An Exploration into Fern Genome Space

Paul G. Wolf1,*, Emily B. Sessa2,3, Daniel Blaine Marchant2,3,4, Fay-Wei Li5, Carl J. Rothfels6, Erin M. Sigel5,9, Matthew A. Gitzendanner2,3,4, Clayton J. Visger2,3,4, Jo Ann Banks7, Douglas E. Soltis2,3,4, Pamela S. Soltis3,4, Kathleen M. Pryer5, and Joshua P. Der8

1Ecology Center and Department of Biology, Utah State University
2Department of Biology, University of Florida
3Genetics Institute, University of Florida
4Florida Museum of Natural History, University of Florida
5Department of Biology, Duke University
6University Herbarium and Department of Integrative Biology, University of California, Berkeley
7Department of Botany and Plant Pathology, Purdue University
8Department of Biological Science, California State University, Fullerton
9Present address: Department of Botany, National Museum of Natural History, Smithsonian Institution, Washington, District of Columbia
*Corresponding author: E-mail: paul.wolf@usu.edu.

Accepted: August 12, 2015

Data deposition: Full data sets for this project, including sequence reads and assemblies have been deposited at the Utah State University Digital Commons with the DOI 10.15142/T39G67 (http://dx.doi.org/10.15142/T39G67). Annotated plastomes have been deposited at GenBank under the accessions KP136829-KP136832, KM052729, and HM535629.

Abstract

Ferns are one of the few remaining major clades of land plants for which a complete genome sequence is lacking. Knowledge of genome space in ferns will enable broad-scale comparative analyses of land plant genes and genomes, provide insights into genome evolution across green plants, and shed light on genetic and genomic features that characterize ferns, such as their high chromosome numbers and large genome sizes. As part of an initial exploration into fern genome space, we used a whole genome shotgun sequencing approach to obtain low-density coverage (~0.4X to 2X) for six fern species from the Polypodiales (Ceratopteris, Pteridium, Polypodium, Cystopteris), Cyatheales (Plagiogyria), and Gleicheniales (Dipteris). We explore these data to characterize the proportion of the nuclear genome represented by repetitive sequences (including DNA transposons, retrotransposons, ribosomal DNA, and simple repeats) and protein-coding genes, and to extract chloroplast and mitochondrial genome sequences. Such initial sweeps of fern genomes can provide information useful for selecting a promising candidate fern species for whole genome sequencing. We also describe variation of genomic traits across our sample and highlight some differences and similarities in repeat structure between ferns and seed plants.

Key words: comparative genomics, plastome, chloroplast, mitochondria, repeat content, transposons.

Introduction

Recent advances in DNA sequencing technology and improvements in assembly strategies are resulting in rapid growth in the availability of genome sequences for nonmodel species. Currently, genome sequences are available for over 100 vascular plants, including one lycopsid, three gymnosperms, and numerous crop and noncrop angiosperms (Michael and VanBuren 2015). However, genomic resources in other major clades of vascular plants are lagging. The sister group of seed plants is the fern clade (Monilophyta sensu Cantino et al. 2007); these two lineages diverged from a common ancestor approximately 380 Ma (Schneider et al. 2004). Ferns in the broad sense include horsetails, whisk ferns and ophioglossoid ferns, marattioid ferns, and leptosporangiate ferns. The latter lineage is by far the most diverse, with about 9,000 species (Smith et al. 2006) that occupy many key ecosystems, and comprise, for example, a significant
component of tropical forest understories and canopies. There are currently no nuclear genome sequences available for any fern, despite the richness of this clade and its key evolutionary position as sister to the seed plants. Having such a reference sequence, or any information about the content and structure of fern genomes, will enable investigation of several perplexing features of fern biology and evolution, and will facilitate comparative analyses of genome structure and function across vascular plants as a whole.

Ferns differ fundamentally from seed plants in several key biological and genomic features. For example, ferns alternate between free-living, independent gametophyte (haploid) and sporophyte (diploid) phases, whereas the gametophyte phase in seed plants is dependent on the sporophyte and is highly reduced. Thus, a large component of the fern genome is presumably expressed only in the haploid phase. Furthermore, most ferns are homosporous (apart from one heterosporous clade), whereas all seed plants are heterosporous. This characteristic is apparently correlated with chromosome number across tracheophytes in that homosporous taxa uniformly have more chromosomes. For example, chromosome numbers average \( n = 57.05 \) for homosporous ferns and lycophytes, \( n = 15.99 \) for flowering plants, and \( n = 13.62 \) for heterosporous ferns and lycophytes (Klekowsk and Baker 1966). The underlying cause of this association between homosporosity and high chromosome number is not understood.

Ferns are the only lineage of land plants for which there is a strong positive correlation between chromosome number and genome size (Nakazato et al. 2008; Bainard et al. 2011). Whether this also extends to lycophytes is not yet clear, but this pattern has not been reported from any other group of eukaryotic organisms, and suggests that fern nuclear genomes may possess unique structural characteristics. Ferns are prone to polyploidization (Wood et al. 2009) but may undergo different polyploidization processes that are distinct from those in other lineages of land plants (Barker and Wolf 2010; Leitch AR and Leitch IJ 2012). Information on the nature and relative proportions of various components of fern genomes will help to establish how this group of plants can be used in studies of genome evolution and dynamics across land plants. If fern genomes respond uniquely to changes in genome size, then they could provide useful control models for the study of genome downsizing following whole genome duplication (Leitch and Bennett 2004).

Gathering information on fern genomes will provide an improved phylogenetic context for investigating evolutionary questions across land plants. For example, knowledge of fern genome content and structure may shed light on the transition from homospory to heterospory that has occurred several times during the evolution of land plants. Ferns are also the most appropriate outgroup for understanding genome structure and evolution in their sister clade, the seed plants.

The research community would benefit from well-assembled, annotated nuclear genomes from several leptosporangiate ferns, as well as representatives of the other early-diverging fern clades (Li and Pryer 2014; Sessa et al., 2014; Schneider et al. 2015). Such nuclear genome sequences are necessary for rigorous tests of most questions about genome and chromosome structure and evolution, and addressing these questions currently awaits completion of one or more fern genome sequencing projects (Li and Pryer 2014; Sessa et al. 2014). Meanwhile, low-coverage genome scans can be used to begin uncovering broad patterns of fern genome content, allowing, for example, preliminary estimates of protein-coding and repetitive content. Here we use such scans at ~0.4–2X coverage for species from six different fern lineages, each representing a major leptosporangiate clade (fig. 1). We use these data to ask how much variation exists in gene and protein content across Ferns, and we compare these with data from existing angiosperm and gymnosperm genome sequencing projects. Although our focus is on nuclear genomes of ferns, the data obtained include many organelar genome sequences, and we use these to assemble plastomes and identify contigs carrying putative mitochondrial genes. The latter are the first such resources available for ferns, for which no mitochondrial genome has been sequenced to date.

Materials and Methods

Samples

We selected six leptosporangiate ferns from across a range of major clades (fig. 1): Dipters conjugata (Gleicheniales), Plagiogyria formosana (Cystheales), Pteridium aquilinum (Dennstaedtiaceae), Ceratopteris richardii (Pteridaceae), Polypodium glycyrrhiza (eupolypods I), and Cyatopteris protrusa (eupolypods II). Details of species used, collections, and vouchers are provided in table 1.

Genome Size Estimation

Genome size estimates for three taxa were derived from the literature: Po. glycyrrhiza (Murray 1985), Pt. aquilinum (Bainard et al. 2011), and Cy. protrusa (Bainard et al. 2011, for the related diploid, Cy. bulbifera). We estimated genome size for Pt. aquilinum by chopping approximately 0.75 cm\(^2\) of fresh fern leaf tissue and 0.5 cm\(^2\) of the standard Vicia faba (26.9 pg) or Pisum sativum (9.09 pg; Doležel et al. 1998) on a chilled surface using a fresh razor blade and adding 500 μl of ice-cold extraction buffer (0.1 M citric acid, 0.5% w/v Triton X-100) (Hanson et al. 2005) with 1% w/v PVP-40 (Yokoya et al. 2000). Tissue was chopped into a semifine slurry, and the resulting mixture was swirled by hand until the liquid obtained a light-green tinge. The suspension was poured through a cell strainer (BD Falcon; Becton, Dickinson and Company, Franklin Lakes, NJ). We added RNase A (1 mg/ml) and 350 μl of propidium iodide staining solution (0.4 M NaPO₄, 10 mM sodium citrate, 25 mM sodium sulfate, 50 μg/ml propidium iodide) to 140 μl of filtrate, incubated...
the solution at 25 °C for 30 min, and then stored it for up to 4 h on ice. We ran the stained solutions on an Accuri C6 (BD Biosciences, Franklin Lake, NJ) using a 488 nm laser and captured 10,000 events. For estimating genome size of Pl. formosana and D. conjugata, we carried out flow cytometry analyses following the protocol of Ebihara et al. (2005) on a BD FACSCan system (BD Biosciences). We calculated the relative genome content using the ratio of the mean fluorescent peak of the sample to the internal standard, multiplied by the genome size of the standard, and converted to an estimate of the number of bases using 1 pg = 980 Mb.

DNA Sequencing

Genomic libraries for Po. glycyrrhiza, Cy. protrusa, Pl. formosana, and D. conjugata were prepared with the KAPA Illumina library preparation kit (KAPA Biosystems, Wilmington, MA) using fragment sizes of 300–400 bp. Barcodes were added with the NEBNext Multiplex Oligos for Illumina kit (New England BioLabs, Ipswich, MA).

Table 1

<table>
<thead>
<tr>
<th>Species</th>
<th>Collection Locality</th>
<th>Voucher (herbarium)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Dipteris conjugata</em> (Kaulf.) Reinw.</td>
<td>Pahang, Malaysia</td>
<td>Schuettpelz 770 (DUKE)</td>
</tr>
<tr>
<td><em>Cystopteris prostrusa</em> (Weatherby) Blasdell</td>
<td>Ashe County, North Carolina, USA</td>
<td>Rothfels 4168 (DUKE)</td>
</tr>
<tr>
<td><em>Plagiogyria formosana</em> Nakai</td>
<td>Nantou County, Taiwan</td>
<td>Schuettpelz 1083A (DUKE)</td>
</tr>
<tr>
<td><em>Ceratopteris richardii</em> Brongn.</td>
<td>Cuba (Accession: Hnn)</td>
<td>dbmarchant01 (FLAS)</td>
</tr>
<tr>
<td><em>Pteridium aquilinum</em> L. Kuhn</td>
<td>Manchester, UK</td>
<td>E. Sheffield S48 (UTC)</td>
</tr>
<tr>
<td><em>Polypodium glycyrrhiza</em> D.C. Eaton</td>
<td>Squamish-Lillooet, British Columbia, Canada</td>
<td>Rothfels 4086 (DUKE)</td>
</tr>
</tbody>
</table>

Fig. 1.—Phylogeny of ferns summarized from Pryer et al. (2004). Numbers of sequenced nuclear genomes are indicated for the lineages that have them. Lineages in pink are the eusporangiate ferns; the leptosporangiate fern clade is in green. Taxa in this study are given in parentheses. Photos of representative ferns are included: (A) *Ceratopteris richardii* (Pteridaceae); (B) *Salvinia* sp., Salviniales (heterosporous water ferns); (C) *Cystopteris prostrusa* (Cystopteridaceae); (D) *Ophioglossum* sp., Ophioglossales (rattlesnake ferns); and (E) *Equisetum* sp. (horsetails). The timescale along the bottom of the phylogeny is in millions of years before present.
England Biolabs, Ipswich, MA). These four taxa were run on a single lane of Illumina HiSeq 2500. Libraries for *Pt. aquilinum* and *Ce. richardii* (average insert size = 300 bp) were prepared using the Illumina TruSeq library preparation kit (Illumina, San Diego, CA) and run together on a second lane of Illumina HiSeq 2500. All Illumina sequencing (paired-end reads) was performed at the Duke University Center for Genomic and Computational Biology, which also performed pre-run library quality control. Illumina data for *Pt. aquilinum* were supplemented with sequences from a Roche 454 GS-FLX Titanium run (Der, 2010).

Data Processing and Assembly

Unwanted adapter sequences were removed from our Illumina reads using cutadapt (Martin 2011; Del Fabbro et al. 2013). We used Sickle (Joshi and Fass 2011) to assess read quality using a sliding window approach. Sections of reads with an average quality score of <Q25 were trimmed, and reads with <50 bp remaining were also removed. Sequence data were sorted by barcode. Quality-trimmed reads for each species were assembled into contigs using CLC Assembly Cell (v4.2.1), specifying a library insert size between 275 and 425 bp (for paired-end reads) and a word size (kmer length) of 31 bp.

Assembly and Analysis of Organellar Genome Components

To assess the structure of the nuclear genomes, we first separated plastid and mitochondrial contigs, which comprised a portion of the assemblies. To identify plastid contigs, we performed BLASTX (Altschul et al. 1997) searches (using an e-value threshold of 1e-10) of the CLC contigs against a custom database of fern proteins extracted from complete sequenced plastomes obtained from GenBank: *Adiantum* (NC_004766), *Pteridium* (NC_014348), *Angiopteris* (NC_008829), *Lygodium* (NC_024153), *Alsophila* (NC_012818), and *Cheilanthes* (NC_014592). The remaining non-plastome contigs were then queried against several plant mitochondrial genomes (there are currently no fern mitochondrial genomes available for such searches): *Zea mays* (NC_007982), *Pleurozia purpurea* (NC_013444), *Nicothiana tabacum* (NC_006581), *Mesostigma vinkei* (NC_008240), *Megaceros aenigmaticus* (NC_012651), *Marchantia polymorpha* (NC_001660), *Cyclas taubungensis* (NC_010303), *Chara vulgaris* (NC_005255), *Arabidopsis thaliana* (NC_001284), *Physcomitrella patens* (NC_007945), and *Vitis vinifera* (NC_012119). We searched against both the above complete mitochondrial genome sequences plus a collection of core mitochondrial genes conserved across 27 plant mitochondrial genomes. Details for extracting these core genes are provided at Digital Commons (http://dx.doi.org/10.15142/T39G67, last accessed September 1, 2015).

The putative plastome contigs from the initial BLASTX search were used to build, iteratively and manually, plastome assemblies in Geneious v7.1 (Biomatters, Auckland, New Zealand). First, we performed manual reference-guided alignments to the most closely related available fern plastome, to orient and order contigs based on the general structure of the reference plastome. The boundaries of the inverted repeat (IR) were manually identified using small cut and paste alignments in Geneious. Next, we used Mauve (Darling et al. 2004) to align these rough plastome assemblies to one or more published reference sequences (listed above). We then transferred preliminary gene annotations from the references to the new assemblies, and manually adjusted reading frames, introns, and putative RNA editing sites. We then used these plastome assemblies as queries in another round of BLASTX searches with the entire CLC contig set as the database, to identify additional possible plastome contigs, or sections of plastid DNA inserted in nuclear or mitochondrial contigs. We then filtered contigs, retaining those with >95% sequence similarity and >90% of the contig length with a match to the plastome assembly. This was to exclude possible small portions of plastid DNA that had been inserted into the nuclear genomes. Any additional contigs not incorporated into the plastome assembly were removed from subsequent analyses of the plastome. Annotated plastomes were deposited on GenBank, and contigs identified as putatively containing mitochondrial genes are available at Digital Commons (http://dx.doi.org/10.15142/T39G67, last accessed September 1, 2015). We further explored possible organellar contigs by examining the average weighted (by contig length) depth of coverage of putative organellar contigs relative to the entire assembly. Because of the higher copy number of organellar relative to nuclear genomes in cells, we expect organellar sequences to be detected at higher depth. Thus, low depth organellar sequences can indicate possible regions that have been transferred to the nucleus. All contigs remaining in the original assemblies after the removal of plastid and mitochondrial DNA were considered to be nuclear genomic DNA and were used to estimate repeat content in the nuclear genome.

Assessing Repeat Content in the Nuclear Genome

To assess repeat structure, we analyzed repeat content in these six fern genomes and made comparisons with six seed plants. We downloaded data sets from six phylogenetically representative seed plants from the NCBI SRA (Amborella trichopoda, V. vinifera, Z. mays, Gnetum gnemnon, Pinus taeda, and Taxus baccata). *Pinus taeda* reads were first trimmed to 100 bp and reads for all 12 taxa were quality-filtered using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), last accessed September 1, 2015) to exclude reads that did not contain at least 70% of the bases with quality scores higher than Q20. Quality-filtered paired-end reads were reassociated and interleaved. Three replicate samples of 0.05X coverage of
the estimated genome size (approximately 5% of each genome) were used in subsequent analyses to identify highly represented repeat clusters using RepeatExplorer (Novák et al. 2013). Default parameters were used in RepeatExplorer analyses, except that paired reads were used and a domain search was performed using an e-value threshold of $1 \times 10^{-5}$. The genome proportions represented by ten sequence-based repeat classes, plus unknown repeats and nonrepetitive sequences, were summed for each species based on the most abundant RepeatMasker (http://www.repeatmasker.org, last accessed September 1, 2015) hits identified for each cluster.

We used standard linear regressions, performed in R (R Core Team, 2014), to test for correlations between genome size and genome proportions inferred to belong to the ten repeat classes, for ferns and seed plants separately. Differences in genomic repeat content between clades were assessed using one-way analysis of variance for a completely randomized design with subsamples. Cline was incorporated in the model as a fixed-effects factor, and species within clades as a random effects factor; triplet samples of each species were considered to be subsamples. Each repeat class was analyzed separately, and proportion data were logit-transformed prior to analysis to better meet assumptions of normality and homogeneity of variance. Data calculations were made using the GLIMMIX procedure in SAS/STAT 13.2 in the SAS System for Windows Version 9.4. This statistical model assumes that our choice of species was a random sample from their respective clades; although not strictly true, we chose species to be representative of the breadth within clades and so are comfortable using these analyses as exploratory indicators.

Assessing Protein-Coding Content

We estimated the size of the protein-coding portion of the nuclear genome in two ways: One based on the proportion of assembled contigs with protein hits, and the other based on the proportion of reads that contained BLASTX hits to a protein database. First, we used BLASTX to query all (previously assembled) nonorganellar contigs against a database of annotated protein sequences from 22 plant species (Amborella Genome Project 2013). We then calculated the total length of BLASTX hits longer than 100 bp and divided this by the total assembly length to get estimates of the proportion of each assembly with protein hits. Because low-copy genes are less likely to be represented in assembled contigs, we also used a second read-based approach. Working with the original read files for each species, we used Bowtie2 (Langmead and Salzberg 2012) to map the reads against the organellar assemblies (plastome assemblies and mitochondrial contigs) as described above. We then removed all reads with organellar hits from the original read file and selected a random sample of the remaining reads to represent an estimated 0.0025X of the genome. We partitioned these random samples of reads into ten equal sets and queried each sequence against a database of annotated protein sequences from 22 plant species (Amborella Genome Project 2013) using BLASTX (e-value $< 1 \times 10^{-5}$). The numbers of reads in each set with hits to known plant proteins were used to calculate mean protein-coding coverage within each set and standard deviation across the ten sets for each species.

## Results

### Genome Sizes and Assemblies

Our estimates of genome size ranged from 2.45 to 15.65 Gb (table 2). Illumina sequencing generated between 40,830,366 and 207,771,644 raw reads per species, and between ~4.0 and 19.4 billion bp of quality-filtered data (table 2). For *Pteridium*, we also included an additional set of 454 data from a previous study (Der 2010), for a total of almost 20 billion bp of data (table 2). Guanine–cytosine content in the assemblies ranged from 37.9% to 620.5 million bp in total genome size (Gb) from flow cytometry.

### Table 2

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Number of Raw Reads</th>
<th>Base Pair of Clean Data</th>
<th>GC Content</th>
<th>Number of N’s in assembly</th>
<th>Genome Size (Gb) from Flow Cytometry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cystopteris</td>
<td>47,005,020</td>
<td>4,580,023,307</td>
<td>0.42</td>
<td>645,022</td>
<td>4.23</td>
</tr>
<tr>
<td>Diphtheris</td>
<td>51,232,072</td>
<td>5,023,794,762</td>
<td>0.42</td>
<td>1,825,278</td>
<td>2.45</td>
</tr>
<tr>
<td>Plagioiugria</td>
<td>58,488,796</td>
<td>5,717,123,738</td>
<td>0.43</td>
<td>1,376,069</td>
<td>14.81</td>
</tr>
<tr>
<td>Polyododium</td>
<td>40,830,366</td>
<td>4,000,482,565</td>
<td>0.43</td>
<td>503,779</td>
<td>10.02</td>
</tr>
<tr>
<td>Ceratopteris</td>
<td>204,001,778</td>
<td>19,445,093,728</td>
<td>0.38</td>
<td>8,287,893</td>
<td>11.25</td>
</tr>
<tr>
<td>Pteridium 454</td>
<td></td>
<td>216,194,085</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pteridium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pteridium</td>
<td>207,771,644</td>
<td>19,437,952,758</td>
<td>0.39</td>
<td>24,208,821</td>
<td>15.65</td>
</tr>
<tr>
<td>Pteridium Both</td>
<td>208,482,822</td>
<td>19,654,146,843</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>610,040,854</td>
<td>58,420,664,943</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.—Clean data refer to reads processed by removing adapters and trimming low-quality regions. GC, guanine–cytosine.
Organellar Genomes

We assembled and annotated plastomes, consisting of 3–4 contigs for each species, and ranging from 123,674 to 158,508 total bp in length (including both copies of the IR; table 4). Plastome sequences are archived in GenBank (see table 4 for accession numbers). Each of the six species sampled here appears to have the same gene order as observed in its nearest relatives with previously published plastomes.

We also detected regions with strong sequence similarity to plastid DNA in contigs that did not appear to be part of the plastome assembly. We infer these to be plastome-like genes residing within the nucleus (or perhaps mitochondrial) genomes. For each fern species, we detected 20–241 contigs containing such regions, ranging from a total of about 10–55 kb of plastome-like sequences (table 4). This amounted to no more than 0.025% of the presumed nuclear genome assembly for each species (table 6). Contigs containing these sequences are available from Digital Commons: http://dx.doi.org/10.15142/T39G67, (last accessed September 1, 2015).

Repeat Content Analyses

We compared genomic repeat content of our sample of ferns with a similar data set of seed plants. We report here possible differences that can be examined further in the future when high coverage assemblies become available for more taxa. Compared with seed plants, ferns had a higher proportion of their genomes in three main repeat classes (fig. 3): DNA transposons (mean ± standard error of 3.2 ± 0.72% in ferns; 0.83 ± 0.19% in seed plants, p(F) = 0.001), long interspersed nuclear elements (henceforth LINES; 2.2 ± 0.75% in ferns; 0.49 ± 0.17% in seed plants, p(F) = 0.006), and simple repeats (15.5 ± 1.5% in ferns; 11.9 ± 0.89% in seed plants, p(F) = 0.007). Satellite DNA (comprising of tandem arrays, including centromeres and telomeres) was on average lower in ferns (0.1 ± 0.03%) compared with seed plants (0.8 ± 0.34%), but both groups in our analyses are low for this class, and the differences were not significant (p(F) = 0.214); differences for all other repeat classes were also not significant (p(F) > 0.1). Figure 4 illustrates relative proportions of the genome for each class of repeat, the two estimates of protein-coding content (see below), and the remaining nonrepetitive component, versus genome size. These plots reveal the similarity across the three samples from each taxon, indicating that our subsampling method is effective. Standard linear regressions revealed that genome size is not significantly correlated with the size of any class of repetitive element.

Protein-Coding Content Analyses

Based on sampling reads representing 0.0025X of each genome, we estimated the protein-coding content as 2.85 ± 0.03% (Pteridium) to 6.61 ± 0.03% (Ceratopteris) of the reads (table 7 and fig. 4). Estimates obtained by examining all nonorganellar contigs in the assemblies were lower, ranging from 1.11% (Pteridium and Ceratopteris) to 1.90%
Discussion

Here we describe the first investigation into the comparative genome content of ferns. Although low-coverage genome scans are unsuitable for obtaining high-quality genome assemblies or revealing higher level aspects of genome structure, these data can provide key initial insights into genome content (Rasmussen and Noor 2009; Weitemier et al. 2014). Based on these first analyses of six species, it appears that several aspects of genome content are consistent across ferns, some of which appear to be similar to seed plants, while other aspects may be unique to ferns.

The six species we sampled have genome sizes typical for homosporous leptosporangiate ferns. Estimates of fern genome sizes range from 0.77 pg for *Azolla microphylla* (heterosporous leptosporangiate) to 65.55 pg for *Ophioglossum reticulatum* and 72.68 pg for *Psilotum nudum* (two eusporangiate ferns; Bennett and Leitch 2001; Obermayer et al. 2002). Our six species have genome sizes of 2.45–15.65 pg, on the lower end of the above range (table 2). The three species for which we performed flow cytometry experiments, *Pl. formosana*, *Ce. richardii*, and *D. conjugata*, had no previous genome size estimates. Although several surveys of genome size have been made in ferns (Bennett and Leitch 2001, 2012; Obermayer et al. 2002; Bainard et al. 2011), published C-value estimates exist for only 104 fern species (Bennett and Leitch 2012) out of the ca. 9,000 known ferns (Smith et al. 2006, 2008). Additional studies are needed to establish the full range of genome sizes in ferns, which will be particularly useful for determining whether currently unsampled species have small genomes that may be suitable candidates for high-coverage sequencing and assembly. The large genome sizes and high chromosome numbers in ferns, and the concomitant challenges they represent for assembly, have contributed significantly to the difficulty in obtaining a reference genome for ferns (Pryer et al. 2002; Sessa et al. 2014).

Although this is the first genome-wide comparative analysis in ferns, several previous studies have made inferences about fern genome structure or content. For example, Pichersky et al. (1990) reported defective copies of chlorophyll a/b binding genes in the homosporous fern *Polystichum munitum*. The authors hypothesized that the defective genes are the result of gene silencing or loss of gene function in duplicated gene copies. Other nuclear genes with silenced copies have been detected in genetically diploid fern genomes (Gastony 1991; McGrath et al. 1994; McGrath and Hickok 1999). Although the presence of these putatively silenced genes appeared consistent with a history of paleopolyploidy in ferns, a high-resolution genetic linkage map generated for...
Ce. richardii (Nakazato et al. 2006) failed to recover evidence of large-scale synteny that would support extensive ancient polyploidy, although 76% loci were duplicated. Nakazato et al. (2006) concluded that small-scale gene duplication was likely the primary mode of duplication in Ce. richardii. Meanwhile, Rabinowicz et al. (2005) examined genome-wide methylation in Ce. richardii and found that, as in other plant groups (Bennetzen et al. 1994; Rabinowicz et al. 2003), gene-rich regions are less methylated than other genomic regions. They also determined that Ce. richardii has roughly the same numbers of genes as angiosperms, but a much lower gene density due to its larger genome size (Rabinowicz et al. 2005).

From calculations based on full, annotated genome sequences, the proportion of a genome that is protein coding varies considerably among taxa. For example, Picea abies (19.6 Gb) is made up of approximately 2.4% of protein-coding genes (Nystedt et al. 2014), whereas Utricularia gibba (83 Mb), the smallest plant genome sequenced to date, may be composed of as much as 97% protein-coding sequence (Ibarra-Laclette et al. 2013). In general, estimates of total gene number vary within an order of magnitude: Usually between 20,000–40,000 per diploid genome (Rabinowicz et al. 2005; Sterck et al. 2007). Thus, the proportion of a genome that is protein coding will tend to reflect the inverse of genome size. The only published estimate of protein-coding content in a fern genome is for Ce. richardii (11 Gb), estimated by Rabinowicz et al. (2005) to be 0.49% or 6%, depending on the approach used. This is very similar to our estimate for the same species (1.1% or 6.61%). Rabinowicz (2005) used fewer than 600 reads at an average length of > 600 bp. Therefore, their estimates of protein-coding content are similar to ours provides us with some confidence in our estimation approaches. Our estimates based on the proportion of reads were three to six times greater for all species than those based on the proportion of all assembled contigs containing protein-coding sequences (fig. 4). With our low coverage, we had expected that the assemblies might underestimate the protein-coding component. This could occur if the assemblies were biased toward repetitive parts of the genome, with the excluded, unassembled regions more likely to be single copy. Despite the difference in results from the two methods, all our estimates fall toward the low end of protein content measured in (nonfern) vascular plants, ranging from 1.11% to 1.90% or 3.07% to 6.61%. These low estimates probably reflect the relatively large genome sizes of the species we sampled. We also suspect that both our estimates could be low because of a lack of reference proteins from ferns. Although genome size and protein-coding gene density may be negatively correlated across plants in general, no such...

Exploration into Fern Genome Space

Table 6
List of Putative Mitochondrial Genes Detected in Six Fern Species

<table>
<thead>
<tr>
<th>Polypodium</th>
<th>Cystopteris</th>
<th>Dipteris</th>
<th>Pliagogyria</th>
<th>Ceratopteris</th>
<th>Pteridium</th>
</tr>
</thead>
<tbody>
<tr>
<td>atp1</td>
<td>atp1</td>
<td>atp1</td>
<td>atp1</td>
<td>atp1</td>
<td>atp1</td>
</tr>
<tr>
<td>atp4</td>
<td>atp4</td>
<td>atp4</td>
<td>atp4</td>
<td>atp4</td>
<td>atp4</td>
</tr>
<tr>
<td>atp8</td>
<td></td>
<td>atp8</td>
<td></td>
<td>atp8</td>
<td></td>
</tr>
<tr>
<td>cob</td>
<td>cob</td>
<td>Cob</td>
<td>cob</td>
<td>Cob</td>
<td></td>
</tr>
<tr>
<td>cox1</td>
<td>cox1</td>
<td>cox1</td>
<td>cox1</td>
<td>cox1</td>
<td></td>
</tr>
<tr>
<td>cox2</td>
<td>cox2</td>
<td></td>
<td>cox2</td>
<td>cox2</td>
<td></td>
</tr>
<tr>
<td>cox3</td>
<td>cox3</td>
<td></td>
<td></td>
<td>cox3</td>
<td></td>
</tr>
<tr>
<td>matR</td>
<td>matR</td>
<td>matR</td>
<td>matR</td>
<td>matR</td>
<td></td>
</tr>
<tr>
<td>nad1</td>
<td>nad1</td>
<td>nad1</td>
<td>nad1</td>
<td>nad1</td>
<td></td>
</tr>
<tr>
<td>nad2</td>
<td>nad2</td>
<td>nad2</td>
<td>nad2</td>
<td>nad2</td>
<td></td>
</tr>
<tr>
<td>nad3</td>
<td>nad3</td>
<td>nad3</td>
<td>nad3</td>
<td>nad3</td>
<td></td>
</tr>
<tr>
<td>nad4</td>
<td>nad4</td>
<td>nad4</td>
<td>nad4</td>
<td>nad4</td>
<td></td>
</tr>
<tr>
<td>nad5</td>
<td>nad5</td>
<td>nad5</td>
<td>nad5</td>
<td>nad5</td>
<td></td>
</tr>
<tr>
<td>nad6</td>
<td>nad6</td>
<td>nad6</td>
<td>nad6</td>
<td>nad6</td>
<td></td>
</tr>
<tr>
<td>nad7</td>
<td>nad7</td>
<td>nad7</td>
<td>nad7</td>
<td>nad7</td>
<td></td>
</tr>
<tr>
<td>nad9</td>
<td>nad9</td>
<td>nad9</td>
<td>nad9</td>
<td>nad9</td>
<td></td>
</tr>
<tr>
<td>rpl16</td>
<td>rpl16</td>
<td>rpl16</td>
<td>rpl16</td>
<td>rpl16</td>
<td></td>
</tr>
<tr>
<td>rpl20</td>
<td></td>
<td>rps12</td>
<td>rps12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rps2A</td>
<td>rps3</td>
<td>rps3</td>
<td>rps3</td>
<td>rps3</td>
<td></td>
</tr>
<tr>
<td>rps4</td>
<td>rps7</td>
<td>rps7</td>
<td>rps7</td>
<td>rps7</td>
<td></td>
</tr>
<tr>
<td>sdh4</td>
<td>sdh4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We detected several differences in the repetitive elements of the genomes of ferns compared with seed plants. As a group, our six fern samples had higher proportions of DNA transposons, LINES, and simple repeats, and lower proportions of satellite DNA than the seed plants examined (figs. 3 and 4). It may be premature to infer that these represent real, biologically significant differences between clades, but given the unusual characteristics of fern genomes it seems reasonable to expect that such differences may exist. Future work should focus on identifying specific subclasses of repeat elements in a broader taxon sampling, to explore patterns of genome repeat structure across land plants in more detail.

Although overall genome coverage was low, coverage for organellar reads was, as expected, much higher (table 5 and fig. 2). Thus, we were able to approach full assembly of fern plastomes and detect the majority of known plant mitochondrial genes. Resolution of assemblies might be improved with the addition of longer read sequence data (such as PacBio) or higher coverage with shorter Illumina reads, or both. Better-resolved assemblies should help to distinguish true plastid and mitochondrial genes from those that have been transferred to a different compartment (Matsuo et al. 2005). However, our low coverage data should be ample for studies requiring just the gene sequences (such as phylogenetic analyses). Currently, even reference-guided assemblies require the manual step of establishing plastome IR boundaries. Given the relative conservation of these positions, it should be possible to automate this assembly step. The plastomes that we assembled here were all from previously sampled major clades of ferns, and we detected no unusual gene order; all plastomes appeared to have structures consistent with other members of the same clades.

Detailed information for mitochondrial genomes is limited from many groups of plants, including ferns. We are aware of three studies that have examined fern mitochondrial genomes. Palmer et al. (1992) isolated restriction fragments of two fern mitochondrial genomes: Those of Equisetum arvense (>200 kb) and Onoclea sensibilis (~300 kb). The authors also detected repeat structure and several known plant mitochondrial genes. In another study, a large fosmid clone (290 kb) of the mitochondrial genome of Gleichenia dicarpa was found to contain fragments of foreign DNA including transposons, retrotransposons, and transposed introns (Grewe 2011). Several fragments of mitochondrial DNA have also been isolated from Adiantum nidis (Panarese et al. 2008). A fragment of almost 21 kb contained IRs and several genes that appear to have been transferred from the plastome. Our analysis of fern mitochondrial DNA identified a large proportion of known plant mitochondrial genes (table 6). However, coverage was not sufficient to assemble large fragments containing more than about three genes. We also cannot be sure if undetected genes are absent from mitochondrial genomes or the result of low coverage. Nevertheless, the sequences of the fragments detected provide an excellent starting point for further studies of fern mitochondrial genes and studies of horizontal gene transfer in plants.

Across sequenced plant genomes, there is a positive correlation between genome size and the proportion of a genome that is made up of repeats (Michael 2014). This is because larger genomes tend to be larger because of the presence of expanding repeat elements. However, we do not observe such a pattern in ferns. One possibility is that the range of genome sizes here (2.45–15.65 pg) is too small, compared with the full range for ferns (0.77–65.55 pg), to detect a relationship, and studies of more fern species may reveal a positive relationship. It is also possible that ferns in general, or the species we chose, do not have many recently expanding repetitive elements. A third possibility is that ferns are indeed different from other plants when it comes to patterns of genome downsizing (Barker and Wolf 2010; Leitch AR and Leitch IJ 2012). In ferns, there is a positive correlation between genome size and chromosome number (Nakazato et al. 2008; Bainard et al. 2011). In most other organisms this relationship...
has been found to be weak (Leitch and Bennett 2004) or even negative (Vinogradov 2001). This lack of a relationship has been explained by rapid (in evolutionary terms) genome downsizing following polyploidy, so that tetraploids (except very recent ones) have genome sizes less than double that of their diploid ancestors (Leitch and Bennett 2004). This downsizing presumably involves the loss of extra genomic material that is not needed. In contrast, the genomic patterns we observe in ferns suggest that they lack the mechanisms for jettisoning excessive and redundant genomic regions. If this pattern holds with examination of other species, then ferns may provide useful control cases for studies of the underlying mechanisms of genome downsizing in other lineages. High-quality assembly and annotation of a fern genome will go a long way to assist researchers in the study of plant genome dynamics.

Our low-coverage genome scans enabled us to make some general statements about the relative content of homosporous, leptosporangiate fern genomes. However, these plants differ in many ways from the heterosporous land plants that have been examined to date, and higher coverage assemblies are critical for detailed comparative analyses of fern and land plant genome structure. Such studies are essential for addressing questions about the evolution of land plant genomes. Furthermore, ferns are the sister group to seed plants (Pryer et al. 2001) and evolutionarily comparative statements about seed plant genomes would benefit from comparisons with a fern genome. Currently, researchers are assembling the first fern genome, that of the heterosporous fern *Azolla* (Li and Pryer 2014). Also underway is higher coverage assembly of the model homosporous fern, *C. richardi* (Sessa et al. 2014; Marchant et al. unpublished data). Meanwhile, here we

![Figure 3](http://gbe.oxfordjournals.org/)

**Fig. 3.—** Genome proportions represented by ten sequence-based repeat classes, plus unknown repeats and nonrepetitive sequences, in six fern and six seed plant taxa, with three samples per taxon.
have presented a first exploration into comparative genome content of ferns.

**Acknowledgments**

This work was supported by the National Science Foundation Doctoral Dissertation Improvement Grants DEB-1407158 (to K.M.P and F.W.L) and DEB-1110767 (to K.M.P and C.J.R.), and a National Science Foundation Graduate Research Fellowship (to F.W.L). The authors thank Li-Yaung Kuo for assisting with flow cytometry, Susan Durham for help with statistical analyses, and Carol Rowe for comments on the manuscript.

**Literature Cited**


**Table 7**

<table>
<thead>
<tr>
<th>Species</th>
<th>Method 1: Mean % Protein Content ± Standard error of the Mean</th>
<th>Method 2: Percent of Assemblies with Blast Hits &gt; 100 bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceratopteris</td>
<td>6.61 ± 0.02</td>
<td>1.11</td>
</tr>
<tr>
<td>Cystopteris</td>
<td>5.22 ± 0.07</td>
<td>1.78</td>
</tr>
<tr>
<td>Dipteris</td>
<td>4.82 ± 0.06</td>
<td>1.12</td>
</tr>
<tr>
<td>Plagiogryia</td>
<td>3.07 ± 0.02</td>
<td>1.90</td>
</tr>
<tr>
<td>Polypodium</td>
<td>4.01 ± 0.03</td>
<td>1.40</td>
</tr>
<tr>
<td>Pteridium</td>
<td>2.85 ± 0.03</td>
<td>1.11</td>
</tr>
</tbody>
</table>

**Fig. 4.** Scatter plots showing the relationship between proportion of different classes of genomic elements and genome size for ferns and seed plants.


