2 × 150 bp reads. A 10% phiX spike was included during sequencing as an internal control.

Data cleaning.—We processed the raw Illumina reads using the Process Radtags pipeline in Stacks (Catchen *et al.*, 2013), retaining all reads with a quality score above 20 and splitting the raw reads by sample and replicate based on the unique *Eco*RI barcode, which was subsequently trimmed along with the cut site. This resulted in roughly 277 million reads. We then used the FAST-X Trimmer from the FAST-X Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) to remove the *Mse*I cut site from each secondary read and to trim the last four bases from the 3′ end of the primary reads to make all reads 127 bp in length. Cleaned, trimmed, forward and reverse reads for each individual were paired used PEAR v. 0.9.2 (Zhang *et al.*, 2014), and all reads from each of the three replicates per sample were pooled together. Unfiltered, demultiplexed sequences have been deposited in NCBI GenBank Short Read Archive (PRJNA542715).

Data analysis.—Processing the raw Illumina data occurred in three stages: 1) creation of a pseudo-reference genome, 2) alignment of reads and variant calling, and 3) admixture analysis. An R Markdown file describing this pipeline and all custom scripts is located here: https://github.com/sylviakinosian/dryopteris_gbs.

Creation of the Pseudo-Reference Genome.—Because there is no reference genome available for *Dryopteris*, we constructed a pseudo-reference using the sequences from the diploid taxa. We chose to use only the diploid taxa because they contain the majority of the sequence variation present in the tetraploids (which are known to be hybrids between the diploids), but harbor none of the possibly divergent sequences that may be present in the tetraploid species. We also performed our analyses separately on each tetraploid clade: *Dryopteris celsa* and its progenitors, and *D. campyloptera* and its progenitors. These two clades are separated by almost 40 million years of evolution (Sessa, Zimmer, and Givnish, 2012b).

We created two pseudo-references (one for each clade) by first clustering highly similar sequences for each diploid taxon separately using VSEARCH v

2.4.2 (Rognes *et al.*, 2016). Clustering was done at 92% similarity to create centroids for further clustering. We then clustered at 84% similarity and removed all sequences that collapsed at this stage, to exclude paralogs. We next combined the two diploid taxa from a given clade (either *D. intermedia* and *D. expansa*, or *D. ludoviciana* and *D. goldiana*) and clustered again using VSEARCH at 84% similarity; the resulting sequences were then used as our two pseudo-reference genomes.

Alignment of Reads and Variant Calling.—Before calling variants for all of the included species, we first had to index the pseudo-reference genome from the previous step, which identifies sequence position points for the alignment. This was done using the INDEX function of BWA v. 0.7.10 (H. Li and Durbin, 2009). Next, we used PicardTools v. 2.9.0 (Broad Institute, 2019) to create a sequence dictionary, and the INDEX function of SAMTOOLS v. 1.5 (H. Li *et al.*, 2009) was then used to create a FASTA index file. We used BWA ALN and SAMSE to align all individual reads to the appropriate pseudo-reference. Next, we used the SAMTOOLS functions VIEW, SORT, and INDEX to prepare all individual reads for variant calling. To call variants, we used two different methods. First, we used the GATK HaplotypeCaller v. 3.8.0 (McKenna *et al.*, 2010) to call variants on all samples as diploids. We then used VCFTOOLS v. 0.1.15 (Danecek *et al.*, 2011) to filter the resulting VCF file. For the second approach, we again used the GATK HaplotypeCaller, but called variants separately on the diploids and tetraploids (HaplotypeCaller allows the user to specify any ploidy). We then used a custom Python script to filter the resulting VCF files (VCFTOOLS does not support polyploids). Finally, we used custom Perl (v. 5.15.3, https://www.perl.org/) scripts to find the intersection of SNPs from the diploids and tetraploids.

Admixture Analysis.—We used the population genetics program ENTROPY v. 1.2 (Gompert *et al.*, 2014), which is very similar to the popular program STRUCTURE (Pritchard, Stephens, and Donnelly, 2000). While both are Bayesian, model-based approaches to population genetics, a key difference between them is that STRUCTURE assumes that individual genotypes are known *a priori*, whereas ENTROPY does not. ENTROPY calculates genotype likelihoods from raw sequence data and quality estimates; these genotype likelihoods are then used as an input for the model. STRUCTURE calculates the likelihoods repeatedly at each MCMC step from prior genotype assignments (Gompert *et al.*, 2014).

The first step of the ENTROPY analysis was to convert our VCF variant file to a Genotype Likelihood (GL) file format using a custom Perl script. A second Perl script was used to convert the GL file to a matrix for input to R v. 3.5.1 (R Development Core Team, 2016). We used the R package ADEGENET v. 2.1.1 (Jombart, 2008) to perform a discriminant analysis of principal components (DAPC) to find the most likely source population for each individual. This analysis is similar to that performed by ENTROPY, but is less complex, and generates starting values that can be used to seed ENTROPY, which helps eliminate label swapping and allows the MCMC analysis to converge more

quickly. We followed the DAPC protocol of Jombart and Collins (available at: http://adegenet.r-forge.r-project.org/files/tutorial-dapc.pdf).

We ran ENTROPY for k = 2, 3, 4, and 5 with 2 chains in each analysis, and we examined the results obtained with and without the DAPC starting values, with high and low burn-in, and with various numbers of iterations. We found that the various values of these parameters did very little to alter the results, and so decided to run the final analysis using the starting values and with a large number of iterations (65,000) and a high burn-in (15,000). We used the program ESTPOST v. 1.2 (Gompert *et al.*, 2014) to extract admixture proportions for each individual, and then used a custom R script and functions to visualize the ENTROPY output.

Finally, we performed a principal component analysis on the ENTROPY output. We used the program ESTPOST to extract genotype probabilities for k = 2–5 and then read those data into R. We averaged the genotype probabilities for all k values, and then performed a PCA using the R function prcomp. All of the R code used to analyze and visualize the data is available as an Rmarkdown file on Github (https://github.com/sylviakinosian/dryopteris_gbs).

Results

The pseudo-reference genomes constructed for the two clades differed in the relative evenness of contributions from the two sets of diploids: for the *intermedia - expansa - campyloptera* clade, 837,895 and 64,631 contigs were retrieved from *D. intermedia* and *D. expansa*, respectively. In the *ludoviciana - goldiana - celsa* clade, the contigs were a bit more evenly divided, with 440,468 retrieved from *D. goldiana* and 132,480 from *D. ludoviciana*.

The first analysis, with variants called on all samples as diploids, retained 1288 SNPs for the *intermedia - expansa - campyloptera* clade and 2122 SNPs for the *ludoviciana - goldiana - celsa* clade. The second analysis, which called variants for diploids and tetraploids separately, retained 3964 SNPs for the *intermedia - expansa - campyloptera* clade, and 4419 SNPs for the *ludoviciana - goldiana - celsa* clade. We performed the ENTROPY analysis on both sets of SNPs, and although calling variants separately on diploids and tetraploids obviously increased the total number of SNPs retained for both clades, we did not see a marked difference in the ENTROPY results between the two variant calling routines. We used the set of SNPs called only as diploids in the final analyses reported here.

ENTROPY analyses of the genotype data generally separate the diploid taxa from one another, but for both clades, the polyploid species contain genomic contributions primarily from one parent or the other (in the case of *Dryopteris celsa*), or they contain a substantial fraction from one parent as well as additional fractions that are not attributed to the second parent (in the case of *D. campyloptera*) (Fig. 2). For the *D. campyloptera* samples, the diploid *D. intermedia* was the dominant contributor, with some small contributions from the other diploid parent, *D. expansa*, and additional fractions unassigned to either diploid.

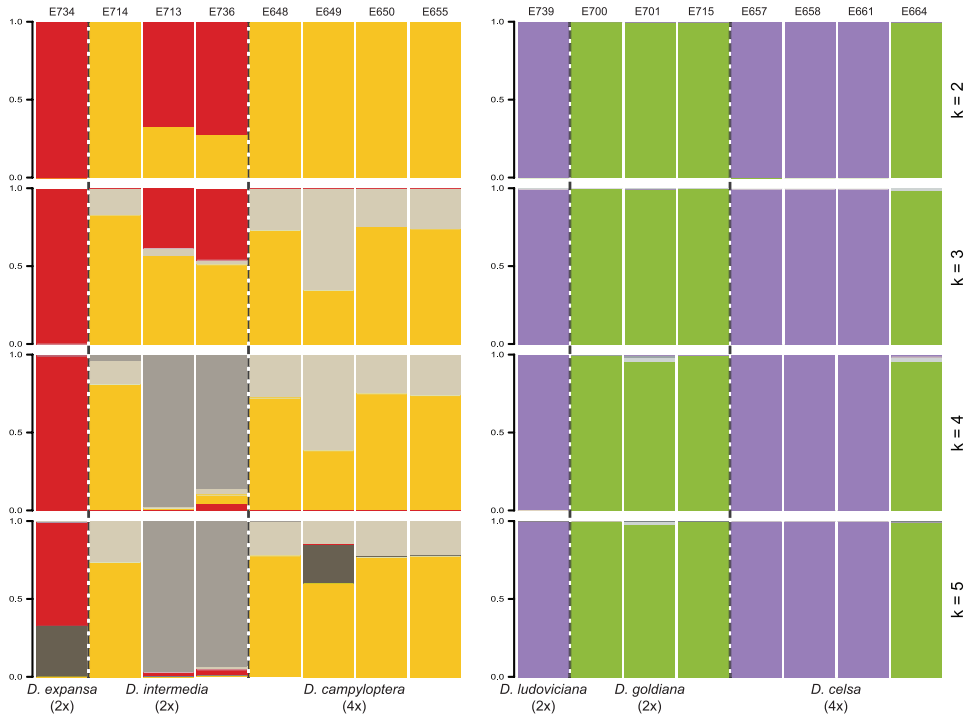

FIG. 2. Results of ENTROPY analysis of genotype data for two polyploid clades in North American *Dryopteris*. For each clade, we investigated admixture proportions between the species of each clade with k = 2–5 ancestral source populations. Our expectation, given that we know the parentage of the tetraploids in both clades, would have been to see equal representation of the two diploid progenitors in the genetic makeup of each tetraploid sample. In contrast to that expectation, we found that for the *D. campyloptera* clade, at all k values, *D. intermedia* was the greatest contributor to the genome of the tetraploid hybrid *D. campyloptera*, and in the *D. celsa clade*, while the two diploid progenitors were both found to contribute to the tetraploid samples, it was always as an overwhelming contribution towards one parent or the other, in each of the samples.

The principal component analysis yielded a similar pattern to the ENTROPY analysis. In the *D. expansa - intermedia - campyloptera* clade, the *D. campyloptera* individuals clustered most closely with the *D. intermedia* individuals (Fig. 3a). In the *D. ludoviciana - goldiana - celsa* clade, *D. celsa* clustered with both progenitor species, although more individuals clustered with *D. ludoviciana* individuals than with *D. goldiana* (Fig. 3b).

DISCUSSION

Dryopteris campyloptera *clade*.—For all values of k tested, SNPs associated with the diploid *Dryopteris intermedia* dominate the genetic complement of the tetraploid samples (Fig. 2, 3). *Dryopteris expansa*, the other diploid parent, has a distinct genotype that rarely appears in any of the tetraploids. We had an

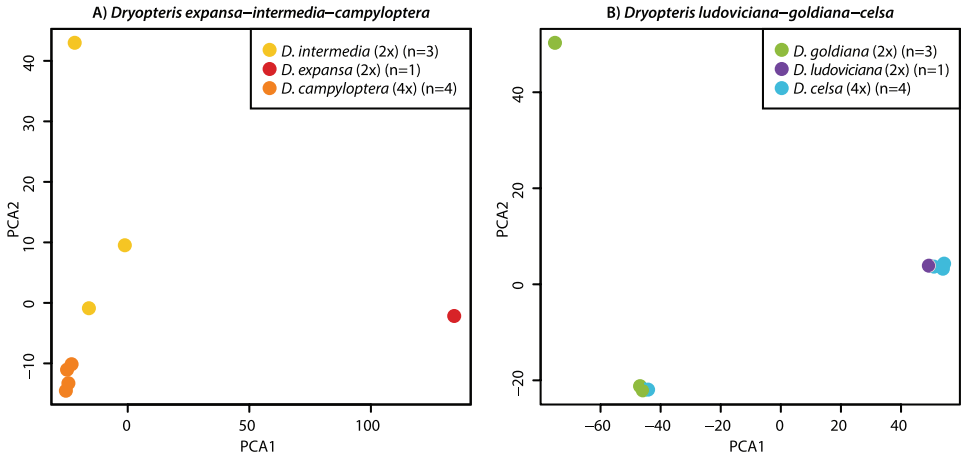

FIG. 3. Principal component analyses for the two clades. Legends indicate species names and number of samples. A) In the *D. expansa - intermedia - campyloptera* clade, the tetraploid *D. campyloptera* clusters most closely with *D. intermedia*, as in the admixture analysis (Fig. 2). There is also reasonable genetic variation in the progenitor *D. intermedia*. B) In the *D. ludoviciana - goldiana - celsa* clade, the tetraploid *D. celsa* clusters with both progenitors. Interestingly one progenitor, *D. goldiana,* shows more variation here than in the admixture plot (Fig. 2). Clusters of points in the lower left-hand corners of both plots have been jittered slightly to make the points distinguishable.

unequal number of samples from the two diploid parents in this group: three from *D. intermedia* and one from *D. expansa*. When constructing the reference genome for this clade, this unevenness would have resulted in a majority of the sequences belonging to the *D. intermedia* genome, as we saw – there were nearly 13 times as many contigs from *D. intermedia* in the pseudo-reference genome as there were from *D. expansa*. Consequently, most of the SNPs called were from this progenitor. Some SNPs associated with *D. expansa* do occur in each of the *D. campyloptera* samples at k = 3, 4, and 5 (visible as extremely thin red lines at the tops of the stacked bars for those k values in Fig. 2), but these are extremely minor contributions to the tetraploid genome samples. At k = 5, a more substantial *D. expansa* contribution occurs in one of the tetraploids (sample E649), further supporting the hypothesis that only a small number of *D. expansa* sequences made it into the reference and subsequent SNP calling.

An additional genetic component found in the *D. campyloptera* samples at k = 3–5 (light grey sections in Fig. 2) may be due to sequence divergence and accumulation of mutations in cut sites that would have occurred subsequent to the formation of the polyploid species, whose earliest inferred age of formation is 4.6 million years ago (Sessa, Zimmer, and Givnish, 2012b). That is an adequate amount of time for numerous mutations to occur that would alter the location and/or frequency of cut sites, and this would potentially produce fragments in the polyploids with lengths and sequence compositions not found in either parent species. Considering this, *D. campyloptera* sample E649 perhaps best represents our *a priori* expectations for a successful result from

this study, as it includes sets of SNPs found in both parents as well as an additional component not found in either diploid.

We also found evidence of genetic diversity in one of the diploid parents in this clade, *D. intermedia*. The three samples of *D. intermedia* differ considerably from one another in both the ENTROPY admixture analysis and the PCA (Figs. 2, 3), and at k = 2 a large portion of SNPs in two of the *D. intermedia* samples are associated with the *D. expansa* sample. At higher k values, the three *D. intermedia* samples become much more distinct from *D. expansa*; at k = 4 and 5, two of the *D. intermedia* samples resemble each other closely, while the third is distinct and resembles the *D. campyloptera* samples. These results suggest that there is substantial genetic diversity within and between populations of *Dryopteris intermedia*, which is unsurprising given its wide geographic range (Montgomery and Wagner, 1993). Both *D. intermedia* and *D. expansa* diverged from their closest diploid relatives in the Miocene (ca. 10–15 mya; Sessa, Zimmer, and Givnish, 2012a), and it therefore seems likely that we might discover additional genetic variation in *D. expansa* as well, if we sampled additional populations across its broad range.

Dryopteris celsa *clade*.—For this clade, we also had three samples from one parent (*D. goldiana*) and one from the other (*D. ludoviciana*), but there was less apparent sampling bias of the under-represented parent in the pseudo-reference assembly. The balance between the number of contigs these species contributed to the reference was more equal than in the *D. campyloptera* clade: after clustering, 440,468 contigs were retrieved from the *D. goldiana* individuals and 132,480 from *D. ludoviciana*. In all of the analyses of this clade (k = 2–5), *D. goldiana* and *D. ludoviciana* are assigned to separate populations; three of the *D. celsa* individuals are assigned to the "*ludoviciana*" population, and one is assigned to the "*goldiana*" population. There is less apparent bias in these assignments than there was in the *D. campyloptera* clade, in that both parental genotypes are present in the tetraploids. However, the results are clearly at odds with our expectations, which would have been for each of the tetraploid samples to show clear evidence of genetic contributions from (at least) two sources, rather than being dominated by only one.

In this clade we also saw some evidence of genetic variation in one progenitor species, *Dryopteris goldiana*. This was most evident from the PCA (Fig. 4), where one of the *D. goldiana* individuals clustered very differently from the other two. *Dryopteris goldiana* has a broad geographic range, but is regionally rare and locally abundant across its distribution, and is found only in rich woods and ravines (Montgomery and Wagner, 1993). This could potentially isolate populations from one another, accounting for the diversity observed in the PCA results. *Dryopteris ludoviciana* has a much narrower range than any of the other three progenitor species in this study (Montgomery and Wagner, 1993), and further sampling would be required to reveal whether it has a similar amount of genetic variation as the other progenitor in this particular clade.

RAD data and analysis of polyploid complexes.—The goal of this research was to assess the utility of RAD sequence data for identifying progenitors in a polyploid complex. Our dataset, which consisted of a small number of samples from a group with known polyploids and no reference genome, is typical of what might be available to researchers interested in polyploid ferns, and thus serves as a test case for these types of analyses. Based on the results discussed above, our success was mixed, and interpretation of our results was greatly facilitated by knowing in advance the progenitors of the allopolyploid species. We suspect that studies attempting to use a RAD-based approach to determine relationships in a polyploid complex where progenitors are *not* known would face substantial challenges and likely obtain at least somewhat ambiguous results. Nonetheless, for researchers interested in these methods despite their potential shortcomings, there are several aspects of our study, including features of the species complex itself and of our experimental design, that are informative and which we discuss below.

As mentioned earlier in the discussion, for both tetraploid complexes there was a disparity in our sampling of the parents: both groups had three samples from one parent and one sample from the other. When building the pseudo-reference genomes for each clade, we attempted to balance the number of contigs contributing to the reference from both diploid species equally. This was done using a Perl script to find the intersection of contigs from two species after the final clustering step. However, at each locus that matched our search criteria, we used the consensus sequence created by VSEARCH. In many cases, these consensus sequences were built from clusters of contigs that had a greater representation of one diploid over the other, making the consensus somewhat biased toward one parent. There are a handful of programmatic ways to remedy this using custom scripts, but perhaps the most practical and effective would be to increase our sample size and better balance sampling from the two progenitors. This would not only help create consensus sequences that are built from similar numbers of contigs from each species, but would also increase the size of the pseudo-reference and consequently the number of SNPs that could be called.

RAD approaches were originally developed to address questions about population genetics (Andrews *et al.*, 2016; Rowe, Renaut, and Guggisberg, 2011), and are therefore most informative across relatively short evolutionary distances, typically for individuals whose maximum divergence is between 5 and 10 mya. GBS approaches have been successful at analyzing hybrid complexes that are more recently diverged than our *Dryopteris* system, for example in *Juglans* (Zhao *et al.*, 2018). In the present study, the two clades of *Dryopteris* are separated by about 40 million years of evolution, and even within the hybrid clades, roughly 15–25 million years have elapsed since the divergence of the diploid progenitors (Sessa, Zimmer, and Givnish, 2012a, 2012b). Mutations can start to accumulate in enzyme cut sites that can result in non-homologous fragments being cut and amplified across species, which becomes more likely the longer ago the species diverged (Eaton *et al.*, 2017). This is perhaps part of the reason that our results did not reflect particularly

well the known relationships between the tetraploid *Dryopteris* species and their progenitors.

To investigate the time scales at which RAD approaches are most effective, Eaton *et al.* (2017) examined RADSeq data across several lineages with a range of divergence times. They found that phylogenetic distance was not a good predictor of the number of SNPs recovered; uneven sequence coverage was found to have the most impact on missing data. They also found that as sample size increased, some loci that were initially found as singletons in small datasets were recovered in more individuals, thereby becoming phylogenetically informative. This suggests that sample size might be the biggest driver of our less than satisfying results, as potentially informative SNPs may only occur in one of the sampled individuals. The small size of our dataset, both in terms of numbers of samples and numbers of species (two sets of three species), also influenced our ability to use additional programs available for analyzing SNP data. For example, phylogenetically-informed approaches such as HyDe and PhyloNet are unlikely to be informative with so few samples and clusters of only three taxa. In some of the datasets examined by Eaton *et al.* (2017), sequencing fewer loci at a higher coverage (essentially the strategy we employed in the present study) resulted in large amounts of missing data for highly divergent lineages. Even though we did not have a large amount of missing data across samples, the high specificity of the double digest method may have been problematic for such a highly divergent group as *Dryopteris*.

Summary and future directions.—Ferns are known to hybridize across vast evolutionary distances (Rothfels, Johnson, *et al.*, 2015; Sessa *et al.*, 2018), and so sequencing tools are needed that are informative across diverse and highly divergent groups. RAD methods may be more informative for studying polyploid or hybrid complexes that have formed more recently than the North American *Dryopteris* complex, but by adjusting the RAD methodology as discussed in Eaton *et al.* (2017), this type of reduced representation may indeed be a good option for future investigations of deeply divergent fern species and their hybrids (assuming that adequate sample size can be achieved).

In addition to reduced representation methods, there are several next generation sequencing techniques that could potentially be utilized to explore complexes involving deeply divergent fern lineages. Ultraconserved elements (UCEs) have proven useful for investigating deep divergences in animal lineages, but are not a viable option for plants due to many ancient polyploidy events (Jiao *et al.*, 2011), which can fracture and rearrange the genome, making UCEs difficult or impossible to isolate (Reneker *et al.*, 2012). Target sequence capture methods (TSC) and exome sequencing currently seem to be the most promising methods for use in this field. TSC methods, which use probes or baits to preferentially amplify and sequence specific regions of the genome (often low or single copy nuclear genes) have been shown to be effective at resolving phylogenetic relationships across ferns (Wolf *et al.*, 2018), and several methodologies are available for developing baits from transcriptome or genome sequences (Wolf *et al.*, 2018; Yang and Smith, 2014; Zimmer and Wen,

2015). Exome capture kits have been designed for several polyploid crop plants (Warr *et al.*, 2015), and could potentially be useful for fern population genetics as well. Both TSC and exome capture methods provide large amounts of data that should be informative across the evolutionary time scales needed to investigate the hybrid complexes that are so common in ferns, such as in North American *Dryopteris*.

Acknowledgements

We thank Paul Wolf and Michael Barker for the invitation to submit a manuscript for this special issue. We also thank Adam Payton and Stuart McDaniel (UF) for their assistance with ddRAD library preparation, and Zach Gompert (USU) for help with the analysis pipeline. EBS is funded by NSF DEB 1541506, IOS 1754911, and DBI 1802134. SPK is funded by an NSF Graduate Research Fellowship.

Literature Cited

Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe. 2016. Harnessing the Power of RADseq for Ecological and Evolutionary Genomics. Nature Reviews. Genetics 17:81–92.

Bainard, J. D., T. A. Henry, L. D. Bainard, and S. G. Newmaster. 2011. DNA Content Variation in Monilophytes and Lycophytes: Large Genomes That Are Not Endopolyploid. Chromosome Research 19:763–75.

Beck, J. B., M. D. Windham, G. Yatskievych, and K. M. Pryer. 2010. A Diploids-First Approach to Species Delimitation and Interpreting Polyploid Evolution in the Fern Genus *Astrolepis* (Pteridaceae). Systematic Botany 35:223–34.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, and E. S. Buckler. 2007. TASSEL: Software for Association Mapping of Complex Traits in Diverse Samples. Bioinformatics 23:2633–35.

Broad Institute. 2019. Picard Toolkit (version 2.9.0). Computer software. Broad Institute, GitHub Repository.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: An Analysis Tool Set for Population Genomics. Molecular Ecology 22:3124–3140.

Danecek, P., A. Auton, G. Abecasis, C.A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and 1000 Genomes Project Analysis Group. 2011. The Variant Call Format and VCFtools. Bioinformatics 27:2156–58.

Eaton, D. A. R., E. L. Spriggs, B. Park, and M. J. Donoghue. 2017. Misconceptions on Missing Data in RAD-Seq Phylogenetics with a Deep-Scale Example from Flowering Plants. Systematic Biology 66:399–412.

Gastony, G. J., and G. Yatskievych. 1992. Maternal Inheritance of the Chloroplast and Mitochondrial Genomes in Cheilanthoid Ferns. American Journal of Botany 79:716.

Gompert, Z., L. K. Lucas, C. A. Buerkle, M. L. Forister, J. A. Fordyce, and C. C. Nice. 2014. Admixture and the Organization of Genetic Diversity in a Butterfly Species Complex Revealed through Common and Rare Genetic Variants. Molecular Ecology 23:4555–73.

Grusz, A. L., M. D. Windham, and K. M. Pryer. 2009. Deciphering the Origins of Apomictic Polyploids in the *Cheilanthes yavapensis* Complex (Pteridaceae). American Journal of Botany 96:1636–45.

Ishikawa, H., Y. Watano, K. Kano, M. Ito, and S. Kurita. 2002. Development of Primer Sets for PCR Amplification of the *PgiC* Gene in Ferns. Journal of Plant Research 115:65–70.

Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, Y. Hu, H. Liang, P. S. Soltis, D. E. Soltis, S. W. Clifton, et al. 2011. Ancestral Polyploidy in Seed Plants and Angiosperms. Nature 473:97–100.

Jombart, T. 2008. Adegenet: A R Package for the Multivariate Analysis of Genetic Markers. Bioinformatics 24:1403–5.

Jorgensen, S. A., and D. S. Barrington. 2017. Two Beringian Origins for the Allotetraploid Fern *Polystichum braunii* (Dryopteridaceae). Systematic Botany 42:6–16.

Li, F.-W., P. Brouwer, L. Carretero-Paulet, S. Cheng, J. de Vries, P.-M. Delaux, A. Eily, N. Koppers, L.-Y. Kuo, Z. Li, M. Simenc, I. Small, et al. 2018. Fern Genomes Elucidate Land Plant Evolution and Cyanobacterial Symbioses. Nature Plants 4:460–72.

Li, H., and R. Durbin. 2009. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. Bioinformatics 25:1754–60.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map Format and SAMtools. Bioinformatics 25:2078–79.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data. Genome Research 20:1297–1303.

Montgomery, J., and W. H. Wagner. 1993. *Dryopteris. in Flora of North America North of Mexico*. Vol. 2. Oxford University Press, New York, New York.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra. 2012. Double Digest RADseq: An Inexpensive Method for de Novo SNP Discovery and Genotyping in Model and Non-Model Species. Plos One 7:e37135.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of Population Structure Using Multilocus Genotype Data. Genetics 155:945–59.

R Development Core Team. 2016. R: A Language and Environment for Statistical Computing. (version 3.3.2). Computer software. R Foundation for Statistical Computing, Vienna, Austria.

Reneker, J., E. Lyons, G. C. Conant, J. C. Pires, M. Freeling, C.-R. Shyu, and D. Korkin. 2012. Long Identical Multispecies Elements in Plant and Animal Genomes. Proceedings of the National Academy of Sciences of the United States of America 109:E1183–91.

Rognes, T., T. Flouri, B. Nichols, C. Quince, and F. Mahé. 2016. VSEARCH: A Versatile Open Source Tool for Metagenomics. PeerJ 4:e2584.

Rothfels, C. J., A. K. Johnson, P. H. Hovenkamp, D. L. Swofford, H. C. Roskam, C. R. Fraser-Jenkins, M. D. Windham, and K. M. Pryer. 2015. Natural Hybridization between Genera That Diverged from Each Other Approximately 60 Million Years Ago. The American Naturalist 185:433–42.

Rothfels, C. J., F.-W. Li, E. M. Sigel, L. Huiet, A. Larsson, D. O. Burge, M. Ruhsam, M. Deyholos, D. E. Soltis, C. N. Stewart, S. W. Shaw, L. Pokorny, et al. 2015. The Evolutionary History of Ferns Inferred from 25 Low-Copy Nuclear Genes. American Journal of Botany 102:1089–1107.

Rothfels, C. J., K. M. Pryer, and F.-W. Li. 2017. Next-Generation Polyploid Phylogenetics: Rapid Resolution of Hybrid Polyploid Complexes Using PacBio Single-Molecule Sequencing. The New Phytologist 213:413–29.

Rowe, H. C., S. Renaut, and A. Guggisberg. 2011. RAD in the Realm of Next-Generation Sequencing Technologies. Molecular Ecology 20:3499–3502.

Schuettpelz, E., A. L. Grusz, M. D. Windham, and K. M. Pryer. 2008. The Utility of Nuclear *GapCp* in Resolving Polyploid Fern Origins. Systematic Botany 33:621–29.

Sessa, E. B., M. Vicent, S. Chambers, and J. Gabriel y Galán. 2018. Evolution and Reciprocal Origins in Mediterranean Ferns: The *Asplenium obovatum* and *A. adiantum-nigrum* Complexes. Annals of the Missouri Botanical Garden 103:13–25.

Sessa, E. B., J. A. Banks, M. S. Barker, J. P. Der, A. M. Duffy, S. W. Graham, M. Hasebe, J. Langdale, F.-W. Li, D. B. Marchant, K. M. Pryer, C. J. Rothfels, et al. 2014. Between Two Fern Genomes. GigaScience 3:15.

Sessa, E. B., and T. J. Givnish. 2014. Leaf Form and Photosynthetic Physiology of *Dryopteris* Species Distributed along Light Gradients in Eastern North America. Functional Ecology 28:108–23.

Sessa, E. B., L.-B. Zhang, H. Väre, and A. Juslén. 2015. What We Do (and Don't) Know about Ferns: *Dryopteris* (Dryopteridaceae) as a Case Study. Systematic Botany 40:387–99.

Sessa, E. B., E. A. Zimmer, and T. J. Givnish. 2012a. Phylogeny, Divergence Times, and Historical Biogeography of New World *Dryopteris* (Dryopteridaceae). American Journal of Botany 99:730–50.

Sessa, E. B., E. A. Zimmer, and T .J. Givnish. 2012b. Unraveling Reticulate Evolution in North American *Dryopteris* (Dryopteridaceae). BMC Evolutionary Biology 12:104.

Sessa, E. B., E. A. Zimmer, and T. J. Givnish. 2012c. Reticulate Evolution on a Global Scale: A Nuclear Phylogeny for New World *Dryopteris* (Dryopteridaceae). Molecular Phylogenetics and Evolution 64:563–81.

Soltis, D. E., C. J. Visger, and P. S. Soltis. 2014. The Polyploidy Revolution Then. . .and Now: Stebbins Revisited. American Journal of Botany 101:1057–78.

Testo, W. L., J. E. Watkins, and D. S. Barrington. 2015. Dynamics of Asymmetrical Hybridization in North American Wood Ferns: Reconciling Patterns of Inheritance with Gametophyte Reproductive Biology. The New Phytologist 206:785–95.

Thiers, B. 2018 [Continuously Updated] Index Herbariorum: A Global Directory of Public Herbaria and Associated Staff. New York Botanical Garden's Virtual Herbarium. Website http://sweetgum.nybg.org/science/ih/. [accessed 18 December 2018].

Vogel, J. C., S. J. Russell, F. J. Rumsey, J. A. Barrett, and M. Gibby. 1998. Evidence for Maternal Transmission of Chloroplast DNA in the Genus *Asplenium* (Aspleniaceae, Pteridophyta). Botanica Acta 111:247–49.

Walker, S. 1955. Cytogenetic Studies in the *Dryopteris spinulosa* Complex I. Watsonia 3:193–209.

Walker, S. 1959. Cytotaxonomic Studies of Some American Species of *Dryopteris*. American Fern Journal 49:104.

Walker, S. 1961. Cytogenetic Studies in the *Dryopteris spinulosa* Complex. II. American Journal of Botany 48:607.

Walker, S. 1962. Further Studies in the Genus *Dryopteris*: The Origin of *D. clintoniana*, *D. celsa*, and Related Taxa. American Journal of Botany 49:497–503.

Walker, S. 1969. Identification of a Diploid Ancestral Genome in the *Dryopteris spinulosa* Complex. British Fern Gazette 10:97–99.

Warr, A., C. Robert, D. Hume, A. Archibald, N. Deeb, and M. Watson. 2015. Exome Sequencing: Current and Future Perspectives. G3 5:1543–50.

Wolf, P. G., T. A. Robison, M. G. Johnson, M. A. Sundue, W. L. Testo, and C. J. Rothfels. 2018. Target Sequence Capture of Nuclear-Encoded Genes for Phylogenetic Analysis in Ferns. Applications in Plant Sciences 6:e01148.

Wood, T. E., N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, and L. H. Rieseberg. 2009. The Frequency of Polyploid Speciation in Vascular Plants. Proceedings of the National Academy of Sciences of the United States of America 106:13875–79.

Yang, Y., and S. A. Smith. 2014. Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. Molecular Biology and Evolution 31:3081–92.

Zhang, J., K. Kobert, T. Flouri, and A. Stamatakis. 2014. PEAR: A Fast and Accurate Illumina Paired-End ReAd MergeR. Bioinformatics 30:614–20.

Zhao, P., H.-J. Zhou, D. Potter, Y.-H. Hu, X.-J. Feng, M. Dang, L. Feng, S. Zulfiqar, W.-Z. Liu, G.-F. Zhao, and K. Woeste. 2018. Population Genetics, Phylogenomics and Hybrid Speciation of Juglans in China Determined from Whole Chloroplast Genomes, Transcriptomes, and Genotyping-by-Sequencing (GBS). Molecular Phylogenetics and Evolution 126:250–65.

Zimmer, E. A., and J. Wen. 2015. Using Nuclear Gene Data for Plant Phylogenetics: Progress and Prospects II. Next-Gen Approaches. Journal of Systematics and Evolution 53:371–79.